Unit-1

Why is Data Analytics important?

Data Analytics has a key role in improving your business as it is used to gather hidden insights, generate reports, perform market analysis, and improve business requirements.

What is the role of Data Analytics?

You can refer below:

- **Gather Hidden Insights** – Hidden insights from data are gathered and then analyzed with respect to business requirements.
- **Generate Reports** – Reports are generated from the data and are passed on to the respective teams and individuals to deal with further actions for a high rise in business.
- **Perform Market Analysis** – Market Analysis can be performed to understand the strengths and weaknesses of competitors.
- **Improve Business Requirement** – Analysis of Data allows improving Business to customer requirements and experience.

What is Data Analytics for Beginners?

Data Analytics refers to the techniques used to analyze data to enhance productivity and business gain. Data is extracted from various sources and is cleaned and categorized to analyze various behavioral patterns. The techniques and the tools used vary according to the organization or individual.

 # Sources of Data for Data Analysis
The actual data is then further divided mainly into two types known as:

1. **Primary data**
2. **Secondary data**

1.Primary data:

The data which is Raw, original, and extracted directly from the official sources is known as primary data. This type of data is collected directly by performing techniques such as questionnaires, interviews, and surveys. The data collected must be according to the demand and requirements of the target audience on which analysis is performed otherwise it would be a burden in the data processing.

Few methods of collecting primary data:

1. **Interview method:**
2. **Survey method:**
3. **Observation method:**
4. **Experimental method:**

2. Secondary data:

Secondary data is the data which has already been collected and reused again for some valid purpose. This type of data is previously recorded from primary data and it has two types of sources named internal source and external source.

Internal source:
These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc. The cost and time consumption is less in obtaining internal sources.

External source:
The data which can't be found at internal organizations and can be gained through external third party resources is external source data. The cost and time consumption is more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labor bureau, syndicate services, and other non-governmental publications.

Other sources:

- **Sensors data:** With the advancement of IoT devices, the sensors of these devices collect data which can be used for sensor data analytics to track the performance and usage of products.
- **Satellites data:** Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information.
- **Web traffic:** Due to fast and cheap internet facilities many formats of data which is uploaded by users on different platforms can be predicted and collected with their permission for data analysis. The search engines also provide their data through keywords and queries searched mostly.

OR

Sources of Data

The sources of data can be classified into two types: statistical and non-statistical. Statistical sources refer to data that is gathered for some official purposes, incorporate censuses, and officially administered surveys. Non-statistical sources refer to the collection of data for other administrative purposes or for the private sector.

What are the different sources of data?

The following are the two sources of data:

1. **Internal sources**

- When data is collected from reports and records of the organisation itself, they are known as the internal sources.
- For example, a company publishes its annual report' on profit and loss, total sales, loans, wages, etc.

2. **External sources**

- When data is collected from sources outside the organisation, they are known as the external sources. For example, if a tour and travel company obtains information on Karnataka tourism from Karnataka Transport Corporation, it would be known as an external source of data.

Types of Data

A) Primary data

- Primary data means first-hand information collected by an investigator.
- It is collected for the first time.
- It is original and more reliable.
- For example, the population census conducted by the government of India after every ten years is primary data.

B) Secondary data

- Secondary data refers to second-hand information.
- It is not originally collected and rather obtained from already published or unpublished sources.
- For example, the address of a person taken from the telephone directory or the phone number of a company taken from Just Dial are secondary data.

Methods of Collecting Primary Data

1. Direct personal investigation
2. Indirect oral investigation
3. Information through correspondents
4. Telephonic interview

5. Mailed questionnaire
6. The questionnaire filled by enumerators

 # Classification of Data Difference Between Structured, Semi-structured, and Unstructured Data

What is Structured Data?

This type of data consists of various addressable elements to encourage effective analysis. The structured form of data gets organized into a repository (formatted) that acts as a typical database. Structured data works with all kinds of data that one can store in the SQL database in a table that consists of columns and rows. These consist of relational keys, and one can easily map them into pre-designed fields. People mostly use and process structured data for managing data in the simplest form during the development process. Relational data is one of the most commendable examples of Structured Data.

What is Semi-Structured Data?

