Webology, Volume 18, Number 4, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I4/107

Heterogeneity of Data: The NoSQL Revolution

Vishan Kumar Gupta[1], Anupriya[2]

[1]Department of Computer Science and Engineering Graphic Era Deemed to be University Dehradun, India
[2]School of Computing Graphic Era Hill University Dehradun, India

ABSTRACT

There is a tremendous growth in the unstructured Data in these past years which has led to the emergence of NOSQL. Several companies and IT firms have unleashed a variety of solutions to support the unstructured & structured data. In this paper we provide an overview of latest trends in SQL, NOSQL & Big-data and we extend the discussion to the importance of suitable Data architecture & key features of each database in depth.

Keywords: NoSQL, SQL, Bigdata, Hadoop, Hive, HBase, MapReduce Relational Database Management Systems (RDBMS).

INTRODUCTION

Prior to 2005 mainly all the essential data was stored on large Data warehouses which were managed by relational database management systems. The most common way to query these RDBMS is using the query language called SQL (Structured Query Language). Huge amounts of data are created daily via web & business applications & large sections of this data is managed by Relational Database systems such as MySQL, Postgre SQL and many more. Relational Databases are generally suited to client server programming & today it is widely used for storing & managing the structured data. Today the IT Companies are hiring more and more professionals to manage RDBMS databases.

The growing market of web applications & accessibility to internet has led to the flood of unstructured data on web. Social media has played a major role in production of unstructured data. According to web statistics report of 2014, there were about 3 billion people who were connected to the world-wide-web and the amount of time they spent online was roughly 35 billion hours per month. And since more than 1/5th of the population is following such behavioral patterns, it is evident that data fetching & storing data like audio, video, images and textual data becomes more and more important. Supporting large number of concurrent users along with dynamic patterns of Data usage has acted as the driving force for the need of new database system. This exponential growth in unstructured data that the world has seen in recent years is one of the main reasons relational databases can no longer meet the company’s needs. The term "NoSQL" was first used in the late 90's but it got its real meaning in 2009. Originally, it was an open-source database which was developed by Carlo Strozzi, all the data was stored as ASCII files shell scripts were used instead of SQL to access data.

For Organizations of every size, managing data has become a critical factor that can determine market leaders. Top IT companies and government bodies are defining new initiatives and reevaluating existing strategies to examine how they can benefit from Big Data. The most widely used Big Data technology is Apache Hadoop. It must not be confused with a type of database, but rather it is a software ecosystem supporting massive parallel computing. It is a framework for processing of various types of NoSQL distributed databases which allows data to be stored in multiple servers. MapReduce is a computational model of Hadoop framework that takes exhaustive data processes and distribute the computation across multiple servers (generally referred to as Hadoop cluster). With MapReduce, a large data procedure which takes only 10 minutes to process when distributed across large Hadoop cluster commodity servers all processing in parallel, might take 24 hours of processing time on a centralized relational database management system.

Big Data can handle any size of data which distinguishes it from the traditional Relational Database System like Sql. Big Data Technologies uses Commodity Hardware for implementation. Other important features of Big Data technologies include Scalability, Distributed Architecture & high availability. Along with more opportunities and added value, Big Data has also brought some challenges. Data Redundancy, Maturity and Durability are some of them.

DATA ARCHITECTURE

Data architecture involves models, policies, standards to govern what data is to be collected, how it is stored, arranged & used in data systems and in organizations. Since Business grows there is always a need to upgrade both hardware and software of a company's infrastructure. It is evident that for any company data is one of the crucial factor which decide its survival and performance in the industry. Thus, having right data architecture is most important. There is no standard solution that fits to all sizes and kinds of IT companies, but there are ways with which an IT manager can determine the right solution. The type of data, volume of data, rate of data generation (Volume, Variety, Velocity), does the data needs to be distributed are some of the factors which plays an important role while deciding a data architecture for an IT firm. Traditionally from last 25 years the data in an IT firm is managed by RDBMS. Relational databases were originally designed to overcome the shortcomings of File system of data management. Later, with the growth of web applications the relational databases were used for storing all the essential data of a company like managing client records and handling data request of client-server-based applications. With the emergence of unstructured data and inability of RDBMS to model the data requirements, the need to model a new database architecture has risen. The NOSQL database systems were developed to manage the limitations of relational databases. NoSQL database engines focuses on performance and Scalability. The main advantage of NOSQL architecture is that the logic can now be code in your own familiar, flexible programming language. Basically, it's all about solving "business" problems. The only thing do is to find the right solution that solves the problem.

