Big Geospatial Data for Environmental and Agricultural Applications

Athanasios Karmas, Angelos Tzotsos, Konstantinos Karantzalos

Abstract Earth observation (EO) and environmental geospatial datasets are growing at an unprecedented rate in size, variety and complexity, thus, creating new challenges and opportunities as far as their access, archiving, processing and analytics are concerned. Currently, huge imaging streams are reaching several petabytes in many satellite archives worldwide. In this chapter, we review the current state-of-the-art in big data frameworks able to access, handle, process, analyse and deliver geospatial data and value-added products. Operational services that feature efficient implementations and different architectures allowing in certain cases the online and near real-time processing and analytics are detailed. Based on the current status, state-of-the-art and emerging challenges, the present study highlights certain issues, insights and future directions towards the efficient exploitation of big EO data for important engineering, environmental and agricultural applications.

1 Introduction

The current generation of space-borne sensors are generating nearly continuous streams of massive earth observation (EO) datasets. Shortly, high-resolution multispectral images will be available almost once a week and in some regions twice per week. In addition to open data satellite missions from national and european organizations (mostly by USA and EU), space industry and startups are increasing significantly as launches get cheaper and technology gets smaller. Several operating small inexpensive satellites (will) make geospatial data widely available for several applications. This new generation of interconnected satellites capture and transmit Remote Sensing (RS) data with a sub-meter resolution daily, allowing the monitoring of earth surface changes with greater frequency than ever before. The

Athanasios Karmas, Angelos Tzotsos, Konstantinos Karantzalos
Remote Sensing Laboratory, National Technical University of Athens, Zographou campus, 15780, Athens, Greece {thanasis.karmas;tzotsos}@gmail.com;karank@central.ntua.gr

satellite-based earth observation industry is witnessing an impressive growth, with around 260 satellite launches expected over the next decade. Moreover, the recent regulatory relaxation to sell very high-resolution satellite imagery data has opened the door for further growth where new industry participants can enter and provide diverse on-demand services which will lead to new market or applications, products and services, and even new business models.

These huge EO streams which are received through satellite downlink channels at gigabit rates, increase therefore at astonishing rates, reaching currently several petabytes in many satellite archives [74],[67],[20],[56]. According to the statistics of the Open Geospatial Consortium (OGC) the global archived observation data would exceed the one Exabyte during 2015.

However, it is estimated that most of datasets in existing satellite imaging archives have never been accessed and processed [64] apart from certain supercomputer centers. Therefore, in order to harness the full potential of these massive earth and environmental datasets, sophisticated methods for organizing and analyzing them are required [87],[67],[27], [46] and processing power, people and tools need to be brought to the data [6],[55],[28],[47].

Therefore, harvesting valuable knowledge and information from big EO data turns out to be extremely challenging [56],[57],[72],[49], while the increasing data volumes are not the only consideration. As the wealth of data increases, the challenge of indexing, searching and transferring increases exponentially as well. Open issues include the efficient data storage, handling, management and delivery, the processing of multimodal and high-dimensional datasets as well as the increasing demands for real-time (or near real-time) processing for many critical geospatial applications [74], [73],[87].

Moreover, RS datasets are multi-modal, acquired from multispectral, radar, lidar, etc. sensors. It is estimated that the archives of NASA include nearly 7000 types of EO datasets. The high dimensional ones (*e.g.*, hyperspectral imagery) contain hundreds of different wavelength data and therefore, tons of information must be stored, processed, transmitted and analysed towards harnessing the potential of these diverse and multi-dimensional datasets.

The development of novel geospatial web services [91],[85], [49] for on-demand remote sensing analysis is a key issue. Geospatial web services enable users to leverage distributed geospatial data and computing resources over the network in order to automate geospatial data integration and analysis procedures. These services should be interoperable and allow for collaborative processing of geospatial data for information and knowledge discovery. The aforementioned features can be accomplished through the utilization of the service computing and workflow technologies [84],[37].

All the aforementioned aspects are actively and intensively discussed in the scientific community and industry towards innovative solutions and cutting-edge new technology. To this end, in this chapter we review the current state-of-the-art on big data frameworks able to handle, process, analyse and deliver geospatial data and value-added products. Certain operational services are detailed with efficient implementations and architectures allowing in certain cases the online and near real-

time processing and analytics. In particular, the tasks of storing, handling, retrieving, analyzing and publishing geospatial data pose significant challenges for several aspects of the state-of-the-art processing systems, including system architectures, parallel programming models, data managing on multilevel memory hierarchy and task scheduling. Current dominating system architectures for data-intensive environmental and agricultural applications are reviewed. Recently developed cutting-edge geospatial tools for cluster-based high performance computing, cloud platforms, parallel file systems and databases are described. Examples with valuable environmental and agricultural geospatial products are also presented. Last but not least, based on the current status, state-of-the-art and emerging new challenges, the present study highlights certain issues, insights and future directions towards the efficient exploitation of big EO data for important engineering, environmental and agricultural applications.

2 Big Geospatial Data

Geospatial data (or geodata) possess qualitative and/or quantitative information along with explicit positioning and location details, such as a road network from a Geographic Information System (GIS) or a geo-referenced satellite image. Geospatial data may include additional attribute data that describe particular features found in a dataset. Geospatial data are mainly available in two formats *i.e.*, vector (2.2) and raster (2.3). The vast majority of geospatial datasets tend to exhibit certain properties that classify them in the most important, interesting and challenging category of datasets, namely, Big Data. In the following sections, various big geospatial data examples are discussed (2.1) in order to establish this perspective. Moreover, the development and usage of open geospatial standards for geospatial content, services, GIS data processing and data sharing (2.4) is presented.

2.1 Multi-V Geospatial Data

Big Data are characterised by what is often referred to as a multi-V model. In particular, focusing on geospatial data, this model encapsulates five fundamental properties of Big EO Data [53]:

- **Volume:** Big datasets are the ones that occupy very large sizes in terms of storage needed.
- **Velocity:** This property accounts for the rapidness not only of the arrival of new data to a system for insertion and processing but also to how quickly a system is able to respond to processing queries on the data submitted by the users.
- **Variety:** With the term variety the different data types, non-alligned data structures as well as inconsistent data semantics are represented.

- **Veracity:** Veracity refers to how much the data can be trusted given the reliability of its source [83].
- **Value:** This property corresponds the monetary worth that a company can derive from employing Big Data computing.

Although the choice of Vs used to explain Big Data is often arbitrary and varies across technical reports and articles across the Web - e.g. as of writing Viability is becoming a new V - variety, velocity, and volume [77], [68] are the items most commonly mentioned.

A goal of this study is to position the multi-V model that characterizes Big Data in regard to geospatial data. This is achieved through the presentation of use case examples derived from real world applications and challenges. This study aims to address both forms of geospatial data *i.e.* raster and vector data models.

Raster geospatial datasets [14] exhibit all of the properties of the multi-V model that characterizes Big Data. More specifically, raster data form big volume datasets. There is an abundance of examples that uphold this statement. In social networks there are cases of tables of incidences that their size is close to $10^8 \times 10^8$. In the earth sciences field, ESA[1] plans to host 10^{12} images that their sizes range from 500MB to 2GB or even beyond. Moreover, the velocity of raster data is tremendous. NASA's instrument MODIS (or Moderate Resolution Imaging Spectroradiometer) that is on board satellites Terra[2] and Aqua[3] captures and transmits daily almost 1TB of raster data. The distributed sensors used in radio astronomy (LOFAR: distributed sensor array farms for radio astronomy) collect 2-3 PB of data per year. Furthermore, the variety of raster data is vast as we obtain data from various sensors that have different properties and operational functionalities. Furthermore, the veracity of raster data plays an important role in their use as the collected and calculated information that is derived from these datasets need to be accompanied with quality metadata. Usually, predefined procedures are used for the calculation of faults in these datasets. However, there are cases where this is happening, and a serious issue in terms of avoiding error propagation exists.

Vector geospatial datasets are also involved in a wide range of application fields that exhibit the properties of the multi-V model. Apart from simple geospatial applications that involve vector data, the most challenging one is the management and exploration of mobility data. Mobility data [70] is ubiquitous, particularly due to the automated collection of time-stamped location information from GPS-equipped devices, from everyday smartphones to dedicated software and hardware in charge of monitoring movement in land (*e.g.* automobiles), sea (*e.g.* vessels) and air (*e.g.* aircrafts). Such wealth of data, referenced both in space and time, enables novel classes of applications and services of high societal and economic impact, provided that the discovery of consumable and concise knowledge out of these raw data collections is made possible. These data collections form big datasets. The numbers speak for themselves as for example the tracking of 933 vessels sailing in the Aegean sea dur-

[1] www.esa.int
[2] http://terra.nasa.gov
[3] http://aqua.nasa.gov

ing a 3 days period resulted in 3 million GPS recordings. All these valuable data need innovative solutions in order to be analysed. Real-time analysis of big geospatial data for the provision of alerts and alternates in case of emergency situations is of critical importance.

It is evident that geospatial data form big datasets. Several challenges are determined that must me tackled in order to be able to utilise geospatial data and create innovative, sustainable and useful geospatial data applications and services.

2.2 *Vector Data*

Many geospatial data are available in the form of vector data structures. Vector data structures are constructed from simple geometrical primitives that are based on mathematical expressions and consist of one or more interlinked nodes to represent images in computer graphics. One node determines a location in space through the use of two or three axes. In the context of geospatial data a location in space is usually determined from its geographical coordinates (*i.e.* longitude, latitude) and its height from sea level. The geometrical primitives can be a point, a line segment, a polyline, a triangle as well as other polygons in 2 dimensions and a cylinder, a sphere, a cube and other polyhedrons in 3 dimensions.

Vector data provide a way for the representation of entities of the real world. An entity can be anything that exists in a place. Trees, houses, roads and rivers are all entities. Each of these entities, apart from its location, has additional information - attributes that describe it. This information can be either in textual or numerical form.

Map data are usually represented in vector form. Roads for example are usually represented from polylines. Other geographical entities such as lakes or even civil entities such as provinces and countries are represented from complex polygons. Some entities can be represented by more than one geometrical primitive depending on the context of the representation. For example a river can be represented by either a curve or a polygon depending on the importance of keeping the river's width.

Utilizing the vector form to represent geospatial data has several advantages [24]. Two of the most important ones are the following:

- Vector data exhibit small demands in terms of storage sizes as the size of their disk footprint does not depend on the dimensions of the object that is represented.
- The focus (zoom in) in a vector data representation can be arbitrary high without altering the visual result of the represention.

