

Data Lake on the AWS Cloud with Talend Big Data Platform, AWS Services, and Cognizant Best Practices

Quick Start Deployment Guide

November 2017

Cognizant Technology Solutions

Talend, Inc.

AWS Quick Start Reference Team

Contents

Overview.....	2
Features	3
Optional Sample Dataset and Talend Jobs	4
Costs and Licenses.....	5
Architecture.....	5
Data Integration Architecture	5
Infrastructure Architecture	7
Prerequisites	9
Specialized Knowledge	9
Technical Requirements.....	9
Deployment Options	10
Deployment Steps	11
Step 1. Prepare Your AWS Account.....	11

Step 2. Upload Your Talend License File	11
Step 3. Launch the Quick Start	11
Step 4. Test the Deployment	23
Deleting the Stacks.....	28
Troubleshooting.....	28
Additional Resources	29
Send Us Feedback	30
Document Revisions	30

This Quick Start deployment guide was created by Amazon Web Services (AWS) in partnership with Cognizant Technology Solutions, an AWS Premier Consulting Partner, and Talend Inc., an AWS Advanced Technology Partner.

[Quick Starts](#) are automated reference deployments that use AWS CloudFormation templates to launch, configure, and run the AWS compute, network, storage, and other services required to deploy a specific workload on AWS.

Overview

This Quick Start reference deployment guide provides step-by-step instructions for deploying a data lake on the AWS Cloud, using AWS services with the Talend Big Data Platform.

The Quick Start provisions Talend Big Data Platform components and AWS services such as Amazon EMR, Amazon Redshift, Amazon Simple Storage Service (Amazon S3), and Amazon Relational Database Service (Amazon RDS) to build a data lake. It also provides an optional sample dataset and Talend jobs developed by Cognizant to illustrate big data practices for integrating Apache Spark, Apache Hadoop, Amazon EMR, Amazon Redshift, and Amazon S3 technologies into a data lake implementation.

Data lakes in the cloud are a key driver of digital transformation initiatives. They enable data and operational agility by enabling access to historical and real-time data for analytics. This Quick Start automates the design, setup, and configuration of hardware and software to implement a data lake in much less time than the traditional approach.

The Quick Start is for users who are evaluating big data in the cloud or looking to accelerate their big data initiative through the adoption of best practices for big data integration. It

illustrates basic big data integration patterns with Amazon S3, Amazon EMR, and Amazon Redshift using Talend technologies and best practices for DevOps and systems development from Cognizant and AWS.

Features

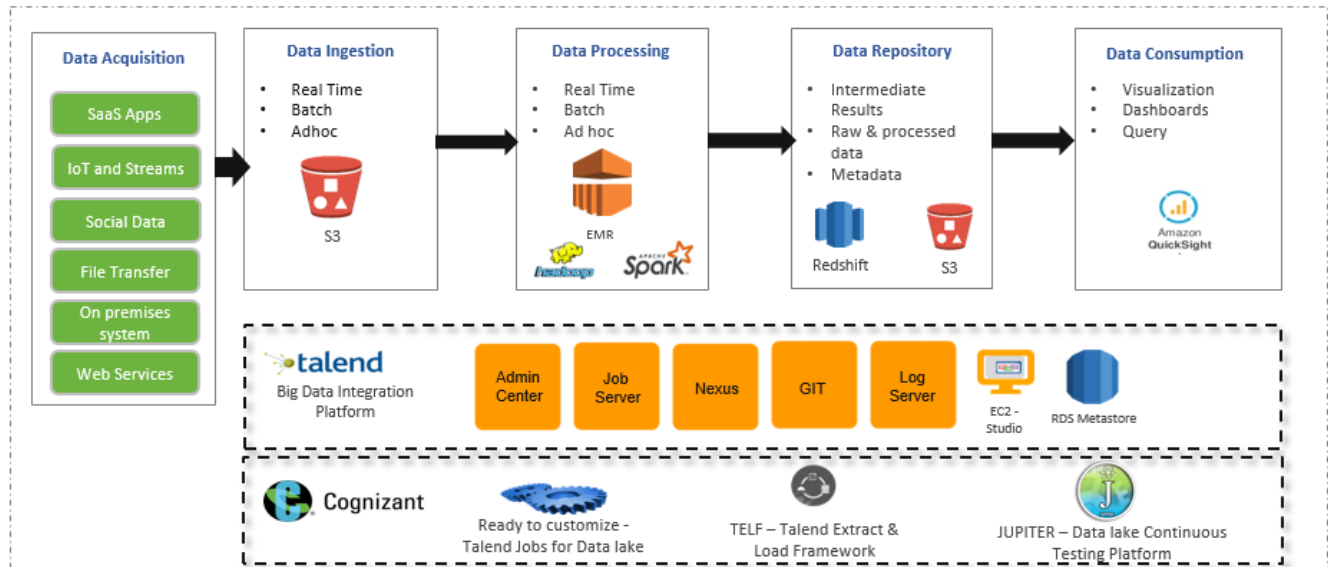


Figure 1: Quick Start features

The Quick Start provides the following features:

- Enables self-service by provisioning required services and components to build a data lake.
- Provides flexibility to spin up environments for development, test, and production.
- Includes an optional sample dataset and prebuilt Talend Spark jobs that help you explore the architecture and understand the stages of the end-to-end dataflow.
- Includes the following, using Talend and Spark capabilities:
 - Ingestion: Loading Amazon S3 data to Hadoop Distributed File System (HDFS) and Apache Hive
 - Data processing: Transformation and aggregation using various Spark and Hadoop features from Talend
 - Data repository: Loading and building a data warehouse using Amazon Redshift
- Optionally offers the Cognizant ingestion framework, big data validation, and DevOps platform to ingest, validate, and deploy big data solutions. These features aren't automated through the Quick Start CloudFormation template.

Optional Sample Dataset and Talend Jobs

The Quick Start dataset includes custom-built fitness data to demonstrate how data is submitted to, and ingested by, the data lake. You can use this data to perform predictive and descriptive analytics.

Figure 2 shows the organization of the sample data.

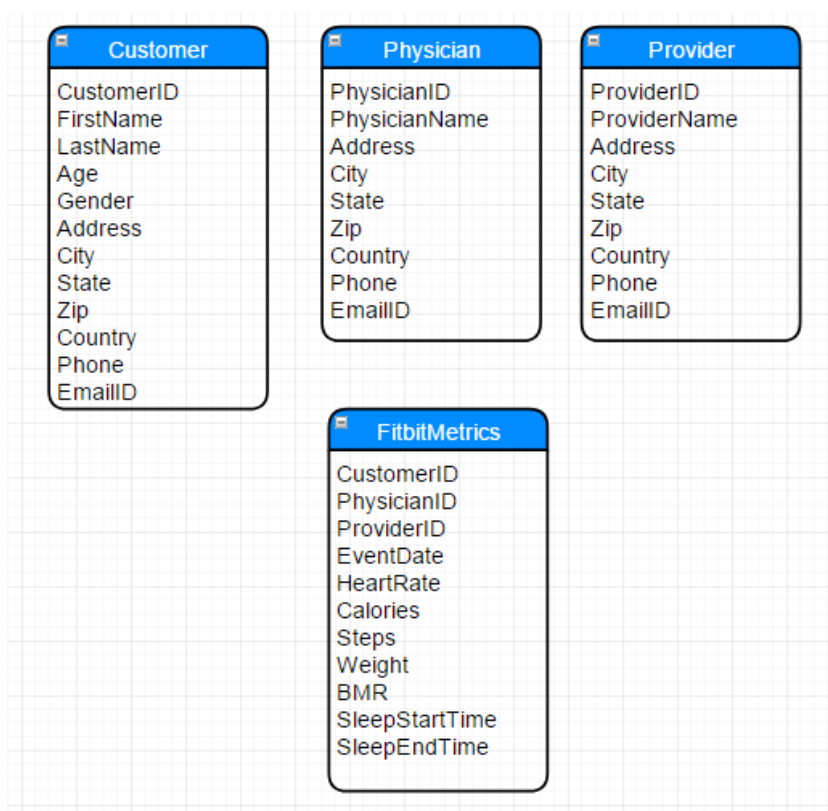


Figure 2: Database tables for fitness sample data

The workflow includes three top-level Talend jobs:

- Sample Talend job that sends the fitness data to Amazon S3 for ingestion
- Sample Talend job that uses Amazon EMR and Spark to perform aggregation, to look up and apply transformation rules on the fitness data, and to load the data into an Amazon S3 harmonized region
- Sample Talend job that loads data from Amazon S3 to an Amazon Redshift data warehouse

For questions or to troubleshoot your environment, use the [Talend AWS Quick Start community forum](#).