Understanding the Data sources and data growth along with the paradigms of Data consumption with respect to growth of company determines the model of data architecture. While designing the data architecture one must take into consideration the Cost of evolving data architecture as business grows. Capacity planning is also an important feature of a good architecture [2].

SQL and NOSQL

Traditionally IT world has been dependent on Relational Database management systems for handling the storage requirements. Huge amounts of data created by web applications are handled by relational databases. Relational databases are well suited for client-server-based applications and today it is predominant technology for handling structured data. Classical relational databases follow the ACID property. That is, all the transactions should be Atomic, Consistent, Isolated and Durable.

- Atomic: Atomicity means "all or nothing". If any part of transaction fails, whole transaction fails and no changes must reflect in database.
- Consistency: It means that database must remain consistent after and before any transaction takes place.
- Isolation: Changes made by a transaction must not hinder with changes made by another transaction.
- Durability: It ensures that after a transaction is successfully completed, it will persist.

Along with these properties, RDBMS have other properties which make it popular. Such as:
Data is organized into two-dimensional tables and relationships are represented by data. Each relational table is a two-dimensional array in which:

- Each element of table of table is data element;
- All the cells of a column are homogenous;
- Normalization is used to reduce redundancy of data;
- They allow greater flexibility and efficiency [7].

SHORTCOMINGS OF SQL

RDBMS can be reliably & efficiently used to store & manipulate the structured data but in today's world the variety and velocity of data is growing exponentially. As in areas like social media, data has no specific structure definitions. All this makes the need to manage unstructured data which is schema-less and non-relational in nature unavoidable. It becomes a challenge for RDBMS to provide cost effective and fast create, read, update, delete (CRUD) operations as it has to deal with maintaining relationship amongst data and overhead of joins. Therefore, to deal with such data a new mechanism is required which could handle data in efficient way.

NoSQL is not a campaign against SQL but it stands for "Not Only SQL". NoSQL can be defined as a collection of various non-relational technologies. NoSQL belongs to broad class of non-relational databases systems that differ from RDBMS in various aspect such as they don't use SQL as query language but provides access by using Application Programming Interfaces. NoSQL has certain properties over traditional RDBMS such as:

- High scalability
- Commodity hardware

- Lower cost
- Distributed Computing;
- Schema Flexibility
- Support semi/unstructured data
- No complex relationships

In contrary to RDBMS NoSQL follows "BASE" systems. Eric Brewer gave this acronym and also formulated CAP theorem whose properties are used by BASE system. According to CAP theorem, a distributed system cannot guarantee all three properties at the same time:

- Consistency- data written once will correspond to all future read requests
- Availability- ensures that database is always accessible
- Partition Tolerance- failure of one part of database will not affect other parts

Brewer originally proposed that only two out of three CAP properties are possible at a time which results into three design options CA, CP and AP which can be described as:

- CA – data must be consistent to all nodes and unless nodes are offline, users can access any node and be sure that data is same on all nodes.
- AP – nodes will always remain online even if communication between nodes is temporarily down and data will be re-sync after resolving partition, but the user is not guaranteed that all nodes have same data.
- CP – consistency of data between all nodes and partition tolerance is maintained by becoming unavailable whenever a node goes down.

BASE system lacks consistency to support greater availability and partition tolerance. BASE system can be defined as:

- Basically, Available which means system guarantees availability
- Soft state indicates that system may change its state over time due to its eventual consistent model.
- Eventual Consistency refers to ability of system to become consistent over the time but system must not receive any input at that time.

TYPES OF NOSQL

There are mainly four types of NoSQL databases available:

- Key- value Databases: These type of data stores might be the most popular technology in NoSQL. A hash table is used which contains a unique key and a pointer to the particular item of data. These databases follow eventually consistent principle. It is the simplest and easiest to implement database. Apache Cassandra is the most famous key-value database which was developed by Facebook. Apart from these other examples include Amazon's DynamoDB, cloudant and Oracle's BDB.

