

Distributed/parallel databases at ESDC

European Space Astronomy Centre, Madrid, Spain

In2p3-ESDC exchange, 06/02/2019

ESA UNCLASSIFIED - For Official Use

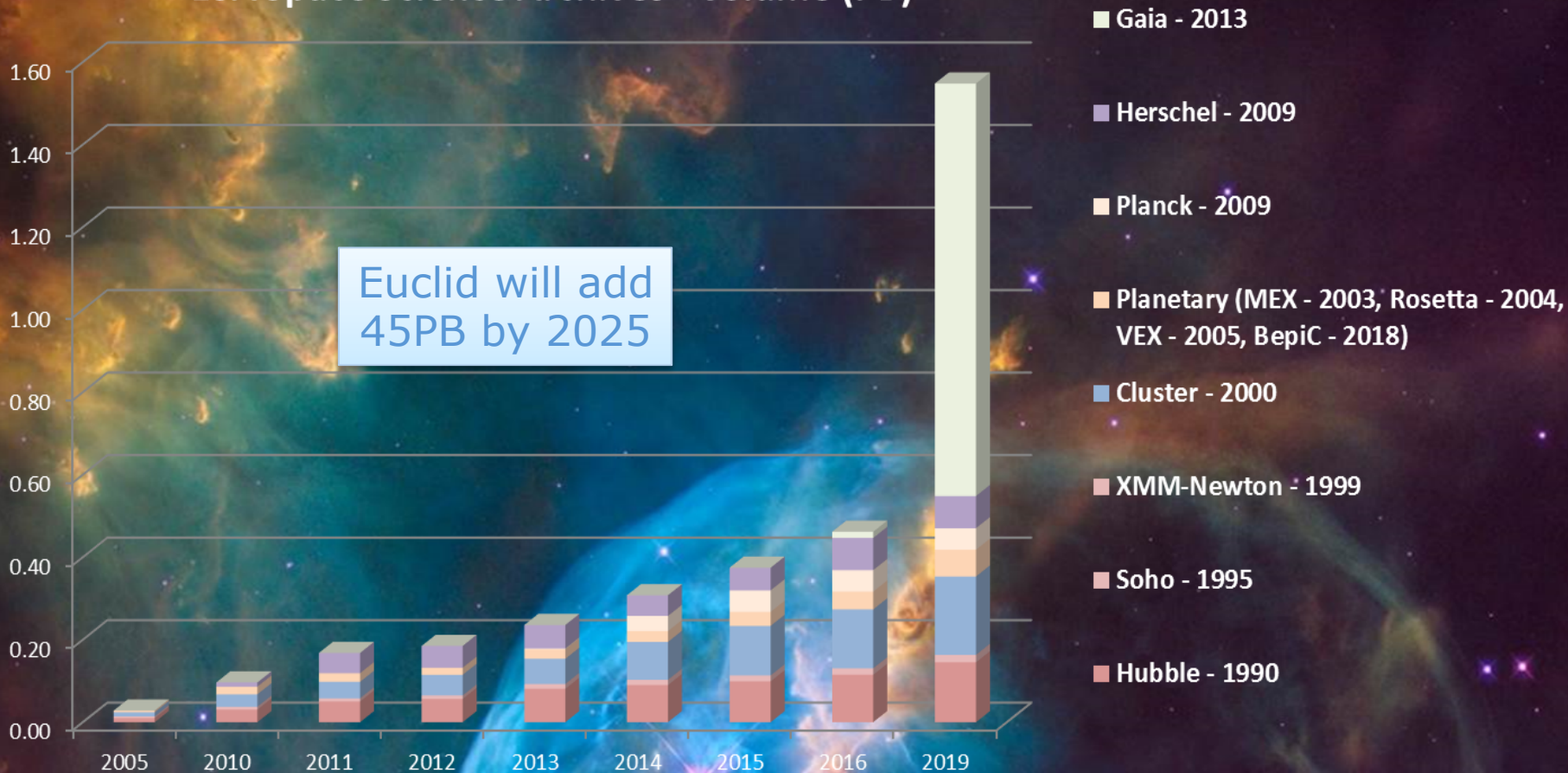


@pteodoro

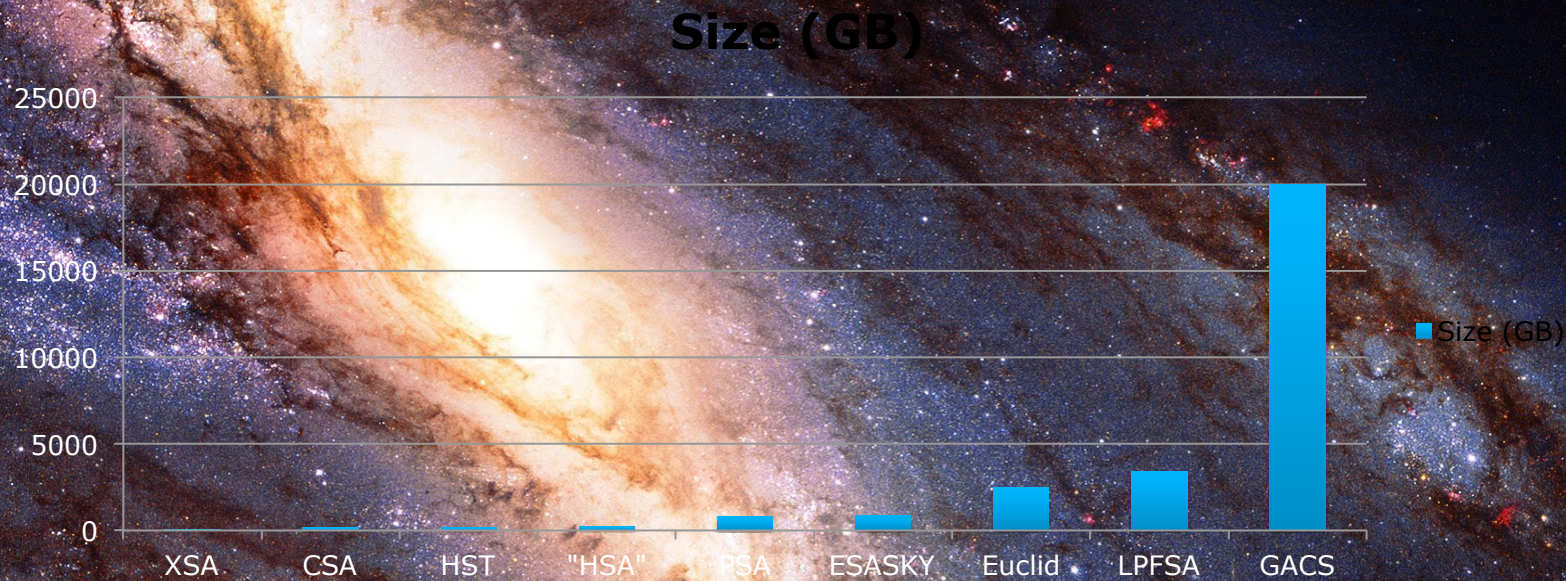


European Space Agency

ESA Space Science Archives - Volume (PB)



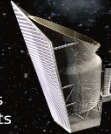
Databases Size at ESDC (January 2019)



→ ESA'S FLEET ACROSS THE SPECTRUM



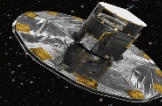
iso
Chemical analysis of celestial objects (1995-1998)



hipparcos
The first astrometry satellite (1989-1993)



gaia
Surveying a billion stars of exoplanets (2013-)



cheops
Sizing and first characterisation



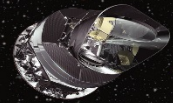
iuue
Analysing ultraviolet light from stars (1978-1996)



xmm-newton
Seeing deeply into the hot and violent Universe (1999-)



herschel
Unveiling the cool and dusty Universe (2009-2013)



planck
Looking back at the dawn of time (2009-2013)



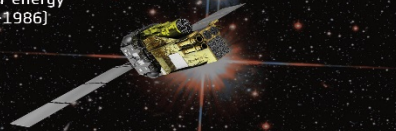
