TALEND CLOUD

Security Architecture Overview

Talend Spring ‘19
AWS V1

talend

SUMMARY

Talend Cloud (https://cloud.talend.com) is a secure and managed cloud integration platform that makes it easy for developers and non-developers to connect, transform, cleanse, and share cloud and on-premises data. Security is our top priority at Talend, whether it is employee, physical, network, infrastructure, platform, or data security. Talend implements a combination of policies, procedures, and technologies to ensure your data is protected and secured. This document provides an overview of the Talend Cloud internal architecture and these policies and procedures for Talend Spring '19.

Talend has a team of dedicated security specialists that follow industry best practices for security and run regular internal security audits. The team also maintains policies that span operations, data security, passwords and credentials, facilities and networks, and connectivity. It should be noted that Talend does not observe, inspect, store, or otherwise interact directly with customer data. These security policies are regularly updated according to evolving industry standards and we reassess compliance through frequent audits. The Talend Cloud platform is hosted on Amazon Web Services, a SSAE 16 certified data center.

TALEND CLOUD OVERVIEW

Talend Cloud is a multi-tenant integration environment allowing you to design, manage, and monitor integration pipelines. Developers use Talend Studio (locally) to design data integration flows (or jobs) or Talend Pipeline Designer (a Talend Cloud app) to design data pipelines. Jobs and pipelines are then run in the Cloud or on-premises. The Talend Cloud infrastructure consists of the following applications as shown in the diagram on the following page: Talend Cloud, Talend Cloud Management Console, Talend Pipeline Designer, Talend Cloud Data Preparation, Talend Cloud Data Stewardship, Talend Cloud API Designer, Talend Cloud API Tester, and Talend Dictionary Service.

Talend Cloud

Talend Cloud is the web application that you access via your web browser on your desktop or mobile device, and it serves as the execution and communications platform to run Talend jobs. Talend Cloud runs as load-balanced application server instances on Amazon Web Services (AWS). After building a job it is executed on one or more "Cloud Engines" or "Remote Engines".

- Cloud Engines are Java-based runtimes deployed via an Amazon Machine Image (AMI) based on CentOS.
- Remote Engines are optional Java-based runtimes deployed by the customer to process data behind the firewall or on a Virtual Private Cloud (VPC), e.g. this can be on-premises or in third party clouds like Google, Azure or AWS.

Talend Cloud Management Console

Talend Cloud Management Console (TCMC) allows you to manage users, roles, groups, and projects for your Cloud applications, e.g. data integration, Data Preparation, Data Stewardship, Pipeline Designer.

Talend Pipeline Designer

Talend Pipeline Designer (TPD) is a Cloud app to design data integration pipelines in your browser and run in the Cloud or on-premises.

- A pipeline is a data integration process (similar to a Talend Job) that extracts data from a dataset (source), transforms data using processors and loads data into one or several datasets (destinations).
- TPD stores input / output connectivity configuration into an inventory of connections and datasets that can be referenced in pipelines.
- Pipelines can be executed interactively in the application or scheduled in Talend Cloud Management Console.
- Pipeline executions are performed by a Remote Engine for Pipelines (hosted on-premises or a 3rd-party Cloud provider).

Talend Cloud Data Preparation

Talend Cloud Data Preparation (TCDP) is a self-service application that enables information workers to simplify and expedite the time-consuming process of preparing data for analysis or other data-driven tasks. In terms of naming, users create, update, delete and share datasets, and will create preparations on top of these datasets. Preparation executions can then be operationalized using Talend Studio.

Talend Cloud Data Stewardship

Talend Cloud Data Stewardship (TCDS) is a team-based, self-service data curation, arbitration and validation app where you quickly identify, manage, and resolve any data integrity issue.

