

# THE GAIA DPCT EXPERIENCE

IN DATA PROCESSING AND DATA MANAGEMENT

AN ASI-INAF FACILITY

**DEBORAH BUSONERO - INAF**  
DPCT SCIENCE MANAGER

Funded by ASI and INAF

ALTEC: ASI N.2016-17-I.0 e atto aggiuntivo N.2016-17-I.1-2018

INAF: accordo attuativo N.2018-24-HH.0

Roma, 17-06-2019

INAF Science Archives & the Big Data Challenge

# OUTLINE

- Overview
- Goals
- INAF-OATo / ALTEC collaboration
  - The Team
- Data Processing Center main activities:
  - Operations
  - The Data
  - The processing pipeline
  - Infrastructure details
  - Data Requests support to the scientific community for Gaia performance papers
  - Gaia Mission Data Exploitation

# OVERVIEW

- Gaia is an ESA Space Mission to chart a three-dimensional map of our Galaxy, the Milky Way, in the process revealing the composition, formation and evolution of the Galaxy.
- Launch in December 19th, 2013
- Nominal mission end after 5 years on July 2019;  
Mission extended to 31 December 2022 (subject to a mid-term review in 2020)

DPCT  one of the 6 Gaia Data Processing Centers belonging to the Gaia Science Ground Segment

Data have been received and processed without interruption since February 2014:

