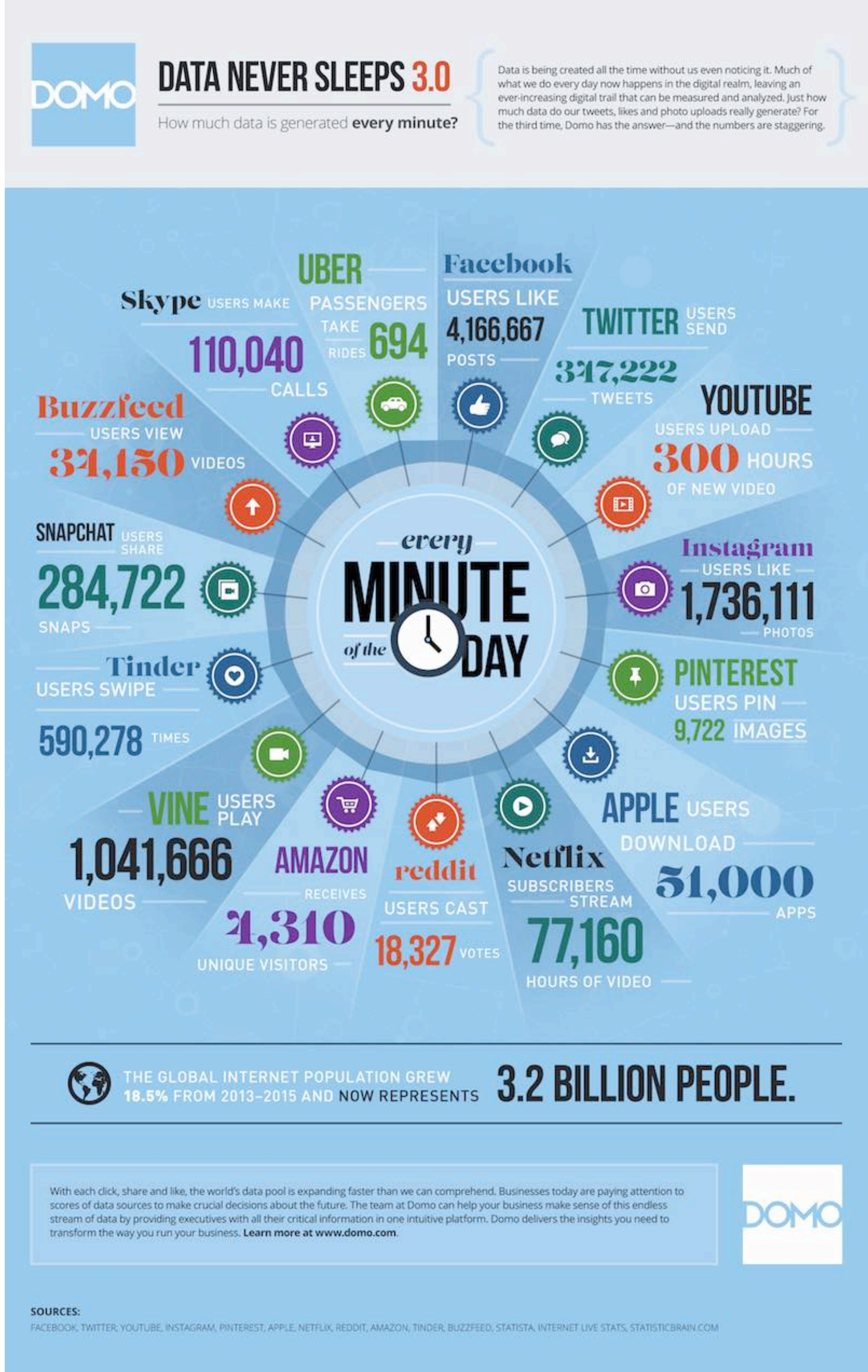
Data Engineering

204426

Big Data

Big Data

- Too large + Complex
- Big data was originally associated with three key concepts: volume, variety, and velocity.



DATA NEVER SLEEPS 3.0

How much data is generated **every minute?**

Data is being created all the time without us even noticing it. Much of what we do every day now happens in the digital realm, leaving an ever-increasing digital trail that can be measured and analyzed. Just how much data do our tweets, likes and photo uploads really generate? For the third time, Domo has the answer—and the numbers are staggering.

DOMO
every MINUTE of the DAY
Skype USERS MAKE 110,040 CALLS
UBER PASSENGERS TAKE 694 RIDES
Facebook USERS LIKE 4,166,667 POSTS
TWITTER USERS SEND 347,222 TWEETS
YOUTUBE USERS UPLOAD 300 HOURS OF NEW VIDEO
Buzzfeed USERS VIEW 34,450 VIDEOS
SNAPCHAT USERS SHARE 284,722 SNAPS
Instagram USERS LIKE 1,736,111 PHOTOS
Tinder USERS SWIPE 590,278 TIMES
PINTEREST USERS PIN 9,722 IMAGES
VINE USERS PLAY 1,041,666 VIDEOS
AMAZON RECEIVES 4,310 UNIQUE VISITORS
reddit USERS CAST 18,327 VOTES
Netflix SUBSCRIBERS STREAM 77,160 HOURS OF VIDEO
APPLE USERS DOWNLOAD 51,000 APPS
THE GLOBAL INTERNET POPULATION GREW 18.5% FROM 2013–2015 AND NOW REPRESENTS 3.2 BILLION PEOPLE.
DOMO

