COVID-19 Tracking with IBM DataStage

Description

IBM DataStage is a data integration tool for designing, developing, and running jobs that move and transform data.

DataStage is one of the data integration components of Watson Studio. The DataStage service is fully integrated into Cloud Pak for Data as a Service as part of the data fabric. It provides a graphical framework for developing the jobs that move data from source systems to target systems. The transformed data can be delivered to data warehouses, data marts, and operational data stores, real-time web services and messaging systems, and other enterprise applications. DataStage supports extract, transform, and load (ETL) and extract, load, and transform (ELT) patterns. DataStage uses parallel processing and enterprise connectivity to provide a truly scalable platform.

This project uses a DataStage flow to join two New York Times COVID-19 public datasets together. The data sets are filtered on case counts in the city of Providence, Rhode Island in the United States of America. The datasets pull data from The New York Times, based on reports from state and local health agencies.

After running this project, you'll be familiar with how to perform filter and join operations within DataStage by using live HTTP connections to the public dataset.

Tip: Download the PDF of these instructions from the Data assets section on the Assets page so you can keep these instructions open while you work.

Data assets

This project contains the following data assets:

- Two data connections
- One DataStage flow
- One DataStage job runtime asset that is tied to the DataStage flow

Before you begin

To complete this project, you must have an instance of DataStage provisioned already. To verify if you already have one, and add one if needed, complete the following steps:

1. From the navigation menu, click **Services** > **Service instances**. The Service instances page appears.
2. If no instances of DataStage are visible, add one by clicking **Add service**, then selecting **DataStage**.

Instructions

Follow these instructions to learn how to perform filter and join operations and tour other features of DataStage.

Use data connections

1. On the **Data Assets** tab, select the pre-configured **COVID Tracking: Colleges** connection.
2. Click **Test connection** to test the connection.

You can create connection assets that connect to a variety of data sources and targets. Check out the sources and targets that are supported by DataStage. A connection asset contains the information necessary to create a connection to a data source or target.

You can use connection assets from within the connectors on the DataStage flow canvas. The same connection can also be used in other products, such as IBM Watson® Studio or Watson Knowledge Catalog.

Work with a DataStage flow

Navigate back to the **Assets** tab in your Project. You will now work with the new DataStage flow canvas. Select the existing flow, **COVID Tracking: Colleges in Providence**, to open the canvas.

Explore the palette of connectors and stages that can be used to build DataStage flows. Interact with the canvas by dragging connectors and stages around, detaching links and double-clicking connectors or stages to see their configuration properties.

This sample DataStage flow extracts data from two HTTP sources, filters the results of one of the sources, joins the two datasets together on a common key, and writes the resulting output dataset to the job log by using a Peek stage.

View the connections

1. When you've become familiar with the canvas, open the **Colleges** HTTP connector by double-clicking the connector on the canvas.

   In the **Properties** section, notice the **Select connection** drop-down menu that you can use to select from existing HTTP connections that were defined within the project. Selecting the connection populates all connection property configurations within the connector.
2. Select the **Output** tab within the connector properties, then open the **Columns** section, then click **Edit**. The **Edit Output Columns** window appears.

   You can click individual columns to select them for editing. You can select multiple columns at the same time to bulk reorder or delete columns.
3. Click **Apply and return** to go back to the HTTP connector properties.
4. Click **Save** to close the connector properties window.

Exploring the filter condition

1. Double-click the Filter stage to open its properties.
2. Open the **Properties** section and note the filter condition, which is located under **Predicates**. Filter conditions support standard SQL expressions. This expression returns all records pertinent to the city of Providence.

Exploring the join condition

1. Double-click the Join stage to open its properties.
2. Open the **Properties** section and note the join key. This the key that will be used to perform a left outer-join of the two input streams, the **College** and **FIPS** datasets.
3. Select the **Output** tab in the Join stage properties, then open the **Columns** section to view the column-mapping details.
4. Click **Edit** to view the **Edit columns** window. This view displays the joined dataset and how the inputs into the join stage will make up the target metadata. From here you can select and map columns from the inputs to the target dataset.

Running a job

1. Click **Run** on the canvas toolbar to save, compile, and run the DataStage flow.
2. Click **Logs** to open the log panel. When the job completes, you see a banner that indicates that the job is finished and successful. The log panel has a type-ahead search and filtering capability that refreshes as new entries come into the log.
3. Return to the project dashboard by clicking the project name in the breadcrumb view on the canvas.

Interacting with jobs and viewing logs

A job is a platform runtime asset that is related to and associated with a flow. Multiple jobs can be associated with the same flow. Jobs can be scheduled or run as needed.

Jobs are automatically created for you when you edit or work with a DataStage flow in the canvas. When you click **Run** on the canvas toolbar, a job is created and invoked. Jobs maintain their past invocations and logs, which you can view on the jobs dashboard.

A flow has a one-to-many relationship with the job asset type, meaning any number of jobs can be associated with a single flow.

Creating a job

1. Go to the **Assets** tab of the project, then go to the **COVID Tracking - Colleges in Providence** flow. Click the overflow menu, then click **Create job**.
2. Follow the steps that the job creation wizard provides. The job creation wizard takes you through the available options for a job, including specifying a scheduled run.

Checking results of a job

Within the project dashboard, select the **Jobs** tab to display all jobs across your project. Select a job from the list to view past executions.

Tip: For further instruction coverage, check out this COVID19 Tracking Blog.

Data connections

- `COVID Tracking: Colleges` : Connection configuration to the latest US colleges data set in the repo.
- `COVID Tracking: FIPS` : Connection configuration to the US counties data set in the repo.

DataStage flow

- `COVID Tracking: Colleges in Providence` : The flow canvas to extract and customize the data sets listed.

Resources

- New York Times COVID-19 Public Datasets
- The New York Times
- DataStage usage documentation

This project comprises of IBM service demonstration generated by approved third-party data and does not contain IBM proprietary information.