Secure Data Commons

Data Analyst User Guide

www.its.dot.gov/index.htm
Draft Report — April 13, 2020
FHWA-JPO-18-xxx

Produced by DTFH6116H00027
U.S. Department of Transportation

Notice

Revision History

#	Name	Version	Revision Date	Revision Description
1	REAN Cloud	1.0	08/02/2018	Initial Draft
2	REAN Cloud	2.0	09/13/2018	Added the Export Functionality
3	REAN Cloud	3.0	10/29/2019	Added the Manage Workstations chapter and updated the Guide per feedback comments
4	SDC Team	4.0	02/06/2020	• Added description of data ingestion and curation to Chapter 1 • Added description and link to instructions for setting up and logging in with Login.gov in Chapter 2 • Added note and example command to the "Download User Data from S3 Bucket through Portal" section under Chapter 2 • Added FMI section and link to GitLab page with more sample queries to Chapter 4 Sample Queries for SDC Datasets • Added note about printing documents in Chapter 5 • Added Chapter 6 Setting Up SDC with GitLab Repositories • Added more commands to the "AWS S3 CLI Commands" section under Chapter 7
5	SDC Team	5.0	04/13/2020	• Added "Connecting to Waze Data in Redshift Using Python" section to Chapter 3 Accessing and Launching Workstations • Removed Chapter 4 Sample Queries for SDC Datasets due to security concerns; transferred all content to GitLab • Added new Step 5 to Chapter 4 Exporting Datasets from the SDC to describe trusted users as well as further instructions on using the auto-export functionality • Added new Q&A about sample queries to Chapter 7 Frequently Asked Questions

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction and Document Overview

The Secure Data Commons (SDC) is a United States Department of Transportation (U.S. DOT) sponsored cloud-based analytical sandbox designed to create wider access to sensitive transportation datasets, with the goal of advancing the state of the art of transportation research and state/local traffic management.

The SDC stores sensitive transportation data made available by participating Data Providers, and grants access to approved researchers to these datasets. The SDC also provides access to open source tools and allows researchers to collaborate and share code with other system users.

The SDC platform is a research environment that allows users to conduct analyses and do development and testing of new tools and software products. It is not intended to be an alternative to any local jurisdiction's traffic management center or local data repository. The current SDC platform provides users with the following data, tools, and features:

- **Data**: The SDC is ingesting several datasets currently. Additional datasets will be added to the environment over time. Users can bring their own data into the environment to use along with the Waze data.
- **Tools**: The environment provides access to open source tools including Python, RStudio, Microsoft R, SQL Workbench, Power BI, Libre Office, and Jupyter Notebook. These tools are available on a virtual machine in the system enabling data analytics in the cloud.
- **Functionality**: Users can access and analyze data within the environment, save their work to a virtual machine, and publish processes and results to share with other SDC users.

The SDC platform supports two major roles:

- **Data Providers** – These are entities that provide data hosted on the SDC platform. The Data Provider establishes the data protection needs and acceptable use terms for the data analysts.
- **Data Analysts** – These are entities that conduct analysis using the datasets hosted within the SDC system. Note that analysts can bring their own data and tools into the SDC system.

During a project's onboarding phase, Data Providers work with the SDC support team to describe their project's data (for example, the type of data, frequency of new data, data file formats, etc., every project's data is unique). Then, Data Providers upload data files to designated "S3 Ingestion Buckets" (a secure, scalable object storage service provided by the SDC platform through Amazon Web Services). We call this the "Data Lake."

As new data arrives to the SDC, policies and procedures established by the Data Provider then govern who, when, and how Data Analysts can access the data. It is common that once new data arrives in the SDC, automated processes "ingest" and "curate" the data, making the data available in other forms. For example, some data may be loaded into our Data Warehouse tools (Redshift or Hadoop databases), whereas other data may be transformed from its source format into other easier-to-use formats, or filtered through a process to identify corrupt, invalid, or duplicate data. Exactly which automated processes a project's data undergoes is determined during the project onboarding processes.

Once data has been ingested and curated, it is then available to Data Analysts through the tools listed above (note that we are always adding new Data Analyst tools based on request). A typical Data Analyst workflow may be:

- Use a tool to develop SQL queries to see a subset of the larger data set.
- Compare the data subset against models to draw unique insights (for example, develop programs utilizing the analytic capabilities of R or Python).
- Use powerful tools that present data and insights in graphical format (some use the power of Python and GeoPandas, others have developed graphical outputs in R Studio, whereas others use Libre Office as an open source alternative to Microsoft Excel). The SDC support team has worked with yet other Data Analysts to install proprietary licensed software to enhance their analytical capabilities.

Finally, there are capabilities by which Data Analysts can export their work out of the SDC, subject to the data use agreements and approval of the Data Providers.

This document provides guidance for the **Data Analyst** role. A similar guide will be prepared for the Data Providers. This document is organized as follows:

- Initial Setup and Validation
- Workstation Access
- Sample Queries
- Exporting Data
- Importing and Exporting Code
- Accessing External Data Sources within SDC
- Setting Up GitLab with SDC
- Technical Support and Contact Information
- Frequently Asked Questions

Prerequisites

Workstation access will not be granted for a Data Analyst user until the user has:

1. Submitted a completed Access Request Form;
2. Received approval for the request;
3. Received an email message with onboarding instructions from the support team; and

4. Received a walkthrough of the system from the support team.

Refer to the Useful Links section later in the document for further information on technologies relevant to SDC.

Chapter 2. Initial Setup and Validation

This chapter provides guidance on the initial setup and validation of the user into the SDC system.

Accessing Secure Data Commons Portal

Users can access the SDC web portal by navigating to https://portal.securedatacommons.com.

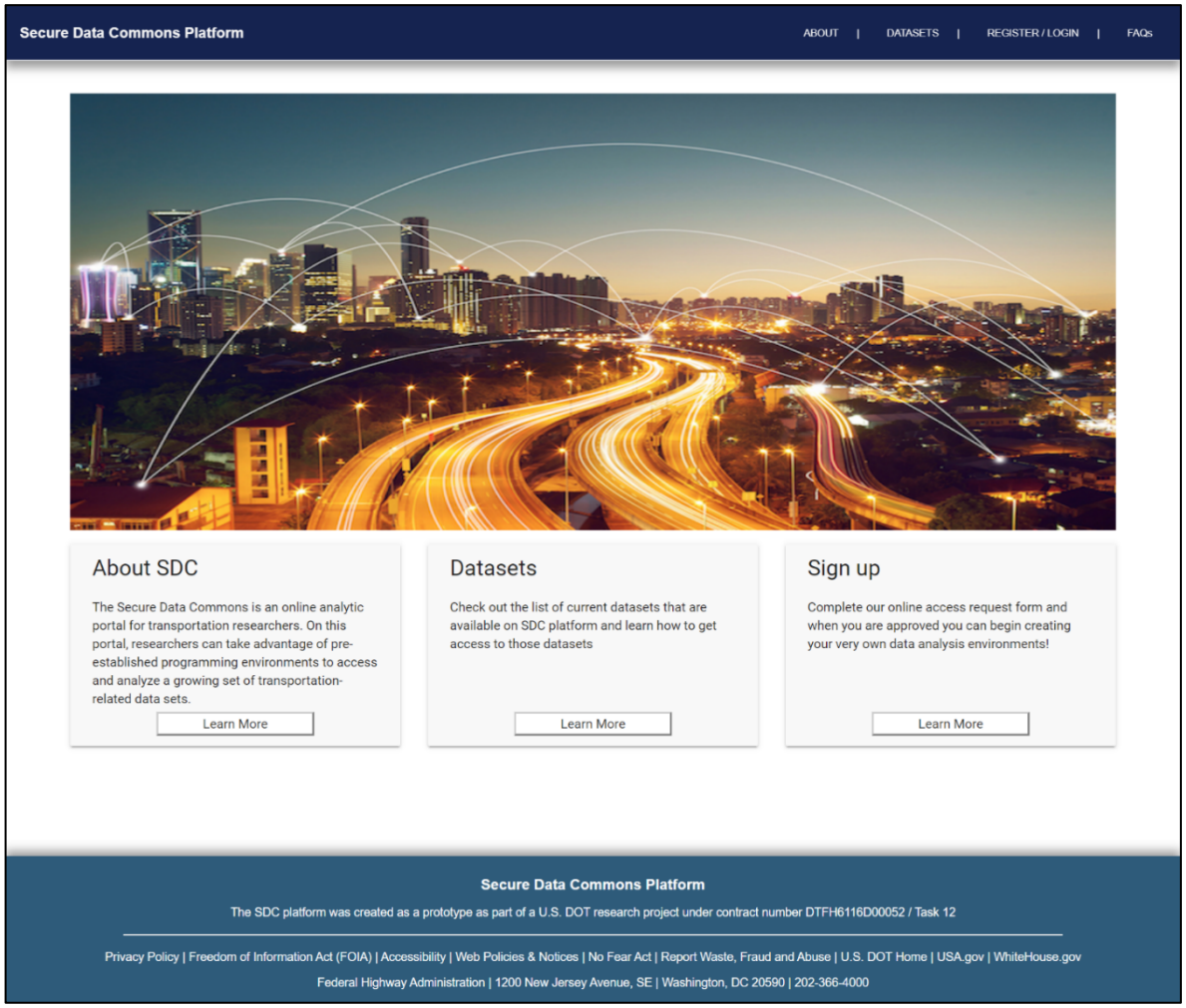


Figure 1: SDC Starting Homepage

Select Register/Login from the top menu to display the Access Request Form and Privacy Policy links, as well as the email address sign-in prompt. The Access Request Form takes you to a PDF form in which you can provide the support team with contact info, a rationale for your access request, your preferred workstation type, and sign-off for the SDC data use agreements.

If you have a DOT email address (@dot.gov), enter it into Email Address and then select Sign In to access the portal. If you do not have a DOT email address, continue to Page 6.

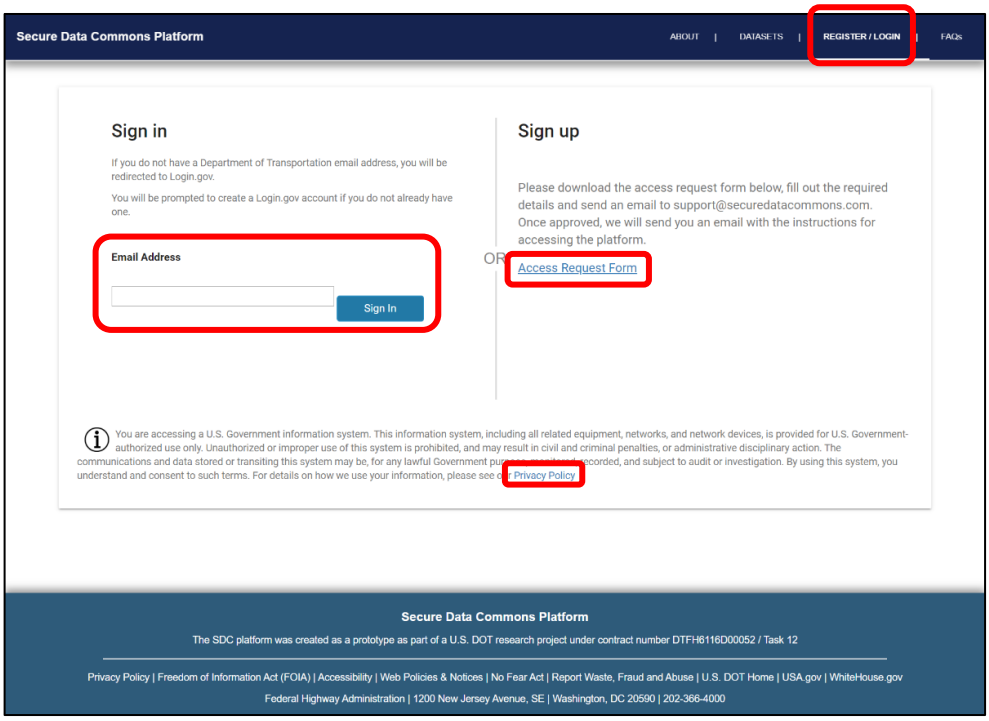


Figure 2: SDC Portal – Register/Login Page

After signing in, you will be redirected to the Secure Data Commons platform login page:

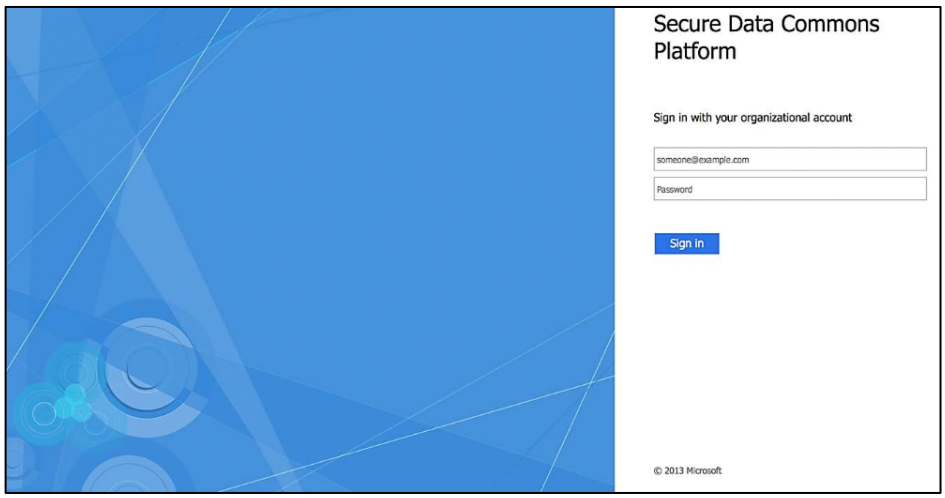


Figure 3: SDC Portal – Login Page

If you do not have a DOT email address, you will need to create a new Login.gov account or log in with your existing one so that it can be linked with your SDC credentials. Linking a Login.gov account with your SDC credentials provides extra security by adding protection you can configure as two layers of authentication. Authentication methods will consist of phone text or call; an app on your phone, tablet, or computer; a security key; a government employee ID; and pre-generated backup codes.

Enter the email address you would like to use as your Login.gov account into the portal's Email Address sign in (Figure 2 above) and then select Sign In to be redirected to the Login.gov website, where you can create a new account. If you already have an existing Login.gov account, enter your credentials and then select Sign in.

If you do not have an existing Login.gov account, select "Create an account" on the page you are redirected to:

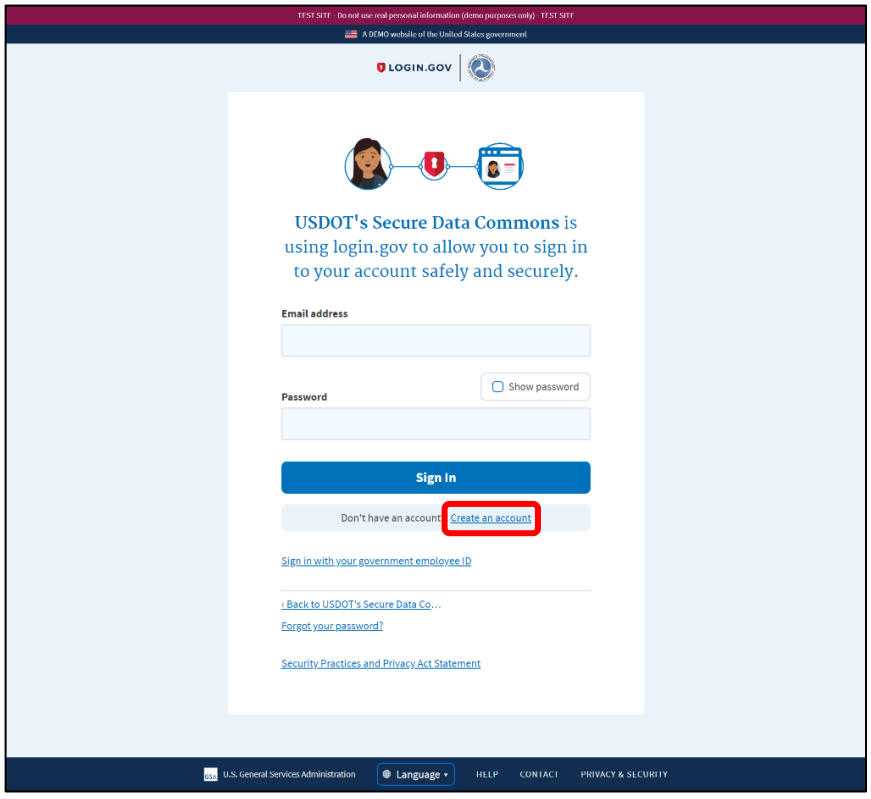


Figure 4: Login.gov – Create New Account

For further instructions on setting up your new Login.gov account and configuring its secure authentication methods, refer to https://www.login.gov/help/creating-an-account/how-to-create-an-account/.

After you have finished setting up all authentication methods or signed in with an existing Login.gov account, you will be redirected to a one-time sign-in form on the portal. Enter your SDC ADFS credentials (provided in your welcome email) and then select Sign in.

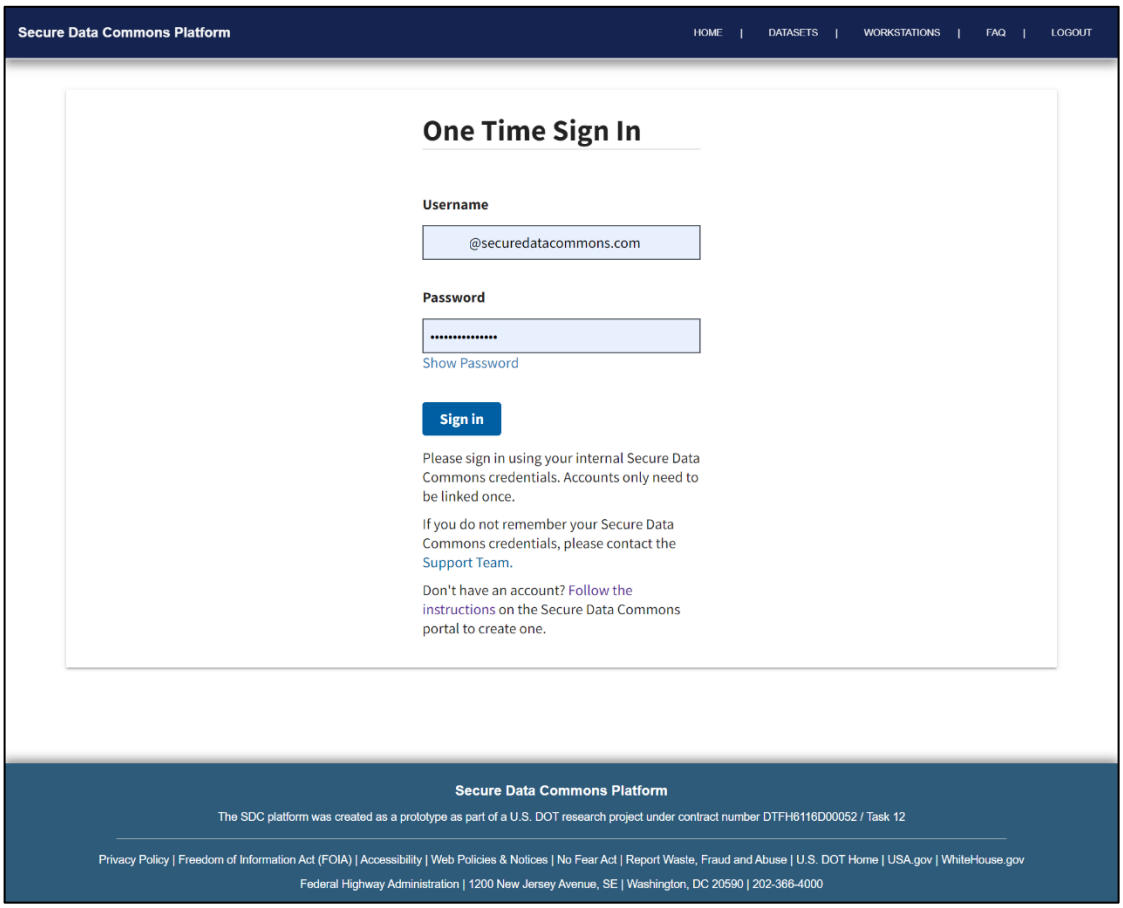


Figure 5: SDC Portal – Synchronize Login.gov Account with SDC

If you are accessing the portal for the first time, you will be prompted to change your password after entering the credentials provided in the welcome email.

