1. About your project.

Currently, I’m working on a project called Certification Factory (CFY). The main goal of this project is to help Société Générale (SG) meet CRR3 or Basel IV regulations, which are important for submitting regulatory reports for all SG entities to the European Central Bank (ECB) as part of their solvency transformation.

To meet these requirements, we developed the CFY project, which is still in the development phase and not yet in production.

Right now, all the data from SG’s entities and subsidiaries is stored in a Teradata Vantage data warehouse, which is a legacy system. This serves as the source for the CFY pipeline. Our work begins once the data is available in this Teradata system. From there, we extract the data and process it through the Certification Factory.

So, if we breakdown the main tasks this project is doing is like:

Data Extraction:

We use an internally developed API, called the Bridge, which is built using Java Spring Boot. The Bridge API has a few key responsibilities:

* Extracting data from Teradata using its Tdload feature, which outputs the data as CSV files.
* Converting these CSV files into Parquet format for better compatibility and efficiency in Databricks.
* Uploading these Parquet files to an S3 bucket (we call this the "landing zone").
* Generating a manifest.yml file, which contains metadata about the extracted data.

Triggering the Lambda Function

Once the data is uploaded to the S3 landing zone, S3 event notification triggers an AWS Lambda function. This function has two main tasks:

Firstly,

* It validates the uploaded manifest.yml file to ensure the data is correct and complete.
* If everything checks out, it moves the datasets to a "raw zone" in the destination S3 bucket (a bucket specifically configured for Databricks access).
* It also registers the dataset with an internal tool called Control Center, which orchestrates the entire project and keeps track of metadata.

Second,

* Once the files are in the destination databricks bucket, it updates the dataset’s status in Control Center and loads the data into raw tables in Databricks environment for further use.

Data Processing in Databricks:

After the data is available in Databricks, a Databricks job will run automatically to apply controls and defaults. This step enriches the data, making it ready for various calculations.

Performing Regulatory Calculations:

The enriched data is then used to perform important regulatory calculations, such as:

* RWA (Risk-Weighted Assets): To measure credit risk.
* LE (Large Exposures): To monitor exposure limits.
* ECL (Expected Credit Loss): To estimate potential losses.

These calculations are based on the requirements provided by Business Analysts (BAs). Depending on the need, we execute specific “runs” to generate the required outputs.

Sending Data to Downstream Systems:

Once the calculations are complete, the data is sent to downstream systems for further processing or reporting.

This project will significantly reduce SG's compliance risks and modernize their regulatory reporting process.