Big Data Overview

Surendra Panpaliya

Agenda

- What’s Big Data?
- Big Data: 3V’s
- Explosion of Data
- What’s driving Big Data
- Applications for Big Data Analytics
- Big Data Use Cases
- Benefits of Big Data

What's Big Data?

Massive complex Structured,

Unstructured data sets

That are rapidly generated and

Transmitted from

A wide variety of sources

What's Big Data?

Collection of large datasets

Cannot be processed

Using traditional computing techniques

What's Big Data?

Data which are very large in size

Peta bytes

10^15 byte size

Big Data: 3V's

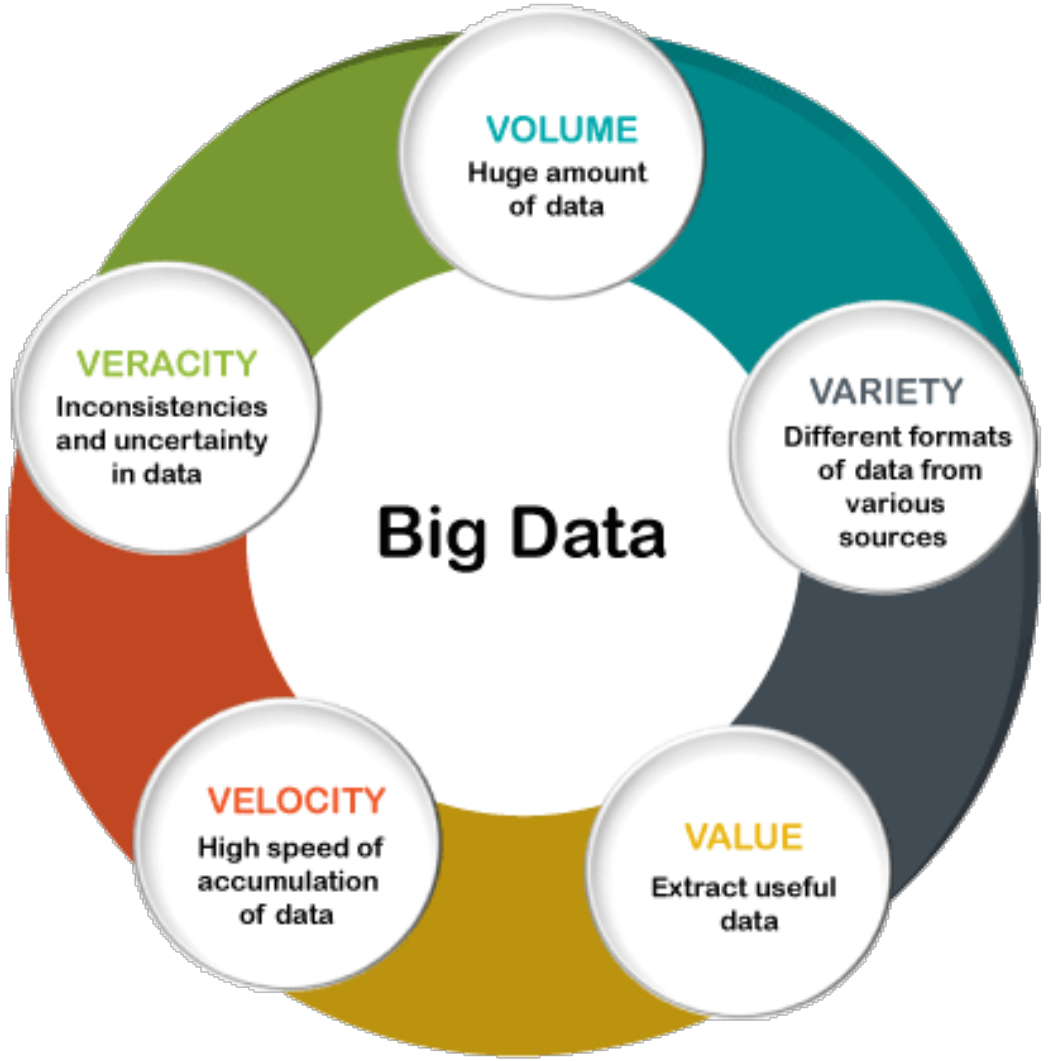
Volume: The huge amounts of data

Velocity: The lightning speed

Variety: The different sources and forms

Big Data: 5 V's

- Volume
- Veracity
- Variety
- Value
- Velocity



Volume

An enormous size.

