**T. Y. B. Tech (Computer Engineering)**
**Professional Elective – I: [CS3203T-A]:- Data Mining and Analytics**

**UNIT-I**

**Introduction Data Mining and Analytics** **06 Hours**

Definition, types, and importance of data analytics, Difference between data analytics, data science, and data mining, Data Types and Sources, Structured, semi-structured, and unstructured data What is Data Analytics? Types of Data Analytics: Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics, Data Analytics Lifecycle: Understanding the analytics lifecycle (data exploration, preparation, modelling, evaluation, and deployment). CRISP-DM (Cross-Industry Standard Process for Data Mining),OSEMN (Obtain, Scrub, Explore, Model, and Interpret),Case studies, Knowledge Representation Methods, Applications Related technologies - Machine Learning, DBMS, OLAP, and Statistics. Tools and Frameworks: Tableau, Excel, and Power BI.

**Definition, types, and importance of data analytics**

**What is Data Analytics?**

In this new digital world, data is being generated in an enormous amount which opens new paradigms. As we have high computing power and a large amount of data we can use this data to help us make data-driven decision making. The main benefits of data-driven decisions are that they are made up by observing past trends which have resulted in beneficial results.

In short, we can say that data analytics is the process of manipulating data to extract useful trends and hidden patterns that can help us derive valuable insights to make business predictions.

**Understanding Data Analytics**

Data analytics encompasses a wide array of techniques for analyzing data to gain valuable insights that can enhance various aspects of operations. By scrutinizing information, businesses can uncover patterns and metrics that might otherwise go unnoticed, enabling them to optimize processes and improve overall efficiency.

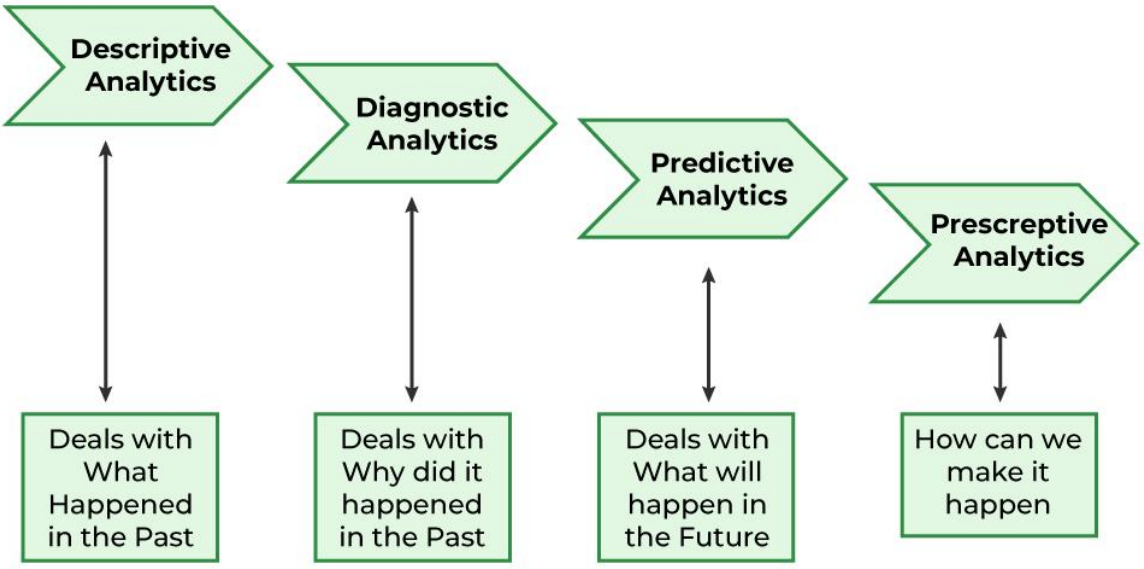
For instance, in manufacturing, companies collect data on machine runtime, downtime, and work queues to analyze and improve workload planning, ensuring machines operate at optimal levels.

Beyond production optimization, data analytics is utilized in diverse sectors. Gaming firms utilize it to design reward systems that engage players effectively, while content providers leverage analytics to optimize content placement and presentation, ultimately driving user engagement.

**Types of Data Analytics**

There are four major types of data analytics:

1. **Predictive (forecasting)**
2. **Descriptive (business intelligence and data mining)**
3. **Prescriptive (optimization and simulation)**
4. **Diagnostic analytics**




Data Analytics and its Types

**Predictive Analytics**

Predictive analytics turn the data into valuable, actionable information. predictive analytics uses data to determine the probable outcome of an event or a likelihood of a situation occurring. Predictive analytics holds a variety of statistical techniques from modeling, machine learning , data mining , and game theory that analyze current and historical facts to make predictions about a future event. Techniques that are used for predictive analytics are:

- Linear Regression
- Time Series Analysis and Forecasting
- Data Mining

**Basic Cornerstones of Predictive Analytics**

- Predictive modeling

- Decision Analysis and optimization
- Transaction profiling

**Descriptive Analytics**

Descriptive analytics looks at data and analyze past event for insight as to how to approach future events. It looks at past performance and understands the performance by mining historical data to understand the cause of success or failure in the past. Almost all management reporting such as sales, marketing, operations, and finance uses this type of analysis.

The descriptive model quantifies relationships in data in a way that is often used to classify customers or prospects into groups. Unlike a predictive model that focuses on predicting the behavior of a single customer, Descriptive analytics identifies many different relationships between customer and product.

**Common examples of Descriptive analytics are company reports that provide historic reviews like:**

- Data Queries
- Reports
- Descriptive Statistics
- Data dashboard

**Prescriptive Analytics**

Prescriptive Analytics automatically synthesize big data, mathematical science, business rule, and machine learning to make a prediction and then suggests a decision option to take advantage of the prediction.

Prescriptive analytics goes beyond predicting future outcomes by also suggesting action benefits from the predictions and showing the decision maker the implication of each decision option. Prescriptive Analytics not only anticipates what will happen and when to happen but also why it will happen. Further, Prescriptive Analytics can suggest decision options on how to take advantage of a future opportunity or mitigate a future risk and illustrate the implication of each decision option.

For example, Prescriptive Analytics can benefit healthcare strategic planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demography, etc.

**Diagnostic Analytics**

In this analysis, we generally use historical data over other data to answer any question or for the solution of any problem. We try to find any dependency and pattern in the historical data of the particular problem.

For example, companies go for this analysis because it gives a great insight into a problem, and they also keep detailed information about their disposal otherwise data collection may turn out individual for every problem and it will be very time-consuming. Common techniques used for Diagnostic Analytics are:

- Data discovery
- Data mining
- Correlations

**The Role of Data Analytics**

Data analytics plays a pivotal role in enhancing operations, efficiency, and performance across various industries by uncovering valuable patterns and insights. Implementing data analytics techniques can provide companies with a competitive advantage. The process typically involves four fundamental steps:

- **Data Mining** : This step involves gathering data and information from diverse sources and transforming them into a standardized format for subsequent analysis. Data mining can be a time-intensive process compared to other steps but is crucial for obtaining a comprehensive dataset.
- **Data Management** : Once collected, data needs to be stored, managed, and made accessible. Creating a database is essential for managing the vast amounts of information collected during the mining process. SQL (Structured Query Language) remains a widely used tool for database management, facilitating efficient querying and analysis of relational databases.
- **Statistical Analysis** : In this step, the gathered data is subjected to statistical analysis to identify trends and patterns. Statistical modeling is used to interpret the data and make predictions about future trends. Open-source programming languages like Python, as well as specialized tools like R, are commonly used for statistical analysis and graphical modeling.
- **Data Presentation** : The insights derived from data analytics need to be effectively communicated to stakeholders. This final step involves formatting the results in a manner that is accessible and understandable to various stakeholders, including decision-makers, analysts, and shareholders. Clear and concise data presentation is essential for driving informed decision-making and driving business growth.

**Steps in Data Analysis**

- **Define Data Requirements** : This involves determining how the data will be grouped or categorized. Data can be segmented based on various factors such as age, demographic, income, or gender, and can consist of numerical values or categorical data.
- **Data Collection** : Data is gathered from different sources, including computers, online platforms, cameras, environmental sensors, or through human personnel.
- **Data Organization** : Once collected, the data needs to be organized in a structured format to facilitate analysis. This could involve using spreadsheets or specialized software designed for managing and analyzing statistical data.
- **Data Cleaning** : Before analysis, the data undergoes a cleaning process to ensure accuracy and reliability. This involves identifying and removing any duplicate or erroneous entries, as well as addressing any missing or incomplete data. Cleaning the data helps to mitigate potential biases and errors that could affect the analysis results.

**Usage of Data Analytics**

There are some key domains and strategic planning techniques in which Data Analytics has played a vital role:

- **Improved Decision-Making** - If we have supporting data in favour of a decision, then we can implement them with even more success probability. For example, if a certain decision or plan has to lead to better outcomes then there will be no doubt in implementing them again.
- **Better Customer Service -** Churn modeling is the best example of this in which we try to predict or identify what leads to customer churn and change those things accordingly so, that the attrition of the customers is as low as possible which is a most important factor in any organization.
- **Efficient Operations -** Data Analytics can help us understand what is the demand of the situation and what should be done to get better results then we will be able to streamline our processes which in turn will lead to efficient operations.
- **Effective Marketing -** Market segmentation techniques have been implemented to target this important factor only in which we are supposed to find the marketing techniques which will help us increase our sales and leads to effective marketing strategies.

**Future Scope of Data Analytics**

- **Retail** : To study sales patterns, consumer behavior, and inventory management, data analytics can be applied in the retail sector. Data analytics can be used by retailers to make data-driven decisions regarding what products to stock, how to price them, and how to best organize their stores.
- **Healthcare** : Data analytics can be used to evaluate patient data, spot trends in patient health, and create individualized treatment regimens. Data analytics can be used by healthcare companies to enhance patient outcomes and lower healthcare expenditures.
- **Finance** : In the field of finance, data analytics can be used to evaluate investment data, spot trends in the financial markets, and make wise investment decisions. Data analytics can be used by financial institutions to lower risk and boost the performance of investment portfolios.
- **Marketing** : By analyzing customer data, spotting trends in consumer behavior, and creating customized marketing strategies, data analytics can be used in marketing. Data analytics can be used by marketers to boost the efficiency of their campaigns and their overall impact.
- **Manufacturing** : Data analytics can be used to examine production data, spot trends in production methods, and boost production efficiency in the manufacturing sector. Data analytics can be used by manufacturers to cut costs and enhance product quality.
- **Transportation** : To evaluate logistics data, spot trends in transportation routes, and improve transportation routes, the transportation sector can employ data analytics. Data analytics can help transportation businesses cut expenses and speed up delivery times.