It is the type of information and data that does not get stored in a relational type of database but has organizational properties that facilitate an easier analysis. In other words, it is not as organized as the structured data but still has a better organization than the unstructured data. One can use some processes for storing this type of data and info in the relational database, and this process can be pretty difficult for some semi-structured data. But overall, they ease the space available for the contained information. XML data is an example of semi-structured data.

What is Unstructured Data?

It is a type of data structure that does not exist in a predefined organized manner. In other words, it does not consist of any predefined data model. As a result, the unstructured data is not at all fit for the relational database used mainstream. Thus, we have alternate platforms to store and manage unstructured data. It is pretty common in IT systems. Various organizations use unstructured data for various business intelligence apps and analytics. A few examples of the unstructured data structure are Text, PDF, Media logs, Word, etc.

Difference Between Structured, Semi-structured, and Untructured Data

Parameters	Structured Data	Semi-Structured Data	Unstructured Data
Data Structure	The information and data have a predefined organization.	The contained data and information have organizational properties- but are different from predefined structured data.	There is no predefined organization for the available data and information in the system or database.
Technology Used	Structured Data words on the basis of relational database tables.	Semi-Structured Data works on the basis of Relational Data Framework (RDF) or XML.	Unstructured data works on the basis of binary data and the available characters.
Flexibility	The data depends a lot on the schema. Thus, there is less flexibility.	The data is comparatively less flexible than unstructured data but way more flexible than the structured data.	Schema is totally absent. Thus, it is the most flexible of all.
Management of Transaction	It has a mature type of transaction. Also, there are various techniques of concurrency.	It adapts the transaction from DBMS. It is not of mature type.	It consists of no management of transaction or concurrency.
Management of Version	It is possible to version over tables, rows, and tuples.	It is possible to version over graphs or tuples.	It is possible to version the data as a whole.
Robustness	Structured data is very robust in nature.	Semi-Structured Data is a fairly new technology. Thus, it is not very robust in nature.	–

Scalability	Scaling a database schema is very difficult. Thus, a structured database offers lower scalability.	Scaling a Semi-Structured type of data is comparatively much more feasible.	An unstructured data type is the most scalable in nature.
Performance of Query	A structured type of query makes complex joining possible.	Semi-structured queries over various nodes (anonymous) are most definitely possible.	Unstructured data only allows textual types of queries.

 # Five Characteristics Of Good Quality Data!

One of the most important things to always remember is that not all data could be considered of fine quality hence making them limited in their usefulness. In order to fully realize the benefits of data, it has to be of high quality. This means that one should look out for certain characteristics in the data. These are:

1. *Data should be precise which means it should contain accurate information. Precision saves time of the user as well as their money.*
2. *Data should be relevant and according to the requirements of the user. Hence the legitimacy of the data should be checked before considering it for usage.*
3. *Data should be consistent and reliable. False data is worse than incomplete data or no data at all.*
4. *Relevance of data is necessary in order for it to be of good quality and useful. Although in today's world of dynamic data any relevant information is not complete at all times however at the time of its usage, the data has to be comprehensive and complete in its current form.*
5. *A high quality data is unique to the requirement of the user. Moreover it is easily accessible and could be processed further with ease.*

 # What is Big Data? Introduction, Types, Characteristics, Examples

What is 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.

What is Big Data?

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

What is an Example of Big Data?

Following are some of the Big Data examples-

The **New York Stock Exchange** is an example of Big Data that generates about ***one terabyte*** of new trade data per day.

Social Media

The statistic shows that ***500+terabytes*** of new data get ingested into the databases of social media site **Facebook**, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.

A single **Jet engine** can generate ***10+terabytes*** of data in ***30 minutes*** of flight time. With many thousand flights per day, generation of data reaches up to many ***Petabytes.***

Types Of Big Data

Following are the types of Big Data:

1. **Structured**
2. **Unstructured**
3. **Semi-structured**

Characteristics Of Big Data

Big data can be described by the following characteristics:

- Volume
- Variety
- Velocity
- Variability

(i) 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 solutions.

(ii) Variety – The next aspect of Big Data is its **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.

(iii) 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.

(iv) Variability – 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.

Advantages Of Big Data Processing

Ability to process Big Data in DBMS brings in multiple benefits, such as-

- 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.

