Qubole on AWS Data Lake

Quick Start Reference Deployment

September 2017
Last update: June 2018 (revisions)

Qubole Team
AWS Quick Start Reference Team

Contents

This Quick Start deployment guide was created by Amazon Web Services (AWS) in partnership with Qubole.

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

Overview

This Quick Start deployment guide provides step-by-step instructions for deploying and configuring a production-ready Qubole Data Service (QDS) environment that is built on a data lake foundation in the AWS Cloud. You can use this Qubole environment to process and analyze your own datasets and implement your own use cases. The Quick Start optionally deploys an environment with prepopulated data, notebooks, and queries to analyze structured and semi-structured data, and gain key business insights into product sales performance for a fictional online retailer.

QDS is a cloud-native, data activation platform that helps operationalize the data lake, reducing the costs and complexities of managing big data. Qubole self-manages and constantly analyzes and learns about the platform's usage with heuristics and machine learning. It provides insights and recommendations to optimize reliability, performance, and costs. Qubole works in concert with AWS services such as Amazon Simple Storage Service (Amazon S3) and Spot instances in Amazon Elastic Compute Cloud (Amazon EC2).

This Quick Start is for data infrastructure professionals (data architects, data administrators, data operators), data engineers, extract, transform, load (ETL) engineers, and data scientists who want to deploy a self-managed and self-optimized autonomous data platform to gain insights into data that resides in a data lake on AWS.

Costs and Licenses

You are responsible for the cost of the AWS services used while running this Quick Start reference deployment. 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. See the pricing pages for each AWS service you will be using for cost estimates.

The Quick Start deploys QDS Business Edition, which allows you to consume up to 10,000 Qubole Compute Usage Hours (QCUH) per month at no cost. However, you are responsible for the cost of AWS resources that Qubole manages on your behalf. To learn more about QDS Business Edition, see the Qubole FAQ.

After you deploy the Quick Start, you can upgrade to QDS Enterprise Edition and use Qubole Cloud Agents, which provide actionable Alerts, Insights, and Recommendations (AIR) to optimize reliability, performance, and costs. To upgrade your license to QDS Enterprise Edition, see the Enterprise Edition upgrade webpage on the Qubole website.

Architecture

Deploying this Quick Start for a new virtual private cloud (VPC) with **default parameters** builds the following Qubole environment in the AWS Cloud.

