แพลตฟอร์มการวิเคราะห์ข้อมูลขนาดใหญ่ด้วยราสเบอร์รี่พายคลัสเตอร์และการประยุกต์ใช้ในการวิเคราะห์เครือข่ายสังคม

ธีรวิทย์ สีฆสมิทธิ[1*], วีระ สอิ้ง[2], ศุภชัย ไทยเจริญ[2]

บทคัดย่อ

การวิเคราะห์ข้อมูลขนาดใหญ่โดยทั่วไปจะทำบนระบบคลัสเตอร์คอมพิวเตอร์หรือบนคลาวด์ สำหรับระบบคลัสเตอร์คอมพิวเตอร์ได้มีงานวิจัยที่นำระบบฝังตัว เช่น บอร์ดราสเบอรี่พาย มาแทนที่เครื่องคอมพิวเตอร์ส่วนบุคคล เพื่อสร้างแพลตฟอร์มทางเลือกสำหรับวิเคราะห์ข้อมูลขนาดใหญ่ ที่มีค่าใช้จ่ายไม่สูง กินพลังงานน้อย เคลื่อนย้ายสะดวก และมีประสิทธิภาพ อย่างไรก็ตาม งานวิจัยเหล่านี้เน้นไปที่การเปรียบเทียบสมรรถนะ โดยไม่ได้ให้รายละเอียดในการใช้แพลตฟอร์มสำหรับวิเคราะห์ข้อมูล ดังนั้น บทความนี้เสนอประสบการณ์การพัฒนาและประยุกต์แพลตฟอร์มการวิเคราะห์ข้อมูลขนาดใหญ่ด้วยบอร์ดราสเบอรี่พายและซอฟต์แวร์แบบเปิดเผยต้นฉบับ Apache Hadoop และ Apache Spark สำหรับวิเคราะห์ข้อมูลเครือข่ายเที่ยวบินเพื่อค้นหาสนามบินที่เกิดความล่าช้าของเที่ยวบินบ่อยที่สุด โดยใช้หลักการวิเคราะห์เครือข่ายสังคม ผลการทดลองแสดงให้เห็นว่าแพลตฟอร์มที่พัฒนาขึ้นด้วยบอร์ดราสเบอรี่พายสามารถใช้ในการวิเคราะห์ข้อมูลขนาดใหญ่ได้จริง และด้วยค่าใช้จ่ายในการพัฒนาที่ไม่สูง ทำให้แพลตฟอร์มนี้เป็นประโยชน์ต่อกลุ่มวิจัยทางด้านการวิเคราะห์ข้อมูลขนาดใหญ่ในสถาบันการศึกษาและกลุ่มวิจัยอิสระ

คำสำคัญ : การวิเคราะห์ข้อมูลขนาดใหญ่, คลัสเตอร์คอมพิวเตอร์, ระบบฝังตัว, การวิเคราะห์เครือข่ายสังคม

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author: Tel: 085-162-4678 E-mail address: teerawit.seekasamit@g.swu.ac.th

A Big Data Analytics Platform Using a Raspberry Pi Cluster and Its Application to Social Network Analysis

Teerawit Seekasamit[1*], Vera Sa-ing[2], Supphachai Thaicharoenn[2]

Abstract

Analyzing big data can typically be conducted on an on-promise computer cluster or on cloud. For an on-promise computer cluster, using embedded systems such as Raspberry Pi instead of desktop computers has been proposed by several studies as a low-cost, low power consumption, portable, and efficient alternative. However, those studies are mostly focused only on performance evaluation, which does not provide sufficient applicable information for target users. Therefore, this paper presents an experience on building a big data analytics platform based on a cluster of raspberry pi boards and open-source big data analytics software libraries, Apache Hadoop and Apache Spark and using the platform to analyze large social network data. The experimental results show that a cluster of raspberry pi board can practically be used as a big data analytics platform. Since its low costs and this system could be beneficial for an academic research group or a private study on big data.

Keywords: Big data analytics, Computer cluster, Embedded system, Social-network analysis

---

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

* Corresponding author: Tel: 085-162-4678 E-mail address: teerawit.seekasamit@g.swu.ac.th

