A Survey on Data Mining and Analysis in Hadoop and Choosing NoSql database using CAP Theorm

Pediredla Praveen Kumar¹ Nambaru Lokeswari² Kamakula Kamaraju³ Talluri Haribabu⁴
Department of computer science and Technology
Anil neerukonda institute of technology and sciences,visakhapatnam,andhra pradesh,india.
{praveen1271kumar,lokeswari538,kkraj7121,talluri.haribabu}@gmail.com

Abstract - Data mining is a process which plays a vital role in the process of knowledge discovery from various types of sources like web,datamarts and warehouse. This process contains several steps like data data integration,data selection,data cleaning ,data transformation. In this work we are discussing how data mining will work on large datasets nothing but BigData. In this large datasets the three main challenges are velocity,variety and volume. Hadoop is a tool which is used to extract knowledge from large datasets and we so many nosql databases to work along with hadoop the section of nosql database depends on our requiremet and CAP theorm. In this paper we are going discuss about the hadoop MapReduce which is a can run by using key/value pair increases the throughput and efficiency and NoSql databases in respect to CAP theorm.

Keywords - Hadoop MapReduce; NoSql databases; CAP theorm; HDFS; Datamining; ACID ; Base

I. INTRODUCTION (HEADING 1)

“Big data” is nothing but an information which fullfills three V’s nothing but Velocity, Variety and Volume to examine or analyze this sort of information traditional tools cannot be adequate. In today's reality the information which is coming inside the oraganizations comes in with undetermined pace they dont know how to get values from the data. Most of the data is in raw form that is unstructured or semi-structured data. Orgnsizations dont know that the information getting inside are worthy or not they are in dailama that they need to store and process that information or not. So this is a major information challenge in present era. So this challenge will need new archieteuctre or techniques or algorithms or analytics to work with this type of data and extract meaningful information or knowledge from it. Volume of data increasing rapidly from tera bytes to peta bytes day-by-day. This data is generating from social media like facebook,twitter and generated from scientifi instruments,mobile phones,sensor data and networked data. To do data mining tasks on generated data and extract knowledge from it will be crucial task with traditional tools. To visualize,summerize,analyze and discover knowledge from this data apache introduced a new framework called Hadoop. Hadoop is a software library that supports processing of large datasets in a scalable and distributed manner across cluster of computers using simple programming model. Hadoop framework consists of two major componenets one is HDFS(hadoop distributed file system) and other one is MapReduce. MapReduce is programming technique to process the data in parallel manner using key/value pairs and HDFS is used to store the data ina persistent place. The input to the MapReduce program is taken from HDFS in the form of key/value pairs and return store the output to the HDFS again.

To analyze the BigData traditional sql databases are not adequate because of the scalability issue. To process or examine this Big volumes of data NoSql databases are needed. NoSql is nothing but “Not Only Sql”. There are so many NoSql database are available today those are CouchDB,MongoDB,Hbase,casendra,AmazonDyna
mo, etc. The major issue behind the relational database are scalability issue if the data volumes increases rapidly performance degrades this leads to new era that is NOSQL data models. Though these services available long ago after the concept of DBaaS (Database as a service). NoSql databases are major competitor of relational database model because of it provides high scalability. NoSql databases are column oriented databases unlike relational databases it does not follow ACID properties (Atomocity, Consistency , Isolation and Durability) instead it follows BASE(Basically Available, Soft state, Eventual consistency). This systems are in compliance with CAP (Consistency , Availability and Parition Tollerance) theorem.

Rest of the paper organized as follows in section II some of the works done previously in section III About Hadoop MapReduce IV types of NOSQL databases in section V selection of NOSQL database using CAP theorem in section VI ends with conclusion

II. RELATED WORK

In [1] Goggle Inc Jeffrey Dean and Sanjay Ghemawat (2004) proposed an MapReduce. MapReduce is a programming model and an associated implementation for processing and generating large datasets. Many real world tasks can be expressed in this model. In this model the programs are automatically parallelized and executed on large commodity of machines. Inter communication between the systems , machine failure , partioning the input data and program execution takes care by the run-time system. Users specify a map function which can process key/value pairs to generate intermediate key/value pairs which is used by the reduce function which is used to merge the intermediate output which are having same key.

