International Journal of Engineering & Technology, 7 (2.32) (2018) 181-186

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Survey of Nosql Database For Analyzing Large Volume Of Data In Big Data Platform

R.S.Raghav[1]*, J. Amudhavel[2], P.Dhavachelvan[3]

[1]*Department of CSE, KL University, Andhra Pradesh, India.*
[2]*Department of CSE, KL University, Andhra Pradesh, India.*
[3] *Department of Computer Science, Pondicherry University, Puducherry, India*
**Corresponding author E-mail: vpmrags@gmail.com*

Abstract

The massive improvement of data in the present era is drastically increased, where millions of data is emerging from variety of applications. Due to this massive flow of data the importance of data becomes a key factor in all applications. In old technology the space occupied by data is very less, but in the present scenario the value of every piece of data plays a vital role. The organization needs some new technology to process large amount of data in an effective way. For that purpose data exploration and visualization systems play a vital role in the Big Data era. It is a complex task for the companies to explore and visualize very large datasets. Every company should follow some protocol to have accurate insight from analysis of large volume of data. This strategy helps organizations to enhance their business functionalities and it also helps to identify the way to improve the quality of their products. The big data [25-29] contains some unique databases to handle massive volume of data. In this survey, we explain the characteristics of database and challenges. It also describes about the different techniques and tools currently used for handling large sets of data and their capabilities to support massive volume of data from variety of data sources.

Keywords: *NoSQL Database, Parallel DBMS, Big Data, MongoDB, Column Stores Database, Document Store Databases*

1. Introduction

The role of database for collecting the data from the raw data is a difficult process. To handle these problems the organization went to a technique for storing and processing the data which is known as Parallel DBMS. It is defined as a DBMS which works on multiple locations across multiple processor and disks. The parallel way of processing is done to enhance performance of system. The main goal of having parallel database is to increase the processing speed and to produce high scalability for large number of users.

The Parallel DBMSs works to overtake the problems faced by single process of systems, where they fail to satisfy some criteria like they are not cost-effective scalability, reliability, and performance [1]. These criteria's are satisfied by a powerful and financially attractive parallel DBMS driven by multiple processors. The Parallel DBMSs consists of more links in smaller machines to produce the same throughput as a single, larger machine. This strategy increases scalability and reliability of the system when compared with single-processor DBMSs. The shared resource management is a key source of using the multiple processors in parallel DBMS. The cluster of nodes is also a kind of architecture executed by parallel database system which is known as "shared-nothing nodes"[2]. This strategy is used for improving the high-speed interconnect among the nodes for quick processing. The horizontal partitioning of relational table and the execution of the partition of SQL queries is used by the parallel DBMS.

The concept of horizontal partitioning is the rows distribution in a relational table across the nodes of the cluster is done for executing the parallel processing of data[3]. The multiple CPU is used for the transaction performed by multiple nodes with the help of parallel data manipulation language. The components of the system like memory, disk and processor needs to handle load balancing, by using the parallel database approach thee constraints can be easily managed. Even though the parallel DBMS consist of many positive aspects, they fail to tackle large data sets [1,3]. These large datasets are gathered in unstructured format such as audio, video, images etc. In order to handle these types of data, the organization should move to some other effective technology like NOSQL database.

2. NOSQL

In big data environment the term NoSQL places a major role for handling different variety of data. As we know the four main V's of big data like (Volume, Variety, Velocity and Value) are the backbone for handling large data sets. The characteristics of NOSQL are it has the ability to handle large data sets which are widely spread across the globe with different database technologies [4]. The main goal of using NoSQL database is, the response and performance of the data in Nosql database is very effective when compared to relational databases.

3. Types NoSQL Database

The NoSQL database is classified into different types of database and this is done by the 4V'S of data collected from different sources. Some of the Nosql database is: 1) Column Store Database 2) Document Store Database 3) Graph Store Database 4) Key-Value Store Database 5) Object Store Database 6) Multimodal Store Database 7) Cloud Store Database 8) XML Store Database 9) Multidimensional Store Database 10) Event Streaming Store Database [5,6].

1. Column Store Database

From the name we can understand the concept of using column based DB. It stores huge volume of datasets in columns instead of using rows. The comparison of each column with rows in an RDBMS table and the identification of row are done with the help of key assigning, where the rows can have multiple columns in it [7]. The adding of column can be carried to any row at any time and the column mapping is done for easy storing. Some of the famous column database is Cassandra, Hyper table, Accumulo and HBase etc.

