**Databand + Next Generation DataStage**

***Hands on Lab***

**This lab is presented by the North America Top team**

*Please reach out to Matt Koscak (*[*matt.koscak@ibm.com*](mailto:matt.koscak@ibm.com)*) with any questions.*

The following material is meant to teach IBMers and Business Partners about the additional monitoring and alerting capabilities that Databand brings to Next Generation DataStage on Cloud Pak for Data as a Service. The DataStage flow we will be observing is the “Multicloud Data Integration” tutorial. We will complete the following exercises:

* Uploading the Multicloud Data Integration flow
* Syncing DataStage/CP4DaaS to Databand
* Viewing the graphical representation of the DataStage job within Databand and the relevant information around this
* Step by step walkthrough functionality of each individual stage
* Dataset metrics and historical trends
* DataStage job inputs and outputs
* Alerting on DataStage jobs

In preparation for this exercise, you must have:

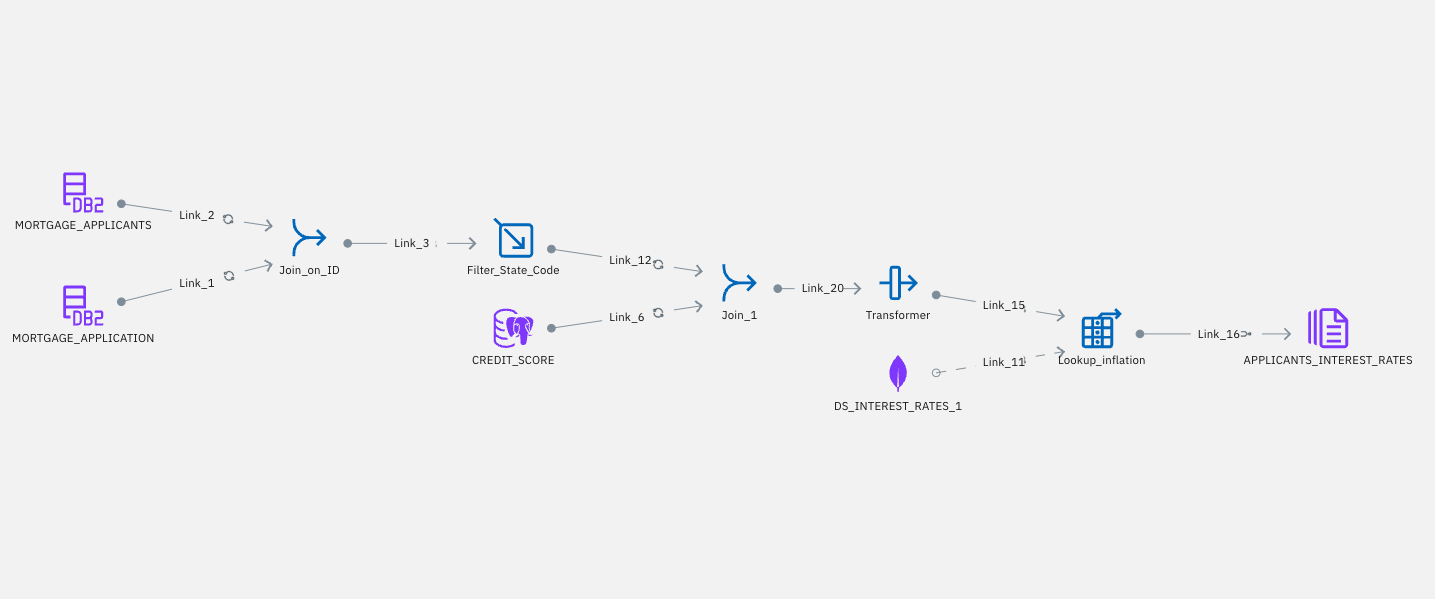
* Gain access to the databand technical sales environment
* Access to a Cloud Pak for Data as a Service (DataStage as a Service) environment
* [The “Multicloud Data Integration.zip” file in the following box folder](https://ibm.box.com/s/icpseagi4rafaa2mmt9et9onmkp6gb86)
* Join the #databand-demo-alerts slack channel in the Data&AI workspace on slack

If you do not have the above prerequisites completed, refer to the links below to gain access to both environments

