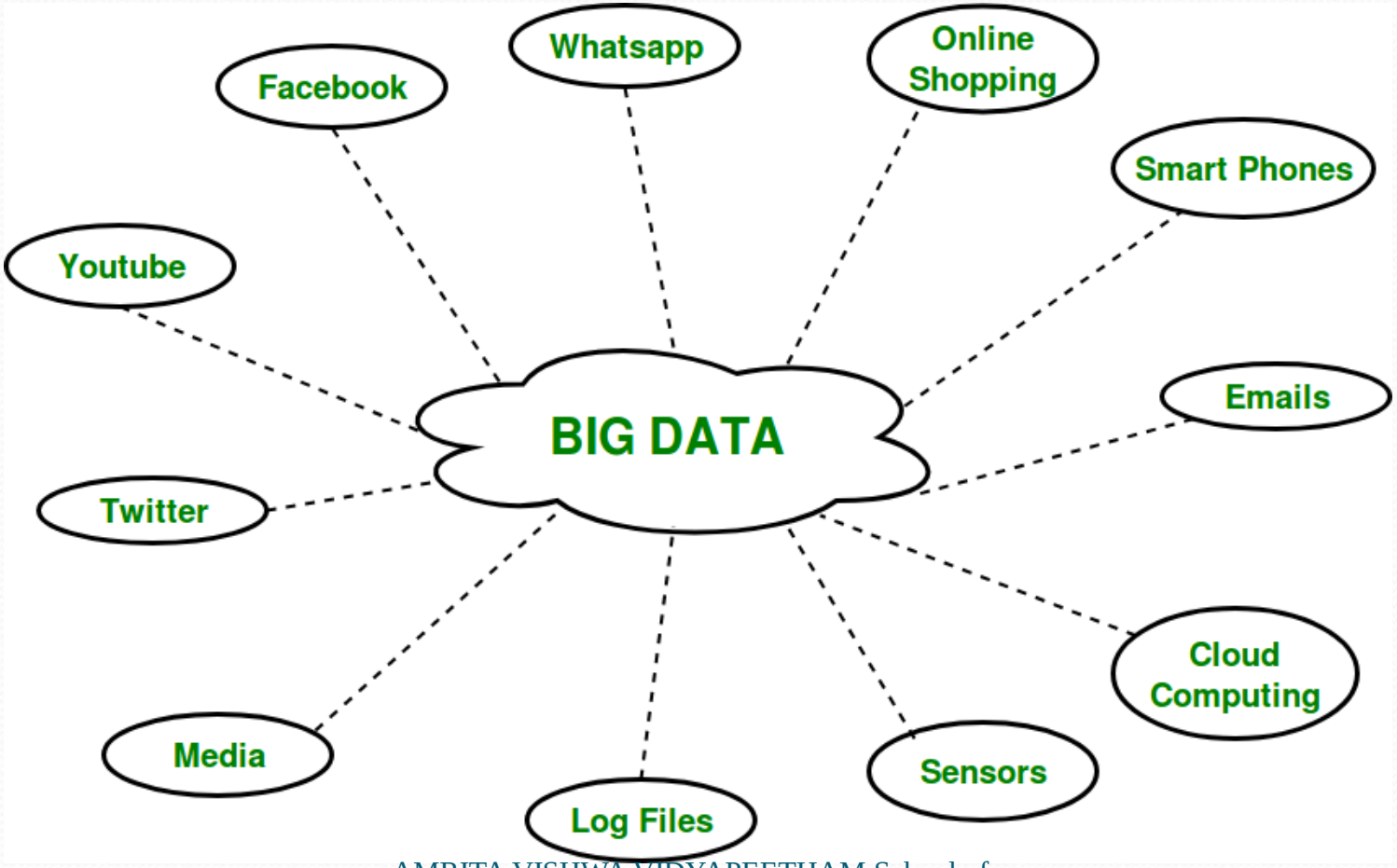
