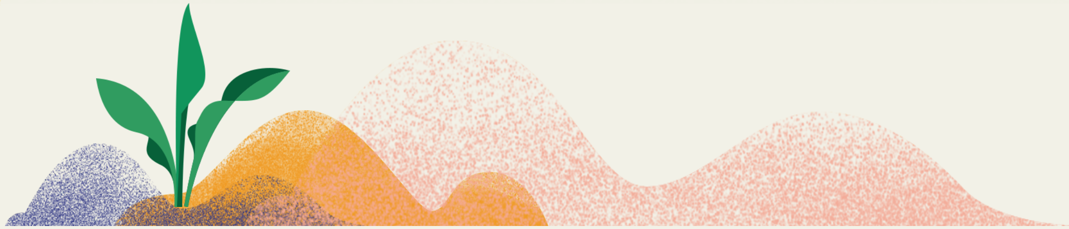
Secure Data Commons 101

U.S. Department of Transportation
Office of the Chief Information Officer

Agenda

1. Overview of the SDC System
2. Project Lifecycle
3. SDC Walkthrough
4. Our Commitment to you
5. Questions & Feedback



1

SDC Overview

2

Project Lifecycle

3

SDC Walkthrough

4

Our Commitment

5

Questions

What is the SDC?

The USDOT Secure Data Commons (SDC) enables collaborative and controlled integration and analysis of research data at the moderate sensitivity level, including personally identifiable information (PII) and confidential business information (CBI). With three types of data transfer updates supported: real-time (streaming), batch (daily, weekly), and ad-hoc (occasional), the SDC offers authorized and controlled access to individual datasets - as well as the metadata - associated with those datasets.

SDC Vision

- The Secure Data Commons will be used as the premier data collaboration source for the transportation challenges we face in the 21st Century
- The SDC will help users collaborate by leveraging open-source code and achieve meaningful insights from their research.

SDC Goals

- Bring together transportation experts, data scientists, and other expert users from academia, industry, federal agencies, state and local governments.
- Share common data sources and help our partners utilize our open-source code.
- Provide new perspectives or analyses that go across modes or a program's portfolio.

