

## ***A Review: Big data Analysis framework***

**Uttam Chauhan<sup>1</sup>, Apurva Shah, Uday Yadav<sup>2</sup>, Mahendra Patel<sup>3</sup>**

*Department of Computer Engineering*

*Vishwakarma Government Engineering college chandkheda, Ahmedabad Gujarat. India.*

**Corresponding Author:** ug\_chauhan@gtu.edu.in<sup>1</sup>

### **Abstract**

*Big data is about high volume and often high velocity data streams with highly diverse data types. To make maximum utilization of massive data, there should be some solid framework to manage the data. In this paper, various architectures have been studied for review. The universal architecture for big data is applicable for common massive data application. In many situations, you may need to react to the current state of data. There is also another class of architecture, real-time big data analysis architecture. Real-time big data architecture is to analyze the stream of online data. Streaming computing is designed to handle a continuous stream of a large amount of unstructured data. Batch mode architecture analyzes the data which are stored earlier. We have also studied RUBA architecture which deals with analysis of real-time data. There is some system which provides recommendation regarding some product or service based on experience of pervious customer or consumer. These types of system collects large amount of data and then analyze them called recommendation system. The architecture of a recommendation system, consumer behavior analysis has been presented here.*

**Keywords:** *Big data, big data architecture, big data framework, Real-time big data*

### **INTRODUCTION**

In general, big data can be defined as the dataset that could not be perceived, acquired, managed and processed by

traditional IT software/hardware tools within a tolerable time. [7]. Currently, Big data generally ranges from several TB to several PB [8]. Big data is flattering one of

the most important technology trends that have the potential for dramatically changing the way organizations use information to enrich the customer experience and change their business model. Big data enables organizations to store, manage, and manipulate huge amounts of data at the right time to gain the right insights. The key to understanding big data is that data has to be managed so that it can meet the business requirement; a given solution is designed to support. Most companies are at an early stage with their big data journey. Many industries are experimenting with techniques that allow them to collect massive amounts of data to determine what kind of hidden pattern exists within that data that might be an early indication of an important change. Some data may indicate that customer or consumer buying patterns are changing or that new elements are in the business that need to be addressed before it is too late. Big data carries a variety of data types which may be very complex in nature. As application and nature of big data varies company to company or case to case, the design of big data storage framework also varies. The big data framework plays a very crucial role in analysis of big data and decision support system. This indicates that big data framework must be composed such a way

that storage and analysis can be achieved optimally. Again big data is not an isolated solution. Implementing a big data solution requires that the infrastructure be in place to support the scalability, distribution, and management of the data.

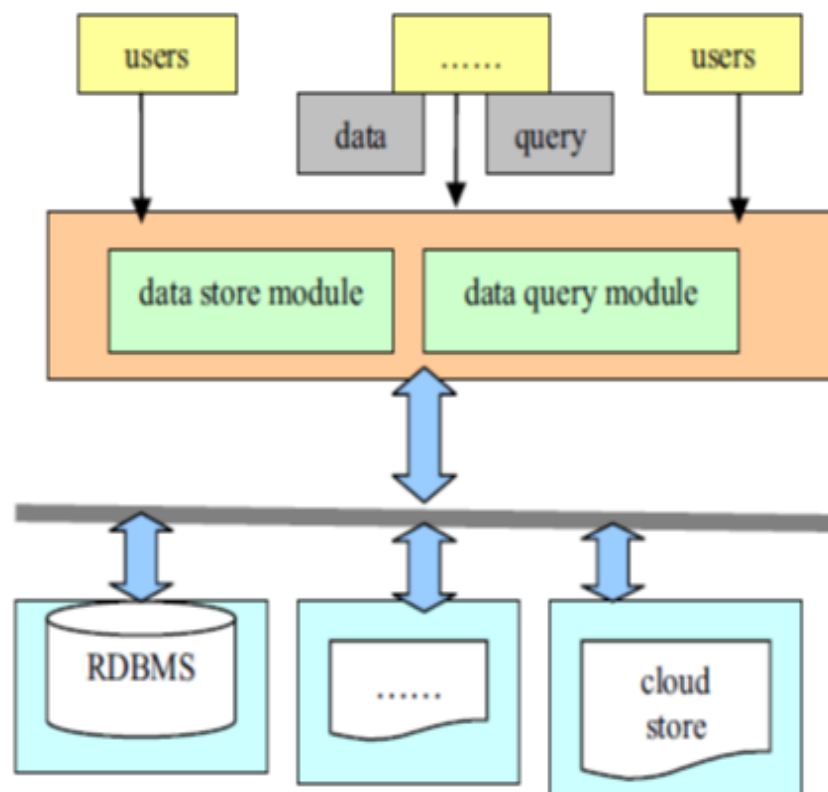
### ***A Universal storage architecture for Big Data in cloud environment***

