การพัฒนาระบบค้นคืนข้อมูลจากโซเชียลเน็ตเวิร์คด้วย PENTAHO

DEVELOPMENT OF SOCIAL NETWORK DATA RETRIEVAL SYSTEM USING PENTAHO

โสรยา สุวรรณธนชัย

สหกิจศึกษานี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์)
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2559

DEVELOPMENT OF SOCIAL NETWORK DATA RETRIEVAL SYSTEM USING PENTAHO

SORAYA SUWANTANACHAI

A COOPERATIVE EDUCATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENT FOR
THE DEGREE OF BACHELOR OF SCIENCE (COMPUTER SCIENCE)
DEPARTMENT OF COMPUTER SCIENCE, FACULTY OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
ACADEMIC YEAR 2016

หัวข้อโครงงานพิเศษ	การพัฒนาระบบค้นคืนข้อมูลจากโซเชียลเน็ตเวิร์คด้วย Pentaho DEVELOPMENT OF SOCIAL NETWORK DATA RETRIEVAL SYSTEM USING PENTAHO
ชื่อนักศึกษา	นางสาวโสรยา สุวรรณธนชัย รหัสนักศึกษา 57050352
ปริญญา	วิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์)
ภาควิชา	วิทยาการคอมพิวเตอร์
ปีการศึกษา	2559
อาจารย์ที่ปรึกษา	ผศ.ดร.วรางคณา กิ้มปาน

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.) อนุมัติให้สหกิจศึกษานี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์) ประจำปีการศึกษา 2559

คณะกรรมการสอบ	ลายมือชื่อ
ดร.รุ่งรัตน์ เวียงศรีพนาวัลย์ ประธานกรรมการ	
ผศ.ดร.วรางคณา กิ้มปาน กรรมการและอาจารย์ที่ปรึกษา	

หัวข้อสหกิจศึกษา การพัฒนาระบบค้นคืนข้อมูลจากโซเชียลเน็ตเวิร์คด้วย PENTAHO
ชื่อนักศึกษา นางสาวโสรยา สุวรรณธนชัย รหัสนักศึกษา 57050352
ปริญญา วิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์)
ภาควิชา วิทยาการคอมพิวเตอร์
คณะ วิทยาศาสตร์
มหาวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2559
อาจารย์ที่ปรึกษา ผศ.ดร.วรางคณา กิ้มปาน

บทคัดย่อ

เนื่องจากในปัจจุบันมีข้อมูลจำนวนมหาศาลที่เกิดขึ้นบนโลกโซเชียลเน็ตเวิร์คบนเว็บไซต์ต่างๆ ซึ่งข้อมูลเหล่านี้เป็นข้อมูลที่ซับซ้อนและไม่ได้อยู่ในรูปแบบโครงสร้างเดียวกัน การจะนำข้อมูลจำนวนมหาศาลที่ได้มาใช้ให้เกิดประโยชน์ ต้องอาศัยเทคโนโลยีการจัดการข้อมูลขนาดใหญ่ หรือบิ๊กดาต้า (Bigdata) เนื่องจากการที่ข้อมูลมีจำนวนมากขึ้นมหาศาล มีหลายรูปแบบ และข้อมูลมีการเพิ่มขึ้นอย่างรวดเร็ว ทำให้องค์กรต่างๆต้องปรับโครงสร้างพื้นฐานด้านข้อมูล (Information Infrastructure) มีการนำเทคโนโลยีใหม่เช่น Hadoop, NoSQL หรือ NewSQL เข้ามาใช้งาน ต้องมีการพัฒนาบุคลากรเพื่อให้เข้าใจการใช้เทคโนโลยีเหล่านี้ รวมถึงมีความรู้ความสามารถในการนำข้อมูลต่างๆไปทำการวิเคราะห์และสรุปผลการวิเคราะห์ให้อยู่ในรูปแบบที่เข้าใจง่าย โดยได้พัฒนาระบบค้นคืนข้อมูลจากโซเชียลเน็ตเวิร์คด้วย Pentaho เก็บลงใน Hadoop โดยเก็บข้อมูลจากโซเชียลเน็ตเวิร์คจากแหล่งต่างๆ เช่น เฟซบุ๊ก (Facebook) ทวิตเตอร์ (Twitter) พันทิป (Pantip) และเว็บข่าว (RSS) ต่างๆ มีการทำให้ข้อมูลมีความถูกต้อง (Data Cleaning) ก่อนการเก็บลงใน Hadoop เพื่อทำให้ผู้ใช้สามารถนำข้อมูลไปใช้ในการวิเคราะห์ได้อย่างสะดวกและรวดเร็ว

คำสำคัญ : ฮาดูป Pentaho เฟซบุ๊ก ทวิตเตอร์ เว็บข่าว

Title	Development of Social Network Data Retrieval System Using Pentaho
Students	Miss Soraya Suwantanachai Student ID 57050352
Degree	Bachelor of Science (Computer Science)
Department	Computer Science
Faculty	Science
University	King Mongkut's Institute of Technology Ladkrabang (KMITL)
Academic Year	2016
Advisor	Asst.Prof.Dr. Warangkhana Kimpan

Abstract

Today, there is a tremendous amount of information available on the social networks on the web. These data are complex and not in the same structure. The vast amount of information that is used to benefit. Bigdata technology is required because of the tremendous amount of data available and the rapid growth of data. New organizations such as Hadoop, NoSQL or NewSQL need to be deployed to improve their knowledge of the use of these technologies. They also have the knowledge and skills to analyze and summarize the results in an easy-to-understand format. The development of data retrieval system from the social network using Pentaho. It will retrieve the required data into Hadoop using Pentaho. The system collects the data from the social networks such as Facebook, Twitter, and RSS (Really Simple Syndication). Data validation was done before being stored in Hadoop. This allows users to easily and quickly analyze data.

Keywords : Hadoop, Pentaho, Facebook, Twitter, RSS

กิตติกรรมประกาศ

สหกิจศึกษาเล่มนี้สำเร็จลุล่วงไปได้ด้วยดี เนื่องจากได้รับความช่วยเหลืออย่างดีจากอาจารย์ที่ปรึกษาสหกิจศึกษา ผศ.ดร.วรางคณา กิ้มปาน ที่คอยดูแลการทำสหกิจศึกษามาโดยตลอดและให้คำปรึกษาแนะนำอย่างใกล้ชิด รวมทั้งเสนอแนะแนวทางแก้ปัญหา ตรวจแก้สหกิจศึกษาฉบับนี้ให้มีความสมบูรณ์มากยิ่งขึ้น ผู้จัดทำสหกิจศึกษารู้สึกซาบซึ้งในความกรุณาของอาจารย์เป็นอย่างยิ่งและขอขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอขอบพระคุณ บริษัท ซีดีจีซีสเท็ม จำกัด ที่ได้ให้ความอนุเคราะห์ในการศึกษาหาข้อมูลในการทำวิจัยรวมถึงขอขอบพระคุณ คุณศริญญา โคมี ที่ได้ให้ความรู้ คำปรึกษาและอำนวยความสะดวกในการดำเนินงานมาโดยตลอด

สุดท้ายนี้ขอกราบขอบพระคุณบิดา มารดา ที่ได้ให้การสนับสนุนด้านการเรียน ให้คำปรึกษาและคอยเป็นกำลังใจที่สำคัญจนการเรียนผ่านพ้นไปได้ด้วยดี ผู้จัดทำสหกิจศึกษาจึงใคร่ขอขอบพระคุณ ทุกท่านเป็นอย่างสูงไว้ ณ ที่นี้

โสรยา สุวรรณธนชัย

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

สารบัญรูป (ต่อ)

สารบัญรูป (ต่อ)

บทที่1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เนื่องจากในปัจจุบันมีข้อมูลจำนวนมหาศาลที่เกิดขึ้นบนโลกโซเชียลมีเดีย (Social Media) บนเว็บไซต์ต่างๆ ซึ่งข้อมูลเหล่านี้เป็นข้อมูลที่ซับซ้อนและไม่ได้อยู่ในรูปแบบโครงสร้างเดียวกัน การจะนำข้อมูลจำนวนมหาศาลที่ได้มาใช้ให้เกิดประโยชน์ ต้องอาศัยเทคโนโลยีการจัดการข้อมูลขนาดใหญ่ หรือบิ๊กดาต้า (Bigdata) เนื่องจากการที่ข้อมูลมีจำนวนมากขึ้นมหาศาล มีหลายรูปแบบ และการมีข้อมูลที่เพิ่มขึ้นอย่างรวดเร็ว ทำให้องค์กรต่างๆต้องปรับโครงสร้างพื้นฐานด้านข้อมูล (Information Infrastructure) มีการนำเทคโนโลยีใหม่เช่น Hadoop, NoSQL หรือ NewSQL เข้ามาใช้งาน ต้องมีการพัฒนาบุคลากรเพื่อให้เข้าใจการใช้เทคโนโลยีเหล่านี้ รวมถึงมีความรู้ความสามารถในการนำข้อมูลต่างๆไปทำการวิเคราะห์และสรุปผลการวิเคราะห์ให้อยู่ในรูปแบบที่เข้าใจง่ายเช่น อยู่ในรูปแบบของ vistualization ดังนั้นโครงงานสหกิจศึกษาที่บริษัท CDGS ในครั้งนี้นั้นได้ใช้เครื่องมือ Pentaho ในการช่วยเก็บข้อมูลจากโซเชียลเน็ตเวิร์คแหล่งต่างๆ เช่น เฟซบุ๊ก (Facebook) ทวิตเตอร์ (Twitter) พันทิป (Pantip) และเว็บข่าว (RSS) ต่างๆ โดยมีการทำให้ข้อมูลมีความถูกต้อง (Data Cleaning) ก่อนการเก็บลงในฮาดูป (Hadoop) เพื่อทำให้ผู้ใช้สามารถนำข้อมูลไปใช้ในการวิเคราะห์ได้อย่างสะดวกและรวดเร็ว

1.2 วัตถุประสงค์

1) เพื่อเก็บข้อมูลที่มีขนาดมหาศาลจากโซเชียลเน็ตเวิร์คแหล่งต่างๆ เช่น เฟซบุ๊ก (Facebook) ทวิตเตอร์ (Twitter) พันทิป (Pantip) และเว็บข่าว (RSS) ต่างๆ
2) เพื่อศึกษาเทคโนโลยีที่ใช้ในการประมวลผลข้อมูลขนาดใหญ่
3) เพื่อจัดเก็บและนำข้อมูลที่ได้มาวิเคราะห์ตามความต้องการ
4) เพื่อศึกษาข้อมูลให้สามารถนำไปพัฒนาต่อยอดได้

1.3 ขอบเขตของสหกิจศึกษา

1) ศึกษาเครื่องมือการดึงข้อมูล และเทคโนโลยีการจัดการข้อมูลขนาดใหญ่ หรือบิ๊กดาต้า (Bigdata)
2) พัฒนาระบบงานดังนี้
   - ดึงข้อมูลจาก Social media แหล่งต่างๆ
   - จัดเก็บข้อมูลที่ได้จาก Social media ลงในที่เก็บข้อมูล Hadoop

- นำข้อมูลที่อยู่ใน Hadoop ออกมาแสดงผลในรูปแบบต่างๆ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) สามารถนำข้อมูลที่ได้ไปวิเคราะห์เพื่อให้เหมาะสมกับความต้องการของผู้ใช้
2) ได้เรียนรู้การจัดการข้อมูลและเทคโนโลยีใหม่ๆ
3) เข้าใจถึงการทำงานของ Hadoop cluster
4) เข้าใจวิธีการดึงข้อมูลจากโซเชียลเน็ตเวิร์คแหล่งต่างๆ
5) เข้าใจวิธีการนำข้อมูลจาก Hadoop มาแสดงในหน้าเว็บ
6) สามารถนำความรู้ไปพัฒนาต่อยอดได้

1.5 เครื่องมือที่ใช้

1) ซอฟต์แวร์ที่ใช้

- Pentaho ใช้ในการดึงข้อมูลจากโซเชียลเน็ตเวิร์คแหล่งต่างๆ
- Cloudera Hadoop ใช้สำหรับเก็บข้อมูลที่ได้จากโซเชียลเน็ตเวิร์ค
- Docker ใช้สำหรับจำลองเครื่องแม่ข่าย (Web Server)
- Ecilp ใช้ในการเขียนโปรแกรมให้บริการเว็บไซต์ (Web Service) เพื่อเชื่อมต่อกับ Hadoop
- Putty ใช้สำหรับ Remote ไปยังเครื่องแม่ข่าย

2) ฮาร์ดแวร์ที่ใช้

- คอมพิวเตอร์แบบพกพา (หน่วยประมวลผลกลาง Intel I7, หน่วยความจำหลัก 8 GB, หน่วยความจำสำรอง 500 GB, ระบบปฏิบัติการ Windows 10)

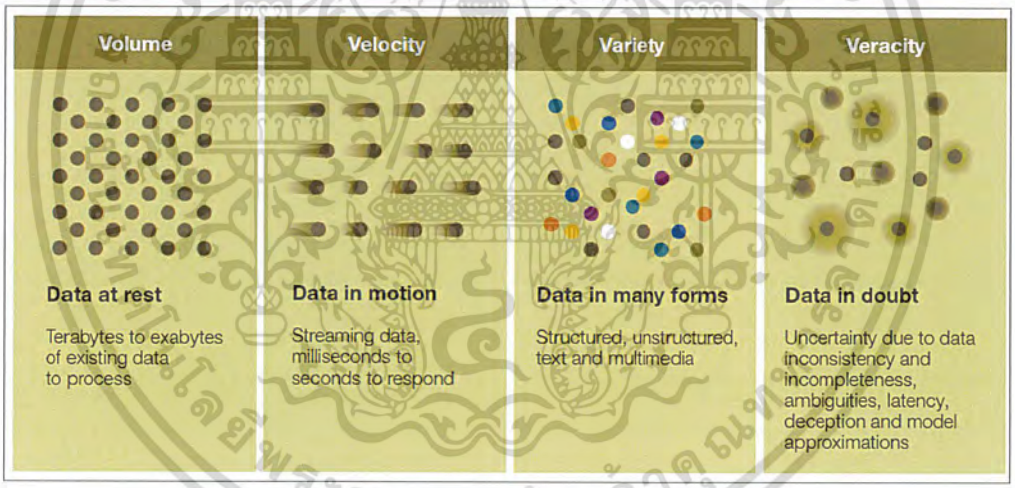
บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 บิ๊กดาต้า

บิ๊กดาต้า [1] คือชุมนุมของชุดข้อมูลที่มีขนาดใหญ่และมีความซับซ้อนมาก จนยากที่จะประมวลผลได้ด้วยเครื่องมือจัดการฐานข้อมูลที่มีอยู่ บิ๊กดาต้ามักรวมถึงชุดข้อมูลที่มีขนาดใหญ่เกินกว่าความสามารถของซอฟต์แวร์ที่ใช้กันอยู่ทั่วไปในการบันทึก จัดการ และประมวลผลข้อมูลดังกล่าวได้ภายในเวลาที่ยอมรับได้ ขนาดของบิ๊กดาต้านั้นอยู่ที่ตั้งแต่ไม่กี่เทราไบต์ไปจนถึงหลายๆ เพตาไบต์ในชุดข้อมูลชุดเดียว ด้วยความยากลำบากนี้ แพลตฟอร์มใหม่สำหรับบิ๊กดาต้าจึงได้เกิดขึ้นเพื่อสามารถจัดการกับข้อมูลจำนวนมาก ตัวอย่างเช่น Apache Hadoop Big Data Platform

2.1.1 ลักษณะของบิ๊กดาต้า

ลักษณะของบิ๊กดาต้ามีการแบ่งออกเป็น 4 ลักษณะ [1] แสดงดังรูป 2.1



รูปที่ 2.1 ลักษณะของบิ๊กดาต้า

1) ปริมาณ (Volume)

ปริมาณของข้อมูลมีการโตขึ้นเรื่อยๆ จนถึงขนาดเทราไบต์ (Terabyte) หรือเพทาไบท์ (Petabyte) เช่น ข้อมูล 12 เทราไบต์จากการ Tweet ในแต่ละวัน

2) ความเร็ว (Velocity)

ปัจจุบันแหล่งของข้อมูลมาจากสื่อสังคมและอุปกรณ์มือถือ รวมทั้งอุปกรณ์คอมพิวเตอร์เคลื่อนที่ กระแสของข้อมูลข่าวสารที่เข้ามายังเครื่องแม่ข่ายเป็นแบบ Real-Time และมีความต่อเนื่อง เช่น Twitter มีการสร้างข้อมูล มากกว่า 7 Terabytes (TB) ต่อวัน Facebook 10 Terabytes ต่อวัน องค์กรขนาดวิสาหกิจทั่วไป 1 Terabytes ต่อชั่วโมงต่อวันทำงาน

3) ความหลากหลาย (Variety)

ข้อมูลที่ใช้งานในปัจจุบัน โครงสร้างข้อมูลมีรูปแบบเพิ่มมากยิ่งขึ้น เริ่มตั้งแต่ข้อความเปล่าๆ ภาพถ่าย แฟ้มข้อมูลเสียงเพลง แฟ้มข้อมูลวิดีโอ ข้อมูลเว็บ ข้อมูล GPS ข้อมูลจากเซ็นเซอร์ต่างๆ ข้อมูลจากฐานข้อมูลเชิงสัมพันธ์ เอกสารทั่วไป ข่าวสาร แฟ้มข้อมูลประเภท PDF แฟ้มข้อมูล Flash และอื่นๆมากมาย

4) Veracity

ข้อมูลที่ได้มีความถูกต้องแม่นยำ เนื่องจากข้อมูลมีความหลากหลายจากแหล่งต่างๆที่อยู่เหนือการควบคุมเช่น Facebook, Twitter, YouTube

2.1.2 เทคโนโลยีสำหรับประมวลผลบิ๊กดาต้า

การที่บิ๊กดาต้าจะเชื่อมโยงไปสู่ระบบการประมวลผลสำหรับข้อมูลปริมาณมาก สามารถจัดแบ่งเทคโนโลยีออกเป็น 4 กลุ่ม ดังนี้

1) กลุ่มที่หนึ่ง คือ กลุ่มซอฟต์แวร์ที่เปิดเผย source code ของโปรแกรม ทำให้สามารถแก้ไขดัดแปลง source code ได้หมด ซึ่งเป็นการให้สิทธิเสรีแก่ผู้ที่จะนำไปใช้เพื่อการพัฒนาซอฟต์แวร์ร่วมกันในลักษณะของสังคม ซอฟต์แวร์ที่มีการศึกษาเพื่อใช้ในการพัฒนาบิ๊กดาต้าคือ Apache Hadoop และ Cloudera

2) กลุ่มที่สอง คือ ระบบฐานข้อมูลที่ไม่ใช้ภาษา SQL (NoSQL Database) เนื่องจากความสามารถที่รวดเร็ว สามารถรองรับข้อมูลแบบกึ่งโครงสร้างและไม่มีโครงสร้างได้ รองรับการขยายตัวในแนวราบ (Horizontal Scaling) ซึ่งสอดคล้องกับสถาปัตยกรรมของ Hadoop ตัวอย่างผลิตภัณฑ์ทางด้าน NoSQL Database ที่เป็นที่นิยมได้แก่ Cassandra, CouchBase, HBase, MongoDB เป็นต้น

3) กลุ่มที่สาม คือ Data Visualization Tools ซึ่งเป็นเครื่องมือที่จะช่วยแปลงข้อมูลบิ๊กดาต้าที่ได้รับการกรองแล้วมาแสดงในรูปของแผนภาพ เพื่อให้ง่ายต่อการเข้าใจและนำไปสู่การตัดสินใจในขั้นถัดไป

4) กลุ่มที่สี่ คือ Analytic Database กลุ่มนี้จะนำไปใช้กับระบบคลังข้อมูลได้ด้วย โดยใช้เทคนิคในการทำงานแบบต่างๆ เพื่อตอบโจทย์ด้านความเร็ว นอกจากนี้ยังมีเทคโนโลยีการบริหารจัดการบิ๊กดาต้าในลักษณะ Cloud Computing

2.2 Apache Hadoop

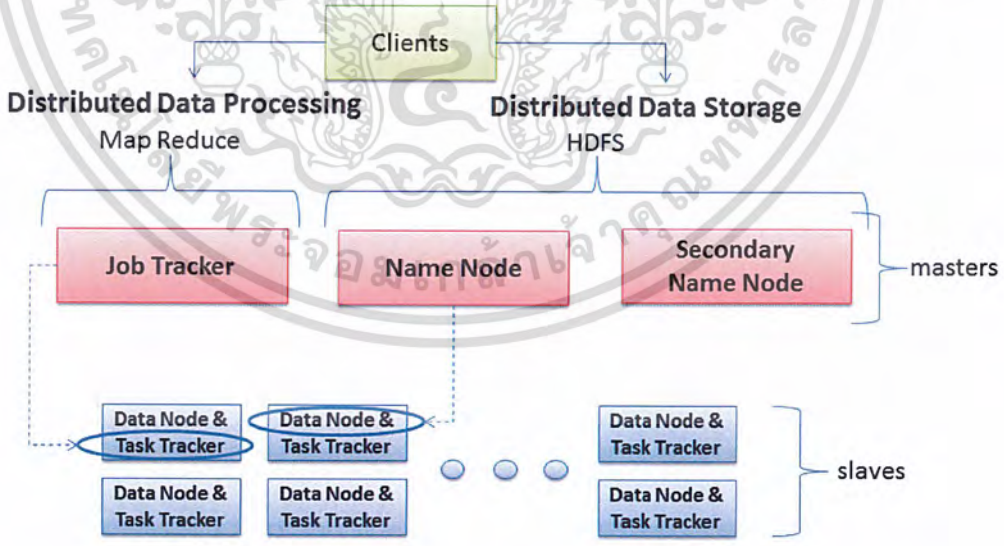
Hadoop [2] เป็นโอเพ่นซอร์สของ Apache สำหรับการเก็บและบริหารข้อมูลขนาดใหญ่ Hadoop เขียนด้วยโปรแกรมภาษาจาวา มีความสามารถในการป้องกันข้อมูลเสียหาย (Fault Tourarent) เพราะจะเก็บข้อมูลซ้ำกันในหลายๆที่ และเป็นระบบที่สามารถเพิ่มจำนวนเครื่องที่ใช้ในการประมวลผล (Horizontal Scale) รันบนเครื่องเซิร์ฟเวอร์ทั่วๆไป (Commodity Server) จำนวนมาก Hadoop เริ่มต้นโดย Doug Cutting และ Mike Cafarella ที่เป็นทีมงานของบริษัท Yahoo ซึ่งต่อมากมีบริษัทอื่นๆนำไปใช้กันอย่างมากทั้ง eBay, Facebook และ Amazon รวมถึงมีบริษัทหลายๆ รายที่นำ Hadoop มาทำ Hadoop Distribution อาทิเช่น Cloudera, MapR, IBM Infoshphere BigInsight, Hortonwork หรือ Amazon Elastic Map Reduce

องค์ประกอบของ Hadoop ด้วยความซับซ้อนของบิ๊กดาต้า จึงทำให้ Hadoop แบ่งออกเป็นโมดูลย่อยๆ ดังนี้

1) Hadoop Distributed File System (HDFS) โมดูลนี้จะเอาไว้ใช้จัดเก็บข้อมูลที่จะนำมาวิเคราะห์ให้อยู่ในรูปที่สามารถเข้าถึงได้อย่างรวดเร็ว รวมไปถึงการสำรองข้อมูลดังกล่าวให้โดยอัตโนมัติ

2) MapReduce โมดูลนี้จะเอาไว้ใช้เกี่ยวกับการประมวลผลข้อมูลปริมาณมหาศาลที่ได้เก็บเอาไว้