Costs and Licenses

You are responsible for the cost of the AWS services used while running this Quick Start reference deployment. There is no additional cost for using the Quick Start.

The AWS CloudFormation template for this Quick Start includes configuration parameters that you can customize. Some of these settings, such as instance type, will affect the cost of deployment. For cost estimates, see the pricing pages for each AWS service you will be using. Prices are subject to change.

You will need to provide your own Talend Big Data Platform license. To request a 30-day free trial license, please fill out the [registration form](#) on the Talend website. You'll receive a unique license key from Talend, which you'll use during the Quick Start deployment process, as explained in [step 2](#) of the deployment instructions.

The code for all Talend jobs included in the Quick Start are released under the [Apache License](#).

Architecture

Data Integration Architecture

The following diagram shows the data integration architecture.

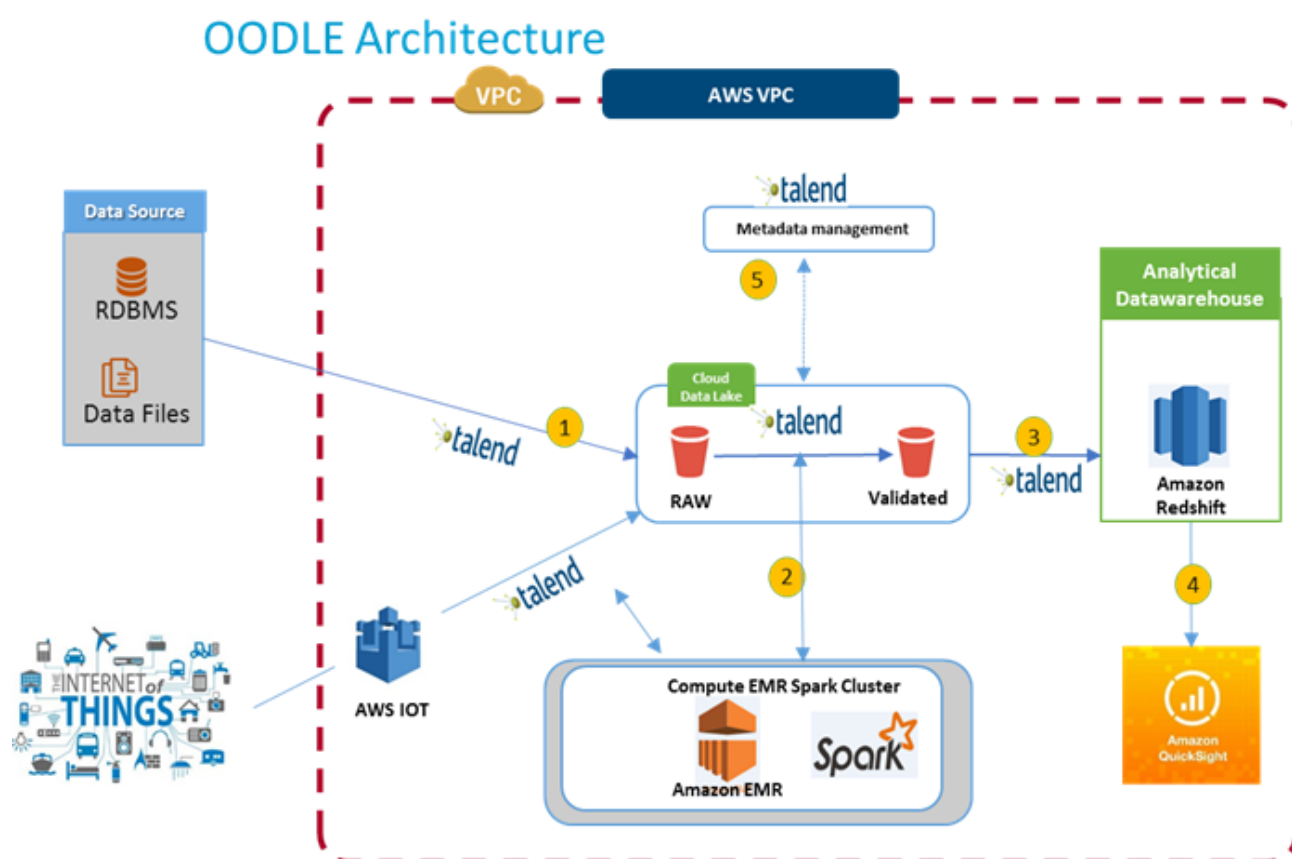


Figure 3: Data integration architecture for the Quick Start

The dataflow includes these steps:

Step 1	Ingest data from various types of sources such as RDBMS, flat files, semi-structured data sources, and streaming data to the raw S3 bucket.
Step 2	Apply data transformation and analytics on raw data by using Talend jobs and Amazon EMR Spark cluster to apply required transformation.
Step 3	Load the data from load-ready files to the analytical data warehouse in Amazon Redshift by using Talend jobs.

The dataflow is designed in Talend Studio and orchestrated by the Talend Big Data Platform.

- Talend Studio helps you create job templates using an easy to understand visual interface. It also provides metadata management capabilities.

- The Talend Big Data Platform then runs these jobs to take the data through the flow detailed in Figure 3.
- You can use the sample, prebuilt jobs included with the Quick Start to test the results of the system. The Quick Start features a number of these prebuilt jobs to demonstrate the flow and use of the system.

Optional AWS Components

You can add the following AWS components to the architecture after deployment. Note that the Quick Start templates do not deploy or configure these services.

- You can add Amazon QuickSight for visualization, analysis, and business insight.
- You can add AWS IoT for Internet of Things (IoT) data source integration, so you can analyze and store data from IoT devices.

Infrastructure Architecture

This Quick Start deploys the AWS and Talend components shown in Figure 4.

