9º Workshop em Engenharia e Tecnologia Espaciais
15 e 16 de Agosto de 2018

A study for the application of OLAP in satellite telemetry data

PEREIRA, Y. M. D. [1 2] **, FERREIRA, M. G. V.** [1] **, SILVA R. R.** [3 4]

[1] National Institute for Space Research, São José dos Campos, SP, Brazil.

[2]Master student of Engineering and Management of Space Systems (CSE).

[3]FATEC Mogi das Cruzes, São Paulo State Technological College, Brazil.

[4]CISUC – Centre for Informatics and Systems of the University of Coimbra, Coimbra, Portugal.

yuri.pereira@inpe.br

Abstract. *The Brazilian National Institute for Space Research (INPE) currently operates three main satellites, each of whom generates their own telemetry data for the engineering department to check the health of the satellite and of its subsystems. This data needs to be analyzed to evaluate whether the satellite will be able to keep nominal operations in the future. Telemetry data is characterized by its high dimensionality, making analysis difficulty and hard to generalize for other satellites. This work aims to present a study on On-Line Analytical Processing (OLAP) architecture alternative: the Data Cube. An overview of the Data Cube structure is presented, together with solutions that it might bring to satellite operations, and concludes by reviewing some alternatives in the literature.*

Keywords: Data Cube; OLAP; Telemetry; Data Warehouse.

1. Introduction

The Satellite Control Center (SCC) located at the Brazilian National Institute for Space Research (INPE) currently monitors and controls three main satellites: the Data Collection Satellite (SCD) family, comprised of SCD1 and SCD2, and the China-Brazil Earth Resources Satellite (CBERS), with the fourth satellite in the family, CBERS-4. These satellites have daily passages with visibility over INPE's ground stations, which are responsible for receiving telemetry and payload data, besides sending telecommands and performing housekeeping activities, like speed measures [Orlando and Kuga 2007].

For each of these passages, telemetry data is acquired and stored at INPE's data centers. These data are comprised of satellite sensors measurements and healthchecks, with information on the satellite battery temperature, current and other subsystem information, to name a few [Azevedo et al. 2011]. Since that these data need to be stored for the entire lifetime of the satellite, they can amount to a big volume over time. For example, the SCD1 alone has been operational for over 20 years [INPE 2013], and there has been 4 satellites operating between 1998 and 2018 from the CBERS family, with one more satellite to come, the CBERS-4A, and SCC is preparing for the first satellite of the Amazonia family, Amazonia-1.

For each of the SCD satellites, we estimate around 20 million telemetry frames for 20 years of operation, and that's because they have about 135 telemetries per frame, with the CBERS family

having over 2 thousand telemetries for each satellite. Due to the number of telemetries, these data are highly dimensional, making an analysis on the relationships between the dimensions difficult.

These data are used by the satellite operators to check the operational capacity of the satellites, see the health of the subsystems and if they're working properly, and that the satellite will continue to perform its duties properly in the near future. In the case of an emergency they might need to check old telemetry data to see if a situation has occured before, and check whether that might prove a danger to the satellite or not [Azevedo et al. 2011].

The historical analysis is very important for satellite operations, as it might unearth rare phenomena and can serve as an early warning that some issue might appear in the future. One example is in the case of CBERS-2, which had the phenomenon of thermal breakdown happening to one of its batteries [Magalhães 2012]. Having the historical telemetries was fundamental in the analysis of the phenomenon, and the lessons learned with it for operations are invaluable.

In this work, we present a data structure called data cubes to make the analysis on satellite telemetry data easier to be performed. Since that a Data Warehouse hasn't been implemented yet for the telemetry data, doing analysis is quite a slow process that involves a lot of manual steps and the creation of complex, not easy to generalize, queries and custom code. The aim of this structure is to make the analysis an action that is easy to perform for the operation of current and future satellites, with good average response times, thus aiming to improve INPE's satellite operations capabilities.

This paper is organized as follows: section 2 presents a background on OLAP, data cubes and the telemetry data; section 3 presents the analysis possible with the data cube and some alternatives and section 4 concludes the paper.

2. Background

The main characteristic of satellite telemetry data is its high dimensionality in the form of having many telemetries for a given time frame, thus making it a time series.

Telemetry data is measured across all the equipments of the satellite, which can measure widly different values like battery voltage, magnetometer influences and payload healthchecks among others, and is therefore highly complex in nature. Some satellites can have thousands of telemetries, with some being received per second continuously during its lifetime, like the Hubble telescope that generated terabytes of telemetry data per year [Miebach 1998].

2.1. Telemetry data

Telemetry data is obtained from the satellites via telecommunication with the ground stations. For SCC, the data is delivered as raw binary data, which is then converted to a human readable format by SATellite Control System (SATCS), a software system developed by the Ground Systems Development Division (DSS) of coordination of Space Engineering and Technology at INPE [Julio Filho et al. 2017].

SATCS also provides means to send telecommand and some basic data visualization, but these only use recent telemetry data and not the full historical database. The system is very useful for daily satellite operations, but it lacks the big data analysis that is necessary for a more comprehensive data mining effort, which lead to the proposal of a data mining architecture for telemetry data at INPE [Azevedo and Ambrósio 2010].

2.2. Data Warehouse and OLAP

A Data Warehouse (DW) is a way to generalize and consolidate data in a multidimensional space, including anything necessary to make the analysis possible on the data, from data cleaning, data integration and transformation to other activities, and its main objective is to provide On-Line Analytical Processing (OLAP) for the analysis of multidimensional data [Han et al. 2011].

The Data Warehouse OLAP technology differs from the common database necessities and requirements, which perform On-Line Transaction Processing (OLTP), whose focus is in the automation of day-to-day tasks, such as banking transactions, user data and similar. These tasks are characterized by their repetitiveness and the following of a set structure, and they consist of short, atomic and isolated transactions [Chaudhuri and Dayal 1997].

In contrast, the focus of the data warehouse is on decision support. This means that it needs to work on historical, summarized data, and the individual transactions aren't as relevant, so the data warehouse uses data from various operational databases and across long periods of time. Due to this, the DW needs to work on orders of magnitude more data than the individual operational databases, and so it needs to prioritize query throughput and response times for highly complex queries, which can be made on millions of records, sometimes performing scans, joins and aggregations [Chaudhuri and Dayal 1997].

OLAP is also not a single concept, as there are different ways of implementing it, like Relational OLAP (ROLAP), Multidimensional OLAP (MOLAP), Spatial OLAP (SOLAP), and Hybrid OLAP (HOLAP), among others [Viswanathan and Schneider 2014].

A Data Warehouse architecture has been proposed for satellite telemetry already on [Azevedo and Ambrósio 2010], however the aim of the article was broader than trying to achieve OLAP, with a focus on the analysis of the telemetry, with OLAP being a necessary step to allow for that to happen.

2.3. Data Cubes

Despite being over 20 years old [Gray et al. 1996], the Data Cube is a data warehouse technology that is extensively used for the storage and analysis of high dimensional data, for example for the geospatial domain in the analysis of satellite images, which are high dimensional by nature [Viswanathan and Schneider 2014].

The Data Cube structure allows for data to be modelled and viewed in multiple dimensions, and it is characterized by dimensions and measures. A dimension is made of the entities in which we keep our records, like the columns of a table and relevant attributes of a model. A measure, also called a fact, is the quantity by which the relationship between dimensions will be analyzed, and they are generally numerical values. And a *fact table* is the representation of those measures with the dimensions, working as the representation of the relationships between them.

Despite the name, the data cube is not necessarily a 3-D geometric structure, but rather an n-dimensional structure. A 2-D data cube, for example, is a table or a spreadsheet in respect to two dimensions. But we can add another dimension to that visualization, and see it as a 3 dimensional cube, as in figure [1]. This process gets harder to visualize with more dimensions, but the general idea is the same: each additional dimension adds more depth to the data, and thus makes it more complex.

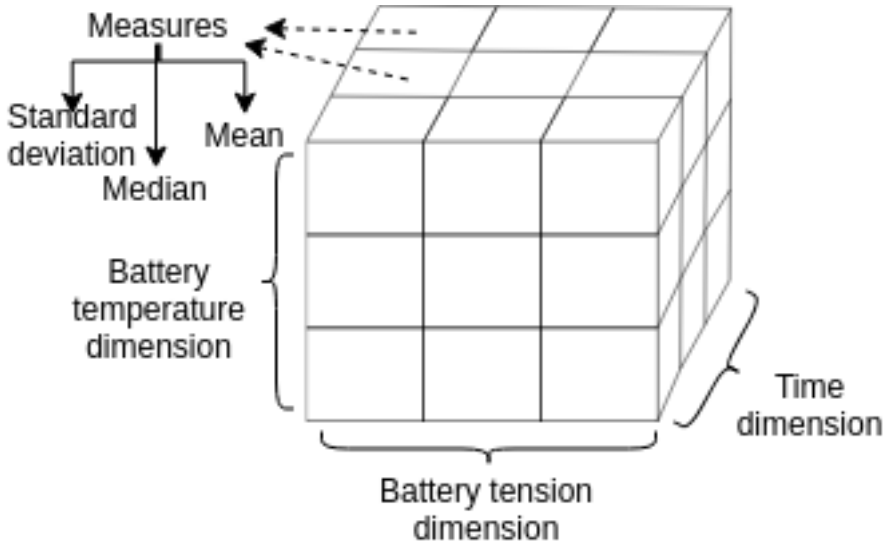

Figure 1. A possible telemetry cube with three dimensions and three measures.

2.3.1. Measures

According to [Han et al. 2011], measures can be classified into three different types: *distributive*, *algebraic* and *holistic*. A **distributive** measure is a measure whose calculation can be partitioned and later joined, and the result would be the same as if the executing the calculation with all the available data at once. In figure [1], the *mean* is a distributive measure. An **algebraic** measure is a measure whose calculation can be made from two or more distributive measures, like an *average* operation, which can be performed with an *sum* and an *count* measures. A measure is **holistic** if it can only be computed without a distributive measure, meaning that it can't be divided into smaller operations and must be computed for all the data, like the measure *standard deviation* in figure [1].

Holistic measures are the hardest to calculate, and so it's an area with increased research attention [Silva 2015].

2.3.2. Concept Hierarchies

A concept hierarchy is used to define a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. It is a form of grouping and discretization, since that it groups values in a way to reduce the cardinality of a dimension [Han et al. 2011]. They are helpful in making the analysis easier to understand, as they do operations like translating the time dimensions in figure [1] to a hierarchy of *day-month-year*, making the queries on the desired level of abstraction easier to be performed and understood.

2.3.3. Operations

Since that the focus is on analysis, an OLAP architecture generally needs to support some common operators: *rollup*, which increases the level of aggregation or climbs a concept hierarchy; *drill-down*, which decreases the level of aggregation, going from more general data to more specific data, sometimes along one or more dimension hierarchies; *slice and dice*, which create selections and projections on the data; and *pivot*, which re-orients the visualization of the data.

There are other operators, like *drill-across* and *drill-through*, but the support and existence of these will depend on the type of OLAP that is necessary [Han et al. 2011].

2.3.4. Cube types

To properly calculate a data cube for some measures and dimensions, you have to count the cardinality of each dimension against the cardinality of all other dimensions. While manageable for a few dimensions, this computation becomes almost impossible for cubes with high dimensions as the number of combinations becomes too much for a single computer to handle. This lead to the development of data cube algorithms that optimize for the most relevant measures in the data cube [Silva 2015].

A *cuboid* is a part of a data cube. For example, if you have three dimensions: temperature, tension and time, a 2-D cuboid could be made from the dimensions temperature and tension, and a 3-D cuboid would be the same as the full data cube, like figure [1]. The data cube algorithms focus on the computation of cuboids, as the every cube is composed of these smaller cuboids. This leads to the existence of the *curse of dimensionality*, as for n of dimensions there will be 2^n possible cuboid computations, making full materialization very difficult after a few dimensions [Han et al. 2011; Silva 2015].

For the algorithms, they can be in three different categories: Computing all the cuboids for the data cube leads to a fully materialized data cube; not computing any cuboid beforehand leads to a non-materialized data cube, and partially computing some cuboids leads to partial materialization.

The non-materialized cube has the lowest amount of required memory, but the highest query response time. The fully materialized cube leads to the lowest query response times, as all combinations are already computed, but it needs the highest amount of memory and is thus very hard to compute. As for the partially materialized cube, it is the main issue for most of the algorithms: how to materialize only the most relevant cuboids, and thus achieve a good compromise between memory usage and query performance?

There's some different types of partially materialized cubes, like the *iceberg*, which is a cube with only cells that have passed a certain condition; *shell fragments* compute only cubes with a few dimensions (from 3 to 5) and aggregate those cubes when a bigger number of dimensions is required and *closed cubes* are cubes whose cells with identical measures are grouped into a single abstraction, also called *closed cells*.

To choose which cuboids to materialize, there's a plethora of different algorithms. Two of the classical ones are *Bottom-up* or *Top-down* strategies. Bottom-up starts from the most specific cuboid, called the base cuboid, and goes to the less specific cuboid. Top-down is the inverse: it starts from the least specific cuboid, called the apex cuboid, and goes to the base cuboid. Most of these are tested and overviewed in [Silva 2015], and won't be repeated here for brevity.

3. Discussion

The focus of this data structure is to make the analysis of high dimensional data not only possible, but also with good performance and capable of handling multiple users working on the data. In SCC's case, it is important to be able to analyze the data quickly during a passage, so that the operators might be able to make any decision concerning the satellite, and whether a certain data trend might be important or not.

From Azevedo and Ambrósio [2010] we know that having that analysis is not only important, but that it might unearth previous unseen information, and from Magalhães [2012] we have an example where having the data in an analysis-ready manner was important, as the occurence of the phenomenon displayed there might be detectable with enough early warning as to allow the satellite operators to make decisions that can extend the usable life of the satellite.

And this kind of structure and analysis is already implemented in other space agencies and satellite tracking stations: from Miebach [1998] we know that the Hubble telescope alone generated terabytes of telemetry data, which is already a problem of the Big Data area and is not easy to analyze. Though they do not give implementation details and whether they're using some kind of Data Cube structure or other type of database, they do have some form to analyze the data quickly, so the idea might have already been implemented.

One interesting approach for the analysis would be to use discovery-driven exploration of Data Cubes, as in Sarawagi et al. [1998], as this might unearth previous unseen or unexpected relationships between telemetries in a visual, user-centric way. This would be possible to implement on top of an already-existing data cube structure for the telemetry data.

Some companies and space agencies already use similar structures, but they're not based on a Data Cube approach, and rather use Cloud Computing, like Boussouf et al. [2018] from AirBus, that uses Apache Hadoop, Spark and HBase for their infrastructure. The important point to note however is that they do have a proper process driven analysis process, and that they have a working structure for the analysis.

However, the data cube is still the only structure that allows for a complete OLAP implementation with all possible dimensional queries, and so is an area worthy of research [Cuzzocrea et al. 2013], which also reveals the interesting research on MapReduce operations for traditional DataBase Management Systems (DBMS) [Abouzeid et al. 2009].

While the advantages and disadvantages of the data cube are explained on this work, the usage of the Data Cube is unclear in some cases for the satellite operation procedures, and it might be a good alternative for INPE's operators it makes a move to a Big Data structure that is ready to perform the analysis of engineering data. However, selecting the best infrastructure for INPE is outside the scope of this article, and will require further study of SCC's specific use cases and requirements.

3.1. Example

Taking the data cube referenced in figure [1], and from inspiration the work in [Magalhães 2012], we can show some possible operator queries in response do a certain telemetry, for example if we wanted to know the mean battery currents on the month of may of 2018, figure [2] shows an answer using the data cube, performing a slice operation on the time dimension, showing transformation from a full data cube to the 2-dimensional table that is of interest. The values were chosen to showcase the cube, but do not have any relationship with reality.

This example is relatively easy to do with access to the data, as it is only a trivial example. But it shows how the complexity of the operations can grow, as to do this for a successive number of dimensions gets exponentially hard, as stated in section 2. Considering that each satellite will have at least a hundred telemetries [Orlando and Kuga 2007], that's already a 100-dimensional data cube to solve for common operations.

In Silva [2015]'s work, a good number of the used algorithms in the literature could not even compute cubes with more than 15 dimensions, and for telemetries we'd need to compute at least $2^{100} = 1.2676506 \times 10^{30}$ different **cuboids**, not considering the number of records, measure

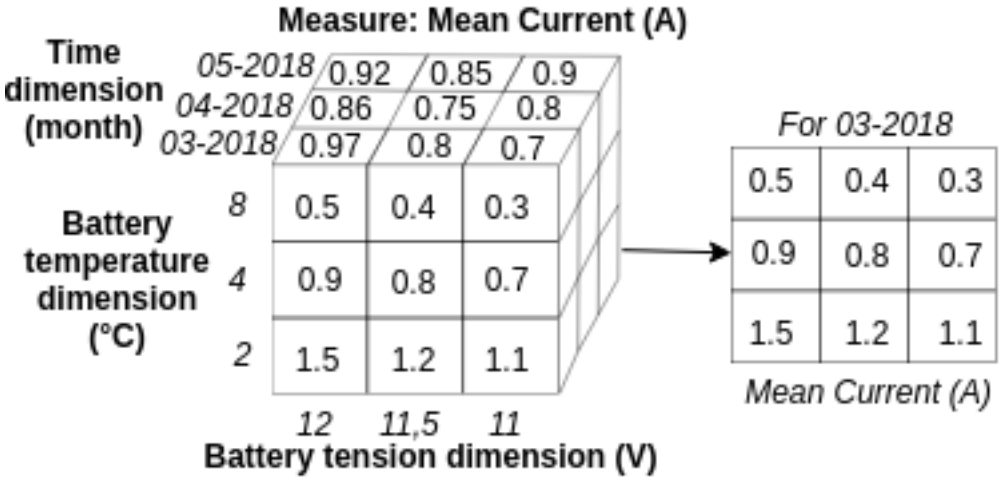

Figure 2. An example DataCube operation for telemetries.

calculations or concept hierarchies for any of the dimensions. It is of note that for the cube in figure [2], there's $2^3 = 8$ possible cuboids, ignoring concept hierarchies.

4. Conclusions and future work

This work presented an overview on a platform based on Data Cubes to implement an OLAP architecture for INPE's satellite control activities.

The use of a data cube structure can lead to an easier data mining effort as it provides a standardized data interface to execute data mining upon. This can be used to acquire engineering knowledge from vast amounts of diverse satellite telemetry data, and is a strategy already used in other areas with similar high dimensional data.

The data structure presented here will be implemented to validate its capability to help with INPE's satellite operation efforts, and to validate the adequability of the data structure to the satellite telemetry data.

Acknowledgements: *The authors would like to thank CAPES for their financial support.*

References

Abouzeid, A., Bajda-Pawlikowski, K., Abadi, D., Silberschatz, A. and Rasin, A. (aug 2009). HadoopDB: An architectural hybrid of MapReduce and DBMS technologies for analytical workloads. *Proceedings of the VLDB Endowment*, v. 2, n. 1, p. 922–933.

Azevedo, D. N. R. and Ambrósio, A. M. (2010). Dependability in Satellite Systems: An Architecture for Satellite Telemetry Analysis. In *Anais...* Workshop em Engenharia e Tecnologia Espaciais, 1. (WETE). INPE.

Azevedo, D. N. R., Ambrósio, A. M. and Vieira, M. (2011). *Estudo sobre técnicas de detecção automática de anomalias em satélites*. Tradução. São José dos Campos: Instituto Nacional de Pesquisas Espaciais.

Boussouf, L., Bergelin, B. and Scudeler, D. et al. (2018). Big Data Based Operations for Space Systems. In:. *2018 SpaceOps Conference*. Tradução. American Institute of Aeronautics and Astronautics..

Chaudhuri, S. and Dayal, U. (mar 1997). An overview of data warehousing and OLAP technol-

ogy. *ACM SIGMOD Record*, v. 26, n. 1, p. 65–74.

Cuzzocrea, A., Bellatreche, L. and Song, I.-Y. (2013). Data Warehousing and OLAP over Big Data: Current Challenges and Future Research Directions. In *Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP*., DOLAP '13. ACM.

Gray, J., Bosworth, A., Lyaman, A. and Pirahesh, H. (1996). Data cube: A relational aggregation operator generalizing GROUP-BY, CROSS-TAB, and SUB-TOTALS.. IEEE Comput. Soc. Press.

Han, J., Kamber, M. and Pei, J. (2011). *Data Mining: Concepts and Techniques, Third Edition*. Tradução. 3 edition ed. Haryana, India; Burlington, MA: Morgan Kaufmann.

INPE (2013). SCD-1 completa 20 anos. Primeiro satélite brasileiro comprova o êxito da engenharia espacial no país.. http://www.inpe.br/noticias/noticia.php?Cod_Noticia=3198.

Julio Filho, A. C., Ambrósio, A. M., Ferreira, M. G. V. and Loureiro, G. (2017). The Amazonia-1 satellite's ground segment - challenges for implementation of the space link extension protocol services. In *Proceedings...* International Astronomical Congress, 68. (IAC).

Magalhães, R. O. de (feb 2012). Estudo de avalanche térmica em um sistema de carga e descarga de bateria em satélites artificiais. Instituto Nacional de Pesquisas Espaciais.

Miebach, M. P. (may 1998). Hubble Space Telescope: Cost reduction by re-engineering telemetry processing and archiving. In *Telescope Control Systems III*.. International Society for Optics and Photonics.

Orlando, V. and Kuga, H. K. (2007). Rastreio e controle de satélites do INPE. In: Othon Cabo Winter; Prado, A. F. B. de A.[Eds.].. *A Conquista do Espaço: Do Sputnik à Missão Centenário*. Tradução. cap. 6 ed. São Paulo: Editora Livraria da Física. p..

Sarawagi, S., Agrawal, R. and Megiddo, N. (1998). Discovery-driven exploration of OLAP data cubes. In: Goos, G.; Hartmanis, J.; Van Leeuwen, J., et al.[Eds.].. *Advances in Database Technology '98*. Tradução. Berlin, Heidelberg: Springer Berlin Heidelberg. v. 1377p. 168–182.

Silva, R. R. (2015). Abordagens para Cubo de Dados Massivos com Alta Dimensionalidade Baseadas em Memória Principal e Memória Externa: HIC e BCubing. Instituto Tecnológico de Aeronáutica.

Viswanathan, G. and Schneider, M. (2014). User-centric spatial data warehousing: A survey of requirements and approaches. *International Journal of Data Mining, Modelling and Management*, v. 6, n. 4, p. 369.