Comparative study between big data technology and relational database query system case study : Healthcare dataset

Rachatha Timasornwichakit[1], Wansa Paoin[2], Worapol Phongpech[1]

[1]Computer and Communication Technology, Faculty of engineering, Dhurakij Pundit University

[2]Faculty of Medicine, Thammasat University

Abstract

This article is aimed to study the proper method to store and analyze the data of illness that gathers the information of service provided by The Ministry of Public Health in a Big Data architecture and the process of using the set of healthcare data to compile with 2 compilation techniques. This is to promote the knowledge and understanding the different of big data technology (Hadoop MapReduce) and the relation database (MySQL) by using experimental research, studying the different between the groups of Mathematics data compilation with different storage architecture and retrieval methods. Testing the set of data of healthcare data with increasing large size and implement the set of questionnaire for the report of illness from 2014, sourcing from 2 reports from the cause of illness as the tool to find the performance of file compilation and query time. Performance evaluation with an analysis of the time by using t-Test tested the hypothesis that given the results of data query when there is a comparison between large data echnologies and relational databases vary significantly that the = 0.05 and assessing the accuracy and precision of results retrieval by percentage. The test results showed that the efficiency of Hadoop and MapReduce use more time to process. Results of the analysis were the P-Value = 0.17 rather than the 0.05 alpha result of the retrieval is not significantly different statistically. This not accept hypothesis predicted in advance. And the retrieval results are accurate synchronization of all data sets 100%.

Keywords: Big Data, Healthcare System

Received 12 September 2016; Accepted 25 November 2016

Correspondence: Rachatha Timasornwichakit, Computer and Communication Technology, Faculty of Engineer, Faculty of engineering, Dhurakij Pundit University, 110/1-4 Prachachuen Road Laksi, Bangkok 10210, Thailand (Tel.: +66-954-7300; E-mail address: timasorn@hotmail.com).

การเปรียบเทียบการค้นคืนข้อมูลบนเทคโนโลยีข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์ กรณีศึกษาตัวอย่างชุดข้อมูลบริการสุขภาพ

รชต ทิมาสรวิชกิจ[1], วรรษา เปาอินทร์[2], วรพล พงษ์เพ็ชร[1]

[1]สาขาเทคโนโลยีคอมพิวเตอร์และการสื่อสาร คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

[2]คณะแพทยศาสตร์ มหาวิทยาลัยธรรมศาสตร์

บทคัดย่อ

บทความนี้มีจุดมุ่งหมายเพื่อศึกษาแนวทางที่เหมาะสมในการจัดเก็บและวิเคราะห์ข้อมูลการให้บริการสุขภาพที่รวบรวมโดยกระทรวงสาธารณสุขบนสถาปัตยกรรมข้อมูลขนาดใหญ่ เพื่อเสริมสร้างความรู้และความเข้าใจในความแตกต่างของเทคโนโลยีระหว่างเทคโนโลยีข้อมูลขนาดใหญ่ฮาดูปและแมพรีดิวกับระบบการจัดการฐานข้อมูลเชิงสัมพันธ์มายเอสคิวแอลที่มีหลักการทางคณิตศาสตร์ที่แตกต่างกัน และมีสถาปัตยกรรมการจัดการข้อมูลที่ไม่เหมือนกัน และรูปแบบวิธีการเรียกใช้งานการค้นคืนที่ต่างกัน ด้วยขั้นตอนวิธีการใช้ชุดข้อมูลบริการสุขภาพนำมาประมวลผลร่วมกับเทคนิควิธีการประมวลผลแบบขนานแมพรีดิวและภาษาสอบถามเชิงโครงสร้างเอสคิวแอล โดยใช้วิธีวิจัยเชิงทดลองนำมาศึกษาความแตกต่าง ทดสอบด้วยชุดข้อมูลบริการสุขภาพที่มีขนาดระเบียนเพิ่มขึ้นเป็นลำดับ และสร้างชุดแบบสอบถามขึ้นจากรายงานสรุปการเจ็บป่วย พ.ศ.2557 จำนวน 2 รายงาน เป็นเครื่องมือที่นำมาใช้หาประสิทธิภาพของเวลาในการประมวลผลการค้นคืนข้อมูล ประเมินประสิทธิภาพด้วยการวิเคราะห์ผลลัพธ์ทางด้านเวลาโดยใช้สถิติ t-Test นำมาทดสอบสมมติฐานที่คาดการณ์ว่าผลลัพธ์ของเวลาการค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างเทคโนโลยีข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์แตกต่างกัน กำหนดค่านัยสำคัญที่ = 0.05 และการประเมินผลลัพธ์ความถูกต้องและแม่นยำการค้นคืนด้วยค่าสถิติร้อยละ ผลการทดสอบพบว่าประสิทธิภาพความเร็วฮาดูปและแมพรีดิวใช้เวลาในการประมวลผลมากกว่าและ ผลการวิเคราะห์สถิติวิจัยที่ได้ P-Value = 0.17 มากกว่าค่าแอลฟ่า 0.05 ผลของเวลาจึงไม่มีความแตกต่างกันทางสถิติ แต่มีผลลัพธ์ถูกต้องแม่นยำตรงกัน 100%

คำสำคัญ: ข้อมูลขนาดใหญ่, ระบบบริการสุขภาพ, การประมวลผลแบบขนาน, ฮาดูป, แมพรีดิว

วันที่รับต้นฉบับ 12 กันยายน 2559; วันที่ตอบรับ 25 พฤศจิกายน 2559

บทนำ

กระทรวงสาธารณสุขมีนโยบายให้ดำเนินการจัดเก็บรวบรวมข้อมูลการให้บริการสาธารณสุขและการแพทย์เข้าไว้ด้วยกันโดยมีกรอบแนวคิด “เครือข่ายบริการที่ไร้รอยต่อ” ที่สามารถเชื่อมโยงระดับบริการปฐมภูมิ ทุติยภูมิ และตติยภูมิเข้าด้วยกัน เพื่อให้สามารถจัดบริการสุขภาพที่มีคุณภาพ และเกิดการใช้ทรัพยากรที่มีอยู่จำกัดอย่างมีคุณภาพ และให้ดำเนินการจัดเก็บรวบรวมข้อมูลการให้บริการจากทุกหน่วยบริการสาธารณสุขเข้าไว้ด้วยกัน[1] ในระบบคลังข้อมูลด้านการแพทย์และสุขภาพ (Health Data Center : HDC) ใช้แฟ้มโครงสร้างมาตรฐานในการจัดเก็บ 43 และ 7 แฟ้มมาตรฐาน ข้อมูลถูกจัดเก็บไว้ในเซิร์ฟเวอร์แต่ละเดือนมีข้อมูลเพิ่มมากขึ้นขนาดใหญ่ขึ้น

ปัญหาสำคัญคือเวลาที่ใช้ในการประมวลผลเพื่อเปรียบเทียบสถิติหรืองานเวชสถิติหรืองานวิเคราะห์ทางการแพทย์ในแต่ละงานต้องใช้เวลาในการประมวลผลยาวนาน ระบบย่อมไม่สามารถสร้างสารสนเทศได้ทันเวลาตามความต้องการ หากมีสารสนเทศที่สามารถใช้ได้ทันตามต้องการ ถูกต้องทันสมัยและข้อมูลที่นำมากลั่นกรองให้ผู้บริหารต้องมีคุณภาพด้วย[2] สารสนเทศเป็นส่วนสำคัญที่จะทำให้ผู้บริหารระดับสูงตัดสินใจได้ถูกต้อง ในปัจจุบันเทคโนโลยีระบบการจัดเก็บแบบกระจายและการประมวลผลแบบขนานที่มีในระบบนิเวศข้อมูลขนาดใหญ่ (Big Data Ecosystem) จะสามารถนำมาประมวล

ผู้นิพนธ์ประสานงาน: รชต ทิมาสรวิชกิจ, สาขาเทคโนโลยีคอมพิวเตอร์และการสื่อสาร คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์ 110/1-4 ถนนประชาชื่น แขวงทุ่งสองห้อง เขตหลักสี่ กรุงเทพฯ 10210 โทร. 0-2954-7300; E-mail address: timasorn@hotmail.com)

ผลข้อมูลขนาดใหญ่จากคลังข้อมูลด้านการแพทย์และสุขภาพที่มีการจัดการฐานข้อมูลเชิงสัมพันธ์ได้หรือไม่

