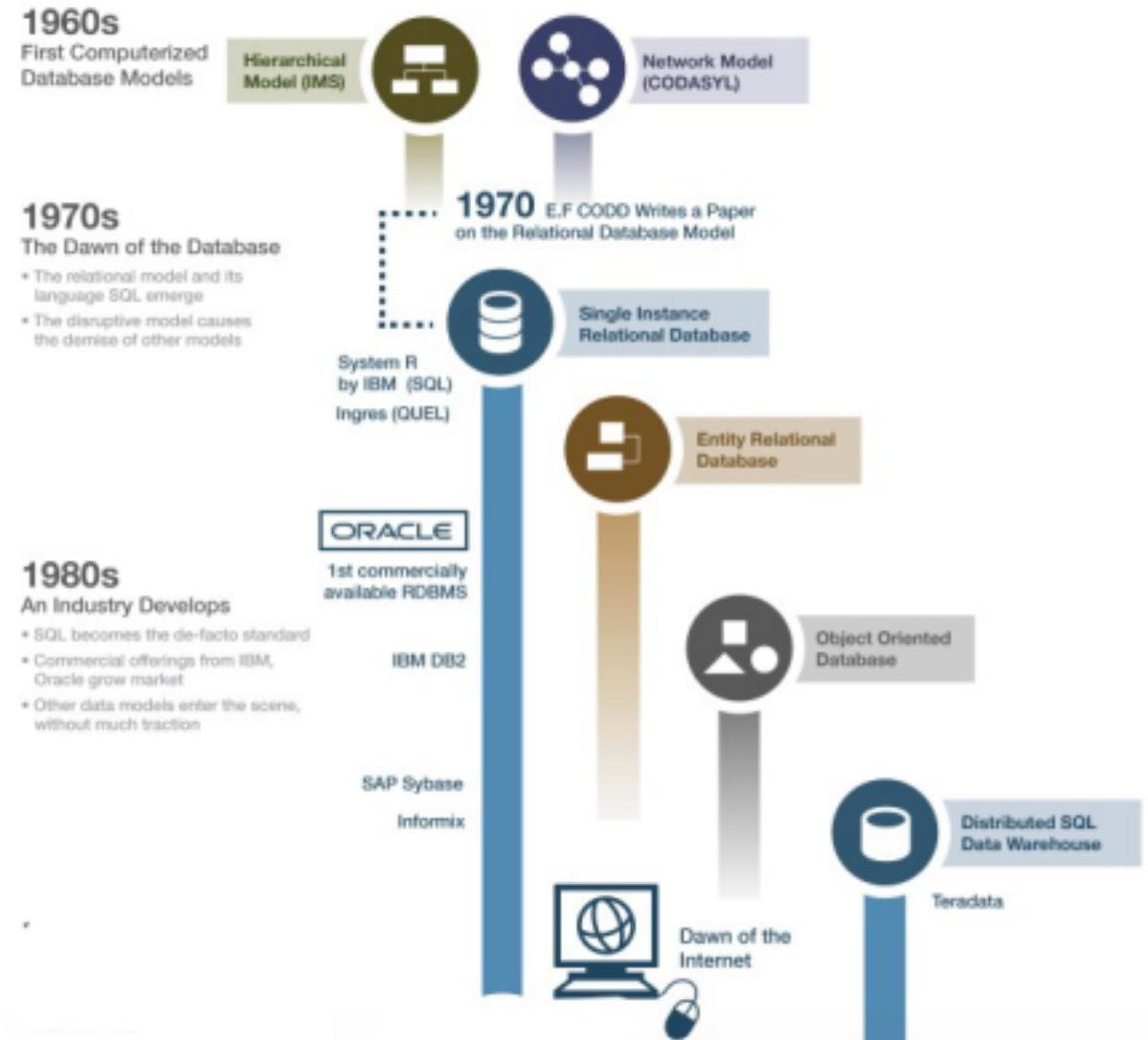
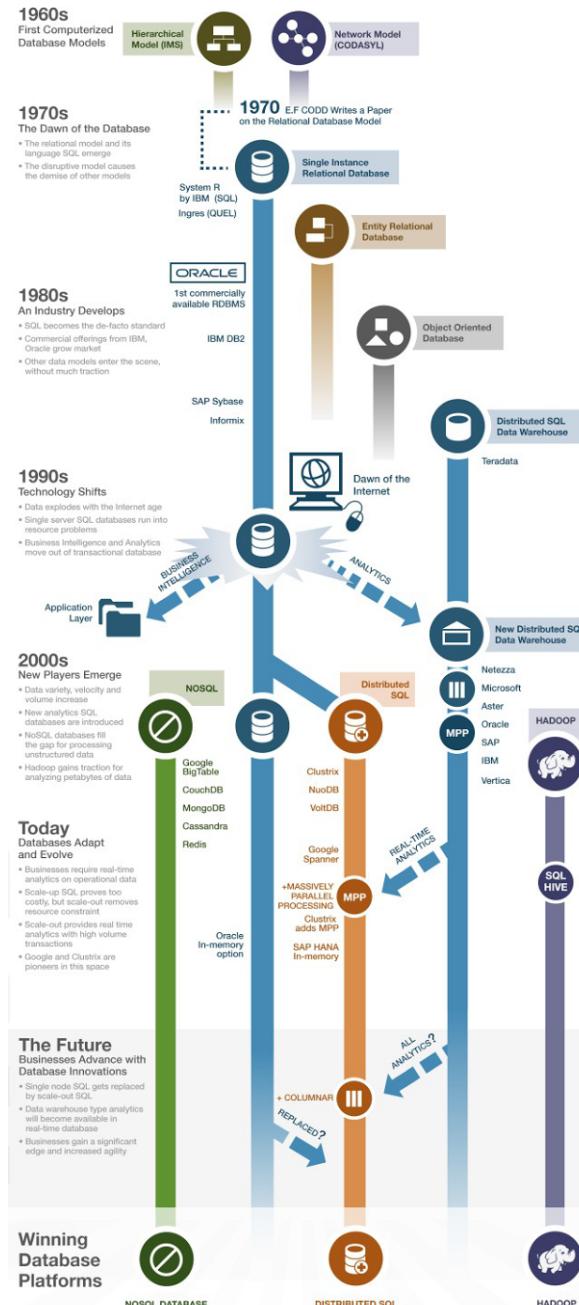
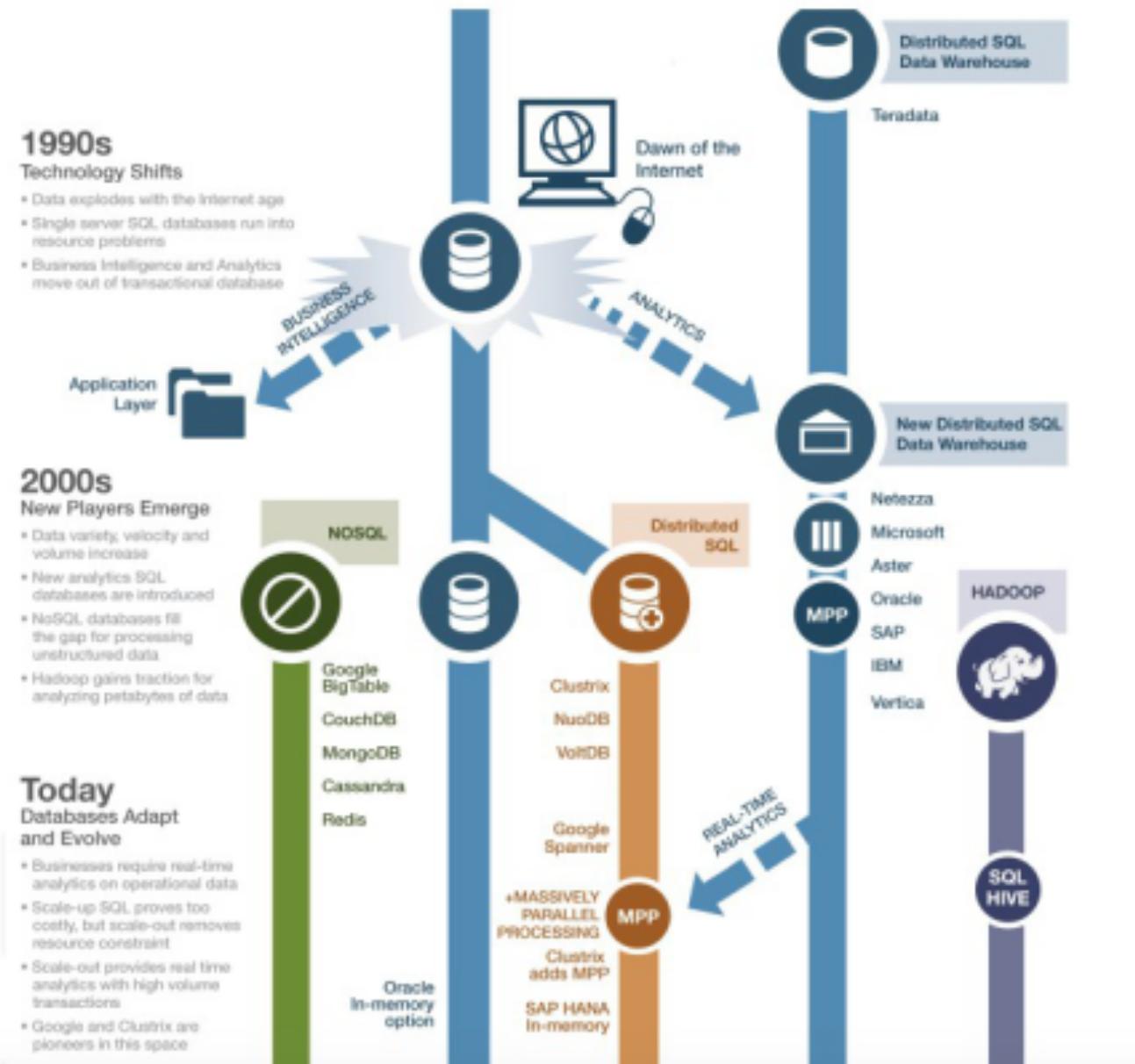
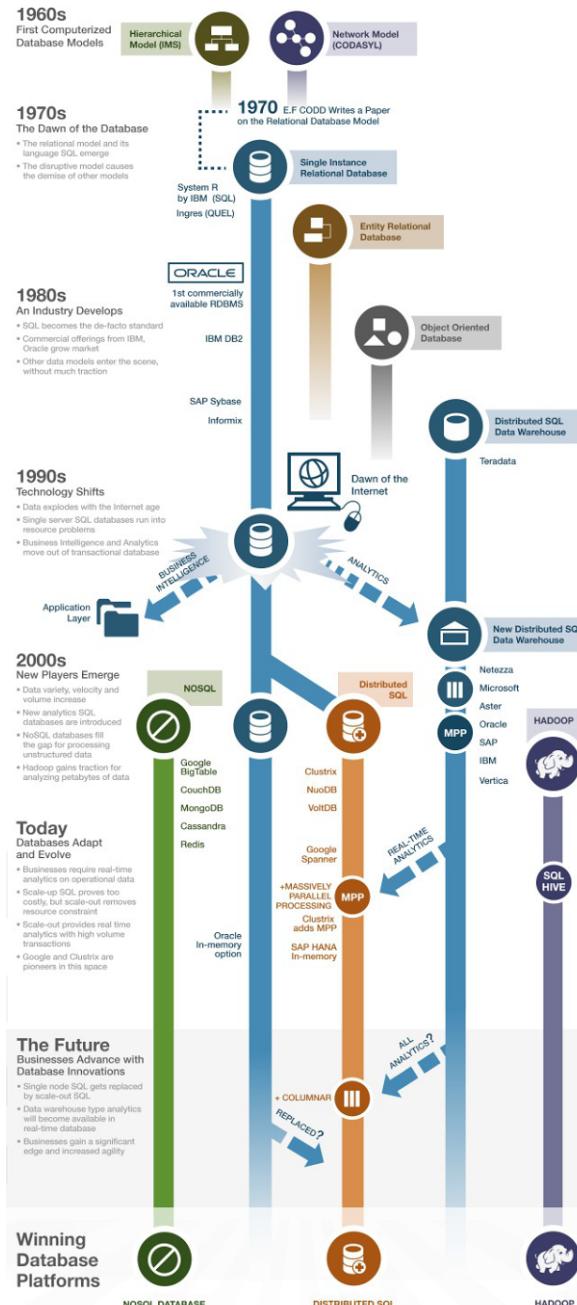


Big Data Technology

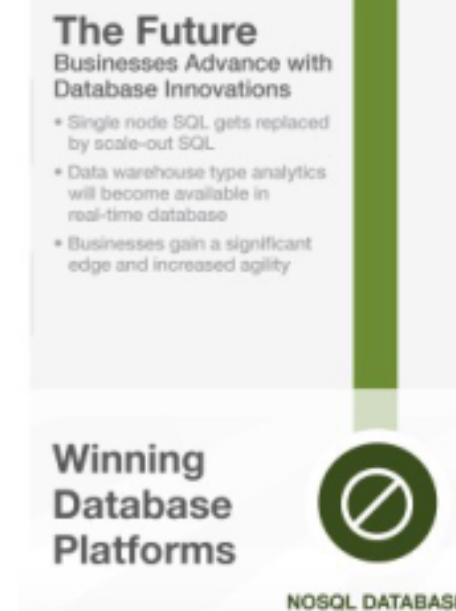
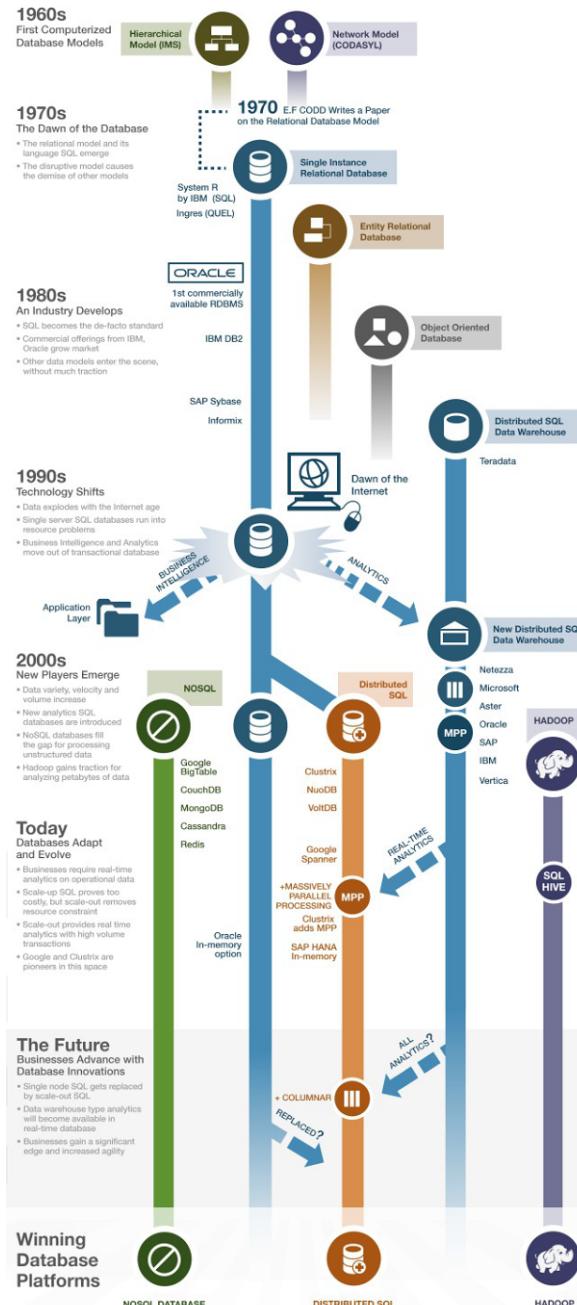
The Future of The DATABASE



