icf.com

we are ICF

DOT FHWA
Secure Data Commons

09/15/2019

SDC Demo

Purpose

- To explain the SDC environment
- To provide a brief overview of the onboarding process
- Demo accessing SDC, importing data, analyzing data and exporting data
- Common FAQ

What is the SDC?

- **The Secure Data Commons (SDC) is an online data warehousing and analysis platform for transportation researchers. On this portal, researchers can take advantage of pre-established programming environments to access and analyze a growing set of transportation-related data sets.**
  - Provides secure access to data and enables the ability to conduct research and analysis on these data sets
  - Security of Data - Moderate level
  - Designed for analysis using programming and statistical tool packages
  - Analysis is performed within the SDC platform through cloud-based resources
- **The SDC platform is being developed as a collaborative environment for traffic engineers, researchers, data scientists, and anyone who is interested in carrying out research and analysis on different datasets related to traffic, weather, crashes, and others.**

Benefits

- Built in architecture for storing and managing data
- Built in architecture for data analyst teams roles
- Ability to rapidly access research data sets for analysis
  - Near-real time data flows
- Controlled access to data provides comfort for data providers.

Users

Projects

- A project is a pairing of Data Providers and Users
- Project Managers must evaluate costs from all three standpoints to assess their need to use the SDC. Costs are provided in three categories:
  - Cloud Consumption
  - IT Services
  - Enablement

Data Providers

- Data Providers can provide data in near-real-time, batch uploads, and ad-hoc uploads
- Data Providers can develop common data formats and fix issues during testing
- Data Providers define the terms of data access and can grant or deny access to specific users or groups.
- Data Providers can grant or deny access to what type of derived data is exported or copied from the system.

