Big Data Analysis using BigQuery on Cloud Computing Platform

Md. Husen Ali[1]*, Md. Sarwar Hosain[1], and Md. Anwar Hossain[1]

[1]Department of Information and Communication Engineering, Pabna University of Science and Technology (PUST), Pabna, Bangladesh.

*Correspondence: husenali.ice@gmail.com (Md. Husen Ali, Research Student, Department of Information and Communication Engineering, PUST, Pabna, Bangladesh).

ABSTRACT

End the age of digitalization, data generated from numerous online and offline sources in every second. The Data are having a considerable amount of size and several properties termed as Bigdata. It is challenging to store, manage process, analyze, visualize, and extract useful information from Bigdata using traditional approaches in local machines. To resolve this cloud computing platform is the solution. Cloud computing has high-level processing units, storage, and applications that do not depend on user devices' performance. Many users can access resources and demanded services remotely from the cloud on a pay-as-use basis. That is why users are not needed to buy and install costly resources locally. Some cloud services providers are Google, AWS, IBM, and Microsoft, and they have their Bigdata analyzing robust systems and products in a cost-efficient manner. There are many Cloud Service Providers (CSP's) having different services of Bigdata analyzing filed. However, we discuss in the paper about an excellent service BigQuery in the Data ware house product of Google to analyze and represent numerous samples of datasets in real-time for making the right decisions within a short time.

Keywords: Bigdata, Cloud computing, Services provider of cloud, Google cloud computing, and BigQuery.

1. INTRODUCTION:

With the advancement of modern technologies, in every moment, data are produced through the internet, business organization, web applications, networking, artificial intelligence, scientific research, and social media very rapidly. Generally, these data types have vast amounts of volume, properties, and categories termed as Bigdata. The significant issues are managing, organizing, processing, presenting, storing, and analyzing Bigdata for decision making in future actions (Rahman and Hasibul, 2019).

Extracting data from multiple sources is challenging by traditional analyzing approaches because of cost, infrastructure, high processing units, sophisticated analysis tools, storages, and robust algorithms. In

this case, the cloud platform is a remarkable solution to managing and handling Bigdata quickly and efficiently within a concise time (Khedekar and Tian, 2020).

The research's main goal is to analyze a large volume of structured or unstructured data defined as Bigdata in the Google cloud platform (GCP) using BigQuery by SQL commands. BigQuery is a product of Google that is a wholly overseen, petabyte-scale, cost-effective for analyzing data in less execution time.

2. Related Works

The researchers did multiple numbers of studies on Bigdata analysis in the field of cloud computing. (Trajanov *et al*., 2016) conducted a study using

BigQuery in a platform of cloud computing for research and educational purposes and recommends only how the platform can be utilized in data analysis. (Tomar and Tomar, 2018) present an overview of Bigdata and cloud computing integration from two sources, i.e., RedBus and Twitter. This paper only discusses the data analysis framework and some methods but does not present a better analysis process in detail.

(Kotecha and Joshiyara, 2018) present a method of managing and handling non-rational data on BigQuery and calculating the processing time. This paper only covers the analysis time with the dataset's size using Google SDK rather than extracting the taken dataset's necessary values. (Harsha, 2017) discuss some important roles and analyzing tools of Bigdata regarding cloud comput-ing technology. However, These papers have not represented any analyzing methods of Bigdata in a clear view. (Bathla, 2018) presents a theoretical discussion on Bigdata management tools in the cloud platform rather than analyzing the non-structural data within a concise time.

(Riahi and Riahi, 2018) discuss the fundamentals of Bigdata, its challenges, and its applications of it in data analytics. The paper covers Hadoop use, an open-source framework to manage data from diff-erent sources and analyze it. Our presented papers highlight cloud-based Bigdata analysis techniques using the BigQuery service of google without the infrastructural development and database adminis-trators. BigQuery has the option of SQL commends in an easy optionof extraction the useful information from Bigdata in realtime.

3. Bigdata

Nowadays, a large volume of data produced from several offline and online sources every second. These data refer to Bigdata. It is difficult to store, process, and analyze Bigdata through traditional database technologies. Bigdata is indistinct and requires substantial processes of classification and conversion knowledge into new insights. Gartner defined Bigdata as high volume, high velocity, and a wide variety of information assets that required new forms of processing to enable enhanced decision making; insight discovery also processes optima-zation" (Harjinder, 2019).

Several researchers define Bigdata from a differing point of view. Bigdata may be distinguished by as 5Vs, i.e., Volume, Variety, Velocity, Veracity, and Value.