* [Request access to the Databand technical sales environment](https://form.asana.com/?k=1zj1vm854D55vFWFVUMwtw&d=655517900334182)
* [Link to the databand environment](https://ibm-sales-sandbox.databand.ai/app/dashboard) (Once you are granted access)
* [CP4DaaS (DataStage) environment signup](https://dataplatform.cloud.ibm.com/registration/stepone?context=cpdaas&apps=cos%2Cdatastage&regions=us-south%2Ceu-de&S_PKG=ov80049&cm_mmca1=10000665&cm_mmca2=000000TF)

In addition to this HOL, you can access TechZone assets that may assist with your Databand sales cycles [HERE](https://techzone.ibm.com/collection/2022-data-and-ai-tech-sales-resources-for-data-observability)

**Section 0: Explaining the Multicloud Data Integration Flow/Job we will be observing**

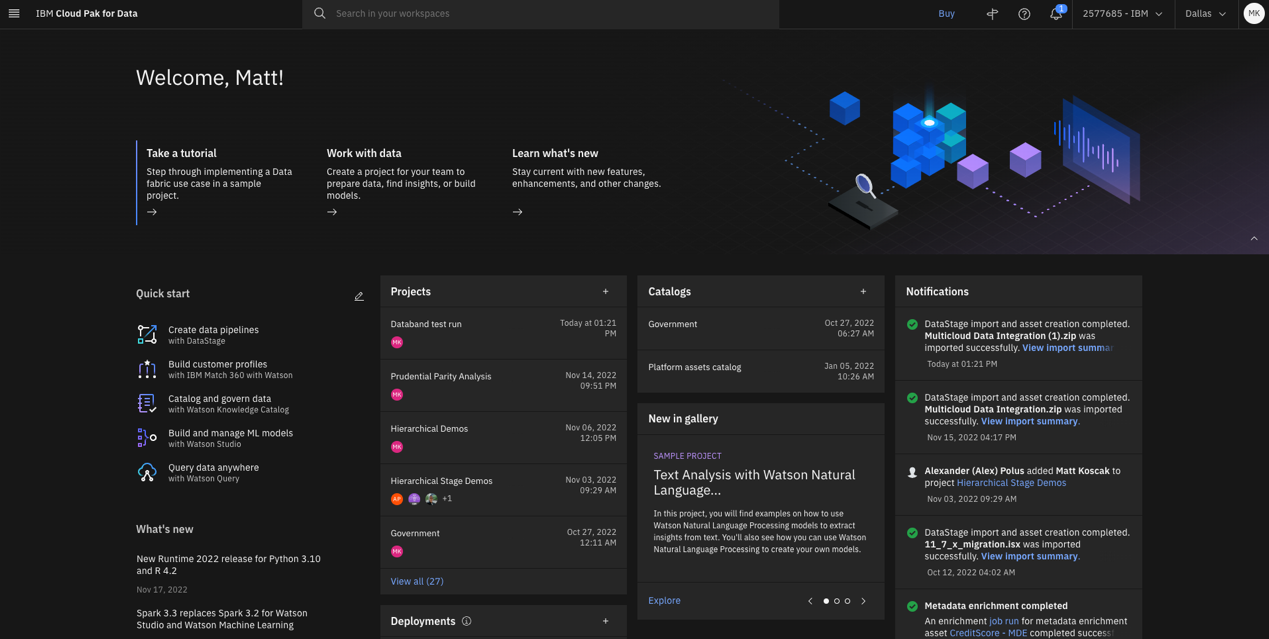


**Screenshot A (above) - Multicloud Data Integration Flow**

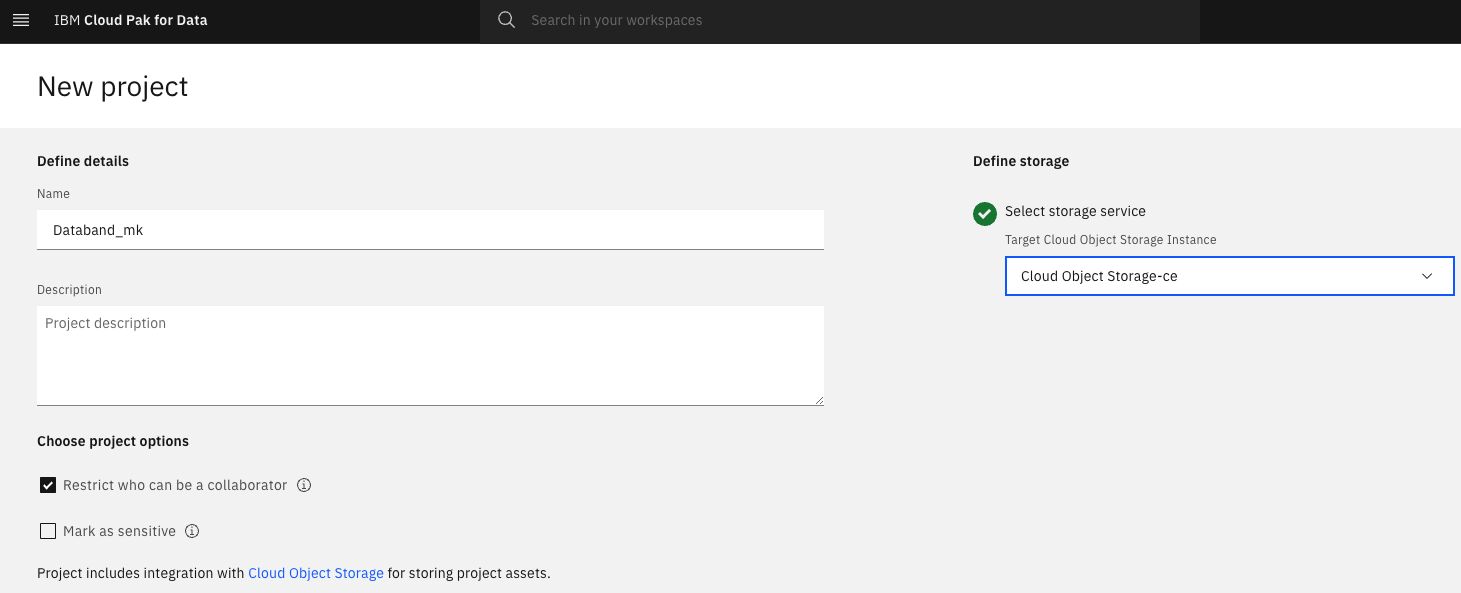
*This Next-Gen DataStage flow integrates data from a Db2 Warehouse on Cloud, Postgres Database, and MongoDB instance. This data is transformed via joining tables, filtering the records by State, calculating a level of debt, and ultimately assigning each individual mortgage applicant an appropriate mortgage rate.*

To begin, follow the procedure below:

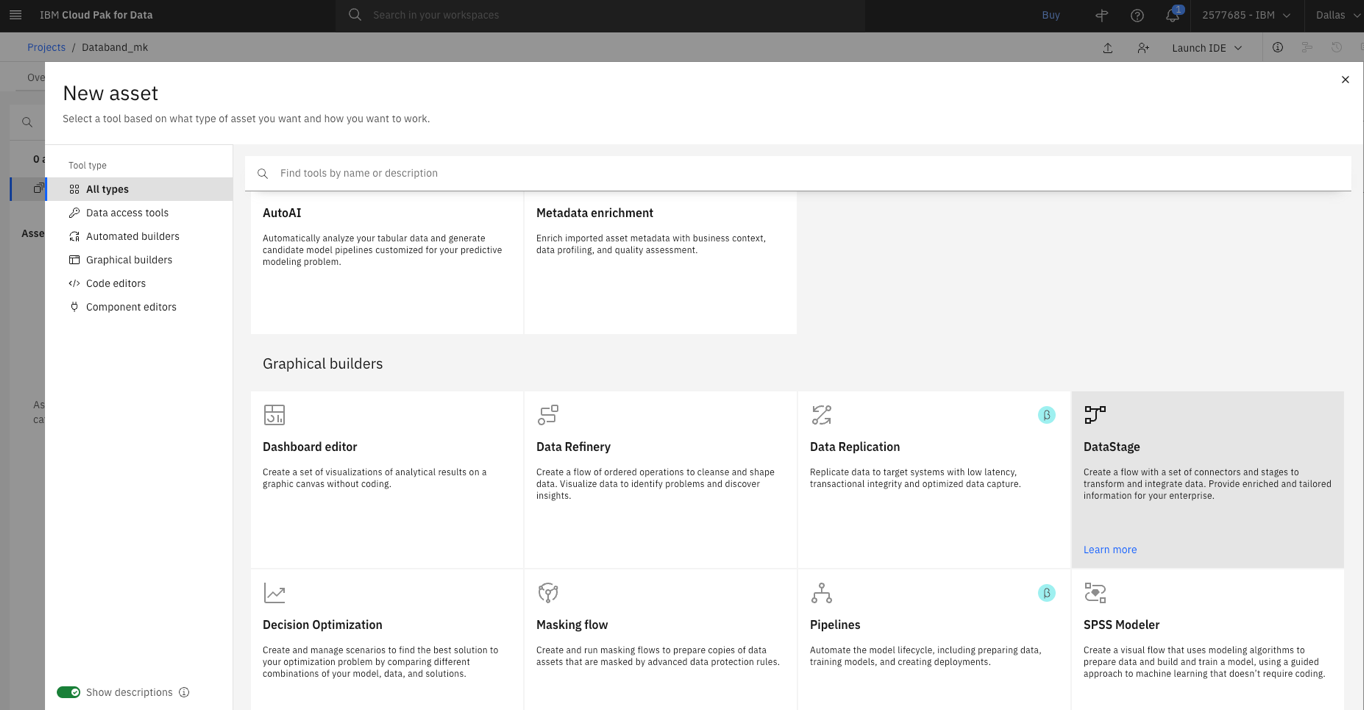
1. Log in to your Cloud Pak for Data as a Service instance on IBM cloud



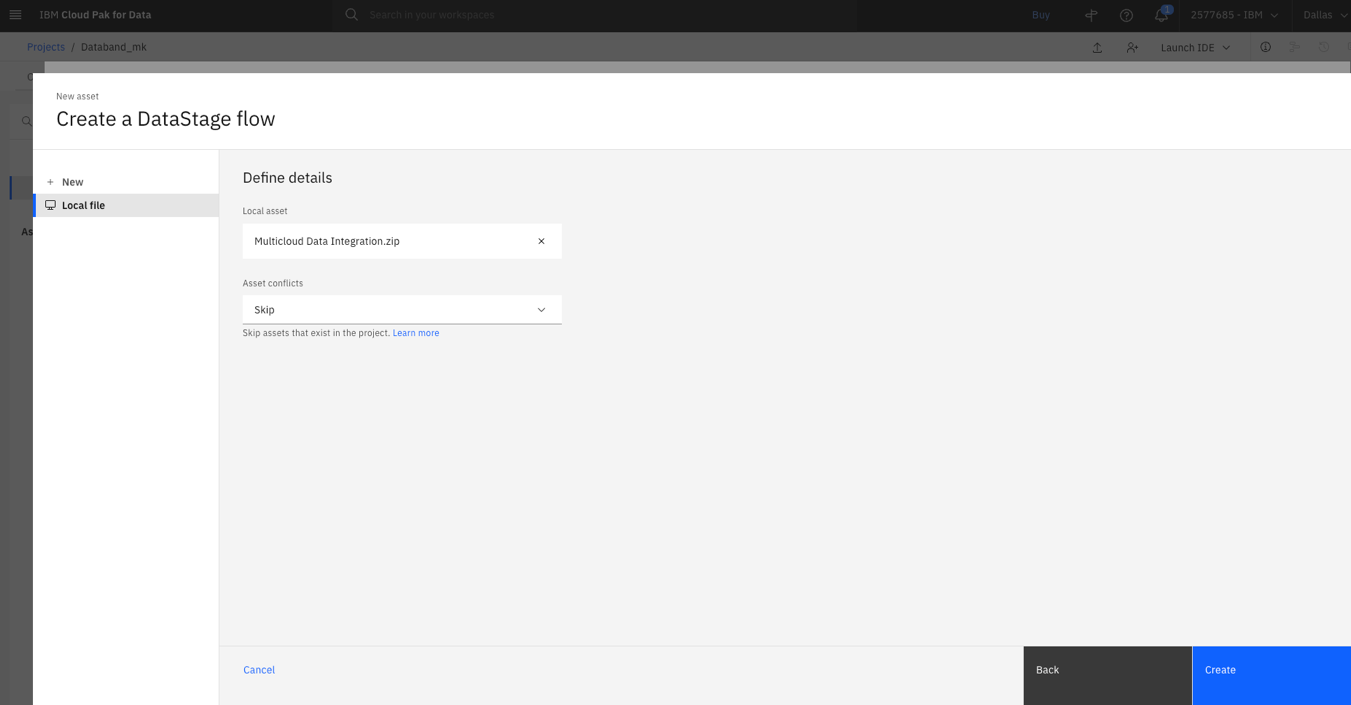
1. Create an empty project named Databand\_yourinitials (I.e. Databand\_mk). Be sure to define the storage service associated with this project. If you do not have a storage service already created please refer to the documentation [HERE](https://cloud.ibm.com/docs/cloud-object-storage?topic=cloud-object-storage-provision).