2. Document Store Database

It consists of document which comprises of pair each key with a difficult structure [8]. The document database is used for storing and retrieving the documents, which is stored in different format such as XML, JSON etc. The hierarchical tree data structure is used for storing these documents. It also consist of maps, scalar values and collections, and the process of storing of documents can be similar to each other document [9]. Some of the document databases are Mongo DB, Couch DB, JSON ODM and TerraStore etc.

3. Graph store Database

The idea of using the graph store database is to handle the data in an accurate way by carrying the information using network [9]**.** The nodes are defined as the instance of an object and relations are referred by the edges. Each node has their own relationships and it also allows represents both relationships between the domain entities. The secondary relationships are referred as path, tress and category for carrying partial indexing. There is no limitation for relationships of a node and they can be represented in the same graph database. The graph store Database is Neo4J, Graph Base, Hyper Graph DB and Infinity Graph etc [10].

4. Key-Value Stores

The simplest and easy way of storing data in **N**oSQL databases is key value DB. The key or attribute name for storing each item in the database with the value of the each item [12]. The user can get the information like by knowing the key value and it is mainly used for just storing purpose. The primary key is used for easy accessing of data and also for storing purpose. Some of the familiar key value store database is voldermort, Amazon DynamoDB, Berkeley DB, Riak and Azure Table storage [11, 13].

5. Object Store Database

It also known as objects database management systems; here Objects are used as attributes and methods for data storage. The attributes describes about the characteristics of an object like such as integers, strings, and real numbers or complex object [14]. The functionality of an object is referred to the methods. Some of the object store database are Velocity DB, HSS DB, Magma, EyeDB and object DB etc.

6. Multimodal Store Database

The data model consists of some process like collecting the data from raw data, organizing the data, storing and processing the data [15]. This type of database supports multiple data models against a single and it is integrated according to the requirements. Some of the multimodal store DB is Arango DB, orient DB, wonder DB, Cortex DB and gunDB.

7. Cloud Store Database

The cloud store databases are mostly used in the cloud environment. It provides service to the user by storing the data in their respective database [16]. It has the ability to achieve optimized scaling, providing high availability, multi-tenancy with effective resource allocation. Some of the cloud store database is Oracle Coherence, Hazelcast and Crate data etc.

8. XML Store Database

These database are used for storing XML based data format and the properties of using XML DB is for fast query processing of the data in the xml format [17]. The XML store DB is EMC DocumentxDB, Sedna, BaseX etc.

9. Multidimensional Store Database

It is also known as multidimensional database management system (MDDBMS), it has the ability to execute the data processing in an effective and easy way[18]. Some of the MDDBMS is Dagger DB, Intersystems Cache,Globals and Minim DB.

10. Time Streaming Store Database

The data are dynamic in nature and nature of data is changed within milliseconds. Inorder to process these types of data, the organization need Time streaming DB. It is used for storing and analyzing high-frequency time-series data at scale. The Time streaming DB is AXIbase and Riak TS etc[19].

4. Merits of NOSQL

The NoSQL have more benefits, when compared to relational database. They are highly scalable and it can show better performance with low cost. Some of the benefits of using Nosql database are displayed below

1. Dynamic Design

The NOSQL db does not follow any design. In rdbmsselection of schema is done before inserting the values in it[20]. But in Nosql database there is no need of any schema process. The user can have quick process and retrieving of the data is carried in short span of time.

2. Scalability

They are highly scalable in nature because they don't have any procedure like scale up and scale out for handling more number of loads. The scale up process can be fit for large volume of data, but the Nosql can be built with an ability of extensible at the time of increase in the load [17].

3. Easy Placement of server

The NOSQL is completely suitable for executing unstructured data in an effective way. And it become expensive and the performance will become slow. The horizontal placement of server is used for handling massive volume of data rather than using the server vertically [19]. This strategy used for prohibiting server failure and the replication process is carried in an effective way to stop data failure.

4. Cost Effective

The main benefit of using Nosql is cost effective, where it doesn't need any special requirement of hardware for carrying large data sets[21]. It has the capacity for carrying the large data sets. The client doesn't require any complex process because of it simple nature.

5. Flexibility

The user can create flexible data models in the Nosql database and they don't have any restrictions for creating the schema [22]. The large data sets can be handled in an effective way without any issues.