of the visible Universe (1990-)



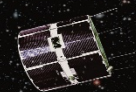
exosat
X-ray survey of high-energy phenomena (1983-1986)



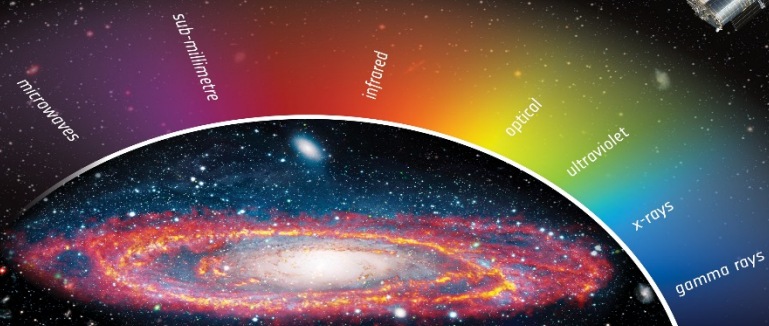
integral
Seeking out the extremes of the Universe (2002-)



cos-b
Surveying the high energy Galaxy (1975-1982)



Databases at ESDC



Database systems at ESDC



Most archives: **PostgreSQL**: 9.4 to 11.1 (15 operational databases)

- Open Source
- Big community
- Spherical queries (pg_sphere, q3c, healpix, postgis)
- Professional companies support (that we do not use)

ISO: **Sybase**

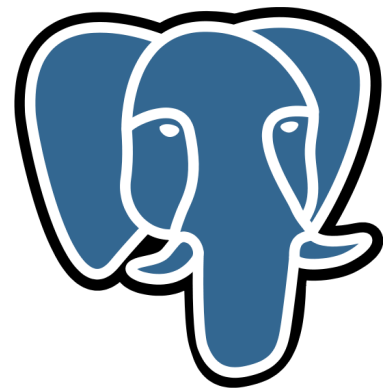
Euclid DPS: **Oracle**



ESASky, PSA (some functionalities): **ElasticSearch**

Euclid Prototype: Spark

Import from Files, MySQL...



elastic

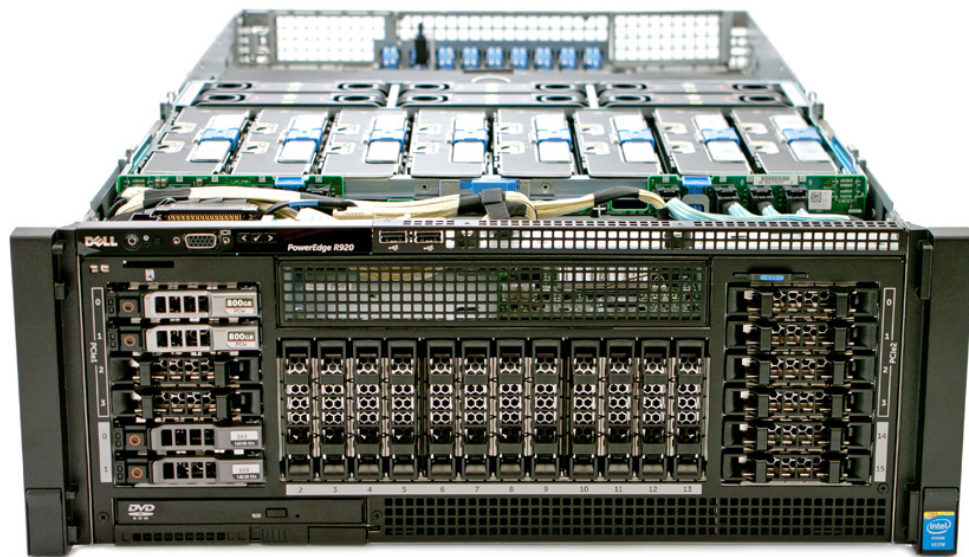
Largest (GACS):

Dell PowerEdge R920 Memory: 1.5TB. 48 cores

7 TB SSD

13 TB SAS

21TB NetApp



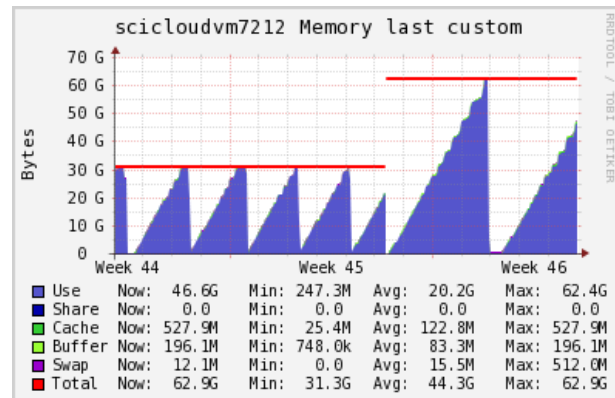
Plan for Scale-out Tests



Objective: Big Datasets queried in efficient times (Requirements).
Ingestion performance.

Pre-Requisites: Resources, data to ingest (Catalogues, Relational tables)

Comparison: ingestion time, retrieving time, administration, problems arisen.



PostgreSQL Scale-out Solutions

Postgres-XL by 2nd Quadrant

- Postgres-10 alpha2 (postgres 10.3)- Euclid
- Postgres-XL 9.5 R1.6 - Gaia



Citus by CitusData

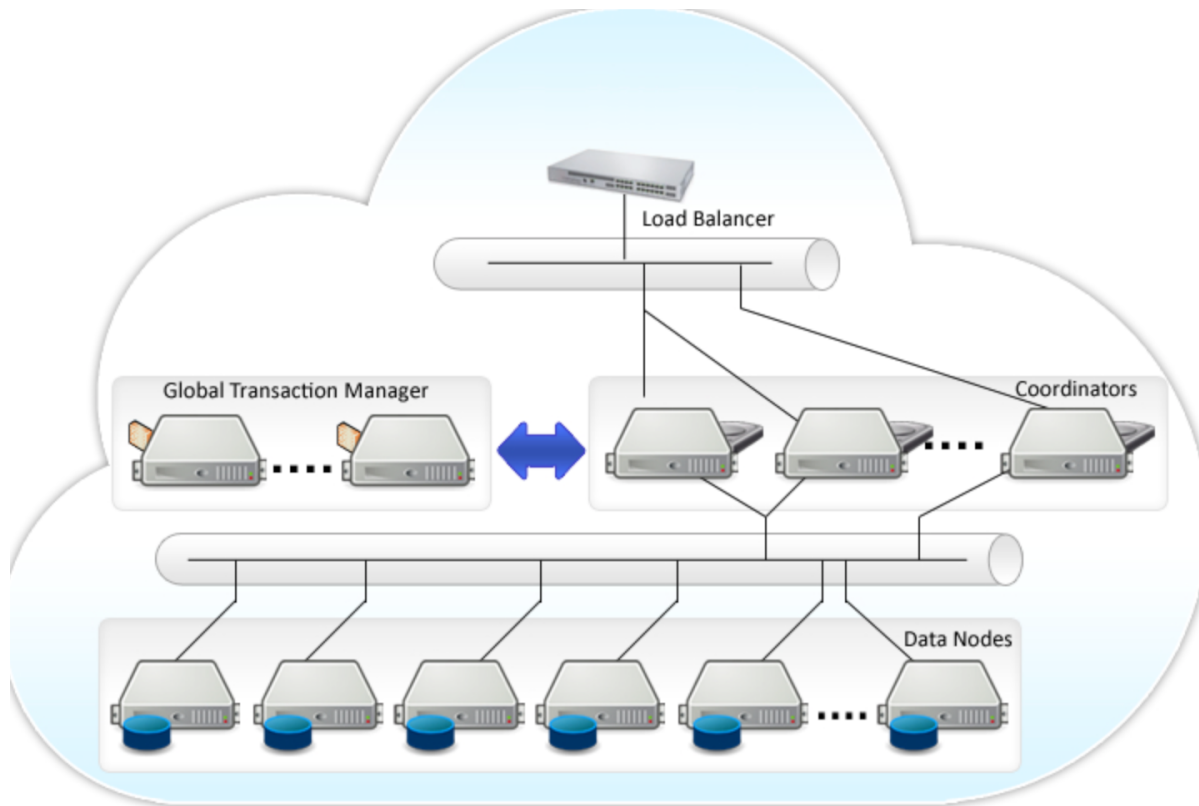
- Citus 7, extension installed with Postgres-10



