CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Introduction To Big Data

Thaveewat Khanan

http://bit.ly/2v7OZTh

bobbylovemovie@gmail.com

Bobby Thaveewat

Dhipaya Insurance PCL (TIP)

Data Science Specialist

Dhipaya Insurance PCL (TIP)

Jul 2018 – Present · 10 mos

Bangkok Metropolitan Area, Thailand

Senior System Analyst(IT-CIS Finance Delivery Big Data)

True Corporation

Jun 2016 – Present · 2 yrs 11 mos

Bangkok Metropolitan Area, Thailand

Programmer Analysis, System Analysis

CDG Systems (CDGS)

May 2005 – Present · 14 yrs

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Big Data

Big Data Analytics

Data Science

Machine Learning

Artificial Intelligence

Deep Learning

“นายกฯ” ชี้ ใช้ "Big Data" บริหารราชการแผ่นดิน

15 May 2018

What is Big Data ?

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

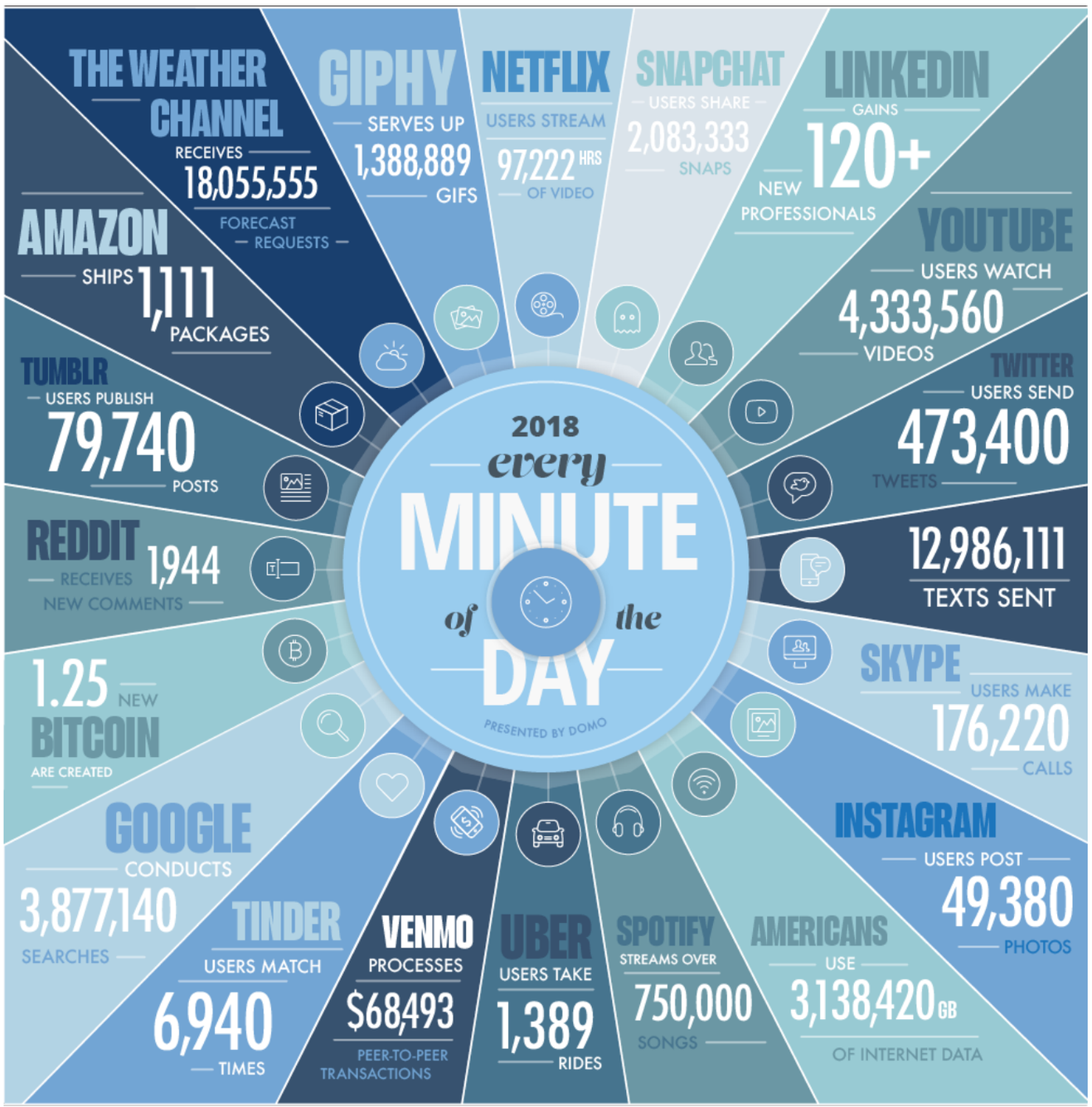
“ Big data is data that exceeds the processing capacity of conventional database systems.
The data is too big, moves too fast,
or doesn’t fit the structures of your database architectures.
To gain value from this data,
you must choose an alternative way to process it. ”

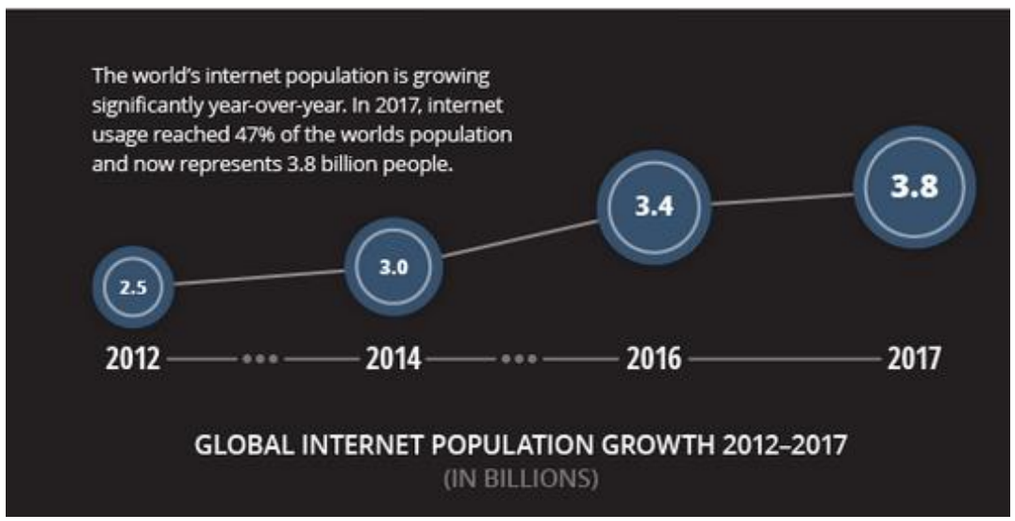
Big Data Now: O'Reilly Media

Why Big Data ?

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

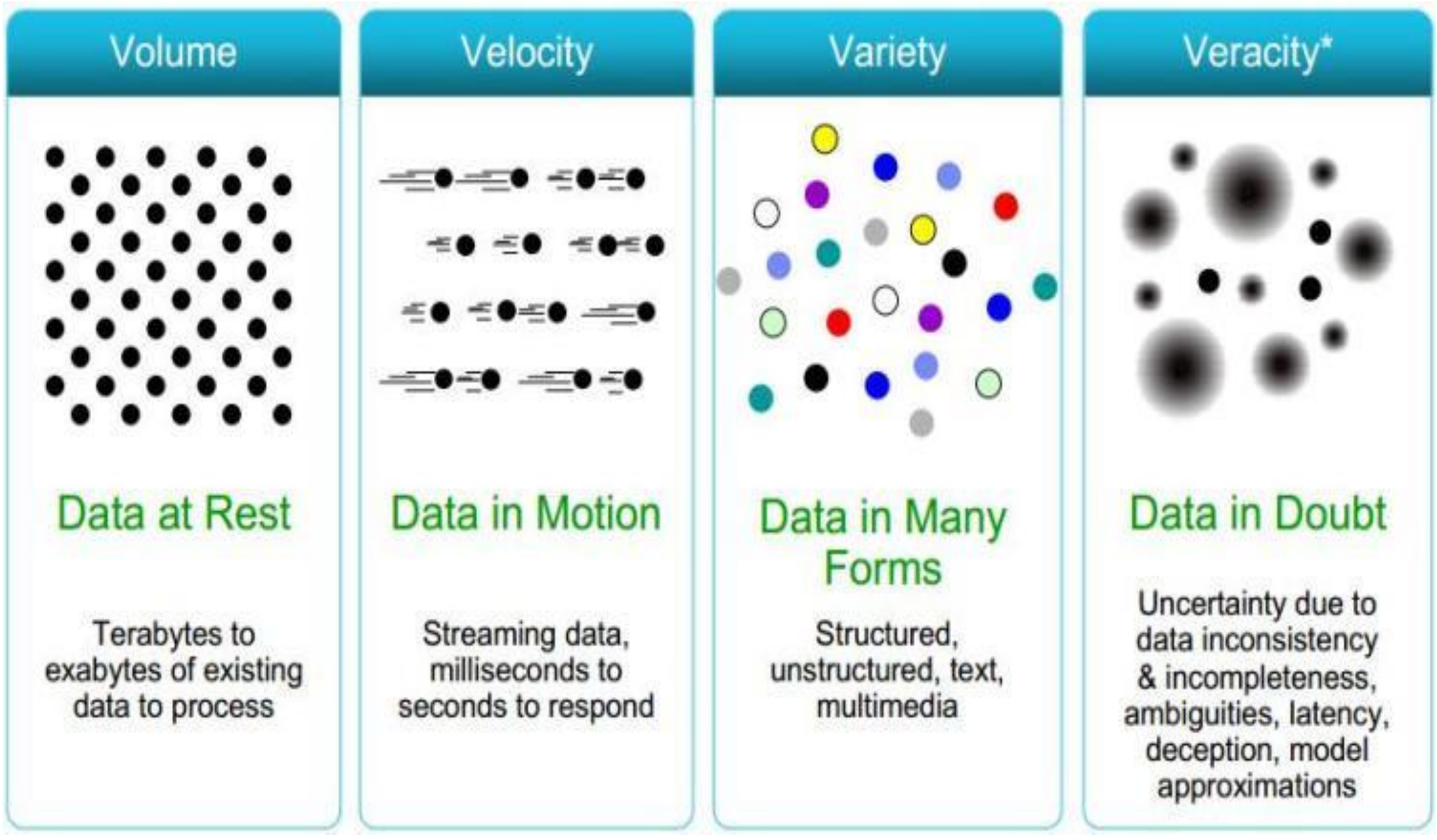
DATA NEVER SLEEP 6.0



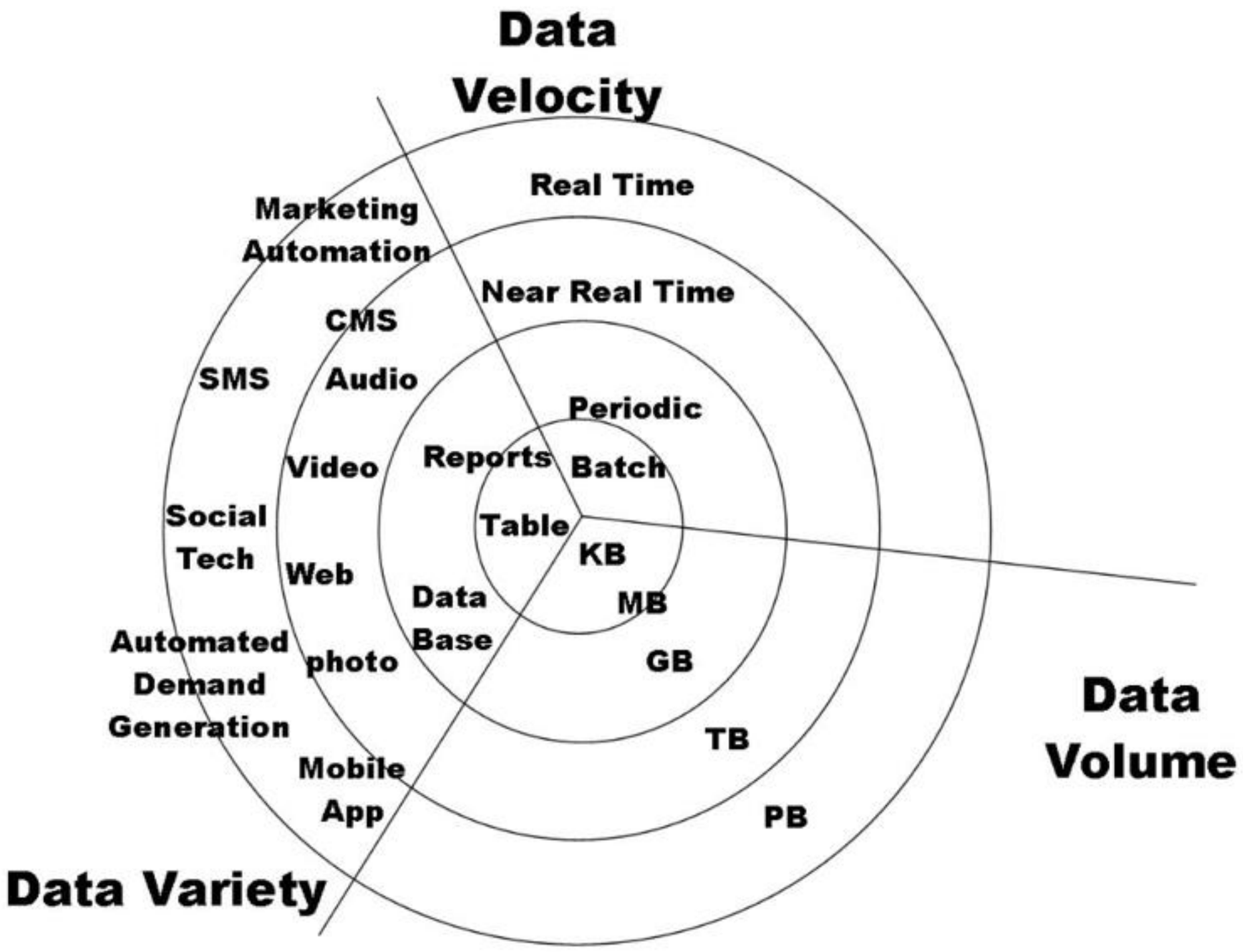


https://www.domo.com/blog/2016/06/data-never-sleeps-4-0/

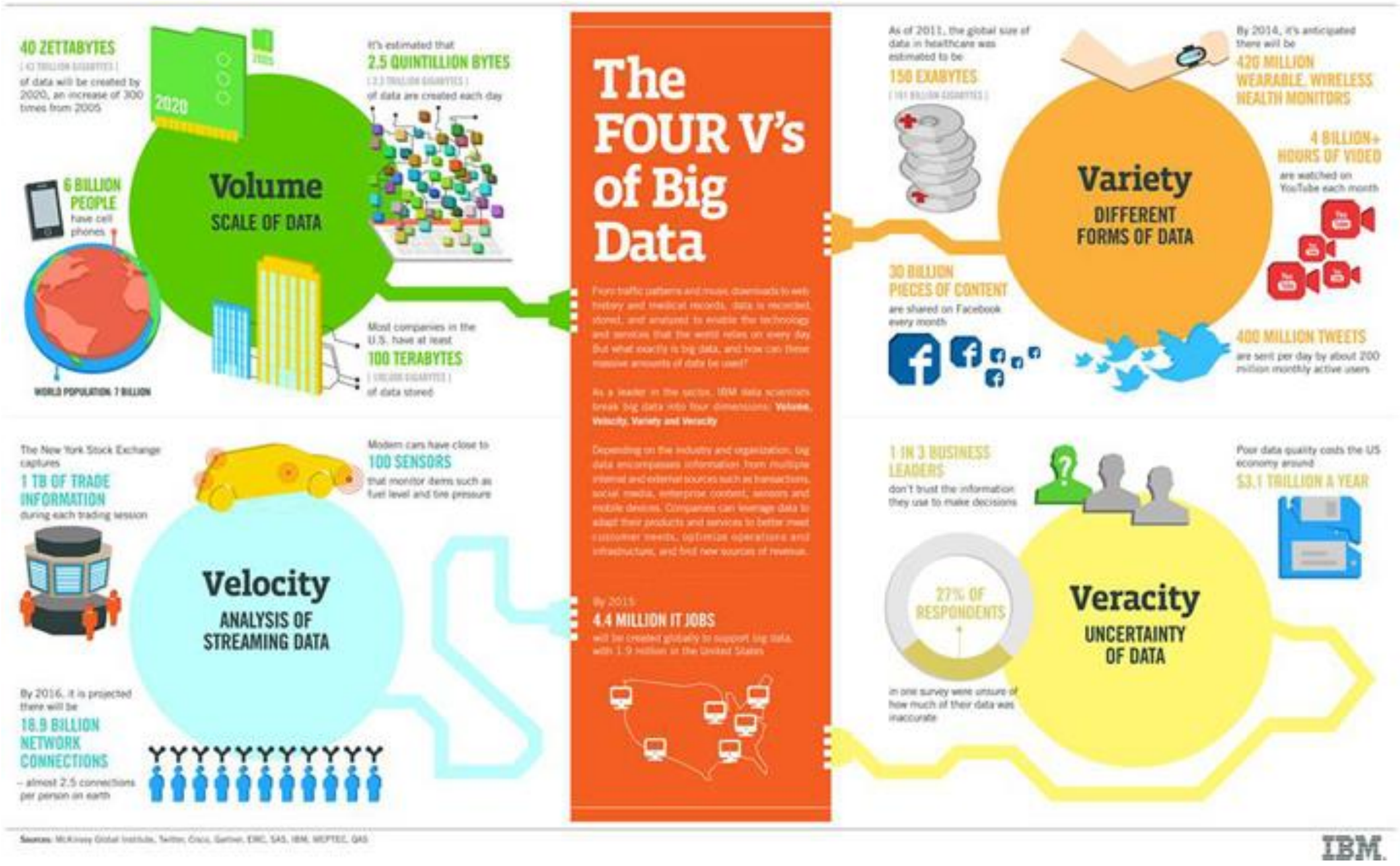
Three Characteristics of Big Data

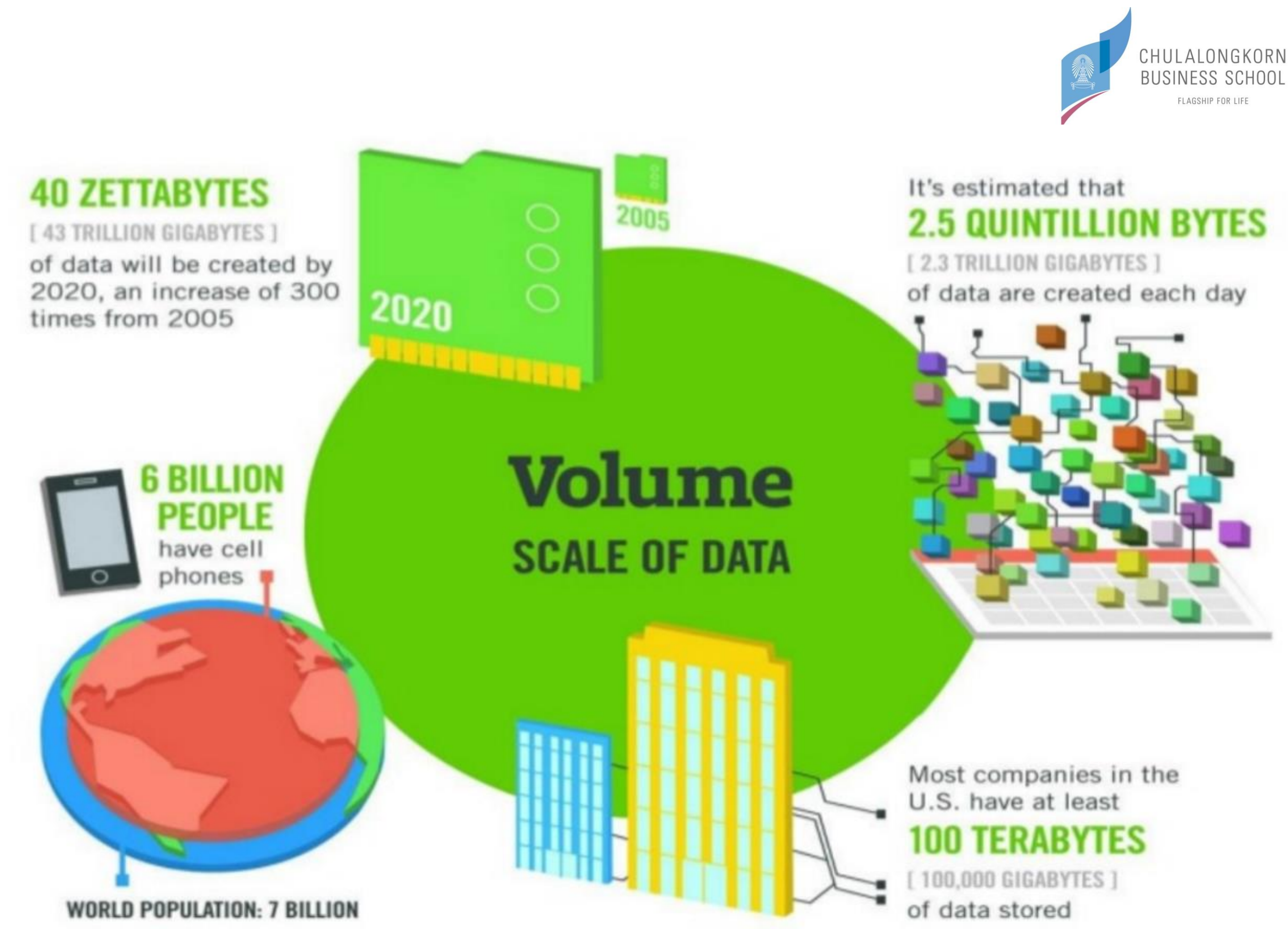


Source Introduction to Big Data: Dr. Putchong Uthayopas


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

4Vs of Big Data




40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
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

