TIBCO Data Science on the AWS Cloud

Quick Start Reference Deployment

December 2018

TIBCO

AWS Quick Start Team

Contents

Quick Links ............................................2
Overview ..............................................2
   TIBCO Data Science on AWS ..................................3
   Costs and Licenses ........................................3
Architecture ..........................................3
Prerequisites .........................................5
   Technical Requirements ..................................5
   Specialized Knowledge ..................................5
Deployment Options ..................................6
Deployment Steps ....................................6
   Step 1. Prepare Your AWS Account ..................6
   Step 2. Subscribe to the TIBCO Data Science AMI ............................7
   Step 3. Launch the Quick Start .........................7
   Step 4. Test the Deployment ............................14
Troubleshooting ......................................21
GitHub Repository ....................................21
Additional Resources ................................22
This Quick Start was created by TIBCO Data Science Group in collaboration with Amazon Web Services (AWS).

Quick Starts are automated reference deployments that use AWS CloudFormation templates to deploy key technologies on AWS, following AWS best practices.

Quick Links
The links in this section are for your convenience. Before you launch the Quick Start, please review the architecture, security, and other considerations discussed in this guide.

- If you have an AWS account, and you’re already familiar with AWS services and TIBCO Data Science, you can launch the Quick Start to build the architecture shown in Figure 1 in a new or existing virtual private cloud (VPC). The deployment takes approximately 80 minutes. If you’re new to AWS or to TIBCO Data Science, please review the implementation details and follow the step-by-step instructions provided later in this guide.

- If you want to take a look under the covers, you can view the AWS CloudFormation templates that automate the deployment.

Overview
This Quick Start reference deployment guide provides step-by-step instructions for deploying TIBCO Data Science (TDS) on the AWS Cloud, using Amazon Elastic File System (Amazon EFS) for shared storage, Amazon Aurora, Application Load Balancers, and an Amazon Elastic Compute Cloud (Amazon EC2) Auto Scaling group.

This deployment guide is for infrastructure architects and TIBCO administrators who want to quickly deploy a TIBCO Data Science system in a fault-tolerant configuration across AWS Availability Zones.
TIBCO Data Science on AWS

TIBCO Data Science is a big data analytics platform for enterprises. The collaborative user interface allows data scientists, data engineers, and business users to work together on data science projects. These cross-functional teams can build machine learning workflows in an intuitive web interface with a minimum of code, while still leveraging the power of big data platforms.

TIBCO Data Science provides an array of tools (from visual workflows to Jupyter Python notebooks) for the data scientist to work with data of any magnitude, and it connects natively to most data sources, including Apache Hadoop, Spark, Hive, and relational databases. The advanced analytic platform provides security and governance. It also enables the analytics team to share and deploy predictive analytics and machine learning insights with the rest of the organization.

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 templates for this Quick Start include 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.

Tip After you deploy the Quick Start, we recommend that you enable the AWS Cost and Usage Report to track costs associated with the Quick Start. This report delivers billing metrics to an S3 bucket in your account. It provides cost estimates based on usage throughout each month, and finalizes the data at the end of the month. For more information about the report, see the AWS documentation.

This Quick Start requires a subscription to the Amazon Machine Image (AMI) for TIBCO Data Science for AWS, which is available from AWS Marketplace. The AMI includes a 7-day free trial of the TIBCO Data Science software. After 7 days, the trial converts to a paid subscription, if not cancelled before then. The TIBCO Data Science software is charged at a flat rate per hour of use, as described in the listing.

Architecture

Deploying this Quick Start for a new virtual private cloud (VPC) with default parameters builds the following TIBCO Data Science environment in the AWS Cloud.
The Quick Start sets up the following:

- A VPC that spans two Availability Zones and includes two public and two private subnets, for security and high availability.*

- An internet gateway to allow access to the internet.*

- In the public subnets, managed NAT gateways to allow outbound internet access for resources in the private subnets.*

- In a public subnet, a Linux bastion host to provide Secure Shell (SSH) access to the TDS instance. The bastion host is in an Auto Scaling group of one, ensuring that there will always be one host available.*

- In a private subnet, a TDS 6.4 instance in an Auto Scaling group of one, ensuring that there will always be one host available.
• Amazon EFS automatically mounted on the TDS instance to ensure high availability. If the TDS instance fails in one Availability Zone, a new server is created in the second Availability Zone and automatically connected to the existing data. Failover typically takes 3-5 minutes, but can be longer.

