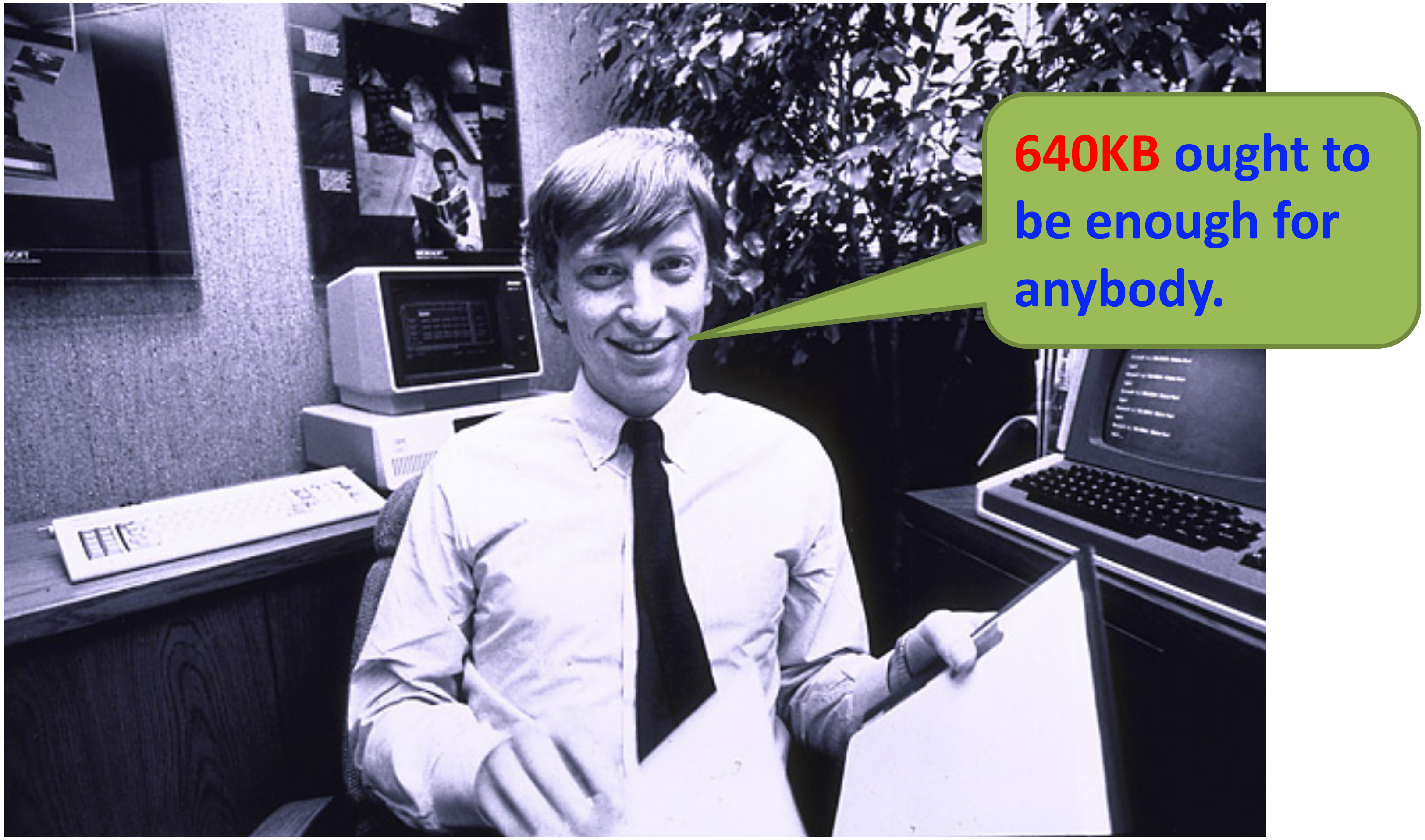
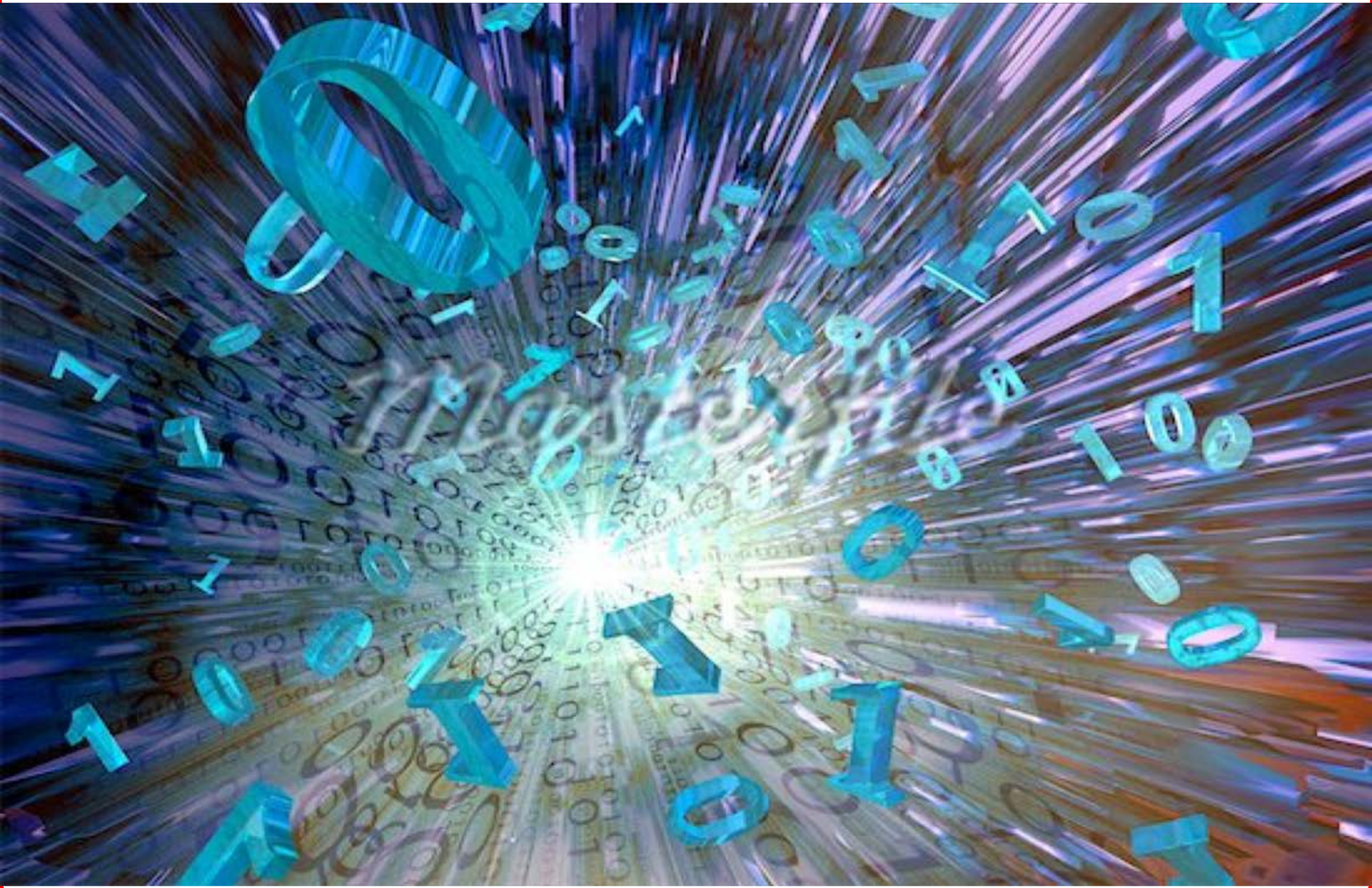
Concepts of Big Data and Hadoop

What amount of data is “big?”