6. Replication

The process of replication plays a important role in every database, because of processing more volume of data. Inorder to avoid failure the DB should have a backup plan to save and retrieve data from crash or dump in the database [23].Most NoSQL databases also support automatic replication, and there is no need of any special server or hardware for carrying replication process. The above mentioned are some of the important merits of using Nosql database for handling more volume of unstructured data.

5. Demerits of NOSQL

The nature of the Nosql database has more benefits, where the user can use this database for every aspect. But they also face some obstacles, which should be avoided before incorporating the database in the system.

1. Skilled Person

The query processing RDBMS is old and the person in organization have deep understand and knowledge about the rdbms. Even though Nosql is easy for implementation it should require some skilled person to know or to have deep insights about the nature of data, inorder to get the effective output [19].

2. Support

The Nosql DB is an upcoming technology for handling big data sets so the support is less when compared to RDBMS. If there is any presence of problem in RDBMS the vendors can easily get the support from the organization.

3. Displaying of Result

It is a major issue in Nosql because once the data is analyzed the displaying of results should be done in an effective way [24]. The data visualization is a different are in big data environment for displaying the results the company should use tools to analyse the results. This process is difficult, if organization fails to choose a proper data visualization tool. The growth of unstructured data is gradually increased, so the company or organization needs to move the database which can solve variety of problems in an effective way. That is the reason most of the organization is moving for Nosql database in big data environment. The fig 1 describes about the classification of NOSQL database. The Table 1 refers to the comparison of variety of NOSQL database used in the big data environment.

6. Discussion

In this section we discuss which type of big data DB is used by the companies to analyze large volume of data. In fig 2 the availability of the column store No Sql database is shown, where bar chart explains the availability of Database to handle massive volume of data type when compared to other type. The usage of multidimensional store and object store database grab the next high level of availability. To handle massive volume of data, the database should have a nature of fault tolerant. If there is any presence of any fault or error, the database should have a backup strategy to tackle the problem. The fig 3 give an idea about the role of key value store and column store NoSQL db for handling fault tolerant problem and this two NoSQL DB was used more when compared with other DB. The performance is one of the key factors, to know about the strategy of every NoSQL DB to handle high amount of variety of data. In fig 4 the performance of each NoSQL DB was shown and the key value DB has the high performance which is also known for its simplest and easy way of storing data in NoSQL databases other type of NOSQL DB such as object store and column store DB falls in second and third position, this is because most of the data can be analyzed easily and efficient way by using the key vale store DB than others. The scalability plays a vital role in every DB; the role of NOSQL DB is mainly used for handling more number of data and high number of users. In Fig 5 the Scalability of Nosql Db Was Shown and The object store DB is used to have high scalability when compared to other NOSQL DB. Even though other NoSQL DB falls close to the object DB, the usage of more number of users will be efficiently carried by using the object store value DB.

