Hybrid Data Lake on the AWS Cloud
with WANdisco Fusion, Amazon S3, and Amazon Athena

Quick Start Reference Deployment

September 2017

Sturdy
AWS Quick Start Reference Team

Contents
Overview..................................................................................................................2
WANdisco Fusion on AWS......................................................................................3
Costs and Licenses.................................................................................................3
Architecture.........................................................................................................3
Design Considerations .......................................................................................5
Prerequisites .......................................................................................................5
Specialized Knowledge .......................................................................................5
Deployment Options ...........................................................................................5
Deployment Steps ...............................................................................................6
Step 1. Prepare Your AWS Account ......................................................................6
Step 2. Subscribe to the WANdisco Fusion AMI ..................................................6
Step 3. Launch the Quick Start ............................................................................8
Step 4. (Optional) Synchronize the Local HDFS with the S3 Bucket .................16
FAQ.....................................................................................................................20
Additional Resources .................................................................................................................. 21
Send Us Feedback ..................................................................................................................... 21
Document Revisions .................................................................................................................. 22

This Quick Start deployment guide was created by Amazon Web Services (AWS) in partnership with Sturdy, an AWS Advanced Consulting Partner that specializes in DevOps, healthcare, and Internet of Things (IoT), and WANdisco.

**Quick Starts** are automated reference deployments for key technologies on the AWS Cloud, based on AWS best practices for security and high availability.

**Overview**

This Quick Start reference deployment guide provides step-by-step instructions for integrating on-premises Hadoop clusters with a data lake on the AWS Cloud using WANdisco Fusion, Amazon Simple Storage Service (Amazon S3), and Amazon Athena. This hybrid data lake architecture combines on-premises components and AWS Cloud components, and supports burst-out processing in the cloud and cloud migration.

The Quick Start is for users who would like to integrate their on-premises Hadoop clusters with a data lake environment on AWS that includes WANdisco Fusion and Amazon S3. The AWS data lake deployment incorporates Amazon Elastic Compute Cloud (Amazon EC2) compute capacity and Amazon S3 storage, deployed within an Auto Scaling group and virtual private cloud (VPC).

The Quick Start provides the option to deploy a Docker container, which represents your on-premises Hadoop cluster for demonstration purposes, and helps you gain hands-on experience with the hybrid data lake architecture. You can also decide to use your own on-premises Hadoop cluster. WANdisco Fusion replicates data from Docker (or your on-premises Hadoop environment) to Amazon S3 continuously, ensuring strong consistency between data residing on premises and data in the cloud. You can use Amazon Athena to analyze and view the data that has been replicated. This configuration provides the capacity required for burst-out processing scenarios.

You can customize this Quick Start to enable a disaster recovery scenario for your on-premises Hadoop cluster. To do this, after you deploy the Quick Start, provision an Amazon EMR cluster that references the data that is replicated into Amazon S3. (The Quick Start doesn’t deploy Amazon EMR.) This could serve as a low-cost disaster recovery environment in the event of a failure of your on-premises Hadoop cluster.
WANdisco Fusion on AWS
WANdisco Fusion is a software application that allows Apache Hadoop deployments to replicate Hadoop Distributed File System (HDFS) data between Hadoop clusters that are running different, even incompatible, versions of Hadoop. It is also possible to replicate data between different vendor distributions and versions of Hadoop, or between Hadoop and Amazon S3, as configured in this Quick Start.

WANdisco Fusion provides:

- A virtual file system for Hadoop, compatible with all Hadoop applications
- A single, virtual namespace that integrates storage from different types of Hadoop deployments, including CDH, Hortonworks Data Platform (HDP), EMC Isilon, Amazon S3, Amazon EMR File System (EMRFS), and MapR
- Storage that can be globally distributed
- WAN replication using WANdisco’s Fusion Active Data Replication technology, which delivers single-copy consistent HDFS data, replicated between geographically disperse data centers

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.

