Distributed Storage and Parallel Computing

Homework 3

12112627 李乐平

3.1 What decides the number of mappers? And What decides the number of reducers?

Solution:

The cluster configuration will mainly decide the upper bound of the number of mappers & reducers. Yet not all mappers and reducers must put into use. The concrete number of mappers depends on the number of input blocks, while the concrete number of reducers depends on the distribution of intermediate key-value pairs. Commonly the number of reducers are 0.95 or 1.75 multiplied by #Nodes * #Maximum Containers per Node.

3.2 It is possible that sometimes the data in different reducers are highly imbalanced. That is, some reducers need to address large volume of data while others do not. What do you think is the reason for this phenomenon? How do you think we can resolve it?

Solution:

The skewed data distribution or inefficient data partitioning would led to imbalance. In this case we can consider the techniques of combining, customizing partitioner or reorganizing the key. Change the number of reducers may also help.

3.3 Please describe the similarity and difference between GET and SCAN of HBase operator.

Solution:

In HBase, both the GET and SCAN operations are used to retrieve data from the distributed and scalable NoSQL database, but they serve different purposes. The GET operation is designed to fetch a specific row based on the provided row key, making it efficient for retrieving individual records. It is suitable for point query where you know the exact row key you want to access. On the other hand, the SCAN operation is more versatile and is used for retrieving a range of rows or the entire table. It allows for scanning through rows based on specified start and stop row keys, column families, and qualifiers, making it suitable for situations where you need to process a subset or the entirety of the data. While GET is focused on fetching specific row, SCAN is designed for broader data retrieval based on specified criteria.