เอกสารประกอบการทำกิจกรรม COP

Basic

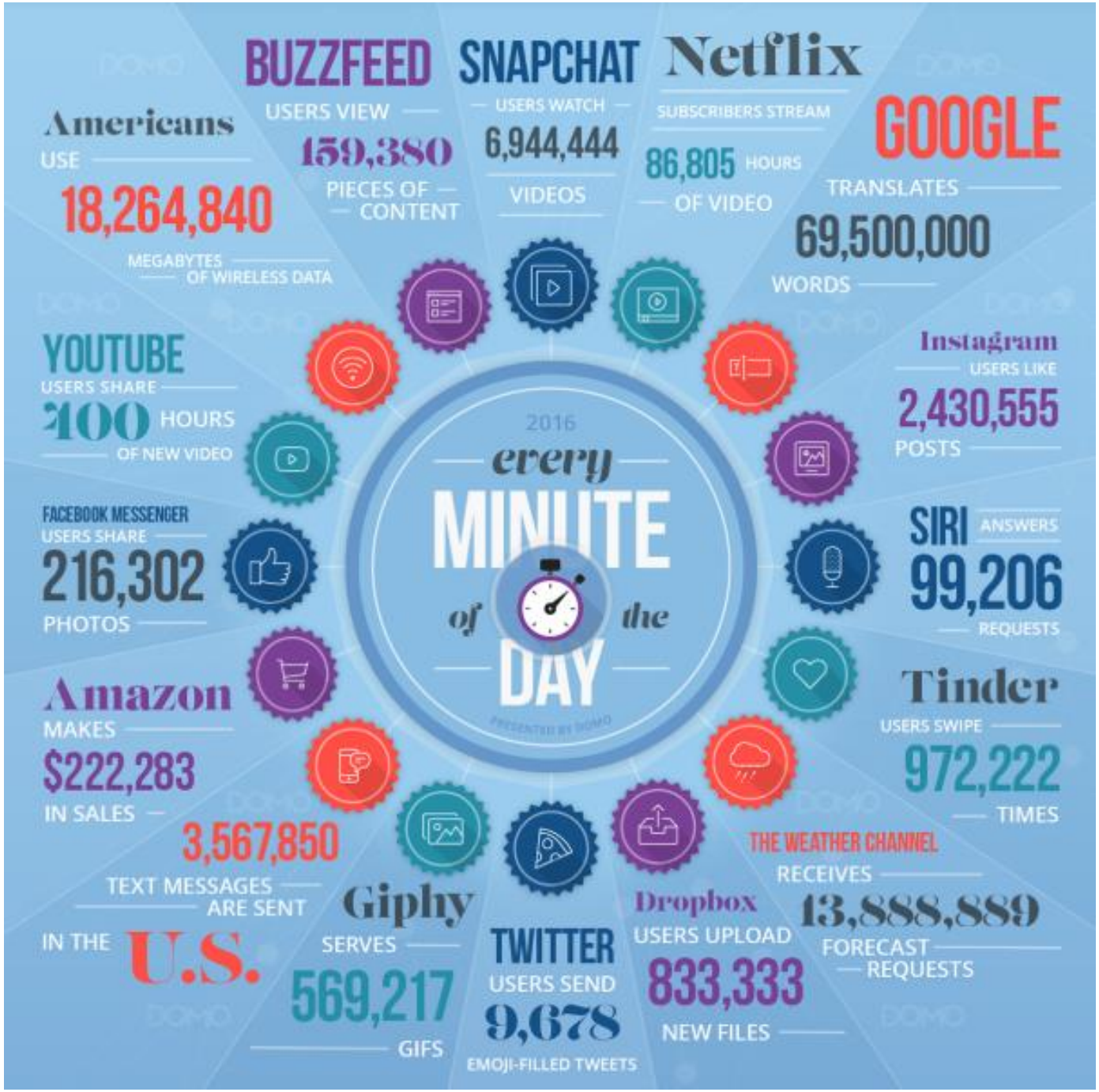
ศส.กรอ.