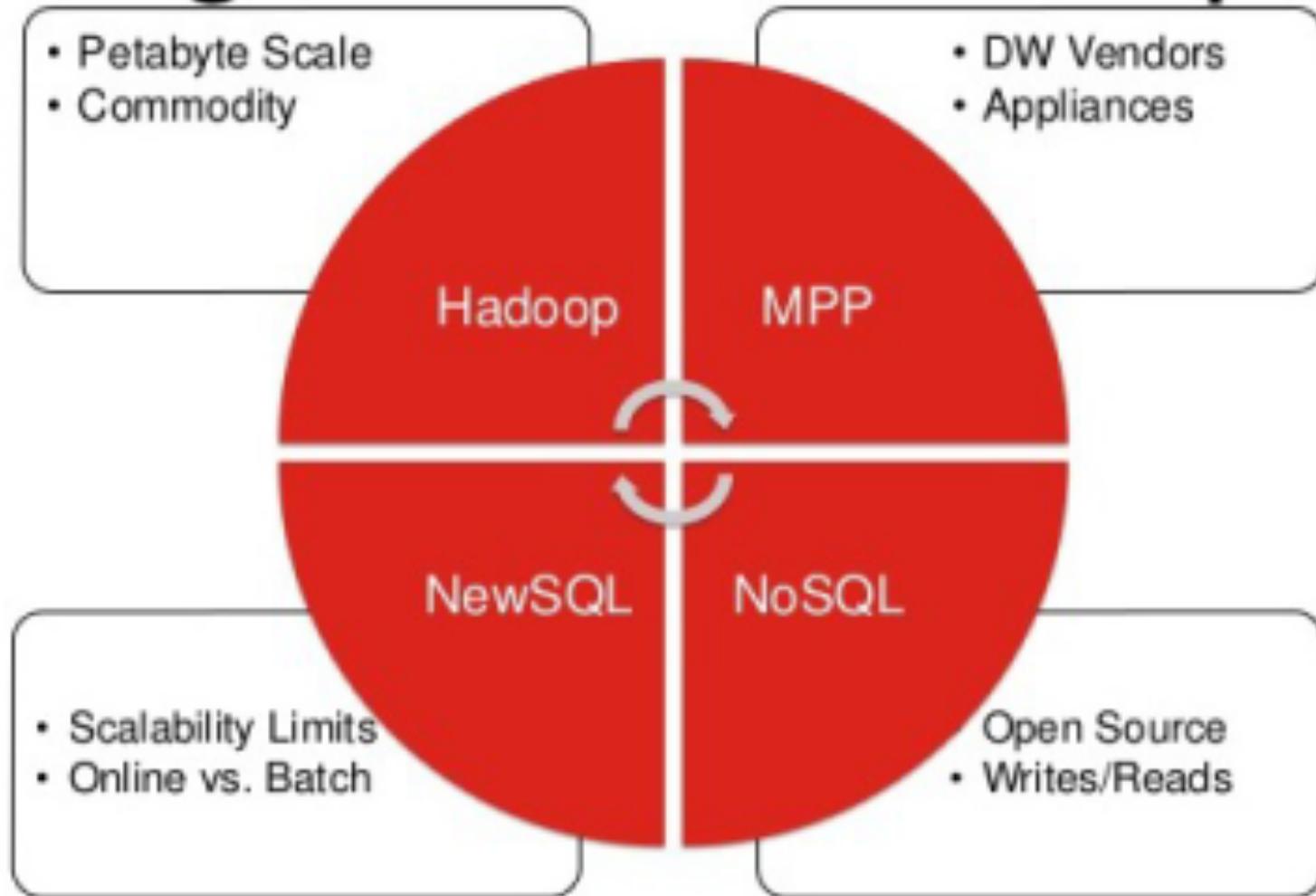
The Future of The DATABASE



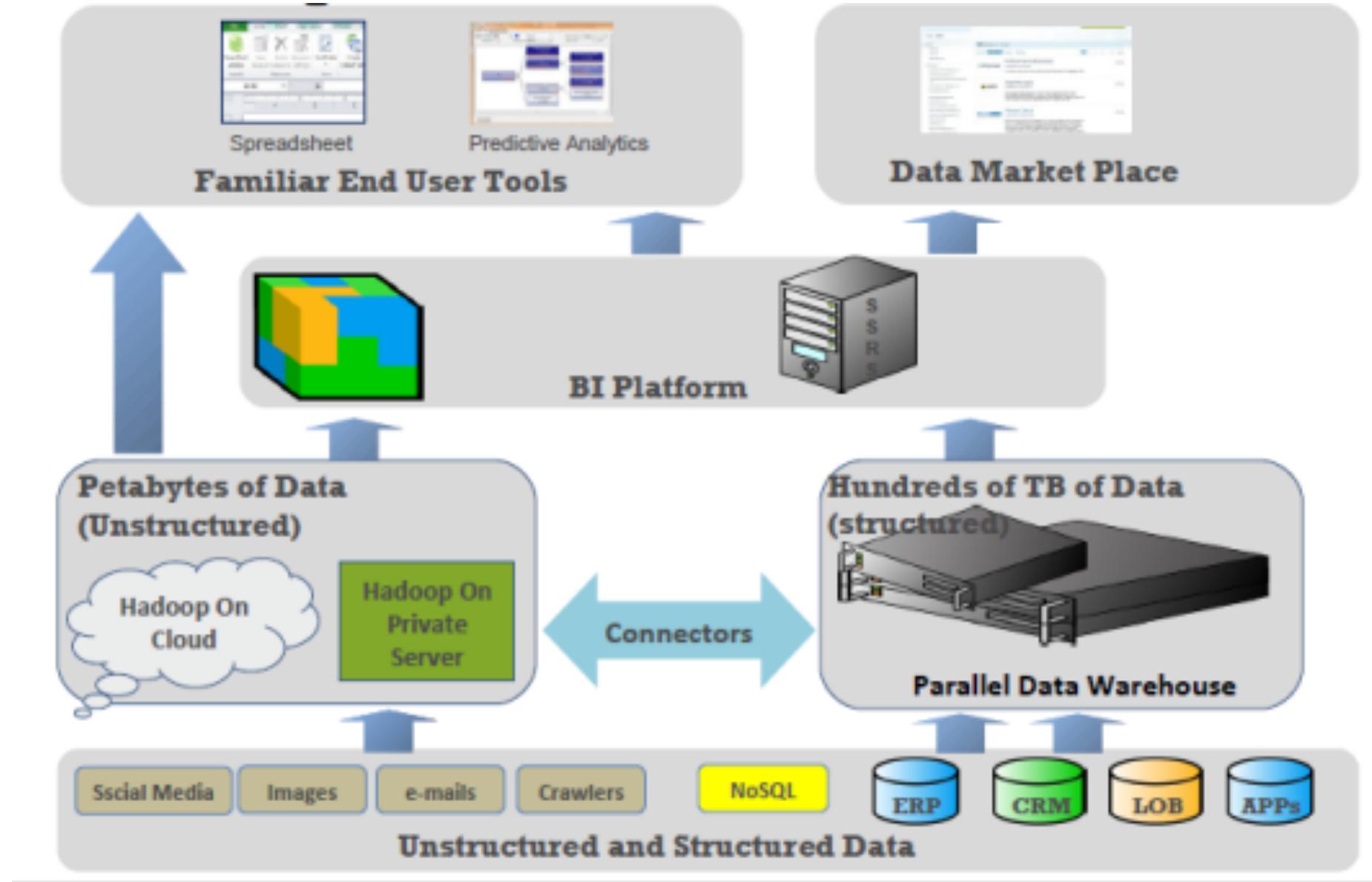
The Future of The DATABASE



Big Data Landscape



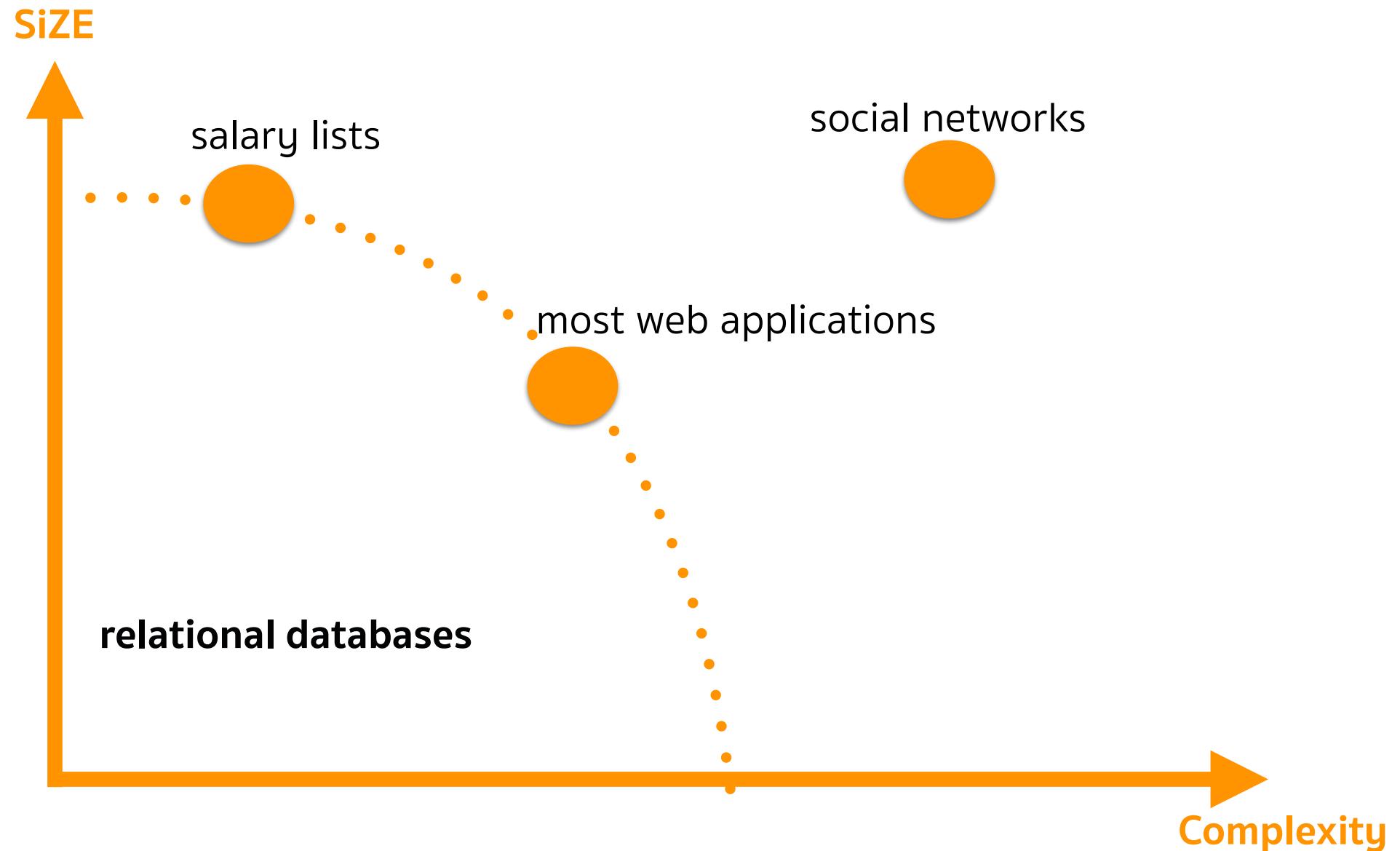
Big Data Future Architecture



What is NoSQL ?

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.



NoSQL PROS AND CONS

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

Column-oriented

Key Value Store

Document Store

Graph

Column-oriented databases

Row Oriented
(RDBMS Model)

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

A callout bubble points from the word "Multi-valued" to the "Interests" column of the third row, and another callout bubble points from the word "null" to the empty "Interests" cell of the second row.

Column Oriented
(Multi-value sorted map)

id	Name
1	Ricky
2	Ankur
3	Sam

id	Age
2	20
3	25

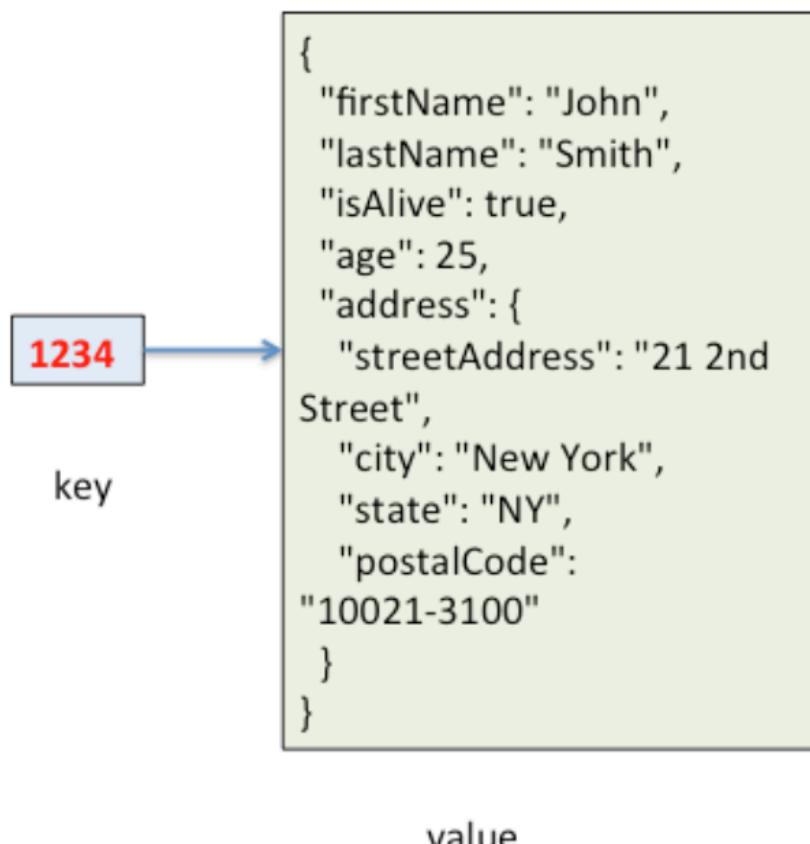
id	Interests
1	Soccer
1	Movies
1	Baseball
3	Music



