The OPS4: towards a legacy Big Data system

A detailed view

Enrico Licata
on behalf of INAF & ALTEC teams

INAF USC VIII - Calcolo Critico
Jun 15 – 16, 2023
Dipartimento di Fisica e Astronomia “Ettore Majorana” Universitá degli Studi di Catania Via S. Sofia, 64

Summary

- The team
- Experience on gaia operations
- Experience on TLS prototype (MITIC)
- OPS4: HW ad Logical Design
- Requirements and Scalability
- Use cases
- Advantages of hybrid data management
- Dynamic Data Modeling
- GAIA-like data exploitation and Oracle technologies
- Resiliency, data security and integrity
- Future plans

The Team

ALTEC

Rosario Messineo

Lorenzo Bramante

Leonardo Tolomei

Daniele Gontero

INAF
ISTITUTO NAZIONALE
DI ASTROFISICA

Deborah Busonero

Raffaella Buzzi

Maria Teresa Crosta

Mario G. Lattanzi

Enrico Licata

Roberto Morbidelli

Alberto Vecchiato

mediaMENTE
consulting

Daniel Procopio
Fabrizio Lupi
Dorianna Tedesco
Alberto Visentin
Cristina Zaccagnini

Experience on GAIA operations 1/2
HW infrastructure

INTERNET LINK: 10Gbps (300 Mbps guaranteed) via GARR

STORAGE CAPACITY: 2.5 PB overall raw disk space distributed on two HP P7400 storage units and one P8400.

COMPUTING: 14 servers HP DL580 G7/G9 with a total of about 600 CPU cores and 4.5TB RAM.

DEV & TEST: 7 servers HP

DB SERVERS: 3 servers HP DL580 G7 (32 cores, 256MB RAM each) based on Oracle RAC technology (DBMS Oracle).

NETWORK CONNECTION: LAN network up to 10 Gbps. SAN network redundant at 8 Gbps.

SECURITY SERVICE: redundant firewall based on pfSense, enabling secure remote access via VPN.

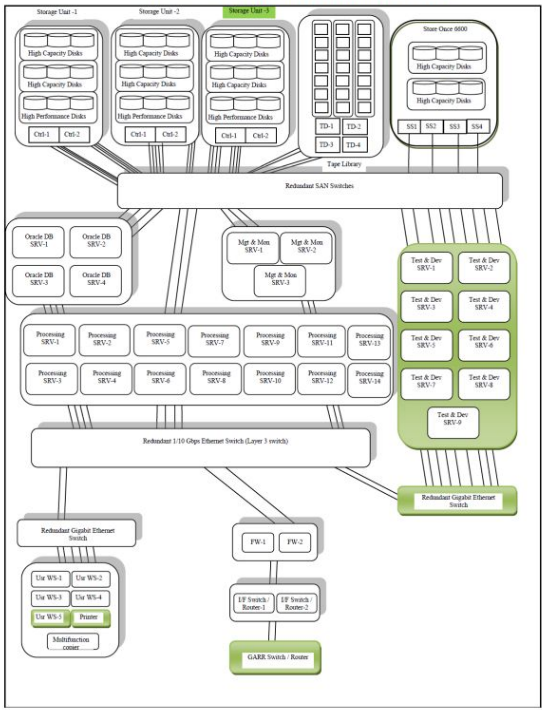
INFRA MONITORING AND MANAGEMENT: services based on VMWare virtual environment configured with two HP DL 580 G7 servers clustered and managed by vCenter Server.

