

Unit III: Business Analytics Ecosystem

1. Differentiate SQL and NoSQL.

SQL(Structured Query Language) is the primary interface used to communicate with Relational Databases. SQL is used to add, update or delete rows of data, retrieving subsets of data for transaction processing and analytics applications, and to manage all aspects of the database. SQL is a programming language for Relational Databases. It is designed over relational algebra and tuple relational calculus. SQL comes as a package with all major distributions of RDBMS(Relational Database Management System).

SQL Commands are:

- 1) DDL(Data Definition Language) – Create, Alter, Drop, Truncate and Rename
- 2) DML(Data Manipulation Language) – Insert, Update and Delete
- 3) TCL(Transaction Control Language) – Savepoint, Commit and Rollback
- 4) DCL(Data Control Language) – Grant and Revoke

NoSQL Database is a non-relational Data Management System that does not require a fixed schema. It avoids joins and is easy to scale. The major purpose of using a NoSQL database is for distributed data stores with homogenous data storage needs. NoSQL is used for Big Data and real- time web apps. For example, companies like Twitter, Facebook and Google collect terabytes of user data every single day. **NoSQL database** stands for “Not Only SQL” or “Not SQL.” Carl Stroz introduced the NoSQL concept in 1998. Traditional RDBMS uses SQL syntax to store and retrieve data for further insights. Instead, a NoSQL database system encompasses a wide range of database technologies that can store structured, semi-structured, unstructured and polymorphic data

2. Explain ACID Properties.

The acronym ACID refers to the four key properties of a transaction are

- Atomicity
- Consistency
- Isolation
- Durability

Atomicity

- A transaction is treated as a single, indivisible unit of work.
- Either all operations within the transaction are successfully executed, or none are.
- If any part of the transaction fails, the entire transaction is rolled back, leaving the system in its original state.

Consistency

- A transaction ensures that the database transitions from one valid state to another valid state.
- The integrity constraints of the database are preserved before and after the transaction.

Isolation

- Transactions occurring concurrently do not interfere with each other.
- The intermediate state of a transaction is invisible to other transactions.

Durability

- Once a transaction is committed, its results are permanent, even in the case of a system failure.
- The changes made by the transaction are saved to a durable storage medium

3. Write about MongoDB

MongoDB is a NoSQL database that provides a flexible, scalable, and high-performance data storage solution. It is designed to handle unstructured and semi-structured data. It stores data in a document-oriented format using BSON (Binary JSON), which allows for dynamic and schema-less data structures. This makes MongoDB ideal for handling unstructured or semi-structured data in applications like real-time analytics, content management, and Internet of Things (IoT) systems. Its distributed architecture supports horizontal scaling through sharding and ensures high availability with replica sets. MongoDB also features a powerful aggregation framework for complex data processing and offers robust indexing options for fast query performance. With its ease of use and ability to handle large-scale data efficiently, MongoDB has become a popular choice for modern, data-intensive applications.

Advantages

- **Flexibility:** No schema restrictions.
- **Performance:** Optimized for read and write operations.
- **Ease of Use:** Intuitive query language

4. Write about Python & R

(Or)

What are the statistical tools in Business Analytics

R and Python are both open-source programming languages with a large community. New libraries or tools are added continuously to their respective catalog. R is mainly used for statistical analysis while Python provides a more general approach to data science.

R and Python are state of the art in terms of programming language oriented towards data science. Learning both of them is, of course, the ideal solution. R and Python requires a time- investment, and such luxury is not available for everyone. Python is a general-purpose language with a readable syntax. R, however, is built by statisticians and encompasses their specific language.

Python

Usage Scenarios

- Data crawling
- Data cleaning
- Data modeling
- Construct data analysis algorithms based on the business scenarios and actual problems
- Data visualization
- Advanced fields of data mining and analysis, such as machine learning and text mining

R Usage Scenarios

The functions of R cover almost any area where data is needed. As far as our general data analysis or academic data analysis work is concerned, the things that R can do mainly include the following aspects.

- Data cleaning and data reduction
- Web crawling
- Data visualization
- Statistical hypothesis testing (t test, analysis of variance, chi-square test, etc.)
- Statistical modeling (linear regression, logistic regression, tree model, neural network, etc.)

5. Write about Big Data

Big Data refers to the massive and complex datasets that traditional data processing tools cannot efficiently handle. It is characterized by the 5 V's: volume (large amounts of data), velocity (high-speed data generation and processing), variety (diverse data types like text, images, and videos), veracity (ensuring data accuracy despite inconsistencies), and value (deriving actionable insights). Technologies like Hadoop, Apache Spark, and NoSQL databases such as MongoDB and Cassandra are commonly used to store and process Big Data. It has transformative applications across industries, including business intelligence, healthcare, finance, government, and IoT, enabling real-time analytics and data-driven decision-making. However, challenges such as data privacy, security, and the need for advanced infrastructure and expertise must be addressed to fully leverage its potential. Big Data continues to revolutionize industries by uncovering trends and insights that drive innovation and efficiency.

Advantages of Big Data

- ❖ **Better Decision-Making:** Enables data-driven strategies and informed decisions.
- ❖ **Improved Customer Experience:** Personalizes interactions and enhances customer satisfaction.
- ❖ **Increased Efficiency:** Optimizes processes like supply chain and resource management.

- ❖ **Cost Savings:** Identifies inefficiencies and reduces operational costs.
- ❖ **Fraud Detection:** Detects anomalies and minimizes financial risks.
- ❖ **Real-Time Insights:** Provides quick responses to market changes or operational needs.
- ❖ **Market Understanding:** Analyses trends and customer behaviour for competitive advantage.
- ❖ **Innovation:** Supports the development of new products and services based on data insights.
- ❖ **Risk Management:** Predicts and mitigates potential risks effectively.
- ❖ **Scalability:** Handles growing volumes of data as businesses expand.

6. What is the purpose of Machine Learning?

The purpose of machine learning is to discover patterns in the data and then make predictions based on often complex patterns to answer business questions, detect and analyses trends and help solve problems. Machine learning in business and other fields is effectively a method of data analysis that works by automating the process of building data models. Machine learning is the ability of a computer system or software platform to teach itself. It is a form of ‘self-programming’ that allows a software platform to come to its own decisions via datasets and outside information without the need for direct human programming.

Machine learning allows a computer, software platform, system, etc. to actively react to new information. This means that when a novel situation or dataset presents itself the machine will modify its process to fit this new information which in turn modifies its results. The existing method of software and system development creates software that cannot ‘think’ of its own results responses and is limited to a set series of possible responses coded into the software by its programmers.

7. What are Visualization Tools?

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions. Our eyes are drawn to colors and patterns. We can quickly identify red from blue, square from circle. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly.

Popular Data Visualization Tools

- ❖ **Tableau:** User-friendly, powerful analytics and interactive dashboards.
- ❖ **Microsoft Power BI:** Affordable, integrates well with Microsoft products, and offers robust features.
- ❖ **Google Data Studio:** Free tool with seamless integration with Google services.
- ❖ **D3.js:** JavaScript library for creating highly customized visualizations.
- ❖ **Excel:** A classic tool for basic visualizations and small-scale data analysis.