Disclaimer

A report submitted to Dublin City University, School of Computing for module CA687I – Big Data Cloud, 2023.

We understand that the University regards breaches of academic integrity and plagiarism as grave and serious.

We have read and understood the DCU Academic Integrity and Plagiarism Policy. We accept the penalties that may be imposed should we engage in practice or practices that breach this policy

We 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.

We declare that this material, which we now submit for assessment, is entirely our 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 we confirm that this assignment, or any part of it, has not been previously submitted by us 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 we confirm that we have read and understood DCU Academic Integrity and Plagiarism Policy (available at: http://www.dcu.ie/registry/examinations/index.shtml)

Name(s): Vinit Saini, Marcio Vieira, Krystian Fikert (Group A)
Date: 25/03/2023

Exploring the Non-Fungible Token Revolution: An Analysis of NFT Transactions

Introduction and motivation

Welcome to the world of Non-Fungible Tokens (NFTs)! As the digital revolution continues to shape the future, NFTs are emerging as a significant force in the digital asset market, offering unprecedented opportunities and potential applications that could revolutionize how we think about asset ownership and value.

The aim of our project is to investigate the NFT revolution by examining the transactions and trades that occurred between 2021 and 2023. We have thoroughly scrutinized and analysed the millions of NFT transactions carried out on various blockchains. Our primary objective was to identify financial and quantitative patterns in order to uncover valuable insights into the NFT market, which was estimated to be worth approximately USD 15.54 billion in 2021, according to Emergen Research.

Our analysis reveals several compelling metrics that offer valuable insights regarding purchase/sale activity, profit/loss margins, average transaction values, fluctuations in buy/sell prices, top-performing NFTs, collections, buyers, sellers, minters, and other noteworthy measures.

Undertaking this project was a fantastic chance for us to delve deeper into the NFT revolution and witness how it is reshaping the digital asset landscape. It is an exciting time to explore the world of NFTs and witness the constant progress and innovation. Our aim is for this project to offer valuable insights and inspire others to participate in this ground-breaking movement.

Data

In our project, obtaining accurate and high-quality data was crucial to extract useful insights. Initially, we utilized a Kaggle dataset (8 GB) that provided valuable information about the nature of NFT transactions, such as the addresses of NFTs, tokens, and wallet addresses. However, it lacked essential information on crucial aspects such as transaction volumes, buyer, seller and minters information for the NFTs. We also needed data on collection names and values to identify patterns and trends in the market and track the value appreciation over time.
To enrich the dataset, we utilized publicly available Moralis APIs to gather additional data based on the initial dataset. Despite the number of restrictions on API calls, we were able to accumulate approximately 137 GB of additional data in just a few weeks. This decision turned out to be the right one, as it provided us with the comprehensive data we needed to derive more meaningful insights and enrich our initial dataset many folds.

Following are the datasets that we accumulated

Datasets **Sourced from different public platforms**			
1.	Kaggle	8 GB	General NFTs related attributes
2.	Moralis	110 GB	Data specific to ERC721 contracts which contains deep information on NFTs, Types, Names and Metadata
3.	Moralis	27 GB	Specific to NFT transfers i.e buyers, sellers, transaction value etc.

Analytics

The analysis for our project involved multiple steps including preparation, processing, querying, and storing of vast and intricate data sets efficiently. The unstructured nature of data and inconsistency was significant issue due to the non-standardization of attributes and names in these datasets. Another challenge with the data set was its diverse and disparate nature, which made traditional methods of analysis ineffective. We developed Python scripts to intelligently and effectively read and parse data from various sources to overcome this challenge. Additionally, we implemented data

cleaning and sanitization techniques to make the data usable for analysis and report generation. By carefully selecting and applying appropriate data transformation techniques, we converted the data into a standardized and consistent format, making it easier to analyse and compare across different sources. These techniques included data aggregation, filtering, transformation, and normalization. Ultimately, by leveraging the power of Spark, Python and other data processing tools, we overcame the challenges posed by the diverse and disparate nature of the data set, enabling us to derive meaningful insights from the data.

