1. **What are the basic steps (show all steps) in building a parallel program? Show at least one example.**

The basic steps in building a parallel program start by first identifying the set of tasks that are able to run concurrently and partitions of data that are able to be processed concurrently

1. **What is MapReduce?**

MapReduce is a programming model or processing model that uses distributed computing when processing a large amount of raw data.

1. **What is map and what is reduce?**

Map is the stage or method of collecting, filtering and sorting data and reduce refers to an algorithm that is used in summary operations like summing up the number of recurrences in data and adding up data etc

1. **Why MapReduce?**

MapReduce is important because it makes the storing and processing large amounts of data cheaper, simpler and faster by breaking down large datasets into much smaller data to be distributed over many nodes.

1. **Show example for MapReduce**

An example for MapReduce would be trying to find people in a particular age group, 25 for example, it would take a long time to group the millions of records as the dataset grows in magnitude. MapReduce provides a cluster based implementation of processing this large dataset into a distributed manner.

.

1. **Explain in your own words how MapReduce model executed?**

The user inputs data for a MapReduce task, this data is stored in input files, and input files stored in HDFS**.** Input data is received by the master who then assigns the data to phase called workers. Now, workersFormat and define how these input files are split and read, then selects the files used for input, then the inputsplits or shards are created by the InputFormat, the data then will be processed by an individual Mapper or workers**.** One map task is created for each split so number of map tasks will be equal to the number of InputSplits. The record reader communicates with the InputSplit in Hadoop MapReduce and converts the data into [key-value pairs](http://data-flair.training/blogs/key-value-pairs-hadoop-mapreduce/) suitable for reading by the mapper/worker. These map key value pairs are buffered in memory. From time to time the buffered pairs are written to local disk by the petitioner, the pairs are passed back to the master who then passes them back to the reducer. The reducer takes the set of intermediate key-value pairs produced by the mappers as the input and then runs a reducer function on each of them to generate an output. The output is then stored in the HDFS. The reduce workers go over the now sorted intermediate data for each unique key, then passes the unique keys to the user reduce function, the output in this computation is appended to the final output.After all map and reduce tasks have been completed, the master passes back the user code after waking up the user program.

1. **List and describe three examples that are expressed as mapReduce computations.**

Some of the expressed examples are set:

The distributed grep, in which the map function helps with the matching of given

patterns, the reduce function is used as a copy function in this case.

Count of url frequencies - in this case, the map function processes the web page log requests and their outputs <URL, 1>. To add together all the values for the same URL and then output a <URL, total count> pair the reduce function is used.

Reverse web link graphs - the map function here outputs the <target, source> pairs for all the links that connect to a target URL found in the “source” page. The reduce function attaches the list of all source URLs which have the same target URL and outputs the <target, list(source)> pair.

**When do we use OpenMP, MPI, and MapReduce (Hadoop) and why?**

OpenMP- is used for parallelism within a multi-core node. OpenMP assigns the task to multiple threads so to split the work amongst them to make the process faster and more efficient.

MPI - is a message-passing parallel programming model. It is used to communicate information from one process to another through operations on each process. It helps communicate data between multiple machines. MPI is used when trying to develop parallel scientific applications.

MapReduce - used to split the input data set into independent data which are processed in a parallel manner by the map function and then the reduce function is applied to the output of that. MapReduce is used especially when there is a large amount of data to process.

1. **In your own words, explain what a Drug Design and DNA problem (no more than 150 words)**

The given problem is very similar to the MapReduce function because the purpose of the Drug Design and DNA problem is to generate ligands and to test them individually and then concatenate the outputs because the computing of scores of these various ligands are distributed to multiple threads to run parallely. And then to search for the ligand with the highest match.