Introduction to Big Data Analytics

Abu Raihan Mostofa Kamal

Professor, CSE Department
Islamic University of Technology (IUT)

November 27, 2023

Chapter Outline

How did big data appear in the picture?

- In between 1990s and 2000s, the growth of the World Wide Web necessitated the development of alternative database technologies capable of efficiently managing the vast and diverse data generated by online activities.
  - ✓ Companies recognized the valuable insights contained in web logs that could enhance companies' understanding of users and enable targeted advertising and marketing campaigns.
  - ✓ Not only does the expansion of user-generated data, in particular, social-media data reach substantial proportions, but also, in contrast to transactional data, these datasets are primarily in raw form or in a semistructured or unstructured format.
- Traditional relational databases, which were designed for handling structured and tabular data within enterprise environments, struggled to cope with the scale and variety of data generated by the web.

How did big data appear in the picture?

- In between 1990s and 2000s, the growth of the World Wide Web necessitated the development of alternative database technologies capable of efficiently managing the vast and diverse data generated by online activities.
  - ✓ Companies recognized the valuable insights contained in web logs that could enhance companies' understanding of users and enable targeted advertising and marketing campaigns.
  - ✓ Not only does the expansion of user-generated data, in particular, social-media data reach substantial proportions, but also, in contrast to transactional data, these datasets are primarily in raw form or in a semistructured or unstructured format.
- Traditional relational databases, which were designed for handling structured and tabular data within enterprise environments, struggled to cope with the scale and variety of data generated by the web.

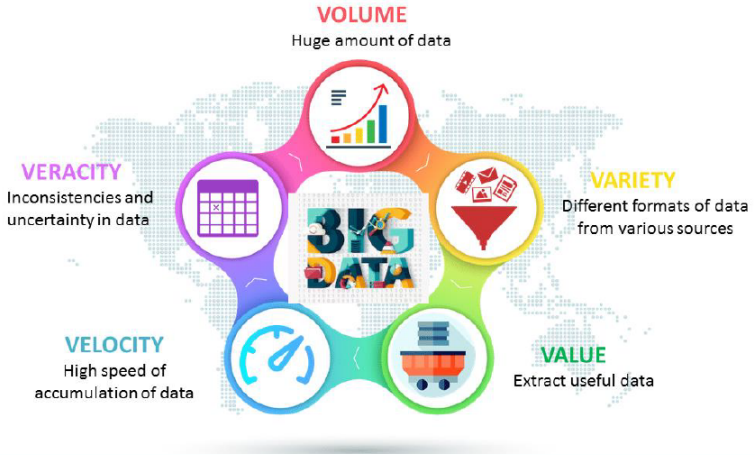
How did big data appear in the picture?

- In between 1990s and 2000s, the growth of the World Wide Web necessitated the development of alternative database technologies capable of efficiently managing the vast and diverse data generated by online activities.
  - ✓ Companies recognized the valuable insights contained in web logs that could enhance companies' understanding of users and enable targeted advertising and marketing campaigns.
  - ✓ Not only does the expansion of user-generated data, in particular, social-media data reach substantial proportions, but also, in contrast to transactional data, these datasets are primarily in raw form or in a semistructured or unstructured format.
- Traditional relational databases, which were designed for handling structured and tabular data within enterprise environments, struggled to cope with the scale and variety of data generated by the web.

How did big data appear in the picture?

- In between 1990s and 2000s, the growth of the World Wide Web necessitated the development of alternative database technologies capable of efficiently managing the vast and diverse data generated by online activities.
  - ✓ Companies recognized the valuable insights contained in web logs that could enhance companies' understanding of users and enable targeted advertising and marketing campaigns.
  - ✓ Not only does the expansion of user-generated data, in particular, social-media data reach substantial proportions, but also, in contrast to transactional data, these datasets are primarily in raw form or in a semistructured or unstructured format.
- Traditional relational databases, which were designed for handling structured and tabular data within enterprise environments, struggled to cope with the scale and variety of data generated by the web.

What makes it Big Data?



Characteristics of Big Data: Volume

- Volume refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit cards, M2M sensors, images, video, and whatnot. We are currently using distributed systems to store data in several locations and bring together by a software Framework like Hadoop.

Facebook alone can generate about billion messages, 45 billion times that the "like" button is recorded, and over 350 million new posts are uploaded each day. Such a huge amount of data can only be handled by Big Data Technologies.