**Conclusion**

Data Analytics act as tool that is used for both organizations and individuals that seems to use the power of data. As we progress in this data-driven age, data analytics will continue to play a pivotal role in shaping industries and influencing future

**Data Analysis Vs. Data Mining Vs. Data Science Vs. Machine Learning Vs. Big Data**

- admin
- April 29, 2023

Table of Contents



**Introduction**

Data science is an interdisciplinary field that involves using statistical, mathematical, and computational techniques to extract insights and knowledge from data. It is a broad field that encompasses many subfields, including data analytics, data analysis, data mining, machine learning, and big data.

**What is Data Analytics?**

**Data analytics** involves examining datasets to extract insights and knowledge from them. It is often used to inform business decisions or identify patterns in data. Data analytics involves both descriptive and diagnostic analysis, which means that it can be used to describe what has happened in the past and diagnose the reasons why it happened.

**What is Data Analysis?**

Data analysis is a more general term that refers to the process of examining data to extract insights and knowledge from it. It can involve various techniques, including statistical analysis, machine learning, and data visualization. Data analysis is often used in scientific research to test hypotheses and draw conclusions from data.

**What is Data Mining?**

Data mining is a specific technique used to extract insights and knowledge from large datasets. It involves using statistical and machine learning algorithms to identify patterns in data that can be used to make predictions or inform business decisions. Data mining is often used in fields like finance, healthcare, and marketing to identify trends and patterns in data.

**What is Data Science?**

Data science is a field that encompasses many different techniques and approaches to working with data. It involves using statistical, mathematical, and computational techniques to extract insights and knowledge from data. **Data science** can involve various subfields, including data analytics, data analysis, data mining, and machine learning.

**What is Machine Learning?**

Machine learning is a specific subfield of data science that involves building models that can learn from data and make predictions or decisions based on that data. It involves training algorithms on large datasets and using them to make predictions or classifications on new data. Machine learning is often used in fields like image and speech recognition, natural language processing, and recommendation systems.

**What is Big Data?**

Big data refers to datasets that are too large and complex to be processed using traditional data processing techniques. Big data involves the use of advanced computing technologies, such as distributed computing and cloud computing, to process and analyze data. Big data is often used in fields like finance, healthcare, and marketing to identify trends and patterns in data.

**Difference Between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, & Big Data**

Although these terms are often used interchangeably, they have distinct differences. Here are some of the key differences between them:

- Data analytics is the process of examining datasets to extract insights and knowledge from them, while data analysis is a more general term that refers to the process of examining data to extract insights and knowledge from it.
- Data mining is a specific technique used to extract insights and knowledge from large datasets using statistical and machine learning algorithms.
- Machine learning is a specific subfield of data science that involves building models that can learn from data and make predictions or decisions based on that data.
- Big data refers to datasets that are too large and complex to be processed using traditional data processing techniques and often involves the use of advanced computing technologies like distributed computing and cloud computing.

**Conclusion**

While data analytics, data analysis, data mining, data science, machine learning, and big data are all related to the management and processing of data, they are different concepts with distinct goals

and objectives. Understanding the differences between these terms is critical to effectively leveraging data and deriving valuable insights.

To summarize, data analytics focuses on extracting insights from data sets, while data analysis involves examining and interpreting data to draw conclusions. Data mining is the process of extracting patterns and insights from data sets, while data science involves the use of scientific methods to extract insights from data. Machine learning is a subset of data science that focuses on building algorithms that can learn from data and make predictions, while big data refers to large, complex data sets that require specialized tools and techniques for processing.

By understanding the differences between these concepts, individuals and organizations can make better decisions about how to leverage data and gain insights into their business and customers. As the importance of data continues to grow, a solid understanding of these concepts will be increasingly critical to success in the digital age.

**Data Science vs. Data Analytics – What's the Key Difference?**

**Introduction**

Today, data is playing a prominent role in the growth of any industry. Many organizations are using this data to gain insights and make effective business decisions to stay ahead in the market competition. This increased demand for data drives the need for skilled professionals who specialize in analyzing and interpreting that data. It's easy to get confused between data science and data analytics. Both are used to analyze information, but there are some key differences between them. In this article, we take a look at the key difference between data science vs data analytics.

**Data Science and Data Analytics are Different**

data science vs data analytics both incredibly important fields in the modern world. Both deal with the use of data to help make decisions and solve problems.

Data analytics and data science are two areas that are often confused with each other. The biggest difference between these two fields is their goals.

Data analytics focuses more on analyzing an existing dataset, whereas data science focuses on creating new models to generate the best outcomes possible.

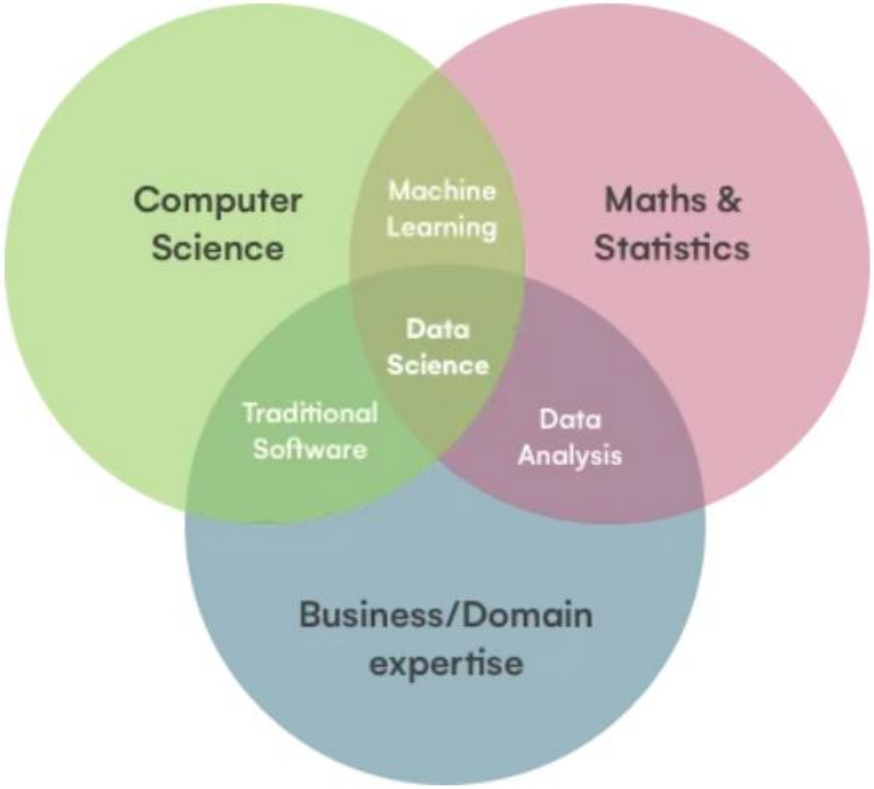
**What is Data Science?**

Data Science is the study of data-driven decision-making. Data Scientists use data to make predictions and to find patterns in data. They create algorithms and models that help them make predictions.

Data science is a broad term that includes many disciplines, including statistics, machine learning, and computer science. These fields are used to create new algorithms and models for processing large amounts of data.

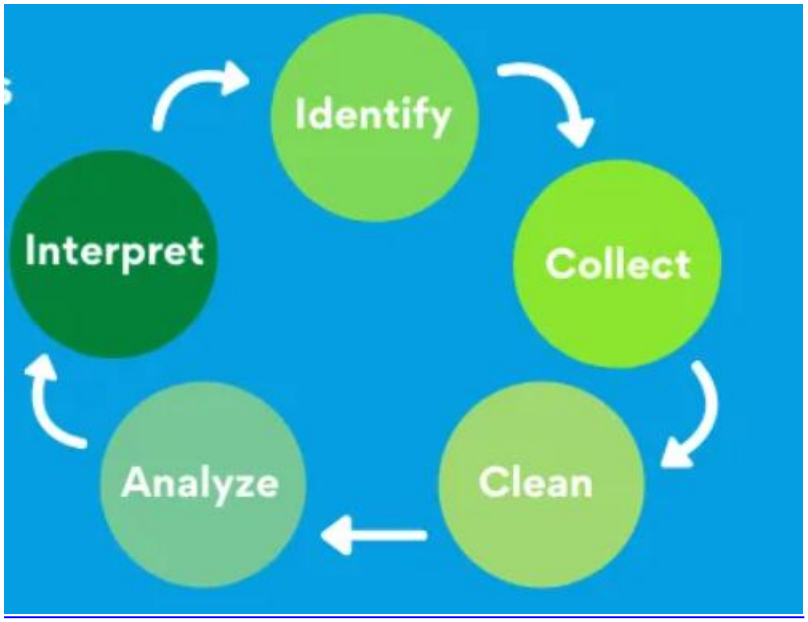
**What is Data Analytics?**

Data analytics is the process of collecting, analyzing, and interpreting data to make better business




decisions.

It's a unique way to get an objective picture of your business, and understand what's working, what isn't and what needs to change. **Data analytics** help find out if your marketing strategy is working or not.




When you're running marketing campaigns on social media or sending newsletters out to your customers, you want to know that you're reaching the right people with the right message. And you want to know that your efforts are paying off.

**1. Data Preparation**

Data preparation involves cleaning up your data so that it is ready to be used in your analysis.

**2. Modeling & Forecasting**

Modeling is the process of transforming your raw data into a format that can be used to generate predictions.

**3. Visualization & Reporting**

Once you've completed the modeling process, it's time to **visualize your results** using tools. It allows for the easy creation of reports with professional-looking graphs and charts. They can be shared with colleagues or clients.

Although data science and data analytics are similar in many ways, they differ in their focus. Data science is more focused on the scientific side of things, while data analytics is more focused on the business side of things.

For example, it's common for someone who works as a "data scientist" to also work as a "data analyst" or vice versa; however, it's not typical for someone who works as an "analyst" to also work as an "expert."

**Examples of Data Science & Data Analytics**

**1. Purchase Decisions**

Let's say you have a dataset of customer purchases over time and want to know what factors affect their purchasing decisions. You can use regression analysis or machine learning to decide which variables most strongly correlate with more purchases. This would be an example of data analytics.

**2. Predicting Customer Behavior**

Whereas, if you want to figure out how to predict how much customers will spend at a particular store next week. Then it would be an example of data science. Because it involves applying algorithms and **machine learning** techniques to generate predictions.

**Skills of Data Scientists**

Data Science is a very broad and diverse field. It's difficult to define the field of Data Science because it consists of interdisciplinary areas that include statistics, computer science, data engineering, and machine learning.

- The best way to understand what a Data Scientist does is to compare it with other fields like programming, information technology (IT), analytics, and business intelligence (**BI**).
- In comparison to IT specialists who use traditional programming languages like Java or C++ for **developing apps** for businesses or corporations, Data Scientists know both IT as well as statistics.
- Data scientists use different statistical techniques on large datasets using tools such as the Hadoop framework for processing big data in parallel clusters.