The Quick Start requires a subscription to the Amazon Machine Image (AMI) for WANdisco Fusion in the AWS Marketplace, as discussed in the deployment steps. The WANdisco Fusion software is provided with the Bring Your Own License model. If no license is provided, the Quick Start will configure the application with a trial key. To continue using WANdisco Fusion beyond the 14-day trial period, you must purchase a license by contacting WANdisco at http://www.wandisco.com/contact.
Architecture

Deploying this Quick Start for a new VPC with default parameters builds the following hybrid data lake environment in the AWS Cloud.

![Figure 1: Quick Start architecture for a hybrid data lake on AWS](image)

The Quick Start sets up the following:

- A VPC configured with public subnets that span multiple Availability Zones for highly availability.*
- An Internet gateway to provide access to the Internet.*
- An IAM role to control access to resources created by the Quick Start. This role is used to control Athena access to Amazon S3 for data analysis, and WANdisco Fusion access to Amazon S3 for data synchronization.
- In the public subnets, WANdisco Fusion server instances in an Auto Scaling group, functioning as a single clustered service. This Quick Start uses Auto Scaling to establish the initial configuration and connectivity between instances in different Availability Zones. Once you replicate data using WANdisco Fusion, moving the Fusion server to a new VM will require manual reconfiguration of some network settings.
• (Optional) An on-premises WANdisco server deployed in a Docker container, to demonstrate the synchronization from HDFS to the S3 bucket in the cloud. The Quick Start uses a sample open dataset consisting of publicly available NYC taxi data.

• (Optional) Amazon Athena to query and analyze the data from the local WANdisco Fusion server, which is synchronized with Amazon S3.

• (Optional) An S3 bucket to store the content that is being synchronized by WANdisco Fusion and the analysis information processed by Athena.

**Design Considerations**
This Quick Start deploys WANdisco Fusion into public subnets to make it easier for the Docker container to communicate with a peer Fusion server in the public network (such as the server in the optional Docker container, as described in the previous section).

AWS recommends deploying workloads into private subnets for security purposes. This choice requires a good understanding of Amazon Virtual Private Cloud (Amazon VPC) and how each service communicates to replicate the data. If you choose to deploy WANdisco Fusion in private subnets, deploy the Quick Start into an existing VPC (see the **Deployment Options** section), and specify an existing private network for the WANdisco Fusion cluster.

**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
- Amazon EC2
- Amazon S3
- Amazon Athena

**Deployment Options**
This Quick Start provides two deployment options:

- **Deploy the hybrid data lake into a new VPC** (end-to-end deployment). This option builds a new AWS environment consisting of the VPC, subnets, security groups, and other infrastructure components, and then deploys the hybrid data lake components into this new VPC.
• **Deploy the hybrid data lake into an existing VPC.** This option provisions the software in your existing AWS infrastructure.

The Quick Start provides separate templates for these options. It also lets you configure CIDR blocks, instance types, WANdisco Fusion settings, and Athena 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 the hybrid data lake on AWS.

3. Create a [key pair](https://aws.amazon.com) in your preferred region.

4. If necessary, request a service limit increase for the `m3.2xlarge` 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 reference deployment.

**Step 2. Subscribe to the WANdisco Fusion AMI**

This Quick Start requires a subscription to the AMI for WANdisco Fusion in the AWS Marketplace.

To subscribe:


2. Type **WANdisco** in the search box, and choose **WANdisco Fusion – BYOL** from the results.
Figure 2: Finding the WANdisco Fusion AMI in AWS Marketplace

3. Open the page for WANdisco Fusion, and choose **Continue**.
4. Use the **Manual Launch** tab. Read the terms and conditions of software usage, and then choose **Accept Software Terms**.

![Figure 4: Accepting license terms](image)

You will get a confirmation page confirming your subscription, and an email confirmation will be sent to the account owner. For detailed instructions, see the [AWS Marketplace documentation](#).

When you subscribe to this AMI, the Quick Start deploys WANdisco Fusion with a 14-day trial license.

