## Table Operations (CQL)

1. **Create Table:**

```
CREATE TABLE employees (
    employee_id int PRIMARY KEY,
    name text,
    commission int,
    total_salary int
);
```

2. **Alter Table:**

```
ALTER TABLE employees ADD
department text;
```

3. **Drop Table:**

```
DROP TABLE employees;
```

4. **Truncate Table:**

```
TRUNCATE employees;
```

5. **Create Index:**

```
CREATE INDEX ON employees
(commission);
```

6. **Drop Index:**

```
DROP INDEX
employees_commission_idx;
```

7. **Batch Operation:** Batch allows multiple commands to be executed together. Example of a batch operation:

```
BEGIN BATCH
    UPDATE employees SET
total_salary = 20000 WHERE
commission = 0;
APPLY BATCH;
```

## CRUD Operations

1. **Create:**

```
INSERT INTO employees
(employee_id, name, commission,
total_salary)
VALUES (1, 'John', 5, 30000);
```

2. **Update:**

```
UPDATE employees SET total_salary
= 25000 WHERE employee_id = 1;
```

3. **Read:**

```
SELECT * FROM employees WHERE
employee_id = 1;
```

4. **Delete:**

```
DELETE FROM employees WHERE
employee_id = 1;
```

## CQL Types

1. **CQL Datatypes:** Common data types include:
    - `text`: Strings
    - `int`: Integer
    - `float`: Floating point
    - `uuid`: Universally Unique Identifier
2. **CQL Collections:** Cassandra supports collections like `list`, `set`, and `map`:
    - **List:** Ordered collection
    - **Set:** Unordered collection with unique elements
    - **Map:** Key-value pairs
3. **User-Defined Datatypes (UDT):** Create UDT in Cassandra to define a structure:

```
CREATE TYPE address (
    street text,
    city text,
    zip_code int
);
```

## Task-Specific CQL Commands

1. **Create Table `employees`:**

```
CREATE TABLE employees (
    employee_id int PRIMARY KEY,
    name text,
    commission int,
    total_salary int
);
```

2. **Update employee’s total salary to 20000 where commission is '0':**

```
UPDATE employees SET total_salary
= 20000 WHERE commission = 0;
```

3. **Create Tables using Collections:**
   - **Teachers and Subjects:**

```
CREATE TABLE teachers_subjects
(
    teacher_name text PRIMARY
KEY,
    subjects list<text>
);
```

   - **Books and Authors:**

```
CREATE TABLE books_authors (
    book_name text PRIMARY KEY,
    authors set<text>
);
```

4. **Batch Operation for Employees Table:** Insert a value, update salary for `employee_id = 03`, and change names to uppercase for employees whose name starts with 'N':

```
BEGIN BATCH
    INSERT INTO employees
(employee_id, name, commission,
total_salary)
    VALUES (3, 'Alice', 5, 18000);
```

```
    UPDATE employees SET
total_salary = 25000 WHERE
employee_id = 3;

    UPDATE employees SET name =
upper(name) WHERE name LIKE 'N%';
APPLY BATCH;
```

5. **Print all values from Books Table:**

```
SELECT * FROM books_authors;
```

## Data-Driven Decisions

**Data-driven decisions** involve making choices based on data analysis and interpretation rather than intuition, personal experiences, or guesswork. In today's world, where large volumes of data are generated across various industries, it is essential to base business strategies, operational decisions, and even day-to-day activities on insights extracted from accurate data.

### *Why Data-Driven Decisions are Important:*

1. **Informed Choices**: Provides a factual basis for decision-making rather than relying on assumptions or gut feelings.
2. **Predictive Insights**: Helps in forecasting trends, behaviors, and outcomes based on historical data.
3. **Improved Efficiency**: Enhances operational processes and minimizes the risk of errors.
4. **Enhanced Customer Experiences**: Data-driven insights help in understanding customer behavior and preferences, enabling personalized services or products.
5. **Competitive Advantage**: Organizations that leverage data effectively can better anticipate market changes, optimize operations, and innovate faster than competitors.

---

## Enterprise Data Management (EDM)

**Enterprise Data Management (EDM)** refers to the strategies, technologies, and processes used to manage, store, organize, and ensure the quality and accessibility of data within an organization. Effective EDM ensures that the right data is available to the right people at the right time.

