JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Evaluation of Information Extraction Architecture in Big data Analytics

[1]Manjunatha Swamy C, [2]Dr.S.Meenakshi Sundaram, [3]Dr.Lokesh M.R
[1]Research Scholar, [2]Professor and Head, [3]Professor
[1]Department of CSE,
[1]GSSS Institute of Engineering and Technology for Women, Mysuru, Affiliated to VTU, Belagavi, India

Abstract : In Big data analytics information extraction and classification plays vital role to process the data in batch mode and in stream mode to support this mode of analysis and processing must require architecture. Architecture is backbone of any application to work successfully it handles data ingestion which collects large data and files generated from different sensors and actuators into storage medium which further represented in report using modern tools. There are many architectures have been introduced from industry and academics, universal acceptance of these approaches for other types data analysis is not feasible. Some architecture solutions are not standard big data processing architectures. In this work we proposed smart Big data architecture to extract information and analyze the performance of information extraction using layered approach model to address data collection, processing, storing and data reporting and visualization of large data both in batch mode and real stream mode. Proposed work shows significant improvement with other architectures.

***Index Terms* - Big data architecture, Feature extraction, Feature classification, Data analysis and Storage, Data visualization, EBBA model, PIMM model.**

I. Introduction

Big data architecture is vital in bringing effective working system to perform information extraction over massive data. Big data architecture is a framework to perform processing of huge data by providing suitable infrastructure and solution to the various problems based on the organization needs. Data is collected from various resources these data is segregated as batch data and real time data which helps to process and represent data in a proper semantic way. Big data architecture is to be designed[12] to handle the ingestion, processing, analysis and reporting of huge data. Data generated is classified into structured data, semi structured and unstructured data, many tools available to handle the data types in well defined way. Batch data is analyzed and processed to data which is already stored by using tools like Apache Hadoop, HBase, Snowflake and Ansible. For example billing and payroll is processed weekly or monthly. Streamed data is real time generated continuously from different sources as data records. Many tools which do processing of data includes from Apache as Kafka, Spark[19], and Azure stream analytics.

Feature extraction is the process of transforming raw data into meaningful data by selecting relevant features attributes in a system. Dimensionality is reduced here by making relevant groups, a distinctive feature of this large data set is that they contain a large number of variables and these variable need additional resources to process it. Selection of particular variables and combining with other set of variables leads to reduction in the data size, the obtained results is evaluated with recall and precision parameters. Principal component analysis (PCA) is used in this dimensionality reduction techniques and it is an unsupervised learning technique. Feature selection also used to highlight the importance of feature in data set to remove less importance features. Data from sources may be linear and non linear in that case to deploy SVM kernel based PCS technique to convert nonlinear data into linear data.

Feature classification is very important to extract the required the data in specific period of time. Classification is must to incorporate if security comes to tackle. Data generated from sources is classified into mainly structured, un structured and semi structured data. Structured data is organized in well defined tabular form for effective analysis like Structured query language SQL format, unstructured data not follows any predefined row and column format even it is very important in analysis process it uses No-SQL format is used to handle unstructured and semi structured large data set even it is not in tabular form. High volume of data is analyzed and processed in real time is possible using NO-SQL and accuracy can be obtained this technique. It grows horizontally to store data and to process key values model in which data is stored as key value pair to make more available data base.

Intelligent data analysis in highly in demand technique to represent unknown valuable information from huge data set and discovers, communicates meaningful patterns of data to make suitable decisions to handle data and later it is visualized in specific format. Based on the type of context analysis also changes it may be predictive analysis, descriptive analysis, diagnostic analysis and prescriptive analysis. Further we analyze the data using correlation method and regression methods. Data visualization is used to represent data in more acceptable format many tools are used to represent data to user in easy understandable way and is very effective in getting patterns, trends and outliers. In this work data visualized using numerical data visualization using discrete data format.

Accuracy of classification in high dimensional data using feature selection model improves the information extraction in binary bat algorithm, selection of optimal subset in available dataset, selected data enhanced to improve classification performance of actual dataset, further noise is eliminated which help the algorithm to execute faster and more efficient[14]. Feature selection involves filters and wrappers approaches to measure the relevance score of each feature, wrapper use classifier to evaluate selected subsets obtained by the algorithm later any modification required also considered. To optimize selected features heuristic approaches is proposed in genetic algorithm, Particle swarm algorithm and Ant colony algorithm. Here proposed bat algorithm as more conventional to select features in bio inspired[1] algorithms, these algorithms used to develop new techniques to increase the robustness of algorithm. Many intelligent algorithms available such as ant colony algorithm, Particle swarm, Genetic algorithms, Grey wolf algorithm and Elephant search algorithm, among these algorithms Bat algorithm is demonstrated as effective algorithm to extract feature information as it is flexible to enhanced. Data mining application supports different phases of application as data mining these includes preprocessing, discovery of pattern and evaluation.

Many real time applications objects[7] will have different representations in different views, the communication between different views as each view is contributing uniquely and multi label prediction is also considered in the proposed model to calculate the gradients as to update weights of network given below using PIMM approach, Later in order to extract specific information of a particular view but we do extract from base information by excluding all shared information, the whole framework is to minimize the loss with respect to parameters of PIMM model. Every iteration verifying with multi label sets if the condition is holds good then fitness of data will be measured using standard function. As observation illustrates that each label is associated with unique feature with data, then label is added with function add() to verify the adaptability, then combination suitability is constructed if not associated then standard function used to generate data with random() function

The flow of paper starts with related work and architecture of the Big data model as low level design and high level design, Algorithm to define process of multi label approach. Mathematical model to support implementation part, later comparison of various attributes like Recall, F-Score, Accuracy, HLoss, HLoss(D,L), Mean and standard deviation to give how comparative best analysis Big data then result and discussion and finally conclusion.

