iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** II **Month of publication:** February 2018

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 6 Issue II, February 2018- Available at www.ijraset.com

Big Data and Predictive Analytics to ERP Systems

Dr. S. Hanumanth Sastry[1] Prof. M. S. Prasad Babu[2]
[1]*Center for ERP, Visakhapatnam Steel Plant, India*
[2]*Dept. of CS & SE, Andhra University, Visakhapatnam India,*

Abstract: Processing large datasets for extracting information and knowledge has always been a challenge. Big Data Analytics has become a valuable technique for organizations to achieve enhanced business value and deeper insights. It facilitates real-time information delivery and enables organizations to run predictive analytics on large volumes of variety data to predict future behavior. Predictive analytics can leverage structured data and improve business processes for reducing costs, increasing sales and bring in a culture of informed decision making. Big Data Analytics can be applied to ERP systems for integrating large amounts of multi-structured data to extract and interpret useful information. Apache Hadoop is the de-facto framework for Big Data processing and R is the most popular data mining language. R can handle data analysis operations such as data loading, exploration, analysis and visualization where as Hadoop can provide a framework for Big Data storage and processing. Hadoop with R gives access to map-reduce programming model to process big datasets and generates actionable intelligence. In this paper, critical issues related to acquisition, storage and processing of Big Data in SAP ERP Data warehouse are identified and some solutions are discussed. An alternate framework for Big Data Analytics in SAP ERP system is also presented using open source R and Hadoop tools.

Index Terms: Predictive Analytics; Big Data; SAP HANA; Decision Management; ERP; R; Hadoop; KPI

I. INTRODUCTION

The main business functions of an Enterprise are production, materials, sales, finance, supply chain logistics, training and human resources. They share a common database to analyze data from multiple perspectives. It includes data about customers, suppliers and products etc. Presently Enterprise Resource Planning (ERP) systems are used to store, process data related various departments of the organizations. Several data mining tools are used to store, process and analyze enterprise structured data to identify potential risks and opportunities. At present, different departments of typical organizations are generating large volumes of variety of high speed structured semi-structured and unstructured data. Existing techniques used in ERP systems are not capable of handling this multi structured data.

In recent days a new technology, namely Big Data Analytics, is appearing in literature, to address the problems arising in handling the large volumes of multi-structured data. Therefore, Big Data processing in ERP systems is an exciting area of research. Here multi-structured large data sets are to be stored or analyzed for identifying hidden patterns. The sources of Big Data in enterprises/industries are operator log files, social media data, machine and sensor data, scientific experimentation data, quality control records and observational data on customers etc. Real time information processing and delivery is a crucial requirement for enterprises/ industries today. Big Data analytics may be used to provide quicker and deeper insights by integrating data from a variety of data sources in near real time. Big Data storage and processing in ERP systems is mainly based on three emerging technical disciplines namely big transaction data, big interaction data and processing of Big Data [5].

Availability of new databases for storing unstructured data and efficient ways to process massive volumes of data have led to the emergence of new technique namely NoSQL (Not only SQL) databases and Hadoop MapReduce framework [35].At the organization level, Big Data Analytics can find correlations in enterprise data on customer feedbacks on company's products, quality complaints etc and can help in improving the customer experience [40].

Apache Hadoop is the de-facto framework for processing Big Data. It includes two major components namely HDFS (Hadoop Distributed File System) and MapReduce programming model. HDFS stores large amount of data and provides high availability of data to user applications running at client machines. MapReduce is a software framework for analyzing and transforming very large data sets into desired output [12]. It splits the input dataset into independent parts for parallel processing by the map tasks. MapReduce program sorts the output of maps and feeds them as input to the reduce tasks.

In this paper the authors have proposed an alternate framework for processing Big Data in SAP ERP Systems. Here R and Hadooptools are integrated for knowledge discovery and identifying hidden patterns from large volumes of data. Some key issues for Big Data processing in ERP systems and appropriate improvised solutions are discussed.

II. RELATED WORK AND BIG DATA PRIMITIVES

Big Data analysis is now driving nearly every aspect of the society. It includes mobile services, retail, manufacturing, financial services, real estate and medical sciences. Some of the recent works closely related to Big Data Analytics are discussed here under.

Ajay Chandramouly et al. [8] discussed about issues such as problems faced during data acquisition, data retention and maintaining the relevant metadata in ERP systems.

Gregory Piatetsky et al. [9] defined Big Data as a new generation of technologies and architectures to economically extract value from very large volumes of data.

Yanchang Zhao et al. [11] identified new aspects in tackling Big Data such as taxonomies, onto logies, schemas, and workflow perspectives.

Paul C et al. [15] identified the relevance of Big Data for everybody in the present context and the way it has changed how business firms operate.

Usama F et al. [23] reviewed the use of various data mining algorithms for knowledge discovery process and presented a unifying framework for knowledge discovery from large databases.

Chen Jiang et al. [24] proposed a new BI architecture with predictive capabilities on IBMCognos Business Intelligence platform mostly suited for structured data.

Madden et al. [25] opined that Big Data means data that is too big, too fast or too hard for the existing tools to process.

Jens Dittrich et al. [26] analyzed the difference between MapReduce and parallel DBMS methods using data optimization techniques such as data layouts and indexes.

Harriet Fryman et al. [27] defined Big Data as the information that goes beyond the organizational ability to leverage from their BI and transactional systems.

Yanchang Zhao et al. [28] conducted a study on visual data exploration tools and concluded that R tools may be used for data visualization.

Jerome L et al. [29] observed that Hadoop and R can complement each other and proposed a new R package RHadoop, which is a collection of three R packages namely RMR, RHDFS and RHBASE.

The characteristics of Big Data, five V's, are shown in fig 1 below.

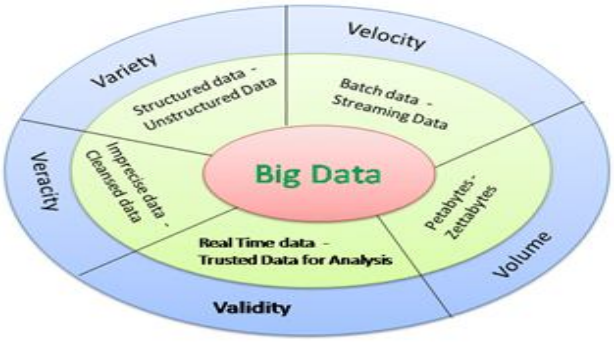


Fig. 1. Characteristics of Big Data

These characteristics are explained with respect to the business functions of a manufacturing industry especially steel plant.

1) *Volume:* This property corresponds to large volumes of data generated every second. This is usually in the range of hundreds of Terabytes, Petabytes or Exabytes of data.
2) *Velocity:* This property refers to the speed at which new data is generated and the way data moves around. Data generated by sources such as machines, sensors, networks and human interaction with social media sites, mobile devices etc flows in a massive, continuous fashion and arrive at a rapid pace.
3) *Variety:* This property refers to the diversity of data from structured, semi-structured and unstructured data sources namely emails, photos, videos, monitoring devices, PDF files and audio files etc.
4) *Veracity:* This property deals with uncertain or imprecise nature of data such as biases, noise and abnormality that are present in data. Due to the high velocity of Big Data, too much time cannot be spent in cleansing it thus allowing some amount of uncertainty in the data.
5) *Validity:* This property refers to timeliness of data i.e. how long the data is valid and for how much duration it should be stored.

6) Value: This property measures the usefulness of data for decision making. Two aspects associated with data value are Data Science and Analytic Science. Data science is exploratory in nature and helps to know the data. Analytic Science encompasses the predictive power of Big Data for performing analytics. The trust we have in the Big Data depends on the value of the data and hence impacts the decisions made based on that data.

III. AN EXTENDED FRAMEWORK FOR SAP ERP BUSINESS WARE HOUSE WITH BIG DATA

Presently SAP ERP systems are processing only structured Data. These systems have several deficiencies and design limitations. Some of them are: inability to handle variety of data, handling data replication to multiple sources, higher data latency and high voluminous data and memory manipulation etc. To address these problems an extended framework is proposed to augment predictive analytic capabilities for the existing ERP Business warehouse by creating a computing environment capable of processing Big Data. Hadoop Distributed File system(HDFS) [50]is used to store Big Data and R functions are used for providing predictive analytics capabilities. Details of the proposed framework are shown in fig 2.

