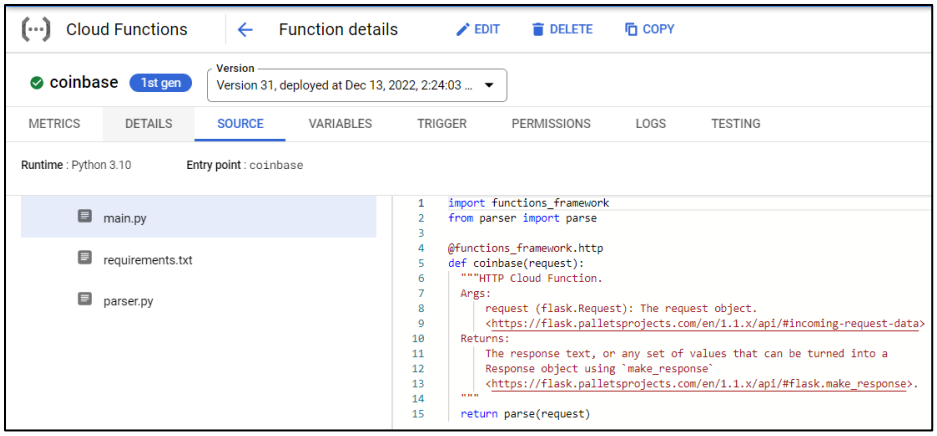
BIG DATA – CAPSTONE PROJECT

Predicting Crypto Market Tendencies – Volatility as an Advantage

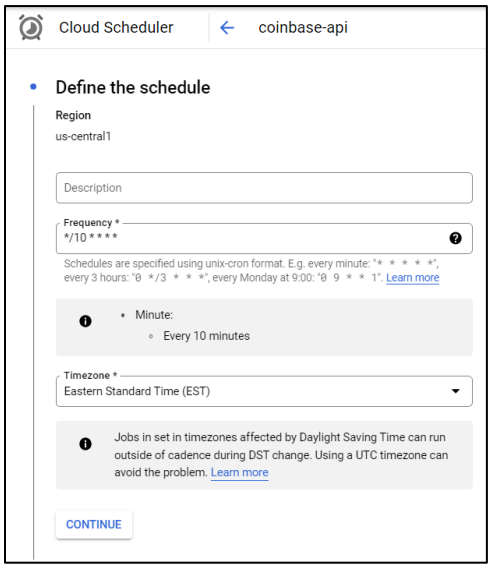
For this project, the team is working on crypto currency data, where archived data is procured from the CoinMarketCap website, and live streaming data is procured from Coinbase API. The goal for the project is to provide a live prediction of market close volume sold across cryptos and closing exchange rate to generate insights that lead to quick profits in the Crypto Currency market.

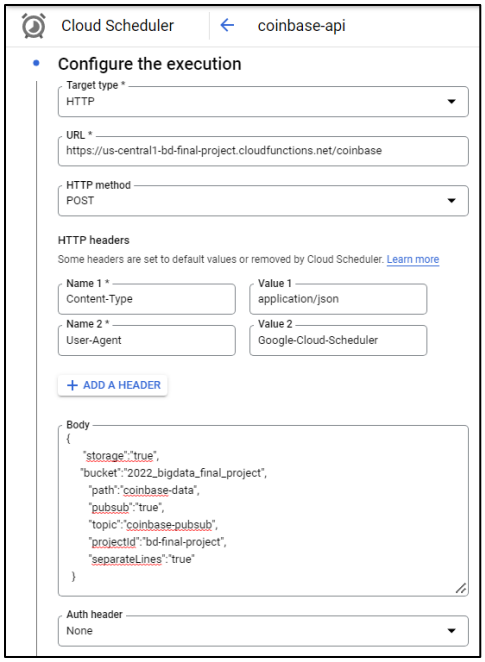
The steps followed by the team are as follows:

1. **Setting up the pipeline**
    a. Create a cloud function to connect to the Coinbase parser from the Coinbase API



