SQOOP Guidelines

1. **Always connect to the Standby database when using Sqoop**, unless you are advised otherwise by your DBA. Please work with your DBA or your SA to get the details of your Standby database. If you are working with ADW database, then this is a **mandatory** requirement (not just a guideline)!

2. While working with Oracle database, always ensure that Oraoop is used and the **-Doraoop.chunk.method** parameter is included in the sqoop command.

   Almost all the tables in our databases are partitioned. So while running sqoop job on these tables, please use the option

   -Doraoop.chunk.method=PARTITION

   As our tables are partitioned, and the databases are tuned to work efficiently with partitions, the generated SQL query by this option will run very fast (about 70-90% faster) with no significant load on the database. If the table you are working on is not partitioned, please work with your DBA and/or Systems Analyst before running the Sqoop job and assess the possible impact. In such a case, please explicitly use the option **-Doraoop.chunk.method=ROWID** eventhough ROWID is the default chunk method. This will safegaurd the applications from potential cluster level default value changes in future.

   To find if a given table is partitioned or not, please run the following query from SQL/Plus or SQL Developer

   select owner, table_name, partitioned from all_tables where table_name = '<*TABLE_NAME*>';

3. If you want to select data only from certain partitions of the table, please use the option

   -Doraoop.import.partitions= <*comma separated list of partitions*>

   This will create the SQL query to fetch the data from the listed partitions only, thereby further improving the job performance.

4. When extracting from Oracle DB, most of tables are with default Parallelism set to 4 at Allstate. This could equate to 4 times number of mapper database sessions (example: if you have 32 mappers,

which equate to 4x32 = 128 database sessions.). So it is recommended to disable the parallel query by setting the option

--Doraoop.oracle.session.initialization.statements="alter session disable parallel query;"

If the performance/SLAs are impacted negatively then try increasing the number of mappers.

5. Disable hints on the oraoop generated query, by using the option
   -Doraoop.import.hint=" "

6. Sqoop can import data using a query option i.e. we can specify a query that joins 'n' tables on the source database and fetch the data to be imported into HDFS. Often this causes load on the database end – as the query as to be run on the DB. Instead, import all the required tables into Hadoop and then create a new table in Hive/Impala by running the query on Hadoop. The general rule of thumb is – ***KNOW ABOUT YOUR DATA***.

7. Sqoop can import data in text, sequence file, AVRO or Parquet format. Of all these formats, unless there is a strong use case not to use Parquet format, Parquet is the preferred file format to store normalized data on Hadoop. However, depending on the data volume and the number of columns in each row, sometimes it may take longer to sqoop import into parquet. In such cases, one alternative is – to import the data as compressed AVRO file and then run a map reduce or spark job to convert the AVRO data into parquet format.
   The general rule of thumb is – ***KNOW ABOUT YOUR DATA***.

8. While sqoop can directly import data into Hive, the recommended option is to import the data as a file into HDFS first and then create hive table structure on it (as a 2 step process). This is just to avoid a situation where any failure on hive side marking the sqoop job as failed and lose the [temporary] imported data. Instead, if we imported the data into a directory and then create the Hive table, any errors on Hive side (like file permissions etc.) will not impact the already imported data (the data is still on HDFS).
   For a production type of job, where everything [like file permissions, Sentry rules etc] is ensured/validated prior hand, it is not required to follow this 2-step process.

9. For Avro format as "--as-avrodatafile " and if there are any Date or Timestamp columns in the source table, please make sure to convert those columns into string format, Since Avro does not work well with the date and timestamp format.
   --map-column-java COLUMN1_DATE=String, COLUMN2_TS=String

10. The default number of mappers in Sqoop is 4 if you do not explicitly specify in your Sqoop statement (like -m 16). Please review with your DBA and understand your Data, before you set the number of mappers in Sqoop command. Sqoop will create the 'number of mappers' many Sqoop's part files on HDFS and while execution, this many sessions/tasks will be created as well.

11. SQOOP Apache Documentation Links:

    https://sqoop.apache.org/docs/1.4.6/SqoopUserGuide.html

12. Sqoop Timezone Behaviors and Inconsistencies. Be aware of situations where the time portion of dates & timestamps can shift as a result of implicit timezone conversion.

    - Sqoop import using Oraoop and write to Parquet
      - The correct representation of the source Oracle date/time will be written to parquet. When querying the parquet datasets via Impala, date/times will be consistent with Oracle but when querying via Hive (Beeline) and Spark SQL, dates/times may shift by the hour differential between your timezone and UTC/GMT.

    - Sqoop import (Oraoop disabled) and write to Parquet
      - The hour differential between the Sqoop job timezone and UTC/GMT will be added to dates/times written to parquet. Note that dates/times queried via Impala will be off by the aforementioned hour differential but dates/times queried via Hive (Beeline) and Spark SQL will be consistent with Oracle as an implicit timezone conversion occurs with Hive & Spark SQL.

    More information and detailed examples can be found at this page:

https://start.allstate.com/sites/BigDataCoE/_layouts/15/start.aspx#/SitePages/Timezone%20Behaviors.aspx

13. Sqoop Password Management Best Practices

    - Please see the following Guidelines and best practices for managing user and application password in your Sqoop jobs:

- https://start.allstate.com/sites/BigDataCoE/SitePages/Password%20Management%20Best%20Practice.aspx