



TRIFACTA

Planning Guide

Version: 9.2
Doc Build Date: 07/29/2022

Copyright © Trifacta Inc. 2022 - All Rights Reserved. CONFIDENTIAL

These materials (the “Documentation”) are the confidential and proprietary information of Trifacta Inc. and may not be reproduced, modified, or distributed without the prior written permission of Trifacta Inc.

EXCEPT AS OTHERWISE PROVIDED IN AN EXPRESS WRITTEN AGREEMENT, TRIFACTA INC. PROVIDES THIS DOCUMENTATION AS-IS AND WITHOUT WARRANTY AND TRIFACTA INC. DISCLAIMS ALL EXPRESS AND IMPLIED WARRANTIES TO THE EXTENT PERMITTED, INCLUDING WITHOUT LIMITATION THE IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT AND FITNESS FOR A PARTICULAR PURPOSE AND UNDER NO CIRCUMSTANCES WILL TRIFACTA INC. BE LIABLE FOR ANY AMOUNT GREATER THAN ONE HUNDRED DOLLARS (\$100) BASED ON ANY USE OF THE DOCUMENTATION.

For third-party license information, please select **About Trifacta** from the Help menu.

1. <i>Install Planning</i>	4
1.1 <i>Product Support Matrix</i>	5
1.2 <i>Product Limitations</i>	11
1.3 <i>System Requirements</i>	15
1.4 <i>Sizing Guidelines</i>	21
1.5 <i>System Ports</i>	23
1.6 <i>System Dependencies</i>	26
1.7 <i>Browser Requirements</i>	35
1.8 <i>Required Users and Groups</i>	38
1.9 <i>Prepare Hadoop for Integration with the Platform</i>	41
1.9.1 <i>Tune Cluster Performance</i>	43

Install Planning

Before you begin installing and deploying Trifacta®, you should review these topics on preparing your environment for Trifacta software installation and integration with your enterprise infrastructure.

Product Support Matrix

Contents:

- *Hosting Infrastructure*
 - *Container deployments*
 - *Platform Integrations*
 - *Cluster Integrations*
 - *On-Premises integrations*
 - *Hadoop Integrations*
 - *AWS Integrations*
 - *Azure Integrations*
 - *Trifacta node*
 - *Trifacta node hardware*
 - *Trifacta node software*
 - *Trifacta databases*
 - *Desktop Browsers*
 - *Connectivity*
-

Before you begin installing Trifacta®, please review the following checklist to verify that the applicable items are available and ready to deploy with the software.

NOTE: Enablement of specific features or integration with external sources may have additional requirements listed in any referenced content. Please be sure to review the Details sections listed below.

NOTE: If the version is listed as `Default`, the supported version is the one that is included with the supported distribution.

For more information on general limitations of your product, see *Product Limitations*.

Hosting Infrastructure

- **On-Premises:**
 - **Cloudera:** Details: *Supported Deployment Scenarios for Cloudera* in the Install Guide.
 - **AWS:** Details: *Supported Deployment Scenarios for AWS* in the Install Guide.
 - **Azure:** Details: *Supported Deployment Scenarios for Azure* in the Install Guide

Container deployments

If you are deploying the Trifacta node to a container, the following versions are supported by the Trifacta platform.

Docker

Supported Versions:

- Docker: 17.12 or higher. Docker version must be compatible with the following version(s) of Docker Compose.
- Docker Compose: 1.24.1

Details: *Install for Docker* in the Install Guide

Platform Integrations

Cluster Integrations

Cluster types

NOTE: Depending on your version of the following cluster platforms, specific versions of Spark may be required.

Hadoop on-premises

Cluster type	Supported Versions	Notes, Limitations and Additional Doc
Cloudera	<ul style="list-style-type: none">CDH 6.3 RecommendedCDH 6.2	<p>NOTE: CDH 6.x requires use of Spark native libraries provided by the cluster. See <i>Configure for Spark</i> in the Configuration Guide.</p> <p>Details: <i>Supported Deployment Scenarios for Cloudera</i> in the Install Guide</p>

Hadoop cloud

Cluster type	Supported Versions	Notes, Limitations and Additional Doc
Cloudera Data Platform	<ul style="list-style-type: none">Cloudera Data Platform 7.1	<p>NOTE: Cloudera Data Platform requires use of Spark native libraries provided by the cluster. See <i>Configure for Spark</i> in the Configuration Guide.</p> <p>Details: <i>Supported Deployment Scenarios for Cloudera</i> in the Install Guide</p>

AWS

Cluster type	Supported Versions	Notes, Limitations and Additional Doc
EMR	<ul style="list-style-type: none">EMR 6.2.1	<p>EMR 6.2.1 is supported only with use of Spark 3.0.1.</p> <p>NOTE: Do not use EMR 6.2.0.</p> <p>Details: <i>Configure for EMR</i> in the Configuration Guide.</p>
EMR	<ul style="list-style-type: none">EMR 5.13 - 5.30.2	<p>EMR 5.28.0 is not supported, due to <i>Spark compatibility issues</i>. Please use 5.28.1 or later.</p> <p>NOTE: Do not use EMR 5.30.0 or 5.30.1.</p> <p>Details: <i>Configure for EMR</i> in the Configuration Guide</p>
AWS Databricks	<ul style="list-style-type: none">AWS Databricks 10.xAWS Databricks 9.1 LTS (Recommended)AWS Databricks 7.3 LTS	<p>Details: <i>Configure for AWS Databricks</i> in the Configuration Guide</p>

For more information, see AWS Integrations below.

Cluster type	Supported Versions	Notes, Limitations and Additional Doc
Azure Databricks	<ul style="list-style-type: none"> Azure Databricks 10.x Azure Databricks 9.1 LTS (Recommended) Azure Databricks 7.3 LTS 	Details: <i>Configure for Azure Databricks</i> in the Configuration Guide

For more information, see Azure Integrations below.

Cluster hardware

See *Sizing Guidelines*.

On-Premises integrations

Base storage layer options

The platform must be configured to integrate with a base storage layer. This layer is used for storage of uploads, samples, and job results. See *Set Base Storage Layer* in the Configuration Guide.

Item	Supported Versions	Notes, Limitations and Additional Doc
HDFS	Default	Details: <i>Configure for Hadoop</i> in the Configuration Guide
S3	n/a	Details: <i>S3 Access</i> in the Configuration Guide

SSO Authentication methods

Item	Supported Versions	Notes, Limitations and Additional Doc
AD-LDAP	n/a	Details: <i>Configure SSO for AD-LDAP</i> in the Configuration Guide
SAML	2.0	Details: <i>Configure SSO for SAML</i> in the Configuration Guide

Hadoop Integrations

Item	Supported Versions	Notes, Limitations and Additional Doc
Hive	<ul style="list-style-type: none"> Hive 1.x Hive 2.x Hive 3.0, 3.1 	Additional support requirements vary with the version of Hive. Details: <i>Configure for Hive</i> in the Configuration Guide
KMS	Default	Additional configuration is required depending on your deployed distribution of Hadoop. Details: <i>Configure for KMS</i> in the Configuration Guide
Sentry	Default	

AWS Integrations

Base storage layer options

Details: *Supported Deployment Scenarios for AWS* in the Install Guide

Item	Supported Versions	Notes, Limitations and Additional Doc
------	--------------------	---------------------------------------

S3	n/a	
HDFS	Default	

SSO Authentication methods

Details: *Configure for AWS* in the Configuration Guide

Item	Supported Versions	Notes, Limitations and Additional Doc
AWS Key-Secret	n/a	
EC2 instance roles	n/a	
IAM roles	n/a	

Azure Integrations

Base storage layer options

Item	Supported Versions	Notes, Limitations and Additional Doc
ADLS Gen2	n/a	Details: <i>ADLS Gen2 Access</i> in the Configuration Guide
ADLS Gen1	n/a	Details: <i>ADLS Gen1 Access</i> in the Configuration Guide
WASBS	n/a	Details: <i>WASB Access</i> in the Configuration Guide

SSO Authentication methods

Item	Supported Versions	Notes, Limitations and Additional Doc
Azure AD	n/a	Details: <i>Configure SSO for Azure AD</i> in the Configuration Guide

Trifacta node

Trifacta node hardware

Tip: For in-place upgrades, there should be at least twice as much available disk space as listed below.

Item	Minimum	Recommended	Notes, Limitations and Additional Doc
Number of Cores	8 cores, x86_64	16 cores, x86_64	
RAM	64 GB	128 GB	
Install disk space	16 GB	24 GB	
Total free disk space	24 GB /opt - 15 GB /var - remainder	100 GB /opt - 15 GB /var - remainder	

Details: *System Requirements*

Trifacta node software

Item	Supported Versions	Notes, Limitations and Additional Doc
Operating System	<ul style="list-style-type: none">CentOS: 7.4 - 7.9, 8.1, 8.4 <div>NOTE: MySQL 5.7 Community is not supported on CentOS/RHEL 8.x.</div> <ul style="list-style-type: none">RHEL: 7.4 - 7.9, 8.1, 8.4Ubuntu: 16.04 (Xenial), 18.04 (Bionic Beaver)	<div>NOTE: There are additional requirements for some of these operation system versions. See <i>System Requirements</i>.</div>
Java	<ul style="list-style-type: none">Java 11 (recommended)Java 8	
NginX	1.20.1	
NodeJS	16.14.0	

Other requirements:

- Edge node:** Platform must be installed on an edge node of the cluster.
- Root access:** Required for installation
- SSL access:** Access to the platform can be limited to SSL only. See *Install SSL Certificate* in the Install Guide.
- Internet access:** If the Trifacta node is not connected to the Internet, you must acquire additional software packages for the installation process. See *Install Dependencies without Internet Access* in the Install Guide.

