ETHEREUM TRANSACTION ANALYSIS WITH BIGQUERY

Abhishek Kesiraju
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
kesiraj@uwindsor.ca

Kalyan Ram Thandu
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
thanduk@uwindsor.ca

Vikram Kumar Kovvuri
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
kovvuri@uwindsor.ca

***Abstract*—Ethereum often considered the sister cryptocurrency for Bitcoin, is the technology that allows people to send cryptocurrency on the world's decentralized public blockchain network. As a result of the pandemic, there has been a wide rise in the demand and number of transactions over these past few years. There are a plethora of sectors for which this coin is creating utility and value such as virtual Real Estate, Healthcare, Finance, Entertainment, etc. There were previously performed Bitcoin and Ethereum analysis using traditional bigdata processing applications such as spark using python. The datasets on which the analysis is performed are limited in size and time. The project aims to perform KDD exploratory data analysis on this "proof of work" blockchain Ethereum transaction data using Big Query. Big Query allows us to perform these operations on this data with standard SQL like syntax and visualize insights with its publicly available cryptocurrency datasets. Big Query is a Google enterprise's cloud data warehouse system that enables us to perform analysis over petabytes of data as a PaaS with support to ANSI SQL**

Keywords—Big Query, Ethereum, Exploratory Data Analysis, Visualization.

I. MOTIVATION

The decentralized Web3 structure use cryptocurrency as its monetary services and Ethereum is the sister cryptocurrency to bitcoin. Blockchain analysis is performed to understand the varying costs of cryptocurrency and to find insights from transactions. The use of Big Query enables us to perform this data exploration with easily optimized queries since this was a long and exhaustive analysis previously performed on Bitcoin using Apache Spark. Data on Web3 over the past 3 years quadrupled and the transactions became innumerable with the introduction of the metaverse and its monetization using ETH.

II. INTRODUCTION

Ethereum being the technology that offers people to send cryptocurrency on the world's decentralized public blockchain has taken advent in transactions recently. it is the second most popular and purchased cryptocurrency in the world. There is a significant rise in terms of the number of these transactions. Google defines "Big Query as a serverless, highly scalable data warehouse that comes with a built-in query engine. The query engine can run SQL queries on terabytes of data in a matter of seconds, and petabytes in only minutes". There is a significant rise in terms of the number of these transactions. The project aims to perform KDD analysis over Ethereum transactions public data using Google's Big Query. It also aims to visualize the findings in a human-understandable format.

III. BACKGROUND STUDY

Block refers to data and states being stored in consecutive groups, a blockchain is a public database updated and shared across the internet. A submitted blockchain transaction for example our Ethereum has the following information:

- Recipient: contains the receiving address.
- Signature: address of the sender
- Value: the amount of ETH
- Data: optional field that includes arbitrary data
- Gaslimit: Units of gas represent the computational resources consumed by the transaction.
- Maxpriorityfeepergas: maximum amount of gas included as a tip.
- maxFeePerGas: maximum amount of gas willing to be paid for the transaction.

'tx' unit represents the signed transaction in JSON form.
A state in Ethereum is made up of objects called accounts. Each of the accounts holds a 20-byte address that contains state transitions which are transfers of value and information from accounts. In general, an account possesses the following four properties.

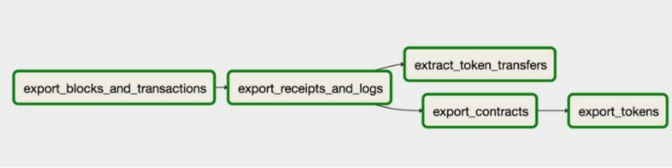
- A nonce, it represents a counter to make sure each transaction is only used once
- The account's current ether balance
- Account's contract code if present.
- The account's storage.

The term transaction in the blockchain contains a signed data package. It typically contains the receiver information, a message by the sender. The amount of ether transferring between the sender and the receiver. An optional data field, with a StartGas price which sets the number of computational steps a transaction could take, and a Gas Price a transaction fee set up for the computational steps.