Upon successfully logging in, you will be redirected to the landing page, which provides an overview of Secure Data Commons and the different actions you can perform from the web portal:

1. Request access to curated and published datasets
2. Access to workstations with programming tools
3. Bring your own datasets / algorithms
4. Publish your datasets / algorithms

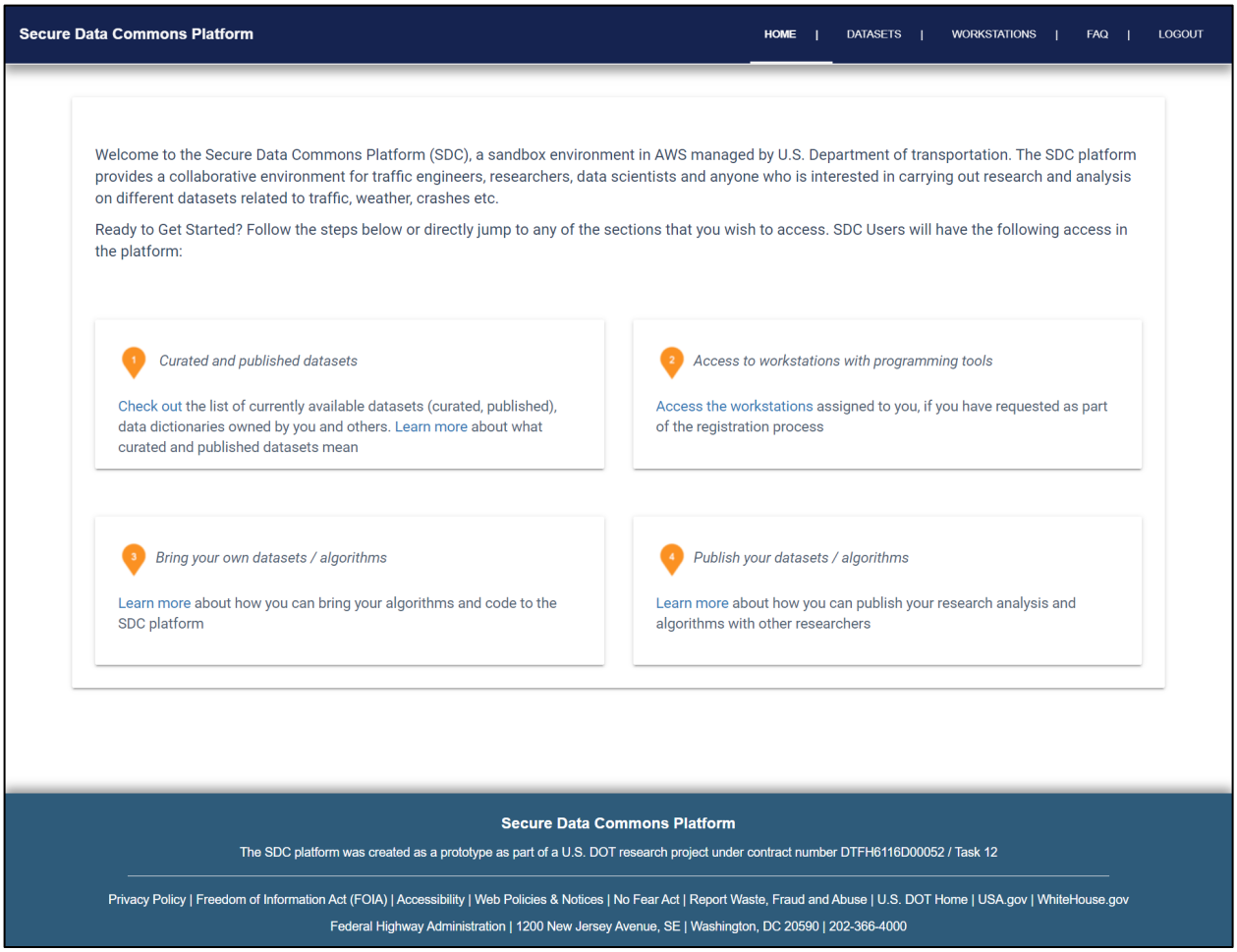


Figure 6: SDC Portal – Landing Page After Login

Request Access to Datasets

Users can request access to the datasets that are available within the SDC platform as published / enabled by the SDC team or published by other users.

Once you are logged in, go to Datasets in the top menu.

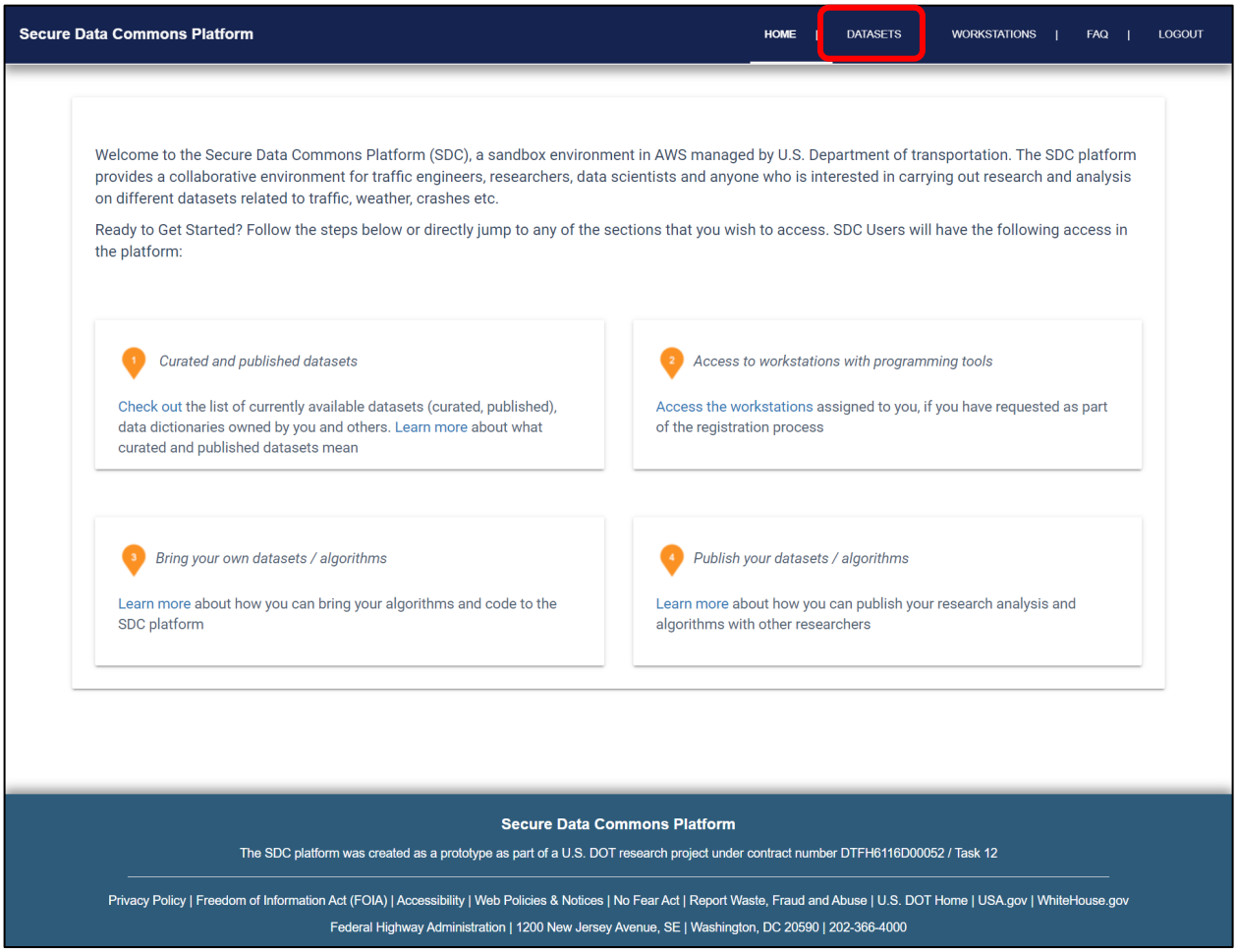


Figure 7: SDC Portal – Datasets Option

Expand the SDC Datasets. You will be able to see all available datasets in the SDC platform. To access a dataset, click on Request.

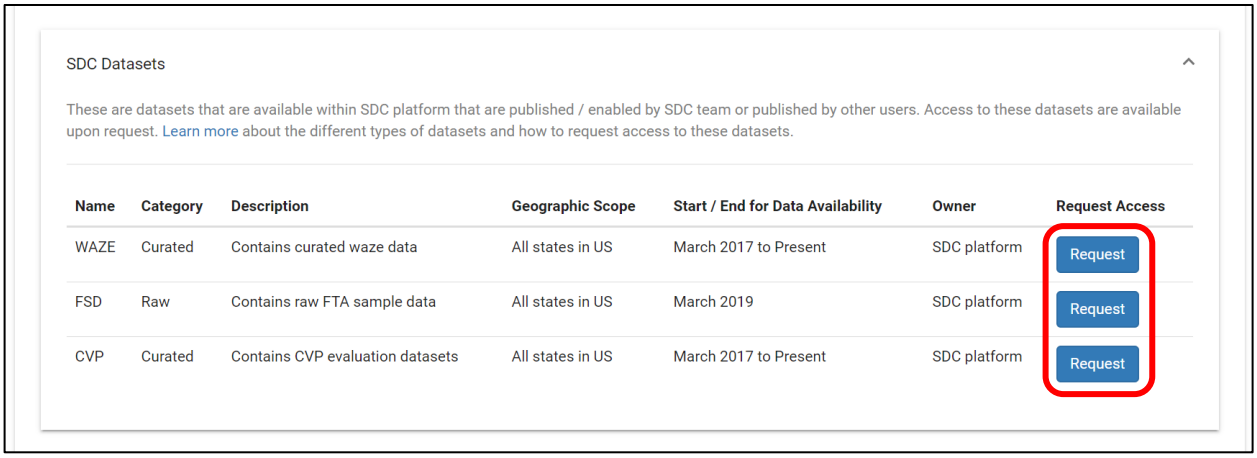


Figure 8: SDC Portal – Request Dataset Access

Complete the SDC Data Access Request form that appears. Once completed, click on Send Request.

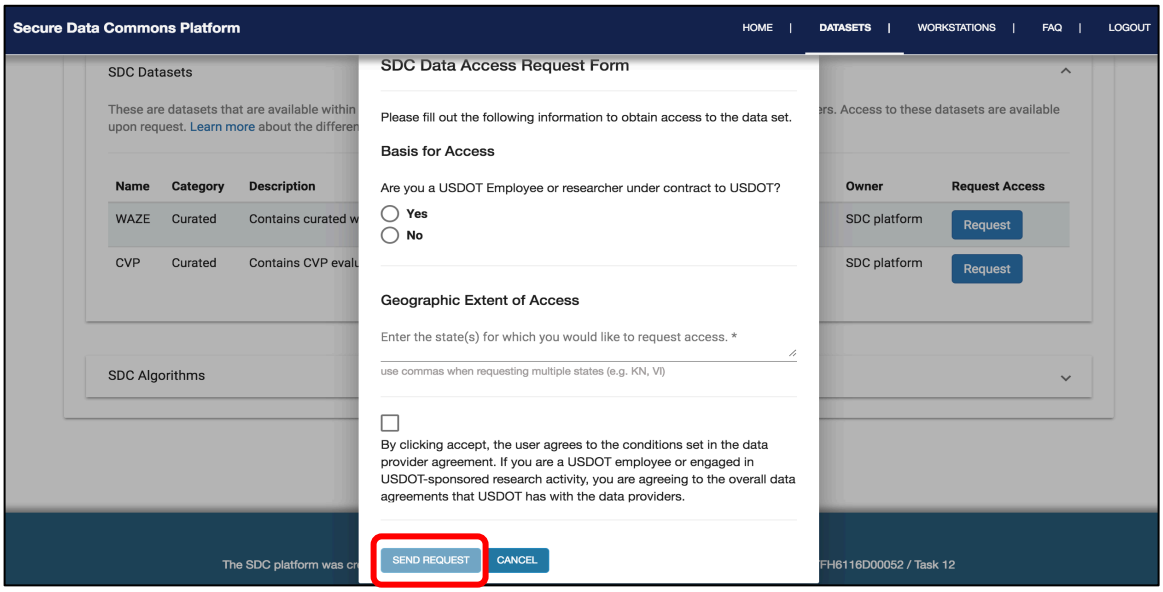


Figure 9: SDC Portal – Send Data Access Request

The request will be sent to the support team and access to the requested dataset will be given upon validation and approval of the information in the form.

Upload User Data to S3 Bucket through Portal

Users who want to share data with other users from their project team can upload their own data to their assigned team/individual buckets through the portal.

1. Click on Datasets from the home page.
2. Click on Upload Files under "My Datasets / Algorithm."
3. A pop-up window appears prompting you to choose one or more files for upload to the assigned bucket. (The assigned bucket name will be displayed on the upload pop-up window.)

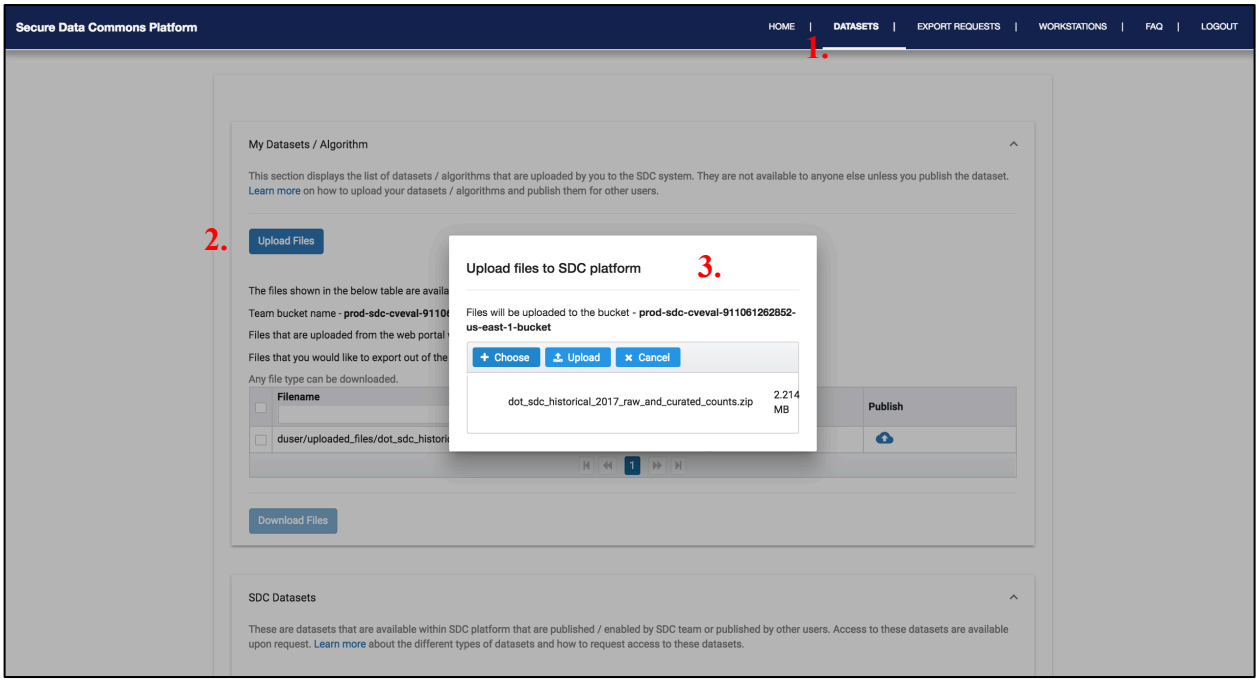


Figure 10: SDC Portal – Upload Files

4. A success message will be displayed upon a successful upload.

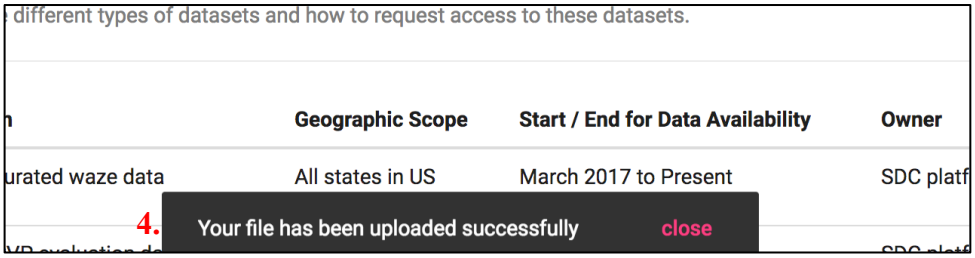


Figure 11: SDC Portal – Upload Success

5. Files that are uploaded from the web portal will be saved in the folder – **username/uploaded_files**
6. To make these files available to members of the project team, users then need to move files from username/uploaded_files to the project team S3 bucket using command-line tools (see AWS S3 CLI Commands for an overview).

7. Users would be able to access only the files that are under the **uploaded_files** and **export_requests** folders.

Download User Data from S3 Bucket through Portal

Users can download their data from their assigned team/individual buckets through the portal.

1. Click on Datasets from the home page.
2. All the available files under **username/uploaded_files** in the assigned bucket will be displayed along with the assigned bucket name under My Datasets / Algorithm.
3. Select the files that you want to download and then click on Download Files.
4. Users should go through the export request workflow to download files that are uploaded under **export_requests** folder. Export requests workflow can be found by clicking here.

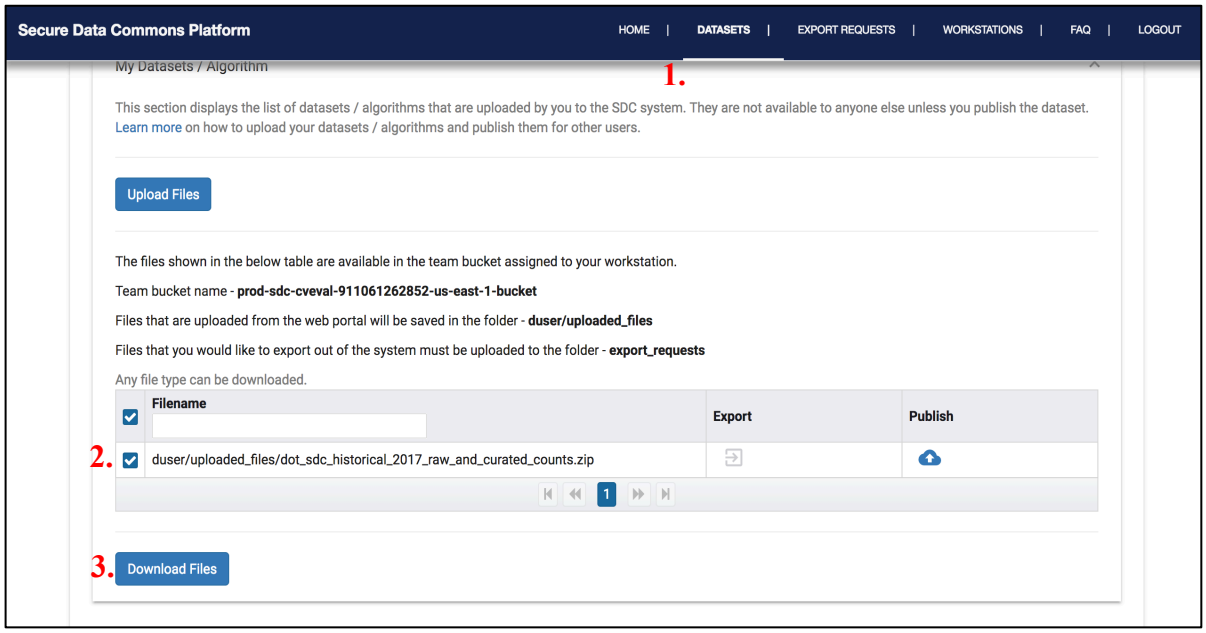


Figure 12: SDC Portal – Selecting Files for Download

Notes:

- Not all the files are downloaded directly. Files with extensions such as .txt, .png, or .pdf will be opened in a separate tab from where they can be downloaded. All other files with extensions like .csv, .zip, etc. can be downloaded directly.
- Files are downloaded individually.
- The Filename box allows searches for partial filenames. This can be used to download all the contents of a sub-folder in an S3 bucket by searching for the sub-folder name and then clicking the box next to Filename to select all objects.
- Files must be copied to and from S3 buckets using the SDC workstation. **NOTE:** Files stored in S3 buckets are not visible through Windows Explorer but can be copied to their SDC workstation.
    - For example: To copy the file shown above in Figure 12 that is hosted in an S3 bucket to your SDC workstation, use the command:

```
aws s3 cp s3://prod-sdc-cveval-911061262852-us-east-1-
bucket/duser/uploaded_files/dot_sdc_historical_2017_ra
w_and_curated_counts.zip
dot_sdc_historical_2017_raw_and_curated_counts.zip
```

Chapter 3. Accessing and Launching Workstations

Users are assigned cloud-based workstations to perform analysis on the datasets. This section provides a description of how to launch and use these workstations.