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"}



redis



Document-oriented database

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

Most of the databases available under this category use

XML, JSON, BSON, or YAML

```
{  
    "EmployeeID": "SML",  
    "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



Author name

Comment Table
Reader Table

Article Table
Author Table

Relational Database approach
Document store approach

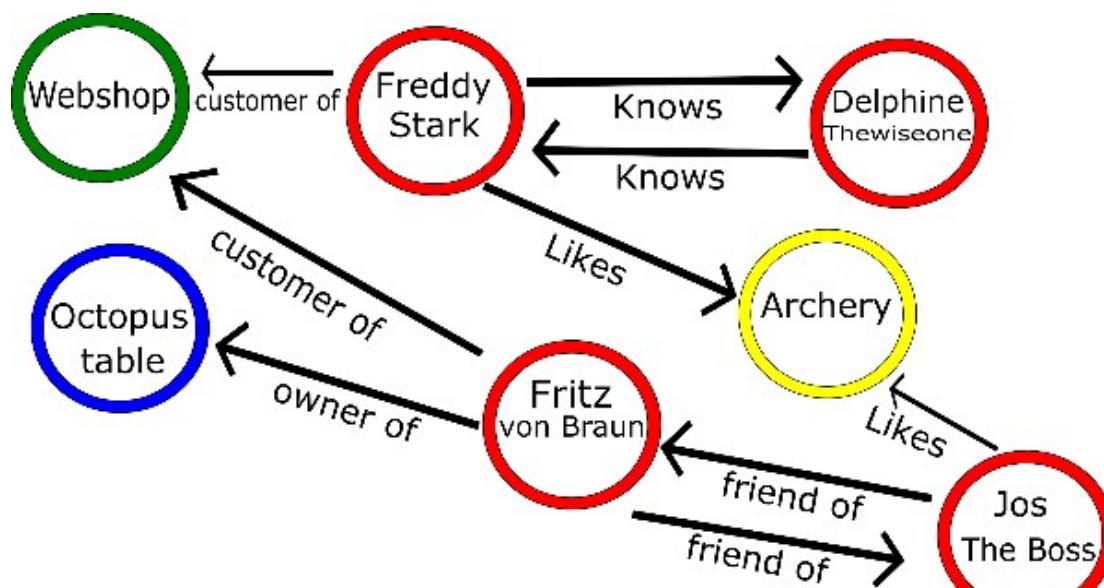
Whereas relational databases chop up data, Document stores save documents as a single entity

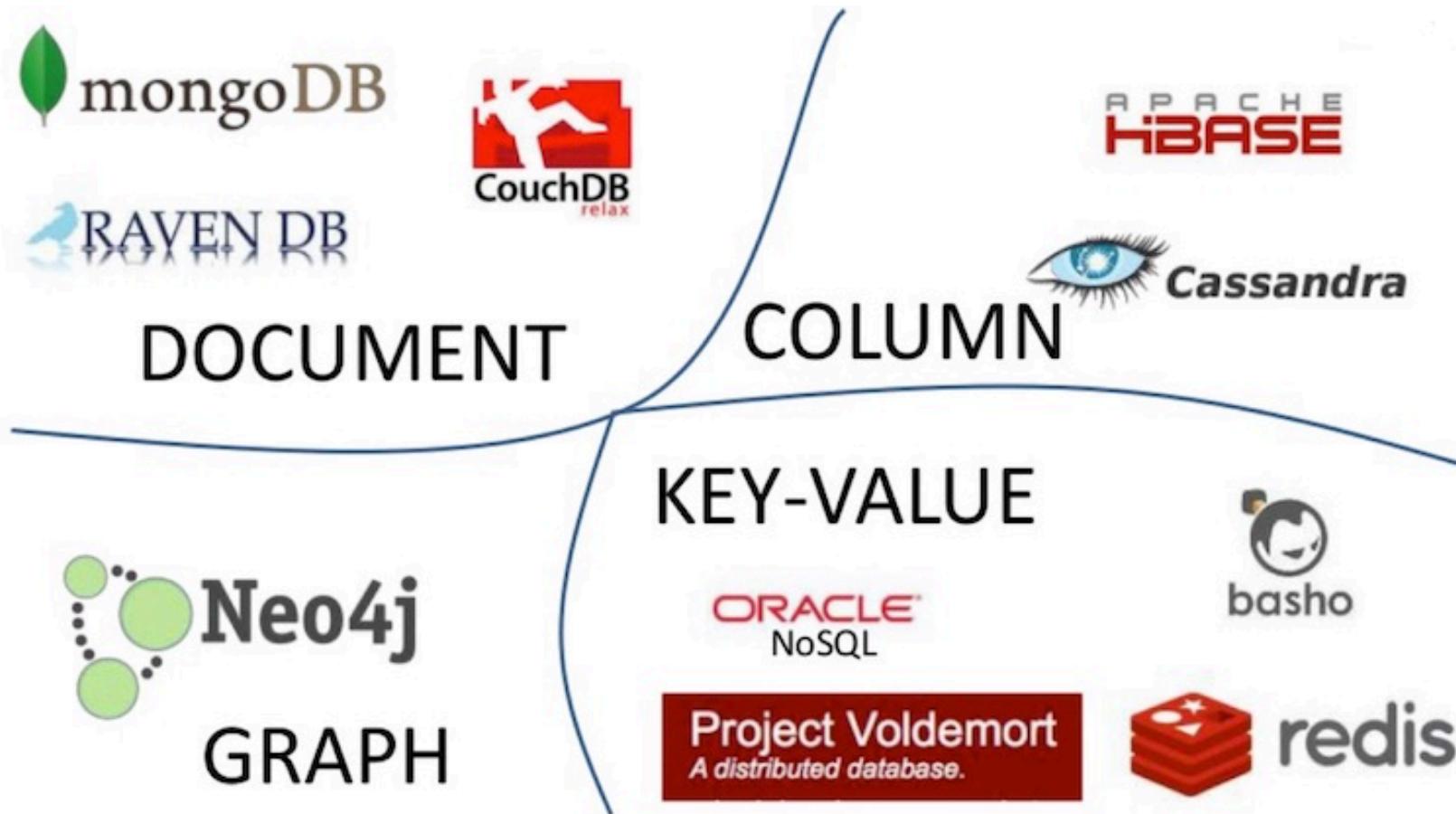
```
{  
  "articles": [  
    {  
      "title": "title of the article",  
      "articleID": 1,  
      "body": "body of the artricle",  
      "author": "Isaac Asimov",  
      "comments": [  
        {  
          "username": "Fritz",  
          "join date": "1/4/2014",  
          "commentid": 1,  
          "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"  
        }  
      ]  
    }  
  ]  
}
```

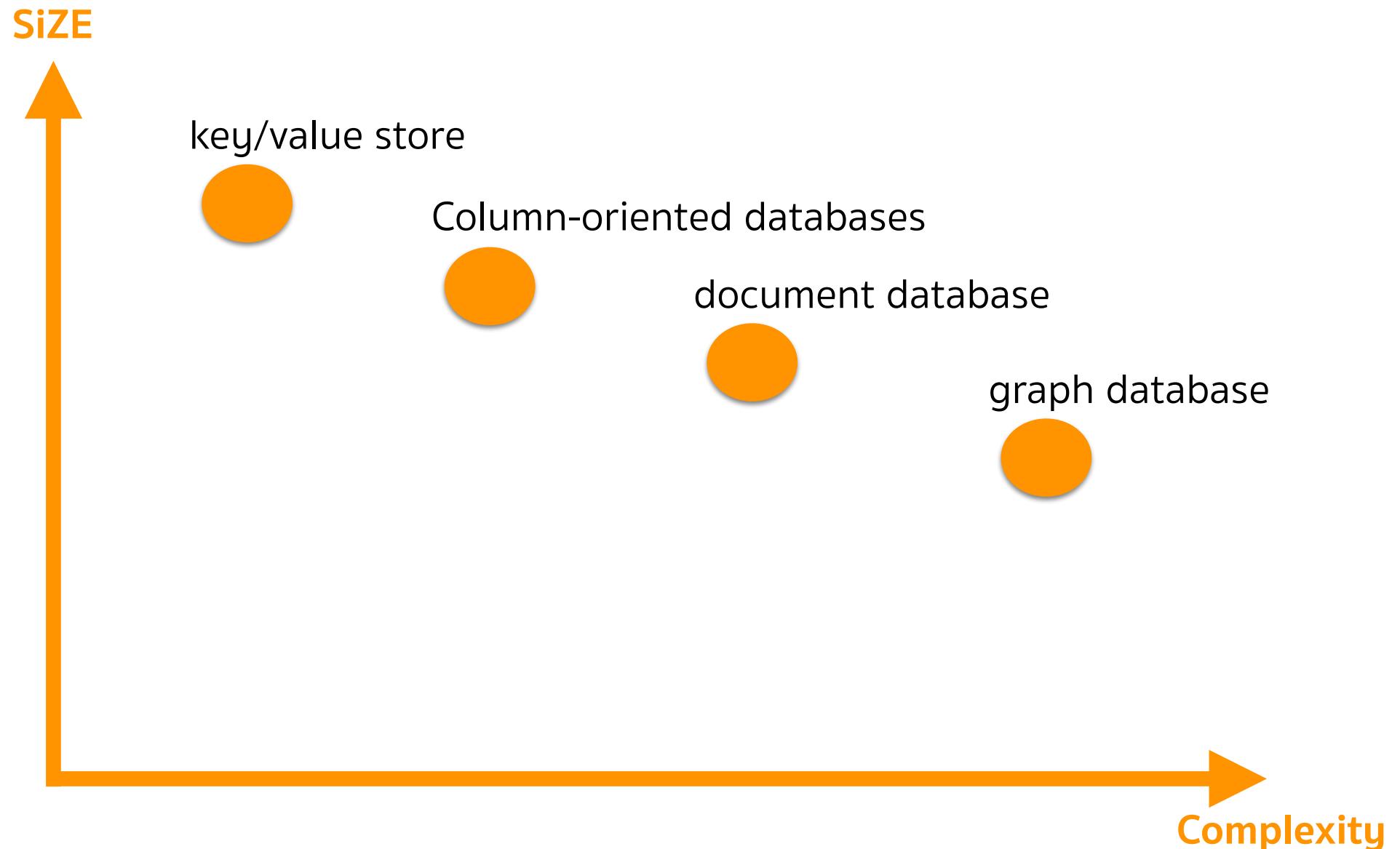


Graph database

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







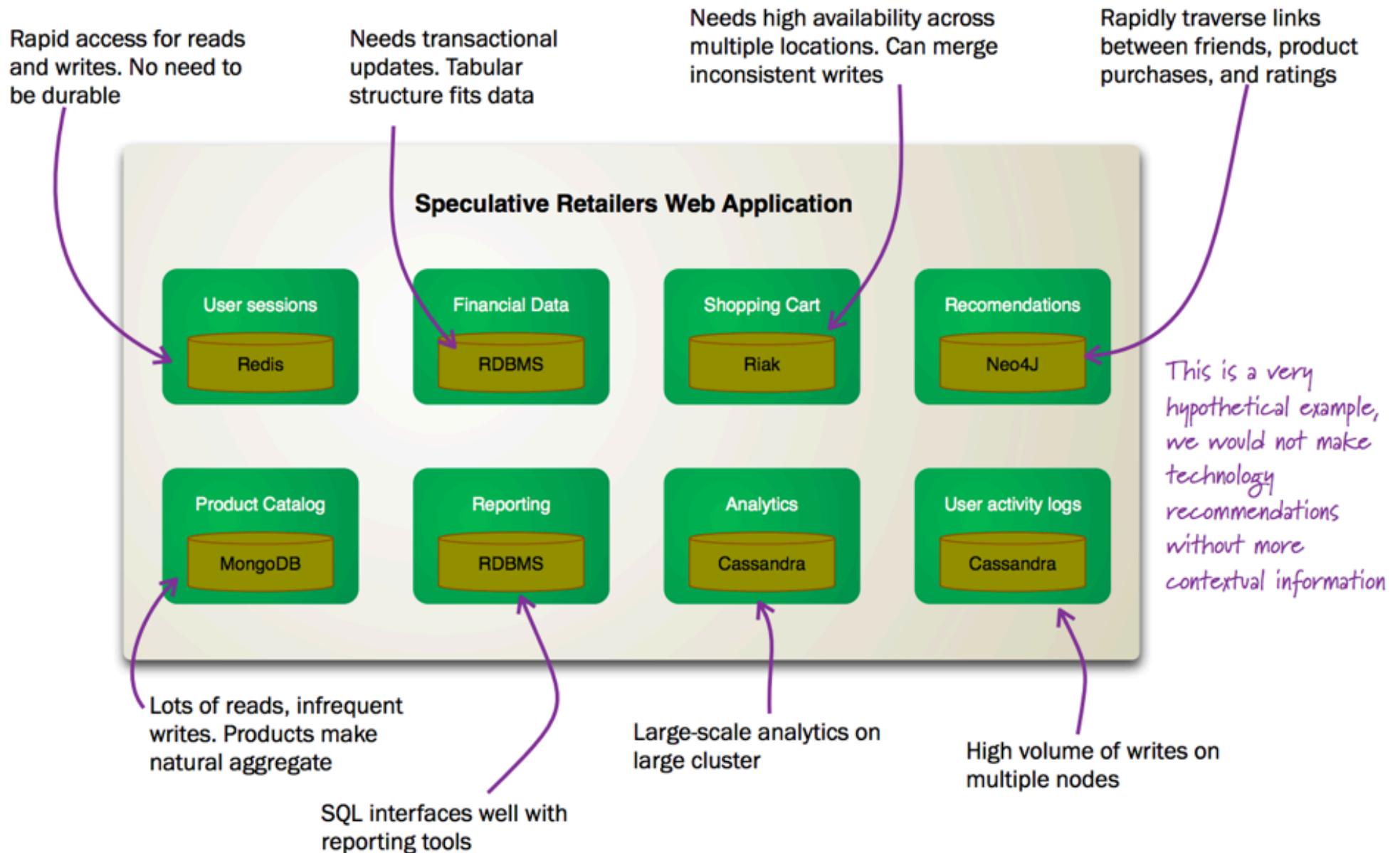
SQL

works great, can't scale for large data

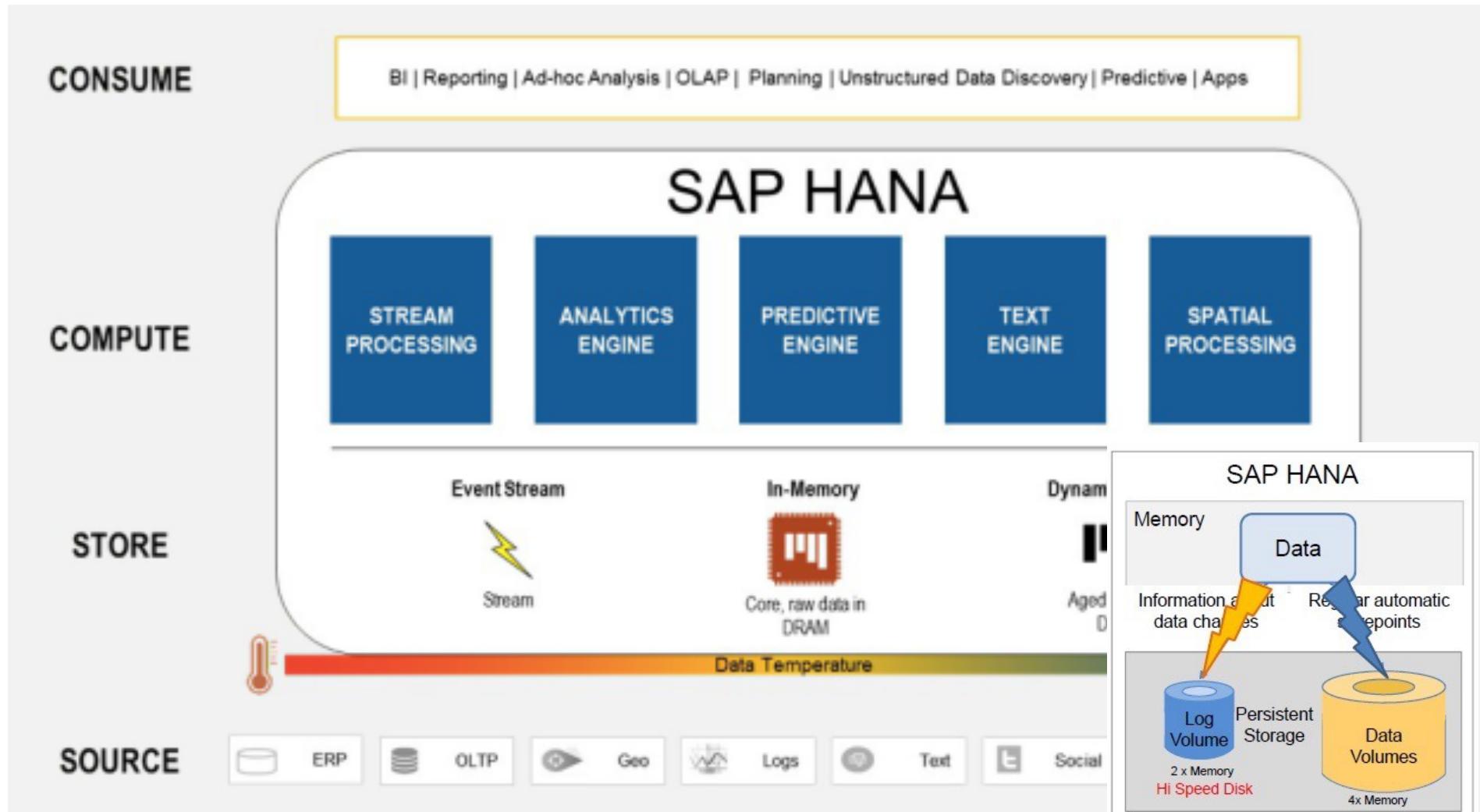
NoSQL

works great, doesn't fit all situations

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



SAP Hana



MPP:Oracle Exadata

Oracle Exadata Database Machine

Extreme Performance for the Cloud

Ellison announces next-generation systems



Ease compliance: OFSAA and Oracle Exadata (PDF)



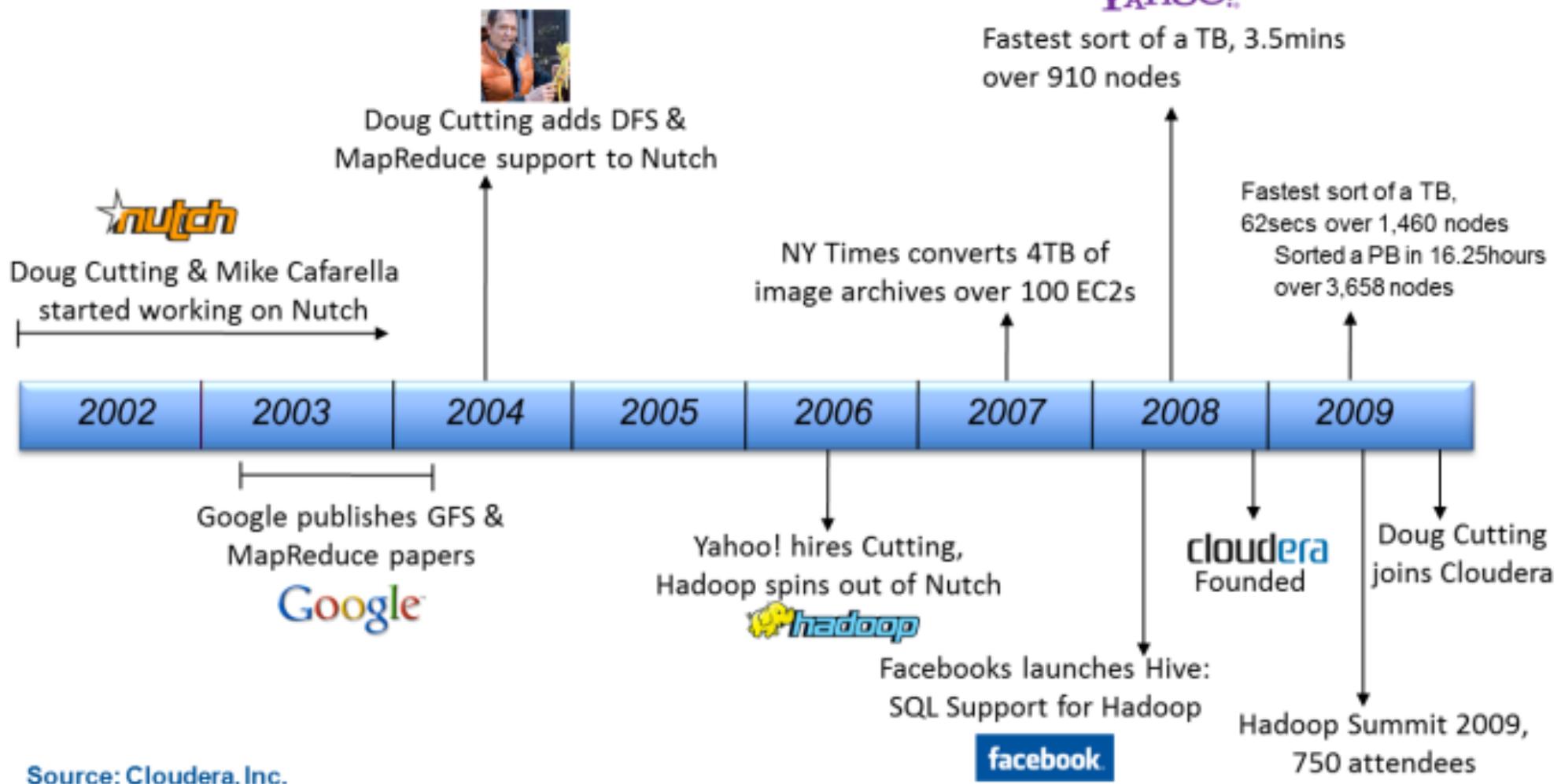
What is Hadoop ?

