iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** VI **Month of publication:** June 2016

DOI:

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Transitioning From Relational Database to Big Data

Edwin Anto
Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai

Abstract -The amount of available data has increased by a huge amount in the past few years because of new social behavior and vast spread of social system. Big data have played a very important role for innovation and growth. This paper provides a brief review about the background of big data; explains the emerging technologies. We also address the challenges and opportunities arising from the use of big data.

I. INTRODUCTION

The amount of information that is being generated and stored is increasing exponentially day by day. It may be doubling every two years, according to the estimate by IDC 2011. New advanced analytics techniques and technologies allow users to connect and interrogate datasets. Big data outline technologies and innovative techniques to distribute, store, capture, manage and analyze larger-sized datasets with high-velocity and numerous structures that traditional data management methods are incapable of handling. Conventional data management and analysis systems are based on the relational database management system (RDBMS). It is clear that the traditional RDBMS could not handle the huge volume and diversity of big data. For outcome of persistence storage and management of large-scale messy datasets, distributed file systems and NoSQL (Not Only SQL) databases are good options. Traditional Data analysis approaches are: Cluster Analysis, Factor Analysis, Regression Analysis, Correlation Analysis, Statistical Analysis, A/B Testing, Statistical Analysis and Data Mining Algorithms. Big Data analysis approaches are: Bloom Filter, Hashing, Index, Triel & Parallel Computing. With the amount of data expanding there need to be number of obstacles which needed to be overcome such as technology challenges, organization limitations, and privacy/trust concerns. While the amount of large datasets is vigorously increasing, it also brings about many challenging problems demanding induce solutions. Big is characterized by the 3 Vs. The three Vs of Big Data are: Variety, Volume and Velocity.

A. Relational Database

A relational database is a group of data items that is organized as a set of formally-described tables from which data can be retrieved or reassembled in many different ways without having to reorganize the database tables. Data in a table can be related based on common keys or concepts, and the ability to extract related data from a table is the base for the term relational database. A Database Management System (DBMS) handles the way data is stored, maintained, and retrieved. In the instance of a relational database, a Relational Database Management System (RDBMS) performs these tasks.

1) *Dimensional*-In a dimensional approach, transaction data are divided into either "facts", which are generally numeric transaction information, or "dimensions", which are the reference information that gives context to the facts. Consider the example, a sales transaction can be split into facts such as the price paid for the products and the number of products ordered, and into dimensions such as order date, customer name, product number, bill-to locations, order ship-to and salesperson accountable for receiving the order. A key advantage of a dimensional approach is its ease of use and well understanding of the data warehouse. The main disadvantages of the dimensional approach are:

a) In order to maintain the integrity of dimensions and facts, loading the data warehouse with data from different operational systems is complex.

b) It is difficult to modify the data warehouse structure if the organization adopting the dimensional approach changes the way in which it works.

2) *Normalized:* In the normalized approach, the data in the data warehouse are stored according to a degree, database normalization rules. Tables are grouped together by subject areas that indicate general data categories (e.g., data on customers, products, finance, etc.). The normalized structure divides data into entities, which generates several tables in a relational database. When applied in large enterprises the result is dozens of tables that are associates together by a web of joints.

Volume 4 Issue VI, June 2016
ISSN: 2321-9653

Furthermore, each of the created entities is transfigured into separate physical tables when the database is executed. The main advantage of this approach is that it is a straightforward approach to add information into the database. A disadvantage of this approach is that, because of the amount of tables involved. It can be difficult for users both to:

a) Join data from different sources into significant information and then

b) Access the data without a precise understanding of the sources of data and data structure of the data warehouse.

Both dimensional and normalized models can be represented in entity-relationship diagrams as both include jointed relational tables. The difference between the two models is the degree of normalization.