Vast 'volumes’ of data

Generated from many sources

Volume

- Sources like,
- Business processes, Machines,
- Social media platforms,
- Networks,
- Human interactions
- Many more

Volume

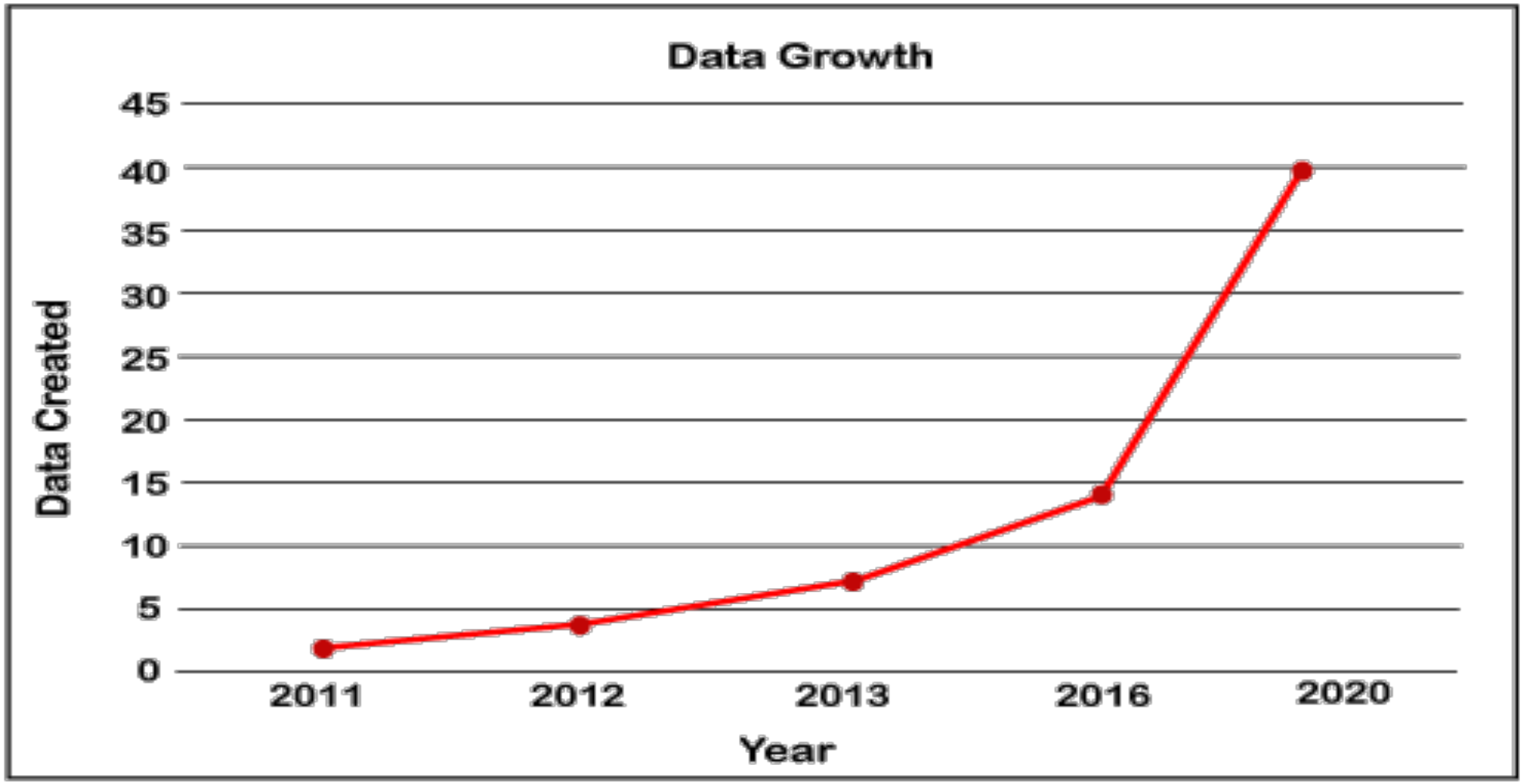
Facebook

Can generate approximately a **billion** messages

4.5 billion times "**Like**" button is recorded,

350 million + New posts are uploaded each day.