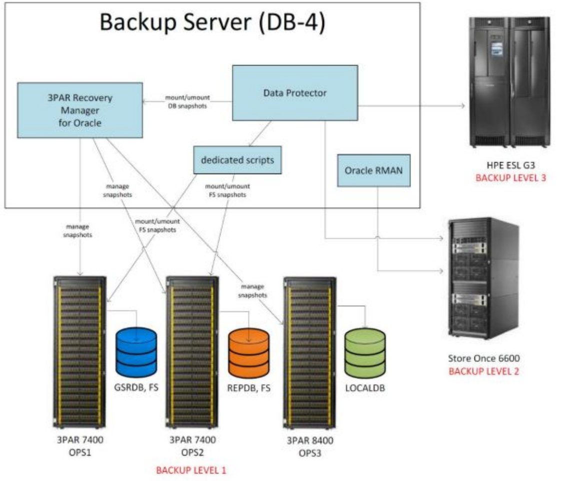
BACKUP SERVERS: HP DL580 G7 dedicated to DB and filesystem backups from data volume snaphots.

3 LEVELS BACKUP: L1 on primary storage array, L2 on disks (StoreOnce 6600) and L3 on tape libraries (HP ESL G3).

HPC INTERCONNECTION: access to HPC super computer at CINECA for dedicated processing

DPCT Overall HW infrastructure





Credits: ALTEC

Experience on GAIA operations 2/2

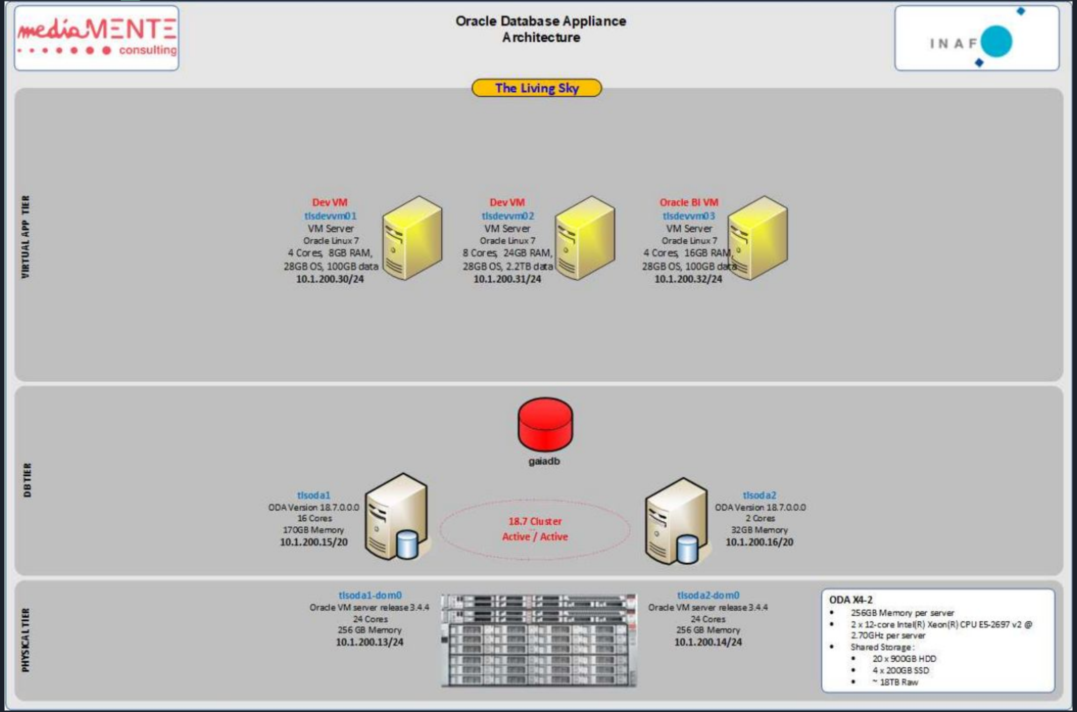
REQUIREMENTS

1. Massive dataset: up to ~10^7 observations /day
2. Pipeline integrated DBMS capable of supporting OLTP and OLAP operations
3. Max processing time 24h
4. Heterogeneous workloads
   - Computation
   - Monitoring
   - Visualization
5. Store mission data (5y) and all produced outputs of both daily and cyclic processing

CHALLENGES

1. Mission extended beyond the expected timeframe
   - Massive databases/tablespaces (~TB) with detrimental effect on DMBS performance and maintenance
2. Daily/cyclic operations require ≠ DBMS paradigms
3. To avoid backlog → code optimization, fast bug fixing, HA environment
4. Heterogeneous requirements for the datamodel and its indexing/partitioning:
   - Hybrid DM: Relational with serialized objects (BLOB)

Experience on TLS Prototype (MITIC) HW Infrastructure



Oracle Database Appliance Architecture

The Living Sky

Dev VM
tlsdevvm01
VM Server
Oracle Linux 7
4 Cores, 8GB RAM, 28GB OS, 100GB data
10.1.200.30/24

Dev VM
tlsdevvm02
VM Server
Oracle Linux 7
8 Cores, 24GB RAM, 28GB OS, 2.2TB data
10.1.200.31/24

Oracle BI VM
tlsdevvm03
VM Server
Oracle Linux 7
4 Cores, 16GB RAM, 28GB OS, 100GB data
10.1.200.32/24

tlsoda1
ODA Version 18.7.0.0.0
16 Cores
170GB Memory
10.1.200.15/20

tlsoda2
ODA Version 18.7.0.0.0
2 Cores
32GB Memory
10.1.200.16/20

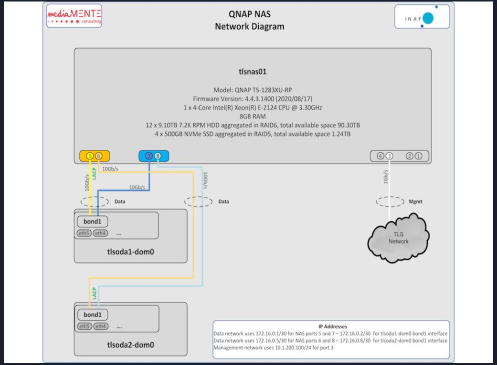
tlsoda1-dom0
Oracle VM server release 3.4.4
24 Cores
256 GB Memory
10.1.200.13/24

tlsoda2-dom0
Oracle VM server release 3.4.4
24 Cores
256 GB Memory
10.1.200.14/24

ODA X4-2
- 256GB Memory per server
- 2 x 12-core Intel(R) Xeon(R) CPU E5-2697 v2 @ 2.70GHz per server
- Shared Storage:
  - 20 x 900GB HDD
  - 4 x 200GB SSD
  - ~ 18TB Raw

TLS Prototype Achitecture: 4th generation Oracle Database Appliance X4-2

Daniel Procopio (Mediamente Consulting S.r.L.)
ed Enrico Licata (INAF-OATo) @ ALTEC datacenter



