

IBM Data Integration

Hands-on Lab Guide

Amin Abou-Gallala

Americas Data Technical Specialist

Bob Reno

WW Data Technical Specialist

DataStage Core Lab

Environment Access

Use the cluster link provided by your instructor

1. Access the **shared cluster 1**:

<https://ibm.biz/BdnLHM>

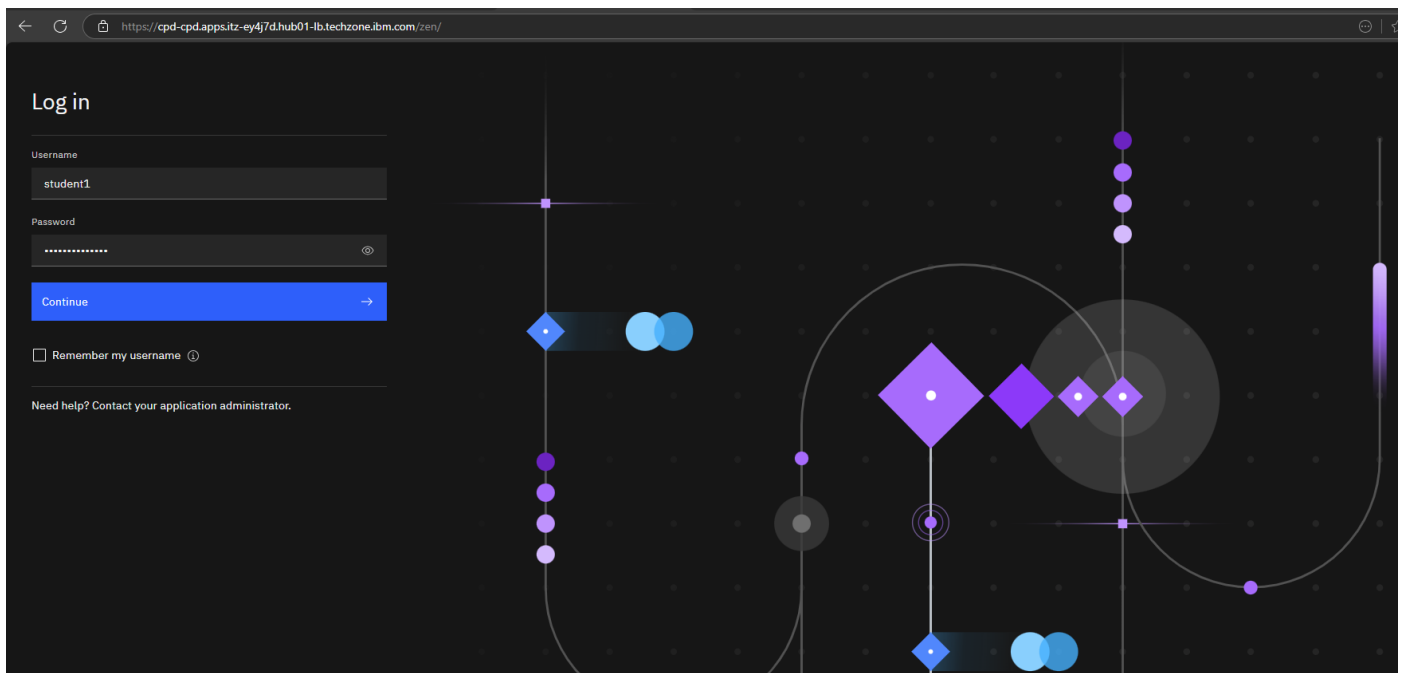
Key in your student id: **student1 – student20** as username

Access the **shared cluster 2**:

<https://ibm.biz/BdntC9>

Key in your student id: **student21 – student40** as username

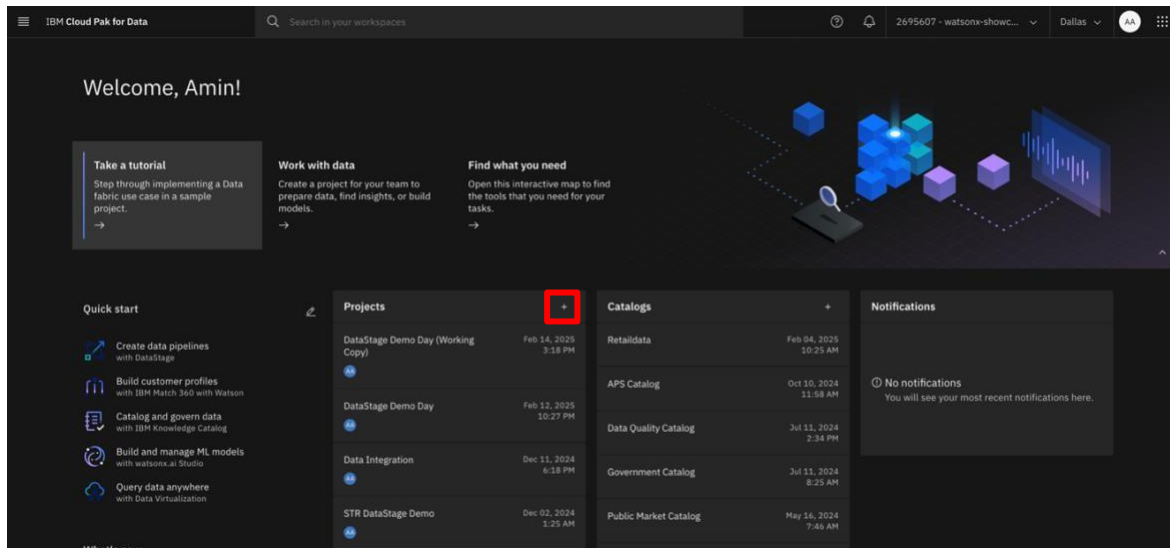
2. Enter **IDtechxchange** for the password.



Lab Setup (Building the flow)

To begin, download the [DataStage Lab Project ZIP file](#) from Box.

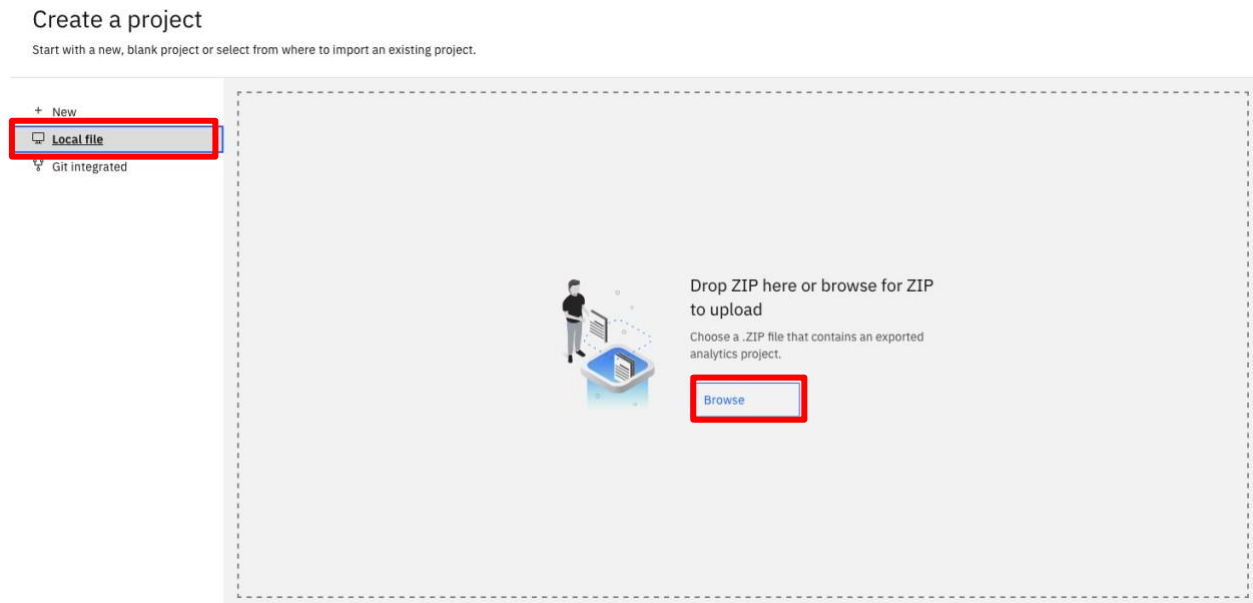
To get started, head to the shared SW cluster above homepage.
Navigate to **projects -> all projects**. Create a new project from file.