สถาปัตยกรรมฮาร์ดแวร์ของระบบ Hadoop [2] แสดงดังรูปที่ 2.2



รูปที่ 2.2 สถาปัตยกรรมฮาร์ดแวร์ของระบบ Hadoop

จากรูปที่ 2.2 สามารถอธิบายฮาร์ดแวร์ของระบบ Hadoop ได้ดังนี้

ระบบ Hadoop ประกอบด้วยเครื่องแม่ข่ายจำนวนมาก โดยจะมีเครื่องหนึ่งทำหน้าที่เป็น Master และมีเครื่องลูกอีกจำนวนมากทำหน้าที่เป็น Slave ปกติ Hadoop กำหนดให้ข้อมูลที่เก็บในเครื่อง Slave มีการเก็บข้อมูลซ้ำกันสามแห่ง ดังนั้นเครื่อง Slave ควรจะมีอย่างน้อยสามเครื่อง ส่วนเครื่อง Master จะทำหน้าที่หลักในการระบุตำแหน่งของข้อมูลและงาน (Task) ที่กระจายในการประมวลผลของ Map/Reduce ดังนั้นเครื่อง Master จึงมีความสำคัญอย่างมาก และต้องมีเครื่อง Secondary Master ในการที่จะสำรองไว้ในกรณีเครื่อง Master ไม่สามารถใช้งานได้ ดังนั้นระบบ Hadoop โดยทั่วไปจะเริ่มต้นที่เครื่องแม่ข่าย จำนวน 5 เครื่อง สำหรับ Master จำนวน 1 เครื่อง Secondary Master จำนวน 1 เครื่อง และ Slave จำนวน 3 เครื่อง หากต้องการเก็บข้อมูลมากขึ้น หรือต้องการประมวลผลข้อมูลให้เร็วขึ้น ต้องเพิ่มจำนวนเครื่อง Slave ให้มากขึ้น ขนาดของข้อมูลที่เก็บได้จะขึ้นอยู่กับขนาดความจุข้อมูลของเครื่อง Slave รวมกันหารด้วยจำนวนข้อมูลที่ต้องการเก็บซ้ำ (ค่าเริ่มต้น คือ 3) ซึ่งการเก็บข้อมูลจำนวนเป็น Petabyte ต้องมีเครื่องเป็นจำนวนมากกว่าร้อยเครื่อง โหนด (Node) หมายถึงเครื่องคอมพิวเตอร์ที่ประกอบไปด้วย CPU, RAM และ Disk ซึ่งโหนดต่างๆ ใน Hadoop จะแบ่งออกเป็น 2 แบบ ดังนี้

1) Data Node เป็นโหนดที่ทำหน้าที่เก็บ Block ของไฟล์เอาไว้ และรับผิดชอบในการประมวลผล Block นั้นๆ แต่ Data Node จะไม่รู้ Block ที่ตัวเองเก็บอยู่เป็นของไฟล์ใด

2) Name Node เป็นโหนดที่ทำหน้าที่รวบรวมผลของการประมวลผล Block ต่างๆ จาก Data Node ซึ่ง Name Node จะรู้ทุกอย่างเกี่ยวกับไฟล์ต้นฉบับ เช่น ชื่อไฟล์ ขนาด ที่อยู่ของแต่ละ Block ที่ถูกกระจายไปตาม Data Node ต่างๆ ซึ่ง Name Node คือ Master ส่วน Data Node คือ Slave

ค่าเริ่มต้นของการทำสำเนาแต่ละ Block จะอยู่ที่ 3 โหนด การจะใช้ Hadoop ได้อย่างมีประสิทธิภาพนั้น ต้องมีอย่างน้อย 5 เครื่อง คือ Data Node จำนวน 3 เครื่อง Name Node จำนวน 1 เครื่อง และ Name Node secondary จำนวน 1 เครื่อง

2.3 Apache Impala

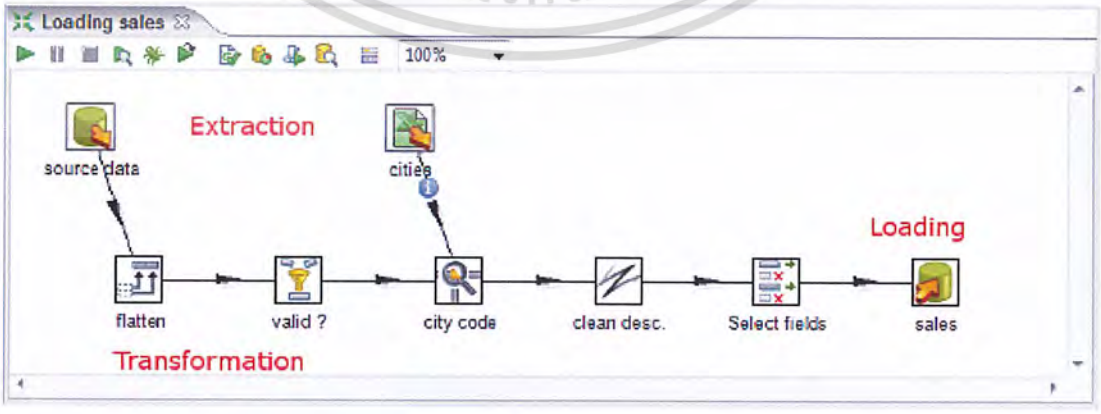
Impala [3] เป็น Massively Parallel Processing (MPP) ที่มีหน่วยประมวลผลที่ใช้ระบบปฏิบัติการและหน่วยความจำของตนเอง มีความสามารถในการค้นหาข้อมูล โดยใช้คำสั่ง SQL สำหรับการประมวลผลข้อมูลปริมาณมากที่ถูกเก็บไว้ในคลัสเตอร์ Hadoop Impala เป็นซอฟต์แวร์โอเพนซอร์ส ภายใต้ Apache License ที่มีประสิทธิภาพสูงเมื่อเทียบกับเครื่องมืออื่น ๆ สำหรับ Hadoop ข้อดีของ Impala สามารถประมวลผลข้อมูลที่เก็บไว้ใน HDFS ได้อย่างรวดเร็วและสามารถเข้าถึงข้อมูลที่ถูกเก็บไว้ใน HDFS, HBase และ Amazon S3 ด้วยคำสั่ง SQL คุณสมบัติของ Impala สามารถรองรับหน่วยความจำในการประมวลผลข้อมูล เช่น เข้าถึงหรือวิเคราะห์ข้อมูลที่เก็บไว้ในโหนดข้อมูล Hadoop โดยไม่ต้องเคลื่อนย้ายข้อมูลและ Impala สามารถใช้งานกับเครื่องมือทางธุรกิจ เช่น Tableau หรือซอฟต์แวร์ Pentaho ได้

2.4 Pentaho Data Integration

Pentaho Data Integration [4] คือเครื่องมือที่ใช้สร้างคลังข้อมูล (Data Warehouse) โดยทั่วไปเครื่องมือดังกล่าวเรียกว่า ETL (Extraction, Transformation and Loading) ดังนี้

- Extraction คือการดึงข้อมูลจากแหล่งต่างๆ ที่ต้องการมาเก็บไว้ใน Data Warehouse โดยจะดึงมาเฉพาะข้อมูลใหม่ที่เพิ่มขึ้นมาหรือข้อมูลที่ถูกเปลี่ยนแปลงแก้ไขโดยข้อมูลที่ดึงมาจะนำมาเก็บพักไว้ก่อน
- Transformation คือการเปลี่ยนแปลงรูปแบบของข้อมูลที่ได้จากการ Extract ให้อยู่ในรูปแบบที่ถูกต้องตามโครงสร้างของ Data Warehouse
- Loading คือการเก็บข้อมูลลงใน Data Warehouse หลังจากทำการแปลงข้อมูลให้อยู่ในรูปแบบที่ถูกต้อง

ความสามารถของ Pentaho Data Integration ประกอบด้วย รองรับมาตรฐานภาษาจาวาและง่ายต่อการใช้งานด้วยเครื่องมือต่างๆ ที่จัดเตรียมไว้ให้ในรูปแบบกราฟิก เพียงลากวางเครื่องมือต่างๆ ตามกระบวนการที่ต้องทำ มีเครื่องมือที่ใช้ในการตรวจสอบความถูกต้องของข้อมูล (Data Quality) อีกทั้งยังรองรับหลากหลายแหล่งข้อมูล (Data Source) เช่น แฟ้มข้อมูล (File Based) แฟ้มข้อความ (Text File) ฐานข้อมูลประเภทต่างๆ เช่น MySQL, PostgreSQL, Oracle เป็นต้น สามารถเชื่อมต่อกับ Pentaho BI ทำให้ใช้ความสามารถอื่นๆร่วมกันได้ เช่น เรื่องของการจัดตารางการทำงาน (Scheduling) ความปลอดภัย (Security) ขั้นตอนการทำงาน (Workflow) เป็นต้น Pentaho Data Integration สามารถนำไปใช้งานในลักษณะต่างๆ ดังนี้ สร้างคลังข้อมูล (Populate Data Warehouse) ส่งออก (Export) ข้อมูลจากฐานข้อมูลไปเป็นแฟ้มข้อความ นำเข้า (Import) ข้อมูลจากแฟ้มข้อความเข้าฐานข้อมูล นำข้อมูลจากฐานข้อมูลหนึ่งไปเข้าอีกฐานข้อมูลหนึ่ง และดูข้อมูลจากฐานข้อมูลที่มีอยู่แล้ว ตัวอย่างการทำงานของ Pentaho Data Integration[4] แสดงดังรูปที่ 2.3



รูปที่ 2.3 ตัวอย่างการทำงานของ Pentaho Data Integration

2.5 Application Programming Interface

Application Programming Interface (API) [5] คือช่องทางการเชื่อมต่อระหว่างเว็บไซต์หนึ่งไปยังอีกเว็บไซต์หนึ่ง หรือเป็นการเชื่อมต่อระหว่าผู้ใช้งานกับเครื่องแม่ข่าย หรือจากเครื่องแม่ข่ายเชื่อมต่อไปหาเครื่องแม่ข่ายด้วยกัน ซึ่ง API เปรียบได้เป็นภาษาคอมพิวเตอร์ที่ทำให้คอมพิวเตอร์สามารถสื่อสารและแลกเปลี่ยนข้อมูลกันได้อย่างอิสระ โดยส่วนมาก API ถูกใช้งานกันอย่างแพร่หลายที่เห็นได้อย่างชัดเจนคือ บริการของ Amazon มี API ที่เปิดให้ผู้ที่สนใจที่จะเป็นตัวแทนขายสินค้าหรือเจ้าของเว็บไซต์ทั่วไป ได้นำสินค้าที่มีขายอยู่ใน Amazon ไปติดไว้ในเว็บไซต์หรือบล็อคของตัวเองได้ โดยเจ้าของเว็บไซต์หรือผู้สนใจจะได้รับคอมมิสชั่นเมื่อมีการคลิกซื้อสินค้าจากเว็บไซต์หรือบล็อคที่นำ API ไปติดตั้ง อีกบริการหนึ่งคือบริการของ PayPal API ซึ่งเจ้าของเว็บไซต์ที่ต้องการเพิ่มช่องทางการชำระเงินให้กับลูกค้าสามารถนำ PayPal API ไปติดตั้งที่เว็บไซต์ที่ต้องการได้ เพื่อเพิ่มความสะดวกสบายให้กับลูกค้าที่ใช้บริการในเว็บไซต์ หน้าที่และประโยชน์ของ API มีดังนี้

2.5.1 หน้าที่ของ API

API ทำหน้าที่ช่วยในการเข้าถึงข้อมูลต่างๆ หรือจะเป็นการนำข้อมูลต่างๆออกจากเว็บไซต์ หรือจะเป็นการส่งข้อมูลเข้าไป โดยเจ้าของเว็บไซต์ที่มี API จะกำหนดขอบเขตในการเข้าถึงบริการต่างๆของทางเว็บไซต์ คำสั่งที่ได้จากฝั่งเครื่องลูกข่าย เรียกว่า Request เมื่อเกิดคำสั่งหรือการร้องขอใดๆ ตัว API จะรับคำสั่งนั้นๆ แล้วนำไปประมวลผลและสรุปเป็นก้อนข้อมูลที่ตรงกับการร้องขอและส่งข้อมูลเหล่านั้นกลับไปที่ส่วนของเครื่องลูกข่ายหรือแอปพลิเคชันอีกครั้ง เรียกการทำงานในขั้นตอนนี้ว่า Response

2.5.2 ประโยชน์ของ API

API ช่วยในการพัฒนาเว็บไซต์หรือแอปพลิเคชันได้ง่ายและรวดเร็วซึ่ง API จะเป็นตัวช่วยที่นักพัฒนาไม่ต้องเข้าไปแก้ไข Code คำสั่ง ทำให้สะดวกสบายในการใช้งาน ผู้ใช้งานเว็บไซต์ต่างๆที่มีการติดตั้ง API ของอีกเว็บไซต์หนึ่ง ไม่ต้องเข้าหน้าเว็บไซต์ที่เป็นเจ้าของ API เพียงเข้ามายังเว็บไซต์ที่มีการติดตั้ง API เท่านั้นทำให้การรับรู้ข่าวสารต่างๆ ทั่วถึงกันและสะดวกในการใช้งานของผู้ใช้งานเว็บไซต์ API สามารถรับส่งข้อมูลข้ามเครื่องแม่ข่ายได้

2.6 Docker

Software Container เป็นแนวความคิดของการสร้างสภาพแวดล้อมเฉพาะให้ซอฟต์แวร์ทำงานได้โดยไม่รบกวนกับซอฟต์แวร์ตัวอื่นบนระบบปฏิบัติการเดียวกัน การจำลองและควบคุมสภาพแวดล้อมสำหรับการรันเฉพาะบางบริการเรียกว่า Container ซึ่ง Container สามารถรันในคอมพิวเตอร์ หรือเครื่องแม่ข่ายเครื่องใดก็จะสารมารถทำงานได้เหมือนเดิม โปรแกรมใน Container ยังทำงานได้ปกติไม่แตกต่างจากเดิม Software Container มีการใช้งานมานาน เช่น LXC (Linux

Container), Solaris Containers, OpenVZ เป็นต้น แต่ไม่เป็นที่แพร่หลายมาก เนื่องจากมีการใช้งานค่อนข้างยาก ปัจจุบันได้มี Engine ชื่อว่า Docker เป็นตัวจัดการ Container ที่ใช้งานได้ง่ายกว่าแบบอื่นๆ ทำให้ได้รับความนิยมและเข้ามามีบทบาทในกลุ่ม Developer และ DevOps หรือ System Admin มากขึ้น

Docker [6] เป็น Software Container ที่ถูกพัฒนาขึ้นมาให้สามารถจัดการ Container ได้ง่าย มี Image ขนาดเล็ก แยกเป็นชั้นๆ สร้างแนวคิด build, ship, run ที่แต่ละรอบของการสร้าง Container เร็วขึ้น ทำให้เป็นที่สนใจและแพร่หลายในกลุ่ม Developer และ System Admin

2.6.1 ความแตกต่างระหว่าง Virtual Machine กับ Container

1) Virtual Machine (VM) เป็นการจำลองสภาพแวดล้อมมาทั้งระบบปฏิบัติการ (OS) รันขึ้นมาเป็นเครื่องแม่ข่ายจำนวน 1 เครื่อง และมีการรันบริการหลายๆบริการใน Virtual Machine เดียวกัน ทำให้แต่ละ Virtual Machine ต้องใช้ทรัพยากรมาก
2) Container เป็นการจำลองและควบคุมสภาพแวดล้อมสำหรับการรันเฉพาะบางบริการ เช่น Contianer ที่รัน nginx ใน ubuntu จะบรรจุสภาพแวดล้อมเหล่านี้ไว้เป็น 1 Contianer และรันบริการเท่าที่จำเป็นต้องใช้ ทำให้ใช้ทรัพยากรน้อยกว่า Virtual Machine

2.6.2 องค์ประกอบต่างๆของ Docker

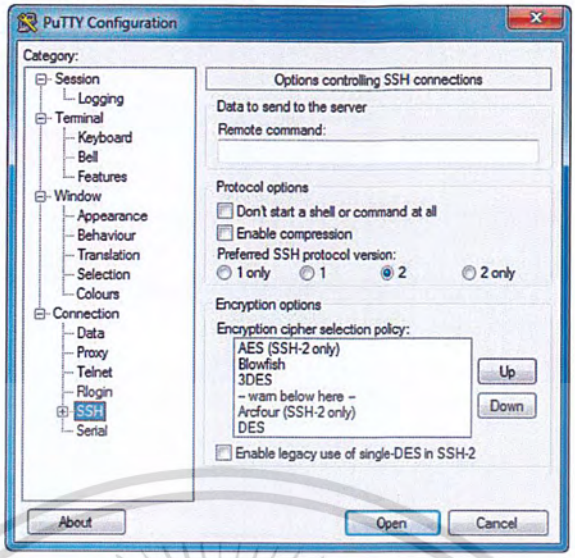
1) Docker image เป็นต้นแบบของ Container ข้างในเป็นลินุกซ์ที่มีการติดตั้งแอปพลิเคชัน และมีการกำหนดค่าเอาไว้ ซึ่งเกิดจากการสร้าง Dockerfile ขึ้นมาเป็น image
2) Docker container ถูกสร้างมาจาก Docker Image ที่เป็นต้นแบบ เกิดเป็น container จะได้บริการ หรือแอปพลิเคชันที่สามารถเรียกใช้งานได้ทันที เมื่อติดตั้ง Docker image นั้นจะได้ Container แอปพลิเคชันอันนั้น โดยใน Container สามารถสั่งเริ่มต้นการทำงาน (Start) หยุดการทำงาน (Stop) ดูการทำงานได้ ซึ่งใน Container แต่ละตัวจะมี RAM, CPU, Users, File
3) Docker registry สามารถสร้าง Docker Image แล้วนำไปเก็บรวมไว้บนเครื่องแม่ข่าย โดย Docker Registry มีให้เลือกใช้งานได้หลากหลาย โดยมี Docker Hub เป็น Docker registry หลัก ในการเรียกใช้ (Pull) Docker Image และนอกจากนี้ยังมีผู้ให้บริการ Docker Registry ของบริษัทอื่นๆด้วย เช่น Gitlab, Quay.io, Google Cloud เป็นต้น

2.6.3 ประโยชน์ของ Docker

1) Docker จะเป็นตัวกลาง ช่วยให้ผู้พัฒนาในทีมรวมถึง Production Server มีสภาพแวดล้อมเดียวกันหมด
2) Docker ช่วยให้จัดการบริการ ที่มีหลากหลายเวอร์ชั่น เช่น php 5.4, php 5.6, php 7 ในเครื่องเดียวกันได้ง่าย
3) Docker ลดเวลาในการติดตั้งบริการลงอย่างมาก ด้วยระบบ Registry ต้องการใช้แอปพลิเคชันสามารถเรียกใช้ (Pull) ลงมาได้เลย
4) สามารถ ลบ และ ติดตั้ง Container ใหม่อย่างรวดเร็วด้วย Docker Image ที่เป็น Template ของบริการ
5) สามารถสร้าง Docker Image เพื่อใช้ในการกำหนดและติดตั้งบริการได้ด้วย Dockerfile
6) Dockerfile สามารถสืบทอดจาก Docker Image ของผู้พัฒนาอื่นได้ เช่น php 5.4 และ php 5.6 สืบทอดมาจาก Apache และ Apache สืบทอดมาจาก Ubuntu
7) สามารถจัดการและควบคุม Container หลายๆตัว เช่นการตั้งค่าให้สามารถทำงานร่วมกัน เริ่มต้นการทำงาน (Start) หยุดการทำงาน (Stop) ทั้งหมดพร้อมกันด้วย Docker Compose
8) สามารถ Scale Container ได้อย่างง่ายดายและรวดเร็วด้วย Docker Swarm

2.7 Putty

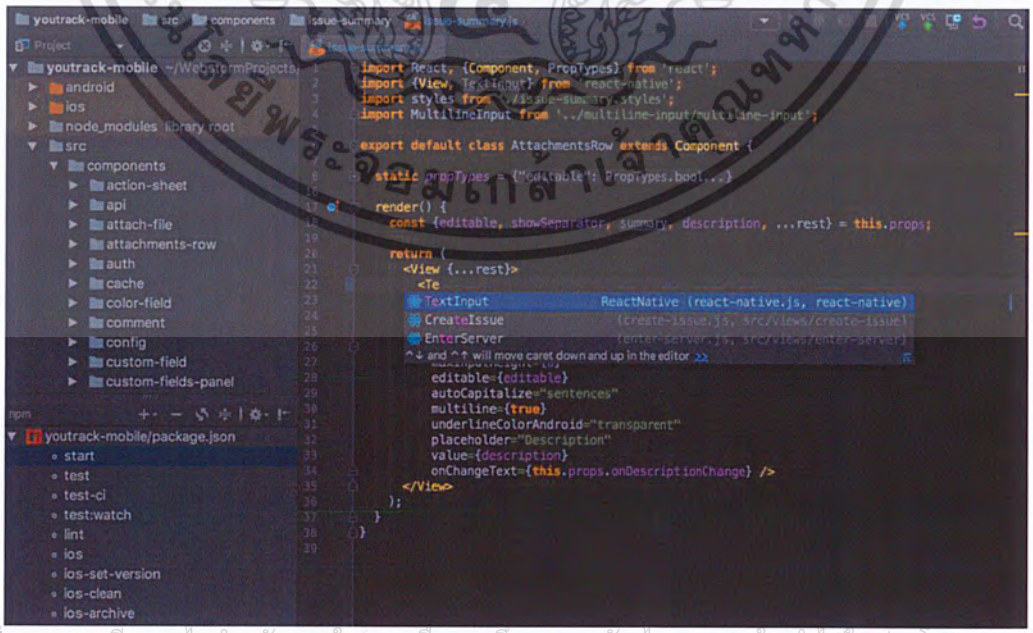
Putty [7] โปรแกรม Remote Server หรือ SSH (Secure Shell) เป็นโปรแกรม SSH client สำหรับการเชื่อมต่อกับเซิร์ฟเวอร์ ผ่านเน็ตเวิร์คภายในหรืออินเทอร์เน็ต ใช้งานในลักษณะสั่งงานเครื่องแม่ข่ายด้วย Command line ติดต่อกับเครื่องแม่ข่ายที่เป็นระบบปฏิบัติการลินุกซ์ เป็นโปรแกรมฟรีแวร์ มีขนาดเล็ก ใช้งานง่าย รองรับการเชื่อมต่อหลากหลายรูปแบบ เช่น Raw, Telnet, Rlogin, SSH, Serial ตัวอย่างการใช้งานโปรแกรม Putty แสดงดังรูปที่ 2.4



รูปที่ 2.4 ตัวอย่างการใช้งานของโปรแกรม Putty

2.8 JetBrains WebStorm

WebStrom [8] มีขนาดเล็ก เบาและมีประสิทธิภาพ WebStrom เป็นโปรแกรม Editor ที่ดีสำหรับการเขียนโปรแกรมที่ซับซ้อนฝั่งไคลเอ็นต์และการพัฒนาเครื่องแม่ข่ายกับ Node.js ซึ่งจะมี plugin ให้สามารถใช้ความสามารถเพิ่มเติมได้ มีประสิทธิภาพและการจัดโครงสร้างสำหรับการเขียนโปรแกรมภาษา JavaScript, TypeScript, stylesheet languages และ frameworks ที่เป็นที่นิยม ตัวอย่างการใช้งานโปรแกรม WebStrom แสดงดังรูปที่ 2.5



รูปที่ 2.5 ตัวอย่างการใช้งานของโปรแกรม WebStrom

2.9 Angular

Angular [9] เป็น Open Source Javascript Framework ที่ออกแบบ และพัฒนาโดยบริษัท Google มีความโดดเด่นในเรื่องของการจัดการโค้ดให้เป็นระบบ และสามารถทำงานได้อย่างรวดเร็วมาก เมื่อเทียบกับ Javascript Framework อื่นๆ และนอกจากนี้แล้ว AngularJS ยังมีความยืดหยุ่นที่สูงมาก ที่สามารถเขียนโค้ดในรูปแบบต่างๆ ได้อย่างง่ายดาย

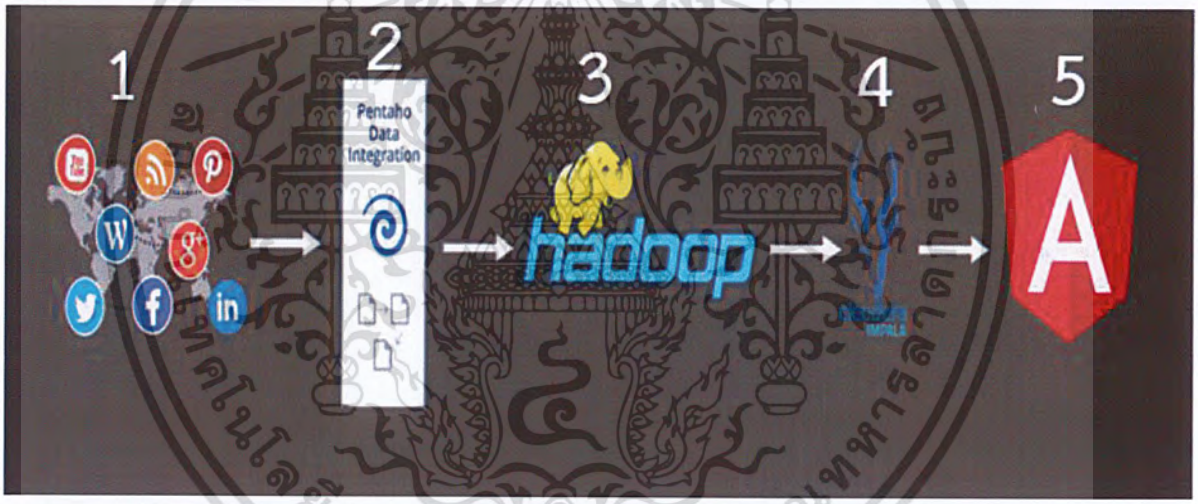
Angular 2 เป็นเฟรมเวิร์คของ Javascript ที่ทำงานฝั่งไคลเอนต์ (Client-side) หรือนำไปต่อยอดใช้สร้างแอปพลิเคชันบนโมบาย หรือ Desktop ได้ด้วย

