Introduction

The following instructions are the requirements and preparation needed to successfully follow along with the data pipeline labs created for the Data Pipeline Exploratorium.

Estimated Duration
20 minutes

No Dependencies

Key Concepts
Workshop prerequisites
Workshop Prerequisites

The following instructions are the requirements and preparation needed to successfully follow along with the data pipeline labs created for the Data Pipeline Exploratorium.

System and Network Requirements
Please come prepared with a laptop that meets the following base requirements:

- Able to create a stable internet connection
- Equipped with a supported browser (Chrome v51+, Mozilla Firefox, Edge 79+)
- A screen resolution of a least 1280x720

Course Environment
StreamSets will provide a link to two virtual machines at the beginning of the workshop that you will use for your Data Pipeline Exploratorium labs. These virtual machines provide access to pre-configured applications, services, and sample data needed to complete the labs.

- **Windows virtual machine** - Provides a Chrome web browser with links to the StreamSets DataOps platform. SQL Server scripts/queries can be executed using SQLPad, which is accessed via the browser.
- **Linux virtual machine** - Provides SQL Server service and the ability to run the StreamSets Data Collector engines as containers

Lab Prerequisites
Before you can work on the labs, you must meet the following requirements:

- **Snowflake Account** - Sign up for a free trial at https://signup.snowflake.com/. We do not recommend using your business Snowflake account, since these can have security configurations that prevent access from outside your network.
- **StreamSets Cloud Account** - Sign up for a free account at: https://cloud.login.streamsets.com/signup. Again, please create a new account just for this training to avoid any security configurations that may have been applied to your StreamSets account. Please note: you will need to use this specific link to create your account, or you will be asked to schedule a meeting before you can login.

Please keep these login credentials at hand for the day of the workshop.
Understanding Data Collector Environments, Deployments, and Engines

StreamSets DataOps Platform uses these three concepts to manage execution.

- **Environment** - a high-level concept such as location (Cloud provider, On-premise cluster, etc.) or usage (Development / Test / Production)
- **Deployment** - Used to manage configurations and dependencies for one or more engines. Engines inherit all configurations from the Deployment they belong to.
- **Engine** - Headless engine that executes the data integration.

Environments are defined and managed in the context of their Environments and Deployments. The Engines are deployed in the data centers to process the data integration tasks.

**ENVIRONMENTS**
StreamSets is ready for scale and includes the option to separate your work by environments: dev, test, prod, etc. This gives you the flexibility to manage projects in separate locations.

**Tip:** The platform also includes a default, self-managed environment to start, so small organizations are ready to go from the start.

**DEPLOYMENT & ENGINES**
StreamSets uses a headless engine deployed into your cloud or on-premise (data plane) using Docker or a Linux install script. Pipelines run on engines to process (SDC) or transform (Transformer) the data.

**Tip:** Using Docker, you can launch your engines and be creating pipelines in minutes.
Understanding Data Collector Environments, Deployments, and Engines

We will be creating a Data Collector environment, deployment, and engine together as part of the workshop.

**ENVIRONMENTS**
Defined by metadata

**DEPLOYMENTS**
Defined by metadata

**ENGINES**
Engines run on Linux. They are configured and managed by Deployments
StreamSets Transformer for Snowflake is a hosted service embedded within the StreamSets DataOps Platform that uses the Snowpark Client Libraries to generate SnowSQL queries executed in Snowflake. Unlike Data Collector, there is nothing to deploy to get started. Instead, when you execute that pipeline, StreamSets generates a DAG. StreamSets then uses the DAG and the Snowpark Client Libraries to generate SnowSQL. That SnowSQL is sent over to Snowflake to be executed in the Warehouse of your choice.

**Transformer for Snowflake**

**How Does It Work?**

Customer designs pipeline in UI

Pipeline runs in StreamSets

**Snowpark** client libraries called

SnowSQL is generated
Configuring Transformer for Snowflake

Under My Account, select the Snowflake Settings tab and enter your Snowflake credentials. Please remember to use the Snowflake account you specifically created for this workshop.

There are two options for authentication: Username/Password or Key Pair Authentication. You can find instructions for setting up Key Pair authentication with Snowflake here. You will want to use a key pair that is not encrypted.

Note that your Snowflake account URL will change depending on the area where you’re located. Make sure you copy paste it up until the “.com”

You also have the option to set pipeline defaults, which is highly recommended. This will prevent you from having to enter your Account URL, Warehouse, Database, Schema, and/or Role for every data pipeline you create. When these values are populated, new Transformer for Snowflake pipelines are created with parameters for those settings pre-populated with the values provided here. These settings can also be overwritten directly in any pipeline.

Stuck?
Please feel free to contact us at roadshow@streamsets.com with any questions.