Big Data Analytics(BTAI14601)

Prepared By Karuna Patel
IT Department
Sarvajanik College of Engineering and Technology
Surat

Unit-1 Introduction to Big Data

Prepared By Prof.Karuna Patel

OUTLINES

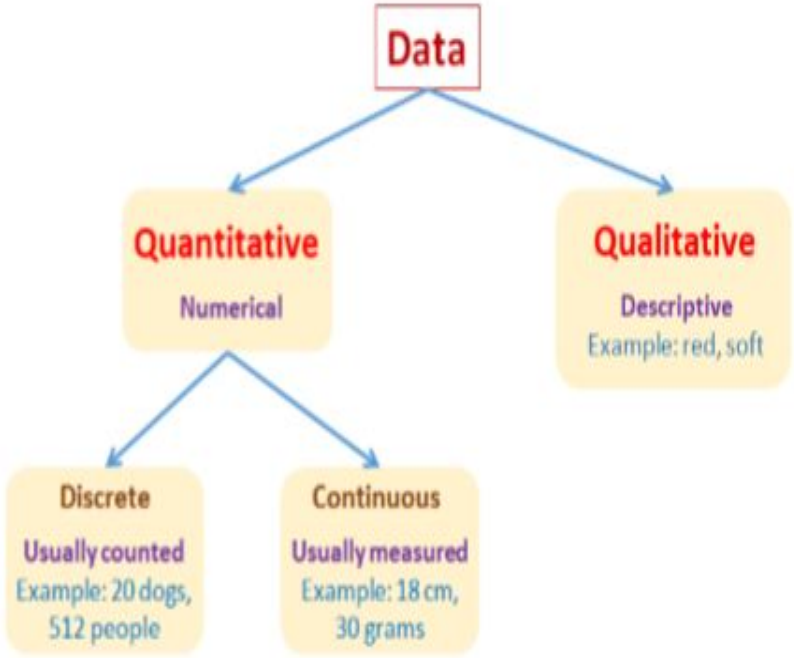
- Introduction to Big Data
- Big Data characteristics
- Challenges of Conventional System
- Types of Big Data
- Intelligent data analysis
- Traditional vs. Big Data business approach
- Challenges in Big Data Analytics – Need of big data frameworks

Big Data:

- The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

Data Comes From

Types of Data




• We create 2.5 quintillions of data everyday.
• Data comes in from various sources and has different formats
• This enormous collection of data is Big Data.
DATA!


BIG
DATA

Definition of Big Data:

- Big Data is a massive collection of data that continues to grow dramatically over time.
- It is a data set that is so huge and complicated that no typical data management technologies can effectively store or process it.
- Big Data is like regular data, but it is much larger.
- A data which are very large in size.
- Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 1015 byte size is called Big Data.

How Big is Big Data

Sources of Big Data

- Photos Videos, Likes and Comments on Social Media.
- Traffic data & GPS Signals
- Emails, Blogs and e-news
- Software logs, camera and microphone
- Huge data from Weather station and satellite that stored and manipulated to forecasting
- Digital Pictures & Videos

Big Data Analytics

- It is a process to extract meaningful information from big data such as hidden patterns, unknown correlation, market trends and customer preferences.
- Advantages:

1. Used for risk management
2. Product development and innovation
3. Helps in quicker and better decision making in organization

Lifecycle of big data analytics

1.Business case Evaluation (Reason and goal behind analytics)