- Early identification of risk to the product/services, if any
- Better operational efficiency

 # Need / Importance of Big Data

Big Data does not take care of how much data is there, but how it can be used. Data can be taken from various sources for analyzing it and finding answers which enable:

- Reduction in cost.
- Time reductions.
- New product development with optimized offers.
- Well-groomed decision making.

Challenges of Big data

- **Rapid Data Growth:** The growth velocity at such a high rate creates a problem to look for insights using it. There no 100% efficient way to filter out relevant data.
- **Storage:** The generation of such a massive amount of data needs space for storage, and organizations face challenges to handle such extensive data without suitable tools and technologies.
- **Unreliable Data:** It cannot be guaranteed that the big data collected and analyzed are totally (100%) accurate. Redundant data, contradicting data, or incomplete data are challenges that remain within it.
- **Data Security:** Firms and organizations storing such massive data (of users) can be a target of cybercriminals, and there is a risk of data getting stolen. Hence, encrypting such colossal data is also a challenge for firms and organizations.

 # Introduction to Big Data platform

1. Big data platform is a type of IT solution that combines the features and capabilities of several big data application and utilities within a single solution.

2. It is an enterprise class IT platform that enables organization in developing, deploying, operating and managing a big data infrastructure/ environment.

3. Big data platform generally consists of big data storage, servers, database, big data management, business intelligence and other big data management utilities.

4. It also supports custom development, querying and integration with other systems.

5. The primary benefit behind a big data platform is to reduce the complexity of multiple vendors/ solutions into a one cohesive solution.

6. Big data platform are also delivered through cloud where the provider provides an all inclusive big data solutions and services.

Features of Big Data analytics platform :

1. Big Data platform should be able to accommodate new platforms and tool based on the business requirement.

2. It should support linear scale-out.

3. It should have capability for rapid deployment.

4. It should support variety of data format.

5. Platform should provide data analysis and reporting tools.

6. It should provide real-time data analysis software.

7. It should have tools for searching the data through large data sets.

 # Steps involved in data analysis are

Data Analysis Process consists of the following phases that are iterative in nature −

- Data Requirements Specification
- Data Collection
- Data Processing
- Data Cleaning
- Data Analysis
- Communication

Data Requirements Specification

The data required for analysis is based on a question or an experiment. Based on the requirements of those directing the analysis, the data necessary as inputs to the analysis is identified (e.g., Population of people). Specific variables regarding a population (e.g., Age and Income) may be specified and obtained. Data may be numerical or categorical.

Data Collection

Data Collection is the process of gathering information on targeted variables identified as data requirements. The emphasis is on ensuring accurate and honest collection of data. Data Collection ensures that data gathered is accurate such that the related

decisions are valid. Data Collection provides both a baseline to measure and a target to improve.

Data is collected from various sources ranging from organizational databases to the information in web pages. The data thus obtained, may not be structured and may contain irrelevant information. Hence, the collected data is required to be subjected to Data Processing and Data Cleaning.

Data Processing

The data that is collected must be processed or organized for analysis. This includes structuring the data as required for the relevant Analysis Tools. For example, the data might have to be placed into rows and columns in a table within a Spreadsheet or Statistical Application. A Data Model might have to be created.

Data Cleaning

The processed and organized data may be incomplete, contain duplicates, or contain errors. Data Cleaning is the process of preventing and correcting these errors. There are several types of Data Cleaning that depend on the type of data. For example, while cleaning the financial data, certain totals might be compared against reliable published numbers or defined thresholds. Likewise, quantitative data methods can be used for outlier detection that would be subsequently excluded in analysis.

Data Analysis

Data that is processed, organized and cleaned would be ready for the analysis. Various data analysis techniques are available to understand, interpret, and derive conclusions based on the requirements. Data Visualization may also be used to examine the data in graphical format, to obtain additional insight regarding the messages within the data.

Statistical Data Models such as Correlation, Regression Analysis can be used to identify the relations among the data variables. These models that are descriptive of the data are helpful in simplifying analysis and communicate results.

The process might require additional Data Cleaning or additional Data Collection, and hence these activities are iterative in nature.

