

Project Report

EQUITY INSIGHT TRENDS EXPLORER

CIS:602 BIG DATA ANALYTICS

Submitted By

Shubham Singh (53)

Special Thanks To

Dr. Ashokkumar Patel

Department of Data Science

Department of Computer & Information Science

August 14, 2024

DESCRIPTION

This project, titled "Equity Insight Trends Explorer," aims to analyze the performance trends and correlations of various global stock exchanges using a real-time equity dataset. By examining diverse stock exchanges from different geographical regions and market sectors, the project provides a comprehensive view of international financial markets. The analysis is supported by a robust data pipeline built on AWS infrastructure, which ensures efficient data storage, processing, and continuous integration for handling large datasets. The primary goal is to gain insights into market trends over time, leveraging advanced AWS services like S3, Glue, CloudWatch, Lambda, Athena, and Sage Maker to deliver accurate and timely analytics.

DATASET DESCRIPTION:

The dataset utilized in this project consists of real-time equity data from a diverse range of global stock exchanges, representing various geographical regions and market sectors. The stock exchanges included are HSI (Hang Seng Index), NYA (NYSE Composite Index), IXIC (NASDAQ Composite Index), 000001.SS (Shanghai Composite Index), N225 (Nikkei 225), N100 (Euronext 100 Index), 399001.SZ (Shenzhen Component Index), GSPTSE (S&P/TSX Composite Index), NSEI (Nifty 50), GDAXI (DAX Index), SSMI (Swiss Market Index), TWII (Taiwan Weighted Index), and J203.JO (Johannesburg Stock Exchange Top 40 Index). The dataset captures key financial metrics, trading volumes, and index values over time, providing a rich source of information for trend analysis and correlation studies across different markets.

TECHNOLOGIES:

The project leverages a suite of AWS cloud services to build and manage the data pipeline and perform analytics on the real-time equity dataset. Key technologies include:

- **AWS S3:** Used for scalable storage of large equity datasets, ensuring secure and durable data storage.
- **AWS Glue:** Employed for data extraction, transformation, and loading (ETL), enabling seamless integration and preparation of data for analysis.
- **AWS CloudWatch:** Utilized for monitoring the performance and operational health of the data pipeline, ensuring real-time visibility into system operations.
- **AWS Lambda:** Provides serverless computing capabilities, triggering automated processes for data processing and analysis without the need for managing servers.
- **AWS Athena:** Allows for interactive query analysis directly on the stored datasets, facilitating quick insights through SQL queries on S3 data.
- **AWS Sage Maker:** Deployed for building, training, and deploying machine learning models to predict and analyze stock market trends based on the processed data.

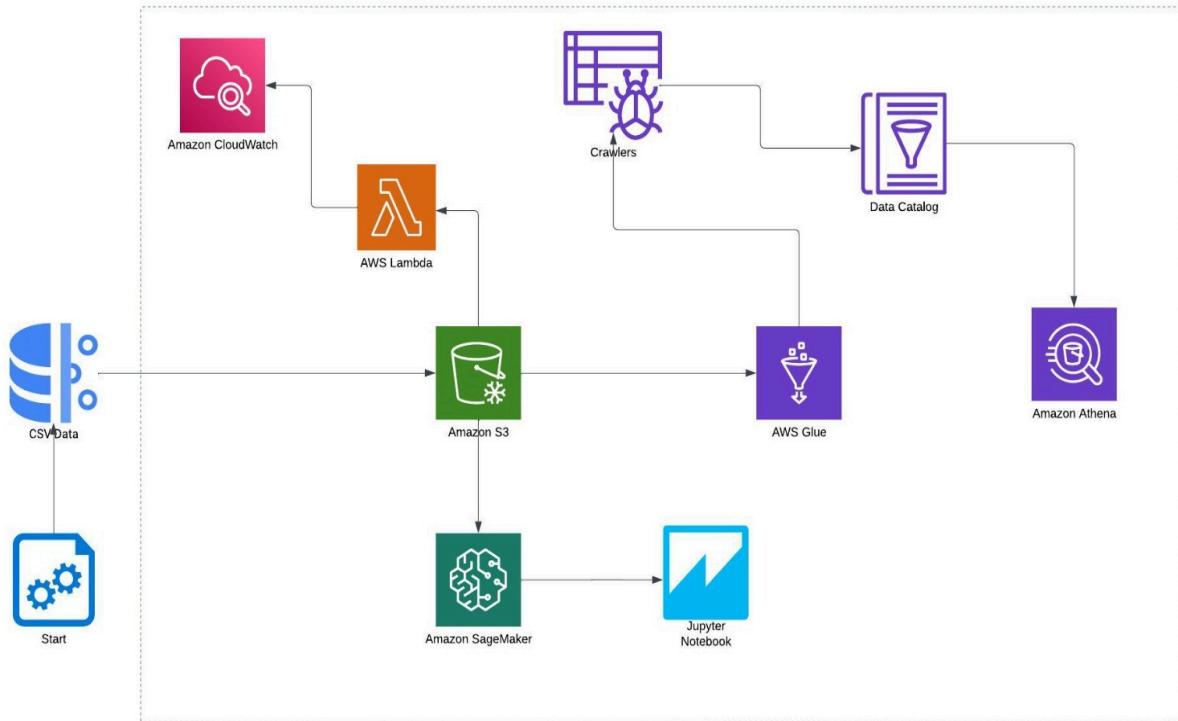
These technologies work together to create a powerful, efficient, and scalable infrastructure capable of handling complex data analytics tasks in the cloud.