See *System Requirements*.

Trifacta databases

The Trifacta platform requires multiple databases to store object metadata and job information. Supported databases:

Item	Supported Versions	Notes, Limitations and Additional Doc
PostgreSQL	12.X 11.X (Azure installs only)	<div>NOTE: Beginning in this release, the latest stable release of PostgreSQL 12 can be installed with the Trifacta platform. Earlier versions of PostgreSQL 12.X can be installed manually.</div> <div>NOTE: PostgreSQL 11 is supported for Azure installs only.</div>
MySQL	5.7 Community <div>NOTE: MySQL 5.7 Community is not supported on CentOS /RHEL 8.x.</div>	Details: <i>System Requirements</i>

See *Install Databases* in the Databases Guide.

Desktop Browsers

Desktop hardware

Item	Supported Versions	Notes, Limitations and Additional Doc
Screen	1280 x 720 pixels is recommended	

Desktop browsers

NOTE: Stable browser versions released after a given release of Trifacta will **NOT** be supported for any prior version of Trifacta. A best effort will be made to support newer versions released during the support lifecycle of the release.

Item	Supported Versions	Notes, Limitations and Additional Doc
Google Chrome	v.91 - v.93, and any stable version that is released prior to the next release of Trifacta.	
Mozilla Firefox	v.90 - v.92, and any stable version that is released prior to the next release of Trifacta.	
Microsoft Edge	v.91 - v.93, and any stable version that is released prior to the next release of Trifacta.	NOTE: This feature is in Beta release.

For more information, see *Browser Requirements*.

Connectivity

For more information, see *Connection Types* in the User Guide.

Product Limitations

Contents:

- *General Limitations*
 - *Data Volume*
 - *Sampling*
 - *Internationalization*
 - *Size Limits*
 - *Limitations by Integration*
 - *General*
 - *LDAP*
 - *Hadoop*
 - *Amazon AMI*
 - *Amazon EMR*
 - *Microsoft Azure*
 - *Redshift*
 - *S3*
 - *Hive*
 - *Spark*
 - *JDBC*
 - *Other Limitations*
-

This section covers key known limitations of Trifacta®.

NOTE: This list of limitations should not be considered complete.

General Limitations

Data Volume

The Trifacta application applies no fixed limits to the number of columns or rows that can be handled during transformation.

NOTE: During transformation, Trifacta is designed to process data volumes of any size.

However, some important considerations:

Soft row limits

- The number of rows that you see within the Trifacta application in the currently selected sample is determined by:
 - Maximum permitted sample size stored on the base storage layer
 - Currently configured sample size for the current recipe

See Sampling below.

Soft column limits

- Soft row limits do not affect the number of columns that are displayed. All available and visible columns are displayed. The number of rows may be affected by the number of columns.

Tip: Avoid creating and working with datasets that are wider than 1000 columns. Datasets that are wider than this recommendation may result in performance impacts in the Trifacta application.

- The number of columns may be limited by:
 - Number of columns permitted in the source datastore.
 - For SQL-based datastores, limits may be placed on the length of individual queries.

Sampling

- Sample sizes are defined by parameter for each available running environment. See *Sample Size Limits* be low.
- All values displayed or generated in the application are based on the currently displayed sample.
 - Transforms that generate new data may not factor values that are not present in the current sample.
 - When the job is executed, transforms are applied across all rows and values in the source data.
 - Transforms that make changes based on data values, such as `header` and `valuestocols`, will still be configured according to sample data at the time of that the step was added, instead at execution time. For example, all of the values detected in the sample are used to determine the columns of a `valuestocols` transform step based on the selected sample when the step was added.
- Random samples are derived from up to the first 1 GB of the source file.
 - Data from later parts of a multi-part file may not be included in the sample.

Internationalization

- The product supports a variety of global file encoding types for import.

For more information, see *Configure Global File Encoding Type* in the Configuration Guide.

- Within the application, UTF-8 encodings are displayed.
 - Limited set of characters allowed in column names.
 - Header does not support all UTF-8 characters.
 - Emoji are not supported in data wrangling operations.
 - Umlauts and other international characters are not supported when filtering datasets in browsers of external datastores.
- States and Zip Code Column Types and the corresponding maps in visual profiling apply only to the United States.
- UTF-8 is generated in output.
- UTF-32 encoding is not supported

NOTE: Some functions do not correctly account for multi-byte characters. Multi-byte metadata values may not be consistently managed.

Size Limits

Job Size Limits

Execution on a Spark running environment is recommended for any files over 5GB in net data size, including join keys.

Limitations by Integration

General

The product requires definition of a base storage layer, which can be HDFS or S3 for this version. This base storage layer must be defined during install and cannot be changed after installation. See *Set Base Storage Layer* in the Configuration Guide.

LDAP

- If LDAP integration is enabled, the LDAP user [`ldap.user` (default=`trifacta`)] should be created in the same realm.
- See *Configure SSO for AD-LDAP* in the Configuration Guide.

Hadoop

- Trifacta Self-Managed Enterprise Edition requires an integration with a working Hadoop cluster.
- See *Running Environment Options* in the Configuration Guide.

Amazon AMI

- For more information, see product documentation on the Amazon Marketplace.

Amazon EMR

- For more information, see product documentation on the Amazon Marketplace.

Microsoft Azure

- For more information, see product documentation on the Azure Marketplace.

Redshift

None.

S3

- S3 integration is supported only over AWS-hosted instances of S3.
- Oracle Java Runtime 1.8 must be installed on the node hosting the product.
- Writing to S3 requires use of S3 as the base storage layer. For more information, see *Set Base Storage Layer* in the Configuration Guide.
- When publishing single files to S3, you cannot apply an `append` publishing action.

Hive

- Only HiveServer2 is supported.
- You can create only one connection of this type.
- When reading from a partitioned table, the product reads from all partitions, which impacts performance.
- For more information, see *Configure for Hive* in the Configuration Guide.

Spark

- None.

JDBC

- The product supports explicit versions of each JDBC source. See *Connection Types* in the Configuration Guide.
- Additional installation may be required. Additional limitations may apply.
- See *Relational Access* in the Configuration Guide.

Other Limitations

- **File Formats:** Limitations may apply to individual file formats. See *Supported File Formats*.
- **Data Type Conversions:** There are some limitations on how data types are converted during import or export/publication. See *Type Conversions*.

System Requirements

Contents:

- *Platform Node Requirements*
 - *Node Installation Requirements*
 - *Hardware Requirements*
 - *Operating System Requirements*
 - *Database Requirements*
 - *Other Software Requirements*
 - *Root User Access*
 - *SSL Access*
 - *Internet Access*
 - *Hadoop Cluster Requirements*
 - *Supported Hadoop Distributions*
 - *Node Requirements*
 - *Hadoop Component Access*
 - *Hadoop System Ports*
 - *Site Configuration Files*
 - *Security Requirements*
 - *Cluster Configuration*
 - *User Requirements*
-

This section contains hardware and software requirements for successful installation of Trifacta®.

Platform Node Requirements

Node Installation Requirements

If the Trifacta platform is installed in a Hadoop environment, the software must be installed on an edge node of the cluster.

- If it is integrated with a Cloudera cluster, it must be installed on a gateway node that is managed by Cloudera Manager.
 - If it is integrated with Cloudera Data Platform, it must be installed on an edge node of the cluster.
- Customers who originally installed an earlier version on a non-edge node will still be supported. If the software is not installed on an edge node, you may be required to copy over files from the cluster and to synchronize these files after upgrades. The cluster upgrade process is more complicated.
- This requirement does not apply to the following cluster integrations:
 - AWS EMR
 - Azure Databricks

NOTE: If you are installing the Trifacta platform into a Docker container, a different set of requirements apply. For more information, see *Install for Docker* in the Install Guide.

Hardware Requirements

Tip: For in-place upgrades, there should be at least twice as much available disk space as listed below.

Minimum hardware:

Item	Required
Number of cores	8 cores, x86_64
RAM	64 GB The platform requires 24 GB of dedicated RAM to start and perform basic operations.
Disk space to install software	16 GB
Total free disk space	24 GB Space requirements by volume: <ul style="list-style-type: none"> • /opt - 15 GB • /var - Remainder

Recommended hardware:

Item	Recommended
Number of cores	16 cores, x86_64
RAM	128 GB The platform requires 24 GB of dedicated RAM to start and perform basic operations.
Disk space to install software	24 GB
Total free disk space	100 GB Space requirements by volume: <ul style="list-style-type: none"> • /opt - 15 GB • /var - Remainder

Operating System Requirements

The following operating systems are supported for the Trifacta node. The Trifacta platform requires 64-bit versions of any supported operating system.

CentOS/RHEL versions:

- CentOS 7.4 - 7.9, 8.1, 8.4

NOTE: MySQL 5.7 Community is not supported on CentOS/RHEL 8.x.

- RHEL 7.4 - 7.9, 8.1, 8.4

Notes on CentOS/RHEL installation:

- Installation on CentOS/RHEL versions 7.4 or earlier requires an upgrade of the RPM software on the Trifacta node. Details are provided during the installation process.
- Disabling SELinux on the Trifacta node is recommended. However, if security policies require it, you may need to apply some changes to the environment.

Ubuntu versions:

- Ubuntu 18.04 (codename Bionic Beaver)
- Ubuntu 16.04 (codename Xenial)

Notes on Ubuntu installation:

- For Ubuntu installations, some packages must be manually installed. Instructions are provided later in the process.

For more information on RPM dependencies, see *System Dependencies*.

