UNIT 1

Introduction to Data Analytics: Sources and nature of data, classification of data (structured, semi-structured, unstructured), characteristics of data, introduction to Big Data platform, need of data analytics, evolution of analytic scalability, analytic process and tools, analysis vs reporting, modern data analytic tools, applications of data analytics.

Data Analytics Lifecycle: Need, key roles for successful analytic projects, various phases of data analytics lifecycle – discovery, data preparation, model planning, model building, communicating results, operationalization.

Introduction to Data Analytics

Data Analytics is a field that encompasses the collection, processing, analysis, and interpretation of data to extract valuable insights and knowledge. The goal of data analytics is to provide decision makers with the information they need to make informed decisions. The following are some of the key components of data analytics:

1. **Data Collection:** The first step in data analytics is to gather data from various sources such as databases, sensors, or customer interactions.
2. **Data Cleaning:** The next step is to clean the data to remove any errors or inconsistencies that may compromise the analysis.
3. **Data Exploration and Visualization:** In this step, data analysts will explore the data using descriptive statistics, graphs, and other visualization tools to understand the underlying patterns and relationships in the data.
4. **Data Modelling:** In this step, data analysts will use statistical models and algorithms to extract insights from the data. Some common models used in data analytics include linear regression, decision trees, and neural networks.
5. **Data Interpretation:** Finally, the results of the analysis are interpreted to provide valuable insights and knowledge that can inform decision making.

There are many tools and technologies used in data analytics, including R, Python, SQL, SAS, and Tableau, among others. Data Analytics can be applied in a wide range of industries, including finance, healthcare, marketing, and e-commerce, to name a few. A career in data analytics requires a strong foundation in mathematics, statistics, and computer science, as well as the ability to think critically and solve problems.

Terminology

1. **Data Analytics:** The process of examining, cleaning, transforming, and modeling data with the goal of discovering useful information, drawing conclusions, and supporting decision making.
2. **Business Intelligence:** The use of data and analytics to support decision-making and strategy development in an organization.
3. **Data Warehousing:** The process of collecting, storing, and managing data from multiple sources for use in data analysis and reporting.

4. **Data Mining:** The process of discovering patterns and relationships in large datasets through the use of algorithms and statistical analysis.
5. **Big Data:** A term used to describe the large volumes of structured, semi-structured, and unstructured data that are generated and need to be analysed.
6. **NoSQL:** A term used to describe database management systems that are designed to handle large volumes of unstructured data and provide flexible data models.
7. **Cloud Computing:** The delivery of computing services over the internet, including storage, databases, and analytics tools.
8. **Predictive Modeling:** The process of using statistical models and algorithms to make predictions about future events.
9. **Data Visualization:** The process of creating graphical representations of data to aid in the exploration, analysis, and presentation of data.
10. **Data Cleaning:** The process of identifying and correcting errors and inconsistencies in data.
11. **Data Transformation:** The process of converting data from one format to another, or changing the structure of data to make it more suitable for analysis.
12. **Data Enrichment:** The process of adding additional information or context to data to make it more useful for analysis.
13. **Regression Analysis:** A statistical method used to model the relationship between a dependent variable and one or more independent variables.
14. **Time Series Analysis:** The analysis of data that is collected over time, with the goal of identifying trends, patterns, and forecasts.
15. **Sentiment Analysis:** The process of analyzing text data to determine the sentiment expressed, such as positive, negative, or neutral.

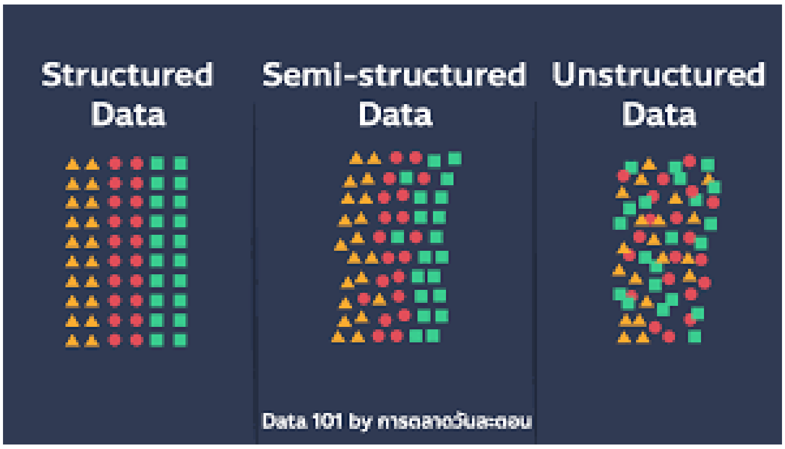
Sources and nature of data
classification of data (structured, semi-structured, unstructured)