Basic

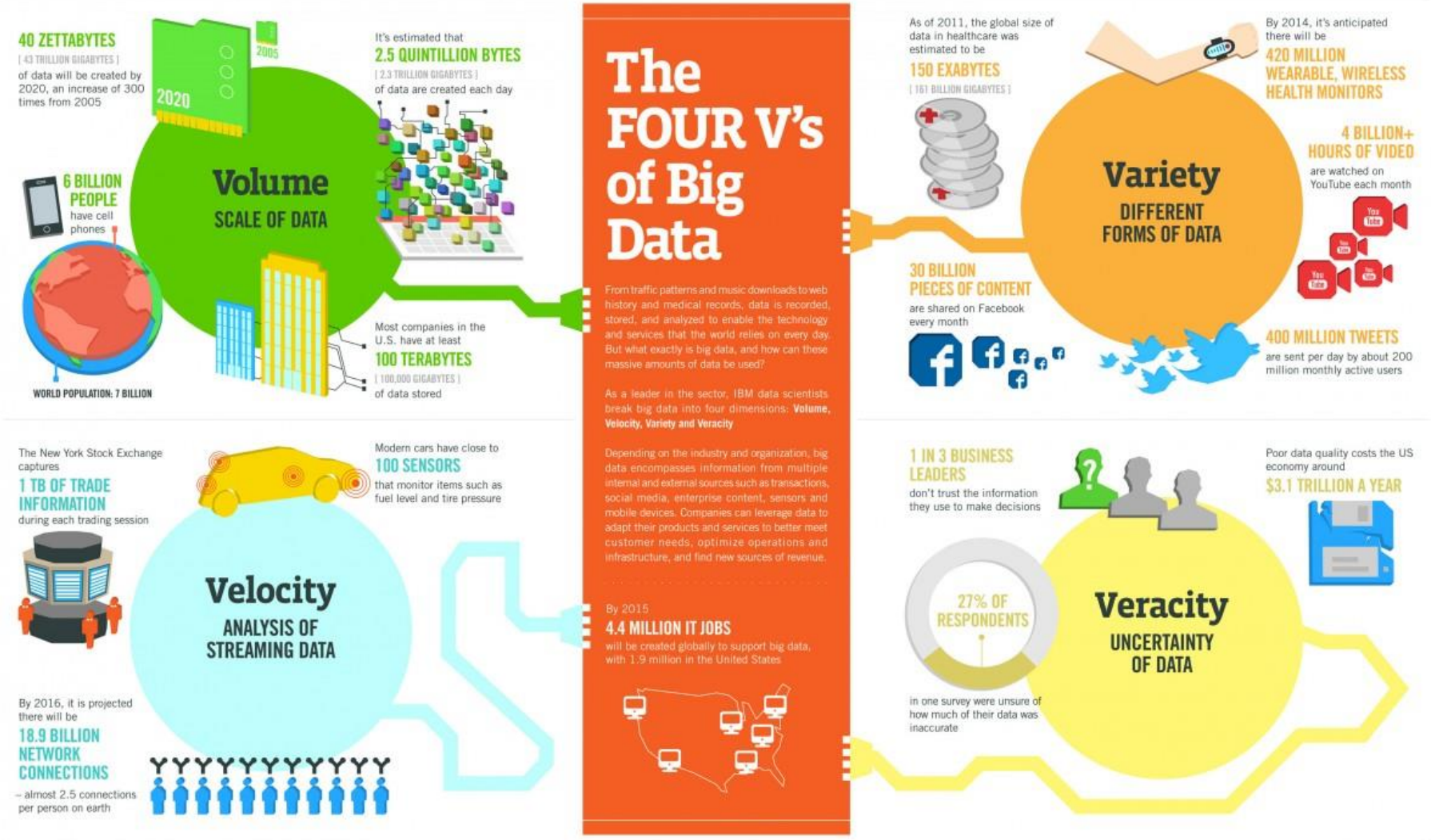
Narongrich Chanprasert

BI Consultant

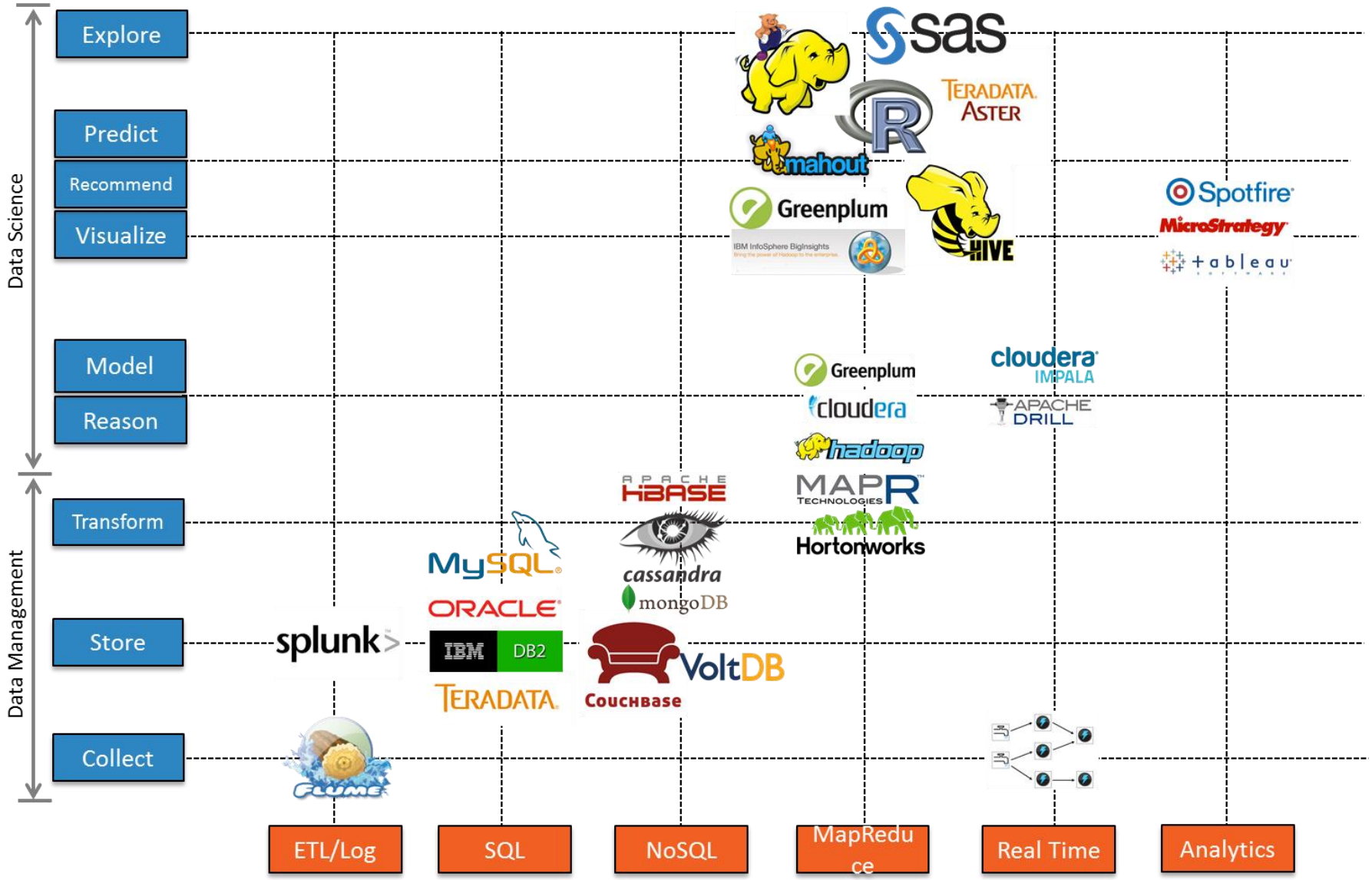
Depth First Co.,Ltd.


