Cloudera Runtime 7.1.8

Tuning Apache Impala

Date published: 2020-11-30
Date modified: 2022-08-25

CLOUDERA

Legal Notice

Contents

Data Cache for Remote Reads

In CDW, frequently accessed data is cached in a storage layer on SSD so that it can be quickly retrieved for subsequent queries.

The ability to quickly retrieve data for subsequent queries boosts performance. Each executor can have up to 200 GB of cache. For example, a Medium-sized Virtual Warehouse can keep 200 * 20 = 4 TB of data in its cache. For columnar formats, such as ORC, data in the cache is decompressed, but not decoded. If the expected size of the hot dataset is 6 TB, which requires about 30 executors, you can over-provision by choosing a Large-sized warehouse to ensure full cache coverage. A case can also be made to under-provision by choosing a Medium-sized warehouse to reduce costs at the expense of having a lower cache hit rate. To offset this, keep in mind that columnar formats allow optimizations such as column projection and predicate push-down, which can significantly reduce cache requirements. In these cases, under-provisioning the cache might have no negative performance effect.