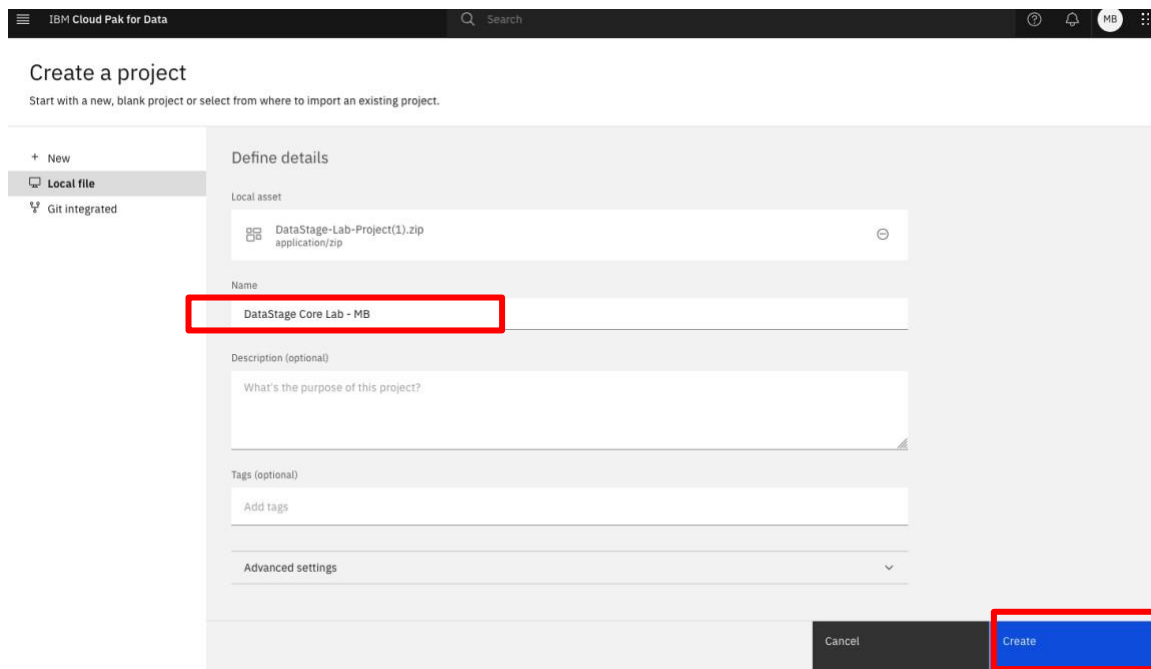
Explainer - DataStage Projects

A **project** is a collaborative workspace in Cloud Pak for Data where you work with data and other assets to accomplish a particular goal. In this case, we are using projects to transform and integrate data with DataStage.

Inside the **Create a Project** view, select **Local file** and upload the downloaded ZIP file.

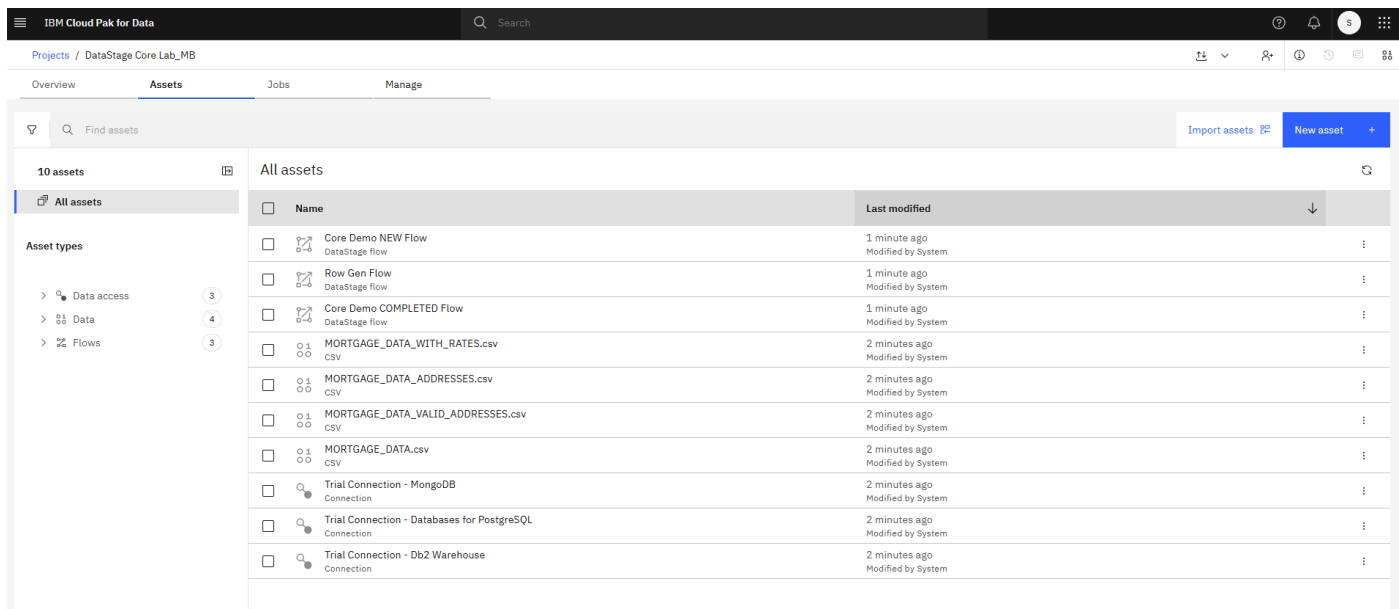


Once the DataStage Core Lab Project ZIP file is uploaded, provide a unique name for your project (i.e. **DataStage Core Lab_Your Initials**)



Click the **Create** button to create the project.

Tip: Once inside the Assets view, sorting the assets by **Name** in ascending order will assist you in navigating the contents inside the Project



Explainer - DataStage Flows

DataStage **flows** are the design-time assets that contain data integration logic.

The basic building blocks of a **flow** are:




- Data sources that read data
- Stages that transform the data
- Data targets that write data
- Links that connect the sources, stages, and targets

This Core Lab section is comprised of two DataStage flows in the project, a **NEW** Flow and a **COMPLETED** Flow.

