AD HOC Data Analytics Platform

[1]Tejas Nakka, [2]Shubham Shikari, [3]Akshaykumar Deekonda, [4]Harshali Rambade,
[1,2,3]Student, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, India
[4]Professor, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, India

Abstract: Currently the data is being generated at a very rapid pace. And analysis of Data is becoming more difficult with increasing volume and it is also very tough to generate real-time data analytics which is currently in demand. Legacy Business Intelligence lacks a very basic need for a tool that is it should be simple to operate by End User. But According to the user of the tool, it requires various certifications needed to be done before we can use the tool itself. Due to that reason, there arises a need for a person who is trained in that tool thus giving rise to another layer of processing through which data needs to be passed before it reaches the end-user and adding more delay. Legacy BI Tools also lack the human perspective of Natural Language Search. When an end-user generates a request for a particular question it takes an awful lot of time to respond with an answer. It is also important to note that the Legacy BI does not have a single source of truth and therefore it brings the ambiguity factor into the picture. Therefore it is high time to shift from Legacy BI to a tool that is capable of solving the problems by providing control to the end-user and has a single source of truth.

Keywords: *Business Intelligence, Legacy BI.*

I. INTRODUCTION

Similar to a very broad concept of management information systems (MIS), business intelligence (BI) is a helpful and convenient software technology for organizations to support the management decision-making process. Whereas management information systems refer to a set of different automotive systems in different information management methods such as decision support systems (DSS) and executive information systems (EIS), business intelligence bears on a category of applications and technologies for data gathering, storing, analyzing, and accessing as a whole.

In recent years, means of management information systems, which collects, transforms, and presents operational data daily into information for effective decision making, has been moved to the term Business Intelligence . This is because many businesses tend to collect an expanded amount of data during their daily operations. Larger quantities of data can be processed and transferred into valuable information that is useful for better business decisions. For example, with a collection of sales, stores, types of products, and types of customers, sales managers will be able to determine commercial strategies to increase the company's sales volume.

Business intelligence is beneficial for organizations in a way of its ability to deal with massive volumes of data from various sources. An essence of data management in business intelligence is to analyze historical data of the business which can be stored in a data warehouse or OLAP. By obtaining knowledge and useful information for future prediction, it provides solutions for transforming data into information along the decision-making support process through computer-based analytical tools to the users.

In AD HOC Data Analytics Platform makes it easy to add or update or connect datasets or databases easily, configure, visualize, trigger, and much more. Any end-user can easily perform analytics no training is necessary to control like BI.

II. LITERATURE SURVEY

A. Comparison of Database Engines for Efficient Reading of Large Datasets

The Website shows a comparison between different Database Engines like InnoDB, MariaDB Columnstore. There was four test case scenario in each test case; no of rows were loaded from the requests obtained to the Wikipedia website. In each test case, a select query was run and found that in each case the column store database engine was 80% more efficient than the conventional InnoDB Engine. It was also found that insert query was not as efficient as InnoDB but bulk Insertion was as same as InnoDB.[1]

B. Comparison of Database Engine with Elastic Search for loading Search Values

The Website displays a comprehensive comparison of performance and technical parameters of elastic search and column store engine. And after the review of the website, it can be concluded that it would be very efficient to use elastic search for search values instead of column store. As the time to connect and time to query are better in elastic search.

C. Selection of Right Chart Type

As the Metrics (Number Columns) and Dimensions (String columns) are selected by the user which chart should be selected so the visualization would be better in terms of information presentation and human understandability. Out of various charts like line, pie, bar, column, scatter, bubble, etc. Few Rules are pre-determined by the system on chart selection.[3]

D. Natural Language Processing (NLP)

NLP is to be used for natural language search queries in which the user will be able to interact with the system in the very natural human way i.e speaking. Basic concepts should be very familiar like a bag of words, tokenization, lemmatizing, etc.[4] But building an NLP model from scratch may be time costly and not scalable. Therefore it is recommended to use some framework. [5]

III. OBJECTIVES

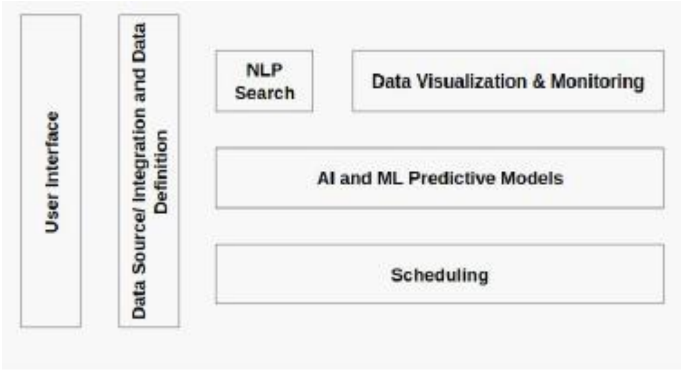
From the view of business users to experienced data analysts, everyone benefits when data and Insights are available by simply asking a question. The goal of better, more accessible data use is more informed decision making for agency leaders and better public access to data for private industry and citizens to innovate solutions for challenges facing our government and communities. Learn how to search and AI-driven analytics is bringing a more democratized approach to data access and management to more effectively and efficiently meet mission goals.

The expected result from this system is a comparison between conventional and in-memory business intelligence features that cause and overcome the difficulty in using business intelligence tools and below are some points.

1. Simple and Customizable User Interface.
2. Any organization can use it easily and in their environment.
3. Users should be able to deploy datasets from any source like Data warehouse, CSV, Database, text files, etc.
4. Organizations can easily create sub-users to give access to the organization.
5. Provide a search feature so that users can easily search on their dataset using NLP.
6. Automatic Prediction of proper visualization based on the user search result.

