BIG DATA & HADOOP OVERVIEW

BIG DATA & HADOOP OVERVIEW

Agenda

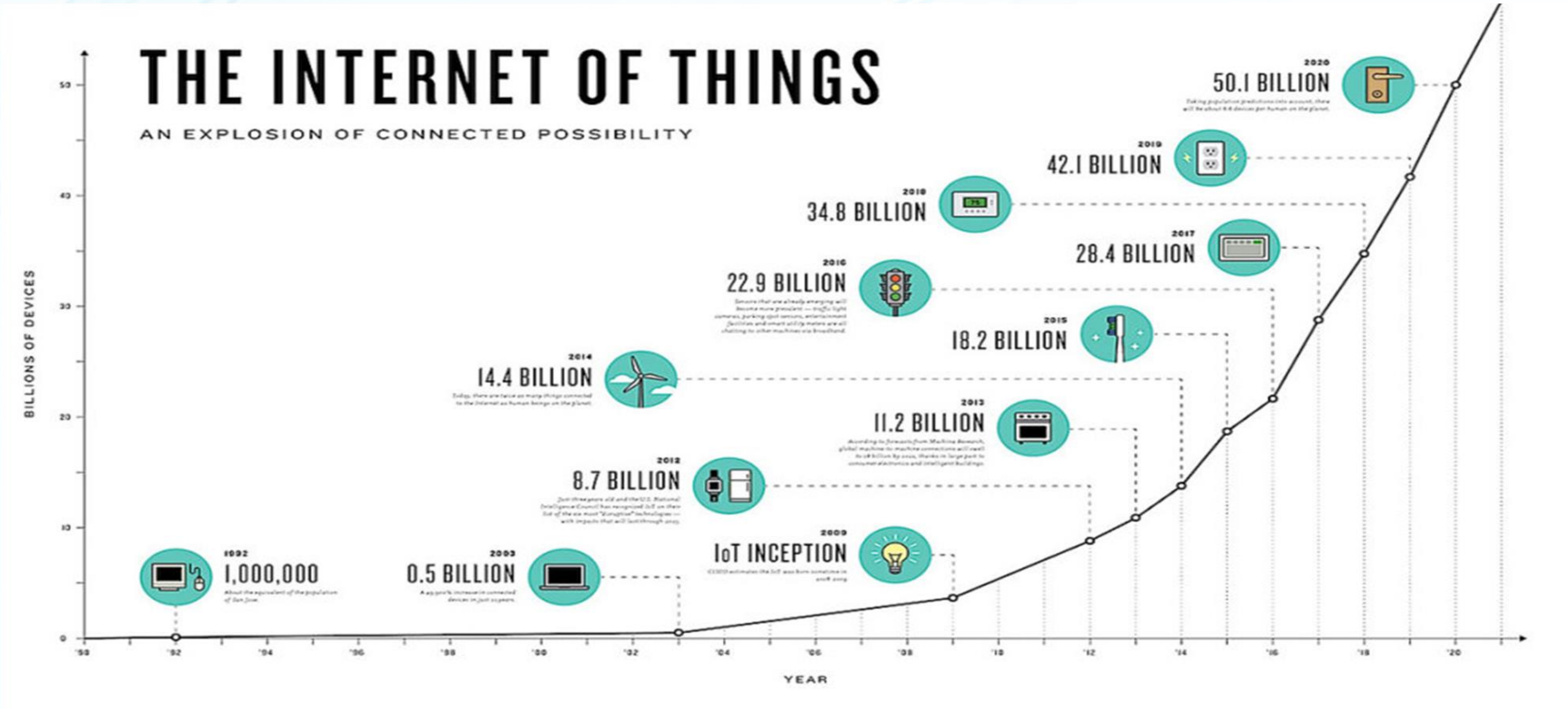
- ✓ Data Explosion
- ✓ Understanding Big Data
- ✓ Big Data Challenges
- ✓ Big Data Opportunity
- ✓ Hadoop as a solution to Big Data
- ✓ Advantages of Hadoop
- ✓ Hadoop Ecosystem

Drivers of Data Explosion - Smart Devices

As devices such as phones and house appliances getting smarter, they also generate huge amount of data.

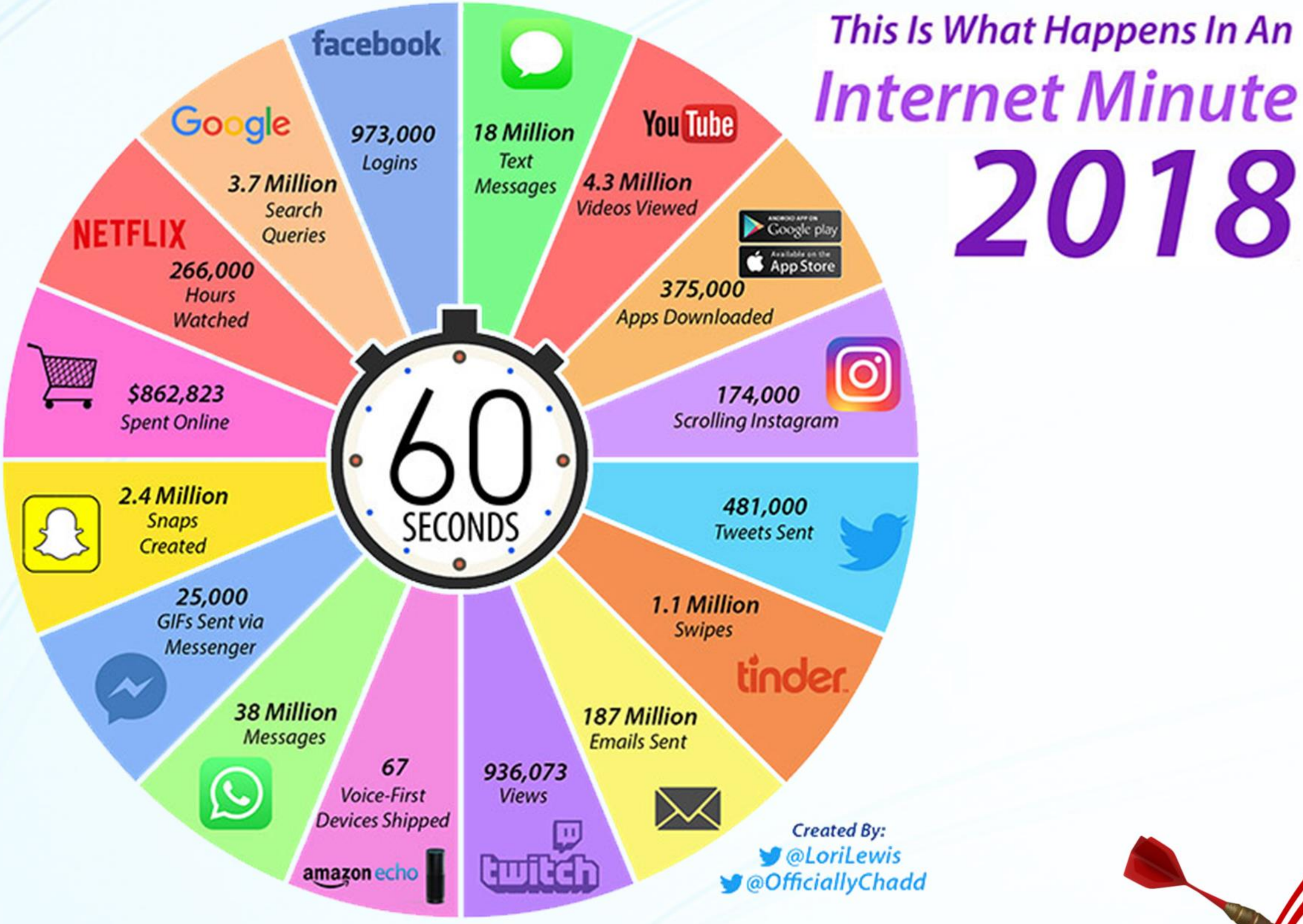
- Smart Phones
- Smart Watches
- Smart Appliances
- Self Driving Cars

