Welcome to Big Data Analytics Module

Instructor : Tushar Kakaiya

Introducing Big Data

Chapter - 01

DATA UNITS

Unit	Value	Size
bit (b)	**0 or 1**	**1/8 of a byte**
byte (B)	**8 bits**	**1 byte**
kilobyte (KB)	**1000^1 bytes**	**1,000 bytes**
megabyte (MB)	**1000^2 bytes**	**1,000,000 bytes**
gigabyte (GB)	**1000^3 bytes**	**1,000,000,000 bytes**
terabyte (TB)	**1000^4 bytes**	**1,000,000,000,000 bytes**
petabyte (PB)	**1000^5 bytes**	**1,000,000,000,000,000 bytes**
exabyte (EB)	**1000^6 bytes**	**1,000,000,000,000,000,000 bytes**
zettabyte (ZB)	**1000^7 bytes**	**1,000,000,000,000,000,000,000 bytes**
yottabyte (YB)	**1000^8 bytes**	**1,000,000,000,000,000,000,000,000 bytes**

- **NOTE:** A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes. This is an important distinction, since a byte is 8x as large as a bit.
- For example, 100 KB (kilobytes) = 800 Kb (kilobits).

BIG DATA

Big data is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large datasets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years.

- Big Data is also **data** but with a **huge size**. Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

Examples of Big Data

- The New York Stock Exchange generates about 4−5 terabytes of data per day.
- Facebook hosts more than 240 billion photos, growing at 7 petabytes per month.
- A single Jet engine can generate 10+terabytes of data in 30 minutes of flight time. With many thousand flights
- Ancestry.com, the genealogy site, stores around 10 petabytes of data.
- The Internet Archive stores around 18.5 petabytes of data.
- The Large Hadron Collider near Geneva, Switzerland, produces about 30 petabytes of data per year.

* This amount of data is approximate figures found from various Internet Sources. Actual data amount can vary.

Types of Big Data

Big Data could be found in three forms:

- **Structured**
- **Unstructured**
- **Semi-structured**

Types of Big Data - Structured

- Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.

- Data stored in Relational DB can be termed as Structured Data set.

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
5467	Tushar Kakaiya	Male	HR	345900
4562	Joshi Pushpa	Female	Admin	123970
8912	Aishwarya Das	Female	Finance	500000

Types of Big Data – Semi-Structured

- Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with.
- Example of semi-structured data is a data represented in an XML file.