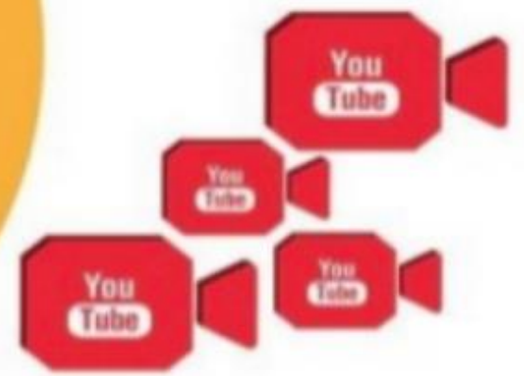
DIFFERENT FORMS OF DATA

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

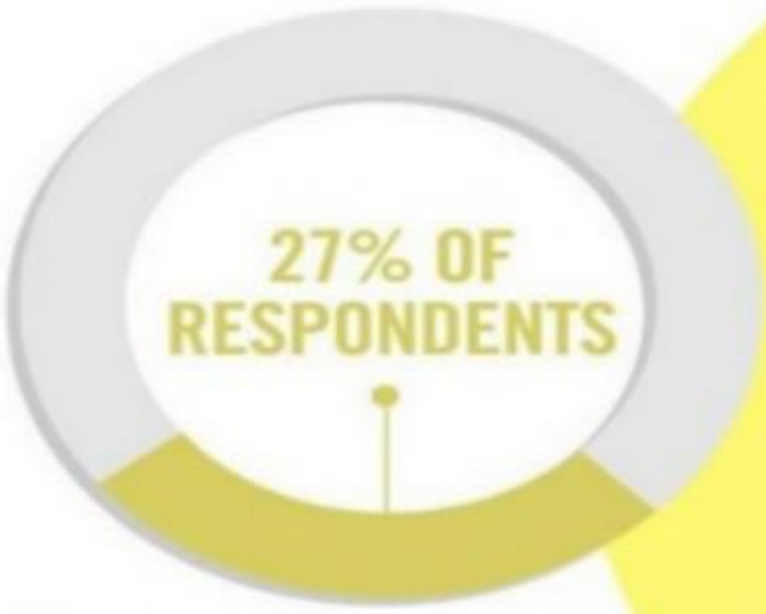
are sent per day by about 200 million monthly active users

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

Visualization

The 'Datafication' of our World;

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

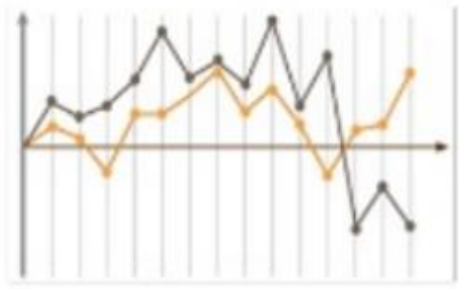
Volume

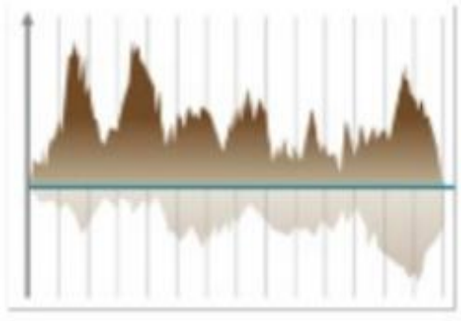
Velocity

Variety

Veracity

Analysis





Analysing Big Data:

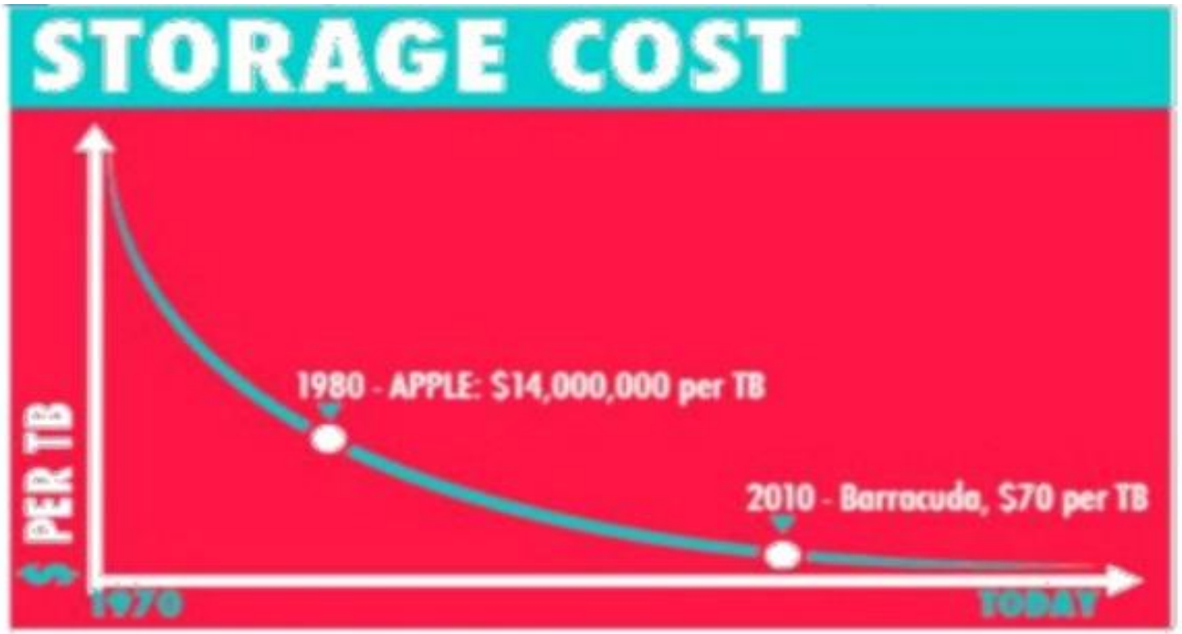
- Text analytics
- Sentiment analysis
- Face recognition
- Voice analytics
- Movement analytics
- Etc.

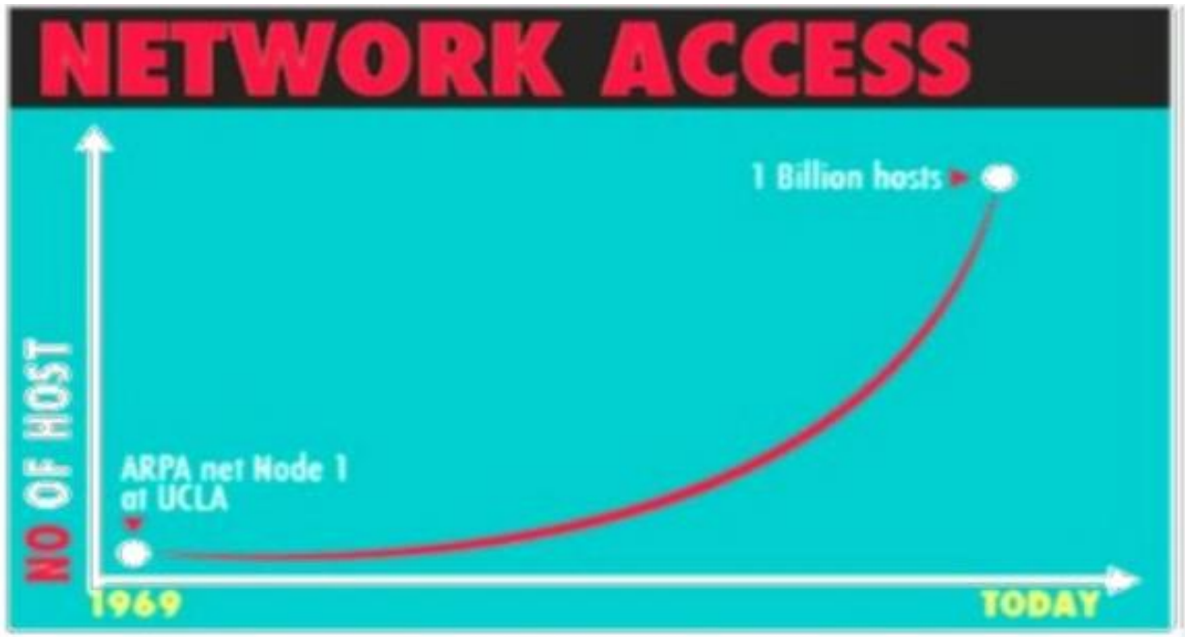
Value

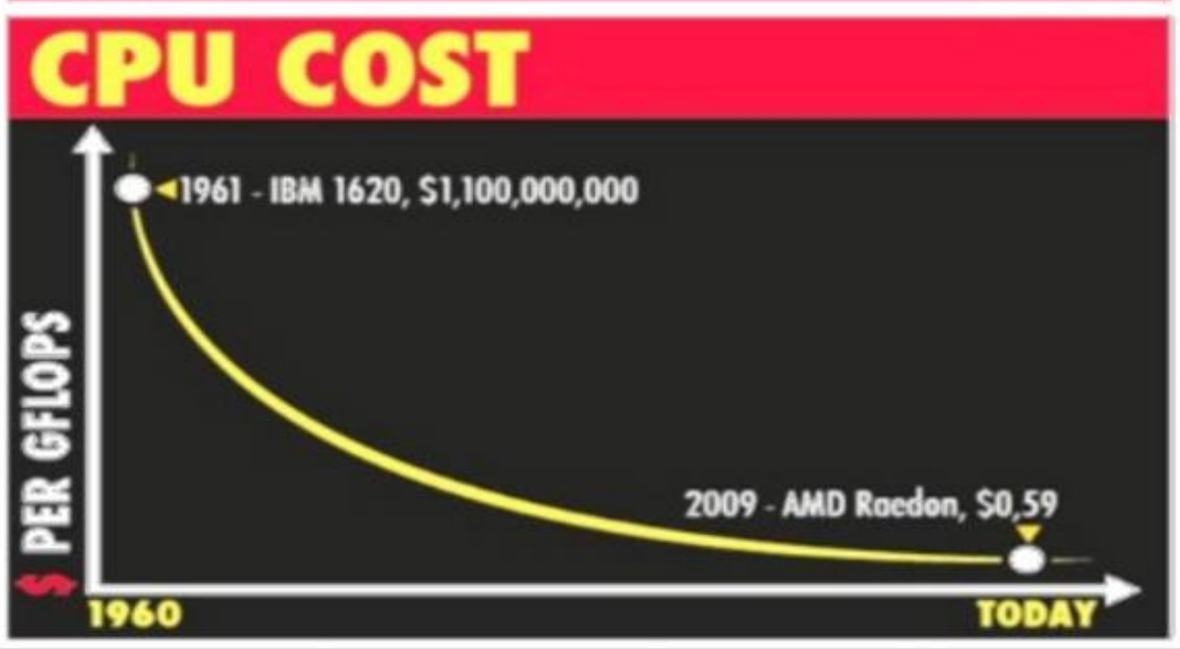
Source: Bernard Marr

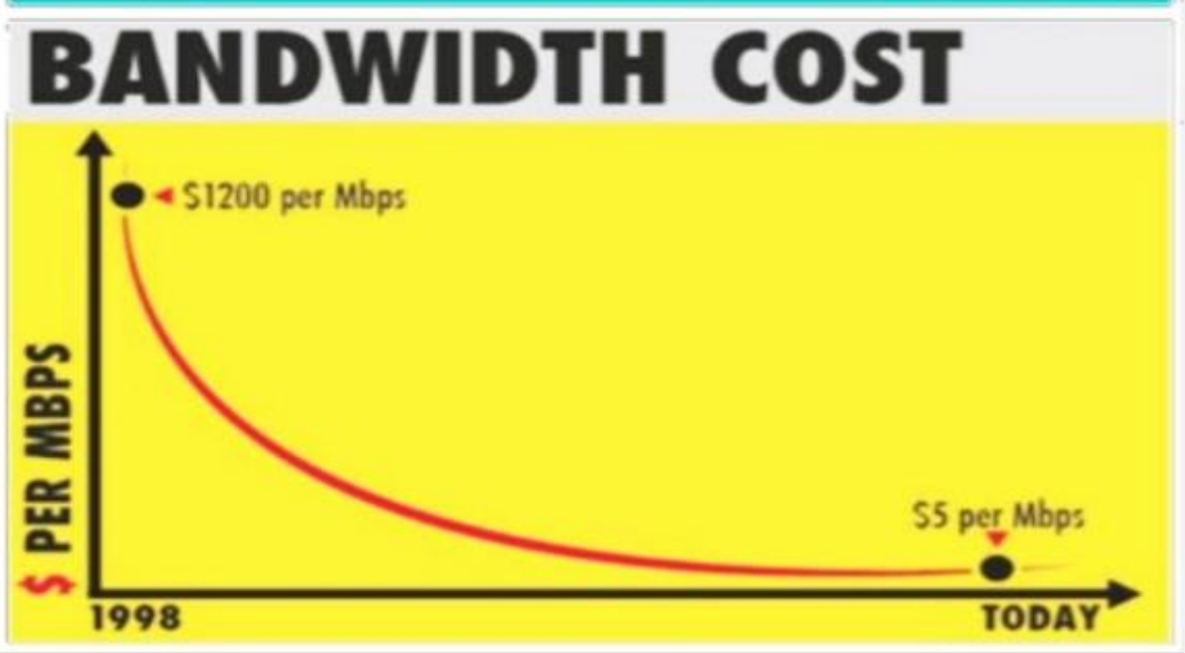
Big Data : Why Now?

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE





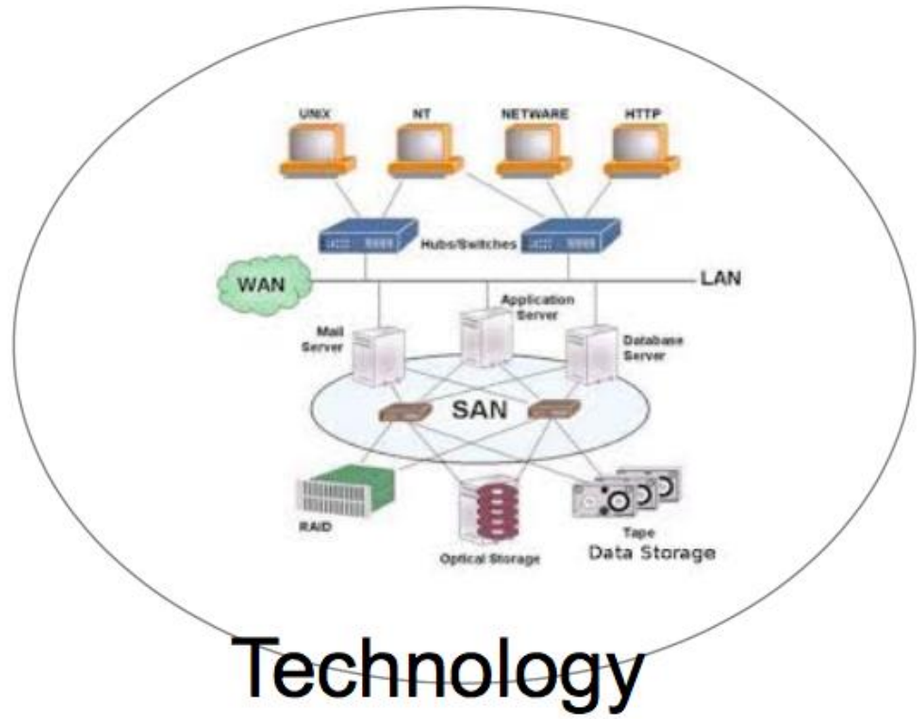




ตัวอย่างอุตสาหกรรมที่มีข้อมูลขนาดใหญ่

- โทรคมนาคม ผู้ให้บริการโทรศัพท์เคลื่อนที่ก็อาจมีข้อมูลลูกค้า หรือข้อมูลการใช้งานโทรศัพท์ของลูกค้า (CDR: Call Detail Record) ที่แต่ละวันอาจเป็นหลาย Terabyte หรือหลายหมื่นสายเรคอร์ด
- การเงินการธนาคาร มีข้อมูลการทำธรกรรมของลูกค้าที่มีจำนวนมาก หรือข้อมูลการซื้อขายหลักทรัพย์ในตลาดทุน
- ค้าปลีก ห้างสะดวกที่มีรายการซื้อสินค้า ข้อมูลสต็อกสินค้า ข้อมูลลูกค้า ที่มีการเคลื่อนไหวในแต่ละวันจำนวนมาก
- การแพทย์ ข้อมูลการรักษาพยาบาลของผู้ป่วย ข้อมูลด้านยา ข้อมูลโรคระบาดต่างๆ

Data Sources

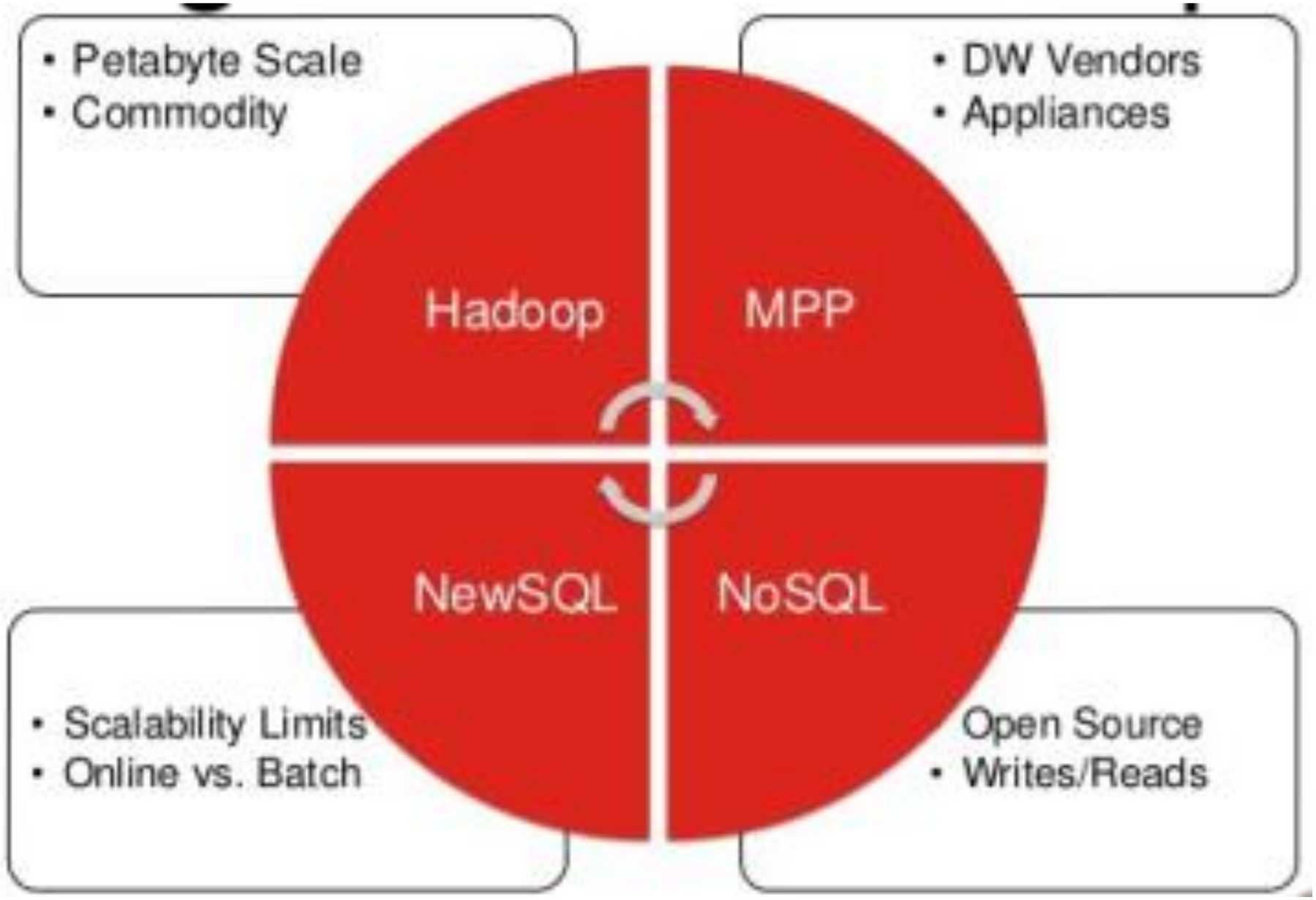


Technology

Analytics

Big Data Technology

Big Data Landscape

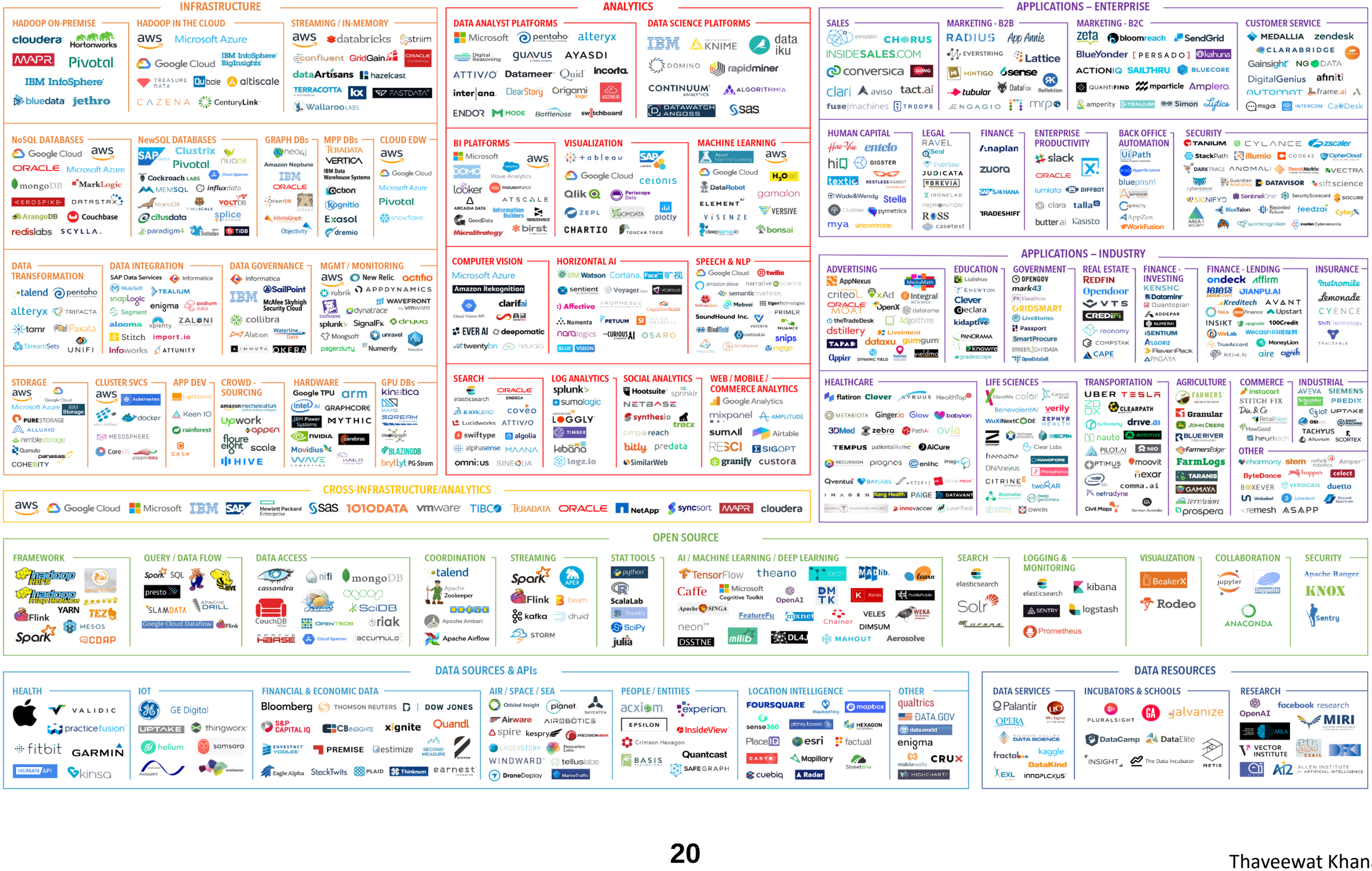


Source: Big Data in the Enterprise. When to Use What?