Few years back down to the line, there was not data generation up to such an extent. With the rapid development in the technology of internet, many commercial and non-commercial applications of internet resulted in the explosion of data generation. This would be continued with exponential rate. In recent years, Internet of Things emerged as a part of development of internet and its application. IoT is collection and interconnection of many embedded devices such as RFID, sensors, actuators, mobile phones etc. with unique IP address assigned to them. In contrast to the traditional database management system, most of the data are unstructured and heterogeneous. Therefore, big data cannot be efficiently stored and analyzed by traditional relational database system. To overcome this situation, researchers have researched and developed NOSQL databases, such as key-value data model, document data model and full-text index

data model. Cloud computing is another alternative which can provides an efficient tool for big data storage and analysis. Cloud computing intended to provide resources as and when required. The use of cloud computing has risen drastically in recent years. The authors have introduced architecture to meet the current issues. They have targeted an ability to store heterogeneous data. **See figure 1.**

In [18], the authors have proposed architecture which can fit many requirement of analysis of big data. The whole architecture consist of mainly two layers 1) Data Analysis layer and 2) Data

Storage layer. There are clusters consists of several nodes in cloud environment. Cluster with strongest computing power is selected as data analysis layer which including data store module and data query module. User may submit query which may generate result consist of vast amount of data without considering the types of data and formats of data. Through unified storage interface data storage module analyze the data. The main task of data storage module is to select appropriate model for normalization. User can submit the query through query interface.



**Figure 1. Universal storage architecture [18]**

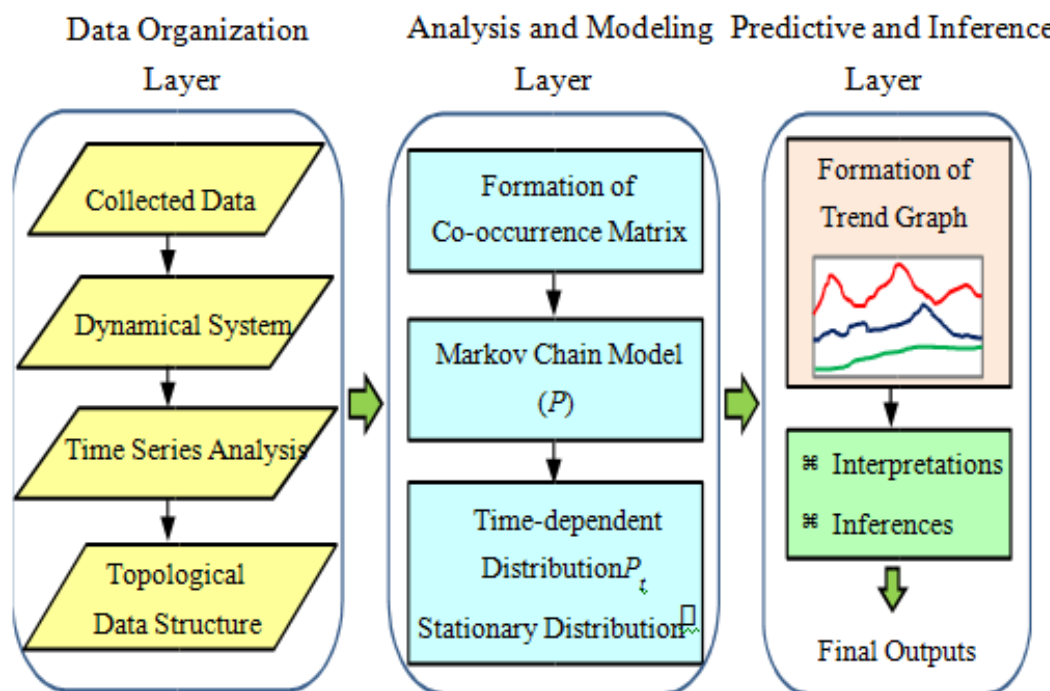


Figure 2. Big Data application framework for consumers [17]

## OVERVIEW OF BIG DATA FRAMEWORK FOR CONSUMER BEHAVIOR ANALYSIS

Now-a-days, in the age of internet, buyers and sellers can explore each piece of information regarding the product or service they buy, sell or use. This information can be availed by many sources like social network platform, network sensors and video servillience system. There may be structured, semi structured or unstructured data. In [17], the Author has presented framework for consumer behavior analysis. The consumer behavior analysis study is about interaction among the consumers, products and operations such as purchasing saving,

brand choice etc. [4]. They have proposed a framework to show how combination of mathematical topology, markov chain theory and co-occurrence analysis can be applied for consumer behavior analysis.