Volume


Data Growth
Data Created
45
40
35
30
25
20
15
10
5
0
2011
2012
2013
2016
2020
Year

Variety

Structured, unstructured, and semi-structured

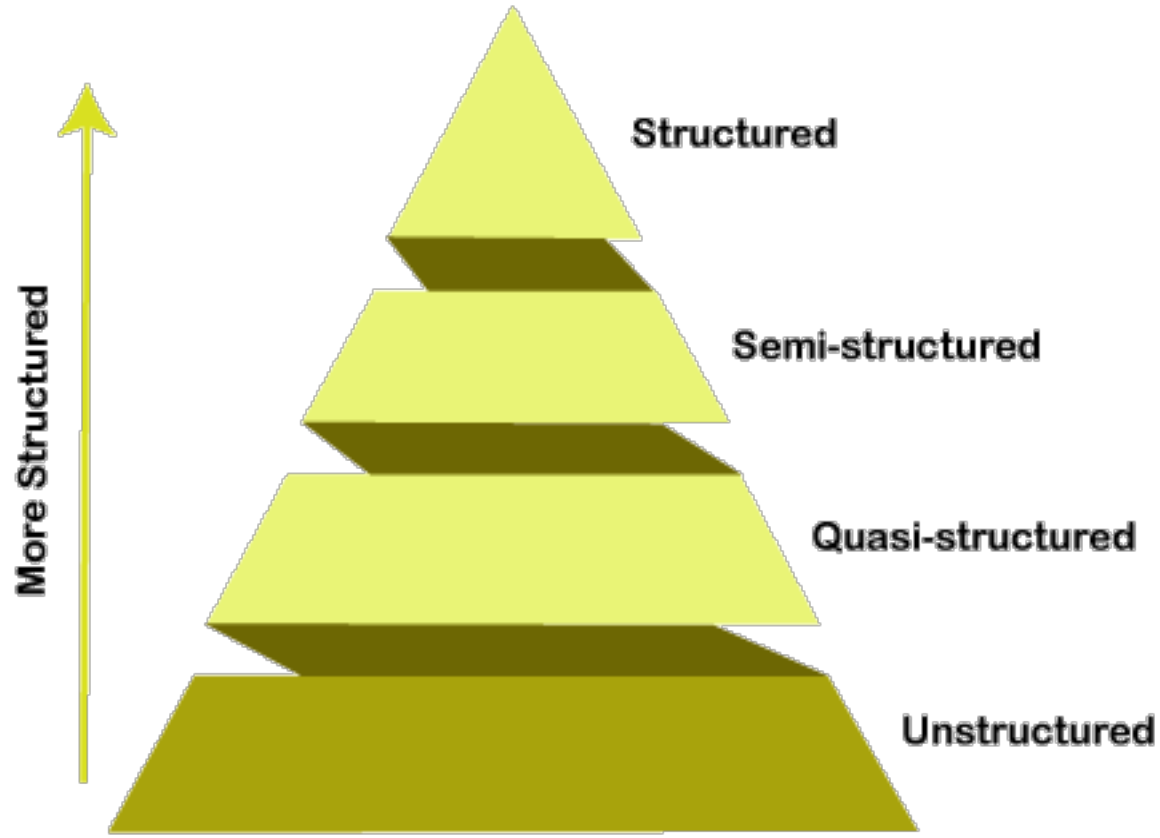
Collected from different sources.

Data will only be collected
from **databases** and **sheets** in the past.

But these days the data will comes in array forms,

PDFs, Emails, audios, SM posts, photos, videos

Variety


More Structured
Structured
Semi-structured
Quasi-structured
Unstructured

Structured data

In Structured schema

Along with all the required columns.

It is in a tabular form.

Stored in the Relational Database Management System.

Semi-structured Data

The schema is not appropriately defined

JSON, XML, CSV, TSV, and **email**.

OLTP (**Online Transaction Processing**) systems

Built to work with semi-structured data.

It is stored in relations, i.e., **tables**.

Unstructured Data

Unstructured files

log files, audio files, and **image** files

Some organizations have much data available,

But they did not know how to **derive** the value of data

Since the data is raw.

Quasi-structured Data

The data format contains textual data

With inconsistent data formats

that are formatted with effort and time with some tools.

Web server logs

Created and maintained by some server that contains a list of **activities**.