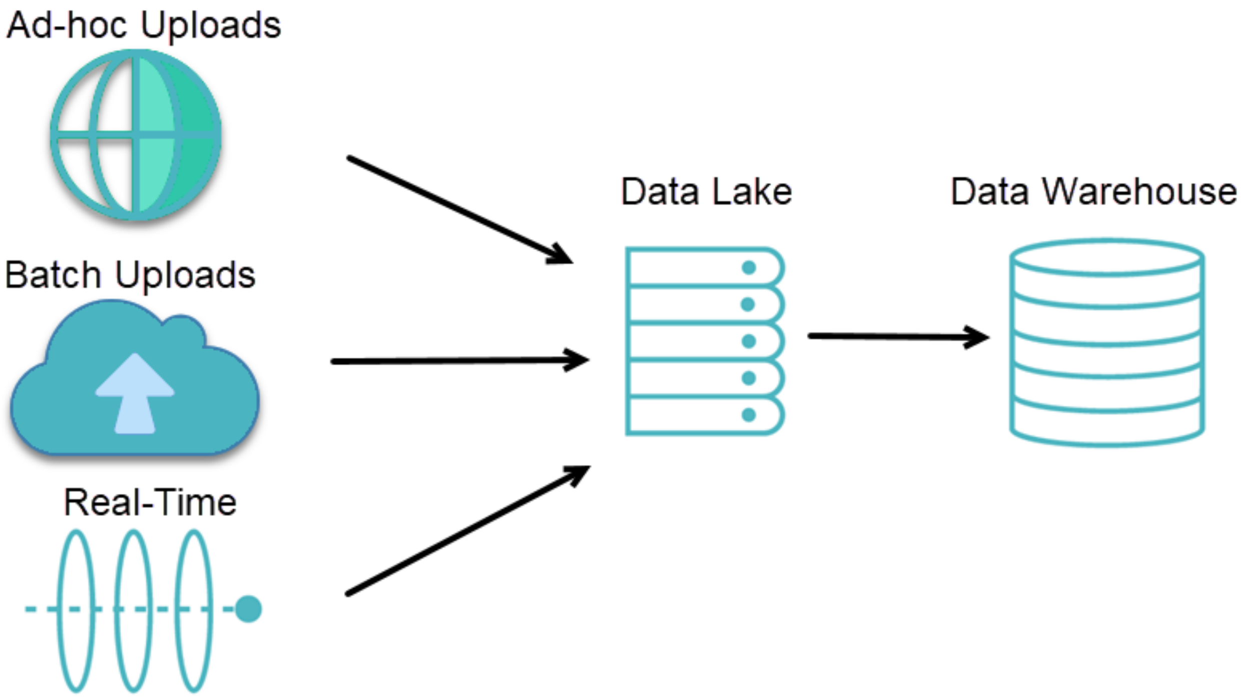
Data Analyst

- Data Analysts work within the SDC Analytic Sandbox. Each Analyst is provided a cloud-based workstation with pre-loaded programming environments and software. The workstations include access to data in the data lake and data warehouse. Data Analysts can…
  - share code and data with each other
  - upload their own datasets

DATA FLOW: Data Provider

- **Data Lake**
  - "Raw" data
  - Can be loosely structured
  - Variable frequency
  - AWS S3 buckets

- **Data Warehouse**
  - Curated data - "lightly" to "highly
  - Subset of data lake data
  - Various technologies