บทนำ

การวิเคราะห์ข้อมูลขนาดใหญ่มีความคล้ายคลึงกับการวิเคราะห์ข้อมูลโดยทั่วไปในแง่ที่ว่า มันเป็นกระบวนการในการค้นหาความรู้ที่ซ่อนอยู่ในข้อมูลในรูปแบบของแพทเทิร์นหรือแนวโน้ม ลักษณะที่แตกต่างที่สำคัญคือการวิเคราะห์ข้อมูลขนาดใหญ่ต้องรับมือกับข้อมูลที่มีขนาดใหญ่ ถูกสร้างขึ้นได้อย่างรวดเร็วต่อเนื่อง รวมถึงมีหลากหลายชนิดและรูปแบบ ด้วยข้อจำกัดของฮาร์ดแวร์และซอฟต์แวร์ การวิเคราะห์ข้อมูลที่มีลักษณะดังได้กล่าวมานี้ ไม่สามารถทำได้บนเครื่องคอมพิวเตอร์เครื่องเดียว และใช้ซอฟต์แวร์สำหรับการวิเคราะห์ข้อมูลทั่วไปที่มีอยู่เพราะถูกพัฒนาขึ้นมาสำหรับการประมวลผลบนข้อมูลที่ถูกเก็บในหน่วยความจำ

การวิเคราะห์ข้อมูลขนาดใหญ่โดยทั่วไปถูกดำเนินการบนคอมพิวเตอร์ที่มีสมรรถนะสูงบนคลัสเตอร์คอมพิวเตอร์ หรือบนระบบคลาวด์ อย่างไรก็ตาม การใช้วิธีเหล่านี้มีค่าใช้จ่าย ไม่ว่าจะเป็นด้านฮาร์ดแวร์ ด้านการดูแลรักษา หรือปริมาณทรัพยากรที่ใช้ ซึ่งอาจจะไม่คุ้มค่าที่จะใช้สำหรับกลุ่มวิจัยในสถาบันการศึกษาหรือกลุ่มวิจัยอิสระ อย่างไรก็ตาม ด้วยความก้าวหน้าทางด้านเทคโนโลยีระบบฝังตัวอุปกรณ์ขนาดเล็ก เช่น บอร์ดราสเบอรี่พาย ได้ถูกพัฒนาขึ้นให้มีสมรรถนะเทียบเท่ากับคอมพิวเตอร์ส่วนตัว ด้วยราคาที่ถูกกว่าและใช้พลังงานน้อยกว่ามาก

งานวิจัยที่มีการใช้บอร์ดราสเบอรี่พายสำหรับสร้างคลัสเตอร์คอมพิวเตอร์ มีดังต่อไปนี้ Diwedi และ Sharm วัดสมรรถนะคลัสเตอร์บอร์ดราสเบอรี่พาย โดยประมวลผลข้อมูลบนคลัสเตอร์ที่ใช้บอร์ดราสเบอรี่พาย 1 ตัว 3 ตัว และ 5 ตัว [1] ซึ่งผลการทดลองพบว่า คลัสเตอร์ที่ใช้จำนวนบอร์ด 5 ตัว จะใช้เวลาในการประมวลผลน้อยสุด Parakyriakou ศึกษาเปรียบเทียบสมรรถนะของคลัสเตอร์ราสเบอรี่พายจำนวน 15 บอร์ดที่ติดตั้งเฟรมเวิร์คอาปาเชฮาดูป (Apache Hadoop) โดยวัดสมรรถนะด้วยชุดเครื่องมือเปรียบเทียบมาตรฐาน TestDFSIO [2] เพื่อวัดความเร็วในการประมวลผลไฟล์ที่เก็บบนระบบไฟล์ HDFS ของอาปาเชฮาดูป จากผลการทดลอง Parakyriakou พบว่าสมรรถนะในการอ่านไฟล์เร็วกว่าการเขียนไฟล์ Nugroho และ Widiyanto ออกแบบแพลตฟอร์มการประมวลผลแบบขนาน (Parallel computing platform) โดยสร้างจากคลัสเตอร์ราสเบอรี่พายและซอฟต์แวร์อาปาเชฮาดูป และประเมินผลแพลตฟอร์มด้วยเครื่องมือ Seige tool เพื่อทดสอบสมรรถนะของการรับมือกับปริมาณการร้องขอข้อมูลผ่าน HTTP request [3] ซึ่งจากผลการทดลอง พบว่า จำนวน HTTP requests ที่ทำให้แพลตฟอร์มที่ออกแบบและสร้างขึ้นทำงานได้อย่างเหมาะสมคือ จำนวน 50 การร้องขอ นอกจากนั้น Nugroho และ Widiyanto ยังพบว่า คลัสเตอร์ที่มีจำนวนราสเบอรี่พาย 4 บอร์ดมีประสิทธิภาพสูงกว่าคลัสเตอร์ที่ใช้ราสเบอรี่พายแค่บอร์ดเดียว ถึง 260% งานวิจัยอื่นๆเกี่ยวกับการประเมินสมรรถนะของคลัสเตอร์ราสเบอรี่พาย [4][5][6] ให้ผลลัพธ์ที่คล้ายคลึงกัน