**Skills of Data Analysts**

- Data analysts use data to solve business problems. They use data to make decisions, present information and make predictions.
- They are also responsible for optimizing business processes by using data. Data analysts are responsible for analyzing data and providing insights to the business.
- They are often called upon to work with other teams to provide these insights, so they must be able to communicate effectively and build relationships.
- Data analysts must also be able to think critically about their work, using logic and reasoning skills to arrive at decision-making.

**Here's a Table that Compares Data Science vs. Data Analytics:**

| Feature | Data Science | Data Analytics |
|---|---|---|
| Coding Language | Python is the most commonly used language for data science along with the use of other languages such as C++, Java, Perl, etc. | The Knowledge of Python and R Language is essential for Data Analytics. |
| Programming Skills | In- depth knowledge of programming is required for data science. | Basic Programming skills is necessary for data analytics. |
| Use of Machine Learning | Data Science makes use of machine learning algorithms to get insights. | Data Analytics doesn't makes use of machine learning. |
| Other Skills | Data Science makes use of Data mining activities for getting meaningful insights. | Hadoop Based analysis is used for getting conclusions from raw data. |
| Scope | Care for your body parts | The Scope of data analysis is micro i.e., small. |
| Goals | Data science deals with explorations and new innovations. | Data Analysis makes use of existing resources. |
| Data Type | Data Science mostly deals with unstructured data. | Data Analytics deals with structured data. |
| Statistical Skills | The statistical skills are necessary in the field of Data Science. | The statistical skills are of minimal or no use in data analytics. |

**Roles & Responsibilities of Data Scientists & Data Analysts**

- Data science is a broad field that covers a wide range of topics. It's the study of data and how it can be used to reach certain goals, like improving business processes or creating better products or services.
- Data analysts are more focused on the analysis of data, but they're not necessarily involved in creating the information from which they analyze.
- Data scientists are more likely to be involved in the creation of data as well as its analysis, but not all data scientists do both tasks equally well.
- There are many different types of professionals who fall under this umbrella term—the difference between them depends on how much emphasis they place on each part of their job. (creation vs analysis)
- It also depends on what **tools they use** for each task and how far along in their career path they've progressed (junior vs senior).

**Conclusion**

As we've seen, start a travel agency business online two very different disciplines. The skills required for each job are of course different. Hence, it's important to understand what these roles & responsibilities entail before hiring them.

Both data science and data analytics are essential entities for the growth of all types of organizations. **Data analytics** is more about the what, while data science is more about the why. Data analytics is more about the present. While data science is more about the future. And finally, data analytics focuses on numbers and facts whereas data science focuses on people and emotio

**Structured Data vs Unstructured Data vs Semi-Structured Data**

**Key Differences, Use Cases & Business Value**

In today's data-driven economy, businesses are inundated with information from thousands of sources, including CRMs, web apps, IoT devices, social media, internal systems, and third-party APIs. To make sense of this information and drive smarter decisions, organizations must understand how data is categorized and handled.

Broadly, data falls into three categories:

- Structured data
- Unstructured data
- Semi-structured data

Each type requires different storage solutions, processing techniques, and analytics tools. Let's break down the characteristics, advantages, and trade-offs, and how to make sure your data stack is ready for all three.

Understanding the difference between structured and unstructured data is critical for building high-performing data architectures. Structured data is clean, organized, and easily queried, perfect for relational databases. Unstructured data is messy but insight-rich, powering AI and ML models. Sitting

between them is semi-structured data, which offers flexibility with some organization. The right approach? Know your use case, and choose the right tools to manage, transform, and analyze your data at scale.

**What is Structured Data?**

Structured data, as it sounds, is the most organized form of data, designed for easy storage, access, and analysis. This data type is typically formatted into predefined rows and columns, making it highly searchable and easily organized within databases or spreadsheets.

Each element in structured data is addressable, which means it can be precisely defined and easily grouped or related to other elements. Typically, structured data is housed in relational database management systems, which allows for complex querying and analysis using SQL.

**Examples of Structured Data include**

- Entity relationship diagrams (e.g., tables, rows, columns, primary keys, foreign keys)
- Financial transactions (e.g., sales data, purchase orders, accounting entries)
- Customer demographic information (e.g., name, address, age, gender)
- Machine logs, like events captured by devices, formatted with time stamps and specific parameters
- Smartphone location data, such as GPS coordinates captured at fixed intervals
- Spreadsheets that are commonly used for various business operations, from inventory to employee tracking
- Structured data is used daily, for instance, in customer order forms used by an e-commerce website. When customers place an order, they fill out a form with specific fields such as name, shipping address, quantity, and price.

Each of these fields is predefined and follows a consistent format. This structured approach ensures that every order is recorded uniformly, making it straightforward to track and process orders, manage inventory, and analyze sales trends efficiently.

**Why use Structured Data?**

The primary reason for using structured data is its simplicity and efficiency in processing. It allows businesses to store vast amounts of information in an organized manner that can be quickly accessed and analyzed.

Plus, relational databases can handle large datasets and allow for complex queries, enabling powerful business intelligence applications.

**Advantages of Structured Data**

- Easy analysis: Because it is highly organized, structured data can be easily analyzed using standard tools like SQL queries.
- Accuracy and consistency: Fixed data fields reduce the chance of errors and provide uniformity.
- Performance: Structured data is optimized for relational databases, making searches and computations fast and efficient.

**Disadvantages of Structured Data**

- Limited flexibility: Structured data requires a predefined schema, which means it can only store information that fits into a rigid format of rows and columns. This makes it difficult to accommodate dynamic or complex data.
- Not suitable for all data types: Real-world data is often complex and doesn't always fit neatly into structured fields. Structured data is not ideal for capturing qualitative information like images, videos, or long text, which can limit its applicability in certain areas.
- Requires upfront planning: Because unstructured data requires a well-defined schema, it often necessitates upfront design and planning, which can slow down agile projects or processes that involve rapid changes.

Structured data is perfect for repeatable processes. It powers dashboards, BI reports, and operational systems. But it can't tell the full story on its own.

**What is Unstructured Data?**

Unstructured data represents the largest category of data, and it's growing exponentially as more digital content is created. Unlike structured data, unstructured data doesn't follow a specific format or schema, which makes it more challenging to store and analyze.

This type of data comes from a wide variety of sources, including emails, social media content, and multimedia files. Due to its lack of structure, unstructured data cannot be stored in traditional row-column databases and often requires more advanced storage solutions like data lakes.

**Example of Unstructured Data**

- Emails—while certain fields, like the sender and timestamp, are structured, the email body itself is unstructured text.
- Photos and videos, because multimedia files are usually stored as raw data and lack predefined fields.
- Audio files (e.g., recordings of customer service calls, podcasts, and music files)
- Text documents (e.g., PDFs, Word documents, and open-ended survey responses)
- Social media content, such as posts, tweets, comments, and other user-generated content, all of which are unstructured and vary widely in format.
- Call center transcripts or recordings, while voice interactions can be analyzed for sentiment or trends, are naturally unstructured.

**Why use Unstructured Data?**

Unstructured data is rich in information but difficult to process with traditional systems. However, advancements in artificial intelligence (AI) and machine learning (ML) have made extracting valuable insights from these vast and complex data sets easier.

**Advantages of Unstructured Data**

- Rich in insights: Unstructured data, especially from sources like social media or customer feedback, often contains nuanced and valuable information.
- Flexibility: Unstructured data can capture complex, real-world scenarios that structured data cannot.
- Sentiment analysis and brand identification: AI algorithms can analyze unstructured data for patterns, trends, and sentiments that structured data may not reveal.
- Versatility: With tools like AI and ML, unstructured data can now be harnessed for applications such as predictive maintenance (from machine logs) and fraud detection.

**Disadvantages of Unstructured Data**

- Difficult to store and manage: Traditional databases cannot handle unstructured data, meaning organizations must invest in alternative storage solutions like data lakes, which require specialized management.
- Challenging to analyze: Extracting useful insights from unstructured data is more difficult and often requires sophisticated tools like AI and ML, which may not be readily available to all organizations.
- Resource-intensive: Processing and analyzing unstructured data can require more computational power, specialized software, and skilled personnel. This makes extracting value from it more costly and time-consuming than extracting value from structured data.
- Quality and consistency issues: Unstructured data is often inconsistent in format and quality, making it harder to standardize and ensure accuracy during analysis. The lack of uniformity in unstructured data can lead to unreliable insights if not processed carefully.

Unstructured data is messy but powerful. It’s where your business hears your customers, predicts failure before it happens, and spots the signals others miss

**Structured vs Unstructured Data: What's the Difference?**

Structured and unstructured data represent two fundamentally different types of information, and understanding the difference is key to selecting the right tools for storage, processing, and analysis.

Structured data is highly organized and easily searchable, typically stored in rows and columns within relational databases. It includes clearly defined fields like names, dates, customer IDs, and transaction amounts, the kind of data that traditional analytics tools thrive on. In contrast, unstructured data lacks a predefined format and doesn’t fit neatly into tables. Think images, videos, emails, PDFs, and social media posts, rich in context but more complex to store, process, and analyze.

As organizations collect more diverse data types from more sources, understanding how to work with both structured and unstructured data becomes critical for generating accurate insights and enabling AI-powered analytics.

**Structured vs. Unstructured Data Examples**

| Structured Data | Unstructured Data |
|---|---|
| Customer database with contact details | Email conversations |
| Product inventory with SKU codes | Product photos or videos |
| Web analytics stored in tables | Social media posts or comments |
| Sales reports in Excel | Customer feedback in Word documents |

While structured data is easier to manage and analyze using traditional BI tools, unstructured data is growing exponentially, and modern data platforms must be able to handle both to deliver complete business insights.

**What is Semi-Structured Data?**

Semi-structured data blends elements of both structured and unstructured data. While it doesn't fit neatly into relational databases, it contains some organizational markers, such as metadata or tags, that make it easier to manage and analyze than fully unstructured data.

Semi-structured data typically doesn't follow rigid schemas but has some structure that tools can leverage to provide useful insights.

In that sense, semi-structured data strikes a balance between structured data's rigidity and unstructured data's flexibility, making it a valuable format for modern businesses looking to work with complex data sources.

**Examples of Semi-Structured Data**

- Web technologies (e.g., HTML)
- NoSQL databases (e.g., MongoDB, CouchDB, CockroachDB)
- DevOps (e.g., log files)
- JSON, XML, and YAML, which are common formats for semi-structured data that contain tags and elements but are not rigidly organized like relational databases.

**Why use Semi-Structured Data?**

Semi-structured data offers a flexible yet management format for businesses that need to work with datasets that have some variability. This data type can be particularly useful for scenarios where partial organization is needed, but the flexibility for unstructured data is also valuable.

