ANALYSIS OF BIG DATA FROM SPACE

J. Tan [a, b, *], B.Osborne [a]

[a] International Space University, 79, Route du Rhin, Illkirch Graffenstaden, Strasbourg, France - (juan.tan, barnaby.osborne)@community.isunet.edu
[b] Beijing Institute of Tracking and Telecommunication Technology, Beiqing Road, Haidian District, Beijing, China – tanjuan3557@163.com

Commission IV, WG IV/3

KEY WORDS: Space Data, Big Data, Cloud Computing, Data Mining, Platform Design, Data Intensive

ABSTRACT:
Massive data have been collected through various space mission. To maximize the investment, the data need to be exploited to the fullest. In this paper, we address key topics on big data from space about the status and future development using the system engineering method. First, we summarized space data including operation data and mission data, on their sources, access way, characteristics of 5Vs and application models based on the concept of big data, as well as the challenges they faced in application. Second, we gave proposals on platform design and architecture to meet the demand and challenges on space data application. It has taken into account of features of space data and their application models. It emphasizes high scalability and flexibility in the aspects of storage, computing and data mining. Thirdly, we suggested typical and promising practices for space data application, that showed valuable methodologies for improving intelligence on space application, engineering, and science. Our work will give an interdisciplinary knowledge to space engineers and information engineers.

1. INTRODUCTION

Since 1960s, mankind has attained vast accomplishments in exploring outer space. The aim of space activities is to facilitate our life on the earth and to explore our frontiers as far as we can. Besides breakthrough in space technologies, massive data has been collected through various mission. To maximize the investment in every space program and mission, the data need to be exploited to the fullest.

Big data technology is the product of information technology which aims to meet the challenges faced by increasing amount of information in various fields. It has characterized features of big data and thus is dealing with problems in storage, processing, distribution, and analysis. Space data has common characteristics of big data 5Vs (volume, velocity, variety, veracity, and value) (Ishwarappa & Anureadha, 2015). But space data also have their unique features, comparing to other daily-life big data such as consumer data, on collecting, storing, organizing, and processing. To use space data efficiently and get most outcomes of space activities, we should display the distinctions of space data both in structures and in application models. A unique platform taking account of these factors is also needed to facilitate space data application.

Apart from these, another field worth of study is the data mining techniques. Many emerging use cases have displayed their value in science research and supportive decision. On the other hand, big data and related technology have also enabled much more applications and researches of data mining technology in professional field such as space engineering and space science.

Realizing the potential value of space data, we try to propose a platform for space data application especially in digging potential knowledge from space data. For better understanding value behind the space data and inspire innovation ideas, we also want to give some typical and promising practices on space data application, which showed valuable techniques for improving intelligence on space application, engineering and science.

Another sub-objective of this paper is to mitigate gaps between space engineering and information engineering which is essential for widespread application of big data from space in information paradigm.

2. RELATED WORK

The conception of big data from space is formed by the gaining momentum of the space data as well as the technological breakthroughs in big data. ESA, the Joint Research Centre(JRC) of European Commission and the European Union Satellite Centre have organized conference on big data from space in 2014 and 2016. The goal of the conferences is to gather the wisdom of researchers, engineers, developers and users in the area of big data from space and focuses on the whole data life cycle, ranging from data acquisition to data management, analysis and exploitation (Soille & Marchetti, 2016).

SSP16 Students of International Space University have carried out their team project study on space big data and presented a comprehensive analysis of activities of stakeholders within the space big data value chain as well as key challenges (ISU SSP16, 2016). In mission and project area, platforms for space data have been tested to take the advantages of big data technology to space mission control. The head of ESAC (ESA's European Space Astronomy Centre) Science Data Center has called for new paradigm to meet the challenges of GAIA

* Corresponding author

archived data (Arviset, 2016). Kaethler also discussed the big data architecture, design and evolution to leverage ESOC (European Space Operation Centre)'s existing big data cluster (Kaethler, 2016).

In general, although researchers have been aware of the importance of the big data from space, and have made great efforts on the procedure, until now in this area no unified platform has been applicable and much of data value hasn't been exploited.

