Faculty of Engineering & Applied Science

Cloud Computing - 74293

Project Milestone 4:

Data Processing: Dataflow-Apache Beam

Taha Hashmat - 100689792

Question:

Google Cloud has another processing service called DataProc. Name another processing service that is usually used in the cloud environment (not necessarily GCP). Compare between it and both Dataflow and DataProc. Your comparison may include but is not limited to the major differences, advantages, disadvantages, and limitations.

Answer:

What is Google Cloud DataPrep

Another processing service that is usually used in the cloud environment is Google Cloud Dataprep. Dataprep is essentially what its name says it is. It is used to prepare structured and unstructured data by filtering it, going through it, and cleaning it for services such as Machine Learning, Data Analytics, etc.

DataPrep vs Dataflow

- Major Differences:

The primary purpose of DataPrep is to process data so that it can be utilized further down the line. Dataflow is mainly considered an analytical service and lets users ingest high volumes of both batch and stream processing of data in real time.

- Advantages:

1. Dataprep is relatively easy to use as compared to dataproc which is considered to be harder
2. Dataprep requires only BigTable and BigQuery for system integration while dataflow requires both of them as well as Apache Beam making the process more tedious

- Disadvantages:

1. Dataprep does not include database replication while in dataflow using SELECT statements, replication can be achieved
2. Dataprep does not give the user to add new data sources whereas dataflow does

- Limitations:

Limitations for Dataprep include that it is only used for the simple processing of data and not much else for ex the stream and batch processing of data

DataPrep vs DataProc

- Major Differences:

The primary purpose of DataPrep is to process data so that it can be utilized further down the line. Dataproc is also a data processing service however dataproc has the ability to process streams and batch data as well. Furthermore data proc is also designed to run on clusters

- Advantages:
  Dataproc is more scalable and flexible mainly due to its ability of processing larger amounts of data simultaneously
- Disadvantages:
  Dataprep cannot handle huge volumes of data as compared to dataproc
- Limitations:
  As we mentioned data prep's inability to handle huge amounts of data, the maximum number of datasets that Dataprep can manage in its workspace is 1000.

Question:

Suggest a practical application using both stream and batch processing that can be applied to a given dataset. It’s expected to use the dataset uploaded in the third milestone but you can use any other dataset. If you decide to use another dataset, It should maintain both variety and huge volume. Your report should include but not limited to:

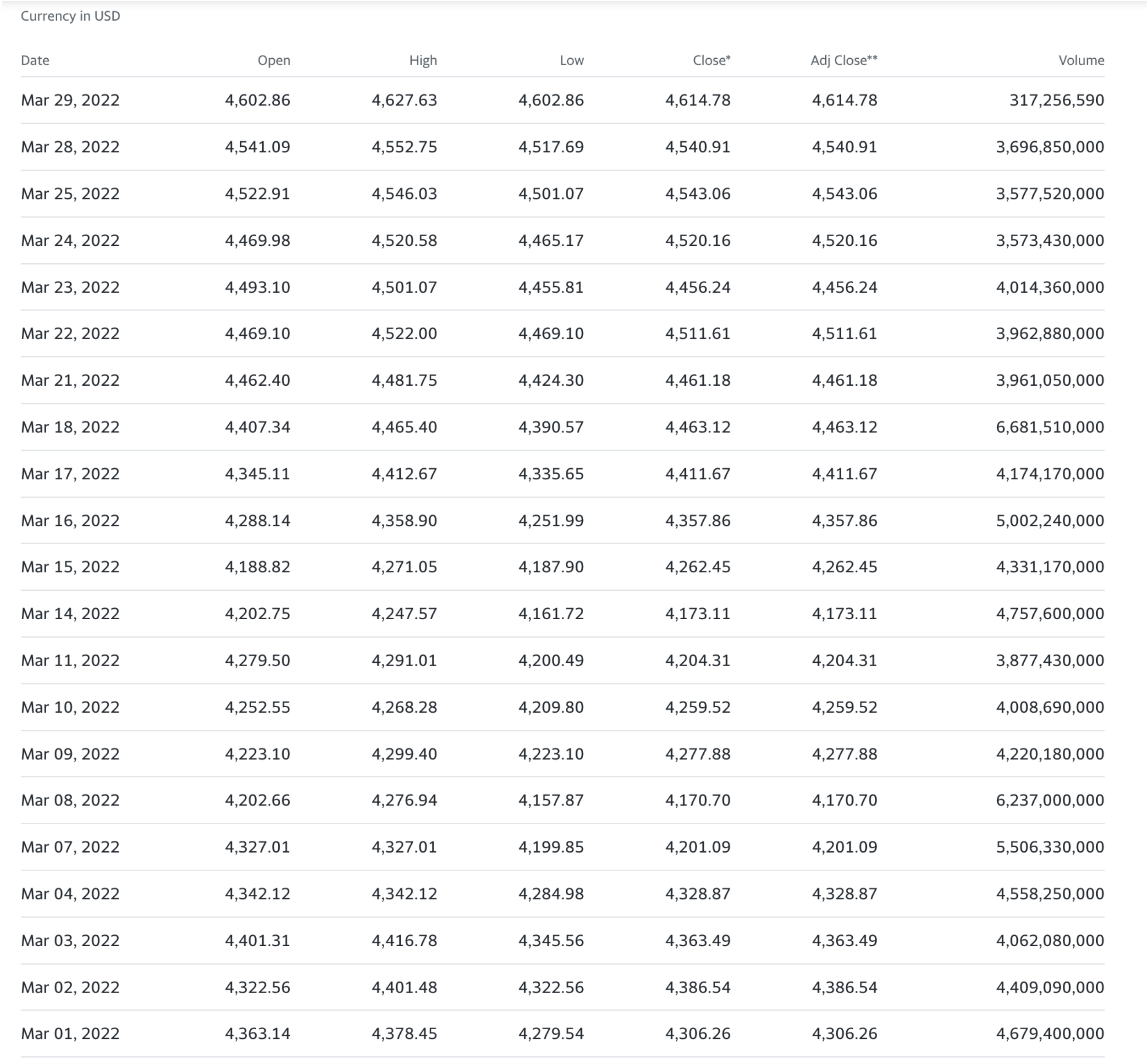
- The application
- Its impact
- The used dataset (size, schema/structure)
- A graph showing the proposed pipeline(s)
- List of other tools (AI, clustering,…) needed to implement that application