Launch Workstations

1. Users can see the assigned workstations by clicking on WORKSTATIONS from the top menu. By default, all the workstations are in an inactive state.
2. Click on Start to start the workstation.

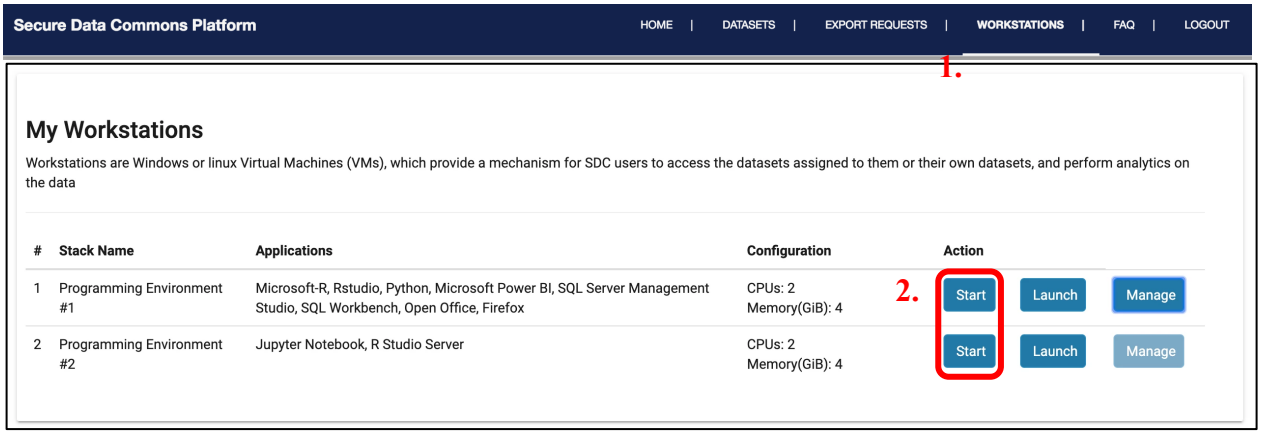


Figure 13: SDC Portal – Starting Workstations

3. The workstation should become available within five minutes; you may not see any change immediately. A message will appear when the workstation has been successfully started.

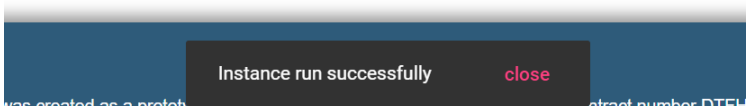


Figure 14: SDC Portal – Run Instance Success

4. Now click Launch for the workstation.

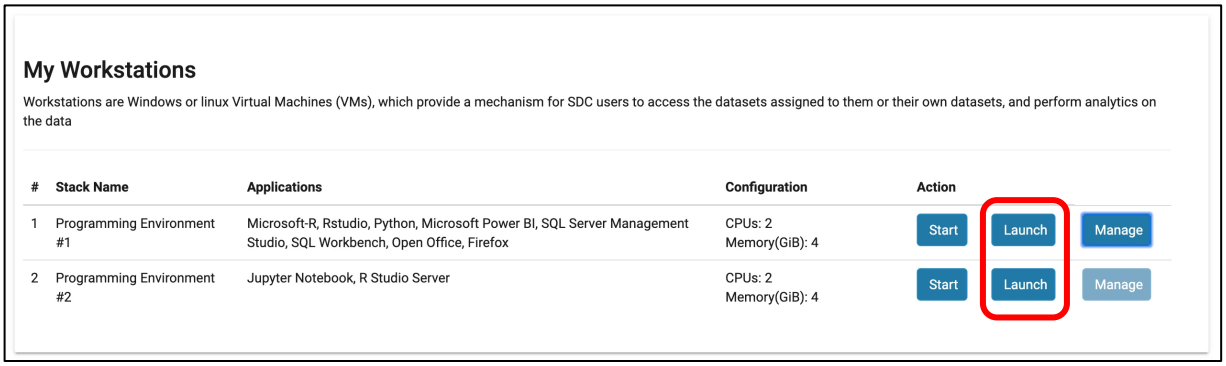


Figure 15: SDC Portal – Launch Workstations

5. This will provide a user access to their workstation within the browser. The workstation may take a few minutes to initialize. When complete, a login screen will appear. User is prompted to re-enter a valid username and password.

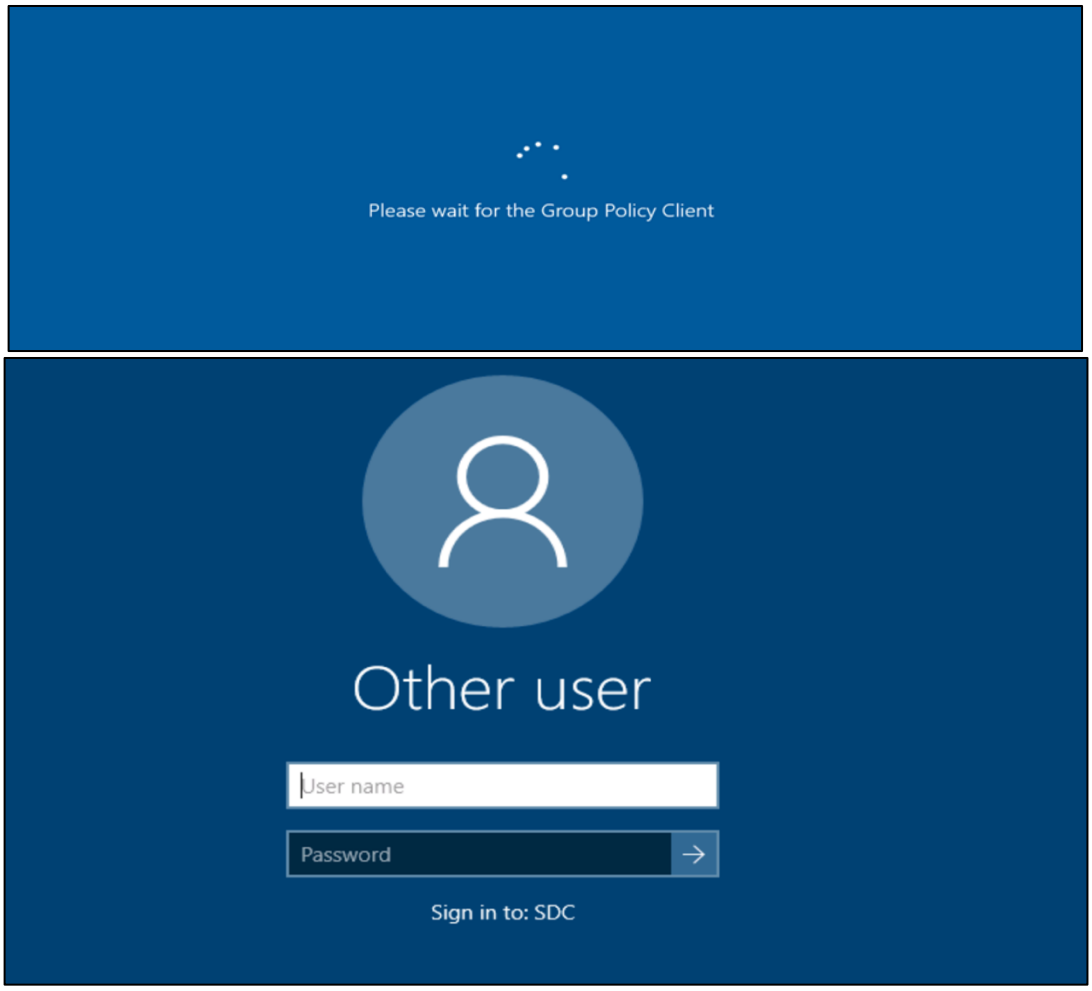


Figure 16: SDC Workstation – Initialization and Login Screens

Software Validation

By default, users will have the following installed on their workstations:

- Java
- Python
- R, RStudio
- SQL Workbench
- Power BI
- AWS CLI
- Adobe
- Libre Office
- Visual Studio
- PuTTY
- Firefox

Connecting to the Data Warehouse

The following sections illustrate how the user can connect to the data stores available to the SDC.

Connecting to Waze Data in Redshift Using SQL Workbench

Launch SQL Workbench by double-clicking the SQL Workbench shortcut on the desktop.

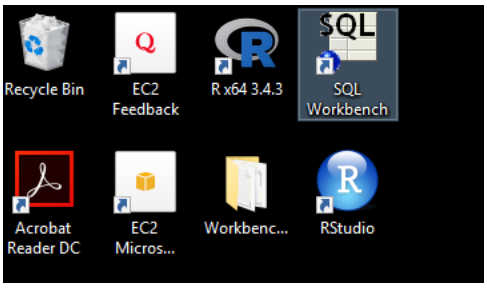


Figure 17: SDC Workstation – SQL Workbench Icon

Create a Redshift connection profile to connect to Waze data:

1. Create a new connection profile by selecting the top left corner icon on the "Select Connection Profile" window.
2. Select "Amazon Redshift Driver" from the Driver drop-down.
3. Update the URL section with the Redshift URL provided in the email from the support desk detailing Redshift login credentials.
4. Provide your username and password received in the welcome email.
5. Click on the Test button at the bottom to test the connection. A pop-up dialog will appear confirming a successful or failed connection.

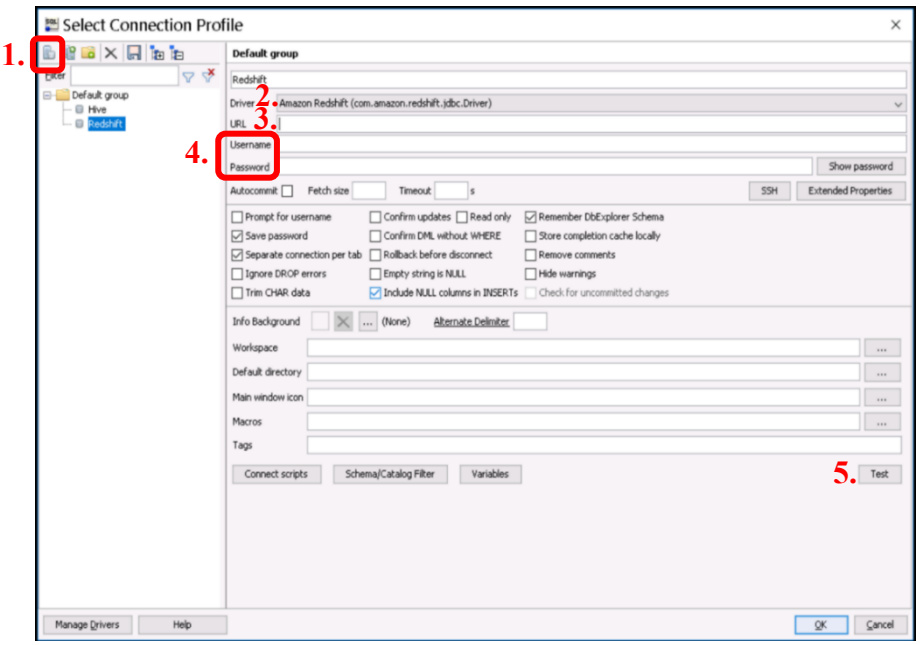


Figure 18: SQL Workbench – Create Redshift Connection Profile

Connecting to Waze Data in Redshift Using Python

NOTE: When you are granted access to Waze data, the SDC support team creates a new Redshift user for you, assigns it with a Redshift password, and emails you with information on the Redshift host you will connect to. This email message provides the `redshiftHost`, `userName`, and `userPassword` values shown below. Your Redshift credentials are only used for connecting to Redshift and NOT for accessing the portal, which uses your separate SDC credentials.

Important: The default version of Python installed on the SDC Windows Workstations is v2.7.4. There are two required Python modules that must be installed prior to attempting to connect to Redshift with Python using the example code below. To install these modules, open a Windows Command Prompt, and enter the following two commands:

```
C:\Users\username> pip install psycopg2
C:\Users\username> pip install numpy
```

The above "pip install …" command(s) only need to be run ONCE on the SDC Windows Workstation. Once the Python modules are installed, they remain available, even across reboots of the workstation.

To test Python connectivity to Redshift, open the IDLE python editor and execute the following:

```
from __future__ import print_function

import psycopg2
import numpy

dbName = 'dot_sdc_redshift_db'
redshiftHost = '[host address]'
redshiftPort = 5439

userName = '[username]'
userPassword = '[password]'

# query = 'select * from dw_waze.alert limit 10;'
query = "select * from dw_waze.alert where
alert_type='ACCIDENT' and city = 'Severance, CO'"

conn = psycopg2.connect(
    dbname=dbName,
    host=redshiftHost,
    port=redshiftPort,
```

```
        user=userName,
        password=userPassword)

    cursor = conn.cursor()

    cursor.execute(query)
    result = cursor.fetchall()

    result = numpy.array(result)

    # print(result)
    for r in result:
        print (r[1], r[8], r[18], r[19], r[22], sep='\t')
```

For further information and examples, refer to the internal SDC GitLab collaboration site.

Connecting to the Hadoop Hive Metastore

Launch SQL Workbench by double-clicking on the SQL Workbench shortcut on the desktop:

1. Create a new connection profile by selecting the top left corner icon on the "Select Connection Profile" window.
2. Select "Hive JDBC" from the Driver drop-down.
3. Update URL section with the Hive URL.
4. Provide your username and password received in the welcome email.
5. Click on the Test button at the bottom to validate your connection. A pop-up dialog will appear confirming a successful or failed connection. If you continue running into a failed connection, contact the SDC support desk for assistance.

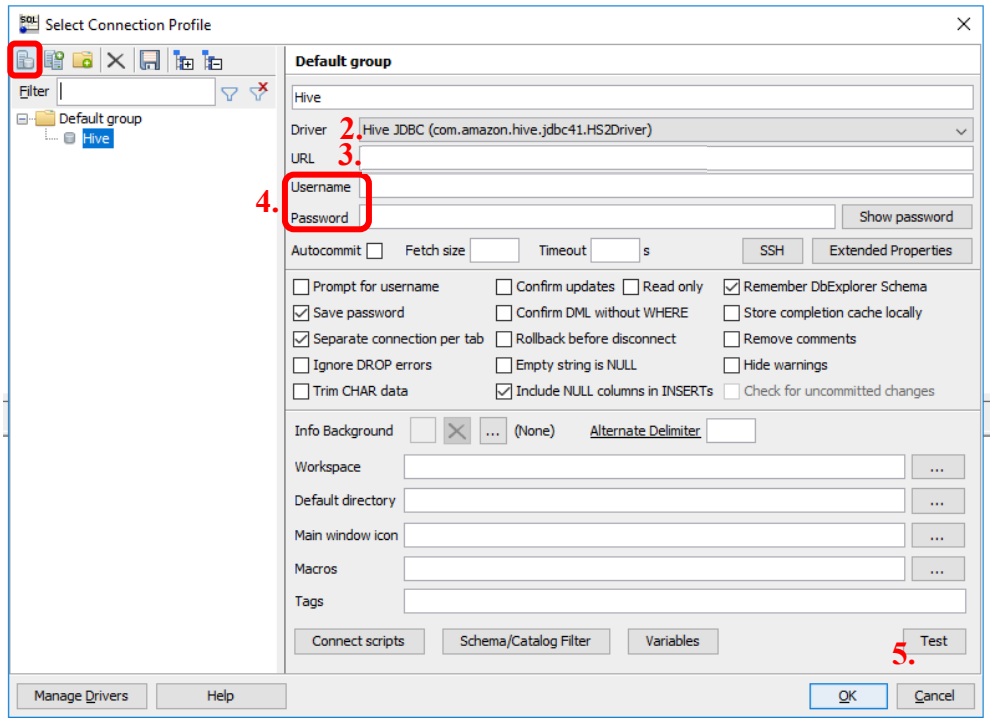
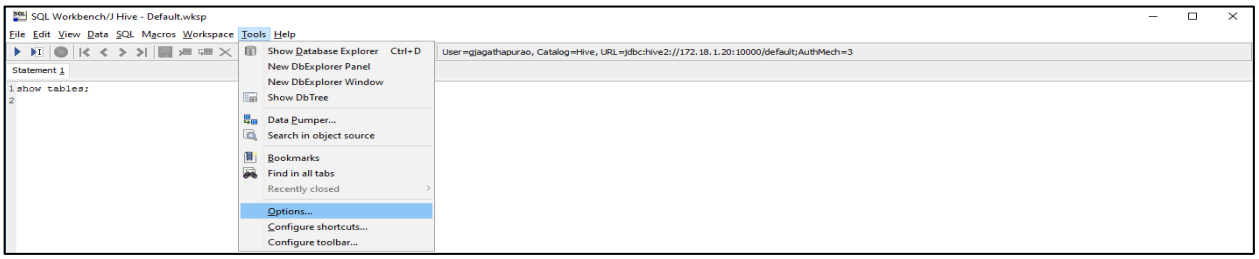


Figure 19: SQL Workbench – Hive Connection Settings

Update Data Formatting Settings in SQL Workbench

Once the connection has been established, navigate to Tools | Options | Data formatting and update the Decimal digits value to 0.



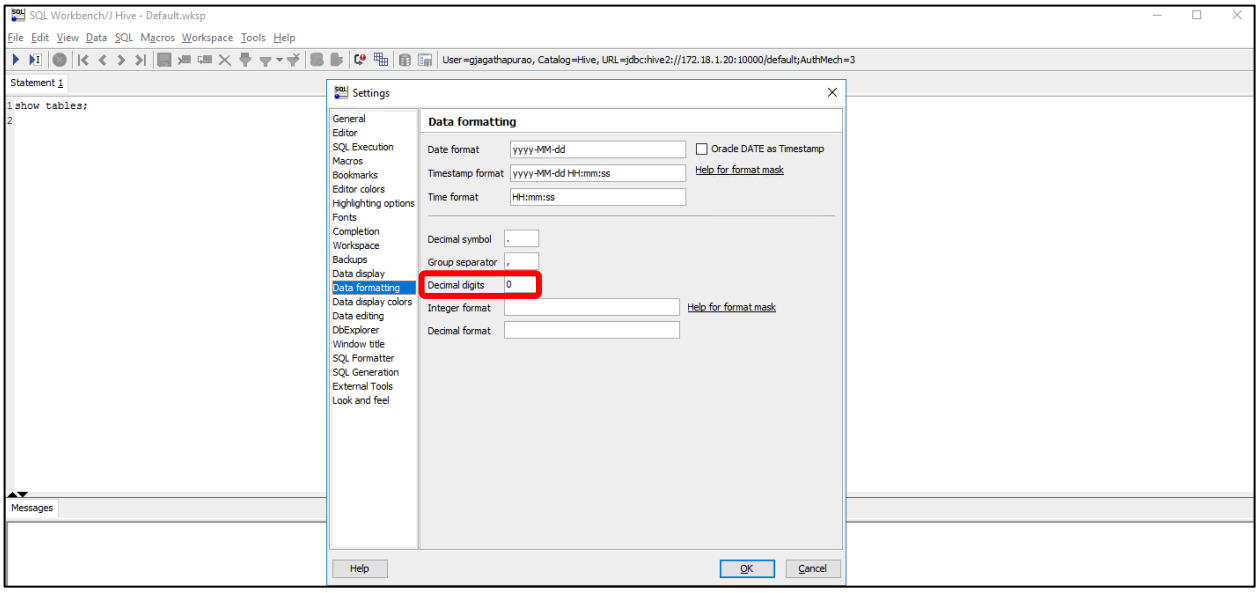


Figure 20: SQL Workbench – Tools → Options → Data formatting

Connecting to the SDC Hadoop Data Warehouse Using Python

Important: The default version of Python installed on the SDC Windows Workstations is v2.7.4. There are two required Python modules that must be installed prior to attempting to connect to Hadoop/Hive with Python using the example code below. To install these modules, open a Windows Command Prompt, and enter the following two commands:

```
C:\Users\username> pip install impyla
C:\Users\username> pip install numpy
```

The above "pip install …" command(s)s only need to be run ONCE on the SDC Windows Workstation. Once the Python modules are installed, they remain available, even across reboots of the workstation.

