



**August 13, 2021**

Project ID: TM30336

# **Spark: R Front End Implementation Within Private Debt Collection**

Ifraz Ahmed

Data Scientist

SB/SE Research, Team 3

# Contents

<b>Project Background</b>	<b>3</b>
Project Objectives	3
<b>Necessary Steps for Setting Up SparkR</b>	<b>3</b>
Connecting to Hadoop	3
Connecting SecureCRT to Hadoop	4
Connecting SecureCRT to Mercury	6
Copying Files to HDFS	7
Creating SparkR Environment	8
Ingesting Files	8
<b>Conclusion</b>	<b>9</b>
<b>Next Steps</b>	<b>9</b>

## Project Background

SB/SE Research has been working on Private Debt Collection (PDC) Program for revenue projections, modeling, and Unique Right Party Contact (RPC) reports. We have been collecting the data since April 2017. Recently the data has grown significantly larger and harder to manipulate than the R platform can handle. Every week, a snapshot of the current PDC inventory is taken at both an entity and module level. Each month these weekly snapshots are appended to an archive dataset that contains weekly snapshots back to the beginning of the program. These monthly files have grown to over 500 million rows and continue to grow.

As the dataset size continues to grow into the territory of big data, it is necessary to adapt to a new storage process. Therefore, Hadoop is being implemented to replace batch processing with parallel processing. Hadoop is comprised of two components, Hadoop Distributed File System (HDFS) and MapReduce. HDFS is the storage mechanism for data stored within the Hadoop. MapReduce is the processing framework that provides the parallelization needed when working with big data. Data scientists can take advantage of Hadoop's infrastructure using languages they are already familiar with, such as SQL (Hive), Python (PySpark), and R (SparkR).

**Purpose Statement:** The focus of this fellowship program is to implement SparkR for faster and more reliable ETL process.

## Project Objectives

1. Create the SparkR environment with a connection to HDFS.
2. Modify the existing base R code to be compatible with the new SparkR environment.
3. Copy data stored within Mercury Server to HDFS.
4. Rebuild the original R code to be compatible with SparkR.

## Necessary Steps for Setting Up SparkR

### Connecting to Hadoop

Get access to Hadoop by performing the following steps:

1. Go to the Compliance Data Warehouse - Data Analysis Tools page.  
(<https://cdw.web.irs.gov/Tools/dataAnalysisTools.aspx>)
2. Request access by selecting "View Details", "Request this Tool", and checking "I have a CDW Account". Fill in the "Software Installation Request" form.

3. Select “1 – 100 GB” for Total Data Set Service and “R Programming” for Project Description in the Software Installation Request Form.

Software Installation Request

I have a CDW account ☒

Please fill out and submit this form to complete your software request.

SEID:

Name:

Email:

Group/Project/Unit:

I am a contractor ☐

Manager/Lead:

Total Data Set Size:

Project Description:

I confirm my request for access to Hadoop ☒

4. After the request has been submitted, CDW staff will contact you about your account. (The initial setup process can take approximately two weeks after submitting the form)
5. Once the initial setup process is completed and you have a password, log in to R Studio (<http://10.207.84.70:8787/>) and enter credentials.

## Connecting SecureCRT to Hadoop

SecureCRT is software used to interact with Hadoop and uses UNIX as the shell scripting language. SecureCRT interacts with two separate locations within Hadoop, the EdgeNode and HDFS. When interacting with the EdgeNode, UNIX commands are entered normally. When interacting with HDFS, commands must have either “*hdfs dfs*” or “*hadoop fs*” prefix added to the command. See the figure below.

```
THIS U.S. GOVERNMENT SYSTEM IS FOR AUTHORIZED USE ONLY!

Use of this system constitutes consent to monitoring, interception,
recording, reading, copying or capturing by authorized personnel of
all activities. There is no right to privacy in this system.
Unauthorized use of this system is prohibited and subject to criminal
and civil penalties, including all penalties applicable to willful
unauthorized access (UNAX) or inspection of taxpayer records (under
18 U.S.C. 1030 and 26 U.S.C. 7213A and 26 U.S.C. 7431).

ATTENTION ALL USERS

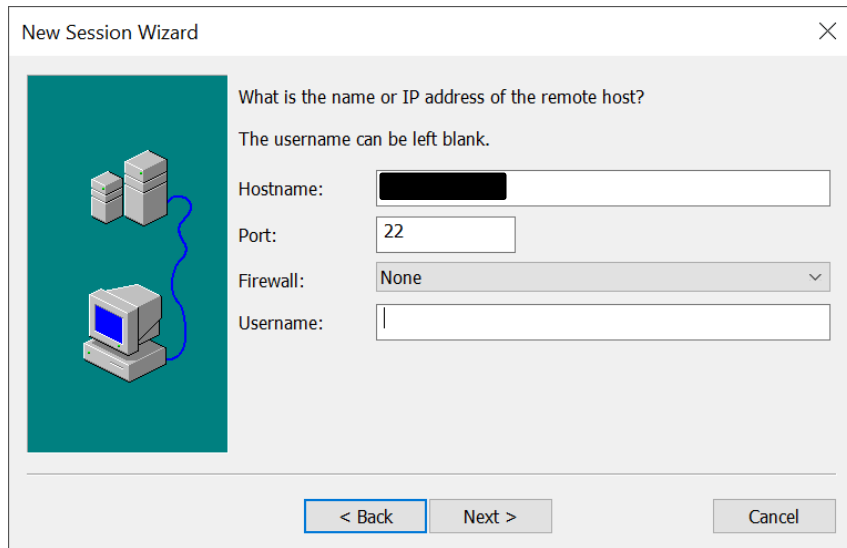
Please delete files from your $HOME to avoid filling up disk space
on the /home partition! Thank you.

Last login: Fri Aug 13 10:08:14 2021 from dci005ma4421998.ds.irsnet.gov
Sourcing SYBASE.sh
[d9ztb@mtb0120ppedge01 ~]$ ls
1.R  EHIST_202051.rds  R  SparkR  HDFS  Dataset  Ingestion.R  test_1.R  testing_set
2  EHIST_202124.csv  spark  SparkR.R  test2.R  working_for_me.R
3.R  final_testing  Sparklyr.txt  SparkR_Test  testing_4
[d9ztb@mtb0120ppedge01 ~]$ hdfs dfs -ls
Found 3 items
drwx----- - d9ztb hdfs 0 2021-07-31 20:00 .Trash
drwx----- - d9ztb hdfs 0 2021-08-13 11:07 .sparkStaging
-rw----- 3 d9ztb hdfs 88734 2021-07-29 12:07 new.inbound.calls.cbe.only.202007.rds
[d9ztb@mtb0120ppedge01 ~]$
```

Note: First command shows files located in the EdgeNode directory and the second command shows files located in HDFS.

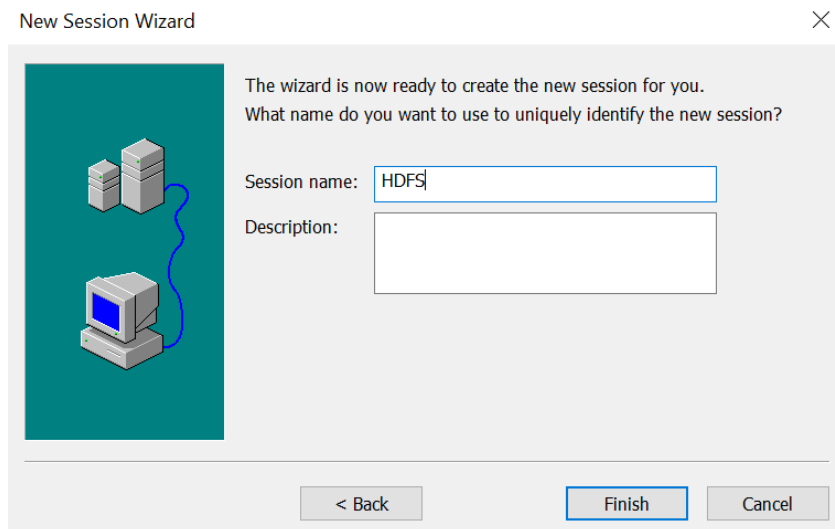
The following steps will connect Hadoop to SecureCRT:

1. Open SecureCRT.
2. Navigate to the Session Manager and click New Session (plus sign).
3. The New Session Wizard will open. Ensure SecureCRT Protocol is SSH2 and click next.
4. In the Host Name field enter [10.207.84.70](#), leave all other fields blank and click next.



The image shows the 'New Session Wizard' dialog box in SecureCRT. On the left is a graphic of a server rack connected to a desktop computer. The text on the right asks 'What is the name or IP address of the remote host?' and 'The username can be left blank.' Below this are input fields for 'Hostname:', 'Port:', 'Firewall:', and 'Username:'. The 'Port' field contains '22' and the 'Firewall' dropdown is set to 'None'. At the bottom are buttons for '< Back', 'Next >', and 'Cancel'.

5. SecureFX Protocol is SFTP, click next.
6. Enter descriptive name for Session Name, preferably "EdgeNode" or "HDFS" and click finish.



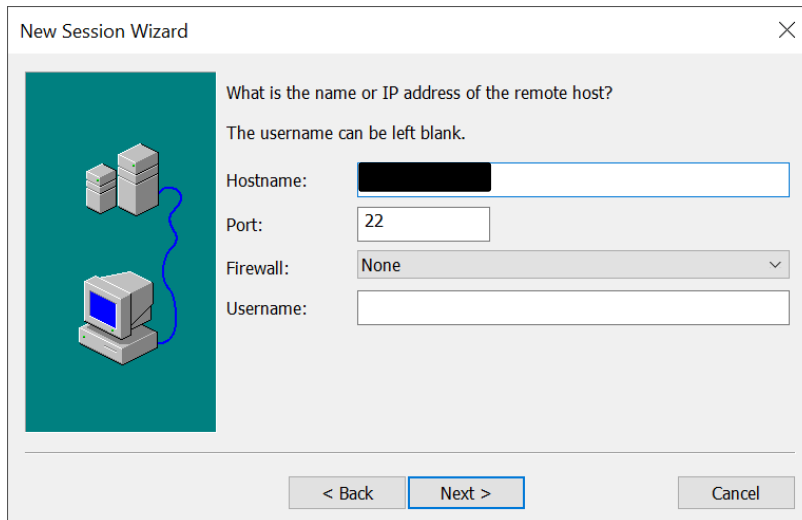
The image shows the second step of the 'New Session Wizard' dialog box. The text on the right says 'The wizard is now ready to create the new session for you. What name do you want to use to uniquely identify the new session?'. Below this are input fields for 'Session name:' and 'Description:'. The 'Session name' field contains 'HDFS'. At the bottom are buttons for '< Back', 'Finish', and 'Cancel'.

7. Double click the Session, enter SEID and password.
8. Once on the terminal enter "ls" (lowercase L and S) and "hdfs" in separate lines to ensure the session is connected properly.

## Connecting SecureCRT to Mercury

The Mercury Server is where the original data files are located (ensure access has been granted to the Mercury Server from CDW before proceeding).

1. Open SecureCRT.
2. Navigate to the Session Manager and click New Session (plus sign).
3. The New Session Wizard will open. Ensure SecureCRT Protocol is SSH2 and click next.
4. In the Host Name field enter **10.207.92.30**, leave all other fields blank and click next.



The image shows the 'New Session Wizard' dialog box in SecureCRT. On the left is a graphic of a server rack connected to a computer. The text on the right asks for the remote host name or IP address and notes that the username can be left blank. The 'Hostname' field is filled with '10.207.92.30'. The 'Port' field is '22'. The 'Firewall' dropdown is set to 'None'. The 'Username' field is empty. At the bottom are buttons for '< Back', 'Next >', and 'Cancel'.

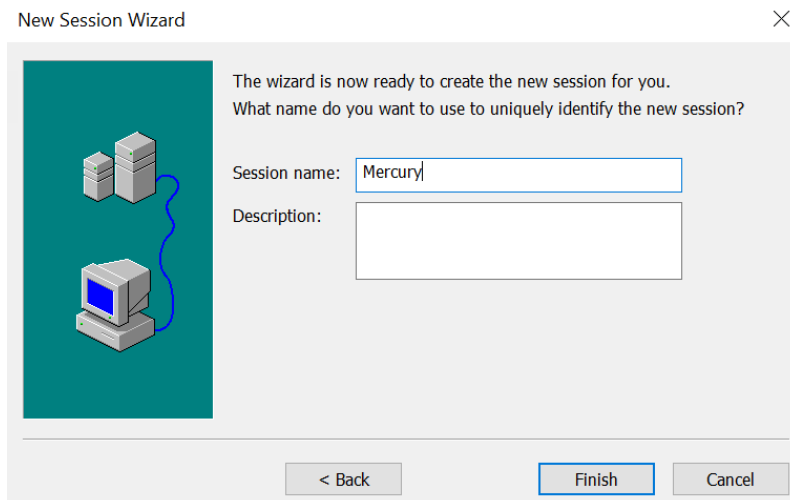
New Session Wizard

What is the name or IP address of the remote host?  
The username can be left blank.

Hostname: 10.207.92.30  
Port: 22  
Firewall: None  
Username:

< Back Next > Cancel

5. SecureFX Protocol is SFTP, click next.
6. Enter descriptive name for Session Name (such as "Mercury").



The image shows the second step of the 'New Session Wizard' dialog box. The text on the right states that the wizard is ready to create the session and asks for a name to uniquely identify it. The 'Session name' field is filled with 'Mercury'. The 'Description' field is empty. At the bottom are buttons for '< Back', 'Finish', and 'Cancel'.

New Session Wizard

The wizard is now ready to create the new session for you.  
What name do you want to use to uniquely identify the new session?

Session name: Mercury  
Description:

< Back Finish Cancel

7. Double click the Session, enter SEID and password.
8. Once on the terminal enter **/s** (lowercase L and S) to ensure the session is connected.

## Copying Files to HDFS

When working with Hadoop, the data must be stored in HDFS. Data are always loaded on to the EdgeNode before getting copied on to HDFS. After the data are copied to HDFS, it is recommended to delete the file(s) from the EdgeNode to save storage for other Hadoop users. Data cannot be stored in the EdgeNode longer than 30 days or it will be deleted by CDW staff to conserve storage.

SecureFX is used to view files using a Graphical User Interface (GUI) while SecureCRT uses a command line interface. You can interact only with the EdgeNode using SecureFX but not HDFS (must use SecureCRT when interacting with HDFS). The following steps will outline how to copy files from Mercury Server to Hadoop:

1. Identify a file in Mercury Server and copy the file path.
2. Open SecureCRT and log in to the Mercury Server.
3. Paste the file path following the syntax - "*REDACTED*" (*REDACTED* is the IP address where you establish a connection to Hadoop. Also, do not forget to include a colon and period at the end).
4. Paste the command in the SecureCRT terminal (connected to Mercury) and press enter.
5. Enter SEID, password, and press enter. The password will not display in the console.

```
THIS U.S. GOVERNMENT SYSTEM IS FOR AUTHORIZED USE ONLY!
Use of this system constitutes consent to monitoring, interception,
recording, reading, copying or capturing by authorized personnel of
all activities. There is no right to privacy in this system.
Unauthorized use of this system is prohibited and subject to criminal
and civil penalties, including all penalties applicable to willful
unauthorized access (UNAX) or inspection of taxpayer records (under
18 U.S.C. 1030 and 26 U.S.C. 7213A and 26 U.S.C. 7431).
Last login: Wed Aug 11 13:05:52 2021 from dc1005ma4421998.ds.irsnet.gov
-----
User Statistics
Running Process: 3
Home Directory: /home/d9ztb
Home Directory Size: (0.8 %) 1.8G Out of 200G
-----
[d9ztb@mtb0120ppcdwmerb ~]$ scp /projects/PDC/program_data/pdc/raas_datasets/202124/EHIST_202124.csv d9ztb@10.207.84.70:.

THIS U.S. GOVERNMENT SYSTEM IS FOR AUTHORIZED USE ONLY!
Use of this system constitutes consent to monitoring, interception,
recording, reading, copying or capturing by authorized personnel of
all activities. There is no right to privacy in this system.
Unauthorized use of this system is prohibited and subject to criminal
and civil penalties, including all penalties applicable to willful
unauthorized access (UNAX) or inspection of taxpayer records (under
18 U.S.C. 1030 and 26 U.S.C. 7213A and 26 U.S.C. 7431).

ATTENTION ALL USERS

Please delete files from your $HOME to avoid filling up disk space
on the /home partition! Thank you.

Password:
EHIST_202124.csv                                     1% 1854MB  86.6MB/s  25:58 ETA
```

6. The transfer speed, percentage completed, and time elapsed will display.

- Open a session connected to Hadoop and enter "`hdfs dfs -ls`". A list of files stored on HDFS under "/home/SEID" folder will display. Ensure the file is included in the list.
- Delete the EdgeNode copy using the following syntax "`rm -r filename.format`".

## Creating SparkR Environment

Setting up the SparkR environment involves code that will download packages, specify the folder for the Spark home environment, and create the SparkContext environment. Script is in SB/SE Research Detroit Server (\OPEN PROJECTS\TM30336 - PDC Program Improvement\Spark and Hadoop)

```

1 library(dplyr)
2 library(sparklyr)
3 Sys.setenv(SPARK_HOME="/opt/software/spark/spark-2.4.5-bin-hadoop2.7/")
4
5 .libPaths(c(file.path(Sys.getenv("SPARK_HOME"), "R", "lib"), .libPaths()))
6 Sys.setenv(SPARK_HOME="/usr/hdp/current/spark2-client")
7 .libPaths(c(file.path(Sys.getenv("SPARK_HOME"), "R", "lib"), .libPaths()))
8 config <- spark_config()
9 options (sparklyr.verbose = TRUE)
10 library(SparkR)
11 sc <- sparkR.session(spark.master="yarn", spark.submit.deployMode="client")
12

```

```

> sc <- sparkR.session(spark.master="yarn", spark.submit.deployMode="client")
Spark package found in SPARK_HOME: /usr/hdp/current/spark2-client
Launching java with spark-submit command /usr/hdp/current/spark2-client/bin/spark-submit --master "yarn" sparkr-shell /tmp/Rtm
puTQfaF/backend_port52ab1f04555f
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
21/08/13 14:20:07 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
21/08/13 14:20:07 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
21/08/13 14:20:07 WARN Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
21/08/13 14:20:08 WARN ResourcePBImpl: Got unknown resource type: yarn.io/gpu; skipping
Warning message:
In sparkR.session(spark.master = "yarn", spark.submit.deployMode = "client") :
  Version mismatch between Spark JVM and SparkR package. JVM version was 2.3.2.3.1.4.0-315 , while R package version was 2.3.1
>

```

Note: It is critical that SparkR version 2.4.5 is loaded.

## Ingesting Files

Use the following script to ingest files into SparkR:

```

variable_name <- file.path("/file/path/in/HDFS/filename.csv")
dataframe_name <- read.df(variable_name.Path, source="csv", header="true")

```

Currently there is no function to read a rds file directly. The workaround is to convert rds file to csv file and use the same script above.



## Conclusion

After testing SparkR with PDC data files it is determined that it may not be the best tool to use. SparkR is limited in the file formats it can read, most notably rds files. A significant amount of data within PDC is stored in rds format. Although a workaround solution of converting a rds file to a csv file exists, the csv file gets larger. We may save time using SparkR but lose the time savings due to the larger csv file size.

Because resources are limited and shared among other users, performance may depend on how busy the server is. Data Administration Office staff may contact you if you have multiple SparkR sessions open and/or a task is taking too long to run. If many people are using the server, performance will degrade.

## Next Steps

- Explore the possibility of using Python.
  - ❖ Python can read and write rds files.
  - ❖ Perform data manipulation in python, save to R, and continue using R script.
- Explore other languages to reduce data processing time.
- Since data are appended weekly, information gets repeated, resulting in considerably large amounts of duplicate information (such as taxpayers that have already paid and left the PDC program).
  - ❖ Explore possibility of reducing the rows in the database while keeping the same quality report.
- Reduce data size while still providing a quality report.