Product Sales Analysis: Analysis Powered by Databricks, Spark and Spark SQL

Sarim[1], Salim Ali[2], Deepak kumar[3], Prof. Ashish Chauhan[4]
[1, 2, 3] *B. Tech Student, Dept. of Computer Engineering, SRGC College, Muzaffarnagar, Uttar Pradesh, India*
[4] *Associate Professor, Dept. of Computer Engineering, SRGC College, Muzaffarnagar, Uttar Pradesh, India*

Abstract— In this 21st century the data is reaching to the skies with fast-growing advancements in E-commerce and product-centric organizations so the data science and the data analytics are the backbones of company. we improve the sales of product after Predicting and analyzing the product for more profit. Big Data application enables these organizations to use prior year's data to better forecast and predict the coming year's sales. It also enables retailers with valuable and analytical insights, especially determining customers with desired products at desired time in a particular store at different geographical location.

Index Terms- Big Data, Data Analytics, Databricks, Spark SQL, Business Intelligence

I. INTRODUCTION

There has been a sudden increase in commerce, and e-service innovations due to technological advancements and greater customer demand. Organization can apply new information technologies using both quantitative and qualitative approaches to develop an understanding the customer demands, predict market patterns and drive revenue generation. Specific ways catering to different organization types – whether they are small businesses, large enterprises or specialized agencies, can increase profitability and drive market success. Toadapt to constantly changing markets, companies are forced to adapt quickly to new and unknown situations. Companies are now increasingly relying on data- driven technologies such as Business Intelligence (BI) to accommodate fast-paced decision-making processes. Business Intelligence can be captured as the process of accumulating, organizing, analyzing, presenting and monitoring information supporting management decisions. The collection of unbiased data present in the market, performing analysis and presenting strategies to increase business efficiency is the primary objective of data analysis. It allows us to gain a significant competitive analysis facilitating future decisions

II. LITERATURE REVIEW

Apache Spark is an open source (free Source) unified analytics engine for huge scale data process has transformative technology in the big data ecosystem. Databricks, founded bythe creators of Spark, extends Spark's capable through its cloud-based platform, integrate with Databricks SQL to offer powerful, scalable data analytics. This evolution, capabilities, and impact of Databricks SQL and Spark in modern data analytics. Apache Spark was introduced in 2009 to address limits in the Apache Hadoop, specifically in terms of speed and ease of use. Spark in memory process capable, general compute model, and helpful for advance analytics, such as machine learning and graphprocessing, have it as a prefer choice for big data processing. Databricks was developed to simplify the deployment and management of Spark clusters. It provides an integrated environment that supports the entire data lifecycle, from data set and storage to real-time analytics. The platform seamless with various cloud services enhances its flexibility and scalability. Databricks SQL, previously known as SQL Analytics, It combines the scalable of Spark with the familiar and easy of SQL.

III. PROBLEM STATEMENT

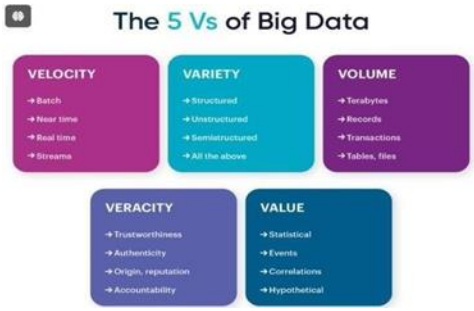
Organization first priority to understand their customers demand to satisfy their needs so that these

customers will return to the store for future needs, thus increasing the product demands for enhance business value These businesses want this information to plan where and when to invest profitably

IV. PROPOSED SYSTEM

1. Big data

The data cannot be handled by traditional tools are called bigdata. The datasets become larger and more difficult to manage. big data also refers to extremely large and diverse collections of structured, unstructured, and semi-structured data that continues to grow exponentially over time. These datasets are so huge and complex in volume, velocity, and variety, that traditional data management systems cannot store, process, and analyze them big data analytics describes the process of uncovering the trends and patterns in large amounts of raw data to help make data-informed decisions.



Volume: Volume refers to the amount of data that gets processed. It speaks about the scale and size of the data that gets processed. Currently data between Exabyte (EB) and Zettabyte(ZB) are considered as big data.