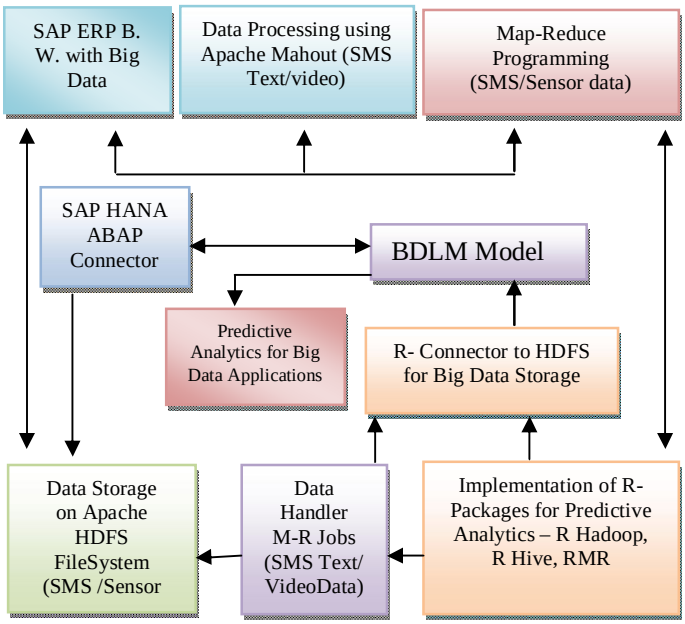


Fig. 2. Extended Framework for ERP Business Warehouse with Big Data

A. *Components of the extended framework*

The components of the extended framework are briefly described below.

1) SAP BW with Big Data: SAP Business Warehouse (BW) is an information repository that stores the cross functional data on a long-term basis from all operational sources of enterprise such as production, sales, materials, finance, costing, human resources, training and development etc. It provides a set of tools to query, analyze and visualize enterprise information. It also facilitates Online Analytical Processing (OLAP) for multiple business perspectives.

2) Data Processing with Apache Mahout: This component Processes Big Data using Hadoop MapReduce programming model. It provides resources for clustering, classification, regression and statistical modeling by using a library of statistical software. Text Analytics may be applied for existing text documents using JSON (JavaScript Object Notation) scripts and assigns unlabeled documents to appropriate category in Hadoop HDFS.

3) MapReduce Programming Model: In this component two separate tasks may be performed on Big Datasets using Hadoop programs. MapReduce programs process large amounts of data and splits the input dataset into independent chunks which are processed by the map tasks in a completely parallel manner. Validation, index creation, or consistency check operations are performed when they will not impair the rate of data insertion into HDFS.

4) BDLM Model: Big Data Life Cycle Management System is a new upcoming technology introduced by IBM in 2012 [30]. It is designed to handle both structured and unstructured data of the enterprise for information access and reporting.IBM provides some commercial packages to handle Big Data at the enterprise level. Some of the known packages are SAP HANA and SAS.

5) *R-Connector to HDFS for Big Data Storage:* R-Connector may be invoked from R computing environment to process Big Data using MapReduce code. R-Connectors are highly extensible, easily readable, and reusable that require less code than Java language. It has objects oriented features and strong graphical capabilities.
6) *Implementation of R Packages:* This component enables execution of R packages such as RHadoop, RHive and RMR, in parallel mode on Big Data. This component also enables faster data retrieval from Hadoop HDFS especially on large clusters.
7) *Data Handler for MapReduce Jobs:* This component creates executable objects with serialization by invoking R predictive algorithms and saves the object as a file in HDFS. This component also makes the file instructions accessible to Mappers and *Reducers in MapReduce programming model*
8) *Data Storage in Apache Hadoop HDFS:* This component stores Big Data on Hadoop HDFS. This data is further distributed on computing nodes namely Name Node and Data Nodes of Hadoop cluster.

B.Critical Issues and Solutions for Big Data Processing in ERP Business Warehouse (BW):

The critical issues identified for implementation of Big Data Analytics in SAP BW systems and the corresponding suggested solutions (Sol) are presented here.

1) *Processing voluminous data:* Big Data Analytic applications require storing large volumes of data and meta-data related to data access, utilization and data governance in real-time. Large volumes of data storage involve challenges such as data scalability, accessibility and fault tolerance etc [24] Sol: Easily scalable Hadoop HDFS may be used to store large volumes of data.
2) *Real time data transportation:* In Big Data analytic applications, data is processed "*inplace*" and only the "*resulting*" information is transmitted to save system /network resources. Therefore, this can be treated as: "*bring the code to data*" unlike the usual method of "*bring the data to the code*". Typical RDBMS systems use a centralized database system and facilitate the later approach. Presently ERP BW systems do not facilitate this kind of high speed data transfer.
3) Sol: Distributed file systems such as Hadoop HDFS are highly suited for the former approach as it streams MapReduce code at high speed to user applications.
4) *Handling data variety for information delivery:* Processing of multi-structured data poses new challenges for SAP BW systems. New ERP BW systems are required to store and process high volumes of complex unstructured data such as web traffic, social media and sensor data in real time. Traditional ERPBW systems are designed to handle only structured data but not the variety of Big Data.
5) Sol: Big Data databases such as Mongo DB may be used to process variety of data and also reduce data latency for Big Data applications. The increase in speed is obtained by putting data into main memory instead of storing on the disk
6) *Database Schema for Big Data Analytics:* Big Data is characterized by either flat-schema or schema-less. Real time analytics is closely related to schema-fewer databases which require stream processing techniques. But traditional RDBMS systems cannot handle them. For Big Data analytics, high velocity data is required to be processed in real time. Traditional ERP BW system cannot process high velocity data. to determine what data can be shared with third parties. Nowadays, user information can be extracted from internet logs or identity management tools and misused for surveillance and marketing purposes. Organizations may use Big Data analytics for efficient information tracking and address major security concerns such as cybercrimes, risk Management and data compliance issues [15]. This necessitates the inclusion of personal information and regulatory mechanisms in Big Data processing systems.
7) *Data Security and identity control:* Data security means protecting data from the actions of unauthorized users. Stringent security controls are applied to SQL based SAP ERP systems such as Enterprise Central Component (ECC). Some important SQL security controls are secure configuration management, multifactor authentication, data classification, data encryption, consolidated auditing or reporting and database firewalls etc. Big data analytic applications need to include secure encryption techniques to protect the enterprise data. Primitive security tools such as HBase and Hive may also be used for big data security.The characteristics of business warehouse Big data are compared with traditional BW relational data and the results are presented in table 1.

TABLEI. comparisons of traditional data vs. sensor data

Attribute	Traditional Data	Big Data	Experimental Data
Data Volume	Gigabytes to Terabytes	Petabytes to Exabytes	Petabytes

Data Storage	Centralized	Distributed	Distributed
Data Variety	Structured	Semi-structured and Unstructured	Unstructured
Database Schema	Stable data model	Flat schema to No Schema	No schema
Data Relationship	Known complex relationships	Few complex relationships	Complex relationships
Data Privacy	High	Medium	Medium
Data Security	HighlySecure	Medium Security	Medium Security

The details of implementation of the extended framework for ERP BW Big Data systems are given in Section IV.

IV. IMPLEMENTATION METHODOLOGY OF EXTENDED FRAMEWORK

A. Process Flow

The Process flow of the implementation of the extended framework to SAP BW system is explained specially by implementing Big Data Tools R and HANA to SMS Data of a steel industry in the following eight steps.

1) Load the SMS text data into SAP BW system and develop Info Objects.
2) Classify the SMS text/audio/video data using Apache Mahout and store the data in Hadoop HDFS.
3) Create serialized objects using the MapReduce programming model component and make it accessible to Mappers and Reducers.
4) Invoke R-connector to process Big Data using MapReduce code.
5) Implement R packages for Hadoop HDFS access
6) Run executable objects created in step by invoking R predictive algorithms.
7) Visualize results of predictive algorithms by implementing BDLM methodology.
8) Invoke SAP HANA ABAP connector for Hadoop HDFS data access.

B. Methodology

The principal techniques applied in the extended framework i.e. Big Data Processing in SAP BW system, BDLM Life cycle mode, HadoopMap Reduce programming, HadoopR connectors and SAP HANA connectors are explained below.

