Build an Incremental ETL Pipeline with AWS CDK

Business Overview

Cryptocurrency refers to digital or virtual currencies that use cryptography for secure financial transactions, control the creation of additional units, and verify the transfer of assets. Cryptocurrencies leverage decentralized technology called blockchain, which is a distributed ledger maintained by a network of computers.

The most well-known and widely used cryptocurrency is Bitcoin, which was introduced in 2009. Bitcoin was the first decentralized cryptocurrency and remains the largest by market capitalization. Since the creation of Bitcoin, thousands of other cryptocurrencies, often referred to as altcoins (alternative coins), have been developed.

Cryptocurrency data analytics refers to the analysis and interpretation of data related to cryptocurrencies and their markets. With the increasing popularity and complexity of the cryptocurrency market, data analytics plays a crucial role in understanding trends, making informed decisions, and identifying opportunities within the crypto space.

Here are some key aspects of cryptocurrency data analytics:

- Market Data Analysis: This involves analyzing historical and real-time market data, including price movements, trading volumes, liquidity, market capitalization, and order book data. By examining these data points, analysts can identify patterns, trends, and market sentiment to gain insights into the market's behavior.
- Blockchain Analysis: Since most cryptocurrencies operate on blockchain technology, blockchain analysis is essential for understanding transaction flows, addresses, and network behavior. It can help identify key addresses, track fund movements, and detect anomalies or suspicious activities.
- Sentiment Analysis: Sentiment analysis involves examining social media posts, news articles, and other textual data to gauge the sentiment and public perception surrounding specific cryptocurrencies. This analysis helps understand market sentiment, public opinion, and the impact of news events on cryptocurrency prices.
- Trading Strategies and Predictive Modeling: Advanced data analytics techniques, such as machine learning and predictive modeling, can be applied to cryptocurrency data to develop trading strategies and forecasting models. These models aim to predict future price movements, identify trading opportunities, and manage risk.
- Risk Assessment and Security: Data analytics can help assess and quantify the risks associated with cryptocurrencies, such as market volatility, liquidity risk, and cybersecurity vulnerabilities. It enables the development of risk management strategies and the identification of potential threats.

Cryptocurrency data analytics can be performed using a variety of tools and platforms that provide access to historical and real-time data, as well as specialized analytics capabilities. These tools often include charting platforms, data visualization tools, sentiment analysis tools, and machine learning libraries.

We aim to develop an incremental Extract, Transform, Load (ETL) solution utilizing AWS CDK to analyze cryptocurrency data. This will involve constructing a serverless pipeline in which lambda functions are utilized to retrieve data from an API and stream it into Kinesis streams. Additionally, we will create another lambda function to consume the data from the Kinesis stream, apply necessary transformations, and store it in DynamoDB.

To perform data analytics on the incoming data within the Kinesis streams, we will leverage Apache Flink and Apache Zeppelin. These tools will enable us to extract insights and derive valuable information from the data. AWS serverless technologies, such as Amazon Lambda and Amazon Glue, will be employed to efficiently process and transform the data from the three different data sources.

Furthermore, we will utilize Amazon Athena, a query service, to analyze the transformed data stored in DynamoDB. This will facilitate efficient querying and exploration of the data, enabling us to extract meaningful insights and make informed decisions based on the cryptocurrency data.

By combining these AWS services and technologies, we aim to create a robust and scalable solution for analyzing cryptocurrency data, allowing for comprehensive data processing, transformation, and analytics.

Dataset Description

Alpha Vantage offers enterprise-grade financial market data via a collection of robust and developer-friendly data APIs and spreadsheets. Alpha Vantage is your one-stop shop for real-time and historical global market data delivered through REST stock APIs, Excel, and Google Sheets, ranging from traditional asset classes (e.g., stocks, ETFs, commodities) to economic metrics, foreign exchange rates to cryptocurrencies, fundamental data to technical indicators.

Tech Stack

➔ Language: Python

➔ Services: AWS S3, Amazon Lambda, Amazon Aurora, AWS Glue, Amazon Athena, Quicksight, AWS CDK

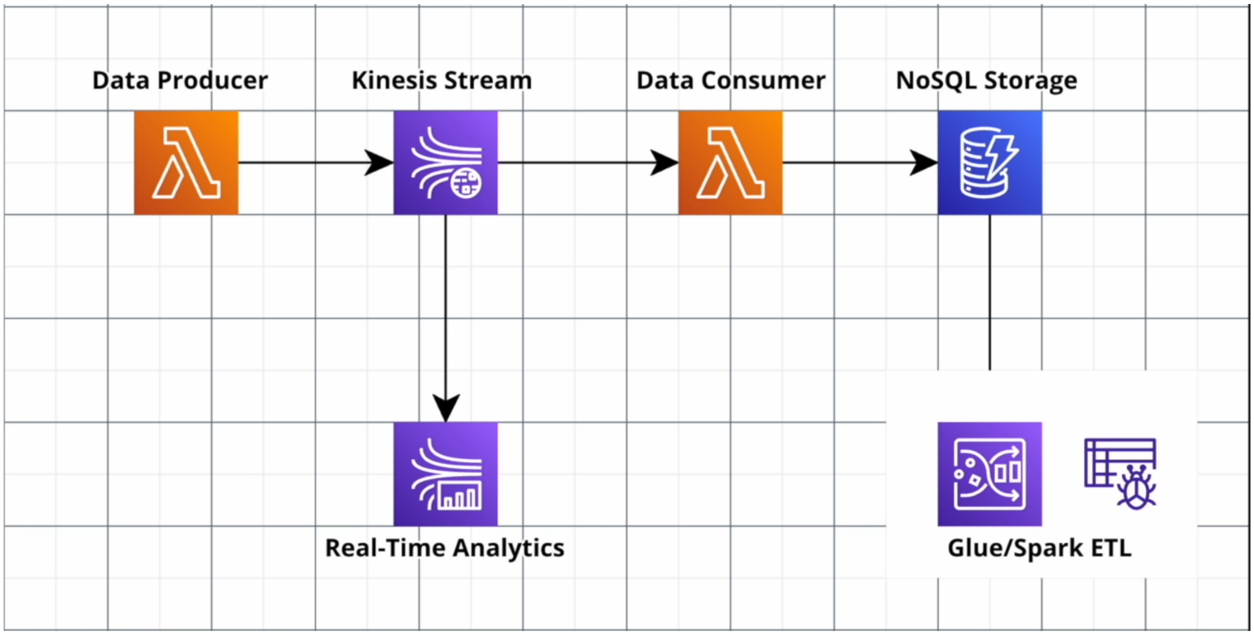
AWS CDK:

The AWS Cloud Development Kit (AWS CDK) is an open-source software development platform for defining cloud architecture in code and provisioning it using AWS CloudFormation. It provides a high-level object-oriented framework for defining AWS resources with the capability of current programming languages. You can quickly include AWS best practices in your infrastructure definition and publish it without worrying about boilerplate logic using CDK's library of infrastructure components.

Key Takeaways

- Understanding the Cryptocurrency dataset
- Understanding the Alpha Vantage API
- Understanding the AWS CDK
- Installation of AWS CDK and its various commands
- Advantages of Serverless technologies
- Creating an AWS Cloud9 environment
- Creating Data Producers Lambda Stack
- Creating Data Consumers Lambda Stack
- Creating Kinesis Data Streams
- Setting up environment variables
- Deployment of AWS CDK
- Performing Data Analytics using Apache Flink and Apache Zeppelin
- Creating ETL jobs using AWS Glue
- Performing analysis using Amazon Athena

Architecture Diagram:


Data Producer
Kinesis Stream
Data Consumer
NoSQL Storage
Real-Time Analytics
Glue/Spark ETL