2. Identification of data (Data source)

3. Data filtering (Remove the corrupt data))

4. Data Extraction (Check compatibility with tool)

5. Data Aggregation(Data integration)

6. Data Analysis (use tool)

7. Final Result

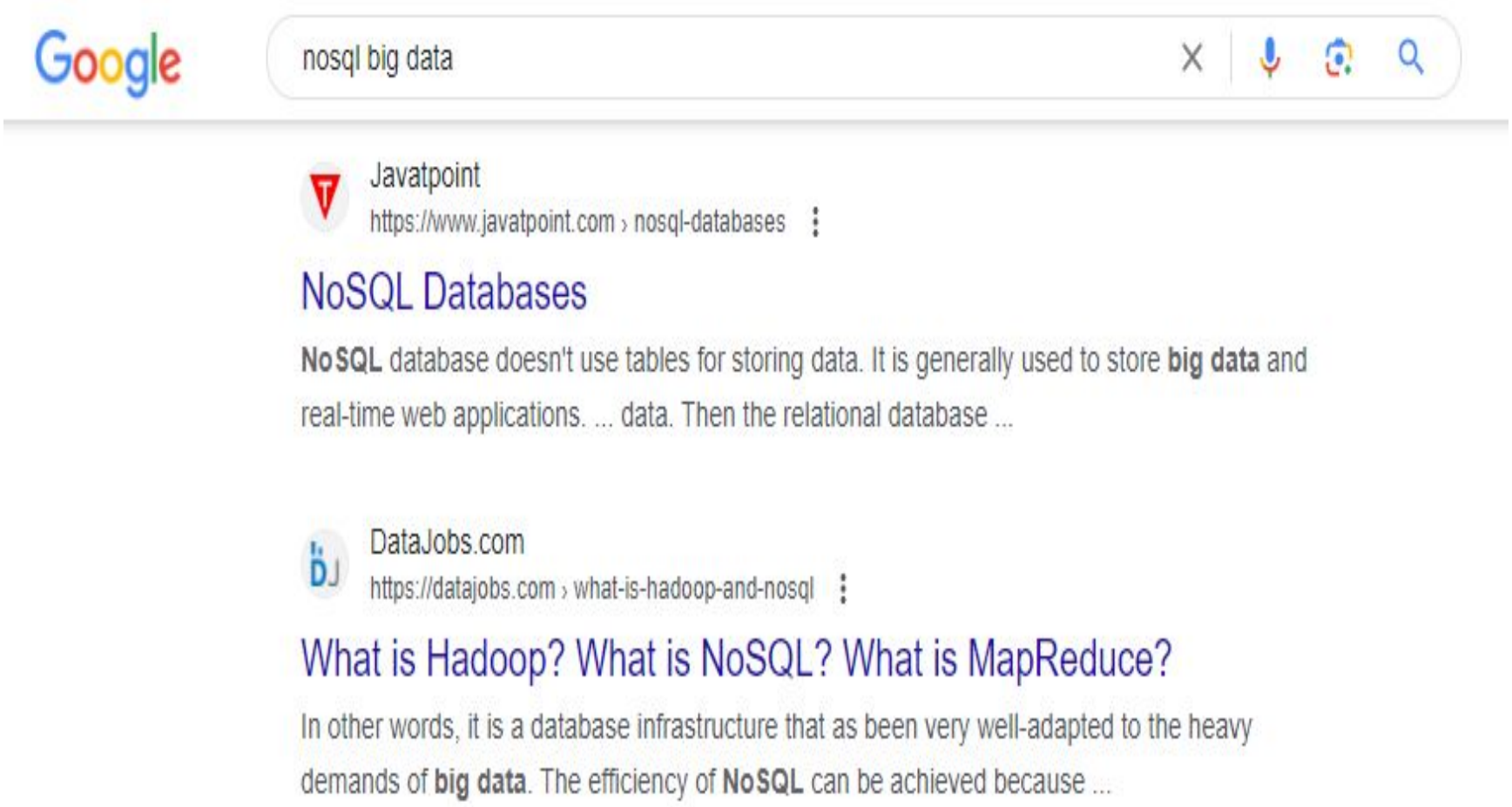
Types of Big Data

1. Unstructured
2. Semi-structured
3. Structured

1.Unstructured:

- Any data with unknown form or the structure is classified as unstructured data.
- In addition to the size being huge, unstructured data poses multiple challenges in terms of its processing for deriving value out of it.
- Typical example of unstructured data is, a heterogeneous data source containing a combination of simple text files, images, videos like search in Google Engine.
- Now a days organizations have wealth of data available with them but unfortunately they don't know how to derive value out of it since this data is in its raw form or unstructured format.

Example


Google
nosql big data
Javatpoint
https://www.javatpoint.com › nosql-databases
NoSQL Databases
NoSQL database doesn't use tables for storing data. It is generally used to store big data and real-time web applications. ... data. Then the relational database ...
DataJobs.com
https://datajobs.com › what-is-hadoop-and-nosql
What is Hadoop? What is NoSQL? What is MapReduce?
In other words, it is a database infrastructure that as been very well-adapted to the heavy demands of big data. The efficiency of NoSQL can be achieved because ...

Examples:

- Text files: Word processing documents, presentations, notes, and PDFs
- Emails: Emails are a type of unstructured data
- Videos: Video files in formats like MP4, AVI, or MOV are unstructured data
- Photos: Photos in formats like JPG, TIFF, GIF, PNG, or RAW are unstructured data
- Audio files: Audio files in formats like MP3, WAV, or FLAC are unstructured data
- Social media posts: Social media posts are unstructured data

- Podcasts: Podcasts are unstructured data
- Machine-generated formats: Machine-generated formats are unstructured data

Structured

- Any data that can be stored, accessed and processed in the form of fixed format is termed as a "Structured" data.
- Over the period of time, talent in computer science have achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also determining value out of it.
- When size of such data grows to a huge extent, typical sizes are being in the range of multiple petabyte.
- Data stored in a relational database management system in one example of a structured data.