The framework they have proposed consists of three layers. There is data organization layers, analysis and modeling layer and predictive and inference layer. Data are collected and organized in dynamical system so that data can be seen as a time series. They have used the topological data structure for big data. The analysis of consumer behavior can be studied using topological structure. The basic idea behind topological structure is

to discover pattern. There are three key properties of topology that make possible to extract insights from the big data. See Figure 3.

### ***Data Analytics Workflow Framework***

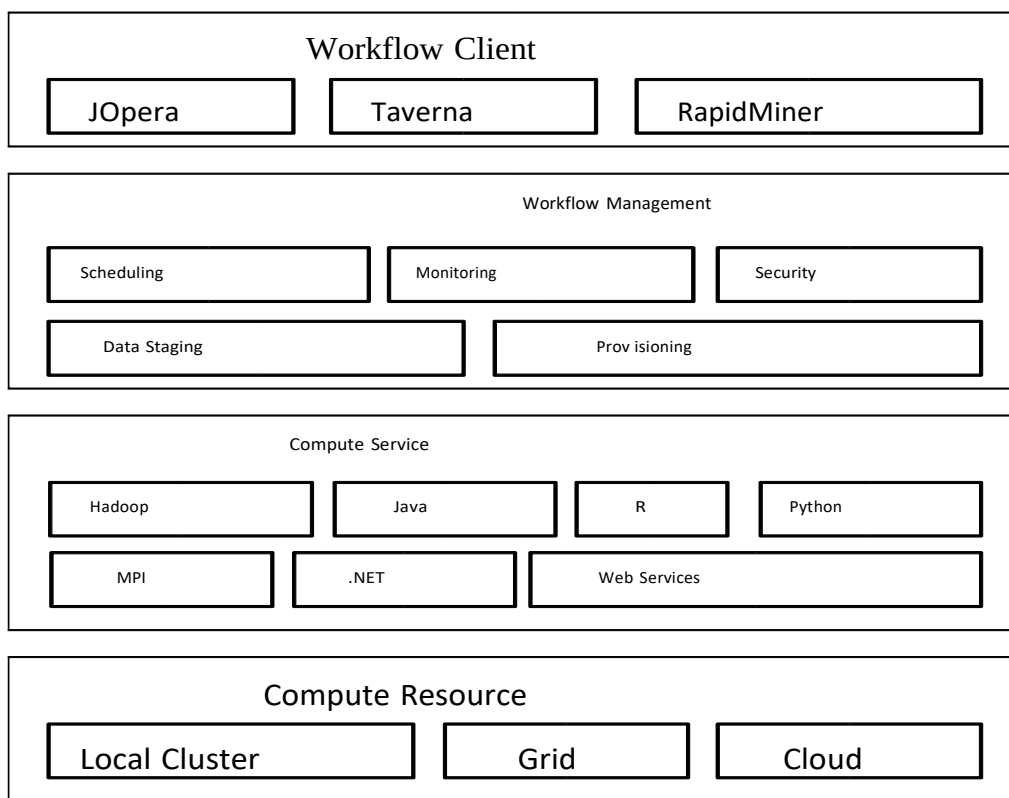
In [19], Authors have proposed the framework for event recognition system in a building based on the patterns of power consumption by lighting, ventilation, electronic equipment, etc. It is a big challenge to recognize what kinds of events happened in a building without additional devices such as camera and motion sensors, etc. Instead, we learn when and how the events happened from the historical record of power consumption. This ERS system generally involves functions of statistics such as correlation, clustering, and regression analysis with a large size of dataset collected in a real time. Massive data processing system like this generally consist of sequence of steps for data manipulation.

A workflow management that includes its composition, optimization, and execution is a significant task of the system. When handling the real-time system, the

workflow management becomes more critical in terms of the system scalability to maintain high throughput or fast speed analysis. The authors have developed a workflow management system which can justify both for remote and parallel execution of workflow submitted by various workflow composers and from various devices such as desktop, laptop and mobiles.