3) *Big Data:* Big data is a term for data sets that are so massive or complex that traditional data processing applications are inadequate. Big data includes both structured and unstructured data that inundates a business on a day-to-day basis. But it's not the amount of data that's important. Challenges include analysis, capture, data, search, sharing, storage, transfer, visualization, querying, updating and privacy. Big data can be analyzed for insights that lead to better strategic and decisions business progress. Big data can be described by three V's. They are as follows:

a) *Volume:* The quantity of stored and generated data. The size of the data decides the value and potential insight- and whether it can actually be considered as big data or not.

b) *Variety:* The type and nature of the data. This helps people who analyze it to effectively and efficiently use the resulting Intuition.

c) *Velocity:* In this context, the speed at which the data is processed and generated to encounter the demands and challenges that lie in the path of growth and development.

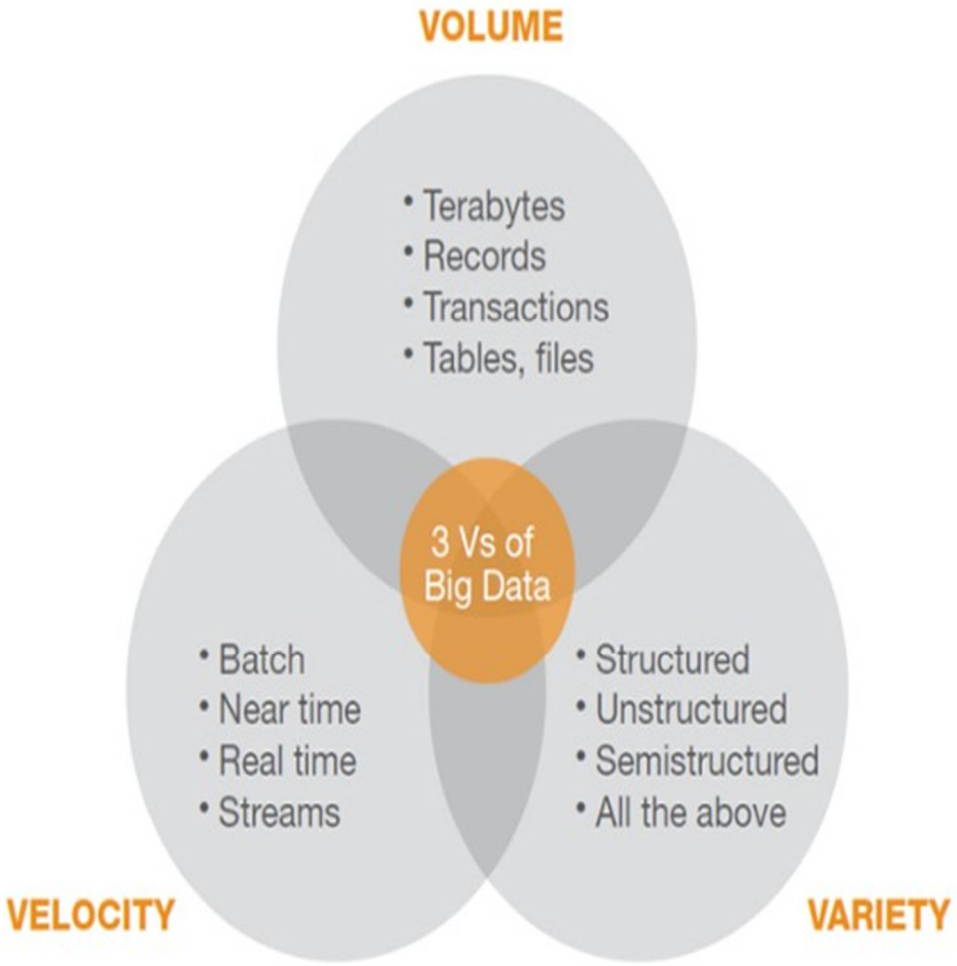


Fig.1 3Vs of Big Data

II. WHY DO WE NEED A TRANSICTION FROM RELATIONAL DATABASE TO BIG DATA?

CHARACTERSTICS	RDBMS	HADOOP
Basic Description	Traditional row-column database used for transactional system, archiving and reporting.	An open source approach for storing data in a file system across a range of commodity system and processing it utilizing parallelism.
Best for Applications	Reads and Writes reasonable data sets (1B rows).	Inexpensive storage of lots of data structured and semi-structured.
Types of Data Supported	Works with structured data only.	Works with structured, semi-structured and unstructured data.
Data Layout	Row-oriented	Column-oriented
Use	RDBMS is generally used for OLTP processing.	HADOOP is currently used for analytical and for processing Big data.

