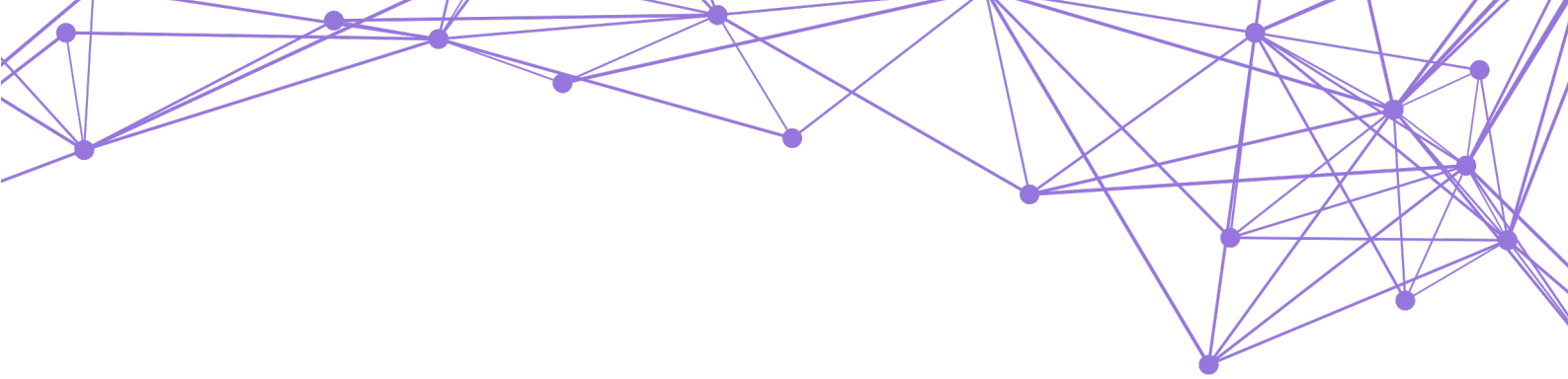
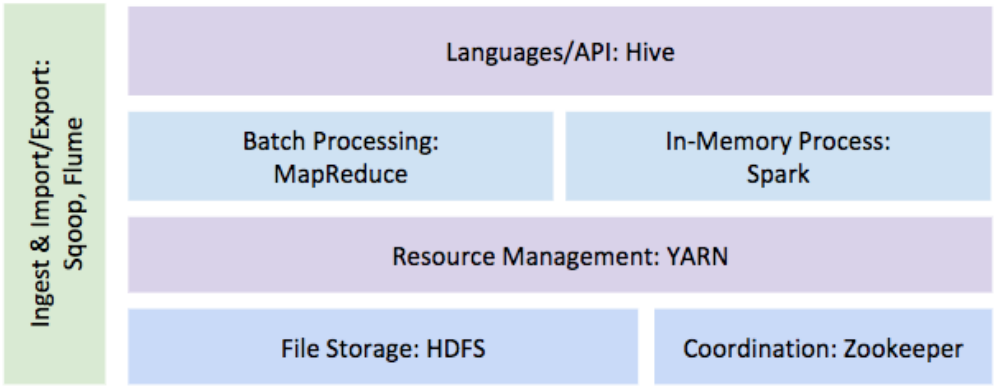


SOFTWARE SPECIFICATION

INTELLIGIST BIG DATA PLATFORM
Version 1.0

September 4, 2017

1. Component



1.1 HDFS (Hadoop Distributed File System) คือระบบจัดการไฟล์ของ Hadoop โดยต้องมีหนึ่งเครื่องเป็น namenode สำหรับเก็บตำแหน่งข้อมูล ส่วนเครื่องอื่นๆ ที่ใช้เก็บข้อมูลจะเรียกว่า datanode และจะมีการสำรองเครื่องไว้เพื่อป้องกันบาง node ล่มไป

1.2 YARN คือตัวจัดการ cluster ในการจัดการ job ที่เกิดขึ้น และยังจัดการ Resource Manager และ Node Manager ดูแลเรื่อง CPU, Memory, Disk, Network

1.3 Mapreduce คือ Distributed Programming Framework สำหรับข้อมูลขนาดใหญ่ และเป็น Batch processing มีการแบ่งการทำงานเป็น 2 ส่วนคือ Map กับ Reduce โดยข้อมูลจะอยู่ในรูปแบบ key-value

1.4 Hive คือเครื่องมือที่ใช้เตรียมข้อมูลที่เป็นลักษณะของคลังข้อมูลบน Hadoop โดยมีการกำหนด Schema เตรียมไว้สำหรับการ Query โดยใช้ภาษาที่เรียกว่า HiveQL ซึ่งมีลักษณะคล้ายกับภาษา SQL

