Secure Data Commons

Data Provider Guide: Version 3

www.its.dot.gov/index.htm
Draft Report — September 24, 2018
FHWA-JPO-18-xxx

U.S. Department of Transportation

Notice

Revision History

#	Name	Version	Revision Date	Revision Description
1	REAN	1.0	08/02/2018	Initial Draft
2	Volpe	1.1	08/08/2018	Formatting
3	REAN	1.2	09/13/2018	Add Export Functionality Documentation

Technical Report Documentation Page

1. Report No. FHWA-JPO-18-xxx	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Secure Data Commons Proof of Concept: Data Provider Guide		**5. Report Date** September 24, 2018 **6. Performing Organization Code**
7. Author(s) ICF, REAN Cloud		**8. Performing Organization Report No.** Task 2 Report
9. Performing Organization Name And Address ICF International, 1725 Eye St NW, Washington DC, 20006 REAN Cloud, 2201 Cooperative Way #302 Herndon, VA 20171		**10. Work Unit No. (TRAIS)** **11. Contract or Grant No.** DTFH61-16-D-00052/Task 12
12. Sponsoring Agency Name and Address U.S Department of Transportation 1200 New Jersey Ave, SE Washington, DC 20590		**13. Type of Report and Period Covered** Design Document, 09-14-2017 to 1/05/2018 **14. Sponsoring Agency Code**
15. Supplementary Notes Ariel Gold (TOCOR), Harry Crump (COR), Bob Brown (CO)		

16. Abstract

The SDC POC is an online cloud-based analytic sandbox that provides users access to data sets and programming environments to a number of transportation related data sets. The project performs the following development activities necessary to support the design, launch, and operation of the Secure Data Commons. The primary SDC concepts are:

- Enable scalable data storage, data analysis, and user access protocol via cloud based platforms (AWS S3, Azure Cloud Storage, etc.);
- Leverage cloud capabilities to share complex (high volume, velocity, and/or variety) transportation datasets with the transportation research community;
- Authorize user access through a data use agreement with revocable access terms to protect the sensitivity of the data;
- Provide users with pre-defined data analysis tools and encourage custom toolsets and open sharing amongst the user community;
- Ensure sensitive data is protected through implementation of DOT IT security standards; and
- Utilize agile development processes to develop increasing product functionality and leverage user feedback.

This document presents guidance for data providers who are contributing data to the SDC platform

17. Key Words Secure Data, Cloud Services, AWS	**18. Distribution Statement** This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161		
19. Security Classif. (of this report) None	**20. Security Classif. (of this page)** None	**21. No. of Pages** 44	**22. Price** NA

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Table of Contents

Chapter 1. Introduction and Document Overview 1
Chapter 2. Initial Setup and Validation 2
Chapter 3. Accessing Workstations 3
Chapter 4. Connected Vehicle Data Ingestion 4
Ingesting Data to S3 4
Ingesting Data to Kinesis Firehose 5
Chapter 5. Exporting Data Out of SDC 6
Approving Export Requests 6
Actions of Review 7
Chapter 6. Upcoming Chapters 8

List of Figures

Figure 1. EXPORT REQUESTS tab of the SDC web portal (Source: REAN CLOUD)..... 6

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 data sets, 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 allow 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 existing SDC provides users with the following data, tools, and features:

- Data: The SDC is ingesting several datasets currently. Additional data sets 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 others.

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 of the datasets hosted in the SDC system. Note that analysts can bring their own data and tools into the SDC system.

This document provides guidance for the data provider role. A similar guide has been prepared for the data analysts which can be accessed by clicking here.

Chapter 2. Initial Setup and Validation

This chapter provides guidance on the initial setup and validation of the user into the SDC system. Please refer to Chapter 2 in the Data Analysts User Guide, which can be accessed by clicking here.

Chapter 3. Accessing 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. Please refer to chapter 3 in data analysts user guide, which can be accessed by clicking here.

Chapter 4. Connected Vehicle Data Ingestion

This chapter provides guidance on how to perform data ingestion to the SDC platform by different providers. Two approaches are currently supported by the SDC platform to ingest Connected Vehicles (CV) data:

- Ingesting Data to S3
- Ingesting Data to Kinesis Firehose

Ingesting Data to S3

Data providers can ingest data to S3 by following the below steps:

- Data providers will receive an instructions email from support@securedatacommons.com with set of instructions, shell script as attachment and below set of values:

```
API_KEY       = REAN CLOUDREAN CLOUD
AUTH_CODE     = REAN CLOUDREAN CLOUD
ROLE_NAME     = REAN CLOUDREAN CLOUD
API_END_POINT = REAN CLOUDREAN CLOUD
```

- Download the shell script attached to instruction email.
- Update the above provided values in the shell script.
- Please ensure you have the following folder and files under your home directory:

```
\--aws
  |--config
  \--credentials
```

- Please make sure **jq** python library is installed on your machine.
- Run the downloaded shell script which will generate a temporary access/secret keys and update ~/.aws/credentials files by creating a new profile named **sdc-token.** If you already have the profile named **sdc-token**, it will overwrite by updating the credentials.
- Use the **sdc-token** profile in your aws commands or python code through which you are uploading data to S3.
- Make sure to run the above shell script before you start uploading data to S3.

Ingesting Data to Kinesis Firehose

Data providers can ingest data to S3 by following the below steps:

- Data providers will receive an instructions email from support@securedatacommons.com with set of instructions, shell script as attachment and below set of values:

```
API_KEY       = REAN CLOUDREAN CLOUD
AUTH_CODE     = REAN CLOUDREAN CLOUD
ROLE_NAME     = REAN CLOUDREAN CLOUD
API_END_POINT = REAN CLOUDREAN CLOUD
```

- Download the shell script attached to instruction email.
- Update the above provided values in the shell script.
- Please ensure you have the following folder and files under home directory:

```
\--aws
  |--config
  \--credentials
```

- Please make sure **jq** python library is installed on your machine.
- Run the downloaded shell script which will generate a temporary access/secret keys and update ~/.aws/credentials files by creating a new profile named **sdc-token** and export access key/secret keys as **AWS_ACCESS_KEY_ID / AWS_SECRET_ACCESS_KEY** environment variables. if you already have the profile named **sdc-token**, it will overwrite by updating the credentials.

Make sure to run the above shell script before you ingest data to kinesis firehose.

Chapter 5. Exporting Data Out of 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. Upon successful export request submission, the request will be sent to appropriate Data Providers. Data providers are responsible for accepting or rejecting the export requests for the data analysts so that they are able to get the data products out of the SDC system.

Approving Export Requests

Data providers can see the requests in the EXPORT REQUESTS tab of the SDC web portal, as seen in Figure 1.

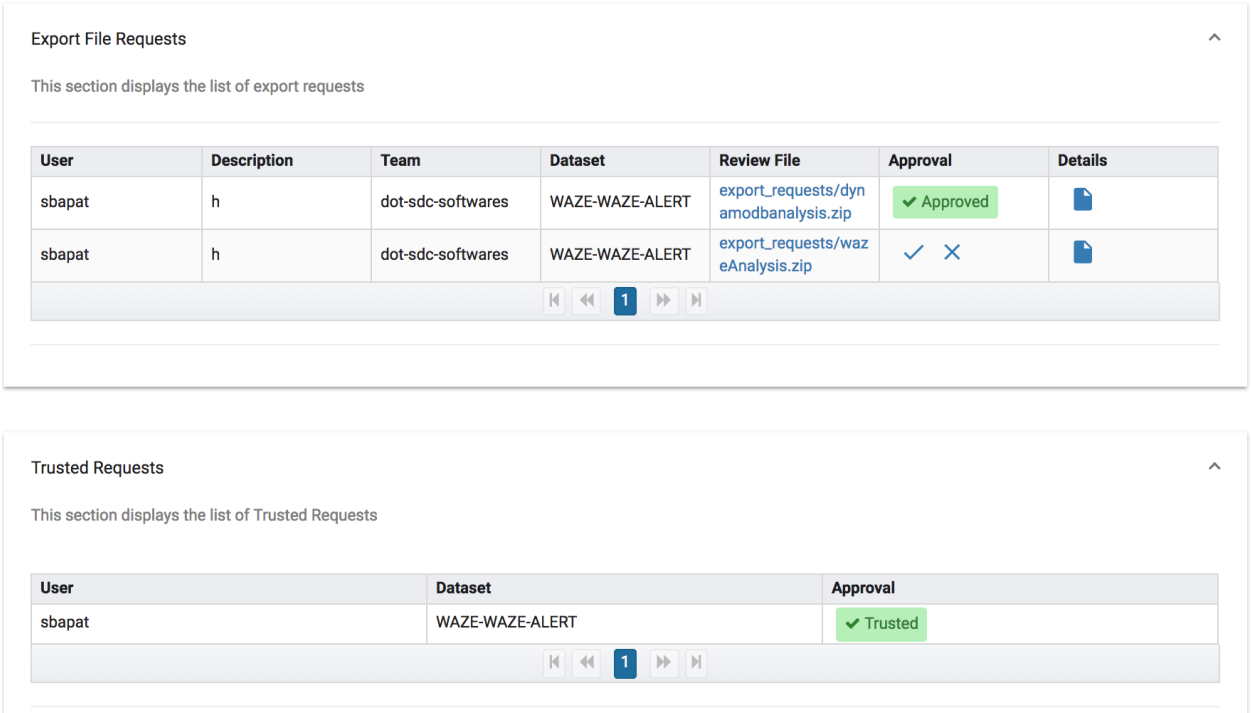

Figure 1. EXPORT REQUESTS tab of the SDC web portal (Source: REAN CLOUD).

There are two sections for the requests made by the data analysts:

1. Export File Requests: These requests correspond to the approval form submitted by the data analyst for exporting any data file out of the SDC system.
   The data provider can accept the request by clicking on the right mark and reject the request by clicking on the wrong mark as shown above.

To get all the details regarding the request, the provider can click on the file symbol under the details column for each request.
To review the file for which the request has been made, the data provider can click on the file link to download the file and review it before giving the access to the analyst.

2. Trusted Requests: These requests correspond to the request for getting the trusted status. The data provider can accept the request by clicking on the right mark and reject the request by clicking on the wrong mark.

Actions of Review

1. Notify: The request of the data analyst for the trusted status or the export request is accepted automatically. The data provider is notified with an email for the request that is accepted.

2. NotifyReview: The request of the data analyst is sent to the data provider for approval by sending him a notification over an email. The data provider has to accept or reject the request from the SDC web portal under the section named “EXPORT REQUESTS”

Chapter 6. Upcoming Chapters

The following chapters will be added to the guide as the SDC system evolves

- Metadata and Data discovery approahces
- Data quality checks and processes

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