International Journal of Leading Research Publication (IJLRP)

E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

Streaming Data Ingestion into Bigquery Using Streamsets

Hareesh Kumar Rapolu

hareeshkumar.rapolu@gmail.com

Abstract

The research paper has shed light on the different ways ingestion of streaming data is carried out in the BigQuery platform by leveraging the power of StreamSets. It can be really important for streamlining all the data silos and getting a holistic view. The study has explored the different ways in which data ingestion is done within the BigQuery platform with the help of StreamSets. Finally, the paper has mentioned a few benefits that a user can achieve through this process.

Keywords: BigQuery, StreamSets, Data ingestion, Google Cloud

I. INTRODUCTION

There is a massive amount of data that is transferred every day. This data needs to be appropriately ingested into platforms like BigQuery through the use of StreamSets. The research paper will critically analyse the different ways in which data ingestion is carried out in BigQuery with the help of StreamSets and its potential benefits.

II. UNDERSTANDING THE CONCEPT OF BIGQUERY

BigQuery is a very important tool that is provided by Google Cloud. This is a fully managed, AI-driven platform that helps an individual or a business to properly manage data. The tool has a number of built-in features like geospatial analysis, machine learning, business intelligence and many others[1]. In addition, the serverless architecture present in BigQuery lets the user incorporate programming languages like Python and SQL. This is extremely important for a business since they are able to operate their data operations and meet their cloud requirements using a zero infrastructure management strategy. It supports uninterrupted data ingestion of streaming data. The entire architecture within the BigQuery primarily consists of two parts. Primarily, there is the storage layer which helps to ingest, store and optimise a volume of data. Subsequently, it has a compute layer that is instrumental in providing analytical insights on the prior operations. The two parts of the BigQuery architecture are independent of one another using Google's extensive petabit-scale network. This is how they are able to communicate with one another whenever necessary. There are different ways in which data is loaded into BigQuery. First, the batch ingestion process incorporates the loading of large volumes of bounded data that do not have to be processed in real-time[2]. Subsequently, the loaded data is ingested at specific frequencies. After that, this data is appropriately queried for creating reports according to the individual or business

requirements. Moreover, the Data Transfer Service (DTS) is a fully managed service that is used to ingest data from Google Saas applications like Google Ads, cloud storage providers like Amazon S3 and shift data from different data warehouses.

Figure 1: Featuring the logo of BigQuery

III. EXPLORING THE STREAMSETS PLATFORM

StreamSets is a widely used data integration platform that is important for helping to build, run and manage numerous data pipelines. The powerful platform is beneficial for accurately overseeing the process of batch and streaming data flows. The StreamSets Data Collector (SDC) is a flexible and easy-to-use data pipeline engine[3]. It allows a business or an individual to simplify the process of data ingestion through a simplified drag-and-drop interface. StreamSets can be considered as an ultimate destination for data ingestion where the data pipeline can be appropriately managed[4]. In this manner, any kind of error can be detected within the pipeline.

IV. HOW STREAMING DATA INGESTION OCCURS IN BIGQUERY USING STREAMSETS

A user can initiate and perform the process of streaming data ingestion in BigQuery with the help of StreamSets. Initially, they need to ensure that the BigQuery platform is properly set up. They also have to ascertain that a particular BigQuery dataset and table is in place where the user wants to ingest streaming data. Subsequently, the BigQuery API is to be enabled from the Google Cloud project. The user needs to further ensure that the different service account credentials are appropriately configured and have the required permissions for successful data ingestion[5]. In the next step, StreamSets Data Collector or StreamSets Cloud can be installed. After that, the Google BigQuery destination is to be changed to StreamSets. After that, the user has to accurately set up the Google Cloud credentials within StreamSets. In the subsequent step, the StreamSets pipeline is to be created by selecting an appropriate real-time data source like Kafka, Google PUB/SUB etc. After this, the accumulated data needs to be cleaned, enriched and transformed with the help of processors like Field Mapper, Expression Evaluator, etc. Furthermore, the Streaming Inserts option needs to be enabled within the BigQuery which is helpful in streaming data ingestion in real-time. In the final step, the user has to utilise StreamSets Control Hub to properly monitor the health of the data pipeline. This is really useful for highlighting any type of error and formulating an appropriate plan for handling them. In this context, it needs to be mentioned that the data ingestion rate within BigQuery is measured by dividing the total amount of data that is ingested by

the total time taken. In addition, data latency is a very important concept within data ingestion that is measured by subtracting the time when the data becomes available in BigQuery from when it was generated.

V. BENEFITS OF STREAMING DATA INGESTION IN BIGQUERY THROUGH STREAMSETS

Low-latency streaming

The StreamSets platform can be directly integrated with the BigQuery API[6]. This is very important for getting data that are available for real-time analysis. Therefore, the lack of latency helps the users to carry out their operations within the stipulated time period.

Simplified schema and data handling

If there are any kind of changes within the BigQuery schemas, StreamSets is able to detect them and update them automatically[7]. The entire operation is carried out in such a way that the data pipeline is not disrupted.

Data quality control

StreamSets plays a big role in eliminating unnecessary data and refining them properly before they are loaded into BigQuery. This improves the quality of the overall data ingestion process and the scope of any kind of error is massively minimised.

Centralised monitoring

The StreamSets Control Hub can help to visually represent the data flow that is happening within the data pipeline. Hence, this is really instrumental for monitoring the overall performance. Moreover, it can also detect latency spikes in real-time.

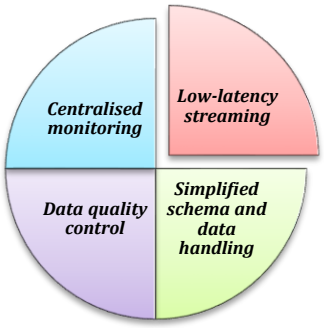

Figure 2: Benefits of streaming data ingestion in BigQuery through StreamSets

VI. CONCLUSION

In conclusion, it can be mentioned that a user needs to properly monitor the process of the ingestion of streaming data into the BigQueryPlatform. StreamSets is a reliable tool that can help to optimise the process and safeguard data security and validity. The user can gain significant insights about their data pipeline which can help them to achieve their desired objectives.

Abbreviations and acronyms

- SQL - Structured Query Language
- DTS - Data Transfer Service
- SDC - StreamSets Data Collector
- API - Application Programming Interface

Units

- Total number of records that are ingested within a second - Rows/sec
- End-to-end latency - Ms or sec
- CPU utilisation - Percentage (%)

Equations

- R=D/T
- L=Tarrival−Tevent

REFERENCES

[1] A. Fard, A. N. Le, G. Larionov, W. Dhillon, and C. Bear, "Vertica-ML: Distributed Machine Learning in Vertica Database," *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, May 2020, doi: https://doi.org/10.1145/3318464.3386137

[2] A. Singhal, T. Winograd, and K. Scarfone, "Guide to Secure Web Services Recommendations of the National Institute of Standards and Technology," NIST Special Publication, Aug. 2007. Available: https://profsite.um.ac.ir/kashmiri/nist/SP800-95.pdf

[3] K. NAGORNY, S. SCHOLZE, A. W. COLOMBO, and J. B. OLIVEIRA, "A DIN Spec 91345 RAMI 4.0 Compliant Data Pipelining Model: An Approach to Support Data Understanding and Data Acquisition in Smart Manufacturing Environments," *Ieee.org*, vol. 8, Dec. 2020, doi: https://doi.org/10.1109/ACCESS.2020.3045111. Available: https://ieeexplore.ieee.org/iel7/6287639/6514899/09296293.pdf

[4] M. Srivastava, C. Sabarinathan, R. Sankineni, and M. TM, "Mining Of Big Data Using Map-Reduce Theorem," *IOSR Journal of Computer Engineering*, vol. 1, no. 17, pp. 49–55, Feb. 2015, doi: https://doi.org/10.9790/0661-17164955

[5] N. Singh, D. P. Singh, and B. Pant, "Big Data Knowledge Discovery Platforms: A 360 Degree Perspective," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 9, no. 2, Dec. 2019, doi: https://doi.org/10.35940/ijeat.B3901.129219

[6] P. Atri, "Enhancing Big Data Interoperability: Automating Schema Expansion from Parquet to BigQuery," *International Journal of Science and Research (IJSR)*, vol. 8, no. 4, pp. 2000–2002, Apr. 2019, doi: https://doi.org/10.21275/SR24522144712%202

[7] T. Mahapatra, "Composing high-level stream processing pipelines," *Journal of Big Data*, vol. 7, Sep. 2020, doi: https://doi.org/10.1186/s40537-020-00353-2