How much data do you need?



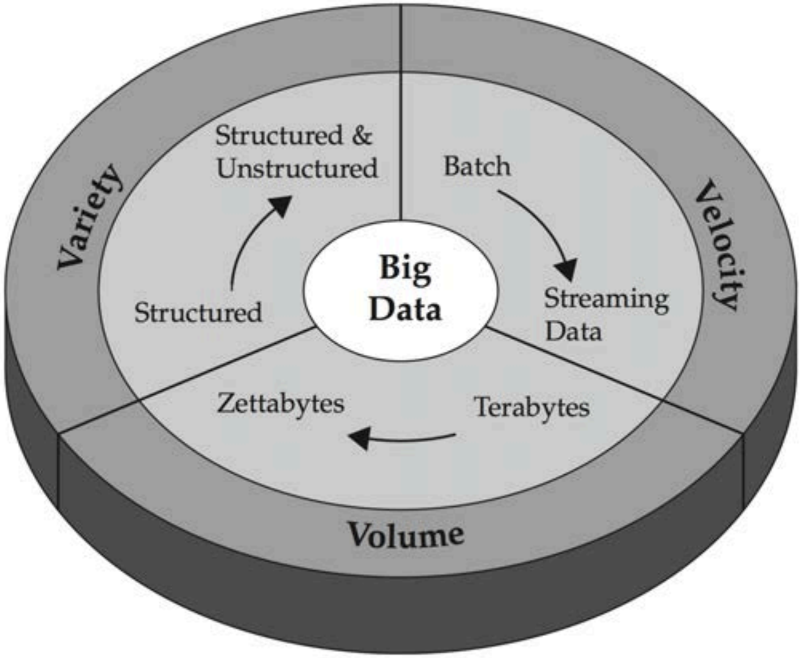
But not for long, data exploded!



What is Big Data?

The term **Big Data** is used to describe the exponential growth and availability of data, both structured and unstructured. It can be characterized with three V's:

- **Volume**: Increased big volume of data.
- **Velocity**: Data is streaming in at a high and unprecedented speed.
- **Variety**: Data comes in all types of formats: structured data in traditional databases, unstructured text documents, web logs, machine log, emails, images, audio and video files …



More About the Name of Big Data

- Strictly speaking, it is loosely-defined or even a bit of a misnomer.
- It implies that pre-existing data is small (but it is not always true). This term can be treated as a new way to look at data.
- Also it implies that only challenge is its sheer size (size is only one of them).
- The three V’s together well define the term “Big Data,” which applies to data sets that are very hard to or cannot be processed and analyzed using traditional methods or tools.

Why Big Data?

- More and more data is accumulated (stored) naturally as the business goes on and people hope to gain something from it later.
- Big data may be critical to business, and for many organizations big data is the reality of doing business.
- What it matters is the potential value that can be derived from big data not just acquiring large amounts of data.
- The hopeful vision is that organizations will be able to take data from any source and analyze it to find answers that enable
  - cost and time reductions,
  - new product development and optimized offerings,
  - better business decision making,
  - new knowledge and technologies,
  - …

Big Data Provide Many Job Opportunities

- Financial services
- Retail
- Insurance
- Manufacturing
- Healthcare
- Web/mobile/social settings
- Government
- Research
- Transportation
- Communication
- Weather services
- Military
- Entertainment
- …

Some Examples of Big Data Applications

- Determine root causes of failures of power grids.
- Optimize routes for many thousands of UPS package delivery vehicles while they are on the road.
- Analyze millions of SKUs to determine prices that maximize profit and clear inventory.
- Generate retail coupons at the point of sale based on the customer's current and past purchases.
- Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.
- Identify customers who matter the most.
- Use large amount of transaction data and data mining to detect fraudulent behavior.
- …

More about the UPS Example

- UPS now tracks data on over 20 million packages per day for about 11 million customers, with an average of 150 million online tracking requests per day.
- Much of its recently acquired big data comes from telematics sensors in more than 50,000 vehicles. The data on UPS trucks includes their speed, direction, braking and drive train performance. The data is used to monitor daily performance and to drive a major redesign of drivers' route structures.
- This has already led to savings, e.g., in 2011 a saving of more than **8.4 million** gallons of fuel by cutting **85 million miles** off of daily routes.
- UPS estimates that saving only one daily mile per driver saves the company **$30 million**.
- UPS is also attempting to use data and analytics to optimize the efficiency of its 2,000 aircraft flights per day.

Examples of the Amount of Data Produced by Different Entities

Facebook: a Big Data Business Example

As of 2012, it had

- 40+ billion photos (100 PB)
- 6 billion messages per day (5-10 TB)
- 900 million users (1 trillion connections)
- ~ 500 TB/day of data ingestion

As of 2014, it had

- 864 million daily active users
- 703 million mobile daily active users
- > 300 PB user data

As of 2015, it had

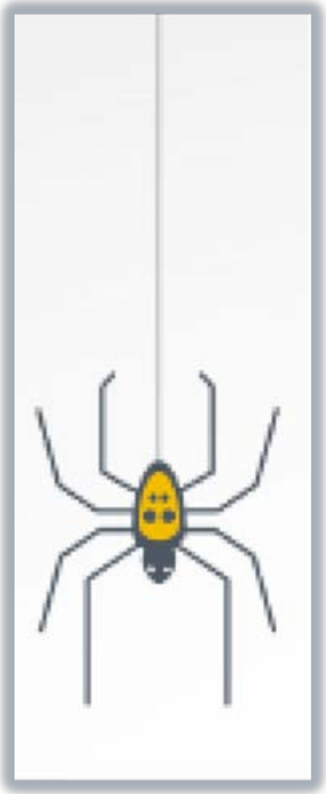
- 1.04 billion daily active users
- 934 million mobile daily active users on average for December 2015

As of 2017, it had

- 2.04 billion monthly active user
- 1.37 billion people on average log onto Facebook daily