Database Requirements

The following database versions are supported by the Trifacta platform for storing metadata and the user's Wrangle recipes.

Supported database versions:

- PostgreSQL 12.X

NOTE: Beginning in this release, the latest stable release of PostgreSQL 12 can be installed with the Trifacta platform. Earlier versions of PostgreSQL 12.X can be installed manually.

- PostgreSQL 11.X

NOTE: PostgreSQL 11 is supported for Azure installs only. Installation on Azure requires PostgreSQL 11. Please follow the database installation instructions for PostgreSQL 12, modifying them for version 11.

- MySQL 5.7 Community

NOTE: MySQL 5.7 Community is not supported on CentOS/RHEL 8.x.

Notes on database versions:

- MySQL 5.7 is not supported for installation in Amazon RDS.

NOTE: If you are installing or upgrading a deployment of Trifacta that uses or will use a remote database service, such as Amazon RDS, for hosting the Trifacta databases, please contact *Alteryx Customer Success Services*. For this release, additional configuration may be required.

- If you are installing the databases into MySQL, you must download and install the MySQL Java driver onto the Trifacta node. For more information, see *Install Databases for MySQL* in the Databases Guide.
- H2 database type is used for internal testing. It is not a supported database.

For more information on installing and configuring the database, see *Install Databases* in the Databases Guide.

Other Software Requirements

The following software components must be present.

Java

Where possible, you should install the same version of Java on the Trifacta node and on the cluster with which you are integrating.

- Java 11 (**runtime only**)
- Java 8

Notes on Java versions:

- OpenJDK 8 is supported.

NOTE: If you are using Azure Databricks as a datasource, please verify that openJDKv1.8.0_302 or earlier is installed on the Trifacta node. Java 8 is required. There is a known issue with TLS v1.3.

- There are additional requirements related to Java JDK listed in the Hadoop Components section listed below.
- If you are integrating your Trifacta instance with S3, you must install the Oracle JRE 1.8 onto the Trifacta node. No other version of Java is supported for S3 integration. For more information, see *S3 Access* in the Configuration Guide.

Other Software

For Ubuntu installations, the following packages must be manually installed using Ubuntu-specific versions:

- NginX: 1.20.1
- NodeJS 16.14.0

Instructions and version numbers are provided later in the process.

Root User Access

Installation must be executed as the root user on the Trifacta node.

SSL Access

(Optional) If users are connecting to the Trifacta platform, an SSL certificate must be created and deployed. See *Install SSL Certificate* in the Install Guide.

Internet Access

(Optional) Internet access is not required for installation or operation of the platform. However, if the server does not have Internet access, you must acquire additional software as part of the disconnected install. For more information, see *Install Dependencies without Internet Access* in the Install Guide.

Hadoop Cluster Requirements

The following requirements apply if you are integrating the Trifacta platform with an enterprise Hadoop cluster.

- For general guidelines on sizing the cluster, see *Sizing Guidelines*.
- If you have upgrades to the Hadoop cluster planned for the next year, you should review those plans with Support prior to installation. For more information, please contact *Alteryx Support*.

Supported Hadoop Distributions

The Trifacta platform supports the following minimum Hadoop distributions.

- The Trifacta platform only supports the latest major release and its minor releases of each distribution.
- The Trifacta platform only supports the versions of any required components included in a supported distribution. Even if they are upgraded components, use of non-default versions of required components is not supported.

Cloudera supported distributions

- CDH 6.3 **Recommended**
- CDH 6.2

NOTE: CDH 6.x requires that you use the native Spark libraries provided by the cluster. Additional configuration is required. For more information, see *Configure for Spark* in the Configuration Guide.

- Cloudera Data Platform 7.1

See *Supported Deployment Scenarios for Cloudera* in the Install Guide.

EMR supported distributions

See *Configure for EMR* in the Configuration Guide.

AWS Databricks supported distributions

See *Configure for AWS Databricks* in the Configuration Guide.

Azure Databricks supported distributions

See *Configure for Azure Databricks* in the Configuration Guide.

Node Requirements

Each cluster node must have the following software:

- Java JDK 8 (some exceptions may be listed below)

Hadoop Component Access

The Trifacta deployment must have access to the following.

Java and Spark version requirements

The following matrix identifies the supported versions of Java and Spark on the Hadoop cluster. Where possible, you should install the same version of Java on the Trifacta node and on the cluster with which you are integrating.

Notes:

- Java must be installed on each node of the cluster. For more information, see https://www.cloudera.com/documentation/enterprise/latest/topics/cdh_ig_jdk_installation.html.
- The versions of Java on the Trifacta node and the Hadoop cluster do not have to match.

	Spark 2.3	Spark 2.4	Spark 3.0.1
Java 8	Required.	Required.	Required

- Support for Spark 3.0.1 has limitations. See *Configure for Spark* in the Configuration Guide.

- If you are integrating with an EMR cluster, there are specific version requirements for EMR. See *Configure for Spark* in the Configuration Guide.

Other components

- HDFS Namenode
 - WebHDFS
 - In HDFS, Append Mode must be enabled. See *Prepare Hadoop for Integration with the Platform*.
 - If you are enabling high availability failover, you must use HttpFS, instead of WebHDFS. See *Enable Integration with Cluster High Availability* in the Configuration Guide.
- For YARN:
 - ResourceManager is running.
 - ApplicationMaster's range of ephemeral ports are open to the Trifacta node.
- HiveServer2:
 - HiveServer2 is supported for metadata publishing.
 - WebHCat is not supported.

Hadoop System Ports

For more information, see *System Ports*.

Site Configuration Files

Hadoop cluster configuration files must be copied into the Trifacta deployment. See *Configure for Hadoop* in the Configuration Guide.

Security Requirements

- **Kerberos supported:**
 - If Kerberos is enabled, a keytab file must be accessible to the Trifacta platform.
 - See *Configure for Kerberos Integration* in the Configuration Guide.
- **If Kerberos and secure impersonation are not enabled:**
 - A user [`hadoop.user` (default=`trifacta`)] must be created on each node of the Hadoop cluster.
 - A directory [`hadoop.dir` (default=`trifacta`)] must be created on the cluster.
 - The user [`hadoop.user`] must have full access to the directory. which enables storage of the transformation recipe back into HDFS.
 - See *Configure for Hadoop* in the Configuration Guide.

Cluster Configuration

For more information on integration with Hadoop, see *Prepare Hadoop for Integration with the Platform*.

User Requirements

Users must access the Trifacta platform through one of the supported browser versions. For more information on user system requirements, see *Browser Requirements*.

Sizing Guidelines

Contents:

- *Requirements for the Trifacta node*
- *Self-Managed Hadoop*
- *Amazon*
 - *Amazon Marketplace AMI*
 - *Amazon EMR*
- *Microsoft Azure*

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*.

Self-Managed 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 *Alteryx 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
Azure Databricks	<p>Please review the Enterprise Hadoop guidelines with <i>Alteryx Customer Success Services</i>.</p> <p>For more information on this integration, see <i>Configure for Azure Databricks</i> in the Configuration Guide.</p>

System Ports

Contents:

- *Trifacta® node Ports*
 - *Internal Service Ports*
 - *Database Ports*
 - *Client Browser Ports*
 - *Hadoop Ports*
 - *Firewall Ports for Hadoop*
 - *EMR Ports*
-

Trifacta® node Ports

Depending on the components enabled or integrated with your instance of the platform, the following ports must be opened on the Trifacta node.

Internal Service Ports

Component	Port
Nginx Proxy	3005
Trifacta application	3006
Java UDF Service	3008
Spark Job Service	4007
Supervisor	4421
ML-Service	5000
Data Service	41912
Java VFS Service	41917
Batch Job Runner	41920
VFS Service	41913
Conversion Service	41914
Job Metadata Service	41915
Artifact Storage Service	41916
Batch Job Runner	41920
Secure Token Service	41921
Connector Configuration Service	41925
Orchestration Service	42424
Time-based trigger Service	43033
Scheduling Service	43143

Database Ports

Component	Port
Postgres	5432

(default)	<div> NOTE: By default, PostgreSQL and the platform use port 5432 for communication. If that port is not available at install/upgrade time, the next available port is used, which is typically 5433. This change may occur if a previous version of PostgreSQL is on the same server. When a non-default port number is used, the platform must be configured to use it. For more information, see <i>Change Database Port</i>. </div>
MySQL	3306

Client Browser Ports

By default, the web client uses port 3005.

NOTE: Any client firewall software must be configured to enable access on this port.

This port can be changed. For more information, see *Change Listening Port* in the Install Guide.

Hadoop Ports

If Trifacta is integrated with a Hadoop cluster, the Trifacta node must have access to the following Hadoop components. Their default ports are listed below.

NOTE: These ports vary between installations. Please verify your environment's ports.

NOTE: In addition to the following ports, you must open any additional ports on Trifacta node for other components and services that are not listed here and are used for running jobs on the running environment cluster.

Hadoop Component	Default Port
HDFS Namenode	Cloudera/HDP: 8020
HDFS Datanode	50020
NOTE: The Trifacta node must be able to access this port on all HDFS datanodes of the cluster.	
HttpFS	14000
WebHDFS	Cloudera/HDP: 50070
YARN ResourceManager	Cloudera: 8032 HDP: 8050
JobTracker	Cloudera/HDP: 8021
HiveServer2 (optional)	TCP connection: 10000 HTTP connection: 10001
Hive Metastore (optional)	9083

Firewall Ports for Hadoop

If the Trifacta node is on a different network from the Hadoop cluster, please verify that these additional ports are opened on the firewall.

