**Lab Submission 05.**

**Data file types**

**Instructor: Basit Ali**

**Differentiate all the studied data file types supported by hadoop/hive/impala.**

The data file types supported by Hadoop/hive/impala are:

1. TEXTFILE
2. AVRO
3. PARQUET
4. ORCFILE
5. SEQUENCEFILE
6. RCFILE

**TEXTFILE**

* A plain text file or CSV is the most common format both outside and within the Hadoop ecosystem.
* The great disadvantage in the use of this format is that it does not support block compression, so the compression of a CSV file in Hadoop can have a high cost in reading.

**AVRO**

* Avro is a row-based storage format.
* This format includes in each file, the definition of the scheme of your data in JSON format, improving interoperability and allowing the evolution of the scheme.
* Avro also allows block compression in addition to its divisibility, making it a good choice for most cases when using Hadoop.

**PARQUET**

* Parquet is a column-based (column-based) binary storage format that can store nested data structures.
* This format is very efficient in terms of disk input / output operations when the necessary columns to be used are specified.
* This format is very optimized for use with Cloudera Impala.

**ORCFILE**

* ORC (Optimized Row Columnar) is considered an evolution of the RCFile format and has all its benefits alongside with some improvements such as better compression, allowing faster queries.
* This format also does not support the evolution of the scheme.

**SEQUENCEFILE**

* The SequenceFile format stores the data in binary format.
* This format accepts compression; however, it does not store metadata and the only option in the evolution of its scheme is to add new fields at the end.
* This is usually used to store intermediate data in the input and output of MapReduce processes.

**RCFILE**

* RCFile (Record Columnar File) is a columnar format that divides data into groups of rows, and inside it, data is stored in columns.
* This format does not support the evaluation of the scheme and if you want to add a new column it is necessary to rewrite the file, slowing down the process.

**---------------------------------------------**