Student Id: 1605029

Question 7-1

Distribution skew is eliminated since there's more virtual nodes than real nodes. Since we're using Round-Robin partitioning technique to distribute the virtual nodes, partition skew doesn't occur, since the partitions are almost of same size.

In this partitioning sk scheme, virtual nodes can be moved from a heavily loaded (real) node to a less heavily loaded node. In this way, the total load is balanced and execution skew is handled.

Question 7-2

In a parallel storage system, the partitioning table is usually stored in the master node. The table is usually replicated and distributed among routers and client nodes for faster query processing.

This particular table helps routers map virtual nodes to real nodes for the ~~querier~~ queries. When a router accepts a read/write request from a client, it forwards the request to the appropriate real node.

Also, dynamic partitioning consistently changes the

partitioning table, which means each update is applied to all instances of the table. Thus, query diversion is achieved.

Question 8-1

Advantages:

1. Reliability and Availability: We are storing the same data in different locations. Hence, failure of any single site (server) does not make the data unavailable.
2. Fast Response: Queries requesting replicated copies of data are always faster (especially read queries).
3. Less Communication Overhead: If a read query requires local data, we don't need to contact other sites.

Disadvantages:

1. Update: Keeping all data current can be a challenge. The more locations we use to store our data, the more we'll have to implement complex systems to keep track of what's what.
2. Storage: We'll need more storage space as our data continues to grow.

Question 8-2

Using 64 MB block size, we can store the 10 GB file in

$\frac{10GB}{64MB} \approx 160$ blocks

If we allocate 16 blocks to a DataNode, we'll need 10 DataNodes. Additionally, two more sets of 10 DataNodes will be needed. Each of those 10 DataNodes will store replicas of the blocks. Hence, total 30 DataNodes will be required for the job. These nodes can be distributed among 3 racks (10 DataNodes per rack).

The NameNode will contain 160 blockId entries against the given file name.