CHULALONGKORN BUSINESS SCHOOL

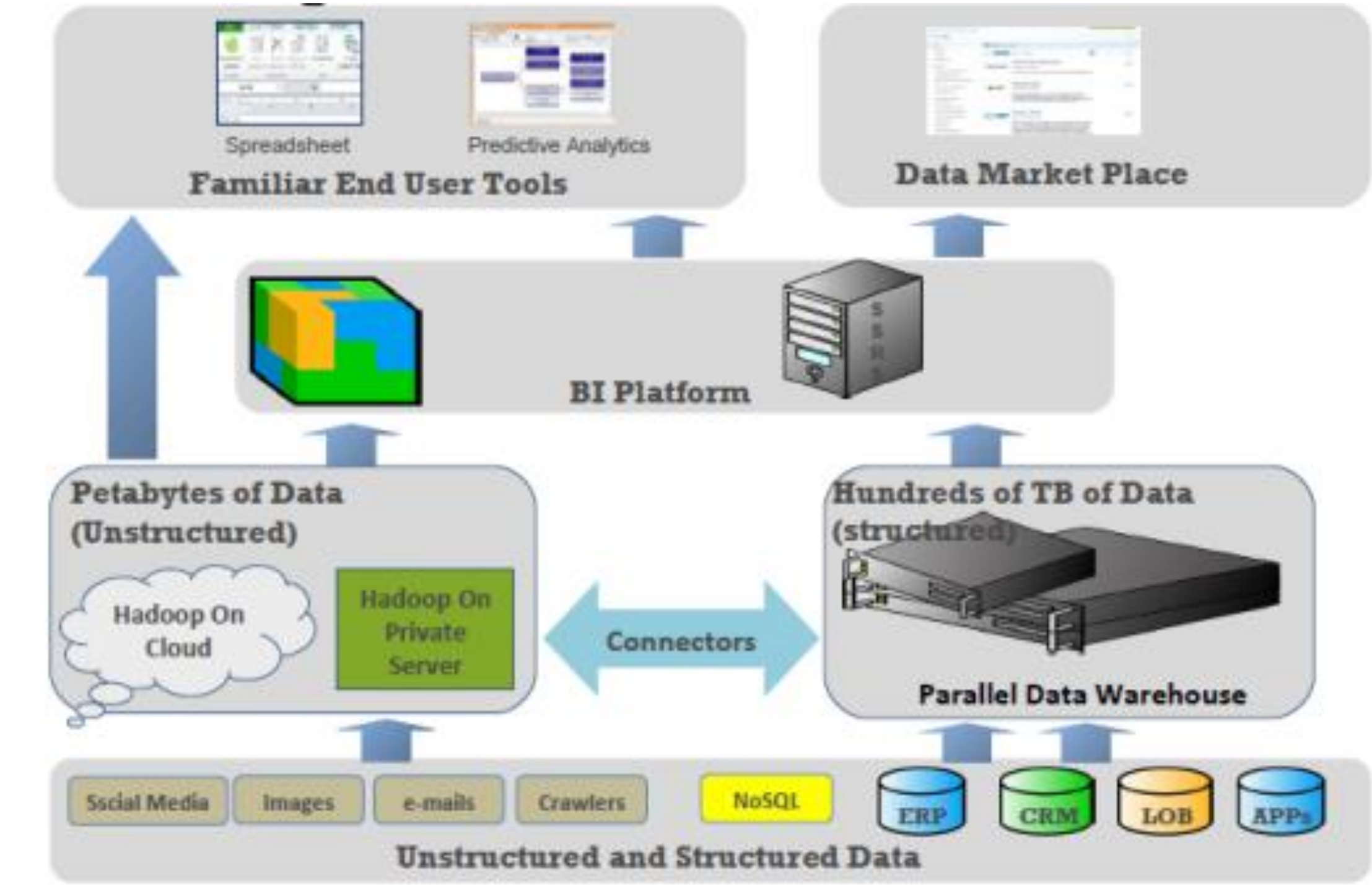
FLAGSHIP FOR LIFE

BIG DATA & AI LANDSCAPE 2018



Thaveewat Khanan

Big Data Future Architecture

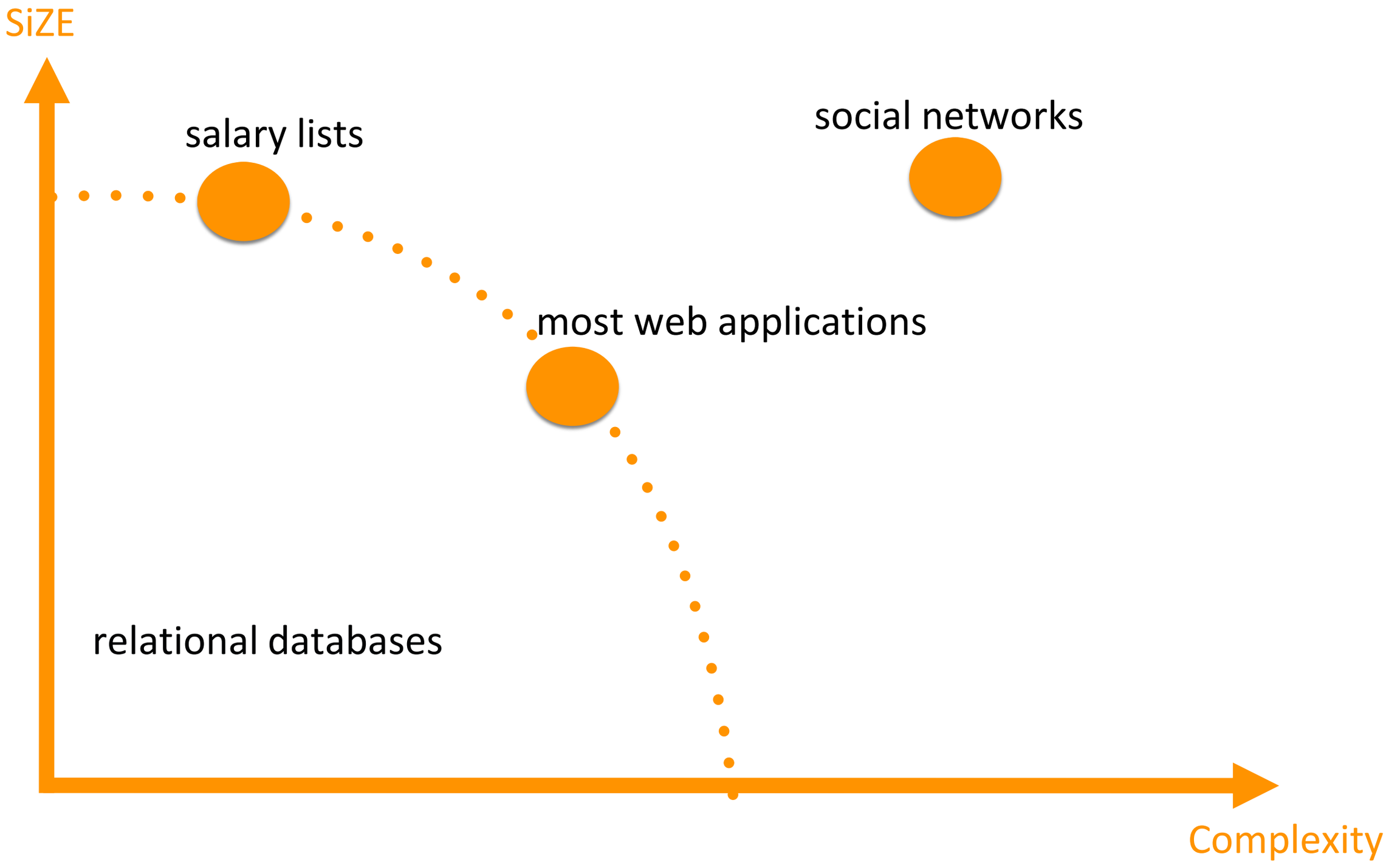


What is NoSQL ?

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

Key Value Store

Column-oriented

Document Store

Graph

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”}

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

Document-oriented database

Designed for storing, retrieving, and managing documentoriented

information, also known as semi-structured data.