Comparison of Relational Database vs Hadoop

III. BIG DATA EMERGING TECHNIQUES AND TECHNOLOGY

For the purpose of processing the huge amount of data, the big data requires exceptional technologies. The various techniques and technologies have been implemented for manipulating, visualizing, and analyzing the big data. There are a number of ways to handle the Big Data, but the Hadoop is one of the prominent used technologies.

A. MapReduce

MapReduce is a programming model and an associated execution for processing and generating large data sets with a parallel, distributed algorithm on a cluster. MapReduce is the heart and soul of Hadoop. It is this programming paradigm that allows for massive scalability across a huge number of servers in a Hadoop cluster. A MapReduce program is composed of a Map() procedure that performs sorting and filtering (such as sorting products by name into queues, one queue for each name) and a Reduce() method that executes operation (such as counting the number of products in each queue, yielding product name frequencies).

B. Hadoop

Hadoop is the implementation of MapReduce, being an entirely open source platform for handling Big Data. Hadoop is an open-source framework for distributed processing and storage of very huge data sets on computer clusters built from hardware. All the modules in Hadoop are designed with a fundamental presumption that hardware failures are quite common and should be automatically handled by the framework.

C. NoSQL Database

Often interpreted as Not only SQL, provides a technique for storage and retrieval of data that is designed in means other than the tabular relations used in relational databases. Motivations for this approach include horizontal scaling, simplicity of design and good control over availability. NoSQL does not have a despotic definition but we can make a set of common observations, such as:

1) Not involve the use of relational model
2) Working well on clusters
3) It is mostly open-source
4) Built for the century world web estates
5) Schema-less

D. Hive

Hive is a data warehouse infrastructure tool to execute structured data in Hadoop. It resides on top of Hadoop to encapsulate Big

Data, and makes analyzing and querying much more easy. Developed by Facebook, Apache Hive is now used and developed by other business organization such as Netflix and the Financial Industry Regulatory Authority (FINRA).

E. PIG

Apache Pig is an important platform for creating programs that execute on Apache Hadoop. The language for this platform is called Pig Latin. Pig can execute its Hadoop jobs in MapReduce, Apache Tez, or Apache Spark. Pig is made up of two parts: the first is the language itself, which is called PigLatin and the second is a runtime environment where PigLatin programs are executed.

IV. SELECTING A BIG DATA TECHNOLOGY: OPERATIONAL VS. ANALYTICAL

A. Operational Big Data

For operational Big Data workload, NoSQL Big Data systems such as document databases have evolved to address a broad set of applications, and other architectures, such as column family stores,key-value stores and graph databases are optimized for more particular applications. NoSQL technologies, which were developed to overcome the shortcomings of relational databases in the present-day computing environment, are faster and inexpensively than relational databases.

1) *NoSQL:* NoSQL (originally refers to "non relational" or "non Structured Query Language") database provides a mechanism for retrieval and storage of information which is modeled in means other than the tabular relations used in relational databases. NoSQL systems are also referred to as "Not only SQL" to show that they may support SQL-like query languages.

NoSQL Database Types:

a) Document databases key pair with a complex data structure known as a document. Documents can comprise many different key-value or key-array pairs, or even nested documents.
b) Graph stores are used to store knowledge about networks of data, such as social connections. Graph stores consists of Neo4J and Giraph.
c) Key-value stores are the easiest NoSQL databases. Every single entity in the database is stored as an attribute name (or 'key'), along with its value. Some of the examples of key-value stores are Riak and Berkeley DB. Some key-value stores, such as Redis, enable each value to have a type, such as 'integer', which adds more functionality.
d) Wide-column stores such as Cassandra and HBase are optimized for queries over huge data sets, and store columns of data together, instead of rows.

