**INTRODUCTION**

The term Big Data refers to collecting huge volume of data from many different sources, processing them using some techniques and storing them in a huge storage area. We need to perform analytics in order to examine huge volume of data to uncover the hidden patterns, their co-relations and some other insights. [1]Organizations have understanding that by capturing the data that stream into their business, collecting past and present data they can perform analytics and predict the future and take some decisions. As and how the data grew the more difficult it was for traditional systems to handle the data and difficult to analyze it, this is where the Big Data come into picture. Today the business can identify the insights for immediate decisions compare to earlier days and this gives organizations a competitive edge.

The big data analytics helps the organizations to grow by the following steps:

* *Boosting the performance via better data analysis*: We can analyze how well the organization is working and how efficient it is working.
* *Understand the market and strategies of competitors*: Helps to understand market sentiment towards the organization and the opponent’s strategies.
* *Reduce risks through deep understanding*: Organization’s risk will be reduced by making proper business decisions.
* *Innovation of new services and products*: Can analyze what the customers want and provide a better service.
* *Precisely monitoring online reputation of brands*: Social media help to get a good idea on the brand’s views.

As big data is becoming more integral part of organization, there is a huge scope of big data analytics experts or professionals and they will be the main people in industries.

As in real world the data would be incomplete, noisy and inconsistent it spoils the quality of data, so to improve the data quality, the data has to be pre-processed. Pre-processing consists of many categories:

1. Data cleaning
2. Data integration
3. Data reduction
4. Numerosity reduction
5. Discretization and concept hierarchy generation
6. **Data cleaning:** This helps in finding missing values (tuples) assuming data is incomplete. It also smoothens noisy data which is nothing but a random error by ‘clustering’ or ‘data binning methods’.

* Binning performs local smoothening, smoothes a sorted data by neighborhood of values.
* Clustering helps in detecting outliers.

1. **Data integration:** Combines the data from multiple sources into a data store (files, databases, etc.,).‘How can the data from multiple sources be matched?’, Entity identification. ‘Redundancy’ is another issue where there will be many unrelated data is loaded. Inconsistency in attributes or dimension naming can cause redundancy.
2. **Data reduction:** Involves following reductions.

* Data cube aggregation: Here aggregation operations are applied to data while constructing data cube.
* Dimensionality reduction: Here we detect and remove irrelevant data and redundant attributes.
* Data compression: Here we make use of encoding mechanism to reduce the data set size.

1. **Numerosity reduction:** The data is replaced or estimated by parametric method where we have smaller data representation (regression models) or non-parametric method (clustering, sampling).
2. **Discretization and concept hierarchy generation:** The raw data values for attributes are replaced by ranges. This hierarchy allows the mining of data at multiple levels of abstractions.

**IMPORTANCE:**

Big data is important as it helps the organizations to tackle their data and use them to identify the new options which in turn lead to smart business, efficient operations, profit in business. Big data is important as it helps in cost reduction where we make use of big data technologies like Hadoop and cloud analytics in order to save cost and to identify more efficient ways to do business.

[3]With Hadoop we can analyze different sources of data immediately as it follows Map Reduce technique, we can make faster and better decisions. Big data is able to track the customer needs based on analytical techniques and based on the results of analytical techniques the company creates new products and provide services as per the customer needs. [8]In the dynamic business atmosphere, aggregating data with couple of projects and applications in different forms into one set for analyzing is a great challenge.

To handle big data we need to consider three key areas:

* To estimate past performance and stopping minute issues from growing huge business analytics make use of facts.
* To identify trends and capture sentiment, customer analytics is significant to know customer conversations.
* Consumption of products, services, determining leading and lagging areas consumption analytics is necessary.

To transform data, one should determine the approach to every analytic group and to do this we have two high level approaches for analysis, they are:

* Real time analytics: Here the analysis is performed on live data which include identifying issues when they occur and repair them before they affect the project.
* Batch analytics: Identifies trends over time or historical analysis where one or more data sets are scheduled for analysis. Both methods are advantageous in different contexts and as per the business needs its use is identified.

Some *Analytical techniques* in big data include:

1. **Descriptive analysis:** This analysis describes or summarizes raw data and makes it understandable to humans. It describes the past and helps analyze future outcomes. For example, reports on financial, sales, company’s production etc., we use these analyses when we need to understand an aggregate level, ex: what is going on in company and describe different aspects of business. It summarizes what happened.
2. **Predictive analysis:** This is the next step for data reduction which makes use of modeling, data mining etc., to study recent and historical data. It helps analysts to predict the future. It tells what might happen in future rather than what will happen in future. All predictive analytics are probabilistic in nature.
3. **Prescriptive analysis:** This identifies best outcomes when given parameters; suggest decisions options to mitigate future risks. It says how to respond to the future events.

