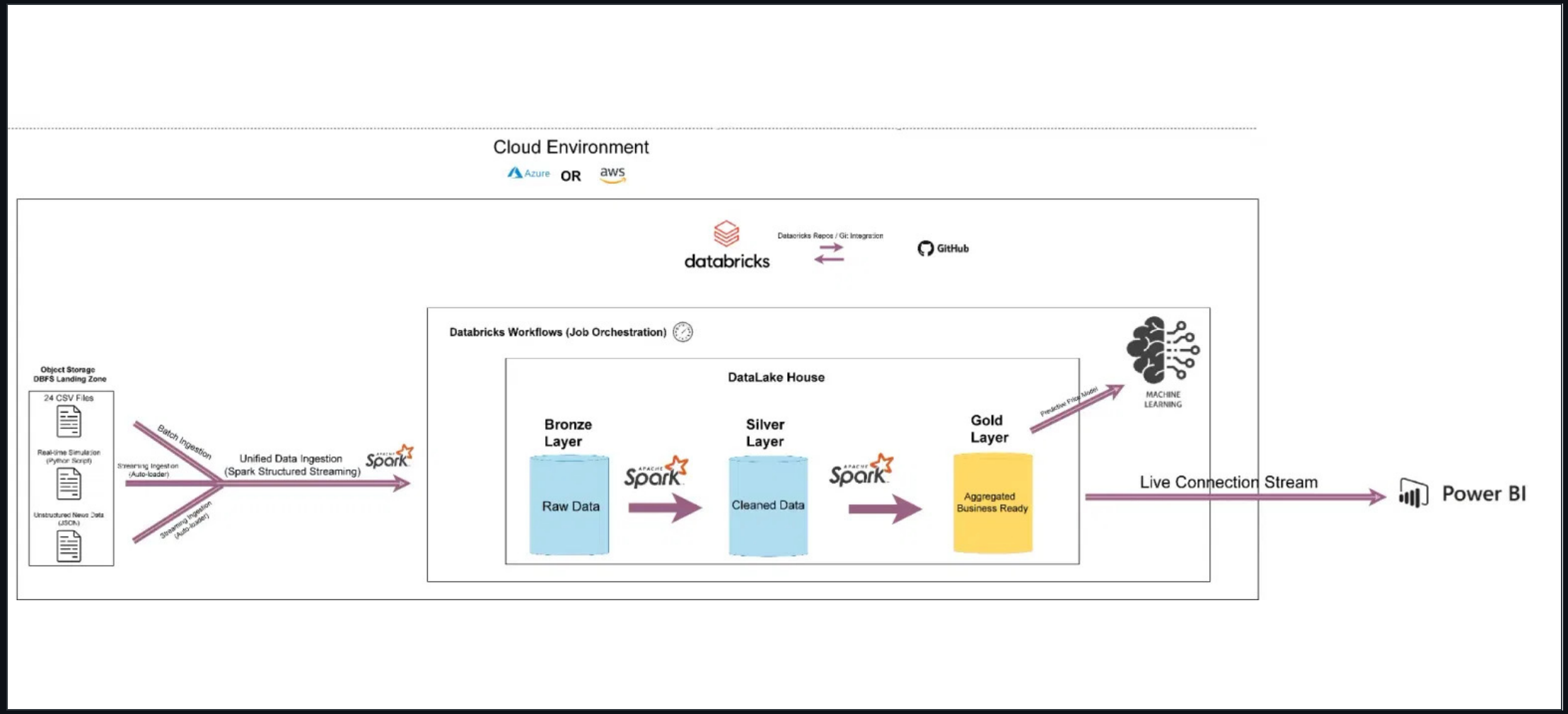
UNIFIED CRYPTO-INTELLIGENCE LAKEHOUSE

Building Real-Time Insights & Predictive Analytics with Databricks and Apache Spark

// PROJECT DESCRIPTION

This project aims to build an **end-to-end data engineering pipeline** to analyze the cryptocurrency market (Bitcoin) in real-time. The system integrates multiple data streams, including **historical price data (Batch)**, **live price updates (Streaming)**, and **unstructured news data (JSON)**. By leveraging the **Medallion Architecture**, the project transforms raw, chaotic data into structured insights. A key highlight is the integration of **Natural Language Processing (NLP)** to perform sentiment analysis on market news, combined with **Machine Learning** to predict price trends based on both financial metrics and public sentiment.

// PROJECT ARCHITECTURE



UNIFIED REAL-TIME DATA ENGINEERING PIPELINE

// MEDALLION ARCHITECTURE

DATA_TIER_01

BRONZE

The "Source of Truth" layer. Data is stored in its original format to allow for full lineage and reprocessing.

- Raw Immutable History
- Ingestion Metadata

DATA_TIER_02

SILVER

The "Engine Room." Data is cleaned, validated, and enriched with NLP sentiment scores for analysis.

- Deduplication & Cleaning
- NLP Sentiment Scoring

DATA_TIER_03

GOLD

The "Analytics Layer." Curated, aggregated datasets optimized for Power BI and Machine Learning.

- Stream-Batch Joins
- Business-Ready Views

// CORE TECHNOLOGY STACK

CLOUD PLATFORM

DATABRICKS

Unified analytics platform for data processing, machine learning, and collaborative engineering.

DATA ENGINE

APACHE SPARK

Distributed engine for high-speed batch and stream processing using PySpark and SQL.

STORAGE LAYER

DELTA LAKE

Providing ACID transactions and scalable metadata handling for the unified lakehouse.

STREAMING

SPARK STREAMING

Real-time data ingestion using the Auto-loader feature for low-latency updates.

AI & NLP

MLFLOW / VADER

Managing ML experiments and performing sentiment analysis on unstructured news data.

VERSION CONTROL

GITHUB

CI/CD integration and code repository management for robust engineering workflows.

ORCHESTRATION

WORKFLOWS

Automating pipeline jobs to ensure 24/7 data flow and operational reliability.

VISUALIZATION

POWER BI

Live-connected dashboards for real-time market monitoring and stakeholder insights.

// DETAILED PROJECT WORKFLOW & ENGINEERING LOGIC

MODULE_01

UNIFIED DATA INGESTION (THE MULTI-MODAL ENTRY)

Our pipeline is designed to be **Source-Agnostic** with a hybrid ingestion strategy:

BATCH INGESTION

Processing 24 high-resolution CSV files containing historical Bitcoin market data.

STREAM INGESTION

Utilizing **Databricks Auto-loader** and "Cloud Files" for event-driven, incremental processing.

</> UNSTRUCTURED DATA

Ingesting JSON feeds for market news, demonstrating non-tabular data handling.

MODULE_02

THE MEDALLION ARCHITECTURE (DATA GOVERNANCE)

A three-tier storage strategy using **Delta Lake** for reliability and consistency:

BRONZE (RAW)

The "Source of Truth" with ingestion timestamps for lineage and reprocessing.

SILVER (ENRICHED)

Cleaning duplicates and performing **NLP Sentiment Scoring** (-1 to +1) on news text.

GOLD (ANALYTICS)

Performing a **Stream-Batch Join** to align sentiment with price movements.

MODULE_03

MACHINE LEARNING & PREDICTIVE INTELLIGENCE

Beyond ETL, integrating a professional predictive component:

FEATURE ENGINEERING

Creating Moving Averages (MA) and Sentiment Indices from Gold tables.

MODEL TRACKING

Using **MLflow** for experiment tracking, versioning, and MLOps workflows.

REAL-TIME INFERENCE

Predicting next-hour trends with "Buy/Sell/Hold" signals from data fusion.

// ROBUST DATA FOUNDATION

PRIMARY DATA SOURCE

KAGGLE DATASET

Cryptocurrency Historical Prices by Sudalairajkumar

DATA FIDELITY

High-resolution historical price data (Open, High, Low, Close, Volume) for BTC, ETH, and other major assets.

kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory

CONTINUOUS INGESTION

DATABRICKS AUTO-LOADER

Efficiently processes new files as they arrive in the landing zone without manual intervention.

SCHEMA EVOLUTION

Automatically handles changes in data structure to ensure pipeline stability and reliability.

DATA INTEGRITY

Ensuring high-fidelity data for robust model training, validation, and real-time simulation.

// TEAM MEMBERS

PROJECT TEAM

- Youssef Hamed Abdelmonim Ahmed
  TEAM LEADER
- Seif Mohamed Fathi Abdelaziz
- Yehia Yasser Yahya Salama
- Nada Mahmoud Hammad Ibrahim
- Nourhan Ahmed Gaber Sharawy