InterSystems
White Paper

The European Space Agency: Charting the Galaxy with the Gaia Satellite and InterSystems Caché

Abstract

The European Space Agency (ESA) has chosen InterSystems Caché® as the database technology for the AGIS astrometric solution that will be used to analyze the celestial data captured by the Gaia satellite.

The Gaia mission is to create an accurate phase-map of about a billion celestial objects. During the mission, the AGIS solution will iteratively refine the accuracy of Gaia's spatial observations, ultimately achieving accuracies that are on the order of 20 microarcseconds.

In preparation of the extreme data requirements for this project, InterSystems recently engaged in a proof-of-concept project which required 5 billion discrete Java objects of about 600 bytes each to be inserted in the Caché database within a span of 24 hours. Running on one 8-core Intel 64-bit processor with Red Hat Enterprise Linux 5.5, Caché successfully ingested all the data in 12 hours and 18 minutes, at an average insertion rate of 112,000 objects/second.

William O'Mullane, Science Operations Development Manager, European Space Agency

Vik Nagjee, Product Manager, InterSystems Corporation

InterSystems White Paper

The European Space Agency: Charting the Galaxy with the Gaia Satellite and InterSystems Caché

Introduction

Space missions are long-term. Generally 15 to 20 years in length, they require robust long-term technologies for data processing, manipulation, and storage. These technologies must also provide critical up-to-date processing feedback in a timely fashion so that adjustments, if any, can be made rapidly to the spacecraft.

The Gaia mission is considered the biggest data processing challenge to date in astronomy. As mentioned in an InterSystems press release in May 2010, the European Space Agency (ESA) has selected InterSystems Caché® to support part of the scientific processing associated with the Gaia mission.

InterSystems and the European Space Astronomy Center (ESAC) have been working together since 2008 to see how InterSystems Caché can offer an advantage for some or all of the Gaia processing needs, and to create an economical computing architecture which can support the massive processing requirements of the Gaia project.

The Gaia Mission

The Gaia Satellite is scheduled for launch from French Guiana aboard a Soyuz-Fregat in 2012. It will spend a couple of months traveling 1.5 million km from Earth to L2, and spend the next 5 years scanning the entire sky. The goal: a phase-space map of our galaxy.

With two fields of view, a gigapixel focal plane and a radial velocity spectrometer, the 2000 kg satellite is a complete surveyor. During its lifetime, Gaia will observe 1 billion sources approximately 80 times each.

In addition to astrometry and photometry for every source, Gaia will measure spectra for approximately 150 million sources. Astrometric accuracies in the final catalogue are expected to be on the order of 20 microarcseconds. Achieving this accuracy requires extremely complex processing.

All Gaia data processing software is written in Java including the core astrometric solution known as the Astrometric Global Iterative Solution (AGIS), which iteratively refines the spatial accuracy of all the Gaia measurements. Since Gaia will be spinning freely and making observations which relate only to other observations made by Gaia, the collected data must be reduced in a self-consistent manner such that all individual observations of celestial sources, the model of each source's position and motion, and Gaia's own attitude, orbit, and velocity are in harmony. Later, the entire system may be aligned with the International Celestial Reference System (ICRS). It is the AGIS problem, about 10% - 50% of the entire Gaia processing, for which InterSystems Caché has been selected.

The scientific goals of Gaia are many-fold, but may be generally classed as unraveling the structure and formation history of our Galaxy.

Technical Challenges and Requirements

Gaia is expected to observe around 10^9 (that is, 1,000,000,000) celestial objects passing its focal plane; for each celestial object, it is expected that Gaia will observe about 100 attributes, totaling 10^{11} (that is, 100,000,000,000) observations. Of these, approximately 10% - 50% are expected to be used to construct a global reference frame using AGIS. Once the data in the global reference frame is calibrated and adjusted for attitude, it is used to update the positions and motions of the other sources in the catalogue. Figure 1 illustrates the high-level workflow of data flow between the Main Database and the AGIS Caché database.

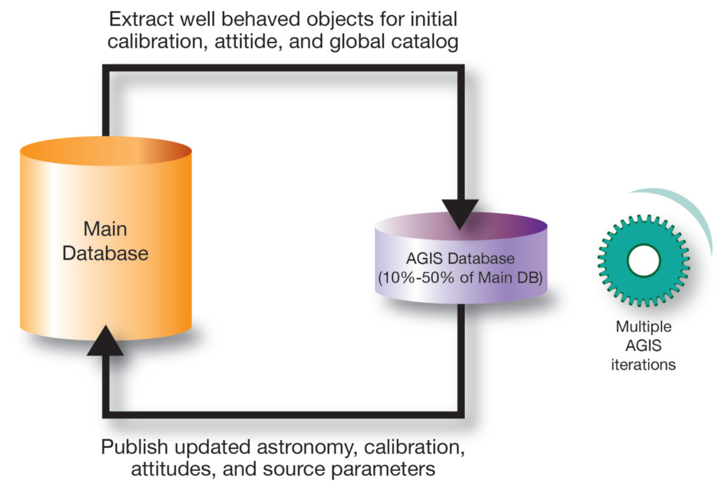
Extract well behaved objects for initial calibration, attitide, and global catalog
Main Database
AGIS Database (10%-50% of Main DB)
Multiple AGIS iterations
Publish updated astronomy, calibration, attitudes, and source parameters

FIGURE 1: HIGH-LEVEL WORKFLOW FOR AGIS EXTRACTION AND PROCESSING

It was previously estimated that the AGIS Caché database would contain the data for roughly 100,000,000 sources (totaling 10,000,000,000 observations). The size of this data was estimated to be in the order of 20 Terabytes. Recently, however, it has been suggested that the AGIS Database could contain up to 500,000,000 sources (totaling 50,000,000,000 observations), yielding a 100 Terabyte database. It is required that this data be ingested (or inserted) into the database within 7 days so that processing can begin immediately.

Once the data is ingested into AGIS, it is expected that around 40 iterations will be required to fully calibrate and adjust the data, and it is required that this be complete within 120 days. At the completion of the adjustment, the data from AGIS is fed back into the Main Database, and the next cycle is initiated. This iterative processing will continue for the life of the mission. Furthermore, the entire Gaia Data processing is iterative – the improved positions from AGIS allow other processes such as photometry and variability to get better results. These in turn are used to improve the next AGIS solution.

Data Ingestion into the AGIS Caché Database

The AGIS Data Model comprises several objects and is defined in terms of Java interfaces. Specifically, AGIS treats each observation as a discrete AstroElementary object. As Figure 2 illustrates, the AstroElementary object contains various properties (mostly of the IEEE long data type) and is roughly 600 bytes on disk.

In addition, the AGIS database contains several supporting indexes which are built during the ingestion phase. These indexes assist with queries during AGIS processing, and also provide fast ad-hoc reporting capabilities.

Using InterSystems Caché, with its Caché eXTreme for Java capability, multiple AGIS Java programs will ingest the 100 Terabytes of data generated by Gaia as 50,000,000,000 discrete AstroElementary objects. This data ingestion, plus the building of supporting indexes, is required to be completed within 5 days, yielding a required sustained ingestion rate of roughly 115,000 AstroElementary objects per second.

```
class AstroElementary {
        long transitTimes[];
        long transitTimeErrors[];
        long HEALPIXID;
        long HTMID;
        double etaObs[];
        double zetaRes;
        double[] sourceParam;
        double[] etaRes;
        double zeta;
        double zetaError;
        float flux;
        float fluxError;
        float bg;
        float bgError;
        long id;
        long telescope;
        long ccdRow;
        short pixelColumns[];
        long detTime;
        long detTimeError;
        int typeFlag;
        long sourceId;
}
```

FIGURE 2: THE ASTROELEMENTARY DATA MODEL

Data Ingestion Proof-of-Concept

As a proof-of-concept, InterSystems and ESAC, working together with NetApp engineers, developed a test-bed to ingest 5 billion (5,000,000,000) AstroElementary objects – roughly 10% of the overall data volume expected in the AGIS Database upon completion of the mission. Per scale, this data would need to be ingested within 12 hours. However, due to non-production-grade hardware being used for this proof-of-concept, ESAC determined that the proof-of-concept would be declared successful if this data were to be ingested within 24 hours.

Table 1 summarizes the specifications of the test system provided for this test:

System	Information / Details
Server	One 8-core Intel-based system
OS	Red Hat Enterprise Linux 5.5 (2.6.18-194.el5), 64-bit
Memory	32GB RAM → 11GB allocated to Caché (global buffers)
File System	ext3
Storage	NetApp FAS3160 with 176 x 1 TB SATA disks @ 7200 RPM
Network	10 GigE, single-port, single-channel, Jumbo Frames enabled
Connection between host and storage	iSCSI over 10 GigE
Caché Version	2010.2, Field Test 6 (plus ad-hoc updates)

TABLE 1: SUMMARY OF TEST SYSTEM ARCHITECTURE

Using the Caché eXTreme for Java capability, the test harness was able to ingest the 5,000,000,000 discrete AstroElementary objects in 12 hours and 18 minutes, yielding an average sustained rate of 112,000 objects per second.

Item	Value
Number of AstroElementary objects ingested	5,000,000,000
Total run time	44,616 seconds (~12.5 hours)
Target (allotted) run time	86,400 seconds (24 hours)
Average ingestion rate	112,000 objects/second

TABLE 2: SUMMARY OF RESULTS FROM INGESTION PROOF-OF-CONCEPT

The test was considered exceptionally successful, especially since the inserts were completed in almost 50% less time than the allotted 24 hours, with a nominal system configuration.

Future ingestion tests will likely include multiple parallel ingestion programs, thereby potentially further increasing the average ingestion rate, and reducing the total amount of time for ingestion.

Conclusion

In a proof-of-concept project conducted by the European Space Agency and InterSystems, sample astrometric data was inserted into the InterSystems Caché database at an average rate of 112,000 objects/second. The entire test was completed in 12 hours and 18 minutes, just over half the allotted time of 24 hours, using nominal test hardware. As a result, Caché continues to prove the right choice as the database technology for the extreme data demands of the Gaia galaxy-mapping satellite.

About Caché

InterSystems Caché® is a high-performance database that allows object, SQL, and multidimensional access to data – without mapping. It powers breakthrough applications around the world in healthcare, financial services, government, telecommunications, retail, and other vertical markets.

About Caché eXTreme for Java

Caché eXTreme for Java is a new capability of the InterSystems Caché database that exposes Caché's enterprise and high-performance features to Java via the JNI (Java Native Interface). It enables "in-process" communication between Java and Caché, thereby providing extremely low-latency data storage and retrieval.

For more information, visit **InterSystems.com/java**.

About ESA

The European Space Agency (ESA) is Europe's gateway to space. Its mission is to shape the development of Europe's space capability and ensure that investment in space continues to deliver benefits to the citizens of Europe and the world. ESA's job is to draw up the European space program and carry it through. ESA's programs are designed to find out more about Earth, its immediate space environment, our Solar System and the Universe, as well as to develop satellite-based technologies and services, and to promote European industries. ESA also works closely with space organizations outside Europe.

ESA has sites in a number of European countries, each of which has different responsibilities. The European Space Astronomy Centre (ESAC) is ESA's centre for space science located near Madrid in Spain. ESAC is the location from which the science operations for space telescopes are conducted, and where all of the scientific data they produce is archived and made accessible to the world.

About InterSystems

InterSystems Corporation is a global software technology leader with headquarters in Cambridge, Massachusetts, and offices in 23 countries. InterSystems provides advanced software technologies for breakthrough applications. InterSystems Caché is a high performance object database that makes applications faster and more scalable. InterSystems Ensemble® is a seamless platform for integration and the development of connectable applications. InterSystems HealthShare™ is a platform that enables the fastest creation of an Electronic Health Record for regional or national health information exchange. InterSystems DeepSee™ is software that makes it possible to embed real-time business intelligence in transactional applications. For more information, visit InterSystems.com.

InterSystems Corporation

World Headquarters
One Memorial Drive
Cambridge, MA 02142-1356
Tel: +1.617.621.0600
Fax: +1.617.494.1631

InterSystems.com

INTERSYSTEMS