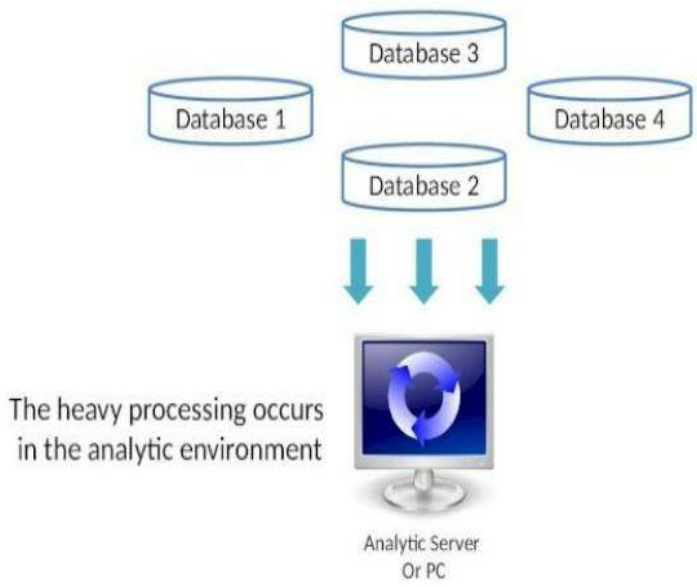
Communication

The results of the data analysis are to be reported in a format as required by the users to support their decisions and further action. The feedback from the users might result in additional analysis.

The data analysts can choose data visualization techniques, such as tables and charts, which help in communicating the message clearly and efficiently to the users. The analysis tools provide facility to highlight the required information with color codes and formatting in tables and charts.

Evolution of analytics scalability

1. In analytic scalability, we have to pull the data together in a separate analytics environment and then start performing analysis.



2. Analysts do the merge operation on the data sets which contain rows and columns.

3. The columns represent information about the customers such as name, spending level, or status.

4. In merge or join, two or more data sets are combined together. They are typically merged / joined so that specific rows of one data set or table are combined with specific rows of another.

5. Analysts also do data preparation. Data preparation is made up of joins, aggregations, derivations, and transformations. In this process, they pull data from various sources and merge it all together to create the variables required for an analysis.

6. Massively Parallel Processing (MPP) system is the most mature, proven, and widely deployed mechanism for storing and analyzing large amounts of data.

7. An MPP database breaks the data into independent pieces managed by independent storage and central processing unit (CPU) resources.

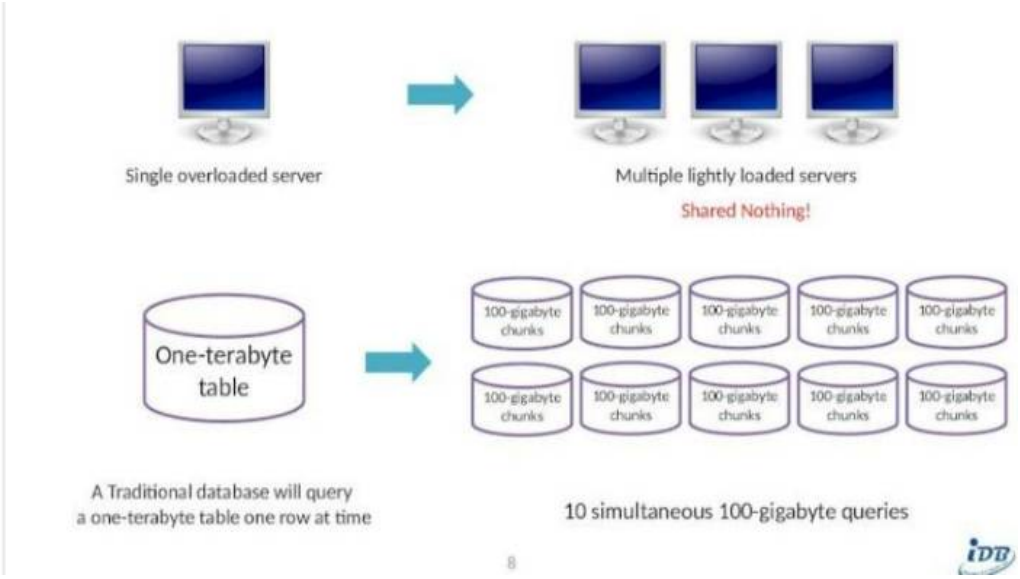


Fig. Massively parallel processing system data storage

8. MPP systems build in redundancy to make recovery easy.

9. MPP systems have resource management tools :

   a. Manage the CPU and disk space

   b. Query optimizer

#Data analytic tools.