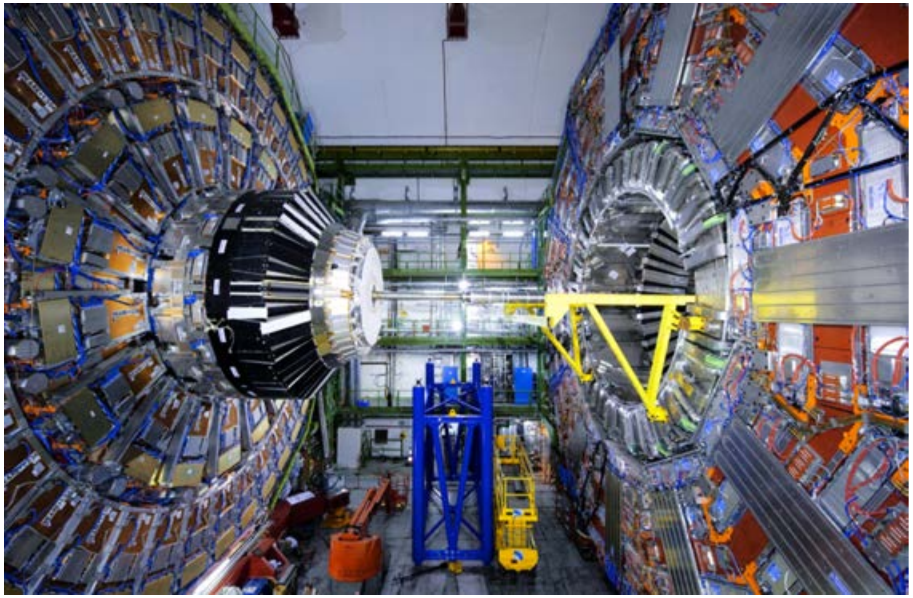
Value	Metric
10^3	Kilobytes (KB)
10^6	Megabytes (MB)
10^9	Gigabytes (GB)
10^{12}	Terabytes (TB)
10^{15}	Petabytes (PB)
10^{18}	Exabytes (EB)
10^{21}	Zettabytes (ZB)
10^{24}	Yottabytes (YB)

Amazon: Amazon Common Crawl, > 5 million web pages and 50 TB of data. Amazon online retail store, as of 2010 more than 42 TB of retail data.

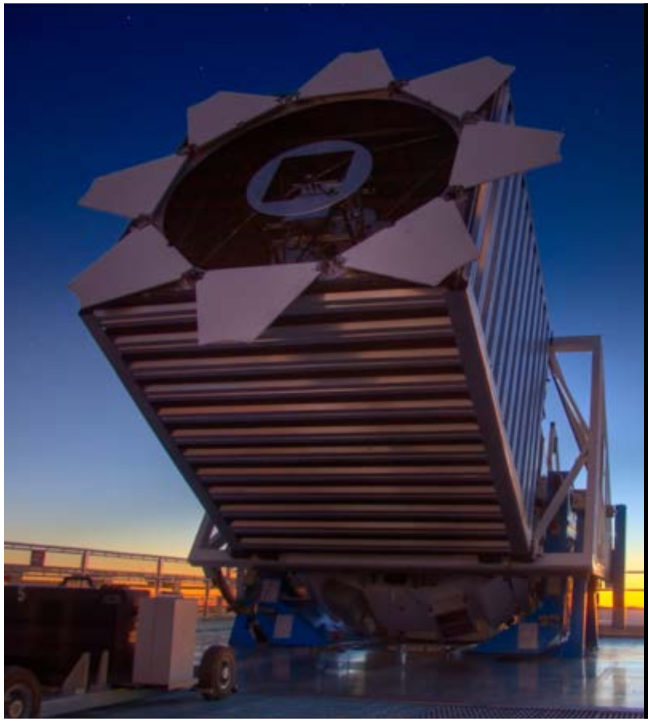


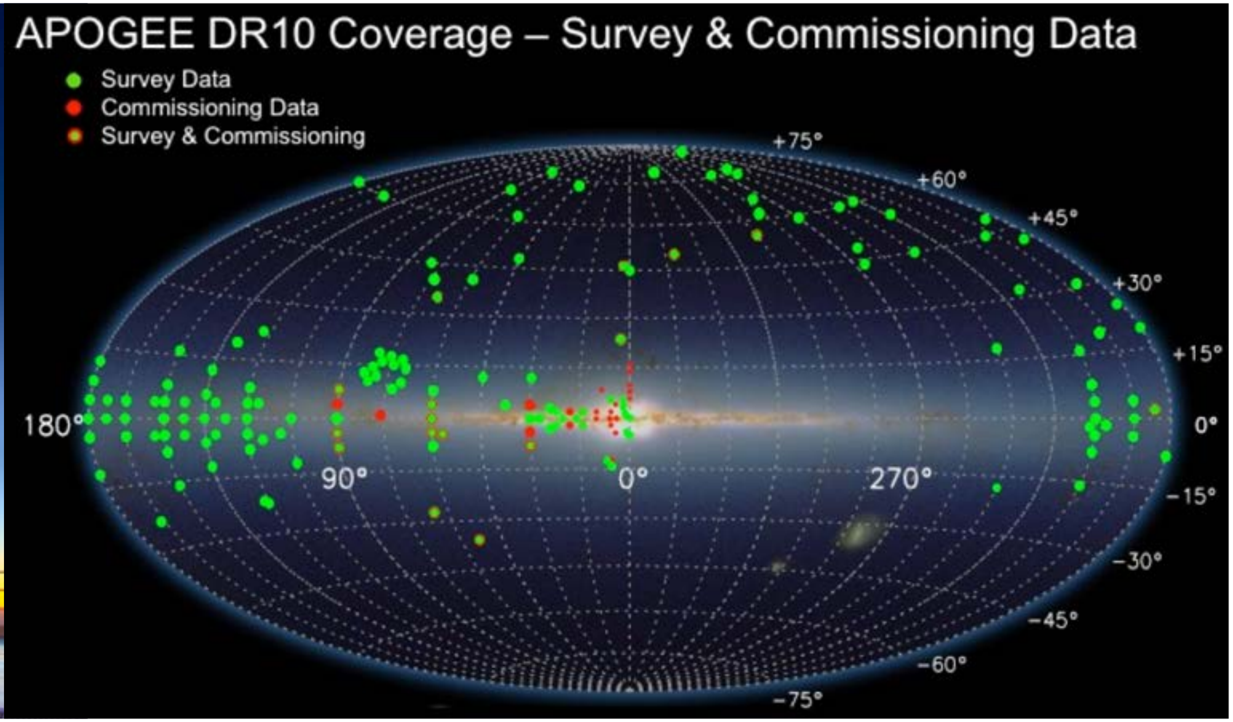
Amazon is using Big Data, AI to accelerate profits