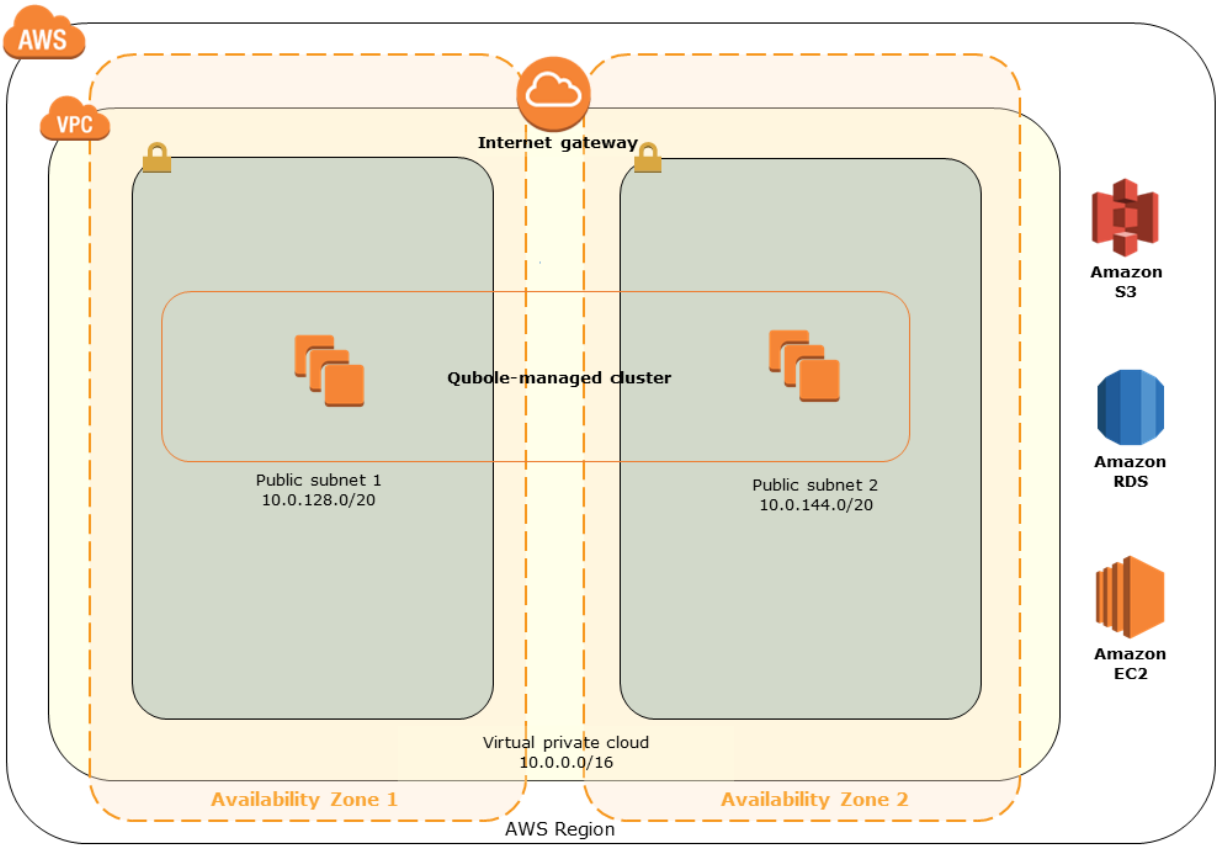


Figure 1: Quick Start architecture for Qubole on the AWS Cloud

This Quick Start adds the following components:

- A standard VPC, which is extended to support communications between instances in the public subnet and Qubole SaaS, and to provide access to the metastore within Qubole SaaS.
- Preconfigured Apache Spark and Hadoop clusters. These clusters are managed by Qubole and are automatically started and scaled depending on the user's workloads.
- Preconfigured data sources that provide access to Amazon Relational Database Service (Amazon RDS) and S3 buckets in the data lake.
- Preconfigured Qubole metastore, notebooks, and queries to show business insights.
- A basic wizard that helps you with Qubole account configuration and dataset deployment.
- Data analysis and visualization, using Qubole's Analyze and Notebooks interfaces.

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 S3
- Amazon EC2
- Amazon RDS
- Amazon VPC

Quick Start Sample Dataset

This Quick Start includes an optional dataset from a fictional online retailer. The dataset includes structured data from the products database hosted in Amazon RDS, and unstructured data from the web logs that record customer interactions with the company website, hosted in Amazon S3. The Quick Start helps you correlate and analyze both datasets to get key business insights.

Deployment Options

This Quick Start provides two deployment options:

- **Deploy the Quick Start 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 configures Qubole to use this new VPC.
- **Deploy the Quick Start into an existing VPC**. This option deploys Qubole clusters and components in your existing AWS infrastructure.

The Quick Start provides separate templates for these options. It also lets you configure CIDR blocks, and Qubole 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 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 data lake foundation on AWS.

3. If necessary, request a service limit increase for the Amazon EC2 **m3.xlarge** 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. Create a Qubole Account

If you don't already have a Qubole account, create one by following the on-screen instructions on the Qubole website. (Please click the preceding link; do not copy/paste the link text into your browser.)

When you sign up, an activation code will be sent to your email address along with a link to confirm the account and to choose your password. (You can skip activation and choose your password immediately if you sign up using Google or LinkedIn.)

