Polyglot Persistence: Handling Multiple Datastores

Ms. Namrata Rawal[1], Ms. Vatika Sharma[2]

[1]PG Student, Network Security, GTU PG School, Ahmedabad, Gujarat, India
[2]I-verve Infoweb Company Ahmedabad, Gujarat, India

ABSTRACT

Handling Big Data means to handle huge databases. Although we have seen that we are handling reltaional data or non-relational data at a time using Map Reduce framework. In other words we can say that handling of multiple datastores cannot be done at a time. So, polyglot persistence come into place to handle it. This paper focuses on Polyglot persistence it is the term that used to describe different data storage technologies to handle multiple data stores at a same time.

Keywords:-*Polyglot persistence, Hadoop, MR, oracle cloud, NoSQL*

1. INTRODUCTION

Handling Big Data means to handle a huge amount of database without any restrictions of relational databases (data storage in form of rows and columns) so; it begins the concept of NoSQL databases to handle non-relational databases on an oracle cloud with no restrictions.

2. MOTIVTAION

Honestly speaking there is nothing new under the Sun. Integration has been a problem that we have faced for 40-plus years in IT. But there are many new ways to handle a problem. So the role is shaped from SOA (Service Oriented Architecture) on the cloud. Another one perspective is big data is also coming in, so the problem is how to integrate that data whether by using historical perspective or from data mining an analytical perspective. So a question arises how we deal or handle huge volumes of data. So to handle big data is a major problem. So I am going to explore many ways to handle these problems by combining past technologies used that are available with us and with the help of new technologies to solve this problem rapidly. This all has been done by making promzise that we do not forget the lessons learned and then trying to deliver a new system whose performance is cost-effective, consistent and to respond to business faster. Here is an overview of important technologies to know about for context around big data infrastructure.

- SQL/Traditional RDBMS (older technology, losing relevance)
- NoSQL Database Systems
- Hadoop, MapReduce, and massively parallel computing

3. NOSQL

It stands for Not Only Structured Query language[1]. It is used to store huge amount of data storage which keeps on increasing day by day. NoSQL is a non-relational database, portable and faster information retrieval databases. It is the database that is

Non-relational- It is non-relational because data is not stored in form of tables

Open source- It is open source because everyone can look into its code freely, can update and everyone can compile it
Distributed- It is distributed because its data can spread onto multiple machines with the concept of data replication (which helps in faults and failures)
Horizontally scalable-It is horizontally scalable as it leads to high performance in linear way

3.1. Basics of NoSQL

ACID: (Atomicity Consistency Isolation Durability) SQL uses the concept of ACID and the consistency feature. NoSQL leads to the concept of ACID free as in the distributed environment data is spread to multiple machines and each machine stores its data. The maintenance of consistency is needed. For example: In SQL, suppose there are two accounts 'A' and 'B'. 'A' account is going to transfer some amount to 'B' account then if in between transaction, failure begin before updating both accounts then it leds to hindrance of consistency property. So with the help of NoSQL the change in one tuple of data can lead to change in each and every machine were the data reside.

BASE*:* BASE stand for Basically Available Soft State and Eventual Consistency[10]. BASE is the opposite of ACID. It is called BASE because each time transaction occurs then the system will not lead it to go to solid state that is it makes its transaction into soft state (temporary).

CAP Theorem*:* It stands for Consistency Availability and partition tolerance[4]. It is the theorem that is based on 3 things.
Data need to be updated on all machines frequently that is consistency
Data need to be available permanently and can be accessed each and every time that is availability
Keeps on doing work (without stopping work during any failure or fault of machine) that is partition tolerance

3.2. Types of NoSQL Data Stores

Key value Databases: It is the fusion of two objects that are Key and Value[1] The key-value data store is the simplest and important model. It stores the data in schema less manner. Key is a unique identifier to certain data entry. Key should not duplicate, repeat or same in database.

Table -1: KV Datastore

Key	Value
K1	V1
K2	V1
K3	V2

Some key value store datastores are Dynamo [1], Voldemort, and Riak[1]

Document datastore: It is the databases that used to store data in form of documents. Document store provides better performance and horizontally scalable. It is somewhat similar to relational database but this is more flexible because it is schema less. Some document datastore are mongodb[1], couchdb[1], redis[1], and lotus notes

Column oriented: Column oriented datastore in NoSQL is hybrid of row/column store instead of relational column databases. It is fit for data mining and analytic applications. Main advantage of column oriented databases is that it is highly compressed, self-indexing and offers high scalability. Some column oriented datastore are BigTable[1], Cassandra[1], and HBase[1].