Most of the databases available u

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

Comment Table
Reader Table
Article Table
Author Table
Author name

Relational Database approach

Document store approach

CouchDB

```
{
  "articles": [
    {
      "title": "title of the article",
      "articleID": 1,
      "body": "body of the article",
      "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

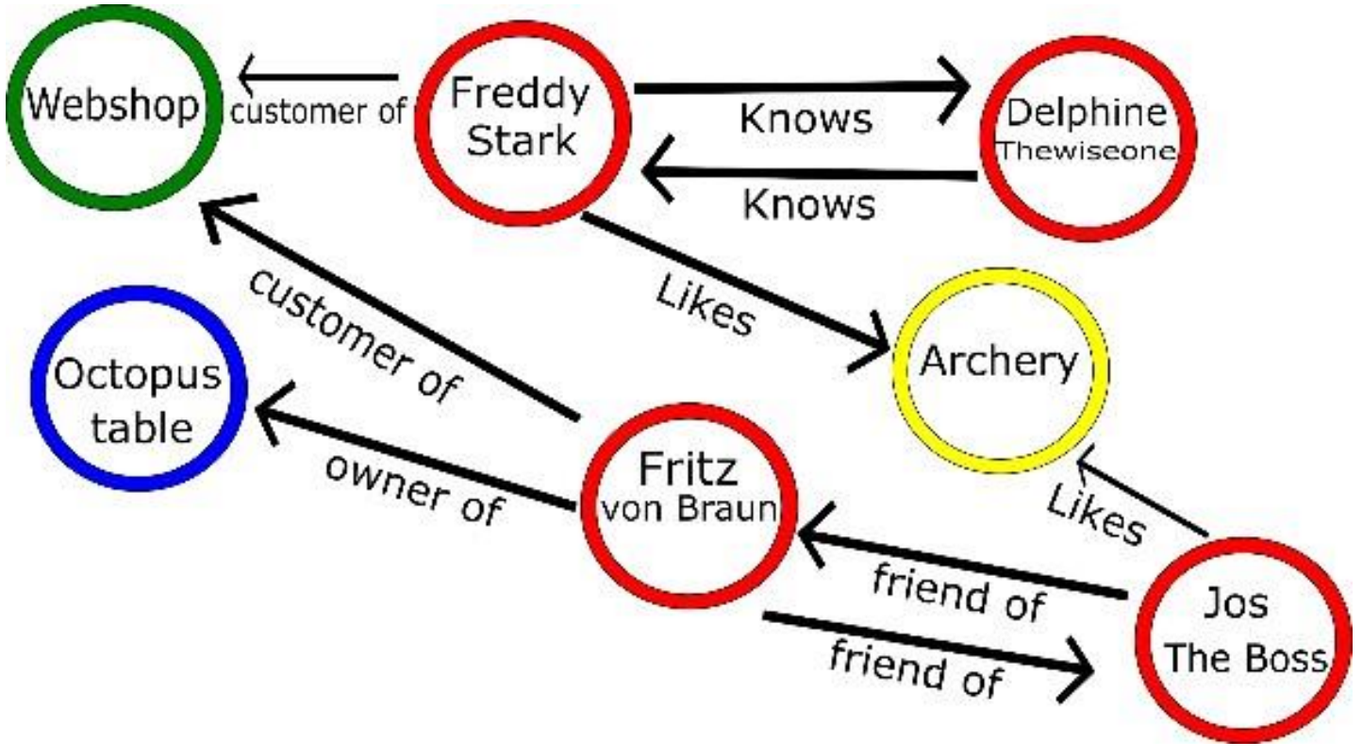
CouchDB relax

mongoDB

Graph database

A database that uses graph structures for semantic queries with

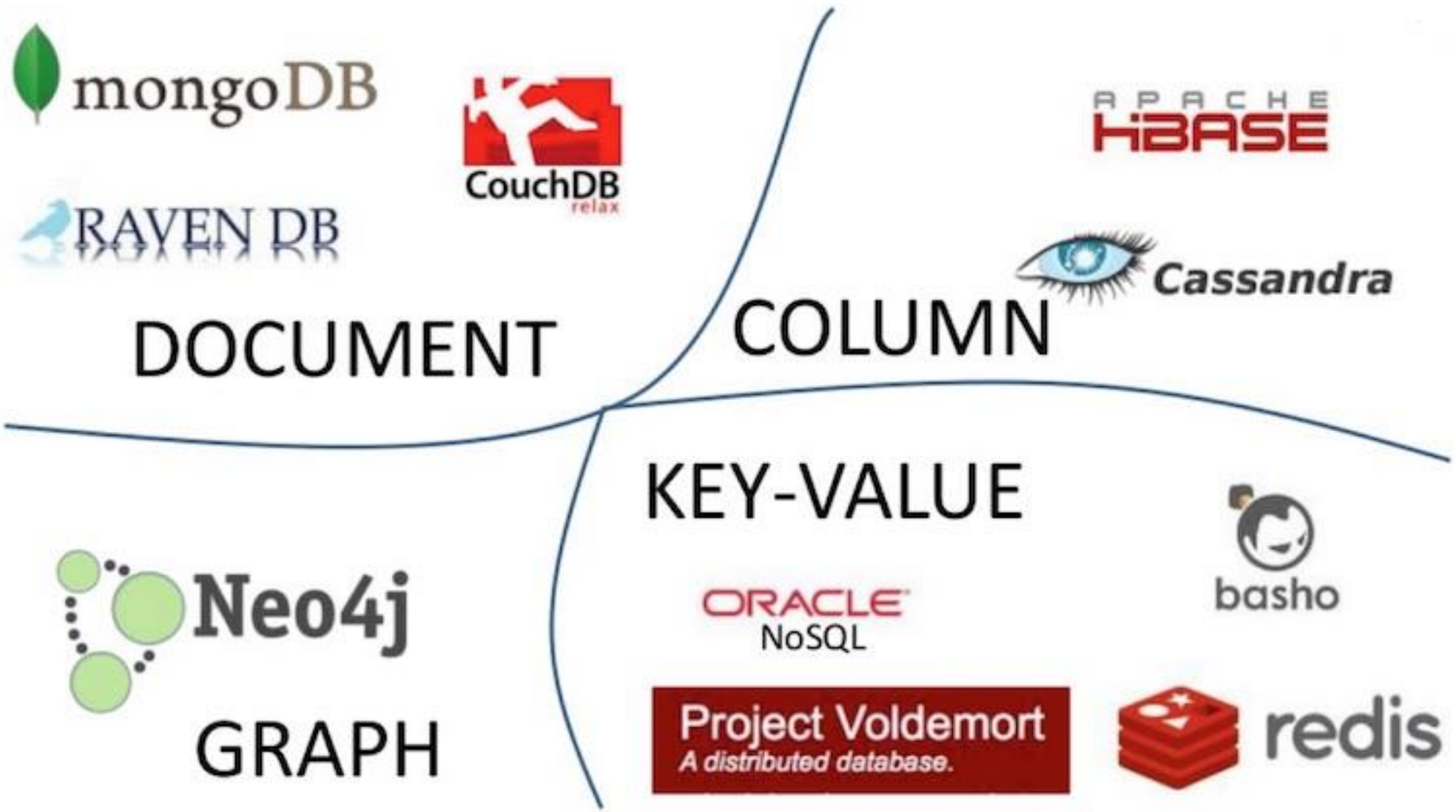
nodes, edges, and properties to represent and store data.

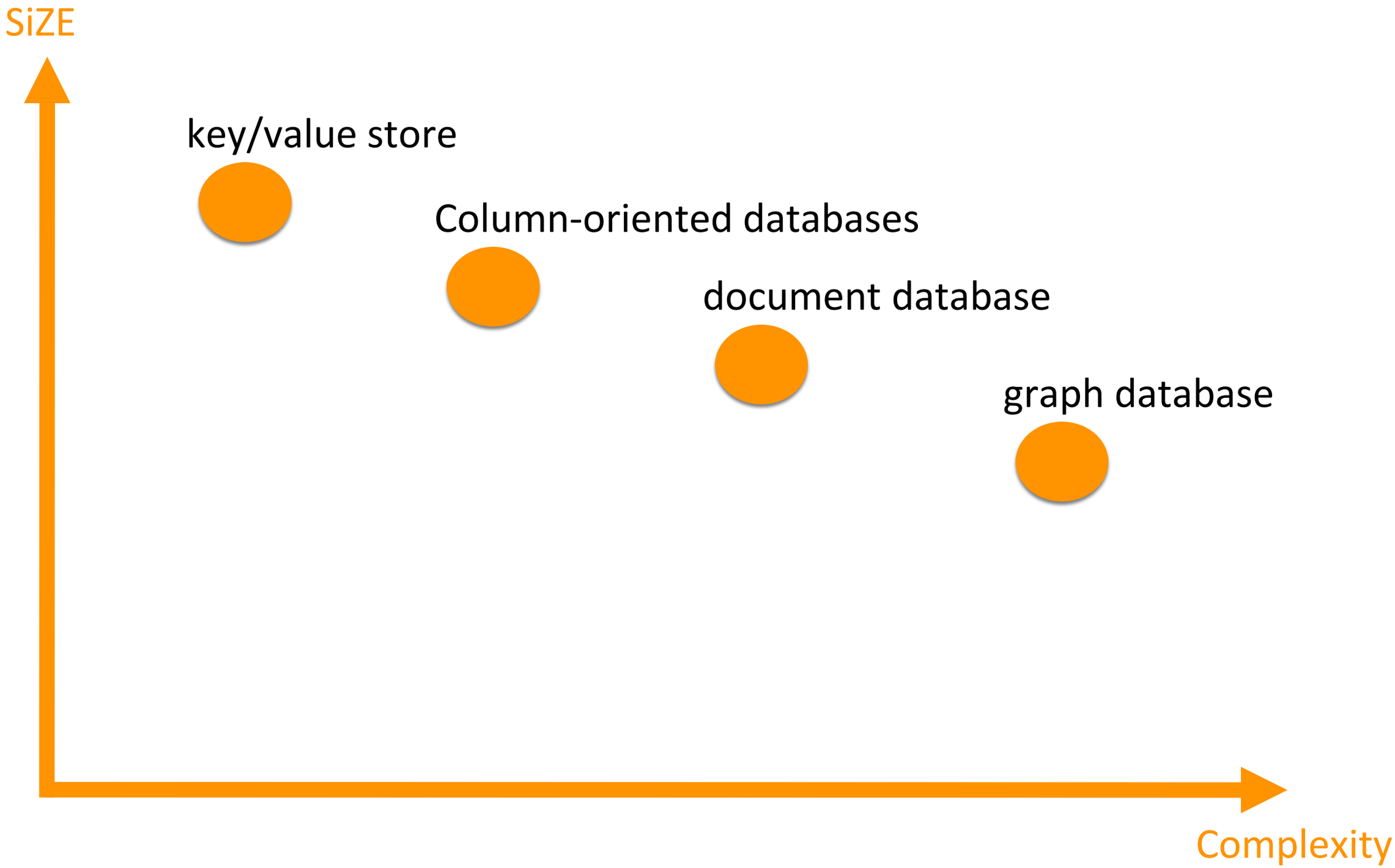


Neo4j
the graph database

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

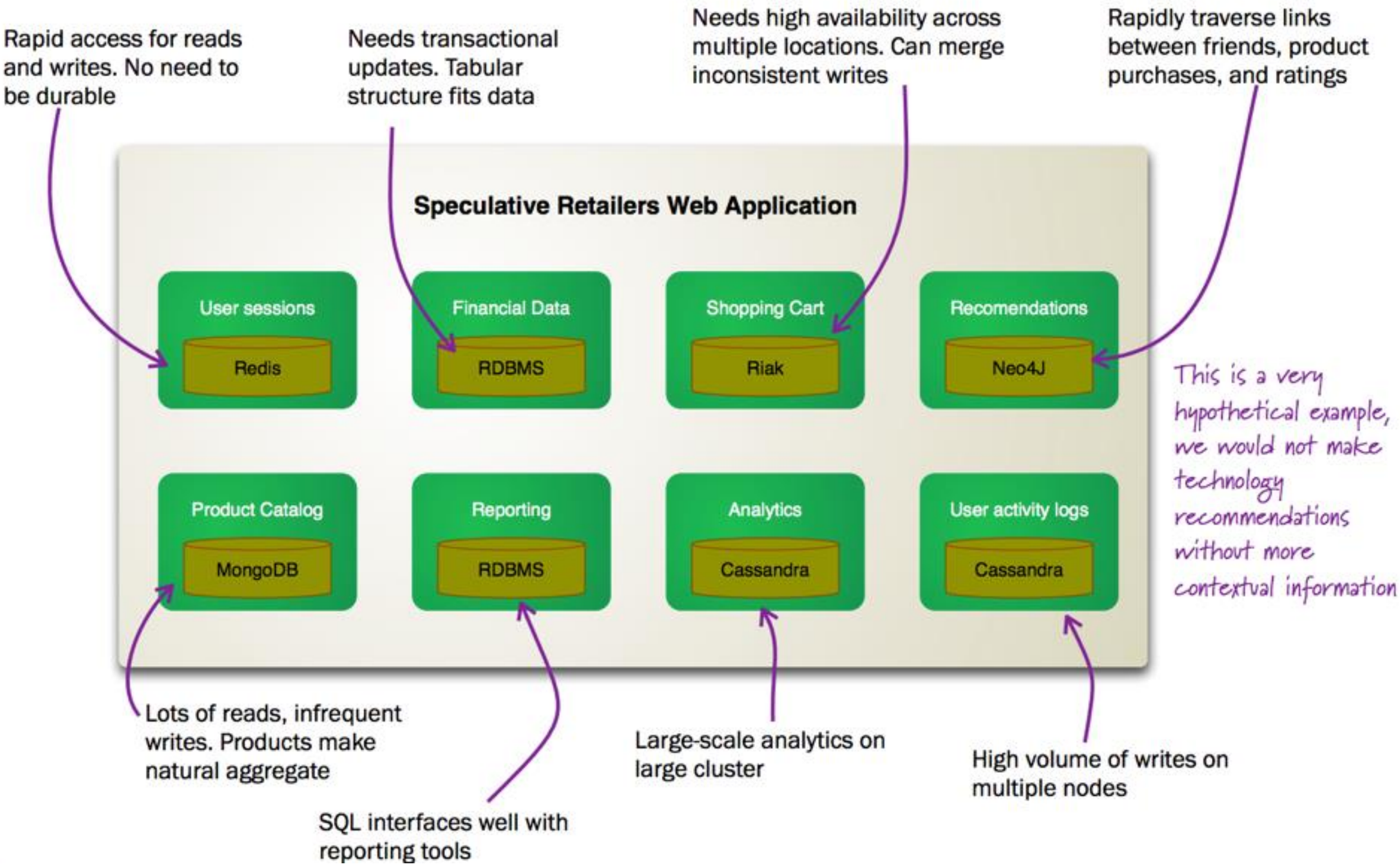
SQL

works great, can’t scale for large data

NoSQL

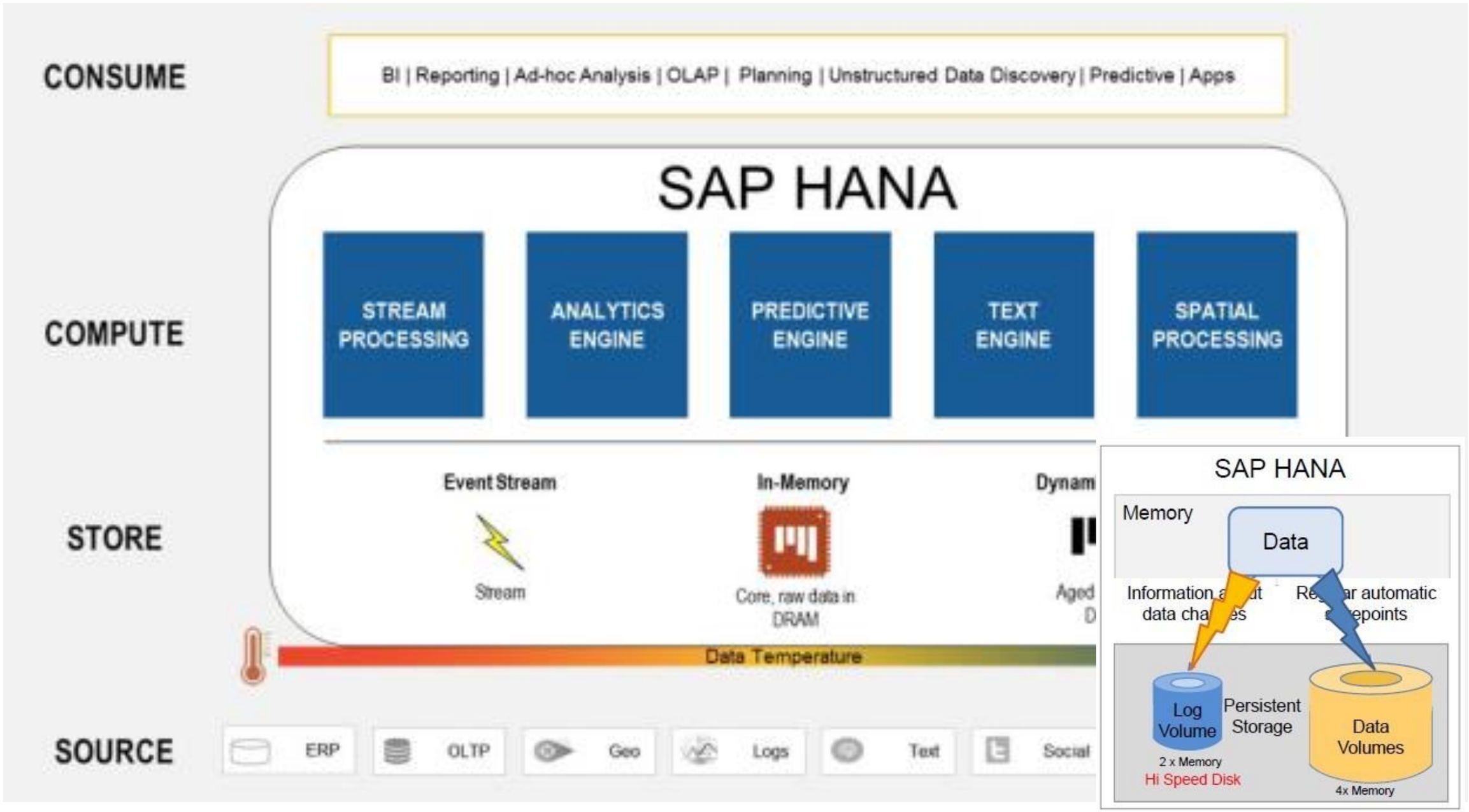
works great, doesn't fit all situations

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

YAHOO!
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

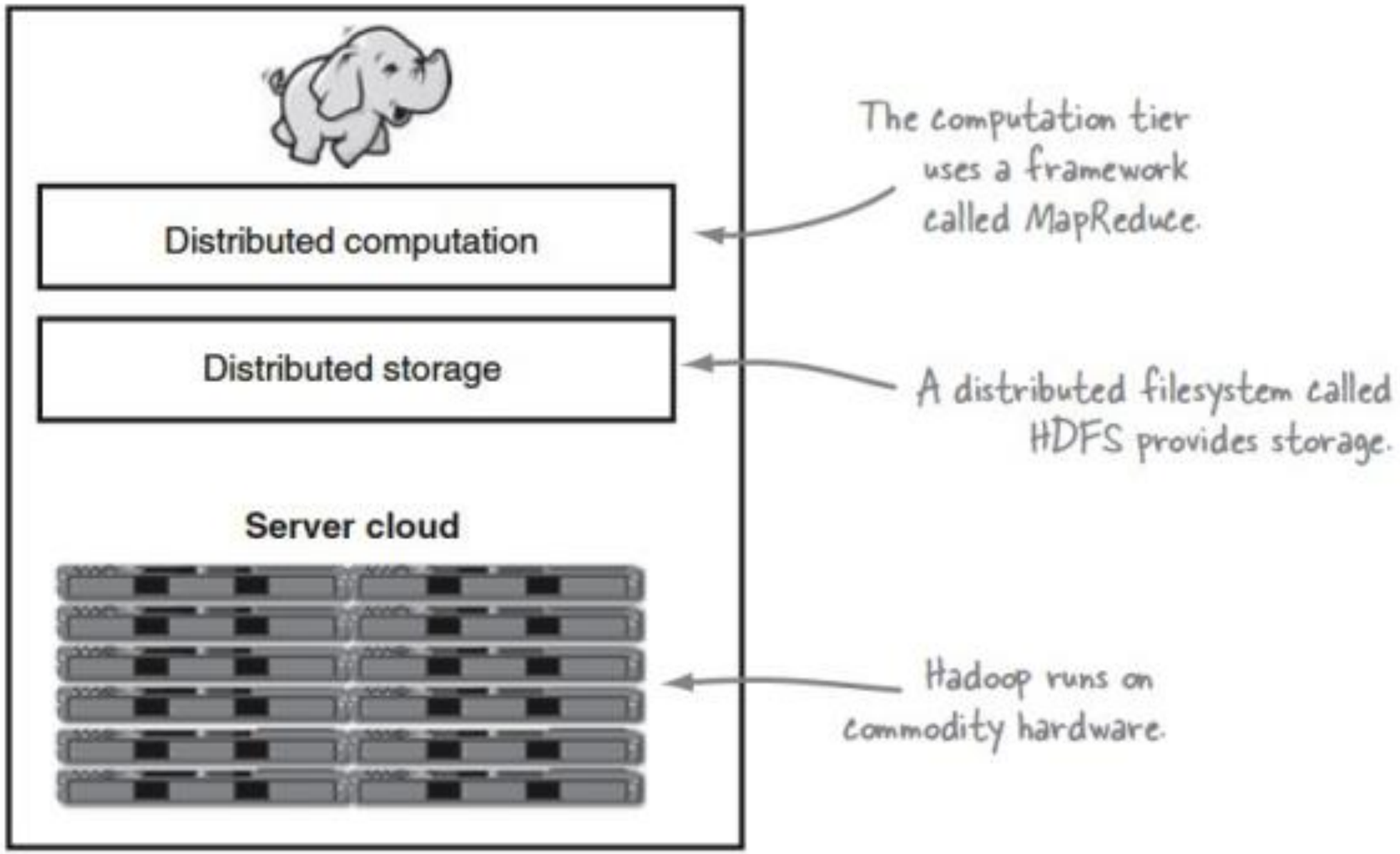
cloudera
Founded

Doug Cutting joins Cloudera

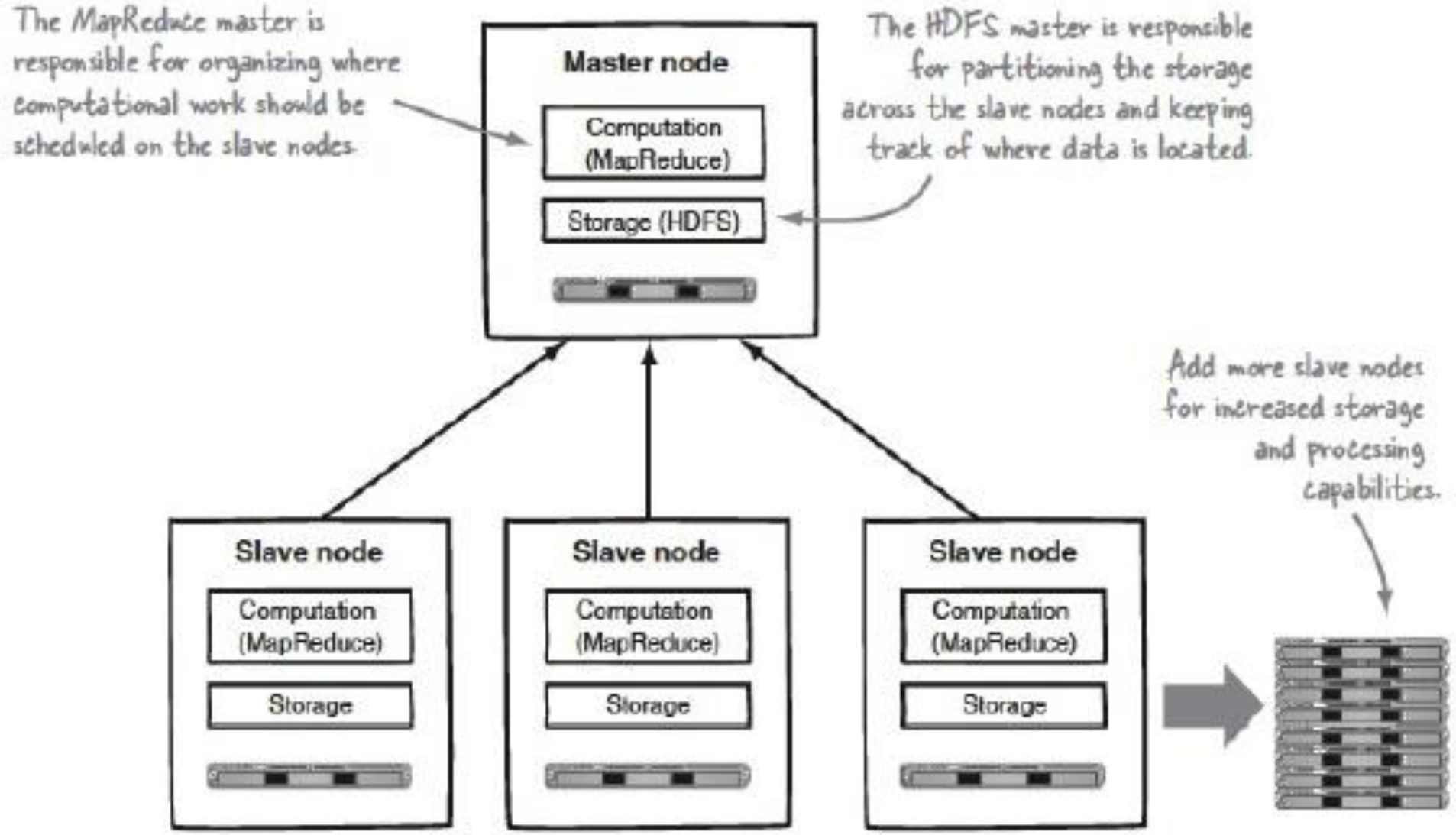
Hadoop Summit 2009, 750 attendees

Source: Cloudera, Inc.

Hadoop Environment



Hadoop Architecture



Major Hadoop Components

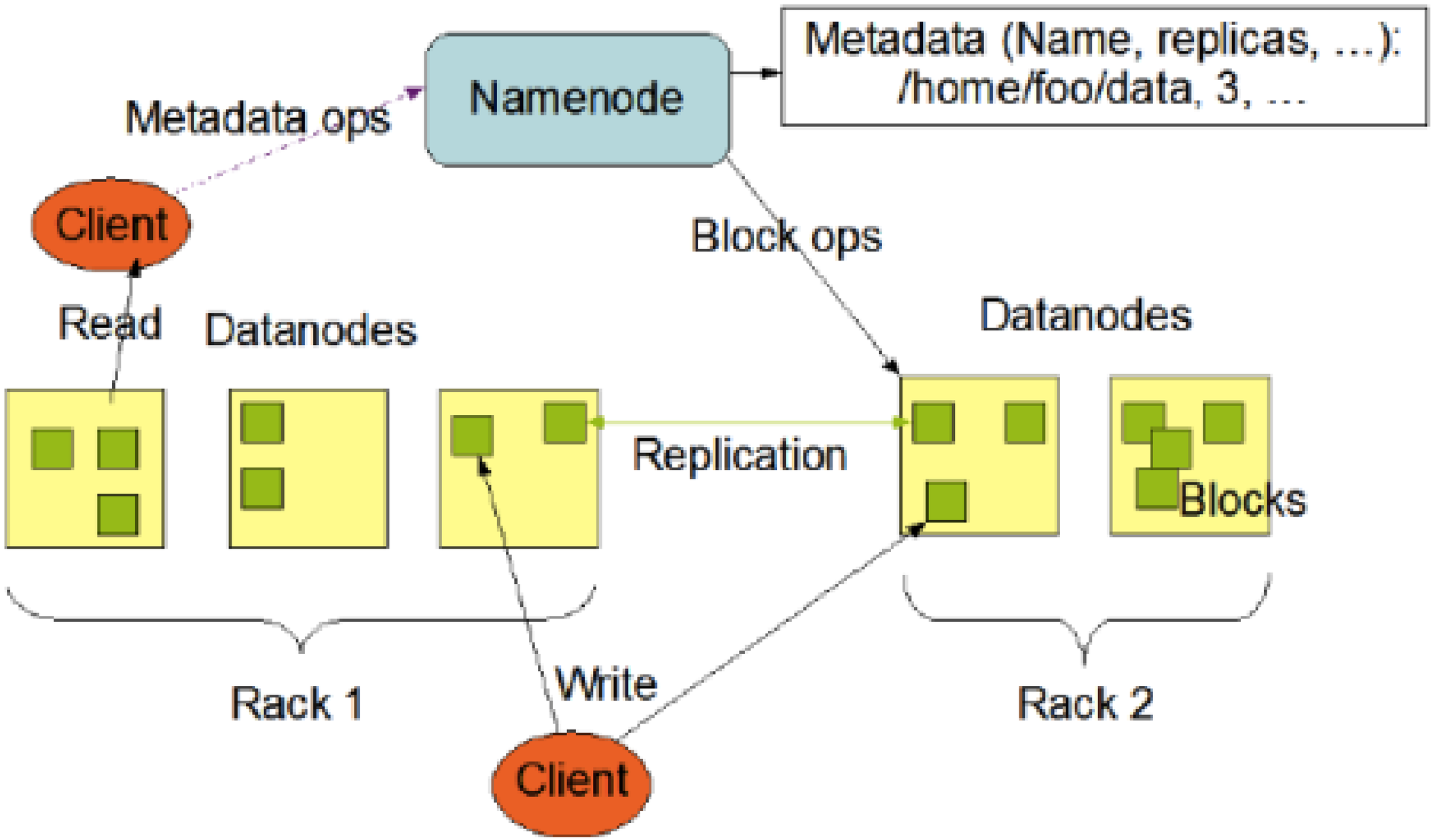
Hadoop Distributed File System (HDFS)

Map/Reduce System

HDFS

- Default storage for the Hadoop cluster
- Data is distributed and replicated over multiple machines
