กระบวนการประมวลผลข้อมูลแบบ Real-Time จากสื่อสังคมออนไลน์บนคลาวด์

CLOUD BASE REAL-TIME DATA PROCESSING FROM SOCIAL MEDIA DATA

พิมพิกา นางาม[1]
ดร. เอกสิทธิ์ พัชรวงศ์ศักดา[2]

บทคัดย่อ

ในปัจจุบันที่การเข้าถึงบริการออนไลน์ต่างๆ เกิดขึ้นเป็นจำนวนมาก ได้ก่อให้เกิดข้อมูลจำนวนมาก ทำให้องค์กรต่าง ๆ ตระหนักถึงความสำคัญในการพยายามเข้าถึงและนำไปวิเคราะห์ผลในเชิงลึก ซึ่งหนึ่งในแหล่งข้อมูลที่ใช้ในการตัดสินใจ 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

---

[1] นักศึกษาหลักสูตรวิศวกรรมมหาบัณฑิต สาขาวิศวกรรมข้อมูลขนาดใหญ่

[2] ที่ปรึกษาสารนิพนธ์หลัก

คำสำคัญ : การประมวลผลข้อมูล, เรียลไทม์ ดาต้า, การประมวลผลข้อมูลแบบเรียลไทม์, ระบบรับส่งข้อความ

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 that can be an important factor for decision and operational processes 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 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 important information can be used for analysis to discover the customer insight.

This study aims to use Twitter's data to describe Real-time data processing on Google Cloud using virtual machines (VM) to connect to the Twitter API and 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 in order to provide scalability computing resources and human recruitment.

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

บทนำ

ในยุคที่เทคโนโลยีและสื่อสังคมออนไลน์มีบทบาทสำคัญในการติดต่อและแลกเปลี่ยนข้อมูลระหว่างบุคคลและองค์กร การใช้ข้อมูลจากแพลตฟอร์มสื่อสังคมต่าง ๆ เพื่อนำมาใช้ในการวิเคราะห์มีความสำคัญอย่างมากต่อธุรกิจในหลายด้าน เช่น การเข้าใจกลุ่มเป้าหมาย, การติดตามและวิเคราะห์แนวโน้ม เป็นต้น ซึ่งมีประโยชน์ต่อองค์กรให้มีความสามารถในการแข่งขันทางธุรกิจ

จากความสำคัญดังกล่าว จึงเป็นที่มาของการศึกษาในครั้งนี้ในการรวบรวมข้อมูลจาก Twitter ซึ่งเป็นแหล่งข้อมูลสำคัญในการใช้วิเคราะห์ข้อมูล โดยการเลือกใช้เครื่องมือที่เหมาะสม สะดวกต่อการดูแลรักษาระบบ สามารถใช้งานได้ง่ายและไม่มีการเกิด ดาวน์ไทม์ (Downtime) หรือช่วงเวลาที่ระบบหยุดทำงาน เนื่องจาก

ข้อมูลใน Twitter เป็นข้อมูลที่มีความเคลื่อนไหวอยู่ตลอดเวลา การรวบรวมและประมวลผลข้อมูลจึงควรใช้เครื่องมือที่เหมาะสมและป้องกันความสูญหายของข้อมูล (Losing Messages) ที่อาจเกิดขึ้นได้

วัตถุประสงค์ของงานวิจัย

เพื่อศึกษากระบวนการการรวมรวมข้อมูลจาก Twitter ในรูปแบบ Real-Time วัดผลเพื่อเลือกโครงสร้างทางคอมพิวเตอร์ที่เหมาะสมต่อการนำไปในการรวบรวมและวิเคราะห์ข้อมูล