Step 3. Obtain a Qubole API Token, Trusted Principal AWS Account ID, and External ID

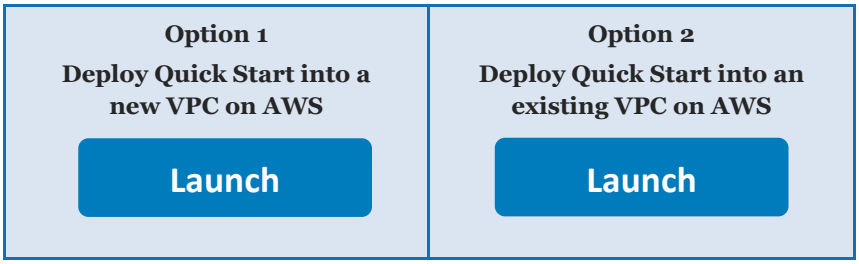
1. Log in to your Qubole account.
2. Prepare your Qubole API token: In the Qubole Control Panel, in the left pane, choose **My Accounts**. Choose **Show** for your account, and copy the API token that is displayed.
3. Prepare your Qubole trusted principal AWS account ID: In the Qubole Control Panel, in the left pane, choose **Account Settings**. In the **Access Mode (Keys/IAM Roles)** section, choose **IAM Role**, and then copy the **Trusted Principal AWS Account ID** that is displayed.
4. Prepare your Qubole external ID: In the Qubole Control Panel, in the left pane, choose **Account Settings**. In the **Access Mode (Keys/IAM Roles)** section, choose **IAM Role**, and then copy the **External ID** that is displayed.

You will use these tokens and IDs for parameter settings in step 4. After you deploy the Quick Start, you will come back to the Qubole Control Panel and provide values from the outputs of the Quick Start, as explained in step 5.

Step 4. Launch the Quick Start

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

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



> **Important** If you're deploying the Quick Start into an existing VPC, make sure that your VPC has two public subnets in different Availability Zones. You'll be prompted for your VPC settings when you launch the Quick Start.

   Each deployment takes about 12 minutes to complete.

2. Check the region that's displayed in the upper-right corner of the navigation bar, and change it if necessary. The template is launched in the US West (Oregon) 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 separately for the two deployment options:

   - Parameters for deploying the Quick Start into a new VPC
   - Parameters for deploying the Quick Start into an existing VPC

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

View template

Network Configuration:

Parameter label (name)	Default	Description
Availability Zones (AvailabilityZones)	*Requires input*	The list of Availability Zones to use for the subnets in the VPC. The Quick Start requires two Availability Zones and preserves the logical order you specify.
VPC Definition (VPCDefinition)	QuickstartDefault	The VPC definition name from the `Mappings` section of the template. Each definition specifies a VPC configuration, including the number of Availability Zones to be used for the deployment and the CIDR blocks for the VPC, public subnets, and private subnets. You can support multiple VPC configurations by extending the map with additional definitions and choosing the appropriate name. If you don't want to change the VPC configuration, keep the default setting. For more information, see the Adding VPC Definitions section.

RDS Configuration:

Parameter label (name)	Default	Description
RDS User Name (RDSUsername)	rdsuser	The user name that is associated with the master user account for the Amazon RDS database that is created. The user name must be lowercase, begin with a letter, contain only alphanumeric characters or underscores, and be less than 128 characters.
RDS Password (RDSPassword)	*Requires input*	The password that is associated with the master user account for the Amazon RDS database that is created. The password must contain 8-64 printable ASCII characters, excluding /, ", \', \ and @. It must contain one uppercase letter, one lowercase letter, and one number.
RDS Database Name (RDSDatabaseName)	qubole	The name of the database created when the RDS instance is provisioned.
RDS Instance Type (RDSInstanceType)	db.t2.small	The instance type of the RDS instance that is created.
RDS port (RDSPort)	3306	The port that the RDS instance will listen on.

Qubole Configuration:

Parameter label (name)	Default	Description
Qubole API token (QuboleApiToken)	*Requires input*	The Qubole account API token, from step 3.
Qubole AWS account ID (QuboleAWSAccountId)	*Requires input*	The Qubole AWS account ID, from step 3. Navigate to https://api.qubole.com/v2/control-panel. In the left pane, choose **Account Settings**. Under **Access Mode (Keys/IAM Roles)**, choose **IAM Role**, and copy the Trusted Principal AWS Account ID that is displayed.
Qubole External ID (QuboleExternalId)	*Requires input*	The Qubole account external ID, from step 3. Navigate to https://api.qubole.com/v2/control-panel. In the left pane, choose **Account Settings**. Under **Access Mode (Keys/IAM Roles)**, choose **IAM Role**, and copy the external ID that is displayed.