Veracity: Veracity represents the quality of data. Data can be incomplete, inconsistent, uncertain and these data sets also categorized as good, bad and wanted, unwanted, defined, or undefined. Accuracy and trust play a huge factor in determining the usability of data. Healthy data sets are collected from trustworthy sources.

Velocity: Data velocity is defined as the speed at which data is being generated or processed.it is also represented in terms of real time, streaming, near real-time and batch.

Variety: Veracity determines the type of data which can be categorized as multimedia, structured, semi-structured and unstructured.

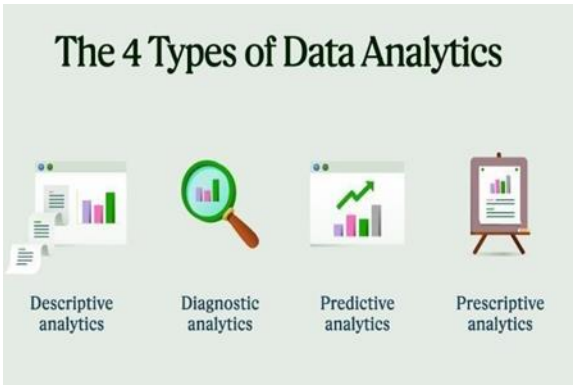
Value: Value represents the usefulness and context of the data for decision-making.

2. DATA ANALYTICS

the primary goal of data analytics is to uncover valuable information from raw data helping organization and individual make informed decisions, solve problem and improve overall performance.

There are four key of data analysis that can help an organization make data driven decisions.

Descriptive analytics: Tell us what happened Diagontic analytics: Tell us why something happend. predictive analytics: Tell us what will likely happen the future prescriptive analytic: Tell us how to act.



3. Databricks is a unified and open analytics platform for Building, deploying, sharing, and maintaining large data. The Databricks Data Intelligence Platform integrates with cloud storage and security in your cloud account, and manages and deploys cloud infrastructure on your behalf. In simple words we can say that databricks is an industry-leading, cloud-based data engineering tool used for processing and transforming massive quantities of data and exploring the data through

Machine learning models Databricks used for

i. Data processing scheduling and management, in particular ETL
ii. Generating dashboards and visualizations
iii. Managing security, high availability, and disaster recovery

iv. Machine learning modeling
v. Generative AI solutions

1. SPARK SQL is a new module in Apache Spark that integrates relational processing with Spark's functional programming API.

We see Spark SQL as an evolution of both SQL-on-Spark and of Spark itself, offering richer APIs and optimizations whilekeeping the benefits of the Spark programming model. Withthe experience from Spark, we wanted to extend relational processing to cover native RDDs in Spark and a much widerrange of data sources.

