Unit-1

Overview of Big Data

Topic covered

- Introduction to Big Data
- Evolution of Big Data
- Structuring of Big Data
- Fundamentals of Big Data
- Big Data Analytics
- Career and Future in Big Data

Introduction to Big Data

- Every seconds, there are around 8,22 tweets on twitter.
- Every minute, nearly 510 comments are posted, 2,93000 statues are updated, and 1,36000 photos are updated on Facebook.
- Every hour, Walmart, a global discount department store chain, handles more then 1 million customers transections.
- Every day, consumers make around 11.5 million payments by using PayPal.
- We live in digital world where data is increasing rapidly because of the ever increasing use of the internet, sensors, and heavy machines at a very high rate.

- The sheer volume, variety, velocity, and veracity of such data is signified by the term ‘Big data is structured, unstructured, semi structured, or heterogeneous in nature.
- It becomes very difficult for computing systems to manage ‘Big Data’ because of the immense speed and volume at which it is generated.
- Traditional data management, warehousing, and analysis systems fizzle to analyze this type of data.
- Due to this complexity, big data is stored in distributed architecture file system.

- Hadoop by Apache is widely used for storing and managing Big Data.
- Analyzing big data is a challenging task as it involves large distributed file systems, which should be fault tolerant, flexible, and scalable.
- According to IBM, “Every day, we create 2.5 quintillion bytes of data – so much that 90% of the data in the world has been created in the last two years alone.
- This comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transection records, and cell phone GPS signals to name a few. This data is big data”

- Data is every where in every industry, in the form of numbers, videos, and text.
- As data continues to grow, so does the need to organize it.
- Collecting such huge amount of data would just be a waste of time, effort, and storage space if it cannot be put to any logical use.
- The need to sort, organize, analyze, and offer this critical data in a systematic manner leads to the rise of the much discusses term, Big Data.
- **The process of capturing or collecting Big Data is known as ‘datafication’. Big Data is datafied so that it can be used productively**

Features of Big Data

- Big Data is a new data challenge that requires leveraging existing systems differently.
- Big data is classified in terms of 4Vs

1. Volume
2. Variety
3. Velocity
4. veracity

- Big data usually unstructured and qualitative in nature.

Real-world example of Big Data

- Consumer product companies and retail organizations are observing data on social media website such as Facebook and twitter. These sites help them to analyze customers behavior, preferences and product perception. Accordingly, the companies can line up their products to gain profits. This phenomena is also known as social media analytics.