บทที่ 3

วิธีการดำเนินงาน

3.1 สถาปัตยกรรมของการทำงาน

ระบบการดึงข้อมูลจาก social network แหล่งต่างๆ เช่น เฟซบุ๊ก (Facebook) ทวิตเตอร์ (Twitter) พันทิป (Pantip) และแหล่งข่าวต่างๆ จะใช้เครื่องมือ Pentaho ในการดึงข้อมูลออกมาและนำเข้าไปเก็บไว้ใน Hadoop แล้วค้นหาข้อมูลที่อยู่ใน Hadoop โดยผ่าน Impala เพื่อนำมาแสดงผลที่เว็บแอปพลิเคชันที่พัฒนาโดย Angular ระบบประกอบไปด้วยการทำงานหลักๆ ด้วยกัน 5 ส่วน ได้แก่ ส่วนศึกษาโครงสร้างข้อมูลของแหล่งต่างๆ ที่ต้องการ ส่วนการดึงข้อมูลด้วย Pentaho Data Integration ส่วนการจัดเก็บข้อมูลลงใน Hadoop ส่วนค้นหาข้อมูลและส่วนของการแสดงผลข้อมูลตามความต้องการ แสดงดังรูปที่ 3.1



รูปที่ 3.1 สถาปัตยกรรมของการทำงาน

จากรูปที่ 3.1 สามารถอธิบายส่วนการทำงานดังนี้

1. ส่วนศึกษาโครงสร้างข้อมูลของแหล่งต่างๆ

เป็นขั้นตอนที่ศึกษาวิธีการที่จะได้ข้อมูลจากแหล่งต่างๆ เช่น ข้อมูลจากเฟซบุ๊กต้องใช้งาน Facebook API ในการเข้าถึงข้อมูล โดยข้อมูลที่ได้จะอยู่ในรูปแบบ JSON หลังจากรู้โครงสร้างข้อมูลแล้วต้องรู้ด้วยว่าจะเก็บข้อมูลอะไรมาบ้าง เช่น เก็บข้อมูลทุกโพส ทุกคอมเม้นท์ จำนวนผู้กดถูกใจ แชร์ คอมเม้นท์

2. ส่วนการดึงข้อมูลด้วย Pentaho

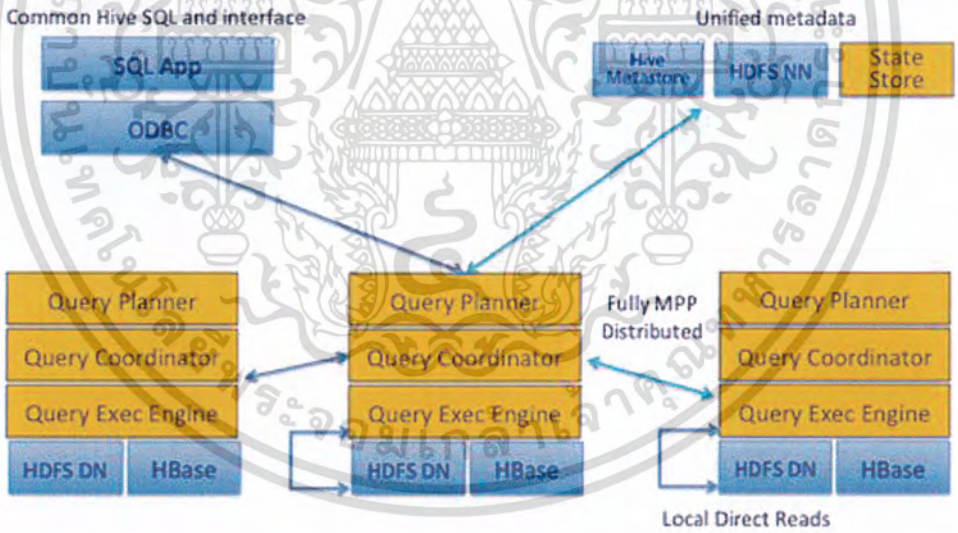
เป็นขั้นตอนที่ใช้ Pentaho ในการดึงข้อมูลจากเฟซบุ๊ก (Facebook) ทวิตเตอร์ (Twitter) พันทิป (Pantip) และเว็บข่าว (RSS) ต่างๆ โดยเครื่องมือแต่ละตัวและการกำหนดค่าสำหรับดึงข้อมูลจากแหล่งต่างๆจะมีการทำงานที่แตกต่างกัน

3. ส่วนการจัดเก็บข้อมูล

เป็นขั้นตอนที่จัดเก็บข้อมูลขนาดใหญ่ด้วย Hadoop มี HDFS ทำหน้าที่เป็นระบบจัดเก็บข้อมูลหลัก โดย HDFS จะสร้างแบบจำลองเป็นบล็อกของข้อมูลบนคลัสเตอร์เพื่อให้การคำนวณผลได้รวดเร็ว โดยการจัดเก็บข้อมูลอยู่ในรูปแบบไฟล์ CSV

4. ส่วนค้นหาข้อมูล

เป็นส่วนที่เขียนโปรแกรมให้บริการเว็บไซต์ แบบ REST ด้วยภาษาจาวา โดยเรียกใช้ผ่านทาง HTTP Method GET และส่งข้อมูลออกมาในรูปแบบของ JSON เชื่อมต่อกับ Impala โดยใช้ JDBC Driver เมื่อเชื่อมต่อกับ Impala แสดงดังรูปที่ 3.2



รูปที่ 3.2 กระบวนการทำงานของ Cloudera Impala

จากรูปที่ 3.2 สามารถอธิบายการทำงานได้ดังนี้

- ส่งคำสั่ง SQL ที่ใช้ค้นหาเข้ามา Impala โดยจะมี Impala Statestone บอก Query Planner ว่าข้อมูลแต่ละที่อยู่ที่ใด

- Query Planner จะจัดการคำสั่ง SQL ที่เข้ามาว่าร้องขอข้อมูลอะไร แล้วจะส่งข้อมูลให้ Query Coordinater ให้ส่งคำสั่งค้นหาไปยัง Query Coordinater ของแต่ละโหนด
- เมื่อแต่ละโหนดได้ข้อมูลแล้วจะส่งข้อมูลกลับมาที่ Query Coordinater และส่งกลับให้โปรแกรมให้บริการเว็บไซต์ แสดงข้อมูลในรูปแบบ JSON

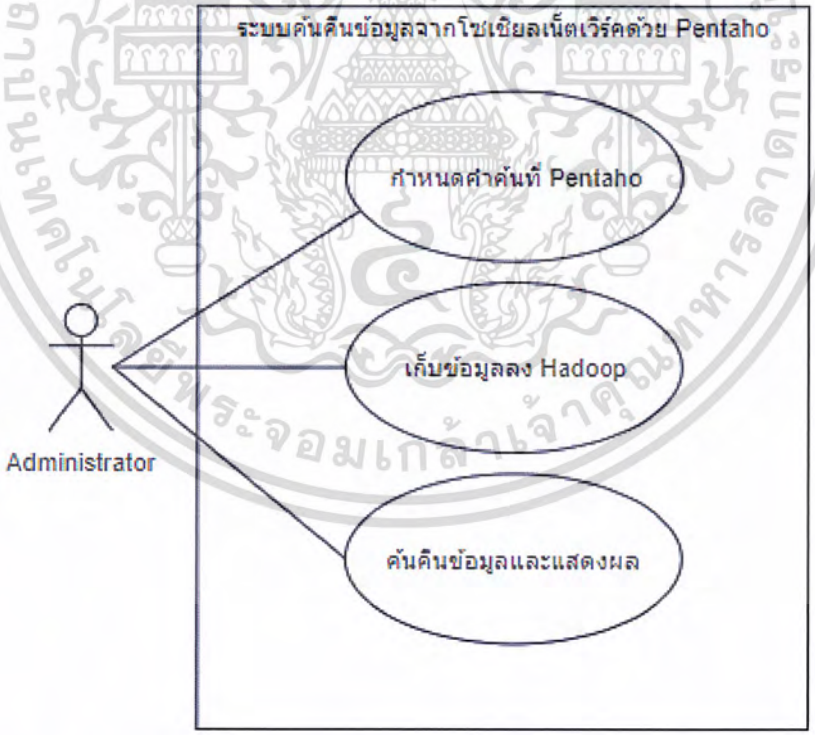
5. ส่วนการแสดงผล

เป็นส่วนที่สร้างเว็บแอปพลิเคชันเพื่อใช้ในการแสดงผลกราฟในรูปแบบต่างๆ จากข้อมูลที่ได้ทำการเก็บมา พัฒนาโดยใช้ Angular

3.2 ความสามารถของระบบ

3.2.1 แผนภาพยูสเคส (Use Case Diagram)

แผนภาพยูสเคสแสดงความสามารถของระบบค้นคืนข้อมูลจากโซเชียลเน็ตเวิร์คด้วย Pentaho แสดงดังรูปที่ 3.3



รูปที่ 3.3 Use Case Diagram ของระบบค้นคืนข้อมูลจากโซเชียลเน็ตเวิร์คด้วย Pentaho

จากรูปที่ 3.3 เป็นการแสดงแผนภาพยูสเคสของระบบค้นคืนข้อมูลจากโซเชียลเน็ตเวิร์คด้วย Pentaho โดยรายละเอียดยูสเคสแสดงในตารางที่ 3.1 ถึง 3.3 ดังต่อไปนี้

ตารางที่ 3.1 กำหนดคำที่ต้องการค้นหา

หัวข้อ	**คำอธิบาย**
Use Casce Name	กำหนดคำที่ต้องการค้นหา
Brief Description	กำหนดคำที่ใช้ในการค้นหา
Actor	Admin
Pre - Conditions	กรอกคำค้นหา
Post - Conditions	แสดงข้อความที่มีคำค้นหา
Flow of Event	1) กรอกคำที่ต้องการค้นหา 2) ข้อมูลถูกส่งไปค้นหาด้วย API 3) แสดงผลข้อความที่มีคำค้นหา
Exception	

ตารางที่ 3.2 บันทึกข้อมูลลง Hadoop

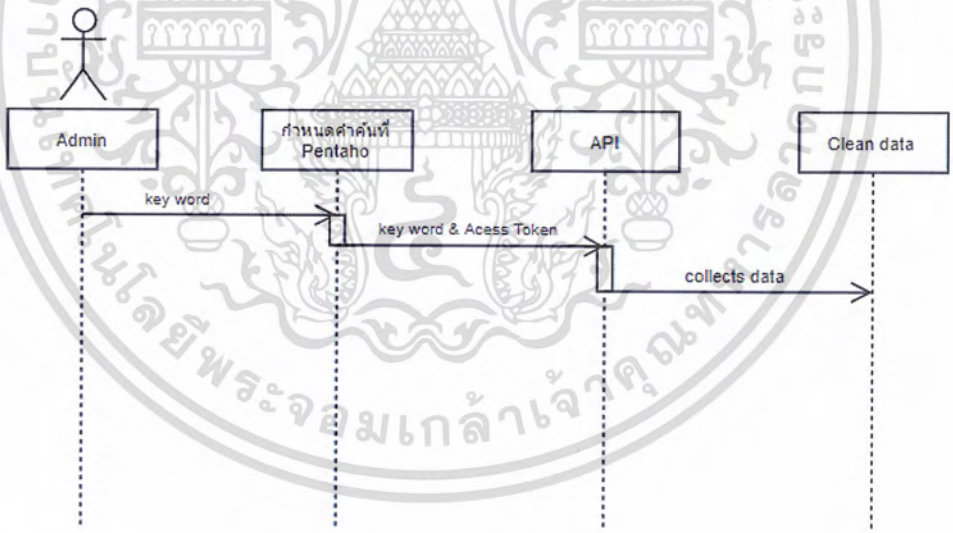
หัวข้อ	**คำอธิบาย**
Use Casce Name	บันทึกข้อมูลลง Hadoop
Brief Description	บันทึกข้อมูลลง Hadoop
Actor	Admin
Pre - Conditions	ข้อมูลที่ได้จากการค้นหา
Post - Conditions	แสดงข้อมูลที่บันทึก
Flow of Event	1) ข้อความที่ได้จากการค้นหา 2) กำหนดที่อยู่ของไฟล์ใน Hadoop ที่ต้องการบันทึก
Exception	

ตารางที่ 3.3 ค้นหาข้อมูลใน Hadoop และแสดงผล

หัวข้อ	**คำอธิบาย**
Use Casce Name	ค้นหาข้อมูลใน Hadoop และแสดงผล
Brief Description	ค้นหาข้อมูลใน Hadoop
Actor	Admin
Pre - Conditions	กรอกคำสั่ง SQL ในการค้นหา
Post - Conditions	แสดงข้อมูลที่ต้องการ
Flow of Event	1) กรอกคำสั่ง SQL 2) ข้อมูลที่ได้สร้างเป็นกราฟ 3) แสดงผล
Exception	

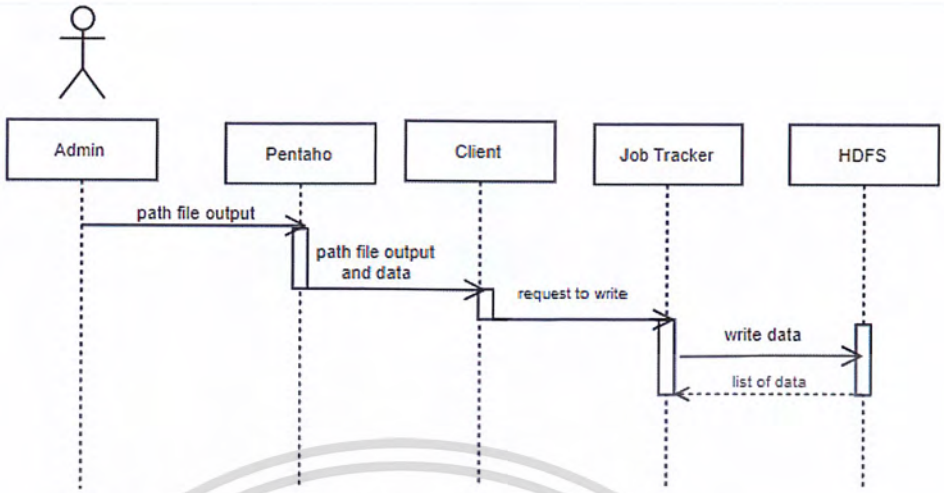
3.2.2 แผนภาพซีเควนซ์ไดอะแกรม (Sequence Diagram)

เป็นแผนภาพ Sequence Diagram การทำงานของระบบ แสดงดังรูปที่ 3.4 ถึง 3.6 ดังต่อไปนี้



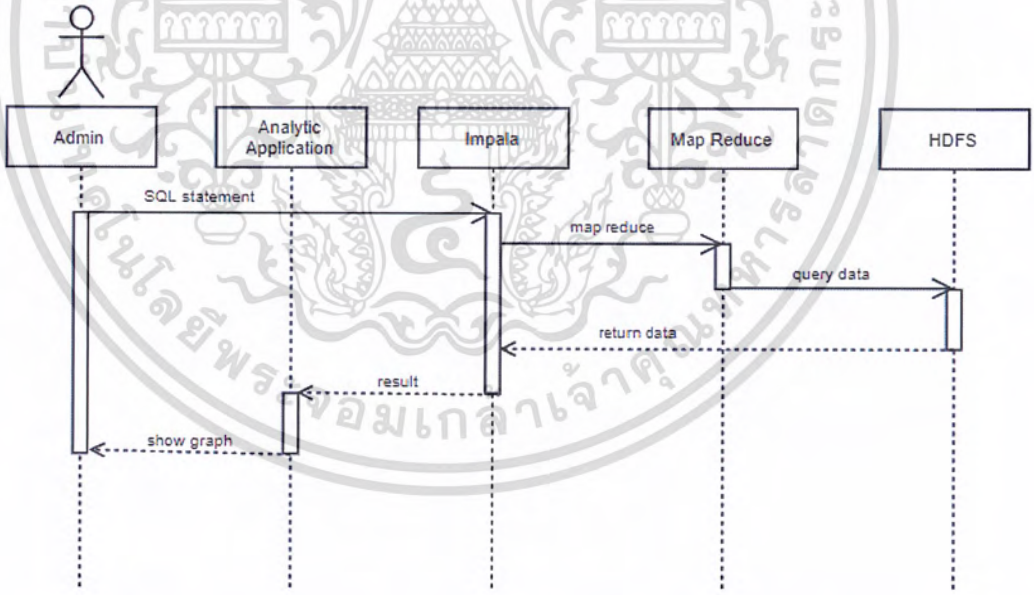
รูปที่ 3.4 Sequence Diagram ของขั้นตอนกำหนดคำค้นหาด้วย Pentaho

จากรูปที่ 3.4 แสดงถึงภาพรวมการทำงานของขั้นตอนกำหนดคำค้นหาด้วย Pentaho โดยผู้ดูแลระบบจะทำการกำหนดคำค้นในเครื่องมือของ Pentaho แล้วเข้าถึงข้อมูลจากโซเชียลเน็ตเวิร์คผ่าน API หลังจากนั้นจะมีการกำหนดเครื่องมือที่ใช้ในการทำความสะอาดข้อมูลให้อยู่ในรูปแบบที่สามารถนำไปใช้งานได้ง่าย



รูปที่ 3.5 Sequence Diagram ของขั้นตอนบันทึกข้อมูลลง Hadoop

จากรูปที่ 3.5 แสดงถึงภาพรวมการทำงานของขั้นตอนบันทึกข้อมูลลง Hadoop โดยหลังจากที่ Pentaho ได้ค้นหาข้อมูลมาแล้ว จะมีการส่งข้อมูลและที่อยู่ของข้อมูลที่ต้องการจัดเก็บลง Hadoop



รูปที่ 3.6 Sequence Diagram ของขั้นตอนการค้นหาข้อมูลและแสดงผล

จากรูปที่ 3.6 แสดงถึงภาพรวมการทำงานของขั้นตอนการค้นหาข้อมูลและแสดงผล โดยผู้ดูแลระบบจะใช้คำสั่ง SQL ในการค้นหาข้อมูล หลังจากนั้นจะนำข้อมูลที่ได้แสดงในกราฟรูปแบบต่างๆ

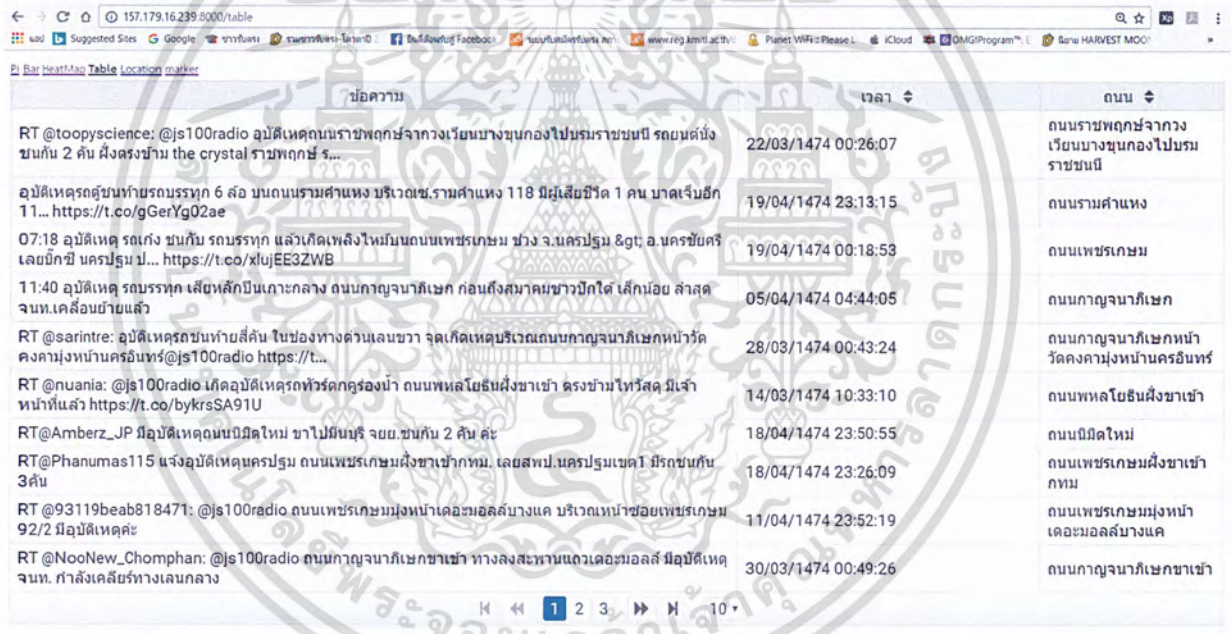
บทที่ 4

ผลการดำเนินงานและการอภิปรายผล

4.1 การแสดงผลโดยกราฟรูปแบบต่างๆ

หลังจากได้ทำการศึกษา เรียนรู้ การใช้เทคโนโลยีการจัดการข้อมูลขนาดใหญ่ การเก็บข้อมูลจากโซเชียลเน็ตเวิร์คและการทำเว็บแอปพลิเคชันที่พัฒนาโดย Angular ในการนำข้อมูลที่จัดเก็บใน Hadoop มาแสดงผลโดยกราฟรูปแบบต่างๆ แสดงได้ดังรูปที่ 4.1 ถึงรูปที่ 4.5

1) ตารางแสดงรายละเอียดของข้อมูลที่ค้นหาจากข้อความ คำว่า อุบัติเหตุ โดยแสดงรายละเอียดของข้อความที่โพส เวลาที่โพส และถนนที่เกิดเหตุ แสดงดังรูปที่ 4.1