With each click, share and like, the world's data pool is expanding faster than we can comprehend. Businesses today are paying attention to scores of data sources to make crucial decisions about the future. The team at Domo can help your business make sense of this endless stream of data by providing executives with all their critical information in one intuitive platform. Domo delivers the insights you need to transform the way you run your business. **Learn more at www.domo.com.**

SOURCES:
FACEBOOK, TWITTER, YOUTUBE, INSTAGRAM, PINTEREST, APPLE, NETFLIX, REDDIT, AMAZON, TINDER, BUZZFEED, STATISTA, INTERNET LIVE STATS, STATISTICBRAIN.COM

3Vs Properties

Volume

- Scale of data
- A huge amount of data
- If the volume of data is very large then it is actually considered as a ‘Big Data’
- Terabyte/Petabyte/Exabyte

Variety

- Different form of data
- Different function of data
- Different data sources

Velocity

- High speed of accumulation of data
- A massive and continuous flow of data.
- The potential of data that how fast the data is generated and processed to meet the demands.

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF 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

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

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

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

Velocity

ANALYSIS OF STREAMING DATA

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

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

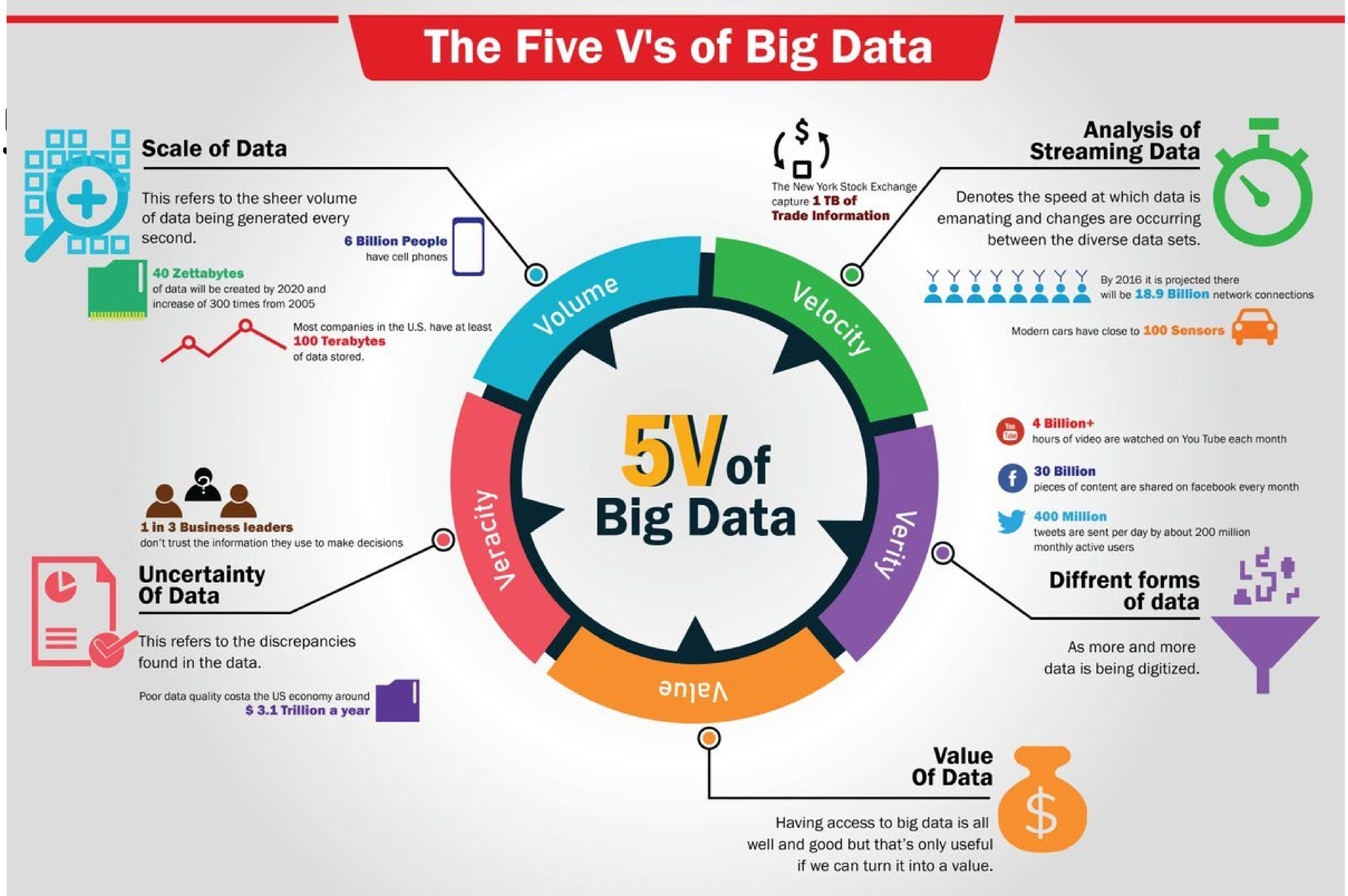
$3.1 TRILLION A YEAR

27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM



The Five V's of Big Data

Scale of Data

This refers to the sheer volume of data being generated every second.

6 Billion People have cell phones

40 Zettabytes of data will be created by 2020 and increase of 300 times from 2005

Most companies in the U.S. have at least **100 Terabytes** of data stored.

Analysis of Streaming Data

Denotes the speed at which data is emanating and changes are occurring between the diverse data sets.

The New York Stock Exchange capture **1 TB of Trade Information**

By 2016 it is projected there will be **18.9 Billion** network connections

Modern cars have close to **100 Sensors**

Diffrent forms of data

As more and more data is being digitized.

4 Billion+ hours of video are watched on You Tube 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

Uncertainty Of Data

This refers to the discrepancies found in the data.