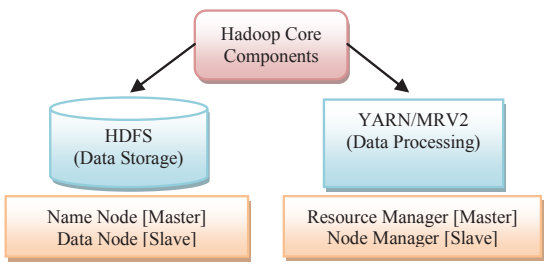
เพื่อทำความเข้าใจและรู้เชิงลึกในเทคโนโลยีข้อมูลขนาดใหญ่มากขึ้น งานวิจัยนี้ใช้วิธีวิจัยเชิงทดลอง กำหนดขั้นตอนการวิจัยเป็นแบบแผนการทดลองจริง มีการเก็บผลการทดลองสามครั้งเพื่อทำการหาค่าเฉลี่ย และมีการวิเคราะห์ผลข้อมูลด้วยสถิติวิจัย ทดสอบด้วยชุดข้อมูลบริการสุขภาพ ใช้วิธีคัดเลือกการสุ่มข้อมูลตัวอย่างด้วยเทคนิควิธีแบบเฉพาะเจาะจงโดยอาศัยการตัดสินใจของผู้เชี่ยวชาญระบบข้อมูลสุขภาพ[3] คัดเลือกแฟ้มผู้ป่วยนอก (OPD)[4] เป็นข้อมูลชุดตัวอย่างและกำหนดชุดแบบสอบถามข้อมูลที่ใช้เป็นโจทย์นำมาจากหนังสือสรุปรายงานการเจ็บป่วย พ.ศ.2557[5] รายงานผู้ป่วยนอก 2 รายงานถูกนำมาใช้ทดลองการค้นคืนข้อมูลด้วยภาษาสอบถามเชิงโครงสร้างเอสคิวแอลเพื่อเปรียบเทียบกับโปรแกรมประมวลผลการค้นคืนแบบขนานด้วยเทคนิคแมพรีดิวที่มีการจัดเก็บข้อมูลแบบกระจายด้วยฮาดูป กำหนดให้มีขนาดระเบียนเพิ่มขึ้นตามลำดับโดยใช้วิธีการเลือกสุ่มข้อมูลแบบโควตา เป็นการสุ่มอย่างง่ายด้วยระบบคอมพิวเตอร์ตามจำนวนที่กำหนดดังนี้ ห้าแสน,หนึ่งล้าน,ห้าล้าน และสิบล้านระเบียนตามลำดับ มีการวิเคราะห์ผลข้อมูลด้วยสถิติวิจัย โดยใช้สถิติเชิงพรรณนาประกอบด้วยค่าเฉลี่ยเลขคณิต, ร้อยละ, รูปตารางและการแสดงผลด้วยกราฟหรือแผนภูมิ[6] และใช้สถิติเชิงอนุมานพิสูจน์คำตอบจากสมมติฐานที่ได้คาดการณ์ไว้ล่วงหน้าด้วยสถิติวิจัย t-Test[7] กับผลการทดลองด้านความเร็วในการค้นคืน และทำการอภิปรายผล

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1. ทฤษฎีที่เกี่ยวข้อง

ระบบบริการสุขภาพจัดเก็บข้อมูลด้วยฐานข้อมูล คือการจัดเก็บข้อมูลอย่างมีระบบใช้คอมพิวเตอร์และซอฟแวร์ช่วยในการจัดการควบคุมการทำงานของฐานข้อมูล [8] เป็นข้อมูลที่มีโครงสร้างตาราง และได้รับการออกแบบลดขนาดความซ้ำซ้อน ข้อมูลมีการจัดการฐานข้อมูลเชิงสัมพันธ์คือข้อมูลในรูปของแถวและคอลัมน์ ที่ถูกจัดเก็บในตารางที่กำหนดให้ตารางหนึ่งสัมพันธ์กับตารางหนึ่งด้วยรีเลชั่นโดยมีโครงสร้างมาตรฐานและใช้หลักพื้นฐานทางคณิตศาสตร์เซตพีชคณิตเชิงสัมพันธ์และแคลคูลัสเชิงสัมพันธ์ที่นำมาใช้จัดการความสัมพันธ์กับข้อมูล [9] และยังมีคุณสมบัติทรานแซคชั่น ACID [10] ในการจัดการฐานข้อมูลเชิงสัมพันธ์นิยมใช้ภาษาสอบถามเชิงโครงสร้างเอสคิวแอล เป็นภาษาโปรแกรมระดับสูงที่สามารถใช้งานร่วมกับฐานข้อมูลด้วยชุดคำสั่งมาตรฐานเพิ่ม,แก้ไข,ปรับปรุง,ลบข้อมูล ในการจัดการฐานข้อมูลสมัยใหม่ยังมีระบบปรับปรุงประสิทธิภาพการสอบถามข้อมูล 3 ขั้นตอนดังนี้ 1) Parsing 2) Query Optimization 3) Query Evaluation [11]

กรอบการทำงานหรือแพลตฟอร์มฮาดูป (Hadoop) เป็นหนึ่งในเทคโนโลยีข้อมูลขนาดใหญ่ เป็นระบบการจัดเก็บข้อมูลแบบกระจาย ฮาดูปพัฒนามาจนถึงปัจจุบันในเวอร์ชั่น 2.6.2 มีขั้นตอนการทำงานโดยการแบ่งไฟล์ออกมาเป็นไฟล์ย่อยๆ หรือบล็อกข้อมูล (Data Block) และมี Name Node (Master) ทำหน้าที่ระบุตำแหน่งเก็บ และมี Data Node (Slave) กระจายไปเก็บในเครื่องอื่นๆ และมี YARN (Yet Another Resource Negotiator) ควบคุมจัดการทรัพยากรและใช้การประมวลผลแบบขนานแมพรีดิวปัจจุบันเป็นเวอร์ชั่น 2 (MRV2)[13] แมพรีดิวจึงเป็นการเขียนโปรแกรมควบคุมความต้องการข้อมูลที่ต้องการค้นคืนผ่านการจับคู่ Key/Value ที่กำหนดไว้[14] YARN Resource Manager ทำหน้าที่ควบคุมคลัสเตอร์ คอยบริหารตารางงานของ Job Tracker หรือ JobHistoryServer ที่ส่งไปยัง Node Manager (Slave) มี YARN Node Manager และ YARN Application Master ทำหน้าที่ควบคุมการทำงานของแมพรีดิวภายในคลัสเตอร์ การจัดการทรัพยากรและการจัดเก็บและประมวลผล YARN/MRV2 แบบใหม่นี้จะกระทำในแต่ละเครื่องคอมพิวเตอร์ เพื่อลดปริมาณการประมวลผลภายในเครือข่ายลง[15] ตาม**ภาพที่ 1**



ภาพที่ 1 กรอบการทำงานของฮาดูปทำงานร่วมกับแมพรีดิว

2. งานวิจัยที่เกี่ยวข้อง

จากการศึกษางานวิจัยที่เกี่ยวข้องกับเทคโนโลยีข้อมูลขนาดใหญ่การจัดจำแนกประเภทเทคโนโลยีของบิ๊กดาต้ายังไม่มีมาตรฐานหรือองค์กรใดจัดตั้งขึ้นมารองรับหรือจัดทำเป็นมาตรฐานสากล แต่จะใช้ลักษณะการเติบโตของข้อมูลนำมาจำแนกประเภทดังนี้ ข้อมูลแบบเชิงสัมพันธ์ข้อมูล (RDBMS) มีข้อมูลเพิ่มขึ้นจะขยายเป็นลักษณะแนวตั้ง (Vertical Scaling) แต่ลักษณะข้อมูลขนาดใหญ่ข้อมูลเพิ่มขึ้นจะขยายเป็นลักษณะแนวนอน (Horizontal Scaling)[16] ดังนั้นผู้วิจัยจึงขอจำแนกเทคโนโลยีข้อมูลขนาดใหญ่จากการศึกษางานวิจัยที่เกี่ยวข้องออกได้เป็น 2 กลุ่มดังนี้

(1). การจัดเก็บข้อมูล (Storage) แบ่งออกได้เป็น 4 ประเภทดังนี้ (a) แบบคอลัมน์ (Columns Oriented) (b) แบบคีย์คู่ (Key-Value Store)[17] (c) แบบเอกสาร (Document

Oriented)[18] (d) แบบกราฟ (Graph Database) ซึ่งการจัดเก็บทั้ง' 4 ประเภทนี้เป็นลักษณะ NoSQL คือการเขียนโปรแกรมมีอัลกอริทึมเพื่อควบคุมการทำงานการอ่าน, เขียน, ลบและแก้ไข แทนการจัดการข้อมูลรูปแบบดั้งเดิมที่มีการใช้ภาษาสอบถามเชิงโครงสร้าง SQL ในการจัดการ[19]