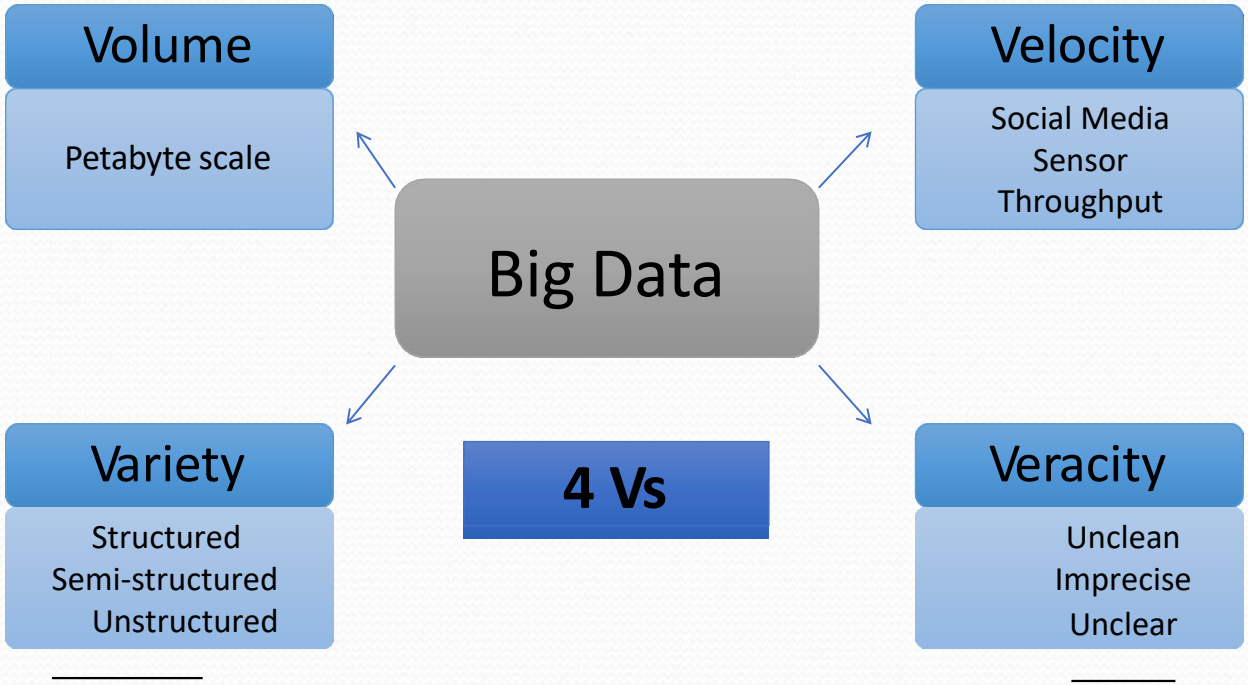
BIG DATA