ขอบเขตงานวิจัย

1. ข้อมูลบ้านมือสองที่มีการลงประกาศขาย ในเดือนพฤษภาคมถึงเดือนมิถุนายน ปี พ.ศ. 2566 บนเว็บไซต์ประกาศขายบ้าน https://www.baania.com จากการทำ web scrapping ข้อมูลในจังหวัดปทุมธานีซึ่งเป็นจังหวัดหนึ่งในเขตปริมณฑล ประเทศไทย ในช่วงราคาไม่เกิน 10 ล้านบาท
2. ข้อมูล POIs ในจังหวัดปทุมธานี ได้แก่ โรงเรียน, สถานีบริการพลังงาน, ร้านสะดวกซื้อ, ไฮเปอร์มาร์เก็ต, ธนาคาร, บริการสาธารณสุข, มหาวิทยาลัย, ร้านกาแฟ, ร้านขายยา, ศูนย์การค้า, สถานีรถไฟฟ้า BTS

ทฤษฎี และผลงานที่เกี่ยวข้อง

1. ทฤษฎี

Data Processing คือการเก็บรวบรวมข้อมูลจากแหล่งต่างๆ เพื่อนำมาใช้ในการประมวลผลหรือวิเคราะห์ เป็นขั้นตอนแรกของการประมวลผลข้อมูล ซึ่งสามารถกระทำได้หลายวิธี นับตั้งแต่ การส่งแบบสอบถาม การสัมภาษณ์ การทดลอง หรือการใช้เครื่องมือที่เหมาะสมช่วยในการรวบรวมข้อมูลที่ต้องการเพื่อให้ได้ข้อมูลที่ถูกต้องตามสภาพความเป็นจริงมากที่สุด

Real-Time Data คือชุดข้อมูลที่มีความสดใหม่และเก็บซ้ำในระดับที่สามารถนำเอาข้อมูลชุดนั้นไปใช้งานได้ก่อนข้อมูลจะมีการเปลี่ยนแปลง หรือถ้าเป็นทางการเก็บข้อมูล ก็คือมีการเก็บข้อมูลทุกครั้งที่มีการเปลี่ยนแปลงเกิดขึ้น ข้อมูลแบบ Real-Time มักเป็นข้อมูลที่เกิดจากการ Search หรือสื่อสังคมออนไลน์ (Social Media) ต่างๆ หรือเป็นข้อมูลจาก Logs การใช้งานต่างๆ เช่น Web Traffic เป็นต้น ข้อมูลเหล่านี้เกิดขึ้นรวดเร็วและมีความถี่มาก จนกลายเป็นข้อมูลที่มีขนาดใหญ่ และมีค่าเปลี่ยนแปลงอยู่ตลอดเวลา

Real-Time Processing 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

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

2. งานวิจัยที่เกี่ยวข้อง

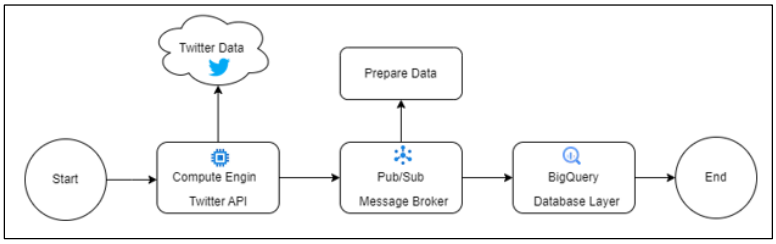
Pervaiz Akhtar and Samir Dani, Real-time big data processing for instantaneous marketing decisions: A problematization approach, (2020) [3] การศึกษาเพื่อหาความเป็นไปได้ในการพัฒนาระบบสนับสนุนการตัดสินใจแบบ Real-Time สำหรับการสร้างเคมเปญทางการตลาดเพื่อตอบสนองและค้นหากลุ่มลูกค้าให้ได้ตรงตามเป้าหมาย โดยการใช้เครื่องมือในรูปแบบคลาวด์ ได้แก่ MapReduce, NOSQL และ HDFS ในการประมวลผลข้อมูลทั้งในรูปแบบ Structure Data และ Unstructured Data เพื่อให้เกิดความได้เปรียบในการแข่งขันในระดับ B2B