• Amazon Aurora (Postgres-compatible 9.6.8) automatically connected to be used as a TDS instance internal database. If the TDS instance fails in one Availability Zone, a new server is created in the second Availability Zone and automatically connected to the existing database. Failover typically takes 3-5 minutes, but can be longer.

• An Application Load Balancer to automatically distribute connections to the active TDS instance.

• An AWS Identity and Access Management (IAM) instance role with fine-grained permissions for access to AWS services necessary for the deployment process.

• Appropriate security groups for each instance or function to restrict access to only necessary protocols and ports. For example, access to HTTP(S) server ports on Amazon EC2 servers is limited to the Application Load Balancer.

• (Optional) Amazon Route 53 as your public Domain Name System (DNS) for resolving your TIBCO Data Science site’s domain name. When you choose to deploy the application with a custom domain and Secure Sockets Layer (SSL) certificate, a new record set in your pre-existing Route 53 hosted zone will be created.

* The template that deploys the Quick Start into an existing VPC skips the tasks marked by asterisks and prompts you for your existing VPC configuration.

**Prerequisites**

**Technical Requirements**

This Quick Start supports TIBCO Data Science version 6.4 on CentOS 7.x.

You need to obtain the TIBCO Data Science for AWS AMI from AWS Marketplace, which includes a free trial period, as discussed in step 2 of the deployment section.

**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 EC2
• Amazon Elastic Block Store (Amazon EBS)
• Amazon Virtual Private Cloud (Amazon VPC)
• AWS CloudFormation
• Amazon Route 53
• Elastic Load Balancing
• Amazon EC2 Auto Scaling
• Amazon EFS
• AWS IAM
• Amazon Aurora

We also recommend that you become familiar with TIBCO Data Science.

Deployment Options

This Quick Start provides two deployment options:

• **Deploy TIBCO Data Science 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 TIBCO Data Science into this new VPC.

• **Deploy TIBCO Data Science into an existing VPC**. This option provisions TIBCO Data Science in your existing AWS infrastructure.

The Quick Start provides separate templates for these options. It also lets you configure CIDR blocks, instance types, and TIBCO Data Science settings, as discussed later in this guide.

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](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 TIBCO Data Science on AWS.

**Important** This Quick Start uses Amazon EFS, which is supported only in the regions listed on the AWS Regions and Endpoints [webpage](https://aws.amazon.com).
3. Create a **key pair** in your preferred region.

4. If necessary, request a **service limit increase** for the Amazon EC2 m5.2xlarge instance type. 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 deployment.

**Step 2. Subscribe to the TIBCO Data Science AMI**

This Quick Start uses AWS Marketplace software from TIBCO and requires that you accept the terms within the AWS account where the Quick Start will be deployed.

2. Open the page for the [TIBCO Data Science AMI](https://aws.amazon.com/marketplace). 
3. Choose **Continue to Subscribe**.

![AMI page for TIBCP Data Science for AWS](image)

**Figure 2. The AMI page for TIBCP Data Science for AWS**

4. Accept the terms of the license agreement, and exit out of AWS Marketplace without further action. **Do not** provision the software from AWS Marketplace—the Quick Start will deploy the AMI for you.

![Accepting software terms in AWS Marketplace](image)

**Figure 3: Accepting software terms in AWS Marketplace**

**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.

   ![Option 1](Launch.png)  ![Option 2](Launch.png)

   **Option 1**
   **Deploy TIBCO Data Science into a new VPC on AWS**

   **Option 2**
   **Deploy TIBCO Data Science into an existing VPC on AWS**

   **Important** If you’re deploying TIBCO Data Science 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](https://aws.amazon.com) in their route tables, to allow the instances to download packages and software without exposing them to the internet. You will also need the domain name option configured in the DHCP options as explained in the [Amazon VPC documentation](https://aws.amazon.com). You will be prompted for your VPC settings when you launch the Quick Start.