<rec><name>Tushar Kakaiya</name><sex>Male</sex><age>29</age></rec>

<rec><name>Seema Rav</name><sex>Female</sex><age>41</age></rec>

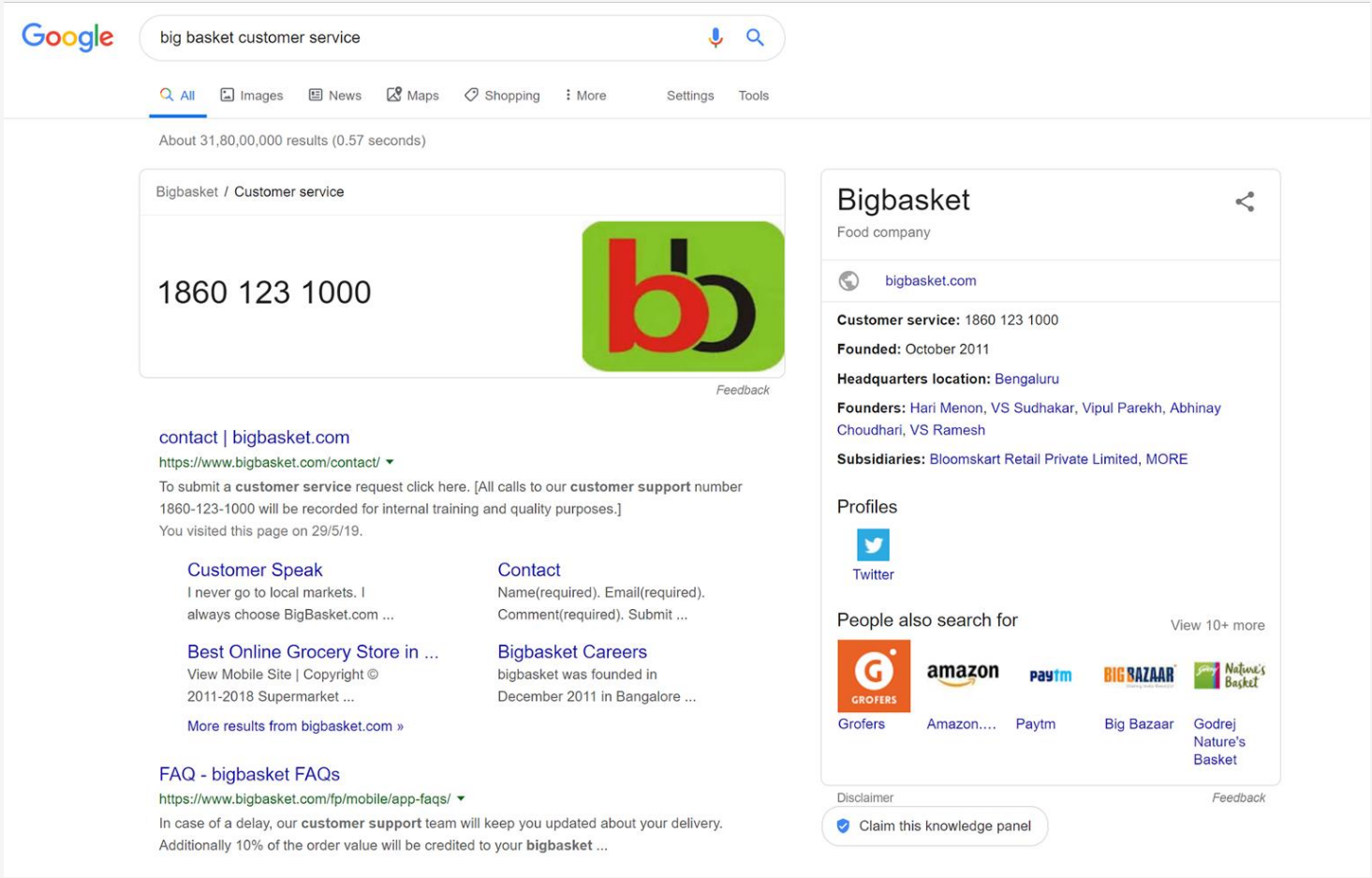
<rec><name>Satish Shah</name><sex>Male</sex><age>29</age></rec>