**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: Deploy the Quick Start into a new VPC on AWS](Launch)

![Option 2: Deploy the Quick Start into an existing VPC on AWS](Launch)

**Important** If you’re deploying the Quick Start into an existing VPC, make sure that your VPC has two public subnets in different Availability Zones.

Each deployment takes about 15 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 the hybrid data lake will be built. The template is launched in the US East (Ohio) Region by default.

**Note** If you want to include Athena in your deployment, choose an [AWS Region where Athena is supported](aws.amazon.com/athena/).

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 the hybrid data lake into a new VPC
- Parameters for deploying the hybrid data lake into an existing VPC
- **Option 1: Parameters for deploying the hybrid data lake into a new VPC**

  **View template**

  **VPC Network Configuration:**

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Availability Zones</strong> (AvailabilityZones)</td>
<td><strong>Requires input</strong></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><strong>VPC CIDR</strong> (VPCCIDR)</td>
<td>10.0.0.0/16</td>
<td>The CIDR block for the VPC.</td>
</tr>
<tr>
<td><strong>Public Subnet 1 CIDR</strong> (PublicSubnet1CIDR)</td>
<td>10.0.128.0/20</td>
<td>The CIDR block for the public (DMZ) subnet located in Availability Zone 1.</td>
</tr>
<tr>
<td><strong>Public Subnet 2 CIDR</strong> (PublicSubnet2CIDR)</td>
<td>10.0.144.0/20</td>
<td>The CIDR block for the public (DMZ) subnet located in Availability Zone 2.</td>
</tr>
</tbody>
</table>

  **WANdisco Fusion Cluster Configuration:**

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cluster Name</strong> (ClusterName)</td>
<td>awsfs</td>
<td>The name of the WANdisco Fusion cluster.</td>
</tr>
<tr>
<td><strong>Cluster Node Type</strong> (ClusterNodeType)</td>
<td>m3.2xlarge</td>
<td>The EC2 instance type for WANdisco Fusion nodes.</td>
</tr>
<tr>
<td><strong>Fusion server storage</strong> (PersistentStorage)</td>
<td>0</td>
<td>The EBS storage to allocate for each block device (in GiB, with 4 devices per node). The default value of 0 (zero) indicates ephemeral storage only. You can specify 0-1024 GiB per device (4 TiB per node).</td>
</tr>
<tr>
<td><strong>Cluster Instance Count</strong> (ClusterInstanceCount)</td>
<td>1</td>
<td>The number of instances in the WANdisco Fusion cluster.</td>
</tr>
<tr>
<td><strong>EC2 key pair name</strong> (KeyName)</td>
<td><strong>Requires input</strong></td>
<td>A 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.</td>
</tr>
<tr>
<td><strong>Local Fusion Access CIDR</strong> (LocalFusionAccess)</td>
<td><strong>Requires input</strong></td>
<td>The CIDR block for local WANdisco Fusion access.</td>
</tr>
<tr>
<td><strong>Administration Access CIDR</strong> (AdminFusionAccess)</td>
<td><strong>Requires input</strong></td>
<td>The CIDR block for administration access.</td>
</tr>
</tbody>
</table>
### WANdisco Fusion Data Storage Configuration:

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replicate to Existing Bucket (S3BucketExisting)</td>
<td>false</td>
<td>Set to <strong>true</strong> if the destination bucket is an existing S3 bucket. If <strong>true</strong>, specify the bucket name in the next (Data Replication S3 Bucket) parameter. The default value of <strong>false</strong> creates a new bucket with the name specified in the next parameter.</td>
</tr>
<tr>
<td>Data Replication S3 Bucket (S3Bucket)</td>
<td>Requires input</td>
<td>The name of the new or existing S3 bucket to use to replicate local HDFS drive data.</td>
</tr>
<tr>
<td>KMS Key (KMSKey)</td>
<td>—</td>
<td>(Optional) The Amazon Resource Name (ARN) for the AWS Key Management Service (Amazon KMS) encryption key ID. Leave this parameter blank to disable AWS KMS encryption.</td>
</tr>
<tr>
<td>S3 Server-side Encryption Algorithm (S3ServerEncryption)</td>
<td>Yes</td>
<td>Set to <strong>No</strong> to disable server-side encryption in Amazon S3.</td>
</tr>
</tbody>
</table>