Benefits & Use Cases of the SDC

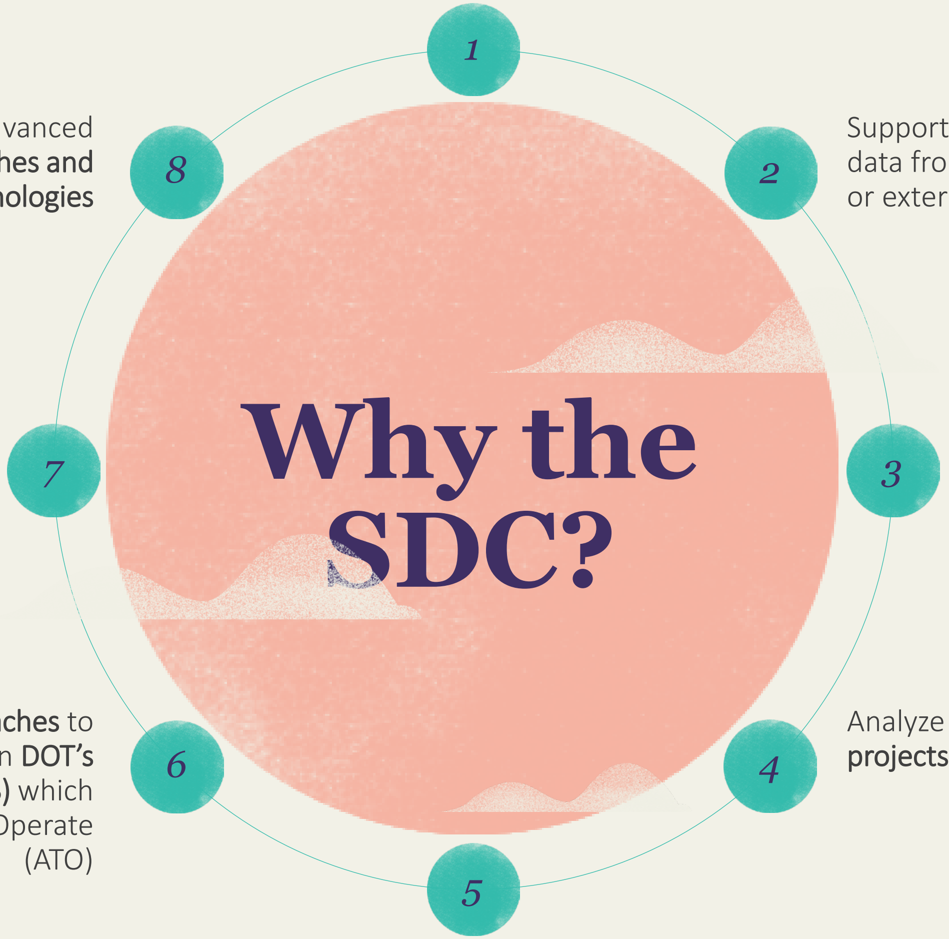
1 SDC Overview

2 Project Lifecycle

3 SDC Walkthrough

4 Our Commitment

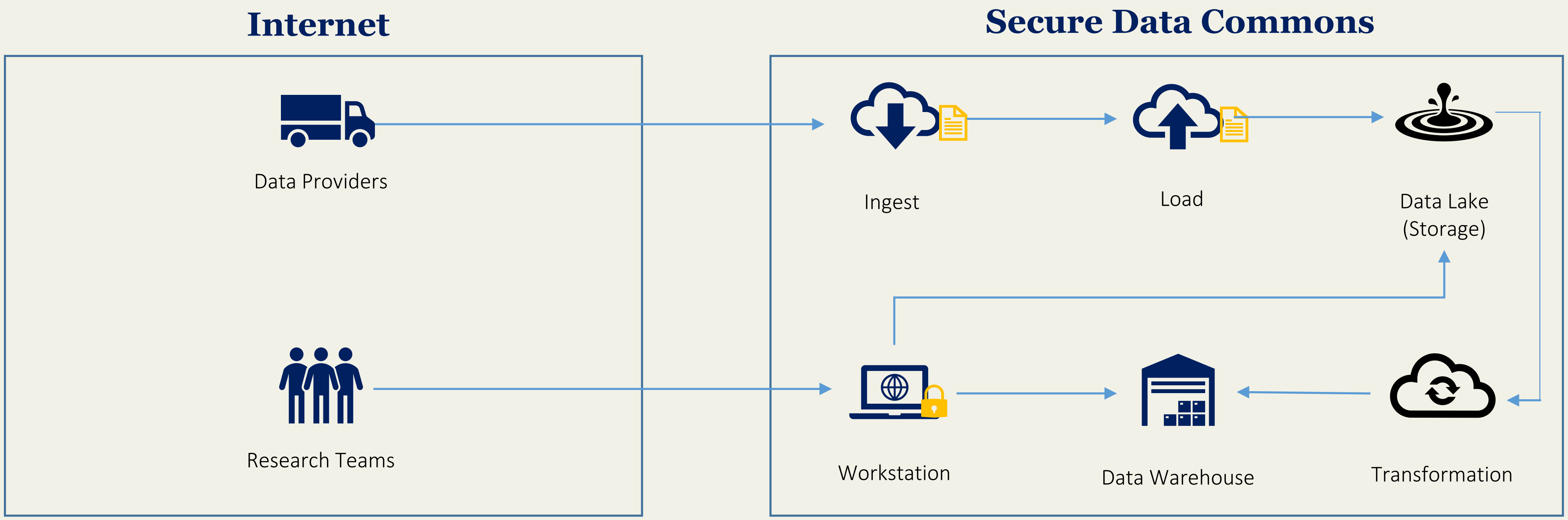
5 Questions



1. For External Analysts to **collaborate** with DOT Analysts using the same data and same analysis tools (DOT Analysts collaborate with External Analysts to gain insights on similar data sets and analysis tools)
2. Support a program's mission by combining data from a **diverse set of sources** (internal or external to DOT)
3. Store and analyze sensitive data with **fine-grained access control**
4. Analyze problems or **gain insights across projects, programs, or modes**
5. Easily scale up and down computational and data storage in **a transparent and cost-effective** manner
6. **Prototype new approaches** to identifying data trends within **DOT's Enterprise Cloud Services (ECS)** which already has an Authority to Operate (ATO)
7. **Leverage cloud capabilities** to share complex transportation datasets with the transportation research community
8. Train staff on the advanced **analytical approaches and technologies**

Example of SDC Architecture

We bring in data from the following data providers into the SDC, using the repeatable pipeline architecture below.



Data Providers:

1. Waze for Cities Program
2. Highway Safety information Systems (FHWA HSIS)
3. Federal Aviation Administration Genomics (FAA)

Ingests and Datasets

1. *Data Ingest: Upload to secure AWS S3 buckets (Data Storage)*
2. *SFTP File Transfer - securely transfer, access, and manage files*
3. *Data Lake: AWS Lambda (Data Transformation) then moves data from ingest to the data lake (Data Storage)*