To test Python connectivity to the data warehouse, open the IDLE python editor and execute:

```
from __future__ import print_function

from impala.dbapi import connect
import numpy

conn = connect(
   host='[host address]',
   port=10000,
   auth_mechanism='PLAIN',
```

```
    user='[your_username]' ,password='[your_password]')

cursor = conn.cursor()

cursor.execute('SHOW TABLES')

result = cursor.fetchall()
result = numpy.array(result)

# print(result)
for r in result:
    print (r)
```

This should result in an array of tables displayed to the user.

Connecting to Redshift from Linux Environments

Credentials to access the Waze Redshift database are communicated from the SDC Administrator (support@securedatacommons.com).

- In R, it is possible to connect to Redshift using multiple packages. The RPostgreSQL package provides a simple method. This package requires the PostgreSQL library to be installed at the system level; if it is not installed, it would be necessary to install as root in the terminal:
  `$ sudo yum install postgresql-devel`
- In R, you may need to `install.packages("RPostgreSQL", dep=T)` if you do not already have the package installed.
- Connect to Redshift using the following code as a guide:

```
library(RPostgres)
# Specify username and password manually, once:
if(Sys.getenv("sdc_waze_username")==""){
  cat("Please enter SDC Waze username and password
manually, in the console, the first time accessing the
Redshift database, using: \n Sys.setenv('sdc_waze_username'
= <see email from SDC Administrator>) \n
Sys.setenv('sdc_waze_password' = <see email from SDC
Administrator>)")
}

redshift_host <- "(details provided by SDC Support to
registered SDC Redshift Users)"
redshift_port <- "5439"
redshift_user <- Sys.getenv("sdc_waze_username")
redshift_password <- Sys.getenv("sdc_waze_password")
redshift_db <- "dot_sdc_redshift_db"
```

```
#drv <- dbDriver("PostgreSQL")
conn <- dbConnect(
  RPostgres::Postgres(),
  host=redshift_host,
  port=redshift_port,
  user=redshift_user,
  password=redshift_password,
  dbname=redshift_db)
```

- A database can then be queried using the `dbGetQuery()` function.

Accessing Jupyter Notebook and RStudio Server

Linux users can access their Jupyter Notebook and RStudio Server using the Firefox web browser through windows workstation using below URLs.

- RStudio – http://<username>-workspace.securedatacommons.internal:8787
- Jupyter Notebook – http://<username>-workspace.securedatacommons.internal:8888

Windows users can click on the "RStudio" 

shortcut icon present on the desktop to open RStudio console.

Manage Workstations

After launching their workstations, users can manage resizing CPU/RAM and scheduling uptime for a workstation by clicking on its Manage button as shown below.

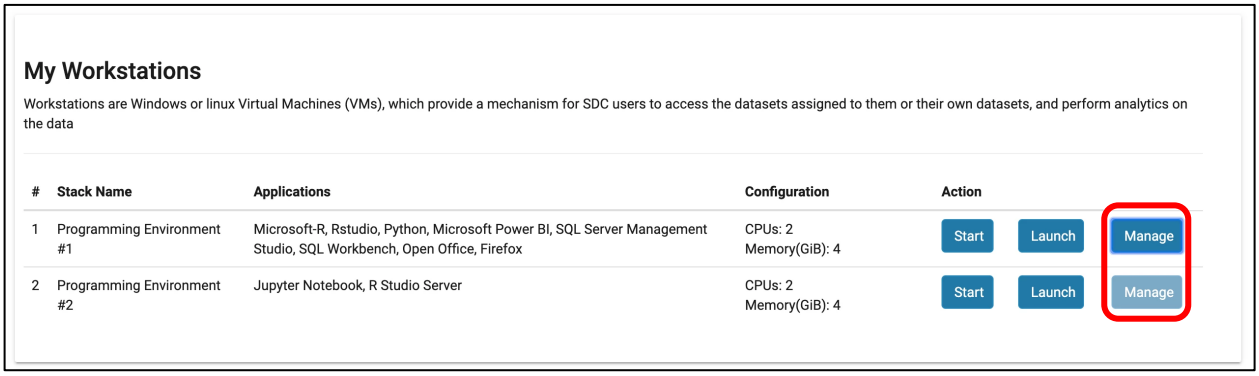


Figure 21: SDC Portal – Manage Workstation

A dialogue window appears with two checkbox options:

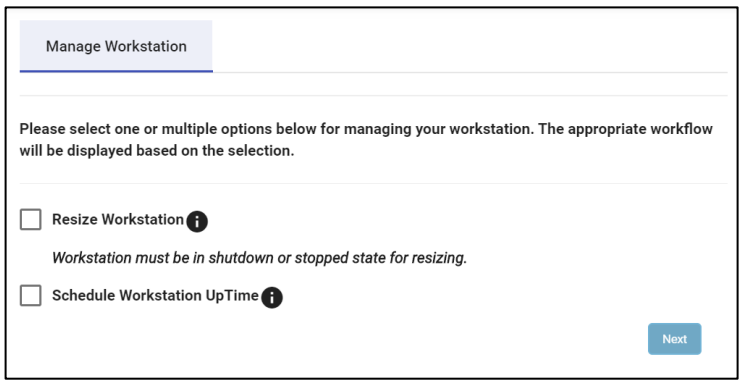


Figure 22: SDC Portal – Manage Workstation Options

Selecting each option renders the appropriate tabs in the dialogue window. The ⓘ icon shown next to each option provides an informational tooltip on their functions.

Resize Workstation

1. To resize the workstation, select the checkbox for Resize Workstation and then Next to continue.

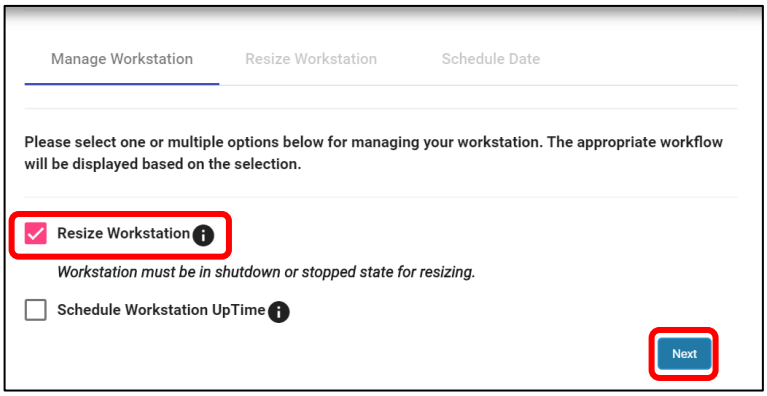


Figure 23: SDC Portal – Resize Workstation Option

2. A message is shown at the bottom of the screen indicating that the workstation will be stopped before applying the resize.

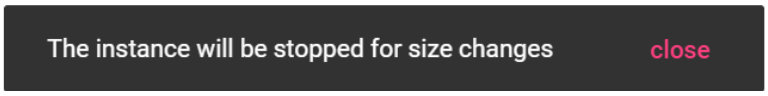


Figure 24: SDC Portal – Workstation Stopped for Resize Changes

3. The Resize Workstation tab allows users to select desired CPU/RAM for their workstation. Current configurations will be grayed out and unavailable. Users can also explore pricing details using the link provided under "click here."
4. Select the "Please start my workstation after resizing to the new configuration" checkbox to automatically start the workstation with the new configuration after saving changes.
5. Select Submit after all details are entered.
6. A Recommended List of instances will appear. Select the desired instance and then the Next button.

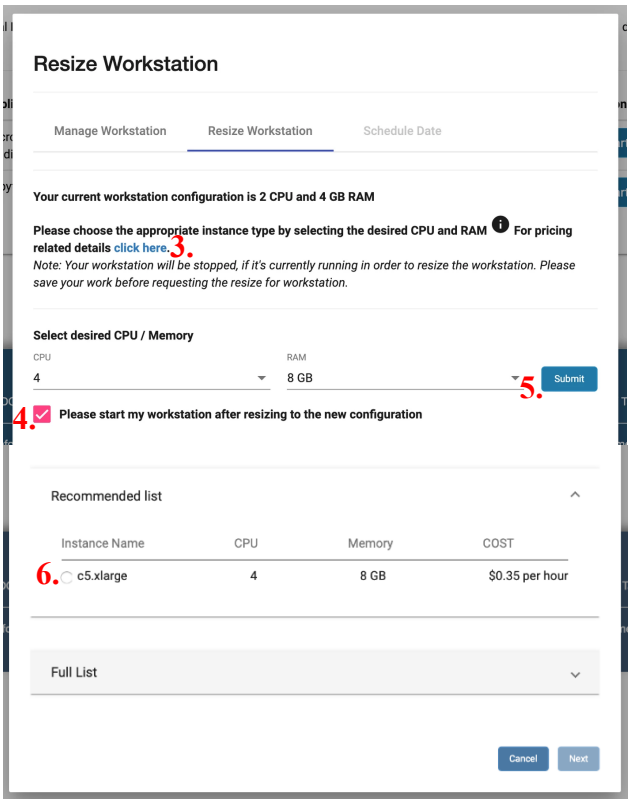


Figure 25: SDC Portal – Resizing Options

7. On the Schedule Date tab, users are prompted to enter a date range for how long the resize should last for the workstation instance. Enter the From and To dates and then select Submit.

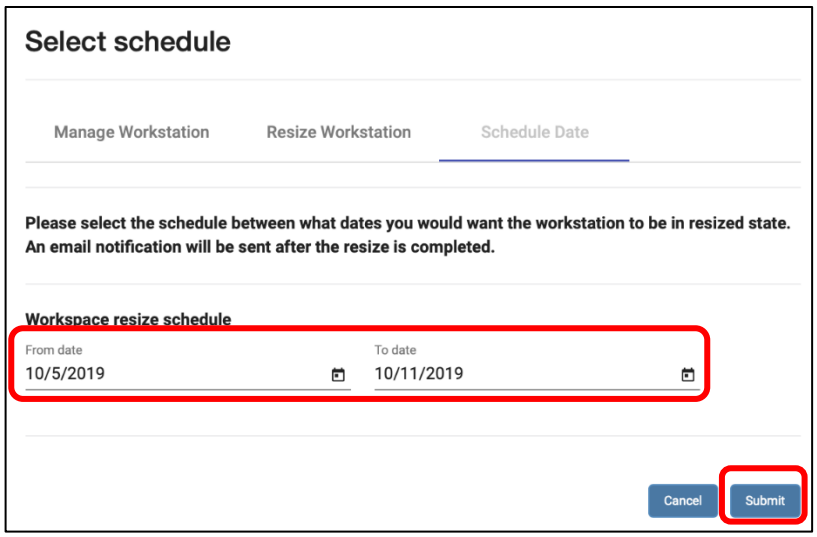


Figure 26: SDC Portal – Schedule Resize

8. Users will be returned to the Workstations tab with updated CPU and memory information. They will also receive a success email message from the system confirming the resize expiration date.

Schedule/Extend Uptime

1. By default, all workstations are shut down at 11 pm EST. If you want to schedule your workstations to be up for a longer period to accommodate analysis runs, select the checkbox for Schedule Workstation Uptime and then Next to continue.

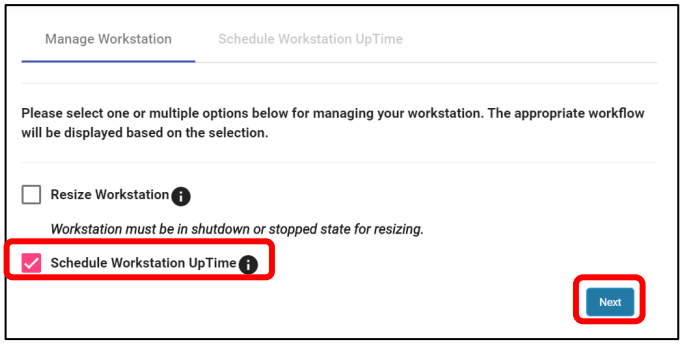


Figure 27: SDC Portal – Schedule Workstation Uptime Option

2. The Schedule Workstation Uptime tab allows users to enter a date range for how long the workstation uptime should last to skip shutdown. Enter the From and To dates and then select Submit.

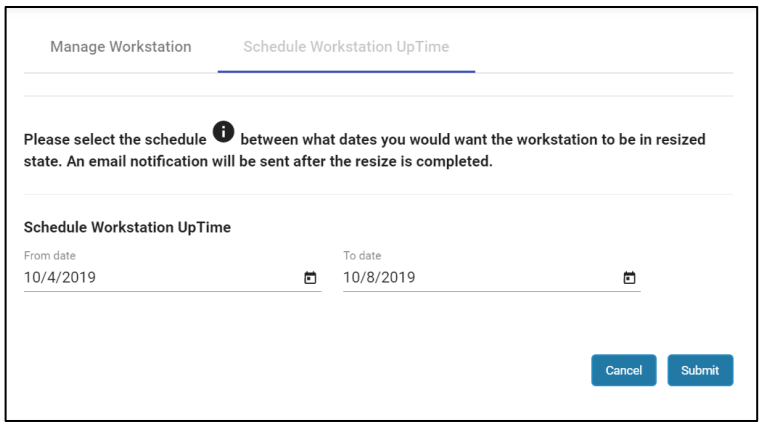


Figure 28: SDC Portal – Schedule Uptime

3. To extend any currently scheduled uptime for the workstation, select the Workstations tab and then select Manage again for the workstation. A new tooltip is now shown for the Schedule Workstation Uptime checkbox on mouse hover that indicates previously scheduled uptime.

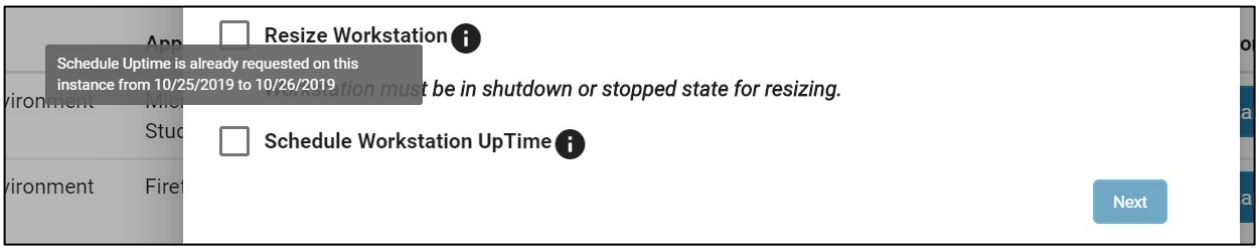


Figure 29: SDC Portal – Tooltip with Existing Scheduled Uptime

4. Repeat steps 1-2. For step 2, the From date will already include the date from the previously scheduled uptime. Add a new To date later in the calendar and then submit the update. The previously scheduled uptime goes inactive while the new one becomes active.
5. After selecting Submit, return to the Workstations tab and then select Manage for the workstation. The tooltip shown on hover for the Schedule Workstation Uptime checkbox now displays the extended uptime.

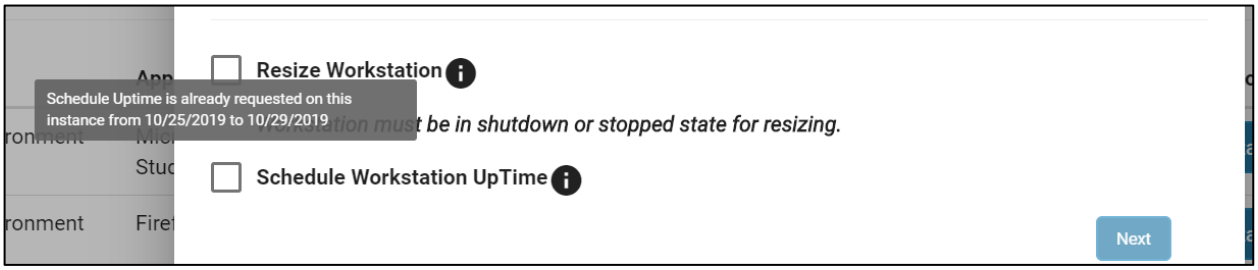


Figure 30: SDC Portal – New Tooltip with Extended Uptime Schedule

Stop Workstations

Users can see the assigned workstations by clicking on the workstations tab on the top right corner of the page. By default, all the workstations are scheduled to stop every day at 11 PM EST. Users can stop the workstations manually by clicking on the Stop button as shown below. A message will appear when the instance is successfully stopped.

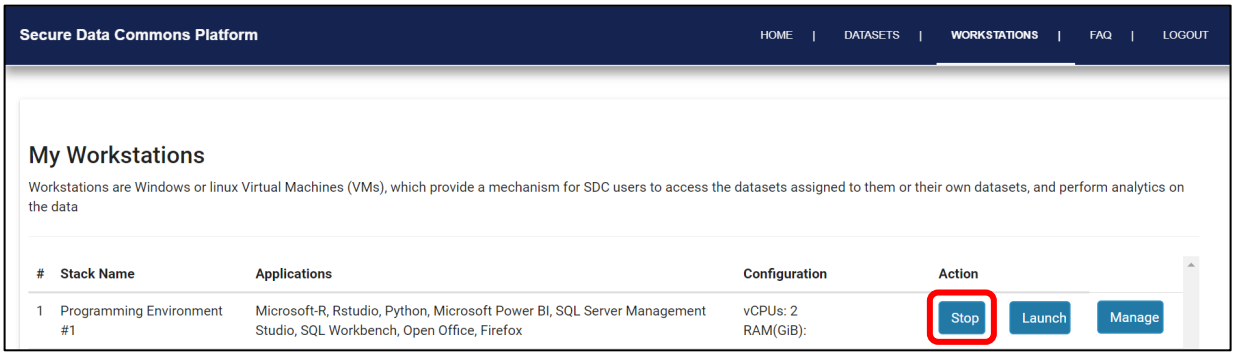


Figure 31: SDC Portal – Stop Workstation

Chapter 4. Exporting Datasets from the SDC

Data Analysts should be able to export the data of the system based on the compliance and data usage policies set forth by a Data Provider.

There are two different types of analysts:

1. **General Analyst:** This type of analyst must provide justification to the Data Provider for each data product that they want to export out of the SDC system. The intent is to ensure that the Data Provider has oversight of the exported data. This type of analyst can also request trusted status from the Data Provider while filling out the approval form.
2. **Trusted Analyst:** This type of analyst already has a trusted status which is provided by the Data Providers. The intent is to reduce the effort for exporting data products of analyses out of the SDC system. A trusted user has a pre-existing and approved relationship with the Data Provider.

Once the Data Analyst completes creating derived datasets, either working on the SDC datasets or combining with other datasets that they import into the system, they can export the derived datasets or share the datasets with other team members.

The following are the steps that the Data Analyst needs to follow to export the data of their analysis from the SDC system to support their research:

1. Each Data Analyst is part of a team bucket which is displayed in the Datasets section. When ready to export, Data Analysts can select the file (or files) that they want to export out of the SDC system and place them in a separate staging folder (i.e., **export_requests)** in their team bucket. Data Analysts can request for exporting a file in this folder by clicking on the export symbol for the file they want to export out of the SDC system. Please note that if Data Analysts want to print out a hard copy of a document, they will need to export it from the SDC workstation to their local machine.

