[Architecture - Talend Data Management 5.6](https://help.talend.com/display/KB/Architecture+-+Talend+Data+Management+5.6)

[Skip to end of metadata](https://help.talend.com/display/KB/Architecture+-+Talend+Data+Management+5.6" \l "page-metadata-end)

[Go to start of metadata](https://help.talend.com/display/KB/Architecture+-+Talend+Data+Management+5.6#page-metadata-start)

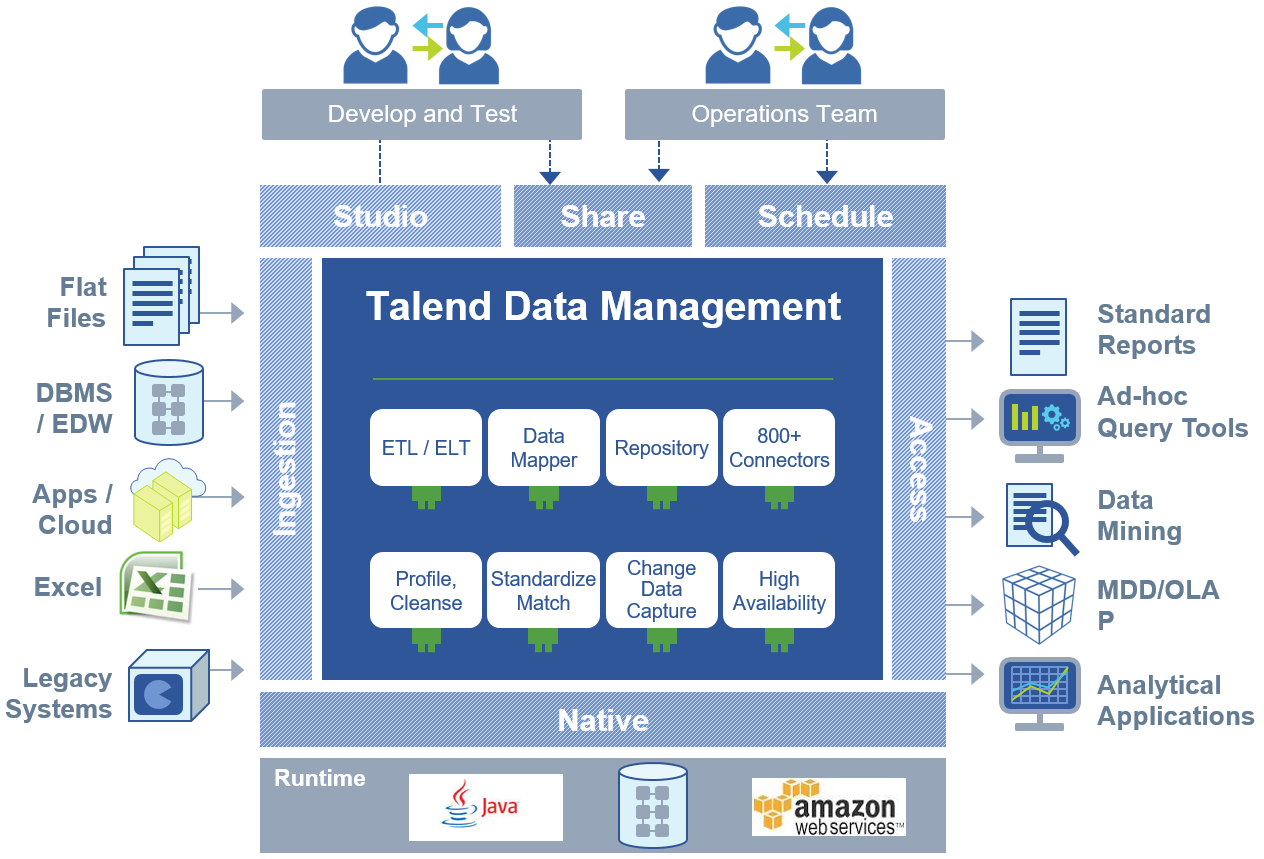
This article describes the logical and physical architecture of Talend Data Management product.