1

SDC Overview

2

Project Lifecycle

3

SDC Walkthrough

4

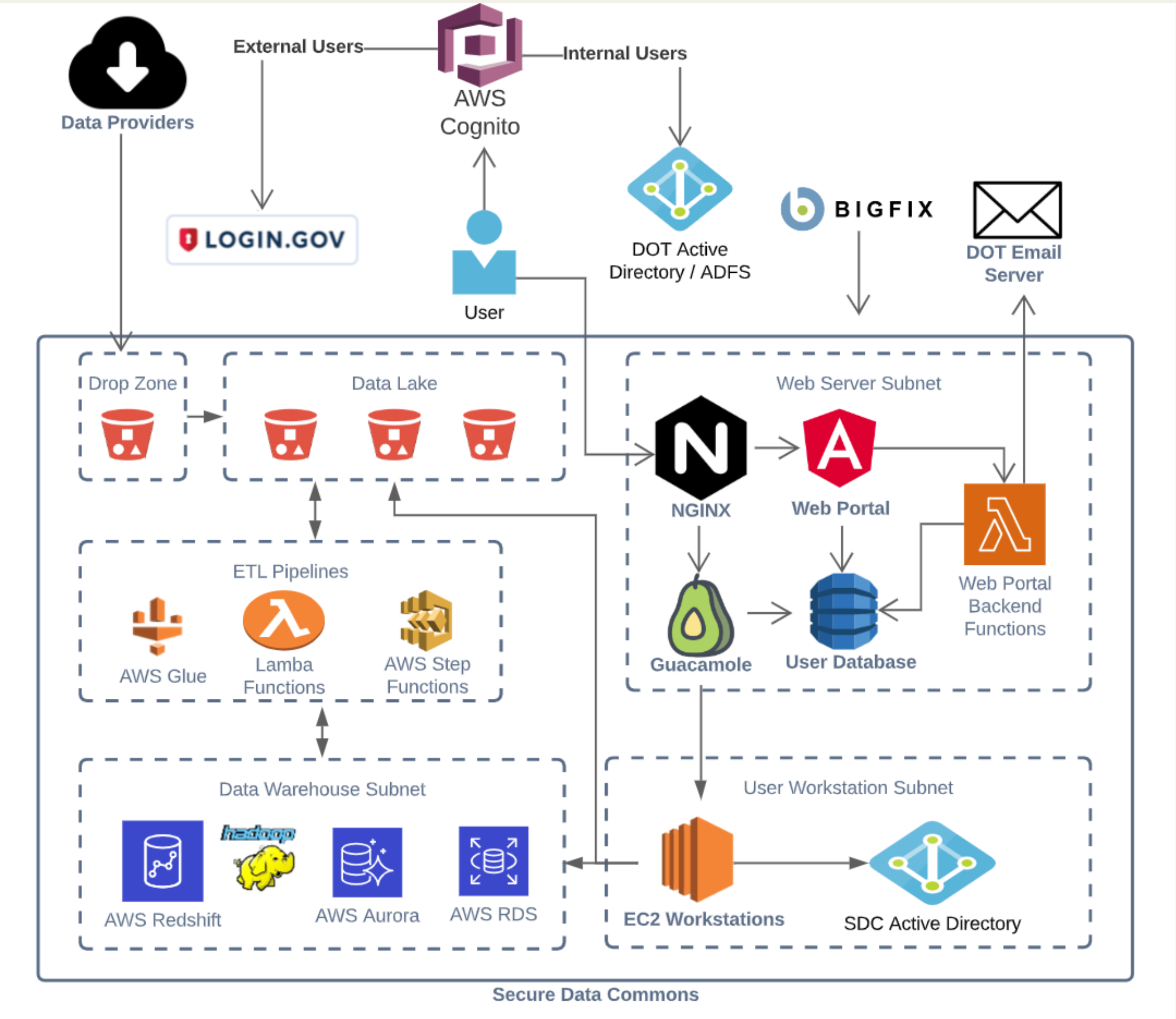
Our Commitment

5

Questions

SDC Platform

SDC enables collaborative and controlled integration and analysis of research data at the moderate sensitivity level, including personally identifiable information (PII) and confidential business information (CBI).



SDC Key Architectural Components

- User Authentication
  - DOT Active Directory or Login.gov (for external DOT users) authentication required to access the SDC platform
  - Local SDC Active Directory authentication required to access workstation
- Resource Isolation
  - Multiple subnets to keep different types of resources separate with only minimally required routes allowed
  - Restricted Internet access for user workstations to greatly reduce threat vectors and data leakage
- Transportation Data Pipeline
  - Standard and Custom Data Pipelines to perform customer-specific ETL logic, data curation, and QA/QC processes
- Multiple Data Warehouse Options
  - Most appropriate solution for each project is selected based on dataset size, customer experience/preferences, and best technical fit
- Self-service Web Portal
  - User workstation access and resizing
  - Access request for data sets and export requests for derived data sets
  - Import data sets into SDC

