A PILOT FOR BIG DATA EXPLOITATION IN THE SPACE AND SECURITY DOMAIN

Sergio Albani (1), Michele Lazzarini (1), Manolis Koubarakis (2), Efi Karra Taniskidou (2), George Papadakis (2), Vangelis Karkaletsis (3), George Giannakopoulos (3)

(1) European Union Satellite Centre, Apdo de Correos 511, 28850 Torrejón de Ardoz, Spain
(2) University of Athens, University Campus Ilisia, 15784 Athens, Greece
(3) National Centre for Scientific Research 'Demokritos', 15310 Aghia Paraskevi Attikis, Greece

ABSTRACT

In the framework of the Horizon 2020 project BigDataEurope (Integrating Big Data, Software & Communities for Addressing Europe's Societal Challenges), a platform comprising key open-source technologies has been set up in order to meet the Big Data requirements of seven communities representing the Horizon 2020 Societal Challenges (Health, Food and Agriculture, Energy, Transport, Climate, Social Sciences and Secure Societies).
The BigDataEurope platform is currently validated by implementing relevant pilots; for the Secure Societies challenge particular importance has been given to the integration and fusion of data coming from remote and social sensing in order to add value to the current data exploitation practices.

Index Terms *- Earth Observation, Big Data, heterogeneous data sources, multi-temporal and multi-sensor analysis, space and security*

1. INTRODUCTION

The rapidly increasing amount and variety of data coming from satellites and other sources is raising new issues such as the management and exploitation of extremely large and complex datasets (Big Data); the main challenge in the Space and Security domain is to improve the capacity to extract in a timely manner operational (i.e. useful and clear) information from a huge amount of heterogeneous data.
A number of public and private initiatives are taking place at European level with the ambition of taking advantage of the opportunities offered by Big Data technologies. The Horizon 2020 BigDataEurope[1] project (Integrating Big Data, Software & Communities for Addressing Europe's Societal Challenges) aims at providing support mechanisms for all the major aspects of the data value chain in terms of employed data and technology assets, the participating roles and the established or evolving processes.
BigDataEurope is focusing on two coordination and support measures:
- Engaging with a diverse range of stakeholder groups representing the Horizon 2020 Societal Challenges[2] Health, Food & Agriculture, Energy, Transport, Climate, Social Sciences and Secure Societies.
- Collecting requirements for the ICT infrastructure needed to design, realize and evaluate a Big Data Aggregator platform infrastructure that comprises open source technologies in the framework of the lambda architecture [1].

The Secure Societies Societal Challenge has been defined for the protection of freedom and security of Europe and its citizens. Key aims of this challenge are to enhance the resilience of our society against natural and man-made disasters, to develop novel solutions for the protection of critical infrastructures, to improve border security and to support the Union's external security policies; a major activity in supporting these aims is the provision of geospatial products and services, mainly resulting from satellite data.
The pilot for Secure Societies implemented to validate the BigDataEurope platform is focusing on the integration and fusion of data coming from remote and social sensing in order to add value to the current data exploitation practices; this is key in the Space and Security domain, where useful information can be derived not only from satellite data but also from data coming from social media and other sources.

[1] BigDataEurope has received funding from the Horizon 2020 programme under grant agreement n° 644564

[2] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges

2. PILOT DESCRIPTION

According to the user requirements collected from a number of stakeholders in the Secure Societies domain during the first phase of the BigDataEurope project, there is a need for exploiting the increasing amount of data coming from space and other sources, with a major contribution of open data and tools. Automatic tools for data management and processing are one of the key aspects, where the adopted solutions have to be integrated in the whole data chain in order to reduce the human effort, to reduce the amount of needed economic resources and to efficiently perform data analysis.

The BigDataEurope pilot for Secure Societies has been developed in the Space and Security domain following the above requirements; it covers all the issues related to Big Data, namely Volume (large satellite images), Variety (heterogeneous data such as satellite images and textual content), Velocity (fast-paced social data and news stream), Veracity (cross-verification of the sources) and Value (adding useful information).

The pilot considers two different workflows of data:

- The first workflow, called the **change detection workflow,** ingests satellite images to detect areas with changes on land cover or land use by using change detection techniques; the identified Areas of Interest are then associated with social media and news items and presented to the end-user for cross-validation.
- The reverse procedure is applied to the second workflow, called the **event detection workflow**. Event detection is triggered by news and social media information, where trending topics (i.e. document clusters) with geospatial connotation constitute a time- and space- localized event; provided such event, the corresponding satellite images are acquired and processed in order to check for changes in land cover or land use.

The pilot has been developed on the basis of existing tools targeted to expert users and suitable for small-scale (i.e. serial) processing; in the context of the pilot, the functionality of these tools becomes fully automated and parallelized according to parallelization and distributed storage principles for higher throughput. For remote sensing, specific tools for image (pre-) processing have been adapted from the Sentinel Application Platform (SNAP)[3]. For social sensing, a set of crawling and machine learning tools lying at the core of the NewSum summarization [2] application were employed; these tools involve specialized techniques for gathering news items and social media data from the Web and for effectively clustering them into events using text mining methods.

3. PILOT ARCHITECTURE

The architecture of the pilot was designed to accommodate both workflows and consists of the three components depicted in Figure 1: the user interface, which is a modified version of the web-application Sextant [3]; the storage component, which involves Apache Cassandra[4] and Strabon [4]; the core component, which consists of two modules, one for change detection in land cover / land use from satellite images and one for event detection in news items and social media.

The first component runs on the client-side (i.e. on the user local computer) and constitutes a web application that can be deployed on a variety of platforms. The last two components run on the server-side (i.e. on the BigDataEurope infrastructure) in order to offer scalability and high efficiency.

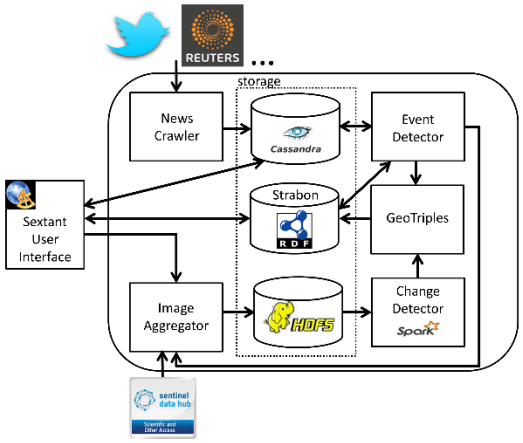

Figure 1: Pilot architecture

For the **change detection workflow**, Sextant offers an interface similar to Google Earth[5], allowing users to select an Area of Interest by forming a rectangle on an Earth map or inserting the geographic coordinates of such area. In both cases, the coordinates of the specified area are extracted and forwarded to the core component of the pilot. The user has the possibility to determine the dates of interest, i.e. one date before and one after the change should have taken place.

The Image Aggregator receives this information as input and issues the appropriate query to the Sentinel Scientific Data Hub[6] repository of satellite images. The parameters of the query are automatically configured, ensuring that only images (within the specified dates) with specific characteristics (e.g. resolution, polarization, orbit and view angle) are downloaded and stored as one file in the local file system.

The selected images are then compared by the Change Detector (implementing the functionalities offered by the ESA SNAP toolbox) in order to identify areas with changes in land cover or land use. At the moment, the

[3] http://step.esa.int/main/toolboxes/snap

[4] http://cassandra.apache.org

[5] https://www.google.com/earth

[6] https://scihub.copernicus.eu

Change Detector is only able to perform a binary classification, distinguishing areas into changed and unchanged ones; in the future it will be enriched with the ability to perform a classification in order to identify the type of change (e.g. an area with forest which becomes urban or an area with water where a manmade structure is built).

In this process, GeoTriples [5] is used in order to convert all the information that have been produced by the Change Detector (e.g. coordinates and type of change) into RDF statements (triples). Subsequently, GeoTriples stores the resulting triples into Strabon, an effective spatiotemporal RDF store that supports the main extensions of SPARQL, offering a large variety of queries over evolving geospatial data.

Finally, Sextant retrieves the stored geocoded areas from Strabon and presents them to the user. It also gets from Strabon the textual events identifiers related to the coordinates and geo-locations of the change. Then it can retrieve from the storage of the event detection workflow (Apache Cassandra) the event content, essentially links to the original news items and content. This is done through a SPARQL query in Strabon, where the crawlers and clusterers of the textual information gather items on a near real-time fashion, updating Strabon with any events detected. Consequently Strabon has all the information to answer queries related to historic data and to current events (e.g. within the last day), given a geo-location specification.

For the **event detection workflow,** the input is automatically given by the news stream, i.e. the RSS feeds from Reuters[7] and the status updates from Twitter which are continuously gathered from the Web by the News Crawler. The Twitter crawler (or *listener*) follows either the Twitter Stream API, or can use using specific keywords (Search API) or user accounts (e.g. news agencies).

The gathered content of the news stream is regularly processed at specific short time intervals by the Event Detector in order to cluster news items into events, based on topic clustering (i.e. news items talking about the same topic form an event cluster). Named Entity Recognition (NER) is used to determine the locations mentioned in the news items as well as in tweets and events; these locations and any explicit location annotation by the news publisher are mapped to geocoded locations through Strabon. Users can specify date and location for retrieved events and, in the future, also topics by appropriate keywords. The geocoded Areas of Interest that correspond to the detected events are forwarded to the change detection workflow in order to verify the events in the satellite images. The rest of the workflow operates as described above, with the changes (if verified) displayed to the user via Sextant.

[7] http://www.reuters.com

4. DATA

Remote and social sensing data have been selected for the pilot to address the fusion of information from heterogeneous sources.

For remote sensing, Sentinel-1 (a satellite launched in 2014 carrying a C-band Synthetic Aperture Radar) Level-1 GRD (Ground Range Detected) images were chosen for this first implementation phase. Sentinel-1 will benefit, among others, services related to: monitoring land-surface for motion risks; mapping for forest, water and soil management; mapping to support humanitarian aid and crisis situations [6]. The data acquisition can be configured with different modes: the default mode over land has a swath width of 250 km and a ground resolution of 5 x 20 m, while the Stripmap mode provides a continuity of ERS and Envisat data offering a 5 x 5 m resolution over a narrow swath width of 80 km. The access and use of Copernicus Sentinel Data and Service Information is regulated under EU law; the free, full and open data policy adopted for the Copernicus programme foresees access available to all users for the Sentinel data products, via a simple pre-registration on the Sentinel Scientific Data Hub.

With regard to social sensing, the focus is on two complementary sources of information: social media, represented by Twitter, and news agencies, represented by Reuters. The former involves plain text (Twitter messages) along with metadata in JSON format, while the latter includes news articles that are made available through RSS feeds in XML format. Recently, Twitter has emerged as a major platform for on-time detection of events, both in research and in industry (e.g. [7]). This pilot uses the free Twitter Public Streams API[8], which provides a random sample of its content. Twitter provides many options for parameterization, thus allowing the pilot to make the most of the retrieved content even if this is a subset of the tweets generated. For example, broad Areas of Interest can be specified, or specific users can be monitored without limitation. No private statuses are retrieved by the Public API. The content provided by news agencies through public RSS feeds is also free of charge. Within the pilot the content is used for the extraction of metadata, while the user is pointed to the original source for more information, to abide by intellectual property rights requirements. Given that all major news agencies release their content as RSS feeds, the pilot approach can be extended straightforwardly to more public content.

5. PRELIMINARY RESULTS

The main effort in this stage of the pilot deployment has been dedicated to the change detection workflow and, in particular, to the Change Detector module. In fact this

[8] https://dev.twitter.com/streaming/public

module has to process in each user request at least two very large satellite images, which occupy several GBs even after compression. The Change Detector module implements in the pilot the functionalities offered by the ESA SNAP toolbox both for the pre-processing step (i.e. co-registration) and the change detection algorithm, achieving high effectiveness in terms of precision and recall. Nevertheless this implementation still requires a significant amount of time for processing a pair of images as a specific parallelization approach to increase efficiency has still to be developed.

To scale the requirement of serving numerous user requests at the same time in the context of Big Data, the functionality of the Change Detector has been parallelized following the straightforward approach that is implemented in Calvalus[9] by the developers of SNAP: a separate node using the MapReduce framework has been assigned to every set of images. This means that if the pilot runs on a cluster with N nodes, this approach is able to process N-1 different user requests in parallel (as one of the nodes operates as the master). The only difference is that the current state-of-the-art in MapReduce, namely Apache Spark[10], has been used, whereas Calvalus employs Apache Hadoop[11] for this purpose. For the event detection workflow, the News Crawler uses the Cassandra storage to persist data and metadata, while the NewSum clustering functions over Apache Spark.

6. Conclusions And Future Works

In the framework of the Horizon 2020 BigDataEurope project, a pilot for Big Data exploitation in the Space and Security domain (deployed on a Big Data platform) has been presented. The pilot considers the fusion and analysis of information from remote sensing (satellite data) and social sensing (news from Reuters and Twitter), where the analysis of satellite images to detect areas with changes on land cover or land use (e.g. construction of critical infrastructures or exploitation of natural resources) can be associated with information provided by social media and news items.

The next steps of the deployment will be focused on the optimization of the whole data management chain. A more elaborate parallelization approach adapting each individual stage in the functionality of the Change Detector to the MapReduce framework will be developed. In this way, the processing time for a pair of images will be reduced to a significant extent that is almost linear to the number of available resources/nodes. This approach involves higher network overhead, due to the distribution of tiles among the available nodes, but its throughput is expected to be higher. Qualitative control on the final output will be performed, e.g. the verification of the change detection accuracy. In the future, once the workflows will be in use, the possibility to ingest other image processing algorithms (e.g. on the Change Detector module) and to process other types of images (e.g. Sentinel 1 Single Look Complex images and Sentinel 2 data) will be considered. Concerning the textual aspect of analysis, online text clustering and summarization methods using a distributed paradigm (e.g. using Apache Spark) will be evaluated in order to increase the throughput and scalability of the system (in alignment to the BigDataEurope aims). Finally, Sextant will be extended with a keyword-search functionality enabling users to filter the detected events according to topics of interest, in addition to the current filtering possibilities which rely on space and time constraints.

[9] http://www.brockmann-consult.de/calvalus

[10] http://spark.apache.org

[11] https://hadoop.apache.org

7. References

[1] N. Marz, and J. Warren, Big Data, *Principles and best practices of scalable realtime data systems*, Manning Publications, USA, 2014.

[2] G. Giannakopoulos, G. Kiomourtzis, and V. Karkaletsis, "NewSum: “N-Gram Graph”-Based." *Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding*, pp. 205-230, 2014.

[3] C. Nikolaou, K. Dogani, K. Bereta, G. Garbis, M. Karpathiotakis, K. Kyzirakos, M. Koubarakis, "Sextant: Visualizing time-evolving linked geospatial data", *Journal of Web Semantics*, pp. 35-52, 2015.

[4] K. Bereta, P. Smeros and M. Koubarakis, "Representing and Querying the Valid Time of Triples for Linked Geospatial Data", *Proceedings of the 10th Extended Semantic Web Conference (ESWC 2013)*, Montpellier, France. May 26-30, 2013.

[5] K. Kyzirakos, I. Vlachopoulos, D. Savva, S. Manegold, M. Koubarakis, "GeoTriples: a Tool for Publishing Geospatial Data as RDF Graphs Using R2RML Mappings", *International Semantic Web Conference (Posters & Demos)*, 2014.

[6] R. Torres, P. Snoeij, D. Geudtner, D. Bibby, M. Davidson, E. Attema, P. Potin, B. Rommen, N. Floury, M. Brown, I. Navas Traver, P. Deghaye, B. Duesmann, B. Rosich, N. Miranda, C. Bruno, M. L'Abbate, R. Croci, A. Pietropaolo, M. Huchler, F. Rostan, “GMES Sentinel 1 Mission”, *Remote Sensing of Environment*, 120, pp. 9-24, 2012.

[7] T. Sakaki, M. Okazaki, Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors", *Proceedings of the 19th international conference on World Wide Web*, New York, USA, 2010.