Hackathon/ Group Activity

Each team will work on the following Activity:

Objective:

Design, implement, and automate an ETL pipeline using Databricks to process the "Superstore" dataset through the Bronze, Silver, and Gold Delta Lake layers. Your team will develop the pipeline, ensuring data quality, performing predictive modeling, and aggregating insights.

Tasks:

1. **Bronze Layer:**
    - Ingest the raw data from Superstore.xls into Delta format.
    - Implement basic transformations such as partitioning the dataset by relevant columns (e.g., Order Date or Region**).**
2. **Silver Layer:**
    - Clean and transform the data for quality and consistency.
    - Apply data quality techniques like handling missing values, removing duplicates, correcting data types, and standardizing formats.
    - Ensure data is ready for aggregation and analysis.
3. **Gold Layer:**
    - Perform aggregations for reporting and analysis (e.g., total sales by region, product category, or customer).
    - Analyze key metrics such as sales performance, profit margins, and product demand.
4. **Automation:**
    - Automate the ETL pipeline using Databricks Workflows, ensuring seamless execution of all layers.
5. **Submission:**

    Please share all the notebooks and screenshots of your entire workflows, pipeline runs as attachments to shatanu.pandey@rpsconsutling.in with your **Team members on CC**