#1 Tableau Public

What is Tableau Public

Tableau, one of the top 10 Data Analytics tools, is a simple and intuitive and tool which offers intriguing insights through data visualization. Tableau Public's million row limit, which is easy to use fares better than most of the other players in the data analytics market.

With Tableau's visuals, you can investigate a hypothesis, explore the data, and cross-check your insights.

Uses of Tableau Public

1. You can publish interactive data visualizations to the web for free.
2. No programming skills required.

3. Visualizations published to Tableau Public can be embedded into blogs and web pages and be shared through email or social media. The shared content can be made available s for downloads.

Limitations of Tableau Public

1. All data is public and offers very little scope for restricted access
2. Data size limitation
3. Cannot be connected to
4. The only way to read is via OData sources, is Excel or txt.

#2 OpenRefine

What is OpenRefine

Formerly known as GoogleRefine, the data cleaning software that helps you clean up data for analysis. It operates on a row of data which have cells under columns, quite similar to relational database tables.

Uses of OpenRefine

1. Cleaning messy data
2. Transformation of data
3. Parsing data from websites
4. Adding data to the dataset by fetching it from web services. For instance, OpenRefine could be used for geocoding addresses to geographic coordinates.

Limitations of OpenRefine

1. Open Refine is unsuitable for large datasets.
2. Refine does not work very well with big data.

#3 KNIME

What is KNIME?

KNIME, ranked among the top Data Analytics tools helps you to manipulate, analyze, and model data through visual programming. It is used to integrate various components for data mining and machine learning via its modular data pipelining concept.

Uses of KNIME

1. Rather than writing blocks of code, you just have to drop and drag connection points between activities.

2. This data analysis tool supports programming languages.
3. In fact, analysis tools like these can be extended to run chemistry data, text mining, python, and R.

Limitation of KNIME

Poor data visualization

#4 RapidMiner

Same as KNIME. Poor data visualization.

What is RapidMiner?

RapidMiner provides machine learning procedures and data mining including data visualization, processing, statistical modeling, deployment, evaluation, and predictive analytics.

RapidMiner, counted among the top 10 Data Analytics tools, is written in the Java and fast gaining acceptance.

Uses of RapidMiner

It provides an integrated environment for business analytics, predictive analysis, text mining, data mining, and machine learning.

Along with commercial and business applications, **RapidMiner** is also used for application development, rapid prototyping, training, education, and research.

Limitations of RapidMiner

1. RapidMiner has size constraints with respect to the number of rows.
2. For RapidMiner, you need more hardware resources than ODM and SAS.

#5 Google Fusion Tables

What is Google Fusion Tables?

When talking about Data Analytics tools for free, here comes a much cooler, larger, and nerdier version of Google Spreadsheets. An incredible tool for data analysis, mapping, and large dataset visualization, Google Fusion Tables can be added to business analytics tools list. Ranked among the top 10 Data Analytics tools, Google Fusion Tables is fast gaining popularity.

Uses of Google Fusion Tables

1. Visualize bigger table data online:
2. Filter and summarize across hundreds of thousands of rows.
3. Combine tables with other data on the web:

You can merge two or three tables to generate a single visualization that includes sets of data. With Google Fusion Tables, you can combine public data with your own for a better visualization.

You can create a map in minutes!

Limitations of Google Fusion Tables

1. Only the first 100,000 rows of data in a table are included in query results or mapped.
2. The total size of the data sent in one API call cannot be more than 1MB.

#6 NodeXL

What is NodeXL?

NodeXL is a free and open-source network analysis and visualization software. Ranked among the top 10 Data Analytics tools, it is one of the best statistical tools for data analysis which includes advanced network metrics, access to social media network data importers, and automation.

Uses of NodeXL

This is one of the best data analysis tools in Excel that helps in:

1. Data Import
2. Graph Visualization
3. Graph Analysis
4. Data Representation

NodeXL integrates into Microsoft Excel 2007, 2010, 2013, and 2016. It opens as a workbook with a variety of worksheets containing the elements of a graph structure like nodes and edges. It can import various graph formats like adjacency matrices, Pajek .net, UCINet .dl, GraphML, and edge lists.

