DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

กระบวนการประมวลผลข้อมูลแบบ Real-Time จากสื่อสังคมออนไลน์บนคลาวด์

พิมพิกา นางาม

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีการศึกษา 2565

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

CLOUD BASE REAL-TIME DATA PROCESSING FROM SOCIAL MEDIA DATA

PHIMPIKA NANGAM

A Thematic Paper Submitted in Partial Fulfillment of the
Requirements for the Degree of Master of Big Data Engineering,
College of Innovative Technology and Engineering,
Dhurakij Pundit University
Academic Year 2022

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ใบรับรองสารนิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่

หัวข้อสารนิพนธ์ กระบวนการประมวลผลข้อมูลแบบ Real-Time จากสื่อสังคมออนไลน์บนคลาวด์

เสนอโดย พิมพิกา นางาม

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาสารนิพนธ์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบสารนิพนธ์แล้ว

(ผู้ช่วยศาสตราจารย์ ดร.ณัฐพัชร์ อารีรัชกุลกานต์) ประธานกรรมการ

(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา) กรรมการที่ปรึกษาสารนิพนธ์

(ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น) กรรมการ

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ รับรองแล้ว

(ดร.ชัยพร เขมะภาตะพันธ์) คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 31 เดือน กรกฎาคม พ.ศ. 2566

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

หัวข้อสารนิพนธ์	การพัฒนาแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องเพื่อพยากรณ์ความสนใจอาคารชุดของผู้บริโภคออนไลน์ด้วยภาพถ่าย
ชื่อผู้เขียน	พิมพิกา นางาม
อาจารย์ที่ปรึกษา	ดร.เอกสิทธิ์ พัชรวงศ์ศักดา
หลักสูตร	วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา	2565

บทคัดย่อ

ในปัจจุบันที่การเข้าถึงบริการออนไลน์ต่างๆ เกิดขึ้นเป็นจำนวนมาก ได้ก่อให้เกิดข้อมูลจำนวนมาก ทำให้องค์กรต่าง ๆ ตระหนักถึงความสำคัญในการพยายามเข้าถึงและนำไปวิเคราะห์ผลในเชิงลึก ซึ่งหนึ่งในแหล่งข้อมูลที่ถือเป็นปัจจัยในการขับเคลื่อน เพื่อการตัดสินใจ รวมถึงส่งผลในระดับปฏิบัติการ ก็คือข้อมูลที่มาจาก Social Media นั่นเอง

โดยหนึ่งใน Social Media ที่ได้รับความนิยมและสามารถนำมาเป็นตัวชี้วัดได้นั้น ได้แก่ Twitter ซึ่งเรามักพบได้จากอ่านบทความที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล รวมถึงเครื่องมือสำเร็จรูปด้านการวิเคราะห์ข้อมูลที่มักจะมีการดึงข้อมูลจาก Twitter มาเป็นส่วนหนึ่งของเครื่องมือนั้นๆ ด้วย แม้ว่าในปี 2022 ที่ผ่านมาจะเป็นปีแห่งการเปลี่ยนแปลงครั้งใหญ่สำหรับ Twitter ก็ตาม แต่เนื่องจากเจ้าของสินค้าหรือผู้ให้บริการ ยังคงมีการใช้งาน Twitter เพื่อเชื่อมโยง และถ่ายทอดเรื่องราวกับลูกค้าหรือผู้ใช้บริการของตนเองอยู่ จึงยังทำให้ข้อมูลที่เกิดขึ้นบน Twitter เป็นแหล่งข้อมูลสำคัญในการนำมาใช้วิเคราะห์เพื่อหาข้อมูลที่เฉพาะเจาะจงของกลุ่มลูกค้าเป้าหมาย (Customer Insight) ได้

การศึกษาครั้งนี้ ผู้วิจัยจะใช้ข้อมูลบน Twitter มาอธิบายกระบวนการรวบรวมข้อมูลในรูปแบบ Real-time โดยการนำไปประมวลผลบน Google Cloud โดยใช้เครื่องคอมพิวเตอร์เสมือน (VM) ในการเขียนคำสั่งเชื่อมต่อกับ Twitter API และส่งข้อมูลเข้าสู่ Pub/Sub ที่เป็นระบบรับส่งข้อความ (Messaging System) ไปจัดเก็บที่ฐานข้อมูล BigQuery และทำการวัดเวลาประมวลผลในการดึงข้อมูล การวัดผลความครบถ้วนของข้อความที่ใช้รับส่ง ซึ่งผลลัพธ์ที่ได้สามารถใช้เป็นตัวช่วยในการตัดสินใจสำหรับการเลือกใช้เทคโนโลยีที่เหมาะสมต่อการประมวลผลและวิเคราะห์ข้อมูล เพื่อให้มีความยืดหยุ่นตามปริมาณการใช้งาน ทั้งยังช่วยลดต้นทุนด้านทรัพยากรคอมพิวเตอร์และการจ้างบุคคลากรด้วย

คำสำคัญ : การประมวลผลข้อมูล, เรียลไทม์ ดาต้า, การประมวลผลข้อมูลแบบเรียลไทม์, ระบบรับส่งข้อความ

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

Thematic Paper Title	CLOUD BASE REAL-TIME DATA PROCESSING FROM SOCIAL MEDIA DATA
Author	PHIMPIKA NANGAM
Thematic Paper Advisor	Dr. Eakasit Pacharawongsakda
Program	Big Data Engineering
Academic Year	2022

ABSTRACT

These days, various online services are known and produce a large amount of information that makes organizations aware of the importance of trying to access and analyze the results in depth. which is one of the data can be important factor for decision and operational process comes from social media.

Twitter is a widely used social media platform that can be used as a metric, which we can often find from reading articles related to data analysis. Including ready-made data analysis tools that often extract data from Twitter as part of that tool as well, though in 2022 will be a year of big changes for Twitter, however the product's owner or service provider still use Twitter to connect and conveying stories to their customers, thus Twitter's data is an important information can be used for analysis to discover the customer insight.

In this study aims to use the Twitter's data for describe Real-time data processing on Google Cloud using virtual machines (VM) to connect to the Twitter API then send data to Pub/Sub which is a messaging system (Messaging System) to store in the BigQuery database, then calculate the execution time for data retrieval and the metric of deliverable messages. In conclusion, this study can be used as a decision-support tool to select relevant technologies for data processing and analysis to provide scalability computing resources and human recruitment.

Keywords: Data Processing, Real-Time data, Real-time processing, Messaging System

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

กิตติกรรมประกาศ

ในงานวิจัยฉบับนี้ ตั้งแต่เริ่มต้นการศึกษา จนสำเร็จลุล่วงได้อย่างสมบูรณ์ ด้วยความกรุณาอย่างยิ่งจาก ดร.เอกสิทธิ์ พัชรวงศ์ศักดา ที่ได้สละเวลาอันมีค่าแก่ผู้วิจัย เพื่อให้คำปรึกษาและแนะนำตลอดจนตรวจทางแกไขข้อบกพร่องต่างๆ ด้วยความเอาใจใส่เป็นอย่างยิ่ง จนงานวิจัยฉบับนี้สำเร็จสมบูรณ์ลุล่วงได้ด้วยดี ผู้วิจัยขอกรอบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้ จากใจจริง

ขอขอบคุณ เหล่าคณะอาจารย์วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่ ที่ได้กรุณาให้คำแนะนำช่วยเหลือ และปรับปรุงงานวิจัยจนเสร็จสมบูรณ์

สุดท้ายนี้ ขออุทิศความดีที่มีในการศึกษาวิจัยนี้แด่ บิดา มารดา ครอบครัวของผู้วิจัย ซึ่งสนับสนุนในทุกด้าน และกำลังใจจากมิตรแท้ทุกท่าน

พิมพิกา นางาม

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญ

หน้า

บทคัดย่อภาษาไทย..... ง

บทคัดย่อภาษาอังกฤษ..... จ

กิตติกรรมประกาศ..... ฉ

สารบัญ..... ช

สารบัญตาราง..... ฌ

สารบัญภาพ..... ญ

บทที่

1. บทนำ..... 1
   - 1.1 ที่มาและความสำคัญของปัญหา..... 1
   - 1.2 วัตถุประสงค์ของการศึกษาหรือวิจัย..... 1
   - 1.3 ขอบเขตของงานวิจัย..... 1
   - 1.4 ประโยชน์ที่คาดว่าจะได้รับ..... 2
2. แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง..... 3
   - 2.1 Data Processing..... 3
   - 2.2 Real-Time Data..... 5
   - 2.3 Real-Time Processing..... 5
   - 2.4 Cloud Computing..... 7
   - 2.5 Google Pub/sub..... 10
   - 2.6 Google BigQuery..... 11
   - 2.7 Google Compute Engine..... 12
   - 2.8 งานวิจัยที่เกี่ยวข้อง..... 16
3. ระเบียบวิธีวิจัย..... 17
   - 3.1 System Flow..... 17
   - 3.2 สร้าง Compute Engine และติดตั้ง Twitter API..... 18
   - 3.3 เตรียมข้อมูลและส่งเข้าสู่ Pub/Sub..... 21
   - 3.4 ส่งข้อมูลเข้าสู่ BigQuery..... 28
4. ผลการวิจัย..... 35
   - 4.1 ผลการดึงข้อมูลจาก Twitter 36
   - 4.2 ผลการทำงานของ Pub/Sub..... 37

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญ (ต่อ)

บทที่ **หน้า**

5. สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ..... 39
5.1 สรุปผลการศึกษา..... 39
5.2 ข้อเสนอแนะ..... 39
บรรณานุกรม..... 40
ประวัติผู้เขียน..... 43

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญตาราง

ตารางที่ **หน้า**

3.1 แสดงรายชื่อแอตทริบิวต์ของ twitter..... 23
3.2 แสดงรายชื่อ Pub/Sub..... 25
3.3 แสดงแสดงรายชื่อ metadata ของ Subscribe..... 30
4.1 ผลการดึงข้อมูลจาก Twitter..... 36

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญภาพ

ภาพที่ **หน้า**

2.1 Real-Time Data Processing..... 7
2.2 ประเภทการให้บริการ Cloud Computing..... 9
2.3 Google Pub/sub..... 11
2.4 Google BigQuery..... 12
2.5 ประเภทของเครื่อง Compute Engine จำแนกตามวัตถุประสงค์การใช้งาน..... 13
2.6 คุณสมบัติของเครื่องในแต่ละประเภท..... 14
2.7 หลักการทำงานของ Load Balancer..... 15
3.1 กระบวนการดำเนินการ..... 17
3.2 เครื่อง Compute Engine..... 18
3.3 สร้างเครื่อง Compute Engine เครื่องที่ 1..... 18
3.4 สร้างเครื่อง Compute Engine เครื่องที่ 2..... 19
3.5 สร้างเครื่อง Compute Engine เครื่องที่ 3..... 20
3.6 แสดงรายชื่อบัญชีผู้ใช้ที่ต้องการดึงข้อมูล..... 21
3.7 แสดงการกำหนดค่าจำนวนโพสต์สูงสุดที่ต้องการดึงต่อบัญชี..... 22
3.8 แสดงข้อมูลแบบ String ที่ได้จาก Twitter API..... 24
3.9 แสดง Pub/Sub สำหรับเครื่องที่ 1..... 25
3.10 แสดง Pub/Sub สำหรับเครื่องที่ 2..... 26
3.11 แสดง Pub/Sub สำหรับเครื่องที่ 3..... 26
3.12 แสดงตัวอย่าง ผลการส่งข้อมูลเข้าสู่ Pub/Sub และบอกเวลาประมวลผลในแต่ละรอบ.. 27
3.13 แสดงการจัดเก็บผลและวาดกราฟแสดงผลลัพธ์การประมวลผล..... 28
3.14 แสดงการกำหนดค่า Subscriber เพื่อส่งข้อมูลเข้าสู่ BigQuery..... 29
3.15 แสดงการสร้าง Dataset..... 31
3.16 แสดงการสร้างตาราง tweet_post สำหรับจัดเก็บข้อมูลจาก Pub/Sub ชุดที่ 1..... 32
3.17 แสดงการสร้างตาราง tweet_post สำหรับจัดเก็บข้อมูลจาก Pub/Sub ชุดที่ 2..... 33
3.18 แสดงการสร้างตาราง tweet_post สำหรับจัดเก็บข้อมูลจาก Pub/Sub ชุดที่ 3..... 34
3.19 แสดงข้อมูลที่จัดเก็บใน BigQuery..... 34

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

4.1 ผลการประมวลผลข้อมูลด้วยเครื่อง Compute Engine ทั้ง 3 เครื่อง..... 36
4.2 แสดงจำนวนการประมวลผล Pub/Sub ที่ 1..... 37
4.3 แสดงจำนวนการประมวลผล Pub/Sub ที่ 2..... 37
4.4 แสดงจำนวนการประมวลผล Pub/Sub ที่ 3..... 38

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในยุคที่เทคโนโลยีและสื่อสังคมออนไลน์มีบทบาทสำคัญในการติดต่อและแลกเปลี่ยนข้อมูลระหว่างบุคคลและองค์กร การใช้ข้อมูลจากแพลตฟอร์มสื่อสังคมต่าง ๆ เพื่อนำมาใช้ในการวิเคราะห์มีความสำคัญอย่างมากต่อธุรกิจในหลายด้าน เช่น การเข้าใจกลุ่มเป้าหมาย, การติดตามและวิเคราะห์แนวโน้ม เป็นต้น ซึ่งมีประโยชน์ต่อองค์กรให้มีความสามารถในการแข่งขันทางธุรกิจ

จากความสำคัญดังกล่าว จึงเป็นที่มาของการศึกษาในครั้งนี้ในการรวบรวมข้อมูลจาก Twitter ซึ่งเป็นแหล่งข้อมูลสำคัญในการใช้วิเคราะห์ข้อมูล โดยการเลือกใช้เครื่องมือที่เหมาะสม สะดวกต่อการดูแลรักษาระบบ สามารถใช้งานได้ง่ายและไม่มีการเกิด ดาวน์ไทม์ (Downtime) หรือช่วงเวลาที่ระบบหยุดทำงาน เนื่องจากข้อมูลใน Twitter เป็นข้อมูลที่มีความเคลื่อนไหวอยู่ตลอดเวลา การรวบรวมและประมวลผลข้อมูลจึงควรใช้เครื่องมือที่เหมาะสมและป้องกันความสูญหายของข้อมูล (Losing Messages) ที่อาจเกิดขึ้นได้

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อศึกษากระบวนการการรวมรวมข้อมูลจาก Twitter ในรูปแบบ Real-Time
2. เพื่อวัดผลการรวบรวมข้อมูลในด้าน เวลาที่ใช้ประมวลผลข้อมูล และความสูญหายของข้อมูล

1.3 ขอบเขตของงานวิจัย

ขอบเขตของการวิจัย การวิจัยครั้งนี้มีขอบเขตดังนี้

1. ดึงข้อมูลจาก Twitter ในรูปแบบ Real-Time โดยใช้ไลบรารี่ Twitter API
2. เตรียมข้อมูลที่ได้รับมาจาก Twitter API ให้อยู่ในรูปแบบที่พร้อมนำไปใช้งาน เช่น การเลือกแอตทริบิวต์ การแปลงรูปแบบข้อความ เป็นต้น
3. ส่งข้อมูลเข้าสู่ Google Pub/Sub Topic เพื่อใช้เป็นที่พักรักษาข้อมูล และกระจายข้อมูล
4. ส่งข้อมูลจาก Pub/Sub Topic เข้าสู่ BigQuery
5. วัดเวลาประมวลผลการทำงาน และความสูญหายของข้อมูล

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

1.4 ประโยชน์ที่ได้รับ

การใช้เทคโนโลยีเพื่อลดภาระงานหรือช่วยในการลดต้นทุนในเรื่องการจัดการกับข้อมูลที่องค์กรจำเป็นต้องใช้ เนื่องจากจะทำให้ทราบความเหมาะสมของขนาดเครื่องที่ใช้ ระยะเวลา และลดจำนวนพนักงานที่ต้องทำหน้าที่นำเข้าด้วยวิธีการรวบรวมข้อมูลด้วยตัวเองด้วยมือแล้วโหลดเข้าสู่ฐานข้อมูลด้วยตัวเอง

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บทที่ 2
ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

2.1 Data Processing

การประมวลผลข้อมูล (Data Processing) คือกระบวนการในการจัดเก็บ รวบรวมข้อมูลจำนวนมาก ให้มาอยู่ในรูปของข้อมูลที่ต้องการหรือเหมาะสมในการนำไปใช้งาน ประกอบด้วยขั้นตอน 6 ขั้นตอน ดังนี้

1) Data Collection คือการเก็บรวบรวมข้อมูลจากแหล่งต่างๆ เพื่อนำมาใช้ในการประมวลผลหรือวิเคราะห์ เป็นขั้นตอนแรกของการประมวลผลข้อมูล ซึ่งสามารถกระทำได้หลายวิธี นับตั้งแต่ การส่งแบบสอบถาม การสัมภาษณ์ การทดลอง หรือการใช้เครื่องมือที่เหมาะสมช่วยในการ รวบรวมข้อมูลที่ต้องการเพื่อให้ได้ข้อมูลที่ถูกต้องตามสภาพความเป็นจริงมากที่สุด

2) Data Preparation เป็นขั้นตอนการเตรียมข้อมูล หลังจากที่ได้รวบรวมข้อมูลมาแล้ว หรือเรียกกันทั่วไปว่า “การประมวลผลล่วงหน้า” เป็นขั้นตอนที่ข้อมูลดิบได้รับการล้างและจัดระเบียบ สำหรับนำไปสู่ขั้นตอนการนำไปวิเคราะห์ผล โดยวัตถุประสงค์ของขั้นตอนนี้มีเพื่อกำจัดข้อมูลที่ไม่ดี เช่น ข้อมูลที่มีความซ้ำซ้อน ไม่สมบูรณ์ หรือไม่ถูกต้อง โดยมีขั้นตอนย่อย 3 ขั้นตอน ได้แก่

2.1) Data Cleansing เป็นการทำงานเกี่ยวกับการทำความสะอาดข้อมูล ได้แก่ตรวจสอบและแก้ไข ให้ได้ข้อมูลที่อยู่รูปแบบที่สมบูรณ์ ขั้นตอนนี้จะทำให้ทราบว่ามีข้อผิดพลาดอะไรเกิดขึ้นกับข้อมูลบ้าง เช่น การจัดเก็บข้อมูลที่มีความซ้ำซ้อนกัน (Duplicate data) ข้อมูลไม่ถูกต้อง (Incorrectly data) ข้อมูลเก่าล้าสมัย (Expired data) ข้อมูลเกิดการเสียหายบางส่วน (Missing Value) มีค่าข้อมูลที่มีความผิดปกติหรือแตกต่างไปจากข้อมูลในกลุ่ม (Outliers) เป็นต้น จากนั้นจึงทำการปรับปรุงข้อมูล โดยอาจใช้วิธีแทนที่เป็นค่าที่ถูกต้องหรือแทนที่ด้วยค่าใหม่ หรือทำการลบข้อมูลที่ไม่ถูกต้องเหล่านั้นออกไป เป็นต้น