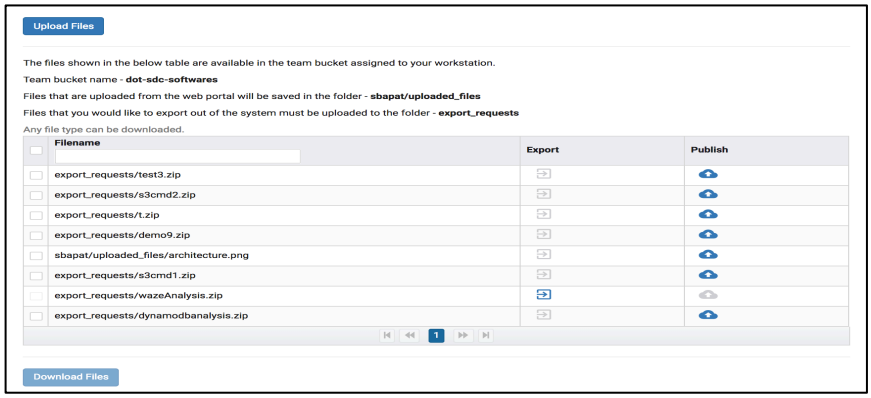


Figure 32: SDC Portal – Request Export

2. Once the export button is selected, a dialog box for requesting the export data will be displayed. The analyst will then need to provide the details of the Project, Data Provider, and Data Type that he has used to create his own dataset and click on the NEXT button once finished.

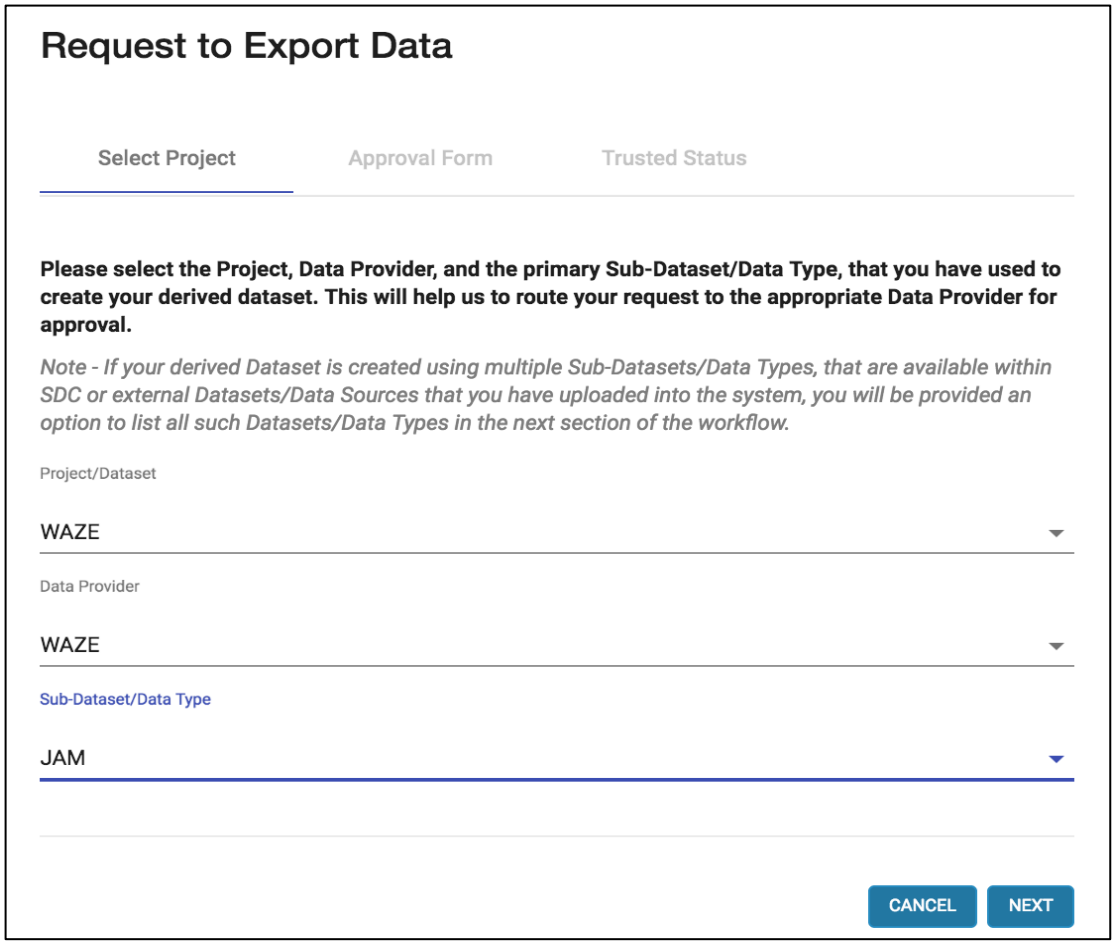


Figure 33: SDC Portal – Request Export Form

3. The additional information regarding the request for exporting the data must be filled out in the approval form below. These details are shared with the Data Providers, which helps them to accept or reject the request made by the Data Analysts.

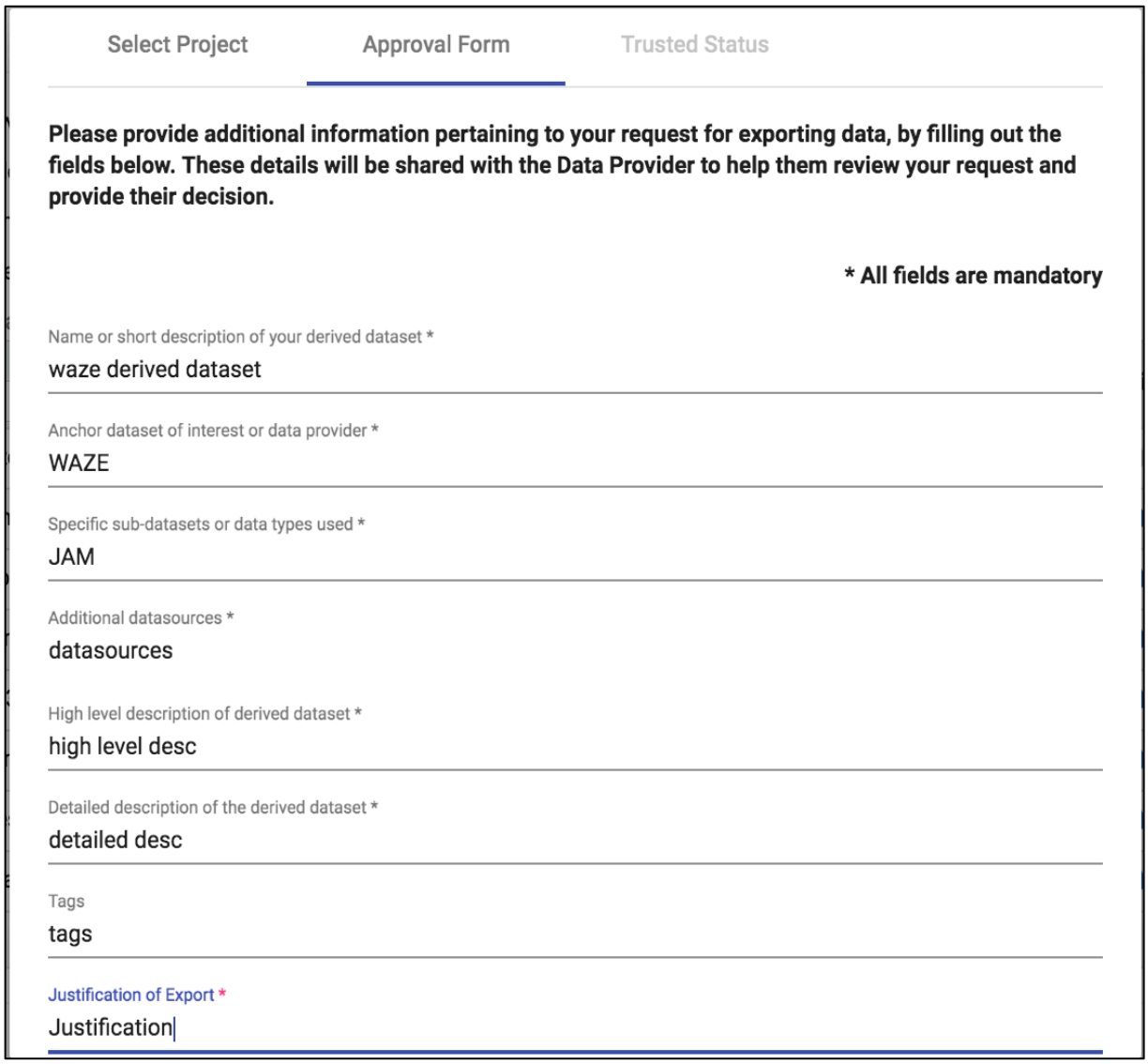


Figure 34: SDC Portal – Approval Form Fields

4. If the user is not a trusted user, he/she will be prompted with the option for requesting the trusted status from the Data Provider. This will allow the analyst to export the data immediately, as opposed to waiting for review and approval from the Data Provider. The user must accept the Acceptable Usage policy for the request to go through to the Data Provider. The form will not be submitted if the user declines.

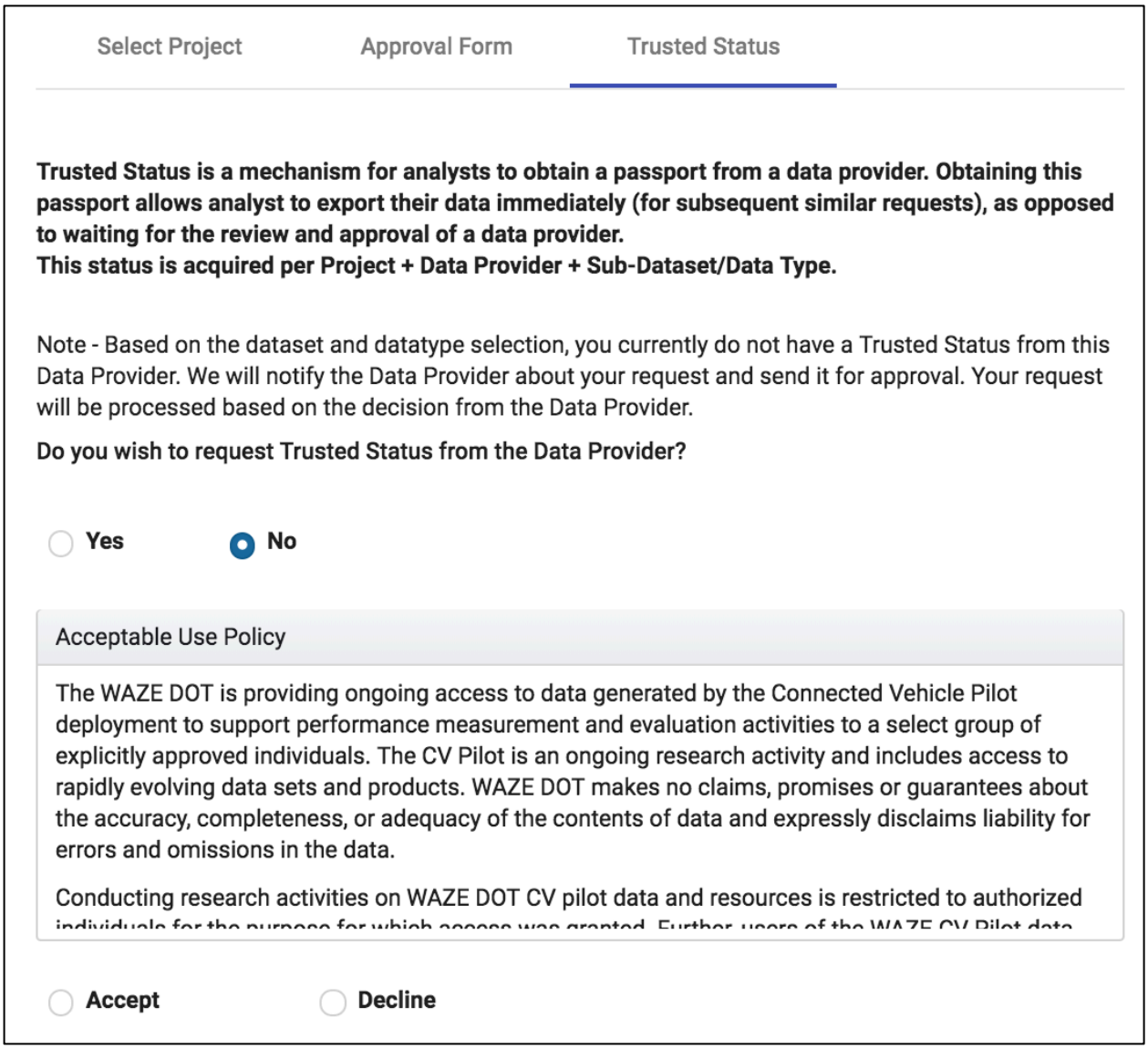


Figure 35: SDC Portal – Acceptable Use Policy

5. If the user is a trusted user, he/she will be prompted with the option for requesting the auto-export approved status from the Data Provider. This will allow the analyst to be able to export data automatically, as opposed to filling out the request form and then going through the approval process before exporting. The analyst enters the derived dataset name of the data type they want to export into 'Specify the derived dataset name' and then the justification for requesting permissions into 'Reason for requesting permission.' The form will still be submitted if the user declines the option to submit a request for auto-export approved status by selecting 'No' under 'Do you wish to request Auto-Export Status from the Data Provider?' This form will be unavailable if the user does not have trusted status yet.

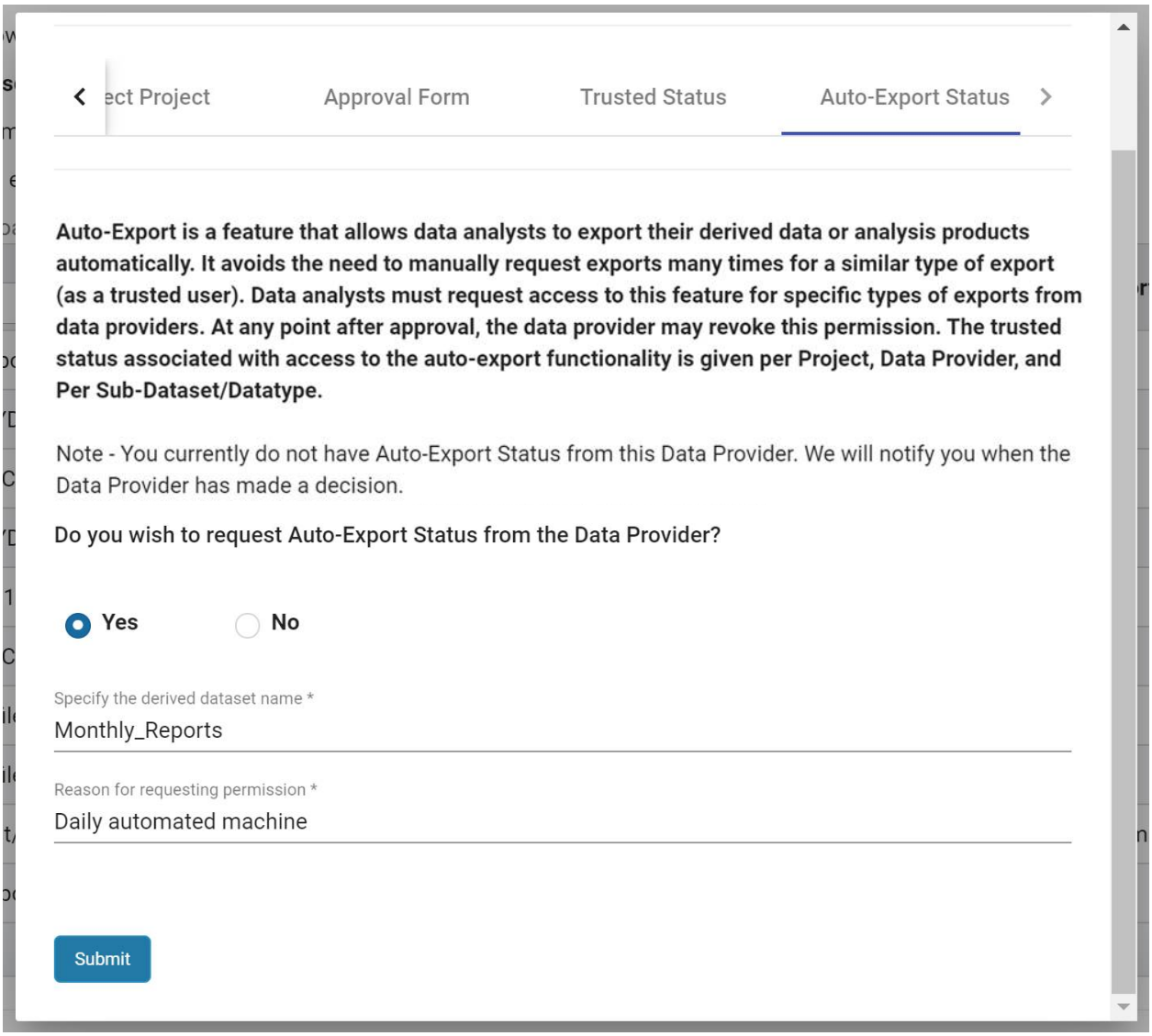


Figure 36: SDC Portal – Auto-Export Status Request Form

6. Upon successful submission, the request will be sent to appropriate Data Providers. Data Providers will be responsible for accepting or rejecting the export requests.

7. Once Data Providers approve the request, Data Analysts will be able to download the dataset out of SDC through portal.
8. Log into the SDC Portal to start and launch the Guacamole workstation. Refer to the instructions in Chapter 3 Accessing and Launching Workstations if needed.
9. Log into the Guacamole workstation.
10. Perform work and save any derived data that needs to be exported to the local machine. For example, the 'Documents' folder would be a good place.
11. Open PowerShell and navigate to the directory that the derived data is located at. **NOTE:** If you are unable to find PowerShell, right-click on the Windows menu, select Search, and enter "powershell."
    a. Use the command `cd` to change the directory, for example, `cd Documents`
12. Copy the file into your team bucket auto_export directory's correct sub-directory that you were approved to auto-export. For example:
    a. If you are a member of your team approved to auto-export 'Monthly_Reports,' then you would copy the file to your team bucket's 'auto_export/Monthly_Reports/' directory
    b. The following command would be run in the PowerShell window:
    ```
    aws s3 cp derivedatatoexport.file s3://team-bucket/auto_export/Monthly_Reports/
    ```
        i. See the AWS S3 CLI Commands section for more information on the `aws s3 cp` command.
13. After the copy has completed, the SDC will run processes to auto-export the derived data to your team bucket's auto-export S3 bucket. This process should only take a few seconds.
14. To see if your derived data has been exported, sign into the AWS CLI on your local machine using the SDC-provided authentication scripts. Make sure to set the role as your auto-export role for the specific derived datatype you are exporting. You will ONLY be able to export the data of the role you have assigned. You must change your role to export other types of derived data if you are approved for multiple datatypes.
    a. If you need access to these authentication scripts, please contact the SDC support desk.

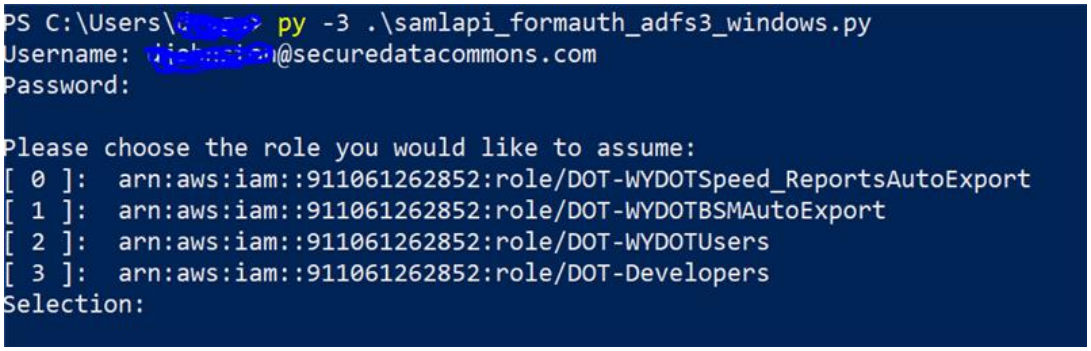


Figure 37: AWS CLI – Set Auto-Export Role

15. Using the AWS CLI, you may list the content of your derived datatype's directory from the team auto-export bucket and download any files located in that directory to your local machine.
    a. Your permissions are very limited, and the commands must be very precise for them to work.
    b. List files: `aws s3 ls s3://team-auto-export-bucket/datatype/`
    c. Download file: `aws s3 cp s3://team-auto-export-bucket/datatype/derivedata.file localmachinelocation`

Chapter 5. Setting Up SDC with GitLab Repositories

The following sections provide instructions for SDC users on setting up GitLab repositories for committing and sharing their code and any other documents.