### WANdisco Fusion Application Configuration:

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zone Name (ZoneName)</td>
<td>AWSCloud</td>
<td>The name used to identify the zone in which the WANdisco Fusion server operates.</td>
</tr>
<tr>
<td>Fusion Admin Username (Username)</td>
<td>admin</td>
<td>The name of the default administrator user for WANdisco Fusion.</td>
</tr>
<tr>
<td>Fusion Admin Password (Password)</td>
<td>Requires input</td>
<td>The password for the WANdisco Fusion administrator. The password must be alphanumeric and must include one of these special characters (!@#$&amp;*)</td>
</tr>
<tr>
<td>ARN Topic to publish messages to (SubscribeARN)</td>
<td>—</td>
<td>(Optional) The ARN of the topic for emailing stack status notifications.</td>
</tr>
<tr>
<td>EMR Version (EMRVersion)</td>
<td>5.4.0</td>
<td>The version of Amazon EMR, if you decide to attach an Amazon EMR cluster to replicate data back to your on-premises server after deployment. The two options are 5.3.0 and 5.4.0.</td>
</tr>
<tr>
<td>Fusion License (FusionLicense)</td>
<td>—</td>
<td>The path to the S3 bucket (in the format s3://bucket-name/path) or the URL of the license key for the WANdisco Fusion license. Leave this parameter blank to use a trial license.</td>
</tr>
</tbody>
</table>
Athena Configuration:

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create Athena Table (AthenaCreateTable)</td>
<td>true</td>
<td>Specify false if you don’t want deploy Athena. By default, the Quick Start creates an Athena table for data analysis.</td>
</tr>
<tr>
<td>Athena Output to Existing Bucket</td>
<td>false</td>
<td>Specify true if you want to place Athena output in an existing S3 bucket. If true, specify the bucket name in the next (Athena Output Bucket) parameter. The default value of false creates a new S3 bucket with the name specified in the next parameter.</td>
</tr>
<tr>
<td>Athena Output Bucket (AthenaBucket)</td>
<td></td>
<td>The name of the new or existing S3 bucket to use for Athena output.</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 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.</td>
</tr>
<tr>
<td>Quick Start S3 Key Prefix (QSS3KeyPrefix)</td>
<td>quickstart-datalake-wandisco/</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 the hybrid data lake 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) where you want to deploy the WANdisco Fusion instances. This VPC must be configured to allow bidirectional network connectivity with all your other WANdisco Fusion servers. The VPC must have an appropriate IP address range, associated public subnet, route table, network gateway, and security settings.</td>
</tr>
</tbody>
</table>
## Parameter label (name) | Default | Description
--- | --- | ---
VPC Subnet ID for Zone 1 (SubnetIdA) | Requires input | The ID of the public subnet in Availability Zone 1 in your existing VPC (e.g., subnet-a0246dcd) where you want to deploy the WANdisco Fusion instances. You must ensure that the IP address range of this subnet is routable from your other Fusion servers so that they can communicate directly with an IP address in this subnet. You must also ensure that the subnet has a route table defined that allows hosts within it to communicate with all other Fusion servers.

VPC Subnet ID for Zone 2 (SubnetIdB) | Requires input | The ID of the public subnet in Availability Zone 2 in your existing VPC (e.g., subnet-b58c3d67) where you want to deploy the WANdisco Fusion instances. You must ensure that the IP address range of this subnet is routable from your other Fusion servers so that they can communicate directly with an IP address in this subnet. You must also ensure that the subnet has a route table defined that allows hosts within it to communicate with all other Fusion servers.

