EARTH OBSERVATION SYMPOSIUM (B1)
Big Data, Data Cubes and new platforms to exploit large-scale, multi-temporal EO Data (6)

Author: Dr. Stefano Speretta
Delft University of Technology (TU Delft), The Netherlands

Mr. Anatoly Ilin
TU Delft, The Netherlands

SCALABLE DATA PROCESSING SYSTEM FOR SATELLITE DATA MINING

Abstract

Distribution and processing of satellite data, being it engineering telemetry or scientific measurements, still poses several challenges. The recent appearance of big constellations dramatically increased the amount of data available, actually making the challenges in data processing and mining even more pressing. A similar problem was also encountered by web companies such as Google and Facebook that had to focus on better and faster archiving and processing the vast amount of data they were collecting.

By looking at the problem from the data processing point of view, in this paper we present the development of a data distribution and processing architecture suited for small satellite operations. By focusing first on the current trends in distributed ground systems for single satellites and massive constellations, we analyze the requirements from the users prospective. This mandated a system that allows users to receive data coming from multiple ground stations and satellites in quasi real-time. Many possible satellite data users require also quick and automatic data retrieval, as compared, for example, to big scientific missions where data can only be selected from a catalog and may be delivered to the users with considerable delays.

This paper shows the evolution of the system from its first version that used custom-developed application running on a single machine, to a fully scalable and distributed architecture. It was found that scalability is critical to accommodate a big number of ground stations, leading to high peak network loads. Our solution benefits heavily from the developments in data analytics developed by web companies such as Apache Kafka and NoSQL databases, which have to process and categorize petabytes of data daily.

Beside a fast and scalable data back-end, we focused on an efficient data processing and dissemination system that makes heavy use of distributed database systems, peer-to-peer communication to deliver processed data to end users within seconds from reception. Tests performed injecting data of previous space missions (to simulate real data reception) showed limited delays between several interconnected nodes. The use of distributed database systems, rather than real-time data links, allowed to couple the system directly to data processing engines, like Apache Hadoop, commonly used to perform data mining. This approach provides a scalable and reliable solution based on existing frameworks and adds an efficient data distribution system for very quick data processing.