**Characteristics of Big Data:**

The characteristics of Big data include

1. Velocity
2. Volume
3. Variety
4. Efficiency
5. Effectiveness
6. Complexity

Variety

Velocity

Volume

Big Data

Variety

Velocity

Volume

**Fig 1: Characteristics of Big Data.**

***Volume:***

It’s a large volume of data, the challenge of this characteristic is to store, process and access the huge data. Contradicting to the traditional storage ( RDBMS), the new technologies of Big Data are outlined to scale with the huge data that is to be stored and processed.

***Velocity:***

It’s the speed for Big data processing in response time. It refers to the data flowing in data infrastructure of enterprise at high rates; it must be processed with less latency. The response time here can be batch or real-time or streamed response.

***Variety:***

It refers to heterogeneity in data collected in order to process and analyze the data. It includes three types of data structured, unstructured, semi-structured. Structured data contain a defined data type, structure and format, the unstructured doesn’t have any structure and semi-structure includes textual data files in XML format.

These are the 3V’s of Big data. [3] Relating to trust and truthfulness of data we have complexity, efficiency and veracity. The key requirement to derive any perceptions or make judgment from the data sets is to estimate and understand the veracity. The quality of data captured differs at different rates thus affecting the accurate analysis.

The next characteristics are effectiveness, efficiency and complexity where the complexity conveys how complex the data is, should the complexity be reduced? how efficient the data? Is dealt with efficiency characteristic and we also see how effective the data is for any operation to be carried out.

Before the advent of Big data, distributed computing was widely used by the data scientists where many time consuming algorithms were replaced with distributed versions, for the current issues, distributed approach is more suitable as no batch architecture is able to tackle the huge data.

**Existing framework/methodology:**

[4]There are many frameworks existing and in use today and the first framework introduced was Map-Reduce framework and the most open source implementation of map-reduce framework is Apache Hadoop which is still in use.

Input key value pairs Input key value pairs

Data store 1 Data store n

Map

Map

Key1(values) Key2(values) Key 3(values) Key1(values) Key2(values) Key 3 (values)

Barrier== Aggregates intermediate values by output key

Key 1 intermediate Key 2 intermediate Key 3 intermediate value

value value value

Reduce

Reduce

Reduce

Final key 1 values Final key 2 values Final key 3 values

**Fig 2: Map-Reduce framework**

The figure 2[5] represents the Map-Reduce framework where the map and reduce are the two functions defined in key, value pairs. Map-Reduce function consists of pair of data with a type in one, data in one particular area and returns a list of pairs.

* Map (k1, v1) -> list (k2,v2)

The Map function is applied to every pair in the input dataset in order to produce a list of pairs for every iteration. The Map-Reduce framework collects all pairs from all lists with same set of key; group them, thereby creating a new group for each key. The reduce function is applied to the set of keys returned by Map function earlier where each group in turn produces collection of values in the same domain.

* Reduce (k2, list (v2)) ->list (v3)

Each reduces call produce either 1 or 0 value of v3. Therefore returns of all iterations are the result list. The Map-Reduce framework is casted by Google establishing Map and Reduce functions in functional programming.

Some of the key attributes involve Resource manager: to identify optimal computing operations.

* ***Optimized scheduling:*** Where it executes jobs based on priority.
* ***Flexibility:*** Any programming language can be used to represent the procedures.
* ***High availability and resiliency:*** Several job and task trackers make sure that jobs crash independently and then restart automatically.
* ***Have scale-out architecture:*** Ability to add up servers in order to increase the processing power

**Hadoop:**

It is an open source apache software framework that assesses the data and converts it into manageable form for the applications to work. [4] It is written in Java language, it can process the given data at a very high speed which is a very big advantage for scalability. Hadoop is a solution for several applications like image processing, analysis of web log, search indexes, indexing textual contents. Hadoop came out as an distributed platform for converting and managing huge datasets and meet many needs of applications listed above in a cost-effective manner. Hadoop distributed file system architecture [HDFS] is shown below in figure 3.

Second Name node

**Job Tracker**

Name node

**Job Tracker**

User

Data node

**Task Tracker**

Data node

**Task Tracker**

Data node

**Task Tracker**

Data node

**Task Tracker**

**Fig 3: HDFS Architecture**

Hadoop consist of two components a distributed file system and a computational framework. The data is stored in Hadoop distributed file system which follows write once and read many times model which breaks the data into blocks and spread across many nodes for fault tolerance and higher performance. It follows master-slave architecture, with an HDFS cluster having a name node which is primary and a data node- secondary. Name node manages the file system and control access to data while the data node manages the storage attached to node boxes that it runs on. There is a second name node which maintains a copy of data in name node in case of any failure.

