Management of Big Data with Riak/MongoDB

Reena Simon
Department of Computer Science
K.M.C., College, Khopoli
reenasimon1@gmail.com

Abstract: **For managing huge amount of data which can't be anticipated in advance that are high-volume, high- velocity and high variety-information assets requires new way of processing to ensure enhanced decision making, insights, and process optimization. Hence managing such type of data becomes a challenge and necessity for today's fast growing business needs and social network.**

Riak supports all, providing high volume (data that is available to read and write), high velocity (easily respond to growth), high variety (can store any type of data as a value). Riak is NoSQL, distributed key-value store is fault tolerant, provides scalability, high availability, operational simplicity and data accuracy. Riak products supports big data, IoT and hybrid cloud applications. Implementation language is Erlang and supports concurrency, solid distributed communication, fault tolerance and run on a virtual machine

MongoDB is NoSQL document store database typically store self describing JSON, XML document. They are key-value store and every value is considered as single document that store all the data. It provides cloud service. It has distributed database approach, high performance, high availability and automatic scaling to the database. Data Curation helps to collect, manage and preserve large scale data. Using various Machine learning models we can train the data sets and make accurate and useful predictions from data sets.

Keywords : Data Management, Distributed Storage, high volume, high velocity, high variety data, Riak.

I. INTRODUCTION

Data has become the most important fact or figure or statistics for problem solving. Analysis can be done on the data to get an expected outcome. We also need to identify the classification of the data that either it is quantitative or qualitative data. It depends on our problem that what type of data we need to consider for obtaining the perfect solution.

The requirement of data and its processing has changed in numerous ways because of the large growing businesses and their competitions of earning more profit than others and the facts and figures that gets generated each single minute the demand to extract the useful insight becomes necessary for the business. Also the type of data that is passed or shared on to the social network sites varies from person to person which gives rise to big data. This all constitutes large data that simply can't be managed with the traditional RDBMS that consists of tables, columns, rows and schemas to store, manage and extract the data. NoSQL doesn't work on structured data and schemas and apply more flexible data model to work.. Hence to manage data that is unstructured data, the concept of NoSQL came into existence which means "not only SQL" that can be used to manage Big data. An appreciable advantage of NoSQL databases is, that it allows the insertion of data without a predefined schema that makes it easy to do application changes in real-time because of which development is faster, code generation is reliable and less database administrator time is required. It supports validation rules for the database, which in turn allows user to enforce governance across the data resulting in maintaining the agility benefits of a dynamic schema.

Riak is a NoSQL, open source, distributed, fault tolerant, key-value store that offers unmatched resiliency with high availability, innovative technology to ensure data accuracy, it never lose a write.

MongoDB is a NoSQL, distributed, open source, document store with high reliability and provides automatic scaling to the database.

Both supports JSON (JavaScript Object Notation) format. JSON is used for managing, retrieving and formatting the data and is language independent. Being a programmer, one always have to think in terms of Objects. Now the interesting thing is that the database also does the same way.

II. BACKGROUND AND RELATED WORK

A. Riak and MongoDB for Modern Applications

This research is based on existing literature of Riak and MongoDB and reviewing a Research on Riak KV performance in Sensor Data. The literature helps to understand the NoSQL concept for distributed data storage.

Riak being a open source, distributed key-value store is fault tolerant, provides scalability, high availability, also gives operational simplicity and data accuracy. It supports all these providing high volume (data that is available to read and write), high velocity (easily respond to growth), high variety (can store any type of data as a value). Riak product ensure big data, IoT and hybrid cloud applications. It is the most resilient NoSQL databases.

Riak KV - flexible key-value data model. Riak TS - for IoT and other Time series.

Basho's Riak focuses on data management and provides many other facilitates. Riak allows to store values of any type and works on key value.

Riak system allows storing hundreds of TBs of data and it handles several GBs daily per node. It has a ring scale out architecture and divides the data around ring. Riak focuses on replication. Replication is automatic in Riak provides security, if a node in the Riak cluster fails, still the data will be available. Data is replicated to nodes, property is set in a bucket's bucket type. Bucket is a flat namespace in Riak. Key name can be the same in multiple buckets. Bucket Type allows a group of buckets to share configuration details. Key is a binary value or string that identifies objects. Value is the actual data stored in Riak. They don't support transactions. Intelligent replication is the feature of Riak KV as by default it makes three replicas of the data on different nodes. Riak is a Basho's product, Apache2 License. Being a Open source It is a multi data center replication with a new S3, API that was released**.**

MongoDB is an open source, distributed document store scaleble database. The collection consists of different documents. of the document can be different in every document of the collection. Each document can have the same or a different structure**.** It has no complex joins and support dynamic queries on document. It is a document database, that is it stores data in JSON like documents. We can format the output in JSON format so that it can improve readability. Mongo has an interesting distribution system when they have masters and slaves called a replica set and they run read and write slaves of that. They have much nicer sets of API's. It also supports deep complex query language. It uses internal memory for storing the working set enabling faster access to data. MongoDB can be used for Data hub, User data management.

Rich JSON documents - supports working with data in a flexible way and provides dynamic schemas.

Powerful query language - Query language supports sorting even if it is nested within a document.

MongoDB assigns a unique id to every document. This unique _id is a 12 byte number which is a hexadecimal number which identifies uniquely every document within the unstructured data. The advantage is that, this _id can be used while working with the data and querying the document.