1) Volume: It is a sign of the enormous data sets generated at high-frequency rates
2) Variety: This deals with the various categories of data, such as structured, semi-structured, and unstructured.
3) Velocity: This means the processing speed of data at which an application might create data.
4) Veracity: This means the accuracy, truth-fulness in the data, also if its authentication.
5) Value: The remarkable attributes of Bigdata that means the ways of finding hidden values from the datasets instantly (Juneja and Das, 2016).

Some vital statistics of Bigdata:

1) In everyday data are adding in the amount over 1 billion through Google queries and sending 294 billion mails
2) By a minute, 65,972 Instagram photos are posted, 448,800 tweets are composed, and 500 hours' worth of YouTube videos is uploaded.
3) By 2020, the number of smartphone users could reach 6.1 billion. Moreover, considering the Internet of Things into account, there could be 26 billion connected devices by then (THORNTECH, 2020).

3.1 Categories of Bigdata

We can classify Bigdata according to these five aspects (a) data sources, (b) content format, (c) data stores, (d) data staging, and (e) data processing (Hashem *et al*., 2015).
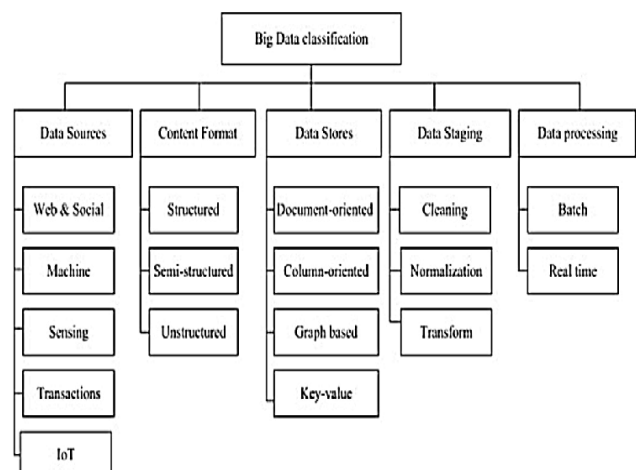

Fig 1: Categories of Bigdata.

3.2 Bigdata Analytics

Analytics of Bigdata is a methodology used to investigate epic data sets containing grouped characteristics of data sorts, such as gigantic data, to uncover every covered model, dark relationship, promote float, customer tendencies, and other steady business information. These definite results could provoke new pay openings, improved customer advantage, upgraded operational expertise, and high grounds over competitor affiliations and distinctive business reimbursement. Analytics may be categorized into the following types:

1. Descriptive analytics: The most direct class of examination, one that grants you to merge big data into humbler, more important pieces of information.
2. Predictive analytics: The accompanying step-up in the decrease of information utilizes a grouping of quantifiable, showing, data mining, and AI procedures to concentrate later and apparent data, allowing specialists to make estimates about what is to come.
3. Prescriptive analytics: Generally, it can be explained as a prescient investigation and need to underwrite a movement so that the business chief can take this information and act (Memon *et al*., 2017).

3.3 Bigdata Applications

Bigdata has many vital applications in the field of technology. Major applications are included below (Memon *et al*., 2017).

a) The Third Eye-Data idea
b) In Banking
c) In Agriculture
d) In Finance
e) In Economy
f) Manufacturing
g) Bioinformatics, etc.

4. Cloud Computing

It is a technology that cans support enormous resources by the requirement of users online in large-scale parameters (Harjinder, 2019). Cloud computing is a paradigm for allowing omnipresent, easy, on-demand network access to a common pool of configurable computing resources (e.g., networks, servers, storage, software, and services) that can be easily provisioned and released with minimal main-tenance effort or interference amongst service providers. Five essential characteristics, service models, and four deployment models compose this cloud model (Pritzker, 2011). Clouds are classified into (Rashid & Chaturvedi, 2019):

1) Private cloud
2) Public cloud
3) Community cloud and
4) Hybrid cloud.

Cloud deployment models are grouped into:

a. Data as a Service (DaaS)
b. Software as a Service (SaaS)
c. Platform as a Service (PaaS) and
d. Infrastructure as a Service (IaaS)

Data as a Service (DaaS) is a paradigm in which data via a cloud-based platform is readily available (Balachandran and Prasad, 2017). Cloud computing has high-level processing units, storages, and applications that do not depend on user devices' performance. Many users can access resources and demanded services remotely from the cloud on a pay-as-use basis. That is why users are not needed to buy and install costly resources locally.

In the digital world, the amount of Bigdata increased very rapidly. It is a significant issue for managing, processing, storing, and analyzing Bigdata by the hardware and applications in local machines. We can resolve the issue by using a cloud computing platform. Some cloud services providers are Google, AWS, IBM, and Microsoft, and they have their Bigdata analyzing grobust systems and products in a cost-efficient manner (Islam and Reza, 2019). However, Google Cloud Computing (GCP) is a com-pletely managed platform that provides excellent services to business users.