II. RELATED WORK

2.1 Big data architecture for various Applications

Architecture varies with different application in which data is processing many application architectures are proposed such as cloud computing architecture, Architecture for smart cities and companies, Enhanced telemedicine systems, digital government[26], Advance metering infrastructure data, Crowd Learning technique, Predictive maintenance of railway points, Search Engine optimization, Electric mobility as a service in smart cities and many more. As noticed all these architectures is used for all the cases of data analysis and representation much.

IOT based edge enabled architectures provides good security but with less scalability. Data heterogeneity and processing time of data is significantly low, enhanced model of architecture to process and store Big data efficiently in the network by using edges to process and store the data at the edges of the network with technology. It compares the architecture with two layered models as IOT edge and cloud processing and YARN technique is used for managing the cluster efficiently.

Mariagrazia speaks about Big data architecture for smart cities[11] and smart companies based on SIBDA approach (System Innovation Big data analytics) has focus on document processing, email application and sensor of networks and as per the requirements with adapted requirements and provides implementation along with experiments and to make architecture best suited for applications.

Mehdi Fahmideh addressed about how to select suitable Big data architecture to satisfy the constraints to handle uncertainty. This method is to handle goal oriented modeling to identify adoption in manufacturing systems and other method is to illustrate with reengineering of collaboration system to a new Big data architecture. Fuzzy logic also incorporated with this approach.

Mert Onuralp, Mohammed Zaki proposed an idea to find Big data architecture to overcome shortcomings and strength in developing architecture for organizational requirement. Available open source Big data architecture for businesses to take capabilities. Main focus is on how to address all challenges in the architecture and to reshape the structure when ever required.

Developing an architecture for spatial data[23] by purnima is about handling wide range of data as batch mode, iterative mode and interactive to find a unified solution to manage massive data. This process involves data preparation, analytics and visualization this model is developed to load, store and process with a query.

2.2 Information Exaction using Enhanced Big Data Architecture

Big data analytical architecture is used in any open source data including massive data. The proposed architecture is designed to provide flexible, scalable, adaptable and cost effective design with existing architecture. Architecture designed with Data collection, Data integration, Data dimensionality reduction and Data transforming with data visualization. In the diagram below data from various sources sensors actuators are collected as the first part of the process. Various tools are used to extract information from various applications it may be Sqoop ETL (Extract, Transform, Load), Amazon redshift, Azure Synapse SQL, Google looker, Zoho Analytics, Hive SQL. Sqoop ETL is popular tool used to roll in roll out massive data from unstructured non Hadoop data store into easily accessible and used by Hadoop and later use by HDFS. It used structured data repositories like already stored data in the form relational data base tables and enterprise data warehouses. It uses many slices to split the data into multiple formats and then load the data into HDFS file system also can be reloaded into database to maintain the schema. To achieve the fault tolerance it deploys YARN (Yet Another Resource Negotiator) is high scale import and export of data in HDFS.

Amazon redshift is big framework which provides the database to store data and allow querying to database in parallel. This tool allows the data store and query to database without any delay and data loss. Redshift uses RA3 (Redundant Array) for clustering and POSTgrey SQL to query the data and it act as open source relational database system. Unstructured data is processed by using columnar data format representation. Azure SQL is a cloud based always up to date fully manageable data base this has multiple services used in big data architecture historical data is stored in data store and it handles different data. System allows handling all type of data as structured, unstructured and semi structured data as well into scalable data which enables storage, processing and analytics.

Google looker is uses insight of the data and automate the process irrespective where the data is located it uses BigQuery with databases which is part of Google cloud platform and provide database as a service. It process the data speed and accuracy uses No-SQL format of data these data is not processed with data in columnar format. Hive SQL also allow the data to read write and manage using SQL. Process the structured data using Meta store and data is analyzed and processed, it uses MapReduce framework for executing queries and OLTP (Online Transaction Processing) facilitate high volume of data in very less time with reliability.

Input is then given to the layer which consists of feature extraction[3][17] and feature classification very important process in which gathered data is converted into feature subset of the data and dimensionality of the data is reduced. Once data set is transformed as feature subset then focus is on only the features set, instead of having all the variables of the data it's better to have only few data features to characterize the few variables in the analysis process and can combine with few other features as well then evaluation is become more precise and accurate by using recall and precision factors as well. There is structured and unstructured data out of which on e of the popular dimensionality reduction method is PCA (Principle component analysis) which is unsupervised algorithm. Feature generation is also important role in which set of features is selected and each feature is evaluated separately in order to make it more accurate by ignoring meaningless interaction. Feature evaluation is well organized to get more work done each feature is utilized for the current needs and unimportant is ignored or reduced to make model more responsive and to avoid inconsistency in the data.

Feature classification is more important in Big data to access more relevantly and consistently. Data security is another aspect in which categorization holds key role with specific period of time. Data is classified into structured, un structured and semi structured. Structured data is in the form of well predefines tabular form having rows and column representation to provide fast access to the data whenever it is required. Unstructured data is not follows ant table like format like SQL format it generates randomly and is more important in data analysis process which uses No-Sql database formats like Word, PDF. Semi structured is combination of structured and u structures some part of the data is easily fit into the model and other part is hard to fit into the model of data lie XML data.