1 in 3 Business leaders don't trust the information they use to make decisions

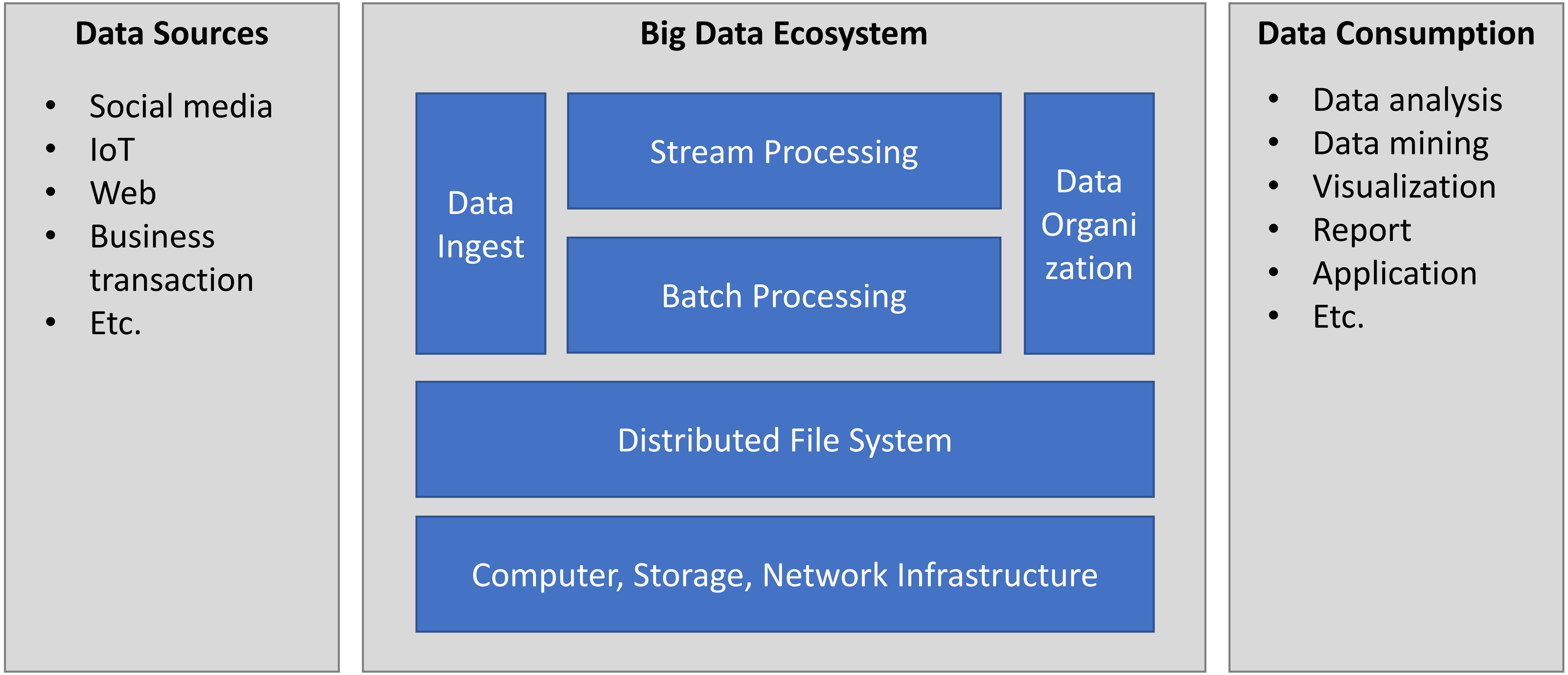
Poor data quality costa the US economy around **$ 3.1 Trillion a year**

Value Of Data

Having access to big data is all well and good but that's only useful if we can turn it into a value.

Source: https://morioh.com/p/ca19c6b8c0fe

Components of the Big Data Ecosystem



Data Sources

- Social media
- IoT
- Web
- Business transaction
- Etc.

Big Data Ecosystem

Data Ingest

Stream Processing

Batch Processing

Data Organi zation

Distributed File System

Computer, Storage, Network Infrastructure

Data Consumption

- Data analysis
- Data mining
- Visualization
- Report
- Application
- Etc.

Components of the Big Data Ecosystem

Data Ingestion

- Transportation of data from assorted sources to a storage medium.
- **Batch processing**
  - Ingestion layer periodically collects and groups source data and sends it to the destination system.
  - Groups may be processed based on any logical ordering, the activation of certain conditions, or a simple schedule.
- **Stream processing**
  - Data is sourced, manipulated, and loaded as soon as it’s created or recognized by the data ingestion layer.

Data Organization

- Database - the method of classifying and organizing data sets to make them more useful.

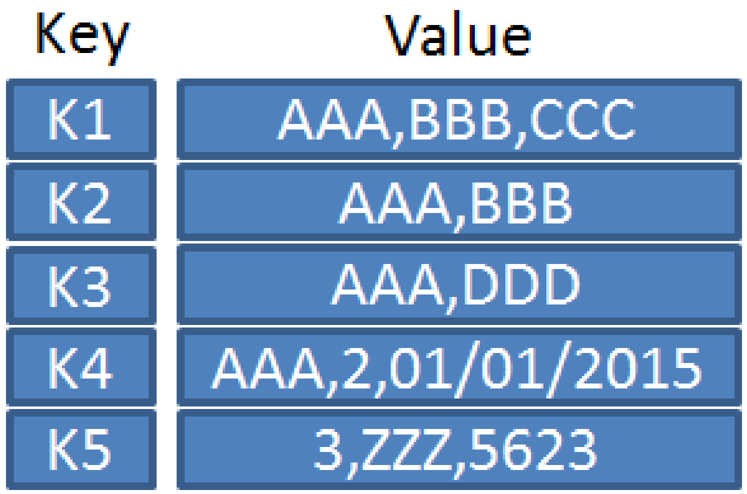
NoSQL

- Non-tabular databases and store data differently than relational tables.
- Types of NoSQL databases:
  - Key-Value Store Databases
  - Document Store Databases
  - Graph Databases
  - Column-Oriented Databases