Demonstration Configuration:

Parameter label (name)	Default	Description
Create Demonstration (CreateDemonstration)	Select option	Set this parameter to **yes** if you want the Quick Start to deploy the Qubole wizard, create an EC2 instance, and load sample data into Amazon RDS. For more information about the wizard, see step 6.
The following parameters are used only if **Create Demonstration** is set to **yes.**		
Wizard User Name (WizardUsername)	QuboleUser	The user name for the wizard, consisting of 1-64 ASCII characters.
Wizard Password (WizardPassword)	*Requires input*	The password for the wizard, consisting of 8-64 ASCII characters. The password must contain one uppercase letter, one lowercase letter, and one number. This password is required, but it will be used only when you launch the Quick Start with **Create Demonstration** set to **yes**.
Key Pair Name (KeyPairName)	*Requires input*	Name of an existing EC2 key pair to enable SSH access to the web server instance.
Remote Access CIDR (RemoteAccessCIDR)	*Requires input*	The CIDR block allowed to access the web server and SSH into the web server instance. You can use http://checkip.amazonaws.com/ to check your IP address. The CIDR block parameter must be in the form x.x.x.x/x (e.g., 96.127.8.12/32, YOUR_IP/32).

AWS Quick Start Configuration:

Parameter label (name)	Default	Description
Quick Start S3 Bucket Name (QSS3BucketName)	aws-quickstart	S3 bucket where the Quick Start templates and scripts are installed. Use this parameter to specify the S3 bucket name you've created for your copy of Quick Start assets, if you decide to customize or extend the Quick Start for your own use. The bucket name can include numbers, lowercase letters, uppercase letters, and hyphens, but should not start or end with a hyphen.
Quick Start S3 Key Prefix (QSS3KeyPrefix)	quickstart-datalake-qubole/	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. It cannot start or end with forward slash (/) because they are automatically appended.

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

View template

Network Configuration:

Parameter label (name)	Default	Description
Existing VPC ID (VPCID)	*Requires input*	The ID of your existing VPC (e.g., vpc-0343606e).
Existing VPC CIDR (VPCCIDR)	*Requires input*	The CIDR block for your existing VPC.
Existing VPC Public Subnet 1 ID (PublicSubnet1ID)	*Requires input*	The ID of the public subnet in Availability Zone 1 (e.g., subnet-a0246dcd).
Existing VPC Public Subnet 2 ID (PublicSubnet2ID)	*Requires input*	The ID of the public subnet in Availability Zone 2 (e.g., subnet -e3246d8e).

RDS Configuration:

Parameter label (name)	Default	Description
RDS User Name (RDSUsername)	rdsuser	The user name that is associated with the master user account for the Amazon RDS database that is created. The user name must be lowercase, begin with a letter, contain only alphanumeric characters or underscores, and be less than 128 characters.
RDS Password (RDSPassword)	*Requires input*	The password that is associated with the master user account for the Amazon RDS database that is created. The password must contain 8-64 printable ASCII characters, excluding /, ", \', \ and @. It must contain one uppercase letter, one lowercase letter, and one number.
RDS Database Name (RDSDatabaseName)	qubole	The name of the database created when the RDS instance is provisioned.
RDS Instance Type (RDSInstanceType)	db.t2.small	The instance type of the RDS instance that is created.
RDS port (RDSPort)	3306	The port that the RDS instance will listen on.

Qubole Configuration:

Parameter label (name)	Default	Description
Qubole API token (QuboleApiToken)	*Requires input*	The Qubole account API token, from step 3.
Qubole AWS account ID (QuboleAWSAccountId)	*Requires input*	The Qubole AWS account ID, from step 3. Navigate to https://api.qubole.com/v2/control-panel. In the left pane, choose **Account Settings**. Under **Access Mode (Keys/IAM Roles)**, choose **IAM Role**, and copy the Trusted Principal AWS Account ID that is displayed.
Qubole External ID (QuboleExternalId)	*Requires input*	The Qubole account external ID, from step 3. Navigate to https://api.qubole.com/v2/control-panel. In the left pane, choose **Account Settings**. Under **Access Mode (Keys/IAM Roles)**, choose **IAM Role**, and copy the external ID that is displayed.