Hadoop Component	Default Port
YARN Resourcemanager Scheduler	8030
YARN Resourcemanager Admin	8033
YARN Resourcemanager WebApp	8088
YARN Nodemanager WebApp	8042
YARN Timeline Service	8188
MapReduce JobHistory Server	10020
HDFS DataNode	50010

For additional details, please refer to the documentation provided with your Hadoop distribution.

EMR Ports

If you are integrating with an EMR cluster, please verify that the following nodes and ports are available to the Trifacta node.

EMR Component	Port
EMR master node	8088

System Dependencies

Contents:

- *Direct Dependencies*
 - CentOS/Redhat 7
 - CentOS/Redhat 8
 - Ubuntu16.04
 - Ubuntu 18.04
 - *Direct and Indirect Dependencies*
 - CentOS/Redhat 7
 - CentOS/Redhat 8
 - Ubuntu 16.04
 - Ubuntu 18.04
-

The following direct and indirect dependencies apply to the Trifacta software that is installed on the edge node for each supported version of the operating system.

NOTE: When dependencies are acquired for versions of Ubuntu, the operating system grabs the latest version, even if it is later than the version on which the software is dependent. In some cases, this mismatch can result in installation errors, which can be fixed by manually installing the dependency with the correct version.

Direct Dependencies

These direct dependencies are packaged with the Trifacta® installer.

CentOS/Redhat 7

supervisor = 3.2.4
nodejs = 2:16.14.0-1nodesource
nginx = 1:1.20.1-1.el7.ngx
gzip >= 1.3.12
bzip2 >= 1.0.5
libgcc >= 4.8.2
liberation-sans-fonts

CentOS/Redhat 8

supervisor = 4.1.0
nodejs = 2:16.14.0-1nodesource
nginx = 1:1.20.1-1.el8.ngx
gzip >= 1.9-9
bzip2 >= 1.0.6-26
libgcc >= 4.8.2
liberation-sans-fonts
libXext
libSM
libXrender

Ubuntu16.04

supervisor = 3.2.4
nodejs = 16.14.0-1nodesource1
nginx = 1.20.1-1~xenial
rlwrap >= 0.37-5
gzip >= 1.4-1ubuntu2
bzip2 >= 1.0.6-1
libgcc1 >= 6.0.1-0ubuntu1
build-essential
libglib2.0-0
libsm6
libxext6
libxrender-dev

Ubuntu 18.04

python-supervisor = 3.2.4
nodejs = 16.14.0-1nodesource1
nginx = 1.12.2-1~trusty
rlwrap = 0.37-5
gzip >= 1.4-1ubuntu2
bzip2 >= 1.0.6-1
libgcc1 >= 4.9.3-0ubuntu4

Direct and Indirect Dependencies

This full list of dependencies is applied during online installs or is included in the offline install package provided to you:

CentOS/Redhat 7

bzip2-1.0.6-13.el7.x86_64.rpm
fontpackages-filesystem-1.44-8.el7.noarch.rpm
java-1.8.0-openjdk-1.8.0.312.b07-1.el7_9.x86_64.rpm
java-1.8.0-openjdk-devel-1.8.0.312.b07-1.el7_9.x86_64.rpm
java-1.8.0-openjdk-headless-1.8.0.312.b07-1.el7_9.x86_64.rpm
liberation-fonts-common-1.07.2-16.el7.noarch.rpm
liberation-sans-fonts-1.07.2-16.el7.noarch.rpm

libc-2.27-4.el7.x86_64.rpm
make-3.82-24.el7.x86_64.rpm
nginx-1.20.1-1.el7ngx.x86_64.rpm
nodejs-16.14.0-1nodesource.x86_64.rpm
openssl-1.0.2k-24.el7_9.x86_64.rpm
openssl-libs-1.0.2k-24.el7_9.x86_64.rpm
postgresql12-12.6-1PGDG.rhel7.x86_64.rpm
postgresql12-libs-12.6-1PGDG.rhel7.x86_64.rpm
postgresql12-server-12.6-1PGDG.rhel7.x86_64.rpm
postgresql96-9.6.10-1PGDG.rhel7.x86_64.rpm
postgresql96-libs-9.6.10-1PGDG.rhel7.x86_64.rpm
postgresql96-server-9.6.10-1PGDG.rhel7.x86_64.rpm
python-backports-1.0-8.el7.x86_64.rpm
python-backports-ssl_match_hostname-3.5.0.1-1.el7.noarch.rpm
python-ipaddress-1.0.16-2.el7.noarch.rpm
python-meld3-0.6.10-1.el7.x86_64.rpm
python-setuptools-0.9.8-7.el7.noarch.rpm
supervisor-3.2.4-1.noarch.rpm

systemd-219-78.el7_9.5.x86_64.rpm
systemd-libs-219-78.el7_9.5.x86_64.rpm
systemd-sysv-219-78.el7_9.5.x86_64.rpm

CentOS/Redhat 8

alsa-lib-1.2.5-4.el8.x86_64.rpm
atk-2.28.1-1.el8.x86_64.rpm
avahi-libs-0.7-20.el8.x86_64.rpm
bzip2-1.0.6-26.el8.x86_64.rpm
cairo-1.15.12-3.el8.x86_64.rpm
chkconfig-1.19.1-1.el8.x86_64.rpm
copy-jdk-configs-4.0-2.el8.noarch.rpm
crypto-policies-20210617-1.gitc776d3e.el8.noarch.rpm
crypto-policies-scripts-20210617-1.gitc776d3e.el8.noarch.rpm
cups-libs-2.2.6-40.el8.x86_64.rpm
dejavu-fonts-common-2.35-7.el8.noarch.rpm
dejavu-sans-fonts-2.35-7.el8.noarch.rpm
fontconfig-2.13.1-4.el8.x86_64.rpm
fontpackages-filesystem-1.44-22.el8.noarch.rpm
freetype-2.9.1-4.el8_3.1.x86_64.rpm
fribidi-1.0.4-8.el8.x86_64.rpm
gdk-pixbuf2-2.36.12-5.el8.x86_64.rpm
gdk-pixbuf2-modules-2.36.12-5.el8.x86_64.rpm
giflib-5.1.4-3.el8.x86_64.rpm
graphite2-1.3.10-10.el8.x86_64.rpm
gtk-update-icon-cache-3.22.30-8.el8.x86_64.rpm
gtk2-2.24.32-5.el8.x86_64.rpm
harfbuzz-1.7.5-3.el8.x86_64.rpm
hicolor-icon-theme-0.17-2.el8.noarch.rpm
jasper-libs-2.0.14-5.el8.x86_64.rpm
java-1.8.0-openjdk-1.8.0.312.b07-2.el8_5.x86_64.rpm
java-1.8.0-openjdk-devel-1.8.0.312.b07-2.el8_5.x86_64.rpm
java-1.8.0-openjdk-headless-1.8.0.312.b07-2.el8_5.x86_64.rpm
javapackages-filesystem-5.3.0-1.module_el8.0.0+11+5b8c10bd.noarch.rpm
jbigkit-libs-2.1-14.el8.x86_64.rpm
libICE-1.0.9-15.el8.x86_64.rpm
libSM-1.2.3-1.el8.x86_64.rpm
libX11-1.6.8-5.el8.x86_64.rpm
libX11-common-1.6.8-5.el8.noarch.rpm
libXau-1.0.9-3.el8.x86_64.rpm
libXcomposite-0.4.4-14.el8.x86_64.rpm
libXcursor-1.1.15-3.el8.x86_64.rpm
libXdamage-1.1.4-14.el8.x86_64.rpm
libXext-1.3.4-1.el8.x86_64.rpm
libXfixes-5.0.3-7.el8.x86_64.rpm
libXft-2.3.3-1.el8.x86_64.rpm
libXi-1.7.10-1.el8.x86_64.rpm
libXinerama-1.1.4-1.el8.x86_64.rpm
libXrandr-1.5.2-1.el8.x86_64.rpm
libXrender-0.9.10-7.el8.x86_64.rpm
libXtst-1.2.3-7.el8.x86_64.rpm
libdatrie-0.2.9-7.el8.x86_64.rpm
liberation-fonts-common-2.00.3-7.el8.noarch.rpm
liberation-sans-fonts-2.00.3-7.el8.noarch.rpm
libfontenc-1.1.3-8.el8.x86_64.rpm
libicu-60.3-2.el8_1.x86_64.rpm
libjpeg-turbo-1.5.3-12.el8.x86_64.rpm
libpkgconf-1.4.2-1.el8.x86_64.rpm

libpng-1.6.34-5.el8.x86_64.rpm
libthai-0.1.27-2.el8.x86_64.rpm
libtiff-4.0.9-20.el8.x86_64.rpm
libxcb-1.13.1-1.el8.x86_64.rpm
lksctp-tools-1.0.18-3.el8.x86_64.rpm
lua-5.3.4-12.el8.x86_64.rpm
lua-libs-5.3.4-12.el8.x86_64.rpm
nginx-1.20.1-1.el8ngx.x86_64.rpm
nodejs-16.14.0-1nodesource.x86_64.rpm
nspr-4.32.0-1.el8_4.x86_64.rpm
nss-3.67.0-7.el8_5.x86_64.rpm
nss-softokn-3.67.0-7.el8_5.x86_64.rpm
nss-softokn-freebl-3.67.0-7.el8_5.x86_64.rpm
nss-sysinit-3.67.0-7.el8_5.x86_64.rpm
nss-util-3.67.0-7.el8_5.x86_64.rpm
pango-1.42.4-8.el8.x86_64.rpm
pixman-0.38.4-1.el8.x86_64.rpm
pkgconf-1.4.2-1.el8.x86_64.rpm
pkgconf-m4-1.4.2-1.el8.noarch.rpm
pkgconf-pkg-config-1.4.2-1.el8.x86_64.rpm
platform-python-pip-9.0.3-20.el8.noarch.rpm
postgresql12-12.6-1PGDG.rhel8.x86_64.rpm
postgresql12-libs-12.6-1PGDG.rhel8.x86_64.rpm
postgresql12-server-12.6-1PGDG.rhel8.x86_64.rpm
postgresql96-9.6.17-1PGDG.rhel8.x86_64.rpm
postgresql96-libs-9.6.17-1PGDG.rhel8.x86_64.rpm
postgresql96-server-9.6.17-1PGDG.rhel8.x86_64.rpm
python3-pip-9.0.3-20.el8.noarch.rpm
python3-setuptools-39.2.0-6.el8.noarch.rpm
python36-3.6.8-38.module_el8.5.0+895+a459eca8.x86_64.rpm
shared-mime-info-1.9-3.el8.x86_64.rpm
supervisor-4.1.0-1.noarch.rpm
ttmkfdir-3.0.9-54.el8.x86_64.rpm
tzdata-java-2021e-1.el8.noarch.rpm
xorg-x11-font-utils-7.5-41.el8.x86_64.rpm
xorg-x11-fonts-Type1-7.5-19.el8.noarch.rpm

Ubuntu 16.04

build-essential_12.1ubuntu2_amd64.deb
bzip2_1.0.6-8ubuntu0.2_amd64.deb
ca-certificates-java_20160321ubuntu1_all.deb
fontconfig-config_2.11.94-0ubuntu1.1_all.deb
fontconfig_2.11.94-0ubuntu1.1_amd64.deb
fonts-dejavu-core_2.35-1_all.deb
fonts-dejavu-extra_2.35-1_all.deb
hicolor-icon-theme_0.15-0ubuntu1.1_all.deb
java-common_0.56ubuntu2_all.deb
libasound2-data_1.1.0-0ubuntu1_all.deb
libasound2_1.1.0-0ubuntu1_amd64.deb
libasyncns0_0.8-5build1_amd64.deb
libatk1.0-0_2.18.0-1_amd64.deb
libatk1.0-data_2.18.0-1_all.deb
libavahi-client3_0.6.32~rc+dfsg-1ubuntu2.3_amd64.deb
libavahi-common-data_0.6.32~rc+dfsg-1ubuntu2.3_amd64.deb
libavahi-common3_0.6.32~rc+dfsg-1ubuntu2.3_amd64.deb
libcairo2_1.14.6-1_amd64.deb
libcups2_2.1.3-4ubuntu0.11_amd64.deb
libdatrie1_0.2.10-2_amd64.deb

libdrm-amdgpu1_2.4.91-2~16.04.1_amd64.deb
libdrm-common_2.4.91-2~16.04.1_all.deb
libdrm-intel1_2.4.91-2~16.04.1_amd64.deb
libdrm-nouveau2_2.4.91-2~16.04.1_amd64.deb
libdrm-radeon1_2.4.91-2~16.04.1_amd64.deb
libdrm2_2.4.91-2~16.04.1_amd64.deb
libelf1_0.165-3ubuntu1.2_amd64.deb
libflac8_1.3.1-4_amd64.deb
libfontconfig1_2.11.94-0ubuntu1.1_amd64.deb
libfreetype6_2.6.1-0.1ubuntu2.5_amd64.deb
libgdk-pixbuf2.0-0_2.32.2-1ubuntu1.6_amd64.deb
libgdk-pixbuf2.0-common_2.32.2-1ubuntu1.6_all.deb
libgif7_5.1.4-0.3~16.04.1_amd64.deb
libgl1-mesa-dri_18.0.5-0ubuntu0~16.04.1_amd64.deb
libgl1-mesa-glx_18.0.5-0ubuntu0~16.04.1_amd64.deb
libglapi-mesa_18.0.5-0ubuntu0~16.04.1_amd64.deb
libgraphite2-3_1.3.10-0ubuntu0.16.04.1_amd64.deb
libgtk2.0-0_2.24.30-1ubuntu1.16.04.2_amd64.deb
libgtk2.0-bin_2.24.30-1ubuntu1.16.04.2_amd64.deb
libgtk2.0-common_2.24.30-1ubuntu1.16.04.2_all.deb
libharfbuzz0b_1.0.1-1ubuntu0.1_amd64.deb
libjbig0_2.1-3.1_amd64.deb
libjpeg-turbo8_1.4.2-0ubuntu3.4_amd64.deb
libjpeg8_8c-2ubuntu8_amd64.deb
liblcms2-2_2.6-3ubuntu2.1_amd64.deb
libllvm6.0_1%3a6.0-1ubuntu2~16.04.1_amd64.deb
libnspr4_2%3a4.13.1-0ubuntu0.16.04.1_amd64.deb
libnss3-nssdb_2%3a3.28.4-0ubuntu0.16.04.14_all.deb
libnss3_2%3a3.28.4-0ubuntu0.16.04.14_amd64.deb
libogg0_1.3.2-1_amd64.deb
libpango-1.0-0_1.38.1-1_amd64.deb
libpangocairo-1.0-0_1.38.1-1_amd64.deb
libpangoft2-1.0-0_1.38.1-1_amd64.deb
libpciaccess0_0.13.4-1_amd64.deb
libpcsclite1_1.8.14-1ubuntu1.16.04.1_amd64.deb
libpixmap-1-0_0.33.6-1_amd64.deb
libpng12-0_1.2.54-1ubuntu1.1_amd64.deb
libpq5_13.3-1.pgdg16.04+1_amd64.deb
libpulse0_1%3a8.0-0ubuntu3.15_amd64.deb
libsm6_2%3a1.2.2-1_amd64.deb
libsndfile1_1.0.25-10ubuntu0.16.04.3_amd64.deb
libthai-data_0.1.24-2_all.deb
libthai0_0.1.24-2_amd64.deb
libtiff5_4.0.6-1ubuntu0.8_amd64.deb
libtxc-dxtn-s2tc0_0~git20131104-1.1_amd64.deb
libvorbis0a_1.3.5-3ubuntu0.2_amd64.deb
libvorbisenc2_1.3.5-3ubuntu0.2_amd64.deb
libwrap0_7.6.q-25_amd64.deb
libx11-6_2%3a1.6.3-1ubuntu2.2_amd64.deb
libx11-data_2%3a1.6.3-1ubuntu2.2_all.deb
libx11-xcb1_2%3a1.6.3-1ubuntu2.2_amd64.deb
libxau6_1%3a1.0.8-1_amd64.deb
libxcb-dri2-0_1.11.1-1ubuntu1_amd64.deb
libxcb-dri3-0_1.11.1-1ubuntu1_amd64.deb
libxcb-glx0_1.11.1-1ubuntu1_amd64.deb
libxcb-present0_1.11.1-1ubuntu1_amd64.deb
libxcb-render0_1.11.1-1ubuntu1_amd64.deb
libxcb-shm0_1.11.1-1ubuntu1_amd64.deb
libxcb-sync1_1.11.1-1ubuntu1_amd64.deb
libxcb1_1.11.1-1ubuntu1_amd64.deb

libxcomposite1_1%3a0.4.4-1_amd64.deb
libxcursor1_1%3a1.1.14-1ubuntu0.16.04.2_amd64.deb
libxdamage1_1%3a1.1.4-2_amd64.deb
libxdmcp6_1%3a1.1.2-1.1_amd64.deb
libxext6_2%3a1.3.3-1_amd64.deb
libxfixed3_1%3a5.0.1-2_amd64.deb
libxi6_2%3a1.7.6-1_amd64.deb
libxinerama1_2%3a1.1.3-1_amd64.deb
libxrandr2_2%3a1.5.0-1_amd64.deb
libxrender-dev_1%3a0.9.9-0ubuntu1_amd64.deb
libxrender1_1%3a0.9.9-0ubuntu1_amd64.deb
libxshmfence1_1.2-1_amd64.deb
libxtst6_2%3a1.2.2-1_amd64.deb
libxxf86vm1_1%3a1.1.4-1_amd64.deb
nginx_1.20.1-1~xenial_amd64.deb
nodejs_16.14.0-1nodesource1_amd64.deb
openjdk-8-jre-headless_8u292-b10-0ubuntu1~16.04.1_amd64.deb
openjdk-8-jre_8u292-b10-0ubuntu1~16.04.1_amd64.deb
postgresql-12_12.7-1.pgdg16.04+1_amd64.deb
postgresql-9.6_9.6.22-1.pgdg16.04+1_amd64.deb
postgresql-client-12_12.7-1.pgdg16.04+1_amd64.deb
postgresql-client-9.6_9.6.22-1.pgdg16.04+1_amd64.deb
postgresql-contrib-9.6_9.6.22-1.pgdg16.04+1_amd64.deb
rlwrap_0.41-1build1_amd64.deb
supervisor_3.2.4_all.deb
tcpd_7.6.q-25_amd64.deb
x11-common_1%3a7.7+13ubuntu3.1_all.deb

