StreamSets Data Collector User Training

Exercise Manual

A1/1709/A

Introduction

This is the Hands-On Exercise Manual for the StreamSets Data Collector User Training course. Our exercise environment runs in the cloud; the machine you will connect to includes a single-node Hadoop cluster and the latest version of StreamSets Data Collector, along with some other software such as MySQL.

SDC Login

You will connect to the StreamSets Data Collector (SDC) via a Web browser, using the IP address given to you by your instructor. SDC is running on port 18630, so if your instructor gives you the IP address 10.9.8.7, then you would point your Web browser at `http://10.9.8.7:18630`.

The default login for the SDC user `admin` (which is a user account allowed to create, delete, view, and run pipelines) has the password `admin`. For security reasons we have changed this, so you should log in with the credentials:
Username: `admin`
Password: `training`

Hands-On Exercise: Build and Run a Pipeline

In this Exercise, you will build a pipeline and observe the metrics generated.

1. Create a new pipeline.
2. Use the Dev Data Generator as your origin. Have it generate three fields: a string, a date, and a long.
3. Use Trash as the destination for the data.
4. Preview the pipeline, and explore how to view the records generated by the source, and those received by the destination.
5. Run the pipeline, and explore the metrics that are generated.
6. When you are done, stop the pipeline.

STOP HERE. THIS IS THE END OF THE EXERCISE.

Hands-On Exercise: Reading Data from Database Tables

In this Exercise, you will read data which is being inserted into a MySQL table. MySQL is running on your local machine, and the MySQL JDBC driver has been installed. We have created a table called `t1` in a database called `sdcdb`. The table has two columns: `id`, a primary key integer, and `chr`, a string. We are running a data generator which writes a new row to the table four times per second.

1. Start a new pipeline
2. Use the JDBC Query Consumer origin. The configuration information you need is as follows:
   - JDBC connection string: `jdbc:mysql://localhost:3306/sdcdb`
   - Username and password (entered in the 'Credentials' tab): `student/student`
   - SQL query: `SELECT id, str FROM t1 WHERE id > ${OFFSET} ORDER BY id`
   - Initial offset: `1`
   - Offset column: `id`
   - Set the query interval to be 5 seconds
3. Use Trash as the destination.
4. Preview, and then run the pipeline. Observe the metrics.
5. When you have finished, stop the pipeline.

If you have more time:

1. Observe what happens when you run the pipeline again. Notice the number of records read. Stop the pipeline, reset the origin, and run it again. Notice the number of records generated this time.

STOP HERE. THIS IS THE END OF THE EXERCISE.

Hands-On Exercise: Writing to a Hadoop Cluster

In this exercise, you will write data to a Hadoop cluster, which is running on a remote server in the cloud. You will use the JDBC origin from the last exercise, but write to files in the Hadoop Distributed File System (HDFS).

1. Duplicate your JDBC pipeline by going to the Pipelines screen, clicking on the 'more' icon next to the pipeline (⋮), and selecting 'Duplicate'.
2. Edit the pipeline. Delete the Trash destination, and instead choose the Hadoop FS from Stage Library CDH 5.12.0 as the destination.
3. Edit the Hadoop FS destination configuration as follows:
    - Hadoop FS tab:
        - Hadoop FS URI: `hdfs://hadoopbox:8020`
        - HDFS User: <Your instructor will give you a username>
        - Hadoop FS Configuration Directory: `/etc/hadoop-conf`
    - Output Files tab:
        - Directory Template: `/user/<your username>/myfiles`
        - Max Records in File: 10
    - Data Format tab:
        - Data Format: Delimited
4. Test, then run your pipeline
5. Open a new browser window, and navigate to `http://http://52.91.95.28:8888`. This is the Hue browser on the Hadoop box. Hue is a tool which allows you to view the HDFS filesystem (amongst many other things). Log in with your username and password (the password is the same as your username). Close the welcome message, and from the "hamburger" menu icon on the top left, choose Browsers→Files. Look in the 'myfiles' directory and you should see multiple files containing the data which is being written by your pipeline.
6. When you have finished, stop the pipeline.

STOP HERE. THIS IS THE END OF THE EXERCISE.

Hands-On Exercise: Modifying Data

In this Hands-On Exercise, you will experiment with some of the StreamSets Data Collector processors.

1. Duplicate the JDBC pipeline.
2. Add a new processor to mask the 'str' field between the source and the trash destination. Preview your pipeline to check that this has worked.
3. Remove the masking processor, and add a processor to create a new field, called ‘combinedinfo’, which contains a concatenation of the ‘id’ and ‘str’ fields. Preview your pipeline to check that this has worked.
4. Change the order of the three fields so that the ‘combinedorder’ field comes first, then the ‘str’ field, and finally the ‘id’ field.

STOP HERE. THIS IS THE END OF THE EXERCISE.