Secure Data Commons

Data Analyst User Guide

www.its.dot.gov/index.htm
Draft Report — October 29, 2019
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	Add Export Functionality
3	REAN Cloud	3.0	10/29/2019	Add Manage Workstations section and update per feedback comments

Table of Contents

Table of Contents iv
Chapter 1. Introduction and Document Overview 1
Prerequisites 2
Chapter 2. Initial Setup and Validation 3
Accessing Secure Data Commons Portal 3
Request Access to Datasets 5
Upload User Data to S3 Bucket Through Portal 8
Download User Data from S3 Bucket Through Portal 9
Chapter 3. Accessing and Launching Workstations 10
Launch Workstations 10
Software Validation 12
Connecting to the Data Warehouse Using SQL Workbench 13
Connecting to Waze Data in Redshift 13
Connecting to CVP Data in Hive Metastore 14
Update Data Formatting Settings in SQL Workbench 14
Connecting to Redshift from Linux Environments 15
Accessing Jupyter Notebook and RStudio Server 16
Manage Workstations 16
Resize Workstation 17
Schedule/Extend Uptime 19
Stop Workstations 21
Chapter 4. Sample Queries for SDC Datasets 22
Connected Vehicle Data 22
Overview 22
WYDOT 22
WYDOT BSM and TIM Metadata RECORDGENERATEDAT vs ODERECEIVEDAT 23
THEA 24

Geospatial Queries 25
Waze Data 25
Overview 25
Chapter 6. Exporting Datasets from the SDC 27
Chapter 7. Technical Documentation and Contact Information 31
Architecture Diagram 31
Workstation Details 31
Tools and Versions 32
Contact Information 32
Useful Links 32
AWS S3 CLI Commands 32
Chapter 8. Frequently Asked Questions 34
How can I get access to the SDC Datasets? 34
How will I understand what a particular dataset consists of? 35
How can I launch a workstation? 35
Where do I store my data? 36
How can I bring my own datasets/algorithm to my workstation? 36
How can I publish my dataset/algorithm? 36

List of Tables

Table 1: Default Workstation Details 31
Table 2: List of Tools Used and Their Versions 32

List of Figures

Figure 1: SDC Starting Homepage 3
Figure 2: Register/Login Page 4
Figure 3: Login Page 4
Figure 4: Landing Page After Login 5
Figure 5: Datasets Option 6
Figure 6: Request Dataset Access 6
Figure 7: Send Data Access Request 7
Figure 8: Uploading Files 8
Figure 9: Upload Success 8
Figure 10: Selecting Files for Download 9
Figure 11: Starting Workstations 10
Figure 12: Run Instance Success 10
Figure 13: Launch Workstations 11
Figure 14: Initialization and Login Screens 11
Figure 15: SQL Workbench Icon 13
Figure 16: Create Redshift Connection Profile 13
Figure 17: Hive Connection Settings 14
Figure 18: Tools → Options → Data formatting 15
Figure 19: Manage Workstation 16
Figure 20: Manage Workstation Options 17
Figure 21: Resize Workstation Option 17
Figure 22: Workstation Stopped for Resize Changes 17
Figure 23: Resizing Options 18
Figure 24: Schedule Resize 19
Figure 25: Schedule Workstation Uptime Option 19
Figure 26: Schedule Uptime 20
Figure 27: Tooltip with Existing Scheduled Uptime 20
Figure 28: New Tooltip with Extended Uptime Schedule 20
Figure 29: Stop Workstation 21
Figure 30: WYDOT BSM Tables 22
Figure 31: Request Export 27
Figure 32: Request Export Form 28
Figure 33: Approval Form Fields 29
Figure 34: Acceptable Use Policy 30
Figure 35: SDC Architecture Overview 31
Figure 36: Datasets Tab 34
Figure 37: Dataset Access Request 35
Figure 38: Workstations Tab 35
Figure 39: Login for Programming Environment 36
Figure 40: Upload Button 37

Figure 41: Publish Dataset Form 37
Figure 42: Algorithm Request Form 38

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, 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.

This document provides guidance for the **Data Analyst** role. A similar guide will be prepared for the Data Providers. The 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

- 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.

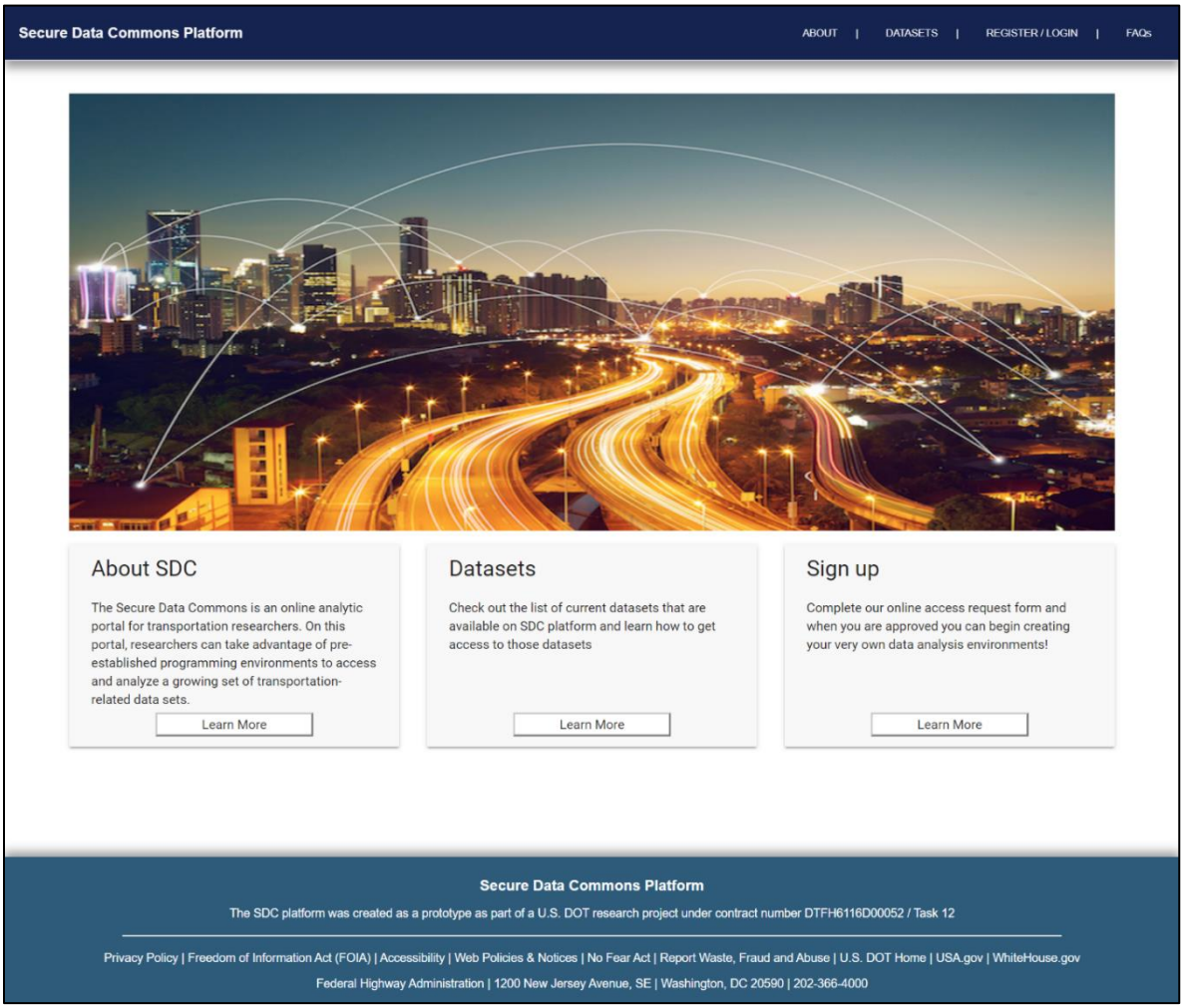


Figure 1: SDC Starting Homepage

Users can click on Register/Login from the top menu to access the Access Request Form and Privacy Policy links, as well as the Login button.

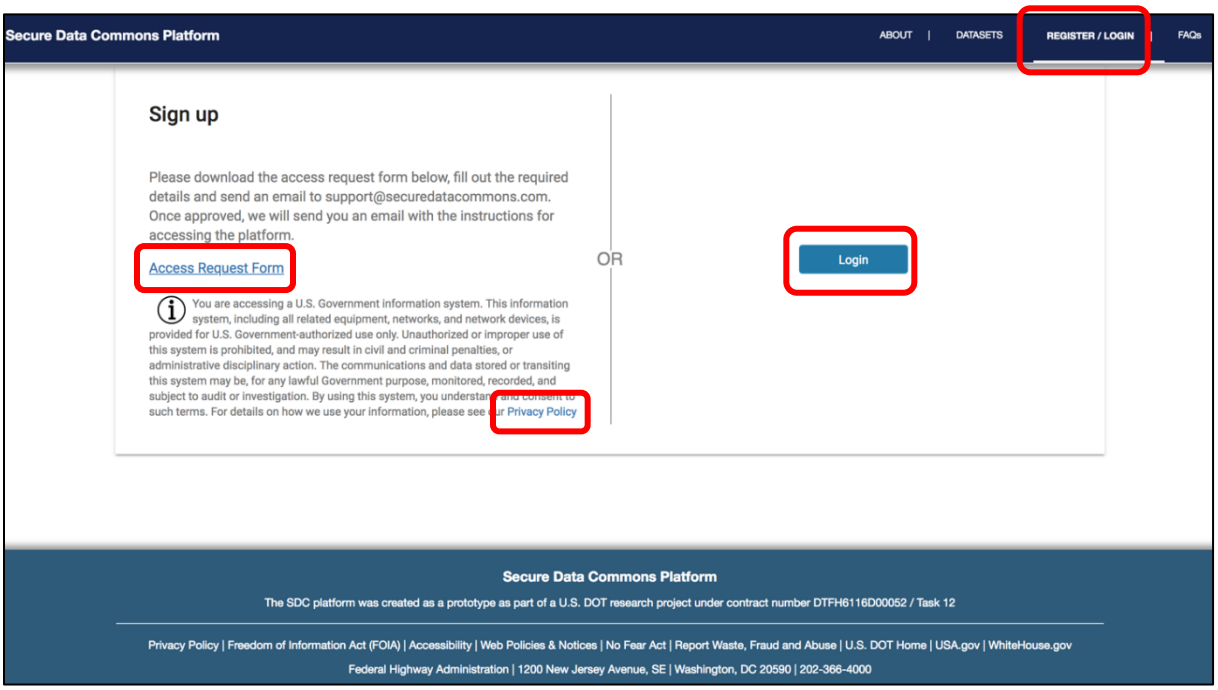


Figure 2: Register/Login Page

Clicking on Login redirects the user to the Secure Data Commons platform login page:

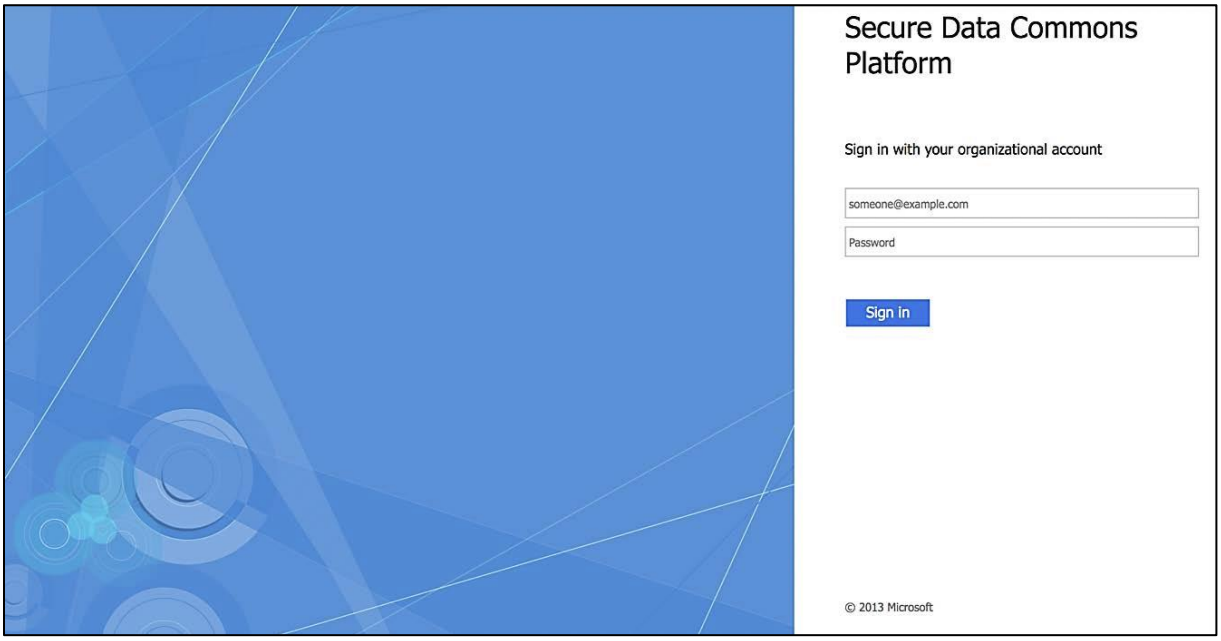


Figure 3: Login Page

If the user is accessing the portal for the first time, they will be prompted to change their password after entering the credentials provided in the welcome email.

