การเปรียบเทียบการค้นคืนข้อมูลบนเทคโนโลยีข้อมูลขนาดใหญ่ (ฮาดูปและแมพรีดิว) กับระบบฐานข้อมูลเชิงสัมพันธ์ (มายเอสคิวแอล) กรณีศึกษา : ชุดข้อมูลบริการสุขภาพ

รชต ทิมาสรวิชกิจ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีคอมพิวเตอร์และการสื่อสาร
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2560

Comparative Study between Big Data Technology (Hadoop / MapReduce) and Relational Database Query System (MySQL) Case study: Healthcare Dataset

Rachatha Timasornwichakit

A Thesis Submitted in Partial Fulfillment of Requirements

For the Degree of Master of Science

Department of Computer and Communication Technology,

College of Innovative Technology and Engineering

Dhurakij Pundit University

2017

หัวข้อวิทยานิพนธ์	การเปรียบเทียบการค้นคืนข้อมูลบนเทคโนโลยีข้อมูลขนาดใหญ่ (ฮาดูปและแมพรีดิว) กับระบบฐานข้อมูลเชิงสัมพันธ์ (มายเอสคิวแอล) กรณีศึกษา : ชุดข้อมูลบริการสุขภาพ
ชื่อผู้เขียน	รชต ทิมาสรวิชกิจ
อาจารย์ที่ปรึกษา	ผศ.ดร.วรพล พงษ์เพ็ชร
อาจารย์ที่ปรึกษาร่วม	รศ.ดร.นพ.วรรษา เปาอินทร์
สาขาวิชา	เทคโนโลยีคอมพิวเตอร์และการสื่อสาร
ปีการศึกษา	2559

บทคัดย่อ

งานวิทยานิพนธ์นี้มีวัตถุประสงค์ในการจัดทำดังนี้ 1) เพื่อศึกษาแนวทางที่เหมาะสมในการจัดเก็บข้อมูลบริการสุขภาพบนสถาปัตยกรรมข้อมูลขนาดใหญ่ 2) เพื่อเสริมสร้างความรู้และความเข้าใจในเทคโนโลยีข้อมูลระหว่างเทคโนโลยีข้อมูลขนาดใหญ่ (ฮาดูปและแมพรีดิว) ซึ่งมีสถาปัตยกรรมการจัดการข้อมูลและใช้หลักการทางคณิตศาสตร์ และเทคนิควิธีการสอบถามค้นคืนข้อมูลที่แตกต่างกันกับระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ (มายเอสคิวแอล) และนำมาประยุกต์ใช้ในการจัดทำสถิติข้อมูลการเจ็บป่วย 3) เพื่อเปรียบเทียบประสิทธิภาพด้านเวลาการประมวลผลและความถูกต้องแม่นยำในการค้นคืนข้อมูล

ทำการศึกษาเปรียบเทียบด้วยการวิจัยเชิงทดลอง มีขั้นตอนวิธีการศึกษาความแตกต่างของเทคนิควิธีการประมวลผลข้อมูล 2 รูปแบบ ด้วยการนำชุดข้อมูลในระบบบริการสุขภาพคัดเลือกแฟ้มผู้ป่วยนอกเป็นชุดข้อมูลตัวอย่าง ทำการแบ่งข้อมูลออกเป็น 4 ชุด มีขนาดระเบียนห้าแสน, หนึ่งล้าน, ห้าล้านและสิบล้านระเบียนตามลำดับ และสร้างชุดแบบสอบถามขึ้นจากรายงานสรุปการเจ็บป่วย พ.ศ.2557 จำนวน 2 รายงาน เป็นเครื่องมือที่นำมาใช้หาประสิทธิภาพของเวลาในการประมวลผลแบบสอบถามค้นคืนข้อมูล

ประเมินประสิทธิภาพความเร็วด้วยการวิเคราะห์ผลค่าเฉลี่ยโดยใช้สถิติ t-Test : Paired Two Sample for Means ทดสอบสมมติฐานที่คาดการณ์ว่าผลลัพธ์ของเวลาการค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างเทคโนโลยีข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์แตกต่างกัน และการประเมินผลลัพธ์ความถูกต้องและแม่นยำการค้นคืนด้วยค่าสถิติร้อยละ

ผลในการวิจัยพบว่า เมื่อข้อมูลเริ่มมีขนาดใหญ่ ที่จำนวนตั้งแต่ข้อมูล 1-10 ล้านระเบียน เทคโนโลยีข้อมูลขนาดใหญ่ (ฮาดูปและแมพรีดิว) จะใช้เวลาในการประมวลผลน้อยกว่าระบบการ

จัดการฐานข้อมูลเชิงสัมพันธ์ (มายเอสคิวแอล) และจากผลการวิเคราะห์สถิติประสิทธิภาพด้านความเร็ว สรุปผลได้ว่าในการวิเคราะห์ชุดข้อมูล 5 แสนระเบียน และ 1 ล้านระเบียน เทคโนโลยีข้อมูลขนาดใหญ่ใช้เวลาน้อยกว่าอย่างมีนัยสำคัญทางสถิติ แต่ในชุดข้อมูล 5 ล้านระเบียน และ 10 ล้านระเบียนใช้เวลาไม่ต่างกัน และประสิทธิผลของผลการสอบถามค้นคืนมีความถูกต้องแม่นยำตรงกันทุกชุดข้อมูล 100% เป็นการยอมรับสมมติฐานที่คาดการณ์ไว้ล่วงหน้า

Thesis Title	Comparative Study between Big Data (Hadoop / MapReduce) and Relational Database Query System (MySQL) Case study: Healthcare Dataset
Author	Rachatha Timasornwichakit
Thesis Advisor	Asst. Prof. Dr.Worapol Pongpech
Co-Thesis Advisor	Assoc.Prof. Dr.Wansa Paoin
Department	Computer and Communication Technology
Academic Year	2016

ABSTRACT

The objectives of this study are 1) to study the proper method to store and analyze the data of illness in Big Data architecture. 2) to compare different big data technology (Hadoop / MapReduce) and the relation database (MySQL) by using experimental research, studying the different between the groups of mathematics data compilation with different storage architecture and retrieval methods. 3) to study different methods of data processing format with the two sets of data in the healthcare system, the patient's sample data set were selected, the sample set was divided into four series with a record of five hundred thousand, one million, five million and ten million records respectively. Create a set of questionnaires and 2 reports of the illness from 2014 as a tool to be used for query processing performance in data retrieval.

Rated and speed of performance were analyzed using t-Test : Paired Two Sample for Means to compare between hadoop/mapreduce and relational databases system. The accuracy and precision of results were also measured.

The test results showed that the efficiency of hadoop/mapreduce use less time to process in one - ten million records. The hadoop/mapreduce used more time to process five hundred thousand records and one million, but one - five million records both technologies performed with no significant statistical different. The accuracy of data processing by both technologies were 100%.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้อย่างสมบูรณ์ โดยได้รับความอนุเคราะห์อย่างดียิ่งจากท่าน ผศ.ดร.วรพล พงษ์เพ็ชร และท่าน รศ.ดร.นพ.วรรยา เปาอินทร์ ผู้วิจัยขอกราบขอบพระคุณและจารึกพระคุณนี้ไว้ในความทรงจำอย่างมิรู้เลือนว่าความสำเร็จของงานวิทยานิพนธ์ฉบับนี้เกิดขึ้นได้เพราะความกรุณาของท่านที่ได้ให้คำแนะนำ ตรวจทาน และแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่เป็นอย่างดี ตลอดจนให้ความรู้ ชี้แนะแหล่งค้นคว้าหาข้อมูล หลักสำคัญในงานวิจัยและสร้างโอกาสในการศึกษาแก่ผู้วิจัยตลอดเวลาที่ได้ศึกษาภายในมหาวิทยาลัยแห่งนี้ และกราบขอบพระคุณท่านคณะกรรมการผู้ทรงคุณวุฒิทุกท่าน ที่ให้คำแนะนำที่มีประโยชน์ในงานวิทยานิพนธ์ฉบับนี้ และกราบขอบพระคุณท่านอาจารย์นุกูล พิมเสน ผู้ดูแลห้องปฏิบัติการวิศวกรรมข้อมูลขนาดใหญ่ ดำเนินการติดตั้งโปรแกรมฮาดูปเครื่องแม่ข่ายพร้อมเครือข่าย และนายนัฐพงษ์ หนูสิงห์ นักศึกษาวิศวกรรมข้อมูลขนาดใหญ่ ผู้เขียนโปรแกรมแมพรีดิวสำหรับงานทดลองในครั้งนี้

ขอขอบคุณเจ้าหน้าที่เลขานุการ คณะวิศวกรรมศาสตร์ทุกท่าน ที่ได้ให้คำแนะนำและช่วยเหลือประสานงานจนวิทยานิพนธ์ฉบับนี้เสร็จลุล่วงไปได้ด้วยดี ตลอดจนพี่น้องนักศึกษาสาขาวิชาเทคโนโลยีคอมพิวเตอร์และการสื่อสารทุกท่านที่ให้กำลังใจ และพี่น้องนักศึกษาสาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่ ที่ร่วมแบ่งปันความรู้กันในห้องทดลอง ขอให้ทุกท่านประสบความสำเร็จในหน้าที่การงานทุกท่าน

ขอขอบคุณคณะผู้บริหารของบริษัท ฮอกไกโด อินเตอร์เนชั่นแนล แฟรนไชส์ จำกัด ที่ให้ความอนุเคราะห์วันและเวลาในการเข้าศึกษาหาความรู้ให้แก่ผู้วิจัยเป็นอย่างดี

สุดท้ายนี้ผู้วิจัยขอมอบคุณความดีที่งานวิจัยนี้ ที่ผู้วิจัยคาดหวังว่าจะมีให้แก่สังคม แด่บิดาและมารดาผู้ให้กำเนิดผู้วิจัยมายังโลกใบนี้ เป็นผู้อยู่เบื้องหลังในความสำเร็จในครั้งนี้ ผู้ซึ่งปลูกฝังความคิดให้ผู้วิจัยใฝ่การศึกษา และไม่ย่อท้อต่ออุปสรรคในชีวิต และเป็นคนดีของสังคมคอยช่วยเหลือตอบแทนสังคม ตลอดจนขอขอบคุณครอบครัว พี่ น้อง และเครือญาติทุกท่านที่คอยเป็นกำลังใจให้พยายามสู้ทำเล่มวิทยานิพนธ์ฉบับนี้ให้สำเร็จจวบจนจบการศึกษาได้

รชต ทิมาสรวิชกิจ

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

บทที่1
บทนำ

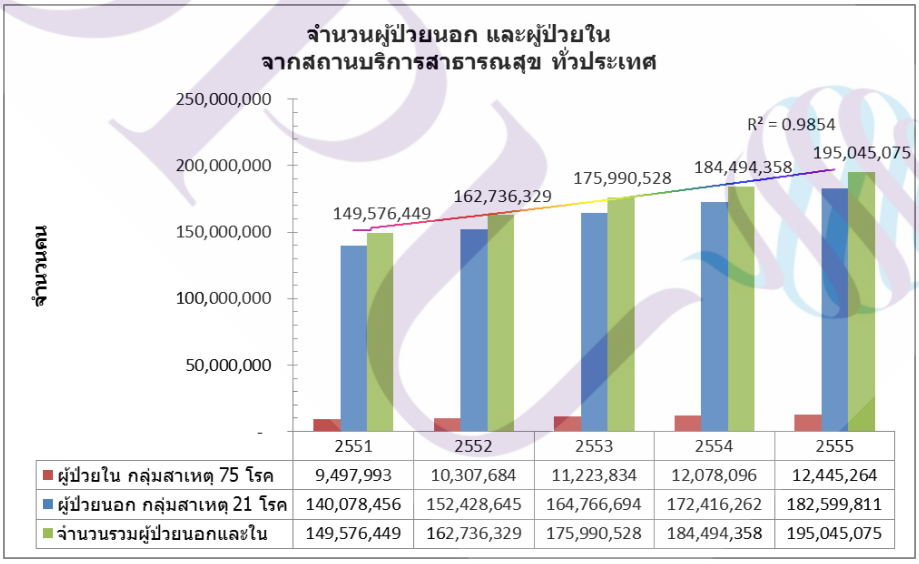
1.1 ที่มาและความสำคัญของปัญหา

กระทรวงสาธารณสุข มีหน้าที่รับผิดชอบดูแลสุขภาพของประชาชน โดยการจัดให้มีระบบบริการสุขภาพที่ครอบคลุมทั้งการส่งเสริมสุขภาพ การป้องกันโรค การรักษาพยาบาล และการฟื้นฟูสภาพ มีการจัดระบบบริการสุขภาพออกเป็นหลายระดับ ได้แก่ บริการระดับปฐมภูมิ (Primary Care) บริการระดับทุติยภูมิ (Secondary Care) และบริการระดับตติยภูมิ (Tertiary Care) โดยมุ่งหวังให้บริการแต่ละระดับมีบทบาทหน้าที่ที่แตกต่างกัน และเชื่อมโยงกันด้วยระบบส่งต่อ เพื่อให้สามารถจัดบริการสุขภาพที่มีคุณภาพ และเกิดการใช้ทรัพยากรที่มีอยู่จำกัดอย่างมีประสิทธิภาพ ตลอดจนเป็นระบบบริการสุขภาพที่มีศักยภาพรองรับปัญหาทางการแพทย์และสาธารณสุขที่มีความซับซ้อนในระดับพื้นที่ได้ (สำนักบริหารการสาธารณสุข [สบรส.], 2555, น. 1) โดยมีกรอบแนวคิด “เครือข่ายบริการที่ไร้รอยต่อ” ที่สามารถเชื่อมโยงบริการทั้ง 3 ระดับเข้าด้วยกัน และมีนโยบายให้ดำเนินการจัดเก็บรวบรวมข้อมูลการให้บริการสาธารณสุขและการแพทย์เข้าไว้ด้วยกัน มีการดำเนินการจัดทำระบบคลังข้อมูลด้านการแพทย์และสุขภาพ (Health Data Center : HDC) ใช้แฟ้มโครงสร้างมาตรฐานในการจัดเก็บ 43 และ 7 แฟ้มมาตรฐาน เพื่อนำข้อมูลการมารับบริการสาธารณสุขรวบรวมจากทุกหน่วยบริการสาธารณสุขในระดับจังหวัด เข้าสู่ศูนย์ข้อมูลระดับจังหวัด และดำเนินการประมวลผลเข้าสู่ระดับเขต และระดับกระทรวงอย่างเป็นลำดับ มีเป้าหมายให้ใช้งานได้ครอบคลุมทั้ง 76 จังหวัด ภายในปี พ.ศ.2558

สำนักงานสถิติแห่งชาติได้รายงานสถิติจำนวนผู้ป่วยใน(In Patient Department : IPD) และจำนวนผู้ป่วยนอก(Out Patient Department : OPD) จำแนกตามกลุ่มสาเหตุการเจ็บป่วย จากสถานบริการสาธารณสุข ของกระทรวงสาธารณสุขทั่วประเทศ ตั้งแต่ปี พ.ศ.2551ถึง พ.ศ.2555 จากรายงานเห็นได้ว่ามีแนวโน้มการเพิ่มขึ้นของข้อมูลผู้ป่วยใน และข้อมูลผู้ป่วยนอกทุกๆ ปีและเมื่อสถานพยาบาลในประเทศไทยทุกสถานพยาบาล ทุกระดับตั้งแต่ระดับปฐมภูมิ ทุติยภูมิ และตติยภูมิต้องจัดส่งข้อมูลเวชระเบียนผู้ป่วยที่มีการตรวจสอบคุณภาพข้อมูลเรียบร้อยแล้วนำเข้าสู่ระบบคลังข้อมูลด้านการแพทย์และสุขภาพ มีวัตถุประสงค์เพื่อการประมวลผลจัดทำรายงานและการวิเคราะห์ เช่น รายงานข้อมูลทรัพยากรสาธารณสุข รายงานจำนวนผู้ป่วย รายงานการกำเนิดและการ

เสียชีวิตทั่วประเทศเป็นประจำทุกๆ เดือน หรือรายงานการป่วย เป็นข้อมูลที่แสดงความชุกของโรคต่างๆ ที่มีผู้มารับบริการรักษาพยาบาลในสถานบริการทุกระดับ ในการรวบรวมข้อมูลการป่วยจากระบบรายงานจากฐานข้อมูลผู้ป่วยรายบุคคลมาประมวลผลและวิเคราะห์ข้อมูลเป็นประจำทุกปี เริ่มปรับเปลี่ยนจากระบบแบบรายงานเดิม รายงานผู้ป่วยนอก ตามกลุ่มสาเหตุ 21 กลุ่มโรค (รง.504) และรายงานผู้ป่วยในรายโรค ตามกลุ่มสาเหตุ 75 กลุ่มโรค (รง.505) ตั้งแต่ปี 2556 (สำนักนโยบายและยุทธศาสตร์ สำนักงานปลัดกระทรวงสาธารณสุข [สนย.สป.สธ.], 2556, น. 1-3) เพื่อนำมาจัดทำเป็นเครื่องชี้วัดทางสุขภาพ (Health Indicators) ทำให้เกิดข้อมูลเป็นจำนวนมากในฐานข้อมูลและส่งผลต่อการประมวลผลข้อมูล (Data Processing) ในการใช้เวลาการคำนวณและพื้นที่จัดเก็บข้อมูล (Storage) ที่ต้องการเพิ่มขึ้น

จากจำนวนข้อมูลผู้ป่วยในรายงานสำนักงานสถิติแห่งชาติ หากอนุมานว่าข้อมูลผู้ป่วย 1 ราย มีการบันทึกข้อมูลลงฐานข้อมูลหนึ่งระเบียน (Record) หรือเท่ากับ 1 กิโลไบต์ (KB) ยกตัวอย่างข้อมูลในปี พ.ศ.2555 มีจำนวนผู้ป่วยนอกและผู้ป่วยในรวมทั้งสิ้น 195,045,075 ราย จะมีขนาดของข้อมูลโดยประมาณเท่ากับ 186 กิกะไบต์ (GB)



	2551	2552	2553	2554	2555
ผู้ป่วยใน กลุ่มสาเหตุ 75 โรค	9,497,993	10,307,684	11,223,834	12,078,096	12,445,264
ผู้ป่วยนอก กลุ่มสาเหตุ 21 โรค	140,078,456	152,428,645	164,766,694	172,416,262	182,599,811
จำนวนรวมผู้ป่วยนอกและใน	149,576,449	162,736,329	175,990,528	184,494,358	195,045,075

ภาพที่ 1.1 กราฟจำนวนผู้ป่วยนอกและผู้ป่วยในตามการจำแนกกลุ่มโรค ปี พ.ศ.2551-2555

ที่มา: สำนักงานสถิติแห่งชาติ [online] : เข้าถึง 28 ต.ค. 2558. จาก http://service.nso.go.th/nso/web/statseries/statseries09.html

ดังนั้นเมื่อนำข้อมูลการมารับบริการสาธารณสุขรวบรวมจากทุกหน่วยบริการสาธารณสุข เข้าจัดเก็บไว้ในเซิร์ฟเวอร์ (Server) ในแต่ละเดือนมีข้อมูลเพิ่มขึ้นและขนาดใหญ่ขึ้น การประมวลผลข้อมูลจำเป็นต้องใช้ทรัพยากรรองรับการประมวลผลข้อมูล เช่นอุปกรณ์จัดเก็บข้อมูล (Storage) และหน่วยประมวลผลกลาง (Central Processing Unit : CPU) และหน่วยความจำหลัก (RAM) เพื่อนำมารองรับในการประมวลผลข้อมูลจำนวนมากเมื่อต้องดำเนินการทำรายงานทางสถิติและดัชนีชี้วัดให้แล้วเสร็จทันเวลาและทันต่อความต้องการก่อให้เกิดค่าใช้จ่ายที่ต้องจัดหาเซิร์ฟเวอร์ที่มีประสิทธิภาพในการประมวลผลสูง การประมวลผลในระบบนี้เรียกว่าการประมวลผลแบบรวมศูนย์ (Centralized Computing)

เมื่อข้อมูลถูกรวบรวมจัดเก็บในคลังข้อมูลด้านการแพทย์และสุขภาพในแต่ละปีเป็นจำนวนหลายเทราไบต์ (TB) จะเกิดเป็นข้อมูลมหาศาลหรือเรียกอีกชื่อว่าข้อมูลขนาดใหญ่ (Big Data) หากต้องการนำมาประมวลผลเปรียบเทียบทางสถิติ งานเวชสถิติ งานวิเคราะห์ทางการแพทย์โดยใช้หลักการทางคณิตศาสตร์ หรือทำดัชนีชี้วัด ปัญหาสำคัญคือ เวลาที่ใช้การประมวลผลในแต่ละงาน (Batch Processing) จะต้องใช้เวลานานอย่างหลีกเลี่ยงไม่ได้ จำเป็นต้องหาแนวทางใหม่นำมาใช้จัดการกับฐานข้อมูลขนาดใหญ่นี้โดยเฉพาะ

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาแนวทางที่เหมาะสมในการจัดเก็บและวิเคราะห์ข้อมูลการเจ็บป่วยที่รวบรวมข้อมูลการให้บริการสุขภาพโดยกระทรวงสาธารณสุข ด้วยสถาปัตยกรรมข้อมูลขนาดใหญ่
2. เพื่อศึกษาทฤษฎีการทำงานของอัลกอริทึมค้นคืนข้อมูลบนเทคโนโลยีข้อมูลขนาดใหญ่และนำมาประยุกต์ใช้ในการจัดทำสถิติข้อมูลการเจ็บป่วย
3. เปรียบเทียบประสิทธิภาพด้านเวลาและความแม่นยำถูกต้องในการค้นคืนข้อมูลบนระบบเทคโนโลยีข้อมูลขนาดใหญ่กับเทคโนโลยีระบบฐานข้อมูลเชิงสัมพันธ์

1.3 สมมติฐานของการวิจัย

1. ผลลัพธ์ของเวลาการค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างเทคโนโลยีระบบข้อมูลขนาดใหญ่กับเทคโนโลยีระบบฐานข้อมูลเชิงสัมพันธ์ มีผลลัพธ์ที่แตกต่างกัน
2. ผลลัพธ์ของความแม่นยำถูกต้องการค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างเทคโนโลยีระบบข้อมูลขนาดใหญ่กับเทคโนโลยีระบบฐานข้อมูลเชิงสัมพันธ์ มีผลลัพธ์ไม่แตกต่างกัน

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้แนวทางที่เหมาะสมในการจัดเก็บและวิเคราะห์ข้อมูลการเจ็บป่วยที่มีการรวบรวมข้อมูลการให้บริการสุขภาพโดยกระทรวงสาธารณสุข ด้วยสถาปัตยกรรมข้อมูลขนาดใหญ่

2. สามารถนำทฤษฎีกรอบการทำงานของระบบข้อมูลขนาดใหญ่ใช้กำหนดเทคนิคและจำลองรูปแบบคำหรือประโยคการค้นคืนข้อมูลเพื่อให้ได้ผลลัพธ์ที่ถูกต้องและแม่นยำได้

3. สามารถนำทฤษฎีกรอบการทำงานของระบบข้อมูลขนาดใหญ่ใช้การค้นคืนข้อมูลเพื่อจัดทำสถิติข้อมูลการเจ็บป่วยได้อย่างมีประสิทธิภาพและประสิทธิผล

4. สามารถนำประสิทธิภาพของเวลาและความเร็วที่ใช้ในการประมวลผลมาวิเคราะห์ เพื่อหาเกณฑ์การประเมินความคุ้มค่าของค่าใช้จ่ายในการประมวลผลสอบถามและการใช้ทรัพยากรได้

1.5 ขอบเขตของการวิจัย

1. ศึกษาโครงสร้างระบบสุขภาพของกระทรวงสาธารณสุขแห่งประเทศไทยเพื่อใช้ในการวางแนวทางงานวิทยานิพนธ์

2. ศึกษาเครื่องมือและกรอบการทำงานในเทคโนโลยีระบบข้อมูลขนาดใหญ่เพื่อใช้ในการนำเข้าข้อมูลจัดเก็บข้อมูลการประมวลผลข้อมูลและการแสดงรายงานข้อมูลการเจ็บป่วย

3. ศึกษาเครื่องมือและกรอบการทำงานในเทคโนโลยีระบบข้อมูลขนาดใหญ่เพื่อใช้กำหนดเทคนิคและจำลองรูปแบบคำหรือประโยคการค้นคืนข้อมูลเพื่อให้ได้ผลลัพธ์ที่ถูกต้องและแม่นยำ

4. ศึกษาเครื่องมือในเทคโนโลยีระบบข้อมูลขนาดใหญ่เพื่อใช้การวิเคราะห์ข้อมูลและจัดทำสถิติการเจ็บป่วยด้วยข้อมูลบริการสุขภาพ

5. ศึกษากระบวนการปรับปรุงประสิทธิภาพการค้นคืนข้อมูลและประเมินความคุ้มค่าของการใช้ทรัพยากร เพื่อปรับปรุงรูปแบบการคืนค้นให้ได้ผลลัพธ์ของเวลาที่ดีที่สุด

6. จัดเก็บข้อมูลเวลาและความแม่นยำถูกต้องในการค้นคืนข้อมูลระบบข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์เพื่อเปรียบเทียบและนำผลทดลองมาวิเคราะห์เชิงสถิติ

1.6 นิยามศัพท์

ผู้ป่วยใน (In Patient Department : IPD) หมายถึง ผู้ป่วยที่ลงทะเบียนเข้ารับการรักษาตัวในโรงพยาบาลหรือสถานพยาล โดยได้รับการวินิจฉัยและคำแนะนำจากแพทย์ต้องนอนพักรักษาตัวในโรงพยาบาลตั้งแต่ 6 ชั่วโมงขึ้นไป

ผู้ป่วยนอก (Out Patient Department : OPD) หมายถึง ผู้ป่วยที่ลงทะเบียนเข้ารับการรักษาที่โรงพยาบาลหรือสถานพยาบาล โดยไม่ต้องนอนพักรักษาตัวในโรงพยาบาลผู้ป่วยสามารถกลับบ้านได้ในวันที่เข้ารับการรักษา

เวชสถิติ (Medical Statistics) หมายถึง สถิติทางการแพทย์ เป็นการเก็บรวบรวมข้อมูลทางการแพทย์เพื่อการนำเสนอข้อมูลทางการแพทย์ นำมาใช้ในงานวิเคราะห์ทางการแพทย์โดยใช้หลักการทางคณิตศาสตร์ สถิติ และนำผลการวิเคราะห์ข้อมูลมาสรุปเพื่อนำไปใช้ในการจัดทำรายงานทางการแพทย์ เพื่อการพัฒนา และการศึกษาวิจัยทางการแพทย์

ระบบสุขภาพ (Health System) หมายถึง ระบบที่มุ่งหวังให้ประชาชนมีสุขภาวะกายและจิตที่ดี ผ่านกระบวนการสร้างเสริมสุขภาพ ป้องกันโรค รักษาโรค ฟื้นฟูการทำงานของร่างกายและกระบวนการสร้างความแข็งแกร่งและความพร้อมของสาธารณสุขในการรับมือโรคติดต่อ โรคไม่ติดต่อ และภัยพิบัติ นอกเหนือการตอบสนองความคาดหวังของประชาชนแล้ว ระบบสุขภาพที่ดีควรเห็นคุณค่าและศักดิ์ศรีในความเป็นมนุษย์ของทุกคน ยึดมั่นในหลักศีลธรรม คุณธรรม จริยธรรมในการดำเนินการ และให้ความเท่าเทียมด้านสิทธิประโยชน์แก่ประชาชนทุกกลุ่มอย่างเหมาะสม ระบบสุขภาพที่สมบูรณ์จึงมีประชาชนเป็นศูนย์กลาง แวดล้อมด้วยกิจกรรมที่มุ่งส่งเสริม ฟื้นฟู และธำรงสุขภาพของประชาชน

ระบบบริการสุขภาพ (Health Care System) หมายถึง ระบบบริการสุขภาพเป็นส่วนหนึ่งของระบบสุขภาพ เป็นระบบบริการต่างๆ ที่จัดขึ้นเพื่อเป็นการดูแลสุขภาพของประชาชนทั้งทางด้านการสร้างเสริมสุขภาพ การควบคุมป้องกันโรค การรักษาพยาบาล และการฟื้นฟูสมรรถภาพ ที่เป็นแบบผสมผสานหรือเฉพาะด้าน เฉพาะเรื่อง เป้าประสงค์ของระบบบริการสุขภาพที่ดีคือ ความเป็นธรรมในการร่วมจ่ายค่าบริการสุขภาพ การให้บริการสุขภาพเพื่อส่งมอบการบริการเพื่อส่งเสริมสุขภาพ, การให้บริการในลักษณะสาธารณสุขมูลฐาน, การให้บริการผู้ป่วยซ้ำซ้อน, การให้บริการสุขภาพเพื่อครอบครัว, การให้บริการระบบส่งต่อ, การให้บริการเฉพาะกลุ่มประชากร ทั้งนี้หากเป็นส่วนที่รัฐจัดขึ้น สนับสนุนให้จัดขึ้น หรืออยู่ภายใต้การควบคุมกำกับของรัฐ เพื่อประชาชนโดยทั่วไป ในอดีตจะเรียกในส่วนนี้ว่า บริการสาธารณสุข

ข้อมูลบริการสุขภาพ (Health Care Data) หมายถึง ข้อมูลที่ได้จากการสำรวจสุขภาพ, ข้อมูลทะเบียนสถิติชีพ, ข้อมูลการเฝ้าระวังโรคและการบาดเจ็บ, ข้อมูลทะเบียนโรค, ข้อมูลบริการสุขภาพจากสถานพยาบาล, แบบบันทึกการรักษาทางการแพทย์ของผู้ป่วย, ข้อมูลเพื่อใช้รวบรวมนำมาวิเคราะห์นำไปใช้เพื่อการบริหารค่าใช้จ่ายและทรัพยากรเพื่อใช้ในการดูแลสุขภาพของประชาชนหรือจะกล่าวสั้นๆ ได้ว่า ข้อมูลที่เกี่ยวข้องกับการเจ็บป่วยที่ถูกจัดเก็บไว้อย่างเป็นระบบ

และระเบียบขั้นตอนแบบแผนในการรวบรวมและจัดเก็บ และมีการตรวจสอบคุณภาพข้อมูลอย่างมีระบบเป็นขั้นตอนก่อนนำเข้าระบบคลังข้อมูลสุขภาพ

ฐานข้อมูลเชิงสัมพันธ์ (Relational Database) หมายถึง ข้อมูลที่ถูกจัดเก็บข้อมูลในรูปของตาราง (Table) หลายๆ ตารางที่มีความสัมพันธ์กัน ในแต่ละตารางแบ่งออกเป็นแถว และในแต่ละแถวจะแบ่งออกเป็นคอลัมน์ (Column) ในทางทฤษฎีใช้แบบจำลองโมเดลเชิงสัมพันธ์ (Relational Database Model) โดยใช้หลักพื้นฐานทางคณิตศาสตร์

ข้อมูลขนาดใหญ่ (Big Data) หมายถึง ปริมาณของจำนวนข้อมูลที่มีมหาศาล ทั้งข้อมูลที่มีโครงสร้างและไม่มีโครงสร้างที่เกิดจากข้อมูลการปฏิบัติงานทุกๆ วันของธุรกิจ ข้อมูลขนาดใหญ่มีคุณลักษณะ 3 ประการ คือ 1.ปริมาณ (Volume) ข้อมูลที่รวบรวมจากแหล่งข้อมูลหลากหลายแหล่งและหลากหลายประเภท 2.หลากหลาย (Variety) ข้อมูลมีลักษณะรูปแบบของข้อมูลที่แตกต่างกัน 3. รวดเร็ว (Velocity) ข้อมูลเกิดขึ้นได้ตลอดเวลาและสามารถรวบรวมข้อมูลได้อย่างทันที งานวิทยานิพนธ์นี้หมายถึงข้อมูลบริการสุขภาพที่มีการจัดเก็บข้อมูลแบบเชิงสัมพันธ์อย่างมีโครงสร้าง

ระบบสอบถามค้นคืนข้อมูลคือ กระบวนการดึงหรือค้นหาข้อมูลย้อนหลังจากที่มีการจัดเก็บไว้ ซึ่งอาจจะหมายถึงการจัดเก็บแบบที่มีโครงสร้างหรือไม่มีโครงสร้างหรือกึ่งโครงสร้างเพื่อนำออกเป็นสารสนเทศตามความต้องการของผู้ใช้

บทที่ 2
ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

งานวิทยานิพนธ์ฉบับนี้นำเสนอแนวคิดการเปรียบเทียบประสิทธิภาพของเวลาในการค้นคืนข้อมูลด้วยภาษาสอบถามเชิงโครงสร้างและผลลัพธ์ความถูกต้องในการประมวลผลข้อมูลระหว่างเทคโนโลยีข้อมูลขนาดใหญ่ (ฮาดูปและแมพรีดิว) กับระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ (มายเอสคิวแอล) ซึ่งมีการใช้ทฤษฎีที่เกี่ยวข้อง 6 ทฤษฎีดังนี้

ก. ทฤษฎีฐานข้อมูลและการจัดการฐานข้อมูลแบบเชิงสัมพันธ์
ข. ทฤษฎีการค้นคืนด้วยภาษาสอบถามเชิงโครงสร้างเอสคิวแอล
ค. ทฤษฎีกระบวนการสอบถามข้อมูล
ง. ทฤษฎีระบบจัดเก็บแบบกระจายข้อมูลบนกรอบการทำงานฮาดูป
จ. ทฤษฎีระบบการประมวลผลแบบขนานด้วยเทคนิคแมพรีดิว
ฉ. ทฤษฎีการวิเคราะห์ข้อมูลและสถิติการวิเคราะห์ข้อมูล

2.1 ทฤษฎีที่เกี่ยวข้อง

ผู้วิจัยมีแนวทางการเขียนงานทฤษฎีที่เกี่ยวข้องโดยการนำวัตถุประสงค์และขอบเขตของงานวิจัยเป็นหัวข้อหลักเพื่อทบทวนวรรณกรรม ผู้วิจัยขอนำเสนอเป็นหัวข้อดังนี้

2.1.1 การทบทวนวรรณกรรมระบบสุขภาพของกระทรวงสาธารณสุขแห่งประเทศไทย

การทบทวนวรรณกรรมระบบบริการสุขภาพจะทำให้ผู้จัดทำวิจัยฉบับนี้เข้าใจและรู้จักกับวงการสาธารณสุขแห่งประเทศไทยให้ดียิ่งขึ้น ด้วยการศึกษาโครงสร้างระบบสุขภาพของกระทรวงสาธารณสุขของประเทศไทย โดยการศึกษาค้นคว้ากับเอกสารงานวิจัย เอกสารงานวิชาการ และแผนนโยบายของกระทรวงสาธารณสุข และบทความต่างๆ ของหน่วยงานที่เกี่ยวข้องกับกระทรวงสาธารณสุข ทำความเข้าใจเรื่องราวในวงการสาธารณสุขของประเทศไทยในปัจจุบัน อีกทั้งยังศึกษาแนวทางในการจัดเก็บและวิเคราะห์ข้อมูลการเจ็บป่วยที่ได้มีการรวบรวมข้อมูลการให้บริการสุขภาพและการค้นคว้าข้อมูลนำมากลั่นกรองประเด็นปัญหาด้วยแนวทางการคิดเชิงวิเคราะห์ ผู้เชี่ยวชาญด้านการคิดเชิงวิเคราะห์กล่าวไว้ว่า หลักการคิดเชิงวิเคราะห์โดยพื้นฐานเกี่ยวข้องกับการจำแนกแจกแจงข้อมูลออกเป็นส่วนๆ ตรวจสอบอย่างละเอียด หาความสัมพันธ์เชิง

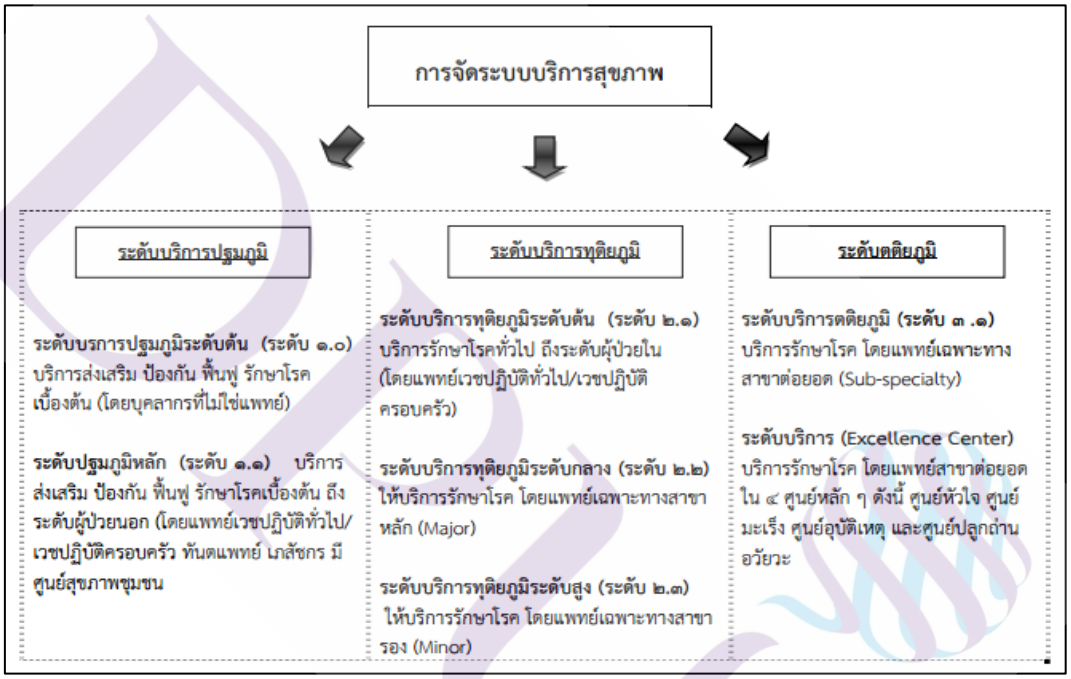
เหตุผล เพื่อทำความเข้าใจ ก่อนที่จะประเมินและตัดสินใจเกี่ยวกับเรื่องนั้น เราต้องเรียนรู้ที่จะมองสิ่งนั้น “ตามเนื้อผ้า” หรือ “มองสิ่งที่เห็นให้เป็นเช่นที่มันเป็นอยู่” เพื่อให้รู้ว่าเรื่องนั้นเกี่ยวกับอะไร ไม่ด่วนสรุปหรือด่วนแสดงทัศนะใดๆ เกี่ยวกับเรื่องนั้น เป็นการแสดงความปรารถนาสืบสาวเรื่องนั้นในระดับลึกลงกว่าเดิม โดยพยายามทำความเข้าใจ หาที่มาที่ไปเกี่ยวกับเรื่องนั้น เชื่อมโยงความสัมพันธ์เชิงเหตุผล เพื่อให้รู้ข้อเท็จจริงก่อนที่จะดำเนินการใดๆ ลงไป (เกรียงศักดิ์ เจริญวงศ์ศักดิ์, 2553, น. 74-75) การคิดเชิงวิเคราะห์จะนำมาใช้ศึกษาทบทวนงานวรรณกรรมระบบสุขภาพ เพื่อค้นหาปัญหาของกระทรวงสาธารณสุข สามารถสรุปความเป็นมาในสิ่งที่เกิดขึ้น ตั้งแต่ก่อนเริ่มระบบบริการสุขภาพและหลังเริ่มระบบบริการสุขภาพ ได้พอสังเขปดังต่อไปนี้

ก่อนเริ่มระบบบริการสุขภาพ ประเทศไทยเริ่มมีการพัฒนารหัสกลุ่มโรคเพื่อให้การบันทึกข้อมูลเป็นไปตามมาตรฐานและหลักการสากลโลกขององค์การอนามัยโลก (Word Health Organization หรือ WHO) ในปี พ.ศ.2543 โดยเริ่มจากสำนักนโยบายและยุทธศาสตร์ สำนักงานปลัดกระทรวงสาธารณสุข เริ่มพัฒนาบัญชีจำแนกโรคฉบับประเทศไทย ฉบับที่ 1 เพื่อใช้เป็นแนวทางในการให้รหัสกลุ่มโรค ทำให้การบันทึกข้อมูลผู้ป่วยที่เข้ารับบริการในสถานพยาบาลในสังกัดกระทรวงสาธารณสุขมีมาตรฐานเป็นไปในแนวทางเดียวกันมีวัตถุประสงค์เพื่อนำข้อมูลการวินิจฉัยทางการแพทย์แปรเปลี่ยนมาเป็นสารสนเทศ สำหรับนำมาวิเคราะห์และวางแผนทางด้านงานสาธารณสุข ดูแลสุขภาพของประชากรชาวไทย อีกทั้งในเรื่องการจัดสรรทรัพยากรทั้งทางด้านบุคลากรและงบประมาณให้เหมาะสมและเพียงพอ ถูกต้องตรงต่อความต้องการในแต่ละส่วนงานและหรือหน่วยงานที่เกี่ยวข้องด้านสาธารณสุข ก่อให้เกิดการพัฒนาระบบสาธารณสุขของประเทศไทยอย่างยั่งยืนซึ่งมีการพัฒนามีการดำเนินการจัดทำและปรับปรุง (Revision) รหัสมาแล้วหลายครั้งหลายฉบับ อย่างต่อเนื่องจนถึง พ.ศ.2559 เป็นฉบับทบทวนและปรับปรุงใหม่ปี พ.ศ.2557

ในปี พ.ศ. 2544 แพทย์สภาได้มีการประกาศใช้คู่มือค่าธรรมเนียมแพทย์ เป็นแนวทางแต่มิใช่เป็นการบังคับอย่างมีบทลงโทษ ซึ่งถือว่าเป็นการใช้รหัส ICD-10-TM ส่วนหัตถการและการผ่าตัดอย่างเป็นทางการครั้งแรก และในระยะเวลาต่อมาได้มีโรงพยาบาลภาครัฐและเอกชนบางส่วนใช้รหัส ICD-10-TM ในกรณีที่ต้องการเก็บข้อมูลการผ่าตัดและหัตถการที่มีรายละเอียดมากกว่า ICD-9-CM และได้เริ่มมีการประกาศใช้รหัส ICD-10-TM for PCU Procedures ทั่วประเทศไทยเป็นครั้งแรกในเดือนพฤษภาคม 2553 (สำนักนโยบายและยุทธศาสตร์ สำนักงานปลัดกระทรวงสาธารณสุข [สนย.สป.สธ.], 2553, น. 1)

กระทรวงสาธารณสุข ได้ออกแผนพัฒนาระบบบริการสาธารณสุขในส่วนภูมิภาค พ.ศ. 2554 เพื่อจัดให้มีระบบบริการสุขภาพที่ครอบคลุมทั้งการส่งเสริมสุขภาพ การป้องกันโรค การรักษาพยาบาล และการฟื้นฟูสภาพ มีการจัดระบบบริการสุขภาพออกเป็น 3 ระดับ ได้แก่ บริการ

ระดับปฐมภูมิ (Primary Care) บริการระดับทุติยภูมิ (Secondary Care) และบริการระดับตติยภูมิ (Tertiary Care) โดยมุ่งหวังให้บริการแต่ละระดับมีบทบาทหน้าที่ที่แตกต่างกัน และเชื่อมโยงกันด้วยระบบส่งต่อ เพื่อให้สามารถจัดบริการสุขภาพที่มีคุณภาพ โดยมีกรอบแนวคิด “เครือข่ายบริการที่ไร้รอยต่อ” ที่สามารถเชื่อมโยงบริการทั้ง 3 ระดับเข้าด้วยกันตั้งแต่ระดับปฐมภูมิ ทุติยภูมิ และตติยภูมิ ในแต่ละจังหวัดจะต้องมีเครือข่ายบริการระดับจังหวัดที่สามารถรองรับการส่งต่อตามมาตรฐานระดับจังหวัดได้อย่างสมบูรณ์ และให้สำนักงานสาธารณสุขจังหวัดเป็นผู้รับผิดชอบการตั้งศูนย์การแพทย์ชุมชนเมืองและเป็นศูนย์กลางประสานงานแม่ข่ายระดับปฐมภูมิ



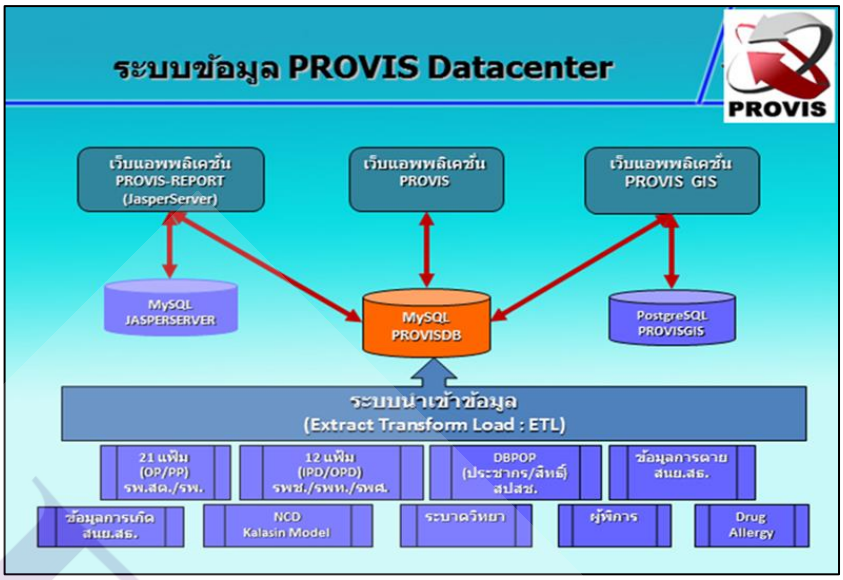
ภาพที่ 2.1 แผนภาพการจัดระดับบริการสุขภาพ กระทรวงสาธารณสุข

ที่มา: ศูนย์มาตรฐานรหัสและข้อมูลสุขภาพแห่งชาติ [online] : เข้าถึง 28 ต.ค. 2558. จาก http://www.thcc.or.th/

หลังเริ่มระบบบริการสุขภาพ กระทรวงสาธารณสุขได้มีการประกาศนโยบายตั้งแต่ปี พ.ศ.2555 ให้ดำเนินการจัดเก็บรวบรวมข้อมูลการให้บริการสาธารณสุขและการแพทย์เข้าไว้ด้วยกัน โดยกำหนดให้มีการดำเนินการจัดทำระบบคลังข้อมูลด้านการแพทย์และสุขภาพ โดยใช้แฟ้มโครงสร้างมาตรฐานในการจัดเก็บ 43 และ 17 แฟ้มมาตรฐาน เพื่อนำข้อมูลการมารับบริการ

สาธารณสุขรวบรวมจากทุกหน่วยบริการสาธารณสุขในระดับจังหวัด เข้าสู่ศูนย์ข้อมูลระดับจังหวัด และดำเนินการประมวลผลเข้าสู่ระดับเขต และระดับกระทรวง เพื่อรวบข้อมูลเข้าไว้ด้วยกันเป็นศูนย์ข้อมูลคลังสุขภาพ (Health Data Center) หรือศูนย์ข้อมูลข่าวสารและสารสนเทศสุขภาพ กระทรวงสาธารณสุข โดยมีเป้าหมายให้มีการใช้งานได้ครอบคลุมทั้ง 76 จังหวัด ภายในปี พ.ศ.2558 มีแบบแผนการทดลอง (Pilot Test) ใช้ในสถานพยาบาลบางแห่งก่อนปรับเปลี่ยนรูปแบบการใช้งานทั่วประเทศ ผู้วิจัยดำเนินการศึกษาค้นคว้าหาข้อมูลจากเนื้อหาข่าวสารในเว็บไซต์ที่เกี่ยวข้องกับระบบสุขภาพบนอินเตอร์เน็ท ทำให้ทราบว่างานเขียนหรือวรรณกรรมที่เกี่ยวข้องด้านเนื้อหาระบบข้อมูลสุขภาพมาจากหน่วยงานในระบบสุขภาพ เช่น หน่วยงานโรงพยาบาล สถานพยาบาลของกระทรวงสาธารณสุข หน่วยงานวิชาการภายในของกระทรวงสาธารณสุข และรวมถึงนักศึกษาสาขางานสาธารณสุขและการแพทย์ โดยขอแบ่งช่วงเวลาออกเป็น 2 ช่วงเวลา จากการประกาศใช้งานโปรแกรมจัดเก็บข้อมูลส่วนกลาง (Data Center) เป็นหลัก

ช่วงเวลาที่ 1) พ.ศ.2555-2558 มีการกล่าวถึงการเก็บข้อมูลสุขภาพในระบบแฟ้มข้อมูล 43 แฟ้ม และ 17 แฟ้มมาตรฐาน ด้วยโปรแกรมฐานข้อมูล (PROVIS) ให้เป็นศูนย์กลางของข้อมูลระดับปฐมภูมิ เป็นระบบฐานข้อมูลสาธารณสุขจังหวัด อีกทั้งยังมีโปรแกรมระบบงานโรงพยาบาลส่งเสริมสุขภาพตำบลและศูนย์สุขภาพชุมชน (JHCIS) ที่ใช้รวบรวบข้อมูลระดับทุติยภูมิ และตติยภูมิ และส่งต่อข้อมูลเข้าสู่ระดับปฐมภูมิหรือจะกล่าวได้ว่าจาก JHCIS รวบรวมเข้าสู่ PROVIS เริ่มมีการใช้งานโปรแกรมอย่างแพร่หลายเมื่อผู้วิจัยได้สืบค้นข้อมูลเพิ่มเติมเพื่อต้องการทราบว่าฐานข้อมูลเหล่านั้นมีการจัดการด้วยโปรแกรมฐานข้อมูลใด เป็นโปรแกรมที่รวบรวมข้อมูลและส่งต่อข้อมูลให้กับหน่วยงานในระดับเขต และระดับกระทรวง พบว่าโปรแกรม PROVIS และ JHCIS ใช้โปรแกรมฐานข้อมูลโอเพ่นซอร์สมายเอสคิวแอล (MySQL) และจัดเก็บฐานข้อมูลแบบเชิงสัมพันธ์ (Relational Database) ซึ่งได้รับการสนับสนุนการเขียนโปรแกรมจากศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร สำนักงานปลัดกระทรวงสาธารณสุขและศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (เนคเทค) ที่มีวัตถุประสงค์เพื่อให้ใช้งานได้ครอบคลุมข้อมูลของโรงพยาบาลทั้งหมดอย่างเป็นระบบ สะดวกรวดเร็ว และมีประสิทธิภาพตอบสนองต่อการจัดทำรายงานต่างๆ ทั้งในระดับเขตระดับกระทรวง และเพื่อให้ได้สารสนเทศตรงตามความต้องการ โดยสามารถใช้งานได้หลายระบบปฏิบัติการ (Operating System) และถูกออกแบบพัฒนาโปรแกรมให้สามารถแลกเปลี่ยนข้อมูลระหว่างโรงพยาบาลผ่านอินเตอร์เน็ตโดยใช้ระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ (RDBMS) และพัฒนาด้วยโปรแกรมภาษาจาวา (Java)



ภาพที่ 2.2 รูปแบบการเชื่อมโยงข้อมูลนำเข้าและส่งออกข้อมูลคลังข้อมูลสุขภาพ PROVIS

ที่มา: สำนักงานปลัดกระทรวงสาธารณสุข [online] : เข้าถึง 8 พ.ย. 2558. จาก http://slideplayer.in.th/slide/2216621/

ช่วงเวลาที่ 2) พ.ศ.2558-2559 กระทรวงสาธารณสุขมีการพัฒนาระบบบริการสุขภาพหรือระบบข้อมูลสุขภาพอย่างต่อเนื่องจนครบทุกสถานพยาบาลในสังกัดกระทรวงสาธารณสุขมีการปรับเปลี่ยนโครงสร้างการเก็บข้อมูลสุขภาพในระบบแฟ้มข้อมูล 43 แฟ้ม และ 7 แฟ้มมาตรฐานหรือ 50 แฟ้มมาตรฐาน กระทรวงสาธารณสุขมีนโยบายการปรับเปลี่ยนโครงสร้างของระบบฐานข้อมูลเพื่อให้บุคลากรมีเวลาในการดูแลให้บริการผู้ป่วยเพิ่มขึ้น และลดภาระการคีย์ข้อมูลของหน่วยงานในระดับปฐมภูมิลงเพื่อให้เกิดประสิทธิภาพในการบริหารจัดการข้อมูลแต่ยังคงไว้ซึ่งข้อมูลที่สามารถนำมาใช้บริหารงานนโยบายและสามารถกำหนดกลยุทธ์หรือยุทธศาสตร์ในการบริหารจัดการ

ข้อมูลในกระทรวงสาธารณสุขมีหลากหลายรูปแบบและมาจากหลากหลายแห่ง สารสนเทศคือการประมวลผลข้อมูล ข้อมูลและสารสนเทศมีความสำคัญมากในการบริหารงานทุกระดับ แต่ข้อมูลและสารสนเทศจะมีการเปลี่ยนแปลงอยู่ตลอดเวลาการเปลี่ยนแปลงของสภาวะสังคมและโลก ซึ่งการเปลี่ยนแปลงที่เกิดขึ้นมีผลกระทบต่อสุขภาพของประชาชนสังคมโลก ประเทศไทย และภายในกระทรวงสาธารณสุข ซึ่งรับผิดชอบดูแลสุขภาพของประชาชนชาวไทย การประยุกต์ใช้ข้อมูลข่าวสารสนเทศเชิงกลยุทธ์จะมีความสำคัญอย่างยิ่งต่อการเปลี่ยนแปลงองค์กร

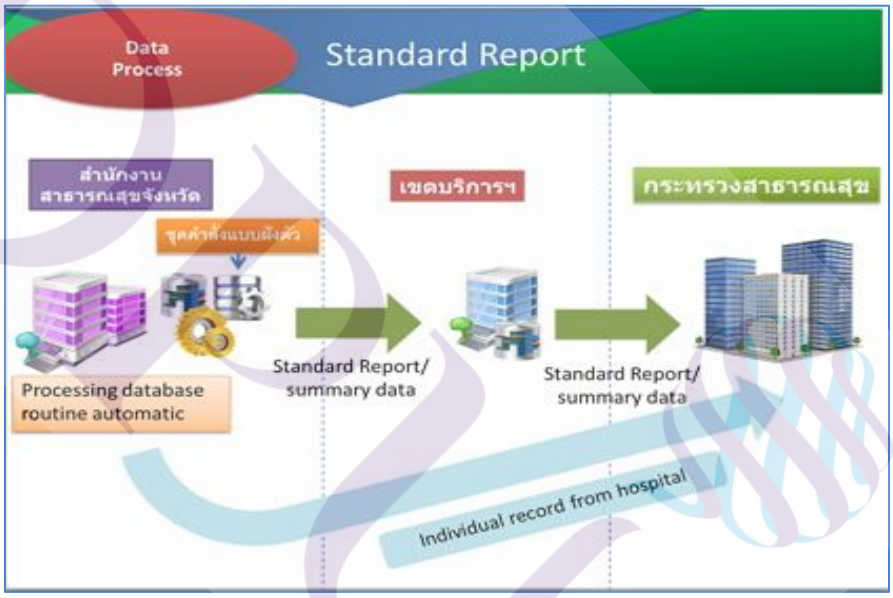
ผู้บริหารระดับสูง และการวางแผนเชิงกลยุทธ์ซึ่งจะแตกต่างกับการวางแผนปฏิบัติการตามปกติ การประยุกต์ใช้ข้อมูลสารสนเทศเชิงกลยุทธ์ในระบบงานสาธารณสุข จะดำเนินการได้อย่างชัดเจนในงาน 3 ด้าน คือ ด้านการบริหาร ด้านวิชาการ และด้านบริการ ซึ่งจะสอดคล้องกับการดำเนินงานในระบบงานสาธารณสุข (เมธี จันท์จารุภรณ์, 2555, น. 2) จะเห็นได้ว่าตามแนวคิดของท่านเมธี สารสนเทศเป็นส่วนสำคัญที่จะทำให้ผู้บริหารระดับสูงกระทรวงสาธารณสุขตัดสินใจได้ถูกต้อง หากมีสารสนเทศที่สามารถใช้ได้ทันตามความต้องการ ถูกต้องและทันสมัย

ผนวกกับแนวคิดที่เห็นว่าข้อมูลที่นำมากลั่นกรองให้ผู้บริหารจะต้องมีคุณภาพด้วย สารสนเทศเป็นทรัพยากรสำคัญที่ต้องมีการบริหารจัดการให้เกิดประโยชน์อย่างคุ้มค่า ส่วนของการเก็บรวบรวมข้อมูล การรักษาคุณภาพข้อมูล การวิเคราะห์ข้อมูล หรือการทำให้ผู้บริหารมีความเข้าใจ รู้จักใช้ระบบและสารสนเทศให้เป็นประโยชน์อย่างเต็มที่ ยังนับได้ว่าน้อยกว่ามากโดยเปรียบเทียบการลงทุนในส่วนของระบบคอมพิวเตอร์และการสื่อสารยังมิได้ถูกจัดการให้เกิดประโยชน์อย่างคุ้มค่าหรือเต็มศักยภาพ นอกจากจะพิจารณาความต้องการเกี่ยวกับระบบสารสนเทศของผู้บริหารหรือผู้ใช้แล้ว ยังมีความจำเป็นที่จะพิจารณาเรื่องของข้อมูลในระบบสารสนเทศว่ามีที่มาอย่างไร หากระบบสารสนเทศไม่มีข้อมูลที่จำเป็นต้องใช้เก็บอยู่ในฐานข้อมูลแล้ว ระบบย่อมไม่สามารถสร้างสารสนเทศได้ทันเวลาตามความต้องการ อีกทั้งยังต้องคำนึงถึงคุณภาพของข้อมูลซึ่งจะมีผลกระทบโดยตรงต่อคุณภาพของสารสนเทศที่สร้างขึ้นจากข้อมูลนั้น รวมทั้งการวิเคราะห์ข้อมูลเพื่อสร้างเป็นสารสนเทศด้วย (สุชาดา กีระนันทน์, 2544, น. 94)