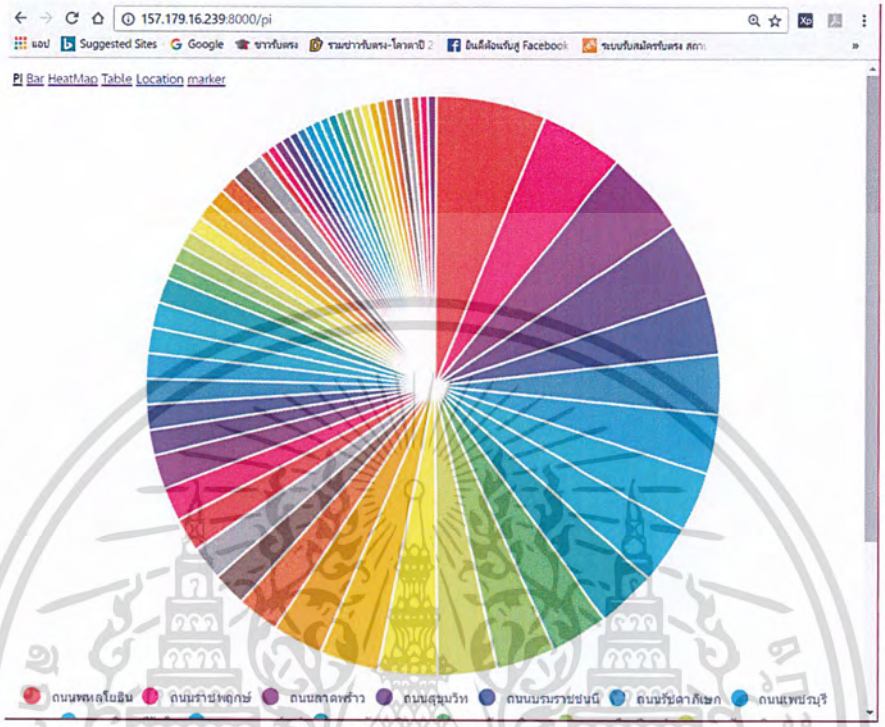


ข้อความ	เวลา	ถนน
RT @toopyscience: @js100radio อุบัติเหตุถนนราชพฤกษ์จากวงเวียนบางขุนกองไปบรมราชชนนี รถยนต์นั่งชนกัน 2 คัน ฝั่งตรงข้าม the crystal ราชพฤกษ์ ร...	22/03/1474 00:26:07	ถนนราชพฤกษ์จากวงเวียนบางขุนกองไปบรมราชชนนี
อุบัติเหตุรถตู้ชนท้ายรถบรรทุก 6 ล้อ บนถนนรามคำแหง บริเวณซ.รามคำแหง 118 มีผู้เสียชีวิต 1 คน บาดเจ็บอีก 11... https://t.co/gGerYg02ae	19/04/1474 23:13:15	ถนนรามคำแหง
07:18 อุบัติเหตุ รถเก๋ง ชนกับ รถบรรทุก แล้วเกิดเพลิงไหม้บนถนนเพชรเกษม ช่วง จ.นครปฐม > อ.นครชัยศรี เลยบิ๊กซี นครปฐม ป... https://t.co/xlujEE3ZWB	19/04/1474 00:18:53	ถนนเพชรเกษม
11:40 อุบัติเหตุ รถบรรทุก เสียหลักบีนเกาะกลาง ถนนกาญจนาภิเษก ก่อนถึงสมาคมชาวปักใต้ เล็กน้อย ล่าสุด จนท.เคลื่อนย้ายแล้ว	05/04/1474 04:44:05	ถนนกาญจนาภิเษก
RT @sarintre: อุบัติเหตุรถชนท้ายสี่คัน ในช่องทางด่วนเลนขวา จุดเกิดเหตุบริเวณถนนกาญจนาภิเษกหน้าวัดคงคามุ่งหน้านครอินทร์@js100radio https://t...	28/03/1474 00:43:24	ถนนกาญจนาภิเษกหน้าวัดคงคามุ่งหน้านครอินทร์
RT @nuania: @js100radio เกิดอุบัติเหตุรถทัวร์ดกคูร้องน้ำ ถนนพหลโยธินฝั่งขาเข้า ตรงข้ามไทวัสดุ มีเจ้าหน้าที่แล้ว https://t.co/bykrsSA91U	14/03/1474 10:33:10	ถนนพหลโยธินฝั่งขาเข้า
RT@Amberz_JP มีอุบัติเหตุถนนนิมิตใหม่ ขาไปมีนบุรี จยย.ชนกัน 2 คัน ค่ะ	18/04/1474 23:50:55	ถนนนิมิตใหม่
RT@Phanumas115 แจ้งอุบัติเหตุนครปฐม ถนนเพชรเกษมฝั่งขาเข้ากทม. เลยสพป.นครปฐมเขต1 มีรถชนกัน 3คัน	18/04/1474 23:26:09	ถนนเพชรเกษมฝั่งขาเข้ากทม
RT @93119beab818471: @js100radio ถนนเพชรเกษมมุ่งหน้าเดอะมอลล์บางแค บริเวณหน้าซอยเพชรเกษม 92/2 มีอุบัติเหตุค่ะ	11/04/1474 23:52:19	ถนนเพชรเกษมมุ่งหน้าเดอะมอลล์บางแค
RT @NooNew_Chomphan: @js100radio ถนนกาญจนาภิเษกขาเข้า ทางลงสะพานแถวเดอะมอลล์ มีอุบัติเหตุ จนท. กำลังเคลียร์ทางเลนกลาง	30/03/1474 00:49:26	ถนนกาญจนาภิเษกขาเข้า

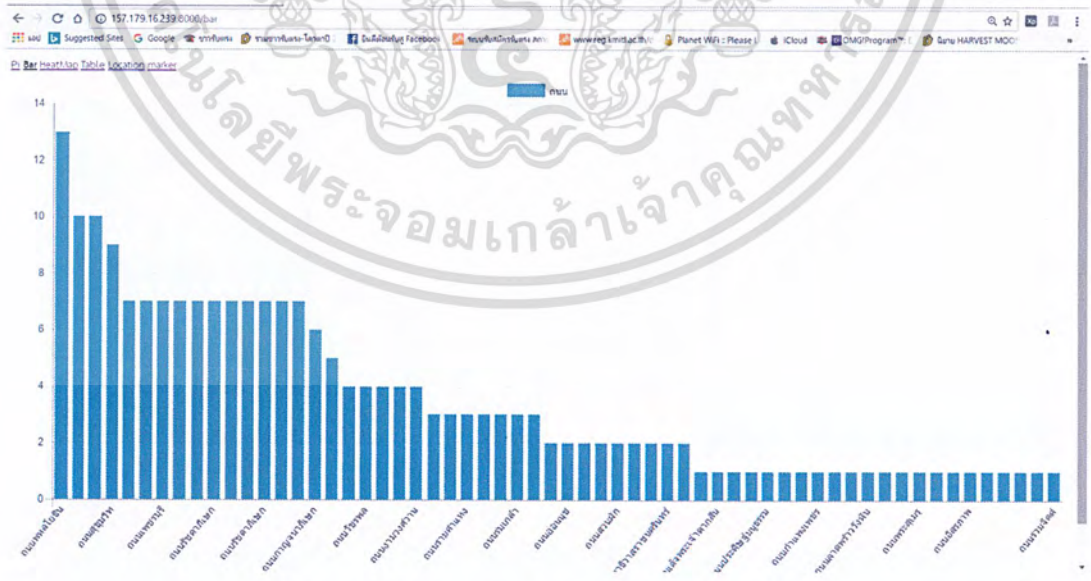
1 2 3 10

รูปที่ 4.1 ตารางแสดงข้อความ เวลา ถนน ที่มีการเกิดอุบัติเหตุ

2) กราฟวงกลมและกราฟแท่งแสดงจำนวนของถนนที่ค้นหาจากข้อความ คำว่า อุบัติเหตุ แสดงดังรูปที่ 4.2 และ 4.3

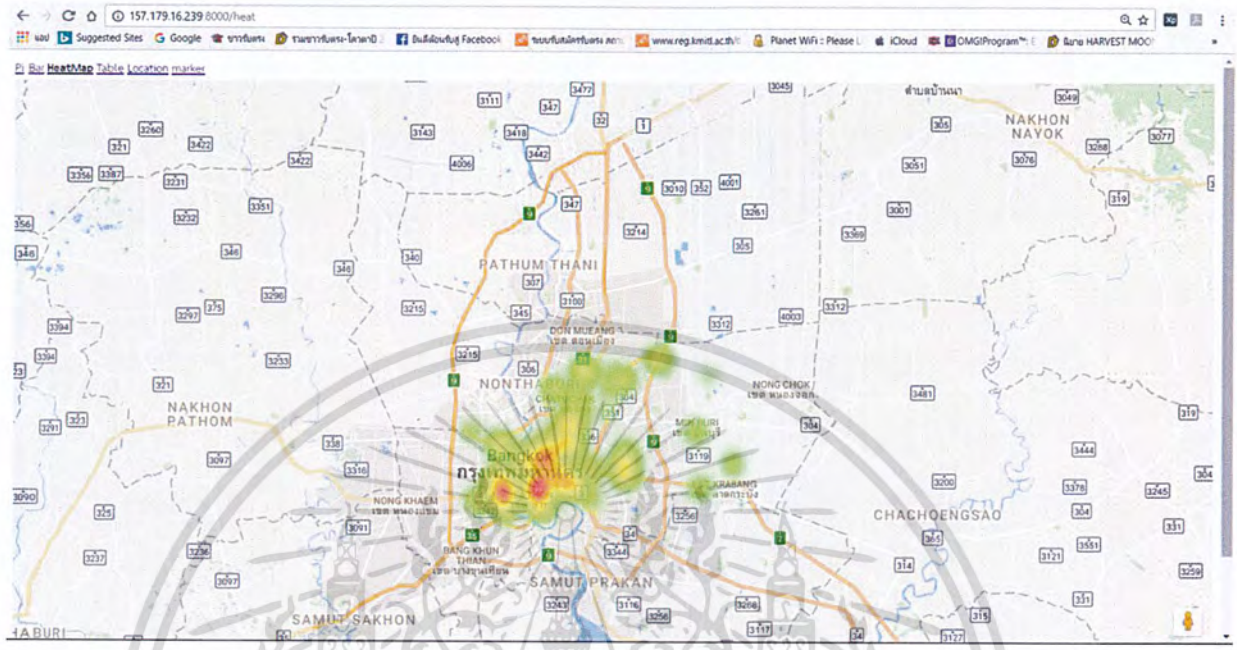


รูปที่ 4.2 กราฟวงกลมแสดงจำนวนอุบัติเหตุที่เกิดบนถนน



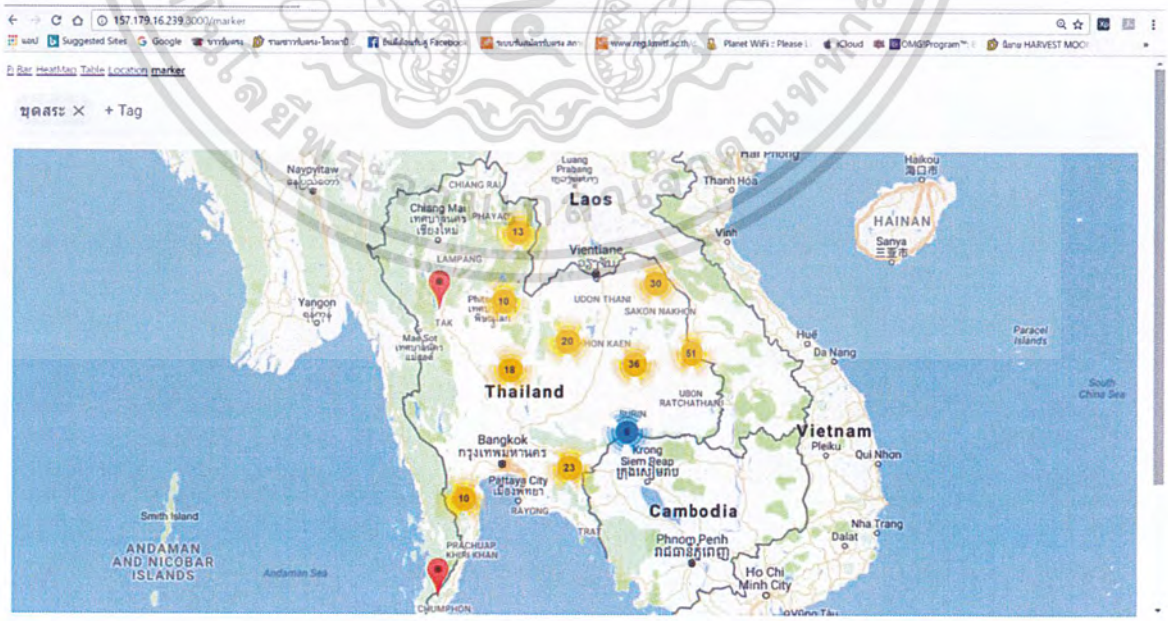
รูปที่ 4.3 กราฟแท่งแสดงจำนวนอุบัติเหตุที่เกิดบนถนน

3) Heat map แสดงความหนาแน่นของถนนที่ค้นหาจากข้อความ คำว่า อุบัติเหตุ โดย สีเขียว แสดงถึงเกิดน้อย สีแดงแสดงถึงเกิดมาก แสดงดังรูปที่ 4.4



รูปที่ 4.4 Heat map แสดงความหนาแน่นของการเกิดข้อมูล

4) Cluster map แสดงกลุ่มความหนาแน่นของข้อมูลที่สามารถระบุคำที่ต้องการค้นหา โดยจัดกลุ่มที่ใกล้เคียงเข้าด้วย แสดงถึงกลุ่มการเกิดของข้อมูลมาก น้อย แสดงดังรูปที่ 4.5

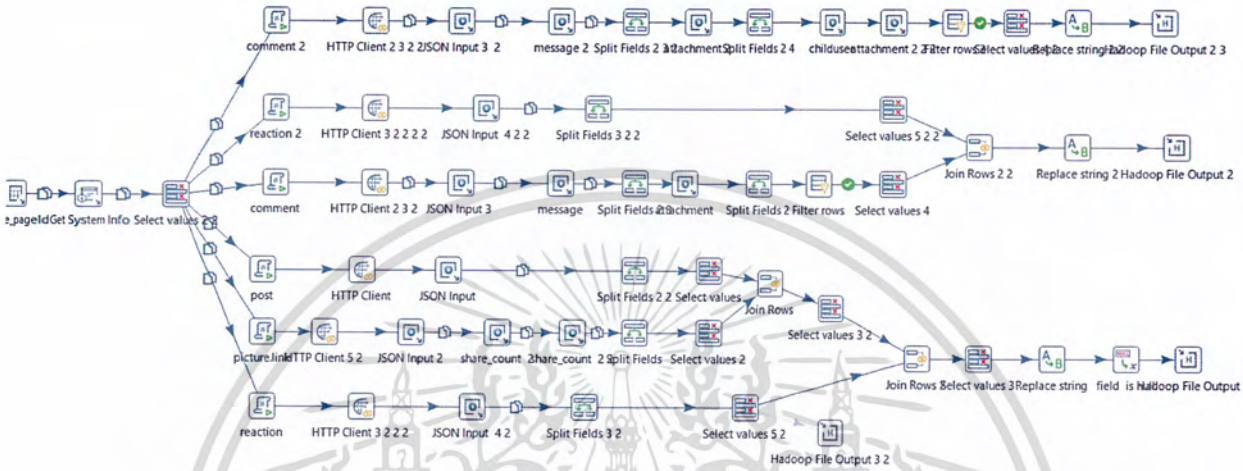


รูปที่ 4.5 Cluster map แสดงกลุ่มความหนาแน่นของการเกิดข้อมูล

4.2 การดำเนินงานของส่วนดึงข้อมูลด้วย Pentaho

วิธีการดำเนินงานในขั้นตอนดึงข้อมูลด้วย Pentaho โดยมีการกำหนดค่าตัวแปรให้แต่ละเครื่องมือทำงานสำหรับดึงข้อมูลจากแหล่งที่ต่างกัน

4.2.1 การทำงานของส่วนการนำข้อมูลจาก Facebook โดย API



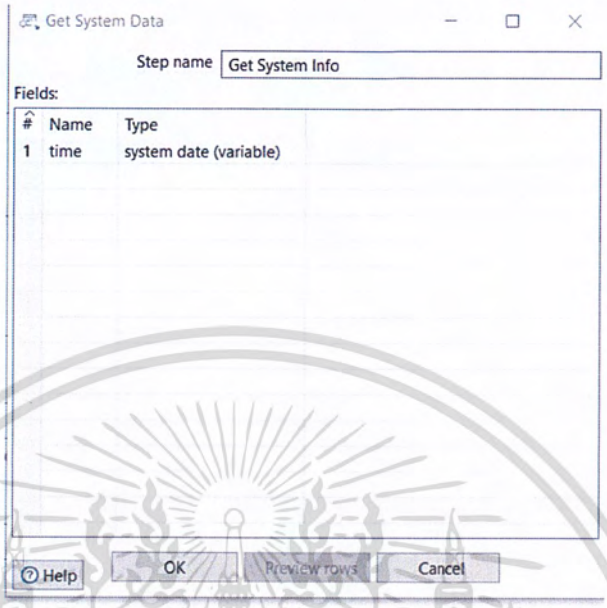
รูปที่ 4.6 การทำงานของส่วนการนำข้อมูลจาก Facebook โดย API

จากรูปที่ 4.6 สามารถอธิบายส่วนของการทำงานดังนี้

1) เขียนคำสั่ง SQL เพื่อค้นหา Id page ที่เก็บอยู่ในฐานข้อมูล แสดงดังรูปที่ 4.7

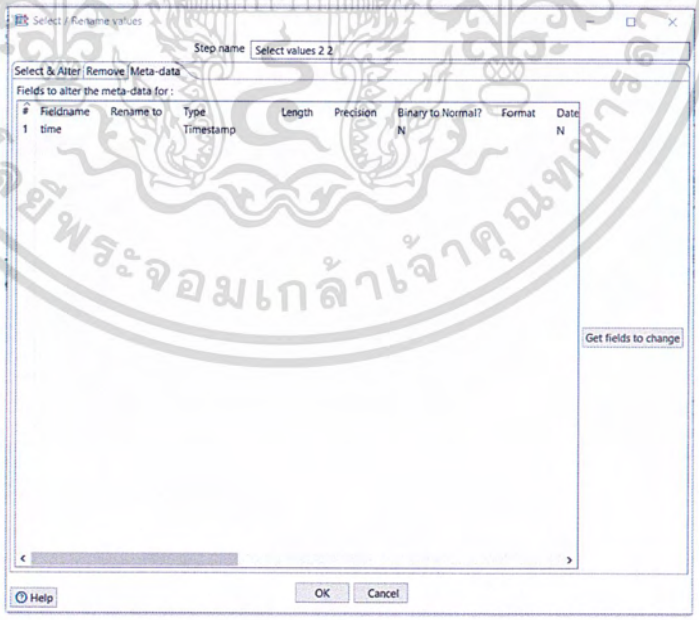
รูปที่ 4.7 คำสั่ง SQL ค้นหา Id page ที่เก็บอยู่ในฐานข้อมูล

2) get system into รับค่าวันที่ตัวแปร Time คือเวลาที่ทำการดึงข้อมูลไว้ใช้ในการเทียบข้อมูลล่าสุด แสดงดังรูปที่ 4.8



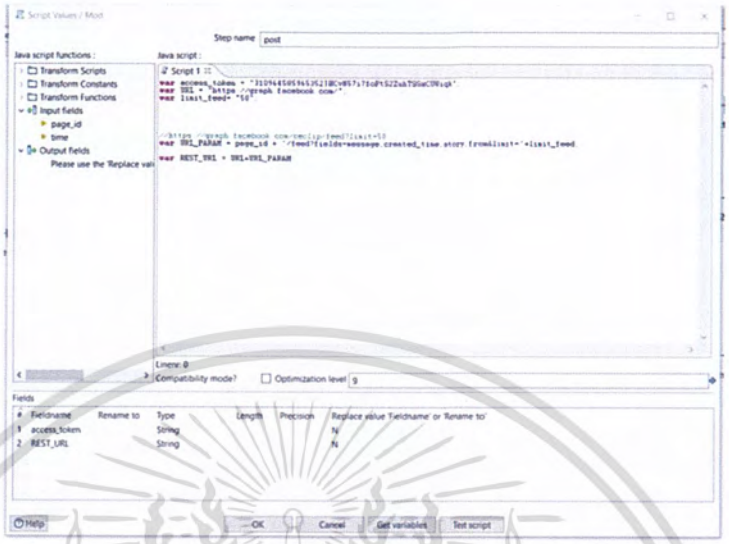
รูปที่ 4.8 กำหนดค่าวันที่ตัวแปร Time

3) select value เพื่อเลือก field ที่ต้องการ แสดงดังรูปที่ 4.9



รูปที่ 4.9 เลือก field ที่ต้องการ

4) เขียน javascript ในการกำหนด URL ที่ใช้เรียกตาม Facebook API โดยกำหนด Accresstoken สำหรับการเรียกใช้ Fackbook API แสดงดังรูปที่ 4.10



รูปที่ 4.10 คำสั่ง javascript ในการกำหนด URL ที่ใช้เรียกตาม Facebook API

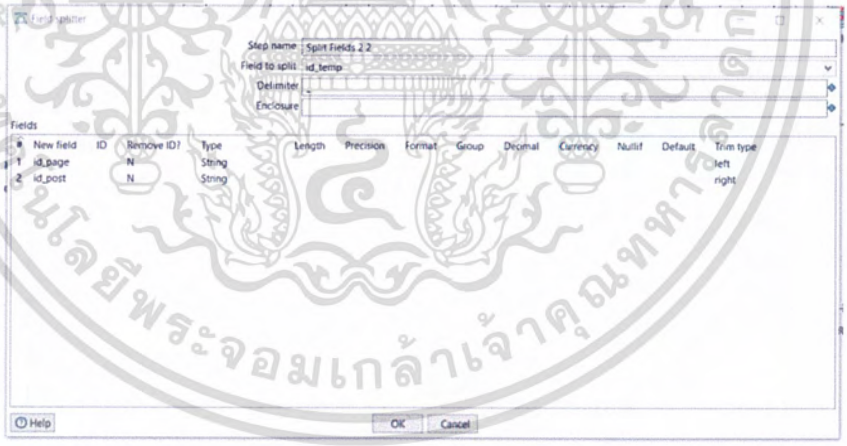
5) http client จะทำการส่ง request ตาม URL ที่รับมา การทำงานต้องรับ Parameter accress_token และ Rest_URL ซึ่งจะเป็นการเรียก Rest_URL&accress_token เช่น https://graph.facebook.com/v2.9/me?fields=id%2Cname&access_token=3134731157144999%7CsmWttLSDQymOlY4ECyHkOlL1qFA แสดงดังรูปที่ 4.11

รูปที่ 4.11 http client จะทำการส่ง request ตาม URL ที่รับมา

6) JSON input เมื่อส่ง request ไป ข้อมูลที่ได้จะอยู่ในรูปแบบ JSON ซึ่งต้องใช้ JSON input ในการอ่านข้อมูล โดยกำหนด path ของข้อมูลนั้นๆ แสดงดังรูปที่ 4.12

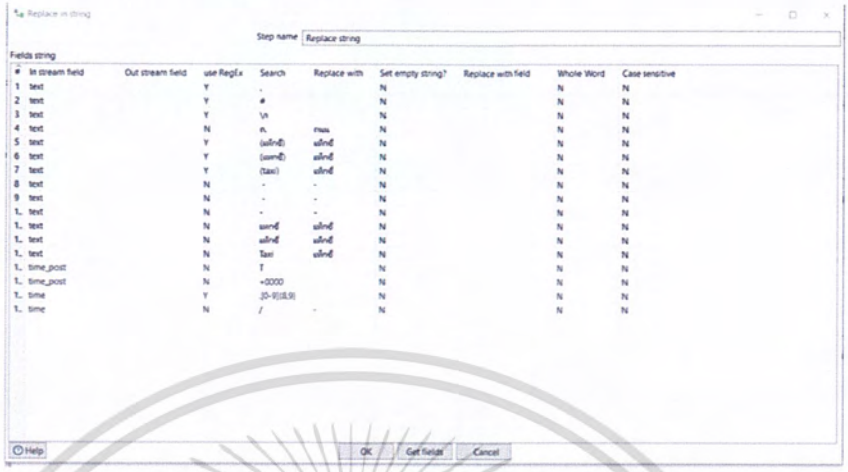
รูปที่ 4.12 เมื่อส่ง request ไป ข้อมูลที่ได้จะอยู่ในรูปแบบ JSON

7) spilt filed ใช้ในการแบ่ง filed เช่นข้อมูล id ประกอบด้วย id page ตามด้วย id post ดังนั้นต้องแยก field เป็น id_page กับ id_post เพื่อใช้ id_post join กับ table อื่นๆ แสดงดังรูปที่ 4.13



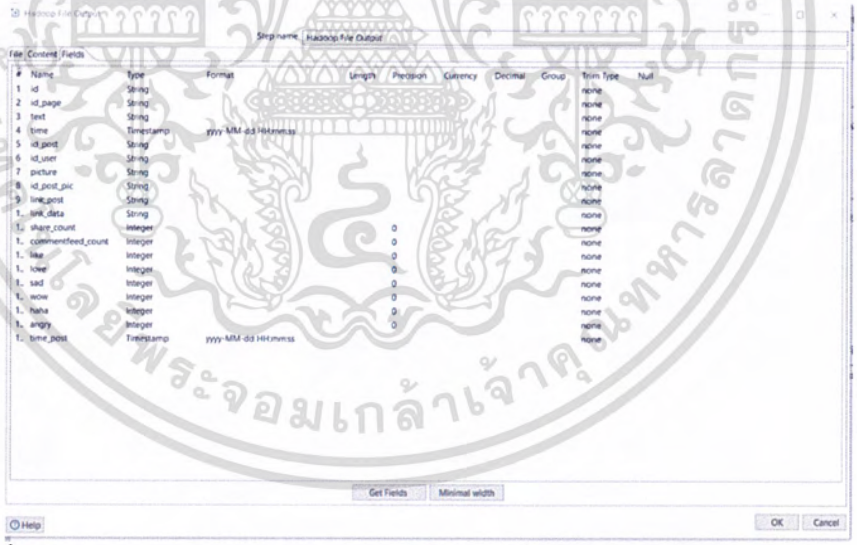
รูปที่ 4.13 spilt filed แยกเป็น id_page และ id_post