Drivers of Data Explosion – IoT

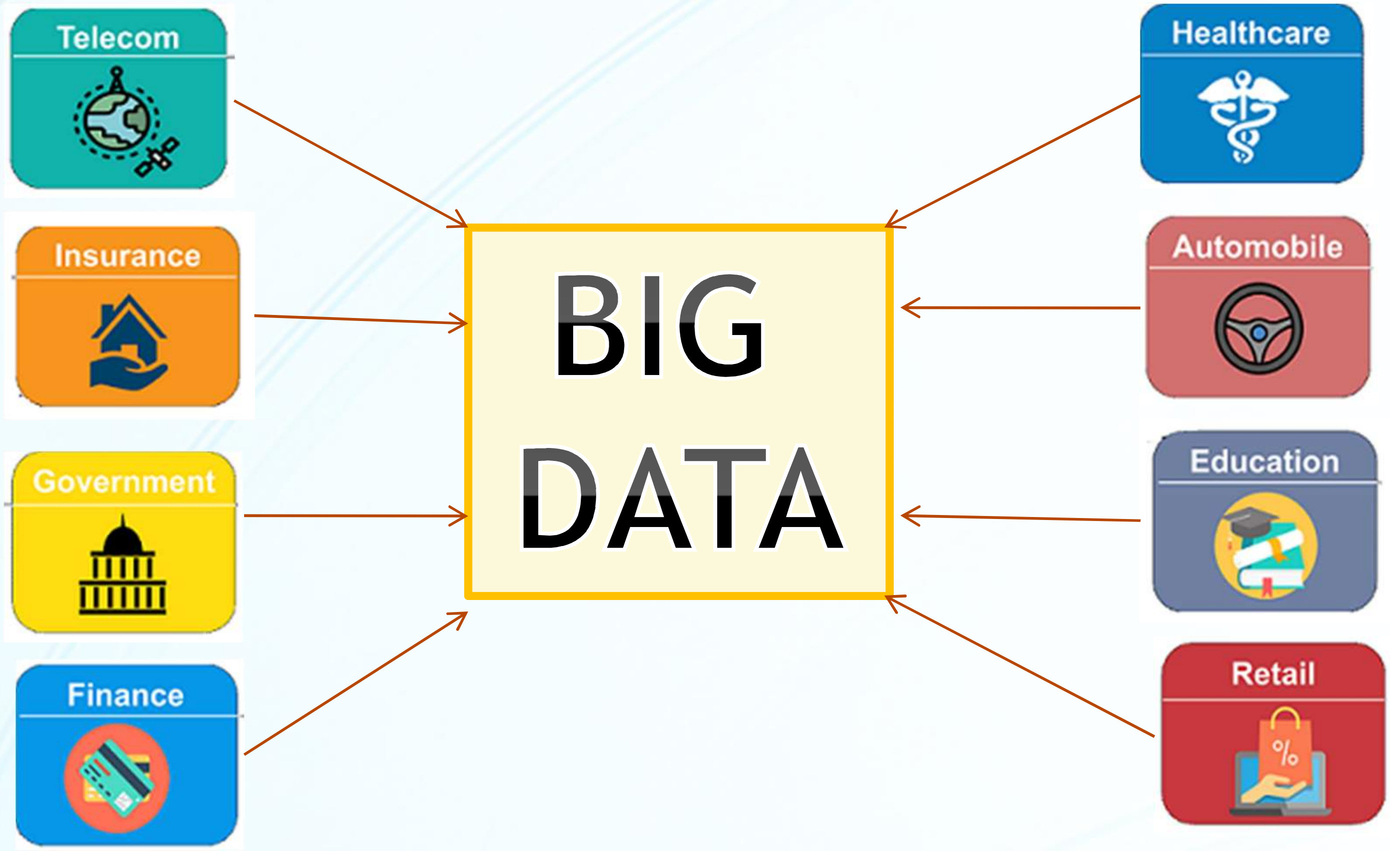


By 2020, there will be 50 billion IoT devices i.e. roughly *6 things online per person* causing huge amount of data getting generated.

