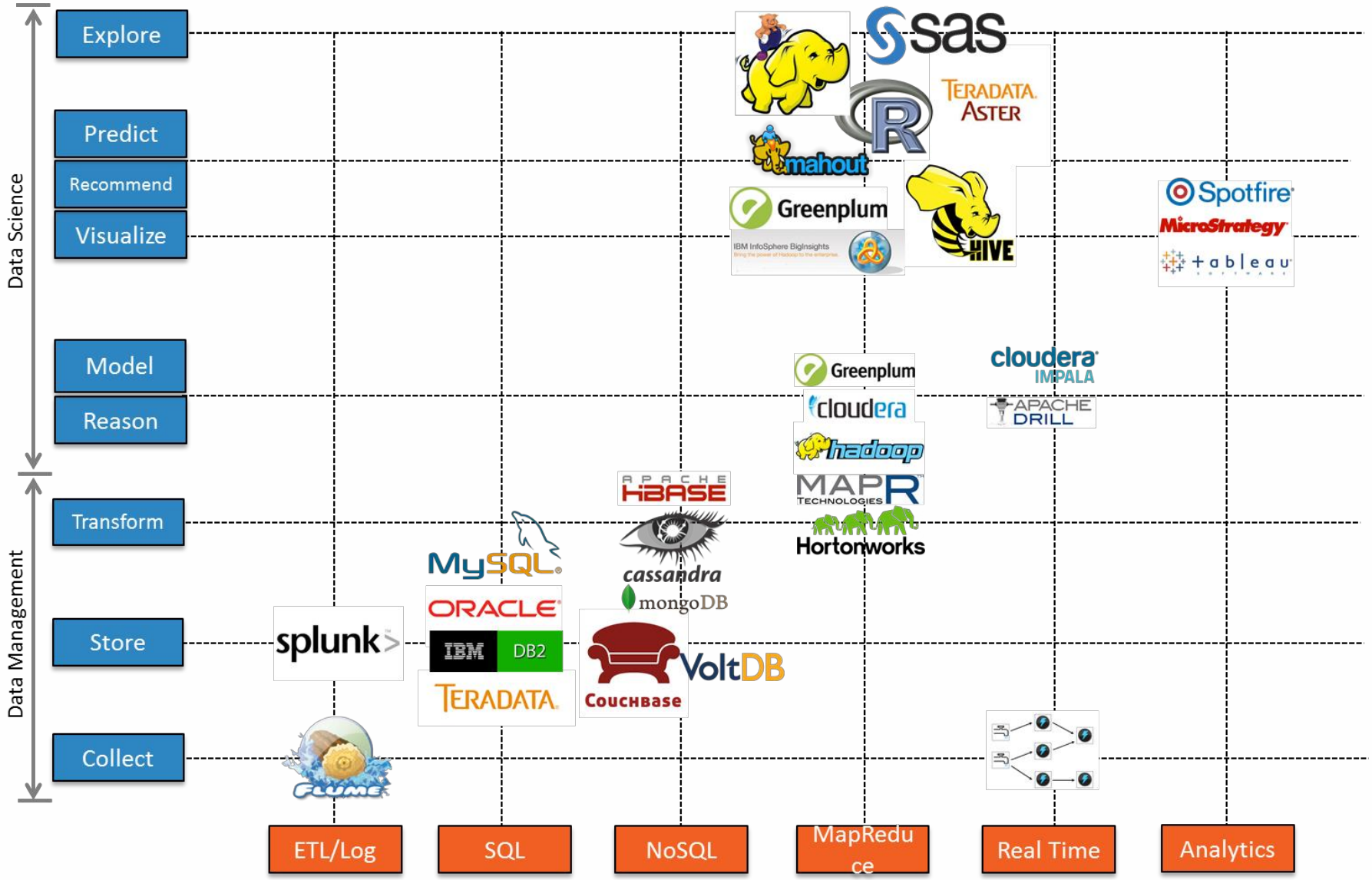
Basic

Pasit Yodsoi

Data Engineer

Big Data Technology



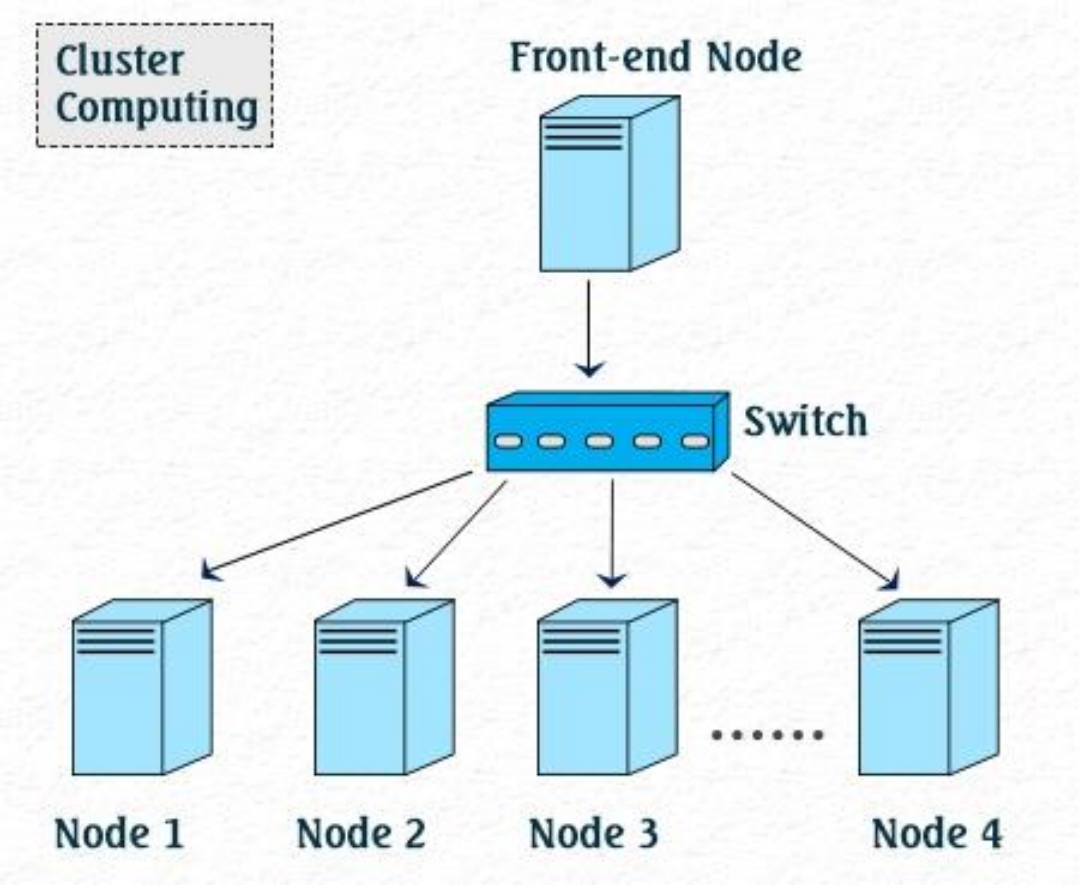
Big Data – Technology, Platforms & Products

NoSQL

MongoDB เป็น open-source document database โดยเป็นฐานข้อมูลแบบ NoSQL คือ ไม่มี relation (ความสัมพันธ์) ของตารางแบบ SQL ทั่วๆไป แต่จะเก็บข้อมูลเป็นแบบ JSON (JavaScript Object Notation) แทน การบันทึกข้อมูลทุกๆ record ใน MongoDB เราจะเรียกมันว่า Document ซึ่งจะเก็บค่าเป็น key และ value จะเห็นว่ามันก็คือ JSON

“Big Data ไม่ใช่ Hadoop
แต่ Hadoop เป็นส่วนหนึ่ง
ของ Big Data”

