Optimization

1. Hive
    a. File formats
        i. Parquet: SNAPPY, GZIP, UNCOMPRESSED
        ii. Avro
        iii. Text
    b. Table statistics
        i. By default stats are auto computed when you create a table.
        To disable, *set hive.stats.autogather = false*
        ii. ANALYZE TABLE <table_name> COMPUTE STATISTICS
2. Spark
    a. Frequently used RDD/DF: persist in memory to avoid re-computation
    b. Avoid any collect operation
    c. Avoid any computation on Driver
    d. Use DF/DS instead of RDD
    e. Delay transformations which cause shuffle operations till the end of computation chain if mandatory to use

Reference:

- https://cwiki.apache.org/confluence/display/Hive/Parquet
- https://cwiki.apache.org/confluence/display/Hive/AvroSerDe