Answer:

The Application:
The application will be a stock market analyzer. Stream Processing can be used to make real time predictions during the periods the stock market is open, by tracking variables in the data such as volume and opening prices. Batch processing can be used to store data from previous data by tracking variables such as closing prices, opening prices and calculating technical indicators from these stored values such as EMA, SMA AND ROI, to predict what stock will do well and which ones won't perform at their best in the coming weeks

Its Impact:
This application will be impactful to mainly 2 groups of people. Number one being day traders and number two being businesses. Day traders will find the aspect of stream processing in the application really useful since it will be providing them real time updates on stock tickers and it will help them with their decisions since they make impulsive calls on the spot. Businesses will find the batch processing aspect of the application really useful since they can smartly invest their assets into predictions made on stocks for the future.

Used Dataset:



Currency in USD

Date	Open	High	Low	Close*	Adj Close**	Volume
Mar 29, 2022	4,602.86	4,627.63	4,602.86	4,614.78	4,614.78	317,256,590
Mar 28, 2022	4,541.09	4,552.75	4,517.69	4,540.91	4,540.91	3,696,850,000
Mar 25, 2022	4,522.91	4,546.03	4,501.07	4,543.06	4,543.06	3,577,520,000
Mar 24, 2022	4,469.98	4,520.58	4,465.17	4,520.16	4,520.16	3,573,430,000
Mar 23, 2022	4,493.10	4,501.07	4,455.81	4,456.24	4,456.24	4,014,360,000
Mar 22, 2022	4,469.10	4,522.00	4,469.10	4,511.61	4,511.61	3,962,880,000
Mar 21, 2022	4,462.40	4,481.75	4,424.30	4,461.18	4,461.18	3,961,050,000
Mar 18, 2022	4,407.34	4,465.40	4,390.57	4,463.12	4,463.12	6,681,510,000
Mar 17, 2022	4,345.11	4,412.67	4,335.65	4,411.67	4,411.67	4,174,170,000
Mar 16, 2022	4,288.14	4,358.90	4,251.99	4,357.86	4,357.86	5,002,240,000
Mar 15, 2022	4,188.82	4,271.05	4,187.90	4,262.45	4,262.45	4,331,170,000
Mar 14, 2022	4,202.75	4,247.57	4,161.72	4,173.11	4,173.11	4,757,600,000
Mar 11, 2022	4,279.50	4,291.01	4,200.49	4,204.31	4,204.31	3,877,430,000
Mar 10, 2022	4,252.55	4,268.28	4,209.80	4,259.52	4,259.52	4,008,690,000
Mar 09, 2022	4,223.10	4,299.40	4,223.10	4,277.88	4,277.88	4,220,180,000
Mar 08, 2022	4,202.66	4,276.94	4,157.87	4,170.70	4,170.70	6,237,000,000
Mar 07, 2022	4,327.01	4,327.01	4,199.85	4,201.09	4,201.09	5,506,330,000
Mar 04, 2022	4,342.12	4,342.12	4,284.98	4,328.87	4,328.87	4,558,250,000
Mar 03, 2022	4,401.31	4,416.78	4,345.56	4,363.49	4,363.49	4,062,080,000
Mar 02, 2022	4,322.56	4,401.48	4,322.56	4,386.54	4,386.54	4,409,090,000
Mar 01, 2022	4,363.14	4,378.45	4,279.54	4,306.26	4,306.26	4,679,400,000

Dataset: https://finance.yahoo.com/quote/%5EGSPC/history/
The dataset being used is from Yahoo Finance. For this application we will focus on the S and P 500 and its companies. It includes all the opening, closing prices, the highs and the lows, as well as the volume of trading. The data is for one calendar year and thus contains 365 rows and 7 columns.

Other Tools:
The other tools that will be used are machine learning models being trained using libraries and resources such as tensorflow and ski citlearn. The model will probably use a hill climbing algorithm as well to optimize the top 20 and bottom 20 predicted performing stocks for the next 2 weeks (for businesses use) and provide real time predictions for day traders.