3. BIG DATA FROM SPACE

Until now, space program and mission can be generally typed as earth observation, space science, microgravity, human flight, while space data can be catalogued as two: operation data and mission data. We are here to give brief classification of these data as well as reveal their characteristics.

3.1 Operation Data

Basically, the data flow over a space link is made of Telemetry(TM) and Telecommand (TC) data, while the TC data is split into reconfiguration command and application-specific command; the TM data include orbit data, payload data, etc (ESA, 2014). For the advantages of discussion, we defined spacecraft data except payload data as operation data. The features of operation data are listed in table 1.

Generating		Any space mission with TT&C subsystem
Storing		Mission operation control centre
5Vs	**Volume**	Low, but massive records
	Velocity	High during the operation
	Variety	Low, but vary among operators
	Veracity	High
	Value	Medium, need to be mined

Table 1. Features of operation data

3.1 Mission Data

Space mission have developed from earth observation program at first stage which was followed with space science exploration to study phenomena occurring in outer space and microgravity experiments (including human space mission) which help us to understand the effects of gravity.

Earth observation is a long-term process and most of its missions are non-real time but need highly accurate data to distinguish objects as required. Getting contact with the data services of earth observation program, we can have features of earth observation data listed in table 2

Generating		Earth observation space mission
Storing		Mission data centre or archived centre
5Vs	**Volume**	high
	Velocity	High, with the development of spacecraft and data sharing
	Variety	Medium, vary among sensor types
	Veracity	High, with auxiliary data
	Value	High, especially with other source data

Table 2. Features of earth observation data

Space faring countries have conducted a series programs and mission to get knowledge about planets and solar system around us which is the expansion of earth observation in sensor apparatus but has less maturity. Getting contact with these space science data services, we can have a view of their features in table 3.

Generating		Solar planets and exoplanet observation system including terrestrial and space observatories
Storing		Specified research centre of principal investigators
5Vs	**Volume**	Medium, but up-growing with the mission
	Velocity	Pulsed velocity, high when get near objectives
	Variety	High, because of different detection methods
	Veracity	Medium, without truth test methods
	Value	High, the only way of research

Table 3. Features of space science data

Study of microgravity is to learn its effect on people and equipment mainly in following ways (NASA, 2012):

- The effect on astronauts' health/safety and how to control the effect.
- Some phenomena which can't take place on earth and its application.

Features of data can be concluded after analysis of microgravity data generated from different experiments which is listed in table 4.

Generating		Experiments carried out in space station, space shuttle, parabolic flight, drop tower and etc
Storing		Specified research centre of principal investigators
5Vs	**Volume**	Medium
	Velocity	Medium
	Variety	High, because of different application fields
	Veracity	High, with controlled experiment environment
	Value	High, for long term human exploration

Table 4. Features of microgravity data

4. APPLICATION MODELS OF SPACE DATA

Data application model is related to problems: how to access data, how to use data, is there any restriction? For various types of space data discussed before, application models are different due to their processing.

4.1 Operation data application model

For operation data, the application model is relatively simple. Usually they are received by ground stations and transmitted to the responsible spacecraft operation centre in real-time; after automatic procession in operation centre they are used locally in real-time without further distribution; they are stored on-line in real-time and after a period of task time they will be stored off-line. Sometimes for emergency diagnosis, operation data will be shared among cooperative operation centres.

4.2 Mission data application model

Similar to operation data, mission data are also received by ground stations and transmitted to operation centre. But after separating from data of platform, mission data will be sent to other professional data processing centres according to their disciplines such as earth observation, space science, microgravity, where they will be processed to basic level data for high level of application by users. Sometimes for the sake of redundancy and reducing risks, a backup data centre will be settled. At the same time, cooperative partners can share the data among their data centres.

The unique feature is that after separation from data of platform, mission data are processed and distributed in near real-time. For some of the cooperative users, they can only get archived data after the direct users or sponsors having attached to the mission data.

5. PLATFORM DESIGNS FOR SPACE DATA APPLICATION

5.1 principle of design

Although system for big data has common basic ideas of design, space data have its special features for consideration. Here we tried to design a system architecture suitable for space data application based on following principles which are needed for big data analytics:

- Emphasize the scalability on architecture
- Balance between “bring the analysis to data” and “bring the data to analysis”
- Balance between the mature and innovative technology
- Support integrating analytical methods and tools

5.2 Platform function design

5.2.1 Data storage and organizing: Big data have present challenged requirement of storage: more storage mediums with higher I/O speed which need innovations on traditional way (Phlip Chen & Zhang, 2014). Underlying requirement of storage is its scalability and reliability. After years of study, researchers and engineers have realized that advance margin on one point hardware for storage will have its limitation in the end, so comes the distributed storage architecture with software of management system which first serves the local system and then provide services for remote users.

Nowadays, three types of mainstream storage are block storage, file storage and object-based storage. In general, all disk arrays are block-based storage which include DAS (Direct Attach Storage) and SAN (Storage Area Network). As typical file-level storage, NAS is a set of network storage devices, usually being directly connected to the network and providing data access services, which is like a data file service system, characterized by cost-effective.

The object storage has the advantages of both SAN’s high-speed access to disk and NAS’s distributed sharing features. Its core idea is to separate the data path (data read or write) and the control path (metadata). Object storage simplifies the organization of the data (such as replacing "tree" and "file" with flattened "ID" and "objects"), and reduces the complexity of protocols and interfaces (such as simplifying complex locking mechanisms, to ensure the ultimate consistency), thereby enhancing the scalability of the system to cope with the massive data challenges in information explosion era. At the same time, the intelligent self-management function of the object can effectively reduce the system maintenance complexity and help the user to reduce the total cost of ownership. The advantages of object storage benefit much more on space data because of their keep-growing volume and velocity as well as their various formats of files.

5.2.2 Data processing and computing: Based on data storage and organizing, we need provide data processing and computing for upper application such as data curation, data analysis, or even data visualization. Computing capability has been evolved from local, individual devices to distributed, virtual, and scalable resources (Shawish & Salama, 2014). Cloud Computing developed on the way of distributed systems to meet the variation of application and requirements, under whose structure reliable services are shared among computing centres with the aid of virtualization technologies and Service Oriented Architectures (SOAs) (Shawish & Salama, 2014). Comparing to former computing paradigms, Cloud Computing has following distinguished advantages (Wang & Laszewski, 2008):
- Provide resources and services according to the users’ demand which means it is customizable.
- Offer guaranteed QoS for users as to computer performance.
- Hardware, software and data inside Clouds are automatically organized and operated for users.
- Scalable and flexible on various dimensions such as distribution, performance, and etc.

Although Cloud Computing is developed in the e-business context, it is also applicable for space industry. The features of on-demand service provisioning, scalability and flexibility greatly meet the demand of space data applications. Unlike in business application of Cloud Computing technology weighting much on economical factor, more concerns in space area are coming from scalability and flexibility demand. So, for the platform of space data, storage as service to guarantee the share of the data and software as service to promote novel algorithms, are commonly adopted service classes, while computing as service is always applicable in the inner part of an agency or in a research association. At the same time, for the space data platform, Private Cloud is firstly used and with the development of cooperation, it will evolve to Community Cloud which share services among the partners in a community. A good Cloud Computing platform should realize the theories of Cloud Computing and provide easy-to-use facilities for customized application. A successful case is Hadoop (open source) which has been widely used in batch computing platforms.

5.2.3 Data mining: For the space data platform, apart from general functions provided such as storage and processing, specific function which should be considered is the data mining tool set. For deep application of space data in areas of engineering decision and science research, it is basis for fully exploit of data value. Two classes of tools: general tools for data mining and professional tools for space data application, will be suggested in the platform for space data application.

Algorithms for general tools include basic techniques of inferring, statistics, classification, clustering and etc. Implementation of these algorithms is not difficult but accuracy and effectiveness can be improved if the algorithms have taken data features into account, such as missing values and numeric attributes. That is to say, even the basic data mining tools adapted to space data can make the outcomes more efficient. Apart from basic algorithms on data mining, there should be some expertise tools to meet the demand of space data application which particularly match the characteristics of space data. For example, much of valuable information is hidden in the images of space data, which can't be mined using basic data mining tools as described before. Another example is automatic fault diagnosis of operation control. Because of so many impact factors, diagnosis model can't be easily built in theory like other general applications. Complement way of analysis is to use archived data to assist diagnosis. By the way, Cloud Computing has enabled the on-demand development of these tools as extended function and even being distributed deployed.

Another point should be mentioned is that: for most of data mining algorithms in space data, they are semi-supervised or unsupervised because lack of supporting theory. What comes out from data mining a re always based on data themselves being analysed which is also the trend of data intensive science.

5.3 General architecture of platform

Base on the analysis of functions of platform, we can construct a general architecture which encompasses layered functions needed for the platform with high scalability and flexibility.

As an ordinary information system architecture, the underlying (first) layer is composed of hardware, system software, and network, as infrastructure of a platform. The second layer is data layer, which is composed of space data as well as other exterior source data for supporting analysis and application, such as GIS data, ground truth data. Although we choose object storage technology for space data as discussed before, we still want the platform to support other heterogenous data and with high compatibility. The third layer is service layer, which provide thorough services for upper layers including data service, computing service, system management service and other general services. The fourth layer is component layer, which include related functional components to support construction of application system. Data mining tools are part of the components which can be configured accordingly. The fifth layer is application layer, which is based on the former layers to implement a system for the user.

As an important feature of space data platform, it can be distributed without limitation of locality in principle, although in practical we should take network resources available into account.

6. PROMISING PRACTICE

Promising practices are used to reveal and invoke potential valuable application area. In general, advent of big data as well as sharing of storage and computing capability has enabled development of data exploration and its support of different space data scenarios which are mainly demonstrated in following aspects (White, 2013):

- Access to new sources of data for improved and more decision making.
- Develop and enhance analytic techniques and technologies to increase the analytic power of existing decision-making solutions.
- Improve analytic performance enabling more sophisticated and new solutions.

6.1 Content based image mining

For space data such as earth observation data, apart from this description information coming from exterior apparatus on spacecraft and payloads, more valuable information is behind the images themselves which always depends on image interpretation software and sometimes personnel. These unstated features are useful for analyst for further study because they may represent a pattern among massive data. With the sharp increase of data volume and velocity of space big data, we should consider the way to extract the information and knowledge from content automatically with less or without personnel in the end. Fortunately, advancement of computing and information technology as well as massive data samples from space has made it feasible.

Content based image mining is to use basic visual content other than description of the image such as colour, texture, pattern, image topology, shape of objects and their layouts and locations (Dubey, Bhargava, & Choubey, 2010). Basic process involved in content based image retrieval is that: first building an image database, setting up feature vectors from images for future use; then calculating vector of the target image and comparing the distance between feature vectors of the target image and image in the database; choose the image in the database if the distance is small enough (KUMAR, 2015). Using this process of image retrieval, researchers can wipe out massive garbage data and confine their work on valuable data. With cloud computing capability, the algorithms of image retrieval can be implemented in no time, and with developed data mining algorithms more valuable knowledge can be extracted. Apart from that, content base features retrieved and justified can be in turn integrated into image's description file for further application.

6.2 Data-intensive space science

As stated before, space science is less mature than earth observation where much of the facts and theory are needed to be explore. Because of limitation of ground truth, observation data and samples are mainly applicable resources for justifying any hypothesis in space science. For example, to classify each object as a particular type of star or galaxy, scientists have measured the attributes of each object on thousands of photographic plates (Read, n.d.). But fortunately, we have entered a new epoch with abundant information which accelerate the method of astronomy research with an unprecedented accuracy which can't be achieved with less data or less computing capacity (Brunner, Djorgovski, Prince, & Szalay, 2001).

An existing project-the Digital Palomar Observatory Sky Survey (DPOSS) is to classify and catalogue objects in the universe using images from Sky Survey mission while some film data should be digitalized firstly before using. Automatical processing has been enabled by the development of SKICAT (SKy Image Cataloging and Analysis Tool). In this tool, database has been built, advanced data mining models and novel artificial intelligence technology have been set up and used, basic classification tools have been implemented, with the outcome of a classification of over 50 million galaxies and a billion stars (Caltech, n.d.). Further research achievement has been got to adjust the survey using the CCD images from Palomar 60-inch telescope (Caltech, n.d.).

6.3 Mission control

Newly achievement in data mining techniques has stimulated the inspiration of study on archived spacecraft telemetry data to extract valuable information especially useful for anomaly detection of the system in real time or afterwards (Iverson, 2008). These data-intensive methods of fault diagnose have more accuracy than human endeavour and make researchers free from tedious work of searching for unusual data.

To explore the strength of massive data-intensive techniques, concurrent analysis of multiple elements in the telemetry data can be applied to extract the relationship between the data sets (Iverson, 2008). A multi-parameters vector is used to express a status of operation in a N-dimensional space, and vector distance can be used as to judge if the tested status is at the boundary of acceptance compared to the reference status or it is abnormal. In this use case, data mining is to some extent preliminary although functional, which decreased human work load with archived data and benefitted the analysts for focusing on analysis work of abnormal data. Another data mining research was also proposed by Iverson (2008), that is to predict abnormal trend based on system model built with normal operation data sets. Those kind of models and tools have been used in real world such as analysing of Space Shuttle monitoring data and real-time health monitoring for ISS gyroscopes (Iverson, 2008).

7. CONCLUSIONS

Inspired by the big data and related technologies, we tried to dive into space data and give systematical research on its characteristics and application to make the full use of space data. We also proposed the platform suitable for the space data application as well as some promising practices invoking potential valuable application of big data from space.

REFERENCES

Arviset, C., 2016. Big data, Big Challenges and new Paradigm for the Gaia Archive http://esaconferencebureau.com/custom/16M05/bids/ALL/01_1230_Arviset.pdf (3 Feb. 2017).

Brunner, R., Djorgovski, S., Prince, T., & Szalay, A., 2001. MASSIVE DATASETS IN ASTRONOMY https://www.cs.princeton.edu/courses/archive/spr04/cos598B/bib/BrunnerDPS.pdf (19 Feb. 2017)

Caltech. n.d., The Palomar Digital Sky Survey (DPOSS) http://www.astro.caltech.edu/~george/dposs/dposs_pop.html(18 Mar. 2017).

Dubey, R., Bhargava, N., & Choubey, R., 2010. Image Mining using Content Based Image Retrieval System. International Journal on Computer Science and Engineering, pp. 2353-2356.

Ishwarappa, & Anureadha. J, 2015. A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. www.sciencedirect.com (14 Feb. 2017).

ISU SSP16., 2016. Space Big Data https://isulibrary.isunet.edu(2 Feb. 2017).

Iverson, D. 2008. Data Mining Applications for Space Mission Operations System Health Monitoring http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.154.7486(17 Feb. 2017).

Kaethler, S., 2016. Big data architecture, design and evolution using ARES and DrMUST https://isulibrary.isunet.edu(26 Jan. 2017).

KUMAR, H., 2015. An overview on content-based image retrieval http://www.computerscijournal.org/?p=1703(16 Mar. 2017).

NASA, 2012. What Is Microgravity? https://www.nasa.gov/audience/forstudents/5-8/features/nasa-knows/what-is-microgravity-58.html(8 Mar. 2017).

Phlip Chen, C., & Zhang, C.-Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data http://dx.doi.org/10.1016/j.ins.2014.01.015(19 Feb. 2017).

Read, B., n.d. Data Mining and Science? https://www.ercim.eu/publication/ws-proceedings/12th-EDRG/EDRG12_Re.pdf(10 Mar. 2017).

Shawish, A., & Salama, M., 2014. Cloud Computing: Paradigms and Technologies http://www.springer.com/cda/content/document/cda_downloaddocument/9783642350153-c2.pdf?SGWID=0-0-45-1429336-p175276227(20 Mar. 2017).

Soille, P., & Marchetti, P., 2016. Proceedings of the 2016 conference on big data from space (BiDS' 16) https://isulibrary.isunet.edu(28 Mar. 2017).

Wang, L., & Laszewski, G., 2008. Scientific cloud computing: Early definition and experience. Proceedings of 10th IEEE International Conference on High Performance Computing and Communications, pp. 825–830.

White, C., 2013. Data Exploration and Discovery: A New Approach to Analytics http://fr.teradata.com/Resources/Analyst-Reports/Data-Exploration-and-Discovery-A-New-Approach-to-Business-Analytics/?LangType=1036&LangSelect=true(12 Mar. 2017).