Ethereum allows developers to create decentralized applications (apps) which share a pool of computing power. Damp has backend code running on a decentralized peer-to-peer network. Transactions over these networks follow consensus mechanisms. These mechanisms are used to maintain a consensus among databases, application servers, nodes, etc. Ethereum like the Bitcoin uses a '**Proof of work'** consensus protocol. Proof of work is the concept of updating the network chain once a miner successfully reaches the end of the block and hence the mining is present in the work itself. As the usage of the ETH network grows, there is an extreme increase in important information on the chain. With this volume of data, aggregating data to report or drive a decentralized application can become a heavy time-consuming endeavor.

This cryptocurrency uses Distributed Ledger Technology (DLT) refers to the approach of sharing data between multiple data stores ledgers. [1] Each node recorded can be mined and updated, so that the ledger increases. The concept of cryptocurrency needs us to understand a few concepts that are important and not to overlook when querying the Ethereum chain.

When looked at in closed detail, a single node contains a lot of further information. In general, it can be classified into three types: A **world state trie**, an **account storage trie**, a **transaction trie, and a receipt trie**. World State Trie gathers all the information which is mapped between accounts and address states. It can be referred to as a global state that is constantly updated by the transaction executions and there is an account state containing information about the account and all the smart contracts possessed by the account. The transaction trie has all the information mentioned above as transaction relevant.
Ether value transfers are precise and direct. The addresses can not only hold balances but also contain the smart contract bytecode. The transaction trie mentioned contains the hash of the contracts present. A successful Ethereum transaction follows the steps: Firstly, a transaction is created. The transaction creator attains an ID or a Hash ID. This can be used to look upon by anyone, the transaction then has to be broadcasted to the nodes in the network and waits for the miner to pick, and verify which then adds a ledger as discussed before. The gas fee set up for the transaction influences this transaction broadcast to the network. All the extract, transform and load operations from the DAG are organized to form the dataset with multiple tables held in the Ethereum Classic dataset which is updated every day by Google.



PROPOSED MODEL

The Ethereum transactions dataset can be obtained from the Google Public Datasets which are found in the google cloud marketplace. On the contrary, we can manually pull data from the Ethereum official website which provides the API from a JSON compute engine. However, organizing data from the Ethereum blockchain into a local desktop or a cloud space is not efficient memory usage.
First and foremost, once we have the Ethereum data, we created a set of business questions that can be answered by transforming and manipulating data to our needs. We have framed 6 business questions, but the project is not limited to them and will continue. They are:

- **What are the topmost valued recent transaction wallets that were performed on the Ethereum chain?**
- **What was the average transaction gas fee over time? When was it the highest?**
- **How much has been the active growth of addresses in the Ethereum network over time?**
- **What is the average value of ether over time?**
- **What was the value of ether transferred per day in the chain in total?**
- **How many successful or unsuccessful are transactions that took place in Ethereum?**

Google Data Analytics considers the approach of answering business questions in the flow Ask, Prepare, Process, Analyze, Share and Act. The data analytics life cycle hence followed is

1. Ask: Asking the right questions, these questions act as problem statements for which a strategy must be made.
2. Prepare: Gather the sources, files, and all requirements for performing preprocessing steps.
3. Process: Involves creation, and transformation of data. Maintain data integrity, test, verify and report on results.
4. Analyze: Use tools to sort, filter, identify patterns and draw conclusions to make data-driven decisions.
5. Share: Bring data to life, and use storytelling to apply visualizations. Communicate to help others understand it to others.
6. Act: Apply insights to solve problems.

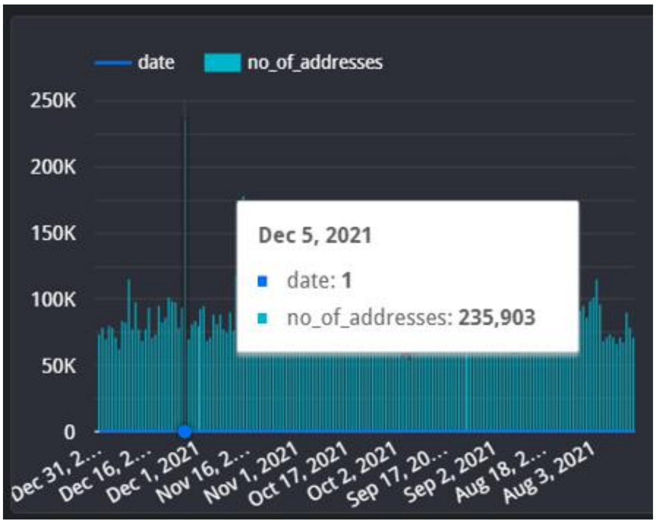
Hence whilst using NumPy to load only a limited space for each question, we have built the queries for each question with an iterative approach to tuning them for utilizing low space.

