Distributed Storage and Parallel Computing

Homework 3

12112627 李乐平

4.1 How to load data into HBase?

Solution:

There are several ways to load data into HBase. A common way is using tools like importtsv or completebultload. Take importtsv as example, first upload the file into HDFS, then call the org.apache.hadoop.hbase.mapreduce.ImportTsv procedure while indicating the separator, column names, table name and filename.

4.2 How to deal with data imbalance (different keys involve different volume of data) in MapReduce?

Solution:

Handling data imbalance in MapReduce involves various strategies based on different scenarios and configurations. In cases where data distribution is uneven, dynamically adjusting data partitioning strategies during the Map phase is recommended. To address the challenge of large intermediate data volumes, utilizing Combiner functions for local aggregation in Map nodes can reduce the network transmission load. For scenarios where certain nodes become performance bottlenecks, dynamic resource allocation and adjustment, possibly through node scaling, are suggested. In situations where Reducer nodes experience imbalanced workloads, approaches such as redistributing data after the Map phase or increasing the number of Reducer nodes can be effective. Additionally, employing adaptive algorithms that dynamically adjust task scheduling and data partitioning based on real-time conditions offers a more flexible solution to maintain system balance throughout MapReduce execution. The choice of specific methods depends on the nature of the task and characteristics of the data.

4.3 What are the differences between ORDER BY and SORT BY in Hive?

Solution:

SORT BY only guarantees the order per reducer, while ORDER BY will maintain the global order across all reducers.