Out of the Box Data Lake

Quick Start Reference Deployment

October 2017

Cognizant Technology Solutions

AWS Quick Start Reference Team

Contents

This Quick Start deployment guide was created by Amazon Web Services (AWS) in partnership with Cognizant Technology Solutions and Talend.

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 using AWS services together with the Talend integration platform in the AWS Cloud in minutes.

The Quick Start goes beyond basic infrastructure to illustrate Big Data best practices with sample jobs developed by Cognizant for integrating Spark, RedShift, hadoop and S3 technologies into a Data Lake implementation.

Out of the Box Data Lake

Data Lake on the cloud plays the role of a key driver for Digital Transformation initiatives for data and operational agility by enabling access to historical and real-time data for analytics. Cognizant in partnership with AWS and Talend brings together a solution that enables customers to build and deploy a Data Lake on AWS in 60% less time.

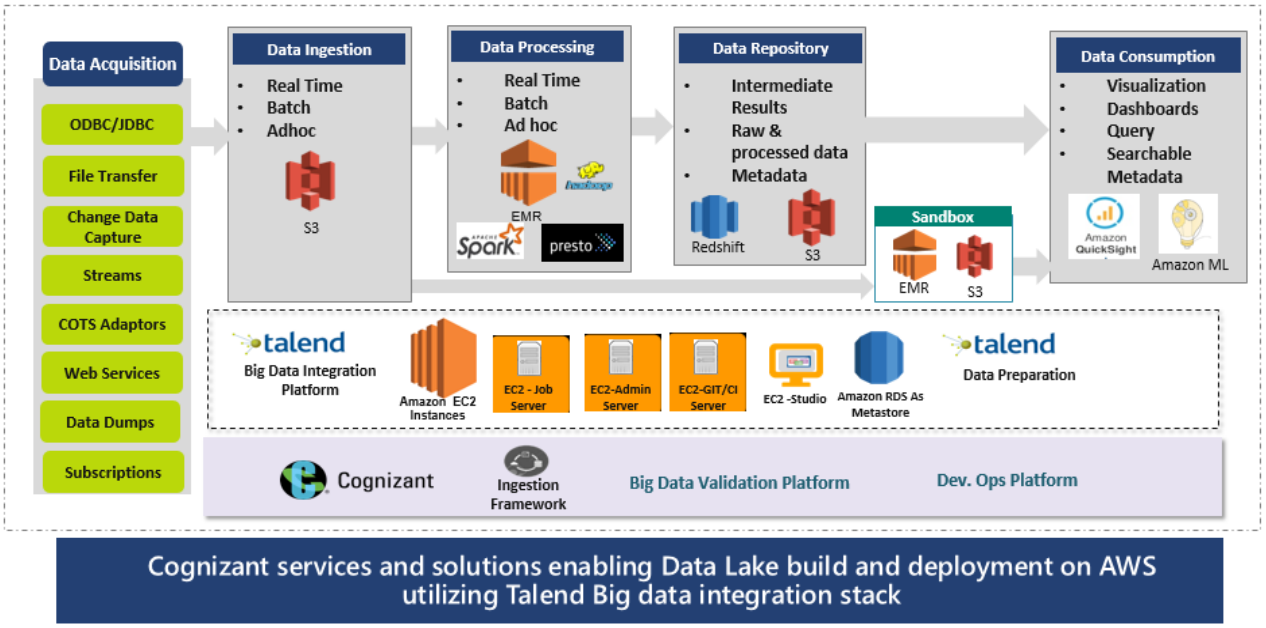
This Quick Start is for anyone evaluating Big Data in the Cloud or looking to accelerate their Big Data initiative through the adoption of best practices for Big Data integration. It illustrates basic Big Data integration patterns with S3, EMR, and Redshift using Talend. The Quick Start uses AWS Cloud Formation scripts to provision Talend and AWS infrastructure using Cognizant best practices for both DevOps and SDLC.

Solution Benefits

- Low cost elastic Infrastructure as a Service from AWS is ideal when dealing with vast amounts of data
- Mature data management practices and SDLC discipline based on Cognizant's Big Data experience
- Talend visual design enables staff to be productive with Hadoop without coding

- Instant provisioning accelerates evaluation and development
- Rapid development accelerates ROI
- Democratizing access to Big Data for all stakeholders broadens adoption across the enterprise, ensuring the success of not just the pilot but your Big Data initiative
- Robust, production ready network environment ensures you can take your pilot right into production

Solution Features



Cognizant services and solutions enabling Data Lake build and deployment on AWS utilizing Talend Big data integration stack

