International Journal of Scientific & Engineering Research, Volume 5, Issue 9, September-2014
ISSN 2229-5518

Innovative Information retrieval Framework on dedicated cloud using Map Reduce & Hummingbird primitives

Dr.Piyush Gupta, Kashinath Chandelkar

Abstract— Cloud computing is an upcoming Technology comprising of IaaS (Infrastructure as a Service), PaaS (Platform as a Service) and SaaS (Software as a Service). A hosted instance of Hadoop Cluster in cloud termed as “Qubole” is tested for a job that helps to retrieve information for end users using Hummingbird Algorithm. A chunk of data called “Big Data” is a challenge to extract information and utilize for real time decision making. Since hummingbird supports voice based real time information retrieval is used in support with map Reduce in dedicated cloud.

Index Terms— IaaS, PaaS, SaaS, Big Data, Qubole, hummingbird, Map Reduce.

1. INTRODUCTION

The organization possesses both physical and logical infrastructure for data management across the departments. As the organization grows, data of the organization also increase exponentially. Understanding this philosophy at W3 (World Wide Web) huge amount of data chunks are piled every day in a network, increasing resources and cost across the network.

The problem of efficient information retrieval from Big Data highlighted in this paper needs introduction to following relevant concepts.

1.1 Cloud Computing

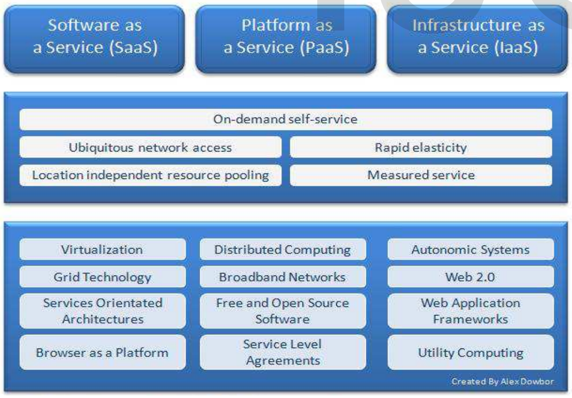


Fig-1: Cloud Architecture **Source**: NIST

Dr. Piyush Gupta Asst. Prof., Dept. of Computer Science & Engineering Birla Institute of Technology, Jaipur, piyush.bitjpr@gmail.com

Chandelkar Kashinath K. Research Scholar, Dept. of Computer Science & Engineering Birla Institute of Technology, Jaipur, Kashinath45@gmail.com

The emerging technology called cloud computing (Fig-1) is an extension of resources managed by a service called a hypervisor. Each virtual machine is migrated on public, private or hybrid cloud serving as IaaS (Infrastructure as a Service), PaaS (Platform as a Service) and SaaS (Software as a Service). Being each virtual machine or VMware has an instance of software running, hypervisor plays an important role to maintain the health and security among the virtual machines. The Elastic Services are measured based on usage for storage, network and other virtual allocated resources.

1.2 Map Reduce Algorithm

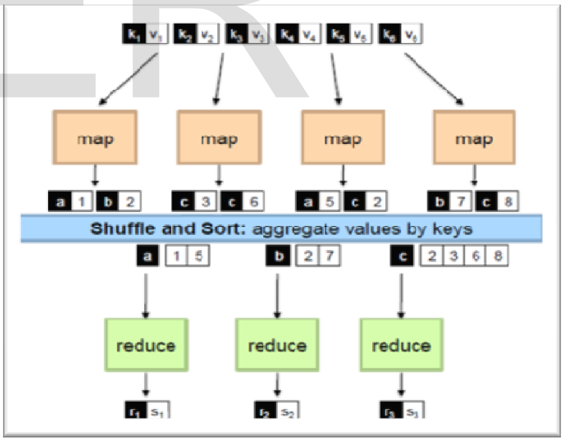


Fig-2: MapReduce Algorithm **Source**: Jimmy Lin

In addition to existing information retrieval algorithms like BFS (Breadth First Search) and centralized approach in traditional architecture, the Map Reduce algorithm has a special potential to work parallel on giving environment. It has two basic components- mappers and reducers.

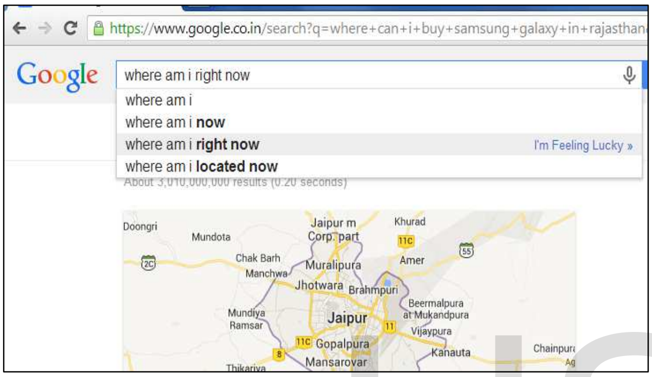
Data input is given to the mapper in the form of a database or set of files. These files are processed by mappers using shuffle and sorting method. The collected output is an input to the reducer. Information collected from reducer being orderly processed and efficiently used for decision making.

