**Image Processing for duplicate image Removal on Hadoop Platform**

Priya B Patel1, Meikang Qiu2

1M.S research student in Computer Information Science, Harrisburg University of Science & Technology, PA, USA

2Professor, Harrisburg University of Science & Technology, PA, USA

***Abstract***

**Apache Hadoop is a system that takes into consideration the dispersed preparing of enormous informational collections crosswise over bunches of item PCs utilizing the straightforward programming model. Hadoop speaks to a Java-based appropriated registering structure that is intended to help applications that are actualized by means of the MapReduce programming model. It contains two centre parts which are HDFS (Hadoop Distributed File System) and MapReduce. MapReduce is a programming model and a related usage for handling and producing enormous informational collections with a parallel, appropriated calculation on a bunch. On Social Media sites like Facebook, Twitter and so on have a few clients who offer and transfer pictures and recordings. At some point Image record with same name, same organization, same goals and same size is shared by numerous clients every one of these Images are put away in the Data Base even though it conveys an equivalent property. This doesn't look great as copy documents consumed unnecessary space in File System.**

**Keywords:**

**Hadoop Distributed File System, MapReduce Algorithm, Hadoop Platform, Virtual Box**

1. **Introduction**

Online networking, the present greatest impact and utilization. Individuals used to transfer pictures, post statues, transfer recordings and some more. Where are every one of these records being spared? The most serious issue or we can say vital issue is to deal with every one of these records and recover them when required. There are several internet-based life and a huge number of client accounts on it. Conquer this issue has given crisp and jejune advancements to the PC world. One of them is Image Processing. There are loads of copy pictures accessible on Internet and they are none of utilization. Evacuate them adequately and cautiously can give us colossal empty space and speed in surfing. The Photos application is one of Facebook's most prevalent highlights. Exceptional, clients have transferred more than 15 billion photographs which makes Facebook the greatest photograph sharing site. For each transferred photograph, Facebook creates and stores four pictures of various sizes, which means a sum

of 60 billion pictures and 1.5PB of capacity. The present development rate is 220 million new photographs for every week, which means 25TB of extra stockpiling expended week by week. At the top there are 550,000 pictures served every second. These numbers represent a critical test for the Facebook photograph stockpiling framework.

**What is Hadoop?**

Apache Hadoop is an open-source programming system written in Java. It is principally utilized for capacity and handling of enormous arrangements of information, also called huge information. It includes a few segments that permit the capacity and handling of huge information volumes in a bunched situation.

Hadoop Distributed File System – Abbreviated as HDFS, it is principally a document framework like a significant number of the officially existing ones. Nonetheless, it is likewise a virtual document framework.

Hadoop MapReduce – MapReduce is for the most part the programming part of Hadoop that permits handling of huge volumes of information.

1. **Literature Survey**

In literature [1] it separates the SPOF existing in fundamental centres of Hadoop and proposes a metadata replication-based response for engage Hadoop high availability. The course of action incorporates three huge stages: in instatement organize, each reinforcement/slave centre point is enrolled to dynamic/basic centre and its hidden metadata, are compensated for wasting time with those of dynamic/basic centre point; [2] displays a parallel planning model reliant on Hadoop arrange for colossal scale pictures getting ready, which hopes to make usage of the advantages of high relentless quality and high flexibility of Hadoop circled organize for scattered memory and coursed figuring, so as to achieve the explanation behind snappy treatment of gigantic scale pictures.

In literature [3] separates the SPOF existing in fundamental centres of Hadoop and proposes a metadata replication-based response for enable Hadoop high availability. The game plan incorporates three huge stages: in instatement arrange, each reinforcement/slave centre point is enrolled to dynamic/fundamental centre point and its hidden metadata, (for instance, version report, record system picture) are compensated for wasting time with those of dynamic/basic centre point; In replication organize which is the inside phase of our answer, the runtime metadata, (for instance, excellent errands, lease states) for failover are copied

This paper [4] shows a parallel planning model subject to Hadoop organize for gigantic scale pictures getting ready, which hopes to make use of the advantages of high unfaltering quality and high adaptability of Hadoop coursed arrange for scattered memory and circled figuring, so as to achieve the purpose behind fast treatment of colossal scale pictures. This model is executed using PC bundle program dependent on Hadoop stage to collect scattered, parallel getting ready of tremendous scale pictures, appeared differently in relation to standard single-centre point picture setting it up has advantages of fast speed and high profitability. Writers [5] show that such duplication is to make certain gigantic, yet that not a wide scope of picture change explored in past composing are evident in web data. Ejection of close duplicates from an aggregation is ridiculous, yet, we recommend that they be removed from sets of answers. We survey our framework for customized recognizing confirmation of close duplicates during inquiry appraisal and demonstrate that it has ensure as a convincing segment for the leading body of close duplication eventually.