1.5 Sqoop คือเครื่องมือที่ทำหน้าที่ในการ Transfer ข้อมูลจากระบบฐานข้อมูลอื่นๆ เช่น Oracle, SQL Server, MySQL เข้ามาเก็บในรูปแบบ HDFS ซึ่งต้องใช้ ODBC/JDBC ในการเชื่อมต่อไปยังระบบฐานข้อมูล

1.6 Zookeeper คือเครื่องมือที่ใช้ในการ replicate ข้อมูลจาก server ต่างๆซึ่งมีกระบวนการทำงานแบบ distribution

1.7 Flume คือเครื่องมือที่ช่วยในการจัดเก็บข้อมูล Log Collector, Event Log หรือ Sensor จากแหล่งต่างๆ โดยทำการการเคลื่อนย้ายข้อมูล จากต้นทางเพื่อทำการไหลข้อมูลผ่านทาง Channel ซึ่งทำงานบน Memory ไปเก็บไว้บน HDFS

1.8 Spark คือเครื่องมือในการประมวลผลแบบ in-memory cluster ที่มีความรวดเร็วและรองรับ API ที่พัฒนาจากภาษา Java, Scala และ Python

1.9 Zeppelin notebook คือเครื่องมือสำหรับช่วยประมวลผลข้อมูลบน HDFS, Hive, Spark เป็นต้น โดยจะทำงานแบบ interpreter นอกจากนี้ยังสามารถ Query และทำ Graph ผ่าน Zeppelin ได้

2. หลักการทำงานของระบบ

2.1 ระบบทำการดึงข้อมูลจาก Datasource (Oracle, MySQL, SQL Server, File System หรืออื่นๆ) เข้ามาในระบบโดยใช้ Component ที่ชื่อว่า Flume และ Sqoop

2.2 เมื่อ Flume และ Sqoop ได้ทำการดึง Datasource มาเรียบร้อยแล้ว Namenode จะทำหน้าที่ในการตัดข้อมูลที่ถูกดึงเข้ามาให้อยู่ในรูปแบบ Block และนำไปเก็บที่ Datanode ตามที่ Namenode ได้ทำการเก็บตำแหน่ง Block ไว้ หรือก็คือการเก็บข้อมูลการเก็บข้อมูลลงบน HDFS

2.3 เมื่อต้องการนำข้อมูลไปใช้ ข้อมูลจะถูกส่งไปทำ ETL และถูก Hive เรียกข้อมูลเพื่อนำข้อมูลไปประมวลผลและทำให้อยู่ในรูปแบบตาราง

2.4 Hive จะมี API สำหรับการเรียกใช้จาก Tools ภายนอกเช่น BI Tools หรือ Analytics Tools ต่างๆ

3. คุณลักษณะของระบบ

3.1 Web-based UI

- รองรับการ monitor การทำงานของแต่ละ node, Resource รวมไปถึงสถานะของ component
- รองรับการจัดการกับ Component ต่างๆ เพื่อเพิ่มประสิทธิภาพของระบบได้โดยง่าย

3.2 ระบบรองรับการจัดเก็บข้อมูล

- ข้อมูลที่มีโครงสร้าง (Structured data) เช่น Oracle, Mysql, MS SQL Server, PostgreSQL รวมไปถึงการเชื่อมโยงกับระบบงานภายในองค์กร เป็นต้น
- ข้อมูลกึ่งโครงสร้าง (Semi-Structured data) เช่น log file, text file, csv, xlsx, xls เป็นต้น
- ข้อมูลแบบไม่มีโครงสร้าง (Unstructure) เช่น MongoDB, Social Media เป็นต้น

3.3 สามารถจัดเก็บข้อมูลในระบบ Hadoop และเรียกใช้ข้อมูลที่จัดเก็บในรูปแบบ ของระบบฐานข้อมูลเชิงสัมพันธ์ (RDBMS) โดยจัดเก็บในรูปแบบมาตรฐานของ Hadoop File System

3.4 ระบบบรองรับการลากวาง (Drag and Drop) ในการ Upload ข้อมูลผ่าน Graphic User Interface

3.5 ระบบบรองรับการแลกเปลี่ยนข้อมูลระหว่างกัน โดยสามารถแยกการให้บริการและสิทธิในการเข้าถึงข้อมูลของแต่ละบุคคล รวมไปถึงการตรวจสอบสิทธิ์ในการนำเข้าข้อมูล (Upload)

3.6 ระบบบรองรับการนำข้อมูลเข้าผ่าน API Gateway และมีการทำโครงสร้างของไฟล์ (Metadata) ซึ่งการทำงานจะมีแต่ละ Component จัดการการนำเข้าข้อมูลต่างชนิด รวมไปถึงการเก็บ Log การทำงานของระบบ

3.7 ระบบบรองรับการดึงข้อมูลและการประมวลผลข้อมูลจาก Apache Hadoop โดยใช้ภาษา ANSI-92 SQL เป็นอย่างน้อยได้