2.3 *Raster Data*

Raster data are the other common form in which geospatial data are created and delivered. Raster data model defines space as an array of equally sized cells (*i.e.*

bitmaps or pixels) arranged in rows and columns and composed of single or multiple bands. Each pixel [86] covers an area of fixed size and contains a value produced by a sensor that describes the conditions for the area covered. A satellite image is a typical example of a 2 dimensional raster image.

Apart from the values of each pixel the data include the area covered by the raster image that is determined for example from the geographical coordinates (*i.e.* longitude, latitude) of its corner pixels, the spatial resolution of the image that is determined by the total number of pixels that the image is composed or more often, as far as geospatial data are concerned, by the total area covered by a single pixel and finally the spectral resolution of the image that is determined from the number of spectral bands for which the image contains information.

The larger the spatial resolution of an image, the better its quality but the cost in storage space increases proportionally. This is in contrast with the vector data model: in order to achieve better image quality more storage resources are needed. Another disadvantage in comparison to vector data is that if one wants to achieve increased zoom quality, the pixel based model imposses limitations: pixel edges appear after a certain zoom level. As a result, the perception of continuity is lost and the various objects that are depicted are not easily distinguished.

Raster data on the other hand are extremely useful [24] when there is need to present information that is continuous in an area and as a result it cannot easily be separated in entities speaking with terms of vector data. For example a valley that has great variety in colors and vegetation density is very difficult to be represented through the vector data model. Either a very simple representation would be utilized, losing valuable information as a result of the simplification, or a very complex one would be utilized so as to digitize every single detail, a task that would require a lot of time and effort. Raster data is therefore the intended solution when we need to represent areas with homogeneous characteristics as the human eye is very capable of interpreting images and distinguishing small details that would be difficult to digitise sufficiently due to their large numbers.

Raster data are not only suitable for representing real world surfaces but can also represent more abstract concepts. For example they can display the rainfall tendencies or the danger for fire manifestation in an area. In such applications every pixel of the raster image represents a different value. In the example with the fire manifestation danger every pixel can contain a value in scale from 1 to 10 for the danger in a particular area.

The common case in geospatial data applications and mainly in their presentation, is to utilize together vector and raster data as it is evident that the two representations complete one another as far as their advantages are concerned. It is common to use raster data as base layers of information and overlay them with information that is derived from vector layers.

2.4 *Open Geospatial Standards*

The vision of the research community [31] is that all available geospatial data should be accesible through the Web and that the Web consists a "place" where geospatial data are in an easy and straightforward way published, interconnected and processed towards the extraction of knowledge and information through interoperable web services and data formats. Towards this direction, the Open Geospatial Consortium (OGC) has standardized several procedures for the analysis of geospatial data through the Web. The most important and popular standards are presented in the following sections (2.4.2).

2.4.1 Open Geospatial Consortium (OGC)

OGC organization [66] was established in 1994 in order to promote collaboration between the various GIS systems and secure interoperability among them. The idea of an open GIS system is to withdraw from the model of developing GIS systems as monolithic software modules and advance towards designing and implementing a modular system that would include many and different software system modules. OGC is the result of the cooperation based on a consensus between public and private sector vendors and organizations, dedicated to the creation and management of an industrial scale architecture towards the interoperable processing of geospatial data. OGC's technical goals [18] are the following:

- A universal space-time data model as well as a processing model on these data that would cover all existing and potential space-time applications. This model is called "OGC data model".
- The definition of the specifications that would apply to all of the important database programming languages in order to enable them to implement the OGC data model.
- The definition of the specifications for each of the most widespread distributed computing environments in order to enable them to implement the OGC processing model.

OGC's technical activities span 3 different categories. These are the development of abstract specifications, the development of implementation specifications as well as the process of revising all existing specifications.

According to the goals that OGC has set, the purpose of the development of abstract specifications [65] is the creation and documentation of a conceptual model responsible for the creation of implementation specifications. Abstract specifications consist of two models: the essential model that establishes the conceptual connection between software and the real world and the abstract model that defines a final software system in a neutral way as far as its implementation is concerned (meaning that the actuals protocols needed are not defined). This gives the potential to data servers and their clients which run processing algorithms to communicate in various environments such as through the Internet, across Intranets or even in the same

workstation. Technical specifications which implement the abstract specifications in each of the most widespread distributed computing environments (*e.g.* CORBA - Common Object Request Broker Architecture environment, DCOM and Java) are available.

All the models that are included in the abstract specifications documents and in the documents of implementation specifications are developed with UML (Unified Modeling Language). The main entity of the OGC model is the "feature" that has a certain type and geometry both of which are defined by OGC itself under the "well-known structures".

2.4.2 OGC Open Standards

OGC's standards[4] are technical documents that define meticulously interfaces and encodings. Software engineers use these documents to construct or integrate open interfaces and encodings in their products and services. These standards are the main products of OGC that have been developed by the community with the purpose of addressing specific interoperability challenges. Ideally, when OGC standards are integrated in products or online services that were the result of the work of two software engineers who worked independently, the produced system components can function immediately in cooperation with each other without the need for any additional work (plug and play). OGC standards along with the technical documents that define them are available to everyone at no cost. The most popular and important interface standards of OGC are presented briefly in the following paragraphs.

2.4.2.1 WCS

OGC's WCS (Web Coverage Service) interface standard [12], defines a standard interface and functionalities that allow for the interoperable access in geospatial data that are in the form of grid coverages. The term coverage typically describes data such as remote sensing images, digital terrain models (DTM) as well as other phenomena that can be represented by numerical values at measurement points. WCS standard is in essence a web data service. It defines a service for accessing data that allows for the retrieval of grid coverages, such as DTMs, through the HTTP protocol. The response of a dedicated web server in a WCS request includes both grid coverage's metadata and the actual data that are encoded in a specific digital image format such as GeoTIFF or NetCDF image formats.

2.4.2.2 WFS

OGC's WFS (Web Feature Service) interface standard [82], defines web functionalities for retrieving and processing vector data. This standard defines procedures that allow for the discovery of available sets of features (GetCapabilities), the description of geographic features (DescribeFeatureType), the retrieval of part of the data

[4] www.opengeospatial.org/standards

through the use of a filter (GetFeature) as well as the addition, the updating or the removal of features (Transaction). All WFS services support data input and output through the utilization of the Geography Markup Language (GML) standard. Some WFS services support additional encodings such as GeoRSS and shapefiles. Users typically interact with WFS services through web browsers or GIS software. These allow them to gain access to data layers from various data sources through the Web.

2.4.2.3 WMS

OGC's WMS (Web Map Service) interface standard [16], provides a simple HTTP interface for the request and retrieval of georeferenced images and maps from one or more distributed spatial databases. The response of the database server to a WMS request is one or more images (in JPEG, PNG, etc. formats) that can be easily presented from either any web browser or from desktop applications running on a personal workstation.

2.4.2.4 CSW

OGC's CSW (Catalog Service for the Web) interface standard [61], specifies a design pattern for defining interfaces to publish and search collections of descriptive information (metadata) about geospatial data, services and related information objects. Providers of resources, such as content providers, use catalogues to register metadata that conform to the providers choice of an information model; such models include descriptions of spatial references and thematic information. Client applications can then search for geospatial data and services in very efficient ways.

2.4.2.5 WPS

OGC's WPS (Web Processing Service) interface standard [80], provides rules for the modelling of input and output data (requests and responses) for geospatial processing services and also describes the access to spatial processing functions through the Web. These functions can include all kinds of algorithms, calculations or models that have been designed to operate on geospatial data that follow either the vector data model or the raster data model. A WPS service can perform simple calculations such as the intersection of two polygons or the addition of two digital images as well as more complex ones such as the implementation of a model for the global climate change. WPS standard defines 3 basic functionalities through which all processing is performed:

- **GetCapabilities**: This functionality requests from a WPS server to respond with the features of the provided service which include the service's metadata as well as the metadata that describe all available processing functions.
- **DescribeProcess**: This functionality requests from a WPS server the description of a WPS function that is available by the service. By the provision of a particular parameter (identifier) the function that will be described is determined while there is potential for requesting the description of more than one functions.

- **Execute**: This functionality submits a request to a WPS server so as to execute a certain processing function with the provided input values and the desired output data. The request, which is an XML file, can be submitted to the server either with the GET method or with the POST method of the HTTP protocol.

WPS interface standard is extremely useful as it provides an abundance of possibilities. Some of them are the reduction of the complexity of the entire procedure of data processing, the creation of processing chains, the simplification of the management of the processing functions as well as the interoperable access to processing functions of great complexity.

2.4.2.6 WCPS

OGC's WCPS (Web Coverage Processing Service) interface standard [11], defines a high level query language that allows for the server-side processing of complex queries that are applied on multidimensional raster data. This language functions as an interface between various clients and a server and can be characterized as the SQL language for data that are in the form of a coverage. WCPS allows for the submission of on demand processing queries to the server aiming at the ad-hoc processing of coverages in order to extract various types of results such as the calculation of a variety of quality indices, the determination of statistical evaluations as well as the creation of various types of charts such as histograms. It is a functional programming language which means that it has no side effects and uses declarative semantics which leaves room for many optimizations on the server side during query execution. This language has been designed to be safe in evaluation, which means that any valid query written in WCPS is guaranteed to terminate in a finite amount of time. This property is very important in client-server environments as it ensures that the system is secured against Denial of Service (Dos) attacks at the level of one isolated processing query. The preservation of this property (*i.e.* safe in evaluation) means that the language loses some capabilities as it has limited expressive power. Explicit use of iteration (*e.g.* for-loop) and recursion are prohibited. These design choices have been taken as a client should not have unlimited power to determine what is executed on the server-side. Even though the potential for certain calculations is lost, a wide range of processing queries are still supported. For instance, transposing a matrix might not be possible but algorithms like general convolutions are still expressible.

3 Big Geospatial Data Frameworks

The increasing amount of geospatial datasets [52] is outstripping the current systems' capacity of exploring and interpreting them. The tasks of storing, handling, retrieving, analyzing and publishing big geospatial data [58] pose significant challenges and create opportunities for several aspects of the state-of-the-art big geospatial data frameworks Figure 1. Several components are included like cloud-based in-

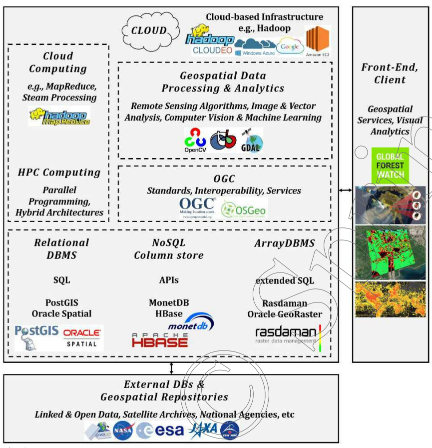

Fig. 1 The current dominating architecture and technology for big geospatial data and analytics.

frastructure, interoperability standards, parallel systems and programming models, Database Management Systems (DBMS) on multilevel memory hierarchy and task scheduling. Furthermore, the growing amount and resolution of geospatial data from remote sensing platforms and traditional geospatial information systems (GIS), as well as the geospatial data from new data sources such as social media and Internet of Things datasets, provide great opportunities to answer new and bigger questions from a geospatial perspective.

The emergence of big geospatial datasets is revolutionizing the techniques for analyzing and extracting valuable insights from these vastly expanding geospatial data streams. Rapid processing of big geospatial data with increasing volumes and complexity forms a great challenge for the existing big data frameworks. There is an urgent need for novel advances in system architecture, especially towards achieving inherent scalability of the underlying hardware and software architecture. In partic-

ular, these data-intensive systems should exhibit potential for linear scaling in order to accommodate processing of geospatial data at almost any volume.

For the purpose of meeting the near real-time processing requirement of some of geospatial applications, easy procedures for adding extra computing resources are also necessary. From a performance efficiency perspective, it is critical for data-intensive platforms to abide by the "move the code to the data" principle [34] so as to minimize data movement. Therefore, a storage hierarchy of system optimized for data-intensive computing would probably reside data locally to reduce network and system overhead introduced by data transferring.

The bulk storage demands for databases, the need for array storage models for large-scale scientific computations and large output files as well as the aggressive concurrency and per server throughput [22] are essential requirements for the applications on highly scalable computing clusters.

Currently, several available high performance platforms are employed in an effort to meet the requirements mentioned above and make sense of these big geospatial data. The most dominant choices of platforms concentrate on, namely, novel database platforms, Cluster-Based HPC (*i.e.* high performance computing) systems or supercomputers as well as Cloud-based platforms.

The aforementioned choices of platforms will be discussed in subsection 3.1. Following this discussion, an effort will occur (subsections 3.2 - 3.5), to present the most popular as well as promising platforms for handling and processing big geospatial data.

3.1 Databases, HPC systems, Cloud-based architectures

Even though the size of Big Data keeps increasing exponentially, current capability to work with big geospatial datasets is only in the relatively lower levels of petabytes, exabytes and zettabytes of data. Moreover, traditional database management tools and platforms [22] are unable to process geospatial datasets that grow so large and complex. Towards addressing this challenge new and innovative solutions have been proposed and are currently utilized. The two most popular, solid and robust ones are Database Management Systems (DBMS) that implement the array data model on the one hand and NoSQL (Not Only SQL) database management platforms on the other.

Array DBMSs are built specifically for serving big raster datasets. As raster datasets consist of pixels forming a regular grid of one, two, or more dimensions, the array data structure is the most appropriate one for modelling big geospatial raster data. Array DBMSs implement the array data model that supports arrays as first class citizens. The model is based on an array algebra [9] developed for database purposes that introduces the array as a new attribute type to the relational model. On top of the model a query language is implemented that extends standard SQL and is enriched with array operators for raster data retrieval and processing. Array DBMSs in essence store and manage structured data. Known array DBMS implementations include *Rasdaman*, MonetDB/SciQL, PostGIS, Oracle GeoRaster and SciDB database platforms.

NoSQL [36] is a current approach for large and distributed data management and database design. A NoSQL database provides a mechanism for storage and retrieval of data that is modelled in means other than the tabular relations used in relational databases. NoSQL systems are either entirely non-relational or simply avoid selected relational functionality such as fixed table schemas and join operations. The reason why mainstream big data platforms adopt NoSQL is to break and surpass the rigidity of normalized relational DBMS schemas. Many NoSQL database platforms though, still use SQL in its database systems, as SQL is a more reliable and simpler query language with high performance in stream big data real-time analytics.

NoSQL database platforms store and manage unstructured data in a way that is contrary to relational databases as it separates data storage and management into two independent parts instead of dealing with these issues simultaneously. This design gives NoSQL databases systems a lot of advantages. The main advantages are the potential for scalability of data storage with high-performance as well as flexibility for data modelling, application developing and deployment [36]. Most NoSQL databases are schema-free. This property enables applications to quickly modify the structure of data without needing to rewrite any tables. Moreover, it allows for greater flexibility when structured data are heterogeneously stored. Well-known NoSQL implementations that serve geospatial data include MongoDB, Google BigTable, Apache HBASE and Cassandra. Companies that use NoSQL database platforms include Google, Facebook, Twitter, LinkedIn and NetFlix.

Apart from database platforms a wide variety of cluster-based HPC systems including grid computing, cluster computing and ubiquitous computing are used to process big geospatial datasets and extract meaningful information. A cluster platform [74], [73] normally performs a large computational problem by the collaborative work of multiple computers (nodes) while offering a single-system image. Currently, cluster platforms are the mainstream architecture for high performance computing and large scale scientific applications. Major organizations and enterprises such as NASA [54] and Google [7] have built large cluster systems consisting of many individual nodes for the processing of geospatial data. This kind of system [58] is able to offer high-level of data capacity, throughput, and availability by virtue of the software fault tolerance, data backup and optimized system management.

HPC systems are evolving towards hybrid and accelerator-based architectures featuring multicore CPUs as well as GPUs [35]. Moreover, they are equipped with high performance Infiniband computing network with high bandwidth for achieving low latency communication between system nodes. HPC systems though, are compute-intensive oriented and both the system architecture and tools are not optimized for data-intensive applications where data availability is the main concern. Thus, in spite of the massive computational capabilities of HPC systems, effective processing of big geospatial data on existing cluster-based HPC systems still remains a challenge. Towards confronting this issue a transition to systems that are arranged in multiple hierarchical levels is taking place. These systems have a higher dimensional connection topology and multilevel storage architecture as well. As far

as performance efficiency is concerned it is of critical importance to take data locality into account. Programming for these multiple levels of locality and routes for a certain dimensionality is yet very challenging and there is a long way down to reaching viable solutions.

The development of virtualization technologies [58] have made supercomputing more accessible and affordable through the utilization of commodity hardware instead of very expensive HPC systems. Powerful computing infrastructures hidden behind virtualization software make systems behave like a true physical computer, enriched with the potential of flexibility on the specification of virtual system's details such as number of processors, memory, disk size and operating system. The use of these virtual computers is known as cloud computing [30], which has been one of the most robust Big Data techniques [78].

For real-time big geospatial data applications timeliness of system response when the volume of data is very large, is at the top priority. Cloud computing ([81], [1], [29]) integrates software, computations and user data to provide remote services through aggregation of multiple different workloads into a large cluster of processors. Cloud computing [22] not only delivers applications and services over the Internet but also has been extended to provide infrastructure as a service (IaaS), for example, Amazon EC2, platform as a service (PaaS), such as Google AppEngine and Microsoft Azure and software as a service (SaaS). Moreover, storage in cloud computing infrastructure provides a tool for storing Big Data with good scalability potential.

Cloud computing is a highly feasible technology and has attracted a large number of researchers to develop it and try to apply its solutions to Big Data problems. There is a need for software platforms and respective programming models that would take full advantage of the cloud computing principles and potential for storage and processing of big data. Towards this direction Apache Hadoop is one of the most well-established software platforms that support data-intensive, distributed and parallel applications. It implements the computational paradigm named Map/Reduce. Apache Hadoop platform consists of the Hadoop kernel, Map/Reduce and Hadoop distributed file system (HDFS) that offers strategic layouts and data replication for fault tolerance and better accessing performance, as well as a number of related projects that on top of it including Apache Hive, Apache HBase, Apache Spark, *etc.*

Map/Reduce [25] is a programming model and an execution scheme for processing and generating large volume of data sets. It was originally introduced and developed by Google and after its release it was also developed by Yahoo and other companies. Map/Reduce is based on the divide and conquer algorithm design paradigm and works by recursively breaking down a complex problem into many sub-problems (*i.e.* Map step), until they are scalable for solving directly. Then the sub-problems are solved in separate and parallel ways (*i.e.* Reduce step). The solutions to the sub-problems are then combined to give a complete solution to the original problem.

All of the well-known companies are utilising cloud computing in order to provide their services. Apart from Google, recently Yahoo has deployed its search engine on a Hadoop cluster. Moreover, Facebook and eBay also develop their large

applications at a scale of exabyte with Hadoop. In addition, for large-scale geospatial data processing, search and accessing, the HadoopGIS [3] framework is also built upon the Hadoop system.

3.2 Rasdaman

Rasdaman is a universal (*i.e.* domain-independent) Array DBMS [8], [10], [15] which offers features for big raster data storage, manipulation and processing. Domain independent means that *Rasdaman* can act as the host database platform in wide range of database applications including online analytical processing (OLAP), statistics, earth and space sciences, medical imagery, wind channels, simulations and multimedia.

Rasdaman supports multi-dimensional arrays of very large sizes and arbitrary number of dimensions that span a remarkably rich manifold of information. From 1-D time series and 2-D images to OLAP data cubes with dozens of dimensions. Due to its design aims and capabilities the system can inherently handle big satellite imaging data. *Rasdaman*'s architecture is based on a transparent array partitioning, called tiling. Conceptually, there is no size limitation for *Rasdaman* as a central DBMS of raster datasets. It features a rich and powerful query language (*i.e.* RasQL) that resembles SQL but is specifically designed and implemented for serving raster datasets. RasQL is a general-purpose declarative query language enriched with internal execution, storage and transfer optimizations. Additionally, *Rasdaman* features parallel server architecture that offers a scalable, distributed environment to efficiently process very large numbers of concurrent client requests and serve distributed datasets across the Web.

Rasdaman has proven[5] ([67], [48], [13], [50]) its efficiency and effectiveness due to its powerful query language, the transparent array partitioning that nullifies a single object's size limitations and allows for scalability as well as the feature of internally supported tile compression for reduced database size.

Moreover, *Rasdaman* implements several OGC standards towards achieving interoperability with other systems. In particular, for WCPS interface standard *Rasdaman* is the reference implementation [11]. WCPS interface standard defines a query language that allows for retrieval, filtering, processing and fast subsetting of multi-dimensional raster coverages such as sensor, simulation, image, and statistics data.

WCPS queries are submitted to the *Rasdaman* database server through PetaScope component [2]. PetaScope is a java servlet package which implements OGC standard interfaces thus allowing on demand submission of queries that search, retrieve, subset and process multidimensional arrays of very large sizes. Moreover, it adds geographic and temporal coordinate system support towards leveraging *Rasdaman*

[5] http://www.copernicus-masters.com/index.php?kat=winners.html&anzeige=winner_t-systems2014.html

into a complete and robust geospatial big data server. The *Rasdaman Community* license releases the server under the GPL license and all client parts in LGPL, thereby allowing the use of the system in any kind of license environment.

3.3 MonetDB

Another framework that has been successfully used in EO applications is MonetDB which is an is an open source column-oriented DBMS. MonetDB [60] was designed to demonstrate high performance when executing complex queries against very large databases. For example when combining tables with hundreds of columns and multi-million rows. MonetDB has been applied in a wide range of high-performance applications such as OLAP, data mining, GIS, streaming data processing, text retrieval and sequence alignment processing. It was employed successfully as a database back-end in the development of a real-time wildfire monitoring service that exploits satellite images and linked geospatial data (Subsection 4.1).

MonetDB's architecture [43] is represented in three layers, each with its own set of optimizers. The front-end is the top layer and provides query interfaces for SQL, SciQL and SPARQL general-purpose programming languages. Queries are parsed into domain-specific representations, like relational algebra for SQL, and are then optimized. The generated logical execution plans are then translated into MonetDB Assembly Language (MAL) instructions, which are passed to the next layer. The middle or back-end layer provides a number of cost-based optimizers for the MAL. The bottom layer is the database kernel, which provides access to the data stored in Binary Association Tables (BATs). Each BAT is a table consisting of an Object-identifier and value columns, representing a single column in the database.

MonetDB's internal data representation also relies on the memory addressing ranges of contemporary CPUs using demand paging of memory mapped files, and thus departing from traditional DBMS designs involving complex management of large data stores in limited memory.

Its architecture is indeed pioneering and also integrates query recycling. Query recycling [45] is an architecture for reusing the byproducts of the operator-at-a-time paradigm in a column store DBMS. Recycling makes use of the generic idea of storing and reusing the results of expensive computations and uses an optimizer to pre-select instructions to cache. The technique works in a self-organizing fashion and is designed to improve query response times and throughput.

Moreover, MonetDB was one of the first databases to introduce Database Cracking. Database Cracking [42] is an incremental partial indexing and/or sorting of the data. It directly exploits the columnar nature of MonetDB. Cracking is a technique that shifts the cost of index maintenance from updates to query processing. The query pipeline optimizers are used to massage the query plans to crack and to propagate this information. The technique allows for improved access times and self-organized behaviour.

Furthermore, MonetDB features the MonetDB/SQL/GIS module which comes with an interface to the Simple Feature Specification of the OGC and thus supports all objects and functions defined in the specification. This opens the route to host geospatial data and thus develop GIS applications. Spatial objects can, however, for the time being only expressed in the Well-Known Text (WKT) format. WKT includes information about the type of the object and the object's coordinates. The implementation of the Simple Feature Specification gives the potential to MonetDB to function as a geospatial database server.

3.4 MrGeo

The National Geospatial-Intelligence Agency (NGA) [62] in collaboration with DigitalGlobe[6], recently released as open source an application that simplifies and economizes the storage and processing of large-scale raster data, reducing the time it takes analysts to search, download, preprocess and format data for analysis.

MapReduce for Geospatial, or MrGeo, is a geospatial toolkit designed to provide raster-based geospatial capabilities (*i.e.* storage and processing) performable at scale by leveraging the power and functionality of cloud-based architecture. The software use, modification, and distribution rights are stipulated within the Apache 2.0 license. NGA has a vision for MrGeo to become the standard for storing, enriching and analyzing massive amounts of raster data in a distributed cloud environment.

MrGeo can ingest and store global datasets in the cloud in an application-ready format that eliminates several data pre-processing steps from production workflows thus freeing the user from all the heavy data logistics previously required in downloading and preprocessing the data on traditional desktop GIS systems. This allows the user to ask bigger questions of the data in the cloud, and receive just the calculated answers for their areas of interest, instead of having to pre-process all the stored data for obtaining the result.

MrGeo provides a general yet robust engine of MapReduce analytics for the processing of georeferenced raster data such as digital elevation models and multispectral as well as hyperspectral satellite and aerial imagery. It also provides a user-friendly command line syntax called Map Algebra interface that enables the development of custom algorithms in a simple scripting API and allows for algebraic math operations, focal operations(*i.e.* slope) and graph operations (*e.g.* cost distance) in order to chain basic operations into pipelines so as to create higher level analytic outputs.

MrGeo [63] is built upon the Hadoop ecosystem to leverage the storage and processing of hundreds of commodity hardware. An abstraction layer between the MapReduce analytics and storage methods provides a diverse set of cloud storage options such as HDFS, Accumulo, HBASE *etc.* Functionally, MrGeo stores large raster datasets as a collection of individual tiles stored in Hadoop to enable large-

6 https://www.digitalglobe.com

scale data and analytic services. The data storage model that maintains data locality via spatial indexing along with the co-location of data and analytics offers the advantage of minimizing the movement of data in favour of bringing the computation to the data; a standard principle when designing Big Data systems. It features a plugin architecture that facilitates modular software development and deployment strategies. Its data and analytic capabilities are provisioned by OGC and REST service end points.

MrGeo has been used to store, index, tile, and pyramid multi-terabyte scale image databases. Once stored, this data is made available through simple Tiled Map Services (TMS) and/or Web Mapping Services (WMS). Even though, MrGeo is primarily built for serving raster datasets, new features have been recently added that allow for vector data storage and processing.

3.5 CartoDB

There is a current need for flexible and intuitive ways to create dynamic online maps and design web geospatial applications. CartoDB is an open source tool that allows for the storage and visualization of geospatial data on the web and aims to become the next generation mapping platform for Big Data that follow the vector data model.

CartoDB [21] is a Software as a Service (SaaS) cloud computing platform that provides GIS and web mapping tools for displaying vector data in a web browser. CartoDB was built on open source software including PostGIS and PostgreSQL. The tool uses JavaScript extensively in the front end web application, in the back end through Node.js based APIs as well as for implementing client libraries. CartoDB platform offers a set of APIs and libraries to help users create maps, manage their data, run geospatial analytics tasks, and other functions through REST services or with client developed libraries.

CartoDB is split into four components. The first is the web application, where users can manage data and create custom maps. Users who are not technically inclined can use an intuitive interface to easily create custom maps and visualizations. Advanced users can access a web interface to use SQL to manipulate data and apply map styles using a cartography language similar to CSS (*i.e.* CartoCSS). The second component is a Maps API that acts as a dynamic tile service, which creates new tiles based on client requests. In addition to the Maps API, a SQL API is provided, where PostgreSQL-supported SQL statements can be used to retrieve data from the database. The SQL API serves data in various formats including JSON, GeoJSON, and CSV. Finally, there is the CartoDB.js library, which can be used for wrapping the Maps and SQL APIs into complete visualizations or for integrating data into other web applications [75].

CartoDB users can use the company's free platform or deploy their own instance of the open source software. With CartoDB, it is easy to upload geospatial data in various formats (*e.g.* Shapefiles, GeoJSON, etc.) using a web form and then make

it public or private. After it is uploaded, one can visualize the data in a table or on a map, search it using SQL and apply map styles using CartoCSS. It is also possible to access it using the CartoDB API and SQL API, or export it to a file.

Known users of the CartoDB platform for the production of dynamic online maps include major organizations and powerful stakeholders in the technology world such as NASA, Nokia [32] and Twitter.

4 Geospatial Environmental and Agricultural Services

Following the discussion about Big Geospatial Data and Big Geospatial Data Frameworks in the previous sections, here we will try to present the big picture about the research goal and direction of the developed and developing technology for geospatial data.

The goal is to put together big geospatial data repositories along with big geospatial data frameworks and create novel, sustainable, useful and cost-effective services. There has been intense efforts from a variety of vendors, organizations, institutions and companies to create Geospatial Environmental and Agricultural Services for a wide range of scientific and industrial fields. Delivery of reliable services through the utilization of big datasets is the main issue nowadays. This is the case as there is an abundance of geospatial data from ubiquitous and various types of sensors as well as several developed geospatial frameworks and programming models towards their efficient handling, processing and extraction of meaningful information. Most of research funding opportunities in the field of geospatial data are turning towards design and implementation of novel services in an effort to motivate the exploitation of all these huge streams of data.

Several projects administered by global partnerships are currently developing innovative Geospatial Environmental and Agricultural Services. For example, the EarthServer project[7] in its phase 1 is creating an on-demand online open access and ad-hoc analytics infrastructure for massive (100+ TB) Earth Science data based on leading-edge Array Database platform and OGC WCPS standard. EarthServer establishes several ,so-called, lighthouse applications, each of which poses distinct challenges on Earth Data analytics . These are Cryospheric Science, Airborne Science, Atmospheric Science, Geology, Oceanography and Planetary Science. In particular, for Planetary Science, PlanetServer application is being developed.

PlanetServer[8] [67], is an online visualization and analysis service for planetary data that demonstrates how various technologies, tools and web standards can be used so as to provide big data analytics in an online environment. PlanetServer focuses on hyperspectral satellite imagery and topographic data visualization and analysis, mainly for mineralogical applications. Apart from the big data analytics part PlanetServer could aid in collaborative data analysis, as it is capable of sharing

7 www.earthserver.eu

8 www.planetserver.eu

planetary data hosted on a database server and querying them from either any web client through any supported web browser or from any online processing service that adheres to OGC standards.

Currently, phase 2 of EarthServer is starting with the ambitious goal that each of the participating data centers will provide at least 1 Petabyte of 3-D and 4-D datacubes. Technology advance will allow real-time scaling of such Petabyte cubes, and intercontinental fusion. This power of data handling will be wrapped into direct visual interaction based on multi-dimensional visualization techniques, in particular: NASA WorldWind.

In this section, innovative Geospatial Environmental and Agricultural Services that span a wide range of activities will be presented in an effort to make more comprehensible the combination of geospatial data and processing platforms of different orientation towards the creation, design and implementation of geospatial services and applications.

4.1 FireHub

The developed FIREHUB[9] project proposes a fully transferable to all sites of Europe fire detection and monitoring service and provides large-scale burn scar mapping capabilities during and after wildfires as well as hourly fire-smoke dispersion forecasting [51]. It is currently fully covering the Greek territory and has already been used by emergency managers for the monitoring of wildfires.

The service has been developed following a database approach to the development of EO applications using scientific database management and linked data technologies and is implemented on top of MonetDB as the central DBMS and Strabon which is a semantic spatiotemporal RDF store. It utilizes SciQL [89] an SQL-based query language for scientific applications with arrays as first class citizens. SciQL allows MonetDB to effectively function as an array database. SciQL is used together with the Data Vault technology, providing transparent access to large scientific data repositories [90].

Data Vault is a database-attached external file repository for MonetDB, similar to the SQL/MED standard. The Data Vault technology allows for transparent integration with distributed/remote file repositories. It is specifically designed for remote sensing data exploration and mining [44]. There is support for the GeoTIFF (Earth observation), FITS (astronomy), MiniSEED (seismology) and NetCDF formats. The data is stored in the file repository, in the original format, and loaded in the database in a lazy fashion, only when needed. The system can also process the data upon ingestion, if the data format requires it. As a result, even very large file repositories can be efficiently analyzed, as only the required data is processed in

9 http://www.copernicus-masters.com/index.php?kat=winners.html&anzeige=winner_bsc2014.html

Fig. 2 The user interface of the FIREHUB service.

the database. Data Vaults map the data from the distributed repositories to SciQL arrays, allowing for improved handling of spatio-temporal data in MonetDB [44].

FIREHUB depends on the real-time processing of satellite images of different spectral and spatial resolutions in combination with auxiliary geo-information layers (land use/land cover data, administrative boundaries and roads as well as infrastructure networks data). The outputs of the service are validated fire-related products (*e.g.* hotspot and burnt area maps) for Southern Europe (Spain, France, Italy, Portugal and Greece).

User benefits include evidence-based decision making (in civil protection and business continuity management) that supports resilience against wildfires with an eye towards societal and economic welfare as well as protection of human lives, private property, and ecosystems.

4.2 Global Forest Watch

Another geospatial platform is the Global Forest Watch Service[10] [38] which is a milestone online framework powered by Google Earth Engine towards the quantification of global forest change which has been lacking despite the recognized importance of forest ecosystem services. It forms a dynamic online forest monitoring and alert system which exploits satellite technology, open data and crowdsourcing to guarantee access to timely and reliable information about forest status globally.

[10] `http://www.globalforestwatch.org/`

In particular, Global Forest Watch (GFW) is an interactive online forest monitoring and alert system designed to empower people everywhere with the information they need to better manage and conserve forest landscapes. GFW uses cutting edge technology and science to provide the timeliest and most precise information about the status of forest landscapes worldwide, including near-real-time alerts showing suspected locations of recent tree cover loss. GFW is free and simple to use, enabling anyone to create custom maps, analyze forest trends, subscribe to alerts, or download data for their local area or the entire world. Users can also contribute to GFW by sharing data and stories from the ground via GFW's crowdsourcing tools, blogs, and discussion groups. Special 'apps' provide detailed information for companies that wish to reduce the risk of deforestation in their supply chains, users who want to monitor fires across Southeast Asia, and more. GFW serves a variety of users including governments, the private sector, NGOs, journalists, universities, and the general public.

Global Forest Watch hosts a wealth of data relating to forests. Some data have been developed by the World Resources Institute or by GFW partner organizations. Other data are in the public domain and have been developed by governments, NGOs, and companies. The data vary in accuracy, resolution, frequency of update, and geographic coverage. Global Forest Watch was launched on February 20, 2014, convening government and corporate leaders to explore how governments, businesses and communities can halt forest loss. Recent reported results based on GFW and earth observation satellite data were used to map global forest loss (2.3 million square kilometers) and gain (0.8 million square kilometers) from 2000 to 2012 at a

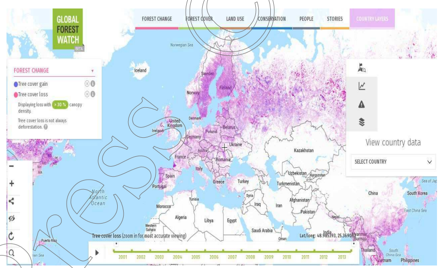

Fig. 3 The user interface of the Global Forest Watch platform powered by the Google Earth Engine.

spatial resolution of 30 meters. The tropics were the only climate domain to exhibit a trend, with forest loss increasing by 2101 square kilometers per year.

4.3 RemoteAgri

During recent decades, earth observation (EO) and remote sensing (RS) have provided valuable insights into agronomic management [76], [39]. Along with geospatial technology, they continue to evolve as an agronomic tool of significant importance which provides information to scientists, consultants and farmers about the status of their crops towards improved and optimal management decisions [88]. By precisely measuring the way in which agricultural fields reflect and emit electromagnetic energy in different spectral regions, EO satellites can quantitatively assess agronomic parameters by monitoring a wide range of variables including surface temperature, photosynthetic activity, soil moisture and weed or pest infestations.

Agricultural maps produced from RS data are useful at several stages in the agricultural value chain and allow the farmer to make rational and comprehensive decisions when planning, planting and growing new crops. Geospatial products and EO information deliver direct benefits in the agriculture sector which stem from: i) cost reductions through optimizing the application of field inputs, ii) profitability through increased yield and iii) potential competitive advantages through ameliorating crop quality and optimal decisions on crop type, variety and land cover/ use. In addition, by reducing field inputs, the run-off of fertilizers and pesticides is reduced and therefore benefit the environment.

Towards this direction, a framework for the online, on the server-side analysis of EO data has been designed and developed [47] for precision agriculture applications. In particular, the core functionality consists of the *Rasdaman* Array DBMS

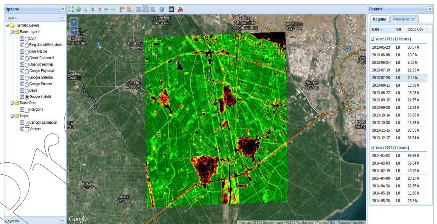

Fig. 4 The user interface of the *RemoteAgri* platform.

for storage, and the OGC WCPS for data querying. Various WCPS queries have been designed and implemented in order to access and process multispectral satellite imagery. The WebGIS client, which is based on the OpenLayers and GeoExt javascript libraries, exploits these queries enabling the online ad-hoc spatial and spectral multispectral data analysis. The developed queries, which are focusing on agricultural applications, can efficiently estimate vegetation coverage, canopy and water stress over agricultural and forest areas. The online delivered remote sensing products have been evaluated and compared with similar processes performed from standard desktop remote sensing and GIS software.

Agricultural Services: The developed services, which have been implemented through WCPS queries, are addressing important tasks like vegetation detection, canopy greenness estimation (green foliage density) and land surface temperature mapping.

Standard vegetation indices have been employed which have been proven highly successful in assessing vegetation condition, foliage, cover, phenology, and processes such as evapotranspiration (ET), primary productivity and fraction of photosynthetically active radiation absorbed by a canopy [4],[71]. Vegetation indices represent, in general, composite properties of the leaf area index (LAI) and canopy architecture, since plant canopy reflectance integrates the contributions of leaf optical properties, plant canopy architecture and the reflectance of underlying materials like soil, plant litter, and weeds [33],[39]. Therefore, the system calculates certain standard broadband indices towards optically measuring and mapping the canopy

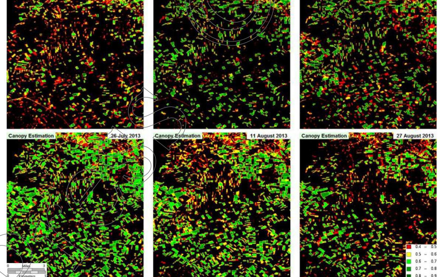

Fig. 5 Multitemporal canopy greenness levels after the application of the developed geospatial service over an agricultural area near Larissa at the Thessaly region, Greece.

greenness which can be considered as a composite property of leaf chlorophyll content, leaf area, green foliage density, canopy cover and structure.

In particular, the following services have been developed and validated in *RemoteAgri*:

1) Color Composites and Vegetation Indices: For the purpose of visualizing the stored earth observation data the creation of color composites upon a request is possible. Apart from natural color composites (e.g., RGB432 for Landsat 8), the online calculation of other composites is also available like RGB543 and RGB654. Standard broadband vegetation indices are also calculated upon request.

2) Vegetation detection: Since the different AOIs that the user can indicate (from single fields to broader agricultural areas) may contain areas with no vegetation, this service is responsible for classifying a certain image into two classes i.e., vegetation and not vegetation. This information is derived through a specific query based on the calculation of NDVI against a threshold value or through a more complex and computationally intensive multiclass classification. The result is delivered and stored in a binary imaging format.

3) Canopy greenness: Based on vegetation detection and NDVI computation, this service delivers canopy greenness maps. Optical observations on LAI can be well correlated with vegetation indices like NDVI for single plant species which are grown under uniform conditions. However, for mixed, dense and multilayered canopies, these indices have non-linear relationships and can only be employed as proxies for crop-dependent vegetation parameters such as fractional vegetation cover, LAI, albedo and emissivity. The service proceeds with a further classifica-

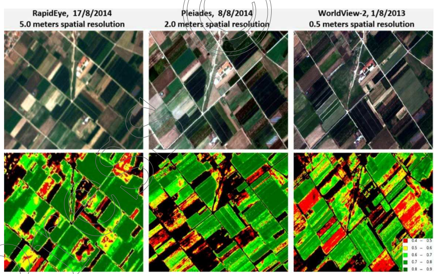

Fig. 6 Canopy greenness estimation on the Vegora agricultural region from different satellite sensors (i.e, RapidEye, Pleiades, WorldView-2) and spatial resolutions. The natural color composites are shown in the top and the estimated canopy greenness levels from the developed system, on the bottom.

tion for those areas that have been detected to contain vegetation towards estimating the different canopy greenness levels which can be associated with the vegetative canopy vigour, biomass, leaf chlorophyll content, canopy cover and structure.

4) Crop-field Surface Temperature: Various studies have demonstrated the relationship between satellite thermal data and actual rates of ET towards quantifying water consumption on specific, individually-irrigated and rainfed fields [41],[19]. Evaporation, in general, cools surfaces, so lower surface temperatures are typically associated with wetter soil and greater ET rates. In-season irrigation issues and patterns can be detected early enough before the symptoms are visually apparent in the canopy. This service calculates surface temperature using narrow band emissivity and corrected thermal radiance towards the quantification of the water use on a field-by-field basis. In particular, precision irrigation requires accurate spatial and temporal monitoring of the actual water use in order to infer water stress for irrigation decisions, aid in yield and assessment of drought conditions. Focusing on a simplified query structure and based on the detected vegetation regions, the information from the available satellite thermal band (e.g., Landsat 8 TIRS) is employed. For practical and visualization purposes, the temperature values are transformed into Celsius degrees and are then classified into 6 to 8 categories with non-overlapping intervals.

In Figure 6, the canopy greenness maps are shown based on different satellite sensors (*i.e,* RapidEye, Pleiades, WorldView-2) and spatial resolutions. Although, acquired at different dates, the information about the spatial in-field variability increases with the level of spatial detail. More specifically, for the vineyards located in this particular area (PDO zone of Amynteo where the Xinomavro variety is dominating) canopy greenness maps, with a spatial resolution between 0.5 and 5m, can depict the spatial variability that is associated with the vegetative canopy vigour, biomass, leaf chlorophyll content, canopy cover and structure.

4.4 RemoteWater

Water quality is a fundamental aspect of global freshwater resources. Information about water quality is needed to assess baseline conditions and to understand trend for water resource management. Therefore, the importance of evaluating and monitoring water quality in terrestrial reservoirs is clear and self-evident. The most commonly used methodology to examine the quality of water is through in-situ sampling and chemical analysis. In-situ sampling lead to accurate estimations but lacks in several other areas.

Therefore, reliable and low cost monitoring methods and techniques are becoming more essential. However, natural inland waters are optically complex due to the interaction of three main parameters, namely chlorophyll, inorganic suspended solids and dissolved organic matter. The estimation of water concentrations in sensitive shallow systems through the use of multispectral remote sensing imagery can be hindered due to possible errors in consistent correlation. The optical complex-

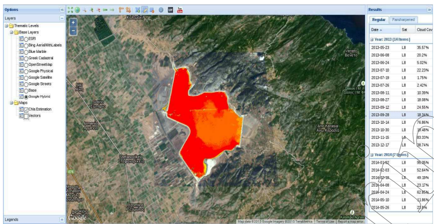

Fig. 7 The user interface of the RemoteWater platform.

ity poses many challenges to the accurate retrieval of biogeochemical parameters. The depth of the lake and the aquatic vegetation levels is of significant importance. Many standard chlorophyll-a retrieval algorithms, which are optically dominated by phytoplankton and their breakdown products, tend to fail when applied to more turbid inland and coastal waters whose optically properties are strongly inuenced by non-covarying concentrations of non-algal particles and coloured dissolved organic matter [40],[69], [79].

Based on a similar framework like the aforementioned *RemoteAgri* tool, a geospatial service able to monitor the quality of inland water was developed. More specifically, the calculation of the Normalized Difference Water Index (NDWI) is the first component of the *RemoteWater* service. Based on the calculation of the NDWI the detection of water bodies can be performed upon request. In particular, the query performs water detection based on the calculation of NDWI on the stored datasets against a threshold value *wd*. The result is delivered in a binary imaging format.

Moreover, the concentration of the photosynthetic green pigment chlorophyll-*a* in inland water bodies is a proven indicator of the abundance and biomass of microscopic plants (phytoplankton) such as unicellular algae and cyanobacteria. Chlorophyll data are useful over a range of spatial scales for monitoring the water quality and environmental status of water bodies.Both the Blue and the Coastal spectral bands, as well as the Green one for lower absorption rates have been employed. For visualisation purposes the output is determined by zoning the different estimated chlorophyll levels.

Experimental results from 6 different acquisition dates are demonstrated for the Chlorophyll Estimation service (in Figure 8). In particular, Figure 8 pictures a sensitive and shallow inland water body *i.e.,* Lake Karla in Magnesia Prefecture, Greece. One can observe that from early spring the Chlorophyll levels are gradually increasing with a peak during the summer period. In particular, during

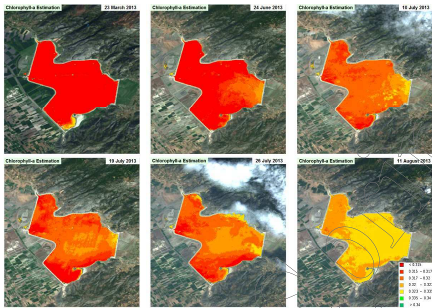

Fig. 8 Multitemporal chlorophyll concentrations after the application of the Chlorophyll Estimation geospatial service.

just a 20 days period (from middle July to middle August) Chlorophyll concentrations have increased significantly. This is in accordance with the in-situ measurements which are periodically performed according to the EU water directives. More dates and results of these particular two services can be found here: `http://users.ntua.gr/karank/Demos/BigGeoServices.html`.

5 Current Challenges

In this section, we discuss current challenges in big geospatial data and the specific issue of their management for analytics.

Data volume: How to handle an always increasing volume of geodata? This challenge is the most fundamental in the field of Big Data, since this is the definitive property of this kind of geospatial data. The most common ways to address this issue is either with data partitioning, or with remote processing of geospatial data. In the first case, geodata are partitioned in smaller, distributed pieces of data, processed in a parallel environment and the results are gathered and combined to perform final analysis. In the second case, when data cannot be partitioned (in cases when the study problem includes global features/characteristics that have to be calculated and shared globally between processes, e.g. knowledge-based classification, image understanding through context information etc.), the solution is to bring the processing to the data, through remote execution of algorithms, and very high code optimization and robust processing algorithms.

Data variety: When the data is unstructured, how to quickly extract meaningful content out of it? How to aggregate and correlate streaming data from multiple sources? Due to the hight variety of the original data, big data experts usually resort to knowledge discovery methods which refers to a set of activities designed to extract new knowledge from complex datasets. Methods like multi-modal data fusion for raster data, help remote sensing experts to extract information from multi-sensor imagery. In a similar way, advanced methods for spatial indexing (like adaptive map tiling schemes within n-dimensional data arrays), can help perform very complex spatial queries to both raster and vector datasets.

Data storage: How to efficiently recognise and store important information extracted from unstructured data? How to store large volumes of information in a way it can be timely retrieved? Are current file systems optimised for the volume and variety demanded by analytics applications? If not, what new capabilities are needed? How to store information in a way that it can be easily migrated/ported between data centres/Cloud providers? Since data storage is becoming cheaper and more efficient by the time, storage systems are becoming more efficient, but in order to solve such big data storage problems, robust distributed systems are needed with the ability to partition data efficiently, duplicate information, search and retrieve data in extremely high speeds. Such state-of-the-art systems are currently evolving, but for geospatial datasets (especially raster) these systems are not yet optimized to handle huge geospatial files in a distributed environment. Recent systems involving storage of Landsat 8 and MODIS imagery show that scalability of storage is not yet in a satisfying level. Hopefully, cloud storage systems like Rados and GlusterFS will become even more efficient in the near future, especially with spatial extensions, in order to smartly distribute information on storage based on spatial information.

Data integration: New protocols and interfaces for integration of data that are able to manage data of different nature (structured, unstructured, semi-structured) and sources. In this area, the work done by open standards in the spatial world are in the best possible direction. Through open standardisation processes, new efficient standards emerge, like GeoJSON, Vector Tiles, Irregular tiles in n-array databases, which are all under some kind of standardization process (either in OGC or under an open source license on GitHub). Given that there is ongoing work in spatial indexes, even for unstructured data (like SOLR spatial indexing or spatial extensions of CouchDB), the challenge of integration of big geospatial data seems to be in a good technological direction.

Data Processing and Resource Management: New programming models optimised for streaming and/or multidimensional data; new backbend engines that manage optimised file systems; engines able to combine applications from multiple programming models (e.g. MapReduce, workflows, and bag-of-tasks) on a single solution/abstraction. How to optimise resource usage in such complex system architectures? This problem is probably the toughest one to solve in the world of big data: by predicting the nature of data, or by robust algorithms that can extract information efficiently from spatial data, a system architect can really optimize the

workflows for data retrieval, data preprocessing, distribution of storage and finally utilization of services and delivery of analytics/useful spatial analysis results.

Visualisation and user interaction: There are many research challenges in the field of Big Data visualisation, especially for geospatial data, due to their dimensionality. First, more efficient data processing techniques are required in order to enable real-time visualisation. Choo and Park [23] appoint some techniques that can be employed with this objective, such as reduction of accuracy of results, coarsely processing of data points, compatible with the resolution of the visualisation device, reduced convergence, and data scale confinement. Methods considering each of these techniques could be further researched and improved. Visualisation for management of computer networks and software analytics [59] is also an area that is attracting attention of researchers and practitioners for its extreme relevance to management of large-scale infrastructure (such as Cloud systems) and software, with implications in global software development, open source software development, and software quality improvements.

There are still, however, many open challenges in this topic [5, 26, 17]. The list above is not exhaustive, and as more research in this field is conducted, more challenging issues will arise.

6 Future Perspectives

Current big geospatial data technologies are working towards, scalability, efficiency of storage and delivery of geospatial analytics in order to deliver spatial analysis information not only to a large audience of GIS professionals but also to users of mobile applications, social media and open data initiatives. The main interest is to move from simple monitoring and reporting on events to a collection of high-end spatial services that can provide useful analysis results to every-day needs of professionals involved in spatial information applications like agriculture, environment, forestry etc. More specifically, we are anticipating moving from data analytics and real-time applications to monitoring changes in real time but also predicting changes through robust modeling and artificial intelligence techniques.

Furthermore in geospatial big data, we are expecting in the following years a paradigm swift to every-day monitoring of the whole planet through micro-satellites, in a spatial resolution of a few meters (in the raster world), but also with the advanced technologies of hyperspectral imagery from drones and satellites. Moreover, we are starting to look at the upcoming technology of video streaming on earth observation sattelites, which will introduce a huge evolution of available spatial data from space, with concurrent explosion in data storage needs and compression rates from new algorithms.

In remote sensing, the introduction of low-cost drones, is expected to boost the acquisition of remote sensing data, especially with ultra high spatial resolution of 10 cm or even less in specific applications. It is easy to realize that this amount of image processing needed will require efforts in automation techniques, especially

for transformation of raw data to calibrated and validated information. It is expected that machine learning techniques will involve in this direction, to provide robust applications for image classification and object recognition, as well as multimodal data registration.

Important future challenges derive from the evolution in the use of maps. A map used to be something that was pictured in a book and described outside locations. This is not the case anymore. Currently, the map is something that people are continuously carrying with them in a smartphone or in a tablet pc and are heavily depending on information that is offered from services based on maps. Moreover, the new tendency lies in maps which describe interior locations such as a museum or a store. As of now portable devices with knowledge of their interior location offer significant business opportunities and as a consequence it is expected that the tremendous availability of maps of interior locations will open a new frontier in the field of indoor location based services. Moreover, technological advances have made possible the quick collection of 3 dimensional information and for this reason 3-D maps are becoming more and more common. The potential for integrating 3-D maps in web browsers in conjunction with the growing interest for the urban environment, provide the necessary boost for the creation of a new range of applications.

Furthermore, Augmented Reality (AR) manages to merge maps with the real world. AR may still be in its infancy but has the potential to change the way we perceive the world. An important aspect of AR is the personalization of maps based on each user. Personalization is based on the rapid and real-time interaction of user's applications with geospatial information. Such computational cartography with the purpose of creating personalized services for users depends on whereas the structure of geospatial information allows for its automatic processing. Evolution in the semantics domain can offer a basic ontology for such automated geospatial computations.

References

[1] Adamov A (2012) Distributed file system as a basis of data-intensive computing. In: Application of Information and Communication Technologies (AICT), 2012 6th International Conference on, pp 1–3, DOI 10.1109/ICAICT.2012.6398484

[2] Aiordachioaie A, Baumann P (2010) Petascope: An open-source implementation of the ogc wcs geo service standards suite. In: Gertz M, Ludascher B (eds) Scientific and Statistical Database Management, Lecture Notes in Computer Science, vol 6187, Springer Berlin Heidelberg, pp 160–168

[3] Aji A, Wang F, Vo H, Lee R, Liu Q, Zhang X, Saltz J (2013) Hadoop gis: A high performance spatial data warehousing system over mapreduce. Proc VLDB Endow 6(11):1009–1020, DOI 10.14778/2536222.2536227, URL `http://dx.doi.org/10.14778/2536222.2536227`

[4] Asrar G, Kanemasu E, Yoshida M (1985) Estimates of leaf area index from spectral reflectance of wheat under different cultural practices and solar angles. Remote Sensing of Environment 17:1–11

[5] Assuncao MD, Calheiros RN, Bianchi S, Netto MA, Buyya R (2014) Big data computing and clouds: Trends and future directions. Journal of Parallel and Distributed Computing (0):–, DOI http://dx.doi.org/10.1016/j.jpdc.2014.08.003, URL `http://www.sciencedirect.com/science/article/pii/S0743731514001452`

[6] Babaee M, Datcu M, Rigoll G (2013) Assessment of dimensionality reduction based on communication channel model; application to immersive information visualization. In: Big Data, 2013 IEEE International Conference on, pp 1–6, DOI 10.1109/BigData.2013.6691726

[7] Barroso L, Dean J, Holzle U (2003) Web search for a planet: The google cluster architecture. Micro, IEEE 23(2):22–28, DOI 10.1109/MM.2003.1196112

[8] Baumann P (1994) Management of multidimensional discrete data. The International Journal on Very Large Data Bases 4(3):401–444

[9] Baumann P (1999) A database array algebra for spatio-temporal data and beyond. In: In Next Generation Information Technologies and Systems, pp 76–93

[10] Baumann P (2009) Array databases and raster data management. In: T. Ozsu, L. Liu (eds.), Encyclopedia of Database Systems, Springer

[11] Baumann P (2010) The OGC web coverage processing service (WCPS) standard. GeoInformatica 14(4):447–479, DOI 10.1007/s10707-009-0087-2

[12] Baumann P (2012) OGC WCS 2.0 Interface Standard-Core: Corrigendum (OGC 09-110r4)

[13] Baumann P (2014) rasdaman: Array databases boost spatio-temporal analytics. In: Computing for Geospatial Research and Application (COM.Geo), 2014 Fifth International Conference on, pp 54–54

[14] Baumann P, Nativi S (2012) Adding big earth data analytics to geoss

[15] Baumann P, Dehmel A, Furtado P, Ritsch R, Widmann N (1998) The multidimensional database system rasdaman. In: In Proceedings of the 1998 ACM SIGMOD international conference on Management of data, ACM Press, pp 575–577

[16] de la Beaujardiere J (2006) OpenGIS Web Map Server Implementation Specification (OGC 06-042)

[17] Begoli E, Horey J (2012) Design principles for effective knowledge discovery from big data. In: Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on, IEEE, pp 215–218

[18] Buehler K, McKee L (2006) The OpenGIS Guide (third edition). In: Technical Committee, version 1, Engineering Specification Best Practices, OGIS TC Doc. 96-001

[19] Cammalleri C, Anderson M, Gao F, Hain C, Ku W (2014) Mapping daily evapotranspiration at field scales over rainfed and irrigated agricultural areas using remote sensing data fusion. Agricultural and Forest Meteorology 186(0):1–11

[20] Cappelaere P, Sanchez S, Bernabe S, Scuri A, Mandl D, Plaza A (2013) Cloud implementation of a full hyperspectral unmixing chain within the nasa web coverage processing service for EO-1. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of 6(2):408–418, DOI 10.1109/JSTARS.2013.2250256

[21] CartoDB (Retrieved 2015) https://cartodb.com/platform

[22] Chen J, Chen J, Liao A, Cao X, Chen L, Chen X, He C, Han G, Peng S, Lu M, Zhang W, Tong X, Mills J (2014) Global land cover mapping at 30m resolution: A POK-based operational approach. International Journal of Photogrammetry and Remote Sensing DOI http://dx.doi.org/10.1016/j.isprsjprs.2014.09.002

[23] Choo J, Park H (2013) Customizing computational methods for visual analytics with big data. Computer Graphics and Applications, IEEE 33(4):22–28

[24] Davis B (1996) GIS: A Visual Approach. OnWord Press

[25] Dean J, Ghemawat S (2008) Mapreduce: Simplified data processing on large clusters. Commun ACM 51(1):107–113, DOI 10.1145/1327452.1327492, URL http://doi.acm.org/10.1145/1327452.1327492

[26] Demchenko Y, Zhao Z, Grosso P, Wibisono A, De Laat C (2012) Addressing big data challenges for scientific data infrastructure. In: Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on, IEEE, pp 614–617

[27] Espinoza-Molina D, Datcu M (2013) Earth-observation image retrieval based on content, semantics, and metadata. Geoscience and Remote Sensing, IEEE Transactions on 51(11):5145–5159, DOI 10.1109/TGRS.2013.2262232

[28] Evangelidis K, Ntouros K, Makridis S, Papatheodorou C (2014) Geospatial services in the cloud. Computers & Geosciences 63(0):116 – 122, DOI http://dx.doi.org/10.1016/j.cageo.2013.10.007, URL http://www.sciencedirect.com/science/article/pii/S0098300413002719

[29] Foster I, Zhao Y, Raicu I, Lu S (2008) Cloud computing and grid computing 360-degree compared. In: Grid Computing Environments Workshop, 2008. GCE '08, pp 1–10, DOI 10.1109/GCE.2008.4738445

[30] Furht B, Escalante A (2011) Handbook of Cloud Computing. Springer

[31] Garcia-Rojas A, Athanasiou S, Lehmann J, Hladky D (2013) Geoknow: Leveraging geospatial data in the web of data. In: Open Data on the Web Workshop, URL http://jens-lehmann.org/files/2013/odw_geoknow.pdf

[32] gigaomcom (Retrieved 2015) Can you predict future traffic patterns? Nokia thinks it can. In: https://gigaom.com/2013/07/02/living-cities-lights-up-traffic-in-5-cities-with-interactive-data-visualization/

[33] Glenn EP, Huete AR, Nagler PL, Nelson SG (2008) Relationship between remotely-sensed vegetation indices, canopy attributes and plant physiological processes: What vegetation indices can and cannot tell us about the landscape. Sensors 8(4):2136, DOI 10.3390/s8042136, URL http://www.mdpi.com/1424-8220/8/4/2136

[34] Gray J (2008) Distributed computing economics. Queue 6(3):63–68, DOI 10.1145/1394127.1394131, URL http://doi.acm.org/10.1145/1394127.1394131

[35] Habib S, Morozov V, Frontiere N, Finkel H, Pope A, Heitmann K (2013) Hacc: Extreme scaling and performance across diverse architectures. In: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, ACM, New York, NY, USA, SC '13, pp 6:1–6:10, DOI 10.1145/2503210.2504566, URL http://doi.acm.org/10.1145/2503210.2504566

[36] Han J, Haihong E, Le G, Du J (2011) Survey on nosql database. In: Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on, pp 363–366, DOI 10.1109/ICPCA.2011.6106531

[37] Han W, Yang Z, Di L, Yue P (2014) A geospatial web service approach for creating on-demand cropland data layer thematic maps. Transactions of the ASABE 57(1):239–247, DOI http://dx.doi.org/10.13031/trans.57.10020

[38] Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, Thau D, Stehman SV, Goetz SJ, Loveland TR, Kommareddy A, Egorov A, Chini L, Justice CO, Townshend JRG (2013) High-resolution global maps of 21st-century forest cover change. Science 342(6160):850–853, DOI 10.1126/science.1244693

[39] Hatfield JL, Prueger JH (2010) Value of using different vegetative indices to quantify agricultural crop characteristics at different growth stages under varying management practices. Remote Sensing 2(2):562, DOI 10.3390/rs2020562, URL http://www.mdpi.com/2072-4292/2/2/562

[40] Hunter PD, Tyler AN, Prsing M, Koves AW, Preston T (2008) Spectral discrimination of phytoplankton colour groups: The effect of suspended particulate matter and sensor spectral resolution. Remote Sensing of Environment 112(4):1527–1544, DOI http://dx.doi.org/10.1016/j.rse.2007.08.003, URL http://www.sciencedirect.com/science/article/pii/S0034425707004051, remote Sensing Data Assimilation Special Issue

[41] Hwang K, Choi M (2013) Seasonal trends of satellite-based evapotranspiration algorithms over a complex ecosystem in east asia. Remote Sensing of Environment 137(0):244 – 263

[42] Idreos S, Kersten ML, Manegold S (2007) Database cracking. In: CIDR 2007, Third Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 7-10, 2007, Online Proceedings, pp 68–78, URL http://www.cidrdb.org/cidr2007/papers/cidr07p07.pdf

[43] Idreos S, Groffen F, Nes N, Manegold S, Mullender S, Kersten M (2012) Monetdb: Two decades of research in column-oriented database architectures. IEEE Data Eng Bull

[44] Ivanova M, Kersten M, Manegold S (2012) Data vaults: A symbiosis between database technology and scientific file repositories. In: Ailamaki A, Bowers S (eds) Scientific and Statistical Database Management, Lecture Notes in Computer Science, vol 7338, Springer Berlin Heidelberg, pp 485–

494, DOI 10.1007/978-3-642-31235-9_32, URL http://dx.doi.org/10.1007/978-3-642-31235-9_32

[45] Ivanova MG, Kersten ML, Nes NJ, Gonçalves RA (2010) An architecture for recycling intermediates in a column-store. ACM Trans Database Syst 35(4):24:1–24:43, DOI 10.1145/1862919.1862921, URL http://doi.acm.org/10.1145/1862919.1862921

[46] Karantzalos K, Bliziotis D, Karmas A (2015) A Scalable Web Geospatial Service for Near Real-Time, High-Resolution Land Cover Mapping. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing Special Issue on 'Big Data in Remote Sensing', (in press)

[47] Karantzalos K, Karmas A, Tzotsos A (2015) RemoteAgri: Processing Online Big Earth Observation Data for Precision Agriculture. In: European Conference on Precision Agriculture

[48] Karmas A, Karantzalos K, Athanasiou S (2014) Online analysis of remote sensing data for agricultural applications. In: OSGeo's European Conference on Free and Open Source Software for Geospatial

[49] Karmas A, Tzotsos A, Karantzalos K (2015) Scalable Geospatial Web Services through Efficient, Online and Near Real-time Processing of Earth Observation Data. In: IEEE Int. Conf. on Big Data Computing Service and Applications, IEEE

[50] Karmas A, Tzotsos A, Karantzalos K (2015) Scalable Geospatial Web Services through Efficient, Online and Near Real-time Processing of Earth Observation Data. In: (BigData Service 2015) IEEE International Conference on Big Data Computing Service and Applications

[51] Koubarakis M, Kontoes C, Manegold S (2013) Real-time wildfire monitoring using scientific database and linked data technologies. In: 16th International Conference on Extending Database Technology

[52] Kouzes R, Anderson G, Elbert S, Gorton I, Gracio D (2009) The changing paradigm of data-intensive computing. Computer 42(1):26–34, DOI 10.1109/MC.2009.26

[53] Laney D (Retrieved 6 February 2001) 3d data management: Controlling data volume, velocity and variety. Gartner

[54] Lee C, Gasster S, Plaza A, Chang CI, Huang B (2011) Recent developments in high performance computing for remote sensing: A review. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of 4(3):508–527, DOI 10.1109/JSTARS.2011.2162643

[55] Liu B, Blasch E, Chen Y, Shen D, Chen G (2013) Scalable sentiment classification for big data analysis using Naive Bayes Classifier. In: Big Data, 2013 IEEE International Conference on, pp 99–104, DOI 10.1109/BigData.2013.6691740

[56] Ma Y, Wang L, Liu P, Ranjan R (2014) Towards building a data-intensive index for big data computing - a case study of remote sensing data processing. Information Sciences DOI http://dx.doi.org/10.1016/j.ins.2014.10.006

[57] Ma Y, Wang L, Zomaya A, Chen D, Ranjan R (2014) Task-tree based large-scale mosaicking for massive remote sensed imageries with dynamic

dag scheduling. Parallel and Distributed Systems, IEEE Transactions on 25(8):2126–2137, DOI 10.1109/TPDS.2013.272

[58] Ma Y, Wu H, Wang L, Huang B, Ranjan R, Zomaya A, Jie W (2014) Remote sensing big data computing: Challenges and opportunities. Future Generation Computer Systems (0):–, DOI http://dx.doi.org/10.1016/j.future.2014.10.029, URL http://www.sciencedirect.com/science/article/pii/S0167739X14002234

[59] Menzies T, Zimmermann T (2013) Software analytics: So what? Software, IEEE 30(4):31–37

[60] MonetDB (Retrieved 2015) https://www.monetdb.org/home/features

[61] Nebert D, Whiteside A, Vretanos P (2007) OpenGIS Catalogue Services Specification (OGC 07-006r1)

[62] NGA (2014) Digitalglobe application a boon to raster data storage, processing

[63] NGA (Retrieved 2015) https://github.com/ngageoint/mrgeo/wiki

[64] Nikolaou C, Kyzirakos K, Bereta K, Dogani K, Giannakopoulou S, Smeros P, Garbis G, Koubarakis M, Molina D, Dumitru O, Schwarz G, Datcu M (2014) Big, linked and open data: Applications in the german aerospace center. In: The Semantic Web: ESWC 2014 Satellite Events, Lecture Notes in Computer Science, Springer International Publishing, pp 444–449, DOI 10.1007/978-3-319-11955-7_64, URL http://dx.doi.org/10.1007/978-3-319-11955-7_64

[65] OGC (Retrieved 20 June 2015) OGC Abstract Specifications. URL http://www.opengeospatial.org/standards/as

[66] OGC (Retrieved 20 June 2015) OGC History. URL http://www.opengeospatial.org/ogc/historylong

[67] Oosthoek J, Flahaut J, Rossi A, Baumann P, Misev D, Campalani P, Unnithan V (2013) Planetserver: Innovative approaches for the online analysis of hyperspectral satellite data from Mars. Advances in Space Research pp 219–244, DOI http://dx.doi.org/10.1016/j.asr.2013.07.002

[68] P Zikopoulos PZ C Eaton (2012) Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Companies, Inc

[69] Palmer SC, Hunter PD, Lankester T, Hubbard S, Spyrakos E, Tyler AN, Prsing M, Horvth H, Lamb A, Balzter H, Tth VR (2015) Validation of envisat {MERIS} algorithms for chlorophyll retrieval in a large, turbid and optically-complex shallow lake. Remote Sensing of Environment 157(0):158 – 169, DOI http://dx.doi.org/10.1016/j.rse.2014.07.024, URL http://www.sciencedirect.com/science/article/pii/S0034425714002739, special Issue: Remote Sensing of Inland Waters

[70] Pelekis N, Theodoridis Y (2014) Mobility Data Management and Exploration. Springer New York

[71] Pettorelli N, Vik J, Mysterud A, Gaillard J, Tucker C, Stenseth N (2005) Using the satellite-derived ndvi to assess ecological responses to environmental change. Trends in Ecology and Evolution 20:503–510

[72] Pijanowski BC, Tayyebi A, Doucette J, Pekin BK, Braun D, Plourde J (2014) A big data urban growth simulation at a national scale: Configuring the GIS and neural network based land transformation model to run in a high performance computing (HPC) environment. Environmental Modelling & Software 51(0):250–268, DOI http://dx.doi.org/10.1016/j.envsoft.2013.09.015

[73] Plaza AJ (2009) Special issue on architectures and techniques for real-time processing of remotely sensed images. J Real-Time Image Processing 4(3):191–193

[74] Plaza AJ, Chang CI (2007) High Performance Computing in Remote Sensing. Chapman & Hall/ CRC Press

[75] Repository CC (Retrieved 2015) https://github.com/cartodb/cartodb.js

[76] Rouse JW Jr, Haas RH, Schell JA, Deering DW (1974) Monitoring Vegetation Systems in the Great Plains with Erts. NASA Special Publication 351:309

[77] Russom P (2011) Big data analytics. TDWI best practices report, The Data Warehousing Institute (TDWI) Research

[78] Sakr S, Liu A, Batista D, Alomari M (2011) A survey of large scale data management approaches in cloud environments. Communications Surveys Tutorials, IEEE 13(3):311–336, DOI 10.1109/SURV.2011.032211.00087

[79] Sass G, Creed I, Bayley S, Devito K (2007) Understanding variation in trophic status of lakes on the boreal plain: A 20 year retrospective using landsat {TM} imagery. Remote Sensing of Environment 109(2):127–141

[80] Schut P (2007) OpenGIS Web Processing Service (OGC 05-007r7)

[81] Vouk M (2008) Cloud computing 2014; issues, research and implementations. In: Information Technology Interfaces, 2008. ITI 2008. 30th International Conference on, pp 31–40, DOI 10.1109/ITI.2008.4588381

[82] Vretanos PPA (2010) OpenGIS Web Feature Service 2.0 Interface Standard (OGC 09-025r1 and ISO/DIS 19142)

[83] Yu P (2013) On mining big data. In: J Wang YIJXJZ H Xiong (ed) Web-AgeInformation Management, Lecture Notes in ComputerScience, Springer-Verlag, Berlin, Heidelberg

[84] Yue P, Gong J, Di L, Yuan J, Sun L, Sun Z, Wang Q (2010) Geopw: Laying blocks for the geospatial processing web. Transactions in GIS 14(6):755–772, DOI 10.1111/j.1467-9671.2010.01232.x, URL http://dx.doi.org/10.1111/j.1467-9671.2010.01232.x

[85] Yue P, Di L, Wei Y, Han W (2013) Intelligent services for discovery of complex geospatial features from remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing 83(0):151 – 164, DOI http://dx.doi.org/10.1016/j.isprsjprs.2013.02.015, URL http://www.sciencedirect.com/science/article/pii/S0924271613000580

[86] Zeiler M (1999) Modeling Our World: The ESRI Guide to Geodatabase Design. ESRI Press

[87] Zell E, Huff A, Carpenter A, Friedl L (2012) A user-driven approach to determining critical earth observation priorities for societal benefit. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of 5(6):1594–1602, DOI 10.1109/JSTARS.2012.2199467

[88] Zhang X, Seelan S, Seielstad G (2010) Digital northern great plains: A web-based system delivering near real time remote sensing data for precision agriculture. Remote Sensing 2(3):861, DOI 10.3390/rs2030861, URL http://www.mdpi.com/2072-4292/2/3/861

[89] Zhang Y, Kersten M, Ivanova M, Nes N (2011) Sciql: Bridging the gap between science and relational dbms. In: Proceedings of the 15th Symposium on International Database Engineering & Applications, ACM, New York, NY, USA, IDEAS '11, pp 124–133, DOI 10.1145/2076623.2076639, URL http://doi.acm.org/10.1145/2076623.2076639

[90] Zhang Y, Scheers B, Kersten MNN Mand Ivanova (2011) Astronomical data processing using sciql, an sql based query language for array data. In: Astronomical Data Analysis Software and Systems XXI, vol 461, p 729

[91] Zhao P, Foerster T, Yue P (2012) The Geoprocessing Web. Computers & Geosciences 47(0):3 – 12, DOI http://dx.doi.org/10.1016/j.cageo.2012.04.021, URL http://www.sciencedirect.com/science/article/pii/S0098300412001446, towards a Geoprocessing Web