2.2) Data Selection หรือ Feature Selection เป็นขั้นตอนของการเลือกข้อมูล (ตัวแปรอิสระ) ที่มีความสอดคล้องหรือสัมพันธ์กับตัวแปรตาม โดยทั่วไปข้อมูลที่ได้มา อาจจะมีแอตทริบิวต์ (Attribute) หรือ ฟีเจอร์ (Feature) เป็นจำนวนมาก ซึ่งอาจจะเกี่ยวข้องกับตัวแปรตามหรือไม่ก็ได้ ดังนั้นเพื่อให้การวิเคราะห์ตัวแปรตามได้ผลลัพธ์ที่ถูกต้องมากที่สุด จึงควรมีการเลือกเฉพาะแอตทริบิวต์หรือฟีเจอร์ที่มีความสัมพันธ์กับตัวแปรตาม หรือผลลัพธ์ที่เราสนใจให้มากที่สุด ซึ่งสามารถทำได้ด้วยวิธีทดสอบความสัมพันธ์ของข้อมูลด้วยหลักทางสถิติเพื่อให้ได้ตัวแปรอิสระที่มีประสิทธิภาพในการวิเคราะห์หรือพยากรณ์ข้อมูลของแบบจำลองที่สร้างขึ้น

2.3) Data Transformation หรือการแปลงข้อมูล เป็นขั้นตอนสำคัญในกระบวนการเตรียมข้อมูลเพื่อนำไปใช้ในการวิเคราะห์ เนื่องจากข้อมูลที่รวบรวมมานั้นอาจมาจากหลายแหล่งที่มีการจัดเก็บในรูปแบบที่แตกต่างกัน เช่น หน่วยวัดระยะทาง ที่มีทั้ง กิโลเมตร เมตร เซนติเมตร เป็นต้น หรือในบางครั้งเรา

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

จำเป็นต้องแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับเทคนิคที่นำมาใช้ในการวิเคราะห์ เช่น การแปลงค่าแบบต่อเนื่องให้เป็นจำนวนเต็มหน่วย (Discretization) การรวมข้อมูล (Aggregation) หรือการทำข้อมูลให้อยู่ในรูปมาตรฐาน (Normalization) เป็นต้น ซึ่งสามารถอธิบายพอสังเขปได้ดังต่อไปนี้

2.3.1) ดิสครีไทเซชัน (Discretization) เป็นกระบวนการในการแปลงข้อมูลที่เป็นตัวเลข (Continuous data) ให้อยู่ในช่วงข้อมูล (Interval) เนื่องจากอัลกอริทึมในการวิเคราะห์ข้อมูล บางครั้งไม่สามารถรองรับข้อมูลที่เป็นตัวเลข หรือแม้อัลกอริทึมจะรองรับการจัดการข้อมูล การแปลงข้อมูลให้อยู่ในรูปของช่วงของข้อมูลอาจจะทำให้ประสิทธิภาพของโมเดลสูงขึ้นได้ เช่น ช่วงอายุ (age) ของคนไข้ที่เก็บรวบรวม อาจจะถูกแปลงเป็นช่วงอายุ เช่น 1-10 11-30 31-60 หรือเป็น เด็ก ผู้ใหญ่ ผู้สูงอายุ เป็นต้น และค่าข้อมูลใหม่นี้จะถูกใช้แทนข้อมูลเดิม ซึ่งวิธีการนี้จะช่วยให้คอมพิวเตอร์ประมวลผลและวิเคราะห์ข้อมูลได้เร็วขึ้นด้วย

2.3.2) แอกกรีเกชัน (Aggregation) เป็นการแปลงข้อมูลให้อยู่ในรูปแบบข้อมูลแบบสรุป (Summary) เนื่องจากการวิเคราะห์ที่มีประสิทธิภาพนั้น ข้อมูลจะต้องมีคุณภาพและมีจำนวนมากพอที่ใช้เพื่อการค้นหารูปแบบของข้อมูลที่ถูกต้อง ดังนั้นบางครั้งการแปลงข้อมูลโดยการรวมข้อมูลหลาย ๆ ข้อมูลเข้าด้วยกัน และใช้ข้อมูลนั้นสำหรับสร้างโมเดลในการพยากรณ์ ก็อาจจะทำให้ประสิทธิภาพของโมเดลพยากรณ์มีประสิทธิภาพสูงขึ้น ตัวอย่างเช่น ข้อมูล การขาย (Sales) ที่เกิดขึ้นในแต่ละวัน อาจจะถูกนำมารวมกัน และใช้ยอดขายรายสัปดาห์ (Weekly) หรือ ยอดขายรายเดือน (Monthly) ไปใช้แทน ตัวอย่างตัวดำเนินการ Aggregation ได้แก่ SUM MIN MAX และ COUNT เป็นต้น

2.3.3) นอร์มอลไลเซชัน (Normalization) การเก็บรวบรวมข้อมูลที่มีหลายแอตทริบิวต์ บางครั้งมีหน่วยของข้อมูล หรือช่วงของข้อมูลที่แตกต่างกันมากๆ ก็อาจจะมีผลต่อผลการวิเคราะห์ข้อมูลของโมเดล เช่น เปลี่ยนจากหน่วย เซนติเมตร เป็น เมตร หากทดลองเปลี่ยนแล้วทำให้ได้ผลทดสอบดีขึ้น แสดงว่าข้อมูลขึ้นอยู่กับหน่วยของข้อมูลที่นำมาใช้ หน่วยของข้อมูลที่เล็กมากจะทำให้ช่วงของข้อมูลกว้างมาก และส่งผลทำให้แอตทริบิวต์นั้นมีผลต่อต่อการทำงานของโมเดล เช่น ช่วงข้อมูลที่ค่า [0, 10,000] จะส่งผลต่อผลลัพธ์ของการวิเคราะห์ข้อมูล มากกว่าช่วงข้อมูล [0.00, 0.10] เป็นต้น ดังนั้นเพื่อป้องกันการขึ้นอยู่กับหน่วยของข้อมูล จึงควรมีการทำนอร์มอลไลเซชั่น (Normalization) หรือ สแตนดาร์ดไดเซชั่น (Standardization) ก่อนที่จะทำการสร้างโมเดลวิเคราะห์ข้อมูล ซึ่งการทำ นอร์มอลไลเซชั่นจะเป็นการแปลงข้อมูลให้อยู่ในช่วงที่กำหนด เช่น 0-100 หรือ 0-1 วิธีการนี้จะเหมาะกับการแปลงข้อมูลสำหรับวิธีการที่ทำงานบนพื้นฐานที่พิจารณาระยะทางระหว่างข้อมูลแต่ละจุด เช่น SVM (Support Vector Machine) หรือ KNN (K-Nearest Neighbors) เช่น การวิเคราะห์ข้อมูลที่ราคาสินค้าในหน่วย ดอลล่าร์ (Dollar) และ เยน (Yen) ซึ่งมีค่าเงินไม่เท่ากันคือ 1 ดอลล่าร์ เท่ากับ 100 เยน หากไม่ทำการแปลงข้อมูล เมื่อใช้โมเดล SVM หรือ KNN จะพิจารณาว่า 1 ดอลล่าร์ มีค่าเท่ากับ 1 เยน ซึ่งไม่ตรงกับความเป็นจริงนั่นเอง ฉะนั้นการทำนอร์มอลไลเซชั่นจะทำให้ช่วงของข้อมูลเปลี่ยนไป แต่การกระจายของข้อมูล (Data Distribution) ยังเหมือนเดิม

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

3) Data Input เป็นการนำเข้าข้อมูลที่ผ่านกระบวนการทำ Data Preparation เข้าสู่ระบบปลายทาง เช่น ระบบต่างๆ ได้แก่ CRM ERP เป็นต้น ซึ่งรูปแบบของข้อมูลที่ได้จะเป็นข้อมูลดิบที่สามารถนำไปใช้งานต่อได้นั่นเอง

4) Data Processing เป็นรันประมวลผลข้อมูลด้วยอัลกอริทึมจากเครื่องคอมพิวเตอร์ หรือปัญญาประดิษฐ์ เพื่อสร้างผลลัพธ์ที่ตรงกับความต้องการนำไปใช้งาน เช่น ตรวจสอบรูปแบบการโฆษณา ผลวินิจฉัยทางการแพทย์ เป็นต้น

5) Data output หรือการตีความข้อมูลที่ได้ผลลัพธ์มาจากขั้นตอนการทำ Data Processing เช่น การแสดงผลลัพธ์ให้อยู่ในรูปแบบกราฟ เช่น ยอดขายประจำปี ผลลัพธ์แนวโน้มยอดขายที่อยู่ในรูปแบบอนุกรมเวลา (Time Series Data) เป็นต้น

6) Data Storage การจัดเก็บข้อมูล เป็นขั้นตอนสุดท้ายของกระบวนการทำ Data Processing ซึ่งสามารถจัดเก็บในรูปแบบของข้อมูลที่เป็นผลลัพธ์สามารถนำไปใช้ได้ทันที หรือจัดเก็บในรูปแบบที่จะถูกนำไปใช้ต่อในอนาคตก็ได้ โดยขึ้นอยู่กับความเหมาะสมและมีการรักษาความปลอดภัยที่เหมาะสม เพื่อให้ผู้ใช้งานสามารถเข้าถึง และเรียกดึงข้อมูลไปใช้ได้อย่างรวดเร็ว

2.2 Real-Time Data

คือชุดข้อมูลที่มีความสดใหม่และเก็บซ้ำในระดับที่สามารถนำเอาข้อมูลชุดนั้นไปใช้งานได้ก่อนข้อมูลจะมีการเปลี่ยนแปลง หรือถ้าเป็นทางการเก็บข้อมูล ก็คือมีการเก็บข้อมูลทุกครั้งที่มีการเปลี่ยนแปลงเกิดขึ้น ข้อมูลแบบ Real-Time มักเป็นข้อมูลที่เกิดจากการ Search หรือสื่อสังคมออนไลน์ (Social Media) ต่างๆ หรือเป็นข้อมูลจาก Logs การใช้งานต่างๆ เช่น Web Traffic เป็นต้น ข้อมูลเหล่านี้เกิดขึ้นรวดเร็ว และมีความถี่มาก จนกลายเป็นข้อมูลที่มีขนาดใหญ่ และมีค่าเปลี่ยนแปลงอยู่ตลอดเวลา ซึ่งการนำข้อมูล Real-Time Data มาวิเคราะห์จะเป็นการประมวลผลแบบทันที (Real-Time Processing) ซึ่งผลลัพธ์จากการประมวลผลในรูปแบบนี้มักถูกนำไปใช้ในการตัดสินใจแก้ปัญหาให้ทันท่วงที เช่น การตรวจจับป้องกันไฟไหม้ (โดยกำหนดว่า ถ้ามีควันมากและอุณหภูมิสูงผิดปกติถือว่าเกิดไฟไหม้) แล้วระบบคอมพิวเตอร์จะต้องนำข้อมูลจริงในช่วงเวลานั้น ไปประมวลผลอยู่ตลอดเวลา ซึ่งถ้าผลการวิเคราะห์ได้ผลลัพธ์เป็นไฟไหม้ ระบบคอมพิวเตอร์ก็จะสั่งให้น้ำยาดับเพลิงที่ติดตั้งไว้ทำงาน เป็นต้น

2.3 Real-Time Processing

การประมวลผลแบบเรียลไทม์ (Real-Time) คือระบบที่สามารถให้การตอบสนองจากระบบอย่างทันทีทันใดเมื่อได้รับอินพุต (Input) เข้าไป ในทางอุดมคติระบบเรียลไทม์นี้จะไม่เสียเวลาในการประมวลผลเลย หรือใช้เวลาประมวลผลเป็นศูนย์ แต่ในการปฏิบัติทางเครื่องคอมพิวเตอร์นั้นไม่สามารถทำได้จริง ทำได้เพียงลดเวลาการประมวลผลให้น้อยที่สุดจนไม่สามาถเห็นความแตกต่างของช่วงเวลาที่ป้อนอินพุตเข้าไปจนได้รับเอาต์พุต (Output) ออกมา เวลาของความแตกต่างนี้เรียกว่า เวลาตอบสนอง (Response Time) ซึ่งสำหรับ

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ผู้ใช้งานทั่วไปย่อมต้องการเวลาตอบสนองให้น้อยที่สุดเพื่อประสิทธิภาพของระบบ การประมวลผลในรูปแบบเรียลไทม์นิยมนำไปใช้ควบคุมกระบวนการทางอุตสาหกรรม ซึ่งปัจจุบันสามารถควบคุมด้วยระบบคอมพิวเตอร์ที่ทันสมัยและใช้เวลาตอบสนองน้อยมาก

เราอาจแบ่งประเภทของของ Real-Time ได้ต่อไปนี้

1) Hard Real-Time เป็นการประมวลผลระบบที่ถูกรับรองว่าจะได้รับการตอบสนองตรงเวลาและหยุดรอไม่ได้ (ทำงานเสร็จตรงตามเวลา) ระบบนี้ไม่มีฮาร์ดิสก์หรือมีขนาดเล็ก การเก็บข้อมูลจะเก็บใน Short-Term Memory หรือ รอม (Rom) มีข้อเสียคือ มีปัญหากับระบบแบ่งส่วนเวลา และไม่มีการซัพพอร์ต (Support)

2) Soft Real-Time เป็นการประมวลผลระบบแบบ Less Restrictive Type ที่สามารถรอให้งานอื่นเสร็จก่อนได้ (ไม่มี Deadline) มีการจำกัด Utility และเสียงในการใช้อุตสาหกรรมควบคุมหรือหุ่นยนต์ มีประโยชน์ในการประยุกต์ใช้ใน Multimedia, Virtual Reality, การสำรวจในน้ำและดาวเคราะห์ ซึ่งต้องการ Features ของระบบปฏิบัติการขั้นสูง

โดยทั่วไปการประมวลผแบบ Real-Time ประกอบด้วย 3 ขั้นหลัก ได้แก่

1) ส่วน Input หรือการนำเข้าข้อมูล มีขั้นตอนย่อยดังนี้

1.1) การรวบรวมข้อมูล (Collection) เป็นขั้นตอนแรกของการประมวลผลข้อมูล คือการรวบรวมข้อมูลที่จะนำมาประมวลผล เช่น ใบสั่งซื้อสินค้า ประวัติพนักงาน ข้อมูลศึกษา เป็นต้น แล้วจึงนำข้อมูลที่รวบรวมมาได้นั้นทำการเตรียมเพื่อนำไปประมวลผล

1.2) การบันทึกข้อมูล (Recording) เป็นขั้นตอนที่ต่อจากการรวบรวมข้อมูล และบันทึกลงในสื่อ (Media) บันทึกข้อมูล เช่น เทปแม่เหล็ก จานแม่เหล็ก (ฮาร์ดิสก์) ฐานข้อมูล เป็นต้น

1.3) การสอบทานข้อมูล (Verification) เป็นการตรวจสอบความถูกต้องของข้อมูล ก่อนที่จะนำไปประมวลผล

2) การ Processing หรือการดำเนินการ เป็นขั้นตอนนำข้อมูลซึ่งผ่านขั้นตอนแรกมาแล้วและนำไปประมวลผล โดยมีกระบวนการย่อยดังต่อไปนี้

2.1) การ Sorting หรือการเรียงลำดับข้อมูล ซึ่งมี 2 วิธีได้แก่ เรียงลำดับจากมากไปน้อย และเรียงจากน้อยไปมาก ในการเรียงลำต้องกำหนดว่าจะใช้ข้อมูลไหนในการจัดเรียง เช่น เรียงตามวันที่สร้างเอกสารจากน้อยไปมาก เรียงจากชื่อผู้ส่งงานตามลำดับตัวอักษร เป็นต้น โดยเราจะเรียกข้อมูลที่ใช้จัดเรียงนี้ว่า คีย์ฟิลด์ (Key Field)

2.2) การทำ Classification หรือทำการจัดกลุ่มข้อมูลที่เหมือนกันไว้ด้วยกัน เช่น จัดกลุ่มพนักงานตามแผนก จัดกลุ่มหนังสือตามประเภท เป็นต้น

2.3) การทำ Selection หรือทำการการเลือกข้อมูลที่มีคุณสมบัติตรงตามต้องการ

2.4) การทำ Calculation หรือการนำข้อมูลไปคำนวณ เช่น การบวก การลบ การคูณ การหาร หรือการยกกำลัง เป็นต้น

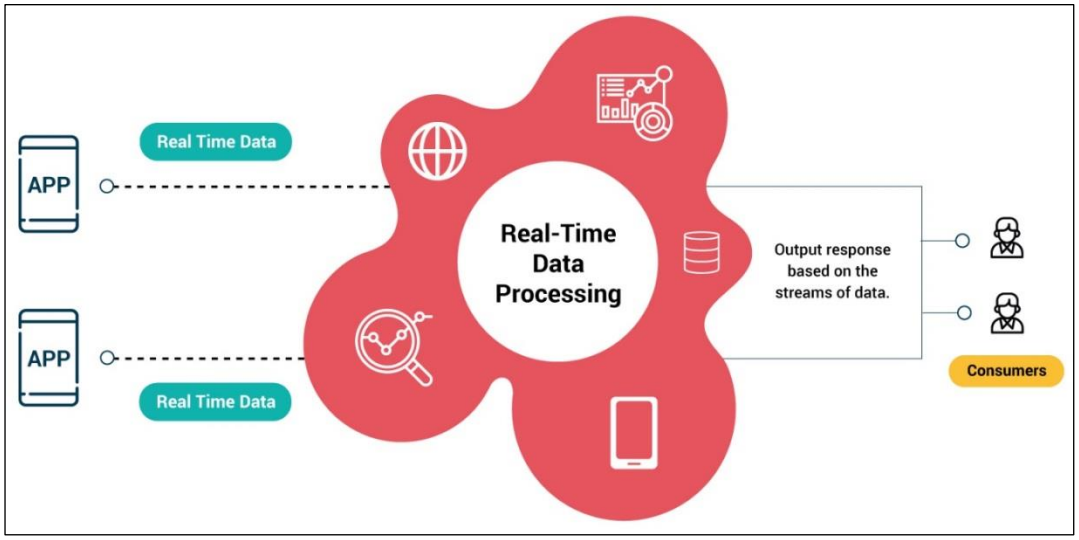
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

2.5) การ Update หรือการปรับปรุงข้อมูลให้ทันสมัยและเป็นปัจจุบันอยู่เสมอ ได้แก่ การเพิ่ม การลบ การแก้ไข โดยต้องคำนึงถึงความถูกต้องของข้อมูลที่เปลี่ยนแปลงนั้นด้วย

3) ส่วน Output หรือผลลัพธ์ของการประมวลผล มีขั้นตอนย่อย ต่อไปนี้

3.1) Report หรือการออกรายงาน เป็นการนำข้อมูลมาแสดงผลตามรูปแบบที่กำหนด โดยอาจทำการออกรายงานทางหน้าจอ Monitor หรือสั่งพิมพ์ทาง Printer ขึ้นอยู่กับวัตถุประสงค์ของการใช้งาน

3.2) การจัดเก็บข้อมูล ควรทำการจัดเก็บผลลัพธ์ที่ได้ในแหล่งที่ค้นหาได้สะดวก ง่ายต่อไปหยิบไปใช้งาน และมีความปลอดภัยด้วย [9]