Once feature extraction and classification is completed then real time data access on demand or steaming and batch mode data access is executed. Data storage also plays a vital role in achieving highly interactive, scalable Big data architecture. Proposed Big data architecture has various tools to support data storage. Hadoop, HBase, Ansible, Amazon S3, Kafka and Google data flow is very much used to support and enable strong data storage. Hadoop store massive data efficiently and process the data here instead of using single large computer clusters of computers is used to do processing parallel. Instead of rely on single computer multiple computers through four components HDFS (Hadoop distributed file system), YARN and Hadoop Mapreduce. Mapreduce is the core of Hadoop system and Name Node captures structure of the file directory and place the data in the form of chunks across the Data nodes to achieve multiprocessing parallel.

HBase is open source used to store non relational database and aloe the access via HBase API similar to SQL.This model consist of client, HBase and Zookeeper, HMaster controls and coordinates many number of region server. Region server is collection of number regions, all region servers are connected to HDFS system which frames HBase architecture. Zookeeper in HDFS is a repository it stores different applications data and stores in it also allows retrieving the data for accessing. HBase handles real time data processing and random read write operation can be performed to it. SQL is handled in it and it uses No-Sql operations and it is implemented on top of the HDFS system and offer fault tolerance.

Ansible is trend in technology provides a platform to perform configuration management, application warehouse and different service orchestration and grouping to offer provisioning. It is a collection various components as core modules, Host inventory, Playbooks, plugins to connect different hosts. Ansible playbook provides reusability if one task to be executed multiple time then write a playbook and keep in source control. YAML representations are used by Ansible to arrange and automate the process. Amazon Kafka is used to process the real time streaming[28] data by creating cluster then it uses Amazon EC2 and Lambda with data analytics connected to producers and consumers to run streaming applications.

Data visualization is very vital process in representing data and capability to handle large data set into visual treat many tools are used to handle these data to represent as a chart, graph forms. Tableau, Cassandra, Grafana are the popular tools in handling data visual. Tableau is end to end analytical platform allow to prepare the data, analyze the data with collaboration to find valuable insight of the data. Starts with data sources and acquisition will lead to data storage then it used for analysis and to produce report in an acceptable pattern. This allow to connect quickly connect, visualize and share the data with encryption. Cassandra provides No-Sql data management to handle large amount of data set across various servers with high availability with no single point of failure concept. It handles structured data with high scalability, fault tolerant and consistent and with all these features it is deployed by big IT companies face book, Twitter. Key elements in Cassandra performs unique functionalities data center is a collection of related nodes with clusters of data centers which maintain a log where all operations is stored. Memory is divided as Mem-table, SS-Table and Bloom filter, once log is create then it is moved to Mem-table once the buffer is filed later it is moved to SS-Table and blooms filter allow the queries to handle it.

Grafana is another tool provides interactive data visualization via graphs, charts and alert system to support data sources also monitors the system. Create explore the data with attractive dash boards is a highlight of this tool. POSTgreSQL is used by grafana to get data to visualize and various plug in are used to perform data connections and offer functionality. ETL process is used by Grafana, event can be streamed using Rudderstack which maintains data warehouse, data transformation and orchestration and finally data visualization using Grafana cloud.

III. BIGDATA ANALYTICAL MODEL TO EXTRACT INFORMATION

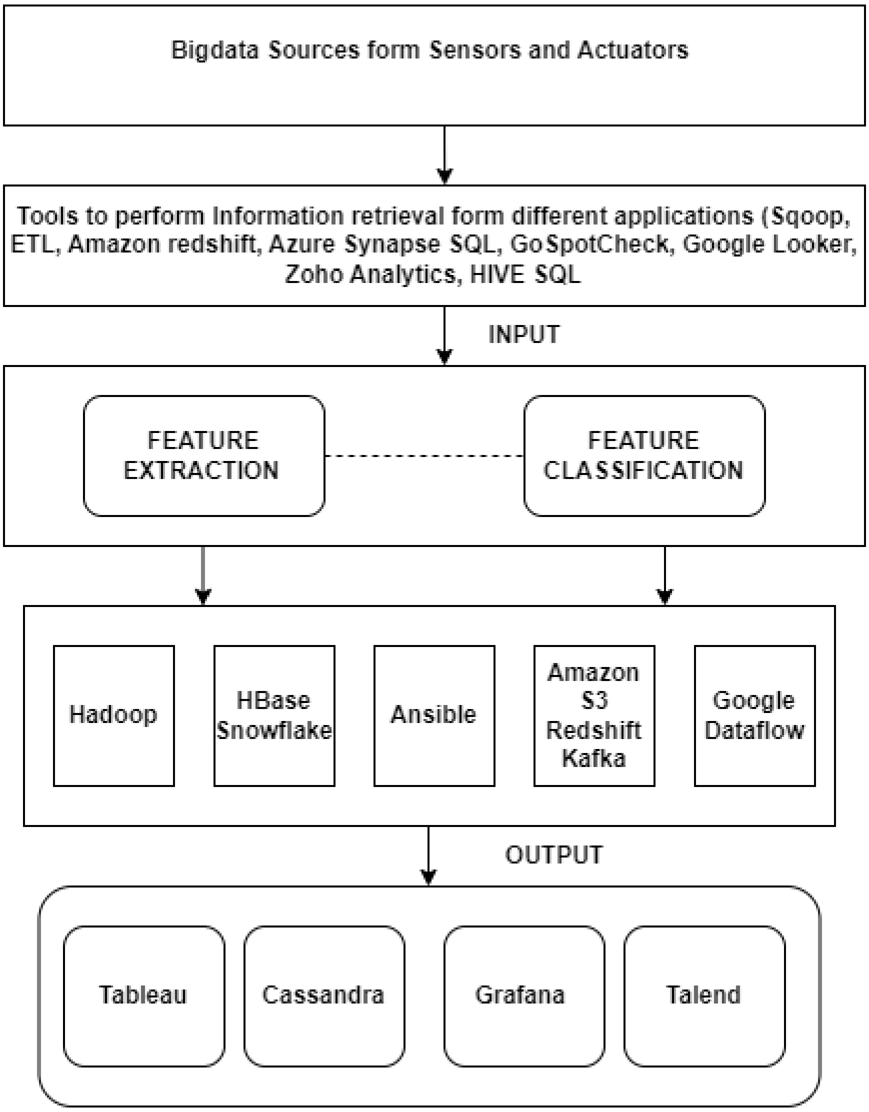
Bigdata Sources form Sensors and Actuators
Tools to perform Information retrieval form different applications (Sqoop, ETL, Amazon redshift, Azure Synapse SQL, GoSpotCheck, Google Looker, Zoho Analytics, HIVE SQL
INPUT
FEATURE EXTRACTION
FEATURE CLASSIFICATION
Hadoop
HBase Snowflake
Ansible
Amazon S3 Redshift Kafka
Google Dataflow
OUTPUT
Tableau
Cassandra
Grafana
Talend