2016
every
MINUTE
of the
DAY
PRESENTED BY DOMO
BUZZFEED
USERS VIEW
159,380
PIECES OF CONTENT
SNAPCHAT
USERS WATCH
6,944,444
VIDEOS
Netflix
SUBSCRIBERS STREAM
86,805 HOURS
OF VIDEO
GOOGLE
TRANSLATES
69,500,000
WORDS
Americans
USE
18,264,840
MEGABYTES OF WIRELESS DATA
YOUTUBE
USERS SHARE
400 HOURS
OF NEW VIDEO
Instagram
USERS LIKE
2,430,555
POSTS
FACEBOOK MESSENGER
USERS SHARE
216,302
PHOTOS
SIRI ANSWERS
99,206
REQUESTS
Amazon
MAKES
$222,283
IN SALES
Tinder
USERS SWIPE
972,222
TIMES
3,567,850
TEXT MESSAGES ARE SENT
IN THE U.S.
Giphy
SERVES
569,217
GIFS
TWITTER
USERS SEND
9,678
EMOJI-FILLED TWEETS
Dropbox
USERS UPLOAD
833,333
NEW FILES
THE WEATHER CHANNEL
RECEIVES
13,888,889
FORECAST REQUESTS

The 4V of Big Data



Big Data Technology

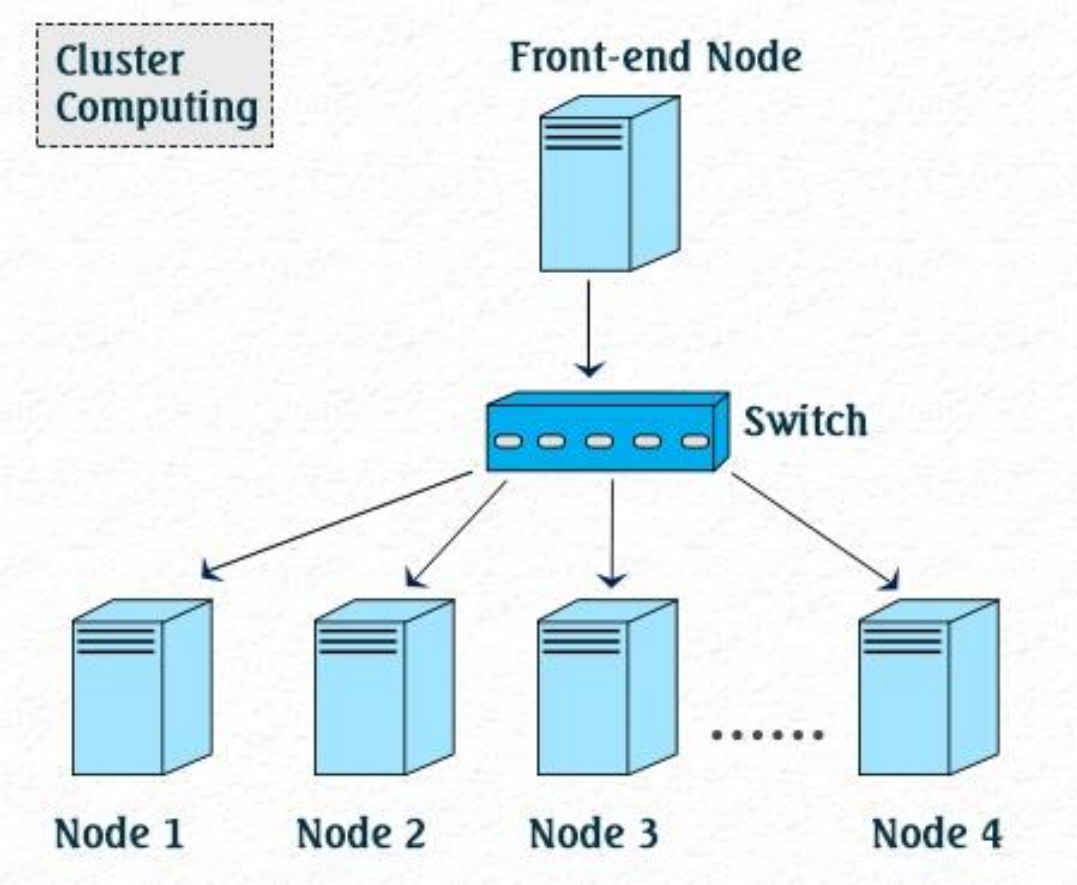


NoSQL

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



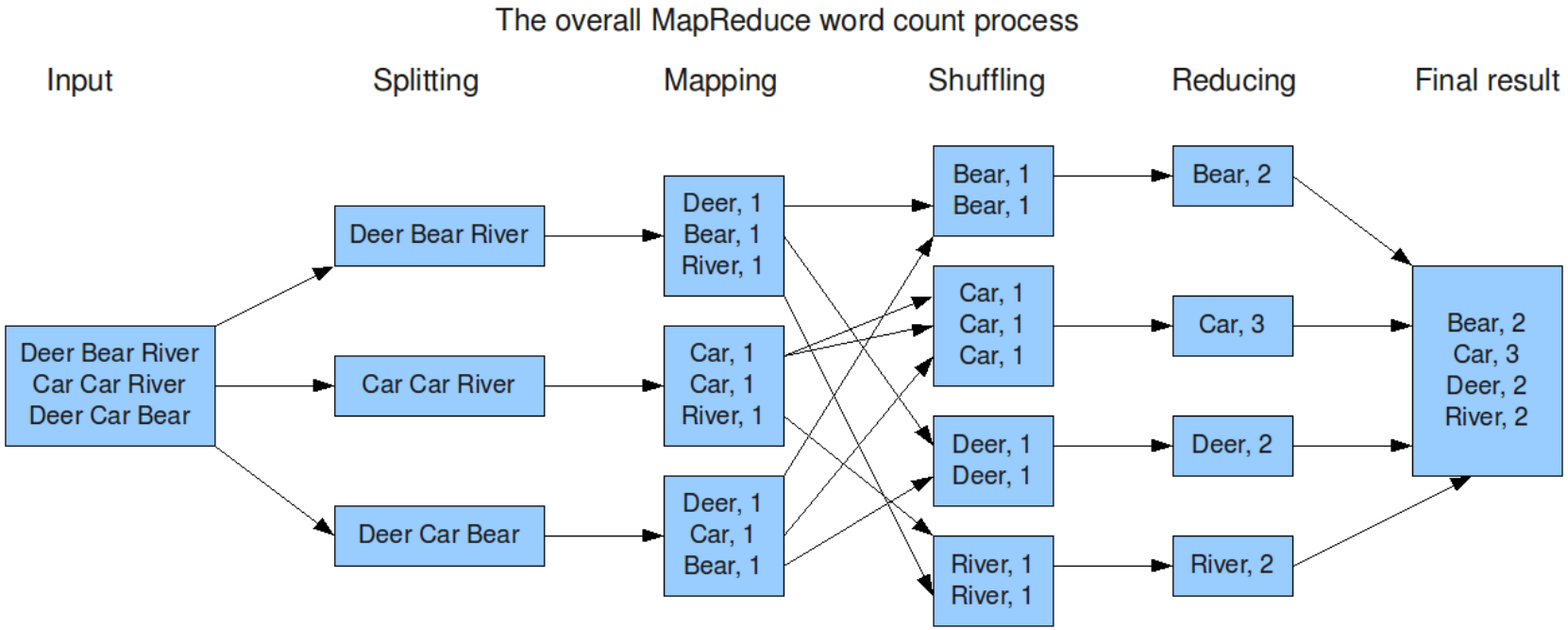
Hadoop Ecosystem

- 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

Cloudera Impala

- เป็นเครื่องมือสำหรับผู้ต้องการสืบค้น (Query) มีการทำงานคล้ายๆ Hive แต่ที่แตกต่างคือ Impala จะทำงานกับข้อมูลที่อยู่บน Memory ซึ่งแน่นอนว่าต้องเร็วกว่าแต่มันก็ยังมีข้อจำกัดในเรื่อง Memory