IMPLEMENTATIONS:

The project involves the implementation of a robust and scalable data pipeline designed to analyze real-time equity data from global stock exchanges. The key implementations include:

1. **Data Ingestion:** Real-time equity data from various stock exchanges is ingested into AWS S3, where it is securely stored in a structured format, ready for processing.
2. **ETL Processes:** AWS Glue is employed to extract, transform, and load (ETL) the ingested data, ensuring it is clean, consistent, and optimized for further analysis. This includes handling missing data, standardizing formats, and aggregating metrics.
3. **Serverless Data Processing:** AWS Lambda functions are used to automate data processing tasks, such as triggering ETL jobs, running periodic analyses, and responding to specific data events, enabling real-time analytics without the need for dedicated servers.
4. **Interactive Querying:** AWS Athena is utilized to perform interactive SQL queries on the processed data directly in S3. This allows for quick and flexible exploration of the dataset to uncover trends and insights.
5. **Machine Learning Integration:** AWS SageMaker is used to build and deploy machine learning models that analyze the processed data to predict market trends and identify significant correlations across different stock exchanges.
6. **Monitoring and Logging:** AWS CloudWatch monitors the entire pipeline, providing real-time logs and performance metrics. This ensures the reliability and efficiency of the system, with alerts set up for any anomalies or failures in the pipeline.

ARCHITECTURE



- **Amazon S3:** Stores raw and processed data. Acts as the central data lake for your data pipeline, where all data is ingested and persisted before further processing.
- **AWS Glue:** Performs ETL (Extract, Transform, Load) operations. It extracts data from S3, transforms it based on business rules, and loads the processed data back into S3 or directly into a data warehouse.
- **AWS Glue Crawler:** Automatically discovers and catalogs metadata from data stored in S3. It updates the AWS Glue Data Catalog with the schema and structure of the data.
- **AWS Glue Data Catalog:** Serves as a centralized metadata repository that stores and manages schema information for your datasets. It enables efficient querying and analysis by keeping track of data structure and location.
- **Amazon Athena:** Allows you to query data directly from S3 using standard SQL queries. It leverages the metadata from the AWS Glue Data Catalog to perform ad-hoc queries and analysis.
- **Amazon SageMaker:** Provides machine learning capabilities. It uses the processed data to build, train, and deploy machine learning models for predictive analytics and insights.
- **Jupyter Notebook:** Used within SageMaker for interactive data analysis and model development. It provides a user-friendly interface for running Python code, visualizing data, and experimenting with machine learning models.
- **AWS Lambda:** Triggers automatic actions based on events. For instance, it can be configured to invoke an ETL job or a data processing workflow whenever new data is uploaded to the S3 bucket.
- **Amazon CloudWatch:** Monitors and logs the operations within the AWS environment. It tracks the performance and health of AWS services, including Lambda functions and ETL jobs, and can trigger alerts based on predefined conditions.

IMPLEMENTATION OF CI/CD PIPELINE

Data Ingestion

Real-time equity data from various global stock exchanges is ingested into AWS S3, where it is securely stored in a structured and scalable format. This step ensures that the data is readily accessible for further processing and analysis in the pipeline.

