HmapS: HyperSpectral Remote Sensing Image Cloud Storage for Large Scale Semantic Search and Distributed Computation

George Vafiadis[1], Vrettos Moulos[1], Dimitris Sykas[1,2]

[1] MapTronic Limited
2 Woodberry Grove, London N12 0DR, UK, {gvafiadis, vrettos, dsykas}@maptronic.co.uk

[2] Laboratory of Remote Sensing, National Technical University of Athens
H. Polytechniou Str. 9, 15780 Zografos Campus, Greece
dimsyk@gmail.com

Abstract: The technological advances in satellite remote sensing technology increased exponentially the volume of remote sensing image data, while the demand for scalable and efficient processing brought many new research challenges. In the current work we propose a computation framework and a storage model for large scale processing of remote sensing image data utilizing the resources of the Cloud. The HmapS is a distributed system based on Hadoop Open Source Software extending the Map-Reduce methodology and implementing a number of core remote sensing algorithms. The Map-Reduce methodology is a programming paradigm that expresses a computation as a sequence of distributed operations on data sets that have the form of key/value pairs. The HmapS framework uses a cluster of machines and executes user defined Map-Reduce jobs across the nodes in the cluster. The storage system is implemented on top of Hadoop's Distributed File System (HDFS). HDFS is designed to reliably store very large files across machines in large clusters. Each file is stored as a sequence of equally sized blocks; these blocks are replicated for fault tolerance and increased reliability.

The storage model implements a multilayer architecture supporting a variety of algorithms for processing and searching. The system is used for storing raw remote sensing image data along with semantically reach Metadata. Collected remote sensing images are uploaded to the cluster of nodes. This process is implemented using a web service in order to facilitate the access. Hyperspectral images are typically stored in files retrieved from a specialized hardware device. These files are spitted to blocks of 64mb and are replicated to different Data Nodes. The replication function is implemented automatically and it is handled by the system in a transparent way. A Data Node hardware failure or availability issues are not affecting the cluster as a whole because the same data are replicated to other nodes.

In the current work, we propose a framework for processing; searching and analysing Remote Sensing Image Data using distributed Map-Reduce Jobs. We provide a method for re-organizing and filtering the data in an efficient and scalable way. The user selects a preferred method of having the data and the extraction algorithm that needs to be applied. The proposed system analyses the request and builds an execution plan, this plan is evaluated and it is optimized before execution. The execution plan has the form of multiple parallel Map-Reduce phases each one implementing a step for the next process.

The proposed system has been evaluated with the implementation of an Unmixing algorithm using Map-Reduce Jobs and performing sematic searches on metadata. The proposed implementation has been tested extensively using different remote sensing datasets.

Keywords: hyper spectral, semantic search, cloud computing, distributed storage, map-reduce