1

SDC Overview

2

Project Lifecycle

3

SDC Walkthrough

4

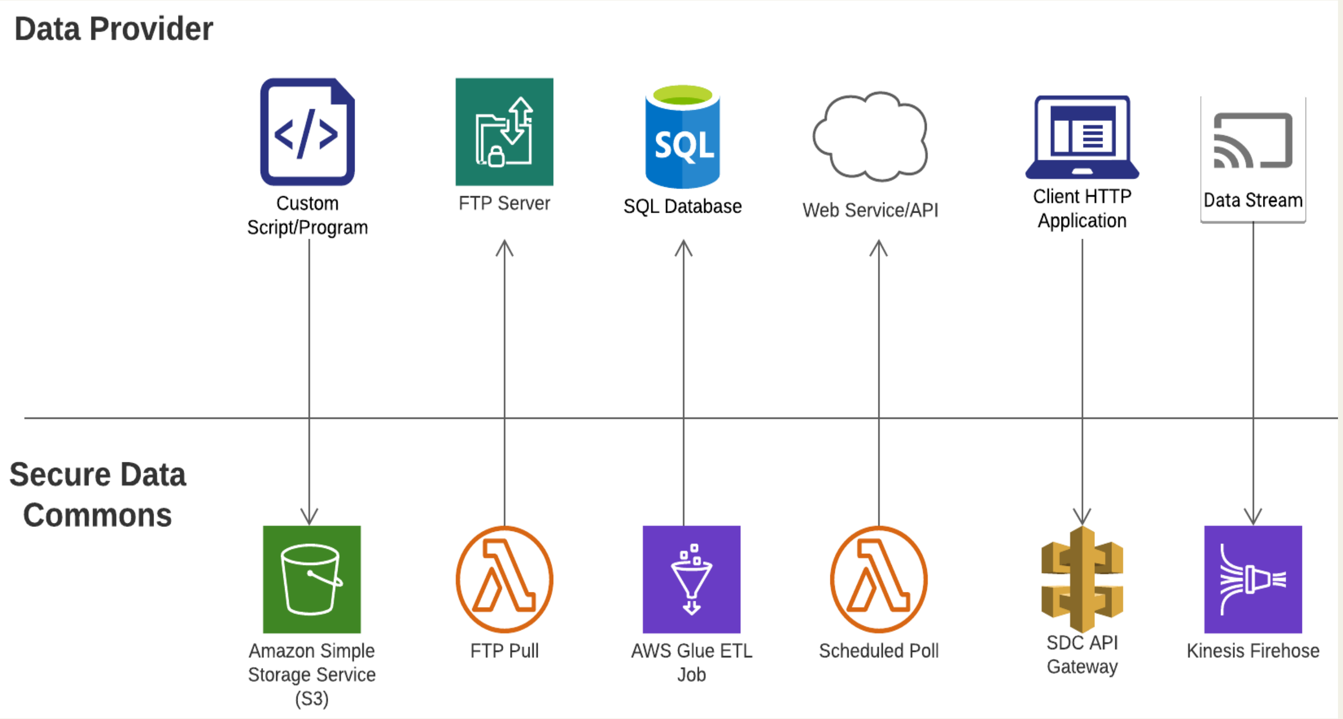
Our Commitment

5

Questions

SDC Dataset Ingestion

Data ingestions are built using cloud native technologies that are FedRAMO approved and FISMA moderate compliant



Data Provider

Custom Script/Program

FTP Server

SQL Database

Web Service/API

Client HTTP Application

Data Stream

Secure Data Commons

Amazon Simple Storage Service (S3)

FTP Pull

AWS Glue ETL Job

Scheduled Poll

SDC API Gateway

Kinesis Firehose

- Data Provider may push to an S3 bucket using a **Custom Script/Program**
- SDC can download from an **FTP Server**
- SDC can read-only connect to a customer's **SQL Database** to download records
- SDC can invoke a **Web Service/API** at scheduled intervals to download data
- Conversely, a data provide my upload data with an **HTTPS Client** to SDC's APIs
- Real time **Data Streaming** can be accommodated using AWS Kinesis Firehose

1

SDC Overview

2

Project Lifecycle

3

SDC Walkthrough

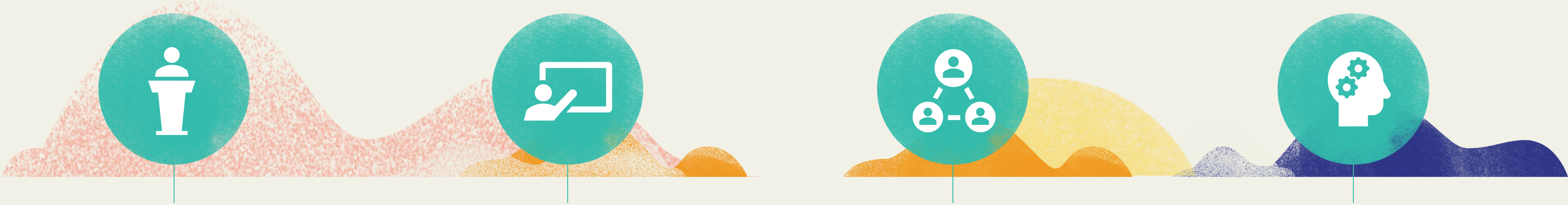
4

Our Commitment

5

Questions

SDC Roles



Project Owners

- The person or organization that has the authority, ability, and responsibility to access, create, modify, store, use, share, and protect data. Project Owners have the right to delegate these privileges and responsibilities to other parties.

Data Stewards

- At the direction of the Project Owner, the Data Steward is a person who is delegated the privileges and responsibilities to manage, control, and maintain the quality of a data asset throughout the data life cycle.

Data Providers

- An individual or team that collects, prepares and/or submits research datasets hosted on the SDC platform. The Data Provider establishes the data protection needs and acceptable use terms for the data analysts.

Research Analysts

- An individual or team 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.

1

SDC Overview

2

Project Lifecycle

3

SDC Walkthrough

4

Our Commitment

5

Questions

Data Steward Responsibilities

The person or organization that is delegated the privileges and responsibilities to manage, control, and maintain the quality of a data asset throughout the data life cycle.

Overview

At the direction of the Project Owner, the Data Steward is a person who is delegated the privileges and responsibilities to manage, control, and maintain the quality of a data asset. The Data Steward may utilize their ability to appropriate protections, restrictions, and other safeguards depending on the nature of the data. In addition, Data Steward is responsible for authorizing access to the data, ensuring appropriate actions and restrictions, authorizing the export of data, and establishing data retention policies that govern when data that is no longer of practical use can be archived or removed from the SDC.

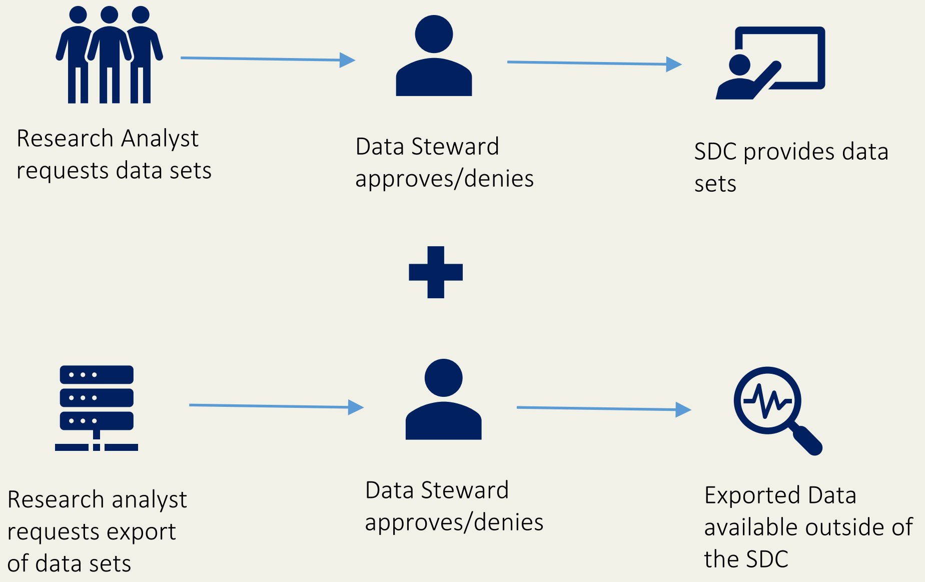
Goals

Ensure the project's data is protected by enforcing data agreements and reviewing access requests to it