Large Hadron Collider, 15 PB of data generated annually



Sloan Digital Sky Survey, 50 TB of data mapped 35% of the sky

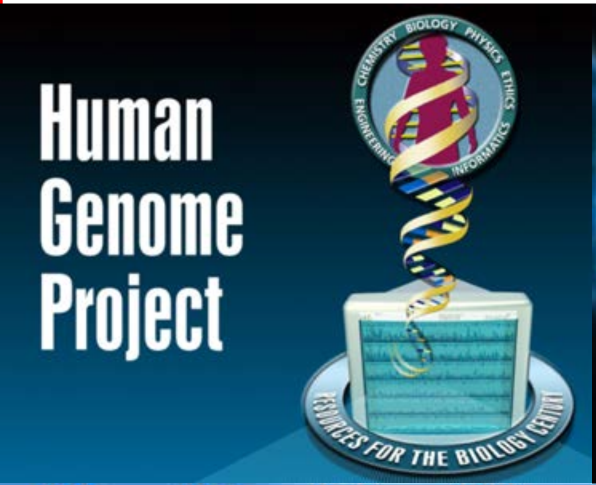


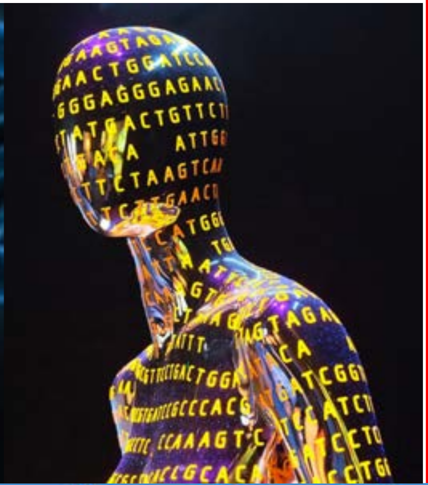


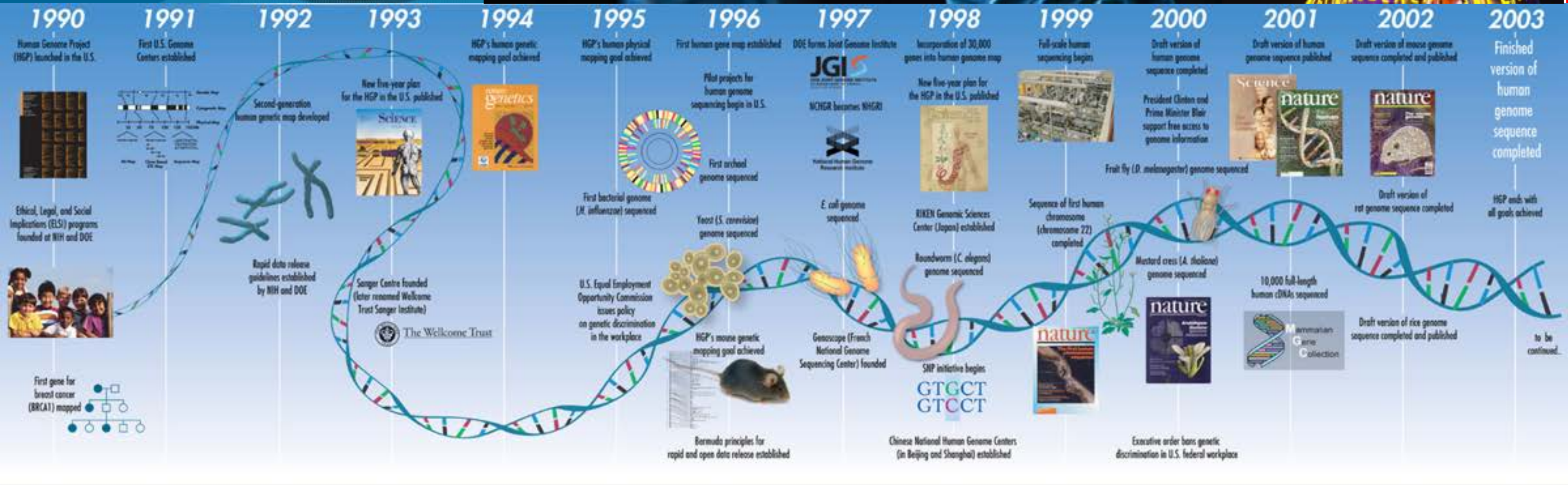
AT&T, as of 2010 323 TB of unique database and 1.9 trillion phone call records

World Data Centre for Climate, 220 TB web data and 6 PB of additional data

Genome projects, > 250 GB sequence generated







1990
Human Genome Project (HGP) launched in the U.S.
Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE
First gene for breast cancer (BRCA1) mapped

1991
First U.S. Genome Centers established

1992
Second-generation human genetic map developed
Rapid data release guidelines established by NIH and DOE

1993
New five-year plan for the HGP in the U.S. published
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)
The Wellcome Trust

1994
HGP's human genetic mapping goal achieved

1995
HGP's human physical mapping goal achieved
First bacterial genome (*H. influenzae*) sequenced
U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace

1996
First human gene map established
Pilot projects for human genome sequencing begin in U.S.
First archeal genome sequenced
Yeast (*S. cerevisiae*) genome sequenced
HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release established

1997
DOE forms Joint Genome Institute
JGI
NCHGR becomes NHGRI
E. coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded

1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the U.S. published
RIKEN Genomic Sciences Center (Japan) established
Roundworm (*C. elegans*) genome sequenced
SNP initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers (in Beijing and Shanghai) established

1999
Full-scale human sequencing begins
Sequence of first human chromosome (chromosome 22) completed

2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (*D. melanogaster*) genome sequenced
Mustard cress (*A. thaliana*) genome sequenced
Executive order bans genetic discrimination in U.S. federal workplace

2001
Draft version of human genome sequence published
10,000 full-length human cDNAs sequenced
Mammalian Gene Collection

2002
Draft version of mouse genome sequence completed and published
Draft version of rat genome sequence completed
Draft version of rice genome sequence completed and published

2003
Finished version of human genome sequence completed
HGP ends with all goals achieved
to be continued...

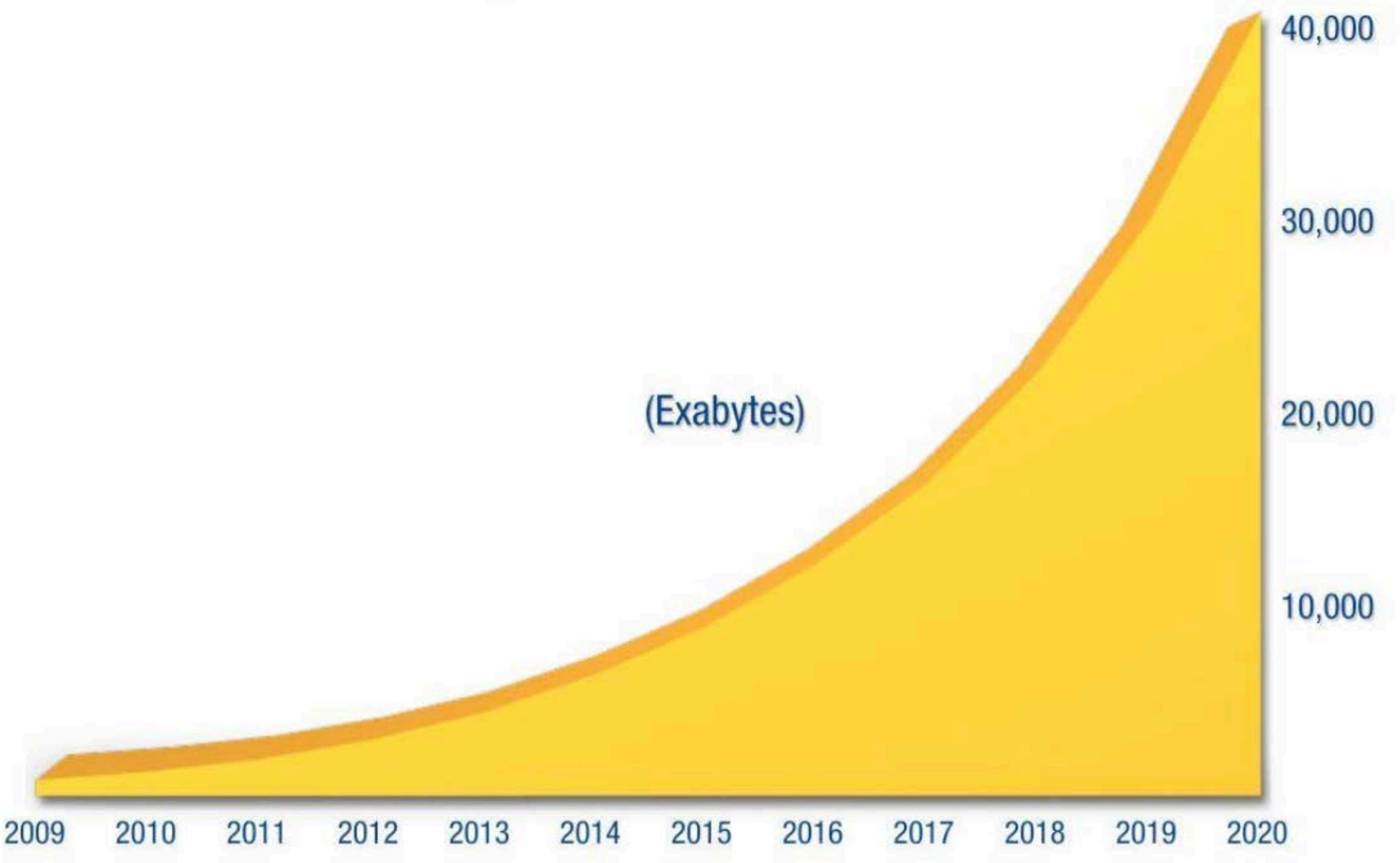
US census data, ~ 200 GB (just year 2000)