Examples:

Financial data: Credit card transactions, bank deposits, wire transfers, and stock information

Customer data: Names, email addresses, phone numbers, and other contact information

Dates and times: Dates can be structured as YYYY-MM-DD, and times can be structured as HH:MM:SS

Spreadsheets:

Semi-structured

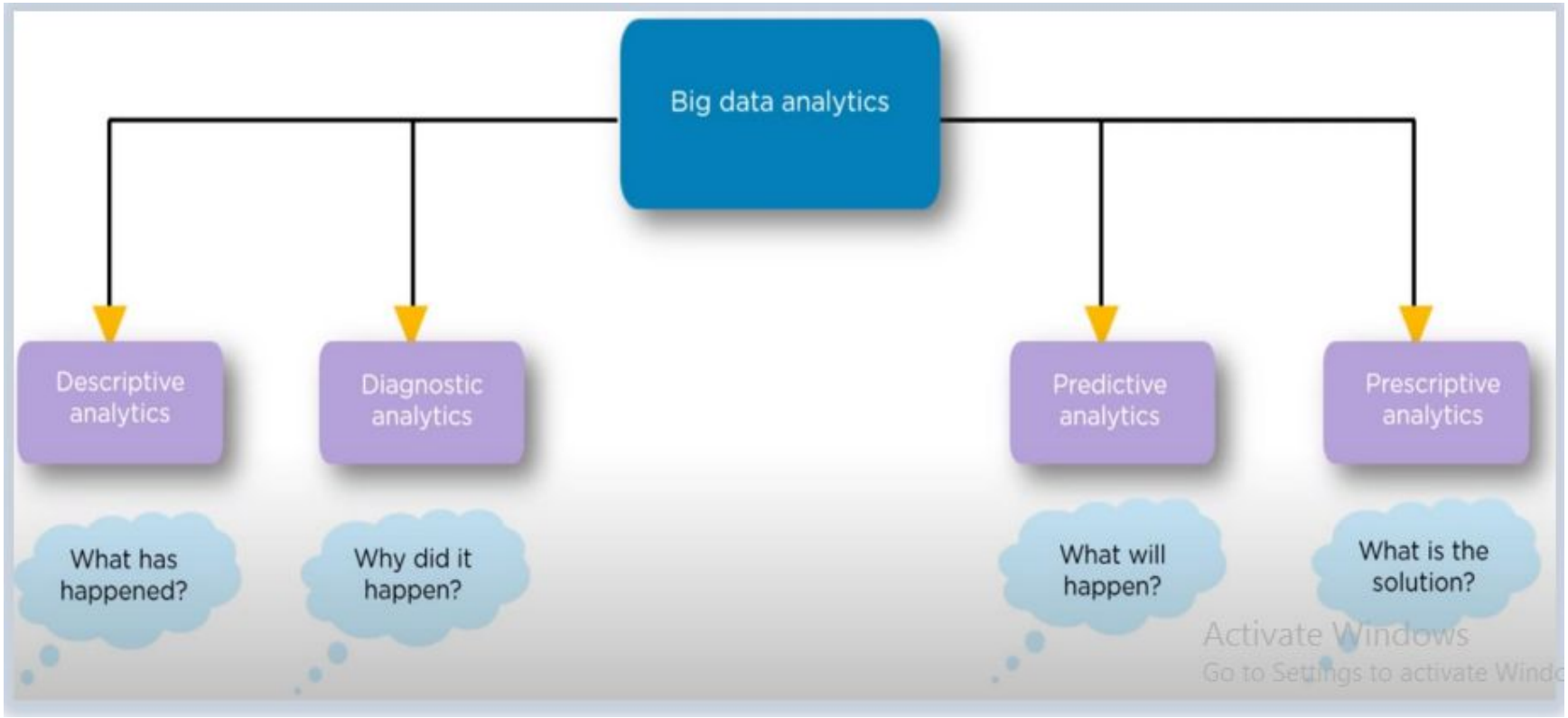
- Semi structured is the third type of big data.
- Semi-structured data can contain both the forms of data.
- Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data.
- To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data.

- Web application data, which is unstructured, consists of log files, transaction history files etc.
- Online transaction processing systems are built to work with structured data wherein data is stored in relations (tables).
- Examples:
- XML,
- Emails,
- HTML,
- Logfiles,
- JSON,
- NOSQL Databases.