Fig 1. Big data Architecture with feature extraction and feature classification

The architecture of web information extraction is shown above which represents how data is received as sample data and collected from multiple sources used to manipulate and analyze representative subset of data points and to find data patterns, which classified into trained data of data set to grow single tree and data to estimate the errors. Data is trained to understand how to apply technologies and to make decisions.

Then feature selection is used to transform raw data collected as inputs to the model then that data is required to some algorithms. This process reduce the number of features by creating new features from existing data which helps to reduce set of features, data applied to select predictors to split data using best predictors so that estimate errors by applying tree to data if not true repeat until stop tree growth is fulfilled. data scraping performed to make data feasible, After estimation with data errors if it as expected go to next level that is random forest by collecting all trees else repeat until specified number of trees to obtained. Sample data and estimation process is correlated with each other, grow tree and feature selection of predictors is related to each other hence random forest algorithm will work as described. Data estimation makes data consistency to avoid failures.

IV. BIGDATA PROCESSING MODEL TO EXTRACT TRANSFORM AND VISUALIZATION

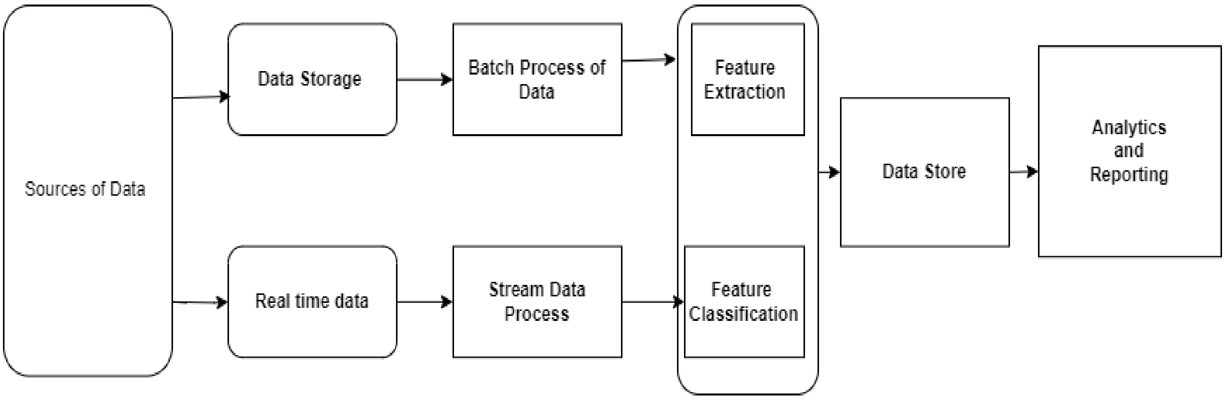

Fig 2. Architecture for Data processing and Reporting

In the above diagram shows analytical processing of the model initially with data gathering from various sources, sensor and actuators. Data storage of all the data which may from already existed data store or it may be real time data as a stream data. Batch processing analysis and processing of data over a set of data is happened over a set of data which is already stored over a period of time like scheduling of batches. Batch processing is executed by many tools available VisualCron, Ansible, Active on batch scheduling. Another part of data collection is real tie stream data which is generated by multiple massive sources as sensors and actuators. Data is sent as multiple data records simultaneously. Amazon S3, Redshift, Google storage, Kafka is important tools to provide the storage of real stream data from sources.

Feature Extraction and classification over the data set collected is applied to impart on the technical features from the data set, the obtained subset of features is evaluated for performance analysis of the model to bring insight of the data value. Classification also plays a key role is grouping data based on the inter similarity or intra similarity. It makes the groups of data by separating and organizing of data based on the relevancy. Once this processing of feature extraction is implanted then data storage is must be provided to do this many tools are available as Hive, Strom, HBase, SparkDrill, YARN representation based on the current requirement is offered b the storage block. Next to represent data in visual form and reporting to end user. Real time decision management, alerts using screen card and dashboards. Data discovery is self service search offered in predictive analysis method. Various analytics is used as statistical analysis, Text analytics, data mining are executed to represent data in highly interactive, visual attractive to make easy to understand of data in a system. Tableau and Grafana is most used framework to handle massive data stored from different sources in effective way.