IV. PROPOSED SYSTEM

Our System is a Progressive Web App based Application which means on a single code base we can generate the applications for Desktop, Mobile, and Web. And System will remain the same but the access type media will change. On Login or Register, the user will be able to select the workspace from the list of workspaces available. After selecting the workspace user will be able to access the features based on his workspace and dataset role. He can create a dataset by uploading a file either of CSV, text, JSON, excel or can pass URL, username, password of the already existing database. If the Existing Database is selected then he can either run select queries on his database server or can schedule pull of data from their database to our system and then we can run queries on our system.


User Interface
Data Source/ Integration and Data Definition
NLP Search
Data Visualization & Monitoring
AI and ML Predictive Models
Scheduling

While the uploading is done it is first loaded into a table with all columns as strings and then the process of optimization is taken place and every column's datatype is determined. Users can also determine which columns are metrics and which are dimensions and configure their formatters, etc. Then we can perform a variety of functions on the uploaded dataset. And through natural search query users can ask direct questions on the dataset and the output would be chart visualization with suitable chart output selected by the chart selecting algorithm. Then users can either download the chart in png, jpg, SVG, and pdf or pin it to the dashboard for future reference. Also can set daily, weekly or monthly notification for visualization.

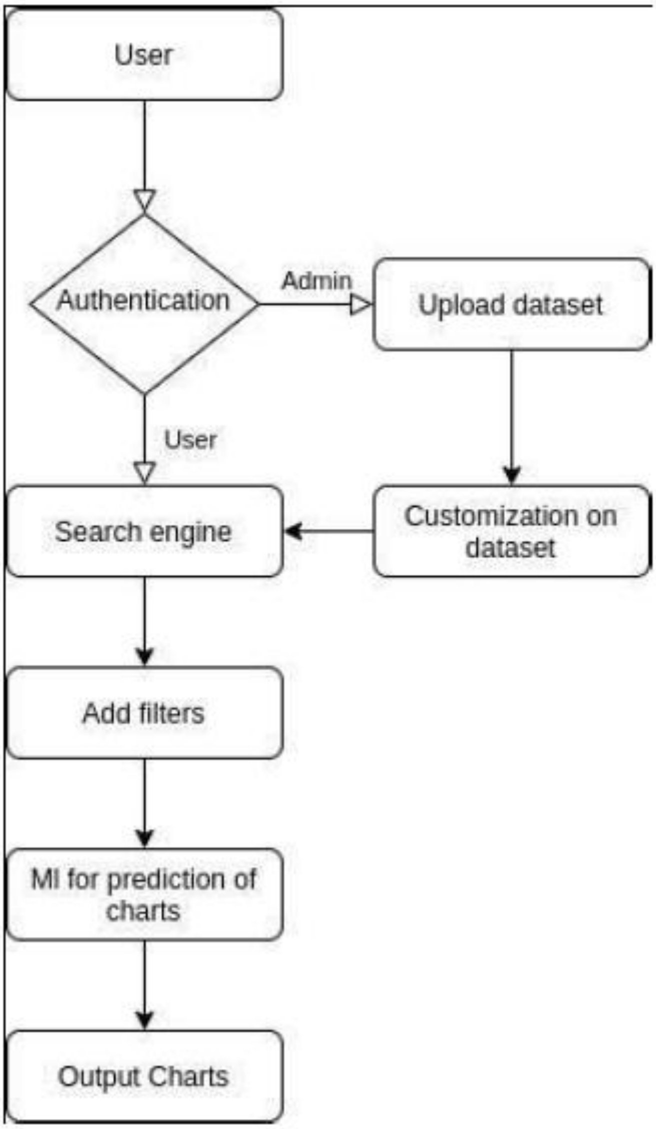


Fig. 2. Flowchart

IV. METHODOLOGY

From all of our findings, we conclude that given the following tasks are need to be done to ensure the quality level of the product is well maintained.

- Data Source and Integration: Uploading data is a major problem. Hence, Our data source integration module provides you the different types of uploading your data like MySQL connection, PostgreSQL SQL connection, files, data warehouse connection, etc.
- Data Definitions: After uploading data the next step is data definitions. This module detects the metric. dimension and date from the uploaded datasets.
- Natural Language Search: Completion of Data source and Integration and Data Definition Now users can perform an analytic by using a natural language search.
- Data Visualization and Monitoring: According to the user asked to query the result of the response will be processed by this module in the form of graphs and charts.
- AI and ML Predictive Module: From the Machine Learning concept this module will be able to automatically detect the best visualizing experience for the user. According to the user data, it can also predict the flow of the data.

- Schedule Reports: Users can set the notification and schedule the reports time to keep tracing on their data.

CONCLUSION

Our proposed system enables end-users to upload or connect their data from various file formats and database types. A user interface that is loaded with the feature built on open-source technology and search enables the end-user to query their dataset on an ADHOC basis. Thus removing the requirement of data analysts and saving plenty of time. Our system can be used on a general-purpose dataset leading to the removal of hardcoded rigidness.

References

[1] Db Performance Comparison: https://www.percona.com/blog/2017/03/17/column-storedatabase-benchmarks-mariadb-columnstore-vs-clickhouse-vs-apache-spark/.

[2] Elasticsearch V/S Columnstore: https://reportserver.net/blog/2016/06/20/mariadb-columnstore-vs-innodb-vs-monetdb/.

[3] Chart Selection: https://infogram.com/page/choose-the-right-chart-data-visualization

[4] NLP Basics: Shttps://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1

[5] Dialog Flow: https://cloud.google.com/dialogflow/es/docs