CHULALONGKORN
BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Introduction To Big Data

Thaveewat Khanan

CHULALONGKORN
BUSINESS SCHOOL
FLAGSHIP FOR LIFE

https://github.com/bobbylovemovie/trainbigdata

bobbylovemovie@gmail.com

Big Data

Big Data Analytics

Data Science

Machine Learning

Artificial Intelligence

Deep Learning


โพสต์ทูเดย์
TODAY
การเมือง
ปมคำถามพ่วง
ตั้งนายกฯคนนอก
"ต่อไปพรรคการเมืองส่งเก่ง"
A4
เศรษฐกิจดิจิทัล
'ข้อมูลคืออำนาจ'
big data
SOCIAL MEDIA
ผลตอบแทนลงทุนคอนโดวูบ


สกู๊ป
โพสต์ทูเดย์
เร่งประสิทธิภาพภาษี
สร้างธุรกิจจาก Big Data
Big Data


10 อาชีพที่ถูกคุกคาม
จาก Big Data และ Machine Learning
โดย...พ.อ.ดร.เศรษฐพงศ์ มะลิสุวรรณ
www.nationtv.tv
NationTV22

What is Big Data ?

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

“ Big data is data that exceeds the processing capacity of conventional database systems.

The data is too big, moves too fast,

or doesn’t fit the structures of your database architectures.

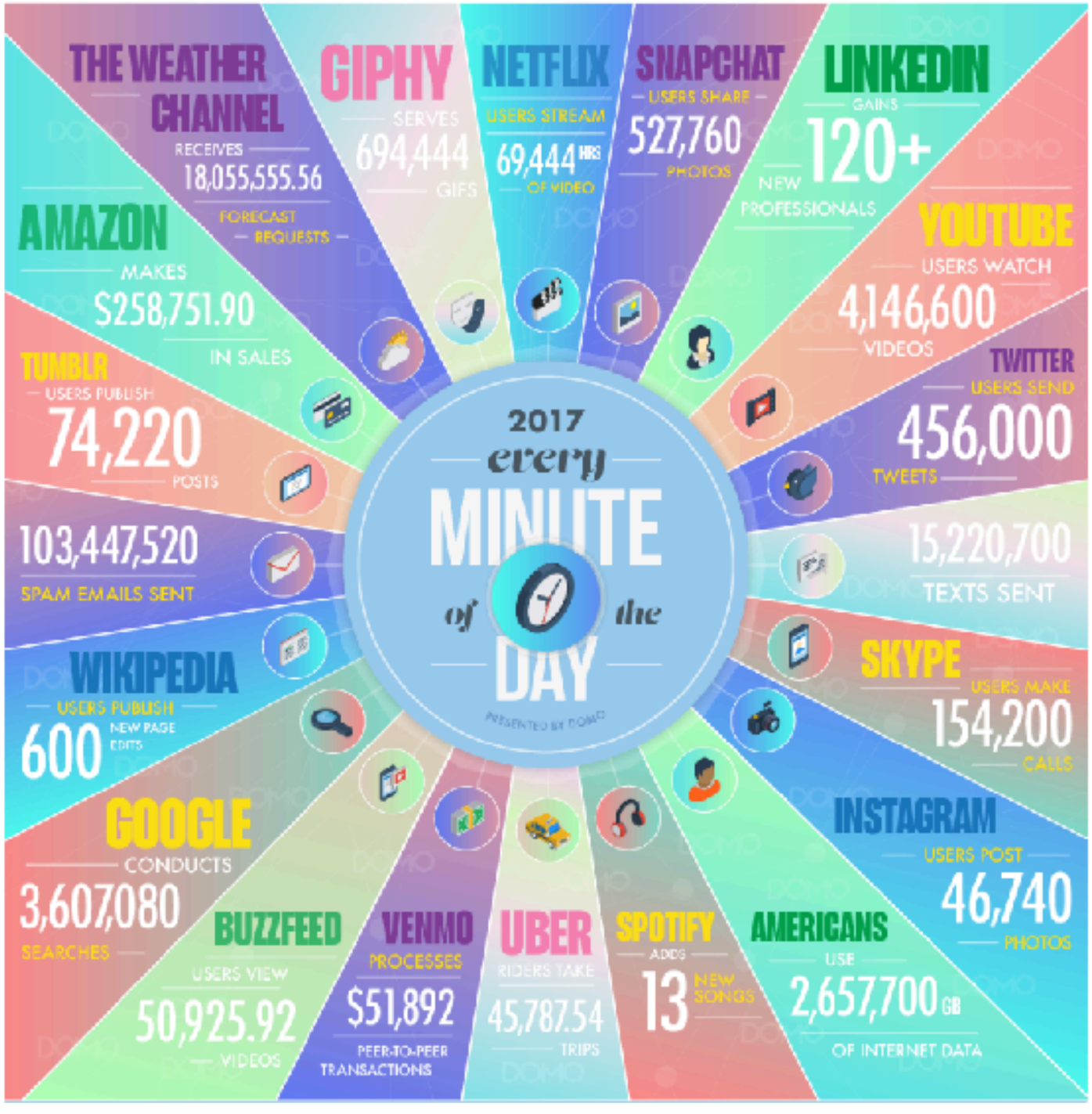
To gain value from this data,

you must choose an alternative way to process it. ”

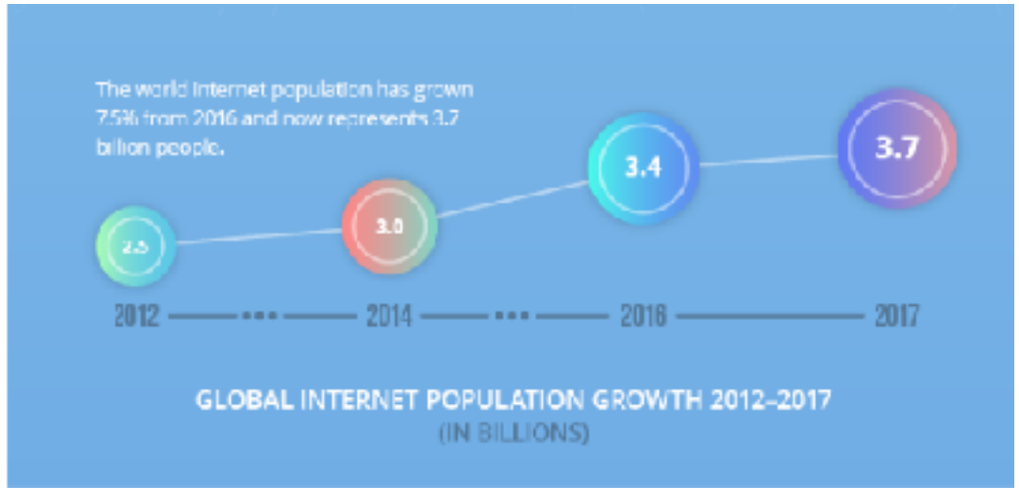
Big Data Now: O'Reilly Media

Why Big Data ?

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



DATA NEVER SLEEP 4.0



https://www.domo.com/blog/2016/06/data-never-sleeps-4-0/

Three Characteristics of Big Data

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Volume

- Volumes of data are larger than those conventional relational database infrastructures can cope with

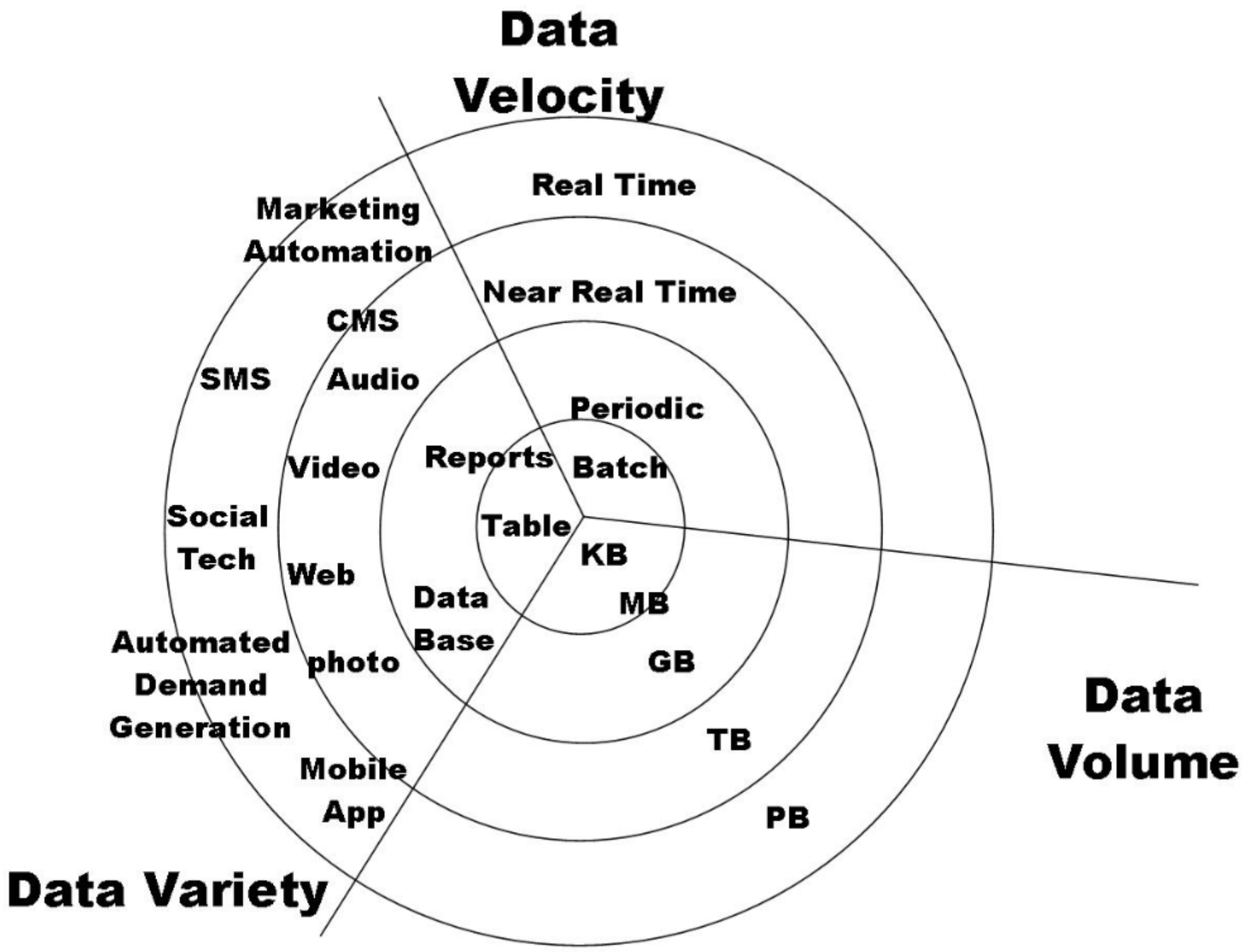
Velocity

- Rate at which data flows in is much faster.
  - Mobile event and interaction by users.
  - Video, image , audio from users

Variety

- the source data is diverse, and doesn't fall into neat relational structures eg. text from social networks, image data, a raw feed directly from a sensor source.

Source Introduction to Big Data: Dr. Putchong Uthayopas

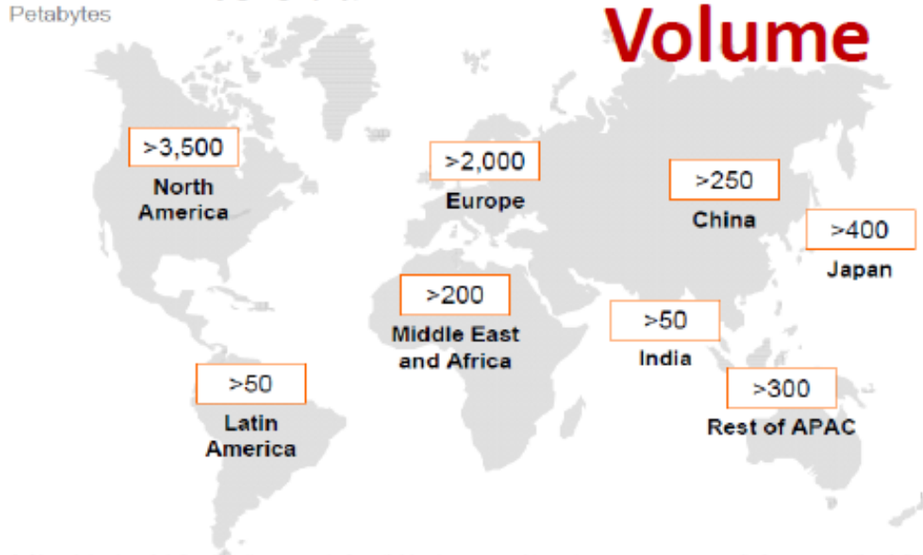

Data
Velocity
Real Time
Near Real Time
Periodic
Batch
Reports
Table
Marketing
Automation
CMS
SMS
Audio
Video
Social
Tech
Web
Data
Base
KB
MB
GB
TB
PB
Automated
Demand
Generation
photo
Mobile
App
Data
Volume
Data Variety

Big Data = Volume, Variety and Velocity (3Vs)

Amount of new data stored varies across geography

New data stored[1] by geography, 2010
Petabytes

Volume



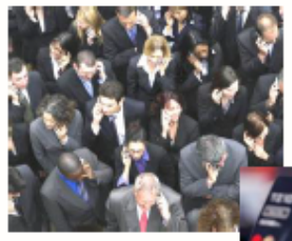
1 New data stored defined as the amount of available storage used in a given year; see appendix for more on the definition and assumptions.

SOURCE: IDC storage reports; McKinsey Global Institute analysis

Variety

People to Machine

People to People



Machine to Machine

Velocity

30 billion pieces of content are shared on Facebook every month.

4 billion hours of video are watched on YouTube each month

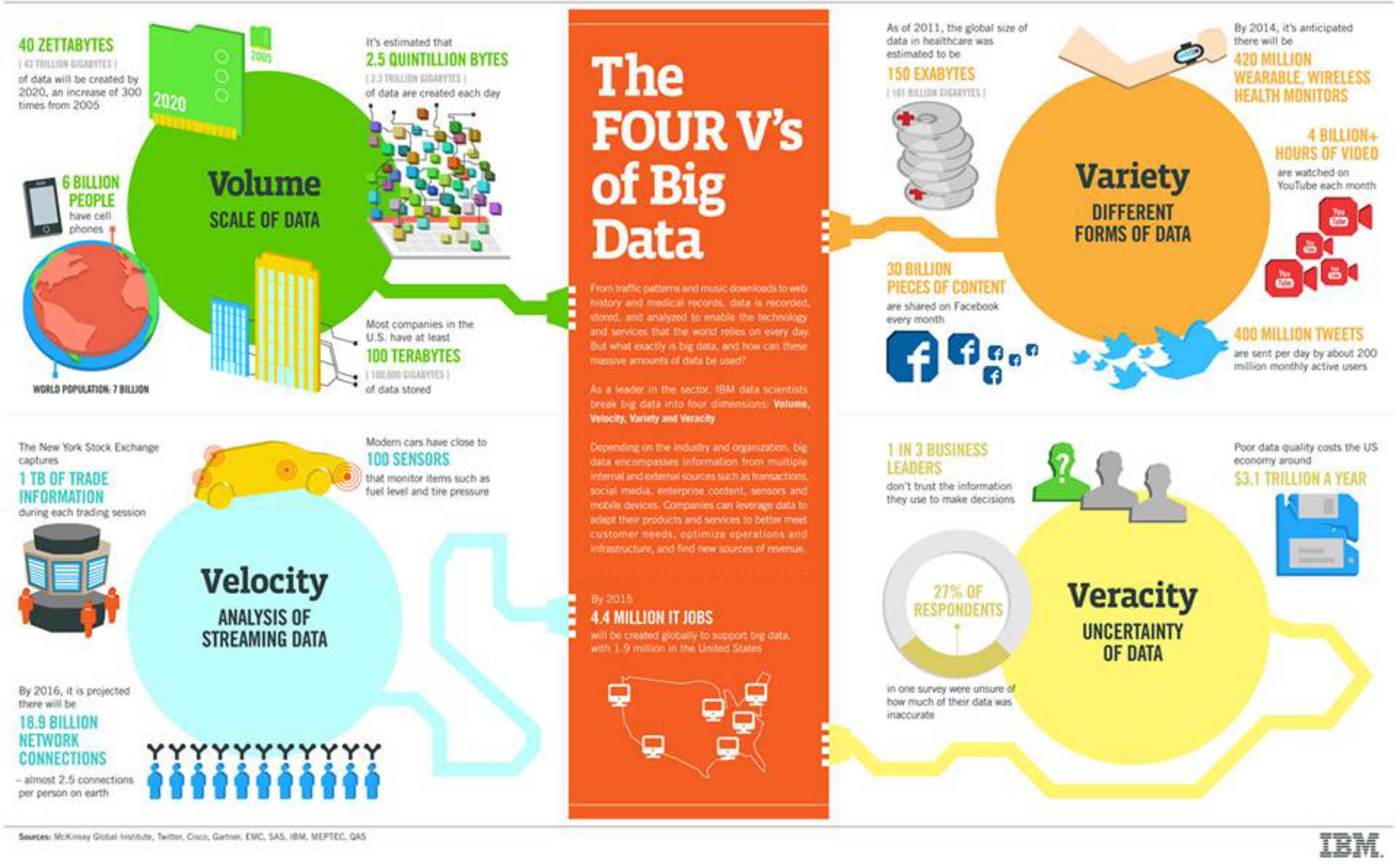
400 Million Tweets are sent per day 200M monthly active users

Source: IBM

Scale

1000 **kilobytes** = 1 Megabyte
1000 **Megabytes** = 1 Gigabyte
1000 **Gigabytes** = 1 Terabyte
1000 **Terabytes** = 1 Petabyte
1000 **Petabytes** = 1 Exabyte
1000 **Exabytes** = 1 Zettabyte
1000 **Zettabytes** = 1 Yottabyte
1000 **Yottabytes** = 1 Bronobyte
1000 **Bronobytes** = 1 Geopbyte

4Vs of Big Data



40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

Volume

SCALE OF DATA

6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

Velocity

ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

Variety

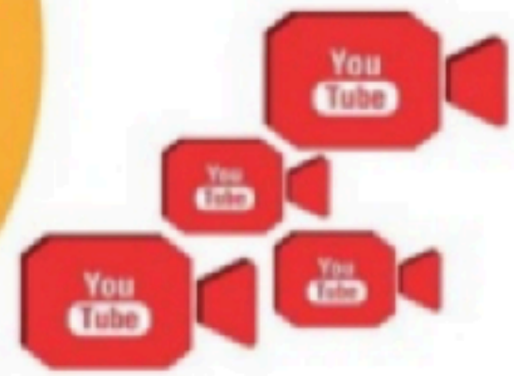
DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month



30 BILLION PIECES OF CONTENT

are shared on Facebook every month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

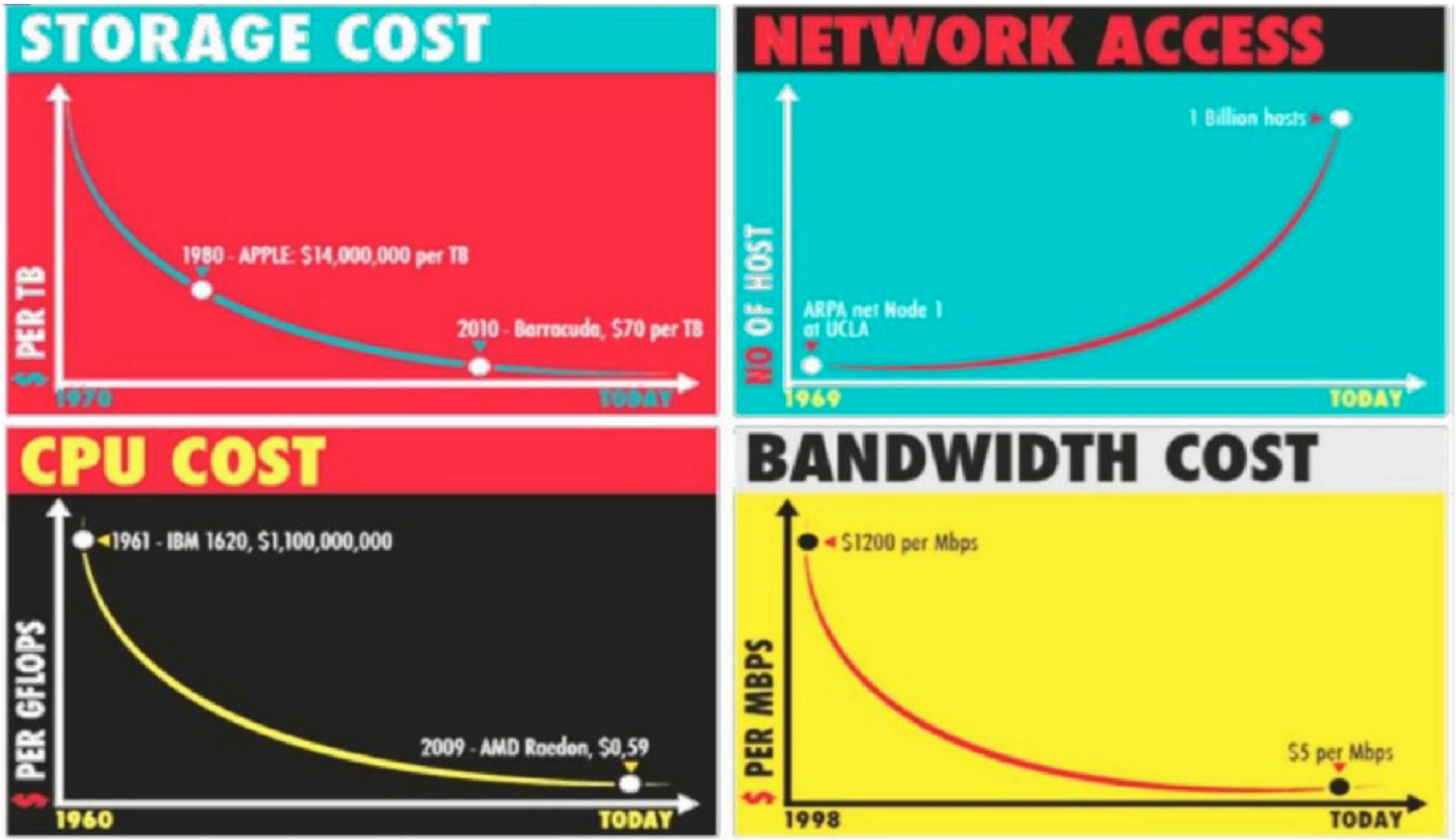
in one survey were unsure of how much of their data was inaccurate

Veracity

UNCERTAINTY OF DATA

Big Data : Why Now?

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Source: William EL KAIM, Enterprise Architecture and Technology Innovation 21

DATA LAKE

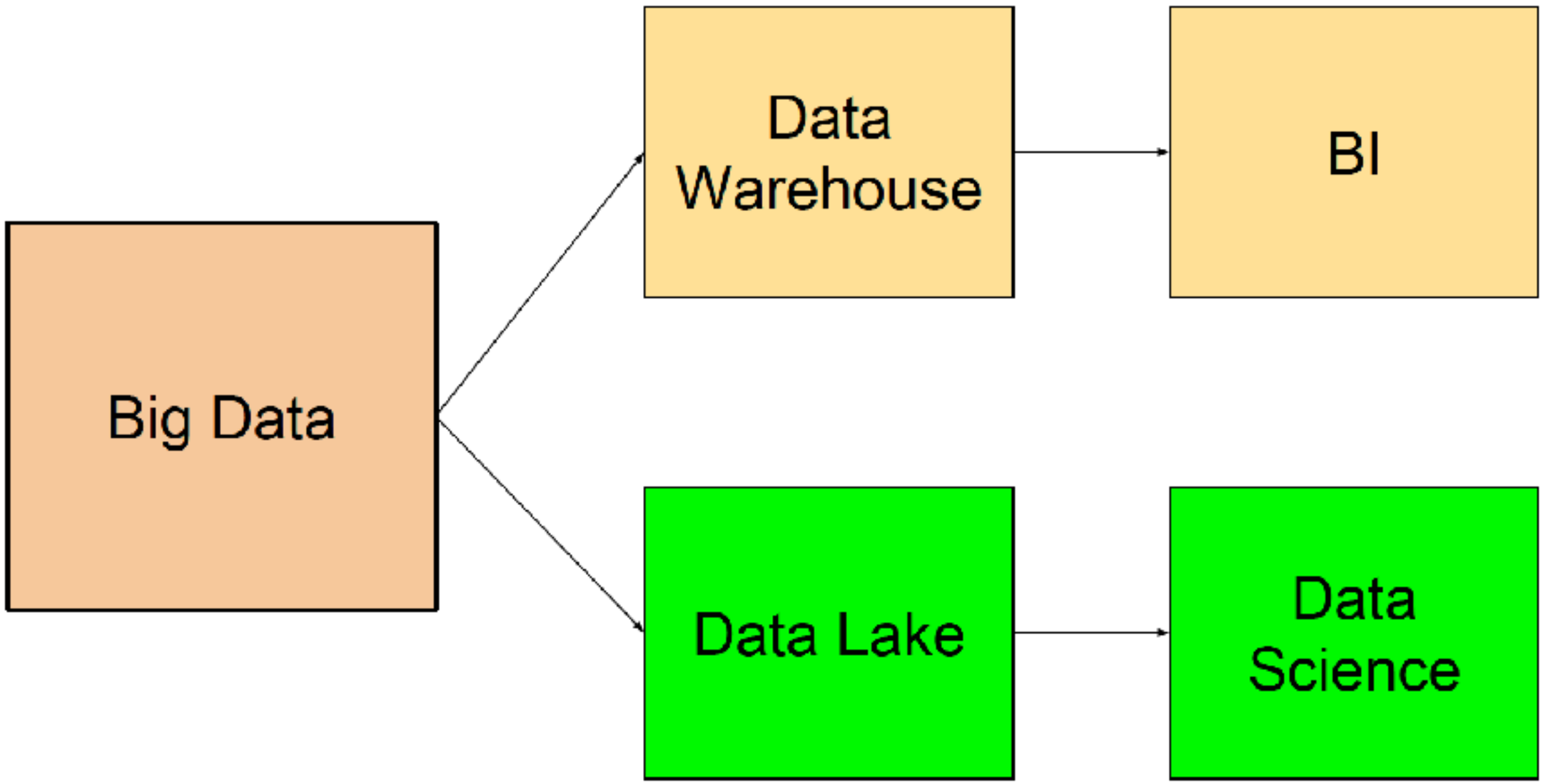
DATA LAKE

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

The old way: Ask, then collect

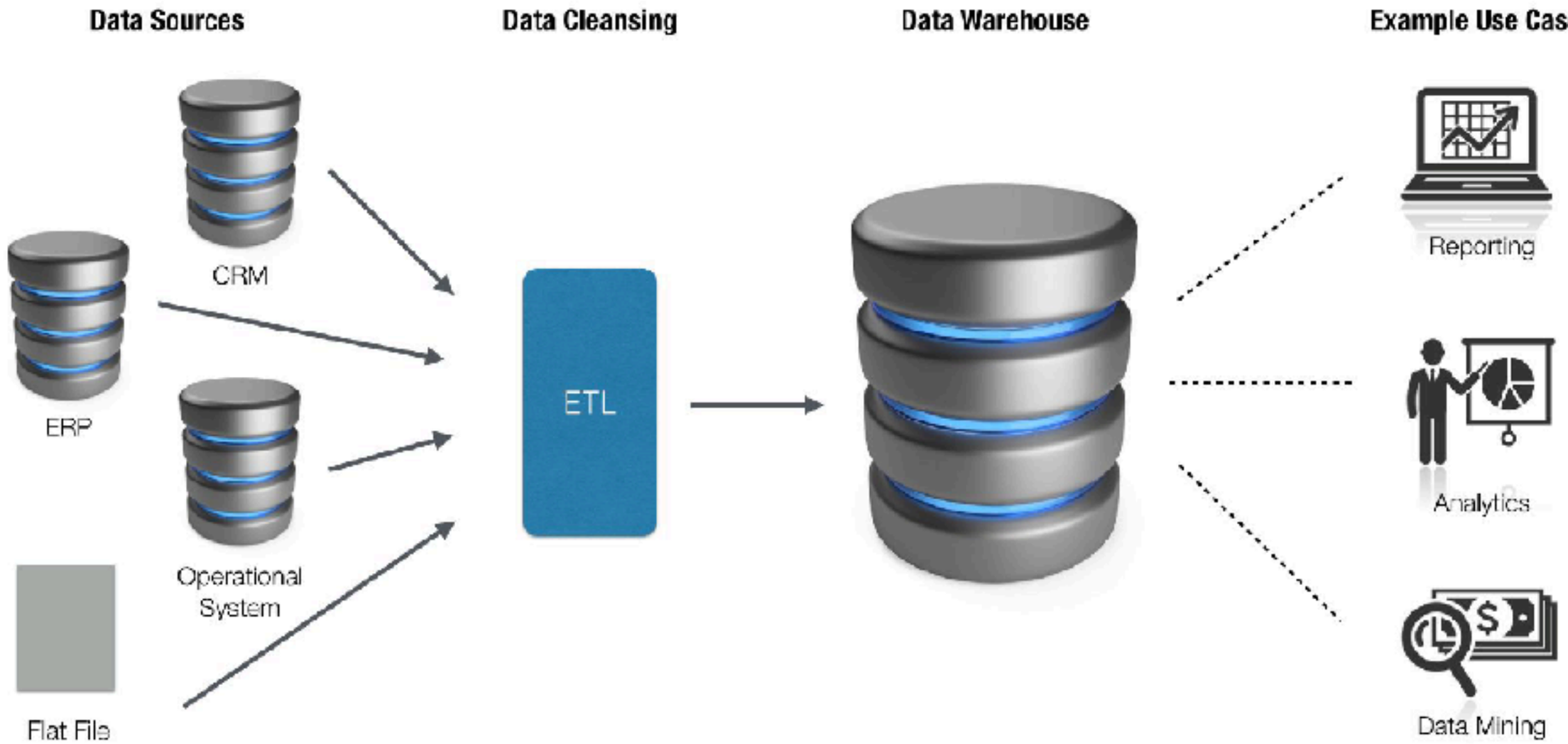
The new way: Collect, then ask

Source: Data Science and Critical Thinking, A, Croll


Big Data
Data Warehouse
BI
Data Lake
Data Science

Data Warehouse

Traditional approach to integrating data for consistency and quality



Data Warehouse Environment

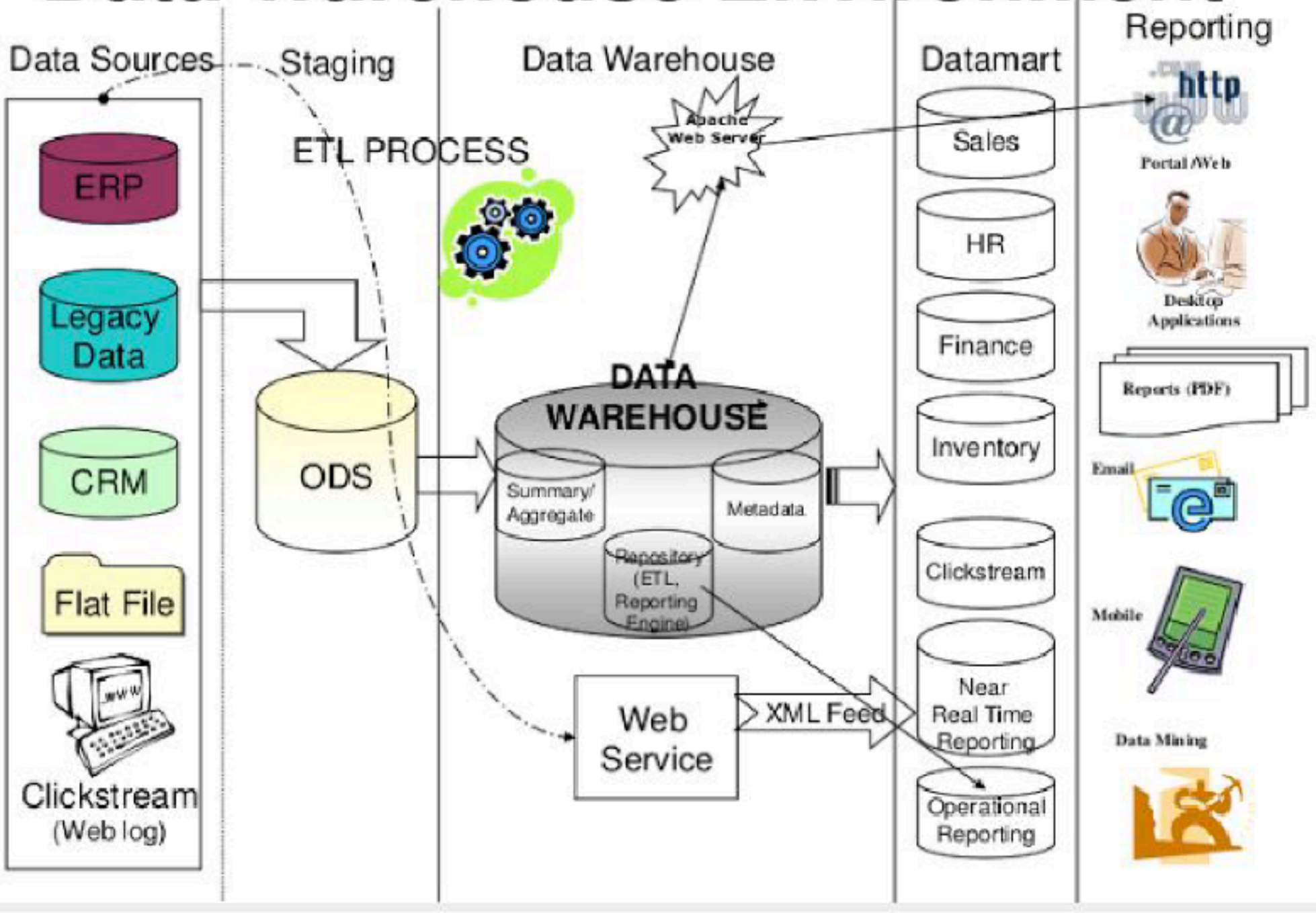
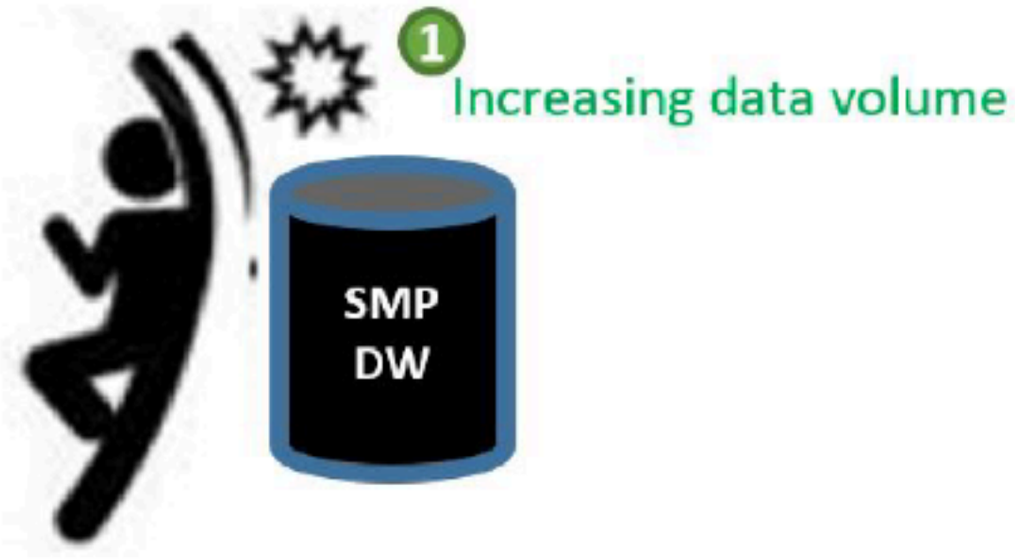
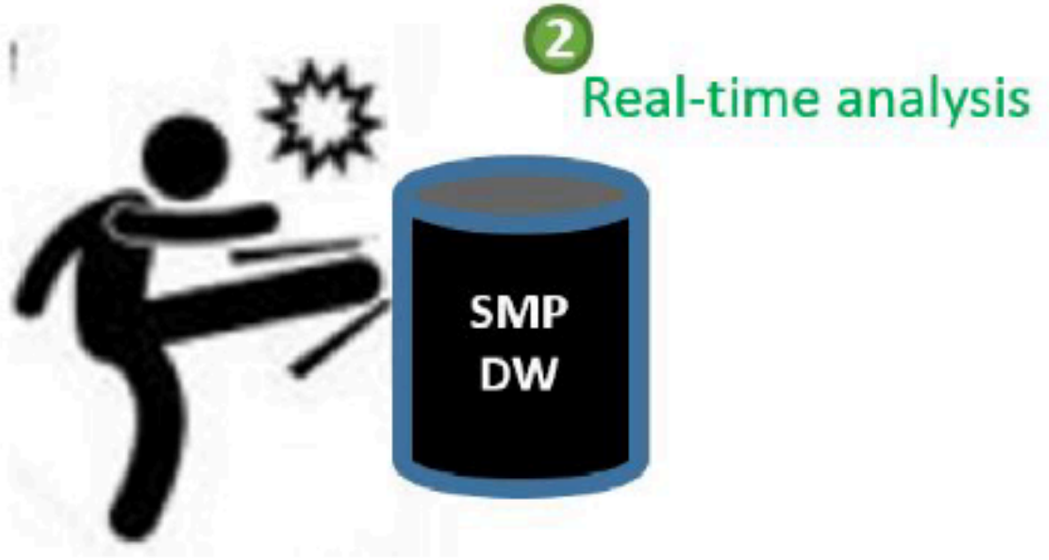
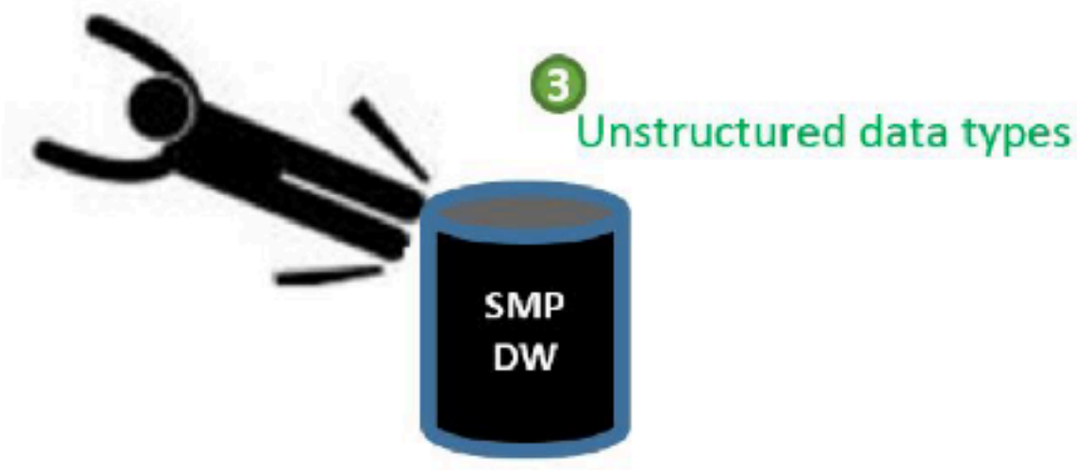


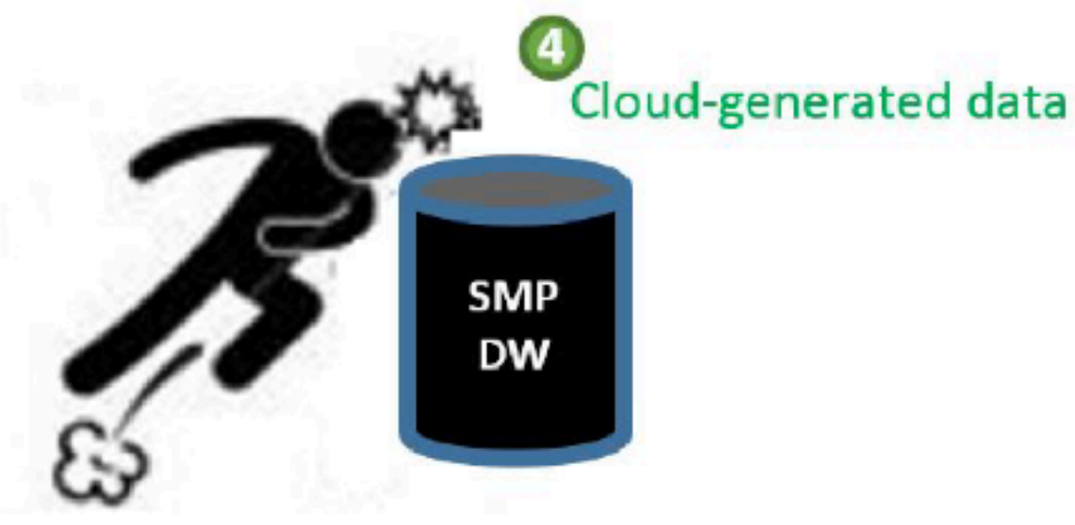
Image: rodneyrohrmann.blogspot.com

Data Warehouse



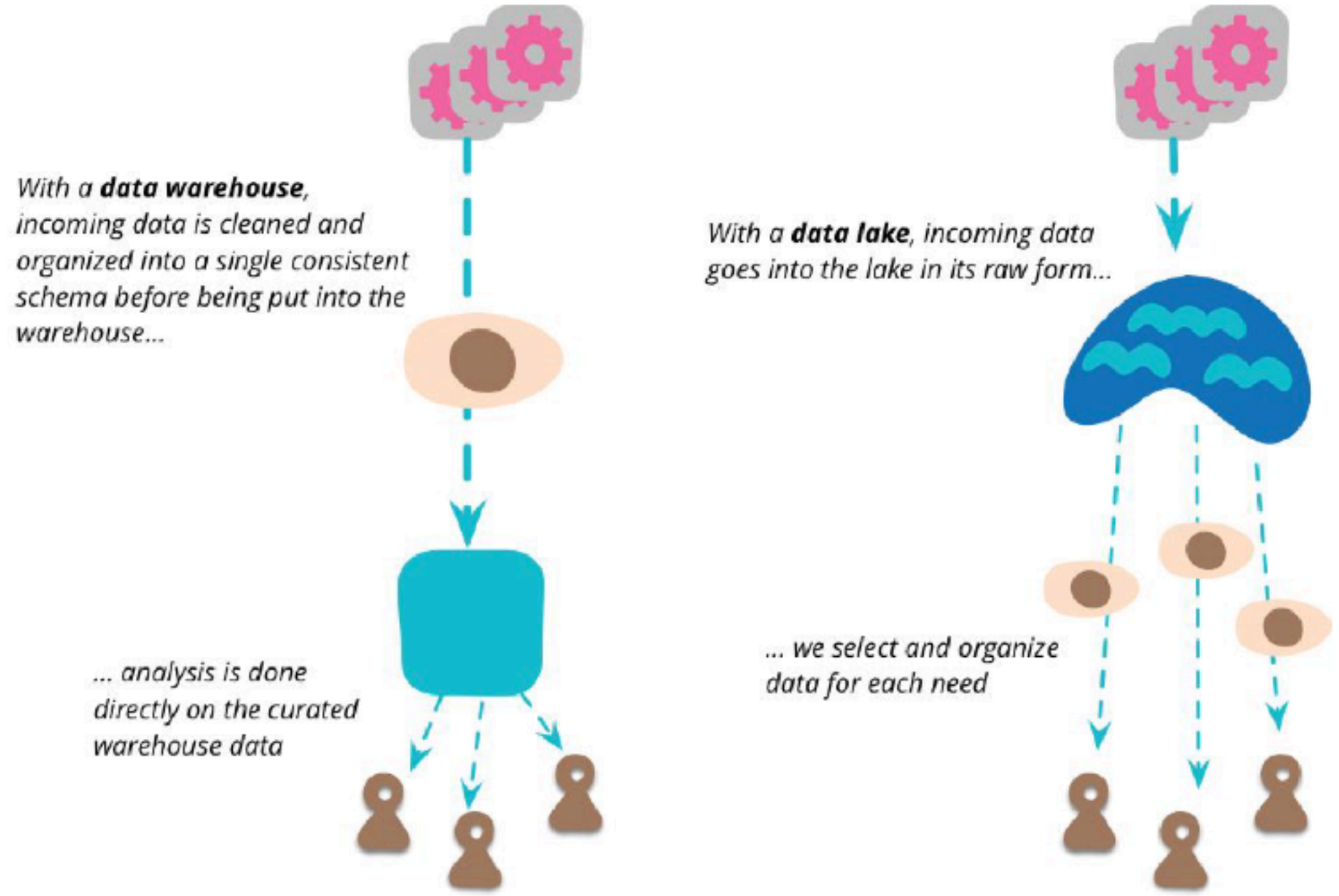




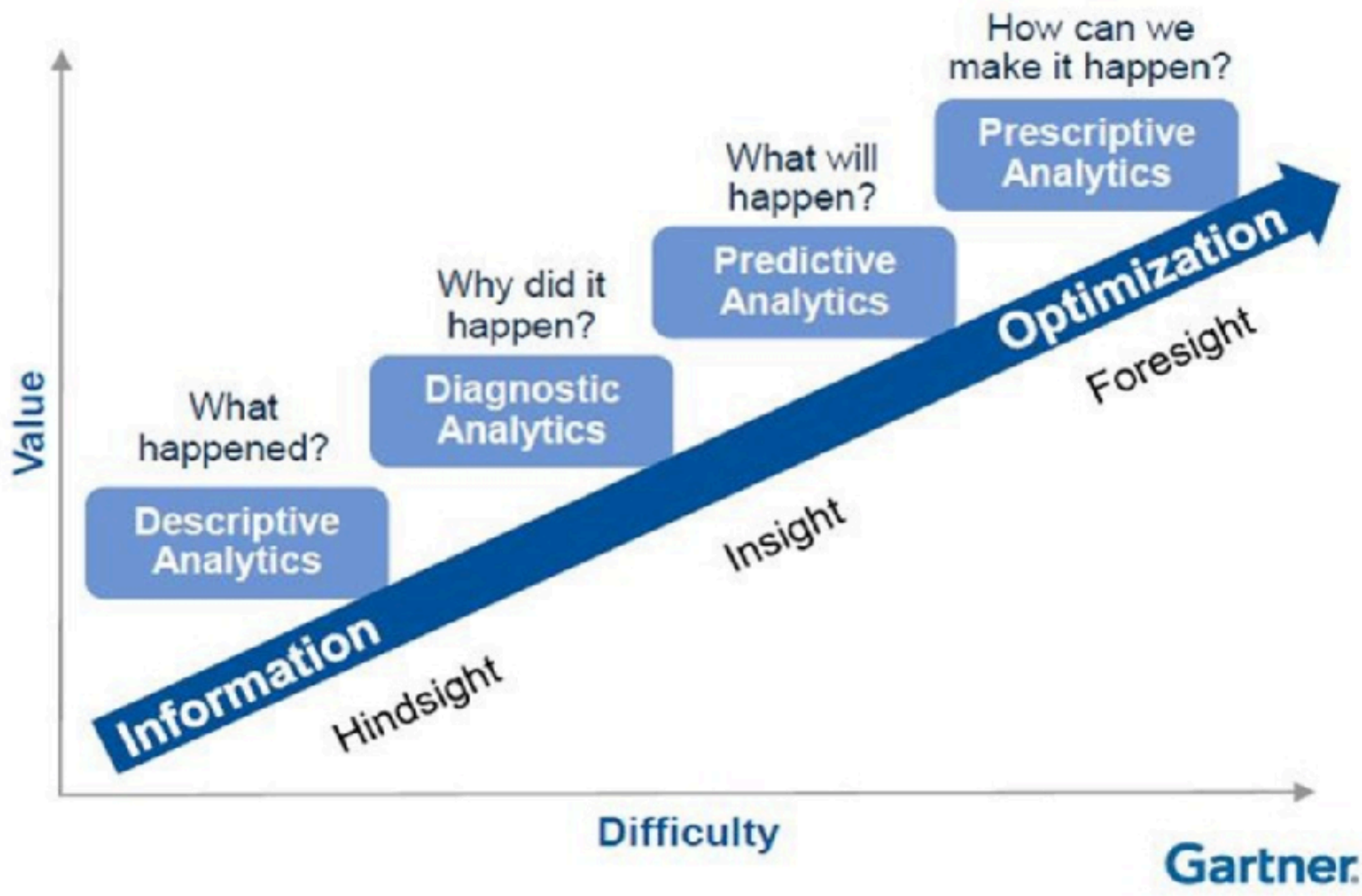


Source: dinesql.blogspot.com

Differences between Data Lake and Data Warehouse

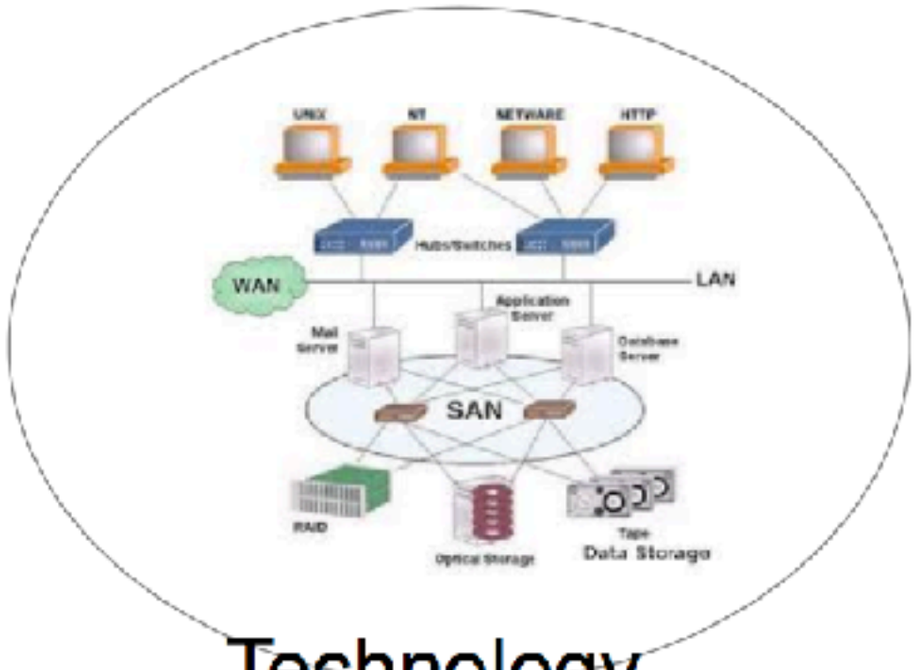


Source: martinfowler.com/bliki/DataLake.html

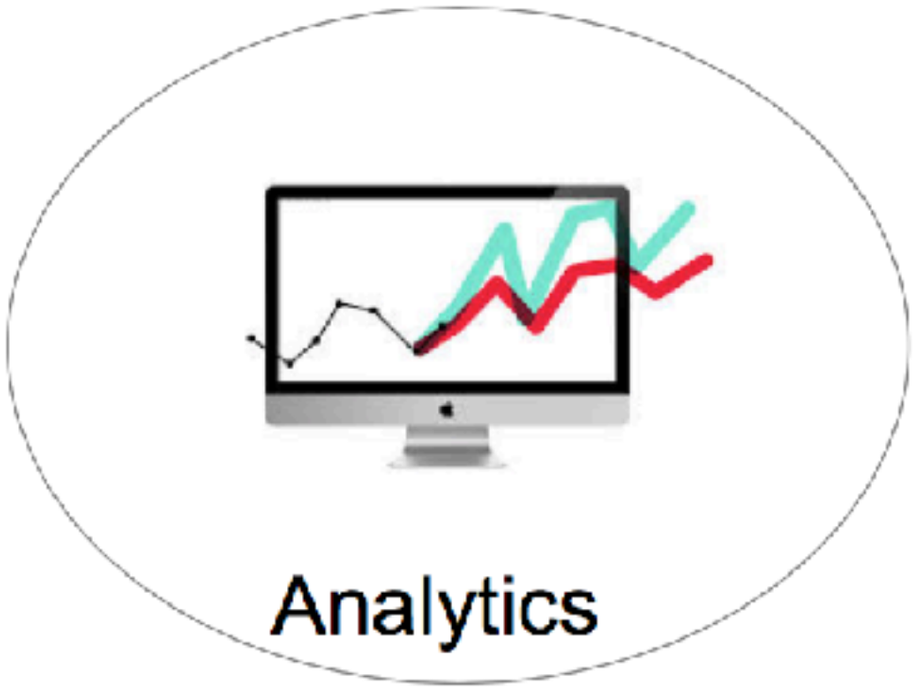

How can we make it happen?
Prescriptive Analytics
What will happen?
Predictive Analytics
Why did it happen?
Diagnostic Analytics
What happened?
Descriptive Analytics
Optimization
Foresight
Insight
Hindsight
Information
Value
Difficulty
Gartner.

	Traditional Analytics (BI)	vs	Big Data Analytics
Focus on	• Descriptive analytics • Diagnosis analytics		• **Predictive analytics** • **Data Science**
Data Sets	• Limited data sets • Cleansed data • Simple models		• Large scale data sets • More types of data • Raw data • Complex data models
Supports	**Causation:** what happened, and why?		**Correlation**: new insight More accurate answers

Data Sources

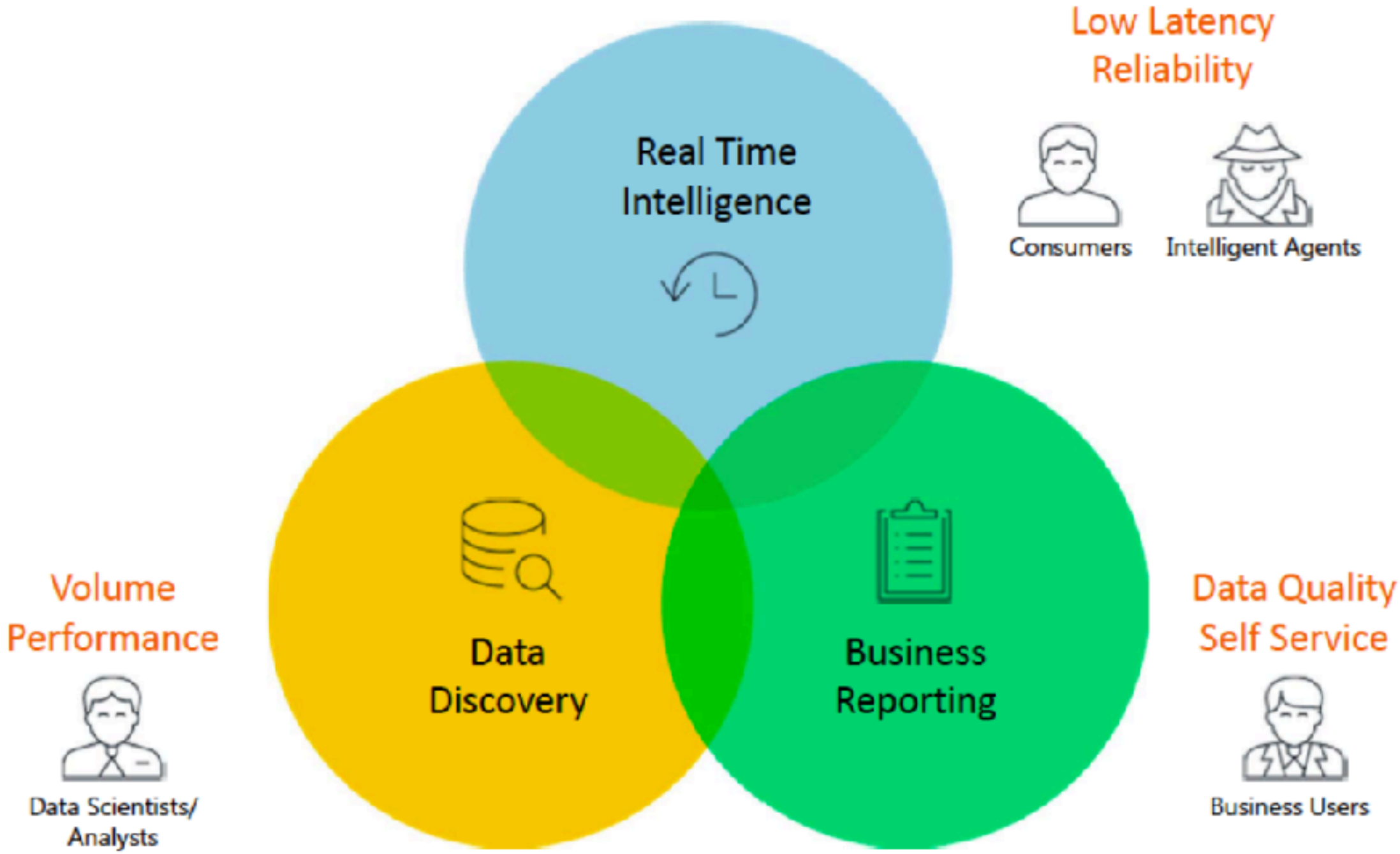


Technology



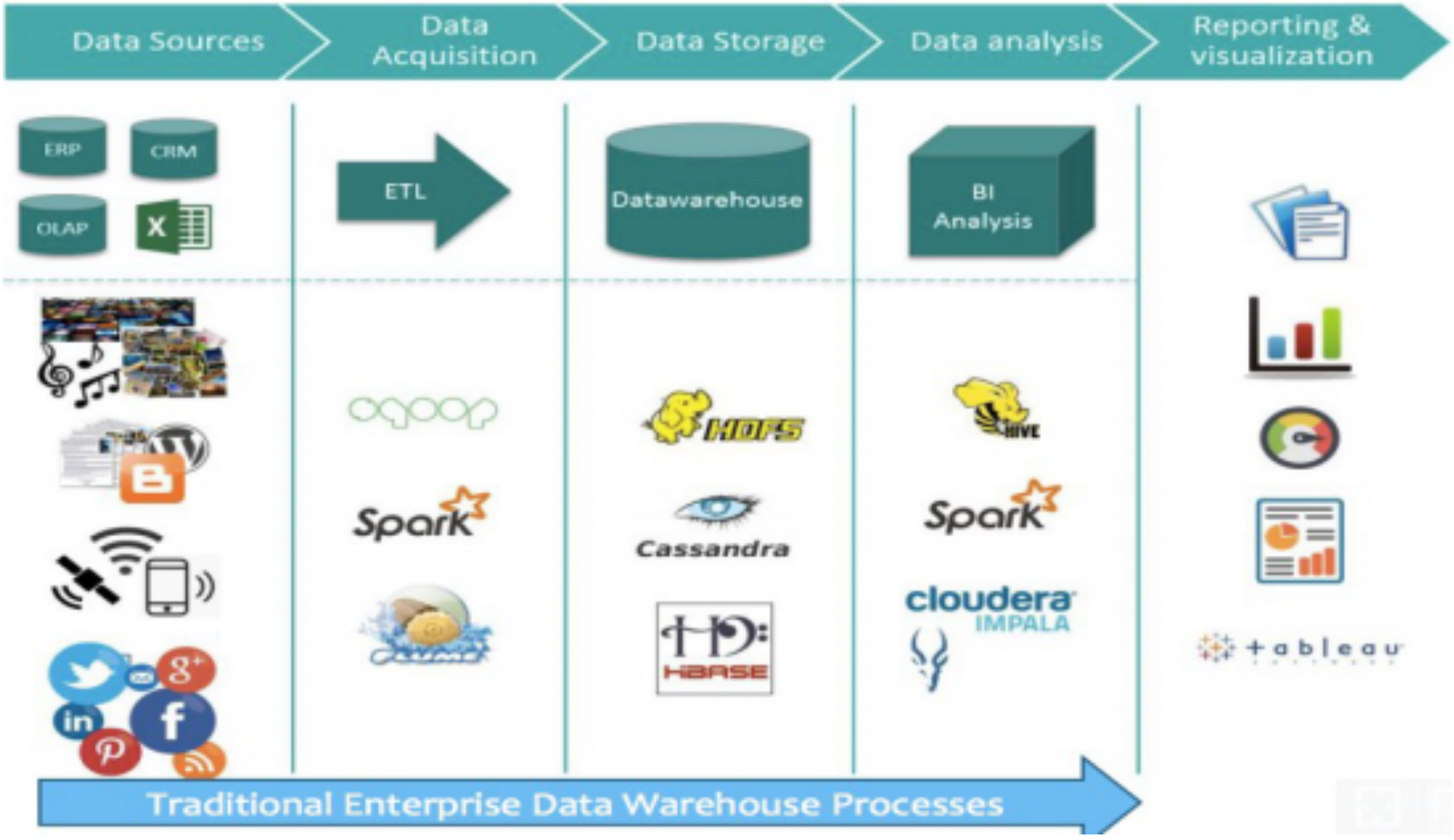
Analytics

Big Data Analytics



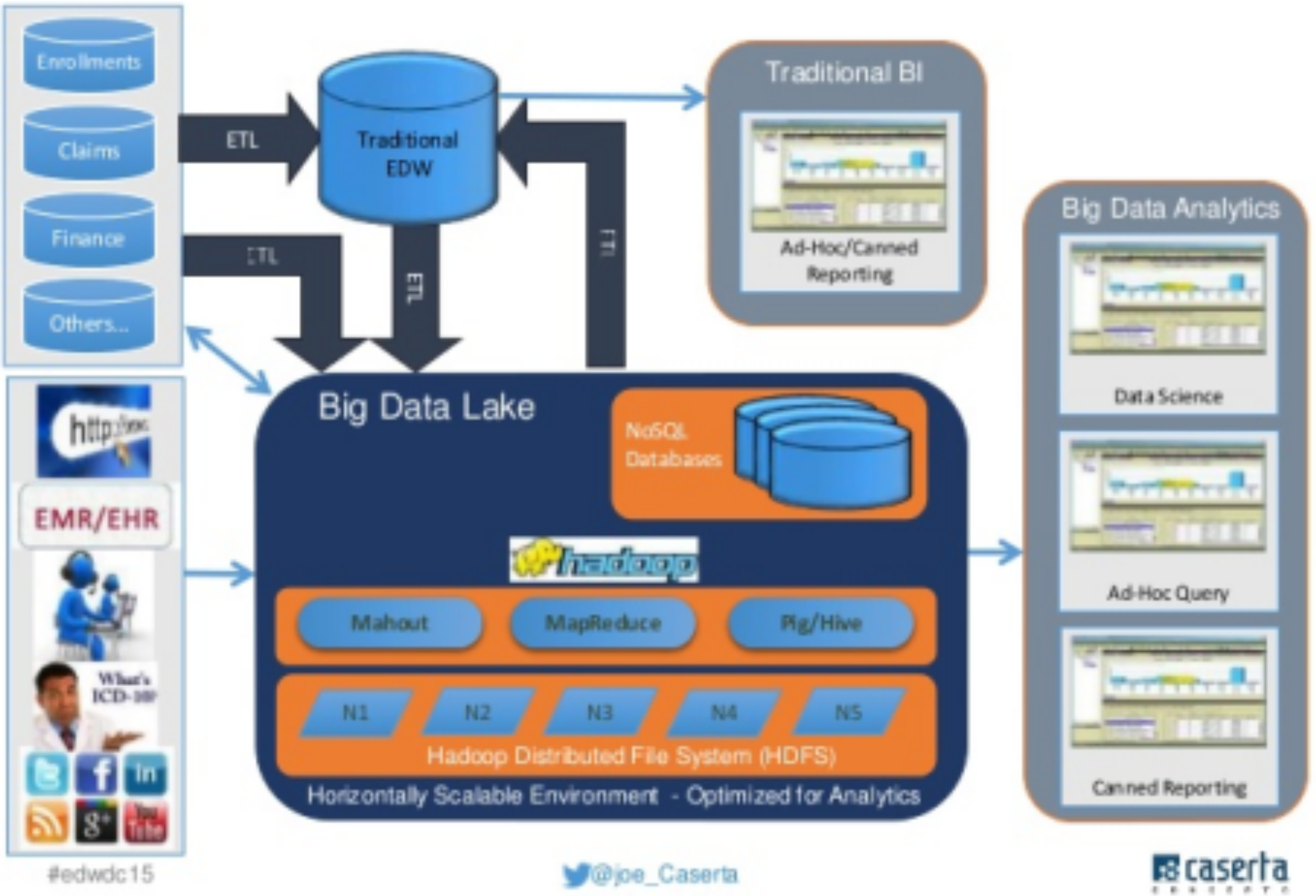
Source: SoftServe

How Data Lake Works?



Source: www.clearpeak.com

Today's business environment requires Big Data



Visualization

The 'Datafication' of our World;

- Activities
- Conversations
- Words
- Voice
- Social Media
- Browser logs
- Photos
- Videos
- Sensors
- Etc.

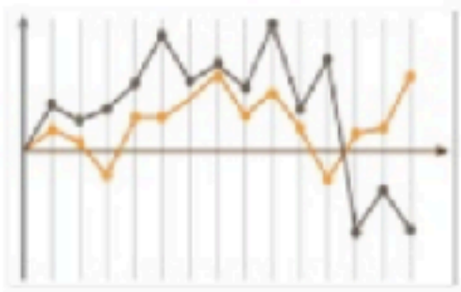
Volume

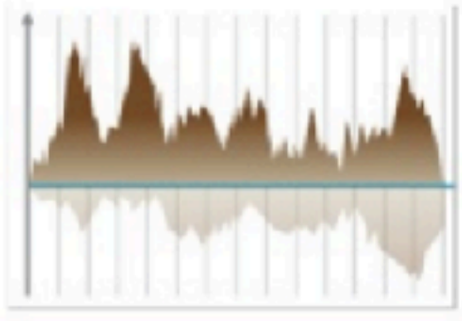
Velocity

Variety

Veracity

Analysis





Analysing Big Data:

- Text analytics
- Sentiment analysis
- Face recognition
- Voice analytics
- Movement analytics
- Etc.

Value

Source: Bernard Marr

Data Lake isn't just a technology
It is an architecture

Data Lake: Key Benefits

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Scale as much as you can

Plug in disparate data sources

Acquire high-velocity data: Store in native format

Don't worry about schema

Unleash your favorite SQL

Advanced algorithms

Administrative resources

Data Lake v.s. Data WareHouse

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Complementary to EDW (not replacement)	Data lake can be source for EDW
Schema on read (no predefined schemas)	Schema on write (predefined schemas)
Structured/semi-structured/Unstructured data	Structured data only
Fast ingestion of new data/content	Time consuming to introduce new content
Data Science + Prediction/Advanced Analytics + BI use cases	BI use cases only (no prediction/advanced analytics)
Data at low level of detail/granularity	Data at summary/aggregated level of detail
Loosely defined SLAs	Tight SLAs (production schedules)
Flexibility in tools (open source/tools for advanced analytics)	Limited flexibility in tools (SQL only)

Data Lake Risks

More Data Sources

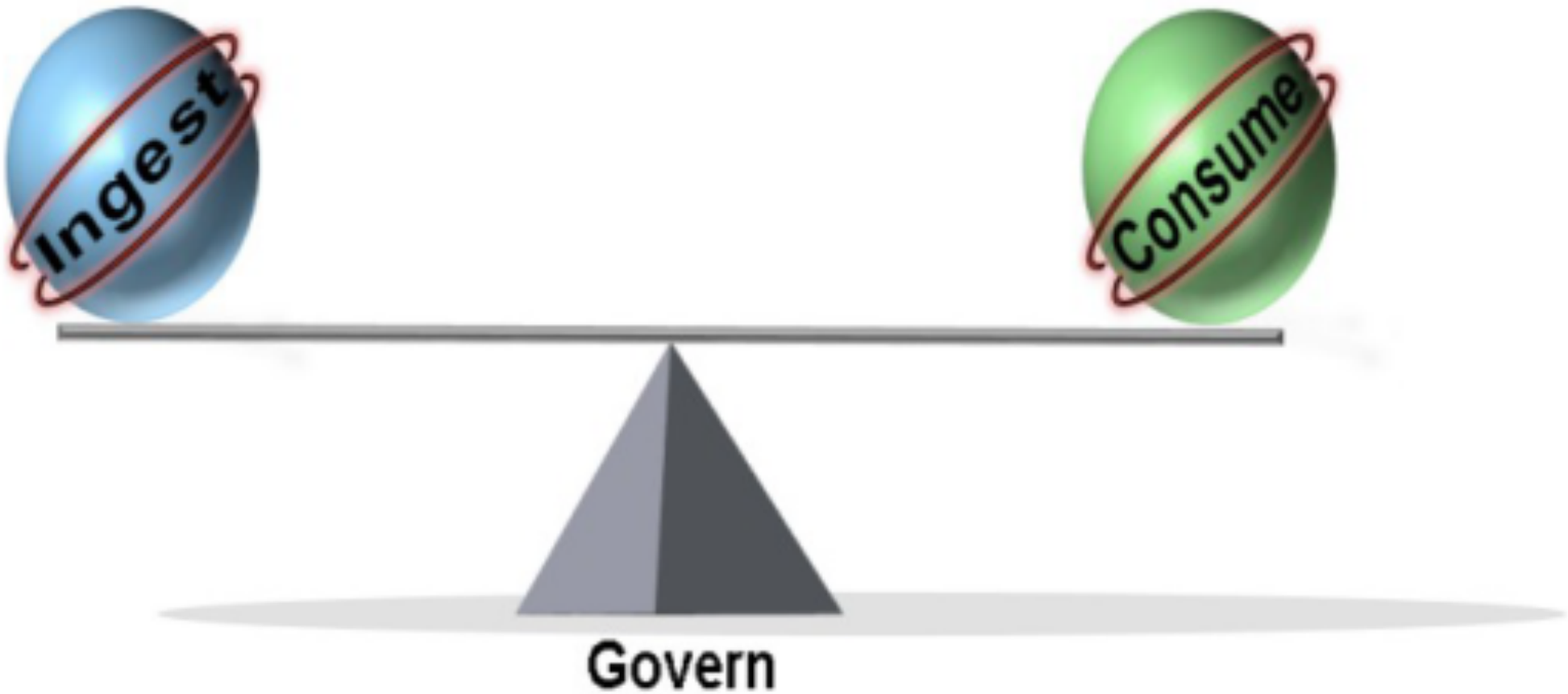
More Applications

More Business Units

More Users

Without proper governance mechanisms
Data lakes risk turning data swamps

Data Lake Governance



Source: What is “Just-Enough” Governance for the Data Lake?

Data Lake Governance

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Fundamental Capabilities

- **The definition of the incoming data from a Business use perspective**
- **Documentation of the context, lineage, and frequency of the incoming data;**
- **Security level classification of the incoming data;**
- **Documentation of creation, usage, privacy, regulatory, and encryption business rules which apply to the incoming data.**

Source: What is “Just-Enough” Governance for the Data Lake?

What can it do for my Data Lake

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

- Where did my data come from ? How is it being transformed ?
- Track usage, resolve anomalies, visualize, optimize and clarify data lineage Search and access data
- Assess data quality and fitness for purpose
- Govern who can/cannot access the data
- Data life cycle management, archiving and retention policies
- Auditing, compliance

Source: Data Governance in the Hadoop Data Lake

Summary

- Big Data: Data lake instead of data warehouse?
- Data Lake is not only a technology, it is an architecture
- Data Lake components: Data acquisition (intake), Data management, Data Storage, Data consumption (Discovery)
- Data Lake governance is very important

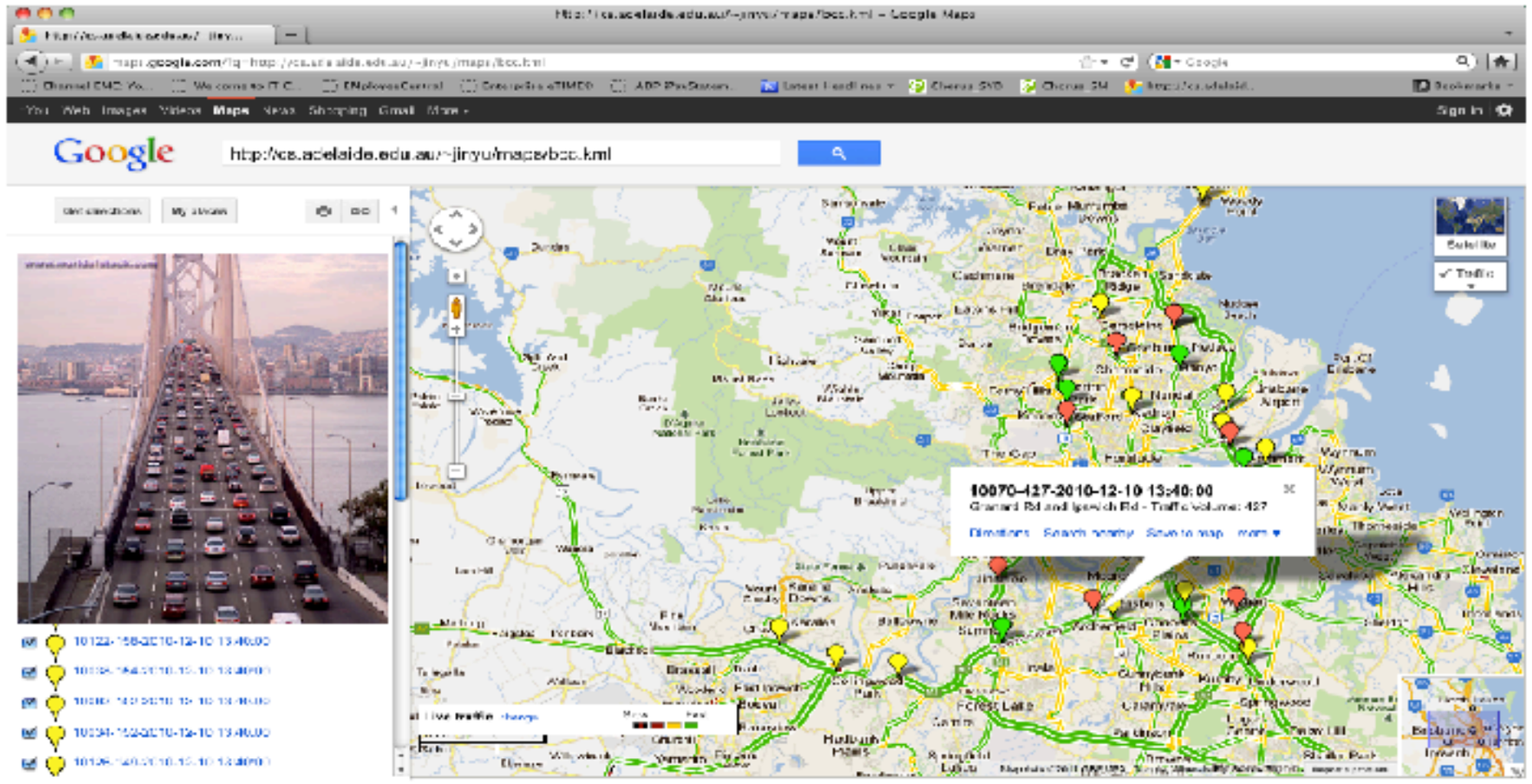
Customer Case Studies on Big Data Lake

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Customer Example: Analytics

Municipal Traffic Analysis to Simulate Traffic Velocity Patterns and Reduce Delays

- Correlate multiple types of data (GPS, weather, sensor, video, social media)
- Simulation techniques to model traffic transition points
- Signal retiming to minimize stops and delays
- Peak delays reduced by 16% and stops reduced by 22%

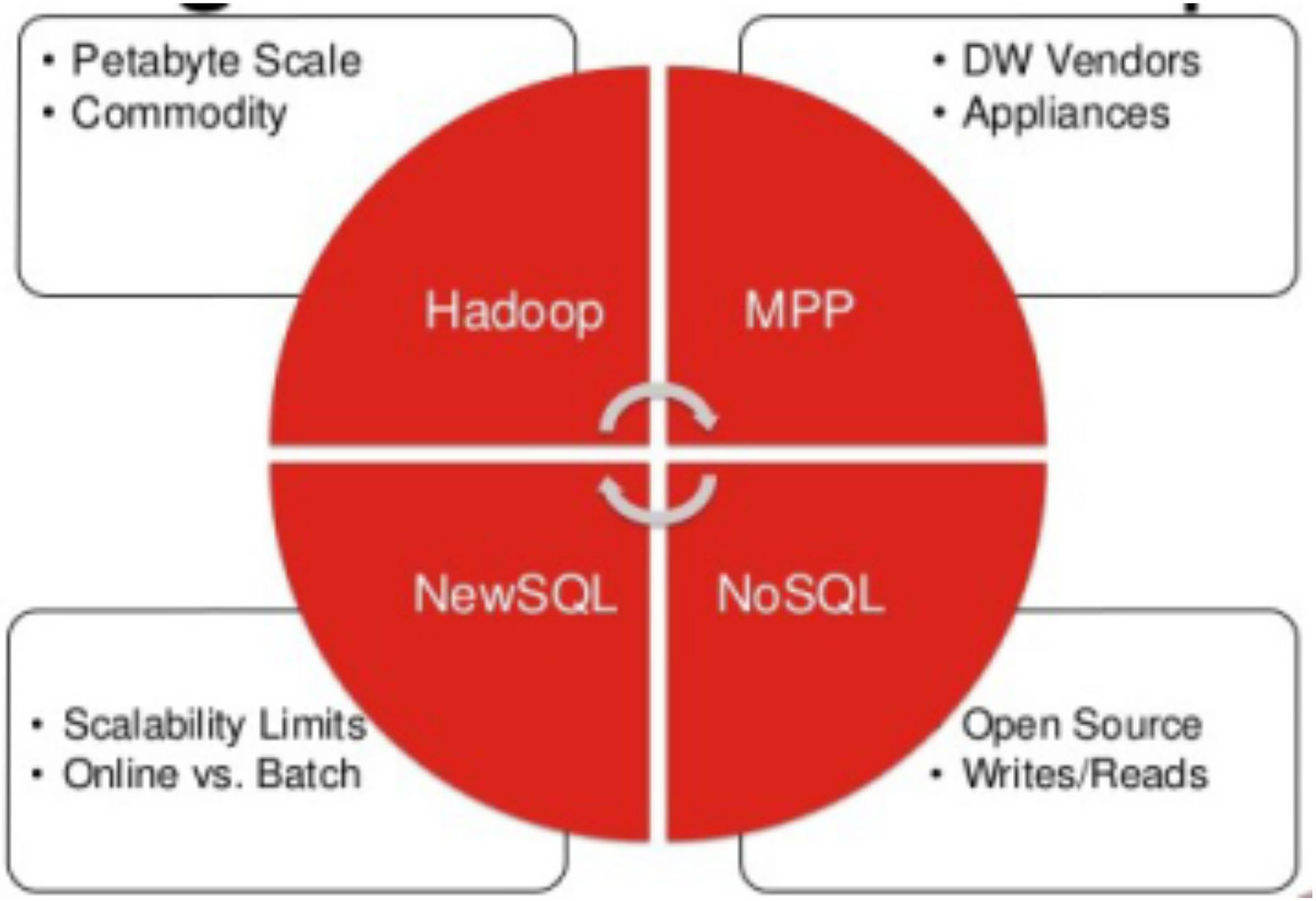


Customer Case Studies on Big Data Lake

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Big Data Technology

Big Data Landscape

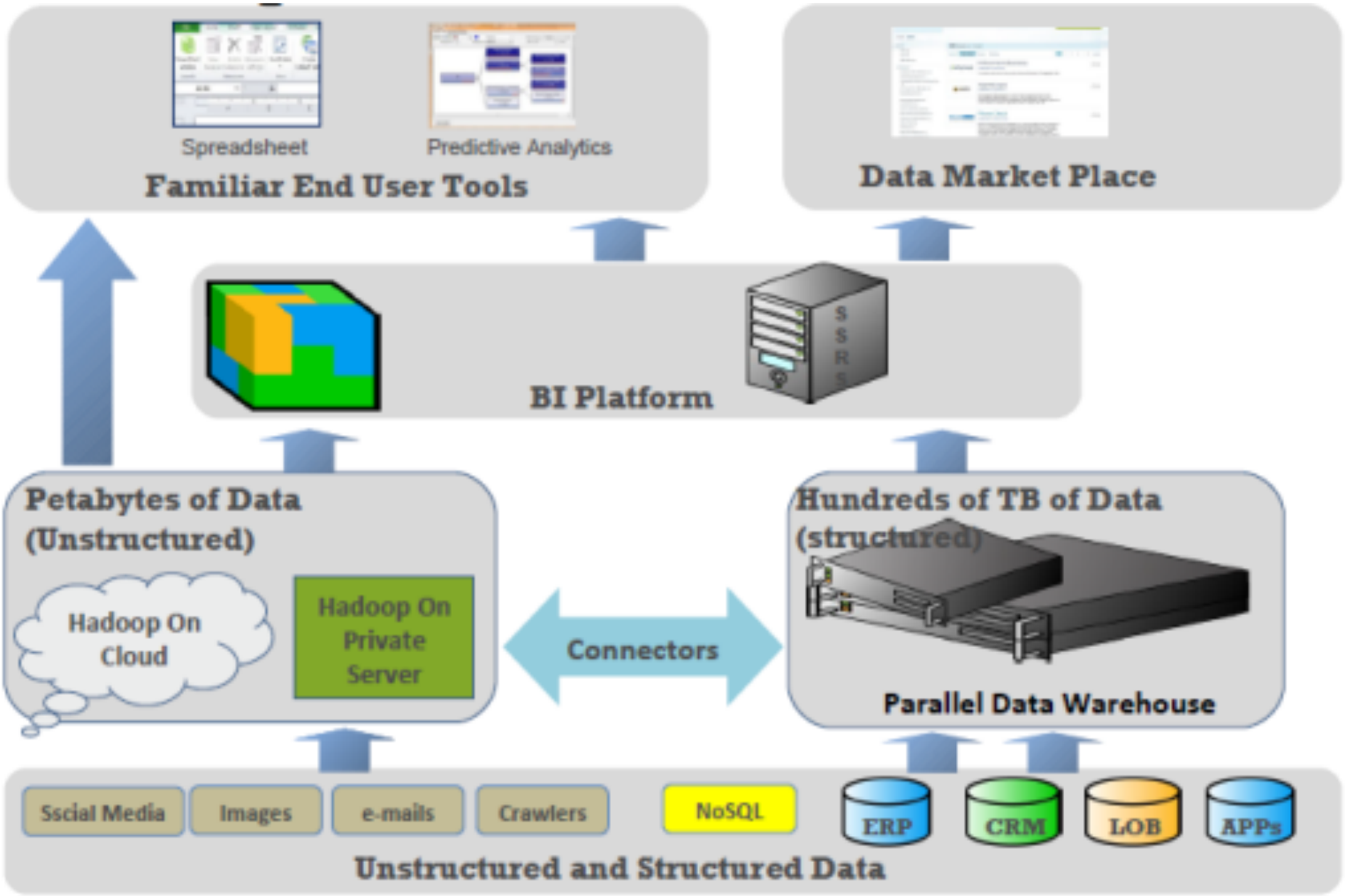


Source: Big Data in the Enterprise. When to Use What?

Big Data Landscape 2017

CHULALONGKORN BUSINESS SCHOOL

Big Data Future Architecture

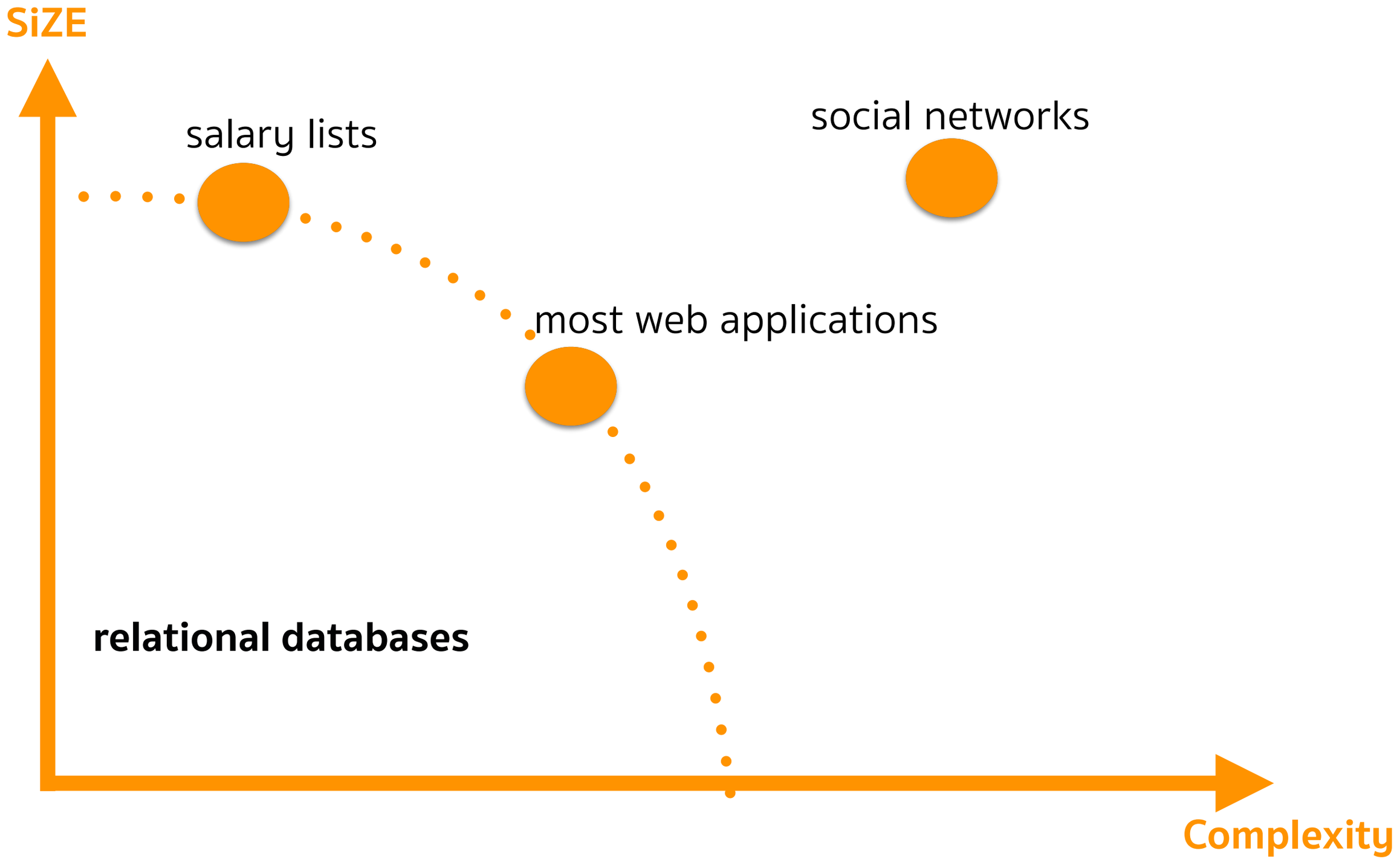


What is NoSQL ?

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

A NoSQL (Not only SQL) database provides
a mechanism for storage and retrieval of data
that is modeled in means other than the tabular relations
used in RDBMS.
Motivations for this approach include
simplicity of design, horizontal scaling,
and finer control over availability.

Source: https://en.wikipedia.org/wiki/NoSQL


SiZE
salary lists
social networks
most web applications
relational databases
Complexity

NoSQL PROS AND CONS

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

PROS

MASSIVE SCALABILITY

HIGH AVAILABILITY

LOWER COST

SCHEMA FLEXIBILITY

STRUCTURED AND SEMI STRUCTURED DATA

CONS

LIMITED QUERY CAPABILITIES

NOT STANDARDISED (PORTABILITY MAY BE AN ISSUE)

STILL A DEVELOPING TECHNOLOGY

Types of NoSQL

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Column-oriented

Key Value Store

Document Store

Graph

Column-oriented databases

Row Oriented
(RDBMS Model)

Multi-valued

id	Name	Age	Interests
1	Ricky		Soccer, Movies, Baseball
2	Ankur	20	
3	Sam	25	Music

null

Column Oriented
(Multi-value sorted map)

id	Name
1	Ricky
2	Ankur
3	Sam

id	Age
2	20
3	25

id	Interests
1	Soccer
1	Movies
1	Baseball
3	Music

Key-value store database

The storage of a value against a key

A key-value store requires the key to be

specified and the key must be known to retrieve the value

Key	Value
Mahesh	{"Mathematics, Science, History, Geography"}
Uma	{"English, Hindi, French, German"}
Paul	{"Computers, Programming"}
Abraham	{"Geology, Metallurgy, Material Science"}

Document-oriented database

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Designed for storing, retrieving, and managing documentoriented information, also known as semi-structured data.

Most of the databases available under this category use XML, JSON, BSON, or YAML

```
{
  "EmployeeID": "SM1",
  "FirstName" : "Anuj",
  "LastName"  : "Sharma",
  "Age"       : 45,
  "Salary"    : 10000000
}
```

```
{
  "EmployeeID": "MM2",
  "FirstName" : "Anand",
  "Age"       : 34,
  "Salary"    : 5000000,
  "Address"   : {
    "Line1"   : "123, 4th Street",
    "City"    : "Bangalore",
    "State"   : "Karnataka"
  },
  "Projects"  : [
    "nosql-migration",
    "top-secret-007"
  ]
}
```

Document-oriented database


```
{
  "articles": [
    {
      "title": "title of the article",
      "articleID": 1,
      "body": "body of the arrticle",
      "author": "Isaac Asimov",
      "comments": [
        {
            "username": "Fritz",
            "join date": "1/4/2014",
            "commentid": 1,
            "body": "this is a great article",
            "replies": [
                {
                    "username": "Freddy",
                    "join date": "11/12/2013",
                    "commentid": 2,
                    "body": "seriously? it's rubbish"
                }
            ]
        },
        {
            "username": "Stark",
            "join date": "19/06/2011",
            "commentid": 3,
            "body": "I don't agree with the conclusion"
        }
      ]
    }
  ]
}
```

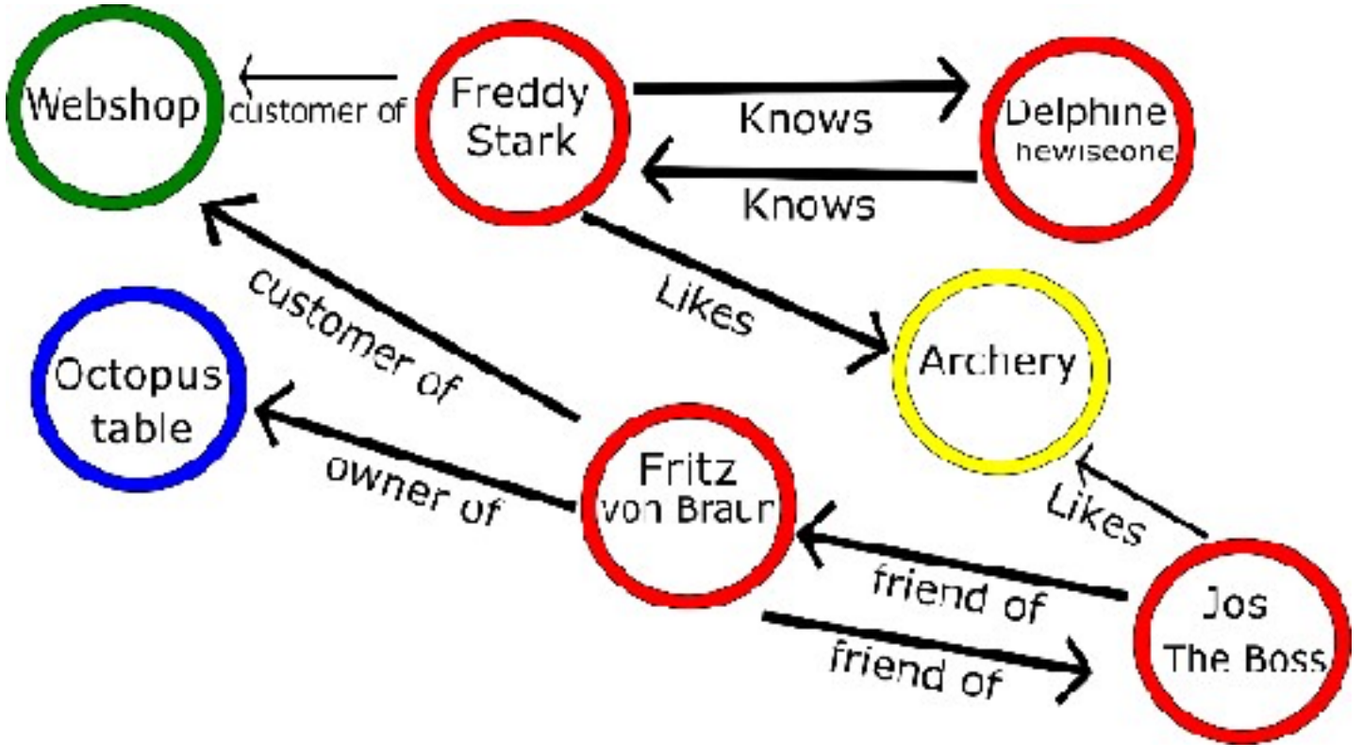
Whereas relational databases chop up data, Document stores save documents as a single entity

CouchDB relax

mongoDB

Graph database

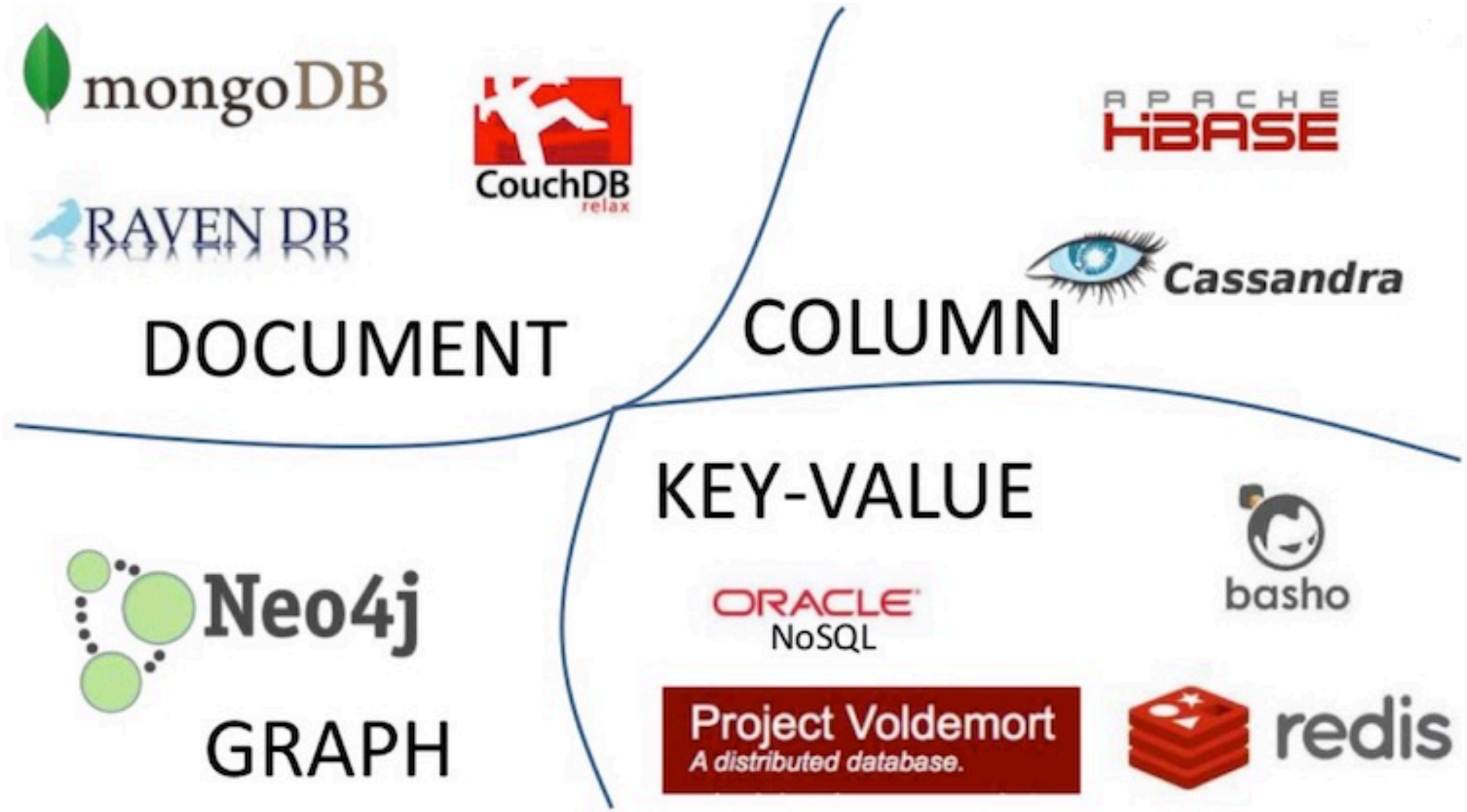
A database that uses graph structures for semantic queries with nodes, edges, and properties to represent and store data.

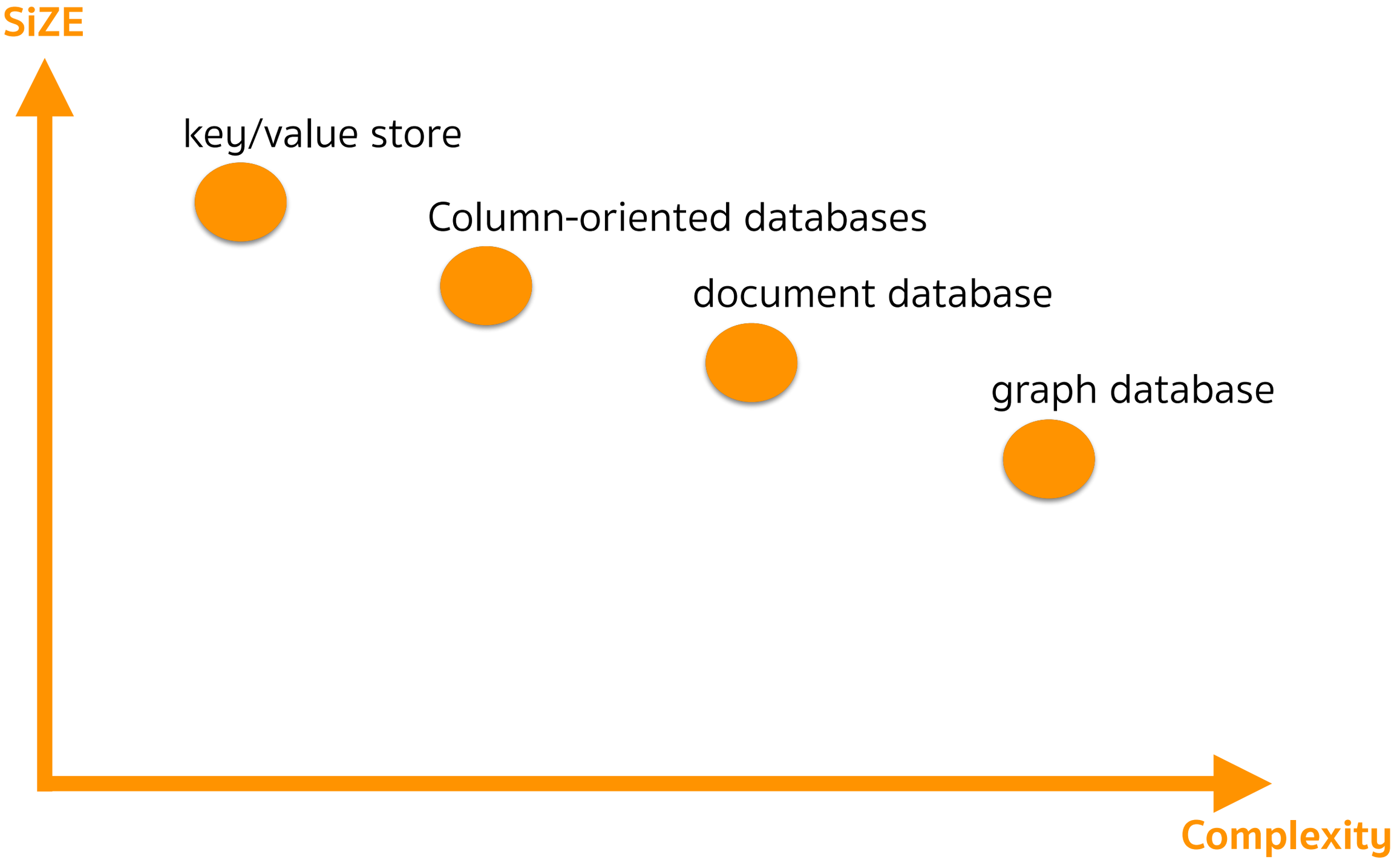


Neo4j
the graph database

Franz Inc.
AllegroGraph

InfiniteGraph


mongoDB
CouchDB
relax
APACHE
HBASE
RAVEN DB
Cassandra
DOCUMENT
COLUMN
KEY-VALUE
Neo4j
GRAPH
ORACLE
NoSQL
basho
Project Voldemort
A distributed database.
redis


SiZE
key/value store
Column-oriented databases
document database
graph database
Complexity

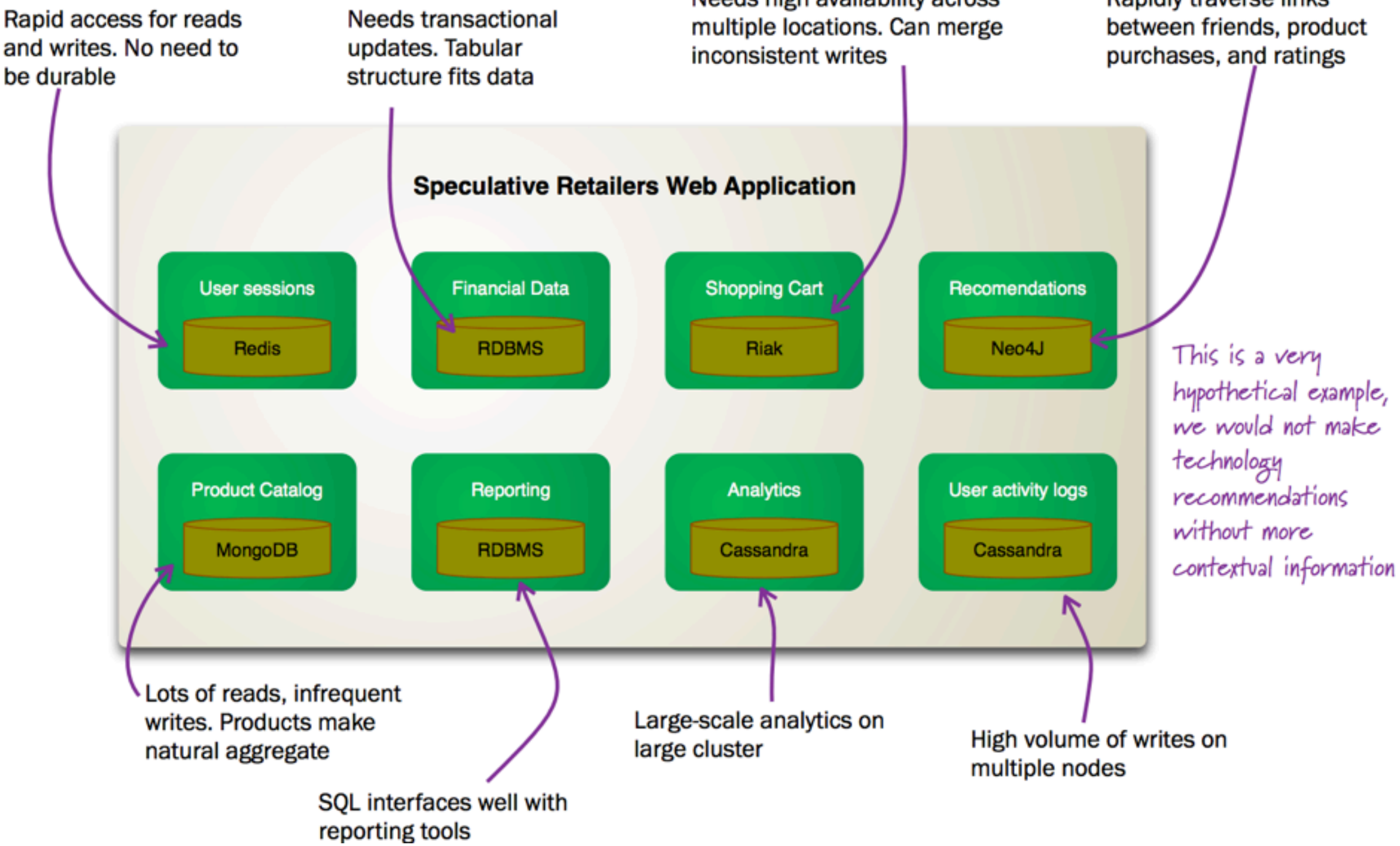
SQL

works great, can’t scale for large data

NoSQL

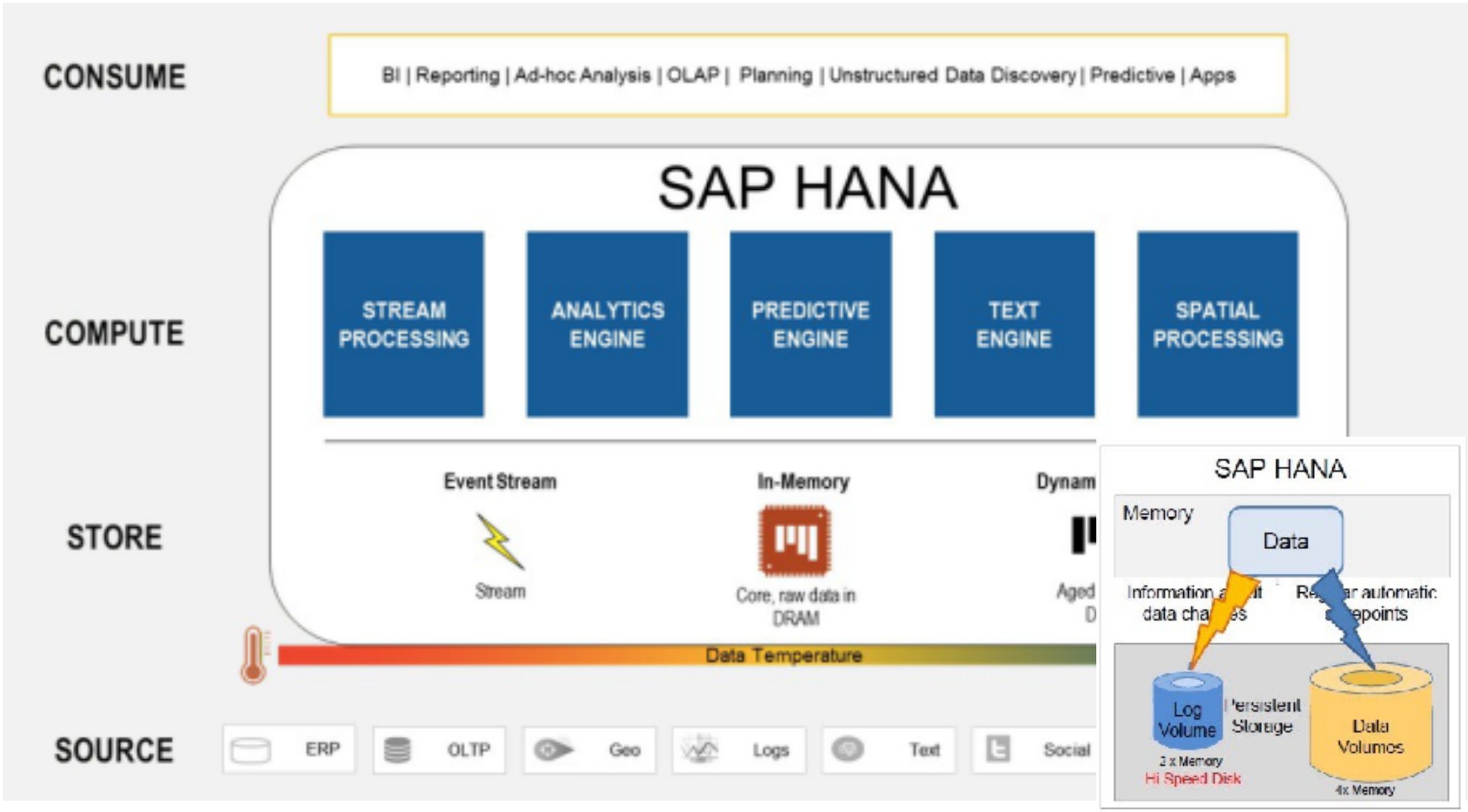
works great, doesn't fit all situations

so use both, but think about when you want to use them!



http://martinfowler.com/bliki/PolyglotPersistence.html

SAP Hana


CONSUME
BI | Reporting | Ad-hoc Analysis | OLAP | Planning | Unstructured Data Discovery | Predictive | Apps
COMPUTE
SAP HANA
STREAM PROCESSING
ANALYTICS ENGINE
PREDICTIVE ENGINE
TEXT ENGINE
SPATIAL PROCESSING
STORE
Event Stream
Stream
In-Memory
Core, raw data in DRAM
Dynam
Aged
Data Temperature
SOURCE
ERP
OLTP
Geo
Logs
Text
Social
SAP HANA
Memory
Data
Log Volume
Persistent Storage
Data Volumes
2 x Memory
Hi Speed Disk
4x Memory

MPP:Oracle Exadata

What is Hadoop ?

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

A scalable fault-tolerant distributed system
for data storage and processing

Completely written in java
Open source & distributed under Apache license

Hadoop Creation History

Doug Cutting adds DFS & MapReduce support to Nutch

Doug Cutting & Mike Cafarella started working on Nutch

NY Times converts 4TB of image archives over 100 EC2s

Fastest sort of a TB, 3.5mins over 910 nodes

Fastest sort of a TB, 62secs over 1,460 nodes
Sorted a PB in 16.25hours over 3,658 nodes

2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009

Google publishes GFS & MapReduce papers

Yahoo! hires Cutting, Hadoop spins out of Nutch

Facebooks launches Hive: SQL Support for Hadoop

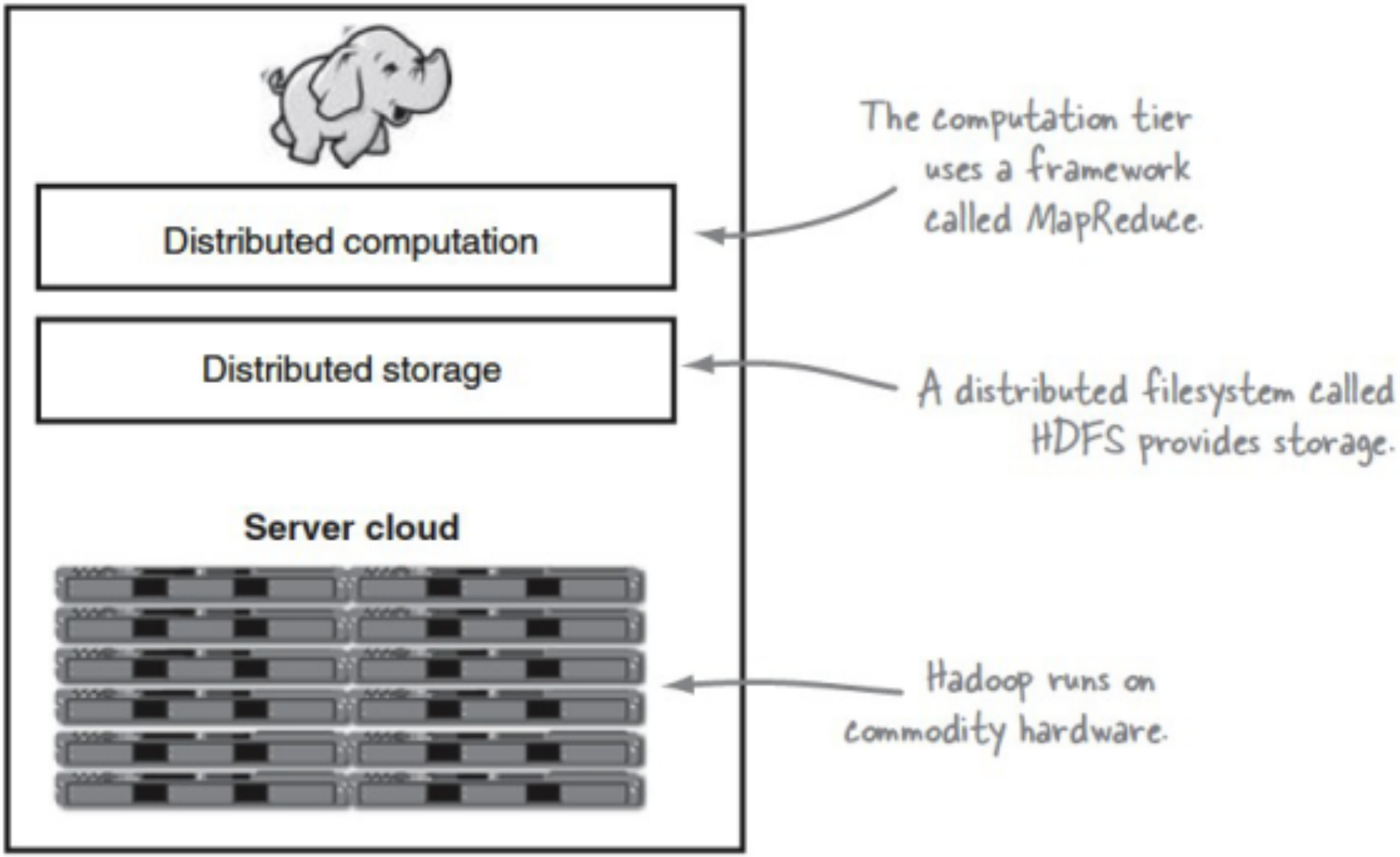
Founded

Doug Cutting joins Cloudera

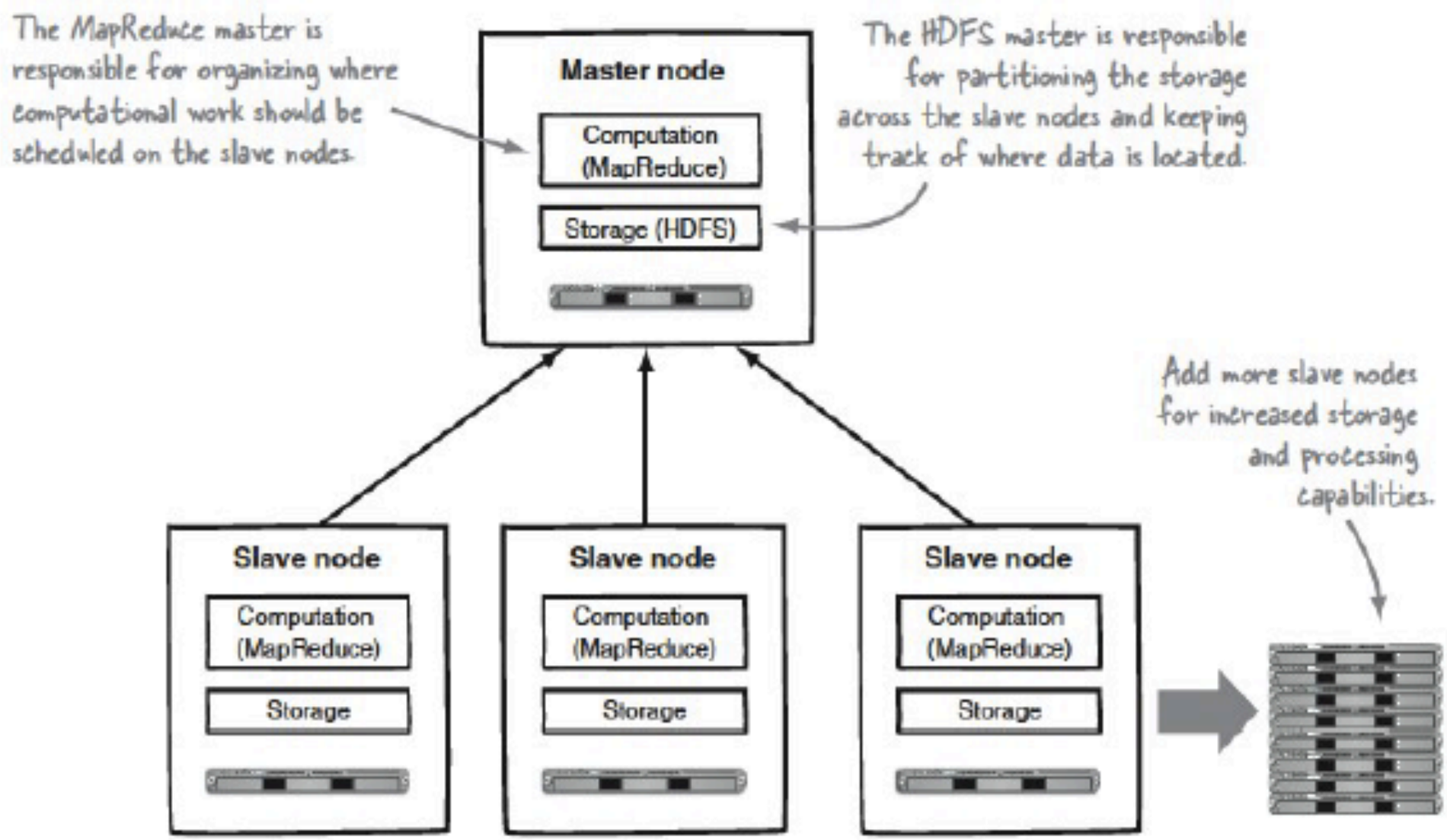
Hadoop Summit 2009, 750 attendees

Source: Cloudera, Inc.

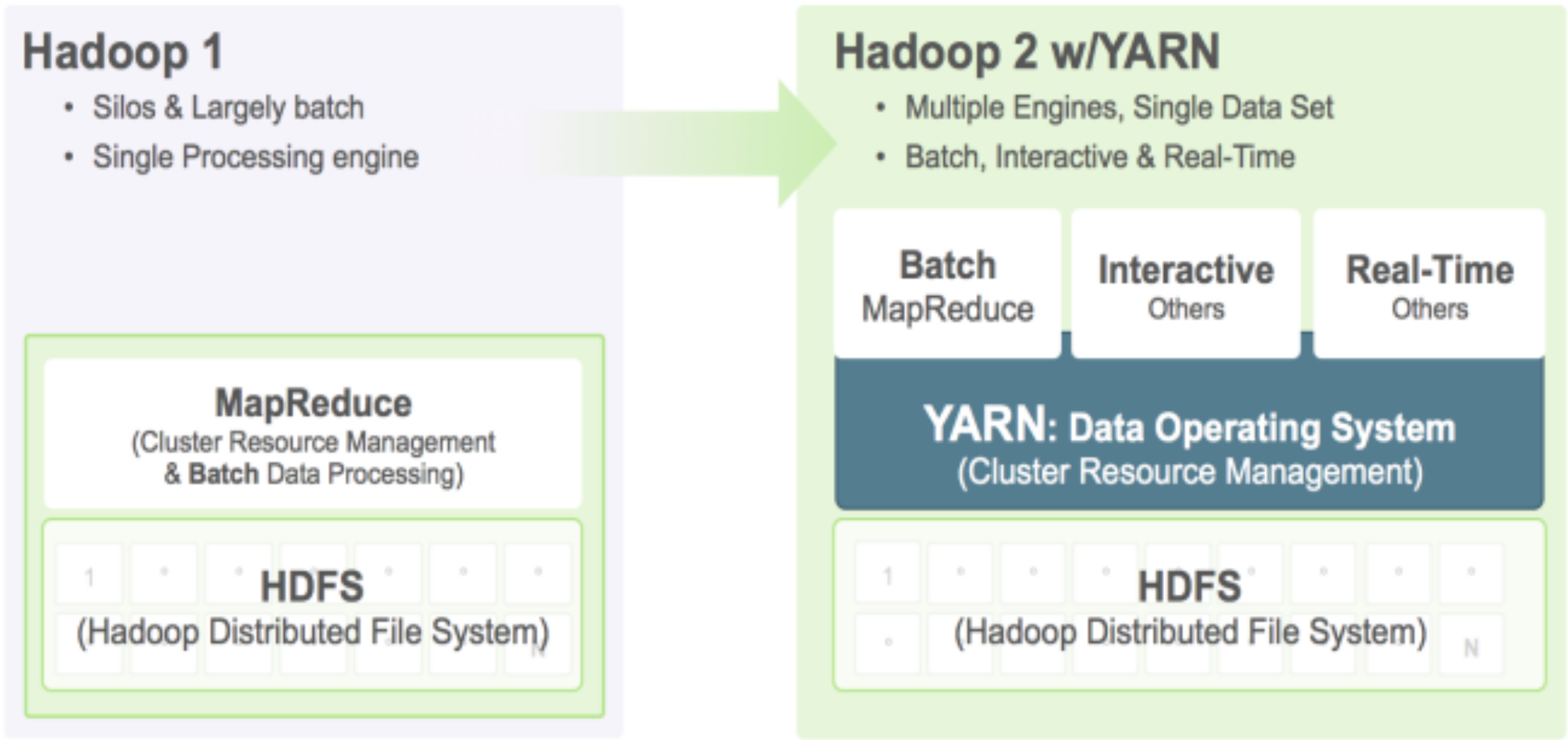
Hadoop Environment



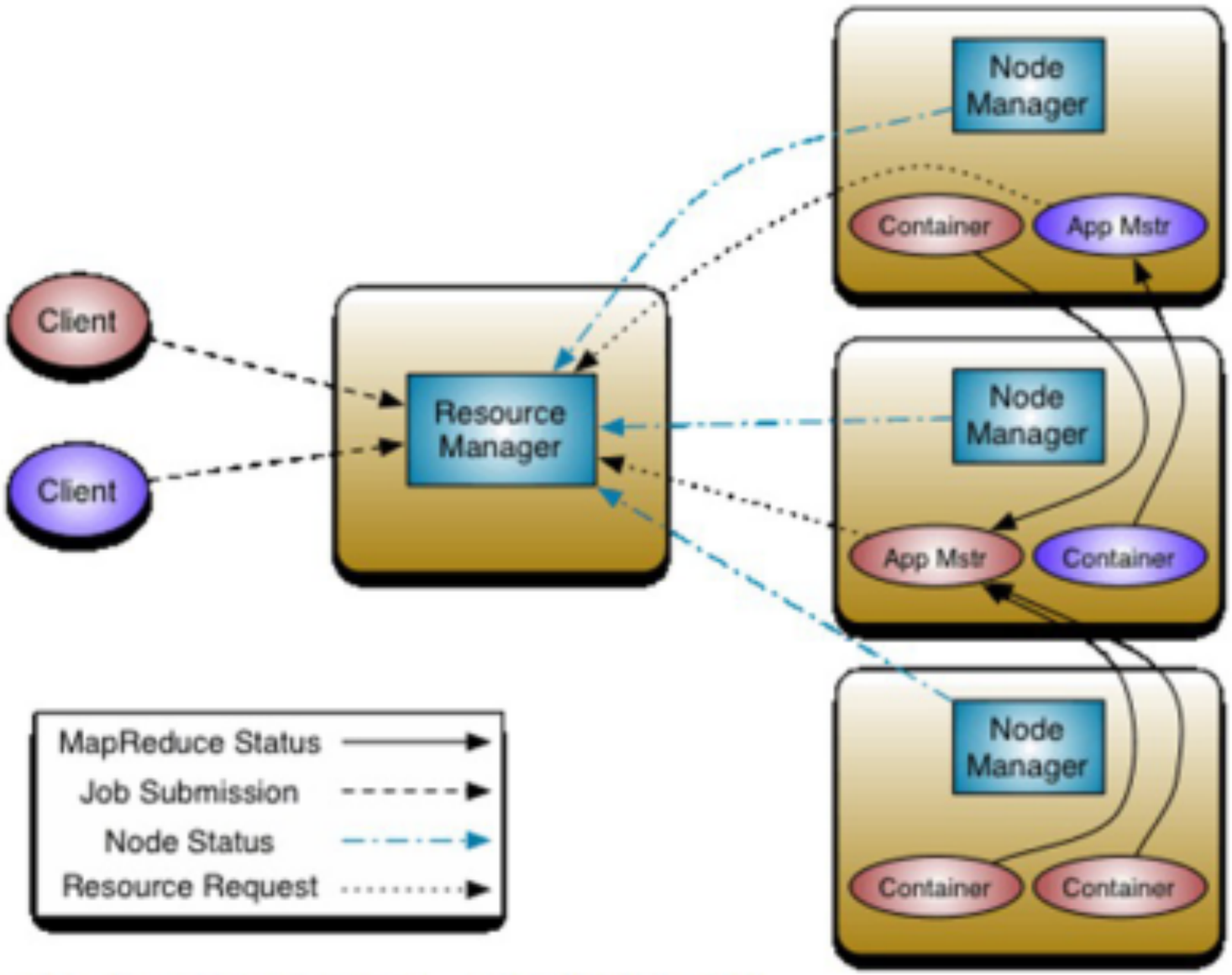
Hadoop Architecture



Hadoop 2.X


Hadoop 1
• Silos & Largely batch
• Single Processing engine
MapReduce
(Cluster Resource Management
& Batch Data Processing)
HDFS
(Hadoop Distributed File System)
Hadoop 2 w/YARN
• Multiple Engines, Single Data Set
• Batch, Interactive & Real-Time
Batch
MapReduce
Interactive
Others
Real-Time
Others
YARN: Data Operating System
(Cluster Resource Management)
HDFS
(Hadoop Distributed File System)

YARN: Yet Another Resource Negotiator

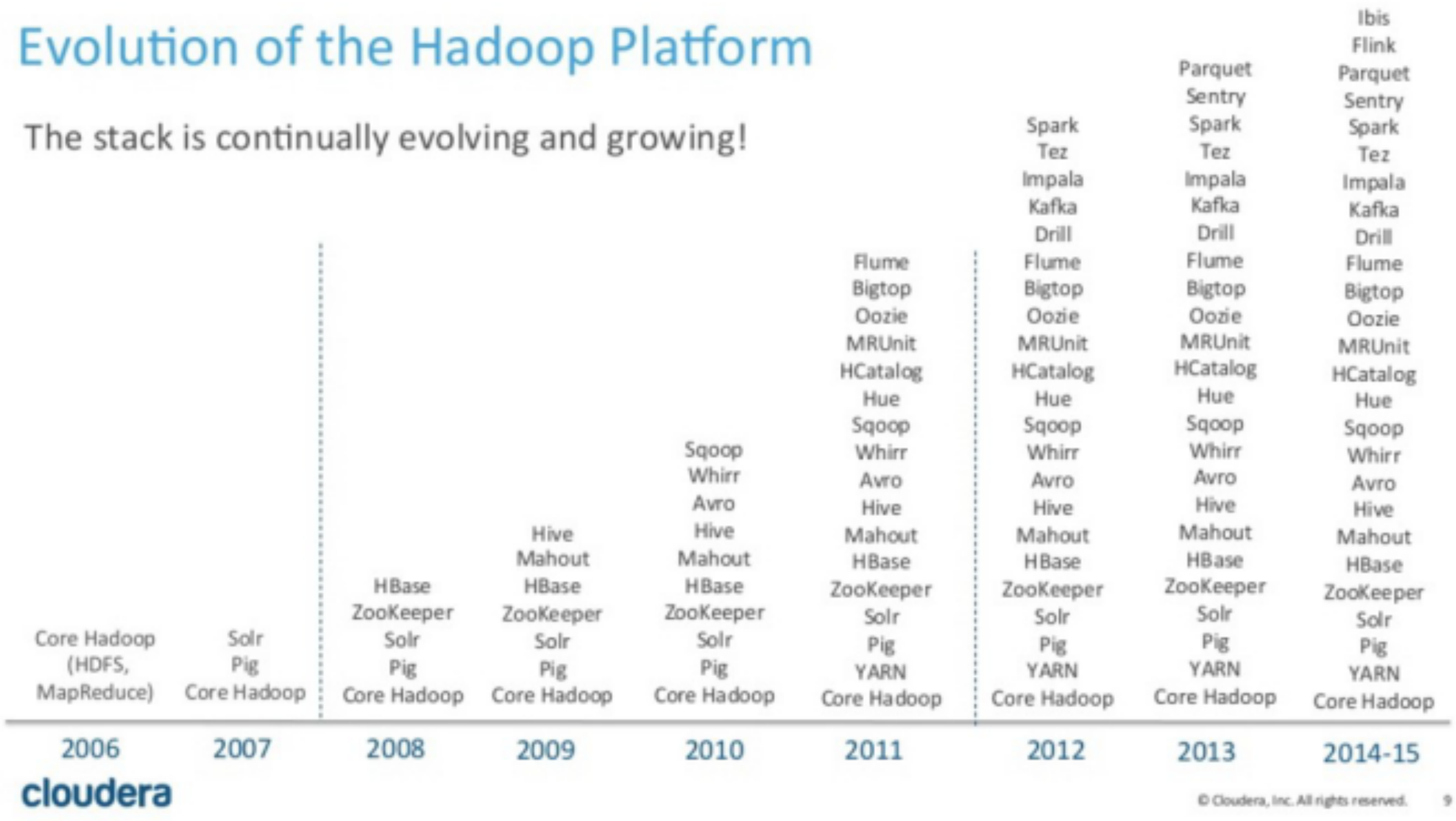


MRV2 maintains API compatibility with previous stable release (hadoop-1.x). This means that all Map-Reduce jobs should still run unchanged on top of MRv2 with just a recompile.

Hadoop.apache.org

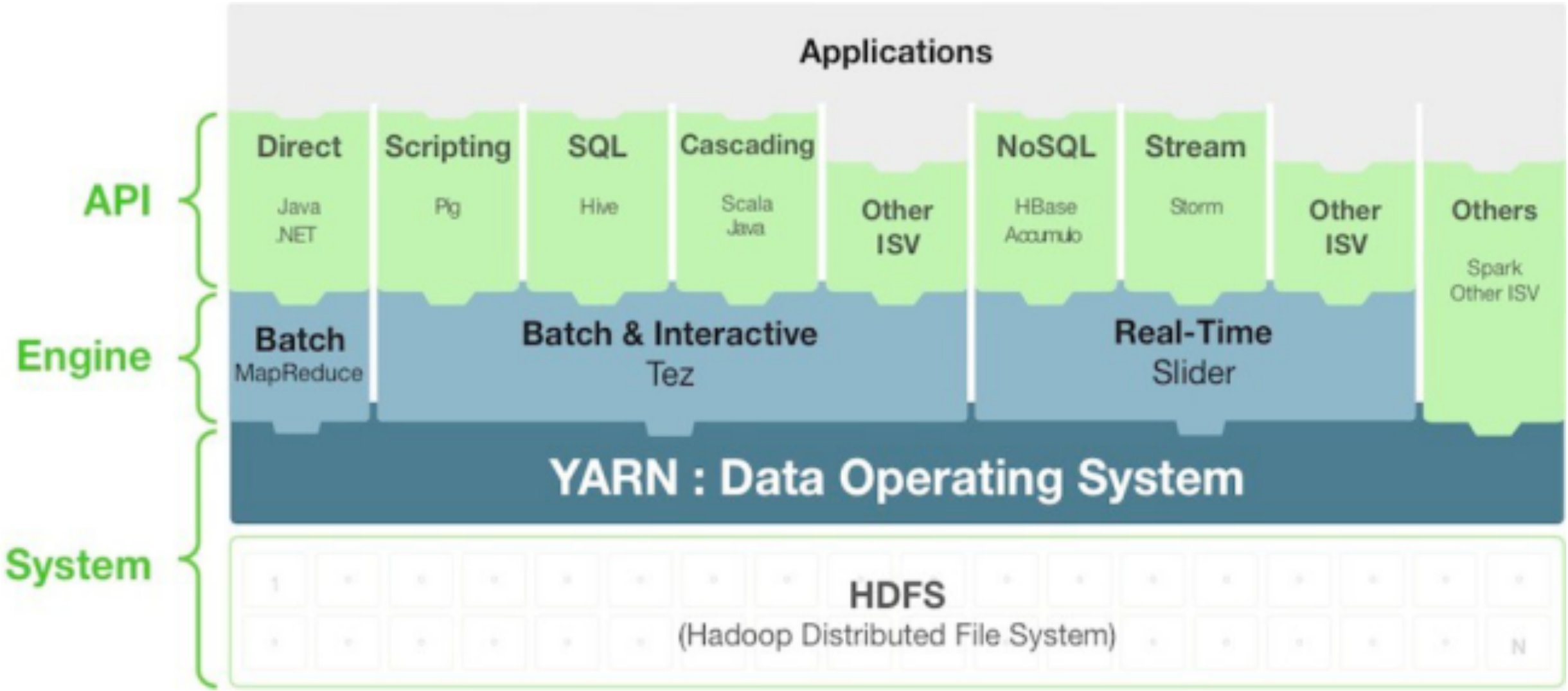
Evolution of the Hadoop Platform

The stack is continually evolving and growing!

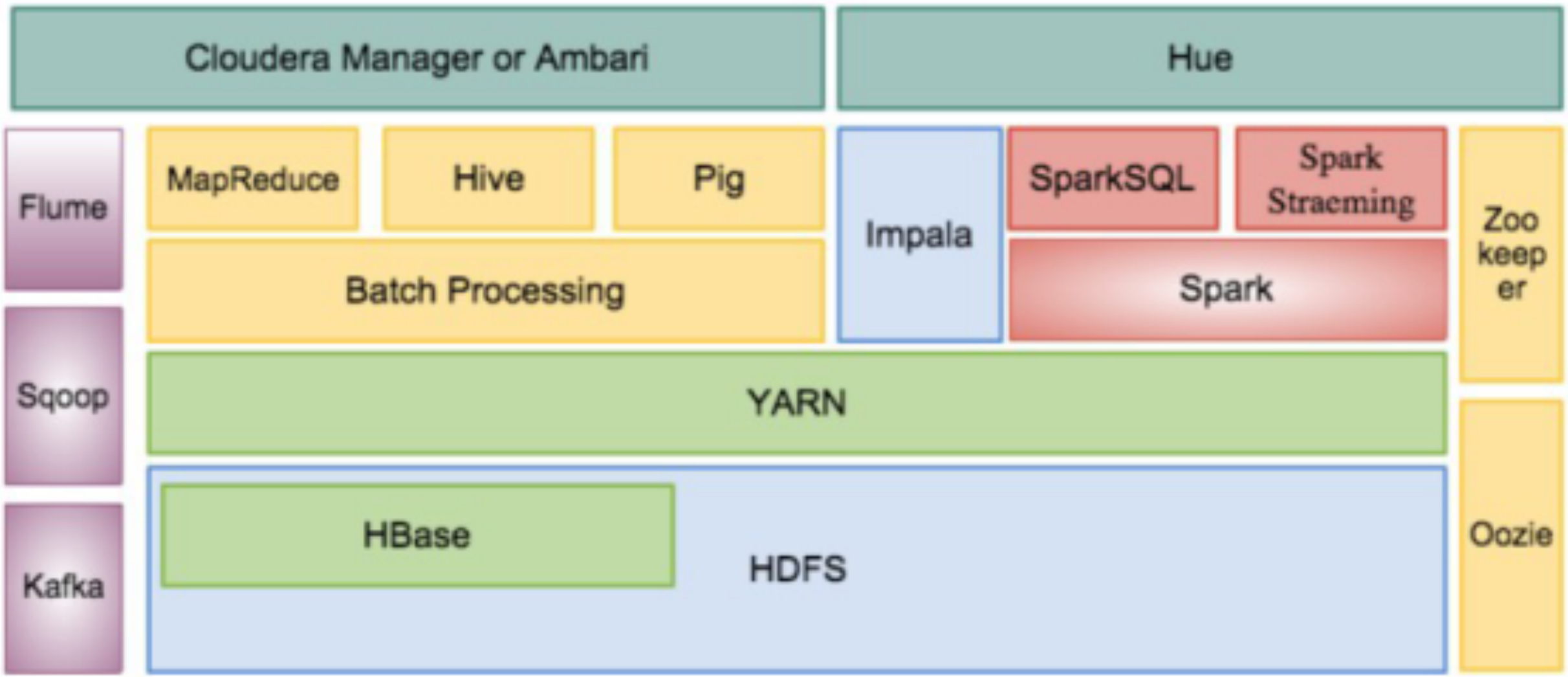


Source: The evolution and future of Hadoop storage: Cloudera

Hadoop 2.x Ecosystems



Hadoop Ecosystems

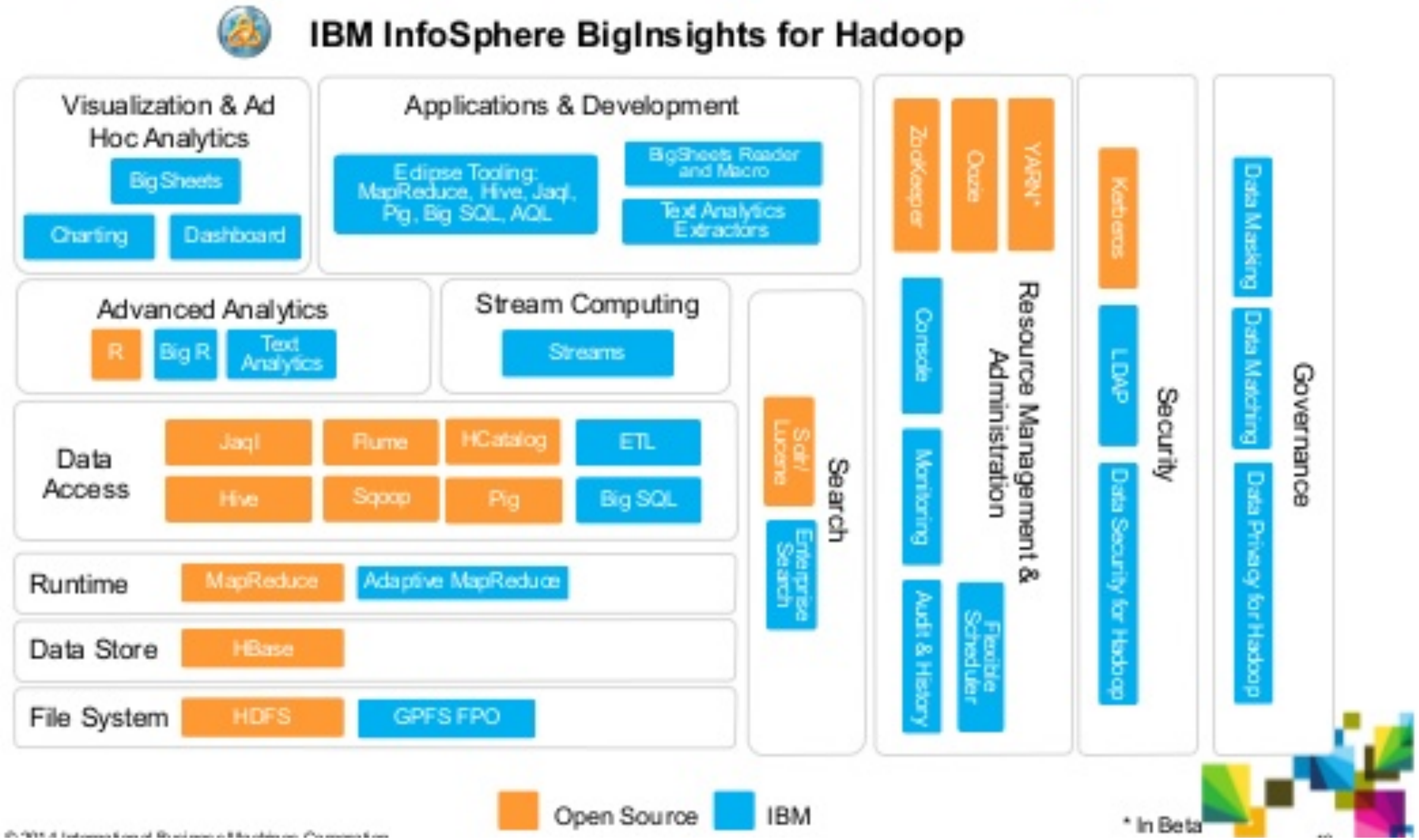


Hadoop Distribution

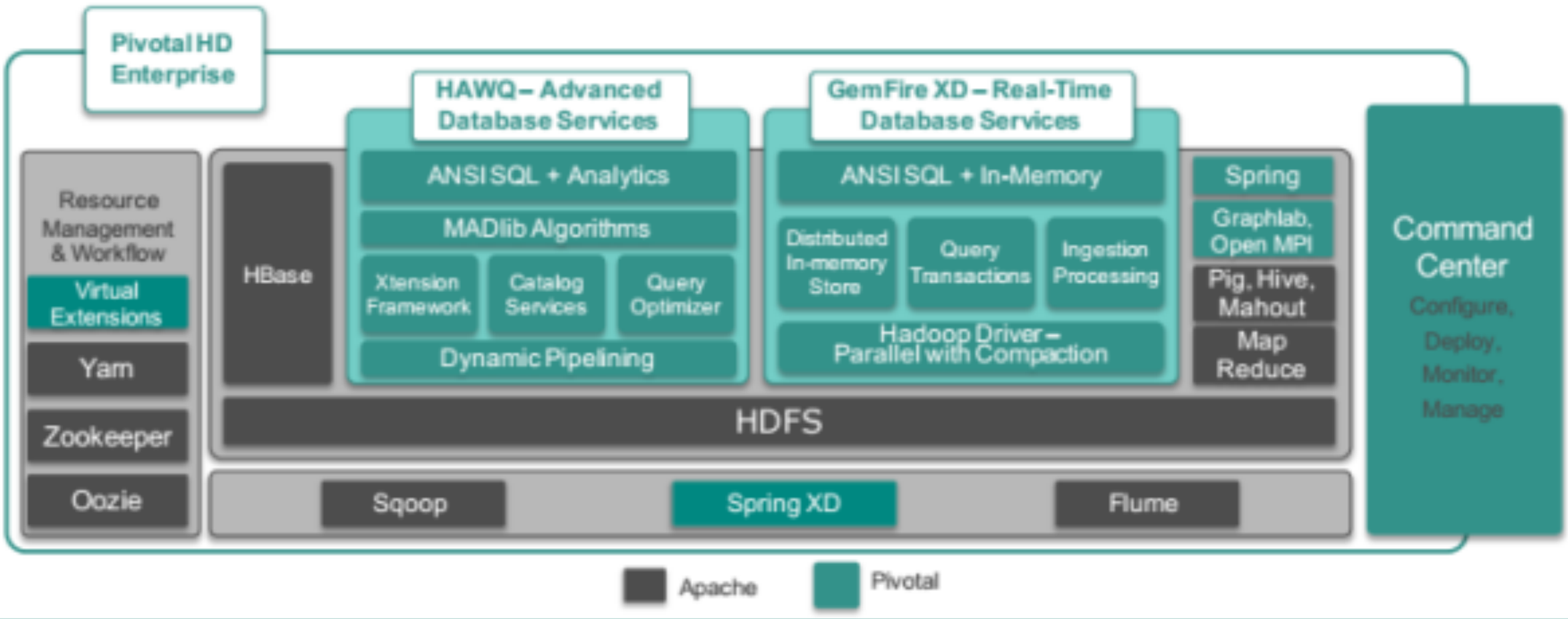
Microsoft Azure

IBM InfoSphere BigInsights

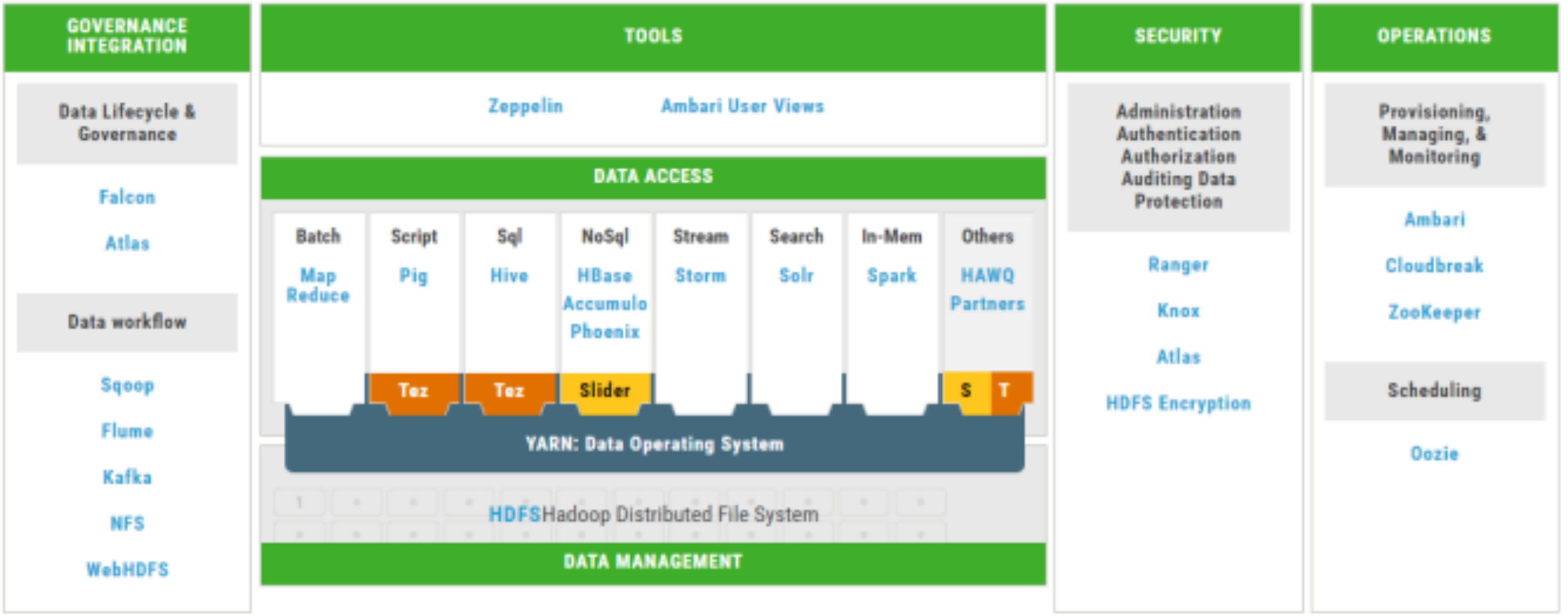
CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Pivotal HD Architecture

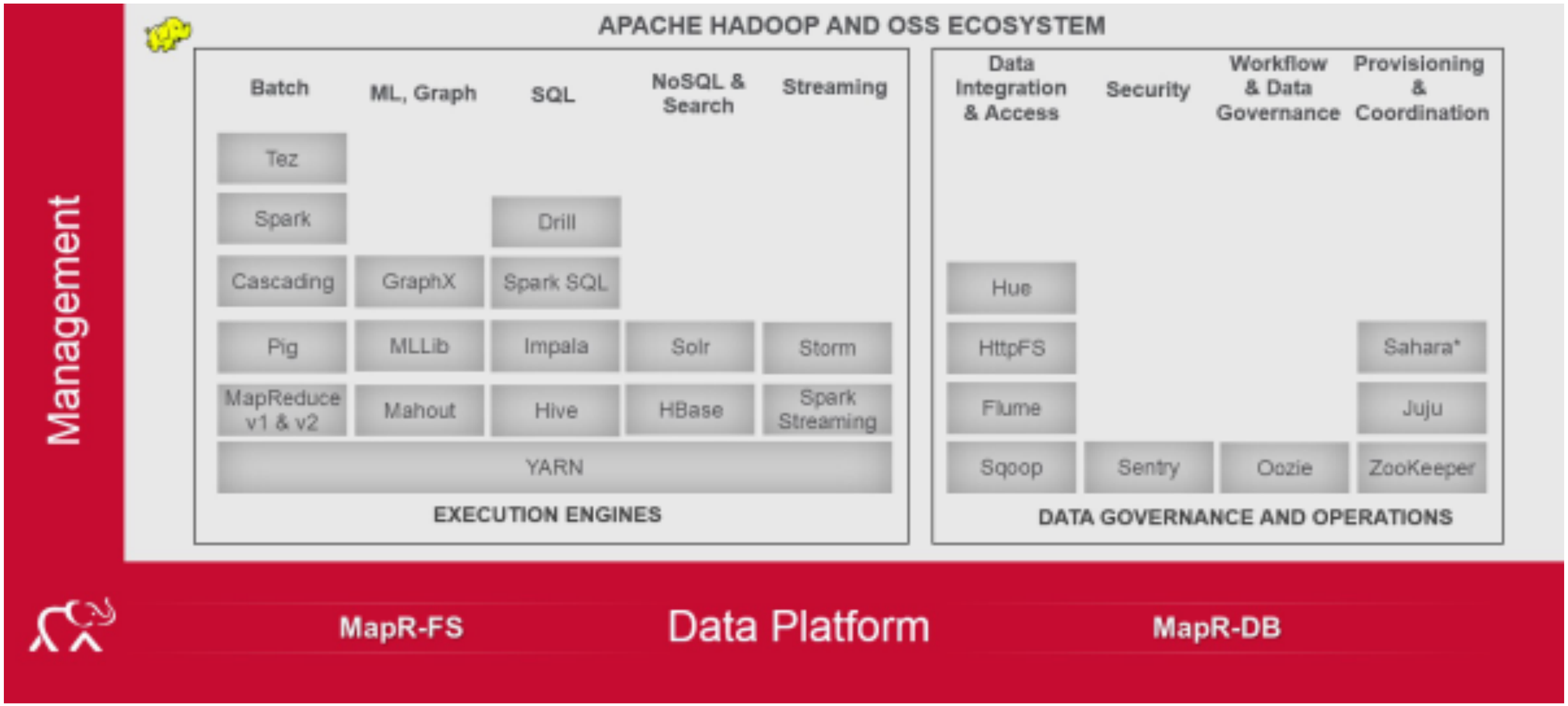


Hortonworks

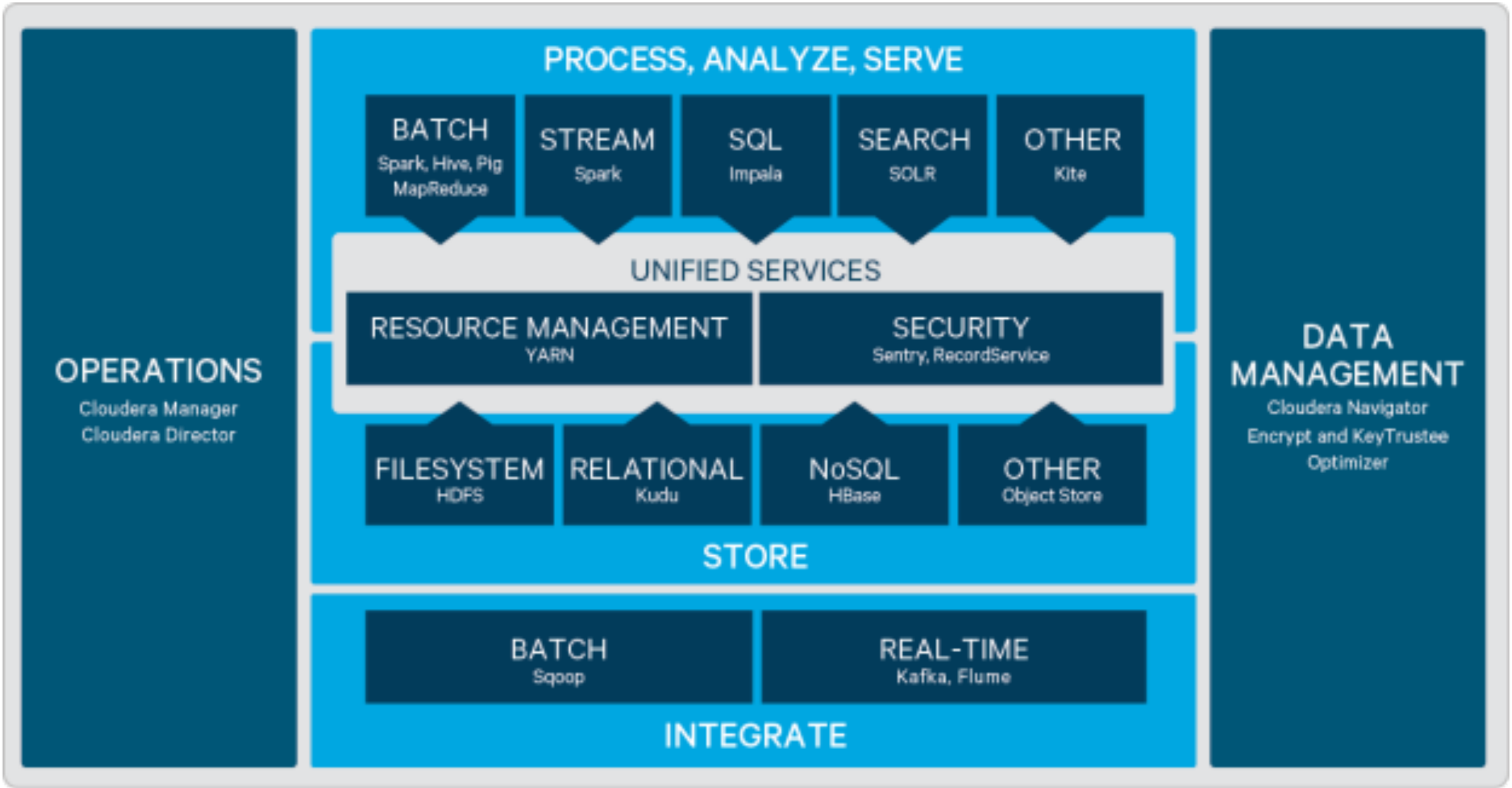


source : http://hortonworks.com/products/data-center/hdp/

MapR


APACHE HADOOP AND OSS ECOSYSTEM
Batch
ML, Graph
SQL
NoSQL & Search
Streaming
Tez
Spark
Drill
Cascading
GraphX
Spark SQL
Pig
MLLib
Impala
Solr
Storm
MapReduce v1 & v2
Mahout
Hive
HBase
Spark Streaming
YARN
EXECUTION ENGINES
Data Integration & Access
Security
Workflow & Data Governance
Provisioning & Coordination
Hue
HttpFS
Sahara*
Flume
Juju
Sqoop
Sentry
Oozie
ZooKeeper
DATA GOVERNANCE AND OPERATIONS
Manage
MapR-FS
Data Platform
MapR-DB

Cloudera


PROCESS, ANALYZE, SERVE
BATCH
Spark, Hive, Pig
MapReduce
STREAM
Spark
SQL
Impala
SEARCH
SOLR
OTHER
Kite
UNIFIED SERVICES
RESOURCE MANAGEMENT
YARN
SECURITY
Sentry, RecordService
FILESYSTEM
HDFS
RELATIONAL
Kudu
NoSQL
HBase
OTHER
Object Store
STORE
BATCH
Sqoop
REAL-TIME
Kafka, Flume
INTEGRATE
OPERATIONS
Cloudera Manager
Cloudera Director
DATA MANAGEMENT
Cloudera Navigator
Encrypt and KeyTrustee
Optimizer

Default Cloudera Services

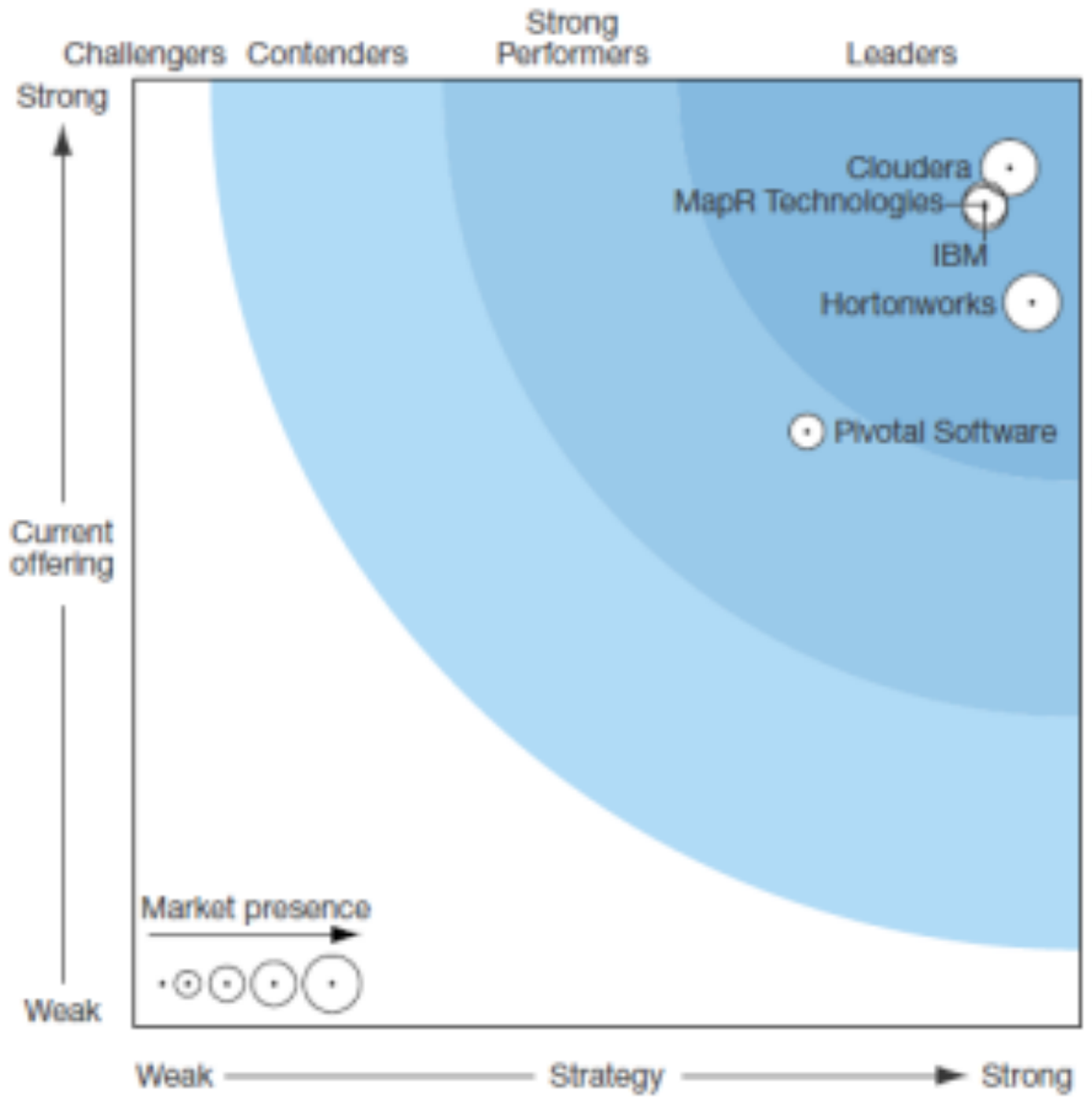
- Cloudera Manager
- HDFS
- YARN
- Apache Hive and Pig
- Apache Flume and Sqoop
- Apache Oozie
- Cloudera Hue
- ZooKeeper

cloudera

The Forrester Wave™

Big Data Hadoop Distributions Q1 2016

Vendor	Product evaluated	Product version evaluated
Cloudera	Cloudera Enterprise	5.50
Hortonworks	Hortonworks Data Platform	2.30
IBM	IBM BigInsights for Apache Hadoop	4.10
MapR Technologies	The MapR Distribution including Apache	5.00
Pivotal Software	HadoopPivotal HD	3.x

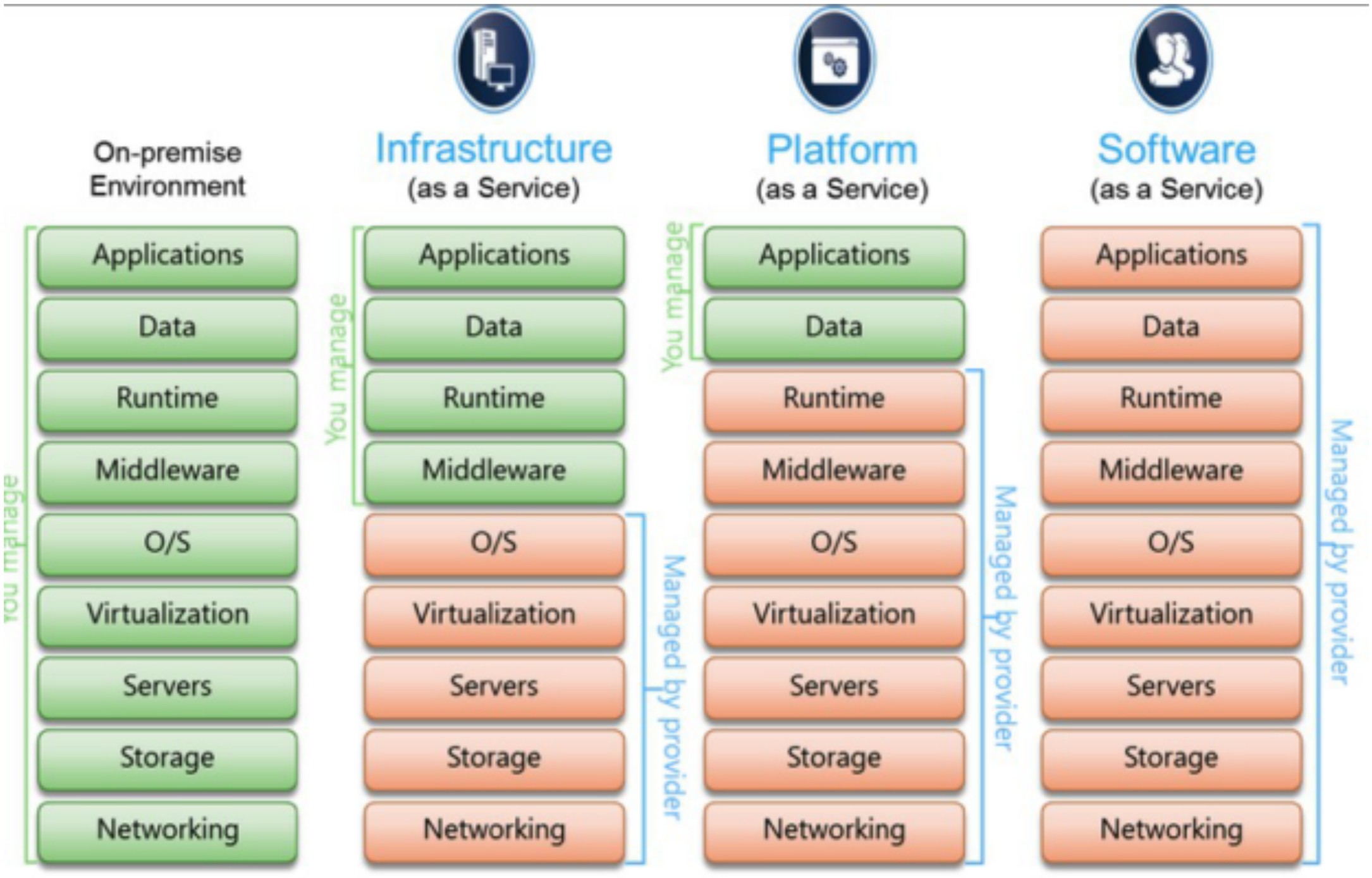


Source: Forrester Research, Inc.; Jan 2016

Issue with Big Data Infrastructure

- Large investment
- Scalability
- ROI
- Business Cases

Cloud Technology

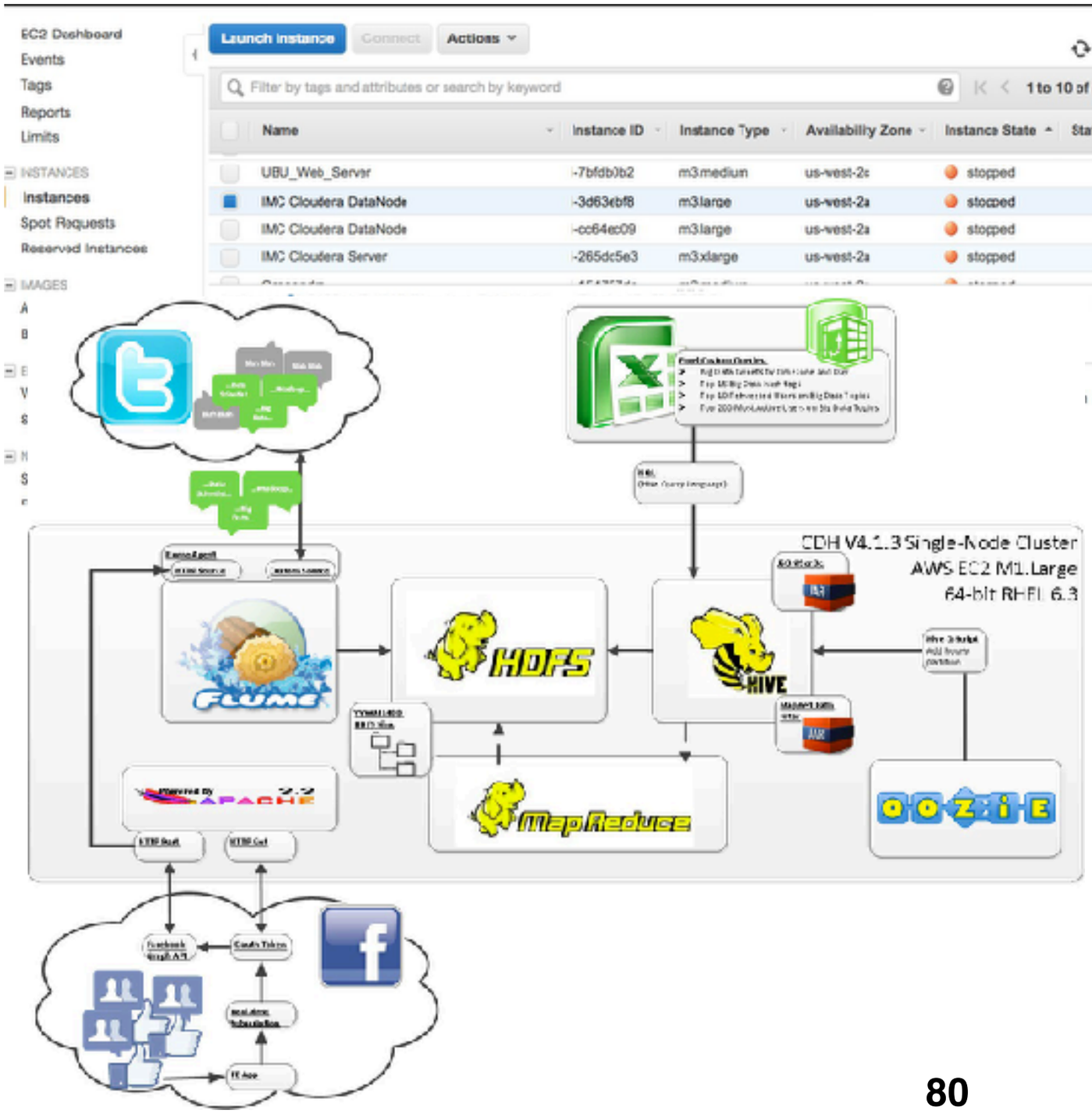


http://www.markerbench.com/blog/2013/01/08/static-blogging/

Big Data on Cloud

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

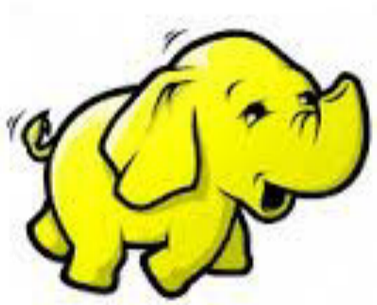
Big Data using IaaS



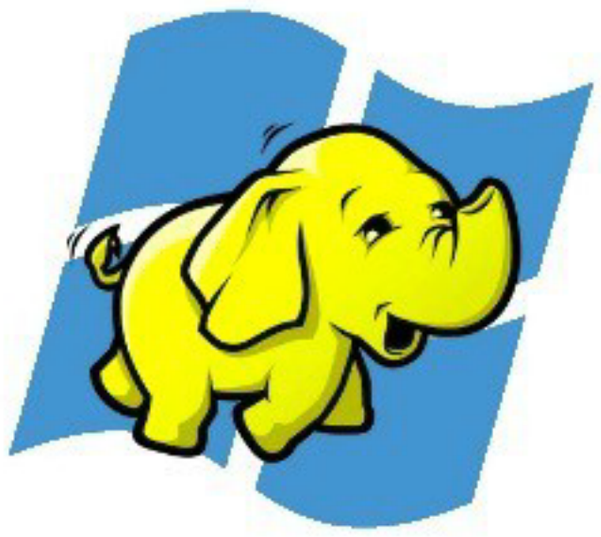
Big Data on Cloud

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Using Big Data as a Services



Amazon
Elastic Mapreduce



Microsoft Azure Hadoop