The Sizes of Digital Universe

- 2010, 1.2 ZB (zettabytes, or trillions of gigabytes)
- 2011, 1.8 ZB
- 2012, 2.8 ZB
- 2013, 4.4 ZB
-
-
-
- 2020, 44 ZB (estimated)

The Growth of the Digital Universe

The Digital Universe: 50-fold Growth from the Beginning of 2010 to the End of 2020



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Many Types of Data

- Structured data (only ~20% of the total data), for example, a table in a relational database.
- Web text data.
- Semi-structured XML data.
- Graph data (social networks, e.g., Facebook, Twitter, …).
- Streaming Data (data that is transmitted and processed in a continuous flow; you can only scan the data once).
- Image data (medical, astronomical, …).
- …

Big Data Usage?

- Unfortunately only a small fraction of it has been analyzed!
- It also provides endless opportunities.
- Data is the **new oil**.
- Mining data is the new century's **Gold Rush**?

What Now?

- Big data has caused a fundamental shift in managing/consuming data.
- The traditional approaches of data management and processing are not suitable for big data any more.
- We need a new class of capabilities to deal with the data we have had and produce today.

The Answer:

The Hadoop Platform!

What is Hadoop?

Hadoop is an open-source platform for storing and processing big data in a distributed fashion on large clusters of computers. It has rapidly emerged as the preferred solution to address business and technology trends that are disrupting traditional data management and processing. Essentially, it accomplishes two tasks:

- Massive data storage.
- Faster data processing.

The term “Hadoop” was the name of a yellow toy **elephant** owned by the son of one of its inventors (Doug Cutting).

A Brief History of Hadoop

- Based on Google's File System (GFS) and Google's MapReduce technologies, which Google published on its white pages in 2003 and 2004. GFS looks very much like HDFS (Hadoop distributed file system).
  - S. Ghemawat, H. Gobioff, and S. Leung. 2003. The Google File System
  - J. Dean and S. Ghemawat. 2004. MapReduce: Simplified Data Processing on Large Clusters
- Doug Cutting was working on an Apache open-source project called Nutch (a web crawler). In 2006 at Yahoo, with the Google's ideas and a portion of Nutch, Cutting spun out a new open-source project written in Java, which he named Hadoop.

The Characteristics of Hadoop

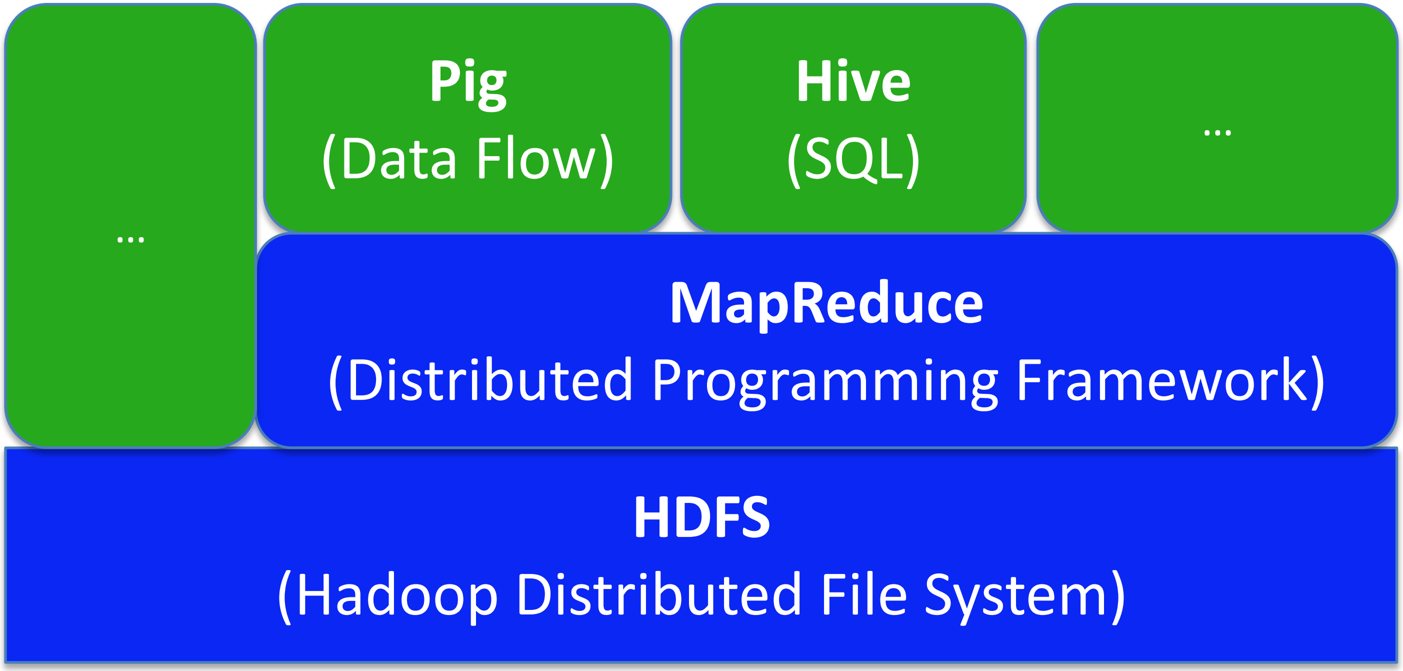
- Optimized to handle massive quantities of data (structured and/semi-structured/unstructured) using relatively low-cost commodity hardware and is good for big data.
- Great for massive parallel computing using **batch processing** but the response time is not immediate.
- Replicates data across different computers, so it can well tolerate machine failures.
- The Hadoop methodology is built around a **function-to-data** model as opposed to a data-to-function model; the analysis programs are sent to the data.

Redundancy in Hadoop

- Redundancy is built into the Hadoop environment.
  - Data redundancy.
  - Programming model redundancy.
- It allows to distribute the data and its associated programming model across a very large cluster of commodity hardware (servers).
- Failures *are expected* and are resolved ***automatically*** by running portions of the program on various servers in the cluster, i.e., it provides *hardware fault tolerance* and a capability for the Hadoop cluster to *heal itself*.