1) *Big Data Processing in SAP Business Warehouse*: The process is explained in the following seven steps.

a) *Creation of info Objects in ERP Business Warehouse:* Big Data processing in ERP BW systems enables the organizations to perform decision making based on predictive analytics. As it is not possible to perform all analytics on a *single platform*, it is necessary to extend this capability. Info Objects such as Info Cubes and Data Store Objects (DSO) are developed in ERP business warehouse by converting logical Entity Relationship Model (ERM) into physical objects. SAP BW system management tasks such as defining new interfaces for DB (Database) connect, UD (Universal Data) connect, BAPI (Business Application Programming Interface) and XML connect are performed by linking relevant metadata. Metadata Repository in SAP BW system stores directory information regarding several object types and maintains Master data of all info Objects[46]. Data is extracted from SAP source systems into BW system and further into data targets such as Info Cubes or Data Store Objects (DSO).

For a given Hardware Rate and Sleek Time, the block size of the data in SAP BW system is computed as given below.
Hardware Rate = 100 MB/s
Sleek Time (S_t) = 10 ms
Transfer Rate (T_r) = 0.01 * S_t
Then, the block size would be 100 MB
Typically the block size of Big Data would be in the range of 64MB to 128MB.

b) *Extract, Transform and Load(ETL) operations on Big Data*: Extract-Transform-Load (ETL) functions used in SAP Business Warehouse are extended to work with Hadoop tools. Since Hadoop is schema-less, it can be used to store and process multi-structured data [42]. ETL operations in ERP data warehouses are: cleansing, pruning, conforming, matching and joining. The streaming unstructured data is pre-processed and standardized with regard to semantics, format and lexicon for analysis. After

performing ETL operations on Big Data, data auditing may be performed for verification and transparency. Here meta-data related to Big Data streams such as origin of data, owner, and time of creation may be linked to Master Data Management (MDM) module of BW systems. MDM contains details of texts, attributes and hierarchies for each Info Object created in ERP BW system. Data in SAP BW system may be reorganized in different data marts to reveal valuable knowledge about various industry relevant topics, products, designs, relationships, opinions and customer sentiments.

c) *Developing Context-aware models in BW system:* Context wareness has gained importance in enterprise systems for real time information delivery due to the increasing mobility of users and devices. For example, in manufacturing industry existing data in ERP system has no relation to the context about the user´s history, location, tasks, habits etc. Login attributes of the special users are also used for generating context aware data from business applications.

d) *Accessing Data from NoSQL Databases*: Unstructured data stored in Not only SQL (NoSQL) databases cannot be directly processed in ERP BW systems since it is not in the required relational format. NoSQL database aggregates the data to be stored into several documents using the JavaScript Object Notation (JSON) format. Each JSON document can be viewed as an object to for processing ERP application. JSON document takes all the data stored in a row that spans 20 or more tables of a relational database and aggregates it into a single document or object [15]. Special data extraction structures are required to load this unstructured data into BW system. File based and Non-RDBMS systems such as Hadoop HDFS are ideal for processing this document oriented data. Not only SQL (NoSQL) database relaxes the constraints of a RDBMS and extends the capabilities of Hadoop clusters by quick object retrieval [8]. Significant time and efforts are spent in moving large amounts of data between the Hadoop clusters. Some examples for NoSQL databases are HBase, Cassandra, Mongo DB, Accumulo, Mark Logic, Aerospike, Riak, Couch DB and Dynamo DBetc[40] [42]

e) *Methodology for Multi-Dimensional Analysis of Big dat*: In ERP BI system, the data is normally stored in databases such as traditional RDBMS, Apache Hadoop and Massively Parallel Processing (MPP) systems. Data in SAP BW system is organized into many Info Cubes, where it is stored as dimensions and facts for multi-dimensional analysis. The difference in database characteristics of these three systems is shown in table 2.

TABLEII. Comparision Of Hadoop, Rdbms, Mpp Systems

Characteristic	Apache Hadoop	RDBMS Systems	MPP Systems
Architecture	Scale-out	Scale-up	Scale-out
Data Storage	Key/Value Pair	Relational Tables	Relational Tables
Data Retrieval	Functional programming	Declarative Queries	Declarative Queries
Data Processing	Offline Batch Processing	Online Transactions	Online Transactions

SAP ERP systems uses RDBMS databases for data processing but Big Data and MapReduce programming model has new requirements for different database parameters as shown in table 3 below.

TABLEIII.COMPARISION OF RDBMS AND MAPREDUCE

DB Parameter	RDBMS	MapReduce
Data Size	Gigabytes to Terabytes	Petabytes to Exabytes
Data Access	Interactive and Batch	Batch
Data Updates	Read/Write many times	Write once, Read many times
Data Structure	Static Schema	Dynamic Schema
Data Integrity	High – Characterized by ACID properties	Low
Data Scaling	Non-linear	Linear
DBA Ratio	1:40	1:3000

C. *Data storage and classification using Hadoop HDFS*

Hadoop software architecture allows the separation of data storage from data processing. Hadoop framework facilitates distributed processing of large data sets across a cluster of computing nodes using MapReduce programming model. The authors propose the use of a staging area before loading unstructured/semi-structured data into ERP Business warehouse [30] to achieve data homogenization. Key components of the Hadoop framework namely Hadoop Distributed File System (HDFS) and MapReduce(MR) implementation framework are described here under.

1)Hadoop Distributed File System (HDFS) for SMS Data: HDFS is a distributed file system that provides high-performance access to data across Hadoop clusters. The number of Mappers generated in Hadoop HDFS usually depends on the size of the input dataset. In HDFS, Name Node manages the file system metadata and Data Nodes store the actual data. Clients contact Name Node for file metadata or meta-data on file modifications to perform actual file I/O directly on the Data Nodes. HDFS schedules one Mapper on each node corresponding to each split of the input data. The process model is explained in Fig 3. In this model, node 1 to node 3 will process video, text and image data respectively. Output of each node is combined through reducer function using a key, value pair.

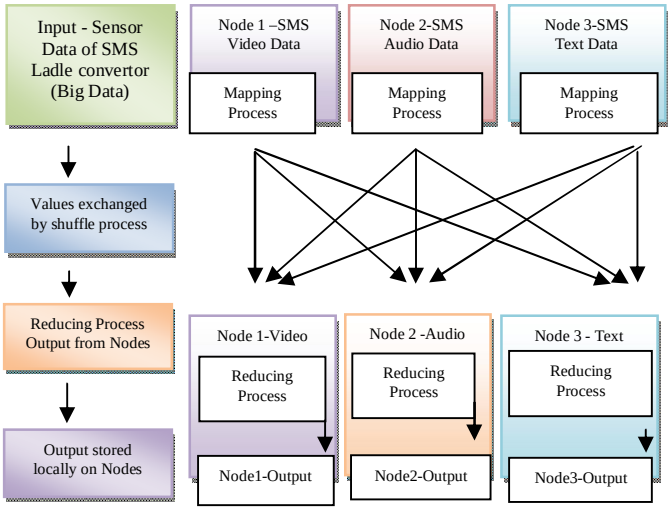


Fig. 3. Mappers and Reducers created in HDFS to process SMS data

D. Namenode and Datanode for processing SMS data:

Hadoop cluster consists of a number of commodity servers called Datanodes that provide automatic replication of data between the Datanodes. HDFS cluster consists of a single Namenode called master server that manages the file system namespace and regulates access to files by several clients. The Namenode executes file system operations such as opening, closing, renaming files and directories. Datanodes are responsible for serving read and write requests from the file system's clients. The details of Namenode and Datanodes applied to SMS converter data are shown in fig 4.

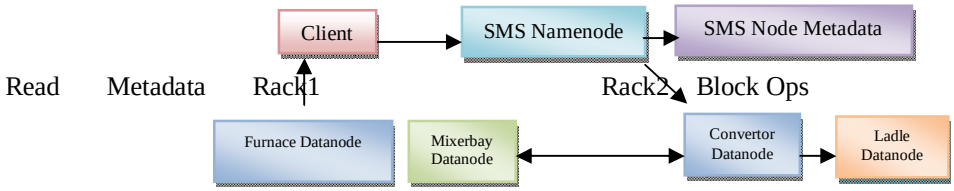


Fig. 4. Namenode and Datanodes in Hadoop HDFS for SMS Data

As shown in fig 4, a central SMS Namenode is connected to multiple Datanodes such as Ladle and furnace Datanodes. The SMS Name node initiates and manages the processing jobs and restarts them in case of any failure. This mechanism guarantees that a Hadoop job always completes its execution. Hadoop software is available for implementation mainly in three modes which are described below [17].

$$\max\left(\min(\text{block}_{\text{size}}), \frac{\text{data}}{\#\text{maps}}\right), \text{min\_split\_size})$$

1) *Reduces*: Reducer has a reduce method that transforms intermediate (key, value) aggregates into any number of output (key, value) pairs. It is calculated as [25] [26]

$$0.95 * num\_nodes * mapred.tasktracker.reduce.tasks.maximum$$

Implementation details of SMS application on Hadoop's stand-alone mode are given in section 5 of this paper.