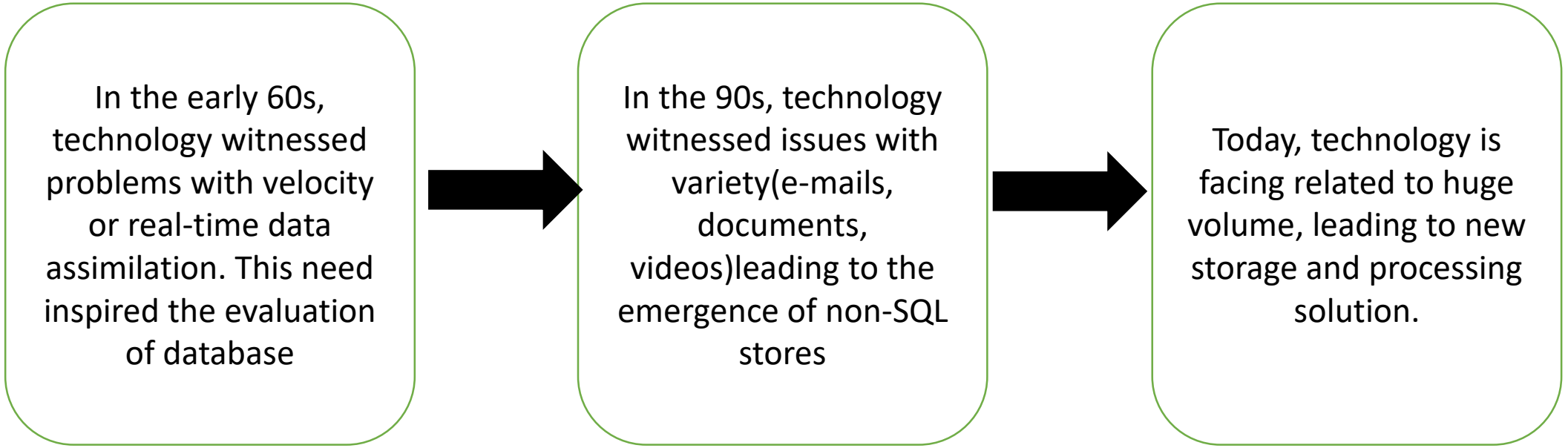
Types and Source of Data

Type	Description	Source
Social Data	Refers to the information collected from various social networking sites and online portals	Facebook, Twitter, LinkedIn
Machine Data	Refers to the information generated from RFID chips, bar code scanners and sensors	RFID chips readings, Global Positioning System(GPS) results
Transactional Data	Refers to the information generated from online shopping sites, retailors, and business to business (B2B) transections	Retail websites like eBay and Amazon

History of Data Management – Evaluation of Big Data

- Big Data is the new term of data evolution directed by the enormous velocity, variety, and volume of data.
- Velocity implies the speed with which the data flows in an organization.
- Variety refers to the varied forms of data, such as structured, or unstructured and volume defines the amount or quantity of data an organization has to deal with it.

- Challenges faced while handling data over past few decades:



Structuring of Big Data

- Structuring of data, simple terms, is arranging the available data in a manner such that becomes easy to study, analyze and derive conclusion from it. But, why is structuring required?
- In daily life you may have come across questions like:

❑How do I use to my advantage the vast amount of data and information I come across?

❑Which news articles should I read the thousands I come across?

❑How do I choose a book of the millions available on my favorite sites or stores?

❑How do I keep myself updated about new events, sports, inventions, and discoveries taking place across the globe?

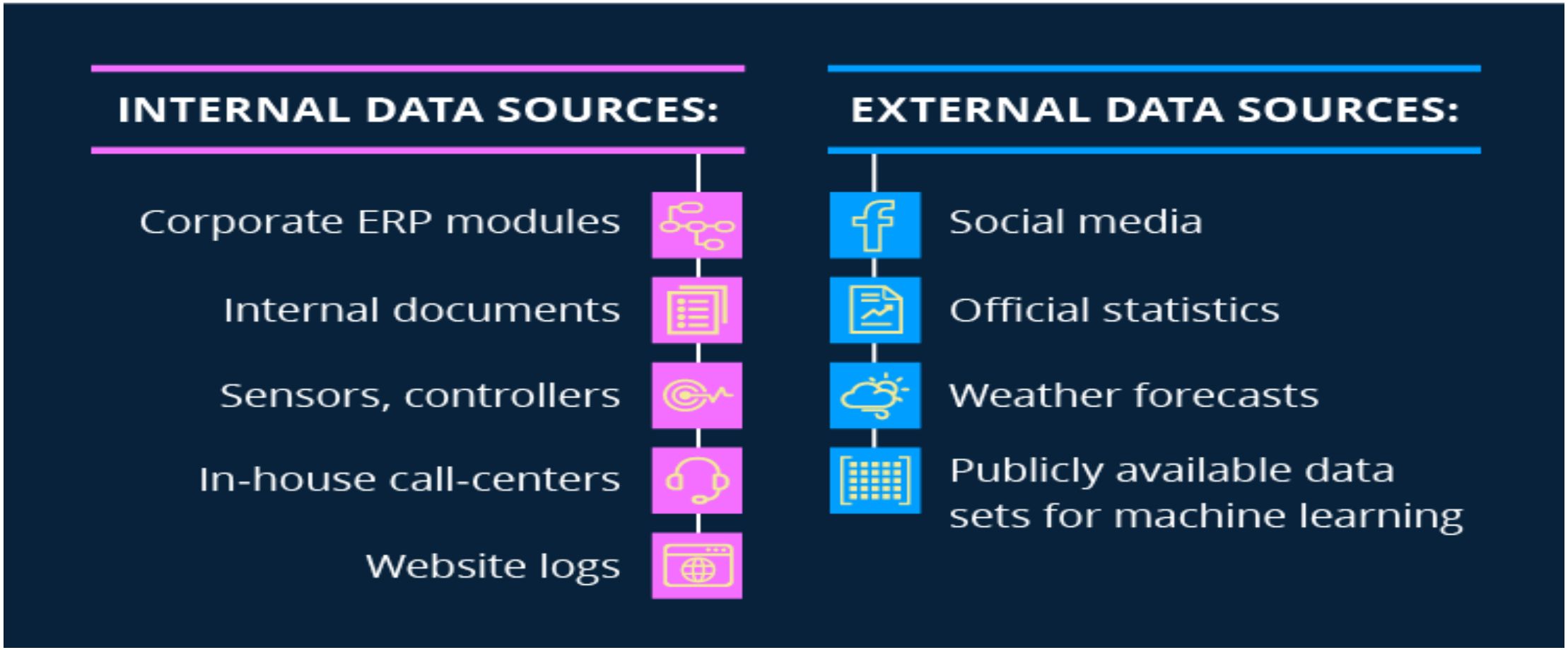
- Today, solution to such questions can be found by information processing systems.
- These systems can analyze and structure a large amount of data specifically for you on the basis of what you can analyze and structure a large amount of data specified for you on the basis of what you searched, what you looked at, and for how long you remained at a particular page or website, thus scanning and presenting you with the customized information as per your behavior and habits.