255600      Workflows

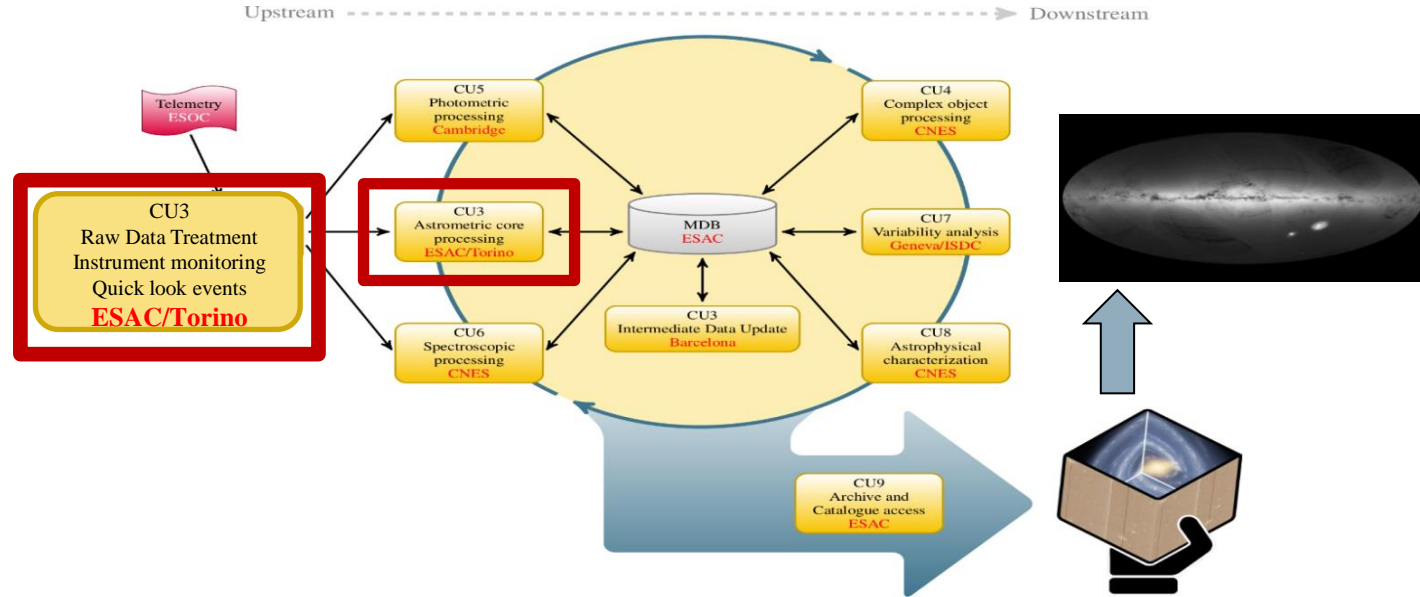
16129      Mission log entries

26497293      Jobs

3800      Daily pipeline reports

1.3 PB      DB size

# ➤ GOALS:



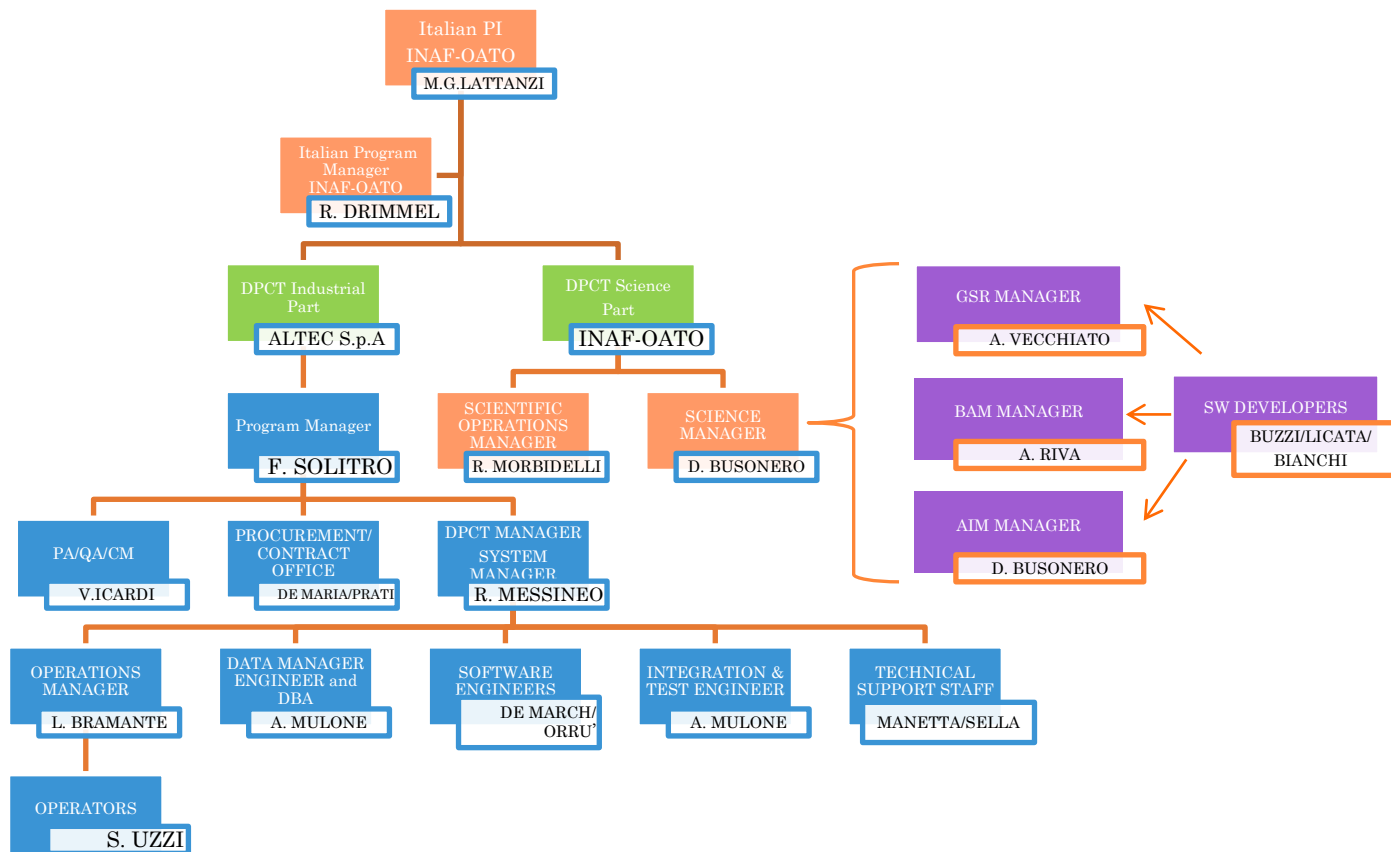
- Develop, implement and running the three independent processing pipeline of Astrometric Verification Unit
- Two daily pipeline AIM and BAM devoted to the raw data treatment and Astrometric Instrument monitoring and calibration
- One DRC pipeline GSR devoted to the astrometric core processing
- Support the GAREQ experiment via Data Requests execution