The information below also applies to all Talend Platform products that include Data Integration (DI) and Data Quality (DQ), namely:

* Talend Platform for Big Data
* Talend Platform for Data Services
* Talend Platform for Data Services with Big Data
* Talend Platform for MDM
* Talend Platform for MDM with Big Data
* Talend Platform for Enterprise Integration
* Talend Platform for Enterprise Integration with Big Data
* Talend Platform - Universal

**High Level Architecture**

The figure below shows the logical architecture for the Talend Data Management product.  It provides a rich set of data integration and data quality features.



The key capabilities are:

* Connects to more than 800 data sources and applications, thereby minimizing coding
* Shared repository facilitates team collaboration
* Advanced management and monitoring tools simplifies deployment and tuning
* Easily deploy batch jobs to Amazon EC2
* Runs natively on Java and SQL (for ELT jobs)

**Talend Software Components**

Talend Data Management is bundled with the following software components:

* Talend Studio
* Talend Administration Center
* Subversion
* CommandLine
* JobServer or Talend Runtime (Runtime contains a JobServer)
* Activity Monitoring Console
* Log Server
* Artifact Repository (Archiva for software update and Nexus/Archiva for publishing compiled artifacts)
* Data Quality Portal
* Data Stewardship Console (DSC)

**Talend Studio**

The Talend Studio is used for developing and building data integration jobs.  The key features are:

* Business oriented process modelling
* Graphical Development
* Broadest Connectivity (800+ components)
* Real-time debugging
* Auto Documentation
* Shared Artifact Repository
* Monitoring of processes
* Data profiling and matching

The Talend Studio uses perspectives to focus developers on the various tasks.  The perspectives regroup a set of features each which are activated by the proper license.

The perspectives are:

* Integration
* Mediation (Routing)
* MDM
* BPM
* Profiling
* Talend Data Mapper
* Component Designer

The studio is based on eclipse 3.6 RCP (Rich Client Platform).  Only eclipse plugins allowed by the Talend license is allowed within the Talend Studio. All features are license activated.

**Talend Administration Center**

The Talend Administration Center is a server component that enables the following:

* Environment Configuration
* User/Role Administration
* Project Administration
* Authorization
* Tasks Scheduling and Execution (Job Conductor)
* Monitoring
* Recovery & Restart of Tasks

The Talend Administration Center (also commonly referred to as TAC) is a web application that can be hosted on Tomcat, JBoss and/or Weblogic.  It is a fully compliant web application and comes packaged as a war file.  It also controls the access to other Talend applications like the Activity Monitoring Console, Drools Guvnor, Kibana, Artifact Repository and others.

In general, only one Talend Administration Center is needed per Talend environment.

The Talend Administration Center will maintain Admin Metadata (users, project definition, authorization, scheduler tasks, configuration, etc.) within a database.  This database is solely accessed by the TAC. The database is generally small in size (less than 1 GB) even with thousands of tasks running on the Job Conductor.

The supported databases for the Talend Administration Center are listed in section [1.8 Compatible Databases](https://help.talend.com/display/TalendPlatformUniversalInstallationandUpgradeGuide56EN/1.8+Compatible+Databases) of the installation and upgrade guide. For Talend 5.6, the databases supported are H2, MySQL, Oracle, SQL Server and PostgreSQL.  Please refer to the installation guide for the supported and recommended versions of these databases.

**Subversion**

Talend uses Subversion as a repository for Talend projects.  Subversion is a software versioning and revision control system from Apache.  It is distributed as free software under the Apache license.  Some common distributions are:

* Visual SVN Server (<http://www.visualsvn.com/server/>)
* CollabNet Subversion Edge (<http://www.collab.net/downloads/subversion>)

Talend stores jobs, connections, schema definitions, custom jars, third party libraries, and properties files in Subversion.  It also provides a versioning system for Talend projects and artifacts within the projects.  Developers can transparently use common functionalities Get, Checkout, Commit without even knowing it.

Generally, Subversion is only needed in the development environment.  The only exception is when using MDM where a subversion instance is needed in test and production environments to properly deploy MDM artifacts.

