LOAD BIG DATA EFFICIENTLY

PART 5: AGGREGATE PUSHDOWN FOR NEXT LEVEL SPEED

- ***Recap predicate Pushdown row and column level***
- ***Aggregate Pushdown***

Summary

- Filter and select push downs work on all data sources
- Aggregate Pushdowns work not on JSON, CSV, AVRO
- Activate aggregate pushdown for Parquet as follows:
  - `spark.conf.set("spark.sql.sources.useV1SourceList", "")`
  - `spark.conf.set("spark.sql.parquet.aggregatePushdown", "true")`
- Aggregate Pushdown has the following limitations:
  - No nested columns and string columns supported for min/max
  - Filter and aggregates are only for partitioned columns supported
- Aggregate Pushdown speeds up the performance significantly of counts, min and max
- V2 Source API seems unclear if more efficient than V1 but SQL interface seems different and Batch Scan is always on.