Drivers of Data Explosion – Social Media



Drivers of Data Explosion - Others



Telecom

Insurance

Government

Finance

BIG DATA

Healthcare

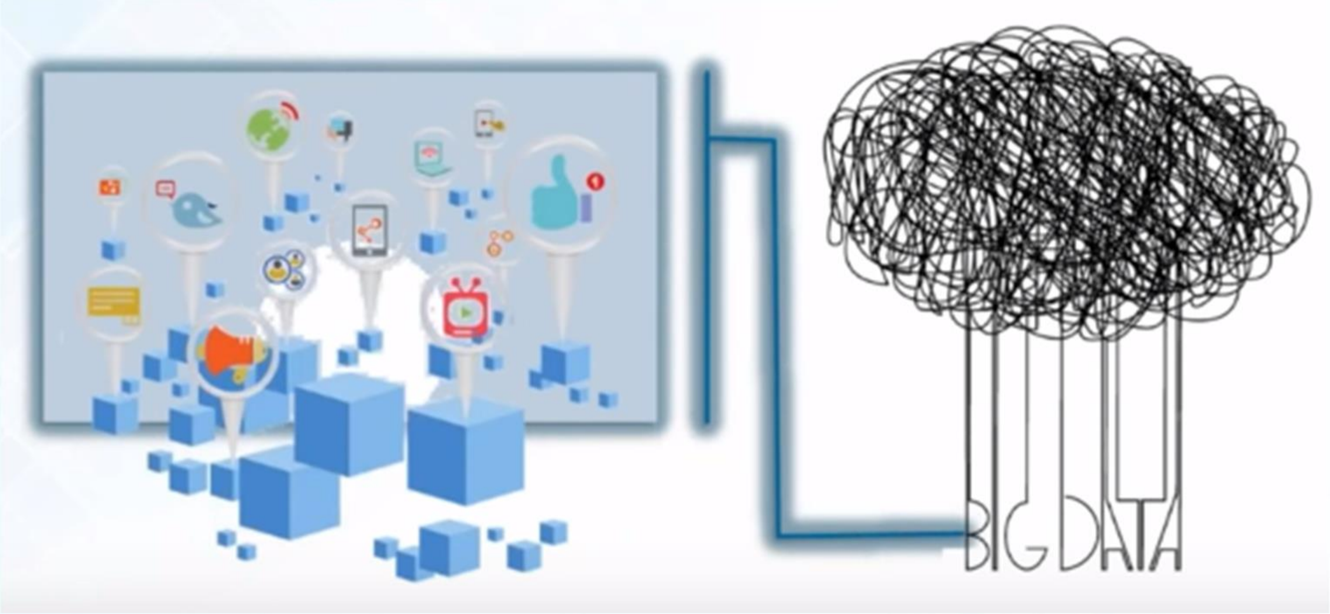
Automobile

Education

Retail

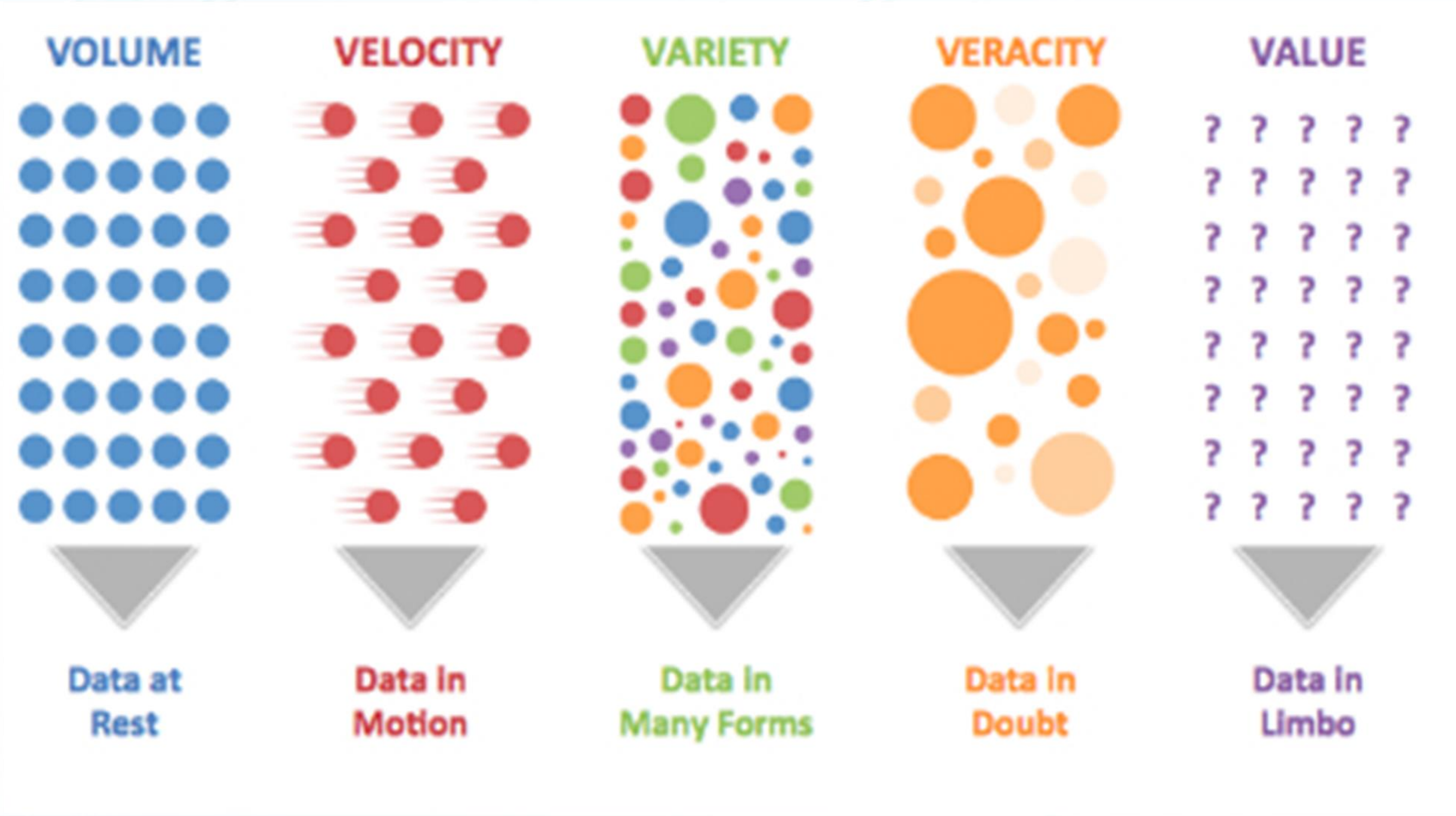
What is Big Data ?

Big Data is a term for a collection of data sets so **large and complex**, that it becomes **difficult to store and process** using on-hand database management tools or traditional data processing applications



How do we classify some data as Big Data?

5 Vs of Big Data



VOLUME

Data at Rest

VELOCITY

Data in Motion

VARIETY

Data in Many Forms

VERACITY

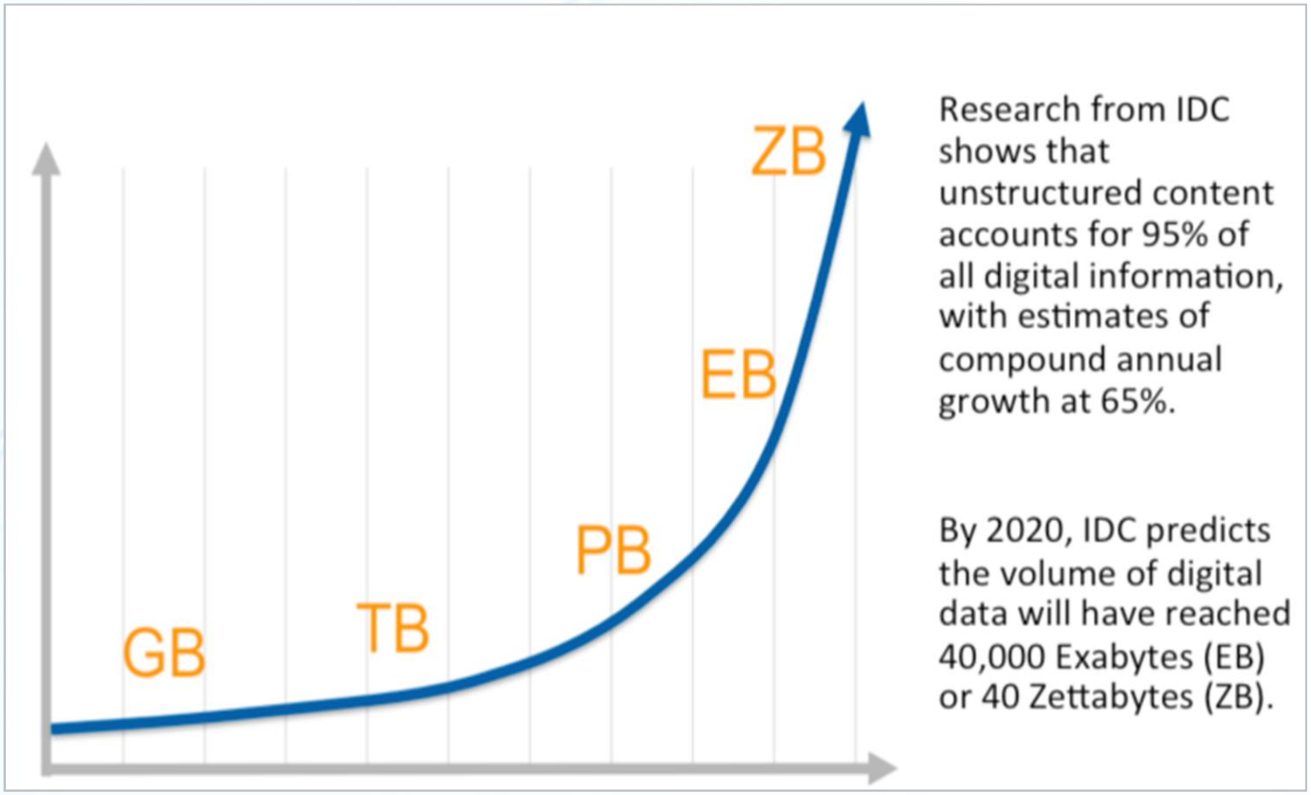
Data in Doubt

VALUE

Data in Limbo

Volume

1 Volume



Research from IDC shows that unstructured content accounts for 95% of all digital information, with estimates of compound annual growth at 65%.

