CHULALONGKORN
BUSINESS SCHOOL
FLAGSHIP FOR LIFE

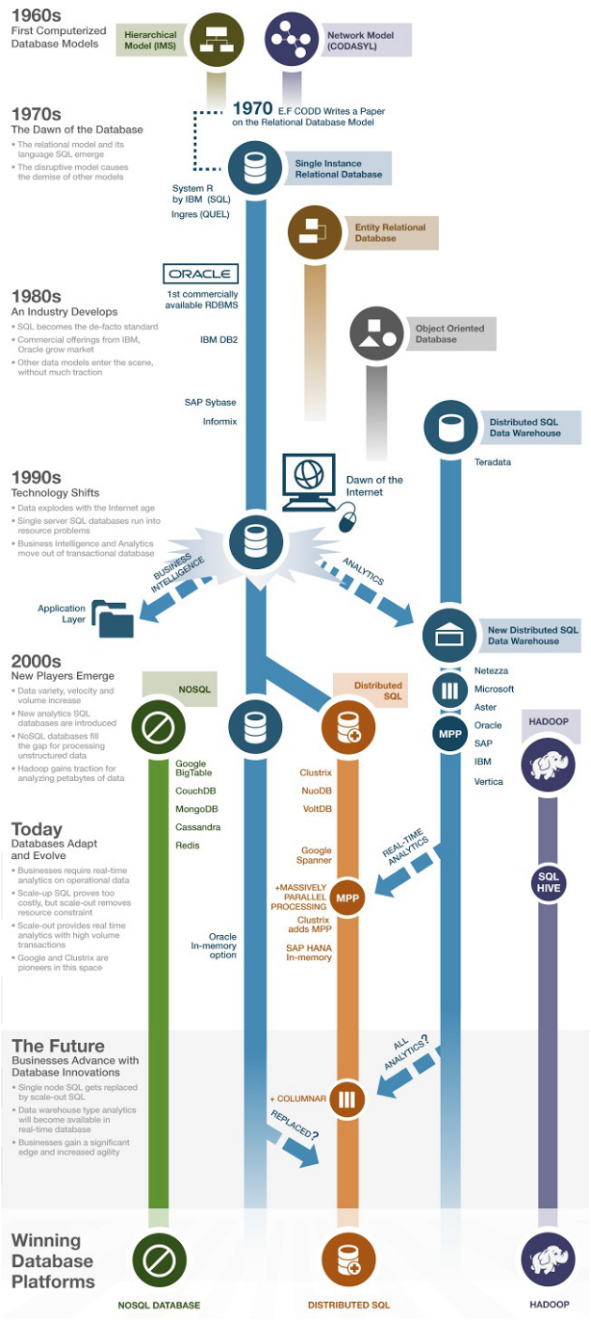
Big Data Technology

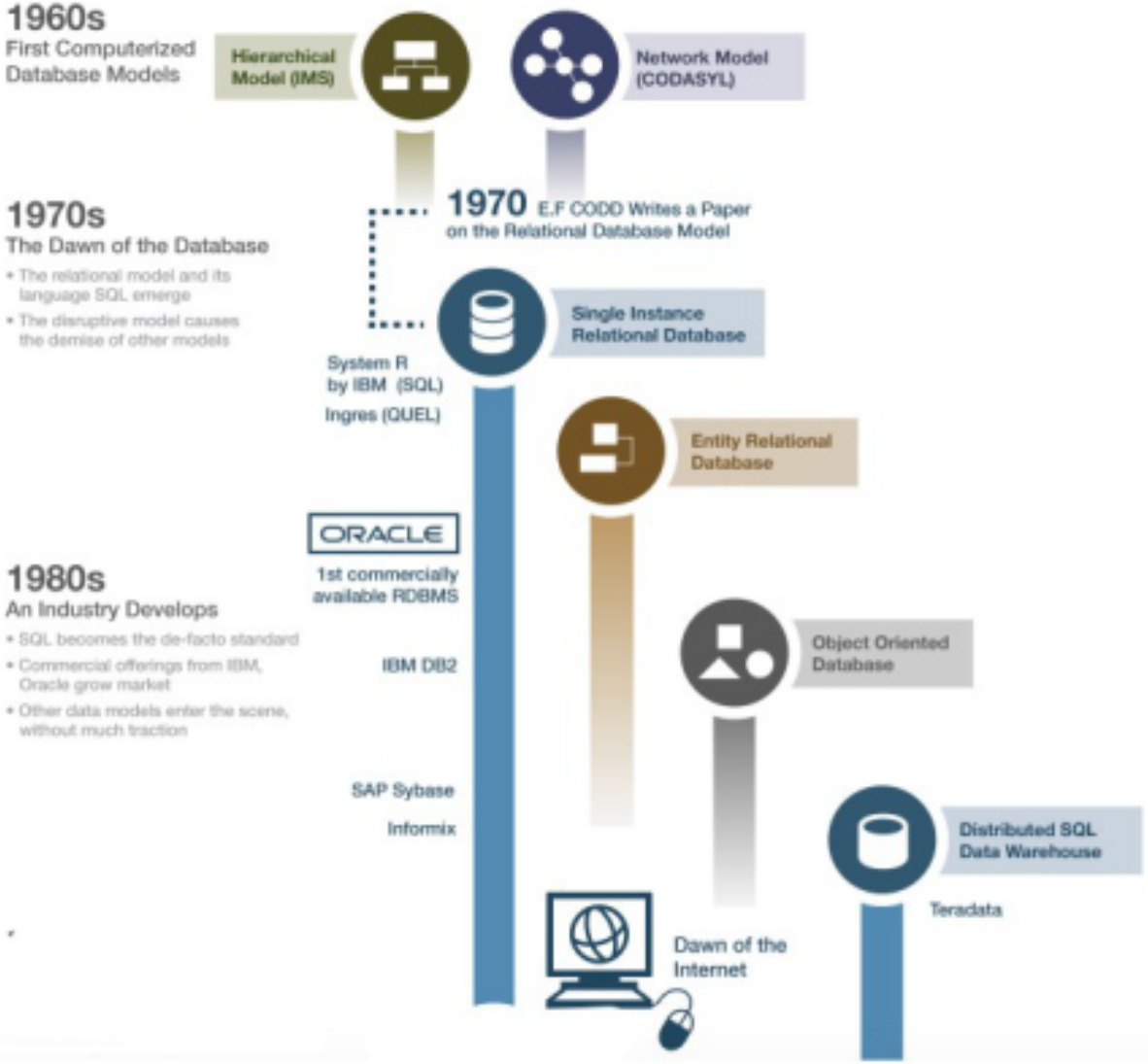
Thaveewat Khanan

2603615 Organization Data Management

The Future of The DATABASE

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

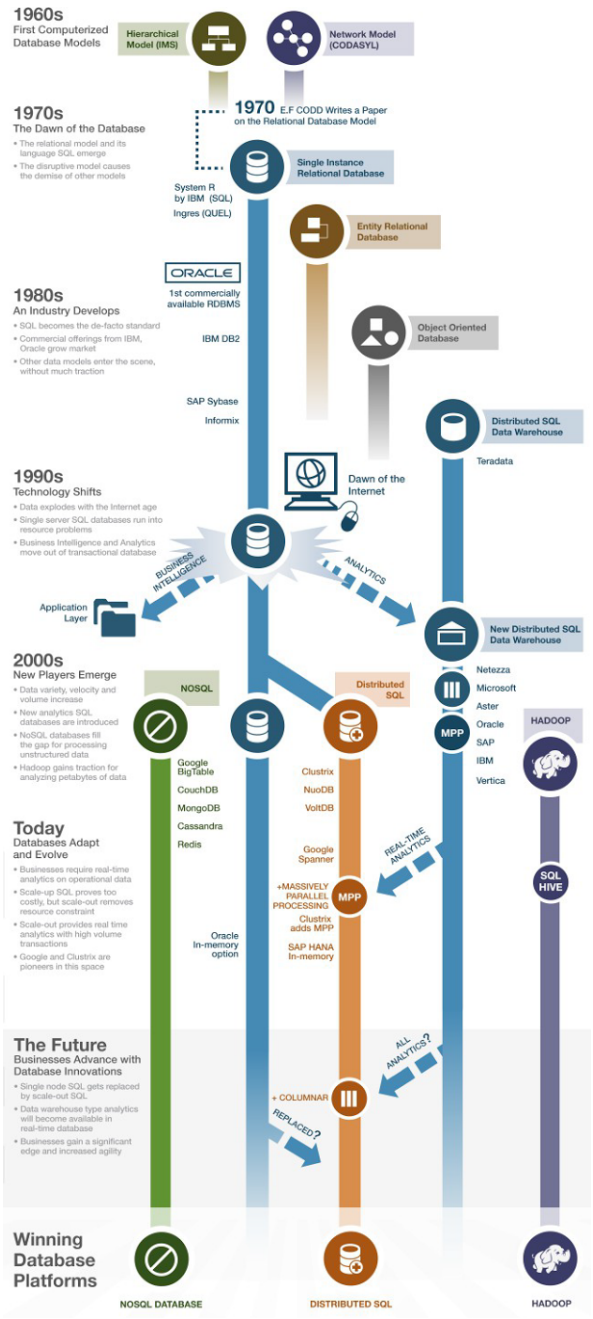




Source: Robin Purohit ; http://insights.wired.com/ 3

The Future of The DATABASE

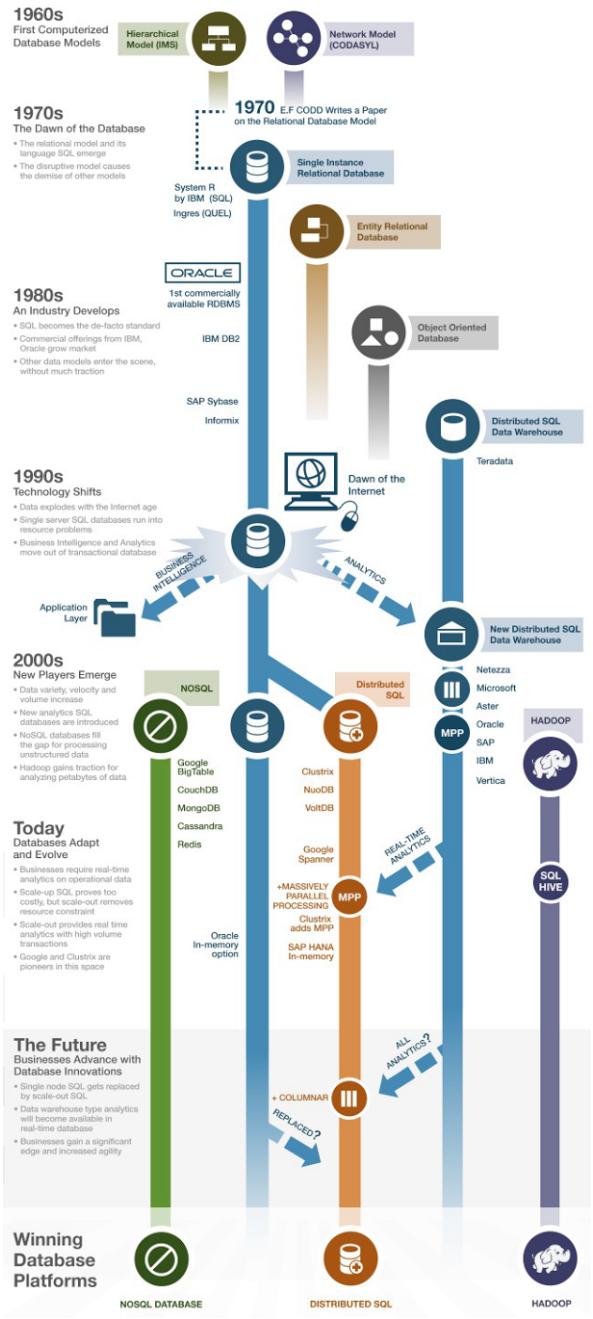
CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Source: Robin Purohit ; http://insights.wired.com/ 3

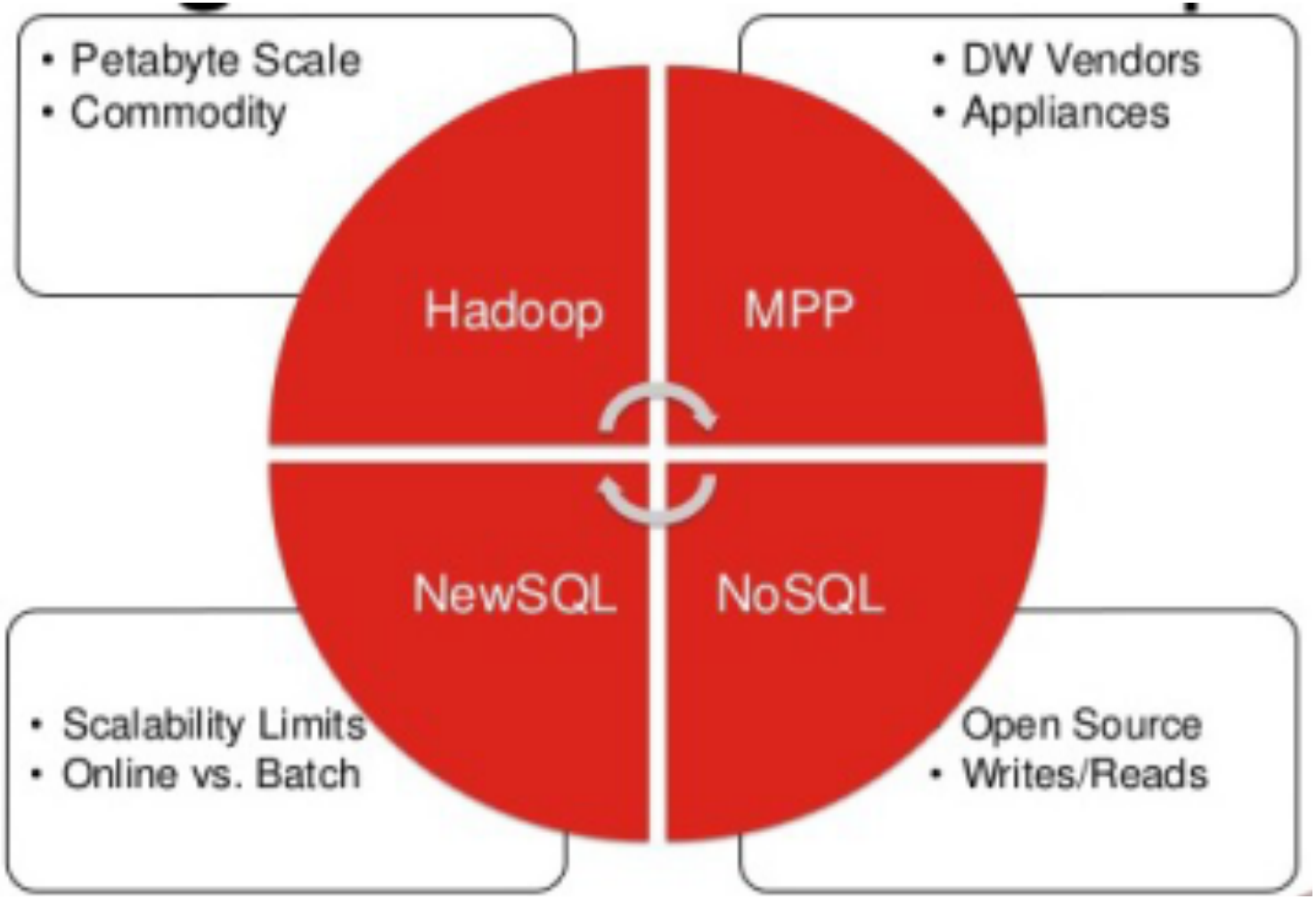
The Future of The DATABASE

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Source: Robin Purohit ; http://insights.wired.com/ 3