(2). การประมวลผล (Processing) เทคโนโลยีการประมวลผลข้อมูลในเทคโนโลยีบิ๊กดาต้ามีหลากหลายรูปแบบเช่นกัน โดยจำแนกออกเป็น 6 ประเภทดังนี้ (a) SQL ยังคงใช้รูปแบบภาษาสอบถามเชิงโครงสร้างในการจัดการข้อมูลแบบเชิงสัมพันธ์ใช้งานร่วมกับการจัดเก็บแบบ HDFS เช่น Hive, Impala หรือ Tajo (b) Key-Value มีลักษณะของการประมวลผลด้วยโปรแกรมมีการเขียนอัลกอริทึมภาษาต่างๆ คอยควบคุมสั่งการ เช่น Java ใช้งานร่วมกับแมพรีดิว[20] (c) NoSQL เป็นลักษณะของการประมวลผลด้วยการเขียนอัลกอริทึมด้วยโปรแกรมภาษาต่างๆ คอยควบคุมสั่งการ เช่น JSON ใช้งานร่วมกับ MongoDB หรือด้วยภาษา Scala, Python ใช้งานร่วมกับ Spark (d) NewSQL เป็นลักษณะของการประมวลผลด้วย SQL แต่นำ NoSQL มาเป็นฐานข้อมูลเพื่อการขยายข้อมูลในลักษณะแนวนอน และในการประมวลผลยังมีการนำหน่วยความจำสำรองมาใช้งานร่วมกับการประมวลผล (In Memory) เช่น VoltDB[21] (e) MPP (Massively Parallel Processing) เป็นการประมวลผลแบบคู่ขนานใช้หน่วยประมวลผล CPU ร่วมกันหลายตัว เช่น Exadata, Greenplumn (f) Graph Processing เป็นลักษณะการประมวลผลแบบกราฟบนเครือข่ายสังคมออนไลน์ด้วยทฤษฎีกราฟ เช่น Neo4j[22,23] ทั้งนี้การประมวลผลทั้ง 6 แบบมีวัตถุประสงค์หลักเพื่อการใช้งานประมวลผลกับชุดของข้อมูลขนาดเทราไบต์และเพตาไบต์[24]

ในยุคข้อมูลขนาดใหญ่กรอบการทำงานฮาดูปและแมพรีดิวเป็นโปรแกรมที่สามารถรองรับได้กับระบบปฏิบัติการหลายระบบ ทำให้สามารถนำมาศึกษาได้ง่าย[25] ผู้วิจัยจึงเลือกศึกษาลักษณะรูปแบบวิธีการใช้งาน และฮาดูปและแมพรีดิวยังได้รับความนิยมนำมาใช้ในวงการการศึกษาวิจัยทั้งภาคอุตสาหกรรมและภาคธุรกิจต่างๆ อย่างกว้างขวางเพื่อพัฒนาให้ระบบมีประสิทธิภาพเพิ่มขึ้น[26] ผู้วิจัยขอแบ่งงานวิจัยที่เกี่ยวข้องออกเป็น 2 กลุ่ม เพื่อง่ายและสะดวกต่อการพิจารณาดังนี้

1) กลุ่มนำเสนอการปรับปรุงประสิทธิภาพของเทคโนโลยีข้อมูลขนาดใหญ่ เช่น แนะนำวิธีการปรับปรุงกระบวนการภายในโปรแกรม จะทำให้เข้าใจการเขียนโปรแกรมมากขึ้นและการกำหนดค่าเริ่มต้นขนาดของไฟล์ข้อมูลเพื่อเพิ่มประสิทธิภาพในการจัดเก็บและเรียกใช้งานในกลุ่มคลัสเตอร์ขนาดเล็ก การนำฮาดูปและแมพรีดิวมาประเมินประสิทธิภาพในกระบวนการเพื่อปรับปรุงกระบวนการใน 2 ขั้นตอนแมพและรีดิว และการทดลองปรับปรุงฮาดูปและแมพรีดิวในการใช้ประมวลผลชุดข้อมูลรูปแบบต่างๆ และนำแมพรีดิวมาสร้างการเชื่อมโยงระหว่างฐานข้อมูลเอกสารด้วย MongoDB [27,28,29]

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

สรุปงานวิจัยที่เกี่ยวข้องกับที่ใกล้เคียงกับจุดประสงค์งานวิจัยนี้ดังนี้ (1) การทดสอบฮาดูปแมพรีดิวด้วยการใช้ข้อมูลขนาดเล็กในการอ่านและเขียนข้อมูลขนาด 512MB, 2GB, 4GB และใช้ขนาดบล็อกข้อมูลที่ 64MB และ 128MB พบว่าในกลุ่มข้อมูลขนาดเล็กจะมีประสิทธิภาพมากหากใช้บล็อกข้อมูล 64MB และมีประสิทธิภาพมากขึ้น 28.6% เมื่ออ่านข้อมูลขนาด 512MB และเมื่ออ่านข้อมูลขนาด 4GB ที่ 25.3% (2) การเขียนแมพรีดิวเชื่อมคอลเลคชั่นฐานข้อมูล MongoDB การเชื่อมหรือ Join ข้อมูลแบบเอกสารด้วยข้อมูลนักเรียนและที่ปรึกษาเพื่อใช้เชื่อมความสัมพันธ์ พบว่าการดำเนินการสามารถเชื่อมข้อมูลได้และมีประสิทธิภาพเพิ่มขึ้นเมื่อกำหนดให้ทำการกรองข้อมูลที่ต้องการเชื่อมไว้ล่วงหน้า ใช้เวลาเพียง 23 วินาที (3) งานวิจัยการปรับปรุงประสิทธิภาพการค้นคืนด้วยแมพรีดิวด้วยหลักการเชื่อมความสัมพันธ์และทำ Query Plan เพื่อให้เผยขั้นตอนการเชื่อมในขั้นตอนที่ดีที่สุด และทำการค้นคืนด้วยภาษาสอบถามเชิงโครงสร้างที่จำนวน 4, 6, 8 เครื่อง ด้วยข้อมูลอันดับเข้าดูเว็บไซต์ โดยทดลองกับ 2 กลุ่ม กลุ่มปรับปรุงประสิทธิภาพได้ผลความเร็วดีขึ้นเมื่อใช้ 8 เครื่อง แต่จะใช้เวลามากกว่ากลุ่มข้อมูลที่ไม่มีการปรับปรุงประสิทธิภาพ เมื่ออัลกอริทึมสั่งเพิ่มข้อมูลที่ระดับ 7-8 รอบ [30] (4) งานวิจัยการปรับปรุงอัลกอริทึมแมพรีดิวเพื่อเพิ่มประสิทธิภาพ เช่น การเชื่อม,การจัดเรียง,การจัดกลุ่มด้วยข้อมูลหลายรูปแบบ ผลรวมการประมวลผลใช้เวลา 9 พันวินาทีหรือ 2 ชั่วโมงครึ่ง ที่ข้อมูลขนาด 500GB ด้วยข้อมูลที่กำหนด

สมมุติขึ้น (Synthetic Data)[31] (5) งานวิจัยที่ใช้แมพรีดิว, ไฮฟ์ และมายเอสคิวแอลทำการทดสอบด้วยข้อมูลการชำระเงินของลูกค้าในธุรกิจขนาดเล็ก มีข้อมูลลูกค้าตั้งแต่ 500-20,000 บัญชี มีขนาดข้อมูลตั้งแต่ 235MB-9GB กับเครื่องจำนวน 1-4 เครื่อง ผลสรุปว่ามายเอสคิวแอลจะใช้เวลามากกว่าแมพรีดิวและไฮฟ์ที่ขนาดข้อมูล 1 หมื่นบัญชี หรือ 5GB ใช้เวลา 25 นาที แมพรีดิวจะใช้เวลาน้อยที่สุดในการประมวลผลทั้ง 1-4 เครื่อง ใช้เวลาโดยประมาณ 80-90 วินาที ในทุกชุดข้อมูลทดสอบ โปรแกรมแมพรีดิวมีประสิทธิภาพสม่ำเสมอและดีที่สุด[32]

สรุปการศึกษาในทุกงานวิจัยมีวัตถุประสงค์คล้ายคลึงกัน คือ เพื่อหาความเหมาะสมและรูปแบบการใช้งาน ที่สามารถนำมาใช้กับข้อมูลในรูปแบบต่างๆ และเพื่อค้นหาแนวทางการเพิ่มประสิทธิภาพให้กับการทำงานบนเทคโนโลยีข้อมูลขนาดใหญ่ ซึ่งได้ผลการเปรียบเทียบที่ให้ผลไปในทิศทางเดียวกันคือ เทคโนโลยีข้อมูลขนาดใหญ่จะมีประสิทธิภาพด้านความเร็วเมื่อข้อมูลมีขนาดใหญ่ขึ้น