Note that Talend does not support the SVN Merge functionality.  Hence, changes are overwritten once the user decide which version to keep.

**CommandLine**

The CommandLine is a server component that is an exact copy of the Talend Studio running in a headless non-GUI mode.  It is a key component to perform continuous integration with Talend.  The CommandLine supports several modes: Server, Interactive Shell and Scripting.

The CommandLine primary purpose is to generate Java code, compile and package job binaries for deployment onto the Job Conductor within the Talend Administration Center.  It is always invoked through commands which can be sent to it.  The same commands can be used as part of a shell scripting approach.

The CommandLine generally runs as a service on the same server as TAC and/or the CI Environment (for example Jenkins, Bamboo, etc.).

**JobServer**

The JobServer is a lightweight agent used for execution and monitoring of Talend tasks deployed through the TAC Job Conductor.  It can also be used by Talend Studio users through the Distant Run functionality.

The JobServer is a server component that runs as a service.  There is no license restrictions on the number of JobServers that a customer can install. The JobServer also monitors the server health (CPU, RAM, Disk Usage).

**Talend Runtime**

The Talend Runtime is an OSGi container based on Apache Karaf project. It allows you to deploy and execute various components and applications.

It can be used to deploy and execute all the services, routes and generic OSGi features created by the Talend Studio.

It provides the following features:

* Embark a JobServer agent for the execution of DI tasks
* Administration and monitoring via jmx
* Control of container via direct shell, ssh or web console

It is recommended to install and configure the Talend Runtime instead of the JobServer agent if there are requirements to build services and routes. However, sometimes it may be preferred to use JobServer agent and Talend Runtime on different execution servers for separation of concerns.

**Activity Monitoring Console**

The Activity Monitoring Console is a set of features that display information about the execution or each tasks.  It is used in conjunction with a database consisting of 3 database tables or 3 files on disk (stats, logs and flow meter).   The schema of each table/file can be extended to add more columns for extra info.  Talend Jobs, if configured, will write to the Activity Monitoring Console tables, and the information can then be access through the Studio or the TAC.

 The volume of data stored in the database tables or files are directly related to the number of tasks and their frequency of executions.  Developers must design additional jobs to manage the size and perform archiving for the data within these 3 tables.  Additional indexes can be added to the 3 tables for enhancing the performance.

 The compatible databases are listed in the installation guide in section [1.8 Compatible Databases](https://help.talend.com/display/TalendPlatformUniversalInstallationandUpgradeGuide56EN/1.8+Compatible+Databases)

**LogServer**

The LogServer is based on ElasticSearch and LogStash (<http://logstash.net/>).  It is used to streamline the capture and storage of logs from TAC, MDM Server, ESB Server and Tasks running through the Job Conductor.

The LogServer runs as a service, generally on the same server as the TAC. The Kibana UI in the TAC connects to the ElasticSearch and enables the administrator/user to query and search the logs.

**Archiva Artifact Repository**

The primary purpose of the Archiva Artifact Repository (based on Apache Archiva) is to receive and store binary distributions for software updates from the Talend website.  It can generally be configured to work through a proxy and can even work without access to the Talend Website.

The Archiva Artifact Repository is generally installed on the same server as the TAC.  It is a web application that is provided as a package with an embedded jetty database.  It is always configured as a service.

Note that the use of Apache Archiva is deprecated in Talend 5.6.2.

**Nexus Artifact Repository**

In Talend 5.6.2, the Nexus artifact repository can be used for storing published artifacts (Jobs, Services, Routes) from the studio.  It cannot be used for software update though.

The Nexus artifact repository is generally configured on a separate server that is accessible from all environments.  The Nexus artifact repository can also be configured to proxy some public repositories which contains third party libraries.

**Talend Data Quality Portal**

The Talend Data Quality Portal allows business users to view Data Quality reports and dashboards via a Web interface.

Technical Details:

* Server component
* Web application – It is accessed via a browser
* There can be many instances of the Talend Data Quality Portal depending on business requirements
* Hosted on Tomcat only (can be same Tomcat as TAC)
* Talend Installer can install the Talend Data Quality Portal and Tomcat (v6) together
* As well as the relational DB for the report data, a HSQL database is also used for environment / user management
* The web application server is typically run as a Service / Daemon

**Talend DQ Data Mart**

The Talend DQ Data Mart is a database that holds the results of the execution of data quality reports.  A data quality report can be executed directly from the Talend Studio Profiling perspective, or within a data integration job that executes a specified report.

Technical Details:

* Server component
* Only MySQL and Oracle databases are supported
* Small – Medium size database: does not hold any actual source data:
* Evolutionary Reports: all results from all report runs
* Basic reports: last run of report

**Talend Data Stewardship Console (DSC)**

The Talend Data Stewardship Console (DSC) provides a Web UI for Data Stewards to resolve issues with records and possible matches when doing data matching.  The DSC can be installed standalone or as part of a Talend MDM installation.  It is used by MDM for integrated and complex matching.

Technical Details:

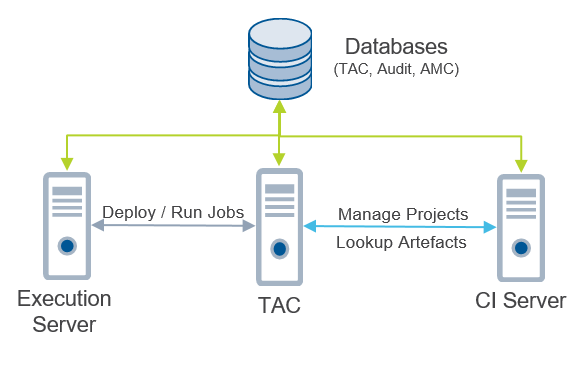
* Server component
* Web application – accessed via browser
* Hosted on Tomcat (if standalone) or JBoss (if MDM)
* If installed with MDM, it uses the MDM authentication system and is a child application of MDM Web UI
* If installed standalone, it provides basic user management via a file
* Talend Installer can install DSC with TAC or MDM
* The web application server is typically run as a Service / Daemon

**Physical Architecture For Data Integration**

Talend recommends that customers plan at least 3 environments i.e. a Development, Test and Production environments. The physical architecture for a typical setup for each of these environments are described below. The architecture team must perform a sizing exercise based on the functional and non-functional requirements of the project(s) and design the correct architecture for development, test and production environments that matches the needs of the business.

**Typical Development Environment Architecture**

A typical development environment architecture is shown in the diagram below.  This architecture is recommended for a small team of 5-10 developers.



Refer to the Talend software [Installation Requirements](https://help.talend.com/display/TalendPlatformUniversalInstallationandUpgradeGuide56EN/1.2+Installation+requirements) for the memory and disk required for installation.  In the above architecture, additional Execution Server(s) may be needed if there are many tasks (> 20) being scheduled at the same time and requiring a significant amount of CPU and memory resources.  Further detailed information about OS, Java, ports used, database engine supported, etc. are provided in the production [installation documentation](https://help.talend.com/display/TalendPlatformUniversalInstallationandUpgradeGuide56EN/Home).

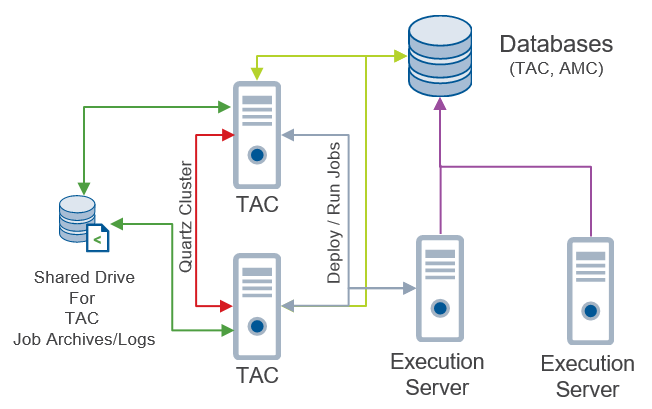
Note: The Developer Workstation is not shown in the above diagram for simplicity.  The typical sizing in the table below describes the typical server and workstation configuration recommended based on the architecture above.  A server may be hosting several components that make up the platform.

