WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Enterprise Data Governance

Tanay Patil[1], Manthan Sarfare[2], Omkar Sodaye[3]

[1]Information Technology, Vidyavardhini's College of Engineering and Technology, Mumbai, India
[1]tanay.192164101@vcet.edu.in,
[2]Information Technology, Vidyavardhini's College of Engineering and Technology, Mumbai, India
[2]manthan.192244101@vcet.edu.in
[3]Information Technology, Vidyavardhini's College of Engineering and Technology, Mumbai, India
[3]omkar.192294101@vcet.edu.in

ABSTRACT—

Data's importance has grown dramatically over the last ten years, going from being a queryable or reportable resource to a genuine company asset. The complexity of the corporate structural framework has also increased significantly at the same time. Due to these factors, a company now needs to analyze and manage its data. Our data governance project visualizes and provides insights from a business' data by streaming the data from an enterprise through an user or employee by Kafka streaming to our Azure database and carrying out the visualization process with PowerBI. Based on the type of data, we made relevant charts and diagrams on PowerBI.

Keywords—analyzing; visualizes; streaming; data; PowerBI

I. Introduction

Enterprise Data is analyzed and insights are drawn in order to make valuable use of that data and improve a business's profits. These insights could simplify crucial decisions regarding the workings of a business. Data is ingestion and streaming is managed by implementing Apache Kafka, which sends the data to our Azure Data Factory Blob storage by implementing a pipeline. Here, the relevant data cleaning is performed and then the data is transferred to PowerBI for visualization.

II. Problem Definition

Data has become an important component in today's world, where a business can thrive or go bankrupt depending upon the way it utilizes its data. Every business needs to gain insights from their data to make important business decisions for maximizing profits. However, the exponential rise in the demand for business insights resulting from advancement in the IT field has not been sufficiently met by the resources available currently. Existing companies offering business insights also charge exorbitant charges for data visualization. Our project aims to bridge the gap between the supply and demand and offer a cheaper alternative to the current competitors in the market.

III. Proposed Approach

A robust and scalable framework can be designed with the help of open-source NiFi-Kafka for continuous data streaming. Data ingestion is an essential part of companies and organizations that collect and analyze large volumes of data. Continuous data streams usually arrive in big data processing and management systems from external sources and are either incrementally processed or used to populate a persisted dataset and associated indexes. To keep pace with massive and fast-moving data, stream processing systems must be able to ingest, process, and persist data on a continuous basis [1]. The architecture of the project is based on this approach.

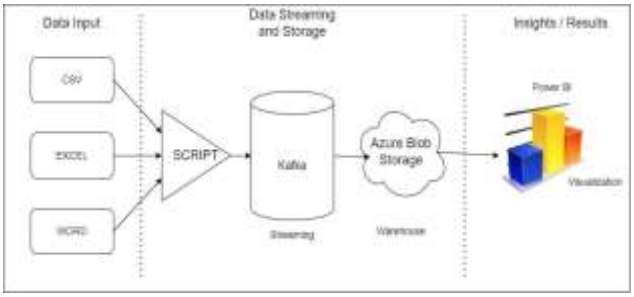

Fig. 1. Architecture

The concept of executing a data warehouse in the cloud is gaining momentum. As cloud storage is economical and scalable, fitting data warehouses in cloud architecture is of tremendous business need. If integrated with cloud computing, data warehouses can be offered as a service. In November 2012, AWS (Amazon Web Service) released Redshift which is a cloud data warehouse. Microsoft also developed Azure SQL DW service. Azure DW can hold a petabyte range of data. It also leverages TSQL to handle relational and non relational data. IBM's dashDB is also another cloud DW service [2]. This research aided in the decision to use cloud in the project.

There are many tools such as Tableau, Plotly, Gephi, Excel 2016, and PowerBI. Microsoft Power BI is a powerful cloud-based business analytics service. Visualizations are interactive and rich. Power BI consists of 3 elements, Power BI Desktop, Service(SaaS), and Apps. Every service is available to us which is why it makes Power BI flexible and persuasive. Amongst the listed tools, we selected Power BI keeping in mind the various features and benefits mentioned above [3].

Comprehensive analysis of various ETL techniques and their relevance in processing dynamic data streams was researched[4]. Key issues associated with data streams, such as high volume, velocity, and varied formats were found and the evaluation of how well the existing ETL techniques can handle these challenges were found[5].

IV. Implementation

The first part of the implementation of this project includes setting up a storage account, which would be used to store data ingested into the database. The Azure Data Lake Storage Gen2 is preferred to create the storage account over its competitors after analyzing the costs associated with its implementation.

