Big Data Analytics Tutorial 2: Map Reduce and HDFS

1. Map Reduce Task

Problem: Find the maximum monthly rented bikes per each station from the following bike share dataset example.

Input: A set of 6 records with format as: `<station name, bike rented per day month>`

- (ChinaTown, 100), (OxfordSt, 85)
- (OxfordSt, 60), (Abraham, 100)
- (ChurchSt, 80), (ChinaTown, 80)

Your Task: Write down the Map and Reduce data flow steps/diagram to solve this problem.

- Assume we split the input data.

2. HDFS Task

Problem: Suppose we have a cluster with 2 data nodes: DN1 and DN2, and we want to distribute 2 files from a local file system to the Hadoop distributed file system (HDFS) cluster.

- File 1 can be divided into three blocks: B1, B2, and B3.
- File 2 is divided into two blocks: B4 and B5.

Your Task: Distribute each of the blocks B1, B2, B3, B4, and B5 across the data nodes DN1 and DN2.

3. Replication Factor of 2 Task

Problem: Consider the data nodes DN1, DN2, and DN3. You are required to distribute 5 blocks (B1, B2, B3, B4, B5) with a replication factor of 2 (each block should be stored twice among all nodes, no more than once on a single node).

Your Task: Distribute these blocks across DN1, DN2, and DN3 such that each block is replicated twice.

4. Replication Factor of 3 Task

Problem: Now consider a cluster with 3 data nodes: DN1, DN2, and DN3. You are required to distribute 7 blocks (B1, B2, B3, B4, B5, B6, B7) with a replication factor of 3.

Your Task: Distribute these blocks across DN1, DN2, and DN3 such that each block is replicated three times.