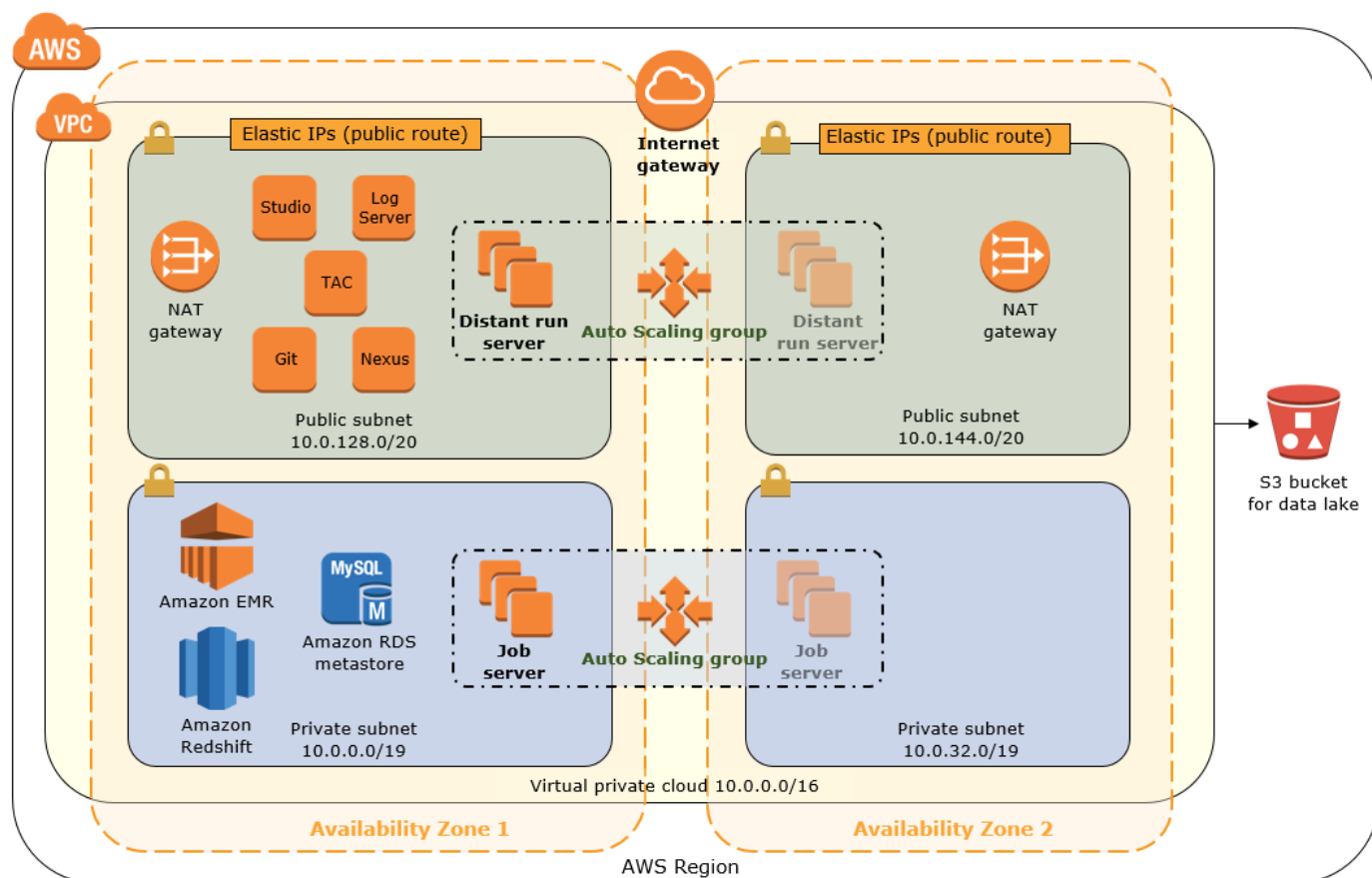


Figure 4: Quick Start architecture

The Quick Start sets up the following:

- A virtual private cloud (VPC) that spans two Availability Zones. Each Availability Zone contains two subnets: a public subnet to allow connecting over the internet and a private subnet for Talend job servers, Amazon Redshift, Amazon RDS, and Amazon EMR. (The private subnet in the second Availability Zone contains only the job servers.)*
- An internet gateway to allow access to the internet. This gateway is used by the bastion hosts to send and receive traffic.*
- In the public subnets, managed network address translation (NAT) gateways to allow outbound internet access for resources in the private subnets.*
- In one or both public subnets, Linux bastion hosts to allow inbound Secure Shell (SSH) access to the resources in the private subnets. You can choose the number of bastion hosts when you launch the Quick Start.*
- In the public subnet in the first Availability Zone:
 - Talend public servers that host the Talend Administration Center (TAC) for administering Talend jobs via the browser.
 - A Talend Studio remote desktop instance available through an X2Go client for users who do not want to run Talend Studio on their laptops.
 - A Nexus artifact repository and Git servers for binary and source configuration management.
 - A Talend log server using Amazon Elasticsearch Service (Amazon ES), Logstash, and Kibana.
- In the private subnet in the first Availability Zone:
 - An Amazon RDS MySQL DB instance to host Talend metadata.
 - An Amazon EMR cluster with Pig, Hive, and Spark that integrates closely with the Talend Big Data Platform and provides Hadoop capability in the data lake.
 - An Amazon Redshift cluster for use as a data warehouse or data mart.
- In the private subnets, Talend job server instances running Talend jobs scheduled by the TAC, in an Auto Scaling group. Auto Scaling allows EC2 instances to be automatically spun up or down to respond to the demand on the Talend job servers. You can configure the desired and maximum number of instances during deployment.
- In the public subnets, Talend distant run job server instances running Talend jobs on behalf of Talend Studio users, in an Auto Scaling group. You can run Talend jobs locally on Talend Studio or on these servers. The Auto Scaling group allows EC2 instances to be

automatically spun up or down to respond to the demand on the Talend job servers. You can set the desired and maximum number of instances during deployment.

- Amazon S3 to ingest data for the data lake.

* You can choose to create a new VPC for the data lake deployment or use your existing VPC on AWS. The template that deploys the Quick Start into an existing VPC skips the components marked by asterisks.

Prerequisites

Specialized Knowledge

Before you deploy this Quick Start, we recommend that you become familiar with the following AWS services. (If you are new to AWS, see the [Getting Started Resource Center](#).)

- [Amazon VPC](#)
- [Amazon Elastic Compute Cloud \(Amazon EC2\)](#)
- [Amazon S3](#)
- [Amazon RDS](#)
- [Amazon Redshift](#)
- [Amazon EMR](#)

You may need to know about the following supporting components of AWS as well:

- [AWS Identity and Access Management \(IAM\)](#)
- [Amazon CloudWatch](#)

You should also read the [documentation on Talend Open Studio for Data Integration](#) on the Talend website.

The [Talend website](#) also provides useful information that will help you get familiar with how Talend powers big data ecosystems.

Technical Requirements

Talend License

You'll need a valid Talend license to deploy this Quick Start. For more information, see the [Cost and Licenses](#) section and [step 2](#) of the deployment instructions.

Talend Studio

You can choose to either run a local instance of Talend Studio on your laptop, or use a remote Talend Studio instance running in the AWS Cloud. The remote instance is accessible through X2Go, which is a remote desktop client that can run on Microsoft Windows or Linux. Using a remote instance minimizes the amount of setup. By default, the Quick Start sets up a single remote Talend Studio instance, so we recommend that you use a local instance on your laptop if you have a larger team. See the [Talend website](#) for details on installing the X2Go client and Talend Studio on your local laptop.

Using Your Own TAC Database and Git Server

By default, the Quick Start provisions a TAC database and a GitLab server for the data lake. However, setting up these components might take a while, so if you prefer, you can choose not to provision these or provide your own TAC database and Git server.

- **TAC database** – The Quick Start uses MySQL as the TAC database. During deployment, you can either configure the parameters under *Talend Administration Center configuration* to provide a running instance of MySQL with an empty database and credentials to be used by TAC, or you can leave these fields empty and the Quick Start will create an Amazon RDS MySQL DB instance and a TAC database for you. If you are planning to do frequent testing, or if you want to have continuity between tests, we recommend that you set up separate instances of these servers and then pass the host and credentials to the stack.
- **Git server** – Like the RDS instances, Git takes a while to provision. If you want to have continuity between your test runs, you may want to run a Git server separately or use a Git service such as GitHub. During the deployment, you can use the parameters under *Talend Git configuration* to specify your choices. If you do not provide a Git server, the Quick Start will provision a GitLab server for you. If you provide your own instance, you can use any of the Git servers supported by Talend. In this case, you need to provide the Git host name as well as credentials.