Witold Maranda, Aneta Poniszewska-Maranda, Data Processing in Cloud Computing Model on the Example of Salesforce Cloud, (2022) [4] เปรียบเทียบวิธีการประมวลผลข้อมูลและจัดเก็บข้อมูล ที่รวบรวมจากแหล่งต่างๆ ด้วยการใช้ระบบคลาวด์ในรูปแบบ SaaS ด้วยผู้ให้บริการ Salesforce cloud เพื่อหารูปแบบที่เหมาะสมในการกระบวนการรวบรวมข้อมูล ประมวลผลข้อมูลทั้งในรูปแบบ Batch Processing และ Real-Time Processing

Minghe Wang, Trever Schirmer, Tobias Pfandzelter and David Bermbach, Lotus: Serverless In-Transit Data Processing for Edge-based Pub/Sub, (2023) [5] ระบบส่งและแปลงข้อมูลที่ชื่อ Lotus ซึ่งทำงานด้วยรูปแบบ Pub/Sub ในการส่งข้อมูลจากอุปกรณ์เซ็นเซอร์ เพื่อให้ทำหน้าที่เป็นตัวกลางในการรวบรวมข้อมูลจากอุปกรณ์ IoT ต่าง และ Publish หรือส่งออกไปในรูปแบบ Topic โดยเป็นการทำงานอยู่บนระบบบคลาวด์แบบ Function-as-a-Service โดยไม่ต้องมีการติดตั้งเซิร์ฟเวอร์ หรือเครื่องมือในการพัฒนาระบบขึ้นมาใช้งานเลย

3. ระเบียบวิธีวิจัย

ผู้วิจัยขั้นตอนการประมวลผลข้อความและดำเนินการสร้างระบบสำหรับตรวจสอบข้อความบนเอกสารได้ตามวัตถุประสงค์ของงานจึงได้กำหนดกระบวนการการดำเนินการเพื่อศึกษาข้อมูลมีขั้นตอน ปรากฏตามภาพที่ 3



ภาพที่ 1 ภาพรวมวิธีดำเนินการวิจัย

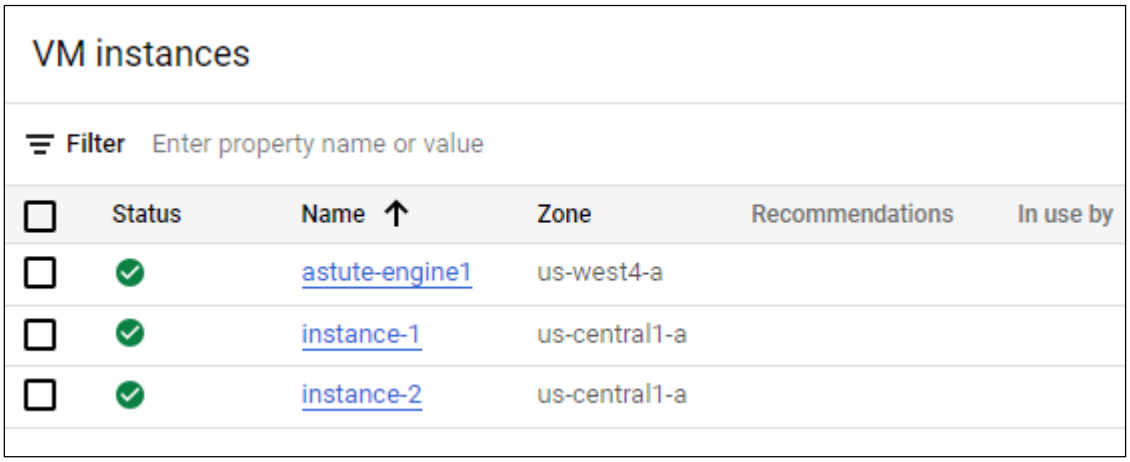
1. สร้าง **Compute Engine** และติดตั้ง **Twitter API**

สร้าง Compute Engine หรือเครื่องคอมพิวเตอร์บนคลาวด์ ทั้งหมด **3** เครื่อง เพื่อใช้รันคำสั่งในการอ่านข้อมูลจาก Twitter ด้วย Twitter API และรันคำสั่งภาษาโปรแกรมในการส่งข้อมูลที่ได้รับ เข้าสู่ Pub/Sub ดังต่อไปนี้