- Enables self-service by provisioning required services and components to build a data lake
- Provides flexibility to spin-up environments for Dev, Test and Prod
- The Quick Start includes a sample dataset and pre-built Talend - Spark jobs that help to explore the architecture and understand the stages of the end to end data lake flow
- Key capabilities include the following using Talend – Spark capabilities
    - Ingestion: Loading S3 data to HDFS / Hive
    - Data Processing: Transformation and aggregation using various Talend's spark and Hadoop features
    - Data Repository: Load and build warehouse using Redshift

- Cognizant offering`s ingestion framework , Big data validation and Dev. Ops platform are optional offerings that is used to ingest, validate and deploy big data solutions but not covered as part of cloud formation template

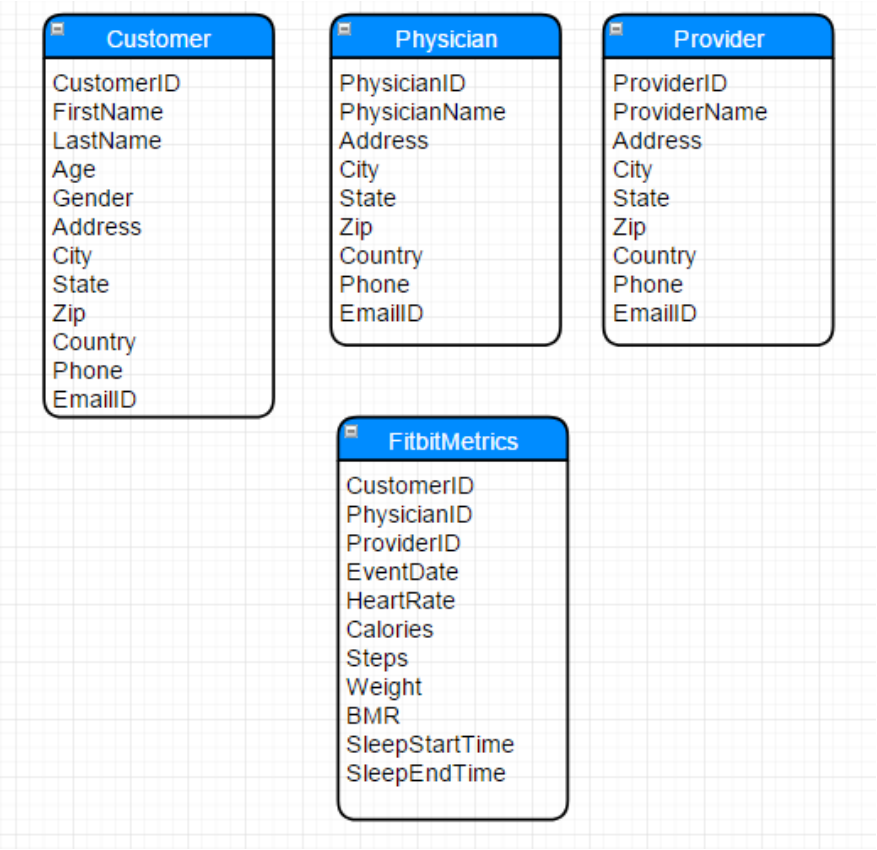
Sample datasets and Talend jobs

The Quick Start dataset includes Fitbit public data to demonstrate how data is submitted to, and ingested by, the Data Lake. This data can then be used to perform predictive and descriptive analytics.

For detailed information about the sample Talend jobs, please refer to the Out of the Box Data Lake User Guide

<<Talend User Guide Hyperlink >>

The sample data included with this Quick Start is organized in the following tables:



The workflow includes three top level jobs

- Sample Talend job that ingests the Fitbit data to S3
- Sample Talend job that uses EMR – Spark framework to perform Aggregation, look up and apply Transformation rules on the Fitbit data and load into S3 harmonized region
- Sample Talend job that loads data from S3 to redshift warehouse

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 license. You should already have received an evaluation license in response to filling the form on the Talend Out of the Box Data Lake Quick Start Landing <<Hyperlink>> page.