Deployment Options

This Quick Start provides two deployment methods:

- **Deploy the Quick Start into a new VPC** (end-to-end deployment). This option builds a new AWS environment consisting of the VPC, subnets, NAT gateways, security groups, bastion hosts, and other infrastructure components, and then deploys the Talend and AWS components into this new VPC.
- **Deploy the Quick Start into an existing VPC**. This option provisions all infrastructure and Talend servers in your existing AWS infrastructure.

Deployment Steps

Step 1. Prepare Your AWS Account

1. If you don't already have an AWS account, create one at <https://aws.amazon.com> by following the on-screen instructions.
1. Use the region selector in the navigation bar to choose the AWS Region where you want to deploy the Quick Start. Check the [AWS Region table](#) to make sure that the AWS services and EC2 instance types you intend to use are available in that specific region.
2. Create a [key pair](#) in your preferred region.
3. If necessary, [request a service limit increase](#) for EC2 instances. You might need to do this if you have an existing deployment that already uses the instance types you're planning to use (T2, M4, and C4 by default; see the parameter tables in [step 3](#)), and you think you might exceed the [default limits](#) with this reference deployment.

Step 2. Upload Your Talend License File

You will need a valid Talend Big Data Platform license to deploy this Quick Start.

You can use your existing Talend Big Data Platform license for this Quick Start deployment, subject to standard licensing terms and conditions.

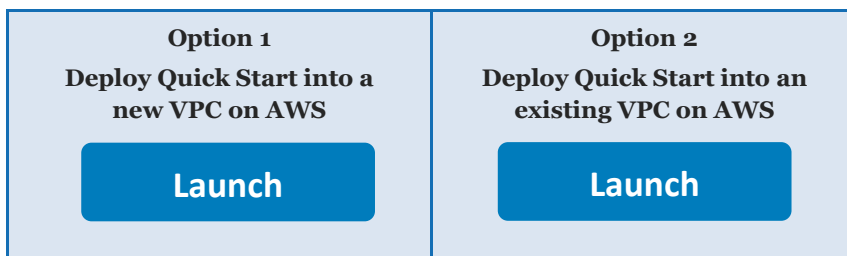
If you're not yet a Talend customer, you can request a 30-day free trial license by filling out the [registration form](#) on the Talend website. You'll receive an email with a unique license key from Talend.

When you've obtained a valid license, create a private S3 bucket, upload the license file to the root of this bucket, and note its URL. You'll need to supply the URL in the **Talend License Bucket** parameter when you launch the Quick Start in step 3.

Step 3. Launch the Quick Start

Note You are responsible for the cost of the AWS services used while running this Quick Start reference deployment. There is no additional cost for using this Quick Start. For full details, see the pricing pages for each AWS service you will be using in this Quick Start. Prices are subject to change.

1. Choose one of the following options to launch the AWS CloudFormation template into your AWS account. For help choosing an option, see [deployment options](#) earlier in this guide.



Important If you're deploying the Quick Start into an existing VPC, make sure that your VPC has two private subnets in different Availability Zones. These subnets require NAT gateways or NAT instances in their route tables, to allow the instances to download packages and software without exposing them to the internet. You'll also need the domain name option configured in the DHCP options as explained in the [Amazon VPC documentation](#). You'll be prompted for your VPC settings when you launch the Quick Start.

Each deployment takes about one hour to complete.

2. Check the region that's displayed in the upper-right corner of the navigation bar, and change it if necessary. This is where the network infrastructure for the data lake will be built. The template is launched in the US East (Ohio) Region by default.
3. On the **Select Template** page, keep the default setting for the template URL, and then choose **Next**.
4. On the **Specify Details** page, change the stack name if needed. Review the parameters for the template. Provide values for the parameters that require input. For all other parameters, review the default settings and customize them as necessary. When you finish reviewing and customizing the parameters, choose **next**.

In the following tables, parameters are listed by category and described for the deployment option:

- [Parameters for deploying the Quick Start into a new VPC](#)
- [Parameters for deploying the Quick Start into an existing VPC](#)

- **Option 1: Parameters for deploying the Quick Start into a new VPC**

[View template](#)

Network configuration:

Parameter label (name)	Default	Description
Availability Zones (AvailabilityZones)	<i>Requires input</i>	The list of Availability Zones to use for the subnets in the VPC. The Quick Start uses two Availability Zones from your list and preserves the logical order you specify.
VPC CIDR (VPCCIDR)	10.0.0.0/16	Classless Inter-Domain Routing (CIDR) block that consists of a range of IPv4 addresses for the new VPC. You can use the default CIDR settings or assign your own CIDR ranges for the VPC and subnets.
Public Subnet 1 CIDR (PublicSubnet1CIDR)	10.0.128.0/20	CIDR block for the public (DMZ) subnet located in Availability Zone 1.
Public Subnet 2 CIDR (PublicSubnet2CIDR)	10.0.144.0/20	CIDR block for the public (DMZ) subnet located in Availability Zone 2.
Private Subnet 1 CIDR (PrivateSubnet1CIDR)	10.0.0.0/19	CIDR block for the private subnet located in Availability Zone 1.
Private Subnet 2 CIDR (PrivateSubnet2CIDR)	10.0.32.0/19	CIDR block for the private subnet located in Availability Zone 2.
Remote Access CIDR (RemoteAccessCIDR)	<i>Requires input</i>	The CIDR IP range that is permitted to access the VPC. We recommend that you use a constrained CIDR range to reduce the potential of inbound attacks from unknown IP addresses. For example, if your IPv4 address is 203.0.113.25, specify 203.0.113.25/32 to list this single IPv4 address in CIDR notation. If your company allocates addresses from a range, specify the entire range, such as 203.0.113.0/24. For details, see VPCs and Subnets in the AWS documentation.

Creation options for existing deployments:

Parameter label (name)	Default	Description
Create Distant Run Stack (CreateDistantRunStack)	true	Set this parameter to false if you want to run Talend jobs locally from Talend Studio instead of submitting them remotely. Keep the default setting to create a Talend job server Auto Scaling group for remote job submission from Talend Studio.
Create Amazon EMR (CreateEmr)	true	Set this parameter to false if you don't want to create a new Amazon EMR instance for data transformation and analytics, or if you have an existing Amazon EMR instance you'd like to use. Keep the default setting to provision a new Amazon EMR instance.

Parameter label (name)	Default	Description
Create TAC Database (CreateTacDatabase)	true	Set this parameter to false if you don't want to create a new Talend Administration Center (TAC) database or if you want to use an existing database. Keep the default setting to set up a new TAC database.
Create Studio Stack (CreateStudioStack)	true	Set this parameter to false if you don't want to use Talend Studio. Keep the default setting if you want to set up Talend Studio during deployment.

EC2 Auto Scaling configuration:

Parameter label (name)	Default	Description
Jobserver Autoscale Desired Capacity (JobserverAutoscaleDesiredCapacity)	1	Desired number of EC2 instances for the Talend job server Auto Scaling group. You can specify up to 10 instances.
Jobserver Autoscale Maximum Capacity (JobserverAutoscaleMaxSize)	5	Maximum number of EC2 instances for the Talend job server Auto Scaling group. You can specify up to 10 instances.
DistantRun Autoscale Desired Capacity (DistantRunAutoscaleDesiredCapacity)	1	Desired number of EC2 instances for the Talend distant run Auto Scaling group. You can specify up to 10 instances.
DistantRun Autoscale Maximum Capacity (DistantRunAutoscaleMaxSize)	5	Maximum number of EC2 instances for the Talend distant run Auto Scaling group. You can specify up to 10 instances.