การประยุกต์ใช้ที่สำคัญของคลัสเตอร์คอมพิวเตอร์ที่แตกต่างจากการใช้คอมพิวเตอร์เครื่องเดียว คือการวิเคราะห์ข้อมูลขนาดใหญ่ ซึ่งในบทความนี้ ได้เลือกชุดข้อมูลเที่ยวบินล่าช้ามาใช้ในการศึกษา เนื่องจากความเที่ยวบินที่ล่าช้าเป็นเหตุฉุกเฉินที่สำคัญสำหรับศูนย์ควบคุมการบินที่สนามบิน และเป็นอุปสรรคใหญ่สำหรับธุรกิจการบิน ปัจจัยเสี่ยงที่ทำให้เกิดความล้าช้ามีความหลากหลายและซับซ้อน ดังนั้นการทำความเข้าใจปัจจัยเหล่านี้สามารถเป็นประโยชน์ต่อการลดความน่าจะเป็นในความล่าช้าของเที่ยวบินได้ ทีมวิจัยนำโดย Wang ได้ประยุกต์ใช้หลักการวิเคราะห์เครือข่ายสังคมในการบ่งชี้ถึงปัจจัยเสี่ยงที่ทำให้เกิดการล่าช้าของเที่ยวบินและความสัมพันธ์ของปัจจัยเหล่านี้กับสนามบิน [7] โดยในการศึกษา Wang และทีมวิจัยได้ใช้ค่าระดับความเป็นศูนย์กลาง (Degree centrality) แทนถึงความสามารถในการควบคุมปัจจัยเสี่ยง ดังนั้น เมื่อออกแบบกลยุทธ์ในการควบคุมและกำจัดปัจจัยเสี่ยง ปัจจัยใดที่มีค่าระดับความเป็นศูนย์กลางสูงจะถูกนำมาพิจารณาเป็นลำดับแรกๆสำหรับปรับปรุงความตรงต่อเวลาของเที่ยวบิน นอกจากงานวิจัยของ Wang และทีมแล้ว ยังมีงานวิจัยอื่นๆที่มีการใช้ค่าระดับความเป็นศูนย์กลางในการบ่งชี้ถึงสนามบินที่มีเที่ยวบินผ่านมากที่สุด [8][9][10]

จากการสำรวจงานวิจัยที่กล่าวมาแล้วข้างต้น งานที่ใช้บอร์ดราสเบอรี่พายสำหรับสร้างคลัสเตอร์วิเคราะห์ข้อมูลขนาดใหญ่ส่วนมากแล้ว ได้เน้นไปที่การประเมินสมรรถนะของคลัสเตอร์ แต่ไม่ได้อธิบายรายละเอียดในการใช้งานคลัสเตอร์ในการวิเคราะห์ข้อมูล ดังนั้น บทความนี้ นำเสนอประสบการณ์ในการพัฒนาคลัสเตอร์ราสเบอรี่พายและการใช้งานแพลตฟอร์มสำหรับวิเคราะห์ข้อมูลขนาดใหญ่ ซอฟต์แวร์ระบบเปิดเผยต้นฉบับ (Open-source software) อาปาเชฮาดูป (Apache Hadoop) และอาปาเชสปาร์ค (Apache Spark) ถูกใช้สำหรับประมวลผลข้อมูลบนคลัสเตอร์ ชุดข้อมูลเที่ยวบินล่าช้าและหลักการวิเคราะห์เครือข่ายสังคมถูกใช้เพื่อบ่งชี้สนามบินที่เกิดเที่ยวบินล่าช้าบ่อยที่สุด ซึ่งวิธีการพัฒนาแพลตฟอร์มการวิเคราะห์ข้อมูลขนาดใหญ่ด้วยราสเบอรี่พายสามารถแบ่งออกเป็น 2 ขั้นตอน คือ การสร้างต่อเชื่อมบอร์ดราสเบอรี่พายและติดตั้งซอฟต์แวร์ที่จำเป็น การปรับแต่งค่าสำหรับใช้งาน สำหรับการวิเคราะห์ข้อมูลด้วยเครือข่ายสังคม แบ่งออกเป็น 2 ขั้นตอนเช่นกัน คือ การขัดเกลาข้อมูลและการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมและการวิเคราะห์ข้อมูล ผลการทดลองพบว่า คลัสเตอร์ราสเบอรี่พายสามารถใช้ในการประมวลผลข้อมูลที่มีขนาดใหญ่ได้และมีประสิทธิภาพสูงกว่าการประมวลผลข้อมูลบนคอมพิวเตอร์ทั่วไป สำหรับผลการวิเคราะห์ข้อมูลความล่าช้าเที่ยวบินด้วยหลักการวิเคราะห์เครือข่ายสังคม พบว่า สนามบิน Chicago O'Hare International Airport (ORD) มีเที่ยวบินล่าช้าจำนวนมากที่สุด วัดจากค่าระดับความเป็นศูนย์กลาง (Degree centrality)