<rec><name>Suneeta Roy</name><sex>Male</sex><age>26</age></rec>

<rec><name>Jeremy Black</name><sex>Male</sex><age>35</age></rec>

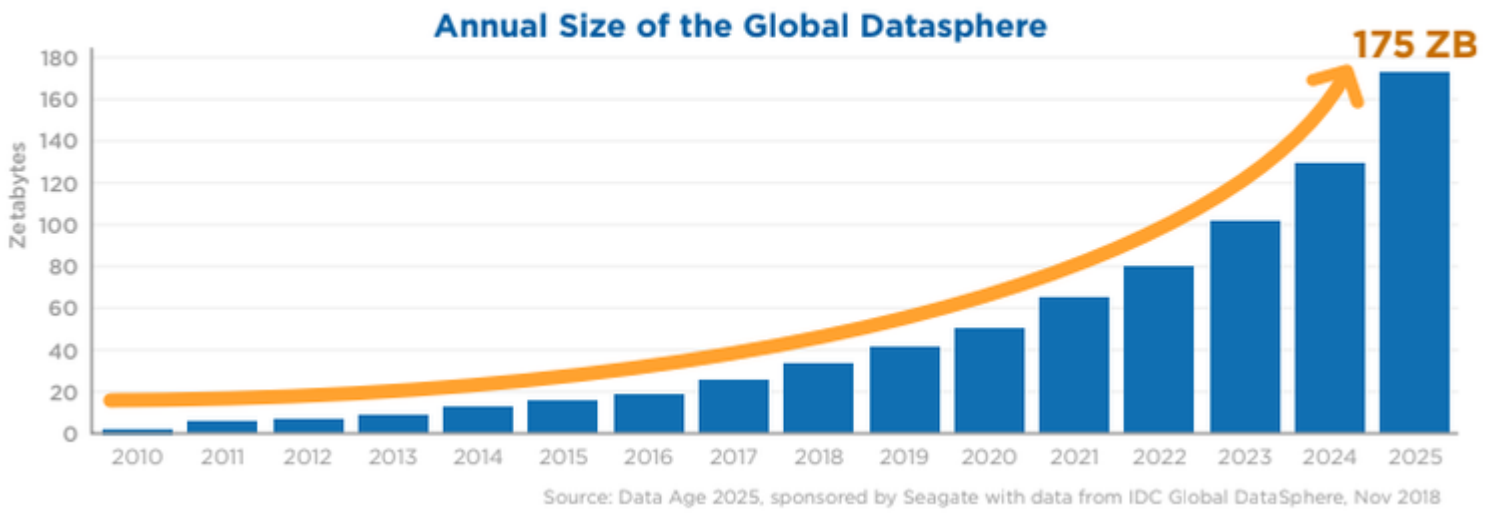
TYPES OF BIG DATA – UNSTRUCTURED

- Any data with unknown form or the structure is classified as unstructured data.
- In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it.
- Heterogeneous data source containing a combination of simple text files, images, audio, videos etc.
- Web site logs, Transaction Logs Files, etc. are also Unstructured in nature

