4.03 Rapporti di Progetto

Numero	
Anno di pubblicazione	2023
Accettazione in OA@INAF	
Titolo	**Definizione dei requisiti per l'esecuzione di Data Request su OPS4@DPCT** **per la Gaia Science Verification di Cycle 4 (DR4) e Cycle 5 (DR5)**
Autori	Mario G. Lattanzi, Roberto Morbidelli
Affiliazione del 1° autore	INAF-O.A.Torino
Progetto	Addendum Accordo Attuativo ASI-INAF n. 2018 24 HH.1 2022
Attività	Gaia DPAC - OPS4@DPCT developements

TABELLA TRACCIAMENTO MODIFICHE DOCUMENTO

ISSUE	DATA	RAGIONI DEL CAMBIAMENTO	PARAGRAFI COINVOLTI
D	20 Giugno 2023	"Template"	Creazione
0.1	21 Giugno 2023	"Draft"	Integrazione
0.2	06 Luglio 2023	"Draft"	Integrazione
0.3	13 Luglio 2023	"Draft"	Integrazione
0.4	14 Luglio 2023	Draft	Integrazione
0.5	18 Luglio 2023	Dopo incontro di MGL e RM con Leonardo Tolomei (DPCT@ALTEC) in sala Gaia	Revisione
0.6	24.Luglio 2023	Verifica pool di tabelle di input	Revisione
0.7	25 Luglio 2023	Redazione testo secondo form OA	Revisione
0.8	26 luglio 2023	Inserimento su Open Access INAF	Referee

Sommario

Introduzione 4

1. Referenze 5

2. Acronimi 7

3. Contesto e finalità 9

4. Casistica 11

a) Science Verification con estrazione dati per sorgenti (in numero Ns ≥ 1) per via del loro identificativo. 11

b) Science Verification con estrazione di aree di cielo circolare (Na ≥ 1) specificate mediante coordinate dei centri, dimensione dei raggi e intervalli temporali 12

c) Science Verification for Gravitational Wave events (con estrazione a seguito di ricerca per aree di cielo (Na > 1) e intervalli temporali) 13

Introduzione

Scopo di questa nota tecnica è di definire alcuni templates di querys "SQL like", da fornire al DPCT@ALTEC di Gaia, da implementare su OPS4 in vista delle attivitá di analisi e verifica per il rilascio della DR4. Questo avvalendosi, anche, delle esperienze condotte ed in corso di realizzazione delle varie PoC pertinenti il prototipo del sistema TLS (The Living Sky Prototype) realizzato in precedenza con risorse INAF presso il DPCT.

In particolare, si forniscono, nel seguito, le informazioni necessarie alla definizione degli input e degli output attesi da un utilizzatore generico ad esito dalle estrazioni dati (strutturati e non) condotte sul sistema OPS4.

Nel documento verrà per primo richiamato, ove funzionale allo scopo, il contenuto delle varie PoCs già esercitate, o solo tentate per insifficienza di risorse, sul sistema proptotipo TLS di cui sopra.

L'idea è che le relative ipotesi di riutilizzo, ove ereditabili, dei metodi sviluppati sul prototipo TLS verso il sistema ingegnerizzato OPS4, residente all'interno dell'infrastruttura operativa del DPCT, possano rendere piú agevole il conseguimento degli scopi qui illustrati.

Tuttavia il documento è da intendersi come propedeutico alla formulazione di nuovi metodi per lo sviluppo "tout court", a partire dalla consistenza dei dati presenti (ovvero da implementare) nel sistema ingegnerizzato OPS4@DPCT.

Si assume che gli input alle query provengano, prevalentemente, dall'utilizzatore, ma non è escluso siano esito di una disponibilità di dati già presenti in OPS4.

Si sottolinea che l'esito di una query, organizzata in un output, sarà costituito da uno o più files csv o FITS. Atteso essere integrazione tra il contenuto del "data lake" (GBIN files e tabelle eventualmente indicizzate) della missione GAIA e, ove esistenti, di ulteriori dati strutturati.

L'output in formato gbin deve essere considerato una eccezione transitoria in previsione di una adozione di gestione dei risultati in files codificati, finalmente in FITS tables.

