**Documentation on ETL in Power BI**

**1. Introduction**

ETL stands for **Extract, Transform, Load**, a process used in data integration and preparation for analytics. In **Power BI**, ETL is handled mainly through **Power Query Editor**, where data is cleaned, transformed, and then loaded into the **data model** for reporting.

**2. ETL Workflow in Power BI**

**🔹 1. Extract**

Extracting data means pulling data from multiple sources. Power BI supports connections to:

* Databases (SQL Server, Oracle, MySQL, PostgreSQL)
* Files (Excel, CSV, XML, JSON, PDF)
* Cloud (Azure, Google Big Query, Salesforce, SharePoint)
* APIs & Web Services

👉 Steps:

* Open Power BI → **Home Tab → Get Data** → Choose Source
* Configure authentication and connection mode (Import, Direct Query, Live Connection).

**🔹 2. Transform**

Transformation is the most critical step in ETL. It makes raw data structured, consistent, and usable for analysis. This is performed in **Power Query Editor**.

Common Transformations:

1. **Remove duplicates & nulls** → Ensures clean dataset.
2. **Change data types** → Dates, numbers, text properly formatted.
3. **Split or merge columns** → Example: Full Name → First Name + Last Name.
4. **Standardization** → Consistent formats (e.g., YYYY-MM-DD for dates).
5. **Unpivot/Pivot data** → Reshape for analysis.
6. **Join & Append queries** → Merge tables or stack multiple datasets.
7. **Create calculated columns** → Add derived values (e.g., Profit = Sales – Cost).
8. **Apply business rules** → Categorize or group values.

👉 All steps are recorded in the **Applied Steps Pane**, which acts as built-in documentation.

**🔹 3. Load**

The final stage is loading transformed data into the **Power BI Data Model**.

Options:

* **Import Mode** → Data is loaded into Power BI (faster but larger file size).
* **Direct Query Mode** → Queries are sent directly to the source (real-time, but slower).
* **Live Connection** → Connects directly to a data source like Analysis Services.

👉 After loading, you can create **relationships, measures (DAX), and hierarchies** for reporting.

**3. Documentation of ETL Process**

To ensure reproducibility and transparency, documentation should cover:

1. **Data Sources**
   * Name, type, location, and refresh frequency.
   * Example: "Sales Data from SQL Server (Updated Daily)".
2. **Data Cleaning & Transformation Steps**
   * Detail each step applied in Power Query.
   * Example:
     + Removed duplicates from Customer\_ID.
     + Changed Order Date to Date format.
     + Merged Sales and Products tables on Productid.
3. **Data Model**
   * Show star schema (Fact & Dimension tables).
   * Document relationships (One-to-Many, Many-to-One).
4. **Load Strategy**
   * Mention whether Import/Direct Query/Live Connection.
   * Example: “Sales data is loaded using Direct Query to ensure real-time reporting.”
5. **Version & Change Log**
   * Keep track of modifications made in ETL.

**4. Best Practices**

* Use **meaningful names** for queries and columns.
* Maintain a **Data Dictionary** with field descriptions.
* Separate queries into **Staging (raw data)** and **Final (transformed data)** layers.
* Validate transformed data against original data to ensure accuracy.
* Use **Power Query M-code** as technical documentation.

**5. Example ETL Scenario in Power BI**

**Business Case:** Creating a Sales Performance Dashboard.

* **Extract**: Connect to Sales.xlsx, Customers.csv, Products (SQL Database).
* **Transform**:
  + Removed duplicates from sales table.
  + Standardized date format (YYYY-MM-DD).
  + Merged sales with customer and product details.
  + Created “Profit” column = Sales – Cost.
* **Load**: Imported data into Power BI model (Import Mode).
* **Model**: Created relationships (Fact Sales → Dim Customer, Dim Product).

**6. Tools Supporting ETL in Power BI**

* **Power Query (M Language)** → Data extraction & transformation.
* **Data Model View** → Relationships and schema design.
* **DAX (Data Analysis Expressions)** → Advanced calculations post-load.
* **Dataflows** → Reusable ETL pipelines in the cloud.