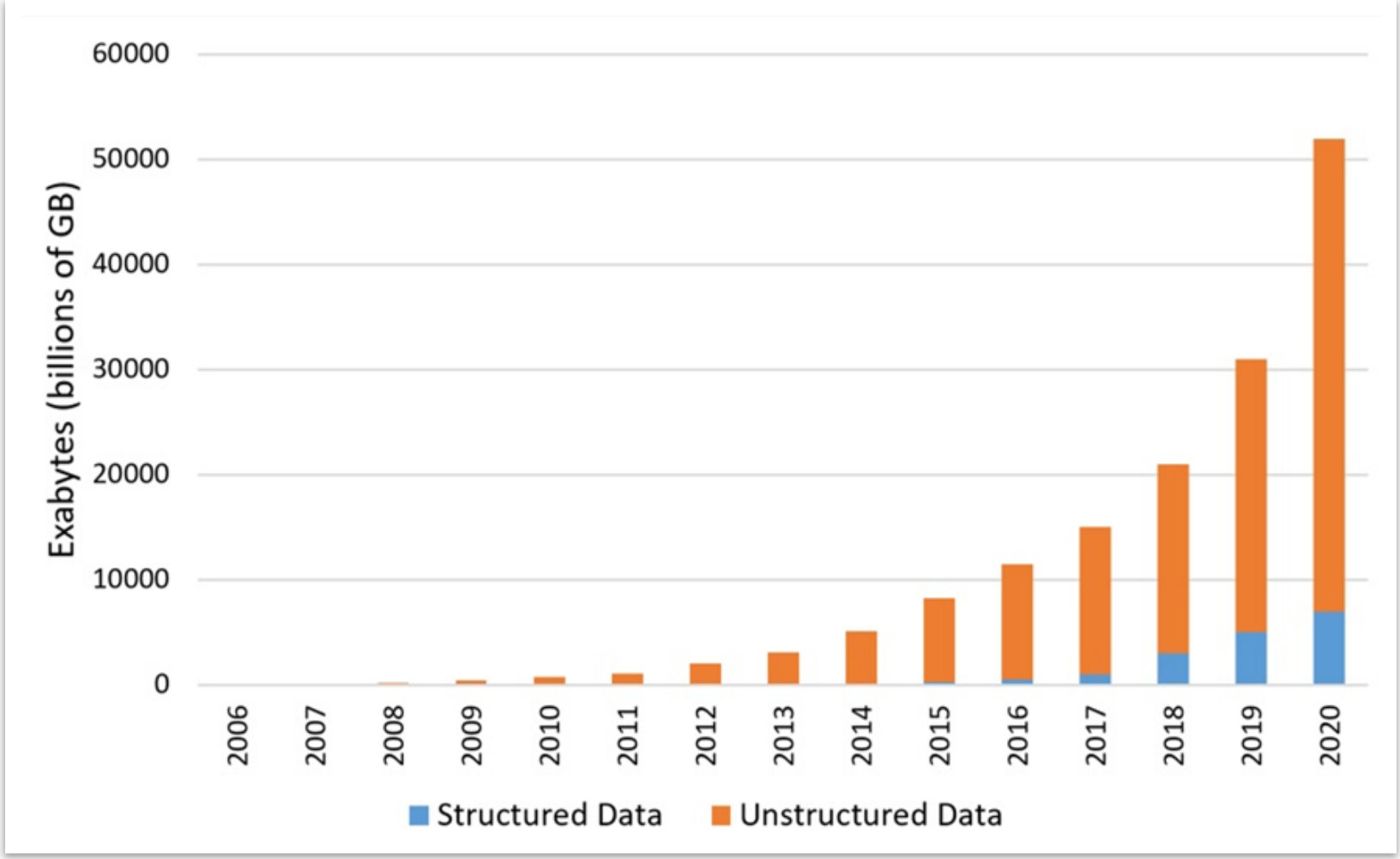


Data growth over the years

Figure 1 – Annual Size of the Global Datasphere



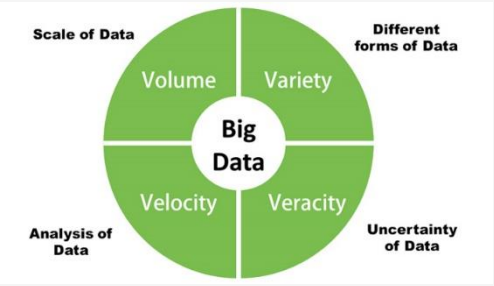
Data growth over the years



Characersitics of Big Data / 4 V’s

- **Volume:**
  - The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, **'Volume'** is one characteristic which needs to be considered while dealing with Big Data.
- **Velocity:**
  - The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.
- **Variety:**
  - Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.
- Veracity:
  - This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

If any of these Properties are satisfied, then we term that data set as BigData



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

It's estimated that
2.5 QUINTILLION BYTES
(2.3 TRILLION GIGABYTES)
of data are created each day

6 BILLION PEOPLE
have cell phones

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)

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

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

Velocity
ANALYSIS OF STREAMING DATA

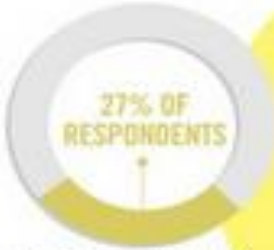
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

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



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

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

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

IBM

Why do we need to process Big Data

- Which product is more popular in my store?
- What is my customer behavioral pattern?
- How do I predict better power consumption?
- How do I detect potential fraud?
- How do I predict customer churn faster?