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

Completely written in java
Open source & distributed under Apache license

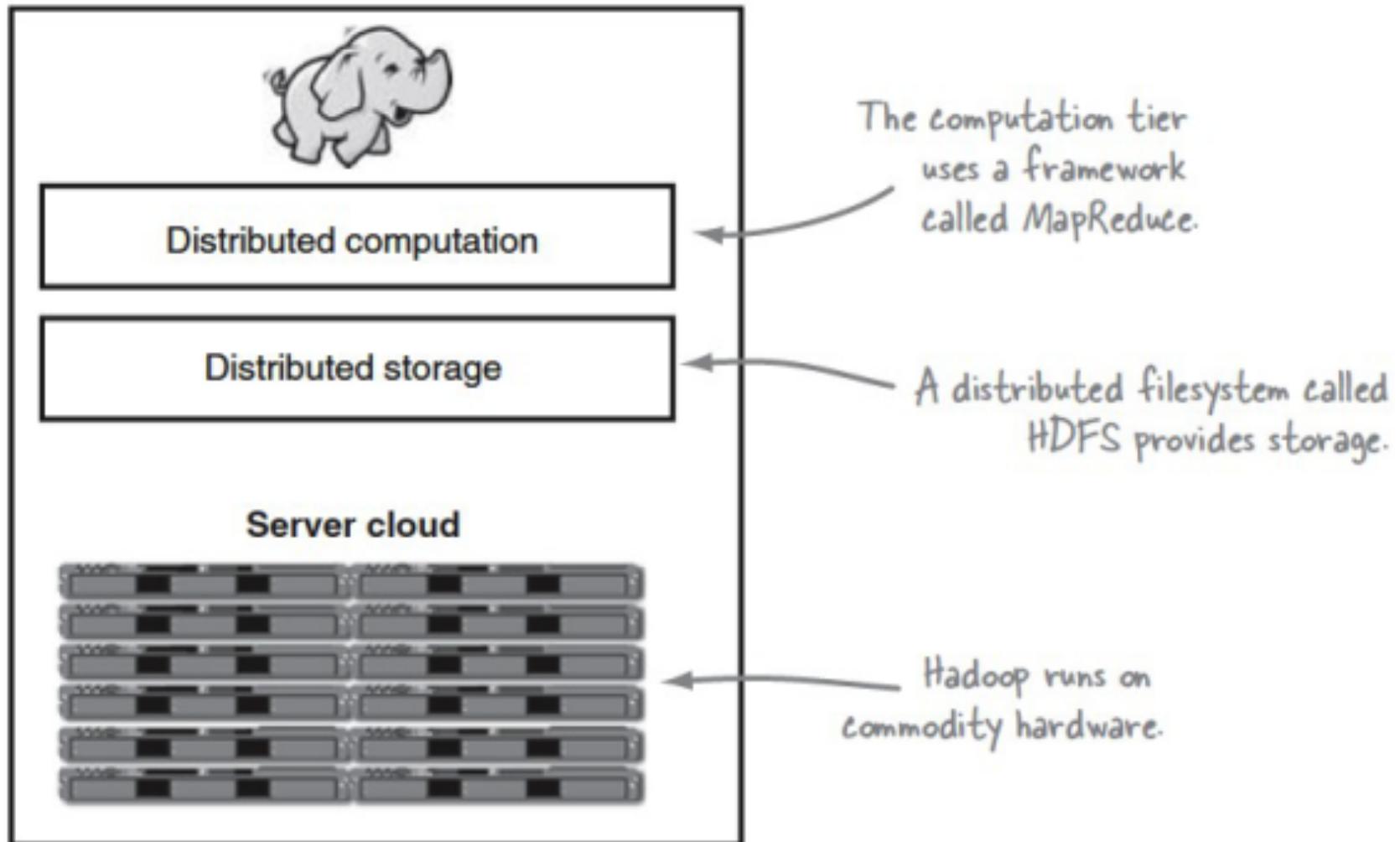


Hadoop Creation History

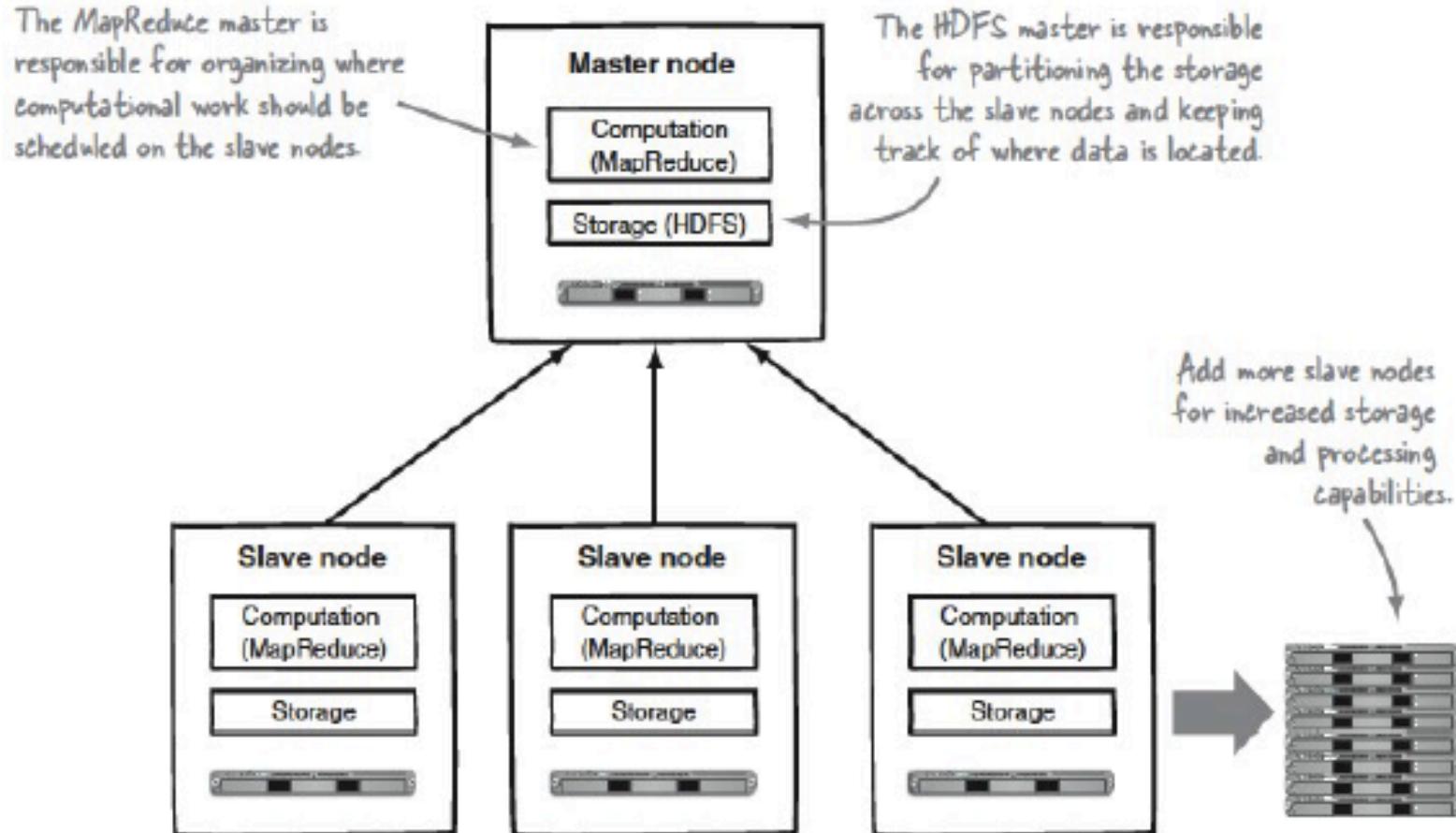


Source: Cloudera, Inc.

Hadoop Environment



Hadoop Architecture



Major Hadoop Components

Hadoop Distributed File System (HDFS)

Map/Reduce System

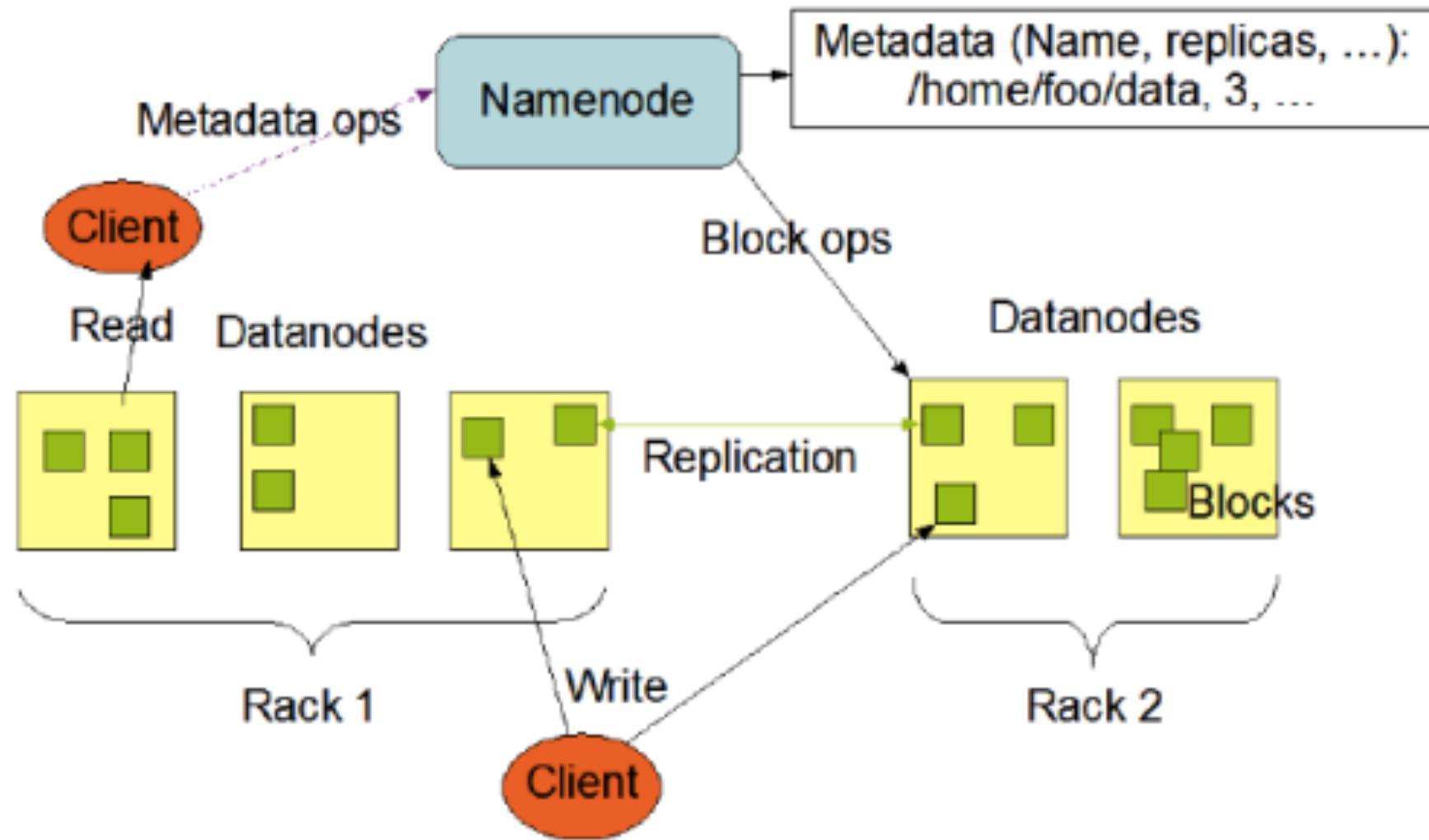


- Default storage for the Hadoop cluster
- Data is distributed and replicated over multiple machines
- Designed to handle very large files with streaming data access patterns.
- NameNode/DataNode
- Master/slave architecture (1 master 'n' slaves)
- Designed for large files (64 MB default, but configurable)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

HDFS Architecture



CHULALONGKORN
BUSINESS SCHOOL
FLAGSHIP FOR LIFE



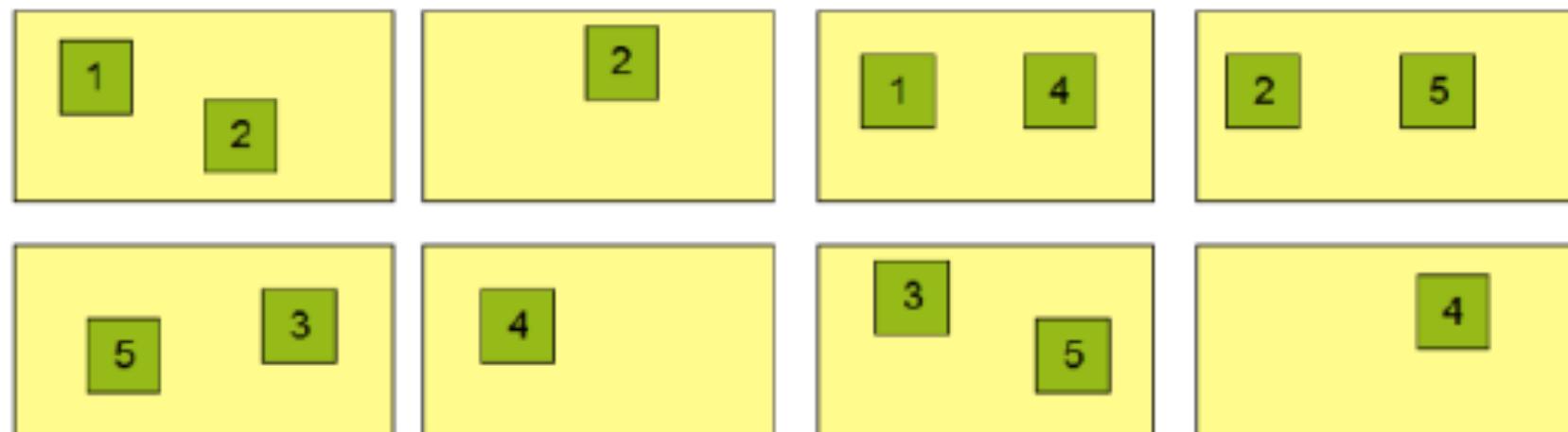
Data Replication in HDFS



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?



A file we want to store on HDFS ...

600 MB

We're raising the question because no one else wants to, because no one else wants to say what needs to be said.

And let's be real, it's the two-ton elephant in the room with nearly every other star's name on the trade rumor radar these days.

We've read over and over again about Nash refusing to ask for a trade, refusing to play the game that so many others have late in their careers.

How does HDFS work?



HDFS Splits file into **blocks** ...

256 MB

We're raising the question because no one else wants to, because no one else wants to say what needs to be said.

256 MB