2) *Hadoop Tools:* Some important Hadoop tools deployed on the extended framework are described below [27]:

a) *Hive:*Hive gives a data warehousing like framework to Hadoop HDFS and makes it possible to use it as a relational database. Although Hive is not fully SQL compliant, it provides a Java Database Connectivity (JDBC) interface which allows it to be *treated as an external database.*

b) *HBase*: HBaseserves as a real time key-value store and provides many API's to store and retrieve multi-megabyte documents using a key. HBaseenables Hadoop HDFS to become a fast, massively scalable data store for random read/write access to Big Data. HBase is used by web sites to store large images, videos or documents for data retrieval and display.

c) *Sqoop:* Sqoop transfers bulk data between Hadoop HDFS and structured data stores such as relational databases. Sqoop has database connectors for MySQL, Postgre SQL, Oracle and IBM DB2.

d) *Ambari:* Ambari deploys and operates a complete Hadoop stack using a Graphical User Interface (GUI). It also manages configuration changes, monitors services and creates alerts for all Name Nodes in Hadoop cluster from a central point.

e) Zookeeper: This tool provides centralized infrastructure and services to enable synchronization across Hadoop clusters. This tool also maintains common database objects needed for large cluster environments. Examples of database objects include configuration information, hierarchical naming space and so on. Applications can use services provided by Zookeeper to coordinate distributed processing across large clusters.

f) *Crowbar:* Crowbar is an open source tool used for Hadoop multi-node deployments. Crowbar uses a modular approach called Barclamp, where each component of the Hadoop stack is deployed as an independent unit.

g) Flume: This tool imports log & text files into Hadoop HDFS

h) Yet another Resource Negotiator (YARN):It is also termed as MapReduce 2.0 (MRv2) or YARN. YARN splits the major functionalities of the Job Tracker, resource management and job scheduling/monitoring into separate daemons. Using YARN tool, Big Data application scan get access to global Resource Manager (RM) and Application Master (AM) interfaces of Hadoop HDFS.

i) *Pig scripts:* This tool generates MapReduce programs written in Pig Latin, which is a procedural language used for data analytics.

E. R data connectors for Hadoop HDFS:

R is a programming language and a software suite used for data analysis, statistical computing and data visualization [21]. Open source R is memory-bound and is designed to have all of its data in memory where as programs in Hadoop (map/reduce) work independently and in parallel mode on individual datasets [35]. Hadoop is batch oriented where jobs are queued and then executed, which may take minutes or hours. Predictive algorithms available in R for Big Data processing are discussed here.

1) *Big Data processing using R Predictive Algorithms*

Different predictive algorithms and tools available for Big Data Analytics in R are briefly discussed here.

a) *Classification Algorithms*: These include decision trees, random forest trees, Bayesian models and SVM (support vector machine). Classifiers are used to improve predictive accuracy measured in terms of AUC (Area under Curve) and ROC (Receiving operating characteristic) curves.

b) *Clustering Algorithms*: They divide data into meaningful groups that share common characteristics. Three types of clustering algorithms are available in R suite for interfacing with Hadoop software framework.

c) *Hierarchical Clustering:* They construct clusters by recursively partitioning the instances in either top-down or bottom-up fashion. Some algorithms of this type are Classit, Cobweb and Discriminant Coordinate Analysis.

d) *Density Based:* They discover clusters with arbitrary shapes and regard clusters as dense regions of objects in the data space. These dense regions are separated by regions of low density that represent noise. Some density based algorithms are DB Scan, Optics and sIB.

e) *Partition Based:* These clusters are created for optimizing a predetermined criterion. Algorithms of this type areK-Means, Expectation-Maximization (EM), EWKM (Entropy Weighing K-Means) and Correlation clusters by Pearson etc.

f) *Regression Algorithms*: R can run Neural Network algorithms such as MLP (Multi-Layer Perceptron) and RBF (Radial Basis Function)on Big Data sets. Linear Regression models can be run on non-Gaussian distributions for glm(), gbm() and lm models

[41]

g) *Data Attribute Analysis in R*: R facilitates Principal Component Analysis (PCA), Pearson Covariance, Benford's Law, data plot for hierarchical clustering and clustering graphs for numerical or non-numerical attributes.

2) Creating MR Jobs in Hadoop HDFS from R Packages

Data process invokes an algorithm in the R library, and distributes subsets of the work to Data nodes in the Hadoop cluster. Once the data nodes finish their tasks, a single reducer process consolidates intermediate results and returns them to the master node. If the master node determines that the algorithm has completed its work, it returns the consolidated results to the script or R command [35] [39].

For K-Means clustering, R program performs a convergence test and checks the number of iterations. If the test is successful, R program returns consolidated results to the script or command, otherwise it repeats the process [44]. The results of each map instance are persisted to main memory and file system, allowing Hadoop to reschedule only failed Mapper instances in the event of incomplete execution.

3) Data Visualization Tools in R

Predictive algorithms available in data mining tools such as R-Rattle, Reducer, and JGR are utilized by ERP systems for visual representation of data. Data mining tools such as Tinn-R, Rpad, RPMG, Red-R and Rattle may be used to visualize the application results in R tool environment [35].

4) R Packages for Hadoop

Integration of Big Data Analytic applications with R and Hadoop is facilitated by the availability of various R packages such as RHIPE, RHadoop and RMR. These packages facilitate the execution of R predictive algorithms over large datasets in a highly scalable manner.

5) Methodology for R integration with Hadoop tools

Client programs can invoke Hadoop jobs from R environment and pass R objects from the client R object space to Mappers and Reducer functions. Testing of MapReduce jobs can be performed locally at client R engine without changing any code (by just switching a system flag). This mechanism makes it easy to debug code before deploying it on the full Hadoop cluster. Using high level R interface, parallel distributed algorithms may be run on HadoopMapReduce model to take advantage of the commodity benefits of the Hadoop cluster. Different R packages and the process of creating MapReduce jobs in Hadoop cluster are briefly explained here.

F.RHadoop, RHDFS, RMR, RHBASE Packages

RHadoop [29] is a collection of three R packages namely RMR, RHDFS and RHBASE. RMR package provides Hadoop MapReduce functionality to R, RHDFS provides HDFS file management to R and RHBASE provides HBase database management functions from within R. RMR contains MapReduce and basic data exchange related functions for HDFS storage. RHadoop runs inside Hadoop and provides direct access to data stored in Hadoop from R interface. RHadoop uses MapReduce code to distribute computational workload across the Namenodes of a Hadoop cluster. Integration of R programs with various Hadoop tools is shown in figure 5 below.

