**Grocery Store Sales Forecasting ETL Pipeline**

**Project Title:**  
Grocery Store Sales Forecasting ETL Pipeline

**Project Description:**  
Ingest daily grocery sales data from CSV files stored in AWS S3, perform necessary cleaning and transformations, and load into Delta Lake tables in Unity Catalog for analytics and forecasting. Common elements include daily batch processing, error handling, data quality checks, and Git integration for version control.

* **Source**: CSV files from the Kaggle dataset “Store Sales - Time Series Forecasting” (Corporación Favorita) in s3://bucket/raw/grocery\_sales/YYYY/MM/DD/, containing store-level sales, items, dates, etc.
* **Destination**:
  + **Bronze**: grocery\_catalog.raw.sales\_transactions (raw data).
  + **Silver**: grocery\_catalog.processed.sales\_cleaned (cleaned and standardized data).
  + **Gold**: grocery\_catalog.analytics.sales\_forecast\_features (aggregated features for forecasting).
* **Transformations**: Parse CSVs, deduplicate, standardize formats, join with dimension tables (e.g., stores, items), compute aggregates (e.g., weekly sales), enrich with external factors (e.g., holidays), etc.
* **Schedule**: Daily batch job at 4 AM UTC using Databricks Workflows.
* **Alerts**: Notify via email for significant data anomalies, log errors to grocery\_catalog.logs.etl\_errors, etc.
* **Technical Specifications**:
  + **Databricks Workspace**: Standard cluster, autoscaling (4–16 workers), m5.large nodes, libraries like pyspark, delta-spark, boto3, etc.
  + **Pipeline Tasks**:
    1. **Data Ingestion**: Read CSVs from S3, infer schema, handle malformed records, partition by year/month, etc.
    2. **Data Cleansing & Transformation**: Deduplicate, standardize dates, join with dimensions, apply custom logic (e.g., holiday flags), etc.
    3. **Data Write**: Write to Bronze, upsert to Silver, create aggregated Gold tables, optimize and vacuum periodically, etc.
    4. **Error Handling**: Log errors, set up alerts for data quality issues, etc.
    5. **Orchestration**: Use Databricks Workflows, integrate with AWS CodeCommit, etc.
  + **Testing**: Unit tests for key logic, validate counts and schemas, test with Kaggle sample data, QA sign-off, etc.

**Objective:**  
Process grocery sales data to enable accurate forecasting for inventory and promotions.

**Tools Used:**  
Databricks, PySpark, Delta Lake, AWS S3, boto3, pytest, etc.

**Outcomes (Aligned with Data Engineering Standards):**  
- **Scalable Data Pipeline**: A robust ETL pipeline meeting industry standards for scalability using Delta Lake and cloud storage (AWS S3).  
- **Data Quality Assurance**: Implements error logging and anomaly detection, ensuring reliable data for downstream analytics.  
- **Performance Optimization**: Leverages partitioning and optimization techniques (e.g., Delta table optimization) for efficient processing.  
- **Version Control and Collaboration**: Git integration ensures reproducible and collaborative development.  
- **Business Impact**: Enables accurate forecasting, reducing stockouts and optimizing inventory costs.