Big Data Analytics

- Big data analytics involves examining large amounts of data.
- This is done so as to uncover the hidden patterns, correlations and also to give insights so as to make proper business decisions.
- Organizations have realized the need for evolving from a knowing organization to a learning organization. Essentially, businesses want to be more objective and data-driven, and so they are embracing the power of data and technology.
- The big data concept has been around for many years. Decades before the first mention of big data, businesses applied analytics on the data they collected so as to gain insights and uncover trends. This involved capturing numbers on a spread sheet and manually examining the numbers.
- Big data analytics is done using advanced software systems. This allows businesses to reduce the analytics time for speedy decision making. Basically, the modern big data analytics systems allow for speedy and efficient analytical procedures. This ability to work faster and achieve agility offers a competitive advantage to businesses. In the meantime, businesses enjoy lower cost using big data analytics software.

- Source Blog: https://www.mentionlytics.com/blog/5-real-world-examples-of-how-brands-are-using-big-data-analytics/

Benefits of Big Data Processing

- Businesses can utilize outside intelligence while taking decisions
- Access to social data from search engines and sites like facebook, twitter are enabling organizations to fine tune their business strategies.
- Improved customer service
- Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses. Foundation of Data Science
- Early identification of risk to the product/services, if any
- Better operational efficiency
- Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the data warehouse. In addition, such integration of Big Data technologies and data warehouse helps an organization to offload infrequently accessed data.

EXAMPLE FOR TARGETED ADVERTS

How Netflix Uses Big Data to Drive Success

January 20, 2018 by Editorial Team 5 Comments

tweet share share email

Netflix has over 100 million subscribers and with that comes a wealth of data they can analyze to improve the user experience. Big data has helped Netflix massively in their mission to become the king of stream.

Big data helps Netflix decide which programs will be of interest to you and the recommendation system actually influences 80% of the content we watch on Netflix. The company even gave away a $1 million prize in 2009 to the group who came up with the best algorithm for predicting how customers would like a movie based on previous ratings. The algorithms help Netflix save $1 billion a year in value from customer retention.

Our friends over at FrameYourTV developed the compelling infographic below, "How Netflix Uses Big Data to Drive Success," that highlights Netflix's use of big data, specifically interesting statistics, how Netflix gathers big data, and how Netflix uses big data.

- With over 100 million subscribers, the company collects huge data, which is the key to achieving the industry status Netflix boosts.
- If you are a subscriber, you are familiar to how they send you suggestions of the next movie you should watch. Basically, this is done using your past search and watch data. This data is used to give them insights on what interests the subscriber most. See the screenshot below showing how Netflix gathers big data.

Example of Brand that uses Big Data Analytics for Risk Management

UOB bank from Singapore is an example of a brand that uses big data to drive risk management. Being a financial institution, there is huge potential for incurring losses if risk management is not well thought of. UOB bank recently tested a risk management system that is based on big data. The big data risk management system enables the bank to reduce the calculation time of the value at risk. Initially, it took about 18 hours, but with the risk management system that uses big data, it only takes a few minutes. Through this initiative, the bank will possibly be able to carry out real-time risk analysis in the near future (Andreas, 2014).

Example of a Brand that uses Big Data for Supply Chain Efficiency

PepsiCo is a consumer packaged goods company that relies on huge volumes of data for an efficient supply chain management. The company is committed to ensuring they replenish the retailers' shelves with appropriate volumes and types of products. The company's clients provide reports that include their warehouse inventory and the POS inventory to the company, and this data is used to reconcile and forecast the production and shipment needs. This way, the company ensures retailers have the right products, in the right volumes and at the right time. Listen to this webinar where the company's Customer Supply Chain Analyst talks about the importance of big data analytics in PepsiCo Supply chain.

Key Takeaway

- Big data analytics is an important investment for a growing business. Through implementing big data analytics businesses can achieve competitive advantage, reduced the cost of operation and drive customer retention. There are various sources of customer data that businesses can leverage. As technological advancements continue, data is becoming readily available to all organizations.

- Technically, it is fair enough to say that organizations already have data at their disposal. It is up to the individual organizations to ensure they implement appropriate data analysis systems that can handle the huge data.