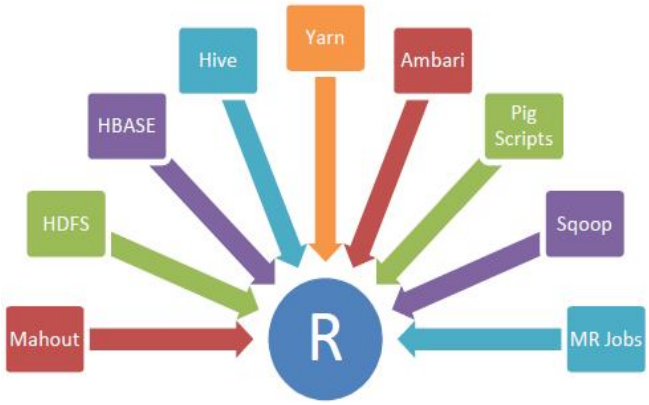


Fig. 5. R Integration with Hadoop tools

1) *RHive:* RHive supports SQL like query language, called Hibernate Query Language (HQL). HQL works with persistent objects and their properties. HQL queries are translated by Hibernate tool into conventional SQL queries which in turn performs

query/update actions on HDFS database. RHive executes R functions through Hive Query

2) *Plyrmr:* This package enables R clientprograms to perform common data manipulation operations on very large data sets stored on Hadoop cluster. Plyrmr package utilizes HadoopMapReducecode to perform its tasks.
3) *Rmongo:*This package is a MongoDB Database interface for R programs. Rmongo package provides many Java calls (API's) to the mongo-java-driver
4) *RExcel:* RExcel is an add-on for Microsoft (MS) Excel that gives access to R packages on Windows 32-bit versions. RExcel uses the statconnDCOM server and rcom package to access R from within Excel.
5) *Radoop:*Radoop provides an interface for editing and running ETL (extract-transform-load), Big Data analytics and machine learning processes on Hadoop clusters.
6) *Issues in R Integration with Hadoop systems:* R program execution is in-memory oriented and does not provide support for distributed computing. Hosting the R analytic software outside of Hadoop cluster creates many deployment problems. Customized R programs are required to implement a predictive model scoring process. Data replication and data movement from R environment can also cause data security problems [39].

G. *Big Data Life Cycle Management (BDLM) Model*

BDLM model consists of two layers namely Data Model and Data Processing that integrate and provide user level access to information for reporting and analysis purpose[35]. In the extended framework for SAP ERP BW systems, visualization of predictive insights and hidden patterns are facilitated by BDLM model. Details of BDLM model in the extended framework for ERP BW system are shown in fig 6.

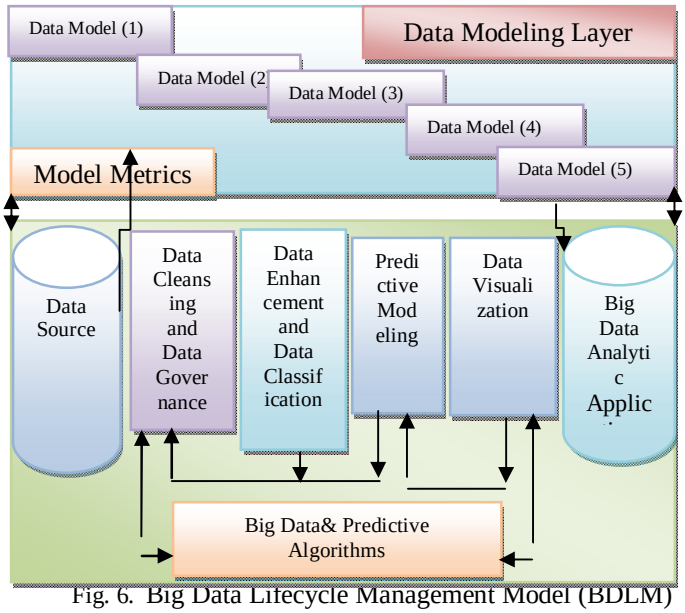


Fig. 6. Big Data Lifecycle Management Model (BDLM)

In figure 6, business case models 1 to 5 represent customer sentiment analysis, human capital management, social media, web-tracking and financial analytics [23]. Business case models help in understanding customers, products and competition for achieving better Return on Investment (ROI). BDLM implementations are available from Cloudera, Map R, Horton works, Apache Hadoop and Data Stax [13]. The components shown in BDLM model are briefly explained hereunder.

1) *Data Sources*: Quality structured and unstructured data from secure and trusted data sources is identified and classified according to different data types such as text, audio, video.
2) *Data cleansing and Data Governance*: Data from source systems is cleansed and loaded into ERP Business Warehouse (BW). To improve business process efficiency, consistent and complete datasets are used for Big Data applications.
3) *Data Enhancement and Data Classification*: Data is categorized according to data type, location and access levels. Data is further classified as semi-structured or poly-structured based on time of data creation, owners of data, time of access and the time of update etc.

4) *Predictive Modeling*: Correlation techniques such as Pearson covariance, Principal Component Analysis (PCA) are applied to find causal relationships between explanatory variables and dependent variables [43]. Predictive techniques are applied to uncover previously hidden patterns and identify classifications, associations, and segmentations in Big Data and facilitate effective decision making. In BDLM models, the focus is on specific business out comes such as customer credit worthiness, fraud risk and customer behavior etc.
5) *Data Visualization*: Big data visualization tools such as heat maps, data relationship trees, geospatial overlays and other interactive tools are used in BDLM model [30] [31].
6) *Big Data Analytic Applications*: Variety of Big Data applications such as online recommendation engines, fraud detection, risk modeling, research and development can be developed using the BDLM model.
7) *Algorithms for Big Data and Predictive Analytics*: Supervised learning methods of Classification and Regression are used for Big Data analytics as shown in figure 7 below. During the training phase in predictive modeling, algorithms try to determine the relationships that exist in data to match the *known* outcome. Using the rules established in the learning phase, they predict outcome for a new unknown set of data [35] [44].

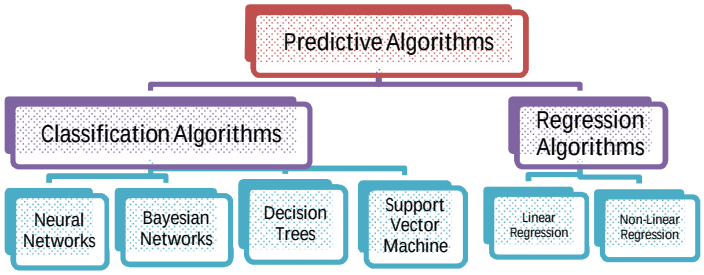


Fig. 7. Supervised algorithms used for Predictive Analytics

H. BDLM Implementation Methodology

In BDLM model, Big Data is processed on Hadoop platform by dividing large datasets into subtasks. Data is replicated on multiple nodes by creating redundant copies of data at different nodes so that in case of failure, a copy of data is always available [42]. BDL Mmodels may be deployed on cloud to aggregate multiple disparate workloads with varying performance goals. In this model, the problem of failure is handled by HDFS and the problem of combining data is handled by MapReduce programming model. MapReduce minimizes the problem of disk reads and writes data by using a key, value pair.

I. SAP HANA ABAP Connector

SAP HANA has pre-configured structured data, integrated from different SAPERP transactional systems. SAP HANA integration with Hadoop tools is performed through SAP ABAP connector as described below.

SAP HANA connector uses pig scripts to transfer jobs containing HDFS source and text transformation algorithms into HDFS file system. MapReduce code to process Big Data is generated by the SAP HANA Data Service. Key functions of SAP Data Service for Hadoop HDFS integration are described below [29] [31].

1) *Data Federation*: The structures of files in Hadoop HDFS are mapped to external database tables in SAP Sybase IQ.
2) *Query Federation*: Queries are executed separately on Hadoop using Hive and the results are combined with other queries executed in SAP Sybase IQ database.
3) *ETL Operation*: Data from a variety of sources such as SAP HANA, SAP IQ and Hadoop HBase is homogenized to give a complete view of information.

V. IMPLEMENTION OF BIG DATA APPLICATIONS ON THE EXTENDED FRAMEWORK

The predictive power of the extended framework for Big Data Analytics in ERP BW systems is explained with the implementation of the following production process applications related to manufacturing industry.

A. Problem Statement

To predict the working hours of equipment in SMS (Steel Melting Shop) with unstructured data gathered from various sensors and PLC's (Programmable Logic Controllers)

Process data from Steel Melting Shop (SMS) which converts hot metal to steel is used. Data generated in this process possesses the following Big Data characteristics.

1) Volume: For SMS application, the number of records received per sampling hour in production process is above17800 records. This corresponds to 15 crore and 38 lakh records per annum. This equals to a data size of 146 Terabytes per annum or nearly 1.5 Petabytes for a 10 year period.

2) Value: In the present SMS application, data contains more than thirty two attributes gathered from sensors about equipment condition.

3) Veracity: SMS data is imprecise in nature with bias and noise.

4) Velocity: In the present SMS application, data velocity is high and data arrives rapidly from nearly 120 sensors with approximately 50 records per second.

5) Validity: In the present SMS application, validity of data varies between different maintenance cycles and usually ranges from 3 to 5 years.

B. Experimental setup

1) *Storage*: The database tables are created in HDFS with required fields and accessed using RMR package in R. In this dataset, monthly working hours for equipment is predicted using K-Means clustering algorithm.

