Emerging No-SQL Technologies for Big Data Processing

Amrinder Kaur
Research Scholar, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra-136119
Email: er.amrinder09@gmail.com
Dr. Rakesh Kumar
Professor, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra-136119
Email: rakeshkumar@kuk.ac.in

----------**ABSTRACT**----------

In modern era everyone is connected with internet and with the usage of information technology, IT tools, data is increasing on exponential rate. The generated data may be structured, semi structured or unstructured in nature, called big data. This demands new techniques and technologies with extensive processing requirement and storage requirement for this voluminous data. Existing technologies and computation facilities are facing challenges in meeting the scale and performance of such a vast data. To scale with big data, organizations are opting diversified solutions like NO-SQL which is proving to be emerging alternative for big data and other fields. This paper discusses big data, no-sql databases, its classifications and comparison of various no-sql technologies.

Keywords – **Big Data, Document Store, Graph Store, key-Value store, NO-SQL Databases, Wide Column Store**

1. Introduction

Under the ambit of big data, the large, unstructured, semi structured or heterogeneous data has gained attention from last few years. Uncontrolled use of social sites like facebook, twitter, linkedin etc is responsible for such a volume of data. According to definition of big data, it consists of large volume (volume) of data with different varieties (variety) which is generating with high velocity (velocity) as shown in fig. 1. These characteristics of big data imposing research challenges.

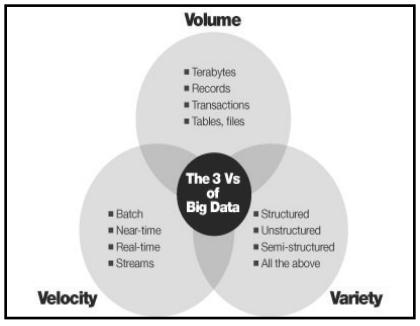

Figure 1: 3 V's of Big Data

As data size increases from terabytes to petabytes but management techniques for such volume of data are not evolving at such a fast pace. Existing technologies like cloud computing, heterogeneous computing etc. could not bear the stress of large volume of data travelling between the computing nodes.

Due to data proliferation challenges, collaborative actions for new technologies are required to handle such a volume of data. Various data outlets are extreme growth in internet users, videos, images, climatic information, social media, sensors etc. This required efficient queries which retrieve the precise information from the universe of data. According to IBM growth of data is about 2.5 quintillion bytes every day [1].

2. NoSQL Database

In 1998, NoSQL was coined by Carlo Strozzi. NOSQL refers to "Not only SQL" is open source, non relational data management system and it is different from traditional relational database in some manner like NOSQL databases are not built upon tables and do not use sql for querying the database and for manipulating the database [2]. These databases are designed for large scale of data which needs distributed data stores and it has fault tolerant architecture. NoSQL database do not use any fixed schema for data storing and it avoid join operations because joins require strong consistency and fixed schemas. For scalability it uses horizontal scaling to clusters of machines. Relational databases are unable to handle such a large volume of data because of its fixed schema and structured data. Relational databases are not able to work with semi structured and unstructured format like audio, videos, emails etc. Relational databases are designed to work upon steady data it is not capable to deal with high velocity of data. All these inabilities lead to emergence of new technologies like NoSQL [3]. Comparison between RDBMS and NoSQL databases is shown in table below [4].

Comparison of RDBMS & NoSQL Database

RDBMS	NOSQL Database
It has Structured & organized data	It has semi structured & unorganized data
It uses Structured Query Language	It uses No Declarative Query Language
It has Predefined Schema	It don't have predefined schema
Consistency is tight	Consistency is Eventual
ACID Transaction	BASE Transaction & CAP Theorem
Data and relationship between data is stored in tables	Data is stored as Key-Value pair storage, Column Store, Document Store or in Graph databases

3. Classification of NoSQL Databases

In this section classification of data models are discussed which offloads the nosql data stress. An ideal nosql model should have following attributes like high availability, high scalability, concurrency, low latency and reduced operational cost. Nosql databases are classified into four categories that is document stored, wide column stored, key-value stored and graph oriented. All these database types are discussed in following subsection:

3.1. Document Oriented Stored

The concept behind document oriented store is to organize and store the data in the form of documents. In this model data is encoded or formatted in XML, JavaScript Option Notation (JSON) and Binary JSON (BSON) etc. It does not enforced to follow any schema due to which it is flexible and easy to change. Each document in this model is assigned a unique key that uniquely identify the document in the store. Documents are organized into numerous ways like collections, tags and directory hierarchies in

order to group diverse kinds of data. MongoDB, CouchDB, couchbase, RavenDB, Cloudant etc. are important data models of this category. Example is shown in fig 2 [4], [5], [6], [7].