GitLab provides an organizational structure of users and repositories managed respectively via "groups" and "projects" to promote collaboration among teams of SDC users. Groups contain members with different access levels to their project repositories based on the assigned roles:

Table 1: GitLab Access Rights by Role

Role	Read-Only Access	Read/Write Access
Guest	✓	
Reporter	✓	
Developer		✓
Maintainer		✓

GitLab Features:

- Share code among a subset of your team, your entire team, or any number of other teams on the SDC and provide comments to improve each other's code
- Simplified collaboration with shared read/write access
- Support for multiple repositories to be kept in a logical structure that separates them from each other and groups by organization.
- Version control and source code management including:
    - File locking to help prevent conflicts
    - Users can work from their local copy of the code
    - Users can branch code to make changes and then quickly merge it after approval

Log In to GitLab

1. Open a web browser, then navigate to http://scm.securedatacommons.internal/, which is the URL of the SDC internal GitLab implementation.

2. In the Sign in page, the LDAP login option is selected by default. Enter your SDC credentials as the LDAP username and password, then click Sign in (in the LDAP Username field, do not enter the "@securedatacommons.com" portion of your SDC username).

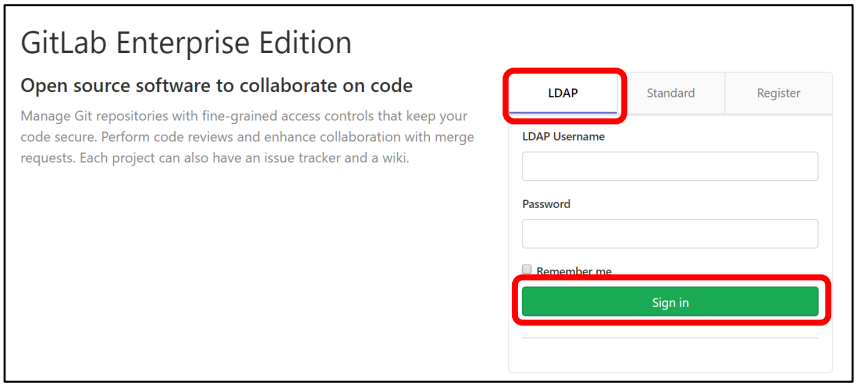


Figure 38: GitLab Login Prompt - LDAP

3. The Welcome to GitLab web page is displayed.

Create a Group

1. In the Welcome to GitLab web page, select Create a group.

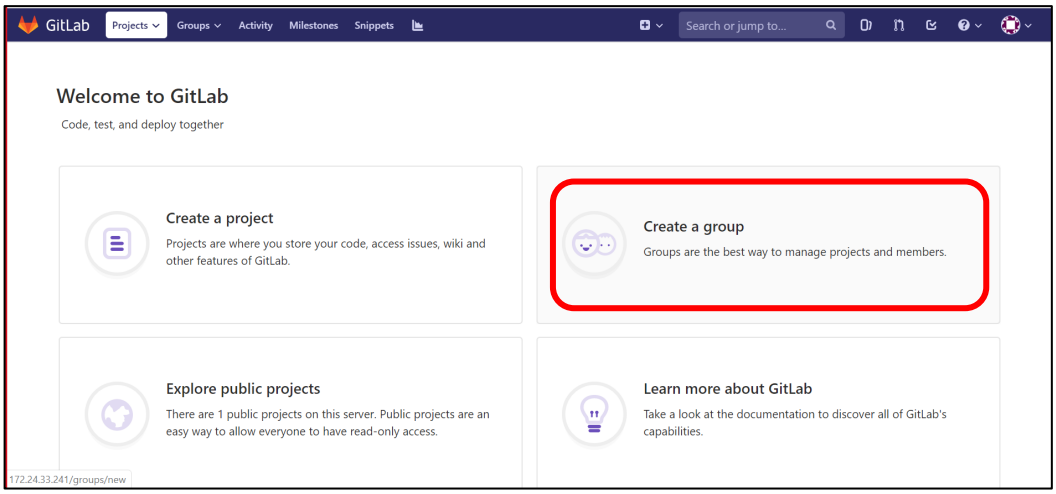


Figure 39: GitLab – Create a Group Option

2. The New group form appears. Enter the name for the new group.
3. Select the Private option under Visibility level to only allow members in the group to see the group and its projects.
4. Select Create group.

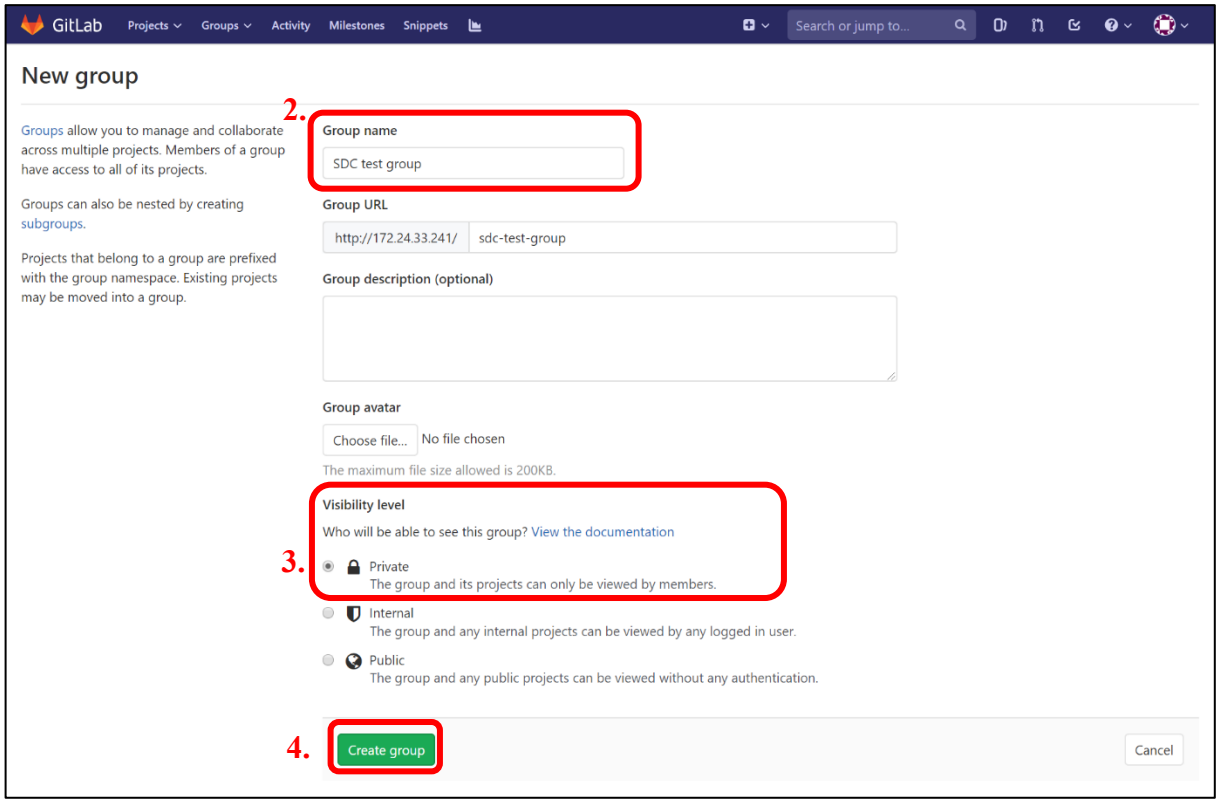


Figure 40: GitLab – New Group Form

5. A success message appears indicating that the group was created. You are now ready to add members to the group.

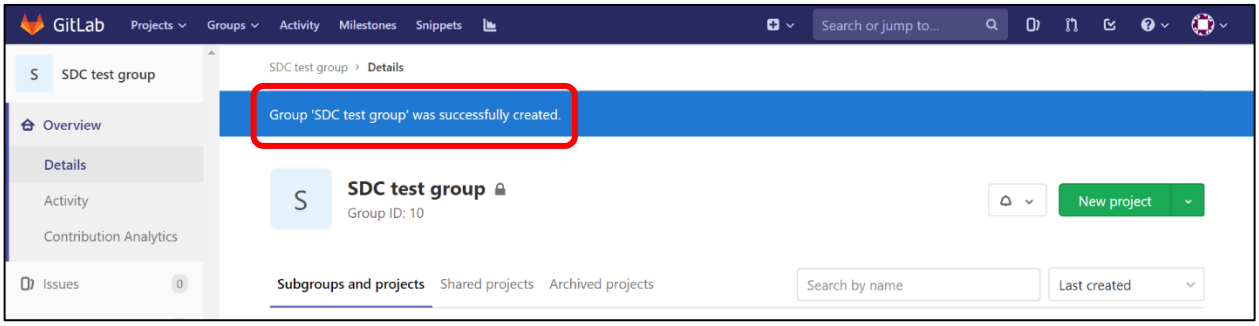


Figure 41: GitLab – Group Creation Success

Add Group Members

Users assigned with the “Developer” or “Maintainer” role can add other GitLab members in the SDC to their groups.

NOTE: SDC users are listed as a GitLab member in the Members panel only after they have logged into GitLab. If you do not see your SDC colleague listed, please ask them to log in to GitLab.

1. In the newly created project web page, click Members from the left navigation menu.

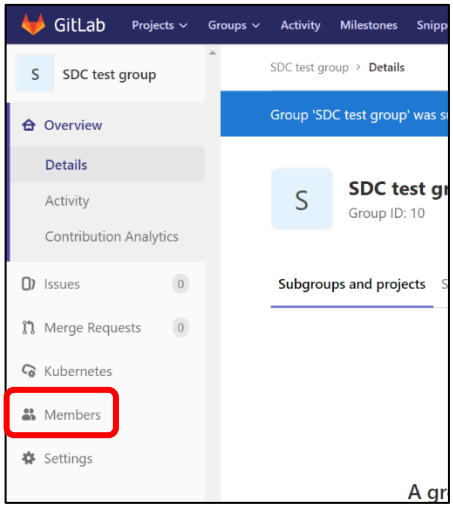


Figure 42: GitLab – Members Left Navigation Menu

2. In the "Search for members" text box, start entering the name of the SDC user you want to add; then select the member's name from the autocomplete list.
3. Select a role from the next drop-down menu. You can select the "Read more" link to view a comprehensive list of the different available permission levels and their access rights. A summary of the roles' access types is also shown in Table 1 at the beginning of this chapter.
4. Select Add to group.

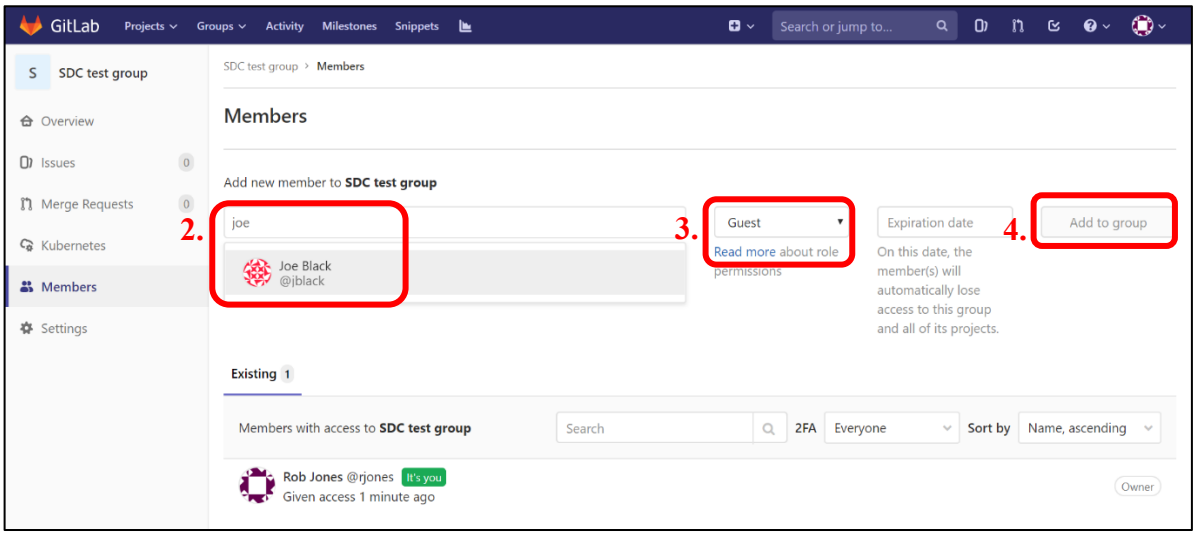


Figure 43: GitLab – Add a New Group Member

Create a Project

1. Select the GitLab icon on the upper left to return to the home page.
2. Select Create a project.

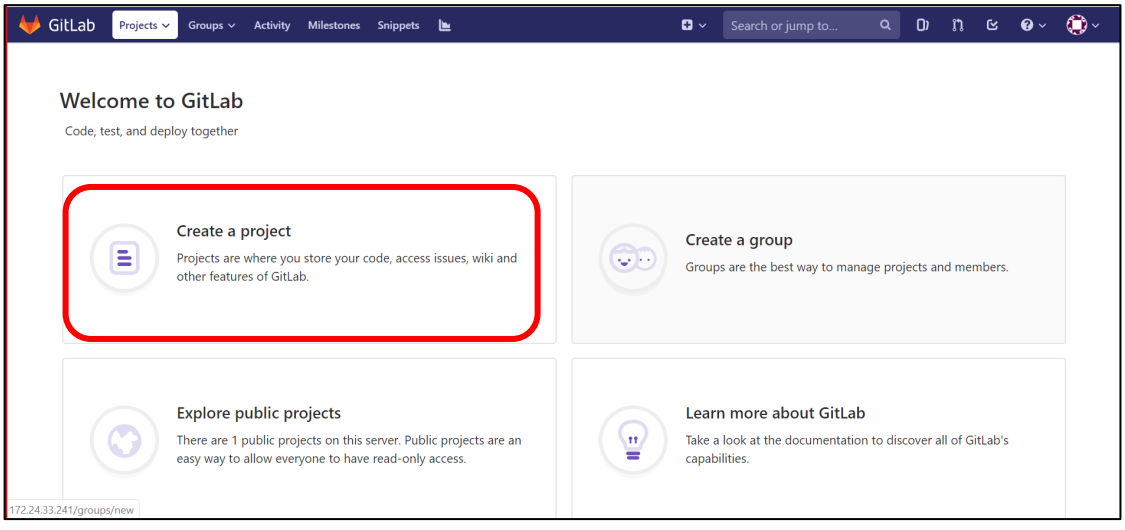


Figure 44: GitLab – Create a Project

3. The New project form appears. Enter a name for the project.
4. Select the Private option under Visibility level, so that the project is only visible to users who are granted explicit access to the project.
5. Select the Initialize repository with README checkbox to automatically generate a readme file that contains the project's repository details.
6. Select Create project.

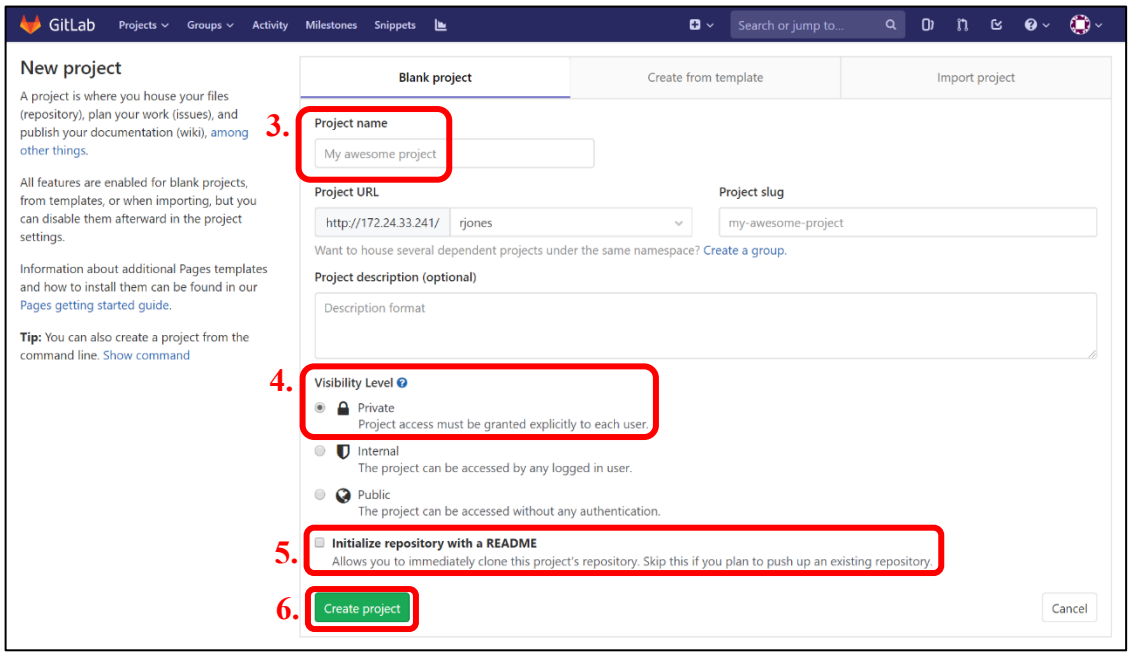


Figure 45: GitLab – New Project Form

7. A success message appears indicating that the project repository was created. You are now ready to add members to the project.

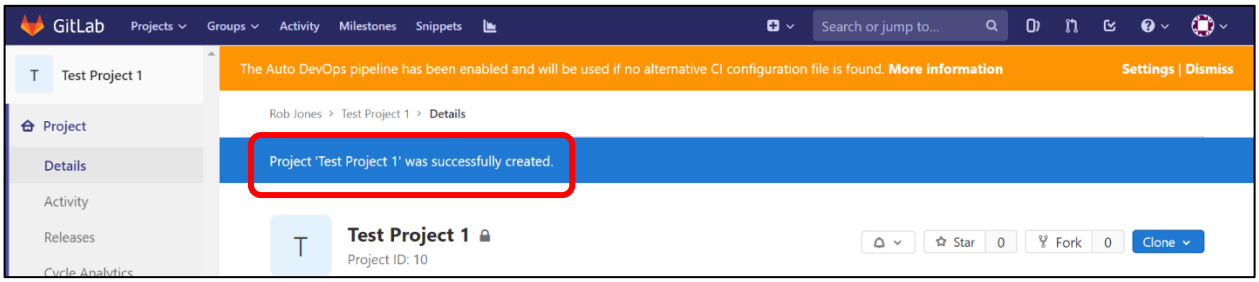


Figure 46: GitLab – Project Creation Success

Add Members or Groups to a Project

1. Select the GitLab icon on the upper left to return to the home page.

2. A list of your projects appears. Select the name of the project you want to add members /groups to.

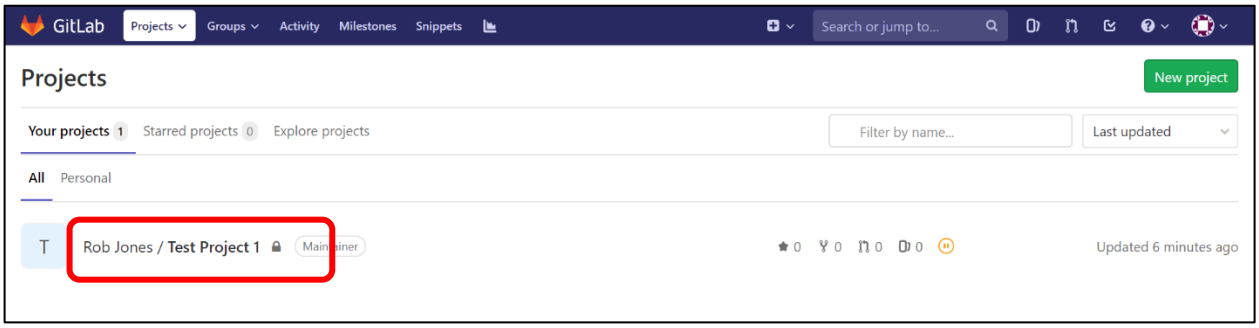


Figure 47: GitLab – Select Project to Add Members

3. Click on Settings and then Members from the left navigation menu.

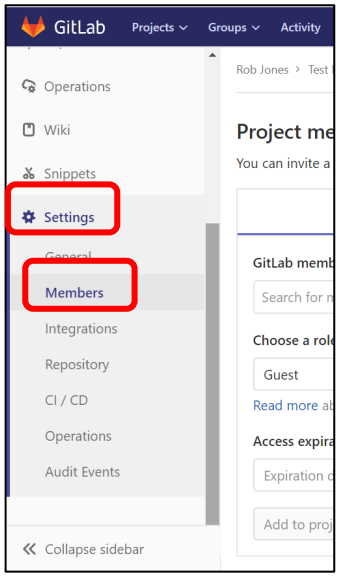


Figure 48: GitLab – Settings, Members Left Navigation Menu

4. You can add individual users or entire groups to the project.
    a. Select the Invite member tab to add members individually to the project.
        i. Start entering the project member's name or email address into GitLab member or Email address to select from the autocomplete list.
        ii. Select a role for the project member from the drop-down menu under Choose a role permission.
        iii. Click Add to project.