Veracity

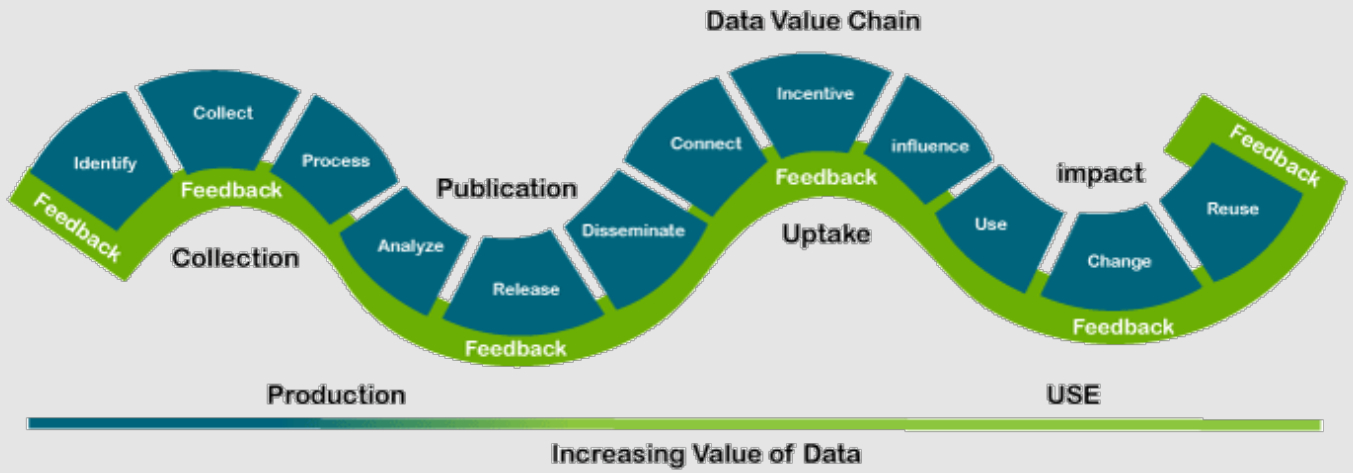
How much the data is reliable.

Ways to filter or translate the data.

It is the process of being able

to handle and manage data efficiently.