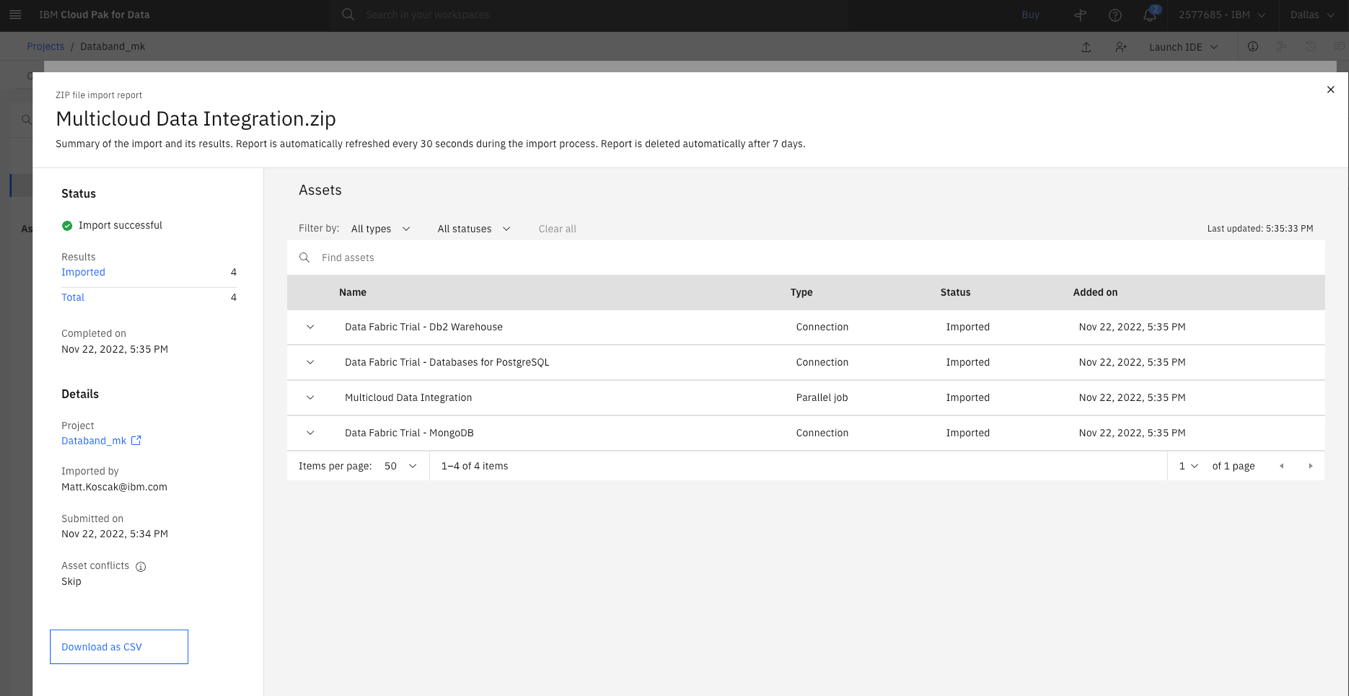
1. Once this project is created, select the ‘Assets’ tab in the project overview screen and click the blue ‘New Asset’ icon