In this paper, [6] a self-loader approach is proposed, which can be used to end possibly certified duplicates in science papers. The guideline responsibilities of the procedure are the following: First, it might be used for the recognizable proof of replicated zones between pictures, yet furthermore for the area of duplicated zones that occur inside a singular picture. In this paper [7] creators associated clustering on the BISAG Dataset. The dataset is a social affair of GeoTIFF pictures each one around 2 MB in size. Bundling was done using the past methodology, which is diverged from gathering without using HIPI. The assessment is associated again and again, each time with different number of pictures to see the modifications in execution when the data size creates. The work is done by making a Virtual Machine (VM) holding Ubuntu OS and Hadoop Standalone Installation with 4 Map Tasks and 1 Reduce Task. They submitted 3 CPUs and 8GB of RAM for the VM, however the encouraging OS is a workstation that has 8CPUs and 16GB RAM.

The proposed framework [8] is gives a way to deal with setup bundling count reliant on the geometric division figuring technique. Here for each datum cases a comparability is enlisted to outline data gatherings. To get this idea various sorts of grouping plans are concentrated, for instance, c-infers, soft c-suggests and others. Outcomes of proposed [9] AMF figuring and existing methods explored the ImageNet dataset tested Hadoop-1.0.3 gathering Results of filtering systems are evaluated the extent that estimations can envision PSNR, MSE, and time capability.

The goal of this paper [10] is to design a structure block for a largescale close duplicate picture web searcher — a single centre point system that significantly improves as far as possible over existing methodologies without growing inquiry time, and for the most part basically, practices search with high conviction, i.e., low false positive rate. The counterfeit positive issue isn't undeniable in customary substance-based picture search as a result of the nonattendance of an objective and unambiguous significance of visual closeness.

1. **Ease of Use**

Apache HDFS or Hadoop Distributed File System is a block organized document framework where each record is partitioned into block of a pre-decided size. These blocks are put away over a cluster of one or a few machines. Apache Hadoop HDFS Architecture pursues a Master/Slave Architecture, where a cluster involves a single Name Node (Master Node ) and the other nodes are Data Nodes (Slave Nodes). HDFS can be conveyed on a broad spectrum of machines that help Java. Even though one can run a few Data Nodes on a single machine, yet in the practical world, these Data Nodes are spread crosswise over different machines.

A green and black text

Description automatically generated

Fig.1, Hadoop Architecture

**Hadoop Distributed File System:**

One of the vital pieces of a Hadoop bunch is the Hadoop Distributed File System (HDFS). Roused by the Google File System, its motivation is to give an issue tolerant capacity structure fit for holding a lot of information, take into consideration quick access of said information, and give an approach to MapReduce to perform calculations on a similar area as the information.

Big Data alludes to the monstrous measures of information gathered after some time that are hard to break down and handle using regular database the board devices. Because of the expanding measure of information that is getting to be accessible consistently and quickened development of data innovation, one can't give an unmistakable meaning of the size and size of enormous information. In any case, a multi-terabyte informational indexes (every terabyte = 1000 gigabytes) to multi-petabytes (every petabyte = 1000 terabytes) are considered as Big information.