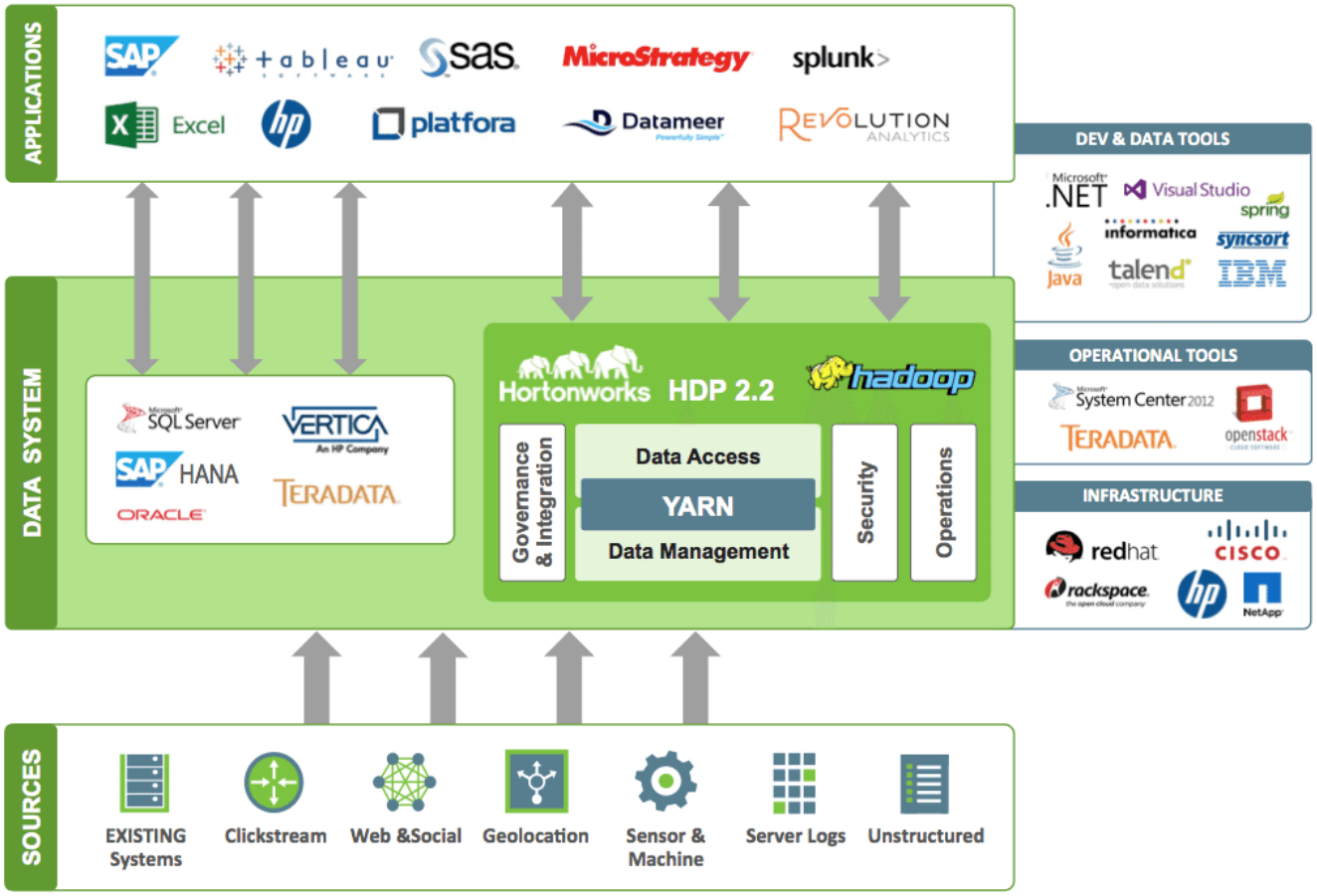
Sqoop

- เป็น Framework ที่จัดการการถ่ายโอนข้อมูลระหว่างข้อมูลรูปแบบ Table บน RDBMS กับข้อมูลรูปแบบ HDFS บน Hadoop

Apache HBase

- เป็น DBMS ที่อยู่บน Hadoop เหมาะสำหรับการ Read/Write ข้อมูลที่เป็น Realtime

Hortonworks

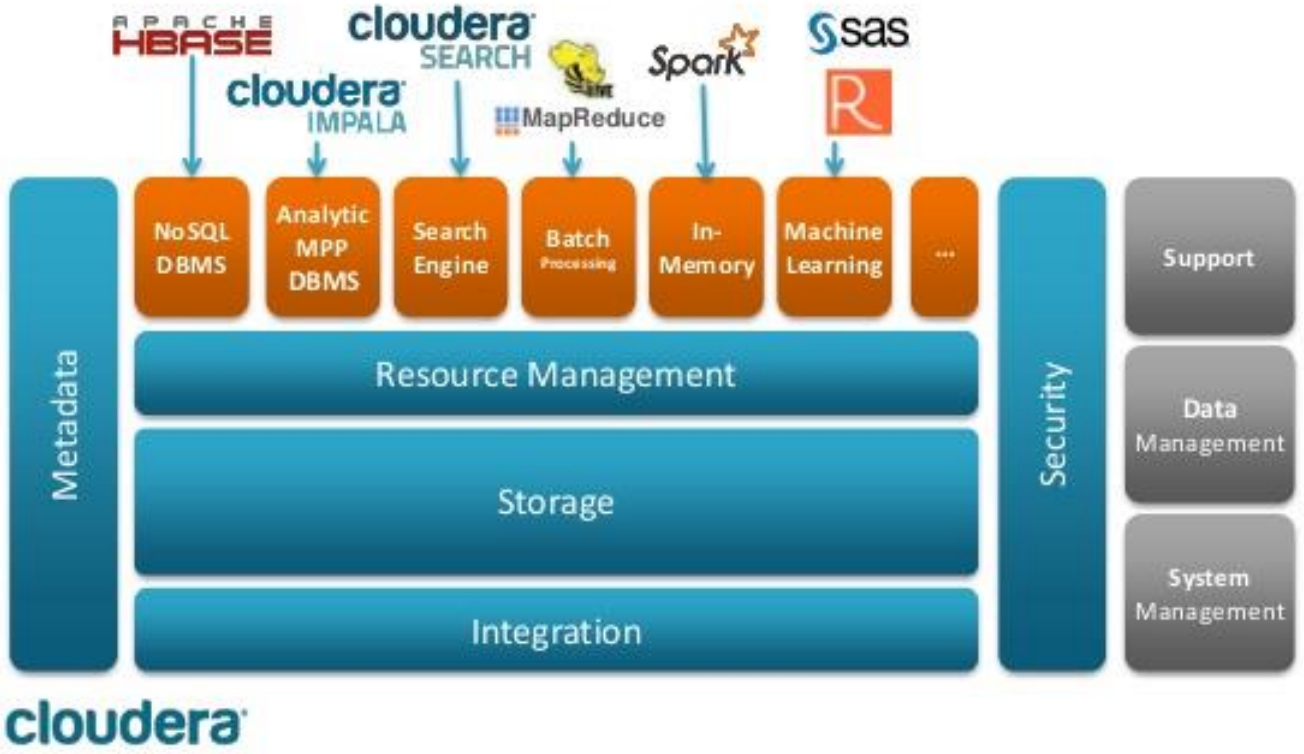


Apache Ambari

- โปรแกรมบริหาร Cluster ทีเป็น Opensource ช่วยทำให้เพิ่มประสิทธิภาพในการบริหาร Server