Types of Data

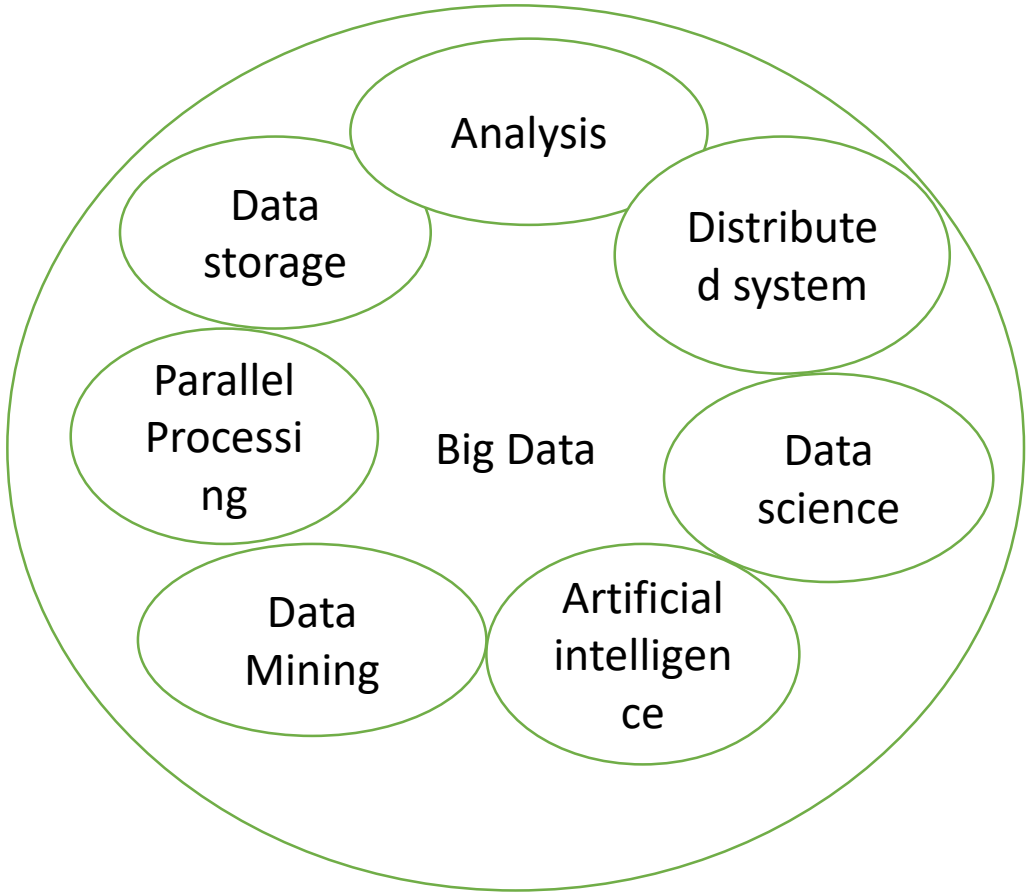
- Data that comes from multiple sources, such as databases, enterprise resource planning (ERP) systems, weblogs, chat history, and GPS maps, varies in its format, different formats of data need to be made consists and clear to be used for analysis. Data is obtained primarily form following types of sources:

❑Internal sources, such as organizational or enterprise data.

