Scientific Journal of Impact Factor(SJIF): 4.14

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Special Issue for ICPCDECT 2016,Volume 3 Issue 1

EFFICIENT IMAGE AND VIDEO UPLODING IN CLOUD COMPUTING USING MAPREDUCE AND SPARK

Chaitanya Deshpande^1 ,Divya Apte^2 ,Pooja Gadiwan^3 ,Nikita Musale^4
^1Computer engineering, ^2Computer engineering,, ^3Computer engineering, ^4Computer engineering,
^1chaitanya.deshpande101294gmail.com,^2divya.apte94@gmaul.com,^3pooja.gadiwan1@gmail.com,^4nktamusale@gmail.com

Abstract —- *Hadoop is a Open source framework used to store and process Big data in distributed and parallel environment. For data processing in Hadoop framework we use one of feature called Map-Reduce used for sorting and filtering input data and then performs summary operation as output. As use of internet and social networking services is growing rapidly multimedia data transmission has become simpler. Techniques like Transcoding and Transmoding are replaced by Hadoop mapreduce used for data processing and reduce the burden on computing infrastructure as data increases. Map reduce limitations are speed and fragmentation and to overcome this, article focuses on new platform apache spark which is more faster than map reduce and also provides more functionality. And thus spark is used as alternative for Hadoop map reduce.*

Keywords- Spark , HDFS, Cloud Computing ,Image uploading, Map reduce

INTRODUCTION

Users access multimedia objects not only from traditional desktops but also from mobile devices, such as smart phones and smart pads, because of its ease, portability and many other features. Thus use of internet and social networking applications are widely used. There are various methods and algorithms used for multimedia data transmission over internet through various devices. There are many factors to be considered while implementing such methodologies like security, speed, reliability etc.

Hadoop Map reduce is one of the emerging technology used for managing big data. As For example Facebook, twitter process a very huge multimedia as well as text data every day and storing and processing becomes complex. Map reduce technique is used to overcome the limitations of transcoding transmoding technique used previously[1]. But also there are few limitations of map reduce like speed, fragmentation, Generality etc. And so this study includes use of new platform Apache spark which provides more features over Hadoop map reduce. Proposed modules consist of HDFS, Hadoop, Map reduce, spark and cloud environment.

1. Hadoop:

Hadoop is an open source software framework written in java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. The core of Hadoop consists of storage part HDFS (Hadoop Distributed File System) and a processing part Map Reduce.

1.1 HDFS-Hadoop File System

HDFS is based on java and provides scalable and reliable primary storage for Hadoop applications. It is designed to run on commodity hardware. HDFS is highly fault tolerant and deployed on low cost hardware. It is subproject of Apache Hadoop[1].

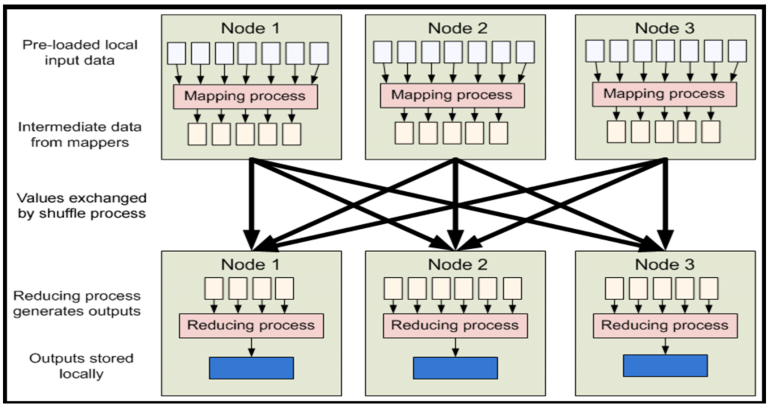

Figure 1. HDFS

1.2 Map Reduce-

Map reduce is programming model which is used for data processing stored in HDFS. It works in distributed and parallel manner because of which execution becomes faster.[1] It handles automatic scheduling, communication, synchronization which has ability to handle huge data and fault tolerance.

There are two main functions used in Map Reduce
1) map()-performs filtering operations

2) reduce() – performs summary operation.

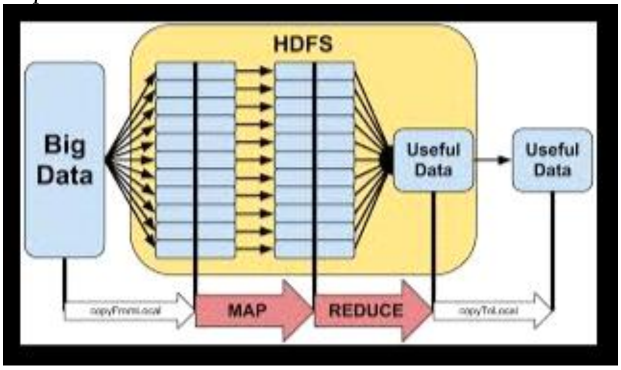

Figure 2.Map Reduce

1.3 Spark

Apache Spark is an open-source cluster computing framework. In contrast to Hadoop's two-stage disk-based Map Reduce paradigm, Spark's in-memory primitives provide performance up to 100 times faster for certain applications. Spark provide 100times faster than Hadoop[5]. By allowing user programs to load data into a cluster's memory and query it repeatedly. Sparks including Hadoop Distributed File System. In this scenario, Spark is running on a single machine with one executor per CPU core. Spark provide a stack of libraries including SQL for machine learning. Spark is used at wide range of organizations to processing large data[3].