Hadoop Concept



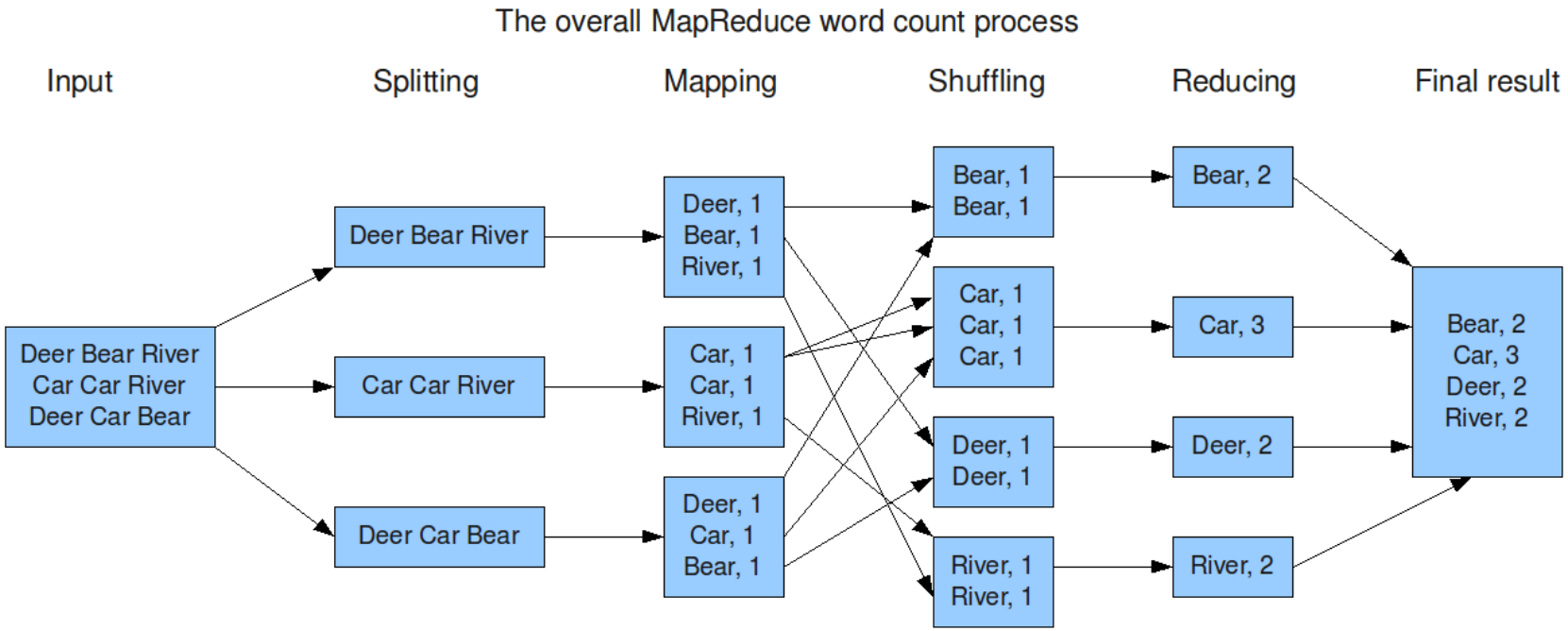
Hadoop Ecosystem

- Apache Hadoop(HDFS + YARN)
- Map Reduce
- Apache Hive
- Spark
- Cloudera Impala
- Sqoop
- Hbase
- Apache Ambari

Apache Hadoop (HDFS + YARN)

- เป็นพื้นฐาน ของ Big Data Tools โดยเฉพาะ HDFS ที่เป็น File System ที่อยู่บน Distributed system

Map Reduce



- เป็น programming model ที่ช่วยในการใช้ทรัพยากรให้คุ้มค่า โดยระบบจะทำการกระจาย Task ไปรันแบบ Parallel บนเครื่องหลายๆ เครื่อง (ลดขั้นตอนการดำเนินการ)

Apache Hive

- เป็นเครื่องมือสำหรับผู้ต้องการสืบค้น (Query) ข้อมูลที่เก็บใน HDFS ด้วยภาษาลักษณะ SQL โดย Hive จะทำหน้าที่ในการแปล SQL like ให้มาเป็น Map/Reduce แล้วก็ทำการรันแบบ Batch