Graph databases: Graph Databases are elaborate with nodes and edges where nodes react as the object and edges react as the relationship between that objects. It is ACID adaptable and provide rollback support some graph data store are neo4j [1], FlockDB, GraphBase

4. HADOOP

Hadoop is an open source framework it is java based programming which maintain the storage and process of huge amount of data in a distributed computing environment.

5. HADOOP CLUSTER

Hadoop cluster defined as a type of computational cluster used for storage of huge amount of unstructured data in distributed environment. Hadoop cluster runs on a low cost. This paper explain

5.1. Architecture of Hadoop Cluster

In hadoop cluster architecture there was main five building blocks into runtime environment. It is bottom-up approach

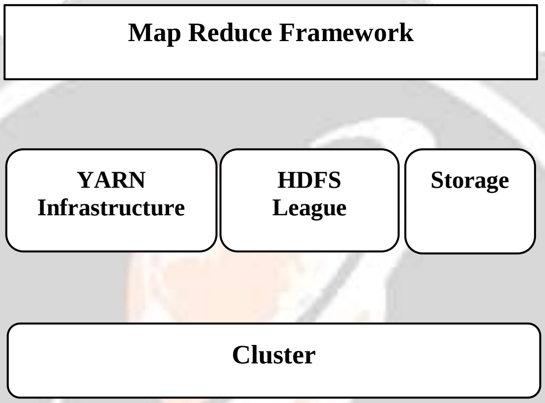

Fig-1: Hadoop Cluster Architecture

Cluster is the hardware section of the infrastructure. It is the set of nodes (host machines).
YARN (Yet another Resource Negotiator) Infrastructure is the computational resource provider. It is responsible for providing resources like CPU, memory, etc needed for application implementation[6].
HDFS League is permanent, reliable and distributed storage provider. It is mainly used for storage of inputs and outputs (no intermediate ones).

MR Framework is the software section where MR paradigm is implemented

5.2. Core components of Hadoop Cluster

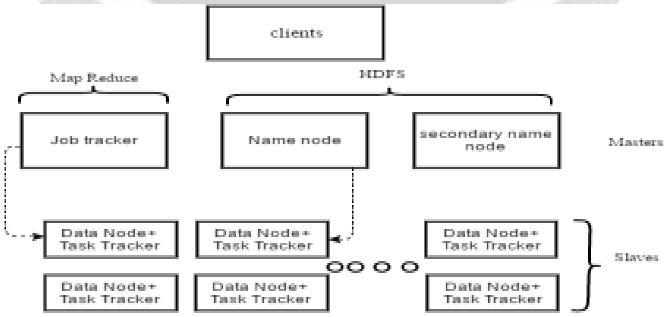

Fig-2: Core components

Client is the neither master nor slave it will submit and retrieve the data
Master consist of main three components Namenode, Secondary node and Jobtracker
Namenode: maintain track of all the file system related details

Secondary node: it is replica of namenode
Jobtracker: Coordinates the parallel processing of data using Map Reduce.
Slave nodes are of huge number in hadoop cluster and they are responsible for data storing and its computational process
Task tracker is slave to the job tracker
Data node is salve to the name node

MR framework works on a master slave architecture. As its name suggests it work in two phases one is map and other is reduce. These two phases signify that whenever any of the problem data given to the master then it is first mapped with all the slaves and the solution data to the clients is given in the reduced form.

6. HISTORY OF POLYGLOT PERSISTENCE

In 2006 Neal Ford [5] strike the concept of polyglot programming[5], to convey the idea that application should be written in a combination of different languages to take benefit of the fact that different languages are suitable to overcome the different problems. This same concept recently fetched by the Martin fowler and Pramod Sadalage [5] and applied to databases; so application can work with multiple databases and giving birth to polyglot persistence

7. POLYGLOT PRSISTENCE

Distinct databases are designed to resolve distinct problems. Working with single database for all of the needs then it will lead to issue of performance for data storage, transaction, caching session information, maintain information and data will create different problems. It can apply over an enterprise or within a single application.Polyglot persistence is process to fragment or divide data into multiple databases to influence their power and offer better scalability and performance.

8. PROS AND CONS OF POLYGLOT PERSISTENCE

PROS

- Response time is fast: Application is influenced by all the features of the databases, which makes the fast response time
- Increase performance
- Increase scalability
- Increase flexibility
- Rich experience: Have a rich experience when have a power of multiple databases at the same time.
- More reliable
- Handling failures and faults
- Manage multiple datastores at a time
- Multitasking

CONS

- Require specialist to integrate different databases
- Executer required to learn different databases
- Resources are required to manage databasesss
- Harsh to test
- More complex
- High cost

9. METHODOLOGY

The below fig:3 methodology shows how to handle bigdata using polyglot persistence. In traditional approach bigdata was initially storing data then access that data using only SQL datastore while in modern approach creating hadoop cluster using master-slaves concept and store it on oracle cloud it will allows to access multiple data like SQL and NoSQL using multiple databases to access data to achieve polyglot persistence

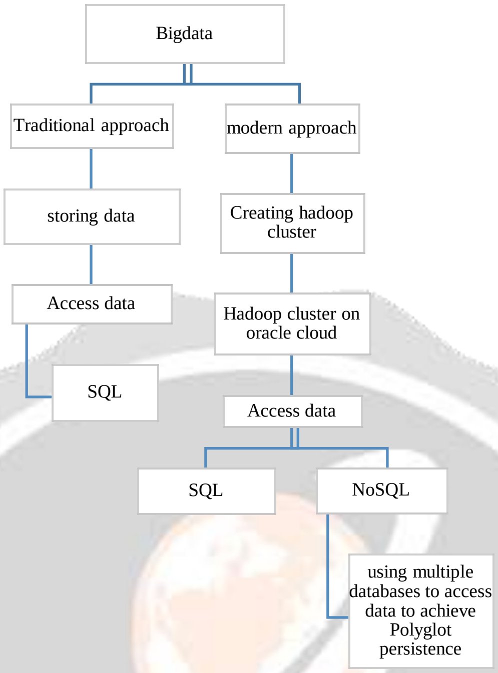

Fig-3: Methodology

10. CONCLUSION AND FUTURE WORK

This paper concludes that accessing of data can be done using multiple datastores on an oracle cloud without any restrictions. By this polyglot persistence technology we have increase the performance metrics in terms of respose time but complexity increases.Future work states to implement the major performance parameters that achieve polyglot persistence.

11. REFERENCES

[1] AmeyaNayak Anil Poriya and DikshayPoojary, “Type of NOSQL Databases and its Comparison with Relational Databases”, in International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 5– No.4, March 2013

[2] Amrit Pal, Pinki Agrawal, Kunal Jain and Sanjay Agrawal, “A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data Using Hadoop”in 2014 Fourth International Conference on Communication Systems and Network Technologies, IEEE 2014

[3] Jinson Zhang and Mao Lin Huang “5Ws Model for BigData Analysis and Visualization”in 2013 IEEE 16th International Conference on Computational Science and Engineering

[4] Prateek Nepaliya, Prateek Gupta, “Performance Analysis of NoSQL Databases” in, International Journal of Computer Applications Volume 127 – No.12 October 2015

[5] Pramodkumar J Sadalage and Martin fowler In NoSQL Distilled:A brief guide to the emerging world of polyglot Persistence; Addison Wesley publication, New-York ,August 2012

[6] Vatika Sharma, Meenu Dave, “NoSQL and Hadoop Technologies On Oracle Cloud”, in Interantional journal of Emerging Trends and Tchnlogy in Computer Science(IJETTCS) Volume 2, Issue 2, March – April 2013

[7] K. Jayasri1, R. Rajmohan2 and D. Dinagaran, “Analyzing the Query Performances of Description Logic based Service Matching using Hadoop” in International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM),August IEEE

[8] CostinJiurjiu, April 9, 2015 “Introducing Oracle Public cloudServices.”retrievedfromwww.oracle.com/webfolder/s/delivery_production/docs/FY15h1/doc17/PAAS-01 Introducing-Oracle.pdf

[9] Aditya B. Patel, Manashvi Birla and Ushma Nair, “Addressing Big Data Problem Using Hadoop and Map Reduce” ,in 2012 Nirma University International Conference On Engineering, Nuicone-2012, 06-08december, 04 April IEE**E**

[10] Konstantin Shvachko, HairongKuang, Sanjay Radia and Robert Chansler. “The Hadoop Distributed File System”, in Mass Storage Systems and Technologies (MSST), 2010 IEEE

[11] Supriya S. Pore and Swalaya B. Pawar, “Comparative Study of SQL &NoSQL Databases”, in International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 5, May 2015

[12] Pragati Prakash Srivastava, SaumyaGoyal and Anil Kumar, “Analysis of Various NoSql Database” ,in 2015 IEEE

[13] Ricardo Jimenez-Peris, Marta Patino-Martinez, Ivan Brondino and Valerio Transactional Processing for Polyglot Persistence”, in 2016 30th International Conference on Advanced Information Networking and Applications Workshops, 19 May 2016 IEEE

[14] Yishan Li and SathiamoorthyManoharan , “A performance comparison of SQL and NoSQLdatabases”,in 2013 IEEEJournal Of Engineering And Computer Science (IJECS) ISSN: 2319-7242 Volume 5 Issue 1 January 2016