And let's be real, it's the two-ton elephant in the room with nearly every other star's name on the trade rumor radar these days.

88 MB

We've read over and over again about Nash refusing to play the game that so many others have late in their careers.

How does HDFS work?



HDFS will create 3replicas of each block ...

3 copies

We're raising the question because no one else wants to, because no one else wants to say what needs to be said.

3 copies

And let's be real, it's the two-ton elephant in the room with nearly every other star's name on the trade rumor radar these days.

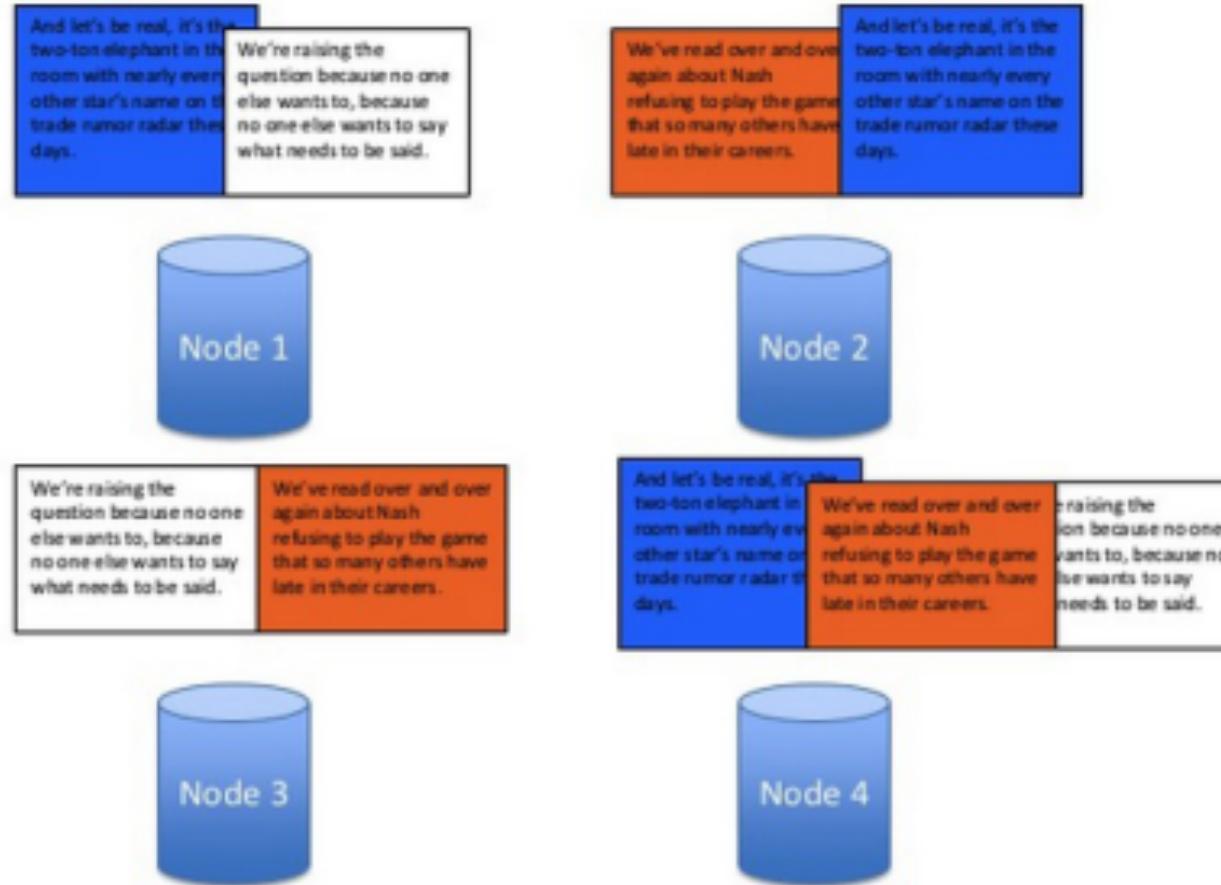
3 copies

We've read over and over again about Nash refusing to play the game that so many others have late in their careers.

How does HDFS work?



HDFS distributes these replicas
across the cluster ...

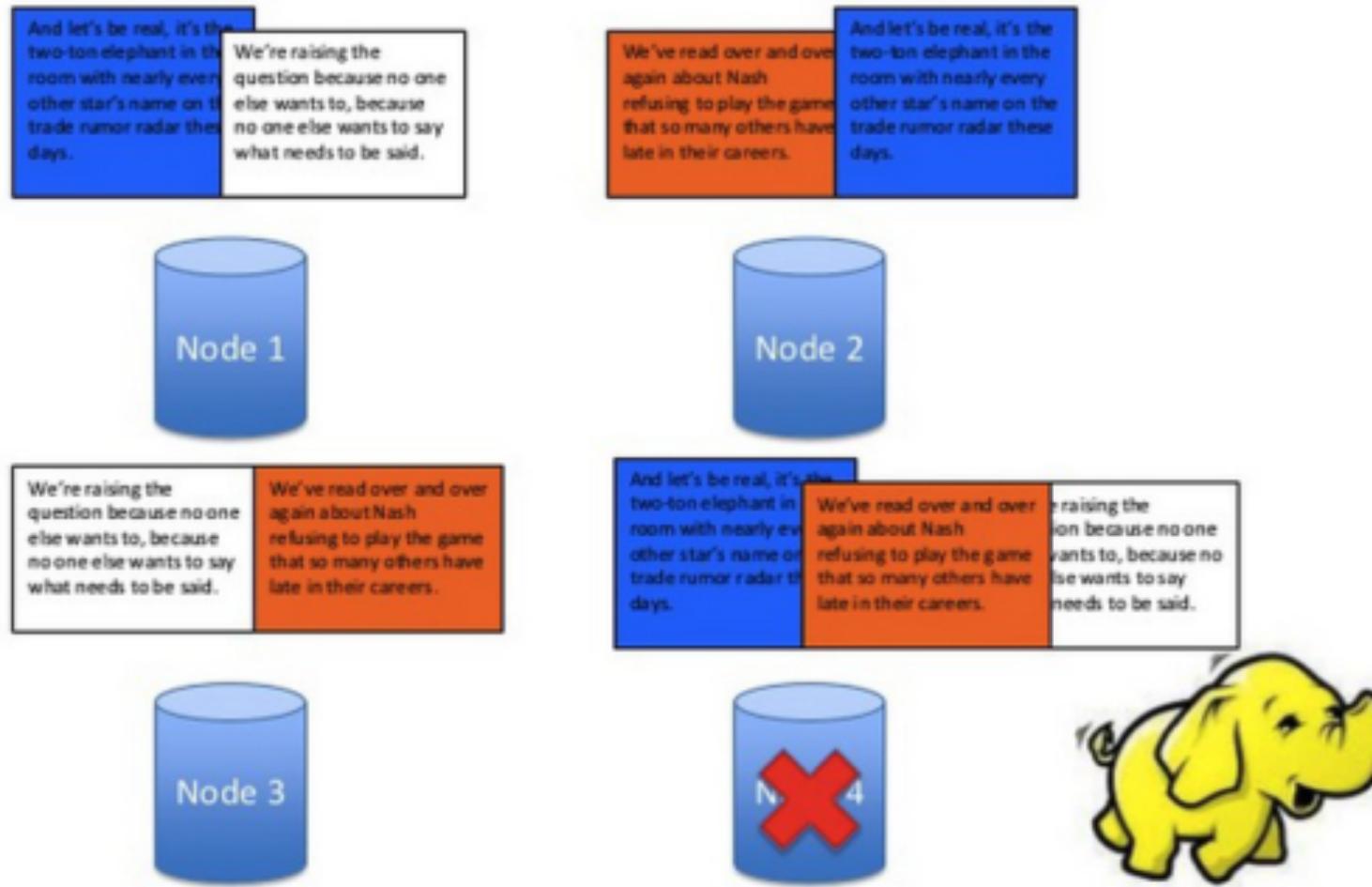


How does HDFS work?

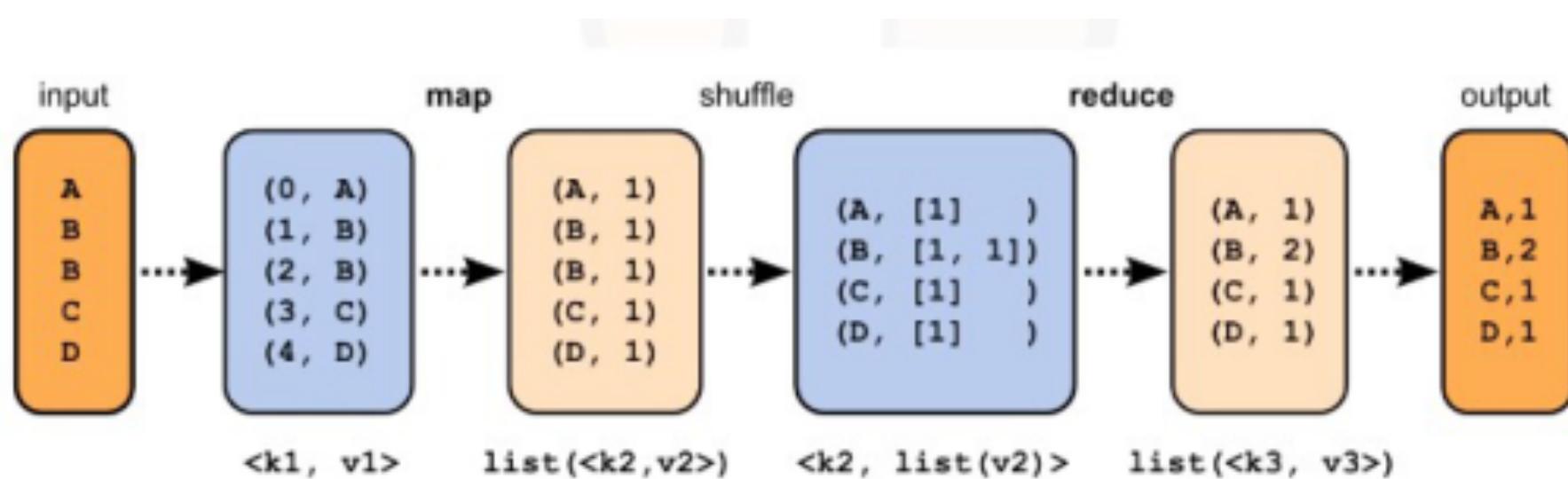


FLAGSHIP FOR LIFE

If a node goes down, we have copies elsewhere



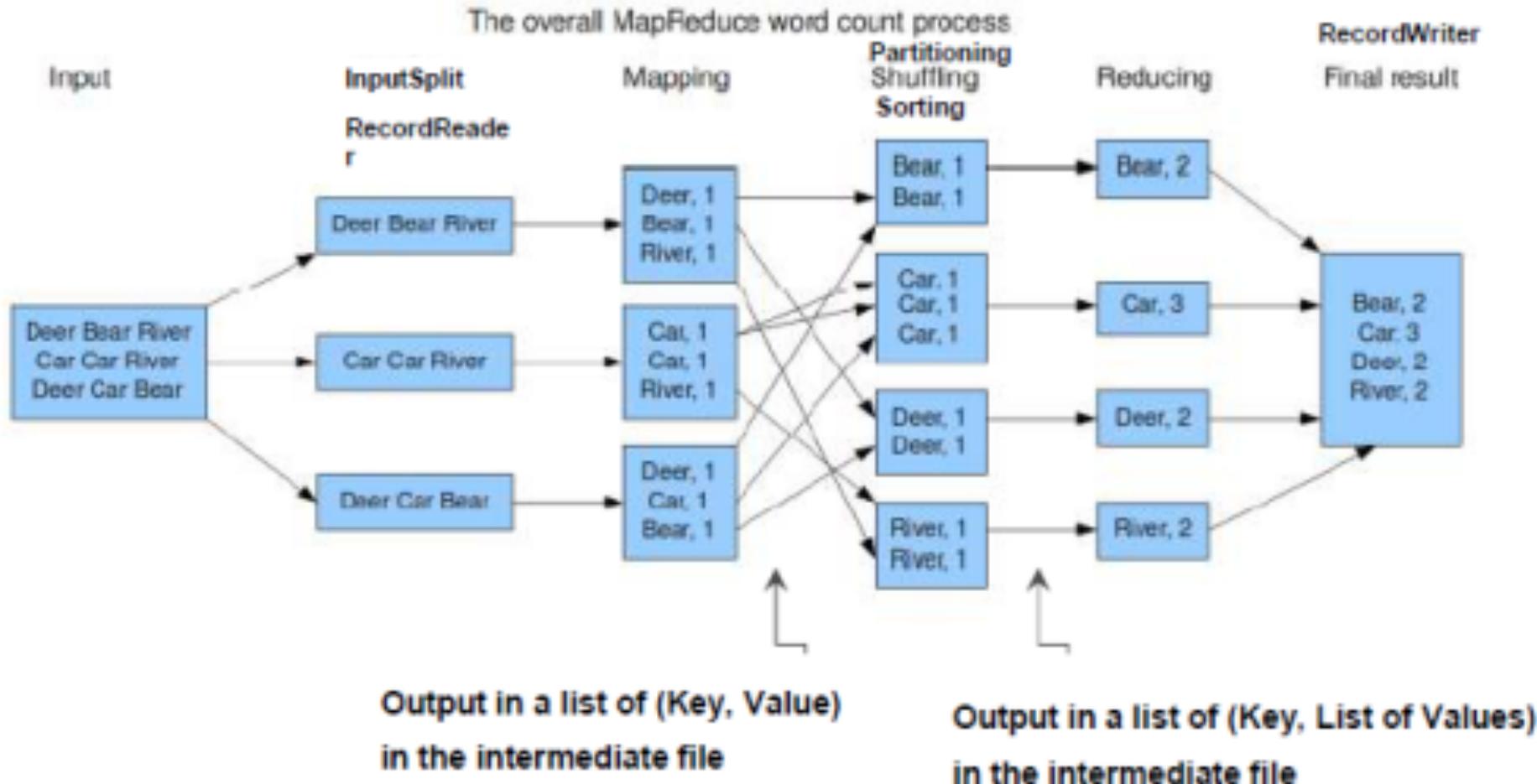
MapReduce Framework



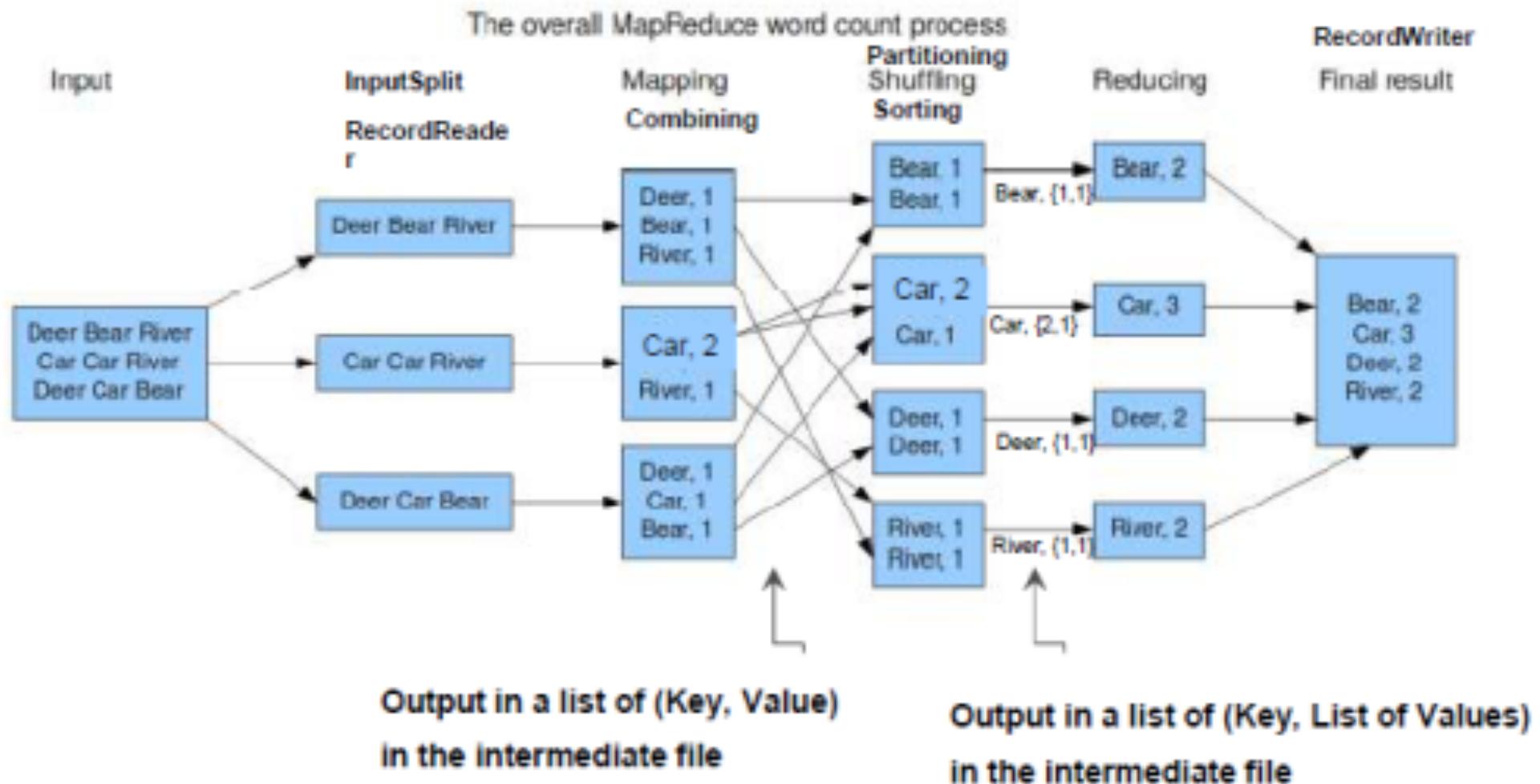
How does the MapReduce work ?



CHULALONGKORN
BUSINESS SCHOOL
FLAGSHIP FOR LIFE



How does the MapReduce work ?



WordCount - MapReduce

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

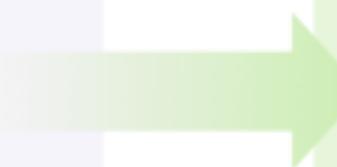
```
public static class Reduce extends MapReduceBase implements  
    Reducer<Text, IntWritable, Text, IntWritable> {  
    public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable>  
        output, Reporter reporter) throws IOException {  
        int sum = 0;  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