ภาพที่ 2.1 Real-Time Data Processing

ที่มา: https://axual.com/top-things-to-know-about-real-time-data-processing/

2.4 Cloud Computing

การประมวลผลแบบคลาวด์ (Cloud Computing) คือเทคโนโลยีทางระบบคอมพิวเตอร์ที่ให้บริการแบบเครือข่ายออนไลน์ทุกรูปแบบ ตั้งแต่ การใช้งานซอฟต์แวร์ พื้นที่จัดเก็บข้อมูล หน่วยประมวลผลข้อมูล ระบบเครือข่าย รวมถึงบริการด้านโครงสร้างพื้นฐานทางด้านไอที ที่จำเป็นต่อการใช้งานขององค์กร มีการกำหนดราคาค่าบริการที่ใช้ตามจริง แทนการซื้อ การเป็นเจ้าของ โดยที่ระบบ Cloud มีการทำงานร่วมกันของเครื่องเซิร์ฟเวอร์ (Server) จำนวนมาก มีการแบ่งชั้นการประมวลผลออกจากชั้นการจัดเก็บ เมื่อมีเซิร์ฟเวอร์ใดเสียหาย ก็ไม่ส่งผลกระทบต่อการใช้งานของผู้ใช้บริการ เพราะมีการสลับไปยังเซิร์ฟเวอร์เครื่องอื่นแทนโดยอัตโนมัติทันที รองการขยายหรือหดตัวเมื่อมีการใช้งานเพิ่มขึ้นหรือลดลง ทำให้การทำงานไม่ติดขัดและมีความมั่นใจได้ตลอดเวลาว่าระบบจะไม่หยุดการทำงานลงโดยไม่มีระบบทดแทน [8] เราสามารถอธิบายคุณสมบัติของ Cloud Computing ได้ดังต่อไปนี้

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

1) บริการตนเองตามความต้องการ (On Demand Self Service) ผู้ใช้งานสามารถเข้าถึงระบบที่ให้บริการได้โดยอัตโนมัติ และสามารถปรับเปลี่ยนการใช้งาน เช่น Server Time และ Storage ได้ตามความต้องการในช่วงเวลาใดก็ได้ ผ่านระบบบริหารจัดการบนเว็บไซต์ที่ผู้ให้บริการจัดหาไว้ให้

2) เข้าถึงได้หลายช่องทาง (Broad Network Access) ผู้ใช้งานสามารถเข้าถึงระบบของของผู้ให้บริการจากอุปกรณ์ประเภทใดก็ได้ เช่น สมาร์ทโฟน (Smart Phone), แท็บเล็ต (Tablet), เครื่องคอมพิวเตอร์ (Computer) หรือโน้ตบุ๊ค (Notebook) เป็นต้น

3) การใช้ทรัพยากรร่วมกัน (Resource Pooling) ความสามารถในการบริหารจัดการระบบเพื่อให้บริการแก่ผู้ใช้บริการจำนวนมากในเวลาเดียวกัน (Multi-tenants) โดยผู้ใช้ไม่จำเป็นต้องรู้ว่าข้อมูลต่างๆ มีการจัดเก็บอยู่ที่ใด

4) ความยืดหยุ่นในการให้บริการสูง (Rapid Elasticity) ระบบที่มีความยืดหยุ่นสูงและหลากหลายความต้องการของผู้ใช้งาน ทำให้มีความสามารถในการเพิ่มหรือลดทรัพยากรได้อย่างรวดเร็ว และไม่มีข้อจำกัดในเรื่องจำนวน ปริมาณ และระยะเวลาในการใช้งาน

5) ระบบการวัดบริการ (Measured Service) ความสามารถในการบริหารจัดการหรือควบคุมการใช้ทรัพยากรที่เกี่ยวข้อง โดยการวัดปริมาณและคิดค่าบริการตามการใช้งานที่เกิดขึ้นจริง หรือ Pay-per-use

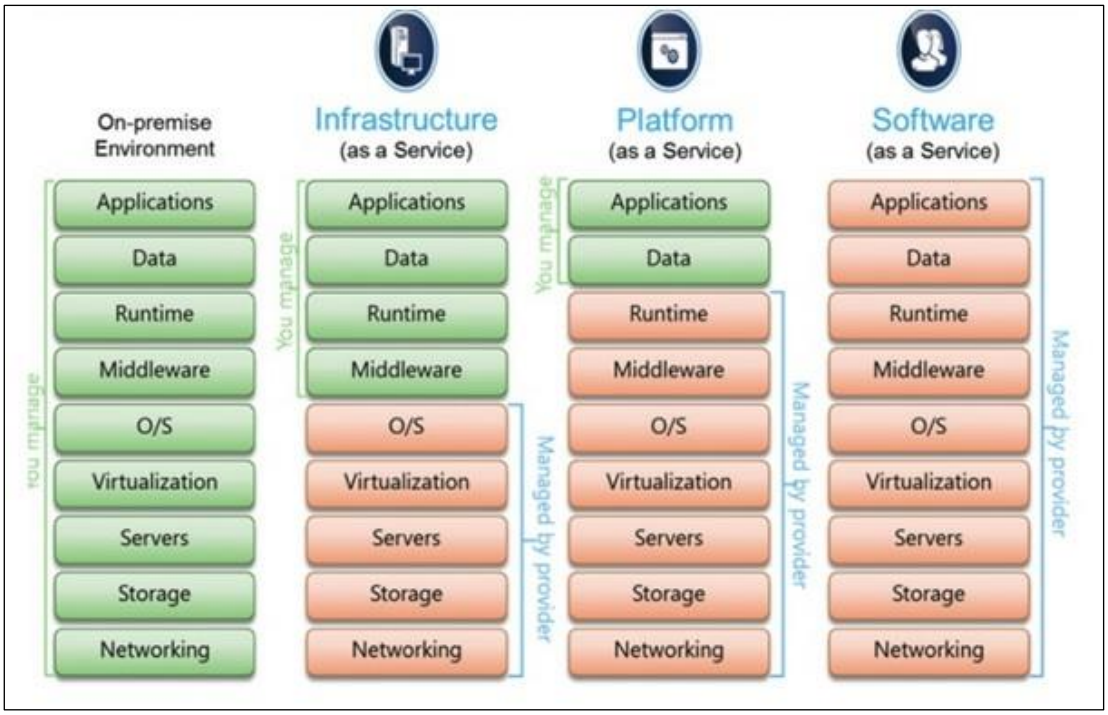
นอกจากคุณสมบัติข้างต้นแล้ว เรายังสามารถแบ่งประเภทการให้บริการ Cloud Computing ได้ดังต่อไปนี้

1) Infrastructure as a Service (IaaS) หมายถึงการให้บริการโครงสร้างพื้นฐานหลักของ Cloud Computing เช่น ระบบประมวลผล ระบบจัดเก็บข้อมูล ระบบเครือข่าย ตลอดจนอุปกรณ์พื้นฐานที่เกี่ยวข้อง เช่น หน่วยความจำ (Storage) เครื่องบริการ (Server) หรือ ระบบปฏิบัติการ (Operation System) เป็นต้น ทำให้ผู้ใช้บริการสามารถใช้ซอฟต์แวร์แอพพลิเคชั่นได้อย่างมีประสิทธิภาพ โดยไม่ต้องบริหารจัดการโครงสร้างพื้นฐานด้วยตนเอง

2) Platform as a Service (PaaS) หมายถึงการให้บริการแพลตฟอร์ม และเครื่องมือเพื่อใช้ในการพัฒนาซอฟต์แวร์แอพพลิเคชั่น เช่น โปรแกรมเบื้องต้น ฐานข้อมูล และระบบที่เอื้อต่อการสร้างแอพพลิเคชั่น โดยผู้ใช้ไม่จำเป็นต้องบริหารจัดการระบบหรือเครื่องมือด้วยตัวเอง แต่ต้องติดตั้ง แก้ไข หรือปรับแต่งแอพพลิเคชั่นที่สร้างขึ้นมาด้วยตนเอง

3) Software as a Service (Saas) หมายถึงการให้บริการซอฟต์แวร์ที่มีความยืดหยุ่นต่อการเข้าถึงการใช้งานได้หลากหลาย โดยผู้ใช้ไม่จำเป็นต้องบริหารจัดการโครงสร้างพื้นฐานเอง ทั้งในส่วนของเครือข่าย เซิร์ฟเวอร์ ระบบปฏิบัติการ การจัดเก็บข้อมูล รวมถึงความสามารถของแอพพลิเคชั่นที่ใช้งาน [10]

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 2.2 ประเภทการให้บริการ Cloud Computing

ที่มา : http://ictstou.blogspot.com/2017/11/cloud-computing-windows-azure-amazon-ec2.html

นอกจากนี้แล้วรูปแบบการให้บริการ Cloud Computing ยังสามารถจำแนกได้ดังนี้

1) Private Cloud หมายถึงบริการ Cloud Computing สำหรับหน่วยงาน หรือองค์กรหนึ่งเพียงองค์กรเดียว (แต่อาจมีผู้ใช้ภายในองค์กรหลายคน) ซึ่งการบริหารจัดการระบบโดยส่วนมากจะกระทำโดยบุคคลากรภายในองค์กร

2) Community Cloud เป็นการรวมกลุ่มของผู้ใช้ที่มีวัตถุประสงค์ ความต้องการ และจุดมุ่งหมายแบบเดียวกัน เช่น การใช้งาน Cloud Computing ของสถาบันการศึกษา เพื่อประโยชน์ในการเรียนการสอน หรือการใช้งานในกลุ่มธุรกิจเฉพาะด้านที่ต้องการโครงสร้างพื้นฐานของระบบ Cloud Computing แตกต่างจากธุรกิจอื่นๆ เช่น โรงพยาบาล สาธารณสุข เป็นต้น โดยทั่วไปแล้วการใช้ Community Cloud ส่วนใหญ่ไม่ได้เป็นการใช้เพื่อหารายได้ แต่มีจุดมุ่งหมายเพื่อให้มีการแบ่งปัน ข้อมูลและประสบการณ์ระหว่างองค์กรที่มีความสนใจในเรื่องเดียวกัน

3) Public Cloud เหมาะสมสำหรับงานบริการแก่บุคคล หรือลูกค้าภายนอกองค์กร รวมถึงระบบงานที่เกี่ยวกับบริการให้ข้อมูลข่าวสาร ที่จำเป็นต้องติดต่อสื่อสารกับบุคคลภายนอกหรือลูกค้าภายนอกองค์กร เช่น ระบบเว็บไซต์ขององค์กร ระบบอีเมล ระบบบริการพนักงานที่ต้องพบปะลูกค้าจำนวนมาก ต้อง

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

อาศัยข้อมูลจากระบบลูกค้าสัมพันธ์ด้วยอุปกรณ์พกพาหลากหลายรูปแบบ เป็นต้น งานประเภทนี้เหมาะสำหรับการใช้ Public Cloud ที่หลายฝ่ายสามารถใช้งานร่วมกันได้

4) Hybrid Cloud องค์กรส่วนใหญ่มีแนวโน้มที่จะมุ่งไปสู่การใช้บริการ Cloud Computing ในลักษณะของการผสมผสาน การใช้บริการหลายๆ รูปแบบเข้าด้วยกัน หรือที่รียกว่า Hybrid Cloud บริการที่จะเป็น Hybrid Cloud ได้นั้น มีเงื่อนไขว่าว่าต้องให้บริการกับงานที่เชื่อมโยงกันหลายฝ่าย มีความจำเป็นในการเข้าถึงข้อมูลจากภายนอกองค์กร แต่จำเป็นต้องมีระบบการรักษาความมั่นคงปลอดภัยที่เข้มงวดกว่าการใช้บริการแบบ Public Cloud เช่น ข้อมูลที่ให้เฉพาะบุคคลกรของบริษัทเข้าถึงได้ มีการกำหนดชั้นความลับของข้อมูลและสิทธิในการเข้าถึงอย่างเป็นระบบ [2]

2.5 Google Pub/sub

เป็นการให้บริการประมวลผลการรับส่งข้อความแบบ Publish-Subscribe หรือการรับส่งข้อความแบบ Pub/Sub บนคลาวด์ของ Google โดยมีการทำงานในรูปแบบการส่งข้อความ (Message) จากผู้ส่ง (Publisher) ไปยังผู้รับ (Subscriber) โดยที่ผู้ส่งและผู้รับ จะตกลงสื่อสารกันด้วยหัวข้อ (Topic) ผ่านตัวกลาง/ช่องทางที่กำหนดไว้แล้ว โดยมีส่วนประกอบดังต่อไปนี้

1) ข้อความ (Message) คือข้อมูลที่ที่ต้องการสื่อสาร จะถูกส่งจากผู้ส่งไปยังผู้รับ โดยอยู่ในรูปแบบใดก็ได้ ตั้งแต่ สตริง (String) วิดิโอ เสียง หรือเนื้อหาดิจิตอลอื่นๆ

2) หัวข้อ (Topic) ทุกข้อความจะมีหัวข้อที่เกี่ยวข้องกับเนื้อหาในข้อความ โดยหัวข้อจะทำหน้าที่เป็นเหมือนช่องทางหรือตัวกลาง ระหว่างผู้ส่งกับผู้รับ ซึ่งเป็นชื่อของแหล่งข้อมูลที่ฝั่งผู้รับจะสามารถเข้าถึงได้

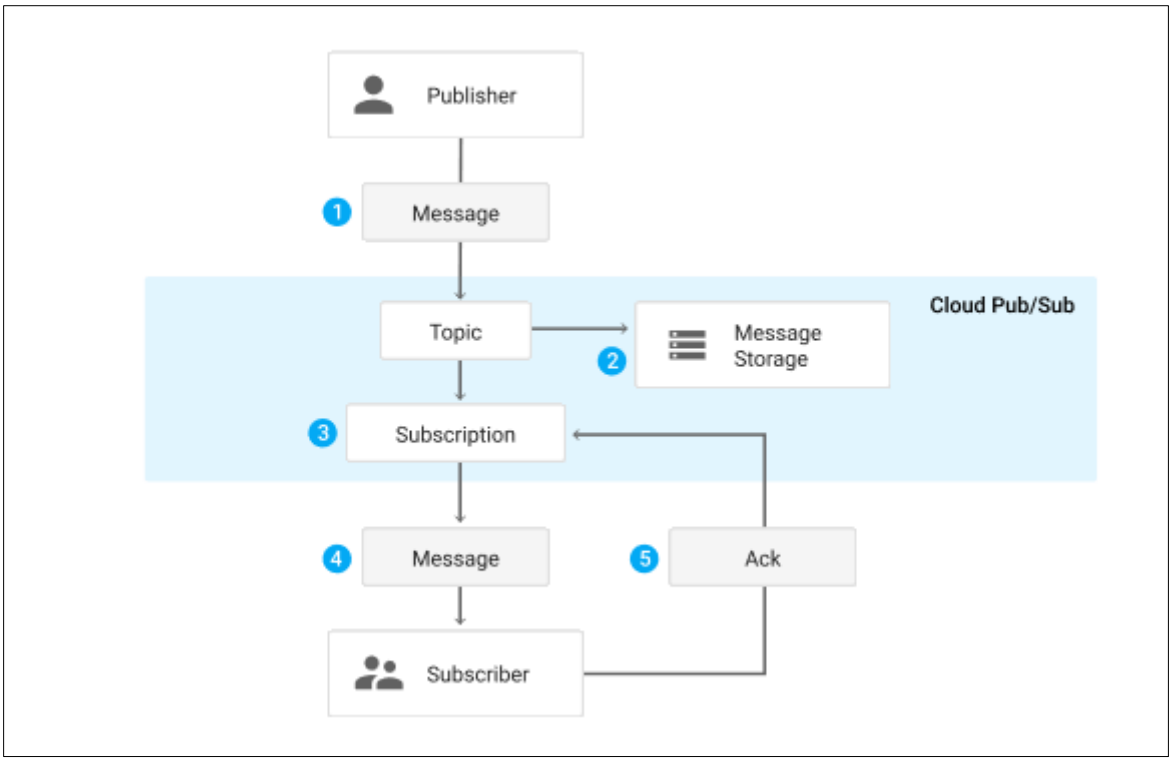
3) ผู้รับข้อความ (Subscriber) คือบุคคลที่รับข้อความ ในหัวข้อที่ตนสนใจ หรือการระบุหัวข้อ (Topic) นั่นเอง

4) ผู้ส่งข้อความ (Publisher) คือผู้ที่หน้าที่ส่งข้อความ โดยมีการตั้งชื่อหัวข้อ (Topic) จากนั้นจะส่งข้อความเหล่านั้นไปให้ผู้รับข้อความของหัวข้อนั้นๆ ทีเดียวทั้งหมด การโต้ตอบระหว่าง Publisher และ Subscriber เป็นความเชื่อมโยงแบบหนึ่งต่อกลุ่ม Publisher ไม่จำเป็นต้องรู้ว่าใครกำลังใช้ข้อมูลที่เผยแพร่ออกไป และ Subscriber ก็ไม่จำเป็นต้องรู้ว่าข้อความนั้นมาจากไหน

5) Acknowledgment (Ack) เป็นสัญญาณที่สร้างจาก Subscriber หลังจากได้รับ Message สำเร็จ และ Acknowledgment จะถูกลบออกจากระบบ Message Queue

6) Push and Pull คือกระบวนการในการส่ง Message ข้อความจาก Publisher ทำการ Push ข้อความหรือ Message แล้ว Subscriber จะได้รับข้อความเหล่านั้นได้ด้วยการทำ Pull ข้อความด้วยเช่นกัน

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



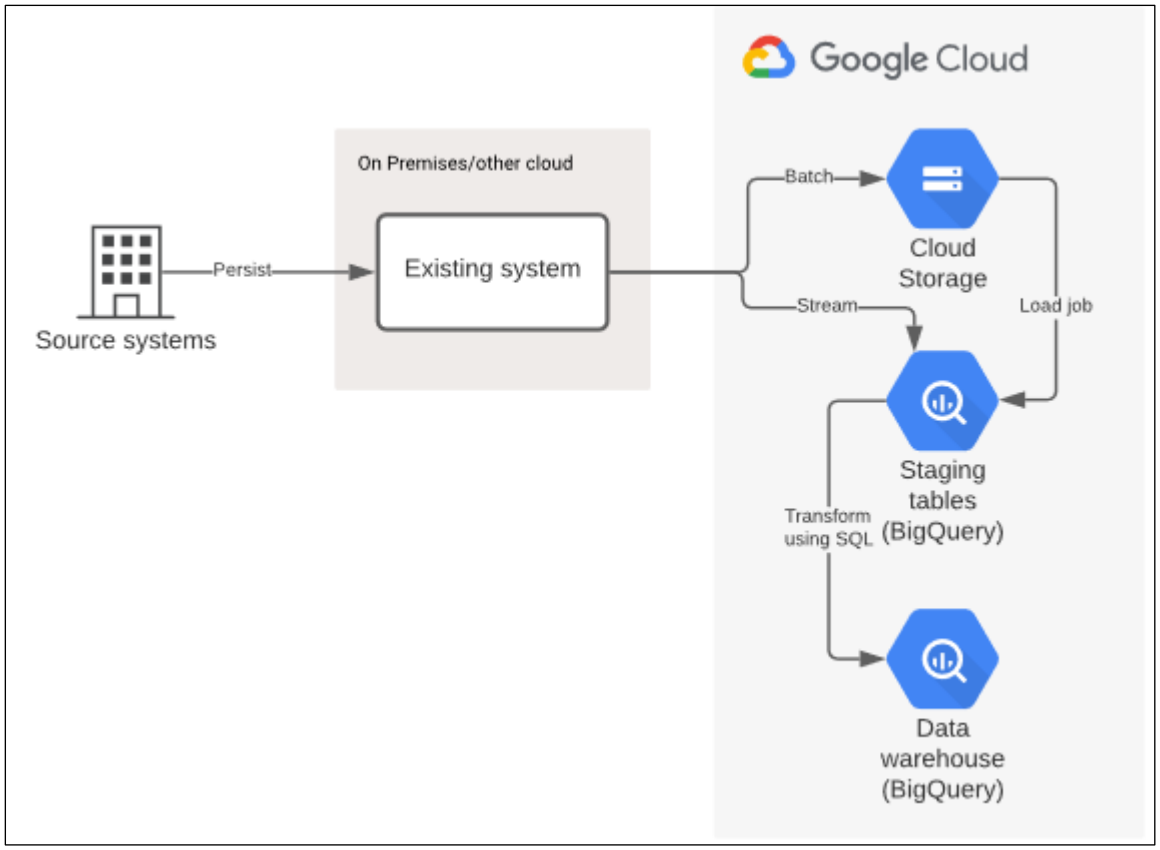
ภาพที่ 2.3 Google Pub/sub

ที่มา: https://cloud.google.com/pubsub/docs/overview

2.6 Google BigQuery

BigQuery เป็นบริการด้านฐานข้อมูลขนาดใหญ่ของ Google ที่เป็นส่วนหนึ่งของการประมวลผลแบบคลาวด์ ที่ทำหน้าที่วิเคราะห์ข้อมูลขนาดใหญ่มากๆ เพื่อหาผลลัพธ์ได้รวดเร็วด้วยการใช้คำสั่งคิวรี่ (Query) ด้วยภาษา Structured Query Language (SQL) และรองรับการเขียนข้อมูลหรือนำเข้าข้อมูลในปริมาณมากได้อย่างมีประสิทธิภาพ โดยรองรับแหล่งข้อมูลในรูปแบบต่างๆ ได้แก่ CSV JSON การนำเข้าแบบสตรีม (Streaming Insert) รวมถึงนำเข้าจากระบบอื่นๆ เช่น Google DoubleClick, AdWords, YouTube เป็นต้น