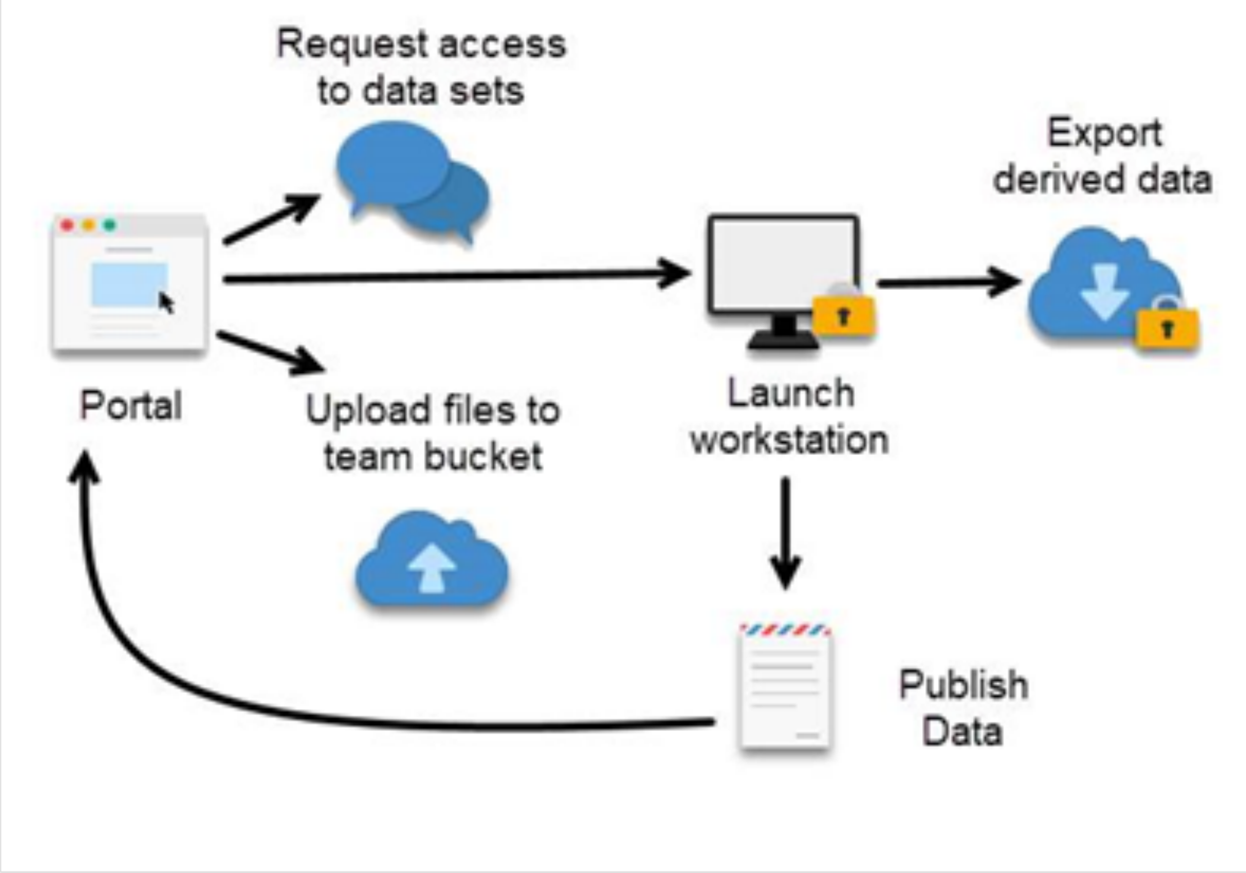


DATA FLOW: Data Analyst

Data Analysts are

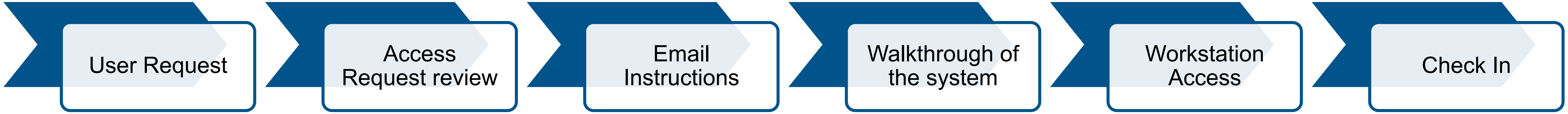
- Provisioned cloud-based workstations within the SDC
- Each workstation comes with pre-loaded programming environments and software
- The workstations include access to data in the data lake and data warehouse

Data Analyst can



Onboarding process

Onboarding a user to the SDC system involves the following steps:



User Request

Access Request review

Email Instructions

Walkthrough of the system

Workstation Access

Check In

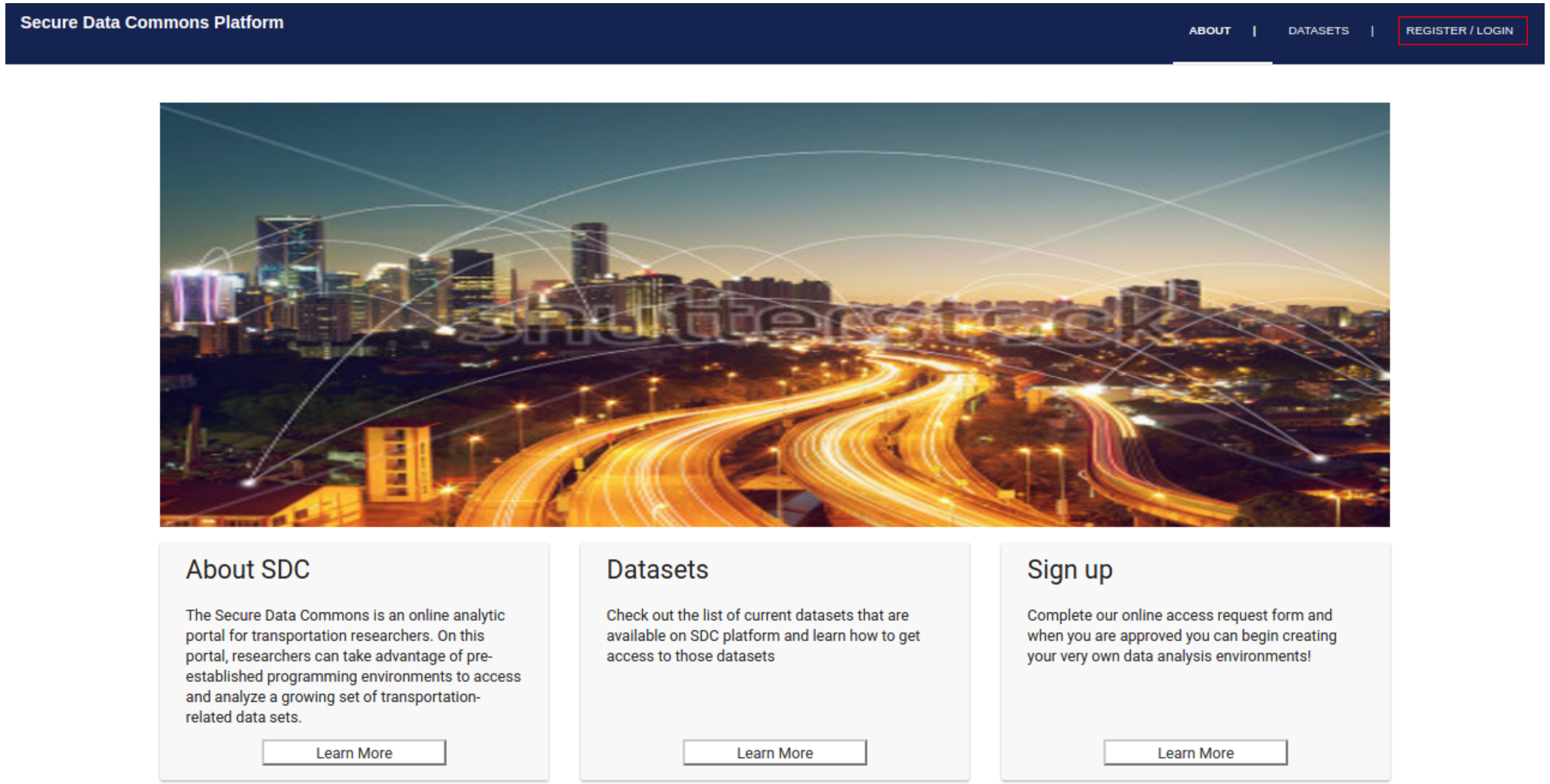
Detail onboarding instructions located: https://portal.securedatacommons.com/faqs