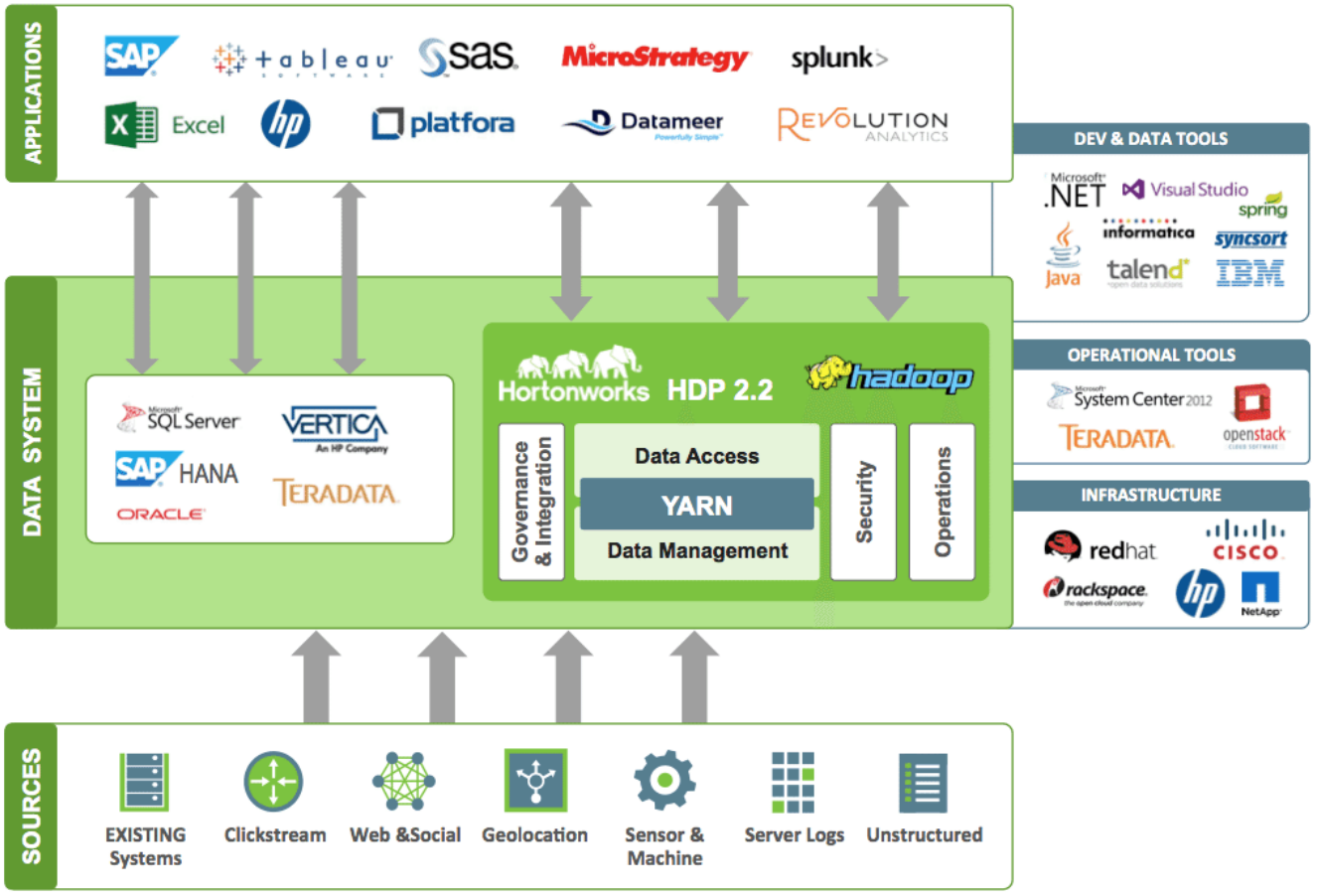
Spark

- เป็น Data Processing Framework ตัวหนึ่งที่นิยมกันแพร่หลาย Spark ทำงานได้รวดเร็วกว่าตัว Hadoop เพราะ Hadoop ทำงานบน Disk แต่ Spark ทำงานบน Memory

Sqoop

- เป็น Framework ที่จัดการการถ่ายโอนข้อมูลระหว่างข้อมูลรูปแบบ Table บน RDBMS กับข้อมูลรูปแบบ HDFS บน Hadoop