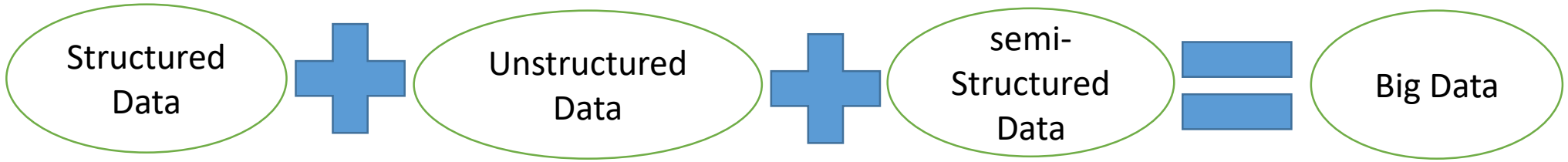
❑External sources, such as social data.


INTERNAL DATA SOURCES:
Corporate ERP modules
Internal documents
Sensors, controllers
In-house call-centers
Website logs
EXTERNAL DATA SOURCES:
Social media
Official statistics
Weather forecasts
Publicly available data sets for machine learning

Concepts of Big Data



- On the basis of the data received from the source mentioned:



Types of Big Data

Structured big data

- In general, structured data in a Big Data environment is **stored in Databases and other well-defined structures and schemas**. Structured data has clearly defined attributes for easy access and is tabular, having rows and columns that clearly outline the data structure

Example of Structured Data

Sample of Structured Data				
Customer ID	Name	Product ID	City	State
1	Vipul Gamit	12	Bardoli	Gujarat
2	Hardik Shah	13	Bardoli	Gujarat
3	Piyush Patel	14	Mahuva	Gujarat

What is unstructured data in big data?

- Unstructured simply means that it is **datasets (typical large collections of files) that aren't stored in a structured database format**. Unstructured data has an internal structure, but it's not predefined through data models. It might be human generated, or machine generated in a textual or a non-textual format.

Challenges Associated with Unstructured Data

- Working with unstructured data certain challenges, which are as follows:

❑Identifying the unstructured data that can be processed.

❑Sorting, Organizing, and arranging un-structured data in different sets and formats.

❑Combing and linking unstructured data in a more structured format to derive any logical conclusions out of the available information.

❑Costing in terms of storage space and human resource(data analytics and scientists) needed to deal with the exponential growth of unstructured data.

What is semi-structured data in big data?

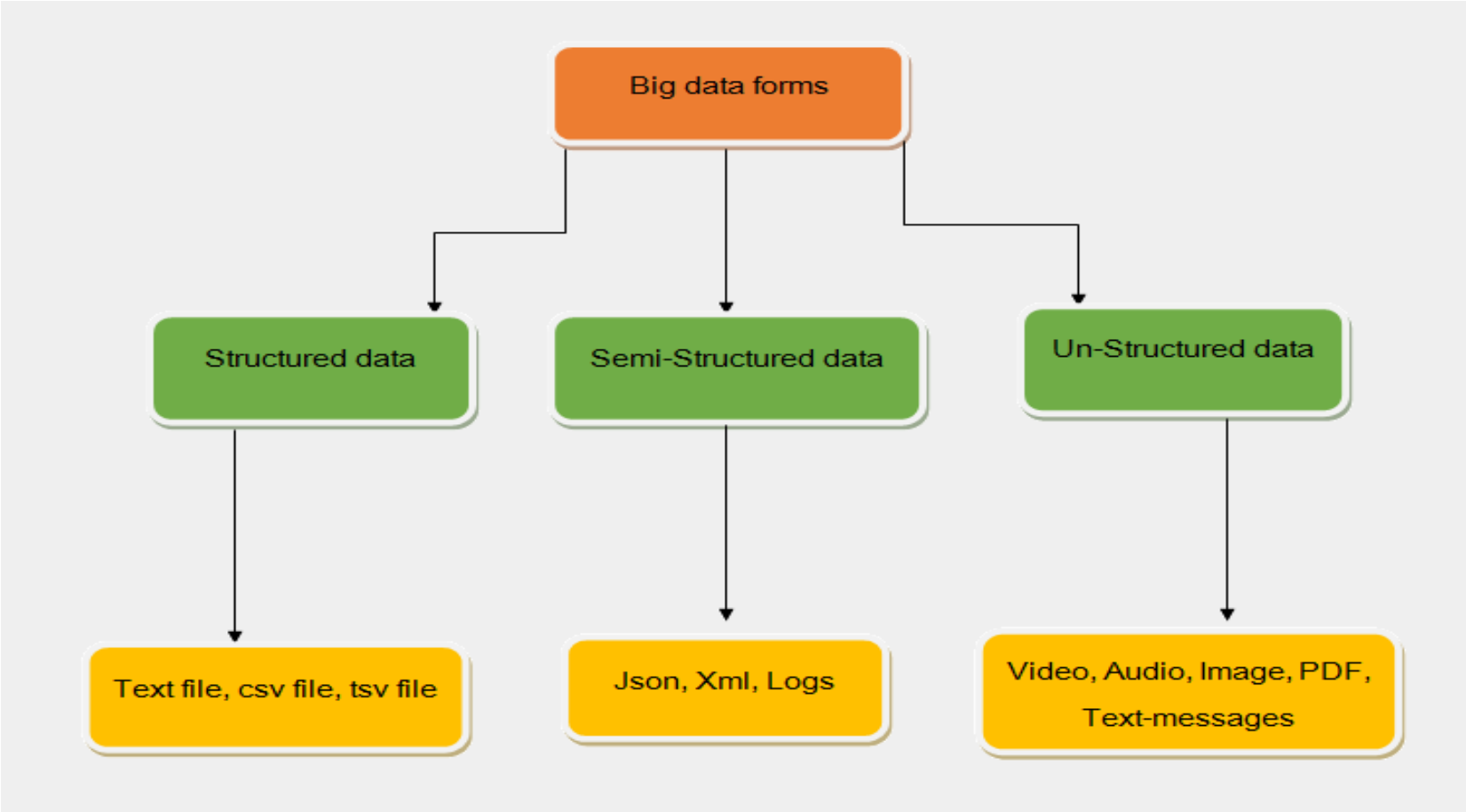
- Semi-structured data refers to **data that is not captured or formatted in conventional ways**. Semi-structured data does not follow the format of a tabular data model or relational databases because it does not have a fixed schema.

