Parallel Conversion of Video File Format using Hadoop

S. M. Srihari Shankar
II M.E, Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.

R.Vidhya Prakash
Assistant Professor, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.

Abstract

Digital data like images, videos contribute the major part of today's life. These data (videos) are available at various formats and in huge size. For the portable usage in many digital devices like pda, mobile phones, the video format has to be converted into a format acceptable by those devices (say 3gp in mobile phones). Conversion from one format to the other, consumes more time. So power of distributed processing is used. Hence HADOOP MAP REDUCE technique performs the conversion operation in less time and in an efficient way.

Keywords— Hadoop distributed file system, Hadoop map reduce, data nodes, name nodes, Tracker.

1. Introduction

AVI stands for Audio Video Interleave. It has become de-factor standard for storing video and audio files on PC. AVI integrates different files of audio and video into single standard container to allow simultaneous playback. One of the major advantage of AVI is its simplified architecture which allows AVI to run on multiple operating systems like Windows, Mac, Linux, Unix and also most of the web browsers supports AVI.

3GP is multimedia container format which is used for the delivery and playback of audio/video files over high-speed wireless networks, especially mobile devices. 3GP file formats were especially designed so as to decrease the bandwidth and storage requirements in order to accommodate mobile devices. Structurally, 3GP's are based on the file format called ISO which is defined in ISO/IEC 14496-12-MPEG-4 Part 12. The 3GP file can store video streams in MPEG-4 Part 2 (also called H.263) and MPEG-4 Part 10 (AVC/H264) formats, and audio streams in AAC-LC, HE-AAC v1, HE-AAC v2, AMR-NB, and AMR-WB+.

For the portable usage file format conversion from one to the other becomes necessary (e.g., For video files access in mobile phones format conversion from avi to 3gp format is needed).The issues regarding the file format conversions are the (1) Quality of conversion.(2)Rate at which the conversion takes place. The existing avi to 3gp conversion tools (format factory, total video convertor, any to any convertor) provide the quality in conversion but they consumes lots of time. The file format conversion process in a single machine takes huge time for completion when the file sizes are in GBs. To overcome this problem in the proposed technique, conversion process is done in Hadoop Distributed File System using the Map Reduce technique.

The remainder of this paper is organized as follows. Section 2 discusses the methodology that should be consider while designing Hadoop framework, Section 3 shows the analysis results, Section 4 concludes the paper.

2.Methodology

Hadoop is a software platform specifically designed to process and handle vast amounts of data. The Hadoop framework consists of the Hadoop Distributed File System (HDFS) that is designed to run on commodity hardware and Map Reduce programming paradigm. Map Reduce divides applications into many small blocks of work that can be executed parallel. Replicas of the data blocks are created by HDFS around the cluster of nodes in order to achieve reliability, and also to increase the computational speed. Map Reduce can then process the data where it is located. HDFS has master/slave architecture[1].

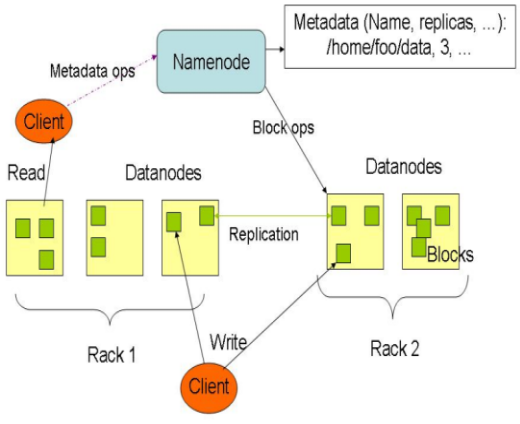

Fig. 2.1 Hadoop Architecture

A HDFS cluster consists of a single NameNode and a number of DataNodes (Fig 1.1).The NameNode is a master server that manages the file system namespace and regulates access to files by clients. The DataNodes manage storage attached to the nodes that they run on. During execution input file is splitted in to number of separate blocks and those blocks are is stored in to individual DataNode. The NameNode performs operations such as opening, closing and also renaming the files and directories by executing the namespace. And also it specifies the DataNodes containing the blocks. When there is a request from the clients DataNodes are responsible for serving the appropriate data to the corresponding clients.