III. METHODOLOGY AND TECHNOLOGY

TABLE I. DIFFERENCES AND SIMILARITIES IN RIAK AND MONGODB

Riak KV	**MongoDB**
Riak is distributed, key-value store, fault tolerant (reads and writes non-stop), operational simplicity, provides scalability.	MongoDB is distributed, document store, available as a cloud service provider and is easy to scale.
Primary Database model is Key- Value store.	Primary Database model is Document store.
Initial release 2009	Initial release 2009
Current release 2.1.0 April 2015	Current release 4.2.2 Dec 2019
Data Schema - Schema free	Data Schema - Schema free
Implementation Language - Erlang	Implementation Language - C++
ServerOperating systems - Linux, OS X	ServerOperating systems - Linux, OS X, Solaris, Windows
Website: www.basho.com/products/	Website: www.mongodb.com

IV. EXAMPLE (DATA CURATION)

Riak TS also support queries. We can install and use Riak shell to create table and run queries. Used Riak shell to create database. It supports various commands like create, delete, describe, explain, insert, select, show. Describe command is for examining table structure and Show is for listing tables.

riak-shell(1)> CREATE TABLE Sensordata (id SIN64 NOT NULL, time TIMESTAMP NOT NULL, value DOUBLE, PRIMARY

KEY(id, QUANTUM(time, 15, 'm')), id, time));

Upon successful creation of table it diplays a message:

Table Sensordata successfully created and activated. It has three columns : id, time, value. The partition key includes the id with a 15 minutes time quantum, and a local key made up of id and time.

riak-shell(2)>SHOW TABLES;

riak-shell(3)>DESCRIBE Sensordata;

riak-shell(4)>INSERT INTO Sensordata(id,time,value) VALUES(1,'2019-12-14 01:00:00Z', 65.0);

riak-shell(5)>SELECT id, time, value from Sensordata; Riak KV 2.0:

offers five datatypes that can reduce some of the complexities of developing, using riak : flags, registers, counters, sets and maps.

In MongoDB we create a database and then collection within the database and documents are inserted in that collection and every document can be different in size, refer the below example. Querying according to the criteria, Sorting, can be done on the data in ascending and descending order based on a particular field of document using sort(). Index helps in searching the data faster. It improves the speed of search operations. createIndex() method is implemented for creating indexes. We can find the indexes within a collection using db.collectionname.getIndexes() and many more commands can be applied on the documents within the collection.

We can download and install MongoDB Used MongoDB shell to create database. create a database customer

>use customer

Display the name of the current database

>db

Insert a document and create collection with a single command

>
db.custdetail.insert({"custno":101,"cname":"anu":"age":25});

>db.custdetail.insert({"custno":102,"cname":"smita","ag
e":30,"city": "pune"});

Note: In above example the documents has different number of fields. This shows that MongoDB allows to create different documents within the same collection.

Print the data in JSON format

>db.custdetail.find().forEach(printjson);

>db.custdetail.find().pretty();

Create collection before inserting the documents

>db.createCollection("Pr
oduct"); Vew all the
collections

>show collections;

Specify a criteria for extracting data

>db.custdetail.find({"age":{$gt:25}}
).pretty(); Display only the age field of
customer

>db.custdetail.find({},{"_id":0,"age":1});

Value 1 will show the field and value 0 mean don't show that field and thus only age will be displayed of all the customer.

V. DISCUSSION

To collect and preserve large unstructured data NoSQL is the best option. As the type of the data that is shared between different users of social networking sites differs and the data that demand changes within a second, according to the needs of the users, which obviously can never be a structured data, hence NoSQL dbs are the better options to manage these large chunks of data, because they allow the data (Unstructured data) to store exactly the way they are keyed in.

Data can be stored and managed with MongoDB, also it can be imported in R for Statistical Analysis.

Riak a scalable distributed data store which is useful for sensor data storage that is time series data which can come from remote sources. It also supports objects, immutable data, MapReduce. It has limits also that it is not good for objects that exceed 1-2 MB in size, objects with complex interdependencies.

Riak is a key value design and deliver powerful, simple models for data that is huge amount of unstructured data. It provides fast and high performance.

We can model data using Riak's key-value design to store key-value pairs comprising objects in buckets that are flat namespaces with some configuration properties, like the replication factor.

Riak is not that popular and hence less used. MongoDB has become more popular and is more preferred as there is ease and variety in performing operations.

VI. CONCLUSIONS

As the data size is growing day-by-day every year within seconds and because of the huge and tremendously varying data that is shared between the different users, which increases difficulty level to manage the data, which is no doubt unstructured, hence NoSQL are the best DBs for maintaining and manipulating data which provides the feature of Data Management.

Riak is a key value store that offers less maintenance and replicaion. Mongodb allows querying documents on random fields but it do not focus much on replication.

Riak and MongoDB NoSQL is used to manage huge amount of unstructured data. Both has its advantages and disadvantages and thus are suitable, according to their features for a particular problem.

If there is a need to perform key-value queries then Riak is suitable. If it needs a lot more code building like indexes, etc then Mongodb is good option for these queries.

VII. REFERENCES

[1] A Little Riak Core Book- Mariano Guerra A Little Riak book - Eric Redmond

[2] Basho-labs/little_riak_book: A Little Riak Book - GitHub A Little Riak Book ACaaSIA

[3] Riak KV Application Guide https://docs.riak.com Create Your First Riak Table https://riak.com Downloads from www.mongodb.com

[4] MongoDB https://.mongodb.com

[5] MongoDB Notes for professionals: https://books.goalkicker.com Doing Data Science, Rachel S. and O'Neil, O'Reilly