EDM involves various stages such as **data preparation**, **data cleaning**, **data integration**, **data governance**, and **data storage** to ensure the overall data lifecycle is well managed and aligned with business goals.

### *Key Components of EDM:*

1. **Data Governance**: Establishes policies, standards, and procedures to ensure the data's accuracy, security, and proper usage.
2. **Data Integration**: Combines data from different sources into a cohesive and usable format.
3. **Data Quality**: Ensures the completeness, accuracy, and timeliness of data.
4. **Master Data Management (MDM)**: Ensures consistency and uniformity in critical business data like customer and product information across the enterprise.

---

## Data Preparation

**Data Preparation** is the process of gathering, cleaning, transforming, and organizing raw data to make it suitable for analysis. This step is critical because it ensures the quality and relevance of data used for making decisions.

### *Key Steps in Data Preparation:*

1. **Data Collection**: Gathering raw data from different sources, including databases, APIs, spreadsheets, or external sources like social media or IoT devices.
2. **Data Profiling**: Understanding the data's structure, quality, and relationships. This involves analyzing the data to identify trends, outliers, missing values, and inconsistencies.
3. **Data Transformation**: Converting data into a consistent format. This can include:
    - **Normalization**: Scaling values to a common range.
    - **Aggregation**: Summarizing data (e.g., weekly sales totals).
    - **Encoding**: Converting categorical data into numerical format (for use in machine learning models).
4. **Data Structuring**: Organizing data into usable formats such as tables, databases, or files. This ensures that the data can be efficiently queried and analyzed.
5. **Data Enrichment**: Augmenting data by adding relevant information from external or additional

sources (e.g., demographic data, geographic information).

### *Importance of Data Preparation:*

- Ensures that data is in the right format and of high quality before analysis.
- Reduces the likelihood of **garbage in, garbage out** situations where incorrect data leads to faulty insights.
- Saves time for analysts by eliminating errors or irrelevant information early in the process.

---

## Data Cleaning

**Data Cleaning** (also known as **data cleansing** or **data scrubbing**) is the process of detecting and correcting (or removing) errors and inconsistencies from the data to improve its quality. This ensures that the analysis based on the data is accurate and reliable.

### *Common Data Cleaning Steps:*

1. **Handling Missing Data**:
    - **Imputation**: Filling missing values using statistical methods (e.g., mean, median) or more advanced methods like machine learning predictions.

    - **Deletion**: Removing records with missing values if they represent a small portion of the dataset or if imputation is not feasible.
2. **Removing Duplicates**:
    - **Duplicate Records**: Identifying and removing identical entries to avoid skewing analysis results.
3. **Correcting Inconsistent Data**:
    - **Standardization**: Ensuring consistency in data formatting, such as dates (DD-MM-YYYY vs MM-DD-YYYY), address formats, and text capitalization.
    - **Normalization**: Ensuring numerical values follow a consistent range or scale.
4. **Fixing Data Entry Errors**:
    - **Typos or Misspellings**: Detecting and correcting typographical errors in text data.
    - **Incorrect Data Types**: Identifying mismatches (e.g., text in numeric columns) and converting them to appropriate types.
5. **Handling Outliers**:
    - **Outlier Detection**: Identifying extreme or abnormal values that might skew analysis results (e.g., sales spikes, sensor errors).
    - **Outlier Treatment**: Depending on the context, outliers can be removed, adjusted, or kept if they have meaningful significance.
6. **Validation**:

    - Ensuring that data values fall within acceptable ranges or categories.
    - **Cross-checking**: Verifying data accuracy by comparing it with external sources or historical data.

### *Importance of Data Cleaning:*

- **Improves Data Accuracy**: Clean data leads to better insights, more accurate predictions, and overall trust in the data.
- **Enhances Efficiency**: Cleaning data upfront reduces the time spent on correcting errors during analysis.
- **Prevents Misleading Results**: Erroneous or inaccurate data can result in faulty decisions or analyses.
- **Improves Data Quality**: High-quality data is essential for data-driven decision-making processes in enterprise environments.

---