Amazon S3

Account snapshot - updated every 24 hours All AWS Regions

Storage lens provides visibility into storage usage and activity trends. [Learn more](#)

[View Storage Lens dashboard](#)

General purpose buckets | Directory buckets

General purpose buckets (2) [Info](#) All AWS Regions

Buckets are containers for data stored in S3.

< 1 > ⌂

Name	AWS Region	IAM Access Analyzer	Creation date
aws-glue-assets-308053217767-us-east-1	US East (N. Virginia) us-east-1	View analyzer for us-east-1	August 10, 2024, 13:57:21 (UTC-07:00)
bigdataprojectettpipeline	US East (N. Virginia) us-east-1	View analyzer for us-east-1	August 10, 2024, 00:02:09 (UTC-07:00)

Services Search [Option+S]

Amazon Redshift EC2 IAM Athena CloudFormation AWS Glue Lambda Amazon SageMaker

N. Virginia v vocabs/user3318216@tfatya@umassd.edu @ 3080-5321-7767

Amazon S3 > Buckets > bigdataprojectettpipeline

bigdataprojectettpipeline [Info](#)

Objects (5) [Info](#)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

< 1 > ⌂

Name	Type	Last modified	Size	Storage class
athena_query_logs/	Folder	-	-	-
orders/	Folder	-	-	-
original_data/	Folder	-	-	-
processed_data/	Folder	-	-	-
raw_data/	Folder	-	-	-

AWS GLUE

ETL Processes

AWS Glue is used to extract, transform, and load (ETL) the ingested data, ensuring it is clean, consistent, and optimized for analysis. This process prepares the data for efficient querying and machine learning tasks.

The screenshot shows the AWS Glue Studio interface. At the top, there's a navigation bar with services like Amazon Redshift, EC2, IAM, Athena, CloudFormation, AWS Glue (which is selected), Lambda, and Amazon SageMaker. The main area has a title "AWS Glue Studio" with a "Info" link. Below it, there's a section titled "Create job" with three options: "Visual ETL" (selected), "Notebook", and "Script editor". Under "Example jobs", there's a "Create example job" button. The "Your jobs (1) Info" section shows one job named "stock_etl" with details: Type: Glue ETL, Last modified: 11/08/2024, 09:56:56, and AWS Glue version: 4.0.

DATA CATALOG

AWS Glue automatically creates a data catalog, organizing metadata from the ingested datasets. This catalog allows for easy discovery, indexing, and management of data, facilitating seamless access and efficient querying across the pipeline.

The screenshot shows the AWS Glue Databases interface. At the top, there's a navigation bar with services like Amazon Redshift, EC2, IAM, Athena, CloudFormation, AWS Glue (selected), Lambda, and Amazon SageMaker. The main area has a title "Databases (1)" with a "Info" link. It shows a database named "db_datapipeline" with the following details: Last updated (UTC): August 12, 2024 at 23:22:27, and Created on (UTC): August 10, 2024 at 07:24:33. There are buttons for "Edit" and "Delete", and an "Add database" button.

TABLE AND SCHEMA

Data is organized into tables with well-defined schemas in AWS Glue, ensuring structured storage and easy access for querying. The schema design supports efficient data retrieval and integration across various analytics and machine learning tasks.

Name: original_data

Database: db_datapipeline

Description: -

Last updated: August 11, 2024 at 02:50:52

Classification: CSV

Location: s3://bigdataproject1/pipeline/original_data/

Connection: -

Deprecated: -

Column statistics: No statistics

#	Column name	Data type	Partition key	Comment
1	index	string	-	-
2	date	string	-	-
3	open	double	-	-
4	high	double	-	-
5	low	double	-	-
6	close	double	-	-
7	adj close	double	-	-
8	volume	double	-	-
9	closeusd	double	-	-

CRAWLER

A crawler is an automated program that systematically explores the internet, indexing web pages for search engines.

Crawlers (4) Info

Last updated (UTC): August 12, 2024 at 23:24:58

Action Run Create crawler

Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last run
orders_project	Ready		Succeeded	August 11, 2024 at 02:57:39	View log	-
original_data	Ready		Succeeded	August 11, 2024 at 02:49:47	View log	1 updated
processed_crawler	Ready		Succeeded	August 11, 2024 at 05:36:55	View log	1 updated
raw_data	Ready		Succeeded	August 11, 2024 at 00:07:24	View log	-

Crawler is working for the original data to transform.