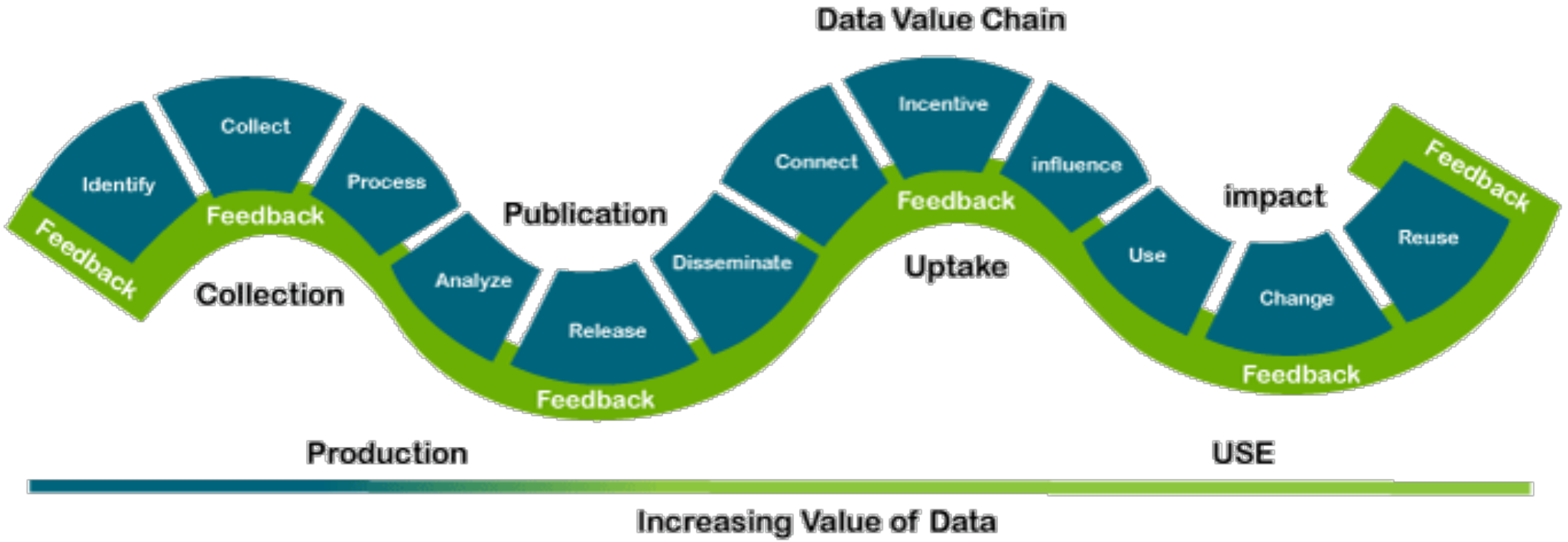
Facebook posts with hashtags



Value

- **Valuable**
- **reliable** data
- that we **store**
- **process**
- **analyze**.

Value


Data Value Chain
Identify
Collect
Process
Analyze
Release
Disseminate
Connect
Incentive
influence
Use
Change
Reuse
Feedback
Feedback
Feedback
Feedback
Feedback
Feedback
Collection
Publication
Uptake
impact
Production
USE
Increasing Value of Data

Velocity

The speed by which the data is created in **real-time**.

It contains the linking of

Incoming **data sets speeds,**

Rate of change

Activity bursts.

Velocity

The primary aspect of Big Data

To provide demanding data rapidly.