NoSQL

Key-Value Store Databases

- Data is represented as a collection of key–value pairs
- The key–value model can be extended to a discretely ordered model that maintains keys in lexicographic order.
- Example:
  - DynamoDB
  - Voldemort
  - Redis

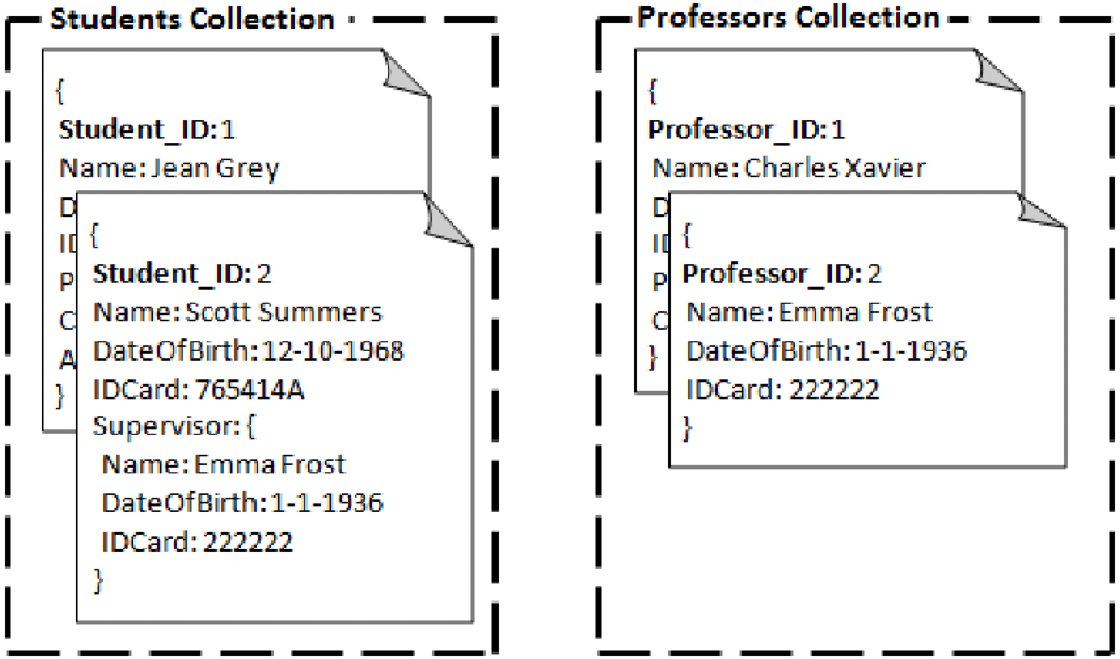


Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

NoSQL

Document Store Databases

- Documents encapsulate and encode data (or information) in some standard format or encoding.
- Encodings in use include XML, YAML, JSON
- Use indexing to read/write data in form of document object.
- Example
  - MongoDB
  - CouchDB



Students Collection

```
{
Student_ID:1
Name: Jean Grey
D
ID
P
C
A
}
```

```
{
Student_ID: 2
Name: Scott Summers
DateOfBirth: 12-10-1968
IDCard: 765414A
Supervisor: {
 Name: Emma Frost
 DateOfBirth: 1-1-1936
 IDCard: 222222
}
```

Professors Collection

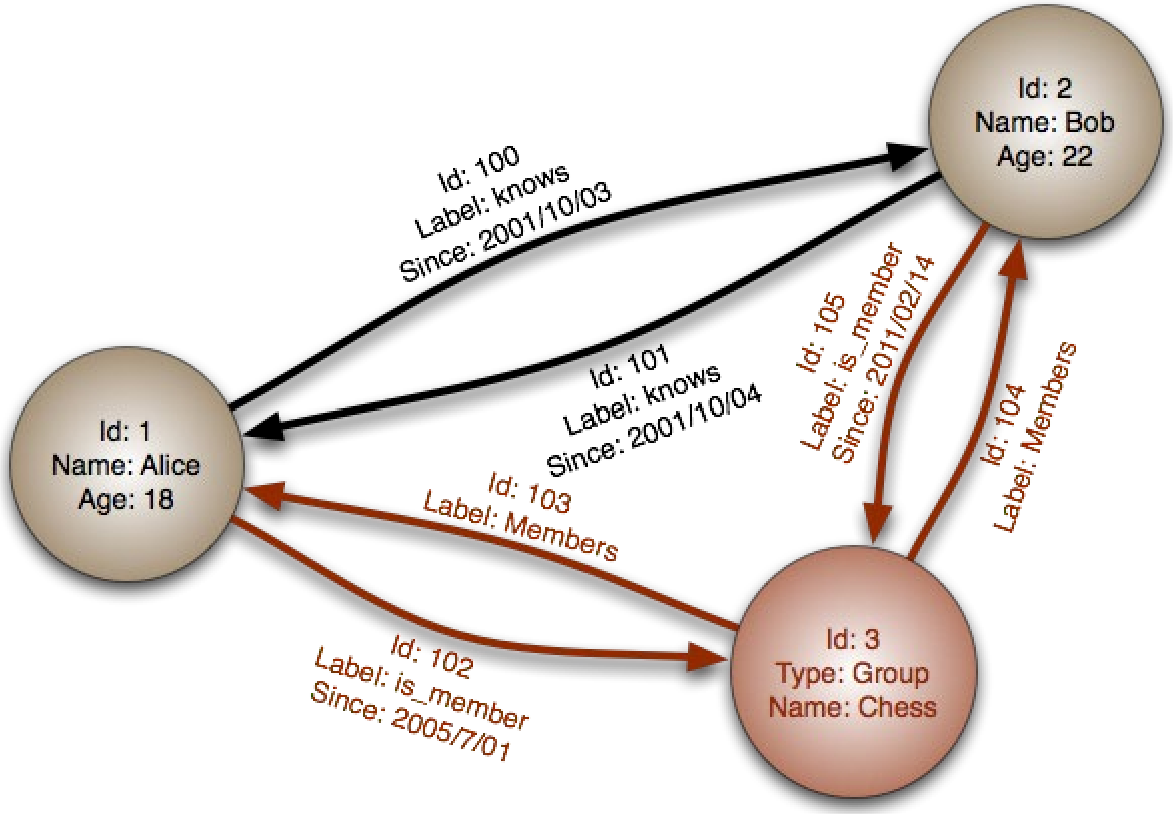
```
{
Professor_ID: 1
Name: Charles Xavier
D
ID
P
C
}
```

```
{
Professor_ID: 2
Name: Emma Frost
DateOfBirth: 1-1-1936
IDCard: 222222
}
```