1) เครื่องที่ 1 Ubuntu เวอร์ชั่น 20.04 ขนาด ประมวลผล CPU 1 หน่วย RAM 3.75 GB HDD 10 GB

2) เครื่องที่ 2 Ubuntu เวอร์ชั่น 20.04 ขนาด ประมวลผล CPU 2 หน่วย RAM 7.50 GB HDD 10 GB

3) เครื่องที่ 3 Ubuntu เวอร์ชั่น 20.04 ขนาด ประมวลผล CPU 4 หน่วย RAM 15 GB HDD 10 GB



ภาพที่ 2 รูปภาพเครื่อง Compute Engine

2. 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 ที่เป็นไลบรารี่กลางที่มีความสามารถในการดึงข้อมูลจากบัญชี Twitter ที่เปิดให้เข้าถึงแบบ Public ได้

หลังจากเชื่อมต่อกับ Twitter API เสร็จเรียบร้อย เราจะทำการเลือกข้อมูลและจัดการกับรูปแบบของข้อมูลที่จะทำการสื่อสารกับ Pub/Sub โดยทำการแปลงรูปแบบให้มีลักษณะแบบ key - value และเลือกแอทริบิวต์ที่เหมาะสมต่อการนำไปใช้งาน ดังต่อไปนี้

ตารางที่ 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

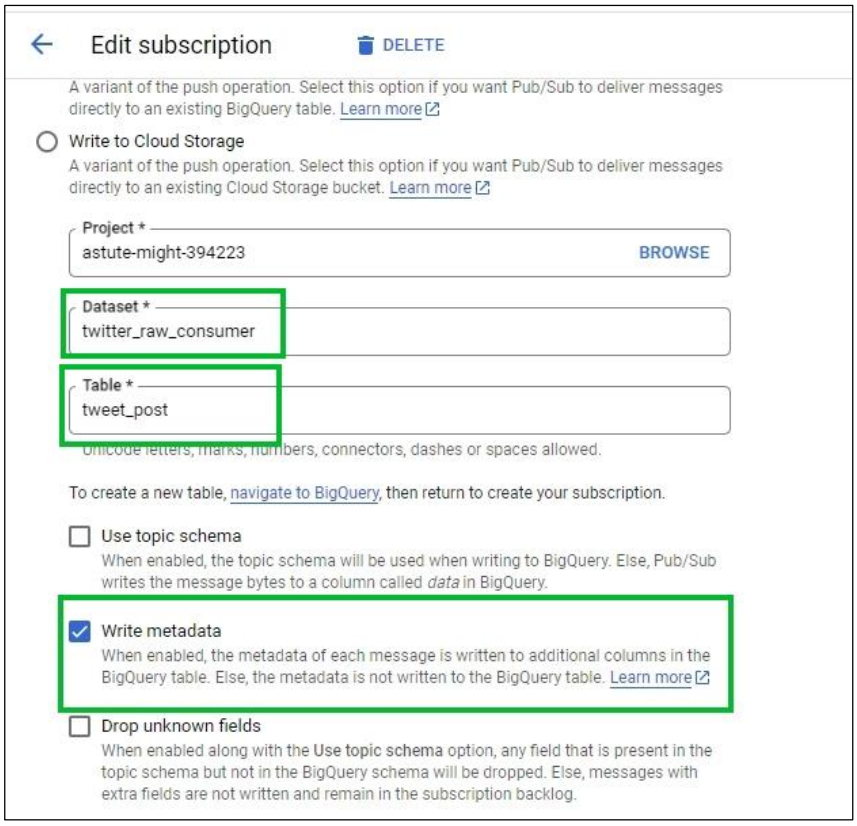
แอตทริบิวต์หลัก	แอตทริบิวต์ย่อย	คำอธิบายข้อมูล
	screen_name	ชื่อ url ของบัญชี
	statuses_count	จำนวนที่โพสต์ทั้งหมด
	verified	การ verified บัญชี
id		Id ของโพสต์
conversation_id		มีค่าเดียวกับ id ของโพสต์
full_text		ข้อความที่โพสต์
reply_count		จำนวนการตอบ Comment
retweet_count		จำนวนการ retweet
favorite_count		จำนวนที่โพสต์ได้รับการกด Like
hashtags		รายการแฮชแท็ก
symbols		สัญลักษณ์ที่อยู่ใน full_text
user_mentions		บัญชีผู้ที่กล่าวถึง
urls		url ที่ปรากฏใน full_text หรือ url ที่อ้างถึงเนื้อหาโพสต์
media		ไฟล์เสียง รูป วิดิโอ 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