4.1 Google Cloud Computing (GCP)

GCP has powerful tools for managing, storing, and efficiently analyzing Bigdata by reducing money and time. BigQuery is a product of Google in Data ware house that used analyzing and representing numerous types of sample data sets in real time for making the right decisions in the case of industries or business purposes (Saif and Wazir, 2018). Big-Query, Cloud Dataflow, Cloud Datalab, Google Cloud Dataproc, Cloud Datalab, Cloud Dataflow, Cloud Pub/Sub, and Google Genomics are the main data analyzing Google services. Among the above

services, BigQuery is as nerveless, user-friendly low-cost Data ware house for analytic (Kumar, 2016).
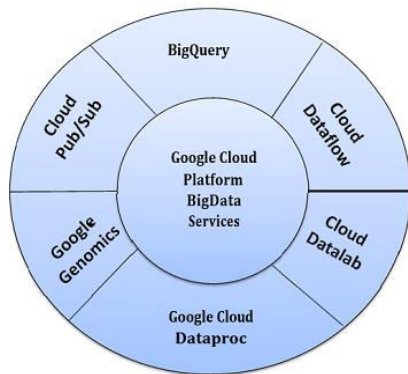

Fig 2: Bigdata Services in Google Cloud Platform.

4.2 BigQuery

BigQuery is the fully controlled data in the cloud platform. The ware house allows carrying out substantial economic queries-data amounts at speeds one would expect from Google. Taking advantage of low pricing and Google's world-class scalability and protection infrastructure provides business insights with strength (Kotecha and Joshiyara, 2018). Big-Query is a petabyte-scale, one of the fastest data warehouse solutions for Bigdata analysis. Without infrastructure and database administrator, one can easily query, represent, and analyze Bigdata as similar SQL commends by BigQery. Hence, most institutions and business organizations are used, from startups to Fortune 500 companies (Kumar, 2016). **Fig 3** shows the data sources that are integrated into BigQuery (Bussiness2Community, 2020).
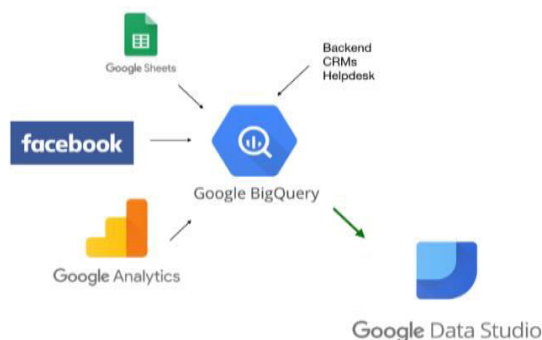

Fig 3: Bigdata from multiple sources into Google BigQuery.

4.3 Bigdata Integration into Cloud

Bigdata and cloud computing and are very closely interrelated. We cannot think about analyzing of Bigdata in our local machine considering processing

time and environmental setup. Besides data in different forms and sources, it is not easy to extract useful information for decision-making. The following **Fig 4** shows the basic architecture of integration of Bigdata into cloud computing from multiple sources (Tomar & Tomar, 2018).
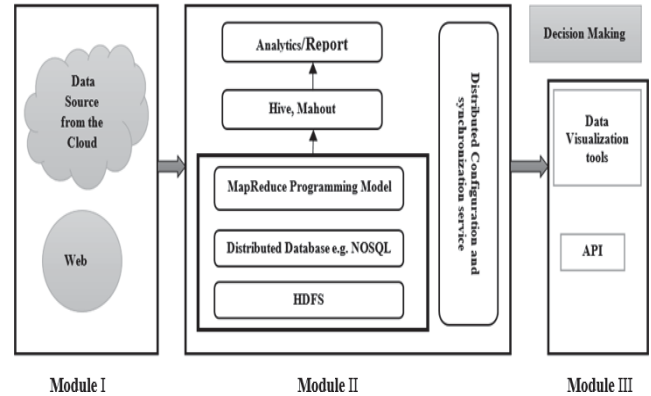

Fig 4: Integration of Bigdata into Cloud.

5. Case Study

We will focus the studies for Bigdata on google cloud. We consider this by a problem statement for the case of dataset 01(ted_main.csv) and dataset 02 (appstore_games.csv). You can load any dataset of the following formats in Google cloud:

- ➢ CSV
- ➢ JSON (newline delimited only)
- ➢ Avro
- ➢ ORC
- ➢ Parquet

Explanation of the steps:

1. Firstly, setting up the environment for BigQuery in Google Cloud -
   - ➢ Login to Console
   - ➢ Login to Big Query
   - ➢ Browsing Publicly available sample tables
2. Real-life case study-Downloading the Publicly available data set
   - ➢ Uploading on Google Big Query
   - ➢ Result set/ Query E
3. Results of the case study
   - ➢ For the case of a public dataset
   - ➢ For the case of real-time dataset

We create a project on Google Cloud Platform and use the service of BigQuery. It is possible to access publicly available datasets and queries through structured query language (SQL) to see various

outputs and data processing speed in BigQuery's data ware house.

1. Accessing publicly available sample data sets in BigQuery Data ware house:
    a) Click on product and services
    b) Click on a product category of Big-Query
    c) Click on bigquery-public-data-sets
2. Browsing publicly available datasets and running some queries with the query editor.
3. After clicking on the tables, for example, Wikipedia and natality, one can see metadata about the table. Metadata represents information about data. In **Fig 5** and **Fig 6**, column details can be seen about a Wikipedia table and the natality table. The tables can be queried by clicking the 'Query Table' button on the top right in the web console.
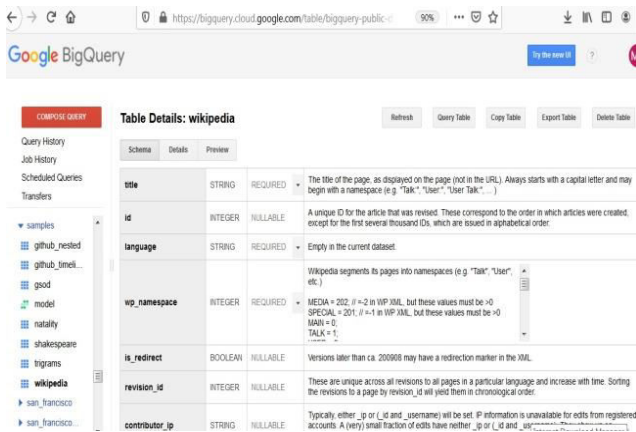
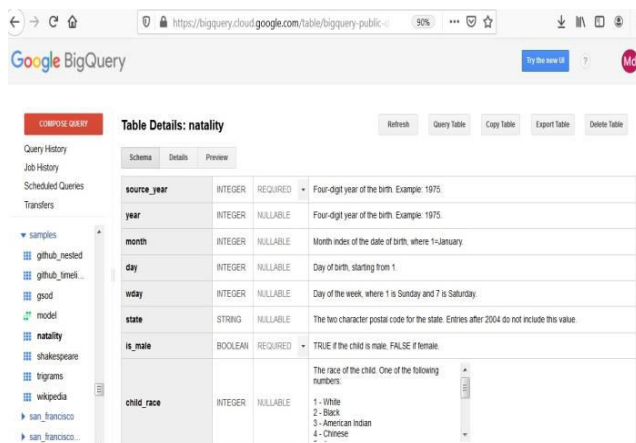**Fig 5:** Sample dataset of Wikipedia on BigQuery.

Fig 6: Sample dataset of natality on BigQuery.

In the following section, data sets of Wikipedia and natality are taken that are publicly available. Data sets are uploaded to BigQuery Data ware house, and then queries are executed to display the above results. This section aims to find a publicly available

dataset, upload it into BigQuery Data ware house and then run a query to find out the result. We take a sample dataset TED talks collected from www.kaggle.com in the format of CSV. These datasets provide metadata on all TED Talk audio-video recordings posted to TED.com's official website until March 2020 (TED, 2020). This dataset downloaded has information about all the recordings which were uploaded on YouTube at different times. TED stands for "Technology, Entertainment, and Design" as a media company that publishes free dissemination talks online under the slogan "ideas worth spreading" (WIKIPEDIA, 2020).

Problem Statement: The main target is to find the top 2000 topics from Ted Talks for dataset 01 (ted_main.csv) at YouTube having maximum views of all time from dataset downloaded and find top 1500 games at App Store counting maximum user rating from for data set 02 (appstore_games.csv). For the solution, the problem statement, we used datasets from external and internal sources.

Dataset 01 termed as ted_main.csv from
www.kaggle.com and Dataset 02 termed as app store_games.csv from Google drive.
URL:https://drive.google.com/file/d/1p2FjJTGMkM mCluhVu0xDchfKUVLIPgnb/view?usp=sharing

Using BigQuery in Data ware house and the following steps were performed to achieve the desired result for the case of data set 01 (ted_main.csv).

1. Finding the Datasets
After some research on google, a website named www.kaggle.com was found to have multiple publicly available datasets. There are two steps needed for dataset download: login to www.kaggle.com by an email id and password. Following the URL, a CSV file with all Ted Main Dataset records was downloaded on the local computer. https://www.kaggle.com/rounakbanik/ted-talks.

2. Need to upload the set of data to BigQuery Data ware house
    a) Logging into BigQuery on the below URL:https://bigquery.cloud.google.com/dataset/bigquery-256708:BigData_on_cloud
    b) Creating new datasets in BigQuery

After logging into BigQuery, clicked on the new project (**Fig 7**). It can be seen that the drop-down menu from 'creative projects' highlights a few options, and the first option is to create a new dataset. The creation of a dataset is a process to upload data on BigQuery Data ware house.
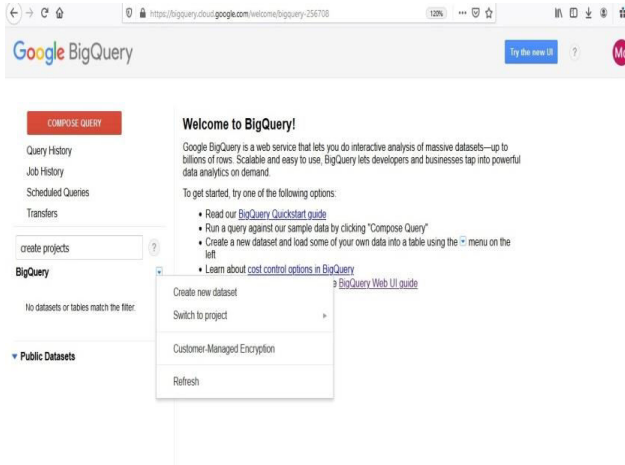

Fig 7: Process to create a new dataset.

For Dataset: 01 ted_main.csv

In **Fig 8**, more details are added for table creation based on available source data, i.e., CSV file termed as ted_main.csv is uploaded from the local computer. In the next row, after setting the table name, click on the table button on the top area of the page is clicked to create the table in BigQuery Data ware house. Following the above steps data set exists on BigQuery. The next step is to upload data sources on BigQuery Data ware house.

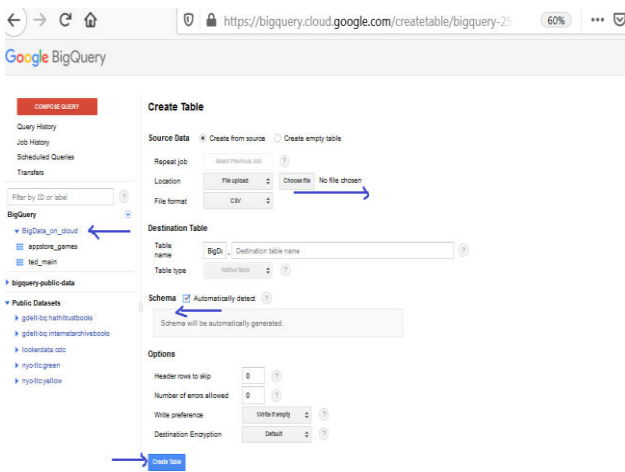In **Fig 9**, the file path is given, which was downloaded from www.kaggle.com.

Fig 8: Uploading file to BigQuery Data ware house.

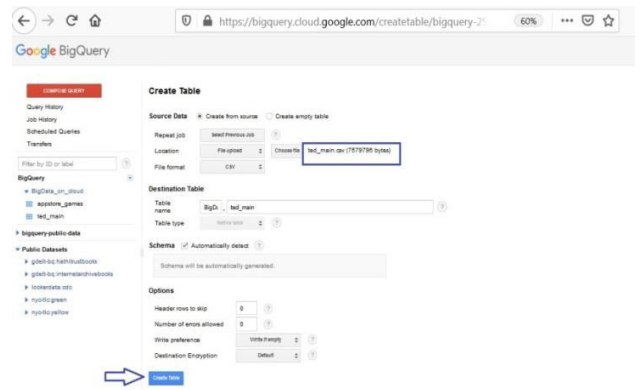In this **Fig 8**, the table name is added, used for querying the data.

Fig 9: Adding the name of the table.

3. Querying table in the editor

The table is ready for query and finding the top 2000 topics viewed by maximum count. This is achieved as per the below query in **Fig 11**. **Fig 10** shows a preview of the taken dataset.
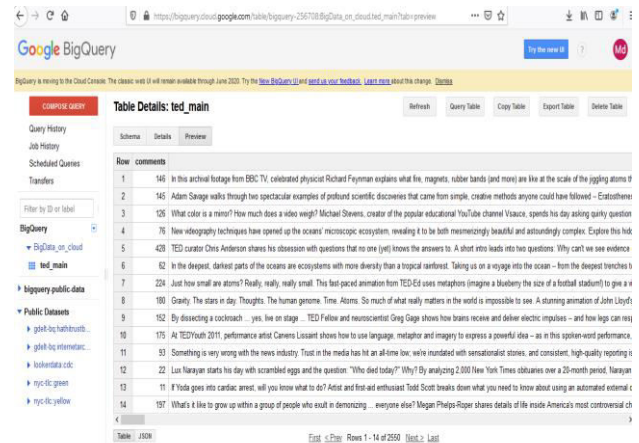

Fig 10: Preview of table details for data set 01.

The final result for Dataset 01: Click query table as per **Fig10**, and writing below SQL query resulted in needed output for finding out the top 2000 topics in the TED event on dataset 01 ted_main.cvs has maximum user rating.
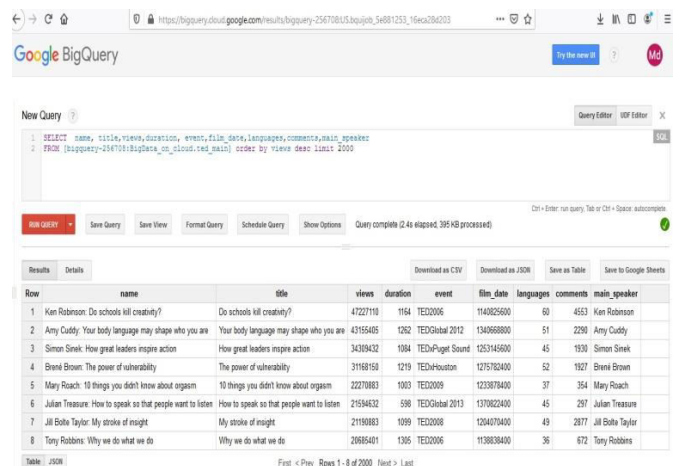

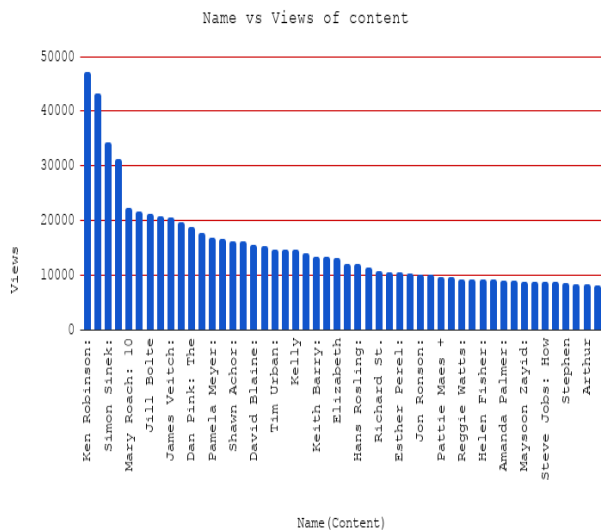**Fig 11:** Displays the results after writing SQL query in query editor on Dataset 01.

Fig 12: Name vs. views of the data on uploaded data set 01.

For Dataset: 02 appstore_games.csv

This dataset termed appstore_games.csvis taken from Google drive by sharing the link because of uploading issues. For the case study, the dataset was taken from Google drive firstly. Linking by the URL to as shown the **Fig 12**, the table was created. Created table details are presented in the **Fig 13** and **Fig 14**. Moreover, finally, the query table is ready as **Fig 14**.
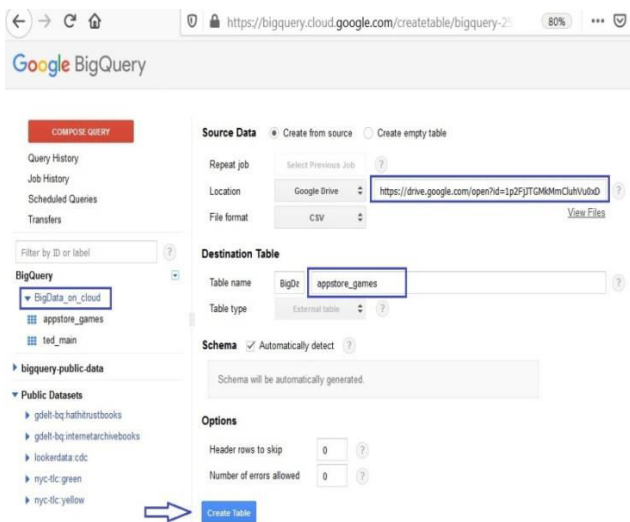

Fig 13: Uploading data set 02 from Google drive.

The final result for Dataset 02: Click query table as per **Fig 14**, and writing below SQL query resulted in needed output for finding out the top 1500 topics in the TED event on data set 02 appstore_games.cvs having maximum user rating.
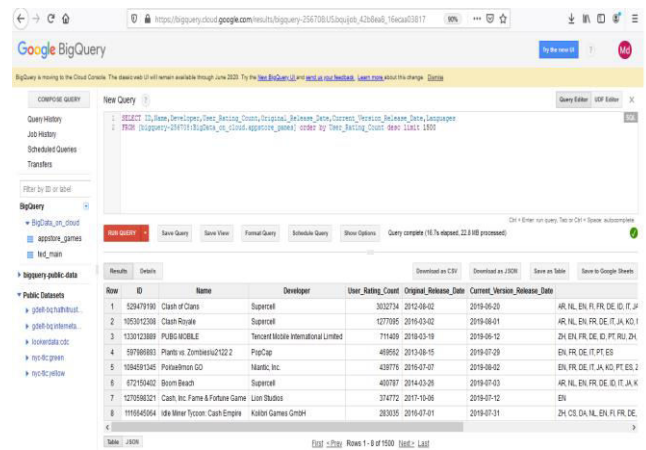

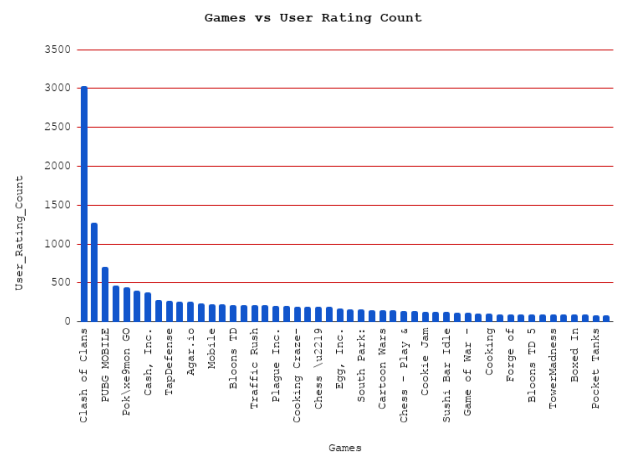**Fig 14:** Displays the results after writing SQL query in Query editor on Dataset 02.

Fig 15: Name vs. Views of the data on Created Data set 02.

6. RESULTS AND DISCUSSION:

This study highlights how easy it is to start the analytics of Bigdata in the field of the cloud. Datasets may be collected from various sources. In our study, data samples are taken from the publicly available datasets on the website www.kaggle.com. The data sets of CSV formed tonate as ted_main was and app store_games. During the upload process of the CSV file, a table is created in BigQuery's Data ware house. Finally, for the dataset ted_main, a SQL query displayed the top 2000 topics' results with maximum view count.

Moreover, for the dataset appstore_games, a SQL query displayed the top 1500 games' results, counting maximum user ratings. In both cases, after querying the data, we saved it to Google sheets for the graphical presentation from where we easily observe the query result within the defined parameters. It is observed from the study on the cloud platform that it is relatively quick and easy. The

main analyzing product of Google BigQuery is used here to have its smooth managing and handling capability. The outcome of this study is analyzing Bigdata (structured, semi-structured, and non-structured) in a real-life scenario. Data from several sources are processed and represented cost-effectively without infrastructure development and database administrators.

7. CONCLUSION AND FUTURE WORK:

Data is an important capital for firms, organizations, and other business areas in technology. The processing of data is so essential to an organization or industry to take the right decision instantly. However, processing and storing a massive amount of different data types (which are in the form of text, audio, and video, etc.) using traditional techniques and methods are so complicated, time-consuming, and costly. Moreover, the traditional servers and database have some limitations in handling these data categories efficiently, which is why evolution to cloud computing began. In our paper, we use the Bigquery service product of google Datawarehouse for solving the issue. Data (structured, semi-structured, and non-structured) from real-time sources is uploaded to the google cloud platform and using BigQuery we can instantly extract necessary information from the data. Moreover, BigQuery is serverless,cost-effective, and easily handled. Without infrastructure development and administrator, we can query, analyze and represent Bigdata within a few seconds.

The paper's main outcome is a representation of big data from a different perspective to instantly take action with the respect of organizations, industry, or any business area. This research will be carried out in the future via a more significant number of experimental datasets. The work will be expanded to BigQueryGIS GIS, rooted in spatial science, incorporates multiple data types. The spatial location is analyzed, and information layers are structured into visualizations using maps and 3D scenes.

8. ACKNOWLEDGEMENT:

9. CONFLICTS OF INTEREST:

The authors declare that they do not have competing interests regarding the publication of the paper.

10. REFERENCES:

1) Balachandran, B., and Prasad, S., (2017). Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Business Intelligence, in International Conference on Knowledge-Based and Intelligent Information and Engineering System, Marseille, France, pp. 1113–1121. https://doi.org/10.1016/j.procs.2017.08.138

2) Bussiness2Community, (2020). Google Big-Query: A Tutorial for Marketer. https://www.business2community.com/marketing/google-bigquery-a-tutorial-for-marketers-02252216

3) Harjinder Kaur, D. M. S. G. (2019). Role of Big Data in Cloud Computing: A Review, *International Journal of Engineering Research & Technology (IJERT)*, **8**(7), pp. 866-869.

4) Harsha, T. (2017). Big Data Analytics in the Cloud Computing Environment. *International Journal of Scientific & Engineering Research*, **8**(8), 393-398. https://cutt.ly/TjX1wv1

5) Hashem, I., Yaqoob, I., Anuar, N., Mokhtar, S., Gani, A., and Ullah Khan, S. (2015). The rise of 'big data' on cloud computing: Review and open research issues, Information Systems, **47**, pp. 98–115. https://doi.org/10.1016/j.is.2014.07.006

6) Islam, M., and Reza, S., (2019). The Rise of Big Data and Cloud Computing, *Internet of Things and Cloud Computing*, **7**(2), 45–53. https://doi.org/10.11648/j.iotcc.20190702.12

7) Juneja, A., and Das, N. (2019). Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application, in 2019 International Conference on Machine Learning, BigData, Cloud and Parallel Computing (Com-IT-Con), India, pp. 559-563. https://doi.org/10.1109/COMITCon.2019.8862267

8) K. Bathla, R., G, S. (2018). Research Analysis of Big Data and Cloud Computing with

Emerging Impact of Testing. *Internat. J. of Engin. & Technol.*, **7**(3.27) 239-243. https://doi.org/10.14419/ijet.v7i3.27.17885

9) Khedekar, V., and Tian, Y. (2020). Multi-Tenant Big Data Analytics on AWS Cloud, in 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, pp. 0647-0653. https://doi.org/10.1109/CCWC47524.2020.9031133

10) Kotecha, B., and Joshiyara, (2018). "Handling Non-Relational Databases on Big Query with Scheduling Approach and Performance Analysis, presented at the 2018 4th International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 118–127. https://doi.org/10.1109/ICCUBEA.2018.8697561

11) Kumar, M. (2016). Google Cloud Platform: A Powerful Big Data Analytics Cloud Platform, *International Journal for Research in Applied Science & Engineering*, **4**(11), 387-392. https://cutt.ly/yjX1kSM

12) Memon, M., Soomro, S., Jumani, A., and Kartio, M. (2017). Big Data Analytics and Its Applications, *Annals of Emerging Technologies in Computing (AETiC)*, **1**(1), pp. 45–54. https://doi.org/10.33166/AETiC.2017.01.006

13) Pritzker Penny, P. D. (2011).NIST Cloud Computing Standards Roadmap, in NIST Special Publication 500-291, Version 2. https://doi.org/10.6028/NIST.SP.500-291r2

14) Rahman, M.M., and Hasibul Hasan, M., (2019). Serverless Architecture for Big Data Analytics, in 2019 Global Conference for Advancement in Technology (GCAT), Bengaluru, India. https://doi.org/10.1109/GCAT47503.2019.8978443

15) Rashid, A., and Chaturvedi, A. (2019). Cloud Computing Characteristics and Services A Brief Review. *International J. of Computer Sciences and Engineering*, **7**(2), 421–426. https://doi.org/10.26438/ijcse/v7i2.421426

16) Riahi, Y., and Riahi, S. (2018). Big Data and Big Data Analytics: concepts, types, and technologies. *International Journal of Research and Engineering*, **5**(9), 524–528. https://doi.org/10.21276/ijre.2018.5.9.5

17) Saif, S., and Wazir, S., (2018). Performance Analysis of Big Data and Cloud Computing Techniques: A Survey, in International Conference on Computational Intelligence and Data Science (ICCIDS 2018). **132**, pp.118-127. https://doi.org/10.1016/j.procs.2018.05.172

18) TED, (2020). TED Recommends. https://www.ted.com

19) THORNTECH, (2020). Big Data in the Cloud. https://www.thorntech.com/2018/09/big-data-in-the-cloud

20) Tomar, D., and Tomar, P. (2018). Integration of Cloud Computing and Big Data Technology for Smart Generation. 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), p.119-124. https://doi.org/10.1109/CONFLUENCE.2018.8443052

21) Trajanov, D., Trajanovska, I., Chitkushev, L. and Vodenska, I. (2016). Using Google Bigquery for Data Analytics in Research and Education. The 12th Annual International Conference on Computer Science and Education in Computer Science, Fulda & Nurnberg, Germany: The Central and Eastern European Online Library, pp. 001-015. https://www.ceeol.com/search/article-detail?id=529712

22) WIKIPEDIA, (2020), TED (conference). https://en.wikipedia.org/wiki/TED_(conference)