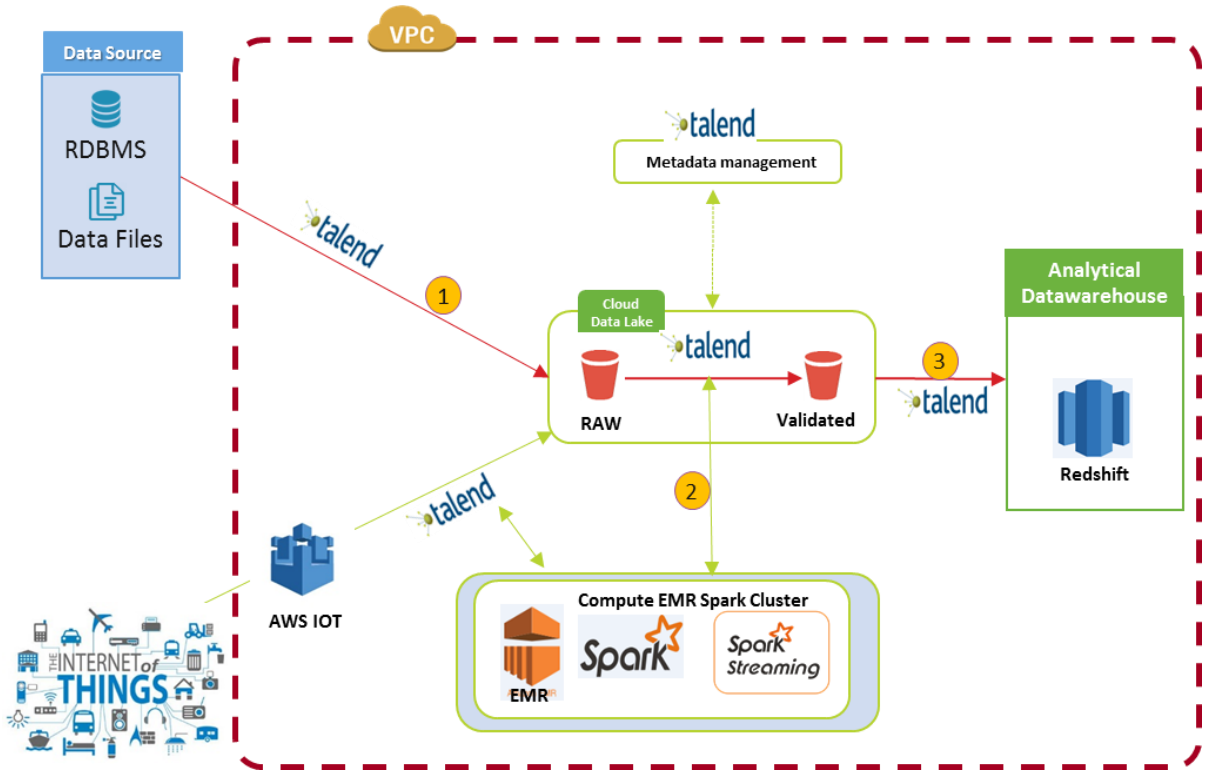
You need to create a private S3 bucket and copy the license file to the root of this bucket. The URL of this bucket is a required parameter to launch the Quick Start.

The code for all Talend jobs included in the Quick Start are released under Apache License.

Architecture

Data Integration Architecture

The following diagram shows the data integration architecture.



The data flow is as follows:

1	Data Ingestion from various types of sources like RDBMS, flat files, semi structured , streaming data to RAW S3 bucket
2	Apply data transformation / Analytics on RAW data using Talend by leveraging EMR spark cluster to apply required transformation.
3	Load the data from load ready files to Analytical Data warehouse (Redshift) using talend

The data flow is designed in Talend Studio and orchestrated by the Talend Data Integration platform.

- Talend Studio allows for the creation of job templates using an easy to understand visual interface. It also provides metadata management capabilities.
- These jobs can then be run by the Talend Data Integration platform to take the data through the flow detailed above.
- Sample jobs are pre built jobs based around typical tasks for Talend that can be used to test the results of the system. The Quick Start features a number of these pre built jobs to demonstrate the flow and use of the system.

Optional AWS Components

These AWS components can be added, but are not configured in the Quick Start templates.

- Integration with Amazon Quicksight for visualization, analysis, and business insight. This can be easily added should the customer desire to use QuickSight for their visualizations.
- IOT data source integration with AWS IoT. The AWS IoT functionality can be easily added, allowing for the analysis and storage of data from IOT devices.

Infrastructure Architecture

The AWS and Talend infrastructure components deployed through this Quick Start are shown in the following diagram.