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

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

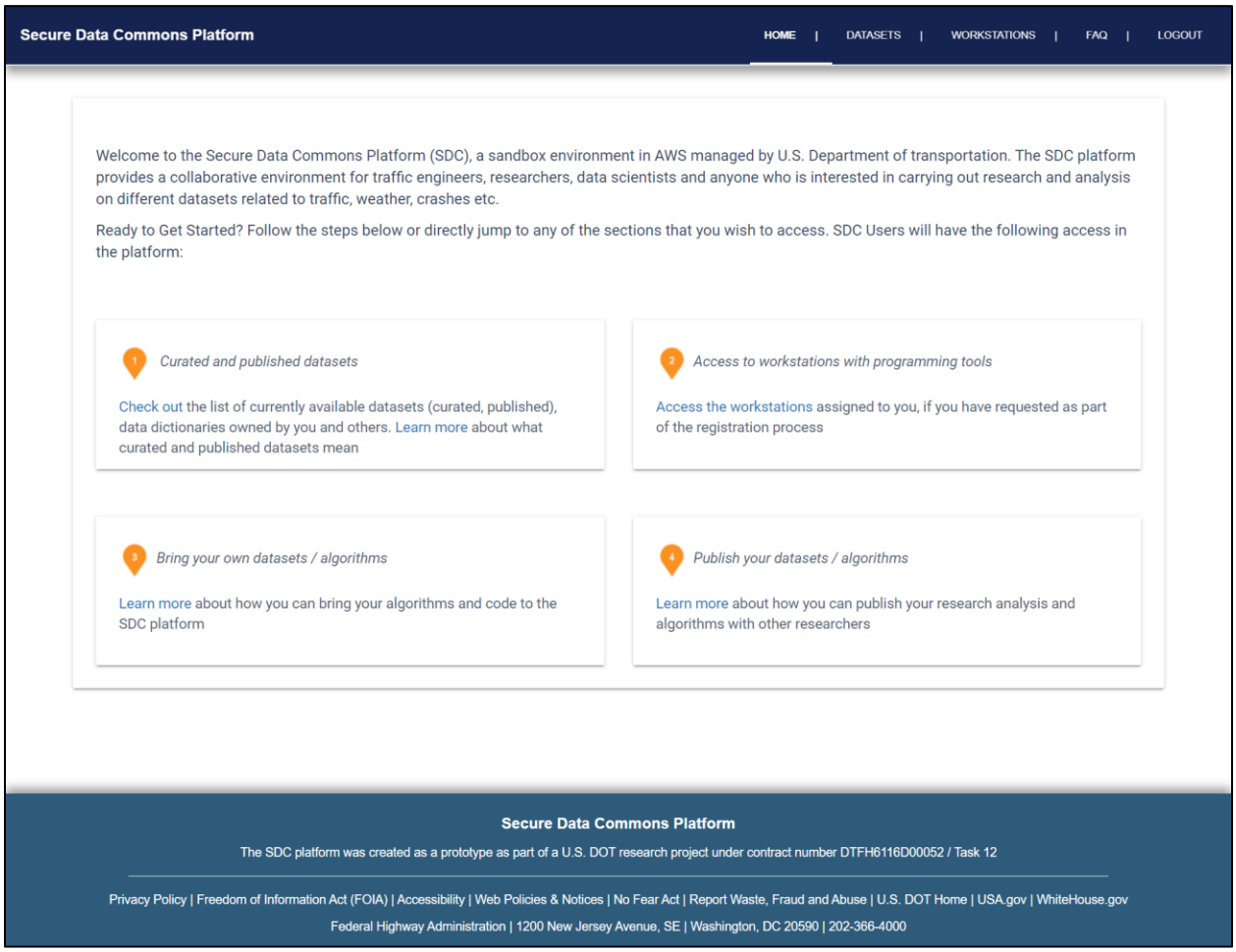


Figure 4: Landing Page After Login

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.

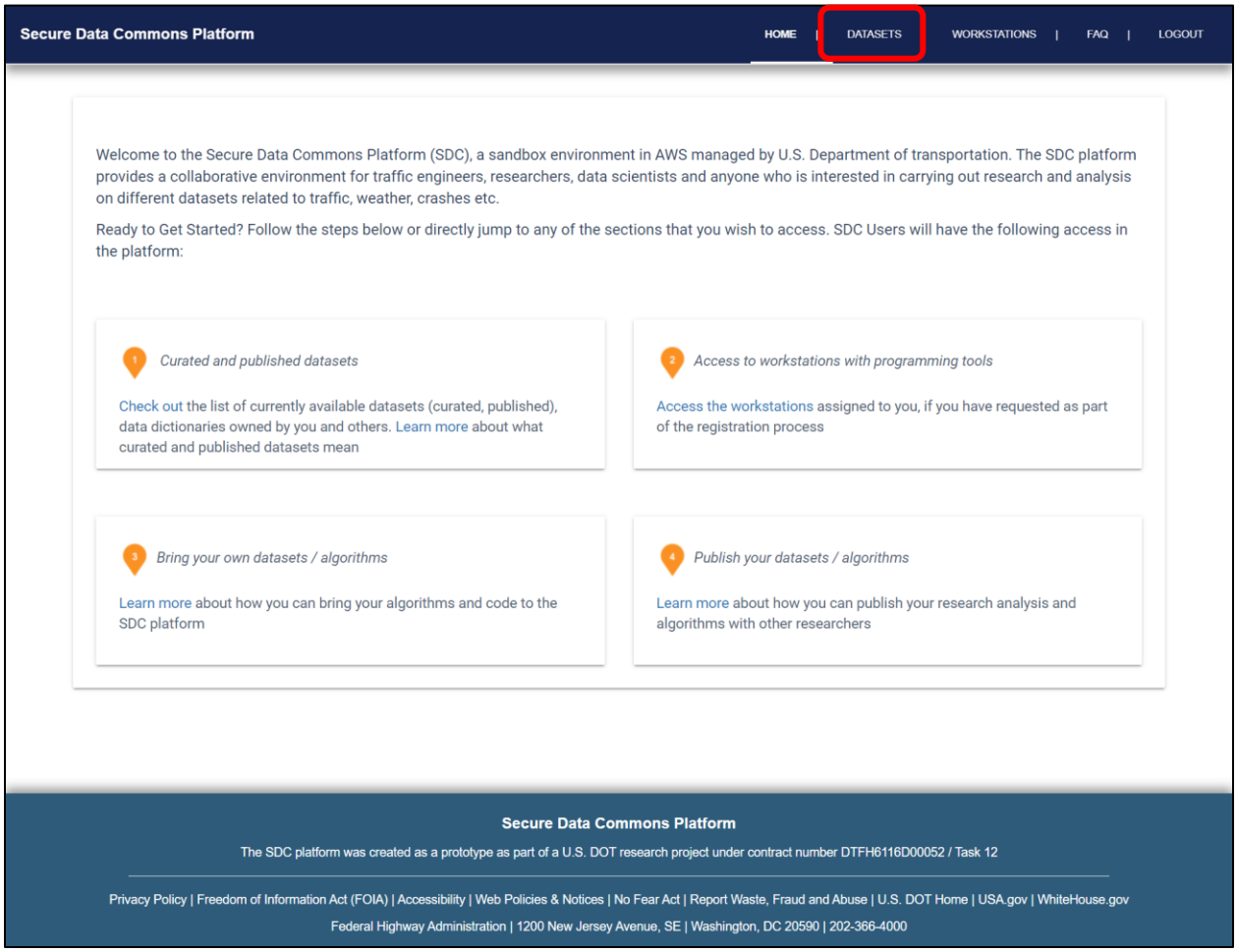


Figure 5: Datasets Option

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

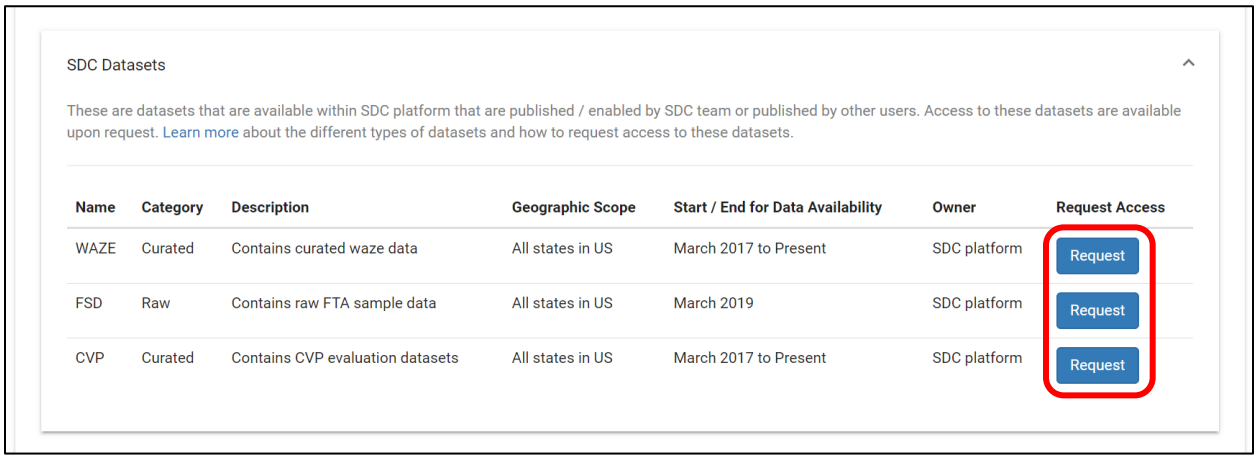


Figure 6: Request Dataset Access

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

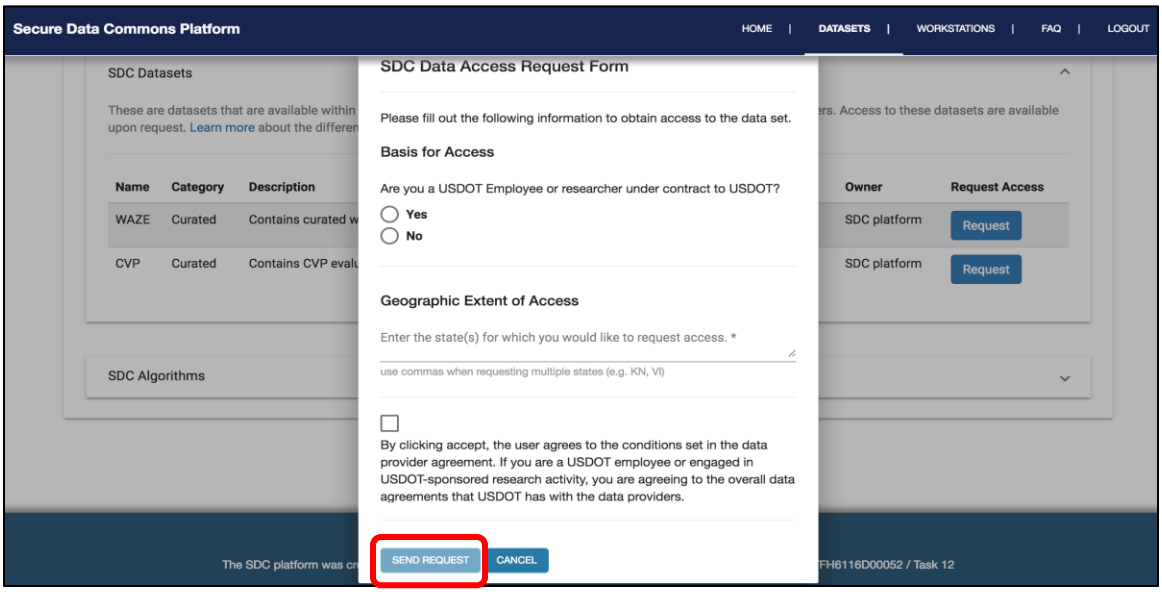


Figure 7: Send Data Access Request

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 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.)

