Cloudera Data Engineering 1.25.0

Managing Performance in Cloudera Data Engineering

Date published: 2020-07-30
Date modified: 2025-11-20

CLOUDERA

Legal Notice

Contents

Cloudera Data Engineering performance tuning

Use local SSD

Using default instances without local SSD does impact performance. Instances with local SSDs generally provide faster performance, but you need to ensure you choose instance types that have large enough local storage for intermediate results.

Avoid instance types with low compute profiles

Make sure there is reasonable headroom (at least 20%) between the per executor memory and core usage and what the instance profile provides.

Disable Spark Analysis

By default, Cloudera Data Engineering collects metrics around CPU, memory, and I/O. This can have a negative impact on performance - especially for long running jobs that take several hours to complete. Profiling metrics are useful in development and testing, but an unnecessary overhead in a production setting.

You can turn this feature off using the Spark Analysis toggle on the job Configuration tab.

Use Spark 3 instead of Spark 2

Spark 3 introduced performance improvements over Spark 2, therefore Cloudera recommends using it whenever possible.