A Data Analysis Framework Using Cloudera Big Data Platform and RapidMiner Radoop

เนติพงษ์ เทพวุฒิสถาพร, สิทธิพร คุณกรธรรมนพ และ ผศ.ดร.ศุภชัย ไทยเจริญ
หลักสูตรวิทยาศาสตร์บัณฑิต สาขาวิทยาการคอมพิวเตอร์

Abstract

Since real-world data produced in this information era is not only heterogeneous and fast generated but also large in volume, techniques and tools that can exploit them must be sufficiently powerful. Although existing tools such as Hadoop, Hive, and Spark together with some programming languages such as Java or Python are a good combination suite of software tools that can be used for this type of tasks, their complexity requires a high learning curve that substantially hampers the learning progress of non-programmer professionals. Accordingly , to encourage data analytic community, a user-friendly framework for analyzing big data that combines Cloudera platform with Radoop extension of RapidMiner Studio. Cloudera platform unifies a number of key big data software components, such as Hadoop, Hive, and Spark, into a single platform and provides GUI Cloudera manager for managing them. Moreover , RapidMiner Radoop contains a set of operators that facilitates data science and big data analysis tasks much easier with minimal programming. Setup guideline are presented in this paper and examples of data analysis using machine learning are demonstrated and implemented.The results indicated that big data analyses are no longer only accessible for highly-technical professionals,but typical technical people can do it as well.

องค์ความรู้ที่เกี่ยวข้อง

Big data (ข้อมูลขนาดใหญ่)[1]

ชุดข้อมูลที่มีขนาดใหญ่เกินกว่าความสามารถของซอฟต์แวร์ที่ใช้กันอยู่ทั่วไป จะจับบันทึก จัดการ และประมวลผลข้อมูลดังกล่าวได้ภายในเวลาที่ยอมรับได้

Big data analytics (การวิเคราะห์ข้อมูลขนาดใหญ่)[2]

กระบวนการวิเคราะห์เซ็ตข้อมูลขนาดใหญ่เพื่อค้นหารูปแบบความสัมพันธ์ของข้อมูลเหล่านั้นที่ซ่อนอยู่ข้างในหาสิ่งเชื่อมโยงที่เชื่อมข้อมูลเหล่านั้นเข้าไว้ด้วยกัน

Hadoop[3]

ซอฟท์แวร์ประเภท Open source ที่จัดทำขึ้นเพื่อเป็นแพลตฟอร์มในการจัดเก็บข้อมูล มีกรอบการทำงานเพื่อการจัดเก็บข้อมูลและประมวลผล Big data ซึ่งสามารถปรับขยาย ยืดหยุ่น เพื่อรองรับ Big data ได้ เพราะมีการกระบวนการประมวลผลข้อมูลแบบกระจายผ่านเครื่องคอมพิวเตอร์ที่ถูกจัดอยู่ในรูปแบบ Cluster

Cloudera[4]

ถูกพัฒนาต่อจาก Hadoop มาเป็น Platform สำหรับใช้ในการวิเคราะห์และจัดการ Big data ทำงานบนระบบปฏิบัติการCentOS

Hive[5]

เป็นระบบ Data Warehouse ซึ่งสร้างอยู่บน Hadoop ใช้สำหรับการวิเคราะห์ข้อมูล โดยจุดเด่นคือการใช้คำสั่งภาษา SQL ในการเรียกข้อมูล ทั้งที่อยู่ในรูปแบบของ Database และไฟล์บน Hadoop ได้

Mapreduce[6]

แบ่งออกเป็น 1.Map คือการจับคู่ข้อมูลที่มี Key เดียวกันมารวมกัน 2.Reduce คือการรวมค่าของ Key นั้นๆ

Spark[7]

ต่อยอดจาก Mapreduce มีความเร็วในการประมวลผลมากกว่าเนื่องจากทำการประมวลผลบนหน่วยความจำส่วน Mapreduce จะทำการประมวลผลบนพื้นที่เก็บข้อมูล

Yarn[8]

ทำหน้าที่ในการจัดการทรัพยากรให้แต่เครื่องต่างๆที่อยู่ในระบบ

Namenode[9]

คอยทำหน้าที่จัดการและควบคุม Datanode จะไม่มีข้อมูลถูกเก็บในนี้ แต่จะเก็บ Metadata ไว้ในนี้

Datanode[9]

เป็นที่ๆเก็บข้อมูลที่ถูกแบ่งออกเป็นบล็อกมาจาก Namenode

RapidMiner[10]

เป็นซอฟท์แวร์ที่ใช้ในการวิเคราะห์ชุดของข้อมูลตั้งแต่การเตรียมการจนถึงการวิเคราะห์ รวมถึงช่วยให้เห็น Visual Workflow (ขั้นตอนการทำงาน) อย่างเป็นรูปเป็นร่างอีกด้วย

