Project Planning

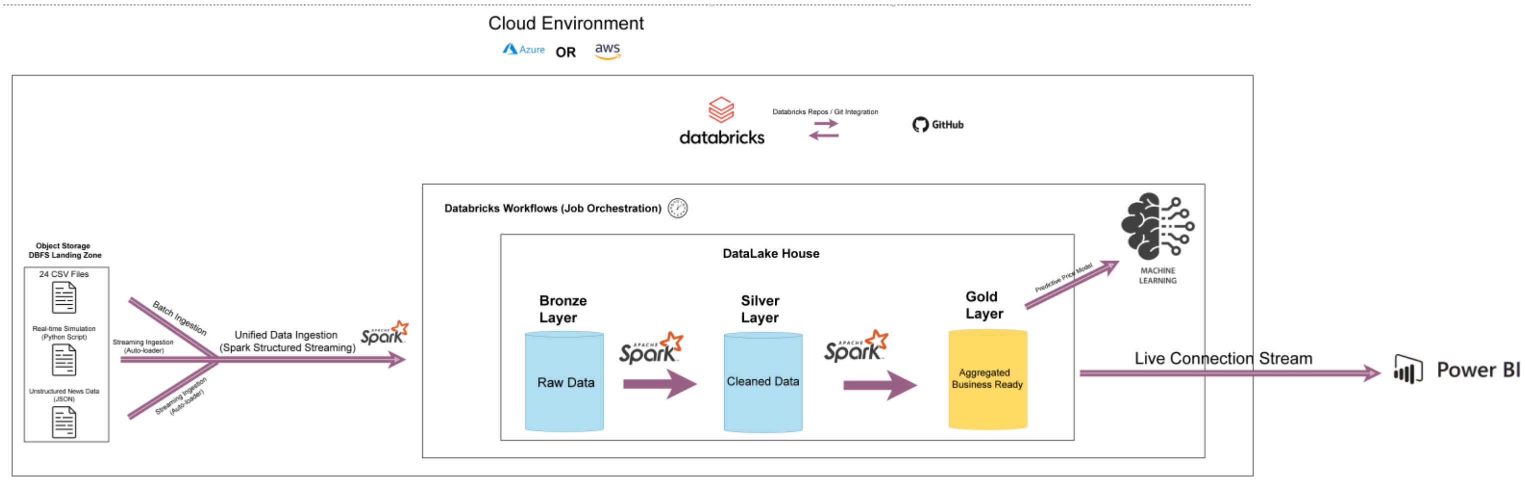
Name: Unified Crypto-Intelligence Lakehouse

1. Project Overview

Title: Building a Unified Real-Time Crypto-Intelligence Lakehouse on Databricks with Predictive Analytics.

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.

2. Project Design



Architectural Note: The design follows the "Decoupling Storage from Compute" principle, using a centralized Data Lakehouse approach to manage the entire data lifecycle from ingestion to visualization and Prediction

3. Technology Stack

Category	Technology Used	Purpose
Cloud Platform	**Databricks** (Community Edition)	The core unified analytics platform for data processing and ML.
Data Engine	**Apache Spark** (PySpark & SQL)	The distributed engine for high-speed batch and stream processing.
Storage Layer	**Delta Lake**	To provide ACID transactions and scalable metadata handling for the Lakehouse.
Streaming	**Spark Structured Streaming**	To handle real-time data ingestion using the **Auto-loader** feature.
AI & NLP	**MLflow & TextBlob/VADER**	For managing ML experiments and performing Sentiment Analysis on news.
Version Control	**GitHub**	For CI/CD integration and code repository management.
Orchestration	**Databricks Workflows**	To automate the pipeline jobs and ensure 24/7 data flow.
Visualization	**Power BI**	To build a live-connected dashboard for real-time market monitoring.

Detailed Project Workflow & Engineering Logic

1. Unified Data Ingestion (The Multi-Modal Entry)

Our pipeline is designed to be **Source-Agnostic**. We implement a hybrid ingestion strategy:

- **Batch Ingestion:** Processing 24 high-resolution CSV files containing historical Bitcoin market data (Open, High, Low, Close, Volume).
- **Stream Ingestion:** Utilizing **Databricks Auto-loader** to provide an event-driven ingestion mechanism. It uses "Cloud Files" to incrementally process new incoming data from the landing zone without manual intervention.
- **Unstructured Data Handling:** Ingesting JSON feeds containing market news and social sentiment, demonstrating the **Lakehouse** ability to handle non-tabular data alongside traditional metrics.

2. The Medallion Architecture (Data Governance)

We implement a three-tier storage strategy using **Delta Lake** to ensure data reliability and consistency:

- **Bronze (Raw Layer):** Acts as the "Source of Truth." Data is stored in its original format with added metadata (ingestion timestamp) to allow for data lineage and reprocessing if needed.
- **Silver (Validated & Enriched Layer):** This is the engine room of the project. We perform:
  - **Data Cleaning:** Removing duplicates and handling null values in price data.
  - **NLP Sentiment Scoring:** Integrating a Natural Language Processing model (TextBlob/VADER) to analyze the `news_text` column. It converts qualitative news into a quantitative `sentiment_score` ranging from -1 (Bearish) to +1 (Bullish).
- **Gold (Analytics Layer):** The final curated layer. Here, we perform a **Stream-Batch Join**, aligning the sentiment scores with price movements based on time-windowing. This creates a high-value dataset ready for Power BI and Machine Learning.

3. Machine Learning & Predictive Intelligence

Beyond simple ETL, the project includes a predictive component:

- **Feature Engineering:** Using the `Gold_Table` to create features like Moving Averages (MA) and the newly generated `Sentiment_Index`.
- **Model Tracking:** Using **MLflow** to track experiments, model versions, and hyperparameters, ensuring a professional MLOps workflow.
- **Inference:** The model predicts the next-hour price trend, providing a "Buy/Sell/Hold" signal based on the fusion of market data and news sentiment.

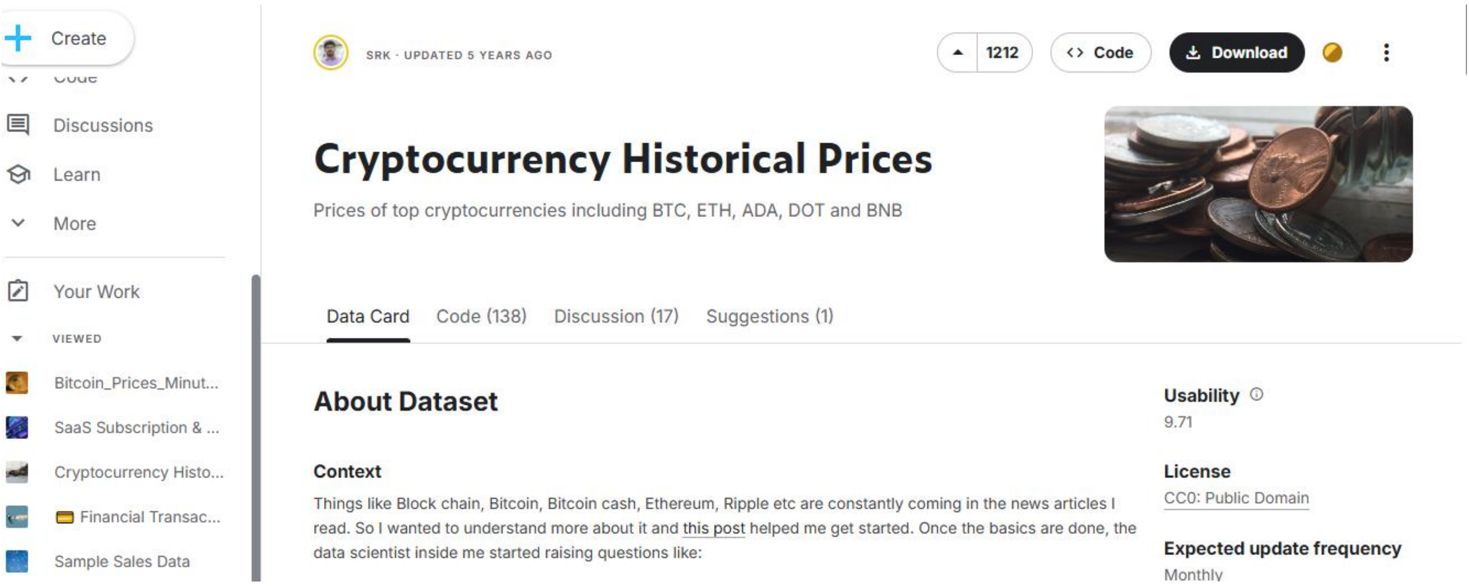
4. Real-Time Visualization (The Business Value)

The final output is served through a **Power BI Dashboard** connected via **Databricks SQL Warehouse**. This allows stakeholders to:

- Monitor live Bitcoin price fluctuations.
- Visualize the correlation between "Market Hype" (Sentiment) and "Price Action."
- Track the accuracy of the Machine Learning predictions in real-time.

4. Data Source

The project utilizes the **Sudalairajkumar Cryptocurrency Dataset** from Kaggle, ensuring high-fidelity historical price data for robust model training and simulation



Cryptocurrency Historical Prices

Prices of top cryptocurrencies including BTC, ETH, ADA, DOT and BNB

About Dataset

Context

Things like Block chain, Bitcoin, Bitcoin cash, Ethereum, Ripple etc are constantly coming in the news articles I read. So I wanted to understand more about it and this post helped me get started. Once the basics are done, the data scientist inside me started raising questions like:

Usability
9.71

License
CC0: Public Domain

Expected update frequency
Monthly

URL : https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory

5. Team Members

- Youssef Hamed Abdelmonim Ahmed (Team Leader)

- Seif Mohamed Fathi Abdelaziz

- Yehia Yasser Yahya Salama

- Nada Mahmoud Hammad Ibrahim

- Nourhan Ahmed Gaber Sharawy