There are lots of benefits of NoSQL in contrast to relational databases. NoSQL databases have good performance and their data model addresses a number of problems that the relational model is not designed to address. Such as:

a) Large volumes of fast changing structured, semi-structured, and unstructured data
b) Agile sprints, quick schema iteration and frequent code push.
c) Object-oriented programming that is flexible and simple to use.
d) Geographically distributed scale-out architecture instead of costly, monolithic architecture.

B. Analytical Big Data

Analytical Big Data workloads tend to be described by MPP database systems and MapReduce. These technologies are also a response to the constraints of traditional relational databases and their lack of capability to scale beyond the resources of a single server. MapReduce also provides a new technique of analyzing data that is compatible to the capabilities provided by SQL.

1) *Massively Parallel Processing:* MPP (massively parallel processing) is the coordinated execution of a program by multiple processors working on different areas of the program. Each processor has their own memory and operating system. MPP speeds the performance of large databases that deal with massive amounts of data. MPP databases use multiple processors and servers, multi-core processors, and storage devices equipped for parallel processing. That collaboration permits reading many pieces of data across many processing units at the same time for improved speed. This method is very much necessary because the frequencies of processors are hitting the limits of the technologies used. Almost all enterprises associated with big data have databases that are massively parallel. An MPP system allows a number of databases to be searched in parallel.

Volume 4 Issue VI, June 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

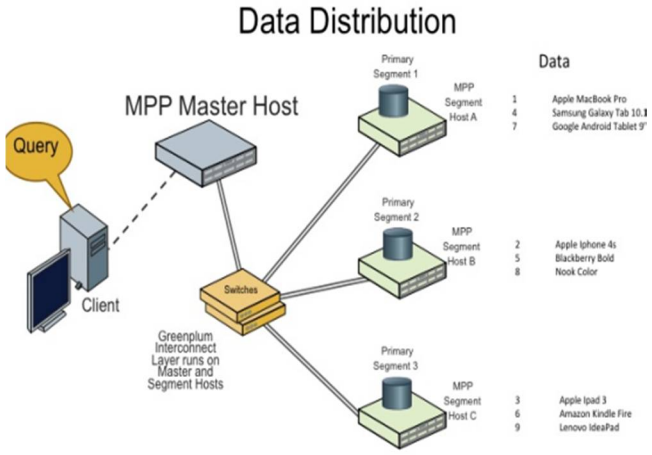


Fig.2 MPP Architecture

V. BIG DATA MIGRATION PROCESS

Migration generally happens in batches of entity. A set of entities is selected, then migrated and finally tested. This cycle goes on and on until all application data is migrated. Big data migration generally involves large volumes of data and multiple source systems. However, most of the organizations lack the open source tools to handle this important task of data migration.

Big Data Migration Process Hadoop as a service is offered by Amazon Web Services (AWS), a cloud computing technique that abstracts the operational challenges of running Hadoop and making large scale data processing accessible, fast, easy and inexpensive. The typical services available include Amazon EMR (Elastic MapReduce) and Amazon S3 (Simple Storage Service).

The migration to the Amazone Web Services (AWS) Hadoop environment involves a three-step process.They are listed below:

A. Cloud service

Physical/Virtual machines are used to connect and extricate the tables from source databases using Sqoop (Sqoop is a tool designed to transfer data between Hadoop and relational databases or any other mainframes), which pushes them to Amazon S3. Amazon S3 (Simple Storage Service) is an online file storage service offered by Amazon Web Services.

B. Cloud storage Amazon S3 cloud storage center is used for all the data that is being sent by virtual machines. It stores data in flat file format system.

C. Data processing

Amazon EMR processes and distributes extensive amounts of data using Hadoop. The data is seized from S3 and stored as Hive.