3.8 ระบบบรองรับการเก็บและการประมวลผลข้อมูลเป็นภาษาไทยได้

3.9 รองรับการนำโปรแกรมที่พัฒนาด้วยภาษา R, Python, Spark ให้ทำงาน และวิเคราะห์ข้อมูลบน Hadoop ได้

3.10 สามารถวิเคราะห์ข้อมูลที่เป็นลักษณะเป็นข้อความ (Text Analytics) ด้วยเครื่องมือทำงานในรูปแบบเว็บ (Web-based Tool)

3.11 ความสามารถในการจัดการฐานข้อมูล คุณสมบัติดังนี้

- มีโมดูลระบบจัดการฐานข้อมูล Structered Data แบบตาราง Columnar
- มีการป้องกันการเสียหายของข้อมูลโดยการจัดเก็บแบบกระจาย

3.12 ความสามารถในการเรียกใช้ฐานข้อมูล คุณสมบัติดังนี้

- สามารถเรียกใช้ฐานข้อมูลได้ด้วยภาษา SQL
- การจัดการข้อมูลด้วยภาษา SQL สามารถจัดการกับข้อมูลได้โดยตรง โดยไม่ต้องทำการเคลื่อนย้ายข้อมูลหรือเปลี่ยนแปลงก่อนทำการประมวลผล
- สามารถอ่าน metadata, ODBC driver and SQL syntax from Apache Hive

3.13 ระบบบรองรับมาตรฐานการเชื่อมต่อแบบ ODBC/JDBC

3.14 ระบบบรองรับการเชื่อมต่อจาก BI Tools และ Analytics Tools เช่น Tableau, Power BI, Knime เป็นต้น

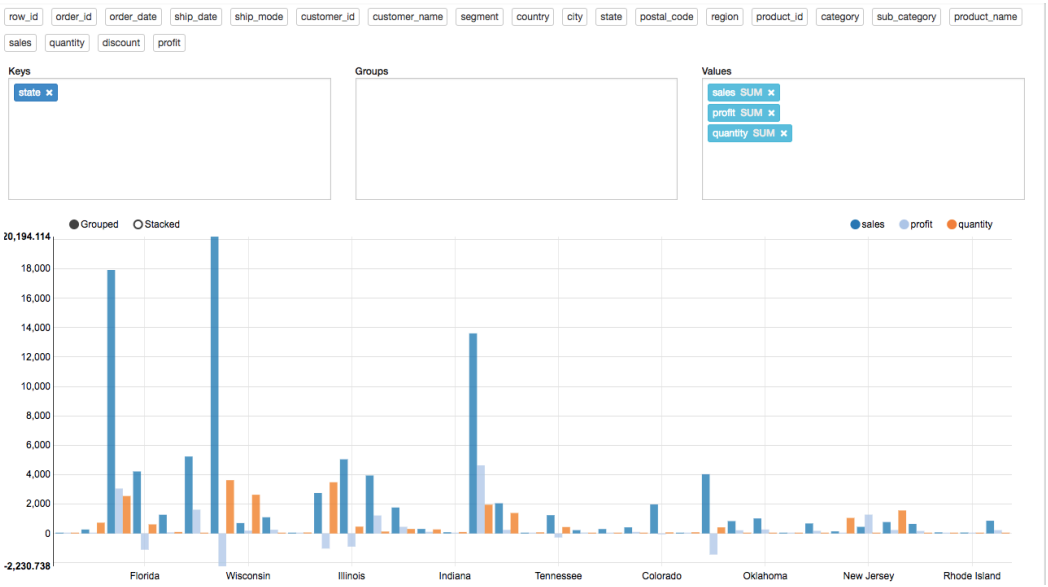
3.15 ระบบรองรับการทำ Machine Learning โดยใช้ภาษา R, Python, Spark เป็นต้น

```
%spark2.r                                                                    FINISHED
df<-read.df('/tmp/wdbc.data', source='csv', header='true');
df1<-as.data.frame(df);
d1<-df1[,2:32]
d1$Diag[d1$Diagnosis=="M"]<-1
d1$Diag[d1$Diagnosis=="B"]<-0
d1$Diag<-as.numeric(d1$Diag)
d1$radius1<-as.numeric(d1$radius1)
d1$radius2<-as.numeric(d1$radius2)
d1$radius3<-as.numeric(d1$radius3)
library(caTools)
s=sample.split(d1$Diag, SplitRatio=0.7)
train<-subset(d1,s==TRUE)
test<-subset(d1,s==FALSE)
m<-glm(Diag~radius1+radius2+radius3,family=binomial(link='logit'),data=train)
nd<-data.frame(radius1=0,radius2=0.00001,radius3=0.0001)
Prob<-predict(m,newdata=test,type="response")
Prob[which(Prob <= 0.9)]<-0
Prob[which(Prob > 0.9)]<-1
table(test$Diag, Prob)

Prob
    0   1
  0 107   0
  1  17  47

Took 6 sec. Last updated by admin at July 20 2018, 4:49:36 PM.
```