1.3 Hummingbird Algorithm

The hummingbird is a new searching algorithm that uses the following steps while retrieving realtime information using voice based search on dynamic node.

1. Input Query using voice based device.
2. Query accepted as a sentence
3. Acquired meaning of the sentence
4. Search for relevant data in the database
5. Realtime voice based results are displayed

Result collected using above steps is shown as under.



https://www.google.co.in/search?q=where+can+i+buy+samsung+galaxy+in+rajasthan

where am i right now

where am i
where am i **now**
where am i **right now**
where am i **located now**

About 3,010,000,000 results (0.20 seconds)

Fig-3: Hummingbird Search **Source**: Google.com

Google Search Engine was asked to use voice search from the computer and by using an Android phone about my current location, to confirm the existence of a hummingbird. A voice based information was collected from an Android phone. Fig-3 shows one of the results collected using Google Chrome on the computer. The results vary based on location and query input.

2. PROBLEM

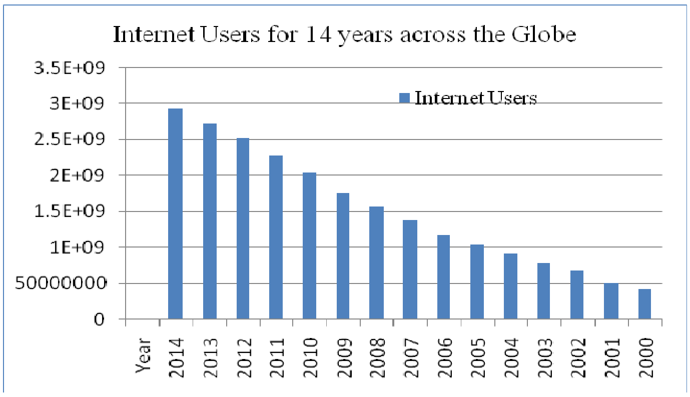


Fig-4: Global Internet users **Source**: W3 Foundation

Data collected till July 1, 2014 helps to understand how internet users are growing exponentially across the globe. These users create data using social media and other sources which require storage, network and secured access throughout.

Since we have limited storage capacity on stand alone computer, the storage limits shall be extended using cloud storage on demand. But extracting relevant information from data chunks requires processing area. Parallel processing is an added advantage. Not only processing cluster helps to extract real time information, delivering it efficiently to end user on dynamic node is a real challenge in Big Data.

3. PROPOSED SOLUTION

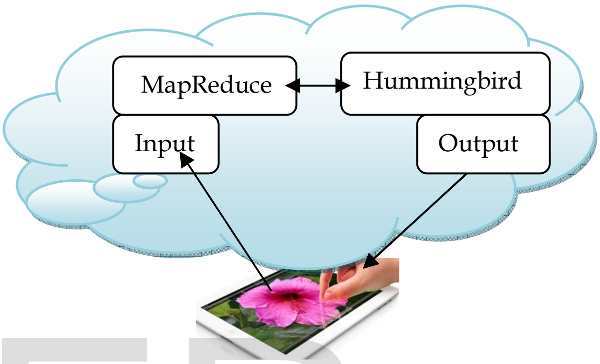
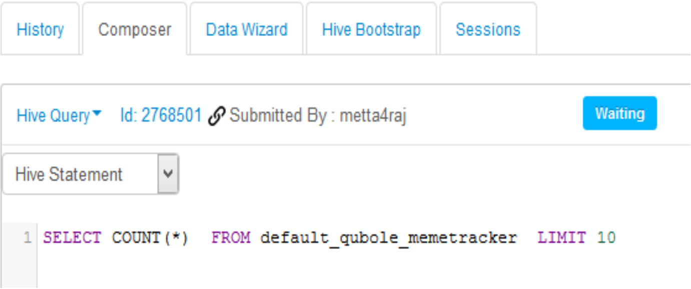


Fig-5: Proposed Information Retrieval System

The proposed hosted solution comprises of private clouds, Map Reducing Algorithms instance supported by the hummingbird algorithm for data delivery. The user inputs data to mapper which is processed using reducer by rearranging as shown in fig-2, and saved in the cloud. The end user has access to stored content which is delivered on request.

4. VALIDATION OF PROPOSED SOLUTION

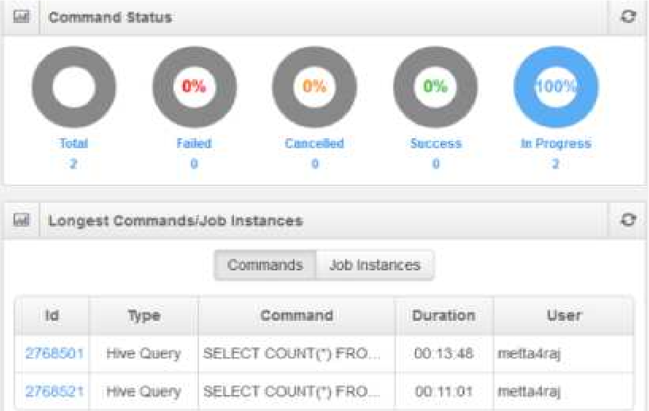


Id: 2768501 Submitted By : metta4raj

```
SELECT COUNT(*)  FROM default_qubole_memetracker  LIMIT 10
```

Fig-6: Query Input to mapper

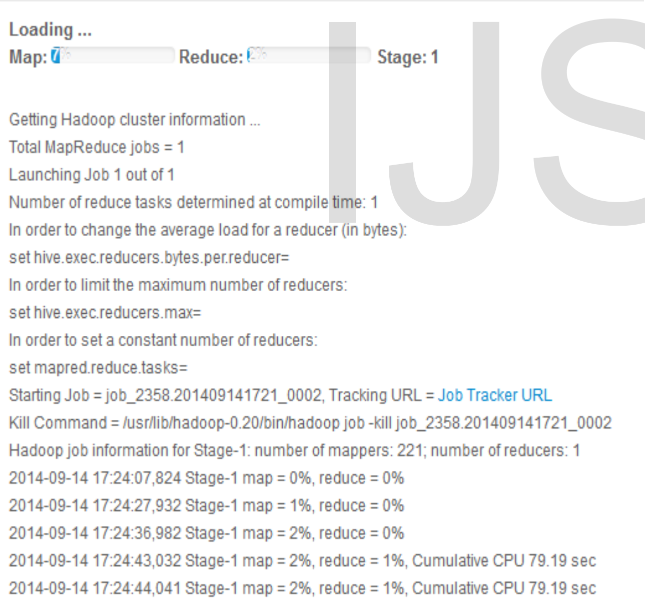
A snapshot of Qubole screen in Fig-6 shows query input to mapper having table input to ten. Hive is used having unique id and user details per session. A set of queries with relevant data shall be given as input for parallel processing. The Amazon S3 Cluster is used to save data for further processing. The setup was further tested by giving two queries as input and time taken to process a query.



Id	Type	Command	Duration	User
2768501	Hive Query	SELECT COUNT(*) FRO...	00:13:48	metta4raj
2768521	Hive Query	SELECT COUNT(*) FRO...	00:11:01	metta4raj

Fig-7: Query dashboard

Fig-7 shows a set of queries being processed simultaneously having unique id and users. The user also gets an intimation if the query fails due to technical issues and shall be recovered under conditions.



```
Loading ...
Map: 7%                Reduce: 2%                Stage: 1

Getting Hadoop cluster information ...
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=
In order to set a constant number of reducers:
set mapred.reduce.tasks=
Starting Job = job_2358.201409141721_0002, Tracking URL = Job Tracker URL
Kill Command = /usr/lib/hadoop-0.20/bin/hadoop job -kill job_2358.201409141721_0002
Hadoop job information for Stage-1: number of mappers: 221; number of reducers: 1
2014-09-14 17:24:07,824 Stage-1 map = 0%, reduce = 0%
2014-09-14 17:24:27,932 Stage-1 map = 1%, reduce = 0%
2014-09-14 17:24:36,982 Stage-1 map = 2%, reduce = 0%
2014-09-14 17:24:43,032 Stage-1 map = 2%, reduce = 1%, Cumulative CPU 79.19 sec
2014-09-14 17:24:44,041 Stage-1 map = 2%, reduce = 1%, Cumulative CPU 79.19 sec
```

Fig-8: Query processing on Hadoop

Fig-8 shows a Hadoop cluster in running state. A single job of ten tables was given as an input to the mapper. The mapper extends the output as input to reducer in a single stage. One shall increase or decrease the amount of mappers and reducers in a single job.

Being the Hadoop instance is running in a cloud environment, having a single DNS, one shall easily scale up the physical and logical resources like storage, network and RAM (Random Access Memory). Results are collected using Amazon S3 Cluster, which needs to be connected after authorization from amazon. The allocated space from amazon is extended to end user.

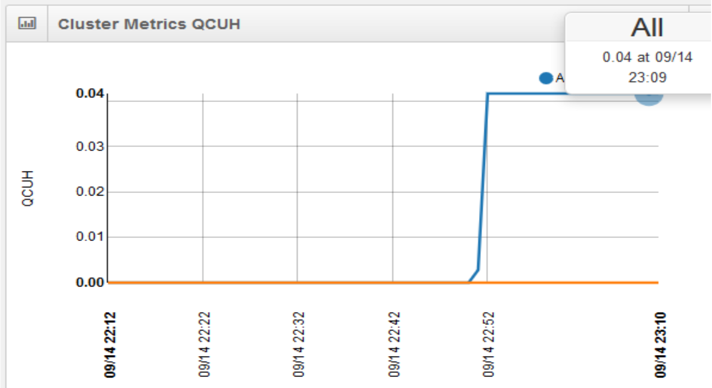


Fig-9: Processed Query

Fig-9 confirms the completion of single query that was accepted as an input. One shall conclude the time spent to run a single query for a single batch.

5. Conclusion and Future work

The experimental setup was designed for dedicated cloud to confirm information retrieved from Big data. The available Map Reduce algorithm is extended with a hummingbird algorithm to retrieve information from S3 cluster hosted on Amazon. A query using Hive was accepted as an input which is also possible using pig script.

A database having limited tables was selected as input, which shall be tested with cluster of information as an input to mapper as future work. Hosting multiple instances of Hadoop in distributed database shall be another area to explore to provide public services.

6. References

[1] Rahul Prasad Kanu , Shabeera T P , S D Madhu Kumar 2014- *Dynamic Cluster Configuration Algorithm in MapReduce Cloud*, International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 4028-4033.

[2] Mr. Kulkarni, N. N., Dr. Pawar V. P., Dr. K.K Deshmukh -2014 *Evaluation of Information Retrieval in Cloud computing based services*, Asian Journal of Management Sciences 02 (03 (Special Issue))

[3] Brian Hellig, Stephen Turner, rich color, long Zheng-2014- beyond *map educe: the next generation of big data analytics* HAMR.Eti.com.

[4] Ismail Hmeidi, Maryan Yatim, Ala' Ibrahim, Mai Abujazouh, 2014 - *Survey of Cloud Computing, Web Services for Healthcare Information Retrieval Systems*, International conference on Computing Technology and Information Management, Dubai, UAE.

[5] Anil Radhakrishnan and Kiran kalmadi -2013- *Big Data Medical engine in the cloud,* Infosys Lab Briefing ,Vol-11,No-1.

[6] Dr. Sanjay Mishra, Dr. Arun Tiwari 2013**-** *A Novel Technique for Information Retrieval Based on Cloud Computing,* International Journal of information technology.

[7] Yu Mon Zaw, Nay Min Tun 2013-*Web Services Based Information Retrieval Agent System for Cloud Computing.* International Journal of Computer Applications Technology and Research, Volume 2– Issue 1, 67-71.

[8] Gautam Vemuganti 2013- *Metadata Management in Big Data*, Infosys lab Briefing.

[9] Aaditya Prakash 2013-*Natured Inspired visualization of unstructured big data*, Infosys lab briefing, Vol-11, No-1.

[10] Xinxin Fan, Guang Gong,Honggang Hu-2011- *Remedying the Hummingbird Cryptographic Algorithm*, IEEE.

[11] Mosashi Inoue 2009- *image retrieval: research and use in the information retrieval,* National Institute of Informatics.

[12] Jeff Dean Google Fellow 2009- *Challenges in Building Large-Scale Information Retrieval Systems.*

[13] Tsungnan Lin, Pochiang Lin, Hsinping Wang,Chiahung Chen-2009- *Dynamic Search Algorithm in Unstructured Peer-to-Peer Networks*, IEEE.

[14]William Hersh -2008 *Future perspectives Ubiquitous but unfinished: grand challenges for information retrieval,* Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, Oregon, USA.

[15] Jeffrey Dean and Sanjay Ghemawat 2004-MapReduce: *Simplified Data Processing on Large Clusters*, Google.com.

[16]Mehran Sahami Vibhu Mittal Shumeet Baluja Henry Rowley 2003-*The Happy Searcher: Challenges in Web Information Retrieval,* google.com

[17]James Allan 2002-*Challenges in Information Retrieval and Language Modeling*, Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst

[18]Amit Singhal 2001- *Modern Information Retrieval: A Brief Overview* IEEE Computer Society Technical Committee on Data Engineering.

[19] tp://www.internetlivestats.com

[20] https://api.qubole.com

[21]Dr. Piyush Gupta, kashinath Chandelkar 2012- *The Need and Impact of Hummingbird Algorithm on Cloud based Content Management System*, vol-2, issue-12, IJARCSSE journal.