Ubuntu 18.04

adwaita-icon-theme_3.28.0-1ubuntu1_all.deb
at-spi2-core_2.28.0-1_amd64.deb
build-essential_12.4ubuntu1_amd64.deb
ca-certificates-java_20180516ubuntu1~18.04.1_all.deb
cron_3.0pl1-128.1ubuntu1_amd64.deb
dbus_1.12.2-1ubuntu1.2_amd64.deb
fontconfig-config_2.12.6-0ubuntu2_all.deb
fontconfig_2.12.6-0ubuntu2_amd64.deb
fonts-dejavu-core_2.37-1_all.deb
fonts-dejavu-extra_2.37-1_all.deb
gtk-update-icon-cache_3.22.30-1ubuntu4_amd64.deb
hicolor-icon-theme_0.17-2_all.deb
humanity-icon-theme_0.6.15_all.deb
java-common_0.68ubuntu1~18.04.1_all.deb
libapparmor1_2.12-4ubuntu5.1_amd64.deb
libasound2-data_1.1.3-5ubuntu0.6_all.deb
libasound2_1.1.3-5ubuntu0.6_amd64.deb
libasyncns0_0.8-6_amd64.deb
libatk-bridge2.0-0_2.26.2-1_amd64.deb
libatk-wrapper-java-jni_0.33.3-20ubuntu0.1_amd64.deb
libatk-wrapper-java_0.33.3-20ubuntu0.1_all.deb
libatk1.0-0_2.28.1-1_amd64.deb
libatk1.0-data_2.28.1-1_all.deb
libatspi2.0-0_2.28.0-1_amd64.deb
libavahi-client3_0.7-3.1ubuntu1.3_amd64.deb
libavahi-common-data_0.7-3.1ubuntu1.3_amd64.deb
libavahi-common3_0.7-3.1ubuntu1.3_amd64.deb
libbsd0_0.8.7-1ubuntu0.1_amd64.deb
libcairo2_1.15.10-2ubuntu0.1_amd64.deb

libcommon-sense-perl_3.74-2build2_amd64.deb
libcroco3_0.6.12-2_amd64.deb
libcups2_2.2.7-1ubuntu2.8_amd64.deb
libdatrie1_0.2.10-7_amd64.deb
libdbus-1-3_1.12.2-1ubuntu1.2_amd64.deb
libdrm-amdgpu1_2.4.101-2~18.04.1_amd64.deb
libdrm-common_2.4.101-2~18.04.1_all.deb
libdrm-intel1_2.4.101-2~18.04.1_amd64.deb
libdrm-nouveau2_2.4.101-2~18.04.1_amd64.deb
libdrm-radeon1_2.4.101-2~18.04.1_amd64.deb
libdrm2_2.4.101-2~18.04.1_amd64.deb
libedit2_3.1-20170329-1_amd64.deb
libelf1_0.170-0.4ubuntu0.1_amd64.deb
libflac8_1.3.2-1_amd64.deb
libfontconfig1_2.12.6-0ubuntu2_amd64.deb
libfontenc1_1%3a1.1.3-1_amd64.deb
libfreetype6_2.8.1-2ubuntu2.1_amd64.deb
libgail-common_2.24.32-1ubuntu1_amd64.deb
libgail18_2.24.32-1ubuntu1_amd64.deb
libgdbm-compat4_1.14.1-6_amd64.deb
libgdbm5_1.14.1-6_amd64.deb
libgdk-pixbuf2.0-0_2.36.11-2_amd64.deb
libgdk-pixbuf2.0-bin_2.36.11-2_amd64.deb
libgdk-pixbuf2.0-common_2.36.11-2_all.deb
libgif7_5.1.4-2ubuntu0.1_amd64.deb
libgl1-mesa-dri_20.0.8-0ubuntu1~18.04.1_amd64.deb
libgl1-mesa-glx_20.0.8-0ubuntu1~18.04.1_amd64.deb
libgl1_1.0.0-2ubuntu2.3_amd64.deb
libglapi-mesa_20.0.8-0ubuntu1~18.04.1_amd64.deb
libglib2.0-0_2.56.4-0ubuntu0.18.04.9_amd64.deb
libglib2.0-data_2.56.4-0ubuntu0.18.04.9_all.deb
libglvnd0_1.0.0-2ubuntu2.3_amd64.deb
libglx-mesa0_20.0.8-0ubuntu1~18.04.1_amd64.deb
libglx0_1.0.0-2ubuntu2.3_amd64.deb
libgraphite2-3_1.3.11-2_amd64.deb
libgtk2.0-0_2.24.32-1ubuntu1_amd64.deb
libgtk2.0-bin_2.24.32-1ubuntu1_amd64.deb
libgtk2.0-common_2.24.32-1ubuntu1_all.deb
libharfbuzz0b_1.7.2-1ubuntu1_amd64.deb
libice6_2%3a1.0.9-2_amd64.deb
libicu60_60.2-3ubuntu3.2_amd64.deb
libjbig0_2.1-3.1build1_amd64.deb
libjpeg-turbo8_1.5.2-0ubuntu5.18.04.4_amd64.deb
libjpeg8_8c-2ubuntu8_amd64.deb
libjson-perl_2.97001-1_all.deb
libjson-xs-perl_3.040-1_amd64.deb
liblcms2-2_2.9-1ubuntu0.1_amd64.deb
libllvm10_1%3a10.0.0-4ubuntu1~18.04.2_amd64.deb
libllvm6.0_1%3a6.0-1ubuntu2_amd64.deb
libnspr4_2%3a4.18-1ubuntu1_amd64.deb
libnss3_2%3a3.35-2ubuntu2.13_amd64.deb
libogg0_1.3.2-1_amd64.deb
libpango-1.0-0_1.40.14-1ubuntu0.1_amd64.deb
libpangocairo-1.0-0_1.40.14-1ubuntu0.1_amd64.deb
libpangoft2-1.0-0_1.40.14-1ubuntu0.1_amd64.deb
libpciaccess0_0.14-1_amd64.deb
libpcsclite1_1.8.23-1_amd64.deb
libperl5.26_5.26.1-6ubuntu0.5_amd64.deb
libpixmap-1-0_0.34.0-2_amd64.deb
libpng16-16_1.6.34-1ubuntu0.18.04.2_amd64.deb

libpopt0_1.16-11_amd64.deb
libpq5_14.2-1.pgdg18.04+1_amd64.deb
libpulse0_1%3a11.1-1ubuntu7.11_amd64.deb
libsvg2-2_2.40.20-2ubuntu0.2_amd64.deb
libsvg2-common_2.40.20-2ubuntu0.2_amd64.deb
libsensors4_1%3a3.4.0-4_amd64.deb
libsm6_2%3a1.2.2-1_amd64.deb
libsndfile1_1.0.28-4ubuntu0.18.04.2_amd64.deb
libthai-data_0.1.27-2_all.deb
libthai0_0.1.27-2_amd64.deb
libtiff5_4.0.9-5ubuntu0.4_amd64.deb
libtypes-serialiser-perl_1.0-1_all.deb
libvorbis0a_1.3.5-4.2_amd64.deb
libvorbisenc2_1.3.5-4.2_amd64.deb
libwrap0_7.6.q-27_amd64.deb
libx11-6_2%3a1.6.4-3ubuntu0.4_amd64.deb
libx11-data_2%3a1.6.4-3ubuntu0.4_all.deb
libx11-xcb1_2%3a1.6.4-3ubuntu0.4_amd64.deb
libxau6_1%3a1.0.8-1ubuntu1_amd64.deb
libxaw7_2%3a1.0.13-1_amd64.deb
libxcb-dri2-0_1.13-2~ubuntu18.04_amd64.deb
libxcb-dri3-0_1.13-2~ubuntu18.04_amd64.deb
libxcb-glx0_1.13-2~ubuntu18.04_amd64.deb
libxcb-present0_1.13-2~ubuntu18.04_amd64.deb
libxcb-render0_1.13-2~ubuntu18.04_amd64.deb
libxcb-shape0_1.13-2~ubuntu18.04_amd64.deb
libxcb-shm0_1.13-2~ubuntu18.04_amd64.deb
libxcb-sync1_1.13-2~ubuntu18.04_amd64.deb
libxcb1_1.13-2~ubuntu18.04_amd64.deb
libxcomposite1_1%3a0.4.4-2_amd64.deb
libxcursor1_1%3a1.1.15-1_amd64.deb
libxdamage1_1%3a1.1.4-3_amd64.deb
libxdmcp6_1%3a1.1.2-3_amd64.deb
libxext6_2%3a1.3.3-1_amd64.deb
libxfixed3_1%3a5.0.3-1_amd64.deb
libxft2_2.3.2-1_amd64.deb
libxi6_2%3a1.7.9-1_amd64.deb
libxinerama1_2%3a1.1.3-1_amd64.deb
libxml2_2.9.4+dfsg1-6.1ubuntu1.5_amd64.deb
libxmu6_2%3a1.1.2-2_amd64.deb
libxmuu1_2%3a1.1.2-2_amd64.deb
libxpm4_1%3a3.5.12-1_amd64.deb
libxrandr2_2%3a1.5.1-1_amd64.deb
libxrender-dev_1%3a0.9.10-1_amd64.deb
libxrender1_1%3a0.9.10-1_amd64.deb
libxshmfence1_1.3-1_amd64.deb
libxslt1.1_1.1.29-5ubuntu0.2_amd64.deb
libxt6_1%3a1.1.5-1_amd64.deb
libxtst6_2%3a1.2.3-1_amd64.deb
libxv1_2%3a1.0.11-1_amd64.deb
libxxf86dga1_2%3a1.1.4-1_amd64.deb
libxxf86vm1_1%3a1.1.4-1_amd64.deb
locales_2.27-3ubuntu1.5_all.deb
logrotate_3.11.0-0.1ubuntu1_amd64.deb
multiarch-support_2.27-3ubuntu1.5_amd64.deb
netbase_5.4_all.deb
nginx_1.20.1-1~bionic_amd64.deb
nodejs_16.14.0-1nodesource1_amd64.deb
openjdk-8-jre-headless_8u312-b07-0ubuntu1~18.04_amd64.deb
openjdk-8-jre_8u312-b07-0ubuntu1~18.04_amd64.deb

perl-base_5.26.1-6ubuntu0.5_amd64.deb
perl-modules-5.26_5.26.1-6ubuntu0.5_all.deb
perl_5.26.1-6ubuntu0.5_amd64.deb
pgdg-keyring_2018.2_all.deb
postgresql-12_12.10-1.pgdg18.04+1_amd64.deb
postgresql-9.6_9.6.24-1.pgdg18.04+1_amd64.deb
postgresql-client-12_12.10-1.pgdg18.04+1_amd64.deb
postgresql-client-9.6_9.6.24-1.pgdg18.04+1_amd64.deb
postgresql-client-common_238.pgdg18.04+1_all.deb
postgresql-common_238.pgdg18.04+1_all.deb
postgresql-contrib-9.6_9.6.24-1.pgdg18.04+1_amd64.deb
rlwrap_0.43-1_amd64.deb
shared-mime-info_1.9-2_amd64.deb
ssl-cert_1.0.39_all.deb
supervisor_3.2.4_all.deb
sysstat_11.6.1-1ubuntu0.1_amd64.deb
tzdata_2022a-0ubuntu0.18.04_all.deb
ubuntu-mono_16.10+18.04.20181005-0ubuntu1_all.deb
ucf_3.0038_all.deb
x11-common_1%3a7.7+19ubuntu7.1_all.deb
x11-utils_7.7+3build1_amd64.deb
xdg-user-dirs_0.17-1ubuntu1_amd64.deb