In [2] Weizhong Zhao, Huifang Ma1 and Qing He 1 The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (2009) proposed an enhanced model based on the standard k-means algorithms. Now a days data has been increasing day by day to clustering large volumes of data still a challenging task. To deal with this kind of problem, many researchers try to contribute parallel efficient algorithms. In this paper they are trying to contribute parallel k-means algorithm using Mapreduce which is simple yet more powerful parallel programming technique.

In [3] the author said how the big data is generating from different sources of the technology. Data in the enterprise have been increasing rapidly in exponential rates. Big data size is moving from tera bytes to zetabytes. Some of the crucial sources where this data is generating are social media, spatial data, satilite information, internal search indexing, phone call records, web logs network data, RFID, biogeochemical, etc.,

In [4] author propose an optimization technique using hadoop mapReduce in this work they propose most popular partial swarm optimization as a parallel working algorithm using MapReduce technique. In this work from the result they found that parallel PSO (MRPSO) gives better result even the data scales out in the size. It also maintain the quality of the cluster.

In [5] author provide a difference between a sql and NOSQL databases and also given the advantages and disadvantages of various NoSql databases and also suggested that before chosing the technique need to go through understanding of advantages and disadvantages of each and every NoSql database.

III. ABOUT HADOOP FRAMEWORK

MapReduce is a great start, but it requires you to expend a significant amount of developer and technology resources to make it work in your organization. This isn’t feasible for most enterprises unless the name on the office building says “Google.” This relative complexity led to the advent of Hadoop.

Hadoop is a well-adopted, standards-based, open-source software framework built on the foundation of Google’s MapReduce and Google
File System papers. It’s meant to leverage the power of massive parallel processing to take advantage of Big Data, generally by using lots of inexpensive commodity servers. Hadoop is designed to abstract away much of the complexity of distributed processing. This lets developers focus on the task at hand, instead of getting lost in the technical details of deploying such a functionally rich environment.

The not-for-profit Apache Software Foundation has taken over maintenance of Hadoop, with Yahoo! making significant contributions. Hadoop has gained tremendous adoption in a wide variety of organizations, including the following:

- Social media (e.g., Facebook, Twitter)
- Life sciences
- Financial services
- Retail
- Government

IV. TYPES OF NOSQL DATABASES

There are so many NoSql databases had been arrived each have its own importance and the selection of NoSql database vary from the source data.

A. RIAK

Ria is an appropriated, fault tolerant, open source database created by Basho technologies utilizing C, Erlang and JavaScript. It implements standards from Amazon’s Dynamo paper. It has an adaptable data schema. It offers high availability, partition tolerance and persistence. Segments of Ria will be Ria Clients, Webmachine, Protocol Buffers, Ria Replication, Ria SNMP/JMX, Ria KV, Ria Search, Ria Pipe and Ria Core.

B. AmazonDynamoDB

Amazon DynamoDB is a recently released completely overseen NOSQL database administration offered by Amazon that gives a quick, highly reliable and cost-effective NOSQL database administration intended for web scale applications. It is executed by using Amazon’s Dynamo model. It offers low, unsurprising latencies at any scale. It stores information on strong state drives (SSD) rather than conventional hard drives subsequently giving speedier access to the information. The information is duplicated synchronously over various AWS Availability Zones in an AWS Region to give high availability and data durability. It replicates data across at least three data centers, thus providing high availability and durability even under complex failure scenarios.
C. **CouchDB**

ApacheCouchDB generally called as CouchDB is a open source project from Apache unlike the relatioanl databases it does store the data in the form of relations. The data is stored in the form of JSON objects. Unlike relational databases it is document-oriented NoSql database which uses javascript for querying using MapReduce. It stores the data in the form of documents each document is maintain its own data and schema. CouchDB implements a form of multi version concurrency control to avoid locking of database during the write operation. One of its recognizing features is multi-master replication, which permits it to scale across machines to build high performance systems. It uses HTTP for its API and it is released in 2005 later in 2008 it has been project of Apache. It is eventually consistent and built for offline usage.

