A New Approach to the Massive Processing of Satellite Images

Wilder Nina†, René Cruz†, Juber Serrano†, Jaime Cuba†, Yoni Huaynacho†,
Alvaro Mamani-Aliaga†‡, Yessenia Yari†‡ and Pablo Yanyachi†
†National University of San Agustín
Arequipa - Peru
‡Catholic University of San Pablo
Arequipa - Peru
Email: see http://www.tunkiproject.org

***Abstract*—Technological advances in the field of Remote Sensing generate large volumes of geospatial data. Current geographic information systems (GIS) doesn't support the massive processing of satellite imagery, two examples of this kind of software are: (i) the brazilian Spring project, GIS and remote sensing image processing system (ii) QGIS, a free and open source GIS. To achieve massive processing, we have HIPI framework, it provides a solution for how to store a large collection of images and works on the Hadoop Distributed File System. Currently, HIPI only supports specific image formats, such as, JPEG, PNG and PPM.**

In this article is presented a new approach to distributed processing of considerable amounts of satellite images. We make an extension of HIPI to support satellite images format, TIFF, this fact helps to preserve the information, process and analyze satellite images massively to have results faster than the traditional way.

***Keywords*—Big Data, Remote Sensing, Hadoop, HIPI, Satellite Images, Procesamiento Masivo de Imágenes Satelitales (PMIS).**

I. Introducción

Nos encontramos en una etapa donde el flujo de la información crece rápidamente en los diferentes campos de la informática como sociedad y seguridad, educación, juegos, medicina, etc. Actualmente el manejo llegando a convertirse a grandes volúmenes de datos llamado *BigData*.

La teledetección es una técnica que ayuda a diversas áreas de la industria, capturando imágenes de la superficie terrestre, en donde la calidad de estas se incrementan exponencialmente a medida que pasan los años, debido al crecimiento y desarrollo tecnológico de los sensores, antenas, velocidad de transmisión de datos, entre otros factores [1], brindanonos mayor cantidad de imágenes y con mejor resolución.

La Computación de Alto Rendimiento (HPC) es una tecnología que reúne diferentes áreas de la computación como arquitectura de computadoras, algoritmos y programas que ayudan a resolver problemas complejos y avanzados de una forma rápida y efectiva. HPC tiene como objetivo satisfacer las crecientes demandas de la velocidad de procesamiento y análisis de enormes conjuntos de datos [2].

Hoy en día existen diversos *frameworks* o técnicas como *Hadoop*, *H2O*, *Spark*, etc. que se utilizan para el procesamiento distribuido y paralelo de grandes volúmenes de datos. *Hadoop* es uno de los *frameworks Open Source* más utilizados para el manejo de *Big Data*, usa un sistema de archivos HDFS[3] y el algoritmo *MapReduce* [4] para procesamiento a gran escala. Para poder trabajar con imágenes es necesario usar HIPI sobre un *Cluster* configurado con *Hadoop*. La desventaja que tiene, es que para analizar imágenes satelitales es necesario realizar algunas modificaciones en HIPI de tal forma que soporte imágenes de formato GeoTiff, lo cual es analizado en este trabajo.

El presente artículo esta organizado de la siguiente manera. En la Sección II se muestra el estado del arte sobre las diversas técnicas y/o *frameworks* referentes al procesamiento de imágenes satelitales. La Sección III, detalla los conceptos referidos al campo de la teledetección, *Big Data* y HIPI. En la Sección IV se explica la propuesta planteada en este trabajo. La sección V muestra los experimentos realizados. Finalmente, en la Sección VI se presenta las conclusiones y los trabajos futuros que se pueden realizar a partir de esta investigación.

II. Estado del arte

El proceso de teledetección tradicional se ve limitada al intentar realizar el procesamiento de una gran cantidad de datos en tiempo real, para este caso Jianbo et al. [5] propone la segmentación y un procesamiento piramidal de estas imágenes mediante un algoritmo el cual que va creando una imagen nueva a partir de la imagen original, gracias al promedio de los píxeles de su propio entorno, de esta forma, estas imágenes, puedan ser procesadas con mayor facilidad.