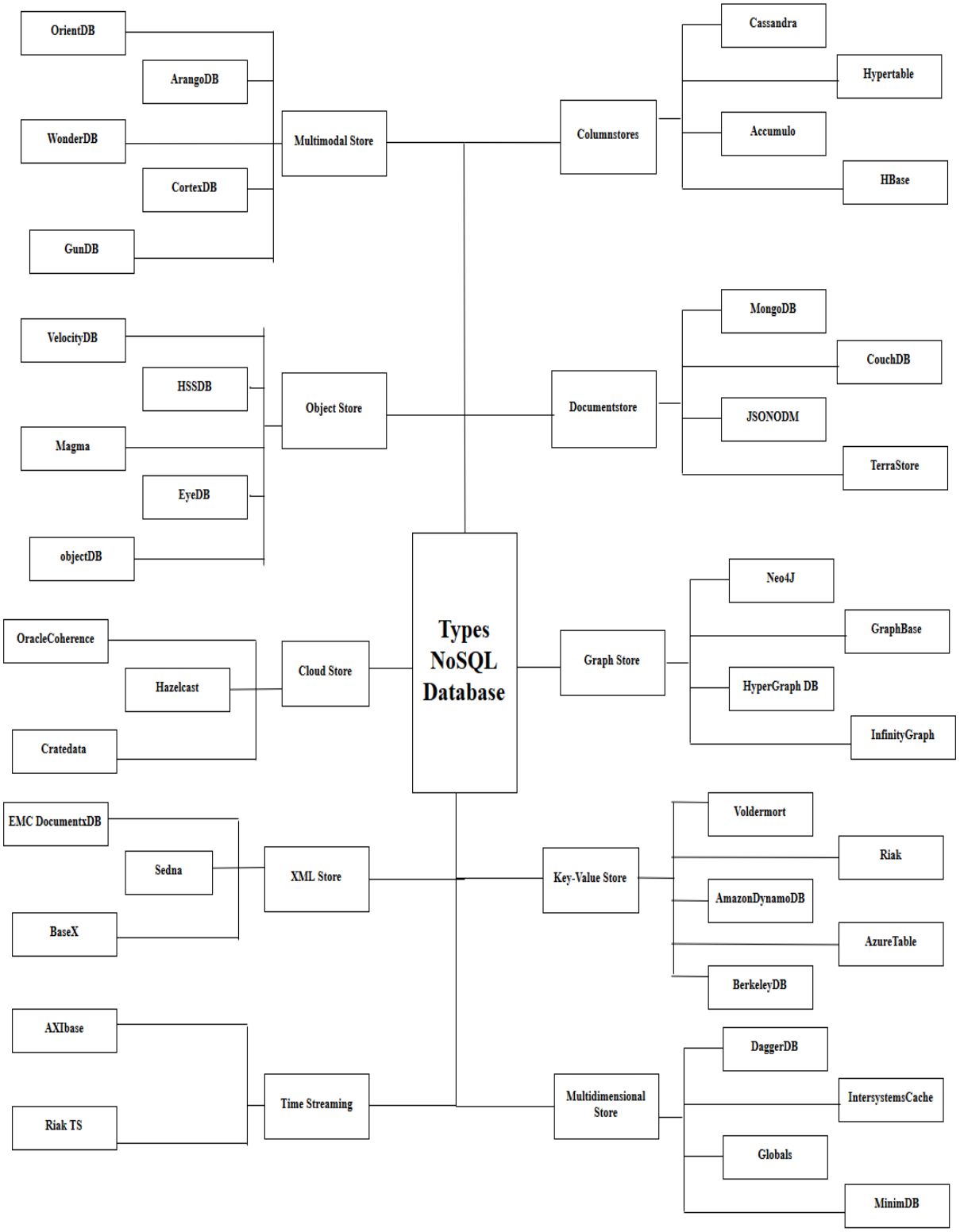

Fig. 1: Classification of NOSQL Database

Table 1: Comparison of NOSQL Database

FEATURES	NOSQL DATABASES															
	OBJECT STORE DATABASE					MULTIMODAL STORE DATABASE					CLOUD STORE DATABASE			XML STORE DATABASE		
	Velocity DB	HSS DB	Magma	Eye DB	Object DB	Arango DB	Orient DB	Wonder DB	Cortex DB	Gun-DB	Oracle Coherence	Hazelcast	Crate data	EMC Doc-umentx DB	Sedna	BaseX
Availability	High	High	High	Avg	High	High	Avg	Low	Avg	High	High	Avg	Low	High	High	Avg
Written In	C#	C#	Magma Programming	C++ and Java.	Java	C/C++/ Javascript	Java	Java	UniPlex	JavaScript	Java,C++	Java	Java	Java	Java, C++	Ja-va,C#,PHP
Query Method	LINQ queries	LINQ Query	Squeak	ODMG OQL	JPA JPQL, JDO JDOQL	AQL	SQL, Gremlin SparQL	Gremlin	SQL, Nosql	JavaScript	Coherence Query Lan-guage	Distributed SQL Query	SQL	XQuery, XPath, XPointer	W3C XQuery	XPath/X Query
Performance	High	High	High	High	High	Avg	Avg	Avg	Avg	High	Avg	Avg	Low	Low	Low	High
Scalability	High	Avg	High	High	High	High	High	High	Avg	High	High	High	Avg	Low	Avg	High
Fault Tolerant	Avg	Avg	High	Avg	High	Avg	Avg	High	Low	High	Avg	Avg	Low	Low	Low	Avg

