Artículo
Eduardo Anglada · 27 mayo, 2021 Lectura de 8 min

IRIS en Astronomía

En este artículo voy a mostrar los resultados de una comparación entre IRIS y Postgress manejando datos Astronómicos.

Introducción

Desde siempre el cielo nocturno nos ha fascinado. Todos hemos soñado con las estrellas y la posibilidad de que haya vida en otros planetas.

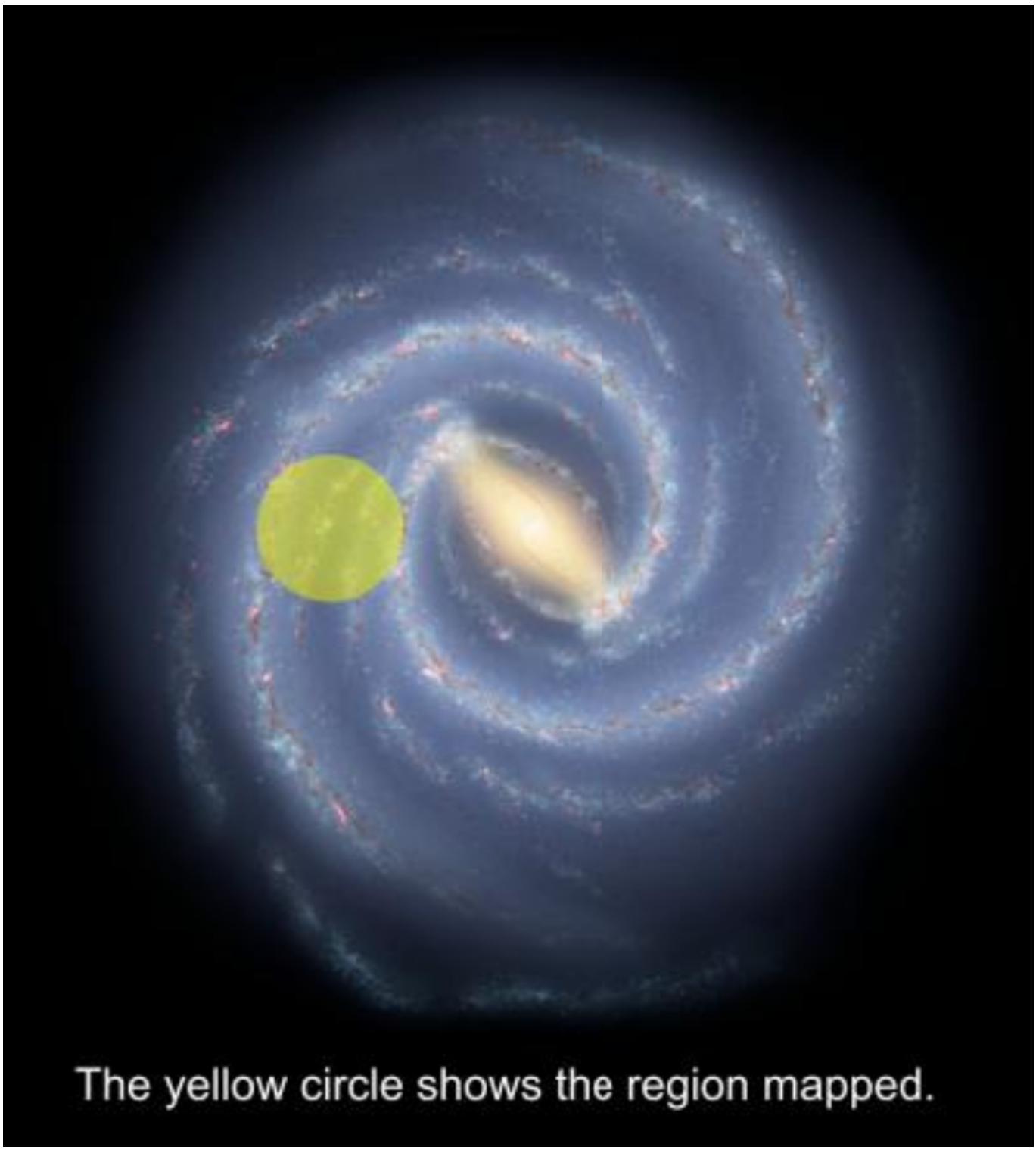
Los astrónomos llevan siglos identificando y clasificando estrellas. Existen catálogos compilados en Mesopotamia y Egipto desde el siglo 2 AC [1]. Durante siglos se han ido compilando catálogos nuevos y recientemente hay dos que sobresalen del resto: Hipparcos [2] y Gaia [3]. Ambos han sido elaborados por la Agencia Europea del Espacio (ESA) que construyó naves espaciales dedicadas al mapeo y caratacterización de estrellas y meteoritos.

Hipparcos fue un proyecto pionero, lanzado en 1989, y permitió determinar la posición de cien mil estrellas con mucha precisión y un millón con una menor precisión. En el año 2000 se publicó el catálogo definitivo que aumentó el número de estrellas a 2.5 millones e incluye el 99% de las estrellas cuya magnitud (medida de brillo sin unidades, ver [4]) es menor que 11. Este catálogo se sigue usando mucho hoy en día.

En 2013 ESA lanzó Gaia [4], una nueva nave espacial dedicada al mapeo de la Vía Láctea y, por supuesto, cualquier otra estrella que detecte. No sólo calcula sus posiciones, también sus velocidades y muchos otros parámetros físicos: brillo en las bandas del rojo y azul, temperaturas superficiales, luminosidad, radio e incluye miles de meteoritos del Sistema Solar. En total el catálogo cuenta con unas 100 columnas que describen las diferentes propiedades y sus errores correspondientes. Todas las estrellas se han clasificado usando el mismo criterio y precisión, pero solo un subconjunto tiene los datos completos. Para el resto, por desgracia, son desconocidos. Durante años los científicos han trabajado duro para que los datos sean lo más fiables posibles.

A lo largo de los años ESA ha publicado varias versiones del catálogo de Gaia y en esta comparativa vamos a usar la versión 2, que es la primera que incluye los resultados para 1.6 billones de estrellas y varios cientos de miles de meteoritos. Los catálogos de Gaia se han usado en miles de artículos científicos.

Los catálogos de Gaia están redifiniendo la astronomía y son un salto cualitativo, tanto en calidad como en calidad, en comparación con los de Hipparcos. Ambos son abiertos y se pueden consultar y descargar desde el servicio de archivos [5]. El Centro Europeo de Astronomía Espacial (ESAC son sus siglas en inglés) es el encargado de custodiar y publicar los catálogos, que están disponibles tanto para la comunidad científica como para el público en general. Para reproducir los resultados se pueden descargar los datos desde el archivo oficial: Gaia Archive (esa.int)



Representación del área de la Vía Láctea que está siendo estudiada por Gaia. (Reproducido gracias a la Licencia Creative Commons By Attribution 4.0 license. Credit: "galaxymap.org, Twitter: @galaxymap").

Gaia pretende cubrir en torno al 1-2 por ciento de las estrellas de la Galaxia.

Cache e IRIS en el procesado de datos diarios de Gaia

La nave Gaia se encuentra en el punto de Lagrange L2 definido por la Tierra y el Sol, a unos 1.5 millones de kilómetros de la Tierra. Siempre está girando y tomando datos, que se reciben en la Tierra usando las antenas localizadas en Cebreros (Ávila, España), New Norcia (Australia) y Malargüe (Argentina). Cada antena envía los datos a Alemania, donde el Centro de Operaciones de Misiones (MOC, en inglés) recibe la telemetría consistente

en los datos científicos y los del estado de la nave. El MOC comprueba el estado de la nave y envía los datos científicos al Centro de Operaciones Científicas (SOC en inglés) en ESAC (Madrid). El promedio diario de datos que recibe el SOC son 40GB, pero fluctúa bastante y puede llegar a los 110GB cuando Gaia está observando el plano de la Galaxia.

InterSystems Caché juega un papel fundamental en el análisis diario de los datos. Gracias a sus capacidades Caché permite acceder en línea a bases de datos de 40TB. Después de ser recibida la telemetría se descomprime e ingesta en InterSystems Caché. El análisis de los datos y las imágenes es llevado a cabo por diferentes programas, pero todos ellos emplean la instancia de Caché para almacenar y leer los resultados. El análisis preliminar de los resultados se tiene que llevar a cabo en menos de 24h para poder crear una alerta científica si, por ejemplo, explotase una Supernova. Caché es capaz de almacenar todos los datos del análisis y gracias a su velocidad y resilencia el SOC es capaz de llevar a cabo su trabajo muy satisfactoriamente. En la actualidad se está llevando a cabo la migración a InterSystems IRIS, la nueva evolución de Caché.

Además los científicos determinan el estado de los diferentes instrumentos y la calidad científica de los resultados. Estos son enviados al resto de centros encargados de su procesamiento (DPAC in inglés), formado por universidades, centros de investigación y observatorios. Los resultados de los estudios realizados por estos centros se guardan en ESAC y se publican como los diferentes catálogos de Gaia.

Rendimiento de IRIS

Vamos a comparar el rendimiento de InterSystems IRIS 2020.1 contra Postgress 12 usando los datos del catálogo Gaia DR2 que han sido descargados en formato CSV del archivo oficial Gaia Archive (esa.int).

Como el catálogo ocupa 1.1 TB y no disponemos de espacio suficiente solo usamos 99GB correspondiente a los archivos CSV cuyo nombre sigue este patrón: GaiaSource1*.csv.

En este repositorio de github se pueden encontrar las instrucciones completas para reproducir estos resultados.

Preparación

Servidor: Ubuntu 20.04 con los últimos parches. 32GB of RAM. Procesador: i7-4790 @ 4.00 GHz 4 cores físicos.

Disco: Sabrent 1TB NVME

Herramienta empleada para llevar a cabo las consultas: DBeaver 21.0.1, usa un driver jdbc apropiado para cada instancia.

IRIS: versión 2020.1, empleando10GB of RAM y archivos de base de datos de 8K.

Postgres: versión 12, instalación por defecto de Ubuntu.

Ingestión de los datos

Los archivos descargados del archivo han sido concatenados en un único archivo.

Ingestión de 115453122 filas con información de las estrellas (94 columnas)

IRIS

Hemos utilizado la utilidad de IRIS SimpleDataTransfer (incluida en la distribución de IRIS). Esta herramienta usa el driver jdbc de IRIS para llevar a cabo la ingestión. Ésta se lleva a cabo en paralelo usando 10 trabajos que ingestan 200000 líneas cada uno.

Postgres

Hemos empleado pgbulkload , que también corre en paralelo y escribe directamente en los archivos de la base de datos sin tener el equivalente a los archivos "journal" activos.

IRIS: 1525 s

Postgres: 2562 s

Consultas simples

Las siguientes consultas son simples y devuelven un conjunto de estrellas que cumplen un requisito dado:

Posiciones: estrellas con un error pequeño en la posición

La consulta es:

```
SELECT * FROM gdr2 WHERE parallax_over_error > 1000
```

Los resultados son:

IRIS: 645 filas .5 s

Postgres: 645 filas 108 s

Estrellas azules

Seleccionar aquellas estrellas con una emisión importante en el azul. La consulta es:

```
select count(*) from gdr2 where bp_rp < -2
```

Los resultados son:

IRIS: 515 filas 2ms

Postgres: 515 filas 150s

Estrellas con movimiento propio importante

Esta consulta devuelve aquellas estrellas con una velocidad transversal importante con respecto a la Tierra:

```
select * from public.gdr2 where (
pmra < -707.1
or pmra > 707.1
or pmdec < -707.1
or pmdec > 707.1
)
and sqrt(pmra * pmra + pmdec * pmdec) > 1e3
```

IRIS 94 filas 34 ms

PG 94 filas 153 s

Resultados de alta calidad (errores pequeños y un número significativo de observaciones)

Anthony Brown el líder de DPAC (el consorcio dedicado al tratamiento de los datos) tiene en su github varios ejemplos. Entre ellos destaca una consulta "simple" que recopila aquellas estrellas con errores pequeños:

```
select source_id, ra, ra_error, dec, dec_error, parallax, parallax_error, parallax_ov
er_error, pmra,
pmra_error, pmdec, pmdec_error, ra_dec_corr, ra_parallax_corr, ra_pmra_corr, ra_pmdec
_corr,
dec_parallax_corr, dec_pmra_corr, dec_pmdec_corr, parallax_pmra_corr, parallax_pmdec_
corr,
pmra_pmdec_corr, radial_velocity, radial_velocity_error,
phot_g_mean_mag, phot_bp_mean_mag, phot_rp_mean_mag, bp_rp, g_rp, bp_g,
2.5/log(10)*phot_g_mean_flux_over_error as phot_g_mean_mag_error,
2.5/log(10)*phot_bp_mean_flux_over_error as phot_bp_mean_mag_error,
2.5/log(10)*phot_rp_mean_flux_over_error as phot_rp_mean_mag_error,
sqrt(astrometric_chi2_al/(astrometric_n_good_obs_al-5)) as uwe
from gaiadr2.gaia_source
where parallax_over_error>5
and radial_velocity is not null
and astrometric_params_solved=31
and rv_nb_transits > 3
```

Resultados

IRIS: En 14m 22s devuelve una lista con 736496 estrellas.

PG: Después de 20m sólo ha encontrado 495895 estrellas.

Mapa de la Galaxia con densidad de estrellas

Desde el año 2009 se publican en https://www.galaxymap.org una serie de mapas de la Vía Láctea. Incluyen mucha información, como un mapa de densidad de estrellas. Mediante isosuperficies (superficies 3D en las que la densidad de estrellas es constante) podemos explorar la Vía Láctea. La última versión emplea los datos de Gaia y el código fuente está disponible en github. Hemos llevado a cabo las mismas consultas en IRIS y Postgress:

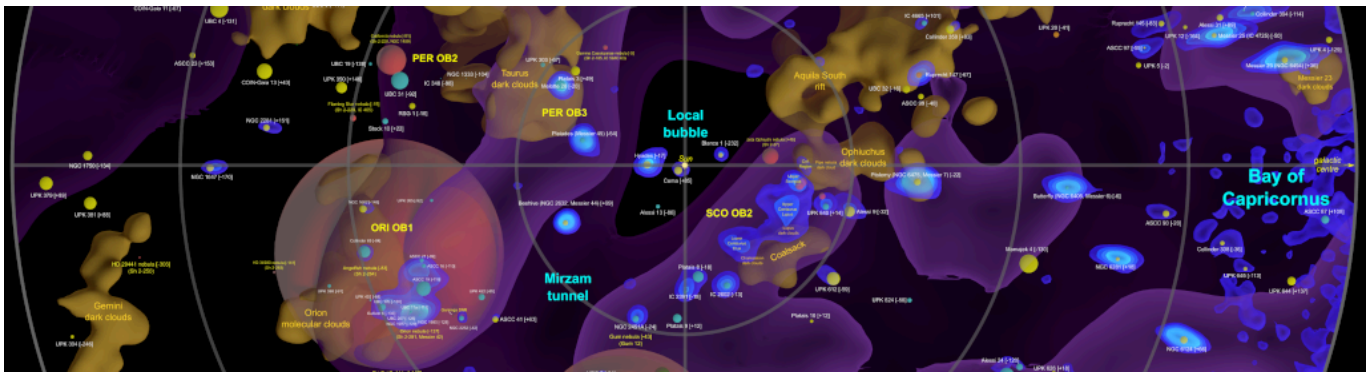
```
SELECT source_id,designation , l , b , parallax , parallax_over_error , phot_g_mean_m
ag , bp_rp , priam_flags , teff_val , a_g_val FROM gdr2 WHERE ( parallax_over_error >
 10 ) and (parallax >= 1.0) and (parallax < 1.1)
```

Resultados:

IRIS: 11m 49s

PG: timeout

Imagen ejemplo, los colores indican las zonas con una densidad de estrellas constante:



Tal y como indica el autor en su blog el archivo oficial de Gaia tiene un tiempo máximo por consulta de 30 minutos, por lo que el autor se vió obligado a realizar muchas consultas y tardó 24 h en obtener todos los datos. IRIS es mucho más rápido y permite obtener los datos en menos tiempo.

Aquí se puede ver una animación de las isosuperficies:

y una descripción completa de los resultados:

Conclusión

IRIS es una plataforma de datos robusta que puede gestionar sin problemas las consultas más complejas y a máxima velocidad.

References

[1] Ancient Star Catalogs. Star chart - Wikipedia

[2] ESA Hipparcos mission. Hipparcos - Wikipedia

[3] ESA Gaia mission. ESA Science & Technology - Gaia

[4] Magnitude: Star brightness. Magnitude (astronomy) - Wikipedia

[5] Parsec: Astronomy unit of length corresponding to 3.26 light years Parsec - Wikipedia

#SQL #InterSystems IRIS