Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

WP4 DMs, Ingestion and Services Updates

Giacomo Coran

Spoke 3 II Technical Workshop, Bologna Dec 17 -19, 2024

Fermi Use Case

- Logical Data Model based on ObsCoreDM
- Physical Data Model as rappresentation of the archiving tabular structure described in two tables
- One table for the scientific data products
- One table for the instrumental data

photon

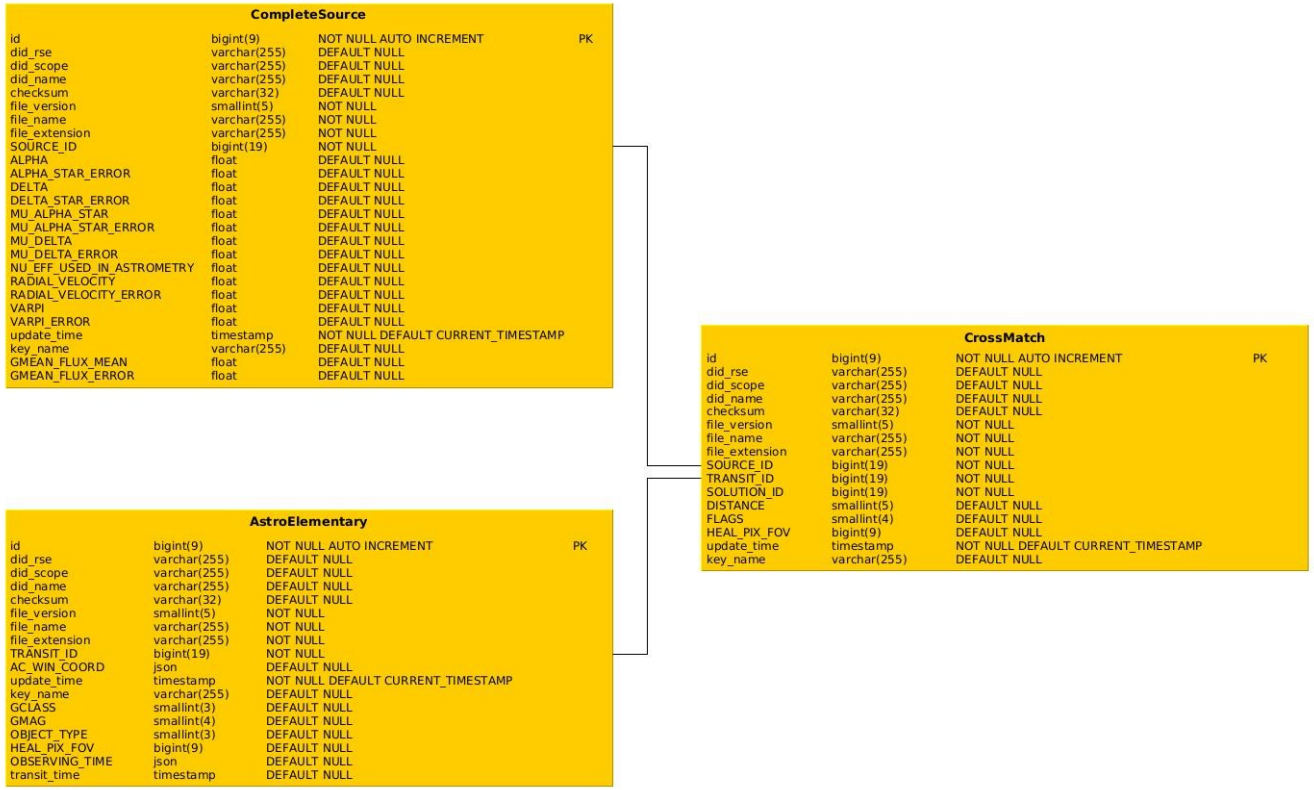
id	mediumint(9)	NOT NULL AUTO INCREMENT	PK
did_rse	varchar(255)	DEFAULT NULL	
did_scope	varchar(255)	DEFAULT NULL	
did_name	varchar(255)	DEFAULT NULL	
checksum	varchar(32)	DEFAULT NULL	
file_version	smallint(5)	NOT NULL	
file_name	varchar(255)	NOT NULL	
file_extension	varchar(255)	NOT NULL	
NAXIS_HDU0	smallint(6)	DEFAULT NULL	
EXTEND	tinyint(1)	DEFAULT NULL	
CHECKSUM_HDU0	varchar(32)	DEFAULT NULL	
TELESCOP	varchar(8)	DEFAULT NULL	
INSTRUME	varchar(8)	DEFAULT NULL	
EQUINOX	float	DEFAULT NULL	
RADECSYS	varchar(8)	DEFAULT NULL	
DATE	timestamp	DEFAULT NULL	
DATE_OBS	timestamp	DEFAULT NULL	
DATE_END	timestamp	DEFAULT NULL	
TSTART	float	DEFAULT NULL	
TSTOP	float	DEFAULT NULL	
TIMEUNIT	varchar(8)	DEFAULT NULL	
TIMEZERO	float	DEFAULT NULL	
TIMESYS	varchar(8)	DEFAULT NULL	
TIMEREF	varchar(8)	DEFAULT NULL	
CLOCKAPP	tinyint(1)	DEFAULT NULL	
GPS_OUT	tinyint(1)	DEFAULT NULL	
MJDREFI	float	DEFAULT NULL	
MJDREFF	float	DEFAULT NULL	
OBSERVER	varchar(255)	DEFAULT NULL	
FILENAME	varchar(255)	DEFAULT NULL	
ORIGIN	varchar(32)	DEFAULT NULL	
CREATOR	varchar(255)	DEFAULT NULL	
VERSION	varchar(8)	DEFAULT NULL	
PROC_VER	smallint(6)	DEFAULT NULL	
DATASUM_HDU0	varchar(32)	DEFAULT NULL	
NAXIS_HDU1	smallint(6)	DEFAULT NULL	
NAXIS1_HDU1	smallint(6)	DEFAULT NULL	
NAXIS2_HDU1	smallint(6)	DEFAULT NULL	
PCOUNT_HDU1	smallint(6)	DEFAULT NULL	
GCOUNT_HDU1	smallint(6)	DEFAULT NULL	
CHECKSUM_HDU1	varchar(32)	DEFAULT NULL	
DATASUM_HDU1	varchar(32)	DEFAULT NULL	
EXTNAME_HDU1	varchar(32)	DEFAULT NULL	
HDUCLASS_HDU1	varchar(32)	DEFAULT NULL	
HDUCLASS1_HDU1	varchar(32)	DEFAULT NULL	
HDUCLASS2_HDU1	varchar(32)	DEFAULT NULL	
PASS_VER	varchar(255)	DEFAULT NULL	
NDIFRSP	smallint(6)	DEFAULT NULL	
DIFRSP0_HDU1	varchar(32)	DEFAULT NULL	
DIFRSP1_HDU1	varchar(32)	DEFAULT NULL	
DIFRSP2_HDU1	varchar(32)	DEFAULT NULL	
DIFRSP3_HDU1	varchar(32)	DEFAULT NULL	
DIFRSP4_HDU1	varchar(32)	DEFAULT NULL	
DSTYP1_HDU1	varchar(32)	DEFAULT NULL	
DSUNI1_HDU1	varchar(16)	DEFAULT NULL	
DSVAL1_HDU1	varchar(32)	DEFAULT NULL	
DSREF1_HDU1	varchar(32)	DEFAULT NULL	
DSTYP2_HDU1	varchar(32)	DEFAULT NULL	
DSUNI2_HDU1	varchar(16)	DEFAULT NULL	
DSVAL2_HDU1	varchar(32)	DEFAULT NULL	
NDSKEYS	smallint(6)	DEFAULT NULL	
NAXIS_HDU2	smallint(6)	DEFAULT NULL	
NAXIS1_HDU2	smallint(6)	DEFAULT NULL	
NAXIS2_HDU2	smallint(6)	DEFAULT NULL	
PCOUNT_HDU2	smallint(6)	DEFAULT NULL	
GCOUNT_HDU2	smallint(6)	DEFAULT NULL	
CHECKSUM_HDU2	varchar(32)	DEFAULT NULL	
DATASUM_HDU2	varchar(32)	DEFAULT NULL	
EXTNAME_HDU2	varchar(32)	DEFAULT NULL	
HDUCLASS_HDU2	varchar(32)	DEFAULT NULL	
HDUCLASS1_HDU2	varchar(32)	DEFAULT NULL	
HDUCLASS2_HDU2	varchar(32)	DEFAULT NULL	
ONTIME	float	DEFAULT NULL	
TELAPSE	float	DEFAULT NULL	
EXTVER_HDU2	smallint(6)	DEFAULT NULL	
WEEK	smallint(3)	NOT NULL	
update_time	timestamp	NOT NULL DEFAULT CURRENT_TIMESTAMP	

spacecraft

id	mediumint(9)	NOT NULL AUTO INCREMENT	PK
did_rse	varchar(255)	DEFAULT NULL	
did_scope	varchar(255)	DEFAULT NULL	
did_name	varchar(255)	DEFAULT NULL	
checksum	varchar(32)	DEFAULT NULL	
file_version	smallint(5)	NOT NULL	
file_name	varchar(255)	NOT NULL	
file_extension	varchar(255)	NOT NULL	
NAXIS_HDU0	smallint(6)	DEFAULT NULL	
EXTEND	tinyint(1)	DEFAULT NULL	
CHECKSUM_HDU0	varchar(32)	DEFAULT NULL	
DATASUM_HDU0	varchar(32)	DEFAULT NULL	
TELESCOP	varchar(8)	DEFAULT NULL	
INSTRUME	varchar(8)	DEFAULT NULL	
EQUINOX	float	DEFAULT NULL	
RADECSYS	varchar(8)	DEFAULT NULL	
DATE	timestamp	DEFAULT NULL	
DATE_OBS	timestamp	DEFAULT NULL	
DATE_END	timestamp	DEFAULT NULL	
TSTART	float	DEFAULT NULL	
TSTOP	float	DEFAULT NULL	
TIMESYS	varchar(8)	DEFAULT NULL	
TIMEUNIT	varchar(8)	DEFAULT NULL	
GPS_OUT	tinyint(1)	DEFAULT NULL	
MJDREFI	float	DEFAULT NULL	
MJDREFF	float	DEFAULT NULL	
OBSERVER	varchar(255)	DEFAULT NULL	
FILENAME	varchar(255)	DEFAULT NULL	
ORIGIN	varchar(32)	DEFAULT NULL	
CREATOR	varchar(255)	DEFAULT NULL	
VERSION	varchar(8)	DEFAULT NULL	
PROC_VER	smallint(6)	DEFAULT NULL	
NAXIS_HDU1	smallint(6)	DEFAULT NULL	
NAXIS1_HDU1	smallint(6)	DEFAULT NULL	
NAXIS2_HDU1	smallint(6)	DEFAULT NULL	
PCOUNT_HDU1	smallint(6)	DEFAULT NULL	
GCOUNT_HDU1	smallint(6)	DEFAULT NULL	
CHECKSUM_HDU1	varchar(32)	DEFAULT NULL	
DATASUM_HDU1	varchar(32)	DEFAULT NULL	
EXTNAME_HDU1	varchar(32)	DEFAULT NULL	
TIMEZERO	float	DEFAULT NULL	
TIMEREF	varchar(8)	DEFAULT NULL	
TASSIGN	varchar(16)	DEFAULT NULL	
CLOCKAPP	tinyint(1)	DEFAULT NULL	
EXTVER	smallint(6)	DEFAULT NULL	
WEEK	smallint(3)	NOT NULL	
update_time	timestamp	NOT NULL DEFAULT CURRENT_TIMESTAMP	

Gaia Use Case

- Physical archive structure divided into three tables
- One table for the observed sources
- One table for the observed transits
- One table to link them all
- Minor adjustments on the transits table to provide better search capabilities



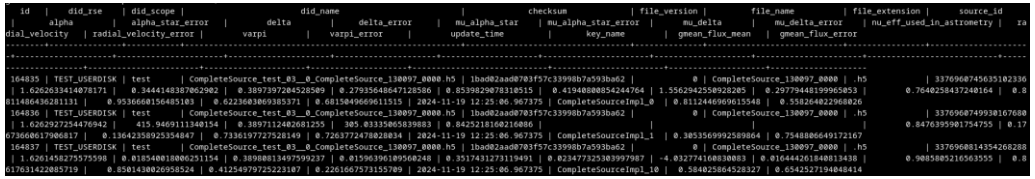
Data Ingestion

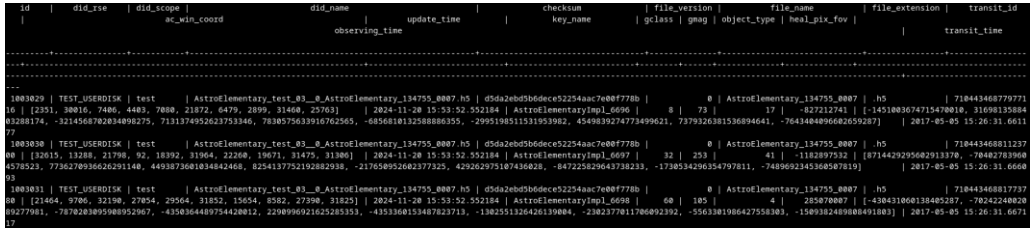
- Minor adjustment to the extracting method to mirror the changes on the file structure and the consequent changes on the archiving structure

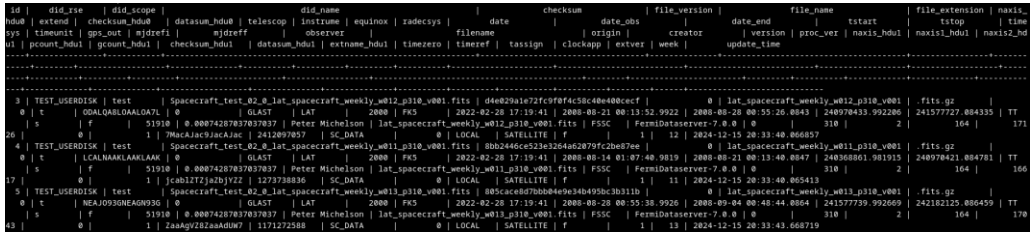
- Data ingestion completed for the provided Gaia datasets

- More than 800k entries in the Gaia Database

- Data ingestion mostly completed also for Fermi



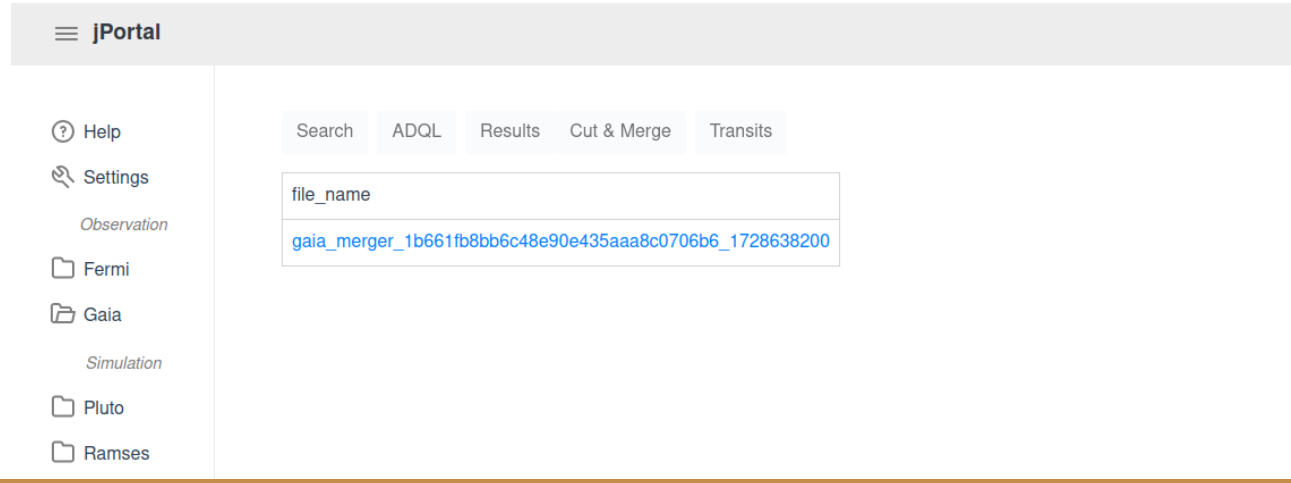




GaiaMerger

- «Cut & Merge» Service
- Generates new temporary files
- For only sources or transit direct copy of the single groupes inside the newly generated file
- For combination of sources and transits implemented file structure so that each transit resides with its related source within one single HDF5 group
- Provides smaller data files to the users with only their selected sources and/or transits
- Provides easy access to each single source to be studied

```
match mergetype:
    case 'sources':
        for ind in range(len(files)):
            with h5.File(files[ind]['path']) as infile:
                if int(files[ind]['sourceId']) == infile[files[ind]['keyName']].attrs['SourceId']:
                    with h5.File(outPath,'a') as ofile:
                        infile.copy(infile[files[ind]['keyName']],ofile,name=f'CompleteSource_{files[ind]["sourceId"]}')
    case 'transits':
        for ind in range(len(files)):
            with h5.File(files[ind]['path']) as infile:
                if int(files[ind]['transitId']) == infile[files[ind]['keyName']].attrs['TransitId']:
                    with h5.File(outPath,'a') as ofile:
                        infile.copy(infile[files[ind]['keyName']],ofile,name=f'AstroElementary_{files[ind]["transitId"]}')
    case 'both':
        sources = []
        transits = []
        for ind in range(len(files)):
            if 'transitId' in list(files[ind].keys()):
                transits.append(files[ind])
            else:
                sources.append(files[ind])
        for ind in range(len(sources)):
            with h5.File(sources[ind]['path']) as infile:
                if int(sources[ind]['sourceId']) == infile[sources[ind]['keyName']].attrs['SourceId']:
                    with h5.File(outPath,'a') as ofile:
                        ofile.create_group(f"Merged_{sources[ind]['sourceId']}")
                        infile.copy(infile[sources[ind]['keyName']],ofile[f"Merged_{sources[ind]['sourceId']}"],name=f'CompleteSource_{sources[ind]["sourceId"]}
        for ind in range(len(transits)):
            with h5.File(transits[ind]['path']) as infile:
                if int(transits[ind]['transitId']) == infile[transits[ind]['keyName']].attrs['TransitId']:
                    with h5.File(outPath,'a') as ofile:
                        infile.copy(infile[transits[ind]['keyName']],ofile[f'Merged_{transits[ind]["sourceId"]}'],name=f'AstroElementary_{transits[ind]["transit
```



Conclusions

- Observative data are easily stored in the constructed archiving system
- Good communication with data providers easily solves fine tuning problems
- Data ingestion is fairly achievable even with few resources
- The observative archive is «search-ready» for the developed portal

Next Steps

- Simulative Logical Data Models are next in line to be prepared
- Simulative Physical DMs still requires the knowledge on the searchable metadata to be stored
- TAP and Datalink implementation to provide access even to standardized clients
- Ingestion of a larger second set of Gaia data
- Implementation of other analysis services, as the connection to the Fermitools

Spoke 3 II Technical Workshop, Bologna Dec 17 -19, 2024