HR Attrition Analytics ETL Pipeline (Bronze → Silver → Gold)

Dataset Link

IBM HR Analytics – Attrition & Performance

Design and implement a complete **end-to-end ETL pipeline** using the IBM HR Analytics dataset.
Your goal is to help the HR Insights team understand **employee attrition patterns**, identify **high-risk employee groups**, and build **clean, analytics-ready data** in the Gold layer.
You will work on Databricks using **Delta Lake**, following industry best practices.

Project Overview
The HR department has provided employee-level data containing demographics, job roles, salaries, performance ratings, and attrition status.
Your task is to build a 3-layer ETL pipeline:

Bronze Layer - Raw Data Ingestion
Requirements:

- Ingest the raw CSV file **as-is** into the Bronze layer.
- Store using **Delta format**.
- Add ingestion metadata:
  - _source_file_name
  - _ingestion_timestamp
- No cleaning or type casting allowed here.

Output:
bronze_hr_employees

Silver Layer - Data Cleaning + Standardization
Perform data cleaning and transformations:

✓ Data Quality Fixes

- Convert categorical columns (e.g., *Attrition*, *Gender*, *JobRole*) to lowercase standardized values.
- Replace null or blank fields with appropriate defaults.
- Validate numeric columns (e.g., *Age*, *MonthlyIncome*, *PercentSalaryHike*).
- Filter unrealistic entries (e.g., age < 18 or > 70).

✓ Data Type Conversions

- Cast numeric columns to correct types.
- Convert "Attrition" into boolean (1 = Yes, 0 = No).

✓ Feature Engineering
Add additional useful fields:

- age_group (18–25, 26–35, 36–45, 46+)
- income_bucket
- years_at_company_bucket
- attrition_flag (1/0)

✓ Standardization

- Convert column names to **snake_case**.

Output:
silver_hr_cleaned

Gold Layer - Analytics Tables
Create 2–3 business-ready aggregated tables that HR can directly use for dashboards.

Gold Table 1 - Attrition Summary by Department
Columns:

- department
- total_employees
- employees_left
- attrition_rate
- avg_monthly_income
- avg_years_at_company

Goal: Identify high-risk departments.

Gold Table 2 - Attrition by Age Group & Job Role
Columns:

- age_group
- job_role
- total_employees
- attrition_count
- attrition_rate

Goal: Understand whether younger or older employees leave more.

Gold Table 3 - Work-Life Balance & Attrition (Optional)
Columns:

- work_life_balance
- avg_over_time_hours
- attrition_rate
- avg_job_satisfaction

Goal: Show impact of work conditions on attrition.

Expected Deliverables
✓ Databricks Notebook
Must include:

- Bronze ingestion
- Silver cleaning + standardization
- Gold analytics tables
- display() outputs

✓ Pipeline Diagram
Bronze → Silver → Gold architecture.

Final Submission Checklist

✅ Notebook: Clean, well-commented, and readable
✅ Dataset: Cleaned & saved as CSV

✅ PPT: Insightful, with visuals and summary points
✅ Zip Folder: Contains all files