Hadoop takes a group of nodes to execute Map-Reduce function in parallel, we have job tracker that schedules all the jobs as an individual task on a group of nodes. The job is split into several tasks in order to execute the worker nodes, task tracker runs on each node and it is responsible for initiating the task and reporting the progress to a job tracker. Lastly the Map-Reduce functions are performed.

[4]Some of the key attribute of HDFS are:

* *High availability*: provision of mission-critical applications and workflows.
* *Fault tolerance* where it recovers from the failure.
* *Scale-out architecture* where it is flexible to add servers in order to increase the capacity.
* *Flexible access*: we have multiple and open frameworks for file systems.
* *Load balancing*: The load is balanced in order to maximize the efficiency and utilization.
* *Replication*: for data protection and recovery.

**ADVANTAGES OF BIG DATA:**

* It allows the businesses to detect errors and thefts as soon as possible.
* It remarkably alleviates against losses.
* Big data analytics provide a major benefit from a competitive view point, allows business to develop more effective strategies towards competitors by making real time analysis.
* Offer businesses a good chance to improve gains and service for customer as the data collected is valuable [6].

**DISADVANTAGES OF BIG DATA:**

* Logistical issue: The need for the modification of entire approach leads the data flow in a company to constant rather than periodic.
* Sophisticated analysis for real time big data demands; by failing leads to implementation of incorrect strategies for entire organization [6].
* Currently used data tools are unable to handle the real time analysis.
* Privacy is the major factor.

**APPLICATIONS OF BIG DATA:**

Big data now a days is used in all areas including agriculture department, banks, bio-informatics, railways, healthcare, finance, enterprise, credit cards, telecom, stock, marketing etc.,[7].Big data plays a major role in all the above fields.

1. *Banking:* There arise some privacy issues regarding customer’s data when used invariably which makes the bankers cautious in their usage of big data due to privacy issues. Distribution of customer data across other branch also leads to some security issues. There are some information of customer getting leaked like personnel details and the amount deposited and withdrew and the bank balance of a customer.
2. *Stock*: Private stock exchanges make use of in-database analytics where inclusive system is established to detect fraud cases and abusive trading.
3. *Credit cards*: These companies rely on speed and accuracy of database analytics in order to identify theft cases.
4. *Enterprise*: Helps the business people make decisions in real time for less expense than that of traditional data analysis tools.
5. *Agriculture*: Sensors can be used to capture the data from fields like temperature, water levels, soil composition, growth of plants, etc., to optimize the crop efficiency. There will be a constant change in the attributes of above data.
6. *Finance*: The banks perform own credit score analysis for the customers who exist using a vast range of data, savings, credit cards and investment data.
7. *Economy*: Deals smartly with commodity hardware, where Hadoop helps the organization transition to low cost servers.
8. *Health care*: The use of big data in this field helps in making treatment decisions independently, using own clinical judgment without relying on any other rules based on big data. The nature of this industry itself creates a lot of challenges, as there are many players there is no easy way to share the data among different providers.
9. *Telecom*: Now a day’s big data is being widely used in telecom too where the service providers are in high competence, more subscribers rely on top players as providers of value added services are concentrated on enhancing the customer experience as a business goal. Researchers can still mine the data to see what treatment is more effective foe a particular disease, identify the drug and dosage of it and actually reduce the cost.
10. *Marketing*: Use of facial recognition to learn how well the success and failure takes in their products. A recent study published what kind of advertisements attracted people and discovered what kind of promotions did induce people to share the advertisements at social Medias and improve sales.

**TESTING TECHNIQUES AND CHALLENGES:**

Big data is a trend across many areas of technology, [9] data is continuously being generated by machines, sensors and many other sources. To examine and correlate datasets of rapidly growing data, characteristics of big data has to be considered. In recent surveys, the current scenario depicts 7.2 billion digital, mobile and social users where 100 TB of data is being processed by Twitter, 600 TB by Face book and 80% of data is unstructured. With such huge data it is difficult for an individual to identify and analyze the relevant data, most of the data is irrelevant, unknown, new and just go undetected which is severe issue in this area thus leading to obsession to industries, researches to collect, analyze, manage and test such voluminous and complex data. Based on many surveys the major concern identified by researchers is detecting unusual behaviors and the root cause of it.

**What is need of testing?**

As the complexity of big data is beyond the traditional processing applications the size of the datasets is variable and will adversely affect if the implementation fail, thus leading to an additional pressure to the testing team to prepare for big data test effort.