Goals for Spark SQL

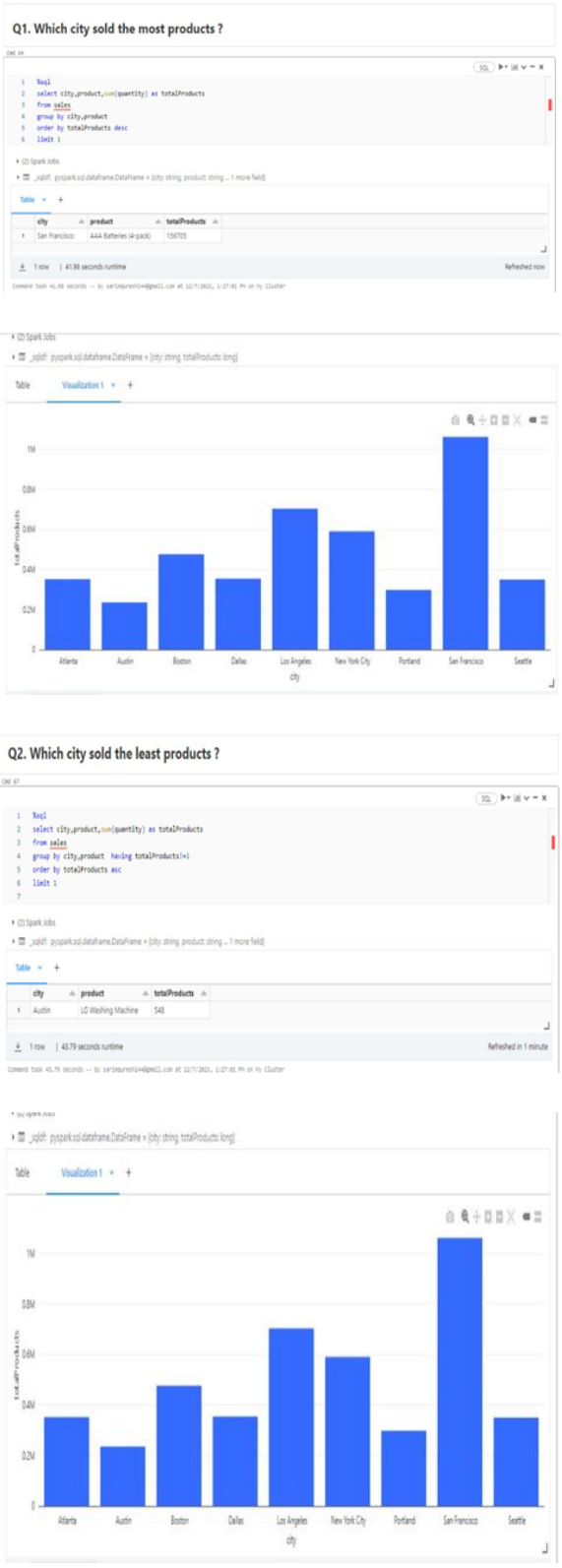
1. Support relational processing both within Spark programs and on external data sources using a programmer friendly API.
2. Provide high performance using DBMS techniques.
3. Easily support new data sources, including semi-structured data and external databases for query
4. Enable extension with advanced analytics algorithms suchas graph processing and machine learning.

V. BUSINESS INTELLIGENCE

Business intelligence analysts transform raw data into meaningful insights that drive strategic decision-making within an organization. BI tools enable business users to access different types of data—historical and current, third- party and in-house, as well as semi structured data and unstructured data such as social media. Users can analyze thisinformation to gain insights into how the business is performing and what it should do next.

CONCLUSION

In this paper, we discussed the sales and marketing integration interface and the impact of big data analytics. We saw the various characteristics of big data and their correlation to different sales data accumulated by organizations. SQL and spark SQL are the revolution in the field of data analysis. Their integration has Access to derive the insights from vast data sets. Continue in these technologies are expected to further drive innovation and efficiency in data-driven decision-making across various industries.



REFERENCES

[1] Apache Avro project.

[2] J. Cohen, B. Dolan, M. Dunlap, J. Hellerstein, and C. Welton. MAD skills: new analysis practices for big data. VLDB, 2009

[3] Ghodsi, A., Zaharia, M., Xin, R. S., Sievert, O., Wendell, P., Das, T., ... & Stoica, I. (2016). Databricks: Unifying Data Analytics. In Proceedings of the 2016 Conference on Innovative Data Systems Research (CIDR '16).

[4] Apache Parquet project. http://parquet.incubator.apache.org.

[5] Apache Spark project. http://spark.apache.org.

[6] G. Graefe. The Cascades framework for query optimization. IEEE Data Engineering Bulletin, 18(3), 1995.

[7] M. Isard and Y. Yu. Distributed data-parallel computing using a high-level programming language. In SIGMOD, 2009

[8] A. Pavlo et al. A comparison of approaches to large-scale data analysis. In SIGMOD, 2009.

[9] M. Zaharia et al. Resilient distributed datasets: a fault-toleran abstraction for in-memory cluster computing. In NSDI, 2012

[10] J. Hegewald, F. Naumann, and M. Weis. XStruct: efficient schema extraction from multiple and large XML documents. in ICDE Workshops, 2006.

[11] A. Spark, "Spark sql and dataframe guide," [Online]. Available: [Accessed Sept 2017. [Online]. Available: www.{Spark}.apache.org/docs/1.5.2/sql-programming-guide.html

[12] R. Xin, M. Armbrust, and D. Liu, "Introducing dataframes in apache Spark for large scale data science," [Online]. Available: [Accessed August 2017], February

[13] https://databricks.com/blog/2015/02/17/introducing-dataframes-in-{Spark}-for-large-scale-data-science.html