3. ส่งข้อมูลเข้าสู่ BigQuery สร้าง Meta Data ใน Subscriber เพื่อส่งข้อมูลเข้าสู่ BigQuery โดยระบุข้อมูลต่อไปนี้

1) ชื่อ Data Set

2) ชื่อ ตาราง

3) ระบุให้จัดเก็บข้อมูลตาม Metadata แบบเดียวกับ Subscriber



ภาพที่ 3 แสดงการกำหนดค่า Subscriber เพื่อส่งข้อมูลเข้าสู่ BigQuery

metadata ของ Topic ซึ่งในรูปแบบ Json ซึ่งประกอบด้วย key, value โดยอธิบายตามตารางที่ 3 และภาพที่ 4 ตามลำดับ

ตารางที่ 2 แสดงรายชื่อ metadata ของ Subscriber

ชื่อแอตทริบิวต์	**ข้อมูลที่จัดเก็บ**
data	คือข้อมูลทุกๆ แอตทริบิวต์ที่ได้รับจาก Twitter API ซึ่งอยู่ในรูปแบบ Json
attributes	รายชื่อ แอตทริบิวต์ทั้งหมด 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

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | CHART PREVIEW | EXECUTION GRAPH

Row	data	attributes	message_id	publish_time	subscription_name
1	{'user': {'created_at': '2014-01-28T06:45:35.000Z', 'description': 'รับชมช่อง 'ไทยรัฐทีวี' ฟรีทั่งประเทศ กด หมายเลข 32",	{"strlen": "1776","inserted_at": "2023-08-22 01:30:05","tweet_id": "1013816568814288896"}	9401426217358280	2023-08-22 01:30:05.522000 U...	projects/117307874281/subsc...
2	{'user': {'created_at': '2014-01-28T06:45:35.000Z', 'description': 'รับชมช่อง 'ไทยรัฐทีวี' ฟรีทั่งประเทศ กด หมายเลข 32",	{"tweet_id": "1014155984200073216","inserted_at": "2023-08-22 01:30:05","strlen": "2103"}	9401463031133573	2023-08-22 01:30:06.031000 U...	projects/117307874281/subsc...
3	{'user': {'created_at': '2018-06-28T03:23:11.000Z', 'description': 'สร้างสรรค์ มี	{"inserted_at": "2023-08-22 01:30:06","tweet_id": "1016599844171890688","strl	8254258885529896	2023-08-22 01:30:06.275000 U...	projects/117307874281/subsc...

ภาพที่ 4 ภาพแสดงข้อมูลที่จัดเก็บใน BigQuery

ผลการศึกษา

1. ผลการดึงข้อมูลจาก Twitter

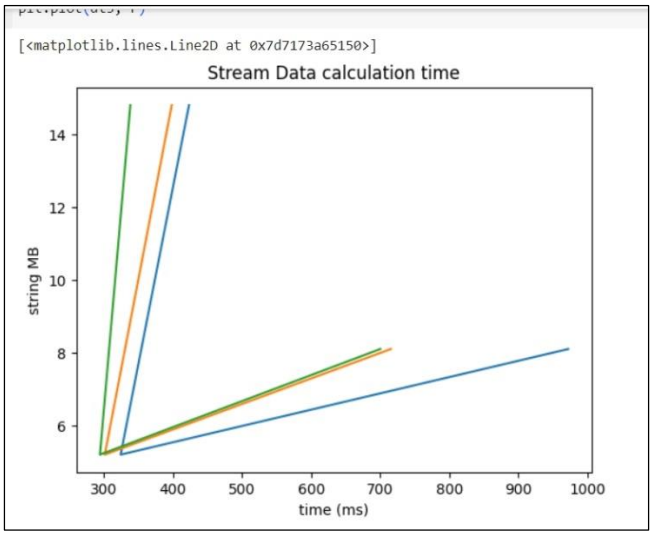
ผลการรันคำสั่งภาษาโปรแกรมในการส่งข้อมูลที่ได้รับจาก Twitter เข้าสู่ Pub/Sub และบอกเวลาประมวลผลข้อมูลบนเครื่องทั้ง **3** เครื่อง โดยตารางเวลาการประมวลผล และขนาดข้อมูล จำนวนรายการ ในการดึงข้อมูล 1 ครั้ง แสดงได้ดังตารางที่ 4