By 2020, IDC predicts the volume of digital data will have reached 40,000 Exabytes (EB) or 40 Zettabytes (ZB).

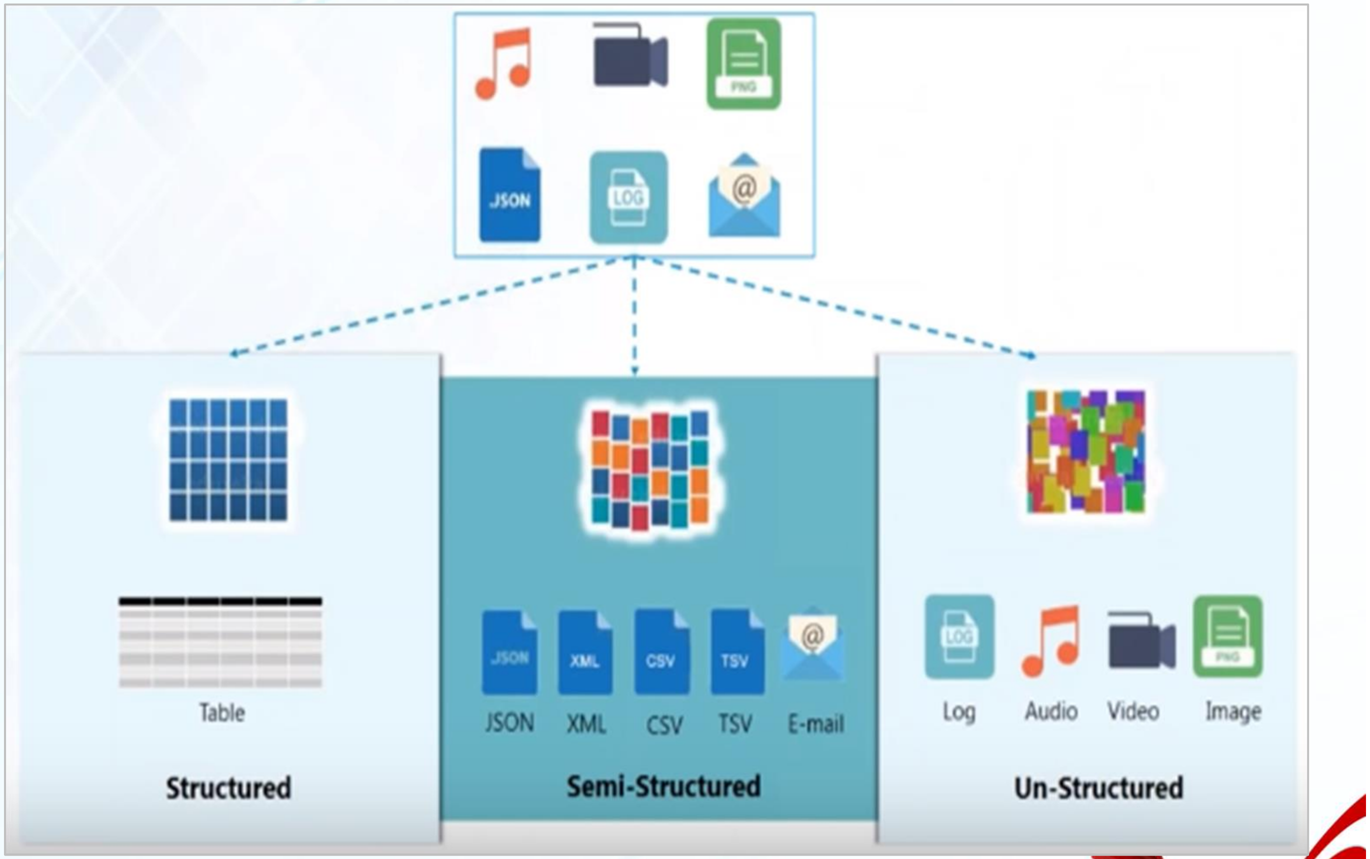
Volume Growth of Unstructured Data

Variety

1 Volume

2 Variety

Data comes in different formats:
Structured, Semi-structured and Unstructured



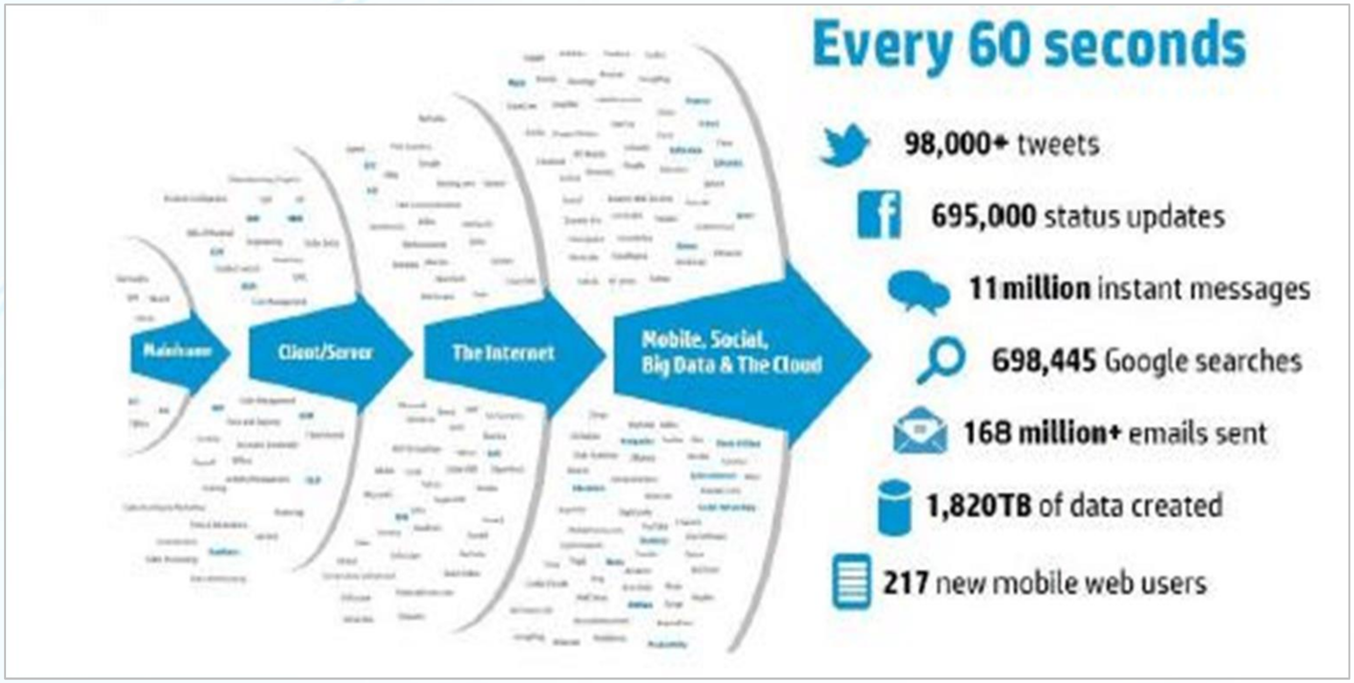
Velocity

1 Volume

2 Variety

3 Velocity

Data is being generated at a very high rate



Every 60 seconds

98,000+ tweets

695,000 status updates