To achieve parallel multiple workflow execution, the task are substituted with parallel subtask. The framework can address workflow/task scheduling, resource provisioning and data staging. The proposed framework claims to overcome many shortcomings. It provides service for remote and parallel workflow execution. It also decreases the complexity for those who are not expert in data mining.

The framework has been tested against dataset collected from over 240 rooms in a building at department of engineering, Tokyo University in 2011. The authors have investigated events occurred in a building from the historical record of power consumption.



**Figure 3: An architecture of the proposed framework [19]**

The architecture mainly consists of four layers. 1) Workflow client layer 2) Workflow management layer 3) Compute service layer 4) Computer resources layer. First layer, workflow client provides user environment where user create workflow using some tools. Those tools provide set of operators. Workflow management layers provides major functions such as scheduling, monitoring, security functions, data staging functions and some others provisioning of this functions are distributed evenly. The third layer provides collection of services. The service can be implemented by various languages/methods e.g. Hadoop, Java, R,

Net and Web services. The fourth layer gives computational platform where executable codes are hosted.

### **RUBA FRAMEWORK FOR REAL TIME BIG DATA ANALYSIS**

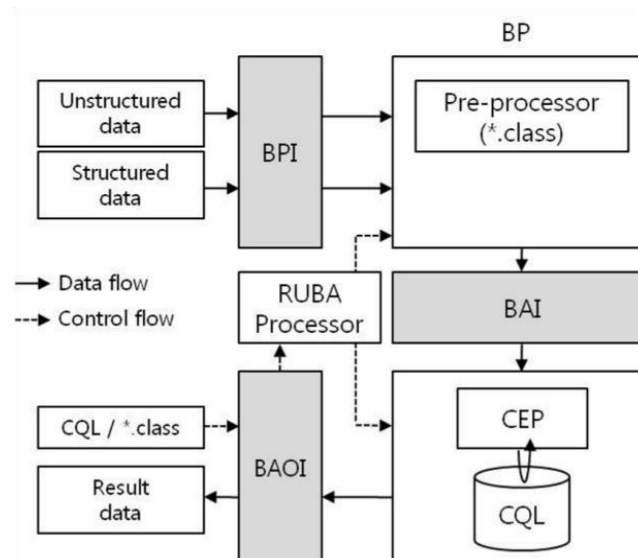
The real time big data analysis to probe crucial information from continuously streaming data. It always difficult to process the data as and when they arrive [1] [2]. The real time analysis is important because data is subject to constant change in their type and values. That change should be identified during analysis phase of data. Real-time big data analysis can be analogues to real-time data mining

methods [9]. There are real-time data mining methods from the work mentioned in [10] [11]. The recent work also has been carried out for the area of real-time big data analysis [12] [13] [14].

The RUBA framework is for real time analysis of big data [9]. The previous architecture are different from RUBA as they are about batch processing of data. RUBA framework follows standard life cycle of data mining technique. The first phase is to evaluate the important from massive dataset and come up with some knowledge. The second phase to derive pattern from the knowledge. There is a serious problem with analysis of big data using batch processing mode. It requires data to be stored and scanned from database. Therefore, this process may be

infeasible for big data analysis. There is also an issue to store big data into the databases. The RUBA framework follows the concepts of continuous query processing. In CQP, query is to be registered in the system in advance. As and when data match with query, the system extracts the data. The CQP is developed as Complex Event Processing (CEP) [15] [16]. Continuous query language is similar to SQL grammar to extract specific data from dataset. If it requires to stop the query from execution, then query either can be paused or deleted from register. Query also can be modified as per requirement.

The core RUBA framework consists of mainly two modules 1) Feature extraction module and 2) classification module.



**Figure 4: Core RUBA framework [9]**

The Pre-processor consists of two modules.

1) Big Data Processing Module (BP)

2) Big Data Analysis Module (BA)

Big Data Processing Module's job is to collect the data from various sources. Business Processing Interface (BPI) collects unstructured or structured data from specific source(s). Then BPI passes the data to business processing Unit (BP). BP prepares the data for analysis. It applies one or other preprocessing techniques to prepare the data for analysis. Once data is ready to be analyzed for particular objective(s), data is sent to the big data Analysis module (BA) through big data Analysis Interface (BAI). BA consists of two components.