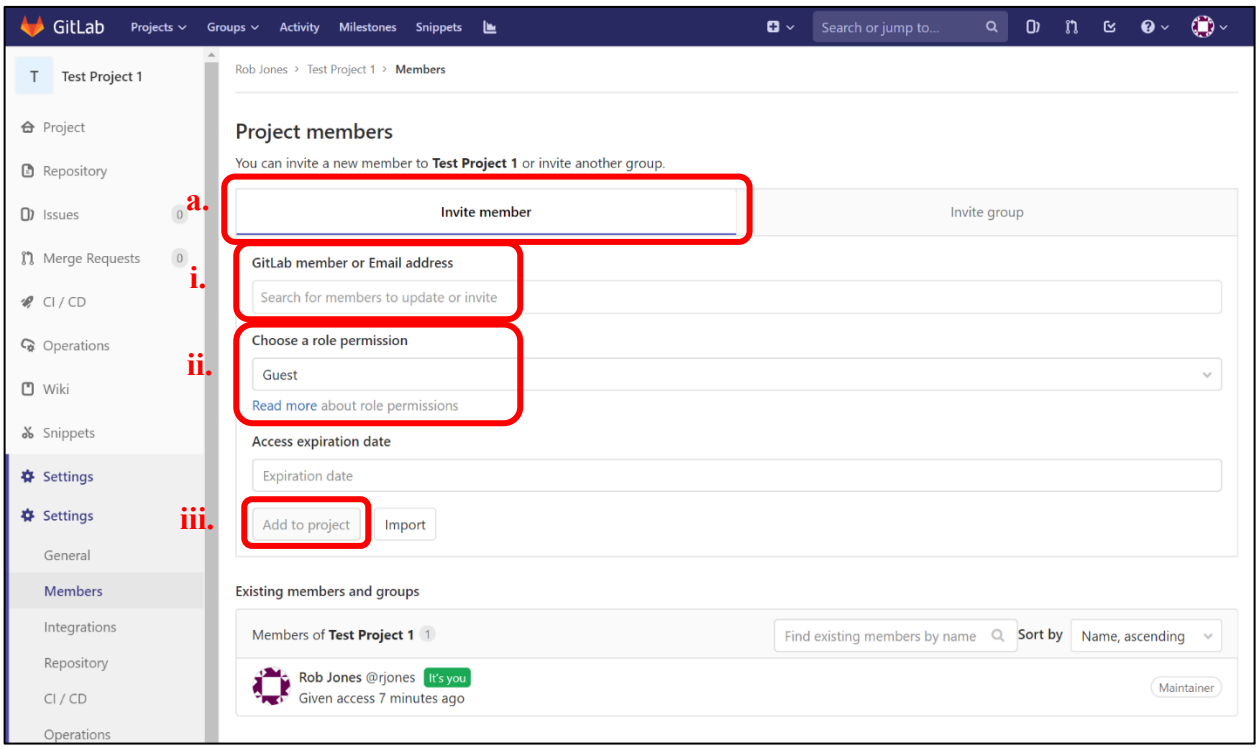


Figure 49: GitLab – Add Member to Project

b. Select the Invite group tab to add an entire group to the project.
    i. Start entering the group's name into Select a group to invite to select from the autocomplete list.
    ii. Select a role to apply for the group from the drop-down menu under Max access level.
    iii. Select Invite to grant the group access to the project repository.

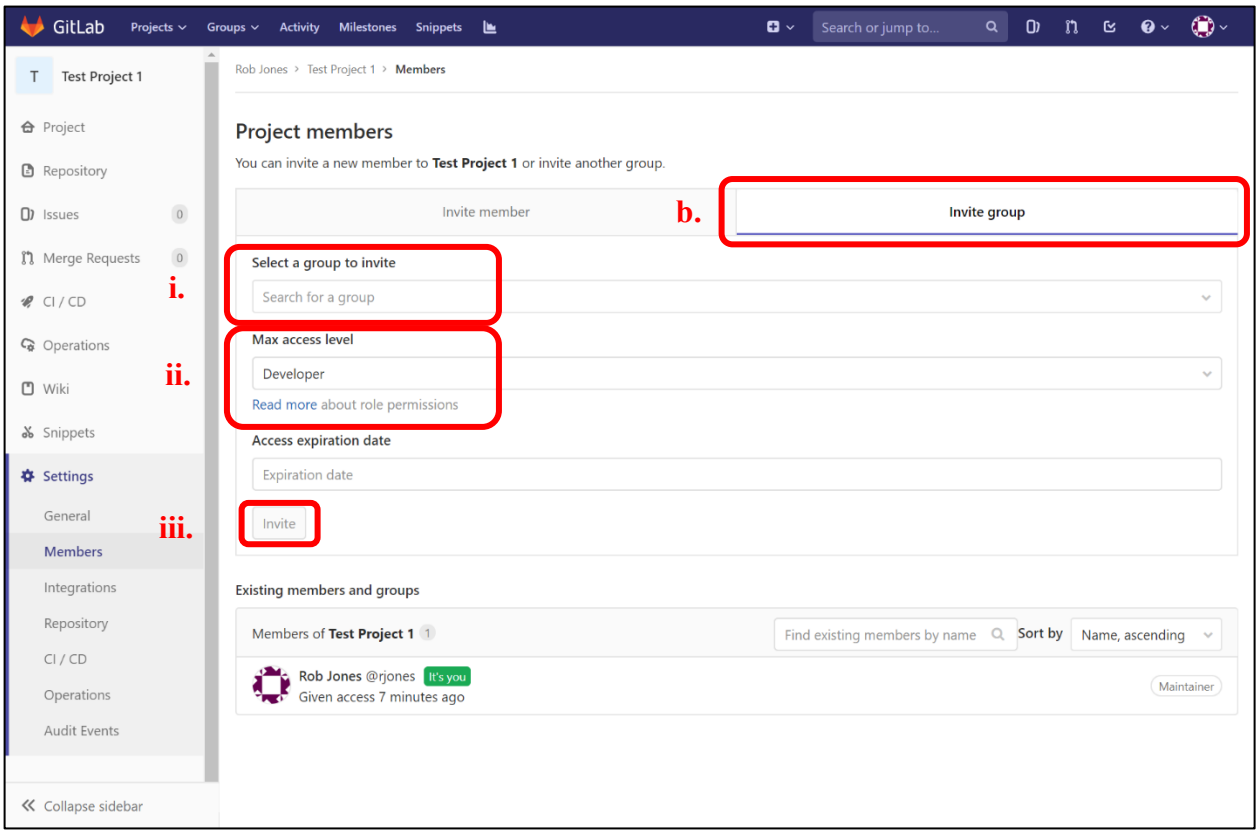


Figure 50: GitLab – Add Group to Project

Generate SSH Key

To set up git integration, you need to have an SSH key on your SDC workstation. This key authenticates your identification to GitLab so that you can push commits to the project repository.

NOTE: For Windows users, perform the following commands in Git Bash. If Git Bash is not installed on your Windows workstation, please copy the Git Bash installer, **Git-2.19.1-64-bit.exe**, from the Z:\software-distribution folder and then install it.

1. To store your authentication credentials for git, use a hidden folder called ".ssh." First, check to see if the folder exists with `ls -a`. If not, create it using the command `mkdir .ssh`:

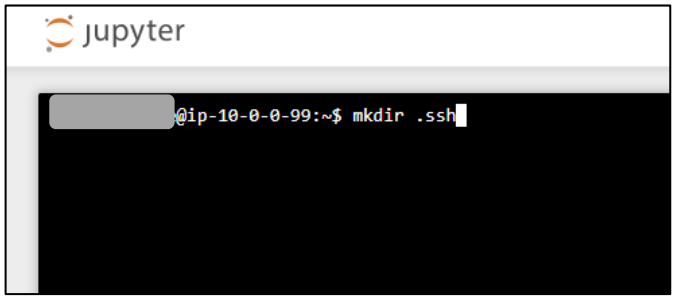


Figure 51: GitLab – Create Folder for Credentials

2. Confirm this works by entering `ls -a` to see all objects in the current directory; you should see the .ssh listed:

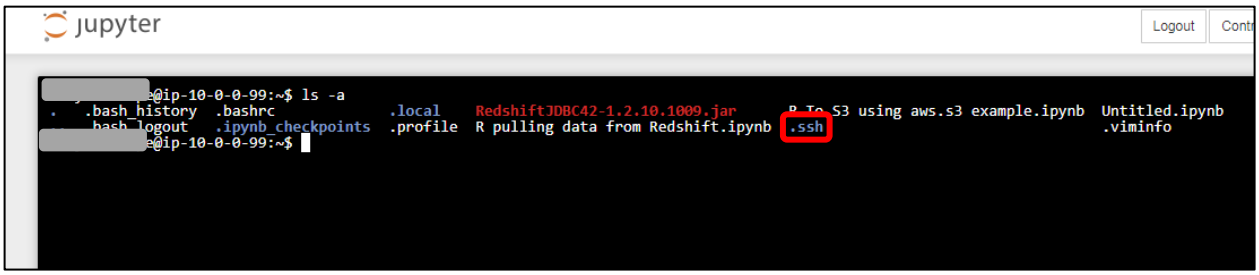


Figure 52: GitLab – View Directory Objects

3. Generate a new key to store in this location by typing the following command and then hitting the Enter key:

```
ssh-keygen -t rsa -b 4096 -C "SDCusername@securedatacommons.com"
```

4. It will output the following message: "*Generating public/private rsa key pair.*"
5. It will ask you where to save the file and to enter a passphrase with the following three prompts. Hit the Enter key each time (no need to enter anything else) to accept the defaults.
    - "*Enter file in which to save the key:*"
    - "*Enter passphrase (empty for no passphrase):*"
    - "*Enter same passphrase again:*"

6. After the following confirmation messages, the new SSH key is now stored in the .ssh folder.
    - "*Your identification has been saved in /c/Users/SDCusername/.ssh/id_rsa.*"
    - "*Your public key has been saved in /c/Users/SDCusername/.ssh/id_rsa.pub.*"

The default key file name is **id_rsa** and is saved under the .ssh folder (e.g., C:/Users/[SDCusername]/.ssh/ in Windows; $HOME/.ssh/id_rsa in Linux)

7. You now need to provide the public part of this key to GitLab by copying the ".pub" part of the key. Here are the different options for copying the public SSH key.

   Option 1:
   On Windows, open the id_rsa.pub file (under the C:/Users/[SDCusername]/.ssh folder) in a text editor (Notepad) and copy the entire key by using Select All then Copy. Be careful not to accidentally change anything!

   Option 2:
   Copy your public SSH key to the clipboard by using one of the commands below depending on your Operating System:

   Git Bash on Windows:

   ```
   cat ~/.ssh/id_rsa.pub | clip
   ```

   WSL/GNU/Linux (requires the xclip package):

   ```
   xclip -sel clip < ~/.ssh/id_rsa.pub
   ```

   If using Jupyter, use the `head` command. Right-click on the output and click **copy**:

   ```
   head ~/.ssh/id_rsa.pub
   ```

8. From the GitLab home page, add your public SSH key to your GitLab account:
   a. In the upper-right corner, click on your avatar image and then click Settings.
   b. From the left navigation menu, click SSH Keys.
   c. In the Key field, paste your public SSH key.
   d. The Title field will be automatically populated with SDCusername@securedatacommons.com.
   e. Click the Add key button.
9. To test whether your SSH key was added correctly, run the following command in your terminal:

   ```
   ssh -T git@scm.securedatacommons.internal
   ```

10. The first time you connect to GitLab via SSH, you will be asked to verify the authenticity of the GitLab host you are connecting to. Type "yes" to add the SDC internal GitLab to the list of trusted hosts.
11. You should see the *"Welcome to GitLab, `@username`!"* message.

NOTE: Once added to the list of known hosts, you will not be asked to validate the authenticity of SDC internal GitLab's host again. Run the above command once more, and you should only receive a "*Welcome to GitLab, `@username`!*" message.

If the Welcome message does not appear, run SSH in verbose mode by replacing `-T` with `-vvvT` to debug the error:

```
ssh -vvvT git@scm.securedatacommons.internal
```

Chapter 6. Technical Documentation and Contact Information

The following sections provide technical resources for SDC users.

Architecture Diagram

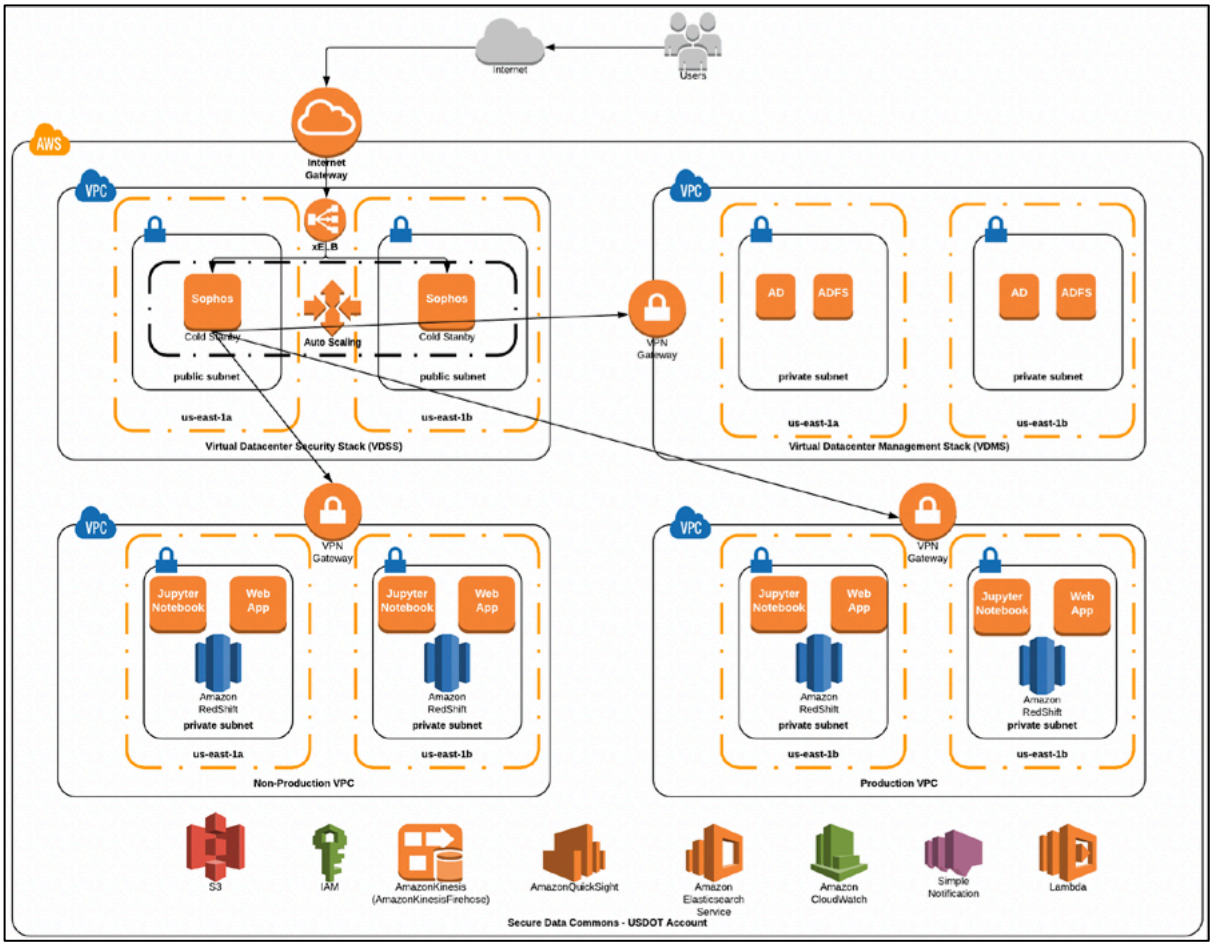


Figure 53: SDC Architecture Overview

Workstation Details

Table 2: Default Workstation Details

Workstation Type	Type
Linux Workstation	t2.medium
Windows Workstation	t2.medium

Note: Workstation size and type can be increased upon user request. Please refer to the Resize Workstation section under Chapter 3 for instructions.

Tools and Versions

Table 3: List of Tools Used and Their Versions

Tool Name	Version	Workstation
Java	1.8.0_162	Linux, Windows
Python	2.7.14	Linux, Windows
SQLWorkbench/J	Build 125	Windows
R	3.4.3	Linux, Windows
RStudio	1.1.423.0	Linux, Windows
Libre Office	5.3.6.1	Windows
Visual Studio	1.20.1.0	Windows
AWS CLI	1.14.46	Windows
7Zip	18.01	Windows
PuTTY	0.70	Windows
Firefox	59.0.2.0	Windows

Contact Information

The SDC support team can be reached at support@securedatacommons.com.

Useful Links

S3: Amazon Simple Storage Service, a place to store data.

Jupyter: An interactive, browser-based programming environment, mostly used for Python scripts but can also run R or other languages and can weave formatted text in with code and results of code into one 'notebook' file.

Redshift: A database system, which can be queried with SQL.

Hive: The Apache Hive™ data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage and queried using SQL syntax.

AWS S3 CLI Commands

The following are a list of helpful commands to work with S3 from the terminal. The AWS_S3_CLI_Cheat_Sheet.pdf file is also available on the desktop of all workstations with useful commands.

In these commands, 'local' refers to the user's SDC workstation (EC2 instance). The listed commands can only be used from within the user's workstation environment.

- **List objects in a bucket -** If there are any files at the bucket level, then this command will return the list of files. If there are only folders/prefixes under the bucket, then it will return the top level folder/prefix names of that bucket.
  `aws s3 ls s3://<bucketName>`

- **List objects under a folder/prefix -** This command will list all the objects/files under that folder or prefix.
  `aws s3 ls s3://<bucketName>/<prefix>/`

- **Copy a local file to S3 -** This command will copy a local file (test.txt) from your workstation over to an S3 bucket:
  `aws s3 cp test.txt s3://<bucketName>/test.txt`

- **Copy an S3 object to a local file -** The below command copies an S3 object to your workstation.
  `aws s3 cp s3://<bucketName>/test.txt test.txt`

Chapter 7. Frequently Asked Questions

How can I get access to the SDC Datasets?

These are datasets that are available within SDC platform that are published / enabled by the SDC team or published by other users. Access to these datasets are available upon request. Once you are logged in, go to 'Datasets' in the top menu.