เนื้อหาส่วนที่เหลือในบทความนี้แบ่งออกเป็นส่วนๆ ดังต่อไปนี้ ส่วนที่ 2 เป็นเนื้อหาสรุปเกี่ยวกับทฤษฎีและหลักการที่ใช้ในบทความนี้ ส่วนที่ 3 เป็นรายละเอียดวิธีการที่นำเสนอ ส่วนที่ 4 เป็นผลการทดลอง และส่วนที่ 5 เป็นบทสรุป

ทฤษฎีและหลักการทางเทคนิค

1. อาปาเชฮาดูป (Apache Hadoop)

อาปาเชฮาดูป เป็นซอฟต์แวร์เฟรมเวิร์คชนิดเปิดเผยต้นฉบับ (Open-source software framework) สำหรับการประมวลผลแบบกระจาย (Distributed processing) โดยสามารถประมวลผลชุดข้อมูลขนาดใหญ่ และจัดการข้อมูลบนแหล่งจัดเก็บแบบกระจายผ่านอัลกอลิทึมการจับคู่และลดส่วน (MapReduce algorithm)

2. อาปาเชสปาร์คและพายสปาร์ค (Apache Spark และ PySpark)

อาปาเชสปาร์ค เป็นเฟรมเวิร์คไลบรารีแบบเปิดเผยต้นฉบับที่ถูกออกแบบมาสำหรับการประมวลผลข้อมูลขนาดใหญ่แบบเรียลไทม์ โดยมีเวอร์ชันสำหรับภาษาหลายภาษา เช่น ภาษาจาวา สกาลา ไพทอน และอาร์ อาปาเช สปาร์คสามารถถูกผนวกรวมเข้ากับอาปาเชฮาร์ดูปได้อย่างดี และสามารถใช้ในการประมวลผลข้อมูลที่อยู่บนระบบไฟล์ของฮาร์ดูป (HDFS)ได้ นอกจากนี้ อาปาเชสปาร์คยังมีไลบรารีอื่นๆที่สำคัญ เช่น เอสคิวแอล การเรียนรู้ของเครื่อง กราฟ และการประมวลผลแบบต่อเนื่อง (Stream processing) เป็นต้น ซึ่งไลบรารีที่นำมาใช้เพิ่มเติมสำหรับงานวิจัยนี้ คือ พายสปาร์ค โดยพายสปาร์ค เป็นเอพีไอภาษาไพทอน (Python API) สำหรับเขียนโปรแกรมประมวลผลข้อมูลแบบขนาน

3. การวิเคราะห์เครือข่ายสังคม (Social Network Analysis)

การวิเคราะห์เครือข่ายสังคมตั้งอยู่หลักการของทฤษฎีกราฟ โดยกราฟหนึ่งๆ ประกอบด้วยชุดของจุดยอด (Vertices) และชุดของเส้นเชื่อมระหว่างจุดยอด (Edges) เมื่อประยุกต์ใช้ทฤษฎีกราฟสำหรับการวิเคราะห์เครือข่ายสังคม จุดยอดแต่ละจุดแทนสิ่งที่อยู่ในความสนใจ เช่น บุคคล สถานที่ หรือสนามบิน เป็นต้น และเส้นเชื่อมระหว่างจุดยอดแทนถึงความสัมพันธ์ระหว่างสิ่งที่สนใจ เช่น สายการบินที่บินระหว่างสนามบินหนึ่งไปอีกสนามบินหนึ่ง โดยเส้นเชื่อมที่อยู่ในกราฟ สามารถแบ่งออกเป็น 2 ชนิดคือ เส้นเชื่อมที่มีทิศทาง (Directed edge) และเส้นเชื่อมที่ไม่มีทิศทาง (Undirected edge) สำหรับเส้นเชื่อมที่มีทิศทาง จะมีการกำหนดจุดต้นทางและจุดปลายทาง และเส้นเชื่อมที่ไม่มีทิศทาง จะไม่มีการกำหนดจุดต้นทางและปลายทาง ดังนั้น จากชนิดที่มีอยู่ของเส้นเชื่อมกราฟๆหนึ่งสามารถแบ่งออกเป็น 3 ชนิด คือ กราฟแบบมีการกำหนดทิศทาง (Directed graph), กราฟแบบไม่มีการกำหนดทิศทาง (Undirected graph), และกราฟแบบผสม (Hybrid graph) ที่มีเส้นเชื่อมทั้งสองชนิดอยู่ในกราฟเดียวกัน