Demonstration Configuration:

Parameter label (name)	Default	Description
Create Demonstration (CreateDemonstration)	Select option	Set this parameter to **yes** if you want the Quick Start to deploy the Qubole wizard, create an EC2 instance, and load sample data into Amazon RDS. For more information about the wizard, see step 6.
The following parameters are used only if **Create Demonstration** is set to **yes**.		
Wizard User Name (WizardUsername)	QuboleUser	The user name for the wizard, consisting of 1-64 ASCII characters.
Wizard Password (WizardPassword)	*Requires input*	The password for the wizard, consisting of 8-64 ASCII characters. The password must contain one uppercase letter, one lowercase letter, and one number. This password is required, but it will be used only when you launch the Quick Start with **Create Demonstration** set to **yes**.
Key Pair Name (KeyPairName)	*Requires input*	Name of an existing EC2 key pair to enable SSH access to the web server instance.
Remote Access CIDR (RemoteAccessCIDR)	*Requires input*	The CIDR block allowed to access the web server and SSH into the web server instance. You can use http://checkip.amazonaws.com/ to check your IP address. The CIDR block parameter must be in the form x.x.x.x/x (e.g., 96.127.8.12/32, YOUR_IP/32).

AWS Quick Start Configuration:

Parameter label (name)	Default	Description
Quick Start S3 Bucket Name (QSS3BucketName)	aws-quickstart	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.
Quick Start S3 Key Prefix (QSS3KeyPrefix)	quickstart-datalake-qubole/	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 deployment is complete.
9. You can view the resources that were created in the **Outputs** tab.

Step 5. Finish the Qubole Configuration

1. Check the **Outputs** tab in the AWS CloudFormation console. It should display the following resources:
   - **QuboleRoleARN**
   - **QuboleLoggingBucketName**

   You will need the values of these resources to complete the configuration of your Qubole account.
2. Change the Qubole access mode and default location for logging:
   a. Open the Qubole Control Panel.
   b. In the left pane, choose **Account Settings**.
   c. In the **Access Mode (Keys/IAM Roles)** section, choose **IAM Role**.
   d. In the **Role ARN** box, type the value for **QuboleRoleARN** from the AWS CloudFormation **Outputs** tab.
   e. In the **Default Location box**, type the value for **QuboleLoggingBucketName** from the AWS CloudFormation **Outputs** tab.
   f. Choose **Save** to save your changes.

Step 6. Test the Deployment

Create Demonstration Parameter Set to No

If you set the **Create Demonstration** parameter to **no** and want to use Qubole to analyze your own data, perform the following steps:

1. If you haven't already done so, log in to your Qubole account.
2. Your Qubole account has default cluster definitions for Spark, Hadoop1, Hadoop2, and Presto. If you want to override the default configurations (node type, VPC, etc.), click the top-left corner menu in the Qubole UI, and choose **Clusters**. Identify the cluster you want to change and, on the right side, choose **Edit**.

3. Qubole clusters are turned off by default, and you don't need to manually start a cluster to use it. Qubole automatically starts a cluster when you submit a command to it using Qubole Analyze, command-line interface, or a REST API, and will shut it down automatically after an hour of inactivity. To use any of the default clusters for analyzing your own data, follow these steps:
    a. Make sure you already have data in Amazon S3.
    b. Follow the instructions in the Qubole documentation:
        - To run a Hadoop workload, see http://docs.qubole.com/en/latest/quick-start-guide/AWS-quick-start-guide/running-hadoop-job.html.
        - To run a Spark workload, see http://docs.qubole.com/en/latest/quick-start-guide/AWS-quick-start-guide/running-spark-app.html.
        - To use Qubole Notebooks, see http://docs.qubole.com/en/latest/quick-start-guide/AWS-quick-start-guide/running-spark-app-in-notebooks.html.
        - To use Hive, see http://docs.qubole.com/en/latest/quick-start-guide/AWS-quick-start-guide/running-hive-query.html.

Create Demonstration Parameter Set to Yes

If you set the **Create Demonstration** parameter to **yes**, you'll see a URL for the wizard in the **Outputs** tab of the AWS CloudFormation console. Follow these steps to use the wizard:

1. Check the **Outputs** tab in the AWS CloudFormation console for the **QuboleWizardWebAppURL** resource.

> **Note** After CREATE_COMPLETE of the stack, it may take one more minute to bootstrap the wizard instances and set up the web server. We recommend that you try the wizard URL at least one minute after the stack creation.