สรุปผล

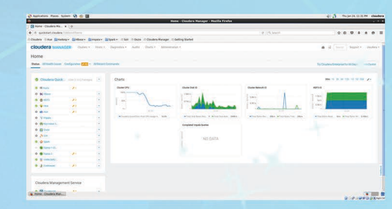
จากการดำเนินงาน ทำให้สรุปผลได้ว่า การวางระบบ Big data system โดยการใช้ Cloudera บนระบบปฏิบัติการ CentOS และมีการใช้ Rapidminer ในการสร้างโมเดลและวิเคราะห์ข้อมูลจากผลลัพธ์มีประสิทธิภาพและระยะเวลาในการประมวลผลที่เทียบเท่ากับอุปกรณ์คอมพิวเตอร์ระดับสูงและมีราคาแพง จำพวกเครื่อง server อีกทั้งจากระบบ Cloudera ที่มี Namenode และ Datanode โดยที่ Datanode สามารถที่มีจะนำมาต่อเพิ่มได้ ทำให้สามารถเพิ่มประสิทธิภาพและช่วยลดระยะเวลาได้ ซึ่งจะเห็นได้ว่าระบบของนี้มีประสิทธิภาพที่เทียบเท่าอุปกรณ์ระดับสูง รวมทั้งยังมีต้นทุนที่น้อยกว่าอีกด้วย ดังนั้นแล้ว ระบบนี้จึงเหมาะสมกับบริษัทขนาดกลางและขนาดเล็กที่มีต้นทุนในการทำงานที่ไม่สูงมากได้ รวมไปยังบุคคลทั่วไปก็สามารถเข้าถึงได้เช่นกัน

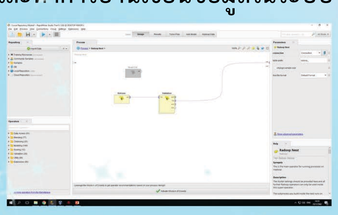
ภาพรวมการดำเนินงาน

วิธีการดำเนินงาน
1.ศึกษาข้อมูลเกี่ยวกับ Big Data, Hadoop และ RapidMiner
2.วิเคราะห์และออกแบบระบบ
โดยจะติดตั้งแบบ 3 เครื่อง จะมี 1 Namenode และ 2 Datanode
3.เตรียมเครื่องมือเพื่อสร้างและพัฒนา
4.ทดสอบการใช้งาน
5.ประเมินผลการใช้งาน

การทดลอง

ทำการทดสอบด้วยการเปิดใช้งานระบบ และทำการอ่านเขียนข้อมูลในระบบ





สามารถเปิดใช้งานระบบได้จริง และสามารถนำข้อมูลขึ้นไปเก็บในระบบและนำออกมาใช้งานได้จริง

อ้างอิง

[1] ข้อมูลขนาดใหญ่. สืบค้นเมื่อ 5 มีนาคม 2562 จาก https://th.wikipedia.org/wiki/ข้อมูลขนาดใหญ่
[2] Big data analytics สำคัญยังไงและช่วยอะไรเราได้บ้าง?. สืบค้นเมื่อ 5 มีนาคม 2562 จาก http://bigdataexperience.org/what-is-big-data-analytics/
[3] มารู้จัก Hadoop เครื่องมือเซเลบในโลกของ Big data. สืบค้นเมื่อ 5 มีนาคม 2562 จาก http://bigdataexperience.org/what-to-know-about-hadoop/
[4] Cloudera. สืบค้นเมื่อ 5 มีนาคม 2562 จาก https://www.cloudera.com/
[5] Ambari #05 การดึงข้อมูลเข้าจาก MySQL เข้าสู่Hive ด้วย Sqoop. สืบค้นเมื่อ 5 มีนาคม 2562 จาก https://sysadmin.psu.ac.th/2017/10/24/ambari-05-sqoop-mysql-hive/
[6] MapReduce คืออะไร?. สืบค้นเมื่อ 5 มีนาคม 2562 จาก https://medium.com/@aorjoa/mapreduce-คืออะไร-62c33f8d9923
[7] Spark 101: What Is It, What It Does, and Why It Matters. สืบค้นเมื่อ 5 มีนาคม 2562 จาก https://mapr.com/blog/spark-101-what-it-what-it-does-and-why-it-matters/
[8] Apache Hadoop YARN. สืบค้นเมื่อ 5 มีนาคม 2562 จาก https://searchdatamanagement.techtarget.com/defifinition/Apache-Hadoop-YARN-Yet-Another-Resource-Negotiator
[9] What is the Difference Between NameNode and DataNode in Hadoop. สืบค้นเมื่อ 5 มีนาคม 2562 จาก http://pediaa.com/what-is-the-difference-between-namenode-and-datanode-in-hadoop/
[10] Rapidminer. สืบค้นเมื่อ 5 มีนาคม 2562 จาก https://rapidminer.com