Big Data Lab04

Group members: Liang Haoxuan, Liu Sihan, Wang Changpeng

Basic Tasks

1/ a/ This decision use Hadoop distributed file system to shorten delays.

Scale-out

Recoverable/reliable

high-speed

b/Slower real-time processing

Complex

Security

2/ a/ Drawback: Large files can slow down node A when reading the file and may also interrupt the reading process

b/ Use HDFS

c/ Data locality refers to the physical proximity between data and its processing logic.In distributed systems, this usually means that tasks should be executed on the nodes that store the data they require whenever possible.

By enhancing data locality, the amount of data transmitted over the network can be reduced, thereby decreasing network bandwidth usage and latency, which in turn improves the execution speed of computational tasks and the overall performance of the system.

3/ i/ Sequential processing refers to the execution of tasks or instructions in a specific order, where each task is completed before moving on to the next one.

ii/ Parallel distributed processing (PDP) is a type of computing where multiple processors work together to complete a task. Each processor has its own local memory and works on a part of the task.

iii/ Use Mapreduce method. Group matrix A and matrix B by Key-value and calculate the value at the same time.

4/ a/ mapreduce, same level part

b/ Hard to process complex algorithm

c/ Different nodes may allocate different amounts of data, and large amounts of data can be costly to transmit

5/ The numbers are divided into groups of key values, and then the groups are divided into 50 nodes. The n-level results are calculated on the nodes again and again.

6/ The operations on the nodes are different.

Medium Tasks

7/a/ The NameNode stores location information, while the DataNode stores the files. When a client needs to access a file, the NameNode returns the file's location information, and the DataNode returns the file.

b/ i/ The file will be split and stored in three blocks, namely S1, S2, and S3. S1 and S2 each store a 64MB file, while S3 stores the remaining 52MB file.

ii/ In HDFS, the default number of replicas of a block is 3. This is achieved by the NameNode allocating the blocks to different DataNodes for storage.

iii/S1：

S1 Replica 1: Node 1

S1 Replica 2: Node 5

S1 Replica 3: Node 7

iv/In order to maintain data availability and integrity, when the NameNode detects that node 5 is about to crash, it will choose a replica from either node 3 or node 7 to replicate and store the copy in a new node, ensuring that S1 has three replicas.

Advanced Tasks

8/ a/ JobTrackr plays a commanding and coordinating role, TaskTracker is the executor of specific tasks. They work together, jointly complete the processing of the MapReduce Job.

b/ Map function is to preprocess and transform the input data.

Reduce function is to summarize and aggregate the intermediate results produced by the Map function.