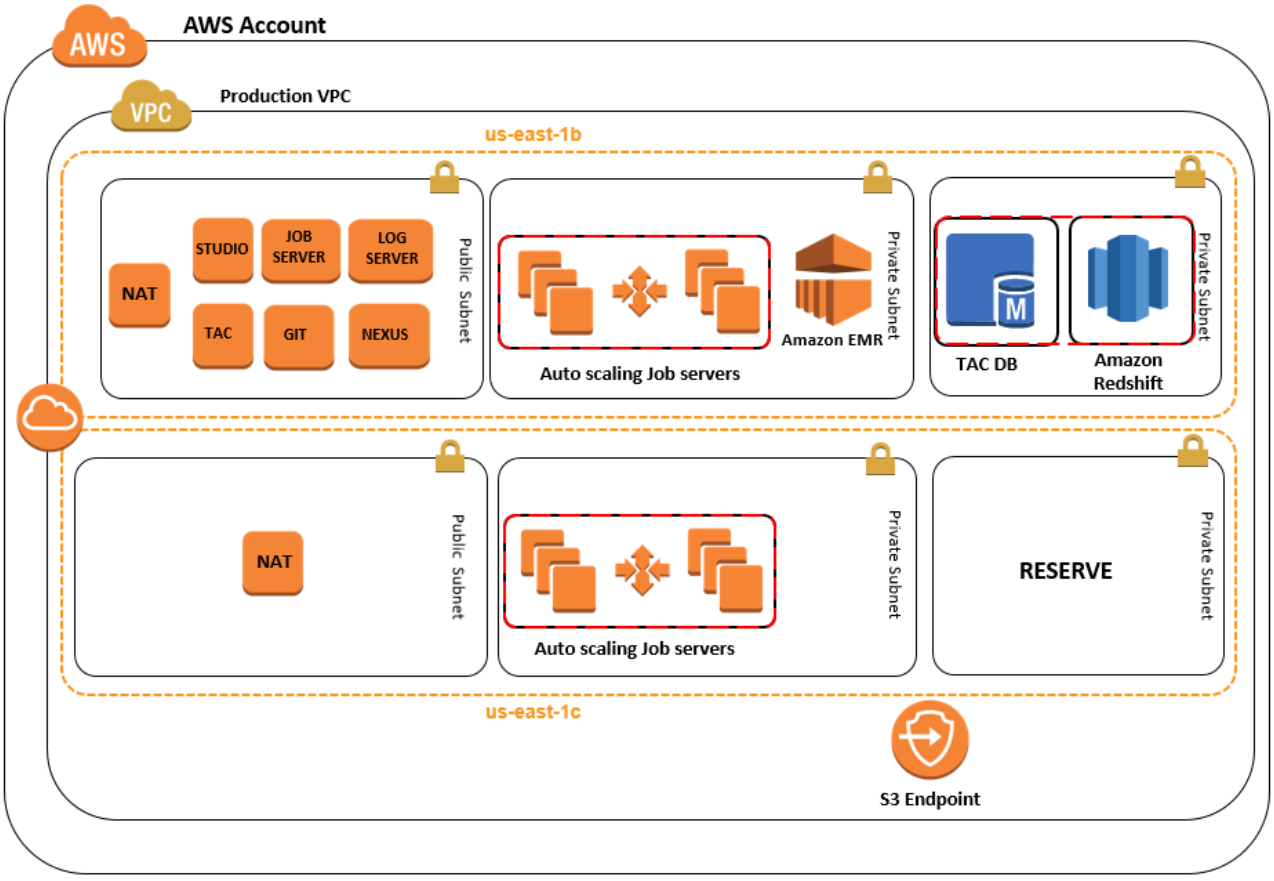


Figure 1: Production Architecture of Out of the Box Data Lake

The Quick Start lets you deploy the stacks on an existing VPC or create a new VPC. It sets up the following:

- VPC spanning two Availability Zones (AZ); in one AZ, two subnets are created – a public subnet to allow connecting over the Internet and a private subnet for Talend jobservers, Redshift, RDS, and EMR. In the second AZ, a private subnet is created, however this is currently not used.

- Network address translation (NAT) gateway as well as a Bastion host.

- 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 X2Go client for users who do not wish to run Talend Studio on their laptops.
- Nexus artifact repository and Git servers for binary and source configuration management.
- Auto-scaling group for Talend JobServer - consisting of Amazon Elastic Cloud Compute (EC2) instances running Talend Jobs scheduled by the TAC. The autoscaling group allows for EC2 instances to be automatically spun up or down to respond to the demand on the Talend JobServers.
- Auto-scaling group for Talend Distant Run JobServer - consisting of Amazon Elastic Cloud Compute (EC2) instances running Talend Jobs on behalf of Talend Studio users. Jobs can be run locally on Studio or on these servers. The autoscaling group allows for EC2 instances to be automatically spun up or down to respond to the demand on the Talend JobServers.
- Relational Database Service (RDS) instance to host Talend metadata in a private subnet.
- Amazon Simple Storage Service (S3) to ingest data and build data lake
- Elastic Map Reduce (EMR) cluster with Pig, Hive and Spark that integrates closely with Talend Big data- provides Hadoop capability to Datalake.
- Redshift cluster for data warehouse and/or data mart
- Talend Logserver using Elastic Search, Logstash, and Kibana.

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 Getting Started with AWS.)