Talend Cloud API Designer

Talend Cloud API Designer (TCAD) is an application that provides visual design and team collaboration tools for defining a contract-first API with consumers. "Live preview" makes it easy to simulate an API, which can be deployed to popular API gateways for load balancing and mediation. Documentation is automatically generated facilitating API use by others. It is deployed as a web application.

Talend Cloud API Tester

Talend Cloud API Tester (TCAT) is a visual tool to discover and debug APIs. You can easily create tests and run scenarios composed of many API requests to simulate real-life usage. Unit tests can be integrated into a managed CI/CD process, ensuring consistent quality.

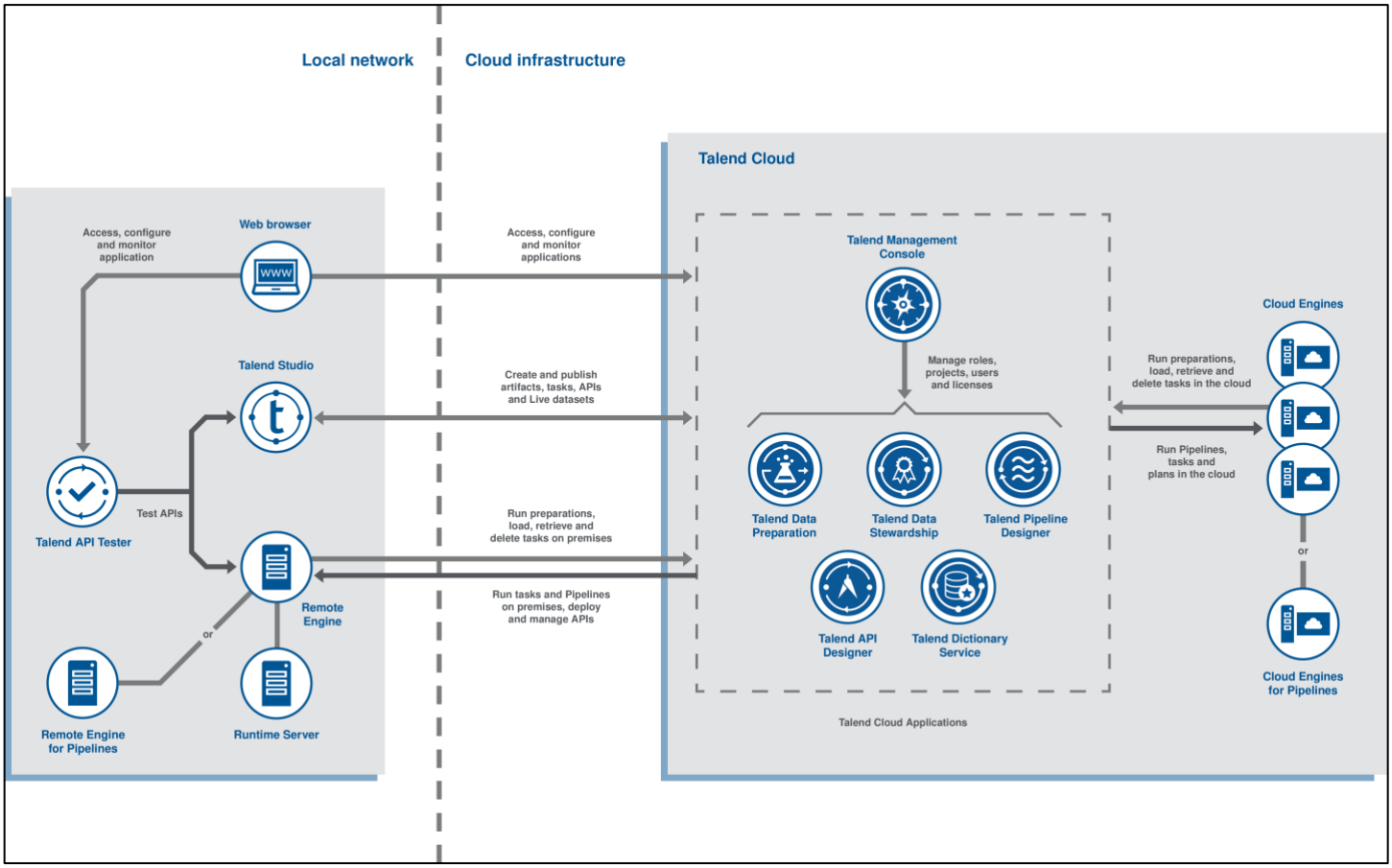


Figure 1 - Talend Cloud functional architecture

Talend Dictionary Service

The Talend Dictionary Service is used to define and manage semantic data types for Talend Cloud Data Preparation and Talend Cloud Data Stewardship. There are two Dictionary Service components: Data Quality (DQ) Dictionary Server and Data Quality (DQ) Dictionary Consumer.

- **DQ Dictionary Server**: Semantic data types can be created, modified, deleted in a dedicated web user interface. Any change is stored in Talend Cloud.

- **DQ Dictionary Consumer:** When modifications are "published", they are propagated from the DQ Dictionary Server to client applications (i.e. TCDP/TCDS) via Apache Kafka events. The client applications rely on a DQ Library that manages these events. The DQ Library creates and maintains a local Apache Lucene copy of the semantic types, so that TCDP/TCDS can perform semantic type discovery and data validation.

Client Deployed Apps

As a Talend Cloud user, you download and install the following applications:

- Talend Studio – the development environment to build integration jobs
- (optional) Remote Engine – a Java-based runtime (similar to a Cloud Engine) to execute Talend jobs, on-premises or another cloud platform that is controlled by the customer
- (optional) Remote Engine for Pipelines – a Java-based runtime to execute Pipeline Designer pipelines, on-premises or another cloud platform that is controlled by the customer
- Web browser – Talend Cloud supports these popular browsers

TALEND CLOUD INFRASTRUCTURE

