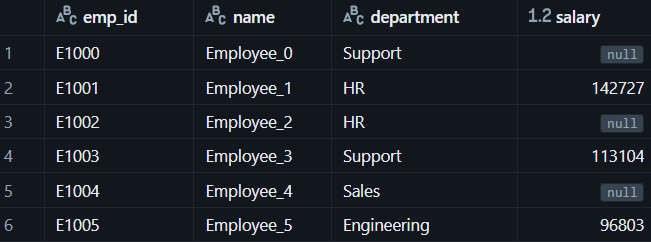
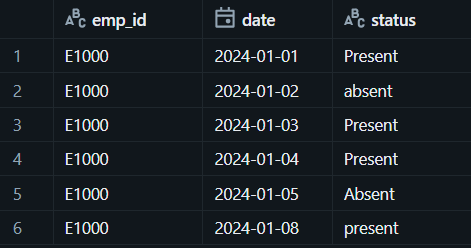
Project Description

This project involves analyzing employee attendance and salary data using PySpark on the Databricks platform. The objective is to clean, transform, and analyze the data to derive attendance-based salary computations

# Input Files

* **employees\_big.csv**: Contains basic employee details.
  + employee\_id: Unique identifier
  + name: Employee name
  + department: Department name
  + salary: Monthly salary (can have missing values)



* **attendance\_big.csv**: Daily attendance records.
  + employee\_id: Matches with employee table
  + attendance\_date: Date of attendance
  + status: Either 'Present', 'Absent', or NULL

# Tools & Technologies Used

| Tool/Platform | Purpose |
| --- | --- |
| Apache Spark (PySpark) | Distributed data processing |
| Databricks | Unified analytics platform for Spark |
| Python | Programming language |
| Jupyter Notebook (Databricks) | Coding and visualization |
| MS Word / PDF | Documentation |

# Data Cleaning

**✅ Salary Imputation**

Missing salaries are filled using **department-wise median** using percentile\_approx(salary, 0.5).

**✅ Attendance Status Cleaning**

Trimmed whitespaces.

Normalized variations ("P", "present", "Present") to "present".

Converted "A", "Absent", "NA", or NULL to "absent" using when().otherwise().

**✅ Date Formatting**

Converted attendance\_date to DateType using to\_date().

Extracted month in "yyyy-MM" format using date\_format().

# **🔄 Data Transformation**

| Transformation | Description |
| --- | --- |
| **Month Extraction** | Extracted month from attendance\_date in yyyy-MM format |
| **Present Days Count** | Counted "present" entries per employee\_id and month |
| **Total Attendance Days** | Counted all entries (present/absent) per employee\_id and month |
| **Attendance Percentage** | Calculated as: (attended\_days / total\_days) \* 100 |
| **Effective Salary** | Salary adjusted to presence: (salary / total\_days) \* attended\_days |
| **Join with Employee Data** | Merged salary & department info into attendance metrics |

# Final Output Columns

| Column |  | Description |
| --- | --- | --- |
| employee\_id |  | Unique ID of the employee |
| department |  | Employee's department |
| month |  | Month in yyyy-MM format |
| attended\_days |  | Number of days marked "present" |
| total\_days |  | Total days recorded |
| attendance\_pct |  | Percentage of presence |
| salary |  | Monthly salary (imputed if missing) |
| effective\_salary |  | Final adjusted salary based on presence |

**Client needs effective average salary each department in given month of the year**

****