Active Growth in Addresses

Active growth in addresses refers to the number of users creating their first account and making a transaction in the Ethereum chain. While writing the query we included consider the value of users or smart contracts performing their first transaction more than 0. For the query to not cross more than 50 Gb, the date is put up from 2019 or changed to a further recent time.
Active growth in Addresses (Accounts creation and their first transaction in ETH blockchain) was highest on Dec 05, 2021 - 235,903

```
SELECT first_tx AS date, COUNT(*) AS no_of_addresses
    FROM (
        SELECT from_address AS address,
MIN(DATE(block_timestamp)) AS first_tx
        FROM `bigquery-public-
data.crypto_ethereum.transactions`
        WHERE value > 0 AND block_timestamp < '2019-01-
01 00:00:00'
```

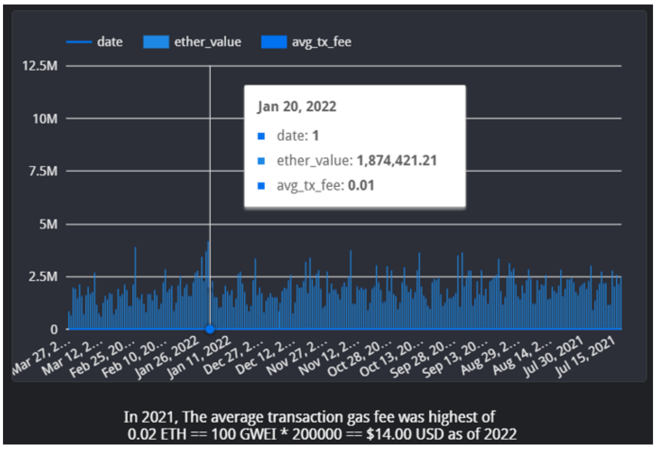
```
    GROUP BY address
    ORDER BY first_tx
  )
  GROUP BY first_tx
  ORDER BY date desc
```



Average Transaction Cost

The transaction cost in Ethereum is obtained by multiplying the amount of gas used with the gas price. One interesting point is that the ether balance is attained in the form of Wei which is the lowest measure of cryptocurrency. To attain the amount of ether, Wei should be raised to the power of 18.

```
SELECT SUM(value/POWER(10,18)) AS
ether_value,DATE(block_timestamp) as date,
AVG(gas_price*(receipt_gas_used/POWER(10,18))) as
avg_tx_fee
FROM `bigquery-public-data.crypto_ethereum.transactions`
WHERE DATE(block_timestamp) BETWEEN '2018-01-01'
AND '2022-03-27'
GROUP BY date
ORDER BY date;
```

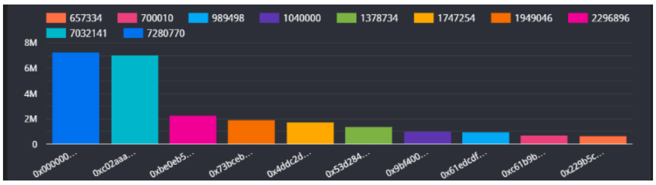


In 2021, The average transaction gas fee was highest of
0.02 ETH == 100 GWEI * 200000 == $14.00 USD as of 2022

Top 10 Ether balances in Wallets

Though we were able to make a simple query from the transactions table which has the highest balance, a developer named Allen at google contributed an efficient query for retrieving the top balances with a large number of dependencies as seen below

```
  SELECT to_address AS address, value AS value
  FROM `bigquery-public-data.crypto_ethereum.traces`
  WHERE to_address IS NOT null
  AND block_timestamp < '2021-09-01 00:00:00'
  AND status=1
  AND (call_type NOT IN ('delegatecall', 'callcode',
'staticcall') OR call_type IS null)

  UNION ALL

  SELECT from_address AS address, -value AS value
  FROM `bigquery-public-data.crypto_ethereum.traces`
  WHERE from_address IS NOT null
  AND block_timestamp < '2021-09-01 00:00:00'
  AND status=1
  AND (call_type NOT IN ('delegatecall', 'callcode',
'staticcall') OR call_type IS null)

  UNION ALL

  SELECT miner as address,
SUM(CAST(receipt_gas_used AS NUMERIC) *
CAST(gas_price AS NUMERIC)) AS value
  FROM `bigquery-public-
data.crypto_ethereum.transactions` AS transactions
  JOIN `bigquery-public-data.crypto_ethereum.blocks` AS
blocks
  ON blocks.number = transactions.block_number
  WHERE block_timestamp < '2021-09-01 00:00:00'
  GROUP BY blocks.miner

  UNION ALL

 SELECT from_address as address, -
(CAST(receipt_gas_used AS NUMERIC) *
CAST(gas_price AS NUMERIC)) AS value
  FROM`bigquery-public-
data.crypto_ethereum.transactions`
  WHERE block_timestamp < '2021-09-01 00:00:00'
)
SELECT address, FLOOR(SUM(value) / power(10,18)) AS
balance
FROM value_table
GROUP BY address
ORDER BY balance DESC
LIMIT 10
```

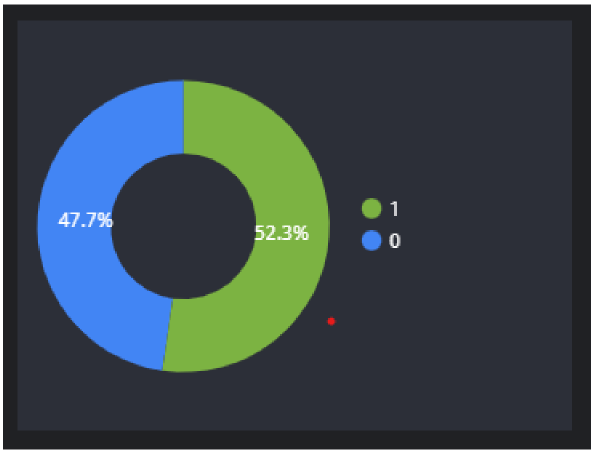


Checking whether a Transaction is Successful or not

This is achieved using the receipt status of a trace in the transaction. A trace of a transaction returns the aggregated summary of post-processing for the requested calls of the transaction. The receipt status if 1 is a success and 0 if not.

The percentages since 2021 were 52% for Successful transactions and 47.7% for Failed Transactions.

```
select
   b.block_timestamp,
   transaction_hash,
   status
from
   (
      select
         block_timestamp,
         `hash`
      from
         bigquery-public-data.crypto_ethereum.transactions
      where true
         and date(block_timestamp) > '2021-01-01'
   ) a
   join
   (
      select
         block_timestamp,
         transaction_hash,
         status
      from
         bigquery-public-data.crypto_ethereum.traces
      where true
         and date(block_timestamp) > '2021-01-01'
         and trace_address is null
   ) b
on a.hash = b.transaction_hash
```

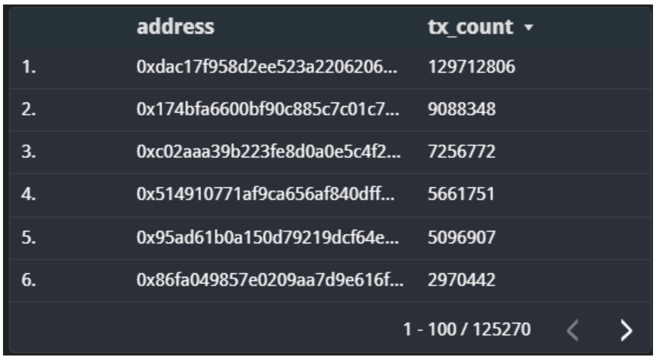


Average Ether Transferred per day
The amount of ether has been increasing over time and it is important to understand the average of it which is being spent transacted on the Ethereum chain per day. Hence this is obtained from the query

```
SELECT
   DATE (block_timestamp) AS date, (SUM(value) /
power(10,18)) AS total_value
FROM
   `bigquery-public-data.crypto_ethereum.transactions` AS
transactions
WHERE
   block_timestamp < '2022-04-01 00:00:00'
GROUP BY date
ORDER BY date
```

Maximum number of transactions done by a person
The tx_count allows us to find the number of transactions done by a person, the highest number of transactions done by an address since 2021 was found to be 129712816. Another interesting feature that we added as a fun fact is the address although has the highest number of transactions has only 1 Ether left now.

```
SELECT contracts Address, COUNT(1) AS tx_count
FROM `bigquery-public-data.crypto_ethereum.contracts`
 AS contracts JOIN `bigquery-public-
data.crypto_ethereum.transactions` AS transactions ON
(transactions.to_address = contracts.address)
WHERE contracts.is_erc20 = TRUE AND
contracts.block_timestamp < '2021-09-01 00:00:00'
GROUP BY contracts.address
ORDER BY tx_count DESC
```



	address	tx_count ▾
1.	0xdac17f958d2ee523a2206206...	129712806
2.	0x174bfa6600bf90c885c7c01c7...	9088348
3.	0xc02aaa39b223fe8d0a0e5c4f2...	7256772
4.	0x514910771af9ca656af840dff...	5661751
5.	0x95ad61b0a150d79219dcf64e...	5096907
6.	0x86fa049857e0209aa7d9e616f...	2970442

REPOSITORY LINK

The manual analysis performed with Python and NumPy to reach answers to the business questions framed above can be found at
https://github.com/Vikram-1798/Eth-bigquery-analysis

LIMITATIONS AND CHALLENGES

Previously performed Bitcoin and Ethereum analyses were performed in traditional big data processing applications such as Spark. With the usage of Big Query, data exploration is highly eased on a recent dataset of approximately 300 GB. Analysis can be performed using the 1Tb free computing storage provided by the Google cloud.

The pipeline made by google with Ethereum ETL makes it easy for the project which has the transaction data updated every 24 hours to make accurate exploration and predictions. This pipeline has two parts, first is the Ethereum Compute engine, an open-source tool that was developed to convert the blockchain into a JSON RPC interface that connects to an Ethereum node, and the exportation of files into Big Query using API. These two phases are coordinated by Google Cloud composer, a workflow orchestration service powered by Apache airflow.

When performing manual analysis with python and NumPy for reaching answers, we had to save the tables created for which we either had to store the data somewhere on the cloud or locally. For instance, when saving the table in the form of a local CSV, the console allows only up to 10 Mb of data i.e.

16000 rows. If the file has to be stored in the form of a google drive, the console allows storage up to 1 GB. To leverage the whole power of the Ethereum dataset, a big query requires pricing to be set up for using the API so that the 'bq helper' module can be utilized for data manipulation operations.

Hence manual analysis operations that were performed to get answers to the questions framed were limited to the amount of space considered.

CONCLUSION AND FUTURE WORKS

Now that we have understood what are the factors that exist in the Ethereum node, the project will continue to find correlating factors that influence the price of Ethereum. We leveraged Google Data Studio to create a dashboard to present all the visualizations. Further studies with Ethereum price prediction analytics are being conducted with comparison of different algorithms like RNN and LSTM networks. Furthermore, the prediction can be compared with accuracy with manual analysis to Google's ML engine.

REFERENCES

[1] Pinna, Ibba, S., Baralla, G., Tonelli, R., & Marchesi, M. (2019). A Massive Analysis of Ethereum Smart Contracts Empirical Study and Code Metrics. IEEE Access, 7, 78194–78213. https://doi.org/10.1109/ACCESS.2019.2921936

[2] Chen, T., Li, Z., Zhu, Y., Chen, J., Luo, X., Lui, J., … Zhang, X. (2020). UnderstandingEthereum via Graph Analysis. ACM Transactions on Internet Technology, 20(2), 1–32. https://doi.org/10.1145/3381036

[3] BigQuery Blog: https://cloud.google.com/blog/topics/developers-practitioners/bigquery-explained-blog-series

[4] GoogleDataStudio: https://blog.hubspot.com/marketing/google-data-studio

[5] Ethereum.org:https://www.preethikasireddy.com/post/the-architecture-of-a-web-3-0-application