Future Generation Computer Systems 81 (2018) 30–40

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A versatile data-intensive computing platform for information retrieval from big geospatial data

P. Soille *, A. Burger, D. De Marchi, P. Kempeneers, D. Rodriguez, V. Syrris, V. Vasilev

European Commission, Joint Research Centre (JRC), Directorate I. Competences. Unit I.3 Text and Data Mining, Via E. Fermi 2749, I-21027 Ispra (Va), Italy

HIGHLIGHTS

- The recent sharp increase of free, full, and open satellite imagery is making Earth Observation truly entering the big data era.
- Novel platforms are needed for timely retrieving information from Earth Observation imagery at scale.
- A versatile platform coping with batch processing of existing scientific workflows as well as interactive visualization and analysis is put forward.
- The versatility of the proposed platform is demonstrated on a variety of actual use cases originating from various application domains.

ARTICLE INFO

Article history:
Received 13 February 2017
Received in revised form 30 October 2017
Accepted 5 November 2017
Available online 22 November 2017

ABSTRACT

The increasing amount of free and open geospatial data of interest to major societal questions calls for the development of innovative data-intensive computing platforms for the efficient and effective extraction of information from these data. This paper proposes a versatile petabyte-scale platform based on commodity hardware and equipped with open-source software for the operating system, the distributed file system, and the task scheduler for batch processing as well as the containerization of user specific applications. Interactive visualization and processing based on deferred processing are also proposed. The versatility of the proposed platform is illustrated with a series of applications together with their performance metrics.

1. Introduction

Geospatial data are playing an increasing role to answer major societal questions such as those related to the environment, climate change, crisis management, and sustainable development goals. At the same time, geospatial data are becoming ubiquitous given the multiplication of digital data sources ranging from individual citizens to private and public organizations. A major source of geospatial data for systematic studies from regional to global scale is provided by Earth observation satellites delivering an ever increasing flow of raster image data since the launch of Landsat-1 in 1972 by the National Aeronautics and Space Administration (NASA) of the United States [1].

In 2008, the United States Geological Survey (USGS) has democratized the use of Landsat data by making all new and archived Landsat imagery accessible over the Internet under a free and open data policy [2]. The open-access to the global Landsat archive [3] has enabled the production of the first ever high-resolution global maps for the dynamics of forest [4], urban [5], and water [6] environments.

A further sharp increase in the availability of free and open geospatial data has recently emerged with the Copernicus Earth observation and monitoring program of the European Union that delivers satellite imagery complemented by in situ observations. The Copernicus Sentinel satellites developed by the European Space Agency (ESA) consists of a series of six missions with the first three missions [7] devoted to land and ocean monitoring: Sentinel-1 [8], Sentinel-2 [9], and Sentinel-3 [10]. These three missions are each composed of a pair of identical satellites (units A and B) to minimize the time difference between two successive observations of the same point on Earth by the same sensor. Fig. 1 shows the estimated yearly volume of data available for download by ESA for the first three Sentinel missions together with those of the current Landsat and MODIS [11] missions from NASA/USGS.

With expected data *volumes* exceeding 10 TB per day with all Sentinel missions at full operational capacity, data *velocity* highlighted by the sensing of the whole globe every six, five, and two days for Sentinel-1, -2, and -3 respectively, and data *variety* resulting from optical and radar sensors at various spatial, spectral, and temporal resolutions, the Sentinel missions contribute significantly to the big data challenges that geospatial applications are nowadays facing.

The extraction of relevant information from big geospatial data streams calls for innovative platforms tackling the data storage,

* Corresponding author.
E-mail address: pierre.soille@ec.europa.eu (P. Soille).

https://doi.org/10.1016/j.future.2017.11.007

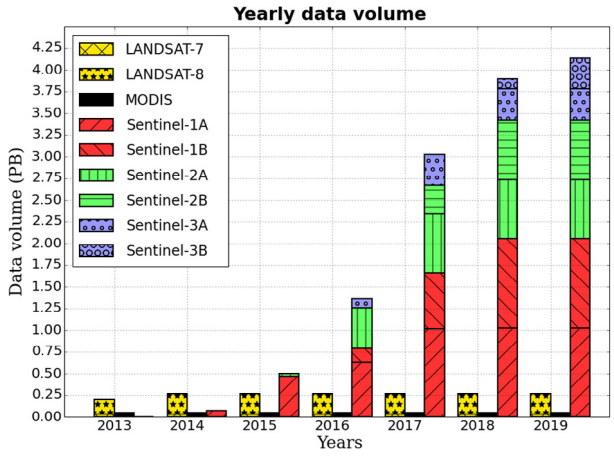

Fig. 1. Estimates of the yearly volume of open and free data for Landsat-7 and Landsat-8, MODIS (Terra and Aqua units), and the three first Sentinel missions. Each reported Sentinel mission is based on a constellation of two identical satellites (units A and B) with full operational capacity reached in 2019. The Sentinel data volume calculations are based on the size of the data packaged for download from the Copernicus delivery services.