Big data velocity deals with

the speed at the data flows

Velocity

Sources like

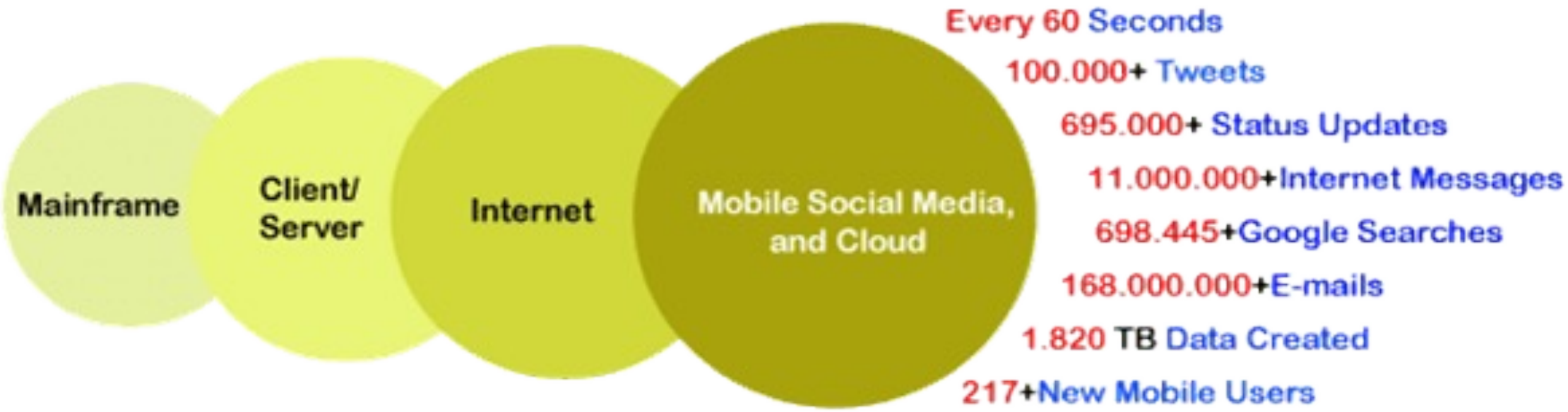
Application logs,

Business processes,

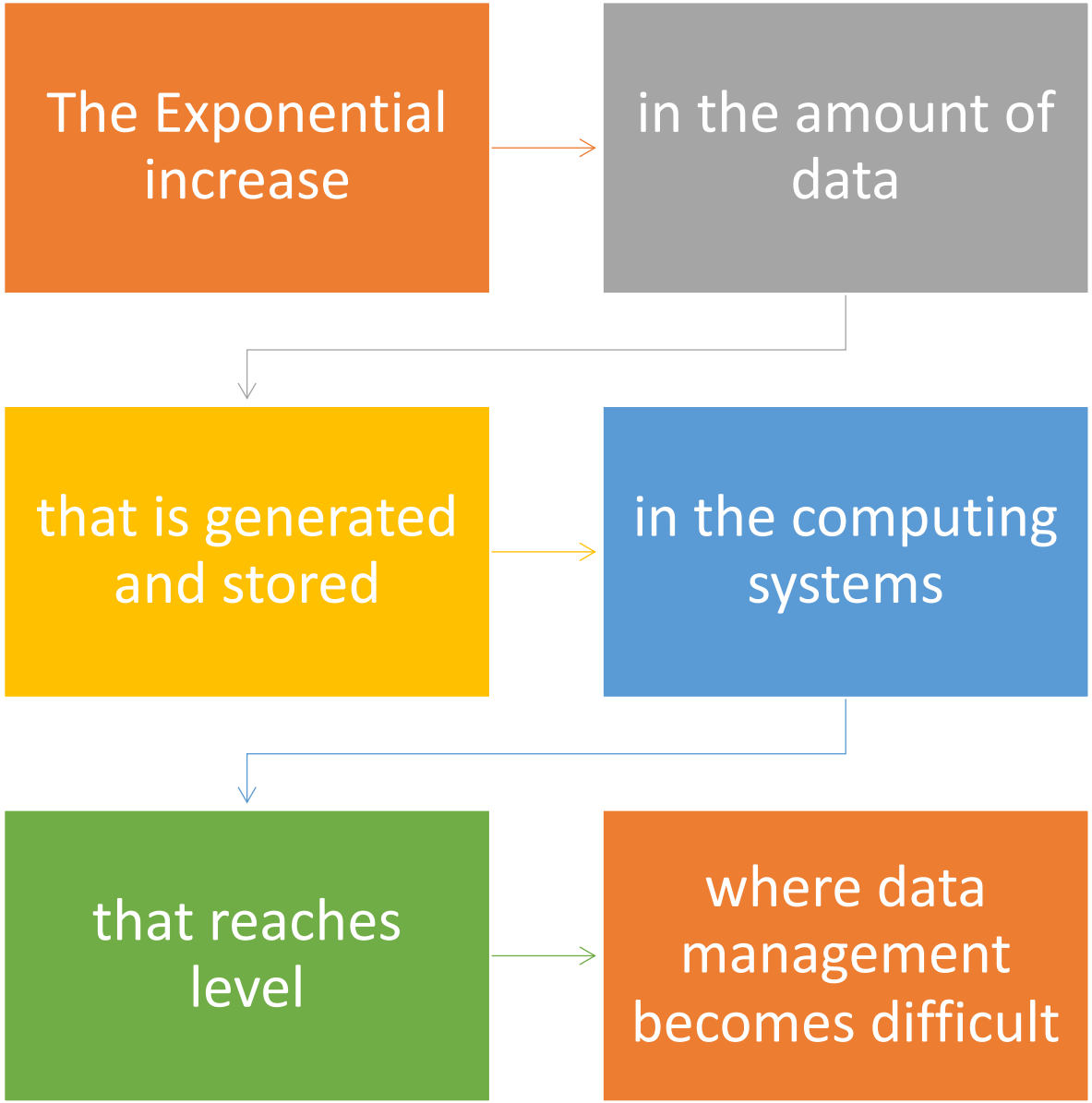
Networks, and social media sites,

Sensors, mobile devices

Velocity


Mainframe
Client/ Server
Internet
Mobile Social Media, and Cloud
Every 60 Seconds
100.000+ Tweets
695.000+ Status Updates
11.000.000+Internet Messages
698.445+Google Searches
168.000.000+E-mails
1.820 TB Data Created
217+New Mobile Users

Data Explosion



Key Drivers of Data Explosion

Increase in storage capacities

Cheaper storage.

Increase in data processing capabilities by modern computing devices.

Data generated and made available by different sectors.

Sources of Big Data

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.

Sources of Big Data

E-commerce site

Sites like Amazon, Flipkart, Alibaba

generates huge amount of logs

from which users buying trends can be traced.

Sources of Big Data

Weather Station

All the weather station and satellite

Gives very huge data

Which are stored and

Manipulated to forecast weather.

Sources of Big Data

- **Telecom Company**
- Telecom giants like Airtel, Vodafone
- study the user trends
- accordingly publish their plans
- for this they store the data of its million users.

Sources of Big Data

Share Market:

Stock exchange across the world

generates huge amount of data

through its daily transaction

What's driving Big Data

The digitization of society

The plummeting of technology costs

Connectivity through cloud computing

Increased knowledge about data science

Social media applications

The upcoming Internet-of-Things (IoT)

1. The digitization of society

- Big Data is largely consumer driven and consumer oriented.
- Most of the data in the world is generated by consumers,
- Who are nowadays ‘always-on’.
- Most people now spend 4-6 hours per day
- Consuming and generating data
- through a variety of devices and (social) applications.