Needs

- Understand and enforce the data agreements
- Control access to data by reviewing requests to access and export data
- Ability to monitor access and usage of data

Process



Research Analyst requests data sets → Data Steward approves/denies → SDC provides data sets

+

Research analyst requests export of data sets → Data Steward approves/denies → Exported Data available outside of the SDC

Data Provider Responsibilities

An individual or team that collects, prepares and/or submits research datasets hosted on the SDC platform.

Overview

An individual or team that collects, prepares and/or submits research datasets hosted on the SDC platform. The Data Provider establishes the data protection needs and acceptable use terms for the Research Analysts.

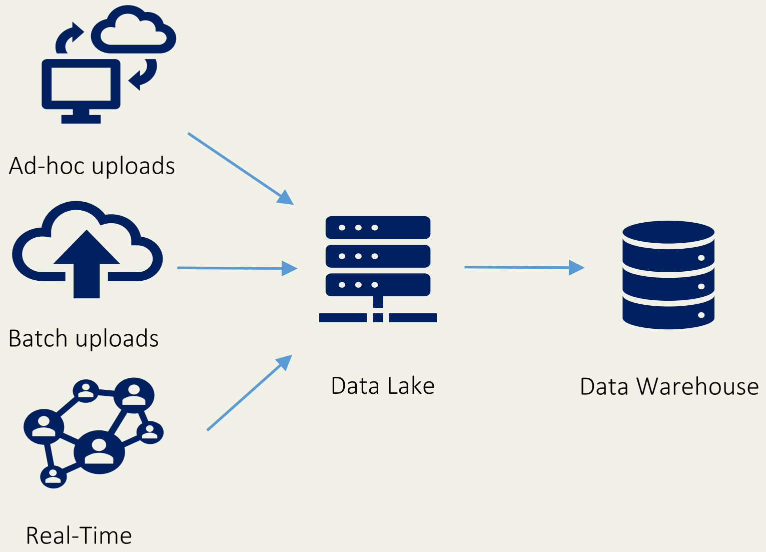
Goals

Get the data to SDC in a way that is comprehensive with the correct levels of access and data definitions with minimal effort

Needs

- Understand how SDC can be used for the project to make sure the right data is made available
- Define metadata, data rules, and agreements so that access to data can be controlled
- Make data available to directly ingest into SDC with minimal effort
- Define access levels to data
- Help ensure data quality

Process



Research Analyst Responsibilities

An individual or team that conduct analysis in SDC.

Overview

The individual or team that conduct complex analysis using the datasets hosted within the SDC system. Note that Research Analysts can bring their own data and tools into the SDC.

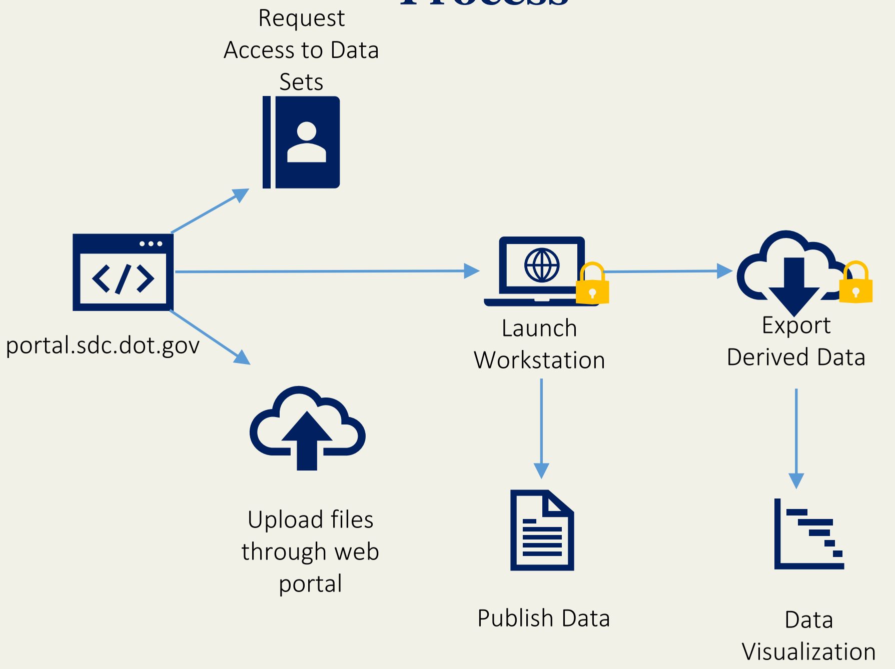
Goals

Use the data and tools available in SDC to create meaningful insights that can be used to inform data-driven research and/or policy.