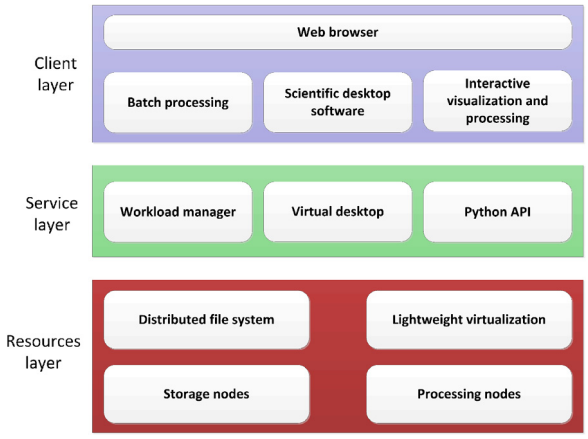

Fig. 2. JEODPP architecture: simplified view with its main layers and components.

processing, analysis, and visualization dimensions. In addition, these platforms need to serve users with very heterogeneous levels of computer literacy given the breadth and depth of the societal questions at stake. In this paper, we propose a versatile data-intensive computing platform meeting these requirements. The platform is versatile in the sense that it accommodates different service levels ranging from large scale batch processing to interactive visualization and analysis. It is in development since 2016 at the Joint Research Centre (JRC) of the European Commission and is called the JRC Earth Observation Data and Processing Platform (JEODPP) [12]. The JEODPP already supports a variety of projects serving policy areas in agriculture, forestry, environment, disaster risk management, development, health, and energy.

The paper is structured as follows. The proposed platform and its components are detailed in Section 2. A comparison with related work is given in Section 3. Applications and performance metrics are presented in Section 4. Conclusions and future directions are given in Section 5.

2. Proposed platform

The velocity and volume of the free and open geospatial data streams that need to be handled call for a solution scalable to the multi-petabyte range. This is achieved on the JEODPP by considering distributed storage coupled with a cluster of computing nodes. Because the targeted applications require high-throughput rather than high-performance computing, commodity hardware for both the storage and processing nodes were considered. The JEODPP accommodates three main services to satisfy the needs of a variety of users: batch processing, provision of legacy environments, and interactive visualization and processing. All services are accessed through a web browser so that no dedicated client software needs to be installed on the devices accessing the platform.

A simplified representation of the JEODPP architecture is shown in Fig. 2 in the form of a three layer stack with the resources layer at its basis, followed by the service layer, and the client layer at its top. The main components of the JEODPP architecture, as well as their interactions, are described hereafter starting from the hardware layer and proceeding with components of increasing abstraction level.

2.1. Hardware layer and distributed file system

A key component for the processing of massive geospatial datasets is a scalable and high-throughput storage sub-system. In the infrastructure set-up used before the implementation of the JEODPP, the processing was using a NetApp appliance as storage backend with file access via NFSv4. The achievable I/O throughput of this solution was not sufficient to cope with the multi-node processing in a cluster environment. In addition, this approach is not financially sustainable for a multi-petabyte storage solution. This is solved on the JEODPP by considering a scalable distributed storage based on commodity servers equipped with one or two attached storage expansion units made of multiple disks (units of 24 disks with 6 or 10 TB per disk). Distributed processing is obtained by an equally scalable solution based on commodity servers outfitted with a number of processing units (either 12 or 40 cores), about 18 GB of random access memory per core, and solid-state drives for fast scratch space used for temporary or intermediate computation results. The processing and storage servers are co-located and connected through standard 10 GB/s switches. Each storage node is connected to two different switches to increase reliability and data access bandwidth from and to the processing nodes.

A unified view of all the files stored on the various disks attached to the storage servers is provided by a distributed file system (DFS). Various DFS types were analyzed based on available features, published performances, development continuity, support, and the possibility to run on commodity hardware [13]. The choice of a suitable DFS is difficult given that strengths and weaknesses are application dependent and often only revealed upon extensive testing in real case scenarios [14]. The systems analyzed were Lustre, EOS, Ceph, GlusterFS, Hadoop, GPFS, BeeGFS, and MooseFS. Out of these, GlusterFS was regarded as a promising candidate and was therefore further tested. While it performed mostly well in the test instance, the inflexibility of the set-up in case of storage extension made it appear less suitable for application domains requiring frequent storage extensions. Drawing an analogy between the ever increasing amount of geospatial data in particular in the form of large (raster image) data files and the data generated by high energy physics, we decided to investigate further the utilization of EOS[1] [15].

EOS is developed by the European Organization for Nuclear Research (CERN). It runs on commodity hardware and it is mainly focused on low latency, high availability, ease of operation, and low

[1] http://information-technology.web.cern.ch/services/eos-service.