Hortonworks

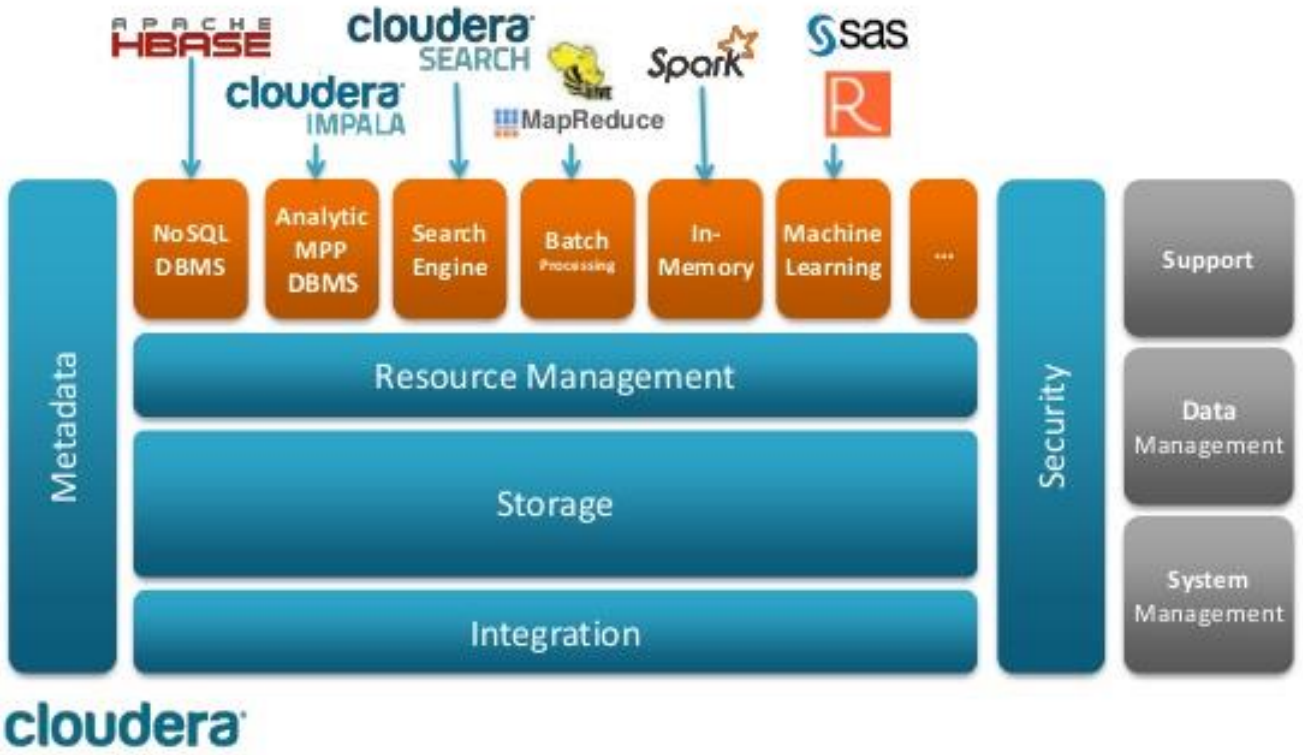


Apache Ambari

- โปรแกรมบริหาร Cluster ทีเป็น Opensource ช่วยทำให้เพิ่มประสิทธิภาพในการบริหาร Server

Cloudera

CDH: the App Store for Hadoop



Basic Command

hdfs dfs -ls /user = List file in hdfs

hdfs dfs -put /tmp/file.csv /user/path/ = Push file to hdfs

Create table

```
CREATE TABLE `table_name `(
   `col1` string,
   `col2` string,
   `col3` string,                                  )
 ROW FORMAT SERDE
   'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'
 WITH SERDEPROPERTIES (
   'field.delim'='|',
   'serialization.format'='|')
 STORED AS INPUTFORMAT
   'org.apache.hadoop.mapred.TextInputFormat'
 OUTPUTFORMAT
   'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
 LOCATION
   'hdfs://BIGDATA/user/hive/warehouse/default/table_name'
 TBLPROPERTIES ("skip.header.line.count"="1");
```