BIG DATA

- Big Data is nothing but lots of data consisting of varieties of data.
- It is the concept of gathering useful insights from such voluminous amounts of structured, semi-structured and unstructured data that can be used for effective decision making in the business environment.
- This data is collected from various sources over a course of time and is cumbersome to be managed by traditional database tools


Whatsapp
Facebook
Online Shopping
Smart Phones
Youtube
BIG DATA
Emails
Twitter
Cloud Computing
Media
Sensors
Log Files

What are the Key Features of Big Data?



Big Data Characteristics

Volume:

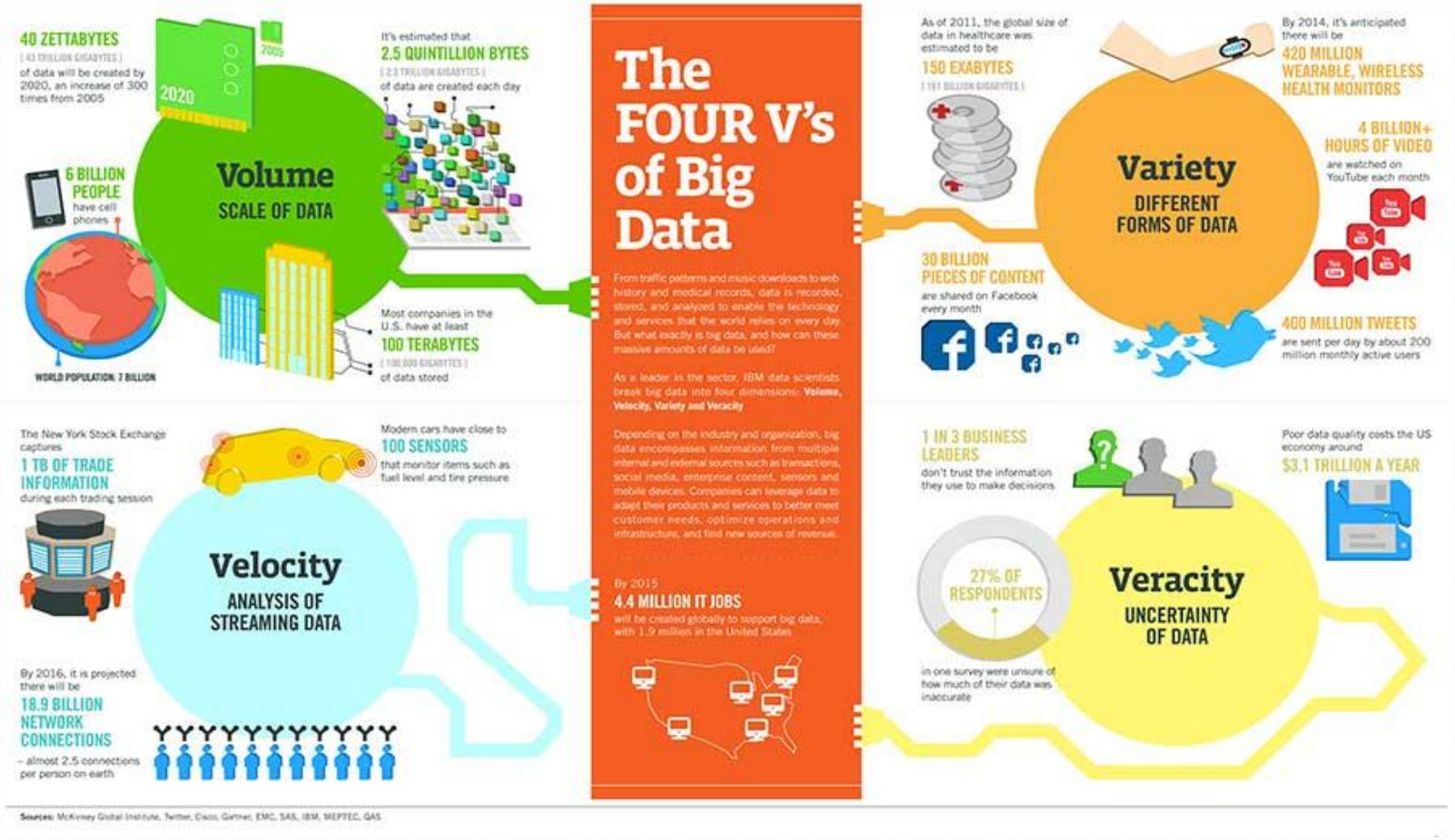
- The name ‘Big Data’ itself is related to a size which is enormous.
- Volume is a huge amount of data.
- To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a ‘Big Data’. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
- Hence while dealing with Big Data it is necessary to consider a characteristic ‘Volume’.
- *Example:* In the year 2016, the estimated global mobile traffic was 6.2 Exabytes(6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 ExaBytes of data.

- **Velocity:**
- Velocity refers to the high speed of accumulation of data.
- In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
- There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
- Sampling data can help in dealing with the issue like ‘velocity’.
- *Example:* There are more than 3.5 billion searches per day are made on Google. Also, FaceBook users are increasing by 22%(Approx.) year by year.

- **Variety:**
- It refers to nature of data that is structured, semi-structured and unstructured data.
- It also refers to heterogeneous sources.
- Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise.
- It can be structured, semi-structured and unstructured.

- **Veracity**
- It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- *Example:* Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

- **Value:**
- After having the 4 V's into account there comes one more V which stands for Value!. The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.
- Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5V's.


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
Volume
SCALE OF DATA
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
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
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
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
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
Variety
DIFFERENT FORMS OF DATA
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
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, GAS

Sources of Big Data

- These data come from many sources like
- **Social networking sites:** Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
- **E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.

- **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
- **Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
- **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.

Who is generating Big Data?

Social

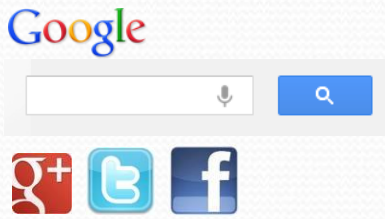
User Tracking & Engagement

Homeland Security

eCommerce

Financial Services

Real Time Search



Types of Big Data

- Big Data could be of three types:
- Structured
- Semi-Structured
- Unstructured
-

Structured

- The data that can be stored and processed in a fixed format is called as Structured Data.
- Data stored in a relational database management system (RDBMS) is one example of ‘structured’ data.
- It is easy to process structured data as it has a fixed schema.
- Structured Query Language (SQL) is often used to manage such kind of Data.