Characteristics of Big Data: Variety

- Big Data is generated in multiple varieties (Structure Data, Semi-structure Data, Unstructure Data). Compared to traditional data, the latest trend of data is in the form of photos, videos, audio, weblogs and many more, making about 80% of the data completely unstructured.
  - ✓ Structured Data owns a dedicated data model. It also has a well-defined structure, follows a consistent order and is designed in such a way that it can be easily accessed and used.
    Example: Database Management (DBMS)
  - ✓ Semi Structured Data inherits a few properties of Structured Data, but the major part of this kind of data fails to have a definite structure and also, it does not obey the formal structure of data models such as an RDBMS.
    Example: Comma Separated Values(CSV)
  - ✓ Unstructured Data neither has a structure nor obeys to follow the formal structural rules of data models. It does not even have a consistent format and it is found to vary all the time.
    Example: Audio Files, Images etc

Characteristics of Big Data: Velocity

- Velocity plays a major role compared to the others, there is no point in investing so much to end up waiting for the data. So, the major aspect of Big Data is to provide data on demand and at a faster pace.

> In the year 2000, Google was receiving 32.8 million searches per day. As for 2018, Google was receiving 5.6 billion searches per day!

Approximate monthly active users as of 2018:

- ✓ Facebook: 2.41 billion
- ✓ Instagram: 1 billion
- ✓ Twitter: 320 million
- ✓ LinkedIn: 575 million

Characteristics of Big Data: Veracity & Value

- Veracity refers to the assurance of quality/integrity/credibility/accuracy of the data. Since the data is collected from multiple sources, we need to check the data for accuracy before using it for business insights.
- Value is the major issue that we need to concentrate on. Just because we collected lots of data, it's of no value unless we garner some insights out of it. Value refers to how useful the data is in decision-making. We need to extract the value of the Big Data using proper analytics.

Source of Big Data

Data come from all the corners–

- Social network
- Environmental data
- Financial data
- Medical data
- Surveillance data

Types of Data & their Usages

- Web data. Customer-level web behavior data such as page views, searches, reading reviews, and purchasing, can be captured. Enhances performance in areas such as next best offer, similar preference, customer segmentation, and targeted advertisement.
- Text data. email, news, documents, etc. It is used to determine the authenticity of news.
- Time and location data. GPS and mobile phone as well as Wi-Fi connection make time and location information. Helps for detecting potential problems, capacity planning, and other related decisions.
- Smart grid and sensor data. Data collected from High-end equipment (Signal, Image, Video). Helps to track the status and predict the chances of problems with the equipment.
- Social network data. Within social network sites like Facebook, LinkedIn, and Instagram, it is possible to do link analysis to uncover the network of a given user. Gives insights into what advertisements might appeal to given users.

Types of Data & their Usages

- Web data. Customer-level web behavior data such as page views, searches, reading reviews, and purchasing, can be captured. Enhances performance in areas such as next best offer, similar preference, customer segmentation, and targeted advertisement.
- Text data. email, news, documents, etc. It is used to determine the authenticity of news.
- Time and location data. GPS and mobile phone as well as Wi-Fi connection make time and location information. Helps for detecting potential problems, capacity planning, and other related decisions.
- Smart grid and sensor data. Data collected from High-end equipment (Signal, Image, Video). Helps to track the status and predict the chances of problems with the equipment.
- Social network data. Within social network sites like Facebook, LinkedIn, and Instagram, it is possible to do link analysis to uncover the network of a given user. Gives insights into what advertisements might appeal to given users.

Types of Data & their Usages

- Web data. Customer-level web behavior data such as page views, searches, reading reviews, and purchasing, can be captured. Enhances performance in areas such as next best offer, similar preference, customer segmentation, and targeted advertisement.
- Text data. email, news, documents, etc. It is used to determine the authenticity of news.
- Time and location data. GPS and mobile phone as well as Wi-Fi connection make time and location information. Helps for detecting potential problems, capacity planning, and other related decisions.
- Smart grid and sensor data. Data collected from High-end equipment (Signal, Image, Video). Helps to track the status and predict the chances of problems with the equipment.
- Social network data. Within social network sites like Facebook, LinkedIn, and Instagram, it is possible to do link analysis to uncover the network of a given user. Gives insights into what advertisements might appeal to given users.