1. The digitization of society

With every click, swipe or message,

new data is created in a database

somewhere around the world.

Because everyone now has a smartphone in their pocket,

the data creation sums to incomprehensible amounts.

1. The digitization of society

Some studies estimate that

60% of data was generated

within the last two years,

which is a good indication of the rate

with which society has digitized.

2. The plummeting of technology costs

Technology related to collecting and processing

massive quantities of diverse (high variety)

Data has become increasingly more affordable.

2. The plummeting of technology costs

The costs of data storage and

Processors keep declining,

Making it possible for small businesses

Individuals to become involved with Big Data.

2. The plummeting of technology costs

For storage capacity,

the often-cited Moore’s Law

Still holds that the storage density (and therefore capacity)

Still doubles every two years.

2. The plummeting of technology costs

A second key contributing factor

to the affordability of Big Data

has been the development of

open source Big Data software frameworks.

2. The plummeting of technology costs

The most popular software framework

Nowadays considered the standard for Big Data

Apache Hadoop

for distributed storage and processing.

2. The plummeting of technology costs

Due to the high availability of these

Software frameworks in open sources,

It has become increasingly inexpensive

to start Big Data projects in organizations.

3. Connectivity through cloud computing

Cloud computing environments

Where data is remotely stored

in distributed storage systems

have made it possible

to quickly scale up or scale down

IT infrastructure and facilitate

a pay-as-you-go model

3. Connectivity through cloud computing

This means that organizations

That want to process

Massive quantities of data

Do not have to invest

In large quantities of IT infrastructure.

3. Connectivity through cloud computing

Instead, they can license the storage

Processing capacity

They need

Only pay for the amounts

They actually used.

3. Connectivity through cloud computing

As a result,

Most of Big Data solutions

Leverage the possibilities of cloud computing

To deliver their solutions

To enterprises.

4. Increased knowledge about data science

In the last decade,

the term data science and data scientist

have become tremendously popular.

In October 2012,

Harvard Business Review called the data scientist

“sexiest job of the 21st century”

4. Increased knowledge about data science

The demand for data scientist

and similar job titles

has increased tremendously

many people have actively become

engaged in the domain of data science.

4. Increased knowledge about data science

Increased knowledge about data science

The knowledge and education

About data science

Has greatly professionalized and

More information

becomes available every day.

4. Increased knowledge about data science

While statistics and data analysis

Mostly remained an academic field previously,

It is quickly becoming a popular subject

Among students

the working population.

5. Social media applications

Everyone understands the impact

that social media has on daily life.

However, in the study of Big Data,

social media plays

a role of paramount importance.

5. Social media applications

Not only because of the sheer volume of data

that is produced everyday through platforms such as

Twitter, Facebook, LinkedIn and Instagram,

but also because social media provides

nearly real-time data

about human behavior.

5. Social media applications

Social media data provides insights

Into the behaviors

Preferences and opinions of ‘the public’

on a scale that has never been known before.

5. Social media applications

Due to this,

it is immensely valuable

to anyone who is able to derive

meaning from these large quantities of data.

5. Social media applications

Social media data can be used

to identify customer preferences

for product development,

target new customers

for future purchases,

or even target potential voters in elections.

5. Social media applications

Social media data

might even be considered

one of the most important

business drivers of Big Data.

6. The upcoming internet of things (IoT)

The Internet of things (IoT)

is the network of physical devices,

vehicles, home appliances and

other items embedded with electronics,

6. The upcoming internet of things (IoT)

Software, sensors,

Actuators,

network connectivity

which enables these objects

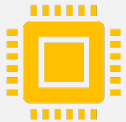
to connect

exchange data.

6. The upcoming internet of things (IoT)

It is increasingly gaining popularity

as consumer goods providers



start including ‘smart’ sensors

in household appliances.

6. The upcoming internet of things (IoT)

Whereas the average household in 2010

had around 10 devices

that connected to the internet,

this number is expected to rise to

50-60 per household by 2025.

6. The upcoming internet of things (IoT)

Examples of these devices include

Thermostats,

Smoke detectors,

Televisions,

Audio systems

Even smart refrigerators.