Hipi, as alternative for satellite images processing

Wilder Nina Choquehuayta
UNSA / Arequipa
wninac@unsa.edu.pe

René Cruz Muñoz
UNSA / Arequipa
rcruzm@unsa.edu.pe

Juber Serrano Cervantes
UNSA / Arequipa
jserranoc@unsa.edu.pe

Alvaro Mamani Aliaga
UNSA / Arequipa
amamani@unsa.edu.pe

Pablo Yanyachi
UNSA / Arequipa
raulpab@unsa.edu.pe

Yessenia Yari
UNSA / Arequipa
yyari@unsa.edu.pe

Abstract

These days, in different fields of both industry and academia, large amounts of data is generated. The use of several frameworks with different techniques is essential , for processing and extraction of data. In the remote sensing field, large volumes of data are generated (satellite images) over short periods of time. Information systems for processing these kind of images were not designed with scalable features. In this paper, we present an extension of the HIPI framework (Hadoop Image Processing Interface) for processing satellite image formats.

KeyWords: Big Data, Remote Sensing, Hadoop, HIPI, MapReduce, Satellite Images

1 Introduction

The field of remote sensing is helpful in different areas of both industry and academia, because it uses images of the earth's surface that are acquired from different sources like antennas and satellites, which provide increasingly better image resolution as technology advances. Nowadays, several open source frameworks are available to process big data, such as **Hadoop** (Shvachko et al., 2010), **H2O** (0xdata:H2O, 2015), **Spark** (Apache:Spark, 2015), etc., which are used for distributed and parallel processing of large volumes of data. HIPI (Hadoop Image Processing Interface) is an image processing library designed to be used with the Apache Hadoop MapReduce parallel programming framework. HIPI facilitates efficient and high-throughput image processing with MapReduce style parallel programs typically executed on a cluster. In the present work the HIPI library was modified, giving additional functionalities to read and process the GeoTIFF format (format provided by USGS -United States Geological Survey- for *Landsat* satellite images).

2 State of the Art

There are different techniques used for image classification in semantic taxonomy categories such as vegetation, water, etc. (Codella et al., 2011), however these methods don't consider scalability as part of its solutions, *Noel C. F. Codella et. al.*. In *Wanfeng Zhang, et. al.* (Zhang et al., 2013) An infrastructure for massive processing of satellite images in a multi-dataCenter environment, consisting of a DataCenter, where *Access Security*, *Information Service* and strategy *Scheduling* for data management us introduced. It's important to consider HIPI (Sweeney et al., 2011) as a state of the art, extensible library for image processing and computer vision applications, which helps to avoid the problem of small files and achieving improvements in memory and response time.

3 Proposal

That is why this paper is a modification of HIPI, to extends its functionality to work with TIFF images or GeoTIFF type. To achieve this, we proceeded as follows:

- It was decided to use the *Tiff* format from satellite images obtained from the USGS since this format unlike others has no compression or data loss (Adobe, 1992).
- JAI API was chosen to read and write the chosen format, JAI has more *codecs* and features available that can be useful for reading multiple formats.
- Classes needed to upload, encode and decode images of *Tiff* type were modified.

Based on the tests, the possibility of using HIPI for processing multispectral and hyperspectral images

was analyzed. For such images, operations such as PCA are important to process all spectral bands, it is proposed to keep them in the same zip file, or as different images belonging to a *tiff* format, and then decode and interpret as a conventional multi-band image.

4 Experiments

The experiments were performed on a Local Heterogeneous *Cluster* depicted in Table 1 where the characteristics of *slaves* and *master* are shown. We used satellite images from *LandSat 7*, only considering the first 4 bands so we compressed a satellite image in .zip then we used the .zip up to 0.5GB, 1GB, 5GB and 10GB. The algorithm was tested about the average of channels which is explained in the official website of HIPI. In each task map, we iterated over each read of band of satellite image as FloatImage, added each value of pixel depending of channel then divided for number of pixels (width x height) and returned the key of the satellite image and array of data calculated. In each reduce task, we only calculated the average of the average of channels from each satellite image.

Node	characteristics
master	*Core i7, RAM 8GB, Disk 100GB, S.O Ubuntu 64 bits*
slave 1	*Core i7, RAM 8GB, Disk 100GB, S.O Ubuntu 64 bits*
slave 2	*Core 2 Duo, RAM 4GB, Disk 100GB, S.O Ubuntu 64 bits*
slave 3	*Core 2 Duo, RAM 4GB, Disk 100GB, S.O Ubuntu 64 bits*

Table 1: Characteristics of the cluster

In Table 2 shows in axis x the amount of data in GBs and in y axis the execution time. The Hadoop configuration was 1 replication of data, chunks of 32MB, 64MB and 128MB, 4096MB in memory for task reduce and map. The java virtual machine for task reduce and map was configured with 4096MB at the most.

5 Conclusions

Based on the review conducted and theoretical experimental tests HIPI modified version of the article concludes as follows: It is possible to perform various image processing operations, such as filters, variance, clustering or dimensionality reduction by using the MapReduce algorithm and also while the information in compressed format occupies less space, this does not necessarily mean faster times when processing, since the matrix calculations are done on the same decompression.

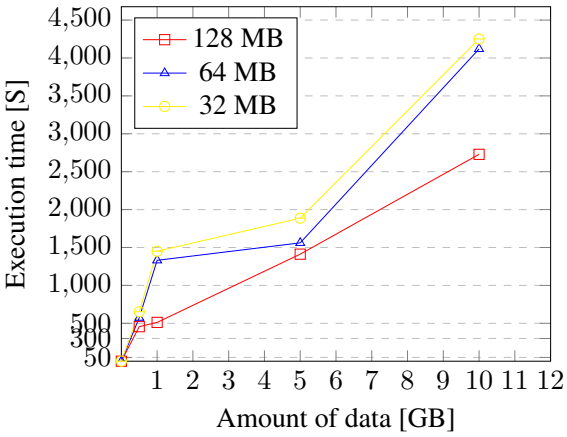

Table 2: Execution time vs Amount of data spatial

References

0xdata:H2O. 2015. H2o@ONLINE. Website. 0xdata:H2O, In: `http://0xdata.com/product/`,accessed(may 2015).

Apache:Spark. 2015. H2o@ONLINE. Website. Spark, Lightning-fast cluster computing , In : `https://spark.apache.org`, `accessed(accessed(2015-05-20))`.

Noel C.F. Codella, Gang Hua, Apostol Natsev, and John R Smith. 2011. Towards large scale land-cover recognition of satellite images. In *Information, Communications and Signal Processing (ICICS) 2011, 8th International Conference on*, pages 1–5. IEEE.

Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. 2010. The hadoop distributed file system. In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*, pages 1–10. IEEE.

Chris Sweeney, Liu Liu, Sean Arietta, and Jason Lawrence. 2011. Hipi: a hadoop image processing interface for image based mapreduce tasks. *Chris,University of Virginia*.

Wanfeng Zhang, Lizhe Wang, Dingsheng Liu, Weijing Song, Yan Ma, Peng Liu, and Dan Chen. 2013. Towards building a multi-datacenter infrastructure for massive remote sensing image processing. *Concurrency and Computation: Practice and Experience*, 25(12):1798–1812.