ในระบบบริการสุขภาพกระทรวงสาธารณสุขมีนโยบายให้จัดทำข้อมูลบริการสุขภาพให้เป็นที่เชื่อถือได้ ในการจะนำมากลั่นกรองเป็นระบบสารสนเทศโดยมีขั้นตอนการผ่านกระบวนการตรวจสอบคุณภาพตามหลักและวิธีการของกระทรวงสาธารณสุข การตรวจสอบคุณภาพข้อมูลผู้ป่วย เป็นเรื่องที่ควรดำเนินการโดยสม่ำเสมอเป็นระยะ เช่น ดำเนินการทุก 3-4 เดือน ปีละ 3-4 ครั้ง เพื่อวัดคุณภาพข้อมูล ให้รู้สถานการณ์ที่เป็นปัญหาอันทำให้เกิดข้อมูลคุณภาพต่ำ เพื่อหาหนทางแก้ไขปัญหา เพื่อให้ข้อมูลมีคุณภาพดีขึ้นและพัฒนาให้ดีขึ้นอย่างต่อเนื่อง สมทบกับผู้เชี่ยวชาญด้านข้อมูลระบบสุขภาพกล่าวว่า ความสำคัญของคุณภาพข้อมูลในระบบบริการสุขภาพข้อมูลในระบบบริการสุขภาพสามารถใช้ประโยชน์ได้มากมาย โดยเกิดประโยชน์ต่อ ประชาชน สถานพยาบาล และกระทรวงสาธารณสุข จึงถือว่าข้อมูลมีความสำคัญอย่างยิ่ง หากข้อมูลนี้มีคุณภาพต่ำ เช่น มีข้อมูลไม่ครบถ้วน หรือมีข้อมูลที่ผิดพลาดจำนวนมาก ก็จะไม่สามารถนำไปใช้ประโยชน์ได้ตามที่ควรจะเป็น ผู้ที่เกี่ยวข้องทุกระดับต้องเข้าใจความสำคัญและมีหน้าที่บันทึกข้อมูลให้มีคุณภาพ มีระบบตรวจสอบคุณภาพข้อมูล และมีกลไกควบคุมคุณภาพข้อมูลให้มีคุณภาพสูงสุด ลักษณะของข้อมูลคุณภาพดี ข้อมูลคุณภาพดีมีลักษณะที่สำคัญ 4 ลักษณะดังนี้

1. ครบถ้วน มีข้อมูลการให้บริการทุกราย มีข้อมูลทุกด้านที่จำเป็น
2. ถูกต้อง ไม่มีข้อผิดพลาด เชื่อถือได้
3. ละเอียด ไม่กำกวม ชัดเจน แยกแยะประเภทต่างๆ ได้
4. ทันสมัย เป็นข้อมูลปัจจุบัน ส่งมาภายในเวลาที่กำหนด

การจัดการให้ข้อมูลมีคุณภาพดี เป็นหน้าที่ของทีมงานที่กำกับดูแลระบบข้อมูล โดยต้องมีกระบวนการตรวจสอบคุณภาพข้อมูลอย่างสม่ำเสมอเป็นระยะปีละ 2-4 ครั้ง โดยถ้าตรวจพบว่าข้อมูลมีปัญหาด้านคุณภาพ ก็ต้องมีกิจกรรมแก้ไขและพัฒนาคุณภาพให้ดีขึ้นอย่างต่อเนื่อง และหากมีข้อมูลคุณภาพดีแล้ว ก็ต้องมีระบบควบคุมคุณภาพให้ดีอย่างต่อเนื่อง (ศูนย์มาตรฐานรหัสและข้อมูลสุขภาพแห่งชาติ [ศมสท.], 2558, น. 6)



ภาพที่ 2.3 รูปแบบการจัดส่งข้อมูลศูนย์ข้อมูลสุขภาพ (Health Data Center)

ที่มา: โรงพยาบาลส่งเสริมสุขภาพตำบลบ้านชวน [online] : เข้าถึง 10 พ.ย. 2558. จาก http://bn210.blogspot.com/2014/10/2558.html

กระทรวงสาธารณสุขเริ่มมีการปรับเปลี่ยนระบบโปรแกรมฐานข้อมูล PROVIS เป็นระบบโปรแกรมศูนย์ข้อมูลคลังสุขภาพ (Health Data Center) บนระบบคลาวน์หรือ HDC On Cloud เป็นการพัฒนาอย่างต่อเนื่อง เพื่อการรวบรวมข้อมูลที่ได้จากระดับจังหวัด และระดับเขต จากนั้นเข้าสู่ระดับกระทรวง ทำให้สามารถดำเนินการประมวลผลออกเป็นดัชนีชี้วัด และรายงานที่จำเป็นได้

แบบทันทีที่ต้องการ ผู้วิจัยได้ทำการศึกษาพบว่าระบบดังกล่าวยังคงใช้ฐานข้อมูลมายเอสคิวแอลเหมือนเช่นเดิม และข้อมูลการบริการสุขภาพมีการประมวลผลบนเซิร์ฟเวอร์ประจำตามกำหนดเวลาที่ผู้ดูแลระบบศูนย์ข้อมูลส่วนกลางเป็นผู้ดำเนินการ

ในปี พ.ศ.2559 ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร กระทรวงสาธารณสุข มีโครงการเพื่อพัฒนาระบบเทคโนโลยีสารสนเทศและการสื่อสาร ตามกรอบยุทธศาสตร์เทคโนโลยีสารสนเทศสุขภาพ มีกลยุทธ์การพัฒนาคลังข้อมูลสุขภาพ กำหนดรูปแบบการบริหารจัดการคลังข้อมูลระบบบริการสุขภาพในระบบข้อมูลขนาดใหญ่ (Big Data Management in Healthcare System) เพื่อให้มีความเหมาะสมในการใช้งานให้กับหน่วยงานแต่ละระดับ (ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร สำนักงานปลัดกระทรวงสาธารณสุข [ศทส.สป.สธ.], 2559, น. 58)

2.1.2 ทฤษฎีฐานข้อมูลและการจัดการฐานข้อมูลแบบเชิงสัมพันธ์

โอภาส เอี่ยมสิริวงศ์ (2558, น. 37-40) กล่าวว่าในลักษณะของการจัดการข้อมูล (Database Management) เกิดจากการรวบรวมจัดเก็บข้อมูลหรือเอกสารต่างๆ ที่รวบรวมไว้ในแฟ้มเอกสาร และนำไปจัดเก็บไว้ในตู้เก็บเอกสาร ซึ่งในปัจจุบันสามารถรวบรวมจัดเก็บไว้ในฐานข้อมูล จากข้อจำกัดของระบบเอกสาร เช่น ปัญหาด้านความซ้ำซ้อนของข้อมูล ความไม่ยืดหยุ่น และความไม่คล่องตัวในหลายๆ ด้าน จึงเกิดเทคโนโลยีฐานข้อมูล

ฐานข้อมูล (Database) คือ การจัดเก็บข้อมูลอย่างมีระบบ ซึ่งผู้ใช้สามารถเรียกใช้ข้อมูลในลักษณะต่างๆ ได้ เช่น การเพิ่มเติมข้อมูล การเรียกดูข้อมูล การแก้ไขหรือลบข้อมูล เป็นต้น โดยทั่วไปการจัดเก็บข้อมูลจะมีการนำระบบคอมพิวเตอร์เข้ามาช่วยในการจัดการฐานข้อมูล ซึ่งในระบบฐานข้อมูลปัจจุบันจะมีการนำระบบการจัดการฐานข้อมูล (Database Management System : DBMS) คือซอฟต์แวร์จัดการฐานข้อมูลที่นำมาใช้เป็นเครื่องมือเพื่อให้ผู้ใช้สามารถโต้ตอบกับฐานข้อมูลได้ ตัวซอฟต์แวร์โดยส่วนใหญ่ประกอบไปด้วยฟังก์ชั่นต่างๆ เพื่อนำมาจัดการกับข้อมูล รวมทั้งภาษาที่ใช้การสั่งงานซึ่งโดยส่วนมากใช้ภาษาขั้นสูงภาษา SQL

ชาญชัย ศุภอรรถการ (2557, น. 125-126) ได้ให้ความหมายของคุณสมบัติ ทรานเซคชั่น (Transactions) ในระบบฐานข้อมูลหรือ ACID หมายถึงการใช้ข้อมูลต่างๆ จะเกิดเป็นขั้นตอนการทำงานซึ่งอาจจะประกอบไปด้วย ขั้นตอนเดียวหรือหลายๆ ขั้นตอนก็ได้ โดยที่ขั้นตอนจะต้องทำให้เสร็จทุกขั้นตอนจึงถือว่าทรานเซคชั่น นั้นเสร็จสมบูรณ์ เพื่อการรับประกันความน่าเชื่อถือของ ACID มีดังนี้

Atomicity คือความสามารถในการรับประกันความถูกต้องของฐานข้อมูล ถ้าส่วนใดส่วนหนึ่งของทรานเซคชั่นไม่สำเร็จ ทรานเซคชั่นทั้งหมดก็จะไม่สำเร็จด้วย

Consistency คือความสอดคล้องของฐานข้อมูล ก่อนและหลังจากการดำเนินการกับทรานเซคชั่นฐานข้อมูลจะยังคงสภาพความถูกต้อง ไม่ว่าการดำเนินการนั้นจะสำเร็จหรือไม่ก็ตามถ้าคำสั่งในทรานเซคชั่นเกิดความผิดพลาดขึ้น ก็จะมีการคืนสภาพกลับไปยังจุดเริ่มต้น

Isolation คือการแบ่งแยกโดยทรานเซคชั่นหนึ่งจะไม่มีผลต่อทรานเซคชั่นอื่น เพราะในแต่ละทรานเซคชั่น จะถูกแยกออกจากกันอย่างสิ้นเชิงทำให้การทำงานในทรานเซคชั่นหนึ่ง จะไม่ไปรบกวนอีกทรานเซคชั่นหนึ่ง

Durability คือมีความทนทาน โดยถ้ามีการดำเนินการกับทรานเซคชั่นหนึ่งจนเสร็จแล้วจะมีการบันทึกอย่างถาวรตามการดำเนินการนั้นลงในฐานข้อมูล

เพราะฉะนั้นฐานข้อมูลประกอบด้วยรายละเอียดของข้อมูลที่เกี่ยวข้องกัน ข้อมูลจะถูกเก็บไว้อย่างมีระบบ เพื่อประโยชน์ในการจัดการและเรียกใช้ข้อมูลได้อย่างมีประสิทธิภาพ เช่น ด้านโรงพยาบาลจะมีฐานข้อมูลที่เกี่ยวข้องกับข้อมูลประวัติคนไข้ ข้อมูลแพทย์เชี่ยวชาญเฉพาะโรค เป็นต้น ฐานข้อมูลเป็นการเก็บรวบรวมข้อมูลให้เป็นศูนย์กลางข้อมูลอย่างมีระบบ สามารถเรียกใช้ร่วมกันได้ ในระบบฐานข้อมูลที่มีประสิทธิภาพควรมีฮาร์ดแวร์ต่างๆ ที่พร้อมจะอำนวยความสะดวกในการบริหารระบบฐานข้อมูลได้อย่างมีประสิทธิภาพ ไม่ว่าจะเป็นขนาดของหน่วยความจำหลัก ความเร็วของหน่วยประมวลผลกลาง อุปกรณ์นำข้อมูลเข้าและออก รวมถึงหน่วยความจำสำรองที่จะรองรับการประมวลผลข้อมูลในระบบได้อย่างมีประสิทธิภาพ ในการประมวลผลฐานข้อมูลอาจจะใช้โปรแกรมที่แตกต่างกัน ทั้งนี้ขึ้นอยู่กับระบบคอมพิวเตอร์ที่ใช้ว่าเป็นแบบใด โปรแกรมที่ทำหน้าที่ควบคุมดูแลการสร้าง การเรียกใช้ข้อมูล การจัดทำรายงาน การปรับเปลี่ยนแก้ไขโครงสร้าง การควบคุม ในเทคโนโลยีฐานข้อมูลจะมีแบบจำลองข้อมูล (Data Model) ซึ่งในส่วนของฐานข้อมูล ซึ่งมีอยู่ 1 แบบจำลองที่นิยมใช้กันมากในระบบฐานข้อมูลปัจจุบันคือแบบจำลองฐานข้อมูลเชิงสัมพันธ์ (Relational Database Model) ซึ่งมีโครงสร้างข้อมูลเชิงสัมพันธ์ที่ประกอบไปด้วย 1.รีเลชั่น 2.แอตตริบิวต์ 3.โดเมน 4.ทัปเพิล 5.ดีกรี 6.คาร์ดินัลลิตี้ ซึ่งในฐานข้อมูลจะประกอบไปด้วยหลายตารางมีการเชื่อมโยงสัมพันธ์ข้อมูลกัน

ฐานข้อมูลเชิงสัมพันธ์ หมายถึงตารางข้อมูลประกอบด้วยจำนวนรีเลชั่นต่างๆ ที่ได้รับการจัดรูปแบบให้เป็นบรรทัดฐาน และโครงสร้างหรือสคีมา (Schema) บางส่วนของฐานข้อมูลจัดการโดยระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ (Relational Database Management System : RDBMS) ซึ่งมีรูปแบบจำลองข้อมูลทางคณิตศาสตร์ที่เกี่ยวกับชุดปฏิบัติการหรือโอเปอเรชั่น (Operations) ที่ใช้จัดการกับข้อมูลเหล่านี้

พีชคณิตเชิงสัมพันธ์ (The Relational Algebra) เป็นแบบจำลองข้อมูล ที่กำหนดโครงสร้างและข้อบังคับแล้วในการเรียกใช้ เป็นชุดปฏิบัติการหรือโอเปอเรชั่นที่นำมาใช้จัดการกับ

ข้อมูล หรือเป็นโอเปอเรชั่นบนแบบจำลองเชิงสัมพันธ์ซึ่งมีโอเปอเรชั่นพื้นฐาน 6 แบบ คือ 1.Selection 2.Projection 3.Cartesian 4.Product 5.Union และ 6.Set Difference และส่วนเรียกดูข้อมูลอีก 3 โอเปอเรชั่นคือ 1.Join 2.Intersection 3.Division

แคลคูลัสเชิงสัมพันธ์ (Relational Calculus) เป็นรูปแบบจำลองข้อมูล ที่ต้องใช้กำหนดการที่ต้องการข้อมูลอะไร สามารถกำหนดรูปแบบการค้นหาในลักษณะของนิพจน์หรือสมการทางคณิตศาสตร์ที่มีตัวแปร ค่าคงที่ ตัวกระทำ และตัวเชื่อมอื่นๆ ซึ่งผลลัพธ์ที่จะได้คือทัปเพิล หรือแถว จากความสัมพันธ์ที่ส่งผลให้ค่าสมการนั้นเป็นจริง

สรุปแคลคูลัสเชิงสัมพันธ์จะยึดหลักเกณฑ์การกำหนดข้อมูลต่างๆ ว่าต้องการอะไรจากรีเลชั่น (What) โดยไม่สนใจวิธีการที่ได้มาซึ่งผลลัพธ์ ในขณะที่พีชคณิตเชิงสัมพันธ์จะมุ่งเน้นถึงวิธีการว่าจะต้องทำอย่างไร (How) เพื่อให้ได้มาซึ่งผลลัพธ์ตามที่ต้องการ (โอภาส เอี่ยมสิริวงศ์, 2558, น. 205-219)

2.1.3 ทฤษฎีการค้นคืนด้วยภาษาสอบถามเชิงโครงสร้างเอสคิวแอล

ภาษาสอบถามเชิงโครงสร้าง (Structured Query Language : SQL) สามารถอ่านออกเสียงได้ว่า เอสคิวแอลซีควล เป็นภาษาที่นิยมใช้งานกับการจัดฐานข้อมูลในการเรียกใช้งานฐานข้อมูล โดยมีมาตรฐาน ANSI ประเภทชุดคำสั่ง SQL มี 3 ประเภทด้วยกัน

1. ภาษานิยามข้อมูล (DDL) ประกอบด้วยกลุ่มคำสั่งที่ใช้สร้างตาราง และลบตาราง และเพิ่ม และลบตาราง รวมถึงแก้ไขแอตตทริบิวต่างๆ ในรีเลชั่น และการสร้างลำดับดัชนี

2. ภาษาจัดการข้อมูล (DML) ประกอบด้วยกลุ่มคำสั่ง อับเดท เพิ่ม ปรับปรุงและเรียกดูข้อมูลในฐานข้อมูล

3. ภาษาควบคุมข้อมูล (DCL) เป็นกลุ่มคำสั่งการอนุญาต หรือยกเลิกสิทธิ์ในการใช้งานฐานข้อมูล ช่วยอำนวยความสะดวกแก่ผู้บริหารฐานข้อมูล ในการควบคุม

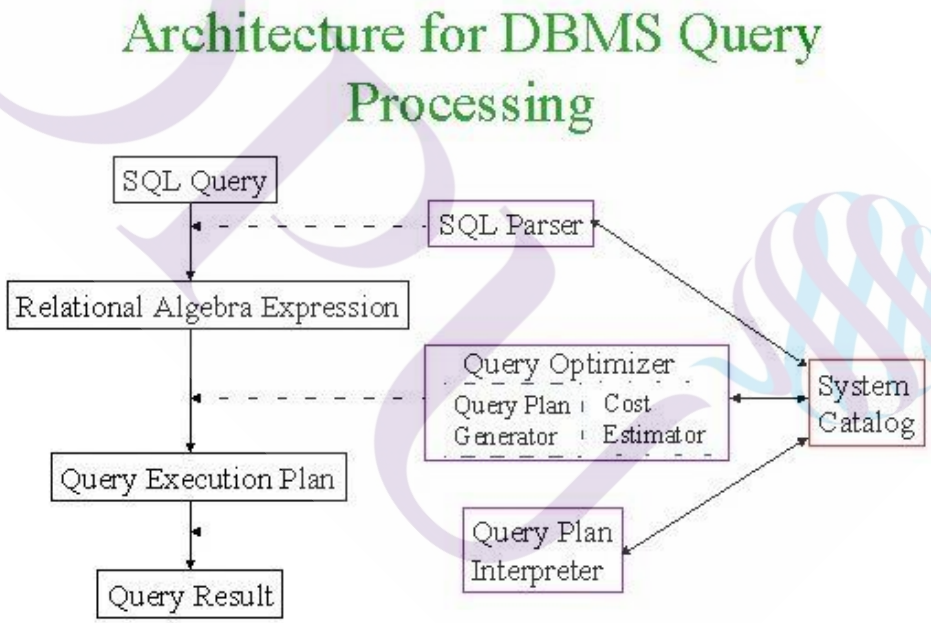
2.1.4 ทฤษฎีกระบวนการสอบถามข้อมูล

ชยาพร แก่นสาร์ (2555, น. 722-724) ได้ให้ความว่ากระบวนการสอบถามข้อมูล (Query Processing) คือการประมวลแบบสอบถามข้อมูล เป็นหนึ่งในขั้นตอนของระบบการจัดการฐานข้อมูล ทำหน้าที่หาคำตอบที่ดีที่สุดและถูกต้องที่สุดให้กับความต้องการของผู้ใช้ได้คำตอบที่เหมาะสมและรวดเร็วประหยัดค่าใช้จ่ายมากที่สุด ซึ่งค่าใช้จ่ายจะเกี่ยวข้องกับการประมวลผลที่หน่วยประมวลผลกลาง (CPU Time) และเวลาที่ใช้ในการดึงข้อมูลจากอุปกรณ์อินพุทและเอาท์พุท (I/O Time) ในกระบวนการประมวลผลแบบสอบถามมี 3 ขั้นตอน ดังนี้

1. Parsing คือการรับคำสั่งมาแยกออกมาเป็นส่วนต่างๆ เพื่อตรวจสอบรูปแบบ และแปลงให้อยู่ในรูปที่กะทัดรัด หรือเรียกว่า Validation สิ่งที่ได้จากขั้นตอนนี้คือ Relational Algebra Tree ที่ประกอบด้วยการเลือก (Selection) การรวม (Joining) การโปรเจคชั่น (Projection)

2. Query Optimization จะเลือกหาวิธีที่ดีที่สุดคือเร็วหรือใช้ต้นทุนน้อยสุดที่ได้จาก Passing สามารถทำได้หลายวิธีที่สามารถ แต่วิธีการไหนเร็วหรือต้นทุนต่ำสุดจะเลือกวิธีนั้นโดยดูจากค่าสถิติต่างๆ สิ่งที่ได้จากขั้นตอนนี้คือ Execution Plan คือรูปแบบหรือวิธีการที่จะเลือกหรือดึงข้อมูลที่ Optimaize ดีที่สุด

3. Query Evaluation จะประกอบด้วย 2 ส่วนคือ Code Generation กับ Runtime Query Execution ในขั้นตอนนี้ก็จะคัดสิ่งที่เลือกในขั้นตอนที่ 2 มาทำการสร้างเป็น Physical Operators ที่ใช้ในการจัดการหรือดึงข้อมูลจากไฟล์ข้อมูลในฐานข้อมูล และการดำเนินการคำสั่งใน Physical Operator หรือที่เรียกว่าการ Execution เพื่อจะได้ผลลัพธ์ของการคิวรี่

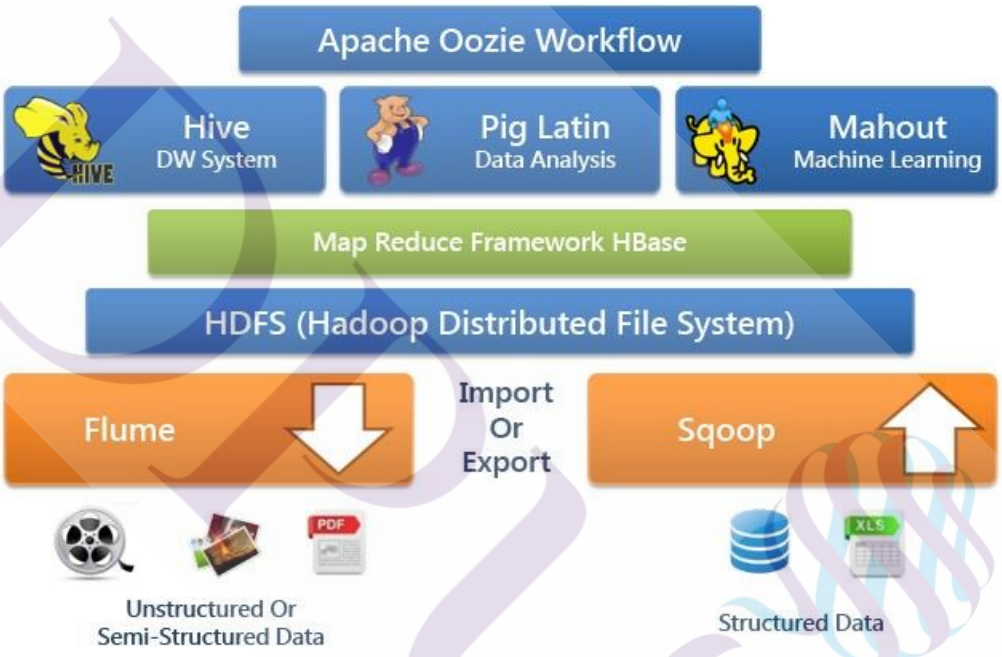


ภาพที่ 2.4 สถาปัตยกรรมกระบวนการสอบถามข้อมูลระบบฐานข้อมูล

ที่มา: Department of Computing Science, University of Alberta [online] : เข้าถึง 15 ก.ย. 2558. จาก https://webdocs.cs.ualberta.ca/~zaiane/courses/cmput391-02/slides/Lect3/sld008.htm

2.1.5 ทฤษฎีระบบจัดเก็บแบบกระจายข้อมูลของฮาดูป

ข้อมูลขนาดใหญ่ (Big Data) เกิดจากความสามารถและประสิทธิภาพที่สูงขึ้นในการสื่อสารผ่านอินเตอร์เน็ตผ่านสื่อออนไลน์ต่างๆ เช่น การค้นหาข้อมูลผ่าน Google การทวิตผ่าน Twitter การโพสต์รูป การกดไลค์ผ่าน Facebook รวมถึงธุรกรรมอื่นๆ ที่ผ่านเครือข่ายอินเตอร์เน็ต เป็นต้น ทำให้เกิดข้อมูลธุรกรรมจำนวนมหาศาล องค์กรต่างๆ จึงหาทางนำข้อมูลต่างๆ มาใช้งานในการวิเคราะห์ข้อมูล เพื่อหาสารสนเทศที่เป็นประโยชน์ และขับเคลื่อนนวัตกรรมขององค์กร โดยมีแนวทางในการดำเนินการและกลยุทธ์ขององค์กร (Gartnet, 2012)

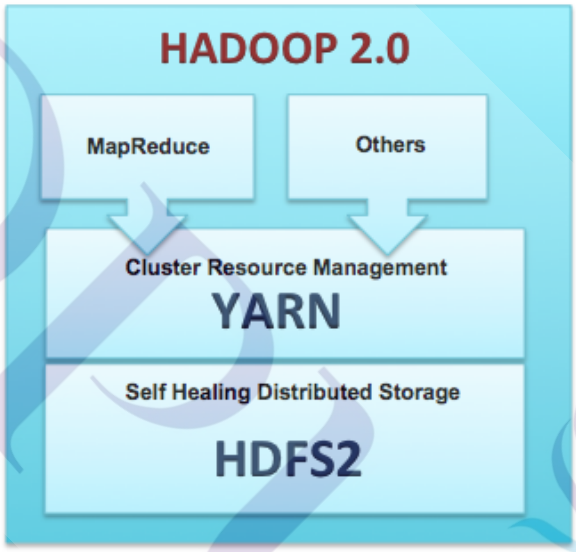


ภาพที่ 2.5 กรอบการทำงาน Hadoop Ecosystem

ที่มา: Introduction to Hadoop, www.stratapps.net [online] : เข้าถึง 15 ก.ย. 2558. จาก http://www.stratapps.net/intro-hadoop.php

กรอบการทำงานหรือแพลตฟอร์มฮาดูป (Hadoop) เป็นหนึ่งในเทคโนโลยีระบบข้อมูลขนาดใหญ่ เป็นระบบการจัดเก็บข้อมูลแบบกระจาย (Distribute System) ฮาดูปพัฒนามาจนถึงปัจจุบันในเวอร์ชั่น 2.6.2 มีขั้นตอนการทำงานโดยการแบ่งไฟล์ออกมาเป็นไฟล์ย่อยๆ หรือบล็อกข้อมูล (Data Block) จัดเก็บในระบบ HDFS (Hadoop Distribute File System) และมี Name Node (Master) ทำหน้าที่ระบุตำแหน่งเก็บ และมี Data Node (Slave) กระจายไปเก็บในเครื่องอื่นๆ และมี

YARN (Yet Another Resource Negotiator) ควบคุมจัดการทรัพยากรและใช้การประมวลผลแบบขนานแมพรีดิว (MRV2) ปัจจุบันเป็นเวอร์ชั่น 2 (White, 2012, pp. 13-14) แมพรีดิวจึงเป็นการเขียนโปรแกรมควบคุมความต้องการข้อมูลที่ต้องการค้นคืนผ่านการจับคู่ Key/Value ที่กำหนดไว้ YARN Resource Manager ทำหน้าที่ควบคุมคลัสเตอร์ คอยบริหารตารางงานของ Job Tracker หรือ JobHistoryServer ที่ส่งไปยัง Node Manager (Slave) มี YARN Node Manager และ YARN Application Master ทำหน้าที่ควบคุมการทำงานของแมพรีดิวภายในคลัสเตอร์ การจัดการทรัพยากรและการจัดเก็บและประมวลผล YARN/MRV2 แบบใหม่นี้จะกระทำในแต่ละเครื่องคอมพิวเตอร์เพื่อลดปริมาณการประมวลผลภายในเครือข่ายลง (Gunarathne & Srinath Perera, 2015, pp. 60-66)



ภาพที่ 2.6 กรอบการทำงานของฮาดูปทำงานร่วมกับแมพรีดิว

ที่มา: Taming Big Data using HDInsight, www.packtpub.com [online] : เข้าถึง 15 ก.ย. 2558. จาก https://www.packtpub.com/books/content/taming-big-data-using-hdinsight

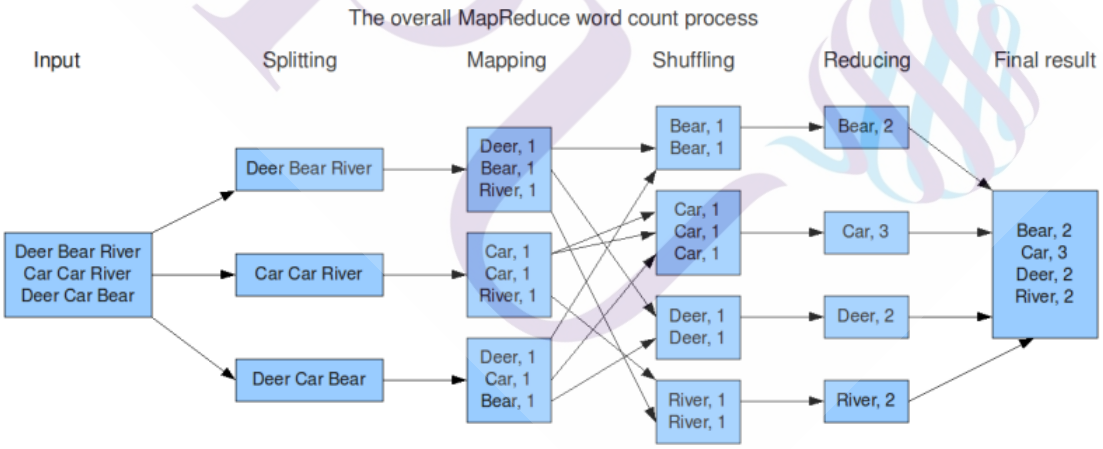
2.1.6 ทฤษฎีระบบการประมวลผลแบบขนานด้วยเทคนิคแมพรีดิว

การประมวลผลข้อมูลแบบขนาน (Parallel Processing) ที่จะนำมาใช้ศึกษาการประมวลผลนี้คือ MapReduce เป็นส่วนประมวลผลข้อมูล ประกอบไปด้วย สองขั้นตอนหลักใหญ่ คือ ขั้นตอนของการ Map และ Reduce

Map หมายถึงการเตรียมข้อมูล ให้อยู่ในรูปของ Key/Value โดยในขั้นตอนการ Map เราสามารถใส่เงื่อนไขเพื่อตรวจสอบค่าก่อนนำมาประมวลผล หรือเปลี่ยนแปลงข้อมูลให้เหมาะสมได้ และปลดปล่อยค่าออกไปในรูปแบบของการจับคู่

Reduce หมายถึงการลดผลลัพธ์ลง จากขั้นตอนการ Map โดย Reduce จะเป็นฟังก์ชั่นที่รับค่า Key/Value และนำมาประมวลผลแสดงค่าผลลัพธ์

สามารถสรุปได้ว่าฮาดูปและแมพรีดิวเป็นสถาปัตยกรรมการจัดการข้อมูลแบบจับคู่ (Key Value) วิธีการจัดเก็บแบบกระจายฮาดูป หรือ HDFS โดยการแบ่งข้อมูลออกเป็นไฟล์ย่อยๆ และมีโปรแกรมควบคุมทำหน้าที่ระบุตำแหน่งเก็บ โดยระบบจะกระจายข้อมูลไปจัดเก็บในเครื่องอื่นๆ ตามที่กำหนด และแมพรีดิวมีวิธีการประมวลผลแบบขนาน มีสองขั้นตอนหลักคือ ขั้นตอนการแมพ (Map) หมายถึงการเตรียมข้อมูลให้อยู่ในรูปคีย์แวลู (Key/Value) โดยในขั้นตอนนี้สามารถเขียนโปรแกรมมีเงื่อนไขเพื่อตรวจสอบค่าก่อนนำมาประมวลผลหรือเปลี่ยนแปลงคุณลักษณะข้อมูลให้เหมาะสมได้ และนำส่งค่าออกไปในรูปแบบของการจับคู่ ขั้นตอนการรีดิว (Reduce) หมายถึงการลดผลลัพธ์ลงจากขั้นตอนการแมพโดยรีดิวจะเป็นฟังก์ชั่นที่รับค่า Key/Value และนำมาประมวลผลแสดงค่าในขั้นตอนสุดท้ายผลลัพธ์จัดเก็บในรูปแบบเท็กซ์ไฟล์ (Text File)



ภาพที่ 2.7 กรอบการทำงาน MapReduce

ที่มา: MapReduce Exercise: Hand On Lab, Calvin College [online] : เข้าถึง 10 ก.พ. 2558. จาก https://cs.calvin.edu/courses/cs/374/exercises/12/lab/

การเขียนโปรแกรมสามารถดำเนินการได้ในหลากหลายรูปแบบ ตัวอย่างตามภาพที่ 2.7 เมื่ออ้างอิงกับทฤษฎีการประมวลผลแบบขนานด้วยแมพรีดิว Miner and Shook (2012, pp. 4-7) ในการกำหนดกลุ่มงานการเขียนโปรแกรมสามารถกำหนดกลุ่มงานย่อยๆ ได้ดังนี้ 1.Record Reader (Input) เขียนโปรแกรมการอ่านข้อมูลนำเข้า 2.Map (Spitting) เขียนโปรแกรมทำการแบ่งข้อมูลออกตามข้อมูลที่ระบุความต้องการ 3.Combiner (Mapping) เขียนโปรแกรมการรวมจำนวนกลุ่มที่มีค่าที่เหมือนกัน 4.Shuffle and Sort (Shuffing/Sorting) เขียนโปรแกรมการจัดเรียงค่าที่เหมือนกันใหม่ 5.Reduce (Reducing) เขียนโปรแกรมการรวมจำนวนค่าที่มีเหมือนกันใหม่ 6.Output Format (Final result) เขียนโปรแกรมการแสดงผลลัพธ์ที่ได้โดยการกำหนดโดยการเขียนโปรแกรมได้ว่าต้องการผลออกมาเป็นไฟล์ฟอร์เมตประเภทใด

2.1.7 ทฤษฎีการวิเคราะห์ข้อมูลและสถิติการวิเคราะห์ข้อมูล

สถิติเพื่อการวิจัยทางสารสนเทศศาสตร์ มีหลักแนวคิดในการค้นคว้าสถิติเพื่อการวิจัยทางสารสนเทศศาสตร์เป็นสถิติที่ใช้เพื่อสรุป และนำเสนอข้อมูลที่เก็บรวบรวมมาได้ตามวัตถุประสงค์ และสมมติฐานของการวิจัย สถิติที่ใช้จะต้องเหมาะสมกับชนิดของตัวแปร วัตถุประสงค์ และรูปแบบของการศึกษาวิจัย การวิเคราะห์ที่น่าเชื่อถือขึ้นอยู่กับคุณภาพของข้อมูล และข้อมูลที่ได้มานั้นมีลักษณะเป็นไปตามข้อตกลงเบื้องต้นของการวิเคราะห์นั้นๆ

สถิติพรรณนาเป็นสถิติที่ใช้เพื่ออธิบายลักษณะของข้อมูลหรือเพื่ออธิบายความสัมพันธ์ระหว่างข้อมูลซึ่งการเลือกใช้สถิติพรรณนาขึ้นอยู่กับวัตถุประสงค์ของการวิจัยและชนิดของตัวแปร

สถิติอ้างอิงเป็นสถิติที่ใช้ในการอ้างอิงข้อสรุปจากกลุ่มตัวอย่างไปยังกลุ่มประชากรเพื่อเปรียบเทียบความแตกต่างของข้อมูลโดยการทำนาย การหาความสัมพันธ์ และการวิเคราะห์ข้อมูลหลายตัวแปร ซึ่งการเลือกใช้สถิติอ้างอิงขึ้นอยู่กับวัตถุประสงค์ของการวิจัย รูปแบบการศึกษา และชนิดของข้อมูล (มหาวิทยาลัยสุโขทัยธรรมาธิราช [มสธ.], 2546, น. 313-350)

ผู้วิจัยมีแนวคิดการคัดเลือกกลุ่มตัวอย่างแบบเจาะจงเป็นการเลือกกลุ่มตัวอย่างให้สอดคล้องกับเรื่องที่วิจัย หรือจะเรียกชื่ออย่างอื่นว่า การเลือกกลุ่มตัวอย่างตามวัตถุประสงค์ การเลือกกลุ่มตัวอย่างตามความมุ่งหมาย และการเลือกกลุ่มตัวอย่างโดยอาศัยการตัดสินใจ (Judgment Sampling) มีการวางแผนกำหนดจำนวนตัวอย่างและเลือกกลุ่มตัวอย่างที่ดีเพื่อไม่ให้เกิดความลำเอียง (สิน พันธุ์พินิจ, 2552, น. 142)

เป็นแนวทางการเลือกแบบความสะดวกในการคัดกลุ่มตัวอย่างจากข้อมูล เนื่องจากต้องการทดลองเพื่อให้ได้ผลลัพธ์ด้านความเร็ว (ประสิทธิภาพ) และผลลัพธ์ความแม่นยำถูกต้อง (ประสิทธิผล) ระหว่างระบบบริหารข้อมูล 2 รูปแบบ ในการเก็บรวบรวมข้อมูล และการวิเคราะห์

และแปลผลสามารถใช้ข้อมูลได้ 2 รูปแบบ คือ ข้อมูลเชิงปริมาณและข้อมูลเชิงคุณภาพ ข้อมูลทั้ง 2 แบบ จะต้องมีคุณสมบัติ ที่เที่ยงตรง และสร้างความเชื่อมั่น และยังกล่าวถึงความไว และความเฉพาะเจาะจงในวิธีการได้มาซึ่งข้อมูล เพื่อเลือกกลุ่มควบคุมเป็นเทคนิค และเกณฑ์ในการคัดกรองคุณภาพ และเป็นสิ่งที่ควรใช้ในการทำการประเมิน การวิเคราะห์และแปลผลข้อมูลเชิงปริมาณ โดยทั่วไปจะแบ่งการวัดดังนี้ 1. การแบ่งกลุ่ม หรือนามมาตร (Nominal Scale) 2. การแบ่งตามตำแหน่งหรือลำดับสิ่งที่ปรากฏ (Ranking) หรืออันดับมาตร (Ordinal Scale) 3. การใช้ช่วงของการวัด (Interval Scale) หรือช่วงมาตร 4. การใช้วัดอัตราส่วนมาตร (Ratio Scale)

การวิเคราะห์ทางสถิติ (Statistical Analysis) ในการประเมินเชิงปริมาณต้องมีการทดสอบทางสถิติของตัวแปร หรือตัวชี้วัดที่เราทำการวัดก่อน และวัดหลังที่มีการปฏิบัติการ ในการประเมินที่จะต้องนำเสนออย่างง่ายคือ การแยกแยะการวิเคราะห์ทีละขั้นตอน 1. นำเสนอและสรุปย่อผลออกมาเป็นตัวเลข โดยการใช้ตารางเปรียบเทียบ หรือใช้กราฟ มีค่ามัชฌิมเลขคณิต ค่ามัธยฐาน ค่าฐานนิยม ช่วงระหว่างค่าสูงสุดและต่ำสุด 2. บอกถึงความเชื่อมั่นในกลุ่มตัวอย่าง 3. ถ้ามีการพิสูจน์สมมติฐานการรับหรือปฏิเสธสมมติฐานเป็นส่วนของผลการประเมิน (นวรัตน์ สุวรรณผ่อง, มธุรส ทิพยมงคลกุล, ทองหล่อ เดชไทย, และนพพร โหวธีระกุล, 2557, น. 183-187)

การนำเสนอข้อคิดเห็นถึงจุดประสงค์ของการอภิปรายผลการพิสูจน์และผลการประเมินประสิทธิภาพคือ การได้มาซึ่งข้อมูลต่างๆ ที่เกี่ยวข้องกับการพิสูจน์ข้อเสนอวิธีการที่นำเสนอ ได้แก่ เหตุผลที่งานวิจัยในอดีตไม่สามารถแก้ปัญหาได้ เหตุผลที่งานวิจัยที่เสนอทำได้ ผลการคำนวณซึ่งเป็นค่าเป้าหมาย ผลการทดสอบหรือผลการทดลองซึ่งเป็นค่าจริง ค่าผิดพลาดซึ่งเป็นค่าความต่างระหว่างค่าเป้าหมายกับค่าจริง ค่าผิดพลาดซึ่งเป็นค่าความต่างระหว่างค่าเป้าหมายกับค่าจริง สาเหตุของค่าผิดพลาดและแนวทางป้องกัน ข้อจำกัดของวิธีการที่เสนอผลข้างเคียง และงานวิจัยในอนาคต (โกสินทร์ จำนงไทย, 2559, น. 231)

2.2 งานวิจัยที่เกี่ยวข้อง

Bhosale and Gadekar (2014) มีงานวิจัยเรื่อง A Review Paper on Big Data and Hadoop ปัญหาของงานวิจัยนี้คือ ข้อมูลขนาดใหญ่เป็นข้อมูลที่มีขนาดความหลากหลายและมีความซับซ้อน ต้องใช้สถาปัตยกรรมใหม่และกรอบการทำงานข้อมูลใหม่ เทคโนโลยีข้อมูลขนาดใหญ่มีอัลกอริทึมและเทคนิคการจัดการข้อมูลและแนวทางการวิเคราะห์ข้อมูล ซึ่งก่อให้เกิดความคุ้มค่าในการสกัดความรู้ที่ซ่อนอยู่ในข้อมูลที่หลากหลายและแตกต่างเหล่านั้นอย่างไร ซึ่งเป็นปัญหาที่ยังไม่กระจ่างแจ้งที่ผู้วิจัยบทความนี้ต้องการหาบทพิสูจน์ ฮาดูป (Hadoop) เป็นแพลตฟอร์มหลักสำหรับการใช้งานเพื่อการสกัดข้อมูลขนาดใหญ่เหล่านั้น

ผู้วิจัยฉบับนี้จึงนำเสนอการวิจัยเชิงสำรวจในเทคโนโลยีข้อมูลขนาดใหญ่และการแก้ปัญหาด้วยเทคโนโลยีข้อมูลขนาดใหญ่ ด้วยการศึกษาทบทวนวรรณกรรมจากงานวิจัยที่เกี่ยวข้อง สมมติฐานของงานวิจัยนี้คือ เทคโนโลยีขนาดใหญ่มีวัตถุประสงค์ในการวิเคราะห์ข้อมูลเพื่อนำข้อมูลที่ถูกวิเคราะห์นำมาใช้ประโยชน์ โดยมีวิธีการวิจัยและรูปแบบการวิจัย ด้วยการศึกษาคุณสมบัติข้อมูลขนาดใหญ่และอะไรที่เรียกว่าข้อมูลขนาดใหญ่ ข้อมูลขนาดใหญ่มีองค์ประกอบ 3 ส่วน หรือเรียกว่า 3Vs คือ 1.ปริมาณของข้อมูล (Volumn) 2.ความหลากหลายของข้อมูล (Variety) 3.ความเร็วของข้อมูล (Velocity) และความหมายของข้อมูลขนาดใหญ่คืออะไร

ในบทความนี้ให้นิยามว่า ข้อมูลขนาดใหญ่หมายถึงชุดข้อมูลหรือข้อมูลที่มีหลากหลายรูปแบบ มีโครงสร้าง, ไม่มีโครงสร้าง, กึ่งโครงสร้าง มีหลายประเภทเช่น ข้อมูลเสียง, ข้อมูลวิดีโอ, ข้อมูลอักษร ที่มาจากหลากหลายแห่งข้อมูล และมีอัตราการเจริญเติบโตของข้อมูลที่รวดเร็ว ทำให้มีความยุ่งยากในการจัดการการประมวลผลหรือการนำมาวิเคราะห์ในเวลาอันจำกัดโดยใช้เทคโนโลยีทั่วไป เช่น ฐานข้อมูลเชิงสัมพันธ์และซอฟต์แวร์ที่สร้างขึ้นเองหรือซอฟต์แวร์ประมวลผลที่ขายในท้องตลาด

มีตัวแปรที่สำคัญ คือฮาดูป (Hadoop) เป็นโครงการฟรีซอฟต์แวร์ที่มีรูปแบบการประมวลผลแบบกระจายช่วยจัดการชุดข้อมูลขนาดใหญ่จัดเก็บไว้ในเครื่องเซิร์ฟเวอร์จำนวนหลายเครื่อง เรียกว่าคลัสเตอร์เซิร์ฟเวอร์ (Cluster Server) เพื่อดำเนินการจัดการข้อมูลจำนวนเหล่านี้ในราคาที่ไม่แพงและมีประสิทธิภาพ และงานวิจัยนี้มีวิธีเก็บข้อมูลด้วยเทคนิควิธีการด้วยการทบทวนวรรณกรรมที่เกี่ยวข้องกับเทคโนโลยีข้อมูลขนาดใหญ่ที่มีโครงสร้างสถาปัตยกรรมฮาดูป (Hadoop) ในการจัดเก็บข้อมูลขนาดใหญ่แบบกระจายและการประมวลผลข้อมูลแบบขนาน

มีเครื่องมือวัดและวิธีวิเคราะห์ข้อมูลด้วยการวิเคราะห์ปัญหาในการประมวลผลข้อมูลขนาดใหญ่ในหัวข้อดังนี้ 1.ข้อมูลที่มีความแตกต่างและไม่ครบถ้วน (Heterogeneity and Incompleteness) 2.ขนาดข้อมูลที่มีการขยายตัว (Scale) 3.ทันเวลาในการเรียกใช้ (Timeliness) 4. ความเป็นส่วนตัว (Privacy) 5.สนับสนุนการทำงานร่วมกันของมนุษย์ (Human Collaboration) และการวิเคราะห์จำแนกองค์ประกอบของเทคโนโลยีข้อมูลขนาดใหญ่ในกลุ่มซอฟต์แวร์ที่ใช้การวิเคราะห์ประกอบไปด้วย HBase, Hive, MongoDB, Redis, Cassandra, Drizzle ตามหัวข้อดังนี้ 1. รายละเอียดการใช้งาน (Description) 2.ภาษาที่ใช้ (Implementation language) 3.รูปแบบฐานข้อมูล (Database Model) 4.แนวคิดความถูกต้องตรงกัน Consistency Concept) 5.การสอดคล้อง (Concurrency) 6.ความทนทาน (Durability) 7.วิธีการทำซ้ำหรือสำเนาข้อมูล (Replication Method)

สรุปผลจากการวิจัย บทความนี้อธิบายแนวคิดของเทคโนโลยีข้อมูลขนาดใหญ่ และปัญหาที่จะเกิดการจากประมวลผลข้อมูลขนาดใหญ่ และความแตกต่างของเครื่องมือในเทคโนโลยี

ข้อมูลขนาดใหญ่ ซึ่งความท้าทายของข้อมูลขนาดใหญ่ไม่ได้มีแต่ปัญหาที่เกิดจากการประมวลผลเท่านั้นแต่ยังมีการสร้างให้เห็นถึงขั้นตอนการวิเคราะห์ข้อมูลที่จะทำให้เกิดประโยชน์ในการนำข้อมูลออกมาตีความจากการทำงานของมนุษย์และใช้งานในหลากหลายโปรแกรม และจะไม่คุ้มค่าเลยถ้าจะทำการวิเคราะห์ในบริบทเดียวและใช้เพียงโดเมน (Domain) หรือกลุ่มการทำงานเดียว

Gurevich (2015) มีงานวิจัยเรื่อง Comparative Survey of NoSQL/NewSQL DB Systems ปัญหาของงานวิจัยนี้คือในการใช้งานฐานข้อมูลรูปแบบใหม่ที่มีความแตกต่างกับแบบเชิงสัมพันธ์รูปแบบดั้งเดิมเติบโตขึ้นอย่างรวดเร็ว และทุกสภาพแวดล้อมของการใช้งานจะมีข้อกำหนดใหม่ๆ สำหรับการจัดเก็บข้อมูลและการประมวลผลที่ยังมีประสิทธิภาพไม่เป็นที่ประจักษ์ ผู้วิจัยฉบับนี้จึงนำเสนอวิธีการวิจัยและรูปแบบการวิจัยที่มีเป้าหมายหลักของการวิจัยคือการสำรวจเชิงเปรียบเทียบของฐานข้อมูล NoSQL ด้วยการศึกษาคุณลักษณะของฐานข้อมูล 4 แบบ คือ 1) Key-value stores 2) Document stores 3) Column family stores 4) Graph databases และของฐานข้อมูล NewSQL เน้นลักษณะทางเทคนิคด้วยการเปรียบเทียบเชิงคุณภาพและเชิงปริมาณ การประเมินผลเชิงคุณภาพด้วยการเปรียบเทียบคุณสมบัติที่ใช้ได้ของฐานข้อมูล SQL กับ NoSQL การเปรียบเทียบคุณสมบัติ RDBMS เป็น ACID และเปรียบเทียบกับคุณสมบัติ NoSQL หรือ BASE (Availability, Graceful degradation, Performance) ที่นิยามโดย Eric Brewer เช่นเดียวกับทฤษฎี CAP ที่นิยามให้กับ Cloud Computing และประเมินผลเชิงปริมาณด้วยข้อมูล 2 ชุดข้อมูล ที่มีจำนวนระเบียนข้อมูลเท่ากัน และขนาดข้อมูลแตกต่างกัน กับเครื่องคอมพิวเตอร์เสมือนที่เท่ากันและใช้วิธีการติดตั้งกับฐานข้อมูลเหมือนกัน และทำการทดลองเพื่อประเมินผลการปฏิบัติด้านประสิทธิภาพการทำงาน

เครื่องมือในการวัดด้วยระบบการเปรียบเทียบ YCSB (Yahoo Cloud Serving Benchmark) และยังเปรียบเทียบฐานข้อมูลด้วย BG (Benchmark Graph) ในระบบเครือข่ายสังคม มีวิธีการคัดเลือกตัวแปรและมีตัวแปรที่สำคัญดังนี้ การคัดเลือกตัวแปรเป็นการเลือกฐานข้อมูลที่เป็นที่นิยมทั้ง NoSQL และ NewSQL ด้วยการเลือกด้วยหัวข้อ 1.DB-Engine Ranking หรือการจัดอันดับของข้อกำหนดในการเลือกดังนี้ 1.Google Trends 2.การจัดอับดับเว็บไซต์ในการใช้ค้นหา Search Engines 3.การตอบคำถามทางเทคนิคบนเว็บไซต์ Stack Overflow และ DBA Stack Exchange 4. การจ้างงานในเว็บไซต์เครือข่ายสังคม LinkedIn 5.การหาประวัติของผู้เชี่ยวชาญในการค้นหาผ่าน Search Engines 6.การค้นหาในเครือข่ายสังคม Twitter ที่มีการพูดถึงฐานข้อมูลที่นิยมกัน ซึ่งได้ผลการจัดอับดับดังนี้ Document Store มีลำดับที่ 1.MongoDB 2.CouchDB และ Graph Database มีลำดับที่นิยมดังนี้ 1.Neo4j 2.OrientDB และลำดับประเภทฐานข้อมูลที่ได้รับความนิยม 1.Graph Database 2.Wide column stores 3.Document stores 4.RDF Stores 5.Search engine 6.Key-value

stores 7.Native XML Database 8.Object oriented Database 9.Multivalue Database 10.Time Series Database

และในบทความนี้ยังมีการเปรียบเทียบลักษณะการทำงานเพื่อการนำไปใช้งานที่เหมาะสม ตามรูปแบบฐานข้อมูลตามที่กล่าวมาในฐานข้อมูล NoSQL ประเภท Key-value stores และ Document stores และ Column family stores และ Graph databases และ NewSQL สมมติฐานของงานวิจัยนี้คือ การออกแบบระบบที่เหมาะสมขึ้นอยู่กับลักษณะของการใช้งานและความต้องการในการสอบถามข้อมูล มีการวิเคราะห์และเปรียบเทียบลักษณะการทำงานกับข้อมูลด้วยหัวข้อดังต่อไปนี้ 1.รูปแบบข้อมูล (Data model) 2.ความเป็นไปได้ของแบบสอบถาม (Querying possibilities) 3.การควบคุมภาวะพร้อมกัน (Concurrency control) 4.การทำซ้ำสำเนา (Replication) 5.การปรับขยาย (Scalability) 6.แบ่งพาร์ติชัน (Partitioning) 7.ความคงเส้นคงวา (Consistency) 8. คุณลักษณะด้านความปลอดภัย (Security features/drawbacks) 9.กรณีการใช้งานที่เหมาะสม (Use cases/Applications suitability) 10.ความนิยม (Popularity) 11.ประสิทธิภาพ (Performance)

โดยมีสภาพแวดล้อมจากคุณสมบัติที่กำหนด การสอบถามที่เป็นไปได้, การควบคุมการทำงานพร้อมกัน, การทำซ้ำสำเนา, การขยายตัว, การแบ่งชุดข้อมูล, การตรวจสอบความถูกต้องตรงกันและความปลอดภัยข้อมูล วิธีวิเคราะห์ข้อมูลด้วยการจัดทำตารางเปรียบเทียบคุณสมบัติโครงสร้างข้อมูลใน Old SQL, NoSQL, NewSQL ซึ่งผลการวิเคราะห์คือ Old SQL มีการเชื่อมความสัมพันธ์ มี SQL และมี ACID ผลการวิเคราะห์รูปแบบข้อมูลของ NoSQL คือ ไม่มีคุณสมบัติโครงสร้างเหมือน Old SQL แต่มีการขยายในแนวนอน และประสิทธิภาพประมวลผลข้อมูลขนาดใหญ่และเป็นข้อมูลที่ไม่ต้องมีการจัดเตรียมโครงสร้างข้อมูลไว้ล่วงหน้า (Schema-less) และคุณสมบัติโครงสร้างข้อมูล NewSQL จะเป็นฐานข้อมูลการผสมผสานระหว่าง NoSQL และ Relational Database เช่น มีการเชื่อมสัมพันธ์, มี SQL, มี ACID, เป็น Horizontal Scalability, ใช้กับข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพแต่ไม่มี Schema-less ซึ่ง NewSQL เกิดจากความต้องการนำ NoSQL มาดำเนินการจัดทำ Online transaction processing (OLTP) ที่เกิดการอ่านและเขียนจำนวนมากด้วยการรับประกันตามมาตรฐาน ACID โดยการนิยามการโดย Matt Aslett กลุ่ม “The 451 group” เช่น VolDB และ NuoDB

ดังนั้นสรุปได้ว่า NewSQL เป็นการนำประโยชน์ของหลักการเชิงสัมพันธ์มาใช้มากกว่าการขยายออกไปตามหลักการขยายแบบแนวนอน และผลการทำตารางความเป็นไปได้ของการหาความเหมาะสมในการใช้งานแบบสอบถามของฐานข้อมูล Key-Value store ใช้ Voldemort, Document store ใช้ MongoDB, Column store ใช้ Cassandra, Graph database ใช้ Neo4j และ New SQL ใช้ VoltDB และ NuoDB ทั้งหมดสนับสนุนการทำ REST API และการสนับสนุนการทำ

MapReduce ยกเว้น Graph database กับ NewSQL และการเปรียบเทียบภาษาสอบถาม (Query Language) มี NewSQL เท่านั้นที่สนับสนุน SQL และการเปรียบเทียบการทำงานเพื่อประเมินผลการปฏิบัติของประสิทธิภาพการทำงานของฐานข้อมูล NoSQL และ NewSQL ด้วยการเปรียบเทียบการปฏิบัติงานด้าน CRUD (Create, Read, Update, Delete) เลือกการ Update, Delete มาทำการทดลองด้วยขนาด 120 ล้านระเบียนเป็นขนาดเล็ก 1 KB แต่ละเรคคอร์ด จำนวน 6 โหนด และ 0.12 TB เป็นการเปรียบเทียบระหว่าง Cassandra, Hbase, Sherpa, MySQL ซึ่งมีผลดังนี้ การอับเดทในระเบียนขนาดใหญ่ Hbase ดีที่สุดและ Cassandra อันดับรองลงมา และในระเบียนขนาดเล็กผลการอ่าน Cassandra ดีที่สุดและ Hbase รองลงมา

ผลจากการวิจัยในระเบียนขนาดเล็ก ผลการอับเดท Hbase ดีที่สุดและ Cassandra รองลงมา และผลการอ่าน MySQL ดีที่สุดและรองลงมาคือ Sherpa บทสรุปของงานวิจัยนี้คือการศึกษาเปรียบเทียบรายละเอียดของ NoSQL และ NewSQL เก็บข้อมูลบนหลายพารามิเตอร์ทั้งทางด้านเทคนิคและไม่ใช่เทคนิค เปรียบเทียบตามในรูปแบบข้อมูล และความสามารถในการสอบถาม, การควบคุมการทำงานพร้อมกัน, การจำลองแบบปรับขยาย, กลยุทธ์การแบ่งรูปแบบสอดคล้องและคุณลักษณะด้านความปลอดภัย นอกจากนี้ยังกล่าวถึงกรณีการใช้งานความเหมาะสมและการใช้งานที่เป็นไปได้ซึ่งในแต่ละประเภทของ NoSQL ฐานข้อมูล NewSQL นอกจากนี้ยังทำการวิเคราะห์และการเปรียบเทียบเชิงปริมาณของเก็บข้อมูลต่างๆ บนพื้นฐานของผลที่ได้รับจากการใช้ระบบการเปรียบเทียบ YCSB แล้วทำการนำเสนอผลจากการทดลอง NoSQL / NewSQL เก็บข้อมูลในสภาพแวดล้อมการลอกเลียนแบบการกระทำของเครือข่ายทางสังคม (ชุดมาตรฐาน BG) ผลในทางปฏิบัติได้ข้อสรุปคือ จำนวนของ NoSQL/NewSQL ฐานข้อมูลเป็นประเภทที่แตกต่างกันทั้งความสามารถในการค้นหาข้อมูล และมีความหลากหลายของการใช้งานที่มีอยู่ และที่เพิ่งเกิดขึ้นใหม่ที่สามารถได้รับประโยชน์จากหลักการและเทคนิคการให้บริการโดยระบบเหล่านั้น การเปรียบเทียบในหมู่ผู้นิยมมากที่สุดเก็บข้อมูล NoSQL/NewSQL พร้อมกับรายละเอียดของกรณีการใช้งานที่เป็นไปได้ให้ไว้ในบทความนี้อาจจะให้ความช่วยเหลือเพิ่มเติมสำหรับ ผู้ปฏิบัติงานเกี่ยวกับการเลือกโซลูชั่นการจัดเก็บข้อมูลที่ดีที่สุดสำหรับความต้องการของโปรแกรมของผู้ใช้ข้อมูล