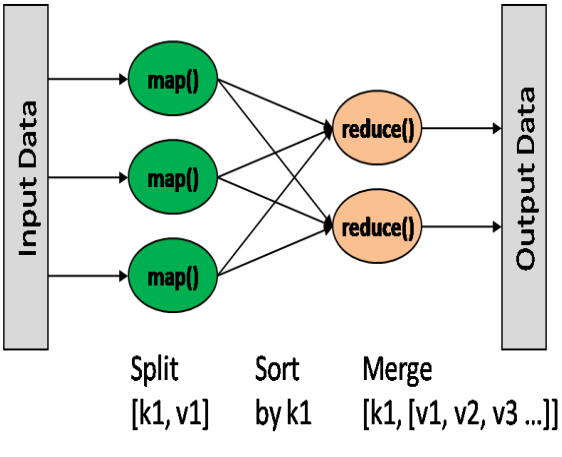

Fig. 2.1 Hadoop Map-Reduce Framework

MapReduce is a programming paradigm (Fig.1.2) that provides parallel operations on data using key/value pairs [4]. A MapReduce computation has a map phase and a reduce phase. The key/value pair is given as the input for the above computation. In the map phase, the framework (record reader) splits the input data set fragments and assigns each fragment to a map task. Each map task consumes key/value (K,V) pairs from its assigned fragment and produces a set of intermediate key/value (K',V') pairs. For each output key-value pair it invokes a user-defined reduce function where the fragments are joined [2].It has a single master server or jobtracker and several slave servers or tasktrackers, one per node in the cluster. Users submit map/reduce jobs to the jobtracker. The jobtracker manages the assignment of map and reduce tasks to the tasktrackers and also it instructs the tasktracker to execute particular task.

A. File splitting

AVI is a derivative of the Resource Interchange File Format (RIFF), which divides a file's data into blocks, or chunks(Fig.1.3). Each chunk is identified by a tag which is called as FourCC. An AVI file takes the input as a RIFF formatted file which is of single chunk and then it splits the chunk into pair of mandatory chunks and one optional chunk. The hdrl tag is used to identify the first sub-chunk, and this chunk holds the file header which provides the information's such as width, height and frame rate about the audio/video file. Another tag called movi tag which represent second sub chunk that holds actual audio/video data. Finally the tag idx1 is used to represent third or the optional sub chunk which holds the offsets of the data chunks within the file.

Basic RIFF File Layout

RIFF: AVI
- LIST: hdrl
  - avih
  - LIST: strl
    - strh
    - strf
    - strd (optional)
    - strn (optional)
- LIST: movi
  - LIST: rec_ chunks 1..n
    - xx data chunks where is the stream number
  - xx data chunks where is the stream number
- idx1

Fig. 2.3 AVI file structure

"Moving the computation is cheaper than moving data" [3], so instead of moving the entire data for conversion, the proposed technique move the process of conversion towards the data. Fig 1.4 Shows the architecture of the proposed system. Once a Hadoop file system has setup the input AVI file that has to be converted is given to the Map phase. As the Map operation is parallelized the input file is first split to several pieces called FileSplits. The splitting is done based on various parameters of the input file like file size, configuration etc. The input avi is splitted into various smaller avi files namely the file splits in the Record reader class. The record reader class after splitting initiates the mapper class. Then for each File split a <key, value> pair is generated. The key can be assigned to identify the order of the filesplits.This can be used to rearrange the filesplit after the conversion process. The value corresponds to the actual avi splits. This key value pair is given as an input to the mapper class. The replication factor may be set to the required value depending upon the domain in which this system is used.

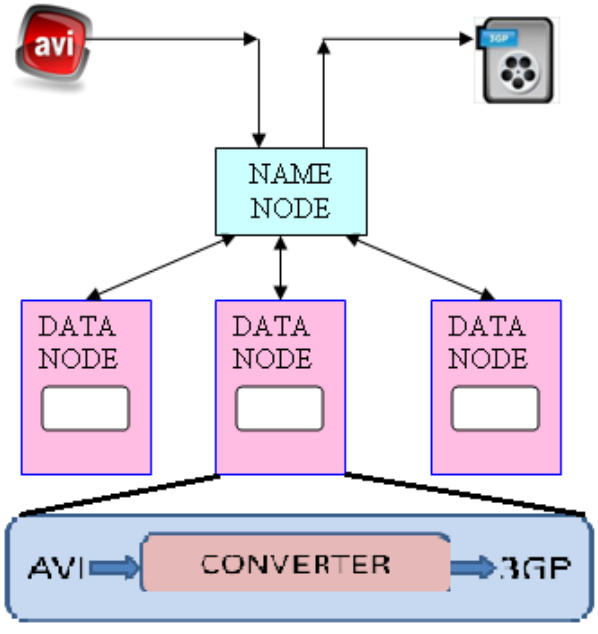

Fig. 2.2 File format conversion using Hadoop Map Reduce technique

B. Conversion

By the end of the splitting process the input file is split into smaller avi files and appropriate key-value pair will be available. Mapper class is used to do any user defined jobs on the key value pair which is given as an input. Once the record reader class initiates the mapper class, the file format is converted to a 3GP format using the video format conversion tool in the Data nodes of the Hadoop cluster. Based upon the requirement of the system and facilities available in the converting tool, the Video format conversion tool can be selected for the conversion process. When the converting tool is invoked it converts the avi data available in each key value pair to 3gp format and stores it in the new key value pair. Now these files form the input to the next phase of the process. The number of Data nodes available in the Hadoop cluster determines the efficiency in reducing the time taken for the video format conversion process.

C. Reduce phase

When a reduce task starts, its input is scattered in many files across all the nodes where map tasks and the conversion tasks ran. These files are then merge sorted so that the key-value pairs for a given key are contiguous. Here based on the key the converted 3GP chunks are ordered and then stored again in the HDFS so that on retrieval Sequential Output is obtained. The 3gp file splits are aggregated with the help of video merging tool like video joiner, total video converter. The details of the converted 3gp files are sent to the name node using the block report of the HDFS which is sent periodically by the data nodes to the name node. The final output is obtained from the name node.

4. Analysis

A set of sample avi files are chosen with varying sizes and used for the conversion process in the stand alone system and in the proposed system with 2 machines in the Hadoop cluster. From the analysis of the avi to 3gp conversion process in both cases we have obtained the following graph which shows the improved format conversion process with less time consumption in the proposed system.

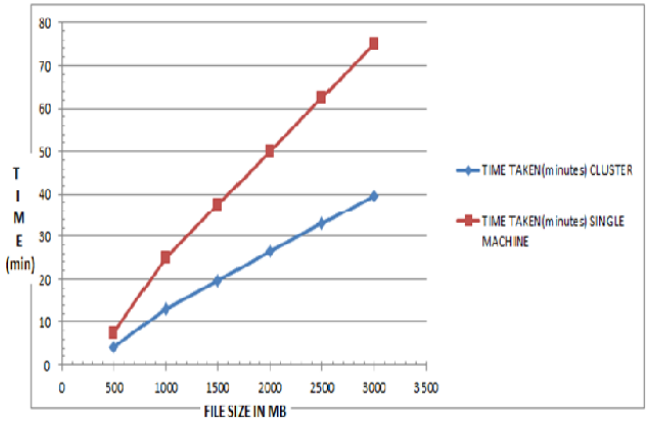

Fig.5. Single Vs Cluster

5. Conclusion

The proposed system uses the hadoop file system-map reduce technique efficiently to convert the given file format (avi to 3gp) in optimized fashion. This system can be used in many domains. Many video lectures are available in the internet and they are in non acceptable formats by digital devices like mobile phones. By using this system the conversion will be done in faster rate and thereby reducing the computation time. Even this system can be extended to convert the videos from any format to any other required format by including the appropriate splitting and merging module for the given input format of the video. This system can be embedded with the web servers so that the available videos for download to clients can be stored in various formats either by converting the video at the time of uploading to server or by the time of request for download in a particular format from the client.

References

[1] IEEE paper on "The Hadoop Distributed File System" By Shvachko, K,; Hairong Kuang; Radia, S.; Chansler, R.;Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on 3-7 May 2010

[2] http://hadoop.apache.org/mapreduce/

[3] IEEE paper on "A Design of Grid Supported Services for Mobile Learning System"by M.Norazizi Sham Mohd Sayuti,Universiti Sains Islam Malaysia (USIM).

[4]J.Dean,S.Ghemawat, MapReduce: simplified data processing on large clusters, Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation, p.10-10, December 06-08, 2004, San Francisco, CA.