## WANdisco Fusion Cluster Configuration:

| Parameter label (name) | Default | Description |
--- | --- | ---
Cluster Name (ClusterName) | awsfs | The name of the WANdisco Fusion cluster. |
Cluster Node Type (ClusterNodeType) | m3.2xlarge | The EC2 instance type for WANdisco Fusion nodes. |
Fusion server storage (PersistentStorage) | 0 | The EBS storage to allocate for each block device (in GiB, with 4 devices per node). The default value of 0 (zero) indicates ephemeral storage only. You can specify 0-1024 GiB per device (4 TiB per node). |
Cluster Instance Count (ClusterInstanceCount) | 1 | The number of instances in the WANdisco Fusion cluster. |
EC2 key pair name (KeyName) | Requires input | A 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. |
Local Fusion Access CIDR (LocalFusionAccess) | Requires input | The CIDR block for local WANdisco Fusion access. |
Administration Access CIDR (AdminFusionAccess) | Requires input | The CIDR block for administration access. |
### WANdisco Fusion Data Storage Configuration:

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replicate to Existing Bucket (S3BucketExisting)</td>
<td>false</td>
<td>Set to <strong>true</strong> if the destination bucket is an existing S3 bucket. If <strong>true</strong>, specify the bucket name in the next (Data Replication S3 Bucket) parameter. The default value of <strong>false</strong> creates a new bucket with the name specified in the next parameter.</td>
</tr>
<tr>
<td>Data Replication S3 Bucket (S3Bucket)</td>
<td>Requires input</td>
<td>The name of the new or existing S3 bucket to use to replicate local HDFS drive data.</td>
</tr>
<tr>
<td>KMS Key (KMSKey)</td>
<td>—</td>
<td>(Optional) The Amazon Resource Name (ARN) for the AWS Key Management Service (Amazon KMS) encryption key ID. Leave this parameter blank to disable AWS KMS encryption.</td>
</tr>
<tr>
<td>S3 Server-side Encryption Algorithm (S3ServerEncryption)</td>
<td>Yes</td>
<td>Set to <strong>No</strong> to disable server-side encryption in Amazon S3.</td>
</tr>
</tbody>
</table>

### WANdisco Fusion Application Configuration:

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zone Name (ZoneName)</td>
<td>AWSCloud</td>
<td>The name used to identify the zone in which the WANdisco Fusion server operates.</td>
</tr>
<tr>
<td>Fusion Admin Username (Username)</td>
<td>admin</td>
<td>The name of the default administrator user for WANdisco Fusion.</td>
</tr>
<tr>
<td>Fusion Admin Password (Password)</td>
<td>Requires input</td>
<td>The password for the WANdisco Fusion administrator. The password must be alphanumeric and must include one of these special characters (!@#$&amp;*).</td>
</tr>
<tr>
<td>ARN Topic to publish messages to (SubscribeARN)</td>
<td>—</td>
<td>(Optional) The ARN of the topic for emailing stack status notifications.</td>
</tr>
<tr>
<td>EMR Version (EMRVersion)</td>
<td>5.4.0</td>
<td>The version of Amazon EMR, if you decide to attach an Amazon EMR cluster to replicate data back to your on-premises server after deployment. The two options are 5.3.0 and 5.4.0.</td>
</tr>
<tr>
<td>Fusion License (FusionLicense)</td>
<td>—</td>
<td>The path to the S3 bucket (in the format s3://bucket-name/path) or the URL of the license key for the WANdisco Fusion license. Leave this parameter blank to use a trial license.</td>
</tr>
</tbody>
</table>
Athena Configuration:

<table>
<thead>
<tr>
<th>Parameter label (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create Athena Table (AthenaCreateTable)</td>
<td>true</td>
<td>Specify <strong>false</strong> if you don’t want deploy Athena. By default, the Quick Start creates an Athena table for data analysis.</td>
</tr>
<tr>
<td>Athena Output to Existing Bucket (AthenaBucketExisting)</td>
<td>false</td>
<td>Specify <strong>true</strong> if you want to place Athena output in an existing S3 bucket. If <strong>true</strong>, specify the bucket name in the next parameter. The default value of <strong>false</strong> creates a new S3 bucket with the name specified in the next parameter.</td>
</tr>
<tr>
<td>Athena Output Bucket (AthenaBucket)</td>
<td><strong>Requires input</strong></td>
<td>The name of the new or existing S3 bucket to use for Athena output.</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 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.</td>
</tr>
<tr>
<td>Quick Start S3 Key Prefix (QSS3KeyPrefix)</td>
<td>quickstart-datalake-wandisco/</td>
<td>The <strong>S3 key name prefix</strong> 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>

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 data lake deployment is ready.
Step 4. (Optional) Synchronize the Local HDFS with the S3 Bucket

Deploy Local WANdisco Fusion Server

Optionally, you can deploy an on-premises WANdisco server in a Docker container on your computer, to see the synchronization from HDFS to the S3 bucket in the cloud.

To deploy this server, follow these steps:

1. **Download** the docker-compose.yml file to your computer and note the folder name.

2. Modify docker-compose.yml and change `AWS_FUSION_HOST` to the remote Fusion server IP. You can choose any of the servers created by the Auto Scaling group. To find this information, open the Amazon EC2 console at [https://console.aws.amazon.com/ec2/](https://console.aws.amazon.com/ec2/), choose **Auto Scaling Groups** in the navigation pane, and search for the Auto Scaling group listed on the **Outputs** tab.
3. Start Docker by navigating to the folder you downloaded docker-compose.yml to, and run `docker-compose up` from the command line.

4. Wait for Docker to start up.
**Set up Replication**

1. In your browser, navigate to `http://localhost:8083/` and access your local WANdisco Fusion instance.

2. Log in with the user name and password you set in [step 3](#) with the **Fusion Admin Username** and **Fusion Admin Password** parameters.

   You should see two bubbles labeled *Fusion* and *Fusion local* in your dashboard. These should be green if they are able to communicate.

![Figure 5: Bubbles in WANdisco Fusion dashboard](image)

3. Navigate to **Memberships** from the top navigation pane.

4. Choose **Create New** to create a new membership.

5. Select your local instance as the **Distinguished Node**, and then choose **Create**. Your local instance should be labeled **AWSCloud**, if you used the default parameter values.
6. Choose the **Replication** tab. In the **Replication Rules** section, choose **Create**.

7. Choose `/user/sample` from the file hierarchy, and choose the two zones.

8. Select the membership you created, and then choose **Create** to finish creating the replication rule.

**Test Replication**

1. Run the following command to download and insert data into HDFS:

   ```bash
docker-compose exec fusion /usr/bin/load.sh
   ```

**Note** Replication may be slow, depending on your connection.
2. Open the Amazon Web Services Management Console for Athena.

3. From the Database list, choose `wandisco_fusion_db_lab_stack_name`.

4. In the left navigation pane, choose `taxi_tripdata`. Choose the action icon (three dots) for `taxi_tripdata`, and then choose Preview table to see the results of the query.

![Figure 7: Using Athena to query the S3 bucket](image)

**FAQ**

**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.

**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.

### Additional Resources

**AWS services**

- Amazon EC2

- AWS CloudFormation
  https://aws.amazon.com/documentation/cloudformation/

- Amazon VPC
  https://aws.amazon.com/documentation/vpc/

**WANdisco Fusion**

- WANdisco product guides
  https://www.WANdisco.com/support/product-guides

- WANdisco Fusion manual
  https://docs.WANdisco.com/bigdata/wdfusion/2.10/

- WANdisco Fusion troubleshooting
  https://docs.WANdisco.com/bigdata/wdfusion/2.10/#troubleshooting

### 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

<table>
<thead>
<tr>
<th>Date</th>
<th>Change</th>
<th>In sections</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 2017</td>
<td>Initial publication</td>
<td>—</td>
</tr>
</tbody>
</table>

© 2017, Amazon Web Services, Inc. or its affiliates, Sturdy, and WANdisco. 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/](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.