The screenshot shows the AWS Glue Crawler configuration page for a crawler named 'processed_crawler'. The crawler is set to run every 1 minute and is currently in a 'READY' state. It is configured to use an IAM role 'LabRole' and a database 'db_datapipeline'. The crawler has run 9 times successfully, with the most recent run completed on August 11, 2024, at 05:38:00. The results table shows 1 table change and 0 partition changes per run. The 'Crawler runs' tab is selected, showing the history of runs from August 10 to August 11, 2024.

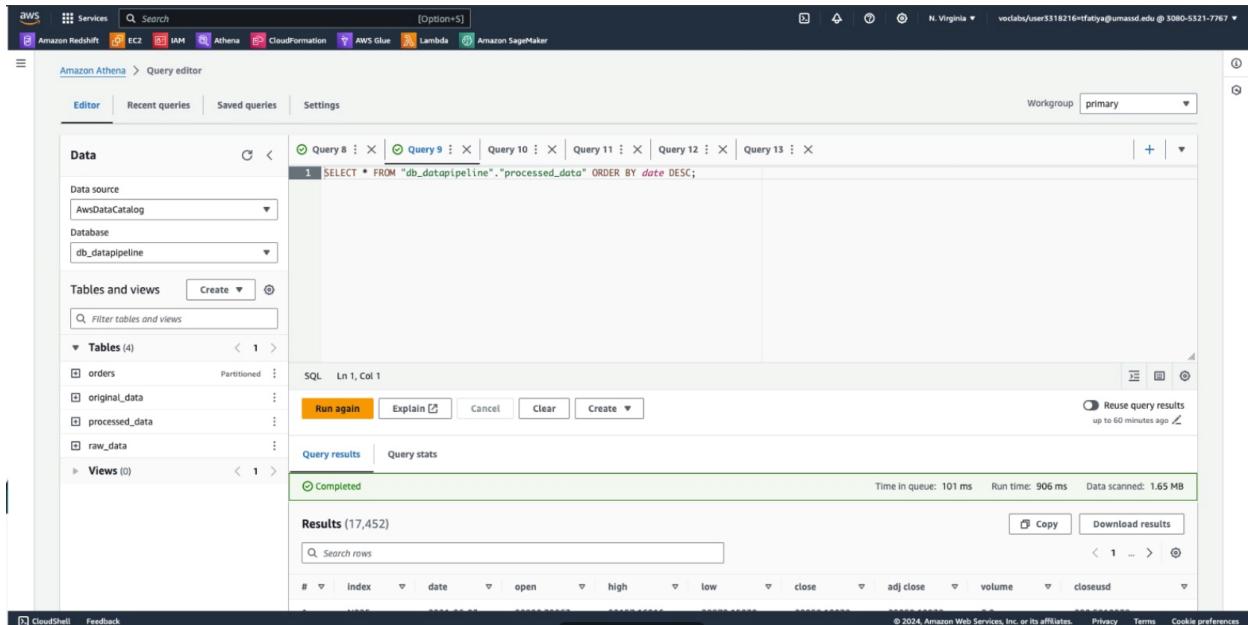
ATHENA

Data populated for the original dataset in which the number of data points are 104,224.

The screenshot shows the Amazon Athena Query editor interface. A single query is running, selecting all columns from the 'original_data' table in the 'db_datapipeline' database. The results show 104,224 rows. The query editor includes a sidebar for data sources, databases, and tables, and a bottom pane for viewing the results.

Transformed Data

Transformed data is data that has been modified or restructured from its original format to serve a specific purpose or improve its usability.



The screenshot shows the Amazon Athena Query editor interface. The query editor tab is selected, and the main area displays the following SQL query:

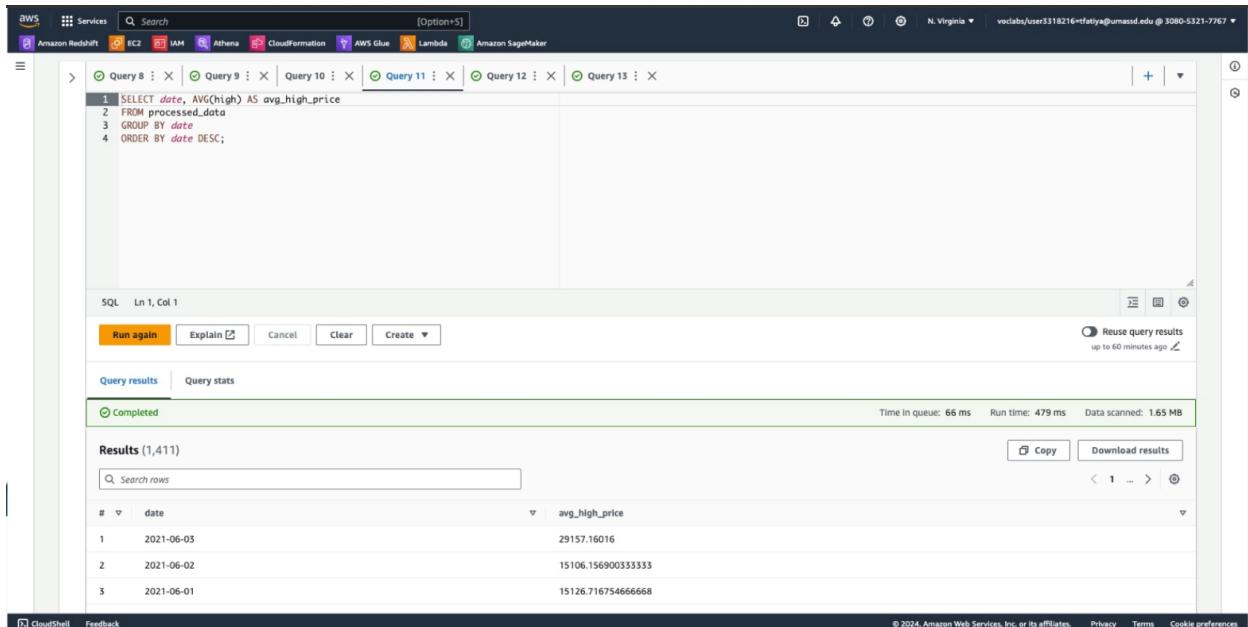
```
1 | SELECT * FROM "db_datapipeline"."processed_data" ORDER BY date DESC;
```

The sidebar on the left shows the data source is set to AwsDataCatalog and the database is db_datapipeline. Under Tables and views, there are four tables listed: orders, original_data, processed_data, and raw_data. The processed_data table is currently selected. There are no views listed.

The results section shows the query completed successfully with 17,452 rows. The columns listed are index, date, open, high, low, close, adj close, volume, and closeusd. Buttons for Copy and Download results are available.

QUERYING

In this process of extracting specific information from a dataset by formulating and executing a request which is showing the average high price of the stock indexes as per the date.



The screenshot shows the Amazon Athena Query editor interface. The query editor tab is selected, and the main area displays the following SQL query:

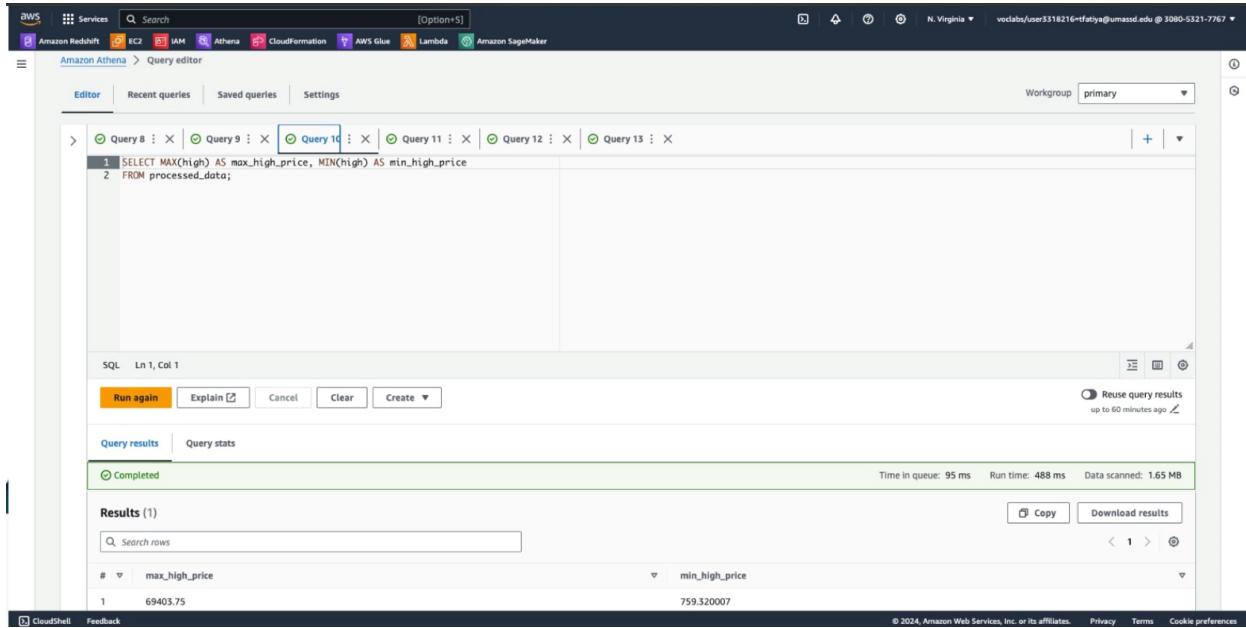
```
1 | SELECT date, AVG(high) AS avg_high_price
2 | FROM processed_data
3 | GROUP BY date
4 | ORDER BY date DESC;
```

The sidebar on the left shows the data source is set to AwsDataCatalog and the database is db_datapipeline. Under Tables and views, there are four tables listed: orders, original_data, processed_data, and raw_data. The processed_data table is currently selected. There are no views listed.

The results section shows the query completed successfully with 1,411 rows. The columns listed are date and avg_high_price. The data is as follows:

date	avg_high_price
2021-06-03	29157.16016
2021-06-02	15106.1569003533333
2021-06-01	15126.7167546666668

In this query we are populating the data showing the maximum and minimum high price of the stock indexes.



The screenshot shows the AWS Management Console with the Athena service selected. In the Query editor, a new query named "Query 10" is open. The SQL code is:

```

1 SELECT MAX(high) AS max_high_price, MIN(high) AS min_high_price
2 FROM processed_data;

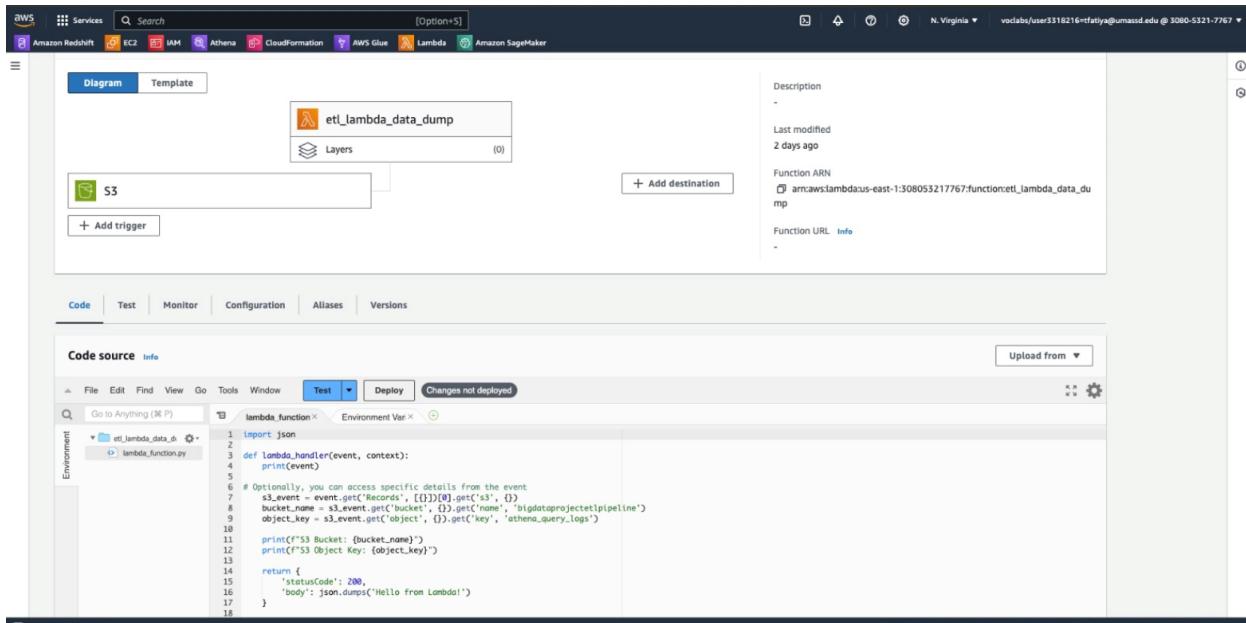
```

Below the code, there are buttons for "Run again", "Explain", "Cancel", "Clear", and "Create". The "Reuse query results" checkbox is checked. The "Query results" tab is selected, showing the results of the completed query:

#	max_high_price	min_high_price
1	69403.75	759.320007

Lambda Function

A Lambda function in AWS is a serverless compute service that automatically runs code in response to events and triggers, scaling effortlessly based on demand. It eliminates the need for server management, allowing developers to focus on writing code rather than managing infrastructure.