- Amazon VPC – The Amazon Virtual Private Cloud (Amazon VPC) service lets you provision a private, logically isolated section of the AWS cloud where you can launch AWS services and other resources in a virtual network that you define. You have complete control over your virtual networking environment, including selection of your own IP address range, creation of subnets, and configuration of route tables and network gateways.
- Amazon EC2 – The Amazon Elastic Compute Cloud (Amazon EC2) service enables you to launch virtual machine instances with a variety of operating systems. You can choose from existing Amazon Machine Images (AMIs) or import your own virtual machine images.
- Amazon S3 – Amazon S3 provides a repository for your data on the internet. Furthermore, S3 is integrated with many AWS services as well as having a simple to use interface to access this data from anywhere on the internet.
- Amazon RDS – Amazon Relational Database Service (Amazon RDS) enables you to set up, operate, and scale a relational database in the AWS cloud. It also handles many database management tasks, such as database backups, software patching, automatic failure detection, and recovery, for database products such as MySQL, MariaDB, PostgreSQL, Oracle, Microsoft SQL Server, and Amazon Aurora. This Quick Start includes a MySQL database by default.
- Amazon Redshift – Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud that makes it simple and cost-effective to analyze all your data using standard SQL, and your existing Business Intelligence (BI) tools. It allows you to run complex analytic queries against petabytes of structured data, using sophisticated query optimization, columnar storage on high-performance local disks, and massively parallel query execution.
- Amazon EMR – Amazon EMR provides a managed Hadoop framework that makes it easy, fast, and cost-effective to process vast amounts of data across dynamically scalable Amazon EC2 instances. EMR can also run other popular distributed frameworks such as Apache Spark, HBase, Presto, and Flink in Amazon EMR, and interact with data in other AWS data stores.

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

- AWS IAM – AWS Identity and Access Management (IAM) is a web service for controlling secure access to AWS services. Secure interaction between Talend Jobservers, S3, EMR and Redshift is governed through IAM policies.

- Amazon CloudWatch – Amazon CloudWatch provides a reliable, scalable, and flexible monitoring solution that you can start using within minutes.

You should also read the documentation on Talend Big Data Integration.

There are useful videos available as well for getting familiar with how Talend powers Big Data ecosystems.

Technical Requirements

Talend Studio

You will need a Talend Studio server to enables users with local Talend Studio client software to connect to Talend Studio running in the Amazon VPC. This provides users secure access to the Talend components without the need to deploy additional Talend Studio software on AWS.

TAC Database

The Quick Start uses MySQL as the TAC DB. You can either 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 RDS instance for you as well as a TAC database. RDS and the Git server take the longest to provision. So if you are going to do frequent testing, or if you wish to have continuity between your tests then it is recommended to set up a separate instance 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. Just as importantly, if you wish continuity between your different runs you may wish to run a Git server separately or use a Git service such as Github. 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. The complete list is in the Talend Installation Guide. 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 Out of the Box Data Lake 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 Talend into this new VPC.

- **Deploy Out of the Box Data Lake 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.
2. Use the region selector in the navigation bar to choose the AWS Region where you want to deploy.
3. Create a key pair in your preferred region.
4. If necessary, request a service limit increase for Amazon EC2 instances. You might need to do this if you already have an existing deployment that uses this instance type, and you think you might exceed the default limit with this reference deployment.

Step 2. 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.

Option 1	Option 2
Deploy Out of the Box Data Lake **into a new VPC on AWS**	**Deploy** Out of the Box Data Lake **into an existing VPC on AWS**
Launch	Launch

> **Important** If you're deploying **Out of the Box Data Lake** into an existing VPC, make sure that your VPC has two private subnets in different Availability Zones for the database instances. 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.

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 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 into a new VPC

   – Parameters for deploying into an existing VPC

- **Option 1: Parameters for deploying Quickstart into a new VPC**

  View template

Components	Parameters	Description
Bastion server	BastionInstanceType	Amazon EC2 instance type for the bastion instances
	EnableTCPForwarding	Enable/Disable TCP Forwarding
	EnableX11Forwarding	Enable/Disable X11 Forwarding
	NumBastionHosts	Enter the number of bastion hosts to create
EMR	CreateEmr	Create EMR.
GIT	GitAdminEmail	Git admin contact email.
	GitAdminPassword	Git password.
	GitAdminUserid	Git user.
	GitHost	Git host.
	GitPort	Git port.
	GitProtocol	Git protocol.
	GitRepo	Git repository.
	GitTacEmail	TAC contact email.
	GitTacPassword	Git TAC password.
	GitTacUserid	Git TAC userid.
RDS	AmcDbPassword	AMC database password.