Cloudera

CDH: the App Store for Hadoop

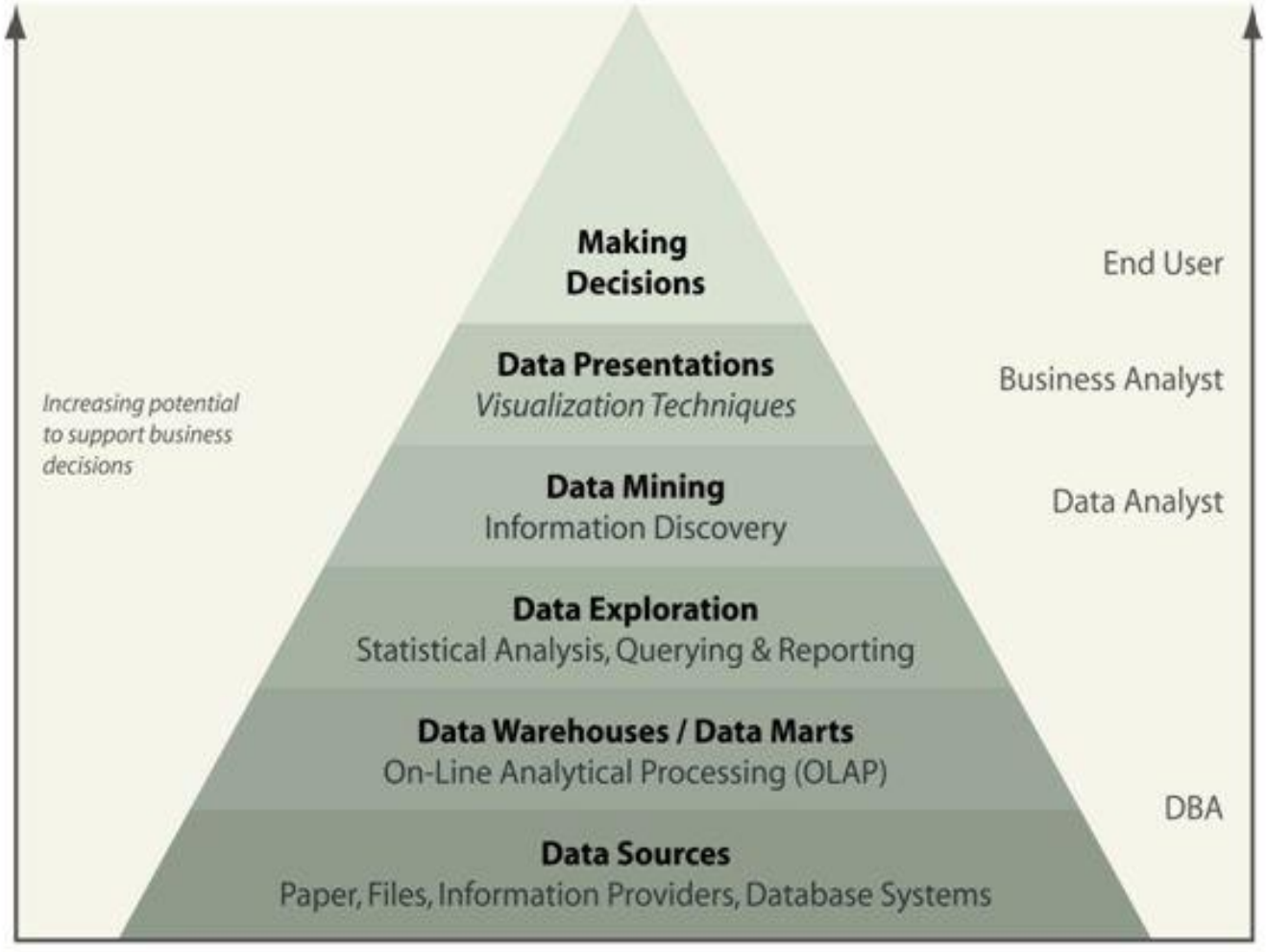


Lab

Basic +ableau

Narongrich Chanprasert
BI Consultant
Depth First Co.,Ltd.