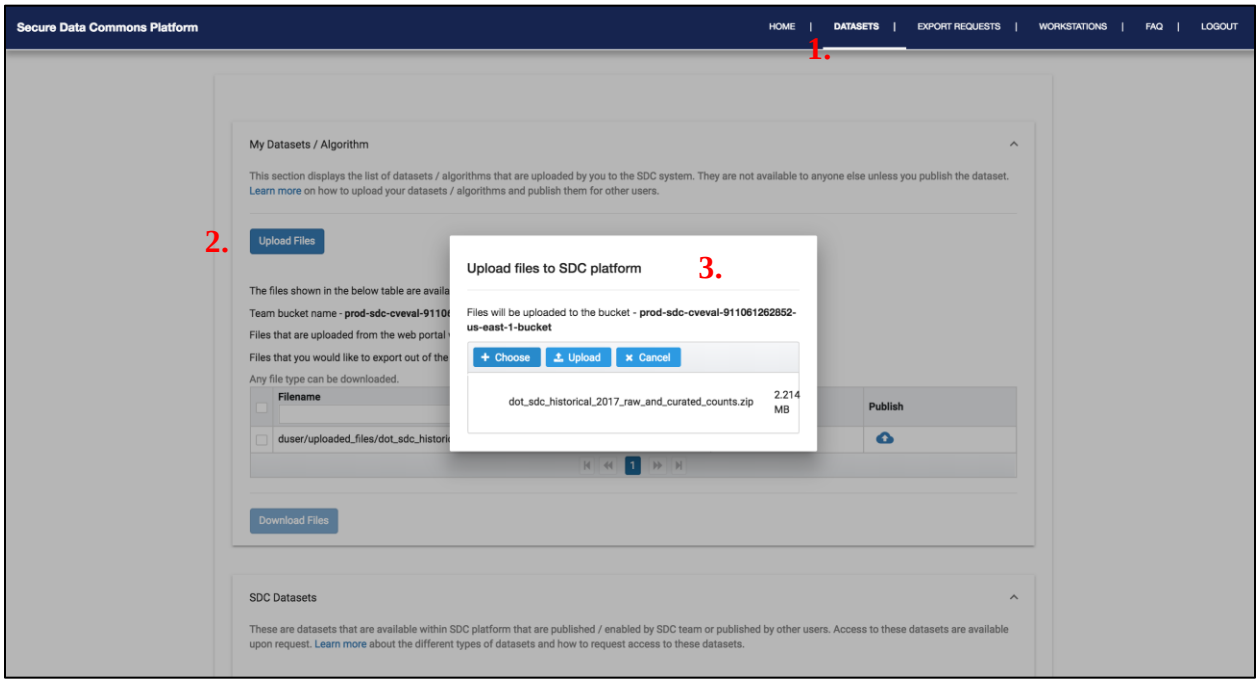


Figure 8: Uploading Files

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

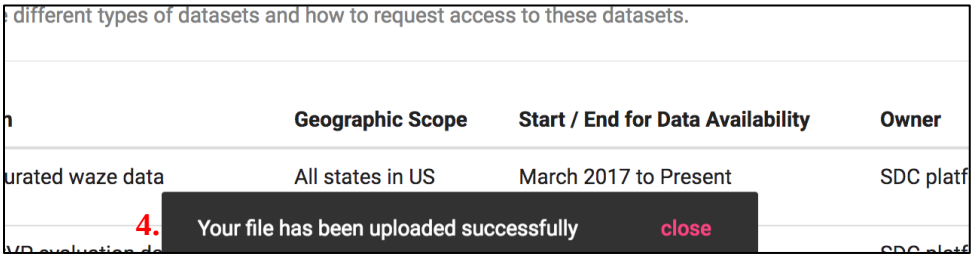


Figure 9: Upload Success

5. Files that are uploaded from the web portal will be saved in the folder – **username/uploaded_files**
6. Users would be able to access only the files that are under **uploaded_files, export_requests** folder.

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.

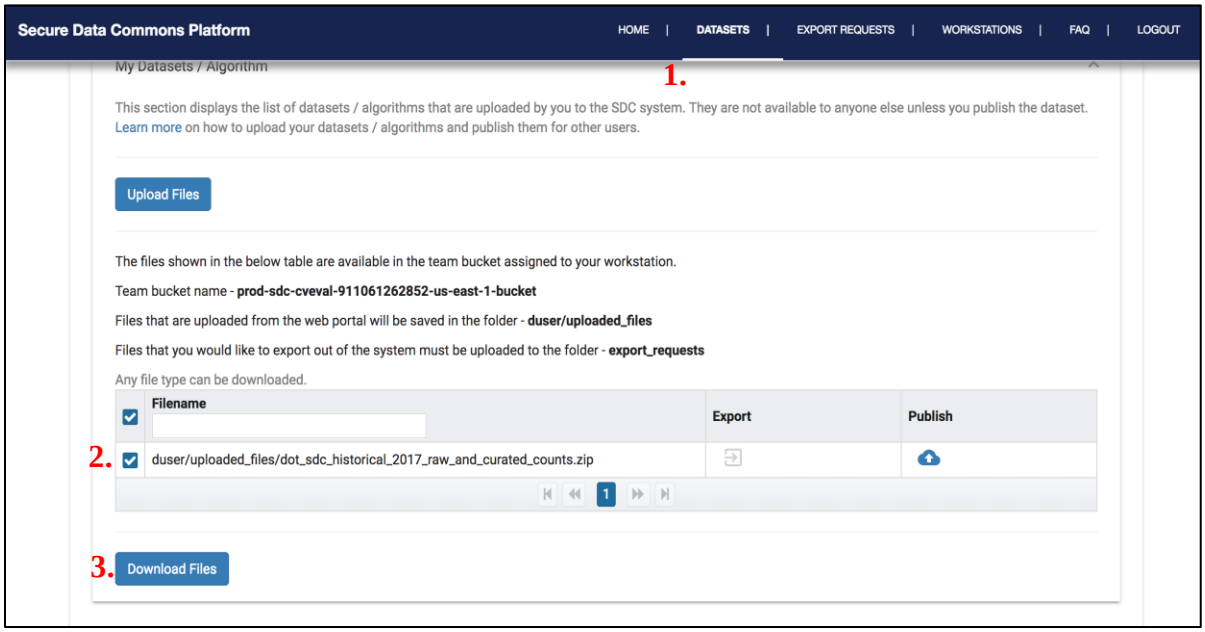


Figure 10: Selecting Files for Download

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.

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.

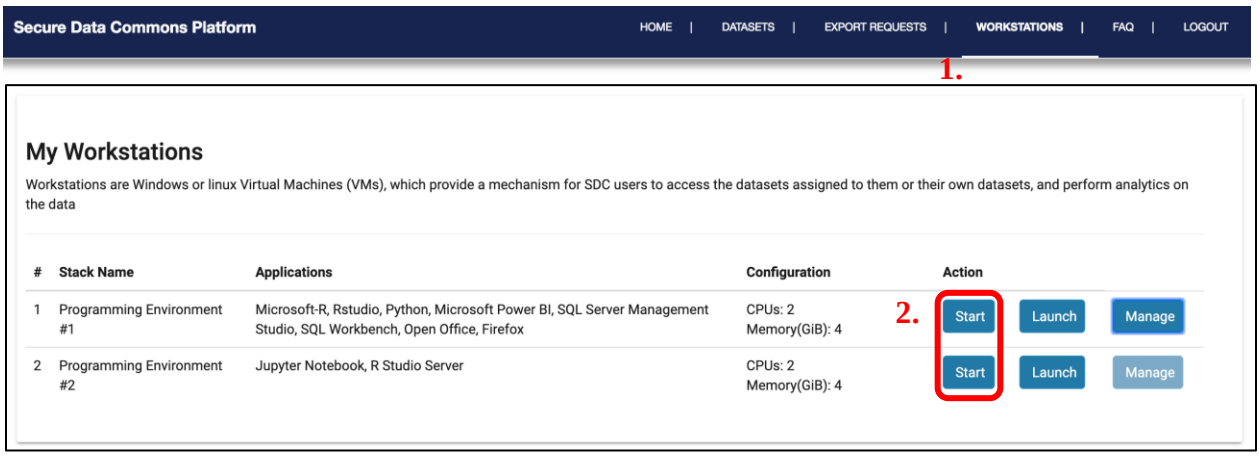


Figure 11: Starting Workstations

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.

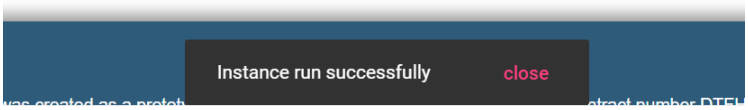


Figure 12: Run Instance Success

4. Now click Launch for the workstation.

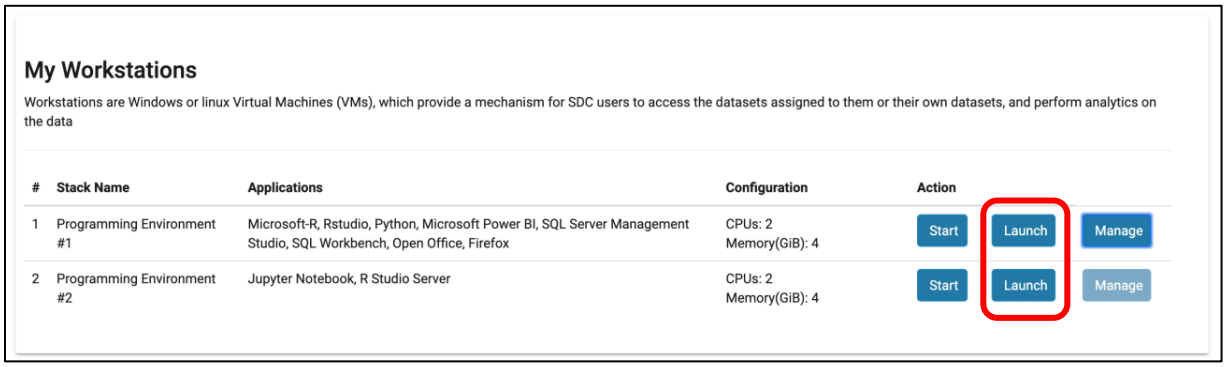


Figure 13: Launch Workstations

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.

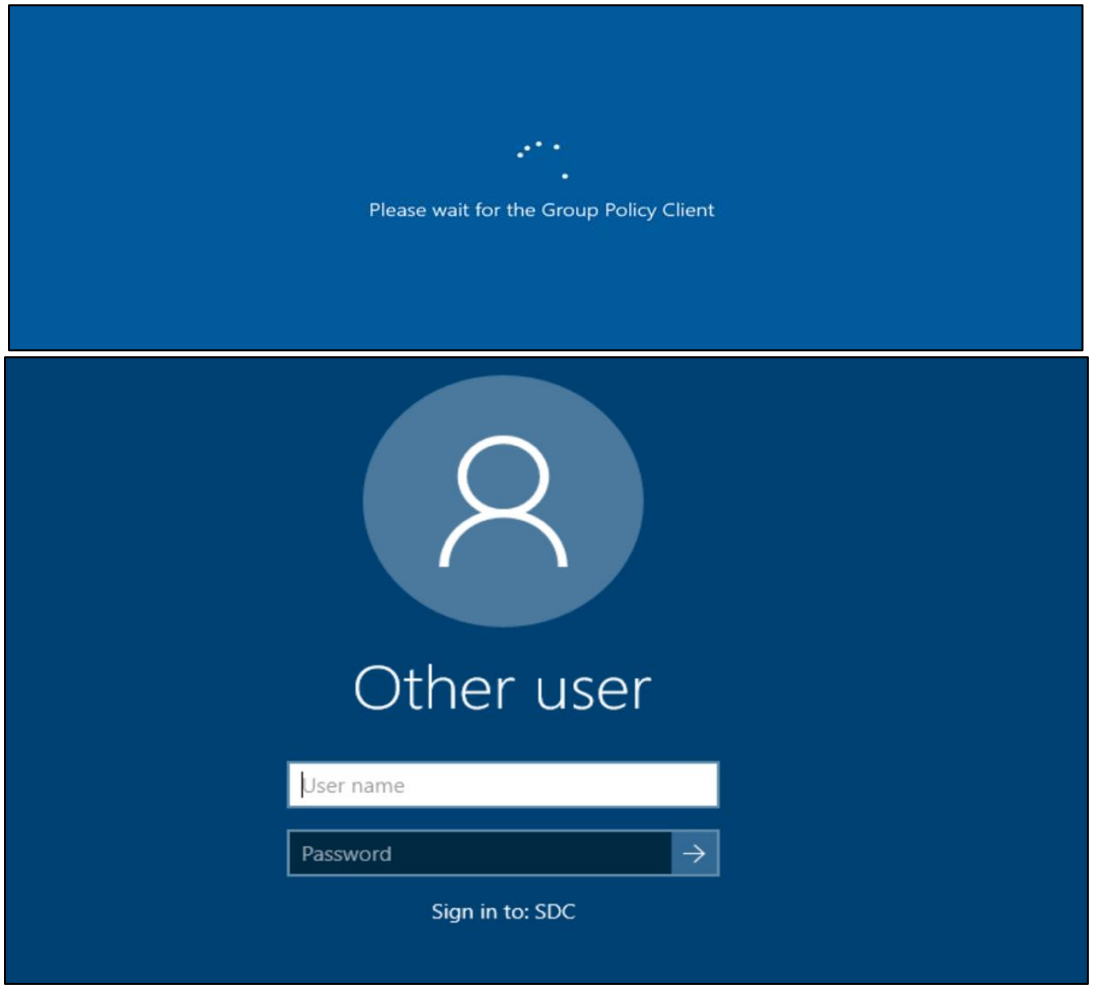


Figure 14: Initialization and Login Screens

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

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

```
from impala.dbapi import connect
conn = connect(
  host='172.18.1.20',
  port=10000,
  auth_mechanism='PLAIN',
  user='<your_username>' ,password='<your_password>')
cursor = conn.cursor()
cursor.execute('SHOW TABLES')
print cursor.fetchall()
```

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

Connecting to the Data Warehouse Using SQL Workbench

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

Connecting to Waze Data in Redshift

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

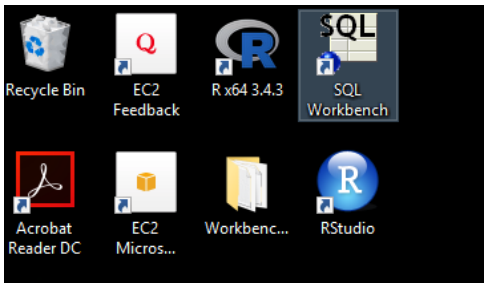


Figure 15: SQL Workbench Icon

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.

