Hadoop Interview Q&A

Part A:

1. Explain Hadoop and its components.
2. What is the difference between HDFS and relational databases like Oracle?
3. What is the difference between HDFS, HBase and Hive?
4. When and how do you perform data modeling in a Hadoop ecosystem?
5. Explain how data is transferred between HDFS and a relational database (and vice versa).

Part B:

1- What is MapReduce?
MapReduce is the framework in Hadoop where the actual data from the HDFS store gets processed efficiently. MapReduce breaks down a big data processing job into smaller tasks for processing.

2- What is Sqoop?
Sqoop acts as the intermediate layer between the RDBMS and Hadoop to transfer data. It is used to import data from the relational database such as MySQL / Oracle to Hadoop Distributed File System (HDFS) and export data from the Hadoop file system to relational databases.

3- What is the metastore in Hive?
Metastore is the central repository of Apache Hive metadata which stores metadata for Hive in a relational database.

4- What is the use of Hcatalog?
HCatalog is used to share data structures with external systems . It offers access to hive metastore to users of other tools on Hadoop so they can read and write data to hive's data warehouse.

5- Which classes are used by the Hive to Read and Write HDFS files.
TextInputFormat/HiveIgnoreKeyTextOutputFormat: To read/write data in plain text file format.
SequenceFileInputFormat/SequenceFileOutputFormat: To read/write data in Hadoop SequenceFile format.