2. Use the URL for **QuboleWizardWebAppURL** to open the Qubole wizard in your browser.

   Log in to the wizard, shown in Figure 2, by using the credentials you set during deployment. Use the **Wizard User Name** value as your login name, and the **Wizard Password** value as your password.

Figure 2: Login page of the wizard

The step-by-step wizard guides you through Qubole features. It includes eight steps, each of which demonstrates and explains a particular Qubole feature. For example, the Notebooks step walks you through the process to visualize data insights interactively. For more information about the sample data used by the wizard, see the appendix.

When you log in, you will see the **Get Started** screen shown in Figure 3. Follow the instructions in the wizard to step through the path from initial data ingest to transformations, to analytics, and finally to visualizations.

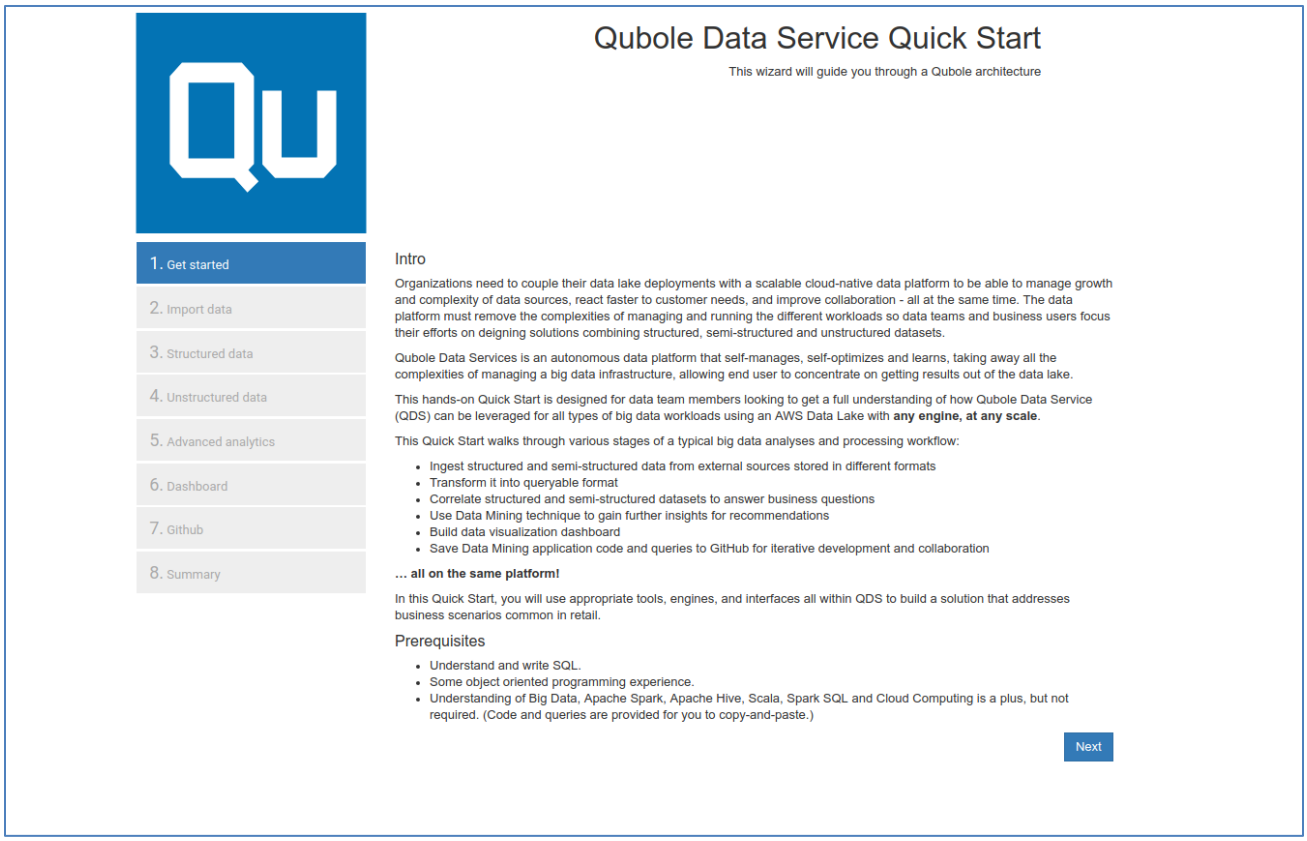


Figure 3: Getting started with the wizard

Optional: Adding VPC Definitions

When you launch the Quick Start in the mode where a new VPC is created, the Quick Start uses VPC parameters that are defined in a mapping within the Quick Start templates. If you choose to download the templates from the GitHub repository, you can add new named VPC definitions to the mapping, and choose one of the named VPC definitions that you have defined when you launch the Quick Start.

The following table shows the parameters defined within each VPC definition. You can define as many VPC definitions as you need within your environments. When you deploy the Quick Start, use the **VPCDefinition** parameter to specify the configuration you want to use.

Parameter	Default	Description
NumberOfAZs	2	Number of Availability Zones to use in the VPC.
PublicSubnet1 CIDR	10.0.1.0/24	CIDR block for the public (DMZ) subnet 1 located in Availability Zone 1.
PublicSubnet2 CIDR	10.0.3.0/24	CIDR block for the public (DMZ) subnet 2 located in Availability Zone 2.
VPCCIDR	10.0.0.0/16	CIDR block for the VPC.

Troubleshooting and FAQ

Wizard Error Messages

The following issues may arise if you launch the Quick Start using an existing Qubole account whose configuration may differ from a new Qubole account. You might also encounter these circumstances if you run the Quick Start multiple times with the same Qubole account.

Q. I chose the **Create clusters and notebooks** button in the Get Started section of the wizard and it says that clusters are starting up. However, I don't see clusters starting in the Qubole UI. What should I do?