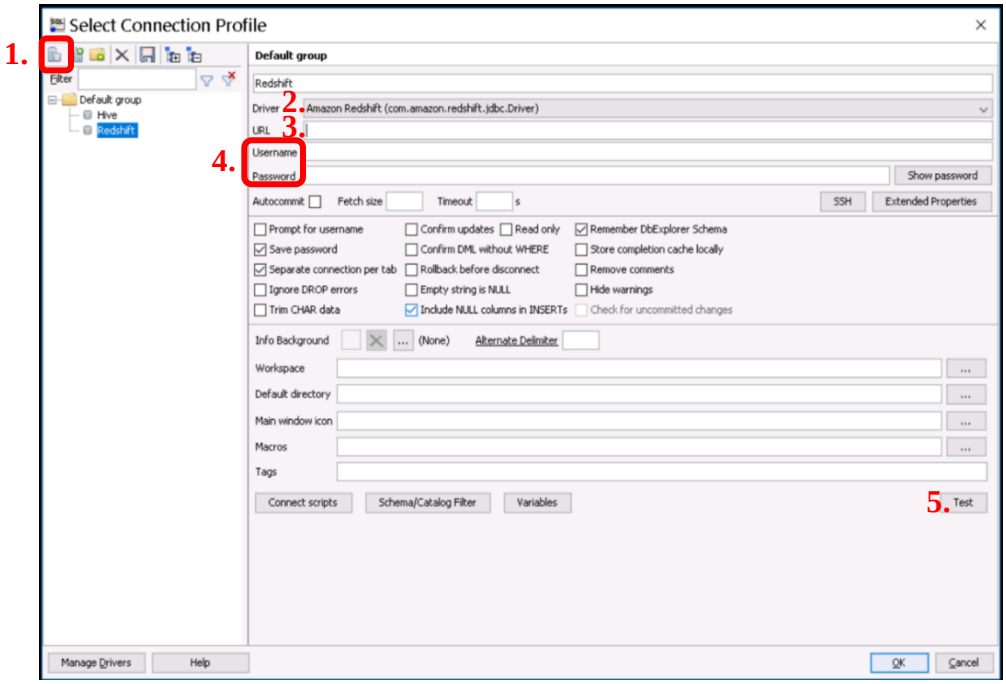


Figure 16: Create Redshift Connection Profile

Connecting to CVP Data in 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 provided in the cheat sheet.
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.

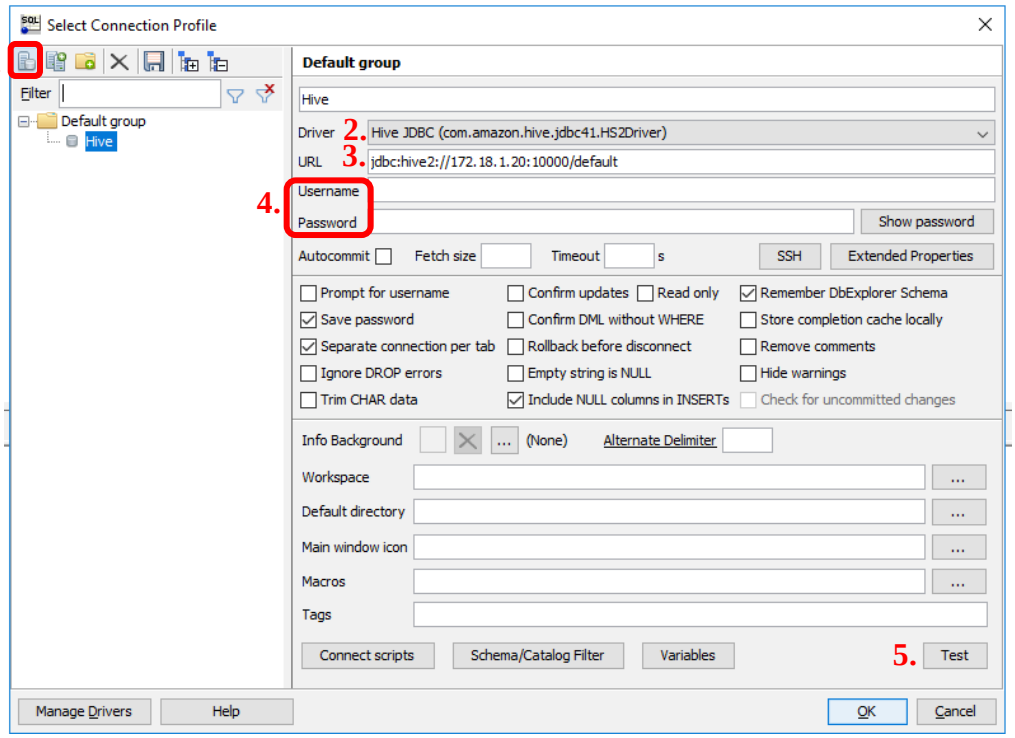
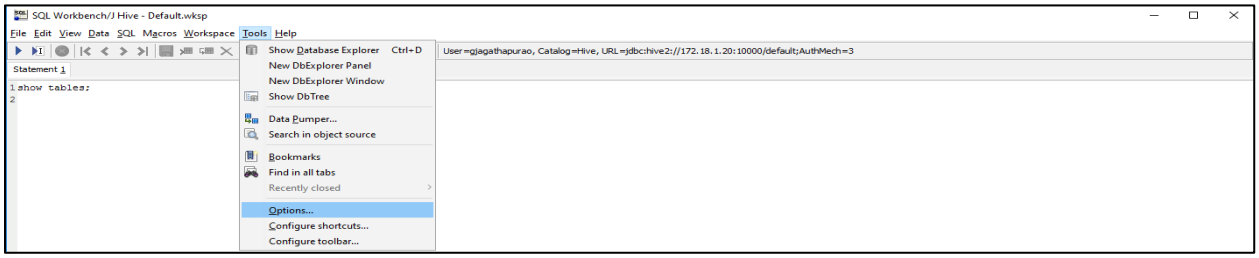


Figure 17: Hive Connection Settings

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.



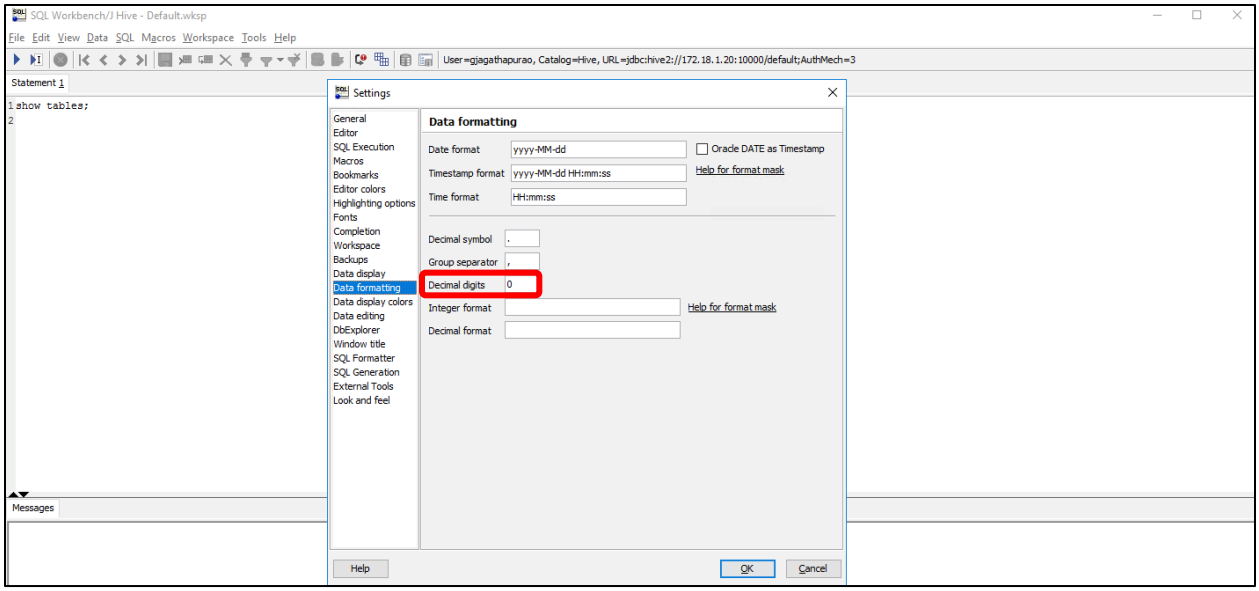


Figure 18: Tools → Options → Data formatting

Connecting to Redshift from Linux Environments

Credentials to access the Waze Redshift database are communicated from the SDC Administrator (support@securedatacommons.com) by a secure email service. Users being granted access will receive an email in the normal email client (such as Outlook) with an "Unlock Message" link. Clicking on this will link to a secure reader that displays the credentials as an email in the web browser.

- 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.

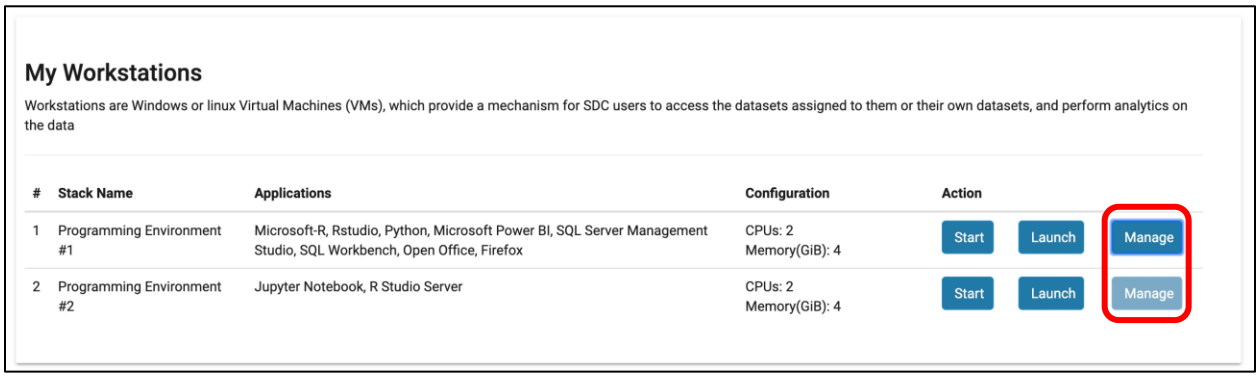


Figure 19: Manage Workstation

A dialogue window appears with two checkbox options:

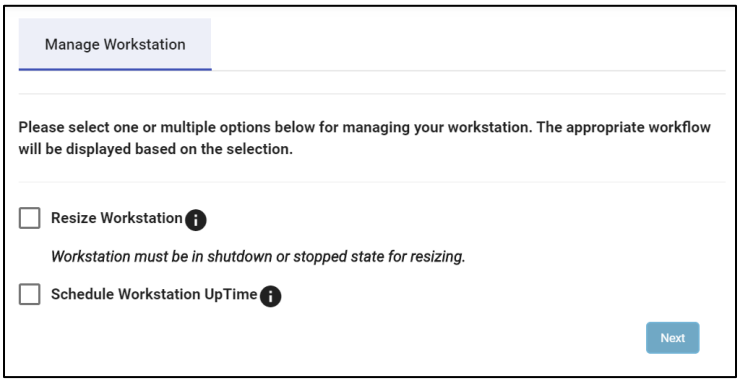


Figure 20: Manage Workstation Options

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.