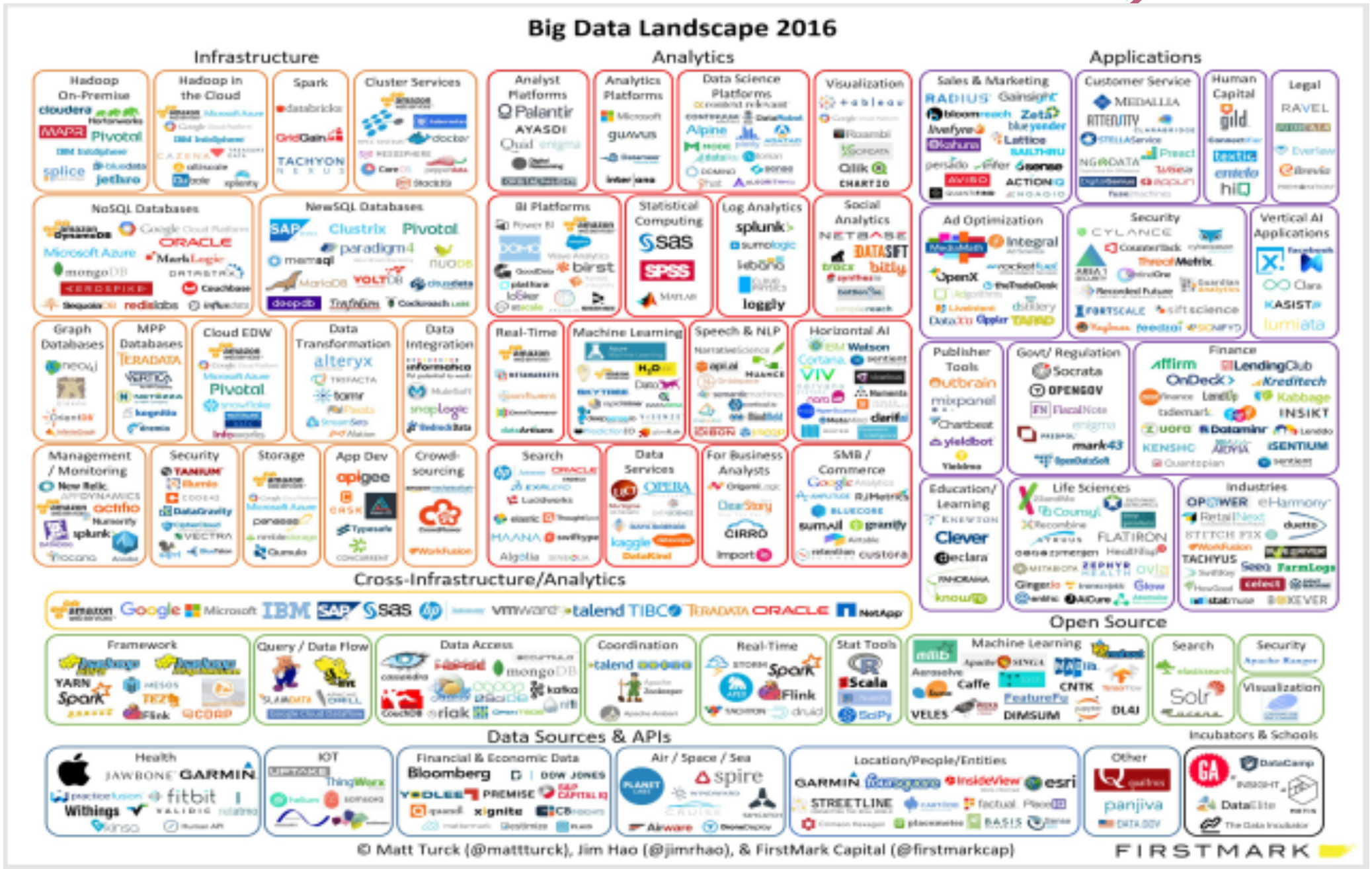
Big Data Landscape



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

Big Data Landscape 2016

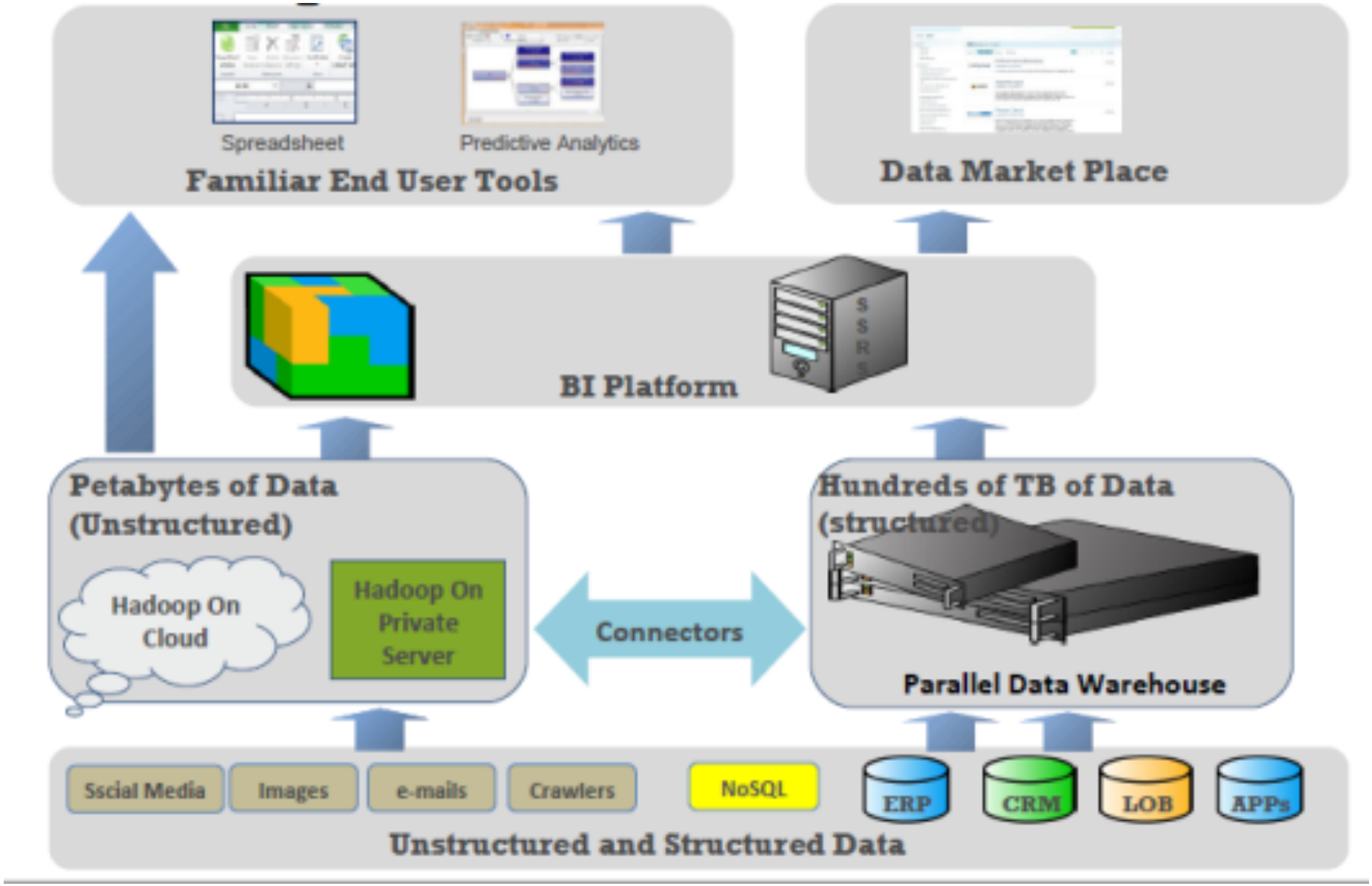
CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Thaveewat Khanan

2603615 Organization Data Management

Big Data Future Architecture

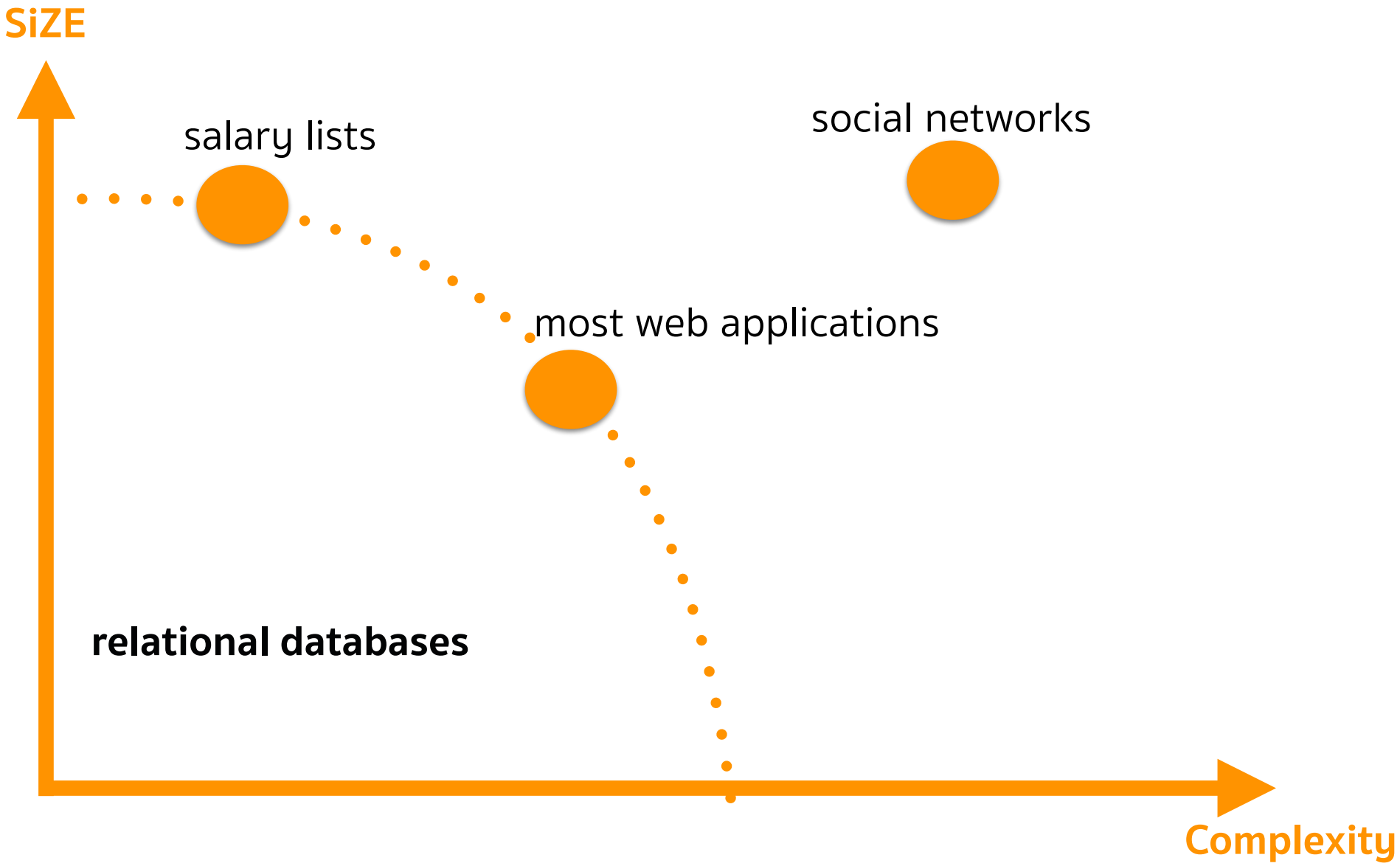


What is NoSQL ?

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

Thaveewat Khanan

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

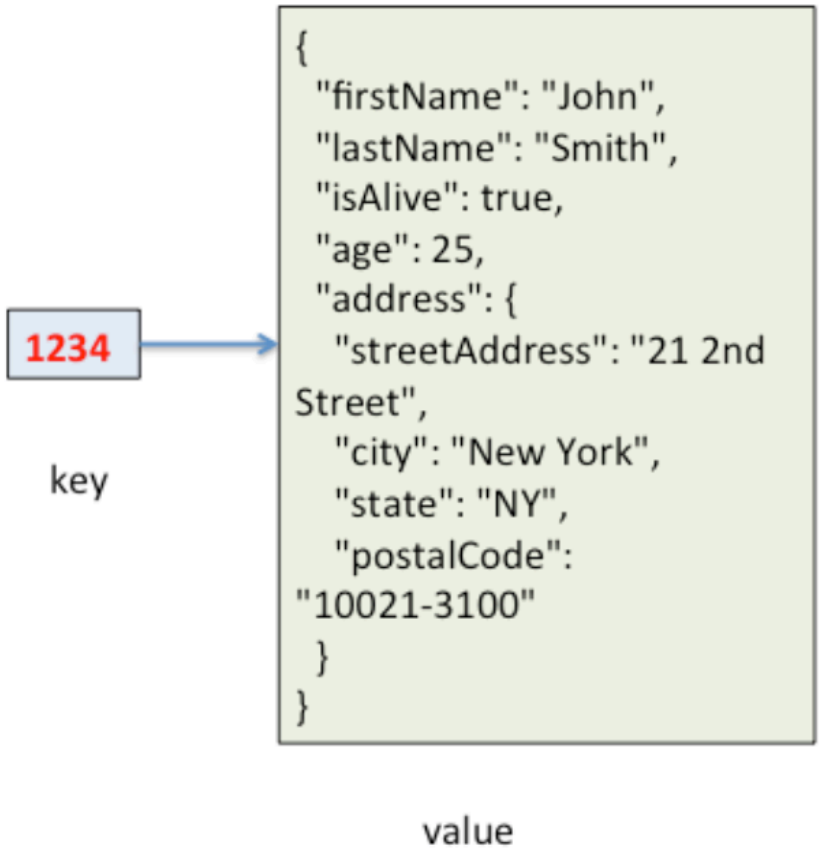
id	Interests
1	Soccer
1	Movies
1	Baseball
3	Music

Key-value store database

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

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 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

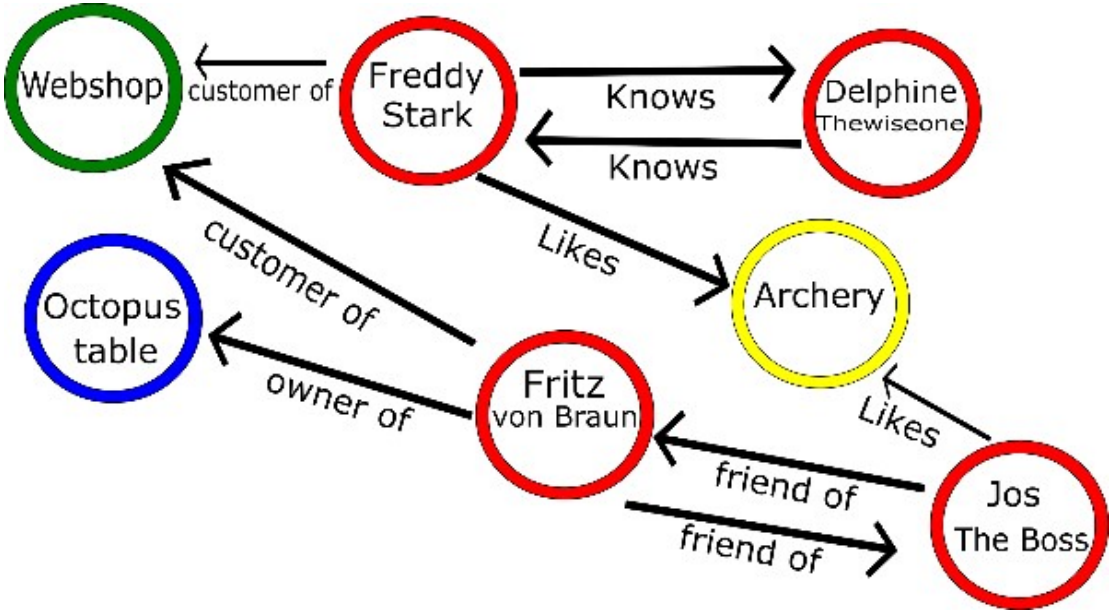
CouchDB relax

mongoDB

Graph database

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

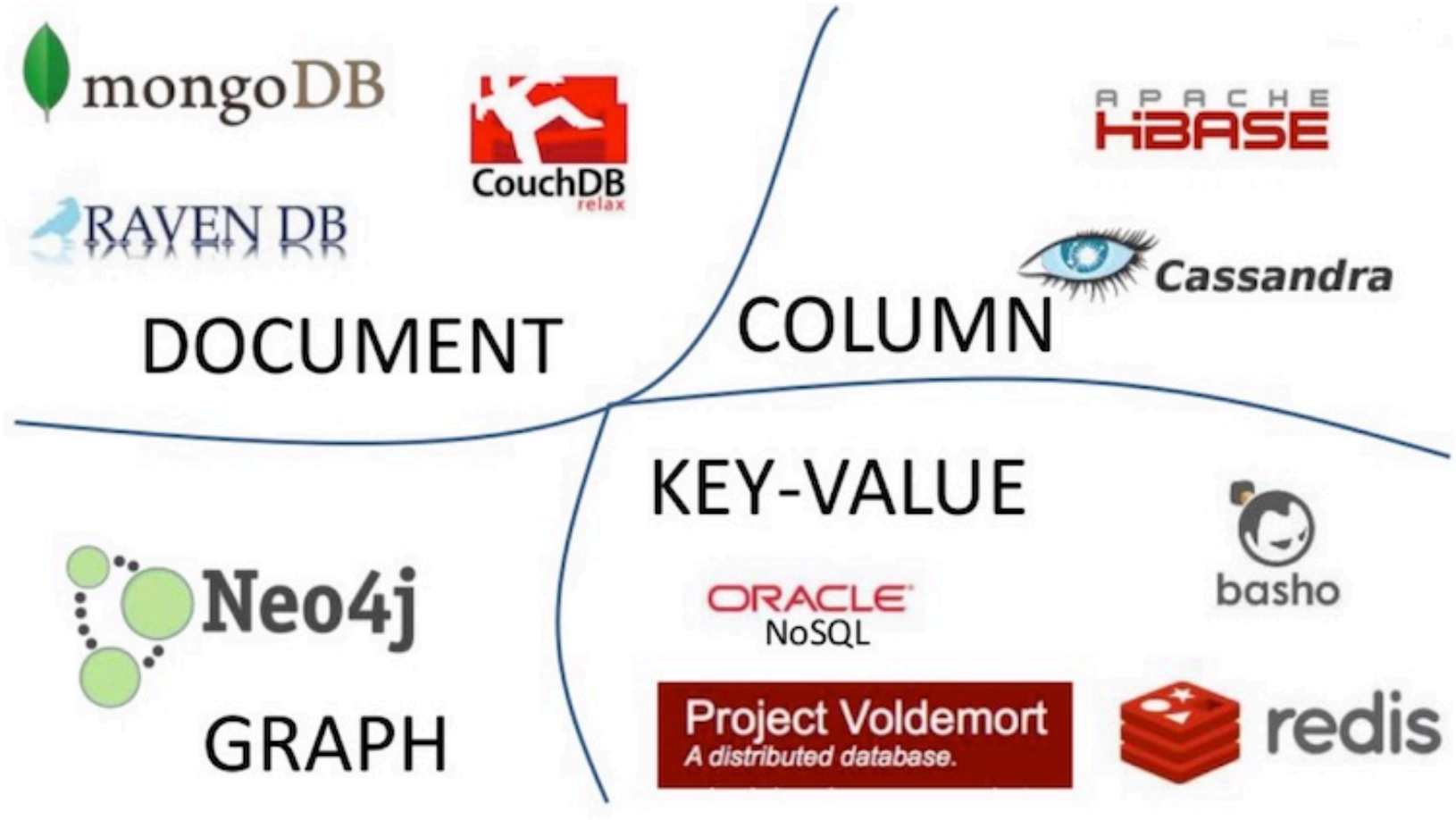
A database that uses graph structures for semantic queries with nodes, edges, and properties to represent and store data.