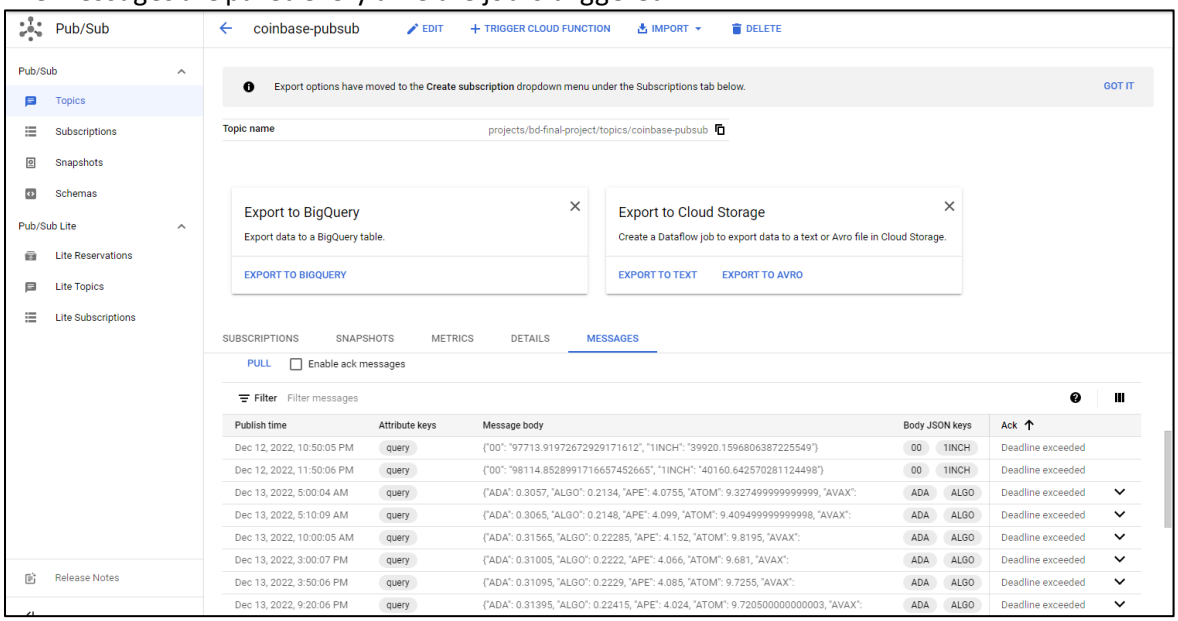
b. Define a job schedule using the Cloud Scheduler with a trigger frequency of 10 minutes. The scheduler triggers the cloud function to transfer the JSON directly to the pubsub and a single dataflow transfers the data to BigQuery. The configuration is shown below.