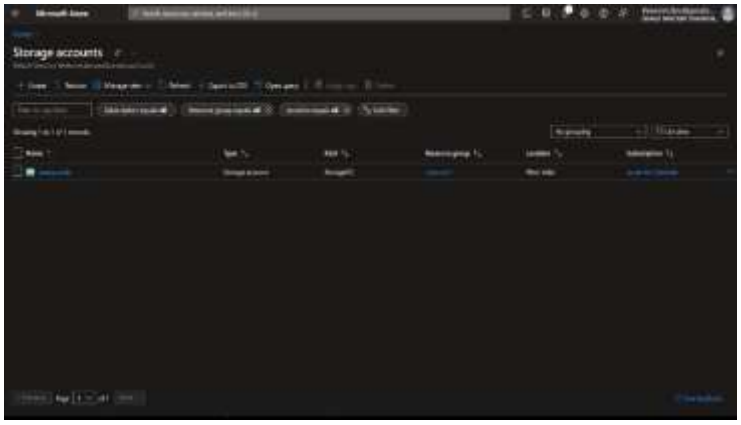

Fig. 2. Setting up the Storage Account

The next part was to create a container inside the Storage Account. A container is just like a folder, which is used to differentiate and classify data in the Storage Account. Basically, it can determine the location where you want to write the data. The Azure Blob Storage is used for storing the data as it can store massive amounts of unstructured data, like text or binary data. We have named the container as e-commerce data, as it contains sample data related to e-commerce sales.

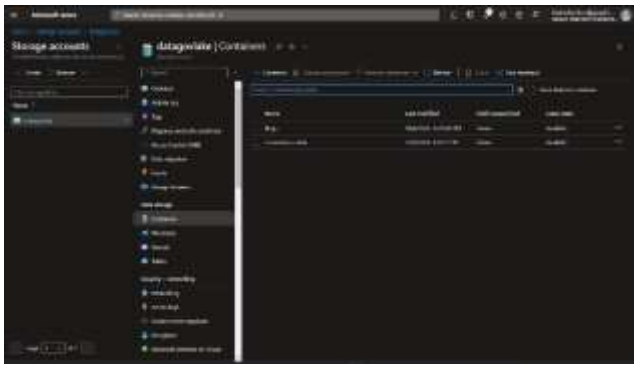

Fig. 3. Creating the container

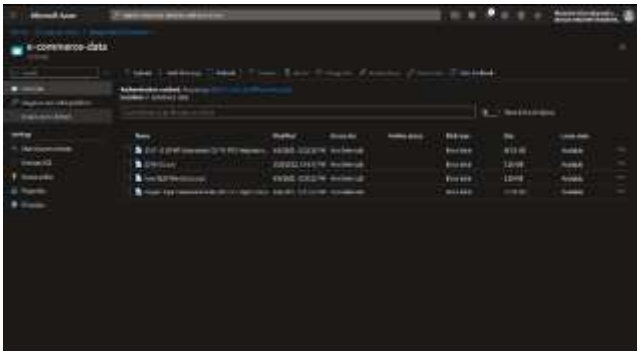

Fig. 4. Data stored in the container

The next stage of the project involved connecting our database to our website, which would enable the user to upload and store the data that they want to get the insights from. Data preprocessing and transforming the data is done by removing duplicate values from the ingested data, filling the null values and missing data with relevant data for overall cleaning of the data. Once the data was cleaned, we moved on to the next stage, which was to visualize the data and gain insights from it.

For visualization, Microsoft PowerBI was preferred, as it allows easier integration with Microsoft Azure. Here, we first used the 'Get Data' function to import the data. Then, we selected Azure Blob Storage from the list of options to select the data from.