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 2.4 Google BigQuery

ที่มา: https://www.analyticsvidhya.com/blog/2022/06/most-frequently-asked-google-big-query-interview-questions/

2.7 Google Compute Engine

Compute Engine คือ Virtual Machine หรือเครื่องคอมพิวเตอร์แบบเสมือน โดยสามารถปรับเพิ่ม หรือลดขนาด และประสิทธิภาพของตัว Hardware ต่าง ๆ ได้อย่างรวดเร็ว เช่น ขนาดหน่วยประมวลผล (CPU) ขนาดหน่วยความจำ (RAM) ขนาดของฮาร์ดิสก์ เป็นต้น โดยคุณสมบัติที่สำคัญของเครื่อง Compute Engine มีดังต่อไปนี้

1) ประเภทของเครื่องที่กำหนดไว้ล่วงหน้า โดย Google มีประเภทของเครื่องรวมถึงขนาดคุณสมบัติของเครื่องในแต่แต่ละประเภทที่ได้กำหนดเอาไว้ตามตามรูปภาพที่ 5

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

Workload type						
General-purpose workloads				Compute-optimized	Memory-optimized	Accelerator-optimized
E2	N2, N2D, N1	C3	Tau T2D, Tau T2A	H3, C2, C2D	M3, M2, M1	A2, G2
Day-to-day computing at a lower cost	Balanced price/performance across a wide range of machine types	Consistently high performance for a variety of workloads	Best per-core performance/cost for scale-out workloads	Ultra high performance for compute-intensive workloads	Highest memory to compute ratios for memory-intensive workloads	Optimized for accelerated high performance computing workloads
• Low-traffic web servers • Back office apps • Containerized microservices • Microservices • Virtual desktops • Development and test environments	• Low to medium traffic web and app servers • Containerized microservices • Business intelligence apps • Virtual desktops • CRM applications • Data Pipelines	• High traffic web and app servers • Databases • In-memory caches • Ad servers • Game servers • Data analytics • Media streaming and transcoding • CPU-based ML training and inference	• Scale-out workloads • Web serving • Containerized microservices • Media transcoding • Large-scale Java applications	• Compute-bound workloads • High-performance web servers • Game servers • High performance computing (HPC) • Media transcoding • Modeling and simulation workloads • AI/ML	• Medium to extra-large SAP HANA in-memory databases • In-memory data stores, such as Redis • Simulation • High Performance databases such as Microsoft SQL Server, MySQL • Electronic design automation	• CUDA-enabled ML training and inference • High-performance computing (HPC) • Massively parallelized computation • BERT natural language processing • Deep learning recommendation model (DLRM) • Video transcoding • Remote visualization workstation

ภาพที่ 2.5 ประเภทของเครื่อง Compute Engine จำแนกตามวัตถุประสงค์การใช้งาน

ที่มา: https://cloud.google.com/compute/docs/machine-resource

และในแต่ละประเภทก็มีการกำหนดคุณสมบัติของเครื่องเอาไว้ให้สำเร็จรูปแบบพร้อมใช้โดยกำหนดคุณสมบัติด้าน CPU RAM รวมถึงพื้นที่ Hard Disk จัดเตรียมให้ใช้เมื่อสร้างเครื่องด้วยเช่นกัน ดังแสดงตามรูปภาพที่ 6

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

	N1	N2	N2D	T2D	T2A	E2	C2	C2D	C3	H3	M
Workload types	General purpose	General purpose	General purpose	General purpose	General purpose	Cost optimized	Compute optimized	Compute optimized	General purpose	Compute optimized	M o
CPU types	Intel Skylake, Broadwell, Haswell, Sandy Bridge, and Ivy Bridge	Intel Cascade Lake and Ice Lake	AMD EPYC Rome and EPYC Milan	AMD EPYC Milan	Ampere Altra	Intel Skylake, Broadwell, and Haswell, AMD EPYC Rome and EPYC Milan	Intel Cascade Lake	AMD EPYC Milan	Intel Sapphire Rapids	Intel Sapphire Rapids	Ir S a B
Architecture	x86	x86	x86	x86	Arm	x86	x86	x86	x86	x86	x
vCPUs	1 to 96	2 to 128	2 to 224	1 to 60	1 to 48	0.25 to 32	4 to 60	2 to 112	4 to 176	88	4
Memory	1.8 to 624 GB	2 to 864 GB	2 to 896 GB	4 to 240 GB	4 to 192 GB	1 to 128 GB	16 to 240 GB	4 to 896 GB	8 up to 1,408 GB	352 GB	9 3
Extended memory	✓	✓	✓	–	–	–	–	–	–	–	–
Local SSD	✓	✓	✓	–	–	–	✓	✓	✓	–	
Max local SSD	9 TB	9 TB	9 TB	0	0	0	3 TB	3 TB	12 TB	0	3

ภาพที่ 2.6 คุณสมบัติของเครื่องในแต่ละประเภท

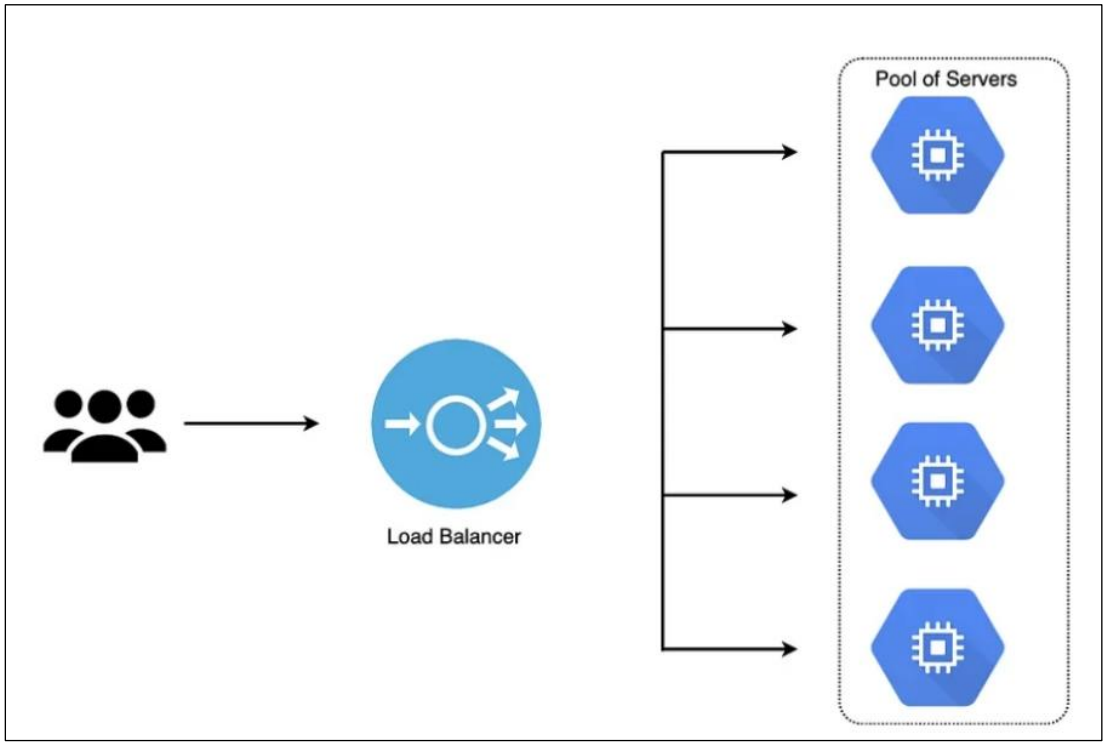
ที่มา: https://cloud.google.com/compute/docs/machine-resource

2) ประเภทของเครื่องที่กำหนดเอง เมื่อสร้างเครื่อง Compute Engine จากคุณสมบัติแบบพร้อมใช้ตามข้อ 1) แล้ว ผู้ใช้บริการยังสามารถกำหนดคุณสมบัติต่างๆ เพิ่มเติมได้ด้วยตนเองอีกด้วย เช่น การเปลี่ยนขนาดหน่วยความ RAM เมื่อตอนสร้างเครื่องขึ้นมาใช้งาน เป็นต้น

3) Load Balancing หรือโหลดบาลานซ์ คือการที่การจัดการระบบของผู้ให้บริการที่จัดสรรทรัพยากรให้เหมาะสมกับกับปริมาณการใช้งานที่เปลี่ยนไปในแต่ละช่วงเวลา เช่น เมื่อมีการเข้าใช้งานพร้อมกันหลายๆ คนจากหลายภูมิภาค เป็นต้น โดยใช้ Load Balancer หรือระบบเซิร์ฟเวอร์ที่มีประสิทธิภาพสูงรองรับการทำงานที่มีปริมาณงาน (Workload) เป็นจำนวนมากได้เป็นอย่างดี เกิดจากการนำเซิร์ฟเวอร์หลายๆ เครื่องที่มีหน้าที่การทำงานเดียวกันมาทำงานร่วมกันเพื่อกระจายปริมาณงานไปยังแต่ละเครื่อง ทำให้ระบบสามารถรองรับการทำงานหนักๆ ได้เป็นอย่างดี มีความยืดหยุ่นสูง เนื่องจากเมื่อความต้องการใช้งานที่เพิ่มขึ้น

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

เพื่อไม่ให้เกิดปัญหา Over Load จนเครื่อง Web Server, Application Server หรือ Database Server ไม่สามารถให้บริการได้นั่นเอง



ภาพที่ 2.7 หลักการทำงานของ Load Balancer

ที่มา: https://levelup.gitconnected.com/load-balancing-on-google-cloud-platform-gcp-why-and-how-a8841d9b70c

4) มีการสนับสนุนระบบปฏิบัติการ Windows และ Linux ซึ่งรองรับผู้ใช้บริการในการสร้างเครื่อง Compute Engine ขึ้นมาใช้งานพร้อมกับให้สามารถกำหนด หรือเลือกระบบปฏิบัติการที่ต้องการใช้ได้ด้วยตนเอง เช่น CentOS, Debian, Ubuntu, CoreOS, SUSE, FreeBSD, Red Hat Enterprise Linux หรือ Windows Server 2012 R2, 2008 R2 เป็นต้น

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

2.8 งานวิจัยที่เกี่ยวข้อง

1) Real-time big data processing for instantaneous marketing decisions: A problematization approach ผู้วิจัยกล่าวถึงการศึกษาเพื่อหาความเป็นไปได้ในการพัฒนาระบบสนับสนุนการตัดสินใจแบบ Real-Time สำหรับการสร้างเคมเปญทางการตลาดเพื่อตอบสนองและค้นหากลุ่มลูกค้าให้ได้ตรงตามเป้าหมาย โดยการใช้เครื่องมือในรูปแบบคลาวด์ ได้แก่ Map Reduce, NOSQL และ HDFS ในการประมวลผลข้อมูลทั้งในรูปแบบ Structure Data และ Unstructured Data เพื่อให้เกิดความได้เปรียบในการแข่งขันในระดับ B2B

2) Data Processing in Cloud Computing Model on the Example of Salesforce Cloud ผู้วิจัยมุ่งไปที่ผลการเปรียบเทียบวิธีการประมวลผลข้อมูลและจัดเก็บข้อมูล ที่รวบรวมจากแหล่งต่างๆ ด้วยการใช้ระบบคลาวด์ในรูปแบบ SaaS ด้วยผู้ให้บริการ Salesforce cloud เพื่อหารูปแบบที่เหมาะสมในการกระบวนการรวบรวมข้อมูล ประมวลผลข้อมูลทั้งในรูปแบบ Batch Processing และ Real-Time Processing

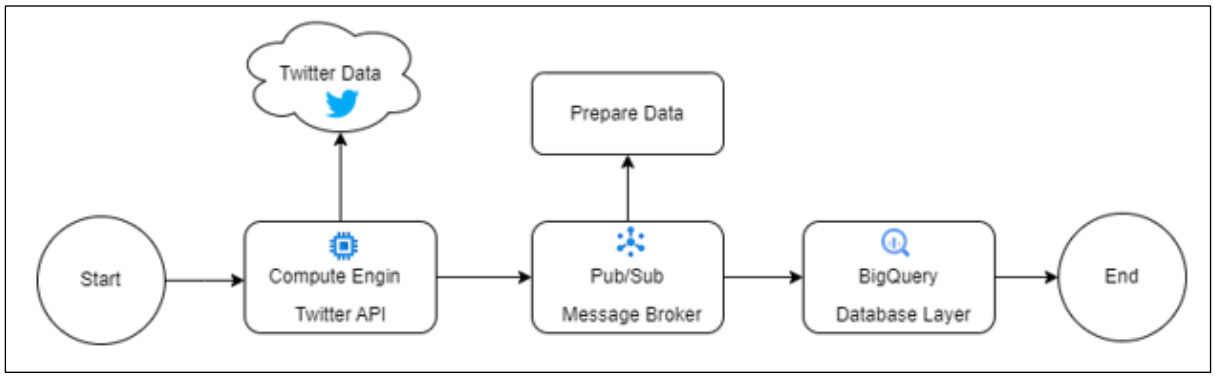
3) Lotus: Serverless In-Transit Data Processing for Edge-based Pub/Sub ผู้วิจัยใช้นำเสนอระบบส่งและแปลงข้อมูลที่ชื่อ Lotus ซึ่งทำงานด้วยรูปแบบ Pub/Sub ในการส่งข้อมูลจากอุปกรณ์เซ็นเซอร์ เพื่อให้ทำหน้าที่เป็นตัวกลางในการรวบรวมข้อมูลจากอุปกรณ์ IoT ต่างๆ และ Publish หรือส่งออกไปในรูปแบบ Topic โดยเป็นการทำงานอยู่บนระบบคลาวด์ แบบ Function-as-a-Service โดยไม่ต้องมีการติดตั้งเซิร์ฟเวอร์ หรือเครื่องมือในการพัฒนาระบบขึ้นมาใช้งานเลย

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บทที่ 3
ระเบียบวิธีวิจัย

3.1 System Flow

ขั้นตอนการประมวลผลข้อความและดำเนินการสร้างระบบสำหรับตรวจสอบข้อความบนเอกสาร ได้ตามวัตถุประสงค์ของงานจึงได้กำหนดกระบวนการการดำเนินการเพื่อศึกษาข้อมูลมีขั้นตอนดังนี้



ภาพที่ 3.1 กระบวนการดำเนินการ

3.2 สร้าง Compute Engine และติดตั้ง Twitter API

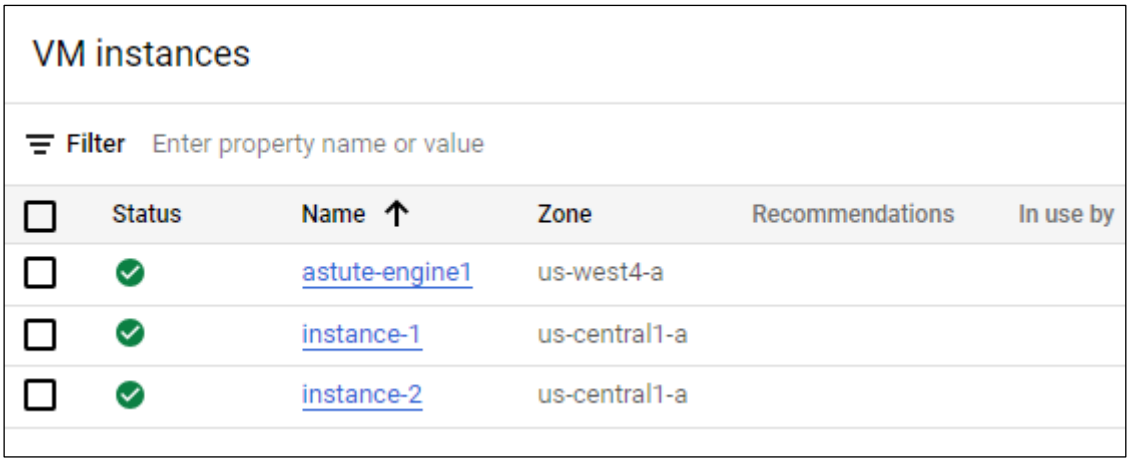
สร้าง Compute Engine หรือเครื่องคอมพิวเตอร์บนคลาวด์ ทั้งหมด 3 เครื่อง เพื่อใช้รันคำสั่งในการอ่านข้อมูลจาก Twitter ด้วย Twitter API ดังต่อไปนี้

1) เครื่องที่ 1 Ubuntu เวอร์ชั่น 20.04 ขนาด ประมวลผล CPU 1 หน่วย RAM 3.75 GB HDD 10 GB โดยใช้ Machine Type เป็น n1-standdard-1