**Advantages of Semi-Structured Data**

- Flexibility with organization: You can store large volumes of data with some structure, making it easier to analyze than fully unstructured data.
- Ideal for web and IoT data: Many modern data formats, such as those used in web apps or IoT devices, are semi-structured, making them more versatile.
- Supports scalability: Scaling semi-structured data storage solutions can be easier than traditional relational databases.

**Disadvantages of Semi-Structured Data**

- Less efficient than structured data: While semi-structured data offers more flexibility, it is still less efficient to store and query than fully structured data, which is optimized for fast, complex queries.
- Requires specialized tools: Semi-structured data can't be easily handled by traditional relational databases, requiring organizations to adopt more specialized tools like NoSQL databases or specific analytics platforms.
- Consistency is harder to ensure: Because semi-structured data doesn't adhere to strict schemas, it can be challenging to maintain consistency across datasets, especially as they grow in size and complexity.
- Limited standardization: Unlike structured data, semi-structured data doesn't have industry-wide standardization, which can lead to compatibility issues when integrating with other systems or platforms.

**Comparison: Structured Data vs Unstructured Data vs Semi-Structured Data**

| **Feature** | **Structured Data** | **Semi-Structured Data** | **Unstructured Data** |
|---|---|---|---|
| **Schema** | Fixed (e.g., SQL) | Flexible (e.g., JSON) | None |
| **Storage** | Relational DBs | NoSQL, object stores | Data lakes, file systems |
| **Querying** | SQL | XQuery, custom scripts | NLP, AI/ML models |
| **Use Cases** | BI, ERP, CRM | APIs, IoT, logs | Media, customer feedback |
| **Scalability** | Medium | High | High |
| **AI/ML Readiness** | Low | Moderate | High |
| **Examples** | Spreadsheets, transactions | JSON logs, HTML files | Emails, videos, audio |

**Why It Matters for Data Integration and Analytics**

Modern data pipelines need to support all three data types to deliver value across the business. From automated ETL to data quality and transformation, each data type requires a tailored approach.

- Structured data powers dashboards and operational reporting.
- Unstructured data feeds AI/ML and customer insights.
- Semi-structured data supports real-time, cloud-native applications.

Without the right integration and orchestration platform, these data types stay siloed, and valuable business insights are lost.

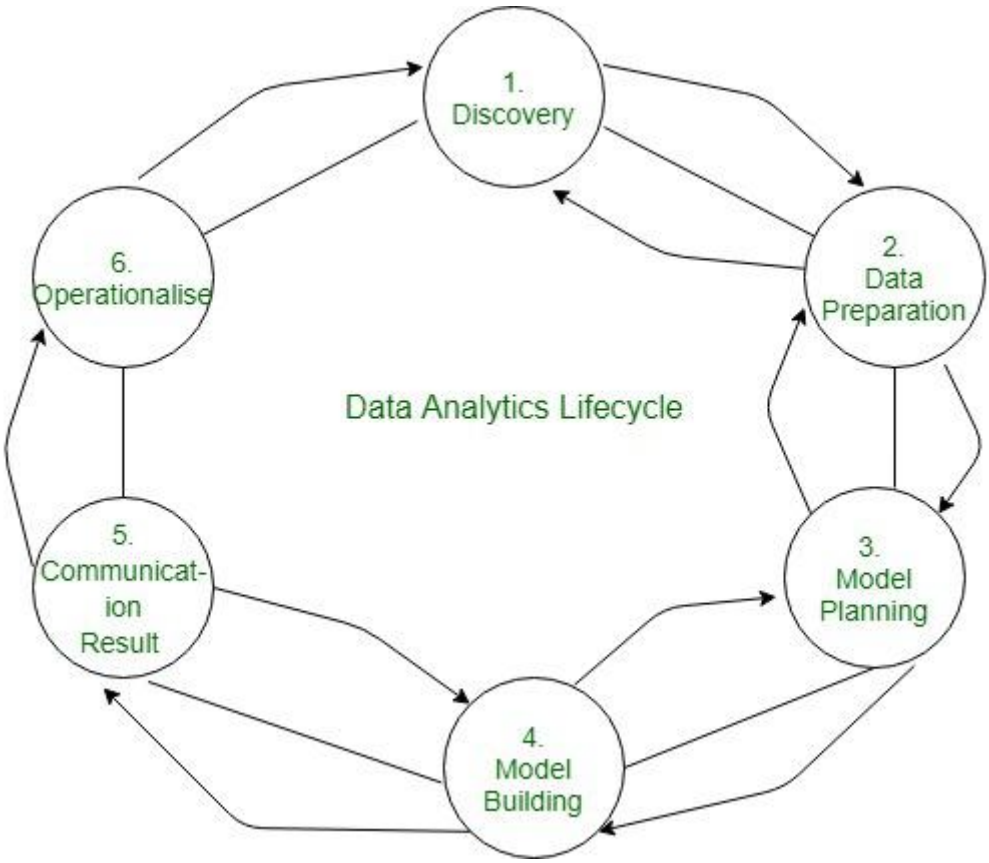
**Life Cycle Phases of Data Analytics**

**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 learns and investigates the problem.
- Develop context and understanding.
- Come to know about data sources needed and available for the project.
- The team formulates the initial hypothesis that can be later tested with data.
- **Phase 2: Data Preparation -**
- Steps to explore, preprocess, and condition data before modeling and analysis.
- It requires the presence of an analytic sandbox, the team executes, loads, and transforms, 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 -**
- The team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
- In this phase, the data science team develops 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 and 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 and 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 which make adjustments before full deployment.
- The team delivers final reports, briefings, codes.
- Free or open source tools - Octave, WEKA, SQL, MADlib.




**Understanding the data analytics lifecycle from end-to-end**

When your competitors are not analyzing their data, being able to transform raw information into actionable insights gives you a crucial competitive advantage. Organizations across industries increasingly rely on data analytics to inform decision-making, optimize operations, identify opportunities, and solve complex problems. However, effective analytics is far more than running numbers through algorithms. The data analytics lifecycle is a structured process with distinct phases that build upon each other to deliver meaningful results.

From an IBM report, "Business analytics (BA) is then a subset of BI, with business analytics providing the prescriptive, *forward-looking* analysis." The core outputs are actionable insights, which are statements of new knowledge that, when implemented, produce business value. The linked report discusses how to derive insights from data by identifying the six attributes of actionable insights: alignment, context, relevance, specificity, novelty, and clarity.

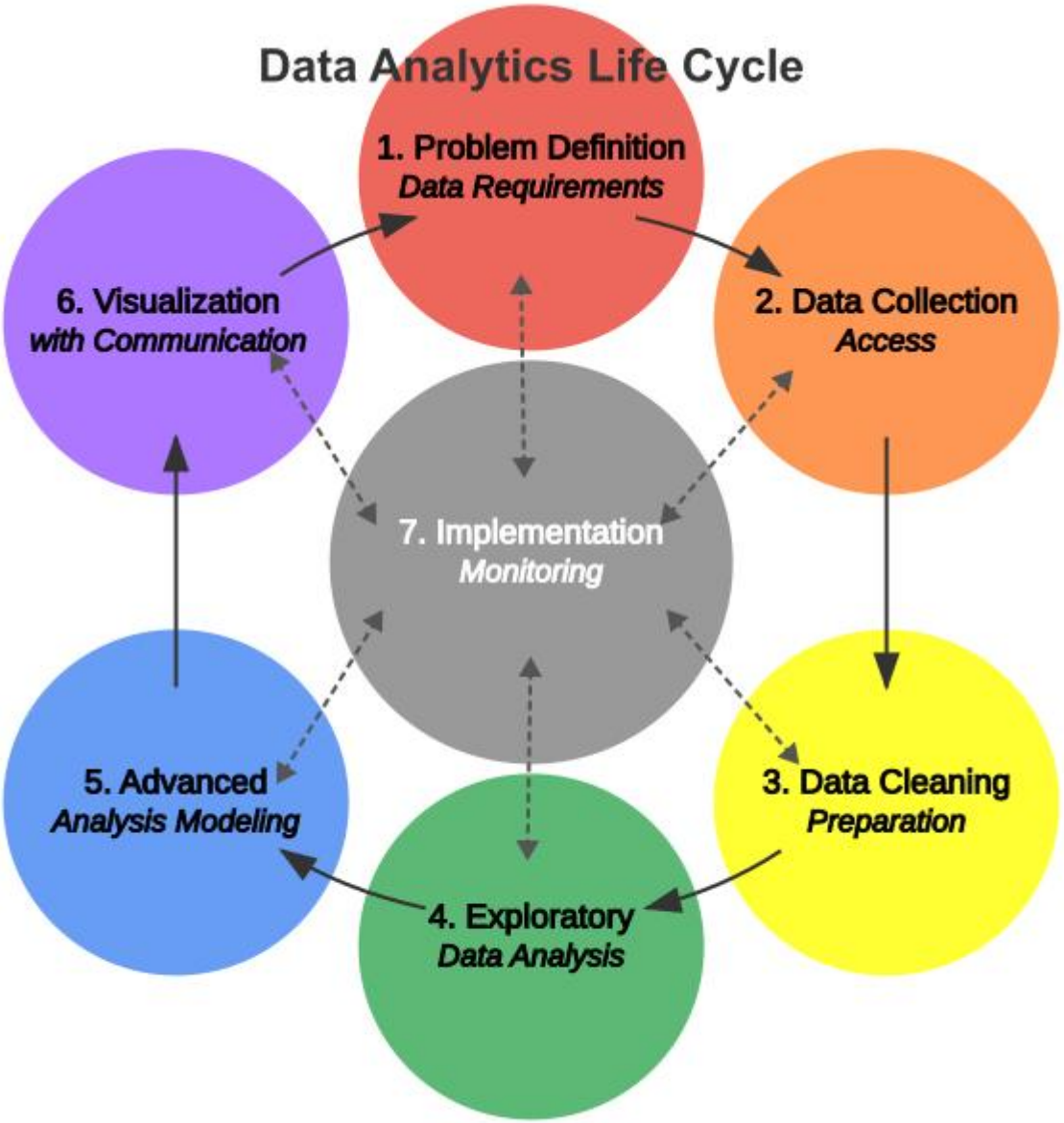
The data analytics lifecycle is a framework of seven phases going from initial data collection to the final presentation of insights. This post describes each phase and its challenges. It also describes the best practices to maximize value in your end-to-end analytics process. AI is rapidly changing the available tools, and Gartner predicts that by 2027, "AI assistants and AI-enhanced workflows incorporated into data integration tools will reduce manual intervention by 60% and enable self-service data management."