แต่งานวิจัยที่เกี่ยวข้องนี้ ผู้วิจัยยังไม่พบงานใดทำการประเมินผลความแม่นยำถูกต้องของผลลัพธ์ข้อมูล งานวิจัยนี้จึงขอนำเสนอ การประเมินประสิทธิภาพด้านความเร็วร่วมกับการประเมินผลความถูกต้องของผลลัพธ์ ด้วยการประมวลผลชุดข้อมูลที่มีการขยายตัวของข้อมูลอย่างเป็นลำดับ เพื่อหาจุดตัดของกราฟด้านผลความเร็ว และผลลัพธ์ที่ถูกต้องตรงกันทุกชุดข้อมูลที่ใช้ในการทดลอง

วิธีดำเนินงานวิจัย

1. เตรียมข้อมูลชุดทดสอบ

ชุดข้อมูลทดสอบเป็นข้อมูลบริการสุขภาพที่เก็บรวบรวมข้อมูลโดยกระทรวงสาธารณสุขด้วยระบบคอมพิวเตอร์เครือข่ายระดับจังหวัด การคัดเลือกข้อมูลเป็นการสุ่มตัวอย่างด้วยเทคนิควิธีแบบเฉพาะเจาะจง เป็นการเลือกกลุ่มตัวอย่างโดยอาศัยการตัดสินใจจากผู้เชี่ยวชาญข้อมูลสุขภาพทำการสุ่มคัดเลือกนำตัวอย่าง กลุ่มข้อมูลชุดตัวอย่าง 1 แฟ้ม จากแฟ้มมาตรฐาน 43 แฟ้ม

ตารางที่ 1 ข้อมูลชุดทดสอบที่ถูกคัดเลือก

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

แฟ้ม Diagnosis_opd

2. สุ่มคัดเลือกข้อมูลชุดทดสอบตามกลุ่มที่กำหนด

สุ่มคัดเลือกข้อมูลชุดทดสอบตามกลุ่มที่กำหนดแบบโควต้า ใช้การสุ่มอย่างง่ายด้วยระบบคอมพิวเตอร์ โดยใช้ฟังก์ชั่น Rand() ในระบบฐานข้อมูลมายเอสคิวแอล แบ่งข้อมูลชุดทดสอบออกเป็น 4 ชุด และทำการนำออกเป็นรูปเท็กซ์ไฟล์ตามจำนวนชุดข้อมูลเพื่อเตรียมนำเข้าในระบบการจัดเก็บแบบกระจายฮาดูป

ตารางที่ 2 แบ่งกลุ่มข้อมูลชุดทดสอบ 4 ชุด

ชื่อแฟ้ม	จำนวนระเบียน	ขนาดไฟล์ (MB)
diagnosis_opd_5h	500,000	34.1
diagnosis_opd_1m	1,000,000	68.2
diagnosis_opd_5m	5,000,000	341.0
diagnosis_opd_10m	10,000,000	682.0

แฟ้ม Diagnosis_opd

3. เตรียมแบบสอบถามชุดทดสอบเอสคิวแอลและแมพรีดิว

กำหนดชุดแบบสอบถามจำนวน 2 ชุด จากรายงานสรุปการเจ็บป่วย พ.ศ.2557

1) รายงาน 10 ลำดับแรก จำนวนผู้ป่วยนอก ตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร)

2) รายงานจำนวนผู้ป่วยนอก รวมตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร)

4. ทดสอบการประมวลผลด้วยชุดแบบสอบถาม

นำรายงานในข้อ 3 สร้างเป็นชุดแบบสอบถามด้วยรูปแบบภาษาสอบถามเชิงโครงสร้างเอสคิวแอล ทำการประมวลผลกับชุดข้อมูลทดสอบในข้อที่ 2 จำนวน 3 ครั้งต่อ 1 รายงานต่อ 1 ชุดข้อมูลในระบบฐานข้อมูลเชิงสัมพันธ์ และสร้างชุดแบบสอบถามด้วยการเขียนโปรแกรมแมพรีดิวใช้ทดสอบกับไฟล์ข้อมูลชุดทดสอบที่จัดเก็บอยู่ในรูปแบบเท็กซ์ไฟล์นำเข้าในโปรแกรมฮาดูปจัดเก็บแบบกระจาย HDFS ทำการประมวลผลจำนวน 3 ครั้งต่อ 1 ชุดข้อมูล

5. บันทึกผลจากการประมวลผล

เก็บบันทึกผลการทดลองมี 2 กลุ่ม คือ กลุ่ม (1) การประมวลผลด้วยภาษาสอบถามเชิงโครงสร้างเอสคิวแอล ทำการบันทึกผลจากการทดลองประมวลผลตามกำหนดข้อ 4 ทุกครั้งจะทำการคืนค่าหน่วยความจำสำรองใหม่ กลุ่ม (2) การประมวลผลด้วยเทคนิคแมพรีดิว บันทึกผลการทดลองประมวลผลตามกำหนดข้อ 4 และกำหนดให้มีสูตรการบันทึกผลการทดลองดังนี้

สูตรการบันทึกผลการค้นคืนด้วยภาษาเอสคิวแอล (1)

Query Time = End Time - Start Time (1)

สูตรการบันทึกผลการค้นคืนด้วยเทคนิคแมพรีดิว (2)

MR Jop1 = (Map Job1 + Reduce Job1)

MR Job2 = (Map Job2 + Reduce Job2)

Total Time = MR Job1 + MR Job2 (2)

6. วิเคราะห์ผลด้วยสถิติวิจัย

นำค่าเฉลี่ยผลความเร็วมาวิเคราะห์ด้วยสถิติ t-Test (dependent) พิสูจน์สมมติฐานเวลาการค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างเทคโนโลยีข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์มีผลลัพธ์แตกต่างกัน กำหนดค่านัยสำคัญที่ 0.05 และใช้สถิติร้อยละประเมินผลความถูกต้องและแม่นยำจากเทคโนโลยีข้อมูล 2 รูปแบบ

7. กำหนดสภาพแวดล้อมการทดสอบ

กำหนดสภาพแวดล้อมการทดสอบด้วยคอมพิวเตอร์เสมือนกำหนดการทดลองฐานข้อมูลเชิงสัมพันธ์เป็นแบบไคลเอนต์เซิร์ฟเวอร์กำหนดเซิร์ฟเวอร์ 1 เครื่อง และไคลเอนต์เป็นแล็ปท็อป 1 เครื่อง สั่งการประมวลผลผ่าน Web Browser และทดลองเทคโนโลยีข้อมูลขนาดใหญ่ในสภาพแวดล้อมการจัดเก็บแบบกระจาย HDFS กำหนดให้ค่าเริ่มต้นของระบบเป็นมาตรฐานดั้งเดิมและให้มี Master จำนวน 1 เครื่อง และมี Slave จำนวน 2 เครื่อง ใช้แล็ปท็อปสั่งการประมวลผลโปรแกรมแบบขนานแมพรีดิวผ่านโปรแกรม SSH สถานที่ทำการทดลองใช้ห้องปฏิบัติการข้อมูลขนาดใหญ่ มธบ. คณะวิศวกรรมศาสตร์ ฮาร์ดแวร์มีคุณสมบัติดังนี้ (ก) CPU 2.4 GHz (ข) RAM 8 GB (ค) HDD 2 TB

ผลการทดลองและสรุปผลงานวิจัย

ผลการทดลองกลุ่ม (1) การนำเข้าในระบบฐานข้อมูลเชิงสัมพันธ์ มีค่าที่นำเสนอผล 2 ส่วน 1.ขนาดหน่วยความจุข้อมูล 2.หน่วยเวลาที่ใช้ในกระบวนการนำเข้า นำเสนอเป็นรูปแบบ **ตารางที่ 3**

ตารางที่ 3 ผลการนำข้อมูลเข้าฐานข้อมูลมายเอสคิวแอล

ชื่อแฟ้ม	จำนวนระเบียน	ขนาดไฟล์ (MB)	เวลาที่ใช้ (Second)
diagnosis_opd_5h	500,000	55.6	2.83
diagnosis_opd_1m	1,000,000	110.6	5.93
diagnosis_opd_5m	5,000,000	552.0	29.70
diagnosis_opd_10m	10,000,000	1,102.0	58.56