The Original Hadoop (1.0) Architecture

- Hadoop has two main components:
  - HDFS (Hadoop Distributed File System)
  - MapReduce
- Hadoop is actually a computing environment built on top of a distributed file system.

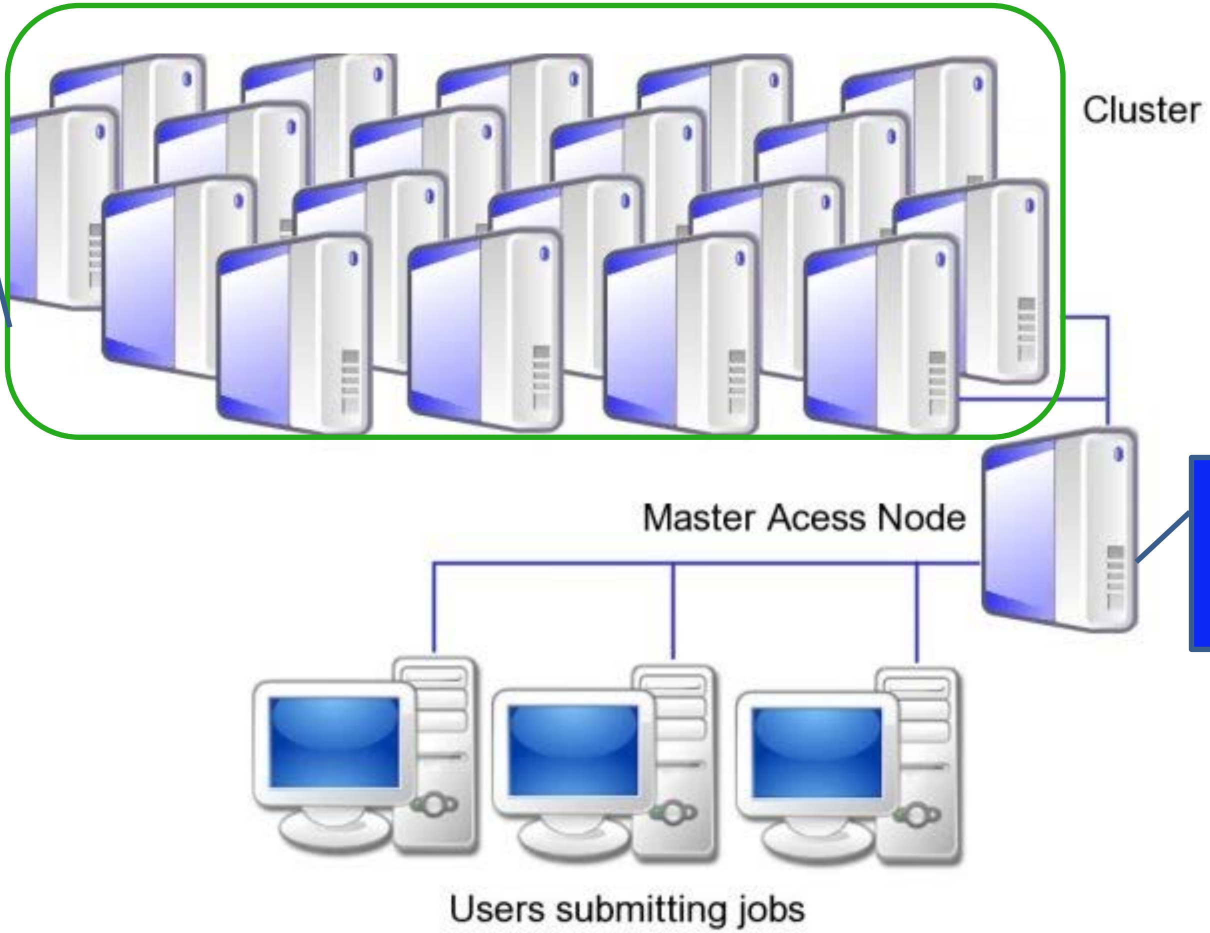

…
Pig
(Data Flow)
Hive
(SQL)
…
MapReduce
(Distributed Programming Framework)
HDFS
(Hadoop Distributed File System)

Hadoop Related Projects

- **Hive** (provides ad hoc SQL-like queries for data aggregation and summarization)
- **Pig** (a high-level Hadoop programming language that provides a data-flow language and execution framework for parallel computation)
- **Cassandra** and **HBase** (databases)
- **Avro** (for data serialization)
- **Mahout** (a machine learning library)
- **ZooKeeper** (provides coordination services for distributed applications)
- …

A Hadoop Cluster

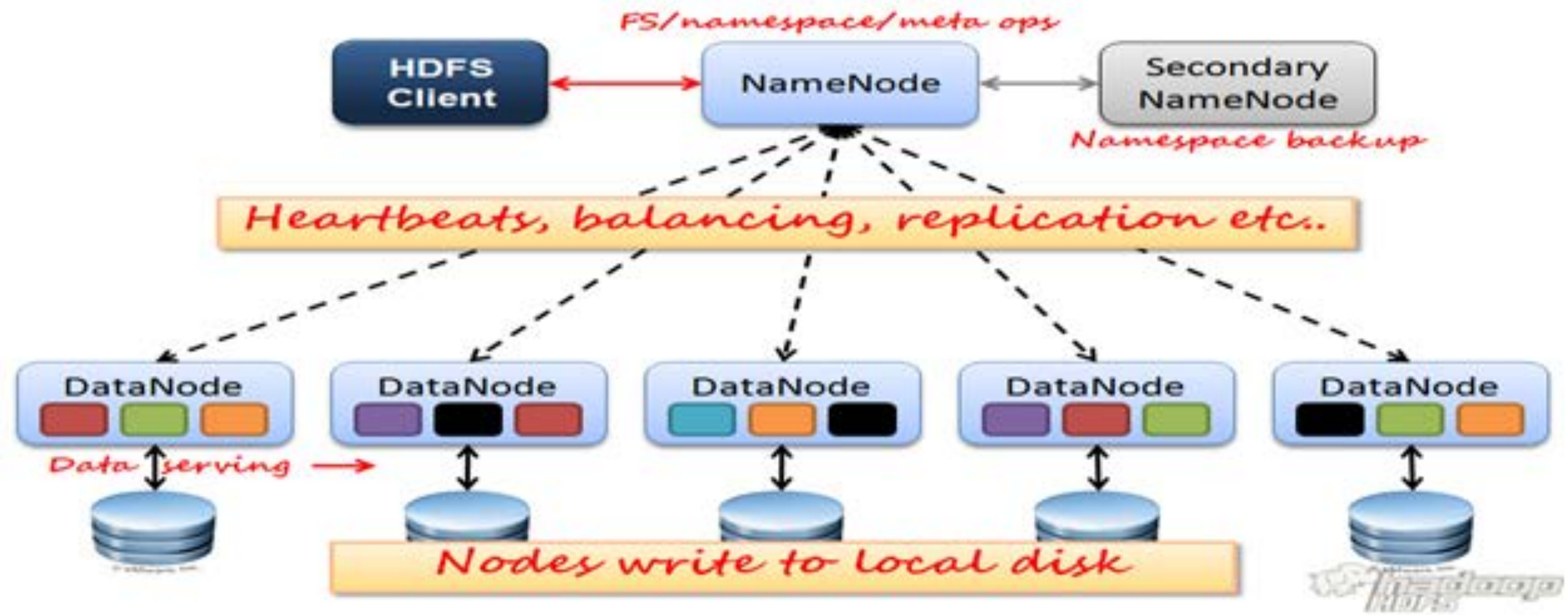
A Hadoop cluster is composed of a Name Node and a cluster of Data Nodes.



