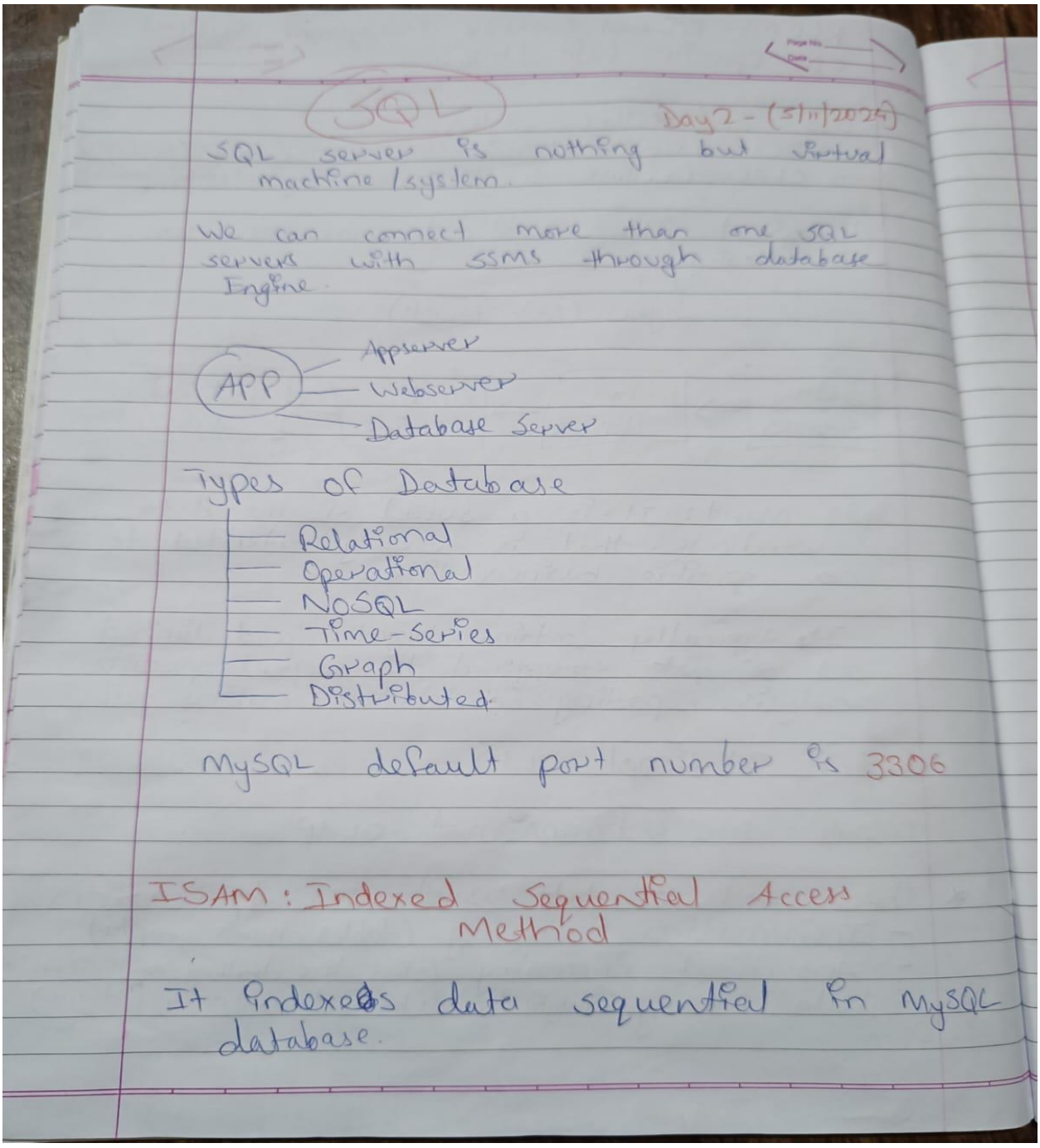
# Summary of Day 2 – Data Engineering

-Sarthak Niranjan Kulkarni (Maverick)

- sarthakkul2311@gmail.com - (+91) 93256 02791

**Day 2 – 5/11/2024 (Tuesday)**

- **Handwritten Notes:-**



Terminal that sets path:-
/var/lib/mysql/bin

Command that connects the server:-
mysql -u root -p

DDL = Data Definition Language (create, ~~delete~~ drop)

DML = Data Manipulation Language (retrieve, update) (select, insert, merge)

DCL = Data Control Language. (Grant, revoke)

Data Cleaning: It is a process of data cleaning involving fixing & identifying inaccurate, incomplete, inconsistent or irrelevant data from a dataset.