1) Complex Event Processing (CEP)

2) Continuous Query Language (CQL)

Big data analysis component is based on CEP [3] and it has input queues of data stream. BA applies analytic techniques and produce the result. The result is transferred to user through big data analysis Input and Output (BAIO)

## DISCUSSION

Its doubtless that big data analysis is heavily depends on the architecture of the particular application. The whole

framework of big data application or the several components of the architecture may differ according to nature of an application of big data. The architecture generally consist of two major parts, preprocessing of data and then analysis of data. The Big Data framework also can be categorized upon batch processing mode and real time processing. The real time processing should be adaptive in nature. Therefore, as and when net datasets becomes available, it must be adapted by existing model of analysis. It should not be in incompatibility state. As a result of this provision, the analysis model can be placed in up to the date state. The other scenario of the same case is one must be able to give small portion of the large dataset as an input in the existing model. Training and testing of the Analysis model must be made on real time data. Therefore, the accuracy and effectiveness can be measured for actual environment. Research can also be carried out to compose batch mode processing and real time processing in one framework. There is open house for the customizable architecture which can fit more than one big data kind of application.

## CONCLUSION

With the increase of use of internet across the globe, the amount of data being

generated also increase with drastic speed. As we have seen some of the framework for big data application, it has been observed that design of framework of big data plays very important part for the efficiency of the system. The accuracy of the system also highly dependent on the components used in the architecture of big data. We have observed that real time analysis requires complex components in the framework. It may be seen that the nature of the application performs very crucial role in the design of the framework of big data.

## REFERENCES

1. A. Arasu, B. Babcock, S. Babu, J. Cieslewicz, M. datar, K. Ito, R. Motwani, U. Srivastava, and J. Widom, "STREAM – 9]The Stanford data Stream Management System", Technical Report, Stanford InfoLab, 2004.
2. B. Babcock, S. Babu, M. Datar, R. Motwani and J. Widom, "Models and Issues in Data Stream systems," in Proc. ACM PODS 2002, 2002.
3. J. I. kim, N. W. Kim, S. K. Yun and B. T. Lee, "A study on CEP performance in mobile embedded system," in Proc. ICTC2012, 2012, pp.15-17.
4. Analyzing customer behavior Predicting-what happens next, Publisher: Across Technology, Editor: Peter Hinssen.
5. Kaufman L, Rousseeuw P J. Finding groups in data: an introduction to cluster analysis [M]. Wiley-Interscience, 2009.
6. Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]. Kdd, 1996.
7. Big Data: A servey Min chen, shiwen Mao, Yunhao Liu, Springer Scinece and Bussiness media New York, 2014
8. Manyika J, McKinsey Global Institute, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute
9. Jaein Kim et al. RUBA: Real-time Unstructured Big Data Analysis Framework, ICTC 2013
10. R. Agrawal and R. Srikant, "Mining sequential Patterns," in Proc. ICDE`95, 1995, pp.3-14.
11. J. Pei, J. Han, B. M. Asl, H. Pinto, Q. Chen, U. Dayal and M. Hus, "PrefixSpan: mining sequential

- patterns efficiently by prefix-projected pattern growth,” in Proc. ICDE’01, 2001, pp.215-226.
12. U. Yun, “WIS: Weighted Interesting Sequential Pattern Mining with a Similar Level of Support and/or Weight,” ETRI Journal, vol. 29, 2007, pp.336-352.
  13. C. F. Ahmed, S. K. Tanbeer and B. S. Jeong, “A Novel Approach for Mining High-Utility Sequential Patterns in Sequence Databases,” ETRI Journal, vol. 32, 2010, pp.676-686.
  14. J. I. Kim, P. S. Choi and B. H. Hwang, “Real-time Sequential Pattern Mining for USN System,” in Proc. ICUIMC`12, 2012.
  15. A. Arasu, B. Babcock, S. Babu, J. Cieslewicz, M. datar, K. Ito, R. Motwani, U. Srivastava, and J. Widom, “STREAM – The Stanford data Stream Management System”, Technical Report, Stanford InfoLab, 2004.
  16. J. I. kim, N. W. Kim, S. K. Yun and B. T. Lee, “A study on CEP performance in mobile embedded system,” in Proc. ICTC2012, 2012, pp.15-17.
  17. Thi Thi Zin, Pyke Tin and Takashi Toriu A Big Data Application Framework for Consumer Behavior Analysis, 2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE)
  18. Qingchen Zhang et al. A Universal storage architecture for big data in cloud Environment, 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE
  19. Chonho Lee et al. A Workflow Framework for Big Data Analytics: Event Recognition in building, 2013 Ninth World Congress on Service