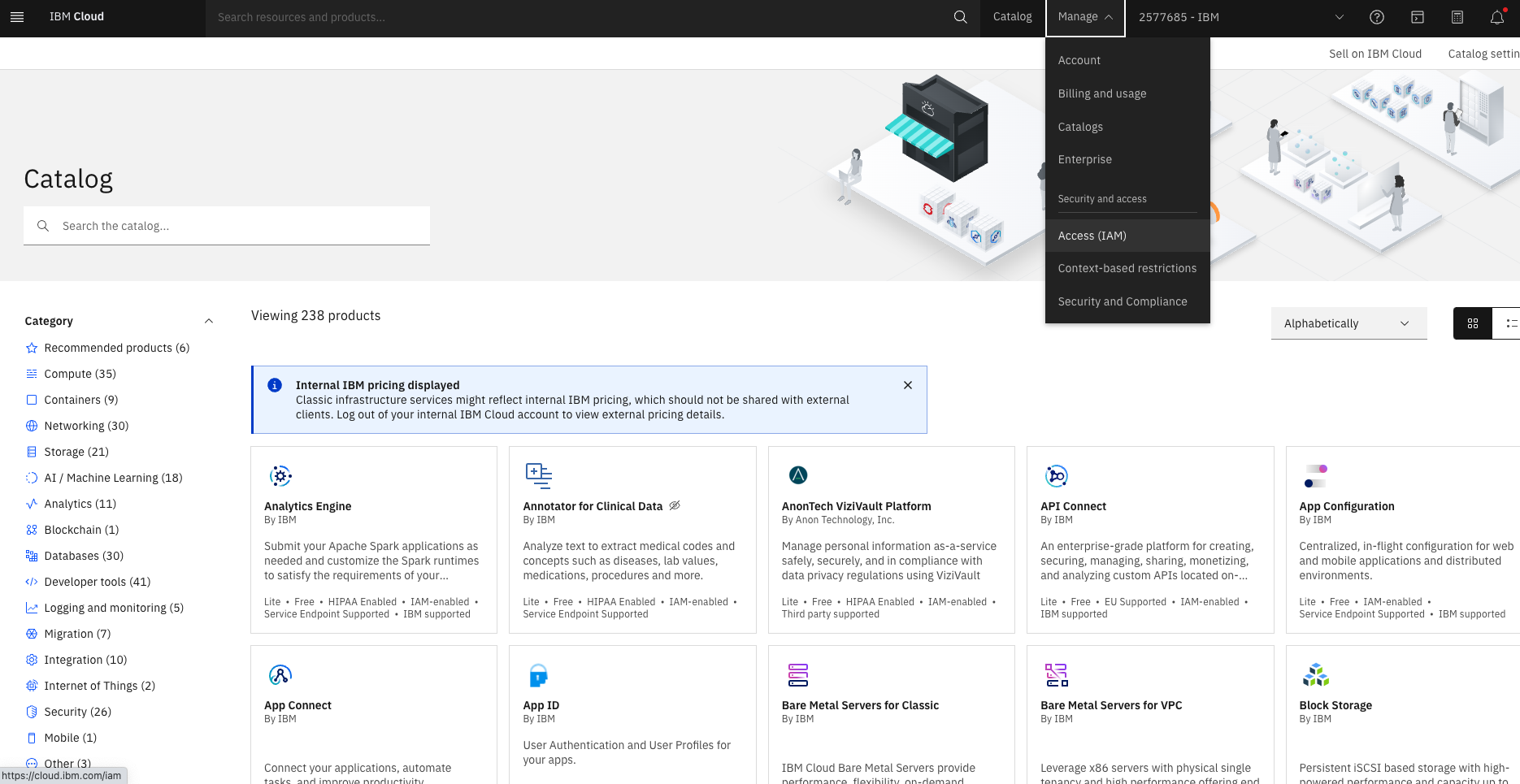
1. Click on the DataStage tile and then click on the “Local file” tab on the lefthand menu. Either drag & drop, or browse for the “Multicloud Data Integration.zip” file that you obtained as a prerequisite. Leave the settings as-is, and press the blue “Create” button. Wait a few moments for the import process to complete.