## ➤ GOALS:



- Providing the SW and HW infrastructure, the reprocessing capabilities, the data and the skills to support the Italian scientific community participation to DPAC and beyond
- Populate and maintain the **DPCT Mission Data Base** from raw data to final data through the intermediate data for scientific exploitation and future data reductions,

# INAF-OATo / ALTEC joint effort: THE TEAM



# OPERATIONS

➤ **Dedicated Team** supervising and verifying:

- Data flow from DPCE to DPCT,
- The whole data management chain from the data receiving, ingestion to the archiving
- The automatic advancement of the scientific pipeline,
- The correct functioning of the operation platform and the evolution of the DPCT operation system



➤ Milestones up to now:

- 5 data segments processed until now; 4 data segments at the end of Mission
- DR1 e DR2 catalogue
- On going the acquisition of the Data Segment-05 and of 3rd data processing cycle which will be used to realize the DR3 delivery

➤ The Operation team schedule depends on the **mission events** (VPU upgrade, Safe Mode, Decontamination, etc.).






# OPERATIONS WEB PORTAL



The current location is: **AIM**









































Welcome Aspera DTS RMS PFS Output Data Management **AIM** BAM Cacti Mission Log SOC Decoders Procedures  
DAAS DataRequests Databases Logs IDL Documentation EAR Whiteboard Query Setup Miscellaneous

Gala DPCT > AIM

AimRun

Edit Download   

OPS  

IDENTIFIER	STATUS	CLOSED	WORKFLOWIDS	ACTIVEPROCESSING	SOLUTIONID	SOLUTIONSIDS PROCESSED	SOLUTIONIDSTOBEPROCESSED	SOLUTIONSUNDERPROCE!
2011	DEFINED	false			0			
2010	DEFINED	false			0			
2009	DEFINED	false			0			
2008	DEFINED	false			0			
2007	DEFINED	false			0			
2006	DEFINED	false			0			
2005	DEFINED	false			0			
2004	DEFINED	false			0			
2003	DEFINED	false			0			
2002	DEFINED	false			0			

Settings Online Friends (1)



# OPERATIONS WEB PORTAL

Welcome Aspera DTS RMS PFS Output Data Management AIM BAM Cacti Mission Log SOC Decoders Procedures DAAS  
DataRequests Databases Logs IDL Documentation **EAR** Whiteboard Query Setup Miscellaneous

Gaia DPCT > EAR

## EventAnomalyReport



Refresh

[Event Anomaly Report](#)