Understanding the data analytics life cycle

The data analysis life cycle represents the end-to-end process of working with data to extract actionable insights. Different organizations may use slightly different terminology or combine certain

phases in an end-to-end data project. The core components of the life cycle of data analysis typically include:

1. Problem Definition and Data Requirements
2. Data Collection and Access
3. Data Cleaning and Preparation
4. Exploratory Data Analysis
5. Advanced Analysis and Modeling
6. Visualization with Communication
7. Implementation and Monitoring




*Legend: The hub-and-spoke model of data analytics where Implementation Monitoring connects with all phases of the analytical process*

This data analytics life cycle diagram illustrates how each phase in the data management lifecycle builds upon the previous one, In practice, however, the process is often iterative rather than strictly linear. The data analysis life cycle diagram below shows the cyclical nature of this process because insights often lead to new questions. The life cycle needs to be carefully analyzed when introducing any type of AI into any of the seven phases.

In this visualization, the outer circles represent the six primary phases of the data analytics lifecycle arranged in a clockwise sequence. Although the life cycle is drawn as an iterative linear process, any phase can feed back into a previous phase and cause change. The double-headed dashed arrows connect each outer phase to the central "Implementation & Monitoring" hub (Phase 7), illustrating that (a) each phase can directly inform or be affected by implementation decisions, and (b) implementation results can trigger work or rework in any of the other phases.

This hub-and-spoke model more accurately represents how data analytics works in practice.

- Insights from any phase might necessitate immediate implementation.
- Implementation and monitoring often reveal needs that require jumping directly to specific phases.
- The process is highly iterative rather than strictly sequential.

The central positioning of Implementation and Monitoring emphasizes its role as both a destination for insights and an origin point for new questions and analytical needs. This creates a dynamic ecosystem where data-driven decisions and their outcomes continuously inform each other.

While this specific presentation is original, the general framework of the data analytics lifecycle is established industry knowledge. This visualization better captures the reality of modern data analytics workflows, where teams frequently move between phases based on findings, business needs, and implementation results, and where AI may be introduced into any or all of the life-cycle phases.

## The seven phases of the life cycle

Let's explore each of the data analysis life cycle phases in detail.

### Phase 1: Problem definition and data requirements

Every effective analytics project begins with a clear definition of the business problem or opportunity. This critical first step in the analysis lifecycle ensures that subsequent analysis will deliver relevant, valuable insights rather than merely interesting but ultimately unusable information.

Strong problem definitions are:

- Specific and focused on a particular challenge or opportunity
- Aligned with broader organizational goals and strategies
- Measurable, with clear criteria for success
- Actionable, leading to potential decisions or intervention

For example, rather than a vague goal like "understand customer behavior," a well-defined problem might be "identify factors contributing to customer churn in our premium subscription tier to develop targeted retention strategies."

Once the business question is clear, the next step in this phase involves identifying what data is needed. This includes:

- Types of data required (demographic, transactional, behavioral, etc.)
- Level of granularity needed (individual, aggregate, etc.)
- Time period of interest (historical depth, frequency of updates)
- Internal and external data sources to be consulted

This phase also involves a preliminary assessment of data availability and accessibility. Are the necessary datasets already available within the organization? Will external data need to be acquired? Are there legal or compliance considerations that could limit data usage?

By investing time upfront in a clear problem definition and requirements, organizations can avoid the common pitfall of collecting massive amounts of data without a clear purpose. Having no clear purpose often leads to "analysis paralysis" or insights that do not address key business needs.

Phase 2: Data collection and access

With requirements defined, the next phase focuses on identifying and accessing data sources. The access mechanism(s) must be established for each source. Common data sources include:

- Internal operational systems (CRM, ERP, marketing automation)
- Data warehouses and data lakes
- External datasets (market research, public datasets, third-party providers)
- Web and social media platforms
- IoT devices and sensors
- Survey and research data

Accessing this data often requires collaboration with IT teams, data engineers, or external providers. Technical considerations include:

- Database connections and query methods
- API access and integration
- File formats and transfer mechanisms
- Authentication and authorization protocols
- Data governance and compliance requirements

Notice that Storage is not listed in the seven phases or on the diagram because data may be transformed either before or after it is loaded into the storage destination. This is often discussed as choosing between an ETL (Extract, Transform, Load) process and an ELT (Extract, Load, Transform) process. For either, loading data into its storage destination may involve:

- Batch processing of historical data
- Setting up streaming pipelines for real-time data
- Creating appropriate storage structures (databases, data lakes, etc.)
- Establishing data catalogs to track available datasets
- Implementing metadata management to document data characteristics

The collection phase often reveals gaps between ideal data requirements and what is actually available. This may necessitate adjusting expectations, finding proxy measures, or initiating new data collection efforts to address critical gaps.

Phase 3: Data cleaning and preparation

Data rarely arrives in an analysis-ready state, and the importance of data cleaning is often undervalued. Raw data typically contains errors, inconsistencies, and structural issues that must be addressed before meaningful analysis can begin. This data cleaning phase is often the most time-consuming part of the data analytics life cycle, with practitioners reporting that it can consume large blocks of their total project time.

Common data quality issues include:

- Missing values: Gaps in the data that must be addressed through deletion, imputation, or flagging
- Duplicate records: Redundant entries that can skew analyses and waste computational resources
- Inconsistent formatting: Variations in how dates, currencies, or categorical values are represented
- Outliers and errors: Values that fall outside expected ranges or contain obvious mistakes
- Structural problems: Issues with how data is organized that complicate analysis

Despite being labor-intensive, thorough data cleaning is essential for reliable results. As the saying goes: "garbage in, garbage out." Even the most sophisticated analytical techniques cannot compensate for poor-quality input data.

Beyond cleaning, raw data typically requires transformation to create analysis-ready datasets. This may involve:

- Standardizing formats and units across different data sources
- Normalizing or scaling numerical variables

- Encoding categorical variables for mathematical analysis
- Creating derived variables that better capture phenomena of interest
- Aggregating or summarizing data tables to appropriate levels
- Restructuring data between wide and long formats

Many analytical projects require combining multiple data sources to create a comprehensive view. This integration process may involve:

- Identifying common keys or matching criteria across datasets
- Resolving entity resolution challenges (e.g., determining when records from different systems represent the same customer)
- Handling conflicting information from different sources
- Establishing temporal alignment between datasets collected at different times

External data enrichment may also add valuable context to internal datasets. For example, augmenting customer data with demographic information, or adding geographic data to retail locations.

Phase 4: Exploratory data analysis (EDA)

Once data is cleaned and prepared, exploratory data analysis (EDA) provides the first opportunity to understand what the data reveals. This critical phase helps analysts:

- Understand the distribution and characteristics of key variables
- Identify relationships and correlations between variables
- Discover patterns, trends, and anomalies
- Generate initial hypotheses for deeper investigation
- Validate assumptions about the data

EDA combines visual and statistical techniques to develop a comprehensive understanding of the dataset. Simple summary statistics such as means, medians, and standard deviations provide a starting point, while data visualizations reveal patterns that numbers alone might miss. Effective exploratory analysis typically employs multiple techniques:

- Univariate analysis examines individual variables through histograms, box plots, and summary statistics.
- Bivariate analysis explores relationships between pairs of variables through scatter plots, correlation coefficients, and contingency tables.
- Multivariate analysis investigates interactions among multiple variables simultaneously.
- Temporal analysis identifies trends, seasonality, and patterns over time.

- Geographic analysis reveals spatial patterns and relationships.

Throughout the EDA phase, the goal is not just to understand what the data contains, but to develop insights relevant to the original business question. Strong EDA maintains the connection between technical exploration and business context.

Phase 5: Advanced analysis and modeling

The advanced analysis and modeling phase includes selecting the appropriate analytical techniques, developing and validating (or choosing) the model, and then interpreting and evaluating the model's results in the context of the original question to be answered or the problem to be solved.

Based on the business question and insights from exploratory analysis, analysts select appropriate advanced analytical techniques. These broadly fall into several categories:

- **Descriptive analytics** summarizes what has happened through aggregations, segmentation, and summarization.
- **Diagnostic analytics** examines why something happened through correlation analysis, factor analysis, and root cause investigation.
- **Predictive analytics** forecasts what might happen through regression, classification, time series, and machine learning approaches.
- **Prescriptive analytics** recommends actions through optimization, simulation, and decision analysis.

The choice of specific techniques depends on:

- The nature of the business question
- Characteristics of the available data
- Required level of statistical confidence
- Available analytical tools and expertise
- Interpretability requirements for stakeholders

For predictive and prescriptive approaches, model development follows a structured process:

- Feature selection: Identifying the most relevant variables for inclusion
- Algorithm selection: Choosing appropriate modeling techniques
- Model training: Using a portion of data to develop initial models
- Hyperparameter tuning: Optimizing model parameters for performance
- Validation: Testing models on held-out data to assess generalizability
- Ensemble methods: Combining multiple models for improved performance

Model validation is particularly important to ensure that results will generalize beyond the specific dataset used for training. Cross-validation techniques, testing on independent datasets, and monitoring for model drift over time help ensure reliable results.

When a model is developed, its results must be interpreted in the context of the original business question. This involves:

- Assessing statistical significance and confidence levels
- Evaluating practical significance and business impact
- Understanding model limitations and constraints
- Identifying potential biases or ethical considerations
- Considering alternative explanations for observed patterns

The goal is to produce accurate models that extract meaningful insights that inform decision-making. Technical performance metrics (accuracy, precision, recall) matter, but business relevance remains the ultimate criterion for success.

Phase 6: Visualization with communication

Effective communication of analytical results requires translating complex findings into formats that stakeholders can easily understand and act upon, typically referred to as last-mile analytics. Data visualization plays a crucial role in this translation, and the choice of type of visualization depends heavily on the nature of the data and the key message.

Beyond individual visualizations, effective communication often requires building a coherent data story that:

- Establishes relevant context for the analysis
- Guides audiences logically through key findings
- Connects analytical results to business implications
- Addresses potential questions or objections
- Leads naturally to recommended actions

Interactive dashboards increasingly complement static reports, allowing stakeholders to explore data dynamically and focus on aspects most relevant to their specific needs with self-serve analytics. These tools provide multiple levels of detail, from high-level summaries to granular exploration. Different stakeholders have different needs and technical backgrounds. Consequently, effective communicators adjust their language, level of detail, and visualization complexity based on audience needs, ensuring that insights are not just presented but understood and applied.

Phase 7: Implementation and monitoring

The implementation phase often requires connecting analytical teams with operational units that will apply the insights in practice. The ultimate value of analytics comes from the actions they enable that produce valuable results. Implementing analytical results involves:

- Translating insights into specific action plans
- Integrating findings into business processes and decision-making
- Developing implementation timelines and responsibility assignments
- Creating procedures to track the impact of data-driven changes
- Establishing feedback mechanisms to refine approaches based on results

Implementing an actionable insight is rarely a one-time event. When insights have been implemented, ongoing monitoring helps:

- Track the impact of changes made based on analytical findings
- Identify when models or insights need to be updated due to changing conditions
- Discover new questions that emerge from initial results
- Refine methodologies based on observed outcomes

This monitoring creates a feedback loop, potentially initiating new iterations of the data analysis life cycle. Changes will cause business questions to evolve and new data to become available.

