Data Management at Gaia Data Processing Centers

Pilar de Teodoro, Alexander Hutton, Benoit Frezouls, Alain Montmory, Jordi Portell, Rosario Messineo, Marco Riello, Krzysztof Nienartowicz

Abstract Gaia is an ESA mission which will deal with large volumes of data that have to be processed at, and transferred between, different data processing centers (DPCs) in Europe. Managing the data and the associated databases will be a significant challenge. This paper presents the different data management configurations that have been evaluated at the Gaia DPCs in order to cope with the requirements of Gaia's complex data handling.

1 Introduction

Conceptually the data that will be generated by the Gaia processing centers can be imagined to be a database with a size of order of 1 Petabyte. This figure represents

Pilar De Teodoro
Serco, ESAC-ESA, Villafranca del Castillo, Madrid, DPCE e-mail: pilar.teodoro@esa.int

Alexander Hutton
Aurora, ESAC-ESA, Villafranca del Castillo, Madrid, DPCE e-mail: alexander.hutton@esa.int

Benoit Frezouls
CNES, DPCC e-mail: Benoit.Frezouls@cnes.fr

Alain Montmory
Thales/CNES, DPCC e-mail: alain.montmory@thalesgroup.com

Jordi Portell
University of Barcelona (DAM-ICCUB-IEEC), DPCB e-mail: jportell@am.ub.es

Rosario Mesineo
ALTEC S.p.a, DPCT, e-mail: rosario.messineo@altecspace.it

Marco Riello
Institute of Astronomy, Cambridge , DPCI e-mail: mriello@ast.cam.ac.uk

Krzysztof Nienartowicz
ISDC, Observatory of Geneva, DPCG e-mail: Krzysztof.Nienartowicz@unige.ch

only the raw data and the reduced data stored in the central database (known as the MDB, or Main Database). If all the Data Processing Centers for the Gaia Mission are considered then the total data size is even larger. The Gaia data handling architecture is already discussed in [1], [2]. This paper builds on the hardware layers mentioned in [1] by describing the database access layers.

The Gaia Data Processing and Analysis Consortium (DPAC) agreed that the database software, or even whether to use an SQL or NoSQL database, would be a free choice for the DPC's. The Gaia Java framework was designed to be as database agnostic as possible. The aim of this paper is therefore to show the different configurations used in the DPC's and how data is handled for processing on each center.

2 DPCE

DPCE is the Data Processing Center at the European Space Astronomy Center(ESAC) in Madrid belonging to the European Space Agency (ESA).

As figure 1 shows the dataflow regarding databases at ESAC includes three databases and a requirement to transfer the data to the other DPCs. There are some processes that take place once the data is downlinked from the satellite. The average data received from the satellite per day is around 30GB. The first database where the data is stored and used for processing is the Initial Data Treatment/ First Look database. The present test environment is an Oracle RAC database which provides high availability. The intention is to replicate this data to another database so that during production tests can be performed independently.

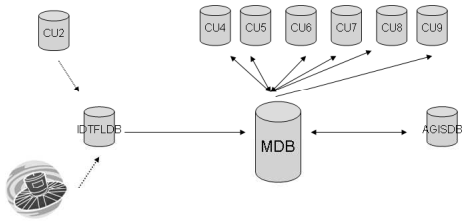


Fig. 1 Gaia Data Flow to and from DPCE, MDBIngestor/Extractor is used to ingest/extract the data from the DPCE databases. Data is transferred between DPCs using Aspera.

The data then is transferred to the MDB database. In the tests done so far this has been an Oracle database. From here the data is transferred to another database to process the data outside of the MDB. The Astrometric Global Iterative Solution, which calculates the precise positions of the sources, is applied and the results are ingested into the MDB again. From the MDB the data is transferred to the DPCs and after the data is processed in each center the data is ingested again in the MDB. It is planned to make a release of the MDB every cycle of 6 months.

However, Oracle is not the only database being used. In DPCE the use of Intersystems Cache Database is growing and for testing Derby and MySQL databases are also being used. The different software versions are shown in Table 1.

Table 1 Database software used at DPCE

Software	version
Oracle	11.1.0.7.4
Intersystems Cache	2010.2
MySQL	5.5
Derby	10.7.1.1

Backup studies are in progress and depending on the database in question different solutions are being analyzed such as Oracle Data Guard or Cache Mirroring. The option of using partitioning is also being studied to improve performance.

3 DPCT

The DPCT system has to be conceived as the set of products, people and processes necessary to achieve CU3 Astrometric Verification Unit (AVU) and GAREQ data processing within Data Processing and Analysis Consortium (DPAC). In addition to these three data processing functions, DPCT will support the hosting and operations of the Initial GAIA Source List (IGSL) database.

Persistent data management is one of the critical point of the overall DPCT tasks. Data to be managed at DPCT are large Terabytes and must be stored for long time. In addition data access must be efficient to avoid that processing is bounded by data access.

The main goal is to design database architecture scalable, with a high availability and performance. In order to address this main objective the database architecture has been designed using the following techniques:

- Storage Virtualization
- Clustering
- Load Balancing
- Redundancy

In the below figure 2 the DPCT database architecture structure is described up on several levels. The level 0 database is the level where data related to Main Database will be stored with a only read access whereas the level 1 database will contain all data strictly tied to processing and infrastructure management. Each hosted scientific module has its schema supporting module pipeline operations and offline data access and analysis.

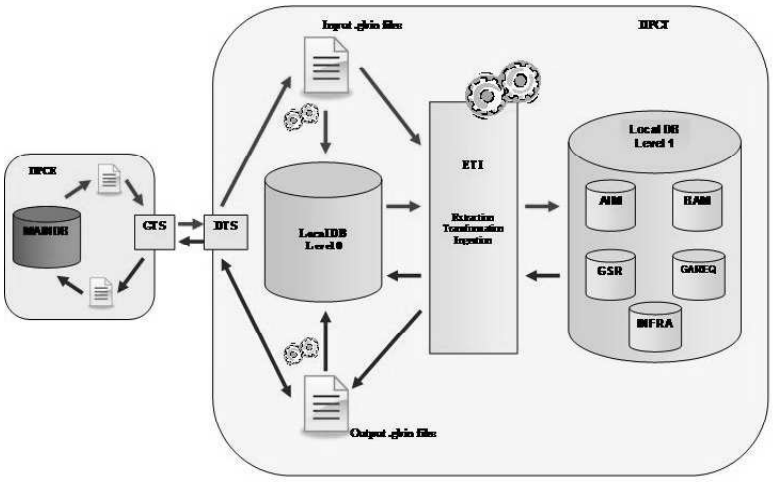


Fig. 2 DPCT architecture diagram. DBMS choice is Oracle that provides advanced availability and capability features.

The Oracle database architecture integrates the following Oracle products (version 11.1.0.7): Oracle Enterprise Server, Oracle RAC on three nodes (Active-Active-Spare), Oracle Partitioning and Oracle ASM to manage storage, provided by a storage arrays. In addition to Oracle products a Mysql instance is dedicated to receive IGSL database and export IGSL tables into Oracle database.

4 DPCC

CNES (French space agency) is hosting the processing center for the CU4 (Objects Processing), CU6 (Spectroscopic processing) and CU8 (Astrophysical Parameters). The foreseen volumetry, at the end of the mission, is of the Petabyte order stored in tables containing around ten billions rows. The current solution is PostgreSQL 8 based but will unfortunately not meet these requirements. Oracle being far too expensive, several alternative solutions are considered. A study aimed at choosing the final operations database system was launched in September 2010 in DPCC. This study is performed by Thales (subcontractor developing the DPCC host framework) and will end in June 2011.

Three distinct phases are identified in the data management in DPCC :

- data ingestion : insert and prepare data files produced by other DPCs (gbin format) to be efficiently accessed by CU4, 6, 8 processings : one data ingestion per CU on both daily and 6 months cycle basis,
- data processing : process the data previously prepared and produce DPCC results,

- data extraction : extract into gbin files the data produced and stored in the database system.

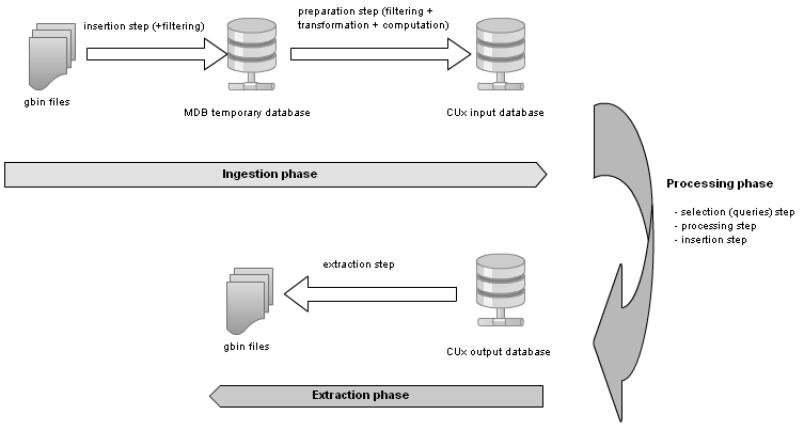


Fig. 3 DPCC Data Flow.

The scenario of the study implements the two dimensioning phases of data ingestion by evaluating the insertion data rate and complex queries between tables and data processing by highlighting the query performance and interface with the processing. The objective is to reach the 10% of the final solution needs in term of volume and number of row stored in the database system (100Tb of data) in order to perform a sensible extrapolation of the obtained results.

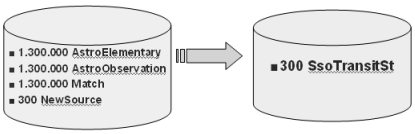
The different solutions are evaluated based on performance, scalability of the solution, data safety, impacts on the existing software, impacts on the hardware architecture, cost of the solution during the whole mission, durability of the solution and administration and monitoring tools.

Five database management systems are tested :

- Apache Cassandra
- Apache Hadoop
- InterSystems Cache
- IBM DB2
- Mixed solution : PostgreSQL 9 + Pl/Proxy 2 + Cassandra

5 DPCG

The processing at DPCG is dedicated to the detection and characterization of variable sources observed by Gaia. DPCG is in charge of the CU7 and their challenges



Existing solution (PostgreSQL 8)	Cassandra	Hadoop	Caché	DB2	PI/proxy+ PostgreSQL9
14min50s	3min37s	3min37s	2min25s	13min55s	6min50s

Fig. 4 Preliminary DPCC Results. Large scale and scalability tests are still to be performed but preliminary results show Hadoop's efficiency.

are relatively complex, evolving Object Model. Analytical queries must be done over sources or processing results (attributes) to support unknown research requirements. Parameter analysis for simulations and configurations changes on historical database are done. ETL-like support must be done for external data. At present Apache OpenJPA is used. Postgress has been used as well. Other alternatives to use are: Hadoop, SciDB, VoltDB and Extensions to PG.

6 DPCI

The Data Processing Center at Cambridge, UK, is responsible for the operation of the main photometric pipeline acting as the operational component of CU5. DPCI tested Hadoop since 2009 Q2 and was fully adopted in 2009 Q4. All data is stored on HDFS, a distributed filesystem, to maximize the network bandwidth and minimize the risk of bottlenecks. The processing tasks are Map/Reduce jobs to minimize the amount of synchronization and the simple abstraction for distributed execution. The tricky bit(i.e. the synchronization) is handled by Hadoop for you. In this way multi-threading issues are eliminated and is resilient against node loss. Running jobs can continue even in case of node loss. The paradigm shift for data model is that Data Types (DTs) are immutable and also the freedom of sharing and composing data types within a process. Hadoop eliminates a large source of potential bugs and encourages clean algorithm definition making the database access layer a lot simpler. DTs are specified via a simple definition language and its specifications are compiled directly to Java bytecode. Configuration based on Java properties has been banned because it easily lend itself to abuse. Instead a configuration approach based on strongly typed configuration items that are validated against a specification at job submission is used. The configuration is also immutable, versioned and pushed to the clients (it's just another form of data). Algorithms are decomposed into a number of processing elements that behave as deterministic functions for predictable behavior (reproducibility) doing one thing and one thing only. The advantage is also that testing is easier and that a scientific recipe is compiled from existing processing elements via a definition language. Also Hadoop has a constant overhead for job submission and (map) task startup: these overheads are negligible when the overall

task execution time is reasonably long (at least a few minutes) which is normally what happens for DPCI processing jobs.

7 DPCB

The Data Processing Center of Barcelona is actually composed of two different institutions, namely, the Supercomputing Center of Catalonia (CESCA) and the Barcelona Supercomputing Center (BSC). Only the latter, which holds the MareNostrum supercomputer, will be used during the Gaia operations, while CESCA is used mainly for development and testing of critical software that will run at ESAC. The combination of these two centers provides all the necessary infrastructure and tools for the successful implementation and test of some of the most important systems for the Gaia data processing.

CESCA features an environment equivalent to the one available at DPCE (ESAC, Madrid), where the near-realtime systems will be run. That is, a very similar computing cluster, a high-performance central filesystem, and an Oracle Database 11g with RAC (three instances) and ASM storage, running in a separate cluster. CESCA is used for the development and testing of the so-called Initial Data Treatment (IDT), as well as for some tests and developments related to the future exploitation of the Gaia catalogue. About 3TB for the filesystem plus 3TB for the Oracle ASM are available for Gaia. Since IDT is the first stage of the Gaia data processing, other centers of DPAC need its output data in order to test their own systems and developments. For this reason, CESCA features a public server (password-restricted and with read-only access) that allows the various DPAC members to access the contents of its central disk and database.

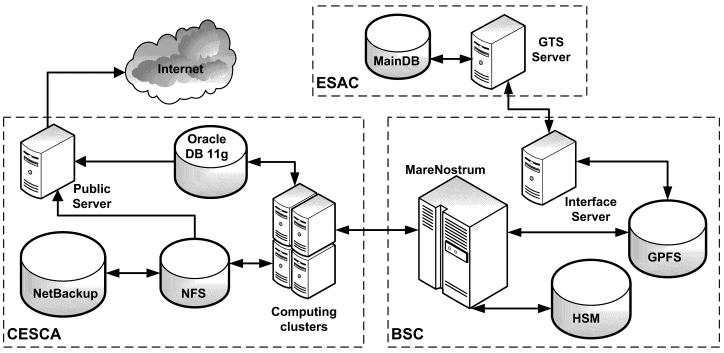


Fig. 5 Overview of the main elements of DPCB, including the computing elements, the storage systems and the communication flow.

On the other hand, BSC features the MareNostrum supercomputer, which offers a huge amount of computing resources and storage capacity. More specifically, it has about 2500 computing nodes (4 cores each) and 120TB in a high-performance highly-scalable central disk (the IBM GPFS solution). This is a perfect environment for the Gaia simulator, which requires a huge amount of computing and storage resources — so this is the main utilisation of BSC for Gaia until launch. Beyond that, MareNostrum will be used for the so-called Intermediate Data Updating (IDU), a very complex and demanding system that will re-process all of the accumulated raw data received from the satellite using the latest calibrations available from other DPAC systems. It means that BSC will have to process about 10TB of data the first time, and about 100TB of data towards the end of operations. The complexity is further increased by the tight relations between the data elements. It must be noted that MareNostrum cannot run any database system, so all the I/O must be done directly on files. For this reason, the DPCB group is developing a set of tools for an optimum usage of the filesystem by the hundreds of nodes that will run during operations for IDU processing. Although GPFS is highly scalable, letting these many nodes access directly and intensively the disk would overload it and decrease its overall performance. A solution based on the excellent Myrinet network of MareNostrum is being implemented, which uses the FMPJ library (a highly efficient Java implementation of MPI). It concentrates the most demanding GPFS accesses in a few nodes, which act as MPI-based data servers and caches for the rest of computing nodes. Finally, regarding the communications with the outside, MareNostrum has a very strict security policy, so the Aspera-based data transfers with DPCE are done through an interface server, which has access to the central GPFS disk of MareNostrum.

We must remark that both CESCA and BSC obviously have adequate backup systems. In the case of CESCA, it is based on the NetBackup solution, while BSC implements a Hierarchical Storage Management system (HSM) with some PB of capacity. We must also note that both centers are interconnected with gigabit network, which makes possible to exchange simulation and test data between them.

Acknowledgements DPCB is supported by the MICINN (Spanish Ministry of Science and Innovation) - FEDER through grant AYA2009-14648-C02-01 and CONSOLIDER CSD2007-00050.

References

1. W. O'Mullane *et al*. 2010. Hardware and Networks. *Gaia: At the Frontiers of Astrometry*.
2. W. O'Mullane. Hardware and Networks.
http://wwwhip.obspm.fr/gaia2010/IMG/pdf/20100607_14_OMullane.pdf