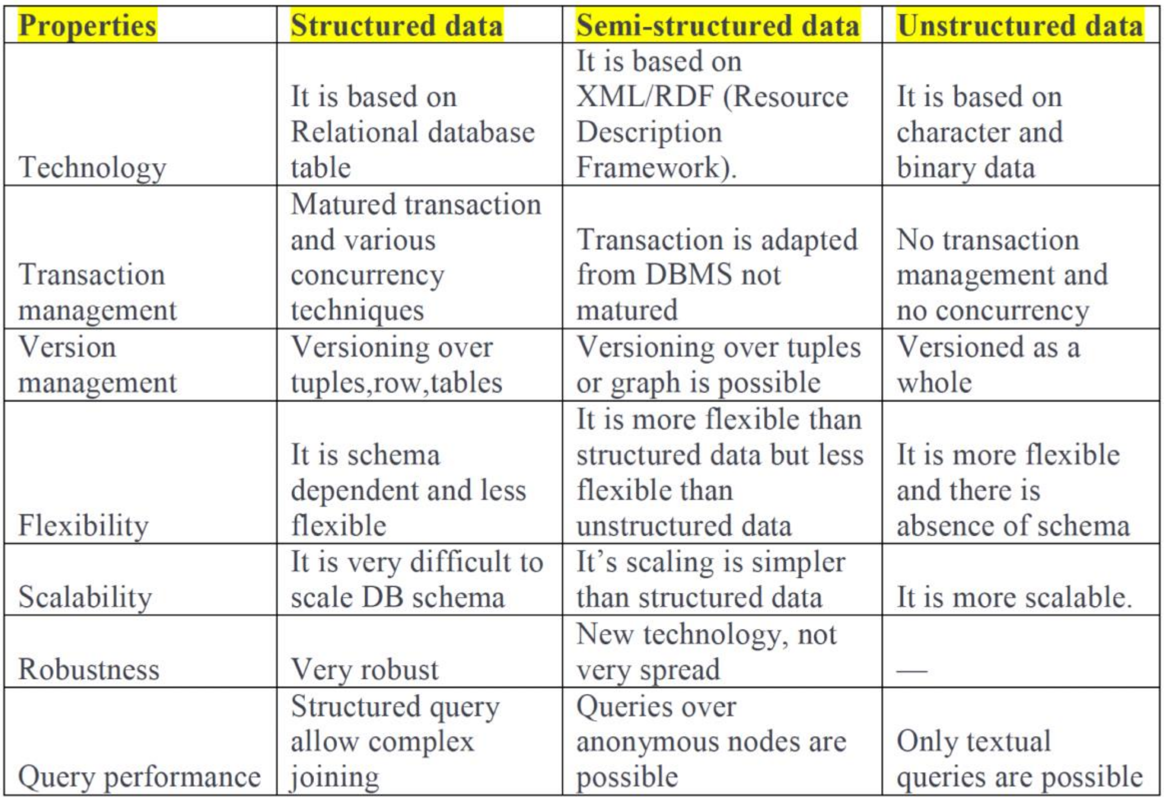
Data is the foundation of data analytics, and its source and nature play a significant role in the quality and usefulness of the analysis. The sources of data can be broadly classified into internal and external sources. Internal data is generated and collected within an organization, while external data is sourced from outside the organization.

In terms of nature, data can be classified into structured, semi-structured, and unstructured data.

1. **Structured Data:** Structured data refers to data that is organized into a fixed format, such as databases or spreadsheets. It is the easiest to analyze and has well-defined fields, columns, and rows. Structured data is often numerical in nature and can be easily processed by computers. Examples of structured data include customer transactions, employee records, and product inventory data.
2. **Semi-Structured Data:** Semi-structured data contains elements of structure, but also has unstructured elements, such as text or images. This type of data is not as easy to analyze as structured data, but it still has some structure that can be leveraged. Examples of semi-structured data include email, customer reviews, and product descriptions.
3. **Unstructured Data:** Unstructured data refers to data that does not have a pre-defined structure and is often more difficult to analyze. This type of data is usually in the form of text, audio, video, or images and does not easily fit into a database or spreadsheet. Examples of unstructured data include audio recordings, video files, and social media posts.