   Each deployment takes about 80 minutes 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 TIBCO Data Science will be built. The template is launched in the US East (Ohio) Region by default.

   **Important** This Quick Start uses Amazon EFS, which is supported only in the regions listed on the AWS Regions and Endpoints [webpage](https://aws.amazon.com).

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 separately for the two deployment options:
- Parameters for deploying TIBCO Data Science into a new VPC
- Parameters for deploying TIBCO Data Science into an existing VPC

**Option 1: Parameters for deploying TIBCO Data Science into a new VPC**

**View template**

**VPC & Bastion configuration:**

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability Zones</td>
<td>Requires input</td>
<td>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.</td>
</tr>
<tr>
<td>Allowed external access CIDR</td>
<td>Requires input</td>
<td>The CIDR IP range that is permitted to access the TIBCO Data Science web console via the Application Load Balancer. We recommend that you set this value to a trusted IP range.</td>
</tr>
<tr>
<td>Bastion AMI Operating System</td>
<td>Amazon Linux HVM</td>
<td>The Linux distribution for the AMI to be used for the bastion instances.</td>
</tr>
<tr>
<td>Bastion instance type</td>
<td>t2.micro</td>
<td>The Amazon EC2 instance type for the bastion instances.</td>
</tr>
</tbody>
</table>

**TIBCO Data Science configuration:**

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key pair name</td>
<td>Requires input</td>
<td>The name of an existing public/private key pair, which allows you to securely connect to your instance after it launches.</td>
</tr>
<tr>
<td>TDS instance owner</td>
<td>tibco-ds-owner</td>
<td>The Owner tag for the TIBCO Data Science instance.</td>
</tr>
<tr>
<td>TDS instance name</td>
<td>TIBCO-DS</td>
<td>The name for the TIBCO Data Science instance that is deployed to EC2.</td>
</tr>
<tr>
<td>TDS instance type</td>
<td>m5.2xlarge</td>
<td>The type of EC2 instance for the TIBCO Data Science 6.4 AMI.</td>
</tr>
<tr>
<td>TDS siteadmin password</td>
<td>Requires input</td>
<td>The password for the siteadmin user to access the TIBCO Data Science web console. Must be 8 characters or more containing at least one uppercase letter, one lowercase letter and one number. It can use any printable ASCII characters (ASCII code 33 to 126) except ‘ (single quote), &quot; (double quote), , , /, @, or space.</td>
</tr>
</tbody>
</table>
Amazon RDS configuration:

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDS database password (RDSPassword)</td>
<td>Requires input</td>
<td>Database password for Amazon RDS Aurora database. Must be 8 characters or more containing at least one uppercase letter, one lowercase letter and one number. It can use any printable ASCII characters (ASCII code 33 to 126) except ‘ (single quote), ” (double quote), , /, @, or space.</td>
</tr>
<tr>
<td>Multi-AZ database (DBMultiAZ)</td>
<td>true</td>
<td>Specifies that the database instance is a multiple Availability Zone deployment.</td>
</tr>
</tbody>
</table>

Amazon EMR configuration:

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMR master instance type (EMRMasterInstanceType)</td>
<td>m4.xlarge</td>
<td>The instance type for the EMR master node.</td>
</tr>
<tr>
<td>EMR core instance type (EMRCoreInstanceType)</td>
<td>m4.xlarge</td>
<td>The instance type for the EMR core nodes.</td>
</tr>
<tr>
<td>EMR cluster name (EMRClusterName)</td>
<td>Requires input</td>
<td>The name for your EMR cluster.</td>
</tr>
<tr>
<td>EMR core nodes (EMRCoreNodes)</td>
<td>1</td>
<td>The number of core nodes. The minimum is 1; the maximum is 500.</td>
</tr>
<tr>
<td>EMR logs bucket name (EMRLogBucket)</td>
<td>—</td>
<td>(Optional) The name of a pre-existing S3 bucket for the EMR logs.</td>
</tr>
</tbody>
</table>

(Optional) TIBCO Site Domain configuration:

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDS Site domain (TSDSSiteDomain)</td>
<td>—</td>
<td>(Optional) The domain name of the TIBCO Data Science site; e.g. example.com. A valid Fully Qualified Domain Name is required when using SSL.</td>
</tr>
<tr>
<td>ALB SSL certificate ARN (ALBSSLCertificateARN)</td>
<td>—</td>
<td>(Optional) The Amazon Resource Name (ARN) of the SSL certificate to be used for the Application Load Balancer.</td>
</tr>
<tr>
<td>Route 53 hosted zone ID (Route53HostedZoneId)</td>
<td>—</td>
<td>(Optional) The Route53 hosted zone ID where the DNS record for TIBCO Data Science site domain will be added.</td>
</tr>
</tbody>
</table>
AWS Quick Start configuration:

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quick Start S3 bucket name (QSS3BucketName)</td>
<td>aws-quickstart</td>
<td>The S3 bucket you have 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.</td>
</tr>
<tr>
<td>Quick Start S3 key prefix (QSS3KeyPrefix)</td>
<td>quickstart-tibco-data-science/</td>
<td>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.</td>
</tr>
</tbody>
</table>

- **Option 2: Parameters for deploying TIBCO Data Science into an existing VPC**

  View template

Network configuration:

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPC ID (VPCID)</td>
<td>Requires input</td>
<td>The ID of your existing VPC (e.g., vpc-0343606e).</td>
</tr>
<tr>
<td>Private subnet 1 ID (PrivateSubnet1ID)</td>
<td>Requires input</td>
<td>The ID of the private subnet in Availability Zone 1 for the TSDS instances (e.g., subnet-a0246dcd).</td>
</tr>
<tr>
<td>Private subnet 2 ID (PrivateSubnet2ID)</td>
<td>Requires input</td>
<td>The ID of the private subnet in Availability Zone 2 for the TSDS instances (e.g., subnet-b1f432cd).</td>
</tr>
<tr>
<td>Public subnet 1 ID (PublicSubnet1ID)</td>
<td>Requires input</td>
<td>The ID of the public subnet in Availability Zone 1 for the Application Load Balancer (e.g., subnet-9bc642ac).</td>
</tr>
<tr>
<td>Public subnet 2 ID (PublicSubnet2ID)</td>
<td>Requires input</td>
<td>The ID of the private subnet in Availability Zone 2 for the Application Load Balancer. (e.g., subnet-e3246d8e).</td>
</tr>
<tr>
<td>Allowed external access CIDR (AdminConsoleAccessCIDR)</td>
<td>Requires input</td>
<td>The CIDR IP range that is permitted to access the TIBCO Data Science web console via the Application Load Balancer. We recommend that you set this value to a trusted IP range.</td>
</tr>
</tbody>
</table>

Bastion configuration:

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bastion Security Group ID (BastionSecurityGroupID)</td>
<td>Requires input</td>
<td>The ID of the bastion host security group to enable SSH connections (e.g., sg-7f16e910).</td>
</tr>
</tbody>
</table>
**TIBCO Data Science configuration:**

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key pair name (KeyPairName)</td>
<td>Requires input</td>
<td>The name of an existing public/private key pair, which allows you to securely connect to your instance after it launches.</td>
</tr>
<tr>
<td>TDS instance owner (TSDSOwner)</td>
<td>tibco-ds-owner</td>
<td>The Owner tag for the TIBCO Data Science instance.</td>
</tr>
<tr>
<td>TDS instance name (TSDSInstanceName)</td>
<td>TIBCO-DS</td>
<td>The name for the TIBCO Data Science instance that is deployed to EC2.</td>
</tr>
<tr>
<td>TDS instance type (TSDSInstanceType)</td>
<td>m5.2xlarge</td>
<td>The type of EC2 instance for the TIBCO Data Science 6.4 AMI.</td>
</tr>
<tr>
<td>TDS siteadmin password</td>
<td>Requires input</td>
<td>The password for the siteadmin user to access the TIBCO Data Science web console. Must be 8 characters or more containing at least one uppercase letter, one lowercase letter and one number. It can use any printable ASCII characters (ASCII code 33 to 126) except ’ (single quote), &quot; (double quote), , /, @, or space.</td>
</tr>
</tbody>
</table>

**Amazon RDS configuration:**

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDS database password</td>
<td>Requires input</td>
<td>Database password for Amazon RDS Aurora database. Must be 8 characters or more containing at least one uppercase letter, one lowercase letter and one number. It can use any printable ASCII characters (ASCII code 33 to 126) except ’ (single quote), &quot; (double quote), , /, @, or space.</td>
</tr>
<tr>
<td>Multi-AZ database</td>
<td>true</td>
<td>Specifies that the database instance is a multiple Availability Zone deployment.</td>
</tr>
</tbody>
</table>

**Amazon EMR configuration:**

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMR master instance type</td>
<td>m4.xlarge</td>
<td>The instance type for the EMR master node.</td>
</tr>
<tr>
<td>EMR core instance type</td>
<td>m4.xlarge</td>
<td>The instance type for the EMR core nodes.</td>
</tr>
<tr>
<td>EMR cluster name (EMRClusterName)</td>
<td>Requires input</td>
<td>The name for your EMR cluster.</td>
</tr>
<tr>
<td>EMR core nodes (EMRCoreNodes)</td>
<td>1</td>
<td>The number of core nodes. The minimum is 1; the maximum is 500.</td>
</tr>
</tbody>
</table>
### Parameter label (name) | Default | Description
--- | --- | ---
EMR logs bucket name (EMRLogBucket) | — | (Optional) The name of the pre-existing S3 bucket for the EMR logs.

### TIBCO Site Domain configuration (Optional):

| Parameter label (name) | Default | Description |
--- | --- | ---
TDS Site domain (TSDSSiteDomain) | — | (Optional) The domain name of the TIBCO Data Science site; e.g., example.com. A valid Fully Qualified Domain Name is required when using SSL.
ALB SSL certificate ARN (ALBSSLCertificateARN) | — | (Optional) The ARN of the SSL certificate to be used for the Application Load Balancer.
Route 53 hosted zone ID (Route53HostedZoneId) | — | (Optional) The Route53 hosted zone ID where the DNS record for the TIBCO Data Science site domain will be added.

### AWS Quick Start configuration:

| Parameter label (name) | Default | Description |
--- | --- | ---
Quick Start S3 Bucket Name (QSS3BucketName) | aws-quickstart | The S3 bucket you have 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) | quickstart-tibco-data-science/ | 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.

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 TIBCO Data Science cluster is ready.
9. Use the URLs displayed in the **Outputs** tab for the stack to view the resources that were created.

**Install TensorFlow libraries for Jupyter Python notebooks**

To use the TensorFlow libraries, run this script located in the following directory after the Image has launched.

1. Connect using SSH to the TIBCO DS instance.
2. Become root user.
3. Download the TensorFlow Install script by using **wget**:

```
```

4. Run the `tf_installer.sh` script as root:

```
./tf_installer.sh
```

5. Restart the Python Notebooks: From inside `/data/notebooks`, run the following:

```
- docker-compose stop
- docker-compose up -d
```

**Step 4. Testing and Post-Deployment Steps**

When the AWS CloudFormation template successfully creates the stack, the EC2 instances will be running in your AWS account, and the TIBCO Data Science software will be configured during launch. Additionally, the TIBCO Data Science application will be running in a fault-tolerant configuration on the server instances.

To verify that TIBCO Data Science is running and accessible, follow these steps:

1. After the TIBCO Data Science AWS CloudFormation template has completed, you can connect to TIBCO Data Science by using the **TDSWebConsoleURL** URL that is provided in the **Outputs** tab.
2. In a browser, go to the TDSWebConsoleURL. Log in using siteadmin as the user name and the password that you provided for the TSDSInstancePassword input parameter.
3. You will see the default view.