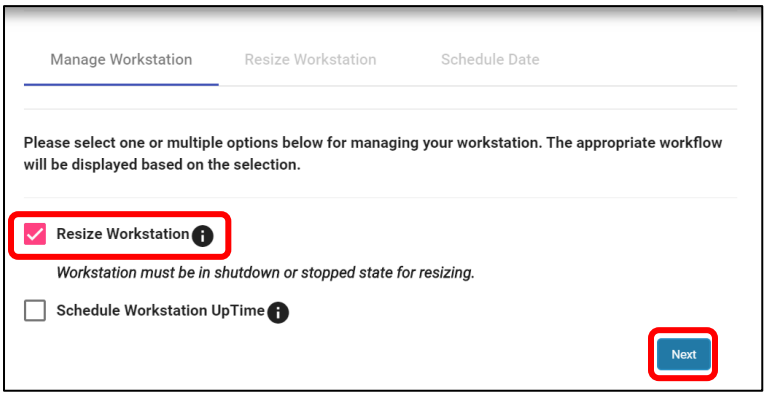


Figure 21: Resize Workstation Option

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

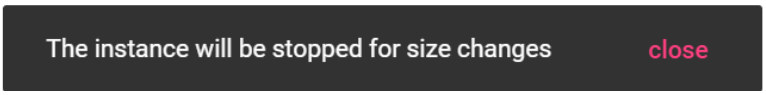


Figure 22: Workstation Stopped for Resize Changes

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.

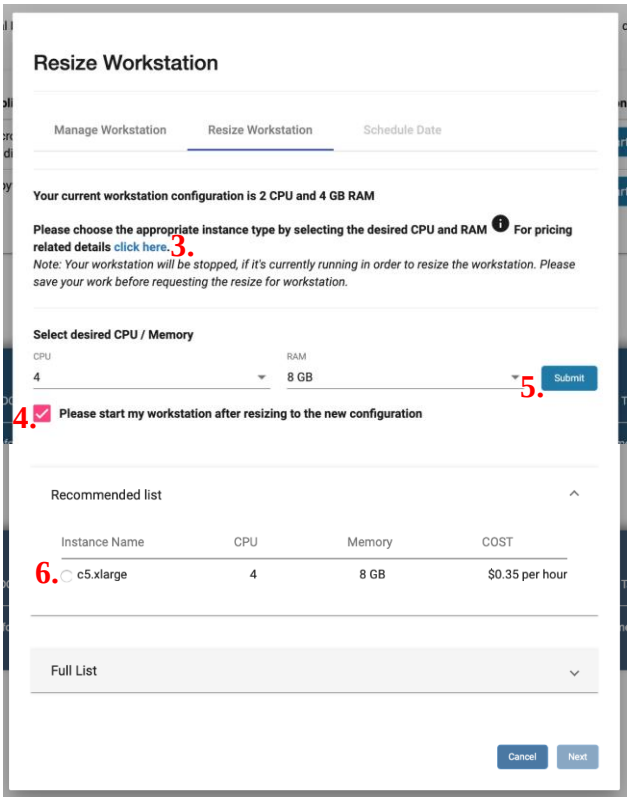


Figure 23: Resizing Options

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.

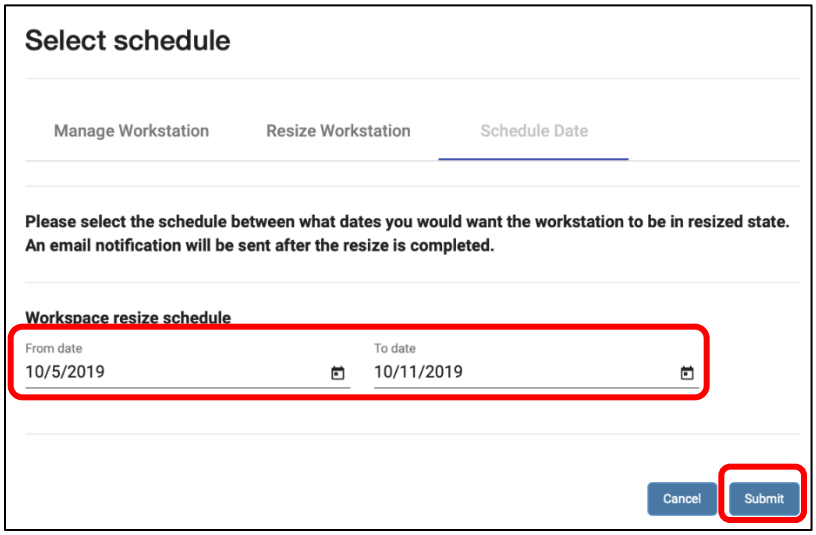


Figure 24: Schedule Resize

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.

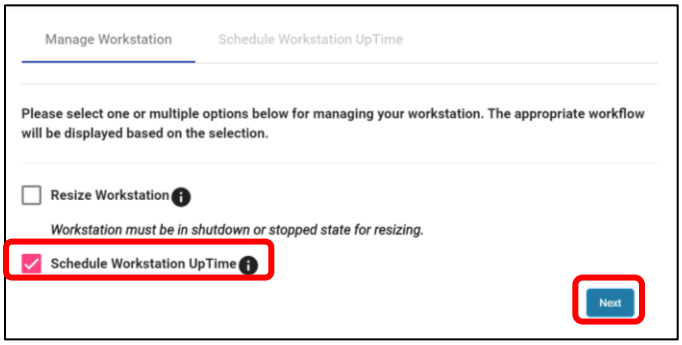


Figure 25: Schedule Workstation Uptime Option

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.

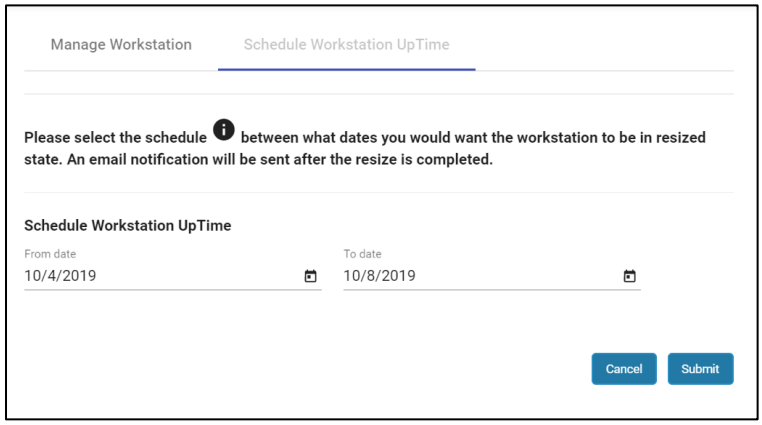


Figure 26: Schedule Uptime

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 27: 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.

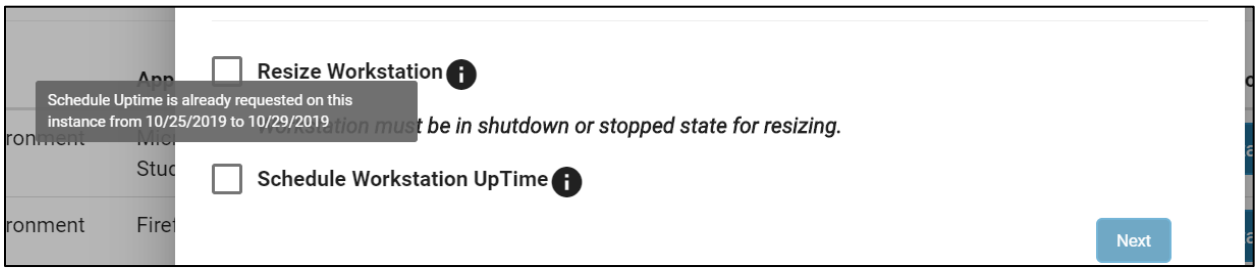


Figure 28: New Tooltip with Extended Uptime Schedule

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.

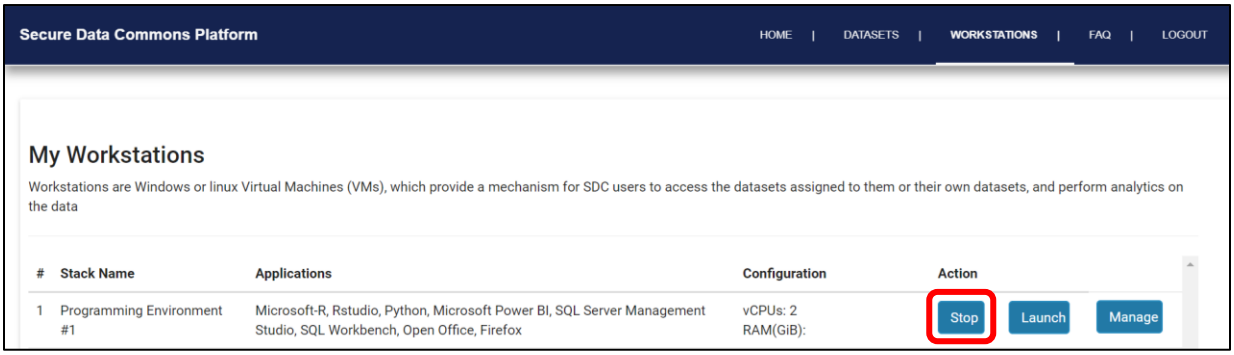


Figure 29: Stop Workstation

Chapter 4. Sample Queries for SDC Datasets

The following section provides some sample queries for anchor datasets hosted by the SDC. Two datasets which are currently available in the SDC system:

- Data from USDOT-sponsored Connected Vehicle Pilot (CVP) program, provided by the following pilot sites:
  - Tampa Hillsborough Expressway Authority (THEA)
  - Wyoming Department of Transportation (WYDOT)
- Data provided by Waze on traffic jams and alerts

Sample queries are provided for each of the two datasets.

Connected Vehicle Data

Overview

The CVP Data Warehouse is built on top of a Hadoop's HDFS cluster and utilizes Hive with a HiveQL querying language as a front-end querying package. Information about the HiveQL language and the language manual can be found online:

- Wikipedia info on Apache Hive: https://en.wikipedia.org/wiki/Apache_Hive
- Language manual: https://cwiki.apache.org/confluence/display/Hive/LanguageManual

Key components of the CVP dataset are the basic safety messages (BSM) and Traveler Information Messages (TIM). Both THEA and WYDOT data include BSM and TIM.

WYDOT

Basic Safety Message (BSM) Relational Tables

WYDOT BSM data are stored in the following set of relational tables.

```
wydot_bsm_core
wydot_bsm_partii
wydot_bsm_partii_crumbdata
```

The following figure illustrates the foreign key relationship of WYDOT BSM tables.

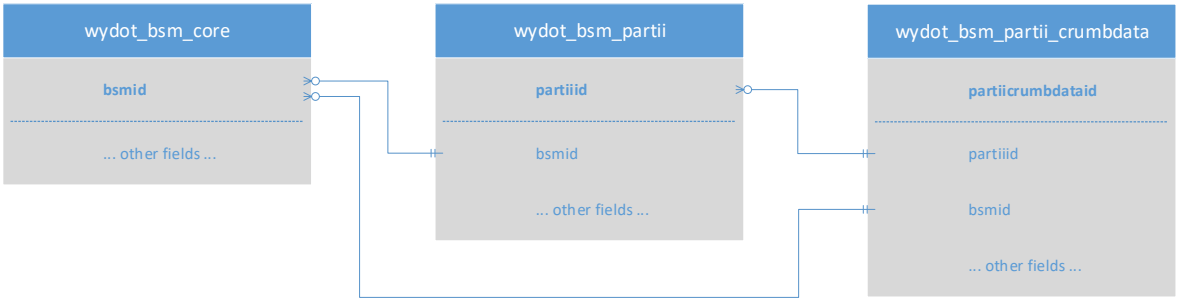


Figure 30: WYDOT BSM Tables

Sample queries using WYDOT BSM tables.
These queries can be run using the SQL Workbench which is configured to connect to the data warehouse. Note: running the last three queries will take longer, because they are joining multiple tables.

```
select * from wydot_bsm_core limit 5;
select * from wydot_bsm_partii limit 5;
select * from wydot_bsm_partii_crumbdata limit 5;

select coredatalatitude, coredatalongitude from
wydot_bsm_core limit 5;

select * from wydot_bsm_core join wydot_bsm_partii where
wydot_bsm_core.bsmid = wydot_bsm_partii.bsmid limit 5;

select * from wydot_bsm_partii join
wydot_bsm_partii_crumbdata where wydot_bsm_partii.partiiid
= wydot_bsm_partii_crumbdata.partiiid limit 5;

select * from wydot_bsm_core join wydot_bsm_partii join
wydot_bsm_partii_crumbdata where wydot_bsm_core.bsmid =
wydot_bsm_partii.bsmid and wydot_bsm_partii.partiiid =
wydot_bsm_partii_crumbdata.partiiid limit 10;
```

WYDOT BSM and TIM Metadata RECORDGENERATEDAT vs ODERECEIVEDAT

WYDOT BSM and TIM messages include a metadata section, appended by the Operational Data Environment (ODE) component. Among others, there are two timestamps of interest are presented in this section:

recordGeneratedAt: Closest time to which the record was created by a Vehicle.
odeReceivedAt: Time ODE received the data in UTC format. This time is the closest to which the CV PEP system received the record.