Historical data, real-time data, IoT data, machine-generated data, social media data, survey data, and satellite data are some other examples of data that can be analyzed in data analytics. The quality, relevance, and volume of data will play a significant role in determining the success of a data analytics project.





Properties	Structured data	Semi-structured data	Unstructured data
Technology	It is based on Relational database table	It is based on XML/RDF (Resource Description Framework).	It is based on character and binary data
Transaction management	Matured transaction and various concurrency techniques	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples,row,tables	Versioning over tuples or graph is possible	Versioned as a whole
Flexibility	It is schema dependent and less flexible	It is more flexible than structured data but less flexible than unstructured data	It is more flexible and there is absence of schema
Scalability	It is very difficult to scale DB schema	It's scaling is simpler than structured data	It is more scalable.
Robustness	Very robust	New technology, not very spread	—
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual queries are possible

Characteristics of Data

The characteristics of data play a crucial role in the data analytics process and impact the results obtained. The following are some key characteristics of data in data analytics:

1. **Volume:** Refers to the amount of data generated and stored, which can be in terabytes, petabytes, or even exabytes. With the rise of digital technology, the volume of data has increased dramatically, requiring new storage and processing solutions.
2. **Velocity:** Refers to the speed at which data is generated and processed. This includes real-time data streams and batch processing. High velocity data requires fast processing and storage solutions to handle the high volume of incoming data.
3. **Variety:** Refers to the different types of data that exist, including structured data, semi-structured data, and unstructured data. Variety impacts the methods and techniques used to process, store, and analyze data.
4. **Veracity:** Refers to the accuracy, trustworthiness, and completeness of data. Veracity is important in ensuring that data is reliable and suitable for use in analytics. Poor data quality can lead to incorrect results and biased insights.

5. **Value:** Refers to the information and insights that can be derived from data. Data must have value to be useful in decision-making and problem-solving. The value of data can vary depending on the context and the problem being solved.
6. **Variability:** Refers to the changes and fluctuations in data over time. This includes trends, patterns, and anomalies in data. Understanding variability is crucial for detecting patterns and trends in data.
7. **Complexity:** Refers to the difficulty in analyzing and processing data, particularly in big data environments. Complexity can arise from the large volume of data, the variety of data sources, and the need for advanced algorithms and computing power.

These characteristics are interdependent and impact the results obtained from data analytics. Understanding and considering these characteristics is essential for effective data analytics.

In data analytics, there are four main types of data attributes: nominal, ordinal, interval, and ratio.

1. **Nominal:** Nominal attributes are categorical and do not have any inherent order or numerical meaning. Examples include gender, hair color, and eye color.
2. **Ordinal:** Ordinal attributes have an inherent order or ranking but no numerical meaning. Examples include education level (high school, college, graduate), and satisfaction ratings (unsatisfied, neutral, satisfied).
3. **Interval:** Interval attributes have an inherent order and numerical meaning but no clear definition of a zero point. Examples include temperature measured in Celsius or Fahrenheit.
4. **Ratio:** Ratio attributes have an inherent order, numerical meaning, and a clear definition of a zero point. Examples include height, weight, and income.

Big Data

Big Data refers to the large and complex data sets that cannot be processed or analyzed using traditional data processing and storage technologies. It is characterized by three main attributes:

1. **Volume:** refers to the sheer size of the data, which can range from terabytes to exabytes and is generated from various sources.

2. **Velocity:** refers to the speed at which data is generated, which can range from real-time to batch processing.
3. **Variety:** refers to the different types of data that are generated, including structured, semi-structured, and unstructured data.

- Requires advanced technologies and techniques for processing and analyzing the data, such as distributed computing, NoSQL databases, and machine learning.
- Enables organizations to process and analyze large amounts of data from various sources and gain valuable insights.
- Presents challenges, including data security and privacy, data governance, and the need for advanced skills and expertise.
- Despite the challenges, the potential benefits of Big Data make it an important area of focus for organizations seeking to gain a competitive advantage.

Big Data platforms

A Big Data platform is a system designed to manage, process, and analyze large and complex data sets that are too big to be handled by traditional data processing tools and technologies. These platforms are designed to handle large volumes of structured and unstructured data, including text, images, videos, and social media data, and to process this data in real-time.