Each Talend Cloud customer is assigned an account to access Talend Cloud services. This account contains as many users as defined by the license purchased. A tenant is defined as an account, which accesses Talend Cloud services, and has its own configuration. This section covers tenant related computation resources, data storage, networking, disaster recovery, business continuity planning and vulnerability management.

Compute

Talend Cloud Management Console, Talend Cloud Data Preparation, and Talend Pipeline Designer are the only Talend Cloud applications that have tenant-specific computation resources.

Talend Cloud Management Console

Execution of an integration flow or job is done using either a Cloud Engine or a Remote Engine. Remote Engines are under control of the customer and run within their own infrastructure. Cloud Engines are started on a dedicated AWS AMI instance, and are not shared with other tenants.

Talend Pipeline Designer

Pipeline executions are performed on a Remote Engine for Pipelines. Remote Engines for Pipelines are under the customer's control and run on-premises or by the customer in the Cloud. The live preview feature of Pipeline Designer (ability to see the result of a processor while designing a pipeline) is executed by the Remote Engine for Pipelines.

Talend Cloud Data Preparation

When preparing data and then doing a "full run" or "fetch more" data process, computation is performed by the Talend *fullrun service* in a thread temporarily dedicated to the tenant's execution. Depending on the user's choice, the results can be stored in one of two ways:

- In an S3 bucket managed directly by the customer when the user selects 'S3' as output type. The credentials are used once and not stored in Talend Cloud.
- In Talend infrastructure, if the user selects 'CSV', 'XLS' or 'Tableau' as the output type. In this case, the results are stored in the bucket/folder as specified by the *Configuration service*.

Data Storage

Talend Cloud, across its applications in the service, stores necessary customer data using services to ensure security and availability. Each tenant has its own database instance. This database is protected by credentials unique to the tenant. Talend provides encryption for data in transit and data at rest following industry best practices.

It should be noted that Talend does not observe, inspect, store, or otherwise interact directly with customer data.