Demo – On SDC Workstation

- **Demo**
  - Accessing web portal
  - Workstations start/launch/stop
  - Import data
  - Analyze data
  - Export data
  - Collaborate
  - Applications available

Access portal

- Go to https://portal.securedatacommons.com
- Go to ‘Register/Login’ menu tab



About SDC

The Secure Data Commons is an online analytic portal for transportation researchers. On this portal, researchers can take advantage of pre-established programming environments to access and analyze a growing set of transportation-related data sets.

Datasets

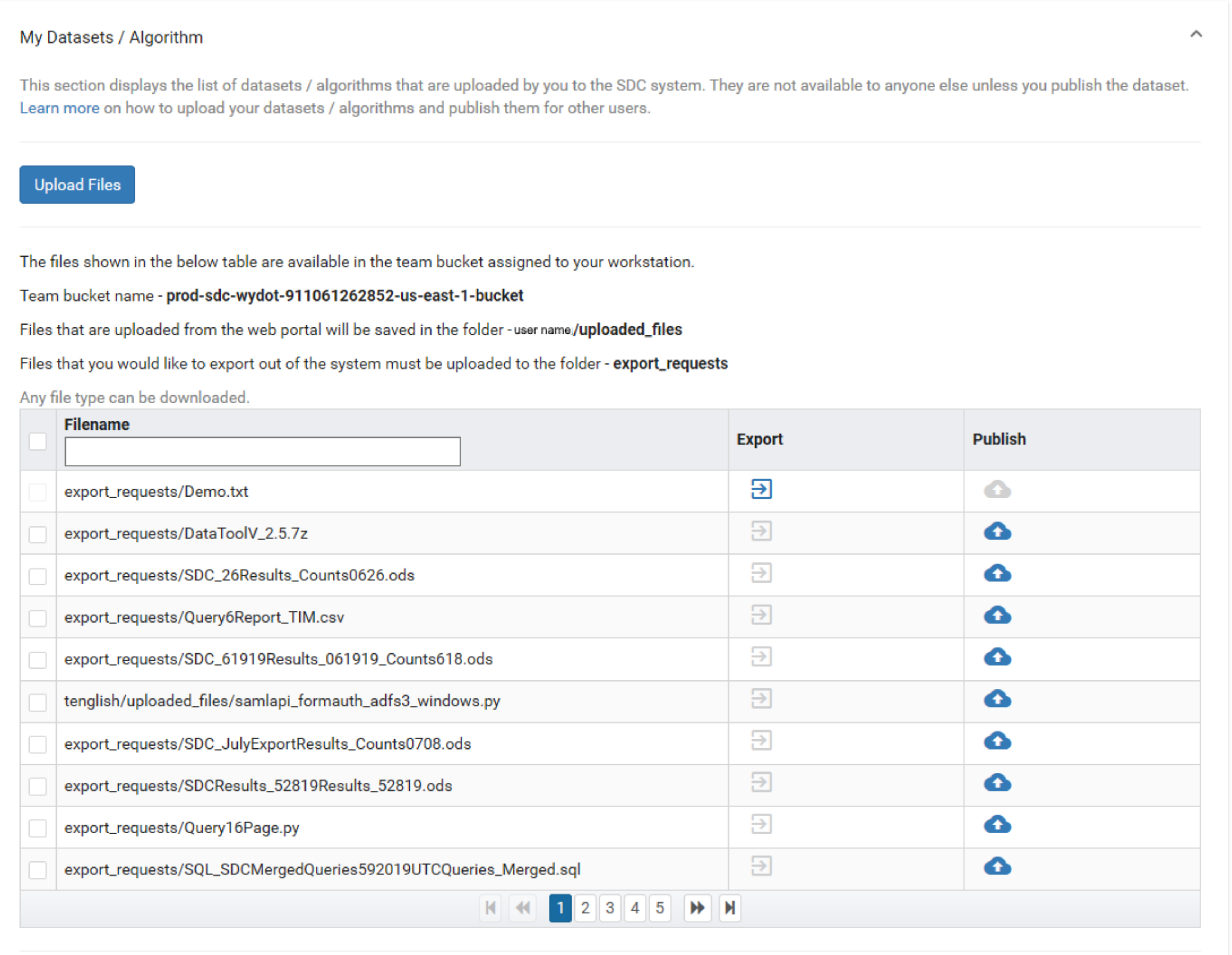
Check out the list of current datasets that are available on SDC platform and learn how to get access to those datasets