- Based on real-life conditions, odeReceivedAt may be days or even weeks after the recordGeneratedAt timestamp.
- The following queries result in distribution of recordGeneratedAt by odeRecevedAt times:
- WYDOT BSM messages:

```
select SUBSTR(metadatarecordgeneratedat, 0, 10) as
RECORDGENERATEDAT
, SUBSTR(metadataodereceivedat, 0, 10) as ODERECEIVEDAT
, count(SUBSTR(metadataodereceivedat, 0, 10)) as CNT
from wydot_bsm_core
group by SUBSTR(metadataodereceivedat, 0, 10)
, SUBSTR(metadatarecordgeneratedat, 0, 10)
order by RECORDGENERATEDAT limit 10000;
```

- WYDOT TIM messages:
```
select SUBSTR(metadatarecordgeneratedat, 0, 10) as
RECORDGENERATEDAT
, SUBSTR(metadataodereceivedat, 0, 10) as ODERECEIVEDAT
, count(SUBSTR(metadataodereceivedat, 0, 10)) as CNT
from wydot_tim
group by SUBSTR(metadataodereceivedat, 0, 10)
, SUBSTR(metadatarecordgeneratedat, 0, 10)
order by RECORDGENERATEDAT limit 10000;
```

WYDOT Speed Data

There are two WYDOT speed datasets in the CV PEP Data Warehouse: wydot_speed_unprocessed and wydot_speed_processed. The following sample query displays average vehicle speed distribution by lane and it will work against either of the tables:

```
select lane, avg(speedmph) as speed_average
from wydot_speed_unprocessed
group by lane
order by lane;
```

THEA

THEA Basic Safety Message (BSM) Relational Tables

THEA BSM data are stored in the following set of relational tables.

THEA BSM Relational Tables
thea_bsm_core thea_bsm_partii thea_bsm_partii_crumbdata

<u>Sample queries using THEA BSM tables.</u>

These queries can be run using the SQL Workbench which is configured to connect to the data warehouse. Note: running the last three queries will take longer, because they are joining multiple tables.

```
select * from thea_bsm_core limit 5;

select * from thea_bsm_partii limit 5;

select * from thea_bsm_partii_crumbdata limit 5;

select coredatalat, coredatalong from thea_bsm_core limit
5;
```

```
select * from thea_bsm_core join thea_bsm_partii where
thea_bsm_core.bsmid = thea_bsm_partii.bsmid limit 5;

select * from thea_bsm_partii join
thea_bsm_partii_crumbdata where thea_bsm_partii.partiiid =
thea_bsm_partii_crumbdata.partiiid limit 5;

select * from thea_bsm_core join thea_bsm_partii join
thea_bsm_partii_crumbdata where thea_bsm_core.bsmid =
thea_bsm_partii.bsmid and thea_bsm_partii.partiiid =
thea_bsm_partii_crumbdata.partiiid limit 10;
```

Geospatial Queries

- The Data Warehouse has geospatial querying capabilities. Functions such as ST_Point, ST_Polygon, ST_Contains and others can be used in queries. For the full list of supported functions see: https://github.com/Esri/spatial-framework-for-hadoop/wiki/UDF-Documentation
- As an example, here is a sample query to retrieve a count of messages generated by vehicles between latitudes 40 and 41 and longitudes between -106 and -105.

```
select count(*) from wydot_bsm_core where
ST_Contains(ST_Polygon(40, -105, 41, -105, 41, -106, 40, -
106),
ST_Point(coredatalatitude, coredatalongitude));
```

Waze Data

Overview

The SDC platform ingests and curates Waze data for all 50 states of United States of America using the Waze API. Users have access only to Waze data pertaining to the specific geographic region specified in their data access policies.

Waze Alert Data

An alert is a User Generated Incident (UGI) reported by a Waze user or group of users, as defined by Waze. In SQL Workbench, the following query can be used to test access to the *alert* table.

```
select distinct alert_uuid from alert where
alert_type='<alert_type>' and pub_utc_timestamp between
<start_time_stamp> and <end_time_stamp> and state
=<state_name>;

select distinct alert_uuid from alert where
alert_type='ACCIDENT' and pub_utc_timestamp between '2018-
```

```
01-01 04:59:43.00' and '2018-01-30 04:59:43.00' and state
='CA';
```

Waze Jam Data

Waze provides information of traffic jams and events that affect road conditions either from wazers or external sources. A traffic jam maybe associated with an alert.

```
select top 10 * from jam;
select top 10 * from jam_point_sequence;
```

Waze Irregularity Data

Irregularities are similar to Jams, where Waze derives these events based on unusual traffic patterns. These could also be a result of an alert or a jam.

```
select top 10 * from irregularity;
select top 10 * from irregularity_point_sequence;
select top 10 * from irregularity_alert;
select top 10 * from irregularity_jam;
```

Chapter 6. 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.

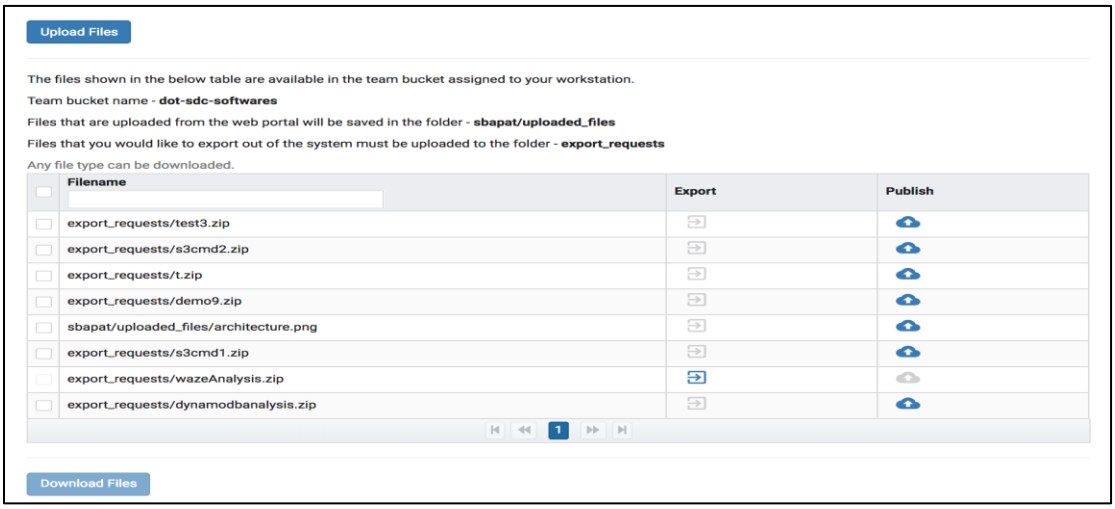


Figure 31: Request Export

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.

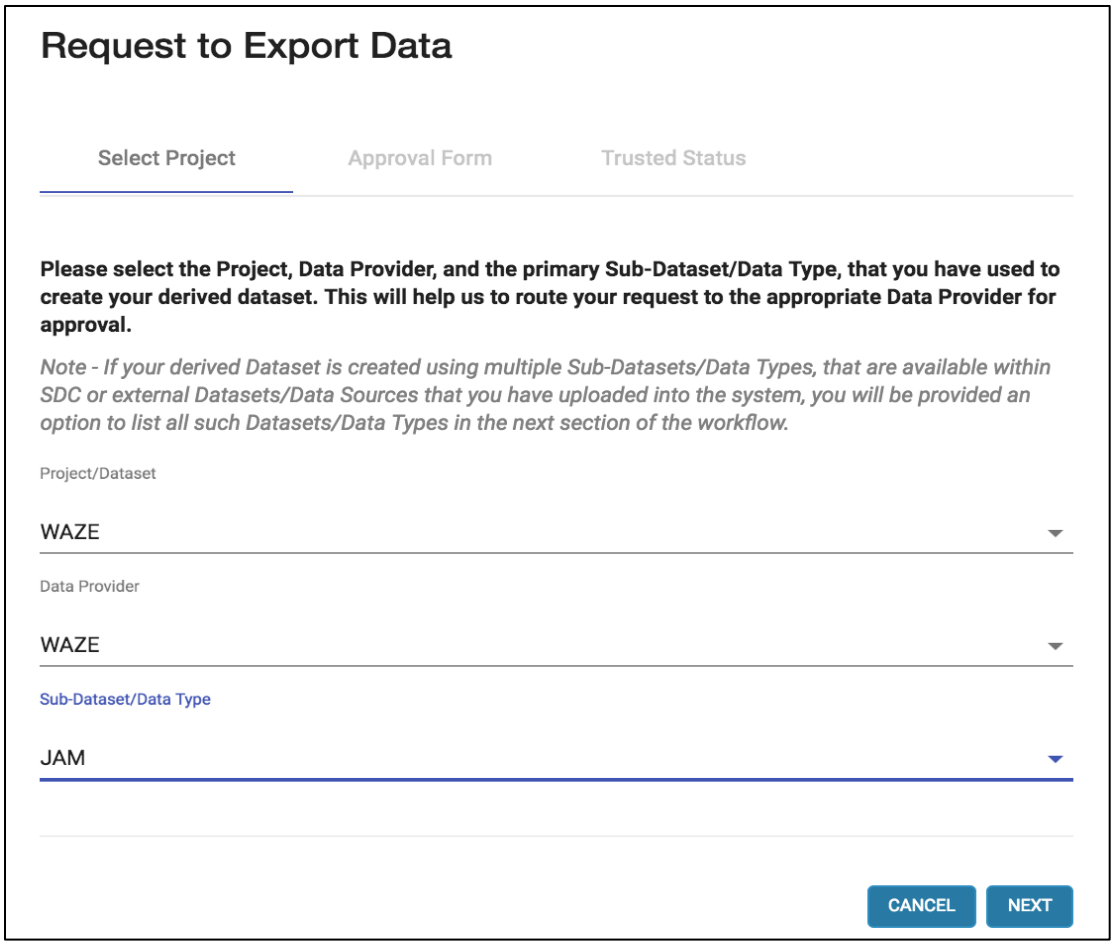


Figure 32: Request Export Form

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.

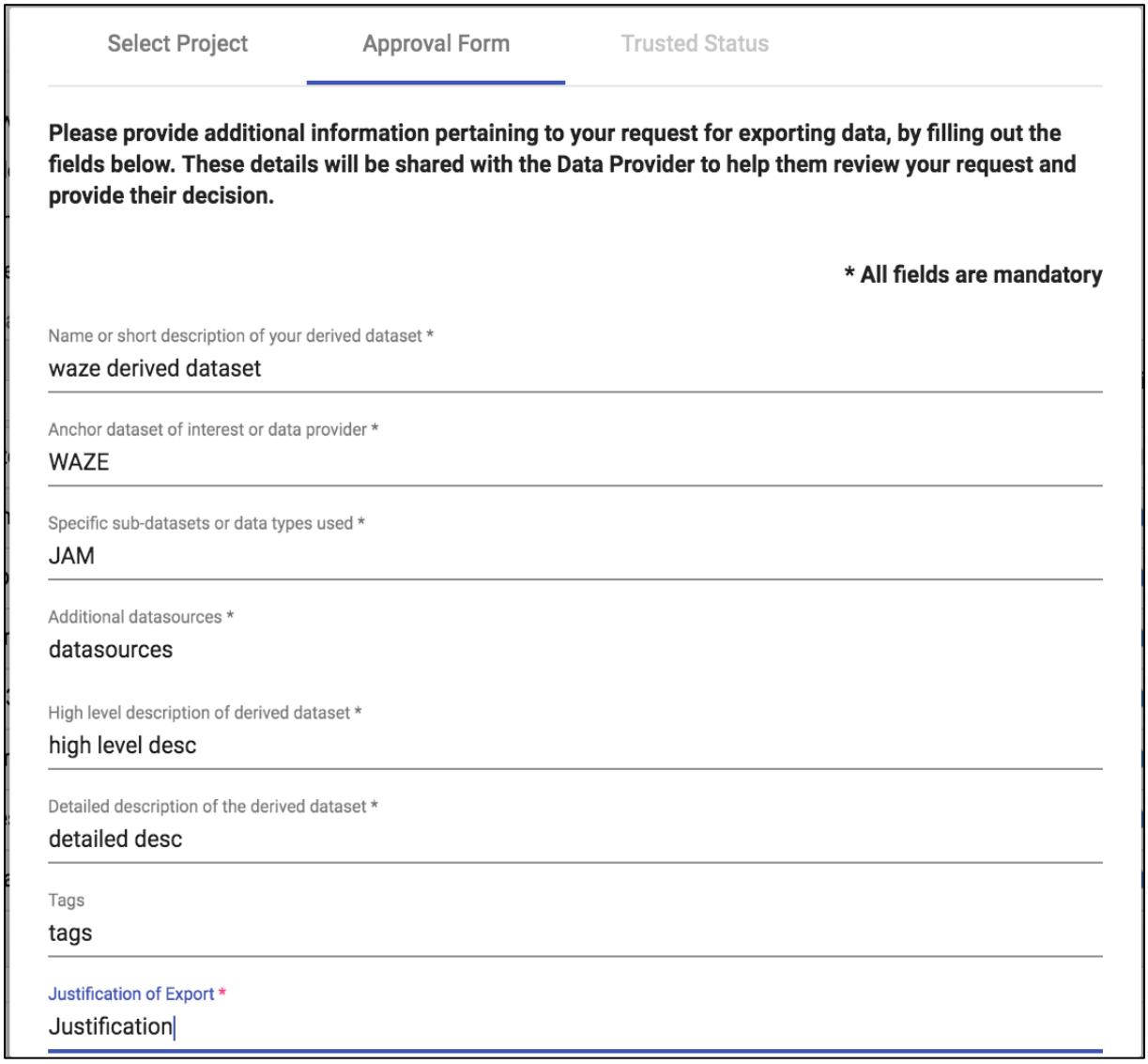


Figure 33: Approval Form Fields

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.