แฟ้ม Diagnosis_opd ในฐานข้อมูล

จากการทดลองนำเข้าข้อมูล สังเกตว่าใน**ตารางที่ 3** มีหน่วยความจุที่เพิ่มขึ้นจากไฟล์ที่จัดเตรียมไว้ก่อนนำเข้าใน**ตารางที่ 2** ซึ่งส่วนนี้เป็นค่าโอเวอร์เฮดในระบบฐานข้อมูลเชิงสัมพันธ์ เป็นส่วนที่นำมาใช้ในการจัดการฐานข้อมูลควบคุมการทำงานของข้อมูล

ผลการทดลองกลุ่ม (2) การนำเข้าในแพลตฟอร์มฮาดูป มีการนำเสนอผล 2 ส่วน เหมือนกลุ่ม (1) นำเสนอเป็นรูปแบบ**ตางรางที่ 4**

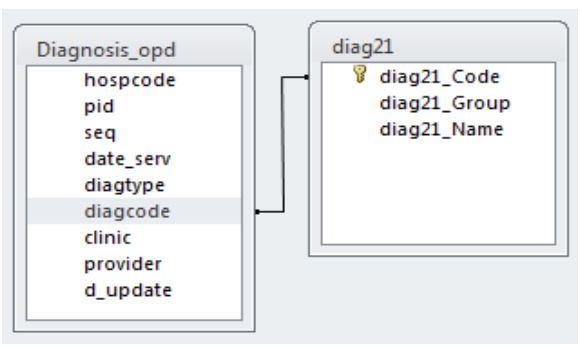
ตารางที่ 4 ผลการนำข้อมูลเข้าจัดเก็บแบบกระจาย HDFS

ชื่อแฟ้ม	จำนวนระเบียน	ขนาดไฟล์ (MB)	เวลาที่ใช้ (Second)
diagnosis_opd_5h	500,000	34.12	0.17
diagnosis_opd_1m	1,000,000	68.23	0.34
diagnosis_opd_5m	5,000,000	341.16	1.07
diagnosis_opd_10m	10,000,000	682.34	2.12

แฟ้ม Diagnosis_opd ในฐานข้อมูล

การทดลองนำเข้าข้อมูลทั้ง 2 กลุ่ม จะสังเกตได้ว่าขนาดของไฟล์และเวลาในการนำเข้าใน**ตารางที่ 4** น้อยกว่าใน**ตารางที่ 3** แต่หากนำไปเปรียบกับ**ตารางที่ 2** จะเพิ่มขึ้นเล็กน้อย ซึ่งระบบฐานข้อมูลมีระบบการจัดการเชิงสัมพันธ์จึงใช้เวลามากกว่า

การเตรียมชุดแบบสอบถามที่ผู้วิจัยคัดเลือกนั้นพบว่าต้องการรายการ 21 กลุ่มโรคหลัก นำมาใช้ในการประมวลผลร่วมกับข้อมูลผู้ป่วยนอก มีโครงสร้างไฟล์ 21 กลุ่มโรคดังนี้ diag21 (diag21_Code Char(3), diag21_Name Char(150), diag21_Group Int(2) index, primary key (diag21_Code)) กำหนดให้มีคีย์หลักเพื่อควบคุมค่าซ้ำกันและทำการเชื่อมความสัมพันธ์เพื่อจัดทำรายงานดัง**ภาพที่ 2**



ภาพที่ 2 เชื่อมความสัมพันธ์ตารางผู้ป่วยนอกกับตาราง 21 กลุ่มโรค

ผู้วิจัยได้จัดเตรียมเครื่องมือการประมวลผลออกเป็นขั้นตอนดังนี้

กลุ่ม 1 ขั้นตอนการประมวลผลด้วยภาษาสอบถามข้อมูลเอสคิวแอล

Query Report1 = 1} Count 2} Group 3} Join 4} Sort 5} Limit

Query Report2 = 1} Count 2} Group 3} Join 4} Sort

กลุ่ม 2 ขั้นตอนการประมวลผลด้วยเทคนิคแมพรีดิว

MR Job1 = 1} Join 2} Map 3} Reduce 4} Output

MR Job2 = 1} Map 2} Combine 3} Reduce 4} Output

ในกลุ่ม 2 ขั้นตอน Combine และ Reduce จะรวมการจัดเรียงจำนวนสูงสุดไว้เป็นลำดับต้น และนำผลทดสอบออกเป็นเท็กซ์ไฟล์และนำเข้าโปรแกรมเอ็กเซล สรุปผลจัดเรียงใหม่ตามรายงานข้อ 3 พร้อมแสดงผลภาพ ซึ่งในขั้นตอนนี้จะไม่นำมารวมบันทึกในเวลาการประมวลผลด้วย

Excel Job = 1} Import Text File 2} Sort 3} Cut

ตารางที่ 5 ผลการค้นคืนข้อมูลระบบฐานข้อมูลเชิงสัมพันธ์ ด้วยภาษาสอบถามข้อมูลเชิงโครงสร้างเอสคิวแอล รายงาน 10 ลำดับแรก จำนวนผู้ป่วยนอก ตามกลุ่มสาเหตุการป่วย และรายงานจำนวนผู้ป่วยนอก รวมตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร)

ชื่อแฟ้ม	จำนวนระเบียน	ขนาดไฟล์ (MB)	1) รายงาน 10 ลำดับแรก จำนวนผู้ป่วยนอก ตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร)				2) รายงานจำนวนผู้ป่วยนอก รวมตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร)				ผลรวมค่าเฉลี่ย
			เวลาที่ใช้ (Second)								
			ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	ค่าเฉลี่ย	ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	ค่าเฉลี่ย	
diagnosis_opd_5h	500,000	55.6	0.8792	0.8723	0.8780	0.8765	0.8697	0.8686	0.8699	0.8694	1.7459
diagnosis_opd_1m	1,000,000	110.6	1.7388	1.7296	1.7451	1.7588	1.7478	1.7465	1.7451	1.7465	3.5053
diagnosis_opd_5m	5,000,000	552.0	8.8761	8.8679	8.8719	8.8720	8.8544	8.8584	8.8823	8.8650	17.7370
diagnosis_opd_10m	10,000,000	1,102.0	17.8370	17.8001	17.8156	17.8176	17.7472	17.7331	17.7646	17.7483	35.5659

ตารางที่ 6 ผลการค้นคืนข้อมูลเทคโนโลยีข้อมูลขนาดใหญ่ ด้วยเทคนิคแมพรีดิว รายงาน 10 ลำดับแรก จำนวนผู้ป่วยนอก ตามกลุ่มสาเหตุการป่วย และรายงานจำนวนผู้ป่วยนอก รวมตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร)

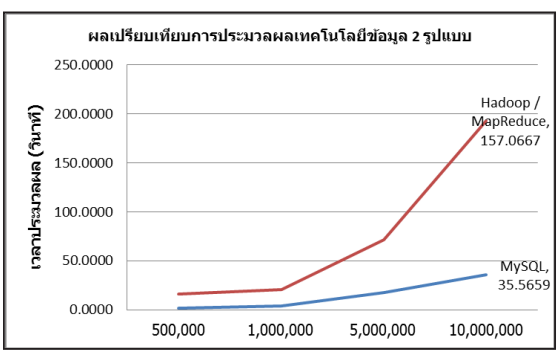
จำนวนระเบียน	ขนาดไฟล์ (MB)	เวลาที่ใช้ (Second)												ค่าเฉลี่ย (Total/3)
		Map Job1			Reduce Job1			Map Job2			Reduce Job2			
		1	2	3	1	2	3	1	2	3	1	2	3	
500,000	34.12	6.245	6.245	6.956	4.144	4.144	2.800	1.920	1.920	1.937	1.933	1.933	1.875	14.0173
1,000,000	68.23	9.234	9.446	9.075	4.097	3.934	4.449	1.901	1.920	1.913	1.950	1.899	1.916	17.2447
5,000,000	341.16	35.298	37.618	38.236	11.884	9.795	14.253	1.983	2.913	1.929	1.926	1.903	1.938	53.2253
10,000,000	682.34	128.192	134.972	108.763	28.792	27.440	29.924	1.875	1.910	1.899	3.560	1.948	1.925	157.0667

ผลจากการทดลองสังเกตได้ว่าใน**ตารางที่ 5** มีค่าเฉลี่ยรายงาน 1) มากกว่ารายงาน 2) เล็กน้อย เนื่องจากมีขั้นตอนการกำหนดลำดับการแสดงผลที่เพิ่มขึ้น และใน**ตารางที่ 6** ผล MR Job1 นั้นจะใช้เวลามากที่สุดเป็นผลจาก 2 ขั้นตอน คือ หลังจากนำข้อมูลเข้าจัดเก็บใน HDFS การเรียกรายงานจะทำการแสดงผลด้วยโปรแกรมสั่งการประมวลผลและรวบรวมผลข้อมูลที่ถูกจัดเก็บในเครื่อง Slave แล้วนำผลลัพธ์ส่งคืนกลับมาให้เครื่อง Master และจะทำทุกครั้งที่มีสั่งการ

ประมวลผล และการ Join ด้วยการนำระเบียนทั้งหมดในแฟ้มผู้ป่วยนอกมาเชื่อมความสัมพันธ์กับแฟ้ม 21 กลุ่มโรค ด้วย รูปแบบ Nested Loop Join หรือการเชื่อมสัมพันธ์โดยไม่มีการจัดเรียงอินเด็กซ์ก่อน ระเบียนมากขึ้นยิ่งใช้เวลามากขึ้นและนำผลรวมที่ได้จัดเก็บในไฟล์ผลลัพธ์เพื่อใช้ในขั้นตอนต่อไปใน MR Job2 แต่สังเกตว่างาน Mj2 และ Rj2 จะใช้เวลาใกล้เคียงกันเนื่องจากขั้นตอนนี้จะใช้ผลจาก MR Job1 นำมาประมวลผลใหม่ด้วยการจับคู่ข้อมูล, รวม, เรียงลำดับใหม่และนำเสนอเป็นผลลัพธ์ใหม่ไว้ในรูปแบบเท็กซ์ไฟล์

สรุปได้ว่าการประมวลผลด้วยเทคนิคแมพรีดิวมีขั้นตอนที่มากกว่าจะส่งผลต่อประสิทธิภาพด้านเวลาในการประมวลผลในขั้นตอน MR Job1 การเชื่อมสัมพันธ์และประมวลผลเพื่อนับจำนวนในแต่ละกลุ่มโรคก่อน จากนั้นสร้างไฟล์ผลลัพธ์จัดเก็บไว้เพื่อส่งต่อให้กับขั้นตอน MR Job2 ในการเรียกประมวลผลรวมแต่ละกลุ่มโรคเพื่อใช้แสดงผลรายงานและจัดเก็บในรูปแบบของไฟล์ผลลัพธ์อีกครั้ง อีกทั้งการปรับปรุงกำหนดค่าเริ่มต้นของระบบฮาดูปให้รองรับการประมวลข้อมูลขนาดเล็กได้ อาจจะส่งผลให้โปรแกรมใช้เวลาในการเรียกคืนไฟล์ข้อมูลขนาดใหญ่จากเครื่องที่ใช้เก็บข้อมูลในเครือข่ายในการเรียกคืนไฟล์ขนาดใหญ่จำนวนมากเพื่อนำมาประมวลผล ทำให้เวลาการประมวลผลเพิ่มขึ้นเมื่อระเบียนมีมากขึ้น

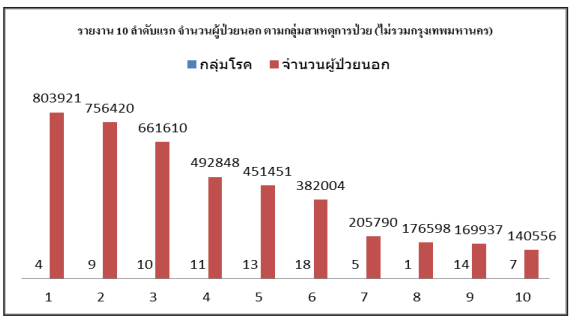
ผลรวมค่าเฉลี่ยทุกขั้นตอนในตารางที่ 6 เมื่อเปรียบเทียบกับตารางที่ 5 ที่รวมค่าเฉลี่ยจาก รายงาน 1) และ 2) มีค่าเฉลี่ยของเวลาประมวลผลมากกว่าทุกชุดข้อมูลทดสอบ ดัง**ภาพที่ 3**



ภาพที่ 3 กราฟผลเปรียบเทียบความเร็วการประมวลผลเทคโนโลยีข้อมูล 2 รูปแบบ

การทดลองแสดงผลภาพการวิเคราะห์ข้อมูลบริการสุขภาพ ด้วยเทคโนโลยีฐานข้อมูลสามารถดำเนินการได้ทันทีด้วยโปรแกรมการจัดการฐานข้อมูลเชิงสัมพันธ์ phpMyAdmin

หากใช้เทคโนโลยีข้อมูลขนาดใหญ่ฮาดูปและแมพรีดิวต้องนำผลที่ได้ในรูปเท็กซ์ไฟล์ออกมาจัดทำในโปรแกรมเอ็กเซล ดัง**ภาพที่ 4**



ภาพที่ 4 แผนภูมิตัวอย่างที่ได้จากรายงาน 1) ชุดข้อมูล 10 ล้าน ด้วย Excel

ผลทดลองการค้นคืนมีความถูกต้องและแม่นยำของเทคโนโลยีข้อมูล 2 รูปแบบ ตรงกัน 100% ในทุกชุดข้อมูลและทุกรายงาน ตามตัวอย่างตารางที่ 7 เปรียบเทียบผลลัพธ์การประมวลผลด้วยรายงาน 1) ซึ่งในรายงาน 2) ไม่ได้นำมาแสดงในบทความนี้ด้วย และรายงานที่ประมวลผลด้วยเทคโนโลยีข้อมูลขนาดใหญ่ยังสามารถประมวลผลผลลัพธ์ออกมาได้ 2 รายงานในคราวเดียว

จากสมมติฐานที่คาดการณ์ไว้ ล่วงหน้าว่าผลลัพธ์ของเวลาในการค้นคืนข้อมูล เมื่อมีการเปรียบเทียบระหว่างเทคโนโลยีข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์แตกต่างกัน มีผลคำนวณสถิติ t-Test ด้วยโปรแกรมเอ็กเซลดังนี้

ตารางที่ 7 ผลการเปรียบเทียบความแม่นยำถูกต้องการค้นคืนข้อมูลจากเทคโนโลยี 2 แบบ ในรายงานที่ 1) รายงาน 10 ลำดับแรก จำนวนผู้ป่วยนอก ตามกลุ่มสาเหตุการป่วย (ไม่รวมกรุงเทพมหานคร)

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

ตารางที่ 8 ตารางผลการคำนวณ T-TEST

	Hapdoop/Mapreduce	*MySQL*
Mean	60.3885	14.63851667
Variance	4469.884773	245.9075516
Observations	4	4
Pearson Correlation	0.9789991	
Hypothesized Mean Differenc	0	
df	3	
t Stat	1.773110822	
P(T<=t) two-tail	0.174320337	
t Critical two-tail	3.182446305	

สรุปผลการวิเคราะห์ค่าเฉลี่ยด้วยสถิติ t-Test สามารถแปลผลได้ดังนี้ เมื่อ P-Value (two tail) ที่ได้เท่ากับ 0.17 มากกว่าค่าแอลฟ่า 0.05 หรือค่าสถิติ t=1.77 มีค่าอยู่ระหว่างจุดวิกฤต (t Critical) -3.18 ถึง 3.18 จึงเป็นการปฏิเสธสมมติฐานที่ได้คาดการณ์ไว้ล่วงหน้าที่ว่าผลลัพธ์ของเวลาประมวลผลการค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างระบบข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์มีผลลัพธ์ที่แตกต่างกัน

อภิปรายผลงานวิจัย

ผลการประเมินประสิทธิภาพวิเคราะห์ด้วยค่าเฉลี่ยเวลาการค้นคืนข้อมูลผล P-Value (two tail) ที่ได้เท่ากับ 0.17 มากกว่าค่าแอลฟ่า 0.05 เป็นการปฏิเสธสมมติฐานที่คาดการณ์ไว้ว่าผลลัพธ์ของเวลาการค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างเทคโนโลยีข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์มีความแตกต่างกัน ซึ่งจากผลการทดสอบด้วยสถิติวิจัยทำให้ทราบว่าผลลัพธ์ด้านความเร็วของเทคโนโลยี 2 รูปแบบไม่มีความแตกต่างกัน

ผลการประเมินประสิทธิผลการค้นคืนมีความถูกต้องและแม่นยำของเทคโนโลยีข้อมูล 2 รูปแบบ พบว่ามีผลลัพธ์ตรงกันในทุกชุดข้อมูลและทุกรายงาน 100% จึงเป็นการยอมรับสมมติฐานที่คาดการณ์ว่า ผลลัพธ์ของความแม่นยำถูกต้อง การค้นคืนข้อมูลเมื่อมีการเปรียบเทียบระหว่างเทคโนโลยีข้อมูลขนาดใหญ่กับระบบฐานข้อมูลเชิงสัมพันธ์มีผลลัพธ์ไม่แตกต่างกัน

เทคโนโลยีข้อมูลขนาดใหญ่ฮาดูปและแมพรีดิวสามารถนำมาใช้งานร่วมกับข้อมูลในระบบฐานข้อมูลเชิงสัมพันธ์แบบมีโครงสร้างได้ สามารถจัดเก็บและนำมาวิเคราะห์ข้อมูลระบบบริการสุขภาพได้ อ้างอิงจากผลการทดสอบในงานวิจัยนี้ที่พบว่าการจัดเก็บแบบกระจายฮาดูปและการประมวลผลแบบขนานแมพรีดิว สามารถนำมาประมวลผลและจัดทำสถิติด้วยข้อมูลบริการสุขภาพได้ และมีความถูกต้องแม่นยำของผลลัพธ์ที่ได้ 100%

การใช้งานโปรแกรมฮาดูปและแมพรีดิวจะมีประสิทธิภาพสูงขึ้นอยู่กับขนาดของข้อมูลในแต่ละงานที่ต้องการประมวลผลและชุดข้อมูลที่เชื่อมสัมพันธ์มีการกรองข้อมูลที่ต้องการไว้ล่วงหน้าจะส่งผลดีต่อความเร็ว แต่ยังไม่เหมาะกับการประมวลผลชุดข้อมูลที่มีขนาดเล็กและคลัสเตอร์ที่มีขนาดเล็ก แต่สามารถใช้ประมวลผลกับชุดข้อมูลที่มีโครงสร้างและสัมพันธ์กันได้ หากจะนำมาใช้ต้องเข้าใจกระบวนการทำงานในโปรแกรมเพื่อการปรับแก้ไขค่าเริ่มต้น และการปรับปรุงวิธีการเขียนโปรแกรมเพื่อประยุกต์ใช้กับงานในชุดข้อมูลเฉพาะที่ต้องการและจำเป็นต้องกำหนดรูปแบบผลลัพธ์ไว้ล่วงหน้าจึงจะเกิดประสิทธิภาพและประสิทธิผลอย่างสูงสุด

ประสิทธิภาพด้านความเร็วในเทคโนโลยีฮาดูปและแมพรีดิวเมื่อเปรียบเทียบกับระบบฐานข้อมูลเชิงสัมพันธ์ในงานวิจัยนี้ พบว่าการใช้ MySQL ประมวลผลข้อมูลขนาดชุดสิบล้านมีความเร็วมากกว่า MapReduce ถึง 4 เท่า แต่หากเปรียบเทียบในการนำเข้าข้อมูลชุดข้อมูลสิบล้าน Hadoop จะเร็วกว่า 27 เท่า ซึ่งหากงานวิจัยนี้เป็นงานที่เปรียบเทียบประสิทธิภาพโดยรวมของการทำงานทั้งหมดฮาดูปและแมพรีดิวจะมีประสิทธิภาพที่ดีกว่ามายเอสคิวแอล และยังไม่เหมาะสมกับการใช้งานการค้นคืนข้อมูลขนาดเล็กและมีโครงสร้าง

แต่เทคโนโลยี NoSQL หรือ Key-Value นี้เป็นการเติมเต็มให้กับฐานข้อมูล RDBMS รูปแบบดั้งเดิม ที่ไม่จำเป็นต้องจัดเก็บแบบโครงสร้าง แต่ยังไม่สามารถนำมาทดแทนได้ในทุกๆ รูปแบบชุดข้อมูล หากนำมาใช้อย่างเหมาะสมตามลักษณะของข้อมูลที่ต้องการจัดเก็บและเรียกใช้งานจะทำให้เกิดประโยชน์สูงสุด

ข้อเสนอแนะ

ข้อจำกัดจากงานวิจัยนี้ผู้วิจัยพบว่าคุณภาพข้อมูลเป็นสิ่งสำคัญ มีข้อควรทราบที่ว่า “สารสนเทศที่ได้จะถูกต้องแม่นยำต้องใช้ข้อมูลคุณภาพ” ด้วยการควบคุมคุณภาพข้อมูลนำเข้าและการจับคู่เชื่อมสัมพันธ์ที่ดีจะทำให้ได้ผลที่ถูกต้อง ในการนำเข้าข้อมูลจำเป็นต้องทำความสะอาดข้อมูลก่อนใช้งาน และข้อมูลที่ใช้ในการเชื่อมความสัมพันธ์ต้องเข้าใจประเภทของข้อมูลที่ใช้ในการจับคู่ความสัมพันธ์ และต้องทำความเข้าใจผลลัพธ์ที่ต้องการว่าต้องการแสดงผลรูปแบบใด เพื่อควบคุมคุณภาพการเขียนโปรแกรมให้ได้ผลลัพธ์ตามที่ต้องการ จึงจะเปลี่ยนข้อมูลให้ออกมาเป็นสารสนเทศที่ถูกต้องในรูปแบบที่ต้องการ สามารถนำไปใช้งานต่อได้อย่างมีคุณภาพ

สำหรับผลด้านประสิทธิภาพความเร็วยังไม่เป็นตามที่คาดหวัง ซึ่งผลที่ได้ขัดแย้งกับงานวิจัยที่เกี่ยวข้องที่พบว่าการประมวลผลด้วยแมพรีดิวจะสามารถประมวลผลข้อมูลขนาดเล็กได้ดี[32] แต่จะสอดคล้องกับทฤษฎีบิ๊กดาต้าเหมาะกับการใช้งานประมวลผลชุดข้อมูลขนาดเทราไบต์ (TB) และเพตาไบต์ (PB) มากกว่า[24]

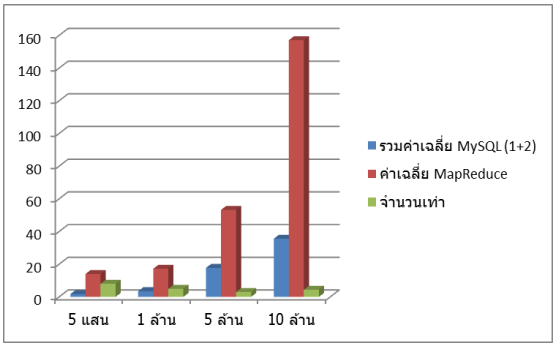
งานวิจัยในอนาคต

จากการทดลองนี้ยังไม่บรรลุวัตถุประสงค์ในการหาจุดตัดของกราฟด้านผลความเร็วว่าจุดใดที่ระบบฐานข้อมูลเชิงสัมพันธ์ไม่สามารถดำเนินการได้หรือมีประสิทธิภาพด้อยกว่าเทคโนโลยีข้อมูลขนาดใหญ่ เพื่อค้นหาจุดที่เริ่มต้องปรับเปลี่ยนระบบ

ตารางที่ 9 เปรียบเทียบค่าเฉลี่ยความเร็วในการประมวลผล

ชุดข้อมูล	รวมค่าเฉลี่ย MySQL (1+2)	ค่าเฉลี่ย MapReduce	จำนวนเท่า
5 แสน	1.7459	14.0173	8.0287
1 ล้าน	3.5053	17.2447	4.9195
5 ล้าน	17.7370	53.2253	3.0024
10 ล้าน	35.5659	157.0667	4.4162

จากผลรวมค่าเฉลี่ยความเร็วดัง**ตารางที่ 9** ที่ใช้ในการประมวลผล หากนำมาคำนวณหาเวลาต่างกันเป็นกี่เท่าจะพบว่ามีแนวโน้มที่ลดลง จึงเป็นไปได้ว่าเมื่อข้อมูลมีเพิ่มขึ้นถึง 50 ล้านระเบียน ประสิทธิภาพด้านความเร็วการประมวลผลด้วยเทคนิคแมพรีดิวจะดีกว่า ผู้วิจัยจึงเห็นว่าสามารถนำมาใช้ในงานทดลองครั้งต่อไป ด้วยการเพิ่มระเบียน และจากการสังเกตในขั้นตอนการทดสอบ เมื่อเทียบขั้นตอนการประมวลผลกันแล้วพบว่าหากใช้ขั้นตอนในการเขียนคำสั่งโปรแกรมแมพรีดิวให้เหมือนกับขั้นตอนของการใช้ภาษาสอบถามเชิงโครงสร้างเอสคิวแอล จะเป็นการลดเวลาในขั้นตอนการประมวลผลส่วนการจับคู่เชื่อมสัมพันธ์ลงได้[29] อีกทั้งการเพิ่มการเรียงลำดับอินเด็กซ์ไว้ล่วงหน้า และการกำหนดขนาดบล็อกข้อมูลใน HDFS จะเป็นการเพิ่มประสิทธิภาพความเร็วได้ จาก**ภาพที่ 5** เมื่อนำ**ตารางที่ 9** มาจัดทำกราฟแสดงผลภาพ



