AMAZON HADOOP FRAMEWORK USED IN BUSINESS FOR BIG DATA ANALYSIS

Ankush Verma*, Dr Neelesh Jain
* Research Scholar Pacific University Udaipur
Professor SIRT Bhopal

DOI: 10.5281/zenodo.801272

KEYWORDS: Amazon Elastic MapReduce, Amazon Elastic Compute Cloud, Hadoop Distributed File System, Amazon Simple Storage Service, Amazon web service.

ABSTRACT

The Amazon MapReduce programming model, introduced by Amazon, a simple and efficient way of performing distributed computation over large data sets on the web especially for e-commerce. Amazon EMR work on Master/Slave Architecture using Amazon EMR for map and reduce big data. Amazon EC2 use cloud computing is a central part of designed web service that provides resizable compute capacity in the cloud. Here we also discuss about the Benefit and limitation of using Amazon EMR. Amazon S3 use easy to store and retrieve any amount of data on web. A Amazon clusters is a set of servers that work together to perform tasks work on distributing database with the servers in parallel. Amazon EMR work in the concept of Hadoop, nodes and cluster. Amazon EMR Model used for analysis big data to increase their business efficiency.

INTRODUCTION

Amazon use Amazon Elastic MapReduce (Amazon EMR) for its e-commerce growing business process and analysis vast amount of data also simplifies big data processing, providing a managed Hadoop framework that makes it easy, fast, and cost-effective for to distribute instance. Amazon EMR work as Master/Slave Architecture like Hadoop basically Amazon EMR works across a cluster of virtual servers running in the Amazon cloud. The cluster is managed using an open-source framework called Hadoop.

Amazon EMR securely and reliably handles your big data use cases, including log analysis, web indexing, data warehousing, machine learning, financial analysis, scientific simulation, and bioinformatics. Hadoop clusters running on Amazon EMR use EC2 instances as virtual Linux servers for the master and slave nodes. It automatically distributes incoming application traffic across multiple Amazon EC2 instances. It enables you to achieve even greater fault tolerance in your applications. Amazon S3 for bulk storage of input and output data and Cloud Watch to monitor cluster performance and raise alarms, can also move data into and out of DynamoDB using Amazon EMR and Hive. Amazon EMR supports many tools on top of Hadoop that can be used for big data analytics and each tool has its own interface.

AMAZON EMR ARCHITECTURE & WORKING

Amazon is flexibility helps organizations mix and match architectures in order to serve their diverse business needs [3]. Amazon EMR Model for analysis big data. The Amazon Elastic MapReduce provides a managed, easy to use analytics platform built around the powerful Hadoop framework that is used by large companies. It also work on Master/Slave architecture whereas master nodes remain master nodes and become part of the master instance group. Slave nodes still run HDFS and become more core nodes and join the core instance group Architectures executes in the following sequence is [7]:

1. A request is sent to Amazon EMR to start a cluster.
2. Amazon EMR creates a Hadoop cluster with a master instance group and core instance group.
3. The master node is added to the master instance group.
4. The slave nodes are added to the core instance group.
5. The legacy cluster runs on a cluster consisting of a single master node and the specified number of slave nodes.

At any time we can add nodes and remove task nodes which hold the Hadoop Distributed File System (HDFS) at any time to increase your processing power and increase HDFS storage capacity. Additionally, we can use Amazon S3 or use Elastic Map Reduce File System (EMRFS) along with or instead of local HDFS. Nodes process Hadoop jobs, but that do not maintain HDFS.

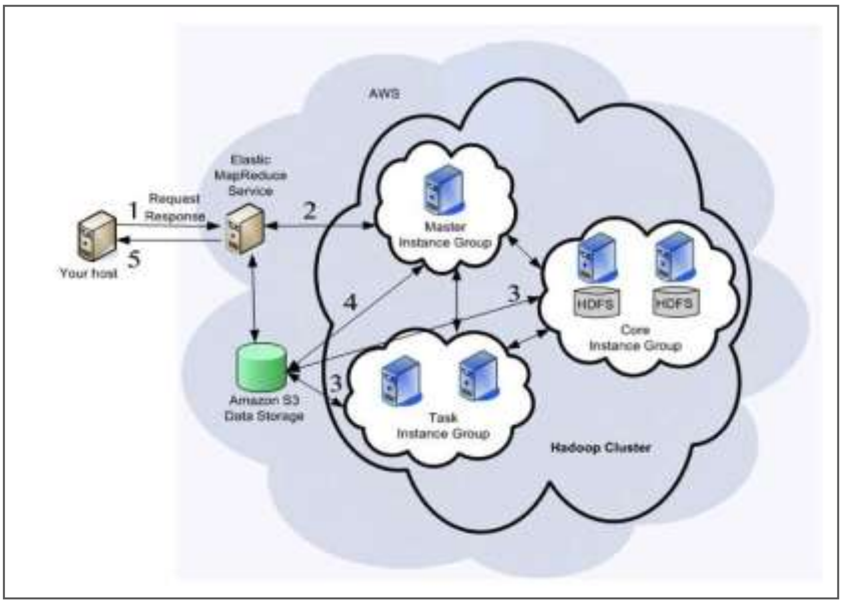

Figure 7.1: Amazon EMR Architecture

Working

Amazon EMR has made enhancements version of Hadoop work seamlessly with Amazon web service (AWS). It is use to managed Hadoop clusters on Amazon Web Services. A Hadoop clusters we know that is a set of servers that work together to perform tasks by distributing the work and data among the servers in parallel. Amazon EMR work in the concept of Hadoop, nodes and cluster.

AMAZON HADOOP

Amazon use the concept of Hadoop an open source framework supports massive data processing across a cluster of mater and slave across multiple servers. Amazon use MapReduce for processing the big data and HDFS for store data.

MapReduce - Amazon EMR's flexible framework reduces large processing problems and data sets into smaller jobs and distributes them across many compute nodes in a Hadoop cluster [10]. A custom Hadoop MapReduce application and run it on Amazon EMR to analyze and work with public large amount of big data sets. To lower the entry barrier for performing MapReduce computations in the cloud, Amazon Web Services provides Elastic MapReduce. The service takes care of resources, configuring and tuning Hadoop, staging data, monitoring job execution, maintain data failure etc.

HDFS - HDFS distributes the data across servers in the cluster, storing multiple copies of data on different servers to ensure that no data is lost if an individual server fails. Amazon S3 is also called HDFS in Amazon the functionality of Amazon S3 is same as HDFS which are used in apache Hadoop.

Jobs and Tasks - In Hadoop a job is a unit of work. Each job may consist of one or more tasks whereas job work as master and task as slave. In Amazon EMR, a step is a unit of work that contains one or more Hadoop jobs. Amazon EMR also provides the option to run multiple instance groups so that you can use on-demand instances in one also together with Instances in another group to have your jobs completed faster and for lower costs. There are many types of logs written to the master node. Apache Hadoop writes logs to report the processing of jobs, tasks, and task attempts. At any time we can add and remove task nodes that can process Hadoop jobs.

AMAZON EMR NODES & CLUSTER

Nodes - A Hadoop cluster created by Elastic MapReduce can be composed of three kinds of nodes [2]:

- **Master Node -** The master node which is unique and acts as a meta data server for HDFS (by running the NameNode service) and schedules MapReduce tasks on other nodes (by running the JobTracker service), It track the status of each task in the clusters.
- **Core Nodes -** Core nodes provide data storage to HDFS (by running the DataNode service) and execute map and reduce tasks (by running the TaskTracker service) also maps to a Hadoop slave node.
- **Task Nodes -** Task nodes execute map and reduce tasks but do not store any data. This task node is optional.

Cluster - In Amazon EMR the cluster is a set of virtual servers running as Elastic Compute Cloud (EC2) instances some work has to done by cluster. Computing resources called *nodes*, which are organized into a group called a *cluster*. Clusters are a process a set amount of data and then terminate when processing is complete.

Each cluster runs an Amazon Redshift engine and contains one or more databases. An Amazon Redshift cluster will automatically detect and replace a failed node in your data warehouse cluster. Amazon Redshift places your existing cluster into read-only mode, provisions a new cluster of your chosen size, and then copies data from your old cluster to your new one in parallel [5].

Each cluster has a leader node and one or more compute nodes. The *leader node* receives queries from client applications, parses the queries, and develops query execution plans. The leader node then coordinates the parallel execution of these plans with the compute nodes and finally returns the results back to the client applications [8]. *Compute nodes* execute the query execution plans and transmit data among themselves to serve these queries. The intermediate results are sent back to the client applications.

AMAZON EC2

Cloud Architectures are designs of software applications that use Internet-accessible on-demand services. Applications built on Cloud Architectures are such that the underlying computing infrastructure is used only when it is needed. Amazon Elastic Compute Cloud (Amazon EC2) is a central part of designed web service that provides resizable compute capacity in the cloud. Amazon EC2's allows obtaining and configuring capacity with minimal friction. It provides complete control of your computing resources and system run on Amazon's computing environment [4].Amazon EC2 has following features:

- Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use.
- Amazon Elastic MapReduce provides a hosted Hadoop framework running on the web-scale infrastructure of Amazon EC2 and allows creating customized JobFlows. JobFlow is a sequence of MapReduce steps [1].
- An Amazon EC2 launched with a choice of two types of storage for its boot disk or root device. The first option is a local "instance-store" disk as a root device. The second option is to use a network-based called Elastic Block Storage (EBS) volume as a root device which can be attached to running instances.
- Amazon Cloud Watch is a web service that provides real-time monitoring to Amazon's EC2 customers on their resource utilization such as CPU, disk and network.
- Static IP addresses for dynamic cloud computing known as *Elastic IP addresses.*
- Amazon EC2 is integrated with most AWS services such as Amazon S3, Amazon Relational Database Service and Amazon Virtual Private Cloud (Amazon VPC) to provide a complete, secure solution for computing, query processing, and cloud storage across a wide range of applications.

AMAZON S3

Amazon Simple Storage Service (Amazon S3) provides access to reliable, fast and inexpensive data storage infrastructure. It is designed to make web-scale computing easy to store and retrieve any amount of data, at any time, from within Amazon EC2 or anywhere on the web [9]. Amazon S3 stores data objects on multiple devices allows concurrent read or write access to these data objects by many separate clients. Customers can also grant access to their Amazon S3 data to all AWS Accounts or to everyone.

Amazon S3 adds the data objects similarly when an object is deleted from Amazon S3, removal of the mapping from the name to the object starts immediately. Once the mapping is removed, there is no remote access to the deleted object and is generally processed across the distributed system within several seconds [6]. Amazon EMR provides custom libraries to move data in and out of Amazon S3 that is Amazon EMR clusters is integrated with Amazon S3, which means that we can store input and output data in Amazon S3 on the cluster in HDFS, or a mix of both. Amazon S3 can also have multiple clusters accessing the same data simultaneously. Data is normally communicated from one step to the next using files stored on the cluster's Hadoop Distributed File System (HDFS). Data stored on HDFS exists only as long as the cluster is running. When the cluster is shut down, all data is deleted.

For store large data Amazon S3 create a bucket in one of the AWS regions on here we can add any number of objects to the bucket according to our storage requirement. The buckets and objects are resources and Amazon S3 provides APIs to manage them [6]. Whereas Amazon S3 buckets names are globally unique, regardless of the AWS region in which you create the bucket.

AMAZON HADOOP MAPREDUCE BENIFITS & LIMITATIONS

Benefits

- Combine with leading BI Tools. It supports business analysis to increase revenue.
- Amazon EMR delivers fast query and I/O performance for virtually any size dataset by using columnar storage technology while parallelizing and distributing queries across multiple nodes.
- As a managed service automation is provided for most of the common administrative tasks which is associated with provision, configuring, monitoring, backup and secure a data warehouse making it easy to manage and maintain.
- Store historical stock trade data.
- Analyze social trends for better analysis big data.
- Analyzing large data sets requires significant capacity that can vary in size based on the amount of input data and the analysis when required.
- Applications can easily up and down base on demands.
- As requirement changes we can easily resize ours situation like horizontally or vertically on AWS to meet your needs without having to wait for additional hardware.
- Get flexible computing infrastructure on a world class access to different geographic regions that as AWS offers.
- Easy-to-use cloud computing platform which need to handle big data analytics workload such as real-time streaming, NoSQL, data warehousing, storage, other analytics tools and data workflow services.
- Combine with other Web Services including S3 as an alternative to HDFS.
- Cost is one of the most in developing and deploying an e-commerce application which can increase the need for hardware and bandwidth. The amount of resources and service that you actually need have to pay for that only.

Limitation

- Every time when create a job flow with more than 20 instances, the creation fails most of the time Amazon MapReduce work with less than 20 instances. It is failed to create a job flow with a large number of instances.
- Amazon EMR is not open source system, so you have limited control over the code.
- The increased of latencies is typical for EMR jobs use data stored in S3 which is process on EC2, moving of data from S3 to EC2 takes time.
- Amazon EMR does not support the latest version of Hadoop, if your application requires using the latest features and upgrade features of Hadoop, EMR may not be the best option and is not easily acceptable by Amazon.

- When data stored in S3 but processed on nodes of the EMR cluster i.e. lack of data locality when it is hit by the small-files problem there is an issues decompression performance.
- For storing large amounts of data on an EC2 cluster quickly it's become expensive to manage.
- Some pig script running on EMR in a workflow, so we need to monitor the status of the job to determine when it is finish and then next it continues. While Amazon needs to monitor all jobs. Problem is for workflow engine which manage dependency, alert and monitoring.
- We found that cluster is busy continuously on doing jobs so we had to build wrappers to periodically shutdown and startup clusters. Thus it creates a problem of starting, configuring and stopping the cluster.

CONCLUSION

It supports ecommerce business analysis to increase revenue and get the knowledge of area of interest for the customers. It is business application architecture over client/server solutions, and now to loosely coupled web services and service-oriented architectures. Amazon support web services individually for all users according to the cost and use. Sometimes Amazon fails to support the large data and unstructured data like images, videos. Amazon Hadoop is a combination of cloud, storage, security and process on data having Technique for each function. It use Amazon Elastic MapReduce model for analysis big data which have some implement on Hadoop for e commerce.

REFERENCES

1. Jinesh Varia, "*Architecting for the Cloud: Best Practices*", January 2011.
2. Pierre Riteau, Ancuta Iordache, Christine Morin, "*Resilin: Elastic MapReduce for Private and Community Clouds*", 13 Oct 2011
3. Jinesh Varia/Sajee Mathew, "*Overview of Amazon Web Services*", January 2014.
4. https://aws.amazon.com/ec2/.
5. https://aws.amazon.com/blogs/aws/amazon-redshift-spectrum-exabyte-scale-in-place-queries-of-s3-data/
6. Developer Guide, "*Amazon Simple Storage Service*", API Version 2006-03-01, Pages 1-5.
7. Management Guide, "*Amazon Elastic MapReduce*", API Version 2009-03-31, Pages 13-14.
8. Management Guide, "*Amazon Redshift*", API Version 2012-12-01.
9. User Guide for Microsoft Windows, "*Amazon Elastic Compute Cloud*", API Version 2015-04-15, Pages 573-574.
10. Developer Guide, "*Amazon Elastic MapReduce*", API Version 2009-03-31, Pages 6-10, 300-306.