**1️. Database Schema Design**

The schema is designed following the **Third Normal Form (3NF)** to ensure **efficient data storage, minimal redundancy, and faster query performance**. The schema consists of **dimension and fact tables**, each serving a specific role.

**Dimension Tables**

Dimension tables store descriptive attributes about business entities. They have **a primary key** that uniquely identifies each record and do not rely on other tables.

**Fact Tables**

Fact tables contain **transactional data** that depend on dimension tables for contextual information. These tables store **measurable numerical values** (e.g., sales amount, inventory count) and have **foreign keys** referencing multiple dimension tables. Fact tables are often **partitioned** by yearmonth concept to optimize query performance.

**Identified Dimension and Fact Tables**

* **Dimension Tables:**
  + Dim\_Plant – Stores information about manufacturing plants.
  + Dim\_Supplier – Contains supplier details such as name, location, and contract terms.
  + Dim\_Employee – Holds employee-related information, including roles and department assignments.
  + Dim\_Time – A dedicated time dimension table **partitioned on YearMonth** to efficiently process time-based queries and partitions.
* **Fact Tables:**
  + Fact\_Inventory – Stores stock levels, product movements, and warehouse tracking details.
  + Fact\_Sales – Captures order transactions, revenue, and product-level sales performance.
  + Fact\_EmployeeShiftTime – Maintains work hours, shift schedules, and employee attendance logs.

**2️. ETL Pipeline Development**

**The Extract, Transform, and Load (ETL) pipeline automates the process of collecting data, ensuring quality, and loading it into the database. This ensures that the latest, clean data is available for reporting and analytics.**

**Extract: Fetch data from source**

* **Data Source:** Excel files stored in a **SharePoint folder** with a well-defined **naming convention** to ensure consistency.
* **Metadata Tracking:**
  + The system monitors the **file metadata** (e.g., creation timestamp, modification date) before processing.
  + If a new file is detected, it is loaded into a **staging table** before transformation.
  + Only the most recent file is processed to **prevent duplicate loads**.

**Transform: Data Cleaning & Validation:**

* **Handling Missing Data:**
  + **Where possible, impute missing values using predefined business rules (can be provided by product owner/ data owner), formula-based calculations, or statistical methods (mean, median).**
  + **If the missing data impacts the quality of reports, consult business stakeholders before taking corrective actions.**
  + **In low-volume cases, it may be preferable to remove records if their absence does not affect business outcomes.**
* **Duplicate Handling:**
  + **Apply DISTINCT queries and partitioning (using the Dim\_Time table) to eliminate duplicate records.**
* **Data Validation:**
  + **Enforce data integrity constraints such as NOT NULL, UNIQUE, FOREIGN KEY to maintain consistency across tables.**
  + **Convert data types as needed to ensure compatibility with the database schema.**

**Load: Sink data to final tables**

* The cleaned and validated data is stored in **final dimension and fact tables** within a **relational database (PostgreSQL/MySQL)**.
* The data is optimized for querying by applying **indexing and partitioning strategies** where applicable.

**Scheduling & Automation**

* **Time-Based Triggers:**
  + **Jobs are scheduled using Apache Airflow or Cron Jobs to execute at predefined times (e.g., every night at 2 AM).**
* **Event-Based Triggers:**
  + **The ETL process automatically triggers when a new file is uploaded to the SharePoint folder by reading the metadata.**

**3️. Standardization & Documentation**

**Data Update Procedures**

* **Dimension Tables: Updated using Upsert Logic, meaning:**
  + **If a record already exists, it is updated.**
  + **If a new record is found, it is inserted.**
* **Fact Tables: Updated using Truncate & Reload Strategy:**
  + **The past three months of data is truncated and reloaded to incorporate changes.**
  + **The specific time window for reloading is determined by business requirements regarding data volatility.**

**Documentation Best Practices**

* **Schema Diagrams:**
  + **A visual representation of all tables and their relationships, highlighting primary and foreign keys.**
* **Workflow Diagrams:**
  + **Step-by-step data flow diagrams to illustrate how data moves from source to final tables.**
* **ETL Pipeline Documentation:**
  + **Detailed explanation of each ETL phase, including dependencies and parallel execution opportunities.**

**4️. Technology Stack Recommendations**

**Database** **Recommended**: **PostgreSQL.**

* **Large data volume processing capacity & indexing for fast queries. It supports JSON, partitioning, and window functions and involves no licensing cost like SQL Server.**

**ETL Tool** **Recommended**: **Apache NiFi.**

* **No-Code UI** – Drag-and-drop functionality similar to **Azure Data Factory (ADF)**. **Supports Real-Time & Batch Processing** and is deal for both streaming and bulk data loads.

**Workflow Orchestration Recommended**: **Apache Airflow**

* **Automated Task Scheduling & Monitoring**. **Extensible & Open-Source** that easily integrates with cloud platforms.

**5️. Security & Compliance**

**Data Encryption**

* **At Rest: Encrypt sensitive data at the table level using built-in encryption techniques such as data masking. (unmasking permission is provided only to specific loader accounts)**
* **In Transit: Ensure secure transmission using SSL/TLS encryption for all database connections.**

**Access Control**

* **Row-Level Security (RLS): Restricts data visibility based on user roles. Only specific rows can be accessed based on the access level.**
  + **Example: Employees from Plant A can only view records related to Plant A.**
* **Column-Level Security (OLS): Controls access to specific attributes based on user roles. Only specific columns can be accessed based on the access level.**
  + **Example: The Sales team can view revenue column data, but HR users cannot.**

**Regulatory Compliance**

* **GDPR (Europe): Implement soft deletes to comply with the right to be forgotten.**
* **ISO 27001: Enforce role-based access control (RBAC) and encrypted backups to protect sensitive information.**
* **SOC 2: Maintain detailed audit logs to track all data modifications.**

**Backup & Disaster Recovery Strategy**

* **Daily incremental backups using a switch mechanism:**
  + **Data is first loaded into a temporary table (fact\_temp), validated, and then switched to the main fact table.**
  + **This ensures that users always access consistent and verified data from the fact table while the fact\_temp is maintained as a backup for recovery.**
* **Weekly full database backups are taken for long-term retention and disaster recovery.**