Data usage

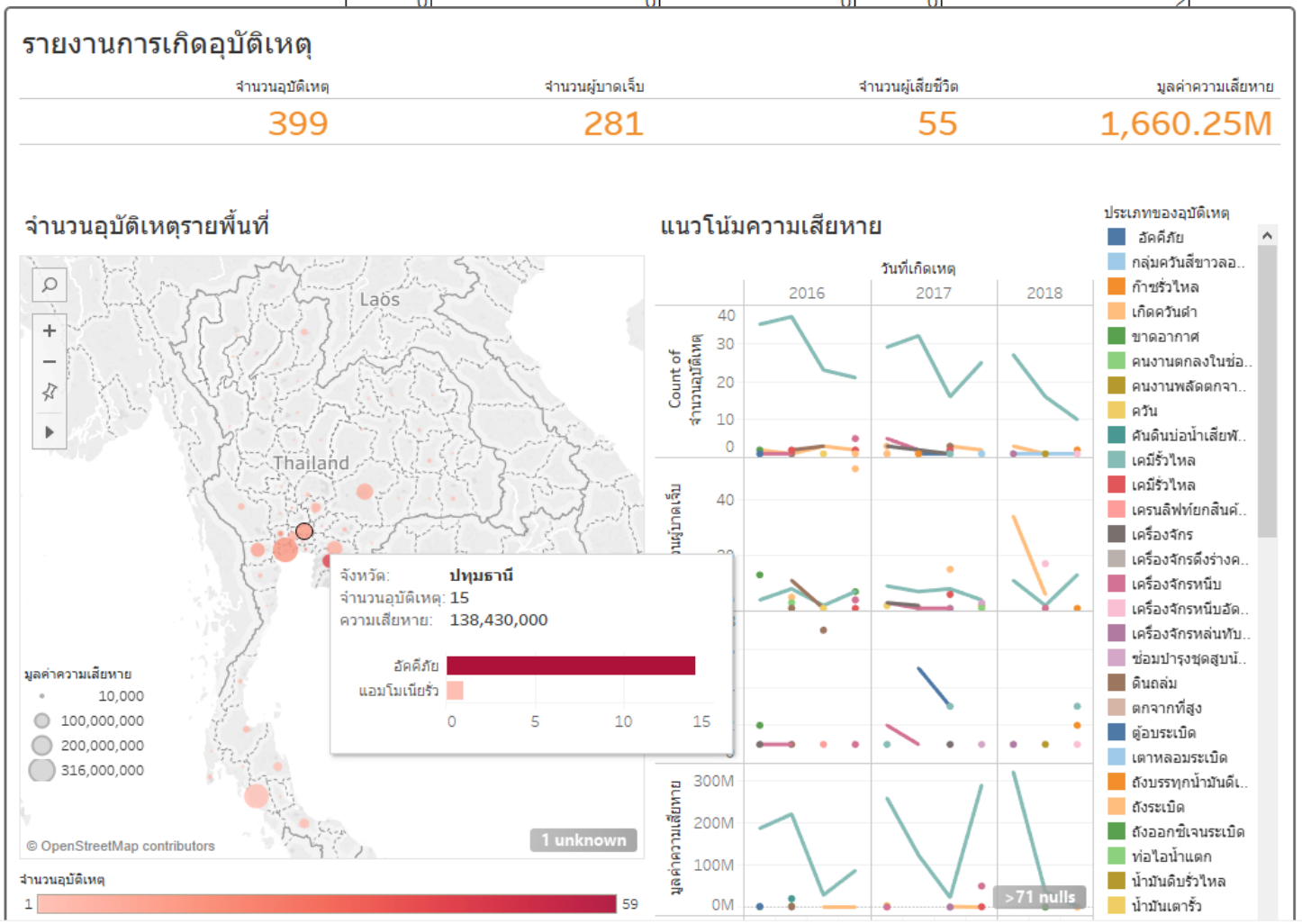

Making Decisions
Data Presentations
Visualization Techniques
Data Mining
Information Discovery
Data Exploration
Statistical Analysis, Querying & Reporting
Data Warehouses / Data Marts
On-Line Analytical Processing (OLAP)
Data Sources
Paper, Files, Information Providers, Database Systems
Increasing potential to support business decisions
End User
Business Analyst
Data Analyst
DBA

รายงานระดับปฏิบัติการ และรายงานสำหรับผู้บริหาร

- รายงานสำหรับระบบงานปรกติ เพื่อให้ผู้ใช้งานระดับปฏิบัติการ
- รายงานนำเสนอในรูปแบบตาราง หรือออกเป็น Spreadsheet
- รายงานนำเสนอข้อมูลที่เป็น Data เป็นข้อมูลที่แสดงรายละเอียด

- รายงานสำหรับระบบงานปรกติ เพื่อให้ผู้ใช้งานระดับผู้บริหาร
- รายงานนำเสนอในรูปแบบกราฟ แผนภาพ หรือตาราง Pivot
- รายงานนำเสนอข้อมูลที่เป็น Information เป็นข้อมูลเชิงสรุป

	A	C	H	I	J	K	L
1	รหัสโรงเรียน	โรงเรียน	จำนวนคอมพิวเตอร์				
2				สพฐ. จัดสรร			เงินนอกงบประมาณ
3			pc ปี 53	สพฐ. จัดสรร เครื่อง Server	สพฐ. จัดสรร เครื่อง PC	เครื่อง PC	เงินนอกงบประมาณเครื่อง PC
4	81010065	บ้านร่าหมาด	0	0	0	0	5
5	81010066	บ้านร่าปู	24	1	24	0	27
6	81010067	บ้านคลองย่าหนัด	0	0	0	0	2
7	81010068	บ้านสิกี					
8	81010069	บ้านปากคลอง					
9	81010070	บ้านนาทุ่งกลาง					
10	81010071	บ้านโล๊ะใหญ่					
11	81010072	บ้านทุ่ง					
12	81010073	บ้านคลองโตนด					
13	81010074	บ้านหลังสอด					
14	81010075	บ้านคลองหิน					
15	81010076	วัดเกาะลันตา					