- The **NEW** Flow is the starting point for building the core capabilities of DataStage in the Core Lab.
- The **COMPLETED** Flow is the final output of the Core Lab, allowing you to have a reference example of the completed flow.

<input type="checkbox"/>	Name	Last modified
<input type="checkbox"/>	Core Demo COMPLETED Flow DataStage flow	9 minutes ago Modified by Service
<input type="checkbox"/>	Core Demo NEW Flow DataStage flow	Now Modified by you

Rename these DataStage flows to include your initials.
This will help in identifying your assets.

<input type="checkbox"/>	Name	↑	Last modified
<input type="checkbox"/>	 Core Demo COMPLETED Flow DataStage flow		 6 minutes ago Modified by Service
<input type="checkbox"/>	 Core Demo NEW Flow_AAG DataStage flow		5 minutes ago Modified by you

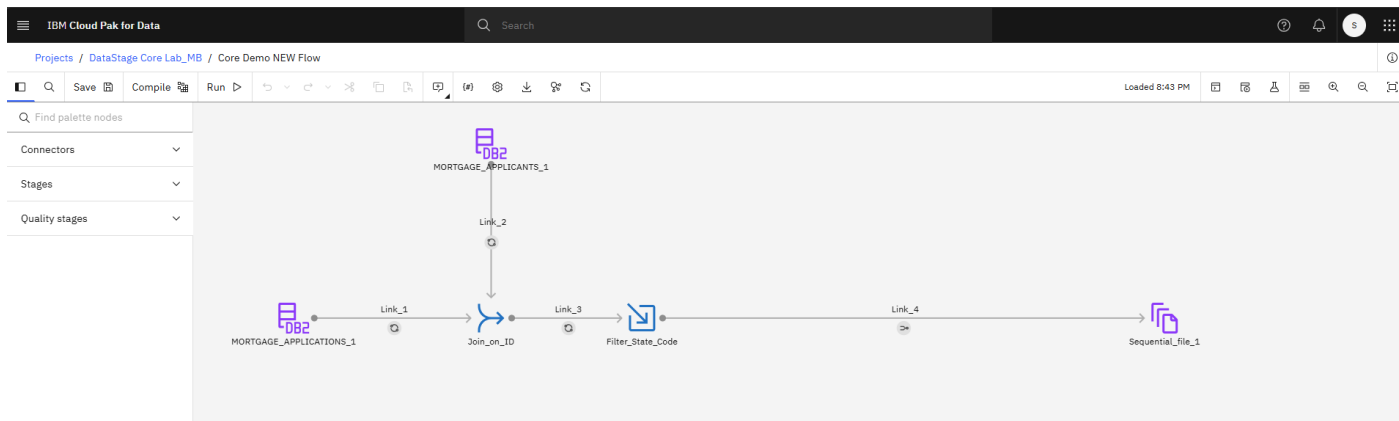
<input type="checkbox"/>	Name	↑	Last modified
<input type="checkbox"/>	 Core Demo COMPLETED Flow_AAG		 5 minutes ago Modified by Service
<input type="checkbox"/>	 Core Demo NEW Flow_AAG DataStage flow		5 minutes ago Modified by you



Click into the NEW Flow to get started building your DataStage job!

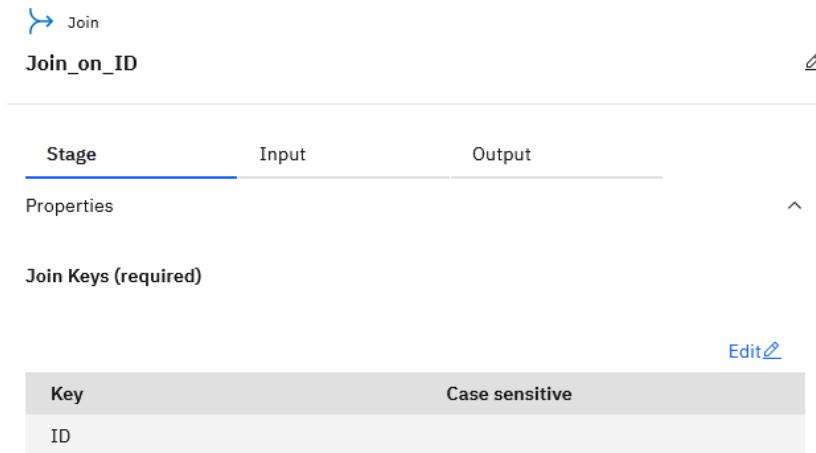
Task 1: Run an existing DataStage Flow

Let's with a basic DataStage flow that joins the mortgage applicants and mortgage applications data sets, then outputs that result to a CSV file in the project. Follow these steps to run the DataStage flow:


1. Start in your **DataStage Core Lab <Your Initials>** project.
If you don't have the project open, follow these steps:
 - a. From the Navigation menu Navigation menu, choose **Projects > View all projects**.
 - b. Open the Data integration project.
2. Click the Assets tab to see all the assets in the project.
3. Click Flows > DataStage flows.
4. Click the **Core Demo NEW Flow_<your initials>** flow in the list to open it. This flow joins the *Mortgage Applicants* and *Mortgage Applications* tables that are stored in Db2 Warehouse, filters the data to those records from the State of California, and creates a sequential file in CSV format as the output.

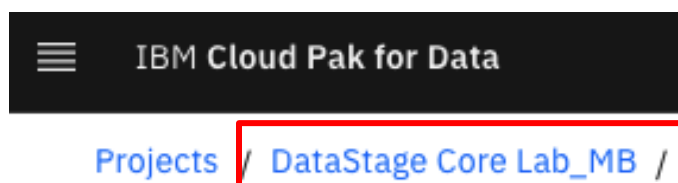


5. Click the **zoom in** icon  zoom in and **zoom out** icon  on the toolbar to set your preferred view of the canvas.
6. Double-click **MORTGAGE_APPLICATIONS_1** node to view the settings.
 - a. Expand the **Properties** section.
 - b. Scroll down to the bottom, then click **Preview data**. This data set includes information that is captured on a mortgage application.
 - c. Click **Close**.
7. Double-click **MORTGAGE_APPLICANTS_1** node to view the settings.
 - a. Expand the **Properties** section.
 - b. Scroll down, and click **Preview data**. This data set includes information about mortgage applicants who applied for a loan.
 - c. Optional: Visualize the data.
 - i. Click the **Chart** panel.
 - ii. In the *Columns to visualize* list, select **STATE**.
 - iii. Click **Visualize data** to see a pie chart showing the distribution of the data by state.
 - iv. Click the **Treemap** icon to see the same data in a treemap chart.
 - d. Click **Close**.
8. Double-click **Join_on_ID** node to view the settings.
 - a. Expand the **Properties** section.
 - b. Note that the join key is the ID column.



c. Click Cancel to close the settings.

9. Click the **Logs** icon  on the toolbar so you can watch the flow's progress.
10. Click **Compile**, and then click **Run**. Alternatively, you can click **Run** which compiles and then runs the DataStage flow. The run can take about one minute to complete.
11. View the logs. You can use the total rows and rows/sec for each step in the flow to visually verify that the filter is working as expected.
12. When the run completes successfully, click **DataStage Core Lab _ < Your Initials>** (your project name may differ) in the navigation trail to return to the project.