11 million instant messages

698,445 Google searches

168 million+ emails sent

1,820TB of data created

217 new mobile web users

Value

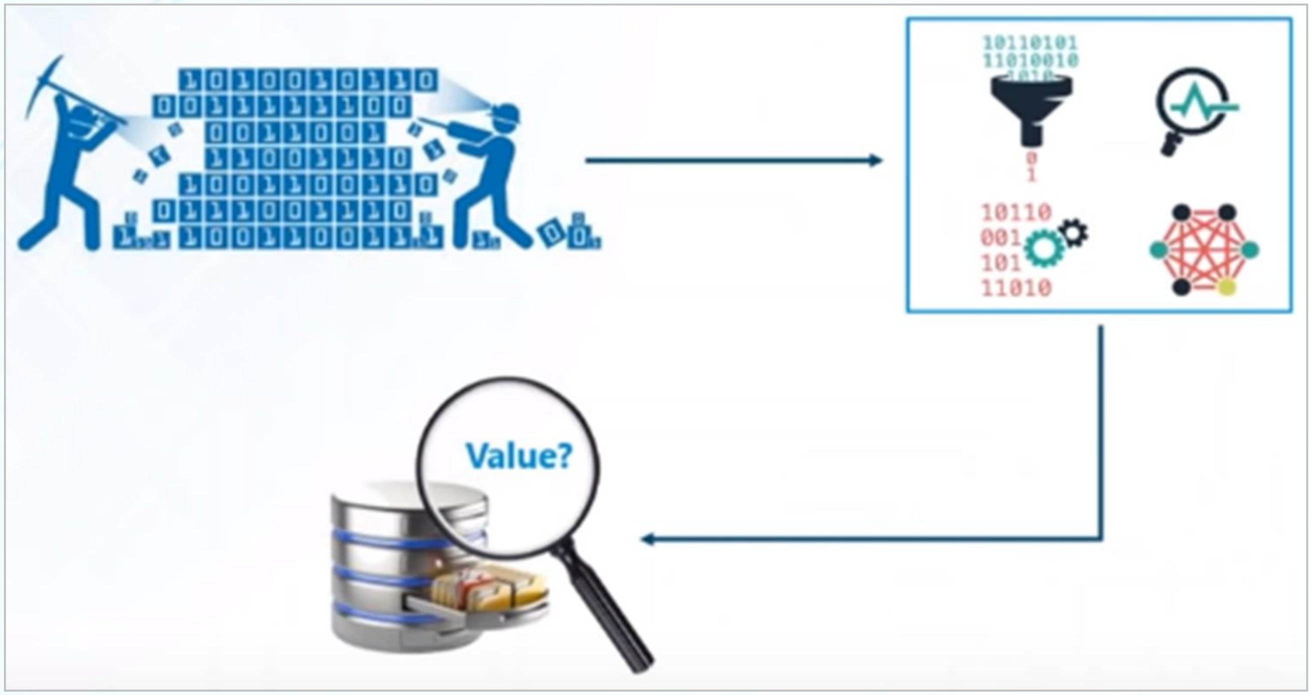
1 Volume

2 Variety

3 Velocity

4 Value

Value refers to the extraction of correct meaning out of the data.



Veracity

1 Volume

2 Variety

3 Velocity

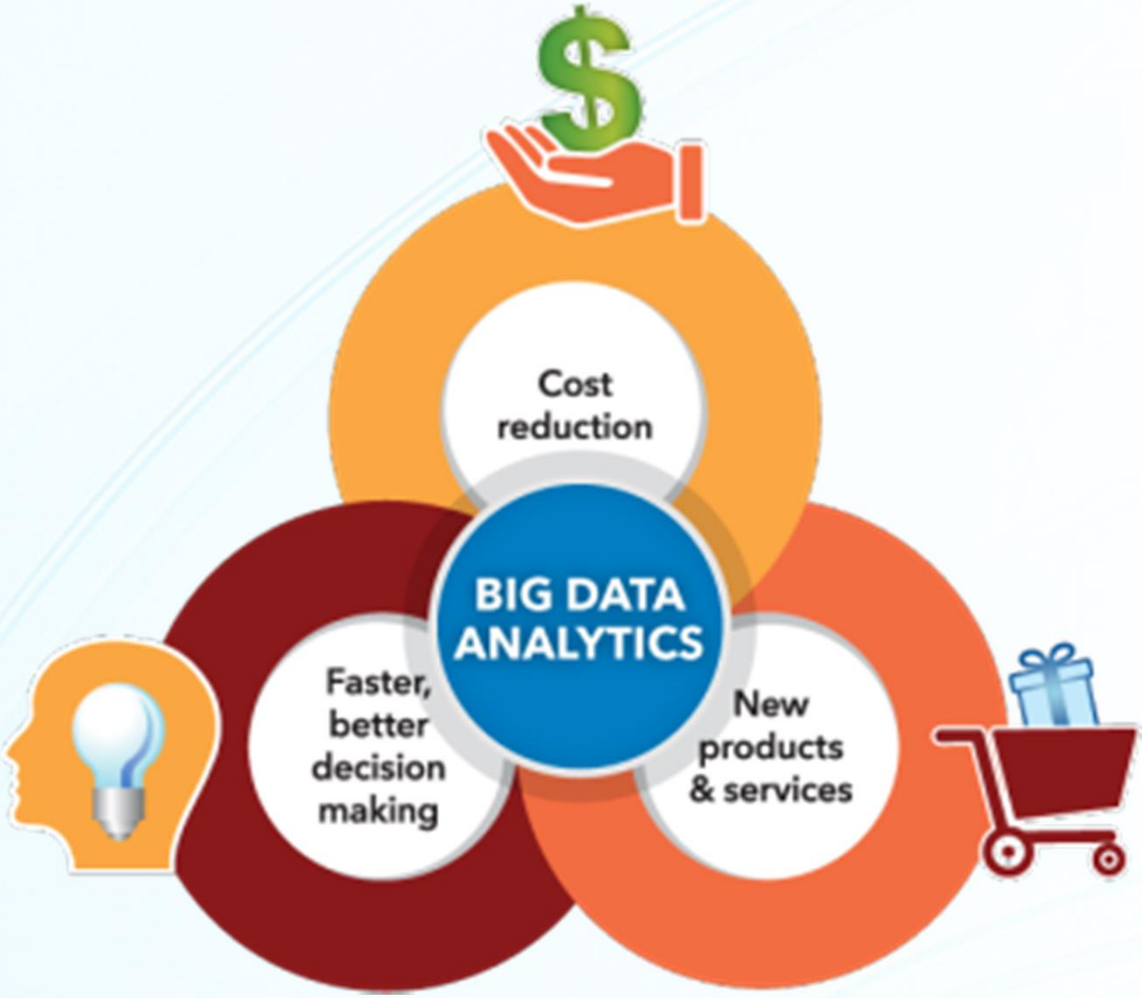
4 Value

5 Veracity

Veracity refers to correctness of data. Data can be ambiguous, uncertain and missing.

Max	Min	Mean	Std. Dev
5.7	4.5	4.8	0.44
10.5	?	11.2	0.32
75000	5.4	9.5	1000
1.5	0.5	?	0.6

Opportunity: Big Data Analytics



BIG DATA ANALYTICS

Cost reduction

Faster, better decision making

New products & services

Problems with Big Data

1. Storage

- Storing exponentially growing huge datasets using traditional ways is a big problem.

2. Processing complex data

- Processing the data that has complex structure is a problem as the data comes as unstructured and semi-structured forms.

3. Processing data faster

- The data is growing at much faster rate than disk IO speeds. Bringing huge amount of data to the computation unit becomes a bottleneck.

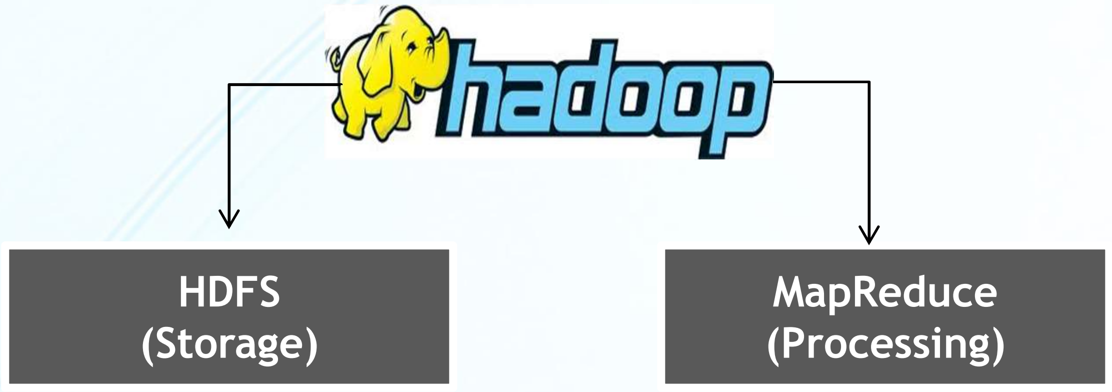
What is the solution to all these Big Data Problems ?

Hadoop is the solution

So lets dive in and look at what Hadoop is all about.

Hadoop

Hadoop is an open-source software framework for storage and processing of datasets on clusters of commodity hardware.



Allows to store any kind of data in a cluster in a distributed fashion

Allows parallel processing of data stored in HDFS

History of Hadoop

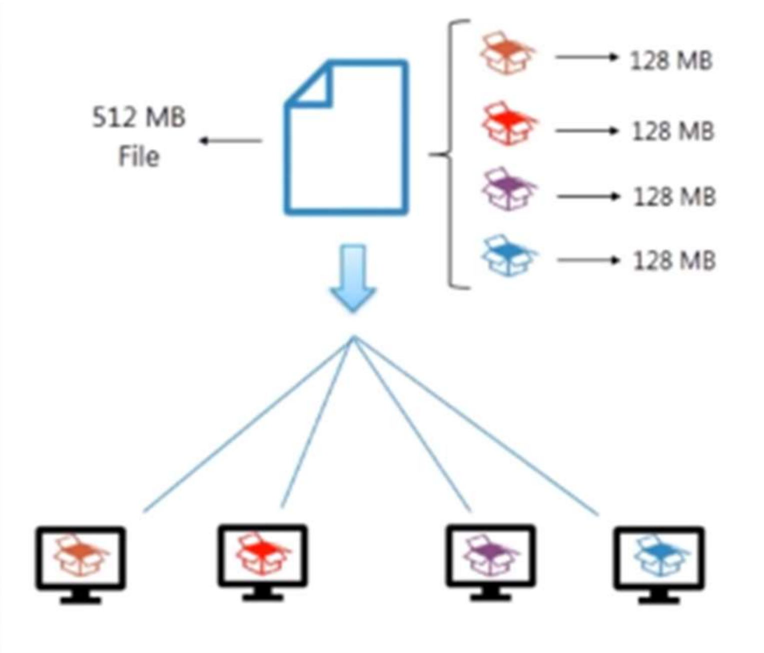
2008	• Hadoop became Apache Top Level Project
2006	• Yahoo! hires Doug Cutting to work on Hadoop with a dedicated team
2005	• Doug Cutting and Nutch team implemented Google's frameworks in Nutch
2004	• Google publishes Google File System (GFS) and MapReduce framework papers

The Hadoop Solution

Problem 1: Storing exponentially growing huge datasets

Solution: HDFS

- Hadoop's storage unit
- Divides files into blocks
- Stores blocks across the cluster
- Replicates blocks as a fail-safe mechanism
- Horizontally Scalable



The Hadoop Solution

Problem 2: Storing unstructured data

Solution: HDFS

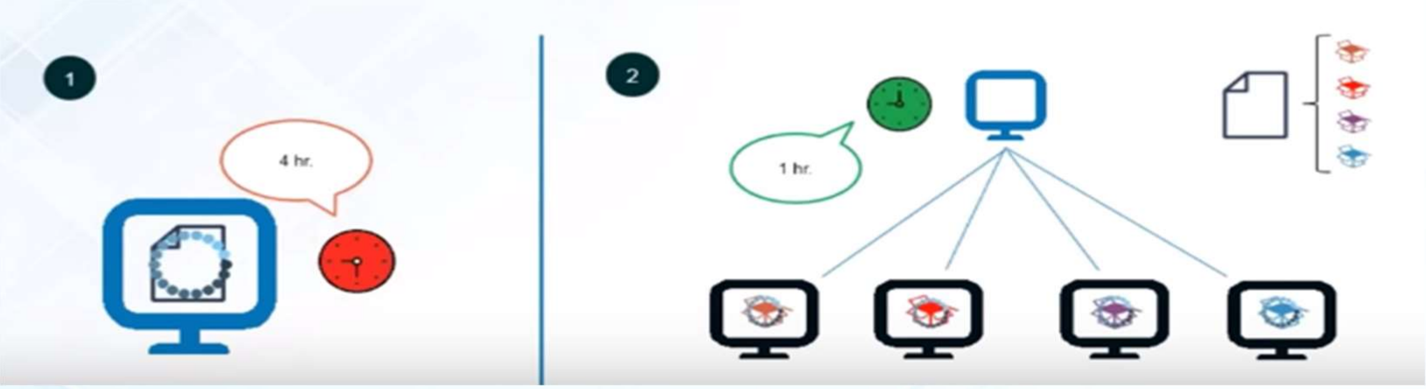
- HDFS allows to store any kind of data – be it structured or not.
- No schema validation done while writing data
- Follows write once and read many times parading (WORM)

The Hadoop Solution

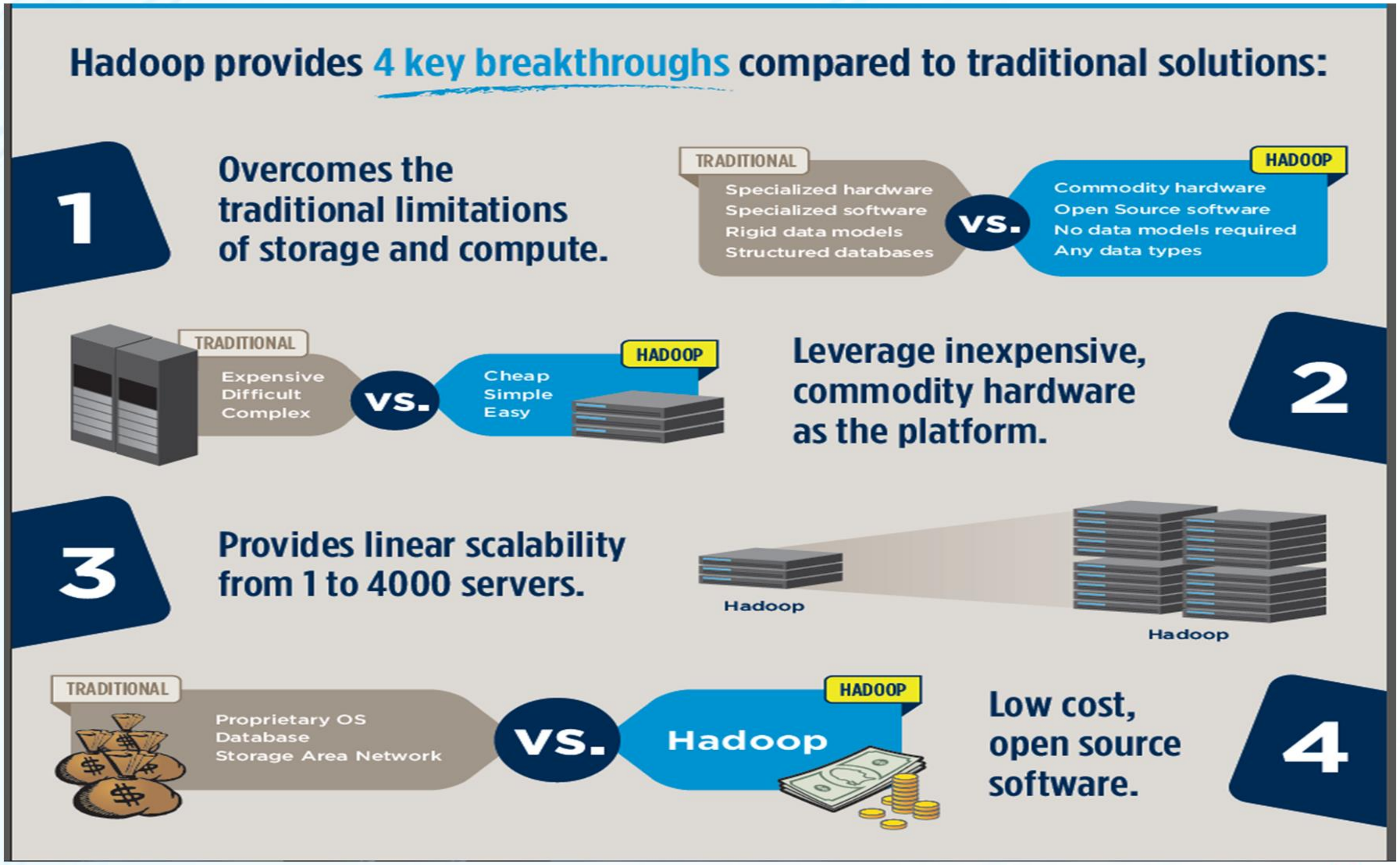
Problem 3: Processing data faster

Solution: MapReduce

- Provides parallel processing of data present in HDFS
- Allows data to be processed locally on each node making use of its resources.
- Brings processing to the data



The Hadoop Solution



Hadoop provides 4 key breakthroughs compared to traditional solutions:

1 Overcomes the traditional limitations of storage and compute.

2 Leverage inexpensive, commodity hardware as the platform.

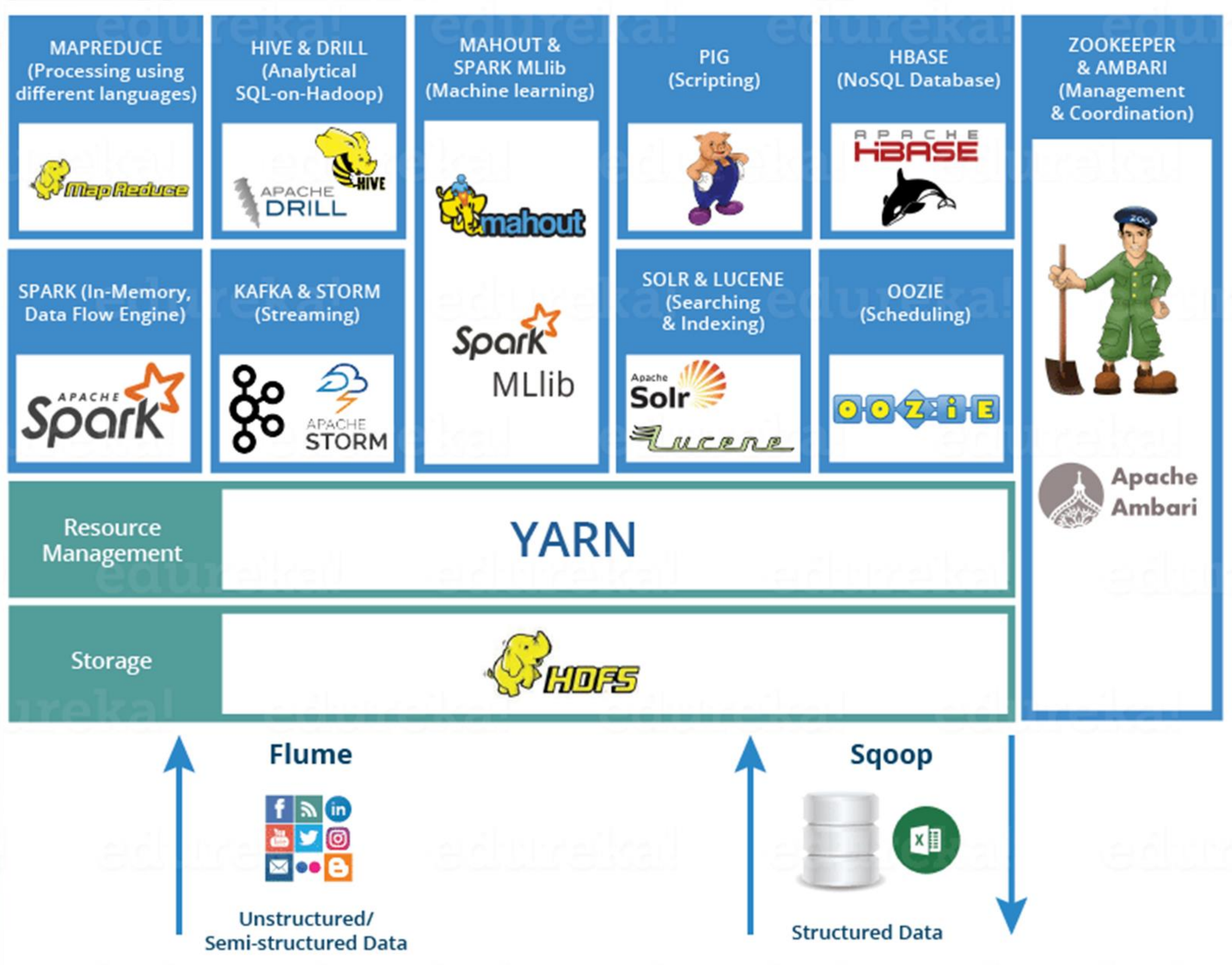
3 Provides linear scalability from 1 to 4000 servers.