To address these challenges, we leveraged GCP's Google DataProc. This end-to-end big data analytics platform provided an efficient and swift way to process and analyse large volumes of data using Spark. Utilizing this cloud-based environment, we collaborated effectively while maintaining data security and efficient processing.

Here is the list of toolsets that we used from the Google Cloud Platform (GCP)

- DataProc (Clusters, Jobs, Workflows)
- MySQL managed services
- Cloud Storage (Bucket)
- Compute Engine (Virtual Machine Instance)
- Metrics Explorer
- Logs Explorer
- IAM and admin

These tools provided a secure and robust environment for data analysis and exploration of cloud services, enabling us to gain valuable insights from the data.

Additionally, DataProc's ability to scale quickly and efficiently meant that we could handle large volumes of data without compromising performance or accuracy. Using the GCP and Google DataProc enabled us to overcome the challenges of analysing diverse and disparate data sets, allowing us to derive meaningful insights to inform our decision-making processes.

We took advantage of Spark's APIs for programming in Java, allowing us to customize our data analysis according to the needs. By utilizing the scalability and flexibility of the cloud, DataProc enabled quick analysis at a cost-effective rate, allowing us to create a cluster in the cloud and run complex analytics Spark jobs.

In addition, we leveraged GCP's managed MySQL service as a reliable storage solution for both our raw and processed data. We utilized it to maintain tables containing the raw data sets, which served as the primary source of information for our Spark jobs. Similarly, we also used it as a destination for storing the processed information under different schema. This processed data was then queried as per our requirements to generate reports and visualizations. Further, To optimize data retrieval speed and efficiency, we created views and implemented indexes. Together, these optimizations made our data retrieval process faster and more efficient.

In order to select the most suitable tool for the presentation layer, we assessed several options, including Tableau, Appsmith, Qlik and Datapine considering factors such as efficiency, user interface, and ease of collaboration. After careful evaluation, and rounds of implementations we found that Datapine Cloud was particularly compelling, as it demonstrated strong capabilities for collaboration, user-friendliness, and efficient handling of large datasets.

Insight

In this section, we present a comprehensive analysis of NFT transactions, including key metrics that shed light on the trends and patterns in the market. We have analyzed a vast dataset consisting of millions of transactions to derive insights into the volume, value, and popularity of NFTs.

Here, we discuss the key metrics that we have produced on the basis of NFT data. By analyzing these charts, investors and collectors can gain a better understanding of the NFT market and make informed decisions about which NFTs to buy or sell.

Collection View **Provides an overview of an NFT collection**		
1.	Number of NFTs	Provides information about the size of an NFT collection.
2.	Number of trades	Indicates the popularity of an NFT collection by showing how many times its NFTs have been traded.
3.	Total value of trades (in ETH)	Represents the total market value of an NFT collection, which is typically produced by an artist or organization.
4.	Min and max trade value	Gives insight into the cheapest and costliest NFTs in the collection.
5.	Average value	Provides information on the average NFT value in the collection.
6.	Variance	Offers insights into the variability or diversity of the NFT collection and can be used to assess its stability.
7.	Timestamp of first transfer and last transfer	Helps to calculate the trading potential of the NFTs in the collection by providing insights into when the first and last trades occurred.
8.	Value of first transfer and last transfer	Indicates the value of the NFTs during the first and last trades and helps to calculate the profit or loss of the NFT collection.

Minters View **Provides insights into NFT content originators**		
1.	Number of NFTs minted by a minter	Shows the biggest minters in the NFT world, which could potentially indicate a monopoly.
2.	Valuation of minters	Provides information on the value of the minters, who can be individuals or organizations.

Traders View **Provides insights into entities involved in NFT trading**		
1.	Top Buyers	Offers insight into the most active NFT buyers.
2.	Top Sellers	Offers insight into the most active NFT sellers.
3.	Top Traders	Provides information on the most active people or organizations in the NFT trading world.

