Capstone Project Proposal – Databricks Lakehouse Implementation

Project Title:
Building a Scalable Data Lakehouse with Databricks for Unified Batch and Streaming Data Processing

Summary:
This project will focus on designing and implementing modern data Lakehouse architecture using Azure Data bricks and Delta Lake. The aim will be to demonstrate how raw data can be ingested, cleansed, transformed, and modeled across bronze, silver, and gold layers while ensuring governance, scalability, and performance optimization. Key aspects of the project will include:

- Data Ingestion: I will leverage Auto Loader to continuously load data from cloud storage into the bronze layer with schema evolution support.
- Data Transformation: I will apply Structured Streaming and Delta Live Tables (DLT) to implement Slowly Changing Dimensions (SCD Type 1 and 2) in the silver and gold layers for accurate historical tracking.
- Comparative Study: A critical component of this project will be the comparison of SCD Type 2 implementation using Spark Structured Streaming vs Delta Live Tables (DLT). The evaluation will cover complexity, maintainability, latency, cost, and pipeline governance, providing insights into when organizations should choose one approach over the other.
- Governance & Security: Access and lineage will be managed using Unity Catalog, external volumes, and storage credentials.
- Optimization: I will apply Z-ORDER, OPTIMIZE, and partitioning strategies to improve query performance and reduce cost.
- Real-Time vs Batch Processing: I will also compare streaming ingestion with batch pipelines to evaluate trade-offs in latency, cost, and complexity.

Expected Outcomes:

- A fully functional end-to-end data pipeline on Databricks demonstrates ingestion, transformation, and consumption.
- A comparative analysis of SCD Type 2 pipelines implemented with Spark Structured Streaming vs Delta Live Tables (DLT).
- Performance benchmarks that will highlight improvements through Databricks optimizations.

Business Value:

This project will mirror a real-world migration from traditional data warehouses to a cloud-native Lakehouse. It will not only showcase best practices for handling streaming + batch workloads but also provide practical guidance on SCD2 implementation choices between Spark Streaming and Databricks' declarative DLT pipelines. The outcomes will help organizations modernize their data platforms, reduce manual overhead, and improve decision-making speed.