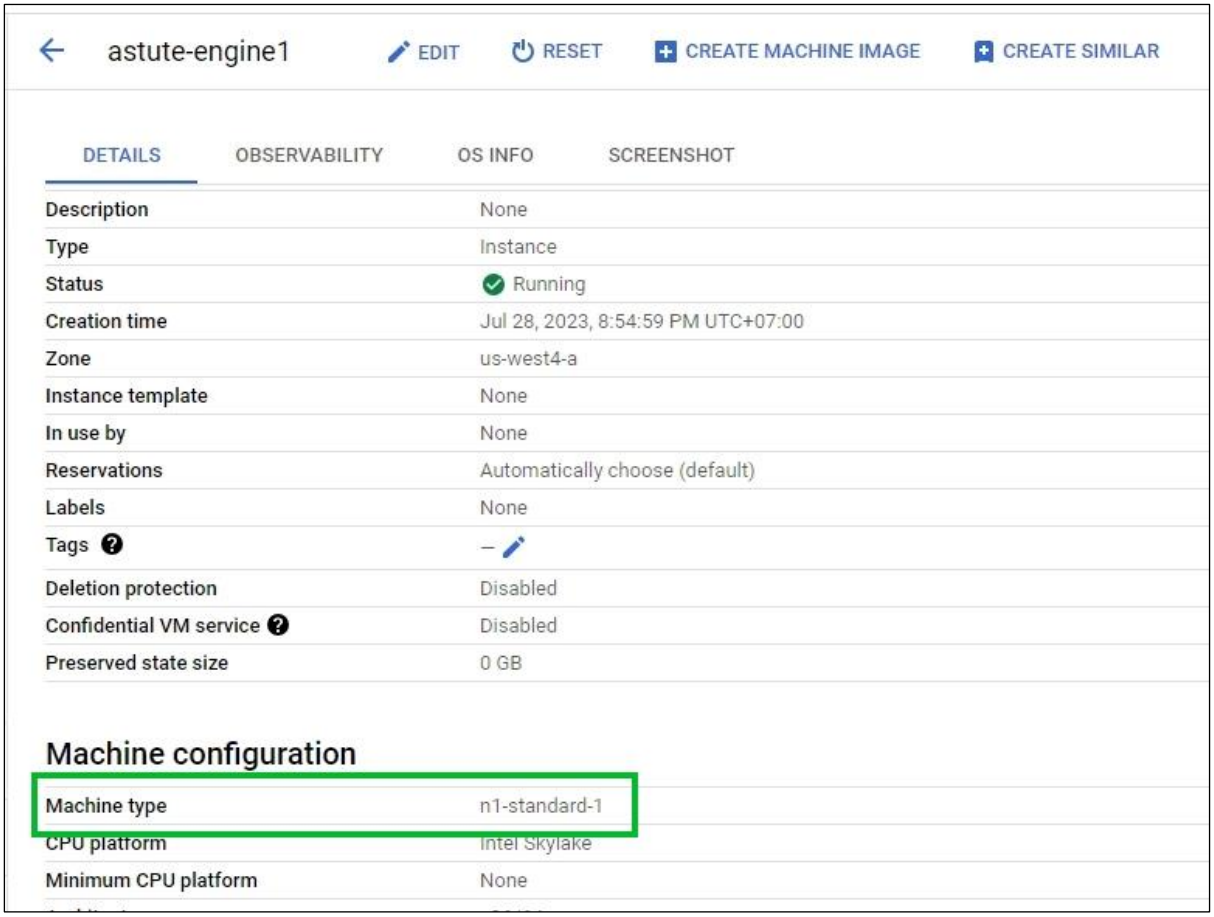
2) เครื่องที่ 2 Ubuntu เวอร์ชั่น 20.04 ขนาด ประมวลผล CPU 2 หน่วย RAM 7.50 GB HDD 10 GB โดยใช้ Machine Type เป็น n1-standdard-2

3) เครื่องที่ 3 Ubuntu เวอร์ชั่น 20.04 ขนาด ประมวลผล CPU 4 หน่วย RAM 15 GB HDD 10 GB โดยใช้ Machine Type เป็น n1-standdard-4

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

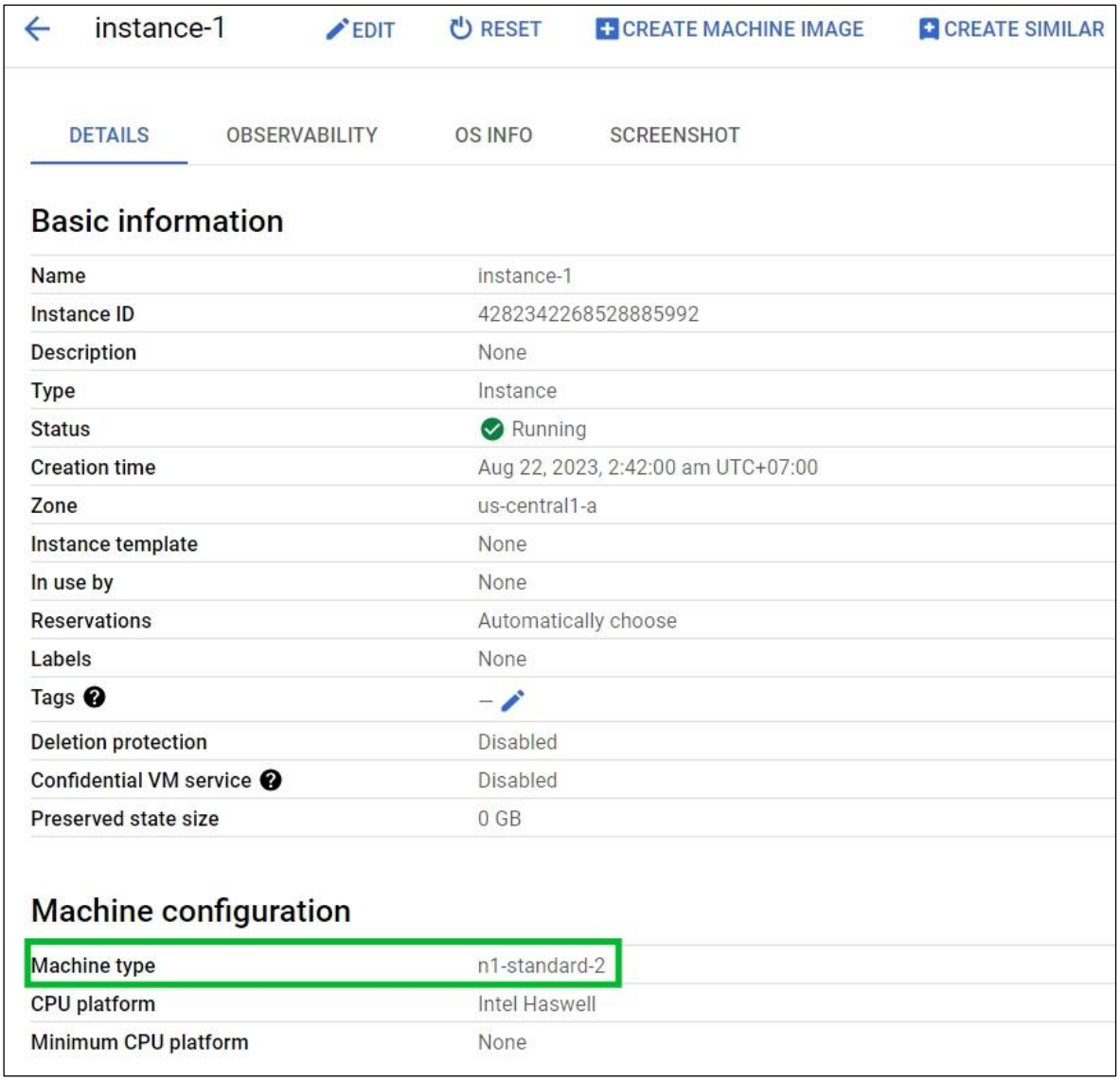


ภาพที่ 3.2 เครื่อง Compute Engine



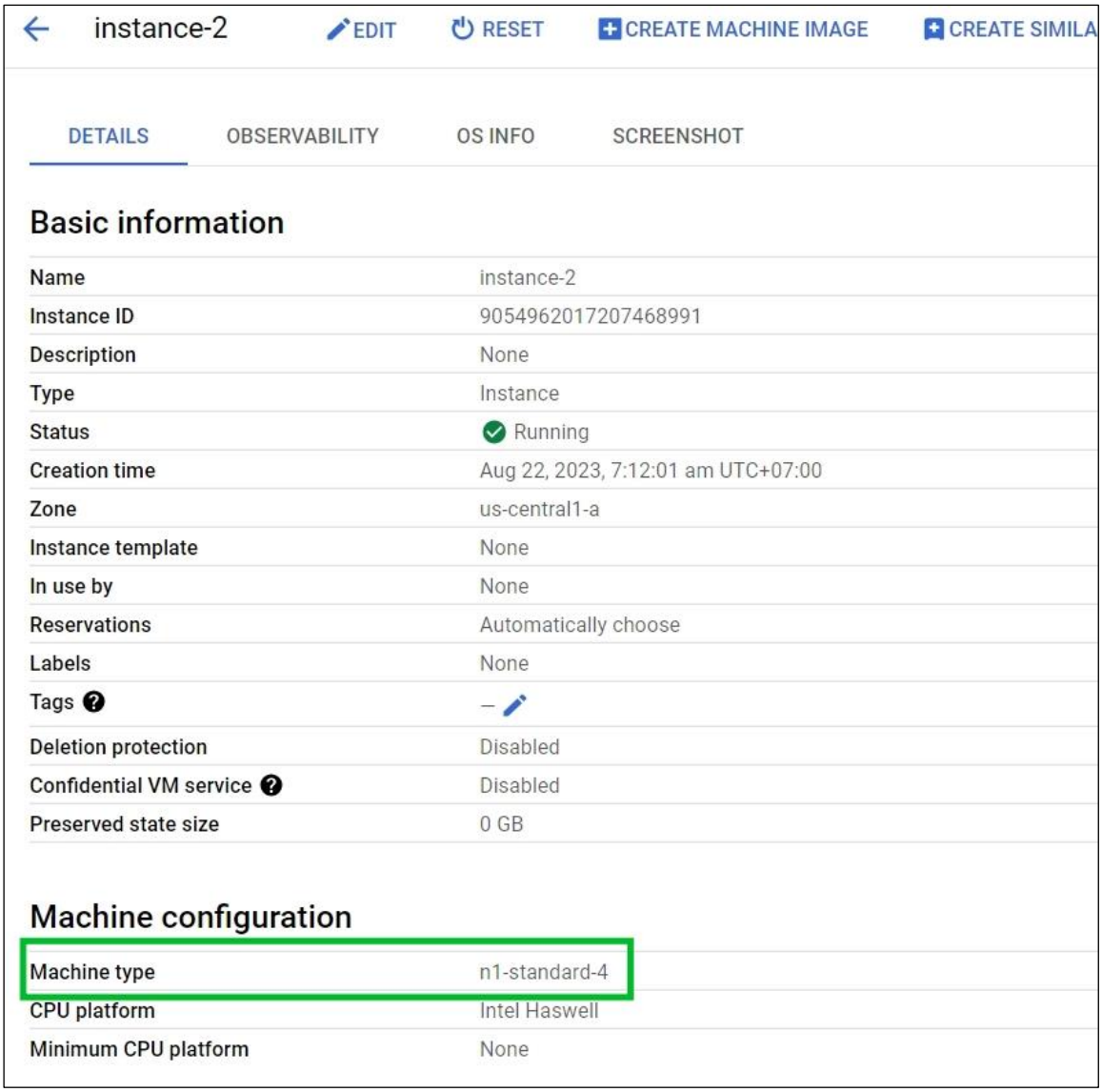
ภาพที่ 3.3 สร้างเครื่อง Compute Engine เครื่องที่ 1

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 3.4 สร้างเครื่อง Compute Engine เครื่องที่ 2

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 3.5 สร้างเครื่อง Compute Engine เครื่องที่ 3

โดยเครื่อง Compute Engine มีไว้เพื่อติดตั้งไลบรารี่ Twitter API ในการอ่านข้อมูลจาก Twitter และรันคำสั่งภาษาโปรแกรมในการส่งข้อมูลที่ได้รับจาก Twitter เข้าสู่ Pub/Sub

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

3.3 เตรียมข้อมูลและส่งเข้าสู่ Pub/Sub

ขั้นตอนการเตรียมข้อมูลเพื่อสร้างเป็น Topic นั้น เราได้ทำการดึงข้อมูลที่ต้องการโดยระบุชื่อบัญชี Twitter ที่เราต้องการทั้งสิ้นจำนวน 80 บัญชี ซึ่งเป็นบัญชีที่เปิดให้เข้าถึงแบบสาธารณะ (Public) และกำหนดปริมาณในการดึงข้อมูลต่อบัญชี สูงสุดต่อครั้ง อยู่ที่ 10,000 โพสต์ โดยทำการดึงข้อมูลทั้งหมด 3 รอบ และใช้ตัวกลางในการดึงข้อมูลชื่อ Apify API ที่มีไลบรารี่กลางที่มีความสามารถในการดึงข้อมูลจากบัญชี Twitter ที่เปิดให้เข้าถึงแบบ Public ได้

โดย Apify เป็นเป็น Platform ที่ต้องจ่ายค่าบริการในการใช้งาน ที่มีการให้บริการหลายแบบ ซึ่งหนึ่งในบริการดังกล่าว คือการดึงข้อมูลจาก Twitter ที่เรียกว่า Twitter Scraper โดยสามารถกำหนดขนาดสูงสุดของจำนวน Post ในแต่ละบัญชีที่ต้องการดึงข้อมูลได้ (แต่จำนวนที่ดึงได้จริงก็ไม่เกินจำนวนโพสต์ที่เจ้าของบัญชี Twitter ได้ทำการโพสต์ใน Twitter)

```python
from apify_client import ApifyClient
import json
import time
from google.cloud import pubsub_v1
from google.oauth2 import service_account
import datetime

API_TOKEN = "apify_api_RrtgGeCZYehKzuee8lM0s0mKYOIZZm3kpm8M"


# Initialize the ApifyClient with your API token
client = ApifyClient(API_TOKEN)

# Prepare the Actor input
run_input = {
    "handles": ["MFPThailand", "PheuThaiParty", "democratTH" , "prachatai", "pprpthailand"
                , "ChartthaiPT", "ChartpattanaKLA","thaisangthai", "khonbhumjaithai", "sereeruamthai"
                ,"thestandardth","VoiceTVOfficial", "WiseKnow"
                ,"ThaksinLive", "Pannika_FWP","Thavisin", "jomquan", "suthichai"
                ,"johnwinyu", "SitaDivari", "Pita_MFP",  "Thanathorn_FWP", "sorrayuth9111"
                ,"HoneKrasae", "EJanNews","Mono29News","MorningNewsTV3","Thairath_TV"
                ,"ati_boons","Noom_Kanchai","Kitti3Miti","thapanee3miti","thepchaiyong","Apirak_BKK"
                ,"paritw92","wirojlak","Chaithawat_MFP","KornGoThailand","Suriyasai"
                ,"KhaosodOnline","thaich8news","onenews31","RDnews_Online","PPTVHD36","amarintvhd"
                ,"TNAMCOT","wpnews23","MatichonOnline",
                "ingshin","RangsimanRome","paritchi","phumtham","Ten_Nattawut","sirotek"
                ,"sudaratofficial","EakThestandard","woodytalk","wongthanong"
```

ภาพที่ 3.6 แสดงรายชื่อบัญชีผู้ใช้ที่ต้องการดึงข้อมูล

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์


```
TwitterStream > GCE > PubSub > tweeterscrap.py > ...
34                  ],
35      "tweetsDesired": 10000,
36      "proxyConfig": { "useApifyProxy": True },
37  }
38
39  start_time = time.perf_counter()
40  # Run the Actor and wait for it to finish
41  run = client.actor("quacker/twitter-scraper").call(run_input=run_input)
42
43  countdataset = client.dataset(run["defaultDatasetId"])
44
45  # Service Account File
46  key_path = "astute-might-394223-cf4bcbc3b568.json"
47  credentials = service_account.Credentials.from_service_account_file(
48      key_path,
49      scopes=["https://www.googleapis.com/auth/cloud-platform"]
50  )
51
52  # Pub/Sub Client
53  pubsub_client = pubsub_v1.PublisherClient(credentials=credentials)
54
```

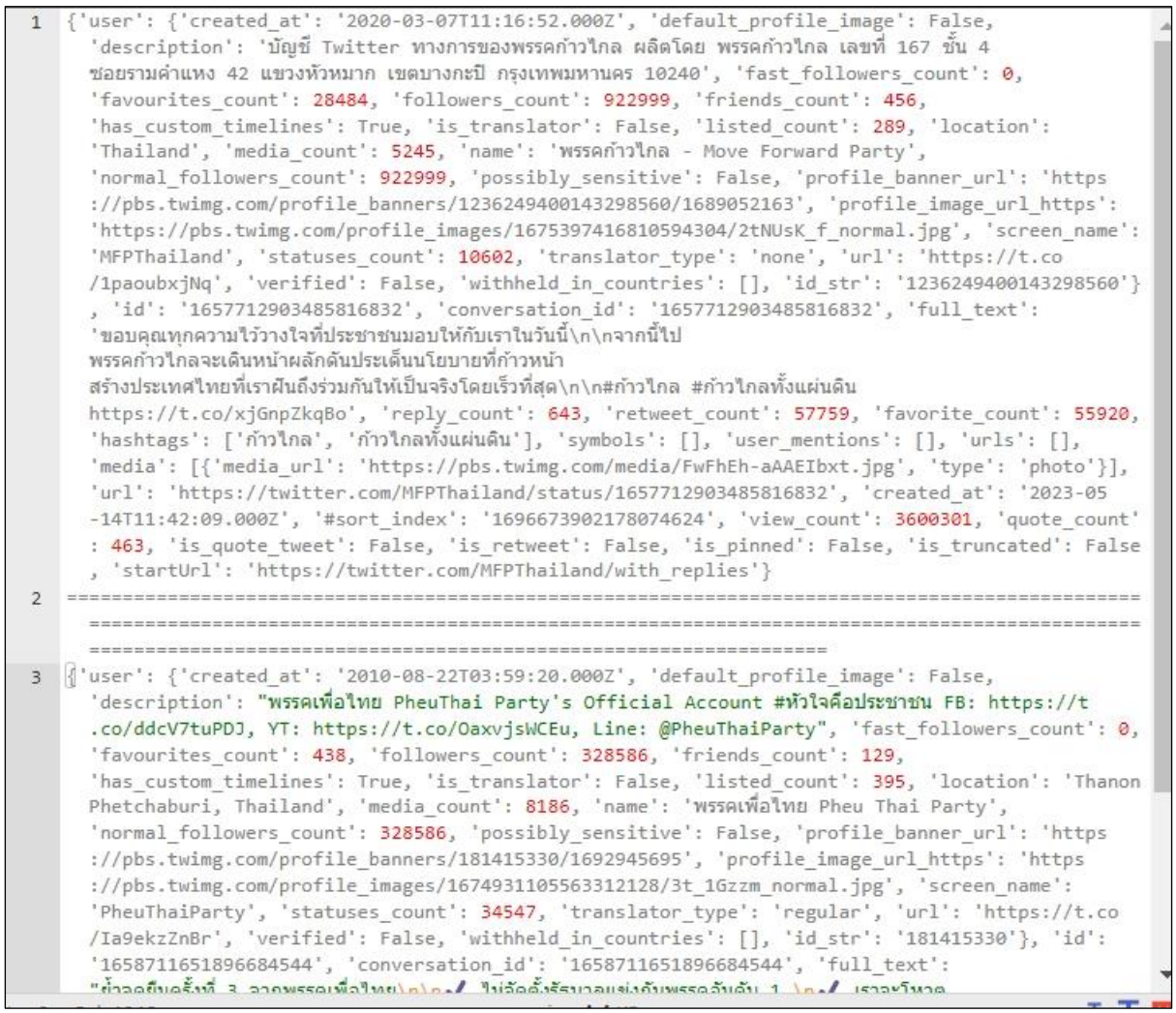
ภาพที่ 3.7 แสดงการกำหนดค่าจำนวนโพสต์สูงสุดที่ต้องการดึงต่อบัญชี

หลังจากเชื่อมต่อกับ Twitter API เสร็จเรียบร้อย เราจะทำการเลือกข้อมูลและจัดการกับรูปแบบของข้อมูลที่จะทำการสื่อสารกับ Pub/Sub โดยทำการแปลงรูปแบบให้มีลักษณะแบบ key - value และเลือกแอทริบิวต์ที่เหมาะสมต่อการนำไปใช้งาน โดยจัดรูปแบบและแปลงข้อมูลแบบ String ที่ได้รับมาจาก API ดังแสดงตามตารางที่ 1 และรูปภาพที่ 15 ตามลำดับ

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 3.1 แสดงรายชื่อแอตทริบิวต์ของ twitter