Sign up

Complete our online access request form and when you are approved you can begin creating your very own data analysis environments!

Import data

- **Select data to import**
  - Curated datasets
  - Raw datasets
  - Published datasets
- **Verify through s3 browser**



My Datasets / Algorithm

This section displays the list of datasets / algorithms that are uploaded by you to the SDC system. They are not available to anyone else unless you publish the dataset. Learn more on how to upload your datasets / algorithms and publish them for other users.

Upload Files

The files shown in the below table are available in the team bucket assigned to your workstation.

Team bucket name - **prod-sdc-wydot-911061262852-us-east-1-bucket**

Files that are uploaded from the web portal will be saved in the folder - user name**/uploaded_files**

Files that you would like to export out of the system must be uploaded to the folder - **export_requests**

Any file type can be downloaded.

□	Filename	Export	Publish
□	export_requests/Demo.txt	⇥	☁
□	export_requests/DataToolV_2.5.7z	⇥	☁
□	export_requests/SDC_26Results_Counts0626.ods	⇥	☁
□	export_requests/Query6Report_TIM.csv	⇥	☁
□	export_requests/SDC_61919Results_061919_Counts618.ods	⇥	☁
□	tenglish/uploaded_files/samlapi_formauth_adfs3_windows.py	⇥	☁
□	export_requests/SDC_JulyExportResults_Counts0708.ods	⇥	☁
□	export_requests/SDCResults_52819Results_52819.ods	⇥	☁
□	export_requests/Query16Page.py	⇥	☁
□	export_requests/SQL_SDCMergedQueries592019UTCQueries_Merged.sql	⇥	☁

Validation of data – Canary Function

- Canary function
  - What is this?