QNAP NAS Network Diagram

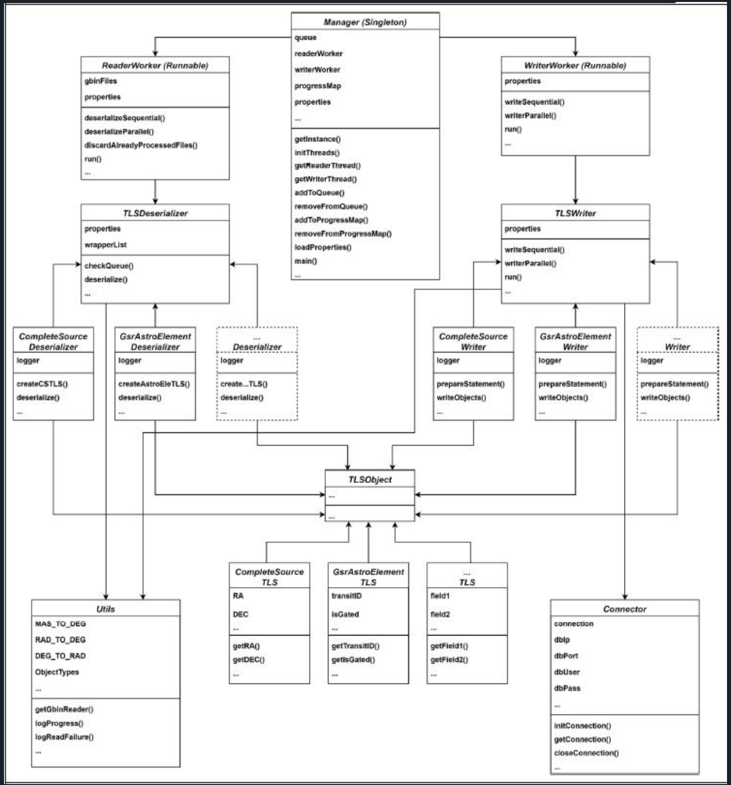
tlsnas01

Model: QNAP TS-1283XU-RP
Firmware Version: 4.4.3.1400 (2020/08/17)
1 x 4-Core Intel(R) Xeon(R) E-2124 CPU @ 3.30GHz
8GB RAM
12 x 9.10TB 7.2K RPM HDD aggregated in RAID6, total available space 90.30TB
4 x 500GB NVMe SSD aggregated in RAID5, total available space 1.24TB

IP Addresses
Data network uses 172.16.0.1/30 for NAS ports 5 and 7 – 172.16.0.2/30 for tlsoda1-dom0 bond1 interface
Data network uses 172.16.0.5/30 for NAS ports 6 and 8 – 172.16.0.6/30 for tlsoda2-dom0 bond1 interface
Management network uses 10.1.200.100/24 for port 3

Storage NAS QNAP TS-1283XU-RP

Experience on TLS Prototype (MITIC)
SW Infrastructure



1. **Manager**: entry point , coordinator, ***Singleton*** design
2. **Read/WriteWorker:** ***Parallel implementation,*** queue
3. **TLSDeserializer:** Abstract, I/O optimized
4. **TLSWriter:** Abstract, **Java Reflection**
5. **TLSObject:** Supertype, **ObjectModel**
6. Connector: ***ojdbc7*****,** manage CPU by limiting

   connections/processes

High level Software design of the TLS Prototype Integrator used to populate the DB

Experience on TLS Prototype (MITIC) 1/2

Prototype based on an Oracle DBMS able to provide an impact assessment, in terms of hardware resources and software tools, of specific scientific "test cases", on selected subsets of astronomical and/or technological data from the Gaia mission. In particular, to explore the possibility of carrying out operations currently not foreseen or not feasible on the "Mission Database" of the Gaia mission, or more generally, with the tools available in the astronomical community, starting from (but not limited to) the VO

REQUIREMENTS

1. **PoC0:** indexing and cone search on the entire GAIA sky (DR2)
   - Cross Match the Complete Sources
   - DM for scientific exploitation (legacy)
2. **Poc1:** GSR Residual computation
   - provide a R&D environment, to perform scientific analysis of results of GSR pipeline
3. **PoC2:** Gravitational Waves: analyze portions of the sky (few square degrees) to identify significant variations of some sources, to verify the effective presence of gravitational waves.

Experience on TLS Prototype (MITIC) 2/2

HIGHLIGHTS

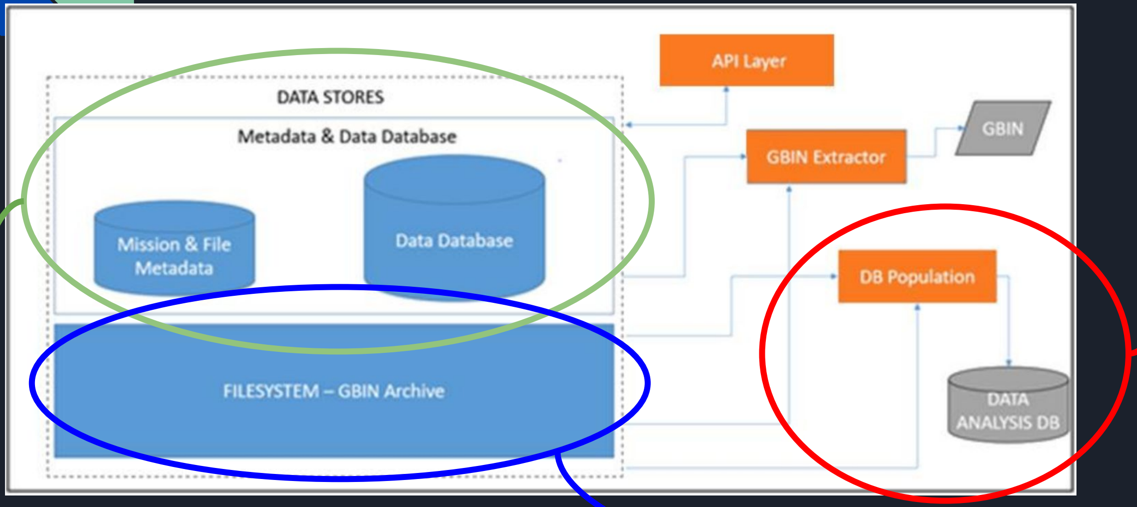
1. Indexing via Oracle Spatial & graph based on the orthodromic formula

   $$\gamma = arccos(cos(ra_2 - ra_1) * cos(dec_2) * cos(dec_1) + sin(dec_2) * sin(dec_1))$$

   with ra_1, ra_2, dec_1, dec_2 the right ascension and declination of bodies 1 and 2 and γ their angular distance
2. TLS Integrator to transform the DM from the Hybrid relational (operations) to a purely relational to allow direct access to scientific data
3. Develop a data Visualization layer to query the data and analyze it by Source and by Time
4. The amount of data required to complete PoC2 was beyond the capabilities of the prototype HW
   - used the data from the GAREQ experiment
   - remodeling of the DM

Poc0 complete reliability was verified with respect to the corresponding VO methods demonstrating quantitatively that it can operate without some "quantitative" limits (VO crash for CS with over 2 million objects extracted) and without time-loss with respect to the corresponding method of the VO (the execution of the Cone Search was often even faster).

OPS4: HW & Logical Design



API Layer

DATA STORES

Metadata & Data Database

Mission & File Metadata

Data Database

FILESYSTEM – GBIN Archive

GBIN Extractor

GBIN

DB Population

DATA ANALYSIS DB

Oracle Database Appliance
X8-2-HA, TAA

HPE Nimble HF40

Oracle ZFS Storage
Appliance Racked
System ZS9-2

Requirements and scalability

REQUIREMENTS

- Store all GAIA Mission data (raw to level 3) and metadata ~2.5 PB
- Hot data: most common queries in *almost real time*
- Cold Data: accessed and made available in *short time*
- Provide flexible platform to perform scientific analysis: Docker, Notebook, VMs , DataAnalysis DB (first phase on ODA) and supply different data formats: binaries, and open standard File Formats like FITS, CSV, XML…

SCALABILITY

- Long term storage provided on an expandable ZFS
- Hybrid approach: scale up / scale out:
  - Nimble & ZFS can be scaled out
  - ODA will require a scale up but…
- The entire ODA is configured as a pluggable component on the data stores, and can easily be replaced
- The design and data management workflows implemented for daily data can be easily replicated to support also cyclic data

All the OPS4 Infrastructure can be upscaled onto an Exadata machine, allowing for great scalability, that will be necessary to provide access also to all the cyclic data

Use Cases

Based on our experience we anticipate 3 main use case categories: by position, by time and by a combination of the two. Caveat: accessing data by position or by time requires the creation of substantially different indexing strategies and DM design.

By Position

- Indexing on healpixes proved ineffective
- Leveraging the capabilities of Oracle Spatial proved to be extremely more effective
- Partitioning strategies, based on healpixes instead could prove to be a useful strategy

By Time

- Usually full table scans indexing on transitIDs
- Leverage a time based partitioning strategy
- Suitable for short time frames (days / months) or full mission limited by XM
- More suited t o perform calibrations and analysis about the instrument’s response

Use cases like the detection of gravitational waves using GAIA Astrometry fall in the third category, requiring to access data both by time and position

We also aim at supporting the possibility to experiment other “dimensions” of investigation , based on other quantities / metadata related to the observation

Advantages of hybrid data management

A paradigm shift is required to move from operations to verification and then exploitation (legacy)

- ***Energy efficiency:*** cold data / cold place → Long term data storage on file system
  - not tape, to fulfill scientific data requests, responsive access is still needed
- ***Fast response times:*** hot data / hot place → 2 DBs
  - 1 Data Database: with a subset of data (vertical cut) to respond quickly to most frequent requests
  - 1 Metadata Database: to store mission metadata and to relate requested data to their respective location on the filesystem: larger and/or less frequent requests
- **Scalability:**
  - easier to expand ZFS storage, keeping high level of performances leveraging Metadata DB
  - DataAnalysis DBs are hosted on engineered machine, pluggable, lift and shift…

Dynamic Data Modeling

Lesson learned on TLS prototype is the importance of a DM suitable for technical/scientific exploitation: the one used by the GAIA operations/pipeline is NOT. The implementation of TLS Integrator Prototype was effective to complete the PoCs but reusability and maintainability can be an issue.

This experience highlights the importance of a DM as dynamic as possible:

The Population of the DataAnalysis DBs can be done leveraging the advantages of **JSON** fields and collections inside a Oracle DBMS:

- JSON fields are inherently schema-less but can be queried, partitioned and indexed as primitive fields (also supporting Oracle Spacial)
- “document like” objects can be created from the relational database, and using SQL syntax , used to create Tables on the DataAnalysis DBs tailored for a specific use case
- This operations could be scripted to automate the construction of the DataAnalysis DBs

Starting from Oracle 23c ***Relational Duality*** will allow the creation of JSON views based on relational tables, allowing the process layer to manage “Object like” data structures, which will be automatically mirrored by relational tables on the DB side

GAIA-like data exploitation and Oracle technologies

In-Memory

- Oracle Database In-Memory implements state-of-the-art algorithms for in-memory scans, joins, and aggregation. These optimizations, enable to run queries at billions of rows per second for each CPU core.
- Perfect application for switching between a Source oriented search by row (space), to a columnar search, by Transit (time) leveraging both indexing methods without the need to duplicate the DB volume

Spatial

- leverage GIS to create a spatial index (the full GAIA sky): defined using orthodromic distance
- "Target" object is quickly placed in relation to the "candidate" objects present in the catalog, allowing only those satisfying the requirement to be identified among all distances.

Relational Duality (starting from Oracle DBMS 23c)

- Using Duality Views, data is still stored in relational tables in a highly efficient normalized format but is accessed by apps in the form of JSON documents. Developers can thus think in terms of JSON documents for data access while using the highly efficient relational model for data storage, without having to compromise simplicity or efficiency.

Resiliency, data security and integrity

Data Resiliency: A comprehensive data resiliency strategy include backups, snapshots, mirrored copies of data, synchronous and asynchronous replication and off-site redundancy.

- At the ***actual*** stage OPS4 DBs and ZFS have redundancy level 1: mirror
- **Composable infrastructure**
- **Global policies**

Data Security: Access to the OPS4 environment is strictly regulated, and access levels are defined on principle of least privilege.

- Access to the ZFS and the Data/Metadata DBs is, ***for now***, only available on the DPCT local network
- Interaction with the DBs and ZFS can happen only through SQL queries. whose results will be made available on a staging area reachable by authorized users via a VPN or on the DataAnalysis DBs, whose access policies are still in development

Data Integrity: aimed at ensuring accuracy, completeness and coherence of data

- All serialized mission data (Gbins) transferred to the ZFS have been checked with MD5 Checksum
- Integrity constraints are enforced by the database for all the data under DBMS
- DM design choices aimed at guaranteeing completeness and coherence of data

Future Plans

Short term (end of 2023):

- put the OPS4 system online, to support the GAIA Operations
- data exploitation for the Verification

Mid term:

- OPS4 system available to users, with its fundamental set of features
- Support on-site data exploitation and analysis

Long term:

- Complete the OPS4 system with the entire GAIA legacy

Thank you for listening