Needs

- Easy access to data and other well curated data sets that can assist in analysis
- Ability to collaborate and share data, code, and analyses with another Research Analyst team member and outside their team within the SDC
- Easy access to tools that Research Analysts are familiar with so there is no need to learn a new set of tools
- Ability to export results of analysis so that Research Analysts can share with others or publish outside SDC
- Ability to scale computational workstations

Process



1 SDC Overview

2 Project Lifecycle

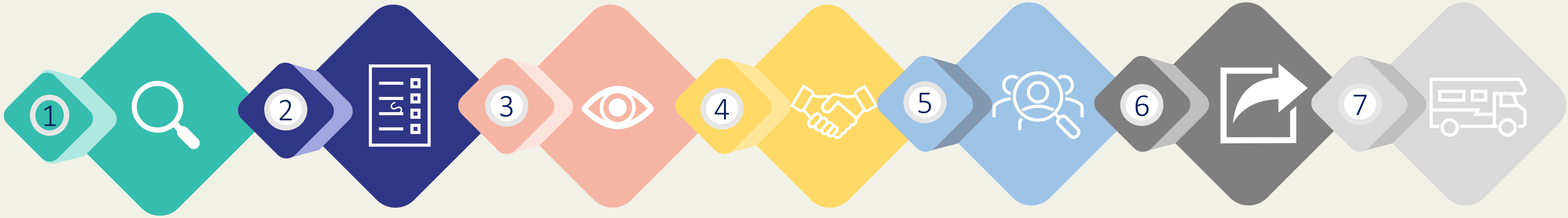
3 SDC Walkthrough

4 Our Commitment

5 Questions

Project Lifecycle

The SDC Team is here to serve you throughout all your Project Lifecycle Needs



Prospective

SDC team to work with research projects who identify the SDC as a potential location for analysis or storage

Projects: OST FLOW

Plan

Project Sponsors work with the SDC team to understand how the SDC can help the Prospective Project objectives, timelines, and needs

Discover

Project Sponsors work with the SDC team to detail requirements for the project, including gathering data set documentation

Onboard

Project Sponsors collaborate with SDC to implement defined requirements for Research analysts to ensure integration and data analysis are fully functional

Active

Project Sponsor and their team utilize the SDC platform to achieve research objectives

Projects: HSIS, FAA Genomics

Export

Project Sponsor and the SDC team work together to conclude active research work in the SDC which allows for any residual analysis to be done (No new data will be coming in from data providers).

Retire

SDC Team archives project data as per records retention schedule and final closeout