Hadoop 2.X



Hadoop 1

- Silos & Largely batch
- Single Processing engine



Hadoop 2 w/YARN

- Multiple Engines, Single Data Set
- Batch, Interactive & Real-Time

Batch
MapReduce

Interactive
Others

Real-Time
Others

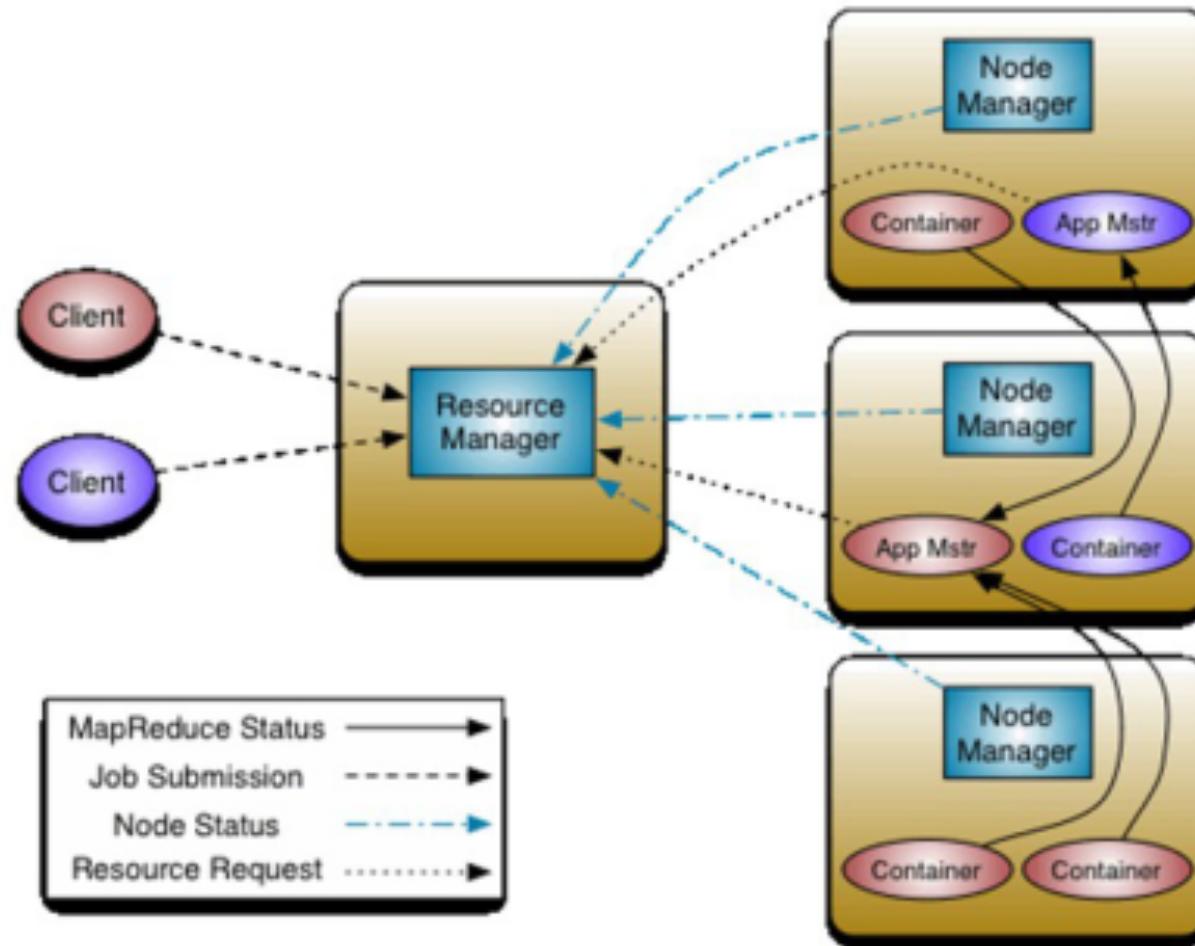
YARN: Data Operating System
(Cluster Resource Management)

MapReduce
(Cluster Resource Management
& Batch Data Processing)

1
HDFS
(Hadoop Distributed File System)

1
HDFS
(Hadoop Distributed File System)
N

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.

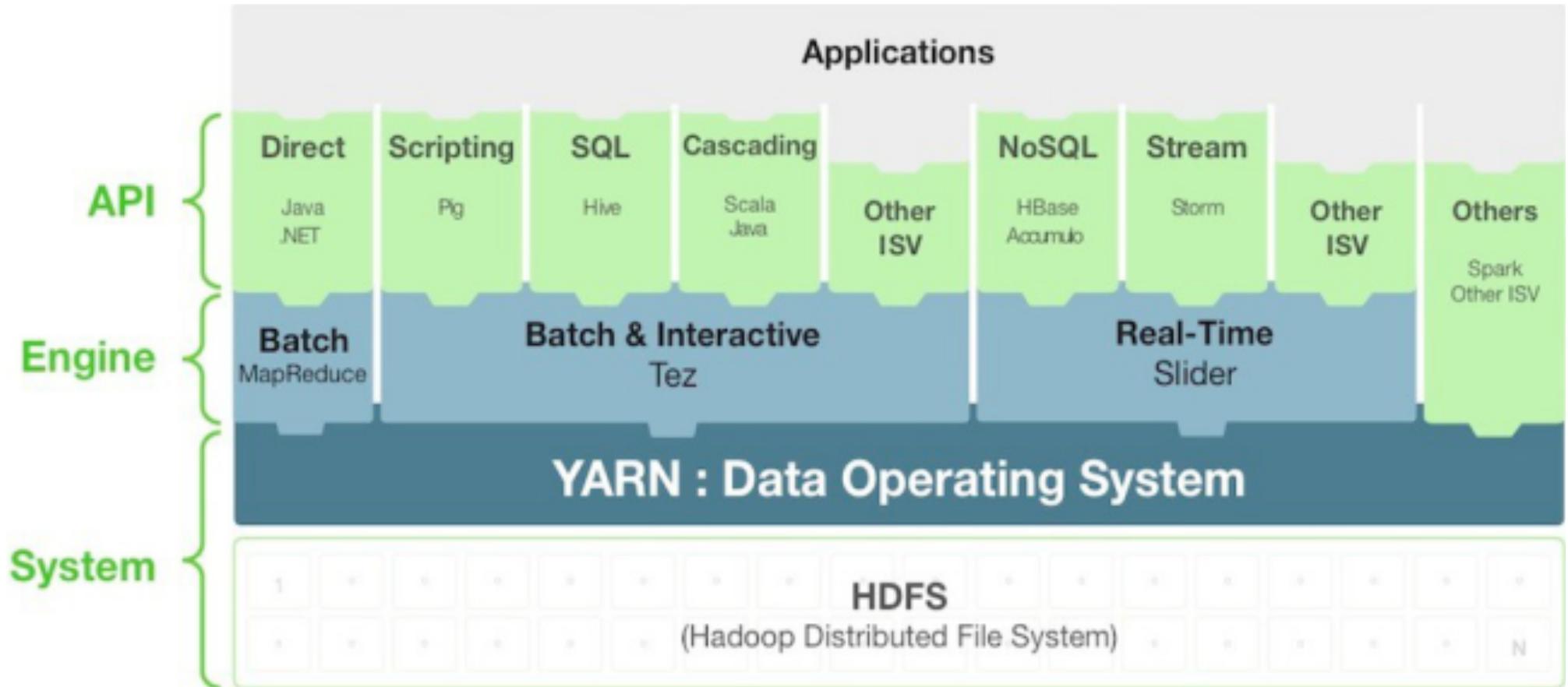
Hadoop.apache.org

Evolution of the Hadoop Platform

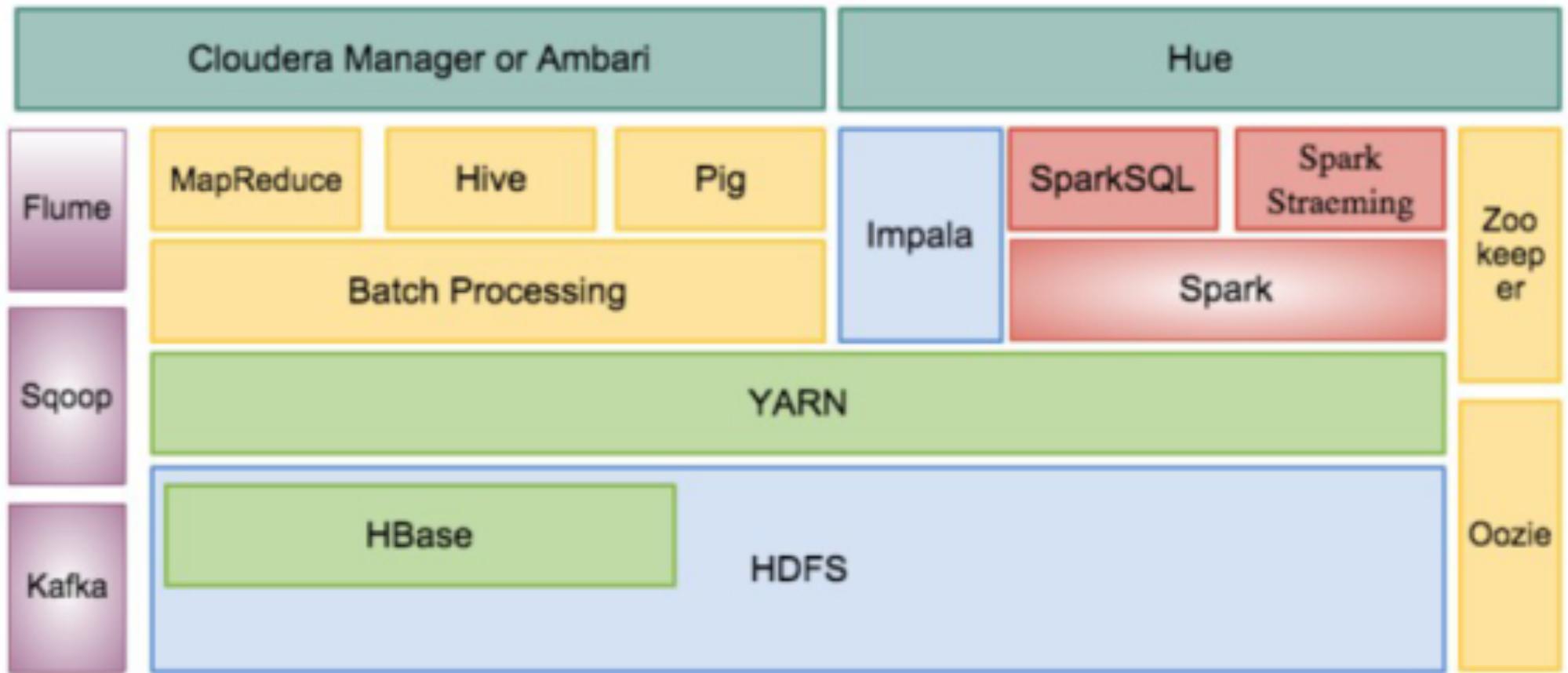
The stack is continually evolving and growing!