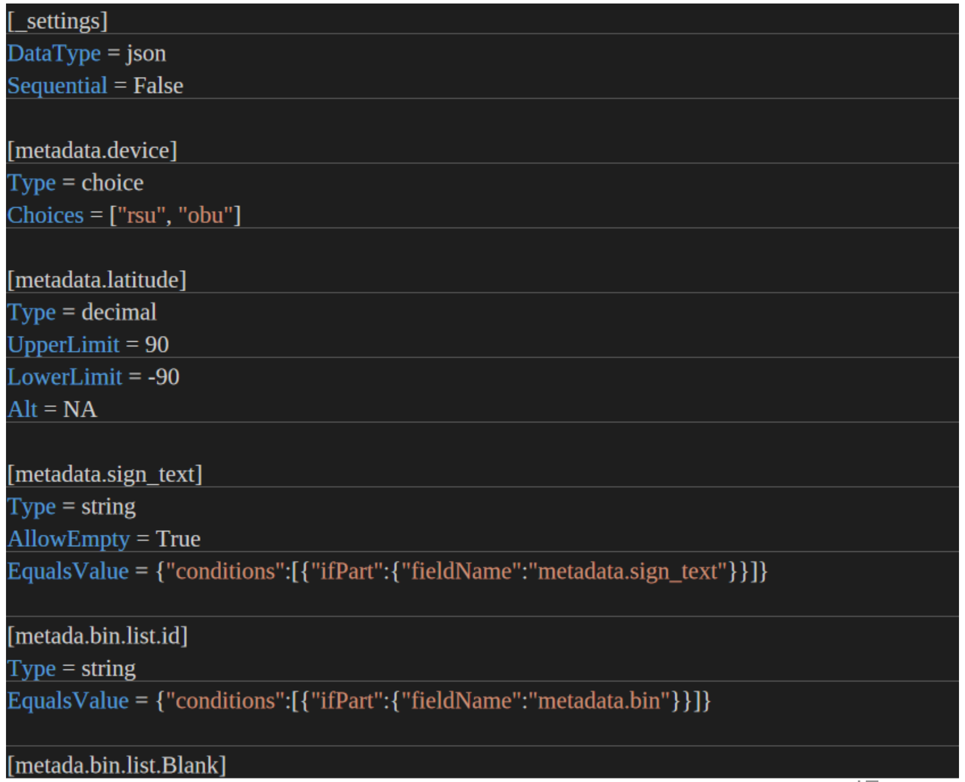
  A sanitization process that validates the content of data provided by Data Providers by using client generated configuration files.

  - Why is it important?

  This process will find problems within data generation on the Data Providers and in the SDC infrastructure.

  - What is a sample .ini/validation file? (CSV/JSON)
  - Validate data or failed data imports?