Projects: CVP, BTS-NOAA, WAZE-SDI, WAZE-COVID, WAZE-BTS, WAZE Academic, WAZE-SLG, FRA ARDS, OSS4ITS, CARMA

1

SDC Overview

2

Project Lifecycle

3

SDC Walkthrough

4

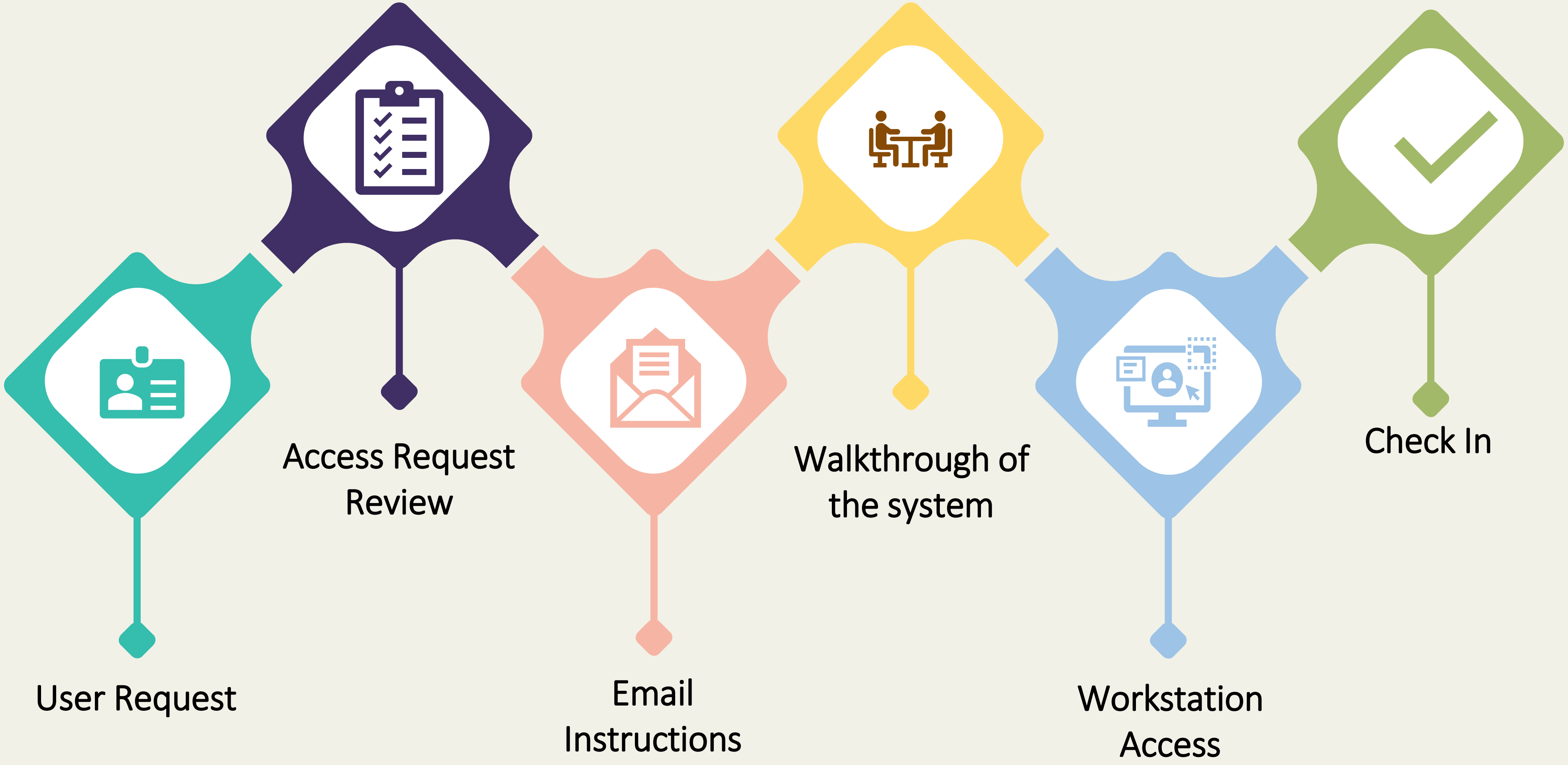
Our Commitment

5

Questions

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

Detailed onboarding instructions located: https://portal.sdc.dot.gov/faqs

1
SDC Overview

2
Project Lifecycle

3
SDC Walkthrough

4
Our Commitment

5
Questions

SDC Walkthrough

SDC PORTAL

- Portal Sign In
- Data Sets
- Workstations

WORKSTATION

- Python 3.12x
- R 4.4.x and Rstudio
- DBeaver
- Anaconda3

EXPORTING DATA

- Uploading Results File
- Request Export in SDC Portal
- Once Approved, download file from the SDC Portal

IMPORTING DATA

- Select data to import
  - Curated Datasets
  - Raw Datsets
  - Published Datasets
- Verify through S3 browser

1

SDC Overview

2

Project Lifecycle

3

SDC Walkthrough

4

Our Commitment

5

Questions

SDC Portal Walkthrough

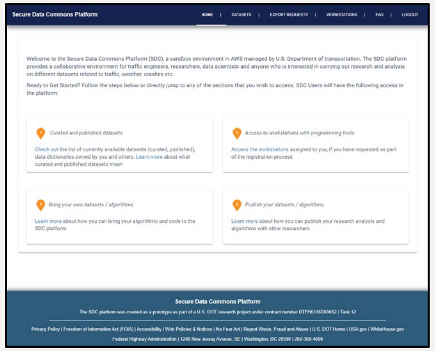
01 SDC Portal Demo

02 SDC Workstation Demo

03 SDC Export Demo

04 SDC Import Demo

- Portal Sign in
  - Login.gov
  - One-time sign in (connects your Login.gov account to your SDC account)
- Datasets
  - My Datasets & Algorithms (upload & download files)
  - SDC Datasets (Request Access)
- Workstations
  - My Workstations


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

Welcome to the Secure Data Commons Platform (SDC), a sandbox environment in AWS managed by U.S. Department of transportation. The SDC platform provides 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 etc.

Ready to Get Started? Follow the steps below or directly jump to any of the sections that you wish to access. SDC Users will have the following access in the platform:

1 *Curated and published datasets*

Check out the list of currently available datasets (curated, published), data dictionaries owned by you and others. Learn more about what curated and published datasets mean

2 *Access to workstations with programming tools*

Access the workstations assigned to you, if you have requested as part of the registration process

3 *Bring your own datasets / algorithms*

Learn more about how you can bring your algorithms and code to the SDC platform

4 *Publish your datasets / algorithms*

Learn more about how you can publish your research analysis and algorithms with other researchers

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

1

SDC Overview

2

Project Lifecycle

3

SDC Walkthrough

4

Our Commitment

5

Questions

SDC Workstation Demo

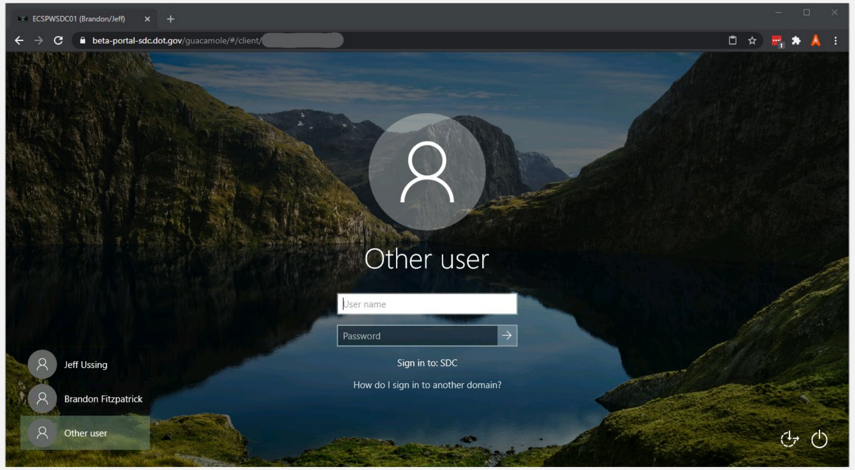
01 SDC Portal Demo

02 SDC Workstation Demo

03 SDC Export Demo

04 SDC Import Demo

- Tool Stack
  - Workstation Scaling
  - Python 3.12.x
  - Notepad++
  - Git and Git Extensions
  - R 4.4.x and RStudio
  - Cyberduck
  - DBeaver
  - Anaconda3
  - On Request: VS Code, Power BI Desktop



ECSPWSDC01 (Brandon/Jeff)

beta-portal-sdc.dot.gov/guacamole/#/client/[redacted]

Other user

User name

Password

Sign in to: SDC

How do I sign in to another domain?

