Sizing Guidelines

Contents:

This section provides general guidelines for cluster sizing and node requirements for effective use of the Trifacta® platform.

> **NOTE:** These guidelines are rough estimates of what should provide satisfactory performance. You should review particulars of the variables listed below in detail prior to making recommendations or purchasing decisions.

Requirements for the Trifacta node

See *System Requirements*.

Enterprise Hadoop

All compute nodes on the cluster (Hadoop NodeManager nodes) should have identical capabilities. Avoid mixing and matching nodes of different capabilities.

Primary variables affecting cluster size:

- Data volume
- Number of concurrent jobs

In the following table, you can review the recommended number of worker nodes in the cluster based on the data volume and the number of concurrent jobs. Table data assumes that each compute node has 16 compute cores (2 x 8 cores), 128GB of RAM and 8TB of disk, with nodes connected via 10 gigabit Ethernet (GbE).

Data Volume \ Number of concurrent jobs	1	5	10	25
1 GB or less	1	1	1	2
10 GB	1	1	2	5
25 GB	1	2	5	10
50 GB	1	5	10	25
100 GB	2	10	20	50
250 GB	5	25	50	125
500 GB	10	50	100	250
1000 GB (1 TB)	20	100	200	500

Additional variables affecting cluster size:

- If you are working with compressed or binary formats, you should use the expanded sizes for your data volume estimates.
- Some workloads are more compute- or memory-intensive and may increase the required number of nodes or capabilities of each node. These include:
  - Scripts with complex steps such as joins (particularly those between large datasets) and sorts
  - Lengthy scripts
- In high availability mode, the total number of connections across all nodes should meet the appropriate requirements in the above table. For each node, please divide the number of connections by the number of Trifacta nodes.

Amazon

Amazon Marketplace AMI

Amazon Marketplace installations support a limited range of installation options for the AMI. For more information, see the install guide available through the Marketplace for Trifacta Wrangler Pro.

Amazon EMR

> **NOTE:** The sizing guidelines listed for Enterprise Hadoop above provide a good estimate for sizing capacity and upper bounds for EMR-based cluster scaling.

For additional details on sizing your EMR cluster, please contact *Trifacta Customer Success Services*.

Microsoft Azure

Microsoft Azure installations support a limited range of installation options, based on the type of cluster integration.

Cluster Type	Description
HDI	Please use the Enterprise Hadoop guidelines listed previously. For more information on this integration, see *Configure for HDInsight* in the Configuration Guide.
Azure Databricks	Please review the Enterprise Hadoop guidelines with *Trifacta Customer Success Services*. For more information on this integration, see *Configure for Azure Databricks* in the Configuration Guide.