4 Low cost, open source software.

Advantages of Hadoop

- Scalable
- Available
- Reliable
- Cost Effective
- Flexible
- Fast
- Fail Safe

Hadoop Ecosystem



MAPREDUCE (Processing using different languages)

HIVE & DRILL (Analytical SQL-on-Hadoop)

MAHOUT & SPARK MLlib (Machine learning)

PIG (Scripting)

HBASE (NoSQL Database)

ZOOKEEPER & AMBARI (Management & Coordination)

SPARK (In-Memory, Data Flow Engine)

KAFKA & STORM (Streaming)

SOLR & LUCENE (Searching & Indexing)

OOZIE (Scheduling)

Resource Management

YARN

Storage

Flume

Unstructured/ Semi-structured Data

Sqoop

Structured Data

Hadoop Ecosystem

Hadoop distributed storage and processing framework for handling big data problems

Apache Hive is a data warehousing tool that provides an SQL like interface to perform data analytics

Apache Pig is a data flow platform for analysing very large datasets. Pig runs on HDFS and MapReduce clusters.

Apache Spark is an in-memory data processing engine for executing batch, streaming, machine learning and SQL workloads.

Apache HBase is a NoSQL database that uses HDFS for its underlying storage, and supports both batch-style computations using MapReduce and point queries.

Hadoop Ecosystem

Apache Sqoop is a tool for efficiently moving structured data between relational databases and HDFS.

Apache Flume is a distributed service for efficiently collecting, aggregating, and moving large amounts of streaming data.

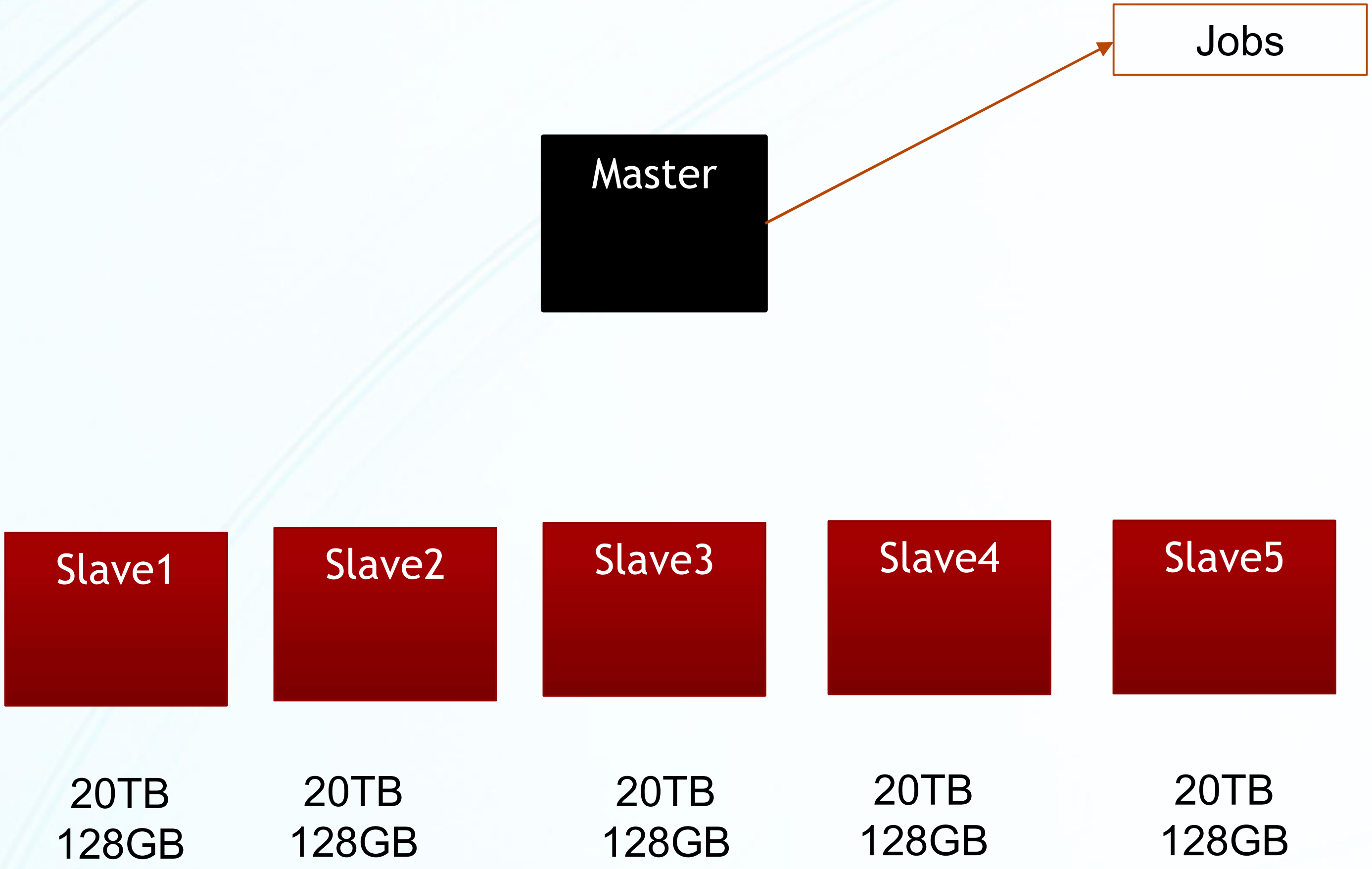
Apache Oozie is a workflow scheduler system to manage Hadoop jobs.

Apache Zookeeper is a distributed coordination service that provides primitives such as distributed locks that can be used for building distributed applications.

Some Big Data Use Cases for Modern Business

- Log Analytics.
- E-Commerce Personalization.
- Recommendation Engines.
- Automated Candidate Placement in Recruiting.
- Insurance Fraud Detection.
- Relevancy and Retention Boost for Online Publishing.
- 360° View of the Customer
- Security Intelligence
- Price Optimization
- Social Media Analysis and Response
- Preventive Maintenance and Support

TIME FOR QUESTIONS
12
1
2
3
4



Jobs

Master

Slave1

Slave2

Slave3

Slave4

Slave5

20TB
128GB

20TB
128GB

20TB
128GB

20TB
128GB

20TB
128GB