1. Referenze

Documenti applicabili

- M. Martino, R. Messineo, R. Morbidelli;

DPCT Development Plan, GAIA-DT-PL-ALT-MM-002-11, 2016-07-12

- B. Garilli et al.;

"Mining The Cosmos: Big Data and Innovative Italian Technology for Frontier Astrophysics and Cosmology", INAF – Roma, 2016-31-10

- R. Messineo, R. De March et al.;

DPCT Internal Interface Control Document; GAIA-DT-SP-ALT-RM-002-20; 16-04-2020

- R. Morbidelli, M. Sarasso, M.G. Lattanzi ;

The Living Sky, Rapporto finale; Allegato al: TLS-MI-VarGroup-03 [minute Teleconferenza come da D.L. 17 marzo 2020, n. 18, art. 87, c. 1 – Misure connesse all'emergenza epidemiologica da COVID-19] INAF, OATO 20/11/2020

Documenti di riferimento

R1 D. Busonero[1], R. Buzzi[1], M. Crosta[1], M. G. Lattanzi[1], E. Licata[1], R. Morbidelli[1], A. Vecchiato[1]
Nota tecnica: The Living Sky POCs data model definition;
INAF, INAF Technical Reports, in press

R2 D. Busonero[1], M. Crosta[1], M. G. Lattanzi[1], E. Licata1, R. Morbidelli[1]
Nota Tecnica: Dettagli sul data model della PoC 2 per Il prototipo "The Living Sky";
INAF, INAF Technical Reports, in press

R3 D. Busonero[1], M.G. Lattanzi[1], E. Licata[1], R. Morbidelli[1], A. Vecchiato[1], R. Buzzi[1], A. Riva[1], M. Sarasso[1], R. Messineo[2] , F. Solitro[2], A. F. Mulone[2], C. Manetta[2], L. Bramante[2]
DPCT OPS4: Architecture Design and Sizing for the Expansion of the Italian Data Processing Center (DPCT), part of the Ground Science Segment for Gaia mission, INAF Technical Reports, in press

R4 Mario G. Lattanzi[1] - The Gaia legacy. From data reduction and analysis to data management and exploitation: HPC, HTC and Big Data issues[3];
INAF USCVIII - Calcolo Critico; 15/6/2023 Dip. di Fisica e Astronomia "Ettore Majorana" Università degli Studi di Catania

R5 Enrico Licata[1] et al. - The OPS4: towards a legacy Big Data system - A detailed view[4];

[1] INAF – OATo (Osservatorio Astrofisico di Torino)

[2] ASI - ALTEC (Aerospace Logistics Technology Engineering Company S.p.A.)

[3] https://indico.ict.inaf.it/event/2366/contributions/15131/

INAF USCVIII - Calcolo Critico; 15/6/2023 Dip. di Fisica e Astronomia "Ettore Majorana" Università degli Studi di Catania

[4] https://indico.ict.inaf.it/event/2366/contributions/15149/

2. Acronimi

Acronimo	*Definizione*
ADQL	Astronomical Data Query Language; IVOA SQL dialect based on SQL92
AGIS	Astrometric Global Iterative Solution
ALTEC	Aerospace Logistics Technology Engineering Company –S.p.A.
CS	Complete Source
CPU	Central Processing Unit
DB	Data Base
DDB	Data Database
DBMS	Data Base Management System
DEC	Declination
DM	Data Model
DPCT	Data Processing Center Torino
DR	Data Release
GDL	Gaia Data Lake
GDR	Gaia Data Release
GMAG	Gaia magnitude
GRAWITA	GRAvitational Wave Inaf TeAm
GSR	Global Sphere Reconstruction

Acronimo	*Definizione*
GW	Gravitational Wave
ID	Identifier (Identification)
IDL	Interactive Data Language
IDU	Intermediate Data Update
INAF	Istituto Nazionale di Astrofisica
MDB	Mission Data Base
OATo	Osservatorio Astrofisico di Torino
OBMT	On-Board Mission Timeline
ODA	Oracle Database Appliance
OGA3	On-Ground Attitude from AGIS
PoC	Proof-of-Concept
RA	Right Ascension
SDO	(Oracle) Spatial Data Option
SQL	Structured Query Language
TLS	The Living Sky
XM	Cross-Matching
XML	eXtensible Markup Language