Types of Data & their Usages

- Web data. Customer-level web behavior data such as page views, searches, reading reviews, and purchasing, can be captured. Enhances performance in areas such as next best offer, similar preference, customer segmentation, and targeted advertisement.
- Text data. email, news, documents, etc. It is used to determine the authenticity of news.
- Time and location data. GPS and mobile phone as well as Wi-Fi connection make time and location information. Helps for detecting potential problems, capacity planning, and other related decisions.
- Smart grid and sensor data. Data collected from High-end equipment (Signal, Image, Video). Helps to track the status and predict the chances of problems with the equipment.
- Social network data. Within social network sites like Facebook, LinkedIn, and Instagram, it is possible to do link analysis to uncover the network of a given user. Gives insights into what advertisements might appeal to given users.

Types of Data & their Usages

- Web data. Customer-level web behavior data such as page views, searches, reading reviews, and purchasing, can be captured. Enhances performance in areas such as next best offer, similar preference, customer segmentation, and targeted advertisement.
- Text data. email, news, documents, etc. It is used to determine the authenticity of news.
- Time and location data. GPS and mobile phone as well as Wi-Fi connection make time and location information. Helps for detecting potential problems, capacity planning, and other related decisions.
- Smart grid and sensor data. Data collected from High-end equipment (Signal, Image, Video). Helps to track the status and predict the chances of problems with the equipment.
- Social network data. Within social network sites like Facebook, LinkedIn, and Instagram, it is possible to do link analysis to uncover the network of a given user. Gives insights into what advertisements might appeal to given users.

Huge Data but very little insight!

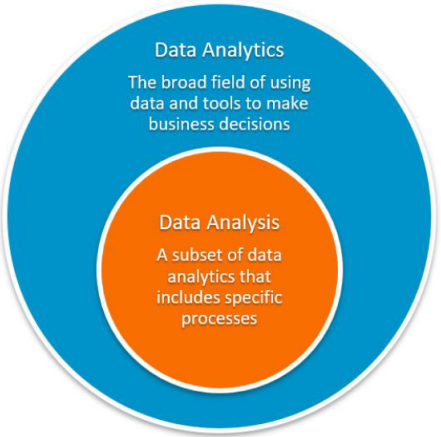
Figure: Percentage of data is increasing while the data can be processed are declining

What is Data Analytics?

- Data analytics is a discipline focused on extracting insights from data.
- Comprises the processes, tools, and techniques of data analysis and management, including the collection, organization, and storage of data.
- The primary aim of data analytics is to apply statistical analysis and technologies on data to find trends and solve problems.
- Data analytics has become increasingly important in the enterprise as a means for analyzing and shaping business processes and improving decision-making and business results.

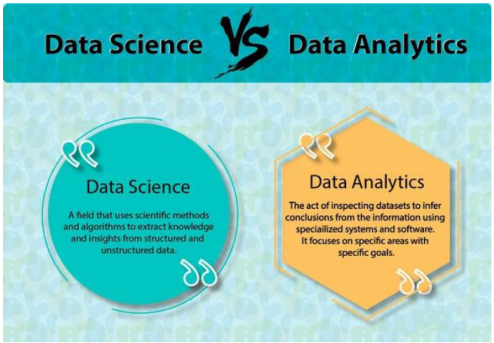
Data Analytics Vs. Data Analysis

- Data analysis is a subset of data analytics concerned with examining, cleansing, transforming, and modeling data to derive conclusions.
- Data analytics includes the tools and techniques used to perform data analysis.



Data Analytics Vs. Data Science

- Data analytics is a component of data science, used to understand what an organization's data looks like.
- Generally, the output of data analytics is reports and visualizations. While data science takes the output of analytics to study and solve problems.
- Data analytics describes the current or historical state of reality, whereas data science uses that data to predict and/or understand the future.



Different types of Data Analytics

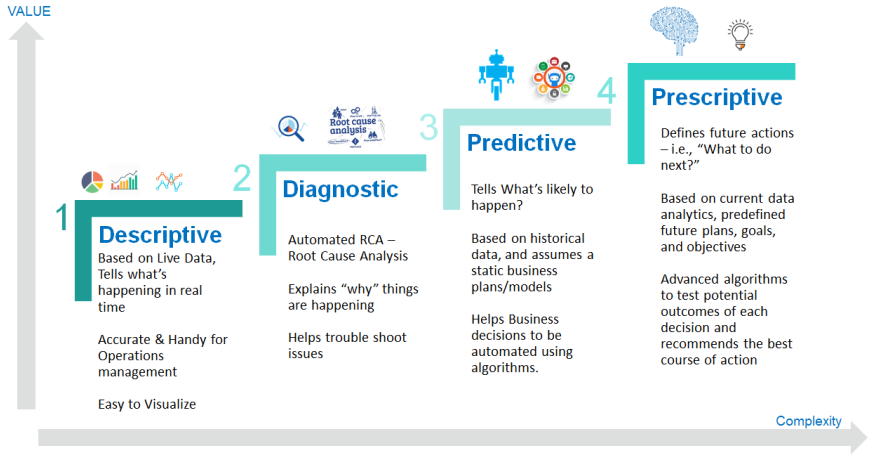


Figure: 4 types of Data Analytics

Data Analytics Methods & Techniques

There are different data analytics methods. Some are:

- Regression analysis
- Monte Carlo simulation
- Factor analysis
- Cohort analysis
- Cluster analysis
- Time Series analysis
- Sentiment analysis

Based on the data type and purpose of task, the method can be chosen.

What is Big Data Analytics?

- Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions.
- Use advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.
- Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable.

A Use Case of Big Data Analysis

Scenario

Let's say, for example, user A is browning a website X for a shirt. Unfortunately, the shirt the user A is searching for is out of stock. However, website X shows similar shirts based on his search; hence user A ends up buying two to three shirts now instead of one. The website X will again send a mail to user A once the shirt he was initially looking for is back in stock. Here the goal of more sales and happy customers is fulfilled.

A Use Case of Big Data Analysis (Cont.)

But how exactly did this happen?

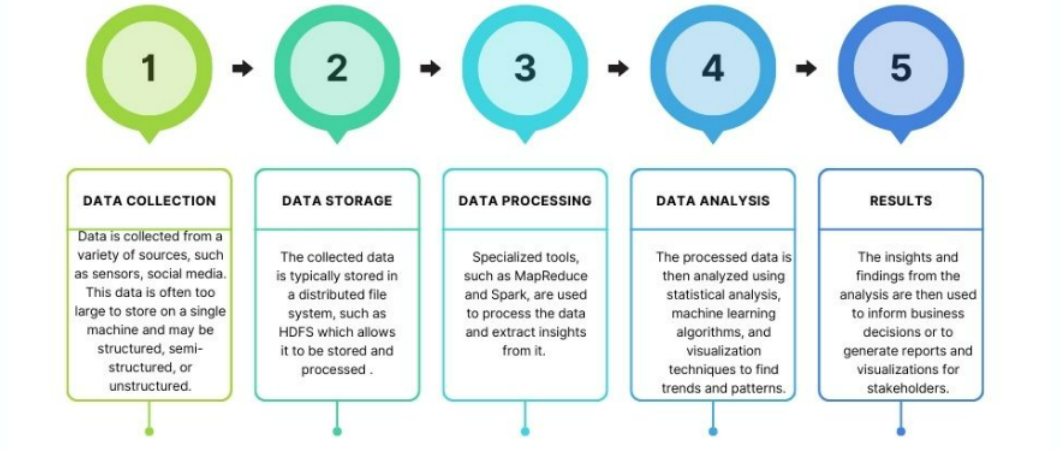
Over time, website X collects lots of data (Volume) about many customers like Users Similarly, the data is collected in the food, games, social media engagement, etc (Variety). Hence website X collects data at different rates (Velocity). Some over a period and a few live data. The algorithms that can analyze the customer's behavior are used in this aspect hence making the best use of the data.

A Use Case of Big Data Analysis (Cont.)

What advantages did this bring about?

- The customer's needs will be fulfilled without extensive search (in our case, shirt).
- Greater Revenue by showing and recommending what customer wants.
- Continuous learning by the system, through which the system suggestions can be completely accurate in the future that might match the customer's likes and wants.

How does Big Data Analysis work?



When to Consider a Big Data Solution?