Big Data platforms are typically composed of a number of different technologies and tools, including data storage systems, data processing engines, and data analytics tools. Some common components of a Big Data platform include:

- **Distributed storage systems**, such as Hadoop HDFS, to store large amounts of data in a distributed manner across multiple nodes
- **Data processing engines**, such as Apache Spark, to process and analyze large amounts of data in parallel
- **NoSQL databases**, such as MongoDB and Cassandra, to store and manage unstructured data
- **Data visualization tools**, such as Tableau and QlikView, to provide graphical representations of the data and insights

- **Machine learning algorithms** and libraries, such as TensorFlow and scikit-learn, to perform predictive analytics and build intelligent applications

The goal of a Big Data platform is to provide organizations with the ability to handle, process, and make sense of vast amounts of data in order to uncover insights and make better data-driven decisions.

OR

Big Data platforms refer to the software, hardware, and technologies that are used to store, process, and analyze large and complex data sets. Some of the most popular Big Data platforms include:

1. **Apache Hadoop:** An open-source Big Data platform that provides a framework for distributed storage and processing of large data sets.
2. **Apache Spark:** An open-source, high-performance Big Data processing framework that is designed for fast and efficient processing of large data sets.
3. **Apache Cassandra:** A highly scalable and highly available NoSQL database that is optimized for handling large amounts of structured data.
4. **Apache Storm:** An open-source real-time Big Data processing framework that is designed for processing large amounts of streaming data.
5. **Apache Flink:** An open-source Big Data processing framework that provides a fast, flexible, and efficient way to process large data sets.
6. **Apache Mahout:** An open-source machine learning library that is optimized for large-scale data processing and analysis.
7. **Apache Hive:** An open-source data warehousing and data analysis platform that provides a way to process and analyze large amounts of structured data.
8. **Amazon Web Services (AWS):** A cloud-based Big Data platform that provides a range of services for storing, processing, and analyzing large data sets.
9. **Google Cloud Big Data Platform:** A cloud-based Big Data platform that provides a range of services for storing, processing, and analyzing large data sets.
10. **Microsoft Azure:** A cloud-based Big Data platform that provides a range of services for storing, processing, and analyzing large data sets.

These are just a few examples of the many Big Data platforms that are available. Each platform has its own strengths, weaknesses, and use cases, so it is important to carefully consider the specific needs and requirements of your organization when choosing a Big Data platform.

Need of Data Analytics

Data analytics is used to identify patterns, trends, and insights in large and complex data sets, and then to use this information to make informed decisions. The need for data analytics arises from the following factors:

1. **Large and complex data sets:** With the growth of data-driven businesses and the increasing amount of data generated by organizations, the need for data analytics has grown.
2. **Business insights:** By analyzing large and complex data sets, organizations can gain valuable insights into customer behavior, market trends, and other important aspects of their business.
3. **Competitive advantage:** Organizations that are able to effectively analyze data are better able to compete in today's fast-paced business environment, as they can quickly identify new opportunities and respond to changing market conditions.
4. **Improved decision making:** Data analytics helps organizations make better decisions by providing them with a clearer picture of what is happening in their business. This includes identifying risks, detecting anomalies, and tracking performance metrics.
5. **Customer engagement:** Data analytics can be used to gain a deeper understanding of customer behavior, preferences, and opinions. This information can be used to improve customer engagement and customer satisfaction.
6. **Cost savings:** Data analytics can help organizations identify areas where they can improve efficiency, reduce waste, and lower costs.
7. **Compliance:** Data analytics can help organizations comply with regulations by tracking and analyzing data related to key performance metrics, such as safety and security.

In summary, the need for data analytics arises from the need to gain business insights, improve decision making, and stay competitive in today's data-driven business environment.

Evolution of Analytic Scalability

The evolution of analytic scalability refers to the development of technologies and methods that allow organizations to analyze increasing amounts of data more efficiently and effectively. This evolution has been driven by a number of factors, including:

1. **Big Data:** The growth of data-driven businesses has led to the development of new technologies and methods for managing, storing, and analyzing big data.
2. **Cloud computing:** Cloud computing has made it possible for organizations to store and analyze large amounts of data at a lower cost and with greater flexibility.
3. **Artificial intelligence:** Artificial intelligence (AI) and machine learning algorithms have made it possible to analyze data at scale and automate many of the tasks involved in data analysis.
4. **Parallel processing:** Parallel processing technology allows organizations to distribute data processing across multiple computers, making it possible to analyze larger data sets in a shorter amount of time.
5. **Real-time analytics:** Real-time analytics technologies have made it possible to process and analyze data in near real-time, which is essential for organizations that need to make data-driven decisions quickly.