Organizations typically encounter several challenges when implementing the data analytics life cycle:

- *Data silos and accessibility issues*: Critical data may be scattered across systems or departments with limited integration.
- *Data quality and governance concerns*: Inconsistent standards for data collection and management can undermine analytical efforts.
- *Skill gaps and resource constraints*: Organizations may lack the specialized skills needed for advanced analytics.
- *Change management hurdles*: Transitioning to data-driven decision-making often requires cultural and process changes.
- *Balancing speed and rigor*: Pressure for quick insights must be balanced with methodological thoroughness.

Incorporating several best practices into the various data analysis life cycle phases can decrease or eliminate some of the challenges:

- *Maintain business alignment*: Keep the original business question at the center of all analytical activities.
- *Embrace iteration*: Recognize that analytics is rarely linear, with insights at one stage often requiring revisiting earlier phases.
- *Document extensively*: Record assumptions, methodologies, and decisions throughout the process for transparency and reproducibility.

- *Foster collaboration*: Build cross-functional teams that combine domain expertise with technical skills.
- *Invest in infrastructure*: Develop a solid data infrastructure and analytics strategy that enables efficient data access, processing, and sharing.
- *Focus on adoption*: Actively work to ensure insights are understood, trusted, and utilized by decision-makers.
- *Prioritize ethical considerations*: Address privacy, fairness, and potential biases throughout the analytics process.

The introduction of AI tools for data analysis can also decrease some of the challenges. For example, using LLMs for data analysis with Quadratic eliminates much of the hassle in data cleaning and preparation. It currently integrates natively with 4 databases and data warehouses as well as APIs for real-time updates, and integrated connections can be requested for other sources.

AI capabilities augment human analysts rather than replace them. The analysts can then focus on higher-value activities requiring judgment and domain expertise. Also, analytics capabilities are expanding to non-specialists because AI for business intelligence provides a natural language interface. For example, Quadratic AI's natural language interface allows non-technical stakeholders to engage directly with data.

Conclusion: From data to insight to action to value

The data analytics life cycle provides a structured framework for transforming raw data into valuable insights and actions. Processes in the phases from problem definition through data collection, preparation, analysis, and communication to implementation can be made more effective and avoid common pitfalls.

However, successful data lifecycle management requires a thoughtful combination of business context, technical expertise, and communication skills. The most successful business analytics from data to insights initiatives recognize that the lifecycle is not just a technical process but a socio-technical one, requiring alignment between people, processes, and technology.

The data analysis process and data life cycle described in this article provide a comprehensive framework for data analytics end-to-end project planning and execution. In a world where data volumes continue to grow exponentially, making the analytics life cycle more efficient and effective represents a crucial advantage that separates data-driven leaders from their competitors.

Knowledge Representation process In data Mining

**KDD Process in Databases**

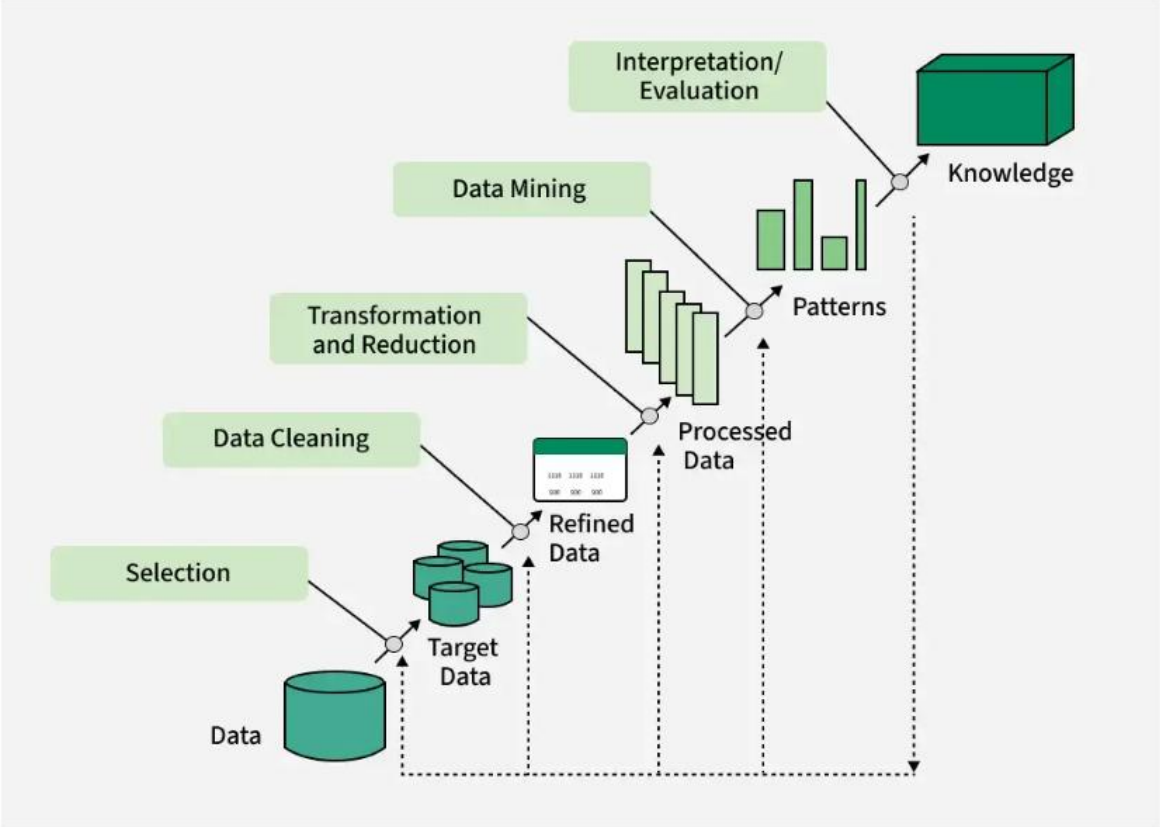
Last Updated : 28 Jan, 2025

- 
- 
- 

Knowledge Discovery in Databases (KDD) refers to the complete process of uncovering valuable knowledge from large datasets. It starts with the selection of relevant data, followed by preprocessing to clean and organize it, transformation to prepare it for analysis, data mining to uncover patterns and relationships, and concludes with the evaluation and interpretation of results, ultimately producing valuable knowledge or insights. KDD is widely utilized in fields like machine learning, pattern recognition, statistics, artificial intelligence, and data visualization.

The KDD process is iterative, involving repeated refinements to ensure the accuracy and reliability of the knowledge extracted. The whole process consists of the following steps:

1. **Data Selection**
2. **Data Cleaning and Preprocessing**
3. **Data Transformation and Reduction**
4. **Data Mining**
5. **Evaluation and Interpretation of Results**




**Data Selection**

Data Selection is the initial step in the Knowledge Discovery in Databases (KDD) process, where relevant data is identified and chosen for analysis. It involves selecting a dataset or focusing on specific variables, samples, or subsets of data that will be used to extract meaningful insights.

- It ensures that only the most relevant data is used for analysis, improving efficiency and accuracy.

- It involves selecting the entire dataset or narrowing it down to particular features or subsets based on the task's goals.
- Data is selected after thoroughly understanding the application domain.

By carefully selecting data, we ensure that the KDD process delivers accurate, relevant, and actionable insights.

**Data Cleaning**

In the KDD process, Data Cleaning is essential for ensuring that the dataset is accurate and reliable by correcting errors, handling missing values, removing duplicates, and addressing noisy or outlier data.

- **Missing Values:** Gaps in data are filled with the mean or most probable value to maintain dataset completeness.
- **Noisy Data:** Noise is reduced using techniques like binning, regression, or clustering to smooth or group the data.
- **Removing Duplicates:** Duplicate records are removed to maintain consistency and avoid errors in analysis.

Data cleaning is crucial in KDD to enhance the quality of the data and improve the effectiveness of data mining.

**Data Transformation and Reduction**

Data Transformation in KDD involves converting data into a format that is more suitable for analysis.

- **Normalization**: Scaling data to a common range for consistency across variables.
- **Discretization**: Converting continuous data into discrete categories for simpler analysis.
- **Data Aggregation**: Summarizing multiple data points (e.g., averages or totals) to simplify analysis.
- **Concept Hierarchy Generation**: Organizing data into hierarchies for a clearer, higher-level view.

Data Reduction helps simplify the dataset while preserving key information.

- **Dimensionality Reduction** (e.g., PCA): Reducing the number of variables while keeping essential data.
- **Numerosity Reduction**: Reducing data points using methods like sampling to maintain critical patterns.
- **Data Compression**: Compacting data for easier storage and processing.

Together, these techniques ensure that the data is ready for deeper analysis and mining.

**Data Mining**

Data Mining is the process of discovering valuable, previously unknown patterns from large datasets through automatic or semi-automatic means. It involves exploring vast amounts of data to extract useful information that can drive decision-making.

Key characteristics of data mining patterns include:

- **Validity**: Patterns that hold true even with new data.
- **Novelty**: Insights that are non-obvious and surprising.
- **Usefulness**: Information that can be acted upon for practical outcomes.
- **Understandability**: Patterns that are interpretable and meaningful to humans.

In the KDD process, choosing the data mining task is critical. Depending on the objective, the task could involve classification, regression, clustering, or association rule mining. After determining the task, selecting the appropriate data mining algorithms is essential. These algorithms are chosen based on their ability to efficiently and accurately identify patterns that align with the goals of the analysis.

**Evaluation and Interpretation of Results**

Evaluation in KDD involves assessing the patterns identified during data mining to determine their relevance and usefulness. It includes calculating the "interestingness score" for each pattern, which helps to identify valuable insights. Visualization and summarization techniques are then applied to make the data more understandable and accessible for the user.

Interpretation of Results focuses on presenting these insights in a way that is meaningful and actionable. By effectively communicating the findings, decision-makers can use the results to drive informed actions and strategies.

**Practical Example of KDD**

Let's assume a scenario that a fitness center wants to improve member retention by analyzing usage patterns.

**Data Selection**: The fitness center gathers data from its membership system, focusing on the past six months of activity. They filter out inactive members and focus on those with regular usage.

**Data Cleaning and Preprocessing**: The fitness center cleans the data by eliminating duplicates and correcting missing information, such as incomplete workout records or member details. They also handle any gaps in data by filling in missing values based on previous patterns.

**Data Transformation and Reduction**: The data is transformed to highlight important metrics, such as the average number of visits per week per member and their most frequently chosen workout types. Dimensionality reduction is applied to focus on the most significant factors like membership duration and gym attendance frequency.

**Data Mining**: By applying clustering algorithms, the fitness center segments members into groups based on their usage patterns. These segments include frequent visitors, occasional users, and those with minimal attendance.

**Evaluation and Interpretation of Results**: The fitness center evaluates the groups by examining their retention rates. They find that occasional users are more likely to cancel their memberships. The interpretation reveals that members who visit the gym less than once a week are at a higher risk of discontinuing their membership.

