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

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

Hive is a data warehouse solution (where data is structured, static and formatted) built on top of Hadoop. It processes structured data in Hadoop cluster. It is designed to enable easy data summarization, ad-hoc querying and analysis of Big data on Hadoop.

Hive provides a SQL like query language called HiveQL. It has a minimal learning curve for people with SQL knowledge. It automatically uses HDFS for storage, but stores all the meta-information about database and table in metadata DB locally to Hive. It is better suited for ad-hoc queries.

Hive is not suited when row wise update and insert is required. It is not well suited to work with hierarchical data. Hive queries have higher latency due to start up overhead. Because of this, queries that would finish in milliseconds on traditional databases would take more time on Hive even on smaller set of data.

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

MapReduce is the default execution engine for Hive. For fine grained and complex processing use MapReduce. It is used when one wants to avoid writing complex MapReduce codes.

1. **What is Hive metastore?**

The Hive metastore service stores the metadata for Hive tables and partitions in a relational database, and provides clients (including Hive) access to this information via the metastore service API. The metastore is the central repository of Hive metadata. It runs in the same JVM as the Hive service and contains an embedded Derby database instance. However, only one embedded Derby database can access the database files on disk at any one time, which means we can have only one Hive session open at a time.

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

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.

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

The Hive server is exposed as a Thrift service, so it’s possible to interact with it using any programming language that supports Thrift. There are third party projects providing clients for Python and Ruby.

Using jdbc driver a Java application will connect to a Hive server running in a separate process at the given host and port.

An ODBC driver allows applications that support the ODBC protocol (such as business intelligence software) to connect to Hive.

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

Partitioning is used for distributing execution load horizontally. As the data is stored in parts, the query response times becomes faster since it has to look up into the partitioned table only as compared to the entire data set. It is useful when there are limited number of partitions and comparatively equal sized partitions.

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

Bucketing function is based on

(hastag function on the bucketed column)mod(total number of buckets)

Records with the same bucketed column will always be stored in the same bucket. ‘Clustered by’ clause is used to divide the table into buckets. Map-side joins will be faster on bucketed tables as the data files are equal sized parts. Bucketing offers efficient sampling.

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

Static Partitioning

* Insert input data files individually into a partition table.
* Usually when loading files (big files) into [Hive tables](http://www.hadooptpoint.com/hive-create-table-examples/) static partitions are preferred.
* Saves your time in loading data compared to dynamic partition
* We can “statically” add a partition in table and move the file into the partition of the table.
* We can alter the partition in static partition
* We can get the partition column value from 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
* Where clause is used to use limit in static partition
* You can perform Static partition on Hive Manage table or external table.

Dynamic Partitioning

* 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 need to use where clause to use limit.
* we can’t perform alter on Dynamic partition
* You can perform dynamic partition on hive external table and managed table
* If you want to use Dynamic partition in hive then mode is in nonstrict mode