Data Manipulation: It refers to the process of adjusting, organizing or transforming data to make it more useful for analysis, reporting or other purposes.

- Fix Structural Error
- Data Cleaning
- Data Transformation
- Aggregation and Summarization.

## ➢ Digital Notes:-

### 1. SQL Server: -

→ SQL Server is a relational database management system by Microsoft designed for storing, managing, and retrieving data efficiently. It uses T-SQL for querying, offers high security, scalability, and performance optimization, and includes tools like SQL Server Management Studio (SSMS) for easy management. Key features include support for data warehousing, business intelligence, and high availability through disaster recovery solutions. SQL Server also integrates well with Microsoft Azure, making it popular for both on-premises and cloud-based applications.

- **Advanced Security**: SQL Server provides robust security features like data encryption, row-level security, dynamic data masking, and advanced auditing to protect sensitive information.
- **High Availability and Disaster Recovery (HADR)**: SQL Server includes features like Always On Availability Groups, failover clustering, and replication to ensure high availability and disaster recovery.
- **Integration and ETL**: SQL Server Integration Services (SSIS) allows users to perform ETL (Extract, Transform, Load) operations for moving and transforming data across different systems.
- **Support for Advanced Analytics**: SQL Server supports in-database analytics with integration for R and Python, enabling users to perform data science and machine learning directly within the database.
- **Memory-Optimized Tables**: SQL Server includes in-memory OLTP to boost performance for transaction processing by using memory-optimized tables and natively compiled stored procedures.
- **Data Warehousing and Big Data**: SQL Server provides columnstore indexes, PolyBase, and integration with Hadoop and Spark, making it suitable for big data and data warehousing workloads.

---

**2. Types Of Database: -**

→

- Relational Databases
- NoSQL Databases
- Hierarchical Databases
- Object-Oriented Databases
- Distributed Databases
- Cloud Databases
- Graph Databases
- Time-Series Databases
- Key-Value Stores
- Column-Family Databases

---

**3. ISAM: -**

→ **Indexed Sequential Access Method (ISAM)** is a data access method developed by IBM that organizes and retrieves records by using both indexes and sequential access. It stores data in a sorted order with a primary index for direct access and optional secondary indexes for alternative retrieval methods. While ISAM allows efficient querying and updating of records, its static structure can lead to performance degradation as data grows, making it less flexible compared to more modern data access methods. ISAM was widely used in early database systems but has largely been replaced by more advanced techniques.

---

**4. DDL :-**

→ **Data Definition Language (DDL)** is a subset of SQL used to define and manage all aspects of database structures. It includes commands that allow users to create, alter, and drop database objects such as tables, indexes, and schemas. Common DDL commands include:

- **CREATE**: To create new database objects.
- **ALTER**: To modify existing objects (e.g., adding or deleting columns).
- **DROP**: To remove objects from the database.

---

5. **DML: -**

→ **Data Manipulation Language (DML)** is a subset of SQL used for managing and manipulating data within a database. It includes commands that allow users to perform various operations on the data, such as:

- **SELECT**: To retrieve data from one or more tables.
- **INSERT**: To add new records to a table.
- **UPDATE**: To modify existing records in a table.
- **DELETE**: To remove records from a table.

---

6. **DCL: -**

→ **Data Control Language (DCL)** is a subset of SQL used to control access to data within a database. It includes commands that define user permissions and access rights, ensuring data security and integrity. The primary DCL commands are:

- **GRANT**: To provide specific privileges to users or roles, allowing them to perform certain operations on database objects.
- **REVOKE**: To remove previously granted privileges from users or roles.

---

7. **Data Cleaning: -**

→ **Data cleaning** is the process of identifying and correcting errors, inconsistencies, and inaccuracies in a dataset to improve its quality and reliability. It involves several key steps, including:

- **Identifying Errors**: Detecting issues such as missing values, duplicates, and outliers.
- **Correcting Errors**: Addressing identified issues by filling in missing values, removing duplicates, and standardizing formats.
- **Validating Data**: Ensuring that the cleaned data meets specific quality criteria and is accurate for analysis.
- **Documentation**: Keeping a record of the cleaning process for transparency and reproducibility.

---

## 8. Data Manipulation: -

→ **Data manipulation** refers to the process of adjusting, organizing, or transforming data to make it more useful for analysis, reporting, or decision-making. Key steps in data manipulation include:

- **Data Collection**: Gathering data from various sources.
- **Data Cleaning**: Identifying and correcting errors or inconsistencies.
- **Data Transformation**: Modifying data formats and structures to suit analysis needs.
- **Filtering and Sorting**: Extracting relevant subsets and organizing data.
- **Aggregation**: Summarizing data using functions like sum and average.
- **Merging**: Combining data from multiple sources or tables.

---