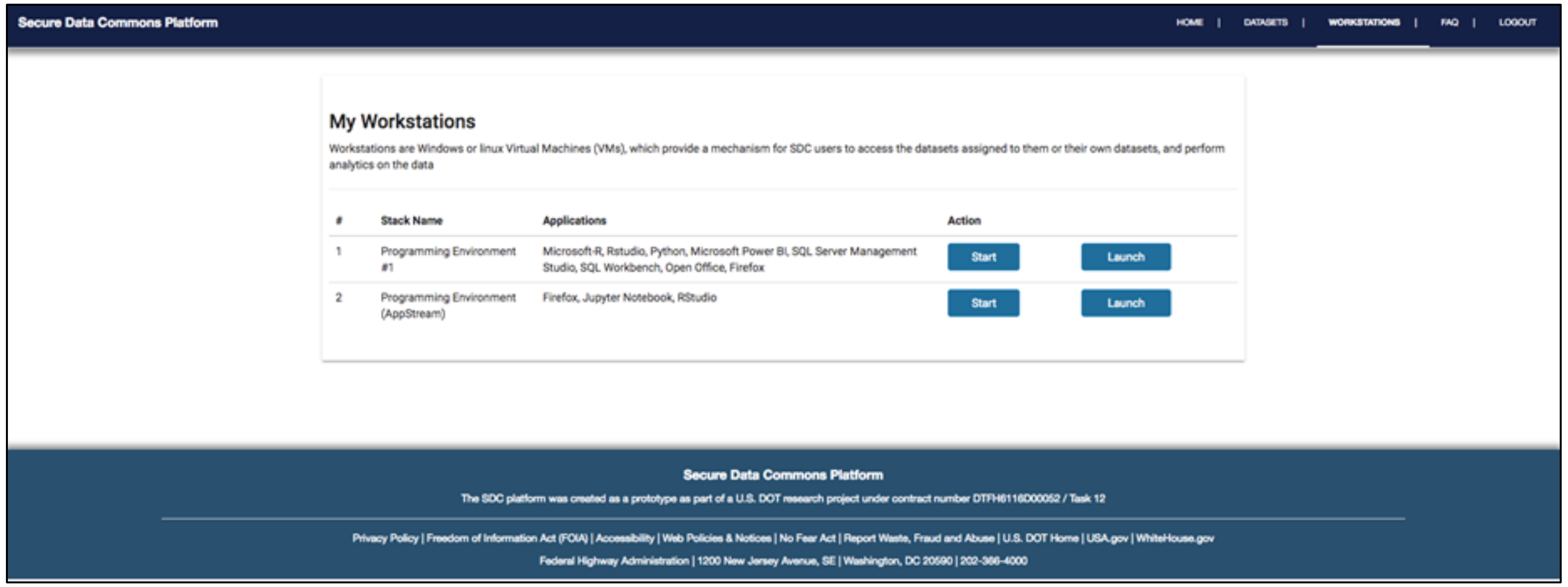
https://console.aws.amazon.com/cloudwatch/home?region=us-east-1#dashboards:name=prod-validator-summary



```
[_settings]
DataType = json
Sequential = False

[metadata.device]
Type = choice
Choices = ["rsu", "obu"]

[metadata.latitude]
Type = decimal
UpperLimit = 90
LowerLimit = -90
Alt = NA

[metadata.sign_text]
Type = string
AllowEmpty = True
EqualsValue = {"conditions":[{"ifPart":{"fieldName":"metadata.sign_text"}}]}

[metada.bin.list.id]
Type = string
EqualsValue = {"conditions":[{"ifPart":{"fieldName":"metadata.bin"}}]}

[metada.bin.list.Blank]
```

Analysis

- **Upload your own data**
  - For individual or team use
- **Derived Data**
  - Results of analysis
  - Can be shared for individual or team use
  - Can request exports from Data Provider
- **Analytical Tools**
  - Virtual Machine Instances
    - Can easily scale up for bigger analytical problems
  - Standard Software
    - E.g. Python, RStudio, SQL
  - Analyst Specific Software


Secure Data Commons Platform
HOME | DATASETS | WORKSTATIONS | FAQ | LOGOUT

My Workstations

Workstations are Windows or linux Virtual Machines (VMs), which provide a mechanism for SDC users to access the datasets assigned to them or their own datasets, and perform analytics on the data

#	Stack Name	Applications	Action	
1	Programming Environment #1	Microsoft-R, Rstudio, Python, Microsoft Power BI, SQL Server Management Studio, SQL Workbench, Open Office, Firefox	Start	Launch
2	Programming Environment (AppStream)	Firefox, Jupyter Notebook, RStudio	Start	Launch