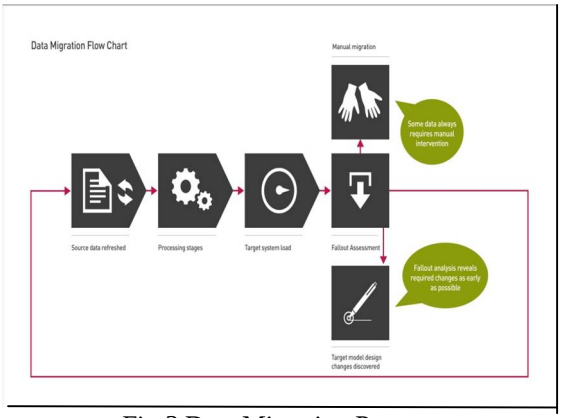


Fig.3 Data Migration Process

VI. CHALLENGES IN BIG DATA

The handling of big data involves very complex processes. Some challenges faced during its integration include uncertainty in data Management, getting data into a big data structure, getting useful information out of the big data, syncing across data sources, volume, big data talent gap, solution cost etc.

A. Getting Data into Big Data Structure

It might be evident that the intent of a big data management require analyzing and processing a huge amount of data .There involves a lot of complication behind the access, transmission and delivery of data and information from a wide scope of resources and then loading these data into a big data platform. The requirement to navigate transformation and extraction is not restricted to traditional relational data sets.

B. Budget and Capacity

Although big data analytics have great prospective to bring value to business organizations, traditional servers in organizations' data centers are not designed to process big data. Analytics servers and in some cases, high performance computing servers and applications will be needed, which requires huge amount of IT investment.

C. Talent Gap in Big Data

There is lack of skills available in the market for big data technologies. The typical specialist has also acquired experience through tool implementation and its use as a programming model, apart from the big data management feature.

D. Technology and Privacy

Technical problems also do exist for big data analytics. Privacy of data is another massive threat and one that increases in the context of Big Data. Business organization must initiate effective procedures, policies, processes and responsibility for big data analytics and use, and incorporate privacy and security restrictions before actually putting them into use.

E. Syncing Across Data Sources

Once you import data into big data platforms you may recognize that these data copies migrated from a wide scope of sources come in at different rates and schedules can quickly get out of the synchronization with the originating system. This indicates that the data coming from one source is not out of date in contrast to the data coming from another source. The order of data migration, extraction and transformation all emerge the situation in which there are risks for data to become unsynchronized.

F. Miscellaneous Challenges

Other problems may arise while integrating big data. Some of the challenges include veracity and validity of data, integration of data, the amount of data, skill availability, the rate of transformation of data, solution cost. It is also a challenge to process a huge amount of data at a fair speed so that data is accessible for users when they are required.

This is all about the big data integration and some problems that one can encounter during the implementation. These points must be examined and should be taken care of if you are going to manage any big data platform.