Apostol et al [6], nos habla sobre una técnica usadas para la clasificación de imágenes satelitales en categorías de taxonomía semántica tales como vegetación, agua, pavimento entre otros, usando implementación distribuida *Hadoop* junto al clasificador *Support Vector Machine* (SVM).

Stathis et al. [7] propone la transmisión de la Geoinformación en diferentes etapas, desde la solicitud del cliente hasta el procesamiento de la data, todo esto realizado en la nube, desarrollando una arquitectura abierta y de interoperabilidad independiente de sus componentes. Teniendo la capacidad de soportar los diferentes formatos de proveedores de Geodata.

Zhang et al. [8] presenta una infraestructura de múltiples centros de datos (MDC) para gestionar y procesar una gran cantidad de imágenes (teledetección). El sistema propuesto se basa en los dos grupos de distribución DCs / clusters, que están

978-1-4673-9143-6/15/$31.00

equipadas con DC o gerente de recursos de clúster. Seguridad de acceso y servicio de información se introducen para apoyar esta arquitectura del MDC.

III. Marco Teórico

En este sección se presentaran los conceptos fundamentales relacionados a la presente investigación.

A. Imágenes Satelitales

En el campo de teledetección la energía que emana la superficie terrestre es medida usando sensores digitales, los cuales se encuentran generalmente ubicados en naves, satélites y plataformas espaciales, dichas mediciones se usan para construir imágenes espectrales. La energía emanada de los cuerpos (ya sea por absorción, dispersión o emisión) por el efecto de la reflexión y emisión de esta radiación es conocida también como reflectancia [9]. Cada valor de reflectancia espectral se registra como un número digital, el conjunto de estos números se transmiten de nuevo a la Tierra donde mediante un procesamiento se convierten en colores o matices de gris para crear una imagen satelital [10].

La imagen satelital resultante consiste en un conjunto de matrices, una por cada canal del sensor, en la que aparecen números en un intervalo de 0 al 255, en donde el valor 0 índica que no ha llegado radiación desde ese punto y el 255 que llega el valor más alto de radiación. En la Figura 1 es mostrado el primer canal de una imagen satelital **LandSat 7** [1] tomada de la provincia de Arequipa, Perú.

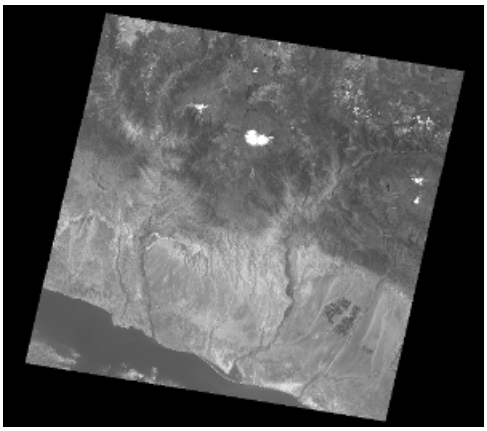

Figure 1: Primer canal de la Imagen Satelital *LandSat 7* de la provincia de Arequipa-Perú

1) Tipos de Imágenes: Existen diferentes tipos de imágenes satelitales de acuerdo con el número de bandas y el valor que toma cada pixel [11], las cuales son:

- **Imágenes pancromáticas**: Son las imágenes captadas mediante un sensor el cual mide la reflectancia de energía en una amplia parte del espectro electromagnético. Estas imágenes suelen abarcar la parte visible e infrarroja cercana del espectro.Son representadas en imágenes blanco y negro.
- **Imágenes Multiespectrales**: Son las imágenes captadas por el sensor en muchas bandas, gracias al conjunto de detectores que este posee, los distintos valores de reflectancia se combinan para crear imágenes de color, este número de bandas se miden desde tres a catorce bandas, dependiendo del sensor.
- **Imágenes hiperespectrales**: Son imágenes captadas con un sensor hiperespectral que mide la reflectancia en cientos de bandas, lo que permite detectar características muy sutiles entre los rasgos de la superficie, especialmente referidos a la vegetación.