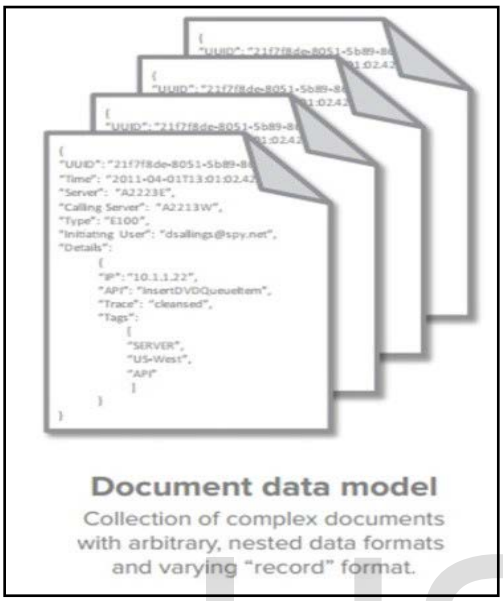

Figure 2: Document Oriented Store

3.2. Wide Column Store

Wide column store are schema less data model and it is specifically works on column. Value of a column is stored contiguously in memory and every column is treated individually. Each column is identified uniquely by a key. This key may be a string or a number. In this store a record can contain billions of columns due to its dynamic nature. It is useful in distributed data storage, large scale batch oriented data processing and in predictive and exploratory analytics. HBASE, Cassandra, bigtable etc. are popular data models. Example is shown in fig 3. [4], [5], [6], [7].

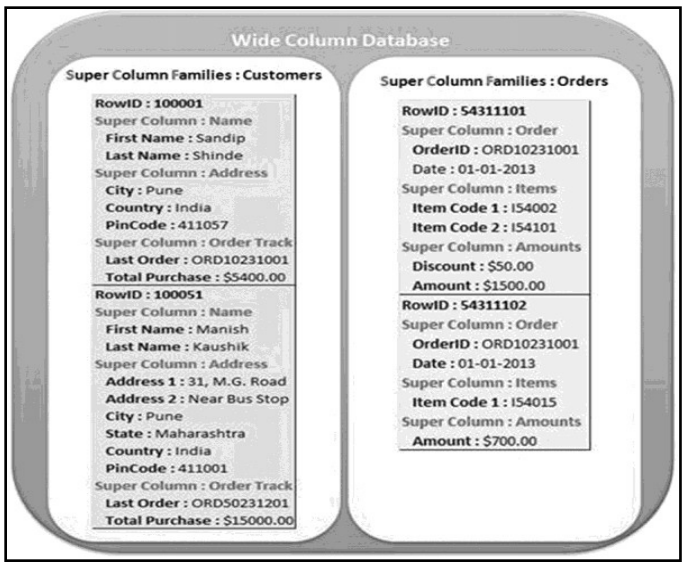

Figure 3: Wide Column Store

3.3. Key-Value Store

The most basic nosql data model is key-value store. It can store any kind of data and data is stored as a collection of key-value pairs where key uniquely identifies the data in the collection. In this model data is searched on the basis of keys not on the basis of data, which limit the search to exact number of matches. It is suitable for fast retrieval of values. Dynamo, Redis, Riak, Orcale NoSQL etc. is popular example of this data model. Example is shown in fig. 4 [4], [5], [6], [7].

Car	
Key	Attributes
1	Make: Nissan Model: Pathfinder Color: Green Year: 2003
2	Make: Nissan Model: Pathfinder Color: Blue Color: Green Year:2005 Transmission: Auto

Figure 4: Key-Value Store

3.4. Graph Oriented Model

In graph oriented model data is represented in the forms of nodes (conceptual objects), edges (node relationships) and properties (attributes as

key-value pair). This is the only data model that provides visual representation of information and due to this visual representation it is human friendly as compared to other nosql data models. It is useful in representing relationship between data. Neo4j, OrientDB, InfoGrid, AllegroGraph are popular graph data model. Example is shown in fig 5. [4], [5], [6], [7].

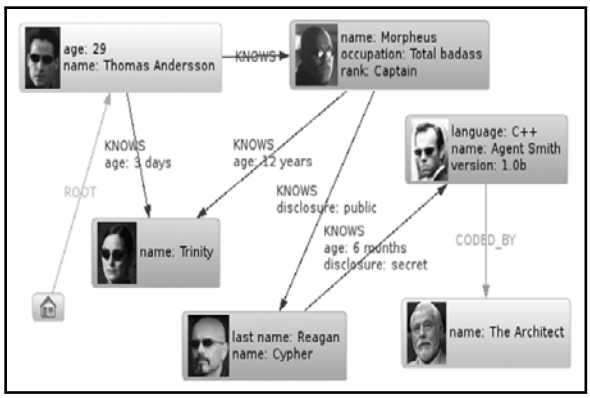

Figure 5: Graph Oriented Model

4. Comparison of NoSQL Databases

Following table will compare the various nosql data models. [5], [6], [8]-[11]

Attributes	NOSQL DATABASES							
	Document Stored		Wide Column Store		Key-Value Store		Graph Oriented Model	
Features	Mongo DB	Couch DB	HBASE	Cassandra	Riak	Oracle NoSQL	Neo4j	Orient DB
Developer	10gen	IBM	Microsoft	Facebook Inc.	Basho Technologies	Oracle Corporation	Neo Technology	Orient Technology
Written in	C++	Erlang	Java	Java	Erlang	Java	Java	Java
Query Language	MongoDB adhoc query Language	Javascript	Pig latin, HQL	CQL	RESTful API	Key access methods	Cypher	SQL
SQL Nature	No	No	No / Yes	Yes	No	No	No	Yes
Data Processing Nature	Batch Processing & Event Streaming	Batch Processing	Batch Processing	Streaming & Atomic Batches	Batch Processing	Batch Processing & Streaming	Batch Processing	Batch Processing
Open Source	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Horizontal Scalable	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Replication Mode	Master-Slave Replication	Master-Slave Replication	Master-Slave Replication	Master-Slave Replication	Multi-master Replication	Multi-master Replication	-	Multi-master Replication
Sharding	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Operating System	Cross platform	Ubuntu, Red Hat Windows Mac OS X	Cross platform	Cross platform	Cross platform	Cross platform	Cross platform	JVM Compatible
High Availability	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
High Scalability	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Relational Nature	No	No	No	Yes	No	No	No	No

5. Conclusion

NoSQL is emerging field which act as a substitute for traditional relational models. These relational models are incapable in handling the big data. Data proliferation by different resources like sensors, internet etc. has rise the demand of robust data processing, high availability and high scalability over the servers. In this paper, fine classification of NoSQL data model and their granular comparison on the basis of few attributes like design, integrity, system etc. were discussed.

References:

[1] What is Big Data? Bringing Big Data to the enterprise,http://www-01.ibm.com/software/data/bigdata/

[2]http://en.wikipedia.org/wiki/NoSQL

[3]Market Realist, RDBMS for data storage‖, available at : http://marketrealist.com/2014/07/traditional-databasesystems- fail-support-big-data/

[4]Aggarwal D. etal., *"Emerging Technologies For Big Data Processing: NOSQL And NEWSQL Data Stores"*, IJECS, Vol 5, Issue 1, Page No. 15598-15604, ISSN: 2319-7242, January 2016

[5] Moniruzzaman AB etal. ," *NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison"*, International Journal of Database Theory and Application Vol. 6, No. 4. 2013

[6] Sharma S., " *An Extended Classification and Comparison of NoSQL Big Data Models"*, Center for Survey Statistics and Methodology, Iowa State University, Dated- 20 April 2016

[7] Sharma V. etal. ," *SQL and NoSQL Databases"*, IJARCSSE, Volume 2, Issue 8, ISSN: 2277 128X, August 2012

[8]http://vschart.com/compare/oracle-nosql-database, Dated 20 April 2016

[9] http://vschart.com/compare/orientdb, Dated 20 April 2016

[10] http://vschart.com/compare/cassandra, Dated 20 April 2016

[11] http://vschart.com/compare/couchdb , Dated 20 April 2016