Limitations of NodeXL

1. Multiple seeding terms are required for a particular problem.
2. Need to run the data extractions at slightly different times.

#7 Wolfram Alpha

What is Wolfram Alpha?

Wolfram Alpha, one of the top 10 Data Analytics tools is a computational knowledge engine or answering engine founded by Stephen Wolfram. With Wolfram Alpha, you get answers to factual queries directly by computing the answer from externally sourced 'curated data' instead of providing a list of documents or web pages.

Uses of Wolfram Alpha

1. Is an add-on for Apple's Siri
2. Provides detailed responses to technical searches and solves calculus problems.
3. Helps business users with information charts and graphs, and helps in creating topic overviews, commodity information, and high-level pricing history.

Limitations of Wolfram Alpha

1. Wolfram Alpha can only deal with the publicly known number and facts, not with viewpoints.
2. It limits the computation time for each query.

#8 Google Search Operators

What is Google Search Operators?

It is a powerful resource that helps you filter Google results instantly to get the most relevant and useful information.

Uses of Google Search Operators

1. Fast filtering of Google results.
2. Google's powerful data analysis tool can help discover new information or market research.

#9 Solver

What is Excel Solver?

The Solver Add-in is a Microsoft Office Excel add-in program that is available when you install Microsoft Excel or Office. Ranked among the best-known Data Analytic tools is a linear programming and optimization tool in excel. This allows you to set constraints. It is an advanced optimization tool that helps in quick problem-solving.

Uses of Solver

The final values found by Solver are a solution to interrelation and decision. It uses a variety of methods, from nonlinear optimization and linear programming to evolutionary and genetic algorithms, to find solutions. It is one of the top 10 Data Analytic tools in use.

Limitations of Solver

1. Poor scaling is one of the areas where Excel Solver lacks.
2. It can affect solution time and quality.
3. Solver affects the intrinsic solvability of your model.

#10 Dataiku DSS

What is Dataiku DSS?

Ranked among the top 10 Data Analytic tools, Dataiku is a collaborative data science software platform that helps the team build, prototype, explore, and deliver their own data products more efficiently.

Uses of Dataiku DSS

It provides an interactive visual interface where they can build, click, and point or use languages like SQL. This data analytics tool lets you draft data preparation and modulization in seconds. Helps you coordinate development and operations by handling workflow automation, creating predictive web services, model health on a daily basis, and monitoring data.

Limitation of Dataiku DSS

1. Limited visualization capabilities
2. UI hurdles: Reloading of code/datasets
3. Inability to easily compile entire code into a single document/notebook
4. Still, need to integrate with SPARK

Process Analysis

Process Analysis is nothing but a review of the entire process flow of an organization to arrive at a thorough understanding of the process. Further, it is also helpful to set up targets for the purpose of process improvement, which is possible by eliminating unnecessary activities, reduce wastage and increasing efficiency. Thus, it ultimately ends up improving the overall performance of the business activities.

Objectives of Process analysis

1. Identify the factors that make it difficult to understand the process.
2. Ascertain completeness of the process.
3. Remove bottlenecks
4. Find redundancies
5. Ascertain the allocation of resources
6. Check out process time

Analytics and Reporting

- Analytics involves data interpreting where reporting involving presenting factual, accurate data.
- Analytics answers *why* something is happening based on the data, whereas, reporting tells *what's* happening.
- Analytics delivers recommendations, but reporting is more about organizing and summarizing data.

#Data Analytics Applications

Some of the different data analytics applications that are currently being used in several organizations across the globe are:

1. Security

Data analytics applications or, more specifically, predictive analysis has also helped in dropping crime rates in certain areas. In a few major cities like Los Angeles and Chicago, historical and geographical data has been used to isolate specific areas where crime rates could surge. On that basis, while arrests could not be made on a whim, police patrols could be increased. Thus, using applications of data analytics, crime rates dropped in these areas.

2. Transportation

Data analytics can be used to revolutionize transportation. It can be used especially in areas where you need to transport a large number of people to a specific area and require seamless transportation. This data analytical technique was applied in the London Olympics a few years ago.
For this event, around 18 million journeys had to be made. So, the train operators and TFL were able to use data from similar events, predict the number of people who would travel, and then ensure that the transportation was kept smooth.

3. Risk detection

One of the first data analytics applications may have been in the discovery of fraud. Many organizations were struggling under debt, and they wanted a solution to this problem. They already had enough customer data in their hands, and so, they applied data analytics. They used 'divide and conquer' policy with the data, analyzing recent expenditure, profiles, and any other important information to understand any probability of a customer defaulting. Eventually, it led to lower risks and fraud.

4. Risk Management

Risk management is an essential aspect in the world of insurance. While a person is being insured, there is a lot of data analytics that goes on during the process. The risk involved while insuring the person is based on several data like actuarial data and claims data, and the analysis of them helps insurance companies to realize the risk. Underwriters generally do this evaluation, but with the advent of data analysis, analytical software can be used to detect risky claims and push such claims before the authorities for further analysis.

5. Delivery

Several top logistic companies like DHL and FedEx are using data analysis to examine collected data and improve their overall efficiency. Using data analytics applications, the companies were able to find the best shipping routes, delivery time, as well as the most cost-efficient transport means. Using GPS and accumulating data from the GPS gives them a huge advantage in data analytics.

6. Fast internet allocation

While it might seem that allocating fast internet in every area makes a city 'Smart', in reality, it is more important to engage in smart allocation. This smart allocation would mean understanding how bandwidth is being used in specific areas and for the right cause.

It is also important to shift the data allocation based on timing and priority. It is assumed that financial and commercial areas require the most bandwidth during weekdays, while residential areas require it during the weekends. But the situation is much more complex. Data analytics can solve it.

For example, using applications of data analysis, a community can draw the attention of high-tech industries and in such cases, higher bandwidth will be required in such areas.

7. Reasonable Expenditure

When one is building Smart cities, it becomes difficult to plan it out in the right way. Remodeling of the landmark or making any change would incur large amounts of expenditure, which might eventually turn out to be a waste. Data analytics can be used in such cases. With data analytics, it will become easier to direct the tax money in a cost-efficient way to build the right infrastructure and reduce expenditure.

8. Interaction with customers

In insurance, there should be a healthy relationship between the claims handlers and customers. Hence, to improve their services, many insurance companies often use customer surveys to collect data. Since insurance companies target a diverse group of people, each demographic has their own preference when it comes to communication.

Data analysis can help in zeroing in on specific preferences. For example, a study showed that modern customers prefer communication through social media or online channels, while the older demographic prefers telephonic communication.

9. Planning of cities

One of the untapped disciplines where data analysis can really grow is city planning. While many city planners might be hesitant towards using data analysis in their favour, it only results in faulty cities riddled congestion. Using data analysis would help in bettering accessibility and minimizing overloading in the city.
Overall, it will generate more efficiency in the planning process. Just erecting a building in a suitable spot will not create an overall benefit for a city since it can harm the neighbors or the traffic in the area. Using data analytics and modelling, it will be easy to predict the outcome of placing a building in a specific situation and therefore, plan accordingly.

10. Healthcare

While medicine has come a long way since ancient times and is ever-improving, it remains a costly affair. Many hospitals are struggling with the cost pressures that modern healthcare has come with, which includes the use of sophisticated machinery, medicines, etc.
But now, with the help of data analytics applications, healthcare facilities can track the treatment of patients and patient flow as well as how equipment are being used in hospitals. It has been estimated that there can be a 1% efficiency gain achieved if data analytics became an integral part of healthcare, which will translate to more than $63 billion in healthcare services.

11. For Travelling

If you ever thought travelling is a hassle, then data analytics is here to save you. Data analysis can use data that shows the desires and preferences of different customers from social media and helps in optimizing the buying experience of travellers. It will also help companies customize their own packages and offer and hence boost more personalized travel recommendations with the help data collected from social media.

12. Managing Energy

Many firms engaging with energy management are making use of applications of data analytics to help them in areas like smart-grid management, optimization of energy, energy distribution, and automation building for other utility-based companies. How does data analytics help here?
Well, it helps by focusing on controlling and monitoring of a dispatch crew, network devices, and management of service outages. Since utilities integrate about millions of data points within the network performance, engineers can use data analytics to help them monitor the entire network.

13. Internet searching

When you use Google, you are using one of their many data analytics applications employed by the company. Most search engines like Google, Bing, Yahoo, AOL, Duckduckgo, etc. use data analytics. These search engines use different algorithms to

deliver the best result for a search query, and they do so within a few milliseconds. Google is said to process about 20 petabytes of data every day.