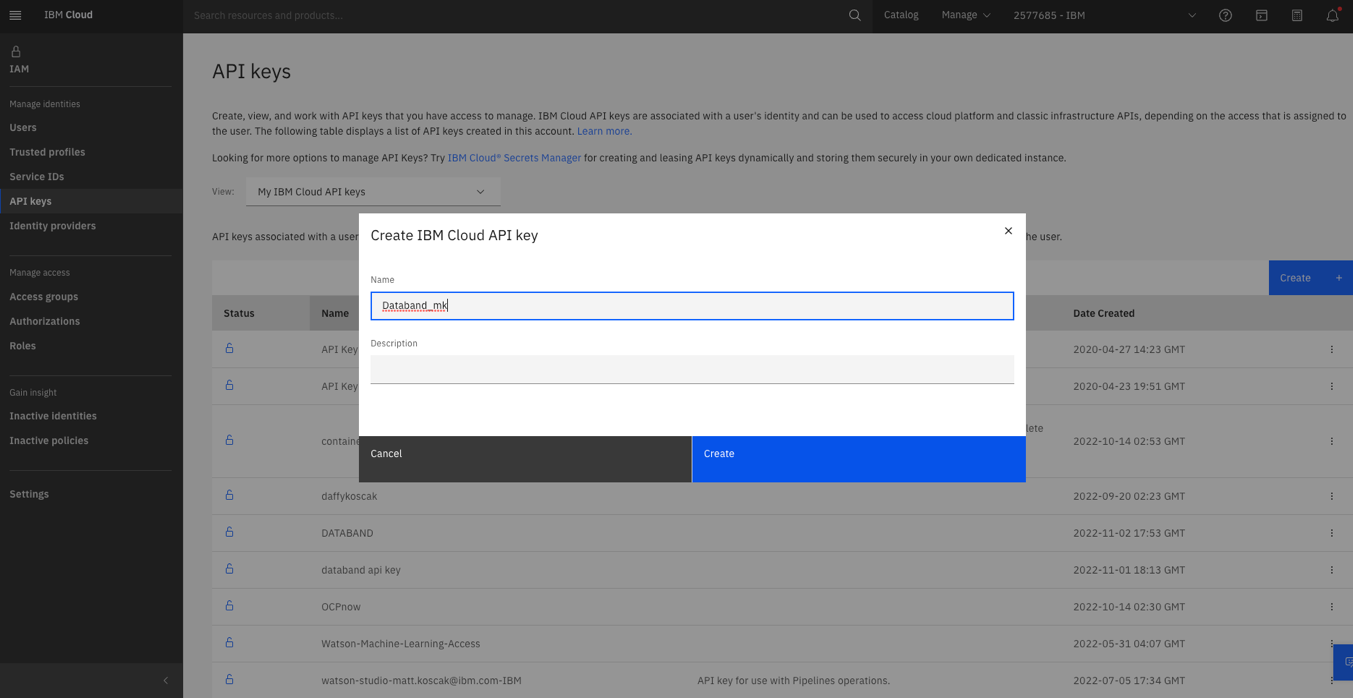
1. After this import process completes, you will see three “Data Fabric Trial” connections, and one Multicloud Data Integration Parallel Job.