ภาพที่ 5 กราฟแสดงผลการเปรียบเทียบค่าเฉลี่ยความเร็ว

กิตติกรรมประกาศ

ขอขอบพระคุณท่าน ผศ.ดร.วรพล พงษ์เพ็ชร คณบดีคณะวิศวกรรมศาสตร์ มธบ. ที่เอื้อเฟื้อและอนุญาตให้ใช้สถานที่ทำการทดลอง และท่านอาจารย์นุกูล พิมเสน ผู้ติดตั้งโปรแกรมฮาดูป และขอบคุณ นายนัฐพงษ์ หนูสิงห์ นักศึกษาวิศวกรรมข้อมูลขนาดใหญ่ผู้เขียนโปรแกรมแมพรีดิวสำหรับใช้ในการทดลองครั้งนี้

เอกสารอ้างอิง

1. คณะกรรมการพัฒนาระบบบริการที่ตอบสนองต่อปัญหาสุขภาพที่สำคัญ (สาขาหลัก). แผนพัฒนาระบบบริการสุขภาพ (Service Plan) กระทรวงสาธารณสุข. กรุงเทพฯ: โรงพิมพ์ชุมนุมสหกรณ์การเกษตรแห่งประเทศไทย จำกัด, 2556, น.1.
2. เมธี จันท์จารุภรณ์. การจัดการเชิงกลยุทธ์ในการพัฒนาสุขภาพ. หน่วยที่ 5 ข้อมูลและสารสนเทศเชิงกลยุทธ์. ครั้งที่ 2. นนทบุรี: มหาวิทยาลัยสุโขทัยธรรมาธิราช, 2556, น.2.
3. สิน พันธุ์พินิจ. เทคนิคการวิจัยทางวิทยาศาสตร์. ครั้งที่ 2. กรุงเทพฯ: พิมพ์ดีการพิมพ์, 2555, น.141-2.
4. สำนักนโยบายและยุทธศาสตร์ สำนักงานปลัดกระทรวงสาธารณสุข. คู่มือการปฏิบัติงานการจัดเก็บและจัดส่งข้อมูลตามโครงสร้างมาตรฐานข้อมูลสุขภาพ กระทรวงสาธารณสุข Version 2.1 (มกราคม 2559) ปีงบประมาณ 2559. ครั้งที่ 1. นนทบุรี: ห้างหุ้นส่วน เอสพี ก๊อปปี้ปริ้น, 2559, น.36.
5. กลุ่มภารกิจด้านข้อมูลข่าวสารสุขภาพ สำนักนโยบายและยุทธศาสตร์ สำนักงานปลัดกระทรวงสาธารณสุข. สรุปรายงานการป่วย พ.ศ.2557. ครั้งที่ 1. นนทบุรี: องค์การสงเคราะห์ทหารผ่านศึกในพระบรมราชูปถัมภ์, 2558, น.5-15.
6. สรชัย พิศาลบุตร. หลักสถิติ. ครั้งที่ 2, กรุงเทพฯ: พิมพ์ดีการพิมพ์, 2555, น.17-21.
7. ชูศรี วงศ์รัตนะ และองอาจ นัยพัฒน์. แบบแผนการวิจัยเชิงทดลองและสถิติวิเคราะห์ แนวคิดพื้นฐานและวิธีการ. ครั้งที่ 1. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์, 2551, น.151.
8. โอภาส เอี่ยมสิริวงศ์. ระบบฐานข้อมูล. กรุงเทพฯ: วี.พริ้นท์(1991), 2558, น.37-40.
9. โอภาส เอี่ยมสิริวงศ์ และสมโภชน์ ชื่นเอี่ยม. คณิตศาสตร์คอมพิวเตอร์. ครั้งที่ 1. กรุงเทพฯ: วี.พริ้นท์(1991), 2558, น.205-19.
10. ชาญชัย ศุภอรรถกร. จัดการฐานข้อมูลด้วย MySQL ฉบับสมบูรณ์. ครั้งที่ 5. กรุงเทพฯ: รีไวว่า, 2557, น.125-6.
11. Schwartz B, Zaitsev P, Tkachenko V, High Performance MySQL Third Edition. O'Reilly Media, 2012, pp.210-38.
12. Plunkett T, Macdonald B, Nelson B, et al. Oracle Big Data Handbook. Oracle Press. 2014, pp.4.
13. White T. Hadoop The Definitive Guide. 3rd.Edition. O'Reilly Media, CA. 2012, pp. 13–4.
14. Miner D, Shook A. MapReduce Design Pattern. O'Reilly Media, CA. 2012, pp. 4-7.
15. Gunarathne T, Perera S. Hadoop MapReduce v2 Cookbook Second Edition. Packt Publishing, Birmingham, UK. 2015, pp.60- 6.
16. Singh D, Reddy CK. A survey on platforms for big data analytics. Journal of Big Data. 2014; 1:8, 1-20.
17. ประกายมาศ ศรีสุขทักษิณ และผสุดี บุญรอด. การเปรียบเทียบความเร็วในการประมวลผลระหว่างฐานข้อมูลเชิงสัมพันธ์และฐานข้อมูลไม่สัมพันธ์แบบเอกสาร. งานประชุมวิชาการ The 10th National Conference on Computing and Information Technology, ภูเก็ต, 8-9 พ.ค. 2557, โรงแรมอังสนา ลากูน่า ภูเก็ต, น.281-286.
18. ผุสดี บุญรอด, ประกายมาศ ศรีสุขทักษิณ. การค้นคืนข้อมูลขนาดใหญ่โดยใช้ภาษาสอบถามแบบไม่มีโครงสร้างร่วมกับเทคโนโลยีเว็บเชิงความหมาย. วารสารวิชาการพระจอมเกล้าพระนครเหนือ, ปีที่ 25, ฉบับที่ 2, น.255-264. พ.ค.-ส.ค. 2558.
19. Sareen P, Kumar P. NoSQL Database and its Comparison with SQL Database. International Journal of Computer Science & Communication Network. 2015;5(5):293-8.
20. Deepika P, Anantha Raman GR. A Study of Hadoop-Related Tools and Techniques. International Journal of Advanced Research in Computer Science and Software Engineering. 2015;5(9): 160-4.
21. Gurevich Y. Comparative Survey of NoSQL/ NewSQL DB Systems, Final Paper Submitted as partial fulfillment of the requirements towards an M.Sc. degree in Computer Science, The Open University of Israel, December 2015, pp.30-31.
22. Kune R, Konugurthi PK, Agarwal, A, et al. The anatomy of big data computing. Wiley Online Library. 2015, 79–105.
23. Vicknair C, Macias M, Zhao Z, et al. A Comparison of a Graph Database and a Relational Database. ACMSE '10, 2010.
24. Bhosale HS, Gadekar, DP. A Review Paper on Big Data and Hadoop. International Journal of Scientific and Research Publications 2014; 4(10):1-7.
25. Dean J, Ghemawat S. MapReduce: Simplied Data Processing on Large Clusters. Google Inc., OSDI 2004, pp.1-13.

26. Appuswamy R, Gkantsidis C, Narayanan D, et al. Scale-up vs Scale-out for Hadoop: Time to rethink?. SoCC'13, 2013.
27. ชูพันธุ์ รัตนโภคา, การออกแบบและพัฒนาระบบค้นหาข้อมูลจราจรทางคอมพิวเตอร์ด้วยวิธี Map/Reduce บนกรอบการทำงานของ Hadoop. วารสารวิชาการเทคโนโลยีอุตสาหกรรมปีที่ 8, ฉบับที่ 3, กันยายน–ธันวาคม 2555, น.18-27.
28. Tien DD. Hadoop Performance Evaluation. https://wr.informatik.uni-hamburg.de/_media/research/labs/2009/2009-12-tien_duc_dinh-evaluierung_von_hadoop-report.pdf. (2009, accessed 12 October 2015).
29. นิรุทธ์ รวยรื่น, เกรียงไกร ปอแก้ว, การใช้แมพรีดิวซ์เชื่อมคอลเลคชันของฐานข้อมูลโนเอสคิวแอลบนมองโกดีบี. วารสารวิจัย มข. (บศ.) 14 (2), เม.ย.-มิ.ย. 2557, น. 23-34.
30. Fegaras L, Li C, Gupta U. An Optimization Framework for Map-Reduce Queries. EDBT 2012, 2012.
31. Tao Y, Lin W, Xiao, X. Minimal MapReduce Algorithms. SIGMOD'13, June 22–27, 2013, NYork:USA, pp.1-13.
32. Rae M. Hadoop and Hive as Scalable Alternatives to RDBMS: A Case Study. Boise State University, Department of Computer Science. 2012, pp.35-44.