1. You're limited by your current platform or environment because you can't process the amount of data that you want to process.
2. You want to involve new sources of data, but you can't, because it doesn't fit into schema-defined rows and columns without sacrificing fidelity or the richness of the data.
3. You need to ingest data as quickly as possible and need to work with a schema-on-demand.
4. You want to analyze not just raw structured data, but also semi-structured and unstructured data from a wide variety of sources.
5. You're not satisfied with the effectiveness of your algorithms or models (when all, or most, of the data, needs to be analyzed or when a sampling of the data isn't going to work).
6. You aren't completely sure where the investigation will take you, and you want elasticity of computing, storage, and the types of analytics that will be pursued.

When to Consider a Big Data Solution?

1. You're limited by your current platform or environment because you can't process the amount of data that you want to process.
2. You want to involve new sources of data, but you can't, because it doesn't fit into schema-defined rows and columns without sacrificing fidelity or the richness of the data.
3. You need to ingest data as quickly as possible and need to work with a schema-on-demand.
4. You want to analyze not just raw structured data, but also semi-structured and unstructured data from a wide variety of sources.
5. You're not satisfied with the effectiveness of your algorithms or models (when all, or most, of the data, needs to be analyzed or when a sampling of the data isn't going to work).
6. You aren't completely sure where the investigation will take you, and you want elasticity of computing, storage, and the types of analytics that will be pursued.

When to Consider a Big Data Solution?

1. You're limited by your current platform or environment because you can't process the amount of data that you want to process.
2. You want to involve new sources of data, but you can't, because it doesn't fit into schema-defined rows and columns without sacrificing fidelity or the richness of the data.
3. You need to ingest data as quickly as possible and need to work with a schema-on-demand.
4. You want to analyze not just raw structured data, but also semi-structured and unstructured data from a wide variety of sources.
5. You're not satisfied with the effectiveness of your algorithms or models (when all, or most, of the data, needs to be analyzed or when a sampling of the data isn't going to work).
6. You aren't completely sure where the investigation will take you, and you want elasticity of computing, storage, and the types of analytics that will be pursued.

When to Consider a Big Data Solution?

1. You're limited by your current platform or environment because you can't process the amount of data that you want to process.
2. You want to involve new sources of data, but you can't, because it doesn't fit into schema-defined rows and columns without sacrificing fidelity or the richness of the data.
3. You need to ingest data as quickly as possible and need to work with a schema-on-demand.
4. You want to analyze not just raw structured data, but also semi-structured and unstructured data from a wide variety of sources.
5. You're not satisfied with the effectiveness of your algorithms or models (when all, or most, of the data, needs to be analyzed or when a sampling of the data isn't going to work).
6. You aren't completely sure where the investigation will take you, and you want elasticity of computing, storage, and the types of analytics that will be pursued.

When to Consider a Big Data Solution?

1. You're limited by your current platform or environment because you can't process the amount of data that you want to process.
2. You want to involve new sources of data, but you can't, because it doesn't fit into schema-defined rows and columns without sacrificing fidelity or the richness of the data.
3. You need to ingest data as quickly as possible and need to work with a schema-on-demand.
4. You want to analyze not just raw structured data, but also semi-structured and unstructured data from a wide variety of sources.
5. You're not satisfied with the effectiveness of your algorithms or models (when all, or most, of the data, needs to be analyzed or when a sampling of the data isn't going to work).
6. You aren't completely sure where the investigation will take you, and you want elasticity of computing, storage, and the types of analytics that will be pursued.

When to Consider a Big Data Solution?

1. You're limited by your current platform or environment because you can't process the amount of data that you want to process.
2. You want to involve new sources of data, but you can't, because it doesn't fit into schema-defined rows and columns without sacrificing fidelity or the richness of the data.
3. You need to ingest data as quickly as possible and need to work with a schema-on-demand.
4. You want to analyze not just raw structured data, but also semi-structured and unstructured data from a wide variety of sources.
5. You're not satisfied with the effectiveness of your algorithms or models (when all, or most, of the data, needs to be analyzed or when a sampling of the data isn't going to work).
6. You aren't completely sure where the investigation will take you, and you want elasticity of computing, storage, and the types of analytics that will be pursued.

When to Consider a Big Data Solution? (Cont.)

If your answer to any of these questions is "yes," you need to consider a Big Data solution.

Application of Big Data Analytics

- Retail
  - ✓ Leading online retail platforms are wholeheartedly deploying big data throughout a customer's purchase journey, to predict trends, forecast demands, optimize pricing, and identify customer behavioral patterns.
  - ✓ Big data is helping retailers implement clear strategies that minimize risk and maximize profit.