Contract View **Provides insights into preferred ways of NFT trading**		
1.	Number of trades per contract type (ERC)	Offers insights into the preferred contract types among NFT traders.

Related Work

Our goal was to analyze the market structure, volatility, top NFT characteristics, and future trends. During our exploration, we discovered fascinating research and applications, including the noteworthy paper "When Big Data Meets NFT: Challenges, Impacts, and Opportunities" (Chen et al., 2023), cited in our reference list. From an application standpoint, we were using https://opensea.io/ (OpenSea) to review NFT collections, pricing, and buyers' and sellers' behaviour.

Challenges and lessons learned

The mix of structured and unstructured data

Incorporating structured and unstructured data into our analysis was challenging. We used Kaggle's structured data but needed Moralis API to obtain unstructured data like NFT Contracts information and NFT Token Transfers, and we collected market data from OpenSea. However, processing such diverse datasets was challenging, especially in correlating the data effectively.

We tried to contact OpenSea several times to request their data set, but there was no reply.

Limitation of resources in free account

Configuring Dataproc to meet our computing needs was challenging due to limitations with available cluster configurations. Our large dataset containing 36 million rows and complex JSON attributes required more processing power than the e2-standard-4 instances with 4vCPU and 16GB memory could provide. Despite our attempts to cap usage, we were unable to increase the cluster size due to quota limitations. However, we found creative solutions to optimize processing power and work within the available resources.

Due to the number of restrictions associated with the Dataproc free account, we had to set up Spark on our machines to overcome data processing challenges.

Responsibility statement

Our team of three members has various backgrounds, and we have used "The Nine Belbin Team Roles" (Belbin) to define our roles and responsibilities.

Name	Role	Tasks
Marcio	Investigator, Completer	- Finding the data - Supplementing data and optimization - Setting up and configuring the data processing infrastructure - Development for fetching data from external APIs - Evaluating Public API for fetching the NFT data - Database level optimizations - Brain-storming - Presentation
Vinit	Shaper, Implementer	- Conducting exploratory data analysis - Defining and calculation of metrics - Cleaning and sanitization of data - High level design - Spark Development and running data processing jobs - Code optimizations - Resource negotiations to run Spark jobs successfully - Brain-storming - Final report
Krystian	Team worker	- Brain-storming

Link to the codebase and other artifacts

Code	GitHub - vsdcu/nft-data-processor
Dashboard	Datapine.com/ - NFT-Market Dashboard
Midway-Report	GitHub - vsdcu/assignment-reports/mid-way-report

Response to feedback

Good problem spec and level of complexity.

Thanks for the assessment of our initial idea.

Be clearer on what you will use each technology for, and why: e.g. do you need cleaning? why? Do you need querying? why? Are you going to use Spark for processing incoming minibatches or are you going to use SparkSQL or SparkML for prediction?

Yes, we did a lot of cleaning and sanitizing of data, this was indeed required as we collected data from various sources and that too was quite unstructured and disparate in nature. We developed python scripts to parse the raw data into the required format. We did try parsing it through big data tools but in the end, we found python scripts most efficient for the job.

Yes, our reports and visualizations are solely based on querying the processed data persisted in RDBMS tables. Every report has specific aspects of data which need selective queries to the database.

Yes, we have used Spark for processing the raw data to find out the relations and patterns in data. We have used various functions like agg(), join(), filter(), distinct(), orderBy(), groupBy and select() are few of them to uncover hidden insights and information.

Yes, we have used SparkSQL as the data was at rest (static) so using SparkSQL was the ideal choice. We didn't use Spark Streaming in our use case.

Similarly, no-use of SparkML, as our use case does not involve any prediction.

It is not clear at the end whether the prediction of trends and pattern analysis is something you think would be enabled by your exploratory analysis or something you plan on providing insights on via SparkML model training.

The goal of our project was to investigate the historical NFT transaction data and uncover insights about the market bias and overall landscape. Our focus was not on using machine learning for prediction purposes, but rather on creating a dashboard that reports on existing data and helps us better understand the NFT market.

By analyzing the data, we were able to gain a deeper understanding of the NFT market and identify trends that could be useful in making informed decisions. Our analysis provided insights into factors that influence the value of NFTs, such as the popularity of the artist, the rarity of the artwork, and the hype surrounding a particular NFT collection.

Our dashboard includes visualizations of the data, such as charts and graphs, to help us easily identify patterns and trends. By presenting the data in a clear and concise way, we were able to draw meaningful insights from it and make informed decisions about the NFT market.

Overall, our project was focused on using data analysis and visualization techniques to gain a better understanding of the NFT market. By uncovering insights about the market bias and overall landscape, Our goal is to furnish valuable insights to stakeholders in the NFT community, which can assist them in making reasonable decisions about the purchase and sale of NFTs.

A good outline of tasks, just be specific on how you are going to divide such tasks (who is going to do what?) as this is part of your self-assessment of individual participation at the end.

We appreciate the feedback regarding the task outline. Our original plan was to distribute specific tasks among team members based on their knowledge, strengths, and interests. However, due to new challenges in the later stages of the project, our approach had to be adjusted. Detailed task division and plan can be found in the appendix section.

Bibliography

Chen, Q., Guo, J., Wei, B., Li, B. and Kelly, J.M., (2023) 'When Big Data Meets NFT: Challenges, Impacts, and Opportunities', *International Journal of Information Systems and Social Change (IJISSC)*, 14(1), pp.1-16. Available at: https://doi.org/10.4018/ijissc.314570

Moralis Web3 - Enterprise-grade Web3 apis (2023) *Moralis Web3 | Enterprise-Grade Web3 APIs.* Available at: https://moralis.io/ (Accessed: March 23, 2023).

Non Fungible Token Market, By Category (Collectibles, Utility, Art, Metaverse, Game), By Application (Real Estate, Medical, Academic, Gaming), and By Region Forecast to 2030 (2022) *Emergen Research.* Available at: https://www.emergenresearch.com/industry-report/non-fungible-token-market (Accessed: March 23, 2023).

OpenSea, *OpenSea the largest NFT Marketplace*. Available at: https://opensea.io/ (Accessed: March 23, 2023).

The Nine Belbin Team Roles. *Belbin*. Available at: https://www.belbin.com/about/belbin-team-roles (Accessed: March 23, 2023).

Zomglings (2022) Ethereum NFTs, *On-chain activity from the Ethereum NFT market.* Available at: https://www.kaggle.com/datasets/simiotic/ethereum-nfts (Accessed: March 17, 2023).

Appendix

Tasks planning

S.no	Tasks	Main Contributor	Completed (on-time)	Notes (if any)
1.	Idea generation and agreement	Team	Yes	NFTs market analysis
2.	Midway report	Team	Yes	
3.	Data finding and downloading	Marcio	Yes	Kaggle
4.	Extracting & parsing complementary data from public APIs	Marcio	Yes	Challenging due to quota limits on REST calls
5.	Data cleaning and pre-processing	Vinit	Yes	
6.	GCP account setup	Krystian	Yes	Consumed over-budget
7.	Development and execution of Spark jobs	Vinit	Yes	Quite complex and time consuming, had to run few of them on local machine
8.	Setting up DataProc Cluster and maintenance	Vinit	Yes	Time consuming due to limitation on resources in free account
9.	GitHub code repository	Vinit	Yes	
10.	Generation of custom views to optimize reports	Marcio, Vinit	Delayed	This was unplanned but became necessary to handle large datasets
11.	Visualization, Reporting	Marcio, Vinit	Delayed	Team evaluated Tableau, Qlick, AppSmith and Datapine (worked)
12.	Presentation and Demo	Marcio, Krystian	Delayed	Due to 10, 11
13.	Final Report	Vinit, Krystian	Delayed	Due to 11, 12