แอตทริบิวต์หลัก	แอตทริบิวต์ย่อย	คำอธิบายข้อมูล
user	created_at	วันที่สร้างบัญชีผู้ใช้
	description	คำบรรยายบัญชี
	followers_count	จำนวน follower
	favourites_count	จำนวนการกด liked
	location	สถานที่ที่บัญชีถูกสร้าง
	name	ชื่อที่ปรากฏที่หน้า Profile ของบัญชี
	screen_name	ชื่อ url ของบัญชี
	statuses_count	จำนวนที่โพสต์ทั้งหมด
	verified	การ verified บัญชี
id		Id ของโพสต์
conversation_id		มีค่าเดียวกับ id ของโพสต์
full_text		ข้อความที่โพสต์
reply_count		จำนวนการตอบ Comment
retweet_count		จำนวนการ retweet
favorite_count		จำนวนที่โพสต์ได้รับการกด Like
hashtags		รายการแฮชแท็ก
symbols		สัญลักษณ์ที่อยู่ใน full_text
user_mentions		บัญชีผู้ที่กล่าวถึง
urls		url ที่ปรากฏใน full_text หรือ url ที่อ้างถึงเนื้อหาโพสต์
media		ไฟล์เสียง รูป วิดีโอ ที่โพสต์
url		url ของโพสต์
created_at		วันที่โพสต์
quote_count		จำนวนครั้งของ Quote
is_quote_tweet		ค่าที่บอกว่าโพสนี้เป็นโพสที่เป็น Quote หรือไม่
is_retweet		ค่าที่บอกว่าโพสต์นี้มีการ Re-tweet หรือไม่
is_pinned		ค่าที่บอกว่าโพสต์นี้ถูกปักหมุดไว้หรือไม่
is_truncated		ค่าที่บอกว่าโพสต์มีความยาวเกินกว่ากำหนดและถูกตัดข้อความบางส่วนออก
startUrl		url เริ่มต้น



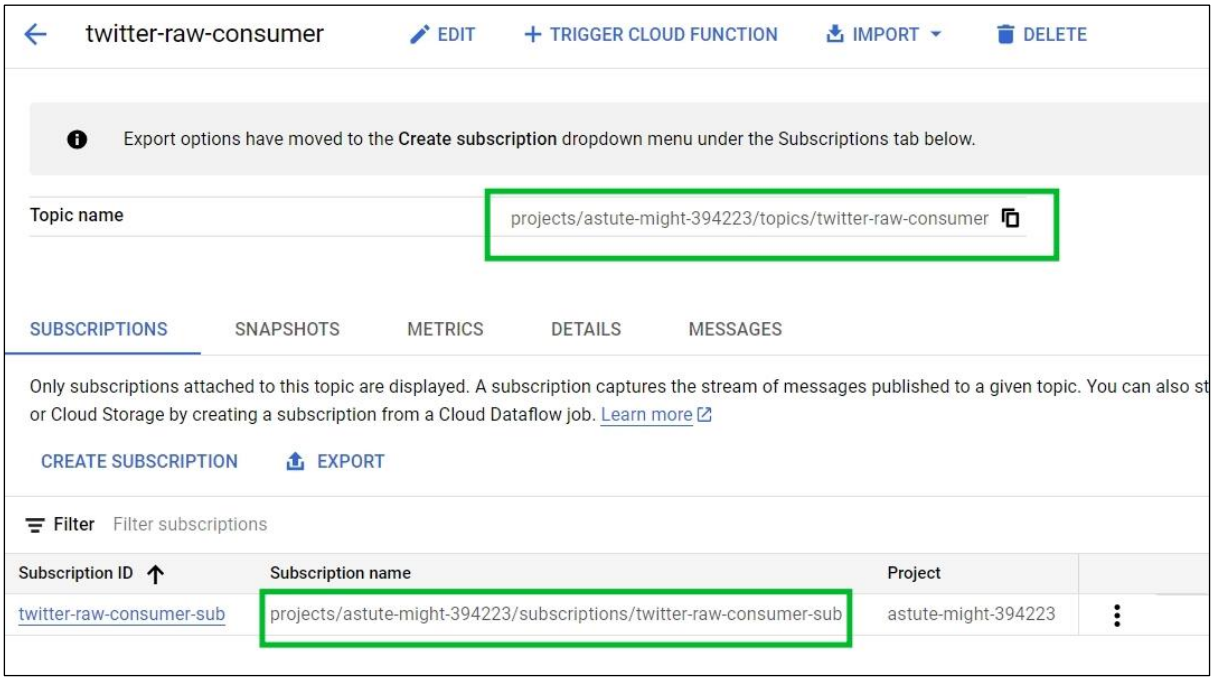
ภาพที่ 3.8 แสดงข้อมูลแบบ String ที่ได้จาก Twitter API

การส่งข้อความที่ได้รับจาก twitter เข้าสู่ Pub/Sub คือการสร้าง Topic ผ่าน API ของ Google ที่ชื่อ Google Cloud pubsub_v1 โดยต้องทำการสร้าง Publisher และ Subscriber ไว้บนคลาวด์ ซึ่งในที่นี้ผู้วิจัยได้สร้าง Publisher ไว้ 3 Publisher และ สร้าง Consumer ไว้ 3 Consumer เพื่อให้รองรับข้อมูลหัวข้อที่มากจากการประมวลผลของทั้ง 3 เครื่องดังต่อไปนี้

ตารางที่ 3.2 แสดงรายชื่อ Pub/Sub

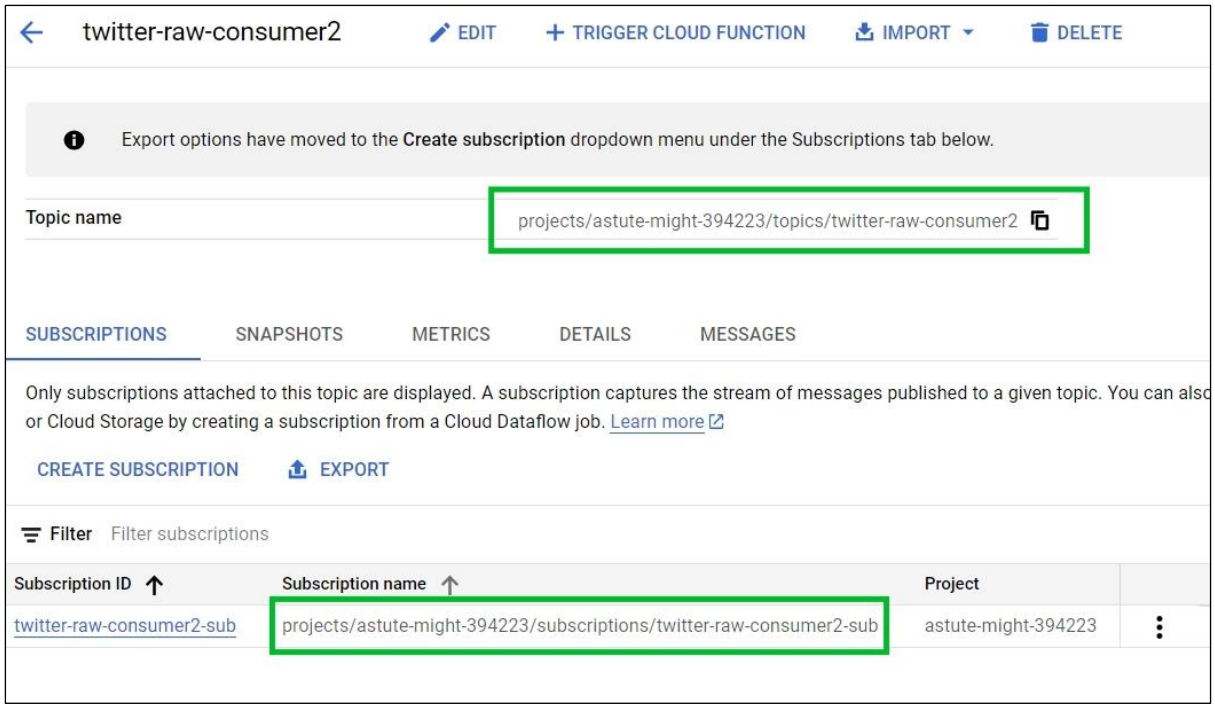
Publisher Topic	Subscriber	Computer Engine ที่ส่งข้อมูล
twitter-raw-consumer	twitter-raw-consumer-sub	เครื่องที่ 1
twitter-raw-consumer2	twitter-raw-consumer2-sub	เครื่องที่ 2
twitter-raw-consumer3	twitter-raw-consumer3-sub	เครื่องที่ 3

โดยแต่ละ Pub/Sub ได้ทำการระบุให้ส่งข้อมูลไปเก็บที่ BigQuery ซึ่งสามารถแสดงรูปภาพตอนสร้าง Pub/Sub ทั้ง 3 Services ได้ตามรูปภาพที่ 16 ถึง ภาพที่18 ตามลำดับ

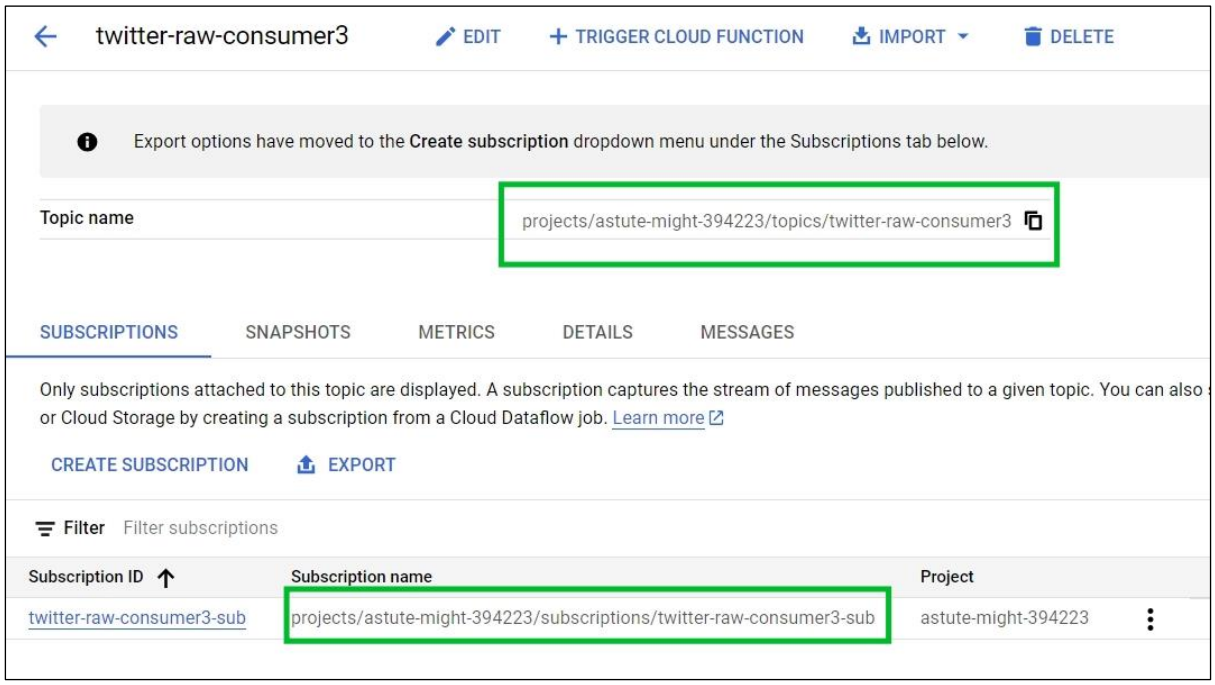


ภาพที่ 3.9 แสดง Pub/Sub สำหรับเครื่องที่ 1

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 3.10 แสดง Pub/Sub สำหรับเครื่องที่ 2



ภาพที่ 3.11 แสดง Pub/Sub สำหรับเครื่องที่ 3

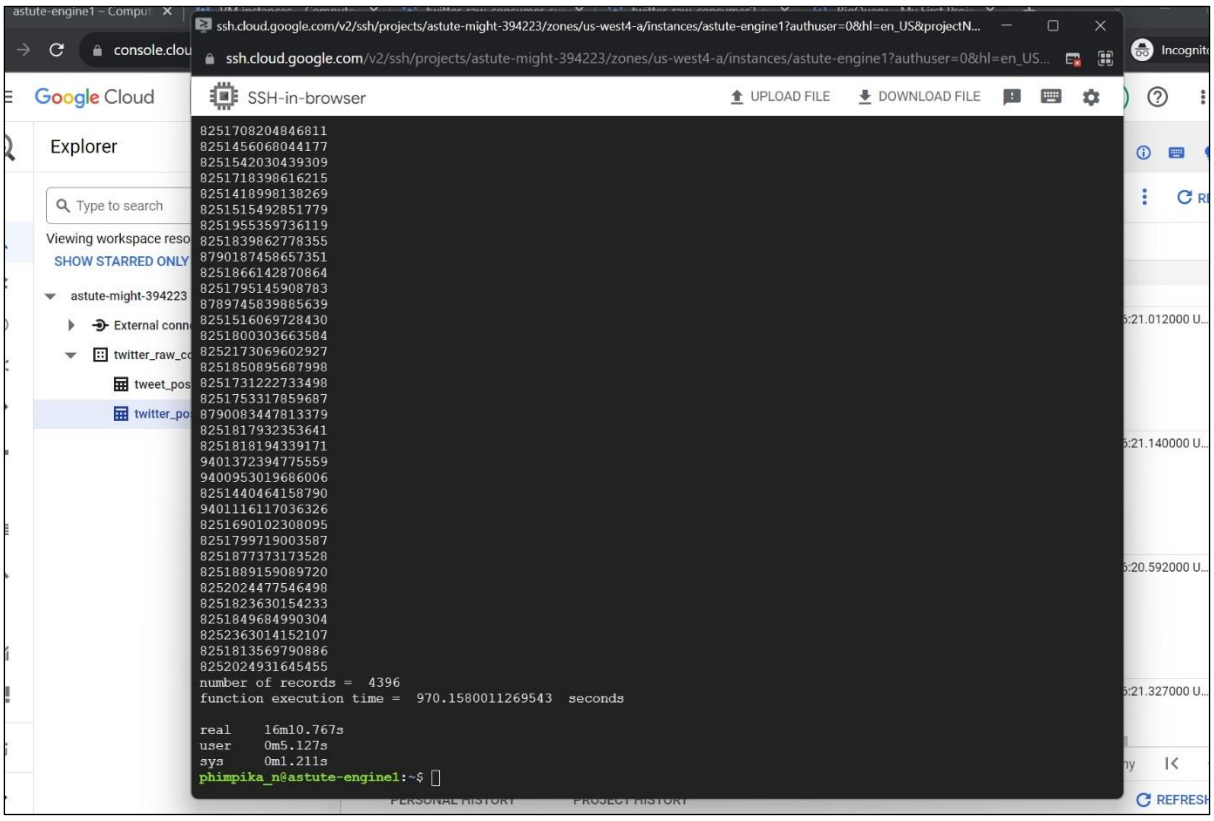
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

เมื่อทำการดึงข้อมูล จนสร้าง Topics ใน Sub/Sub ได้มีการเขียนคำสั่งให้เครื่องแสดงเวลาประมวลผลทั้งหมดออกมาด้วย เพื่อต้องการทราบเวลาในประมวลผลของแต่ละเครื่อง โดยมีปัจจัยที่ควบคุมให้ใกล้เคียงกัน ได้แก่

1) จำนวนบัญชี Twitter ที่ดึงข้อมูล 80 บัญชี

2) จำนวนข้อมูลโพสต์ที่ดึงของแต่ละบัญชี จำกัดไว้เกิน 10,000 โพสต์

3) ดึงต่อเนื่องกัน 3 รอบ เพื่อให้จำนวนโพสต์ที่ดึงได้ในแต่ละรอบมีจำนวนเท่ากัน หรือใกล้เคียงกันมากๆ เนื่องจากเป็นข้อมูล Real-Time จึงไม่อาจทราบจำนวนข้อมูลต้นทางที่ถูกสร้างขึ้นมาว่ามีความรวดเร็วเท่าใด หรือเปลี่ยนแปลงเป็นจำนวนเท่าไหร่ในแต่ละครั้งที่ดึงข้อมูล



ภาพที่ 3.12 แสดงตัวอย่าง ผลการส่งข้อมูลเข้าสู่ Pub/Sub และบอกเวลาประมวลผลในแต่ละรอบ

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์


```
r = mb_string # byte from twitter to string
dt = caltime1

dt2 = caltime2
dt3 = caltime3

plt.xlabel("time (ms)")
plt.ylabel("string MB")
plt.title('Stream Data calculation time')
plt.plot(dt, r)
plt.plot(dt2, r)
plt.plot(dt3, r)
```

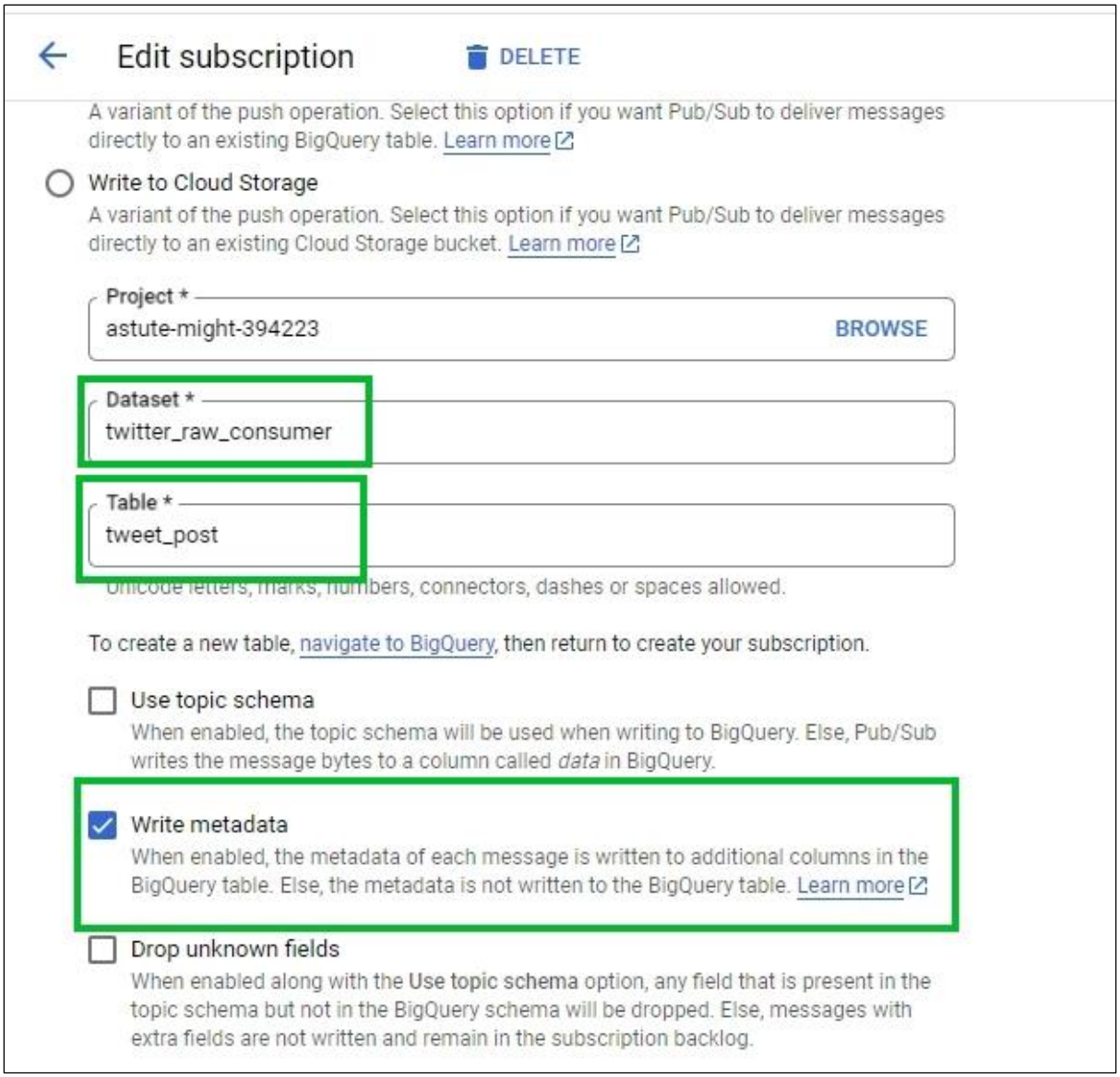
ภาพที่ 3.13 แสดงการจัดเก็บผลและวาดกราฟแสดงผลลัพธ์การประมวลผล

