



TRIFACTA

Planning Guide

Version: 6.8.2
Doc Build Date: 04/14/2020

Copyright © Trifacta Inc. 2020 - 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. *Install Planning* . 4
 - 1.1 *Product Support Matrix* . . 4
 - 1.2 *Product Limitations* . 9
 - 1.3 *System Requirements* 12
 - 1.4 *Sizing Guidelines* . 18
 - 1.5 *System Ports* . 19
 - 1.6 *System Dependencies* 22
 - 1.7 *Desktop Requirements* 25
 - 1.8 *Supported File Formats* . 28
 - 1.9 *Required Users and Groups* . 31
 - 1.10 *Prepare Hadoop for Integration with the Platform* . 35
 - 1.10.1 *Tune Cluster Performance* . 37

Install Planning

Before you begin installing and deploying Trifacta® Wrangler Enterprise, 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*
 - *Desktops*
 - *Connectivity*
-

Before you begin installing Trifacta® Wrangler Enterprise, 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.
 - **Hortonworks:** Details: *Supported Deployment Scenarios for Hortonworks* 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.

On-premises

Cluster type	Supported Versions	Notes, Limitations and Additional Doc
Cloudera	<ul style="list-style-type: none">• CDH 6.3 Recommended• CDH 6.2• CDH 6.1• CDH 5.16 Recommended	<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>
Hortonworks	<ul style="list-style-type: none">• HDP 3.1 Recommended• HDP 3.0• HDP 2.6	<p>NOTE: HDP 3.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 Hortonworks</i> in the Install Guide</p>

AWS

Cluster type	Supported Versions	Notes, Limitations and Additional Doc
EMR	<ul style="list-style-type: none">• EMR 5.8 - 5.27	Details: <i>Configure for EMR</i> in the Configuration Guide

For more information, see AWS Integrations below.

Azure

Cluster type	Supported Versions	Notes, Limitations and Additional Doc
Azure Databricks	<ul style="list-style-type: none">Azure Databricks 5.3 - 5.5 LTS	Details: <i>Configure for Azure Databricks</i> in the Configuration Guide
HDI	<ul style="list-style-type: none">HDI 3.6	Details: <i>Supported Deployment Scenarios for Azure</i> in the Install 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>Enable 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.xHive 2.x	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	
Ranger	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	n/a	Details: <i>Enable ADLS Access</i> in the Configuration Guide
WASBS	n/a	Details: <i>Enable 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

Item	Minimum	Recommended	Notes, Limitations and Additional Doc
Number of Cores	8 cores	16 cores	
RAM	64 GB	128 GB	
Install disk space	4 GB	16 GB	
Total free disk space	16 GB /opt - 10 GB /var - remainder	100 GB /opt - 10 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: 6.4 - 6.x, 7.1, 7.2, 7.4 - 7.6RHEL: 6.4 - 6.x, 7.1, 7.2, 7.4 - 7.6Ubuntu: 14.04 (Trusty), 16.04 (Xenial)	NOTE: There are additional requirements for some of these operation system versions. See <i>System Requirements</i> .
Java	Java 1.8	
NginX	1.12.2	
NodeJS	10.13.0	

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	9.6	
MySQL	5.7 Community	Details: <i>System Requirements</i>

See *Install Databases* in the Databases Guide.

Desktops

For more information on the following requirements, see *Desktop Requirements*.

Desktop hardware

Item	Supported Versions	Notes, Limitations and Additional Doc
CPU	<ul style="list-style-type: none">Intel Pentium 4 or newerAMD Opteron	
RAM	<ul style="list-style-type: none">8 GB recommended4 GB minimum	
Screen	1280 x 720 pixels or better	
Connection	DSL or better	

Desktop browsers

NOTE: Stable browser versions released after a given release of Trifacta Wrangler Enterprise will **NOT** be supported for any prior version of Trifacta Wrangler Enterprise. 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.76 - v.78, and any stable version that is released prior to the next release of Trifacta Wrangler Enterprise.	
Mozilla Firefox	v.68 - v.70, and any stable version that is released prior to the next release of Trifacta Wrangler Enterprise.	NOTE: This feature is in Beta release.

