**Assignment 1**

**Give brief answers to the below questions:**

1. **When Hive is best suited and when is it not?**

**Ans. Hive is best for analyzing the structured data in the big data ecosystem. Data set should be in the large volume then Hive will give the best optimize result on the other hand querying small dataset using hive will be very slow. Hive is best suited for batch processing.**

1. **When should one use Hive over MapReduce?**

**Ans. To reduce the complexity of the writing map reduce program and not being expert of java language one should use hive. Hive also uses map reduce to process the data but it hides the complexity of the map reduce program. It’s a sql type environment which sits on top of Hadoop. It is query language and easy to understand.**

1. **What is Hive metastore?**

**Ans. Hive metastore is table that stored the metadata information in hive .Matastore table lies in the rdbms.Default rdbms for storing metadata information in hive is derby.There are three types of metastore in hive 1.Embedded metastore 2. Local metastore 3. Thrift metastore**

1. **How can Hive improve performance with orc file format tables?**

Ans. **A file format is a way in which information is stored or encoded in a computer file. In Hive it refers to how records are stored inside the file.  ORC (optimized record columner) file format reduces the size of the original data up to 75%. As a result the speed of data processing also increases**

1. **What is thrift server and client, jdbc and odbc driver importance in hive?**

**Ans. Using Thrift Server and client, any programming language that supports thrift service can interact with hive server.** There are third party projects providing clients for Python and Ruby**.Using Jdbc driver a java application can connect and interact with hive server.Using ODBC ,application that support ODBC protocol can connect and interact with hive server.**

**6. What is the importance of partition in hive?**

**Ans. Partition is used for optimizing the query result in hive, speed up the query performance and helps in organizing data in logical manner so that every time hive does not need to scan complete dataset. It is used for distributing load horizontally. There are two types of partition is hives 1. Static partition (when we know data belongs to which partition) 2. Dynamic partition(we can create as many number of partition with single hive statement).**

**7.What is the use of bucketing in hive?**

**Ans. When we need to create lot of tiny partitions for values of a column in a table then we prefer using bucketing. Bucketing overcome the over partitioning on a table and decompose data in to more manageable part called bucket.**

**8. What is the difference between static partitioning and dynamic partitioning in hive?**

**Dynamic Partition:**

**single insert to partition table is known as dynamic partition**

**Usually dynamic partition load the data from non partitioned table**

* **Dynamic Partition takes more time in loading data compared to static partition**
* **When you have large data stored in a table then Dynamic partition is suitable.**
* **If you want to partition number of column but you don’t know how many columns then also dynamic partition is suitable**
* **Dynamic partition there is no required where clause to use limit.**
* **we can’t perform alter on Dynamic partition**
* **Dynamic partition is in NonStrict Mode**

**Static Partition:**

* **Insert input data files individually into a partition table is Static Partition**
* **Usually when loading files (big files) into**[**Hive tables**](http://www.hadooptpoint.com/hive-create-table-examples/)**static partitions are preferred**
* **Static Partition saves your time in loading data compared to dynamic partition**
* **You “statically” add a partition in table and move the file into the partition of the table.**
* **We can alter the partition in static partition**
* **You can get the partition column value form the filename, day of date etc without reading the whole big file.**
* **If you want to use Static partition in hive you should set property set hive.mapred.mode = strict  This property set by default in hive-site.xml**
* **Static partition is in Strict Mode(default mode)**
* **You should use where clause to use limit in static partitio**