3.4 ส่งข้อมูลเข้าสู่ BigQuery

สร้าง Meta Data ใน Subscriber เพื่อส่งข้อมูลเข้าสู่ BigQuery โดยระบุข้อมูลต่อไปนี้

1) ชื่อ Data Set
2) ชื่อ ตาราง
3) ระบุให้จัดเก็บข้อมูลตาม Metadata แบบเดียวกับ Subscriber

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 3.14 แสดงการกำหนดค่า Subscriber เพื่อส่งข้อมูลเข้าสู่ BigQuery

กระบวนการทำงานของ Pub/Sub จะเป็นไปโดยอัตโนมัติเมื่อเรากำหนดค่าให้ตัว Subscriber ทำการส่งข้อมูลเข้าสู่ BigQuery ทันทีที่ Topic ได้ทำการส่งออกข้อความ ซึ่งขั้นตรงนี้จะลดภาระงานในการเขียนโปรแกรมนำเข้าข้อมูลด้วยตัวเองได้ ซึ่งข้อมูลที่ส่งเข้าสู่ BigQuery มีรูปแบบตาม metadata ของ Topic โดยอธิบายตามตาราง

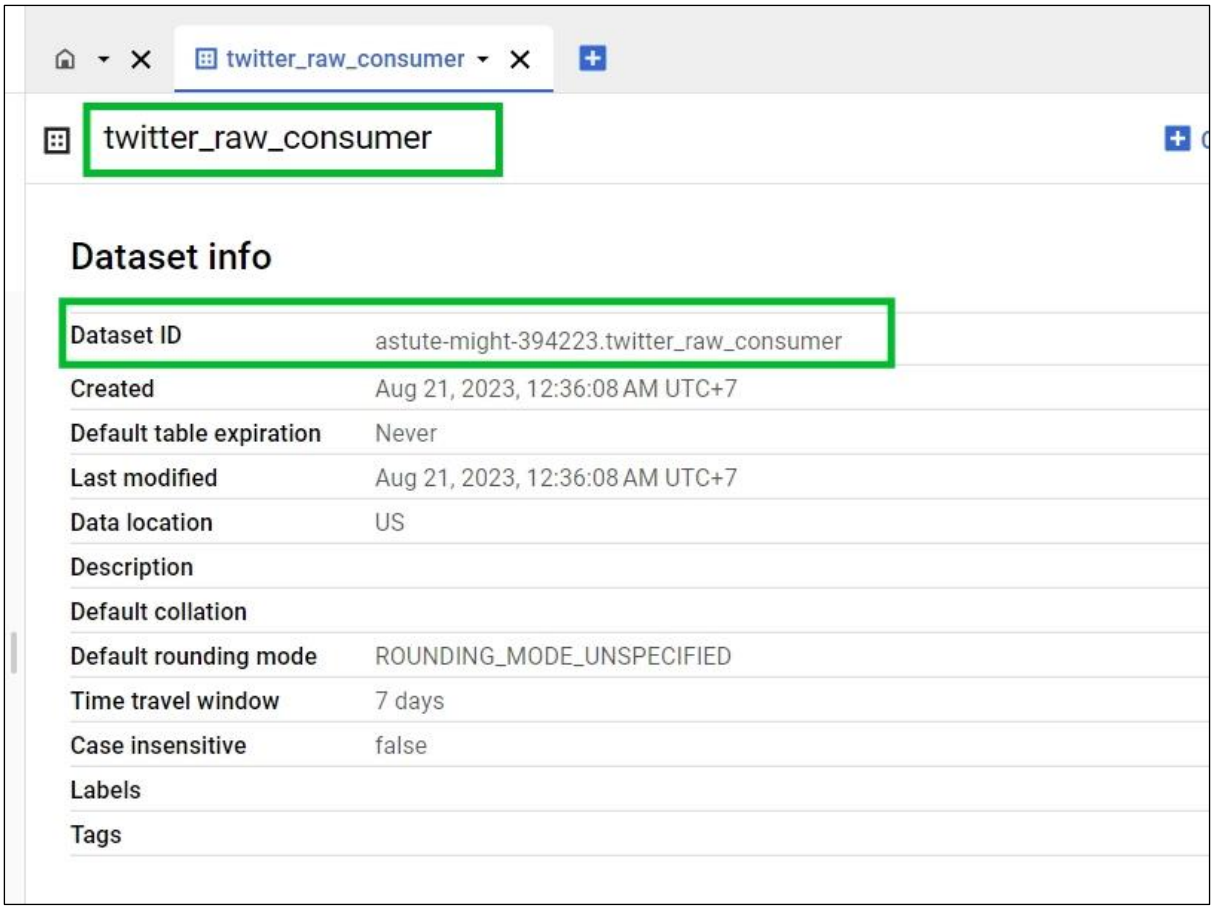
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ตารางที่ 3.3 แสดงรายชื่อ metadata ของ Subscriber

ชื่อแอตทริบิวต์	ข้อมูลที่จัดเก็บ
data	คือข้อมูลทุกๆ แอตทริบิวต์ที่ได้รับจาก Twitter API ซึ่งอยู่ในรูปแบบ Json
attributes	รายชื่อ แอตทรบิวต์ทั้งหมด ที่ได้ประกาศไว้ใน Source Code ตอนสร้าง Topic ในกรณีต้องการสร้างแอตทริบิวต์เพิ่มเติม หรือจัดระเบียบข้อมูล หรือคำนวนค่าบางอย่างก่อนส่งเข้า Topic
message_id	เป็นรหัสหรือหมายเลขประตัวของข้อความ (Message) ที่ Pub/Sub ได้ทำการสร้างให้อัตโนมัติ
publish_time	เป็นเวลาที่ข้อความ (Message) ถูก Publish
Subscription_name	ชื่อ Subscriber ที่ดึงข้อมูลจาก Topic และนำเข้าสู่ฐานข้อมูล

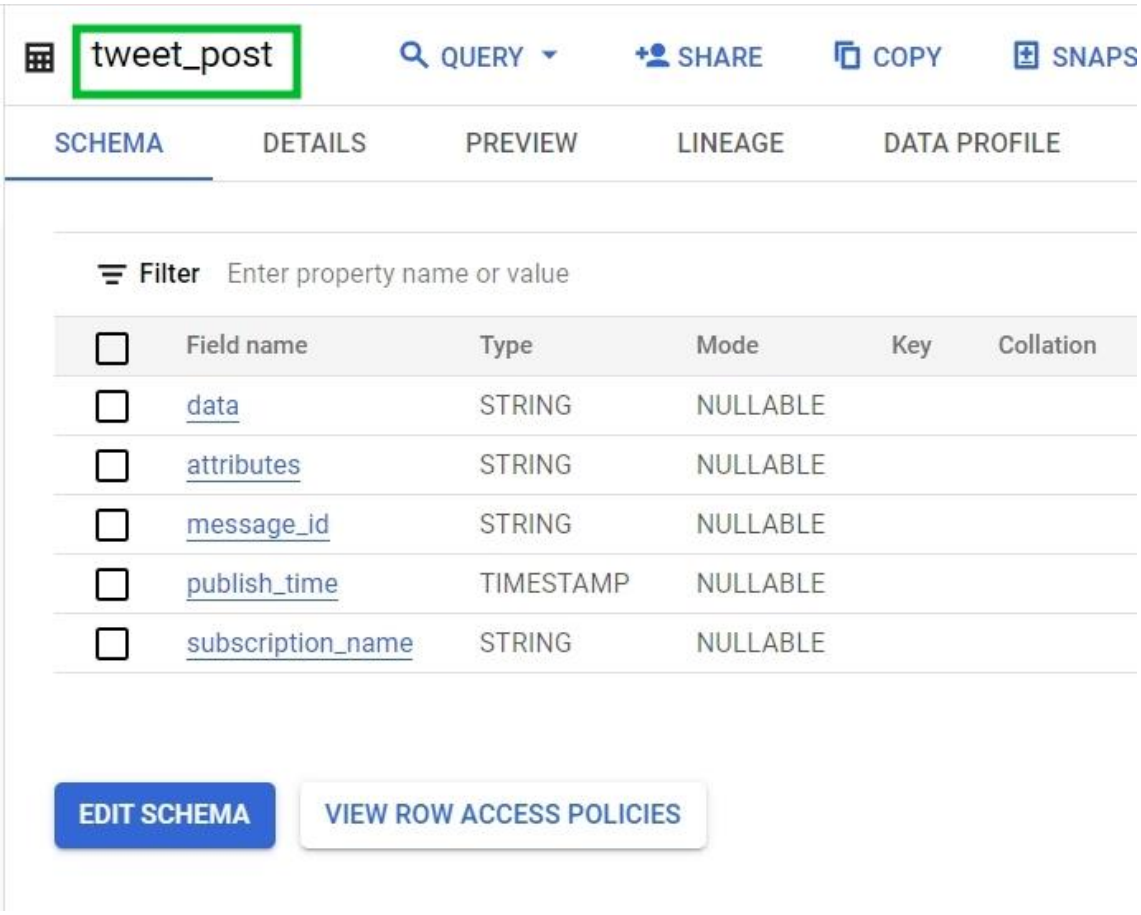
ในการเก็บข้อมูลใน BigQuery จะต้องทำการสร้าง Dataset และตาราง (Table) สำหรับจัดเก็บข้อมูลด้วย โดยสร้าง Dataset จำนวน 1 Dataset และสร้างตาราง ไว้ภายใต้ Dataset นั้นจำนวน 3 ตาราง ให้เก็บข้อมูลที่ Subscriber จะทำการส่งเข้ามา ดังแสดงได้ตามรูปภาพที่ 22 ถึงภาพที่ 26 ต่อไปนี้

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



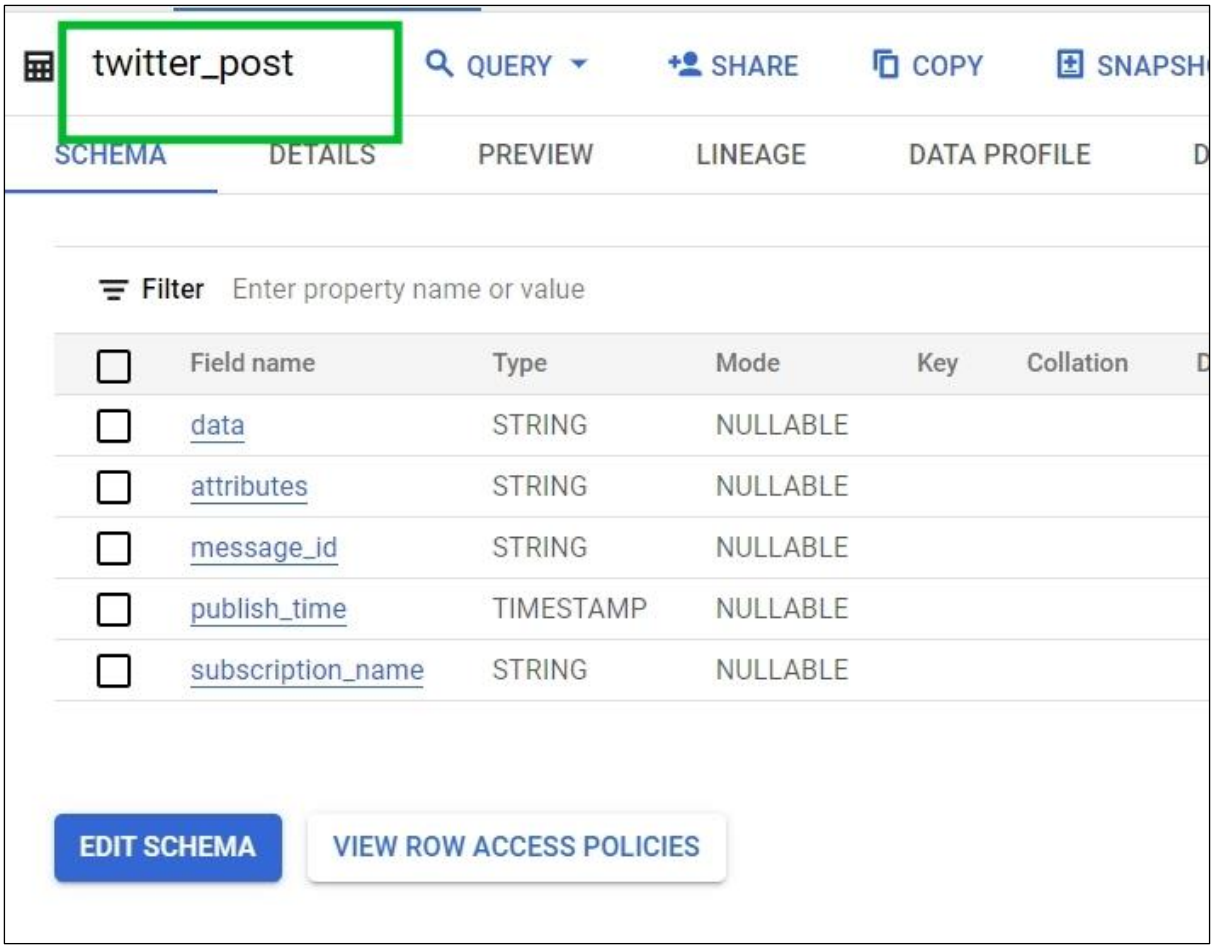
ภาพที่ 3.15 แสดงการสร้าง Dataset

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



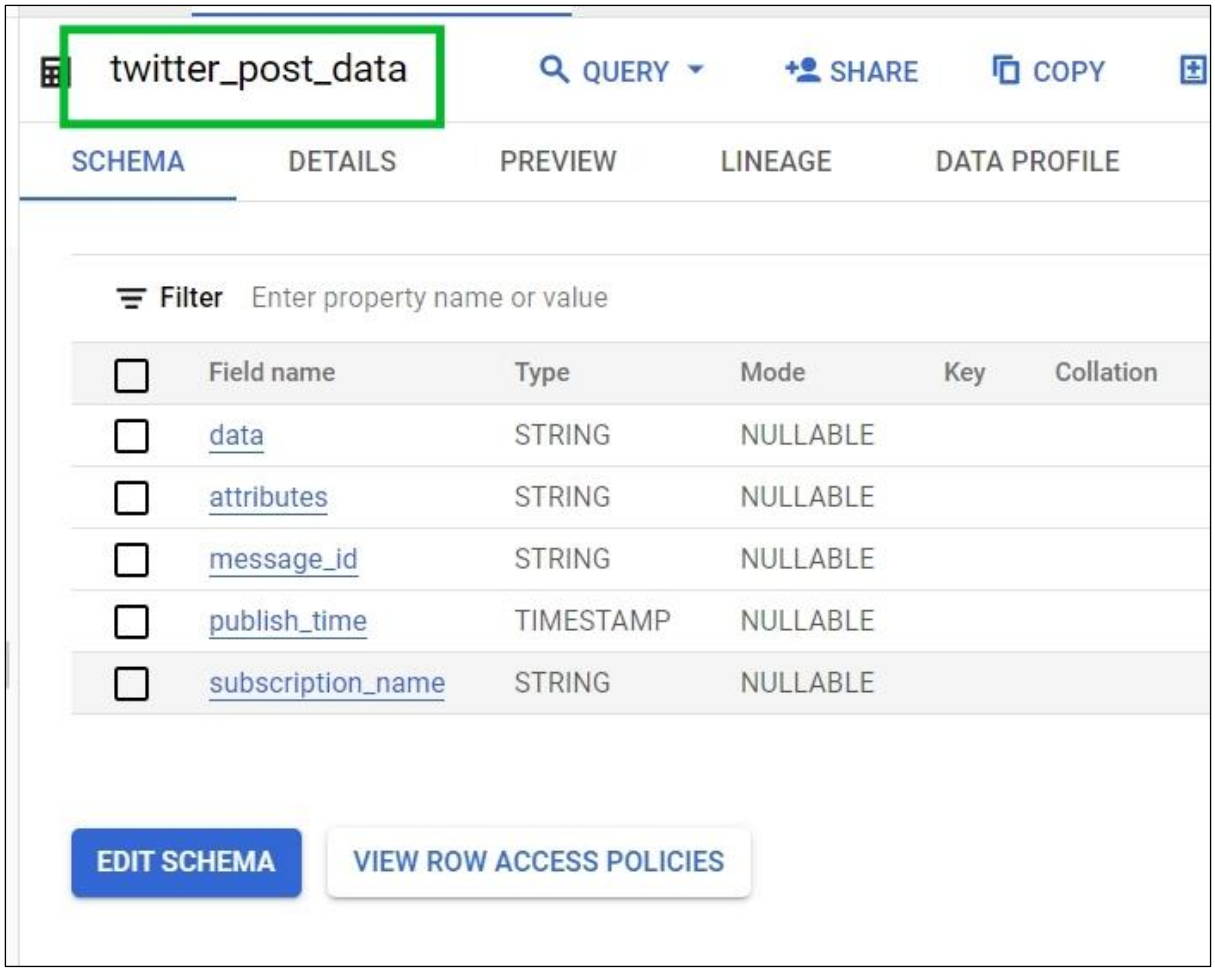
ภาพที่ 3.16 แสดงการสร้างตาราง tweet_post สำหรับจัดเก็บข้อมูลจาก Pub/Sub ชุดที่ 1

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



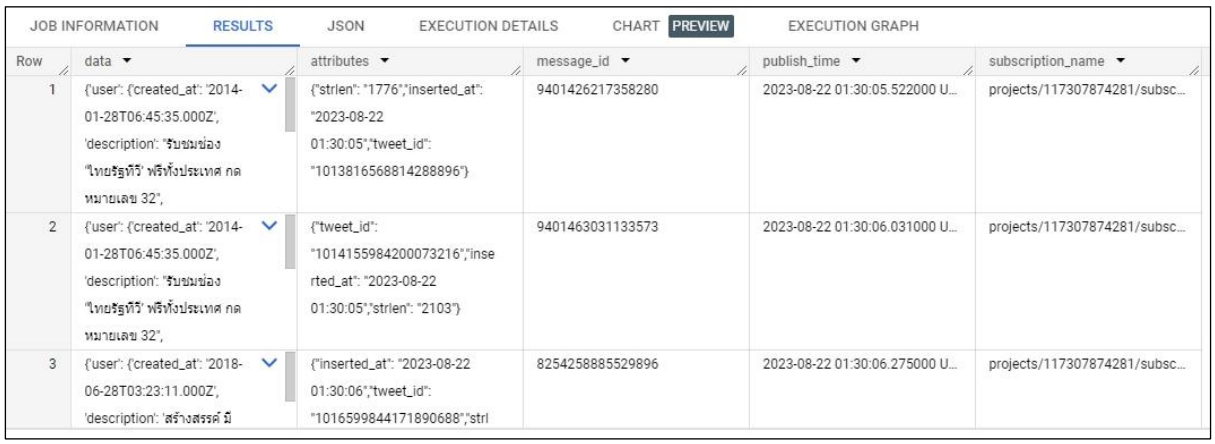
ภาพที่ 3.17 แสดงการสร้างตาราง twitter_post สำหรับจัดเก็บข้อมูลจาก Pub/Sub ชุดที่ 2