These advancements have allowed organizations to analyze data at an increasingly large scale, which has led to more accurate and actionable insights. The continued evolution of analytic scalability will likely result in even more sophisticated and efficient methods for analyzing data in the future.

The evolution of analytic scalability can be divided into the following stages:

1. **Traditional Analytics:** Early analytics relied on traditional databases and centralized systems, limiting scalability.
2. **Distributed Analytics:** The rise of big data led to the development of distributed systems such as Hadoop and NoSQL databases to handle the increasing volume of data.
3. **Cloud Analytics:** The adoption of cloud computing has enabled the use of cloud-based analytics platforms for greater scalability and cost-effectiveness.

4. **Real-time Analytics:** The need for real-time data processing has driven the development of streaming analytics and technologies such as Apache Kafka and Spark Streaming.
5. **Multi-cloud Analytics:** Organizations are increasingly using multiple cloud services, leading to the development of multi-cloud analytics solutions for seamless data and analytics integration across multiple platforms.

Analytic Process and Tools

The analytics process typically consists of the following steps:

1. **Data Collection:** Gathering relevant data from various sources, such as databases, spreadsheets, and APIs.
2. **Data Preparation:** Cleaning, transforming and organizing the data to make it usable for analysis.
3. **Data Exploration:** Examining the data to understand its structure, patterns, and relationships.
4. **Modeling:** Building mathematical models to identify relationships and make predictions.
5. **Evaluation:** Testing and validating the models to determine their accuracy and effectiveness.
6. **Deployment:** Implementing the models and integrating them into business processes and decision-making.
7. **Monitoring:** Continuously monitoring the models to ensure their continued relevance and accuracy.

Some popular analytics tools used in each of these steps include:

1. **Data Collection:** SQL, Python, R.
2. **Data Preparation:** Excel, KNIME, RapidMiner.
3. **Data Exploration:** Tableau, Power BI, Google Analytics.
4. **Modeling:** SAS, R, Python, KNIME.
5. **Evaluation:** SAS, R, KNIME.
6. **Deployment:** Tableau, Power BI, Google Analytics.
7. **Monitoring:** Power BI, Google Analytics, Tableau.

Ultimately, the best analytics tool will depend on the specific requirements and goals of the organization and project.

Analysis vs Reporting

Data analysis and reporting are two essential stages in the data analytics process that work together to turn raw data into actionable insights. Here is a detailed comparison between the two:

Data Analysis:

1. **Goal:** The goal of data analysis is to gain a deeper understanding of the data and make informed decisions based on the insights generated. It involves evaluating data to identify patterns, relationships, and trends, as well as predicting future trends and outcomes.
2. **Techniques:** Data analysis techniques include statistical analysis, machine learning, data visualization, and predictive modeling. These techniques help to uncover meaningful insights and patterns in the data that can be used to inform decision-making.
3. **Output:** The output of data analysis is often a set of findings, insights, and predictions, which are then used to make decisions or inform further analysis.

Data Reporting:

1. **Goal:** The goal of data reporting is to communicate the insights and results of data analysis to stakeholders in a clear and concise format. Data reporting is used to provide a summary of the findings and insights generated through data analysis, and to present the information in a way that is easy to understand and use.
2. **Techniques:** Data reporting techniques include creating charts, tables, dashboards, and other visual representations of the data. These techniques are used to present the data in a clear and concise format, making it easier for stakeholders to understand and use the information to inform decision-making.
3. **Output:** The output of data reporting is a visual representation of the data and insights generated through data analysis. This information is used to communicate the results of the analysis to stakeholders and support data-driven decision-making.

In conclusion, data analysis and reporting work together to turn raw data into actionable insights. Data analysis provides the insights, while data reporting presents those insights in a clear and concise format for stakeholders to review and use.