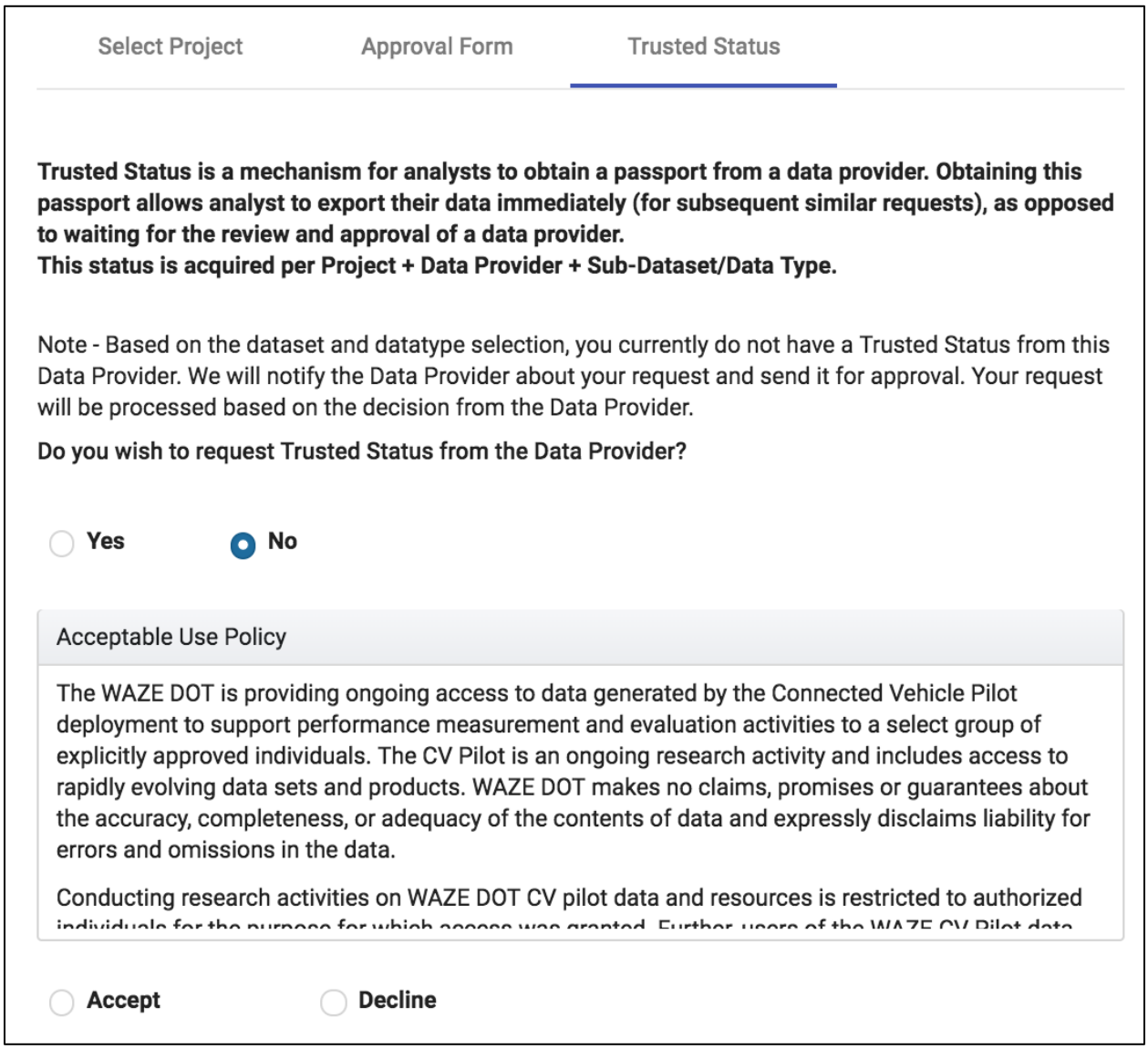


Figure 34: Acceptable Use Policy

5. Upon successful submission, the request will be sent to appropriate Data Providers. Data Providers will be responsible for accepting or rejecting the export requests.
6. Once Data Providers approve the request, Data Analysts will be able to download the dataset out of SDC through portal.

Chapter 7. Technical Documentation and Contact Information

The following sections provide technical resources for SDC users.

Architecture Diagram

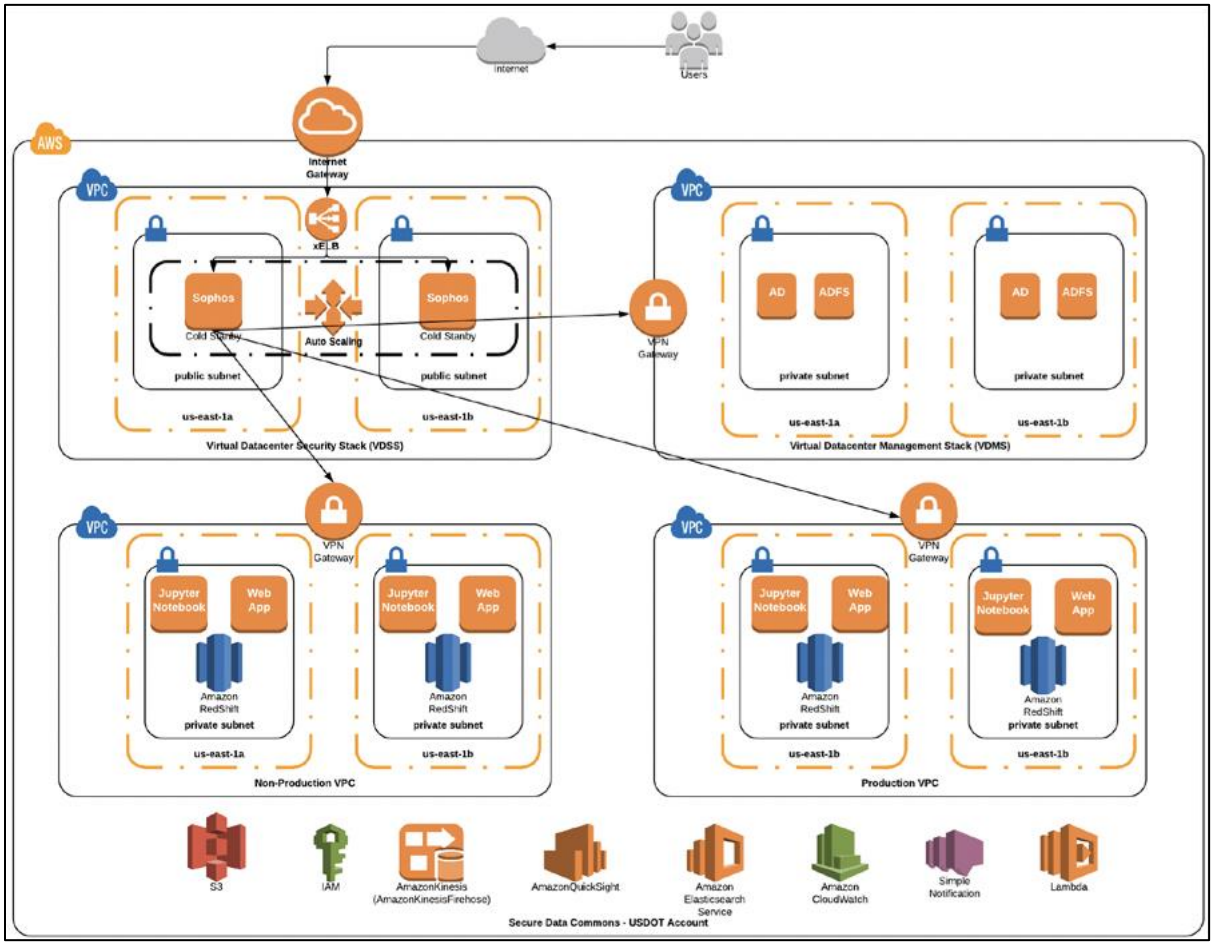


Figure 35: SDC Architecture Overview

Workstation Details

Table 1: 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.

Tools and Versions

Table 2: 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

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

Useful Links

S3: 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 EC2 instance being run inside SDC:

- **List objects in a bucket -** If there are any files at the bucket level then the below 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 -** The below command will list all the objects/files under that folder or prefix.
  aws s3 ls s3://<bucketName>/<prefix>/

- **Copy a local file to an s3 bucket -** This command will copy the file at the root of the bucket .
  aws s3 cp nohup.out s3://<bucketName>/
- **Copy a local file to a specific folder/prefix in an s3 bucket -** To create a new output folder and copy to that destination, simply add the desired path to the end, following bucket/.
  aws s3 cp <filename> s3://<bucketName>/<prefix>/
- **Get a file from S3 to the ec2 instance -** This command will help you get a file from S3 to the EC2 instance:
  aws s3 cp s3://<bucketName>/<prefix>/<fileName> .

Chapter 8. Frequently Asked Questions

- ***How can I get access to the SDC Datasets?***
- ***How will I understand what a particular dataset consists of?***
- ***How can I launch a workstation?***
- ***Where do I store my data?***
- ***How can I bring my own datasets/algorithm to my workstation?***
- ***How can I publish my dataset/algorithm?***

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 'Dataset' in the top menu.

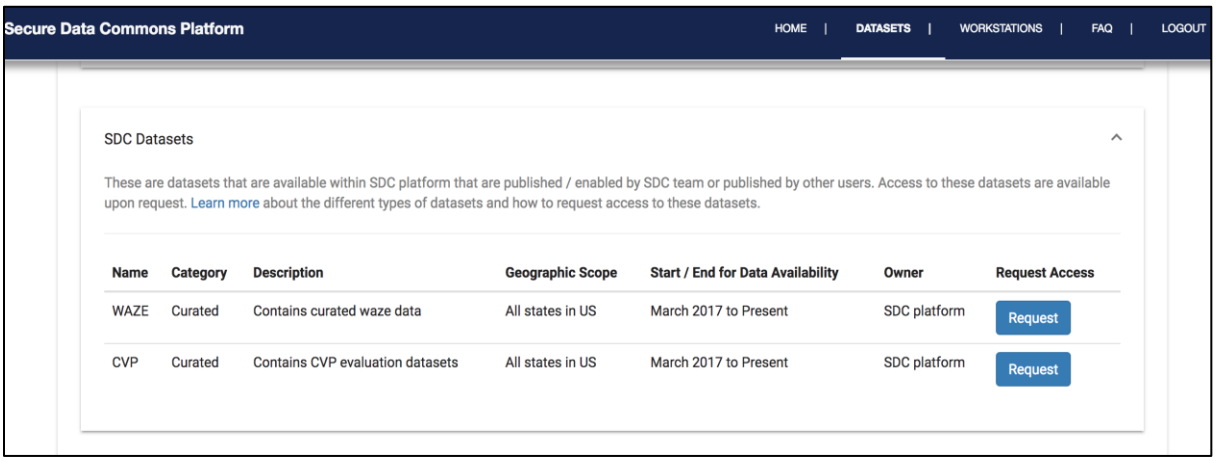


Figure 36: Datasets Tab

Click on 'SDC Dataset' to view all the available datasets.
To access a dataset, you need to click on the 'Request' button.
A form will pop up. Fill in that form and click on 'Send Request' button.