13. On the *Assets* tab, click **Data > Data assets**.
14. Open the **MORTGAGE_DATA.CSV file**. You can see that this file contains the columns from both the mortgage applicants and mortgage applications data sets.

Check your progress

The following image shows resulting CSV file. The next task is to edit the DataStage flow.

ID	INCOME	APPLIEDONLINE	RESIDENCE	YRS_AT_CURRENT_ADDRESS	YRS_WITH_CURRENT_EMPLOYER	NUMBER_OF_CARDS	CREDITCA
100272	43593	NO	Owner Occu...	13	0	1	2315
100306	43650	NO	Public Housi...	8	12	1	267
100377	51945	YES	Public Housi...	19	0	1	25
100388	48511	YES	Owner Occu...	21	25	1	793
100415	53791	NO	Owner Occu...	1	15	1	1844
100442	44271	NO	Private Renti...	17	20	1	284
100451	46949	NO	Public Housi...	0	2	2	1634
100453	45869	NO	Private Renti...	21	8	2	71
100473	43750	YES	Private Renti...	13	6	1	988
100477	43761	YES	Owner Occu...	10	6	1	1418
100507	43633	NO	Public Housi...	14	16	1	6100
100512	45621	NO	Owner Occu...	1	19	1	1878
100516	45849	YES	Living with p...	13	9	2	364
100521	44394	YES	Owner Occu...	20	17	1	2041
100529	50075	NO	Private Renti...	4	18	2	3768
100530	43303	NO	Owner Occu...	2	7	2	866
100536	45450	NO	Owner Occu...	15	24	1	4217
100600	45537	NO	Private Renti...	22	21	2	2698

Overview: Edit the DataStage Flow

Now that you joined the mortgage applicant and application data, you are ready to edit the DataStage flow to:

- Task 2: Specify a key column for the Join stage.
- Task 3: Add credit score data from a PostgreSQL database.
- Task 4: Add a Join stage to join the credit score data with the applicant and application data.
- Task 5: Add a Transformer stage to calculate total debt.
- Task 6: Add interest rate data from a MongoDB database.
- Task 7: Add a Lookup stage to look up interest rates for applicants based on their credit scores and Golden Bank's daily interest rate ranges.

Explainer - DataStage Stages

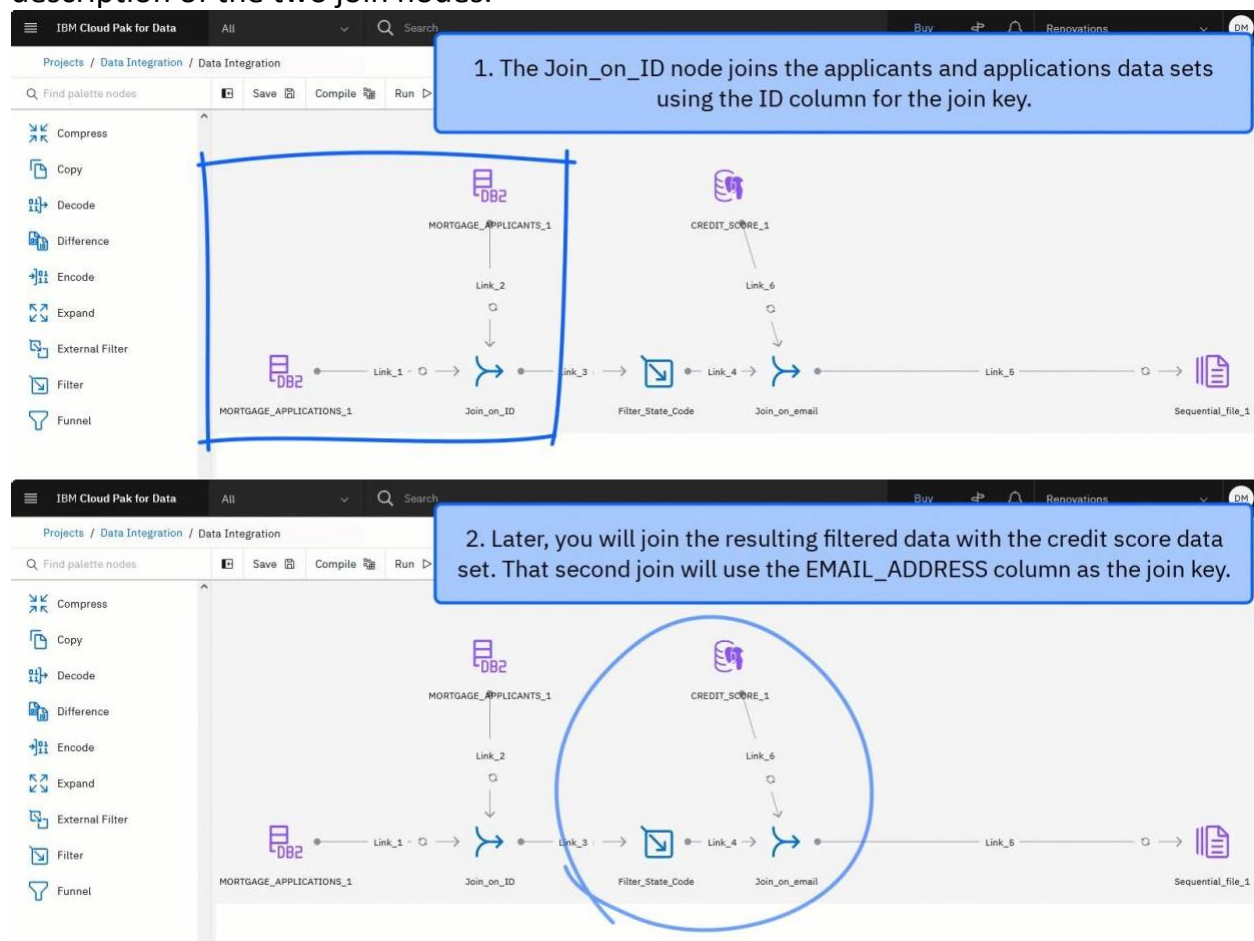
A DataStage flow consists of **stages** that are linked together, which describe the flow of data from a data source to a data target. A stage describes a data source, a processing step, or a target system. The stage also defines the processing logic that moves the data from the input links to the output links.

A stage usually has at least one data input or one data output. However, some stages can accept more than one data input, and output to more than one stage. The following [documentation link](#) lists the available stages and gives details on their functions.

Task 2: Specify the key column for the Join stage

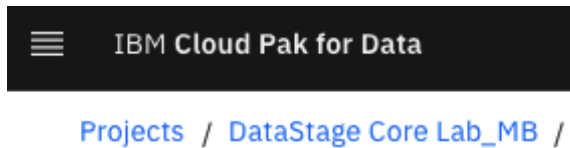
Identifying a key column indicates to DataStage that column contains unique values. The **Join_on_ID** node joins the mortgage applicants and mortgage applications data sets using the ID column for the join key. The next phase is to join the resulting data set with the credit score data. Later, you will join the resulting filtered data with the credit score data set. The second join will use the EMAIL_ADDRESS column as the join key. In this task, you edit the DataStage flow to specify the EMAIL_ADDRESS column as the key column for the resulting data set when it is joined with the credit score data.

The following images provide a visual representation as an alternative to the description of the two join nodes.