NoSQL

Graph Databases

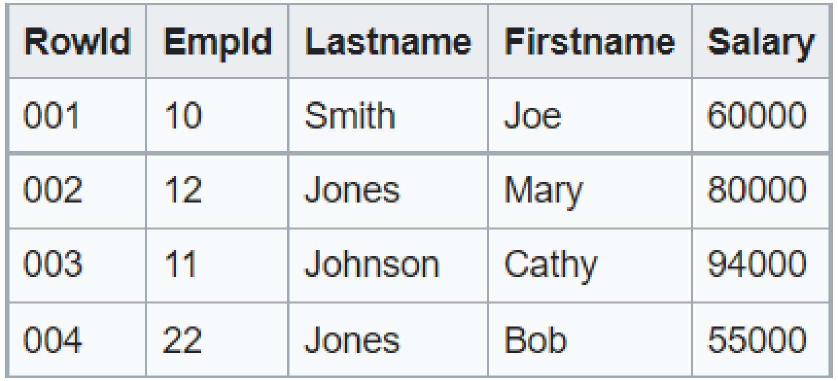
- Use graph structures for semantic queries with nodes, edges, and properties to represent and store data.
- The graph relates the data items in the store to a collection of nodes and edges, the edges representing the relationships between the nodes.
- Example
  - Neo4J
  - FlockDB



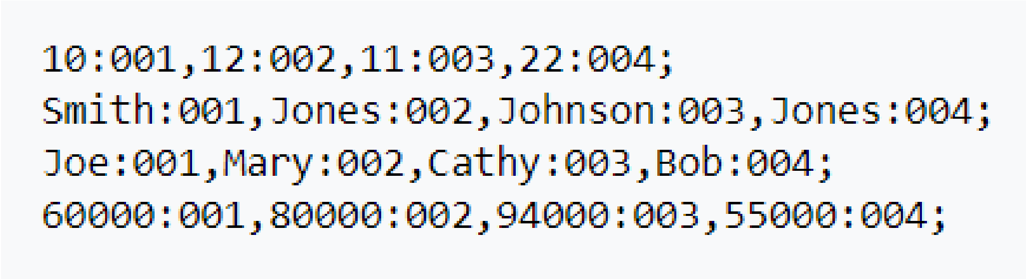
NoSQL

Column-Oriented Databases

- Stores data tables by column rather than by row.
- By storing data in columns rather than rows, the database can more precisely access the data it needs to answer a query rather than scanning and discarding unwanted data in rows.
- Stores each column continuously. i.e. on disk or in-memory each column on the left will be stored in sequential blocks.



RowId	EmpId	Lastname	Firstname	Salary
001	10	Smith	Joe	60000
002	12	Jones	Mary	80000
003	11	Johnson	Cathy	94000
004	22	Jones	Bob	55000