3.16 ระบบรองรับการวิเคราะห์ข้อมูลเบื้องต้น เช่น การทำกราฟ โดยการลากวาง (Drag and Drop)



3.17 รองรับเทคโนโลยี CPO (CPO Technology) ซึ่งเป็นเทคโนโลยี การเปรียบเทียบประสิทธิภาพ การวิเคราะห์ ข้อความ โดยใช้กระบวนการประมวลผลธรรมชาติ เพื่อเตรียมข้อมูลสำหรับทำ Social Network Analytic ที่มีความซับซ้อนได้ มีคุณสมบัติดังนี้

- มี API สำหรับดึงข้อมูลจาก Social Media Platform และสามารถกำหนดคำสืบค้น (Keyword) เพื่อสนับสนุนการวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพ
- รองรับการเชื่อมต่อกับแหล่งข้อมูลที่หลากหลาย ทั้งแบบมีโครงสร้าง(Structured) และแบบไร้โครงสร้าง (Unstructured)
- ใช้กระบวนการทำเหมืองข้อความ (Text Mining) เพื่อเปลี่ยนข้อความที่มีโครงสร้างประโยคที่ไม่แน่นอน (Unstructured Data) ให้อยู่ในรูปแบบข้อมูลที่มีโครงสร้างประโยคที่แน่นอน (Structure Data) เพื่อใช้ ในการสร้างตัวแบบการเรียนรู้

- สามารถนำเข้าข้อมูลที่อยู่ในรูปแบบต่างๆ เช่น Spreadsheet/Microsoft Excel, Text File หรือ CSV
- สามารถส่งออกข้อมูลในรูปแบบต่างๆ เช่น Spreadsheet/ Microsoft Excel, Text File หรือ CSV
- สนับสนุนข้อมูลภาษาไทย ตั้งแต่การนำเข้าข้อมูล การวิเคราะห์ข้อมูล และ ส่งออกข้อมูล
- ใช้กระบวนการประมวลผลภาษาธรรมชาติ เพื่อเตรียมข้อมูล (Data Preparation) ให้มีความถูกต้องสมบูรณ์ก่อนนำไปใช้วิเคราะห์ โดยไม่เปลี่ยน/แก้ไข/ส่งผลกระทบกับข้อมูลนำ เข้า โดยใช้วิธีในการเตรียมข้อมูล ดังนี้
    - การตัดคำ (Word Segmentation)
    - การกำจัดคำหยุด (Stop-Word List Removal)
    - การแทนข้อความ (Text Representation)
- มี package สำหรับตัดคำเพื่อปรับรูปแบบข้อมูลให้อยู่ในรูปฟีเจอร์เวกเตอร์
- มี package สำหรับนำคำที่ไม่มีนัยสำคัญออก โดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลง เพื่อกำจัดคุณลักษณะที่ไม่เป็นประโยชน์และลดขนาดของดัชนีลง
- คำนวณค่าดัชนีของคำในเอกสาร (Term Weighting) เพื่อสร้างตัวแทนเนื้อหาเอกสาร (Document Representation) สำหรับใช้ในกระบวนการเรียนรู้ของเครื่องมือการเรียนรู้ ด้วย TF-IDF
- สามารถทำการวิเคราะห์ข้อมูลโดยใช้เทคนิค Classification Algorithms
- มี package สำหรับการวิเคราะห์ข้อมูลโดยใช้เทคนิค Classification Algorithms เช่น Support Vector Machine (SVM)
- สามารถแก้ไขปัญหาความไม่สมดุลของข้อมูล (Class Imbalance) ด้วยวิธีการสุ่มข้อมูลเพิ่มต่างๆ (Oversampling Technique) ได้แก่ SMOTE, Borderline-SMOTE, ADASYN และ Safe-level SMOTE เป็นต้น
- มี package สำหรับแก้ปัญหาความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มข้อมูลเพิ่มต่าง ๆ ได้แก่ SMOTE, Borderline-SMOTE, ADASYN และ Safe-level SMOTE เป็นต้น
- สามารถให้คะแนน Dataset ที่ยังไม่เคยถูกนำเข้าระบบมาก่อนด้วย Trained Model
- มีเทคนิคการทดสอบความถูกต้องของโมเดล ด้วย Cross Validation เช่น Holdout method
- มีตัวชี้วัดการประเมินความแม่นยำของการทำโมเดล Classification ด้วยค่าต่างๆ ดังนี้ Accuracy, Precision, Recall, False Positive, False Negative, True Positive, True Negative