Roadmap to embrace Business Intelligence

Predictive Analysis
Forecasting

OLAP Reports
What-If Analysis

Simple
Reports

Data

Common Business Intelligence Process

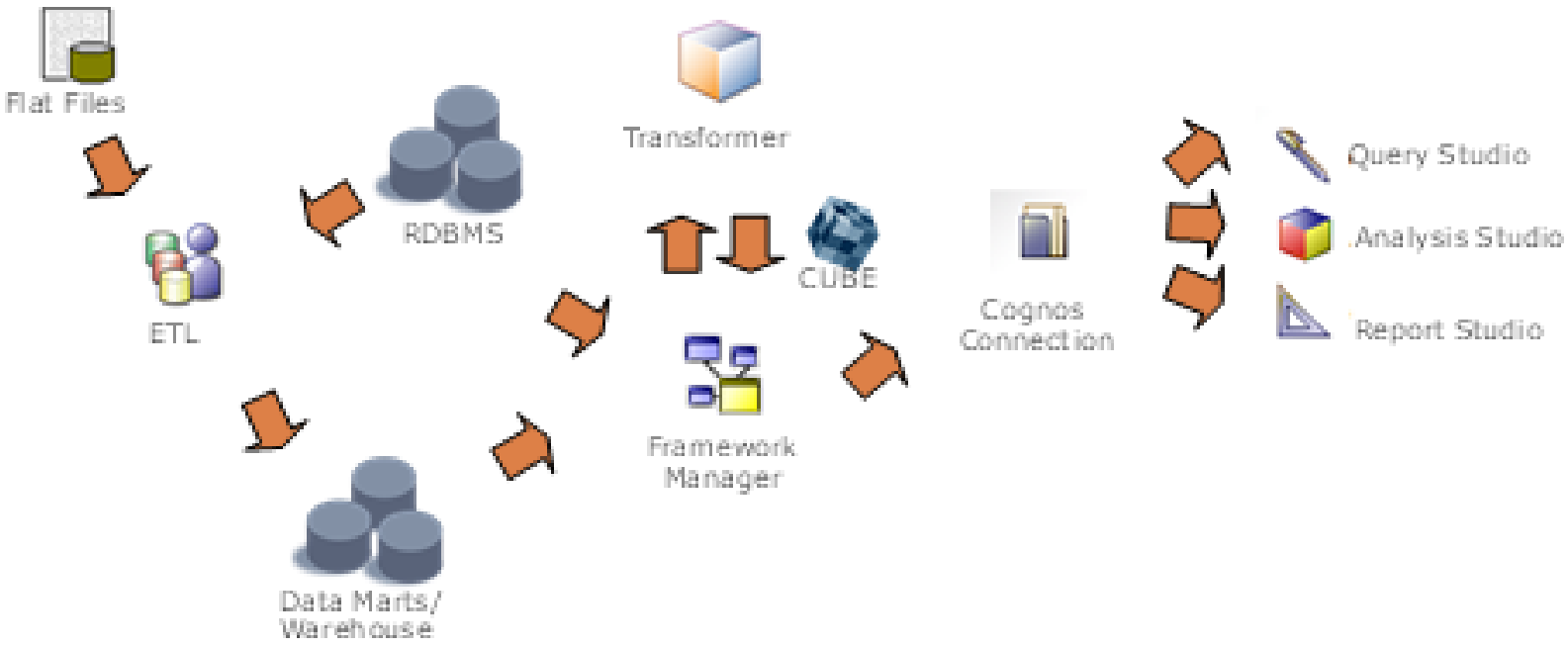
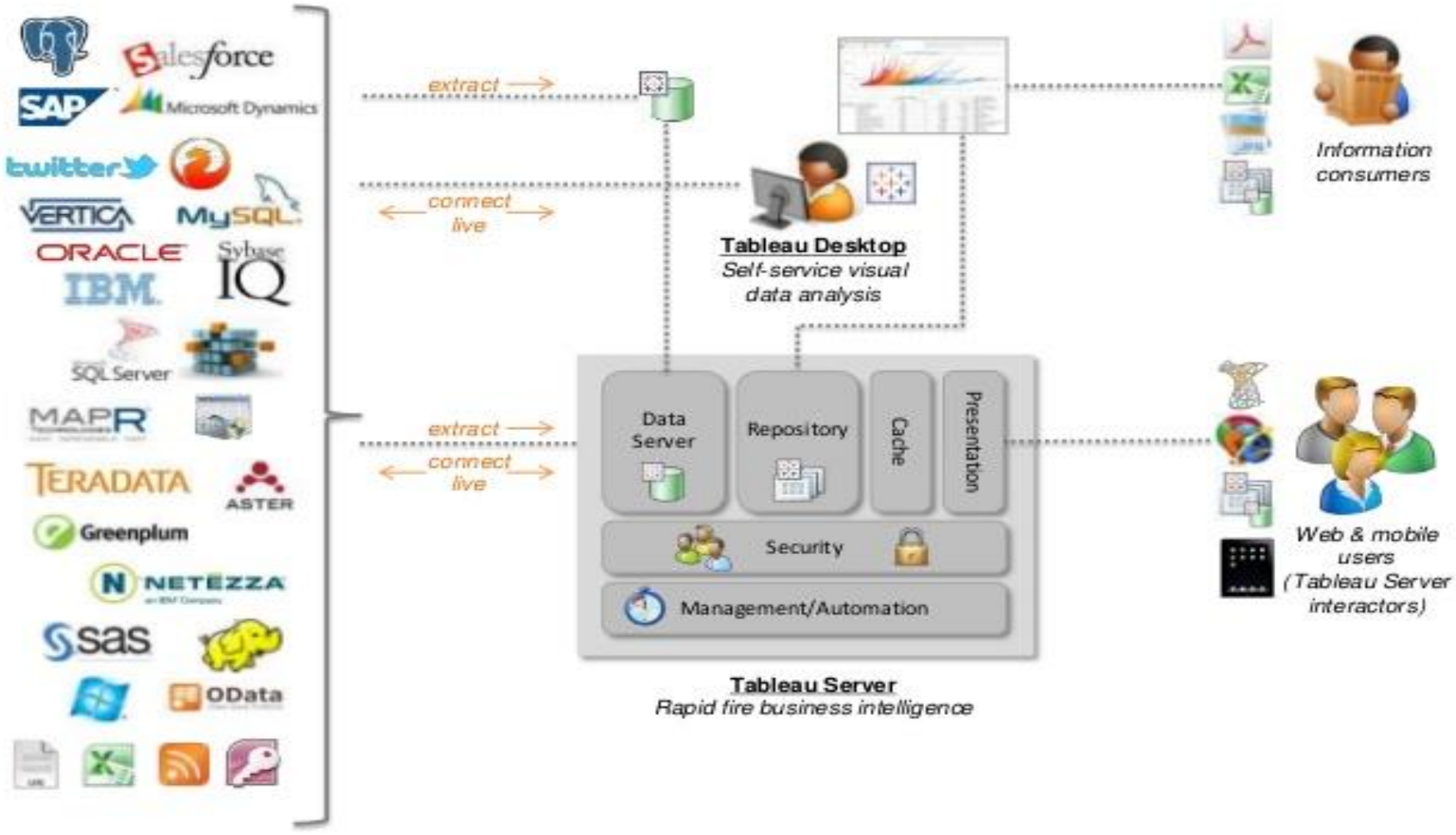


Tableau Architecture



Basic Data Warehouse Concept

- Dimension
  - คือ มุมมองในการวิเคราะห์ข้อมูล ส่วนใหญ่เป็นข้อมูลเชิงคุณภาพ เช่น ข้อมูลชื่อ หรือว่ารายละเอียด ต่างๆ
- Measure
  - คือ ข้อมูลเชิงปริมาณที่มีค่าวัดมากน้อยได้ สามารถนำมาประมวลผลเชิงผลรวม Aggregation ได้ เช่น ผลรวม, ค่าเฉลี่ย, min, max เป็นต้น