CLOUD WATCH

Amazon CloudWatch is a monitoring and observability service that provides real-time insights into AWS resources and applications. It helps track metrics, collect logs, and set alarms to ensure the health and performance of your AWS infrastructure.

The screenshot shows the AWS CloudWatch Log groups interface. The left sidebar includes sections for Dashboards, Alarms, Logs, Log groups, Log Anomalies, Live Tail, Logs Insights, Contributor Insights, Metrics, X-Ray traces, Events, Application Signals, Network monitoring, and Insights. The main area displays log events for the path /aws/lambda/etl_lambda_data_dump. A specific log entry from August 12, 2024, at 28:29:22Z is expanded, showing JSON data related to an AWS Lambda execution. The log entry includes fields like 'Records' and detailed information about the event, such as the principal ID and request ID.

AWS SAGEMAKER

Amazon SageMaker is a fully managed service that simplifies building, training, and deploying machine learning models at scale. It offers integrated tools for data preparation, model training, and inference.

The screenshot shows the AWS SageMaker Notebooks and Git repos interface. The left sidebar includes sections for Getting started, Applications and IDEs (Studio, Canvas, RStudio, TensorBoard, Profiler, Notebooks), Admin configurations (Domains, Role manager, Images, Lifecycle configurations), SageMaker dashboard, Search, JumpStart (Foundation models, Computer vision models, Natural language processing models), and Governance. The main area displays a promotional banner for JupyterLab in SageMaker Studio, followed by a table of Notebook instances. One instance, named 'stockdataanalysis', is listed with details: Name (stockdataanalysis), Instance (ml.t3.medium), Creation time (8/10/2024, 1:18:21 PM), Status (Stopped), and Actions (Start). A 'Create notebook instance' button is also visible.

CONCLUSION

The project "Equity Insight Trends Explorer" effectively showcases the ability of AWS cloud technologies to handle and evaluate intricate, real-time equity data sourced from international stock exchanges. Through the utilization of an advanced AWS service such as S3, Glue, Lambda, Athena, and SageMaker, along with a well-designed data pipeline, the project offers significant insights into market patterns and correlations across many geographies and industries. The application of machine learning models, interactive querying, and serverless computing significantly improves the project's capacity to provide precise and rapid insights.

The initiative also shows how democratizing access to advanced financial analytics tools is possible. Innovative financial technology is promoted by the use of cloud technologies, which enable smaller businesses or lone academics to undertake sophisticated data analysis that was previously limited to major financial firms.

The "Equity Insight Trends Explorer" project successfully demonstrates the power of AWS cloud technologies in managing and analyzing complex, real-time equity data from global stock exchanges. By leveraging a well-designed data pipeline and advanced AWS services like S3, Glue, Lambda, Athena, and SageMaker, the project provides valuable insights into market trends and correlations across different regions and sectors. The implementation of serverless computing, interactive querying, and machine learning models further enhances the project's ability to deliver accurate and timely analytics. This project not only highlights the capabilities of modern cloud infrastructure in financial data analysis but also sets a strong foundation for future enhancements and scalability, enabling deeper market insights and more sophisticated predictive models.

By harnessing the power of cloud computing and advanced analytics, it contributes to the ongoing digital transformation of the financial sector, promising more informed decision-making and deeper understanding of global market dynamics.

REFERENCES

- [1] Short-term stock market price trend predictions Journal of Big Data
<https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6>
- [2] Stock Analysis: Different Methods for Evaluating Stocks By James Chen Updated March 11, 2023 Reviewed by Gordon Scott Fact checked by Yarilet Perez
<https://www.investopedia.com/terms/s/stock-analysis.asp#:~:text=Fundamental%20analysis%20is%20a%20method,find%20under%2D%20r%20overvalued%20stocks.&text=Trend%20analysis%20is%20a%20technique,on%20recently%20observed%20trend%20data..>
- [3] QuantInsti Blog "Stock Market Data: Obtaining Data, Visualization & Analysis in Python"
<https://blog.quantinsti.com/stock-market-data-analysis-python/>