- Some source for semi-structured data include:

❑File systems such as Web data in the form of cookies.

❑Data exchange formats such as JavaScript Object Notation(JSON) Data.

Semi Structured Data		
Sl.No	Name	E-mail
1.	Sams Jacobs	smj@xyz.com
2.	First Name: David Last Name: Brown	davidb@xyz.com


Big data forms
Structured data
Semi-Structured data
Un-Structured data
Text file, csv file, tsv file
Json, Xml, Logs
Video, Audio, Image, PDF, Text-messages

Elements of Big Data

- According to Gartner, data is growing at the rate of 59% every year.
- This growth can be depicted in terms of the following four Vs:

❑ Volume

❑ Velocity

❑ Variety

❑ Veracity

Volume

- Volume is the amount of data generated by organizations or individuals.
- Today, the volume of data to reach in most organizations is approcing Exabyte's. Some experts predicts the volume of data to reach zettabytes in the coming years.
- The internet generates a huge amount of data. The followings figures help us to get an idea of the internet traffic:

❑ Internet has around 14.3 trillion live Web pages, and 48 billion Web pages are indexed by Google Inc; 14 billion Web pages are indexed by Microsoft Bing.

- Internet has around 672 Exabyte's of accessible data.
- Total world-wide internet traffic in the year 2013 was 43,639 perabytes.
- Over 9,00,000 servers are owned by Google Inc., which is the largest in the world.
- Total data stored on the internet is over 1 yottabyte.

Velocity

- Velocity describes the rate at which data is generated, captured and shared.
- Enterprise can capitalize on data only if it is captured and shared in real time.
- Information processing such as CRM and ERP face problem with data, which keeps adding up but cannot be processed quickly.
- The sources of high velocity data include the following:

❑ IT devices, including routers, switches, firewalls, etc., constantly generate valuable data.

❑Social media, including Facebook posts, and other social activities, create huge amount of data, which is to be analyzed instantly at a speed because the value degrades quickly with time.

❑Portable device, including mobile, PDA, etc., also generate data at a high speed.

Variety

- We know the data is being generated at a very fast pace, now this data is generated from different types of source, such as internal, external, social and behavioral, and comes in different formats, such as images, text, videos, etc.
- Even a single source can generate data in varied formats for example, GPS and social networking sites, such as Facebook, produce data of all types, including text, images, videos, etc.

Veracity

- Veracity generally refers to the uncertainty of data, i.e., whether the obtained data is correct or consistent. Out of the huge amount of data that is generated in almost every process, only the data that is correct and consistent can be used for further analysis.
- Data when processed becomes information; however, a lot of effort goes in processing the data.
- Big data, especially in the unstructured and semi structured forms, is messy in nature, and it takes a good amount of time and expertise to clean and make it suitable for analysis.

Big Data Analytics

Thank You !