Follow these steps to change the Join node settings:

Click **DataStage Core Lab_ <Your Initials>** (your project name may differ) in the navigation trail to return to the project.



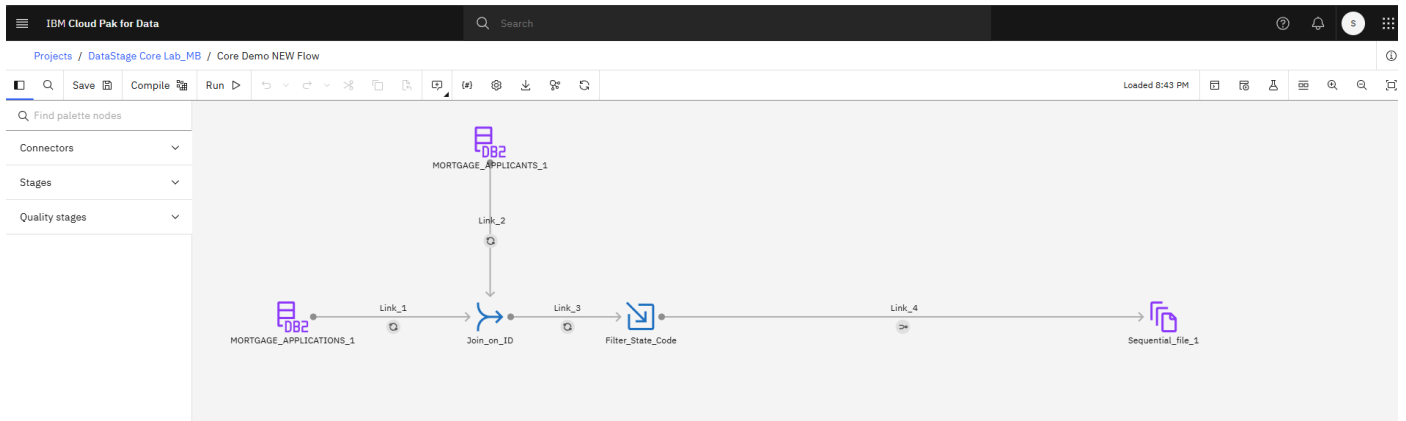
1. On the **Assets** tab, click **Flows > DataStage flows**.
2. Open the **Core Demo NEW Flow_<your initials>** flow.
3. Double-click the **Join_on_ID** node to edit the settings.
4. Click the **Output** tab, and expand the **Columns** section to see a list of the columns in the joined data set.
5. Click **Edit**.
6. For the *EMAIL_ADDRESS* column name, select **Key**.

<input type="checkbox"/>	Column name	Data type	Length	Scale	Key
<input type="checkbox"/>	CITY	NVARCHAR	1024	–	<input type="checkbox"/>
<input type="checkbox"/>	STATE	NVARCHAR	1024	–	<input type="checkbox"/>
<input type="checkbox"/>	STATE_CODE	NVARCHAR	1024	–	<input type="checkbox"/>
<input type="checkbox"/>	ZIP_CODE	NVARCHAR	1024	–	<input type="checkbox"/>
<input type="checkbox"/>	EMAIL_ADDRESS	NVARCHAR	1024	–	<input checked="" type="checkbox"/>

7. Click **Apply and return** to return to the *Join_on_ID* node settings.
8. Click **Save** to save the *Join_on_ID* node settings.

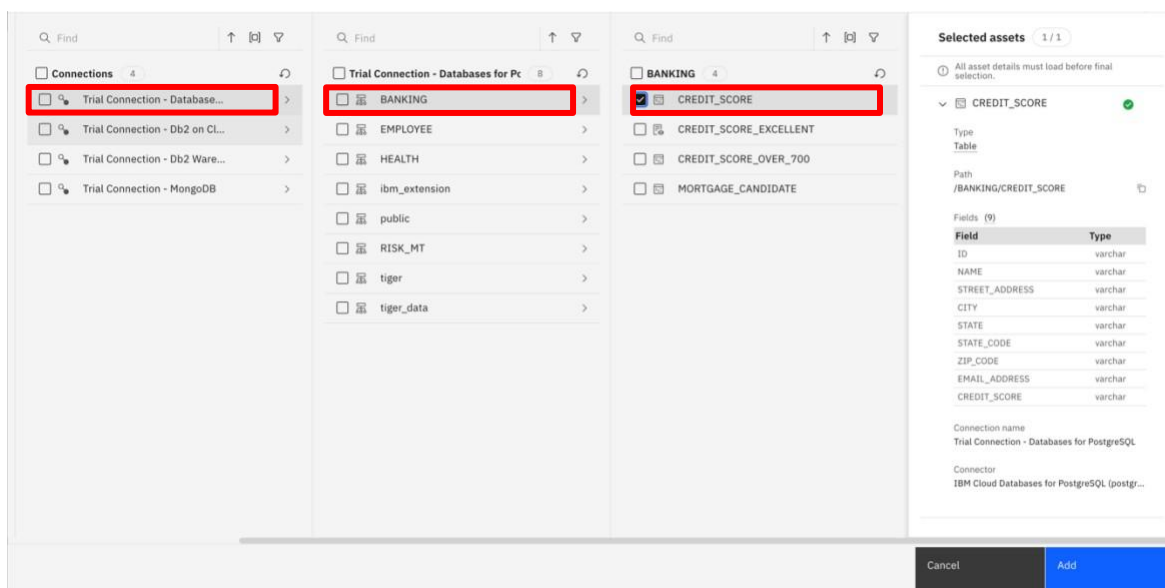
Check your progress

The following image shows the DataStage flow with the edited Join_on_id stage. Now that you identified the EMAIL_ADDRESS column as the key column, you can add the **PostgreSQL** data containing the applicants credit scores.



Explainer - Asset Browser for DataStage

The Asset browser is used to search for connections and assets and add them to your DataStage flows. When you open the asset browser from the palette, you can use it to browse connectors, DataStage subflows, and data assets (.csv, .txt, .xls., .xlsx, .xml, .json files). The data can then be previewed before being onboarded to the DataStage flow – this feature saves developers time searching for the right data and is an important benefit.

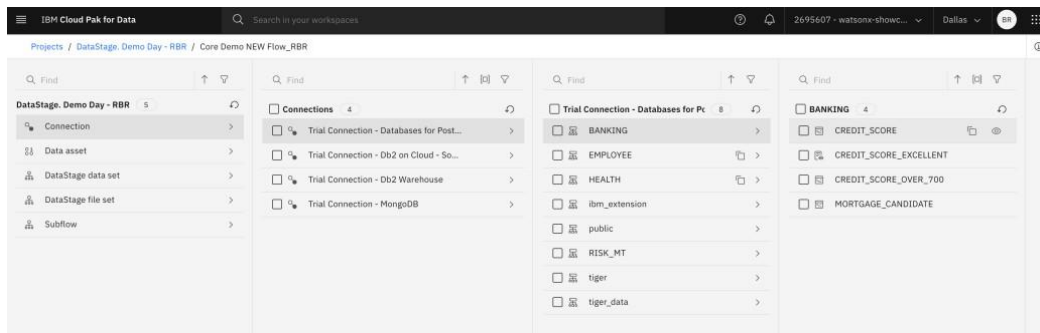


Task 3: Add credit score data from a PostgreSQL database

Follow these steps to add the credit score data that is stored in a PostgreSQL database to the DataStage flow:

1. In the node palette, expand the **Connectors** section.
2. Drag the **Asset browser** connector to the canvas beside the *MORTGAGE_APPLICANTS_1* node.
3. Locate the asset by selecting **Connection > Trial Connection - Databases for PostgreSQL > BANKING > CREDIT_SCORE**.

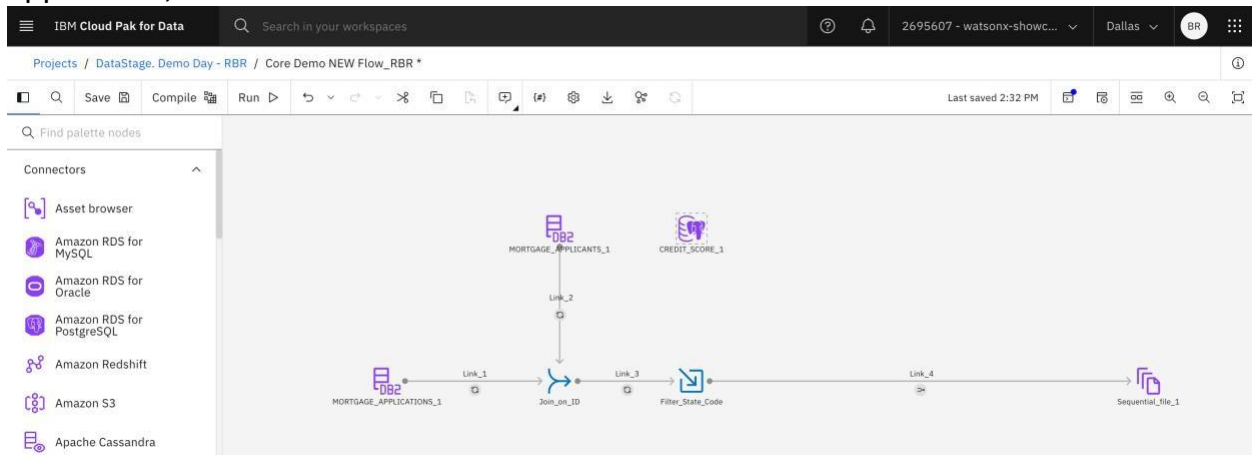
Note: Click the connection or schema name instead of the checkbox to expand the connection and schema.



4. Click the **Preview** icon (🔍) to preview the credit score data for each applicant.
5. Click **Add**.

Check your progress

The following image shows the DataStage flow with the credit score asset added. Now that you added the credit score data to the canvas, you need to join the applicant, application, and credit score data.



Task 4: Add a Join stage to join the credit score data with the applicant and application data

Follow these steps to add another Join stage to join the filtered mortgage application and mortgage applicant joined data with the credit score data in the DataStage flow:

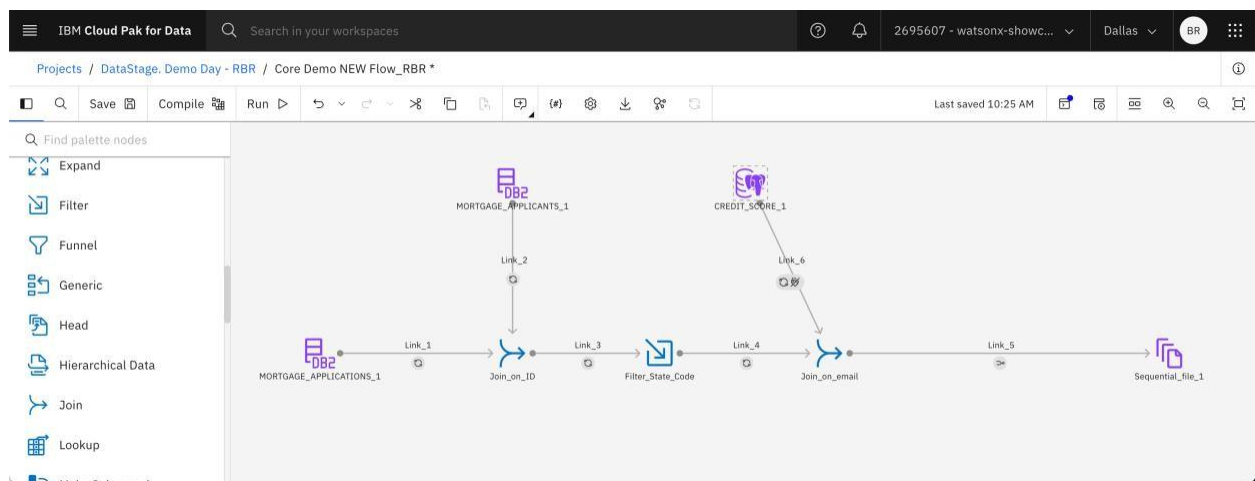
1. In the node palette, expand the **Stages** section.
2. Drag the **Join** stage on to the canvas, and drop the node on the link line between the *Filter_State_Code* and *Sequential_file_1* nodes.
3. Hover over the **CREDIT_SCORE_1** connector to see the arrow. Connect the arrow to the **Join** stage.
4. Double-click the **CREDIT_SCORE_1** node to edit the settings.
 - a. Click the **Output** tab, and expand the **Columns** section to see a list of the columns in the joined data set.
 - b. Click **Edit**.
 - c. For the *EMAIL_ADDRESS* and *CREDIT_SCORE* column names, select **Key**.
 - d. Click **Apply and return** to return to the *CREDIT_SCORE_1* node settings.
 - e. Click **Save** to save the *CREDIT_SCORE_1* node settings.
5. Double-click the **Join_1** node to edit the settings.
 - a. Expand the **Properties** section.
 - b. Click **Add key**.
 - i. Click **Add key** again.
 - ii. Select **EMAIL_ADDRESS** from the list of possible keys.
 - iii. Click **Apply**.



- c. Click **Apply and return** to return to the *Join_1* node settings.
- d. Change the *Join_1* node name to **Join_on_email**.
- e. Click **Save** to save the *Join_1* node settings.



Check your progress

The following image shows the DataStage flow with a second Join stage added. Now that you joined the application, applicant, and credit score data, you need to add a Transformer stage to calculate each applicant's total debt.