Components	Parameters	Description
	AmcDbUser	AMC database user.
	CreateAmcDatabase	Create AMC Database (true) or use an existing AMC database (false)
	CreateTacDatabase	Create a new TAC Database (true) or use an existing TAC database (false)
	DbAllocatedStorage	Allocated Storage (in GB) for RDS instance
	DbClass	Instance class of RDS instance
	MasterDbPassword	Master user database password. Only needed if creating the TAC or AMC databases.
	MasterDbUser	The master or root user used to create TAC and AMC databases and the TAC user. Only needed if creating the TAC or AMC databases.
	TacDbHost	Specify an external mysql database hostname
	TacDbPassword	TAC database password.
	TacDbSchema	TAC database schema.
	TacDbUser	TAC database user.
	TacPassword	TAC application password for tadmin account.
Redshift	RedshiftDbName	RedShift Database name
	RedshiftHost	Specify an external Redshift database hostname
	RedshiftPassword	RedShift Password: Can only contain alphanumeric characters or the following special characters !^*-_+
	RedshiftUsername	RedShift Username
S3	QSS3BucketName	S3 bucket name for the Quick Start assets. Quick Start bucket name can include numbers
	QSS3KeyPrefix	S3 key prefix for the Quick Start assets. Quick Start key prefix can include numbers
	TalendLicenseBucket	Bucket holding Talend license
	TalendResourceBucket	Talend S3 resources bucket.
Talend Jobserver Autoscale	CreateDistantRunStack	Create Jobserver stack in public subnet for use by Studio Distant Run capability.
	CreateJobserverAutoscaleStack	Create JobserverAutoscale stack.
	DistantRunAutoscaleDesiredCapacity	Talend DistantRun autoscale maximum size
	DistantRunAutoscaleMaxSize	Talend DistantRun autoscale maximum size
	JobserverAutoscaleDesiredCapacity	Talend Jobserver autoscale maximum size
	JobserverAutoscaleMaxSize	Talend Jobserver autoscale maximum size
	JobserverInstanceType	Jobserver EC2 instance type
Talend Logserver EC2	LogserverInstanceType	Logserver EC2 instance type
Talend Nexus	NexusAdminPassword	Nexus password.
	NexusAdminUserid	Nexus administrator userid.
	NexusInstanceType	Nexus EC2 instance type

Components	Parameters	Description
Talend Studio EC2	CreateStudioStack	Create Studio stack.
	StudioInstanceType	Studio EC2 instance type
Talend TAC EC2	TacInstanceType	TAC EC2 instance type
VPC	AvailabilityZones	List of Availability Zones to use for the subnets in the VPC. Note: The logical order is preserved and only 2 AZs are used for this deployment.
	KeyPairName	Public/private key pairs allow you to securely connect to your instance after it launches
	PrivateSubnet1CIDR	CIDR block for private subnet 1 located in Availability Zone 1.
	PrivateSubnet2CIDR	CIDR block for private subnet 2 located in Availability Zone 2.
	PublicSubnet1CIDR	CIDR Block for the public DMZ subnet 1 located in Availability Zone 1
	PublicSubnet2CIDR	CIDR Block for the public DMZ subnet 2 located in Availability Zone 2
	RemoteAccessCIDR	Allowed CIDR block for external SSH access to the bastions
	VPCCIDR	CIDR Block for the VPC

5. 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**.
6. 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.
7. Choose **Create** to deploy the stack.
8. Monitor the status of the stack. When the status is **CREATE_COMPLETE**, the cluster is ready.
9. Use the URLs displayed in the **Outputs** tab for the stack to view the resources that were created.

Step 3. Test the Deployment

The Quick Start launches a single VPC with two Availability zones. Within each Availability Zone are two subnets, one public subnet and one private subnet. The entry point for the Quick Start is the oodle-master.template in the templates directory.

As a user of the Quick Start, most of your time will be spent accessing the TAC either directly through the browser or indirectly through the Talend Studio. Talend Studio will likely be running on your desktop. The TAC server but also the Nexus and Logservers are all running in the public subnet. However, access to these servers is till constrained by the Remote Access CIDR. So be sure to set the Remote Access CIDR to your IP address. The

typical Remote Access CIDR is your 4 octet IP address followed by the 32 bit specifier, e.g. 71.120.28.163/32.

In addition, accessing these servers through the browser, you can use SSH with the specified Key Pair Name to access the TAC, Nexus, or Logservers from the shell.

In contrast, the Jobserver runs on the private subnet and is therefore not directly accessible. However, you can SSH to its private IP address from the TAC. If you wish, you can set up SSH tunneling as well.