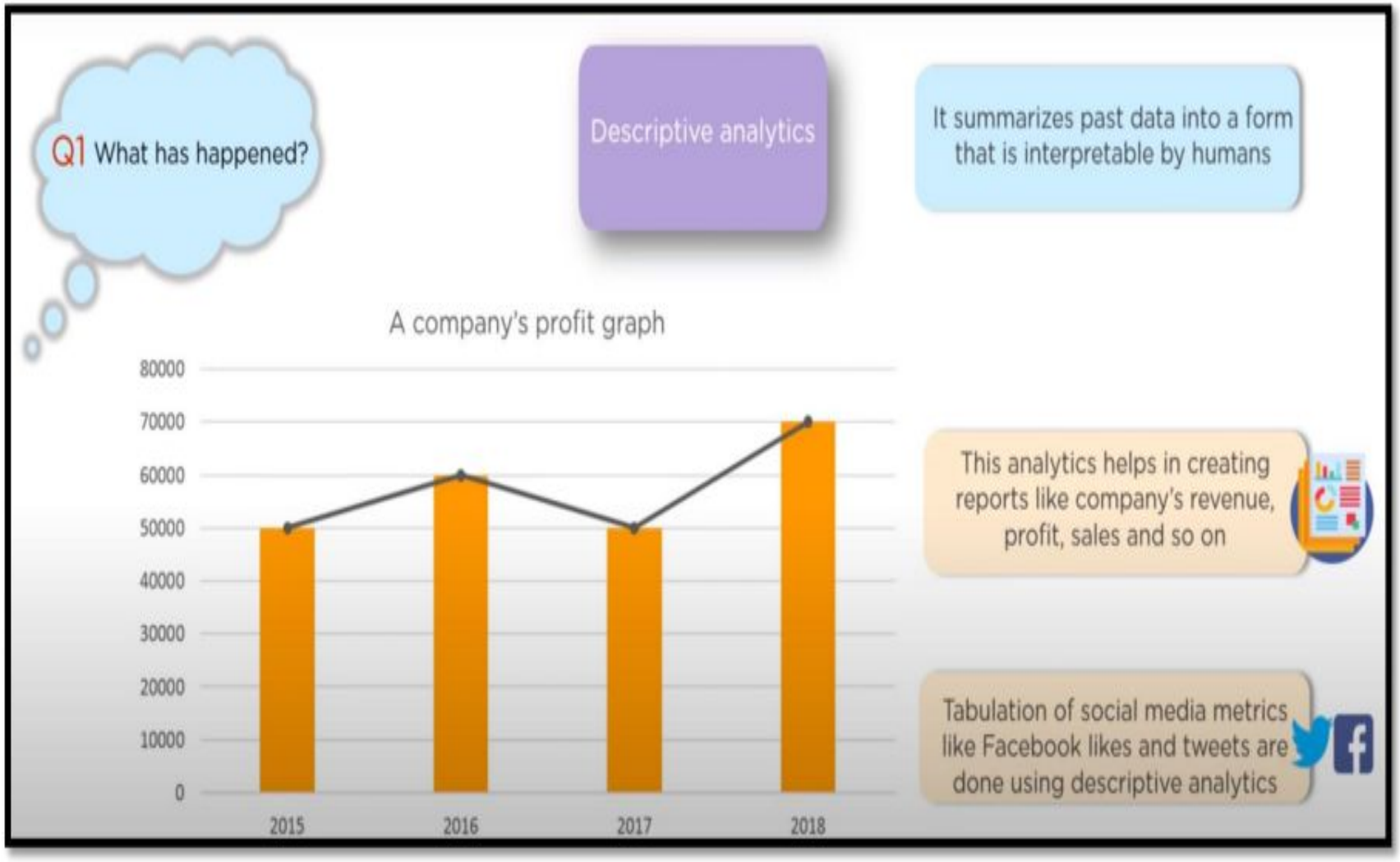
Difference between types of big data

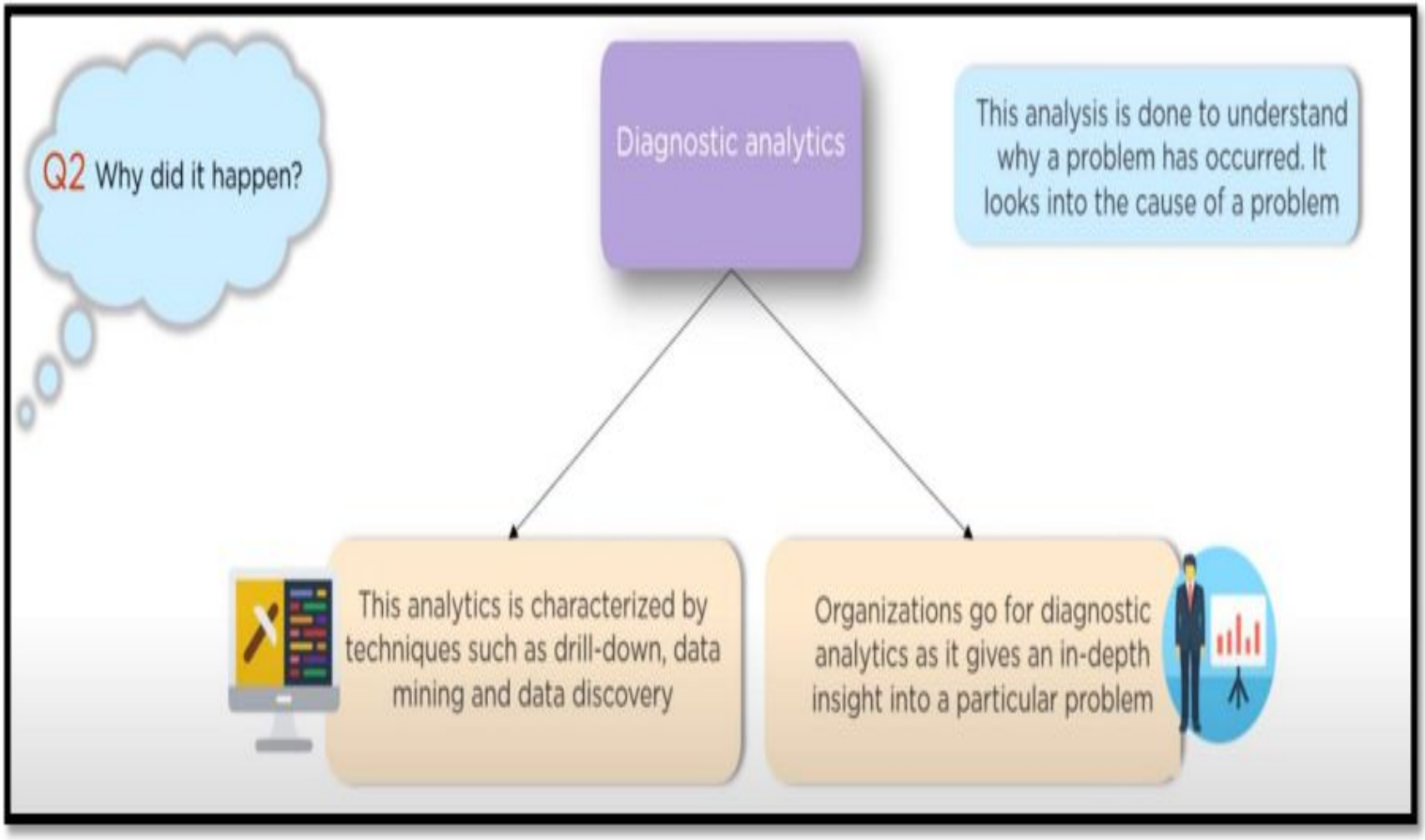
Characteristic	Structured	Unstructured	Semi Structured
Mapping	Data is mapped by Relational Database	Data is mapped on the basis of binary and simple character	Data is mapped with XML/RDF
Scalability	Data is schema dependant, which makes it less flexible and scalable	Due to no dependency, unstructured data is flexible and scalable as well	Semi structure data is more scalable and flexible than structures data, but less flexible when
Performance	Guarantees highest performance with structure query	Only Textual query is executable	Only allows anonymous queries
Organized	Highly Organised data	Unorganised Data	Partially organized Data