Vicknair, Macias, Zhendong, Zhao, Nan, Chen, and Wilkins (2010) มีงานวิจัยเรื่อง A comparison of a Graph Database and Relational database ปัญหาของงานวิจัยนี้คือ ฐานข้อมูลเชิงสัมพันธ์เป็นเทคโนโลยีฐานข้อมูลทางเลือกสำหรับการจัดเก็บข้อมูลแบบดั้งเดิมที่มีผู้นิยมใช้กันมากที่สุดและการเรียกใช้งานข้อมูลจำนวนมาก การสืบค้นมักจะใช้ SQL เป็นภาษาแบบสอบถามที่ใช้งานได้ดี ยกเว้นแต่ข้อมูลที่มีการกำหนดความสัมพันธ์ ให้ใช้ร่วมกันของตารางข้อมูล 2 ตารางขึ้นไปจะมีขนาดใหญ่ การใช้วิทยาการคอมพิวเตอร์เข้ามาช่วยแก้ไขปัญหา เช่น มีการใช้หลักการ

คณิตศาสตร์กราฟนำมาประยุกต์ใช้เริ่มแพร่หลายในสาขาชีววิทยา, เคมี, พันธุกรรม เป็นต้น ซึ่งเป็นแบบจำลองโครงสร้างที่มีประโยชน์แต่มาตรฐานนี้จะนำมาใช้งานได้อย่างไร

ผู้วิจัยฉบับนี้จึงนำเสนอการศึกษามาตรฐานของโครงสร้างฐานข้อมูลกราฟ ด้วยโปรแกรม Neo4j เพื่อนำมาใช้ในการจัดเก็บและการสอบถาม แล้วเปรียบเทียบกับฐานข้อมูลแบบเชิงสัมพันธ์ที่มีความนิยม MySQL ก่อนที่จะนำมาใช้ในการตัดสินใจเลือกใช้ในการจัดเก็บฐานข้อมูลใหม่ เพื่อใช้เป็นเทคโนโลยีพื้นฐานในการพัฒนาซอฟต์แวร์ระบบการบันทึกและการสอบถามข้อมูล สมมติฐานของงานวิจัยนี้คือ ที่มาของข้อมูลสามารถอยู่ในระดับรายละเอียดที่แตกต่างกัน รากของกราฟสามารถดูฐานข้อมูลทั้งหมดในระเบียน (Tuples) ในแฟ้มได้ และเป็นเรื่องง่ายที่จะใช้งานแบบสอบถามและมีประสิทธิภาพด้วยรูปแบบ Direct acyclic graph (DAG)

โดยมีวิธีการวิจัยและรูปแบบการวิจัยด้วยการเปรียบเทียบมาตรฐานต่างๆ เพื่อการตัดสินใจตามหัวข้อดังนี้ 1.การยอมรับและการสนับสนุนของผู้ใช้และผู้ผลิต (Maturity / Level of Support) 2.โปรแกรมใช้งานง่าย (Ease of Programming) 3.ความยืดหยุ่น (Flexibility) 4.ความปลอดภัยข้อมูล (Security) และทดลองเพื่อประเมินผลด้านผลความเร็วกับชุดข้อมูลที่กำหนดไว้ล่วงหน้า และทดสอบการสืบค้นข้อมูลด้วยการใช้ชุดแบบสอบถามที่กำหนดขึ้น มีตัวแปรที่สำคัญจำนวนข้อมูลที่เพิ่มขึ้น, รูปแบบจัดเก็บข้อมูลที่แตกต่างกัน, แบบสอบถามเหมือนกัน

และมีเครื่องมือวัดและวิธีเก็บข้อมูลด้วยเทคนิควิธีการ สร้างฐานข้อมูล MySQL และ Neo4J จำนวนเท่ากัน 12 ฐานข้อมูล ด้วยข้อมูลน้ำหนักการบรรทุก Neo4J มี Node (Nodeid, Payload) และ Edge (Source, Sink) และ MySQL เป็นรูปแบบข้อมูลกึ่งโครงสร้างมี 2 ตาราง ด้วยการใช้ XML และ JSON ในการเชื่อมสัมพันธ์การจัดเก็บข้อมูลและมีโครงสร้างเหมือนกับ Neo4J และมีจำนวนแต่ละชุดข้อมูลดังนี้ 1,000, 5,000, 10,000 และ 100,000 ตามลำดับ และในการประเมินด้วยแบบสอบถามโดยการใช้สุ่มเลือกจากตัวเลข และตัวอักษรแบบ 8KB และตัวอักษรแบบ 32KB และยังเก็บข้อมูลการใช้พื้นที่จัดเก็บในฮาร์ดดิสก์ที่จำเป็น และทำการสร้างดัชนีไว้ทั้งสองฐานข้อมูล และกำหนดให้แบบสอบถามมีชุดคำสั่งสำหรับค้นหาชุดตัวเลขดังนี้ 1.นับจำนวนโหนดที่มีน้ำหนักบรรทุกข้อมูลที่มีค่าเท่ากัน 2.นับจำนวนโหนดที่มีน้ำหนักบรรทุกน้อยกว่าค่าที่กำหนดเป็นตัวแปรตั้งต้น 3.แบบสอบถามมีชุดคำสั่งสำหรับค้นหาชุดตัวอักษรดังนี้ นับจำนวนโหนดที่มีข้อมูลตามที่กำหนดตั้งต้น กำหนดความยาวตัวอักษร 4-8 ตัว แล้วทำการทดสอบแบบสอบถามชุดละ 10 ครั้งในแต่ละฐานข้อมูลแล้วนำมาหาค่าเฉลี่ย เพื่อทำให้มั่นใจว่ากระบวนการแคชหรือระบบไม่ส่งผลกระทบต่อเวลาการค้นหาข้อมูล โดยมีสภาพแวดล้อมด้วยเครื่องเซิร์ฟเวอร์ที่มี OS เป็น Ubuntu Linux เวอร์ชั่น 9.10 และมี CPU 2 Duao 3.00 GHz และมี RAM 4GB และทำการเชื่อมต่อแต่ละเครื่องเข้ากับระบบอินเตอร์เน็ต และโปรแกรม MySQL เป็นรุ่น 5.1.421 และโปรแกรม Neo4J เป็น

รุ่น 4.0 b112 วิธีวิเคราะห์ข้อมูลด้วยการสร้างตารางเก็บผลการทดลองและนำวิเคราะห์เปรียบเทียบผลที่ได้จากการทดลอง

ผลจากการวิจัยทั้งสองระบบดำเนินการได้กับชุดแบบสอบถามที่ดำเนินการขึ้นแต่ลักษณะโดยทั่วไปฐานข้อมูลกราฟจะได้ผลประสิทธิภาพที่ดีกว่าในแบบสอบถามชนิดฐานข้อมูลเชิงสัมพันธ์ในค้นหาตัวอักษรข้อความแบบเต็มอย่างมีนัยสำคัญดีกว่าฐานข้อมูลเชิงสัมพันธ์ ซึ่งกลไกในการทำดัชนี Lucene มีผลกับตัวอักษรมากกว่าแบบตัวเลข ที่ฐานข้อมูลเชิงสัมพันธ์มีประสิทธิภาพมากกว่าฐานข้อมูลกราฟ ซึ่งเป็นปัญหาจากการทำดัชนีที่ยังเป็นข้อด้อยของฐานข้อมูลกราฟ ในปัจจัยอื่นที่สำคัญในการเลือกใช้งานคือการรักษาความปลอดภัยที่ฐานข้อมูลกราฟยังไม่ได้รับการสนับสนุนใน Neo4J ภายใต้มาตรการรักษาความปลอดภัย ACL Based (Access Control List) ก็ยังเป็นข้อด้อยอีกจุดหนึ่งที่ยังต้องมีการปรับปรุงจากผู้ผลิต

Appuswamy, Gkantsidis, Narayanan, Hodson, and Rowstron (2013) มีงานวิจัยเรื่อง Scale-up vs Scale-out for Hadoop: Time to rethink ปัญหาของงานวิจัยนี้คือ การวิเคราะห์ข้อมูลด้วยการใช้แมพรีดิวและฮาดูปเป็นการใช้งานการกระจายไฟล์ไปตามเครื่องคอมพิวเตอร์แบบส่วนบุคคลที่เป็นการเก็บข้อมูลในฮาร์ดดิสก์ที่ไม่น่าเชื่อถือ จึงมีโอกาสเกิดความผิดพลาดสูง หากแต่การใช้เก็บข้อมูลกับผู้ให้บริการที่มีเซิร์ฟเวอร์ประสิทธิภาพสูงจะมีผลลัพธ์ที่ดีกว่าในข้อมูลระดับเอ็กซ์ตาไบต์หรือเพตาไบต์หรือเทราไบต์

ผู้วิจัยฉบับนี้จึงมีคำถามว่ามันควรจะขยายออกในแบบกระจายหรือแบบขนาน (Scale out) หรือแบบขยายขึ้นดีกว่ากัน (Scale up) ผู้วิจัยฉบับนี้จึงนำเสนอ การวิจัยเชิงทดลองที่นำฮาร์ดแวร์ที่มีคุณสมบัติที่มีประสิทธิภาพสูงนำมาเปรียบเทียบกับฮาร์ดแวร์ที่มีประสิทธิภาพต่ำที่ใช้ฮาดูปและแมพรีดิว สมมติฐานของงานวิจัยนี้คือการใช้เซิร์ฟเวอร์ที่มีประสิทธิภาพเครื่องเดียวจะดีกว่าการมีเครื่องส่วนบุคคลที่มีหลายเครื่องที่ใช้ฮาดูป เนื่องจากฮาดูปจะมีประสิทธิภาพด้อยลงเมื่อข้อมูลมีจำนวนมากขึ้น และสมมุติฐานที่สองคือการปรับปรุงประสิทธิภาพให้กับเครื่องเซิร์ฟเวอร์ที่มีฮาดูปเพื่อรองรับการใช้งานในเครื่องเดียวและใช้ได้ดีกว่าคลัสเตอร์คอมพิวเตอร์

และการวิเคราะห์ข้อมูลด้วยฮาดูปและแมพรีดิวได้รับการออกแบบมาเพื่อการจัดการข้อมูลระดับเพตาไบต์ทำการประเมินผลกันระหว่างแบบขยายออกและแบบขยายขึ้น โดยมีวิธีการวิจัยโดยวิธีการทดลอง ทำการทดลองกับเครื่องเวิร์คสเตชั่นและเครื่องเซิร์ฟเวอร์สเป็คสูง และเซิร์ฟเวอร์ที่มีแกนสมอง 32 แกน หน่วยความจำ 512 GB เปรียบเทียบกับเวิร์คสเตชั่น 8 โหนด นำเสนอการประเมินผลว่าเซิร์ฟเวอร์ที่มีการขยายเพิ่มขึ้นมีค่าใช้จ่ายด้านพลังงานและการใช้พื้นที่การจัดเก็บเมื่อเทียบกับแบบ 16 โหนดคลัสเตอร์จะใช้ต้นทุนมากกว่ากันเท่าไร หรือจุดที่ต้องเริ่มมีการปรับเปลี่ยนจากการขยายขนาดขึ้นไปเป็นการขยายตามแนวนอน

เครื่องมือในการทดลองมีคลัสเตอร์คอมพิวเตอร์มีจำนวน 2 กลุ่ม คือจำนวน 8 โหนด และ 16 โหนด เป็นเวิร์คสเตชั่น ในแบบการขยายออก และแบบการขยายขึ้นใช้เซิร์ฟเวอร์ที่มีประสิทธิภาพสูงจำนวน 1 โหนด วิธีเก็บข้อมูลด้วยเทคนิควิธีการใช้ข้อมูลการใช้งาน (Log) ข้อมูลการวิเคราะห์แบบสอบถามจากข้อมูลการเข้าถึงเว็บไซต์ ข้อมูลจากมาตรฐานข้อมูล TeraSortข้อมูลจากการจัดเรียงและข้อมูลการเรียนรู้ของเครื่อง (Machine Learning) โดยการใช้โปรแกรม Mahout ที่มีขนาดข้อมูลแตกต่างกันไป โดยมีสภาพแวดล้อมจากการใช้แพลตฟอร์มของฮาดูป ทั้งในการทดสอบทั้งแบบการขยายออกและแบบการขยายขึ้น การเพิ่มประสิทธิภาพกับหน่วยจัดเก็บข้อมูล (Storage) โดยใช้ SSD และใช้ HDFS ในการจัดเก็บไฟล์และทำการกำหนดแบบกระจายในเครื่องเดียวด้วยแบบจำลองเครื่องในการขยายขึ้นและการกระจายไปที่เครื่องจริงแบบขยายออก และยังทำการกำหนดค่าที่เหมาะสมสำหรับงานต่อชุดข้อมูลจำนวนสูงสุดที่ 4 GB ต่อการทำแมพและรีดิวและทำการประเมินผล วิธีวิเคราะห์ข้อมูลด้วยการใช้ข้อมูลประสิทธิภาพผลการดำเนินงาน (Throughput) และหน่วยค่าใช้จ่ายฮาร์ดแวร์ทั้งเครื่องแบบเวิร์คสเตชั่นและเซิร์ฟเวอร์ และค่าใช้จ่ายด้านพลังงานในการใช้งาน เพื่อให้เข้าใจถึงข้อดีข้อเสียของการปรับแบบการขยายขึ้นเพื่อเทียบกับการปรับแบบขยายออก

ผลจากการวิจัยพบว่าการปรับปรุงกระบวนการฮาดูปในการขยายแบบขึ้นมีนัยสำคัญในการประเมินด้านการใช้พลังงานและการเพิ่มชั้นการจัดเก็บและต่อค่าใช้จ่าย และยังมีส่วนของการใช้งานหน่วยความจำที่มีประสิทธิภาพ และในการขยายแบบขึ้นยังดีกว่ากลุ่มคอมพิวเตอร์จำนวน 8 เครื่องในบางงานด้วย และการปรับปรุงกระบวนการโอนข้อมูลในระบบ HDFS ไป SSD มีผลต่อการปรับปรุงประสิทธิภาพและสิ่งที่กระบวนการแย่ที่สุดคือการนำเข้าข้อมูลมีผลกระทบต่อประสิทธิภาพมากที่สุด และในการทำทดสอบการทำเครื่องขยายแบบขนานระบบคลาวด์คอมพิวเตอร์มีการใช้แบนด์วิธในเครือข่ายทำให้ระบบไม่มีประสิทธิภาพในการทำงาน และการใช้การขยายออกเป็นแบบเฉพาะงานจะเป็นตัวเลือกที่ดีที่สุด

ในบทความวิจัยนี้จึงแสดงให้เห็นทิศทางตรงกันข้ามกับความรู้ดั้งเดิมว่างานวิเคราะห์โดย Hadoop และ MapReduce มักจะมีบริการที่ดีขึ้น โดยเซิร์ฟเวอร์ในการขยายขึ้นดีกว่ากลุ่มคลัสเตอร์ที่มีการทำงานการขยายออก ในกลุ่มการวิเคราะห์โดยทั่วไปควรจะมีการจัดเตรียมเซิร์ฟเวอร์แบบขยายขึ้นเป็นตัวเลือกที่ดีสำหรับงานจำนวนมาก ไม่ว่าจะเป็นในคลัสเตอร์ส่วนตัวหรือในบนระบบคลาวน์ การเพิ่มประสิทธิภาพที่นำเสนอในบทความนี้ให้เป็นจุดเริ่มต้นที่ดีสำหรับการปรับปรุงประสิทธิภาพการทำงานเซิร์ฟเวอร์ให้เป็นแบบขยายขึ้น (Scale up)

Singh and Reddy (2014) มีงานวิจัยเรื่อง A survey on platforms for big data analytics ปัญหาของงานวิจัยนี้คือ เมื่อข้อมูลขนาดใหญ่ส่งผลให้มีการเปลี่ยนแปลงต่อการการวิเคราะห์ข้อมูล

รูปแบบดั้งเดิม แพลตฟอร์มหรือกรอบการทำงานของซอฟต์แวร์และฮาร์ดแวร์แบบเดิมๆ ไม่สามารถดำเนินการวิเคราะห์ใดๆ ที่เกี่ยวข้องกับข้อมูลขนาดใหญ่และซับซ้อนได้ เช่น ฮาร์ดแวร์ต้องทำการปรับเปลี่ยนให้ทำงานได้ภายใต้การทำงานของข้อมูลขนาดใหญ่ และซอฟต์แวร์ที่มีแพลตฟอร์มสำหรับการวิเคราะห์ข้อมูลสำหรับการตัดสินใจที่สำคัญ ซึ่งปัญหาดังกล่าวเป็นพื้นฐานขององค์กรก่อนที่จะทำการตัดสินใจเลือกใช้แพลตฟอร์มของเทคโนโลยีที่ถูกต้อง โดยปกติเมื่อผู้ใช้ต้องการตัดสินใจเลือกแพลตฟอร์มที่เหมาะสม ผู้ใช้จะต้องตรวจสอบแพลตฟอร์มที่ต้องการใช้ในขั้นตอนและวิธีการใช้งานร่วมกับข้อมูลของพวกเขา เพื่อให้ตรงตามความต้องการของผู้ใช้และทำให้เกิดความพึงพอใจในการวิเคราะห์ข้อมูลให้ใช้เวลาที่เหมาะสม

ผู้วิจัยฉบับนี้จึงนำเสนอการแก้ปัญหาที่อยู่บนพื้นฐานของแพลตฟอร์มเฉพาะทางนี้ โดยการศึกษาอย่างละเอียดและทำความเข้าใจต่อแพลตฟอร์มข้อมูลขนาดใหญ่ที่กำลังเป็นที่นิยมและถูกนำมาทดลองการใช้งานอย่างกว้างขวาง เช่น โปรแกรม Apache Hadoop และ MapReduce และ Apache Pig และ Spark เป็นต้น

กลุ่มนักวิจัยจำนวนมากได้พยายามสร้างกรอบการทำงาน สร้างเทคนิคการวิเคราะห์ข้อมูลขนาดใหญ่มากขึ้นและทำการวิจัยกันมาอย่างต่อเนื่องเพื่อจะนำมาใช้งานได้จริง การพัฒนามีขั้นตอนวิธีการที่แตกต่างกันและความหลากหลายของฮาร์ดแวร์และซอฟต์แวร์ในการใช้งานข้อมูลขนาดใหญ่ที่มีอยู่มีลักษณะและการปฏิบัติที่แตกต่างกัน การเลือกแพลตฟอร์มเทคโนโลยีที่เหมาะสมต้องมีความรู้ในเชิงลึกเกี่ยวกับความสามารถของแพลตฟอร์มเหล่านี้ทั้งหมดและงานวิจัยนี้ยังเน้นให้เห็นถึงข้อดีและข้อเสียของแต่ละแพลตฟอร์ม โดยเฉพาะอย่างยิ่งความสามารถของแพลตฟอร์มที่จะปรับให้เข้ากับข้อมูลเฉพาะทางขององค์กรที่มีแนวโน้มเพิ่มขึ้นให้ได้ การประมวลผลจึงมีบทบาทสำคัญที่จะนำมาใช้ในการตัดสินใจว่าแพลตฟอร์มใดเหมาะสมที่จะนำมาใช้สร้างการวิเคราะห์ในทางปฏิบัติหรือไม่

สมมติฐานของงานวิจัยนี้คือ 1) จะทำอย่างไรให้ได้ผลลัพธ์ที่รวดเร็ว 2) มีขั้นตอนการดำเนินการอย่างไรกับข้อมูลขนาดใหญ่ 3) การสร้างต้นแบบต้องทำซ้ำหลายครั้งหรือสามารถทำครั้งเดียวได้ 4) มีความจำเป็นที่จะต้องใช้ความสามารถในการประมวลผลข้อมูลที่เริ่มจะมีมากขึ้นในอนาคตหรือไม่ 5) โปรแกรมควรให้ความสำคัญต่อการคำนวณอัตราการโอนถ่ายข้อมูล (I/O) หรือไม่ 6) โปรแกรมต้องคอยควบคุมจัดการความล้มเหลวของฮาร์ดแวร์หรือไม่

โดยมีรูปแบบการวิจัยและวิธีการวิจัยมุ่งเน้นไปที่การเปรียบเทียบแพลตฟอร์มทั้งหมดดังนี้ Scaling หรือการปรับขนาดของระบบให้ได้ตามความต้องการในแง่ของการประมวลผลข้อมูลสามารถแบ่งได้ออกเป็น 2 รูปแบบ คือ

1) Horizontal Scaling (การขยายตามแนวนอน) จะเป็นการกระจายภาระงานให้กับเซิร์ฟเวอร์จำนวนมาก ทรัพยากรในการขยายตามแนวนอนมีดังนี้

1.1) เครือข่ายในกลุ่มการขยายในแนวนอนระบบเครือข่าย Peer-to-Peer Network (เครือข่ายเพียร์ทูเพียร์) หรือ TCP/IP เป็นการเชื่อมต่อเครือข่ายแบบกระจายตามคอมพิวเตอร์หรือโหนด ซึ่งเครื่องคอมพิวเตอร์หรือโหนดในเครือข่ายสามารถเป็นได้ทั้งไคลเอนต์และเซิร์ฟเวอร์ในเครื่องเดียวและใช้ทรัพยากรร่วมกันได้ มีใช้การสื่อสารรูปแบบ Massage Passing Interface (MPI) ในการแลกเปลี่ยนข้อมูลและเหมาะกับการประมวลผลซ้ำหลายครั้ง (Iterative Processing) ซึ่งเป็นรูปแบบเครือข่ายที่เหมาะกับการพัฒนาอัลกอริทึมในการวิเคราะห์ข้อมูลขนาดใหญ่ แต่ก็ยังมีข้อบกพร่องในการทำการป้องกันความล้มเหลวของการสื่อสารหรือ (Fault Tolerance) ดังนั้นจึงต้องใช้การจัดการทางซอฟต์แวร์ป้องกันความผิดพลาดแทน จึงทำให้ Apache Hadoop ซึ่งมีคุณสมบัติดังกล่าวมีประสิทธิภาพมากและกลายเป็นที่นิยมแพร่หลายในงานวิจัย ซอฟต์แวร์ที่จัดอยู่ในกลุ่มการขยายตามแนวนอนที่ทำการศึกษามี

1.2) Apache Hadoop ซึ่งมีองค์ประกอบสำคัญคือ Hadoop Distributed File System (HDFS) หรือระบบกระจายไฟล์ที่ใช้ในการจัดเก็บ และ Hadoop YARN ในการจัดการทรัพยากรและจัดตารางงานของกลุ่มคลัสเตอร์คอมพิวเตอร์

1.3) การประมวลผล MapReduce เป็นการเขียนโปรแกรมการประมวลผลข้อมูลที่ใช้ใน Hadoopที่มีลักษณะการทำงาน 2 ลักษณะคือ Map และ Reduce ที่อ่านข้อมูลจาก HDFS ซึ่งเป็นการประมวลผลข้อมูลแบบขนานจากกลุ่มโหนดในคลัสเตอร์แล้วทำการรวบรวมข้อมูลมาแสดงผลลัพธ์สุดท้าย

1.4) MapReduce Wrappers เป็นการเขียนควบคุมพัฒนาดีขึ้นมากกว่า MapReduce ด้วยสภาพแวดล้อมของ SQL เช่น Apache Pig ถูกพัฒนาด้วย Yahoo และ Apache Hive ที่ถูกพัฒนาโดย Facebook เพื่อให้มีมาตรฐานการประมวลผลที่ดีกว่าลดความซับซ้อนของการเขียนโปรแกรมแบบ MapReduce และยังมี DryadLINQ ที่ถูกพัฒนาให้มีความยืดหยุ่นในการใช้งานมากขึ้นใช้งานร่วมกับภาษา C# และ LINQ ที่ใช้พัฒนาร่วมกับภาษา Visual Studio.NET และนักวิจัยบางกลุ่มยังทำการพัฒนา Apache Mahout เพื่อการเรียนรู้เครื่องจักร (Machine Learning) โดยอาศัยรูปแบบการทำงานหรือกระบวนทัศน์การทำงานแบบ MapReduce แต่ทั้งนี้ MapReduce ยังมีข้อจำกัดในด้านการประมวลผลที่ไม่สามารถสร้างอัลกอริทึมการประมวลผลแบบซ้ำแล้วซ้ำอีกได้มันเป็นการประมวลผลแบบกลุ่ม (Batch Processing) หรือจะกระทำใหม่ทุกครั้งที่มีการสั่งรันโปรแกรมใหม่ทำให้ประสิทธิภาพในการเข้าถึงข้อมูลลดลงจึงถือว่าเป็นค่าใช้จ่ายที่เกิดขึ้นในอนาคต จึงมีการพัฒนาแก้ปัญหาดังกล่าว เช่น HaLoop เป็นปรับปรุงประสิทธิภาพด้วยการสร้างอัลกอริทึมให้

สามารถใช้งานหน่วยความจำสำรอง (Cache) มาทำงานร่วมกันเพื่อเก็บข้อมูลที่ต้องการเรียกใช้บ่อย หรือ iMapReduce ของ Twister ที่ทำงานในลักษณะการเก็บในหน่วยความจำสำรอง

1.5) Spark เป็นรูปแบบการประมวลผลข้อมูลขนาดใหญ่แบบใหม่ได้มีการพัฒนาโดยกลุ่มนักวิจัยมหาวิทยาลัยแคลิฟอร์เนียทำงานร่วมกับทีมงานกลุ่มพัฒนา Hadoop เพื่อปรับปรุงประสิทธิภาพในการทำงานของระบบ I/O ให้ดีขึ้น มีลักษณะการทำงานร่วมกับหน่วยความจำ ด้วยการคำนวณในหน่วยความจำและมีภาษาทำงานร่วมกันได้คือ Java, Scala, Python และมีการพิสูจน์แล้วว่ามีความเร็วกว่า MapReduce ถึง 100 เท่า เมื่อข้อมูลใช้งานในหน่วยความจำและถึง 10 เท่า เมื่อข้อมูลอยู่บนฮาร์ดดิสก์

1.6) BDAS (Berkeley Data Analytics Stack) เป็นกรอบการทำงานในกระบวนทัศน์ของ Spark ที่พัฒนาขึ้นมาเพื่อทำการวิเคราะห์ข้อมูลหรือจะเรียกอีกชื่อว่า Tachyon มีประสิทธิภาพที่ใช้หน่วยความจำในระดับ I/O มากขึ้นสามารถอ่านไฟล์บ่อยและเก็บในหน่วยความจำแคชจึงลดการเข้าถึงฮาร์ดดิสก์ในงานที่แตกต่างกันได้ และสนับสนุนการจัดการตารางข้อมูลหลายร้อยคอลัมน์และสามารถรองรับการประมวลผลแบบทันที (Real-time stream processing) และการตั้งเวลาการทรัพยากรได้ (Multi-resource scheduling capabilities) และยังใช้งานร่วมกับ Amazon Elastic ได้อีกด้วย จึงเป็นที่นิยมในการใช้งานเพิ่มมากขึ้น

2) Vertical Scaling (การขยายตามแนวตั้ง) จะเป็นการติดตั้งหน่วยประมวลผลมากขึ้นอีกทั้งหน่วยความจำและฮาร์ดแวร์ให้เร็วขึ้นโดยทั่วไปจะทำภายใต้เซิร์ฟเวอร์เดียว

2.1) High Performance Computing (HPC) Clusters, Blades หรือ Super Computer เป็นคอมพิวเตอร์ระดับสูงที่ไม่สามารถนำมาใช้งานร่วมกับ Hadoop หรือ Spark ได้ มีค่าใช้จ่ายในการจัดซื้ออุปกรณ์สูง

2.2) Multicore CPU หรือเครื่องที่มีหน่วยประมวลผลหลัก CPU จำนวนมาก ทำงานแบบคู่ขนานมีหลายแกนสมอง เนื่องจากมีการพัฒนาบอร์ดแบบใหม่ๆ ที่มารองรับเพิ่มขึ้น และมีการเรียกใช้งานผ่านโปรแกรมที่นิยม เช่น Java เป็นต้น แต่ยังมีข้อด้อยคือเมื่อขนาดของข้อมูลเกินกว่าหน่วยจำของระบบและการเข้าถึงฮาร์ดดิสก์ (I/O) จะกลายเป็นคอขวดขนาดใหญ่ (Huge Bottleneck) ต้องใช้การทำงานร่วมกันกับ DDR5 ห้ามใช้ DDR3 และใช้ร่วมกับ GPU จะช่วยเพิ่มความเร็วในการเข้าถึงข้อมูล

2.3) Graphics processing unit (GPU) หรือหน่วยประมวลผลกราฟฟิกเป็นการสร้างภาพในเฟรมบนบัฟเฟอร์ในการประมวลผลภาพของวิดีโอและรูปภาพ ถูกนำมาใช้งานร่วมกับขั้นตอนวิธีการเรียนรู้ของเครื่องจักรให้รวดเร็วขึ้น ซึ่งข้อจำกัดคือการต้องใช้หน่วยความจำสูงสุดได้

เพียง 12 GB จึงไม่เหมาะกับการจัดการข้อมูลขนาดเทราไบต์ จึงเป็นจุดคอขวดหากทำการประมวลผลข้อมูลขนาดใหญ่

2.4) Field Programmable Gate Arrays (FPGA) เป็นฮาร์ดแวร์ที่สร้างขึ้นสำหรับการใช้งานเฉพาะและใช้แพลตฟอร์มภาษา Hardware Descriptive Language (HDL) ตัวอย่างการใช้งานจริงคือการนำมาใช้งานป้องกันเครือข่ายหรือไฟร์วอลล์ฮาร์ดแวร์ มีการทำงานกับข้อมูลปริมาณมากในการสแกนข้อมูลบนเครือข่าย มีวิธีการประเมินผลด้วยการวิเคราะห์ตามสมมติฐานและมีสภาพแวดล้อมการวิเคราะห์ภายใต้การใช้งานการวิเคราะห์ข้อมูลขนาดใหญ่แล้วให้คะแนนแต่ละแพลตฟอร์มข้อมูลขนาดใหญ่ ขึ้นอยู่กับลักษณะที่เหมาะสมต่างๆ เหล่านี้ กลุ่มระบบและแพลตฟอร์ม ก) การขยายขีดความสามารถ (Scalability) ข) ประสิทธิภาพของการรับเข้า/ส่งออกข้อมูล (Data I/O performance) ง) ความคงทนต่อความล้มเหลวของการทำงาน (Fault tolerance) กลุ่มซอฟต์แวร์และอัลกอริทึม ก) การประมวลผลแบบทันที (Real-time processing) ข) ขนาดของข้อมูลที่รองรับ (Data size supported) ค) กระบวนการทำงานซ้ำหรือเก็บข้อมูลในหน่วยความจำสำรอง (Iterative task support) และให้คำแนะนำบางอย่างเกี่ยวกับความเหมาะสมของแพลตฟอร์มที่แตกต่างสำหรับทุกชนิดของสถานการณ์ที่เกิดขึ้นขณะที่กำลังทำวิเคราะห์ข้อมูลขนาดใหญ่ในทางปฏิบัติเพื่อที่จะให้ความเข้าใจที่ครอบคลุมมากขึ้นในแง่มุมที่แตกต่างกันของปัญหาข้อมูลขนาดใหญ่และวิธีการที่พวกเขาจะถูกจัดการ

งานวิจัยนี้ใช้กรณีศึกษาเกี่ยวกับการดำเนินงานของ K-Mean หมายถึงขั้นตอนวิธีการจัดกลุ่มต่างๆ แพลตฟอร์มข้อมูลขนาดใหญ่ K-Mean Clustering วิธีการจัดกลุ่มได้รับการคัดเลือกที่นี่ไม่เพียงเพราะของมันนิยม แต่ยังเกิดจากมิติต่างๆ ของความซับซ้อนที่เกี่ยวข้องกับขั้นตอนวิธี เช่น การทำซ้ำคำนวณจำนวนมากและมีความสามารถในการทำคู่ขนานบางส่วนของการคำนวณ และจัดให้มี Pseudo Code รายละเอียดของการดำเนินการของขั้นตอนวิธี K-Mean กับฮาร์ดแวร์และซอฟต์แวร์ที่แตกต่างแพลตฟอร์มและจัดให้มีการวิเคราะห์ในเชิงลึกและข้อมูลเชิงลึกในรายละเอียดขั้นตอนวิธีวิเคราะห์ข้อมูลขนาดใหญ่ในทางปฏิบัติ

ผลจากการวิจัย การเปรียบเทียบระบบและแพลตฟอร์มกลุ่มการขยายตามแนวนอน ตามหัวข้อการทดลองดังนี้ 1) ขีดความสามารถ (Scalability) ได้รับคะแนนดีที่สุดคือเครือข่าย Peer-to-Peer, MapReduce และ Spark กลุ่มการขยายตามแนวตั้ง HPC ได้การประเมินดีที่สุดแต่ได้รับคะแนนน้อยกว่าการขยายตามแนวนอน 2) การรับเข้า/ส่งออกข้อมูล (I/O) กลุ่มการขยายแนวตั้งได้คะแนนในกลุ่ม GPU, FPGA มากกว่าการขยายแนวนอนเป็นเพราะการที่การประมวลผลแบบขนานอย่าง Hadoop ต้องดำเนินการถ่ายโอนข้อมูลมากกว่าโดยไม่มีการใช้หน่วยความจำและการประมวลผล Spark มีคะแนนดีกว่า MapReduce 3) การคงทนต่อความล้มเหลวของระบบ (Fault

tolerance) กลุ่มการขยายตามแนวนอนได้คะแนนดีกว่าแต่จะแตกต่างกับกลุ่มการขยายตามแนวตั้งเพียงเล็กน้อย ซึ่งคะแนนของ Spark และ MapReduce ซึ่งใช้กรอบการทำงานของ Hadoop ซึ่งมีกลไลควบคุมความผิดพลาดมีคะแนนเท่ากัน และในกลุ่มขยายตามแนวตั้ง 4) การประมวลผลแบบทันที (Real-time processing) กลุ่มการขยายตามแนวตั้งในกลุ่ม GPU, FPGA (HDL) ได้ดีกว่ากลุ่มการขยายในแนวนอนเพราะการประมวลผลร่วมกับหน่วยความจำของเครื่องจึงเหมาะกับการประมวลผลในเวลาจริงมากกว่า 5) การสนับสนุนขนาดของข้อมูล (Data size supported) กลุ่มการประมวลแบบขนานมีคะแนนที่ดีกว่าซึ่งระดับสูงที่สุดเป็นกลุ่ม Peer-to-Peer (TCP/IP) ซึ่งในทางทฤษฎีจะรองรับข้อมูลได้ไม่จำกัด และ MapReduce กับ Spark สามารถรับรองการทำงานได้หลายหมื่นโหนดและยังสามารถประมวลผลและจัดการชุดข้อมูลขนาดใหญ่ได้ที แต่ HPC ในกลุ่มการขยายแบบแนวตั้งสามารถรองรับข้อมูลขนาดเทราไบต์ได้ซึ่งมีคะแนนเท่ากับ MapReduce กับ Spark 6) การสนับการทำงานซ้ำ (Iterative tasks support) กลุ่มการขยายแบบแนวตั้งจะได้คะแนนดีกว่าทุกกลุ่ม HPC, Multicore, GPU, FPGA จะมีคะแนนเท่ากันทุกกลุ่ม แต่จะแตกต่างกว่ากลุ่มการขยายแนวนอนอย่าง Spark เพียงเล็กน้อย และ Spark มีคะแนนที่ดีกว่า MapReduce เพราะการทำงานเป็นลักษณะการทำงานจะต้องเขียนผลข้อมูลลงดิสก์ทุกครั้ง ในกลุ่มการขยายแบบแนวตั้งจึงเหมาะสมกับการทำงานซ้ำมากกว่า

ผลการวิเคราะห์ด้วย K-Mean เพื่อใช้ในการวิเคราะห์ข้อมูลขนาดใหญ่สำหรับการตัดสินใจมีการเลือกโดยปัจจัย 1.ขนาดของข้อมูล (Data size) 2.ความเร็วหรือการเพิ่มประสิทธิภาพการส่งข้อมูล (Speed or throughput optimization) 3.การฝึกอบรมหรือการใช้แบบจำลอง (Training / Applying a model) 4.ทดสอบข้อมูลการปฏิบัติงานจริง (Practical implications) การทดสอบการประมวลผลเซนทรอยด์ (Centroids) ด้วยข้อมูลจาก Datapoints ด้วยการทดสอบบนแพลตฟอร์ม MapReduce, MPI, GPU

สรุปการวิเคราะห์ด้วย K-Mean จะดีที่สุดเพราะลักษณะของการประมวลผลที่ต้องการทำซ้ำหลายครั้งเพื่อให้เซนทรอยด์มาบรรจบกัน สรุปข้อดีข้อเสียของการปรับขนาดของแพลตฟอร์มความสามารถในการปรับขนาดของแพลตฟอร์มแบบแนวตั้งจะต้องมีการลงทุนทางการเงินเพื่อจัดการปริมาณงานในอนาคตด้วยการปรับเพิ่มฮาร์ดแวร์แต่เนื่องจากข้อจำกัดของเซิร์ฟเวอร์ที่มีช่องสำหรับเพิ่มหน่วยความจำหรือฮาร์ดดิสก์หรือหน่วยประมวลผลที่ไม่สามารถเพิ่มขึ้นได้อีก แต่ความสามารถของการปรับขนาดของแพลตฟอร์มแบบแนวนอนนั้นช่วยในการเพิ่มประสิทธิภาพการทำงานทีละน้อยและช่วยลดเงินลงทุน เป็นการเพิ่มขยายเพิ่มขึ้นได้ ตามความต้องการขยายเซิร์ฟเวอร์ตามความจำเป็น แต่ยังมีข้อเสียคือความพร้อมของซอร์ฟแวร์ที่จะนำมาใช้รองรับการทำงานให้เกิดประสิทธิภาพและเกิดประโยชน์มากที่สุด การเลือกแพลตฟอร์มที่เหมาะสมสำหรับ

การใช้งานเฉพาะขึ้นอยู่กับความต้องการใช้งานข้อมูลเฉพาะองค์กร หรืออาจจะใช้หลายแพลตฟอร์มร่วมกัน เช่น ใช้ Hadoop ร่วมกับ GPU เป็นต้น

Sareen and Kumar (2015) มีงานวิจัยเรื่อง NoSQL Database and Its Comparision with SQL Database ปัญหาของงานวิจัยนี้คือ ฐานข้อมูล NoSQL เป็นทางเลือกที่เกิดขึ้นใหม่ มีกลไกสำหรับการจัดเก็บและการดึงข้อมูลที่มีการสร้างแบบจำลองในวิธีการอื่นๆ ไม่เหมือนกับที่ใช้กันในฐานข้อมูลเชิงสัมพันธ์บางครั้งโครงสร้างข้อมูลที่ใช้ฐานข้อมูล NoSQL จะถูกมองว่ามีความยืดหยุ่นมากกว่าตารางฐานข้อมูลเชิงสัมพันธ์ นำไปใช้กับข้อมูลขนาดใหญ่และใช้งานกับเว็บแบบ Real-time อุปสรรคคือการยอมรับในการใช้ภาษาที่จะนำมาแทนที่ SQL และขาดการเชื่อมสัมพันธ์ที่ไม่สามารถทำการ Join ข้ามตารางได้ การใช้งานได้บางส่วนจึงเกิดปัญหาของการเลือกใช้ฐานข้อมูล NoSQL ว่าจะเลือกใช้อย่างไรจึงจะมีความเหมาะสม

ผู้วิจัยฉบับนี้จึงนำเสนอแนวคิดการหาความสอดคล้องในการออกแบบของฐาน NoSQL กับข้อมูลเพื่อปรับให้สามารถใช้งานได้ สมมติฐานของงานวิจัยนี้คือ ฐานข้อมูล SQL และ NoSQL แตกต่างกันอย่างไรและมีอะไรบ้างที่แตกต่าง และอรรถประโยชน์ของฐานข้อมูล NoSQL มีอะไรบ้าง โดยมีวิธีการวิจัยและรูปแบบการวิจัยด้วยการศึกษาประเภทและโครงสร้างของฐานข้อมูล NoSQL แล้วทำการเปรียบเทียบฐานข้อมูลที่นิยมและเป็นที่รู้จักกันอย่างดี เช่น Microsoft SQL และ MongoDB

มีวิธีวิเคราะห์ข้อมูลด้วยการเปรียบเทียบฐานข้อมูล MS SQL กับ MongoDB โดยมีหัวข้อในการประเมินผลพิจารณาในการเลือกกรอบการทำงาน NoSQL ดังนี้ 1) Workload diversity หรือความหลากหลายของภาระงาน เช่น การทำ Real-time และวิเคราะห์ข้อมูลได้ทันที 2) Scalability หรือการขยายของระบบสามารถรองรับและยืดหยุ่นในสถานการณ์จำเป็น 3) Performance หรือการใช้งานได้รวดเร็วมีประสิทธิภาพในการทำงาน 4) Continuous Availability หรือความพร้อมใช้งานได้อย่างต่อเนื่องข้อมูลสามารถใช้งานได้ตลอดเวลา 24 ชั่วโมง 5) Manageability หรือการบริหารจัดการในการพัฒนาและการเก็บรักษาข้อมูลหรือย้ายข้อมูลเข้าฐานข้อมูล NoSQL 6) Cost หรือค่าใช้จ่ายในการโยกย้ายหรือการขึ้นใช้งานและการพัฒนาเพื่อการใช้งานฐานข้อมูล NoSQL 7) Strong Community หรือชุมชนผู้ใช้ระบบที่องค์กรเหล่านั้นต้องการใช้งานเพื่อที่จะได้มีบุคลากรมารองรับและสนับสนุนช่วยเหลือทางด้านเทคนิคการใช้งาน ให้สามารถใช้ทรัพยากรที่ขึ้นระบบได้อย่างคุ้มค่า

มีการอภิปรายผลประเภทของฐานข้อมูล NoSQL แบ่งเป็น 4 หัวข้อดังนี้ 1) Document Database 2) Graph stores 3) Key-value stores 4) Wide-column stores และโครงสร้างของ MongoDB เป็นการจัดเก็บข้อมูลด้วยรูปแบบระเบียนเอกสารเก็บไว้ในไบนารี JSON มีสคีมาและ

หรือคอลเลคชั่นที่เขียนควบคุมด้วยภาษา JSON มีโครงสร้างข้อมูลและรูปแบบอาร์เรย์ที่คล้ายของฐานข้อมูลเชิงสัมพันธ์ และยังสามารถทำดัชนีและแบบสอบถามค้นคืนข้อมูล ดัชนีสามารถประกาศในช่องที่ไม่ซ้ำกันทั้งดัชนีเดียวหรือหลายดัชนีได้ และสามารถอยู่ในเขตข้อมูลที่มีโครงสร้างซ้อนกันได้ และยังรองรับการทำป้องกันความล้มเหลวของข้อมูลด้วยการทำเซิร์ฟเวอร์รองรับไว้ 3 เครื่องได้ และอรรถประโยชน์ของ NoSQL แบ่งเป็นหัวข้อดังนี้ 1) Elastic scaling หรือการปรับความยืดหยุ่นได้ผู้บริหารข้อมูลไม่จำเป็นต้องซื้อเซิร์ฟเวอร์ใหม่และไม่จำเป็นต้องใช้ฮาร์ดแวร์ที่มีประสิทธิภาพสูงเพื่อรองรับการทำงาน 2) Big Data หรือรองรับข้อมูลขนาดใหญ่ที่เริ่มจะมีแนวโน้มปริมาณข้อมูลที่เพิ่มขึ้น รองรับการทำงานข้อมูลได้โดยไม่จำเป็นต้องใช้ RDBMS 3) No DBAs หรือการไม่ต้องใช้ผู้เชี่ยวชาญดูแลฐานข้อมูล ที่ต้องมีการบริหารจัดการเมื่อฐานข้อมูลเริ่มมีขนาดใหญ่ขึ้น หรือไม่จำเป็นต้องใช้ผลิตภัณฑ์ซอฟต์แวร์ฐานข้อมูลที่มีประสิทธิภาพระดับสูง 4) Economics หรือด้านเศรษฐกิจการใช้ NoSQL ไม่จำเป็นต้องใช้เซิร์ฟเวอร์ที่มีราคาแพงและระบบจัดเก็บข้อมูลราคาแพงค่าใช้จ่ายต่อกิกะไบต์หรือการประมวลผลรายการต่อวินาทีถือได้ว่าน้อยกว่า RDMS มาก 5) Flexible data models หรือรูปแบบของข้อมูลมีความยืดหยุ่นการดูแลรักษาระบบข้อมูล RDBMS นั้นแม้มีการเปลี่ยนแปลงโครงสร้างข้อมูลเพียงเล็กน้อยจะกระทบต่อข้อมูลทั้งหมดต้อง เช่น การเพิ่มคอลัมน์ แต่หากเป็นระบบข้อมูล NoSQL สามารถเพิ่มได้ทันทีโดยไม่ต้องหยุดระบบทั้งหมด

สรุปผลจากการวิจัยฐานข้อมูล NoSQL มีความสำคัญมากขึ้น และจะเป็นส่วนหนึ่งของภูมิทัศน์ฐานข้อมูลและเมื่อใช้อย่างเหมาะสมทำให้เกิดประโยชน์ที่แท้จริง อย่างไรก็ตามผู้ประกอบการควรดำเนินการด้วยระมัดระวัง ด้วยการเรียนรู้ถึงรูปแบบของข้อดีข้อเสียและข้อจำกัดกับขององค์กร และควรจะศึกษาการใช้งานให้ถูกต้องตามกฎหมายในประเด็นที่เกี่ยวข้องกับฐานข้อมูลเหล่านี้

ผุสดี บุญรอด และ ประกายมาศ ศรีสุขทักษิณ (2558) มีงานวิจัยเรื่อง การค้นคืนข้อมูลขนาดใหญ่โดยใช้ภาษาสอบถามแบบไม่มีโครงสร้างร่วมกับเทคโนโลยีเว็บเชิงความหมาย ปัญหาของงานวิจัยนี้คือการค้นคืนข้อมูลให้ตรงตามความต้องการของผู้ใช้มีความจำเป็นอย่างมากในปัจจุบันเนื่องจากข้อมูลมีปริมาณมากขึ้นทำให้การประมวลช้าและการค้นคืนไม่ตรงตามความต้องการของผู้ใช้งาน ผู้วิจัยฉบับนี้จึงนำเสนอฐานข้อมูลไม่สัมพันธ์นำมาประยุกต์ใช้กับการจัดการข้อมูลขนาดใหญ่โดยใช้แบบภาษาสอบถามแบบไม่มีโครงสร้าง และโครงสร้างออนโทโลยีที่เป็นเทคโนโลยีเว็บเชิงความหมายมาช่วยค้นคืนข้อมูลให้มีประสิทธิภาพ สมมติฐานของงานวิจัยนี้คือแนวคิดของเทคโนโลยีเว็บเชิงความหมายสามารถนำมาช่วยในการค้นคืนข้อมูลให้ตรงตามความต้องการของผู้ใช้งานมากยิ่งขึ้น

โดยมีวิธีการวิจัยการรวบรวมข้อมูลและวิเคราะห์ปัญหาในการจัดการข้อมูลขนาดใหญ่ โดยการใช้ภาษาสอบถามแบบไม่มีโครงสร้าง ซอฟต์แวร์ที่ใช้ในการพัฒนาฐานข้อมูลไม่สัมพันธ์ และวิธีการค้นคืนข้อมูล และมีรูปแบบการวิจัย โดยการสร้าง 3 ส่วนหลักในการทำงาน ชั้นติดต่อกับผู้ใช้งานใช้บันทึกข้อมูลจัดเก็บลงในฐานข้อมูล และส่วนของเว็บเชิงความหมาย เป็นส่วนของการประมวลผลการค้นคืนข้อมูลโดยใช้โครงสร้างออนโทโลยีกำหนดความสัมพันธ์ระหว่างข้อมูล และส่วนสุดท้ายการบันทึกข้อมูลลงฐานข้อมูล MongoDB มีตัวแปรที่สำคัญคือข้อมูลลูกค้า ประเภทลูกค้า ผลิตภัณฑ์ ประเภทบริการ และหน่วยงานที่รับผิดชอบ และโครงสร้างออนโทโลยี เครื่องมือวัดวิธีเก็บข้อมูลด้วยเทคนิควิธีการด้วยการพัฒนาโครงสร้างออนโทโลยีเพื่อนำจัดเก็บและรองรับการค้นคืนข้อมูลขนาดใหญ่และการสร้างภาษาสอบถามแบบไม่มีโครงสร้างด้วยเทคโนโลยีเว็บเชิงความหมาย โดยมีสภาพแวดล้อมจากเทคโนโลยีออนโทโลยีตามวงจรชีวิต Ontology Life Cycle ใช้ซอฟต์แวร์ Hozo Ontology Editor เวอร์ชั่น 5.2.36 และเก็บโครงสร้างออนโทโลยีในฐานข้อมูล MongoDB เวอร์ชั่น 2.4.5 และทำการออกแบบโครงสร้างเอกสารในรูปแบบ JSON แบ่งข้อมูลออกเป็น 3 ส่วนชื่อ (Subject), คุณสมบัติ (Predicate), วัตถุ (Object)

วิธีวิเคราะห์ข้อมูลด้วยการทดสอบความเร็วในการประมวลผลแบ่งการทดสอบโดยใช้ 3 ตัวดำเนินการ คือการอ่าน (Select), การแก้ไข (Update) และการลบ (Delete) และทดสอบกับจำนวนเธรดที่แตกต่างกัน 1, 10 , 100 และการทดสอบการค้นคืนกับจำนวนระเบียนที่แตกต่างกัน 100, 1,000, 10,000, 100,000 ตามลำดับ และทำการคำนวณผลจากเวลาทั้งหมดในการประมวลผลและใช้การหาค่าเฉลี่ยเลขคณิตจากการทดลองจำนวน 3 ครั้ง

ผลจากการวิจัยโดยพิจารณาจากจำนวนเธรดว่าสามารถทำงานได้มากกว่าหนึ่งโปรแกรมในเวลาเดียว และจำนวนข้อมูล อีกทั้งใช้เวลาทั้งหมดในการประมวลผล ความเร็วในการอ่านข้อมูลจากฐานข้อมูลไม่สัมพันธ์แต่การอ่านข้อมูลจำนวนเธรด 100 กับจำนวนข้อมูล 100,000 แถว ใช้เวลาในการประมวลผลมากที่สุด แต่ในส่วนที่เหลือใช้เวลาใกล้เคียงกัน และในส่วนของการแก้ไขข้อมูลใช้เวลาในการประมวลผลกับฐานข้อมูลไม่สัมพันธ์ใช้เวลาในทุกข้อมูลและทุกเทรดได้ใกล้เคียงกัน และในส่วนของการลบข้อมูลใช้เวลาในการประมวลผลกับฐานข้อมูลไม่สัมพันธ์ใช้เวลาในลบข้อมูล ของฐานข้อมูลไม่สัมพันธ์ใช้เวลาในทุกข้อมูลและทุกเทรดได้ใกล้เคียงกัน จึงสรุปได้ว่าการออกแบบและพัฒนาฐานข้อมูลไม่สัมพันธ์ เพื่อรองรับการจัดเก็บข้อมูลและการค้นคืนข้อมูลขนาดใหญ่ โดยใช้ภาษาสอบถามแบบไม่มีโครงสร้างร่วมกับโครงสร้างออนโทโลยีที่เป็นเทคโนโลยีเว็บเชิงความหมาย ใช้กรณีศึกษาการออกแบบใบแจ้งค่าใช้บริการทำการจัดเก็บข้อมูลลงในฐานข้อมูลไม่สัมพันธ์แบบเอกสารโดยการจัดเก็บในรูปแบบเอกสาร JSON และได้ทำการทดลองเพื่อตรวจสอบประสิทธิภาพความเร็วในการประมวลผล จาก 3 ตัวดำเนินการ ได้แก่การอ่าน, แก้ไข

และการลบ ผลปรากฏว่าที่มีการแตกต่างกันมากมีเพียงแต่การดำเนินการด้านการอ่านเท่านั้น ฐานข้อมูลไม่สัมพันธ์จึงมีความเหมาะสมในการจัดเก็บและค้นคืนข้อมูลขนาดใหญ่ ซึ่งสามารถประมวลผลได้อย่างรวดเร็ว มีรูปแบบโครงสร้างคล้ายคลึงกับฐานข้อมูลเชิงสัมพันธ์ แต่ยังไม่เหมาะสมกับการใช้งานร่วมกับเว็บเชิงความหมายซึ่งต้องผู้วิจัยทำการปรับปรุงการออกแบบเพื่อให้สอดคล้องต่อไป

นิรุทธ์ รวยรื่น และ เกรียงไกร ปอแก้ว (2557) มีงานวิจัยเรื่อง การใช้แมพรีดิวซ์เชื่อมคอลเลคชั่นของฐานข้อมูลโนเอสคิวแอล (NoSQL) บนมองโกดีบี (MongoDB) งานวิจัยฉบับนี้ มีปัญหางานวิจัยดังนี้ปัจจุบันปริมาณจำนวนข้อมูลที่จัดเก็บในฐานข้อมูลมีอัตราการเติบโตที่สูง เมื่อข้อมูลที่จัดเก็บมีจำนวนมาก ทำให้การจัดเก็บลงฐานข้อมูลแบบเชิงสัมพันธ์และการจัดการข้อมูลไม่มีประสิทธิภาพเท่าที่ควร เนื่องจากรูปแบบการจัดเก็บข้อมูลมีลักษณะโครงสร้างที่ซับซ้อนต้องใช้เวลานานในการจัดการเข้าถึงข้อมูลที่มีปริมาณมาก จึงมีการนำเทคโนโลยีฐานข้อมูล NoSQL เข้ามาใช้ในการจัดเก็บและจัดการข้อมูลที่มีจำนวนปริมาณมากนั้น การเก็บข้อมูลในฐานข้อมูล NoSQL นั้นจะเป็นการเก็บข้อมูลในลักษณะไม่มีความสัมพันธ์เชิงโครงสร้าง ซึ่งฐานข้อมูล NoSQLไม่รองรับการเชื่อมความสัมพันธ์ของข้อมูล งานวิจัยนี้จึงได้เสนอแนวคิดสร้างความสัมพันธ์ของข้อมูลในฐานข้อมูล NoSQL ด้วยฐานข้อมูล MongoDB โดยนำเทคนิคหลักการเขียนโปรแกรมแบบแมพรีดิวมาประยุกต์ใช้ในการเชื่อมความสัมพันธ์ข้อมูล และดำเนินการทดลองวัดผลเพื่อนำค่ามาวิเคราะห์วัดประสิทธิภาพในการจัดการข้อมูล ในการเชื่อมคอลเลคชั่นของฐานข้อมูล MongoDB คุณสมบัติคอลเลคชั่นของ MongoDB ปกติไม่รองรับการเชื่อมคอลเลคชั่น แต่เพื่อเป็นการนำมาประยุกต์ใช้ โดยการนำข้อดีของฐานข้อมูล NoSQLในเรื่องการอ่านและเขียนข้อมูลที่รวดเร็ว มาใช้งานรวมกับแนวคิดของฐานข้อมูลเชิงสัมพันธ์ที่ลดความซ้ำซ้อนของข้อมูลด้วยการแยกตาราง ซึ่งหากเปรียบเทียบกับ MongoDB คือการแยกคอลเลคชั่น และประยุกต์ใช้วิธีการประมวลผลข้อมูลด้วยวิธีแมพรีดิวซ์มาช่วยในการเชื่อมคอลเลคชั่นของฐานข้อมูล MongoDB โดยเพิ่มขั้นตอนการกรองข้อมูลของคอลเลคชั่น แล้วนำมาสร้างเป็นคอลเลคชั่นชั่วคราว เพื่อใช้ในการเชื่อมคอลเลคชั่น มีวิธีการเชื่อมคอลเลคชั่น 5 รูปแบบดังนี้ (1) Join Direct (2) MR Join (3) MR Filter Join 4) MR Left Filter Join 5) MR Right Filter Join และข้อมูลที่ใช้ในการทดสอบนี้ 2 ชุด คือ ชุดที่ 1 Student มี 16 Fields และเอกสาร 9,665 Documents และ ชุดที่ 2 Advisor มี 11 Fields 5และเอกสาร 954 Documents

สรุปผลจากการทดลองได้ว่า การใช้วิธี 1) Join Direct จะใช้ซีพียูและเวลาในการเชื่อมคอลเลคชั่นนานกว่าวิธีอื่น 2) MR Join ใช้วิธีแมพรีดิวซ์ในการเชื่อมอย่างเดียวจะใช้เวลานาน เนื่องจากแมพรีดิวซ์จะอ่านข้อมูลทั้งหมดในคอลเลคชั่น ต้องมีเงื่อนไขเพื่อกรองข้อมูลเพื่อช่วยการ

ประมวลผลให้รวดเร็วขึ้น 3) MR Filter Join เป็นการสร้างคอลเลคชั่นชั่วคราว ก่อนที่จะเชื่อมคอลเลคชั่นซึ่งวิธีนี้จะทำให้การเชื่อมทำได้เร็วขึ้นและน้อยที่สุดในการทดลอง 4) MR Left Filter Join จะเป็นการสร้างคอลเลคชั่นชั่วคราวของพารามิเตอร์ในการเชื่อม ซึ่งได้ความเร็วในการเชื่อมดีกว่าวิธีที่ (1) และ (2) แต่ทั้งนี้การประมวลผลวิธีนี้ก็ขึ้นอยู่กับปริมาณข้อมูลและเงื่อนไขในการกรองข้อมูล ซึ่งผลของเวลาวิธี (3) จะใกล้เคียงกัน 5) MR Left Filter Join จะเป็นการสร้างคอลเลคชั่นชั่วคราวของพารามิเตอร์ในการเชื่อม ผลสรุปได้เหมือนกับวิธีที่ (4) สรุปวิธีที่ (4) และ (5) ผลหมือนกัน

Fegaras, Li, and Gupta (2012) มีงานวิจัยเรื่อง An Optimization Framework for Map-Reduce Queries ปัญหาของงานวิจัยนี้คือ การใช้แบบสอบถามด้วยแมพและรีดิวกับสถาปัตยกรรมของฮาดูปมีการพัฒนาให้สามารถใช้ร่วมกับภาษา SQL อย่างเช่น HiveQL และ Pig Latin เป็นการให้ผู้ใช้ทำการ Plug in เพิ่มเข้ามาในระบบเพื่อทำการใช้งานแบบสคริปต์การสอบถามได้ แต่หากว่าผู้ใช้งานจำเป็นต้องมีความเชี่ยวชาญในการเขียนโปรแกรมแบบ Declaratively ในการสอบถามเพื่อเรียกใช้ข้อมูล และการรักษารหัสการเขียนโปรแกรม (Source Code) ยุ่งยาก ซึ่งอาจจะส่งผลก่อให้เกิดความผิดพลาดของผลลัพธ์ที่ต้องการได้ ผู้วิจัยฉบับนี้จึงนำเสนอการเพิ่มประสิทธิภาพให้กับการเขียนโปรแกรมแบบแมพและรีดิวด้วยการลดคำสั่งการใช้งานแมพและรีดิว ด้วยการใช้แบบสอบถามที่มีพีชคณิตและอกรีบา (Algebra) เข้ามาช่วยเพื่อให้สามารถใช้งานภาษาสอบถามแบบ SQL ได้หรือเรียกว่า MRQL

สมมติฐานของงานวิจัยนี้คือการปรับปรุงเวิร์กโฟลว์ (Work Flow) ให้มีประสิทธิภาพในการทำแบบสอบถามเพื่อค้นคืนข้อมูลสามารถนำรูปแบบของฐานข้อมูลเชิงสัมพันธ์ที่มีรูปแบบทางคณิตศาสตร์พีชคณิตและอกรีบามาใช้งานได้และมีประสิทธิภาพ โดยมีวิธีการวิจัยและรูปแบบการวิจัย ในกระบวนปรับปรุง MR Job หรือการทำงานของแมพรีดิว (MapReduce Job) มีขั้นตอนการทำงานย่อยๆ ดังนี้ 1.ปรับการทำงานโอเปอเรชั่นแมพรีดิว (The MapReduce Operation) 2.ลดฟังก์ชั่นเชื่อมสัมพันธ์ (Reduce-Side Join) 3.การปรับส่วนการทำซ้ำของการเชื่อมสัมพันธ์ (Fragment-Replicate Join) 4.การปรับการดำเนินการทางกายภาพอื่น (Other Physical Operations) และยังมีการปรับแบบสอบถามให้ออกเป็นรูปแบบอกรีบา (Algebra) และทำการปรับปรุงกรอบการทำงานทั้งหมดเพื่อให้ MRQL ลดความซ้ำซ้อนของกระบวนการสอบถามแมพรีดิวทางกายภาพของการเขียนโปรแกรมเพื่อให้ใช้งานร่วมกันกับฐานข้อมูลเชิงสัมพันธ์

โดยกำหนดให้มีขั้นตอนดังนี้ 1.ลดความซ้ำซ้อนของแบบสอบถาม 2.สร้างกราฟแบบสอบถาม 3.รูปแบบที่มีการปรับเปลี่ยนเป็นแบบพีชคณิต 4.แผนผังในการวางแผนการปรับปรุงผลให้เป็นแบบพีชคณิตและทำลองและปรับปรุงประสิทธิภาพให้ดีขึ้น 5.สังเคราะห์ฟังก์ชั่นการรวม (Combine) ในจากกระบวนการลดงาน มีตัวแปรที่สำคัญ รูปแบบหรือโมเดล MRQL และภาษา

ไวยากรณ์ที่ใช้ เครื่องมือวัดวิธีเก็บข้อมูลด้วยเทคนิควิธีการเก็บข้อมูลจากเว็บการป้อนข้อมูลจากข้อมูล Log Running บนเครื่องเครือข่ายกลุ่มเมฆที่มีการเข้าใช้งานของผู้ใช้ใช้ข้อมูลในระบบ และตารางลูกค้าตามมาตรฐาน TPCH ใช้จำนวน 5 ชุดข้อมูล ขนาดเพิ่มขึ้นทีละ 1 เท่าตัวเริ่มตั้งแต่ 4 GB ไปจนถึง 20 GB โดยมีสภาพแวดล้อมจากการใช้คลัสเตอร์เซิร์ฟเวอร์ขนาดเล็กจำนวน 9 เครื่อง ใช้ระบบปฏิบัติการลีนุกซ์ CentOS 5.4 ใช้ระบบเครือข่ายกิกะบิตสวิทซ์ และทำการทดลองกับระบบข้อมูลขนาดใหญ่ Hadoop 0.20.2 การจัดการบนพื้นฐานคำสั่งของ Clouderaโดยการควบคุมการทำงานของ NameNode และ JobTracker และส่วนของเซิร์ฟเวอร์อีกจำนวน 8 โหนดทำหน้าที่ DataNodes และ Trackers ซึ่งเซิร์ฟเวอร์แต่ละเครื่องจะมี CPU 4 Core Xeon 3.2 GHz กับหน่วยความจำ RAM 4 GB และทำการกำหนดค่าให้แตกต่างใน Hadoop ดังนี้ 8 โหนด 32 แกน 6 โหนด 24 แกน และ 4 โหนด 16 แกน ในรูปแบบการจัดเก็บแบบ HDFS และมีวิธีวิเคราะห์ข้อมูลโดยการกำหนดแบบสอบถามขึ้นมาแบบ MRQL ซึ่งเป็นการผสมผสานระหว่าง SQL และ MapReduce โดยการกำหนดให้มี Select , Like, GroupBy เป็นต้น

ผลสรุปจากการวิจัยเวลา การกำหนดกลุ่มทั้ง 3 กลุ่มคลัสเตอร์ 4, 6, 8หลังจากการปรับปรุงประสิทธิภาพด้วยการใช้ MRQL ในการกำหนด MapReduceแล้วได้ประสิทธิภาพ 50% และ 65% แต่การประเมินประสิทธิภาพกับ PageRank ยังไม่มีประสิทธิภาพที่ชัดเจนกับการสังเคราะห์ด้วยกราฟ ดังนั้นผลการสร้างกรอบการทำงานให้เพิ่มประสิทธิภาพของแบบสอบถามด้วย MRQL สามารถนำมาดำเนินการได้เป็นการลดค่าใช้จ่ายในการประมวลผลในส่วนของการลด (Reduce)

Khanam and Agarwal (2015) มีงานวิจัยเรื่อง Map-Reduce Implementations : Survey And Performance Comparison ปัญหาของงานวิจัยนี้คือการใช้งานแมพรีดิวในภาคการวิจัยและในสถาบันการศึกษาและภาคอุตสาหกรรมเพื่อการวิเคราะห์ข้อมูลขนาดใหญ่ มีการใช้งานในหลากหลายรูปแบบแตกต่างกัน เช่น การทำเหมืองวิเคราะห์ข้อมูลขนาดใหญ่ที่มีการใช้พารามิเตอร์แตกต่างกัน การที่จะใช้พารามิเตอร์ให้มีประสิทธิภาพและประสิทธิผลนั้นจึงต้องทำความเข้าใจในทางเทคนิคของกรอบการทำงานแมพรีดิว

ผู้วิจัยฉบับนี้จึงนำเสนอการวิจัยเชิงสำรวจเพื่อทำความเข้าใจทางเทคนิคของกรอบการทำงานแมพรีดิวและคุณสมบัติต่างๆ ของแพลตฟอร์มที่มีการใช้งานคล้ายกันและยังมีการสร้างเปรียบเทียบพารามิเตอร์การใช้งานที่แตกต่างกัน และเปรียบเทียบเทคโนโลยีที่ใช้แพลตฟอร์มต่างกันและคุณสมบัติต่างกันด้วย และภาษาที่ใช้ในการเขียน โดยมีวิธีการวิจัยด้วยการสร้างตารางเปรียบเทียบแสดงเทคโนโลยีมีดังนี้ Hadoop, Spark, Phoenix++, MARISSA, MARIANE, MapReduce-MPI, Disco, SASReduce, BitDew, MARLA, DRYAD, DRYADLINQ, Themis,

MR4C และคุณสมบัติการใช้งานด้วยเครื่องมือโปรแกรมใด และมีฟังก์ชั่นใดที่ใช้งาน ซึ่งรูปแบบการวิจัยโดยการใช้การวิจัยเชิงสำรวจด้วยการศึกษางานวิจัยที่เกี่ยวของเทคโนโลยีที่กล่าวมาแล้วข้างต้น มีเครื่องมือวัดวิธีเก็บข้อมูลด้วยเทคนิควิธีการทำการหาข้อดีและข้อเสียจากเทคโนโลยี มีวิธีวิเคราะห์ข้อมูลจากการใช้งานนำมาเทียบกับแมพและรีดิว

ผลจากการวิจัยเทคโนโลยี Dryad พัฒนาโดยไมโครซอฟท์ เป็นการดำเนินการแบบขนานในรูปแบบอัลกอริทึมแบบกราฟดำเนินการกับระบบไฟล์ที่มีลักษณะคล้ายคลึงกับ Google MapReduce โดยการเรียงลำดับ Map/Distribute/Sort/Reduce เมื่อเปรียบเทียบกับแมพรีดิวแล้วมีความซับซ้อนมากกว่า และเทคโนโลยี DryadLINQ พัฒนาโดยไมโครซอฟท์เช่นกันเป็นรูปแบบการเขียนแบบเชิงวัตถุด้วย Visual.Net ควบคุมการทำงานของ MapReduce และการใช้ SQL ที่ใช้งานจาก LINQ (Language Integrated Query) เมื่อไปเทียบกันแล้วจะใช้เวลามากกว่าแมพรีดิวและเทคโนโลยี Spark เป็นเทคโนโลยีที่นำมาใช้งานหากเทียบกับ Hadoop แล้ว Spark จะเร็วกว่า จากที่ Spark มีการเก็บข้อมูลในหน่วยความจำ และใช้ในการโต้ตอบกับแบบสอบถามชุดข้อมูลขนาดใหญ่และมีประสิทธิภาพดีกว่าในการใช้งานแบบสอบถาม และยังพบว่าหลายที่ต้องการใช้งานแบบสอบถามแบบโต้ตอบ และหลายช่องทางการใช้งาน จึงมีการพัฒนาออกแบบเรียกว่าความยืดหยุ่นแบบกระจายชุดข้อมูล (Datasets) RDD ซึ่งจะอ่านและเขียนเร็วกว่าระบบไฟล์แบบกระจาย และเทคโนโลยี MARISSA ให้ประสิทธิภาพดีกว่าการใช้งาน Hadoop ที่กำหนดเป็นรูปแบบสตรีมมิ่งและสามารถเพิ่มประสิทธิภาพการใช้งานร่วมกับแมพรีดิวได้ และเทคโนโลยี SAS มีลักษณะการทำงานเหมือนกับแมพและรีดิว แต่มีข้อเสียคือจะดำเนินการได้เฉพาะในการใช้งานได้กับเครื่องคอมพิวเตอร์ส่วนบุคคลเท่านั้น และจะใช้ได้ดีเฉพาะการเรียกใช้งานข้อมูลบนตารางด้านบนเท่านั้น และยังมีข้อเสียที่ไม่สามารถจัดการกับความล้มเหลวของ MPI และประสิทธิภาพของแมพรีดิวจะทำงานได้ดีกว่า

สรุปงานวิจัยฉบับนี้จากการศึกษางานวิจัยที่เกี่ยวข้องยังพบว่าบริษัทขนาดใหญ่นำ Hadoop และ MapReduce นำไปใช้งานในหลากหลายลักษณะของงานตั้งแต่ การประมวลผลภาพถ่ายดาวเทียมขนาดใหญ่ และภูมิสารสนเทศข้อมูล วิทยาศาสตร์ และมีประสิทธิภาพของวิธีการพัฒนาอัลกอริทึม มีความยืดหยุ่นและปรับขนาดขยายได้ใน Hadoop วัตถุประสงค์หลัก ออกแบบเพื่อให้สามารถใช้งานข้อมูลแบบ DBMS กับแมพรีดิวให้สามารถใช้งานได้ง่าย และยังสนับสนุนการเขียนภาษาสคริปต์เพื่อใช้งานแบบสอบถาม SQL ซึ่งพัฒนาอยู่ในกรอบของแมพรีดิวซึ่งอย่างไรก็ตามวิธีการทั้งหมดนี้สามารถใช้งานได้ในรูปแบบที่แตกต่างกัน และลักษณะชุดข้อมูลที่แตกต่างกันก็มีผลในการเลือกใช้วิธีการ และยังมีการสนับสนุนฐานข้อมูลใช้งานร่วมกับ MapReduce ได้เช่น

MongoDB, Aster เป็นต้น ซึ่งในพื้นฐานของระบบไฟล์เหล่านี้จะดำเนินการได้รวดเร็วกว่าระบบไฟล์ HDFS แต่จะไม่ดีกว่าในระบบคลัสเตอร์คอมพิวเตอร์จำนวนมาก

Tao, Lin, and Xiao (2013) มีงานวิจัยเรื่อง Minimal MapReduce Algorithms ปัญหาของงานวิจัยนี้คือจากข้อจำกัดของการใช้งานแมพรีดิวที่มีการใช้เวลาในการถ่ายโอนข้อมูลบน CPU และ I/O และเครือข่ายในแต่ละเครื่องและถึงแม้ว่าจะมีการพัฒนาอัลกอริทึมที่เน้นการจัดการเรื่องเหล่านี้โดยเฉพาะแล้ว แต่เป็นการเขียนอัลกอรึทึมที่มีจำนวนมากยังเป็นปัญหาในการจัดทำและการตรวจสอบและอีกทั้งการแก้ไขเกิดข้อผิดพลาดได้ ผู้วิจัยฉบับนี้จึงนำเสนอการเขียนอัลกอริทึมที่มีจำนวนน้อยแต่ยังคงไว้ซึ่งประสิทธิภาพ 0

สมมติฐานของงานวิจัยนี้คือ หน่วยงานที่สะสมข้อมูลไว้เป็นจำนวนมาก หากต้องการจะนำข้อมูลมาประมวลผลข้อมูลมหาศาลในขนาดเทราไบต์หรือสูงกว่ามาใช้งานแบบเร่งด่วน การเขียนโปรแกรมให้มีขั้นตอนที่สั้นจะช่วยให้ลดขั้นตอนจากการจับคู่ (Map), การสุ่ม (Shuffle) และการลด (Reduce) โดยมีวิธีการวิจัยด้วยการศึกษารูปแบบขั้นตอนวิธีการประมวลผลที่มีประสิทธิภาพในแง่มุมที่แตกต่างกันด้วยการกำหนดคุณสมบัติ การส่งออกข้อมูลขั้นต่ำในพื้นที่จัดเก็บที่ในแต่ละโหนด และการจัดการทราฟฟิคในแต่ละรอบของการจัดส่ง และอัลกอริทึมต้องหยุดการทำงานหลังจากทำงานครบรอบ และการหาเวลาการคำนวณที่เหมาะสมในแต่ละรอบ (รอบคงที่)

มีรูปแบบการวิจัยโดยการลดขั้นตอนการสับเปลี่ยนการดูแลภายในเครือข่ายที่เกิดขึ้นจากการถ่ายโอนข้อมูลมีการลดขั้นตอนการสื่อสารแต่ละเครื่องจะคำนวณจากในแต่ละเครื่องใน 1 รอบของการจัดเรียงข้อมูลเพราะหากงานยังไม่เสร็จสิ้นจะทำการคำนวณอีกรอบทำให้ใช้เวลาในการคำนวณใหม่ มีตัวแปรที่สำคัญคือชุดข้อมูลเริ่มต้นและจำนวนเครื่องเครื่องมือวัดวิธีเก็บข้อมูลและเทคนิควิธีการโดยการจัดเรียงข้อมูลให้น้อยที่สุด โดยการสร้างอัลกอริทึม S เป็นชุดข้อมูล และ n ดึงข้อมูลออกมาจากโดเมน และ T เป็นการแจกจ่ายไปยังเครื่องในเครือข่าย

โดยมีสภาพแวดล้อมจากการใช้งานแมพรีดิวในบริบทเครือข่ายกลุ่มเมฆวิธีวิเคราะห์ข้อมูลจากสถิติการเรียงลำดับ (TeraSort), การเลือก (Choice of ρ), การลดการจราจรบนเครือข่าย (Removing the Broadcast Assumption)และผลจากอัลกอรึทึมในฐานข้อมูลโดยการกำหนดขั้นตอนการเรียงลำดับน้อยที่สุดการจัดอันดับกลุ่มโดย 1.กึ่งเข้าร่วมข้อมูล(Semi Join) 2.การจัดอันดับและเส้นขอบ (Ranking and Skyline) 3.การจัดกลุ่ม (Group By) และผลของการใช้การรวมแถบเลื่อน (Sliding Aggregation) 1.การจัดเรียงที่สมบูรณ์ (Sorting with Perfect Balance) 2.การเลื่อนรวม (Sliding Aggregate Computation) ทั้งหมดมีเป้าหมายการทำงานรอบเดียว มีขั้นตอนการทดลองโดยการใช้คลัสเตอร์คอมพิวเตอร์จำนวนเครื่องแม่ 1 Master และจำนวนเครื่องลูก 56 Slave มีเครื่อง

คอมพิวเตอร์ระดับเซิร์ฟเวอร์มีคุณสมบัติ CPU Xeon 2.4 GHz และ Ram 24 GB. และทำการติดตั้งโปรแกรมโอเพ่นซอร์สฮาดูปเวอร์ชั่น 1.0 และใช้ Java Virtual Machine แต่ละโหนดมี RAM 4 GB. โดยกำหนดพารามิเตอร์ fs.block.size มีขนาด 128 MB และ io.sort.mb มีขนาด 512 MB และ io.sort.record.percentage เท่ากับ 0.1 และ io.sort.spill.percentage เท่ากับ 0.9 และ io.sort.factor เท่ากับ 300 และ dfs.replication เท่ากับ 3 และใช้ข้อมูลในการทดลองจริงที่มีชื่อว่า LIDAR6 และ PageView7 ขนาด 514 GB และมีขนาดระเบียนของ LIDAR มี 7,350,000,000 และกลุ่ม PageView เป็น 332 GB และมีขนาดระเบียน 11,800,000,000 Tuples

ผลจากการวิจัยในกลุ่มที่ 1.การเรียงลำดับ ข้อมูล (Sort) Terasort มีประสิทธิภาพดีกว่า HS มีความแตกต่างอย่างมีนัยสำคัญ ในกลุ่มที่ 2 Skylineประสิทธิภาพของ HS ไม่มีค่าใช้จ่ายในการกระจายของระเบียนสูง 2.เส้นขอบ (Skyline) การใช้ MR-SFS ที่พัฒนาจากการคำนวณใน MapReduce ซึ่งเมื่อเปรียบเทียบการใช้ข้อมูล LIDAR ที่มีขนาดข้อมูลเพิ่มขึ้น Minimal-Sky มีประสิทธิภาพดีกว่า MR-SFS และมีค่าใช้จ่ายน้อยกว่าเพราะ MR-SFS มีขั้นตอนที่ติดต่อกันโหนดมากกว่าทำการประมวลผลมีเวลาที่มากกว่า 3.การจัดกลุ่ม (Group By) ผลของ Minimal-GB จะใช้งานอย่างมีประสิทธิภาพดีกว่า Base-GB 4.กึ่งเข้าร่วมข้อมูล (Semi Join)ผลของ Minimal-SJ ดีกว่าเพราะการกระจายงานไปยังโหนดต่างๆ ใช้เวลารวมน้อยกว่า จึงมีประสิทธิภาพมากว่า 5.การเลื่อนรวม (Sliding Aggregate Computation) ผลของ Minimal-SA มีประสิทธิภาพดีกว่าทุกชุดข้อมูล

ได้บทสรุปจากการทดลองดังนี้ MapReduce ได้เติบโตขึ้นเป็นสถาปัตยกรรมที่นิยมอย่างมากกับการคำนวณแบบขนานกับข้อมูลขนาดใหญ่ ถึงแม้ว่าจะมีวิธีการที่มีความหลากหลายของการพัฒนาสำหรับ MapReduce มีไม่กี่วิธีที่สามารถจะบรรลุเป้าหมายที่เหมาะสมในการทำประมวลผลแบบคู่ขนานและการเกิดภาระงานที่สมดุลอีกทั้งการทำงานข้ามเครื่องที่ร่วมอยู่ภายในเครือข่าย และเพิ่มความเร็วให้มากขึ้นกับการลำดับขั้นตอนวิธีการเชิงเส้นของจำนวนเครื่องผลงานที่สำคัญของการวิจัยนี้คือขั้นตอนวิธีการ 4 แบบจากการทดลองทำงานได้อย่างมีประสิทธิภาพรวดเร็วมากขึ้น ด้วยเงื่อนไขของ Minimality และสามารถนำไปใช้ประโยชน์ได้

ประกายมาศ ศรีสุขทักษิณ และ ผุสดี บุญรอด (2557) มีงานวิจัยเรื่อง การเปรียบเทียบความเร็วในการประมวลผลระหว่างฐานข้อมูลเชิงสัมพันธ์ และฐานข้อมูลไม่สัมพันธ์แบบเอกสาร ปัญหาของงานวิจัยนี้คือข้อมูลข่าวสารที่เพิ่มขึ้นเกินขีดความสามารถของฐานข้อมูลเชิงสัมพันธ์การบริหารจัดการข้อมูลที่ดีจึงเป็นสิ่งสำคัญ การมีตัวเลือกของเทคโนโลยีฐานข้อมูลไม่สัมพันธ์เกิดขึ้นซึ่งเทคโนโลยีนี้จะนำมาใช้งานได้จริงหรือไม่

ผู้วิจัยฉบับนี้จึงนำเสนอ การทดสอบความเร็วในการประมวลผลของฐานข้อมูลไม่สัมพันธ์แบบเอกสารที่มีการประมวลผลเครื่องเดียว และการประมวลผลแบบกระจายเปรียบเทียบ

กับฐานข้อมูลเชิงสัมพันธ์ ในด้านการเขียน, อ่าน, แก้ไข และลบข้อมูล สมมติฐานของงานวิจัยนี้คือ ฐานข้อมูลไม่สัมพันธ์เหมาะสมกับการใช้งานเนื่องจากมีต้นทุนต่ำในการพัฒนาและมีประสิทธิภาพดีกว่า โดยมีวิธีการวิจัยและมีรูปแบบการวิจัยการนำฐานข้อมูลเชิงสัมพันธ์ MySQL ที่มีการเพิ่มประสิทธิภาพด้วยการใช้งานอินเด็กซ์ (Index) ข้อมูลไว้ก่อนล่วงหน้า และฐานข้อมูลไม่สัมพันธ์แบบเอกสาร MongoDB แบบเครื่องเดียวและแบบกระจายนำมาประมวลผลเปรียบเทียบกันในคุณลักษณะการใช้งานอ่าน (Read), เขียน (Insert), แก้ไข (Update) และลบ (Delete) กับข้อมูลที่มีการจัดเตรียมไว้ที่จำนวนระเบียนตั้งแต่ 500, 5,000, 50,000, 500,000 ซึ่งให้เธรดในการติดต่อฐานข้อมูล 1, 10, 100 เธรด ตามลำดับ มีตัวแปรที่สำคัญ จำนวนระเบียนข้อมูลและจำนวนเธรดข้อมูล และการทดสอบจะทำการประมวลผลจำนวน 3 ครั้งและทำการหาค่าเฉลี่ย เครื่องมือวัดวิธีเก็บข้อมูลด้วยเทคนิควิธีการคำนวณเวลาที่ใช้ความเร็วในการประมวลผลของฐานข้อมูลแต่ละแบบ โดยมีสภาพแวดล้อมจากเครื่องที่ใช้ทำการทดสอบมีหน่วยประมวลผลกลาง (CPU) แบบ Intel Core i5 2.27 GHz มีหน่วยความจำ RAM 8 GB และ HD มีความจุ 5800GB ที่ใช้เป็นทั้งเครื่อง Server จำนวน 1 เครื่อง และ Client จำนวน 3 เครื่อง และทำการทดสอบกับฐานข้อมูลเชิงสัมพันธ์ MySQL เวอร์ชั่น 5.0.51b และ ฐานข้อมูลไม่สัมพันธ์ใช้ MongoDB เวอร์ชั่น 2.4.5 วิธีวิเคราะห์ข้อมูลโดยการสร้างกราฟแท่งผลการทดสอบในแต่ละแบบมีการเปรียบเทียบผลลัพธ์ด้านความเร็ว

ผลจากการวิจัยในการอ่านเมื่อจำนวนเธรดมากขึ้น MongoDB แบบประมวลผลเครื่องเดียวและแบบกระจายทำงานได้ใกล้เคียงกัน ทำงานดีกว่า MySQL ประมาณ 3 เท่า และในการเขียนข้อมูลมีผลเช่นเดียวกันกับการอ่าน และในการแก้ไขข้อมูลเมื่อจำนวนเธรดมากขึ้น MongoDB แบบประมวลผลเครื่องเดียวและแบบกระจายได้เวลาใกล้เคียงกัน แต่การทำงานในจำนวนข้อมูล 500,000 ระเบียน จะทำงานได้ดีกว่า MySQL ถึง 40 เท่า และในการลบข้อมูลเมื่อจำนวนเธรดเพิ่มขึ้น MongoDB จะใช้เวลาประมวลผลการลบข้อมูลมากขึ้นทั้งในแบบเครื่องเดียวและแบบกระจาย แต่การทำงานในจำนวนข้อมูล 500,000 ระเบียนใช้ 100 เธรด 100 จะทำงานได้ดีกว่า MySQL ถึง 70 เท่า ผลสรุปหากข้อมูลจำนวนเธรดน้อยฐานข้อมูลจะมีความสามารถในการประมวลผลทั้งหมดได้ใกล้เคียงกัน และเมื่อจำนวนเธรดมากขึ้นและจำนวนระเบียนมากขึ้นจะเริ่มเห็นความแตกต่างการประมวลผลที่ชัดเจนมากยิ่งขึ้น ดังนั้นฐานข้อมูลไม่สัมพันธ์จึงมีประโยชน์ มีประสิทธิภาพและมีค่าใช้จ่ายที่ต่ำกว่า

ชูพันธ์ รัตนโภคา (2555) มีงานวิจัยเรื่อง การออกแบบและพัฒนาระบบค้นหาข้อมูลจราจรทางคอมพิวเตอร์ (Log) ด้วยวิธี Map/Reduce บนกรอบการทำงานของ Hadoop ปัญหาของงานวิจัยนี้คือผู้ให้บริการเครือข่ายคอมพิวเตอร์จำเป็นต้องเก็บรักษาข้อมูลการจราจรคอมพิวเตอร์ไม่น้อยกว่า 90 วัน ทำให้ผู้ให้บริการเครือข่ายคอมพิวเตอร์ต้องเก็บข้อมูลเป็นจำนวนมาก หากมีการ

สืบค้นคืนข้อมูลเพื่อระบุตัวผู้ใช้บริการต้องใช้เวลานานในการค้นหา ผู้วิจัยฉบับนี้จึงนำเสนอการออกแบบและพัฒนาระบบค้นหาข้อมูลจราจรคอมพิวเตอร์โดยนำ Hadoop มาประยุกต์ใช้ในการเก็บข้อมูลจราจรคอมพิวเตอร์ และใช้วิธีการค้นหาข้อมูลด้วย Map/Reduce เพื่อให้มีประสิทธิภาพในการค้นหาข้อมูลได้รวดเร็ว

สมมติฐานของงานวิจัยนี้คือ เทคโนโลยี Hadoop นำมาประยุกต์ใช้ในการจัดเก็บข้อมูลขนาดใหญ่ โดยมีวิธีการวิจัยเชิงทดลองทำการแบ่งข้อมูลขนาดใหญ่ออกเป็นส่วนย่อยๆ แล้วกระจายไปยังเครื่องคอมพิวเตอร์ต่างๆ ที่เชื่อมต่อกันแล้วใช้ Map/Reduce ส่งคำสั่งค้นหาข้อมูลกระจายไปยังเครื่องคอมพิวเตอร์ทุกเครื่องโดยไม่จำเป็นต้องมีการย้ายข้อมูลระหว่างการประมวลผล มีรูปแบบการวิจัยที่นำข้อมูลจราจรทางคอมพิวเตอร์เป็นแฟ้มข้อมูลตัวอย่างและมีการเก็บรวบรวมจากคอมพิวเตอร์แม่ข่ายที่เป็นส่วนเชื่อมต่ออินเตอร์เน็ตภายนอกกับเครือข่ายภายใน (NAT) ของวิทยาลัยเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ และใช้การพัฒนาโปรแกรมส่วนติดต่อกับผู้ใช้หรือ GUI ด้วยภาษาจาวาเชื่อมต่อกับไฟล์ HDFS บนระบบปฏิบัติการลีนุกซ์ และค้นหาข้อมูลจากไฟล์ HDFS ด้วยการเขียนโปรแกรมภาษาจาวาค้นหาข้อมูลด้วยเทคนิควิธี Map/Reduce ขนาดของตัวอย่างในการวิจัยมีขนาด 5, 10, 20, 30, 40, 50 GB ตามลำดับ

มีตัวแปรที่สำคัญ ตัวแปรต้นคือข้อมูลจราจรทางคอมพิวเตอร์และตัวแปรตามคือเวลาในการประมวลผล เครื่องมือวัดวิธีเก็บข้อมูลด้วยเทคนิควิธีการบันทึกข้อมูลจากการวัดหรือนับโดยใช้อุปกรณ์คอมพิวเตอร์ในการตรวจนับ โดยมีสภาพแวดล้อมจากการใช้จำนวนเครื่องคอมพิวเตอร์ที่ทำการทดสอบจำนวน 11 เครื่อง เป็นเครื่อง Name node (Master) จำนวน 1 เครื่อง และ Data node (Slave) จำนวน 10 เครื่อง เป็นเครื่องที่มีคุณลักษณะเหมือนกันทั้ง CPU, RAM, Hard disk วิธีวิเคราะห์ข้อมูลด้วยการนำเข้าข้อมูล ด้วยการประยุกต์ใช้ HDFS ทำหน้าที่ในการเก็บข้อมูลแบบกระจาย สามารถที่จะรองรับการขยายตัวของจำนวนเครื่องเก็บข้อมูลได้ และทำการบันทึกผลของเวลาในการนำเข้าตามจำนวนขนาดข้อมูลตัวอย่างและจำนวนคอมพิวเตอร์ในการจัดเก็บ และทำการเก็บผลการวิจัยการค้นหาข้อมูลจราจรทางคอมพิวเตอร์ด้วยวิธี Map/Reduce ที่ผ่านการออกแบบพัฒนาโปรแกรมที่ช่วยค้นหาข้อมูลขนาดใหญ่

ซึ่งผลการทดสอบความเร็วในการค้นหาข้อมูลจำนวนเครื่อง 10 เครื่อง และค้นหาข้อมูลจราจรคอมพิวเตอร์ที่มีขนาดสูงสุด 50 กิกะไบต์ ผลจากการวิจัยปรากฏว่าเวลาที่ใช้ในการค้นหาข้อมูลจะมีความเร็วเพิ่มขึ้นประมาณ 10 เท่า เมื่อเทียบกับการค้นหาที่ใช้เครื่องคอมพิวเตอร์เพียงเครื่องเดียว

Dean and Ghemawat (2008) มีงานวิจัยเรื่อง MapReduce: simplified data processing on large clusters ปัญหาของงานวิจัยนี้คือการประมวลผลเพื่อคำนวณข้อมูลการใช้งานเว็บไซต์ของ Google เพื่อรวบรวมทำดัชนีการเข้าถึง จำนวนหน้าที่เข้าถึงและอื่นๆ จะต้องทำการคำนวณในเครื่องเซิร์ฟเวอร์หลายร้อยเครื่อง และต้องทำการกระจายการคำนวณไปยังเครื่องต่างๆ และยังพบว่าข้อมูลมีขนาดใหญ่ จนทำให้เกิดความล้มเหลวในการคำนวณ เกิดเป็นปัญหาในการจัดการกับข้อมูลที่มีขนาดใหญ่เหล่านี้ ผู้วิจัยฉบับนี้จึงนำเสนองานวิจัยที่มีการออกแบบการคำนวณแบบง่ายๆ กับข้อมูลขนาดใหญ่ ด้วยการคำนวณแบบขนาน และยังทนทานต่อความผิดพลาด และเป็นการกระจายภาระงานในการประมวลผลข้อมูล โดยได้รับแรงบันดาลใจจากการจับคู่ (Map) และการลด (Reduce) และยังทำการออกแบบการบันทึกข้อมูลในแบบที่ง่ายต่อการคำนวณแบบชุดคู่คีย์ (Key-value) ซึ่งหลังจากมีการปรับเปลี่ยนรูปแบบแล้วทำให้สามารถคำนวณแบบคู่ขนานได้อย่างง่ายดาย

สมมติฐานของงานวิจัยนี้คือการออกแบบโปรแกรมแบบคู่คีย์หรือรูปแบบโปรแกรมที่เรียกว่า Map และ Reduce สามารถดำเนินการกับข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพและง่ายดายและใช้งานได้กับข้อมูลหลากหลายรูปแบบ โดยมีวิธีการวิจัยและรูปแบบการวิจัยด้วยวิธีการวิจัยเชิงทดลองด้วยการศึกษาตัวอย่างการเขียนโปรแกรมตามกำหนดหัวข้อดังนี้ 1.การกระจายการค้นหา (Distributed Grep) 2.นับจำนวนการเข้าถึงเว็บไซต์ (Count of URL Access Frequency) 3. การย้อนกลับเข้าดูเว็บไซต์จากจุดเชื่อมโยงของเว็บไซต์ (Reverse Web-Link Graph) 4.การสรุปคำสำคัญหรือคำค้นหา (Term-Vector per Host) 5.การตัดคำเพื่อสร้างดัชนีย้อนกลับ (Inverted Index) 6.การกระจายการจัดเรียง (Distributed Sort) และยังมีการปรับแต่งการเขียนโปรแกรมหลายรูปแบบเพื่อให้เกิดประโยชน์ที่ใช้งานได้จริงกับการทดลอง มีตัวแปรที่สำคัญคือโปรแกรมแมพรีดิวและรูปแบบการเขียนโปรแกรมขั้นพื้นฐาน เครื่องมือวัดคือโปรแกรมที่มีอัลกอริทึมสำหรับการใช้งานตามหัวข้อที่กำหนด 7 หัวข้อ ที่กล่าวไว้แล้ว

วิธีเก็บข้อมูลด้วยเทคนิควิธีการการใช้อินเตอร์เฟซที่แตกต่างกันออกไปตามความเหมาะสมของสภาพแวดล้อม เช่น เหมาะกับเครื่องที่มีหน่วยความจำขนาดเล็กแต่มีหลายหน่วยประมวลผล (Multi Processor) หรือเหมาะกับกลุ่มคอมพิวเตอร์ขนาดใหญ่ที่มีในเครือข่าย โดยมีสภาพแวดล้อมจากกลุ่มคอมพิวเตอร์ขนาดใหญ่ที่ใช้งานในการทำงานการให้บริการ Search Engine ของ Google ซึ่งเป็นคอมพิวเตอร์ที่มีเครือข่ายขนาดใหญ่จำนวนหลายพันเครื่อง ซึ่งเครื่องโดยทั่วไปจะมีหน่วยประมวลผล (CPU) x86 ใช้ระบบปฏิบัติการ Linux มีหน่วยความจำสำรอง 2-4 GB. และใช้ระบบเครือข่ายด้วยการ์ดแลนขนาดความเร็ว 100/1000 Mbps. ในแต่ละเครื่อง ซึ่งจำนวนเครื่องมีตั้งแต่ 100-1,000 เครื่อง และมีการจัดเก็บข้อมูลในฮาร์ดดิสก์แบบ IDE และมีระบบไฟล์แบบ

กระจายที่ถูกพัฒนาขึ้นใน Google หรือเรียกว่า GFS (Google File System) และสภาพแวดล้อมสุดท้ายเป็นการกำหนดการตั้งเวลาในการส่งชุดข้อมูลให้กับระบบการประมวลผล

มีวิธีวิเคราะห์ข้อมูลด้วยการใช้ข้อมูลการใช้งานของผู้ใช้เว็บไซต์บนเซิร์ฟเวอร์ HTTP มีขนาดประมาณ 1 เทราไบต์ในครั้งแรกก่อนจะขึ้นใช้งานทดลองกับข้อมูลจริง แล้วทำการวิเคราะห์ด้วยโปรแกรม R วิเคราะห์ด้วยการแยกชุดข้อมูลออกเป็นบล็อกย่อยๆ ขนาด 64 เมกะไบต์ ควบคุมด้วยเครื่องของผู้ใช้ และเริ่มการประมวลผลด้วยการสร้างชุดข้อมูลหลายชุดข้อมูลในแต่ละกลุ่มเครื่อง และใช้การแยกการวิเคราะห์ข้อมูลแต่ละชุดข้อมูลในแต่ละเครื่องด้วยการทำคีย์คู่จากการทำแมพและรีดิว ซึ่งแต่ละชุดข้อมูลจะทำการประมวลผลจากบัฟเฟอร์ เมื่อทำการประมวลผลแล้วจะทำการส่งผลกลับไปยังในแต่ละเครื่องของผู้ใช้แต่หากเป็นคอมพิวเตอร์เครือข่ายขนาดใหญ่จะมีการทำการแบ่งกลุ่ม (Partition) ของแต่ละที่ควบคุมการทำงานด้วยการทำงานระยะไกลและลำดับงานต่อไปจะทำการจัดเรียงใหม่ในแต่ละเครื่องด้วยการทำแมพและรีดิวเช่นกัน ซึ่งการทำดำเนินการทั้งหมดจะถูกส่งเป็นไฟล์ที่มีการดำเนินการแล้วกลับไปหาโปรแกรม R ของผู้ใช้

ผลการวิจัยการป้อนข้อมูลเข้ามีอัตราสูงสุดมากกว่า 30 Gbps. 1,764 Worker ในการแมพ และเริ่มลดจนเป็น 0 วินาที ซึ่งในชุด GFS มีการนำเข้า 1,000 ไฟล์ ใช้เวลาทั้งหมดประมาณ 150 วินาที และมีผลการจัดเรียงข้อมูลขนาด 1 เทราไบต์ จากการแบ่งพาร์ติชั่นการส่งออกเป็น 4,000 ไฟล์ ซึ่งใช้เวลาสูงสุดที่อัตรา 13 Gbps. ใช้เวลาทั้งสิ้น 200 วินาที และทั้งหมดในการทำงานที่ใช้เวลาตั้งแต่เริ่มต้นการคำนวณจะใช้เวลา 891 วินาที เป็นการอ่านข้อมูลจากฮาร์ดดิสก์จากเซิร์ฟเวอร์ที่แต่ละภูมิภาคและทำข้ามเครือข่ายขนาดใหญ่ระหว่างประเทศ และในการทดลองการใช้งานจริงใช้กับ 1,000 เครื่องดำเนินการเสร็จทั้งหมดภายในครึ่งชั่วโมง และในการจัดทำดัชนีข้อมูลขนาดใหญ่กับข้อมูลมากกว่า 20 เทราไบต์ใช้เวลาไม่กี่วัน

สรุปผลจากการวิจัยรูปแบบการเขียนโปรแกรมแมพรีดิวประสบความสำเร็จและได้รับการใช้งานในกูเกิ้ล (Google) มีวัตถุประสงค์เพื่อการใช้งานในแต่ละงานที่แตกต่างกัน แต่ไม่ใช่ว่าการเริ่มต้นแล้วจะสำเร็จทั้งหมดยังมีข้อผิดพลาดในช่วงเริ่มต้นมีการปรับปรุงการเขียนโปรแกรมและฮาร์ดแวร์ด้วย เช่น การทำงานการคำนวณที่ช้าเนื่องจากข้อผิดพลาดจากฮาร์ดดิสก์เสีย หรือการเขียนโค้ดที่ผิดพลาดที่ก่อให้เกิดหน่วยความจำแคชหยุดการทำงานเป็นต้น ซึ่งหากเป็นปัญหาที่ฮาร์ดแวร์หรือหน่วยความจำไม่เพียงพอทำการแก้ไขด้วยการเพิ่มทรัพยากรซึ่งเป็นจำนวนน้อยแต่เป็นการแก้ปัญหาที่ทำให้การประมวลผลด้วยแมพรีดิวทำงานได้ดีขึ้นอย่างมีนัยสำคัญ ซึ่งความสำเร็จนี้มีเหตุผลหลายประการรุ่นแรกนี้ใช้งานง่ายแม้จะเป็นโปรแกรมเมอร์ที่ไม่มีประสบการณ์กับระบบแบบขนานและระบบแบบกระจายเพราะมันซ่อนรายละเอียดของการทนต่อความผิดพลาดบนเครือข่าย และยังมีปัญหาความหลากหลายของงาน และความหลากหลายความ

ต้องการข้อมูลที่แก้ไขปัญหาเหล่านี้ได้ด้วยแมพรีดิว และอีกหนึ่งข้อคือได้มีการพัฒนาการดำเนินงานของแมพรีดิวการเพิ่มเครื่องในคลัสเตอร์ขนาดใหญ่ของเครื่องประกอบไปด้วยหลายพันเครื่อง การดำเนินงานทำให้การใช้ทรัพยากรอย่างมีประสิทธิภาพเครื่องเหล่านี้ระบบมีการกำหนดเป้าหมายในการลดปริมาณของข้อมูลที่ส่งผ่านเครือข่ายสามารถใช้ในการลดผลกระทบของเครื่องช้าและจะจัดการกับความล้มเหลวของเครื่องและการสูญเสียข้อมูลได้อย่างมีประสิทธิภาพ

Hollingsworth (2012) มีงานวิจัยเรื่อง Hadoop and Hive as Scalable Alternatives to RDBMS : A Case Study ปัญหาของงานวิจัยนี้คือ การใช้งานโซลูชั่นการจัดการข้อมูลอย่างเช่น Hadoop วัตถุประสงค์เพื่อใช้งานสำหรับการวิเคราะห์ข้อมูลขนาดใหญ่ แต่ถ้าเป็นธุรกิจขนาดกลางและขนาดเล็กที่ยังมีความต้องการการใช้งานระบบการจัดการข้อมูลขนาดใหญ่ต้นทุนต่ำ ซึ่งข้อมูลขององค์กรเหล่านี้นับวันก็จะมีการสะสมจำนวนมาก

ผู้วิจัยฉบับนี้จึงนำเสนองานวิจัยที่มีวัตถุประสงค์ในการเปรียบเทียบการขยายตัวของระบบการจัดการฐานข้อมูลเชิงสัมพันธ์และการจัดการข้อมูลแบบกระจายสำหรับข้อมูลขนาดเล็กและขนาดกลางโดยใช้เครื่องมือในการวิจัยนี้คือโปรแกรมฐานข้อมูล MySQL และโปรแกรม MapReduce และโปรแกรม Hive ด้วยการใช้ข้อมูลประวัติการชำระเงินของบัญชีลูกค้า สมมุติฐานของงานวิจัยนี้คือ ค่าใช้จ่ายจะสูงขึ้นและเพิ่มมากขึ้นสำหรับการใช้งานในฐานข้อมูลเชิงสัมพันธ์ที่ใช้สำหรับการจัดการข้อมูล หากเปรียบเทียบกับการจัดการข้อมูลแบบกระจายที่มีประสิทธิภาพดีกว่าและค่าใช้จ่ายต่ำกว่าและพิสูจน์ด้วยการตั้งคำถามดังนี้ 1.ขนาดของข้อมูลของจำนวนบัญชีหรือของลูกค้า จะมีวิธีการแก้ปัญหาที่ดีกว่ากัน 2.โครงสร้างข้อมูลจะทำงานร่วมกับแต่ละวิธีได้ดีขึ้นหรือไม่ 3.ในการแก้ไขปัญหาข้อมูลขนาดใหญ่เหล่านี้จะมีค่าใช้จ่ายในการปรับเปลี่ยนอย่างไร โดยมีวิธีการวิจัยด้วยการทดลองระหว่างระบบฐานข้อมูลเชิงสัมพันธ์ (RDBMS) โดยการตรวจสอบการทำงานของซอฟต์แวร์โอเพ่นซอร์ส MySQL เป็นตัวแทนของระบบการจัดการข้อมูลเชิงสัมพันธ์และการจัดการข้อมูลแบบกระจาย (DDMSs) โดยการตรวจสอบการทำงานของซอฟต์แวร์โอเพ่นซอร์ส Hadoop เป็นตัวแทนของระบบการจัดการข้อมูลแบบกระจาย และทำการวิเคราะห์ข้อมูลในชุดข้อมูลขนาดกลางที่มีการใช้งานจริงของสถานประกอบการธุรกิจขนาดกลาง ด้วยโปรแกรมโอเพ่นซอร์ส MySQL ในการจัดการข้อมูลเชิงสัมพันธ์และ MapReduce และ Hive ในการจัดการข้อมูลแบบกระจายและใช้ฮาร์ดแวร์และซอร์ฟแวร์ดังนี้ในระบบ MySQL รุ่น 5.1 ในลีนุกซ์ RedHat-GNU มี Maximum Buffer ที่ 16GB และ Maximum Package 16MB และทำการลบหน่วยความจำหลังการทดลองทุกครั้ง ติดตั้งเครื่องที่มีหน่วยประมวลผล 4 CPU และ Hadoop เวอร์ชั่น 0.20.2 ทำงานร่วมกับ Java เวอร์ชั่น 1.6.0.21 ด้วยการทำงาน 4 โหนด และลักษณะฮาร์ดแวร์เป็นเช่นเดียวกับ MySQL และการติดตั้ง Hive รุ่น 0.7.0 บนการทำงานของ HDFS

มีรูปแบบการวิจัยดังนี้การนำข้อมูลจากฐานข้อมูลธุรกิจจากระบบธุรกิจอัจฉริยะ (Business Intelligence) มีตัวแปรที่สำคัญโปรแกรมสำหรับการจัดการข้อมูล 3 โปรแกรมได้แก่ MySQL, MapReduce, Hive และจำนวนข้อมูลการทำธุรกรรมบัญชีการเงินและประวัติการชำระเงินของลูกค้า ที่มีการทดลองกับข้อมูลตั้งแต่ 200MB, 500MB, 1GB, 5GB และ 10GB ตามลำดับ และใช้ข้อมูลในการวิเคราะห์จาก 500 ระเบียน ถึง 20,000 ระเบียน ตัวแปรต้นคือ จำนวนระเบียนของบัญชีลูกค้าโดยการสุ่มข้อมูลบัญชีลูกค้า และขนาดข้อมูลที่ทำการทดลอง

เครื่องมือวัดวิธีเก็บข้อมูลด้วยเทคนิควิธีการ ทำกระบวนการสคีมาการชำระเงินจากฐานข้อมูล MySQL ดึงเข้าที่ HDFS และ MySQL และสคีมาแบบสอบถามการวิเคราะห์ทั้งแบบ MySQL และ MapReduce และใช้ Hive ดำเนินการโหลดประวัติการชำระเงินไปฐานข้อมูล HDFS สำหรับการวิเคราะห์โดย Hive โดยใช้ข้อมูลการทดลองแต่ละขนาดที่กำหนดไว้แล้วข้างต้นในการทดลองโดยการทดลองจำนวน 3 ครั้งต่อชุดข้อมูลโดยมีสภาพแวดล้อมจาก คอมพิวเตอร์คลัสเตอร์ Master จำนวน 1 เครื่อง และ Slave 32 เครื่อง ในมหาวิทยาลัย Boise State และใช้เครือข่าย Gigabit Ethernet เครื่อง 64 Bit Intel Core 2 Duo 3.0GHz 2GB Ram และ Harddisk 160 GB และ Master เป็น Intel Xeon 2.4 GHz Hyper-threading 8 Processing threads 4 Core with 2 threads per core Harddisk SCSI with RAID-6 วิธีวิเคราะห์ข้อมูลการวิเคราะห์เชิงพยากรณ์ในการวิเคราะห์ประวัติชำระทางการเงิน ผลจากการวิจัยในการทดลองกับชุดข้อมูลตั้งแต่ 200 MB จนถึง 10 GB

ผลในการศึกษาพบว่าการใช้งานเซิร์ฟเวอร์เดียว MySQL ทำงานได้ดีที่สุด สำหรับขนาดการทดลองตั้งแต่ 200 MB จนถึง 1 GB การใช้งาน MySQL มีประสิทธิภาพที่ด้อยกว่า MapReduce ที่มีการใช้งานบนชุดข้อมูลขนาดใหญ่เกินกว่า 1 GB ขึ้นไปและ MapReduce ยังมีประสิทธิภาพดีกว่า Hive และ MySQL บนชุดข้อมูลขนาดใหญ่เกินกว่า 2 GB จึงสามารถสรุปการแก้ปัญหาทั้งหมดในงานวิจัยนี้คือ MapReduce มีประสิทธิภาพมากและดีที่สุดในทุกชุดข้อมูลขนาดเล็กตั้งแต่ 200MB ถึง 10GB

2.3 สรุปงานวิจัยที่เกี่ยวข้อง

สรุปจากการศึกษางานวิจัยที่เกี่ยวข้องกับเทคโนโลยีข้อมูลขนาดใหญ่การจัดจำแนกประเภทเทคโนโลยีของบิ๊กดาต้ายังไม่มีมาตรฐานหรือองค์กรใดจัดตั้งขึ้นมารองรับหรือจัดทำเป็นมาตรฐานสากล แต่จะใช้ลักษณะการเติบโตของข้อมูลนำมาจำแนกประเภทดังนี้ ข้อมูลแบบเชิงสัมพันธ์ข้อมูล (RDBMS) มีข้อมูลเพิ่มขึ้นจะขยายเป็นลักษณะแนวตั้ง (Vertical Scaling) แต่ลักษณะข้อมูลขนาดใหญ่ข้อมูลเพิ่มขึ้นจะขยายเป็นลักษณะแนวนอน (Horizontal Scaling)

2.3.1 การจำแนกเทคโนโลยีข้อมูลขนาดใหญ่ ผู้วิจัยจึงขอจำแนกเทคโนโลยีข้อมูลขนาดใหญ่จากการศึกษางานวิจัยที่เกี่ยวข้องออกได้เป็น 2 กลุ่มดังนี้

2.3.1.1 การจัดเก็บข้อมูล (Storage) การจัดเก็บนี้เป็นลักษณะ NoSQL คือการเขียนโปรแกรมมีอัลกอริทึมเพื่อควบคุมการทำงานการอ่าน, เขียน, ลบและแก้ไขแทนการจัดการข้อมูลรูปแบบเดิมที่มีการใช้ภาษาสอบถามเชิงโครงสร้าง SQL ในการจัดการ แบ่งออกเป็น 4 ประเภท ดังนี้

ก) แบบคอลัมน์ (Columns Oriented)

ข) แบบคีย์คู่ (Key-Value Store)

ค) แบบเอกสาร (Document Oriented)

ง) แบบกราฟ (Graph Database)

2.3.1.2 การประมวลผล (Processing) เทคโนโลยีการประมวลผลข้อมูลในเทคโนโลยีบิ๊กดาต้ามีหลากหลายรูปแบบเช่นกัน ทั้งนี้การประมวลผลทั้งหมดมีวัตถุประสงค์หลักเพื่อการใช้งานประมวลผลกับชุดของข้อมูลขนาดเทราไบต์และเพตาไบต์โดยจำแนกออกเป็น 6 ประเภทดังนี้

ก) SQL ยังคงใช้รูปแบบภาษาสอบถามเชิงโครงสร้างในการจัดการข้อมูลแบบเชิงสัมพันธ์ ใช้งานร่วมกับการจัดเก็บแบบ HDFS เช่น Hive, Impala หรือ Tajo

ข) Key-Value มีลักษณะของการประมวลผลด้วยโปรแกรมมีการเขียนอัลกอริทึมภาษาต่างๆ คอยควบคุมสั่งการ เช่น Java ใช้งานร่วมกับแมพรีดิว

ค) NoSQL เป็นลักษณะของการประมวลผลด้วยการเขียนอัลกอริทึมด้วยโปรแกรมภาษาต่างๆ คอยควบคุมสั่งการ เช่น JSON ใช้งานร่วมกับ MongoDBหรือด้วยภาษา Scala, Python ใช้งานร่วมกับ Spark

ง) NewSQL เป็นลักษณะของการประมวลผลด้วย SQL แต่นำ NoSQLมาเป็นฐานข้อมูลเพื่อการขยายข้อมูลในลักษณะแนวนอน และในการประมวลผลยังมีการนำหน่วยความจำสำรองมาใช้งานร่วมกับการประมวลผล (In Memory) เช่น VoltDB

จ) MPP (Massively Parallel Processing) เป็นการประมวลผลแบบคู่ขนานใช้หน่วยประมวลผล CPU ร่วมกันหลายตัว เช่น Exadata, Greenplumn

ฉ) Graph Processing เป็นลักษณะการประมวลผลแบบกราฟบนเครือข่ายสังคมออนไลน์ด้วยทฤษฎีกราฟ เช่น Neo4j

2.3.2 การจำแนกงานวิจัยที่เกี่ยวข้อง ผู้วิจัยได้วิเคราะห์งานวิจัยในยุคข้อมูลขนาดใหญ่มีโปรแกรมที่เกี่ยวข้องจากงานวิจัยมีความหลากหลาย ดังนั้นผู้วิจัยขอแบ่งงานวิจัยที่เกี่ยวข้องออกเป็น 2 กลุ่ม เพื่อง่ายและสะดวกต่อการพิจารณา ดังนี้

2.3.2.1 กลุ่มนำเสนอการปรับปรุงประสิทธิภาพของเทคโนโลยีข้อมูลขนาดใหญ่ เช่น แนะนำวิธีการปรับปรุงกระบวนการภายในโปรแกรม จะทำให้เข้าใจการเขียนโปรแกรมมากขึ้นและการกำหนดค่าเริ่มต้นขนาดของไฟล์ข้อมูลเพื่อเพิ่มประสิทธิภาพในการจัดเก็บและเรียกใช้งานในกลุ่มคลัสเตอร์ขนาดเล็ก การนำฮาดูปและแมพรีดิวมาประเมินประสิทธิภาพในกระบวนการเพื่อปรับปรุงกระบวนการใน 2 ขั้นตอนแมพและรีดิว และการทดลองปรับปรุงฮาดูปและแมพรีดิวในการใช้ประมวลผลชุดข้อมูลรูปแบบต่างๆ และนำแมพรีดิวมาสร้างการเชื่อมโยงระหว่างฐานข้อมูลเอกสารด้วย MongoDB

2.3.2.2 กลุ่มประเมินผลเปรียบเทียบประสิทธิภาพ เช่น ใช้เทคโนโลยีข้อมูลขนาดใหญ่ทำการเปรียบเทียบเชิงทดลองกับกลุ่มฐานข้อมูลแบบดั้งเดิมหรือเปรียบเทียบกันในกลุ่มฐานข้อมูลรูปแบบใหม่ ตัวอย่างเช่น ใช้ NoSQL เปรียบเทียบกับ RDBMS ด้านการเขียน, อ่าน, แก้ไขและลบข้อมูล ทั้งแบบเครื่องเดียวและแบบหลายเครื่อง หรือ NoSQL เปรียบเทียบกับ SQL แบบเชิงสาเหตุด้านความแตกต่างของโครงสร้างและรูปแบบการจัดการ หรือ NoSQL เปรียบเทียบเชิงทดลองกับ NewSQL ในด้านการนำเข้า, อ่าน, เขียนและค้นหาบนคอมพิวเตอร์กลุ่มเมฆเน้นด้านโปรแกรมบนกลุ่มเครือข่ายสังคม หรือ Graph DB เปรียบเทียบเชิงทดลองกับ RDBMS ในด้านการค้นหาตามขนาดของตัวอักษรและขนาดข้อมูล หรือแมพรีดิวเปรียบเทียบเชิงทดลองกับ HiveQL และ RDBMS ด้านการเขียนข้อมูล ด้วยข้อมูลขนาด 200MB-10GB ทั้งแบบเครื่องเดียวและแบบหลายเครื่อง

2.3.3 การสรุปงานวิจัยที่เกี่ยวข้องกับที่ใกล้เคียงกับจุดประสงค์งานวิจัยนี้ดังนี้

2.3.3.1 การทดสอบฮาดูปแมพรีดิวด้วยการใช้ข้อมูลขนาดเล็กในการอ่านและเขียนข้อมูลขนาด 512MB, 2GB, 4GB และใช้ขนาดบล็อกข้อมูลที่ 64MB และ 128MB พบว่าในกลุ่มข้อมูลขนาดเล็กจะมีประสิทธิภาพมากหากใช้บล็อกข้อมูล 64MB และมีประสิทธิภาพมากขึ้น 28.6% เมื่ออ่านข้อมูลขนาด 512MB และเมื่ออ่านข้อมูลขนาด 4GB ที่ 25.3%

2.3.3.2 การเขียนแมพรีดิวเชื่อมคอลเลคชั่นฐานข้อมูล MongoDB การเชื่อมหรือ Join ข้อมูลแบบเอกสารด้วยข้อมูลนักเรียนและที่ปรึกษาเพื่อใช้เชื่อมความสัมพันธ์ พบว่าการดำเนินการสามารถเชื่อมข้อมูลได้และมีประสิทธิภาพเพิ่มขึ้นเมื่อกำหนดให้ทำการกรองข้อมูลที่ต้องการเชื่อมไว้ล่วงหน้า ใช้เวลาเพียง 23 วินาที

2.3.3.3 งานวิจัยการปรับปรุงประสิทธิภาพการค้นคืนด้วยแมพรีดิวด้วยหลักการเชื่อมความสัมพันธ์และทำดิวรีแพลนเพื่อให้เผยขั้นตอนการเชื่อมในขั้นตอนที่ดีที่สุด และทำการค้นคืนด้วยภาษาสอบถามเชิงโครงสร้างที่จำนวน 4, 6, 8 เครื่อง ด้วยข้อมูลอันดับเข้าดูเว็บไซต์ โดยทดลองกับ 2 กลุ่ม กลุ่มปรับปรุงประสิทธิภาพได้ผลความเร็วดีขึ้นเมื่อใช้ 8 เครื่อง แต่จะใช้เวลามากกว่ากลุ่มข้อมูลที่ไม่มีการปรับปรุงประสิทธิภาพ เมื่ออัลกอริทึมสั่งเพิ่มข้อมูลที่ระดับ 7-8 รอบ

2.3.3.4 งานวิจัยการปรับปรุงอัลกอริทึมแมพรีดิวเพื่อเพิ่มประสิทธิภาพ เช่น การเชื่อม, การจัดเรียง, การจัดกลุ่มด้วยข้อมูลหลายรูปแบบ ผลรวมการประมวลผลใช้เวลา 9 พันวินาที หรือ 2 ชั่วโมงครึ่ง ที่ข้อมูลขนาด 500GB ด้วยข้อมูลที่กำหนดสมมุติขึ้น (Synthetic Data)

2.3.3.5 งานวิจัยที่ใช้แมพรีดิว, ไฮฟ์และมายเอสคิวแอลทำการทดสอบด้วยข้อมูลการชำระเงินของลูกค้าในธุรกิจขนาดเล็ก มีข้อมูลลูกค้าตั้งแต่ 500-20,000 บัญชี มีขนาดข้อมูลตั้งแต่ 235MB-9GB กับเครื่องจำนวน 1-4 เครื่อง ผลสรุปว่ามายเอสคิวแอลจะใช้เวลามากกว่าแมพรีดิวและไฮฟ์ที่ขนาดข้อมูล 1 หมื่นบัญชี หรือ 5GB ใช้เวลา 25 นาที แมพรีดิวจะใช้เวลาน้อยที่สุดในการประมวลผลทั้ง 1-4 เครื่อง ใช้เวลาโดยประมาณ 80-90 วินาที ในทุกชุดข้อมูลทดสอบ โปรแกรมแมพรีดิวมีประสิทธิภาพสม่ำเสมอและดีที่สุด

ทุกงานวิจัยมีวัตถุประสงค์คล้ายคลึงกันคือ เพื่อหาความเหมาะสมและรูปแบบการใช้งาน ที่สามารถนำมาใช้กับข้อมูลในรูปแบบต่างๆ และเพื่อค้นหาแนวทางการเพิ่มประสิทธิภาพให้กับการทำงานบนเทคโนโลยีข้อมูลขนาดใหญ่ ซึ่งได้ผลการเปรียบเทียบที่ให้ผลไปในทิศทางเดียวกันคือ เทคโนโลยีข้อมูลขนาดใหญ่จะมีประสิทธิภาพด้านความเร็ว เมื่อข้อมูลมีขนาดใหญ่ขึ้น

แต่ทว่างานวิจัยที่เกี่ยวข้องนี้ ผู้วิจัยยังไม่พบงานใดทำการประเมินผลความแม่นยำถูกต้องของผลลัพธ์ข้อมูล งานวิจัยนี้จึงขอนำเสนอ การประเมินประสิทธิภาพด้านความเร็วร่วมกับการประเมินผลความถูกต้องของผลลัพธ์ ด้วยการประมวลผลชุดข้อมูลที่มีการขยายตัวของข้อมูลอย่างเป็นลำดับ เพื่อหาจุดตัดของกราฟด้านผลความเร็ว และผลลัพธ์ที่ถูกต้องตรงกันทุกชุดข้อมูลที่ใช้ในการทดลอง

การคัดเลือกเครื่องมือที่ใช้ในการทดลองนี้จากการอ่านงานวิจัยที่เกี่ยวข้องกับการใช้งานเทคโนโลยีข้อมูลขนาดใหญ่ จึงสรุปได้ว่าโปรแกรมฮาดูปและแมพรีดิวมีความเหมาะสมที่จะนำมาใช้งานในการวิจัยในครั้งนี้เนื่องด้วยสาเหตุที่ว่า เครื่องมือนี้สามารถค้นหาองค์ความรู้ทางวิชาการและสามารถค้นคว้าทำการศึกษาหลักการทำงานขั้นพื้นฐานได้ง่ายและสะดวก

ในยุคข้อมูลขนาดใหญ่กรอบการทำงานฮาดูปและแมพรีดิวเป็นโปรแกรมที่สามารถรองรับได้กับระบบปฏิบัติการหลายระบบ การคัดเลือกศึกษาลักษณะรูปแบบวิธีการใช้งานสามารถศึกษาได้เป็นจำนวนมากทำให้สามารถนำมาศึกษาได้ง่าย อีกทั้งฮาดูปและแมพรีดิวยังได้รับความนิยมนำมาใช้ในวงการการศึกษาวิจัยทั้งภาคอุตสาหกรรมและภาคธุรกิจต่างๆ อย่างกว้างขวางเพื่อพัฒนาให้ระบบมีประสิทธิภาพเพิ่มขึ้น

ฮาดูปและแมพรีดิวยังมีข้อดีในการใช้งานง่ายและสามารถเขียนได้โดยโปรแกรมเมอร์ที่ไม่มีประสบการณ์กับระบบแบบขนานและระบบแบบกระจาย ด้วยการสนับสนุนจากเจ้าของ

ผลิตภัณฑ์ทำให้โปรแกรมเมอร์สามารถเขียนโปรแกรมได้ง่ายขึ้น เพราะโปรแกรมฮาดูปและแมพรีดิวมีการซ่อนรายละเอียดของการทนทานต่อความผิดพลาดบนเครือข่าย

แต่ทั้งนี้โปรแกรมแมพรีดิวยังมีข้อจำกัดในด้านการประมวลผลที่ไม่สามารถสร้างอัลกอริทึมการประมวลผลแบบซ้ำแล้วซ้ำอีกได้ เป็นโปรแกรมการประมวลผลแบบกลุ่ม (Batch Processing) หรือจะกระทำใหม่ทุกครั้งที่มีการสั่งรันโปรแกรมใหม่ ทำให้ประสิทธิภาพในการเข้าถึงข้อมูลลดลงจึงถือว่าเป็นค่าใช้จ่ายที่เกิดขึ้นในอนาคตที่ต้องใช้โปรแกรมเสริมเพิ่มเติมเข้ามาจัดการโดยเฉพาะ เช่น โปรแกรมเทคโนโลยีข้อมูลขนาดใหญ่ เช่น Spark หรือ Tajo เป็นต้น

บทที่ 3
แนวคิด และวิธีดำเนินงานวิจัย

งานวิทยานิพนธ์นี้มีแนวคิดการวิจัยเพื่อศึกษาเหตุการณ์ที่เกิดขึ้นจากการทดลองว่าเกิดขึ้นได้อย่างไร มีสาเหตุมาจากอะไร และทำไมจึงเป็นเช่นนั้น การศึกษาความสัมพันธ์ของชุดข้อมูลบริการสุขภาพกับเทคนิควิธีการประมวลผลข้อมูล 2 รูปแบบ ระหว่างเทคโนโลยีข้อมูลขนาดใหญ่ที่มีรูปแบบการจัดเก็บแบบกระจาย (ฮาดูป) และการประมวลผลแบบขนาน (แมพรีดิว) นำมาศึกษาความแตกต่างระหว่างกลุ่มการประมวลผลข้อมูลที่มีหลักการทางคณิตศาสตร์ที่แตกต่างกัน และมีสถาปัตยกรรมการจัดการข้อมูลที่ไม่เหมือนกัน และเทคนิควิธีการสอบถามค้นคืนข้อมูลที่แตกต่างกันกับระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ (มายเอสคิวแอล)

การประมวลผลข้อมูลชุดตัวอย่างเดียวกันเพื่อลดข้อแตกต่างและข้อขัดแย้งของข้อมูล ทำการวิจัยด้วยวิธีการเชิงทดลองด้วยชุดข้อมูลบริการสุขภาพ และทำสร้างชุดแบบสอบถามขึ้นจากรายงานสรุปการเจ็บป่วย พ.ศ.2557 รายงานผู้ป่วยนอก จำนวน 2 รายงาน เป็นเครื่องมือที่นำมาใช้หาประสิทธิภาพทางด้านความเร็วและประสิทธิผลทางด้านความแม่นยำถูกต้องของสารสนเทศ อีกทั้งมีการทดสอบการปรับปรุงประสิทธิภาพและประสิทธิผลในกระบวนการสอบถามข้อมูลในการเรียกคืนข้อมูลเพื่อให้ได้ประสิทธิภาพด้านเวลาการค้นคืน และประสิทธิผลด้านความถูกต้อง

การวิเคราะห์ผลทางสถิติโดยใช้สถิติและระเบียบวิธีวิจัยทางด้านวิทยาศาสตร์และเทคโนโลยีสารสนเทศมาประยุกต์ร่วมกัน และนำผลลัพธ์ทางด้านเวลามาวิเคราะห์ผลข้อมูลทางด้านประสิทธิภาพของเวลาและประสิทธิผลของผลลัพธ์ มีรายละเอียดดังนี้

3.1 กรอบแนวคิดการออกแบบงานวิจัย

กรอบแนวคิด (Conceptual Framework) ก่อนเริ่มออกแบบกระบวนการทดลองผู้วิจัยมีแนวคิดเชิงวิเคราะห์ ด้วยการเก็บรวบรวมข้อมูลจากแหล่งข้อมูลทุติยภูมิ (Secondary Data) เช่น การค้นคว้าจากเอกสารทางราชการและหลักฐานในงานวิชาการด้านสาธารณสุข ได้จำแนกช่วงเวลาสถานการณ์ด้านเทคโนโลยีสารสนเทศของกระทรวงสาธารณสุขออกเป็น 2 ช่วงเวลา คือก่อนกระทรวงสาธารณสุขเริ่มรวบรวมข้อมูลด้านการเจ็บป่วย และหลังกระทรวงเริ่มรวบรวมข้อมูลด้านการเจ็บป่วย ซึ่งผู้วิจัยได้กล่าวถึงแล้วในบทที่ 1 หน้าที่ 1 และบทที่ 2 หน้าที่ 8-14 ข้อสรุปจากการ

วิเคราะห์ข้อมูลเชิงกราฟ (Graphical Analysis) ตามภาพที่ 1.1 จากกราฟผลเป็นที่ประจักษ์ ข้อมูลมีแนวโน้มเพิ่มขึ้นขนาดใหญ่ขึ้นทุกปี เป็นปัญหาอีกแง่มุมหนึ่งของผู้ใช้ข้อมูลกระทรวงสาธารณสุข เมื่อผู้วิจัยตั้งคำถามว่า ควรเริ่มต้นจากจุดใด หากพิจารณาจากปัญหาของกระทรวงสาธารณสุข ที่เป็นโจทย์ในงานวิจัยคือ ปัญหาการรวบรวมข้อมูลการเจ็บป่วยจากสถานพยาบาลทุกหน่วยงานในสังกัดกระทรวงสาธารณสุขจากระดับอำเภอนำข้อมูลเข้าสู่ระดับจังหวัดเพื่อประมวลผลเข้าสู่ส่วนกลางระดับเขตและกระทรวงสาธารณสุข มีข้อมูลจำนวนเพิ่มขึ้น ขนาดใหญ่ขึ้นในทุกๆ ปี การเรียกใช้ข้อมูลการมารับบริการทางการแพทย์ เพื่อทำการประมวลผลนำสารสนเทศมาใช้งานการวิเคราะห์ข้อมูลทางการแพทย์ทำได้ช้าลง

คำถามเชิงวิเคราะห์ การเริ่มต้นจากคำถาม เพื่อนำไปสู่กระบวนการค้นหาคำตอบในการแก้ไขปัญหา โดยใช้เครื่องมือการวิเคราะห์ปัญหาด้วยเทคนิค 5W1H เพื่อวิเคราะห์ข้อมูล แจกแจงหาแนวทางการแก้ไข

เมื่อผู้วิจัยพิจารณาจากการทบทวนวรรณกรรมระบบบริการสุขภาพได้มีแนวทางการคิดเชิงวิเคราะห์เพื่อการเตรียมวิธีการทดลอง ได้พิจารณาแนวคิดเกี่ยวกับการวางแผนการออกแบบการทดลอง การเตรียมวิธีการทดลอง จากคำถามหรือโจทย์ของปัญหาที่ได้กล่าวมาแล้วข้างต้น สามารถสรุปการแจกแจงได้ดังนี้

Who ใคร (ในเรื่องนั้นมีใครบ้าง) ผู้บันทึกข้อมูลสถานพยาบาล เจ้าหน้าที่ผู้เกี่ยวข้อง ผู้ใช้ข้อมูลจัดทำรายงานหรืองานวิจัย เจ้าหน้าที่สาธารณสุขจังหวัดและเขตและกระทรวง

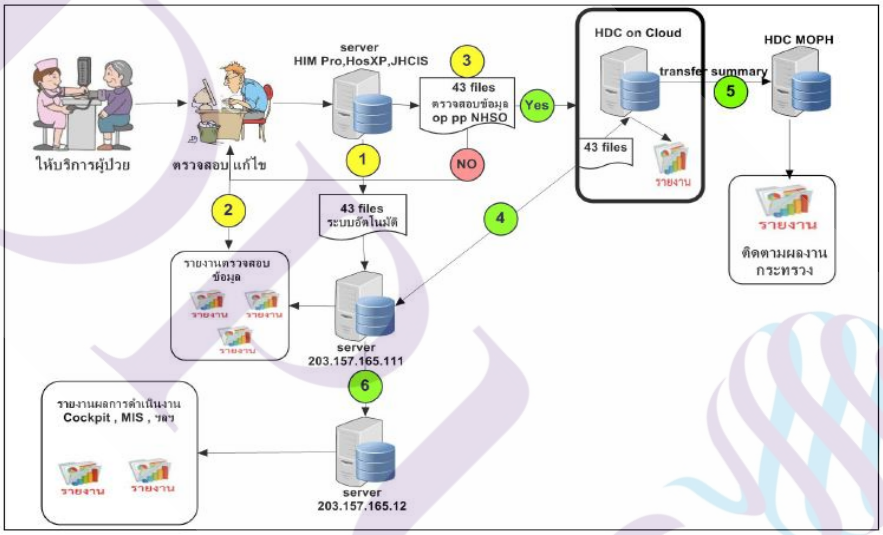
What ทำอะไร (แต่ละคนทำอะไรบ้าง) ผู้บันทึกข้อมูลบริการสุขภาพเข้าระบบโปรแกรมฐานข้อมูลของสถานพยาบาล เจ้าหน้าที่ผู้เกี่ยวข้องตรวจสอบปรับปรุงแก้ไขข้อมูล 43+7 แฟ้ม เจ้าหน้าที่ผู้เกี่ยวข้องส่งข้อมูลของสถานพยาบาลเก็บรวบรวมข้อมูลการเจ็บป่วยนำส่งข้อมูล 43+7 แฟ้มเข้าเครื่องแม่ข่ายระดับจังหวัด 43+7 แฟ้ม เจ้าหน้าที่สาธารณสุขระดับเขตและกระทรวงใช้ข้อมูลทำรายงานสถิติการแพทย์และตรวจสอบรายงานดัชนีชี้วัด

Where ที่ไหน (เหตุการณ์หรือสิ่งที่ทำนั้นอยู่ที่ไหน) การจัดส่งข้อมูลจัดทำบนเครื่องลูกข่ายระดับอำเภอเข้าสู่เครื่องแม่ข่ายระดับจังหวัด การจัดส่งข้อมูล 43+7 แฟ้ม จัดทำบนเครื่องแม่ข่ายระดับจังหวัดสู่เครื่องแม่ข่ายระดับเขตและกระทรวง การประมวลผลเพื่อจัดทำรายงานสถิติและดัชนีชี้วัดบนเครื่องแม่ข่ายระดับจังหวัด การประมวลผลเพื่อจัดทำรายงานผลการดำเนินงาน รายงานสถิติการเจ็บป่วยและดัชนีชี้วัดบนเครื่องแม่ข่ายระดับเขตและกระทรวง

When เมื่อไหร่ (เหตุการณ์หรือสิ่งที่ทำนั้นทำเมื่อวันเดือนปีใด) การส่งข้อมูลจากเครื่องลูกข่ายระดับอำเภอเข้าสู่เครื่องแม่ข่ายระดับจังหวัดต้องดำเนินการจัดทำไม่เกิน 7 วันนับจากวันให้บริการ การตรวจสอบและแก้ไขเพื่อการประมวลผลจัดทำรายงานในเครื่องแม่ข่ายระดับจังหวัด

จัดทำทุกสิ้นเดือน การส่งข้อมูลเข้าเครื่องแม่ข่ายระดับเขตและส่วนกลางระดับกระทรวง จะกระทำภายในวันที่ 15 ของเดือนถัดไปและส่งแก้ไขข้อมูลย้อนหลังภายในเดือนนั้น

Why ทำไม (เหตุใดจึงได้ทำสิ่งนั้น หรือเกิดเหตุการณ์นั้นๆ) การส่งข้อมูล 43+7 แฟ้ม เป็นข้อมูลการบริการทางการแพทย์ของสถานพยาบาลภายใต้สังกัดกระทรวงสาธารณสุขเข้าสู่เครื่องแม่ข่ายระดับจังหวัด ระดับเขตและระดับกระทรวงซึ่งเป็นฐานกลางข้อมูลสุขภาพส่วนกลางระดับประเทศ เพื่อรวบรวมข้อมูลการเจ็บป่วยนำมาประมวลผลจัดทำรายงานสถิติทางการแพทย์ รายงานผลการดำเนินงานเพื่อการบริหารจัดการนำมาใช้ในการบริหารงานวางนโยบายของกระทรวงสาธารณสุขในการจัดบริการทางการแพทย์ อุปกรณ์ทางแพทย์ หรือเพื่อการบริหารจัดการระบบสุขภาพเพื่อประโยชน์ของประชากรชาวไทยทั้งประเทศ



ภาพที่ 3.1 กระบวนการระบบการส่งข้อมูลบริการสุขภาพ

ที่มา: สำนักงานสาธารณสุขจังหวัดศรีสะเกษ [online] : เข้าถึง 26 ก.พ. 2559. จาก http://www.khukhanph.com/2016/02/2559.html

How อย่างไร (เหตุการณ์หรือสิ่งที่ทำนั้น ทำเป็นอย่างไรบ้าง) กระทรวงสาธารณสุขมีนโยบายใช้โปรแกรมที่พัฒนาขึ้นโดยหน่วยงานศูนย์เทคโนโลยีสารสนเทศและการสื่อสารกระทรวงสาธารณสุขร่วมกับศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (เนคเทค) และโปรแกรมที่ได้รับการรับรองจากกระทรวงสาธารณสุขให้สามารถใช้งานบันทึกข้อมูลได้ในสถานพยาบาลระดับตำบล อำเภอและจังหวัด มีวัตถุประสงค์เพื่อจัดเก็บข้อมูลการบริการสุขภาพใน

ฐานข้อมูลที่มีการออกแบบแฟ้มข้อมูลที่มีมาตรฐานโครงสร้างตามที่กระทรวงสาธารณสุขกำหนด และเชื่อมโยงกันแบบมีความสัมพันธ์ โดยใช้โปรแกรมโอเพ่นซอร์สระบบการจัดการฐานข้อมูลมายเอสคิวแอลเป็นฐานข้อมูลระดับจังหวัด ระดับเขต ระดับกระทรวง ใช้งานได้ครอบคลุมทั้ง 76 จังหวัด ในปี พ.ศ.2558 และศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร สำนักงานปลัดกระทรวงสาธารณสุข ยังมีโครงการพัฒนาระบบเทคโนโลยีสารสนเทศและการสื่อสารตามกรอบยุทธศาสตร์เทคโนโลยีสารสนเทศสุขภาพ มีกลยุทธ์การพัฒนาคลังข้อมูลสุขภาพ กำหนดรูปแบบการบริหารจัดการคลังข้อมูลระบบบริการสุขภาพในระบบข้อมูลขนาดใหญ่ (Big Data Management in Healthcare System) เพื่อให้มีความเหมาะสมในการใช้งานให้กับหน่วยงานแต่ละระดับ

สรุปการวิเคราะห์ 5W1H ที่เกี่ยวข้องกับองค์ประกอบของปัญหาได้ดังนี้ เมื่อข้อมูลบริการสุขภาพหรือข้อมูลการเจ็บป่วยของผู้ป่วยของสถานพยาบาลในสังกัดกระทรวงสาธารณสุขมีเพิ่มขึ้นและขนาดใหญ่ขึ้น เทคโนโลยีฐานข้อมูลมายเอสคิวแอลที่ใช้อยู่ปัจจุบันไม่เพียงพอกับจำนวนข้อมูลที่เพิ่มขึ้นเป็นภาระให้กับผู้ดูแลระบบที่จะต้องจัดตารางเวลาการประมวลผลรายงานสถิติและผลการดำเนินการออกจากตารางเวลาการประมวลผลข้อมูลการปฏิบัติงานประจำวัน และการดูแลปรับปรุงฐานข้อมูล จึงยังไม่ตอบสนองการเรียกใช้ข้อมูลหรือสารสนเทศได้ทันต่อความต้องการ อีกทั้งโอเพ่นซอร์สมายเอสคิวแอลได้มีการเปลี่ยนแปลงภายในองค์กรซึ่งปัจจุบันขึ้นตรงกับออราเคิล จึงเกิดความเสี่ยงหากเจ้าของโปรแกรมเรียกเก็บเงินค่าบริการราคาสูง อีกทั้งผู้ที่เกี่ยวข้องในการดำเนินการจัดทำสารสนเทศกระทรวงสาธารณสุขมีโครงการพัฒนาระบบเทคโนโลยีสารสนเทศด้วยการนำโปรแกรมเทคโนโลยีบิ๊กดาต้าเข้ามาบริหารจัดการข้อมูลคลังข้อมูลระบบบริการสุขภาพ ทั้งนี้ยังไม่ทราบแน่ชัดในส่วนของโปรแกรมบริหารจัดการข้อมูลขนาดใหญ่ที่กระทรวงเลือกใช้ ซึ่งจากการศึกษาเทคโนโลยีบิ๊กดาต้านั้น ปัจจุบันมีผลิตภัณฑ์หลากหลายรูปแบบ หลากหลายผู้ผลิต แต่ทุกผลิตภัณฑ์ โดยส่วนมากมีสถาปัตยกรรมภายในโปรแกรมที่เหมือนกันคือ การใช้กรอบการทำงานฮาดูปและแมพรีดิวเป็นโครงสร้างพื้นฐานของโปรแกรม

ในปัจจุบันเทคโนโลยีระบบแบบกระจาย (Distributed System) และการประมวลผลแบบขนาน (Parallel Processing) เช่น โปรแกรมโอเพ่นซอร์สอาปาเช่ฮาดูป (Apache Hadoop) เป็นระบบการจัดเก็บแบบกระจาย (Hadoop Distributed File System : HDFS) และการประมวลผลแบบขนานด้วยเทคนิคแมพรีดิว (MapReduce) เทคโนโลยีนี้จะสามารถนำมาช่วยเหลือการประมวลผลข้อมูลขนาดใหญ่ จากคลังข้อมูลด้านการแพทย์และสุขภาพ ที่มีการจัดการฐานข้อมูลแบบเชิงสัมพันธ์ได้หรือไม่ และมีแนวทางในการดำเนินการอย่างไร (How) หากต้องปรับเปลี่ยนวิธีการประมวลผล หรือหากจะต้องประยุกต์ใช้กับสิ่งที่มีอยู่เดิมต้องทำอย่างไร (How) ซึ่งการทำงานกับ

ข้อมูลขนาดใหญ่จะมีปัญหาในการจัดเก็บข้อมูล การโอนย้ายข้อมูล การสำรองข้อมูล และการสืบค้นคืนข้อมูล จะมีวิธีการอย่างไร (How) ที่จะช่วยทำให้การจัดการสืบค้นคืนข้อมูล และนำข้อมูลจำนวนมากเหล่านี้มาใช้ประโยชน์ได้ภายในเวลาอันรวดเร็วอย่างมีประสิทธิภาพมากที่สุด แต่ยังคงไว้ให้ได้ซึ่งความถูกต้องของข้อมูลที่ได้รับการสืบค้นคืน ด้วยความสำคัญของคุณภาพข้อมูลในระบบบริการสุขภาพจากลักษณะสำคัญ 4 ส่วนคือ ครบถ้วน ถูกต้อง ละเอียด และทันสมัย อีกทั้งยังสร้างโอกาสในการพัฒนาสถานพยาบาลให้เจริญก้าวหน้าต่อไปในอนาคตด้วยค่าใช้จ่ายในการลงทุนทรัพยากรและต้นทุนความเป็นเจ้าของที่ต่ำ

เทคโนโลยีที่จะนำมาใช้ศึกษาวิจัยในงานวิทยานิพนธ์ฉบับนี้จากปัญหาที่กล่าวมาผู้วิจัยจึงเลือกใช้เทคโนโลยีฮาดูปและแมพรีดิวเป็นโปรแกรมที่นำมาใช้เพื่อการทดลองหาความเหมาะสมในการใช้งานร่วมกับข้อมูลในระบบบริการสุขภาพ

3.2 ขั้นตอน และวิธีการดำเนินงานวิจัย

ผู้วิจัยนำแนวคิดหลักการออกแบบการวิจัยการทดลองขั้นพื้นฐานของการวิจัยเชิงทดลอง 3 ประการ โดยการกำหนดตัวแปร (Variable) และแบบแผนการทดลอง (Experimental design) ที่ใช้สำหรับการออกแบบการทดลองดังนี้

ภาพที่ 3.2 การกำหนดตัวแปรที่ใช้ในการทดลอง

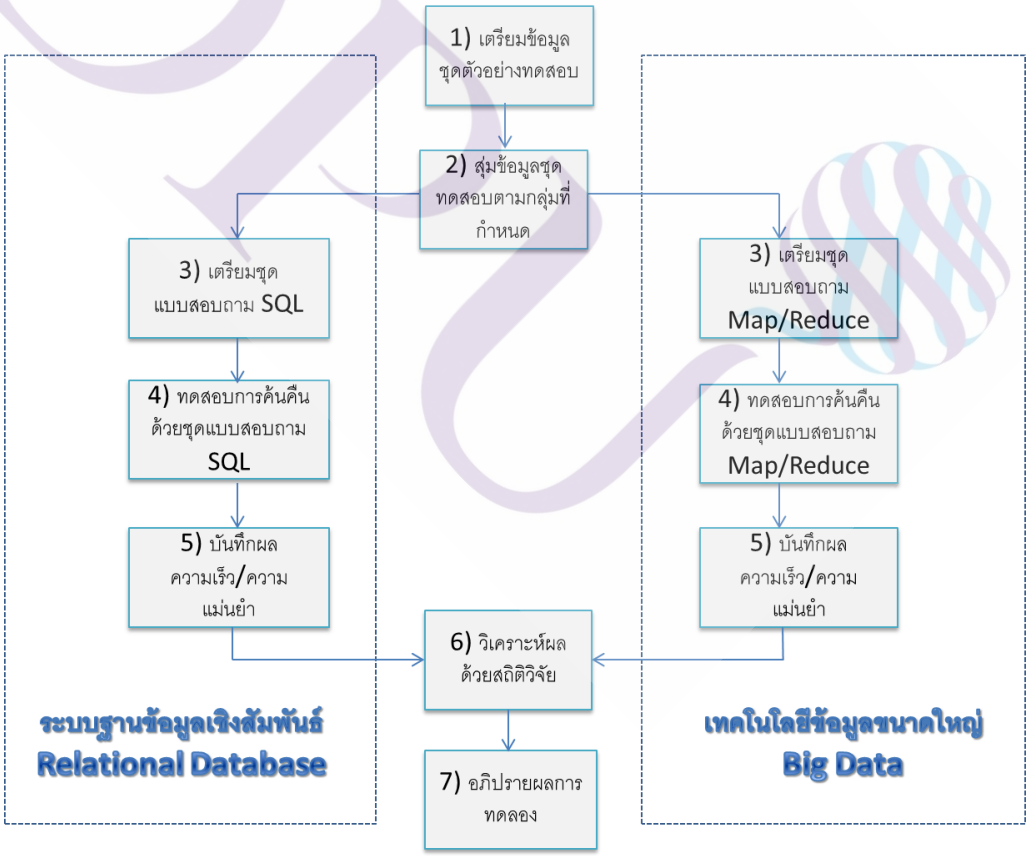
ประการที่ 1 การเพิ่มความแปรปรวนของการทดลองให้มากที่สุด (Maximization of experimental variance) ออกแบบการเพิ่มความแปรปรวนของการทดลองด้วยการออกแบบชุดข้อมูลที่มีจำนวนระเบียนเพิ่มขึ้น ใช้วิธีแบบง่ายกำหนดช่วงข้อมูล โดยการกำหนดข้อมูลเริ่มต้น ห้าแสนระเบียนเพิ่มเป็นหนึ่งล้านระเบียน และห้าล้านระเบียน และสิบล้านระเบียนตามลำดับ

ประการที่ 2 การลดความแปรปรวนของความคลาดเคลื่อนให้น้อยที่สุด (Minimization of error variance) ออกแบบการลดความแปรปรวนของความคลาดเคลื่อนในด้านผลของความเร็วการค้นหา ด้วยสภาพแวดล้อมและเครื่องมือชนิดเดียวกันในขณะทำการทดสอบ

ประการที่ 3 การควบคุมตัวแปรแทรกซ้อน (Control of extraneous variables) ออกแบบการควบคุมตัวแปรแทรกซ้อนโดยการทดสอบด้วยกลุ่มตัวอย่างและชุดแบบสอบถามในการค้นหาชุดเดียวกัน เพื่อลดความคลาดเคลื่อนของผลลัพธ์ของการทดลอง

หลักการออกแบบขั้นพื้นฐานใช้เป็นกรอบแนวคิดนำกระบวนการทดลอง กำหนดแบบแผนการทดลองจริง (True-experimental design) มีการเก็บผลการทดลองจากการประมวลผลจำนวน 3 ครั้ง และหาค่าเฉลี่ย ทำการเปรียบเทียบระหว่างกลุ่มการทดลองประมวลผล 2 รูปแบบ

การกำหนดขั้นตอนและวิธีการดำเนินการทดลอง ผู้วิจัยคำนึงถึงการประมวลผลของระบบเทคโนโลยีในการเปรียบเทียบยึดหลักเครื่องมือที่ใช้อยู่ในปัจจุบันกับเครื่องมือที่ต้องการใช้ในอนาคตจะรองรับการปรับเปลี่ยนหรือไม่ กำหนดแนวทางของขั้นตอนและวิธีการทดลองไว้ดังนี้



ภาพที่ 3.3 ขั้นตอนและวิธีดำเนินการทดลอง

1) เตรียมข้อมูลชุดทดสอบ
2) สุ่มคัดเลือกข้อมูลชุดทดสอบตามกลุ่มที่กำหนด
3) เตรียมชุดแบบสอบถามทดสอบเอสคิวแอลและแมพรีดิว
4) ทดสอบการประมวลผลด้วยชุดแบบสอบถาม
5) บันทึกผลลัพธ์จากการประมวลผล
6) นำผลลัพธ์ที่ได้นำมาวิเคราะห์สถิติเปรียบเทียบการประมวลผล
7) สรุปผลที่ได้จากการวิเคราะห์สถิติ

3.2.1 เตรียมข้อมูลชุดทดสอบ

ชุดข้อมูลทดสอบเป็นข้อมูลบริการสุขภาพ เก็บรวบรวบข้อมูลจากกระทรวงสาธารณสุข ด้วยระบบคอมพิวเตอร์เครื่องแม่ข่ายระดับจังหวัด การคัดเลือกนำชุดข้อมูลทดสอบทำการคัดเลือกการสุ่มตัวอย่างด้วยเทคนิควิธีแบบเฉพาะเจาะจง (Purposive Sampling) การคัดเลือกกลุ่มตัวอย่างแบบเจาะจงเป็นการเลือกกลุ่มตัวอย่างโดยอาศัยการตัดสินใจ (Judgment Sampling) การคัดเลือกนี้ใช้การตัดสินใจจากผู้เชี่ยวชาญข้อมูลสุขภาพเป็นผู้คัดเลือกข้อมูล ทำการคัดเลือกด้วยเหตุผลในด้านข้อมูลผู้ป่วยเป็นความลับ ข้อมูลการเจ็บป่วยไม่สามารถเปิดเผยได้ตามกฎหมายสาธารณสุขที่ว่าด้วยข้อมูลการเจ็บป่วยของผู้ป่วยเป็นความลับ ผู้เชี่ยวชาญดำเนินการตัดข้อมูลชื่อและนามสกุลออกจากข้อมูลการทดสอบ และทำการสุ่มคัดเลือกนำตัวอย่างกลุ่มข้อมูลทดสอบเพียงบางส่วนตามจำนวนที่ผู้วิจัยต้องการทดสอบโดยการคัดเลือกกลุ่มตัวอย่าง 3 จังหวัด จากทั้งสิ้น 76 จังหวัด และคัดเลือกข้อมูลชุดตัวอย่าง 2 แฟ้ม จากแฟ้มมาตรฐาน 43 แฟ้ม มีแฟ้มข้อมูล DIAGNOSIS_OPD คือ แฟ้มข้อมูลวินิจฉัยโรคของผู้ป่วยนอกและผู้มารับบริการ

ตารางที่ 3.1 จำนวนข้อมูลชุดทดสอบแฟ้ม Diagnosis_opd

แฟ้ม Diagnosis_opd

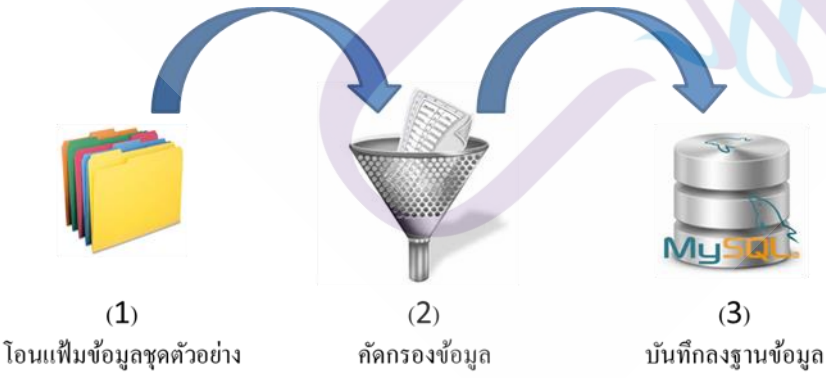
ชื่อแฟ้ม	จำนวนระเบียน	ขนาดไฟล์ (MB)
diagnosis_opd_1.txt	899,972	65.2
diagnosis_opd_2.txt	5,558,268	374.0
diagnosis_opd_3.txt	6,308,357	454.0
รวม	12,766,597	893.2

ข้อมูลบริการสุขภาพ สำนักนโยบายยุทธ์ศาสตร์และยุทธศาสตร์ กระทรวงสาธารณสุข มีการกำหนดโครงสร้างมาตรฐานข้อมูลด้านการแพทย์และสุขภาพไว้เป็นมาตรฐาน ผู้วิจัยดำเนินการสร้างตารางในฐานข้อมูลทดสอบที่มีลักษณะโครงสร้างไว้ดังนี้

ตารางที่ 3.2 โครงสร้างแฟ้มข้อมูลมาตรฐานของตาราง Diagnosis_opd

No	Field	Caption	Primary	Type Data	Not Null
1	HOSPCODE	รหัสสถานบริการ	Y	CHAR(5)	Y
2	PID	ทะเบียนบุคคล	Y	CHAR(15)	Y
3	SEQ	ลำดับที่	Y	CHAR(16)	Y
4	DATE_SERV	วันที่ให้บริการ		DATE	Y
5	DIAGTYPE	ประเภทการวินิจฉัย		CHAR(1)	Y
6	DIAGCODE	รหัสการวินิจฉัย	Y	CHAR(6)	Y
7	CLINIC	แผนกที่รับบริการ		CHAR(5)	Y
8	PROVIDER	เลขที่ผู้ให้บริการ		CHAR(15)	
9	D_UPDATE	วันเดือนปีที่ปรับปรุง		DATETIME	Y

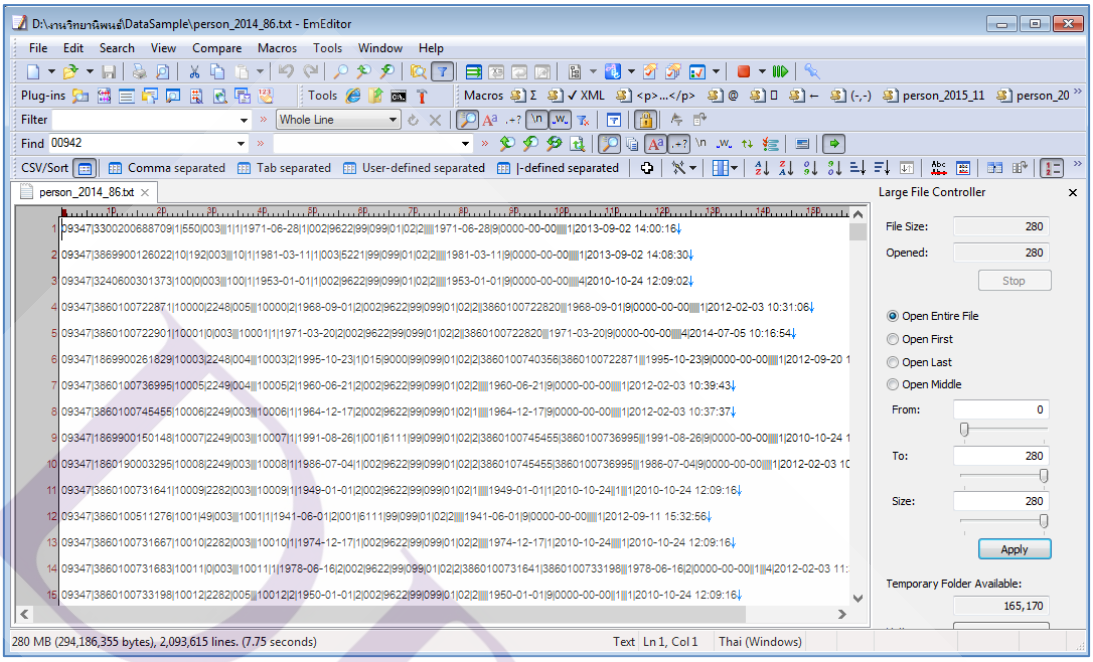
เริ่มต้นการคัดกรองข้อมูลชุดทดสอบที่ได้จากขั้นตอนการเตรียมชุดข้อมูลก่อนนำเข้าระบบฐานข้อมูลมายเอสคิวแอล



ภาพที่ 3.4 ขั้นตอนการคัดกรองตรวจสอบข้อมูลชุดทดสอบ

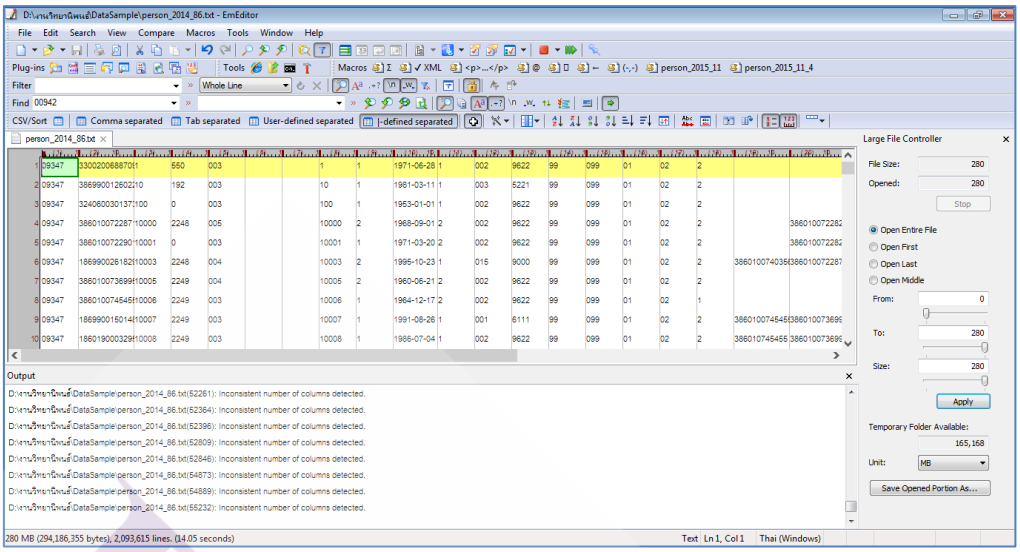
การผ่านขั้นตอนการคัดกรองตรวจสอบคุณภาพข้อมูล มีวัตถุประสงค์ในการตรวจเช็คความไม่สมบูรณ์ของข้อมูลที่ได้มาจากการบันทึก หรือการเกิดการสูญหายของข้อมูลในกระบวนการประมวลผลหรือส่งต่อข้อมูล เพื่อให้เกิดความเชื่อถือได้ในข้อมูลชุดก่อนทดสอบ การ

คัดกรองข้อมูลเมื่อพบข้อมูลไม่สมบูรณ์ จะทำการตัดข้อมูลที่ไม่สมบูรณ์ของแถวหรือระเบียนนั้นออกจากข้อมูลชุดทดสอบ



ภาพที่ 3.5 ภาพการแสดงข้อมูลชุดตัวอย่างก่อนคัดกรองด้วยโปรแกรม EmEditor

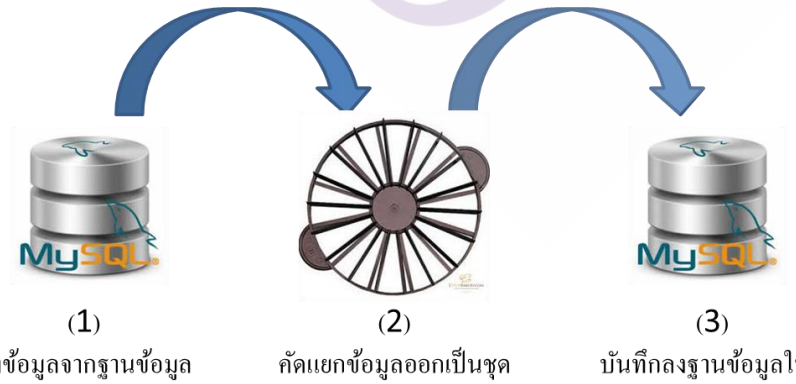
การตัดข้อมูลที่ไม่สมบูรณ์ (Define Separate) ด้วยวิธีการตรวจสอบพิสูจน์ตัวอักษร (Regular Expression) และการใช้อักขระพิเศษ (Wildcard) ในไฟล์ชุดตัวอย่าง ด้วยโปรแกรมอีเอ็มอีดิเตอร์ (EmEditor) เป็นเครื่องมือที่ใช้ในการตรวจเช็คข้อมูล ทำการจัดแบ่งคอลัมน์ด้วยอักขระพิเศษ เพื่อจัดรูปแบบระเบียนซึ่งคั่นคอลัมน์ด้วยอักขระพิเศษเส้นตั้งหรือไพป์ (Pipe) ใช้สัญลักษณ์เส้นในแนวตั้ง (|) ไฟล์ที่ผ่านการคัดกรองแล้วจะนำเข้าฐานข้อมูลมายเอสคิวแอล ตัวอย่างระเบียนข้อมูลแฟ้มข้อมูลทั่วไปของประชาชน 00933|3101490000023|000002|846| 003 |||0000002|1|1956-05-20|3|014|5152|099|099|01|03|2|||||2005-01-20|1||||||1|2012-07-25 16:47:17



ภาพที่ 3.6 ภาพการแสดงข้อผิดพลาดข้อมูลชุดตัวอย่างด้วยโปรแกรม EmEditor

3.2.2 สุ่มคัดเลือกข้อมูลชุดทดสอบตามกลุ่มที่กำหนด

การกำหนดขนาดตัวอย่างโดยการใช้วิธีการเลือกตัวอย่างแบบกำหนดจำนวนหรือแบบโควต้า (Quota Sampling) กำหนดจำนวนตัวอย่างไว้อย่างชัดเจนแล้วทำการคัดเลือกสุ่มข้อมูลชุดทดสอบตามจำนวนที่กำหนด หรือการเลือกแบบง่ายหรือแบบสะดวก (Simple Random Sampling) ใช้การสุ่มอย่างง่ายด้วยระบบคอมพิวเตอร์ โดยใช้ฟังก์ชั่น Rand() ในระบบฐานข้อมูล เป็นการสุ่มระเบียนออกจากฐานข้อมูลนำมาเก็บแยกเป็นตารางข้อมูลชุดทดสอบออกเป็น 4 ชุด การคัดเลือกจำนวนในชุดดังนี้ ชุดที่ 1 มีจำนวน 500,000 ระเบียน ชุดที่ 2 มีจำนวน 1,000,000 ระเบียน ชุดที่ 3 มีจำนวน 5,000,000 ระเบียน ชุดที่ 4 มีจำนวน 10,000,000 ระเบียน ตามลำดับ



ภาพที่ 3.7 ขั้นตอนการสุ่มข้อมูลการทดสอบออกเป็น 4 ชุดข้อมูล

ในขั้นตอนการสุ่มคัดเลือกข้อมูลชุดทดสอบตามกลุ่มที่กำหนด มีชุดคำสั่งฟังก์ชั่น Rand( ) เป็นชุดคำสั่งการสุ่ม การสุ่มในระบบฐานข้อมูลเพื่อทำการดึงระเบียนบันทึกแยกตารางจากระบบฐานข้อมูลเดิม วัตถุประสงค์เพื่อการควบคุมจำนวนระเบียนในการทดสอบ ลดตัวแปรแทรกซ้อนและความคลาดเคลื่อนที่จะส่งผลกระทบต่อเวลาการค้นหา และผลลัพธ์ที่จะได้จากการทดสอบและการตรวจสอบคุณภาพของเครื่องมือ ใช้การตรวจสอบโดยผู้เชี่ยวชาญข้อมูลระบบสุขภาพ

3.2.3 เตรียมชุดแบบสอบถามทดสอบเอสคิวแอลและแมพรีดิว

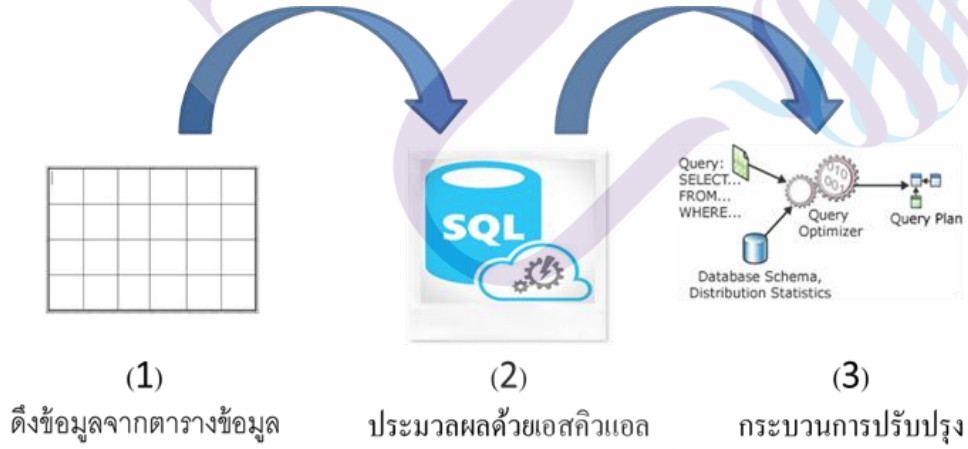
การเตรียมคำสั่งชุดทดสอบหรือชุดแบบสอบถามทดสอบเอสคิวแอลและแมพรีดิว ด้วยการใช้ชุดแบบสอบถามที่คำนึงถึงหลักการใช้งานการค้นคืนเพื่อจัดทำสถิติทางการแพทย์จากสรุปรายงานการป่วย พ.ศ. 2557 จำนวน 2 รายงาน ดังนี้

1) รายงาน 10 ลำดับแรกจำนวนผู้ป่วยนอกตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร) พ.ศ.2557 (รายงานตัวอย่าง ในภาคผนวก ก)

2) รายงานจำนวนผู้ป่วยนอกรวมตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร) พ.ศ.2557 (รายงานตัวอย่าง ในภาคผนวก ก)

3.2.4 ทดสอบการประมวลผลด้วยชุดแบบสอบถาม

ทดสอบการประมวลผลด้วยชุดแบบสอบถามเอสคิวแอลและแมพรีดิว ขั้นตอนการทดสอบในระบบฐานข้อมูลเชิงสัมพันธ์ เป็นการจัดรูปแบบภาษาสอบถามข้อมูลเอสคิวแอล



ภาพที่ 3.8 ขั้นตอนการประมวลผลและปรับปรุงกระบวนการแบบสอบถามข้อมูล

สำหรับการทดสอบประมวลผล และดำเนินการปรับปรุงประสิทธิภาพการค้นคืนด้วยการปรับปรุงการเพิ่มดัชนี ซึ่งในการทดลองนี้การทำดัชนีคือการทำคีย์หลัก (Primary Key) ซึ่งคีย์

หลักถือเป็นข้อมูลที่เป็น Unique และ Nut Null และ Index รูปแบบหนึ่ง หลังจากนั้นทำการประมวลผลด้วยชุดแบบสอบถามที่จัดเตรียมไว้แล้วนำผลที่ได้เก็บลงในตารางบันทึกผลการทดลอง และทำการปรับปรุงการจัดชุดคำถามรูปแบบภาษาสอบถาม (Query Optimizer) เพื่อค้นหาชุดแบบสอบถามที่ใช้เวลาในการประมวลผลน้อยที่สุด เก็บข้อมูลผลที่ดีที่สุดที่ปรับปรุงแล้วจัดเก็บลงในตารางบันทึกผลการทดลอง

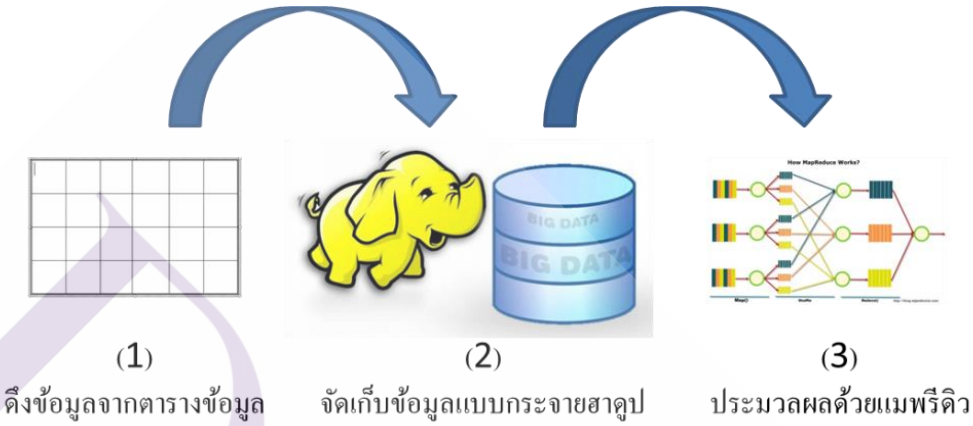
ขั้นตอนการทดสอบในระบบเทคโนโลยีข้อมูลขนาดใหญ่ ด้วยการนำชุดแบบสอบถามเอสคิวแอล ทำการปรับเปลี่ยนเป็นการจัดชุดแบบสอบถามรูปแบบคำสั่งตามภาษาโครงสร้างโปรแกรมฮาดูปและแมพรีดิว นำมาทำการประมวลผลด้วยชุดแบบสอบถามโปรแกรมแมพรีดิว ซึ่งชุดแบบสอบถามแมพรีดิวใช้ทดสอบกับไฟล์ที่จัดเก็บอยู่ในรูปแบบเท็กซ์ไฟล์ (Text File) หรือรูปแบบซีเอสวีไฟล์ (CSV) ตามข้อมูลตัวอย่างตารางที่ 3.3

ตารางที่ 3.3 ตัวอย่างชุดข้อมูลจากตาราง Diagnosis_opd ที่นำเข้าทดสอบ

HOSPCODE	PID	SEQ	DATE_SERV	DIAGTYPE	DIAGCODE	CLINIC	PROVIDER	D_UPDATE
00933	000003	636436	6/11/2014	1	Z123	00000	0004	6/11/2014 15:20
00933	000004	634919	19/11/2014	1	Z123	00000	0004	25/11/2014 18:21
00933	000005	634920	19/11/2014	1	Z123	00000	0004	25/11/2014 18:21
00934	022886	482116	22/6/2015	1	I10	00000	0009	11/7/2015 16:53
00934	022886	482527	28/6/2015	1	Z099	00000	0005	30/6/2015 14:56
00934	022886	483062	5/7/2015	1	Z251	00000	0017	5/7/2015 14:35
00934	022886	487831	14/9/2015	1	I10	00000	0005	8/10/2015 11:59
00941	062537	647821	13/4/2015	1	Z235	00000	0028	10/5/2015 17:57
00941	062537	647821	13/4/2015	4	Z236	00000	0028	10/5/2015 17:57
00941	062538	657548	11/5/2015	1	J069	00000	0006	11/5/2015 10:02
00941	062538	657548	11/5/2015	4	Z133	00000	0006	11/5/2015 10:02
00941	062539	639673	12/3/2015	1	Z001	00000	0005	28/3/2015 13:25
00937	027620	640775	21/2/2015	4	Z012	01100	0022	21/2/2015 17:18
00937	027620	640775	21/2/2015	4	Z133	01100	0010	21/2/2015 17:18
00937	027620	652230	11/5/2015	4	Z123	00000	0008	11/5/2015 16:54
00937	027620	652230	11/5/2015	4	Z133	00000	0008	11/5/2015 16:54

กระบวนการในขั้นตอนการทดสอบการทำงานของระบบการจัดเก็บข้อมูลแบบกระจายฮาดูปและการประมวลผลแบบขนานแมพรีดิวนั้น ผู้วิจัยเลือกใช้ขั้นตอนในการดึงข้อมูลออกจากตารางข้อมูลในฐานข้อมูลมายเอสคิวแอลโดยจะใช้เครื่องมือของการจัดการฐานข้อมูลเชิงสัมพันธ์ในระบบโปรแกรมมายเอสคิวแอลทำการส่งออกข้อมูลจากตารางที่ได้ทำการคัดกรองไว้เรียบร้อยแล้วออกเป็นไฟล์รูปแบบซีเอสวีไฟล์ (CSV) หลังจากนั้นทำการนำข้อมูลเข้าสู่ระบบเทคโนโลยี

ข้อมูลขนาดใหญ่ด้วยการนำเข้าในระบบจัดเก็บข้อมูลแบบกระจายหรือ HDFS ในระบบฮาดูป เป็นการนำส่งไฟล์ที่จัดเตรียมไว้เข้าระบบเพื่อรอการเรียกใช้งาน เมื่อดำเนินการจัดเตรียมข้อมูลในรูปแบบการกระจายเรียบร้อย จะดำเนินการทดสอบประมวลผลชุดแบบสอบถามที่จัดเตรียมไว้ด้วยเทคนิคแมพรีดิว ตามขั้นตอนดังภาพที่ 3.9



ภาพที่ 3.9 ขั้นตอนการประมวลผลด้วยเทคโนโลยีข้อมูลขนาดใหญ่

ตัวอย่างรูปภาพที่ 3.10 สามารถแสดงขั้นตอนการประมวลผลแมพรีดิวเพื่อค้นหารหัสสถานพยาบาลที่ให้บริการสุขภาพ ในตัวอย่างเป็นชุดข้อมูลจากตาราง Diagnosis_opd เพื่อค้นหาจำนวนสถานพยาบาลที่มีรหัส 00933 ในตารางข้อมูล โดยมีขั้นตอน Input เพื่อนำข้อมูลเข้าโปรแกรมจะทำการจับ Key/Value ตามที่ได้ออกแบบไว้ Key = ชื่อแฟ้ม และ Value = ชุดข้อมูลในระเบียน แล้วทำการแปลงข้อมูลให้อยู่ในรูปแบบ Key/Value ที่สามารถใช้ในขั้นต่อไปได้ ขั้นตอน Map จะทำการจับคู่สิ่งที่ต้องการค้นหาโดยการป้อนรหัสสถานพยาบาลที่ต้องการนับ ตัวอย่างเช่น (Diagnosis_opd, 00933) หลังจากนั้นจะทำ Shuffle & Sort ใหม่หรือการจัดเรียงตามกลุ่มที่กำหนดตามกระบวนการทำงานของโปรแกรมเพื่อส่งไปยังขั้นตอนต่อไป ขั้นตอนการ Reduce จะทำการนำข้อมูลที่ได้ทำการจับคู่ Key/Value ใหม่ แล้วนำมาทำการนับจำนวนรหัสสถานพยาบาลที่ต้องการแล้วทำการแสดงผลบนหน้าจอ

