**ASSIGNMENT 28.3**

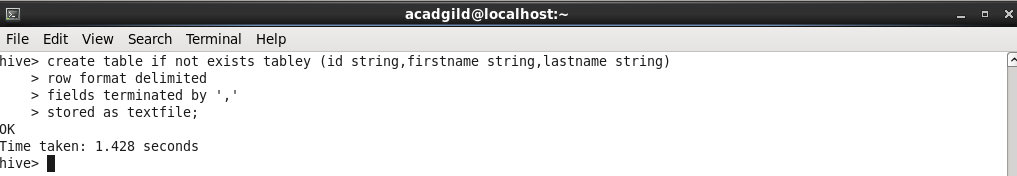
**Explain File Formats in Hive in Brief with an example.**

There are some specific file formats which Hive can handle such as,

• TEXTFILE  
• SEQUENCEFILE  
• RCFILE  
• ORCFILE

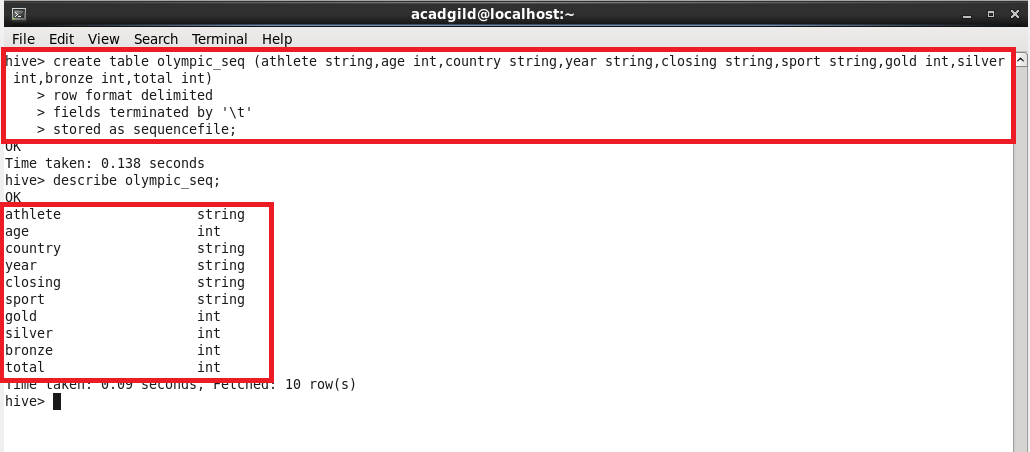
**TEXTFILE**:

* TEXTFILE format is a famous input/output format used in [Hadoop](https://acadgild.com/big-data/big-data-development).
* In Hive if we define a table as TEXTFILE it can load data of from CSV (Comma Separated Values), delimited by Tabs, Spaces, and JSON data.
* This means fields in each record should be separated by comma or space or tab or it may be JSON(JavaScript Object Notation) data.
* By default, if we use TEXTFILE format then each line is considered as a record.
* Without specifying anything, hive stores all the files as text file.



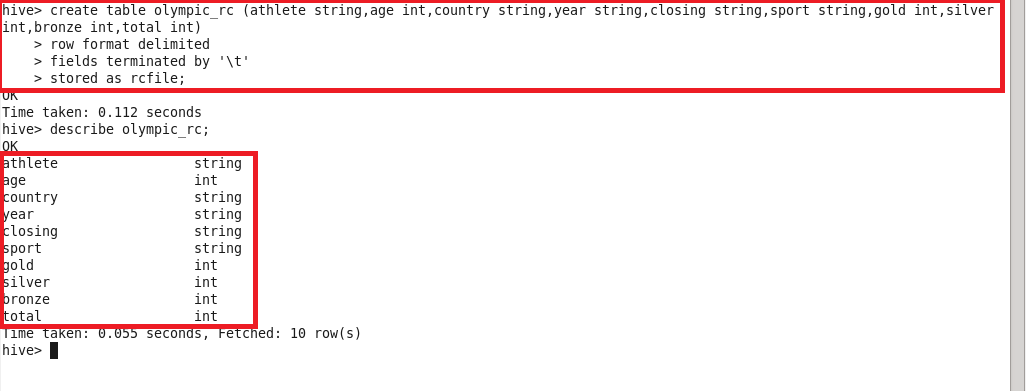
**SEQUENCEFILE:**

* We know that Hadoop’s performance is drawn out when we work with a small number of files with big size rather than a large number of files with small size.
* If the size of a file is smaller than the typical block size in Hadoop, we consider it as a small file.
* Due to this, a number of metadata increases which will become an overhead to the NameNode.
* To solve this problem sequence files are introduced in Hadoop. Sequence files act as a container to store the small files.
* Sequence files are flat files consisting of binary key-value pairs. When Hive converts queries to MapReduce jobs, it decides on the appropriate key-value pairs to be used for a given record.
* Sequence files are in the binary format which are able to split and the main use of these files is to club two or more smaller files and make them as a one sequence file.
* In Hive we can create a sequence file by specifying STORED AS SEQUENCEFILE in the end of a CREATE TABLE statement.

****

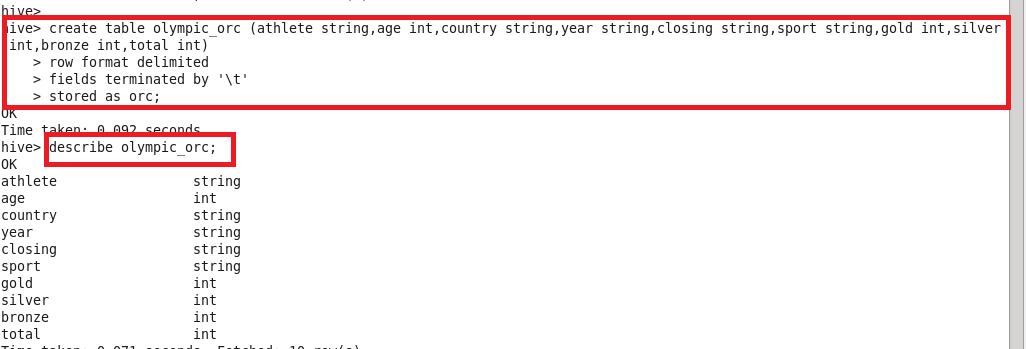
**RC FILE:**

* RCFILE stands of Record Columnar File which is another type of binary file format which offers high compression rate on the top of the rows.
* RCFILE is used when we want to perform operations on multiple rows at a time.
* RCFILEs are flat files consisting of binary key/value pairs, which shares much similarity with SEQUENCEFILE.
* RCFILE stores columns of a table in form of record in a columnar manner.
* It first partitions rows horizontally into row splits and then it vertically partitions each row split in a columnar way.
* RCFILE first stores the metadata of a row split, as the key part of a record, and all the data of a row split as the value part.
* This means that RCFILE encourages column oriented storage rather than row oriented storage.
* This column oriented storage is very useful while performing analytics.
* It is easy to perform analytics when we “hive’ a column oriented storage type.



**ORCFILE:**

* ORC stands for Optimized Row Columnar which means it can store data in an optimized way than the other file formats.
* ORC reduces the size of the original data up to 75%.
* As a result the speed of data processing also increases. ORC shows better performance than Text, Sequence and RC file formats.
* An ORC file contains rows data in groups called as Stripes along with a file footer.
* ORC format improves the performance when Hive is processing the data.



Thus you can use the above four file formats depending on your data.

For example,

\* If your data is delimited by some parameters then you can use TEXTFILE format.  
\* If your data is in small files whose size is less than the block size then you can use SEQUENCEFILE format.  
\* If you want to perform analytics on your data and you want to store your data efficiently for that then you can use RCFILE format.  
\* If you want to store your data in an optimized way which lessens your storage and increases your performance then you can use ORCFILE format.