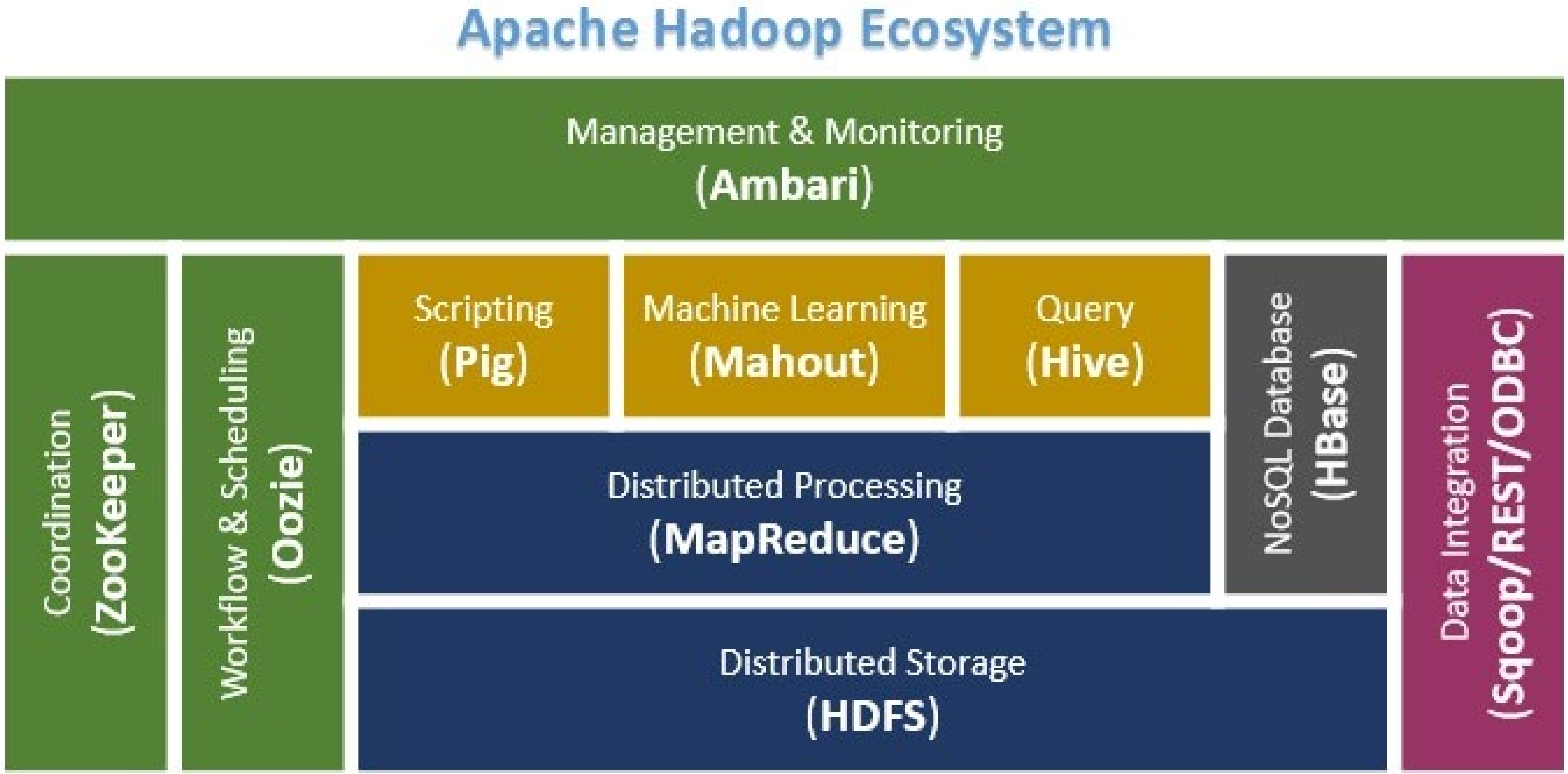


```
10:001,12:002,11:003,22:004;
Smith:001,Jones:002,Johnson:003,Jones:004;
Joe:001,Mary:002,Cathy:003,Bob:004;
60000:001,80000:002,94000:003,55000:004;
```

Distributed File System

- File system that is distributed on multiple file servers or multiple locations.
- Allows programs to access or store isolated files as they do with the local ones, allowing programmers to access files from any network or computer.

Hadoop: A bigdata framework



Apache Hadoop Ecosystem

Management & Monitoring
(**Ambari**)

Coordination
(**ZooKeeper**)

Workflow & Scheduling
(**Oozie**)

Scripting
(**Pig**)

Machine Learning
(**Mahout**)

Query
(**Hive**)

Distributed Processing
(**MapReduce**)

NoSQL Database
(**HBase**)

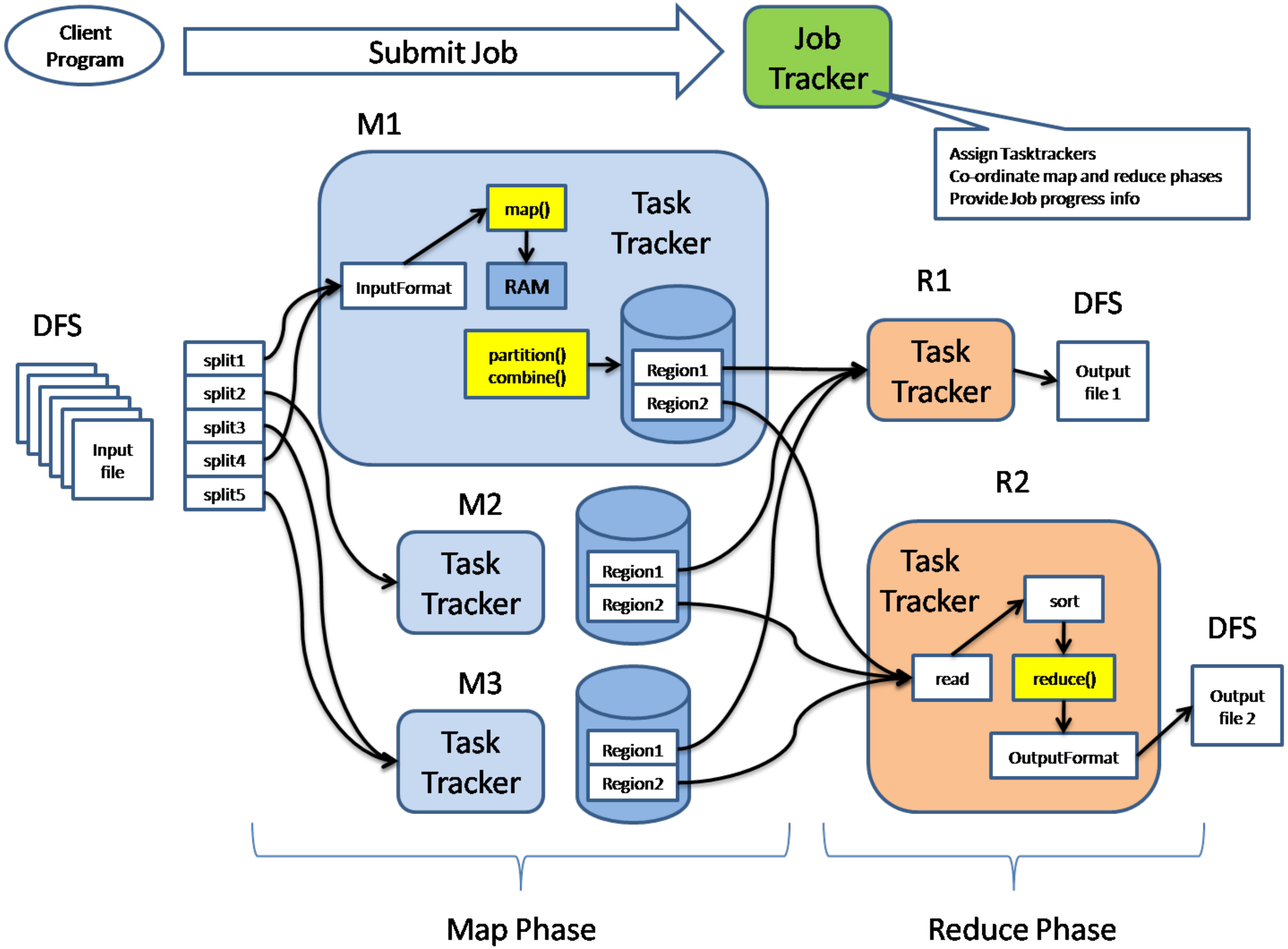
Distributed Storage
(**HDFS**)