The key points that lead the testing team to success are:

* Testers should know the need of leveraging data.
* Testers should enhance their knowledge on data warehousing and business intelligence testing and know the differences between them.
* Designers and developers both should work with testers in order to gain leading edge.
* Testers should have knowledge on the framework that is being used which help them to test big data more effectively.

With increasing nature of data, its analysis, storage, transfer, security, visualizatuion and testing was always an issue so to overcome these issues there was an mapping between information and testing where an organized testing methodology was introduced. This led to enhancement of testing quality which helped in determining the faults early and eliminate the expense of usage.

The resources need to built and tested and the huge information testing require skills like coding, white-box testing abilities, skill of examining information in order to distinguish quality issues in information. Mohit Dayal and Bheemasankaram came up with the deficiency of testing hypothesis with wide specimen to present an approach that is more effective, productive, and proper. They identified that the discourage is at hypothetical level of asymptotic hypothesis and unidentified with proper demo. Incorporating those results with functional parts evoked model building approach which varies in great range with that of traditional approach.

Vast specimens are crucial for examining theories as the high measurable power and declined testing of flaws were categorized as type1 and type 2 rates. This super power method includes whole data investigation, displaying and identifying conclusions. Shirish Bhale came up with possibility for execting testing and introduced percentage of devices accessible that fit to understand the difficulties and exhibited how MongoDB, Cassandra can be checked and anxity tried making use of particular instruments. Alexander, Christhope and Volker introduced a method of regression testing, where workload currently in testing and workbench group is unable to catch unpredictability of volume data and identified the challenges of parallel information and their processing.

**Problems faced during testing data**:

1. Dealing with huge data and executing it across several nodes is a high risk as there could be a bad data and there could be issue in every stage regarding quality of data.
2. Integrating huge data that is available is also a huge issue which forces the organizations to have reliable and clean data; this can be ensured by testing all data sources end to end.
3. Continuous data collection and deployment is also one of the issues where one should test the data before live deployment.
4. There are some real time scalability issues where a basic flaw in structural components representation of huge data engenders terrible circumstances.

**Testing steps for verifying applications of Big data:**

A high level overview of phases in Testing Big Data Applications is shown in fig4.

Test Entry Points

BI

Target Data Ware house

ETL Process

Source Hadoop

Data Source (RDBMS, Social Media, etc,)

**Fig4: Big Data Testing**

**Stage 1: *Data Stating Validation***

This first stage is also referred as pre-hadoop stage it consists of process validation

1. From the various sources like RDBMS, media, weblogs, the data is collected, which should be validated before pulling it to into the system.
2. Comparison between pushed data and source data should be done to verify whether they match or not.
3. Verification should be done such that right data is extracted and loaded to HDFS.

**Stage 2: *Validation of “MapReduce”***

Here, the validation on every node is been done to verify the business logic it ensures that;

1. Process MapReduce works correctly.
2. Data segregation and aggregation rules are implemented on data.
3. Generation of key values
4. Validation of data after MapReduce

**Stage 3: *Validation of Output Phase***

This is the final stage, where it is called output validation process. Here, output data files are generated and then they are moved to EDW or to other systems based on the requirement.

1. Checks the transformation rules.
2. Checks the data integrity and successful data loaded into the target system.
3. Checks the data corruption.

**CHALLENGES IN TESTING ARE:**

1. Performance
2. Scalability
3. Continuous availability and security of data
4. Meeting data speed, understanding it and addressing its quality
5. Failure of node
6. **Performance:**

As the data is more volatile and unstructured manner performance testing is of no use and to access detailed information from this huge data at very high speed and with high granularity makes the challenge more worse.

1. **Scalability:**

It’s the ability of framework to manage workloads by extending the framework in a clear way and on the other hand, it is difficult to estimate the situations that will profit by profit by adapting frameworks. Working with huge data in parallel with various machines becomes a tedious task and in different machine environment disappointment becomes a concern in designers stress. If the machine is corrupted there is regaining of data so synchronization between machines remain a challenge.

1. **Continuous availability and security of data:**

For business applications the data should be available at all times but in NoSQL systems have little downtime built in default and some technique is required to overcome this issue. The data may contain the confidential information which has to be stored securely but providing high security to such huge data is also a threat as there could be an involvement of third party and NoSQL big data have few mechanisms that could tackle with this issue.

1. **Meeting data speed, understanding it and addressing its quality:**

It is very much required to analyze and find the required data in this competitive environment but understanding big data is itself a major challenge. To make visualization better, addressing quality of data is also a major challenge, the data should be in a proper format where the users can easily use it if not they it will affect the organizations to take decisions.