- Column Oriented: These databases were originally created for storing and processing of large amounts of data which is distributed over multiple machines. These databases also contain keys but they point to multiple columns. Columns are arranged according to column family. Examples include Hbase, etc.

- Document Databases: These databases comprise of wide array of binary encodings and formats. Standard formats like JSON and XML stores semi-structured data and combines with binaries like Microsoft's Word and Adobe's PDF. These databases are the next level of key-value, which allows nested values associated with each key. Key lookup, use of metadata and tagging is used for successful querying of document database. Examples include Apache CouchDB, MongoDB uses JSON markup for querying purposes.

- Graph Databases: This type of database uses flexible graph model instead of tables and rows or rigid structure of SQL which can be further scalable across multiple machines. These databases use matrix view of data. Social networking is one of its applications [5][6].

SHORTCOMINGS OF SQL

In a recent interview to IT Trends and Analysis Dave Rosenthal, Cofounder of FoundationDB said that the lines between SQL and NoSQL will tend to blur this year. IBM's decision of buying Cloudant, a NoSQL cloud Database startup indicates blurring of lines has already started. According to 2014 state of database Technology survey, the most popular database solutions include relational databases (RDBMS) from Oracle (46%), Microsoft's SQL Server (34%), and IBM's DB2(46%) dominates the market with MongoDB is just 5% and SAP HANA being 3%. MongoDB was the only NoSQL vendor to reach top 10 databases securing 10th place and rest all are relational databases. The growth of NoSQL market is greatly influenced by the increase in adoption of big data solutions but the lack of awareness about potentials of NoSQL and lack of basic infrastructure facilities to support NoSQL are major challenges NoSQL has to overcome. All these forecasts are drawing a lot of interests and money for NoSQL vendors. MongoDB received US$ 150 million with DataStax, FoundationDB and CouchDB earning revenues about $45 million, $17 million and $25 million respectively in 2014 while SQL is growing by around 8% said by Nick Heudecker, research director at Gartner. In November 2014, Rosenthal from FoundationDB said that company is focusing in a database technology having ACID properties of SQL and distributed architecture of NoSQL. In 2013, the company acquired Akiban and it is going to launch its SQL Database engine which runs on top of FoundationDB. A true ANSI SQL database running as a module on top of NoSQL engine and supporting data models like graph or document. Rosenthal said that there are over 150 NoSQL systems in market and it's common for developers to have 7, 8, 11 different databases which they have to adapt. The question arises "How many to adopt?" certainly not one but also not all of them. It seems pretty clear that FoundationDB will lead the charge to bridge the gap between NoSQL and SQL [4].

CONCLUSION

NoSQL might be gaining popularity with enormous speed but this does not mean that relational databases will come to an end. Although, Monopoly of relational databases has come to an end. It is an era of polyglot persistence, which means an era of using different needs of different data warehouse. There is no denying in the impact made by non-relational databases in industry but it also needs improvement. Since, Necessity is the mother of invention and in case of NoSQL, technologies like Big Data, distributed computing and growth of social networking an1d other technologies contained within the NoSQL movement are standing at the pinnacle of what databases can accomplish today.

REFERENCES

1. Internet live stats, [online] available, http://www.internetlivestats.com/internet-users.
2. Xiaoquan Li; Fujiang Zhang; Yongliang Wang. 2013, Research on Big Data Architecture, Key Technologies and Its Measures, Dependable, Autonomic and Secure Computing (DASC), IEEE 11th International Conference.
3. Alexey Vasiliev. 2013. World of NoSQL databases, [online] available, http://leopard.in.ua/2013/11/08/nosql-world/.
4. Steve Wexler. 2014. Blurring the lines between SQL and NoSQL, [online] available, http://it-tna.com/2014/03/17/blurring-lines-sql-nosql-databases/.
5. Paul Williams, 2012. The NoSQL Movement-What is it? [Online] available, http://www.dataversity.net/the-nosql-movement-what-is-it/.
6. Bill Vorhies. 2013. A Brief History of Big data Technologies – From SQL to NoSQL to Hadoop and Beyond, [online] available. http://data-magnum.com/a-brief-history-of-big-data-technologies-from-sql-to-nosql-to-hadoop-and-beyond/.
7. Martin Fowler. 2012. Introduction to NoSQL. GOTO Aarhus Conference.