8) replace string ใช้ในการจัดรูปแบบของคำให้ตรงตามความต้องการเช่น ค้นหาคำว่า แท็กซี่ ให้แทนที่ด้วย แท็กซี่ แสดงดังรูปที่ 4.14



รูปที่ 4.14 replace string ใช้จัดรูปแบบของคำ

9) hadoop file output ใช้ในการระบุที่อยู่ของ file ปลายทางใน Hadoop และระบุ field ที่ต้องการ แสดงดังรูปที่ 4.15



รูปที่ 4.15 hadoop file output ใช้ในการระบุที่อยู่ของ file ปลายทางใน Hadoop

4.2.2 การทำงานของส่วนการนำข้อมูลจาก Twitter โดย API

รูปที่ 4.16 การทำงานของส่วนการนำข้อมูลจาก Twitter โดย API

จากรูปที่ 4.16 สามารถอธิบายส่วนของการทำงานดังนี้

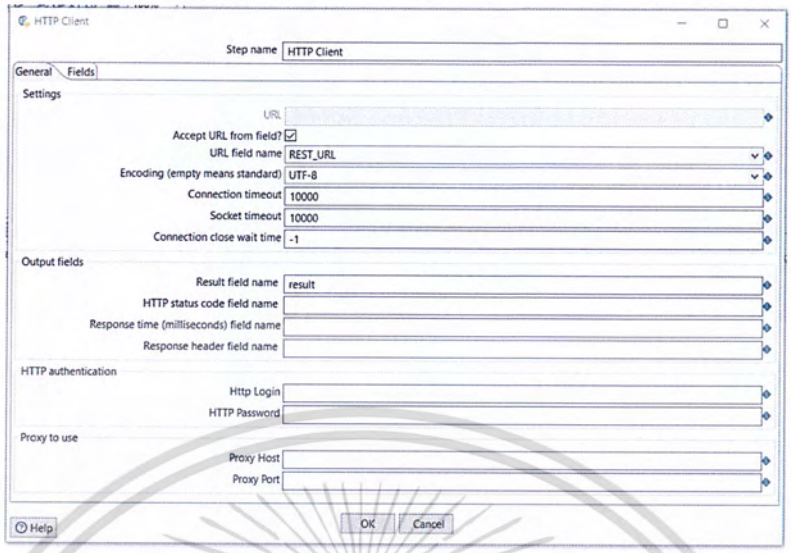
1) table searchWord ใช้ในการเขียนคำสั่ง SQL ค้นหาคำที่เก็บอยู่ในฐานข้อมูล แสดงดังรูปที่ 4.17

รูปที่ 4.17 เขียนคำสั่ง SQL ค้นหาคำที่เก็บอยู่ในฐานข้อมูล

2) เขียนคำสั่ง javascript ในการกำหนด URL ที่ใช้เรียกตาม Twitter API โดยกำหนด Accresstoken สำหรับการเรียกใช้ Twitter API แสดงดังรูปที่ 4.18

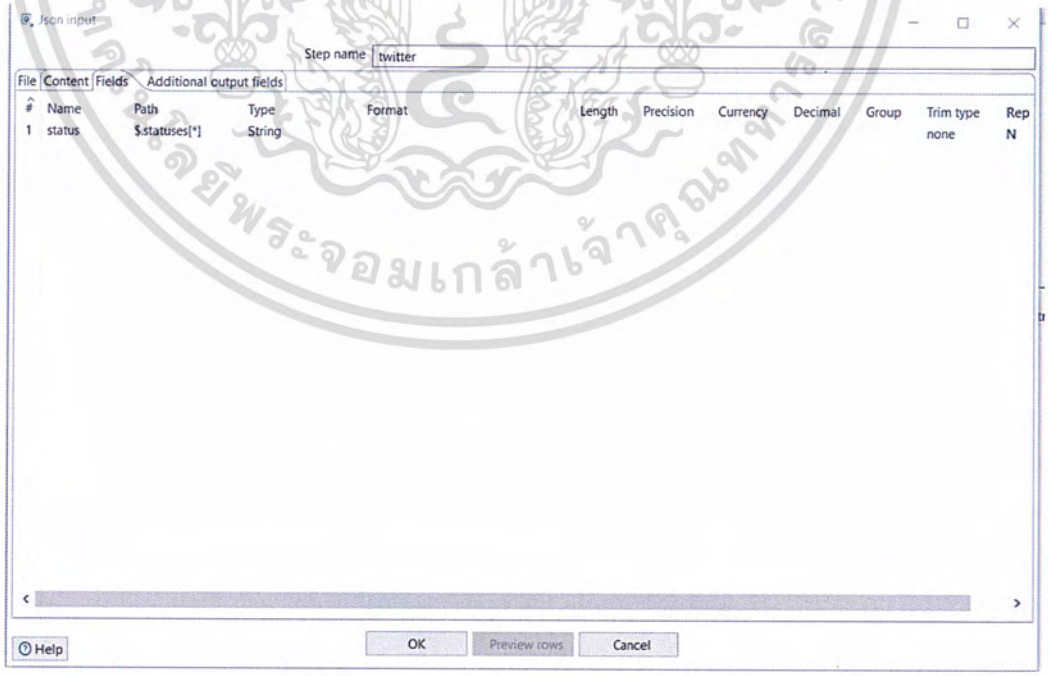
รูปที่ 4.18 คำสั่ง javascript ในการกำหนด URL ที่ใช้เรียกตาม Twitter API

3) http client จะทำการส่ง request ตาม URL ที่รับมา แสดงดังรูปที่ 4.19

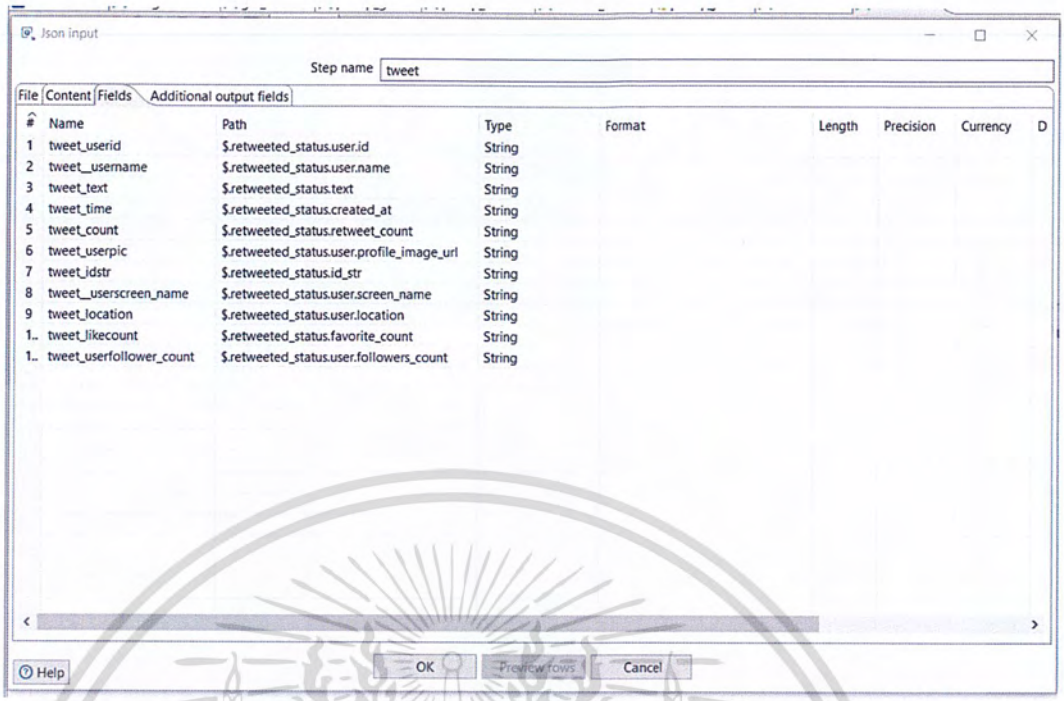


รูปที่ 4.19 http client จะทำการส่ง request ตาม URL ที่รับมา

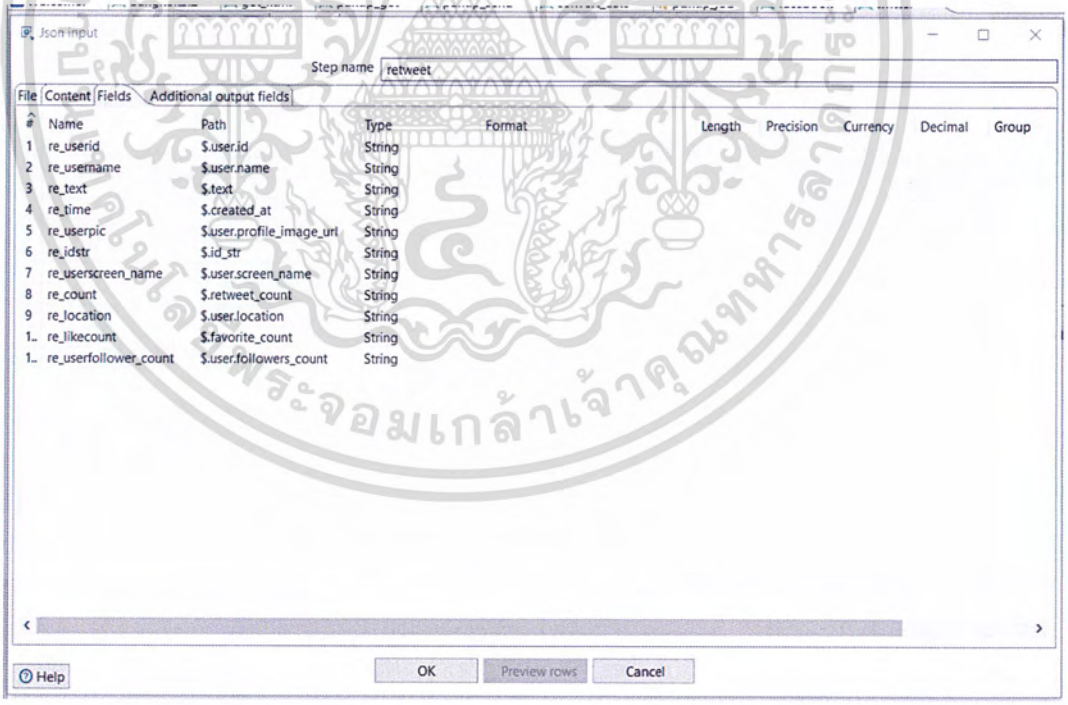
4) JSON input เมื่อส่ง request ข้อมูลที่ได้จะอยู่ในรูปแบบ JSON ซึ่งต้องใช้ JSON input ในการอ่านข้อมูล เนื่องจากมี array ของ object สองตัวคือ tweet กับ retweet จึงต้องทำการเรียก object ตัว parent ก่อน แล้วใช้ JSON input อ่านข้อมูลของ child แสดงดังรูปที่ 4.20 ถึง 4.22



รูปที่ 4.20 เรียก path ของ parent เพื่อให้ได้ child

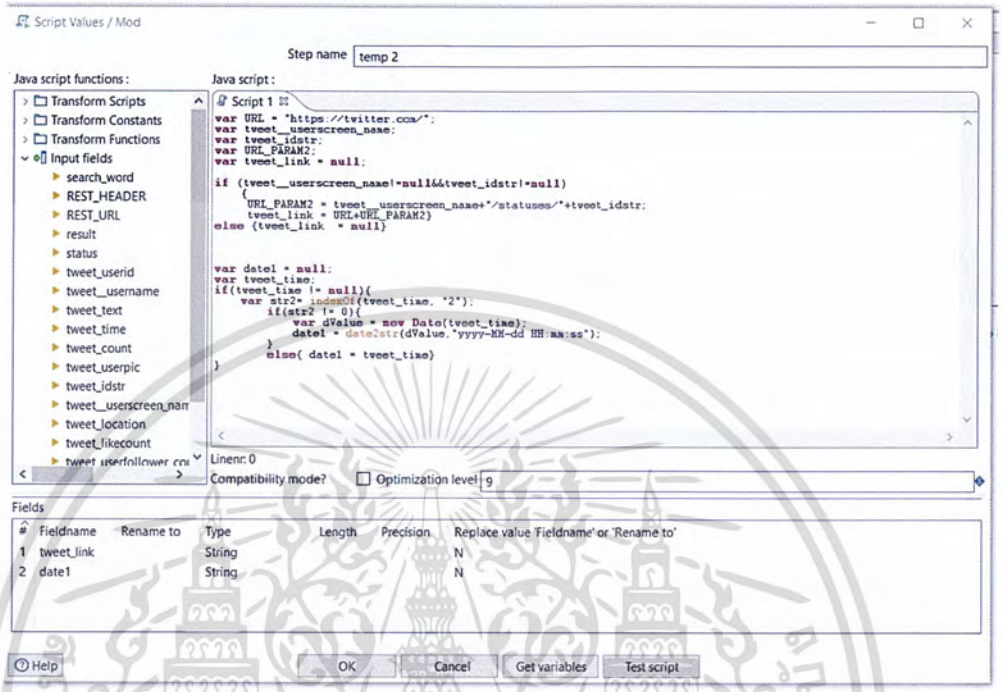


รูปที่ 4.21 เรียก path ของ child object tweet



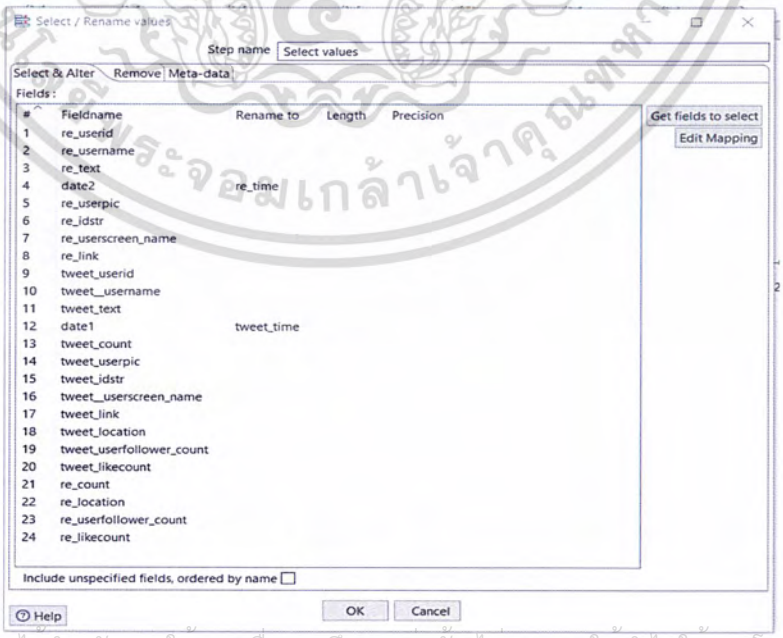
รูปที่ 4.22 เรียก path ของ child object Retweet

5) เขียน javascript ในการแปลงรูปแบบของวันที่ให้อยู่ในรูป 2017-06-19 12:22:03 และเพื่อกำหนด URL ของหน้าที่ tweet ข้อความเช่น https://twitter.com/user_screnname /statuses/tweet_idstr แสดงดังรูปที่ 4.23



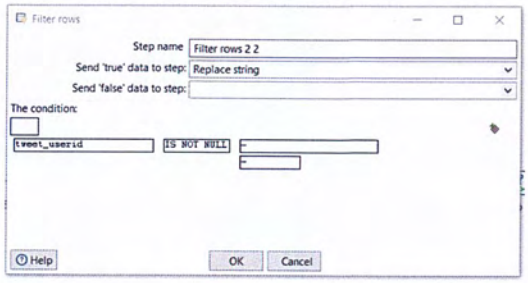
รูปที่ 4.23 คำสั่ง javascript ในการแปลงรูปแบบของวันที่และกำหนด URL

6) select value เพื่อเลือก field ที่ต้องการ แสดงดังรูปที่ 4.24



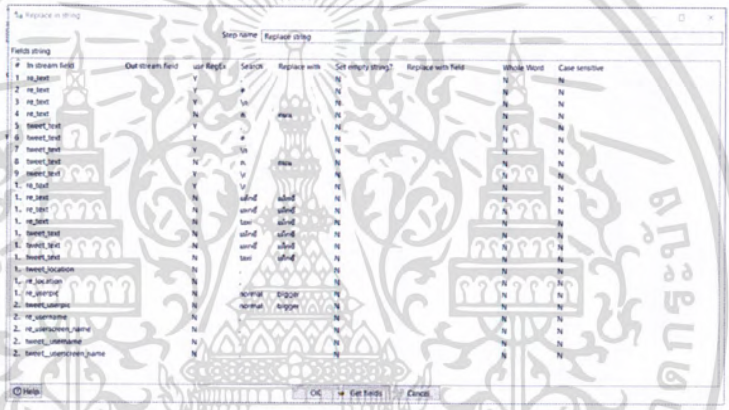
รูปที่ 4.24 select value เพื่อเลือก field ที่ต้องการ

7) filter row ใช้กรอง row ที่ไม่มีค่า แสดงดังรูปที่ 4.25



รูปที่ 4.25 filter row ใช้กรอง row ที่ไม่มีค่า

8) replace string ใช้ในการจัดรูปแบบของคำให้ตรงตามความต้องการเช่น ค้นหาคำว่า แท็กซี่ ให้แทนที่ด้วย แท็กซี่ แสดงดังรูปที่ 4.26



รูปที่ 4.26 replace string จัดรูปแบบของคำ

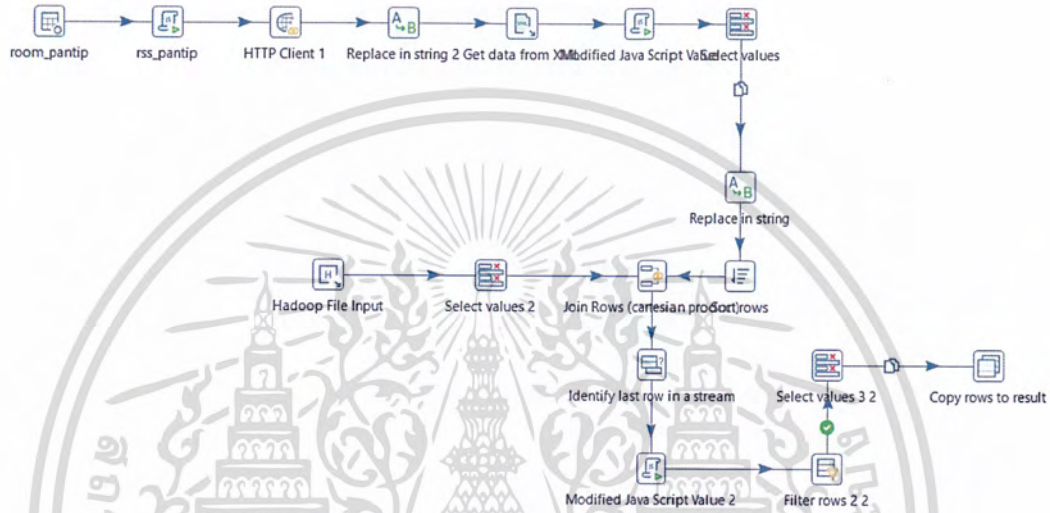
9) hadoop file output ใช้ในการระบุที่อยู่ของ file ปลายทางและระบุ field ที่ต้องการ แสดงดังรูปที่ 4.27

รูปที่ 4.27 hadoop file output กำหนดที่อยู่ของ file ปลายทางและ field ที่ต้องการ

4.2.3 การทำงานของส่วนการนำข้อมูลจาก Pantip โดย RSS

การะบวนการทำงานของส่วนการนำข้อมูลจาก Pantip โดย RSS จะเก็บข้อมูลแบบไม่ซ้ำกันทำให้ต้องแยกการทำงานออกเป็นสองส่วนคือส่วนการนำข้อมูลจาก RSS และส่วนการนำข้อมูลเข้า Hadoop ซึ่งจะมีการทำงานเป็น Job

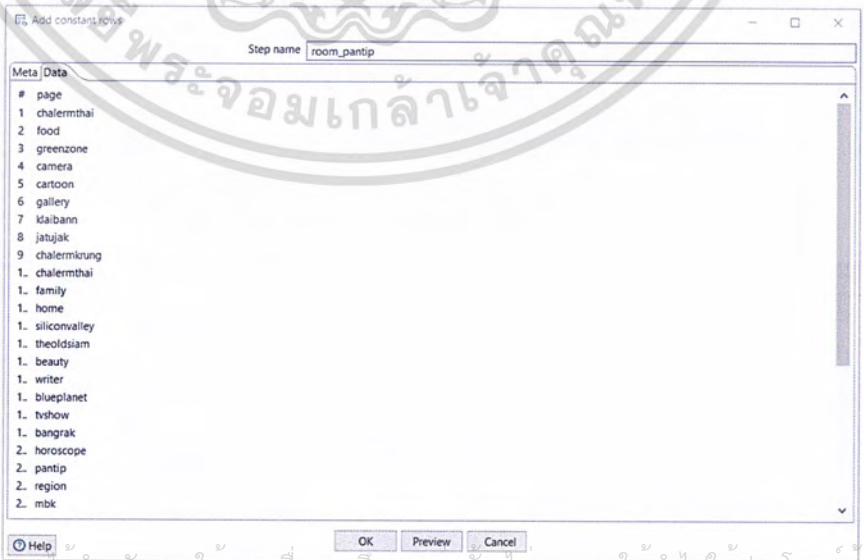
1. ส่วนการนำข้อมูลจาก RSS



รูปที่ 4.28 สถาปัตยกรรมของการทำงานส่วนการนำข้อมูลจาก RSS

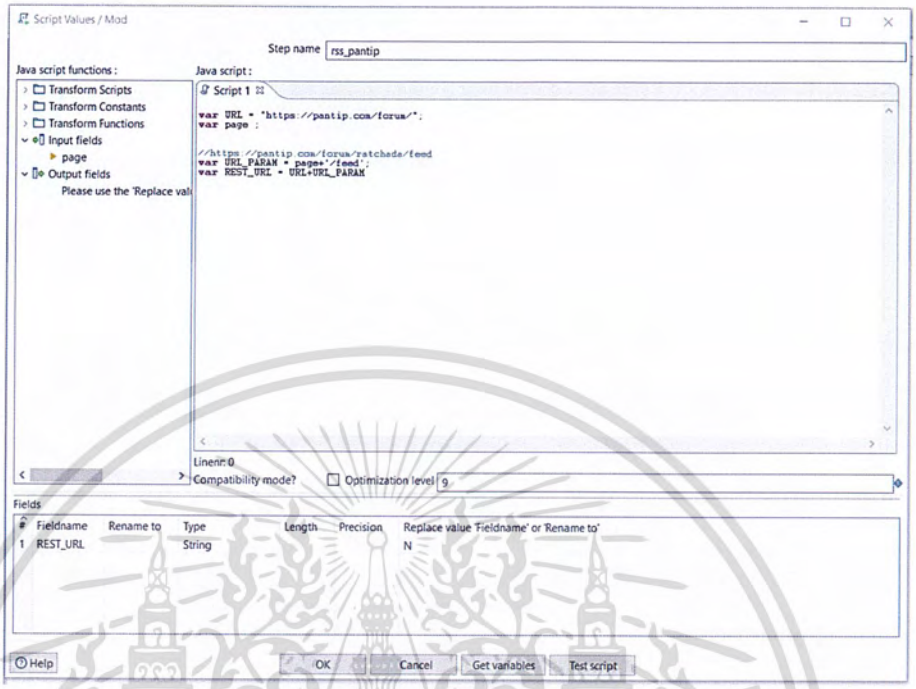
จากรูปที่ 4.28 สามารถอธิบายส่วนของการทำงานดังนี้

1) room_pantip ใช้เก็บชื่อห้องต่างๆของ Pantip แสดงดังรูปที่ 4.29



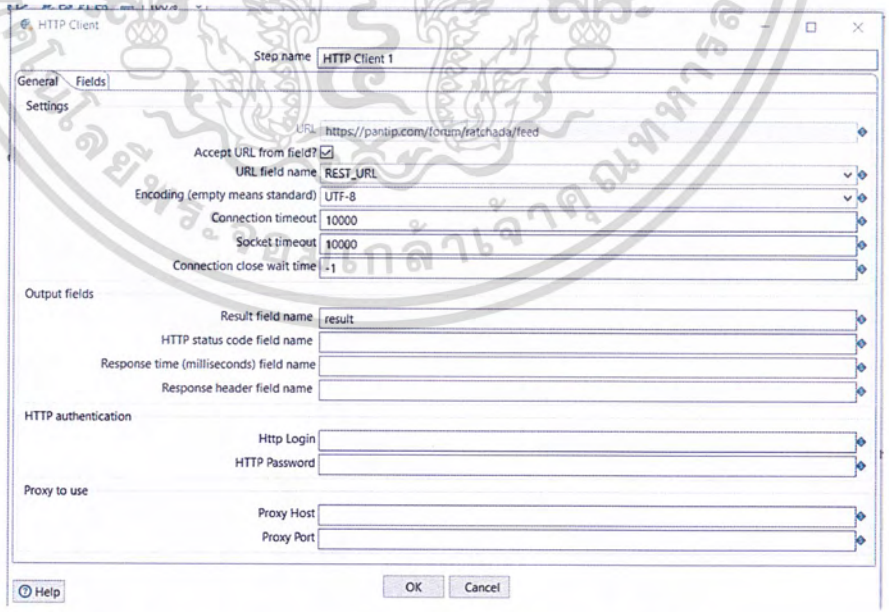
รูปที่ 4.29 กำหนดชื่อห้องต่างๆของ Pantip

2) เขียนคำสั่ง javascript ในการกำหนด URL โดยนำชื่อห้องต่างๆของ Pantip ตามด้วย /feed แสดงดังรูปที่ 4.30



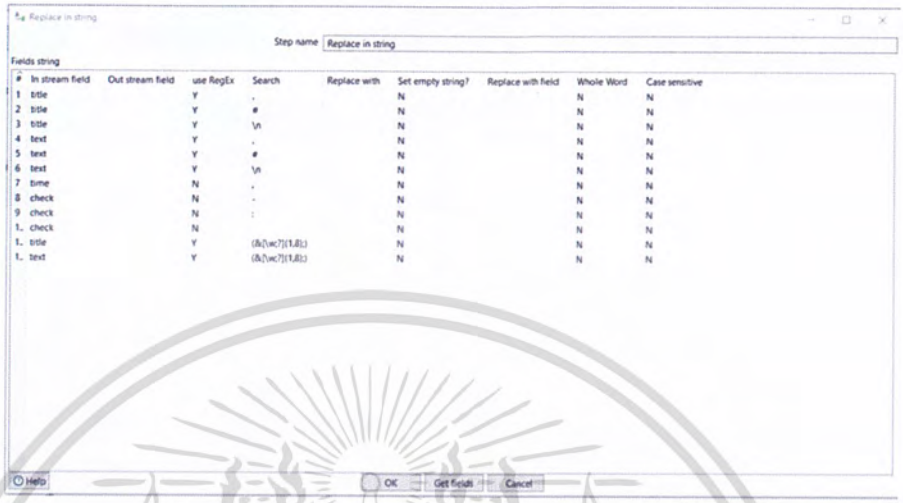
รูปที่ 4.30 คำสั่ง javascript ในการกำหนด URL

3) http client จะทำการส่ง request ตาม URL ที่รับมา แสดงดังรูปที่ 4.31



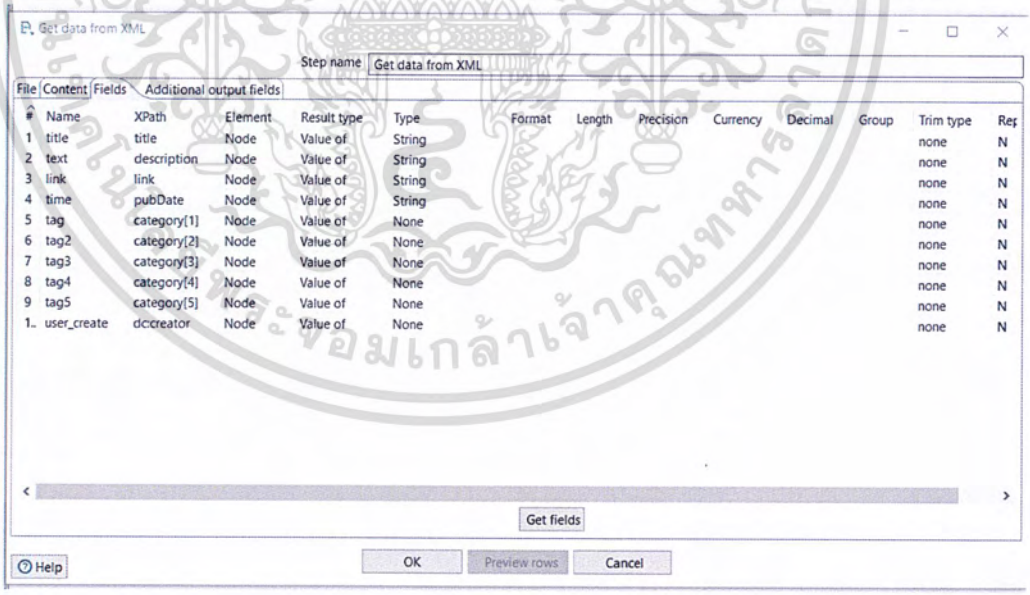
รูปที่ 4.31 http client ส่ง request ตาม URL ที่รับมา

4) replace string ใช้ในการจัดรูปแบบของคำให้ตรงตามความต้องการเช่น ค้นหารูปแบบคำที่ไม่ต้องการโดยใช้ Regular Expression ในการกำหนดรูปแบบของคำค้นหา แล้วแทนที่ด้วยค่าว่าง แสดงดังรูปที่ 4.32



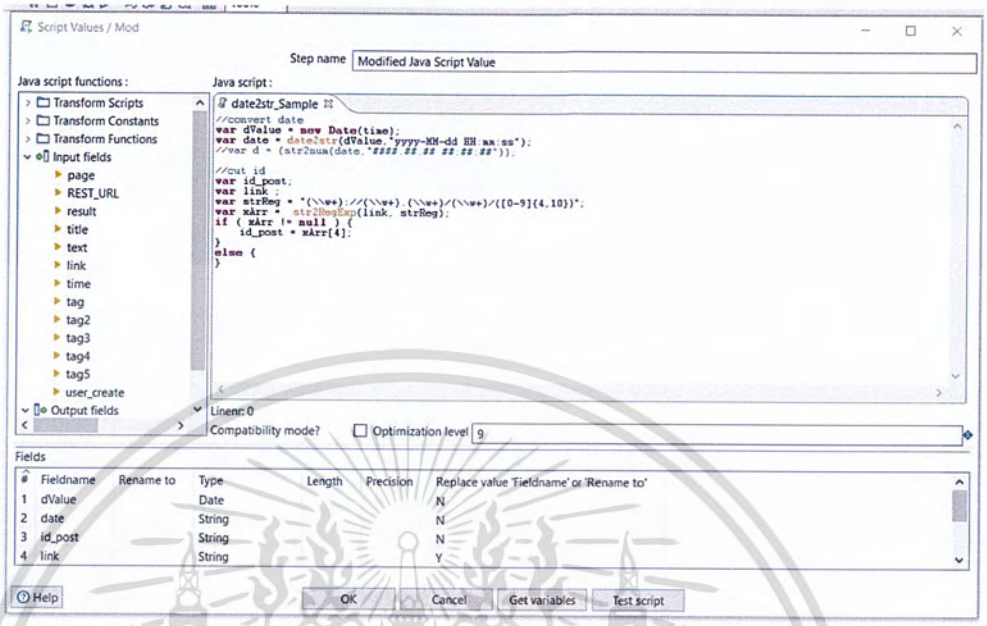
รูปที่ 4.32 replace string จัดรูปแบบของคำ

5) get data from XML เมื่อส่ง request ไป ข้อมูลที่ได้จะอยู่ในรูปแบบ XML ซึ่งต้องใช้ get data from XML ในการอ่านข้อมูล โดยเรียกตาม path ของข้อมูล แสดงดังรูปที่ 4.33



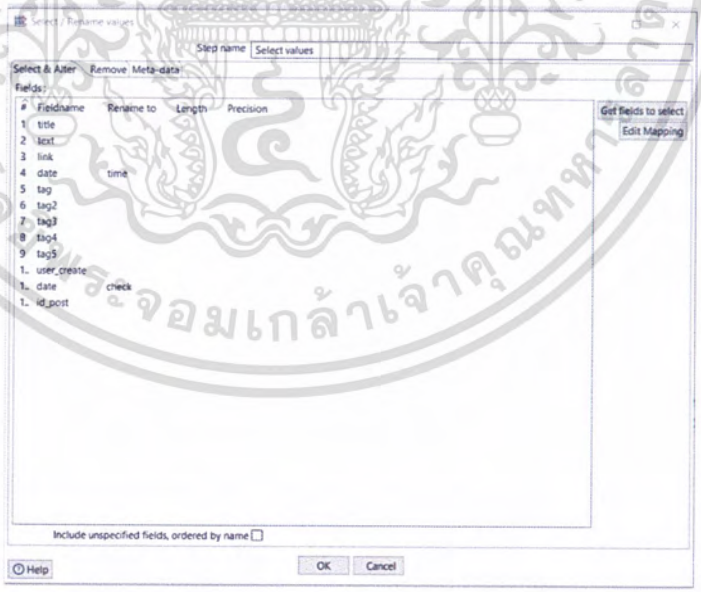
รูปที่ 4.33 get data from XML ใช้อ่านข้อมูลตาม path ของ XML

6) เขียนคำสั่ง javascript ในการแปลงรูปแบบของวันที่เป็น 2017-06-25 25:26:00 และเขียน Regular Expression ในการหาค้นหาค่า id จาก URL แสดงดังรูปที่ 4.34



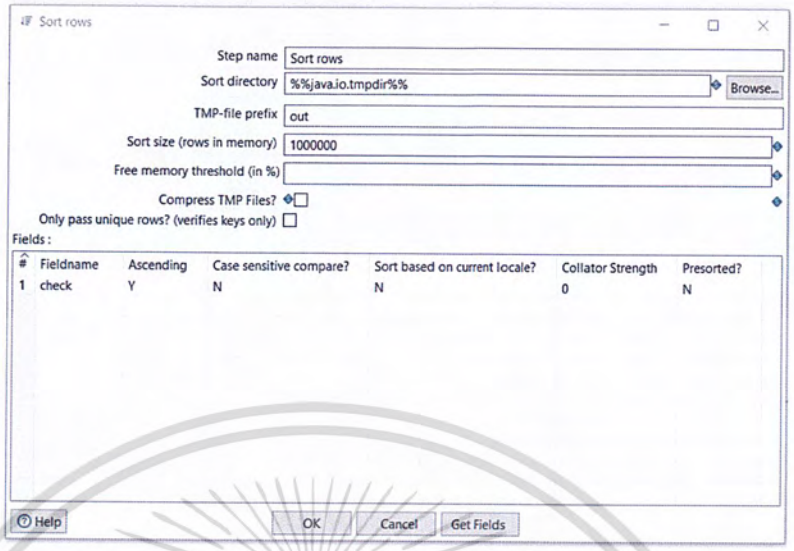
รูปที่ 4.34 คำสั่ง javascript ในการแปลงรูปแบบวันที่ และค้นหา id

7) select value เพื่อเลือก field ที่ต้องการ แสดงดังรูปที่ 4.35



รูปที่ 4.35 select value เลือก field ที่ต้องการ

8) sort row ใช้ในการเรียงลำดับของข้อมูลตามเวลา แสดงดังรูปที่ 4.36

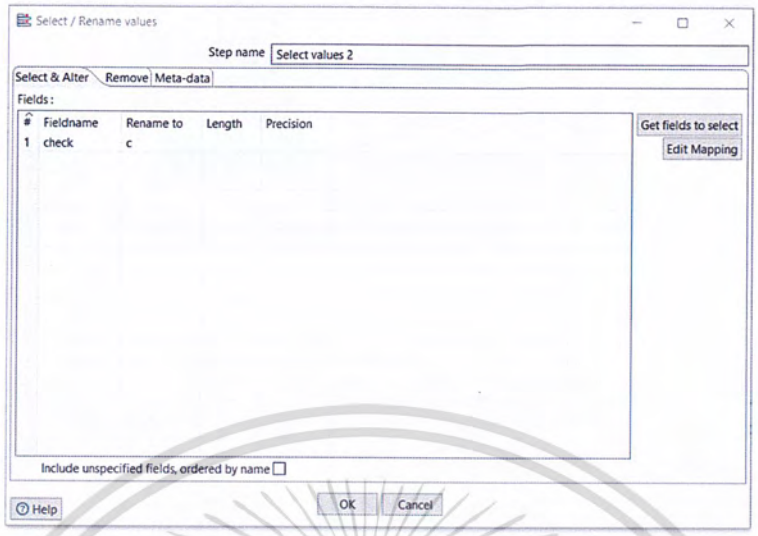


รูปที่ 4.36 sort row เรียงลำดับของข้อมูลตามเวลาจากน้อยไปหามาก

9) hadoop file input ใช้ในการอ่านค่าของข้อมูลใน file ที่กำหนด โดยอ่านจาก file ที่ใช้เก็บเวลาล่าสุด เช่น file check จะเก็บค่าของวันที่ล่าสุดที่ถูกเก็บข้อมูลเข้า Hadoop แสดงดังรูปที่ 4.37

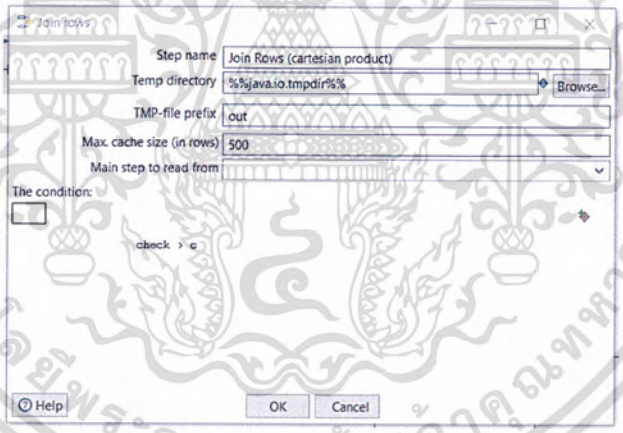
รูปที่ 4.37 hadoop file input อ่านค่าของข้อมูลใน file ที่เก็บเวลาล่าสุด

10) select value เพื่อเลือก field ที่ต้องการ แสดงดังรูปที่ 4.38



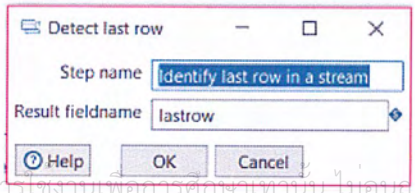
รูปที่ 4.38 select value เลือก field ที่ต้องการ

11) join row ใช้ในการเชื่อมข้อมูลสองชุดเข้าด้วยกัน แสดงดังรูปที่ 4.39



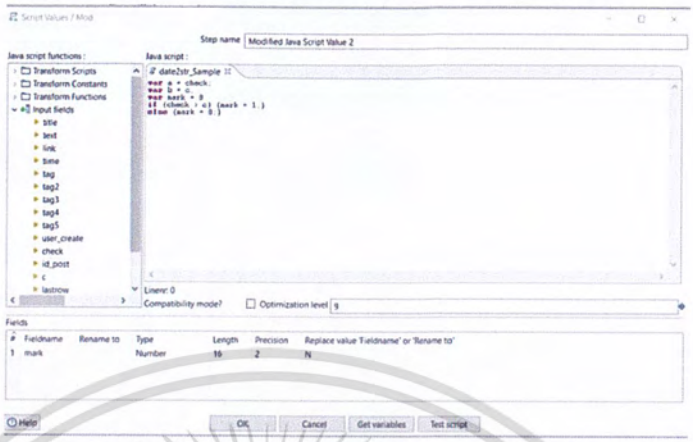
รูปที่ 4.39 join row เพื่อเชื่อมข้อมูลสองชุดเข้าด้วยกัน

12) identify last row ใช้ในการกำหนด row สุดท้าย หรือ time ล่าสุด โดยจะเพิ่ม field หนึ่ง field ซึ่งมีค่า Y แสดงว่าคือ row สุดท้าย แสดงดังรูปที่ 4.40



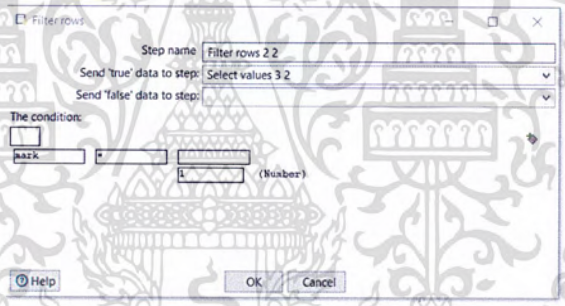
รูปที่ 4.40 identify last row กำหนด row สุดท้าย

13) เขียนคำสั่ง javascript ในการตรวจสอบข้อมูลที่รับมาว่า time มากกว่า time ล่าสุด โดยเพิ่ม field mark ถ้าเท่ากับ 1 แสดงว่าเป็นข้อมูลใหม่ แสดงดังรูปที่ 4.41



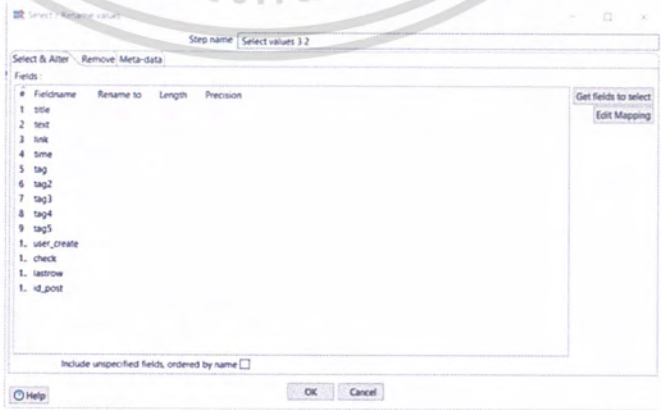
รูปที่ 4.41 คำสั่ง javascript ในการตรวจสอบ time ที่เก็บมาใหม่

14) filter row ใช้การกรอก row ที่ time มีค่ามากกว่า time ข้อมูลเก่า แสดงดังรูปที่ 4.42



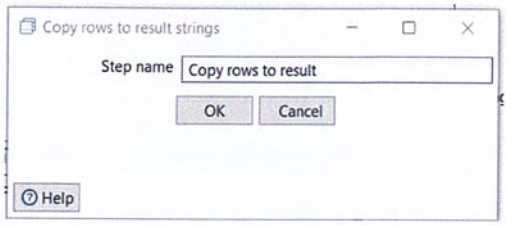
รูปที่ 4.42 filter row กรอก row ที่ time มีค่ามากกว่า time ข้อมูลเก่า

15) select value เพื่อเลือก field ที่ต้องการ แสดงดังรูปที่ 4.43



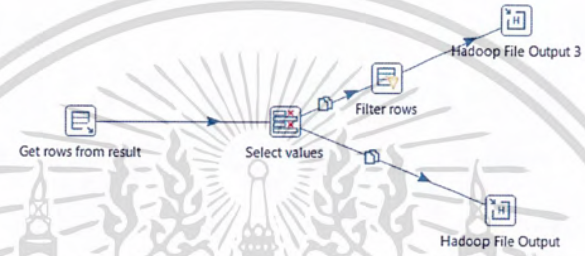
รูปที่ 4.43 select value เลือก field ที่ต้องการ

16) copy row to result เป็นการคัดลอกค่าเพื่อไว้ใช้ในส่วนการทำงาน แสดงดังรูปที่ 4.44



รูปที่ 4.44 คัดลอกค่าเพื่อไว้ใช้ในส่วนการทำงานถัดไป

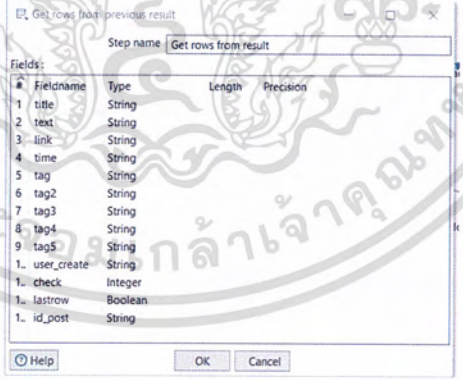
2. ส่วนการนำข้อมูลเข้า Hadoop



รูปที่ 4.45 สถาปัตยกรรมของการทำงานส่วนการนำข้อมูลเข้า Hadoop

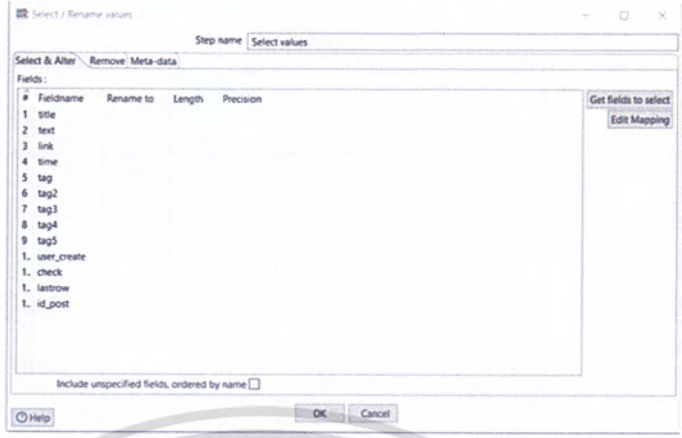
จากรูปที่ 4.45 สามารถอธิบายส่วนของการทำงานดังนี้

1) get row from result เป็นการรับค่ามาจากส่วนการทำงานก่อนหน้า แสดงดังรูปที่ 4.46



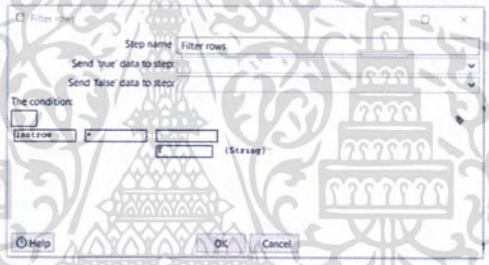
รูปที่ 4.46 get row from result รับค่ามาจากส่วนการทำงานก่อนหน้า

2) select value เพื่อเลือก field ที่ต้องการ แสดงดังรูปที่ 4.47



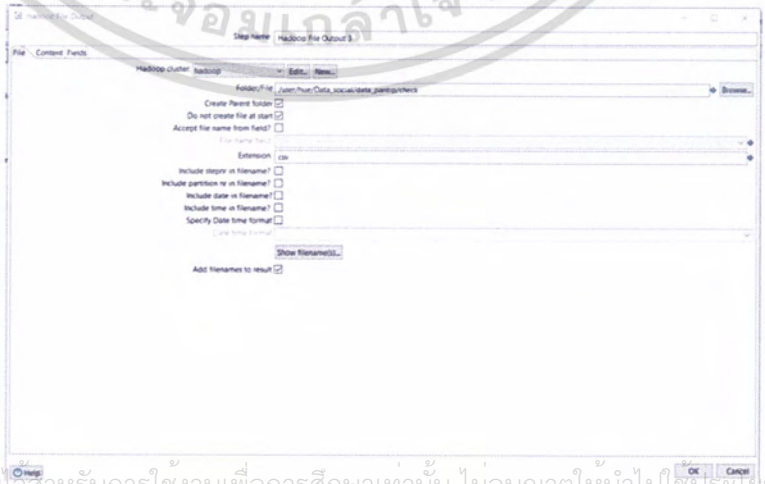
รูปที่ 4.47 select value เลือก field ที่ต้องการ

3) filter row ใช้กรอก row สุดท้ายซึ่งจะมีค่าของ time ล่าสุด เพื่อเอาไปเก็บที่ file check แสดงดังรูปที่ 4.48



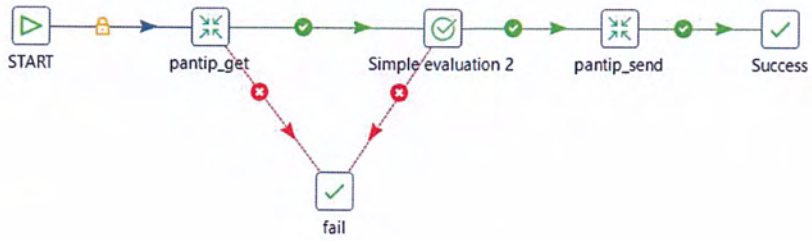
รูปที่ 4.48 filter row กรอก row ที่มีค่าของ time ล่าสุด

4) hadoop file output ใช้ในการระบุที่อยู่ของ file ปลายทางและระบุ field ที่ต้องการ แสดงดังรูปที่ 4.49



รูปที่ 4.49 hadoop file output ระบุที่อยู่ของ file ปลายทางและ field ที่ต้องการ

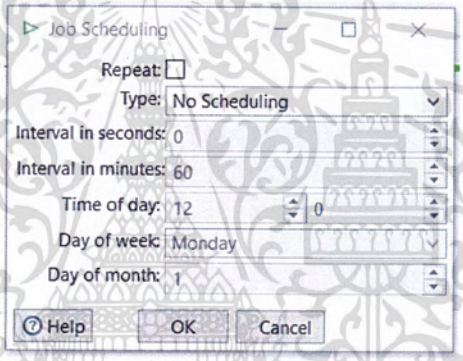
3. ส่วนการทำงานของ Job



รูปที่ 4.50 สถาปัตยกรรมของการทำงานส่วนการทำงานของ Job

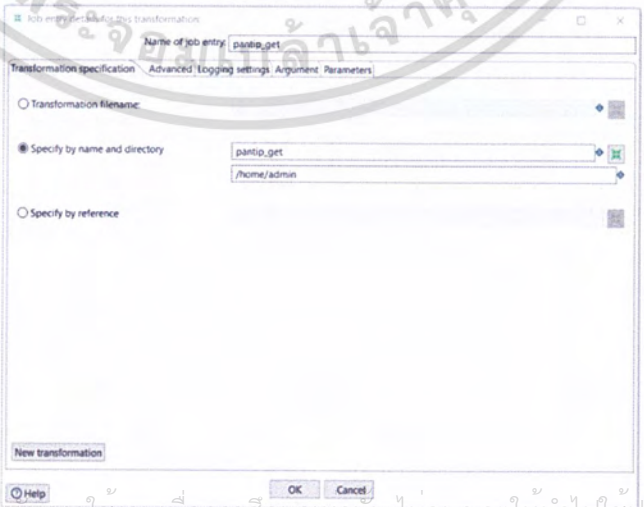
จากรูปที่ 4.50 สามารถอธิบายส่วนของการทำงานดังนี้

1) start เป็นส่วนเริ่มการทำงาน สามารถกำหนดเวลาในการเริ่มทำงานได้ แสดงดังรูปที่ 4.51



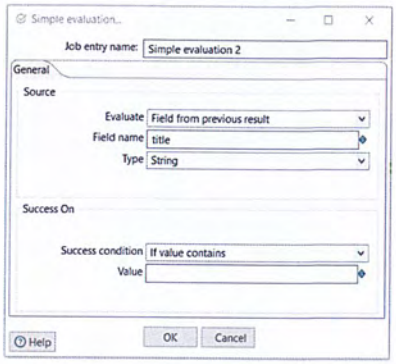
รูปที่ 4.51 start เป็นส่วนเริ่มการทำงาน

2) pantip_get เป็นส่วนการทำงานของการนำข้อมูลจาก Pantip แสดงดังรูปที่ 4.52



รูปที่ 4.52 pantip_get ใช้เป็นส่วนการทำงานของการนำข้อมูลจาก Pantip

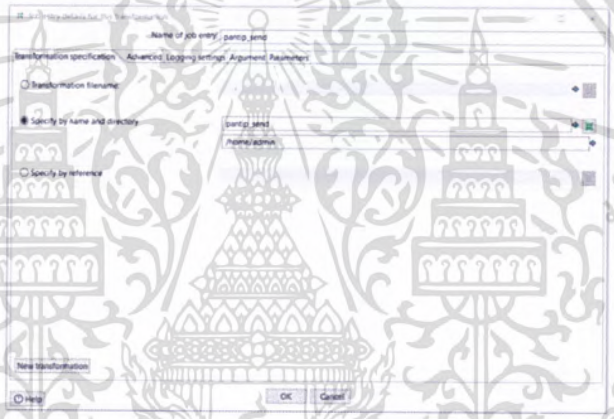
3) simple evaluation เป็นส่วนที่ตรวจสอบการทำงานก่อนหน้ามีข้อมูล แสดงดังรูปที่ 4.53



รูปที่ 4.53 simple evaluation ตรวจสอบว่าการทำงานก่อนหน้ามีข้อมูล

4) fail ใช้สำหรับการจบการทำงาน เมื่อเกิดข้อผิดพลาดจากส่วนการทำงานก่อนหน้า

5) pantip_send เป็นส่วนการทำงานนำข้อมูลเข้า Hadoop แสดงดังรูปที่ 4.54

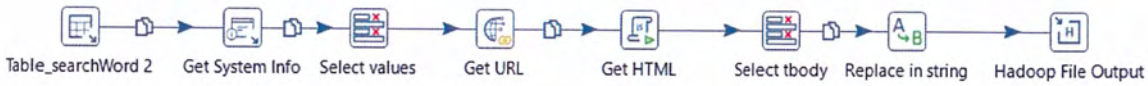


รูปที่ 4.54 pantip_send เป็นส่วนการทำงานนำข้อมูลเข้า Hadoop

6) success ใช้สำหรับการจบการทำงาน เมื่อการทำงานส่วนก่อนหน้าเสร็จ

4.2.4 การทำงานของส่วนการนำข้อมูลจากเว็บต่างๆโดยอ่านจากหน้า HTML

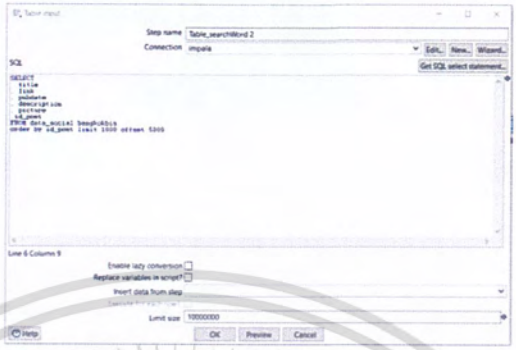
กระบวนการทำงานของส่วนการนำข้อมูลจากเว็บต่างๆโดยอ่านจากหน้า HTML จะเป็นกระบวนการที่เก็บจำนวนของผู้เข้าชม จำนวนคนถูกใจ จำนวนคนแชร์ เว็บข่าว แสดงดังรูปที่ 4.55



รูปที่ 4.55 สถาปัตยกรรมของส่วนการนำข้อมูลจากเว็บต่างๆโดยอ่านจากหน้า HTML

จากรูปที่ 4.55 สามารถอธิบายส่วนของการทำงานดังนี้

1) table search เขียนคำสั่ง SQL ใช้ในการค้นหา id และ URL จาก table ข่าวต่างๆ แสดงดังรูปที่ 4.56

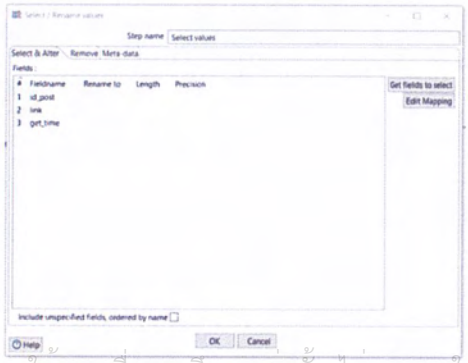


รูปที่ 4.56 table search เขียนคำสั่ง SQL ค้นหา id และ URL จาก table ข่าวต่างๆ

2) get system into รับค่าวันที่ เวลาที่ทำการดึงข้อมูลไว้ใช้ในการเทียบข้อมูลล่าสุด แสดงดังรูปที่ 4.57

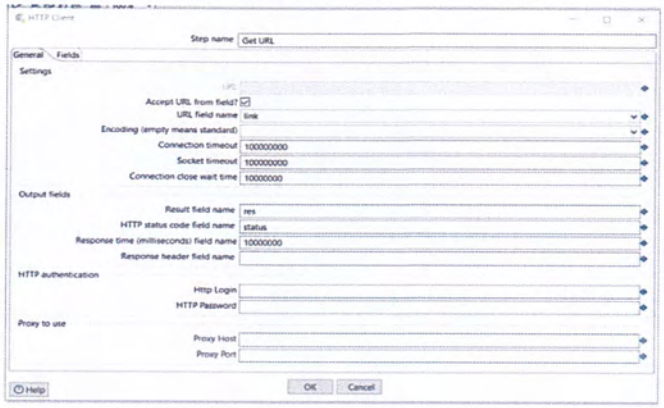
รูปที่ 4.57 get system into รับค่าวันที่ เวลาที่ทำการดึงข้อมูลไว้ใช้ในการเทียบข้อมูลล่าสุด

3) select value เพื่อเลือก field ที่ต้องการ แสดงดังรูปที่ 4.58



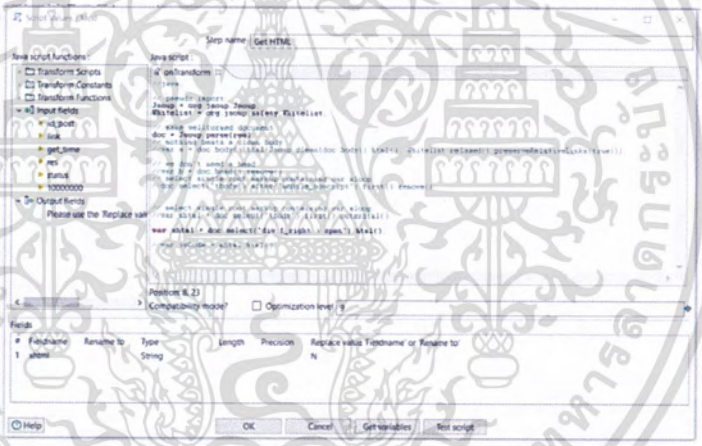
รูปที่ 4.58 select value เลือก field ที่ต้องการ

4) http client จะทำการส่ง request ตาม URL ที่รับมา แสดงดังรูปที่ 4.59



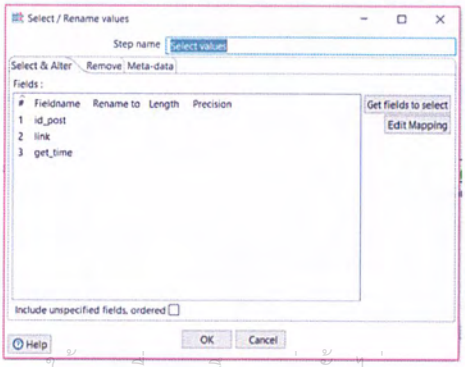
รูปที่ 4.59 http client ส่ง request ตาม URL ที่รับมา

5) get html เขียนคำสั่ง javascript โดยใช้ library Jsoup ในการอ่านข้อมูลจาก HTML ซึ่งก่อนการเรียนใช้ library ต้องทำการโหลด file library มาติดตั้งก่อน แสดงดังรูปที่ 4.60



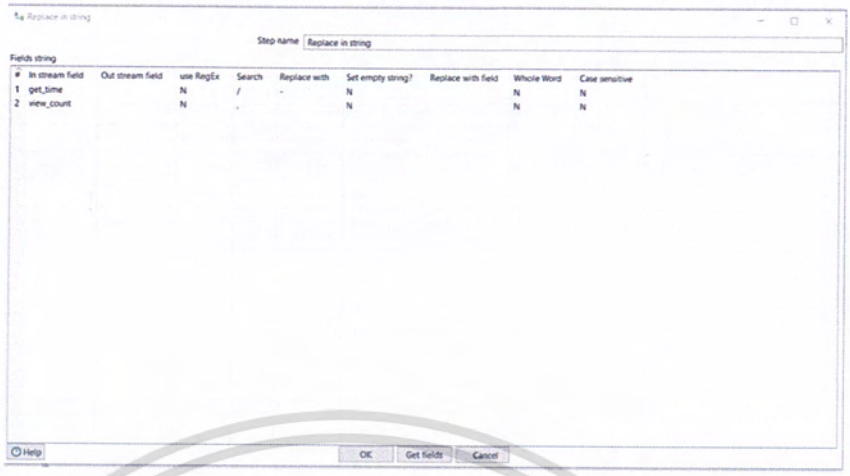
รูปที่ 4.60 get html เขียนคำสั่ง javascript ใช้ library Jsoup ในการอ่านข้อมูลจาก HTML

6) select value เพื่อเลือก field ที่ต้องการ แสดงดังรูปที่ 4.61



รูปที่ 4.61 select value เลือก field ที่ต้องการ

7) replace string ใช้ในการจัดรูปแบบของคำให้ตรงตามความต้องการ แสดงดังรูปที่ 4.62



รูปที่ 4.62 replace string จัดรูปแบบของคำ

8) hadoop file output ใช้ในการระบุที่อยู่ของ file ปลายทางและระบุ field ที่ต้องการ แสดงดังรูปที่ 4.63

รูปที่ 4.63 hadoop file output ระบุที่อยู่ของ file ปลายทางและ field ที่ต้องการ

บทที่ 5

สรุปผลการดำเนินงานและข้อเสนอแนะ

5.1 สรุปผลการดำเนินงาน

การพัฒนาระบบการดึงข้อมูล (ETL) จาก Data Source ต่างๆ เข้าสู่ Hadoop มีจุดประสงค์เพื่อดึงข้อมูล ที่มีขนาดมหาศาลจาก Social media แหล่งต่างๆและจัดเก็บข้อมูล ให้เหมาะสม เพื่อการนำข้อมูลไปใช้ วิเคราะห์ได้อย่างสะดวก

ผู้จัดทำสามารถพัฒนาระบบการดึงข้อมูล (ETL) จาก Data Source ต่างๆ เข้าสู่ Hadoop และจัดเก็บข้อมูลได้สำเร็จตามขอบเขตงานที่กำหนดไว้ และสามารถนำไปใช้งานได้เป็นอย่างดี มีความถูกต้อง ซึ่งทำให้ผู้ที่รับผิดชอบในการนำข้อมูลไปใช้วิเคราะห์สามารถทำงานได้อย่างสะดวก และรวดเร็ว

5.2 ปัญหาและข้อจำกัด

1) ผู้จัดทำใช้เวลาในการศึกษา Hadoop, Pentaho, Docker, Angular และรูปแบบการดึงข้อมูลของแต่ละแหล่งมากพอสมควรเนื่องจากไม่มีพื้นฐานและความรู้ในด้านที่เกี่ยวข้อง
2) ไม่สามารถพัฒนาตามที่ออกแบบไว้ได้ในบางรายการ ด้วยเวลาที่จำกัด และความรู้ความสามารถของนักศึกษาสหกิจยังมีไม่เพียงพอ

5.3 ข้อเสนอแนะและแนวทางในการพัฒนา

1) ควรมีการตรวจสอบความต้องการอีกครั้ง เมื่อจะพัฒนาต่อ
2) ควรศึกษารูปแบบการดึงข้อมูลของแต่ละแหล่งการจัดรูปแบบข้อมูลที่ดึงให้มีความถูกต้องสมบูรณ์มากยิ่งขึ้น

เอกสารอ้างอิง

[1] SURANART NIAMCOME, **Big Data คืออะไร วิธีใช้ Hadoop/Spark บน Cloud Dataproc**, [Online], Available: http://www.siamhtml.com/getting-started-with-big-data-and-hadoop-spark-on-cloud-dataproc/,เข้าถึงเมื่อวันที่ 17 สิงหาคม 2560.

[2] ธนชาติ นุ่มนนท์, **hadoop-ecosystem-สำหรับการพัฒนา-big-data**, [Online], Available: https://thanachart.org/2014/10/18/hadoop-ecosystem-สำหรับการพัฒนา-big-data/, เข้าถึงเมื่อวันที่ 17 สิงหาคม 2560.

[3] Cloudera.com, **Impala Concepts and Architecture**, [Online], Available: https://www.cloudera.com/ documentation/cdh/5-0-x/Impala/Installing-and-Using-Impala/ciiu_concepts.html, เข้าถึงเมื่อวันที่ 17 สิงหาคม 2560.

[4] Goingjesse.com, **Pentaho Open Source BI**, [Online], Available: http://www.goingjesse.com/pentaho-solution, เข้าถึงเมื่อวันที่ 20 ตุลาคม 2560.

[5] it.mcu.ac.th, **API คืออะไร ทำหน้าที่อะไร ประโยชน์ของ API มีอะไรบ้าง**, [Online], Available: http://www2.it.mcu.ac.th/?p=3748, เข้าถึงเมื่อวันที่ 13 ธันวาคม 2560.

[6] Pattanapong Cherthong, **ทำความรู้จัก Docker และ Software Container**, [Online], Available: https://medium.com/thothzocial-engineering/ทำความรู้จัก-docker-และ-software-container-c6338629da11, เข้าถึงเมื่อวันที่ 2 พฤศจิกายน 2560.

[7] dtv.mcot.net, **ควบคุม Raspberry Pi ผ่าน LAN ใช้ SSH Secure Shell,** [Online], Available:http://dtv.mcot.net/data/up_show.php?id=1453477569&web=epost, เข้าถึงเมื่อวันที่ 2 พฤศจิกายน 2560.

[8] jetbrains.com, **Feature,** [Online], Available: https://www.jetbrains.com/webstorm/features/,เข้าถึงเมื่อวันที่ 17 มกราคม 2561.

[9] angular.io, **What is Angula**, [Online], Available: https://angular.io/docs#what-is-Angular, เข้าถึงเมื่อวันที่ 17 มกราคม 2561.

ภาคผนวก

ภาคผนวก ก

ผลงานที่ได้รับรางวัล

การประชุมวิชาการระดับปริญญาตรีด้านคอมพิวเตอร์ระดับภูมิภาคอาเซียนครั้งที่ 6 (The 6 ASEAN Undergraduate Conference in Computing: AUCC 2018)

ส่งเข้าแข่งขันในประเภทการนำเสนอด้วยวาจา (Oral Presentation) รูปถ่ายและผลงานมีดังต่อไปนี้

1) ได้รับรางวัลผลงานดีมาก (Very Good Paper Award) ในหมวด Knowledge and Data Management แสดงดังรูปที่ ก.1

The 6th ASEAN Undergraduate Conference in Computing (AUC² 2018)

Certificate of Oral Presentation

presented to

โสรยา สุวรรณธนชัย, สุทัศน์ศิณี จิตต์ซื่อ, ศิรัญญา พวงดี
และ วรางคณา กิ้มปาน

for a contribution in the field of computing entitled

การพัฒนาระบบค้นคืนข้อมูลจากโซเชียลเน็ตเวิร์คด้วย Pentaho

upon with the success of *Very Good Paper Award*

held on 23rd – 25th March 2018 at Faculty of Science,
King Mongkut's Institute of Technology Ladkrabang, Thailand

Assoc. Prof. Dr. Ittipol Jangchud
General Chair of AUC² 2018

รูปที่ ก.1 รางวัลที่ได้รับ

2) ผลงานที่นำไปแข่งขัน

การพัฒนาระบบค้นคืนข้อมูลจากโซเชียลเน็ตเวิร์คด้วย Pentaho

DEVELOPMENT OF SOCIAL NETWORK DATA RETRIEVAL SYSTEM USING PENTAHO

โสรยา สุวรรณธนชัย สุทัศน์ศิณี จิตต์ซื่อ ศิรัญญา พวงดี และวรางคณา กิ้มปาน
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
Emails: 57050352@kmitl.ac.th, 57050357@kmitl.ac.th, 57050333@kmitl.ac.th, knwarang@kmitl.ac.th

บทคัดย่อ

บทความนี้นำเสนอการพัฒนาระบบค้นคืนข้อมูลจากโซเชียลเน็ตเวิร์คด้วย Pentaho เก็บลงใน Hadoop โดยเก็บข้อมูลจากโซเชียลเน็คเวิคจากแหล่งต่างๆ เช่น เฟซบุ๊ก (Facebook) ทวิตเตอร์ (Twitter) พันทิป (Pantip) และเว็บข่าว (RSS)ต่างๆ มีการทำให้ข้อมูลมีความถูกต้อง (Data Cleaning) ก่อนการเก็บลงใน Hadoop เพื่อทำให้ผู้ใช้สามารถนำข้อมูลไปใช้ในการวิเคราะห์ได้อย่างสะดวกและรวดเร็ว

ABSTRACT

This paper presents the development of data retrieval system from the social network using Pentaho. It will retrieve the required data into Hadoop using Pentaho. The system collects the data from the social networks such as Facebook, Twitter, and RSS (Really Simple Syndication). Data validation was done before being stored in Hadoop. This allows users to easily and quickly analyze data.

คำสำคัญ— Hadoop; Pentaho; Facebook; Twitter; RSS;

1. บทนำ

เนื่องจากในปัจจุบันมีข้อมูลจำนวนมหาศาลที่เกิดขึ้นบนโลกโซเชียลเน็คเวิร์คบนเว็บไซต์ต่างๆ ซึ่งข้อมูลเหล่านี้เป็นข้อมูลที่ซับซ้อนและไม่ได้อยู่ในรูปแบบโครงสร้างเดียวกัน การจะนำข้อมูลจำนวนมหาศาลที่ได้มาใช้ให้เกิดประโยชน์ ต้องอาศัยเทคโนโลยีบิ๊กดาต้า (Bigdata) เนื่องจากการที่ข้อมูลมีจำนวนมหาศาล (Volume) มีหลายรูปแบบ (Variety) และข้อมูลมีการเพิ่มขึ้นอย่างรวดเร็ว (Velocity) ทำให้องค์กรต่างๆ ต้องปรับโครงสร้างพื้นฐานด้านข้อมูล (Information Infrastructure) ในปัจจุบันได้มีการนำเทคโนโลยีใหม่ เช่น Hadoop, NoSQL หรือ NewSQL เข้ามาใช้งาน ซึ่งจำเป็น ต้องมีการพัฒนาบุคลากรเพื่อให้เข้าใจการใช้เทคโนโลยีเหล่านี้ รวมถึงมีความรู้ความสามารถในการนำข้อมูลต่างๆ ไปทำการวิเคราะห์และสรุปผลการวิเคราะห์ให้อยู่ในรูปแบบที่เข้าใจง่าย

2. ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีที่นำมาใช้ในการพัฒนาระบบดึงข้อมูลจากโซเชียลเน็ตเวิร์คด้วยPentaho มีดังนี้

2.1. Cloudera Hadoop

Cloudera Hadoop [1] เป็นโอเพ่นซอร์สแพลตฟอร์ม ที่มีการพัฒนาต่อยอดจาก Apache Hadoop ซึ่งเป็นซอฟต์แวร์มาตรฐานในวงการบิ๊กดาต้า (Bigdata) โดยใช้ชื่อผลิตภัณฑ์ว่า CDH (Cloudera Distribution Hadoop) มีการจัดหาเครื่องมือสำหรับสำหรับการจัดเก็บข้อมูลขนาดใหญ่ เช่น Hadoop, Spark, Hive, Impala, Mahout, Sqoop, Flume, Oozie, HBase เป็นต้น

2.2. Hadoop

Hadoop [2] เป็น โอเพ่นซอร์สโปรเจค ของ Apache สำหรับการเก็บและบริหารข้อมูลขนาดใหญ่ Hadoop เขียนด้วยโปรแกรมภาษาจาวา มีความสามารถในการป้องกันข้อมูล

เสียหาย (Fault Tolerance) เพราะจะเก็บข้อมูลซ้ำกันในหลายๆ ที่และเป็นระบบที่เป็นสามารถเพิ่มจำนวนของเครื่องที่ใช้ในการประมวลผล (Horizontal Scale) ที่รันบนเครื่องเซิร์ฟเวอร์ทั่วๆไป (Commodity Server) จำนวนมาก องค์ประกอบของ Hadoop แบ่งออกเป็นโมดูลย่อยๆ ดังนี้

1) Hadoop Distributed File System (HDFS) ใช้จัดเก็บข้อมูลที่จะนำมาวิเคราะห์ให้อยู่ในรูปแบบที่สามารถเข้าถึงได้อย่างรวดเร็ว รวมไปถึงการสำรองข้อมูลดังกล่าวให้โดยอัตโนมัติ

2) MapReduce ใช้สำหรับการประมวลผลข้อมูลปริมาณมหาศาลที่ได้เก็บเอาไว้ โดยใช้ฟังก์ชันการ Map และ Reduce ระบบจะกระจายงานไปรันแบบ Parallel บนเครื่องหลายๆ เครื่อง