Data Integration
(**Sqoop/REST/ODBC**)

MapReduce

- Enhances the processing of massive data using dispersed and parallel algorithms in the Hadoop ecosystem.
- Process large datasets across computer clusters.
- Two primary tasks in MapReduce: Map and Reduce
- In the map job
  - Split the input dataset into chunks.
  - Task processes these chunks in parallel.
  - Use outputs as inputs for the reduce tasks.
- For reducer
  - Process the intermediate data from the maps into smaller tuples, that reduces the tasks, leading to the final output of the framework.

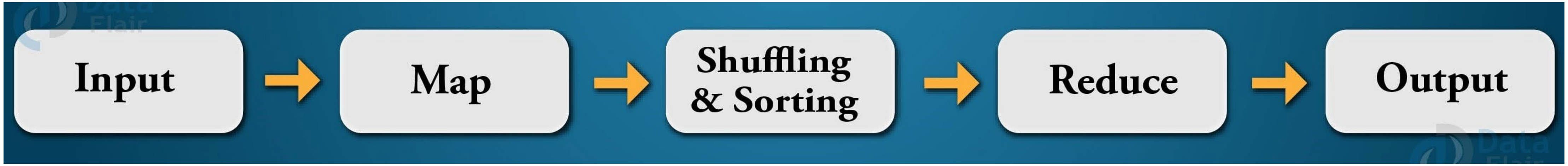
MapReduce



Source: http://a4academics.com/images/hadoop/Hadoop-Mapreduce-Architecture.png

MapReduce

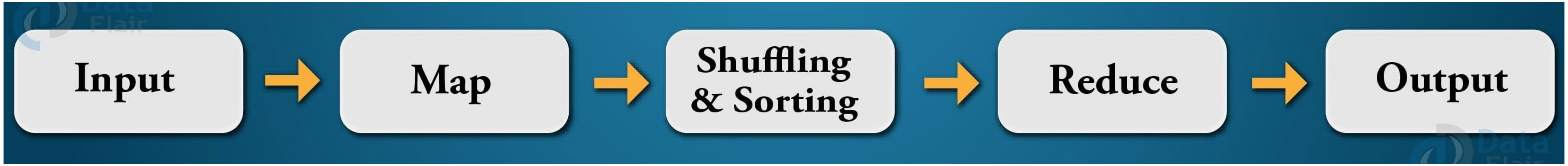
Simplified flow diagram for the MapReduce program



- A dataset is split into equal units called chunks (input splits) in the splitting step.
- Hadoop consists of a RecordReader that uses TextInputFormat to transform input splits into key-value pairs.
- The key-value pairs are then used as inputs in the mapping step.
- The mapping step contains a coding logic that is applied to these data blocks.
- The mapper processes the key-value pairs and produces an output of the same form (key-value pairs).

MapReduce

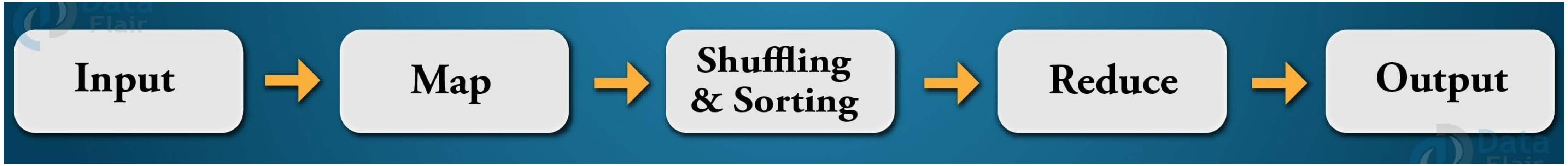
Simplified flow diagram for the MapReduce program



- It consists of two main steps: sorting and merging.
- In the sorting step, the key-value pairs are sorted using the keys. Merging ensures that key-value pairs are combined.
- The shuffling phase facilitates the removal of duplicate values and the grouping of values.
- Different values with similar keys are grouped. The output of this phase will be keys and values, just like in the Mapping phase.