Deleting the Stacks

If you want to decommission the Quick Start modules from your AWS infrastructure, you can delete the stacks created through the Quick Start templates. Deleting a stack, either via CLI and APIs or through the AWS CloudFormation console, will remove all the resources created by the template for that stack. The exceptions are:

- **EMR security group:** When the stack is deleted, the EMR deletion happens normally but default EMR security groups do not get deleted. No error message/failure occurs as they are not part of the template, however the VPC fails to delete because security groups still exist. The default EMR security groups either fail to delete because they are not part of the cloud formation, or because they are self-referencing. For more information on how to work around this issue, see the Troubleshooting section.

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

Best Practices Using Out of the Box Data Lake

- Multi-AZ architecture intended for high availability
- Isolation of instances between private/public subnets
- Security groups limiting access to only necessary services
- Network access control list (ACL) rules to filter traffic into subnets as an additional layer of network security

- A secured bastion host instance to facilitate restricted login access for system administrator actions
- Standard IAM policies with associated groups and roles, exercising least privilege
- Monitoring and logging; alerts and notifications for critical events
- S3 buckets (with security features enabled) for logging, archive, and application data
- Implementation of proper load balancing and Auto Scaling capabilities
- Amazon RDS database backup and encryption

Security

- This Quick Start follows standardized architecture on the AWS Cloud
- The network layer of this Quick Start uses the AWS VPC Quick Start.

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.

For additional information, see Troubleshooting AWS CloudFormation on the AWS website or contact us on the AWS Quick Start Discussion Forum.

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

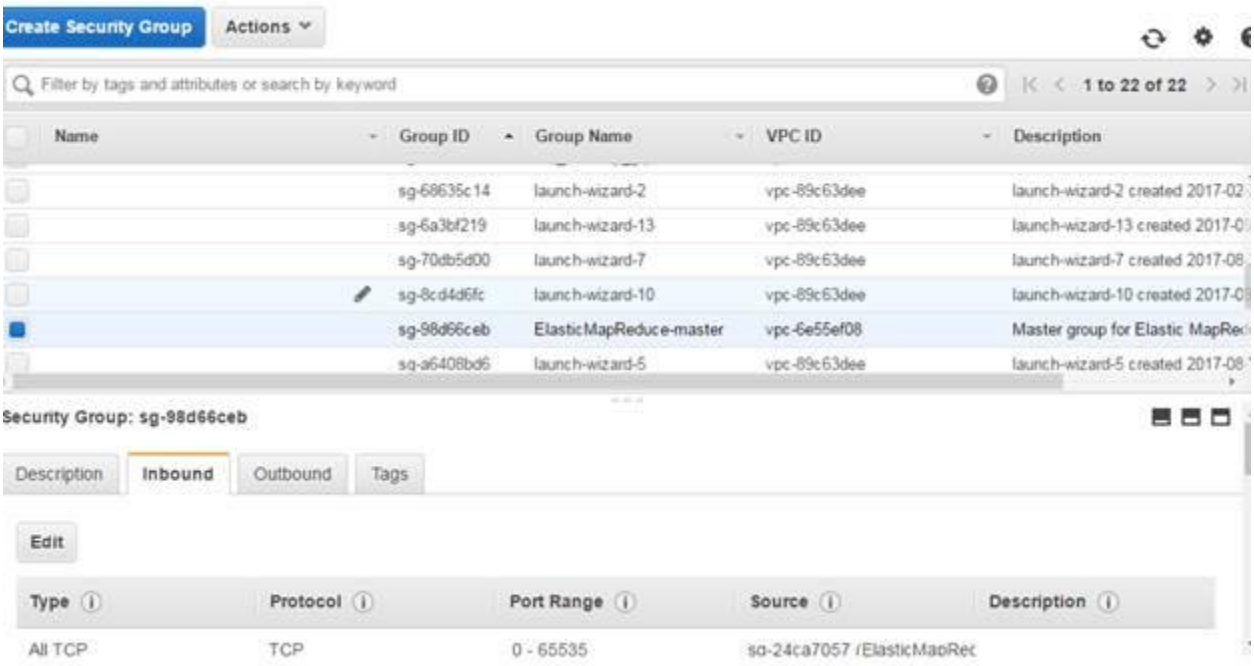
A. We recommend that you launch the Quick Start templates from the location we've provided or from another S3 bucket. If you deploy the templates from a local copy on your computer or from a non-S3 location, you might encounter template size limitations when you create the stack. For more information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