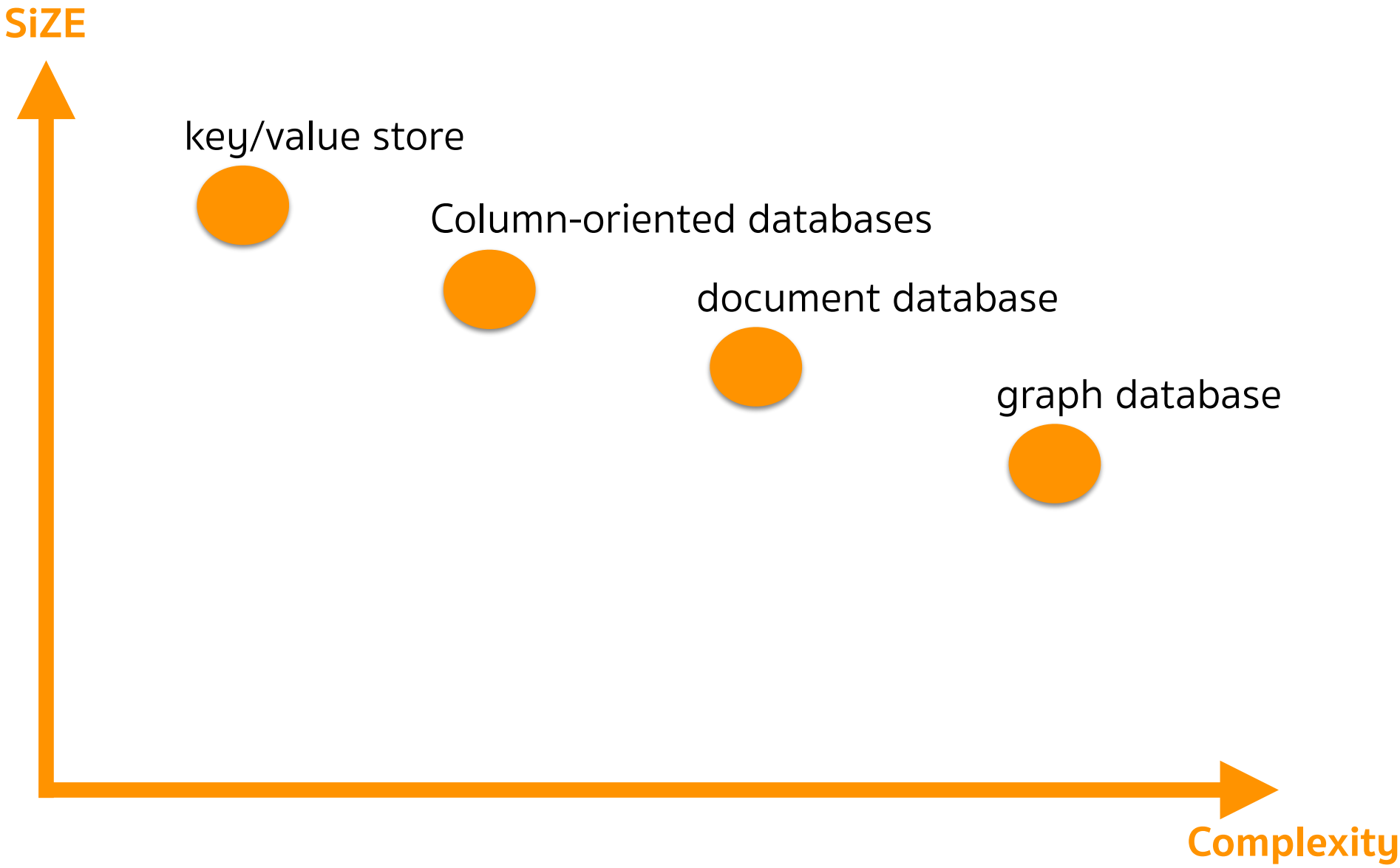


Neo4j
the graph database

Franz Inc.
AllegroGraph

InfiniteGraph


mongoDB
CouchDB
relax
RAVEN DB
APACHE
HBASE
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

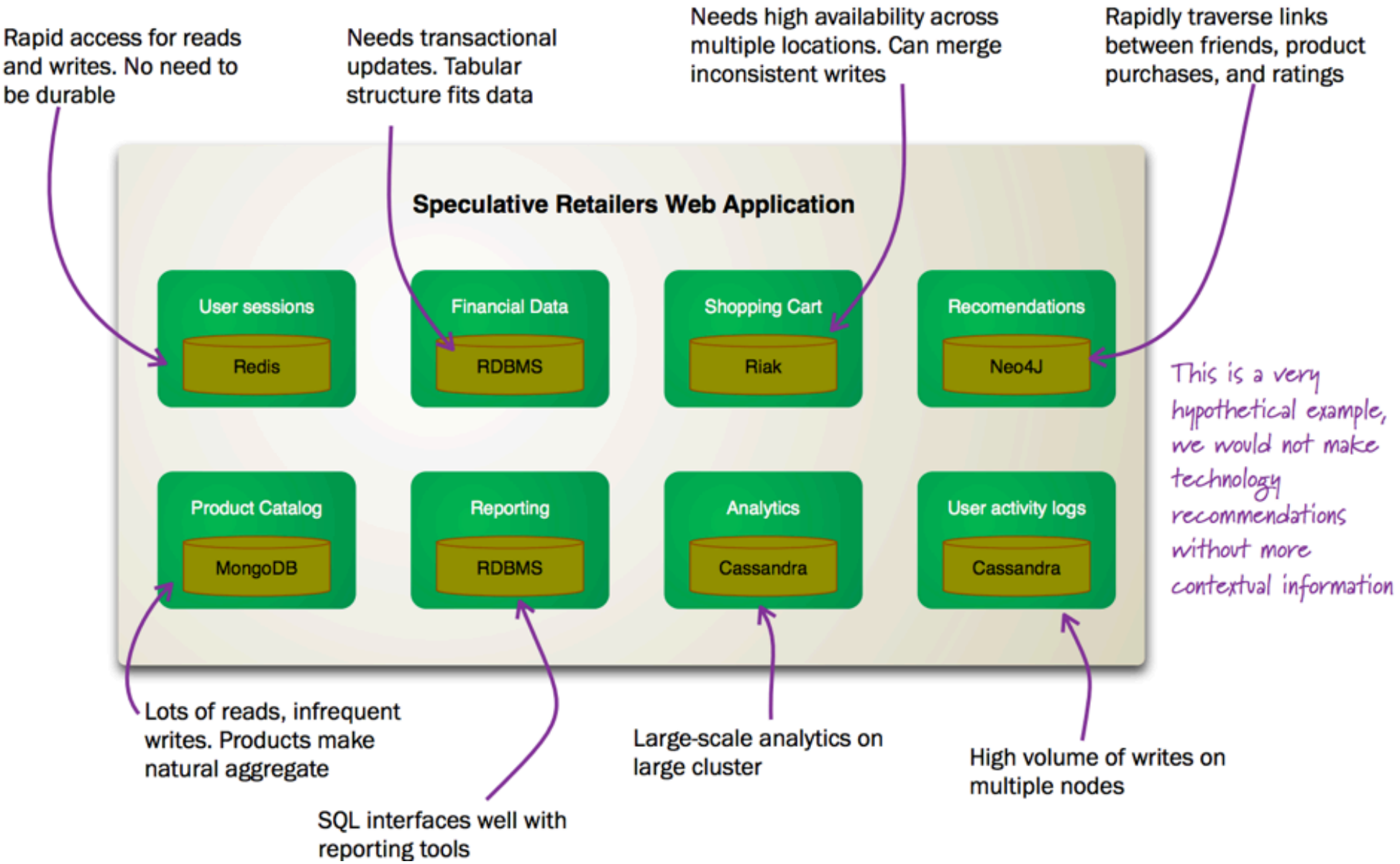
works great, can’t scale for large data

NoSQL

works great, doesn't fit all situations

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

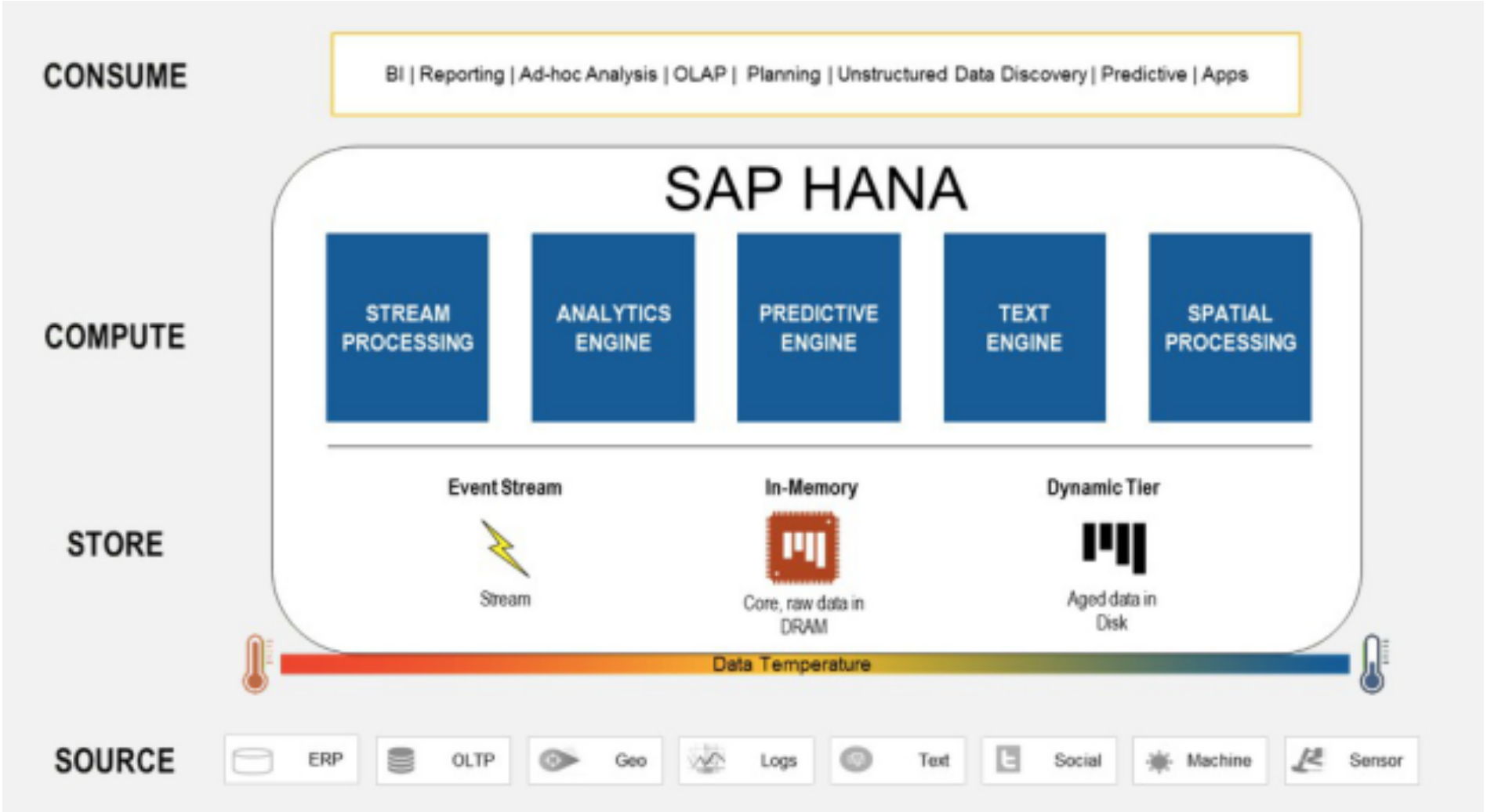
Thaveewat Khanan



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

MPP: SAP Hana

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Oracle Exadata

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Thaveewat Khanan

2603615 Organization Data Management

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

Google publishes GFS & MapReduce papers

NY Times converts 4TB of image archives over 100 EC2s

Yahoo! hires Cutting, Hadoop spins out of Nutch

Facebooks launches Hive: SQL Support for Hadoop

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

cloudera Founded

Doug Cutting joins Cloudera

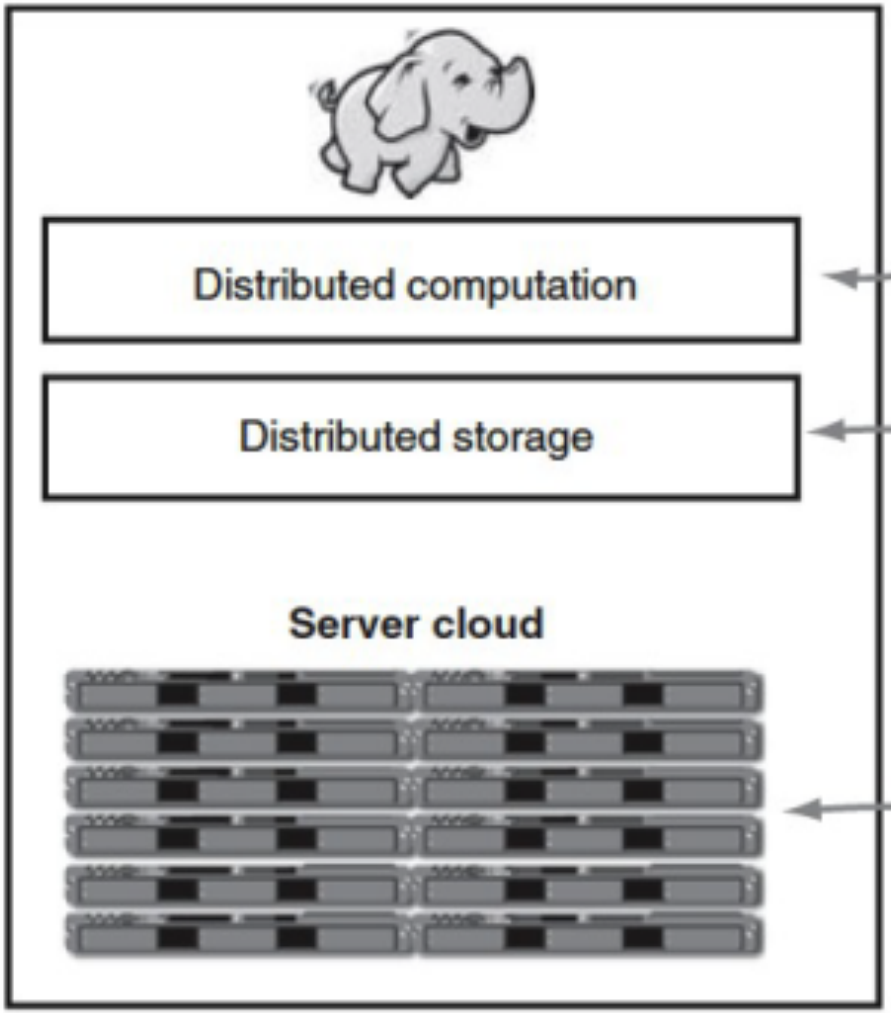
Hadoop Summit 2009, 750 attendees

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

Source: Cloudera, Inc.