1. **Software Requirement**
2. Virtual Machine
3. Any Linux Operating System, i.e. Redhat, Centos, Ubuntu, Lubantu, etc
4. Hadoop Cluster setup with Min 6 GB RAM, 16 GB Harddisk
5. Master-slave configuration (Min one Master 1 slave)
6. Eclipse IDE with Hadoop Libraries Installed

**Configuration:**

Core-site.xml

Hdfs-site.xml

Mapred-site.xml

1. **Architecture detailed Design**

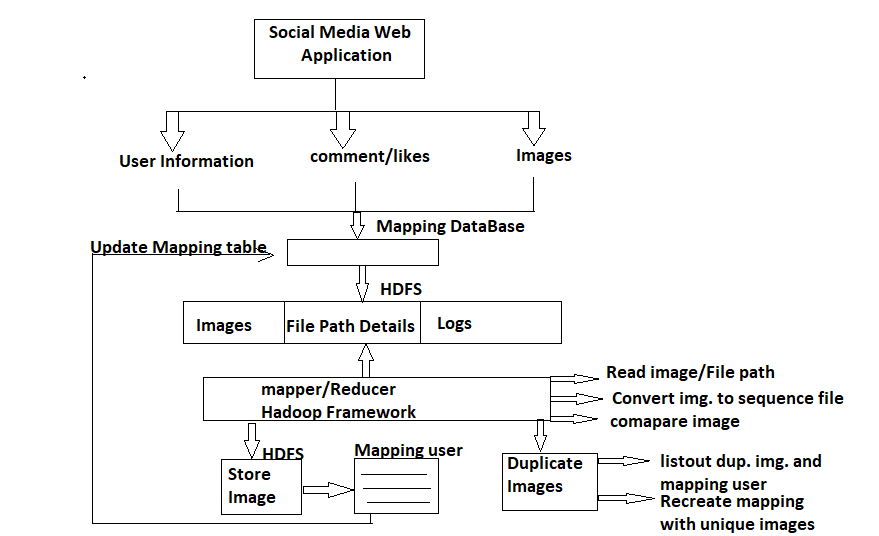
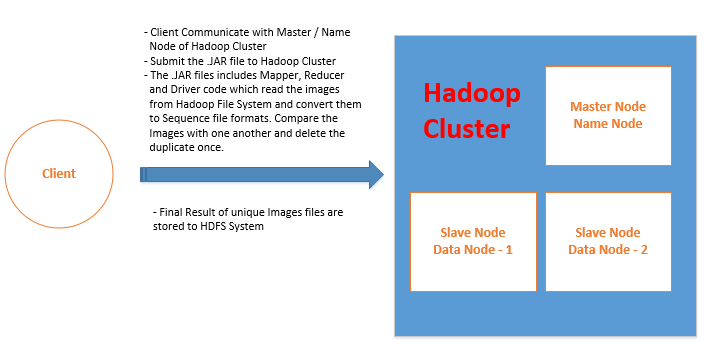
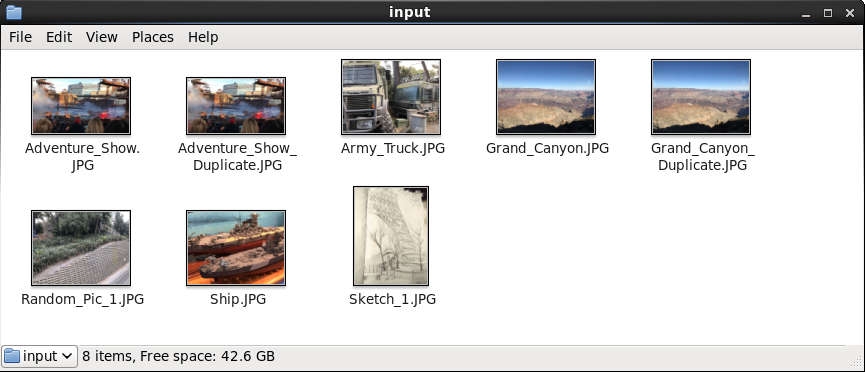


Fig.2 Detailed Design



1. **Experiments Setup**

Step 1: we took a certain number of Input Images and stored it into Hadoop File System. Here we used Java Map reduce code.



Step 2: All the Image which we took, we converted those in sequence file.

As Hadoop MapReduce structure was intended to store and process enormous documents, we are utilizing Sequence record organization to change over all the picture records (little records) into one single huge document of paired document type for handling in the MapReduce calculation.

A Sequence document is a level record where the key worth sets of the yield document will be as a double information type.

When we get the Sequence record with all the paired information, we can utilize MD5 calculation to produce an extraordinary key for each picture and contrast this key all together with locate the repetitive documents.

**Method Map**

The Mapper usage, TextinputFormat accepts the balance an incentive as key and whole way of the picture line by line as worth.

Whole worth is perused in the support and composed back to bytearray yield stream and radiates the way of the pictures as key and the whole record contained in bytearray yield stream as worth.

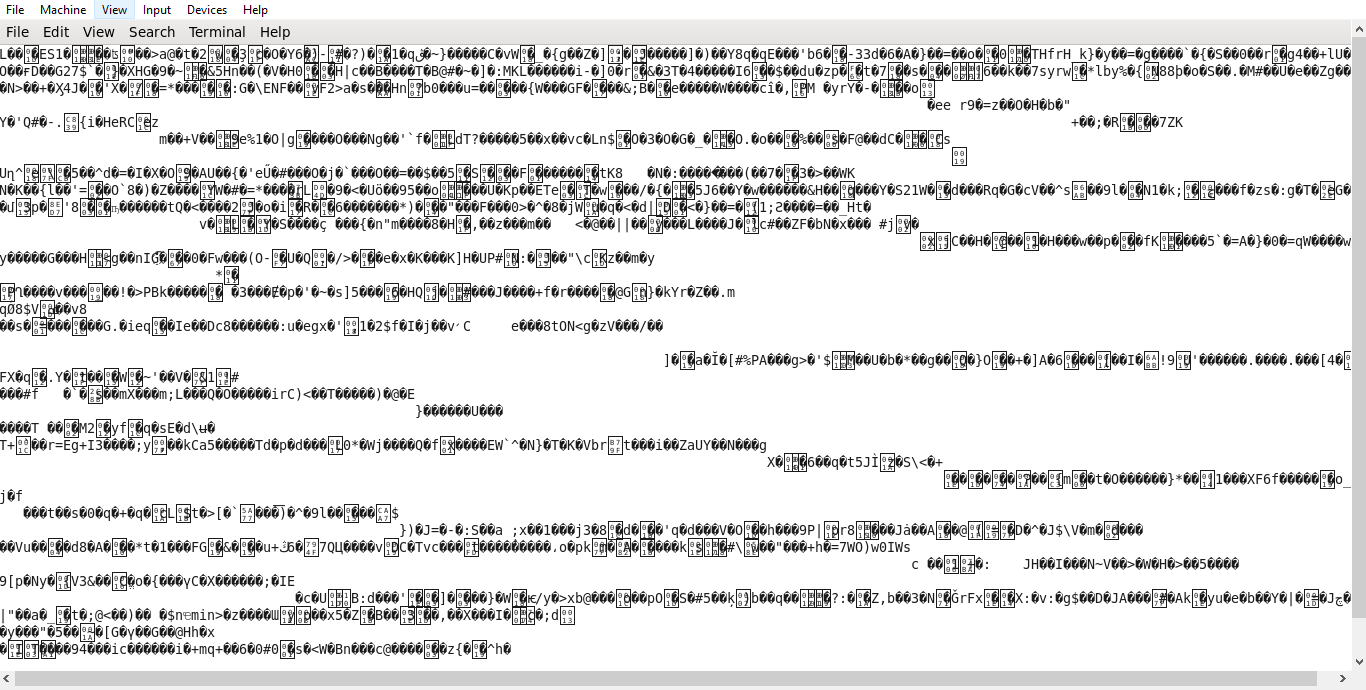
**Main Method**

In the main technique, we characterize succession arrangement class which takes the key and qualities from mapper and stores them as a grouping record with every one of the sets of keys and qualities in hdfs.

From the above program, we can change over a Binary picture document into sequence file format.

Next, we can refer beneath Map Reduce program to remove duplicate records from the above sequence document as an info and to store the names of one of a kind picture records in the output format.

This image is showing the output of the program execution. Its nothing but a sequence file output of the saved image files.



Step 3: Comparison of sequence file with one another.

Step 4: Identified the duplicate Images.

Step 5: Removed the duplicate images from Hadoop File System.

The Mapper code will peruse the twofold picture information of all the picture records and will make MD5 string ( code ) for each document. It will pass this information to the reducer where the key will be the MD5 string while the worth will be the filename. In this manner, all the indistinguishable pictures will be gathered by the Hadoop system.

The key here is the md5 hash while the values are all the picture records that are related with it. for each md5 esteem, we have to take just one record (the first). In the outcome document, the key will be again the picture record way.

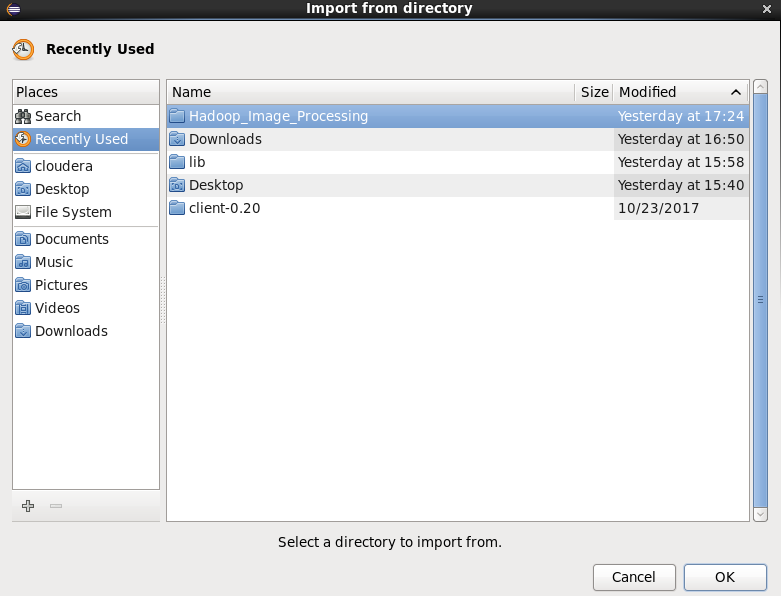
1. **Execution Procedure**

We have already written two programs.

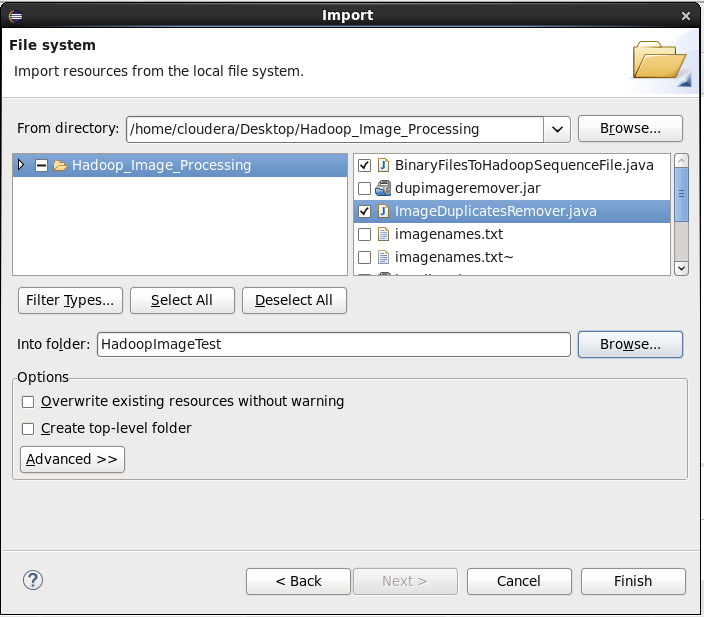
1. BinaryFilesToHadoopSequenceFile

2. ImageDuplicateRemover

So, first go on the new project we have created “HadoopImageTest”. Right Click on it and select Import. Now, from the file system choose a directory in which our program is saved. Open that directory and click OK.

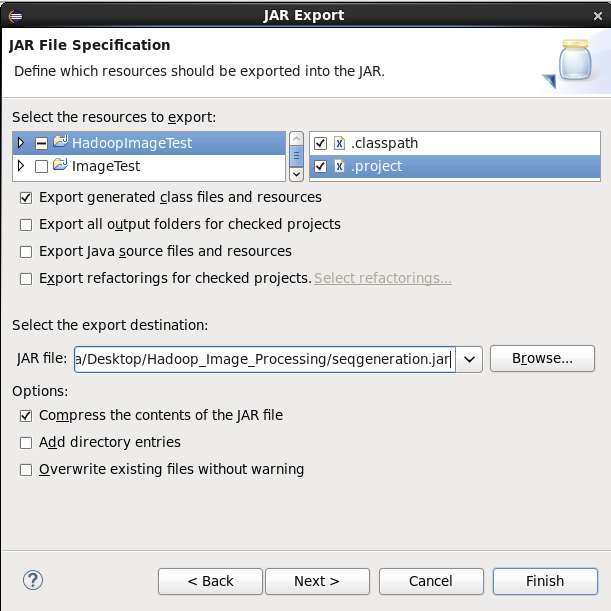


Now select both the files we needed for this project and name the project and Finish.

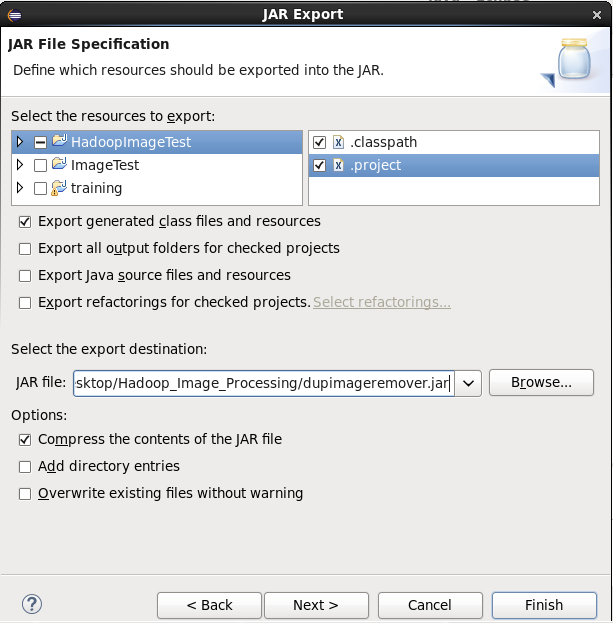


Now we need to convert those files in jar files to process in the command line.

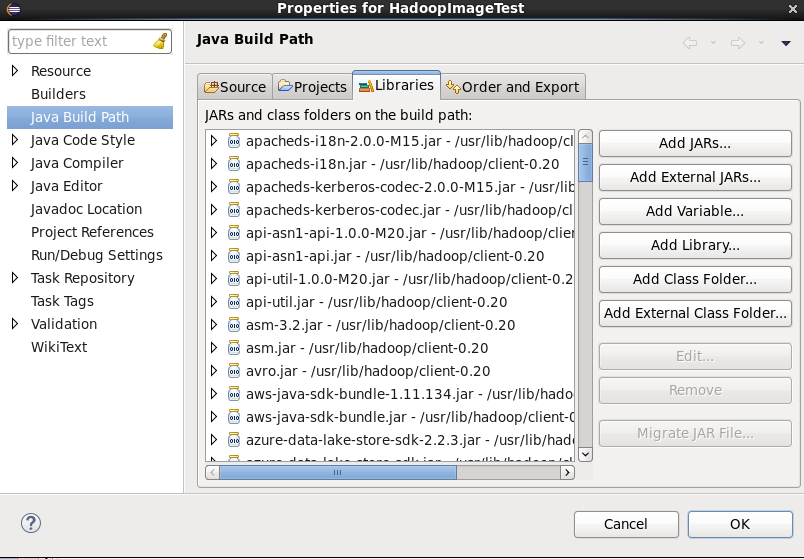
First right clicks on Project name, select export, go to the folder JAVA and select jar file. Then go to next and change the name of jar file according to you. Here my jar file is seqgeneration.jar



So the same thing for another java program and jar file for the second program is named as dupimageremover.jar



To run the Hadoop project on java platform, you need some Hadoop libraries to be imported in the project to get it execute completely. For that you have to right click on project and go to the properties then on the left side select the “java Build path” then go to the libraries and select all to add.



Now run the java program

**$ hadoop fs -cat /imagetest\_file/filename/ imagenames.txt**

Then we will get all the information about images and we keep it on HDFS

$ **Hadoop jar seqgeneration.jar /imagetest\_file/ filename /imagetest\_file/seqfile\_output**

Now lets run the second program to remove the duplicate images.

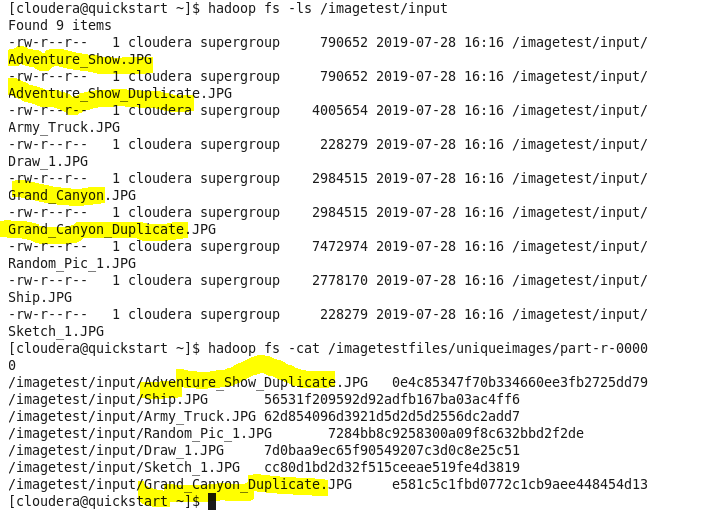
**$ Hadoop jar dupimageremover.jar/ imagetest\_file/seqfile\_output /imagetest\_file/ uniqueimages**

Output:

You must write two command two compare the input and output images

**$ Hadoop fs -ls /imagetest/input**

**$ Hadoop fs -cat /imagetestfiles/uniqueimages/ part-r-0000**



1. **Conclusion**

The amount of picture data has developed impressively as of late because of the development of interpersonal interaction, surveillance cameras, and satellite pictures. In any case, this development isn't restricted to mixed media information. Hadoop is most recent framework used to handle and process Big amount of data which is mostly adopted by social media and financial industries. Using this MapReduce job, we can reduce time of Image Processing and remove duplicate Images in fraction of seconds from huge data sets. There are lots of duplicate images available on Internet and they are none of use. Remove them effectively and carefully can give us huge vacant space and speed in surfing.

1. **References**

[1] Gongrong Zhang, Qingxiang Wu, Zhiqiang Zhuo, Xiaowei Wang, Xiaojin Lin “A Large-scale Images Processing Model Based on Hadoop Platform” ACM ICCC’13, December 1–2, 2013, Wuhan, China.

[2] Vigneshwari K1, Dr K.Kalaiselvi 2 “ADAPTIVE MEDIAN FILTER BASED NOISE REMOVAL ALGORITHM FOR BIG IMAGE DATA” 4th International Conference on Cyber Security (ICCS) 2018 (IJASSR).

[3] Purval Rathore, Deepak Shukla “Analysis and performance improvement of K-means clustering in Big Data environment” International Conference on Communication Networks (ICCN) 2015 IEEE.

[4] Jun Je Foo, Justin Zobel, Ranjan Sinha, S.M.M. Tahaghogi “Detection of Near-duplicate Images for Web Search” 2007 ACM, July 9–11, 2007, Amsterdam, The Netherlands.

[5] Markus Zlabingr, Allan Hanbury “Finding duplicate images in biology papers” 2017 ACM, April 03 - 07, 2017, Marrakech, Morocco.

[6] Feng Wang, Bo Dong, Je Qiu, Xinhui Li, Je Yang, Ying Li “Hadoop High Availability through Metadata Replication” 2009 ACM, November 2, 2009, Hong Kong, China.

[7] Wei Dong, Zhe Wang, Moses Charikr, Kai Li “High-Confidence Near-Duplicate Image Detection” 2012 ACM June 5-8, Hong Kong, China

[8] Mosab Shaheen1\* and Dr Madhukar B. Potdar2† “Image Processing in Hadoop Distributed Environment” ICRISET2017(Kalpa Publications in Computing, vol.2), pp.188{195}

[9] The Apache Software Foundation, "What is Apache Hadoop?", 2016, hadoop.apache.org (accessed on Dec 5, 2016).

[10] Jia Li, Knhu Lin, and Jingjin Wang "Design of the mass multimedia files storage architecture based on Hadoop" IEEE 8th International Conference on Computer Science & Education (ICCSE). April 26-28, 2013. Colombo, Sri Lanka.