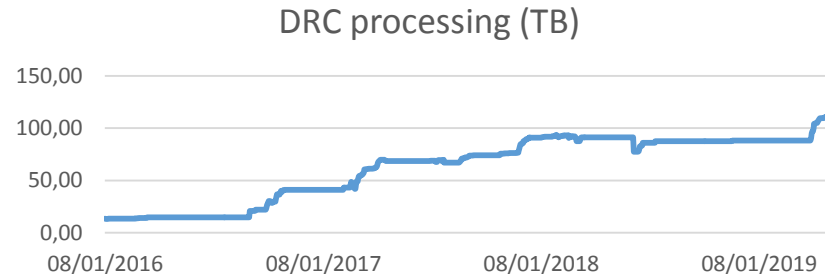
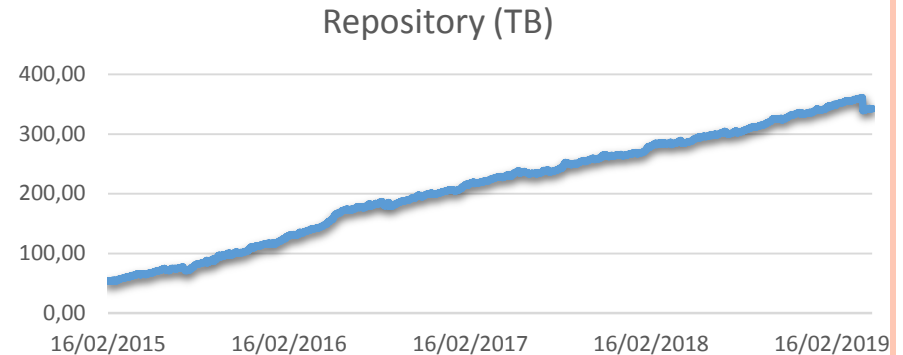
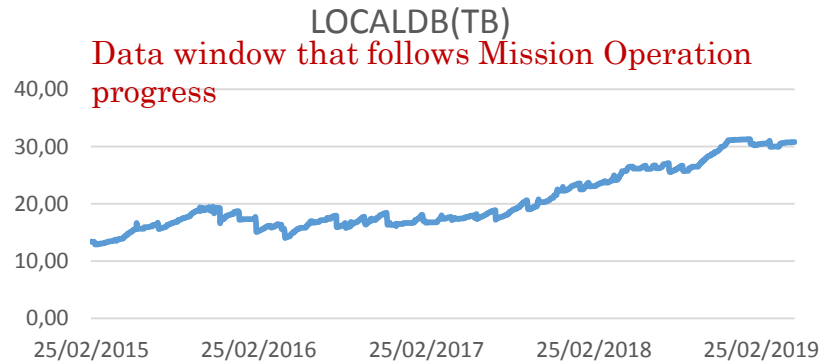
SpaceCraft Event Anomaly Report

DPCE Event Anomaly Report

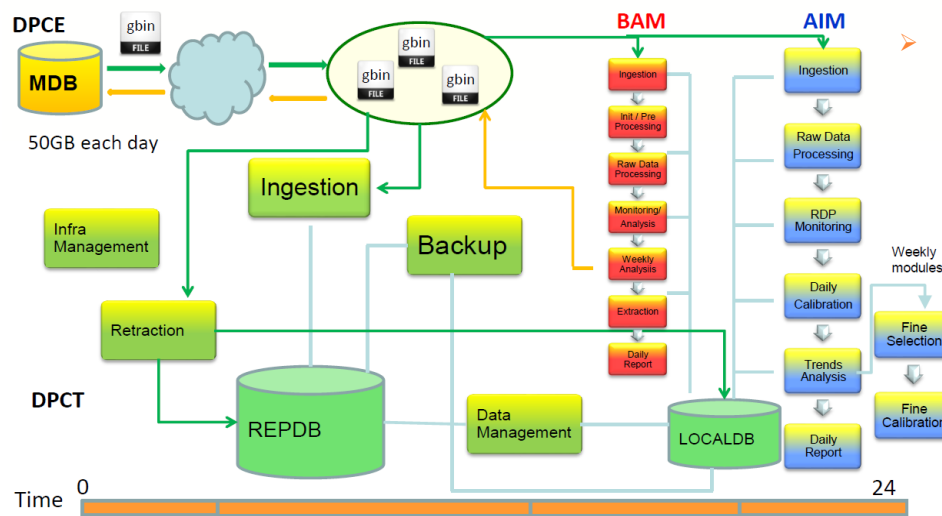
Solutionid	Starttime	Endtime	Reporter	Report
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
275001052246310912	23292998211919880	0	ncheek	DPAC data processing data segment 0 sta
275005450292822023	27521198213183800	27607598213209628	rguerra	A strong solar flare (X1.6) took place at 17:
275009848339333120	27747998224313360	27974798224552324	rguerra	Incorrect CDB update (0028891) resulted
275011497606774784	23292998211919880	0	rguerra	Data range for OR#5 starts on July 25, 201
275011497606774786	0	26097849998208300	rguerra	Data range for OR#5 end on 2014-08-26T
275014246385844227	23547265920000000	23547565920000000	rguerra	A major hit (amplitude <-10 mas/s in AL, ?
275014246385844228	23055258960000000	23055558960000000	rguerra	A major hit (amplitude -13.3 mas/s in AL, +
275014246385844229	23453284320000000	23453584320000000	rguerra	A major hit (amplitude 0 mas/s in AL, +24