Task 5: Add a Transformer stage to calculate total debt

Follow these steps to add a Transformer stage that creates a new column by summing the LOAN_AMOUNT and CREDITCARD_DEBT columns:

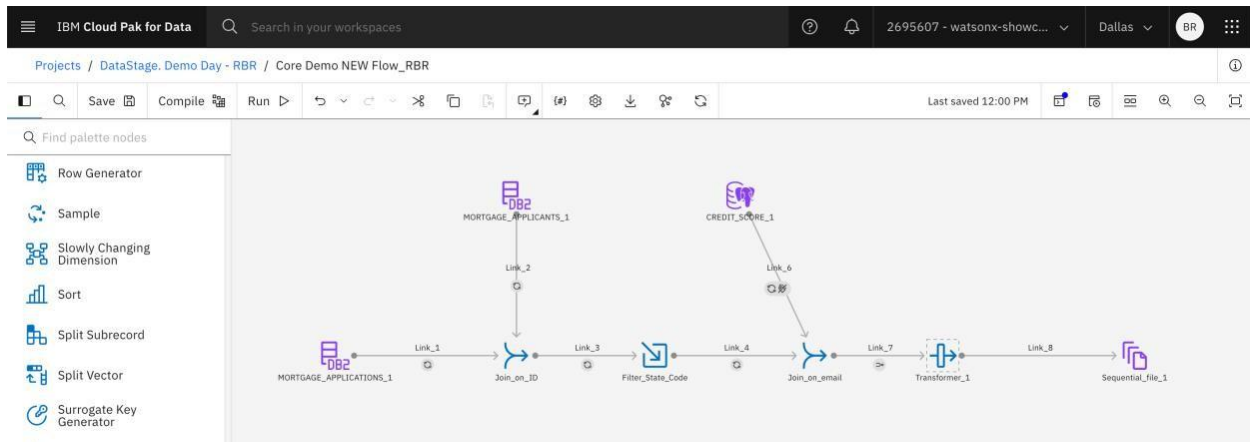
1. In the *Stages* section, drag the **Transformer** stage on to the canvas, and drop the node on the link line between the *Join_on_email* and *Sequential_file_1* nodes.
2. Double-click the **Transformer** node to edit the settings.
3. Click the **Output** tab.
 - a. Click **Add column**.
 - b. Scroll down in the list of columns to see the new column.
 - c. Name the column **TOTAL_DEBT**.
 - d. Click the **Edit** icon  in the row's *Derivation* column.
 - e. Click the Calculator icon  in the *Derivation* column to open the expression builder.
 - f. Search for **LOAN_AMOUNT**, and double-click the column name to add it to the expression. Note that the link number is appended to the column name.
 - g. Type a plus sign **+**.
 - h. Search for **CREDITCARD_DEBT**, and then double-click the column name to add it to the expression. Note that the link number is appended to the column name.
 - i. Verify that the final expression is **Link_7.LOAN_AMOUNT + Link_7.CREDITCARD_DEBT**.

Note: Your link number may be different.

 - j. Click **Apply and return** to return to the *Transformer* page. If you see an error, change the column **Data type** to **nvarchar**.
 - k. For the *CREDIT_SCORE* column name, scroll to the right and select Key.
4. Click the **Stage** tab.
 - a. Select the **Advanced** page.
 - b. Change the *Execution mode* to **Sequential**.
5. Click **Save and return** to return to the canvas.


Check your progress

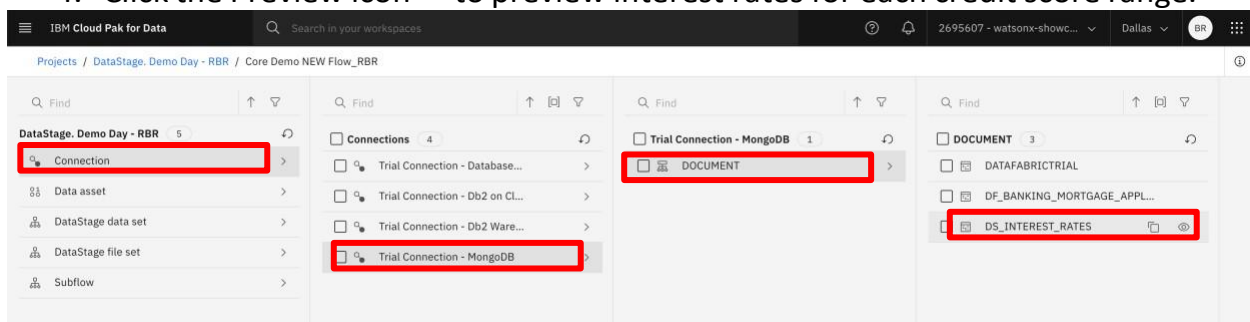
The following image shows the DataStage flow with the Transformer stage added. Now that you calculated each applicant's total debt, you need to add the table of interest rates to offer based on credit score ranges.



Task 6: Add interest rate data from a MongoDB database

Follow these steps to include the interest rates in the flow by adding a data asset connector to a MongoDB database:

1. In the node palette, expand the **Connectors** section.
2. Drag the **Asset browser** connector on to the canvas beside the **CREDIT_SCORE_1** node.
3. Locate the asset by selecting **Connection > Trial Connection - MongoDB > DOCUMENT > DS_INTEREST_RATES**.
4. Click the Preview icon  to preview interest rates for each credit score range.

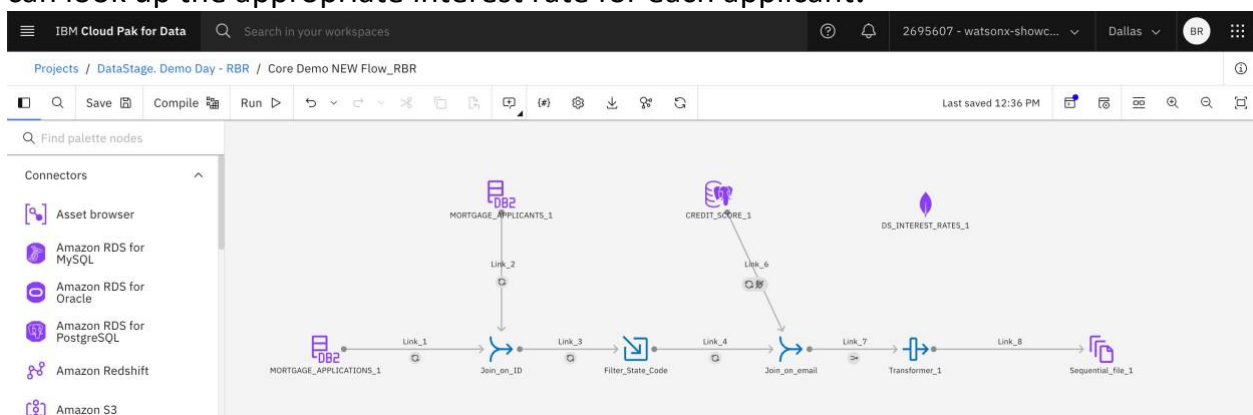


You can use the values in the `STARTING_LIMIT` and `ENDING_LIMIT` columns to look up the appropriate interest rate based on the applicant's credit score. The `ID` column is not needed, so you will delete that column in the next step.

5. Click **Add**.

Check your progress


The following image shows the DataStage flow with the interest rates data asset added from the MongoDB external source. Now that you added the interest rates table, you can look up the appropriate interest rate for each applicant.



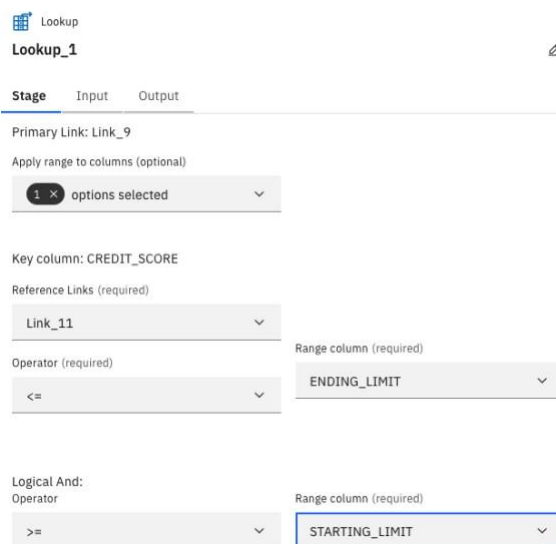
Task 7: Add a Lookup stage to look up interest rates for applicants

Based on each applicant's credit score, you want to look up the appropriate interest rate. Follow these steps to add a Lookup stage and specify the range for starting and ending credit score limits for each interest rate:

1. In the **Stages** section, drag the **Lookup** stage on to the canvas, and drop the node on the link line between the **Transformer_1** and **Sequential_file_1** nodes.
2. Connect the **DS_INTEREST_RATES_1** connector to the **Lookup_1** stage.
3. Double-click the **DS_INTEREST_RATES_1** node to edit the settings.
4. Click the **Output** tab.
 - a. Expand the **Columns** section, and click **Edit**.
 - b. Select the **_ID** column.