This analysis helps the fitness center implement effective retention strategies, such as offering tailored incentives and creating engagement programs aimed at boosting the activity of occasional users.

**Difference between KDD and Data Mining**

| Parameter | KDD | Data Mining |
|---|---|---|
| **Definition** | KDD is the overall process of discovering valid, novel, potentially useful, and ultimately understandable patterns and relationships in large datasets. | Data Mining is a subset of KDD, focused on the extraction of useful patterns and insights from large datasets. |
| **Objective** | To extract valuable knowledge and insights from data to support decision-making and understanding. | To identify patterns, relationships, and trends within data to generate useful insights. |
| **Techniques Used** | Involves multiple steps such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge representation. | Includes techniques like association rules, classification, clustering, regression, decision trees, neural networks, and dimensionality reduction. |
| **Output** | Generates structured knowledge in the form of rules, models, and insights that can aid in decision-making or predictions. | Results in patterns, relationships, or associations that can improve understanding or decision-making. |
| **Focus** | Focuses on the discovery of useful knowledge, with an emphasis on interpreting and validating the findings. | Focuses on discovering patterns, relationships, and trends within data without necessarily considering the broader context. |
| **Role of Domain Expertise** | Domain expertise is important in KDD, as it helps in defining the goals of the process, choosing appropriate data, and | Domain expertise is less critical in data mining, as the focus is on using algorithms to detect patterns, often without prior domain-specific |

| Parameter | KDD | Data Mining |
|---|---|---|
| | interpreting the results. | knowledge. |

**Difference Between OLAP and OLTP in Databases**

OLAP (Online Analytical Processing) and OLTP (Online Transaction Processing) are both integral parts of data management, but they have different functionalities.

- OLTP focuses on handling large numbers of transactional operations in **real time**, ensuring data consistency and reliability for daily business operations.
- OLAP is designed for complex queries and data **analysis**, enabling businesses to derive insights from vast datasets through multidimensional analysis.

Let's learn about the differences between them in detail:

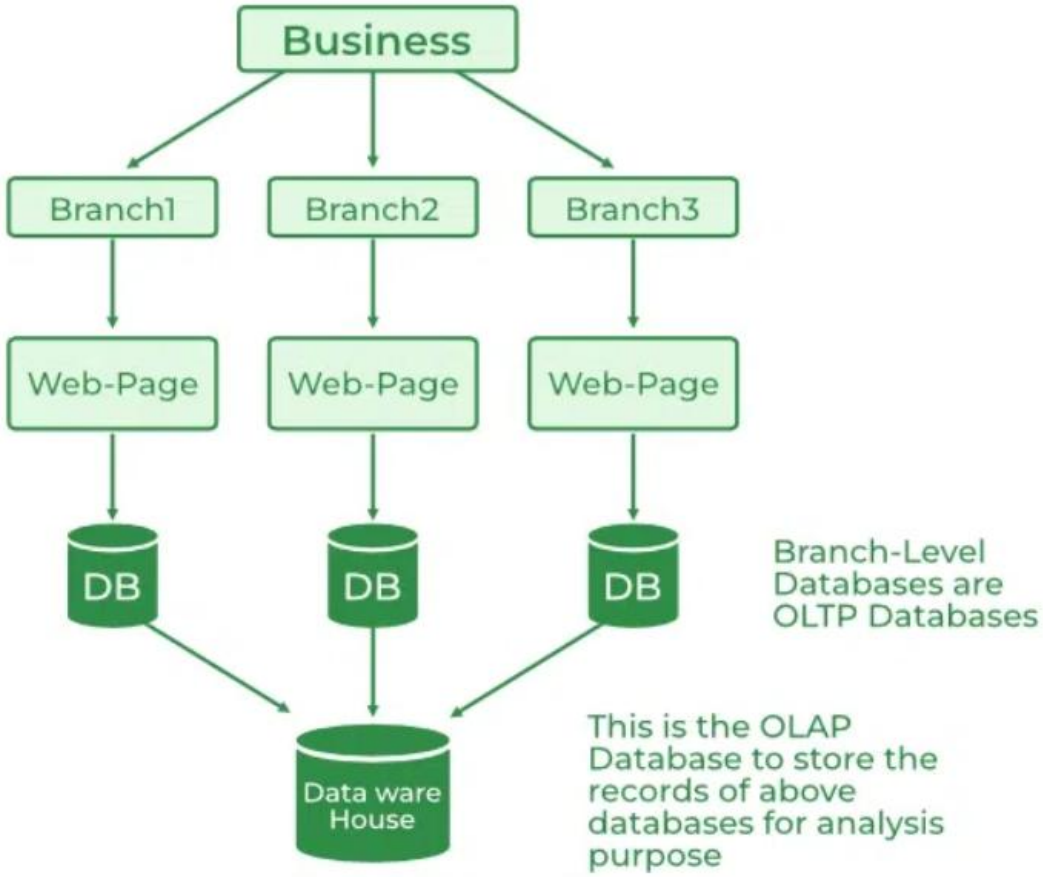
**Online Analytical Processing (OLAP)**

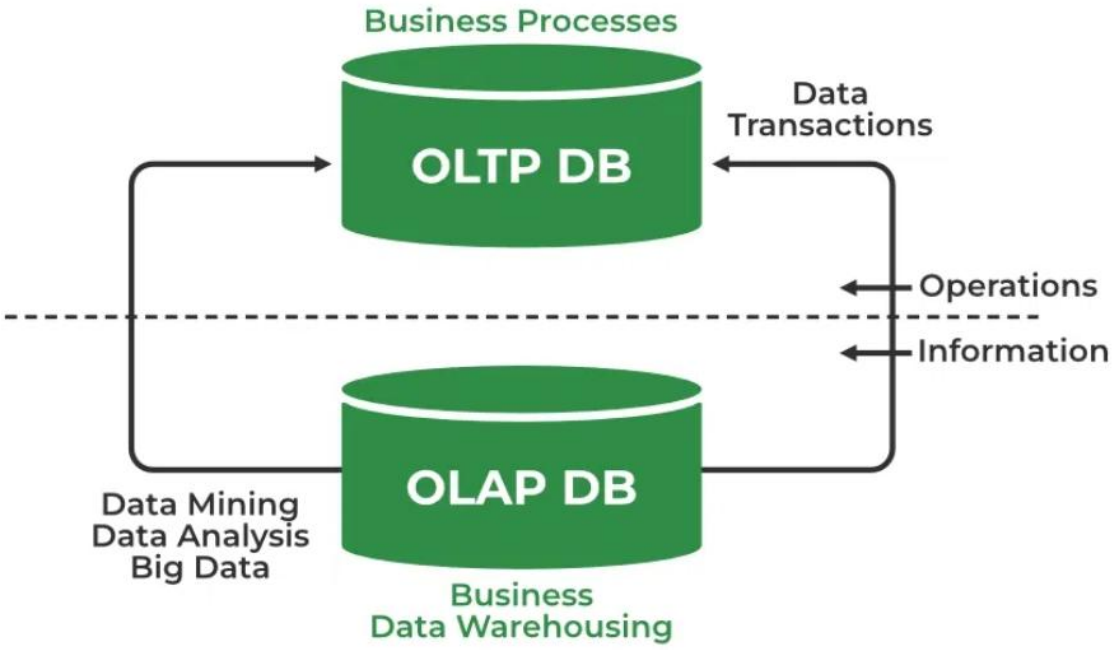
Online Analytical Processing (OLAP) refers to software tools used for the analysis of data in business decision-making processes. OLAP systems generally allow users to extract and view data from various perspectives, many times they do this in a multidimensional format which is necessary for understanding complex interrelations in the data. These systems are part of data warehousing and business intelligence, enabling users to do things like trend analysis, financial forecasting, and any other form of in-depth data analysis.

**OLAP Examples**

Any type of Data Warehouse System is an OLAP system. The uses of the OLAP System are described below.

- Spotify personalizes homepages with custom songs and playlists based on user preferences.
- Netflix movie recommendation system.




**Benefits of OLAP Services**

- Helps in keeping consistency and performing calculation on data.

- Can store planning, analysis, and budgeting for business analytics within one platform.
- Efficiently handle large volumes of data, making them suitable for enterprise-level business applications.
- Assist in applying security restrictions for data protection.
- Provide a multidimensional view of data, which helps in applying operations on data in various ways.

**Drawbacks of OLAP Services**

- Requires professionals to handle the data because of its complex modeling procedure.
- Expensive to implement and maintain in cases when datasets are large.
- Data analysis occurs only after extraction and transformation, leading to system delays.
- Not efficient for decision-making, as it is updated on a periodic basis.

**Online Transaction Processing (OLTP)**

Online Transaction Processing, commonly known as OLTP, is a data processing approach emphasizing real-time execution of transactions. The majority of OLTP systems are meant to manage numerous short atomic operations that keep databases in line. To maintain transaction integrity and reliability, these systems support ACID (Atomicity, Consistency, Isolation, Durability) properties. It is through this that numerous unavoidable applications run their critical courses like online banking, reservation systems etc.

**OLTP Examples**

An example considered for OLTP System is ATM Center a person who authenticates first will receive the amount first and the condition is that the amount to be withdrawn must be present in the ATM. The uses of the OLTP System are described below.

- ATM center is an OLTP application.
- OLTP handles the ACID properties during data transactions via the application.
- It's also used for Online banking, Online airline ticket booking, sending a text message, add a book to the shopping cart.

**Benefits of OLTP Services**

- Allow users to quickly read, write, and delete data operations.
- Support an increase in users and transactions for real-time data access.
- Provide better data protection through multiple security features.
- Aid in decision-making with accurate, up-to-date data.

- Ensure data integrity, consistency, and high availability.

**Drawbacks of OLTP Services**

- Limited analysis capability, not suited for complex analysis or reporting.
- High maintenance costs due to frequent updates, backups, and recovery.
- Susceptible to disruption during hardware failures, impacting online transactions.
- Prone to issues like duplicate or inconsistent data.