Network

The network communications for Talend Cloud works as follows. All Cloud Engines are started in a separate AWS Virtual Private Cloud. This VPC segment is isolated from all other segments within Talend Cloud. Only outgoing Secure Socket Shell (SSH) communication from Talend Cloud to the Internet is enabled. No incoming connections are allowed from the Internet to that VPC segment.

Disaster Recovery/Business Continuity Plan

Talend Cloud's DR/BC Plan includes annual reviews and updates, mirrored data centers in the US, mirrored data centers in EMEA, and hourly data synchronization from production to warm backup sites. The DR/BC Plan is fully tested for each major Talend Cloud release. The testing and process is audited annually by 3rd party SOC auditors.

Vulnerability Management

Talend Cloud components and applications are scanned prior to deployment for vulnerabilities in third party code via industry leading scanning tools. All detected vulnerabilities are then addressed according to Talend's internal vulnerability management policy. Talend follows the Security Content Automation Protocol (SCAP) framework and vulnerabilities are rated according to the Common Vulnerability

Scoring System (CVSS) v3.0 equation. Vulnerabilities are resolved depending on their severity rating and their potential impact on the infrastructure.

Access to Talend Cloud is closed off from the outside in; only critical service ports are allowed, and these are secured via industry standard security protocols.

CERTIFICATION AND COMPLIANCE

Data Center and Talend Cloud Platform

Talend Cloud leverages the comprehensive security and certifications provided by the Amazon Web Services (AWS) data center which it is hosted on:

- SSAE 16 certified data center
- Comprehensive and regularly tested Disaster Recovery and Business Continuity plan
- See the complete list of AWS compliance here

All metadata sits in the region in which the customer is hosted. The AWS Data Processing Agreement meets the requirements of the EU Data Protection Directive (http://aws.amazon.com/compliance/eu-data-protection).

Talend Cloud Platform is certified to comply with the following standards and frameworks:

- SSAE 16 SOC 2 Type II certified
- ISAE 3402 certified
- Cloud Security Alliance (Level 1)

Talend Software Development Practices

Talend implements a Top Ten Open Web Application Security Project (OWASP) awareness program during application development, and schedules regular internal and external audits to assess compliance with OWASP best practices.

Talend leverages 3rd party security services to perform external penetration tests, which are scheduled twice a year and prior to a new application deployment inside Talend Cloud. The penetration tests cover a wide range of security aspects of the application and address modern web best practices.

Any issue found is logged by the Talend quality assurance and security department and is resolved by Talend R&D in a timely manner according to the severity of the issue. Issue status is tracked by the development project management team, and reports are available upon request at Talend's discretion. When an issue is considered resolved, a re-test session is planned to validate that the patch correctly remediates the issue.

AUTHENTICATION AND AUTHORIZATION

On the client's side, tenants are isolated by their respective HTTPS connection to the Talend Cloud infrastructure. Client-side authentication is performed using tenant's specific credentials for Basic Authentication (e.g. login and password combination).

On the server side, X.509 certificates are deployed to authenticate Transport Layer Security (TLS) endpoints.

Once successfully authenticated, traditional session management through session cookies is used. All tenants share an identical Same-Origin Policy context at the web browser level.

Talend Cloud platform administrative access requires management review and approval. Elevated privilege access requires the same level of approval by management. Access to the platform console requires secure key authentication as well as user credentials. Console access to AWS requires multifactor authentication and user credentials. User account passwords follow industry standard best practice policies. User accounts are reviewed quarterly and annually by 3rd party SOC auditors.

Access to the AWS console is restricted to Talend Site Reliability Engineering (SRE) team members where new account creation must follow an approval process.

Server access is performed using a SSH private key. Public keys are automatically deployed with the Talend Cloud configuration management tool.

Enterprise Identity and Access Management (IAM)

Talend Cloud Management Console provides enterprise Identity and Access Management (IAM), including single sign-on (SSO) support for Okta and the SCIM (System for Cross-domain Identity Management) service. SCIM is an open standard that allows for the automation of user provisioning to make user data more secure and simplify the user experience. SSO with other identity providers can be set up with a plug-in.

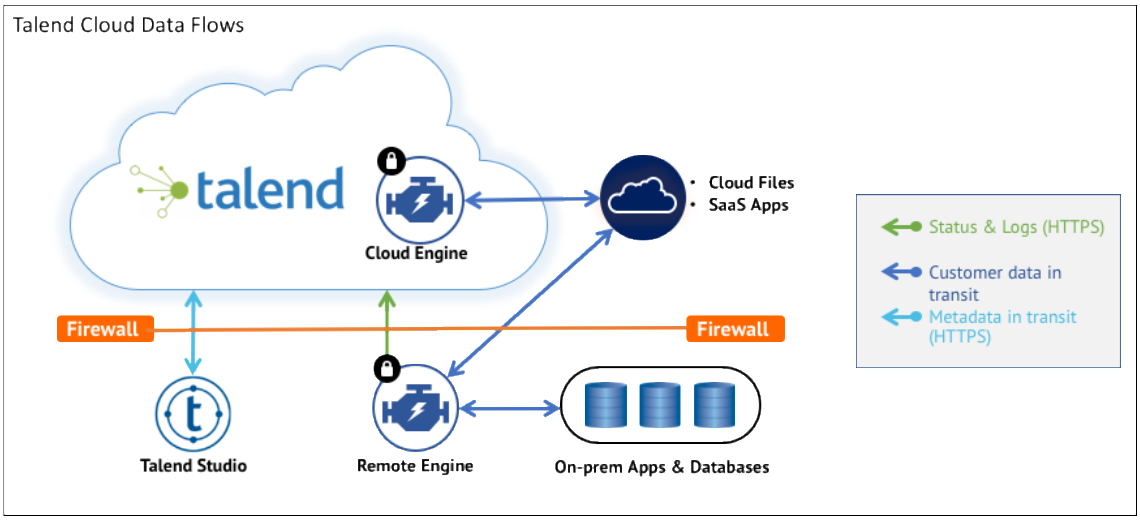
Key Management

The AWS services used by Talend Cloud rely on AWS-managed Customer Master Keys (CMK) for encryption. Talend Cloud uses its own AWS CMK to generate unique Data Encryption Keys (DEKs) for each tenant. The DEKs are managed through a Talend-developed Crypto service. Talend generates the keys and rotates them on the behalf of AWS every 3 years.

Front-end TLS endpoints are managed through the AWS Certificate Manager (ACM). The private key is generated by Talend's SRE team and the associated certificate signed by Talend's approved Certificate Authority (CA). User generated content hosting (API Documentation & API Mocks) is managed through 2 dedicated domains (resp: *.[dc].api-documentation.com and *.[dc].api-mocks.com). Each of them is protected by a dedicated wildcard ACM generated certificate.

SUMMARY OF DATA FLOWS

This section summarizes the data flows between Talend Cloud Management Console and Talend Studio or Remote Engines.



Data Flows Between Talend Studio and Talend Cloud

A question often asked is "What data is captured and transferred between Talend Studio and Talend Cloud Management Console?"

The kind of data being transferred is:

a) Task Artifact Binaries
b) Job Metadata (e.g. context variables and parameters)
c) Talend API Designer definitions

The metadata is transferred to Talend Cloud via our inventory service. It is reachable over HTTPS under the URL:

- US: https://tmc.us.cloud.talend.com/inventory
- Europe: https://tmc.eu.cloud.talend.com/inventory
- APAC: https://tmc.ap.cloud.talend.com/inventory

Endpoints providing the list of API designs:

- https://api-apid-service.us.cloud.talend.com/external/projects
- https://api-apid-service.eu.cloud.talend.com/external/projects
- https://api-apid-service.ap.cloud.talend.com/external/projects

Endpoints providing a given API design:

- https://api-apid-service.us.cloud.talend.com/external/projects/{projectId}
- https://api-apid-service.eu.cloud.talend.com/external/projects/{projectId}
- https://api-apid-service.ap.cloud.talend.com/external/projects/{projectId}

The user's login and password are required to authorize the transfer.

Data Flows Between Remote Engine and Talend Cloud

Another question asked is, "What data is captured and transferred between a Remote Engine and Talend Cloud Management Console?"

The Remote Engine ALWAYS initiates a TCP connection, and Talend Cloud NEVER initiates one. Everything is HTTPS.

The kind of data being transferred is:

a) Status Information and Metrics
b) Lifecycle Commands
c) Job Metadata
d) Logs
e) Task Artifact Binaries

The next sections discuss each data type in the scope of REST service URLs that are being targeted (and the corresponding systems behind them).

- **msg.us.cloud.talend.com** or **msg.eu.cloud.talend.com** or **msg.ap.cloud.talend.com** – this service is the primary gateway to Talend's ActiveMQ cluster. Data of type a) to d) is being transferred using this channel.
  Remote Engine status information and lifecycle commands is the first kind of data sent over the wire. This path is a control path to schedule flow deployments and capture execution status (success, fail). Other information transferred are the number of rows successfully processed or being rejected. This also includes the final success message.

- **repo.us.cloud.talend.com** or **repo.eu.cloud.talend.com** or **repo.ap.cloud.talend.com** – this service is the primary access point for job and action binaries. Behind this REST service, you will find the customer-specific Nexus repositories, which are only accessible via HTTPS and the customer's unique Nexus credentials. These credentials are propagated to the Remote Engine during Remote Engine pairing.

- **pair.us.cloud.talend.com** or **pair.eu.cloud.talend.com** or **pair.ap.cloud.talend.com** – this service is used during initial pairing of the Remote Engine to its account. Further, it is used to send the heartbeat, availability, and idle status of the engine itself. As with all other services, it is only accessible via HTTPS.

- **dts.us.cloud.talend.com** or **dts.eu.cloud.talend.com** or **dts.ap.cloud.talend.com** – this data transfer service is a token generation service. It is used to create one-time, time-limited tokens to authorize file uploads to Talend Cloud. The file transfer is an HTTPS POST from the Remote Engine to Talend Cloud, e.g log files or resource files.

Data Flows between Remote Engine for Pipelines and Talend Cloud

The Remote Engine ALWAYS initiates a TCP connection and Cloud NEVER initiates one. Everything is HTTPS.

The kind of data being transferred is:

a) Lifecycle Commands
b) Executable pipelines definition including all properties set at design time in pipeline designer (secrets are sent encrypted)
c) Dataset sample data (1,000 first records of datasets)
d) Preview requests and responses (including data)
e) Connectors metadata (connection, dataset, connectors, processors forms definition) and connectors dynamic function execution requests and response
f) Logs
g) Status information and metrics
h) Vault commands

The following Talend Cloud services are reached by Remote Engine for Pipelines: (replace <region> by either us, eu, ap)

- **pair.<region>.cloud.talend.com** is reached by the Remote Engine for Pipelines during its initiation phase in order to pair it with Talend Cloud. This endpoint is also reached when the Remote Engine for Pipelines is running in order to send a heartbeat to Talend Cloud. This service is only accessible via HTTPS.
- **vault-gateway.<region>.cloud.talend.com** is used by the Remote Engine for Pipelines to communicate with Vault. This service is only accessible via HTTPS.
- **engine.<region>.cloud.talend.com** is used by the Remote Engine for Pipelines to establish a Websocket connection with Talend Cloud. This Websocket connection will be used during all the engine's lifetime to exchange requests / responses both ways between the Remote Engine for Pipelines and Talend Cloud.