Hadoop Environment

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



The computation tier uses a framework called MapReduce.

A distributed filesystem called HDFS provides storage.

Hadoop runs on commodity hardware.

Major Hadoop Components

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Map/Reduce System

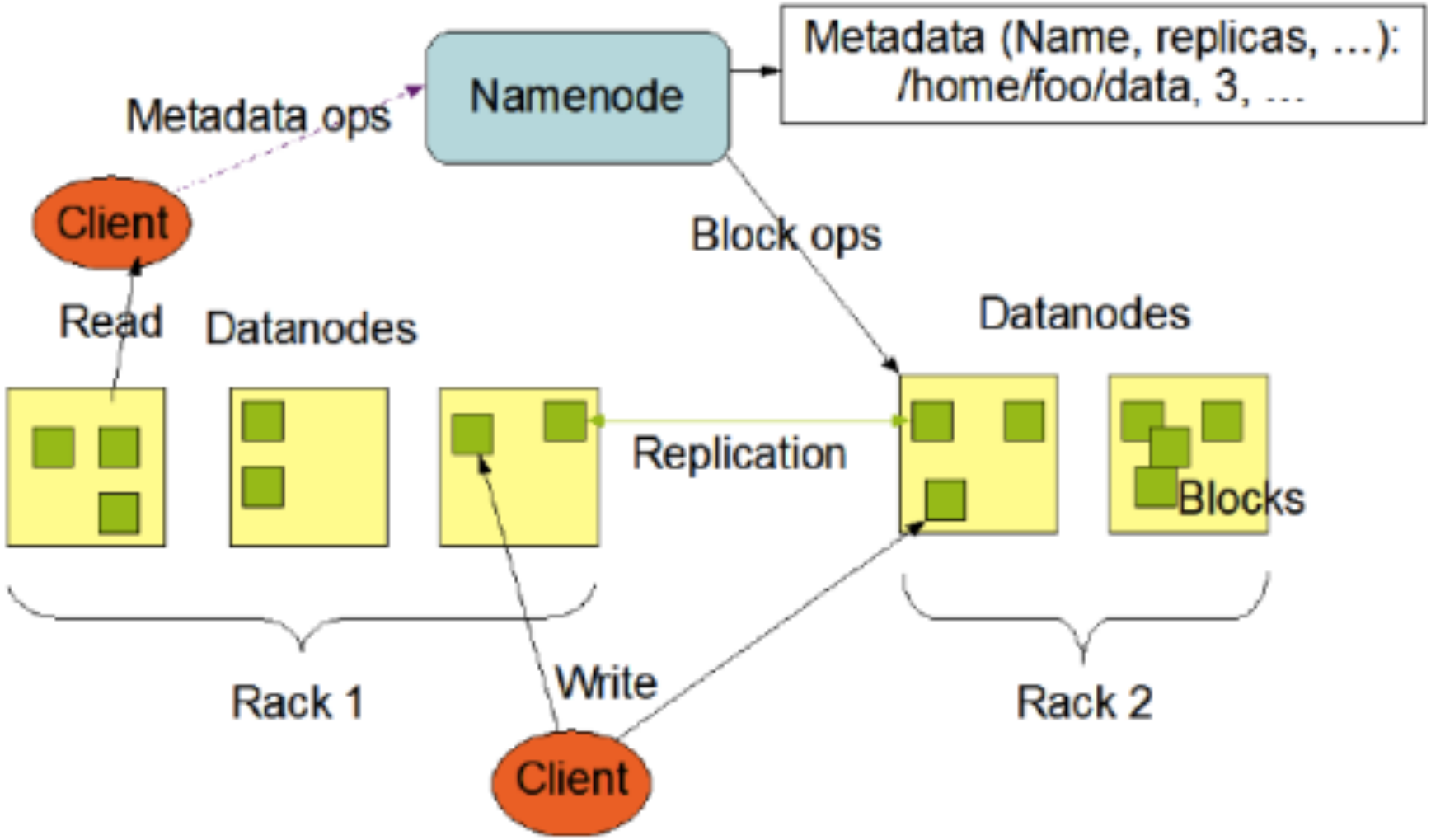
Hadoop Distributed File System (HDFS)

HDFS

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

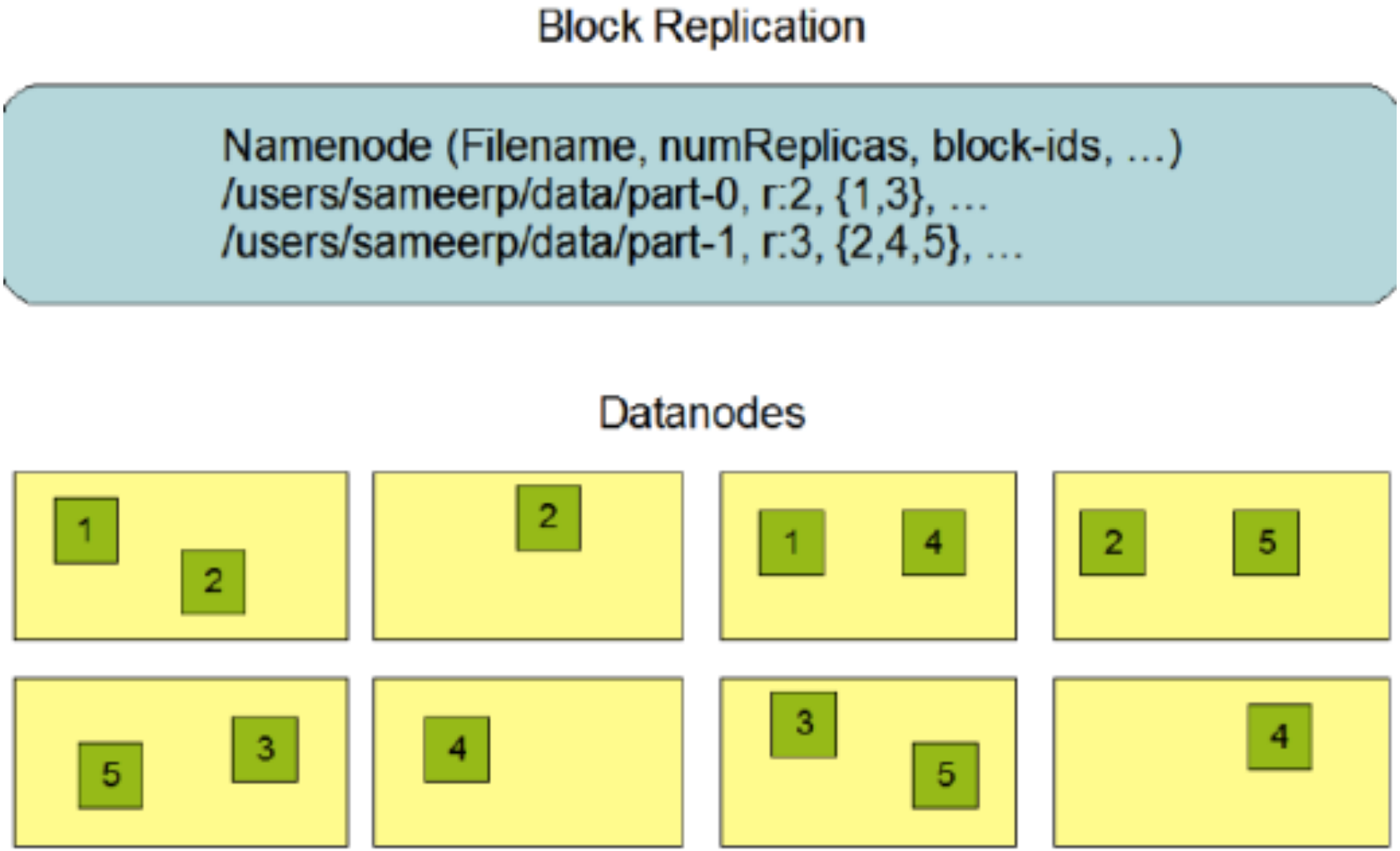
- Default storage for the Hadoop cluster
- Data is distributed and replicated over multiple machines
- Designed to handle very large files with streaming data access patterns.
- NameNode/DataNode
- Master/slave architecture (1 master 'n' slaves)
- Designed for large files (64 MB default, but configurable)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

HDFS Architecture

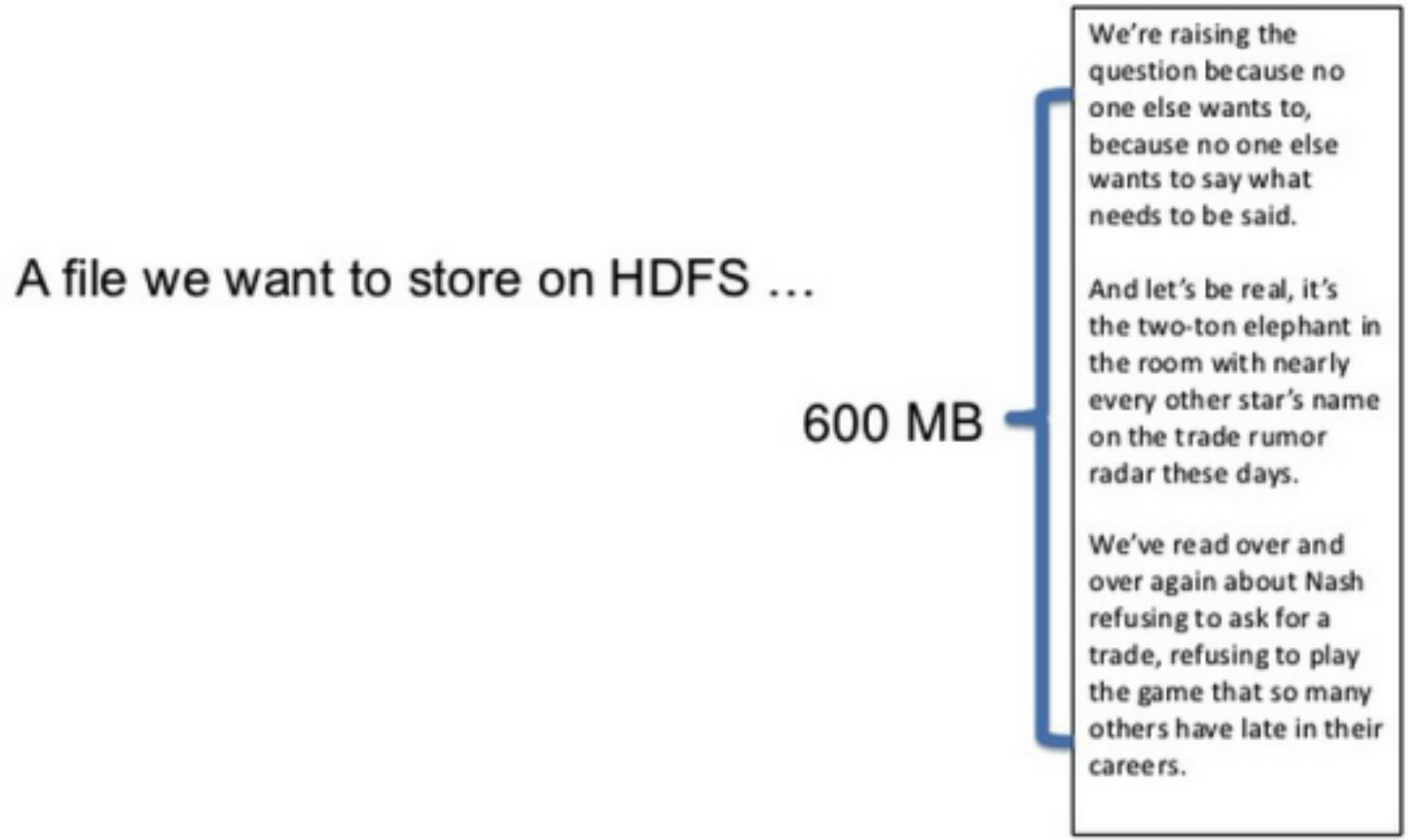


Source Hadoop: Shashwat Shriparv

Data Replication in HDFS

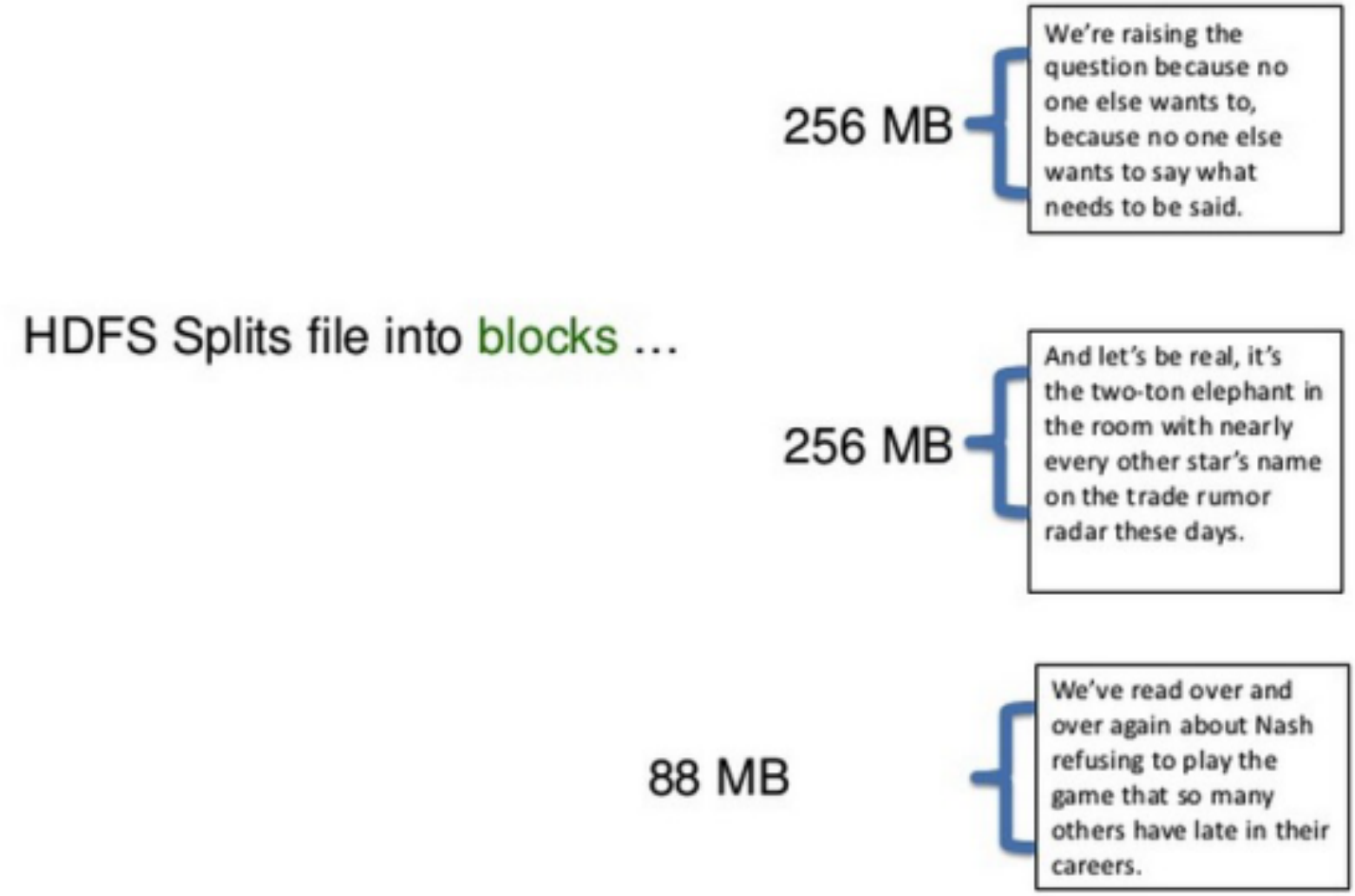


How does HDFS work?



Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

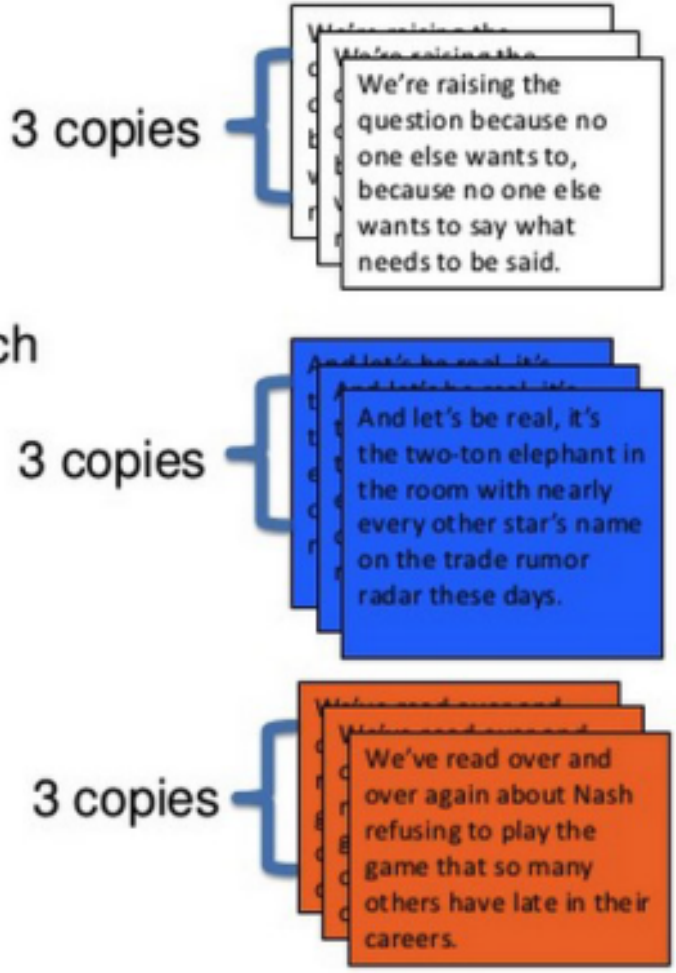
How does HDFS work?



Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

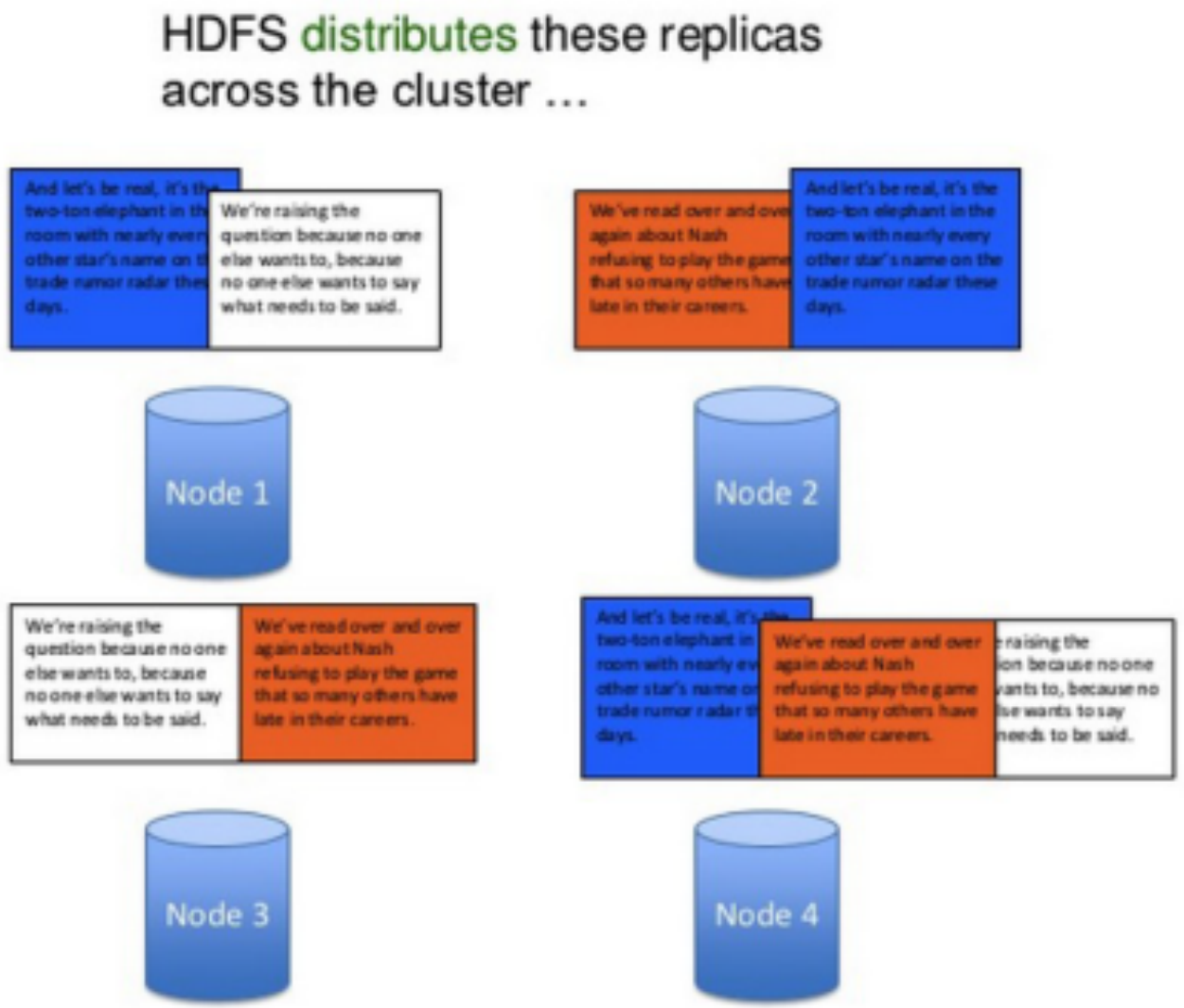
How does HDFS work?

HDFS will create 3replicas of each block …



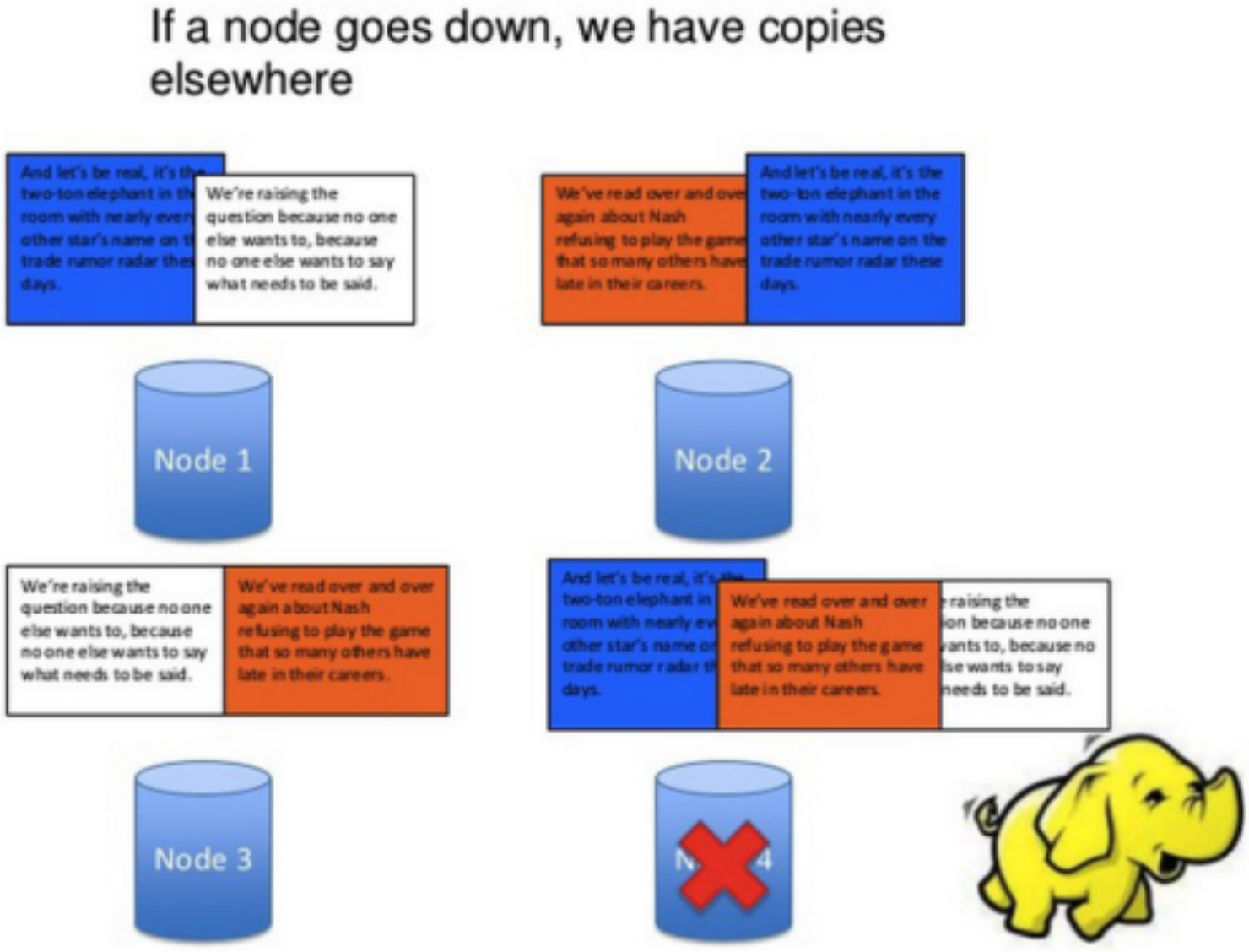
Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

How does HDFS work?



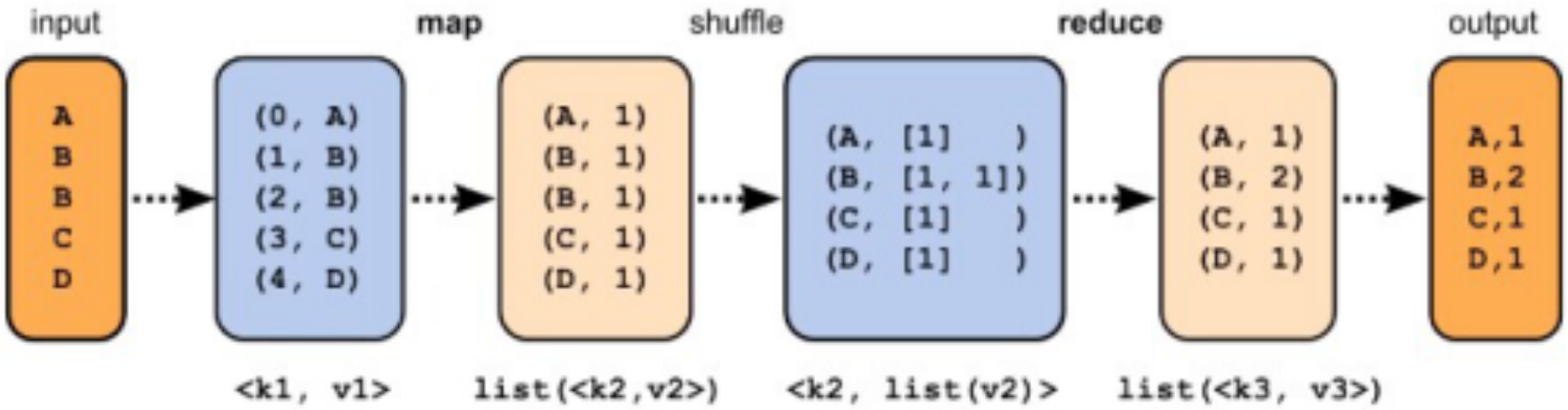
Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

How does HDFS work?



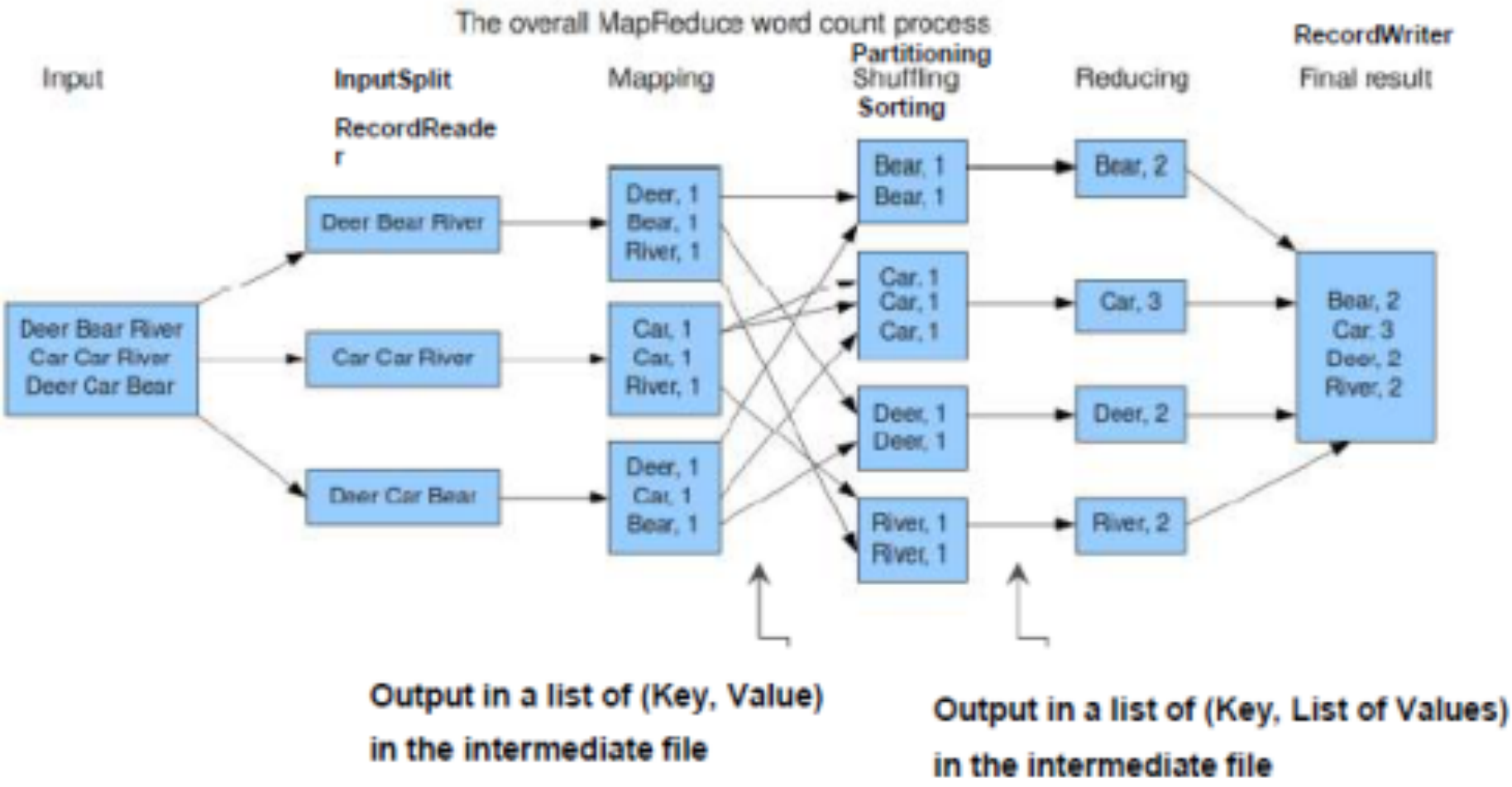
Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