Desktop application

For Trifacta Wrangler Enterprise, the Wrangler Enterprise desktop application can be deployed to desktops that do not have a supported browser installed on them, provided that they meet the following requirements:

Item	Supported Versions	Notes, Limitations and Additional Doc
RAM	8 GB minimum	
Operating System	Windows 7 (Service Pack 1), 8 or 10 (64-bit only)	
Packages	Visual C++ for Visual Studio 2015.	Download: https://www.microsoft.com/en-us/download/details.aspx?id=48145

Connectivity

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

Product Limitations

Contents:

- *General Limitations*
 - *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® Wrangler Enterprise.

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

General Limitations

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

Sample Size Limits

Defaults for each running environment:

- For the Trifacta Photon running environment, samples are limited to 10 MB.

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 Wrangler Enterprise 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.
 - Some additional installation may be required.
 - See *Enable Relational Connections* in the Configuration Guide.
- Jobs for JDBC sources must be executed on the Trifacta Photon running environment.
- Writing to JDBC sources is not supported in this release.

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* in the User Guide.

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*
 - *I/O Requirements*
-

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

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 a Hortonworks cluster, it must be installed on an Ambari/Hadoop client that is managed by Hortonworks Ambari.
- If it is integrated with an HDI cluster, it must be installed on an edge node.
- 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

Minimum hardware:

Item	Required
Number of cores	8 cores, x86_64
RAM	64 GB The platform requires 12GB of dedicated RAM to start and perform basic operations.
Disk space to install software	4 GB
Total free disk space	16 GB Space requirements by volume: <ul style="list-style-type: none">• /opt - 10 GB• /var - Remainder

Recommended hardware:

Item	Recommended
Number of cores	16 cores, x86_64
RAM	128 GB The platform requires 12GB of dedicated RAM to start and perform basic operations.
Disk space to install software	16 GB

Total free disk space	100 GB Space requirements by volume: <ul style="list-style-type: none">• /opt - 10 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 6.4 - 6.x, 7.1, 7.2, 7.4 - 7.6
- RHEL 6.4 - 6.x, 7.1, 7.2, 7.4 - 7.6

Notes on CentOS/RHEL installation:

- If you are installing on CentOS/RHEL 7.1, you must be connected to an online repository for some critical updates. Offline installation is not supported for these operating system distributions.
- For security reasons, RHEL 7.3 is not supported for installation of Release 5.0 or later of the Trifacta platform. Please upgrade to RHEL 7.4 or a later supported release.
- Installation on CentOS/RHEL versions 7.4 or earlier may require 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 14.04 (codename Trusty) and 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 9.6
- MySQL 5.7 Community

Notes on database versions:

- MySQL 5.7 is not supported for installation in Amazon RDS.
- 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 1.8

Notes on Java versions:

- OpenJDK 1.8 is officially supported. It is installed on the Trifacta node during the installation process.
- 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 *Enable S3 Access* in the Configuration Guide.

Other Software

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

- NginX 1.12.2
- NodeJS 10.13.0

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 *Trifacta 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
- CDH 6.1

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.

- CDH 5.16 **Recommended**

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

Hortonworks supported distributions

- HDP 3.1 **Recommended**
- HDP 3.0

NOTE: HDP 3.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.

- HDP 2.6

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

EMR supported distributions

See *Configure for EMR* in the Configuration Guide.

HDInsight supported distributions

See *Configure for HDInsight* 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 1.8 (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 2.2	Spark 2.3	Spark 2.4
Java 1.8	Required.	Required.	Required.

- 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 *Desktop Requirements*.

I/O Requirements

See *Supported File Formats* in the User Guide.

Sizing Guidelines

Contents:

- *Requirements for the Trifacta node*
 - *Enterprise 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*.

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

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 <i>Configure for HDInsight</i> in the Configuration Guide.
Azure Databricks	Please review the Enterprise Hadoop guidelines with <i>Trifacta Customer Success Services</i> . For more information on this integration, see <i>Configure for Azure Databricks</i> in the Configuration Guide.

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
Time-based trigger Service	43033
Scheduling Service	43143

Database Ports

Component	Port
Postgres (default)	5432 <div style="border: 1px solid #ccc; padding: 10px; margin-top: 10px;"> <p>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>.</p> </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 Wrangler Enterprise 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.

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 6*
 - *CentOS/Redhat 7*
 - *Ubuntu 14.04*
 - *Ubuntu 16.04*
- *Direct and Indirect Dependencies*
 - *CentOS/Redhat 6*
 - *CentOS/Redhat 7*
 - *Ubuntu 14.04 / Ubuntu 16.04*

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 6

python-supervisor = 3.0

nodejs = 2:10.13.0

nginx = 1.12.2-1

gzip = 1.3.12-22

bzip2 = 1.0.5-7

openldap-clients

CentOS/Redhat 7

supervisor = 3.1.3

nodejs = 2:10.13.0

nginx = 1:1.12.2-1

gzip = 1.5-8

bzip2 = 1.0.6-13

openldap-clients

Ubuntu 14.04

python-supervisor = 3.0

nodejs = 10.13.0-1nodesource1

nginx = 1.12.2-1~trusty

rlwrap = 0.37-5

gzip = 1.4-1ubuntu2

bzip2 = 1.0.6-1

ldap-utils

Ubuntu 16.04

supervisor = 3.2.0-2

nodejs = 10.13.0-1nodesource1

nginx = 1.12.2-1~xenial

rlwrap = 0.37-5

gzip = 1.4-1ubuntu2

bzip2 = 1.0.6-1

ldap-utils

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 6

alsa-lib-1.1.0-4.el6.x86_64.rpm
fontconfig-2.8.0-5.el6.x86_64.rpm
freetype-2.3.11-17.el6.x86_64.rpm
giflib-4.1.6-3.1.el6.x86_64.rpm
java-1.8.0-openjdk-1.8.0.121-0.b13.el6_8.x86_64.rpm
java-1.8.0-openjdk-devel-1.8.0.121-0.b13.el6_8.x86_64.rpm
java-1.8.0-openjdk-headless-1.8.0.121-0.b13.el6_8.x86_64.rpm
jpackage-utils-1.7.5-3.16.el6.noarch.rpm
libICE-1.0.6-1.el6.x86_64.rpm
libSM-1.2.1-2.el6.x86_64.rpm
libX11-1.6.3-2.el6.x86_64.rpm
libX11-common-1.6.3-2.el6.noarch.rpm
libXau-1.0.6-4.el6.x86_64.rpm
libXext-1.3.3-1.el6.x86_64.rpm
libXfont-1.5.1-2.el6.x86_64.rpm
libXi-1.7.4-1.el6.x86_64.rpm
libXrender-0.9.8-2.1.el6_8.1.x86_64.rpm
libXtst-1.2.2-2.1.el6.x86_64.rpm
libfontenc-1.1.2-3.el6.x86_64.rpm
libjpeg-turbo-1.2.1-3.el6_5.x86_64.rpm
libpng-1.2.49-2.el6_7.x86_64.rpm
libxcb-1.11-2.el6.x86_64.rpm
make-3.81-23.el6.x86_64.rpm
nginx-1.12.2-1.el6_4.ngx.x86_64.rpm
nodejs-10.13.0-1.nodesource.el6.x86_64.rpm
nspr-4.11.0-1.el6.x86_64.rpm
nss-3.21.3-2.el6_8.x86_64.rpm
nss-softokn-3.14.3-23.3.el6_8.x86_64.rpm
nss-softokn-freebl-3.14.3-23.3.el6_8.x86_64.rpm
nss-sysinit-3.21.3-2.el6_8.x86_64.rpm
nss-tools-3.21.3-2.el6_8.x86_64.rpm
nss-util-3.21.3-1.el6_8.x86_64.rpm
openldap-clients-2.4.40-16.el6.x86_64
openssl-1.0.1e-48.el6_8.4.x86_64.rpm
pkgconfig-0.23-9.1.el6.x86_64.rpm
python-meld3-0.6.10-1.noarch.rpm
python-setuptools-0.6.10-3.el6.noarch.rpm
python-supervisor-3.0-1.noarch.rpm
ttmkfdir-3.0.9-32.1.el6.x86_64.rpm
tzdata-java-2017a-1.el6.noarch.rpm
xorg-x11-font-utils-7.2-11.el6.x86_64.rpm
xorg-x11-fonts-Type1-7.2-11.el6.noarch.rpm

CentOS/Redhat 7

alsa-lib-1.1.1-1.el7.x86_64.rpm
bzip2-1.0.6-13.el7.x86_64.rpm
chkconfig-1.7.2-1.el7.x86_64.rpm
copy-jdk-configs-1.2-1.el7.noarch.rpm
fontconfig-2.10.95-10.el7.x86_64.rpm
fontpackages-filesystem-1.44-8.el7.noarch.rpm
giflib-4.1.6-9.el7.x86_64.rpm
java-1.8.0-openjdk-1.8.0.121-0.b13.el7_3.x86_64.rpm
java-1.8.0-openjdk-devel-1.8.0.121-0.b13.el7_3.x86_64.rpm
java-1.8.0-openjdk-headless-1.8.0.121-0.b13.el7_3.x86_64.rpm
javapackages-tools-3.4.1-11.el7.noarch.rpm
libICE-1.0.9-2.el7.x86_64.rpm
libSM-1.2.2-2.el7.x86_64.rpm
libX11-1.6.3-3.el7.x86_64.rpm
libX11-common-1.6.3-3.el7.noarch.rpm
libXau-1.0.8-2.1.el7.x86_64.rpm
libXcomposite-0.4.4-4.1.el7.x86_64.rpm
libXext-1.3.3-3.el7.x86_64.rpm
libXfont-1.5.1-2.el7.x86_64.rpm
libXi-1.7.4-2.el7.x86_64.rpm
libXrender-0.9.8-2.1.el7.x86_64.rpm
libXtst-1.2.2-2.1.el7.x86_64.rpm
libfontenc-1.1.2-3.el7.x86_64.rpm
libjpeg-turbo-1.2.90-5.el7.x86_64.rpm
libpng-1.5.13-7.el7_2.x86_64.rpm
libxcb-1.11-4.el7.x86_64.rpm
lksctp-tools-1.0.17-2.el7.x86_64.rpm
nginx-1.12.2-1.el7_4ngx.x86_64.rpm
nodejs-10.13.0-1nodesource.el7.centos.x86_64.rpm
nspr-4.11.0-1.el7_2.x86_64.rpm
nss-3.21.3-2.el7_3.x86_64.rpm
nss-softokn-3.16.2.3-14.4.el7.x86_64.rpm
nss-softokn-freebl-3.16.2.3-14.4.el7.x86_64.rpm
nss-sysinit-3.21.3-2.el7_3.x86_64.rpm
nss-tools-3.21.3-2.el7_3.x86_64.rpm
nss-util-3.21.3-1.1.el7_3.x86_64.rpm
openldap-clients-2.4.40-13.el7.x86_64
openssl x86_64 1:1.0.2k-12.el7
python-javapackages-3.4.1-11.el7.noarch.rpm
python-meld3-0.6.10-1.el7.x86_64.rpm
python-setuptools-0.9.8-4.el7.noarch.rpm
supervisor-3.1.3-3.el7.noarch.rpm
ttmkfdir-3.0.9-42.el7.x86_64.rpm
tzdata-java-2016j-1.el7.noarch.rpm
xorg-x11-font-utils-7.5-20.el7.x86_64.rpm
xorg-x11-fonts-Type1-7.5-9.el7.noarch.rpm

Ubuntu 14.04 / Ubuntu 16.04

Please contact *Trifacta Support*.

Desktop Requirements

Contents:

- *Browser Requirements*
 - *Google Chrome Requirements*
 - *Mozilla Firefox Requirements*
 - *Ports*
 - *Desktop Requirements*
 - *General Requirements*
 - *Wrangler Enterprise desktop application Requirements*
-

Browser Requirements

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

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 loading Trifacta datasets. You may need to disable the extension, particularly if you are loading marketing data.

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

NOTE: If you are using the Wrangler Enterprise desktop application, an installed instance of one of the supported browsers is not required. Additional requirements are listed below.

Google Chrome Requirements

Trifacta Wrangler Enterprise requires the use of Google-supported client extensions. Supported desktop Google Chrome clients:

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.
 - Google has not yet implemented multi-threaded WebAssembly.
 - 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.

Browser versions

Version: Google Chrome v.76-v.78, and any stable version that is released prior to the next release of Trifacta Wrangler Enterprise.

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

NOTE: To access Trifacta Wrangler Enterprise you must use a supported desktop version of the Google Chrome browser, unless you are using Wrangler Enterprise desktop application.

NOTE: Mobile browsers and Google Chromebook are not supported.

Mozilla Firefox Requirements

The following versions of the Mozilla Firefox browser are supported:

NOTE: This feature is in Beta release.

Version: Mozilla Firefox v.68-v.70, and any stable version that is released prior to the next release of Trifacta Wrangler Enterprise.

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

Ports

By default, the web client uses port 3005.

For more information on required client ports, see *System Ports*.

Desktop Requirements

The following requirements apply to your local system.

General Requirements

- Intel Pentium 4 or AMD Opteron processor or newer (SSE2 or newer is required)
- 4 GB RAM

Tip: At least 8 GB RAM is recommended.

- 2 GB hard disk space
- 1280 x 720 screen resolution and above
- Internet connection (DSL or better)

NOTE: Trifacta® Wrangler Enterprise does not support connection through non-transparent or unauthenticated proxies. If you receive "Remote server timed out" error messages, you may need to re-connect through a network without a proxy.

Wrangler Enterprise desktop application Requirements

NOTE: The Wrangler Enterprise desktop application may be deprecated in a future release, when additional browsers are supported.

The Wrangler Enterprise desktop application can be used locally to interact with Trifacta Wrangler Enterprise and can be installed in your local environment. Please verify that the following requirements are met.

System requirements

In addition to the general desktop requirements, the following requirements apply to the Wrangler Enterprise desktop application:

- 8 GB RAM minimum

Windows requirements

- Windows 7 (Service Pack 1), 8, or 10. 64-bit versions.

NOTE: The Wrangler Enterprise desktop application requires a 64-bit version of Microsoft Windows.

- Visual C++ for Visual Studio 2015 (`vc_redist.x86.exe`). To download: <https://www.microsoft.com/en-us/download/details.aspx?id=48145>.

Windows execution

When the installation is completed, most users should be able to launch the local application without issues. If you are experiencing issues launching the application, please verify that all DLL and EXE files in the following directory are executable:

Tip: This requirement typically applies to more secure environments. Most desktop users should not have to modify permissions.

```
C:\Users\\AppData\Local\Trifacta\Trifacta Wrangler Enterprise\
```

Supported File Formats

Contents:

- *Native Input File Formats*
 - *Native Output File Formats*
 - *Compression Algorithms*
 - *Read Native File Formats*
 - *Write Native File Formats*
 - *Additional Configuration for File Format Support*
 - *Publication of some formats requires execute permissions*
-

This section contains information on the file formats and compression schemes that are supported for input to and output of Trifacta® Wrangler Enterprise.

NOTE: To work with formats that are proprietary to a desktop application, such as Microsoft Excel, you do not need the supporting application installed on your desktop.

Native Input File Formats

Trifacta® Wrangler Enterprise can read and import directly these file formats:

- Excel (XLS/XLSX)

Tip: You may import multiple worksheets from a single workbook at one time. See *Import Excel Data* in the User Guide.

- PDF

NOTE: PDF support may need to be enabled in your environment. See *Import PDF Data*.

- PDF is supported for import only.
- See *Import PDF Data* in the User Guide.

- CSV
- JSON, including nested

NOTE: Trifacta Wrangler Enterprise requires that JSON files be submitted with one valid JSON object per line. Consistently malformed JSON objects or objects that overlap linebreaks might cause import to fail. See *Initial Parsing Steps* in the User Guide

- Plain Text
- LOG
- TSV
- Parquet

NOTE: When working with datasets sourced from Parquet files, lineage information and the `$sourcerownumber` reference are not supported.

- XML

NOTE: XML files can be ingested as unstructured text. XML support is not enabled by default. For more information, please contact *Trifacta Customer Success Services*.

- Avro

For more information on data is handled initially, see *Initial Parsing Steps* in the User Guide.

Native Output File Formats

Trifacta Wrangler Enterprise can write to these file formats:

- CSV
- JSON

- Tableau (TDE)

NOTE: Publication of results in TDE format may require additional configuration. See below.

- Avro

NOTE: The Trifacta Photon and Spark running environments apply Snappy compression to this format.

- Parquet

NOTE: The Trifacta Photon and Spark running environments apply Snappy compression to this format.

Compression Algorithms

NOTE: Importing a compressed file with a high compression ratio can overload the available memory for the application. In such cases, you can decompress the file before uploading. If decompression fails, you should contact your administrator about increasing the Java Heap Size memory.

NOTE: Publication of results in Snappy format may require additional configuration. See below.

NOTE: GZIP files on Hadoop are not split across multiple nodes. As a result, jobs can crash when processing it through a single Hadoop task. This is a known issue with GZIP on Hadoop.

Where possible, limit the size of your GZIP files to 100 MB or less, or use BZIP2 as an alternative compression method. As a workaround, you can try to run the job on the unzipped file. You may also disable profiling for the job. See *Run Job Page* in the User Guide.

Read Native File Formats

	GZIP	BZIP	Snappy
CSV	Supported	Supported	Supported
JSON	Supported	Supported	Supported
Avro			Supported
Hive			Supported

Write Native File Formats

	GZIP	BZIP	Snappy
CSV	Supported	Supported	Supported
JSON	Supported	Supported	Supported
Avro			Supported; always on
Hive			Supported; always on

Additional Configuration for File Format Support

Publication of some formats requires execute permissions

When job results are generated and published in the following formats, the Trifacta platform includes a JAR, from which is extracted a binary executable into a temporary directory. From this directory, the binary is then executed to generate the results in the proper format. By default, this directory is set to `/tmp` on the Trifacta node.

In many environments, execute permissions are disabled on `/tmp` for security reasons. Use the steps below to specify the temporary directory where this binary can be moved and executed.

Steps:

1. Login to the application as an administrator.
2. From the menu, select **Settings menu > Settings > Admin Settings**.
3. For each of the following file formats, locate the listed parameter, where the related binary code can be executed:

File Format	Parameter	Setting to Add
Snappy	"data-service.jvmOptions"	-Dorg.xerial.snappy.tmpdir=<some executable directory>
TDE	"batch-job-runner.jvmOptions"	-Djna.tmpdir=<some executable directory>

4. Save your changes and restart the platform.
5. Run a job configured for direct publication of the modified file format.

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*
 - *HDInsight*
 - *Cluster*
-

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:** `trifactaactiviti`
- **Documentation reference:** `[db.jobs.user]`

Scheduling database

Storage of schedules, including datasets to execute.

- **Required configuration:** Yes
- **Default user:** `trifactaschedulingservice`
- **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:** `trifactaconfigurationservice`
- **Documentation reference:** `[db.configuration.user]`

Artifact Storage Service database

Storage for feature-specific usage data such as 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]`

HDInsight

Cluster

When integrating with HDInsight, you must specify a user for the platform to use with its connections to HDI.

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

NOTE: The Trifacta platform requires access to a set of Hadoop components. See *System Requirements*.

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*

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. Trifacta 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-allocation-vcores</code>	Per Container	Minimum allocation virtual CPU cores per container; requests lower than will be increased to this value.
<code>yarn.scheduler.maximum-allocation-vcores</code>	Per Container	Maximum allocation virtual CPU cores per container; requests higher than this will be capped to this value
<code>yarn.scheduler.increment-allocation-vcores</code>	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
<code>yarn.scheduler.maximum-allocation-vcores</code>	3	6	13	32	52
<code>yarn.scheduler.increment-allocation-vcores</code>	1	1	1	1	1
<code>spark.executor.memory</code>	6GB	6GB	16GB	20GB	20GB
<code>spark.executor.cores</code>	2	2	4	5	5
<code>spark.driver.memory</code>	4GB	4GB	4GB	4GB	4GB
<code>spark.driver.cores</code>	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



Copyright © 2020 - Trifacta, Inc.
All rights reserved.