14. Digital advertisement

Data analytics has revolutionized digital advertising, as well. Digital billboards in cities as well as banners on websites, that is, most of the advertisement sources nowadays use data analytics using data algorithms. It is one of the reasons why digital advertisements are getting more CTRs than traditional advertising techniques. The target of digital advertising nowadays is focused on the analysis of the past behaviour of the user.

Key roles for a successful analytics project :

1. **Business User :**
    - The business user is the one who understands the main area of the project and is also basically benefited from the results.
    - This user gives advice and consult the team working on the project about the value of the results obtained and how the operations on the outputs are done.
    - The business manager, line manager, or deep subject matter expert in the project mains fulfills this role.

2. **Project Sponsor :**
    - The Project Sponsor is the one who is responsible to initiate the project. Project Sponsor provides the actual requirements for the project and presents the basic business issue.
    - He generally provides the funds and measures the degree of value from the final output of the team working on the project.
    - This person introduce the prime concern and brooms the desired output.

3. **Project Manager :**
    - This person ensures that key milestone and purpose of the project is met on time and of the expected quality.

4. **Business Intelligence Analyst :**
    - Business Intelligence Analyst provides business domain perfection based on a detailed and deep understanding of the data, key performance indicators (KPIs), key matrix, and business intelligence from a reporting point of view.
    - This person generally creates fascia and reports and knows about the data feeds and sources.

5. **Database Administrator (DBA) :**
    - DBA facilitates and arrange the database environment to support the analytics need of the team working on a project.
    - His responsibilities may include providing permission to key databases or tables and making sure that the appropriate

security stages are in their correct places related to the data repositories or not.

6. **Data Engineer :**
    - Data engineer grasps deep technical skills to assist with tuning SQL queries for data management and data extraction and provides support for data intake into the analytic sandbox.
    - The data engineer works jointly with the data scientist to help build data in correct ways for analysis.

7. **Data Scientist :**
    - Data scientist facilitates with the subject matter expertise for analytical techniques, data modelling, and applying correct analytical techniques for a given business issues.
    - He ensures overall analytical objectives are met.
    - Data scientists outline and apply analytical methods and proceed towards the data available for the concerned project.

 # Stages of data analytics /Life cycle of data analytics

Phases of Data Analytics Lifecycle
Phase 1: Data Discovery and Formation
Phase 2: Data Preparation and Processing
Phase 3: Design a Model
Phase 4: Model Building
Phase 5: Result Communication and Publication
Phase 6: Measuring of Effectiveness

Data Analytics Lifecycle :
The Data analytic lifecycle is designed for Big Data problems and data science projects. The cycle is iterative to represent real project. To address the distinct requirements for performing analysis on Big Data, step – by – step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing, and repurposing data.

Phase 1: Discovery –

- The data science team learn and investigate the problem.
- Develop context and understanding.
- Come to know about data sources needed and available for the project.
- The team formulates initial hypothesis that can be later tested with data.

Phase 2: Data Preparation –

- Steps to explore, preprocess, and condition data prior to modeling and analysis.

- It requires the presence of an analytic sandbox, the team execute, load, and transform, to get data into the sandbox.
- Data preparation tasks are likely to be performed multiple times and not in predefined order.
- Several tools commonly used for this phase are – Hadoop, Alpine Miner, Open Refine, etc.

Phase 3: Model Planning –

- Team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
- In this phase, data science team develop data sets for training, testing, and production purposes.
- Team builds and executes models based on the work done in the model planning phase.
- Several tools commonly used for this phase are – Matlab, STASTICA.

Phase 4: Model Building –

- Team develops datasets for testing, training, and production purposes.
- Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
- Free or open-source tools – Rand PL/R, Octave, WEKA.
- Commercial tools – Matlab , STASTICA.

Phase 5: Communication Results –

- After executing model team need to compare outcomes of modeling to criteria established for success and failure.
- Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
- Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

Phase 6: Operationalize –

- The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
- This approach enables team to learn about performance and related constraints of the model in production environment on small scale , and make adjustments before full deployment.
- The team delivers final reports, briefings, codes.
- Free or open source tools – Octave, WEKA, SQL, MADlib.