Types of Big Data Analytics

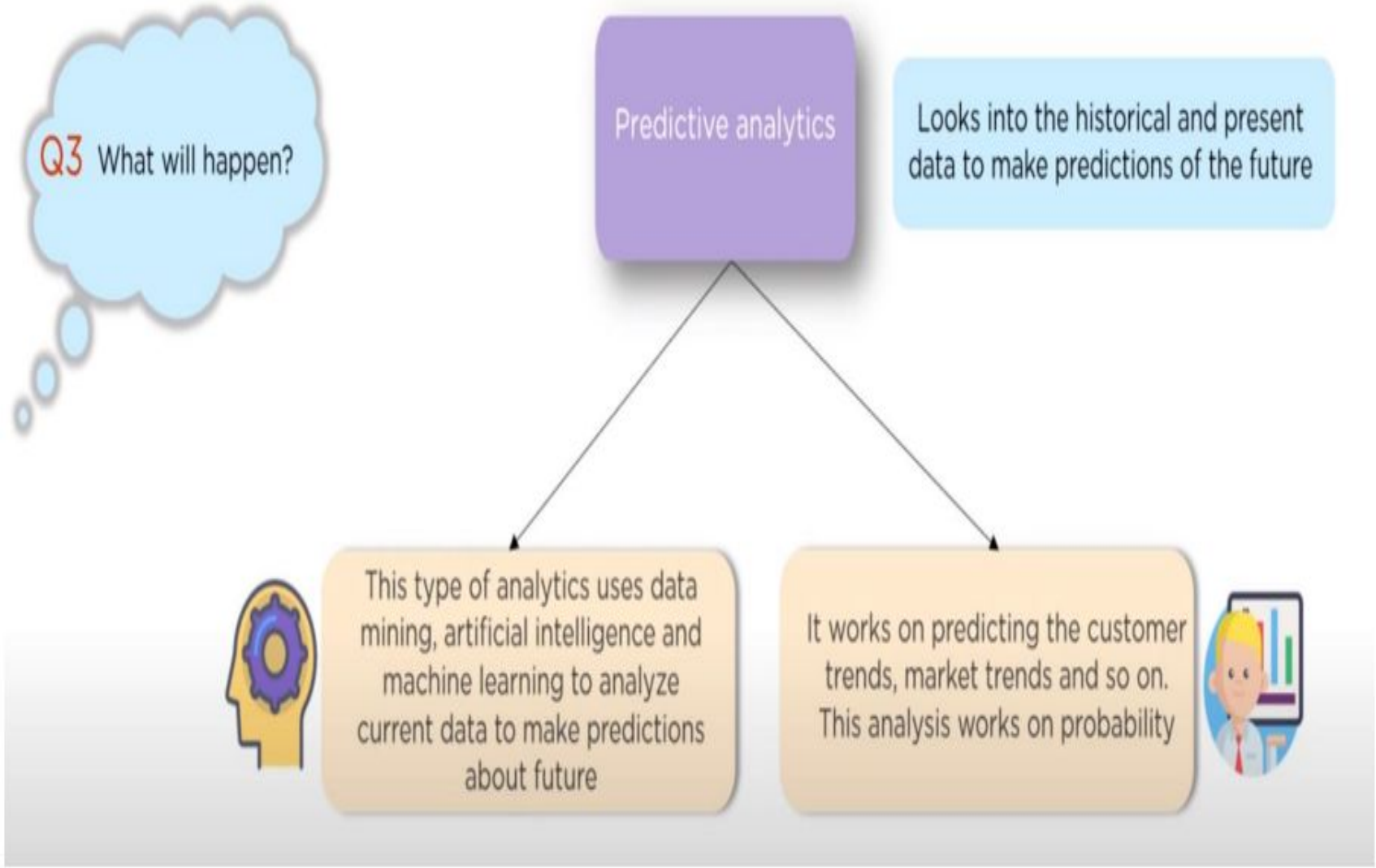


Types of Big Data Analytics




Q2 Why did it happen?
Diagnostic analytics
This analysis is done to understand why a problem has occurred. It looks into the cause of a problem
This analytics is characterized by techniques such as drill-down, data mining and data discovery
Organizations go for diagnostic analytics as it gives an in-depth insight into a particular problem


Diagnostic analytics
Why did we make fewer online sales although we had good marketing?
An ecommerce company's report shows that their sales have reduced although customers are adding products to the cart
A lot of things could have gone wrong :
• The form didn't load correctly
• The shipping fee was too high
• Not enough payment options available
404


Q3 What will happen?
Predictive analytics
Looks into the historical and present data to make predictions of the future
This type of analytics uses data mining, artificial intelligence and machine learning to analyze current data to make predictions about future
It works on predicting the customer trends, market trends and so on. This analysis works on probability

Q3 What will happen?