2) *Processing*: Hadoop virtual machine with Mahout predictive algorithms are used for data processing

3) R program code is executed from client workstation.

4) Tools used:

a) HDFS (Hadoop Distributed File System) is a distributed, scalable, java based file system used for data storage in distributed files/nodes. It provides a storage layer for large volumes of unstructured data.

b) MapReduce (MR) program splits the input dataset into independent parts for parallel processing by the map and reduce tasks.

c) R provides a set of predictive algorithms called packages that can create data processing jobs in HDFS.

Big Data storage and processing methodologies in the extended framework, explained in sections IIIand IV are implemented in the following ten steps.

Step 1. Create data storage in Hadoop HDFS: Database tables namely sms_conv_lp and sms_conv_work are created in HDFS with 32 fields related to temperature, gas pressure and gas flow etc.

Step 2. Create MapReduce jobs in Hadoop: Process sensor data received from SMS convertor shop using MapReduce program code and split it into independent jobs in a completely parallel manner.

Step 3. Monitor MapReduce jobs: Use HadoopJob Tracker and TaskTracker functions to monitor MapReduce jobs. Pass the location of the input data file and the processing instructions for Mappers and Reducers using JobTracker

Step 4. Create Job and Task Trackers: Create a large set of TaskTrackers for each Hadoop SMS DataNode to monitor MapReduce jobs. Also pass the file credentials and its HDFS location to SMS Namenode.

Step 5. Combine jobs from Hadoop clusters: Invoke Mapper program from HadoopTaskTracker function and pass the following parameters namely input data split location and instruction file location and await the completion of Mapper or Reducer.

Step 6. Pass the Result set: Combine data from different Mappers into HDFS file segments through a Java process and pass the result objects back to Hadoop master process.

Step 7. Check for data convergence: Check the master process for the consolidated result object for data convergence and if the test succeeds push the final results to R programming interface. If the test fails, repeat the master process execution with new instructions. End the process when the data convergence test succeeds or the number of iterations reaches a pre-defined limit of 10 iterations.

Step 8. Monitor results from R: Invoke parallel algorithm from R packages for data processing and distribute subsets of the work to data nodes in the Hadoop cluster.

Step 9. Consolidate the results: Consolidate the intermediate results and returns the control file to the master node, after the DataNodes finish their tasks.

Step 10. Return the Results: Return the consolidated results to the script or R command, if the master node determines the successful execution of the algorithm.

The input and Output types of a MapReduce job is represented as:

(input) <k1, v1> -> map -><k2, v2> -> combine -><k2, v2> -> reduce -><k3, v3> (output)

The executable code for Map-Reduce functions using RMR package in R is given below.

```
Sapply (data, function)
Mapreduce(data, map = function)
Call: mapreduce(input, output, map, reduce)
Map = function(k,v) if (hash (k)) %% 10 = = 0) keyval(k,v)
Reduce = function(k, vv) keyval(k,length(vv))
Condition = function(x) x > 10
Out = mapreduce(input=input, map=function(k,v) if (condition(v)) keyval(k,v))
x= from.dfs(hdfs.object)
hdfs.object = to.dfs(x)
Insert overwrite table sms_conv_work
Select sub_eqpt, mechanism, make, rating, equipmt_make from sms_conv_lp group by brigade
Mapreduce (input = mapreduce(input= “mechanism”,
Map = function(k,v) keyval(v [‘sub_eqpt’], v[‘rating’],
  Reduce = function (work_hours, manuf_type)
Lapply(unique (ac_dc), function(g) keyval (null,g)),
Output = “equipmt_history”,
Map = function(x, equipmt_make)  keyval (work_hours,1)
Reduce = function(work_hours, counts) keyval (k, sum (unlist(counts)))
```

C. Program for implementing K-Means Clustering Algorithm

The executable code for the implementation of K-Means clustering algorithm is given below.

```
K-Means = function(points, ncenters, iterations = 10,
Distfun = function(a,b) norm(as.matrix9a-b), type= ‘F’)) {
Newcenters = kmeans.iter(points, distfun, ncenters=ncenters)
For (i in 1: iterations ) {
Newcenters = kmeans.iter(points, distfun, centers=newcenters)}
Newcenters}
Kmeans.iter= function(points,distfun, ncenters = dim(centers) [1] centers= null) {
From.dfs(Mapreduce ( input = points,
Map = if(is.null(centers)){
Function(k,v) keyval(equppmnt_make(1:ncenters,1), v) }
Else{
Function(k,v){
Distance = apply(centers, 1, function(c) distfun(c,v))
Keyval(centers(which.min(distances), 1, 1)}},
Reduce = function (k,vv) keyval(null, apply(do.call(rbind,vv), 2,mean))), to.data.frame= T)}
```

D. Big Data Application on Convertor Control Equipment of Steel Melting Shop (SMS)

Big Data is processed on a Hadoop single node server to reveal useful information on pressure, temperature and flow parameters. The results of MapReduce program execution on single node Hadoop implementation are shown in table 4 below. In this table tag name, acceptable value range and results found from data are indicated against tag types namely temp (^oC), pressure (m^3/hr, kg/cm^2) flow (m^3/sec).

TABLEIV. Deviation Detection From Sensor Data

Tag Name	Acceptable Value Range	Results from Operational Data	Tag Type
GCP3_ANALOG\COC	0 - 100	100	Temperature
GCP3_ANALOG\COCF	0 – 90	90	Temperature
GCP3_ANALOG\FT03C	0 – 100	50-65	Pressure
GCP3_ANALOG\FT05C	0 - 2500	1040-1390	Flow

GCP3_ANALOG\FT05CF	0 – 2300	2000- 2500	Flow
GCP3_ANALOG\FT30C	0 – 195	100-240	Pressure

VI.RESULTS AND DISCUSSION

Important results for top five sub-equipment categories observed from R command prompt in client system are recorded after the completion of MapReduce jobs in Hadoop system and are shown in table 5. This study is performed to analyze monthly working hours of different equipment and their failure patterns that are observed from the SMS operational data.

TABLEV. K-Means Algorithm Results In Hadoop

Sub_Eqpmt	Mechanism	Make	Rating(KW)	Work_hours
Dummy Bar Conveyor	Dummy Bar Conveyor	Motorola	11	156.2
CNC Machine,408	Tacho for Spindle Motor	Siemens	75	155.7
Mould Cooling	Mould Water Com &S1-4 Valves(67)	GEC	60	163.4
Tundish Blower	Tundish Blower (Spare Mtr)	NGEF	4	178.9
Ladle Furnace	Fes Rotary Feeder	Flender	30	175.8

The above information helps in minimizing equipment breakdowns but is not available in ERP business warehouse. The comparison of above data is graphically shown in fig 8.

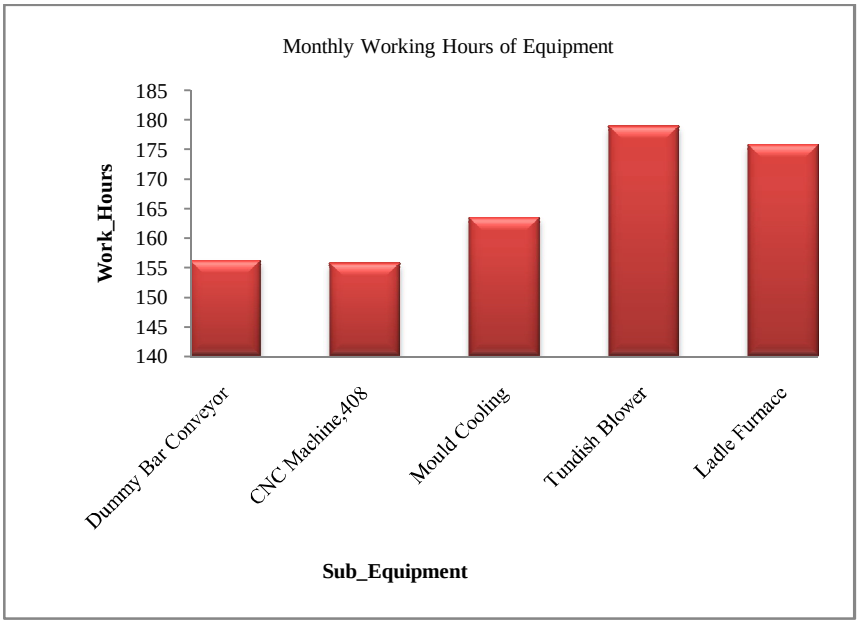


Fig. 8.Monthly working hours of equipment

A. Comparative study with SAP Reference Architecture for Big Data Analytics

SAP reference architecture for Big Data Analytics in ERP systems is shown in figure 9 below [35] [42].

