This tutorial shows you how to run RStudio Server (https://www.rstudio.com/products/rstudio/#Server) on a Dataproc cluster and access the RStudio web user interface (UI) from your local machine.

This tutorial assumes that you are familiar with the R language and the RStudio web UI, and that you have some basic understanding of using Secure Shell (SSH) tunnels, Apache Spark, and Apache Hadoop running on Dataproc.

Note: RStudio® is a trademark of RStudio, Inc (https://www.rstudio.com/about/trademark/). and is not affiliated with Google. For more information, see the RStudio site (https://www.rstudio.com).

This tutorial walks you through the following procedures:

- Connect R through Apache Spark to Apache Hadoop YARN running on a Dataproc cluster.
- Connect your browser through an SSH tunnel to access the RStudio, Spark, and YARN UIs.
- Run an example query on Dataproc using RStudio.

This tutorial uses the following billable components of Google Cloud:

- Dataproc (/dataproc/pricing)
- Cloud Storage (/storage/pricing)

To generate a cost estimate based on your projected usage, use the pricing calculator (/products/calculator). New Google Cloud users might be eligible for a free trial (/free-trial).

1. Sign in (https://accounts.google.com/Login) to your Google Account.

   If you don't already have one, sign up for a new account (https://accounts.google.com/SignUp).

2. In the Cloud Console, on the project selector page, select or create a Cloud project.

★ **Note**: If you don't plan to keep the resources that you create in this procedure, create a project instead of selecting an existing project. After you finish these steps, you can delete the project, removing all resources associated with the project.

Go to the project selector page (https://console.cloud.google.com/projectselector2/home/dashboard)

3. Make sure that billing is enabled for your Google Cloud project. Learn how to confirm billing is enabled for your project (/billing/docs/how-to/modify-project).

4. Enable the Dataproc and Cloud Storage APIs.

   Enable the APIs (https://console.cloud.google.com/flows/enableapi?apiid=dataproc,storage_component)

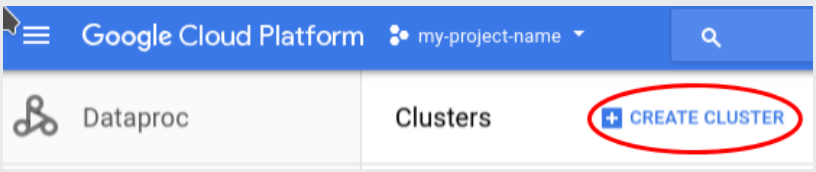
5. Install and initialize the Cloud SDK (/sdk/docs/).

When you finish this tutorial, you can avoid continued billing by deleting the resources you created. See Cleaning up (#clean-up) for more information.

1. In the Cloud Console, go to the **Dataproc Clusters** page:

   GO TO THE CLUSTERS PAGE (https://console.cloud.google.com/dataproc/clusters)

2. Click **Create Cluster**:



3. Name your cluster, and click **Create**.

For this tutorial, the default cluster sizes are adequate. Note the zone that you created the cluster in, because you will need that information in later steps.

To create a user account to log in to the RStudio UI, follow these steps.

1. Create a new user account, replacing `[USER_NAME]` with the new username:

2. When you are prompted, enter a password for the new user.

You can create multiple user accounts on the master node to give users their own RStudio environment. For each user that you create, follow the sparklyr and Spark installation steps (#sparklyr).

RStudio Server runs on the Dataproc master node and is accessible only from the GCP internal network. To access the server, you need a network path between your local machine and the master node on the GCP internal network.

You can connect by port forwarding through an SSH tunnel, which is more secure than opening a firewall port to the master node. Using an SSH tunnel encrypts your connection to the web UI, even though the server uses simple HTTP.

There are two options for port forwarding: Dynamic port forwarding using SOCKS (https://man.openbsd.org/ssh#D) or TCP port forwarding (https://man.openbsd.org/ssh#L).

Using SOCKS, you can view all internal web interfaces that are running on the Dataproc master node; however, you need to use a custom browser configuration to redirect all browser traffic over the SOCKS proxy.

TCP port forwarding does not require a custom browser configuration, but you can only view the RStudio web interface.

Note: For either port forwarding method, the SSH session must remain active to access the web interface. If the session closes, the RStudio web UI will become non-responsive until you reconnect the SSH session.

To create an SSH SOCKS tunnel and connect by using a specially configured browser profile, follow the steps in Connecting to the web interfaces (/dataproc/docs/concepts/accessing/cluster-web-interfaces#connecting_to_web_interfaces).

After you connect, use the following URLs to access the web interfaces.

- To load the RStudio web UI, connect your specially configured browser to `http://[CLUSTER_NAME]-m:8787`. Then log in by using the username and password that you created.

- To load the YARN resource manager web UI, connect your specially configured browser to `http://[CLUSTER_NAME]-m:8088`.
- To load the HDFS NameNode web UI, connect your specially configured browser to `http://[CLUSTER_NAME]-m:9870`.

To install the sparklyr package and Spark, in the RStudio R console, run the following commands:

These commands download, compile, and install the required R packages and a compatible Spark instance. Each command takes several minutes to complete.

Each time you restart an R session, follow these steps:

1. Load the libraries and set up the necessary environment variables:

2. Connect to Spark on YARN, using the default settings:

The `sc` object references your Spark connection, which you can use to manage data and execute queries in R.

If the command succeeds, then skip to Checking the status of the Spark connection. (#checking_the_status_of_the_spark_connection).

If the command fails with an error message starting with:

Then there is an incompatibility (https://issues.apache.org/jira/browse/SPARK-15343) between the version of YARN and the version of Spark that RStudio uses. You can avoid this incompatibility by disabling the Yarn time service.

3. In the menu of the RStudio web-UI, navigate to **Tools** > **Shell** to create a new Terminal tab.
4. In the Terminal tab, enter the following command to disable the service causing the incompatibility.

5. Close the Terminal tab, and in the menu navigate to **Session** > **Restart R**.

Now repeat steps 1 and 2 again and you will connect to Spark successfully.

The `sc` object created above is the reference to your Spark connection. You can execute the following command to confirm that the R session is connected:

If your connection is established, the command returns the following:

You can tune the connection parameters by using a configuration object that can be passed to `spark_connect()`.

For more details on sparklyr connection parameters and tuning Spark on YARN, see the following:

- sparklyr documentation (http://spark.rstudio.com/)
- Running Spark on YARN (https://spark.apache.org/docs/latest/running-on-yarn.html)
- How-to: Tune Your Apache Spark Jobs, Part 1 (http://blog.cloudera.com/blog/2015/03/how-to-tune-your-apache-spark-jobs-part-1/) and Part 2 (https://blog.cloudera.com/blog/2015/03/how-to-tune-your-apache-spark-jobs-part-2/)

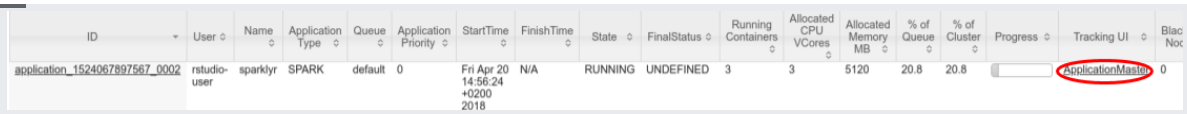
To verify that everything is working, you can load a table onto the Dataproc cluster and perform a query.

1. In the R console, install the example dataset, a list of all New York City flights in 2013, and copy it into Spark:

2. If you are not using SOCKS port forwarding, skip to step 3. Otherwise, use the Spark UI to verify that the table was created.

    a. In the browser that you configured, load the YARN resource manager:

In the application list, a row for the sparklyr app will appear in the table.



ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blac Noc
application_1524067897567_0002	rstudio-user	sparklyr	SPARK	default	0	Fri Apr 20 14:56:24 +0200 2018	N/A	RUNNING	UNDEFINED	3	3	5120	20.8	20.8		ApplicationMaster	0

b. In the **Tracking UI** column, on the right side of the table, click the **ApplicationMaster** link to access the Spark UI.

   In the **Jobs** tab of the Spark UI, you will see entries for the jobs that copied the data to Spark. In the **Storage** tab, you will see an entry for **In-memory table 'flights'**.

RDD Name	Storage Level	Cached Partitions	Fraction Cached	Size in Memory	Size on Disk
In-memory table `flights`	Memory Deserialized 1x Replicated	1	100%	22.5 MB	0.0 B

3. In the R console, run the following query:

   This query creates a list of the average departure delay per flight by airline in descending order, and produces the following result:

If you go back to the **Jobs** tab in the Spark UI, you can see the jobs that are used to execute this query. For longer-running jobs, you can use this tab to monitor progress.

Thanks to Mango Solutions (https://www.mango-solutions.com/) for their assistance in preparing certain technical content for this article.

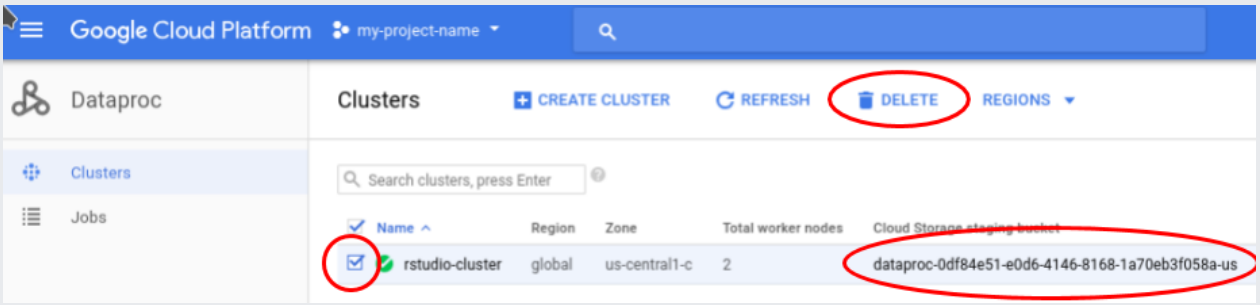
To avoid incurring charges to your Google Cloud Platform account for the resources used in this tutorial:

- Delete the Dataproc cluster.
- If you have no other Dataproc clusters in the same region, you also need to delete the Cloud Storage bucket that was automatically created for the region.

1. In the Cloud Console, go to the Dataproc Clusters page:

   GO TO THE CLOUD DATAPROC CLUSTERS PAGE (https://console.cloud.google.com/dataproc/clusters)

2. In the cluster list, find the row for the Dataproc cluster that you created, and in the **Cloud Storage staging bucket** column, make a note of the bucket name, which begins with the word **dataproc**.



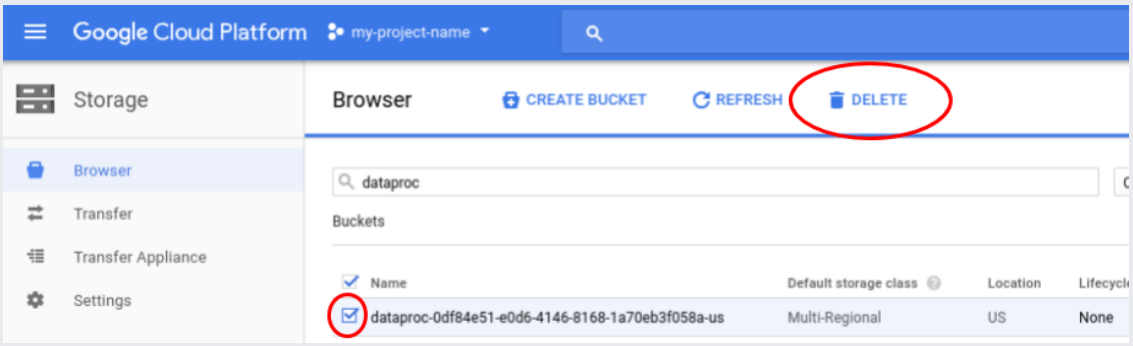
3. Select the checkbox next to **rstudio-cluster,** and click **Delete**.
4. When you are prompted to delete the cluster, confirm the deletion.

1. To delete the Cloud Storage bucket, go to the Cloud Storage Browser:

   GO TO THE CLOUD STORAGE BROWSER (https://console.cloud.google.com/dataproc/clusters)

2. Find the bucket that is associated with the Dataproc cluster that you just deleted.

3. Select the checkbox next to the bucket name, and click **Delete**.



4. When you are prompted to delete the storage bucket, confirm the deletion.

- For other ways of interacting with Dataproc, see Samples and Tutorials (/dataproc/docs/tutorials).
- Try out other Google Cloud features for yourself. Have a look at our tutorials (/docs/tutorials).