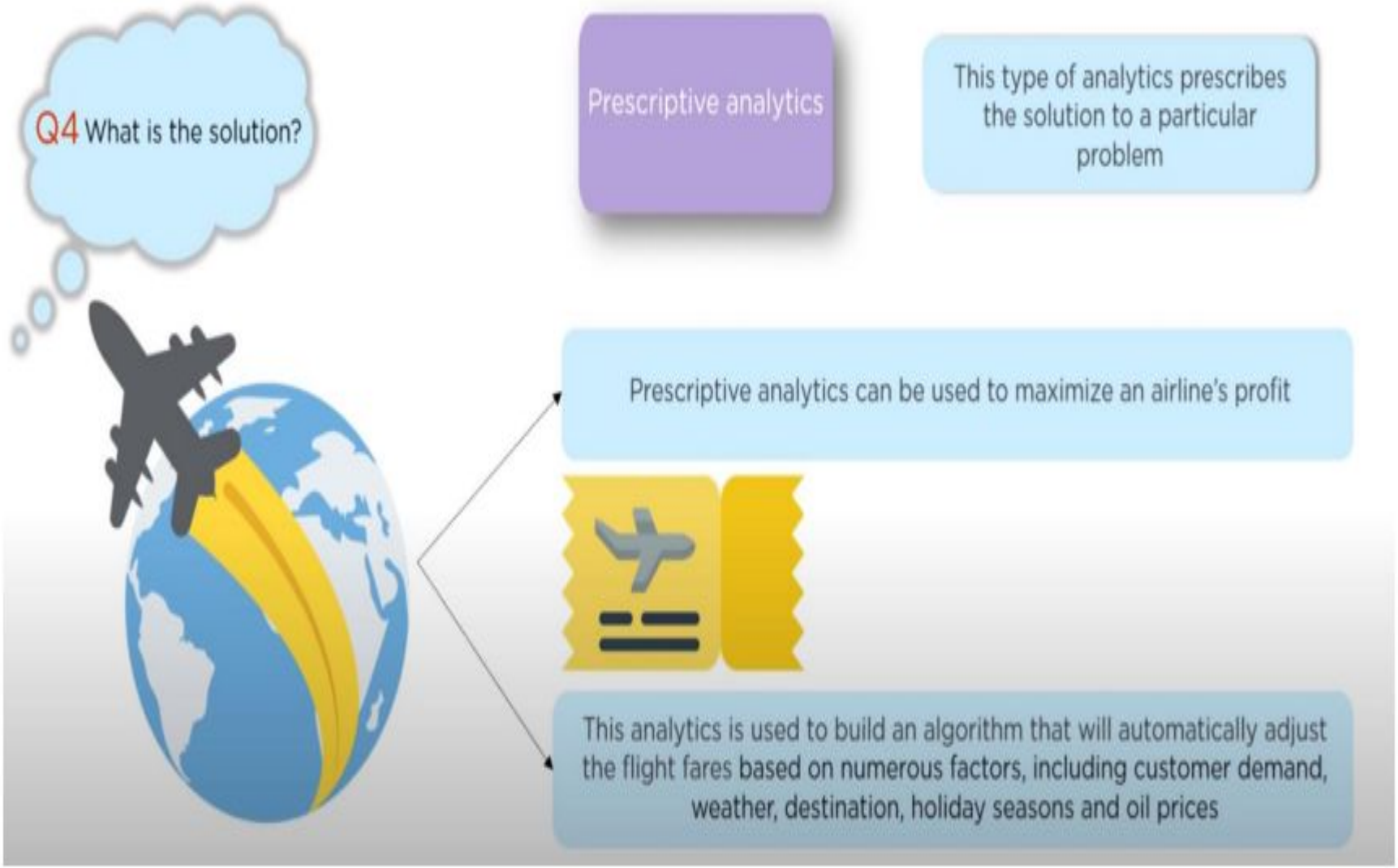
Predictive analytics

Looks into the historical and present data to make predictions of the future

PayPal

Paypal determines what kind of precautions they have to take to protect their clients against fraudulent transactions

Using predictive analytics, the company uses all the historical payment data, the user behavior data and builds an algorithm which predicts fraudulent activities


Q4 What is the solution?
Prescriptive analytics
This type of analytics prescribes the solution to a particular problem
Prescriptive analytics can be used to maximize an airline's profit
This analytics is used to build an algorithm that will automatically adjust the flight fares based on numerous factors, including customer demand, weather, destination, holiday seasons and oil prices

Big Data characteristics

5 V's of Big Data:

1. Volume: Represents the amount of data that is growing at a higher rate.
2. Veracity: Refers to uncertainty of available data. Veracity arises due to the high volume of data that brings incompleteness and inconsistency.
3. Variety: Refers to different data types i.e different data formats like text, audios, videos etc.
4. Velocity: Rate at which data grows. Social media plays a major role in the velocity of growing data.
5. Value: It refers to turning data into value.

1.Volume: Volume refers to the ‘amount of data’, which is growing day by day at a very fast pace.

- The size of data generated by humans, machines and their interactions on social media itself is massive.
- The volume of data refers to the size of the data sets that need to be analyzed and processed, which are now frequently larger than terabytes and petabytes.
- The sheer volume of the data requires distinct and different processing technologies than traditional storage and processing capabilities.

- In other words, this means that the data sets in Big Data are too large to process with a regular laptop or desktop processor.
- An example of a high-volume data set would be all credit card transactions on a day within a country.

2.Velocity: Velocity refers to the speed with which data is generated.

- High velocity data is generated with such a pace that it requires distinct (distributed) processing techniques. This flow of data is massive and continuous.

- There are 1.03 billion Daily Active Users (Facebook DAU) on Mobile as of now, which is an increase of 22% year-over-year.
- This shows how fast the number of users are growing on social media and how fast the data is getting generated daily.
- If you are able to handle the velocity, you will be able to generate insights and take decisions based on real-time data.
- Example: Amazon web service - Kinesis

3.Variety: As there are many sources which are contributing to Big Data, the type of data they are generating is different.

- It can be structured, semi-structured or unstructured.
- The variety in data types frequently requires distinct processing capabilities and specialist algorithms.
- Earlier, we used to get the data from excel and databases, now the data are coming in the form of images, audios, videos, sensor data etc.

- Hence, this variety of unstructured data creates problems in capturing, storage, mining and analyzing the data.
- Example: CCTV audio and video file that are generated at various locations.

4. Veracity :

- Veracity refers to the quality of the data that is being analyzed.
- High veracity data has many records that are valuable to analyze and that contribute in a meaningful way to the overall results.

- Low veracity data, on the other hand, contains a high percentage of meaningless data.
- The non-valuable in these data sets is referred to as noise. An example of a high veracity data set would be data from a medical experiment or trial.
- Example: Data sets like Medical Experiment

5.Value: It refers to turning data into value. By turning accessed big data into values, businesses may generate revenue.

- Value is the end game. After addressing volume, velocity, variety, variability, veracity, and visualization – which takes a lot of time, effort and resources – you want to be sure your organization is getting value from the data.
- For example, data that can be used to analyze consumer behavior is valuable for your company because you can use the research results to make individualized offers.

