EPSC Abstracts
Vol. 13, EPSC-DPS2019-1997-1, 2019
EPSC-DPS Joint Meeting 2019

EPSC
EPSC-DPS Joint Meeting 2019
15–20 September 2019 | Geneva, Switzerland

Distributed framework for Space Weather forecasts

Angelo Fabio Mulone (1), Marta Casti (1), Roberto Susino (2), Rosario Messineo (1), Ester Antonucci (2), Gabriele Chiesura (1), Daniele Telloni (2), Ruben De March (1), Enrico Magli (3), Alessandro Bemporad (2), Gianalfredo Nicolini (2), Silvano Fineschi (2), Filomena Solitro (1) and Michele Martino (1)
(1) ALTEC S.p.A., Corso Marche 79, 10146 Torino, Italy (angelo.mulone@altecspace.it)
(2) INAF-OATo, Via Osservatorio 20, 10025 Pino Torinese, Torino, Italy (roberto.susino@inaf.it)
(3) Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy (enrico.magli@polito.it)

Abstract

HDS (Heliospheric Data System) is a system designed and implemented to provide space weather services. The main system goal is to reduce the time between the space weather services definition and their activation in operating environment. It is capable to manage and process near-real time data. Tens of different data sources, related to past and current missions, have been integrated. Data managed by the system have been described using standard data models. Big data technologies have been exploited to deal with the challenges of big data management and processing. The first version of the system provided medium and short-term forecast of geo-effective space weather events like the coronal mass ejections (CMEs).

1. Introduction

Sun activities can generate disturbances on Earth magnetosphere, ionosphere and thermosphere which can influence the functioning and the reliability of space and ground based systems and services. This group of activities is considered ‘space weather’. To predict solar phenomena, it is important to observe those signals that can be found in different datasets acquired by space and ground-based instruments. To support solar phenomena prediction, HDS was developed within the Space Weather Center project, a joint effort between ALTEC and INAF-OATo. The management framework used in HDS is the ALTEC Space Data Processing (ASDP) that is characterized by flexibility and extensibility. It is distributed framework easily extensible in terms of software to integrate, supported metadata and processing resources. Three different processing pipelines have been integrated in HDS: remote-sensing, in-situ and deep learning. The first one provides medium-term forecasts (by less than ~2 days) while the other two short-term forecast (by less than ~2 hours).

2. System Architecture

HDS exploits ASDP framework, whose architecture, reported in Figure 1, foresees different data stores:

- Data stores for input data
- Data store for metadata
- Data stores for processing and output

The metadata repositories are based on Elasticsearch that was chosen due to its flexibility.

The processing and output data stores are dependent from the processing itself, ASDP is able to support filesystems or databases (SQL or NoSQL). All data stores are handled by the Product Manager component that provides a transparent data access service.

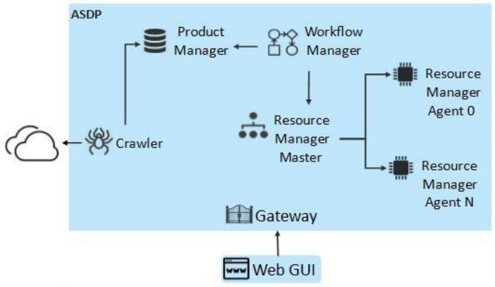

Figure 1: System Architecture.

2.1 ALTEC Space Data Processing

ASDP is a distributed data processing framework designed for providing a flexible system capable to handle and process a large variety and amount of data. ASDP allows integrating both existing and new coded algorithms, enabling automatic processing of large datasets and complex pipelines. ASDP exploits container technology and its container can be deployed through docker-swarm or Kubernetes. This simplifies the deployment in any distributed environments and allows the expansion at runtime. ASDP uses:

- AKKA framework for messaging and cluster managing
- Elasticsearch as metadata repository
- Apache Cassandra as data store
- Apache Spark as advanced data analysis framework
- Jupyter framework as tool for offline analysis.

3. Near real time data management

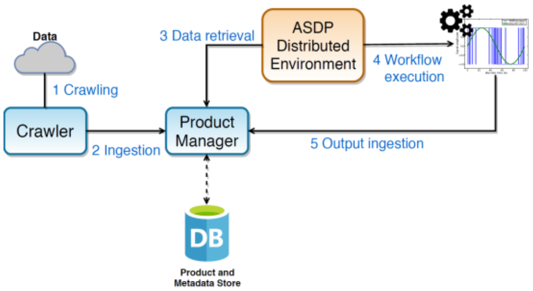

Figure 2: Data Flow.

Figure 2 shows the data flow that starts with crawling and ingestion. The crawler component queries remote repositories for data products and triggers internal ingestion and processing pipelines. The products are stored in an archiving file system shared with all the nodes of ASDP processing cluster, while metadata are stored in Elasticsearch where it is possible to enable and schedule snapshots of the repository.

3.1. Data Integration and data model

During data products ingestion metadata are extracted. Metadata is a key factor for organizing solar dataset and data model used is SPASE. It is a set of terms and values along with the relationships between them that allows describing all the resources in a heliophysics data environment. SPASE aims at unifying and improving existing Space and Solar Physics data models. In order to better describe the output generated by the pipelines, we integrated SPASE with ESPAS data model. As a matter of fact, ESPAS allows describing better the processing, while SPASE describes better the structure of the data.

4. Data Processing

Three different pipelines are integrated in HDS: remote sensing, in-situ and neural networks. This have been possible thanks to ASDP flexibility. Remote sensing pipeline exploits the SolarSoft, a system based on IDL built from libraries related to several solar mission. The ingestion of the latest available images starts the pipeline that is composed by the following steps: calibration, detection, identification and propagation. The images are calibrated in the first step, during the detection step images are analyzed and an event of interest is detected. The identification classifies an event and the propagation extracts the physical parameters and computes its propagation time until it reaches L1 and the probability of impact on earth. After the ingestion of real time data the in-situ pipeline starts. Each pipeline run processes the latest 28 days of data. The remote sensing and the in situ pipeline outcomes are then compared in order to improve the final event forecast. The recurrent neural network pipeline uses TensorFlow library and python scripts. This pipeline processes data in real time using a model previously trained to perform predictions.

5. Summary and Conclusions

The capacity to manage different data products, different data stores, different frameworks and libraries allows a lot of improvements in terms of algorithms and services. New versions of both remote sensing and in situ algorithms will be integrated as soon as possible. Moreover the crosscheck between these two pipelines shall be developed. Furthermore, new deep learning pipelines could be developed and integrated, processing also image products instead of only on numerical data.

Acknowledgements

The authors would like to acknowledge Prof. Mauro Messerotti, Senior Advisor for Space Weather, INAF Science Directorate.

References

[1] Mulone, A. F., Casti, M., Susino, R. et al.: Near-real time data management and processing system, to develop and validate space weather services, Proceedings of the Conference on Big Data from Space (BiDS'19), Munich, Germany, 19-21 February 2019.

[2] Gormley, C. and Tong, Z., Elasticsearch: The Definitive Guide, O'Reilly Media, 2015.