1. **Failure of node:**

In case of Hadoop, Cassandra the data is widely distributed among several nodes where there are chances of node failure or network failure which leads to data loss and this has to be tested against all nodes which is also a great challenge.

Hadoop is the one which can handle all these issues by overcoming all traditional issues regarding storage and computation as it is least expensive and easy to use. To overcome performance issue, cost and high availability issue we can make use of NoSQL solutions. Scalability issue can be overcome by clustering technique where we group similar objects in one cluster. There are testing solutions proposed to various issues and some are as follows

1. Testing approach for performance issues
2. Testing approach for scalability issues
3. ***Testing approach for performance issues:***

We have three solutions for performance testers to increase the performance, in parallelism there are two ways where the database accomplish parallelism, "Threading with locks” and “shared memory". In the prior case, various client operations are performed simultaneously, yet in the event that two operations require the same information so they must serialize.

The latter approach to accomplish parallelism is to intelligently segment application information where work is separated in light of the information the work must get to. The two operations require the same legitimate information, they still must serialize which is an issue. So NoSQL provide a solution to this issue by supporting all types of data, providing read-anywhere and write-anywhere functions and sort the unstructured data. Cassandra is another NoSQL solution that accommodates large and complex requirements of Big Data workloads.

It enables sub response time with linear scalability in order to increase throughput linearly with the increased number of nodes and to deliver very high speed to customers. Indexing is also a solution where one can sort records in multiple files and NoSQL provide a separate database for indexing. Binary search technique is used to access the indexed data which increases the performance greatly.

**Performance Testing Approach**

Performance testing for big data application involves testing of huge volumes of structured and unstructured data, and it requires a specific testing approach to test such massive data.

Optimum configuration

Execution and Analysis

Prepare individual clients

Identify and Design workload

Set up the big data application

True components and deployment

Objectives not met

**Fig5: Performance testing approach**

1. ***Testing approach for scalability issues:***

* The techniques to resolve this issue are
* Clustering technique
* Data portioning

**Clustering technique:**

Here the data is distributed across nodes of a cluster as and how loaded. Large data files are divided into pieces which are than handled by different nodes in a cluster. Each data piece is mirrored across multiple machines to resolve the issue of data loss. Ex: Hadoop. In Hadoop HDFS (Hadoop Distributed File System) divides huge data into pieces which are than handled by nodes in a cluster. Here file pieces are mirrored across many machines to form a single namespace. Hadoop is specifically designed to have a very flat scalability curve. Once a Hadoop program is written for functioning of ten nodes, very little re-work is required for the same program to run on a larger amount of hardware. Adding nodes to the cluster can increase both the capacity and performance of the database.

**Data partitioning:**

Here the system treats CPU centers as individual from a group and highlights the authorized allotment of information at each CPU center. This method have few pros and cons, it does scale well for couple of machines with numerous centers per machine, or for bigger numerous centre bunches. It is particularly favorable for composing workloads, which are hard to support with normal reserving frameworks. For people who use this approach, the information administration and control part is done in lock free, single-strung code. Apart from velocity, this makes the code more solid and simpler to investigate.

Another approach where a tester can test the data quality is to perform functional test which make sure that data is clean and identify related issues like coding errors and node configuration errors and at every stage it is required to ensure that data is being processed errorless. While performing this approach, the tester will perform validation for the following data extraction, validate whether the data is correctly loaded in database, is the data replicated correctly across all nodes in a network, validate that whether the data output report is generated properly for the processed data, validate load in target system, validate aggregation of data and integrity of data.

This can also be done by data sampling methods where the raw data is converted into required format in order to compare it with actual output data and to identify the difference in structured data compare tools are used in case of structured form of data, in case of semi-structured the data is first converted into structured form and than compare the expected and actual results. In case of unstructured the data is parsed.

If all V’s of big data is tested efficiently than testing of big data can be tested efficiently. Many tools can also be applied to test data format and its similarity. To test big data more efficiently, functional and incremental load testing is required.

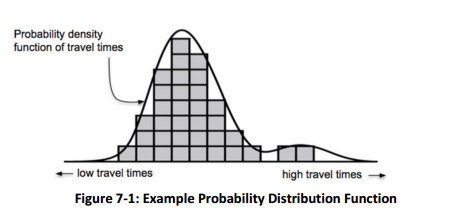
**METRICS IN BIG DATA:**

1. Big data performance metrics:

[10]New technologies and performance approaches are transforming conjestion monitoring practices all over nation which resolve and improve monitoring program by letting access more data for less cost thus enhancing the scope of conjestion analysis. Considering the commercial speed data from INRIX- the global SaaS company that provide real-time traffic information, state traffic analytics, traffic collisions, car services data etc., the performance metrics and congestion duration are computed and analyzed for Alameda country’s freeway network.