A. This can occur if Qubole is not able to communicate with instances in your VPC. You can troubleshoot this problem by looking at cluster startup logs. To access the logs, choose **Clusters** on the top-left menu, choose the cluster ID of the cluster you want to troubleshoot (e.g. "38096"), and then choose **Cluster Start Logs**.

Q. I chose the **Create clusters and notebooks** button in the **Get Started** section of the wizard and received the error message: "Notebook dashboard_quickstart Validation failed: Name has already been taken." What should I do?

A. The Qubole account already has a notebook named "dashboard_quickstart" and you must manually remove the notebook. From the Qubole UI, choose the Qubole Notebooks menu, choose the **Common** tab, and remove the notebook. Then redeploy the Quick Start by selecting the top-level AWS CloudFormation stack, deleting it, and launching the Quick Start again.

Q. I chose the **Create clusters and notebooks** button in the **Get Started** section of the wizard and received the error message: "Cannot delete cluster with ID <*xxxxx*> because it is running. Please terminate it and try again." What should I do?

A. Terminate the Hadoop2 and Spark clusters from the Qubole UI. At the top-left, choose **Clusters**, locate the Hadoop2 or Spark cluster that corresponds to the cluster ID in the message, and on the right side choose **Stop**. Redeploy the Quick Start by selecting the top-level AWS CloudFormation stack, deleting it, and launching the Quick Start again.

General 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.

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, you might encounter template size limitations when you create the stack. For more information about AWS CloudFormation limits, see the AWS documentation.

Datasets and Upgrades

Q. Can I use the QuickStart with my own data?

A. Yes. The Qubole environment configured in this Quick Start is production-ready and can be extended for additional big data use cases through custom datasets. However, the transformations, analytics, and visualizations featured by the Quick Start were developed for the sample dataset. If you're using your own dataset, transformations, analytics, and visualizations may be different. You can use the instructions in step 6 (**Create Demonstration** parameter set to **no**) to use the Quick Start with your own dataset.

Q. The Quick Start uses QDS Business Edition, but I want to extend to use it with other datasets and I will likely use more than the 10,000 QCUH included. How can I upgrade to the next version?

A. To upgrade to Qubole Enterprise Edition, log in to your Qubole account and open the **Control Panel**. Choose **Subscription and Payment**, and then choose **Contact us to upgrade to Enterprise Edition**. A Qubole sales representative will contact you to discuss your options.