ตารางที่ 3 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

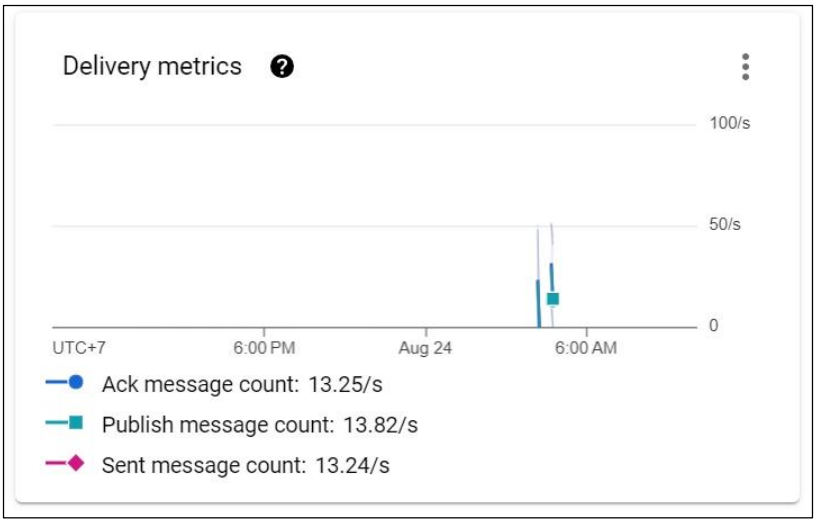
จากตารางผลลัพธ์ ผู้เขียนได้ทำการจัดเก็บผลดังกล่าวไว้ในรูปแบบตัวแปรและวาดออกมาในรูปแบบกราฟแสดงผล ซึ่งแสดงให้เห็นว่าจำนวนข้อมูลที่ดึง กับเวลาประมวลผลแตกต่างกันในแต่ละเครื่อง ที่ได้กำหนดขนาดเครื่องเอาไว้ต่างกัน ดังภาพที่ 4



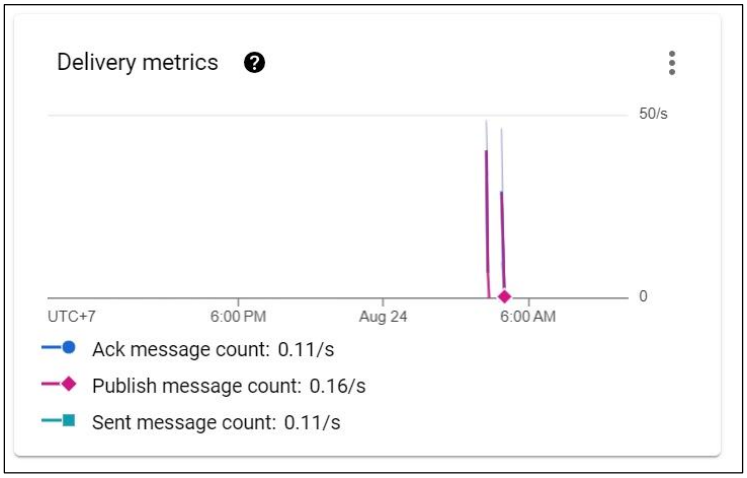
ภาพที่ 5 ผลการประมวลผลข้อมูลด้วยเครื่อง Compute Engine ทั้ง 3 เครื่อง

2. ผลการทำงาน ของ Pub/Sub

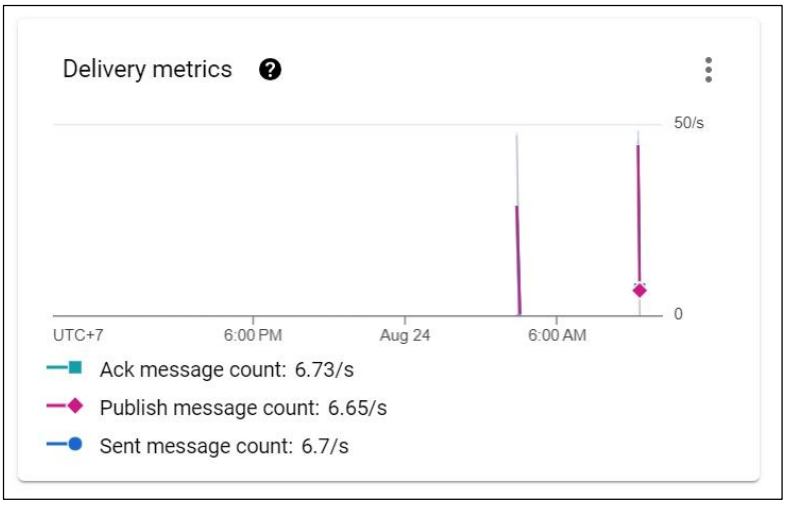
ผลการทำงานของ Pub/Sub ทั้ง 3 Pub/Sub พบว่าไม่เกิด Downtime หรือการสูญหายของข้อความ (Message) โดยกราฟ ที่แสดงจำนวนเวลาของการดึงและรับข้อความ แต่ไม่มีปรากฏกราฟที่แสดงจำนวนข้อความที่สูญหายเลย ดังแสดงตามภาพที่ 6 ถึง ภาพที่ 8 ตามลำดับ



ภาพที่ 6 แสดงจำนวนการประมวลผล Pub/Sub ที่ 1



ภาพที่ 7 แสดงจำนวนการประมวลผล Pub/Sub ที่ 2



ภาพที่ 8 แสดงจำนวนการประมวลผล Pub/Sub ที่ 3

สรุปและอภิปรายผลการวิจัย

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

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

5.2 การส่งข้อมูลเข้าสู่ Pub/Sub ไม่พบว่าข้อมูลที่ได้รับจาก Twitter API มีการสูญหาย

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

5.4 จำนวนข้อมูลที่ BigQuery ได้รับตรงกันกับจำนวนข้อมูลที่ถูก Subscribe ในแต่ละครั้ง

ข้อเสนอแนะ

ในการศึกษานี้ทำการศึกษาเพียงจังหวัดเดียวปทุมธานี เพื่อให้ขอบเขตการศึกษากว้างขึ้น และเพิ่มโอกาสในการปรับปรุงประสิทธิภาพในการประเมินราคา จึงมีข้อเสนอแนะดังต่อไปนี้

1. เนื่องจากการดึงข้อมูลเป็นการดึงแบบ Real-Time ทำให้ข้อมูลที่ได้รับมาอาจมีความซ้ำกัน ดังนั้นก่อนจะนำไปประมวลผลข้อมูลต้องทำการกำจัดค่าที่ซ้ำกันเพื่อให้ไม่เกิดความซ้ำซ้อน

2. การวัดผลเพื่อตัดสินใจเรื่องการติดตั้งเครื่อง Compute Engine อาจต้องทำการทดสอบเพิ่มด้วยการเพิ่มปริมาณข้อมูลจากแหล่งอื่นๆ ร่วมด้วย เนื่องจากการ Stream ข้อมูลผ่าน Twitter API เองก็มีข้อจำกัดในการดึงต่อครั้ง จึงทำให้ไม่พบการโหลดข้อมูลจำนวนมากได้

บรรณานุกรม

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

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.

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: 2303.0777