The reliability metrics considers the travel time variability where it is significant to determine the time to allow a trip to arrive on time with some degree of certainity and unreliable travel times can occur because of normal demanding fluctuations, inclement weather, incidents and some special events. These factors cause significant changes in journey time. Reliability concept is depicted by probability distribution where awareness may be obtained by

* High point in the graph aligning the experienced travel time
* Left and right parts of distribution aligning with minimum and maximum experienced travel time
* Range or difference of travel time between minimum and maximum occurring travel time.



**Fig6: Probability Distribution function with reliability metrics**

Performance measures are defined to compare reliability are:

*Planning time:*

This tells how much time is allowed for a trip to ensure 95% on-time arrival and its equal to 95 percentile of experienced travel time. To express we use following formula

Planning time = 95th percentile travel time

*Planning time index:*

the planning time index is a ratio of the 95th percentile travel time to the free flow travel time in order to allow for comparison across different routes and different trip lengths. It can be represented as

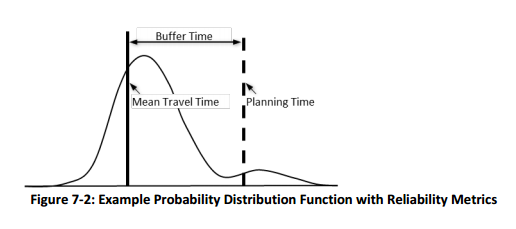
Planning time index = 95th percentile travel time/free flow travel time

*Buffer time/Index:*

This represent extra buffer that allow in addition to average time travelled to account for any delay. It can be represented as

Buffer time = 95th percentile travel time – mean travel time

Buffer index = 95th percentile travel time – mean travel time/ mean travel time

****

**Fig7: Probability Distribution function with reliability metrics**

**EXPERIMENTAL RESULTS:**

[15]Here we are considering two real world databases (Accidents and Shop4) , by using pattern-growth++ the periodic-frequent patterns are produced.

***Experimental Setup:***

1. Database of Shop4: here the data of seven online stores of ECML/PKDD discovery challenge is produced by Czech company, here the data of product category is taken for consideration, it states the each transaction with the transactional database which represents the people who visited the set of web pages at a particular *minute interval.* It consist of 155distinct items and 59,240 transactions
2. Database of Accidents: Here the data of aircraft damages by Federal Aviation Authority (FAA) is collected; this was made available in Aviation Safety Information Analysis and Sharing (ASIAS) for the purpose of safety improvement. Here the collected database is from Created database of ASIAS from 1/1/1978 to 31/12/24.

Tmax, TAvg, and Tmin represent the maximum, minimum and average number of items within a transaction. It is represented in table 1.1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Database** | **Tmin** | **TAvg** | **Tmax** | **Size** | **Items** |
| **Shop-4** | 1 | 2.4 | 82 | 59,240 | 155 |
| **Accidents** | 3 | 8.9 | 9 | 96,864 | 9,290 |

**Table1.1: Database statistics of Experimental Results**

**CASE STUDIES:**

***1. Precision big data in agriculture Weather forecasting:***

Precision big data in agriculture Weather forecasting for the future beneficial of farming. This mainly deals with the details of agricultural data by using big data analytics approach. Here we have a situation where we will be using the information and communication technology (ICT) in the environment of agricultural big data for getting huge amount of data. Here it provides detailed instructions on weather forecasting decisions, productivity with the improved yield methodology, it also prevents the unwanted cost issues during harvesting, using fertilizers and pesticides.

Here previously generated called historical data of real time data with more structured and unstructured data are used and analyzed. Detailed information about agriculture data produces knowledgeable information of data which will be useful for them in future. The future farming is very useful in improvement in agricultural products quality. It also provides wide range of functions and intelligence discovery and details of data to address variety of importance of agricultural problems and decisions.

The overall data in the sector of agriculture are produced by using the on-sight farming. The important aspects in present scenarios are to provide the effective development in the production of agriculture and world population feeding, and also managing disaster. Here the data types collected are both structured and unstructured data types, where both heterogeneous and homogeneous technologies are used to gather these data.

To improve yield production, millions of dollars has been invested in precision agriculture. For better decision in agriculture, green zone or dry zone production is necessary. The data that is collected to analyze the weather are the rainfall in (mm), temperature in degree (C), humidity in (%) for nearly ten previous years in order to analyze and predict weather. Hadoop, which is a parallel and distributed platform developed by Apache is used to process big data and its components help to procure data, manage and visualize applications in ICT. To side by side store and process data, to analyze data that do not fit in memory, Map-Reduce programming model is used and to store this huge data HDFS (Hadoop distributed file system) is used.

The predictive model here consists of Map-reduce for processing big data and linear regression analysis for analyzing and predicting the data. The map reduce model solve the problem of big data by scattering it over distributed file system and to do this GFS (Google file system) is used to scatter data over clustered network. Predictive analytics is used to extract knowledge from the existing data, determine patterns, predict future and come up with some outcomes.

**Analytical model:**

**[10]**The supervised leaning method for predicting future patterns from historical data and to handle large linear and non-linear data, regression model is used. In regression analytical model we have ‘x’ and ‘y’ numerical variables where the vector ’x’ consists of ‘n’ numerical variables (x1,x2,x3,…xn) and ‘n’ is number of attributes of each data bit in dataset. Linear regression is used when there is a dependency between the two variables. The equation

= f(x) +

It gives the difference between the real and forecasted value and is denoted by sigma, the forecasted value of y is f(x) and is denoted by y cap. The equation used to represent ‘y’ as a function of ‘x’ is given by

I = a0 + a1xi +

To calculate coefficients a0 and a1 we use the following formulas

a1 =

a0 = a1

The span of the x and a1 vectors are equal to the number of measurements in the feature space and the blunder term sigma is the distinction between the real and anticipated assessment of the target variable. The variable ‘y’ is used to represent the forecasted value of target variable through model and we have y = a0 + a1x formula to predict target variable and the algorithms to learn a0 and a1 estimations of the items present in the dataset. The goal is to limit the distinction between the real and predicted value for all data items as this difference could be measured by limiting the whole of square of the real and anticipated target values.

**Challenges faced are:**

1. Knowledge discovery and association from past records.
2. Getting to know well about Big data and unstructured format of data.
3. Managing large video data.
4. Crop monitoring data management through sensors and their communications.
5. It’s expensive for agriculturist to adopt and access latest technologies.
6. This lack of technological knowledge, the community of agriculture requires more
7. Training and security to the ICT instruments.

***2. MapReduce implementation on Health care Perspective analytics:***

A type of business which uses analytics that focuses on detecting the best action of course for a scenario is called Perspective analytics. “In the last few years the worlds 90% of data is generated”

The main preliminary requisites of perspective analytics are sharing of data among the nodes which are used for computing, where it is very difficult to achieve by using MapReduce. Thus it states that it is not possible to solve all problems of optimization based on MapReduce framework.

***Big Data analytics for clinical data:***

A new area called “Big Data” is emerged in the field of information technology because of the exponential production of data in the recent years. From the data of large-scale laboratory information system (LIS) these datasets are emerged. By using the MapReduce framework the data types like big data sets of biomedical, clinical, and biometric data are successfully processed. The promise of big data analytics in bioinformatics and health care is general and is more in descriptive form. However this extends to the MapReduce framework application and it also allows to implementation of Hadoop to the range of clinical data with:

Combined BLAST results

BLAST(results for subsets)

All genes

Subset of genes

NODE5

NODE4

NODE3

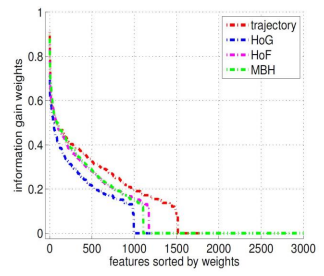
NODE2

NODE1

**Fig83: Big data block diagram**

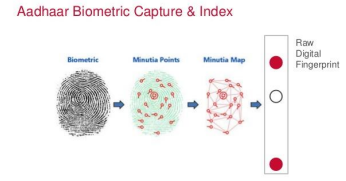
**Public datasets:**

To the proposal of common adverse drug event (ADE) the MapReduce algorithm is used. The main aspect of using MapReduce algorithm is to find out the possibility that the MapReduce algorithm usage to speed up the mining of biomedical data mining by using the pharma covigilance case. The MapReduce distributed architecture and high dimensionality compression via Markov boundary feature selection have been used to identify unproven cancer treatments on the World Wide Web.



**Biometrics:**

To classify the biometrics measurements the MapReduce programming is been used. The Hadoop platform is used for face matching, iris recognition and also for fingerprint recognition. For the security purpose while accessing the cloud data, a biometric-capture mobile phone application is having been introduced. During the standard web session the recognition and biometric capture are done. For the establishment of connection between the server and a user is done by using Hadoop platform.



