**Configuration used:**

--num-executors 8

--executor-cores 2

--driver-memory 4g

--executor-memory 6g

**Performance metrics:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Input records | Parameters | Avro+Snappy | Parquet-GZIP | Parquet-Snappy |
| 220million -ACT data | Size in HDFS | 85GB | 35GB | 55GB |
| all columns | Read | 1.2min | 25sec | Not able to read.  Error: \_metadata is not a Parquet file (too small) |
| all columns | Write | 6min | 2min | Not able to read |
| all columns | record count | 2min | 30sec | Not able to read |
| 60 columns | Read data Select few columns  using Data Frames Group by on ACS table-(min(score)) Join ACT with ACS Write data | 5min | 2min | Not able to read |

**Other metrics:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| File format | compression codec supported | Default codec | Ideal use case | File format | Remarks |
| Avro | Snappy(faster); Deflate(compact) | none | Schema evaluation | Binary Row | Schema changes can be handled easily. Cross-language Interoperability(Java,C,C++,C#). use schema to generate Java beans or vice-versa. |
| Parquet | Snappy;LZO;GZIP | GZIP | Read random columns | Binary Column | Less I/O. More choices of compression codec. Provides basic aggregations for each column(min,max&count). Column-wise compression. |