Analytic Benchmarking of DBs

Goal: Compare the benchmark results between Redshift, Redshift Spectrum and Athena

Athena

Description:

- Interactive query service to analyze data directly in Amazon S3 using standard SQL
- Uses Presto with full standard SQL support and works with a variety of standard data formats, including CSV, JSON, ORC, Apache Parquet and Avro
- Serverless
- **Pricing**: $5.00 per TB of data scanned and the S3 storage costs
    - **Note**: compression of the data will reduce Athena cost

Observations

- Does not support all SQL column types and functions e.g. datetime or dateadd()
- Does not support co-related or nested queries (you can use WITH clauses to achieve the same outcome)-
- Athena error message are cryptic , for example:

```
" line 5:17: mismatched input ‘in’ expecting {<eof>, ‘group’, ‘order’, ‘having’, ‘limit’, ‘approximate’, ‘or’, ‘and’, ‘union’, ‘except’, ‘intersect’}
(service: amazonathena; status code: 400; error code: invalidrequestexception; request id: 11fa5fc1-4bbe-11e7-9300-7f0d24e9f51a)”
```

Redshift Spectrum

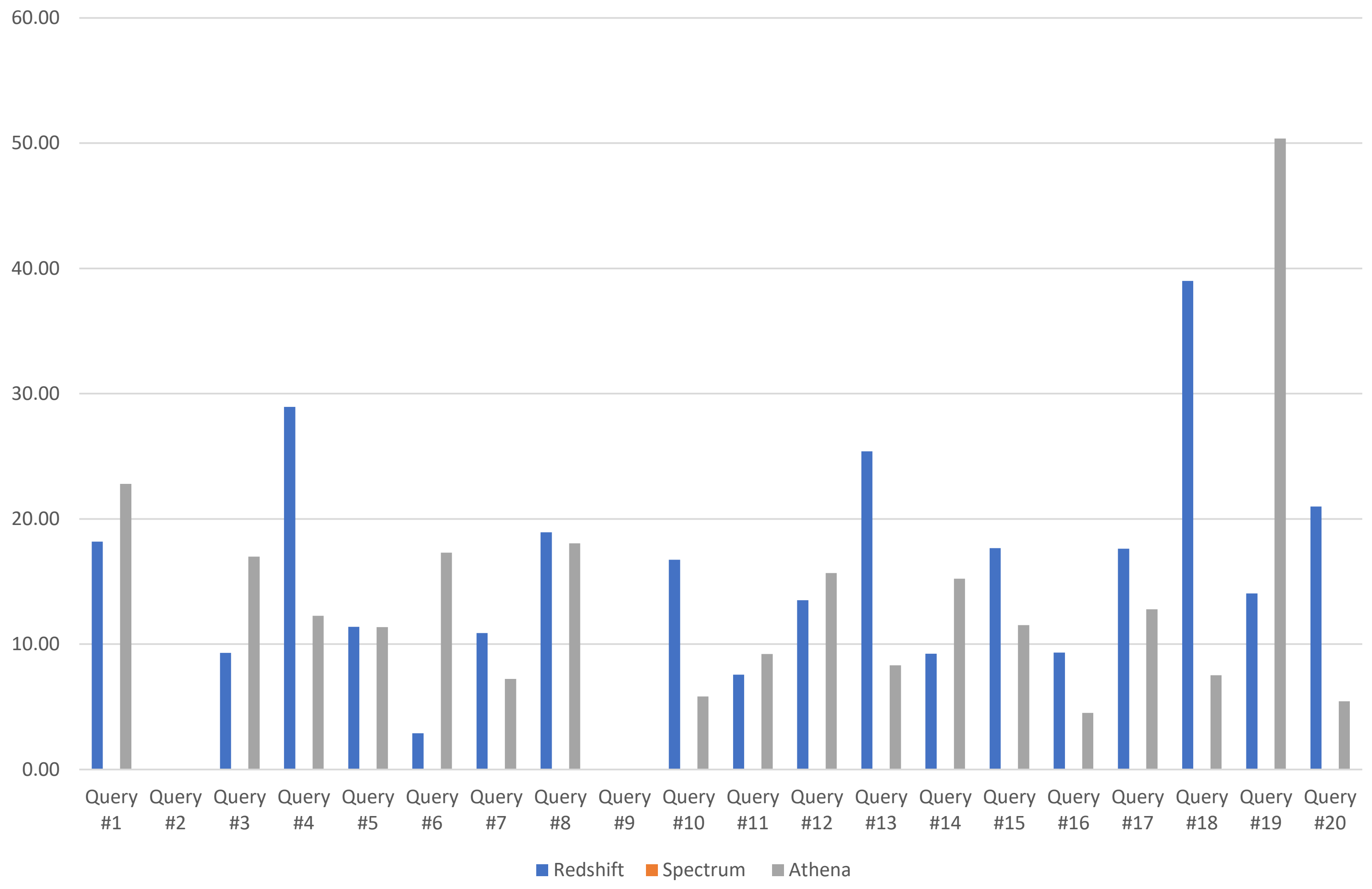
Description:

- Can run queries against exabytes of unstructured data in Amazon S3, with no loading or ETL required
- Redshift Spectrum directly queries data in Amazon S3 using the open data formats (e.g. CSV, TSV, Parquet)
- AWS documentation states that it supports the same SQL syntax of Amazon Redshift; however, inserts are not supported
- **Pricing**: based on Compute node hours, backup storage, data transfer (*to/from Redshift and S3*), data scanned (*same as Athena*)
    - **Note**: 4 node Redshift cluster (dc1.large) was used (~$732 monthly)

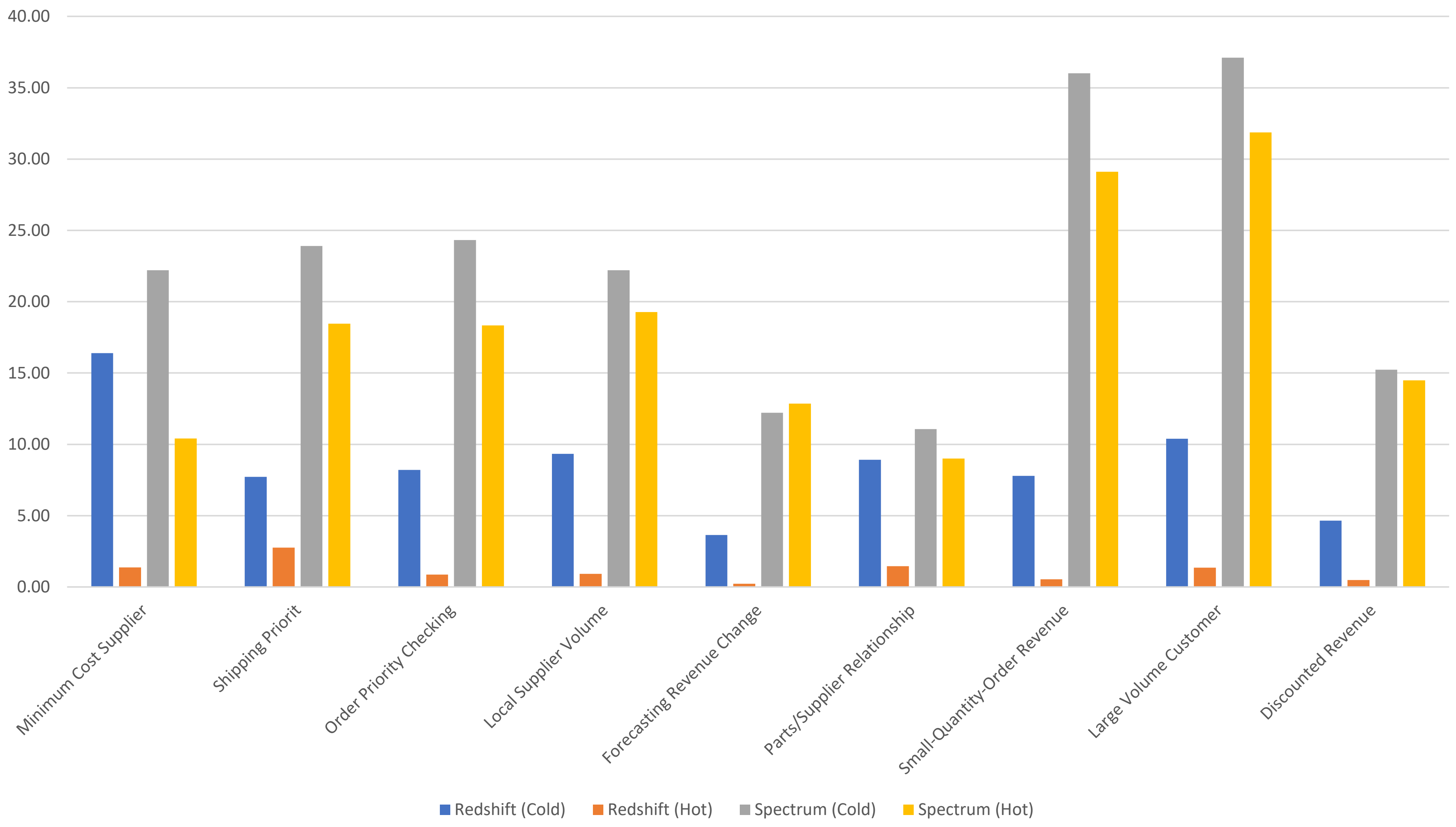
Observations

- Probably used Athena (uses the Athena Full access role), if so, are the Athena limitations inherited by Redshift Spectrum?
- Redshift does not cache used tables, thus re-running query should be the same; however, there is an improvement, which could be a result of Athena caching capabilities

Benchmark Test Results (50GB dataset with cold queries)



Benchmark Test Results (1GB dataset with cold and hot-cache queries)


Minimum Cost Supplier
Shipping Priorit
Order Priority Checking
Local Supplier Volume
Forecasting Revenue Change
Parts/Supplier Relationship
Small-Quantity-Order Revenue
Large Volume Customer
Discounted Revenue
Redshift (Cold)
Redshift (Hot)
Spectrum (Cold)
Spectrum (Hot)