- Designed to handle very large files with streaming data access patterns.
- NameNode/DataNode
- Master/slave architecture (1 master 'n' slaves)
- Designed for large files (64 MB default, but configurable) across all the nodes

HDFS Architecture



Source Hadoop: Shashwat Shriparv

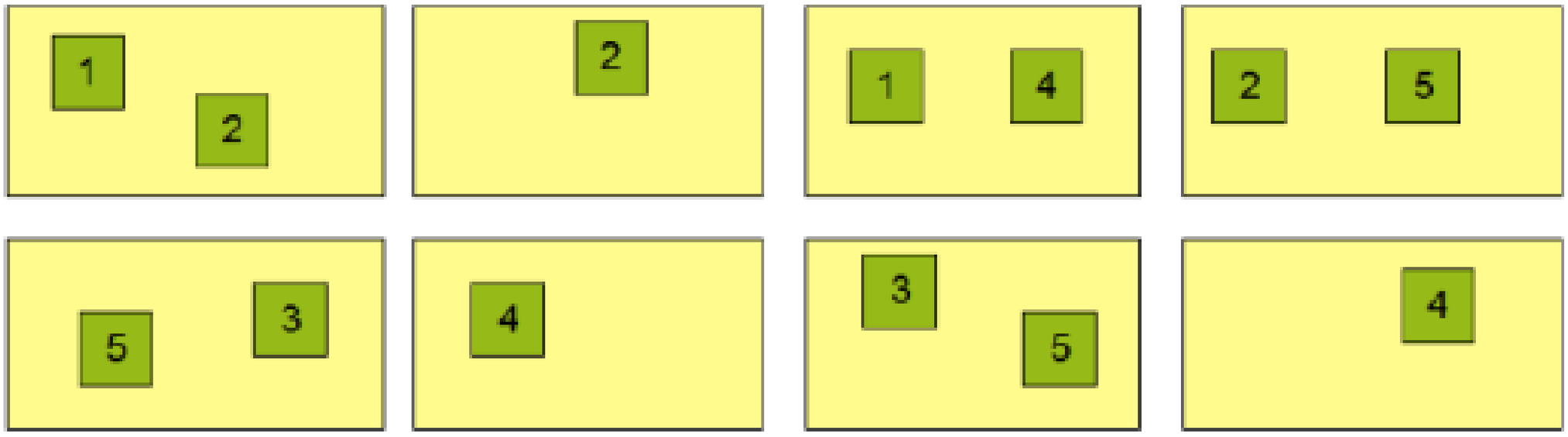
Data Replication in HDFS

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

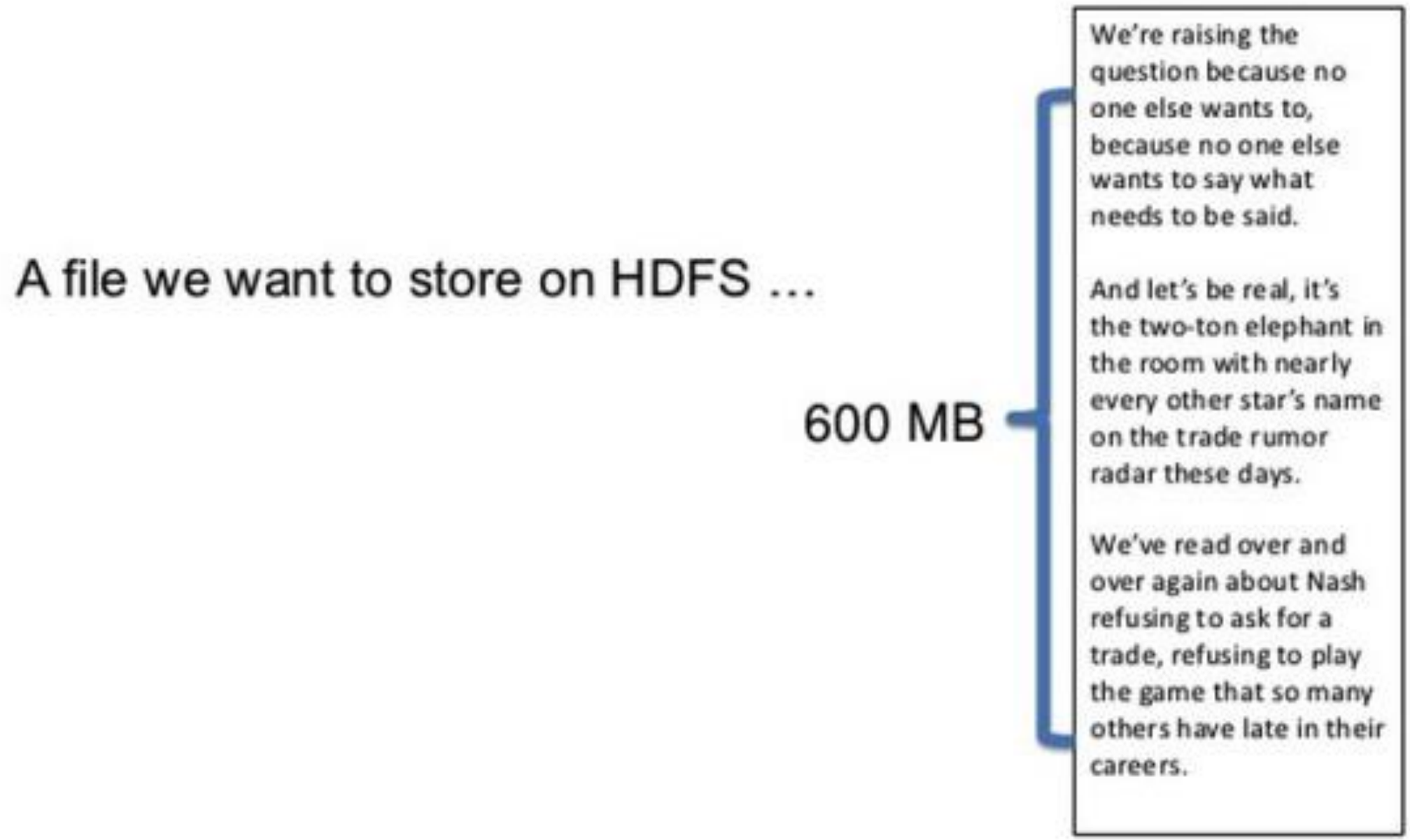
Block Replication

Namenode (Filename, numReplicas, block-ids, …)
/users/sameerp/data/part-0, r:2, {1,3}, …
/users/sameerp/data/part-1, r:3, {2,4,5}, …

Datanodes



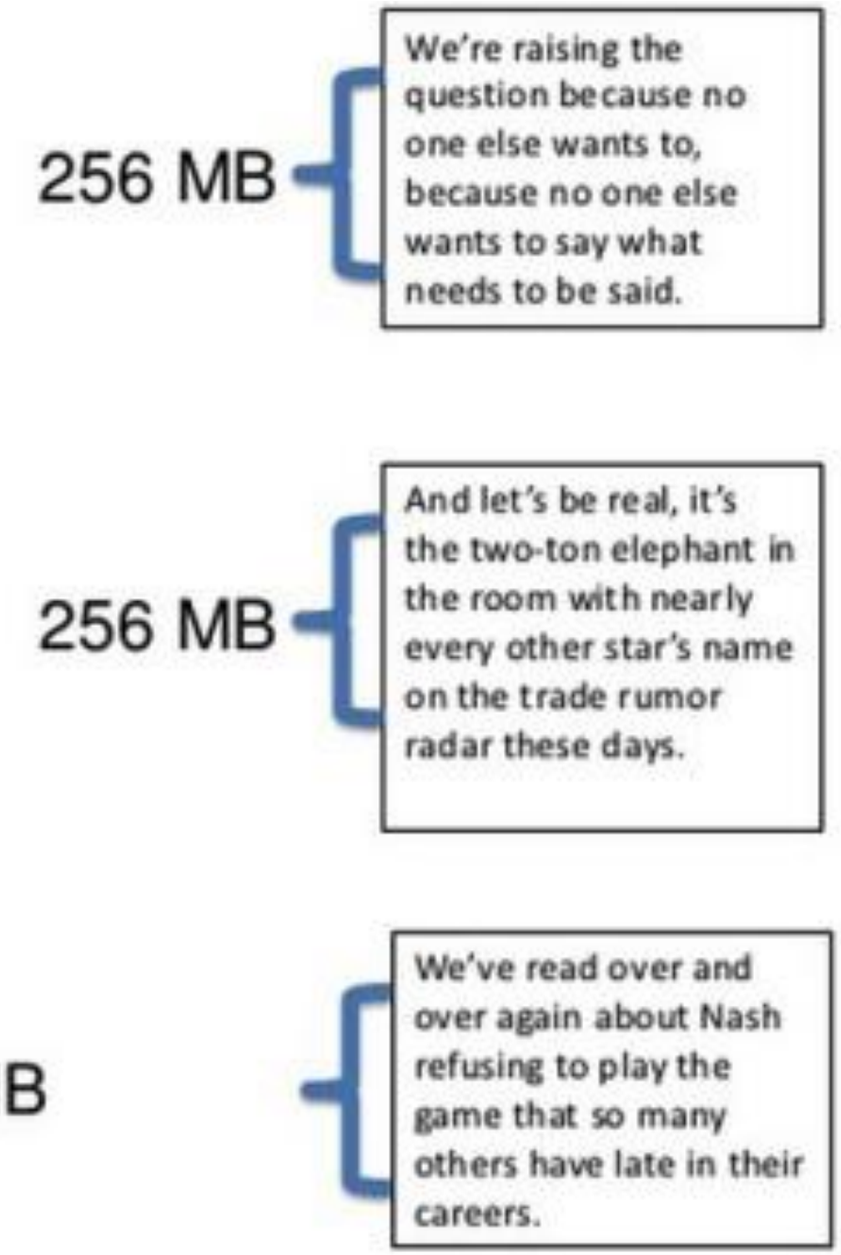
How does HDFS work?



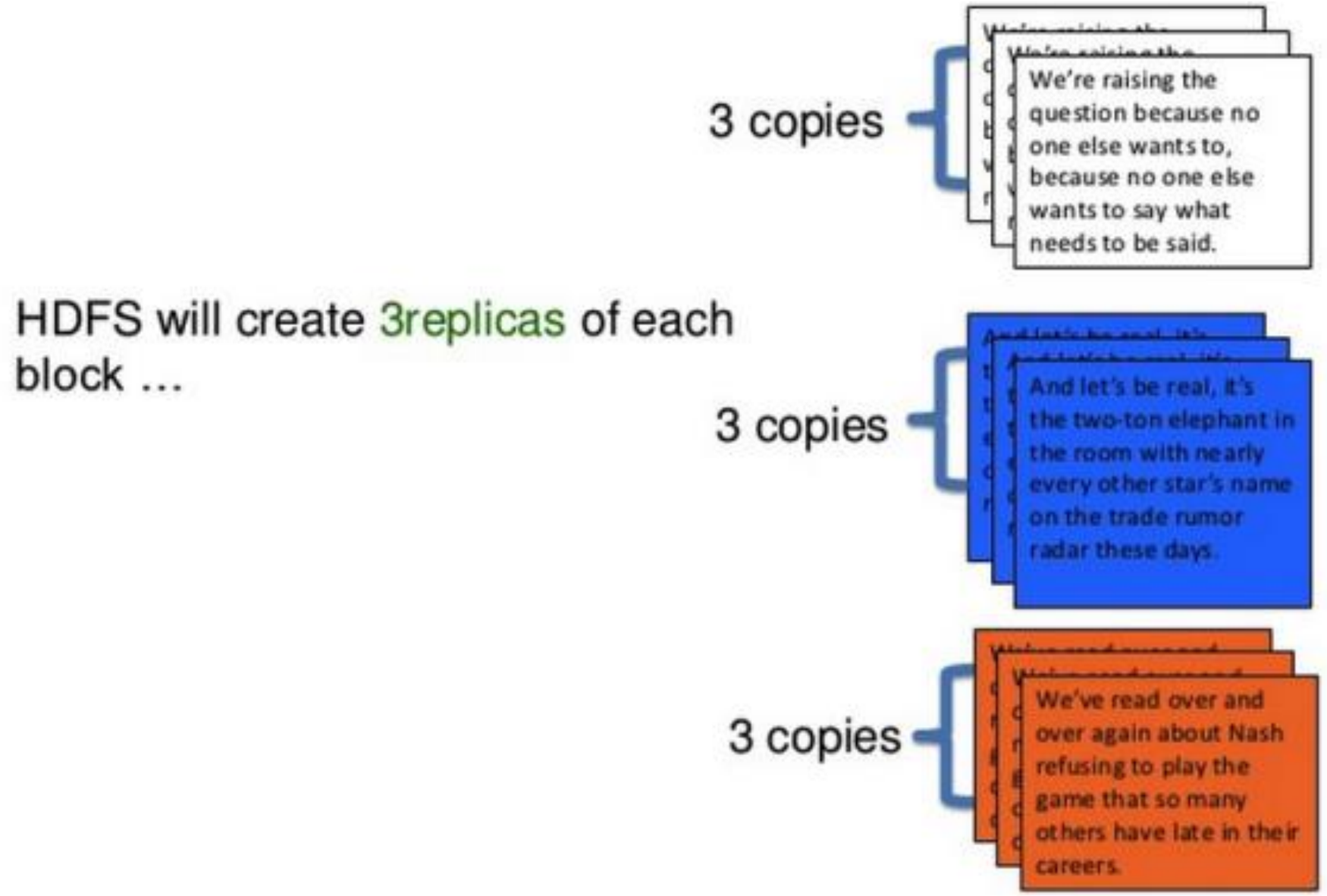
Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

How does HDFS work?

HDFS Splits file into blocks …

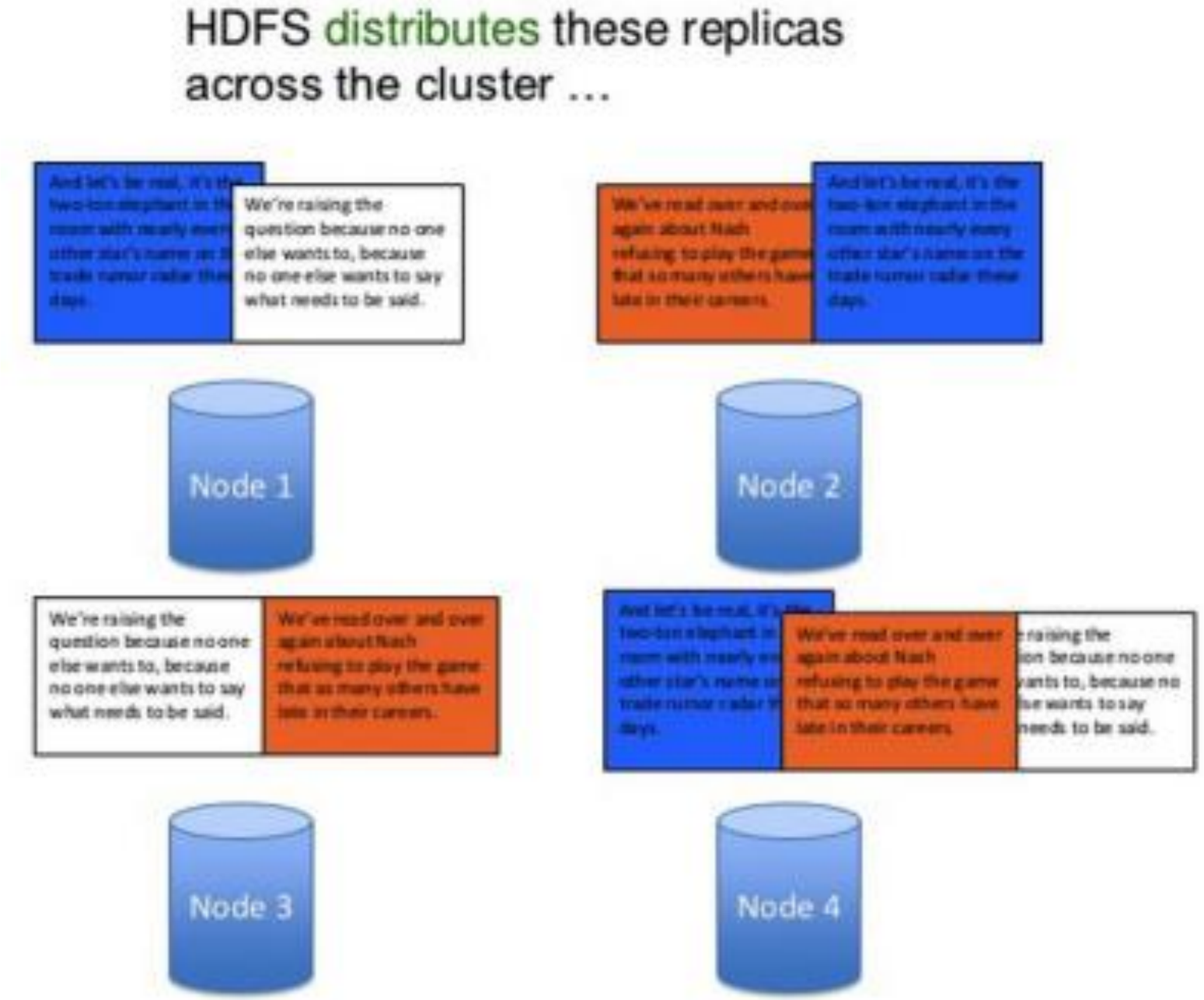


How does HDFS work?



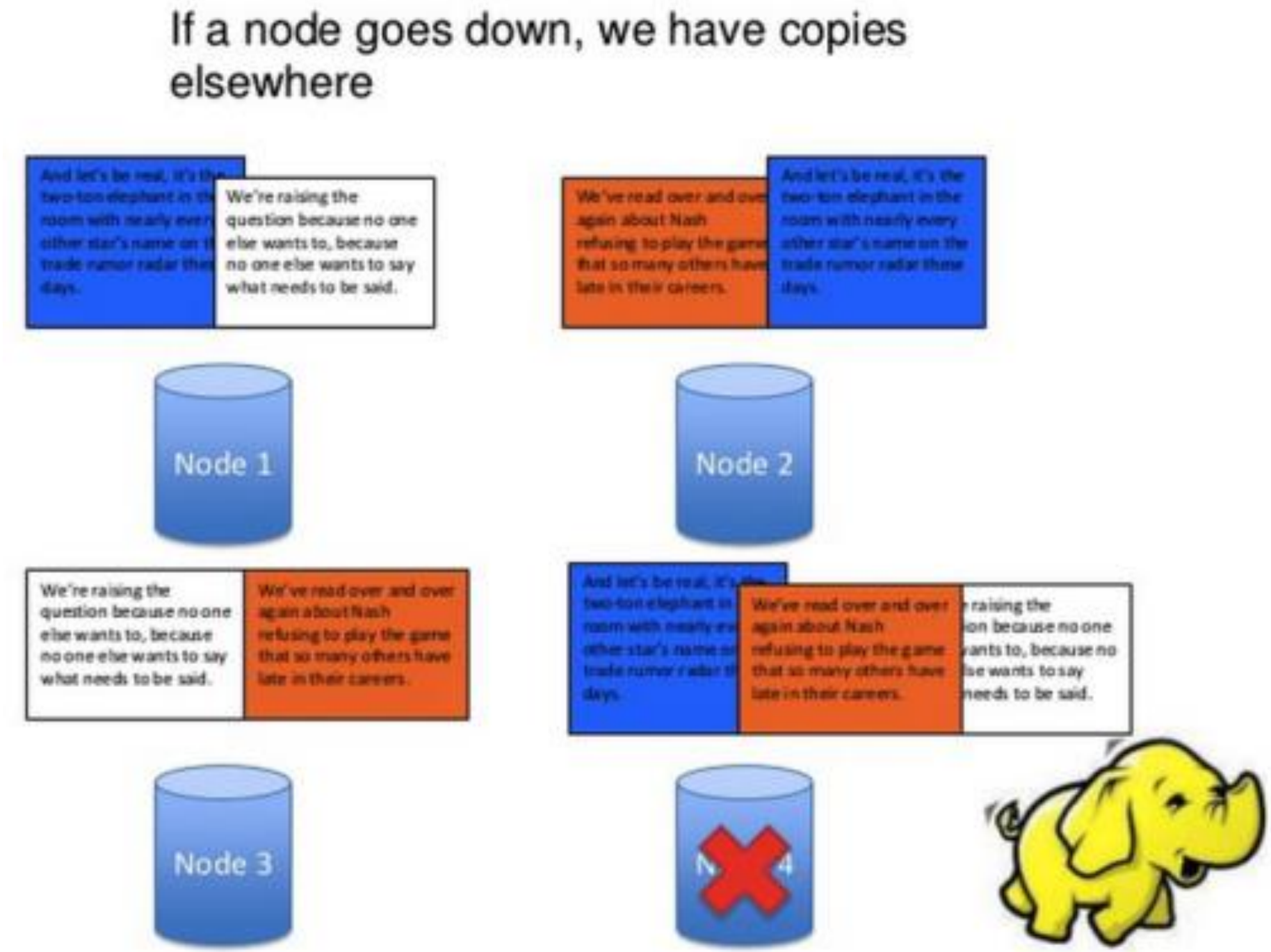
Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

How does HDFS work?



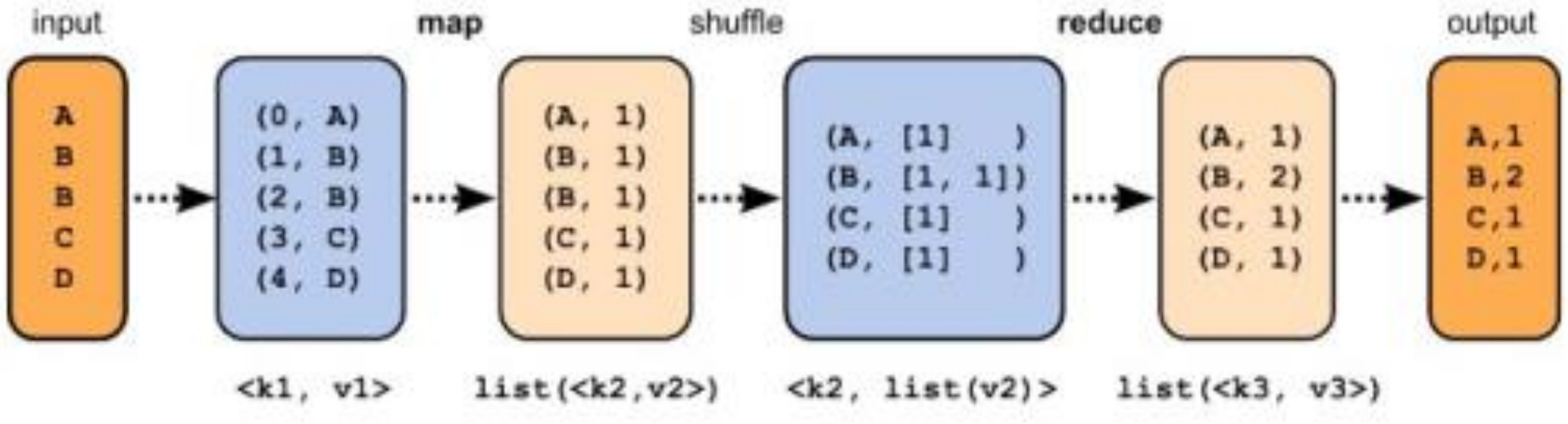
Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

How does HDFS work?



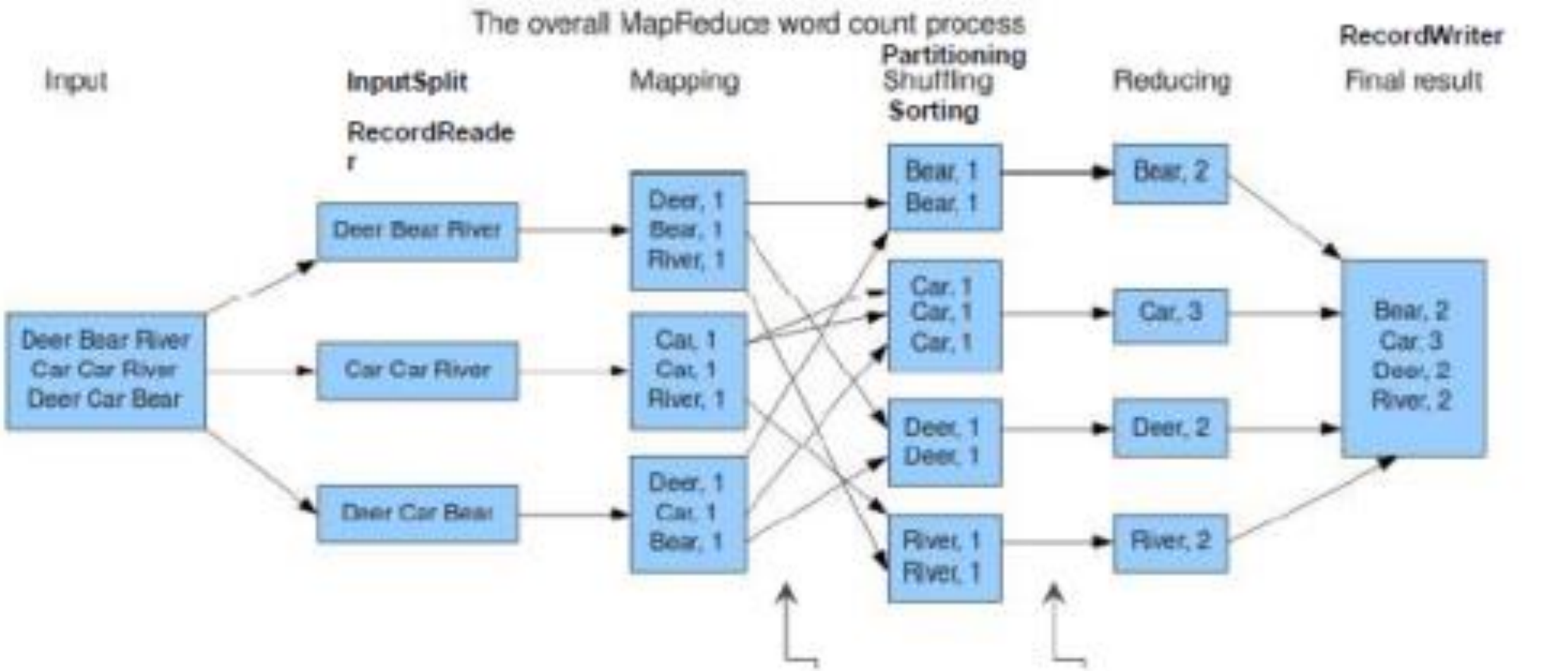
Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

MapReduce Framework

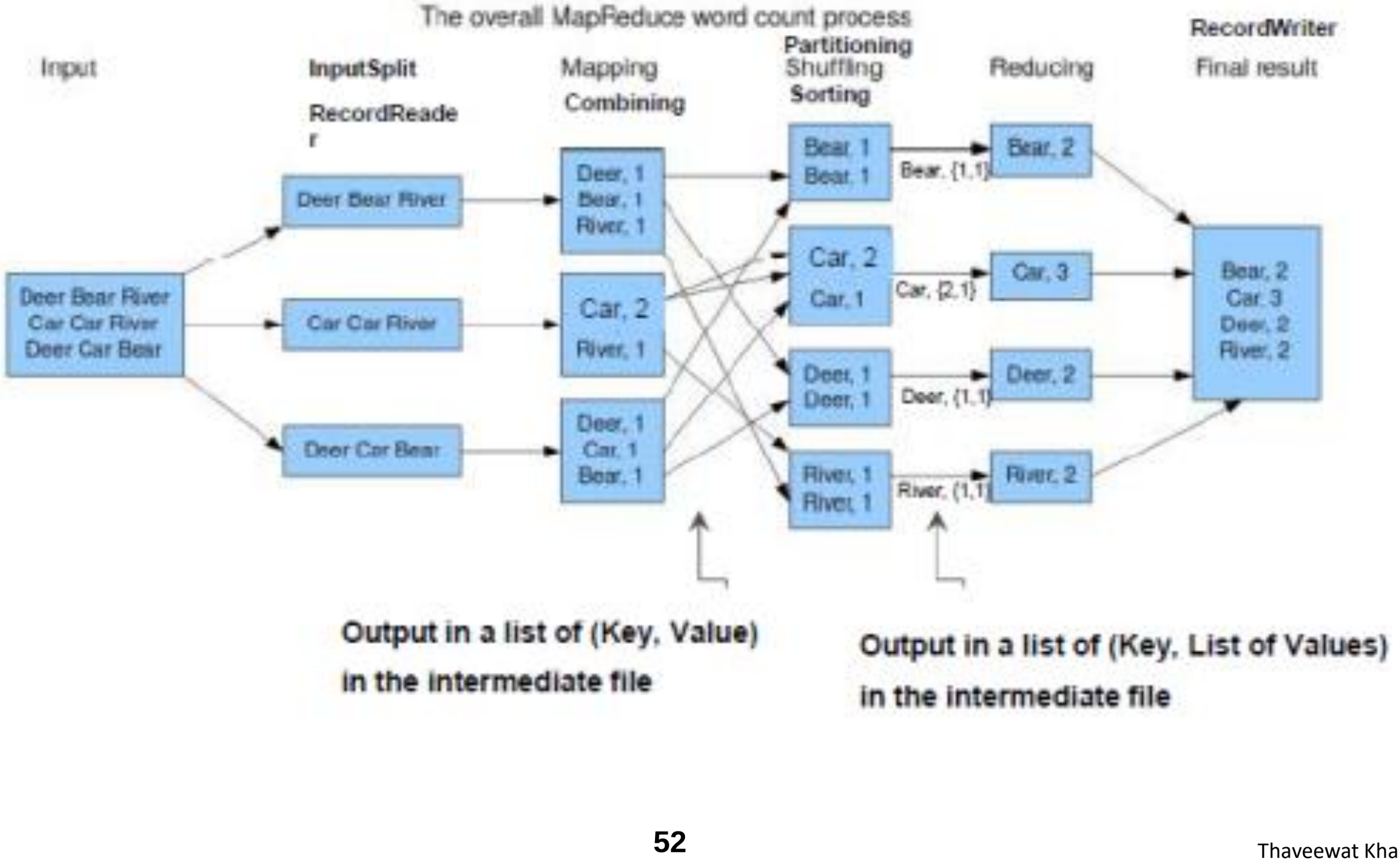


Source: www.bigdatauniversity.com

How does the MapReduce work ?



How does the MapReduce work ?



WordCount - MapReduce

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

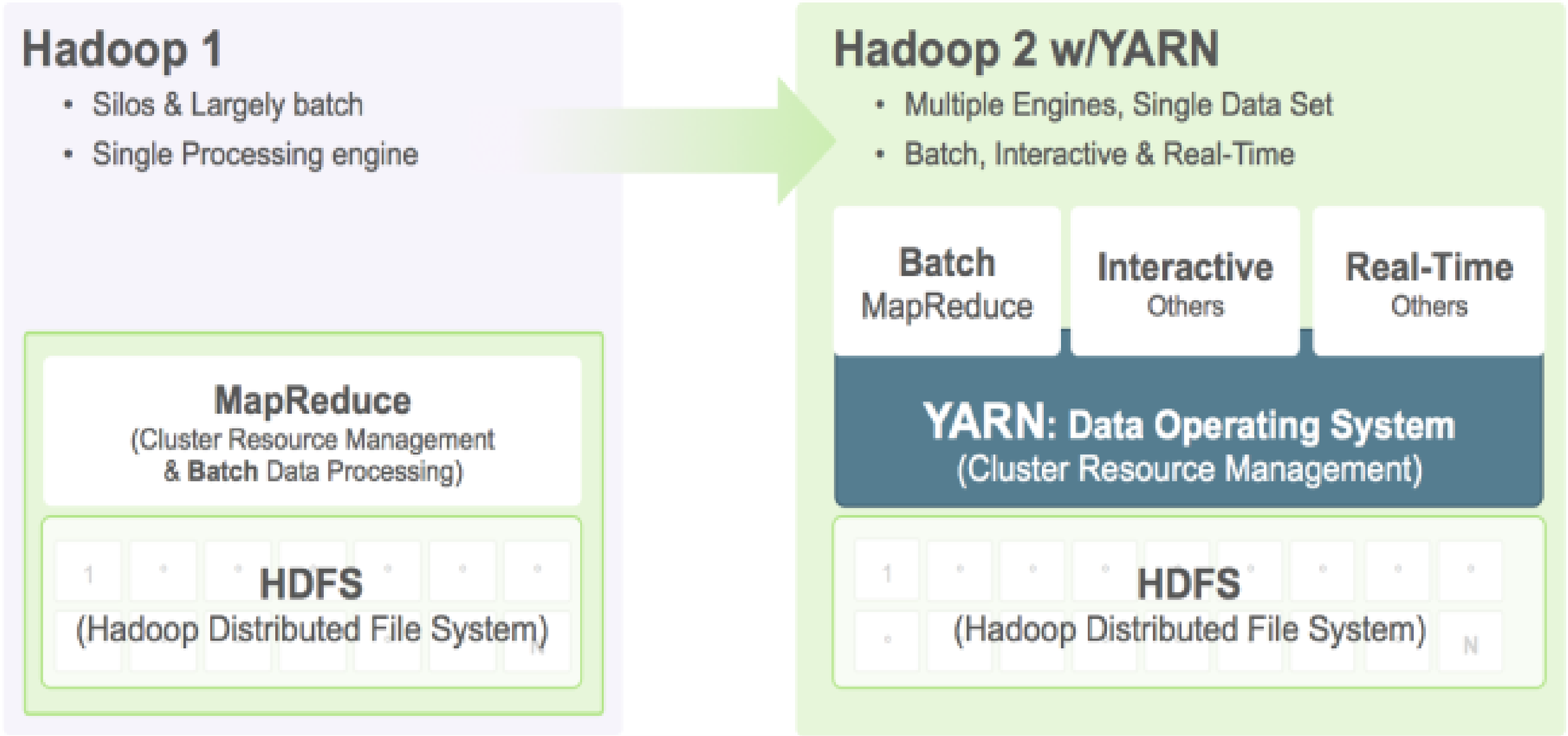
Map Function