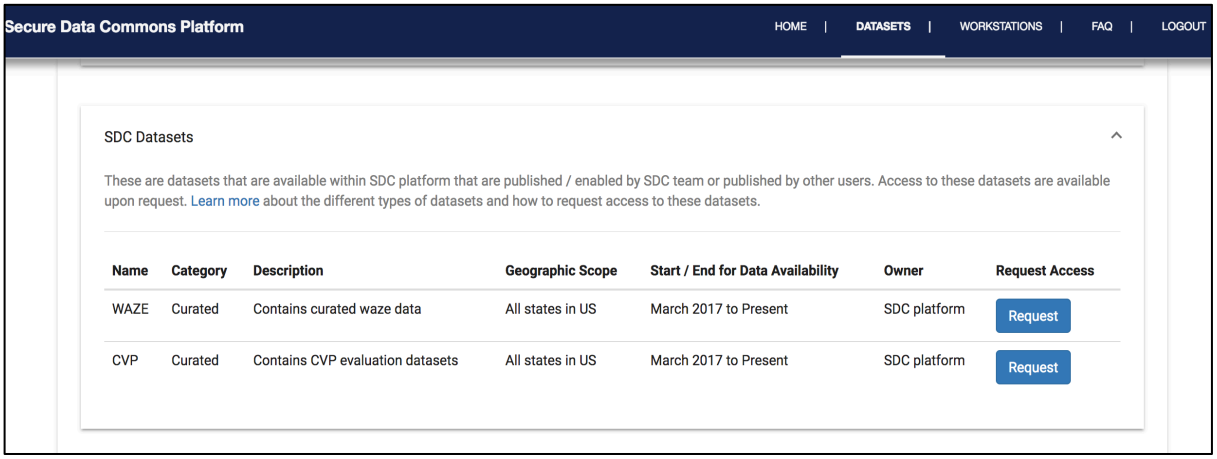


Figure 54: SDC Portal – Datasets Tab

All the available datasets are listed under the ‘SDC Datasets’.
To request access to a dataset, click on the ‘Request’ button.
A form will pop up. Fill out the form and click on ‘Send Request’ button.

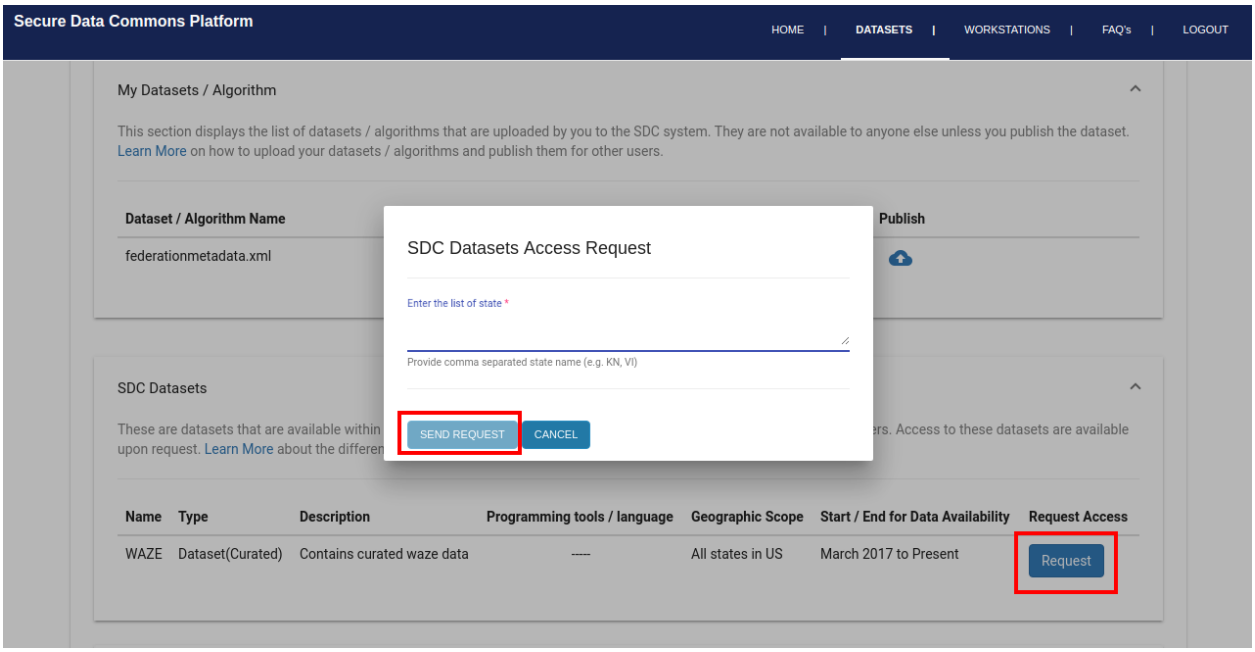


Figure 55: SDC Portal – Dataset Access Request

The request will be sent to the SDC support team and access to the requested dataset will be given upon approval.

How will I understand what a particular dataset consists of?

Click on Name of Dataset, you can see README of that particular dataset below it.

How can I launch a workstation?

Click on ‘Workstations’ and click on the ‘Launch’ button of any workstation you want to access.

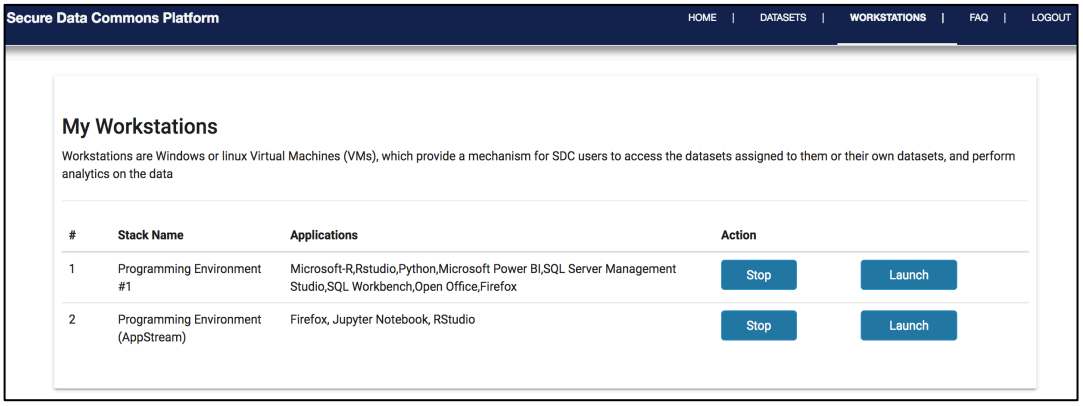


Figure 56: SDC Portal – Workstations Tab

For Programming Environment #1, you will be prompted with username and password to log in to the Windows workstation.

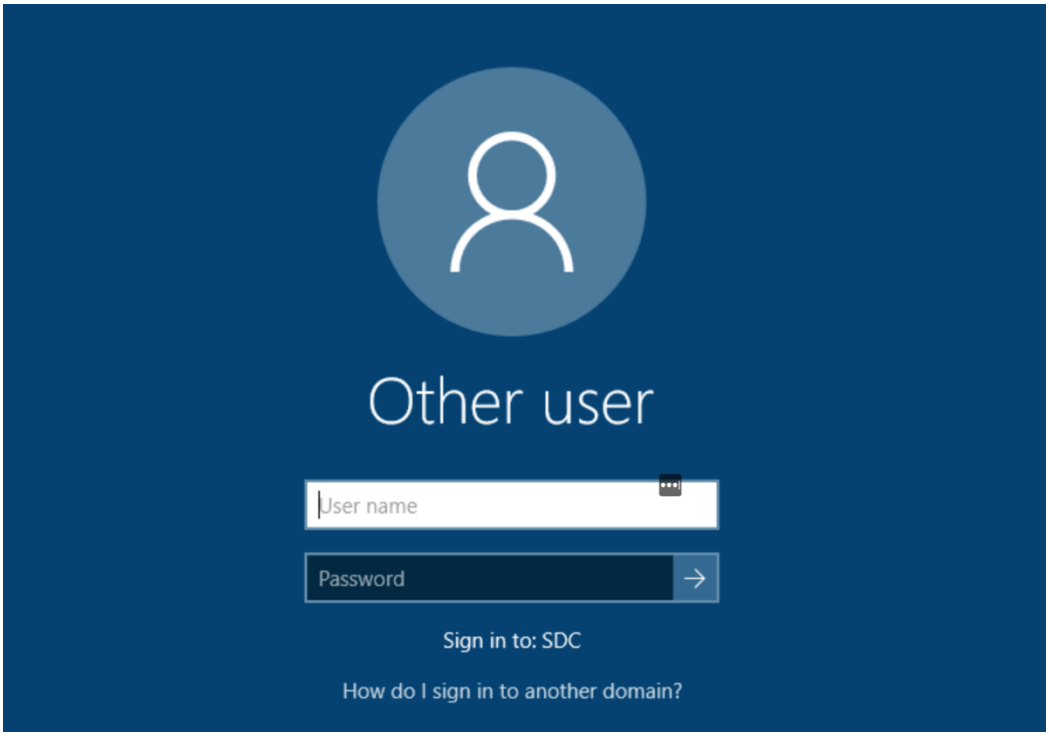


Figure 57: SDC Workstation – Login

Where do I store my data?

You can store your data in your team/individual bucket. Please refer to Upload User Data to S3 Bucket Through Portal

How can I bring my own datasets/algorithm to my workstation?

Please refer to Upload User Data to S3 Bucket Through Portal to bring your own datasets/algorithm to workstation.

How can I publish my dataset/algorithm?

Follow the below steps to publish your datasets / algorithms and share with other SDC Users.

1. Navigate to the Datasets page.
2. Click on the Publish button for the dataset/algorithm you wish to publish.

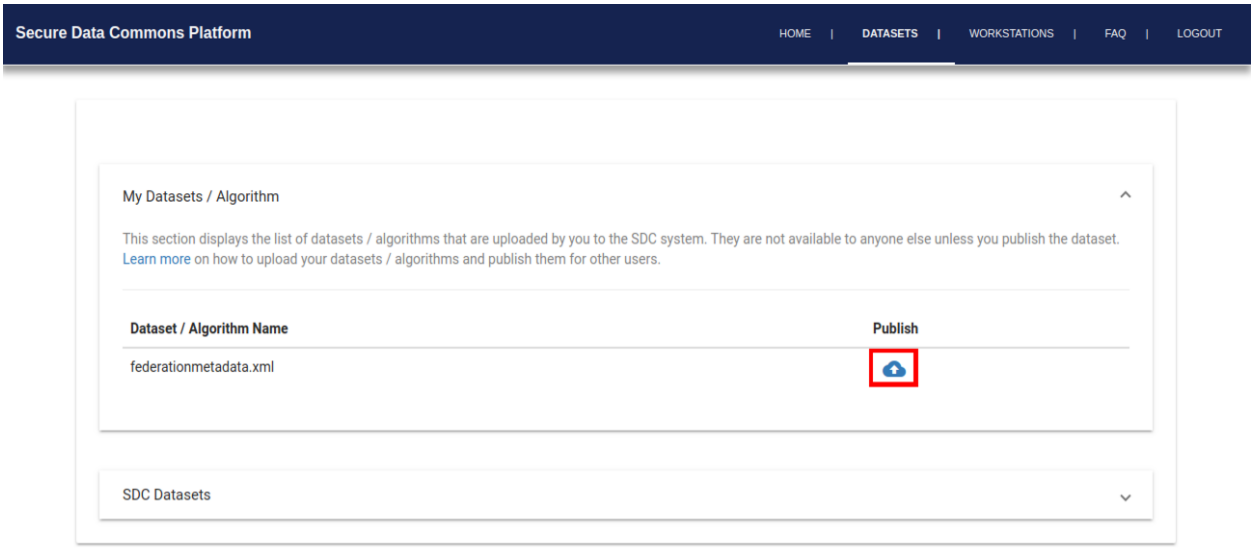


Figure 58: SDC Portal – Publish Button

3. In the pop-up window, there are two options for the Type: either a **Dataset** or **Algorithm**.
   a. From the Type drop-down menu, select **Dataset**:

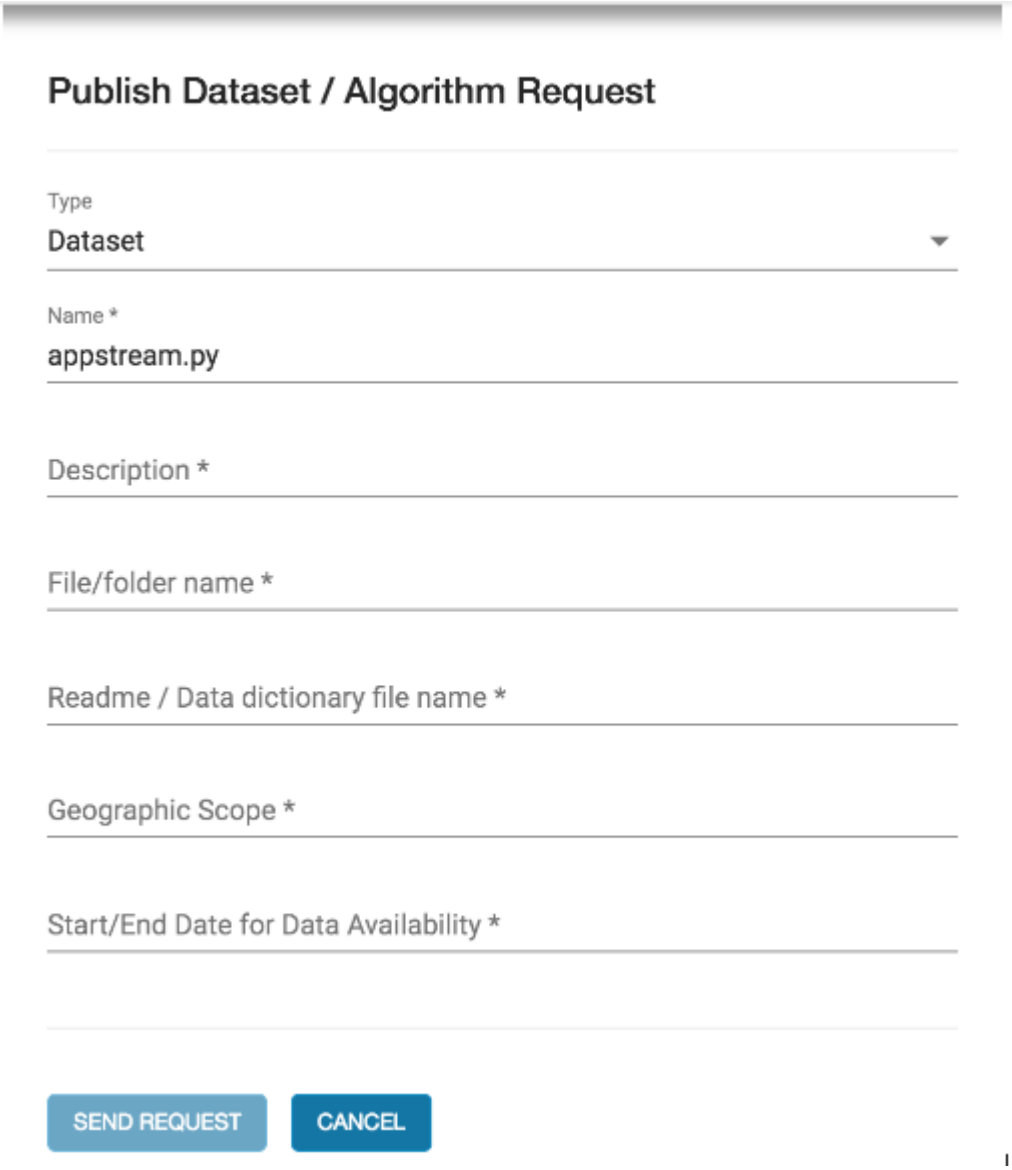


Figure 59: SDC Portal – Publish Dataset Request Form

i. Name - Name of the dataset, which you wish to call it. Users will see your dataset with this name under SDC Datasets section.
ii. Description - Provide a short description so users can get an idea about your dataset.
iii. File/folder name - Name of the file or folder where your dataset resides in your S3 Bucket. We need this information, so the support team can publish this dataset and make it available to other users.
iv. Readme / Data dictionary file name - This file should provide detailed instructions about your dataset, how it was created or any relevant information that helps user to understand and use the dataset. Save this file in your home folder relative to the dataset file/folder name.

v. Geographic scope - Indicate the geographic scope for your dataset whether it belongs to a specific state, region, country etc.
vi. Start/End Date for data availability - Provide the start and end dates of the data that belongs in your dataset. For example, your dataset may contain data from March 2017 to August 2017.

b. From the Type drop-down menu, select **Algorithm**.

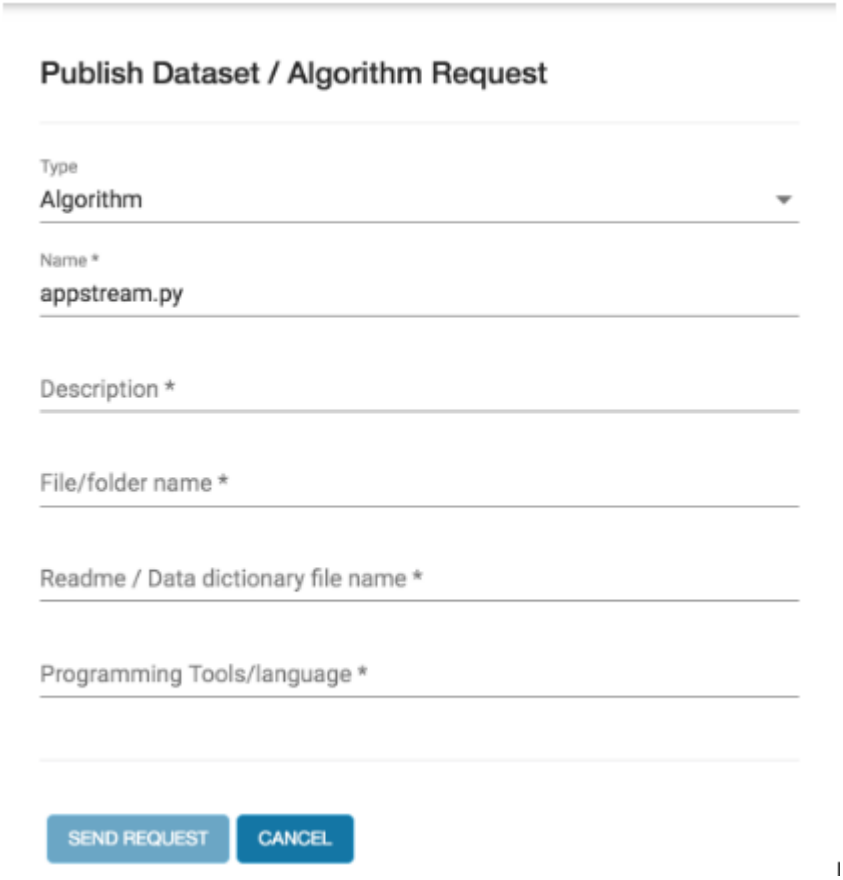


Figure 60: SDC Portal – Publish Algorithm Request Form

i. Name - Enter the name for your algorithm. Users will see your algorithm with this name under SDC Datasets section
ii. Description - Provide a short description about your algorithm
iii. File/Folder name - Name of the file or folder where your algorithm resides in your S3 bucket. We need this information, so SDC support team can publish this algorithm and make it available to other users
iv. Readme / Data dictionary file name - This file should provide detailed instructions about your algorithm, how it was created or any relevant information that helps user to understand and use the algorithm. Save this file in your home folder relative to the algorithm file/folder name

v. Programming Tools/language - Provide the details of programming tools and/or languages that were used to create this algorithm, so users can leverage the same to run your program.

Where can I find sample queries for my dataset(s)?

Sample queries are provided for each of the datasets on a GitLab page we have set up for code sharing: http://scm.securedatacommons.internal/sdc_code_collaboration/sdc-hadoop-sql-query-examples. Please refer to the wiki page of a specific dataset to find its related queries there.

U.S. Department of Transportation
ITS Joint Program Office-HOIT
1200 New Jersey Avenue, SE Washington, DC 20590

Toll-Free "Help Line" 866-367-7487
www.its.dot.gov

FHWA-JPO-18-XXX

U.S. Department of Transportation