From the below reference architecture, it is observed that integration with Hadoop is performed through SAP products like HANA, BO (Business objects). SAP Business Warehouse (BW) is merely positioned as a data repository above which third party Big Data analytic tools are integrated. SAP Predictive analysis (SAP PA) tool is built on open source R packages for extending predictive functionality to SAP ERP system [43].

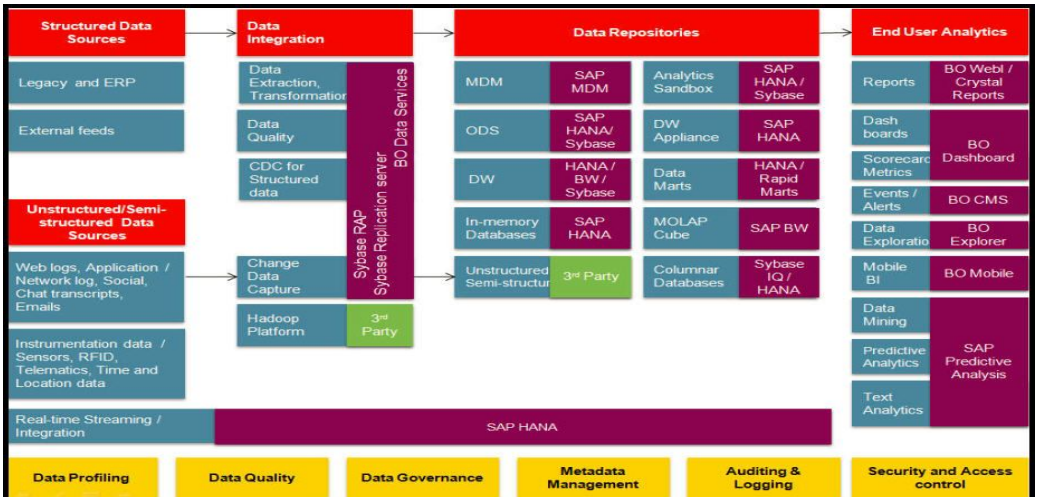


Fig. 9.SAP Big Data Reference Architecture (Courtesy:www.sap.com)

Predictive Analysis replaces SAP's *Predictive Workbench tool* which is based on SPSS PASW modeler engine.After SPSS was acquired by IBM in 2009, SAP no longer supports this tool.

The working methodology of SAP PA framework is briefly explained below [40] [43].

1) Organizing data: In this stage, business issues are identified and data is extracted from RDBMS database servers, CSV/Excel files etc. Based on business requirements, creation of formulae and other key metrics is also performed.

2) Defining Model: The support for modeling is available in three types of libraries namely PA native library, HANA predictive library and R open source library. The models can be shared via Predictive Model Markup Language (PMML) into data mining tools such as Weka, R and Rapid Miner etc.

3) Analyzing results: The model is executed with valid training data and the results are viewed graphically. SAP PA tool can be run on desktop or integrated with SAP HANA in-memory computing platform.

4) Publishing and sharing model results: The results of analysis are published to ERP enterprise portals.

B. Advantages of proposed methodology for Big Data Analytics in ERP BW systems over the existing approaches:

1) *Flexibility in handling arbitrary data structures:* The proposed extended framework uses Hadoop tools which interact with non-relational databases and commodity servers to handle data sizes in the Exabyte range. These sizes are much higher than the 100 TB ranges that SAP HANA and Relational Database Management Systems (RDBMS) can handle [33]. Current versions of Hadoop can handle arbitrary data structures more easily and usually at much lower hardware storage costs per terabyte than SAP HANA system.

2) *Big Data Governance and Data latency in ERP Systems:* For effective data governance in enterprises, the business rules have to be applied consistently to both real-time and batch data by using a hub approach (Business Warehouse). When blending structured and unstructured data, standard BI content in SAP BW is explored for business purpose and the results are shared faster. It is observed that current versions of SAP HANA connectors for Hadoop HDFS storage are significantly slower than Hadoop YARN tools, taking minutes or hours to provide analytic results. In Hadoop YARN, data is filtered by using customer specific key words, categories and brands etc

3) *Flexibility in data process in:* Hadoop tools used in the extended framework provide a flexible approach in processing the Big Data stored in Hadoop HDFS file system. To fully benefit from this, data in Hadoop HDFS is kept in its original raw form wherever practical, so that the original information is always available for new analysis. Atomic data is put in Hadoop and summarized data is transported into ERP Business Warehouse (BW).solutions offered by SAP HANA [40], SAS or IBM SPSS.

VII. CONCLUSIONS

Hadoop framework combines heterogeneous data sources and facilitates the processing of Big Data in ERP BW systems. Hadoop tools divide Big Data into multiple parts so that each part can be processed and analyzed at the same time. Hadoop framework has limited capabilities for performing advanced analytics because of the non-availability of parallel predictive algorithms required for running Big Data applications. This gap is filled by R tools which provide many predictive algorithms that can interface with

Hadoop HDFS. In the present paper, the authors have discussed some new methodologies to integrate R software tools with SAP ERP BW systems and to process Big Data, where ERP business warehouse tools are not available.

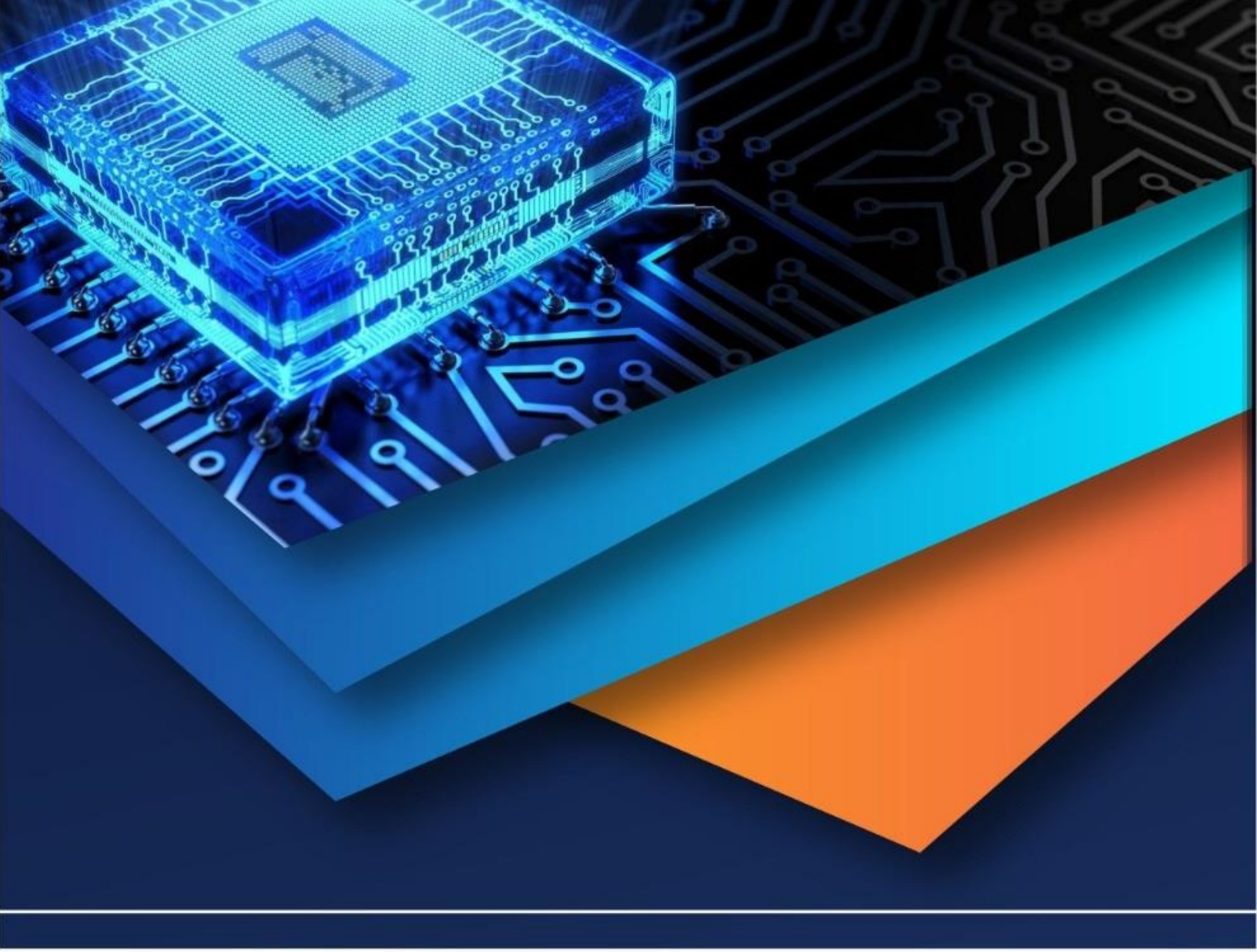
The proposed methodology shown in section III (figure 2) keeps Big Data in Hadoop system and allows analysis without much data movement by using R connectors. This scheme reduces the total cycle time to build and deploy predictive models, and eliminates potential security issues from data movement or replication. This framework uses the R connector to interface with Hadoop for read or write access to Big Data, which is an advantage over SAP HANA system that requires proprietary ABAP (Advanced Business Application Programming) connectors. Using the extended framework, Big Data applications are implemented at a much lesser cost than commercially licensed products such as SAP Data Services and SAP Predictive Analysis. The results of analysis were used in predicting working hours of equipment and monitoring equipment condition in a steel plant.