Application of Big Data Analytics (Cont.)

- Healthcare
  - ✓ Big data is revolutionizing the healthcare industry, especially the way medical professionals in the past diagnosed and treated diseases.
  - ✓ In recent times, effective analysis and processing of big data by machine learning algorithms provide significant advantages for the evaluation and assimilation of complex clinical data, which prevent deaths and improve the quality of life by enabling healthcare workers to detect early warning signs and symptoms.

Application of Big Data Analytics (Cont.)

- Financial Services and Insurance
  - ✓ The increased ability to analyze and process big data is dramatically impacting the financial services, banking, and insurance landscape.
  - ✓ In addition to using big data for swift detection of fraudulent transactions, lowering risks, and supercharging marketing efforts, few companies are taking the applications to the next level.

Application of Big Data Analytics (Cont.)

- Manufacturing
  - ✓ Advancements in robotics and automation technologies, modern-day manufacturers are becoming more and more data-focused, heavily investing in automated factories that exploit big data to streamline production and lower operational costs.
  - ✓ Top global manufacturers are also integrating sensors into their products, capturing big data to provide valuable insights on product performance and its usage.

Application of Big Data Analytics (Cont.)

- Energy
  - ✓ To combat the rising costs of oil extraction and exploration difficulties because of economic and political turmoil, the energy industry is turning toward data-driven solutions to increase profitability.
  - ✓ Big data is optimizing every process while cutting down energy waste from drilling to exploring new reserves, production, and distribution.

Application of Big Data Analytics (Cont.)

- Logistics & Transportation
  - ✓ State-of-the-art warehouses use digital cameras to capture stock-level data, which, when fed into ML algorithms, facilitates intelligent inventory management with prediction capabilities that indicate when restocking is required.
  - ✓ In the transportation industry, leading transport companies now promote the collection and analysis of vehicle telematics data, using big data to optimize routes, driving behavior, and maintenance.

Application of Big Data Analytics (Cont.)

- Government
  - ✓ Cities worldwide are undergoing large scale transformations to become “smart”, through the use of data collected from various Internet of Things (IoT) sensors.
  - ✓ In the transportation industry, leading transport companies now promote the collection and analysis of vehicle telematics data, using big data to optimize routes, driving behavior, and maintenance.

Big Data Analytics Challenges

- Exponential data growth rate is one of the biggest challenges. Managing it will be very challenging.
- Unstructured Data is again a big problem. Data like Images, Videos, audio files, emails, and other types of files come under Unstructured Data, which is very difficult to search unless and until advanced artificial techniques are available.
- Integrating data from different sources is again a challenging task while dealing with a lack of coordination is highly possible.
- Data storage, processing, and maintaining the data quality and Data Security, using the right tools and technologies are a few other challenges in Big Data Analytics.

Big Data Analytics Challenges

- Exponential data growth rate is one of the biggest challenges. Managing it will be very challenging.
- Unstructured Data is again a big problem. Data like Images, Videos, audio files, emails, and other types of files come under Unstructured Data, which is very difficult to search unless and until advanced artificial techniques are available.
- Integrating data from different sources is again a challenging task while dealing with a lack of coordination is highly possible.
- Data storage, processing, and maintaining the data quality and Data Security, using the right tools and technologies are a few other challenges in Big Data Analytics.

Big Data Analytics Challenges

- Exponential data growth rate is one of the biggest challenges. Managing it will be very challenging.
- Unstructured Data is again a big problem. Data like Images, Videos, audio files, emails, and other types of files come under Unstructured Data, which is very difficult to search unless and until advanced artificial techniques are available.
- Integrating data from different sources is again a challenging task while dealing with a lack of coordination is highly possible.
- Data storage, processing, and maintaining the data quality and Data Security, using the right tools and technologies are a few other challenges in Big Data Analytics.

Big Data Analytics Challenges

- Exponential data growth rate is one of the biggest challenges. Managing it will be very challenging.
- Unstructured Data is again a big problem. Data like Images, Videos, audio files, emails, and other types of files come under Unstructured Data, which is very difficult to search unless and until advanced artificial techniques are available.
- Integrating data from different sources is again a challenging task while dealing with a lack of coordination is highly possible.
- Data storage, processing, and maintaining the data quality and Data Security, using the right tools and technologies are a few other challenges in Big Data Analytics.

The End