Hadoop มีการทำงานโดยแบ่งไฟล์ออกเป็น Block ย่อยๆ เช่นไฟล์ CSV ขนาด 1TB การประมวลผลไฟล์ขนาดใหญ่จะช้า ดังนั้น HDFS จะแบ่งไฟล์ออกเป็นไฟล์ย่อยๆเรียกว่า Block แล้วนำไปเก็บกระจายตามโหนด (Node) ต่างๆ ใน Cluster ทำให้โหนดต่างๆ สามารถช่วยกันประมวลผลไฟล์ CSV แบบขนานได้ นอกจากนั้น HDFS ยังช่วยทำสำเนา(replicate) แต่ละ Block เอาไว้ที่โหนดอื่นๆ ด้วย (default คือ ทำสำเนาไป 3 โหนด) เช่น Block A ของ Node 1 ไม่สามารถใช้งานได้แต่สามารถมั่นใจได้เลยว่า Block A จะยังมีสำรองอยู่ในโหนดอื่นๆ อย่างแน่นอน ข้อมูลที่ถูกเก็บอยู่ใน HDFS จะไม่ใช่รูปแบบของตารางอย่างที่เก็บในฐานข้อมูลเชิงสัมพันธ์ (RDBMS)

2.3. Pentaho Data Integration

Pentaho Data Integration[3] คือเครื่องมือที่ใช้สร้างคลังข้อมูล (Data Warehouse) โดยทั่วไปเครื่องมือดังกล่าวเรียกว่า ETL (Extraction, Transformation and Loading)

1) Extraction คือการดึงข้อมูลจากแหล่งต่างๆ ที่ต้องการมาเก็บไว้ใน Data Warehouse โดยจะดึงมาเฉพาะข้อมูลใหม่ที่เพิ่มขึ้นมาหรือข้อมูลที่ถูกเปลี่ยนแปลงแก้ไขโดยข้อมูลที่ดึงมาจะนำมาเก็บพักไว้ก่อน

2) Transformation คือการเปลี่ยนแปลงรูปแบบของข้อมูลที่ได้จากการ Extract ให้อยู่ในรูปแบบที่ถูกต้องตามโครงสร้างของ Data Warehouse

3) Loading คือการเก็บข้อมูลลงใน Data Warehouse หลังจากทำการแปลงข้อมูลให้อยู่ในรูปแบบที่ถูกต้อง

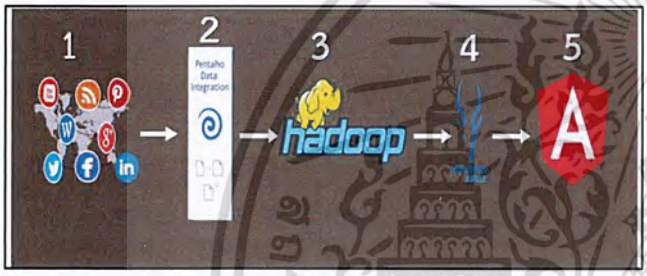
ความสามารถของ Pentaho Data Integration ประกอบด้วย รองรับมาตรฐาน Java และง่ายต่อการใช้งาน ด้วยเครื่องมือต่างๆ ที่จัดเตรียมไว้ให้ในรูปแบบกราฟิก เพียงลากวางเครื่องมือต่างๆ ตามกระบวนการที่ต้องทำ มีเครื่องมือที่ใช้ในการตรวจสอบความถูกต้องของข้อมูล (Data Quality) อีกทั้งยังรองรับหลากหลายแหล่งข้อมูล (Data Source) ไม่ว่าจะเป็น file base (DBF), text file, excel file, ฐานข้อมูลประเภทต่างๆ เช่น MySQL, PostgreSQL, Oracle เป็นต้น สามารถเชื่อมต่อกับ Pentaho BI ทำให้ใช้ความสามารถอื่นๆ ร่วมกันได้เช่น เรื่องของการจัดตารางการทำงาน (Scheduling), ความปลอดภัย (Security), ขั้นตอนการทำงาน (Workflow) เป็นต้น Pentaho Data Integration สามารถนำไปใช้งานในลักษณะต่างๆ ดังนี้ สร้างคลังข้อมูล (Populate Data Warehouse) ส่งออก (Export) ข้อมูลจากฐานข้อมูลไปเป็น text file นำเข้า (Import) ข้อมูลจาก text file เข้าฐานข้อมูล นำข้อมูลจากฐานข้อมูลหนึ่งไปเข้าอีกฐานข้อมูลหนึ่งและดูข้อมูลจากฐานข้อมูลที่มีอยู่

2.4. Impala

Impala[4] เป็น Massively Parallel Processing (MPP) ที่มีหน่วยประมวลผลที่ใช้ระบบปฏิบัติการและหน่วยความจำของตนเอง มีความสามารถในการค้นหาข้อมูล โดยใช้คำสั่ง SQL สำหรับการประมวลผลข้อมูลปริมาณมากที่ถูกเก็บไว้ในคลัสเตอร์ Hadoop Impala เป็นซอฟต์แวร์โอเพนซอร์ส ภายใต้ Apache License ที่มีประสิทธิภาพสูงเมื่อเทียบกับเครื่องมืออื่น ๆ สำหรับ Hadoop ข้อดีของ Impala สามารถประมวลผลข้อมูลที่เก็บไว้ใน HDFS ได้อย่างรวดเร็วและสามารถเข้าถึงข้อมูลที่ถูกเก็บไว้ใน HDFS, HBase และ Amazon S3 ด้วยคำสั่ง SQL คุณสมบัติของ Impala สามารถรองรับหน่วยความจำในการประมวลผลข้อมูล เช่น เข้าถึงหรือวิเคราะห์ข้อมูลที่เก็บไว้ในโหนดข้อมูล Hadoop โดยไม่ต้องเคลื่อนย้ายข้อมูลและ Impala สามารถใช้งานกับเครื่องมือทางธุรกิจเช่น Tableau หรือซอฟต์แวร์ Pentaho ได้อีกด้วย

3. การทำงานของระบบ

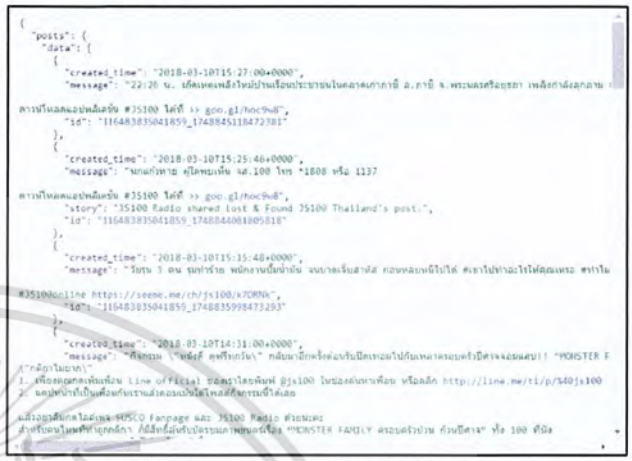
ระบบการดึงข้อมูลจาก social network แหล่งต่างๆ เช่น เฟซบุ๊ก (Facebook) ทวิตเตอร์ (Twitter) พันทิป (Pantip) และแหล่งข่าวต่างๆ จะใช้เครื่องมือ Pentaho ในการดึงข้อมูลออกมาและนำเข้าไปเก็บไว้ใน Hadoop แล้วค้นหาข้อมูลที่อยู่ใน Hadoop โดยผ่าน Impala เพื่อนำมาแสดงผลที่เว็บแอปพลิเคชันที่พัฒนาโดย Angular ระบบประกอบไปด้วยการทำงานหลักๆ ด้วยกัน 5 ส่วน ได้แก่ ส่วนศึกษาโครงสร้างข้อมูลของแหล่งต่างๆ ที่ต้องการ ส่วนการดึงข้อมูลด้วยPentaho Data Integration ส่วนการจัดเก็บข้อมูลลงในHadoop ส่วนค้นหาข้อมูลและส่วนของการแสดงผลข้อมูล ตามความต้องการ แสดงดังรูปที่ 1



รูปที่ 1. ส่วนการทำงานของระบบ

จากรูปที่ 1 สามารถอธิบายการทำงานได้ดังนี้

1) ส่วนศึกษาโครงสร้างของข้อมูลแหล่งต่างๆ เป็นขั้นตอนที่ศึกษาวิธีการที่จะได้ข้อมูลจากแหล่งต่างๆ เช่น ข้อมูลจาก Facebook ต้องใช้งานFacebook API ในการเข้าถึงข้อมูล โดยข้อมูลที่ได้จะอยู่ในรูปแบบ JSON แสดงดังรูปที่ 2

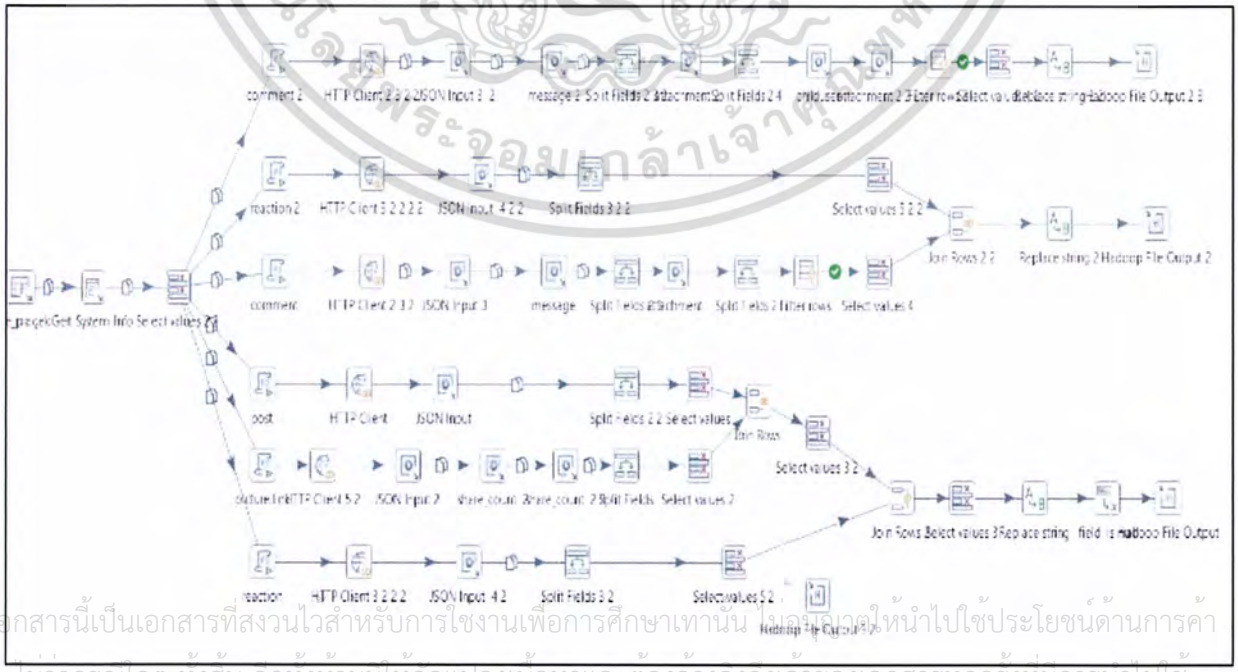


รูปที่ 2. ข้อมูลของ Facebook API ที่ค้นหา

2) ส่วนการดึงข้อมูลด้วย Pentaho เป็นขั้นตอนที่ใช้เครื่องมือในการดึงข้อมูลโดยเครื่องมือแต่ละตัวจะมีการทำงานที่แตกต่างกัน เช่น ขั้นตอนการดึงข้อมูลจาก Facebook แสดงดังรูปที่ 3

จากรูปที่3 สามารถอธิบายการทำงานได้ดังนี้

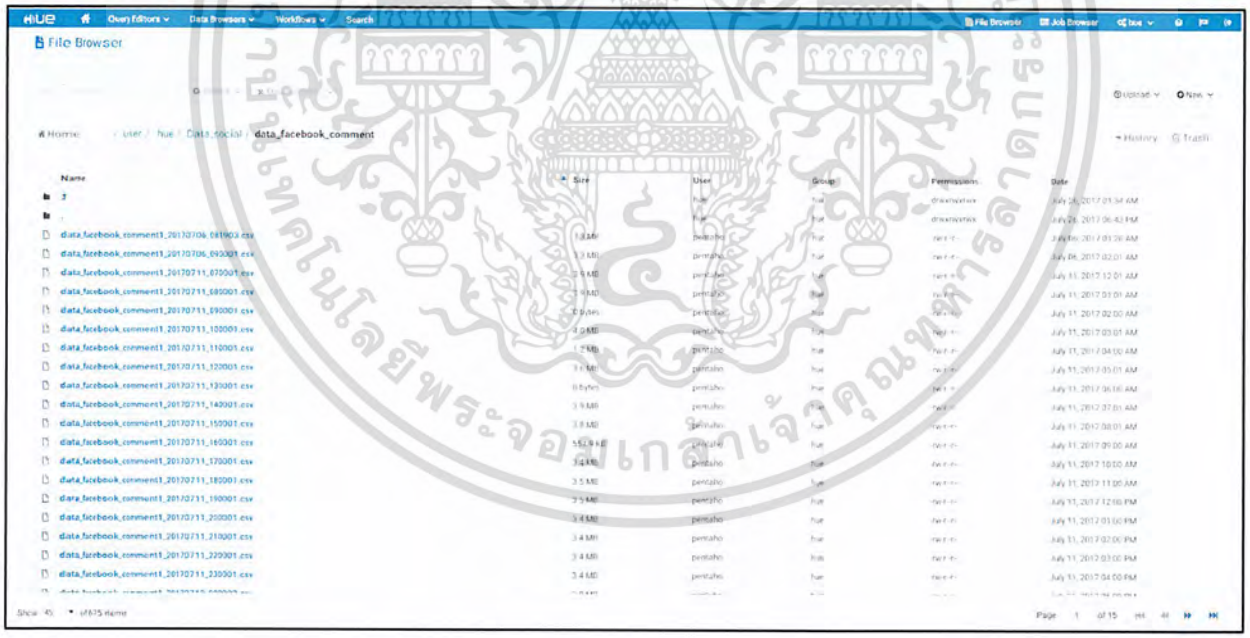
- เครื่องมือเขียนคำสั่ง SQL เพื่อค้นหา Id page ของ Facebook ที่ต้องการข้อมูล ซึ่งเก็บอยู่ในฐานข้อมูล



รูปที่ 3. ขั้นตอนการทำงานของส่วนดึงข้อมูล

- เครื่องมือ get system into รับค่าวันที่ตัวแปร Time คือเวลาที่ทำการดึงข้อมูลไว้ใช้ในการเทียบข้อมูลล่าสุด
- เครื่องมือ select value เพื่อเลือก field ที่ต้องการ
- เขียน javascript ในการกำหนด URL ที่ใช้เรียกตาม Facebook API โดยกำหนด Accresstokenสำหรับการเรียกใช้ Facebook API
- เครื่องมือ http client จะทำการส่ง request ตาม URL ที่รับมา การทำงานต้องรับ Parameter accress_token และ Rest_URL ซึ่งจะเป็นการเรียก Rest_URL&accress_token เช่น https://graph.facebook.com/v2.9/me?fields=id%2Cname&access_token= 3134731157144499% 7CsmWttLSDQymOlY4ECyHkOlL1qFA
- เครื่องมือ JSON input เมื่อส่ง request ไป ข้อมูลที่ได้จะอยู่ในรูปแบบ JSON ซึ่งต้องใช้ JSON input ในการอ่านข้อมูล โดยกำหนด path ของข้อมูลนั้นๆ
- เครื่องมือ spilt filed ใช้ในการแบ่ง filed เช่นข้อมูล id ประกอบด้วย id page ตามด้วย id postดังนั้นต้องแยก field เป็น id_page กับ id_post เพื่อใช้ id_post join กับ table อื่นๆ
- เครื่องมือ replace string ใช้ในการจัดรูปแบบของคำให้ตรงตามความต้องการเช่น ค้นหาคำว่า “แท็กซี่” แล้วให้แทนที่ด้วย แท็กซี่
- เครื่องมือ hadoop file output ใช้ในการระบุที่อยู่ของ file ปลายทางใน Hadoop และระบุ field ที่ต้องการโดยเก็บเป็นไฟล์ CSV

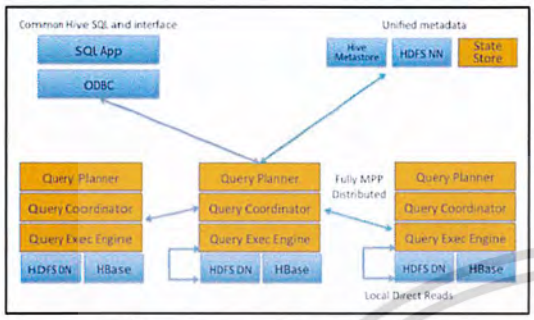
3) ส่วนการจัดเก็บข้อมูล เป็นขั้นตอนที่จัดเก็บข้อมูลขนาดใหญ่ด้วย Hadoop มี HDFS ทำหน้าที่เป็นระบบจัดเก็บข้อมูลหลัก โดย HDFS จะสร้างแบบจำลองเป็นบล็อกของข้อมูลบนคลัสเตอร์เพื่อให้การคำนวณผลได้รวดเร็ว โดยการจัดเก็บข้อมูลอยู่ในรูปแบบไฟล์ CSV แสดงดังรูปที่ 4



รูปที่4 . ข้อมูล Facebook ที่เก็บใน Hadoop

4) ส่วนค้นหาข้อมูล เป็นขั้นตอนที่เขียน Web Service ด้วยภาษา Java โดยเรียกใช้ผ่านทาง HTTP Method GET เชื่อมต่อกับ Impala โดยใช้ JDBC Driver เมื่อเชื่อมต่อกับ Impala แล้ว จะส่งคำสั่ง SQL ที่ใช้ในการค้นหาเข้ามา Impala โดยจะมีการทำงานดังนี้ Impala Statestone ติดต่อกับ Query Planner ว่าข้อมูลอยู่ที่ไหน เมื่อ Query Planner รู้ที่อยู่ของข้อมูล จะจัดการคำสั่ง SQL ที่เข้ามาว่าร้องขอข้อมูลอะไรบ้างแล้วจะส่งบอก Query Coordinater ให้ส่งคำสั่งค้นหาไปยัง Query Coordinater ของแต่ละโหนด หลังจากนั้นเมื่อแต่ละ

โหนดได้ข้อมูลแล้วจะส่งข้อมูลกลับมาที่ Query Coordinater และส่งกับให้ Web Service แสดงข้อมูลในรูปแบบ JSON แสดงดังรูปที่ 5



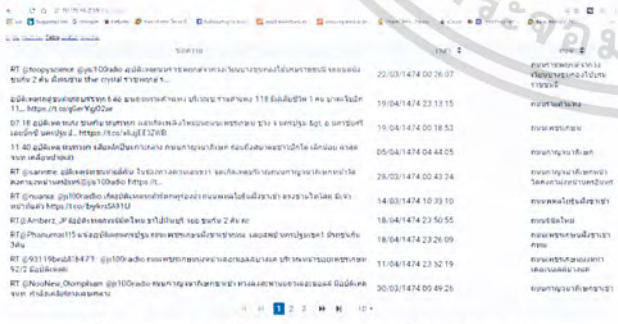
รูปที่ 5. กระบวนการทำงานของ Impala

5) ขั้นตอนการแสดงผล เป็นขั้นตอนที่สร้างเว็บแอปพลิเคชันเพื่อใช้ในการแสดงผลกราฟในรูปแบบต่างๆ จากข้อมูลที่มี พัฒนาโดย Angular

4. ผลการทดลอง

หลังจากได้ทำการศึกษา เรียนรู้ การใช้เทคโนโลยีบิ๊กดาต้า การเก็บข้อมูลจากโซเชียลเน็ตเวิร์คและการทำเว็บแอปพลิเคชันที่พัฒนาโดย Angular ในการนำข้อมูลที่จัดเก็บใน Hadoop มาแสดงผลโดยกราฟรูปแบบต่างๆ แสดงได้ดังรูปที่ 6 ถึงรูปที่ 10

รูปที่ 6 เป็นตารางแสดงรายละเอียดของข้อมูลที่ค้นหาจากข้อความ คำว่า อุบัติเหตุ โดยแสดงรายละเอียดของข้อความที่โพส เวลาที่โพส และถนนที่เกิดเหตุ



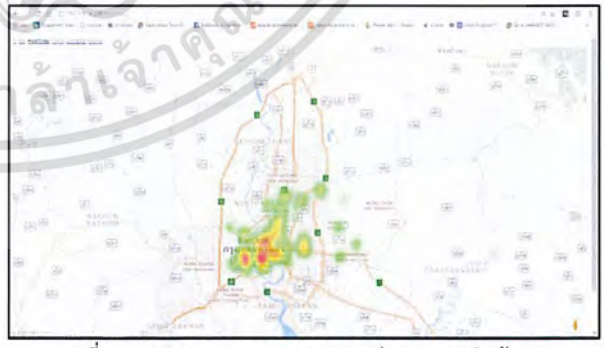
รูปที่ 6. ตารางแสดงข้อความ เวลา ถนน ที่มีการเกิดอุบัติเหตุ

กราฟวงกลมและกราฟแท่งแสดงจำนวนของถนนที่ค้นหาจากข้อความ คำว่า อุบัติเหตุ แสดงดังรูปที่ 7 และ 8

รูปที่ 7. กราฟวงกลมแสดงจำนวนอุบัติเหตุที่เกิดบนถนน

รูปที่ 8. กราฟแท่งแสดงจำนวนอุบัติเหตุที่เกิดบนถนน

Heat map แสดงความหนาแน่นของถนนที่ค้นหาจากข้อความ คำว่า อุบัติเหตุ โดย สีเขียวแสดงถึงเกิดน้อย สีแดงแสดงถึงเกิดมาก แสดงดังรูปที่ 9



รูปที่ 9. Heat map แสดงความหนาแน่นของการเกิดข้อมูล

Cluster map แสดงกลุ่มความหนาแน่นของข้อมูลโดยสามารถระบุคำที่ต้องการค้นหาได้ โดยจัดกลุ่มที่ใกล้เคียงเข้าด้วยกัน สีส้มแสดงถึงกลุ่มการเกิดของข้อมูลมาก สีฟ้าแสดงถึง

กลุ่มของข้อมูลน้อย สีแดงแสดงถึงข้อมูลมีตำแหน่งเดียว แสดงดังรูปที่ 10

รูปที่ 10. Cluster map แสดงกลุ่มความหนาแน่นของการเกิดข้อมูล

5. สรุป

การพัฒนาระบบการค้นคืนข้อมูลจากแหล่งต่างๆ เข้าสู่ Hadoop ด้วย Pentaho เพื่อเก็บข้อมูลที่มีขนาดมหาศาลจาก โซเชียลเน็ตเวิร์คแหล่งต่างๆ เนื่องจากข้อมูลจากโซเชียลเน็ตเวิร์คนั้นเป็นข้อมูลที่มีความหลากหลายทั้งโครงสร้างข้อมูลและแหล่งที่มาของข้อมูล การจัดเก็บข้อมูลจึงมีความสำคัญ จากการพัฒนาระบบจะเห็นได้ว่าก่อนการจัดเก็บข้อมูลต้องศึกษาโครงสร้างของข้อมูลและวิธีในการเข้าถึงข้อมูลก่อนจากนั้นมีการทำให้ข้อมูลมีความถูกต้องและจัดเก็บข้อมูลให้เหมาะสมเพื่อช่วยให้ผู้ใช้ที่ต้องการใช้ข้อมูลสามารถนำข้อมูลมาใช้งานได้อย่างสะดวกและรวดเร็ว

เอกสารอ้างอิง

[1] Cloudera.com, CDH Component, [Online], Available: https://www.cloudera.com/products/open-source/apache-hadoop/key-cdh-components.html, เข้าถึงเมื่อวันที่ 8 กันยายน 2560.

[2] SURANART NIAMCOME, Big Data คืออะไร วิธีใช้ Hadoop/Spark บน Cloud Dataproc, [Online], Available: http://www.siamhtml.com/getting-started-with-big-data-and-hadoop-spark-on-cloud-dataproc/,เข้าถึงเมื่อวันที่ 17 ธันวาคม 2561.

[3] Goingjesse.com, Pentaho Open Source BI, [Online], Available: http://www.goingjesse.com/pentaho-solution, เข้าถึงเมื่อวันที่ 20 ตุลาคม 2560.

[4] Cloudera.com, Impala Concepts and Architecture, [Online], Available: https://www.cloudera.com/documentation/cdh/5-0-x/Impala/Installing-and-Using-Impala/ciiu_concepts.html, เข้าถึงเมื่อวันที่ 17 ธันวาคม 2561.