- c. Click the **Delete** icon  to delete the **_ID** column.
 - d. Click **Apply and return** to return to the **DS_INTEREST_RATES_1** node settings.
 - e. Click **Save** to save the changes to the **DS_INTEREST_RATES_1** node.
5. Double-click the **Lookup_1** node to edit the settings.
6. Expand the **Properties** section.
 - a. For the **Apply range to columns** field, select **CREDIT_SCORE**. The **Reference Links**, **Operator**, and **Range** column fields display.
 - b. For the **Reference Links**, select **Link_9**.

Note: Your link number may be different.
 - c. For the first **Operator**, select **<=**.
 - d. For the first **Range column**, select **ENDING_LIMIT**.
 - e. For the second **Operator**, select **>=**.
 - f. For the second **Range column**, select **STARTING_LIMIT**.



Lookup
Lookup_1

Stage Input Output

Primary Link: Link_9

Apply range to columns (optional)

1 x options selected

Key column: CREDIT_SCORE

Reference Links (required)

Link_11

Operator (required)

<=

Range column (required)


ENDING_LIMIT

Logical And: Operator

>=

Range column (required)

STARTING_LIMIT

7. Click the **Output** tab.
 - a. Expand the **Columns** section, and click **Edit**.
 - b. Select the **STARTING_LIMIT** and **ENDING_LIMIT** columns.
 - c. Click the **Delete** icon  to delete these unnecessary **STARTING_LIMIT** and **ENDING_LIMIT** columns.

- d. Click **Apply and return** to return to the **Lookup_1** node settings.
- e. Click **Save** to save the changes to the Lookup_1 node.

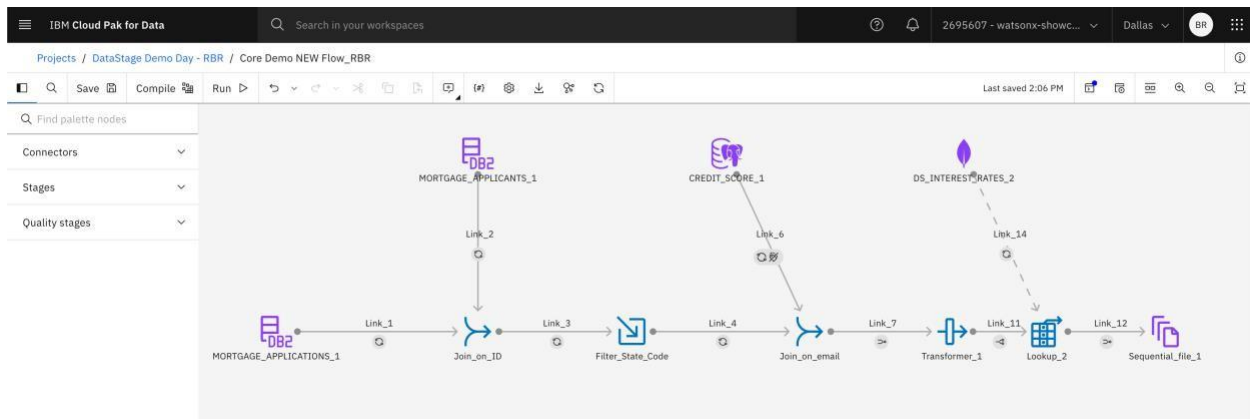
Explainer - Column metadata change propagation

When you add/remove columns or change a column's metadata, **Column metadata change propagation** automatically propagates these changes downstream. For example, when the **STARTING_LIMIT** and **ENDING_LIMIT** columns were deleted, these changes are propagated to the output Sequential File automatically, so those columns will not be seen as a part of the input.

Note: Changes made **upstream** do not apply to a column once you modify its metadata. If you delete a column, modifying the column in a later stage will not add the column back.

Check your progress

The following image shows that the DataStage flow with the Lookup stage added. The DataStage flow is now complete. The last task before running the flow is to specify the name for the output file.



Task 8: Edit the Sequential file node and run the DataStage flow

Follow these steps to edit the Sequential file node to create a final output file as a data asset in the project, and then compile and run the DataStage flow:

1. Double-click the **Sequential_file_1** node to edit the settings.

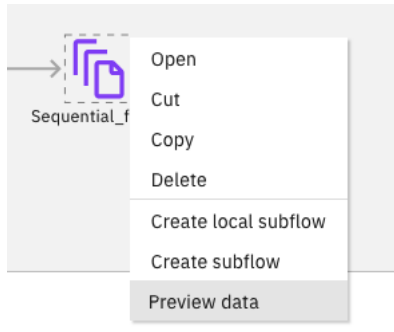
- Click the **Input** tab.
- Expand the **Properties** section.
- For the Target File, copy and paste **MORTGAGE_APPLICANTS_INTEREST_RATES.CSV** for the file name.
- Select **Create data asset**.
- For the First line is column names field, select **True**.
- Click **Save**.
- Click **Run** which compiles and then runs the DataStage flow. The job takes about 1 minute to complete.
- Click **Logs** on the toolbar to watch the flow's progress. It is normal to see warnings during the run, and then you see that the flow ran successfully.

Check your progress

The following image shows that the DataStage flow ran successfully!

The screenshot displays the IBM Cloud Pak for Data interface. At the top, the breadcrumb navigation shows 'Projects / DataStage Demo Day - RBR / Core Demo NEW Flow_RBR'. The toolbar includes icons for Save, Compile, Run, and other actions. A green banner at the top of the workspace area states 'Run successful with warnings.' Below this, a DataStage flow diagram is visible, showing a sequence of jobs: 'MORTGAGE_APPLICANTS_1', 'CREDIT_SCORE_1', and 'DS_INTEREST_RATES_2'. The bottom pane is set to the 'Logs' tab, displaying a list of log entries. The logs include information about job initialization, environment variables, and a warning message: 'WARNING IIS-DSEETFP-00063 <APT_JoinSubOperatorNC in Join_on_email> When checking operator: Dropped component "NAME" because of a prior component with the same name.'

- Hover over the sequential file then click the ellipses to preview the data. The CSV file includes all applicants and their interest rates.



Conclusion and Next Steps

This section covered the core functionalities of DataStage on Cloud Pak for Data, including how to create projects and modify an existing flow to provide a personalized interest rate for each mortgage applicant. Key concepts and features of DataStage's modernized developer experience were covered – including the Asset Browser and Column Metadata Change Propagation, which are new, key features in DataStage on Cloud Pak for Data.

This concludes the lab guide.