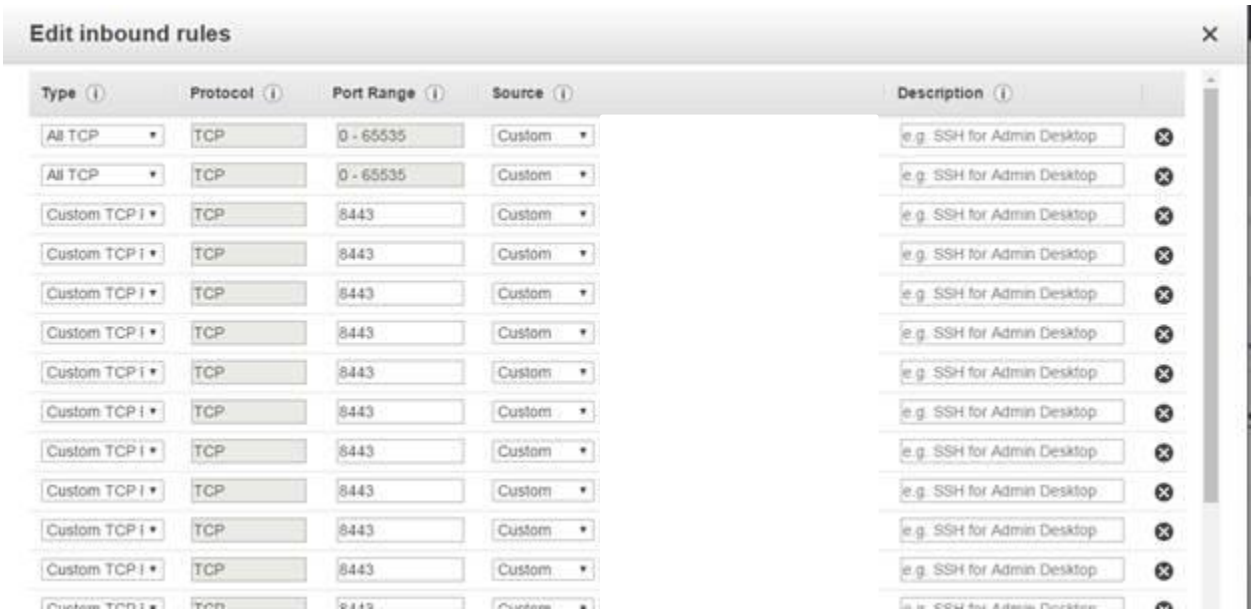
Q. EMR security group does not get deleted when the stack is deleted

A. When the stack is deleted EMR deletes normally but default EMR security groups do not delete. No error message/failure occurs as they are not part of the template but VPC fails to delete because security groups still exist. The default EMR security groups either fail to delete because they are not part of the cloud formation, or because they are self-referencing.

If they are deleted manually in the console, the self-referencing ingresses need to be removed before they can be deleted.

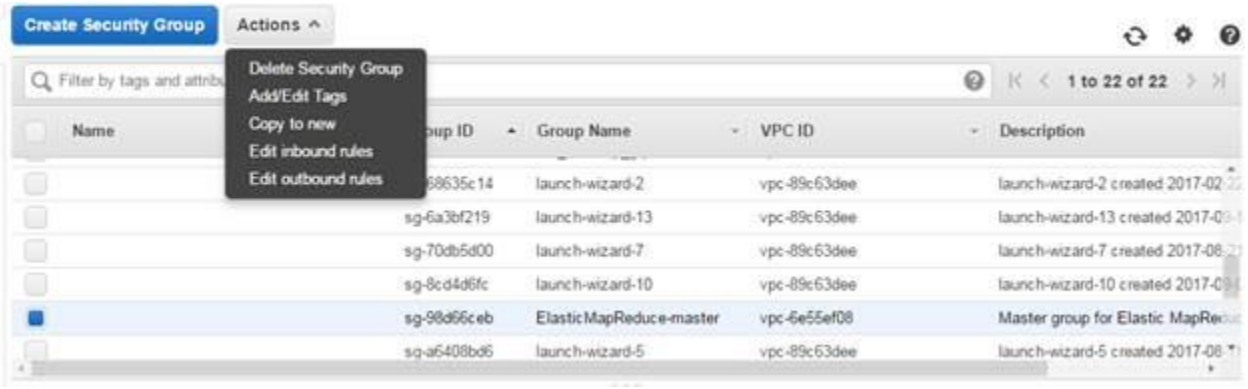
This issue can be resolved by editing the default EMR security group ingresses, by deleting all ingresses and egresses as shown in the picture below





Delete Ingress:

Once all ingresses and egresses have been removed from the service, slave and master security groups, they can be deleted as normal.



Q. EMR cluster creation fails with error "EMR_DefaultRole is invalid."

A. This Quick Start template relies on AWS managed IAM roles for creating the EMR cluster. If the default IAM roles do not exist or are not properly configured, the EMR cluster may fail to initialize. Manually recreate the EMR_DefaultRole IAM role with default configuration and launch the CloudFormation stack again.

For detail information on resolution, see this AWS knowledge center article.

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

Quick Start reference deployments

- AWS Quick Start home page
  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
October 2017	Initial publication	NA

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.