2) Bandas y Sensores: Un sensor es el aparato que reúne la tecnología necesaria para captar imágenes a distancia y que es transportado por un satélite. Puede captar información para diferentes regiones del espectro y cada una de estas regiones se denomina canal o banda. Las mas utilizadas son: ***Teledetección del infrarrojo reflectivo y el visible*** generalmente usado para el reconocimiento de la vegetación, monitoreo de agua e identificación de suelo y roca, ***Teledetección en el infrarrojo lejano*** usado para detectar la emisión térmica emitida por los cuerpos, usada en detección de incendios forestales y ***Teledetección en la zona de las microondas***, usado de dos formas, mediante la percepción remota pasiva y activa, ambas para la detección de microondas.

El sensor de satélite cuenta con miles de detectores diminutos que miden la cantidad de radiación electromagnética que refleja la superficie terrestre, entre estos: objetos y/o seres vivos. De esta forma los sensores son clasificados de la siguiente forma: (i) **sensores activos**, los cuales generan su propia radiación y la reciben rebotada; y (ii) **sensores pasivos** los cuales reciben radiación emitida o reflejada por la Tierra.

Dentro de los sensores pasivos están los sensores del tipo fotográfico, óptico-electrónico que combina una óptica similar a la fotográfica y un sistema de detección electrónica (detectores de barrido y empuje), espectrómetro de imagen, y de antena (radiómetro de microondas). Por lo que se refiere a los sensores activos, actualmente se dispone del Radar y el Lidar [2] [12].

Los diferentes canales se pueden caracterizar en función de variables [12]:

- Amplitud espectral, región del espectro para la cual capta datos.
- Resolución radiométrica, número de intervalos de intensidad que puede captar.
- Resolución espacial, tamaño de píxel.
- Resolución temporal, tiempo que tarda el satélite en pasar dos veces por el mismo sitio.

3) Formatos de las Imágenes Satelitales: La información obtenida puede representar diferentes tipos de imágenes, los formatos en que son guardados los archivos se refieren a una estructura lógica utilizada para almacenar la información, se aprecian dos grandes grupos de formatos, raster y vectorial.

[1] LANDSAT (LAND=tierra y SAT=satélite), primer satélite enviado por los Estados Unidos para el monitoreo de los recursos terrestres.

[2] Laser Imaging Detection and Ranging, basado en tecnología láser

a) Formatos vectoriales: Los formatos vectoriales son comunes en las imágenes satelitales, son usados de diversas formas, siendo posible almacenar coordenadas, atributos por ejemplo.

b) Formatos raster: Son usados para almacenar información de la imagen ya sean imágenes escaneadas, fotografías aéreas, o información capturada por satélites este último denominado teledetección, en contraste con las imágenes donde la dimensión se da en puntos por pulgada o cantidad de celdas, en las imágenes teledetectadas cada celda indica el área que cubre en metros, a continuación algunos formatos de este tipo:

- *Arc Digitized Raster Graphics (ADRG)*: Es un formato usado por la armada milicia para almacenar imágenes a partir de mapas físicos.
- *Band Interleaved by Line (BIL), Band Interleaved by Pixel (BIP), y Band Sequential (BSQ)*: Son formatos producidos por los sistemas de teledetección. La principal diferencia entre ellos es la técnica utilizada para almacenar valores de brillo capturadas simultáneamente en cada una de sus bandas espectrales.
- *Digital Elevation Model (DEM)*: Modelo digital de elevaciones, es usado por la USGS para almacenar información de la elevación en la superficie terrestre en vez de intensidades como otros tipos de imágenes.
- *PC Paintbrush Exchange (PCX)*: es un formato común producido por diferentes programas de computadora o escáneres de información.
- *Spatial Data Transfer Standard (SDTS)*: es un formato estándar para transferencia de información en formato raster.
- *Tagged Image File Format (TIFF)*: Al igual que PCX, es un formato común producido por diferentes programas de computadora o escáneres de información.

B. Big Data

Segun la EMC en su articulo [14] define a *Big Data* como una nueva generación de tecnologías y arquitecturas, diseñados para extraer, analizar económicamente valores de volúmenes muy grandes de una amplia variedad de datos a alta velocidad. También la podemos definir como ”grandes cantidades de datos” lo que se traduce a cantidades masivas (petabytes, exabytes, zettabytes, etc) [15].

El Instituto Global McKinsey estimó que el volumen de datos está creciendo 40% por año, y crecerá 44% entre los años del 2009 al 2020 [16], [17].

1) Tipos de datos: La Big Data se clasifica dentro de diferentes categorías los cuales son fuentes de datos (*Web & Social*, *Machine*, *Sensing*, *Transaction*, *IoT*), formato de Contenido (Estructurado, SemiEstructurado y No Estructurado) , almacenamiento de datos (Orientado al documento, Orientado a la columna, Basado en grafos y Llave-valor) , etapas de la data (Limpieza, Normalización y Transformación) y procesamiento de datos (Batch y Tiempo Real) [18].

2) Algoritmo Mapreduce: El algoritmo *MapReduce* tiene su origen en las operaciones primitivas *map* o *reduce* presentes en *Lisp* y en otros lenguajes funcionales [4]. *MapReduce* de *Hadoop* se basa en el modelo de programación de *Google* en el 2004. En el *MapReduce* de *Hadoop* hay un solo *master* que gestiona un número de *slaves* o *workers*, donde el archivo de entrada reside en un sistema de archivos distribuido en todo el *Cluster*, este se divide en trozos llamados *Chunks* de tamaño uniforme replicados a través del *Cluster* para tolerancia a fallas y acceso rápido de lectura. Como se muestra en la Figura 2 cada *Chunk* de entrada se procesa primero por una tarea *Map*, que da salida a una lista de pares clave-valor generado por una función de mapa definido por el usuario. Las salidas de tareas *maps* se dividen en *buckets* basado en la clave. Cuando todos los mapas han terminado, se comienza las tareas *reduce* que aplica una función de reducir a la lista de salidas del mapa con cada clave.

C. Arquitectura de HIPI

El *framework* HIPI es una librería libre y extensible para el procesamiento de imágenes y aplicaciones de visión computacional, basado sobre *MapReduce* [19]. HIPI se presenta a los usuarios como una interfaz intuitiva donde las aplicaciones que se hagan serán altamente escalables, trabándose de manera distribuida y paralela.

HIPI trabaja sobre un tipo de dato HIB que consiste de 2 archivos: un archivo de datos conteniendo las imágenes concatenadas y otro archivo de índice conteniendo la *metada*. HIB (*HipiImageBundle*) presenta muchos beneficios con respecto al formato *Hadoop's Sequence file* el cual se desempeña mejor que las aplicaciones estándar pero el tipo de lectura que utiliza (serial) y el tiempo que se toma en generar una es una gran desventaja. HIPI también presenta beneficios sobre *Hadoop Archive* (HAR) el cual mayormente es usado como *Backup* donde el tiempo de lectura es muy lenta [19].

En la Figura 3 se muestra la arquitectura de HIPI el cual tiene como entrada un HIB, el cual pasa sobre una fase *Cull* el cual permite que las imágenes se filtren basado a sus propiedades, como por ejemplo el tamaño o la resolución. Esto se realiza en base a su *Header*. Después la salida de cada imagen ya filtrada es convertida en un tipo *Float Image* el cual es la entrada para los diferentes *Task Maps* que se procesaran en cada *Slave* o *Worker* en el *Cluster*. Siguiendo con el modelo de programación *MapReduce* la salida de cada *Task Map* sera un par intermedio de índice y valor, el cual sera guardado en disco, después entra la fase de *Shuffle* en donde se emparejará de acuerdo al índice los diferentes valores. Finalmente se ejecutaran diferentes *Task Reduces* sobre el *Cluster* donde cada entrada es un índice con un iterador sobre los valores emparejados y salida es una operación sobre cada índice.

En la Figura 4 es mostrada una comparación para el ejemplo de *The Principal Components of Natural Images*, o mas conocido como PCA, donde competirán *small files* y los 2 formatos HIB y *Hadoop Sequence File*, donde en el eje x representan la cantidad de imágenes y el eje y el tiempo que se

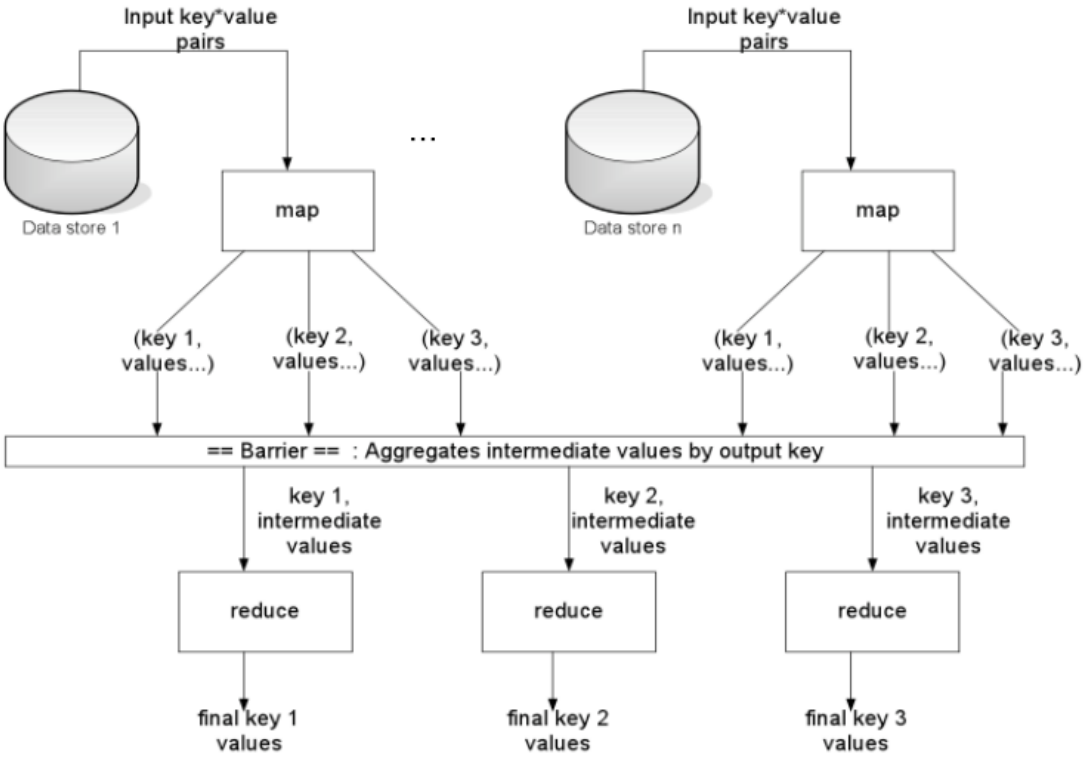

Figure 2: Modelo de Trabajo de *MapReduce* [13]

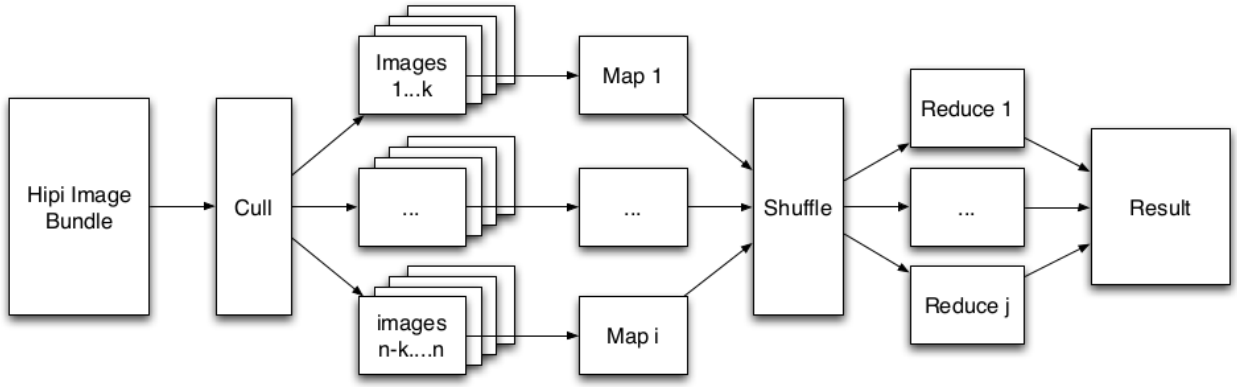

Figure 3: Arquitectura de HIPI basado en Hadoop [19]

demora en procesar. Como se observa para una entrada de 100 imágenes el *small files* muere debido a la creación de varios *chunks*, donde cada *chunk* se le asigna un espacio en memoria que provoca un *overload* en *Hadoop*, a diferencia de *HIB* que tiene una ventaja con respecto a *Hadoop Sequence File* para imágenes mayores $100,000$.

D. Java Advanced Imaging (JAI)

El API de JAI provee un conjunto de interfaces orientadas a objetos, soporta un simple modelo de alto nivel de programación, que permite la manipulación de imágenes de una manera sencilla en Aplicaciones de Java, JAI va más allá de la funcionalidad de las APIs de imágenes tradicionales para proporcionar un alto rendimiento y un *framework* de procesamiento de imágenes extensible e independiente de la plataforma [20].

IV. Propuesta

La generanción de un gran volumen de información (imágenes satelitales) son el resultado de la teledetección, la cual invita a realizar el procesamiento de dichas imágenes tomando en cuenta conceptos de alto rendimiento y escalabilidad. Como se revisó en la Sección III HIPI es una herramienta que nos ayuda con el procesamiento masivo sin embargo esta sólo soporta tres formatos básicos: JPEG, PNG y PPM.

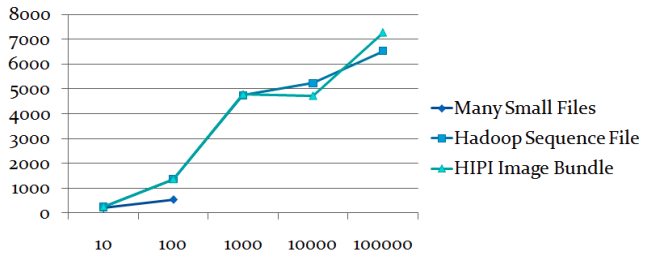

Figure 4: Ejemplo de Covarianza con HIPI, Hadoop Sequence y Small Files [19]

Una imagen satelital esta dada en diferentes formatos como ADRG, BIP, BSQ, DEM, PCX, SDTS y TIFF, siendo este último el que se presenta en las galerías de imágenes de USGS. El presente trabajo propone una modificación de HIPI, extendiendo su funcionalidad para poder trabajar con imágenes del tipo tiff o geotiff, se usan estos formatos ya que a diferencia de los demás no presenta compresión ni pérdida de datos, lo cual hace que el análisis sobre estas imágenes sean mas exactas.

Se propone hacer esta modificación de la siguiente manera:

- Se eligió el API de JAI para la lectura y escritura del formato elegido, además este cuenta con mas *codecs* y funciones disponibles que pueden ser útiles para la lectura de múltiples formatos especificados en la sección VI.
- Se modificó los archivos de compilación de HIPI para agregar las librerías actuales de JAI al *classpath*, obtenidas de la página : http://www.oracle.com/technetwork/java/current-142188.html.
- Se modificó las clases necesarias para subir, codificar y decodificar imágenes del tipo *Tiff*.

V. Experimentos realizados

El ambiente (*cluster*), compuesto por 10 recursos computacionales, utilizado para los experimentos tienen las siguientes características:

- Procesador Corei7;
- memoria RAM 4GB;
- almacenamiento en disco 30GB SATA;
- S.O. Ubuntu 64bits

Para dicha actividad se ejecutó la técnica de análisis de componentes principales, PCA sobre un conjunto de 20 imágenes satélitales *LandSat* cada una de ellas conformada por 7 bandas. Dicha técnica fue utilizada para reducir la dimensionalidad del conjunto de datos.

VI. Conclusiones

En este artículo se propone un nuevo enfoque para el procesamiento de imágenes satelitales, usando HIPI como una alternativa para manipularlas. El hecho de añadir el nuevo formato al *framework* HIPI, formato (*tiff*), ayuda a que se pueda preservar la información, procesar y analizar imágenes satelitales en forma masiva para tener resultados más rápidos en comparación con la teledetección tradicional.

En base a las pruebas realizadas, como trabajo futuro se tiene la posibilidad de usar el *framework* HIPI para el procesamiento de imágenes multiespectrales e hiperespectrales. Si bien inicialmente HIPI fue diseñado para el procesamiento de imágenes que representan intensidades de color, ya sea en una, dos, tres o cuatro bandas, puede soportar un numero indefinido de estas. Para procesar todas las bandas espectrales, se propone mantenerlas en un mismo archivo comprimido ZIP, o como diferentes imágenes pertenecientes a un *tiff*, para luego ser decodificadas e interpretadas como una imagen convencional de múltiples bandas.

Agradecimiento

Los autores agradecen al “Fondo Nacional de Desarrollo Científico y Tecnológico”, Fondecyt - Perú, por el financiamiento del proyecto.

References

[1] Huawu Deng, Shicun Huang, Qi Wang, Zhiqiang Pan, and Yubin Xin. Building high-performance system for processing a daily large volume of chinese satellites imagery, 2014.

[2] Anshu Pallav Sivakumar V Mamta Bhojne, Abhishek Chakravarti. High performance computing for satellite image processing and analyzing – a review. *International Journal of Computer Applications Technology and Research Volume 2– Issue 4, 424 - 430, 2013*, 2013.

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

[4] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters, osdi'04: Sixth symposium on operating system design and implementation, san francisco, ca, december, 2004. *S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, ad A. Tomkins, Self-similarity in the Web, Proc VLDB*, 2001.

[5] Liu Liang Li, Fu Chen. Technology research of the ground real-time data processing for long range image and information broadcast of remote sensing satellite. In *Geoscience and Remote Sensing (IITA-GRS), 2010 Second IITA International Conference on (Volume:1)*, pages 1–5. IEEE, 2010.

[6] Apostol Natsev John R. Smith Noel C. F. Codella, Gang Hua. 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, 2011.

[7] Stathis Makridis Constantine Papatheodorou Konstantinos Evangelidis, Konstantinos Ntouros. Geospatial services in the cloud. *Computers & Geosciences 63(2014)116–122*, 2014.

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

[9] John A Richards and JA Richards. *Remote sensing digital image analysis an Introduction*, volume 5. Springer, 2013.

[10] S Oprisescu and M Dumitrescu. On the regularization of segmented satellite images. In *Signals, Circuits and Systems, 2005. ISSCS 2005. International Symposium on*, volume 1, pages 83–86. IEEE, 2005.

[11] Sensores Remotos et al. Guía básica sobre imágenes satelitales y sus productos. *Nuevas tecnologías en la gestión de espacios naturales*, 2013.

[12] Ernesto Gómez Vargas, Nelson Obregón Neira, and Diego Fernando Rocha Arango. Métodos de segmentación de nubes en imágenes satelitales. *Tecnura*, 17(36):96–110, 2013.

[13] Matei Zaharia, Andy Konwinski, Anthony D. Joseph, Randy Katz, and Ion Stoica. Improving mapreduce performance in heterogeneous environments. In *Proceedings of the 8th USENIX Conference on Operating Systems Design and Implementation*, OSDI'08, pages 29–42, Berkeley, CA, USA, 2008. USENIX Association.

[14] John Gantz and David Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the Future*, 2007:1–16, 2012.

[15] Jainendra Singh. Big data analytic and mining with machine learning algorithm. *International Journal of Information and Computation Technology4*, pages 33–40, 2014.

[16] Yousuf Syed, Vishal Sonawane, Raghav Gupta, Harsha Mare, and Mahesh Pawaskar. A survey of sentiment analysis on big data.

[17] Andrew McAfee and Erik Brynjolfsson. Big data: the management revolution. *Harvard business review*, (90):60–6, 2012.

[18] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of "big data" on cloud computing: Review and open research issues. *Inf. Syst.*, 47:98–115, 2015.

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

[20] Java advanced imaging disponible en. http://web.archive.org/web/20080207010024. Accedido: 2015-05-20.