MapReduce Framework

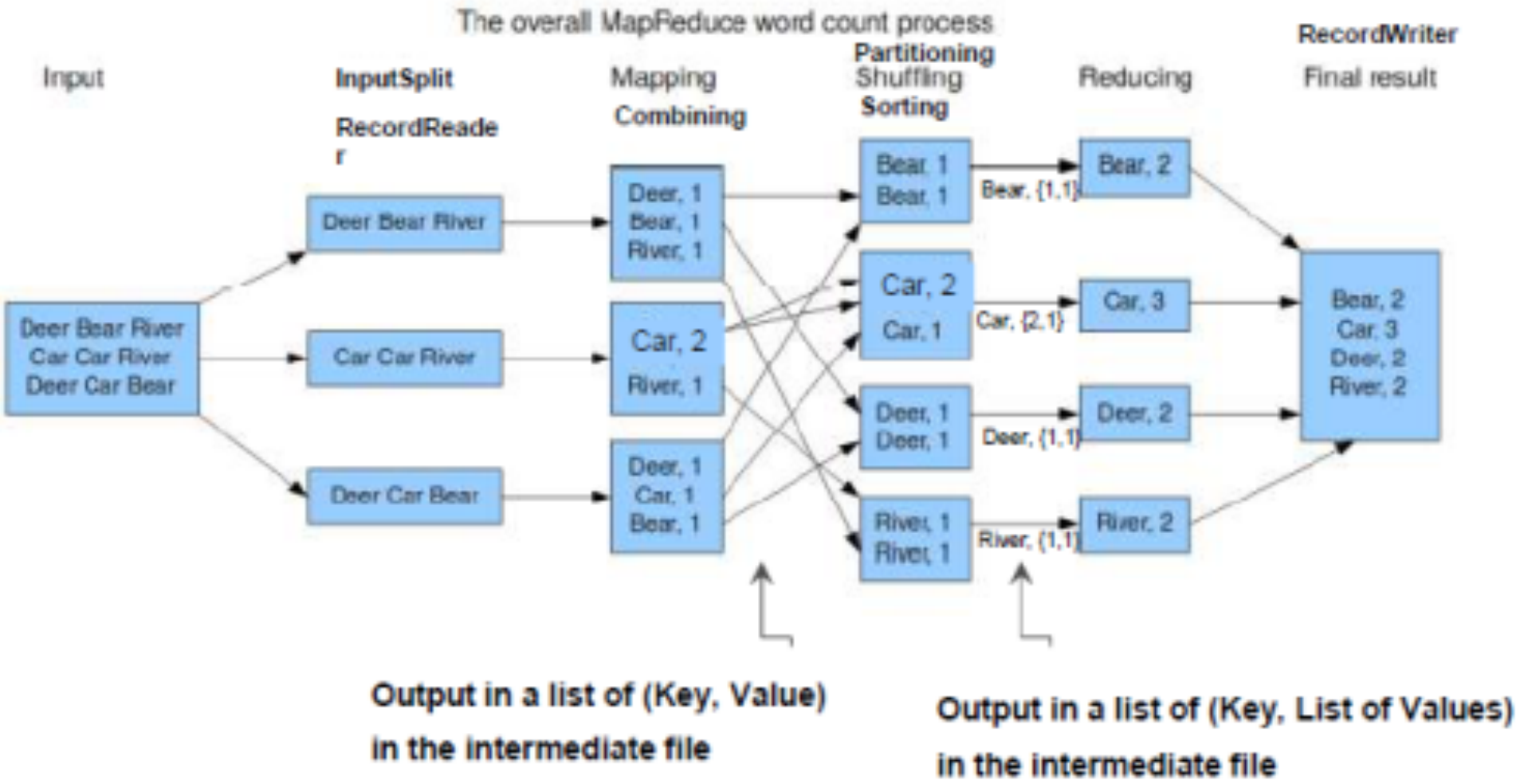


Source: www.bigdatauniversity.com

How does the MapReduce work ?



How does the MapReduce work ?



WordCount - MapReduce

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

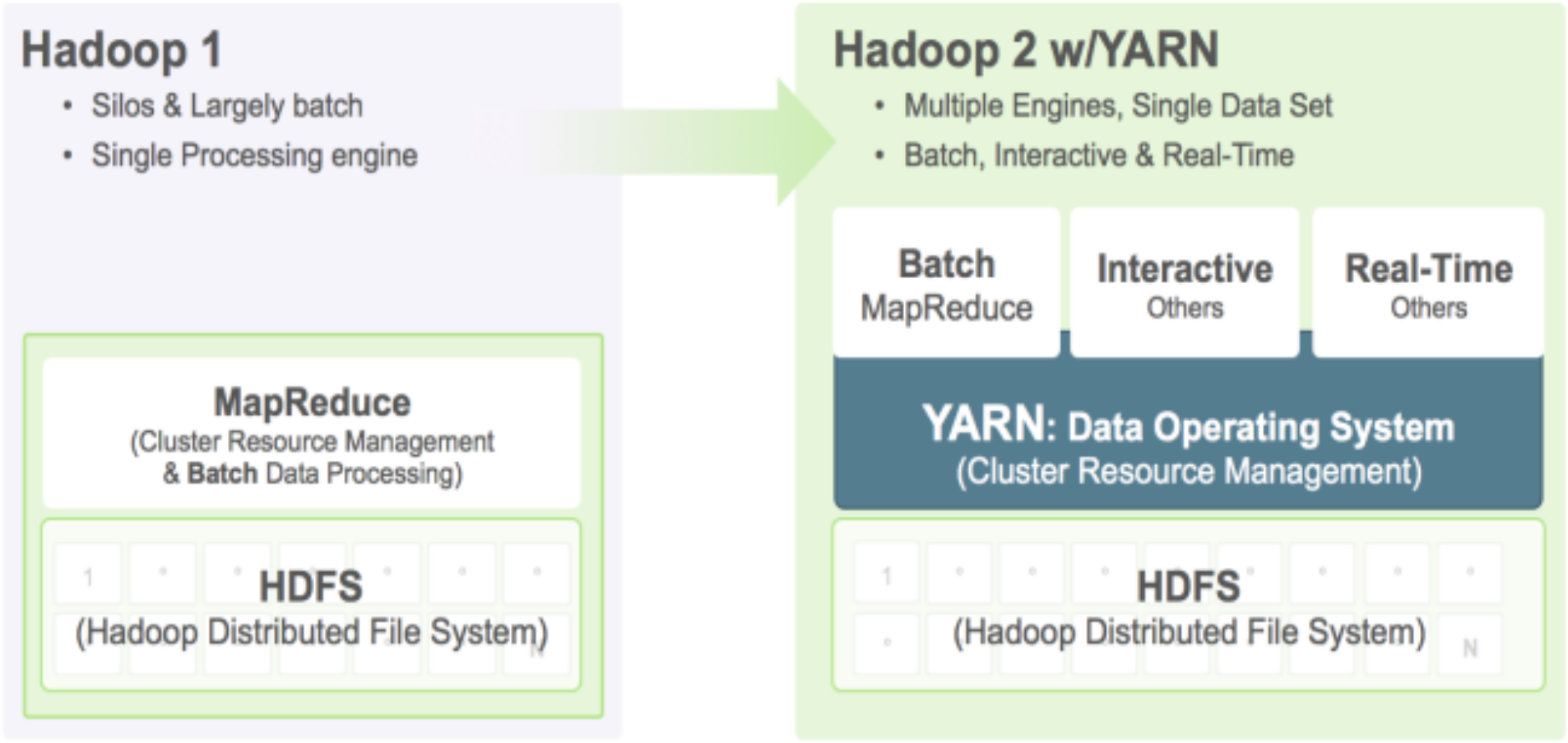
Map Function

```
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);
  }}}
```

