GENERIC USER EVENT ANALYSIS AND PREDICTION

Shreyas Kulkarni[1], Kiran Mokashi[2], Shivam Bawane[3], Shailesh Bagade[4]

1 Computer Engineering (B.E.), P.I.C.T, Pune, Maharashtra, India
2 Computer Engineering (B.E.), P.I.C.T, Pune, Maharashtra, India
3 Computer Engineering (B.E.), P.I.C.T, Pune, Maharashtra, India
4 Computer Engineering (B.E.), P.I.C.T, Pune, Maharashtra, India

---***---

Abstract - *This paper presents a prototype of design and implementation of a system which carries out data analysis and prediction that allows clients to configure the system according to their application. The proposed scheme consists of event collector, database, search engine and recommendation engine. After filtering the raw data it is stored in formatted database. The elastic search assists to build the search engine which helps us to parse through the data and SPARK to perform various data analytics. Machine learning is used to build recommendation engine and analysis for the versatile functionalities on available data set to provide suggestions enhancing user experience. Analytics UI helps providing proper understanding of search patterns and user events to increase efficiency of the application in which it is used.*

Key Words: Data Analysis, data prediction, Internet of things, Elastic Search, Kafka, Machine Learning

1. INTRODUCTION

In present era, interactions of customers is happening mostly using internet. Users perform various activities like commerce, businesses, and entertainment using online services more than the traditional methods. Enormous amount of data is generated everyday by their online activities. This data can be utilized in generalizing patterns of user interests depending on their categories, which can be further used by service providers for increasing quality of services and profits. They can use this data for analysis and prediction purpose and for the further suggestion to their users.

In this paper, we are building a system application in which we are analyzing user's data according to the activities performed by them. By extracting the respective field data, applying machine learning operations and data prediction algorithm, we are suggesting vendors trends and patterns of user activities and provide them with suggestions to increase productivity and recommendations for end users related to their previous searches.

2 MOTIVATION OF THE PROJECT

Let us take the example of YouTube. It needs to suggest videos according to user interest or in the case of e-commerce sites vendors need to know what their users need and patterns to increases their sales.

There are very few analysis and recommendation engine developed by considering vendor's needs. Those systems lacks in proper analysis. So, recommendation engine may contain some shortcoming and may not provide optimized result.

3. IMPLEMENTATION METHODOLOGY

Our proposed system uses multiple open source technologies like Kafka, Elastic Search, Cassandra, D3 Charts, and Spark.

3.1 Apache Kafka:

Data will be collected using Apache Kafka-open source tool used for event collection. It is horizontally scalable, fault-tolerant, wicked fast, and runs in production in thousands of companies. [3]

3.2 Apache Cassandra:

It is linear scalability and proven fault-tolerance on commodity hardware as well as cloud infrastructure make it the perfect platform for mission-critical data. System will store all raw data collected from apache Kafka in Cassandra. [6]

3.3 Elastic search:

It is used for efficiently managing data stored in Cassandra and also extracting useful data from raw data. Everything is indexed in elastic search. So, searching operations becomes more efficient. It is schema-free JSON document.[2]

3.4 Apache Spark:

This is analytical tool used for analysis of data. Data stored in Cassandra retrieved by elastic search will be analyzed using Machine Learning algorithm such as WEKA (Waikato Environment for Knowledge Analysis). Several types of analysis will be carried out using Spark. It runs program 100 times faster than Hadoop MapReduce in memory and 10 times on the disk. [1]

3.5 D3 Charts:

D3 stands for Data-Driven Documents. It is a java script library for producing runtime and interactive designing of UI.