Evolution of the Hadoop Platform									
The stack is continually evolving and growing!									
Core Hadoop (HDFS, MapReduce)		Core Hadoop		Core Hadoop		Core Hadoop		Core Hadoop	
2006	Solr	Solr	ZooKeeper	HBase	Hive	MRUnit	Flume	Parquet	Ibis
	Pig	Pig	ZooKeeper	HBase	Mahout	HCatalog	Sentry	Sentry	Flink
2007	Core Hadoop	Core Hadoop	Core Hadoop	ZooKeeper	Sqoop	Hive	Bigtop	Spark	Parquet
					Whirr	MRUnit	Oozie	Tez	Sentinel
2008					Avro	HCatalog	Impala	Tez	Spark
					Hive	Hue	Kafka	Drill	Tez
2009					Mahout	MRUnit	Flume	Drill	Parquet
					HBase	HCatalog	Bigtop	Flume	Sentinel
2010					ZooKeeper	ZooKeeper	Oozie	Oozie	Spark
					Solr	Solr	Impala	Impala	Tez
2011					Pig	MRUnit	Kafka	Kafka	Parquet
					YARN	Hue	Drill	Drill	Sentinel
2012					YARN	Sqoop	Flume	Flume	Spark
					Core Hadoop	Whirr	Bigtop	Bigtop	Tez
2013					Core Hadoop	Avro	Oozie	Oozie	Parquet
					Core Hadoop	Hive	Impala	Impala	Sentinel
2014-15					Core Hadoop	Mahout	Hive	Hive	Spark
					Core Hadoop	HBase	Hive	Hive	Tez
					Core Hadoop	ZooKeeper	Hive	Hive	Parquet