Additional Resources

AWS services

- AWS CloudFormation
  http://aws.amazon.com/documentation/cloudformation/
- Amazon EBS
  http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AmazonEBS.html
- Amazon EC2
  http://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/
- Amazon VPC
  http://aws.amazon.com/documentation/vpc/
- Amazon S3
  https://aws.amazon.com/documentation/s3/
- Amazon Elasticsearch Service (Amazon ES)
  https://aws.amazon.com/documentation/elasticsearch-service/

Qubole

- Qubole
  https://qubole.com/

Quick Start reference deployments

- AWS Quick Start home page
  https://aws.amazon.com/quickstart/
- Quick Start for Data Lake Foundation on the AWS Cloud
  https://fwd.aws/J8xBm

Appendix: Sample Dataset

The Quick Start includes an optional sample dataset that the Qubole wizard uses to demonstrate how Qubole can transform, query and analyze data for a fictional online retailer. The wizard shows how to correlate structured data (from the products database) with unstructured data (from web logs) to analyze product sales performance. QDS helps you analyze the dataset to answer key business questions, such as:

- Which products do customers like to buy?
  - What are the top 10 most popular product categories?
  - What are the top 10 revenue generating products?
- Do the most viewed products also sell the most?
  - Which products are viewed a lot but not purchased?
- What are the top 10 two-product combinations purchased together?
- What are the top 5 products with total transactions per order status?

The key data domains for the fictional retailer include:

Categories Data

- Category_id
- Category_department_id
- Category_name

Customers Data

- Customer_id
- Customer_name
- Customer_lname
- Customer_email
- Customer_password
- Customer_street
- Customer_city
- Customer_state
- Customer_zipcode

Departments Data

- Department_id
- Department_name

Order Items Data

- Order_item_id
- Order_item_order_id
- Order_item_product_id
- Order_item_quantity
- Order_item_subtotal
- Order_item_product_price

Order Data

- Order_id
- Order_date
- Order_customer_id
- Order_status

Products Data

- Product_id
- Product_category_id
- Product_name
- Product_description
- Product_price
- Product_image

Web_logs – Semi-structured data like the following:

79.133.215.123 - - [14/Jun/2014:10:30:13 -0400] "GET /home HTTP/1.1" 200 1671 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"

The Qubole wizard walks you through the following flow:

- Qubole architecture overview
  - Customizing a QDS Business Edition account
- Ingesting structured data from a MySQL database
  - Creating a data store in Qubole that connects to a MySQL database in Amazon RDS
  - Creating Apache Hive tables in Qubole and importing structured data stored in the MySQL database
- Querying structured data in Qubole

  - Querying the top 10 most popular products
  - Querying the top 10 revenue generating products
- Correlating structured data with semi-structured data
  - Creating Hive tables in Qubole for semi-structured web logs data stored in Amazon S3
  - Querying top viewed products
  - Determining top viewed products that are not being sold
- Advanced analytics -- gaining insights into product relationships
  - Creating an Apache Spark application in Scala; using the FPGrowth data mining MLlib algorithm to mine a set of frequent patterns
  - Querying top 10 two-product combinations purchased together
- Building a dashboard in Qubole Notebooks
  - Discover the total orders by date
  - Interactive chart with total orders by month and year
- Saving the Apache Spark application to GitHub
  - Creating a new GitHub repository and token
  - Configuring the GitHub token in Qubole
  - Linking a Spark Notebook with your GitHub profile
  - Committing the Spark Notebook to GitHub

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
June 2018	Revised architecture to remove AWS Lambda, Amazon Kinesis Data Firehose, Amazon Redshift, bastion hosts, and private subnets For Quick Start deployment into a new and existing VPC, removed parameters for configuring Hadoop, Spark, Amazon Redshift, and Amazon Kinesis Data Firehose For deployment into an existing VPC, removed parameters for configuring bastion hosts and private VPC	Architecture; Option 1: Parameters for deploying the Quick Start into a new VPC; Option 2: Parameters for deploying the Quick Start into an existing VPC
September 2017	Initial publication	—

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.