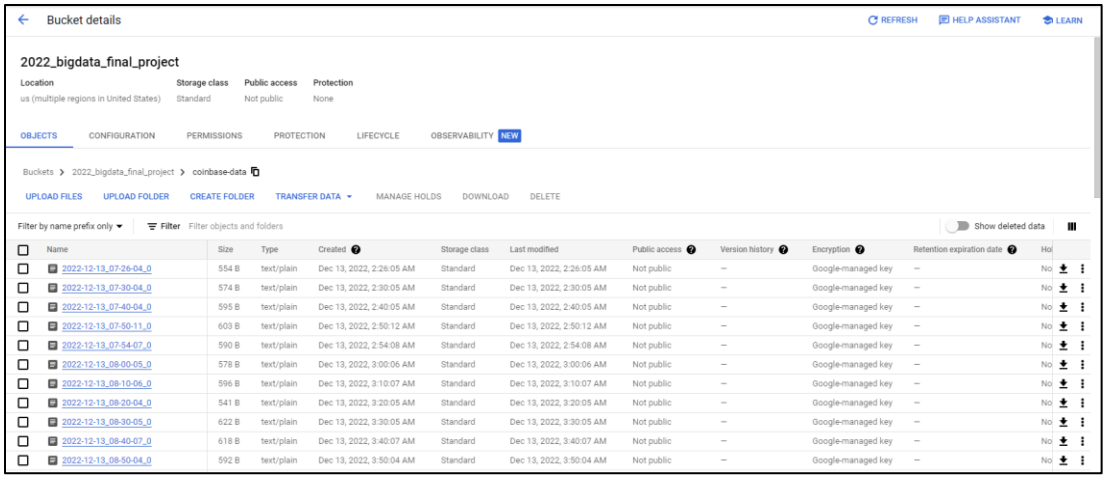
c. Create a Pubsub

The messages are pulled every time the job is triggered

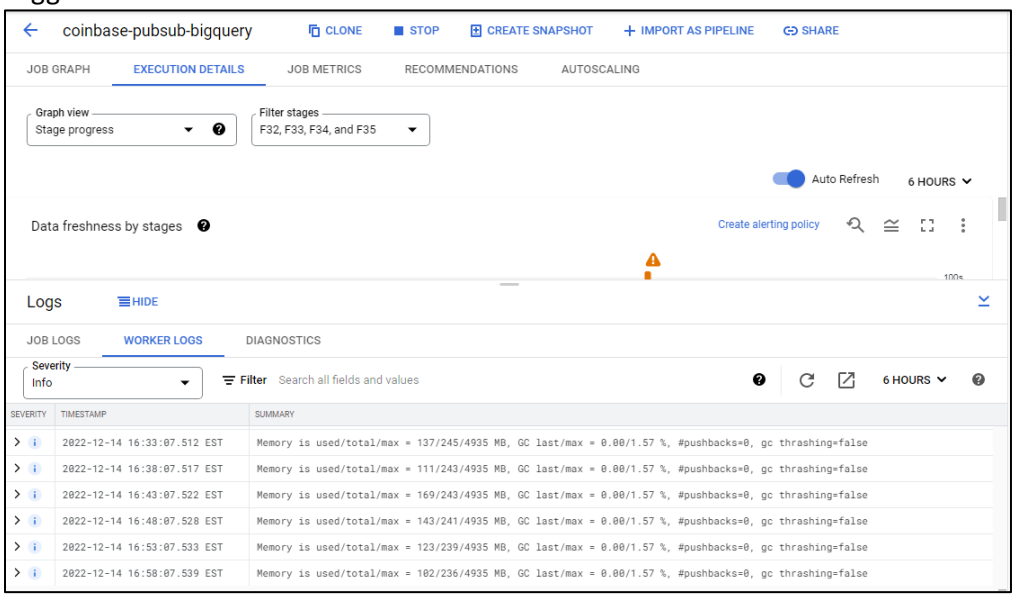


d. Store data in GCS

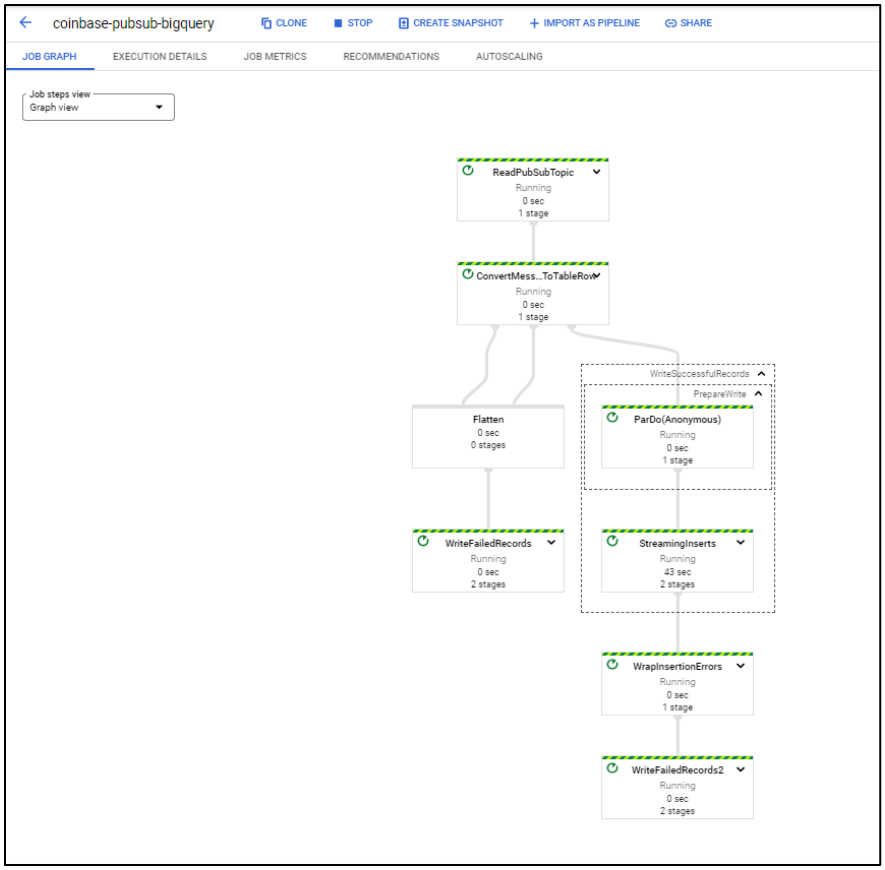
With the given configuration in the scheduler, the streaming records are pushed into the GCS buckets as files.



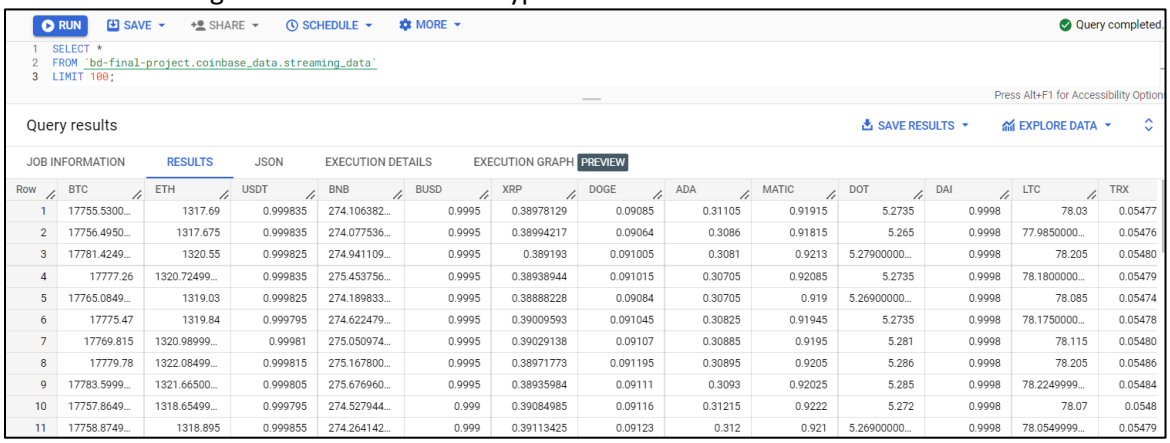
e. Create a dataflow: The debug logs depict the data being transferred to Big Query at each trigger



f. Dataflow jobs move data from GCS to BigQuery



g. BigQuery data set with live data for analysis
Below is a screengrab of the live stream crypto dataset



Row	BTC	ETH	USDT	BNB	BUSD	XRP	DOGE	ADA	MATIC	DOT	DAI	LTC	TRX
1	17755.5300...	1317.69	0.999835	274.106382...	0.9995	0.38978129	0.09085	0.31105	0.91915	5.2735	0.9998	78.03	0.05477
2	17756.4950...	1317.675	0.999835	274.077536...	0.9995	0.38994217	0.09064	0.3086	0.91815	5.265	0.9998	77.9850000...	0.05476
3	17781.4249...	1320.55	0.999825	274.941109...	0.9995	0.389193	0.091005	0.3081	0.9213	5.27900000...	0.9998	78.205	0.05480
4	17777.26	1320.72499...	0.999835	275.453756...	0.9995	0.38938944	0.091015	0.30705	0.92085	5.2735	0.9998	78.1800000...	0.05479
5	17765.0849...	1319.03	0.999825	274.189833...	0.9995	0.38888228	0.09084	0.30705	0.919	5.26900000...	0.9998	78.085	0.05474
6	17775.47	1319.84	0.999795	274.622479...	0.9995	0.39009593	0.091045	0.30825	0.91945	5.2735	0.9998	78.1750000...	0.05478
7	17769.815	1320.98999...	0.99981	275.050974...	0.9995	0.39029138	0.09107	0.30885	0.9195	5.281	0.9998	78.115	0.05480
8	17779.78	1322.08499...	0.999815	275.167800...	0.9995	0.38971773	0.091195	0.30895	0.9205	5.286	0.9998	78.205	0.05486
9	17783.5999...	1321.66500...	0.999805	275.676960...	0.9995	0.38935984	0.09111	0.3093	0.92025	5.285	0.9998	78.2249999...	0.05484
10	17757.8649...	1318.65499...	0.999795	274.527944...	0.999	0.39084985	0.09116	0.31215	0.9222	5.272	0.9998	78.07	0.0548
11	17758.8749...	1318.895	0.999855	274.264142...	0.999	0.39113425	0.09123	0.312	0.921	5.26900000...	0.9998	78.0549999...	0.05479

2. Procuring archive data

Archive data for crypto exchange rates was procured from the CoinMarketCap website. The data includes name of the crypto currency, the date for which data is procured, the opening exchange rate on the day, the closing exchange rate on the day, the daily high of the exchange rate, the daily low of the exchange rate, the adjusted closing exchange rate on the day and the volume sold on the day.

A screengrab of the data is available below.

```
SELECT *
FROM `bd-final-project.coinbase_data.archive`
LIMIT 100;
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH PREVIEW

Row	Name	Date	Open	High	Low	Close	Adj_Close	Volume
1	ADA	2021-12-13	1.347895	1.357773	1.202574	1.225348	1.225348	144879170...
2	ADA	2021-12-14	1.22446	1.269876	1.201422	1.222835	1.222835	159620268...
3	ADA	2021-12-15	1.266263	1.329269	1.207161	1.311847	1.311847	163406653...
4	ADA	2021-12-16	1.311818	1.33101	1.236784	1.240534	1.240534	129331159...
5	ADA	2021-12-17	1.240744	1.258695	1.187881	1.219892	1.219892	137613128...
6	ADA	2021-12-18	1.220517	1.267222	1.201947	1.242534	1.242534	105798936...
7	ADA	2021-12-19	1.242394	1.309385	1.241757	1.244661	1.244661	122019647...
8	ADA	2021-12-20	1.243437	1.260699	1.202716	1.23824	1.23824	136530648...
9	ADA	2021-12-21	1.236419	1.289425	1.229115	1.280859	1.280859	114907800...
10	ADA	2021-12-22	1.280337	1.366798	1.278285	1.328041	1.328041	152880498...
11	ADA	2021-12-23	1.328843	1.489489	1.311121	1.474691	1.474691	206852429...
12	ADA	2021-12-24	1.473877	1.490347	1.38348	1.392367	1.392367	133779971...

Callouts

- The exchange rates provided in the datasets are against USD.
- Data for the top 28 crypto currencies, according to coinmarketcap, are available in the dataset.
- The open and close values in archive data are recorded at 00 hr at day change. All our metrics and predictions are based on it.

Metrics:

Fibonacci Retracement: Fibonacci retracement is a technical analysis tool that uses horizontal lines to indicate areas of support or resistance at the key Fibonacci levels before the price continues in the original direction. These levels are derived from the Fibonacci sequence and are commonly used in conjunction with trend lines to find entry and exit points in the market. We have calculated these levels at 38.5%, 50%, 61.8% and 75% respectively which serve as checkpoint for investing and selling for investors.

3. **Data treatment**

To appropriately run a prediction model, the stream data needed to be pivoted, and additional columns such as open exchange rate on the day, as well as highs and lows needed to be queried into the dataset. The query below accomplishes the same and creates a new table to be used for modeling.

```
CREATE OR REPLACE TABLE
 `bd-final-project.coinbase_data.stream` AS
WITH time_plus_unpivot AS
(
 SELECT currency, exchange_rate, new_time
 FROM
 (
  SELECT BTC, ETH, USDT, BNB, BUSD, XRP, DOGE, ADA, MATIC, DOT, DAI, LTC, TRX, SHIB, SOL, UNI, AVA
X, WBTC, LINK, XMR, ATOM, TON, ETC, XLM, BCH, CRO,
  ALGO, APE, TIMESTAMP(timestamp) AS new_time
  FROM `bd-final-project.coinbase_data.streaming_data`
 ) AS p
```

```
 UNPIVOT
 (exchange_rate FOR currency IN (BTC, ETH, USDT, BNB, BUSD, XRP, DOGE, ADA, MATIC, DOT, DAI, LTC,
TRX, SHIB, SOL, UNI, AVAX, WBTC, LINK, XMR, ATOM,
 TON, ETC, XLM, BCH, CRO, ALGO, APE)
 ) AS unpvt
 ORDER BY currency, new_time DESC
)

SELECT currency, EXTRACT(DATE FROM new_time) AS date_extract,
exchange_rate, MAX(open_temp) OVER(PARTITION BY currency, date_) AS open, high, low, new_time,
((high-low)*0.382+low) AS fibo_level_1, ((high-low)*0.5+low) AS fibo_level_2, ((high-low)*0.618+low)
AS fibo_level_3, ((high-low)*0.75+low) AS fibo_level_4
FROM
(
 SELECT currency, exchange_rate, high, low,
 CASE WHEN open_flag IS NULL THEN exchange_rate
 ELSE 0 END AS open_temp,
 date_, new_time
 FROM
 (
  SELECT currency, exchange_rate,
  MAX(exchange_rate) OVER(PARTITION BY currency, date_) AS high,
  MIN(exchange_rate) OVER(PARTITION BY currency, date_) AS low,
  LAG(currency) OVER(PARTITION BY currency, date_ ORDER BY new_time) AS open_flag,
  date_, new_time
  FROM
  (
   SELECT currency, exchange_rate, EXTRACT(DATE FROM new_time) AS date_, new_time
   FROM time_plus_unpivot
  )
 )
);
```

4. **Modeling**

   The goal of the project is to predict the close volume and the close exchange rate for a given crypto currency.

   Linear and boosted tree models were used to run predictions for the required variables. The model evaluation is below.

   Linear Model for Volume Prediction

Mean absolute error	54,448,883,154.7131
Mean squared error	3.137E+24
Mean squared log error	1.2828
Median absolute error	356,270,648.6463
R squared	0.0059

Boosted Tree Model for Volume Prediction

Mean absolute error	42,861,256,666.3831
Mean squared error	3.1573E+24
Mean squared log error	1.6115
Median absolute error	271,168,904
R squared	-0.0006

Linear Model for Close Prediction

Mean absolute error	22.6399
Mean squared error	11,106.2911
Mean squared log error	0.0002
Median absolute error	0.0511
R squared	0.9999

Boosted Tree Model for Close Prediction

Mean absolute error	129.8037
Mean squared error	347,401.6136
Mean squared log error	0.0146
Median absolute error	0.3516
R squared	0.9955

The Linear Model is selected for Volume prediction and the Boosted Tree Model is selected for Close prediction.

5. **Recommendations**

Product Marketing

Market the product to crypto investors and users of Coinbase as a lower risk alternative to market intelligence for crypto investing. Augment the reliability of the prediction with Fibonacci retracement levels to validate and add value to the product.

Product Improvement

Model experimentation – Implement models such as LSTM that are known to work well with real-time predictions to improve the quality and reliability of predictions.

Data improvements – Develop data collection quality to improve additional variables in the real time stream to enhance the model's capability.

6. **Visualization**
   - **Looker Studio link:** https://datastudio.google.com/reporting/59719bab-af88-4dfb-b5e5-ae6ae6e8a0ac