1

SDC Overview

2

Project Lifecycle

3

SDC Walkthrough

4

Our Commitment

5

Questions

SDC Export Demo

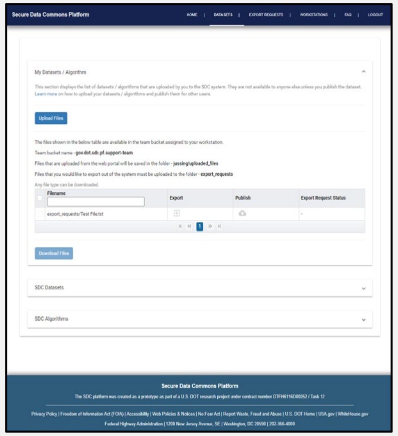
01 SDC Portal Demo

02 SDC Workstation Demo

03 SDC Export Demo

04 SDC Import Demo

- Upload your results file to:
  - s3://sdc.dot.gov.team.your-team/export_requests/
- Go to the SDC Portal
  - Datasets → My Datasets / Algorithm → Request Export
- After your request is approved return to the SDC Portal
  - Datasets → My Datasets / Algorithm to download your file


Secure Data Commons Platform
HOME | DATASETS | EXPORT REQUESTS | WORKSTATIONS | FAQ | LOGOUT
Upload Files
Download Files
SDC Datasets
SDC Algorithms

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.

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

Team bucket name - **gov.dot.sdc.pf.support-team**

Files that are uploaded from the web portal will be saved in the folder - **jussing/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 Request Status
export_requests/Test File.txt			-

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

SDC Import Demo

1

SDC Overview

2

Project Lifecycle

3

SDC Walkthrough

4

Our Commitment

5

Questions

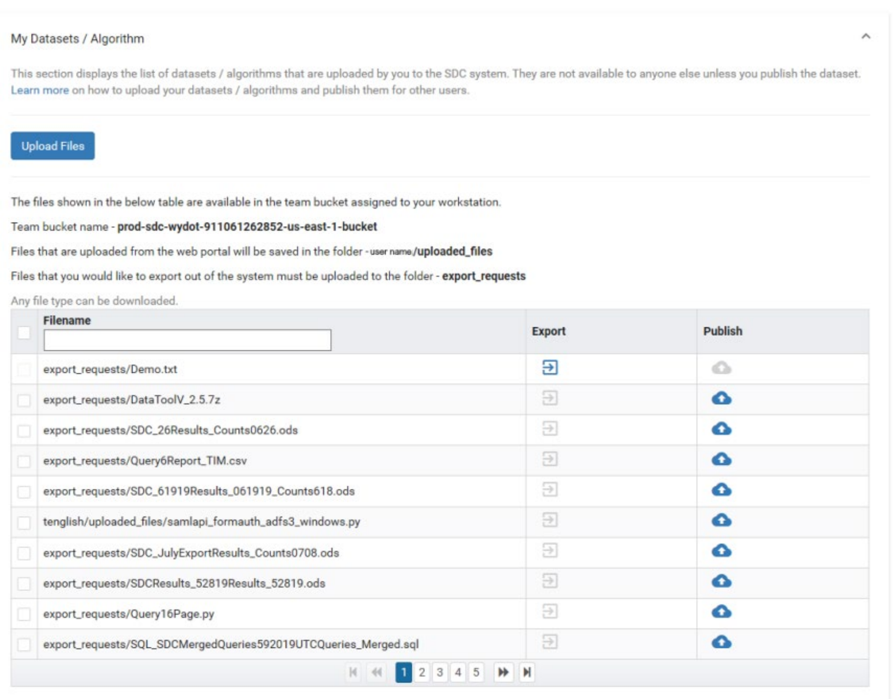
01 SDC Portal Demo

02 SDC Workstation Demo

03 SDC Export Demo

04 SDC Import Demo

- 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		

1

SDC Overview

2

Project Lifecycle

3

SDC Walkthrough

4

Our Commitment

5

Questions

How We Plan to Serve You

Customer Focus

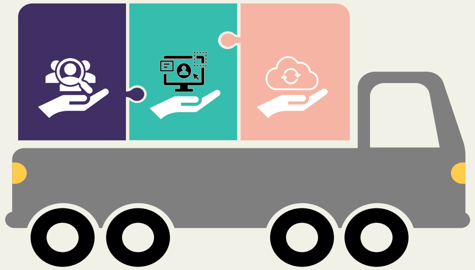
- *SDC product roadmap* that focuses on project and user needs
- Quarterly *Executive Briefing* meetings to communicate different project successes using SDC and upcoming project pipeline
- Quarterly *Customer Advisory Board* meetings to solicit feedback, discuss new features, and collaborate with the community
- Customized design/control of curation of raw data and exporting of derived data sets
- *Chargeback model* that is transparent and easy to understand – allowing for straightforward budgeting
- *Responsive* and **knowlegable** customer support to quickly work through any user roadblocks or problems

Enhanced Enablement Services

- Customer centered *data preparation processes* for each onboarded research team
- *Data modeling and harmonization* to assist users in cross-SDC project research
- Enhanced communication on *data availability* to allow users to perform timely SDC research
- Project owner dashboards to highlight usage and billing in real time

SDC Engineering Improvements

- *Automated builds and deployments* to release new features quickly and reliably
- Use of cutting-edge *cloud-based* technologies to provide an optimal user experience while paying for only what you consume
- Dedicated Information Security Services Officer (ISSO) to ensure that platform remains compliant and data stays protected
- Enhanced monitoring to rapidly detect data anomolies



Contact Us

Shyla Morisetty- shyla.morisetty@dot.gov

SDC Support- sdc-support@dot.gov

SDC Website- https://www.transportation.gov/data/secure

Thank you