วิธีการวิเคราะห์เครือข่ายสังคมที่รู้จักกันดี คือการคำนวณหาค่าความเป็นศูนย์กลางของจุดยอดที่อยู่ในกราฟ โดยค่าความเป็นศูนย์เป็นเครื่องวัดความสำคัญของจุดยอดนั้นๆ ซึ่งแบ่งออกเป็น 4 ชนิดใหญ่ คือ

3.1 ระดับความเป็นศูนย์กลาง (Degree centrality)

3.2 ระดับความใกล้ศูนย์กลาง (Closeness centrality)

3.3 ระดับค่าคั่นกลาง (Betweenness centrality)

3.4 ระดับค่าอิทธิพล (Eigenvector centrality)

วิธีการที่นำเสนอ

เนื้อหาที่นำเสนอในบทความนี้แบ่งออกเป็น 2 ส่วน คือ (i) วิธีการพัฒนาแพลตฟอร์มการวิเคราะห์ข้อมูลขนาดใหญ่โดยคลัสเตอร์บอร์ดราสเบอรี่พาย และ (ii) การประยุกต์ใช้แพลตฟอร์มที่สร้างสำหรับวิเคราะห์เครือข่ายสังคม

1. การพัฒนาแพลตฟอร์มการวิเคราะห์ข้อมูลขนาดใหญ่ด้วยคลัสเตอร์บอร์ดราสเบอรี่พาย

การสร้างแพลตฟอร์มการวิเคราะห์ข้อมูลขนาดใหญ่ด้วยบอร์ดราสเบอรี่พาย แบ่งออกเป็นขั้นตอนดังต่อไปนี้

1.1 การเชื่อมต่อฮาร์ดแวร์และติดตั้งซอฟต์แวร์

ในงานวิจัยชิ้นนี้ บอร์ดราสเบอรีพายรุ่น 4B ที่มีขนาดหน่วยความจำ 4 GB และพื้นที่จัดเก็บ 16 GB จำนวน 5 บอร์ด ถูกนำมาใช้ และเชื่อมต่อกันเป็นคลัสเตอร์แบบไร้สายโดยใช้เราเตอร์ไวไฟ โดยบอร์ดหนึ่งทำหน้าที่เป็นโหนดมาสเตอร์ (Master node) สำหรับบริหารจัดการงานต่างๆ และบอร์ดที่เหลือสี่บอร์ดทำหน้าที่เป็นโหนดคนงาน (Worker node) โหนดแต่ละโหนดจะถูกกำหนดค่า IP address ในไฟล์ /etc/hosts และติดตั้งระบบปฏิบัติการราสเบียน (Raspbian OS), อาปาเชฮาดูป (Apache Hadoop), อาปาเชสปาร์ค (Apache Spark), และซอฟต์แวร์พื้นฐานอื่นๆ เช่น PuTTy

1.2 การปรับแต่งค่าคลัสเตอร์

หลังจากสร้างคลัสเตอร์และติดตั้งระบบซอฟต์แวร์ที่จำเป็นแล้ว ขั้นตอนที่สำคัญมากอันหนึ่งคือ การปรับแต่งค่าสำหรับการประมวลผลบนคลัสเตอร์ โดยสำหรับอาปาเชฮาดูป การปรับแต่งค่าต่างๆจะทำในไฟล์นามสกุล XML ดังต่อไปนี้

- core-site.xml: สำหรับปรับแต่งค่าเพื่อแจ้ง Hadoop deamon เกี่ยวกับตำแหน่งที่ namenode รันอยู่บนคลัสเตอร์
- hdfs-site.xml: สำหรับแต่งค่าตำแหน่งของ namenode, datanode และการทำ data replication
- yarn-site.xml: สำหรับปรับแต่งค่าของ YARN เพื่อบริหารจัดการการใช้งาน CPU, memory, applications, และลำดับงานต่างๆ
- mapred-site.xml: สำหรับปรับแต่งค่าของ mapreduce daemons, job tracker, และ task tracker

2. การวิเคราะห์ข้อมูลการล่าช้าของสายการบินด้วยเครือข่ายสังคม

การวิเคราะห์ข้อมูลการล่าช้าของสายการบิน แบ่งออกเป็น 4 ขั้นตอน คือ (i) การทำความเข้าใจข้อมูล (ii) การกรองและขัดเกลาข้อมูล (iii) การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม และ (iv) การวิเคราะห์เครือข่ายสายการบินที่ล่าช้า