There is a great change in analytics world today with the convergence of Big Data, predictive analytics and real-time information delivery. Big Data processing in ERP BW Systems is likely to be influenced by the availability of new R packages for predictive analytics. The new predictive models can efficiently interface with Hadoop framework and provide an enterprise level framework fordecision making. Future scope exists for empirical validation of the Big Data methods discussed in this paper with emerging Big Data reference architectures such as SAS Enterprise Miner, IBM SPSS, Oracle Big Data appliance etc.

REFERENCES

[1] Bendoly and Elliot. "Theory and support for process frameworks of knowledge discovery and data mining from ERP systems." Information & Management, vol. 40, no. 7, pp. 639-647, Aug. 2003
[2] Botta-Genoulaz, Valerie, P-A. Millet, and Bernard Grabot. "A survey on the recent research literature on ERP systems." vol. 56, no. 6 pp. 510-522, Aug. 2005
[3] Injazz J. Chen, "Planning for ERP systems: analysis and future trend", Business Process Management Journal, vol. 7, no. 5, pp. 374 – 386, 2001
[4] Symeonidis, Andreas L, Dionisis D. Kehagias, and Pericles A. Mitkas, "Intelligent policy recommendations on enterprise resource planning by the use of agent technology and data mining techniques." Expert Systems with Applications, vol. 25, no. 4, pp. 589-602, Nov. 2003
[5] Chuck Schaeffer, “Big Data is Big Deal”, http://www.socialerp.com/, last accessed May 2014
[6] http://www.crmsearch.com/big-data-use-cases.php,last accessed December 2013
[7] http://whitepapers.inside-bigdata.com, last accessed May 2014
[8] Ajay Chandramouly et al, “Predictive Analytics and Interactive Queries on Big Data”, https://software.intel.com/, Intel paper on Big data Analytics
[9] Gregory Piatetsky, “Top Languages for analytics, data mining, data science”, http://www.kdnuggets.com/2013, Aug 27, 2013.
[10] Hai Qian, “PivotalR: A Package for Machine Learning on Big Data”, The R Journal, Vol. 6, No. 1, June 2013
[11] Yanchang Zhao, “R and Data Mining: Examples and Case Studies”, http://www.RDataMining.com, April 26, 2013
[12] Jeff Kelly, “Big Data: Hadoop, Business Analytics and Beyond”, Feb 05, 2014
[13] Revolution Analytics, RHadoop , 2011
[14] Olha Hrytsay, “Building Predictive Analytics on Big Data Platforms”, SoftServe Innovation Conference in Austin, Texas 2013
[15] Paul C et al, “Understanding Big Data: Analytics for Enterprise class Hadoop and Streaming Data”, McGraw-Hill, 2012, pp. 50-75
[16] http://mahout.apache.org/
[17] Julio Gonzalez, “Big Data Analytics Research Report”, http://www.slideshare.net/calidadgmv/big-data-analytics-research-report
[18] Challenges and Opportunities with Big Data - A community white paper developed by leading researchers across the United States
[19] Big Data A New World of Opportunities, http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf, last accessed December 2012
[20] Big Data and Analytics, http://www.paradigm4.com/wp-content/uploads/2012/01/MassTLC_BDR, last accessed Dec, 2013
[21] Survey of Recent Research Progress and Issues in Big Data, http://www.cse.wustl.edu/~jain/cse570-13/ftp/bigdata2.pdf
[22] Research Trends, Special Issue on Big Data, http://www.researchtrends.com/wp-content, last accessed Sep, 2012
[23] Usama Fayyad, Grogory p, Smyth P, “Knowledge Discovery and Data Mining:Towards a Unifying Framework”, Springer 2013
[24] Chen Jiang, David L. Jensen, Heng Cao, Tarun Kumar, “ Building Business Intelligence applications having prescriptive and predictive capabilities”, Proc.11th International conference on Web-age Information Management, Berlin, pp. 376-385, 2010
[25] Madden, Sam. "From databases to Big Data." IEEE Internet Computing Vol. 16, no. 3, 2012
[26] Dittrich, Jens, and Jorge-ArnulfoQuiané-Ruiz. "Efficient Big Data processing in HadoopMapReduce." Proceedings of the VLDB Endowment 5, no. 12, 2012
[27] Harriet Fryman and Linda Briggs, “TDWI E-Book: Big Data Analytics”, The Data Warehousing Institute, Sep. 2012
[28] Yanchang Zhao, “R and Data Mining: Examples and Case Studies”, http://www.RDataMining.com, April 26, 2013
[29] David Rich, David Smith, “Big Data and Revolution R”, Revolution Analytics, october 2011
[30] David Loshin, “Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL” , Elsevier Publishers, pp. 110-132, August 2013
[31] Krish Krishnan, Newnes , “Data Warehousing in the Age of Big Data”, pp. 245 – 275, May 2013
[32] Nathan Marz, James Warren, “Big Data: Principles and Best Practices of Scalable Realtime Data Systems”, Manning Publications Company, pp. 215-238, Sep. 2013
[33] Steve LaValle, Eric Lesser et al, “ Big Data, Analytics and the Path From Insights to Value”, MIT Sloan Management Review, Dec. 2010
[34] Stephen, Kaisler, Frank Armour, "Big Data: Issues and challenges moving forward." 46th Hawaii International Conference on System sciences (HICSS), pp. 995-1004, IEEE, Sep. 2013
[35] Big Data Analytics, http://www.ericsson.com/ wp-big-data.pdf, last accessed August 2013
[36] Tackling Big Data, http://csrc.nist.gov/, last accessed March 2014

[37] Scholarly Research in Marketing: Trends and Challenges in the Era of Big Data, http://web.uri.edu/business/files/Encycl-Communication-DataMining-n-Marketing-.pdf, last accessed Apr. 2014
[38] Scaling Big Data Mining Infrastructure: The Twitter Experience, http://www.kdd.org/sites/default/files/issues/14-2-2012-12/V14-02-02-Lin.pdf, last accessed Jun. 2014
[39] Carl W. Olofson and Dan Vesset, "Big Data: Trends, Strategies, and SAP Technology**",** Aug. 2012
[40] Tom White , O' Reilly, "Hadoop: The Definitive Guide", pp. 545 – 578, Jan. 2012
[41] Chris Eaton et al, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data", 2012,Mc-Graw-Hill, pp. 69-72
[42] Frank Ohlhort, "Big Data Analytics – Turning Big Data into Big Money", John Wiley & Sons, pp. 152-169, Feb. 2013
[43] Michael Minelli, Michele Chambers, AmbigaDhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Business", John Wiley & Sons, pp. 55-75, Dec. 2012
[44] Jay Liebowitz, "Big Data and Business Analytics", CRC Press, pp. 181-194, Jun.2013



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY