Redshift vs Postgres Sampling & Approximate Query Processing comparison

Khirod Sahoo, Amrit Bhat

BE BOUNDLESS

Motivation

In the era of Big data, analysts seek to perform quick analytics and often sampling of data is done to estimate results for decision making.
We are comparing difference in Approximate Query Processing (AQP) between Redshift and local system on full and sample dataset.

Objectives

1. Check random sampling query execution time.
2. Compare Approximate Query Processing with full vs sampled dataset
3. Overall query performance difference between Redshift and Postgres

Data

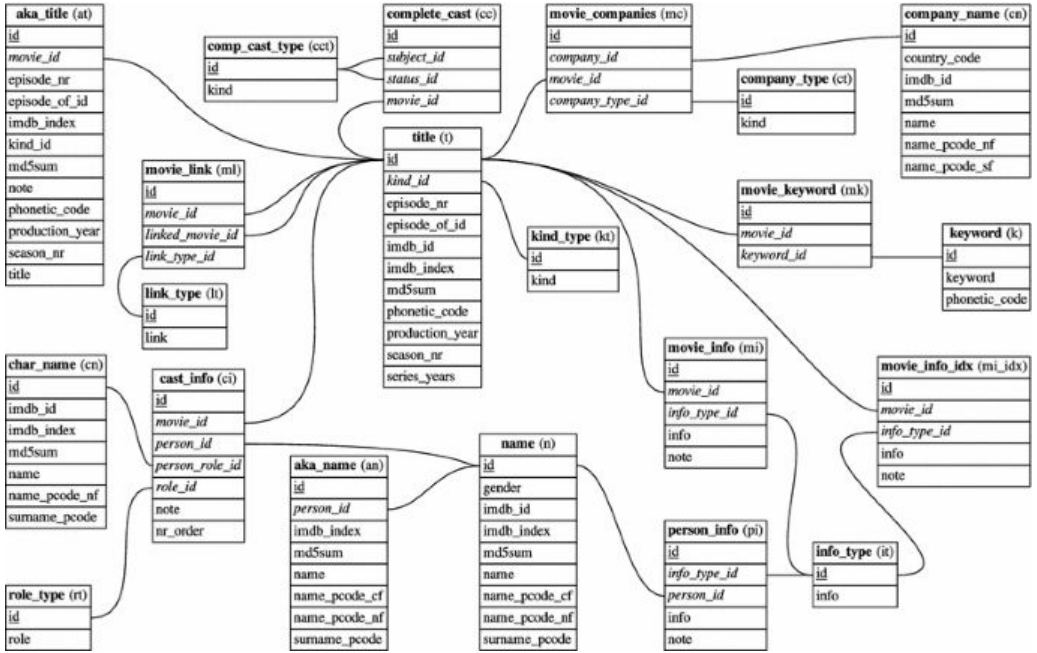
IMDB Join Order Benchmark dataset dump

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/2QYZBT

Size of IMDB_pg dump file- 1.31 GB

Total number of tables - 21

Total size of files- 3.61 GB

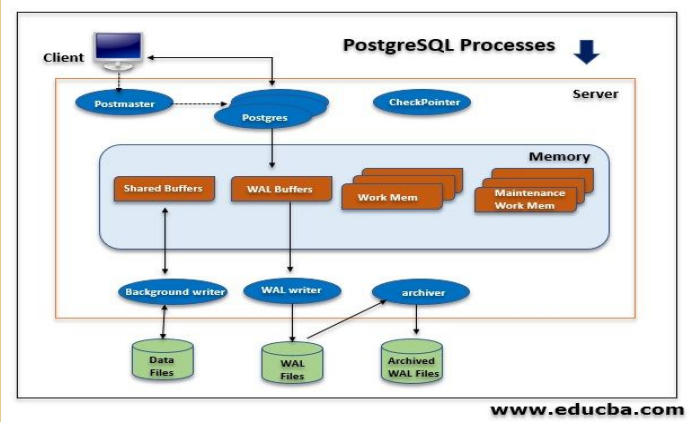


Challenges

- IMDB dataset first downloaded from https://datasets.imdbws.com/ had many data issues while loading in S3 bucket of AWS and hence, we used IMDB_pg dump from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/2QYZBT.

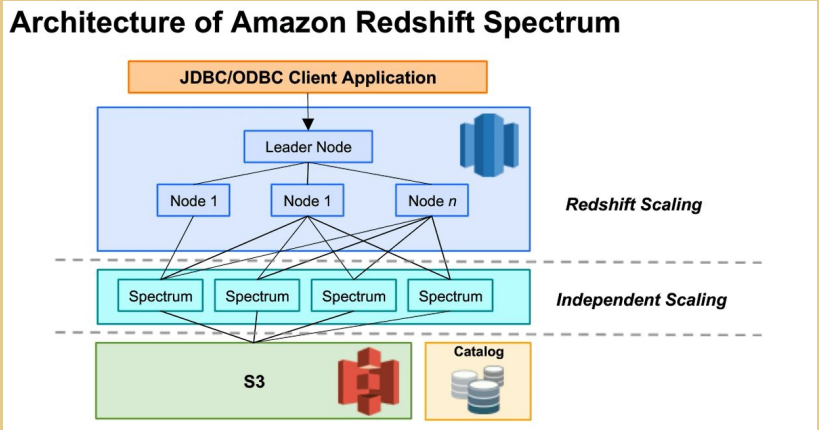
- CLI interface for PostGres was not very user friendly. So we used a PostGre client- DBeaver for running SQL queries on local machine.

System description



Postgres on Windows Acer Nitro 5 Configuration:

RAM - 24 GB
Cores - 4



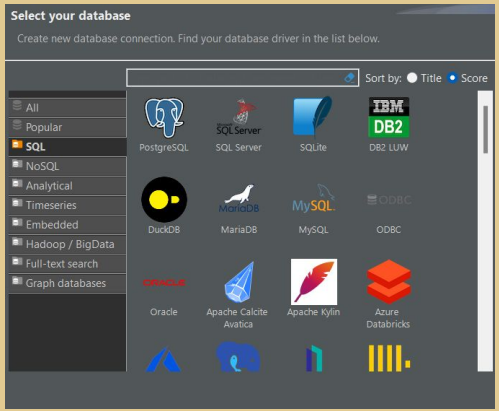
Redshift cluster configuration

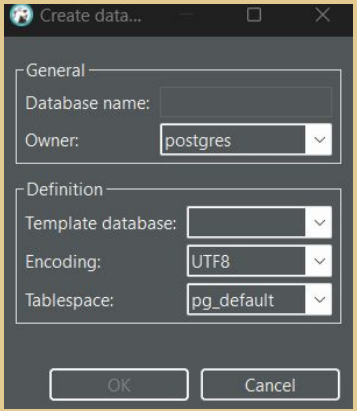
No. of nodes - 2
RAM per node - 15 GB
No. of CPU per node - 2

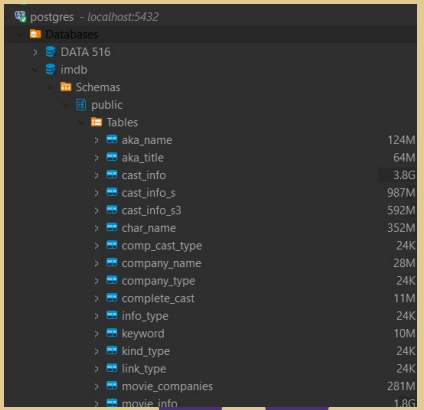
Postgres setup

PostgreSQL (Postgres for short) is a free and open-source relational database management system (RDBMS). It is optimal for a range of workloads, from single machines to data warehouses with concurrent users

Software installation → **Connect to Postgres server** → **Create database** → **Create schemas and tables**



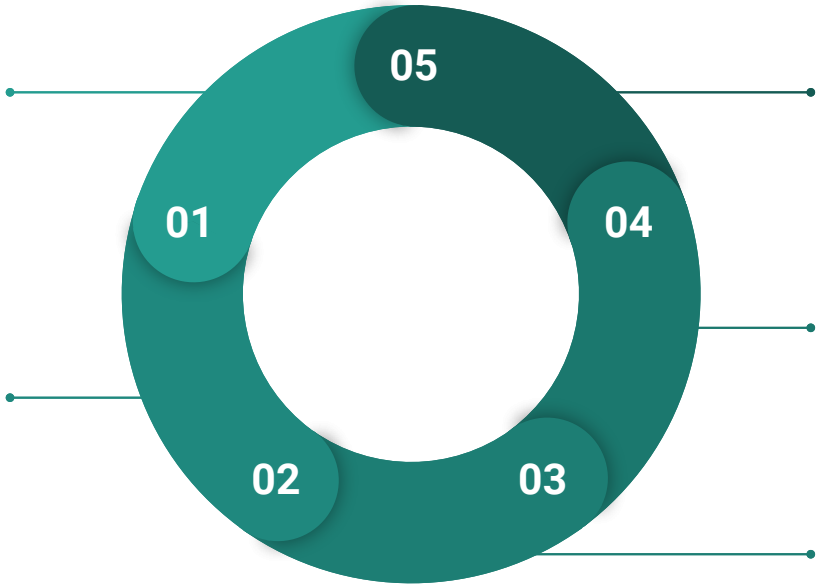




Methods and setup

Configure Dbeaver for Postgres, establish connection to Postgres client, create databases using the sql dump file.

Export the table from Postgres to csv format.



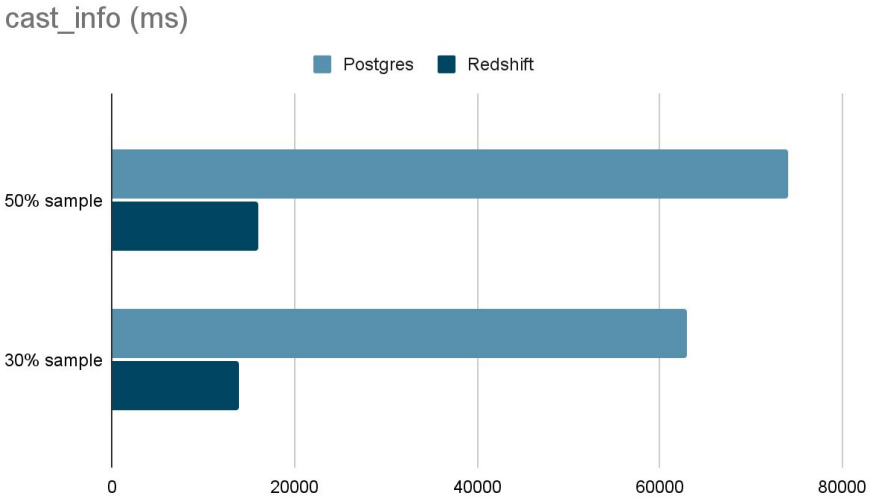
Run queries on both the systems

Create databases on Redshift spectrum after connecting to S3

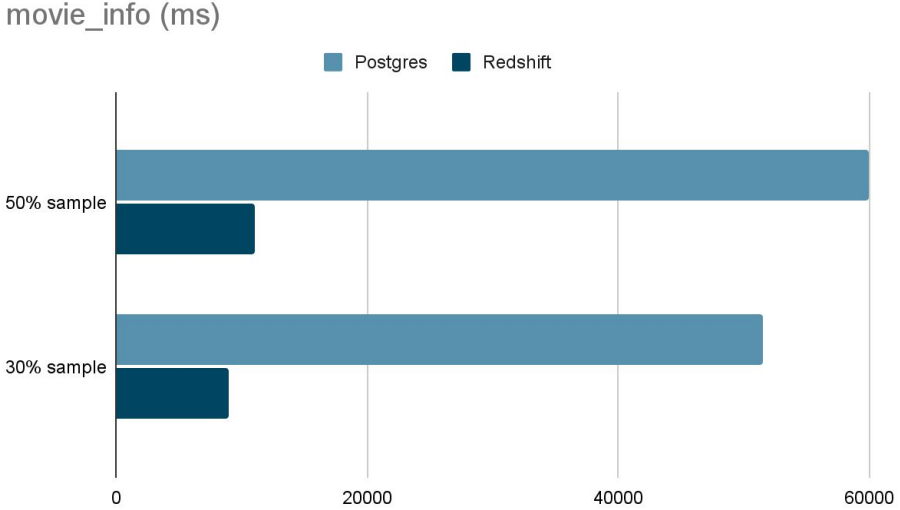
Load the csv files from previous step into AWS S3 bucket

Results - Sampling execution time

~36 million rows

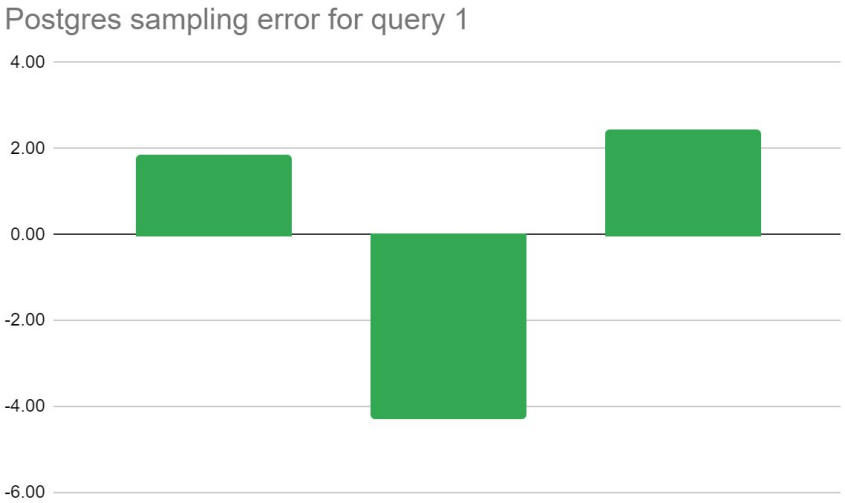


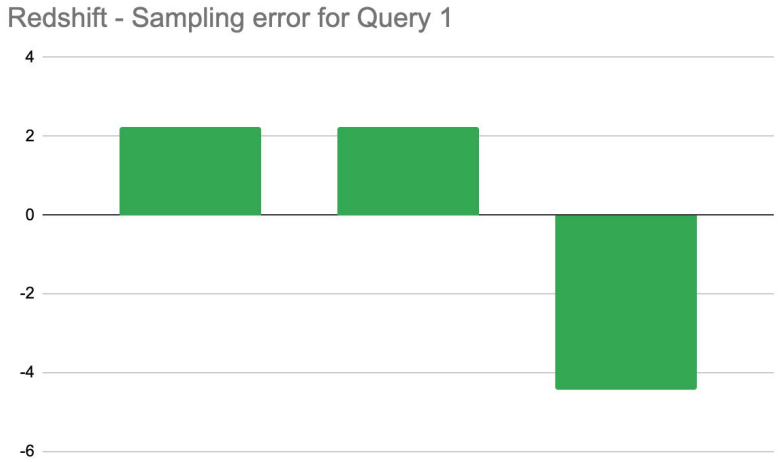
~14 million rows



Redshift is 6X faster than Postgres in Random Sampling

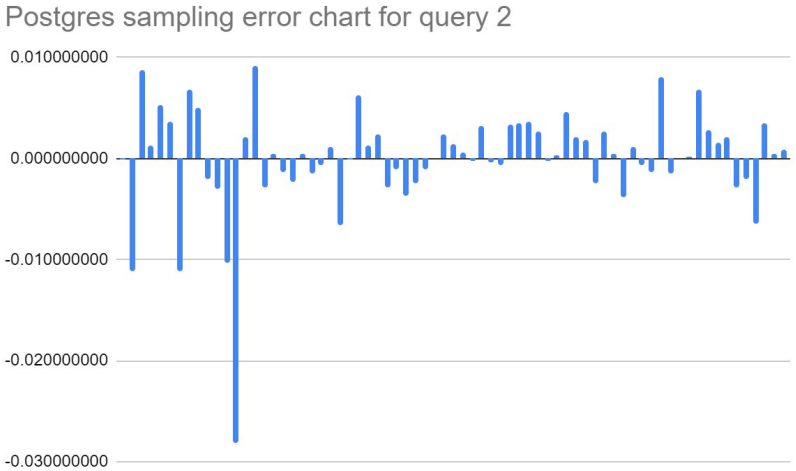
Results - AQP comparison (for Full - 50% sample)

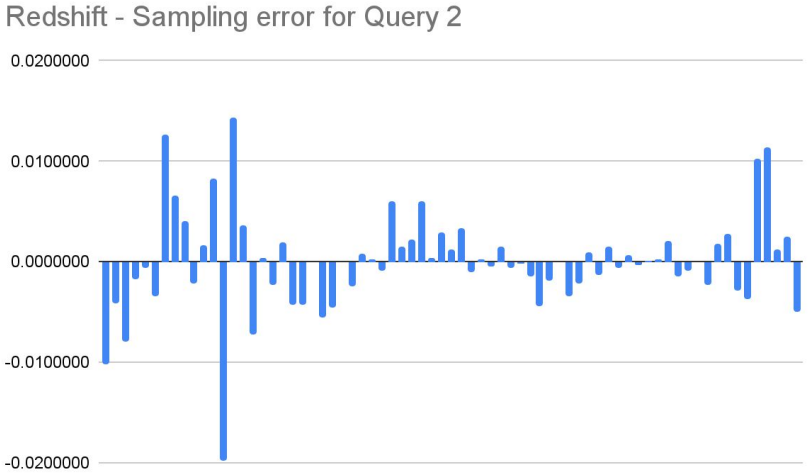




The sampling error is close to 0 in both Redshift and Postgres

Results – AQP comparison (for Full vs 50% sample)

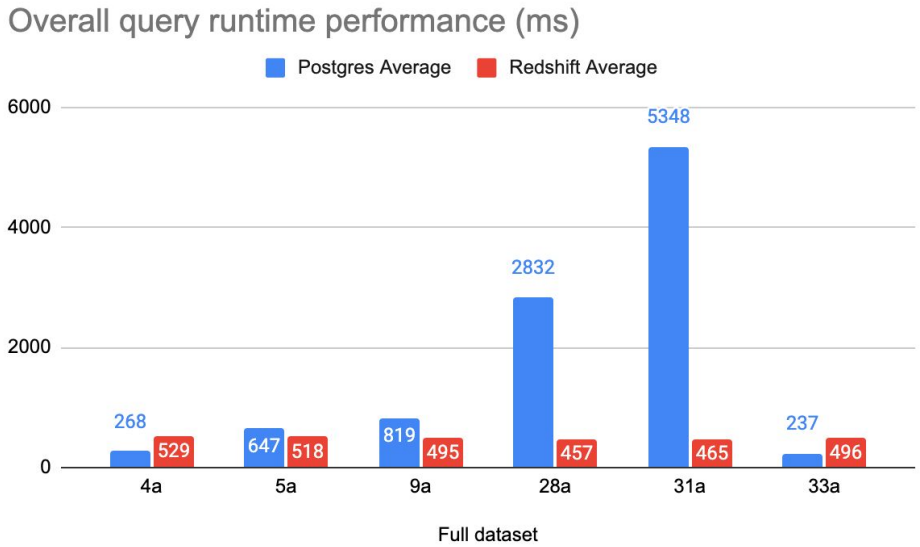




The sampling error is close to 0 in both Redshift and Postgres

Results - Overall query performance

- Postgres execution time increases with increase in number of joins and table sizes
- Redshift runtimes remained approximately same in all the queries



Takeaways

1. Check random sampling query execution time.
    - Redshift 6x faster than Postgres.

2. Compare Approximate Query Processing for full dataset vs sampled dataset
    - Sampling error is close to 0 in both Redshift and Postgres
    - In Big data, we can use sampled data to perform quick analytics

3. Overall query performance difference between Redshift and Postgres
    - Postgres execution time increases linearly with join order but Redshift execution time remains same
    - Postgres does better than Redshift when executing queries on small data!

Thank you!

BE BOUNDLESS