ภาพที่ 3.10 ขั้นตอนการประมวลผลชุดข้อมูลตัวอย่างด้วยเทคนิคแมพรีดิว

3.2.5 บันทึกผลลัพธ์จากการประมวลผล

การบันทึกผลจากการทดลองผู้วิจัยมีแนวคิดการดำเนินการจัดเก็บผลการทดลองใน 3 ขั้นตอนการทดลองดังนี้

1) ขั้นตอนการเตรียมข้อมูล เป็นขั้นตอนการนำไฟล์ข้อมูลที่มีการเก็บรวบรวมไว้เป็นรูปแบบเท็กซ์ไฟล์จากกระทรวงสาธารณสุขนำมาคัดเลือกสำหรับการทดลอง

2) ขั้นตอนการนำเข้าข้อมูล เป็นขั้นตอนการคัดกรองข้อมูลตามจำนวนกลุ่มของข้อมูลที่มีการขยายตัวอย่างเป็นลำดับ ตามที่กำหนดไว้สำหรับใช้ในการทดลอง

3) ขั้นตอนการประมวลผล เป็นขั้นตอนที่ประมวลผลด้วยชุดแบบสอบถามที่จัดเตรียมไว้จากรายงานการเจ็บป่วย และทำการประมวลผลกับเทคโนโลยีข้อมูล 2 รูปแบบ

การเก็บบันทึกผลการทดลองเพื่อนำใช้ในการวิเคราะห์ข้อมูลด้วยสถิติ ด้วยการจัดเก็บผลลัพธ์ของเวลาในการประมวลผลหลังจากดำเนินการปรับปรุงกระบวนการประมวลผลแล้วจำนวน 3 ครั้ง ทำการจัดเก็บข้อมูลผลการทดลองตามกลุ่มจำนวนระเบียน ด้วยการประมวลผลกับชุดแบบสอบถาม จากรายงานการเจ็บป่วย พ.ศ.2557 มี 2 กลุ่ม ดังนี้

กลุ่ม 1 การประมวลผลด้วยเทคโนโลยีระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ด้วยภาษาสอบถามเชิงโครงสร้างเอสคิวแอล ทำการบันทึกผลการทดลองหลังจากมีการปรับปรุงกระบวนการสอบถามข้อมูล จำนวน 3 ครั้ง ตามกลุ่มจำนวนระเบียน บันทึกผลเป็นหน่วยของเวลาวินาที

กลุ่ม 2 การประมวลผลเทคโนโลยีระบบข้อมูลขนาดใหญ่ด้วยเทคนิคแมพรีดิว ทำการบันทึกผลการทดลองหลังจากมีการปรับปรุงกระบวนการสอบถามข้อมูล จำนวน 3 ครั้ง ตามกลุ่มจำนวนระเบียน บันทึกผลเป็นหน่วยของเวลาวินาที

3.2.6 นำผลลัพธ์ที่ได้นำมาวิเคราะห์สถิติเปรียบเทียบการประมวลผล

หลังจากทำการประมวลผลกับเทคโนโลยีข้อมูล 2 รูปแบบ การนำผลลัพธ์ที่ได้นำมาวิเคราะห์ทางสถิติเพื่อเปรียบเทียบผลลัพธ์จากการประมวลผล มีวัตถุประสงค์เพื่อนำผลลัพธ์มาอภิปรายผลทางสถิติ เลือกใช้สถิติเชิงพรรณนา (Descriptive Statistics) ใช้เพื่อค้นหาคำตอบจาก

ผลลัพธ์ที่ได้จากขั้นตอนการเตรียมข้อมูล และนำเข้าข้อมูล และจากการประมวลผล ประกอบด้วย ค่าเฉลี่ย (Mean) ร้อยละ (เปอร์เซ็นต์) และทำการแสดงผลด้วยกราฟหรือแผนภูมิ และสถิติเชิงอนุมาน (Inferential Statistics) ใช้เพื่อหาคำตอบจากสมมติฐานที่ได้คาดการณ์ไว้ล่วงหน้า

การวิเคราะห์สถิติ ด้วยวิธีการวิเคราะห์ทางสถิติ t-Test Paired Two Sample for Means ด้วยการใช้โปรแกรม Excel มาตรการวัด (Measurement Scale) คือ มาตราอัตราส่วน (Ratio Scale) เป็นข้อมูลเวลาที่ใช้ในการประมวลผล มีการกำหนดให้ค่านัยสำคัญที่ $\alpha = 0.05$ ที่จะนำมาใช้ทดสอบสมมติฐาน ซึ่งเป็นการกำหนดความน่าจะเป็นที่ผู้วิจัยจะยอมให้เกิดความคลาดเคลื่อนประเภทที่ 1 (α) จากการปฏิเสธสมมติฐานหลักที่เป็นจริง

สมมติฐานที่ผู้วิจัยได้สันนิษฐานไว้ สามารถแสดงได้ดังนี้

1) ผลลัพธ์ของเวลาการค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างเทคโนโลยีข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์ มีผลลัพธ์ที่แตกต่างกัน

ทดสอบจากค่าเฉลี่ยของกลุ่มข้อมูลผลของเวลา $H = \mu 1 \neq \mu 2$

ค่าสมมติฐานที่ถูกพิสูจน์ด้วยสถิติ t-Test Paired Two Sample for Means จากข้อตกลงเบื้องต้น คือใช้สำหรับการทดสอบค่าเฉลี่ยของ 2 กลุ่ม เพื่อวิเคราะห์ความแตกต่างของประชากรมีหรือไม่ และมีนัยสำคัญหรือไม่ สามารถใช้กับข้อมูลมาตราอัตราส่วน และใช้กับข้อมูลจำนวนน้อย และเพื่อใช้กับกลุ่มข้อมูลจำนวน 2 กลุ่ม

2) ผลลัพธ์ของความแม่นยำถูกต้องการค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างระบบข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์ มีผลลัพธ์ไม่แตกต่างกัน

มีการเปรียบเทียบผลการทดลองจากข้อมูลโดยการให้ค่าอัตราเปอร์เซ็นต์ความถูกต้องเปรียบเทียบกันระหว่าง ผลลัพธ์ความถูกต้องของการประมวลผลด้วยเทคโนโลยีระบบฐานข้อมูลมายเอสคิวแอลและระบบเทคโนโลยีข้อมูลขนาดใหญ่ฮาดูปและแมพรีดิวมีผลลัพธ์ที่ถูกต้องตรงกันมีความแม่นยำเหมือนกัน เทคโนโลยีข้อมูลขนาดใหญ่นี้สามารถนำมาใช้งานทดแทนกันได้หรือไม่

3.2.7 สรุปผลที่ได้จากการวิเคราะห์สถิติ

สรุปผลที่ได้จากการวิเคราะห์ ผลลัพธ์ที่จะได้จากการทดลองสามารถแสดงผลได้หลากหลายรูปแบบ ผู้วิจัยมีแนวคิดการเลือกการสรุปผลออกเป็น 3 ส่วนคือ

1) ผลลัพธ์ที่ได้จากการวิเคราะห์ผลด้วยสถิติเชิงพรรณนา

2) ผลลัพธ์ที่ได้จากการวิเคราะห์ผลด้วยสถิติอนุมานที่จะใช้พิสูจน์ผลจากการตั้งสมมติฐานว่าเป็นจริงตามที่สันนิษฐานไว้หรือไม่

3) นำผลที่ได้จากการการวิเคราะห์ข้อมูลเชิงสถิติพรรณนานำไปเปรียบเทียบกับงานวิจัยที่เกี่ยวข้อง

3.3 เครื่องมือดำเนินงานวิจัย

3.3.1 ฮาร์ดแวร์ (Hardware)

เครื่องเซิร์ฟเวอร์ จำนวน 3 เครื่อง สำหรับใช้เป็นเครื่องมื่อในการทดลอง กำหนดให้เป็นเครื่องแม่ (Master) 1 เครื่อง และเครื่องลูก (Slave) จำนวน 2 เครื่อง ซึ่งมีคุณสมบัติเหมือนกันดังนี้

1) CPU Intel® XEON 2.4 GHz
2) RAM 8 GB DDR3
3) HDD 2 TB

3.3.2 ซอฟต์แวร์ (Software)

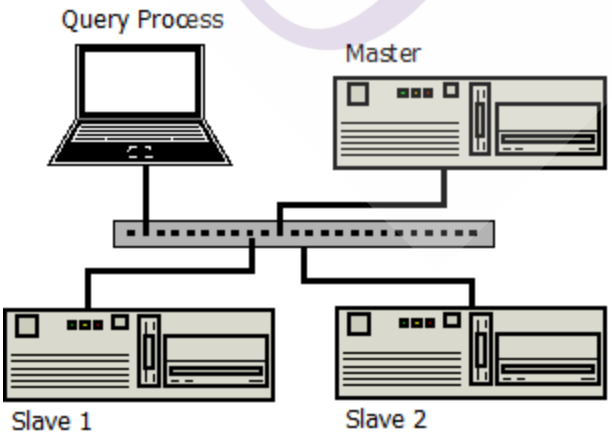
ที่ใช้ในงานวิจัย มีรายละเอียดดังนี้

1) โปรแกรม EmEditor เวอร์ชั่น 15.8.1
2) โปรแกรม Ubuntu Server 14.04.4 LTS
3) โปรแกรม phpMyAdmin 5.5.9
4) โปรแกรม MySQL 5.5.47
5) โปรแกรม Apache Hadoop 2.7.2
6) โปรแกรม Eclipse Standard/SDK (Kepler Service Release 2)
7) โปรแกรม Microsoft Excel Professional Plus 2010

3.3.3 เครือข่าย (Network)

การเชื่อมต่อใช้รูปแบบการเชื่อมต่อแบบสตาร์ (Star) หรือการเชื่อมโยงคอมพิวเตอร์ทั้งหมดเข้าด้วยผ่านอุปกรณ์เครือข่ายสวิตซ์ (Switch)

1) Ethernet Switch 24 Port 10/100/1000 Gigabitgh6h



ภาพที่ 3.11 รูปแบบเครือข่ายคอมพิวเตอร์ที่ใช้ในงานวิจัย

3.4 สถานที่ทำงานวิจัย

สถานที่ทดลองที่ใช้ในงานวิจัย ห้องทดลองข้อมูลขนาดใหญ่ คณะวิศวกรรมข้อมูลขนาดใหญ่ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

บทที่ 4
ผลงานวิจัย และสรุปผลงานวิจัย

ผลการวิจัยจากการทดลองเพื่อการศึกษาและทำความเข้าใจในเทคโนโลยีข้อมูลขนาดใหญ่ที่มีรูปแบบการจัดเก็บแบบกระจายฮาดูป และการประมวลผลแบบขนานแมพรีดิว ด้วยการนำชุดข้อมูลในระบบบริการสุขภาพ เป็นเครื่องมือที่นำมาใช้หาประสิทธิภาพของเวลาในการประมวลผลการค้นคืนข้อมูล ประเมินประสิทธิภาพความเร็วด้วยการวิเคราะห์ผลโดยใช้สถิติ t-Test ทดสอบสมมติฐานที่คาดการณ์ไว้ล่วงหน้า มีรายงานผลงานวิจัยดังนี้

4.1 ผลการเตรียมข้อมูลชุดทดสอบ

ผลการเตรียมข้อมูลชุดทดสอบ ทำการคัดเลือกข้อมูลด้วยเทคนิควิธีการสุ่มแบบเฉพาะเจาะจง การคัดเลือกกลุ่มตัวอย่างแบบเจาะจงเป็นการเลือกกลุ่มตัวอย่างโดยอาศัยการตัดสินใจโดยอาศัยเกณฑ์การตัดสินใจจากผู้เชี่ยวชาญข้อมูลในระบบบริการสุขภาพ และทำการสุ่มคัดเลือกกลุ่มข้อมูลชุดตัวอย่างจากแฟ้มมาตรฐาน 43 แฟ้ม คัดเลือกแฟ้มผู้ป่วยนอกเป็นข้อมูลชุดตัวอย่าง และทำการสุ่มคัดเลือกนำตัวอย่างกลุ่มข้อมูลทดสอบเพียงบางส่วนตามจำนวนที่ผู้วิจัยต้องการทดสอบ โดยการคัดเลือกกลุ่มตัวอย่าง 3 จังหวัด จากทั้งสิ้น 76 จังหวัด และทำขั้นตอนการทำความสะอาดคัดกรองข้อมูลที่มีอักขระผิดพลาดหรือแถวข้อมูลที่มีคอลัมน์เกินหรือขาดตัดออก ได้ขนาดระเบียนดังตารางที่ 4.1

ตารางที่ 4.1 การคัดกรองคัดเลือกข้อมูลชุดทดสอบแฟ้ม Diagnosis_opd

	แฟ้ม Diagnosis_opd							
ชื่อแฟ้ม	จำนวนระเบียน ก่อนคัดกรอง	ขนาดไฟล์ (MB)	จำนวนระเบียน หลังคัดกรอง	ขนาดไฟล์ (MB)	จำนวนระเบียน ที่ถูกตัดออก	ขนาดไฟล์ ลดลง (MB)	เปอร์เซ็นต์ ถูกตัดออก	เวลาที่ใช้ (วินาที)
diagnosis_opd_1.txt	899,972	65.2	899,294	65.1	678	0.1	0.0753%	2,880.00
diagnosis_opd_2.txt	5,558,268	374.0	5,555,064	373.0	3,204	1.0	0.0576%	12,960.00
diagnosis_opd_3.txt	6,308,357	454.0	6,304,556	442.0	3,801	12.0	0.0603%	18,720.00
รวม	12,766,597	893.2	12,758,914	880.1	7,683	13.1	0.0602%	34,560.00

ตารางที่ 4.2 แบ่งการคัดกรองชุดทดสอบ 4 ชุด เข้าระบบฐานข้อมูลแฟ้ม Diagnosis_opd

แฟ้ม Diagnosis_opd

ชื่อแฟ้ม	จำนวนระเบียน ก่อนนำเข้าฐานข้อมูล	ขนาดไฟล์ (MB)	จำนวนระเบียน หลังนำเข้าฐานข้อมูล	ขนาดไฟล์ (MB)	จำนวนระเบียน ที่ถูกตัดออก	ขนาดไฟล์ เพิ่มขึ้น (MB)	เปอร์เซ็นต์ ถูกตัดออก	เวลาที่ใช้ (วินาที)
diagnosis_opd_1.txt	899,294	65.1	899,294	99.6	0	34.5	0.0000%	25.84
diagnosis_opd_2.txt	5,555,064	373.0	5,555,064	613.0	0	240.0	0.0000%	166.46
diagnosis_opd_3.txt	6,304,556	442.0	6,304,556	696.0	0	254.0	0.0000%	182.67
รวม3 แฟ้ม	12,758,914	880.1	12,758,914	1,408.6	0	528.5	0.0000%	374.97
diagnosis_opd_all.txt	12,758,914	881.0	12,758,914	1,406.0	0	525.0	0.0000%	375.18

การดำเนินการคัดกรองข้อมูล เมื่อทำการคัดเลือกข้อมูลเรียบร้อยแล้ว และมีการนำเข้าข้อมูลลงสู่ระบบการจัดการฐานข้อมูลเชิงสัมพันธ์มายเอสคิวแอลพร้อมกับบันทึกเวลาที่ใช้ในการนำเข้าเป็นวินาที และบันทึกจำนวนระเบียนที่สูญเสียไปกับการปฏิเสธจากระบบการจัดการฐานข้อมูลมายเอสคิวแอล ยกตัวอย่าง เช่น มีค่าซ้ำกันของคีย์หลัก (Primary Key) เป็นต้น จึงทำให้จำนวนระเบียนลดลงแต่จะไม่กระทบต่อเป้าหมายของการทดสอบหรือจำนวนไม่น้อยกว่าข้อมูลที่กำหนดไว้สำหรับการทดลอง

4.2 ผลการสุ่มคัดเลือกข้อมูลชุดทดสอบตามกลุ่มที่กำหนด

การบันทึกผลในขั้นตอนที่ 4.2 การสุ่มคัดเลือกข้อมูลชุดทดสอบตามกลุ่มที่กำหนด มีการเก็บบันทึกเวลาที่ใช้ในการสุ่มคัดเลือกข้อมูลชุดทดสอบแล้วนำเข้าฐานระบบการจัดการฐานข้อมูลมายเอสคิวแอลใหม่อีกครั้ง เพื่อตรวจสอบเวลาที่ใช้กับจำนวนระเบียนโดยเฉลี่ย

เมื่อทำการสุ่มข้อมูลเข้าระบบการจัดการฐานข้อมูลมายเอสคิวแอลเป็นที่เรียบร้อยแล้ว จะทำการนำข้อมูลดังกล่าวในแต่ละชุดข้อมูลออกมาในรูปแบบไฟล์ซีเอสวี (CSV) เพื่อทำการนำข้อมูลเข้าในระบบเทคโนโลยีข้อมูลขนาดใหญ่ฮาดูปหรือ HDFS และทำการเปรียบเทียบด้วยการจัดทำตาราง

สุ่มคัดเลือกข้อมูลชุดทดสอบตามกลุ่มที่กำหนด การกำหนดขนาดตัวอย่างโดยการใช้วิธีการเลือกตัวอย่างแบบกำหนดจำนวนหรือแบบโควต้า ทำการคัดเลือกสุ่มข้อมูลชุดทดสอบตามจำนวนที่กำหนด หรือการเลือกแบบง่ายหรือแบบสะดวก ใช้การสุ่มอย่างง่ายด้วยระบบคอมพิวเตอร์เข้าระบบฐานข้อมูลมายเอสคิวแอลในตารางแฟ้มข้อมูลที่กำหนดให้มีโครงสร้างตามมาตรฐานในระบบบริการสุขภาพ โดยใช้ฟังก์ชั่น Rand() ในระบบฐานข้อมูล เป็นการสุ่มระเบียนออกจากฐานข้อมูลนำมาเก็บแยกเป็นตารางข้อมูลชุดทดสอบออกเป็น 4 ชุด การคัดเลือกจำนวนในชุดดังนี้

ชุดที่ 1 มีจำนวน 500,000 ระเบียน ชุดที่ 2 มีจำนวน 1,000,000 ระเบียน ชุดที่ 3 มีจำนวน 5,000,000 ระเบียน ชุดที่ 4 มีจำนวน 10,000,000 ระเบียน ตามลำดับ แบ่งข้อมูลออกเป็น 4 ชุด หลังจากที่ได้ข้อมูลทั้งหมดจะถูกนำข้อมูลออกเป็นไฟล์ CSV และทำการนำเข้าระบบ HDFS ดังตารางที่ 4.3

ตารางที่ 4.3 นำเข้าข้อมูลชุดทดสอบ 4 ชุด แฟ้ม Diagnosis_opd

แฟ้ม Diagnosis_opd

		CSV	Import MySQL		Import Hadoop	
ชื่อแฟ้ม	จำนวนระเบียนก่อนนำเข้าฐานข้อมูล	ขนาดไฟล์ (MB)	ขนาดไฟล์ (MB)	เวลาที่ใช้ (วินาที)	ขนาดไฟล์ (MB)	เวลาที่ใช้ (วินาที)
diagnosis_opd_5h	500,000	34.1	55.6	3.01	34.12	0.17
diagnosis_opd_1m	1,000,000	68.2	110.6	6.13	68.23	0.34
diagnosis_opd_5m	5,000,000	341.0	552.0	30.37	341.16	1.07
diagnosis_opd_10m	10,000,000	682.0	1,102.0	58.56	682.34	2.12
รวม 4 แฟ้ม	16,500,000	1,125.3	1,820.2	98.07	1,125.85	3.70

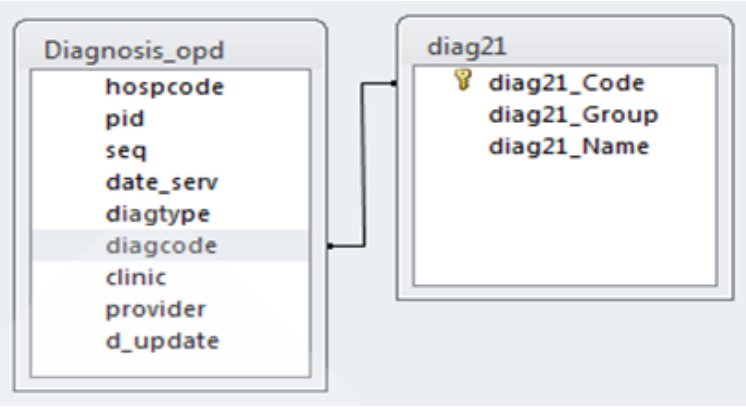
4.3 ผลการเตรียมคำสั่งชุดทดสอบเอสคิวแอลและแมพรีดิว

เตรียมชุดคำถามทดสอบการประมวลผลด้วยภาษาสอบถามข้อมูลเอสคิวแอลและการประมวลผลด้วยเทคนิคแมพรีดิว ด้วยการใช้ชุดคำถามที่คำนึงถึงหลักการใช้งานการค้นคืนเพื่อจัดทำสถิติทางการแพทย์จากสรุปรายงานการป่วย พ.ศ. 2557 จำนวน 2 รายงาน

1) รายงาน 10 ลำดับแรกจำนวนผู้ป่วยนอกตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร) พ.ศ.2557 (รายงานตัวอย่าง ในภาคผนวก ก)

2) รายงานจำนวนผู้ป่วยนอกรวมตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร) พ.ศ.2557 (รายงานตัวอย่างอยู่ ในภาคผนวก ก)

จากการเตรียมชุดแบบสอบถามตามขั้นตอนการทดสอบข้อ 4.3 พบว่าต้องการรายการ 21 กลุ่มโรคหลัก นำมาใช้ในการประมวลผลร่วมกับข้อมูลผู้ป่วยนอก แฟ้ม 21 กลุ่มโรค กำหนดให้มีคีย์หลักเพื่อควบคุมค่าซ้ำกันและทำการเชื่อมความสัมพันธ์เพื่อจัดทำรายงาน ดังรูปที่ 4.1



ภาพที่ 4.1 การเชื่อมโยงความสัมพันธ์ระหว่างแฟ้มรหัสกลุ่มโรคและแฟ้มผู้ป่วยนอก

และในการกำหนดการเชื่อมโยงตารางข้อมูแฟ้มรหัสกลุ่มโรคกับแฟ้มผู้ป่วยนอก ซึ่งมีจำนวนข้อมูลในการนำเข้าทั้งสิ้น 2,136 ระเบียน ตัวอย่างสามารถแสดงได้ดังตารางที่ 4.4 และได้ดำเนินการสร้างแฟ้มโครงสร้างในระบบฐานข้อมูลมายเอสคิวแอล ดังตารางที่ 4.5

ตารางที่ 4.4 ตัวอย่างข้อมูลในแฟ้มรหัสกลุ่มโรค จำนวน 2,136 ระเบียน

รหัส ICD-10	ชื่อกลุ่มโรค	ชื่อโรค
A00	1	โรคติดเชื้อและปรสิต
A01	1	โรคติดเชื้อและปรสิต
B00	1	โรคติดเชื้อและปรสิต
B01	1	โรคติดเชื้อและปรสิต
C00	2	เนื้องอก
C01	2	เนื้องอก
D00	2	เนื้องอก
D01	2	เนื้องอก
D50	3	โรคเลือดและอวัยวะสร้างเลือด และความผิดปกติเกี่ยวกับภูมิคุ้มกัน
D51	3	โรคเลือดและอวัยวะสร้างเลือด และความผิดปกติเกี่ยวกับภูมิคุ้มกัน
D52	3	โรคเลือดและอวัยวะสร้างเลือด และความผิดปกติเกี่ยวกับภูมิคุ้มกัน
.	.	.....
.	.	.....
.	.	.....
Y88	21	สาเหตภายนอกอื่น ๆ ที่ทำให้ป่วยหรือตาย
Y89	21	สาเหตภายนอกอื่น ๆ ที่ทำให้ป่วยหรือตาย

ตารางที่ 4.5 โครงสร้างตารางแฟ้ม 21 กลุ่มโรค

No	Field	Caption	Primary	Type Data	Not Null
1	Diag21_Code	รหัสการวินิจฉัย	Y	CHAR(6)	Y
2	Diag21_Group	รหัสกลุ่มโรค		CHAR(20)	Y
3	Diag21_Name	ชื่อโรค		CHAR(100)	Y

ในการประมวลในระบบฐานข้อมูลมายเอสคิวแอล ได้ดำเนินการจัดทำชุดคำสั่งสำหรับเตรียมการทดลองการประมวลผลแบบสอบถามข้อมูล รายงานที่ 1 และ รายงานที่ 2 ดังนี้

ชุดคำสั่งการนำเข้าข้อมูลเพื่อจัดเก็บเข้าฐานข้อมูลมายเอสคิวแอลตามจำนวนที่กำหนด ตัวอย่าง ชุดคำสั่งในการนำเข้าข้อมูลจัดเก็บเข้าฐานข้อมูลมายเอสคิวแอล ด้วยภาษาเอสคิวแอล

Insert into diagnosis_opd_5h (hospcode, pid, seq, date_serv, diagtype, diagcode, clinic, provider, d_update) Select * from diagnosis_opd_all order by rand() limit 500000;

ชุดคำสั่งการเคลียร์ค่าหน่วยความจำสำรองในระบบฐานข้อมูลมายเอสคิวแอล เพื่อทำการล้างค่าข้อมูลจากระบบในการเริ่มต้น เพื่อเก็บผลการทดลองใหม่ให้ครบตามจำนวน 3 ครั้ง ดังนี้

RESET QUERY CACHE;

FLUSH QUERY CACHE;

FLUSH TABLES;

ชุดคำสั่งสำหรับการประมวลผลรายงานที่ 1 ในการประมวลผลด้วยภาษาสอบถามแบบมีโครงสร้างเอสคิวแอล ในระบบฐานข้อมูลเชิงสัมพันธ์ ตัวอย่าง ชุดคำสั่งสำหรับการประมวลผลการค้นคืนข้อมูลด้วยภาษาสอบถามเอสคิวแอล รายงานที่ 1 และรายงานที่ 2 ตามลำดับ

ตัวอย่างภาษาสอบถามแบบมีโครงสร้างเอสคิวแอล รายงานที่ 1

SELECT opd_diag.opd_diag3 AS กลุ่มโรค, opd_diag.opd_diag4 AS สาเหตุกลุ่มการป่วย, COUNT(opd_diag.opd_diag3) AS จำนวนผู้ป่วยนอก FROM `diagnosis_opd_10m`, `opd_diag` WHERE LEFT(diagnosis_opd_10m.diagcode, 3) = opd_diag.opd_diag1 GROUP BY opd_diag.opd_diag3 ORDER BY จำนวนผู้ป่วยนอก DESC

ตัวอย่างภาษาสอบถามแบบมีโครงสร้างเอสคิวแอล รายงานที่ 2

SELECT opd_diag.opd_diag3 AS กลุ่มโรค, opd_diag.opd_diag4 AS สาเหตุกลุ่มการป่วย, COUNT(opd_diag.opd_diag3) AS จำนวนผู้ป่วยนอก FROM `diagnosis_opd_10m`,

`opd_diag` WHERE LEFT(diagnosis_opd_10m.diagcode, 3) = opd_diag.opd_diag1 GROUP BY opd_diag.opd_diag3 ORDER BY จำนวนผู้ป่วยนอก DESC LIMIT 10

ลักษณะการ Join ตาราง มีหลายรูปแบบตามทฤษฎี มี 6 ลักษณะ ดังนี้ 1) Left Join 2) Inner Join 3) Full outer join 4) Right Join 5) Left Join (If Null) 6) Right Join (If Null) ซึ่งในการทดลองนี้ใช้คุณลักษณะการ Join แบบที่ 1 Left Join

ชุดคำสั่งสำหรับการนำเข้าข้อมูลในการจัดเก็บในระบบไฟล์ข้อมูล HDFS เพื่อทำการจัดเก็บแบบกระจายไปยังแหล่งจัดเก็บในเครื่องเซิร์ฟเวอร์ในกลุ่มคลัสเตอร์ ดังมีตัวอย่าง ชุดคำสั่งในการนำเข้าข้อมูลจัดเก็บในระบบเทคโนโลยีข้อมูลขนาดใหญ่ฮาดูปมีดังนี้

```
hdfs dfs -put /home/node1/Documents/testhelloworld.txt /user
```

ชุดคำสั่งสำหรับการประมวลผลแบบขนาน ด้วยเทคนิคแมพรีดิวเป็นขั้นตอนการดำเนินการในขั้นตอนการประมวลผลข้อมูลจำนวน 3 ครั้ง ตัวอย่างชุดคำสั่งมีดังนี้

```
hadoop jar report2-0.1-SNAPSHOT-jar-with-dependencies.jar -1 healthcare/data/
diag_opd/ healthcare/data/test/keng/opd10m/ healthcare/data/output/keng/report210m
```

ชุดคำสั่งของภาษาจาวา สำหรับการประมวลผลข้อมูลออกมาเป็นรูปแบบ CSV ไฟล์ต้องนำผลลัพธ์โดยการใช้งานคำสั่งดังนี้ ตัวอย่างชุดคำสั่งการประมวลผลและทำการค้นหาผลลัพธ์และนำผลลัพธ์เข้าจัดเก็บในระบบ HDFS ดังมีตัวอย่างซอร์สโค๊ดดังนี้

```
public class Report2 {
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        String[] otherArgs = new GenericOptionsParser(conf,
        args).getRemainingArgs();
        if (otherArgs.length != 4) {
          System.err.println("Command need 4 parameters but your input are "
           + otherArgs.length);
          System.err.println("Usage:  \n\t[1.]n - Top n (-1 for all) \n" + "\t[2.]
           Master Diagnosis OPD Input Path \n"
           + "\t[3.] Diagnosis OPD Input Path\n" + "\t[4.] Output Path\n");
          System.exit(2);
        }
    System.out.println("-1".equals(args[0]) ? "All records" : ("Top " + args[0]));
```

```
        conf.set(AppConstant.MAX_KEY, args[0]);
        conf.set(AppConstant.INPUT_SEPERATOR_KEY, "\\|");
    //* * Job 1 - */
        Job job1 = Job.getInstance(conf, "Report2 - Joining OPD and Diagnosis");
        job1.setJarByClass(Report2.class);
        job1.setMapOutputValueClass(Text.class);
        job1.setReducerClass(FinalReducer.class);
        job1.setOutputKeyClass(Text.class);
        job1.setOutputValueClass(IntWritable.class);
    // Master Diagnosis
        MultipleInputs.addInputPath(job1, new Path(args[1]), TextInputFormat.class,
        DiagnosisMasterMapper.class);
        // Diagnosis OPD
        MultipleInputs.addInputPath(job1, new Path(args[2]), TextInputFormat.class,
        DiagnosisOPDMapper.class);
        FileOutputFormat.setOutputPath(job1, new Path(args[3] + "/joined"));
        job1.waitForCompletion(true);
    //* * Job 2 - */
        Job job2 = Job.getInstance(conf, "Report2 - Counting");
        job2.setJarByClass(Report2.class);
        job2.setMapperClass(JoinedDataMapper.class);
        job2.setCombinerClass(TopNReducer.class);
        job2.setReducerClass(TopNReducer.class);
        job2.setOutputKeyClass(Text.class);
        job2.setOutputValueClass(IntWritable.class);
        FileInputFormat.addInputPath(job2, new Path(args[3] + "/joined"));
        FileOutputFormat.setOutputPath(job2, new Path(args[3] + "/final"));
        System.exit(job2.waitForCompletion(true) ? 0 : 1);
    }
}
```

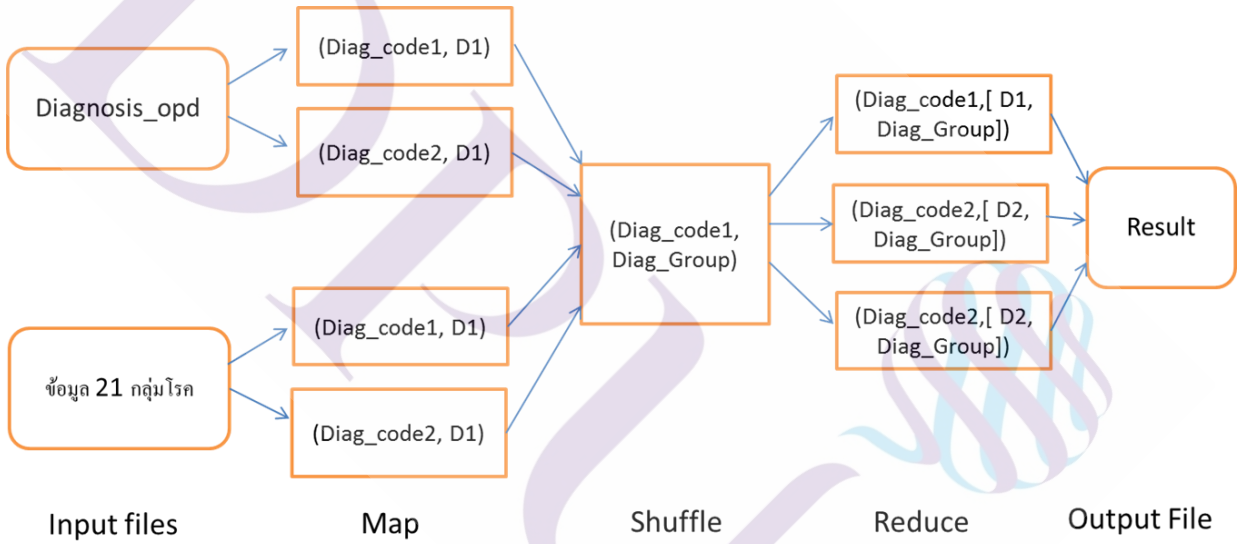
ชุดคำสั่งภาษาจาวาดังกล่าวมีการทำงานใน 2 ขั้นตอน ดังนี้

ขั้นตอนงานที่ 1 คือ การ Join Data ในกระบวนการย่อยของการ Join Data มี 3 ขั้นตอน ดังนี้

ขั้นตอนที่ 1.1 Map phase คือการทำ Map ฟังก์ชั่น ทำการจับคู่คีย์ รหัสกลุ่มโรคในแต่ละระเบียนเพื่อการเรียกคืนข้อมูล ตัวอย่างคู่คีย์ (Diag_Code, D1) เป็นต้น ซึ่งจะดำเนินการในขั้นตอนพร้อมกัน 2 ไฟล์ CSV คือ ไฟล์ Diagnosis_opd และข้อมูลรหัสกลุ่มโรค

ขั้นตอนที่ 1.2 Shuffle phase คือ การจัดทำการรวบรวม Grouping และการจัดเรียงลำดับ Sorting ใหม่ ตัวอย่างคู่คีย์ (Diag_Code, Diag_Group)

ขั้นตอนที่ 1.3 Reduce phase คือ การลดจำนวนคู่คีย์เพื่อรวบรวมเป็นผลลัพธ์ Text File สำหรับการเรียกใช้ในขั้นตอนที่ 2 Counting ตัวอย่างคู่คีย์ (Diag_Code, [D1, Diag_Group])



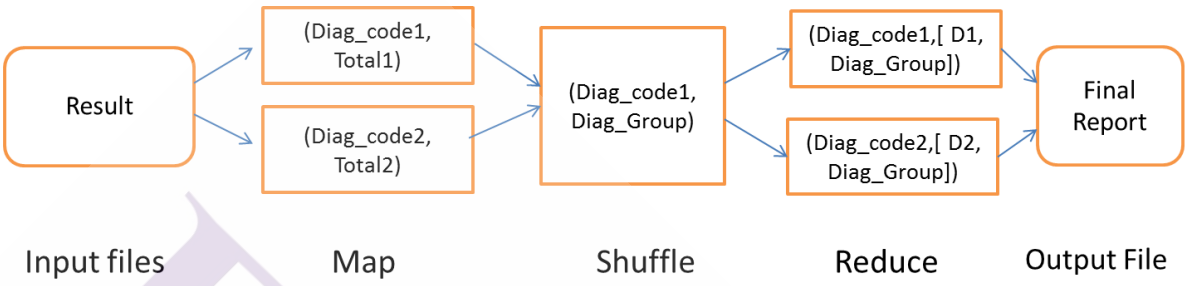
ภาพที่ 4.2 ขั้นตอนการ Join Data โปรแกรมแมพรีดิว

ขั้นตอนงานที่ 2 คือ การ Counting และSorting Data ในกระบวนการย่อยของการ Counting และ Sorting Data มี 3 ขั้นตอนดังนี้

ขั้นตอนที่ 1.1 Map phase คือการทำ Map ฟังก์ชั่น ทำการจับคู่คีย์ รหัสกลุ่มโรคในแต่ละระเบียนจากไฟล์ผลลัพธ์ที่ได้จากขั้นตอนที่ 1 เพื่อการเรียกคืนค่าข้อมูล ตัวอย่างคู่คีย์ (Diag_Code, Total) เป็นต้น

ขั้นตอนที่ 1.2 Shuffle phase คือ การจัดทำการรวบรวม Grouping และการจัดเรียงลำดับ Sorting ใหม่ ตัวอย่างคู่คีย์ (Diag_Code, Diag_Group)

ขั้นตอนที่ 1.3 Reduce phase คือ การลดจำนวนคู่คีย์เพื่อรวบรวมเป็นผลลัพธ์เท็กซ์ไฟล์ Text File เป็นผลการค้นคืนทั้งหมดในแต่ละรายงาน ตัวอย่างคู่คีย์ (Diag_Code, [D1, Diag_Group])

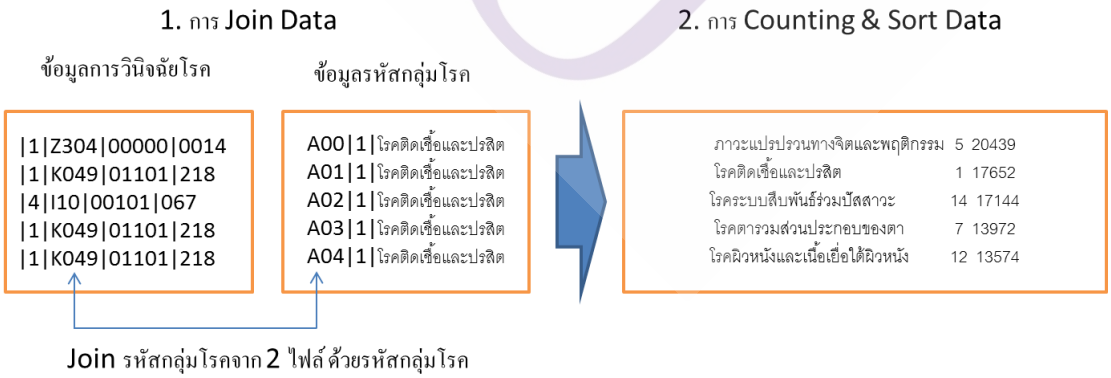


ภาพที่ 4.3 ขั้นตอนการ Counting และ Sorting Data โปรแกรมแมพรีดิว

ชุดคำสั่งของเทคโนโลยีฮาดูปและแมพรีดิว สำหรับการนำผลลัพธ์ในการประมวลผลข้อมูลออกโดยการใช้งานคำสั่งดังนี้ ตัวอย่างชุดคำสั่งการนำผลลัพธ์ที่ถูกจัดเก็บไว้ใน HDFS นำออกเป็นรูปแบบเท็กซ์ไฟล์ (Text File) และนำเข้าในโปรแกรมเอ็กเซล (Excel) เพื่อจัดเรียงข้อมูล

hadoop fs -copyToLocal /user/hadoopuser/healthcare/data/person/person_2015_17.txt person_2015_17.txt

สรุปขั้นตอนในภาพรวมของวิธีการแมพและรีดิว ดังนี้



ภาพที่ 4.4 ขั้นตอนการค้นคืนข้อมูลด้วยโปรแกรมแมพรีดิว

4.4 ผลการทดลองการประมวลผลด้วยชุดแบบสอบถามเอสคิวแอลและแมพรีดิว

นำรายงานและเครื่องมือที่ได้จัดเตรียมไว้ในข้อ 4.3 นำมาทำการประมวลผลใหม่ สร้างเป็นชุดแบบสอบถามด้วยรูปแบบภาษาสอบถามเชิงโครงสร้างเอสคิวแอล ทำการประมวลผลกับชุดข้อมูลทดสอบในข้อที่ 2 จำนวน 3 ครั้งต่อ 1 รายงาน ต่อ 1 ชุดข้อมูลในระบบฐานข้อมูลเชิงสัมพันธ์ และสร้างชุดแบบสอบถามด้วยการเขียนโปรแกรมแมพรีดิวใช้ทดสอบกับไฟล์ข้อมูลชุดทดสอบที่จัดเก็บอยู่ในรูปแบบเท็กซ์ไฟล์นำเข้าในโปรแกรมฮาดูปจัดเก็บแบบกระจาย HDFS ทำการประมวลผลจำนวน 3 ครั้งต่อ 1 ชุดข้อมูล ผู้วิจัยเลือกใช้วิธีการตรวจสอบความเที่ยงตรง (Validity) ชุดแบบสอบถาม 2 รูปแบบ ด้วยการทดลองใช้ประมวลผลในชุดข้อมูลขนาดเล็ก เพื่อตรวจเช็คผลลัพธ์ที่ไม่ถูกต้องตรงกันของข้อมูล เมื่อพบผลที่ผิดพลาดจะทำการตรวจเช็คหาจุดที่ผิดพลาดและทำการแก้ไขปรับปรุงข้อมูลทั้ง 4 ชุด ก่อนประมวลผลเพื่อบันทึกผลการทดลอง

ในการทดสอบความเที่ยงตรง เป็นเตรียมการก่อนทดสอบผู้วิจัยพบว่าในระบบฐานข้อมูลมายเอสคิวแอล ในขั้นตอนการนำข้อมูลเข้าข้อมูลในฐานข้อมูล และขั้นตอนการประมวลผลจะมีผลต่อประสิทธิภาพการทำงานด้านความเร็วอย่างยิ่งกับข้อมูลที่มีปริมาณมากขึ้นในระดับ 10 ล้านระเบียน ใช้เวลาในการนำเข้ามากกว่า 2 ชั่วโมงขึ้นไป ดังนั้นผู้วิจัยจึงได้ทำการศึกษาเพิ่มเติมในส่วนของความสามารถและคุณสมบัติเฉพาะด้านของระบบฐานข้อมูลมายเอสคิวแอล พบว่าในระบบฐานข้อมูลมาสเอสคิวแอลมีระบบการจัดการฐานข้อมูลเชิงสัมพันธ์หรือมี Storage Engine หลายรูปแบบ ผู้วิจัยเลือกใช้คุณสมบัติที่มีความนิยมในการใช้งานเพื่อทดสอบดังนี้

InnoDB คือคุณสมบัติการจัดการฐานข้อมูลแบบทรานเซคชั่น (Transaction) เป็นการเขียนโปรแกรมการจัดการฐานข้อมูลแบบมีมาตรฐาน ACID เป็นหลัก หรือเรียกว่าการจัดการในระบบทรานเซคชั่น และสนับสนุนการทำงานแบบ Foreign Key แต่มีข้อเสียคือจะทำงานได้ช้ากว่า MyISAM

MyISAM คือ การจัดการแบบดั้งเดิมที่ถูกออกแบบมาโดยการใช้แนวคิดการทำงานที่ตารางมากกว่า และจะถูกอ่านมากกว่าการอับเดท แต่มีข้อเสียคือคุณสมบัตินี้ไม่สนับสนุนการทำงานแบบทรานเซคชั่นจะไม่สามารถเรียกคืนข้อผิดพลาด (Rollback) ได้

จากการทดสอบผู้วิจัยจึงเลือกใช้คุณสมบัติฐานข้อมูลแบบ InnoDB เนื่องจากฐานข้อมูลนี้มีความนิยมการใช้งานโดยทั่วไปในโปรแกรมระบบข้อมูลสุขภาพ ซึ่งคุณสมบัตินี้เป็นการจัดการระดับทรานเซคชั่น ในหน่วยงานระดับปฐมภูมิจะมีการใช้งานกันมาก และจากการตัดสินใจเพื่อเลือกผลลัพธ์เสมือนในการปฏิบัติงานจริงการเรียกใช้รายงานการเจ็บป่วยนี้ จะทำเพียงปีละ 1 ครั้ง หรือเดือนละ 1 ครั้งเท่านั้น

ตารางที่ 4.6 ผลการเปรียบเทียบ Database Engine ของฐานข้อมูลมายเอสคิวแอล

แบบสอบถามข้อมูล	ชุดข้อมูล	จำนวนระเบียนข้อมูล	ค่าเฉลี่ย InnoDB	ค่าเฉลี่ย MyISAM
รายงาน 10 ลำดับแรก จำนวนผู้ป่วยนอก ตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร)	diagnosis_opd_5h	500,000	1.2307	2.4603
	diagnosis_opd_1m	1,000,000	2.6791	4.1068
	diagnosis_opd_5m	5,000,000	61.1114	12.4483
	diagnosis_opd_10m	10,000,000	150.7065	27.5282
รายงานจำนวนผู้ป่วยนอก รวมตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร)	diagnosis_opd_5h	500,000	1.2118	2.3861
	diagnosis_opd_1m	1,000,000	2.6439	4.2724
	diagnosis_opd_5m	5,000,000	60.7291	12.2554
	diagnosis_opd_10m	10,000,000	142.8611	25.2866

ในขั้นตอนการเตรียมการทดสอบการประมวลผลเพื่อปรับปรุงประสิทธิภาพการประมวลผลด้วยแมพรีดิวผู้วิจัยได้ศึกษาจากงานวิจัยที่เกี่ยวข้องจึงทราบว่าขนาดของบล็อกไซค์ (Block Size) มีส่วนสำคัญที่จะส่งผลให้ประสิทธิภาพการประมวลผลขึ้นอยู่กับระบบเครือข่ายด้วย ดังนั้น ผู้วิจัยจึงทำการทดสอบด้วยบล็อกไซค์ ขนาดมาตรฐานของระบบเทคโนโลยีข้อมูลขนาดใหญ่ฮาดูป โดยการคัดเลือกการทดสอบกับข้อมูลขนาดสิบล้านระเบียนด้วยเหตุผลที่ว่าข้อมูลขนาดใหญ่ขึ้นจะเหมาะสมกับการประเมินประสิทธิภาพมากที่สุด ได้ผลทดสอบตารางที่ 4.7 ผู้วิจัยจึงทำการคัดเลือกขนาดบล็อกไซค์เป็น 128MB เพื่อทำการทดสอบเปรียบเทียบเทคโนโลยีข้อมูล

ตารางที่ 4.7 ผลการเปรียบเทียบบล็อกไซค์ของฮาดูปและแมพรีดิว

เวลาการประมวลผลแมพรีดิว (วินาที)

ขนาด	ฮาดูปบล็อกไซต์		
ชุดข้อมูล	มาตรฐาน	64 เมกะไบต์	128 เมกะไบต์
10m	125.59	232.88	94.75

4.5 บันทึกผลลัพธ์จากการประมวลผล

การบันทึกผลในขั้นตอนทดสอบการประมวลผลด้วยชุดคำถามเอสคิวแอลและแมพรีดิว มีการเก็บบันทึกเวลาที่ใช้ในการประมวลผลข้อมูลชุดทดลอง การบันทึกข้อมูลการทดลองกำหนดข้อมูลที่ต้องการบันทึกไว้เป็นมาตรการวัด คือมาตราอัตราส่วน เป็นข้อมูลเวลาที่ใช้ในการประมวลผลเป็นวินาที เก็บบันทึกผลการทดลองในเทคโนโลยีข้อมูล 2 รูปแบบ

1) ผลของการค้นคืนแบบสอบถามข้อมูล ด้วยภาษาสอบถามแบบเอสคิวแอล เมื่อทำทดสอบ 1 ครั้งจะทำการคืนค่าหน่วยความจำสำรองใหม่ และบันทึกเวลาการทดสอบจำนวน 3 ครั้ง และให้มีสูตรบันทึกผลการทดลองดังนี้โดยมีสูตรการบันทึกดังนี้

สูตรการบันทึกผลการค้นคืนด้วยภาษาเอสคิวแอล

Query Time=End Time – Start Time

ตารางที่ 4.8 ผลของการค้นคืนแบบสอบถามข้อมูล ด้วยภาษาสอบถามแบบเอสคิวแอล

จำนวนระเบียน	1) รายงาน 10 ลำดับแรก จำนวนผู้ป่วยนอกตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร)			2) รายงานจำนวนผู้ป่วยนอก รวมตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร)		
	เวลาที่ใช้ (วินาที)					
	ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3
500,000	1.8054	0.9399	0.9467	1.7556	0.9423	0.9376
1,000,000	4.2826	1.8771	1.8776	4.1826	1.8726	1.8765
5,000,000	164.1651	9.5381	9.6311	162.9354	9.6163	9.6357
10,000,000	413.7557	19.1024	19.2613	390.376	19.1001	19.1073

2) ผลของการค้นคืนแบบสอบถามข้อมูล ด้วยเทคนิคแมพรีดิว จำนวน 3 ครั้ง การเก็บผลไม่รวมขั้นตอนการนำผลออกเพื่อจัดทำแสดงผลภาพด้วยโปรแกรมเอ็กเซลและให้มีสูตรบันทึกผลการทดลองดังนี้โดยมีสูตรการบันทึกดังนี้

สูตรการบันทึกผลการค้นคืนด้วยเทคนิคแมพรีดิว

MR j1 = Map j1 + Reduce j1

MR j2 = Map j2 + Reduce j2

Total Time = MR j1 + MR j2

ในขั้นตอนการบันทึกผลของการประมวลผลเทคโนโลยีข้อมูลขนาดใหญ่ ด้วยเทคนิควิธีแมพรีดิวนั้น โดยการเก็บขั้นจากการรายงานผลของโปรแกรมเทคโนโลยีข้อมูขนาดใหญ่ของแต่ละงาน (Job) ดังตัวอย่างนี้

Map j1 : Total time spent by all maps in occupied slots (ms) = 9444

Reduce j1 : Total time spent by all reduces in occupied slots (ms) = 3813

ตารางที่ 4.9 ผลของการค้นคืนแบบสอบถามข้อมูล ด้วยเทคนิคแมพรีดิว

จำนวนระเบียน	เวลาที่ใช้ (วินาที)											
	Map j1			Reduce j1			Map j2			Reduce j2		
	1	2	3	1	2	3	1	2	3	1	2	3
500,000	6.245	6.245	6.956	4.144	4.144	2.800	1.920	1.920	1.937	1.933	1.933	1.875
1,000,000	9.234	9.446	9.075	4.097	3.934	4.449	1.901	1.920	1.913	1.950	1.899	1.916
5,000,000	35.298	37.618	38.236	11.884	9.795	14.253	1.983	2.913	1.929	1.926	1.903	1.938
10,000,000	128.192	134.972	108.763	28.792	27.440	29.924	1.875	1.910	1.899	3.560	1.948	1.925

จากผลการทดลองที่ได้เพื่อให้สามารถทำการเปรียบเทียบประสิทธิภาพด้านเวลาการประมวลผลด้วยภาษาเอสคิวแอลกับวิธีแมพรีดิว ได้ใกล้เคียงกับจัดทำรายงานให้กระบวนการจัดทำรายงานจริง ซึ่งวิธีแมพรีดิวสามารถดำเนินการได้เพียงครั้งเดียวสามารถทำประมวลผลข้อมูลได้ผลลัพธ์ออกรายงานได้ 2 รายงาน จึงทำการสรุปการเปรียบเทียบการประมวลผลด้วยเอสคิวแอลนำผลรายงานทั้งสองมารวมกัน รายงานตัวอย่างที่ 1 + รายงานตัวอย่างที่ 2 และรูปแบบการประมวลผลแบบขนานเทคนิคแมพรีดิว ในงาน MR j1 ผลรวมกับ MR j2 ดังสามารถสรุปได้ดังตารางที่ 4.10

ตารางที่ 4.10 ผลการเปรียบเทียบผลรวมจากการประมวลผลเทคโนโลยีข้อมูล 2 รูปแบบ

จำนวนระเบียน	ภาษาสอบถามเชิงโครงสร้างเอสคิวแอล รายงานที่ 1 + 2 หน่วยเวลาเป็น (วินาที)			ภาษาสอบถามด้วยแทคนิคแมพรีดิว MR j1 + MR j2 หน่วยเวลาเป็น (วินาที)		
	ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3
500,000	3.561	1.8822	1.8843	14.2420	14.2420	13.5680
1,000,000	8.4652	3.7497	3.7541	17.1820	17.1990	17.3530
5,000,000	327.1005	19.1544	19.2668	51.0910	52.2290	56.4460
10,000,000	804.1317	38.2025	38.3686	162.4190	166.2700	142.5110

ผลการเปรียบเทียบประสิทธิผลด้านความแม่นยำถูกต้องทั้ง 2 ระบบ ทำการบันทึกในตารางบันทึกและได้ดำเนินการเปรียบเทียบด้วยค่าร้อยละ (เปอร์เซ็นต์) ดังยกตัวอย่างรายงานที่นำมาแสดงตารางที่ 4.11

ตารางที่ 4.11 สรุปผลเปรียบเทียบความแม่นยำถูกต้องจากการประมวลผล

รายงาน 10 ลำดับแรก จำนวนผู้ป่วยนอก ตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร)

กลุ่มโรค	สาเหตุกลุ่มการป่วย	จำนวนผู้ป่วยนอก								ผลความแม่นยำถูกต้องร้อยละ (%)
		5 แสน		1 ล้าน		5 ล้าน		10 ล้าน		
		SQL	MR	SQL	MR	SQL	MR	SQL	MR	
4	โรคเกี่ยวกับต่อมไร้ท่อ...	40475	40475	80330	80330	402100	402100	803921	803921	100%
9	โรคระบบไหลเวียนเลือด...	38240	38240	75508	75508	378424	378424	756420	756420	100%
10	โรคระบบหายใจ...	32872	32872	66079	66079	329670	329670	661610	661610	100%
11	โรคระบบย่อยอาหาร...	24665	24665	49333	49333	245614	245614	492848	492848	100%
13	โรคระบบกล้ามเนื้อ...	22666	22666	45231	45231	226363	226363	451451	451451	100%
18	อาการแสดงและสิ่งผิดปกติ...	19362	19362	37941	37941	190885	190885	382004	382004	100%
5	ภาวะแปรปรวนทางจิต...	10265	10265	20439	20439	102636	102636	205790	205790	100%
1	โรคติดเชื้อและปรสิต...	8916	8916	17652	17652	87787	87787	176598	176598	100%
14	โรคระบบสืบพันธ์รวม...	8319	8319	17144	17144	84858	84858	169937	169937	100%
7	โรคตารวมส่วนประกอบ...	6943	6943	13972	13972	69832	69832	140556	140556	100%

4.6 นำผลลัพธ์ที่ได้มาวิเคราะห์สถิติ

นำผลลัพธ์ที่ได้นำมาวิเคราะห์สถิติ มีวัตถุประสงค์เพื่อนำผลมาอภิปรายผลทางสถิติ เลือกใช้สถิติเชิงพรรณนา ประกอบด้วยค่าเฉลี่ย, ร้อยละ(เปอร์เซ็นต์) และการแสดงผลด้วยกราฟหรือแผนภูมิ และสถิติเชิงอนุมานจะใช้ค้นหาคำตอบสมมติฐานที่ได้คาดการณ์ไว้ล่วงหน้า

การรวบรวมผลการวิจัยเป็นวินาที จากการประมวลผลจากเทคโนโลยีระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ (RDBMS) จากตารางที่ 4.10 เป็นรายงานที่รวมผลโดยการรวบรวมรายงานตัวอย่างที่ 1 สรุปรวมกับรายงานตัวอย่างที่ 2 ตามกลุ่มจำนวนระเบียน และรูปแบบการจัดเก็บแบบกระจายฮาดูปและการประมวลผลแบบขนานแมพรีดิว ในงาน MR Job1 ผลรวมกับ MR Job 2 จากตาราง ดังตารางที่ 4.10 ทำการวิเคราะห์ผลทางสถิติ t-Test: Paired Two Sample for Means ด้วยโปรแกรมเอ็กเซล ได้สรุปผลการวิเคราะห์ด้วยสถิติตามกลุ่มจำนวนระเบียนดังนี้

ตารางที่ 4.12 ผลการวิเคราะห์ t-Test: Paired Two Sample for Means (5 แสนระเบียน)

ชุดข้อมูล 5 แสนระเบียน

จำนวนการประมวลผล	แมพรีดิว (MR j1+MR j2)	มายเอสคิวแอล (รายงาน 1+2)
ครั้งที่ 1	14.242	3.561
ครั้งที่ 2	14.242	1.8822
ครั้งที่ 3	13.568	1.8843

t-Test: Paired Two Sample for Means (500,000 Record)

	แมพรีดิว (MR j1+MR j2)	*มายเอสคิวแอล (รายงาน 1+2)*
Mean	14.01733333	2.4425
Variance	0.151425333	0.93828279
Observations	3	3
Pearson Correlation	0.49906095	
Hypothesized Mean Difference	0	
df	2	
t Stat	23.73471762	
P(T<=t) two-tail	0.001770424	
t Critical two-tail	4.30265273	

ตารางที่ 4.13 ผลการวิเคราะห์ t-Test: Paired Two Sample for Means (1 ล้านระเบียน)

ชุดข้อมูล 1 ล้านระเบียน

จำนวนการประมวลผล	แมพรีดิว (MR j1+MR j2)	มายเอสคิวแอล (รายงาน 1+2)
ครั้งที่ 1	17.182	8.4652
ครั้งที่ 2	17.199	3.7497
ครั้งที่ 3	17.353	3.7541

t-Test: Paired Two Sample for Means (1,000,000 Record)

	แมพรีดิว (MR j1+MR j2)	*มายเอสคิวแอล (รายงาน 1+2)*
Mean	17.24466667	5.323
Variance	0.008874333	7.40507047
Observations	3	3
Pearson Correlation	-0.575440961	
Hypothesized Mean Difference	0	
df	2	
t Stat	7.437026779	
P(T<=t) two-tail	0.017604107	
t Critical two-tail	4.30265273	

ตารางที่ 4.14 ผลการวิเคราะห์ t-Test: Paired Two Sample for Means (5 ล้านระเบียน)

ชุดข้อมูล 5 ล้านระเบียน

จำนวนการประมวลผล	แมพรีดิว (MR j1+MR j2)	มายเอสคิวแอล (รายงาน 1+2)
ครั้งที่ 1	51.091	327.1005
ครั้งที่ 2	52.229	19.1544
ครั้งที่ 3	56.446	19.2668

t-Test: Paired Two Sample for Means (5,000,000 Record)

	แมพรีดิว (MR j1+MR j2)	*มายเอสคิวแอล (รายงาน 1+2)*
Mean	53.25533333	121.8405667
Variance	7.959026333	31598.73333
Observations	3	3
Pearson Correlation	-0.664156311	
Hypothesized Mean Difference	0	
df	2	
t Stat	-0.661260698	
P(T<=t) two-tail	0.576433824	
t Critical two-tail	4.30265273	

ตารางที่ 4.15 ผลการวิเคราะห์ t-Test: Paired Two Sample for Means (10 ล้านระเบียน)

ชุดข้อมูล 10 ล้านระเบียน

จำนวนการประมวลผล	แมพรีดิว (MR j1+MR j2)	มายเอสคิวแอล (รายงาน 1+2)
ครั้งที่ 1	162.419	804.1317
ครั้งที่ 2	166.270	38.2025
ครั้งที่ 3	142.511	38.3686

t-Test: Paired Two Sample for Means (10,000,000 Record)

	แมพรีดิว (MR j1+MR j2)	*มายเอสคิวแอล (รายงาน 1+2)*
Mean	157.0666667	293.5676
Variance	162.6081243	195506.7821
Observations	3	3
Pearson Correlation	0.363323549	
Hypothesized Mean Difference	0	
df	2	
t Stat	-0.540169293	
P(T<=t) two-tail	0.643184955	
t Critical two-tail	4.30265273	

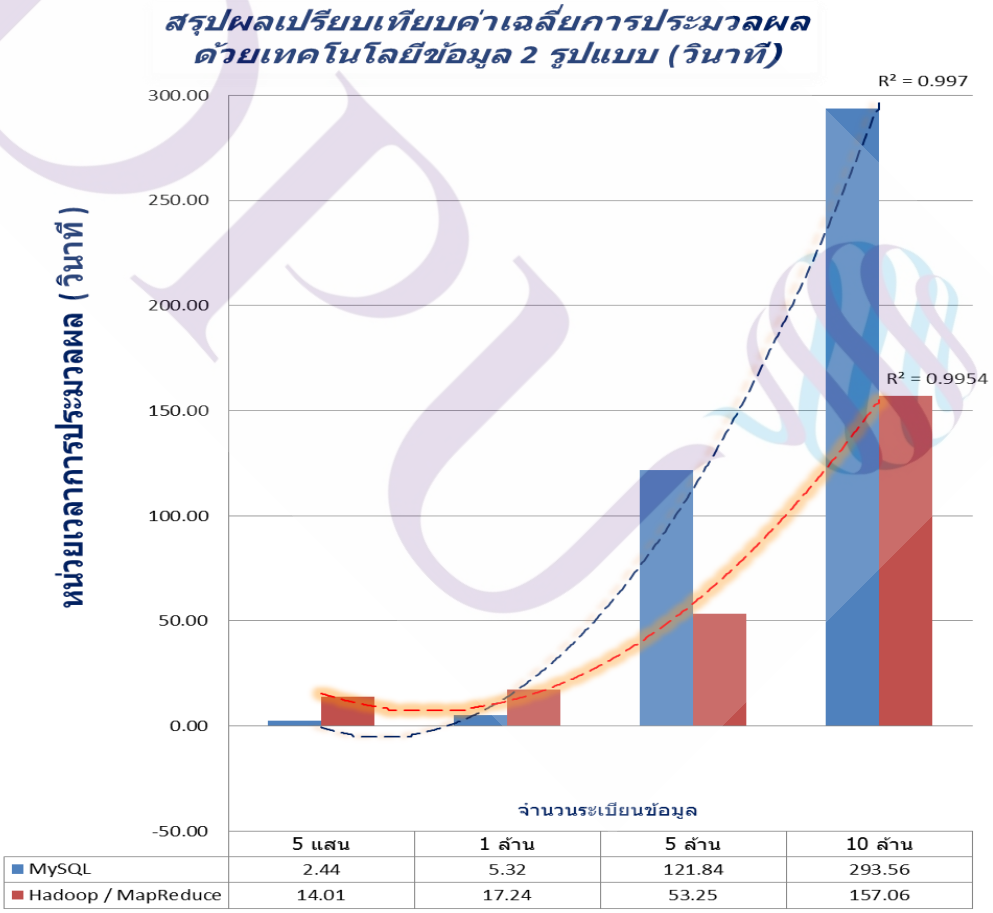
จากผลค่าเฉลี่ยทางสถิติของทุกชุดข้อมูลสามารถนำมาดำเนินการจัดตารางสรุปผลค่าเฉลี่ยทางสถิติ ดังตารางที่ 4.16

ตารางที่ 4.16 ตารางสรุปผลเวลาเฉลี่ยการประมวลผลเทคโนโลยีข้อมูล 2 รูปแบบ (วินาที)

ตารางสรุปผลเปรียบเทียบค่าเฉลี่ยเวลาการประมวลผล (วินาที)

	ค่าเฉลี่ย (Mean)			
จำนวนระเบียนข้อมูล	5 แสน	1 ล้าน	5 ล้าน	10 ล้าน
MySQL	2.44	5.32	121.84	293.56
Hadoop / MapReduce	14.01	17.24	53.25	157.06

จากตารางสรุปผลค่าเฉลี่ยสามารถนำมาดำเนินการจัดทำกราฟเพื่อการวิเคราะห์ผลค่าเฉลี่ยเปรียบเทียบเทคโนโลยีข้อมูล 2 รูปบแบบ ดังภาพที่ 4.5



ภาพที่ 4.5 กราฟแสดงผลเปรียบเทียบการประมวลผลเทคโนโลยีข้อมูล 2 รูปแบบ

4.7 สรุปผลที่ได้จากการวิเคราะห์สถิติ

จากสมมติฐานที่คาดการณ์ไว้ล่วงหน้าว่าผลลัพธ์ของเวลาในการค้นคืนข้อมูล เมื่อมีการเปรียบเทียบระหว่างเทคโนโลยีข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์แตกต่างกัน กำหนดค่านัยสำคัญที่ $\alpha = 0.05$ มีผลคำนวณสถิติ t-Test : Paired Two Sample for Means ด้วยโปรแกรมเอ็กเซลดังนี้

ตารางที่ 4.12 สรุปผลการวิเคราะห์ค่าเฉลี่ยด้วยสถิติ t-Test : Paired Two Sample for Means กลุ่มข้อมูล 5 แสนระเบียน สามารถแปลผลได้ดังนี้ เมื่อ P-Value (two tail) ที่ได้เท่ากับ 0.001 น้อยกว่าค่าแอลฟ่า 0.05 หรือค่าสถิติ t = 23.73 มีค่ามากกว่าจุดวิกฤต (t Critical) -4.30 ถึง 4.30

ตารางที่ 4.13 สรุปผลการวิเคราะห์ค่าเฉลี่ยด้วยสถิติ t-Test: Paired Two Sample for Means กลุ่มข้อมูล 1 ล้านระเบียน สามารถแปลผลได้ดังนี้ เมื่อ P-Value (two tail) ที่ได้เท่ากับ 0.017 น้อยกว่าค่าแอลฟ่า 0.05 หรือค่าสถิติ t = 7.43 มีค่ามากกว่าจุดวิกฤต (t Critical) -4.30 ถึง 4.30

ตารางที่ 4.14 สรุปผลการวิเคราะห์ค่าเฉลี่ยด้วยสถิติ t-Test: Paired Two Sample for Means กลุ่มข้อมูล 5 ล้านระเบียน สามารถแปลผลได้ดังนี้ เมื่อ P-Value (two tail) ที่ได้เท่ากับ 0.576 มากกว่าค่าแอลฟ่า 0.05 หรือค่าสถิติ t = -0.661 มีค่าระหว่างจุดจุดวิกฤต (t Critical) -4.30 ถึง 4.30

ตารางที่ 4.15 สรุปผลการวิเคราะห์ค่าเฉลี่ยด้วยสถิติ t-Test: Paired Two Sample for Means กลุ่มข้อมูล 10 ล้านระเบียน สามารถแปลผลได้ดังนี้ เมื่อ P-Value (two tail) ที่ได้เท่ากับ 0.643 มากกว่าค่าแอลฟ่า 0.05 หรือค่าสถิติ t = -0.54 มีค่าระหว่างจุดวิกฤต (t Critical) -4.30 ถึง 4.30

จากผลทดลองการสอบถามค้นคืนด้วยเทคโนโลยีข้อมูล 2 รูปแบบ มีความถูกต้องและแม่นยำตรงกัน 100% ในทุกชุดข้อมูลและทุกรายงาน ตามตัวอย่างตารางที่ 4.11 เปรียบเทียบผลลัพธ์การประมวลผลด้วยรายงานตัวอย่างที่ 1 และรายงานตัวอย่างที่ 2 (ภาคผนวก ข)

4.8 สรุปผลจากการทดลอง

สรุปผลที่ได้จากการวิเคราะห์ ผลลัพธ์ที่จะได้จากการทดลอง ผู้วิจัยมีแนวคิดการเลือกการอภิปรายผลออกเป็น 3 ส่วนคือ

4.8.1 การวิเคราะห์ผลด้วยสถิติเชิงพรรณนาจากผลการทดลอง

จากผลการทดลองการนำเข้าในระบบฐานข้อมูลเชิงสัมพันธ์มีค่าที่นำเสนอผล 2 ส่วน ดังนี้ 1) ขนาดหน่วยความจุข้อมูล 2) หน่วยเวลาที่ใช้ในกระบวนการนำเข้า จากการทดลองนำเข้าข้อมูลทั้ง 2 กลุ่ม จะสังเกตได้ในตารางที่ 4.3 มีหน่วยความจุที่เพิ่มขึ้นจากไฟล์ที่จัดเตรียมไว้ก่อนนำเข้าในตารางที่ 4.2 ซึ่งส่วนนี้เป็นค่าโอเวอร์เฮดในระบบฐานข้อมูลเป็นส่วนที่นำมาใช้ในการจัดการฐานข้อมูลควบคุมการทำงานของข้อมูล และสังเกตได้ว่าขนาดของไฟล์และเวลาในการ

นำเข้าระหว่างกลุ่มเทคโนโลยีระบบข้อมูลขนาดใหญ่ และเทคโนโลยีระบบข้อมูลฐานข้อมูลเชิงสัมพันธ์ (RDBMS) ในตารางที่ 4.3 น้อยกว่า แต่หากนำไปเทียบกับตารางที่ 4.2 จะเพิ่มขึ้นเล็กน้อย

จากตารางที่ 4.8 และ 4.9 เปรียบเทียบผลการทดลองการประมวลผลข้อมูล 2 รูปแบบ จากผลการทดลองหากเปรียบเทียบแต่ละชุดข้อมูลมีผลค่าเฉลี่ยรายงานตัวอย่างที่ 1 มากกว่ารายงาน 2 เล็กน้อย เนื่องจากมีขั้นตอนการกำหนดลำดับการแสดงผลที่เพิ่มขึ้น และผลกลุ่ม MR j1 มีผลค่าเฉลี่ยมากที่สุดเป็นผลจาก 2 ขั้นตอน คือหลังจากนำข้อมูลเข้าจัดเก็บแบบกระจาย HDFS การเรียกค้นคืนข้อมูลโปรแกรมจะสั่งการประมวลผลและรวบรวมผลข้อมูลที่ถูกจัดเก็บในแต่เครื่อง Slave แล้วนำผลลัพธ์ส่งคืนกลับมาให้เครื่อง Master และจะทำทุกครั้งที่มีการประมวลผล และมีคำสั่งการ Join หรือการเชื่อมสัมพันธ์จะนำระเบียนทั้งหมดในแฟ้มผู้ป่วยนอกมาเชื่อมกับแฟ้ม 21 กลุ่มโรคหลักด้วยรูปแบบ Nested Loop Join หรือการเชื่อมสัมพันธ์โดยไม่มีการจัดเรียง (Index) ก่อน ระเบียนมากขึ้นยิ่งใช้เวลามากขึ้น แล้วนำผลรวมที่ได้จัดเก็บในไฟล์ผลลัพธ์เพื่อใช้ในขั้นตอนต่อไป ซึ่งสังเกตได้ว่างาน MR j2 ใช้เวลาใกล้เคียงกันทุกชุดข้อมูล เนื่องจากขั้นตอนนี้ใช้ผลจาก MR j1 นำมาประมวลผลใหม่ด้วยการจับคู่ข้อมูล รวม เรียงลำดับใหม่ และบันทึกเป็นผลลัพธ์ใหม่ไว้ในรูปแบบเท็กซ์ไฟล์ (Text File)

จากตารางที่ 4.10 ผลของการค้นคืนแบบสอบถามข้อมูล ด้วยภาษาสอบถามแบบมีโครงสร้างเอสคิวแอล จะสังเกตุได้ว่าผลของครั้งที่ 1 ของการประมวลผลจากรายงานตัวอย่างที่ 1 และรายงานที่ 2 จะมีผลมากกว่าการประมวลผลครั้งที่ 2 และครั้งที่ เนื่องจากในระบบการจัดการข้อมูลเชิงสัมพันธ์มีการจัดการที่เรียกว่าแคช (Cache) ซึ่งในการทดลองได้มีการใช้คำสั่งเคลียร์ค่าแคชด้วยการใช้ Query Cache เป็นการเคลียร์ค่าที่หน่วยความจำสำรอง (Ram) แต่ยังคงมีผลกับการประมวลผลในครั้งแรกทำให้การประมวลผลในครั้งใช้เวลามากกว่าครั้งที่ 2 และครั้งที่ 3 ซึ่งในการทดสอบผู้วิจัยได้ทดสอบการเคลียร์ค่าแคชด้วยวิธีการรีสตาร์ทเครื่องเซิร์ฟเวอร์ใหม่พบว่าในทุกครั้งที่มีการปิดและเปิดเครื่องใหม่การประมวลผลเริ่มต้นจะใช้เวลามากกว่าอยู่เสมอ แต่ทว่าในการสอบถามค้นคืนแบบเทคนิควิธีแมพรีดิวจะใกล้เคียงกันทุกครั้งเนื่องจากในระบบเทคโนโลยีข้อมูลขนาดใหญ่ที่ใช้ในการทดสอบไม่มีฟังก์ชั่นการจัดเก็บข้อมูลที่ทำการประมวลผลในหน่วยความจำจึงทำให้มีผลใกล้เคียงกันทุกครั้ง

นำผลค่าเฉลี่ยทางสถิติจากตารางที่ 4.12 และ 4.13 และ 4.14 และ 4.15 นำมาดำเนินการจัดตารางสรุปผลค่าเฉลี่ย ดังตารางที่ 4.16 และจัดทำกราฟเพื่อการวิเคราะห์แผนภาพ เมื่อนำผลรวมค่าเฉลี่ยทุกขั้นตอน ระหว่างผลความเร็วการประมวลผลของฐานข้อมูลเชิงสัมพันธ์มายเอสคิวแอลด้วยภาษาสอบถามเชิงโครงสร้างเอสคิวแอล (SQL) ที่รวมค่าเฉลี่ยจากรายงานตัวอย่างที่ 1 และรายงานตัวอย่างที่ 2 เปรียบเทียบการประมวลผลเทคโนโลยีข้อมูลขนาดใหญ่ฮาดูปและแมพรีดิวที่

รวมค่าเฉลี่ย 2 ขั้นตอนแมพและรีดิว มีผลค่าเฉลี่ยมากกว่าทุกชุดข้อมูลทดสอบและมีแนวโน้มเพิ่มมากขึ้นตามชุดข้อมูล ดังภาพกราฟที่ 4.5 เพื่อทำการหาจุดตัดของข้อมูลที่เทคโนโลยีระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ (RDBMS) มีข้อมูลจำนวนระเบียนเท่าใด ที่ประสิทธิภาพของ RDBMS ทำงานมีประสิทธิภาพน้อยกว่าเทคโนโลยีข้อมูลขนาดใหญ่ สามารถนำมาเป็นจุดที่ต้องการปรับเปลี่ยน หรือนำเสนอการเตรียมความพร้อมการปรับปรุงจากระบบเดิมไปสู่ระบบใหม่ในการใช้งานร่วมกับข้อมูลในระบบข้อมูลสุขภาพสำหรับผลด้านประสิทธิภาพความเร็วที่คาดหวังว่าจะค้นพบจุดตัดระหว่างเพื่อเตรียมการปรับเปลี่ยนระบบ และเล็งเห็นถึงจุดที่ต้องควรป้องกันจากการประมวลผลเพื่อเรียกรายงานได้ไม่ทันต่อความต้องการใช้งาน จากการทดลองนี้สามารถหาจุดตัดของกราฟด้านผลความเร็วว่าจุดใด อยู่ที่จำนวนระเบียนข้อมูลโดยประมาณ 1 ล้านระเบียน ที่ระบบฐานข้อมูลเชิงสัมพันธ์ดำเนินการได้ต่ำกว่าหรือมีประสิทธิภาพด้อยกว่าเทคโนโลยีข้อมูลขนาดใหญ่ฮาดูปแลแมพรีดิว เป็นจุดที่ต้องเริ่มมีการพิจารณาการปรับเปลี่ยน

แต่ทั้งนี้การพิจารณานี้เป็นส่วนของการเรียกใช้งานการประมวลผลที่ตารางแฟ้มข้อมูลหลักจำนวน 9 คอลัมน์และแฟ้มข้อมูลเชื่อมสัมพันธ์ 1 คอลัมน์ และ 1 การเชื่อมสัมพันธ์ (Join) เท่านั้น การใช้เวลาในการประมวลผลอาจจะมากขึ้นตามลำดับหากมีตารางแฟ้มข้อมูลในการเชื่อมโยงมากกว่า 1 การเชื่อมโยงและหรือจำนวนคอลัมน์ในแฟ้มข้อมูลมีจำนวนมากขึ้น

4.8.2 การวิเคราะห์ผลด้วยสถิติอนุมานที่จะใช้พิสูจน์ผลจากการตั้งสมมติฐาน

จากสมมติฐานที่คาดการณ์ว่าผลลัพธ์ของเวลาการค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างระบบข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์ มีผลลัพธ์ที่แตกต่างกัน

สรุปผลจากการทดลอง จากสมมติฐานที่คาดการณ์ไว้ ล่วงหน้าว่าผลลัพธ์ของเวลาในการค้นคืนข้อมูล เมื่อมีการเปรียบเทียบระหว่างเทคโนโลยีข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์แตกต่างกัน ในทุกๆ ชุดข้อมูล ดังนี้

ตารางที่ 4.12 กลุ่มข้อมูล 5 แสนระเบียน เมื่อ P-Value (two tail) ที่ได้เท่ากับ 0.001 น้อยกว่าค่าแอลฟ่า 0.05 หรือค่าสถิติ t = 23.73 มีค่ามากกว่าจุดวิกฤต (t Critical) -4.30 ถึง 4.30 จึงเป็นการยอมรับสมมติฐานว่าเทคโนโลยีข้อมูลทั้ง 2 รูปแบบ แตกต่างกัน

ตารางที่ 4.13 กลุ่มข้อมูล 1 ล้านระเบียน เมื่อ P-Value (two tail) ที่ได้เท่ากับ 0.017 น้อยกว่าค่าแอลฟ่า 0.05 หรือค่าสถิติ t = 23.73 มีค่ามากกว่าจุดวิกฤต (t Critical) -4.30 ถึง 4.30 จึงเป็นการยอมรับสมมติฐานว่าเทคโนโลยีข้อมูลทั้ง 2 รูปแบบ แตกต่างกัน

ตารางที่ 4.14 กลุ่มข้อมูล 5 ล้านระเบียน เมื่อ P-Value (two tail) ที่ได้เท่ากับ 0.002 น้อยกว่าค่าแอลฟ่า 0.05 หรือค่าสถิติ t = -0.661 มีค่ามากกว่าจุดวิกฤต (t Critical) -4.30 ถึง 4.30 จึงเป็นการปฏิเสธสมมติฐานว่าเทคโนโลยีข้อมูลทั้ง 2 รูปแบบ ไม่แตกต่างกัน เนื่องจากมายเอสคิวแอลใช้

เวลาในการประมวลผลครั้งที่ 1 สูง แต่ในครั้งที่ 2 และครั้งที่ 3 ลดลงเป็นจำนวนมาก เพราะมีระบบการสร้างหน่วยความจำสำรอง (Cache) จึงทำให้การวิเคราะห์ด้วยสถิติมีผลไม่แตกต่างกัน

ตารางที่ 4.15 กลุ่มข้อมูล 10 ล้านระเบียน เมื่อ P-Value (two tail) ที่ได้เท่ากับ 0.003 น้อยกว่าค่าแอลฟ่า 0.05 หรือค่าสถิติ t = -0.540 มีค่ามากกว่าจุดวิกฤต (t Critical) -4.30 ถึง 4.30 จึงเป็นการปฏิเสธสมมติฐานว่าเทคโนโลยีข้อมูลทั้ง 2 รูปแบบ ไม่แตกต่างกัน 30 จึงเป็นการปฏิเสธสมมติฐานว่าเทคโนโลยีข้อมูลทั้ง 2 รูปแบบ ไม่แตกต่างกัน เนื่องจากมายเอสคิวแอลใช้เวลาในการประมวลผลครั้งที่ 1 สูง แต่ในครั้งที่ 2 และครั้งที่ 3 ลดลงเป็นจำนวนมาก เพราะมีระบบการสร้างหน่วยความจำสำรอง (Cache) จึงทำให้การวิเคราะห์ด้วยสถิติมีผลไม่แตกต่างกัน

จากสมมติฐานที่คาดการณ์ไว้ล่วงหน้าว่าผลลัพธ์ของเวลาการค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างระบบข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์ มีผลลัพธ์ที่แตกต่างกัน นั้นการประเมินผลทางสถิติมีผลที่ยอมรับสมมติฐานเฉพาะในกลุ่มชุดข้อมูล 5 แสนและ 1 ล้านระเบียน แต่มีการปฏิเสธสมมติฐานในกลุ่มชุดข้อมูล 5 ล้านและ 10 ล้านระเบียน

ผู้วิจัยได้วิเคราะห์สาเหตุจากการปฏิเสธสมมติฐาน ผลกับการประมวลผลในครั้งแรกทำให้การประมวลผลในครั้ง 1 ใช้เวลามากกว่าครั้งที่ 2 และครั้งที่ 3 ซึ่งในการทดสอบผู้วิจัยได้ทดสอบการเคลียร์ค่าแคชด้วยวิธีการรีสตาร์ทเครื่องเซิร์ฟเวอร์ใหม่ พบว่าในทุกครั้งที่มีการปิดและเปิดเครื่องใหม่ เมื่อการประมวลผลเริ่มต้นจะใช้เวลามากกว่าครั้งที่ 2 และ 3 อยู่เสมอ จึงคาดว่าเป็นสาเหตุในการทดสอบสมมติฐานทางสถิติเกิดความคลาดเคลื่อน

เนื่องจากในระบบการจัดการฐานข้อมูลเชิงสัมพันธ์มีระบบฟังก์ชั่นเพื่อช่วยในการจัดเก็บข้อมูลที่ต้องการเรียกใช้งานบ่อยครั้ง ในฮาร์ดแวร์ (แรม, ฮาร์ดดิสก์) เพื่อทำให้การค้นหาในครั้งต่อไปสะดวกและรวดเร็วมากยิ่งขึ้น นอกจากนี้ระบบการจัดการฐานข้อมูลมายเอสคิวแอลยังมีการจัดการแคชอีกหลายประเภท เช่น Table Cache หรือ Thread Cache (Schwartz et al., 2012, pp. 353-354)

จากตารางที่ 4.11 ผลทดลองการเปรียบเทียบผลการสอบถามค้นคืนข้อมูลมีความถูกต้องและแม่นยำตรงกันของเทคโนโลยี 2 รูปแบบ พบว่ามีผลลัพธ์ถูกต้องตรงกันในทุกชุดข้อมูลและทุกรายงาน 100% ดังตัวอย่างที่นำมาแสดงเป็นรายงานที่ 1 ชุดข้อมูล 5 แสนระเบียน และ 1 ล้านระเบียน และ 5 ล้านระเบียน และ 10 ล้านระเบียน

สรุปผลได้ว่าจากสมมติฐานที่คาดการณ์ไว้ล่วงหน้าว่าผลลัพธ์ของความแม่นยำถูกต้องการค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างระบบข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์ มีผลลัพธ์ไม่แตกต่างกัน เป็นการยอมรับสมมติฐาน

4.8.3 การวิเคราะห์ผลด้วยสถิติเชิงพรรณนาจากผลการทดลองนำไปเปรียบเทียบกับงานวิจัยที่เกี่ยวข้อง

การเปรียบเทียบกับงานวิจัยที่เกี่ยวข้องที่มีความใกล้เคียงกับการวิจัยทดลองนี้ ผลการทดลองที่ได้ต่างกับงานวิจัยที่เกี่ยวข้องที่มีการใช้แมพรีดิว, ไฮฟ์และมายเอสคิวแอล ทำทดสอบด้วยข้อมูลการชำระเงินของลูกค้าในธุรกิจขนาดเล็ก มีขนาดข้อมูลตั้งแต่ 235MB – 9GB กับเครื่องจำนวน 1-4 เครื่อง ผลงานวิจัยนี้สรุปว่ามายเอสคิวแอลจะใช้เวลามากกว่าแมพรีดิวและไฮฟ์ที่ขนาดข้อมูลหนึ่งหมื่นบัญชี 5GB ใช้เวลา 25 นาที แมพรีดิวจะใช้เวลาประมาณ 80 - 90 วินาที และในทุกชุดข้อมูลทดสอบโปรแกรมแมพรีดิวมีประสิทธิภาพสม่ำเสมอและดีที่สุด สามารถประมวลผลข้อมูลขนาดเล็กได้ดี (Hollingsworth, 2012, pp. 43-44)

ซึ่งสามารถสรุปผลที่ได้จากการทดลองสอดคล้องกับบทความวิจัยนี้ที่พบว่าขนาดข้อมูล 682MB และมีระเบียนข้อมูลสิบล้านจะใช้เวลาในการประมวลด้วยเทคโนโลยีระบบข้อมูลขนาดใหญ่ฮาดูปและแมพรีดิวที่มีผลการทดลองที่ 157 วินาที น้อยกว่าการประมวลผลมายเอสคิวแอลที่ใช้เวลา 293 วินาที จึงสามารถสรุปผลการทดลองได้ว่าเมื่อข้อมูลเริ่มมีขนาดใหญ่ขึ้น การประมวลด้วยเทคโนโลยีข้อมูลขนาดใหญ่ อย่างโปรแกรมโอเพ่นซอร์สฮาดูปและแมพรีดิวสามารถนำมาใช้งานได้อย่างเป็นอย่างดีกับข้อมูลในระบบข้อมูลสุขภาพ

ผลการทดลองที่ได้เทคโนโลยีข้อมูลขนาดใหญ่เหมาะกับการใช้งานประมวลผลชุดข้อมูลขนาดใหญ่ (ชูพันธ์ รัตนโภคา, 2555, น. 27) เมื่อจำนวนข้อมูลเริ่มมีขนาดใหญ่ ในการพิจารณาควรพิจารณาข้อมูลตั้งแต่ระดับข้อมูลจำนวนข้อมูลกิกะไบต์ (GB) และเทราไบต์ (TB) และเพตาไบต์ (PB)

บทที่ 5
อภิปรายผลงานวิจัย และข้อเสนอแนะ

ในปัจจุบันเทคโนโลยีระบบการจัดเก็บแบบกระจายและการประมวลผลแบบขนานที่มีในระบบนิเวศข้อมูลขนาดใหญ่ เช่น โปรแกรมโอเพ่นซอร์สฮาดูปและแมพรีดิวจะสามารถนำมาประมวลผลข้อมูลระบบสุขภาพขนาดใหญ่จากคลังข้อมูลด้านการแพทย์และสุขภาพที่มีการจัดการฐานข้อมูลเชิงสัมพันธ์ได้หรือไม่ และมีแนวทางอย่างไรหากต้องปรับเปลี่ยนวิธีการประมวลผลหรือหากจะต้องประยุกต์ใช้กับสิ่งที่มีอยู่เดิมต้องทำอย่างไร ซึ่งการทำงานกับข้อมูลขนาดใหญ่จะมีปัญหาในการจัดเก็บข้อมูล การโอนย้ายข้อมูล และการสำรองข้อมูล อีกทั้งการค้นคืนข้อมูลจะมีวิธีการอย่างไร ที่จะช่วยทำให้การจัดการค้นคืนข้อมูล และนำข้อมูลจำนวนมากเหล่านี้มาใช้ประโยชน์ได้ภายในเวลาอันรวดเร็วและมีประสิทธิภาพมากที่สุด แต่ยังคงไว้ให้ได้ซึ่งความแม่นยำถูกต้องของข้อมูลที่ได้รับการค้นคืน ด้วยความสำคัญของคุณภาพข้อมูลในระบบบริการสุขภาพจากลักษณะสำคัญ 4 ส่วนคือ ครบถ้วน ถูกต้อง ละเอียด และทันสมัย

5.1 อภิปรายงานวิจัย

การอภิปรายผลการวิจัยจากการทดลองเพื่อค้นหาคำตอบตามจุดประสงค์งานวิจัย 1) เพื่อศึกษาแนวทางที่เหมาะสมในการจัดเก็บข้อมูลบริการสุขภาพบนสถาปัตยกรรมข้อมูลขนาดใหญ่ 2) เพื่อเสริมสร้างความรู้และความเข้าใจในเทคโนโลยีข้อมูล ระหว่างเทคโนโลยีข้อมูลขนาดใหญ่ที่มีรูปแบบการจัดเก็บแบบกระจาย (ฮาดูป) และการประมวลผลแบบขนาน (แมพรีดิว) มีสถาปัตยกรรมการจัดการข้อมูลและใช้หลักการทางคณิตศาสตร์ และรูปแบบเทคนิควิธีการสอบถามค้นคืนข้อมูลที่แตกต่างกันกับ ระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ (มายเอสคิวแอล) และนำมาประยุกต์ใช้ในการจัดทำสถิติข้อมูลการเจ็บป่วย 3) เพื่อเปรียบเทียบประสิทธิภาพด้านเวลาการประมวลผลและความถูกต้องแม่นยำในการค้นคืนข้อมูล ผลการศึกษาในงานวิจัยที่เกี่ยวข้องและจากผลการทดลองผู้วิจัยขอแบ่งการอภิปรายผลเป็นประเด็นหัวข้อตามวัตถุประสงค์งานวิจัย ดังนี้

5.1.1 เพื่อศึกษาแนวทางที่เหมาะสมในการจัดเก็บข้อมูลบริการสุขภาพบนสถาปัตยกรรมข้อมูลขนาดใหญ่ แบ่งการอภิปรายผลดังนี้

เทคโนโลยีการจัดเก็บแบบกระจายฮาดูปและการประมวลผลแมพรีดิวสามารถนำมาประยุกต์ใช้ร่วมกันกับคลังข้อมูลด้านการแพทย์และสุขภาพได้ ข้อมูลในระบบบริการสุขภาพเป็นข้อมูลที่มีการจัดการข้อมูลแบบมีโครงสร้างเป็นแฟ้มที่มีโครงสร้างมาตรฐานในการจัดเก็บ 43+7 แฟ้มมาตรฐาน และเป็นชุดข้อมูลที่มีการเชื่อมความสัมพันธ์กัน โดยพิจารณาจากผลจากการทดลองดังนี้

การจัดเก็บข้อมูลโดยการดำเนินการจับเก็บในปัจจุบันมีการการจัดส่งข้อมูลจัดทำบนเครื่องลูกข่ายระดับอำเภอเข้าสู่เครื่องแม่ข่ายระดับจังหวัด การจัดส่งข้อมูล 43+7 แฟ้ม จัดทำบนเครื่องแม่ข่ายระดับจังหวัดสู่เครื่องแม่ข่ายระดับเขตและกระทรวง การประมวลผลเพื่อจัดทำรายงานสถิติและดัชนีชี้วัดบนเครื่องแม่ข่ายระดับจังหวัด และบนเครื่องแม่ข่ายระดับเขต และกระทรวงสาธารณสุขในระดับประเทศ เป็นไฟล์ข้อมูลรูปแบบเท็กซ์ไฟล์ ซึ่งสามารถนำข้อมูลเข้าใช้งานในกรอบการทำงานฮาดูปได้ทันที โปรแกรมฮาดูปยังเป็นโปรแกรมที่สามารถรองรับได้กับระบบปฏิบัติการหลายรูปแบบ มีลักษณะการจัดเก็บแบบกระจายที่มีเท็กซ์ไฟล์เป็นไฟล์ข้อมูลหลักสามารถจัดเก็บได้ทั้งแบบเครื่องเดียวและหลายเครื่อง จากผลการทดลองฮาดูปใช้พื้นที่จัดเก็บน้อยกว่าและใช้เวลานำเข้าน้อยกว่า เนื่องจากในการจัดการฐานข้อมูลเชิงสัมพันธ์มีระบบควบคุมการทำงาน และผลของเวลาในการจัดเก็บในระบบฐานข้อมูลเชิงสัมพันธ์ ดังนั้นระบบฐานข้อมูลมีการจัดการเชิงสัมพันธ์จึงเป็นต้นทุนอีกทางหนึ่ง หากเทียบกับเทคโนโลยีฮาดูปที่มีต้นทุนการจัดเก็บน้อยกว่า และเหมาะสมกับการจัดเก็บข้อมูลแบบเท็กซ์ไฟล์ในระบบงานปัจจุบัน

การประมวลผลโดยกรอบการทำงานแมพรีดิวมีรูปแบบวิธีการใช้แบบจับคู่ เป็นการเขียนโปรแกรมประมวลผลด้วยอัลกอริทึมควบคุมความต้องการข้อมูลผ่านการจับคู่ Key/Value จากผลการทดลองสามารถนำมาใช้กับชุดข้อมูลระบบบริการสุขภาพได้ โดยชุดข้อมูลสิบล้านระเบียนมายเอสคิวแอลใช้เวลาในการประมวลผล 2 รายงาน ใช้เวลาการสอบถามค้นคืนข้อมูลมากกว่าเปรียบเทียบกับการประมวลผลด้วยเทคนิคแมพรีดิว แม้ว่าแมพรีดิวใช้การประมวลผลครั้งเดียวได้ 2 รายงาน จากกราฟเปรียบเทียบรูปที่ 4.5 จะเห็นได้ว่าเมื่อชุดข้อมูลมีระเบียนเพิ่มขึ้นจะส่งผลต่อเวลาในการประมวลผลเป็นลำดับและมีแนวโน้มเพิ่มมากขึ้น เมื่อข้อมูลเริ่มมีจำนวนมากขึ้นการประมวลผลจึงเหมาะสมที่จะเริ่มมีการปรับเปลี่ยนมาใช้เทคโนโลยีข้อมูลขนาดใหญ่ และผลลัพธ์ที่ได้การประมวลผลด้วยโปรแกรมฮาดูปและแมพรีดิวหากเทียบผลกับระบบฐานข้อมูลเชิงสัมพันธ์ มีผลลัพธ์ถูกต้องตรงกันในทุกชุดข้อมูลและทุกรายงานผลลัพธ์จากการทดลอง (ภาคผนวก ข.)

ผู้วิจัยวิเคราะห์ปัจจัยหรือสาเหตุความเหมาะสมในการนำเทคโนโลยีข้อมูลขนาดใหญ่มาใช้งานกับข้อมูลระบบบริการสุขภาพจากการทดลองพบว่า การประมวลผลด้วยแมพรีดิวนั้นมีขั้นตอนมากกว่าไม่ส่งผลต่อประสิทธิภาพด้านเวลาในการประมวลผล เช่น ขั้นตอน MR j1 ชุดสิบ

ล้านใช้เวลาเฉลี่ย 157.06 วินาที. ใช้เวลามากในขั้นตอนการเชื่อมสัมพันธ์ และประมวลผลเพื่อนับจำนวนในแต่ละกลุ่มโรคก่อน ต่างกับขั้นตอน MR j2 ชุดสิบล้าน ใช้ประมวลผลรวมแต่ละกลุ่มโรคจากไฟล์ที่รับจาก MR j1 ดังนั้นในขั้นตอน MR j1 มีผลต่อเวลาการประมวลผลเพื่อค้นคืนข้อมูลมากที่สุด ซึ่งเปรียบเทียบกับระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ (RDBMS) ที่เมื่อข้อมูลเพิ่มมากขึ้น การเชื่อมสัมพันธ์หรือการ Join จะใช้เวลาในการประมวลผลเพิ่มสูงขึ้น หากมีการปรับปรุงประสิทธิภาพการดำเนินการดังกล่าวจะทำให้มีประสิทธิภาพทางด้านเวลามากขึ้นได้สารสนเทศที่ทันสมัย สถาปัตยกรรมนี้เป็นที่นิยมอย่างมากกับการคำนวณแบบขนานใช้งานกับข้อมูลขนาดใหญ่ ถึงแม้ว่าจะมีวิธีการ ที่มีความหลากหลายของการพัฒนาสำหรับเทคนิคแมพรีดิว มีไม่กี่วิธีที่สามารถจะบรรลุเป้าหมายที่เหมาะสมในการทำประมวลผลแบบคู่ขนาน และการเกิดภาระงานที่สมดุลอีกทั้งการทำงานข้ามเครื่องที่ร่วมอยู่ภายในเครือข่าย และเพิ่มความเร็วให้มากขึ้น (Tao et al., 2013)

โปรแกรมฮาดูปและแมพรีดิวจะมีประสิทธิภาพสูง ขึ้นอยู่กับขนาดของข้อมูลในแต่ละงานที่ต้องการประมวลผล และชุดข้อมูลที่เชื่อมสัมพันธ์ต้องมีการกรองข้อมูลไว้ล่วงหน้าจะส่งผลดีต่อความเร็ว และสามารถใช้ประมวลผลกับชุดข้อมูลที่มีโครงสร้างเชิงสัมพันธ์ได้ หากการนำมาใช้งานต้องเข้าใจกระบวนการทำงานในโปรแกรมเพื่อปรับการตั้งค่าเริ่มต้นให้เหมาะสม และปรับปรุงวิธีเขียนโปรแกรมเพื่อประยุกต์ใช้กับงานในชุดข้อมูลเฉพาะที่ต้องการ และจำเป็นต้องกำหนดรูปแบบผลลัพธ์ไว้ล่วงหน้าจึงจะเกิดประสิทธิภาพอย่างสูงสุด ผลจากการวิจัยพบว่าการปรับปรุงกระบวนการฮาดูปในการขยายแบบขึ้นมีนัยสำคัญ ในการประเมินด้านการใช้พลังงานและการเพิ่มชั้นการจัดเก็บ (Rack) และด้านค่าใช้จ่าย อีกทั้งยังมีส่วนของการใช้งานภายในหน่วยความจำที่มีประสิทธิภาพ (Appuswamy et al., 2013)

แต่ยังมีข้อจำกัดที่ต้องใช้ทรัพยากรบุคคลที่มีความเชี่ยวชาญในการจัดทำโปรแกรมเฉพาะทางนี้ไว้ให้ในแต่ละความต้องการข้อมูลหรือแต่ละองค์กรโดยเฉพาะ แต่ทว่าผู้ผลิตสนับสนุนการจัดทำโปรแกรมเพื่อให้สามารถโปรแกรมเมอร์ประสบการณ์น้อยใช้งานได้ง่ายและสะดวก (Dean & Ghemawat, 2008, p. 1) รูปแบบวิธีการค้นคืนด้วยโปรแกรมแมพรีดิวใช้รูปแบบการจัดเก็บลักษณะของแถวหรือระเบียนข้อมูลและทำการนับจากจำนวนตัวอักษร (Digit) 3 ตัวแรกของข้อมูลรหัส ICD10 เพื่อทำการเชื่อมสัมพันธ์กับข้อมูลหลัก 21 กลุ่มโรค และทำการเรียกคืนข้อมูลด้วยวิธีการนับจำนวน ซึ่งเป็นวิธีการที่ทำได้อย่างมีประสิทธิภาพและประสิทธิผลในการทดลอง แต่การใช้ชุดแบบสอบถามข้อมูลหากมีการใช้คำสั่งหรือโปรแกรมคำสั่งที่ไม่เข้าใจผลลัพธ์ที่ต้องการ และชุดข้อมูลที่ถูกจัดเตรียมไว้ไม่ถูกต้องและมีคุณภาพ หรือตรงตามผลลัพธ์ที่ต้องการ อีกทั้งการเชื่อมสัมพันธ์ของข้อมูลหากไม่มีคุณภาพและถูกต้อง ตัวแปรต้นเหล่านี้จะส่งผลสูง ทำให้ตัวแปรตาม คือ ผลลัพธ์การประมวลผลผิดพลาดได้

สำหรับผลด้านประสิทธิภาพความเร็วตามแนวคิดที่คาดหวังว่าจะค้นพบจุดที่การประมวลผลด้วยระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ (มายเอสคิวแอล) มีประสิทธิภาพด้อยกว่าเทคโนโลยีข้อมูลขนาดใหญ่ที่มีรูปแบบการจัดเก็บแบบกระจาย (ฮาดูป) และการประมวลผลแบบขนาน (แมพรีดิว) เพื่อเตรียมการปรับเปลี่ยนระบบ หรือเล็งเห็นถึงจุดที่ต้องควรป้องกันจากการประมวลผลเพื่อเรียกรายงาน จากการทดลองนี้สามารถหาจุดตัดของเส้นกราฟด้านผลความเร็วว่าข้อมูลจำนวน 1 ล้านระเบียน ระบบฐานข้อมูลเชิงสัมพันธ์ดำเนินการได้หรือมีประสิทธิภาพด้อยกว่าเทคโนโลยีข้อมูลขนาดใหญ่ฮาดูปและแมพรีดิว

5.1.2 เพื่อเสริมสร้างความรู้และความเข้าใจในเทคโนโลยีข้อมูล ระหว่างเทคโนโลยีข้อมูลขนาดใหญ่ที่มีรูปแบบการจัดเก็บแบบกระจาย (ฮาดูป) และการประมวลผลแบบขนาน (แมพรีดิว) มีสถาปัตยกรรมการจัดการข้อมูลและใช้หลักการทางคณิตศาสตร์และรูปแบบเทคนิควิธีการสอบถามค้นคืนข้อมูลที่แตกต่างกันกับ ระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ (มายเอสคิวแอล) และนำมาประยุกต์ใช้ในการจัดทำสถิติข้อมูลการเจ็บป่วย ระบบฐานข้อมูลเชิงสัมพันธ์ (RDBMS) มีข้อแตกต่างกับเทคโนโลยีข้อมูลขนาดใหญ่ แบ่งการอภิปรายผลได้ดังนี้

1) ใช้ขั้นตอนที่ต่างกันในการค้นคืนข้อมูล เช่น RDBMS จัดเรียงก่อนแสดงผลคืนข้อมูล แต่แมพรีดิว ขึ้นอยู่กับแต่ละการเขียนโปรแกรมเพื่อการสอบถามค้นคืนข้อมูล

2) ใช้หลักทางคณิตศาสตร์ต่างกันในการค้นคืนข้อมูล เช่น RDBMS ใช้ทฤษฎีเซต แต่แมพรีดิวใช้แบบอาร์เรย์

3) รูปแบบการเรียกคืนต่างกัน เช่น RDBMS ใช้ภาษามีโครงสร้างเอสคิวแอลเพียงอย่างเดียว แต่แมพรีดิวมีการเขียนโปรแกรมแบบมีอัลกอริทึม และสามารถเขียนได้หลากหลายรูปแบบ

4) การจัดการข้อมูลต่างกัน เช่น RDBMS มีการจัดการฐานข้อมูลตามคุณสมบัติ ACID และแบบตารางสัมพันธ์ แต่แมพรีดิวใช้รูปแบบการจับคู่ข้อมูล (Key/Value)

5.1.3 เพื่อเปรียบเทียบประสิทธิภาพด้านเวลาการประมวลผล และความถูกต้องแม่นยำในการค้นคืนข้อมูล แบ่งการอภิปรายผลเป็น 2 หัวข้อ ดังนี้

เทคโนโลยีการจัดเก็บแบบกระจายฮาดูปและการประมวลผลแมพรีดิวสามารถนำมาประยุกต์ใช้เพื่อการประมวลผลข้อมูลจากคลังข้อมูลด้านการแพทย์และสุขภาพให้ได้ผลลัพธ์ได้รวดเร็วขึ้น การใช้ชุดแบบสอบถามข้อมูลหากมีการเรียกใช้คำสั่งที่ไม่เข้าใจผลลัพธ์ที่ต้องการล่วงหน้า และมีการปรับปรุงกระบวนการค้นคืนเพื่อเพิ่มประสิทธิภาพ พบว่าตัวแปรต้นกลุ่มนี้มีความสัมพันธ์โดยตรงที่จะส่งผลต่อตัวแปรตาม คือเวลาที่ใช้ในการประมวลผลข้อมูล

เทคโนโลยีการจัดเก็บแบบกระจายฮาดูปและการประมวลผลแมพรีดิวสามารถนำมาประยุกต์ใช้เพื่อการประมวลผลข้อมูลจากคลังข้อมูลระบบบริการสุขภาพได้ และสามารถใช้งาน

ได้ผลลัพธ์ในการค้นคืนที่ถูกต้อง และหากมีการปรับปรุงกระบวนการค้นคืนเพื่อเพิ่มประสิทธิภาพ พบว่าตัวแปรต้นเหล่านี้ไม่ส่งผลต่อตัวแปรตาม คือผลลัพธ์การประมวลผล

5.2 ข้อเสนอแนะ

5.2.1 การจับคู่เชื่อมความสัมพันธ์ที่ดีในการเขียนโปรแกรมแมพรีดิวจะทำให้ได้ผลลัพธ์ที่ถูกต้อง และข้อมูลที่ใช้ในการเชื่อมความสัมพันธ์ต้องเข้าใจประเภทของข้อมูลที่ใช้ในการจับคู่ความสัมพันธ์ และต้องทำความเข้าใจผลลัพธ์ที่ต้องการว่าต้องการแสดงผลรูปแบบใด เพื่อควบคุมคุณภาพการเขียนโปรแกรมให้ได้ตามผลลัพธ์ที่ต้องการ จึงจะเปลี่ยนข้อมูลให้ออกมาเป็นสารสนเทศที่ถูกต้องในรูปแบบที่ต้องการ เพื่อให้ผู้ใช้สามารถนำไปใช้งานต่อได้อย่างมีคุณภาพ เช่น ในการเรียกใช้ตามรายงานตัวอย่างในการทดลองต้องการตามรหัสกลุ่มโรค ซึ่งเป็นรหัสที่รวมรหัส ICD10 หลายรหัสเข้าด้วยกัน เช่น A00-A99 เป็นต้น

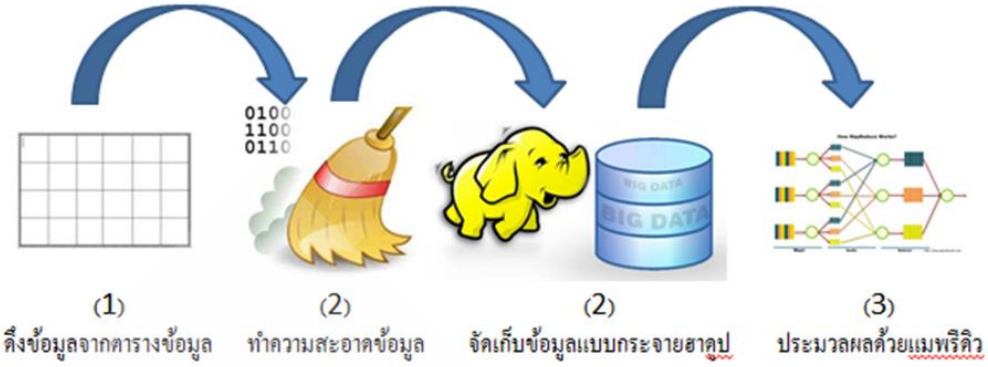
5.2.2 ผลของการค้นคืนแบบสอบถามข้อมูล ด้วยภาษาสอบถามแบบมีโครงสร้างเอสคิวแอล จะสังเกตุได้ว่าผลของครั้งที่ 1 ของการประมวลผลจากรายงานตัวอย่างที่ 1 และรายงานที่ 2 จะมีผลมากกว่าการประมวลผลครั้งที่ 2 และครั้งที่ เนื่องจากในระบบการจัดการข้อมูลเชิงสัมพันธ์มีการจัดการที่เรียกว่าแคช (Cache) ซึ่งในการทดลองได้มีการใช้คำสั่งเคลียร์ค่าแคชด้วยการใช้ Query Cache เป็นการเคลียร์ค่าที่หน่วยความจำสำรอง (Ram) แต่ยังคงมีผลกับการประมวลผลในครั้งแรก ทำให้การประมวลผลในครั้ง 1 ใช้เวลามากกว่าครั้งที่ 2 และครั้งที่ 3 ซึ่งในการทดสอบผู้วิจัยได้ทดสอบการเคลียร์ค่าแคชด้วยวิธีการรีสตาร์ทเครื่องเซิร์ฟเวอร์ใหม่ พบว่าในทุกครั้งที่มีการปิดและเปิดเครื่องใหม่ เมื่อการประมวลผลเริ่มต้นจะใช้เวลามากกว่าครั้งที่ 2 และ 3 อยู่เสมอ จึงเป็นสาเหตุในการทดสอบสมมติฐานทางสถิติเกิดความคลาดเคลื่อนได้ ผู้วิจัยจึงนำเสนอหากมีงานวิจัยเพื่อการเปรียบเทียบกับงานวิจัยนี้ ควรจะมีการปรับปรุงในส่วนของหน่วยความจำดังกล่าว และการกำหนดแผนการทดสอบ โดยที่ยังไม่ได้คำนึงถึงผลของปัจจัยอื่นๆ ของระบบฐานข้อมูลเชิงสัมพันธ์อาจจะทำให้ผลการทดสอบของสมมติฐานทางสถิติเกิดคลาดเคลื่อนได้ หรือหาสถิติวิจัยอื่นที่เหมาะสมมากกว่าเพื่อทำการวิจัยต่อไป

5.2.3 ในด้านความปลอดภัยของระบบจัดเก็บรูปแบบเท็กซ์ไฟล์เป็นอีกหนึ่งเรื่องที่ต้องนำมาพิจารณาในการป้องกันการเข้าถึง ควรกำหนดขอบเขตให้ผู้ใช้ ใช้ได้เฉพาะส่วนงานที่เกี่ยวข้องและควรใช้อยู่ในพื้นที่ศูนย์ข้อมูลจะปลอดภัยสูง การรักษาความปลอดภัยเป็นหนึ่งในประเด็นที่สำคัญที่สุดของการจัดเก็บข้อมูลสมัยใหม่ ประเภทของข้อมูลที่สำคัญหรือข้อมูลส่วนบุคคลสามารถที่เก็บไว้ใน NoSQL ควรมีความปลอดภัยจากหัวข้อที่ควรตรวจสอบดังนี้ 1.การรับรองความถูกต้อง (Authentication) 2.การให้สิทธิ์ (Authorization) 3.การตรวจสอบ (Auditing) 4.การเข้ารหัส

(Encryption) กลายเป็นความกังวลหลักสำหรับบริษัทที่จะเลือกเทคโนโลยีข้อมูลขนาดใหญ่ที่จะนำมาใช้นี้ (Gurevich, 2015, pp. 52-54) จึงสามารถนำมาเป็นงานการวิจัยเพิ่มเติมต่อไป

5.2.4 จากผลรวมค่าเฉลี่ยความเร็ว ผู้วิจัยสังเกตในขั้นตอนการทดสอบ เมื่อเทียบขั้นตอนการประมวลผลกันแล้วพบว่าหากใช้ขั้นตอนในการเขียนคำสั่งโปรแกรมแมพรีดิวให้เหมือนหรือใกล้เคียงกับขั้นตอนของการใช้ภาษาสอบถามเชิงโครงสร้างเอสคิวแอล จะเป็นการลดเวลาในขั้นตอนการประมวลผลได้ ในส่วนการจับคู่เชื่อมความสัมพันธ์ลงได้ อีกทั้งการเพิ่มการเรียงลำดับอินเด็กซ์ไว้ล่วงหน้า การสร้างคอลเลคชันชั่วคราวขึ้นมาก่อนที่จะเชื่อมกันนั้นจะทำให้การเชื่อมคอลเลคชั่นทำได้เร็วขึ้น (นิรุทธ์ รวยรื่น และ เกรียงไกร ปอแก้ว, 2557, น. 28) และการกำหนดขนาดบล็อกข้อมูลใน HDFS จะเป็นการเพิ่มประสิทธิภาพความเร็วได้ ในการกำหนดการใช้งานควรจะกำหนดคุณสมบัติของโปรแกรมการประมวลผลแบบกระจายฮาดูปไว้ที่ขนาดบล็อกไซต์ 128MB ซึ่งศึกษาวิธีการเพิ่มประสิทธิภาพเพิ่มเติมได้

5.2.5 ผู้วิจัยจึงขอนำเสนอแนวทางการนำไปใช้ร่วมกับข้อมูลบริการสุขภาพ เทคโนโลยีข้อมูลขนาดใหญ่สามารถนำมาใช้งานร่วมกับข้อมูลในระบบฐานข้อมูลเชิงสัมพันธ์แบบมีโครงสร้างได้ สามารถจัดเก็บข้อมูลและนำมาวิเคราะห์ข้อมูลระบบบริการสุขภาพได้ อ้างอิงจากผลการทดสอบในงานวิจัยนี้ที่พบว่าการจัดเก็บข้อมูลแบบกระจายฮาดูป และการประมวลผลแบบขนานแมพรีดิวนำมาใช้ประมวลผลข้อมูลบริการสุขภาพมีความถูกต้องแม่นยำของผลลัพธ์ สามารถนำจัดทำแบบ ETL (Extract Transform Load) ได้ หรือการแตกไฟล์เท็กไฟล์ออกและทำความสะอาดข้อมูลและทำการนำเข้าในระบบโปรแกรมฮาดูป ในปัจจุบันมีระบบเทคโนโลยีสนับสนุนหลายรูปแบบที่สามารถนำมาใช้การประมวลผลนอกเหนือจากแมพรีดิว เช่น ในระบบข้อมูลขนาดใหญ่หรือ Big Data Eco System เช่น HBase หรือ Hive หรือโปรแกรมอื่นๆ เช่น Spark หรือโปรแกรม Tajo เป็นโปรแกรมที่สามารถนำมาใช้เป็นแบบสอบถามค้นคืนข้อมูลร่วมกับเทคโนโลยีการจัดเก็บแบบกระจายฮาดูปได้ ที่สามารถใช้งานรูปแบบภาษาสอบถามเอสคิวแอลได้ ความพร้อมของซอร์ฟแวร์ที่จะนำมาใช้รองรับการทำงานให้เกิดประสิทธิภาพและเกิดประโยชน์มากที่สุด การเลือกแพลตฟอร์มที่เหมาะสมสำหรับการใช้งานเฉพาะขึ้นอยู่กับความต้องการใช้งานข้อมูลเฉพาะองค์กรหรืออาจจะใช้หลายแพลตฟอร์มร่วมกัน (Singh and Reddy, 2014)



ภาพที่ 5.1 แนวทางที่ผู้วิจัยนำเสนอการประมวลผลแบบ ETL (Extract Transform Load)

บริษัทขนาดใหญ่นำฮาดูปและแมพรีดิว นำไปใช้งานในหลากหลายลักษณะของงาน และมีประสิทธิภาพของงาน ด้วยวิธีการพัฒนาอัลกอริทึมให้มีความยืดหยุ่นและปรับขนาดขยายได้ในฮาดูป วัตถุประสงค์หลักการออกแบบเพื่อให้สามารถใช้งานข้อมูลแบบฐานข้อมูลใช้งานร่วมกับแมพรีดิวให้สามารถใช้งานได้ง่าย และยังสนับสนุนการเขียนภาษาสคริปต์เพื่อใช้งานแบบสอบถามเชิงโครงสร้างเอสคิวแอล ซึ่งพัฒนาอยู่ในกรอบของแมพรีดิว ซึ่งอย่างไรก็ตามวิธีการทั้งหมดนี้สามารถใช้งานได้ในรูปแบบที่แตกต่างกัน และลักษณะชุดข้อมูลที่แตกต่างกันก็มีผลในการเลือกใช้วิธีการ แต่ยังมีการใช้งานในแมพรีดิวร่วมกับสนับสนุนฐานข้อมูลได้ (Khanam & Agarwal, 2015, p. 124)

5.3 งานวิจัยในอนาคต

5.3.1 เทคโนโลยีการจัดเก็บแบบกระจายฮาดูปและการประมวลผลแมพรีดิว สามารถนำมาประยุกต์ใช้เพื่อการประมวลผลแบบทันทีเมื่อมีอินพุตข้อมูลเข้ามาใหม่ได้หรือไม่ โดยการนำข้อมูลระบบบริการสุขนำมาใช้งานเปรียบเทียบกับระบบเทคโนโลยีข้อมูลอื่น ที่มีความคุณสมบัติที่สามารถประมวลผลได้แบบทันที (Real-time processing) เช่น เทคโนโลยีข้อมูลขนาดใหญ่ประเภท NewSQL ที่สามารถรองรับกับภาษาสอบถามเชิงโครงสร้างเอสคิวแอลได้ หรือเทคโนโลยีข้อมูลอื่นๆ ที่สามารถใช้งานในระบบการจัดเก็บแบบกระจายฮาดูปได้

5.3.2 การทดลองครั้งนี้มีการเรียกใช้งานการประมวลผลที่มีตารางแฟ้มข้อมูลหลัก จำนวน 9 คอลัมน์ และแฟ้มข้อมูลเชื่อมสัมพันธ์ 1 คอลัมน์ และ 1 การเชื่อมสัมพันธ์ (Join) เท่านั้น การใช้เวลาในการประมวลผลอาจจะมากขึ้นตามลำดับหากมีตารางแฟ้มข้อมูลในการเชื่อมโยงมากกว่า 1 การเชื่อมโยงและหรือจำนวนคอลัมน์ในแฟ้มข้อมูลมีจำนวนมากขึ้น ควรมีการทดลองเพิ่มเติมในส่วน

ของการใช้แฟ้มข้อมูลที่มากกว่า 2 ตาราง และมีการเชื่อมความสัมพันธ์มากกว่า 1 ความสัมพันธ์เพื่อดูผลกระทบที่เกิดขึ้นจากการประมวลผลว่ามีผลต่อเทคโนโลยีข้อมูลขนาดใหญ่หรือไม่

5.3.3 เทคโนโลยีการจัดเก็บแบบกระจายฮาดูปมีวิธีการคำนวณค่าใช้จ่ายในการใช้ทรัพยากรหลัก เช่น ซีพียู และแรม และสตอเรจ หรือ I/O หรือไม่ สามารถนำมาใช้เพื่อหาความคุ้มค่าในการใช้งานได้หรือไม่ งานวิจัยในอนาคตควรจะใช้ทำการศึกษาเพิ่มเติมได้ ด้วยการเพิ่มจำนวนเครื่องเซิร์ฟเวอร์ในระบบเครือข่ายตามหลักการใช้งานจริงทั้งในศูนย์คอมพิวเตอร์ของกระทรวงและหรือของระดับเขตและระดับจังหวัดเพื่อสามารถนำมาประยุกต์ตามการใช้งานจริงได้

5.3.4 เนื่องจากรายงานตัวอย่างทั้ง 2 รายงาน ที่ใช้ประโยชน์ในการสืบค้น มีการออกรายงานเพียงปีละ 1 ครั้ง หรือเดือนละ 1 ครั้ง เท่านั้น ทำให้ดูเหมือนว่าเวลาที่ระบบการจัดการฐานข้อมูลมายเอสคิวแอลที่ประมวลผลด้วยภาษาสอบถามเชิงโครงสร้างเอสคิวแอลใช้ในการออกรายงานที่มากกว่า จะไม่เป็นปัญหากับหน่วยงานในสาธารณสุขเท่าใด จึงควรเลือกโจทย์เป็นรายงานที่ออกทุกวันหรือทุกสัปดาห์ เพื่อให้ทราบว่าจะมีผลกระทบมากกว่าในการปฏิบัติงานจริงหรือไม่ในงานวิจัยครั้งต่อไป

5.3.5 การทดลองแบบสอบถามอื่นๆ เพิ่มเติม แบบสอบถามที่มีความหลากหลายของการเรียกใช้รายงาน สมควรจะมีการทดลองเพิ่มเติมเนื่องจากในลักษณะการปฏิบัติงานจริงจะมีรายงานที่หลากหลายรูปแบบ กับคำสั่งที่ใช้สอบถามค้นค้นที่หลากหลายมากกว่า เพื่อเปรียบเทียบให้เห็นความแตกต่าง และประสิทธิภาพที่ชัดเจน เนื่องจากบางคำสั่งภาษาสอบถามเอสคิวแอลอาจจะทำได้รวดเร็วมากกว่า แต่ในบางสถานการณ์หรือบางรายงานหรือบางชุดแบบสอบถามอาจจะ ไม่เหมาะกับการใช้งานด้วยโปรแกรมการประมวลผลด้วยเทคนิคแมพรีดิว

บรรณานุกรม

บรรณานุกรม

ภาษาไทย

กระทรวงสาธารณสุข. สำนักบริหารการสาธารณสุข. (2555). *แผนพัฒนาระบบบริการสุขภาพ (Service Plan).* กรุงเทพฯ: ชุมชนสหกรณ์การเกษตร. น. 1.

กระทรวงสาธารณสุข. ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร. (2559). *ยุทธศาสตร์เทคโนโลยีสารสนเทศสุขภาพ ปี 2559–2556 (ฉบับร่างไม่มีตีพิมพ์).* สำนักงานปลัดกระทรวงสาธารณสุข, นนทบุรี. น. 58.

กระทรวงสาธารณสุข. สำนักนโยบายและยุทธศาสตร์. (2559). *คู่มือการปฏิบัติงานการจัดเก็บและจัดส่งข้อมูลตามโครงสร้างมาตรฐานข้อมูลสุขภาพ กระทรวงสาธารณสุข Version 2.1 (มกราคม 2559) ปีงบประมาณ 2559.* กรุงเทพฯ: เอสพี ก๊อปปี้ปริ้น. น. 36.

กระทรวงสาธารณสุข. สำนักนโยบายและยุทธศาสตร์. (2556). *สรุปรายงานการป่วย พ.ศ.2557.* นนทบุรี: องค์การสงเคราะห์ทหารผ่านศึก. น. 5-15.

กระทรวงสาธารณสุข. สำนักนโยบายและยุทธศาสตร์. (2553). *บัญชีจำแนกโรคระหว่างประเทศ ฉบับประเทศไทย (อังกฤษ-ไทย) ICD-10-TM for PCU : ตารางการจัดกลุ่มและดรรชนีรหัสหัตถการ.* นนทบุรี: องค์การสงเคราะห์ทหารผ่านศึก. น. 1.

เกรียงศักดิ์ เจริญวงศ์ศักดิ์. (2553). *การคิดเชิงวิเคราะห์* (พิมพ์ครั้งที่ 5). กรุงเทพฯ: ซัคเซสมีเดีย. น. 74-75.

โกสินทร์ จำนงไทย. (2559). *การทำวิจัยและเขียนบทความวิจัย ในสายวิศวกรรมศาสตร์ เทคโนโลยี และวิทยาศาสตร์.* กรุงเทพฯ: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย. น. 231.

ชาญชัย ศุภอรรถกร. (2557). *จัดการฐานข้อมูลด้วย MySQL ฉบับสมบูรณ์* (พิมพ์ครั้งที่ 5). กรุงเทพฯ: รีไวว่า. น. 125-126.

ชยาพร แก่นสาร์. (2555, กันยายน–ธันวาคม). การวิเคราะห์แผนการสืบค้นเพื่อประเมินประสิทธิภาพการทำงานของอ๊อพติไมเซอร์ที่มีต่อคำสั่งเอสคิวแอลแบบซีเล็กชัน. วารสารวิชาการพระจอมเกล้าพระนครเหนือ, 22(3), 721-734.

ชูพันธ์ รัตนโภคา. (2555, กันยายน–ธันวาคม). การออกแบบและพัฒนาระบบค้นหาข้อมูลจราจรทางคอมพิวเตอร์ด้วยวิธี Map/Reduce บนกรอบการทำงานของ Hadoop. วารสารวิชาการเทคโนโลยีอุตสาหกรรม, 8(3), 18-27.

นิรุทธ์ รวยรื่น, เกรียงไกร ปอแก้ว, (2557, เมษายน-มิถุนายน). *การใช้แมพรีดิวซ์เชื่อมคอลเลคชันของฐานข้อมูลโนเอสคิวแอลบนมองโกดีบี*. วารสารวิจัย มข.(บศ.), 14(2), 23-34.

นวรัตน์ สุวรรณผ่อง, มธุรส ทิพยมงคลกุล, ทองหล่อ เดชไทย, และนพพร โหวธีระกุล. (2557). *นโยบายสุขภาพ : การจัดทำ วิเคราะห์และประเมินผล*. นครปฐม: สำนักพิมพ์มหาวิทยาลัยมหิดล. น. 183-187.

ประกายมาศ ศรีสุขทักษิณ, ผสุดี บุญรอด, (2557, พฤษภาคม). *การเปรียบเทียบความเร็วในการประมวลผลระหว่างฐานข้อมูลเชิงสัมพันธ์และฐานข้อมูลไม่สัมพันธ์แบบเอกสาร*. The Tenth National Conference on Computing and Information Technology, 281-286.

ผุสดี บุญรอด, ประกายมาศ ศรีสุขทักษิณ, (2558, พฤษภาคม-สิงหาคม). *การค้นคืนข้อมูลขนาดใหญ่โดยใช้ภาษาสอบถามแบบไม่มีโครงสร้างร่วมกับเทคโนโลยีเว็บเชิงความหมาย*. วารสารวิชาการพระจอมเกล้าพระนครเหนือ, 25(2), 255-264.

เมธี จันท์จารุภรณ์. (2556). *การจัดการเชิงกลยุทธ์ในการพัฒนาสุขภาพ : หน่วยที่ 5 ข้อมูลและสารสนเทศเชิงกลยุทธ์* (พิมพ์ครั้งที่ 2). นนทบุรี: มหาวิทยลัยสุโขทัยธรรมาธิราช. น. 2.

มหาวิทยาลัยสุโขทัยธรรมาธิราช. บัณฑิตศึกษา . (2546). *การวิจัยทางสารสนเทศศาสตร์*. นนทบุรี: มหาวิทยาลัยสุโขทัยธรรมาธิราช. น. 313-350.

สุชาดา กีระนันทน์. (2544). *เทคโนโลยีสารสนเทศสถิติ : ข้อมูลในระบบสารสนเทศ* (พิมพ์ครั้งที่ 4). กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย. น. 94

สิน พันธุ์พินิจ. (2555). *เทคนิคการวิจัยทางวิทยาศาสตร์* (พิมพ์ครั้งที่ 2). กรุงเทพฯ: พิมพ์ดีการพิมพ์. น. 141-142.

โอภาส เอี่ยมสิริวงศ์. (2558). *ระบบฐานข้อมูล : ฉบับปรับปรุงเพิ่มเติม*. กรุงเทพฯ: ซีเอ็ดยูเคชั่น. น. 37-40.

โอภาส เอี่ยมสิริวงศ์, และสมโภชน์ ชื่นเอี่ยม. (2558). *คณิตศาสตร์คอมพิวเตอร์*. กรุงเทพฯ: วี.พริ้นท์(1991). น. 205-219.

ภาษาต่างประเทศ

Appuswamy R., & Gkantsidis C., & Narayanan D., Hodson O., & Rowstron, A. (2013). *Scale-up vs Scale-out for Hadoop: Time to rethink?*. SoCC'13. Retrieved June 27, 2016, from http://dl.acm.org/citation.cfm?id=2523629

Bhosale S. H., & Gadekar P. D. (2014). *A Review Paper on Big Data and Hadoop*. International Journal of Scientific and Research Publications, 4(10), 1-7.

Dean J., & Ghemawat S. (2004). *MapReduce: Simplied Data Processing on Large Clusters*. Google Inc., OSDI 2004, 51(1), 107-113. Retrieved June 28, 2016, from http://dl.acm.org/citation.cfm?id=1327492

Fegaras L., & Li C., & Gupta, U. (2012). *An Optimization Framework for Map-Reduce Queries*. EDBT2012. Retrieved June 28, 2016, from http://dl.acm.org/citation.cfm?id=2247601

Gunarathne, T., & Perera S. (2015). *Hadoop MapReduce v2 Cookbook Second Edition*. (2nd eds.) Birmingham, UK: Packt Publishing. pp. 60- 66.

Gurevich Y., (2015). *Comparative Survey of NoSQL / NewSQL DB Systems*. (Department Computer Science, The Open University of Israel, Ra'anana). 30-31. Retrieved July 3, 2016, from http://www.openu.ac.il/lists/mediaserver_documents/academic/cs/ComparativeSurvey.pdf

Hollingsworth R. M., (2012). *Hadoop and Hive as Scalable Alternatives to RDBMS: A Case Study*. (Department of Computer Science, Boise State University, Boise). Retrieved June 28, 2016, from http://scholarworks.boisestate.edu/cs_gradproj/2/

Khanam, Z., & Agarwal, S. (2015). *Map-Reduce Implementations: Survey And Performance Comparison.* International Journal of Computer Science & Information Technology (IJCSIT), 7(4), 119-126.

Miner, D., & Shook, A. (2012). *MapReduce Design Pattern*. CA: O'Reilly Media. pp. 4-7.

Sareen P., & Kumar P. (2015). *NoSQL Database and its Comparison With SQL Database*. International Journal of Computer Science & Communication Networks, 5(5), 293-298.

Schwartz, B., Zaitsev, P., & Tkachenko, V. (2012). *High Performance MySQL Third Edition.* CA: O'Reilly Media. pp. 210-238.

Singh D., & Reddy K. C. (2014). *A survey on platforms for big data analytics*. Journal of Big Data, 1(8), 1-20. Retrieved July 7, 2016, from http://www.journalofbigdata.com/content/1/1/8

Tao Y., & Lin W., & Xiao, X. (2013). *Minimal MapReduce Algorithms*. SIGMOD'13, 529-540. Retrieved June 28, 2016, from http://dl.acm.org/citation.cfm?id=2463719

Vicknair C., Macias M., Zhao Z., Nan X., Chen Y., & Wilkins, D. (2010). *A Comparison of a Graph Database and a Relational Database*. ACM SE'10, 15(17). Retrieved June 29, 2016, from http://dl.acm.org/citation.cfm?id=1900067

White, T. (2012). *Hadoop The Definitive Guide*. (3rd ed.). CA: O'Reilly Media. pp. 13–14.

ภาคผนวก

ภาคผนวก ก

รายงานการเจ็บป่วย พ.ศ.2537

ข้อมูลการป่วยผู้ป่วยนอก

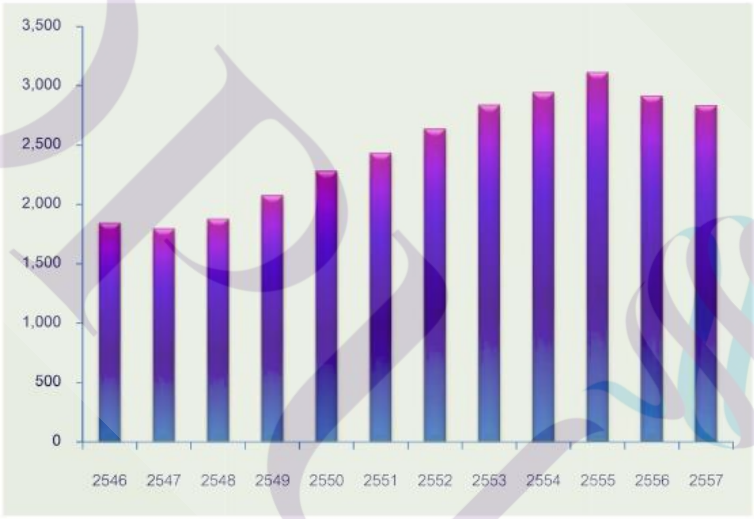
ข้อมูลการป่วยของผู้ป่วยนอก

ข้อมูลการป่วยของผู้ป่วยนอก

ในปี พ.ศ. 2557 การนำเสนอข้อมูลการป่วยของผู้ป่วยนอก ได้ประมวลผลจากฐานข้อมูลผู้ป่วยนอกรายบุคคล ตามรูปแบบ 21 แฟ้ม แต่การจำแนกกลุ่มโรคยังคงแบ่งเป็น 21 กลุ่มสาเหตุตาม รง.504 โดยแสดงข้อมูลการป่วยของผู้มารับบริการรักษาพยาบาลในภาพรวมระดับประเทศและภาคตามกลุ่มโรคต่าง ๆ ซึ่งสถานการณ์การเจ็บป่วยของผู้ป่วยนอก พ.ศ.2556 - 2557 จะพบว่ามีแนวโน้มโรคที่เปลี่ยนแปลงไป กล่าวคือ ยอดรวมของอัตราป่วยลดลง เนื่องจากพบว่าบางกลุ่มโรคมีอัตราป่วยลดลงซึ่งอาจเนื่องจากเดิมใช้รูปแบบการรายงาน ที่สำนักงานสาธารณสุขทุกจังหวัดรวบรวมหรือประมวลผลส่งเป็นรายเดือนมาเป็นส่วนกลางนำฐานข้อมูลมาตรวจสอบและประมวลผล นอกจากนี้ข้อจำกัดของข้อมูลผู้ป่วยนอก คือได้รับข้อมูลเฉพาะสถานบริการสาธารณสุขในสังกัดสำนักงานปลัดกระทรวงสาธารณสุขไม่มีข้อมูลผู้ป่วยของสถานบริการสังกัดอื่น ๆ และกรุงเทพมหานคร

สถานการณ์การป่วยจากรายงานผู้ป่วยนอก

ภาพ 1 อัตราผู้ป่วยนอก ต่อประชากร 1,000 คน พ.ศ.2546 – 2557



แหล่งข้อมูล ปี พ.ศ. 2546-2555 รายงานผู้ป่วยนอกตามกลุ่มสาเหตุ (รง.504) สำนักนโยบายและยุทธศาสตร์
ปี พ.ศ.2556-2557 ฐานข้อมูลผู้ป่วยนอกรายบุคคลตามรูปแบบ 21 แฟ้ม สำนักนโยบายและยุทธศาสตร์

อัตราผู้ป่วยนอกในภาพรวมประเทศ ตั้งแต่พ.ศ.2546 มีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่องในทุกปี โดยที่อัตราผู้ป่วยนอกในช่วงปี พ.ศ. 2549 - 2555 มีอัตราเพิ่มขึ้นมากกว่าในช่วงปี พ.ศ. 2546 – 2548 สำหรับปีพ.ศ.2556-2557 มีอัตราป่วยลดลงเล็กน้อย ซึ่งอาจมีผลจากการเปลี่ยนรูปแบบ การวิเคราะห์ข้อมูลเป็นการนำข้อมูลผู้ป่วยนอกรายบุคคลจากฐานข้อมูล 21 แฟ้มมาประมวลผลที่ส่วนกลางแทนการวิเคราะห์ข้อมูลจากรายงานที่สำนักงานสาธารณสุขทุกจังหวัดรวบรวมจัดส่งเป็นรายเดือนจากสถานบริการสาธารณสุขทุกแห่งในสังกัดเท่านั้น (ภาพที่ 1)

จากตารางที่ 1 อัตราผู้ป่วยนอกตามกลุ่มสาเหตุ 10 ลำดับแรกในช่วง 8 ปีที่ผ่านมา กลุ่มโรคที่มีอัตราผู้ป่วยนอกสูงเป็น ลำดับที่ 1 อันดับที่ 2 และ ลำดับที่ 3 จะอยู่ในกลุ่มโรคกลุ่มเดียวกัน ซึ่งเป็นโรคระบบทางเดินหายใจ โรคระบบไหลเวียนเลือด และกลุ่มโรคเกี่ยวกับต่อมไร้ท่อ โภชนาการและเมตาบอลิซึม

เมื่อพิจารณาตารางที่ 2 อัตราผู้ป่วยนอกตามกลุ่มสาเหตุ 10 ลำดับแรก รายภาค พ.ศ.2557 พบว่า โรคระบบไหลเวียนเลือดเป็นลำดับแรกในภาพรวมประเทศ เช่นเดียวกับภาคกลาง ภาคเหนือและภาคใต้ ส่วนภาคตะวันออกเฉียงเหนือ พบว่าโรคระบบทางเดินหายใจสูงเป็นลำดับแรก

ตาราง 1 10 ลำดับแรก อัตราผู้ป่วยนอก ต่อประชากร 1,000 คน ตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร) พ.ศ. 2550 - 2557

กลุ่มโรค	สาเหตุการป่วย (ชื่อโรค)	2550		2551		2552		2553		2554		2555		2556		2557	
		ลำดับ	อัตรา	ลำดับ	อัตรา	ลำดับ	อัตรา	ลำดับ	อัตรา	ลำดับ	อัตรา	ลำดับ	อัตรา	ลำดับ	อัตรา	ลำดับ	อัตรา
10	โรคระบบหายใจ (J00-J99)	1	444.09	1	457.41	1	498.16	1	499.53	1	489.16	1	473.34	1	418.25	3	409.98
11	โรคระบบย่อยอาหาร รวมโรคในช่องปาก (K00-K93)	2	263.10	3	282.66	4	288.64	3	312.51	4	333.54	5	333.54	6	326.92	5	323.44
9	โรคระบบไหลเวียนเลือด (I00-I99)	3	255.76	2	283.78	2	306.42	2	345.03	2	398.02	2	425.72	2	414.01	1	438.34
13	โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม (M00-M99)	4	237.92	4	264.16	3	290.73	4	311.12	5	333.39	4	354.91	4	351.66	4	345.13
4	โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม (E00-E90)	5	188.58	5	218.10	5	246.72	5	288.35	3	348.37	3	385.76	3	390.52	2	414.58
1	โรคติดเชื้อและปรสิต (A00-A99,B00-B99)	6	119.16	6	122.60	6	129.35	6	131.94	6	128.55	6	127.53	7	116.71	9	86.97
12	โรคผิวหนังและเนื้อเยื่อใต้ผิวหนัง (L00-L99)	7	89.10	7	91.68	7	95.53	7	104.29	7	102.92	7	104.93	9	98.64	7	92.26
14	โรคระบบสืบพันธุ์รวมปัสสาวะ (N00-N99)	8	63.97	8	70.86	8	74.36	8	86.07	8	96.07	8	103.33	8	104.70	10	81.25
7	โรคตารวมส่วนประกอบของตา (H00-H59)	9	59.86	9	61.86	9	70.42	9	71.85	9	76.70	9	80.67	10	75.62	8	88.88
5	ภาวะแปรปรวนทางจิตและพฤติกรรม (F00-F99)	10	42.27	10	45.89	10	49.00	10	57.21	10	60.74	10	60.74	12	59.77	12	50.76

ตาราง 2 10 ลำดับแรก อัตราผู้ป่วยนอก ต่อประชากร 1,000 คน ตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร) พ.ศ. 2557

ลำดับ	กลุ่มโรค	สาเหตุการป่วย	ทั้งประเทศ	ภาคเหนือ	ภาคตะวันออกเฉียงเหนือ	ภาคกลาง	ภาคใต้
1	9	โรคระบบไหลเวียนเลือด(I00-I99)	438.34	594.67	377.60	435.95	385.30
2	4	โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม(E00-E90)	414.58	488.74	416.50	398.98	342.29
3	10	โรคระบบหายใจ(J00-J99)	409.98	418.54	457.60	364.33	367.56
4	13	โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม(M00-M99)	345.13	424.54	392.15	291.89	226.30
5	11	โรคระบบย่อยอาหาร รวมโรคในช่องปาก(K00-K93)	323.44	353.67	370.05	273.78	262.68
6	12	โรคผิวหนังและเนื้อเยื่อใต้ผิวหนัง(L00-L99)	92.26	98.81	96.05	85.57	86.81
7	7	โรคตารวมส่วนประกอบของตา(H00-H59)	88.88	85.17	111.44	74.26	66.21
8	1	โรคติดเชื้อและปรสิต(A00 - A99,B00 - B99)	86.97	87.68	97.78	73.48	84.52
9	14	โรคระบบสืบพันธุ์รวมปัสสาวะ(N00 - N99)	81.25	97.73	85.09	66.66	77.01
10	5	ภาวะแปรปรวนทางจิตและพฤติกรรม(F00-F99)	50.76	62.84	52.56	42.03	46.58

แหล่งข้อมูล ปี พ.ศ. 2550-2555 รายงานผู้ป่วยนอกตามกลุ่มสาเหตุ (ร.ง.504) สำนักนโยบายและยุทธศาสตร์
ปี พ.ศ.2556-2557 ฐานข้อมูลผู้ป่วยนอกรายบุคคลตามรูปแบบ 21 แฟ้ม สำนักนโยบายและยุทธศาสตร์

หมายเหตุ การจัดลำดับโรค ไม่นำกลุ่มโรคที่ 18 และ21 มาจัดลำดับ (ดูรายชื่อกลุ่มโรคอาด แบบรายงานผู้ป่วยนอกตามกลุ่มสาเหตุ ในภาคผนวก)

ตาราง 3 จำนวน และอัตราผู้ป่วยนอก อัตราต่อประชากร 1,000 คน ทั้งประเทศ และรายภาค (ไม่รวมกรุงเทพมหานคร) พ.ศ.2557

กลุ่มโรค	สาเหตุการป่วย	ภาคเหนือ		ภาคตะวันออกเฉียงเหนือ		ภาคกลาง		ภาคใต้		ทั้งประเทศ	
		จำนวน	อัตรา	จำนวน	อัตรา	จำนวน	อัตรา	จำนวน	อัตรา	จำนวน	อัตรา
	รวม	**38,927,121**	**3,288.79**	**64,638,007**	**2,963.64**	**42,366,037**	**2,575.53**	**22,194,732**	**2,420.37**	**168,125,897**	**2,836.80**
1	โรคติดเชื้อและปรสิต (A00-A99,B00-B99) Certain infectious and parasitic diseases	1,037,818	87.68	2,132,704	97.78	1,208,753	73.48	775,078	84.52	5,154,353	86.97
2	เนื่องอก (รวมมะเร็ง) (C00-C97,D00-D84) Neoplasma	321,918	27.20	580,449	26.61	333,902	20.30	183,527	20.01	1,419,796	23.96
3	โรคเลือดและอวัยวะสร้างเลือด และความผิดปกติเกี่ยวกับภูมิคุ้มกัน (D50-D89) Diseases of the blood and blood forming organs and certain disorders involving tne immune mechanism	249,947	21.12	447,202	20.50	261,788	15.91	133,188	14.52	1,092,125	18.43
4	โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม (E00-E90) Endocrine, mutritional and metabolic diseases	5,784,858	488.74	9,084,033	416.50	6,562,969	398.98	3,138,765	342.29	24,570,625	414.58
5	ภาวะแปรปรวนทางจิตและพฤติกรรม (F00-F99) Mental and behavioural disorders	743,808	62.84	1,146,290	52.56	691,432	42.03	427,099	46.58	3,008,629	50.76
6	โรคระบบประสาท (G00-G99) Diseases of the mervous system	702,731	59.37	1,023,164	46.91	851,238	51.75	355,008	38.71	2,932,141	49.47
7	โรคตารวมส่วนประกอบของตา (H-H59) Diseases of the eye and adnexa	1,008,047	85.17	2,430,600	111.44	1,221,584	74.26	607,133	66.21	5,267,364	88.88
8	โรคหูและปุ่มกกหู (H60-H95) Diseaese of the ear and mastoid process	391,433	33.07	434,919	19.94	322,404	19.60	179,905	19.62	1,328,661	22.42
9	โรคระบบไหลเวียนเลือด (I00-I99) Diseases of the circulatory system	7,038,721	594.67	8,235,599	377.60	7,171,103	435.95	3,533,143	385.30	25,978,566	438.34
10	โรคระบบหายใจ (J00-J99) Diseases of the respiratory system	4,953,916	418.54	9,980,511	457.60	5,993,062	364.33	3,370,548	367.56	24,298,037	409.98
11	โรคระบบย่อยอาหาร รวมโรคในช่องปาก (K00-K93) Diseases of the digestive system	4,186,092	353.67	8,070,888	370.05	4,503,532	273.78	2,408,729	262.68	19,169,241	323.44
12	โรคผิวหนังและเนื้อเยื่อใต้ผิวหนัง (L00-L99) Diseases of the skin and subcutaneous tissue	1,169,582	98.81	2,094,814	96.05	1,407,544	85.57	796,061	86.81	5,468,001	92.26
13	โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม (M00-M99) Diseases of the musculoskeletal system and connective tissue	5,024,961	424.54	8,553,013	392.15	4,801,435	291.89	2,075,203	226.30	20,454,612	345.13
14	โรคระบบสืบพันธุ์รวมปัสสาวะ (N00-N99) Diseases of the genitourinary system	1,156,742	97.73	1,855,879	85.09	1,096,530	66.66	706,188	77.01	4,815,339	81.25
15	ภาวะแทรกในการตั้งครรภ์ การคลอด และระยะหลังคลอด (O00-O99) Complication of pregnancy, childbirth and the puerperium	89,747	7.58	226,906	10.40	175,953	10.70	156,290	17.04	648,896	10.95
16	ภาวะผิดปกติของทารกแรกเกิดขึ้นในระยะปริกำเนิด (อายุครรภ์ 22 สัปดาห์ขึ้นไป จนถึง 7 วันหลังคลอด) (P00-P96) Certain conditions originating in the perinatal period	27,985	2.36	74,342	3.41	37,057	2.25	14,461	1.58	153,845	2.60

ตาราง 3 จำนวน และอัตราผู้ป่วยนอก อัตราต่อประชากร 1,000 คน ทั้งประเทศ และรายภาค (ไม่รวมกรุงเทพมหานคร) พ.ศ.2557 (ต่อ)

กลุ่มโรค	สาเหตุการป่วย	ภาคเหนือ		ภาคตะวันออกเฉียงเหนือ		ภาคกลาง		ภาคใต้		ทั้งประเทศ	
		จำนวน	อัตรา	จำนวน	อัตรา	จำนวน	อัตรา	จำนวน	อัตรา	จำนวน	อัตรา
17	รูปร่างผิดปกติแต่กำเนิด การพิการจนผิดรูปแต่กำเนิดและโครโมโซมผิดปกติ (Q00-Q99) Congenital malformations, deformations and chromosomal abnormalities	33,866	2.86	79,613	3.65	45,313	2.75	34,109	3.72	192,901	3.25
18	อาการ, อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการที่ไม่สามารถจำแนกโรคในกลุ่มอื่นๆ ได้ (R00-R99) Symptoms, signs and abnormal clinical and labooratory findings, not elsewhere classified	4,063,943	343.35	6,394,250	293.18	4,489,426	272.92	2,636,251	287.49	17,583,870	296.69
19	การเป็นพิษและผลที่ตามมา (X00-X49, X60-X69) Poisoning, toxic effect, and their sequelae	16,009	1.35	24,555	1.13	15,675	0.95	7,356	0.80	63,595	1.07
20	อุบัติเหตุจากการขนส่ง และผลที่ตามมา (V01-V99,Y85) Transport accidents and their sequelae	178,687	15.10	381,501	17.49	258,905	15.74	145,726	15.89	964,819	16.28
21	สาเหตุภายนอกอื่น ๆ ที่ทำให้ป่วยหรือตาย (W00-W99, X00-X19, X20-X29, X30-X39, X50-X59, X70-X84, X91-X99, Y00-Y09, Y20-Y39, Y40-Y84, Y86-Y89) Other external causes of morbidity and mortality (eg: accidents, injuries, intentional self-harm, assault, animals and plants, complications of medical and surgical care and other unspecified causes)	746,310	63.05	1,386,775	63.58	916,432	55.71	510,964	55.72	3,560,481	60.08

แหล่งข้อมูล ฐานข้อมูลผู้ป่วยนอกรายบุคคลตามรูปแบบ 21 แฟ้ม สำนักนโยบายและยุทธศาสตร์

ภาคผนวก ข

ผลลัพธ์การทดลองแบบสอบถามค้นคืนข้อมูล

1. การแสดงผลลัพธ์ SQL รายงาน 1 จำนวนข้อมูล 500,000 ระเบียน

ผลลัพธ์ SQL

โฮสต์: localhost
ฐานข้อมูล: helpcare
เวลาในการสร้าง: 17 ก.ค. 2016 17:16น.
สร้างโดย: phpMyAdmin 4.0.10deb1 / MySQL 5.5.49-0ubuntu0.14.04.1
คำสั่ง SQL: SELECT opd_diag.opd_diag3 AS กลุ่มโรค, opd_diag.opd_diag4 AS สาเหตุกลุ่มการป่วย, COUNT(opd_diag.opd_diag3) AS จำนวนผู้ป่วยนอก FROM diagnosis_opd_5h, opd_diag WHERE LEFT(diagnosis_opd_5h.diagcode, 3) = opd_diag.opd_diag1 GROUP BY opd_diag.opd_diag3 ORDER BY จำนวนผู้ป่วยนอก DESC LIMIT 10;
แถว: 10

This table does not contain a unique column. Grid edit, checkbox, Edit, Copy and Delete features are not available.

กลุ่มโรค	สาเหตุกลุ่มการป่วย	จำนวนผู้ป่วยนอก
4	โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม	40475
9	โรคระบบไหลเวียนเลือด	38240
10	โรคระบบหายใจ	32872
11	โรคระบบย่อยอาหาร รวมโรคในช่องปาก	24665
13	โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม	22666
18	อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการที่ไม่สามารถจำแนกโรคในกลุ่มอื่	19362
5	ภาวะแปรปรวนทางจิตและพฤติกรรม	10265
1	โรคติดเชื้อและปรสิต	8916
14	โรคระบบสืบพันธ์ร่วมปัสสาวะ	8319
7	โรคตารวมส่วนประกอบของตา	6943

2. การแสดงผลลัพธ์ MapReduce รายงาน 1 จำนวนข้อมูล 500,000 ระเบียน

ผลลัพธ์ SQL

โฮสต์: localhost
ฐานข้อมูล: helpcare
เวลาในการสร้าง: 17 ก.ค. 2016 17:18น.
สร้างโดย: phpMyAdmin 4.0.10deb1 / MySQL 5.5.49-0ubuntu0.14.04.1
คำสั่ง SQL: SELECT opd_diag.opd_diag3 AS กลุ่มโรค, opd_diag.opd_diag4 AS สาเหตุกลุ่มการป่วย, COUNT(opd_diag.opd_diag3) AS จำนวนผู้ป่วยนอก FROM diagnosis_opd_5h, opd_diag WHERE LEFT(diagnosis_opd_5h.diagcode, 3) = opd_diag.opd_diag1 GROUP BY opd_diag.opd_diag3 ORDER BY จำนวนผู้ป่วยนอก DESC LIMIT 0, 30 ;
แถว: 21

This table does not contain a unique column. Grid edit, checkbox, Edit, Copy and Delete features are not available.

กลุ่มโรค	สาเหตุกลุ่มการป่วย	จำนวนผู้ป่วยนอก
4	โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม	40475
9	โรคระบบไหลเวียนเลือด	38240
10	โรคระบบหายใจ	32872
11	โรคระบบย่อยอาหาร รวมโรคในช่องปาก	24665
13	โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม	22666
18	อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการที่ไม่สามารถจำแนกโรคในกลุ่มอื่	19362
5	ภาวะแปรปรวนทางจิตและพฤติกรรม	10265
1	โรคติดเชื้อและปรสิต	8916
14	โรคระบบสืบพันธ์ร่วมปัสสาวะ	8319
7	โรคตารวมส่วนประกอบของตา	6943
12	โรคผิวหนังและเนื้อเยื่อใต้ผิวหนัง	6919
21	สาเหตภายนอกอื่น ๆ ที่ทำให้ป่วยหรือตาย	4295
6	โรคระบบประสาท	4254
8	โรคหูและปุ่มกกหู	2010
3	โรคเลือดและอวัยวะสร้างเลือด และความผิดปกติเกี่ยวกับภูมิคุ้มกัน	1732
2	เนื้องอก	1691
15	ภาวะแทรกในการตั้งครรภ์ การคลอด และระยะหลังคลอด	1125
20	อุบัติเหตุจากการขนส่ง และผลที่ตามมา	1035
16	ภาวะผิดปกติของทารกแรกเกิดขึ้นในระยะปริกำหนด	355
17	รูปร่างผิดปกติแต่กำเนิด การพิการจนผิดรูปแต่กำเนิดและโครโมโซมผิดปกติ	263
19	การเป็นพิษและผลที่ตามมา	47

3. การแสดงผลลัพธ์ MapReduce รายงาน 1 และ รายงาน 2 จำนวนข้อมูล 500,000 ระเบียน

ICD-10	Grou	Total
โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม	4	40,475.00
โรคระบบไหลเวียนเลือด	9	38,240.00
โรคระบบหายใจ	10	32,872.00
โรคระบบย่อยอาหาร รวมโรคในช่องปาก	11	24,665.00
โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม	13	22,666.00
อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการที่ไม่สามารถจำแนกโรคในกลุ่มอื่	18	19,362.00
ภาวะแปรปรวนทางจิตและพฤติกรรม	5	10,265.00
โรคติดเชื้อและปรสิต	1	8,916.00
โรคระบบสืบพันธ์ร่วมปัสสาวะ	14	8,319.00
โรคตารวมส่วนประกอบของตา	7	6,943.00
โรคผิวหนังและเนื้อเยื่อใต้ผิวหนัง	12	6,919.00
สาเหตภายนอกอื่น ๆ ที่ทำให้ป่วยหรือตาย	21	4,295.00
โรคระบบประสาท	6	4,254.00
โรคหูและปุ่มกกหู	8	2,010.00
โรคเลือดและอวัยวะสร้างเลือด และความผิดปกติเกี่ยวกับภูมิคุ้มกัน	3	1,732.00
เนื้องอก	2	1,691.00
ภาวะแทรกในการตั้งครรภ์ การคลอด และระยะหลังคลอด	15	1,125.00
อุบัติเหตุจากการขนส่ง และผลที่ตามมา	20	1,035.00
ภาวะผิดปกติของทารกแรกเกิดขึ้นในระยะปริกำหนด	16	355.00
รูปร่างผิดปกติแต่กำเนิด การพิการจนผิดรูปแต่กำเนิดและ โครโมโซมผิดปกติ	17	263.00
การเป็นพิษและผลที่ตามมา	19	47.00

4. การแสดงผลลัพธ์ SQL รายงาน 1 จำนวนข้อมูล 1,000,000 ระเบียน

ผลลัพธ์ SQL

โฮสต์: localhost
ฐานข้อมูล: helpcare
เวลาในการสร้าง: 17 ก.ค. 2016 17:15น.
สร้างโดย: phpMyAdmin 4.0.10deb1 / MySQL 5.5.49-0ubuntu0.14.04.1
คำสั่ง SQL: SELECT opd_diag.opd_diag3 AS กลุ่มโรค, opd_diag.opd_diag4 AS สาเหตุกลุ่มการป่วย, COUNT(opd_diag.opd_diag3) AS จำนวนผู้ป่วยนอก FROM diagnosis_opd_1m, opd_diag WHERE LEFT(diagnosis_opd_1m.diagcode, 3) = opd_diag.opd_diag1 GROUP BY opd_diag.opd_diag3 ORDER BY จำนวนผู้ป่วยนอก DESC LIMIT 10;
แถว: 10

This table does not contain a unique column. Grid edit, checkbox, Edit, Copy and Delete features are not available.

กลุ่มโรค	สาเหตุกลุ่มการป่วย	จำนวนผู้ป่วยนอก
4	โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม	80330
9	โรคระบบไหลเวียนเลือด	75508
10	โรคระบบหายใจ	66079
11	โรคระบบย่อยอาหาร รวมโรคในช่องปาก	49333
13	โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม	45231
18	อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการที่ไม่สามารถจำแนกโรคในกลุ่มอื่	37941
5	ภาวะแปรปรวนทางจิตและพฤติกรรม	20439
1	โรคติดเชื้อและปรสิต	17652
14	โรคระบบสืบพันธ์ร่วมปัสสาวะ	17144
7	โรคตารวมส่วนประกอบของตา	13972

5. การแสดงผลลัพธ์ SQL รายงาน 2 จำนวนข้อมูล 1,000,000 ระเบียน

ผลลัพธ์ SQL

โฮสต์: localhost
ฐานข้อมูล: helpcare
เวลาในการสร้าง: 17 ก.ค. 2016 17:18น.
สร้างโดย: phpMyAdmin 4.0.10deb1 / MySQL 5.5.49-0ubuntu0.14.04.1
คำสั่ง SQL: SELECT opd_diag.opd_diag3 AS กลุ่มโรค, opd_diag.opd_diag4 AS สาเหตุกลุ่มการป่วย, COUNT(opd_diag.opd_diag3) AS จำนวนผู้ป่วยนอก FROM diagnosis_opd_1m, opd_diag WHERE LEFT(diagnosis_opd_1m.diagcode, 3) = opd_diag.opd_diag1 GROUP BY opd_diag.opd_diag3 ORDER BY จำนวนผู้ป่วยนอก DESC LIMIT 0, 30 ;
แถว: 21

This table does not contain a unique column. Grid edit, checkbox, Edit, Copy and Delete features are not available.

กลุ่มโรค	สาเหตุกลุ่มการป่วย	จำนวนผู้ป่วยนอก
4	โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม	80330
9	โรคระบบไหลเวียนเลือด	75508
10	โรคระบบหายใจ	66079
11	โรคระบบย่อยอาหาร รวมโรคในช่องปาก	49333
13	โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม	45231
18	อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการที่ไม่สามารถจำแนกโรคในกลุ่มอื่	37941
5	ภาวะแปรปรวนทางจิตและพฤติกรรม	20439
1	โรคติดเชื้อและปรสิต	17652
14	โรคระบบสืบพันธ์ร่วมปัสสาวะ	17144
7	โรคตารวมส่วนประกอบของตา	13972
12	โรคผิวหนังและเนื้อเยื่อใต้ผิวหนัง	13574
21	สาเหตภายนอกอื่น ๆ ที่ทำให้ป่วยหรือตาย	8795
6	โรคระบบประสาท	8451
8	โรคหูและปุ่มกกหู	3947
3	โรคเลือดและอวัยวะสร้างเลือด และความผิดปกติเกี่ยวกับภูมิคุ้มกัน	3435
2	เนื้องอก	3405
15	ภาวะแทรกในการตั้งครรภ์ การคลอด และระยะหลังคลอด	2181
20	อุบัติเหตุจากการขนส่ง และผลที่ตามมา	1995
16	ภาวะผิดปกติของทารกแรกเกิดขึ้นในระยะปริกำหนด	790
17	รูปร่างผิดปกติแต่กำเนิด การพิการจนผิดรูปแต่กำเนิดและโครโมโซมผิดปกติ	600
19	การเป็นพิษและผลที่ตามมา	97

6. การแสดงผลลัพธ์ MapReduce รายงาน 1 และ รายงาน 2 จำนวนข้อมูล 1,000,000 ระเบียน

ICD-10	Group	Total
โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม	4	80,330
โรคระบบไหลเวียนเลือด	9	75,508
โรคระบบหายใจ	10	66,079
โรคระบบย่อยอาหาร รวมโรคในช่องปาก	11	49,333
โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม	13	45,231
อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการที่ไม่สามารถจำแนกโรคในกลุ่มอื่	18	37,941
ภาวะแปรปรวนทางจิตและพฤติกรรม	5	20,439
โรคติดเชื้อและปรสิต	1	17,652
โรคระบบสืบพันธ์ร่วมปัสสาวะ	14	17,144
โรคตารวมส่วนประกอบของตา	7	13,972
โรคผิวหนังและเนื้อเยื่อใต้ผิวหนัง	12	13,574
สาเหตภายนอกอื่น ๆ ที่ทำให้ป่วยหรือตาย	21	8,795
โรคระบบประสาท	6	8,451
โรคหูและปุ่มกกหู	8	3,947
โรคเลือดและอวัยวะสร้างเลือด และความผิดปกติเกี่ยวกับภูมิคุ้มกัน	3	3,435
เนื้องอก	2	3,405
ภาวะแทรกในการตั้งครรภ์ การคลอด และระยะหลังคลอด	15	2,181
อุบัติเหตุจากการขนส่ง และผลที่ตามมา	20	1,995
ภาวะผิดปกติของทารกแรกเกิดขึ้นในระยะปริกำหนด	16	790
รูปร่างผิดปกติแต่กำเนิด การพิการจนผิดรูปแต่กำเนิดและ โครโมโซมผิดปกติ	17	600
การเป็นพิษและผลที่ตามมา	19	97

7. การแสดงผลลัพธ์ SQL รายงาน 1 จำนวนข้อมูล 5,000,000 ระเบียน

ผลลัพธ์ SQL

โฮสต์: localhost
ฐานข้อมูล: helpcare
เวลาในการสร้าง: 17 ก.ค. 2016 17:16น.
สร้างโดย: phpMyAdmin 4.0.10deb1 / MySQL 5.5.49-0ubuntu0.14.04.1
คำสั่ง SQL: SELECT opd_diag.opd_diag3 AS กลุ่มโรค, opd_diag.opd_diag4 AS สาเหตุกลุ่มการป่วย, COUNT(opd_diag.opd_diag3) AS จำนวนผู้ป่วยนอก FROM diagnosis_opd_5m, opd_diag WHERE LEFT(diagnosis_opd_5m.diagcode, 3) = opd_diag.opd_diag1 GROUP BY opd_diag.opd_diag3 ORDER BY จำนวนผู้ป่วยนอก DESC LIMIT 10;
แถว: 10

This table does not contain a unique column. Grid edit, checkbox, Edit, Copy and Delete features are not available.

กลุ่มโรค	สาเหตุกลุ่มการป่วย	จำนวนผู้ป่วยนอก
4	โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม	402100
9	โรคระบบไหลเวียนเลือด	378424
10	โรคระบบหายใจ	329670
11	โรคระบบย่อยอาหาร รวมโรคในช่องปาก	245614
13	โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม	226363
18	อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการที่ไม่สามารถจำแนกโรคในกลุ่มอื่	190885
5	ภาวะแปรปรวนทางจิตและพฤติกรรม	102636
1	โรคติดเชื้อและปรสิต	87787
14	โรคระบบสืบพันธ์ร่วมปัสสาวะ	84858
7	โรคตารวมส่วนประกอบของตา	69832

8. การแสดงผลลัพธ์ SQL รายงาน 2 จำนวนข้อมูล 5,000,000 ระเบียน

ผลลัพธ์ SQL

โฮสต์: localhost
ฐานข้อมูล: helpcare
เวลาในการสร้าง: 17 ก.ค. 2016 17:19น.
สร้างโดย: phpMyAdmin 4.0.10deb1 / MySQL 5.5.49-0ubuntu0.14.04.1
คำสั่ง SQL: SELECT opd_diag.opd_diag3 AS กลุ่มโรค, opd_diag.opd_diag4 AS สาเหตุกลุ่มการป่วย, COUNT(opd_diag.opd_diag3) AS จำนวนผู้ป่วยนอก FROM diagnosis_opd_5m, opd_diag WHERE LEFT(diagnosis_opd_5m.diagcode, 3) = opd_diag.opd_diag1 GROUP BY opd_diag.opd_diag3 ORDER BY จำนวนผู้ป่วยนอก DESC LIMIT 0, 30 ;
แถว: 21

This table does not contain a unique column. Grid edit, checkbox, Edit, Copy and Delete features are not available.

กลุ่มโรค	สาเหตุกลุ่มการป่วย	จำนวนผู้ป่วยนอก
4	โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม	402100
9	โรคระบบไหลเวียนเลือด	378424
10	โรคระบบหายใจ	329670
11	โรคระบบย่อยอาหาร รวมโรคในช่องปาก	245614
13	โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม	226363
18	อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการที่ไม่สามารถจำแนกโรคในกลุ่มอื่	190885
5	ภาวะแปรปรวนทางจิตและพฤติกรรม	102636
1	โรคติดเชื้อและปรสิต	87787
14	โรคระบบสืบพันธ์ร่วมปัสสาวะ	84858
7	โรคตารวมส่วนประกอบของตา	69832
12	โรคผิวหนังและเนื้อเยื่อใต้ผิวหนัง	68727
21	สาเหตภายนอกอื่น ๆ ที่ทำให้ป่วยหรือตาย	43910
6	โรคระบบประสาท	42980
8	โรคหูและปุ่มกกหู	19171
3	โรคเลือดและอวัยวะสร้างเลือด และความผิดปกติเกี่ยวกับภูมิคุ้มกัน	17340
2	เนื้องอก	16922
15	ภาวะแทรกในการตั้งครรภ์ การคลอด และระยะหลังคลอด	11142
20	อุบัติเหตุจากการขนส่ง และผลที่ตามมา	10130
16	ภาวะผิดปกติของทารกแรกเกิดขึ้นในระยะปริกำหนด	3822
17	รูปร่างผิดปกติแต่กำเนิด การพิการจนผิดรูปแต่กำเนิดและโครโมโซมผิดปกติ	2933
19	การเป็นพิษและผลที่ตามมา	469

9. การแสดงผลลัพธ์ MapReduce รายงาน 1 และ รายงาน 2 จำนวนข้อมูล 5,000,000 ระเบียน

ICD-10	Group	Total
โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม	4	402,100
โรคระบบไหลเวียนเลือด	9	378,424
โรคระบบหายใจ	10	329,670
โรคระบบย่อยอาหาร รวมโรคในช่องปาก	11	245,614
โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม	13	226,363
อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการที่ไม่สามารถจำแนกโรคในกลุ่มอื่	18	190,885
ภาวะแปรปรวนทางจิตและพฤติกรรม	5	102,636
โรคติดเชื้อและปรสิต	1	87,787
โรคระบบสืบพันธ์ร่วมปัสสาวะ	14	84,858
โรคตารวมส่วนประกอบของตา	7	69,832
โรคผิวหนังและเนื้อเยื่อใต้ผิวหนัง	12	68,727
สาเหตภายนอกอื่น ๆ ที่ทำให้ป่วยหรือตาย	21	43,910
โรคระบบประสาท	6	42,980
โรคหูและปุ่มกกหู	8	19,171
โรคเลือดและอวัยวะสร้างเลือด และความผิดปกติเกี่ยวกับภูมิคุ้มกัน	3	17,340
เนื้องอก	2	16,922
ภาวะแทรกในการตั้งครรภ์ การคลอด และระยะหลังคลอด	15	11,142
อุบัติเหตุจากการขนส่ง และผลที่ตามมา	20	10,130
ภาวะผิดปกติของทารกแรกเกิดขึ้นในระยะปริกำหนด	16	3,822
รูปร่างผิดปกติแต่กำเนิด การพิการจนผิดรูปแต่กำเนิดและโครโมโซมผิดปกติ	17	2,933
การเป็นพิษและผลที่ตามมา	19	469

10. การแสดงผลลัพธ์ SQL รายงาน 1 จำนวนข้อมูล 10,000,000 ระเบียน

ผลลัพธ์ SQL

โฮสต์: localhost
ฐานข้อมูล: helpcare
เวลาในการสร้าง: 17 ก.ค. 2016 17:14น.
สร้างโดย: phpMyAdmin 4.0.10deb1 / MySQL 5.5.49-0ubuntu0.14.04.1
คำสั่ง SQL: SELECT opd_diag.opd_diag3 AS กลุ่มโรค, opd_diag.opd_diag4 AS สาเหตุกลุ่มการป่วย, COUNT(opd_diag.opd_diag3) AS จำนวนผู้ป่วยนอก FROM diagnosis_opd_10m, opd_diag WHERE LEFT(diagnosis_opd_10m.diagcode, 3) = opd_diag.opd_diag1 GROUP BY opd_diag.opd_diag3 ORDER BY จำนวนผู้ป่วยนอก DESC LIMIT 10;
แถว: 10

This table does not contain a unique column. Grid edit, checkbox, Edit, Copy and Delete features are not available.

กลุ่มโรค	สาเหตุกลุ่มการป่วย	จำนวนผู้ป่วยนอก
4	โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม	803920
9	โรคระบบไหลเวียนเลือด	756420
10	โรคระบบหายใจ	661610
11	โรคระบบย่อยอาหาร รวมโรคในช่องปาก	492848
13	โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม	451449
18	อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการที่ไม่สามารถจำแนกโรคในกลุ่มอื่	382004
5	ภาวะแปรปรวนทางจิตและพฤติกรรม	205790
1	โรคติดเชื้อและปรสิต	176598
14	โรคระบบสืบพันธ์ร่วมปัสสาวะ	169937
7	โรคตารวมส่วนประกอบของตา	140556

11. การแสดงผลลัพธ์ SQL รายงาน 2 จำนวนข้อมูล 10,000,000 ระเบียน

ผลลัพธ์ SQL

โฮสต์: localhost
ฐานข้อมูล: helpcare
เวลาในการสร้าง: 17 ก.ค. 2016 17:17น.
สร้างโดย: phpMyAdmin 4.0.10deb1 / MySQL 5.5.49-0ubuntu0.14.04.1
คำสั่ง SQL: SELECT opd_diag.opd_diag3 AS กลุ่มโรค, opd_diag.opd_diag4 AS สาเหตุกลุ่มการป่วย, COUNT(opd_diag.opd_diag3) AS จำนวนผู้ป่วยนอก FROM diagnosis_opd_10m, opd_diag WHERE LEFT(diagnosis_opd_10m.diagcode, 3) = opd_diag.opd_diag1 GROUP BY opd_diag.opd_diag3 ORDER BY จำนวนผู้ป่วยนอก DESC LIMIT 0, 30 ;
แถว: 21

This table does not contain a unique column. Grid edit, checkbox, Edit, Copy and Delete features are not available.

กลุ่มโรค	สาเหตุกลุ่มการป่วย	จำนวนผู้ป่วยนอก
4	โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม	803920
9	โรคระบบไหลเวียนเลือด	756420
10	โรคระบบหายใจ	661610
11	โรคระบบย่อยอาหาร รวมโรคในช่องปาก	492848
13	โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม	451449
18	อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการที่ไม่สามารถจำแนกโรคในกลุ่มอื่	382004
5	ภาวะแปรปรวนทางจิตและพฤติกรรม	205790
1	โรคติดเชื้อและปรสิต	176598
14	โรคระบบสืบพันธ์ร่วมปัสสาวะ	169937
7	โรคตารวมส่วนประกอบของตา	140556
12	โรคผิวหนังและเนื้อเยื่อใต้ผิวหนัง	137255
21	สาเหตภายนอกอื่น ๆ ที่ทำให้ป่วยหรือตาย	87206
6	โรคระบบประสาท	86051
8	โรคหูและปุ่มกกหู	38637
3	โรคเลือดและอวัยวะสร้างเลือด และความผิดปกติเกี่ยวกับภูมิคุ้มกัน	34405
2	เนื้องอก	33841
15	ภาวะแทรกในการตั้งครรภ์ การคลอด และระยะหลังคลอด	22267
20	อุบัติเหตุจากการขนส่ง และผลที่ตามมา	20216
16	ภาวะผิดปกติของทารกแรกเกิดขึ้นในระยะปริกำหนด	7700
17	รูปร่างผิดปกติแต่กำเนิด การพิการจนผิดรูปแต่กำเนิดและโครโมโซมผิดปกติ	5939
19	การเป็นพิษและผลที่ตามมา	918

12. การแสดงผลลัพธ์ MapReduce รายงาน 1 และ รายงาน 2 จำนวนข้อมูล 10,000,000 ระเบียน

ICD-10	GROUP	Total
โรคเกี่ยวกับต่อมไร้ท่อ โภชนาการ และเมตะบอลิสัม	4	803,920
โรคระบบไหลเวียนเลือด	9	756,420
โรคระบบหายใจ	10	661,610
โรคระบบย่อยอาหาร รวมโรคในช่องปาก	11	492,848
โรคระบบกล้ามเนื้อ รวมโครงร่าง และเนื้อยึดเสริม	13	451,449
อาการแสดงและสิ่งผิดปกติที่พบได้จากการตรวจทางคลินิกและทางห้องปฏิบัติการ	18	382,004
ภาวะแปรปรวนทางจิตและพฤติกรรม	5	205,790
โรคติดเชื้อและปรสิต	1	176,598
โรคระบบสืบพันธ์ร่วมปัสสาวะ	14	169,937
โรคตารวมส่วนประกอบของตา	7	140,556
โรคผิวหนังและเนื้อเยื่อใต้ผิวหนัง	12	137,255
สาเหตภายนอกอื่น ๆ ที่ทำให้ป่วยหรือตาย	21	87,206
โรคระบบประสาท	6	86,051
โรคหูและปุ่มกกหู	8	38,637
โรคเลือดและอวัยวะสร้างเลือด และความผิดปกติเกี่ยวกับภูมิคุ้มกัน	3	34,405
เนื้องอก	2	33,841
ภาวะแทรกในการตั้งครรภ์ การคลอด และระยะหลังคลอด	15	22,267
อุบัติเหตุจากการขนส่ง และผลที่ตามมา	20	20,216
ภาวะผิดปกติของทารกแรกเกิดขึ้นในระยะปริกำหนด	16	7,700
รูปร่างผิดปกติแต่กำเนิด การพิการจนผิดรูปแต่กำเนิดและ โครโมโซมผิดปกติ	17	5,939
การเป็นพิษและผลที่ตามมา	19	918

ประวัติผู้เขียน

ชื่อ – นามสกุล	นายรชต ทิมาสรวิชกิจ
ประวัติการศึกษา	พ.ศ. 2543 ปริญญาตรี การจัดการสารสนเทศคอมพิวเตอร์ มหาวิทยาลัยเซนต์จอห์น
ตำแหน่งและสถานที่ทำงานปัจจุบัน	ประกอบธุรกิจส่วนตัว
ประสบการณ์การทำงาน	IT Support Supervisor บริษัท ฮอกไกโด อินเตอร์เนชั่นแนล จำกัด IT Manager บริษัท บูโอโน่ (ประเทศไทย) จำกัด Senior POS Officer บริษัท ไทยแฟรนไชซิ่ง จำกัด System Support บริษัท เซ็นทรัล เรสตอรองส์ กรุ๊ป จำกัด System Develop บริษัท ไฮไฟ โอเรียนท์ (ไทย) จำกัด