MapReduce

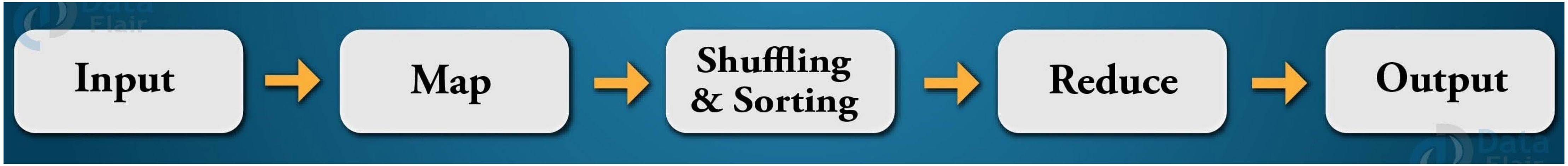
Simplified flow diagram for the MapReduce program



- The output of the shuffling phase is used as the input.
- The reducer processes this input further to reduce the intermediate values into smaller values.
- It provides a summary of the entire dataset. The output from this phase is stored in the HDFS.

MapReduce

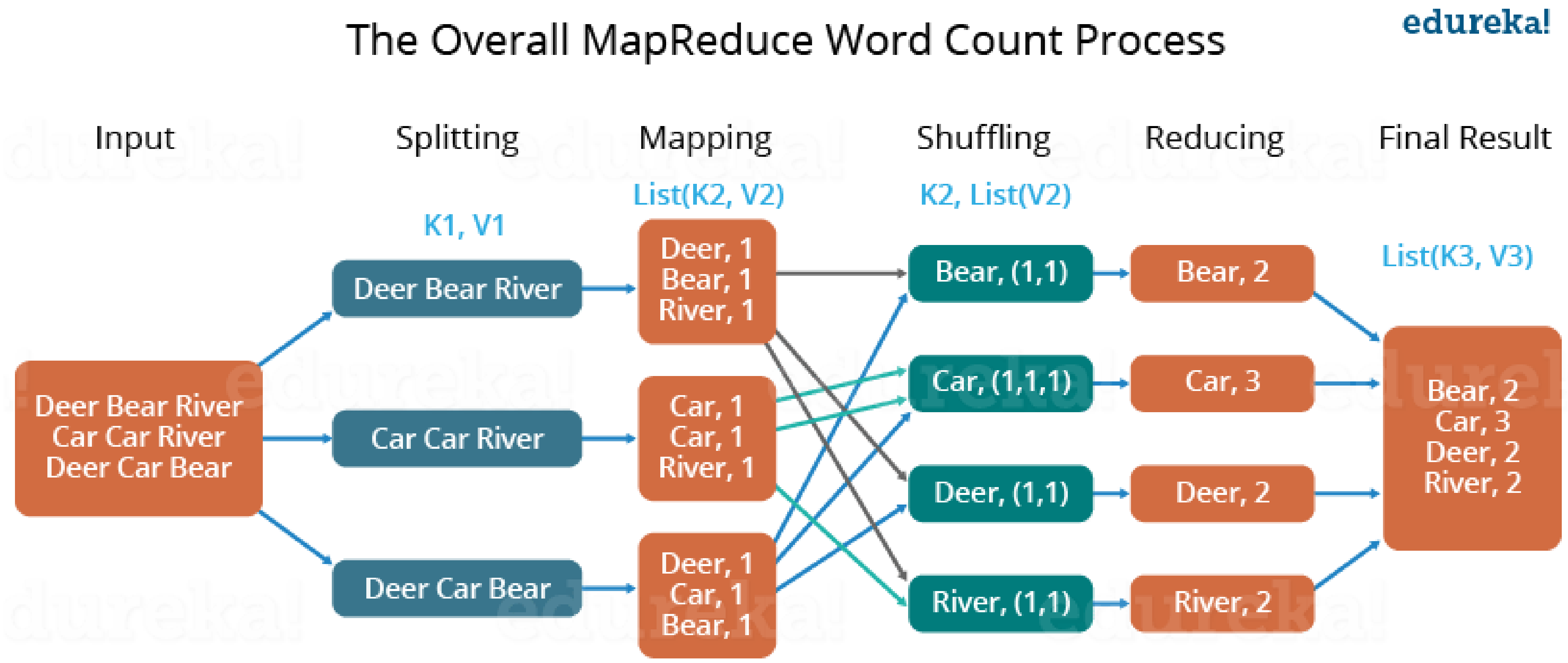
Simplified flow diagram for the MapReduce program



Combiner phase

- Optional phase that’s used for optimizing the MapReduce process.
- It’s used for reducing the pap outputs at the node level.
- In this phase, duplicate outputs from the map outputs can be combined into a single output.
- The combiner phase increases speed in the Shuffling phase by improving the performance of Jobs.

MapReduce



Source: https://4zy7s42hws72i51dv3513vnm-wpengine.netdna-ssl.com/wp-content/uploads/2018/02/MapReduce-Way-MapReduce-Tutorial-Edureka.png

References

- วราภรณ์ พรหมวิอินทร์ (2562). Big Data Analytics. นนทบุรี: คอร์ฟังก์ชั่น
- https://www.section.io/engineering-education/understanding-map-reduce-in-hadoop/
- https://www.stitchdata.com/resources/data-ingestion/