```
public class WordCount {

  public static class Map extends MapReduceBase implements
                 Mapper<LongWritable, Text, Text, IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable>
                    output, Reporter reporter) throws IOException {
      String line = value.toString();
      StringTokenizer tokenizer = new StringTokenizer(line);
      while (tokenizer.hasMoreTokens()) {
        word.set(tokenizer.nextToken());
        output.collect(word, one);
}}}
```

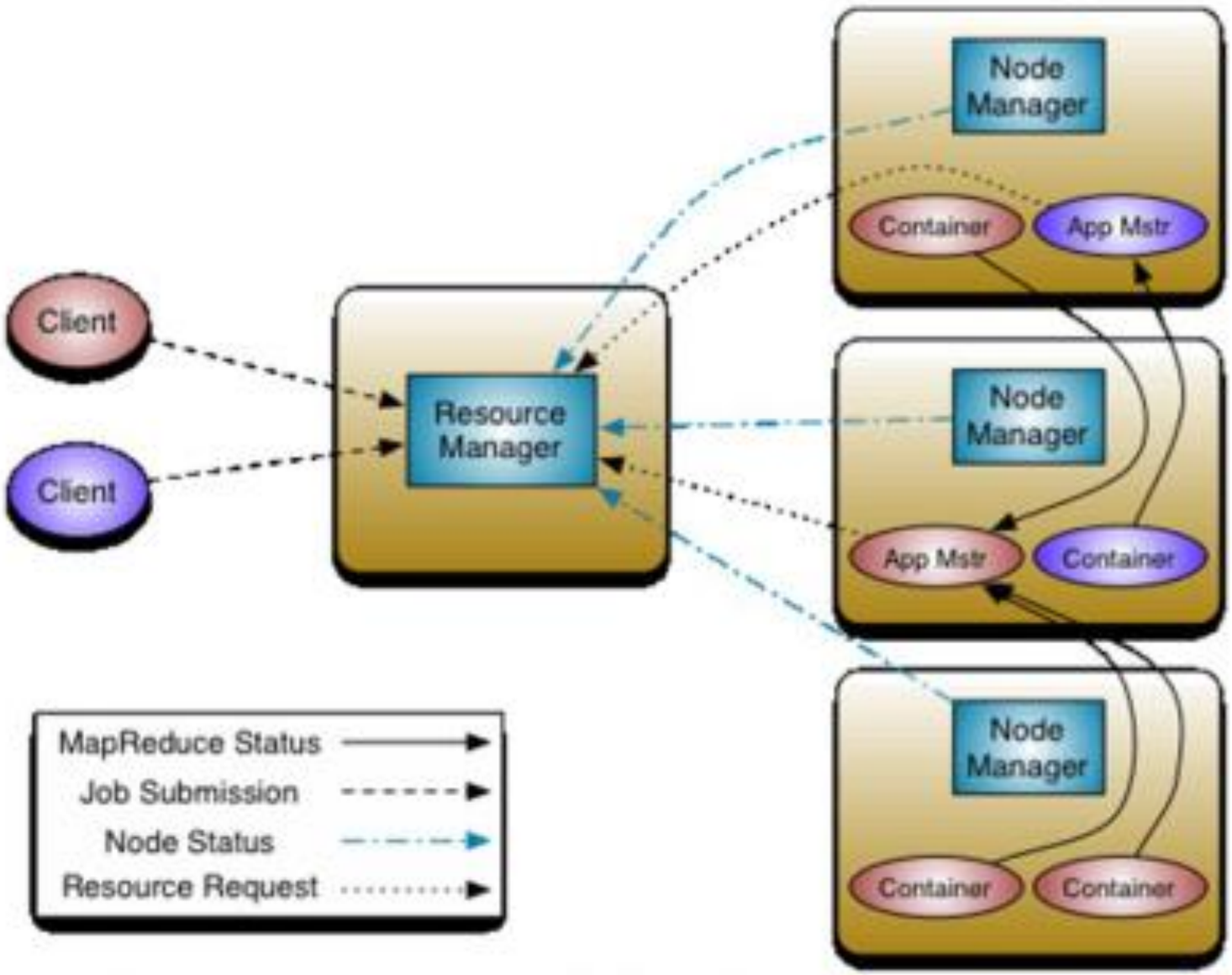
Reduce Function

```
public static class Reduce extends MapReduceBase implements
                Reducer<Text, IntWritable, Text, IntWritable> {
  public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text,
                     IntWritable> output, Reporter reporter) throws IOException {
    int sum = 0;
    while (values.hasNext()) { sum += values.next().get(); }
    output.collect(key, new IntWritable(sum));
}}
```

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.

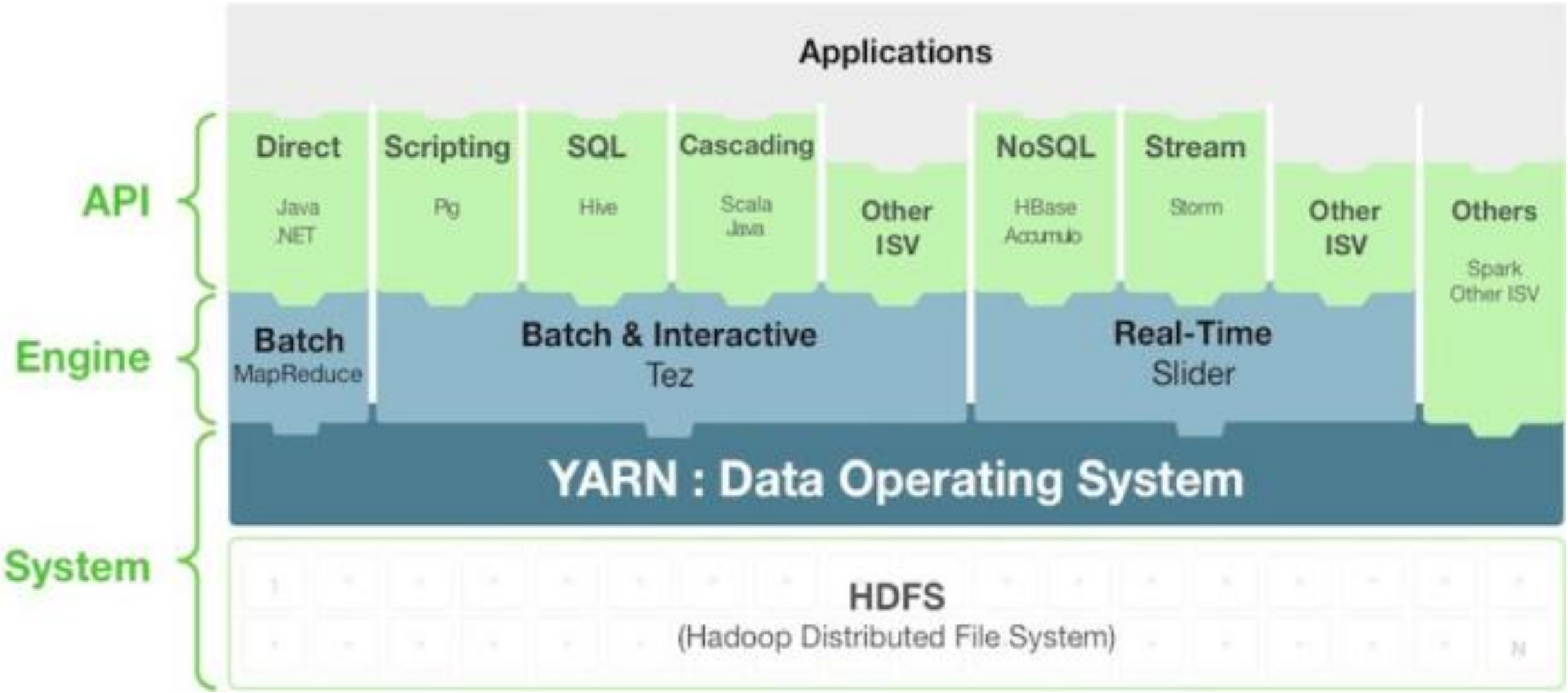
Hadoop.apache.org

Evolution of the Hadoop Platform

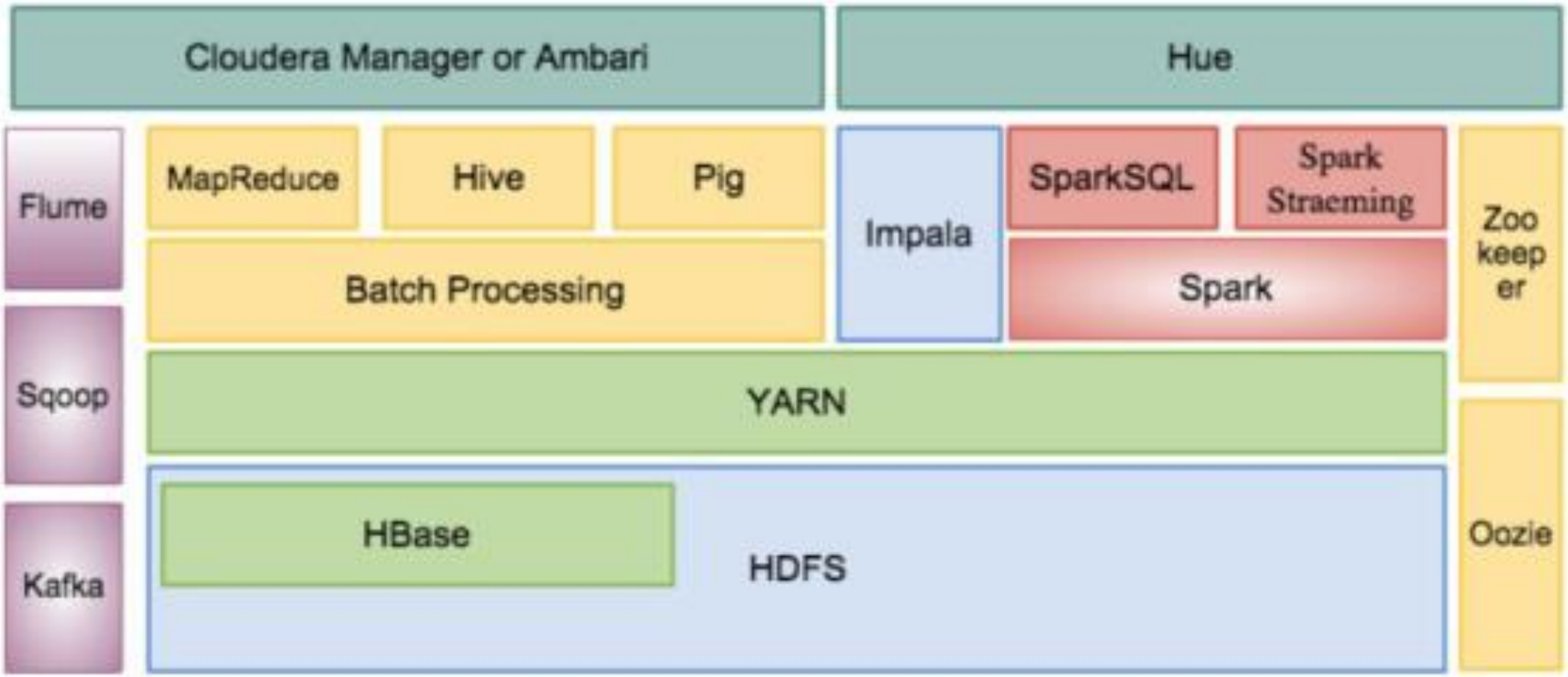
The stack is continually evolving and growing!

cloudera

Hadoop 2.x Ecosystems



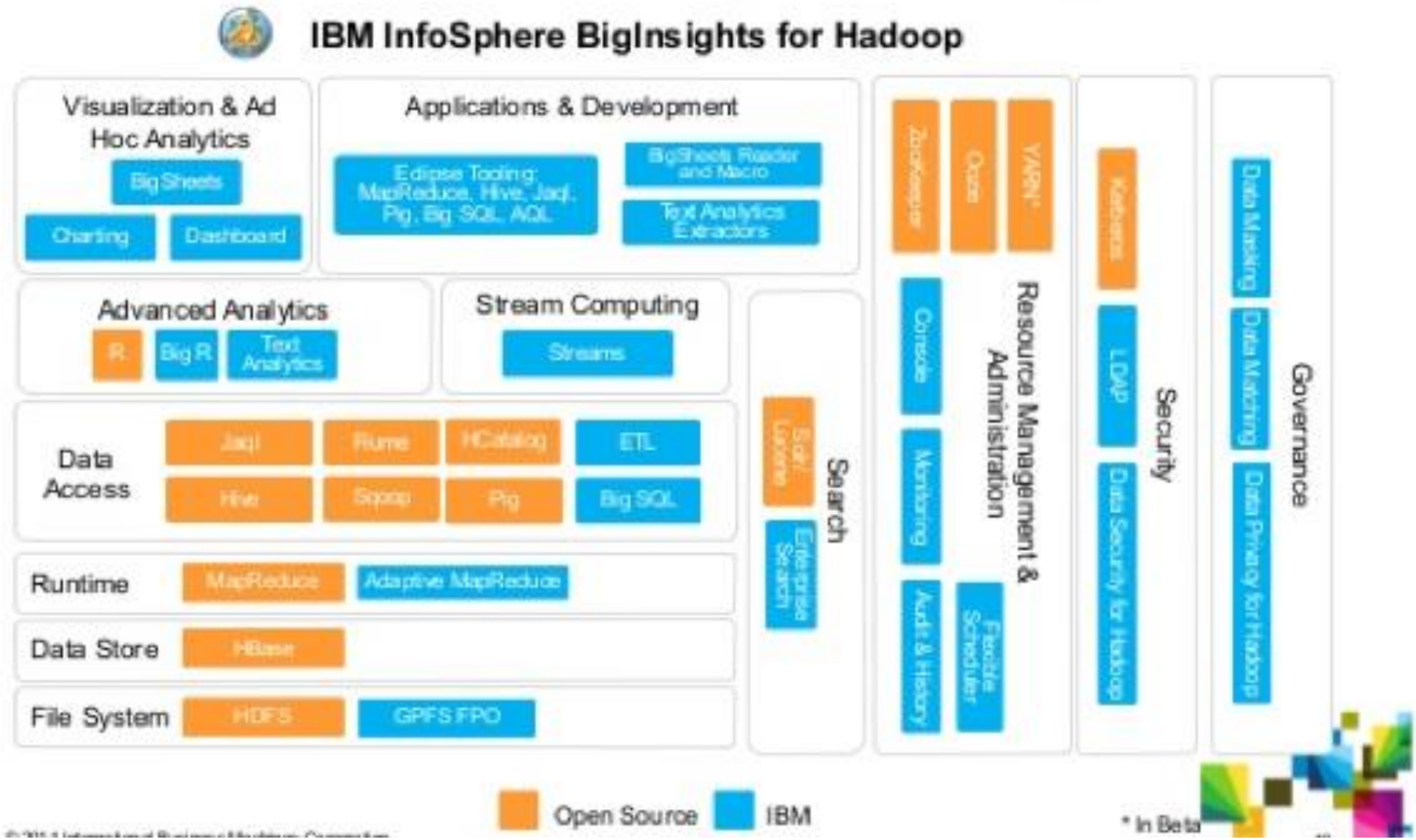
Hadoop Ecosystems



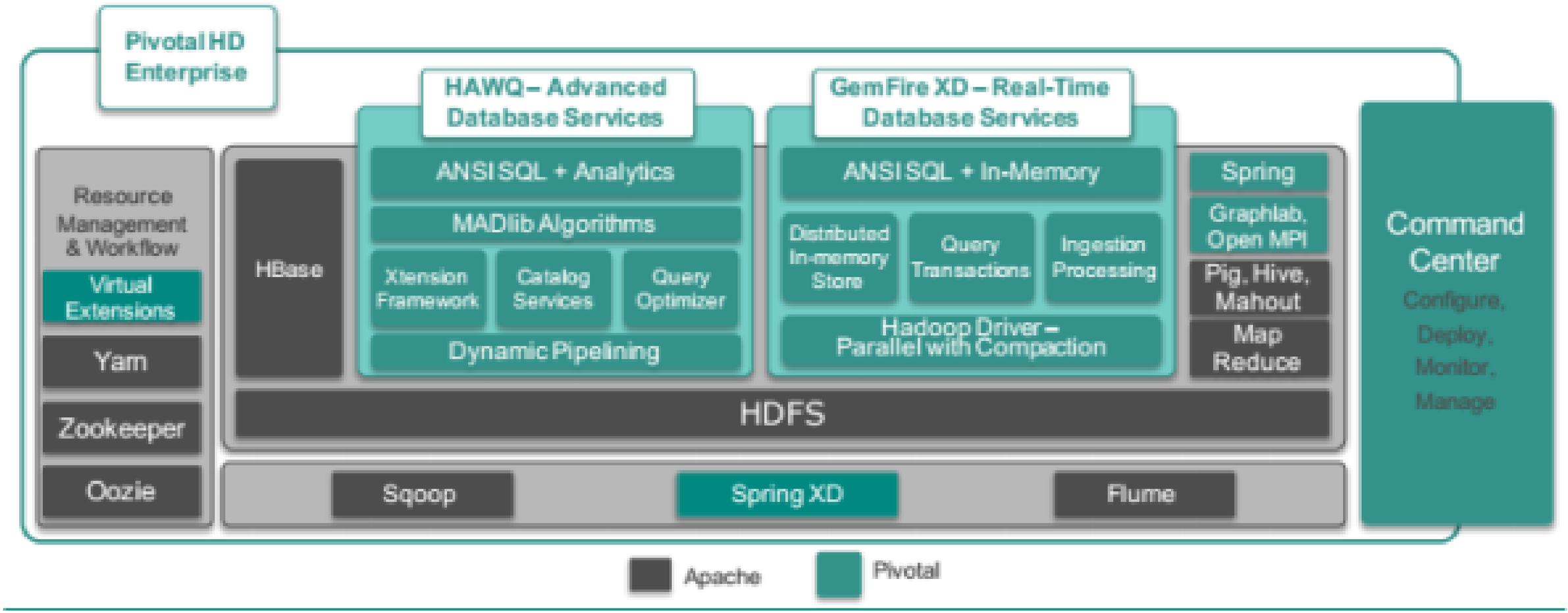
Hadoop Distribution

Microsoft Azure

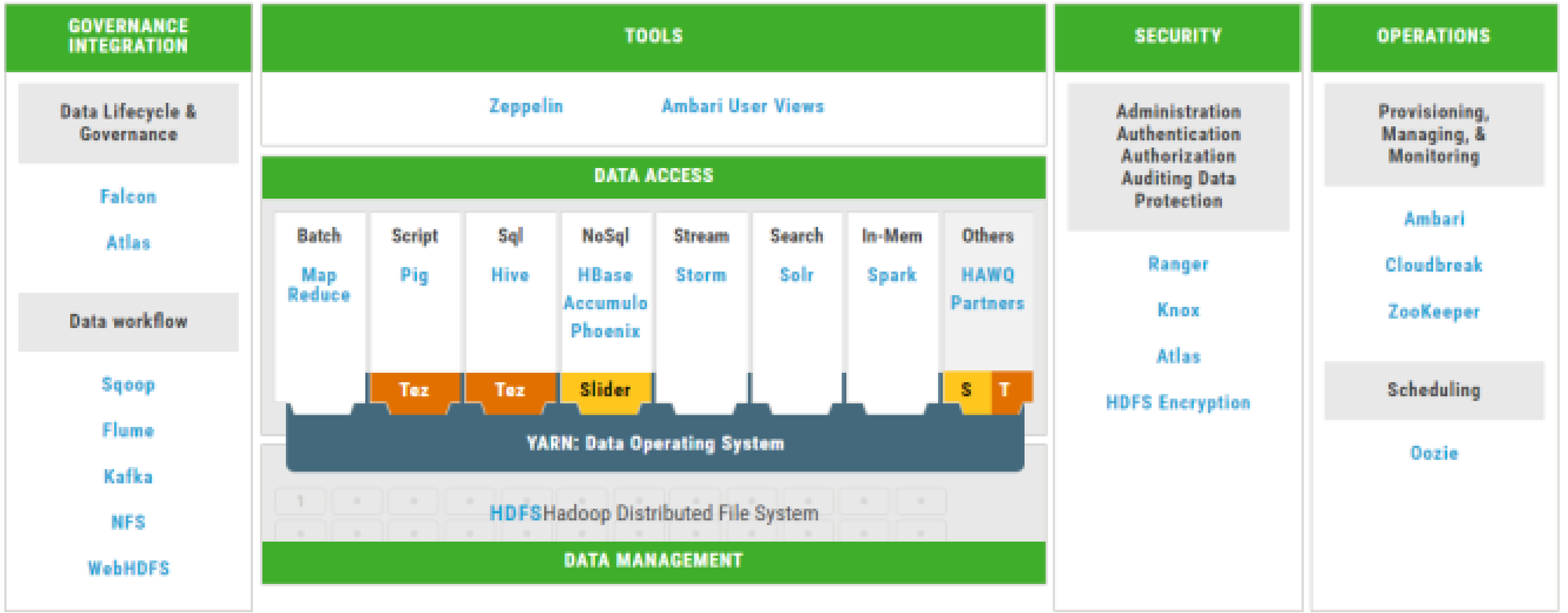
IBM InfoSphere BigInsights



Pivotal HD Architecture

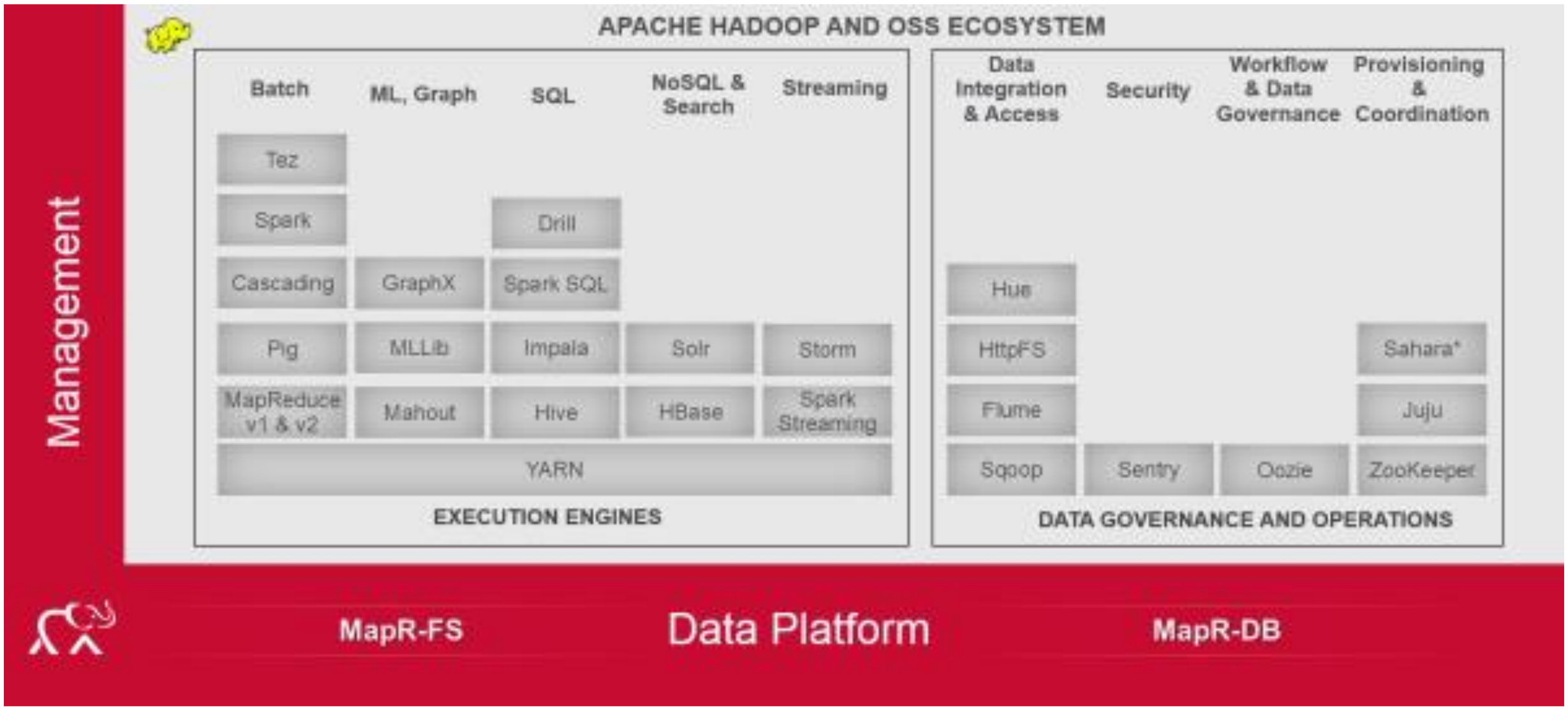


Hortonworks

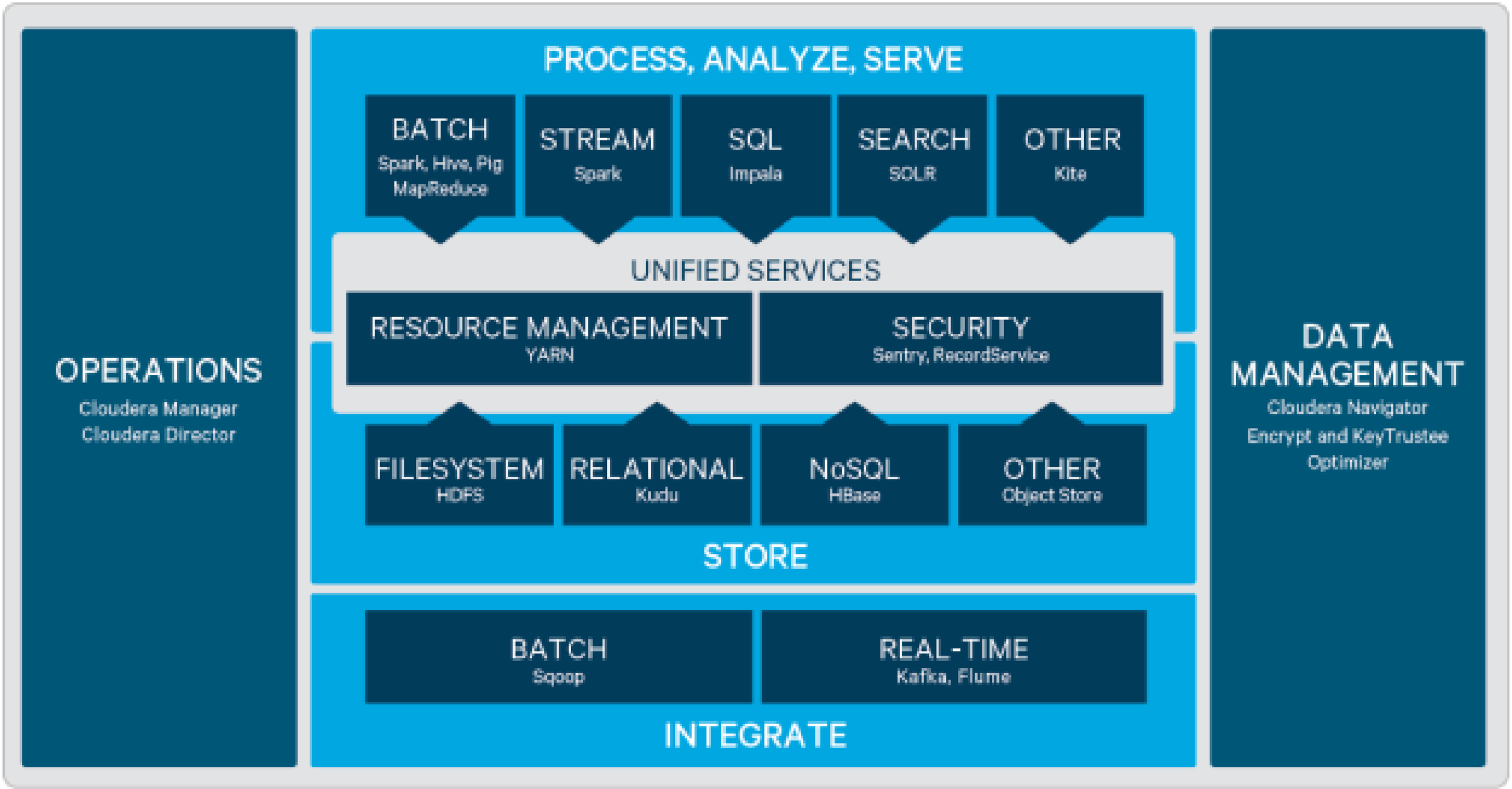


source : http://hortonworks.com/products/data-center/hdp/

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
Management
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

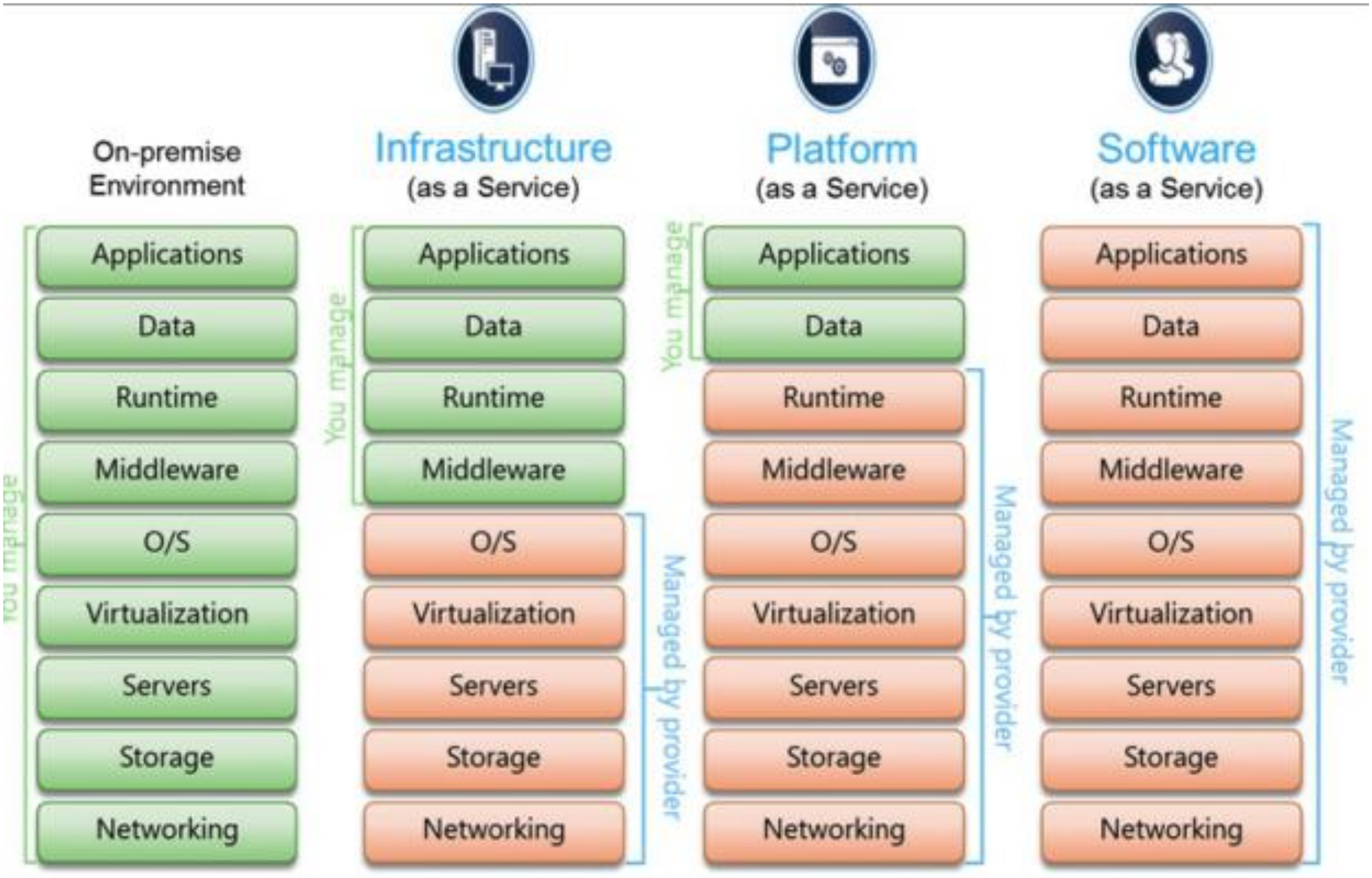
Big Data Using Public Cloud

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

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

Google Cloud

Google app engine

amazon web services™

salesforce 1 Platform

CLOUD

the rackspace cloud

OPENSHIFT

Microsoft Azure

CLOUD FOUNDRY

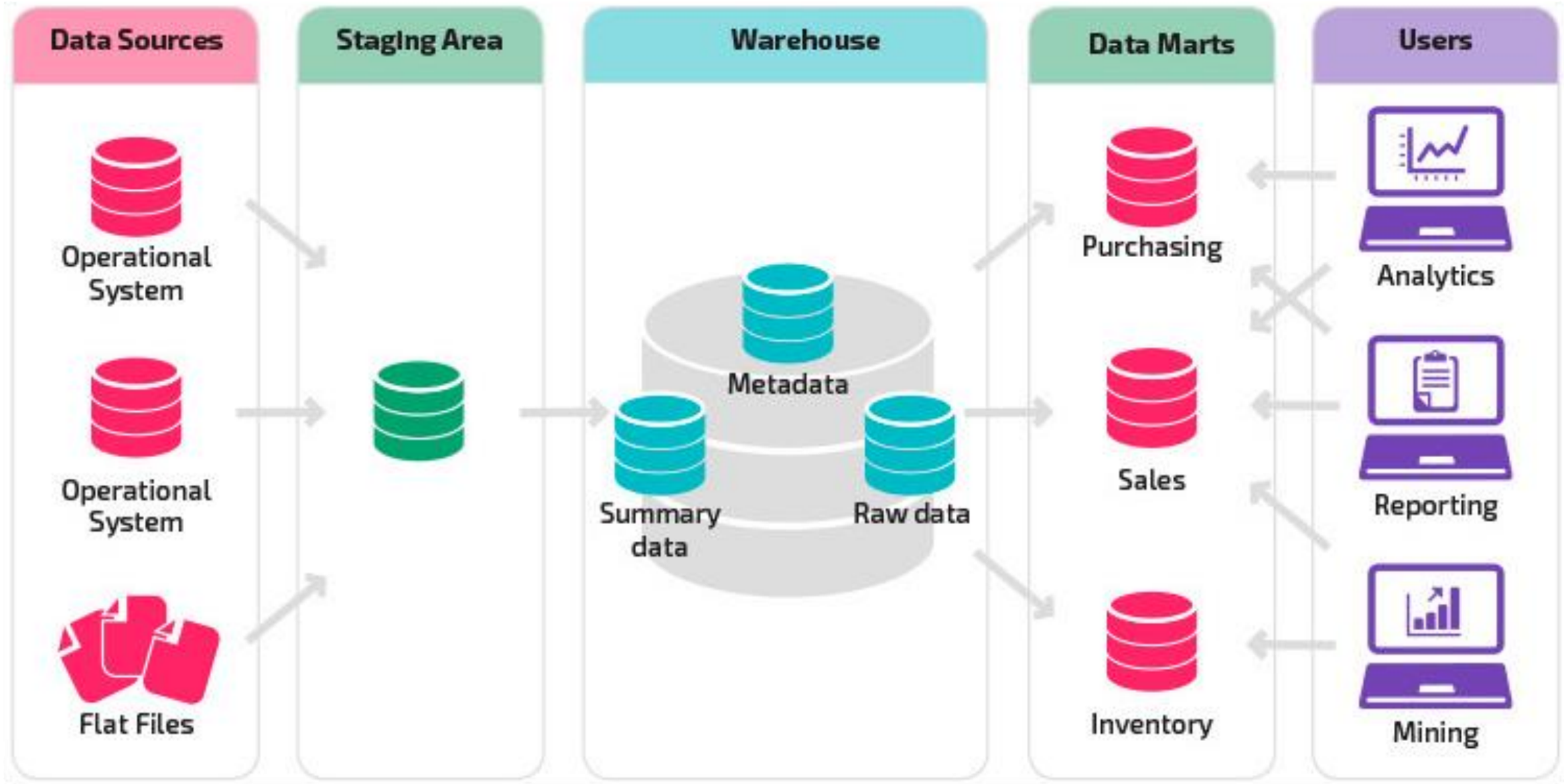
DATA LAKE

DATA LAKE

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Data Warehouse

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

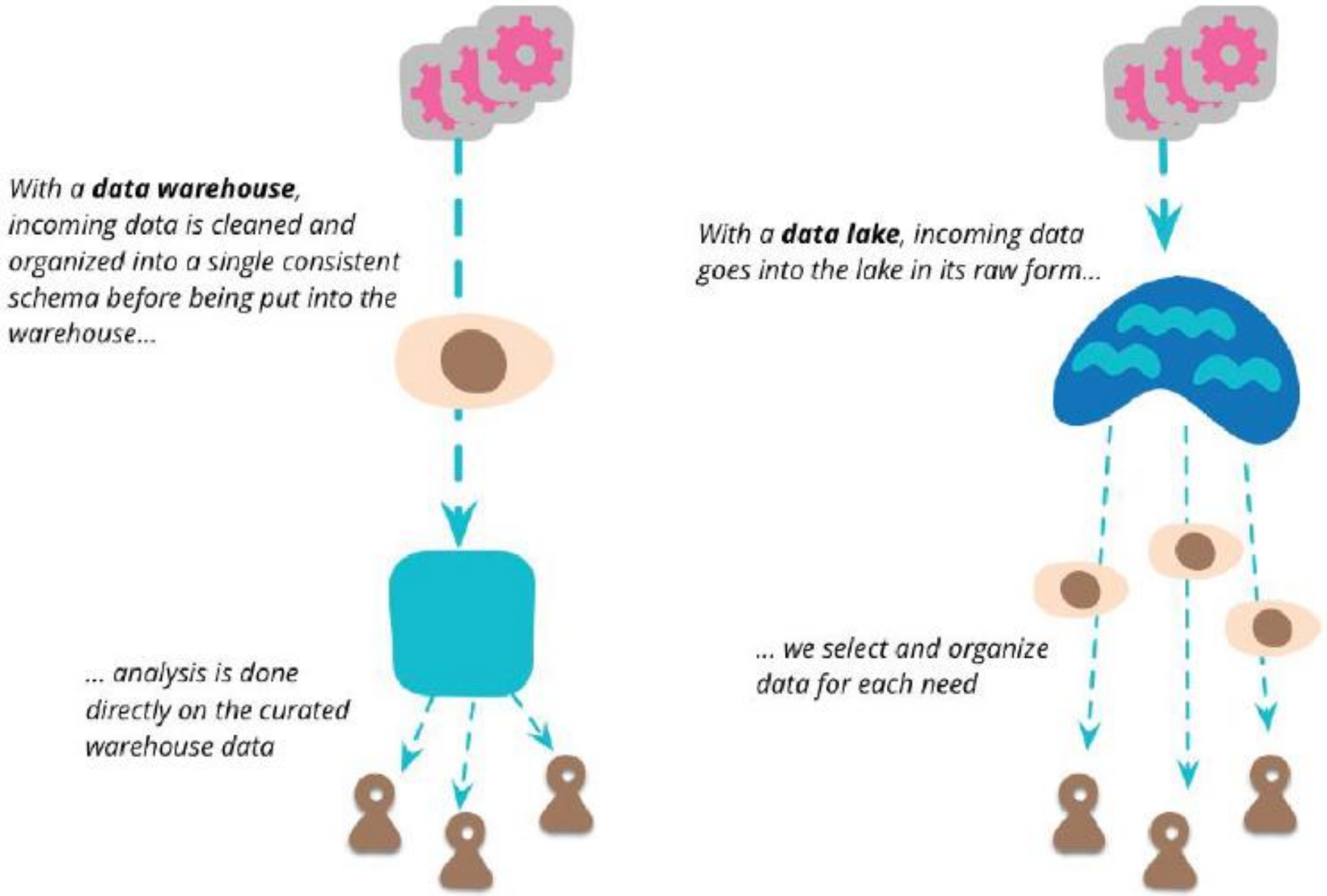


The old way: Ask, then collect

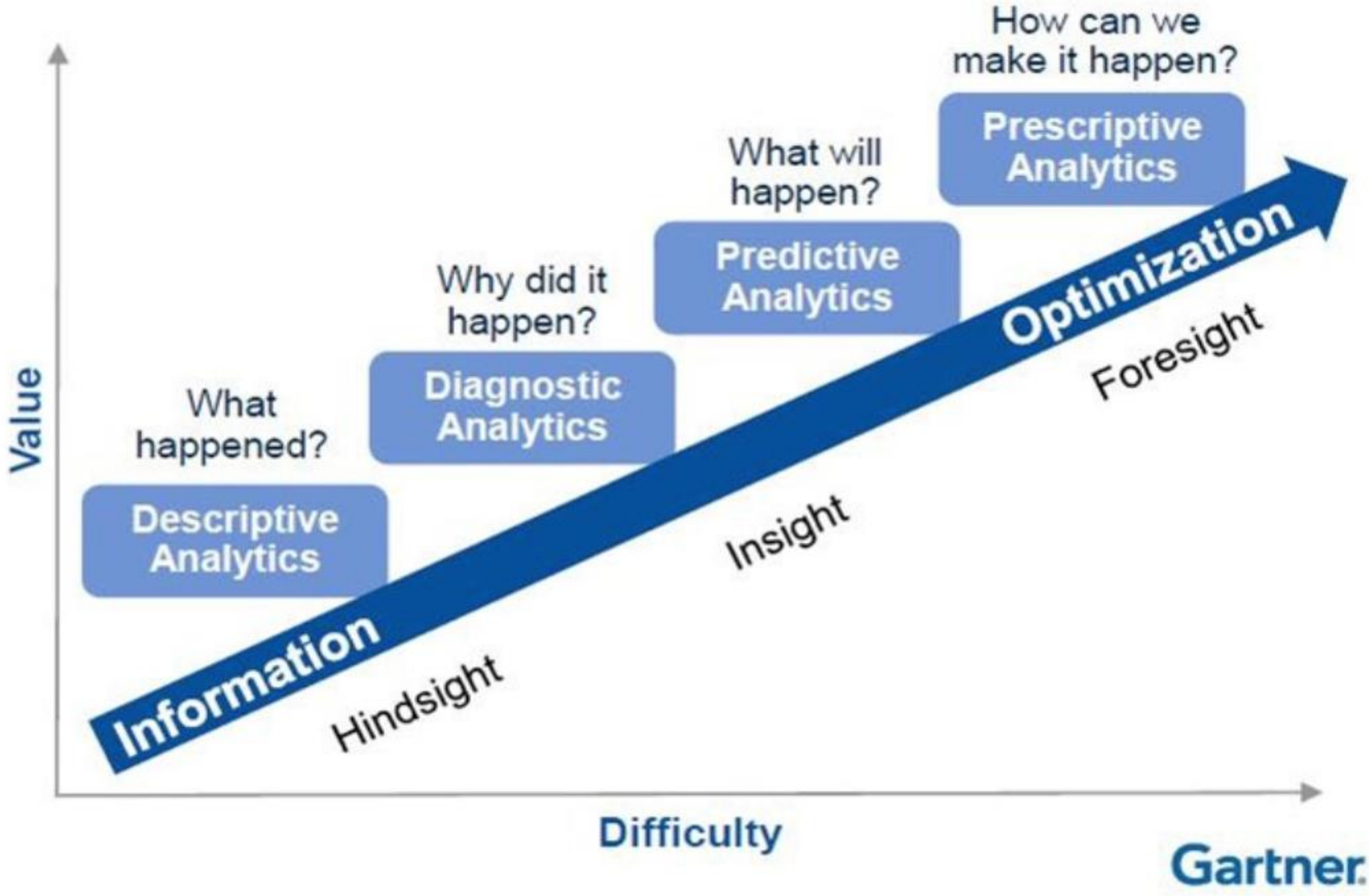
The new way: Collect, then ask

Source: Data Science and Critical Thinking, A, Croll

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

Business Intelligence

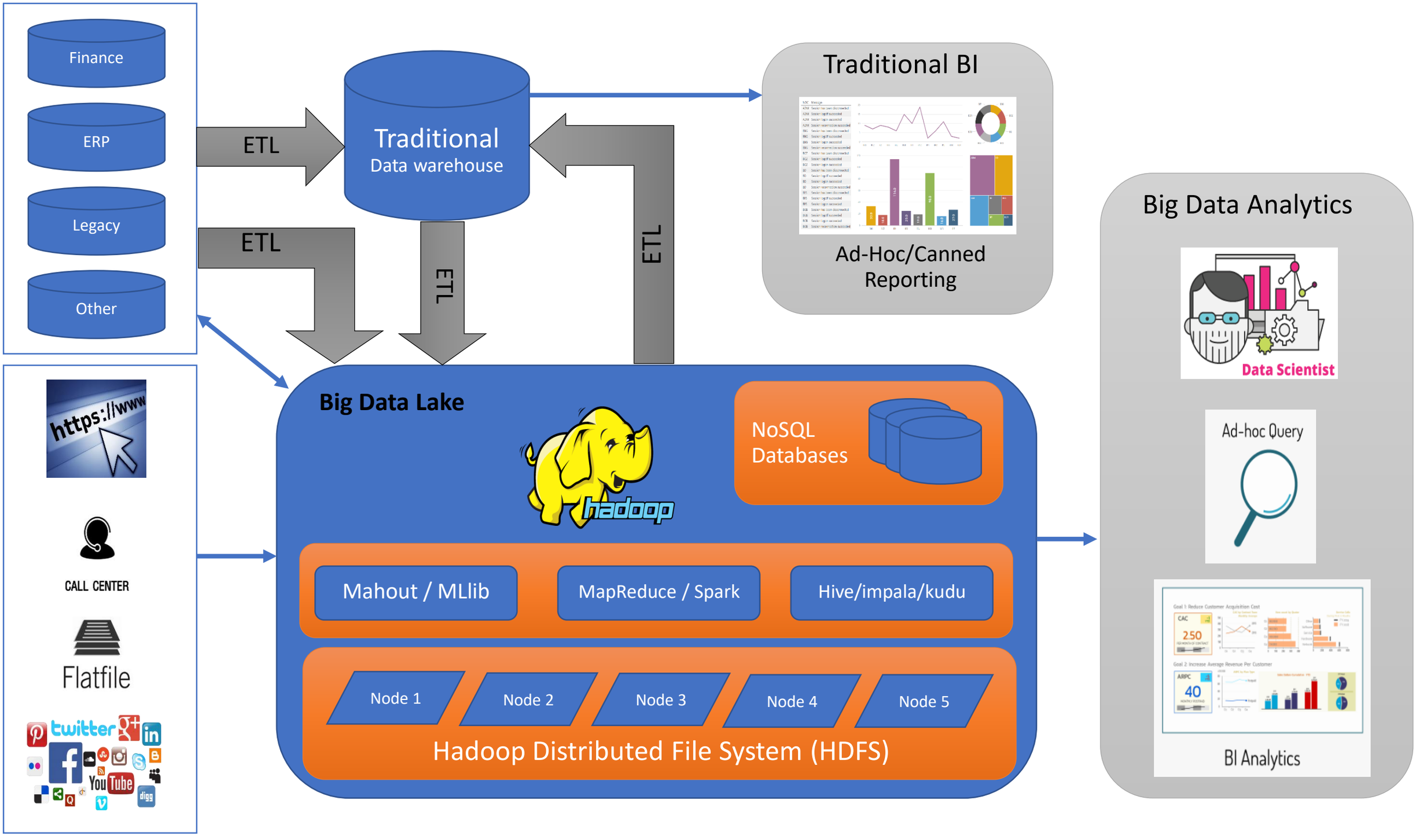
- เน้น Descriptive Analytics และ Diagnostic Analytics
- เน้นความรู้สถิติ
- ลักษณะข้อมูล
  - มีจำกัด
  - Cleaned data
  - Structure
  - Simple model
- เทคโนโลยี
  - Data warehouse

Big Data Analytics

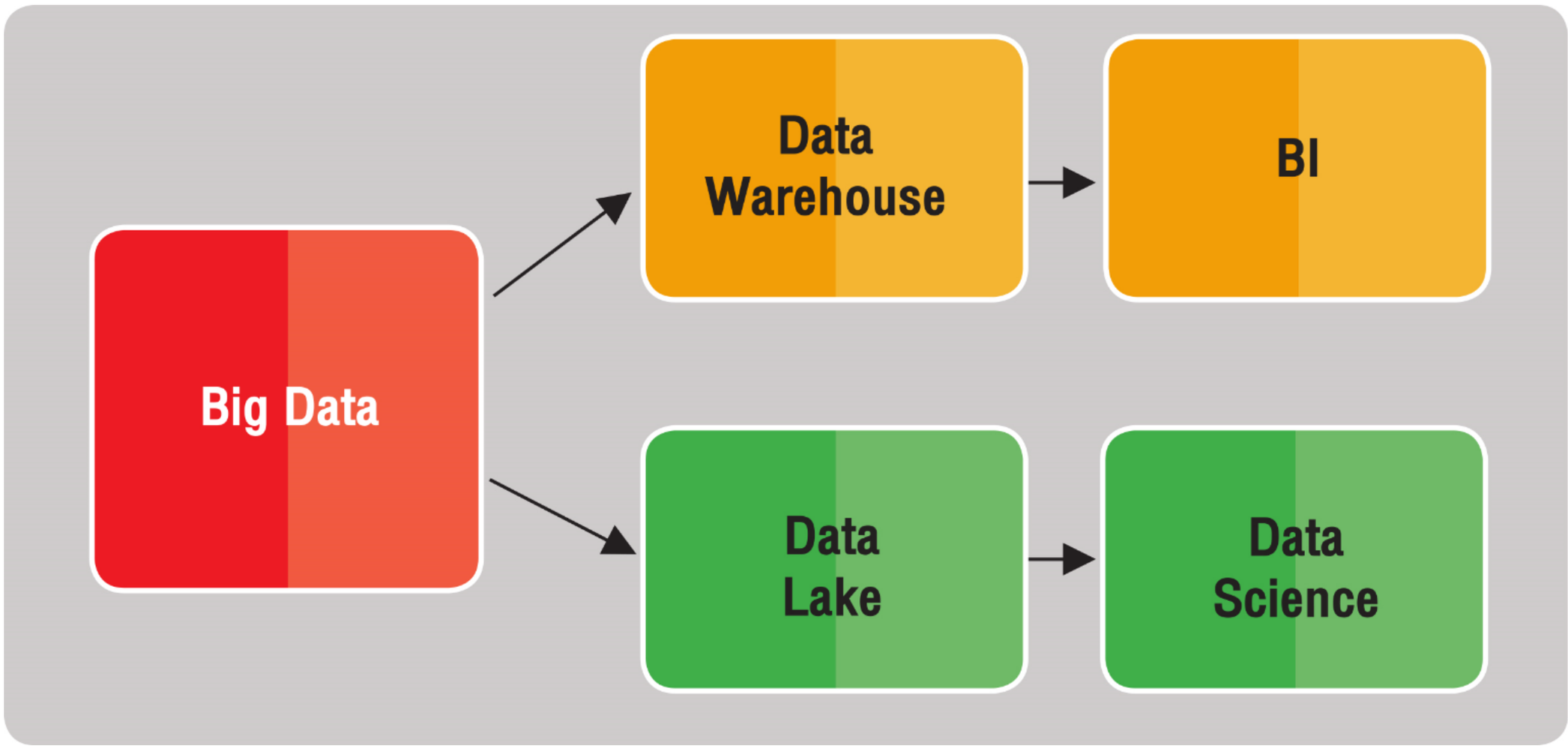
- เน้น Predictive Analytics และ Prescriptive Analytics
- เน้น Data Science
- ลักษณะข้อมูล
  - ขนาดใหญ่
  - ข้อมูลดิบ
  - หลากหลายชนิด
  - Complex model
- เทคโนโลยี
  - Data Lake

Today's business environment requires Big Data

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Analytics



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

3 jobs in data

	Data Engineer	Data Scientist	Business and Data Analyst
			
หน้าที่ (Responsibilities)	ออกแบบช่องทางของข้อมูล และวิธีการจัดเก็บ และใช้งาน	ออกแบบโมเดลจากข้อมูล เพื่อหาช่องทางใหม่ๆ ให้องค์กร	วิเคราะห์ และออกแบบการนำเสนอข้อมูล เพื่อแก้ปัญหาในส่วนต่างๆ ของธุรกิจ
การศึกษา (Education)	- วิศวกรรมศาสตร์คอมพิวเตอร์ - เทคโนโลยีระบบข้อมูล	- วิศกรรมอุตสาหการ - การวิจัยเชิงปฏิบัติการ - วิทยาศาสต์คอมพิวเตอร์ - สถิติ	- การตลาด - เศรษฐศาสตร์ - การเงิน - สถิติ - โลจิสติกส์ และอื่นๆ
คุณสมบัติ (Skills)	- Database design - Production coding - Data warehousing - Data transformation - Cloud computing - Big data design	- Mathematics - Propability - Statistics - Machine Learning - Programming - Data Visualization - Business	- Analysis - Statistics - Business - Communication - Project management
เครื่องมือ (Program)	- Oracle - Hadoop - SQL - NoSQL - Pig - Hive - Spark - MongoDB - SAP - Pentaho - Cloudera - JavaScript	- Python - R - SAS - MATLAB - Julia - Spark - Tableau	- R - SPSS - SAS - Tableau - Microsoft Excel