![TIBCO Data Science console home](image)

**Figure 6: TIBCO Data Science console home**

4. Configure the EMR Data Source:
a. Choose the menu at top left.

![Home](image)

**Figure 7: Accessing the menu by choosing icon at top left**
b. Choose **Data**.

![Image of TIBCO Spotfire Data Science interface with Data menu highlighted]

**Figure 8:** Choose Data option to configure EMR Data Source

C. Choose **Add Data Source**, choose Hadoop Cluster as the **Data Source Type** and then fill the values as shown in Figure 9 and Table 1.
Figure 9: Adding EMR Data Source
Table 1: EMR Data Source configuration – required properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Source Name</td>
<td>Provide a name for the data source</td>
</tr>
<tr>
<td>Hadoop Version</td>
<td>Amazon EMR5</td>
</tr>
<tr>
<td>Name Node Host</td>
<td>Output of EMRPublicDNS as shown in Figure 4</td>
</tr>
<tr>
<td>Name Node Host Port</td>
<td>8020</td>
</tr>
<tr>
<td>Resource Manager Host</td>
<td>Output of EMRPublicDNS as shown in Figure 4</td>
</tr>
<tr>
<td>Resource Manager Host Port</td>
<td>8032</td>
</tr>
<tr>
<td>Data Source User</td>
<td>hdfs</td>
</tr>
<tr>
<td>Group</td>
<td>supergroup</td>
</tr>
</tbody>
</table>

d. Choose Configure Connection Parameters, and then choose Load Configuration from Resource Manager. Enter the EMRPublicDNS value from the Outputs tab and choose Fetch. Choose Save to save the connection parameters.

![Configure Connection Parameters](image)

Figure 10: Fetch and save connection parameters

e. You can also choose Test Configuration, although we expect the following will not pass:

   - DNS Resolve Name Node
   - DNS Resolve Resource Manager

f. Choose Add Data Source.
5. To connect to another data source that is available in your network, see the TIBCO documentation.

Troubleshooting

Q. I encountered a CREATE_FAILED error when I launched the Quick Start.
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. (For Windows, look at the log files in %ProgramFiles%\Amazon\EC2ConfigService and C:\cfn\log.)

Important When you set Rollback on failure to No, you will continue to incur AWS charges for this stack. Please make sure to delete the stack when you finish troubleshooting.

For additional information, see Troubleshooting AWS CloudFormation on the AWS website.

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 links in this guide 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. None of the TIBCO Data Science services are up and the URL is unreachable.
A. Log in to the TDS instance via your bastion hosts. You can use the sudo su - command to switch to the tsds user. Then follow these instructions to restart the TDS instance.

Q. Python doesn’t start.
A. Log into the TDS instance via your bastion hosts. You can use the sudo su - command to switch to the tsds user. Then follow these instructions to restart Python notebooks.

GitHub Repository

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.

**Additional Resources**

**AWS services**
- Amazon EBS
- Amazon EC2
  [https://aws.amazon.com/documentation/ec2/](https://aws.amazon.com/documentation/ec2/)
- Amazon VPC
  [https://aws.amazon.com/documentation/vpc/](https://aws.amazon.com/documentation/vpc/)
- AWS CloudFormation
  [https://aws.amazon.com/documentation/cloudformation/](https://aws.amazon.com/documentation/cloudformation/)
- Amazon Route 53
- Elastic Load Balancing
  [https://docs.aws.amazon.com/elasticloadbalancing/latest/application/introduction.html](https://docs.aws.amazon.com/elasticloadbalancing/latest/application/introduction.html)
- Amazon EC2 Auto Scaling
- Amazon EFS
  [https://docs.aws.amazon.com/efs/latest/ug/whatisefs.html](https://docs.aws.amazon.com/efs/latest/ug/whatisefs.html)
- AWS Identity and Access Management (IAM)
- Amazon Aurora
  [https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/CHAP_AuroraOverview.html](https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/CHAP_AuroraOverview.html)

**TIBCO Data Science documentation**
- [https://alpine.atlassian.net/wiki/spaces/V6/overview](https://alpine.atlassian.net/wiki/spaces/V6/overview)

**Quick Start reference deployments**
- AWS Quick Start home page
  [https://aws.amazon.com/quickstart/](https://aws.amazon.com/quickstart/)
© 2018, Amazon Web Services, Inc. or its affiliates, and TIBCO. 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.