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 3.18 แสดงการสร้างตาราง twitter_post_data สำหรับจัดเก็บข้อมูลจาก Pub/Sub ชุดที่ 3

โดยข้อความที่ส่งเข้าสู่ BigQuery จะเป็นข้อความในรูปแบบ Json ซึ่งประกอบด้วย key, value



ภาพที่ 3.19 แสดงข้อมูลที่จัดเก็บใน BigQuery

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บทที่ 4
ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษากระบวนการการรวบรวมข้อมูลจาก Twitter ในรูปแบบ Real-Time เพื่อวัดผลการรวบรวมข้อมูลในด้านเวลาที่ใช้ประมวลผลข้อมูลและความสูญหายของข้อมูล โดยมีรายละเอียดของผลการศึกษาดังต่อไปนี้

4.1 ผลการดึงข้อมูลจาก Twitter

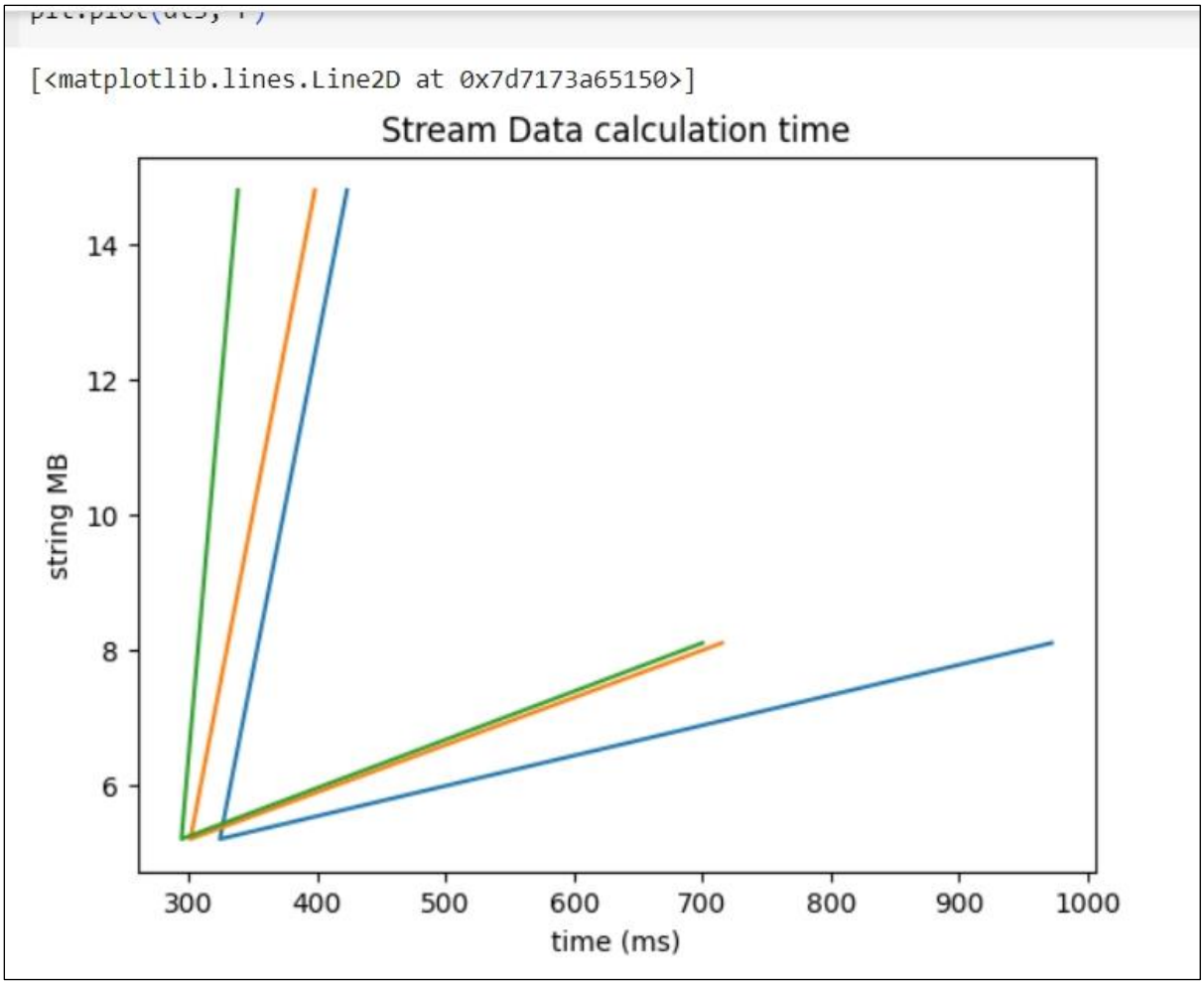
ผลการรันคำสั่งภาษาโปรแกรมในการส่งข้อมูลที่ได้รับจาก Twitter เข้าสู่ Pub/Sub และบอกเวลาประมวลผลข้อมูลบนเครื่องทั้ง 3 เครื่อง โดยตารางเวลาการประมวลผล และขนาดข้อมูล จำนวนรายการในการดึงข้อมูล 1 ครั้ง แสดงได้ต่อไปนี้

ตารางที่ 4.1 ผลการดึงข้อมูลจาก Twitter

เครื่อง	ขนาดเครื่อง	ครั้งที่	จำนวนรายการ	ขนาดข้อมูลทั้งหมด (BYTE)	เวลาประมวลผล (Seconds)
เครื่องที่ 1	CPU 1 หน่วย RAM 3.75 GB	1	4419	8173051	972.15
เครื่องที่ 2	CPU 2 หน่วย RAM 7.50 GB	1	4720	7750797	715.23
เครื่องที่ 3	CPU 4 หน่วย RAM 15 GB	1	4147	6442673	700.03
เครื่องที่ 1	CPU 1 หน่วย RAM 3.75 GB	2	6920	5203529	324
เครื่องที่ 2	CPU 2 หน่วย RAM 7.50 GB	2	6920	8,223,103	301
เครื่องที่ 3	CPU 4 หน่วย RAM 15 GB	2	6920	9,934,597	294
เครื่องที่ 1	CPU 1 หน่วย RAM 3.75 GB	3	6920	14803529	423
เครื่องที่ 2	CPU 2 หน่วย RAM 7.50 GB	3	6920	15,123,103	398
เครื่องที่ 3	CPU 4 หน่วย RAM 15 GB	3	6920	16,834,597	338

จากตารางผลลัพธ์ ผู้เขียนได้ทำการจัดเก็บผลดังกล่าวไว้ในรูปแบบตัวแปรและวาดออกมาในรูปแบบกราฟแสดงผล ซึ่งแสดงให้เห็นว่าจำนวนข้อมูลที่ดึง กับเวลาประมวลผลแตกต่างกันในแต่ละเครื่อง ที่ได้กำหนดขนาดเครื่องเอาไว้ต่างกัน ดังภาพที่ 27

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

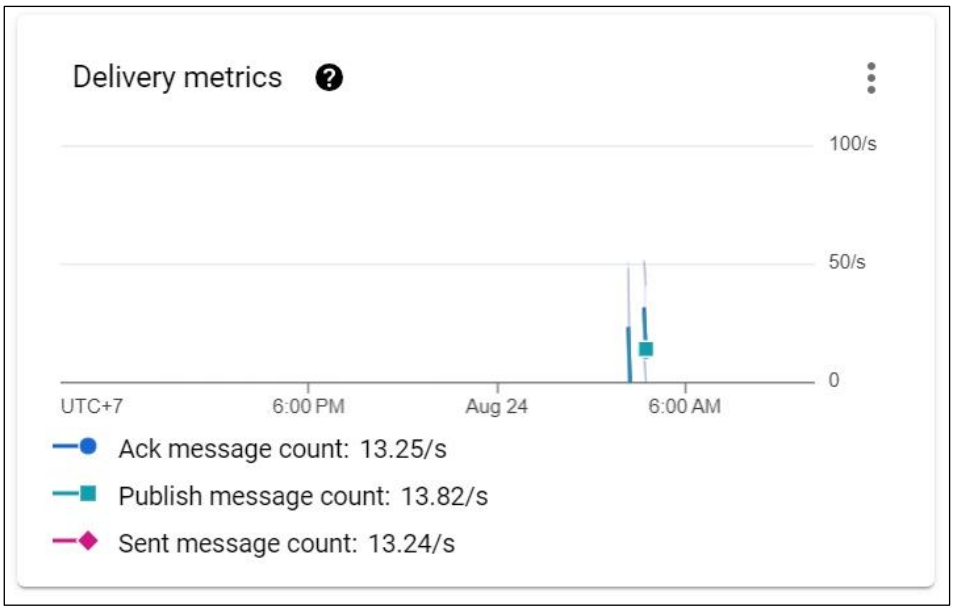


ภาพที่ 4.1 ผลการประมวลผลข้อมูลด้วยเครื่อง Compute Engine ทั้ง 3 เครื่อง

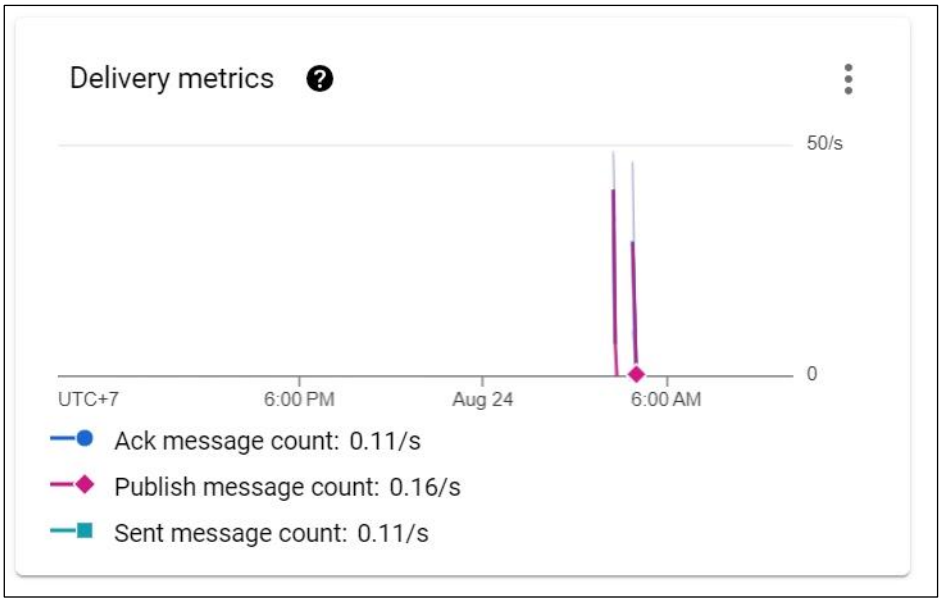
DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

4.2 ผลการทำงานของ Pub/Sub

ผลการทำงานของ Pub/Sub ทั้ง 3 Pub/Sub พบว่าไม่เกิด Downtime หรือการสูญหายของข้อความ (Message) โดยกราฟที่แสดงจำนวนเวลาของการดึงและรับข้อความ แต่ไม่มีปรากฏกราฟที่แสดงจำนวนข้อความที่สูญหายเลย

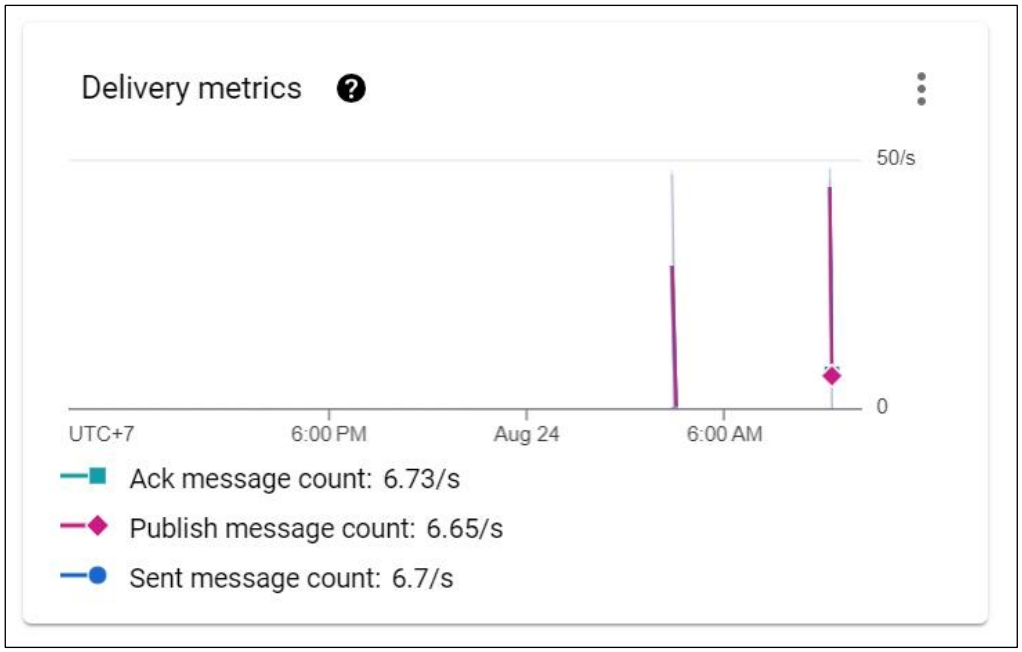


ภาพที่ 4.2 แสดงจำนวนการประมวลผล Pub/Sub ที่ 1



ภาพที่ 4.3 แสดงจำนวนการประมวลผล Pub/Sub ที่ 2

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์



ภาพที่ 4.4 แสดงจำนวนการประมวลผล Pub/Sub ที่ 3

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จากการดำเนินงานเพื่อจัดเก็บข้อมูลจาก Twitter แบบ Real-Time ด้วยการใช้ Pub/Sub เข้าสู่ฐานข้อมูล BigQuery สรุปได้ดังนี้

5.1.1 ระยะเวลาในการประมวลผลของเครื่องคอมพิวเตอร์ทั้ง 3 เครื่องยังไม่พบว่ามีความแตกต่างกันมากจนเกินไป เมื่อเทียบกับจำนวนข้อความที่ดึงในแต่ละรอบ ทำให้เมื่อนำไปใช้งานอาจสามารถใช้เครื่องประมวลผลที่มีคุณสมบัติ (Spec ของเครื่อง) ระดับ CPU 1 ตัว และ Memmory ขนาด 1.75 ในการดึงข้อมูลได้ เนื่องจากการดึงแต่ละรอบได้กำหนดจำนวนสูงสุดของข้อความที่ได้จาก Twitter API ไว้แล้ว

5.1.2 การส่งข้อมูลเข้าสู่ Pub/Sub ไม่พบว่าข้อมูลที่ได้รับจาก Twitter API มีการสูญหาย

5.1.3 การใช้เวลา 1 รอบของ Pub/Sub โดยนับจาก Publish ไปจนดึงนำเข้าสู่ BigQuery พบว่าใช้เวลาน้อยมาก เมื่อเทียบกับปริมาณข้อมูลตามที่ทำการทดสอบ

5.1.4 จำนวนข้อมูลที่ BigQuery ได้รับตรงกันกับจำนวนข้อมูลที่ถูก Subscribe ในแต่ละครั้ง

5.2 ข้อเสนอแนะ

5.2.1 เนื่องจากการดึงข้อมูลเป็นการดึงแบบ Real-Time ทำให้ข้อมูลที่ได้รับมาอาจมีความซ้ำกัน ดังนั้นก่อนจะนำไปประมวลผลข้อมูลต้องทำการกำจัดค่าที่ซ้ำกันเพื่อให้ไม่เกิดความซ้ำซ้อน

5.2.2 การวัดผลเพื่อตัดสินใจเรื่องการติดตั้งเครื่อง Compute Engine อาจต้องทำการทดสอบเพิ่มด้วยการเพิ่มปริมาณข้อมูลจากแหล่งอื่นๆร่วมด้วย เนื่องจากการ Stream ข้อมูลผ่าน Twitter API เองก็มีขีดจำกัดในการดึงต่อครั้ง จึงทำให้ไม่พบการโหลดข้อมูลจำนวนมากได้

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บรรณานุกรม

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บรรณานุกรม

[1] TANMOY KANUNGO, "Processing Streaming Data using Google Pub/sub and BigQuery", [ออนไลน์]. https://www.linkedin.com/pulse/processing-streaming-data-using-google-pubsub-bigquery-tanmoy-kanungo. (เข้าถึงเมื่อ: 16 สิงหาคม 2566)

[2] SARIN, "Solving BigData with BigQuery", [ออนไลน์]. https://blog.cleverse.com/2018/03/29/solving-bigdata-with-bigquery-1/. (เข้าถึงเมื่อ: 16 สิงหาคม 2566)

[3] Pervaiz Akhtar and Samir Dani, "Real-time big data processing for instantaneous marketing decisions: A problematization approach", *Industrial Marketing Management, vol. 90,* pp. 563–565. *Accessed on: Aug. 16, 2023.* [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019850118307454?via%3Dihub

[4] Witold Maranda, Aneta Poniszewska-Maranda, and Małgorzata Szymczyńska, "Data Processing in Cloud Computing Model on the Example of Salesforce Cloud", MDPI, 13(2), pp. 3-4. Accessed on: Aug. 17, 2023. [Online]. Available: https://www.mdpi.com/2078-2489/13/2/85.

[5] Minghe Wang, Trever Schirmer, Tobias Pfandzelter and David Bermbach, "Lotus: Serverless In-Transit Data Processing for Edge-based Pub/Sub" arXiv, pp. 2-3, Mar, 2023, doi: 2303.07779

[6] Pub/Sub, (ม.ป.ป.). [ออนไลน์] Available: https://cloud.google.com/pubsub.

[7] What is BigQuery, (ม.ป.ป.). [ออนไลน์]. https://cloud.google.com/bigquery/docs/introduction.

[8] axual, What is BigQuery, 2020 [ออนไลน์]. Available: https://axual.com/top-things-to-know-about-real-time-data-processing/.

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

บรรณานุกรม (ต่อ)

[9] วัฒนา เมธาวศิน, Real-Time Programming and System, wachum, pp. 11-18. เข้าถึงเมื่อ: 17 สิงหาคม 2566. [ออนไลน์]. Available: http://www.wachum.com/dewey/000/com21.pdf

[10] สำนักส่งเสริมอุตสาหกรรมเทคโนโลยีสารสนเทศและการสื่อสาร, “คู่มือการเลือกใช้บริการ Cloud Computing”, [ออนไลน์]. https://www.mdes.go.th/content/download-detail/2907. (เข้าถึงเมื่อ: 18 สิงหาคม 2566)

DPU DHURAKIJ PUNDIT UNIVERSITY มหาวิทยาลัยธุรกิจบัณฑิตย์

ประวัติผู้เขียน

ชื่อ – นามสกุล พิมพิกา นางาม

ประวัติการศึกษา

พ.ศ. 2552 - ปริญญาวิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์)
มหาวิทยาลัยราชภัฏอุบลราชธานี

ประสบการณ์ทำงาน

พ.ศ. 2566 - Senior Programmer,
บริษัท เอ สตาร์ เทคโนโลยี จำกัด