Project Statement: Generative AI Data Processing in Databricks

Duration: 8 Hours

Objective: Transform raw text data into structured insights using the Bronze, Silver, and Gold layer architecture in Databricks, leveraging GPT-2-based Generative AI models for enrichment, summarization, and analytics.

Project Overview

Participants will ingest, clean, enrich, and analyze text data using Databricks Delta Lake and Generative AI (GPT-2). The workflow will progress through the Bronze, Silver, and Gold layers, demonstrating AI-powered text processing and structured transformations.

Key Activities

1. Bronze Layer (Raw text ingestion and storage).

2. Silver Layer (Preprocessing, cleaning, and AI-driven text augmentation).

3. Gold Layer (Summarization, embeddings, and structured analytics).

4. Model Training & Fine-tuning (Optional for advanced teams).

5. Visualization & Reporting (Querying structured insights).

Project Breakdown (8 Hours)

Phase 1: Bronze Layer – Data Ingestion & Storage (1.5 Hours)

- Load raw text files (TXT, JSON, CSV, or scraped data) into Databricks Delta Lake.

- Store data in the Bronze Table without modifications.

- Log metadata, file schema, and perform basic validation.

- Apply basic data quality checks (e.g., missing values, schema validation).

- Create a data ingestion pipeline using Databricks Auto Loader or PySpark.

Sample Starter Code: Bronze Layer Ingestion

Disclaimer: This is a sample code snippet for reference only. It may require adjustments based on specific project requirements and dataset formats.

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import input_file_name

# Initialize Spark Session
```

```
spark = SparkSession.builder.appName("BronzeLayerIngestion").getOrCreate()

# Define Source Path (Update as needed)
source_path = "dbfs:/mnt/raw_text_files/"

# Read Raw Text Files
df_bronze = (spark.read.text(source_path)
          .withColumn("source_file", input_file_name())  # Capture source metadata
         )

# Save as Delta Table (Bronze Layer)
df_bronze.write.format("delta").mode("overwrite").saveAsTable("bronze_text_data")

# Display Sample Data
df_bronze.show(5)
```

Deliverables

- Bronze Table storing raw text.
- Metadata logs for tracking ingestion.

Phase 2: Silver Layer – Data Cleansing & AI Enrichment (2 Hours)

- Perform text preprocessing such as lowercasing, punctuation removal, tokenization.
- Handle stopwords, special characters, and formatting inconsistencies.
- Apply GPT-2 for text augmentation including missing text generation, readability improvement, and text expansion.
- Store cleaned and AI-enhanced data in the Silver Table.
- Track data lineage to show transformations.

Phase 3: Gold Layer – Summarization & AI-Driven Insights (2.5 Hours)

- Apply GPT-2-based summarization to generate key insights.
- Perform topic modeling (LDA, BERT embeddings) for text classification.
- Conduct sentiment analysis.
- Generate text embeddings for AI-powered search.
- Store final structured dataset in the Gold Table.

Phase 4: Model Training & Fine-Tuning (1.5 Hours)

- Fine-tune GPT-2 on custom.

- Train a simple text classification model (e.g., classifying reviews, topics).

- Compare pre-trained vs fine-tuned model outputs.

- Save trained models in MLflow for tracking.

Phase 5: Visualization, Queries & Reporting (1.5 Hours)

- Query structured insights using Databricks SQL & PySpark.

- Build a basic visualization dashboard using Databricks Notebooks & Plotly.

- Validate data lineage from Bronze → Silver → Gold.

- Discuss real-world applications of this pipeline (e.g., news summarization, chatbot training).

Expected Final Deliverables

All of the below should be uploaded in the group's github repository.

1. Databricks Notebooks for each transformation phase.

2. Bronze, Silver, and Gold Delta Tables with respective data stages (screenshots).

3. AI-enriched, structured insights stored in Gold Layer (screenshots).

4. Basic ML model & embeddings (print the model and embeddings in a file)

5. Final queries, dashboards, and visualizations.