HR Data Analysis using Microsoft Fabric (Short Summary)

Objective:

Analyze employee data by applying the Medallion Architecture: **Staging → Bronze → Silver → Gold**, and visualize insights using **Power BI**.

Tools Used:

- Microsoft Fabric
- PySpark (Notebooks)
- Dataflow
- Power BI

Data Source:

HR_DATA.csv with employee details like name, gender, department, dates, etc.

Key Steps:

1. **Staging Layer:**
   Raw file uploaded to OneLake; no changes applied.

2. **Bronze Layer:**
   Structured table created by promoting headers and validating data types.

3. **Silver Layer Processing:**
   - Cleaned and converted columns (e.g., removed dashes from ID, converted date strings to date type).
   - Renamed cleaned ID to employee_id.
   - Stored result as a Delta Table.

4. **Gold Layer Tables:**

- **DimEmployee:** Selected columns, renamed fields, and added full_name.
- **DimDepartment:** Listed distinct departments with generated department_id.
- **FactEmployee:** Grouped by department & location to compute:
  - Total employees
  - Average age
  - Gender distribution
  - Turnover rate
  - Assigned surrogate

- fact_id: Unique identifier for each row in the fact table.
- employee_id: Synthetic identifier to simulate employee linkage (for dimensional modeling).
- Rearranges columns into a clean and logical order for storage and querying.

- Displays the final transformed DataFrame containing employee statistics for review.

Output Schema

Column Name	Description
fact_id	Unique ID for each fact row
employee_id	Synthetic ID representing employee entity
location_id	Foreign key referencing location dimension
department_id	Foreign key referencing department dimension
total_employees	Count of employees in the department-location group
avg_age	Average age of employees
gender_distribution	Gender ratio in format: "Male: x%, Female: y%, Other: z%"
turnover_rate	Percentage of employees who left the organization