Modern Data Analytic Tools

1. **Hadoop and Apache Spark:** Distributed systems for big data processing and storage.
2. **Tableau:** A data visualization tool that allows users to create interactive dashboards and reports.
3. **Power BI:** A business intelligence tool that allows users to create and share data visualizations and reports.
4. **Google Analytics:** A web analytics service that provides insights into website traffic and user behavior.
5. **Alteryx:** A data analysis and visualization tool that helps users prepare, blend, and analyze data from various sources.
6. **SAP Lumira:** A data visualization tool for exploring and visualizing large data sets.
7. **QlikView:** A data discovery and analytics platform that allows users to create and explore data visualizations.
8. **IBM Cognos Analytics:** A business intelligence and analytics platform that provides a full range of data analytics capabilities.
9. **TIBCO Spotfire:** A data visualization and discovery tool that helps users explore and analyze large data sets.

These are just a few examples of the many modern data analytic tools available today. The right tool for a particular organization will depend on its specific needs, data sources, and resources.

Applications of data analytics

Data analytics has a wide range of applications across various industries, including:

1. **Marketing:** Analyzing customer data to understand buying patterns and preferences, target advertising and promotions, and measure campaign effectiveness.
2. **Finance:** Analyzing financial data to detect fraud, manage risk, and make informed investment decisions.

3. **Healthcare:** Analyzing medical data to improve patient outcomes, reduce healthcare costs, and support clinical decision-making.
4. **Retail:** Analyzing sales data to optimize inventory management, improve supply chain efficiency, and personalize customer experiences.
5. **Manufacturing:** Analyzing production data to improve quality control, increase efficiency, and reduce waste.
6. **Telecommunications:** Analyzing network data to optimize network performance, detect network issues, and improve customer satisfaction.
7. **Energy:** Analyzing energy usage data to optimize energy usage, reduce costs, and reduce environmental impact.
8. **Sports:** Analyzing player performance data to optimize training, support tactical decision-making, and improve player performance.
9. **Transportation:** Analyzing transportation data to optimize routes, reduce fuel consumption, and improve safety.

These are just a few examples of the many applications of data analytics in various industries. The use of data analytics is constantly expanding as new data sources become available and new tools are developed to make sense of that data.

Key Roles for Successful Analytic Projects

For a successful data analytics project, the following key roles are crucial:

1. **Project Manager:** Responsible for managing the project timeline, resources, and budget.
2. **Data Engineer:** Responsible for data preparation, storage, and management.
3. **Data Analyst:** Conducts data exploration and analysis, creates visualizations, and generates insights.
4. **Business SME (Subject Matter Expert):** Provides domain expertise and helps to ensure that the insights generated are relevant to the business.
5. **Data Scientist:** Develops complex algorithms and models to analyze data and extract insights.

6. **Decision Maker:** Uses the insights generated from the data analytics project to make informed decisions.
7. **IT/Technical Support:** Ensures that the technology infrastructure and systems are in place to support the data analytics project.

It's important to note that the specific roles and responsibilities may vary depending on the size and complexity of the project, and that cross-functional collaboration and effective communication are crucial for success.

Data Analytics Lifecycle

The various phases of the data analytics lifecycle are as follows:

1. **Discovery:** In this phase, the purpose and goals of the data analytics project are established. This may involve defining the problem to be solved, determining the key metrics to be measured, and identifying the relevant data sources.
2. **Data Preparation:** This phase involves collecting and cleaning the data to make it suitable for analysis. This may include data integration, data transformation, and data profiling. It's crucial to ensure that the data is accurate, complete, and consistent before it is used for analysis.
3. **Model Planning:** In this phase, the type of analysis to be performed is determined and the appropriate data models are planned. This may involve selecting the appropriate statistical methods, algorithms, and tools to be used. It's also important to define the assumptions and limitations of the models in this phase.
4. **Model Building:** In this phase, the data models are constructed and the data is analyzed to generate insights. This may involve creating predictive models, performing clustering, or conducting hypothesis testing.
5. **Communicating Results:** In this phase, the insights and findings from the data analysis are communicated to stakeholders. This may involve creating visualizations, writing reports, or presenting the results to key decision-makers. The goal is to communicate the insights in a way that is easily understood by the intended audience.
6. **Operationalization:** This phase involves putting the insights and findings into action. This may involve updating processes and systems, automating data-driven processes, or monitoring the results to ensure that they align with expectations. It's important to

track and measure the impact of the data analytics project to determine if the goals have been met.

It's important to note that the data analytics lifecycle is an iterative process and may require revisiting previous phases to refine the analysis and improve the results. Effective communication and collaboration among team members is also crucial for success.