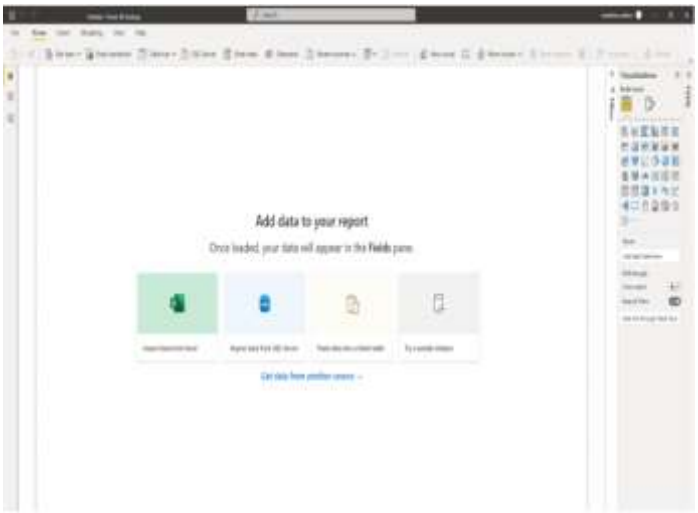

Fig. 5. Getting The Data

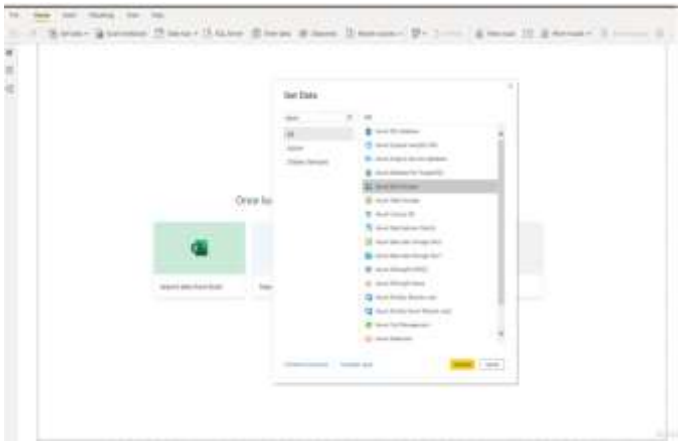

Fig. 6. Selecting Blob Storage

After selecting the storage, we have to select the container from which we are importing the data, which is e-commerce in this case.

Fig. 7. Selecting the container

V. Results and Discussion

For this project, we are using the Azure Data factory with Apache Kafka for creating a pipeline to get our data from any source and store it in Azure Blob Storage in JSON format. The stored data is further Loaded into the Power-BI tool. The loaded data is transformed to match our requirements. The transformed data is further visualized to draw insights. The data results of each stage are divided and given below.

Fig. 8 Loading the data in fields

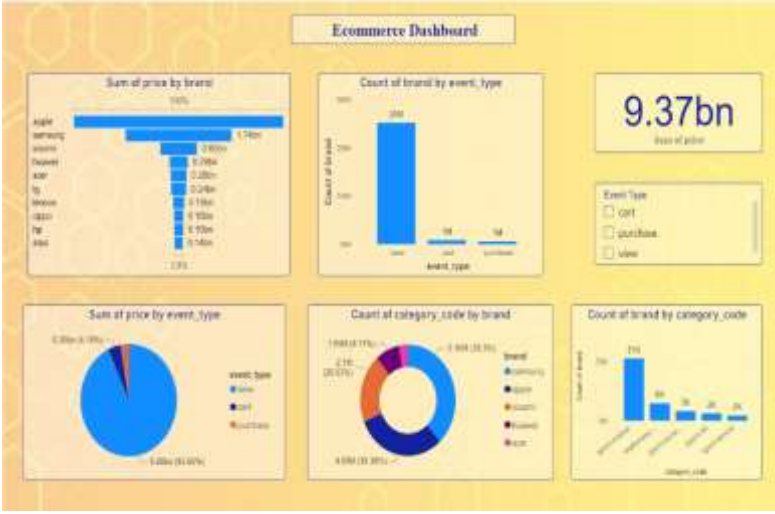

Fig. 9 Visualized Data in PowerBI

VI. Conclusion

While implementing this project, we successfully created a data factory in Azure to store data collected from various sources through Apache Kafka for performing data visualization and getting different insights from that data. Power-BI desktop and Power-BI service is used for visualization and report analysis.

VII. Further Work

In this project, we have used only one data source as an example for implementing the data ingestion, cleaning, and visualization process. In the future, we can cater to the demands of our clients by tweaking our methodology and working process by collecting different types of data from various sources. We can personalize this experience and help a variety of businesses to perform visual analyses of their data. This report is further published to a webpage by embedding the Power-BI dashboard in our react app to enhance UI, which can be viewed by the user.

References

[1] A Scalable and Robust Framework for Data Stream Ingestion by - Haruna Isah, Farhana Zulkernine - 2018

[2] A Comparative Review Of Data Warehousing ETL Tools With New Trends And Industry In- sight by - Rajendrani Mukherjee, Pragma Kar - 2017

[3] Big Data Visualization: Tools and Challenges by - Syed Mohd Ali, Noopur Gupta, Gopal Kr- ishna Nayak, Rakesh Kumar Lenka - 2016

[4] A Big Data Perspective of Current ETL Techniques by K V Phanikanth, Sithu D. Sudarsan - 2016

[5] Variety of data in the ETL processes in the cloud: state of the art by Papa Senghane Diouf, Aliou Boly, Samba Ndiaye - 2018