Browser Requirements

Contents:

- *Google Chrome Requirements*
 - *Browser versions*
 - *WebAssembly client extension*
 - *Mozilla Firefox Requirements*
 - *Browser versions*
 - *Microsoft Edge Requirements*
 - *Browser versions*
 - *Other Requirements*
 - *Screen*
 - *Ports*
-

These requirements apply to Trifacta®, which interacts with the platform through the browser. Access to the platform requires one of the supported browser versions listed below.

Tip: 64-bit versions of all supported browsers are recommended. Depending on your datasets, you may encounter memory issues running the Trifacta application in a 32-bit browser.

NOTE: Parts of the application may become hidden or distorted unless zoom level is set to 100%.

NOTE: In some cases, ad blocking extensions in your browser, such as Adblock, can interfere with features of the product. If you are experiencing issues with some Trifacta features, you may need to disable any ad blockers.

NOTE: Multiple browser tabs or windows open to different versions of the product is not supported.

Google Chrome Requirements

Browser versions

Version: Google Chrome v.101-v.103, and any stable version that is released prior to the next release of Trifacta.

NOTE: Stable browser versions released after a given release of Trifacta will **NOT** be supported for any prior version of Trifacta. A best effort will be made to support newer versions released during the support lifecycle of the release.

NOTE: Mobile browsers and Google Chromebook are not supported.

For more information, please see the requirements for installing and using the browser on your operating system:
<https://support.google.com/chrome/a/answer/7100626?hl=en>.

WebAssembly client extension

No other configuration is required.

Limitations:

In this release, the following limitations apply to use of WebAssembly:

- The current implementation of WebAssembly in this release is single-threaded, and performance may be impacted. When multi-threading is available, the Trifacta Photon client will feature multi-threading.
- Progress bars are not displayed for actions in the Transformer page. This is a known issue.

Mozilla Firefox Requirements

Browser versions

Mozilla Firefox v.101-v.103, and any stable version that is released prior to the next release of Trifacta.

NOTE: Stable browser versions released after a given release of Trifacta will **NOT** be supported for any prior version of Trifacta. A best effort will be made to support newer versions released during the support lifecycle of the release.

For more information, please see the requirements for installing and using the browser on your operating system:
<https://www.mozilla.org/en-US/firefox/releases/>.

Microsoft Edge Requirements

NOTE: This feature is in Beta release.

Browser versions

See Google Chrome above.

NOTE: Stable browser versions released after a given release of Trifacta will **NOT** be supported for any prior version of Trifacta. A best effort will be made to support newer versions released during the support lifecycle of the release.

For more information, please see the requirements for installing and using the browser on your operating system:
<https://docs.microsoft.com/en-us/previous-versions/windows/edge-legacy/about-microsoft-edge>.

Other Requirements

The following requirements also apply.

Screen

- Screen resolution of 1280 x 720 is recommended.

Ports

By default, the web client uses port 3005. For more information on required client ports, see *System Ports*.

Required Users and Groups

Contents:

- *Installation node*
 - *Install*
 - *Running Services*
 - *Active Directory/LDAP*
 - *Databases*
 - *Main database*
 - *Jobs database*
 - *Scheduling database*
 - *Time-based Trigger database*
 - *Configuration Service database*
 - *Artifact Storage Service database*
 - *Job Metadata Service database*
 - *Hadoop*
 - *Hadoop User*
 - *Kerberos*
 - *Hive*
-

The following users may be required for installation of the Trifacta® platform and integration with other components in the environment. In some cases, you must also designate a group in which the user or users must belong.

NOTE: Except as noted, you may substitute your own usernames for the default usernames. These substitutions are identified in the documentation references.

In this sections below, you can review the user requirements for various aspects of platform installation and integration.

Legend:

- **Required configuration:** If Yes, then the configuration and the relevant user are required for all installations of the platform.
- **Default user:** Default or expected username for the user.
- **Documentation reference:** How the user is referenced in the documentation.

Installation node

Install

NOTE: The software must be installed on the node using the `root` account.

Running Services

After installation, you can run the platform as the `trifacta` user.

Active Directory/LDAP

When enabling Single Sign-On, you must specify an Active Directory user to serve as the admin for provisioning users within the Trifacta platform.

- **Required configuration:** No
- **Defaults:**
 - User: `trifacta`
 - Group: `trifactausers`
- **Documentation reference:**
 - User: `[ldap.user]`
 - Group: `[ldap.group]`

Databases

The Trifacta platform installs and maintains two databases.

Main database

The Main database is used for managing Trifacta metadata.

- **Required configuration:** Yes
- **Default user:** `trifacta`
- **Documentation reference:** `[db.main.user]`

Jobs database

The Jobs database is used for tracking batch execution jobs initiated by the platform.

- **Required configuration:** Yes
- **Default user:** `trifactaactivities`
- **Documentation reference:** `[db.jobs.user]`

Scheduling database

Storage of schedules, including datasets to execute.

- **Required configuration:** Yes
- **Default user:** `trifactascheduling-service`
- **Documentation reference:** `[db.scheduling.user]`

Time-based Trigger database

Storage of triggering information.

- **Required configuration:** Yes
- **Default user:** `trifactatimebasedtriggerservice`
- **Documentation reference:** `[db.tbts.user]`

Configuration Service database

Storage of parameter settings at the workspace level.

- **Required configuration:** Yes
- **Default user:** `trifactaconfiguration-service`
- **Documentation reference:** `[db.configuration.user]`

Artifact Storage Service database

Storage for feature-specific usage data such value mappings.

- **Required configuration:** Yes
- **Default user:** `trifactaartifactstorageservice`
- **Documentation reference:** `[db.artifact.user]`

Job Metadata Service database

Storage of metadata on job execution.

- **Required configuration:** Yes
- **Default user:** `trifactajobmetadataservice`
- **Documentation reference:** `[db.metadata.user]`

Hadoop

Hadoop User

When the platform interacts with the Hadoop cluster, all actions are brokered through the use of a single Hadoop user account.

NOTE: This user account is specified and used in multiple configurations for integration with the Hadoop cluster.

- **Required configuration:** Yes
- **Defaults:**
 - User: `trifacta`
 - Group: `trifactausers`
- **Documentation references:**
 - User: `[hadoop.user]`
 - Group: `[hadoop.group]`

Kerberos

If Kerberos is enabled on your cluster, you must specify the principal of the Hadoop user for the Trifacta platform. Depending on the other components available in the cluster, you may need to specify other Kerberos principals.

- **Required configuration:** No
- **Default user:** `trifacta`
- **Documentation reference:** `[hadoop.user.principal]`

Hive

You must specify a user that Hive uses to connect to HDFS.

- **Required configuration:** No
- **Defaults:**
 - User: `hive`
 - Group: `trifactausers`
- **Documentation references:**
 - User: `[hive.user]`
 - Group: `[hive.group]`

Prepare Hadoop for Integration with the Platform

Contents:

- *Create Trifacta user account on Hadoop cluster*
- *HDFS directories*
- *Kerberos authentication*
- *Acquire cluster configuration files*

Before you deploy the Trifacta® software, you should complete the following configuration steps within your Hadoop environment.

Create Trifacta user account on Hadoop cluster

The Trifacta platform interacts with Hadoop through a single system user account. A user for the platform must be added to the cluster.

NOTE: In a cluster without Kerberos or SSO user management, the `[hadoop.user (default=trifacta)]` user must be created on each node of the cluster.

If LDAP is enabled, the `[hadoop.user]` user should be created in the same realm as the cluster.

If Kerberos is enabled, the `[hadoop.user]` user must exist on every node where jobs run.

For POSIX-compliant Hadoop environments, the user IDs of the Trifacta user accessing the cluster and the Hadoop user must match exactly.

UserID:

If possible, please create the user ID as: `trifacta`

This user should belong to the group: `trifactausers`

User requirements:

- Access to HDFS
- Permission to run YARN jobs on the cluster.

Verify that the following HDFS paths have been created and that their permissions enable access to the Trifacta user account:

NOTE: Depending on your Hadoop distribution, you may need to modify the following commands to use the Hadoop client installed on the Trifacta node.

Below, change the values for `trifacta` to match the `[hadoop.user]` user for your environment:

```
hdfs dfs -mkdir /trifacta
hdfs dfs -chown trifacta /trifacta
hdfs dfs -mkdir -p /user/trifacta
hdfs dfs -chown trifacta /user/trifacta
```

HDFS directories

The following directories must be available to the `[hadoop.user]` on HDFS. Below, you can review the minimum permissions set for basic and impersonated authentication for each default directory. Secure impersonation is described later.

NOTE: Except for the `dictionaries` directory, which is used to hold smaller reference files, each of these directories should be configured to permit storage of a user's largest datasets.

Directory	Minimum required permissions	Secure impersonation permissions
<code>/trifacta/uploads</code>	700	770 Set this to 730 to prevent users from browsing this directory.
<code>/trifacta/queryResults</code>	700	770
<code>/trifacta/dictionaries</code>	700	770
<code>/trifacta/tempfiles</code>	770	770

You can use the following commands to configure permissions on these directories. Following permissions scheme reflects the secure impersonation permissions in the above table:

```
$ hdfs dfs -mkdir -p /trifacta/uploads
$ hdfs dfs -mkdir -p /trifacta/queryResults
$ hdfs dfs -mkdir -p /trifacta/dictionaries
$ hdfs dfs -mkdir -p /trifacta/tempfiles
$ hdfs dfs -chown -R trifacta:trifacta /trifacta
$ hdfs dfs -chmod -R 770 /trifacta
$ hdfs dfs -chmod -R 730 /trifacta/uploads
```

If these standard locations cannot be used, you can configure the HDFS paths. You can apply this change through the *Admin Settings Page* (recommended) or `trifacta-conf.json`. For more information, see *Platform Configuration Methods*.

```
"hdfs.pathsConfig.fileUpload": "/trifacta/uploads",
"hdfs.pathsConfig.batchResults": "/trifacta/queryResults",
"hdfs.pathsConfig.dictionaries": "/trifacta/dictionaries",
```

Kerberos authentication

The Trifacta platform supports Kerberos authentication on Hadoop.

NOTE: If Kerberos is enabled for the Hadoop cluster, the keytab file must be made accessible to the Trifacta platform. See *Configure for Kerberos Integration* in the Configuration Guide.

Acquire cluster configuration files

The Hadoop cluster configuration files must be made available to the Trifacta platform. You can either copy the files over from the cluster or create a local symlink to them.

For more information, see *Configure for Hadoop* in the Configuration Guide.

Tune Cluster Performance

Contents:

- *YARN Tuning Overview*
- *Spark Tuning Overview*
 - *Spark Performance Considerations*
 - *Limiting Resource Utilization of Spark Jobs*
- *Tuning Recommendations*
- *Spark Job Property Overrides*

This section contains information on how you can tune your Hadoop cluster and Spark specifically for optimal performance in job execution.

YARN Tuning Overview

This section provides an overview of configuration recommendations to be applied to the Hadoop cluster from the Trifacta platform.

NOTE: The recommendations in this section are optimized for use with the Trifacta platform. These may or may not conform to requirements for other applications using the Hadoop cluster. Alteryx assumes no responsibility for the configuration of the cluster.

YARN manages cluster resources (CPU and memory) by running all processes within allocated containers. Containers restrict the resources available to its process(es). Processes are monitored and killed if they overrun the container allocation.

- Multiple containers can run on a cluster node (if available resources permit).
- A job can request and use multiple containers across the cluster.
- Container requests specify virtual CPU (cores) and memory (in MB).

YARN configuration specifies:

- **Per Cluster Node:** Available virtual CPUs and memory per cluster node
- **Per Container:** virtual CPUs and memory for each container

The following parameters are available in `yarn-site.xml`:

Parameter	Type	Description
<code>yarn.nodemanager.resource.memory-mb</code>	Per Cluster Node	Amount of physical memory, in MB, that can be allocated for containers
<code>yarn.nodemanager.resource.cpu-vcores</code>	Per Cluster Node	Number of CPU cores that can be allocated for containers
<code>yarn.scheduler.minimum-allocation-mb</code>	Per Container	Minimum container memory, in MBs; requests lower than this will be increased to this value
<code>yarn.scheduler.maximum-allocation-mb</code>	Per Container	Maximum container memory, in MBs; requests higher than this will be capped to this value
<code>yarn.scheduler.increment-allocation-mb</code>	Per Container	Granularity of container memory requests
<code>yarn.scheduler.minimum-</code>	Per	Minimum allocation virtual CPU cores per container; requests lower than

allocation-vcores	Container	will increased to this value.
yarn.scheduler.maximum-allocation-vcores	Per Container	Maximum allocation virtual CPU cores per container; requests higher than this will be capped to this value
yarn.scheduler.increment-allocation-vcores	Per Container	Granularity of container virtual CPU requests

Spark Tuning Overview

Spark processes run multiple executors per job. Each executor must run within a YARN container. Therefore, resource requests must fit within YARN's container limits.

Like YARN containers, multiple executors can run on a single node. More executors provide additional computational power and decreased runtime.

Spark's dynamic allocation adjusts the number of executors to launch based on the following:

- job size
- job complexity
- available resources

You can apply this change through the *Admin Settings Page* (recommended) or `trifacta-conf.json`. For more information, see *Platform Configuration Methods*.

The per-executor resource request sizes can be specified by setting the following properties in the `spark.props` section :

NOTE: In `trifacta-conf.json`, all values in the `spark.props` section must be quoted values.

Parameter	Description
<code>spark.executor.memory</code>	Amount of memory to use per executor process (in a specified unit)
<code>spark.executor.cores</code>	Number of cores to use on each executor - limit to 5 cores per executor for best performance

A single special process, the application driver, also runs in a container. Its resources are specified in the `spark.props` section:

Parameter	Description
<code>spark.driver.memory</code>	Amount of memory to use for the driver process (in a specified unit)
<code>spark.driver.cores</code>	Number of cores to use for the driver process

Spark Performance Considerations

Optimizing "Small" Joins

Broadcast, or map-side, joins materialize one side of the join and send it to all executors to be stored in memory. This technique can significantly accelerate joins by skipping the sort and shuffle phases during a "reduce" operation. However, there is also a cost in communicating the table to all executors. Therefore, only "small" tables should be considered for broadcast join. The definition of "small" is set by the `spark.sql.autoBroadcastJoinThreshold` parameter which can be added to the `spark.props` section of `trifacta-conf.json`. By default, Spark sets this to 10485760 (10MB).

NOTE: We recommend setting this parameter between 20 and 100MB. It should not exceed 200MB.

Checkpointing

In Spark's driver process, the transformation pipeline is compiled down to Spark code and optimized. This process can sometimes fail or take an inordinately long time. By checkpointing the execution, Spark is forced to materialize the current table (in memory or on disk), thereby simplifying the segments that are optimized. While checkpointing can incur extra cost due to this materialization, it can also reduce end-to-end execution time by speeding up the compilation and optimization phases and by reusing materialized columns downstream.

NOTE: To increase the checkpointing frequency, set `transformer.dataframe.checkpoint.threshold` in the `spark.props` section of `trifacta-conf.json`.

Limiting Resource Utilization of Spark Jobs

With Spark's dynamic allocation, each job's resource utilization can be limited by setting the maximum number of executors per job. Set `spark.dynamicAllocation.maxExecutors` in the `spark.props` section of `trifacta-conf.json`. When applied, the maximum job memory is then given (approximately due to small overhead added by YARN) by:

```
spark.dynamicAllocation.maxExecutors * (spark.driver.memory + spark.executor.memory)
```

The maximum number of cores used per job is given (exactly) by:

```
spark.dynamicAllocation.maxExecutors * (spark.driver.cores + spark.executor.cores)
```

To limit the overall cluster utilization of Trifacta jobs, YARN queues should be configured and used by the application.

Tuning Recommendations

The following configuration settings can be applied through Trifacta platform configuration based on the number of nodes in the Hadoop cluster.

NOTE: These recommendations should be modified based on the technical capabilities of your network, the nodes in the cluster, and other applications using the cluster.

	1	2	4	10	16
Available memory (GB)	16	32	64	160	256
Available vCPUs	4	8	16	40	64
<code>yarn.nodemanager.resource.memory-mb</code>	12288	24576	57344	147456	245760
<code>yarn.nodemanager.resource.cpu-vcores</code>	3	6	13	32	52
<code>yarn.scheduler.minimum-allocation-mb</code>	1024	1024	1024	1024	1024
<code>yarn.scheduler.maximum-allocation-mb</code>	12288	24576	57344	147456	245760
<code>yarn.scheduler.increment-allocation-mb</code>	512	512	512	512	512
<code>yarn.scheduler.minimum-allocation-vcores</code>	1	1	1	1	1

yarn.scheduler.maximum-allocation-vcores	3	6	13	32	52
yarn.scheduler.increment-allocation-vcores	1	1	1	1	1
spark.executor.memory	6GB	6GB	16GB	20GB	20GB
spark.executor.cores	2	2	4	5	5
spark.driver.memory	4GB	4GB	4GB	4GB	4GB
spark.driver.cores	1	1	1	1	1

The specified configuration allows, maximally, the following Spark configuration per node:

CoresxNode	Configuration Options
1x1	(1 driver + 1 executor) or 1 executor
2x1	(1 driver + 2 executor) or 3 executors
4x1	(1 driver + 3 executors) or 3 executors
10x1	(1 driver + 6 executors) or 6 executors
16x1	(1 driver + 10 executors) or 10 executors

Spark Job Property Overrides

You can enable a set of Spark properties that users are permitted to override on individual jobs. For more information, see *Enable Spark Job Overrides*.



Copyright © 2022 - Trifacta, Inc.
All rights reserved.