- Pivotal **GreenPlum**

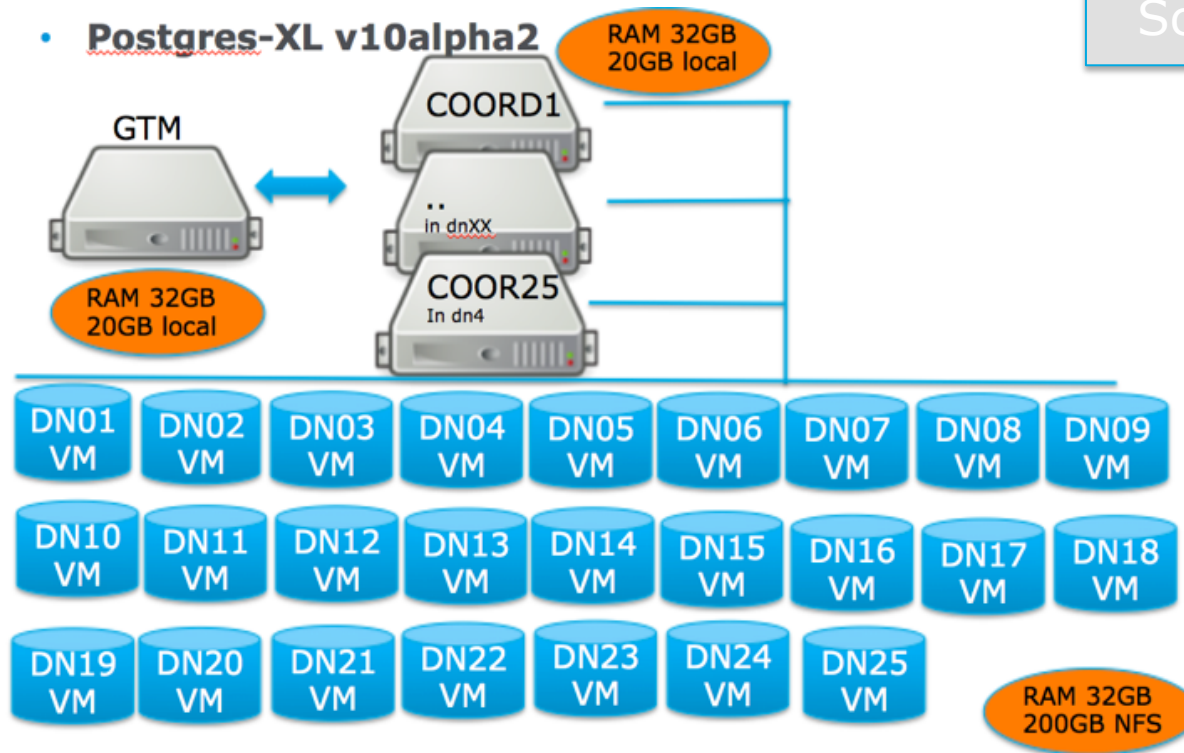


Postgres-XL architecture



Postgres-XL Architecture for Testing

Scicloud-27 VMs



October 2017

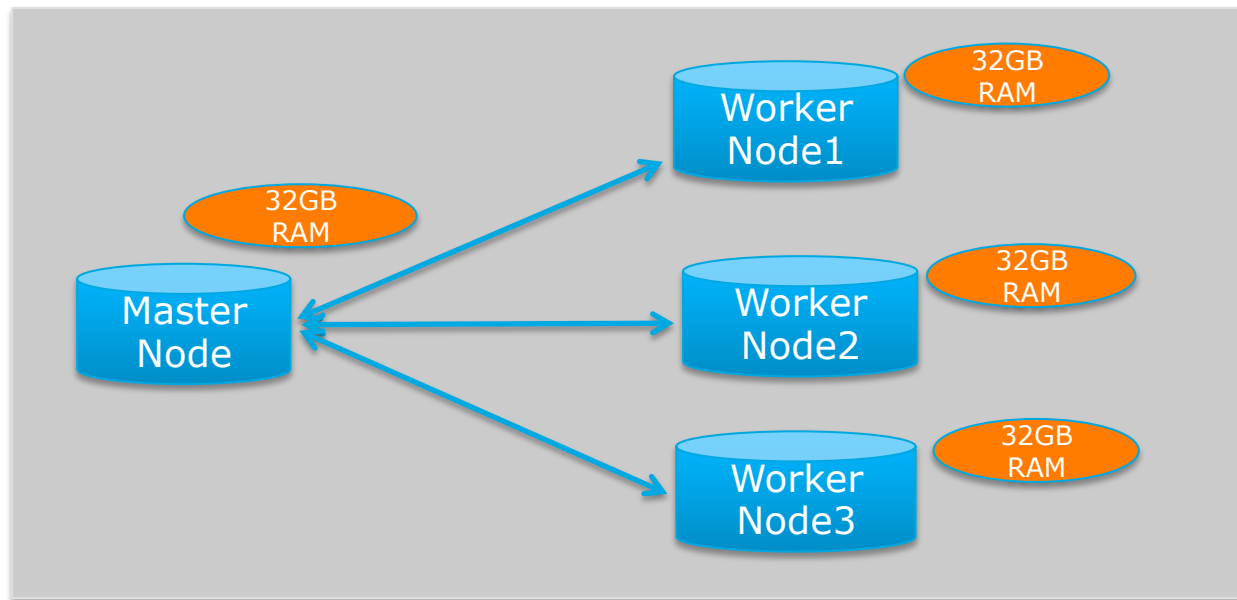
Citus architecture



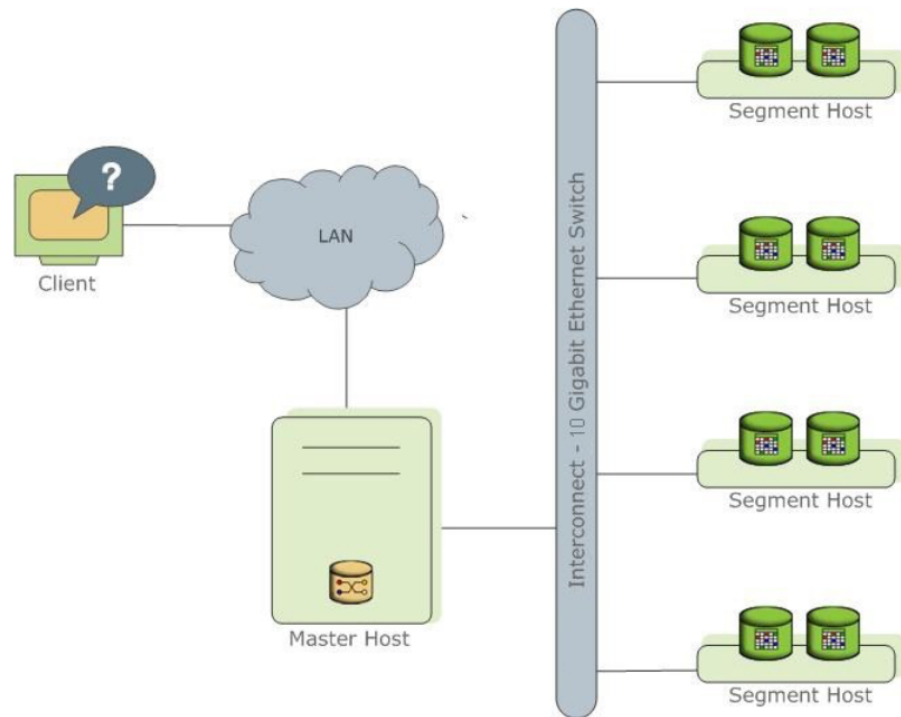
Citus Architecture for Testing

- PostgreSQL-10
- Citus extension:7.1

Scicloud-4VMs



Greenplum architecture



Greenplum architecture for Testing

Scicloud VMs:



Netapp

Scicloud Bare Metal



SAS Drive/
Netapp

Dell Cluster



SSD

Simple Query Tests: Postgres-XL vs Citus



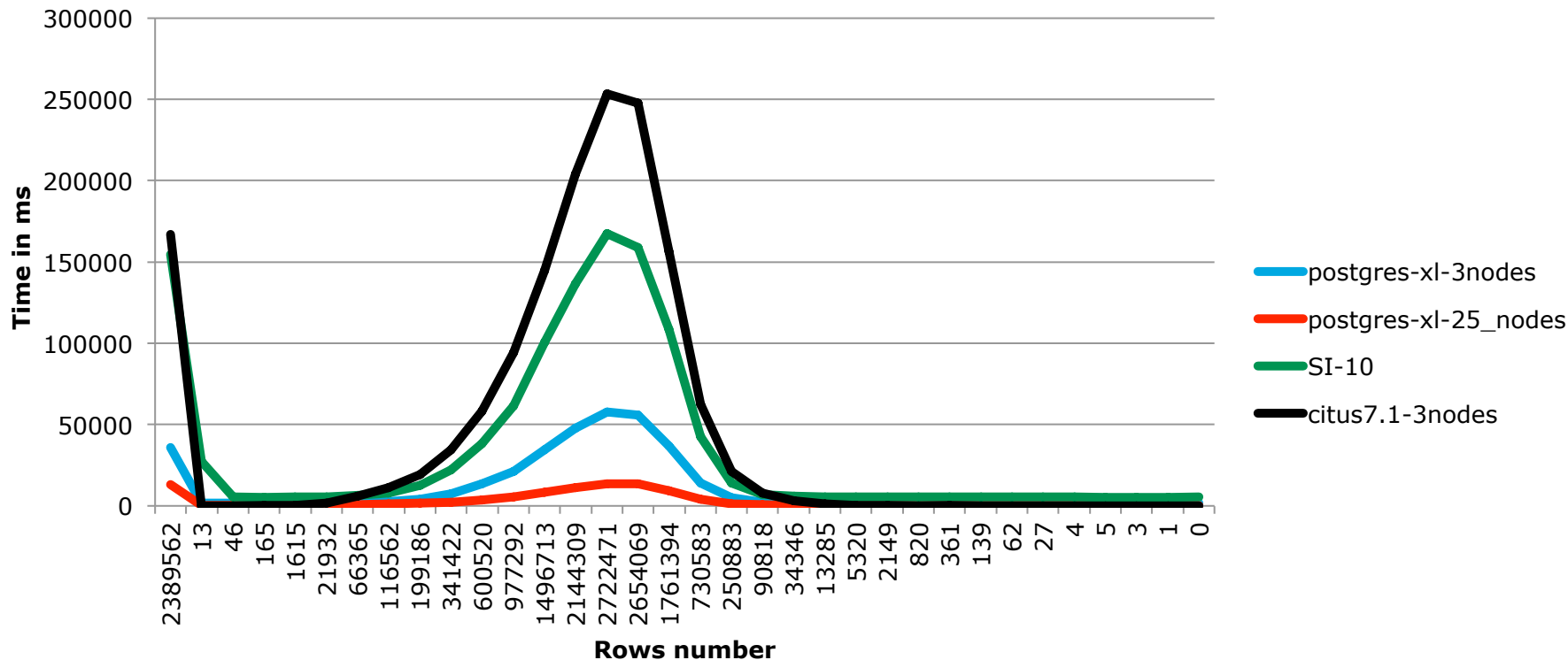
On Kids catalogue:

- ❖ 16M rows
- ❖ 12GB total
- ❖ Comparison for Single instance(SI) on postgres v10 vs postgres-xl, citus running on 3 nodes and postgres-xl running on 25 nodes. **Time grows with the number of rows.**

```
Select *  
from kids_mb_catalogue  
Where mag_auto_i_trunc=XX
```

Filter	Rows	Filter	Rows
10	238956	27	730583
11	1	28	250883
12	4	29	90818
13	16	30	34346
14	161	31	13285
15	2193	32	5320
16	6636	33	2149
17	11656	34	820
18	19918	35	361
19	34142	36	139
20	60052	37	62
21	97729	38	27
22	149671	39	4
23	214430	40	5
24	272247	41	3
25	265406	42	1
26	176139	43	0

Postgres Single Instance vs XL vs Citus (hash distribution)

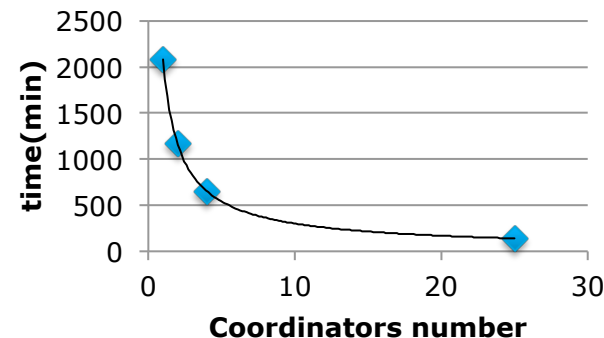


Ingestion scalable tests in Postgres-XL

Ingestion:flagship_mock catalogue: 1,2TB;~10M rows/file;260 files

Ingesting in parallel in X#coordinators:

Coordinators number	Time(min)	Number of files ingested per node	Total time(min)
1	8	260	2080
2	9	130	1170
4	10	65	650
25	14	10	140



Ingesting in parallel with 12-15 coordinator should be enough according to plot.

Hardware testing for `Select count(*) from flagship_mock;`



Environment	Time
Netapp (10 partitions)	3h 46m
Netapp (FDW postgres-xl 25 nodes)	1h 40m
SSD (10 partitions)	1h 17m
Postgres-XL 25 nodes (Netapp scicloud)	35m
Postgres-XL 10 nodes(VM on physical nodes-Netapp)	2h 35min

Flagship_mock table: 1,1TB

Rows#: 2,5B

Greenplum testing



- Ingestion
- Expansion
- Performance
- High Availability
- Resource Management
- Community edition vs Commercial edition

Community edition does not include:

- Support from Pivotal Greenplum team
- Product packaging and installation script.
- Support for QuickLZ compression. QuickLZ compression is not provided in the open source version of Greenplum Database due to licensing restrictions.
- Support for managing Greenplum Database using Pivotal Greenplum Command Center.
- Support for full text search and text analysis using Pivotal GPText.
- Support for data connectors:
 - Greenplum-Spark Connector
 - Greenplum-Informatica Connector
 - Greenplum-Kafka Integration
 - Gemfire-Greenplum Connector
- Data Direct ODBC/JDBC Drivers
- gpcopy utility for copying or migrating objects between Greenplum systems.
- The PXF Apache Ignite connector.

Ingestion:

gpfdist: *nohup **gpfdist** -d /spv02/ -p 4000 > **gpfdist**_4000.log 2>&1 &*

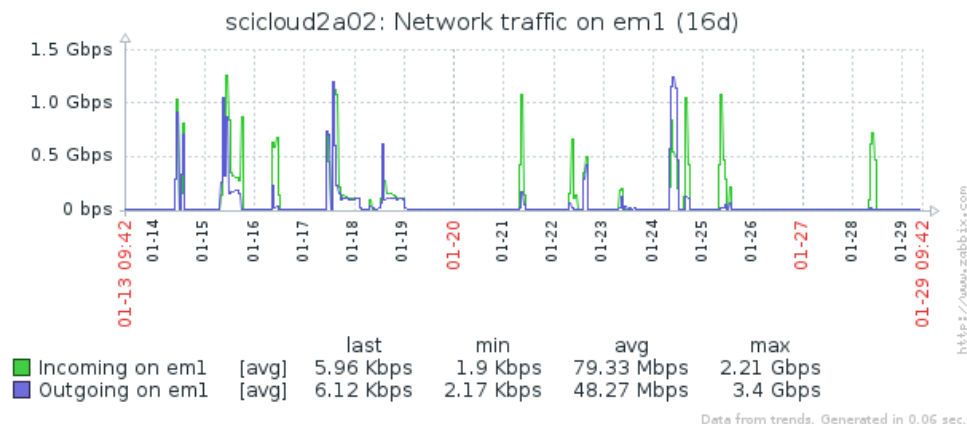
Create external table flag_ext()

location ('gpfdist://172.25.4.244:4000/') FORMAT 'CSV';*

3TB in 3 hours (NetApp)

Expansion:

GPEXpand: Expanding the cluster to new datanodes or same ones increasing number of segments.



GP Testing: Performance



Performance:

- Gaia query profiling tests: Catalogues xmatch
- Euclid query profiling tests: 1,3TB catalogue

What to test:

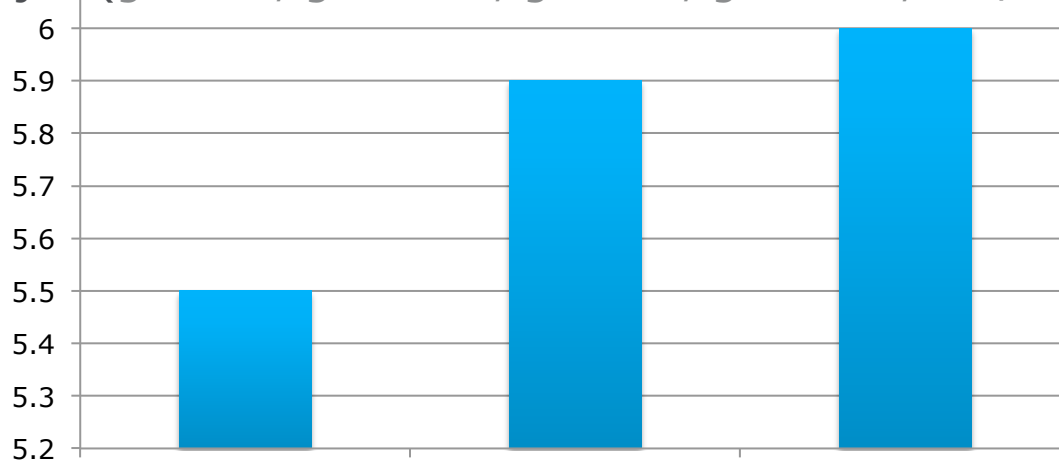
- Table formats
- Scalability

GP testing: Performance

Most relevant queries:

Gaia query:

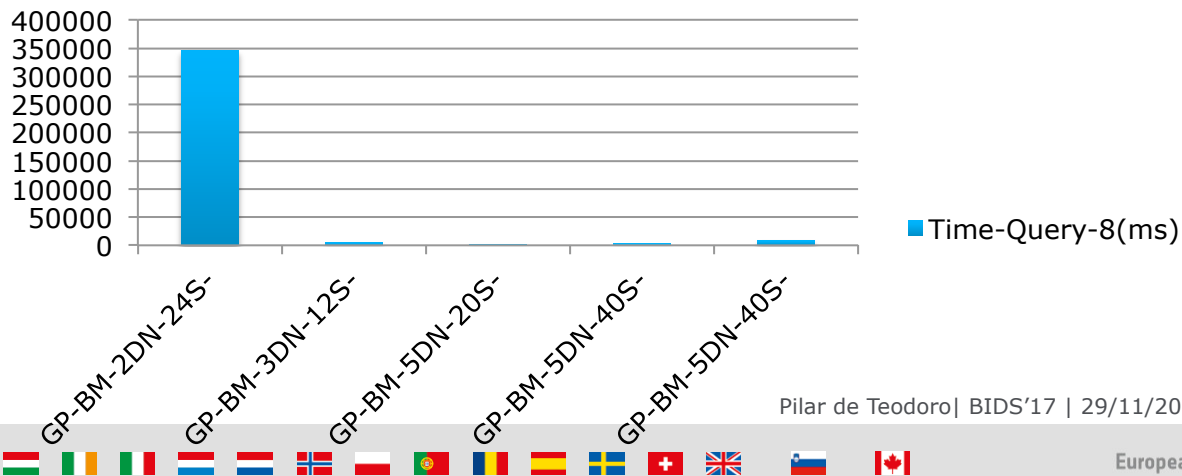
```
SELECT gdr1.source_id AS iddr1, gdr2.source_id AS iddr2  
FROM gaiadr1.gaia_source AS gdr1, gaiadr2.gaia_source AS gdr2  
WHERE q3c_join(gdr2.ra, gdr2.dec, gdr1.ra, gdr1.dec, 0.5/3600)
```



Euclid query:

```
SELECT q3c_dist(  
  euclid.flag."ra_gal",euclid.flag."dec_gal",30,30)*3600 AS dist , euclid.flag."ra_gal" AS ra_gal , euclid.flag."dec_gal" AS dec_gal ,  
  euclid.flag."euclid_vis" AS euclid_vis , euclid.flag."euclid_nisp_y" AS euclid_nisp_y , euclid.flag."euclid_nisp_j" AS euclid_nisp_j ,  
  euclid.flag."euclid_nisp_h" AS euclid_nisp_h , euclid.flag."observed_redshift_gal" AS observed_redshift_gal  
FROM euclid.flag  
WHERE (q3c_join(30,30,euclid.flag."ra_gal",euclid.flag."dec_gal",0.01677)) = '1'  
ORDER BY dest ASC;
```

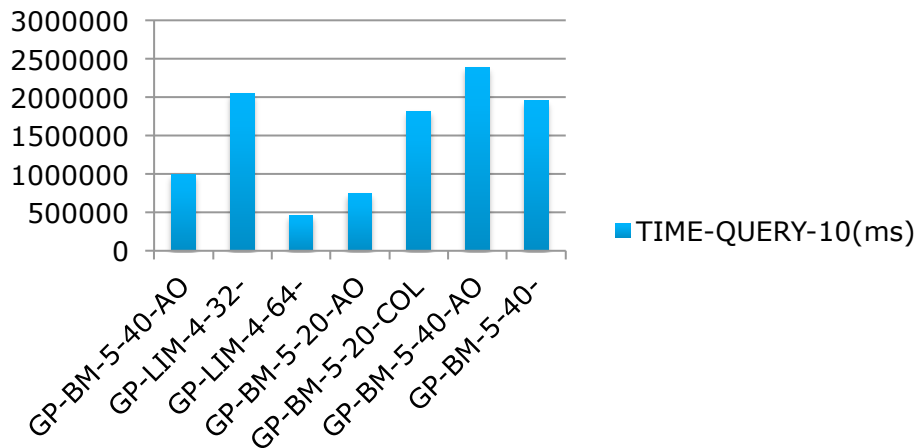
Time-Query-8(ms)



GP testing: Euclid query

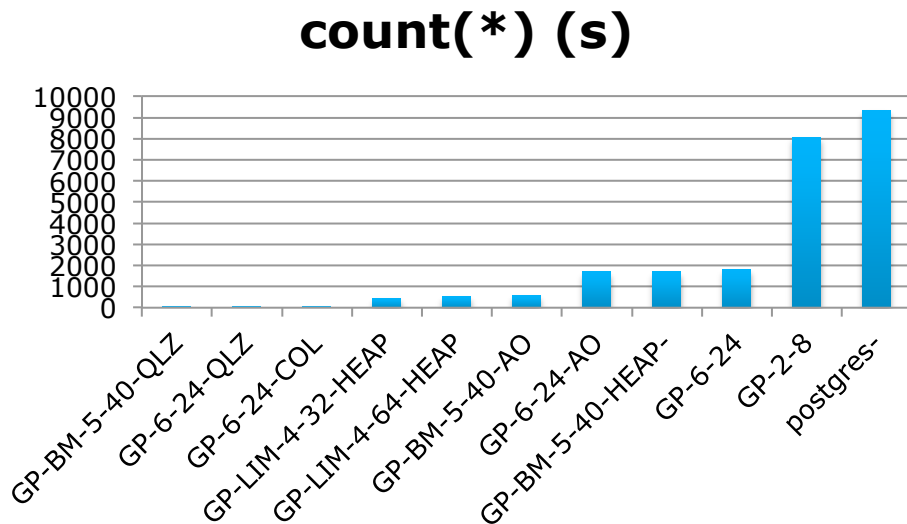
```
SELECT q3c_dist(public.flagship_mock."ra_gal",public.flagship_mock."dec_gal",220.,0.0) AS dist , * FROM public.flagship_mock euclid WHERE public.flagship_mock."galaxy_id" = 2 AND public.flagship_mock."gr_cosmos" > 1;
```

TIME-QUERY-10(ms)



GP Testing: Performance

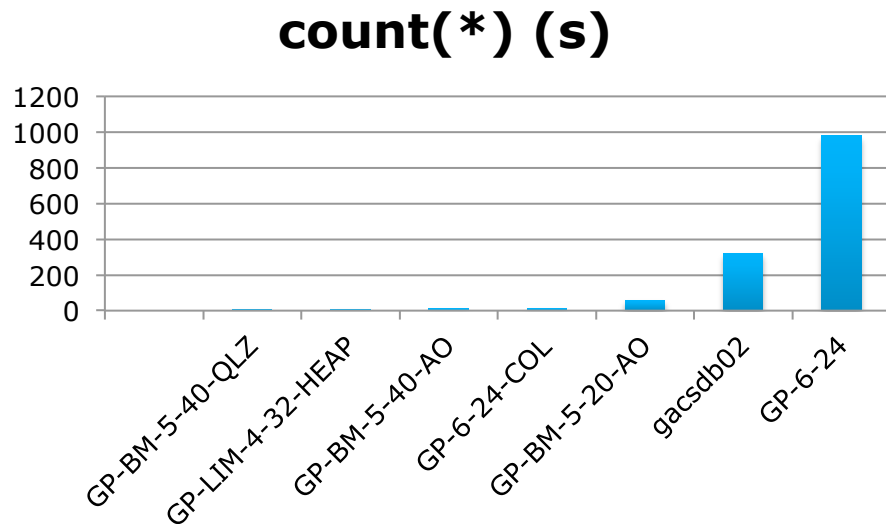
select count(*) from euclid.flagship mock;



GP-BM-5-40-QLZ	14
GP-6-24-QLZ	22
GP-6-24-COL	33
GP-LIM-4-32-HEAP	457
GP-LIM-4-64-HEAP	533
GP-BM-5-40-AO	581
GP-6-24-AO	1693
GP-BM-5-40-HEAP-C	1695
GP-6-24	1808
GP-2-8	8055
postgres-XL-10nodes	9324

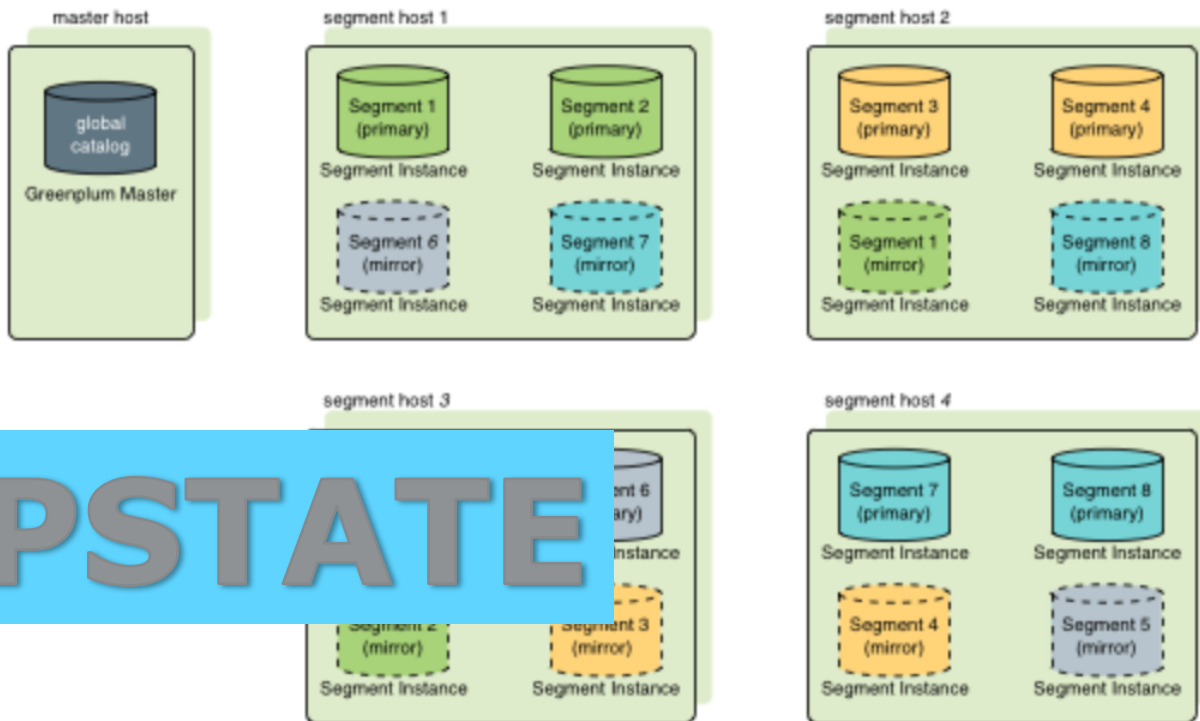
GP Testing: Performance

Select count(*) from gaiadr2.gaia_source: 1692919135 rows



GP-BM-5-40-QLZ	6
GP-LIM-4-32-HEAP	7
GP-BM-5-40-AO	10
GP-6-24-COL	14
GP-BM-5-20-AO	59
gacbdb02	323
GP-6-24	980

GP Testing: High Availability



- Resource groups: set and enforce CPU, memory, and concurrent transaction limits in Greenplum Database. Use Linux Control Groups (cgroups) to manage CPU resources.
- Resource queues: manage the degree of concurrency in a Greenplum Database system.
 - manage the number of active queries that may execute concurrently,
 - the amount of memory each type of query is allocated,
 - priority of queries.

```
esdc=# CREATE RESOURCE QUEUE adhoc WITH (ACTIVE_STATEMENTS=3);  
CREATE QUEUE  
esdc=# ALTER RESOURCE QUEUE adhoc WITH (PRIORITY=LOW);  
ALTER QUEUE
```

Lessons Learnt and what is missing(I)



- ❖ Yes, Postgres-XL is scaling but:
 - ❖ still some issues about stability: GTM memory leak, workaround GTM proxy
 - ❖ does not work with foreign data wrappers
 - ❖ GTM does not rotate logs
 - ❖ not direct monitoring tool
 - ❖ Reshuffling not working
 - ❖ JDBC bug identified by ESDC not resolved yet
- ❖ Citus:
 - ❖ as it is an extension can make some workers to behave as single instance.
 - ❖ Will need to work on performance to retrieve large amount of rows.

Lessons Learnt and what is missing(II)



- ❖ GP Most mature postgres-flavour solution
 - Active developers community
 - GP 5 based on postgres 8.3 use “reserved words”, TAP problem, resolved in future GP6
 - Going to GP 6 based on postgres 9
 - Community edition is enough for ESDC currently
 - Use of existing HW
 - Resource management if needed
 - Best performance: load big catalogues in different format depending on the query and when parsing it decide which one to use using TAP.
 - Relational datamodels must be adapted. No FKs are enforced.

Near Future Tests



Euclid: postgres-xl will be replaced by GP.

We will continue ingesting catalogues in this system.

Summary Table

Solution	Greenplum	Postgres-XL	Citus
Support Company	Pivotal	2ndQuadrant	Citusdata
Last version released*	5.6.15	9.5	8_11
Tested version	5.6.14	10 alpha	7_10
PostgreSQL version	8.3	9.5	10
Allow multiple coordinators?	No	Yes	Yes
Data nodes Mirroring?	Yes	Yes	Yes
Possibility to work on individual nodes?	No	Yes	No
Foreign Data Wrappers allowed?	PXF tools	No (in development)	No
Is it a PostgreSQL fork?	Yes	Yes	No
Commercial support offered	Yes	Yes	Yes
SQL, ACID	Yes	Yes	Yes
Open source community	Yes	Yes	Yes
Support answering speed	Ongoing	Slow	Very slow
Performance on ESDC testing	Good	Good but bug found	Poor
Expansion-reshuffling (3TB)	Good	Bad – crashed	-
Ingestion	Parallelize with GPFDIST	Use multiple coordinators	-
Monitoring tool	Yes (commercial)	No	No

Thank you