**Challenges faced are:**

Lack of data utilization and unsustainable costs are the two major things by which health care system suffers from. Therefore there is need for finding the solutions for this reduction of costs. Enhancement in health care quality and the measures of costs mainly depends on the utilization of large integrated databases to produce outline of patterns. It’s not only the lack of data but also the information that can be used for utilization to support critical decision-making. These things lead to the following challenges:

* *Straggling of technology*- For the purpose of redesigning Health care is resistant and also technologies approval for the development of health care system.
* *Dispersion of data*- By the utilization of many sources the clinical data is generated this supports the purpose of data maintaining mechanism and data integration to have the in the flexible form of data warehouse.
* *Privacy and other Security issues*- By sharing the data between scholars and researchers there are lots of benefits, and because of some privacy and security issues these benefits are restricted and regulated by the laws of data privacy and access.
* *Regulations and Standards*- The architecture of big data should be more flexible so that it can be adoptable to manage the growth of regulations and standards also to manage variety of dispersed sources

**Future Scope:**

Big data is key feature which have the ability to unlock the whole health care value chain. It changes the traditional perspective of health care systems from finding new drugs to patient-centric health care for better clinical outcomes and with increased efficiency. [11] For the enhancement and accelerating interactions among clinicians, lab directors, researchers, logistic managers and administrator’s big data have a very bright future by creating better efficient and reducing the costs, improving personalized care and reducing the risks.

The following list produces the future applications:

* *E-clinics:* based on the retrieval applications on text analytics. The collected data is large and is mostly unstructured data and the aim is to discover knowledge by mining these types of text data.
* *Applications of Genotyping:* The Genomic data represents a specific amount of sequencing genes data and applications that requires the analysis and understands the sequence for the betterment of patient treatment.
* *Analysis and mining of biosensors applications:* Streamed data home monitoring, tele-health, handheld and sensor-based wireless are well established data sources for clinical data.
* *Application of social media analytics:* The communication between physician, patients and communities can be increased by these social medias, because of this requirement of analytics emerges to analyze the data and outbreak the feedback of patient satisfaction.
* *Applications of Business and organizational modeling:* The exponentially growing data like Administrative data (scheduling, billing and other non-healthcare present data) as source of data, and analyzing and optimizing these data may reduce the large amount of costs and also increases the sustainability of health care facility.

The above mentioned types of clinical data sources provide rich environment for many future applications and research. The future applications can be analyzed for better patient treatment and also for sustainable health care system.

REFERENCES:

[1] Bigdata analytics “importance-of-big-data-analytics-for-business-growth” may- 2014.

[2] Naveen Garga, “insights and analytics of big-data-analytics” jul-2014.

[3] Hu Shuijing, “Big Data Analytics: Key challenges and technologies”,IEEE 2016 International Conference on Robots & Intelligent System.

[4] E.Sivaraman, Dr.R.Manickachezian, “High Performance and Fault Tolerant Distributed File

System for Big Data Storage and Processing using Hadoop”, © 2014 IEEE.

[5] Mr.S.Ramamoorthy, Dr.S.Rajalakshmi, “ Optimized Data Analysis in Cloud Using Big Data Analytics Techniques”, IEEE July 4-6, 2013.

[6] chicklum, “pros-and-cons-of-big-data”, Sep 30-2015.

[7] Kuchipudi Sravanthi, Tatireddy Subba Reddy, “Applications of Big data in Various Fields”, Kuchipudi Sravanthi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, 2015.

[8]Computer Business Review “What is big data and and why is it important” 01-may-2016.

[9] Naveen Garga , Dr. Sanjay Singlab, Dr. Surender Jangrac, “Challenges and techniques for testing of Big data”, 1877-0509 © 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license.

[10] Joshua D. Woodard “Data Science and Management for Large Scale Empirical Applications in Agricultural and Applied Economics Research” Applied Economic Perspectives and Policy (2016) volume 38, number 3, pp. 373–388.

[11] Tim McGuire, James Manyika, Michael Chui, “Why Big Data is the new competitive advantage” July/august 2012.

[12] Ambar lee Dennis “Data Governance Success: Implementing the Appropriate Approach” April 12,2017.

[13] Ambar lee Dennis “case-study-implementing-data-governance-data-lakes-big-data” April 6, 2017

[14]Charles Roe “category and data-topics of data-governance” May 3, 2017

[15] Jessica Folliet “The Importance of Big Data and Data Visualization” January 12, 2016

[16] [William](http://www.tandfonline.com/doi/abs/10.1080/08982112.2014.846119) Q.Meker, “Rehabilitees meets big data – Challenges and opportunities”

[17] Xing Wu, “Reliabilities of Big data” IEEE 20-21 Dec. 2014