A cluster data nodes (Slaves/ workers)

Name Node (Master)

More Details of a Hadoop Cluster



Hadoop at Yahoo

The largest production user with an application running on a Hadoop cluster, consisting of approximately 10,000 Linux machines. It is also the largest contributor to the Hadoop open source project. Hadoop was created there when …

A cluster of comupters running Hadoop at Yahoo

Parallel Processing in Hadoop

- Parallel processing is Complicated:
  - How to assign tasks to different nodes?
  - How to check in on these nodes?
  - What if there are more tasks than slots?
  - How do you know if a task is still alive, dead, or successfully completed?
  - How to deal with newly submitted tasks if the system is already at the capacity? To queue them up? Where to queue them, in the memory or on a disk?
  - How to deal with reliability if some nodes goes down?
  - What happen when tasks fail?
  - How do you know if a task fails?
  - How to handle distributed synchronization?
  - …

Important Features that Hadoop Provides

- Redundant, fault-tolerant data storage.
- Parallel computation framework.
- Job coordination.
- It frees the developers from having to worry about the parallel mechanism. The Hadoop system deals with these automatically.
- One operator can take care of a cluster of thousands of nodes.

Hadoop is Adopted by

- Amazon
- AOL
- AT&T
- China Mobile
- eBay
- ETSY
- Facebook
- Foursquare
- GE
- HP
- Google
- IBM
- LinkedIn
- Microsoft
- NASA
- Netflix
- SAS
- Twitter
- Yahoo!
- …

Hadoop in Action: Example 1

In the media industry **The New York Times** : it wanted to host on their website all public domain articles from 1851 to 1922.

They converted articles from 11 million image files to 1.5TB of PDF documents only by one employee who ran a job in 24 hours on a 100-instance Amazon EC2 Hadoop cluster at a very low cost.

Hadoop in Action: Example 2

In the telecommunications industry **China Mobile** : It produces about 5-8 TB of Call Data Records daily. A Hadoop cluster was built to perform data mining on the data.

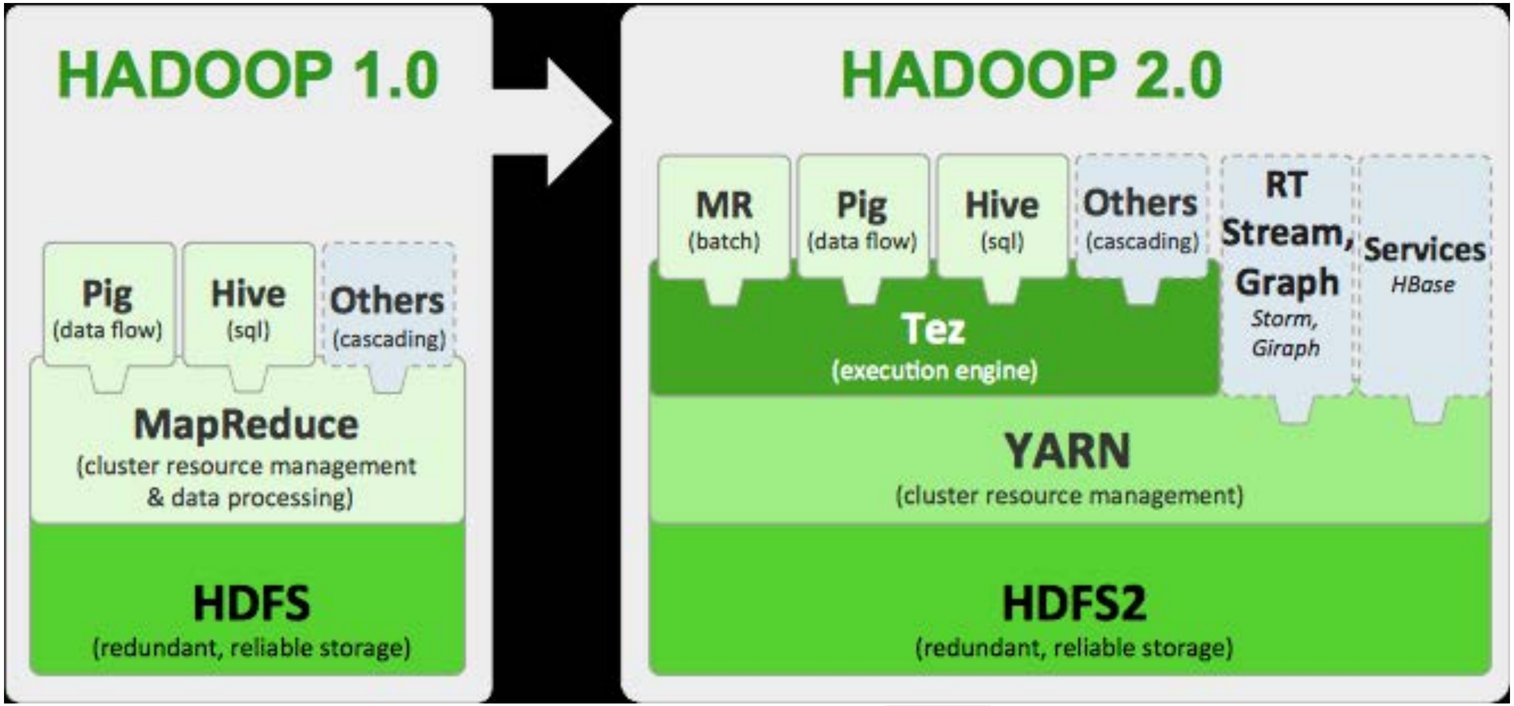
中国移动通信
CHINA MOBILE

They were able to process 10 times as much data at one fifth of the cost as when using their old system.

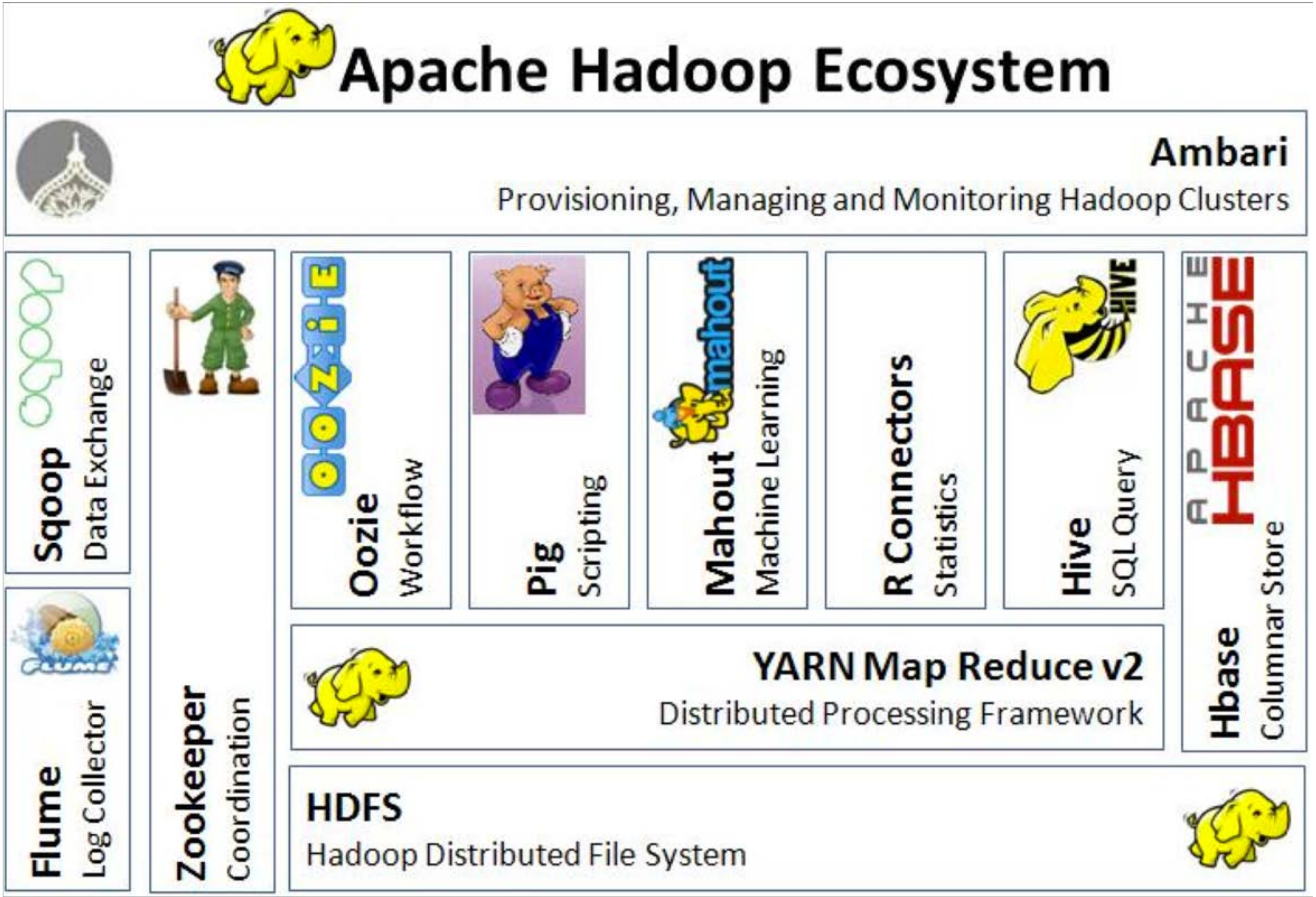
Hadoop is NOT a Magic Bullet

- Currently Hadoop alone is **not good**
  - for processing small data files
  - when work cannot be parallelized
  - low latency data access
  - intensive calculations with little data (other systems are more efficient)
  - real- time warehousing, or blazing transactional speeds
  - online transaction processing where data are randomly accessed on structured data, often a relational database (Hadoop is not a replacement for a relational database system)

Developments in Hadoop



HADOOP 3.0



Apache Hadoop Ecosystem

Ambari
Provisioning, Managing and Monitoring Hadoop Clusters

Sqoop
Data Exchange

Flume
Log Collector

Zookeeper
Coordination

Oozie
Workflow

Pig
Scripting

Mahout
Machine Learning

R Connectors
Statistics

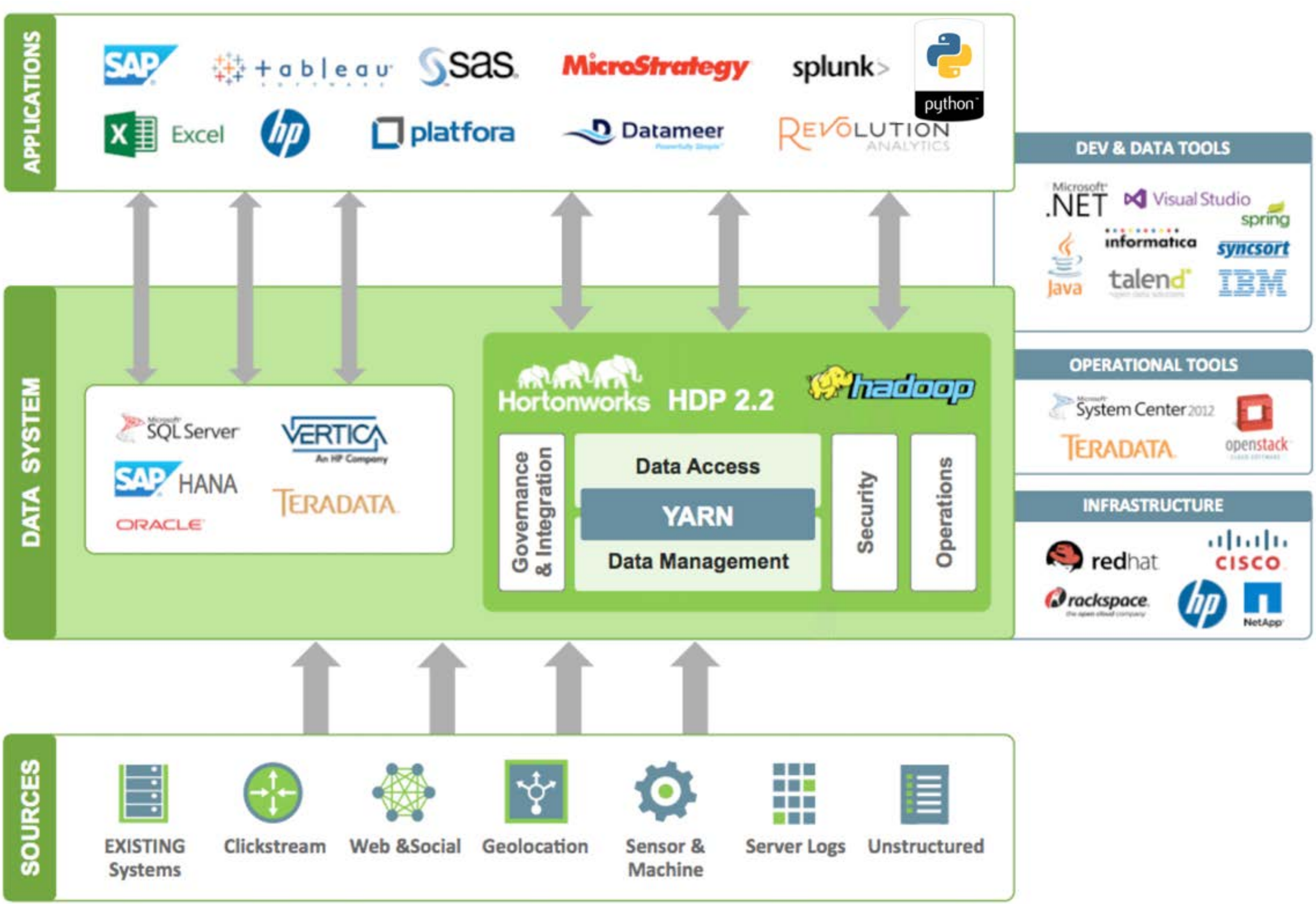
Hive
SQL Query

Hbase
Columnar Store

YARN Map Reduce v2
Distributed Processing Framework

HDFS
Hadoop Distributed File System

Hadoop in a Modern Data Architecture



Main Hadoop Distribution Vendors

cloudera
MAPR™
TECHNOLOGIES
Hortonworks