Secure Data Commons Platform

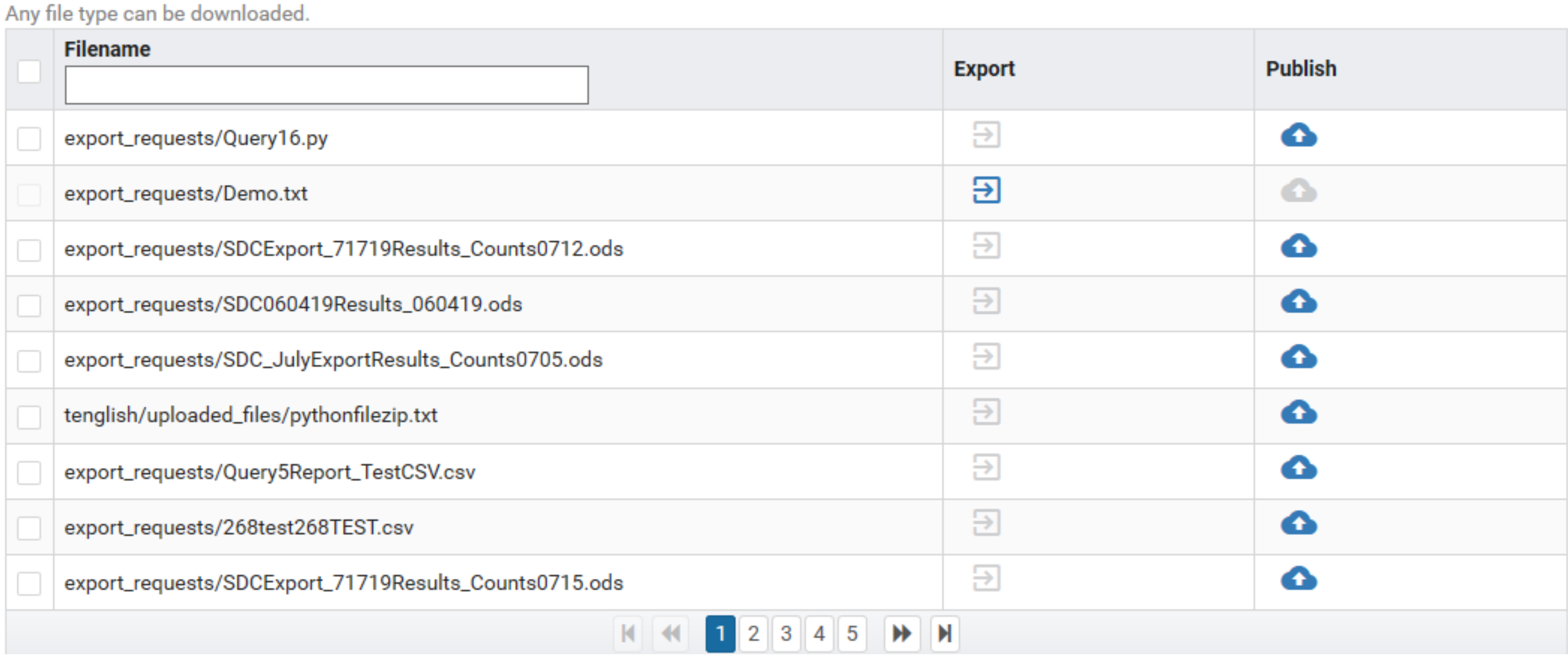
The SDC platform was created as a prototype as part of a U.S. DOT research project under contract number DTFH6116D00052 / Task 12

Privacy Policy | Freedom of Information Act (FOIA) | Accessibility | Web Policies & Notices | No Fear Act | Report Waste, Fraud and Abuse | U.S. DOT Home | USA.gov | WhiteHouse.gov

Federal Highway Administration | 1200 New Jersey Avenue, SE | Washington, DC 20590 | 202-366-4000

Export data

- Select data to export
- Export from portal
- View exported data



Any file type can be downloaded.

□	Filename	Export	Publish
□	export_requests/Query16.py	⇥	☁
□	export_requests/Demo.txt	⇥	☁
□	export_requests/SDCExport_71719Results_Counts0712.ods	⇥	☁
□	export_requests/SDC060419Results_060419.ods	⇥	☁
□	export_requests/SDC_JulyExportResults_Counts0705.ods	⇥	☁
□	tenglish/uploaded_files/pythonfilezip.txt	⇥	☁
□	export_requests/Query5Report_TestCSV.csv	⇥	☁
□	export_requests/268test268TEST.csv	⇥	☁
□	export_requests/SDCExport_71719Results_Counts0715.ods	⇥	☁

Management Tools

- **Confluence Project Dashboard**
  - Backlog to Users Stories, features
  - Meetings
  - Personas
- **High Level Roadmap**
- **Support Functionality**
  - Quarterly Newsletter/Meetings
  - Release Notes
  - Frequently Asked Questions
  - Support Tickets sent to SDC Admin email and process in our JIRA visible to you as the end users

Q&A and FAQ

- Why cant I access my favorite websites?
- Why cant I cut & paste
- Why cant I go to Git Hub?

Other FAQ:
https://portal.securedatacommons.com/faqs

Secure Data Commons Website

- Learn more about the SDC from our Website
- https://its.dot.gov/data/secure/index.html

Questions