V. PROPOSED ALGORITHM FOR INFORMATION EXTRACTION

In the proposed approach, information extraction using the Bat algorithm with K-Mean classification is described below:

5.1 Enhanced Binary Bat Algorithm for Information Extraction

An algorithm shown below demonstrates how a random tree classifier[16] which classifies data to make decisions, training with replacement of data to produce a new data set. Later, a new tree is constructed with instances. If only one instance belongs to that tree, it returns that itself; otherwise, it selects randomly by splitting available features in the set. The tree with nodes N, Frequency F is modified into possible value sub child nodes and the procedure is repeated until trained error-free content is obtained. The bat algorithm[18] is provided below, with input, output, and various steps involved.

Extraction process is improved based on the factors used as frequency, loudness factors to generate new solutions to demonstrate how random tree classify data to make decisions by selecting suitable classifier then training of data with replacement to produce new data set.

Input: A trained dataset with the attributes frequency and loudness
Output: Preexistence Dataset, Optimized Data Set
Step 1: Data set in the population with specific parameters for frequency, velocity, and loudness.
Set the pulse rate. Pi and loudness Ai

Step 2: Generate new solutions by updating frequency, velocity and positions, while (t< maximum iteration)
Step 3: Apply random function concept.if(rd>ri)Select solution among best solutions randomly obtain Ls solution around BsEndif
Step 4: Generate Rs using Randfly() Randfly( )if(Rd<Ai)& f(xi)<f(B)Get Solution increase Ri and Ai
Repeat Endif Obtain Rank and Get Final best Fb.

The algorithm clearly states that the data is a randomly trained data set that is then checked for reproducibility with a preexisting data set. Possible features of the data are divided to create an N number of child nodes as an instance with relevance to the dataset taken from the build tree to optimize the information. The proposed algorithm avoids limitations in data lists with repetitive occurrences.

5.2 Adopted multi label based classification Algorithm for Big data Extraction

An algorithm shown below, demonstrate how ensemble technique which ensure better performance obtained from any of the dataset and to compare two or more different analytical model and to synchronize results too increase accuracy of data retrieval methods with respect to boost random forest model is an best approach, *also to increase classification* performance of a model. Every iteration verifying with multi label if the condition is holds good then fitness of data will be formulated using function fitness(). Each label is a*ssociated with unique feature* with data then label is added with function add(), then combination suitability is constructed if not associated then

sub_child=cross_over(p1,p2) mutation=TRUE

mutation mut_child=mutation(p1, p2) To get new possibilities.

Generate sub nodes of set, as p1,p2….pn if, here F is associated with (F1…,Ff)

do for I range from i=1 to f

Recall to function Cross_over(p1,p2) if ends

for ends

In the algorithm [30] it is very clearly specified that data is randomly trained and rechecked for reproduce with preexisted data set. Using label approach data is combined with different patterns and possible features of data set create N number of child nodes as instance with relevance to the dataset taken build tree to resolve the efficiency and to optimize information. Algorithm proposed avoids limitations in with data lists with repetitive occurrences.

VI. RESULT AND DISCUSSION

6.1 Software Requirements

To demonstrate the significance improvement in the proposed big data architecture by using Tableau software every system is associated with Windows 10 operating system, 4 core CPU, 16GB RAM, 2GB free disk space and SSE4 (Streaming SIMD Extension 4), virtual environment as Microsoft Hyper V. Tableau desktop is Unicode enable and compatible with data stored in any language. Grafana is also one of the best tools to use with minimum memory of 255MB and 1 CPU with data base SQL. It uses set of commands grafana>conf uses 50GB of Logs is framed to visualize the stored data. Test on data set is using comparative methods and evaluation metrics. The data set considered is ionosphere.csv file with 34 features set, 7450rows of data and validation factor considered as 70 to 30 over the split up data using metrics as xtrain, xtest, ytrain and ytest. Other parameters as K value in KNN, number of variables, maximum number of iterations to perform feature selection. The data is modeled with selected features as number of train, number of validation to increase the accuracy and to obtain data convergence and used python libraries.

In general, three approaches are usable: web mining usage, content mining on the web, and structure mining. In addition to the combined tags and value similarity, DOM (Document object model) tree structures can also be used. Redundant data records RDR rule, QRR query-related record extraction, operator used The DOM model, the Machine learning method, and the successive steps of the proposed method, the iForest anomaly detection algorithm are listed.

Random Forest and Multi-Layer Perceptron (MLP) classifiers are also used in order to address web information extraction and make it more efficient. Fitness of attributes has been obtained for a number of iterations, as shown in section 4, which describes how algorithm yields better fitness with number of iterations, the algorithm affects fitness attributes over n number of iterations using -1 iteration values and in figure 6 elaborated to describe KNN attributes using N iterations in each case it significantly shows the improvement in obtaining information extraction.

6.2 Comparative Discussion

Traditional Big data architectures systems provide less data accessibility with time factor. Proposed Big data architecture performing better with other architecture approaches. Enhanced Big data architecture is compared with edge cloud computing, Detection of Intrusions, Mobility, Engine optimization, Enhanced systems and with proposed model. Significant model enhancement is performing data accessibility better with other models and advanced system models use automated methods to extract information. The contents and model structure play a vital role in establishing relations between a page and page level attributes. If there are any changes in the model, then the wrapper function will enable the deployment to make the function work. Compared to our model, it's not possible to ensure the efficiency and here more complication and achieve better extraction.

Dom Tree Model, the Mining data records MDR Algorithm, the improved HMM (Hidden Markov Model), and the Long short text classification method LSTM used in neural networks are compared for attribute accuracy and flexibility in selecting designs when changes occur in websites. Clustering algorithms are applied to formatted and unformatted records in data slots. If any modifications in layout design are not accommodated, these shortfalls are used to generate layout designs with different patterns.

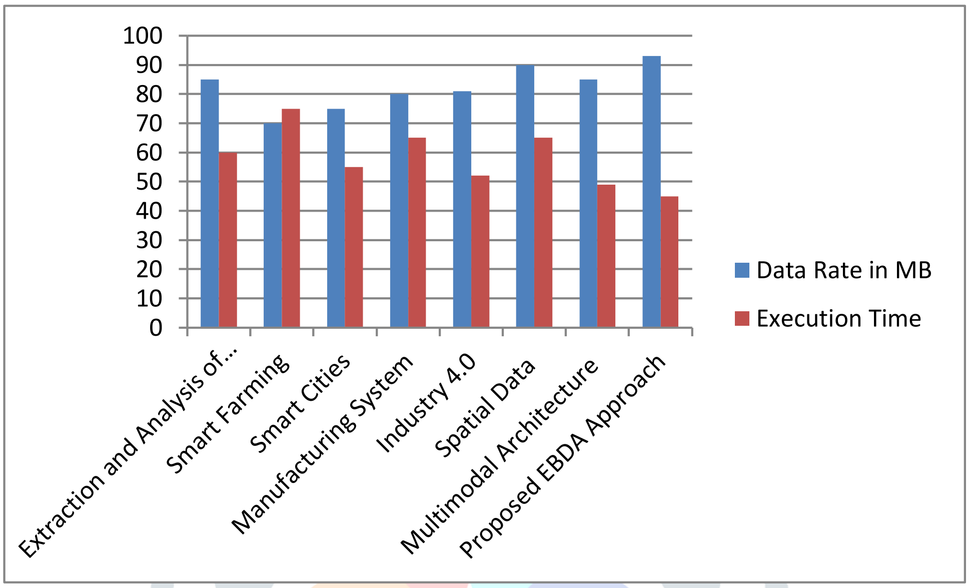

Fig 3. Performance comparison with two attributes.

In figure 3 Data rate is the processing abilities of architecture over the collected data. Data is collected from various sources which is structured or unstructured some data is stored in batch mode which consumes more time than the real stream data. Real stream data takes time in minutes and already stored data takes more time in hours. Various architectures shows average data processing on collected data and execution time is measured in minutes. Proposed architecture significantly shows betterment in data processing and execution of data in 40 min time units and table below shows all different approaches attributes are tabulated.

Table 1. Comparison of different approaches with two attributes

Approach Used	Data Rate in MB	Execution Time
Extraction and Analysis of EHR Data	85	60
Smart Farming	70	75
Smart Cities	75	55
Manufacturing System	80	65
Industry 4.0	81	52
Spatial Data	90	65
Multimodal Architecture	85	49
Proposed EBDA Approach	93	40

Figure 4 Describes performance of existing approaches with the proposed approach using data recognition rate here incoming data is segregated based on the type of data here data may be unstructured and structured. Classification of data in the collected data set is streamlined using tools like business intelligence, neural network and random forest algorithms. Feature selection model is used to find the subset of data to focus on meaningful data using PCA (principal component analysis model). Table shows all architectures compared to analyze the performance of data recognition rate, classification and feature selection of various architectures. In the table and graph it's observed that proposed model shows high rate in all three attributes of data.

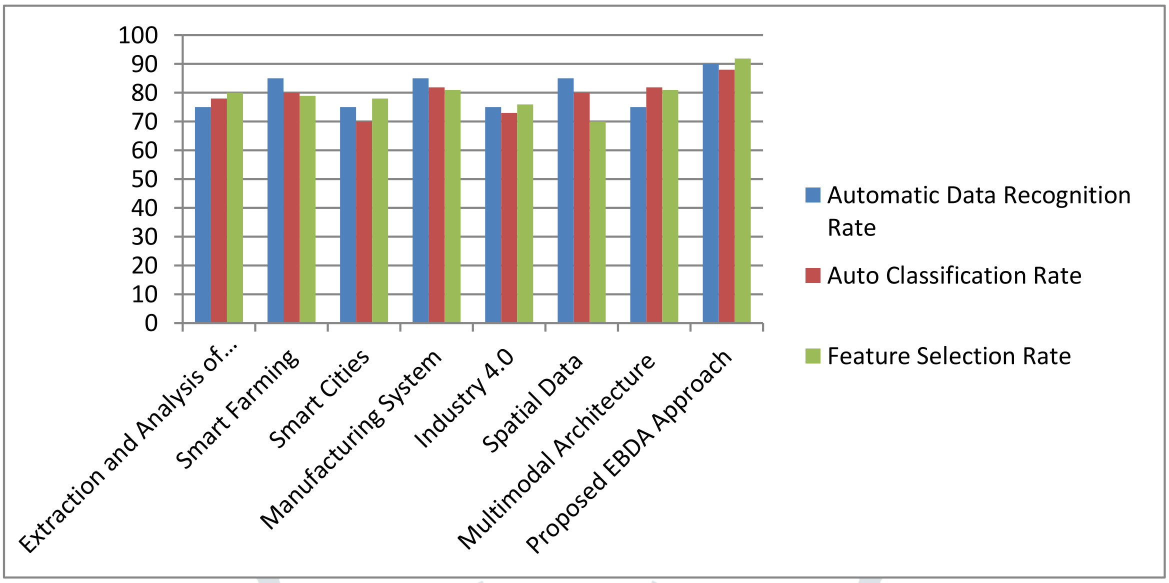

Fig 4. Analysis of Data Recognition rate, Classification rate with Feature selection