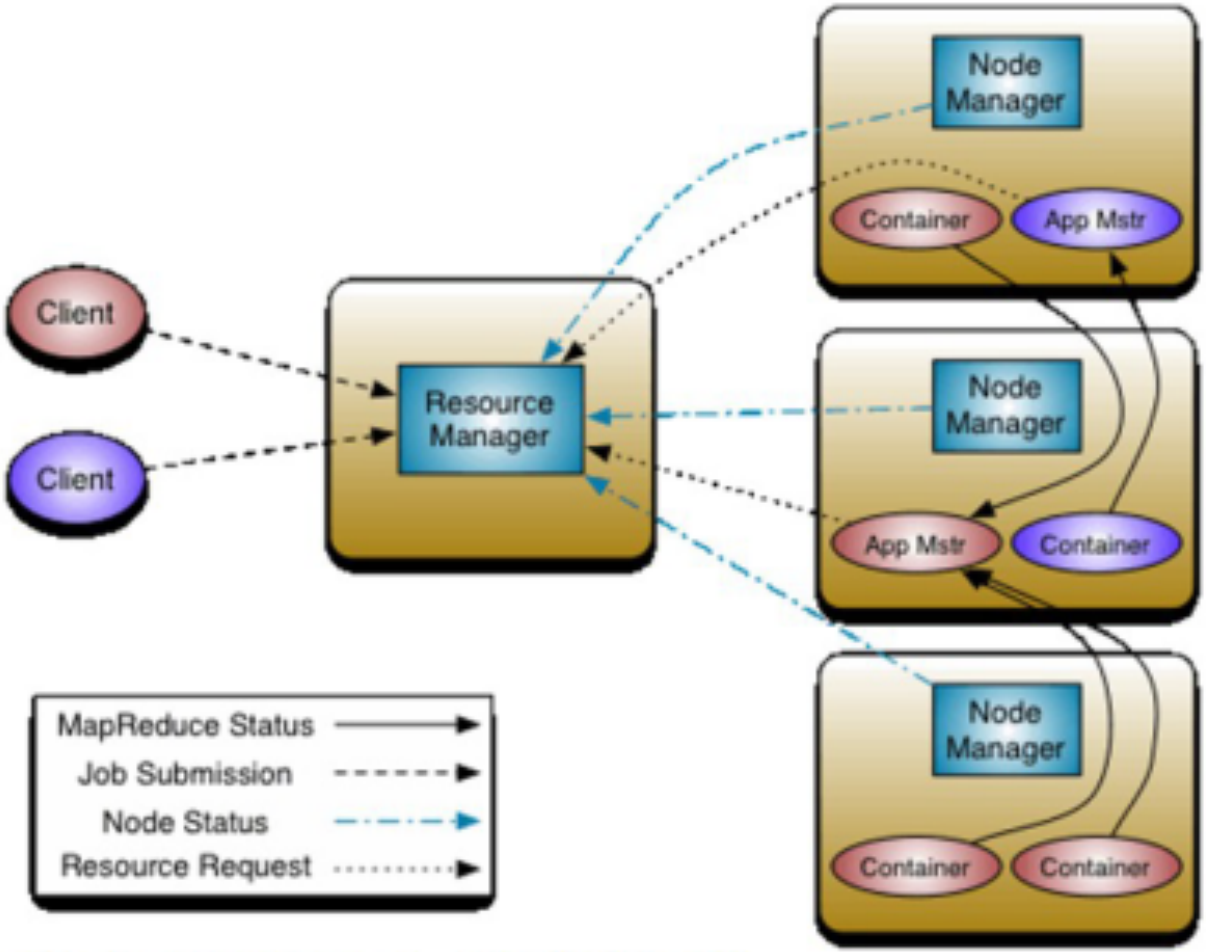
Reduce Function

```
  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

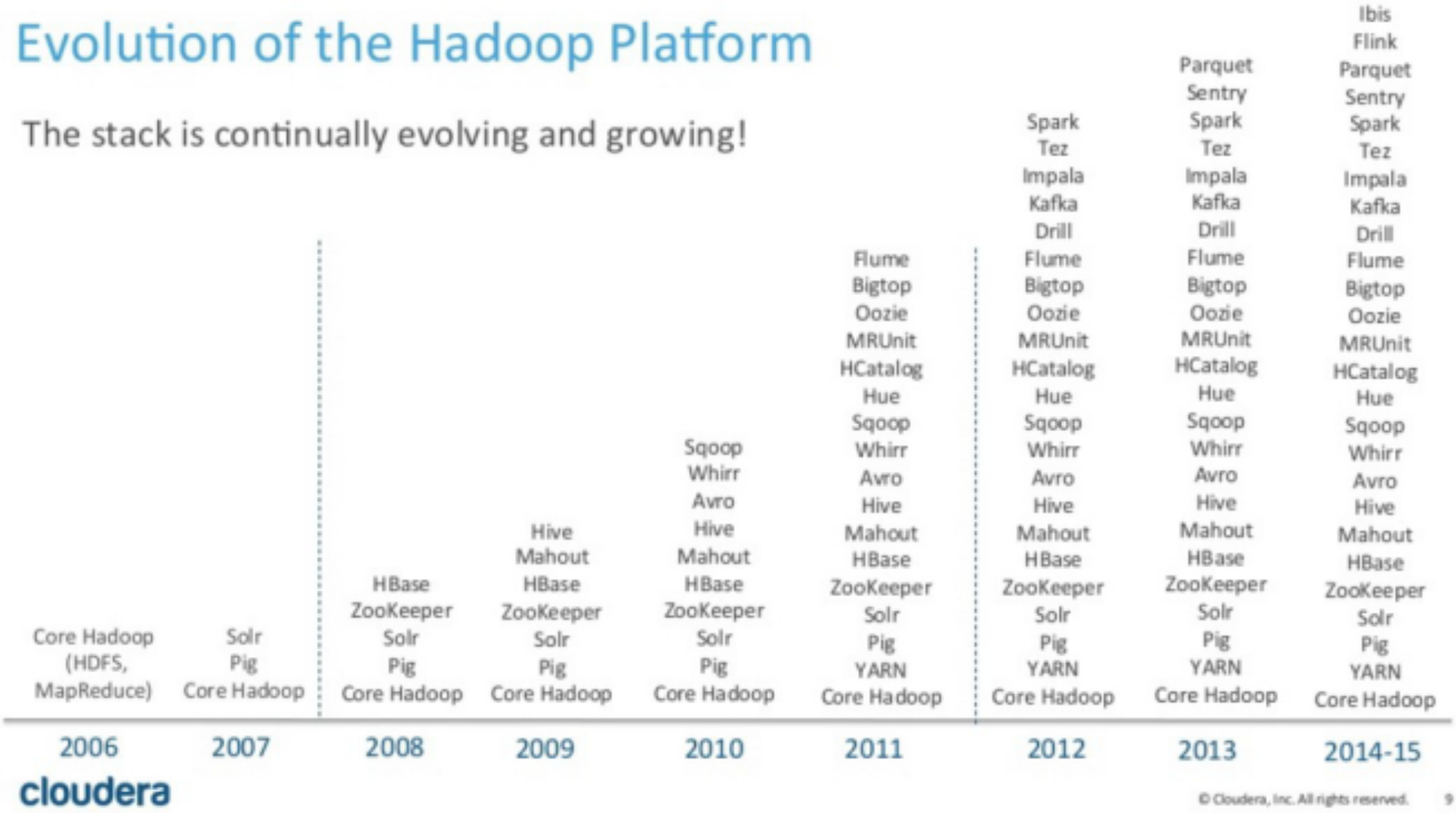


MRV2 maintains API compatibility with previous stable release (hadoop-1.x). This means that all Map-Reduce jobs should still run unchanged on top of MRv2 with just a recompile.

Hadoop.apache.org

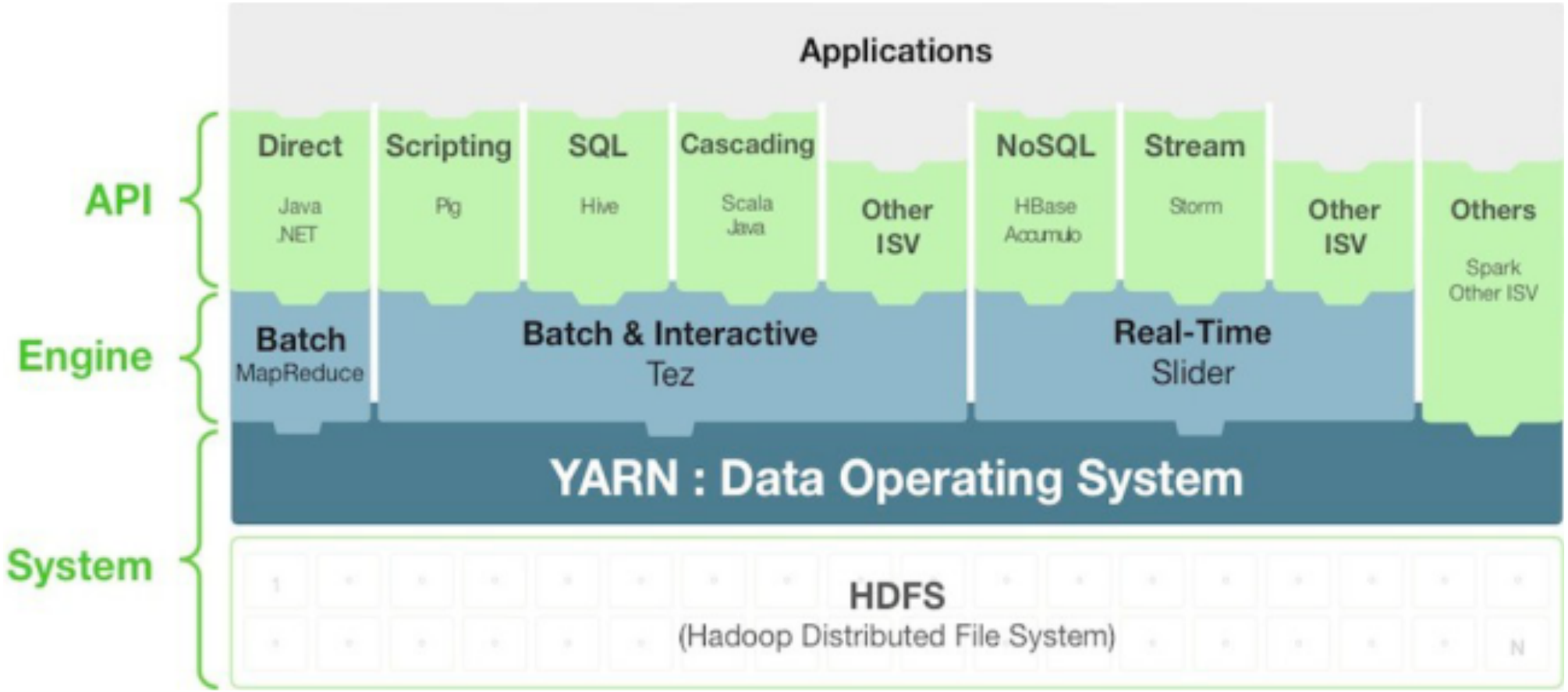
Evolution of the Hadoop Platform

The stack is continually evolving and growing!



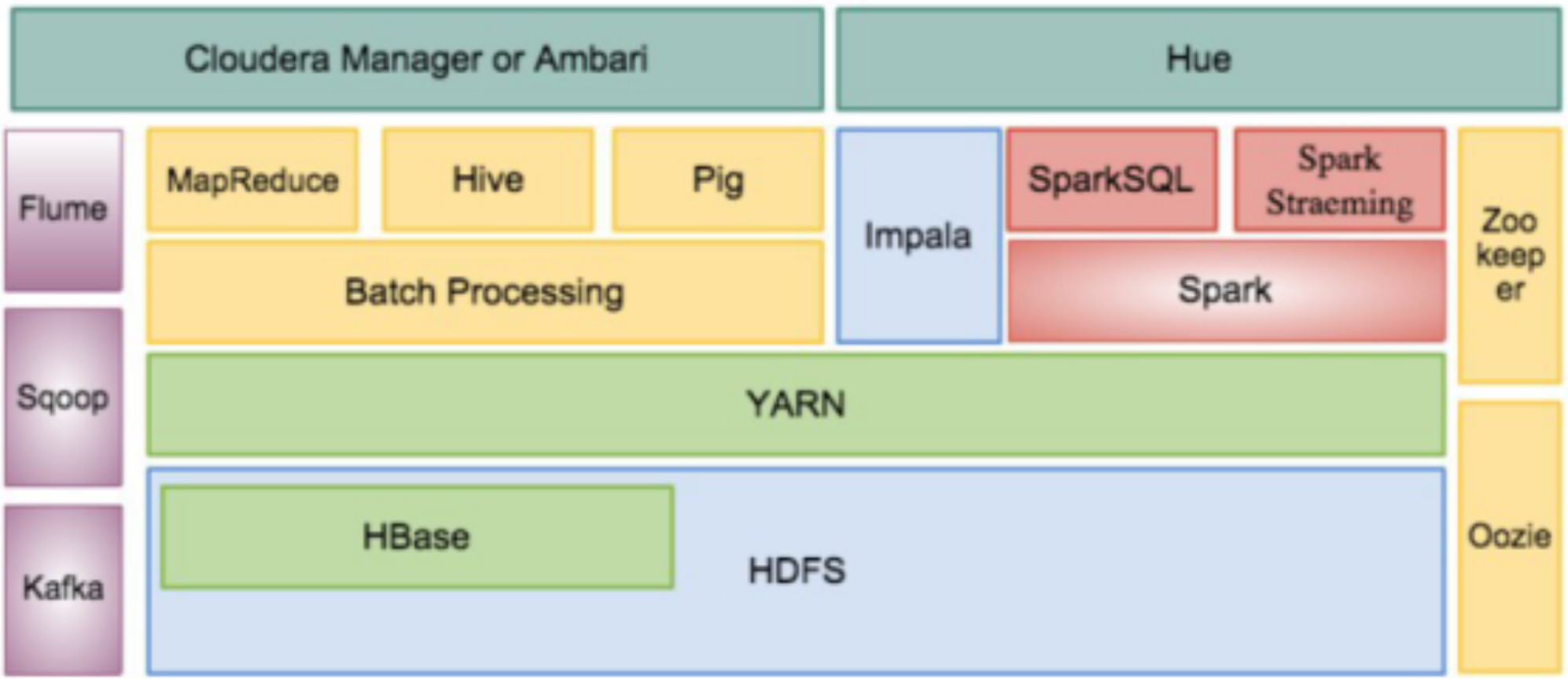
Source: The evolution and future of Hadoop storage: Cloudera

Hadoop 2.x Ecosystems



Hadoop Ecosystems

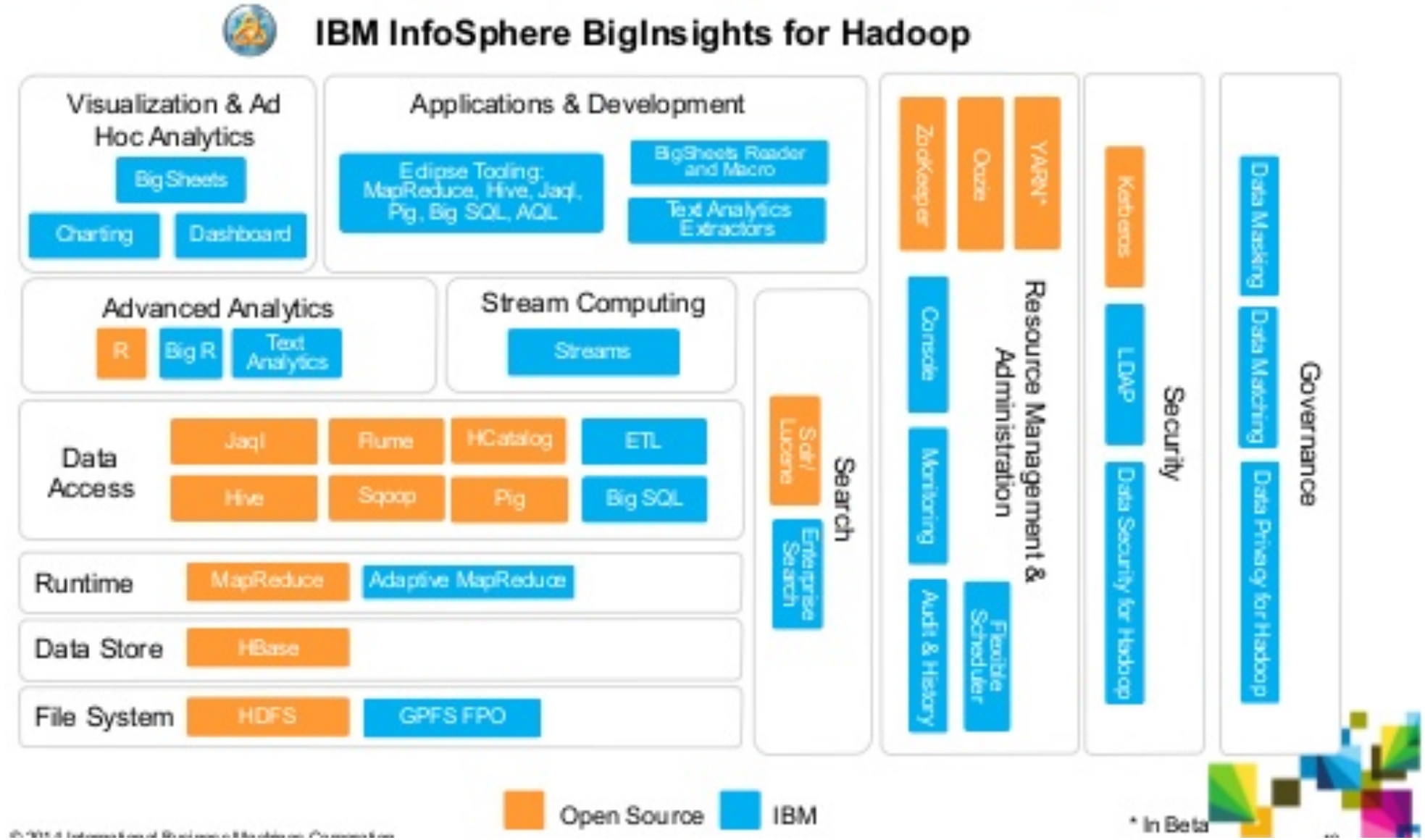
CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



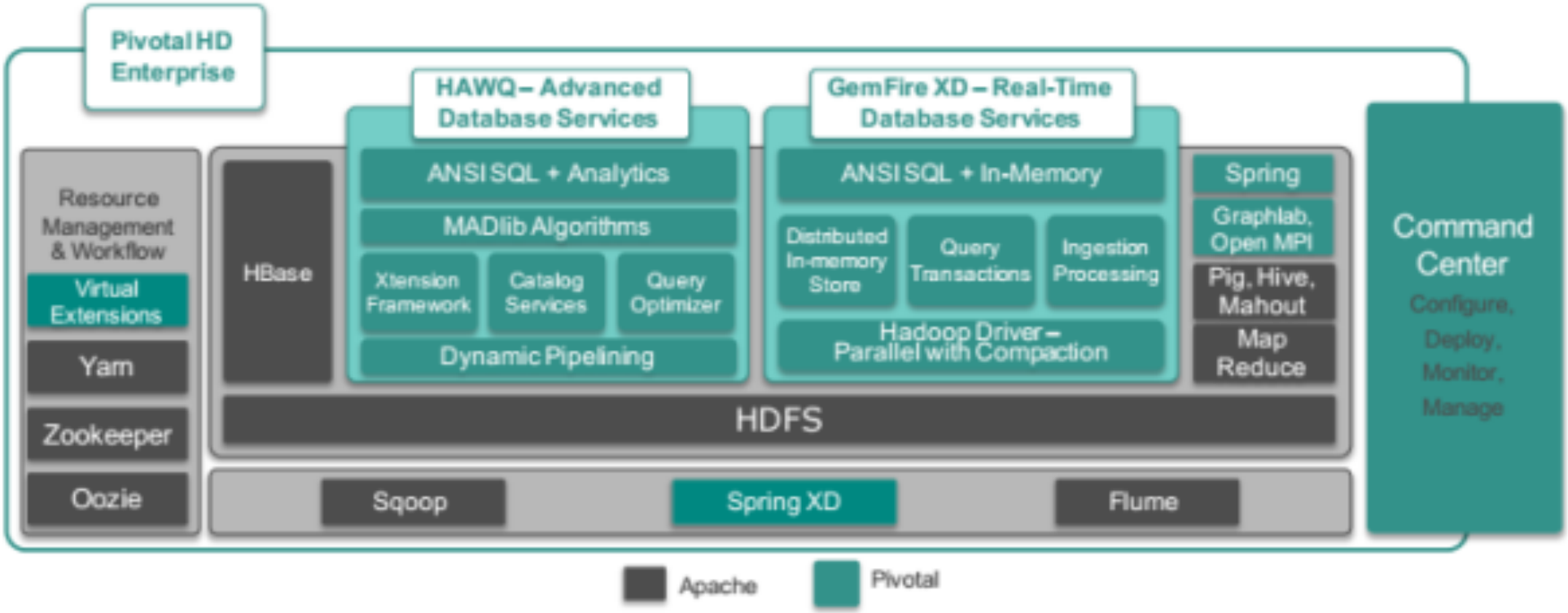
Hadoop Distribution

IBM InfoSphere BigInsights

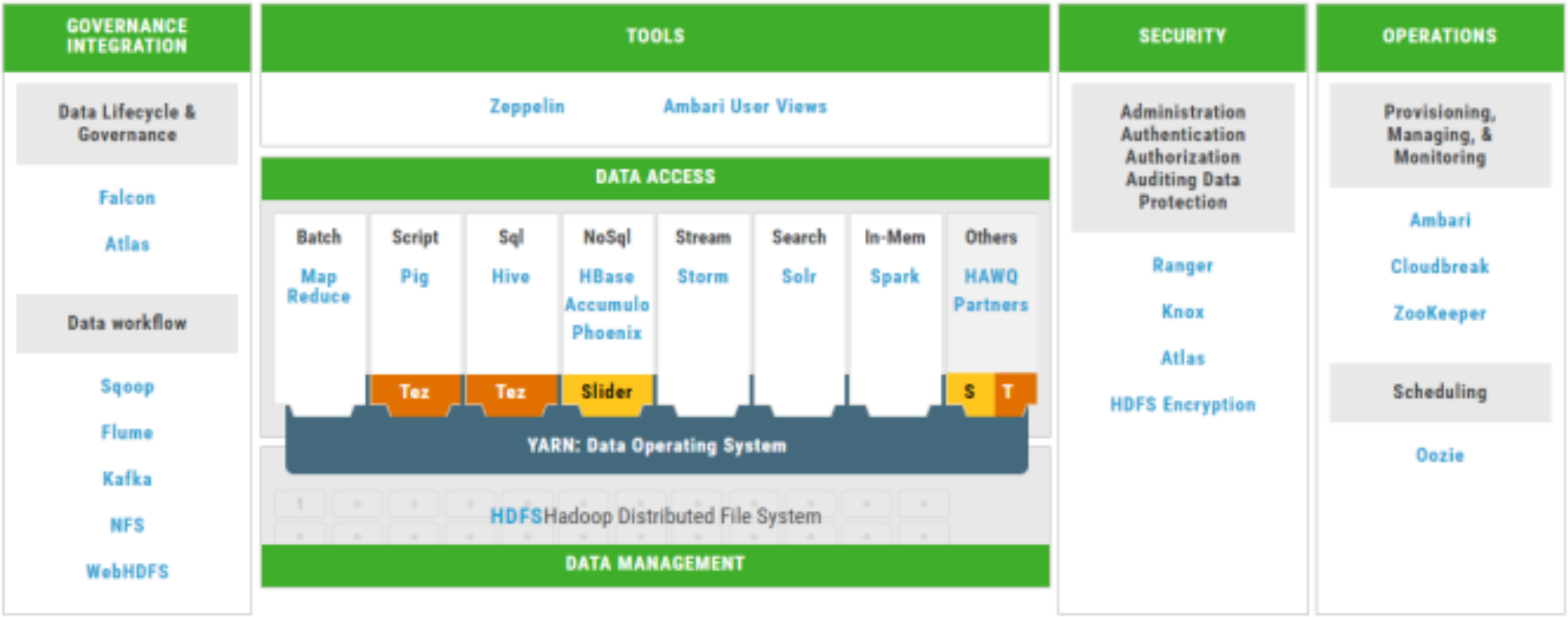
CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Pivotal HD Architecture

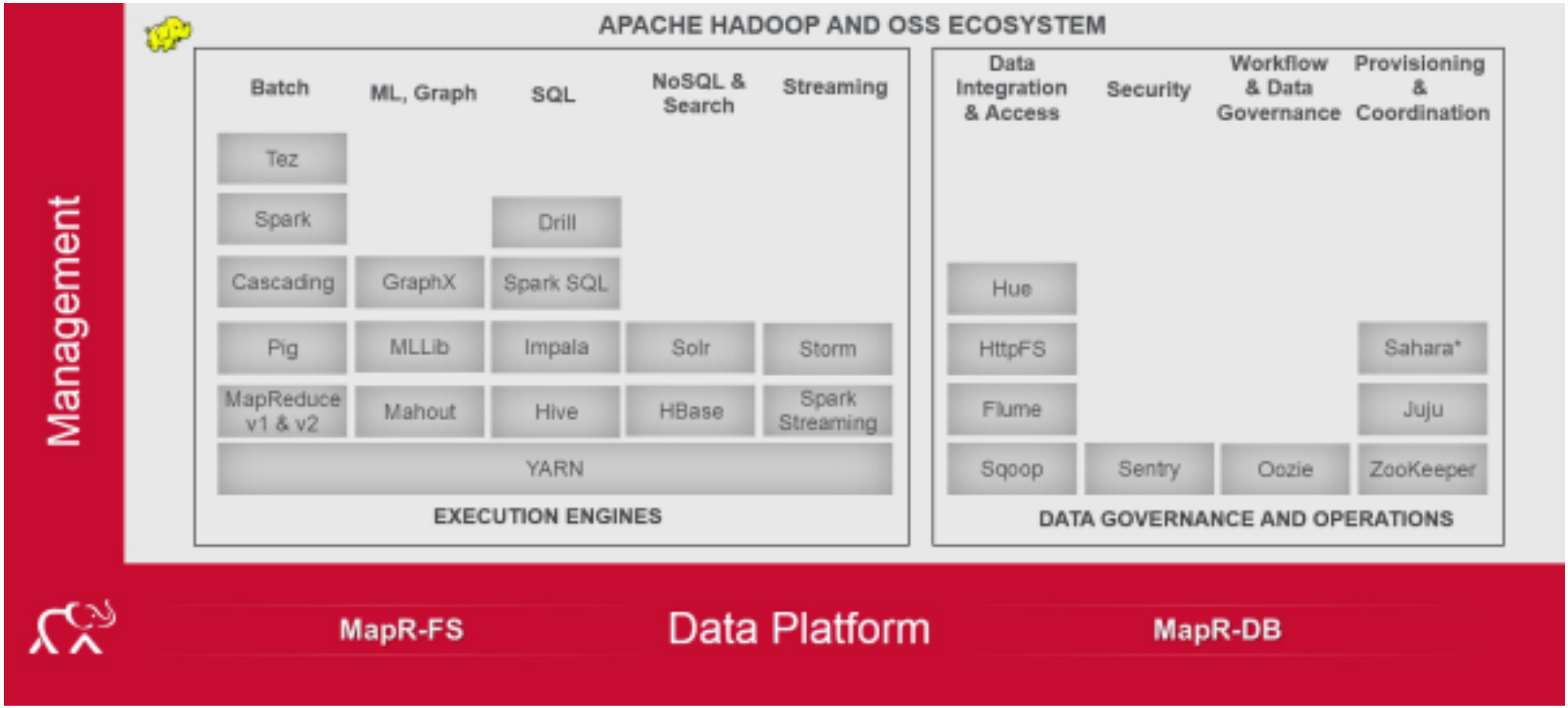


Hortonworks



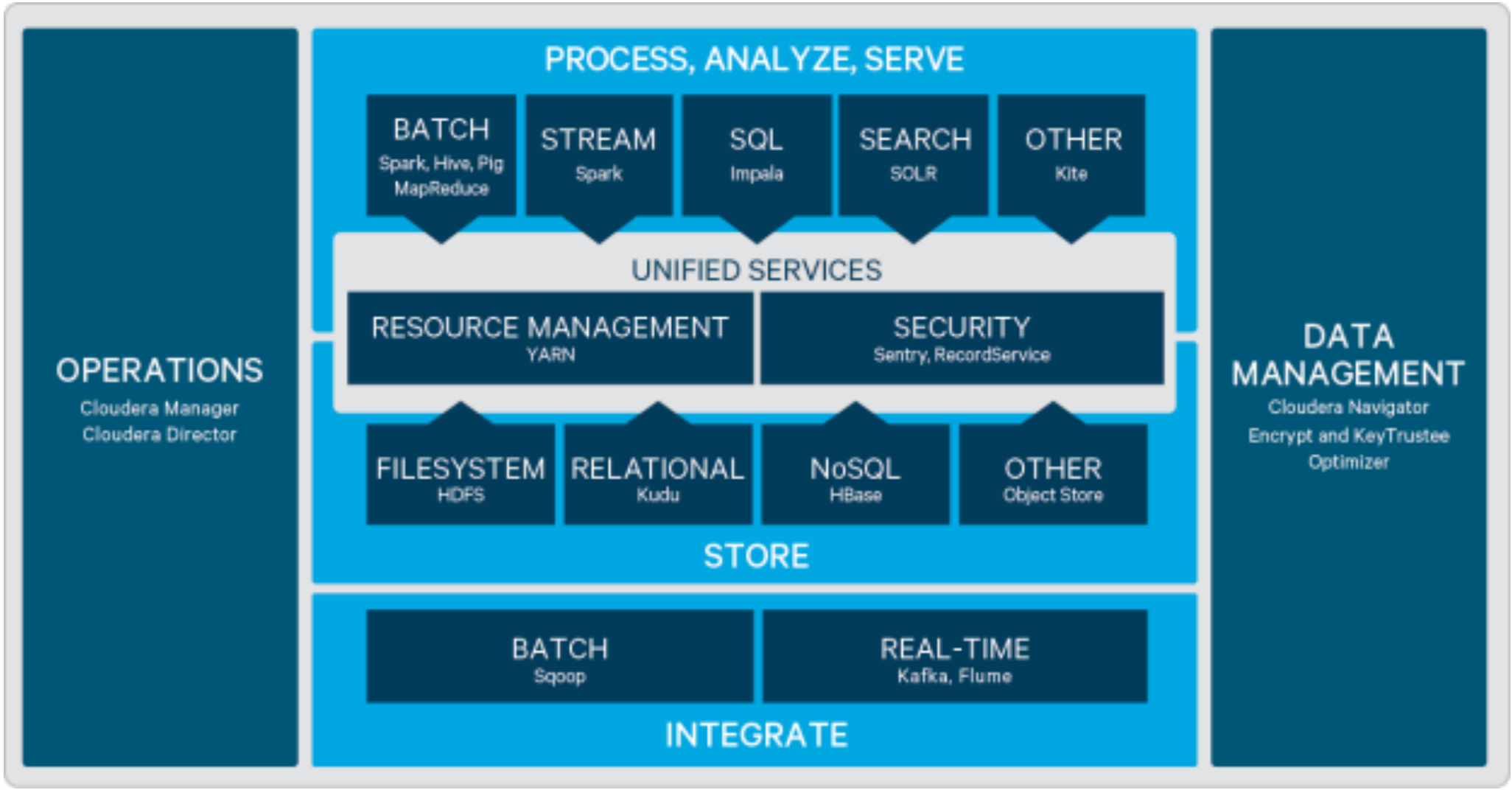
source : http://hortonworks.com/products/data-center/hdp/

MapR