| **Workstation/Server Role** | **Description** | **Typical Sizing** |
| --- | --- | --- |
| Execution Server | This is where all Talend jobs deployed as tasks on the Job Conductor in TAC will be executed for testing. | OS: Windows/Linux (See Installation Guide)  CPU: 8 Cores Minimum  RAM: 32+ GB RAM (16 GB Minimum)  Disk Size: 100 GB |
| TAC | The TAC server will host the following:   * Talend Administration Center * AMC * Kibana * Archiva Artifact Repository (Installed locally for software updates. Archiva is deprecated in Talend 5.6.2) | OS: Windows/Linux (See Installation Guide)  CPU: 4 Cores Minimum  RAM: 16 GB RAM (8 GB Minimum)  Disk Size: 100 GB |
| CI Server | The CI Server or sometimes also referred to as repository server will host the following:   * Subversion * Nexus * CommandLine * CI Tools like Jenkins and Maven | OS: Windows/Linux (See Installation Guide)  CPU: 4 Cores Minimum  RAM: 16 GB RAM  Disk Size: 300 GB |
| Developer Workstation | The developer workstation is not shown in the above diagram to keep the diagram simple.  The following are generally installed on the developer workstation:   * Java JDK * Talend Studio * Web Browser * Other tools like text editors and SoapUI | OS: Windows/Linux/Mac OS (See Installation Guide)  CPU: 4 Cores Minimum  RAM: 16 GB RAM  Disk Size: 512 GB Recommended |

**Typical Test Environment Architecture**

It is recommended to plan one or more test environments based on the needs of the business.  There can be multiple test environments for various business requirements like System Integration Test, User Acceptance Test, Performance Test, etc.

The diagram below shows the architecture for a typical test environment.



The main differences here from the development environment are:

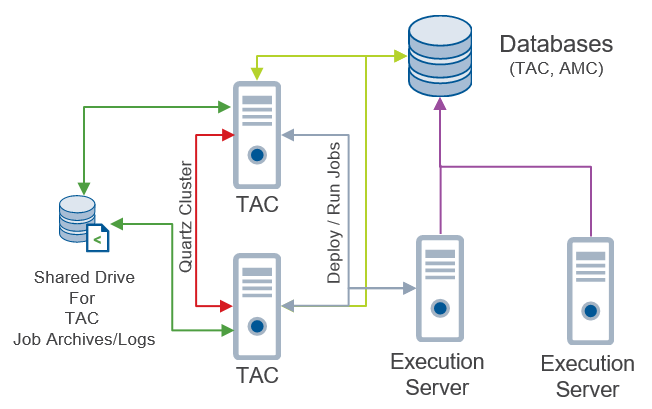
* 2 TACs configured with Quartz Scheduler clustering. Note that this is only available in Platform products. A shared drive is used between the 2 TACs for storing the job archives and log generated by each task run.  This way both TAC see exactly the same configuration.
* Minimum of 2 execution servers to mimic production and enables testing on a production like configuration

The TAC in the test environments will need access to the nexus snapshots and releases repositories configured in the development environment.  Tasks on the job conductor will be created using the nexus deployment functionality.  The job binaries will be download from either the snapshots or releases repository.

| **Workstation/Server Role** | **Description** | **Typical Sizing** |
| --- | --- | --- |
| Execution Server | This is where all Talend jobs deployed as tasks on the Job Conductor in TAC will be executed for testing. | OS: Windows/Linux (See Installation Guide)  CPU: 8 Cores Minimum  RAM: 32+ GB RAM (16 GB Minimum)  Disk Size: 100 GB |
| TAC | The TAC server will host the following:   * Talend Administration Center * AMC * Kibana * Archiva Artifact Repository (Installed locally for software updates. Archiva is deprecated in Talend 5.6.2) | OS: Windows/Linux (See Installation Guide)  CPU: 4 Cores Minimum  RAM: 16 GB RAM (8 GB Minimum)  Disk Size: 100 GB |
| Tester Workstation | The tester workstation is not shown in the above diagram to keep the diagram simple.  The following are generally installed on the developer workstation:   * Java JDK * Web Browser * Other tools like text editors and SoapUI | OS: Windows/Linux/Mac OS (See Installation Guide)  CPU: 4 Cores Minimum  RAM: 16 GB RAM  Disk Size: 512 GB Recommended |

**Typical Production Architecture**

The diagram below shows a typical production environment.  It is very similar to the test environment described above because it is usually recommended that a test environment is setup similar to the production one for User Acceptance and Performance testing.



The 2 TACs are configured such that the quartz scheduler is clustered for high availability.  It is possible to have more than 2 TAC and more than 2 execution servers in production. These needs will be driven by the functional and non-functional requirements.

The TAC in the production environment will need access to the nexus releases repository configured in the development environment.  Tasks on the job conductor will be created using the nexus deployment functionality.  The job binaries will be download from the releases repository.

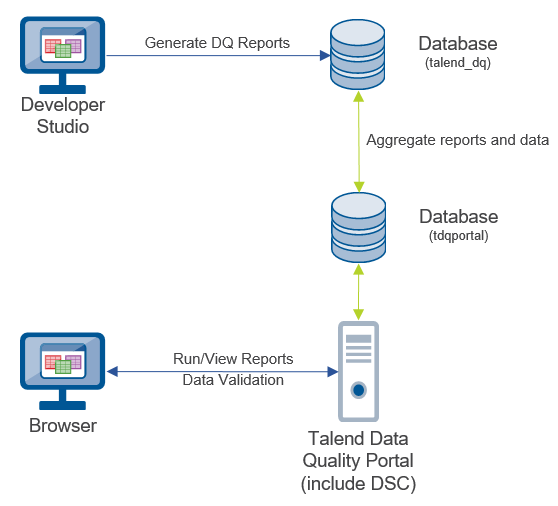
| **Workstation/Server Role** | **Description** | **Typical Sizing** |
| --- | --- | --- |
| Execution Server | This is where all Talend jobs deployed as tasks on the Job Conductor in TAC will be executed in development environment for testing. | OS: Windows/Linux (See Installation Guide)  CPU: 8 Cores Minimum  RAM: 32+ GB RAM (16 GB Minimum)  Disk Size: 100 GB |
| TAC | The TAC server will host the following:   * Talend Administration Center * AMC * Kibana * Archiva Artifact Repository (Installed locally for software updates. Archiva is deprecated in Talend 5.6.2) | OS: Windows/Linux (See Installation Guide)  CPU: 4 Cores Minimum  RAM: 16 GB RAM (8 GB Minimum)  Disk Size: 100 GB |
| Support/Admin Person Workstation | The support/admin person workstation is not shown in the above diagram to keep the diagram simple.  The following are generally installed on the administrator workstation:   * Java JDK * Web Browser * Other tools like text editors and SoapUI | OS: Windows/Linux/Mac OS (See Installation Guide)  CPU: 4 Cores Minimum  RAM: 16 GB RAM  Disk Size: 512 GB Recommended |

**Physical Architecture For Data Quality**

The diagram below shows the main components needed for data quality features of the platform.  This architecture may be replicated as is in the development, test and production environments.  Each environments will follow the same architecture design.

The following are needed:

* 2 Databases/Schemas
* 1 Server that will host the Data Quality Portal and the Data Stewardship Console web applications. The 2 web applications will need to be accessible to business users and data stewards.  Hence, the security requirements on this server may be different due to a need for access to other users than developers and administrators.



| **Workstation/Server Role** | **Description** | **Typical Sizing** |
| --- | --- | --- |
| Talend Data Quality Portal | This server will host the Talend Data Quality Portal and the Data Stewardship console web applications. | OS: Windows/Linux (See Installation Guide)  CPU: 8 Cores Minimum  RAM: 16 GB RAM  Disk Size: 100 GB |