# DATA FLOW AND STORAGE CAPABILITY

- **Data/day** (50 GB) and **Data/cycle** (10-60 TB) received from DPCE
- 3 data stores designed with different characteristics for supporting the 3 different phases of data management ➡ 3 different DBMS used for the I / O of the pipelines and for archiving.
- **DB SERVERS:** 3 HP DL580 G7 dedicated to the database cluster based on Oracle RAC technology.



# DAILY PIPELINES



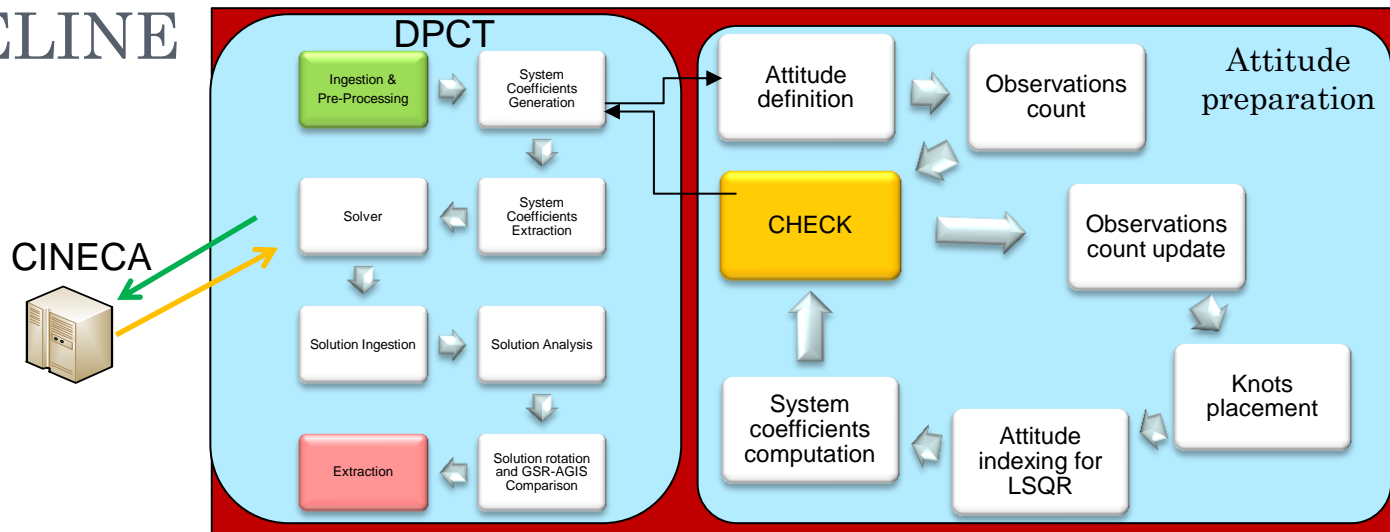
➤ **AIM pipeline:** raw data processing, image parameter determination, LSF/PSF modelization and calibration, astrometric instrument monitoring and diagnostics throughout the mission lifetime

- 1920 daily runs
- 24 hours of raw data each run: **from  $2 \times 10^6$  to  $15 \times 10^6$  raw images**
- Complex structure of the pipeline: 10 sw modules managed by a coordinator in an automatic way, the output of one run become the input of the next one
- **6 hours** of time execution on the DPCT Operation platform **for each run**

➤ **AVU/BAM pipeline:** raw data coming from the Basic Angle Monitoring (BAM) instrument, i.e. fringes, for monitoring and analyzing the instrument behaviour throughout the mission and performing the BAV calibration.

- 1950 daily runs
- 24 hours of raw data each run: almost  **$8 \times 10^4$  images**
- **1-2 hours for each run**
- The pipeline output is sent to DPCE and ingested into the MDB
- AVU/BAM runs also a cyclic version of the software aiming to fringes reprocessing for calibration improvement

# DRC PIPELINE



- **The Global Sphere Reconstruction** (GSR) solves a linearized system of equations whose result gives the global astrometric reference system (position, parallax, proper motions). This solution is compared with that of AGIS
  - GSR in Operations since the beginning of this year
  - **Starting from  $10^7$  to  $10^8$  objects for each run**
  - Very complex pipeline structure
  - Final GSR output sent to DPCE in the MDB
  - The Solver module run at CINECA which is managed as one processing node of the DPCT
  - The whole process could be iterated for Non-Linearity
  - **One run takes from 3 to 6 days** on  $10^7$  objects.

# HW INFRASTRUCTURE

- Operation and test & validation platform committed to the DPCT project
- Procurement performed incrementally according to mission needs

**INTERNET LINK** : 1Gbps (300 Mbps guaranteed) via GARR

**STORAGE CAPACITY**: **1.5 PB overall raw disk space**

distributed between 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 12.2c to 18.5).

**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 snapshots.

**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.



# USER ORIENTED SUPPORT FOR THE PERFORMANCE VERIFICATIONS CALLS

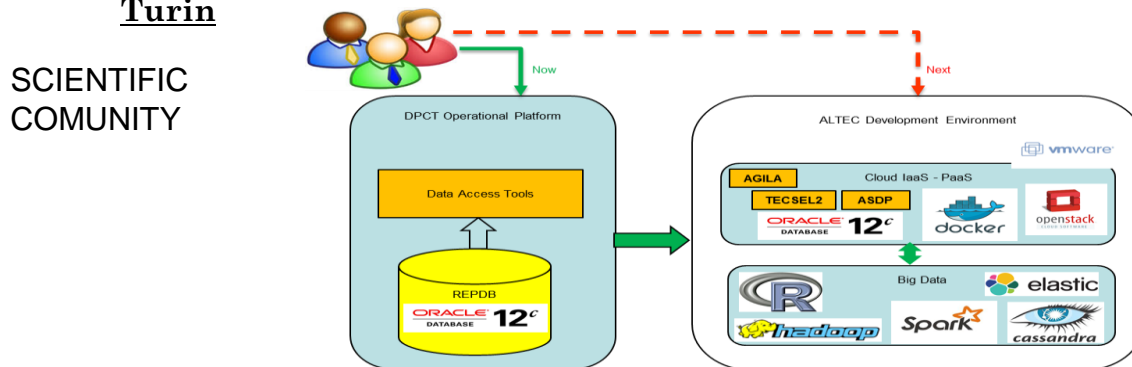
## ➤ Handling of the Data Request from the DPAC community

- Using the Gaia Jira web portal and the DPCT web portal
- DPCT Operation Manager, DPCT DBA and INAF Astronomer domain experts for the identification, preparation and extraction of the required data
- ICSR (International Celestial Reference System) has been implemented at RDBMS level
- Implementation of metadata structures to enable the DB exploitation by spatial criteria (for spatial-based data, like the sources)
- Source-only data are not enough
- A whole universe of intermediate data (**1200 billion of entry**) with not-trivial relations among them are needed. An in-depth knowledge of the Gaia data model and rules is required

## ➤ Up to now 27 Data Requests completed

# EXPLOITATION OF MISSION DATA

- DPCT hosts a fully consistent astrometric Mission Data Base, from raw to intermediate to final data
- Availability of the Gaia data extends for tens of years
  - Execution of new calibrations
  - New scientific experiments
  - Support to new Missions
  - Multidisciplinary activities
- The Living Sky” (TLS-MITIC) (see Lattanzi talk on Wednesday)
  - First step for a data exploitation platform.
  - “Big Data” techniques and methodologies developments in collaboration with the Politecnico of Turin



DATA EXPLOITATION  
PLATFORM

Page  
15