Table 2. Comparison of different approaches with three attributes

Approach Used	Automatic Data Recognition Rate	Auto Classification Rate	Feature Selection Rate
Extraction and Analysis of EHR Data	75	78	80
Smart Farming	85	80	79
Smart Cities	75	70	78
Manufacturing System	85	82	81
Industry 4.0	75	73	76
Spatial Data	85	80	70
Multimodal Architecture	75	82	81
Proposed EBDA Approach	90	88	92

In figure 5 the proposed Big data architecture model is compared with six approaches using five attributes and it's proved that EBDA model reliable in all used parameters data processing rate, execution rate to process the data based on feature selection model with multi view multi label approach. Data type recognition rate is much better with other model, classification of data based on proximity and principle component analysis is a key solution in the given model. Feature selection reduces data dimensionality to make it as more simple and accurate to process much faster since we focus on reduced feature set rather than massive data set. Overall analysis of the work is tabulated in below table to give facts and figures according to the experimental setup carried out in the process.

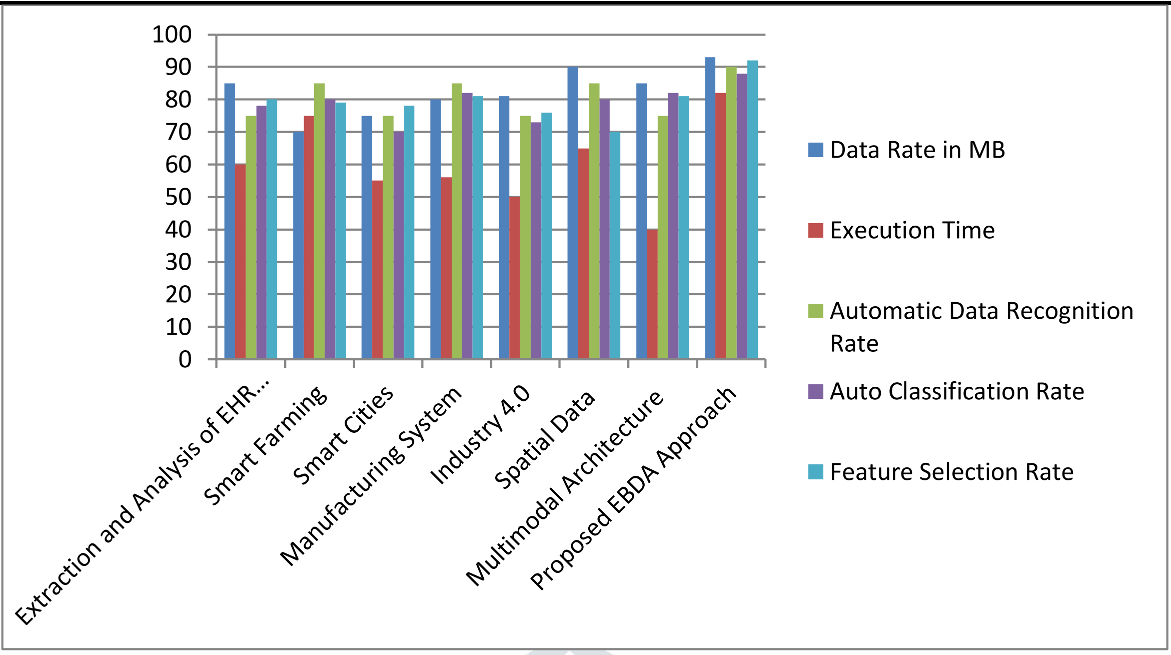

Fig 5. Comparison of different approaches with proposed EBDA approach.

Table 3. Comparison of different approaches with three attributes

Approach Used	Data Rate in MB	Execution Time	Automatic Data Recognition Rate	Auto Classification Rate	Feature Selection Rate
Extraction and Analysis of EHR Data	85	60	75	78	80
Smart Farming	70	75	85	80	79
Smart Cities	75	55	75	70	78
Manufacturing System	80	56	85	82	81
Industry 4.0	81	50	75	73	76
Spatial Data	90	65	85	80	70
Multimodal Architecture	85	40	75	82	81
Proposed EBDA Approach	93	82	90	88	92

VII. CONCLUSION

In Big data analytics information extraction and classification plays vital role to process the data in batch mode and in stream mode to support this mode of analysis and processing must require architecture. Architecture is backbone of any application to work successfully it handles data ingestion which collects large data and files generated from different sensors and actuators into storage medium which further represented in report using modern tools.

In this work Enhanced Big data architecture is compared with existing system models with data accessing rate with respect time used in seconds there is comparatively improvement in the accessibility shown in the graph and table. There are many architectures have been introduced from industry and academics, universal acceptance of these approaches for other types data analysis is not feasible.