- Visualization: Big data visualization is the process of displaying data in charts, graphs, maps, and other visual forms.
- It is used to help people easily understand and interpret their data at a glance, and to clearly show trends and patterns that arise from this data.
- Raw data comes in a different formats, so creating data visualizations is process of gathering, managing, and transforming data into a format that's most usable and meaningful.

- Big Data Visualization makes your data as accessible as possible to everyone within your organization, whether they have technical data skills or not.

- Virality: Virality describes how quickly information gets spread across people to people (P2P) networks.
- It is measures how quickly data is spread and shared to each unique node.
- Time is a determinant factor along with rate of spread.

Challenges of Conventional System

- There are main three challenges of conventional system, which are as follows:

1. Volume of Data
2. Processing and Analyzing
3. Management of Data

1.Volume of Data: The volume of data increasing day by day, especially the data generated from machine, telecommunication service, airline services, data from sensors, etc…

- The rapid growth in data every year is coming with new source of data which are emerging.
- As per survey, the growth in volume of data is so rapid that it is expected by IBM that by 2020 around 35 zettabyte of data will get stored in the world.

2.Processing & Analyzing:

- Processing of such large volume of data is major challenge and is very difficult.
- Organization make use of such large volume of data by analyzing in order to achieve their business goals.
- Taking out insights from such large amount of data is time consuming and it also takes lot of effort to do.
- Processing and analyzing of data is also costly since the data is in different format and is complex.

3.Management of Data:

- As the data gathered have different formats like structured, semi-structured and unstructured, it is very challenging to manage such different variety of data.

- Intelligent data analysis:

Intelligent Data Analysis (IDA) is one of the major issues in the field of artificial intelligence and information.

- Intelligent data analysis reveals implicit, previously unknown and potentially valuable information or knowledge from large amounts of data.
- It also helps in making a decision.
- All zones of data visualization, data pre-preparing(combination, altering, change, separating, examining), data engineering, database mining procedure,

- devices and applications, use of domain knowledge in data analysis, big data applications, developmental algorithms, etc…

- It includes three major steps:

1. Data Preparation
2. Rules finding or data mining
3. Result validation and explanation

1.Data Preparation:

- It includes extracting or collecting relevant data from source and then creating a data set.

2. Rules finding or Data mining:

- It is working out rules contained in the dataset by means of certain methods or algorithms.

3.Result Validation and Explanation:

- This result validation means examining these rules.
- And Result explanation is giving intuitive, reasonable, and understandable description using logical reasoning.

- IDA is to extract useful knowledge, the process demands a combination of extraction, analysis, conversion, classification, organization, reasoning, and so on.
- We can imply machine learning and deep learning concept for IDA.
- It will helps in many area:
- Banking & Securities, Communications, Media, & Entertainment
- Healthcare Providers

Difference between Traditional Data & Big Data

Parameters	Traditional Data	Big Data
Structure of data	Structures are defined	Mix of Structured, semi-structured and unstructured data
Data Volume	Based on business volumes and extent of digitization	Very high, in petabytes and even more
Variety of Data Sources	Data source from database systems	Besides data from business information systems, text (emails, documents), weblogs, sensors, RFID, etc.
Velocity	Low to moderate based on volume of business	High velocity
Flow	Fixed	Continuous round the clock accumulation of data
Structured Data	Structured Data	Structured, Semi-structured and Unstructured data
Sources of data	Organizational data, trading partners data	Organizational data, RFID, Sensor data, Google searches, Social media (Linked in, Facebook, Twitter, Whatsapp, etc.)
Analytics	Provide historical view, status reports	Real-time, direct feedback from the consumer, sentiment analysis, opinions
Functions	Advises senior executives on internal business decisions, focused on analyzing data for	Customer facing functions get direct market feedback which can be used for planning market strategies, planning etc.,

Confidentiality & Data Accuracy

- Confidentiality involves setting up a set of rules and restrictions to limit access to confidential data.
- It's generally treated with access control and cryptographic mechanisms.
- The areas of research to improve the confidentiality of data in Big Data are concerned with issues such as:
- Merging and integrating of access control policies,
- Automatic management of these policies,
- Automatic administration of authorizations,
- Application of access control on Big Data platforms, etc.

Flexibility

- A traditional database is based on a fixed schema that is static in nature.
- It could only work with structured data that fit effortlessly into relational databases or tables. In reality, most data is unstructured.
- The extensive variety of unstructured data requires new methods to store and process.
- Some examples include movies and sound files, images, documents, data, text, weblogs, strings, and web content.

Data Storage Size

- The size of limit in data is huge.
- In traditional data, it's hard to store a large amount of data.
- The primary certain entirely can be taken care of data.

Where are businesses finding uses for Big Data ?

Thank You