**Difference Between OLAP and OLTP**

| Category | OLAP (Online Analytical Processing) | OLTP (Online Transaction Processing) |
|---|---|---|
| Definition | It is well-known as an online database query management system. | It is well-known as an online database modifying system. |
| Data source | Consists of historical data from various Databases. | Consists of only operational current data. |
| Method used | It makes use of a data warehouse. | It makes use of a standard database management system (DBMS). |
| Application | It is subject-oriented. Used for Data Mining, Analytics, Decisions making, etc. | It is application-oriented. Used for business tasks. |
| Normalized | In an OLAP database, tables are not normalized. | In an OLTP database, tables are normalized (3NF). |
| Usage of data | The data is used in planning, problem-solving, and decision-making. | The data is used to perform day-to-day fundamental operations. |
| Task | It provides a multi-dimensional view of different business tasks. | It reveals a snapshot of present business tasks. |

| Category | OLAP (Online Analytical Processing) | OLTP (Online Transaction Processing) |
|---|---|---|
| Purpose | It serves the purpose to extract information for analysis and decision-making. | It serves the purpose to Insert, Update, and Delete information from the database. |
| Volume of data | A large amount of data is stored typically in TB, PB | The size of the data is relatively small as the historical data is archived in MB, and GB. |
| Queries | Relatively slow as the amount of data involved is large. Queries may take hours. | Very Fast as the queries operate on 5% of the data. |
| Update | The OLAP database is not often updated. As a result, data integrity is unaffected. | The data integrity constraint must be maintained in an OLTP database. |
| Backup and Recovery | It only needs backup from time to time as compared to OLTP. | The backup and recovery process is maintained rigorously |
| Processing time | The processing of complex queries can take a lengthy time. | It is comparatively fast in processing because of simple and straightforward queries. |
| Types of users | This data is generally managed by CEO, MD, and GM. | This data is managed by clerksForex and managers. |
| Operations | Only read and rarely write operations. | Both read and write operations. |
| Updates | With lengthy, scheduled batch operations, data is refreshed on a regular basis. | The user initiates data updates, which are brief and quick. |

| Category | OLAP (Online Analytical Processing) | OLTP (Online Transaction Processing) |
|---|---|---|
| Nature of audience | The process is focused on the customer. | The process is focused on the market. |
| Database Design | Design with a focus on the subject. | Design that is focused on the application. |
| Productivity | Improves the efficiency of business analysts. | Enhances the user's productivity. |

**Excel vs Tableau vs Power BI: Choosing the Right Tool for Your Data Analysis Needs**

Suramaparna

6 min read

.

Feb 18, 2024

--

**Introduction**

**Microsoft Excel:** The Classic Spreadsheet Powerhouse

Originally released 34 years ago, Microsoft Excel is one of the most widely used spreadsheet applications in the world. Excel is part of Microsoft Office and features calculations, graphing tools, pivot tables, and a macro programming language called Visual Basics for Applications (VBA).

It is is likely the first tool many will choose for starting financial, mathematical, and statistical analysis in many different settings — personal, commercial, and educational.

**Tableau:** Elevating Data Visualization to New Heights

Founded in 2003, Tableau is an interactive data visualization software. It is recognized as the fastest growing data visualization tool mainly because of its ability to help users see and understand data. It simply converts raw data into a comprehensible visual that transforms the way people use data for problem solving and decision making.

**Power BI:** Microsoft's Comprehensive Business Intelligence Solution

Get Suramaparna's stories in your inbox

Join Medium for free to get updates from this writer.

Subscribe

Released in 2014. Microsoft Power BI is a self-service analytics and business intelligence (BI) platform that connects and visualizes any data using a unified, scalable platform. The original Power BI was based on Excel's add-ins — Power Query, Power Pivot and Power View.

**Some key features of Excel, Tableau, and Power BI:**

**Excel:**

1. Spreadsheet Functionality: Excel provides powerful spreadsheet capabilities, including data entry, manipulation, and calculation using formulas and functions.
2. PivotTables: Excel offers PivotTables for summarizing and analyzing large datasets quickly.
3. Charts and Graphs: Users can create various types of charts and graphs to visualize data, such as bar charts, line graphs, and pie charts.
4. Data Analysis Tools: Excel includes built-in data analysis tools like regression analysis, data tables, and scenario management.
5. Customization: Users can customize charts, graphs, and formatting to suit their needs.
6. Integration: Excel integrates with other Microsoft Office applications, such as Word and PowerPoint, for seamless data sharing and reporting.

**Tableau:**

1. Interactive Dashboards: Tableau allows users to create interactive dashboards and reports with drag-and-drop functionality.
2. Data Connection: Tableau connects to various data sources, including databases, spreadsheets, and cloud services, enabling users to blend and analyze data from multiple sources.
3. Advanced Visualizations: Tableau offers a wide range of advanced visualization options, such as heatmaps, tree maps, and geographic maps, to uncover insights in data.
4. Data Preparation: Tableau Prep provides tools for data preparation, cleaning, and shaping before analysis.
5. Collaboration: Tableau Server and Tableau Online allow for easy sharing and collaboration on dashboards and reports within organizations.
6. Mobile Compatibility: Tableau dashboards are compatible with mobile devices, enabling users to access and interact with data on the go.

**Power BI:**

1. Data Modeling: Power BI offers robust data modeling capabilities, allowing users to create relationships between different data tables and define measures and calculations.
2. Data Visualization: Power BI enables users to create interactive and visually appealing reports and dashboards with a variety of visualization options.
3. Integration: Power BI seamlessly integrates with other Microsoft products, such as Excel, Azure, and SQL Server, as well as third-party applications and services.
4. AI-powered Insights: Power BI provides AI-powered insights and natural language querying capabilities to help users discover trends and patterns in their data.
5. Data Connectivity: Power BI connects to a wide range of data sources, both on-premises and in the cloud, allowing users to analyze data from diverse sources.
6. Collaboration and Sharing: Power BI Service allows for easy sharing and collaboration on reports and dashboards within organizations, with features like scheduled data refresh and row-level security.

These are just some of the key features offered by Excel, Tableau, and Power BI. Each tool has its strengths and is suited to different use cases and user preferences.

**Limitations of each tool**

Excel:

1. Performance: Excel can become slow and unstable when dealing with large datasets or complex calculations.
2. Limited Data Visualization: While Excel offers basic charting capabilities, it may not be suitable for creating advanced or interactive visualizations.
3. Data Integrity: Excel lacks built-in features for ensuring data integrity and consistency, which can lead to errors in analysis.
4. Version Control: Managing multiple versions of Excel files and ensuring data consistency across different versions can be challenging.
5. Scalability: Excel may not scale well for enterprise-level data analysis and reporting needs, leading to inefficiencies and data management challenges.
6. Limited Collaboration: Collaborating on Excel files can be cumbersome, especially when multiple users need to work on the same file simultaneously.

Tableau:

1. Cost: Tableau can be expensive for organizations, particularly for larger deployments or enterprise-level usage.
2. Steep Learning Curve: While Tableau offers powerful capabilities, mastering its features may require significant time and effort.

3. Performance Issues: Tableau may experience performance issues when dealing with extremely large datasets or complex visualizations.

4. Data Preparation: Tableau's data preparation capabilities are not as robust as some other tools, requiring users to perform data cleaning and shaping outside the platform.

5. Mobile Compatibility: While Tableau offers mobile compatibility, creating dashboards optimized for mobile devices can be challenging and may require additional development effort.

6. Limited Customization: Tableau's customization options for visualizations may be limited compared to other tools, restricting users' ability to tailor visualizations to their specific needs.

Power BI:

1. Complexity: Power BI can be complex, especially for users with limited technical skills, requiring training and expertise to fully utilize its capabilities.

2. Cost: While Power BI offers a free version, more advanced features and capabilities require a paid subscription, which can be costly for organizations.

3. Data Connectivity: Power BI's connectivity options may be limited for certain data sources, requiring additional connectors or custom solutions.

4. Performance: Power BI may experience performance issues when dealing with large datasets or complex calculations, especially in the desktop version.

5. Data Model Limitations: Power BI has limitations on the size and complexity of data models, which may restrict users' ability to handle extremely large datasets or complex relationships.

6. Dependency on Microsoft Ecosystem: Power BI's close integration with other Microsoft products may be a limitation for organizations using non-Microsoft technologies or platforms.

These limitations highlight the potential challenges and constraints that users may encounter when working with Excel, Tableau, and Power BI. Understanding these limitations can help organizations make informed decisions about which tool best suits their specific needs and requirements.

**When to choose which Tool, and Why**

**Choose Excel When:**

1. Basic Data Analysis: Excel is suitable for basic data analysis tasks, such as creating simple charts, performing calculations, and organizing data.

2. Small Datasets: Excel is ideal for handling small to medium-sized datasets that can be comfortably managed within a spreadsheet.

3. Familiarity: If your team is already familiar with Excel and requires a quick and familiar solution for data analysis, Excel may be the right choice.

4. Limited Budget: Excel is often the most cost-effective option, especially for small businesses or individuals with limited budgets.

5. Ad-hoc Reporting: Excel is well-suited for ad-hoc reporting and one-off analysis tasks that do not require advanced visualization or complex data modeling.

**Choose Tableau When:**

1. Advanced Data Visualization: If your project requires creating visually appealing and interactive dashboards, Tableau excels in this aspect.

2. Large Datasets: Tableau can handle large and complex datasets more efficiently than Excel, making it suitable for enterprise-level analysis.

3. Exploratory Analysis: If your analysis involves exploring data from multiple angles and uncovering insights through visual exploration, Tableau's interactive capabilities are beneficial.

4. User-Friendly Interface: Tableau's intuitive drag-and-drop interface makes it accessible to users with varying levels of technical expertise, enabling quick and easy visualization creation.

5. Real-time Analytics: If your project requires real-time data streaming and analysis, Tableau offers robust capabilities for real-time analytics.

**Choose Power BI When:**

1. Comprehensive Business Intelligence: If you need a comprehensive business intelligence solution that integrates seamlessly with other Microsoft products, Power BI is a strong choice.

2. Data Modeling: Power BI's robust data modeling capabilities make it suitable for defining relationships and calculations within the data model, making it ideal for complex analysis.

3. AI-powered Insights: If your project requires leveraging AI capabilities for uncovering trends and patterns in data, Power BI offers advanced AI-powered insights.

4. Enterprise-level Integration: Power BI integrates seamlessly with other Microsoft products and third-party applications, making it suitable for organizations with existing Microsoft infrastructure.

5. Collaboration and Sharing: If your project involves collaboration and sharing of reports and dashboards within the organization, Power BI Service provides robust collaboration features.

Ultimately, the choice between Excel, Tableau, and Power BI depends on your specific needs, preferences, and the capabilities required for your project or organization. Consider factors such as data volume, complexity, visualization requirements, budget, and team expertise when making your decision.

**Conclusion: Finding Your Perfect Fit**

In conclusion, the choice between Excel, Tableau, and Power BI boils down to your specific requirements, project scope, and organizational needs. Excel remains the versatile workhorse, ideal for basic analysis and cost-effective solutions. Tableau emerges as the visualization maestro, empowering users with advanced visual storytelling capabilities. Meanwhile, Power BI stands tall as the comprehensive business intelligence solution, seamlessly integrating with Microsoft's ecosystem and offering robust data modeling and AI-powered insights.

Understanding the strengths and limitations of each tool is crucial in making an informed decision. Whether you opt for Excel's familiarity, Tableau's visualization prowess, or Power BI's comprehensive suite of features, rest assured that each tool empowers you to unlock the insights hidden within your data, driving informed decisions and strategic growth for your organization