Table2: Comparison of NOSQL Database

FEATURES	NOSQL DATABASES																
	COLUMN STORES DATABASE				DOCUMENT STORE DATABASES				GRAPH STORES DATABASE				KEY-VALUE STORES				
	Cassandra	Hyper table	Accumulo	H Base	Mongo DB	Couch DB	JSON ODM	Terra Store	Neo4J	Graph Base	Hyper Graph	Infinity Graph	Volder-mort	Amazon Dyna-moDB	Berkeley DB	Riak	Azure Table
Availability	High	High	High	High	High	Avg	Low	Avg	High	Avg	Low	Avg	High	High	Avg	High	High
Written In	Java	C++	Java	Java	C++	Erlang	Ja-vaS-cript	Java	Java	Java	Java	Java (Core C++)	Java	Java, .NET	Java	Erlang	C#
Query Method	CQL and Thrift	HQL, native Thrift API	Accumulo Shell	MapR educe Java	dynamic object-based lan-guage & MapRe-duce	MapRe-duceR of JavaS-cript Funcs	grem lin	Range queries, Predi-cate	SparQL, na-tiveJa-vaAPI, JRuby	SPARQL, Gremlin	Java	Graph Nav-igation API, Predicate Language Qualifica-tion	BDB-JE, MySQL	Java, .NET	No query method	MapRe duce term match-ing	LINQ syntax
Performance	High	High	Avg	High	High	Avg	Avg	Avg	High	Avg	Low	High	High	High	High	High	Avg
Scalability	High	High	High	High	High	High	Low	Avg	High	High	Avg	Avg	High	High	High	High	Avg
Fault Tolerant	High	Avg	High	High	High	Avg	Low	Low	High	Avg	Low	Low	High	High	Avg	High	Avg

Table3: Comparison of NOSQL Database

FEATURES	Dagger DB	Intersystem Cache	Globals	Minim DB	AXIbase	Riak TS
Availability	High	High	High	Avg	High	AvAvg
Written In	C#	Java	Java / .NET	Perl, .NET, ActiveX, Java	R, Java, Ruby, Python	Java
Query Method	LINQ queries	SQL capable JDBC, ODBC	Squeak	MUMPS-based query	SQL	SQL, Gremlin, SparQL
Performance	High	High	Avg	High	Avg	High
Scalability	High	Avg	High	High	High	High
Fault Tolerant	Avg	Avg	Avg	Avg	Avg	Avg

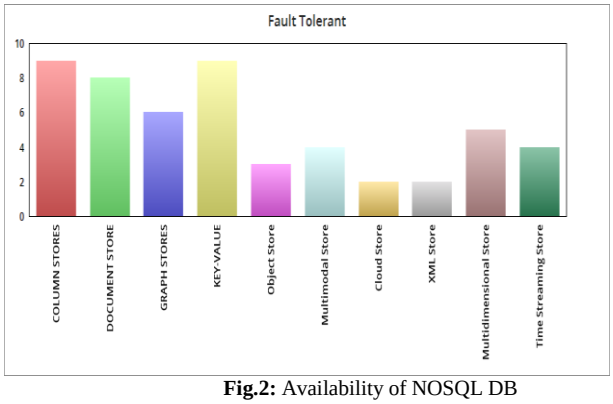

Fig.2: Availability of NOSQL DB

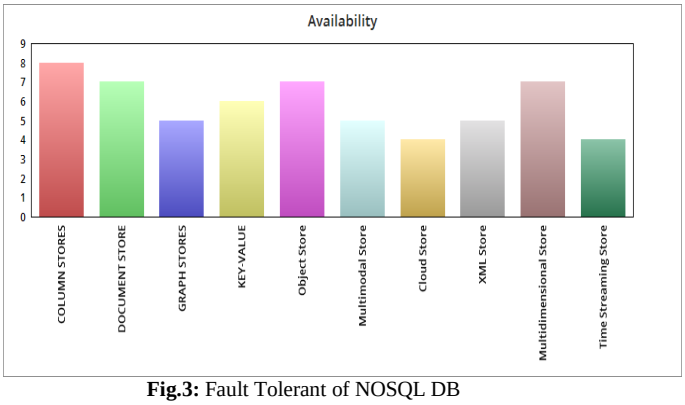

Fig.3: Fault Tolerant of NOSQL DB

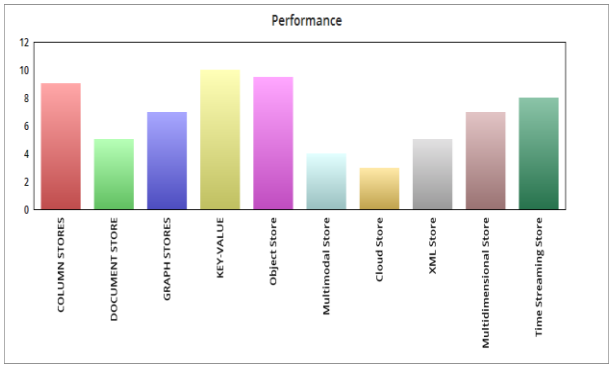

Fig.4: Performance of NOSQL DB

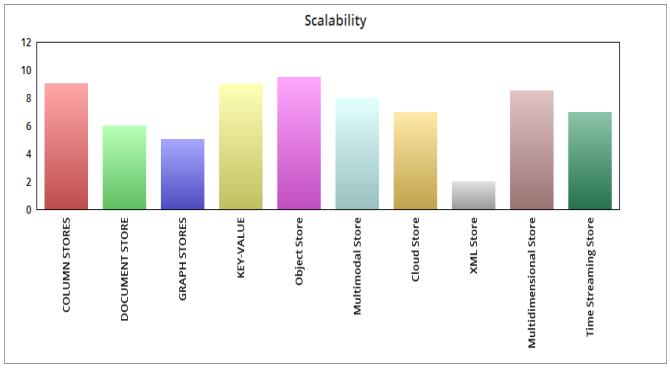

Fig.5: Scalability of NOSQL DB

7. Conclusion

The value of data in present world is high, where a normal database couldn't have the capacity to handle massive volume of data. The handling of data and analyzing the large volume of data need to carry pre process like proper extraction of data from variety of data sources. They should know the 3v's of data such as volume, variety, velocity and value etc. According to it, the company should select the proper database, process, scripting language and last the proper data visualization tool for carrying effective analyses. In this paper we discussed about the variety of database and its features. These strategies help the business people to know the value of each data and how to process the data and analyze it and how to improve their business value. By selecting a perfect database the company can gain deep insights about the data and quick decisions on the advantageous business opportunities can be done effectively.

References

[1] Poonam S. Patil, Rajesh N.Phursule "Survey Paper on Big Data Processing and Hadoop Components" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064

[2] Salim Raza Qureshit, AnkurGupta "Towards Efficient Big Data and Data Analytics: A Review"

[3] Franks, B., Analytics on Web Data: The Original Big Data, in Enterprise Analytics, T.H. Davenport, Editor. 2013, Pearson Education, Inc.: Upper Saddle River, New Jersey.

[4] D. Jiang, B. C. Ooi, L. Shi, and S. Wu, "The performance of mapreduce: An in-depth study," in Proceedings of the 36th International Conference on Very Large Data Bases (VLDB), vol. 3,no. 1, 2010, pp. 472-483. S

[5] 5R. Chen, H. Chen, and B. Zang, "Tiled-mapreduce: optimizing resource usages of data-parallel applications on multicore with tiling," in Proceedings of the 19th international conference on Parallel architectures and compilation techniques, ser. PACT '10. New York, NY, USA: ACM, 2010, pp. 523-534

[6] Hao Zhang, Gang Chen, Kian-Lee Tan "In-Memory Big Data Management and Processing: A Survey"Ieee Transactions On Knowledge And Data Engineering, Vol. 27, No. 7, July 2015.

[7] N. Udipi, N. Muralimanohar, N. Chatterjee, R.Balasubramonian, A.Davis ,and N.P.Jouppi,"Rethinking DRAM design and organization for energy-constrained multi-cores," in Proc. 7th Annu. Int. Symp. Comput. Archit., 2010, pp. 175–186.

[8] R. Stoica and A. Ailamaki, "Enabling efficient os paging for main-memory oltp databases," in Proc. 9th Int. Workshop Data Manag. New Hardware, 2013, pp. 7:1–7:7.

Pavlo, C. Curino, and S. Zdonik, "Skew-aware automatic database partitioning in shared-nothing, parallel OLTP systems," in Proc. ACM SIGMOD Int. Conf. Manag. Data, 2012, pp. 61–72.

[9] V. Raman, G. Swart, L. Qiao, F. Reiss, V. Dialani, D. Kossmann, I. Narang, and R. Sidle, "Constant-time query processing," in Proc. IEEE 24th Int. Conf. Data Eng., 2008, pp. 60–69.

[10] H. Zhang, B. M. Tudor, G. Chen, and B. C. Ooi, "Efficient inmemory data management: An analysis," in Proc. Int. Conf. Very Large Data Bases, 2014, pp. 833–836.

[11] T. Lahiri, M.-A. Neimat, and S. Folkman, "Oracle timesten: An inmemory database for enterprise applications," IEEE Data Eng. Bull., vol. 36, no. 2, pp. 6–13, Jun. 2013.

[12] Anita Brigit MathewS. D. Madhu Kumar, "Analysis of Data Management and Query Handling in Social Networks using NoSQL Databases" 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCl).

[13] W. Kim, "Web data stores (aka NoSQL databases): a data model and data management perspective,"z T vol. 10, no. 1, pp. 100-110, 2014.

[14] Raghav, R. S., J. Amudhavel, and P. Dhavachelvan. "a survey on tools used in big data Platform." advances and applications in mathematical sciences 17, no. 1 (2017): 213-229.

[15] R. Paivarinta, R. and Yrjo, "Performance evaluation of nosql cloud database in a telecom environment," 2011.

[16] J. Kuhlenkamp, M. Klems, and O. Ross, "Benchmarking scalability and elasticity of distributed database systems," Proc. VLDB Endow.

[17] Min-Gyue Jung, Seon-A Youn, Jayon Bae, Yong-Lak Choi"A Study on Data Input and Output Performance Comparison of MongoDB and PostgreSQL in the Big Data Environment" 2015 8th International Conference on Database Theory and Application.

[18] Prathibha.P.G, Dileesh.E.D "Design of a Hybrid Intrusion Detection System using Snort and Hadoop" International Journal of Computer Applications (0975 – 8887), Vol. 73– No.10, July 2013.

[19] Min Chen · Shiwen Mao · Yunhao Liu "Big Data: A Survey" Springer Science, Business Media New York, Vol. 08, PP. 171–209, 22 January 2014.

[20] Amudhavel, J., D. Sathian, R. S. Raghav, Dhanawada Nirmala Rao, P. Dhavachelvan, and K. Prem Kumar. "Big Data Scalability, Methods and its Implications: A Survey of Current Practice." In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), p. 56. ACM, 2015.

[21] Pal SK, Talwar V, MitraP (2002) "Web mining in soft com puting framework, relevance, stateofthe art and futuredirec tions". IEEE Transac Neural Netw 13(5):1163–1177.

[22] Patterson S, Elmore AJ, Nawab Fetal."Serializability, not serial: Concurrency control and Zavailability in multi data center data stores". In Proc. the38thInternational Conference on Very Large Data Bases, July 2012, pp 1459–1470.

[23] Cipar J, Ganger G, Keeton Ketal. Lazy Base: "trading freshness for performance in a scalable database". In Proc. the 7thACM European Conference on Computer Systems, April 2012, pp 169-182.

[24] Raghav, R. S., Sujatha Pothula, T. Vengattaraman, and Dhavachelvan Ponnurangam. "A survey of data visualization tools for analyzing large volume of data in big data platform." In Communication and Electronics Systems (ICCES), International Conference on, pp. 1-6. IEEE, 2016.

[25] Padmapriya, V., Gowri, V., Lakshmipriya, K., PremKumar, K., Thiyagarajan, B., "Perspectives, motivations and implications of big data analytics", (2015) ACM International Conference Proceeding Series, 06-07-March-2015, art. no. 2743099,

[26] Rao, D.N., Sathian, D., Dhavachelvan, P., Raghav, R.S., Prem Kumar, K., "Big data scalability, methods and its implications: A survey of current practice", (2015) ACM International Conference Proceeding Series, 06-07-March-2015, art. no. 2743121,

[27] Karthikeyan, P., Sathian, D., Raghav, R.S., Abraham, A., Dhavachelvan, P., "A comprehensive survey on variants and its extensions of BIG DATA in cloud environment", (2015) ACM International Conference Proceeding Series, 06-07-March-2015, art. no. 2743097,

[28] Padmapriya, V., Gowri, V., Lakshmipriya, K., Vinothini, S., PremKumar, K., "Demystifying challenges, opportunities and issues of Big data frameworks", (2015) ACM International Conference Proceeding Series, 06-07-March-2015, art. no. 2743110,

[29] Bandi, R., Gouse, S., "A comparative analysis for big data challenges and big data issues using information security encryption techniques1, 2", (2017) International Journal of Pure and Applied Mathematics, 115 (8 Special Issue), pp. 245-251.