Report Title	Mid-way Report – Application of cloud technologies
Name	Vishal Padwal, Vipul Popat, Margarita Tsekvava.
student ID number	21268168, 21267549, 21267404
email address	vishal.padwal2@mail.dcu.ie, vipul.popat2@mail.dcu.ie, margarita.tsekvava2@mail.dcu.ie
program of study	MCM - M.Sc. in Computing - Blockchain
module code	CA687I
date of submission	09/02/2022
word count	781

An assignment submitted to Dublin City University, School of Computing for module CA687 Cloud Systems.

I understand that the University regards breaches of academic integrity and plagiarism as grave and serious. I have read and understood the DCU Academic Integrity and Plagiarism Policy. I accept the penalties that may be imposed should I engage in practice or practices that breach this policy.

I have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations, paraphrasing, discussion of ideas from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references.

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work. By signing this form or by submitting this material online I confirm that this assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study. By signing this form or by submitting material for assessment online I confirm that I have read and understood DCU Academic Integrity and Plagiarism Policy available here.

Name: Margarita Tsekvava,Vishal Padwal, Vipul Popat

Date: 16 February 2022

TABLE OF CONTENTS

MID-WAY REPORT – APPLICATION OF CLOUD TECHNOLOGIES

DATA SET:

For this assignment, we have chosen the cryptocurrency's historical data ranging from 2019 to 2021, which shows the per minute trade history of the following cryptocurrencies: ADA, BCH, BNB, BTC, DOGE, EOS, ETC, ETH, IOTA, LTC.

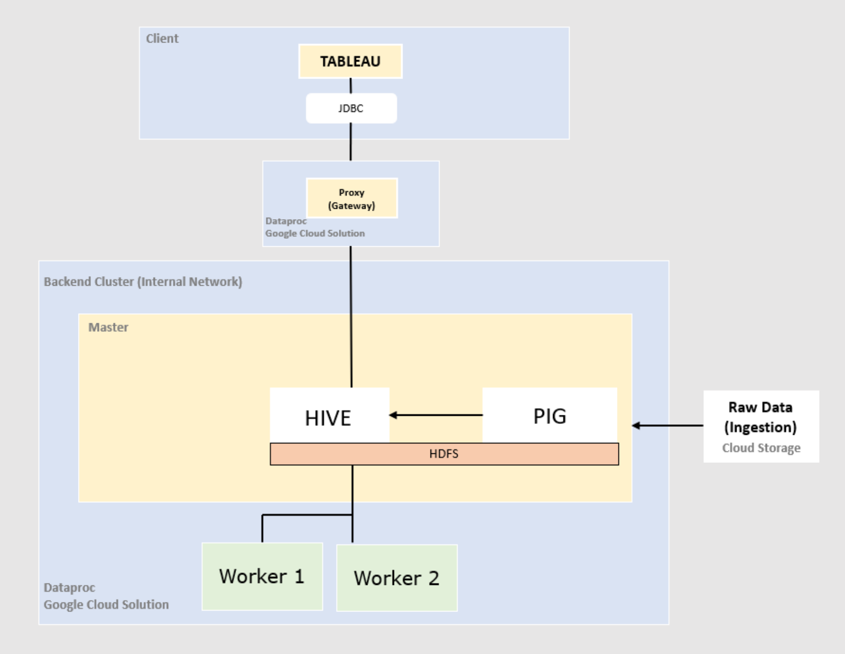
Raw Dataset URL: https://www.kaggle.com/lucasmorin/raw-crypto-1m-data-from-binance

PREFERRED TECHNOLOGY:

- Platform: Google Cloud Provider (GCP)
- Dataproc: A service package offered by Google for Hadoop Cluster
- Hadoop ecosystem (Yarn, MapReduce, HDFS)
- Pig and Hive framework
  - To perform ETL which is nothing but Extract -> Transform -> Load
  - Pig is used to read data sets, transform data and load it into Hive for further analysis.
  - The Hive is used for query & reporting data communication with visualization software Tableau to produce the analytics.
- SQL Cloud: To store Hive metadata Externally
- Cloud Storage: To store all raw Data (Note: HDFS is part of the Hadoop ecosystem and not to be mixed with raw data storage)
- Tableau – for visualisation

  Note: Pig and Hive framework were preferred over Storm and Spark since we are analysing static historical data, not the Real-time streaming data.