Se*mi-Structured*

- Semi-Structured Data is a type of data which does not have a formal structure of a data model,
- i.e. a table definition in a relational DBMS, but nevertheless it has some organizational properties like tags and other markers to separate semantic elements that makes it easier to analyze.
- XML files or JSON documents are examples of semi-structured data.

Unstructured

- The data which have unknown form and cannot be stored in RDBMS and cannot be analyzed unless it is transformed into a structured format is called as unstructured data.
- Text Files and multimedia contents like images, audios, videos are example of unstructured data.
- The unstructured data is growing quicker than others, experts say that 80 percent of the data in an organization are unstructured.

Examples of Big Data

- Daily we upload millions of bytes of data. 90 % of the world's data has been created in last two years.

Advantages:

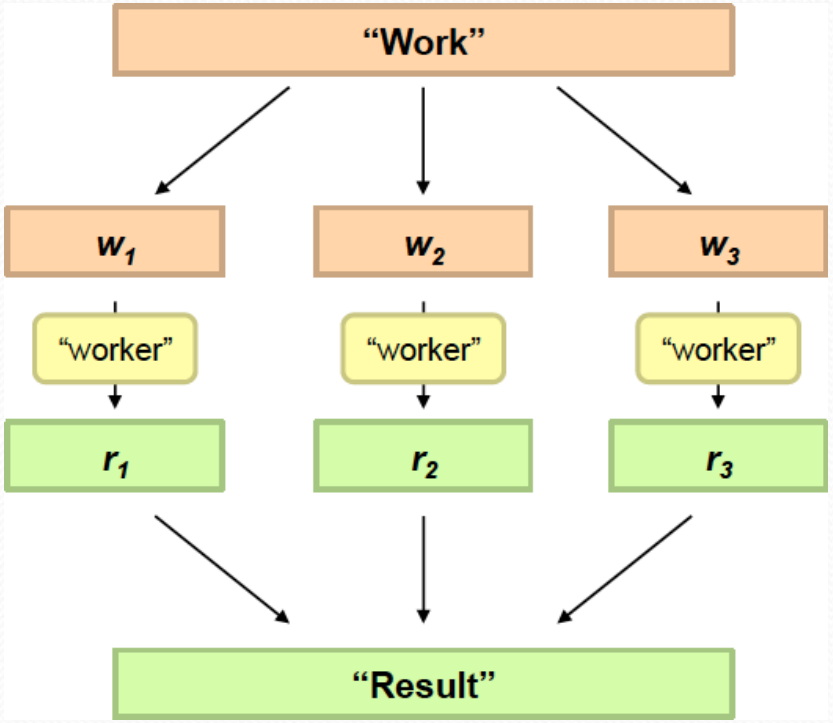
- Errors inside the business are known immediately.
- Higher conversion rate and additional income.
- The plan of action of your opposition is seen promptly.
- Extortion can be recognized the minute it happens and legitimate measures can be taken to restrict the harm.
- The principal points of interest of Big Data include the increased speed, capacity, and scalability of storage and having the measures and tools to deal with the data all the more proficiently.

Disadvantages:

- Data is collected from every source possible over a certain course of time. The data collected is raw, inconsistent and therefore subjected to more noise.
- Security is one of the key issues that Big Data is still struggling with, especially on the social media front.
- Most of the data a user is looking for analysis and interpretation purposes is hidden behind firewalls and private cloud that can only be accessed by having the technical knowledge and expertise to turn the raw data into relevant information.

- Philosophy to Scale for Big Data?

Divide and Conquer



Divide Work

Combine Results

Distributed processing is non-trivial

- How to assign tasks to different workers in an efficient way?
- What happens if tasks fail?
- How do workers exchange results?
- How to synchronize distributed tasks allocated to different workers?

Big data storage is challenging

- Data Volumes are massive
- Reliability of Storing PBs of data is challenging
- All kinds of failures: Disk/Hardware/Network Failures
- Probability of failures simply increase with the number of machines …