Architecture solutions are not standard Big data processing architectures hence In this work we proposed smart Big data architecture to extract information and analyze the performance of information extraction using layered approach model to address data collection, processing, storing and data reporting and visualization of large data both in batch mode and real stream mode. Proposed work shows significant improvement with other architectures.

REFERENCES

[1] Andreou, Andreas G. et al. “Bio-inspired system architecture for energy efficient, BIGDATA computing with application to wide area motion imagery.” 2016 IEEE 7th Latin American Symposium on Circuits & Systems (LASCAS) (2016): 1-6.

[2] Amini, Sasan et al. “Big data analytics architecture for real-time traffic control.” 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) (2017): 710-715.

[3] Al-Thanoon, N. A., Algamal, Z. Y., &Qasim, O. S. (2021). Feature selection based on a crow search algorithm for big data classification. Chemometrics and Intelligent Laboratory Systems, 212, 104288.7.4.37 (2018): 168.

[4] Abdulhamit subasi, Esrra Molah, Fatin Almkallawi, “Intelligent website detection using random forest classifier”,ICCIS, 2019.

[5] Bleeker, Maurits. “Multi-modal Learning Algorithms and Network Architectures for Information Extraction and Retrieval.” Proceedings of the 30th ACM International Conference on Multimedia (2022): n. pag.

[6] Bil Yuchen Lin, Ying Sheng, “FreeDom A Transferable Neural architecture for structured information extraction on web documents”, Pages 1092-1102, 2020..

[7] Chauhan, Vikas Singh et al. “Real-time BigData and Predictive Analytical Architecture for healthcare application.” Sādhanā 44 (2019): n. pag.

[8] Chandra, Vajja Vignesh et al. “An Efficient Framework for Load Balancing using MapReduce Algorithm for Bigdata.” 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (2022): 791-794.

[9] Cravero, Ania, et al. "Agricultural Big Data Architectures in the Context of Climate Change: A Systematic Literature Review." Sustainability 14.13 (2022): 7855.

[10] Chen, Chao, et al. "Analysis of the Development Trend and Scheme of Agricultural Electrification Intelligence Based on Big Data Mining and OLAP Tool Analysis Algorithm." International Transactions on Electrical Energy Systems 2022 (2022).

[11] Fugini, Maria Grazia et al. “A Big Data Analytics Architecture for Smart Cities and Smart Companies.” Big Data Res. 24 (2021): 100192.

[12] Fahmideh, Mahdi and GhassanBeydoun. “Big data analytics architecture design - An application in manufacturing systems.” Compute. Ind. Eng. 128 (2019): 948-963.

[13] Gohar, Moneeb et al. “A Big Data Analytics Architecture for the Internet of Small Things.” IEEE Communications Magazine 56 (2018): 128-133.

[14] Gattoju, Saritha and V Nagalakshmi. “AN EFFICIENT APPROACH FOR BIGDATA SECURITY BASED ON HADOOP SYSTEM USING CRYPTOGRAPHIC TECHNIQUES.” Indian Journal of Computer Science and Engineering (2021

[15] Gill, S. S., &Buyya, R. (2019). Bio-inspired algorithms for big data analytics: a survey, taxonomy, and open challenges. In Big data analytics for intelligent healthcare management (pp. 1-17). Academic Press.

[16] Hiranandani, P., Pilli, E. S., Chand, N., Ramakrishna, C., & Gupta, M. (2018, January). Big Data Analytics Using Multi-Classifier Approach with Rhadoop. In 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 478-484). IEEE.

[17] Ji, B., Lu, X., Sun, G., Zhang, W., Li, J., & Xiao, Y. (2020). Bio-inspired feature selection: An improved binary particle swarm optimization approach. IEEE Access, 8, 85989-86002.

[18] Jeyasingh, S., &Veluchamy, M. (2017). Modified bat algorithm for feature selection with the wisconsin diagnosis breast cancer (WDBC) dataset. Asian Pacific journal of cancer prevention: APJCP, 18(5), 1257.

[19] Liu, Dan. “Big Data Analytics Architecture for Internet-of-Vehicles Based on the Spark.” 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS) (2018): 13-16.

[20] Maddumala, Venkata Rao and R. Arunkumar. “A Weight Based Feature Extraction Model on Multifaceted Multimedia Bigdata Using Convolutional Neural Network.” Ingénierie des Systems d Inf. 25 (2020): 729-735.

[21] Mohan, M. M., Augustin, S. K., &Roshni, V. K. (2015, December). A BigData approach for classification and prediction of student result using MapReduce. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 145-150). IEEE.

[22] Shah, Gita and K. Chandrashekar Shet. “DESIGN AN EFFICIENT BIGDATA ANALYTIC ARCHITECTURE FOR RETRIEVAL OF DATA BASED ON W EB SERVER IN CLOUD ENVIRONMENT.” International Conference on Cloud Computing (2019).

[23] Shah, Purnima. “Developing Big Data Analytics Architecture for Spatial Data.” PhD@VLDB (2019).

[24] Santos, Maribel Yasmina et al. “A Big Data Analytics Architecture for Industry 4.0.” WorldCIST (2017).

[25] Tidke, Bharat et al. “Real-Time Bigdata Analytics: A Stream Data Mining Approach.” (2018).

[26] Tekinerdogan, Bedir, and Burak Uzun. "Design of variable big data architectures for E-Government Domain." Software Engineering for Variability Intensive Systems. Auerbach Publications, 2019. 251-274.