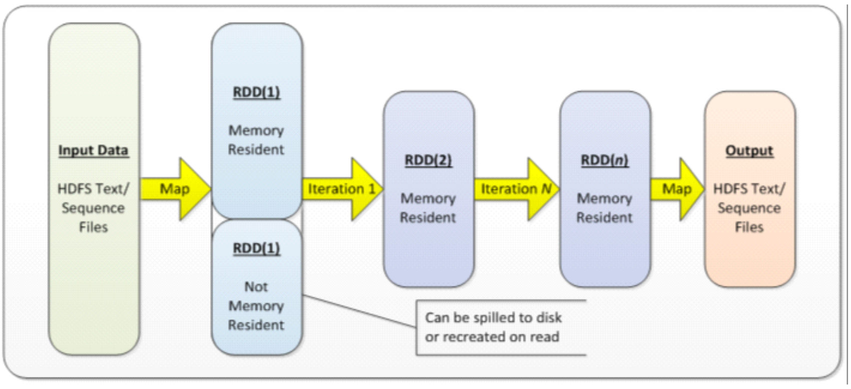

Figure 3.Spark architecture

Proposed System:

In this project our task is to convert the image in cloud so that there will not be any overhead problem in the infrastructure. Basic aim is to convert the image and to maintain the quality. As we are trying to process the real time data from internet there are more chances of failure of internet. This requires more amount of time to process and network failure is to be considered.

To avoid these problems of time consumption and network overhead normal technique of image conversion is used by overcoming all the disadvantages of cloud environment. This technique will scale the image by maintaining the quality of the multimedia data concurrently while accessing the same data as well.

In this model we will be using Hadoop Distributed file system to store the data in the Hadoop cluster[3]. Whatever input we will be giving will be processed in accordance with the Map reduce Function where the data will be sent in different clusters. For scaling the image and converting it in its normal form some inbuilt libraries are used for example here we are using java interface library. The data which is processed is stored in Hadoop Distributed File System[6].

A. Image conversion function:

In this module actual implementation is done where the real time data is processed so as to maintain the quality of image in a parallel computing environment.

Firstly input is taken from the user it can be any image or video who has the size more than 50GB. This input will be given as a key value pair from client to namenode. Namenode will take the input in a specific format like text input or as a sequence input and will assign the task to datanode. Secondary namenode is used to keep the data replication. Namenode will assign the task through jobtracker and then job tracker will assign task to tasktracker.[6]

The input format which is given will be passed will be passed on to mapper as a set of key value pair. Image coversion module will scale and resize the image in specific format so that it will be compatible on all the devices.

In Hadoop data is stored in chunks so that chunks of mapper class will be combined and shuffled to Reducer class for this we are using the java interface library where all the task of mapper and reducer is done.[5] In this number of input splits will be equal to number of mappers and number of reducers will be equal to the output. The input format which has been used the same output format will be generated and stored in Hadoop File System.

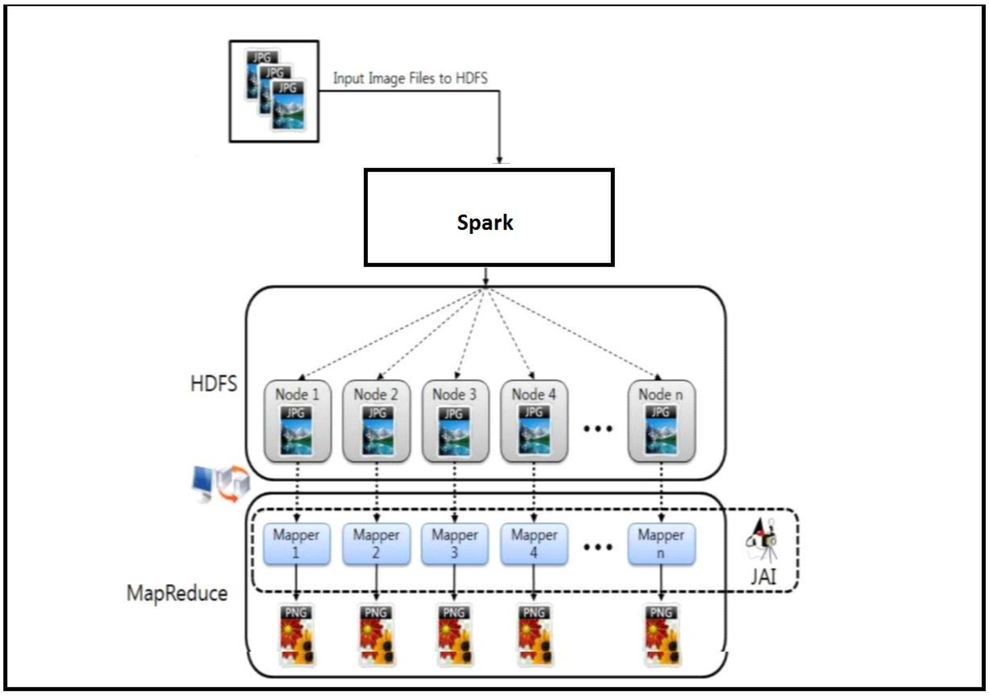

Fig4.architecture Diagram

B. video conversion function:

In this we are going to upload a video in hdfs on spark environment where the working will be the same for all the multimedia processing. Data will be efficiently handled and processed so as to maintain the quality of the uploaded video. Spark environment is very efficient to handle big data and processing and analysis of data is very fast in this environment. In spark we use RDD's i.e Resilient distributed datasets. In this the videos will be splitted in a key value pair like the same in MapReduce function.[2] It avoids data bottleneck so that all the overheads will be handled and will make the system work efficiently. In spark for storing we are using HDFS and spark environment acts as an API. RDD's are unit of data in spark. It relies on JVM. In this we can do the scala programming. As there is a map reduce function in Hadoop same we will be there in Spark only first we need to filter the data and Map function and Reduce Function is implemented. In spark operations are performed in two ways i.e on disk and in memory. For disk there is 100 times faster than memory whereas memory is 10times faster.[5]

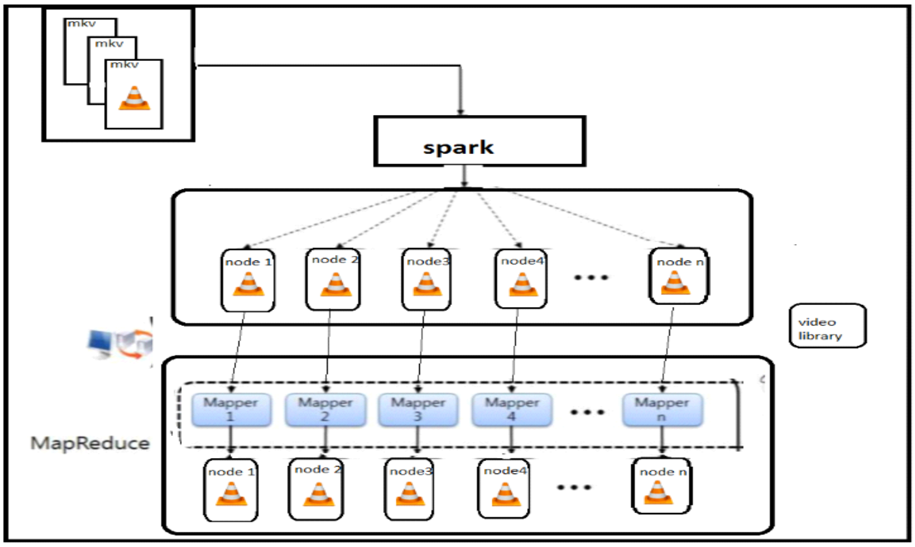

Fig 5.video architecture

Conclusion:

Hence, we will be implementing the map reduce function on video and image uploading efficiently so that the quality of the multimedia data will be maintained and for this implementation we will be using spark environment as it is the new technology to efficiently handle all the operation of big data.

REFERENCES

*[1] Hyeokju Lee Myoungjin Kim, Joon Her, and Hanku Lee" Implementation of MapReduce-based Image Conversion Module in Cloud Computing Environment" Konkuk University
Seoul Korea (2014)*

[2] Ardiana Garcia and Hari Kalva "Cloud Computing for Mobile Video Content Delivery"

[3] Hari Kalva, Aleksander "Parallel Programming for multimedia Applications"

[4] Reynold S. Xin,Joseph E. Gonzalez,Michael J. Franklin, "GraphX: A Resilient Distributed Graph System on Spark " Ion Stoica AMPLab, EECS, UC Berkeley

*[5] A. Radenski L. Ehwerhemuepha, K. Anderson Schmid "From in-disk to in-memory big data with Hadoop: Performance experiments with nucleotide sequence data"
College of Science and Technology, Chapman University, Orange, California, U.S.A. ,ABDA'15*

*[6] Chao-Hsuan Shen, Patrick Loomis,"Spark Application on AWS EC2 Cluster" ImageNet dataset classification using Multinomial Naive Bayes Classifier John Geevarghese
Tasks" University of Virginia*