Management
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
MapR-FS
Data Platform
MapR-DB

Cloudera

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

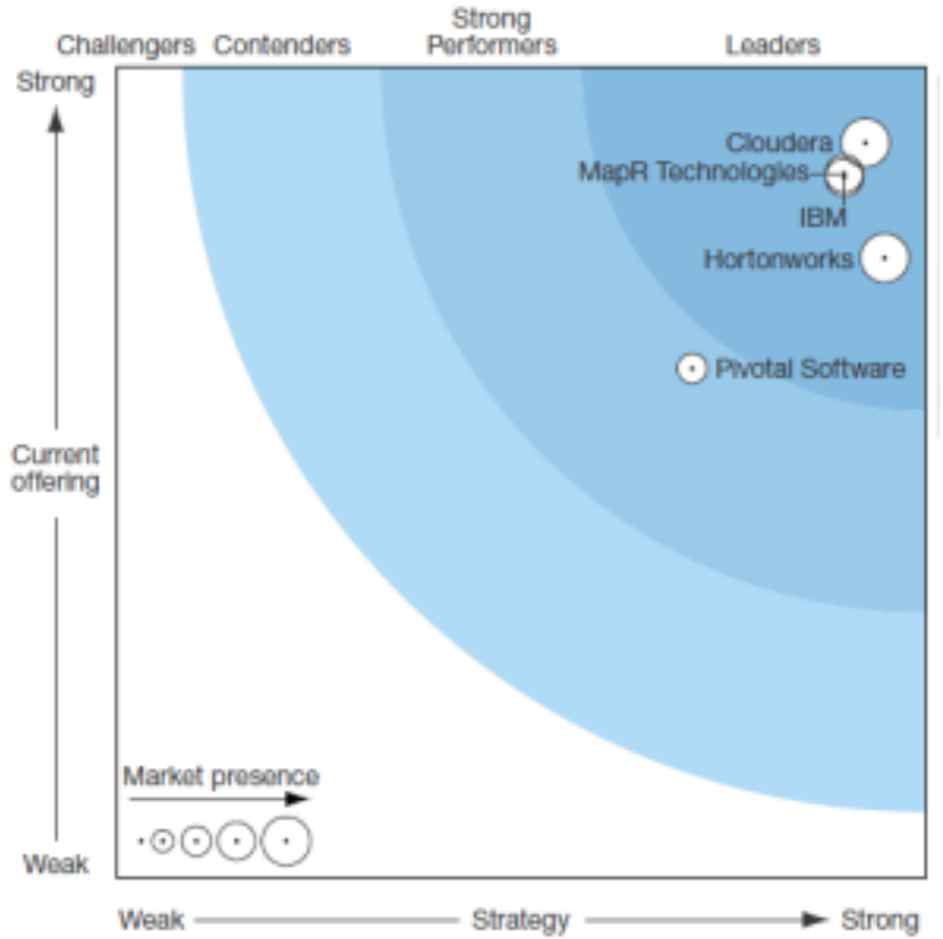


The Forrester Wave™

Big Data Hadoop Distributions Q1 2016

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

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

Big Data Using Public Cloud

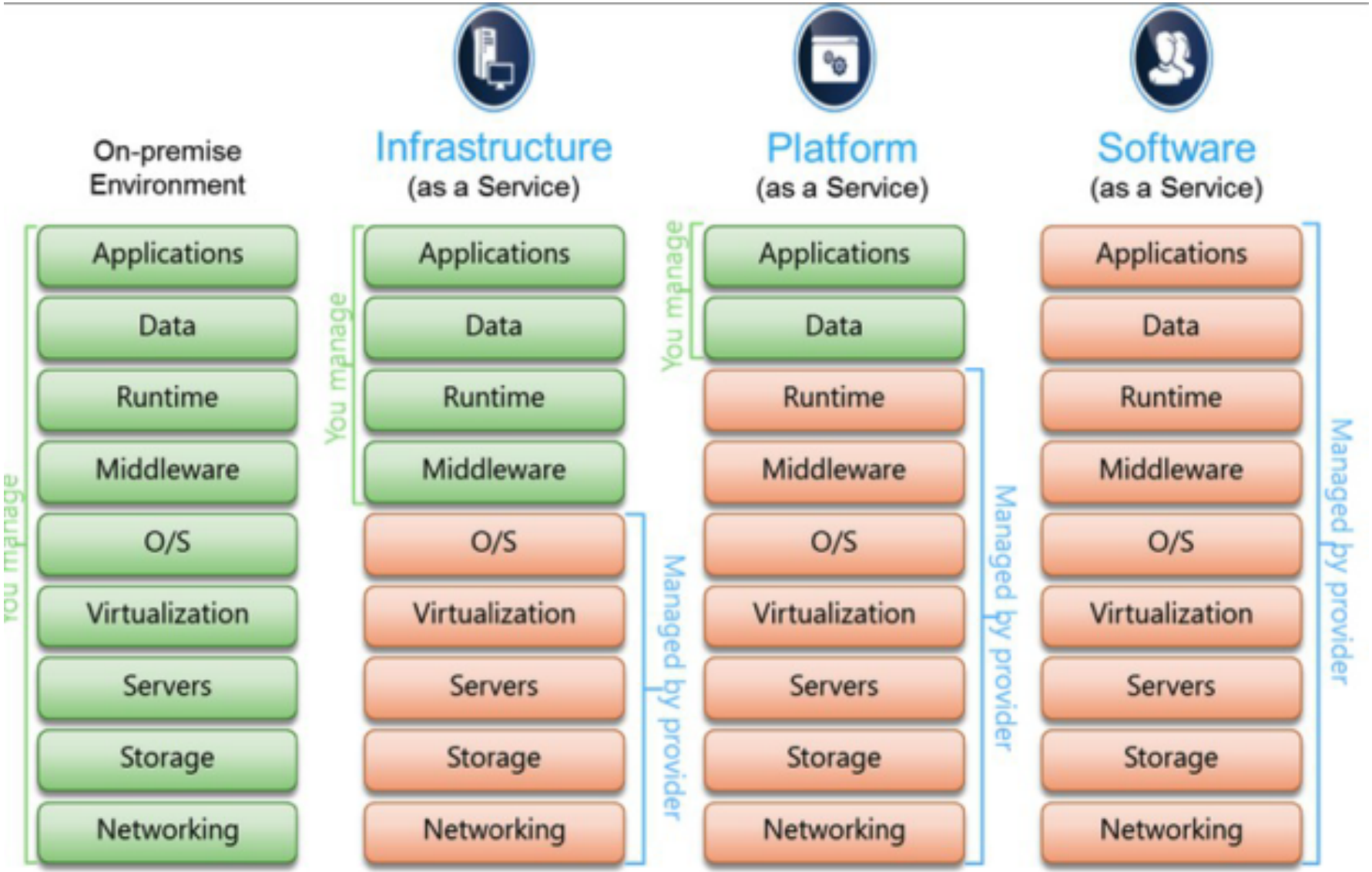
CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Issue with Big Data Infrastructure

- Large investment
- Scalability
- ROI
- Business Cases

Cloud Technology

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

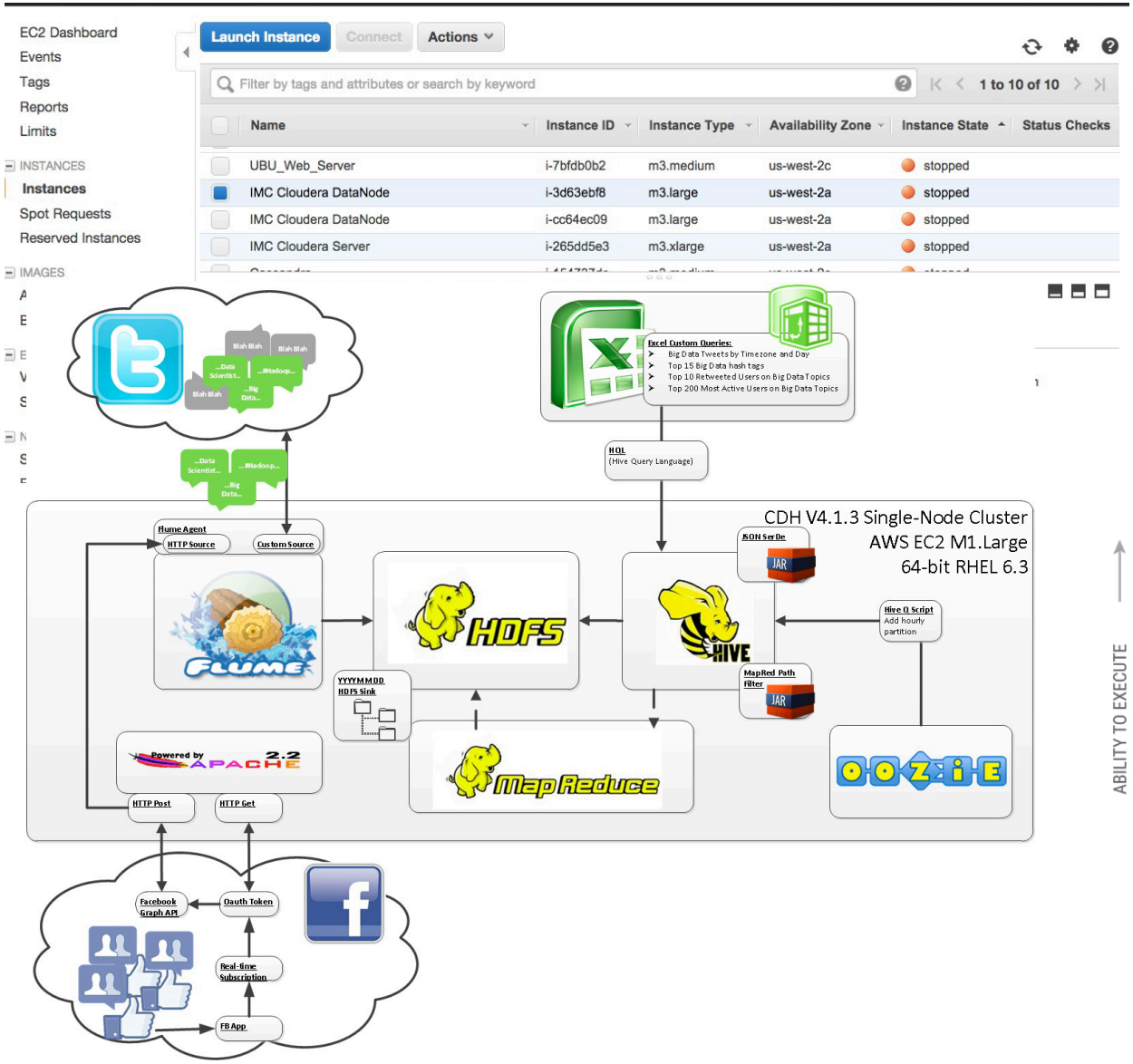


http://www.markerbench.com/blog/2013/01/08/static-blogging/

Big Data on Cloud

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Big Data using IaaS



Thaveewat Khanan

2603615 Organization Data Management

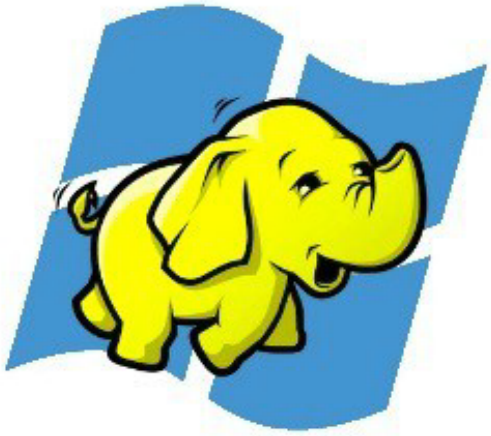
Big Data on Cloud

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Using Big Data as a Services

Amazon

Elastic Mapreduce



Microsoft Azure Hadoop