VII. CONCLUSION

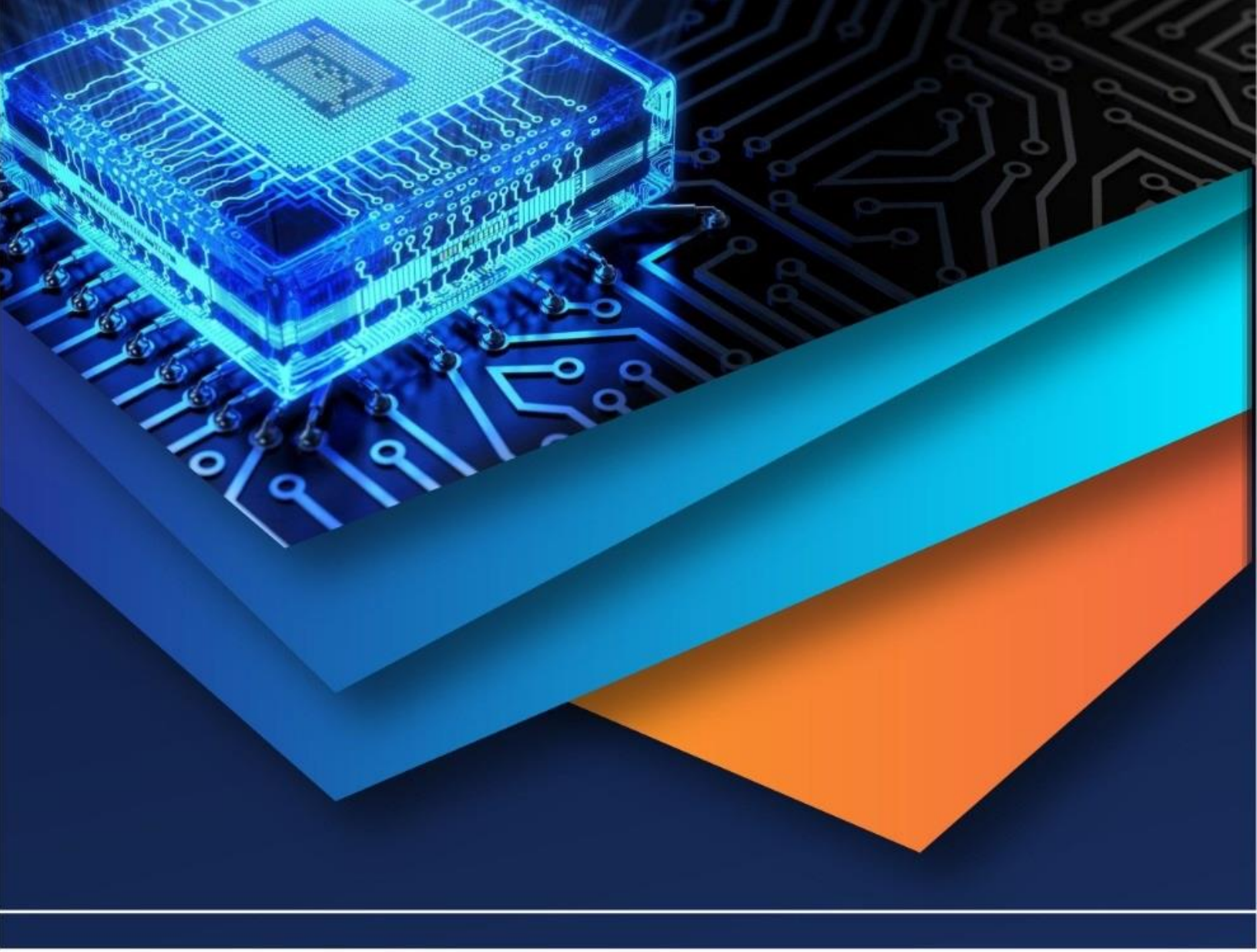
We have entered into a period of Big Data. The amount of information and data that can be extracted from the digital universe is expected to expand exponentially as users come up with new ways to massage and process data. Big data is not structurally and formally well defined. The big data technology is still in the stages of development. The analysis of big data needs to face with many challenges. This paper is a collaborative research effort of examining the traditional view of data analytics and big data analytics, gives brief introduction about new technologies and techniques to handle big data. We have recognized some major challenges concerning big data which big data users specifically face. We must support and encourage fundamental research towards overcoming these technical challenges if we are to achieve the promised benefits of Big Data. My future research will give more emphasis on developing a more complete understanding of challenges associated with big data.

REFERENCES

[1] From Relational Database Management to Big Data: Solutions for Data Migration Testing

Volume 4 Issue VI, June 2016
ISSN: 2321-9653

[2] World's data will grow by 50X in next decade, IDC study predicts http://www.computerworld.com/s/article/9217988/World_s_data_will_grow_by_50X_in_next_decade_IDC_study _predicts
[3] NoSQL Architecture a blog by Kris Zyp http://www.sitepen.com/blog/2010/05/11/nosql-architecture/
[4] From Databases to Big Data by Sam Madden – Article published in IEEE Internet Computing magazine
[5] A Nevins Partners, "Why is BIG Data Important?"White Paper, May 2012.
[6] "Big Data: Volume, Velocity, Variability, Variety", http://nosql.mypopescu.com/post/6361838342/bigdata-volume-velocity-variability-variety, Accessed April 2015.
[7] D. Gosain, "A survey and comparison of relational and non-Relational Databases|. IJERT,Vol 1,Issue 6,2012.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY