

Finanziato dall'Unione europea NextGenerationEU







# Spoke 3 Archive Infrastructure Massimo Costantini - INAF - OATs

Spoke 3 III Technical Workshop, Perugia 26-29 Maggio, 2025

ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









The High-Performance Computing, Big Data, and Quantum Computing Research Center, established and managed by ICSC, is one of five National Centers created under the Italian National Recovery and Resilience Plan (PNRR) and funded by the European Union.



#### SPOKE 3 ASTROPHYSICS & COSMOS OBSERVATIONS

ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









The main goals of the Spoke 3 project are to leverage state-of-the-art solutions in High-Performance Computing (HPC) and Big Data processing and analysis, to address challenges in the fields of Astrophysics and Cosmic Observation.



#### SPOKE 3 ASTROPHYSICS & COSMOS OBSERVATIONS

ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









Spoke 3 is structured into several Work Packages (WPs) and this presentation focuses on WP4, which is dedicated to challenges related to Big Data Management, Storage, and Archiving.



#### SPOKE 3 ASTROPHYSICS & COSMOS OBSERVATIONS

ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









This "actors/actions" schema illustrates the aggregation of multiple data sources, including both observational and simulation-based data.

There are two main data workflows: Ingestion and Extraction.











The Ingestion workflow starts by separating metadata from raw data by the "Importer", a software component implemented as a device and also monitored within the TANGO Controls distributed control system framework.











The initialized "Importer" instance retrieves the required data description model from a local MariaDB database and, using mapping tables, extracts key-value pairs through functions specifically developed for the various data formats used in our use cases.











Metadata is written to a PostgreSQL database while the raw data is managed by a Rucio infrastructure for data storage and replication using a MinIO-based S3 object storage backend. At this point, the Extraction workflow can begin.











A custom portal allows astronomers to create queries to search for specific metadata and download only the files they are interested in. The portal also supports operations like Cut & Merge (in the Gaia use case) and provides access to Fermi Tools (in the Fermi use case).













Through a form-based interface, users can, for example, select an objectName and retrieve the corresponding alpha, delta, and epsilon values, or select other metadata to generate a query. An integrated editor also allows users to write and modify queries. [1]

$\equiv$ Spoke 3 Por	tal		
⑦ Help 왕 Settings	Resolve Edit	Search Cut & Merge	Reset
Observation			
🗅 Fermi	Form		
🗁 Gaia	objectName	-	sourceld
Simulation	M32		
C Pluto	alpha	delta	
Ramses			
	epsilon		
	alpha		alphaStarError
	Min: 1.61364	Max: 1.63109	Min: 0.0119842 Max: 971.737

ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









Through a form-based interface, users can, for example, select an objectName and retrieve the corresponding alpha, delta, and epsilon values, or select other metadata to generate a query. An integrated editor also allows users to write and modify queries. [2]

$\equiv$ Spoke 3 Porta	al							
Help	Resolve Edit	Search Cut & Merge	Reset					
Not Settings	completesource -							
Observation								
🗅 Fermi	Form							
🔁 Gaia	alpha		alphaStarE	alphaStarError				
Simulation	1.61815	1.61815 1.61835		19842 Max: 971.737				
D Pluto	delta		deltaError					
C Ramses	0.40185	0.40205	Min: 0.01	01086 Max: 734.994				
	muAlphaStar		muAlphaStarError					
	Min: -280.5	Max: 287.5	Min: 0.0	Max: 206.8				
	muDelta		muDeltaEr	muDeltaError				
	Min: -148.85	Max: 200.139	Min: 0.0	Max: 206.1				

#### ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing











Through a form-based interface, users can, for example, select an objectName and retrieve the corresponding alpha, delta, and epsilon values, or select other metadata to generate a query. An integrated editor also allows users to write and modify queries. [3]

$\equiv$ Spoke 3 Portal	
<ul> <li>Help</li> <li>Settings</li> <li>Observation</li> <li>Fermi</li> <li>Gaia</li> <li>Simulation</li> <li>Pluto</li> <li>Ramses</li> </ul>	Resolve       Edit       Search       Cut & Merge       Reset         Editor       Normal ♥       B I U ♥       = = = = = ??        >?        > = = = = A ※ ● II I I       II I









In the Gaia use case, the query returns two tables. The first (Sources) is actually the result of merging two queries: one to PostgreSQL to retrieve metadata, including the Rucio Data Identifier (DID), and one to Rucio to locate the corresponding data file. The second table contains Transits.

Sourc	es					
	file_name	source_id	alpha	alpha_star_error	delta	
	CompleteSource_130097_0003	3425114617586286848	1.618320372472243	0.0252838789312494	0.40193937806911	
	CompleteSource_130097_0003	3425114617586288128	1.6182774194347807	0.08808795660406121	0.4018892840683867	
	CompleteSource_130097_0003	3425114617587515008	1.6182973429850582	2.864232984901267	0.4019187701892775	
	CompleteSource_130097_0003	3425114823742974720	1.6182802384704114	20.517780727768972	0.4020303469978408	
CompleteSource_130097_0003		3425114823745120512	1.6182794166233065	0.534499046224391	0.4020499999301908	
Fransi	its					
	file_name	source_id	transit_id	ac_win_coord	transit_time	
	AstroElementary_134755_0006	3425114617586286848	71044382629193047	[ 6008, 10765, 7849, 23375, 5877, 2960, 16268, 27836, 28160, 4277 ]	2017-05-05 15:27:27.522425	
-	AstroElementary_134756_0006	3425114617586286848	71048474888300196	[ 9139, 14738, 21180, 27859, 3839, 32384, 20020, 18213, 13261, 19935 ]	2017-05-05 17:14:01.677219	
	AstroElementary_134757_0006	3425114617586286848	71058215738283403	[ 32345, 14328, 19099, 11576, 21558, 31161,	2017-05-05	
lessa	ages					
earc	h started: 9:16:48 AM					
otal	PostgreSQL time: 145.00 ms					

Merge time: 514.00 ms Render time: 4.00 ms

Total time: 1147.00 ms

ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









# At this point, users can select one or more rows and perform a Cut & Merge operation, producing a significantly smaller file to download. [1]

	file_name	source_id	alpha	alpha_star_error	delta
-	CompleteSource_130097_0003	3425114617586286848	1.618320372472243	0.0252838789312494	0.40193937806911
	CompleteSource_130097_0003	3425114617586288128	1.6182774194347807	0.08808795660406121	0.4018892840683867
-	CompleteSource_130097_0003	3425114617587515008	1.6182973429850582	2.864232984901267	0.4019187701892775
	CompleteSource_130097_0003	3425114823742974720	1.6182802384704114	20.517780727768972	0.4020303469978408
	CompleteSource_130097_0003	3425114823745120512	1.6182794166233065	0.534499046224391	0.4020499999301908
rans	its				
-	file_name	source_id	transit_id	ac_win_coord	transit_time
2	AstroElementary_134755_0006	3425114617586286848	71044382629193047	[ 6008, 10765, 7849, 23375, 5877, 2960, 16268, 27836, 28160, 4277 ]	2017-05-05 15:27:27.522425
⊐.	AstroElementary_134756_0006	3425114617586286848	71048474888300196	[ 9139, 14738, 21180, 27859, 3839, 32384, 20020, 18213, 13261, 19935 ]	2017-05-05 17:14:01.677219
_	AstroElementary_134757_0006	3425114617586286848	71058215738283403	[ 32345, 14328, 19099, 11576, 21558, 31161, 19624, 26578, 20740, 5042	2017-05-05
loca	ages				

Total Rucio time: 580.00 ms Merge time: 514.00 ms

Render time: 4.00 ms Total time: 1147.00 ms

#### ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









At this point, users can select one or more rows and perform a Cut & Merge operation, producing a significantly smaller file to download. [2]

Cut & Merged	
file_name	
gaia_merger_4390e96bc45c19	1c225133de0a2e7792_1747639922.27298











Gaia data is not publicly accessible, so users must log into the Spoke 3 Portal via the RAP (Remote Authentication Portal) to obtain the necessary authentication and authorization credentials.











Additional features of the Spoke 3 Portal include:

- the ability to automatically correct the selection of transits that do not match any source;
- comprehensive logging of all operations (queries and downloads), collected using Elasticsearch and visualized through Kibana.









Key Points So Far:

- aggregations of both observational and simulation-based data through Ingestion and Extraction workflows;
- metadata is stored in PostgreSQL; raw data is managed by Rucio over a MinIO-based S3 object storage backend;
- a custom portal allows querying metadata and downloading only relevant files, supporting operations like Cut & Merge (for Gaia) or Fermi Tools access;
- users interact with the Spoke 3 Portal via forms or query editor;
- authentication via RAP is required to access private Gaia data;
- features include automatic correction of unmatched transits and full operation logging with Elasticsearch/Kibana.









## But what if the astronomer doesn't like my portal...?



#### ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









**IVOA (International Virtual Observatory Alliance):** 

- one of the microservices used by the Spoke 3 Portal is a technically compliant TAP (Table Access Protocol) exposing a DataLink table to enable all the custom services available for the query results;
- this allows users to perform the same operations previously described, but using other clients, like TOPCAT.









## By appending the /tap endpoint to the Spoke 3 Portal URL, users can access the TAP schema and the PostgreSQL tables directly from TOPCAT. The integrated editor also allows query writing and editing. [1]

	Table Access Protocol (TAP) Query	- o ×
Indow IAP Authentication Registry Edit Interop Help		
Select Service Use Service Resume Job Running Jobs		
Locate TAP Service		
By Table Properties By Service Properties		
Keywords:		And
Match Fields: 🕑 Table Name 🕑 Table Description 🕑 Service		Cancel Find Services
<ul> <li>TAPVizieR (60284) - ivo://cds.vizier/tap</li> <li>TAPVizieR (60284) - ivo://cds.vizier/tap</li> <li>Data Lab TAP (2481) - ivo://clas.aheasarc/services/xamin</li> <li>TEASATR (1028) - ivo://masa.heasarc/services/xamin</li> <li>TESA TAP (974) - ivo://masa.heasarc/services/xamin</li> <li>GAIA (248) - ivo://main-heidelberg.de/gaia/tap</li> <li>GAIA (248) - ivo://esavo/esasky/tap</li> <li>TESA TAP_CAT (112) - ivo://esa.org/tap_cat</li> <li>ESO TAP_CAT (112) - ivo://eso.org/tap_cat</li> <li>StyMapper TAP (96) - ivo://archive.stsci.edu/tap/gaiadr3</li> <li>StyMapper TAP (96) - ivo://archive.stsci.edu/tap/lajaddr3</li> <li>StyMapper TAP (96) - ivo://archive.stsci.edu/tap/lajaddr3</li> <li>YouCat (71) - ivo://archive.stsci.edu/ps1dr1tap</li> <li>YouCat (71) - ivo://archive.stsci.edu/ps1dr1tap</li> <li>PS1 DR1 TAP (69) - ivo://archive.stsci.edu/ps1dr1tap</li> <li>Euclid (66) - ivo://archive.stsci.edu/ps1dr1tap</li> <li>Euclid (66) - ivo://esavo/tesavo/tap_obs</li> <li>Herschel (53) - ivo://esavo/tesavo/tap</li> <li>Hotsber (747) (49) - ivo://esavo/tap</li> <li>Hubble/TAP (49) - ivo://esavo/tap</li> <li>Hubble/TAP (49) - ivo://esavo/tap</li> </ul>	ww.plate-archive.org/tap	
<ul> <li>PSA (47) - ivo://esavo/psa/epntap</li> <li>ASTRON VO TAP (45) - ivo://astron.nl/tap</li> <li>J-PAS-EDR (44) - ivo://cefca/j-pas/j-pas-edr</li> <li>MINJPASPDR201912 (42) - ivo://cefca/minijpas/minij-pas-pdr201912</li> <li>PAD F GAVO TAP (41) - ivo://cafca/o.org/tap</li> <li>PADC TAP (41) - ivo://padc.obspm.planeto/tap</li> <li>J-PLUS-DR3 (40) - ivo://cefca/j-plus/gr3</li> </ul>		-
Selected TAP Service		
TAP URL: https://131.154.99.146.myip.cloud.infn.it/tap		<b>.</b>
		Use Service

#### ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









## By appending the /tap endpoint to the Spoke 3 Portal URL, users can access the TAP schema and the PostgreSQL tables directly from TOPCAT. The integrated editor also allows query writing and editing. [2]

					Table Access Prot	ocol (TAP) Query				- o ×
Mindow IAP Authentication Reg	gistry <u>E</u> dit <u>I</u> nterop <u>F</u>	<u>H</u> elp								
Select Service Use Service R	esume Job Running Jo	obs								
Find: Or	O Service @ ADQL	O Schema O Ta	ble O Columns C	FKeys Hints						
Name Descrip     Sort:      Sort:      Alphabetic     TAP Service ○ Alphabetic     TAP Setvice ○ Alphabetic     TaP_schema.(10)     TAP_schema.key_colum     田 gaia.astroelementary     田 gaia.completesource     田 gaia.crossmatch	Name	]Ţy	уре	Unit	Indexed	Description	UCD	Utype	Xtype	Flags
Service Capabilities Query Language: ADQL-2.0 V N ADQL Text	Aax Rows:	Uploads:	unavailable							Log In/Out Z
Query Mode: Synchronous									• 44	
SELECT * FROM gaia.completesource WHERE alpha < 1.61835 AND alp	ha > 1.61815 AND del	ta < 0.40205 AND d	delta > 0.40185;							
Examples										Info 년

#### ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









In the result table, it is possible to select one or more rows, exactly as previously described with the Spoke 3 Portal, and create a subset.