Hadoop 2.x Ecosystems



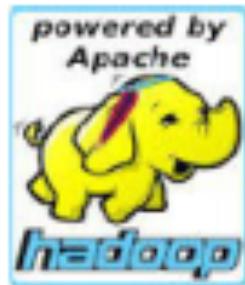
Hadoop Ecosystems



Hadoop Distribution



CHULALONGKORN
BUSINESS SCHOOL
FLAGSHIP FOR LIFE



MAPR

cloudera



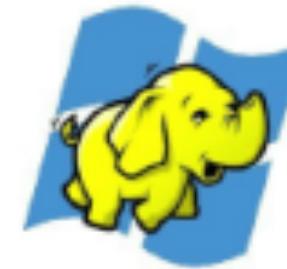
Pivotal™



TERADATA

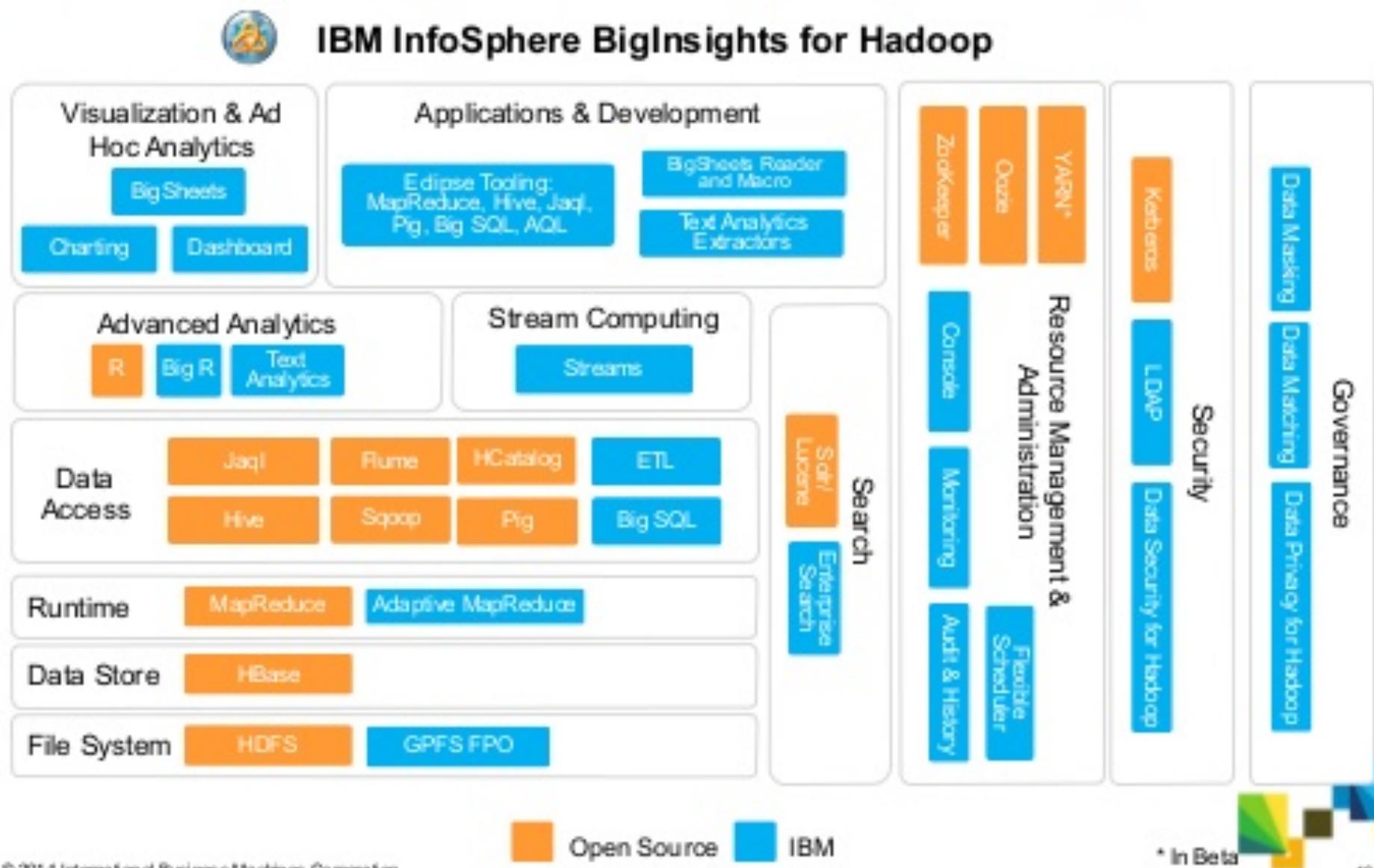


amazon
web services™

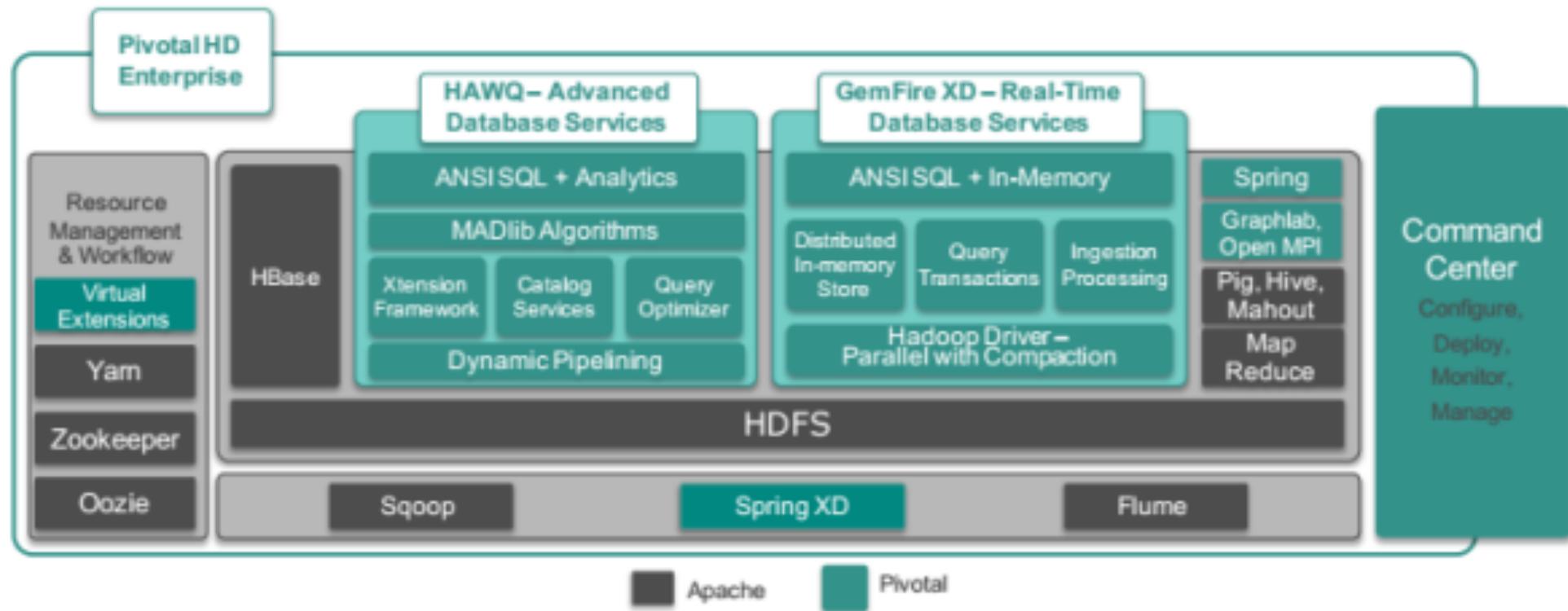


Microsoft Azure

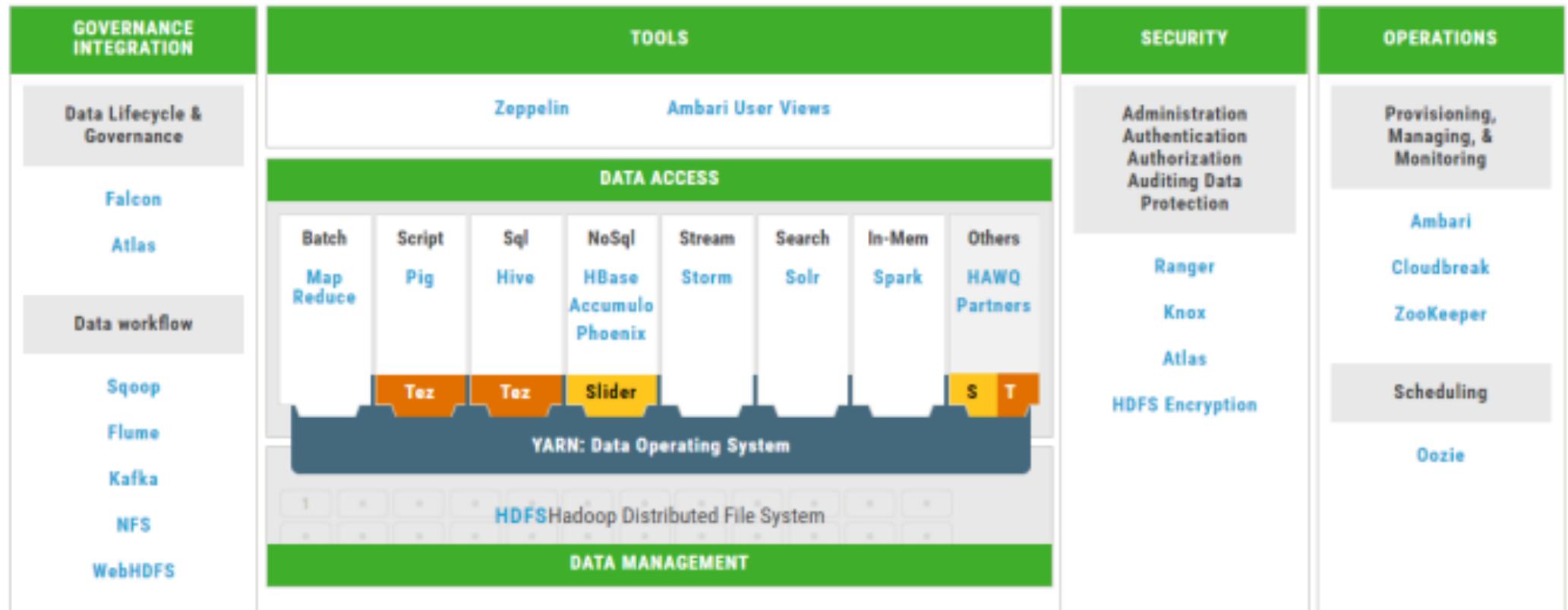
IBM InfoSphere BigInsights



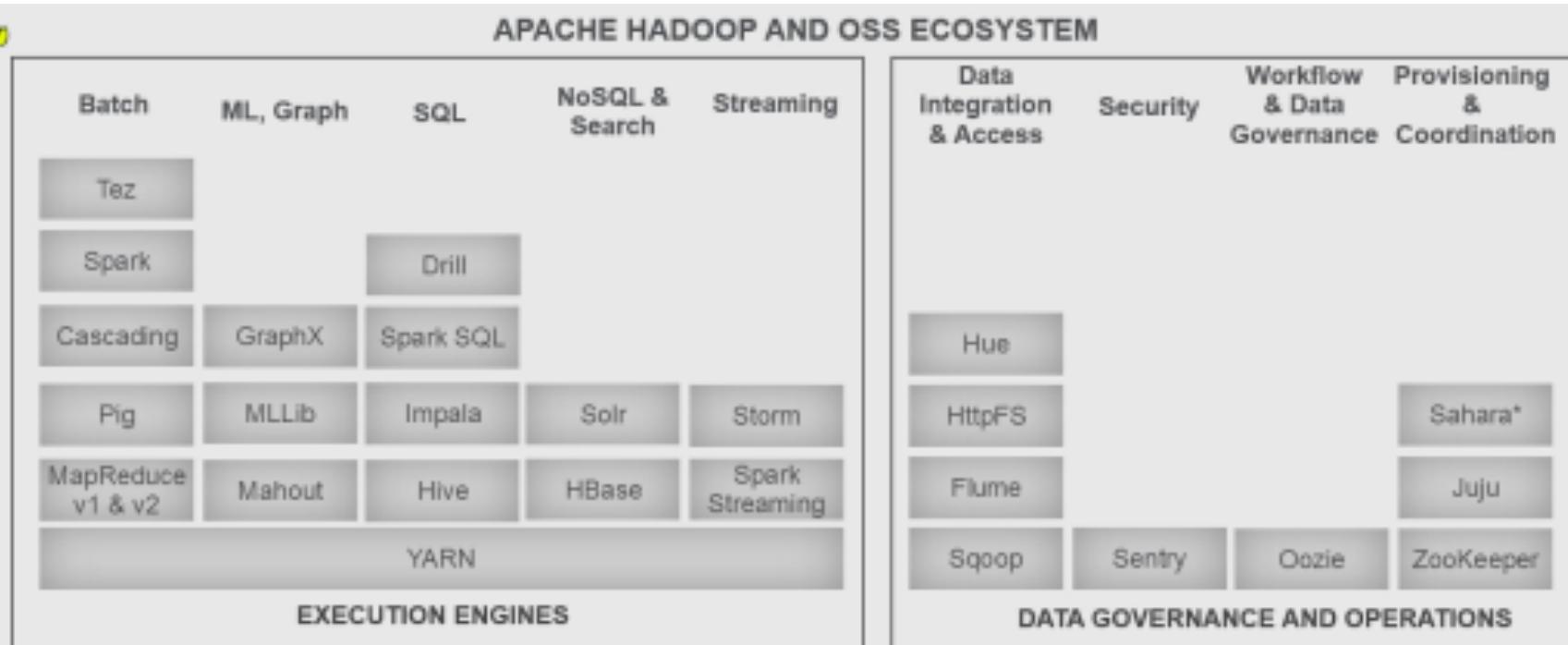
Pivotal HD Architecture



Hortonworks



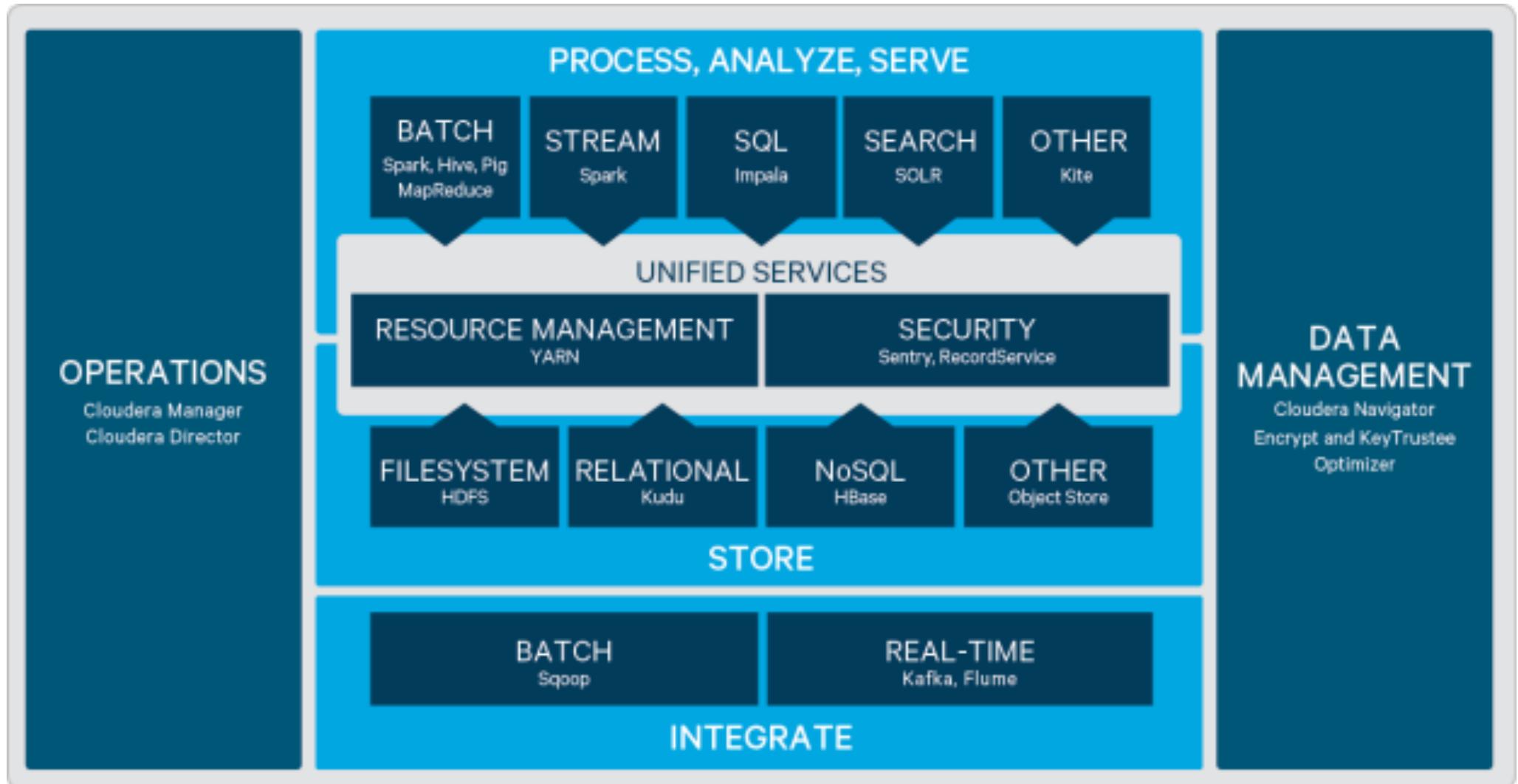
Management



MapR-FS

Data Platform

MapR-DB



Default Cloudera Services

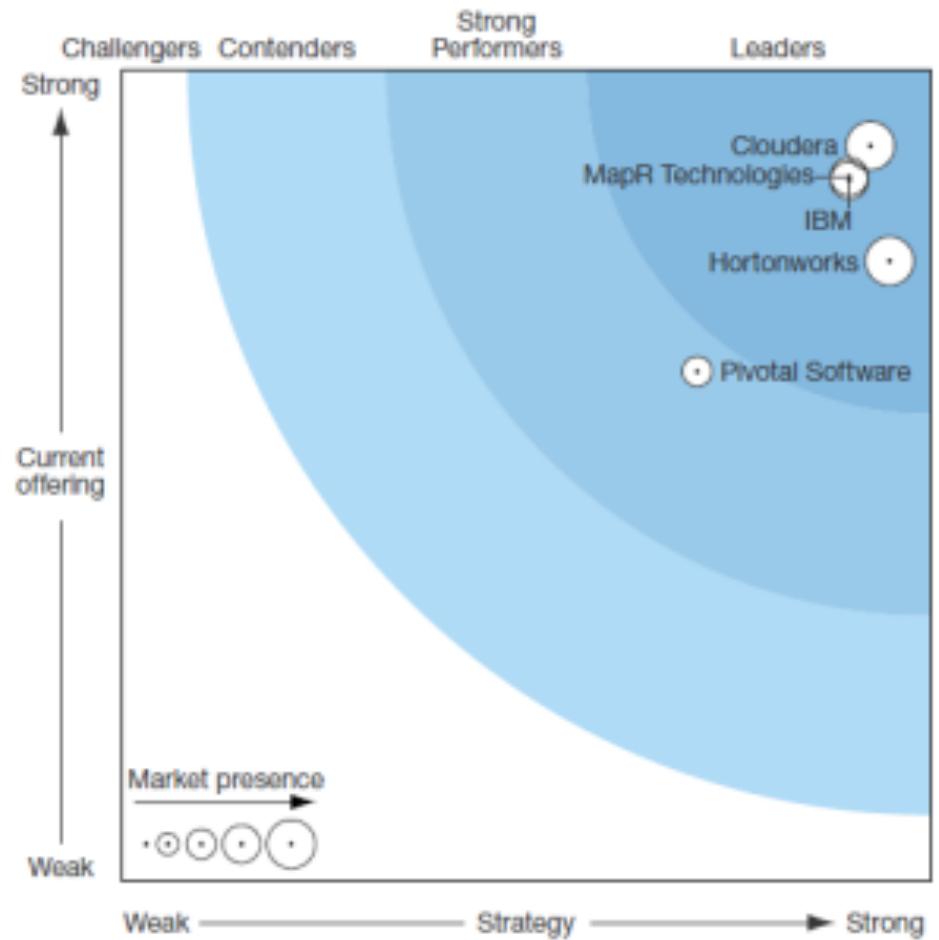
- Cloudera Manager
- HDFS
- YARN
- Apache Hive and Pig
- Apache Flume and Sqoop
- Apache Oozie
- Cloudera Hue
- ZooKeeper



cloudera

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

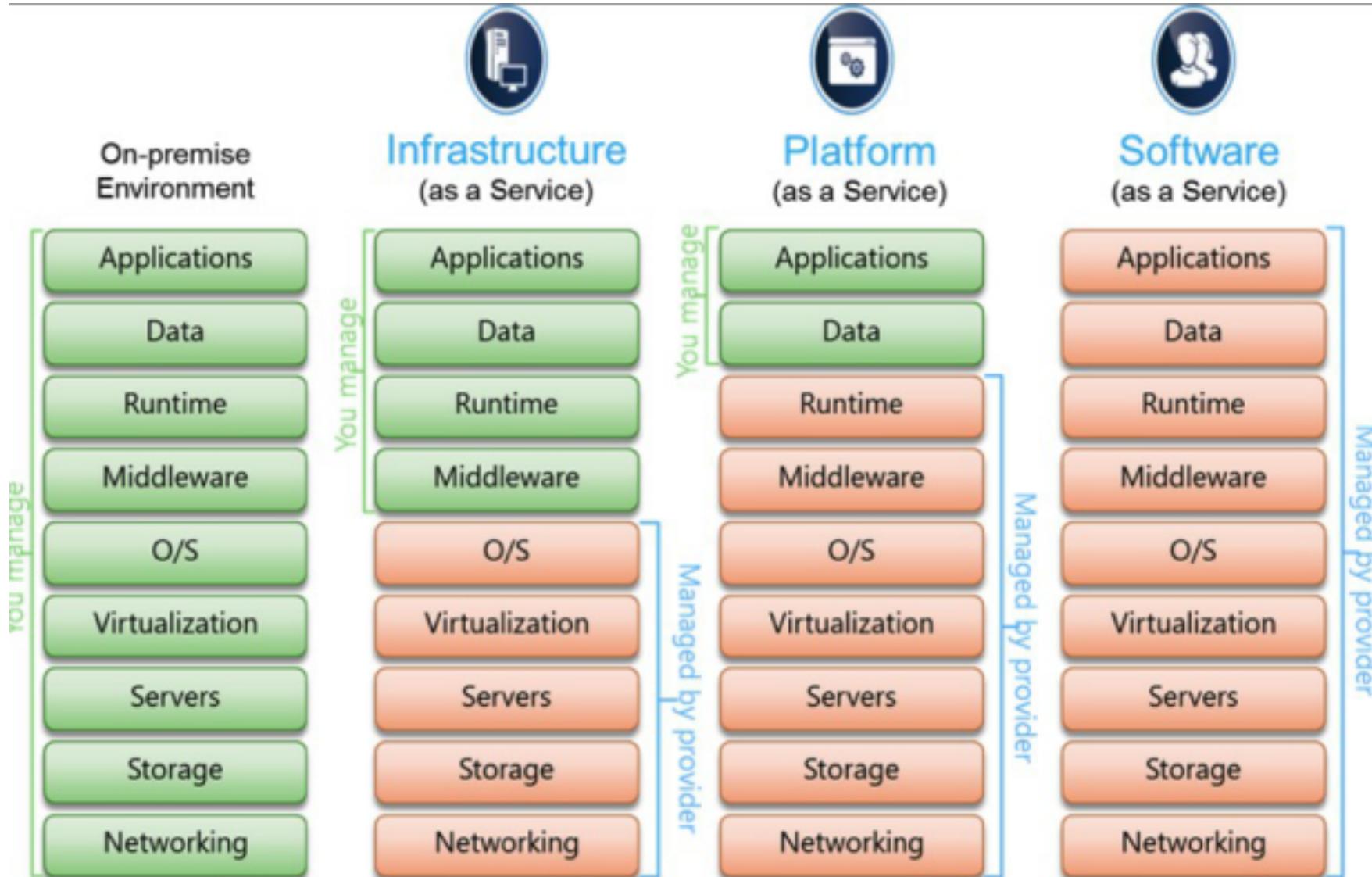


Big Data Using Public Cloud

Issue with Big Data Infrastructure

- Large investment
- Scalability
- ROI
- Business Cases

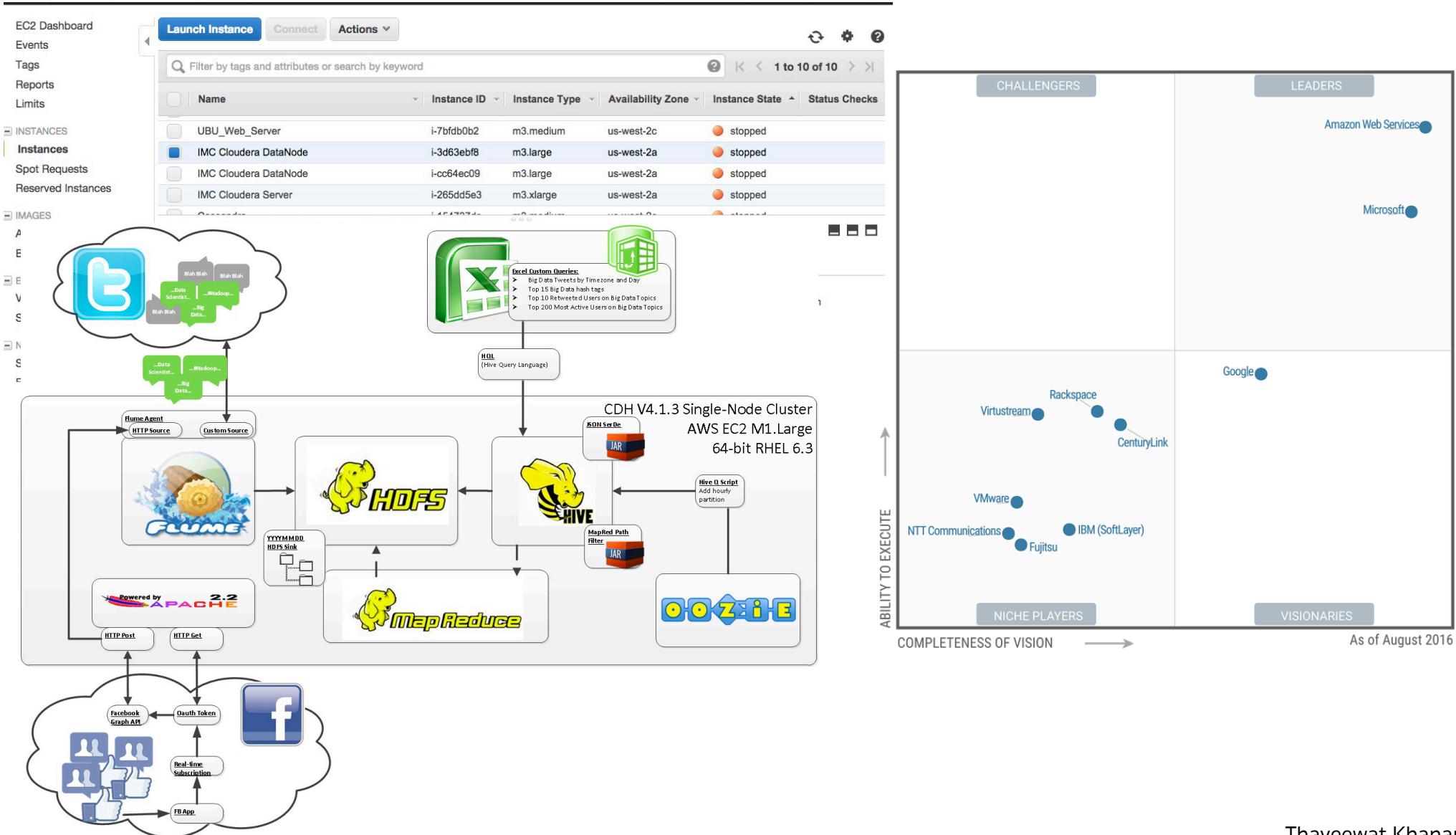
Cloud Technology





Big Data on Cloud

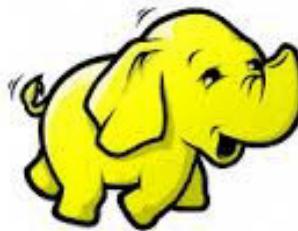
Big Data using IaaS



Big Data on Cloud

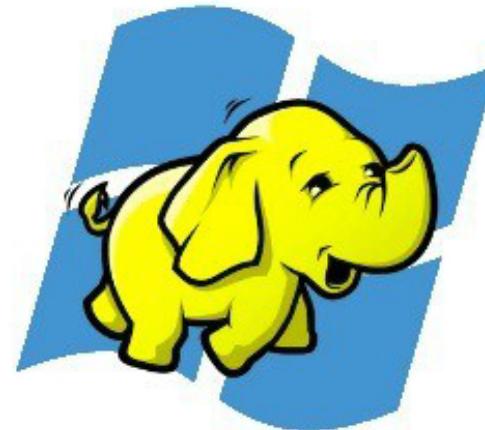
Using Big Data as a Services

Amazon Web Services



Amazon

Elastic Mapreduce



Microsoft Azure Hadoop