D. **MangoDB**

MongoDB was developed by Neo Technology and was initially released in 2007. It is a cross-platform document oriented database. Unlike traditional database tables it stores the data in the form of JSON like document with dynamic schema.it calls this format as BSON(binary JSON) making the integration of data in certain types of application easier and faster. It is released as platform as a service model later company shifted as open source project. MongoDB is currently being used by MTV networks, Foursquare, The Guardian etc. It is also being used in projects like CERN’s LHC, UIDAI Aadhaar which is India's unique identification project. The disadvantages are that it can be unreliable and indexing takes up lot of ram. It is a fourth most popular database as of july 2015.

E. **BigTables**

Google’s Big Tables is initially released in the year 2005 which works on top of google file System (GFS). It is a column oriented database and compressed high performance database. Big Tables are implemented by using c and c++. It provides consistancy persisance and fault tollerant. It is designed to scale across thousand of machines it is very easy to add more machines. Big Tables are used by number of google applications like gmail,youtube and google earth. The Big Table implementation has three major components: a library that is linked into every client, one master server, and many tablet servers. Tablet servers are used to manage a set of tablets (same as tables in RDBMS). The master server handles schema changes, performs tasks like assigning tablets to tablet servers, balancing tablet server load, garbage collection etc. Big Table is not distributed outside Google, but it is available as a part of Google app engine.

F. **Cassandra**

Cassandra was produced by Apache Software Foundations and was released in 2008. It was implemented utilizing Java. It depends on both Amazon’s Dynamo model and Google’s Big table. Hence it includes ideas of both key-value stores and column-oriented. It offers features like high accessibility, partition tolerance, persistence, high scalability and so on. It has a dynamic schema. It can be utilized for a mixed bag of utilization like partition tolerance, persistence, high scalability and so forth. Cassandra is being utilized by Adobe, Digg, eBay, Twitter and so on. The disadvantage of Cassandra is that reads are relatively slower than writes. Compared to remaining NoSql databases it is faster in both the scenario.

V. **Nosql Selection Using CAP Theorm**

CAP theorem expresses that there are 3 basic requirements which exists in a special relation when designing applications for a distributed archietecture.

Consistency : This means that the data in a database remains consistancy after the execution of an operation.

Availability : This means that the system is always on.

Partition Tollerance : This means that the system continous to function even the communication among the server is unreliable i.e the servers may be partitioned into multiple groups that cannot communicate one another.

CAP provides a basic requirements for the distributed system to fallow 2 of the 3 requirements. In theoritically it is impossible to acquire all the 3 requirements.
Then for all the current NoSQL databases follows the different combination of the C,A,P from the CAP Theorm.

The pictorial representation of CAP Theorm with different combination of C,A,P is shown below

![CAP Theorm chart]

In the above diagram shows that CA and CP follows ACID properties, and AP follows BASE properties

![CAP Implications]

CA – Single site cluster therefore all nodes are always in contact when a partition occurs the system blocks.

CP – some data may not be accurate but the rest is still consistent.

AP – System is still available under partitioning, but some of the data returned may be inaccurate.

So from the CAP Theorm we can say that the Selection of NoSQL database is depending on dataset and also the implications of CAP theorem.

But Casandra NoSQL database is one of the best NoSQL database because it have the capability of key-value and column oriented databases. Column oriented databases are most suited for the datamining analysis.

VI. CONCLUSION

Now a days data increasing from terabytes to petabytes traditional techniques like sql database are failed to process this huge ammount of data due to its scalability issues so we need to incorporate the NoSql databases along with hadoop like BigData processing tools. We had seen so Many NoSql databases among them casandra NoSql database is preferable for datamining analysis because of it is column oriented as well as key-value database nature. Casandra is very easy to install,maintain and code.

REFERENCES


[4] Ibrahim Aljarah and Simone A. Ludwig, “Parallel Particle Swarm Optimization Clustering Algorithm based on MapReduce Methodology,” 978-1-4673-4768-6 c 2012IEEE.