Tabl	e Browser for 1: TAP_1_gaia.con	npletesource							2	a					23	
	file_name	source_id	alpha	alpha_star_error	delta	delta_error	mu_alpha_star	mu_alpha_star_error	mu_delta	mu_delta_error	nu_eff_used_in_astrometry	radial_velocity	radial_velocity_error	varpi	varpi_error	access_url
1	CompleteSource_130097_0003	3425114617586286848	1.61832	0.02528	0.40194	0.02384	0.64278	0.02619	-4.59132	0.02006	0.45894	0.09275	0.01782	0.1542	0.92586	https://131.154.99.146.myip.cloud.infn.it/d
2	CompleteSource_130097_0003	3425114617586288128	1.61828	0.08809	0.40189	0.09428	0.9021	0.09593	-0.88858	0.07311	0.30315	0.84391	0.90766	0.98055	0.15312	https://131.154.99.146.myip.cloud.infn.it/d
3	CompleteSource_130097_0003	3425114617587515008	1.6183	2.86423	0.40192	2.50148	0.48707	4.89259	5.54169	6.53078	0.26648	0.7674	0.56905	0.81822	0.0022	https://131.154.99.146.myip.cloud.infn.it/d
4	CompleteSource_130097_0003	3425114823742974720	1.61828	20.51778	0.40203	18.78652	0.16148				0.09746	0.28924	0.61654	0.0078	0.94949	https://131.154.99.146.myip.cloud.infn.it/d
5	CompleteSource_130097_0003	3425114823745120512	1.61828	0.5345	0.40205	0.6226	0.89111	0.59489	-0.23023	0.4512	0.28749	0.50805	0.17135	0.91088	0.50136	https://131.154.99.146.myip.cloud.infn.it/d
6	CompleteSource_130097_0003	3425114823745878272	1.61832	718.11256	0.40205	696.69769	0.75758				0.2297	0.41527	0.03135	0.06855	0.44425	https://131.154.99.146.myip.cloud.infn.it/d
7	CompleteSource_130097_0003	3425114819446323584	1.61825	0.0889	0.40195	0.09096	0.03078	0.09331	-0.90564	0.07075	0.38733	0.76744	0.54766	0.48679	0.03204	https://131.154.99.146.myip.cloud.infn.it/d
8	CompleteSource_130097_0003	3425114819446325376	1.61818	0.60828	0.40197	0.61666	0.90372	0.54045	-3.21546	0.47557	0.07449	0.39592	0.38584	0.43062	0.53997	https://131.154.99.146.myip.cloud.infn.it/d

	New Subset
?	New Subset Name: 3 rows
	Add Subset
	Add and Set Current Subset

ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









It is worth noting the column named access\_url, which is required by TOPCAT to invoke the DataLink. This column embeds two UUIDs (Universally Unique Identifiers): the first is used to store the selected rows and queries, the second to create a personal cache for each user accessing TOPCAT. Both operations are managed by Redis (the NoSQL in-memory database).

access_url
https://131.154.99.146.myip.cloud.infn.it/datalink?id=e7d8f16d-b7dc-4fc2-bd22-df0b3d1efd83&clientId=d8a25848-e8a8-468c-a1db-c5c0ae356802
https://131.154.99.146.myip.cloud.infn.it/datalink?id=cb372fee-d99d-4f24-be5c-2b314fc2e595&clientId=d8a25848-e8a8-468c-a1db-c5c0ae356802
https://131.154.99.146.myip.cloud.infn.it/datalink?id=1e8b4111-a35b-495d-88d0-5ba65992e4ab&clientId=d8a25848-e8a8-468c-a1db-c5c0ae356802









## In the Activation Actions menu, it is necessary to select "View Datalink Table"...

	TOPCAT(1): Activation Actions	– ø ×
<u>W</u> indow <u>A</u> ctions <u>H</u> elp		
+ m m 4 %		
Delay Display Cutout Image Display HiPS cutout Display image	Completesource Update other TOPCAT windows (e.g. Cone Search) with sky coordinates	
Display image region Download URL Execute code Invoke Datalink Row	RA Column:     radial_velocity <ul> <li>degree</li> <li>degree</li> <li>degree</li> <li>degree</li> </ul>	
Load Table Plot Table Run system command		
SAMP Message Send FITS Image Send HiPS cutout Send Sky Coordinates	- Status	
Send Spectrum Send VOTable Send row index		
Use Sky Coordinates in TOPCAT View Datalink Table View in Web Browser	View a referenced DataLink table	





ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing





Missione 4 • Istruzione e Ricerca

### ...and finally, invoke the DataLink.

	TOPCAT(1): Activation Actions	- ø ×
<u>W</u> indow <u>A</u> ctions <u>H</u> elp		
<b>₽</b> 🖻 🖻 👂 🧏		
Activation Actions for 1: TAP_1_gai	Perform the currently selected action on every row in the current subset in turn	
Actions	Description	
🛊 🗔 Use Sky Coordinates in T	View a referenced DataLink table	
🛢 🔲 Send Sky Coordinates	Configuration	
🛢 🔲 Display HiPS cutout	Datalink Table URL	
🛢 🔲 Send HiPS cutout	Linke Table Location: access ud	
🛊 🗌 Display image		
🛊 🔲 Display image region		
🛊 🗔 Load Table		
韋 🗔 Plot Table		
🛢 🔤 Send FITS Image	Datalink ID:	
🛢 🔲 Send Spectrum		
Download URL		
🛊 🗌 View in Web Browser	r Status	
🛊 🗌 Delay 🔍	Invoke now on row 1	
Results		
Seg Row Status Message		









By selecting #this, a Cut & Merge operation can be performed, exactly as previously described with the Spoke 3 Portal. Since the content\_type is HDF5, the file will be downloaded directly through the browser....

	ТОР	CAT(1): Activation - View Datalink Table	— <b>a</b> ×
DataLink Table			
semantics	description	content_type	access_url
1 #this	Cut & Merge (extracted from the progenitor dataset)	application/x-hdf5	https://131.154.99.146.myip.cloud.infn.it/datalink/gaiamerger?id=e116d238-d45a-450e
2 #progenitor	Original progenitor dataset	application/x-hdf5	https://minio.131.154.99.166.myip.cloud.infn.it:443/spoke3/test/ee/b6/CompleteSource
3 #counterpart	Link to Astroelementary entries	application/x-votable+xml	https://131.154.99.146.myip.cloud.infn.it/datalink/transits?id=e7d8f16d-b7dc-4fc2-bd22
Row Link Type			
Fixed Access URL			
- Row Detail			
Access URL: https://131.154.99.146.myip.cloud.infn.it/datali	nk/gaiamerger?id=e116d238-d45a-450e-a75c-af14227d4292&clientId=d8a258	48-e8a8-468c-a1db-c5c0ae356802	
Content Type: application/x-hdf5			
Content Length:			
Description: Cut & Merge (extracted from the progenitor da	ataset)		
Semantics: #this			
URL: https://131.154.99.146.myip.cloud.infn.it/datalink/gaiamerg	ger?id=e116d238-d45a-450e-a75c-af14227d4292&clientId=d8a25848-e8a8-46	8c-a1db-c5c0ae356802	
Type: UNKNOWN 🔻 Guess 🖌	Action: Show web page		Auto-Invoke Invoke
Result:			









– ō ×

## ... if #progenitor is selected, the original file containing all sources or all transits will be downloaded...

TOPCAT(1): Activation - View Datalink Table

DataLink Table			
semantics	description	content_type	access_url
1 #this	Cut & Merge (extracted from the progenitor dataset)	application/x-hdf5	https://131.154.99.146.myip.cloud.infn.it/datalink/gaiamerger?id=e116d238-d45a-450e
2 #progenitor	Original progenitor dataset	application/x-hdf5	https://minio.131.154.99.166.myip.cloud.infn.it:443/spoke3/test/ee/b6/CompleteSource
3 #counterpart	Link to Astroelementary entries	application/x-votable+xml	https://131.154.99.146.myip.cloud.infn.it/datalink/transits?id=e7d8f16d-b7dc-4fc2-bd22
Row Link Type			
Fixed Access URL			
Row Detail			
Access URL: https://minio.131.154.99.166.myip.	cloud.infn.it:443/spoke3/test/ee/b6/CompleteSource_test_2025-02-16_0_CompleteSource_	130097_0003?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=spok	e3%2F20250519%2Fus-west-2%2Fs3%2Faws4_request&X-Amz-Date=20250519T090902Z&X-Amz-Expires=
Content Type: application/x-hdf5			
Content Length:			
Description: Original progenitor dataset			
Semantics: #progenitor			
URL: https://minio.131.154.99.166.myip.cloud.infn.it.	443/spoke3/test/ee/b6/CompleteSource_test_2025-02-16_0_CompleteSource_130097_000	3?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=spoke3%2F2025	0519%2Fus-west-2%2Fs3%2Faws4_request&X-Amz-Date=20250519T090902Z&X-Amz-Expires=3600&X-Am
Type: UNKNOWN 🔻 Guess 🖌	Action: Show web page		Auto-Invoke Invoke
Result:			









...and finally, if #counterpart is selected, the TAP is accessed again to retrieve the Transits table, and then the entire process can be repeated: selecting one or more rows, creating a subset, and invoking the DataLink to either download the original file or perform a new Cut & Merge on the selected rows.

	TOP	PCAT(1): Activation - View Datalink Table	×
DataLink Table			
semantics	description	content_type	access_url
1 #this	Cut & Merge (extracted from the progenitor dataset)	application/x-hdf5	https://131.154.99.146.myip.cloud.infn.it/datalink/gaiamerger?id=e116d238-d45a-450e
2 #progenitor	Original progenitor dataset	application/x-hdf5	https://minio.131.154.99.166.myip.cloud.infn.it:443/spoke3/test/ee/b6/CompleteSource
3 #counterpart	Link to Astroelementary entries	application/x-votable+xml	https://131.154.99.146.myip.cloud.infn.it/datalink/transits?id=e7d8f16d-b7dc-4fc2-bd22
Row Link Type			
Fixed Access URL			
Row Detail			
Access URL: https://131.154.99.146.myip.cloud.infn.it/datalin	k/transits?id=e7d8f16d-b7dc-4fc2-bd22-df0b3d1efd83&clientId=d8a25848-e8	8a8-468c-a1db-c5c0ae356802	
Content Type: application/x-votable+xml			
Content Length:			
Description: Link to Astroelementary entries			
Semantics: #counterpart			
URL: https://131.154.99.146.myip.cloud.infn.it/datalink/transits?id	=e7d8f16d-b7dc-4fc2-bd22-df0b3d1efd83&clientId=d8a25848-e8a8-468c-a1	db-c5c0ae356802	
Type: TABLE 🔽 Guess 🖌	Action: Load Table		Auto-Invoke Invoke
Result:			









Key Points So Far:

- a separated microservice exposes TAP and a DataLink table, usable also via other clients like TOPCAT;
- the /tap endpoint is used to access the TAP schema and PostgreSQL tables;
- the queries return the Sources table with an access\_url column containing two UUIDs, for Redis storage and user cache;
- users can select rows, create subsets, and invoke DataLink directly from TOPCAT;
- #this (to trigger a Cut & Merge), #progenitor (to download the full file), and #counterpart (to
  access the Transits table), are vocabulary terms that allow automated identification of the type
  of action to perform on the described query result;
- the Transits table provides a recursive DataLink solution, enabling the same workflow again.









## And what if the astronomer doesn't like TOPCAT either...?



ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









This cell initializes the Python client by importing required libraries, setting the base URL of the TAP server, and defining the query. It is equivalent to setting the /tap endpoint and compose a query in TOPCAT's ADQL panel.

: # 1. Import lib	raries, define base URL, and set up the ADQL query	
import aighttp		
import pandas a	as pd	
from astropy.io	votable import parse single table	
from io import	BytesIO	
from urllib.par	se import urlparse, parse_qs	
# Base URL of t	the TAP & DataLink service	
BASE_URL = "htt	p://localhost:3000"	
# ADQL query to	o run	
adql_query = ""		
SELECT *		
FROM gaia.compl	etesource	
WHERE alpha < 1	61835 AND alpha > 1.61815	
AND delta < 0	0.40205 AND delta > 0.40185	









This cell queries the /tables endpoint to retrieve available schemas, tables, and columns. The XML response is parsed into a Pandas DataFrame. It is equivalent to TOPCAT's metadata loading when connecting to the TAP service. [1]

2. Fetch available tables from the IAP service (/tab/tables)	
mport xml.etree.ElementTree as ET	
sync def fetch tap tables():	
async with aichttp.ClientSession() as session:	
<pre>async with session.get(f"{BASE_URL}/tap/tables") as resp:</pre>	
<pre>xml_text = await resp.text()</pre>	
# Parse <tableset> XML structure (VOSI format)</tableset>	
<pre>root = ET.fromstring(xml_text)</pre>	
<pre>ns = {"vosi": "http://www.ivoa.net/xml/VOSITables/v1.0"}</pre>	
rows = []	
<pre>for schema in root.findall("vosi:schema", ns):</pre>	
<pre>schema_name = schema.findtext("vosi:name", default="", namespaces=ns)</pre>	
<pre>for table in schema.findall("vosi:table", ns):</pre>	
<pre>table_name = table.findtext("vosi:name", default="", namespaces=ns)</pre>	
<pre>for column in table.findall("vosi:column", ns):</pre>	
<pre>col_name = column.findtext("vosi:name", default="", namespaces=ns)</pre>	
<pre>col_type = column.find("vosi:dataType", ns)</pre>	
<pre>datatype = col_type.text if col_type is not None else ""</pre>	
rows.append({	
"schema": schema_name,	
"table": table_name,	
"column": col_name,	
"datatype": datatype	
3) ({	
return pd.DataFrame(rows)	
Call the function and display the result	
ables = await fetch tap tables()	
(isplay(tables)	

#### ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









This cell queries the /tables endpoint to retrieve available schemas, tables, and columns. The XML response is parsed into a Pandas DataFrame.

It is equivalent to TOPCAT's metadata loading when connecting to the TAP service. [2]

datatype	column	table	schema	
char	arraysize	TAP_schema.columns	TAP_schema	0
char	datatype	TAP_schema.columns	TAP_schema	1
char	column_name	TAP_schema.columns	TAP_schema	2
char	table_name	TAP_schema.columns	TAP_schema	3
char	ucd	TAP_schema.columns	TAP_schema	4
char	did_rse	gaia.crossmatch	gaia	211
int	id	gaia.crossmatch	gaia	212
int	file_version	gaia.crossmatch	gaia	213
int	flags	gaia.crossmatch	gaia	214
int	heal_pix_fov	gaia.crossmatch	gaia	215

216 rows × 4 columns









This cell sends the query to the /sync endpoint and retrieves the result as a VOTable. The response is parsed by Astropy and displayed as a Pandas DataFrame for inspection in the Jupyter notebook. This mirrors executing a query and viewing results in TOPCAT's table viewer. [1]

# 3. Submit an ADQL query using the /tap/sync endpoint and display the result as a DataFrame	
<pre>async def run_adql_query(adql, client_id=None):</pre>	
payload = {	
"QUERY": adql,	
"request": "doQuery"	
}	
<pre>if client_id:</pre>	
<pre>payload["clientId"] = client_id</pre>	
<pre>async with aiohttp.ClientSession() as session:</pre>	
<pre>async with session.post(f"{BASE_URL}/tap/sync", data=payload) as resp:</pre>	
<pre>votable_bytes = await resp.read()</pre>	
<pre>table = parse_single_table(BytesIO(votable_bytes)).to_table()</pre>	
df = table.to pandas()	
return df	
# Run the ADOL query defined earlier in Cell 1	
df = await run addl guery(adgl guery)	
display(df)	









## This cell sends the query to the /sync endpoint and retrieves the result as a VOTable. The response is parsed by Astropy and displayed as a Pandas DataFrame for inspection in the Jupyter notebook. This mirrors executing a query and viewing results in TOPCAT's table viewer. [2]

id	did_rse	did_scope	did_name	checksum	file_version	file_name	file_extension	key_name	
<b>0</b> 39432	TEST_USERDISK	test	CompleteSource_test_2025-02-16_0_CompleteSourc	df04ee8881a1be3d3762ed41fe3551ca	0	CompleteSource_130097_0003	.h5	CompleteSourceImpl_5369	34251 <sup>,</sup>
<b>1</b> 39438	TEST_USERDISK	test	CompleteSource_test_2025-02-16_0_CompleteSourc	df04ee8881a1be3d3762ed41fe3551ca	0	CompleteSource_130097_0003	.h5	CompleteSourceImpl_5370	34251 <sup>,</sup>
<b>2</b> 39441	TEST_USERDISK	test	CompleteSource_test_2025-02-16_0_CompleteSourc	df04ee8881a1be3d3762ed41fe3551ca	0	CompleteSource_130097_0003	.h5	CompleteSourceImpl_5371	34251 <sup>.</sup>
<b>3</b> 39635	TEST_USERDISK	test	CompleteSource_test_2025-02-16_0_CompleteSourc	df04ee8881a1be3d3762ed41fe3551ca	0	CompleteSource_130097_0003	.h5	CompleteSourceImpl_5430	34251 <sup>,</sup>
<b>4</b> 39640	TEST_USERDISK	test	CompleteSource_test_2025-02-16_0_CompleteSourc	df04ee8881a1be3d3762ed41fe3551ca	0	CompleteSource_130097_0003	.h5	CompleteSourceImpl_5431	34251
<b>5</b> 39642	TEST_USERDISK	test	CompleteSource_test_2025-02-16_0_CompleteSourc	df04ee8881a1be3d3762ed41fe3551ca	0	CompleteSource_130097_0003	.h5	CompleteSourceImpl_5432	34251 <sup>.</sup>
<b>6</b> 40112	TEST_USERDISK	test	CompleteSource_test_2025-02-16_0_CompleteSourc	df04ee8881a1be3d3762ed41fe3551ca	0	CompleteSource_130097_0003	.h5	CompleteSourceImpl_5573	34251
<b>7</b> 40118	TEST_USERDISK	test	CompleteSource_test_2025-02-16_0_CompleteSourc	df04ee8881a1be3d3762ed41fe3551ca	0	CompleteSource_130097_0003	.h5	CompleteSourceImpl_5574	342511

#### ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









This cell extracts the access URL from the TAP result row and queries the /datalink endpoint. It retrieves links to #this, #progenitor, and #counterpart replicating TOPCAT's behavior when invoking DataLink on selected table rows. [1]

# 4. Fetch DataLink services for a result row	, <del>,</del> ,
<pre>async def fetch_datalink_services(row_id, client_id):</pre>	
async with aiohttp.ClientSession() as session:	
<pre>async with session.get(f"{BASE_URL}/datalink", params={</pre>	
"id": row_id,	
"clientId": client_id	
<pre>}) as resp:</pre>	
<pre>votable_bytes = await resp.read()</pre>	
<pre>table = parse_single_table(BytesIO(votable_bytes)).to_table()</pre>	
<pre>df = table.to_pandas()</pre>	
# Reorder columns to match standard DataLink order (like in TOPCAT)	
ordered_columns = ["semantics", "description", "content_type", "access_url"]	
<pre>df = df[[col for col in ordered_columns if col in df.columns]]</pre>	
return df	
# Use the first row from previous TAP query to extract id and clientId	
access_url = df.iloc[0]["access_url"]	
parsed_url = parse_qs(urlparse(access_url).query)	
row id = parsed url["id"][0]	
:lient_id = parsed_url["clientId"][0]	
# Fetch and display DataLink result	
Jatalink = await fetch datalink services(row id, client id)	
display(datalink)	









This cell extracts the access URL from the TAP result row and queries the /datalink endpoint. It retrieves links to #this, #progenitor, and #counterpart replicating TOPCAT's behavior when invoking DataLink on selected table rows. [2]

	semantics	description	content_type	access_url
0	#this	Cut & Merge (from progenitor)	application/x-hdf5	http://localhost:3000/datalink/gaiamerger?id=7
1	#progenitor	Original progenitor dataset	application/x-hdf5	http://localhost:3000/api/files?name=CompleteS
2	#counterpart	Link to transits	application/x-votable+xml	http://localhost:3000/datalink/transits?id=799

ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing







Key Points So Far:

• As the entire workflow can be performed with TOPCAT, the same goes for Python with Astropy, which is simply another example of VO-aware client.









### Modern tools. Younger astronomers...!



ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









## Summary:

- the Spoke 3 Archive Infrastructure project ingests and extracts both observational and simulation-based data;
- metadata is stored in PostgreSQL, while raw data is managed using Rucio on a MinIO-based S3 storage backend;
- a custom portal enables advanced querying and custom operations like Cut & Merge (for Gaia) or access to Fermi Tools;
- the same features are exposed via TAP/DataLink microservices, compatible with other clients like TOPCAT;
- Python with Astropy is just another example of a VO-aware client.









What's Next:

- comparative performance testing between Oracle (for Gaia) and PostgreSQL, including high-load scenarios;
- evaluation of sharding strategies for distributing data across multiple databases or servers, improving performance and scalability;
- future extension of the portal to support new astrophysical datasets and use cases;
- furthermore, resources and services described are not registered, data concerning Pluto and Ramses are still being analyzed, ADQL features are not yet implemented.









### Any questions...?



ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing