Part -A

1. What is The Globus Toolkit Architecture (GT4)

The Globus Toolkit, started in 1995 with funding from DARPA, is an open middleware library for the grid computing communities. The toolkit addresses common problems and issues related to grid resource discovery,management, communication, security, fault detection, and portability. The library includes a rich set of service implementations.

2. What is GT4 library?

The high-level services and tools, such as MPI, Condor-G, and Nirod/G, are developed by third parties for generalpurpose distributed computing applications. The local services, such as LSF, TCP, Linux, and Condor, are at the bottom level and are fundamental tools supplied by other developers.

3. What is meant by Globus Container ?

The Globus Container provides a basic runtime environment for hosting the web services needed to execute grid jobs.

4. What are the Functional Modules in Globus GT4 Library ?

- Global Resource Allocation Manager
- Communication
- Grid Security Infrastructure
- Monitory and Discovery Service
- Health and Status
- Global Access of Secondary Storage
- Grid File Transfer

5. What is meant by input splitting ?

For the framework to be able to distribute pieces of the job to multiple machines, it needs to fragment the input into individual pieces, which can in turn be provided as input to the individual distributed tasks. Each fragment of input is called an input split.

6. What are the five categories of Globus Toolkit 4 ?

- Common runtime components
- Security
- Data management
- Information services
- Execution management

7. What are the are the available input formats?

- KeyValueTextInputFormat
- TextInputFormant
- NLineInputFormat
- MultiFileInputFormat
- SequenceFIleInputFormat

8. What is meant by HDFS ?

Hadoop comes with a distributed filesystem called HDFS, which stands for Hadoop Distributed Filesystem. HDFS is a filesystem designed for storing very large files with streaming data access patterns, running on clusters of commodity hardware.

9. What is meant by Block

A disk has a block size, which is the minimum amount of data that it can read or write.Filesystems for a single disk build on this by dealing with data in blocks, which are an integral multiple of the disk block size. Filesystem blocks are typically a few kilobytes in size, while disk blocks are normally 512 bytes. HDFS, too, has the concept of a block, but it is a much larger unit—64 MB by default.

10. Differentiate Namenodes and Datanodes

An HDFS cluster has two types of node operating in a master-worker pattern: a namenode (the master) and a number of datanodes (workers). The namenode manages the filesystem namespace. It maintains the filesystem tree and the metadata for all the files and directories in the tree. This information is stored persistently on the local disk in the form of two files: the namespace image and the edit log. The namenode also knows the datanodes on which all the blocks for a given file are located.

11. List the various Hadoop filesystems ?

Local,HDFS, HFTP, HSFTP, WebHDFS.

12. What is meant by FUSE?

Filesystem in Userspace (FUSE) allows filesystems that are implemented in user space to be integrated as a Unix filesystem. Hadoop's Fuse-DFS contrib module allows any Hadoop filesystem (but typically HDFS) to be mounted as a standard filesystem.

13. What is Hadoop File system ?

Hadoop is written in Java, and all Hadoop filesystem interactions are mediated through the Java API. The filesystem shell, for example, is a Java application that uses the Java FileSystem class to provide filesystem operations.

14. How to Reading Data from a Hadoop URL

One of the simplest ways to read a file from a Hadoop filesystem is by using a java.net.URL object to open a stream to read the data from. The general idiom is:

```
InputStream in = null;
try {
in = new URL("hdfs://host/path").openStream();
// process in} finally {
IOUtils.closeStream(in);
}
```

15. How to write data in Hadoop?

The FileSystem class has a number of methods for creating a file. The simplest is the method that takes a Path object for the file to be created and returns an output stream to write to:

public FSDataOutputStream create(Path f) throws IOException

16. How are Deleting Datas are Deleted in Hadoop ?

Use the delete() method on FileSystem to permanently remove files or directories:

public boolean delete(Path f, boolean recursive) throws IOException

If f is a file or an empty directory, then the value of recursive is ignored.

17. Illustrate MapReduce logical data flow

18. What are two types of nodes that control the job execution process?

a jobtracker and a number of tasktrackers controls the job execution process. The jobtracker coordinates all the jobs run on the system by scheduling tasks to run on tasktrackers. Tasktrackers run tasks and send progress reports to the jobtracker, which keeps a record of the overall progress of each job. If a task fails, the jobtracker can reschedule it on a different tasktracker.

19. Illustrate MapReduce data flow with a single reduce task

20. Illustrate MapReduce data flow with multiple reduce tasks

Part -B

1. Explain the Globus Toolkit Architecture (GT4)
2. Explain MapReduce Model in detail
3. Explain Map & Reduce function?
4. Explain HDFS Concepts in detail?
5. Explain Anatomy of a File Read?
6. Explain Anatomy of a File write?