2.1 การทำความเข้าใจข้อมูล

ชุดข้อมูลที่ใช้คือ “Airline Delay and Cancellation Data, 2009-2018” เป็นข้อมูลเที่ยวบินที่เกิดความล่าช้าหรือถูกยกเลิกของเที่ยวบินภายในประเทศอเมริกาในช่วงเวลา 10 ปีจากปี ค.ศ.2009-2018 ข้อมูลนี้สามารถค้นหาและดาวน์โหลดได้จากเว็บ Kaggle ตารางที่ 1 แสดงข้อมูลจำนวนเที่ยวบินที่ล่าช้า แถวข้อมูล และขนาดข้อมูลของชุดข้อมูลในแต่ละไฟล์

ตารางที่ 1 รายละเอียดเชิงสถิติของชุดข้อมูล

File name	No. delayed flights	No. records	Size (MB)
2009.csv	1,170,501	6,429,338	774
2010.csv	1,174,884	6,450,117	775
2011.csv	1,110,531	6,066,650	729
2012.csv	1,015,158	6,096,762	757
2013.csv	1,269,277	6,369,482	769
2014.csv	1,170,501	6,429,338	702
2015.csv	1,063,439	5,819,079	702
2016.csv	964,239	5,617,658	678
2017.csv	1,029,473	5,674,621	685
2018.csv	1,352,710	7,213,446	872
All	11,390,740	61,556,964	7,443

2.2 การกรองและการขัดเกลาข้อมูล

จากข้อมูลในแต่ละไฟล์ ตารางข้อมูลเที่ยวบินพร้อมกับเวลาที่ล่าช้าถูกสร้างขึ้น ดังแสดงในตารางที่ 2 โดยคอลัมน์เวลาที่ล่าช้าเป็นผลรวมของคอลัมน์เวลาล่าช้าในไฟล์ข้อมูล 5 คอลัมน์คือ “CARRIER_DELAY”, “WEATHER_DELAY”, “NAS_DELAY”, “SECURITY_DELAY”, และ“LATE_AIRCRAFT_DELAY” จากนั้นทำการลบเที่ยวบินที่มีผลรวมของเวลาที่ล่าช้าเป็นศูนย์ออกไป

ตารางที่ 1: ข้อมูลเวลาล่าช้าของแต่ละเที่ยวบิน

Origin	Destination	Flight delay time (min)
EWR	ORD	27
IAH	BHM	27
IAH	CLT	15
IAH	CVG	32
COS	IAH	27
MAF	IAH	184
CMH	EWR	24
CLE	RIC	60
MSP	EWR	15
IAH	MEM	15

2.3 การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม

เนื่องจากวัตถุประสงค์ของการวิเคราะห์ในบทความนี้ คือต้องการบ่งชี้ถึงสนามบินที่มีเที่ยวบินที่มีความล่าช้าเกิดขึ้นบ่อยที่สุด ดังนั้นตารางเที่ยวบินและเวลาที่ล่าช้าในขั้นตอนก่อนหน้านี้ ถูกแปลงไปเป็นเที่ยวบินและจำนวนครั้งที่เกิดความล่าช้า โดยใช้ฟังก์ชัน GroupBy ใน PySpark ดังแสดงในตารางที่ 3

ตารางที่ 3: ข้อมูลในรูปแบบที่เหมาะสมสำหรับสร้างกราฟเครือข่ายสังคม

Origin	Destination	count
SNA	PHX	5739
ORD	PDX	5086
PHL	MCO	11571
EWR	STT	361
ATL	GSP	5425
SPI	ORD	2095
LAS	LIT	709
SMF	BUR	3709
ROC	CLE	377
MSP	AVL	2

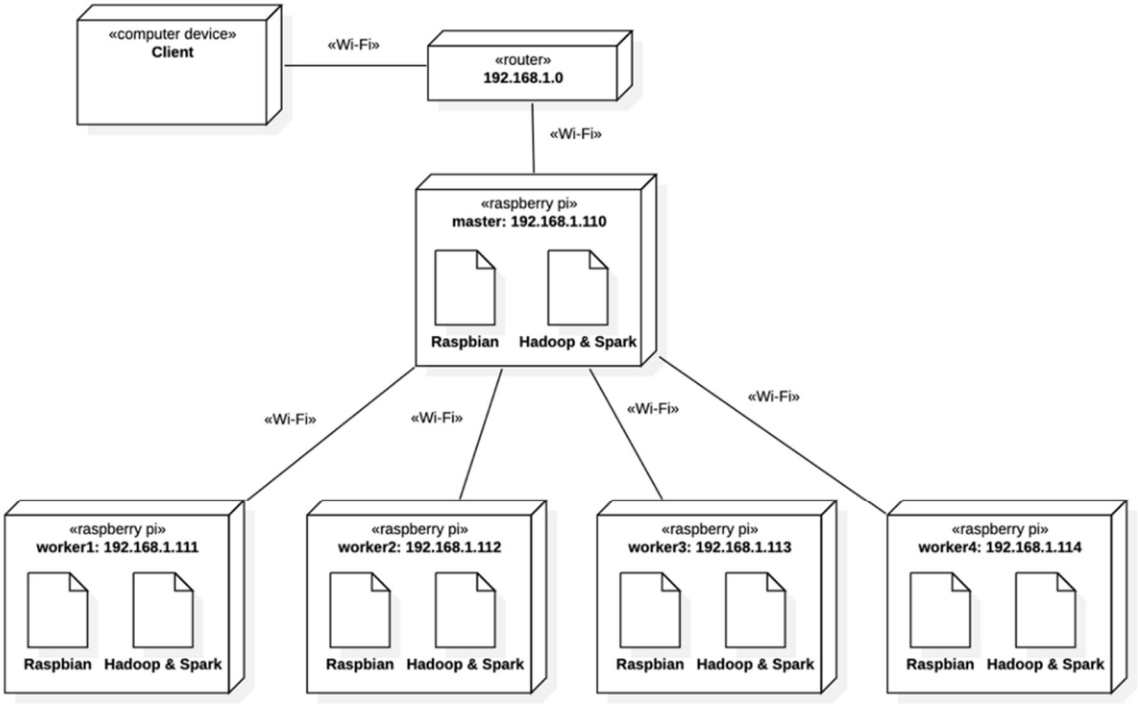
2.4 การวิเคราะห์เครือข่ายเที่ยวบินที่ล่าช้า

เมื่อได้ข้อมูลในรูปแบบที่เหมาะสมแล้ว กราฟเครือข่ายสังคมของเที่ยวบินและจำนวนครั้งที่ล่าช้าสามารถถูกสร้างขึ้นได้ โดยโหนดเริ่มต้นแทนถึงสนามบินต้นทาง โหนดสิ้นสุดแทนถึงสนามบินปลายทาง และจำนวนเที่ยวบินที่ล่าช้าแทนถึงน้ำหนักของการเชื่อมต่อระหว่างสนามบิน สุดท้ายค่าความเป็นศูนย์กลางของสนามบินถูกคำนวณเพื่อบ่งชี้ถึงสนามบินที่มีจำนวนเที่ยวบินล่าช้าบ่อยที่สุด

ผลการการวิจัยและอภิปรายผลการวิจัย

1. แพลตฟอร์มการวิเคราะห์ข้อมูลขนาดใหญ่ด้วยคลัสเตอร์บอร์ด

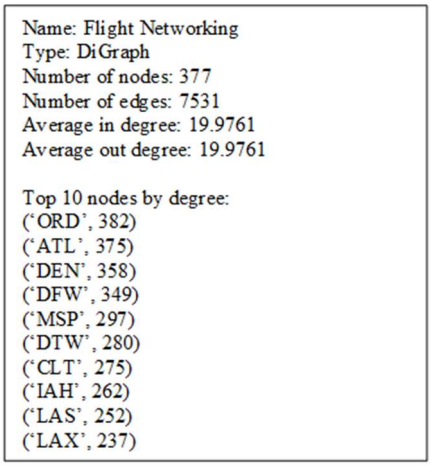
ภาพที่ 1 แสดงภาพรวมของแพลตฟอร์มการวิเคราะห์ข้อมูลขนาดใหญ่ประกอบไปด้วยเราเตอร์ไวไฟจำนวน 1 เครื่อง และบอร์ดราสเบอรี่พายจำนวน 5 บอร์ดเชื่อมต่อกันผ่านอินเตอร์เน็ตไวไฟ โดยแต่ละบอร์ดมีการกำหนดค่า IP Address เฉพาะของแต่ละบอร์ด และติดตั้งระบบปฏิบัติการราสเบียน อาปาเชฮาดูป และอาปาเชสปาร์ค ตามลำดับ



ภาพที่ 1 ภาพรวมแพลตฟอร์มการวิเคราะห์ข้อมูลขนาดใหญ่

2. ผลการวิเคราะห์เครือข่ายสังคม

ภาพที่ 2 แสดงผลลัพธ์ของการวิเคราะห์กราฟเครือข่ายเครือข่ายเที่ยวบินที่ล่าช้าแบบมีการกำหนดทิศทาง ซึ่งกราฟเครือข่ายประกอบด้วย 377 โหนด และ 7,531 จุดยอด (Edges) แต่ละโหนดแทนถึงสนามบินและจุดยอด (Edges) แทนถึงเส้นทางจากเที่ยวบินขาออกของสนามบินไปยังเที่ยวบินขาเข้าของสนามบิน สนามบินที่มีค่าระดับศูนย์กลางสูงสุด 10 อันดับแรกได้เรียงตามลำดับ โดยสนามบินที่มีความเป็นศูนย์กลางสูงสุดคือ สนามบิน ORD จึงอนุมานได้ว่าสนามบินแห่งนี้มีจำนวนเที่ยวบินล่าช้ามากที่สุด



ภาพที่ 2 รายละเอียดข้อมูลของผลการวิเคราะห์กราฟเครือข่ายเที่ยวบินที่ล่าช้าแบบมีการกำหนดทิศทาง

สรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อแสดงศักยภาพของการใช้คลัสเตอร์บอร์ดราสเบอรี่พายสำหรับการวิเคราะห์ข้อมูลขนาดใหญ่ โดยอธิบายขั้นตอนในการสร้างแพลตฟอร์มขึ้นมา และการประยุกต์ใช้แพลตฟอร์มในการวิเคราะห์ข้อมูล ซึ่งได้เลือกข้อมูลความล่าช้าของสายการบินมาทำการวิเคราะห์เครือข่ายสังคมเพื่อหาสนามบินที่มีจำนวนเที่ยวบินล่าช้ามากที่สุด

นอกจากนี้ยังช่วยให้ประมวลผลข้อมูลขนาดใหญ่ได้อย่างรวดเร็วบนอาปาเชฮาดูป (Apache Hadoop) และอาปาเชสปาร์ค (Apache Spark) ได้ง่ายขึ้น

ในอนาคต สามารถนำชุดคลัสเตอร์ของบทความนี้เป็นชุดเทมเพลต นำไปต่อยอดในด้านอื่นๆของแพลตฟอร์มการวิเคราะห์ข้อมูลขนาดใหญ่ได้ เช่น แพลตฟอร์มการวิเคราะห์ข้อมูลขนาดใหญ่ด้วยเทคนิคการเรียนรู้เชิงลึก (Deep learning) เพื่อทำนายโรคติดเชื้อไวรัสโคโรนา ทำให้แพลตฟอร์มนี้เป็นประโยชน์ต่อกลุ่มวิจัยทางด้านการวิเคราะห์ข้อมูลขนาดใหญ่ในสถาบันการศึกษาและกลุ่มวิจัยอิสระ

กิตติกรรมประกาศ

การจัดทำวิจัยได้รับการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

เอกสารอ้างอิง

[1] Diwedi, D.V. and Sharma, “Development of a low-cost cluster computer using raspberry pi,” IEEE global conference on wireless computing and networking (GCWCN), pp. 11–15, 2018

[2] Papakyriakou, “Benchmarking raspberry pi 2 hadoop cluster,” International Journal of Computer Applications, vol. 178, pp. 37–47, 2019

[3] Nugroho, S. and Widiyanto, “Designing parallel computing using raspberry pi clusters for IoT servers on apache Hadoop,” Journal of Physics: Conference Series, vol. 1517, pp. 012070, 2020

[4] Kaewkasi, C. and Srisuruk, “A study of big data processing constraints on a low-power Hadoop cluster,” International Computer Science and Engineering Conference (ICSEC), pp. 267–272, 2014

[5] Kaewkasi, C. and Srisuruk, “Optimizing performance and power consumption for an ARM-based big data cluster,” IEEE Region 10 Conference, pp. 1–6, 2014.

[6] Marković, D. “Image Processing on Raspberry Pi Cluster,” IJEEC - INTERNATIONAL JOURNAL OF ELECTRICAL ENGINEERING AND COMPUTING, vol. 2, pp. 83-90, 2019.

[7] Wang, Y. and Li, Y. “Complexity Analysis on the Influence Factors of the Flight Delay Risk Based on SNA,” Open Journal of Social Sciences, vol. 08, pp. 54–71, 2020

[8] Sapre, M. and Parekh, N. “Analysis of centrality measures of airport network of India,” Pattern Recognition and Machine Intelligence, pp. 376–381, 2011.

[9] Wang, J. “Exploring the network structure and nodal centrality of China’s air transport network: A complex network approach,” Journal of Transport Geography, vol. 19, pp. 712–721, 2011.

[10] Yang, Y. “A novel centrality of influential nodes identification in complex networks,” IEEE Access, 2020.