Amazon Redshift configuration:

Parameter label (name)	Default	Description
Amazon Redshift Host (optional) (RedshiftHost)	—	DNS name or IP address of the master node of an existing Amazon Redshift cluster that you intend to use for the Talend sample jobs. Leave this parameter blank to create a new Amazon Redshift cluster.
Amazon Redshift Username (RedshiftUsername)	tadmin	User name for the Amazon Redshift database.
Amazon Redshift Password (RedshiftPassword)	<i>Requires input</i>	Password for the Amazon Redshift database. This must be an 8-28 character string that contains only alphanumeric characters or the following special characters: ! ^ * - _ +
Amazon Redshift Database Name (RedshiftDbName)	—	Database name for Amazon Redshift.

Talend Administration Center configuration:

Parameter label (name)	Default	Description
TAC Database Host (optional) (TacDbHost)	—	Name or IP address of the host for an existing MySQL database that you intend to use as the TAC database. Leave this blank to create a new MySQL database for TAC. For additional information, see Using Your Own TAC Database and Git Server earlier in this guide.
TAC Master Database User (MasterDbUser)	tadmin	The master or root user used to create TAC and Activity Monitoring Console (AMC) databases and the TAC user. This parameter is needed only if you're creating a new TAC or AMC database.
TAC Master Database Password (MasterDbPassword)	<i>Requires input</i>	Master user database password. This parameter is needed only if you're creating a new TAC or AMC database.
TAC Database Schema (TacDbSchema)	tac_quickstart	Existing database schema for the TAC database.
TAC Password (TacPassword)	<i>Requires input</i>	TAC application password for the <i>tadmin</i> account.
TAC Database Username (TacDbUser)	tac	Existing database user name for TAC.
TAC Database Password (TacDbPassword)	<i>Requires input</i>	Existing database password for TAC.
Database Instance Class (DbClass)	db.t2.medium	Instance class of the Amazon RDS MySQL DB instance that will be used as the TAC database.
Database Allocated Storage (DbAllocatedStorage)	20	Allocated storage (in GiB) for the RDS instance.
AMC Database Username (AmcDbUser)	amc	Database user name for AMC.
AMC Database Password (AmcDbPassword)	<i>Requires input</i>	Database password for AMC.
Talend Resource Bucket (TalendResourceBucket)	repo-quickstart-talend	The S3 bucket that contains Talend resources. The default bucket is prepopulated with Talend installation binaries, so no additional action is required from the user.
Talend License Bucket (TalendLicenseBucket)	<i>Requires input</i>	The S3 bucket that contains the Talend license, from step 2 .

Talend Nexus configuration:

Parameter label (name)	Default	Description
Nexus Admin User ID (NexusAdminUserid)	admin	Administrator user ID for the Nexus server.
Nexus Admin Password (NexusAdminPassword)	<i>Requires input</i>	Password for the Nexus server.

Talend Git configuration:

Parameter label (name)	Default	Description
Git Protocol (GitProtocol)	http	Git protocol.
Git Host (optional) (GitHost)	—	Host name of your Git server. Leave this parameter blank if you want the Quick Start to provision a GitLab server. For additional information, see Using Your Own TAC Database and Git Server earlier in this guide.
Git TCP Port (GitPort)	80	Port number of the Git TCP port.
Git Repository (GitRepo)	oodlejobs	Name of Git repository.
Git Admin User ID (GitAdminUserid)	tadmin	User ID of Git administrator.
Git Admin Password (GitAdminPassword)	<i>Requires input</i>	Password for Git administrator.
Git Admin Email (GitAdminEmail)	—	Email address of Git administrator.
Git TAC User ID (GitTacUserid)	tac	User ID for Git TAC.
Git TAC Password (GitTacPassword)	<i>Requires input</i>	Password for Git TAC.
Git TAC Email (GitTacEmail)	—	Contact email address for TAC.

Sizing configuration:

Parameter label (name)	Default	Description
Key Pair Name (KeyPairName)	<i>Requires input</i>	Public/private key pair, which allows you to connect securely to your instance after it launches. When you created an AWS account, this is the key pair you created in your preferred region.

Parameter label (name)	Default	Description
TAC Instance Type (TacInstanceType)	t2.medium	EC2 instance type for the Talend Administration Center.
Logserver Instance Type (LogserverInstanceType)	t2.medium	EC2 instance type for the log server.
Jobserver Instance Type (JobserverInstanceType)	t2.large	EC2 instance type for Talend job servers.
Nexus Instance Type (NexusInstanceType)	t2.medium	EC2 instance type for Nexus server.
Studio Instance Type (StudioInstanceType)	m4.xlarge	EC2 instance type for Talend Studio.
Git Instance Type (GitInstanceType)	m4.large	EC2 instance type for the Git server.
Bastion Instance Type (BastionInstanceType)	t2.micro	EC2 instance type for the bastion host.
Number of Bastion Hosts (NumBastionHosts)	1	The number of Linux bastion hosts to run. Auto Scaling will ensure that you always have this number of bastion hosts running. You can specify 0, 1, or 2 hosts.
Amazon Redshift Node Type (RedshiftNodeType)	dc2.large	EC2 instance type for the Amazon Redshift cluster nodes.
Number of Amazon Redshift Nodes (RedshiftNumberOfNodes)	1	The number of nodes in the Amazon Redshift cluster.
Amazon EMR Master Node Instance Type (EmrMasterInstanceType)	c4.xlarge	Instance type for the Amazon EMR master node.
Amazon EMR Core Node Instance Type (EmrCoreInstanceType)	c4.xlarge	Instance type for the Amazon EMR core nodes.
Number of Amazon EMR Core Nodes (EmrCoreNodes)	2	Number of Amazon EMR core nodes. You can specify 1-500 nodes.

Quick Start Configuration:

Parameter label (name)	Default	Description
Quick Start S3 Bucket Name (QSS3BucketName)	quickstart-reference	S3 bucket where the Quick Start templates and scripts are installed. Use this parameter to specify the S3 bucket name you've created for your copy of Quick Start assets, if you decide to customize or extend the Quick Start for your own use. The bucket name can include numbers,

Parameter label (name)	Default	Description
		lowercase letters, uppercase letters, and hyphens, but should not start or end with a hyphen.
Quick Start S3 Key Prefix (QSS3KeyPrefix)	datalake/cognizant /talend/latest/	The S3 key name prefix used to simulate a folder for your copy of Quick Start assets, if you decide to customize or extend the Quick Start for your own use. This prefix can include numbers, lowercase letters, uppercase letters, hyphens, and forward slashes.

- **Option 2: Parameters for deploying the Quick Start into an existing VPC**

[View template](#)

Network configuration:

Parameter label (name)	Default	Description
Remote Access CIDR (RemoteAccessCIDR)	<i>Requires input</i>	The CIDR IP range that is permitted to access the VPC. We recommend that you use a constrained CIDR range to reduce the potential of inbound attacks from unknown IP addresses. For example, if your IPv4 address is 203.0.113.25, specify 203.0.113.25/32 to list this single IPv4 address in CIDR notation. If your company allocates addresses from a range, specify the entire range, such as 203.0.113.0/24. For details, see VPCs and Subnets in the AWS documentation.
VPC ID (VpcId)	<i>Requires input</i>	ID of your existing VPC where the AWS and Talend resources will be deployed (e.g., vpc-0343606e).
Public Subnet 1 ID (PublicSubnetId1)	<i>Requires input</i>	ID of the public subnet in Availability Zone 1 in your existing VPC (e.g., subnet-a0246dcd).
Public Subnet 2 ID (PublicSubnetId2)	<i>Requires input</i>	ID of the public subnet in Availability Zone 2 in your existing VPC.
Private Subnet 1 ID (PrivateSubnetId1)	<i>Requires input</i>	ID of the private subnet in Availability Zone 1 in your existing VPC (e.g., subnet-b58c3d67).
Private Subnet 2 ID (PrivateSubnetId2)	<i>Requires input</i>	ID of the private subnet in Availability Zone 2 in your existing VPC.

Creation options for existing deployments:

Parameter label (name)	Default	Description
Create Distant Run Stack (CreateDistantRunStack)	true	Set this parameter to false if you want to run Talend jobs locally from Talend Studio instead of submitting them remotely. Keep the default setting to create a Talend job server Auto Scaling group for remote job submission from Talend Studio.

Parameter label (name)	Default	Description
Create Amazon EMR (CreateEmr)	true	Set this parameter to false if you don't want to create a new Amazon EMR instance for data transformation and analytics, or if you have an existing Amazon EMR instance you'd like to use. Keep the default setting to provision a new Amazon EMR instance.
Create TAC Database (CreateTacDatabase)	true	Set this parameter to false if you don't want to create a new Talend Administration Center (TAC) database or if you want to use an existing database. Keep the default setting to set up a new TAC database.
Create Studio Stack (CreateStudioStack)	true	Set this parameter to false if you don't want to use Talend Studio. Keep the default setting if you want to set up Talend Studio during deployment.

Sizing configuration:

Parameter label (name)	Default	Description
Key Pair Name (KeyPairName)	<i>Requires input</i>	Public/private key pair, which allows you to connect securely to your instance after it launches. When you created an AWS account, this is the key pair you created in your preferred region.
TAC Instance Type (TacInstanceType)	t2.medium	EC2 instance type for the Talend Administration Center.
Logserver Instance Type (LogserverInstanceType)	t2.medium	EC2 instance type for the log server.
Jobserver Instance Type (JobserverInstanceType)	t2.large	EC2 instance type for Talend job servers.
Nexus Instance Type (NexusInstanceType)	t2.medium	EC2 instance type for Nexus server.
Studio Instance Type (StudioInstanceType)	m4.xlarge	EC2 instance type for Talend Studio.
Git Instance Type (GitInstanceType)	m4.large	EC2 instance type for the Git server.
Bastion Instance Type (BastionInstanceType)	t2.micro	EC2 instance type for the bastion host.
Amazon EMR Core Node Instance Type (EmrCoreInstanceType)	c4.xlarge	Instance type for the Amazon EMR core nodes.
Amazon EMR Master Node Instance Type (EmrMasterInstanceType)	c4.xlarge	Instance type for the Amazon EMR master node.

Parameter label (name)	Default	Description
Amazon Redshift Node Type (RedshiftNodeType)	dc2.large	EC2 instance type for the Amazon Redshift cluster nodes.
Number of Amazon EMR Core Nodes (EmrCoreNodes)	2	Number of Amazon EMR core nodes. You can specify 1-500 nodes.
Number of Bastion Hosts (NumBastionHosts)	1	The number of Linux bastion hosts to run. Auto Scaling will ensure that you always have this number of bastion hosts running. You can specify 0, 1, or 2 hosts.
Number of Amazon Redshift Nodes (RedshiftNumberOfNodes)	1	The number of nodes in the Amazon Redshift cluster.

Auto Scaling configuration:

Parameter label (name)	Default	Description
Jobserver Autoscale Desired Capacity (JobserverAutoscaleDesiredCapacity)	1	Desired number of EC2 instances for the Talend job server Auto Scaling group. You can specify up to 10 instances.
Jobserver Autoscale Maximum Capacity (JobserverAutoscaleMaxSize)	5	Maximum number of EC2 instances for the Talend job server Auto Scaling group. You can specify up to 10 instances.
DistantRun Autoscale Desired Capacity (DistantRunAutoscaleDesiredCapacity)	1	Desired number of EC2 instances for the Talend distant run Auto Scaling group. You can specify up to 10 instances.
DistantRun Autoscale Maximum Capacity (DistantRunAutoscaleMaxSize)	5	Maximum number of EC2 instances for the Talend distant run Auto Scaling group. You can specify up to 10 instances.

Amazon Redshift configuration:

Parameter label (name)	Default	Description
Amazon Redshift Host (optional) (RedshiftHost)	—	DNS name or IP address of the master node of an existing Amazon Redshift cluster that you intend to use for the Talend sample jobs. Leave this parameter blank to create a new Amazon Redshift cluster.
Amazon Redshift Username (RedshiftUsername)	tadmin	User name for the Amazon Redshift database.
Amazon Redshift Password (RedshiftPassword)	<i>Requires input</i>	Password for the Amazon Redshift database. This must be an 8-28 character string that contains only

Parameter label (name)	Default	Description
		alphanumeric characters or the following special characters: ! ^ * - _ +
Amazon Redshift Database Name (RedshiftDbName)	—	Database name for Amazon Redshift.

Talend Administration Center configuration:

Parameter label (name)	Default	Description
TAC Database Host (optional) (TacDbHost)	—	Name or IP address of the host for an existing MySQL database that you intend to use as the TAC database. Leave this blank to create a new MySQL database for TAC. For additional information, see Using Your Own TAC Database and Git Server earlier in this guide.
TAC Master Database User (MasterDbUser)	tadmin	The master or root user used to create TAC and Activity Monitoring Console (AMC) databases and the TAC user. This parameter is needed only if you're creating a new TAC or AMC database.
TAC Master Database Password (MasterDbPassword)	<i>Requires input</i>	Master user database password. This parameter is needed only if you're creating a new TAC or AMC database.
TAC Database Schema (TacDbSchema)	tac_quickstart	Existing database schema for the TAC database.
TAC Password (TacPassword)	<i>Requires input</i>	TAC application password for the <i>tadmin</i> account.
TAC Database Username (TacDbUser)	tac	Existing database user name for TAC.
TAC Database Password (TacDbPassword)	<i>Requires input</i>	Existing database password for TAC.
Database Instance Class (DbClass)	db.t2.medium	Instance class of the Amazon RDS MySQL DB instance that will be used as the TAC database.
Database Allocated Storage (DbAllocatedStorage)	20	Allocated storage (in GiB) for the RDS instance.
AMC Database Username (AmcDbUser)	amc	Database user name for AMC.
AMC Database Password (AmcDbPassword)	<i>Requires input</i>	Database password for AMC.
Talend Resource Bucket (TalendResourceBucket)	repo-quickstart-talend	The S3 bucket that contains Talend resources. The default bucket is prepopulated with Talend installation binaries, so no additional action is required from the user.

Parameter label (name)	Default	Description
Talend License Bucket (TalendLicenseBucket)	<i>Requires input</i>	The S3 bucket that contains the Talend license, from step 2 .

Talend Nexus configuration:

Parameter label (name)	Default	Description
Nexus Admin User ID (NexusAdminUserid)	admin	Administrator user ID for the Nexus server.
Nexus Admin Password (NexusAdminPassword)	<i>Requires input</i>	Password for the Nexus server.

Talend Git configuration:

Parameter label (name)	Default	Description
Git Protocol (GitProtocol)	http	Git protocol.
Git Host (optional) (GitHost)	—	Host name of your Git server. Leave this parameter blank if you want the Quick Start to provision a GitLab server. For additional information, see Using Your Own TAC Database and Git Server earlier in this guide.
Git TCP Port (GitPort)	80	Port number of the Git TCP port.
Git Repository (GitRepo)	oodlejobs	Name of Git repository.
Git Admin User ID (GitAdminUserid)	tadmin	User ID of Git administrator.
Git Admin Password (GitAdminPassword)	<i>Requires input</i>	Password for Git administrator.
Git Admin Email (GitAdminEmail)	—	Email address of Git administrator.
Git TAC User ID (GitTacUserid)	tac	User ID for Git TAC.
Git TAC Password (GitTacPassword)	<i>Requires input</i>	Password for Git TAC.
Git TAC Email (GitTacEmail)	—	Contact email address for TAC.

Quick Start Configuration:

Parameter label (name)	Default	Description
Quick Start S3 Bucket Name (QSS3BucketName)	quickstart-reference	S3 bucket where the Quick Start templates and scripts are installed. Use this parameter to specify the S3 bucket name you've created for your copy of Quick Start assets, if you decide to customize or extend the Quick Start for your own use. The bucket name can include numbers, lowercase letters, uppercase letters, and hyphens, but should not start or end with a hyphen.
Quick Start S3 Key Prefix (QSS3KeyPrefix)	datalake/cognizant/talend/latest/	The S3 key name prefix used to simulate a folder for your copy of Quick Start assets, if you decide to customize or extend the Quick Start for your own use. This prefix can include numbers, lowercase letters, uppercase letters, hyphens, and forward slashes.

- On the **Options** page, you can [specify tags](#) (key-value pairs) for resources in your stack and [set advanced options](#). When you're done, choose **next**.
- On the **Review** page, review and confirm the template settings. Under **Capabilities**, select the check box to acknowledge that the template will create IAM resources.
- Choose **Create** to deploy the stack.
- Monitor the status of the stack. When the status is **CREATE_COMPLETE**, the cluster is ready.

You can use the URLs displayed in the **Outputs** tab for the stack to view the resources that were created, as described in the next step.

Step 4. Test the Deployment

To test the deployment, you'll need information from the **Outputs** and **Parameters** tabs of the AWS CloudFormation console.

Outputs – In the AWS CloudFormation console at <https://console.aws.amazon.com/cloudformation>, choose the **Outputs** tab for the parent stack.

▼ Outputs			
Key	Value	Description	Export Name
GitPublicDnsName		Git DNS	testoodlestack:GitPublicDnsName
OodleStack		Nested Oodle stack	testoodlestack:OodleStack
RedshiftJDBC		Redshift JDBC Url	testoodlestack:RedshiftJDBC
TalendSourceBucket		Talend Source Bucket	testoodlestack:TalendSourceBucket
EmrMasterPublicDns		EMR public DNS	testoodlestack:EmrMasterPublicDns
VpcStack		Nested VPC stack	testoodlestack:VpcStack
NexusPublicDnsName		Nexus public DNS	testoodlestack:NexusPublicDnsName
RedshiftEndpoint		Redshift Endpoint	testoodlestack:RedshiftEndpoint
EmrLogBucket		Emr Log Bucket	testoodlestack:EmrLogBucket
StudioPublicDnsName		Studio public URL	testoodlestack:StudioPublicDnsName
CredentialBucket		Talend Credential Bucket	testoodlestack:CredentialBucket

Figure 5: Quick Start stack outputs

You'll see the server host name and the following details, which you'll need to test the deployment:

- Talend credential bucket
- Talend source bucket
- Talend target bucket
- TAC URL
- Git URL
- Nexus URL
- X Window Studio DNS
- Amazon EMR master node DNS
- Amazon Redshift host

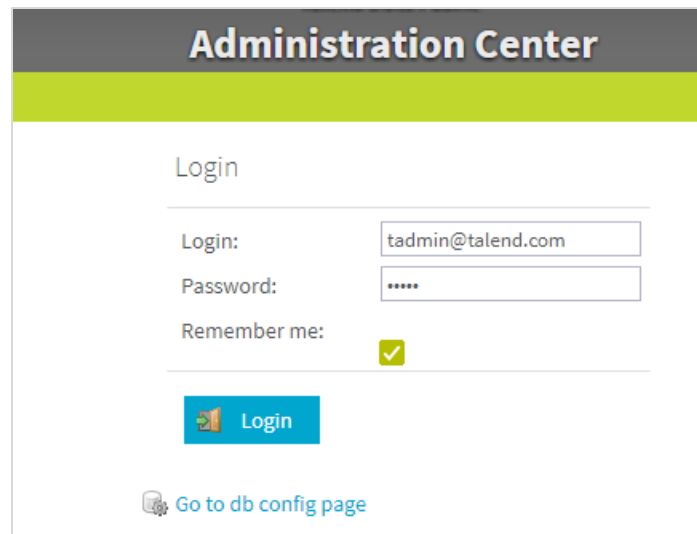
Parameters – In the AWS CloudFormation console, choose the **Parameters** tab for the main stack. This tab displays the TAC user name, job server, and details for Amazon Redshift, Git, and the Nexus server. These details will also be required to test this deployment.

Overview	Outputs	Resources	Events	Template	Parameters	Tags	Stack Policy	Change Sets
RedshiftDbName						redshiftdbname		
RedshiftHost								
RedshiftNodeType						dc1.large		
RedshiftNumberOfNodes						1		
RedshiftPassword						****		
RedshiftUsername						tadmin		
RemoteAccessCIDR						0.0.0.0/0		
StudioInstanceType						c4.xlarge		
TacDbHost								
TacDbPassword						****		
TacDbSchema						tac_quickstart		
TacDbUser						tac		
TacInstanceType						t2.medium		

Figure 6: Stack parameters

To test the deployment:

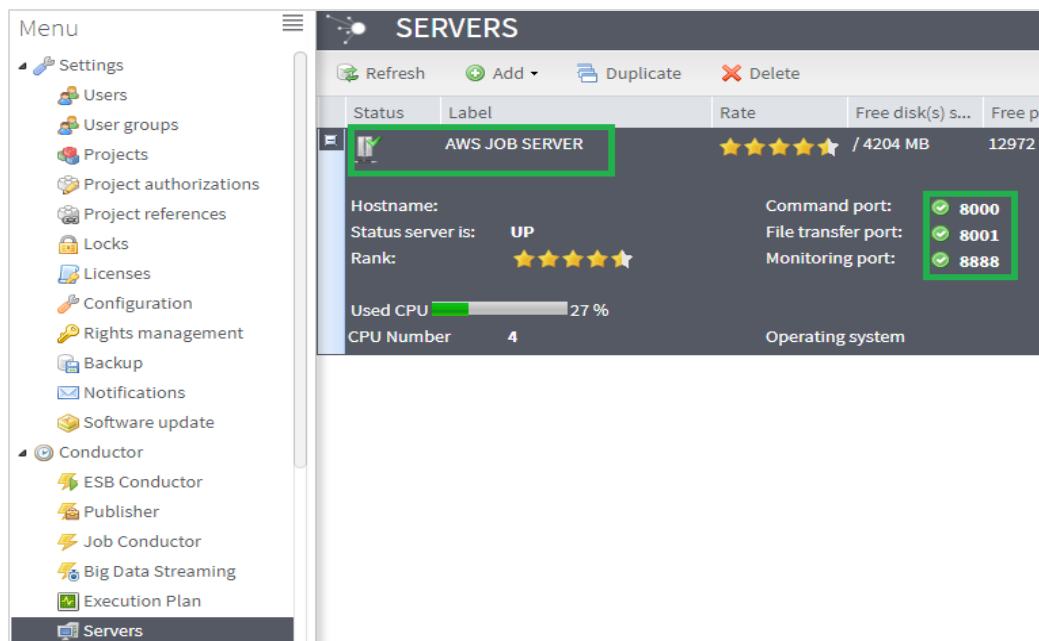
1. Check the status of all stacks associated with the Quick Start in the AWS CloudFormation console to make sure they're **CREATE_COMPLETE** and display no errors.
2. Open TAC:
 - a. Connect to <http://tacdns:8080/tac> and make sure that the TAC login page is displayed as shown in Figure 7.



The image shows the 'Administration Center' login page. It has a green header with the title 'Administration Center'. Below the header is a 'Login' section with fields for 'Login:' (containing 'tadmin@talend.com'), 'Password:' (masked with dots), and a 'Remember me:' checkbox which is checked. A blue 'Login' button is below the fields. At the bottom, there is a link 'Go to db config page' with a small icon.

Figure 7: TAC login page

- b. Log in to TAC using the user name and password you specified during deployment in step 3, and make sure that the TAC home page is displayed.
3. Check job servers: Go to the **Servers** tab in TAC and make sure that server status is in green, as shown in Figure 8.



The image shows the 'SERVERS' tab in the TAC interface. On the left is a 'Menu' sidebar with various options like Settings, Users, Projects, etc. The main area shows a table of servers. The first server is 'AWS JOB SERVER', which is highlighted with a green box. Below the table, the server's details are shown: Hostname, Status (UP), Rank (5 stars), Used CPU (27%), CPU Number (4), Command port (8000), File transfer port (8001), and Monitoring port (8888). The ports are highlighted with green boxes.

Status	Label	Rate	Free disk(s) s...	Free p
UP	AWS JOB SERVER	★★★★★	/ 4204 MB	12972

Hostname:
Status server is: UP
Rank: ★★★★★
Used CPU: 27 %
CPU Number: 4
Command port: 8000
File transfer port: 8001
Monitoring port: 8888
Operating system:

Figure 8: Job server status

4. Check the Nexus server: Go to the **Configuration** tab in TAC and check if the fields of the artifact repository are all marked in green, as shown in Figure 9.

Artifact Repository (7 Parameters)		
Artifact repository type:	NEXUS	✓
Nexus url:	http://localhost:8081/nexus	✓
Nexus username:	admin	✓
Nexus password:	change password	✓
Nexus Default Release Repo:	releases	✓
Nexus Default Snapshot Repo:	snapshots	✓
Nexus Default Group ID:	org.example	✓

Figure 9: Nexus configuration

5. Check the Git server:
 - a. Connect to the Git URL from the **Outputs** tab of the AWS CloudFormation console, using the user name and password you specified during deployment in step 3 to ensure successful login.
 - b. Go to the **Configuration** tab in TAC to make sure that Git fields are marked in green, as shown in Figure 10.

Git (5 Parameters)		
Branches whitelist:	true	✓
Git server url:	https://[redacted]:[redacted]@github.com:443/	✓
Username:	[redacted]	✓
Password:	change password	✓
Commit Log Pattern:	{0}	✓

Figure 10: Git configuration

6. Check the distant run server: SSH into the distant run server and make sure that you can connect successfully.
7. Check Amazon Redshift: Connect to the Amazon Redshift cluster by following the steps in the [Amazon Redshift documentation](#), and check if the connection is successful.

8. Amazon EMR: Connect to the Hue (Hadoop User Experience) interface by using `http://<master-node-hostname>:8888` and check if the Hue login page is displayed.
9. (Optional) Run the Talend sample jobs by following the steps in the [user guide on the Talend website](#) to test end-to-end data integration.

Deleting the Stacks

If you want to decommission the Quick Start modules from your AWS infrastructure, you can delete the stacks that were created through the Quick Start templates. Deleting a stack, either by using the AWS Command Line Interface (AWS CLI) and APIs, or through the AWS CloudFormation console, will remove all the resources created by the template for that stack. The only exception is the S3 bucket for Talend credentials, which you must delete manually. (The S3 source and target buckets are automatically deleted with the stack. Note that you won't get any error messages or notifications about the credentials bucket not being deleted.)

1. Sign in to the AWS Management Console and open the Amazon S3 console at <https://console.aws.amazon.com/s3/>.
2. Select the credentials bucket, and then choose **Delete bucket**.

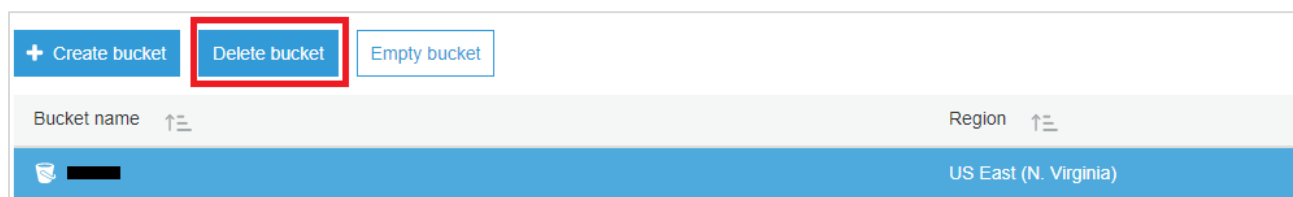


Figure 11: Deleting the Talend credentials bucket

Important This Quick Start deployment uses nested AWS CloudFormation templates, so deleting the main stack will remove the nested stacks and all associated resources.

Troubleshooting

Q. I encountered a `CREATE_FAILED` error when I launched the Quick Start. What should I do?

A. If AWS CloudFormation fails to create the stack, we recommend that you relaunch the template with **Rollback on failure** set to **No**. (This setting is under **Advanced** in the

AWS CloudFormation console, **Options** page.) With this setting, the stack's state will be retained and the instance will be left running, so you can troubleshoot the issue. (You'll want to look at the log files in %ProgramFiles%\Amazon\EC2ConfigService and C:\cfn\log.)

Important When you set **Rollback on failure** to **No**, you'll continue to incur AWS charges for this stack. Please make sure to delete the stack when you've finished troubleshooting.

Q. I encountered a size limitation error when I deployed the AWS CloudFormation templates.

A. We recommend that you launch the Quick Start templates from the location we've provided or from another S3 bucket. If you deploy the templates from a local copy on your computer or from a non-S3 location, you might encounter template size limitations when you create the stack. For more information about AWS CloudFormation limits, see the [AWS documentation](#).

If you have additional questions, please use the [Talend AWS Quick Start community forum](#).

Additional Resources

AWS services

- Amazon EC2
<https://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/>
- AWS CloudFormation
<https://aws.amazon.com/documentation/cloudformation/>
- Amazon VPC
<https://aws.amazon.com/documentation/vpc/>
- Amazon RDS
<https://aws.amazon.com/documentation/rds/>
- Amazon Redshift
<https://aws.amazon.com/documentation/redshift/>
- Amazon EMR
<https://aws.amazon.com/documentation/emr/>

- Amazon S3
<https://aws.amazon.com/documentation/s3/>

Talend

- Talend documentation
<http://doc.talend.com>

Quick Start reference deployments

- [AWS Quick Start home page](https://aws.amazon.com/quickstart/)
<https://aws.amazon.com/quickstart/>

Send Us Feedback

You can visit our [GitHub repository](#) to download the templates and scripts for this Quick Start, to post your comments, and to share your customizations with others.

Document Revisions

Date	Change	In sections
November 2017	Initial publication	—

© 2017, Amazon Web Services, Inc. or its affiliates, and Cognizant Technology Solutions. All rights reserved.

Notices

This document is provided for informational purposes only. It represents AWS's current product offerings and practices as of the date of issue of this document, which are subject to change without notice. Customers are responsible for making their own independent assessment of the information in this document and any use of AWS's products or services, each of which is provided "as is" without warranty of any kind, whether express or implied. This document does not create any warranties, representations, contractual commitments, conditions or assurances from AWS, its affiliates, suppliers or licensors. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

The software included with this paper is licensed under the Apache License, Version 2.0 (the "License"). You may not use this file except in compliance with the License. A copy of the License is located at <http://aws.amazon.com/apache2.0/> or in the "license" file accompanying this file. This code is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.