D3 Chart will display analyzed result in the form of charts, tables, graphs

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

4. ARCHITECTURE

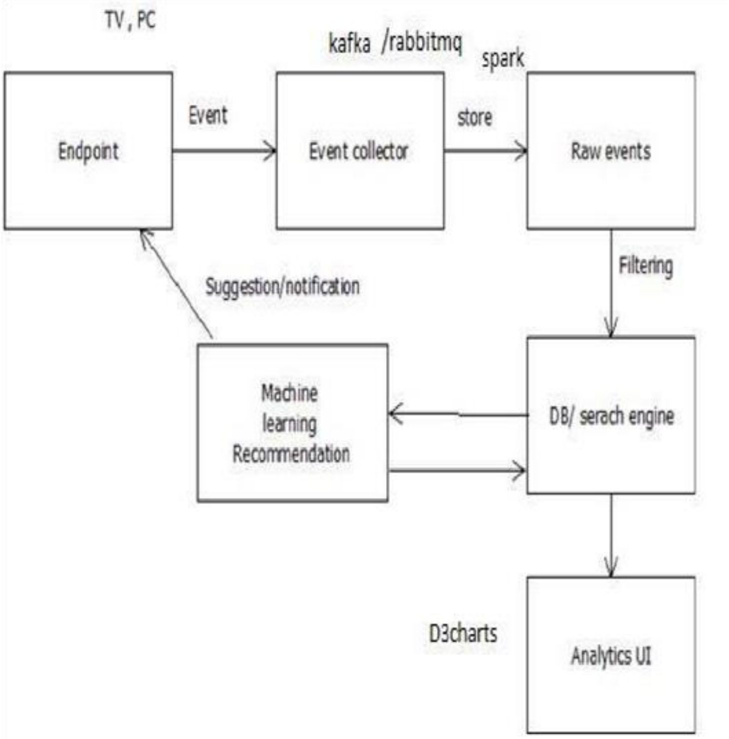

Fig.2 Architecture diagram

Multiple activities performed by users are collected at endpoint using event collector, Kafka. The events may be in form of searches, playing videos, songs etc. These events will be collected using event collector in raw data format and stored in the Cassandra as raw events collection. Cassandra will hold all the collected raw data and data extraction will be performed to get required data form large set of available data.

Data filtering, searching, indexing will be performed over the data by elastic search engine. Apache Spark is used for performing analysis of required data. Only 10% of collected data is used for analysis.

Machine learning algorithm like WEKA (Waikato Environment for Knowledge Analysis) will be applied on data for prediction and recommendation result accordingly.

Analyzed result will be displayed by using D3 charts in the form of charts, tables, graphs etc.

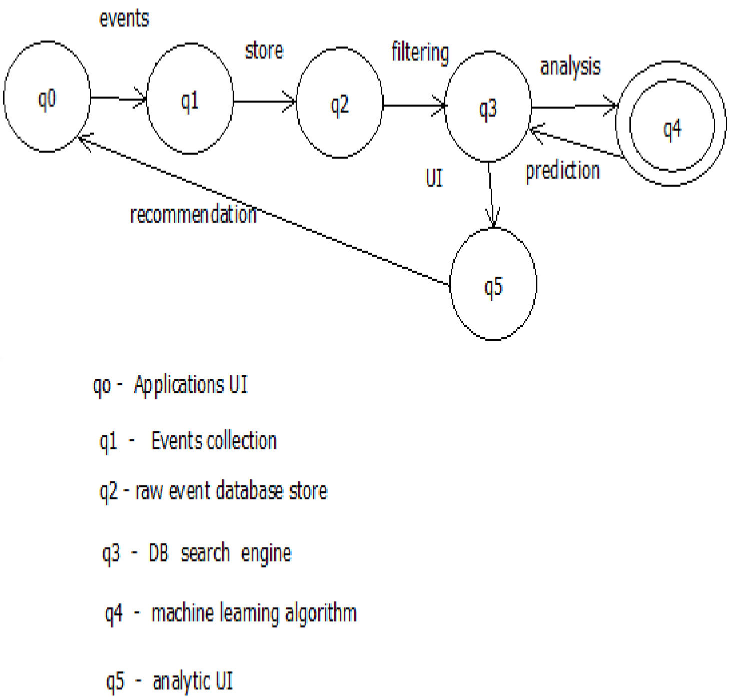

Fig.1 State diagram

5 CONCLUSION

In this paper, we used abstractive summarization instead of extractive to obtain a more precise and accurate summary of multiple documents. The designed optimization framework operates on the summary level so that more complementary semantic content units can be incorporated. The technology stack used can be used to optimize vendor's performance efficiently. Analysis can be correctly performed based on various factors and criteria's which depends on the argument provided to this framework. It is generic framework so performance depends on the application and their arguments. Elastic search can be used for efficient search using indexing and sorting methods. WEKA algorithm can recommend vendors using the analysis performed on the raw data. So, this framework can be used almost everywhere as recommendation engine and analysis displayed using various types of pie charts and graphs.

REFERENCES

[1] http://spark.apache.org

[2] Oleksii Kononenko, Olga Baysal, Reid Holmes, and Michael W. Godfrey
"Mining Modern Repositories with Elasticsearch"

[3] Jay Kreps, Neha Narkhede Jun Rao
"Kafka: a Distributed Messaging System for Log Processing"

[4] https://en.wikipedia.org

[5] Yiannakis Sazeides and James E. Smith
"The Predictability of Data Values"

[6] https://www.elastic.co/products/elasticsearch

[7] https://en.wikipedia.org/wiki/Apache_Cassandra

[8] https://d3js.org/

[9] Hina Gulati "Predictive Analytics Using Data Mining Technique"