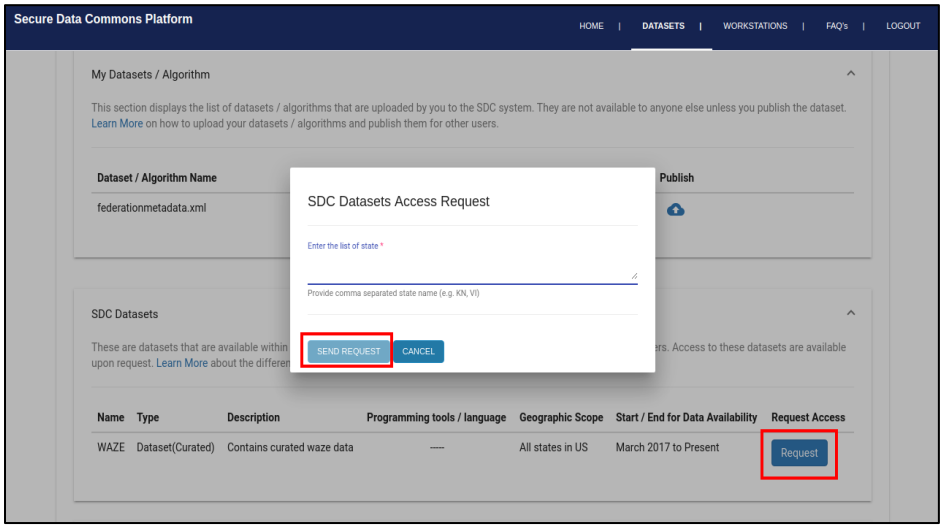


Figure 37: Dataset Access Request

The request will be sent to 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 'Workstation' and click on the 'Launch' button of any workstation you want to access.

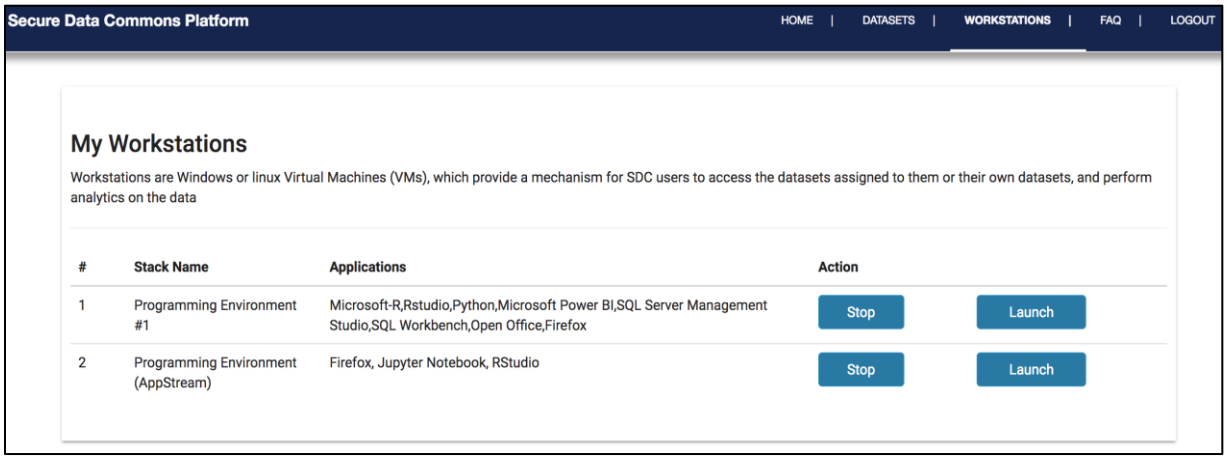


Figure 38: Workstations Tab

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

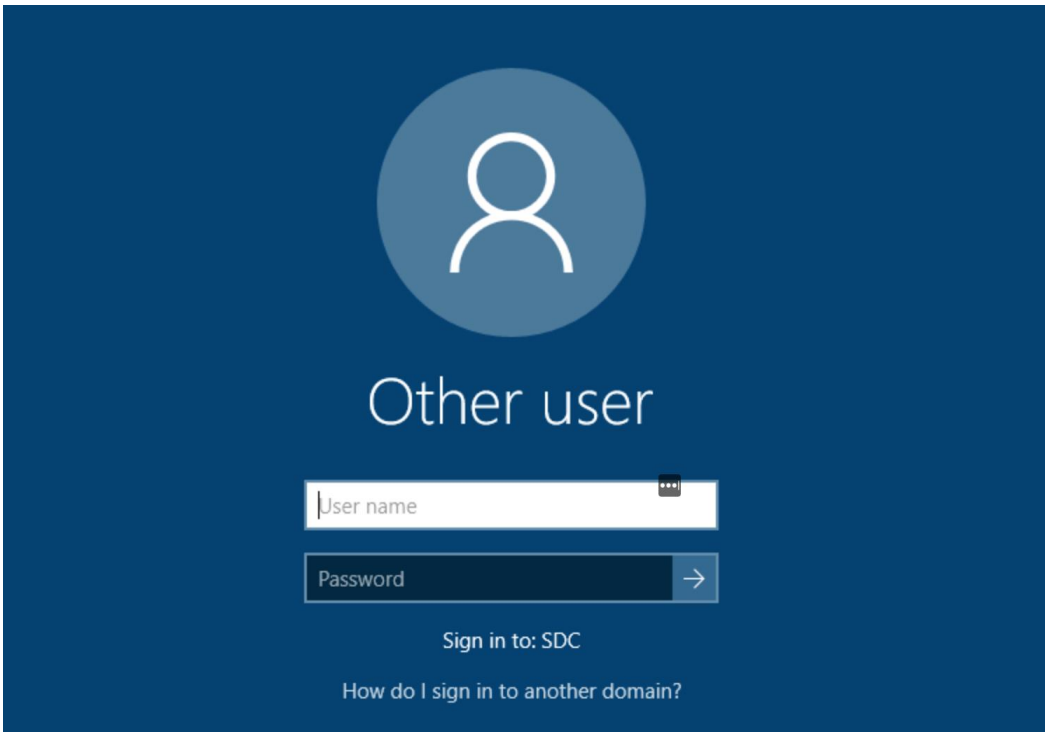


Figure 39: Login for Programming Environment

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 upload icon under publish for the dataset/algorithm you wish to publish.

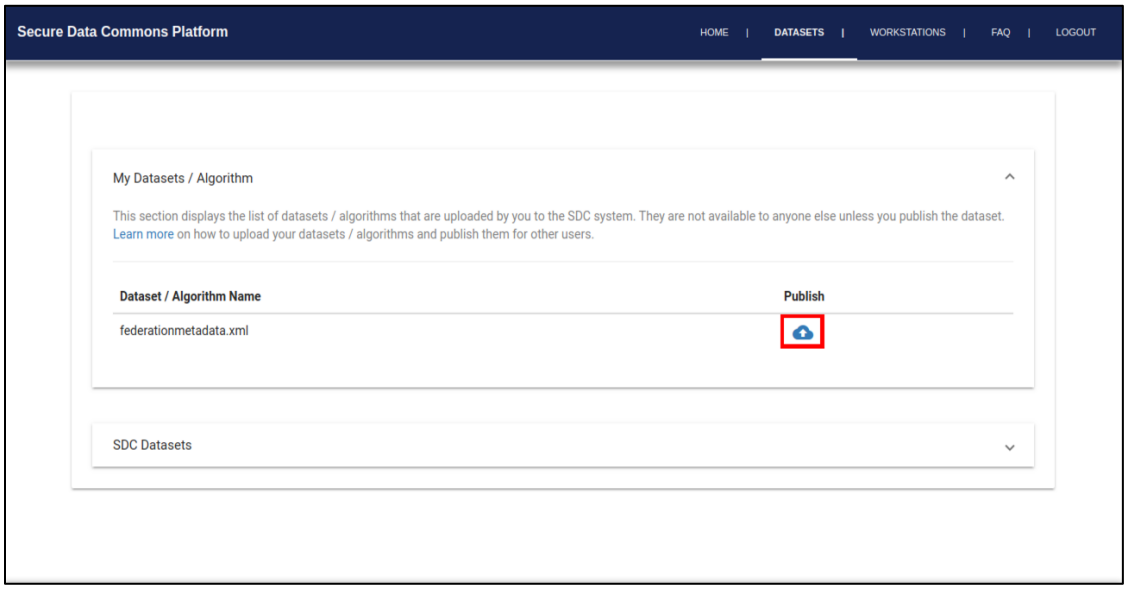


Figure 40: Upload Button

3. In the pop-up window, you will have two options, either a Dataset or Algorithm.
   a. Select the drop-down value as Dataset and fill in the required values.

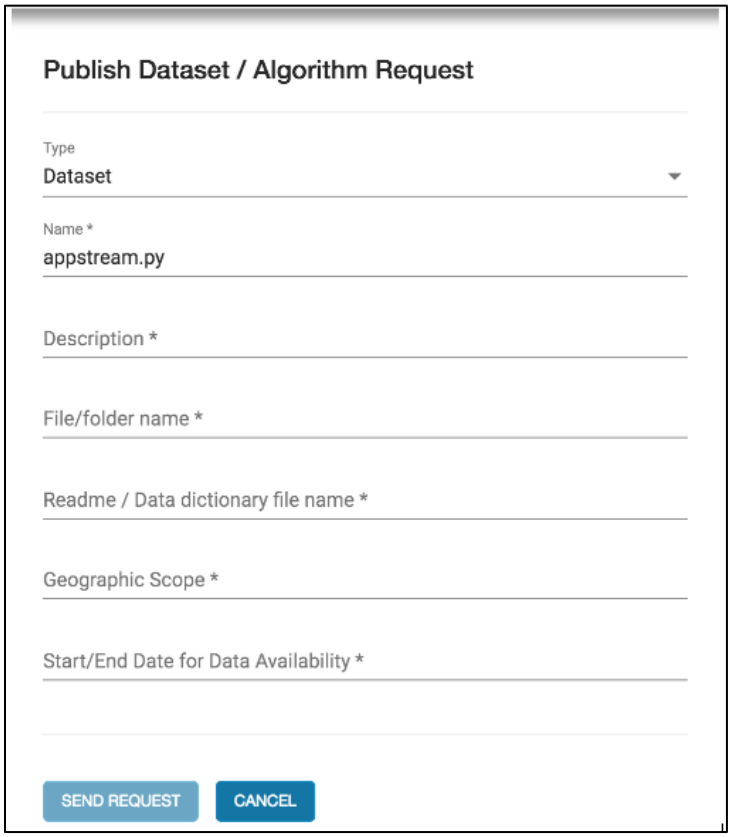


Figure 41: Publish Dataset Form

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. Select the drop-down value as Algorithm.

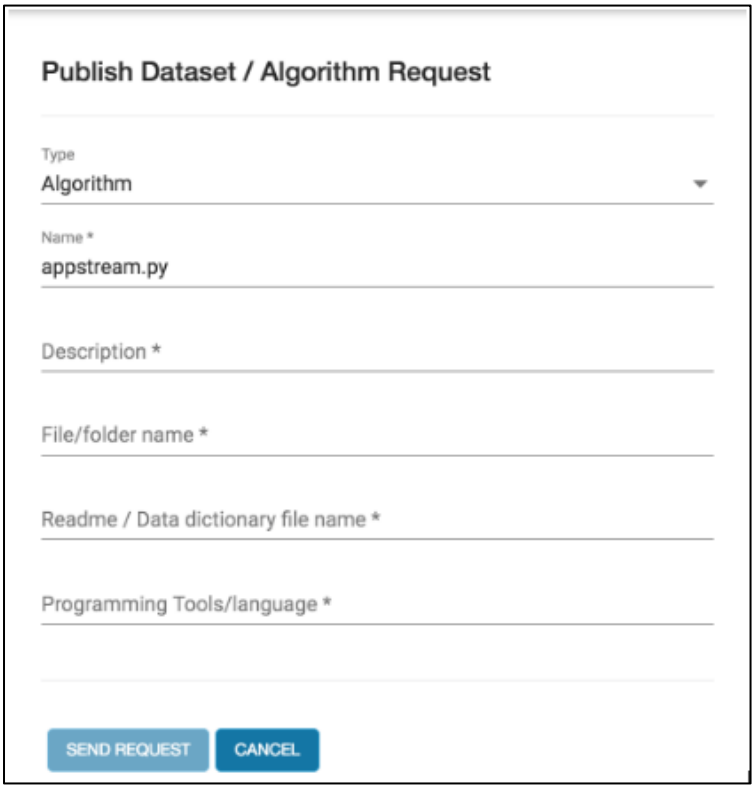


Figure 42: Algorithm Request Form

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.

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