Proposed Architecture:



ANALYTICS:

The goal is to analyze the insights from the data set for the specified currencies to depict the entire picture of trading history, including the highlights listed below.

- Annul Asset Volume per year, quarter, Month, day, time & currency
- Highest and lowest price values year, quarter, Month, day, time & currency
- Open and close price values year, quarter, Month, day, time & currency
- Currency Trend with respect to number of trades
- Market trend for all or individual currencies.

A different approach entails analysing and describing the financial performance of cryptocurrencies from a statistical standpoint.

These approaches demonstrate that it is possible to categorize cryptocurrencies based on their financial performance.

This assignment aims to propose a new methodology that will assist us in organizing and understanding the market's main trends at a glance, based on the financial behavior of cryptocurrencies.

Based on the data presented above, we can deduce market trends, peak trading volumes, and the performance of individual cryptocurrencies.

ANALYTICS - DATA TRANSFORMATION

Row Data:

- **Open_time –** Trade Opening timestamp for the minute covered now
- **Open** - The USD price at the beginning of the minute
- **High** - The highest USD price during the minute
- **Low** - The lowest USD price during the minute
- **Close** - The USD price at the end of the minute
- **Volume** - The number of crypto asset units traded during the minute
- **Close_time –** Trade Close timestamp for the minute covered
- **Quote_asset_valume** - The volume-weighted average price for the minute
- **Number_of_trades –** Number of trades for the minute covered
- **Takere_busy_base_**
- **Taker_buy_quote_**
- **Ignore -**

After Transformation - Populate Hive table with following data:

- **currency** - cryptocurrency name
- **trade_date –** date of the trade
- **open_time –** trade open time of the day
- **close_time –** trade close time of the day

- **open_price –** The USD price at the first minute of the day
- **close_price –** The USD time at the last minute of the date
- **volume –** The number of crypto assets units traded during the day
- **high_price –** The highest USD price during the day
- **low_price –** Lowest USD price during the day
- **trade_number –** Sum Number of trades of the day

RELATED WORK

Using Hive and Spark with DataBrick was used in the financial company where we work mainly used by equity traders for analysing the data, trades, pattern's and etc.

CHALLENGES AND LESSONS LEARNED

- The backend connection to the front end was the most challenging part – proxy set up and JDBC connection
- Because of the large amount of data in the collection, deciding what to visualize in the limited time allotted for the task was difficult.
- All technologies were new for us, so some learning curves were required.
- If would have more time, we would use hive partition for better performance. In addition, in the case of a longer time frame, more dashboard visualization might be supplied.

ROLES AND TASK'S:

No.	Task's	Contributors	Status
1	Requirement gathering- Dataset finalisation	MT, VP1, VP2	Done
2	Design and Architecture proposal	MT, VP1, VP2	Done
3	System design, Using the established requirements	MT, VP1	Done
4	Data insights and analysis	MT, VP1	Done
5	Backend Development	MT	Done
6	Pipeline to established connectivity	MT, VP2	Done
7	Front End Development	VP1	Done
8	Visual Presentation - Video	VP1, MT, VP2	Done
9	Final report	MT, VP1, VP2	Done

MT – Margarita Tsekvava, VP1 – Vishal Padwal, VP2 – Vipul Popat

RESPONSE TO PEER FEEDBACK

The Peer feedback was crisp, fair and professionally drafted.

GITLAB REFERENCE

https://gitlab.computing.dcu.ie/popatv2/group1-ca687i-assignment1