3. Contesto e finalità

Il sistema OPS4 è una infrastruttura ingegnerizzata ibrida (Oracle e HPE), basata su DBMS Oracle 19 G.
È costituita, in termini di hardware, da un sistema Oracle ZFS Storage Appliance Racked System ZS9-2: ATO model family collegato ad un Oracle Database Appliance X8-2-HA, TAA, questi due elementi sono interconnessi a loro volta, con uno storage HPE Nimble H40.
Nel sistema ZFS sono immagazzinati i file gbin (contenenti sia dati grezzi che elaborati, il nostro "data lake" o GDL come lo abbrievieremo in seguito), mentre i metadati (organizzati in tabelle Oracle) sono alloggiati nel Data Database (DDB) realizzato, insieme al Mission Data Base (MDB) che contiene i metadati relativi, e.g., ai dati orbitali e strumentali utili per l'analisi), sull'unitá Nimble a costituire il nuovo paradigma della gestione dati al DPCT come riportato nella nota tecnica " DPCT OPS4: Architecture Design and Sizing for the Expansion of the Italian Data Processing Center (DPCT), part of the Ground Science Segment for Gaia mission, INAF Technical Reports" (R3).
Nuovo paradigma che viene realizzato integrando pienamente OPS4 nell'infrastruttura operativa del DPCT in conseguenza del fatto che la presa dati in orbita è stata estesa di quasi 6 anni, dal luglio del 2019 al giugno del 2025, per un totale di vita operativa di 11 anni (invece degli iniziali 5 previsti).

Dal punto di vista delle queries di analisi e validazione per le future DR4 e DR5, attività fondamentali che riguardano sia l'analisi di dati giornalieri calibrati che quelli ciclici, il ruolo fondamentale dell'ODA è quello di ripristinare queste attivitá che erano state sospese a causa della saturazione dello spazio disco (sotto DBMS in ambiente operativo) derivante dalle ripetute estensioni della durata della presa dati in orbita di Gaia.

Il sistema delle "Gaia operations" opera secondo schemi strutturalmente legati alle "pipe" di processamento sulla base sia del flusso dei dati che quotidianamente giungono dal satellite (pipelines giornaliere) che su calibrazioni "cicliche" che richiedono il processamento di dati raccolti e/o prodotti nell'arco di periodi temporali maggiori rispetto a quelli giornalieri.
Pertanto, i "templates delle data request" qui di seguito proposti vanno a descrivere metodologicamente, anche se non sintatticamente, casi di estrazione dati diversi da quelli operati sia dalle pipelines gironaliere che cicliche. In ogni caso, queste estrazioni sono comunque connesse sia al GDL che ai metadati, di cui sopra, presenti nel DDB e nel MDB.

Queste estrazioni si articoleranno:
nella selezione di un elenco di parametri, ritenuti scientificamente/tecnicamente rilevanti; nella individuazione delle tabelle e/o files in cui questi parametri sono presenti come tali od in forme da cui derivarli; nell'attuazione delle operazioni atomiche che sono eseguite per sodisfare quanto formulato da un qualsiasi utilizzatore, membro del Consorzio DPAC, per tramite del Team INAF@DPCT, partecipante alla fase di verifica e validazione precedente le Gaia Data Releases (GDR).

Di seguito, dunque, i casi "notevoli" finalizzati a:

1) Definire e consolidare la dimensione dei dati interessati sulla base dei Data Model e/o delle struture dei files (gbins) a supporto;
2) Verificare la consistenza dei dati presenti in OPS4 (siano essi contenuti in tabelle o gbin);
3) Definire e realizzare le operazioni di "data mining", basati sulla formulazione in linguaggio SQL (ma non solo quello, come ad esempio è il caso di dati non strutturati) avvalendosi, quanto più possibile, di metodi ancillari già esistenti per il DBMS Oracle (Spatial, BI, Inmemory ecc…) al fine di ottimizzare il conseguimento degli output attesi;
4) Definire e strutturare gli output delle queries.

Di seguito i tre casi notevoli presentati nella sezione successiva nella forma di pseudo queries:

a) Science Verification con estrazione dati per sorgenti (in numero Ns ≥ 1) per via del loro identificativo.
b) Science Verification con estrazione di aree di cielo circolare (Na ≥ 1) specificate mediante coordinate dei centri e dimensione dei raggi.
c) Science Verification for Gravitational Wave events (con estrazione a seguito di ricerca per aree di cielo (Na > 1) e intervalli temporali)

4. Casistica

A prerequisito dei casi esposti di seguito va sottolineato che gli output possono essere la totalità, ovvero un sottoinsieme, di quella che nel seguito viene definita come una selezione di variabili, pertinenti, esistenti nel GDL ovvero esito di esecuzione diretta su di essi di operazioni e pertanto da esse derivate. Esito per queste attività saranno comunque dei files i cui formati saranno preferenzialmente testuali (es. ASCII UTF 8 o UTF 16) contenenti dati in unità di misura chiaramente specificate.

Come estrema ratio, ove materialmente impossibile procedere nell'immediato ad una tale definizione dell'output, sarà temporaneamente ammesso l'utilizzo del formato disponibile, ivi incluso il GBIN.
In nessun caso, però, l'output di una query potrà essere costituito da metadati o da dati che richiedano per la loro interpretazione ulteriori operazioni di elaborazione in OPS4.

Da privilegiare sono, in tal senso, la produzione di tabelle in formato csv e, finalmente, FITS.

a) Science Verification con estrazione dati per sorgenti (in numero Ns ≥ 1) per via del loro identificativo.

Data una o piu' sorgenti (Ns), per le quali viene specificato il Gaia Source ID, si ottengono per essa i dati della/e sorgente/i e quelli d'epoca presenti all'interno di intervallo temporale specificato.

Strategia: select single/multiple object oriented + tempo

Select (lista variabili), from (GDL), where SOURCEID = ####;

Select (lista variabili), from (GDL), where SOURCEID = #### & TIME_START>=##### & TIME_END<=######;

Output attesi dalle seguenti tabelle e/o files su file system[5]:

1. CompleteSource
2. Astroelementaries, Photoelementaries [e Spectroelementaries]
3. Astroobservations [Photoobservations, Spectroobservations]
4. AuxData

b) Science Verification con estrazione di aree di cielo circolare (Na ≥ 1) specificate mediante coordinate dei centri, dimensione dei raggi e intervalli temporali

Data una o piu' areee di cielo (Na) specificate da centro e raggio (in gradi decimali), vengono estratti i dati per le sorgenti in essa(e) contenute e quelli d'epoca presenti all'interno di intervallo temporale specificato[6].

Questo tipo di query fa riferimento a situazioni dove l'utenza conosce solo l'esistenza di eventi in una o piu' direzioni della volta celeste ma nulla delle eventuali sorgenti Gaia coinvolte. Lo scopo è proprio quello di contribuire alla caratterizzazione di questi eventi comunque segnalati.

Strategia: cone search oriented (già indagata nella POC0 TLS (R1) ma da estendere a criteri di selezioni temporali.

Select (lista variabili), from (GDL), where RA = ####, DEC = #### & distance(*) <= ####;

[5] Seppure non riportati in questo caso come nei successivi non bisogna escludere la possibilità di utilizzo dei dati spettroscopici quantomeno a partire da quelli contenuti nella tabella delle Spectroelementaries

[6] Fino anche a copertura totale della sfera celeste.

Select (lista variabili), from (GDL), where RA = ####; DEC=##### & distance (*) <= ### & TIME_START>=##### & TIME_END<=######;

Output attesi dalle seguenti origini:

1. CompleteSource
2. Astroelementaries, Photoelementaries [e Spectroelementaries]
3. Astroobservations [Photoobservations, Spectroobservations]
4. AuxData (orbit, attitude, Calibrations..)

c) Science Verification for Gravitational Wave events (con estrazione a seguito di ricerca per aree di cielo (Na > 1) e intervalli temporali)

Il risultato di una query di questo tipo sarà di ottenere, per un evento GW dato, tutto quanto necessario per eseguire le operazioni di correlazione parziale o totale sulla sfera celeste delle distanze angolari tra oggetti puntiformi eventualmente condizionandone la definizione degli archi ad un dato intervallo di magnitudine, colore, tempo, ecc..

A titolo di esempio, nel caso di GW da sorgenti Galattiche, il centro di estrazione sara' orientato nella direzione indicata dall'evento 'alert' e la dimesione estratta sara' almeno pari ad una calotta sferica di almeno 90 gradi con polo nella direzione data.

In questo caso, quindi, da lista, in input, di sorgenti (ad esempio da catalogo) si darà luogo ad un output pertinente diverse aree della volta celeste (Na>1) all'interno delle quali, per ogni oggetto soddisfacente la query, è specificato: il Gaia Source ID e le coordinate ad un equinozio di riferimento nonché tutti i dati d'epoca specificati all'interno di un intervallo temporale (transiti), presenti internamente ai GBIN coinvolti.

Strategia: per partizionamenti della volta celeste e intervalli temporali (es. Healpix level =### o intervalli temporali commisurati alla legge di scansione).

Select (lista variabili), from (GDL), where RA = ####; DEC=##### & distance[7] <= ### & TIME_START>=##### & TIME_END<=###### & [min < magnitude < max];

Output attesi dalle seguenti origini:

1. CompleteSource
2. Astroelementaries, Photoelementaries [e Spectroelementaries]
3. Astroobservations [Photoobservations, Spectroobservations]
4. AuxData (orbit, attitude, Calibrations..)

[7] L'intervallo temporale puo' corrispondere alla totalità della durata della missione come potrebbe essere, di nuovo, il caso qui menzionato di "alert" per le onde gravitazionali.