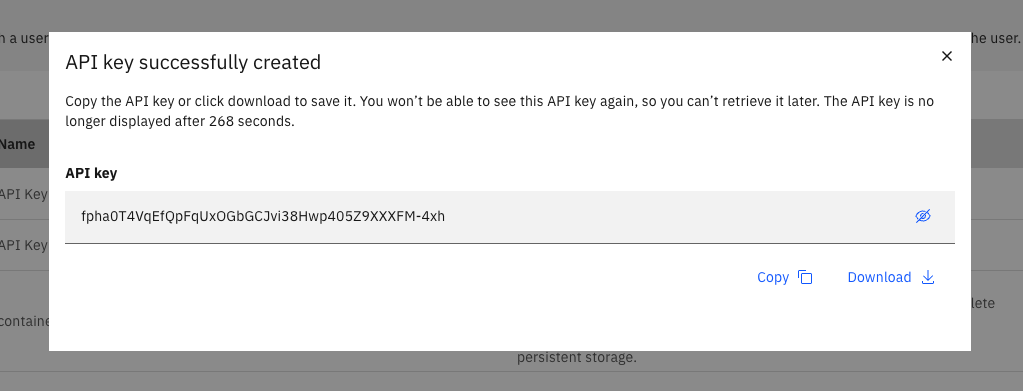


1. Exit the import screen and open the DataStage flow titled “Multicloud Data Integration”. Your DataStage flow should look like the flow in Screenshot A.
2. Your DataStage environment is ready to be integrated with Databand. Open a new web browser tab and go to your IBM cloud console at the weblink [HERE](https://cloud.ibm.com/).
3. After logging in to IBM Cloud, you will now create an API Key for your cloud account. This can be done by clicking the “Manage” dropdown on the top menu bar, and clicking “Access (IAM)”. This is shown in the screenshot below.

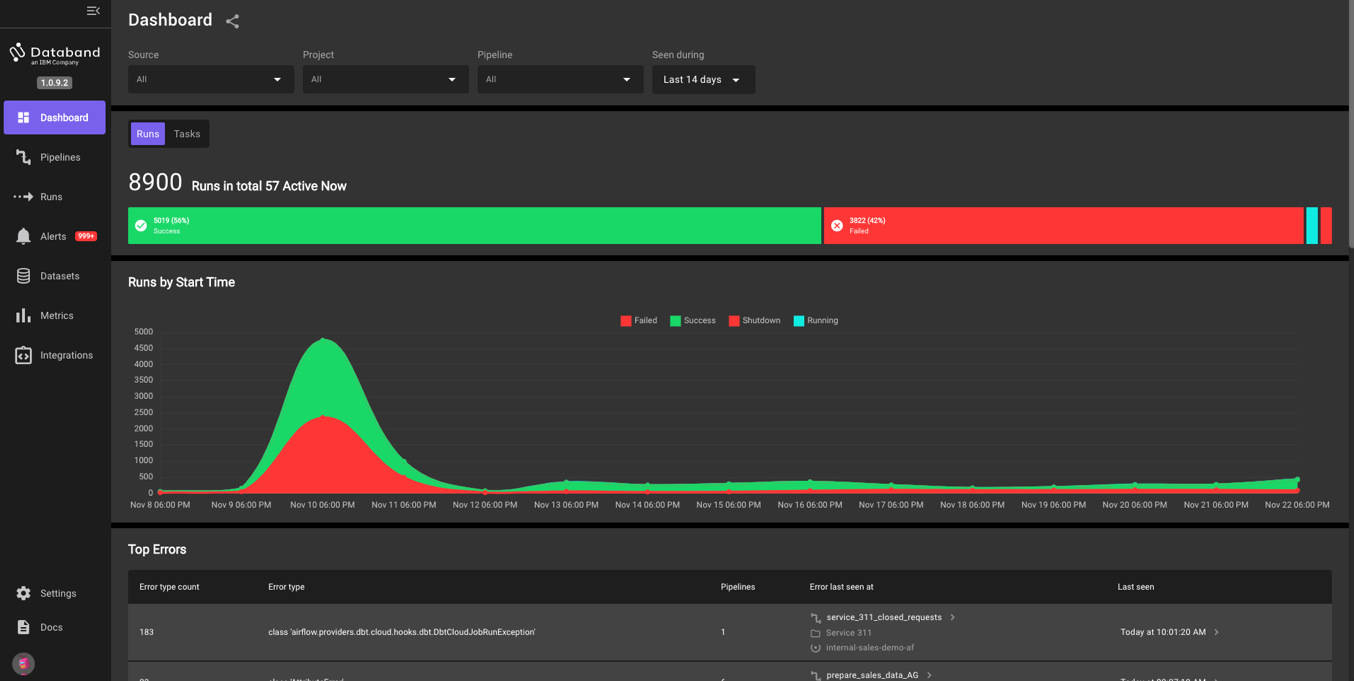


1. On the IAM screen, click the “API keys” tab on the lefthand menu. Click the blue “Create” button, name your API key “Databand\_yourinitals” and click the blue “Create” button. Your API Key will be generated, save this key to a text pad or note elsewhere, as you will need this to create your DataStage Syncer in Databand.
   1. NOTE – You will not be able to see the API key again. If you happen to exit the screen before saving this key, simply delete the key you created and make a new one.

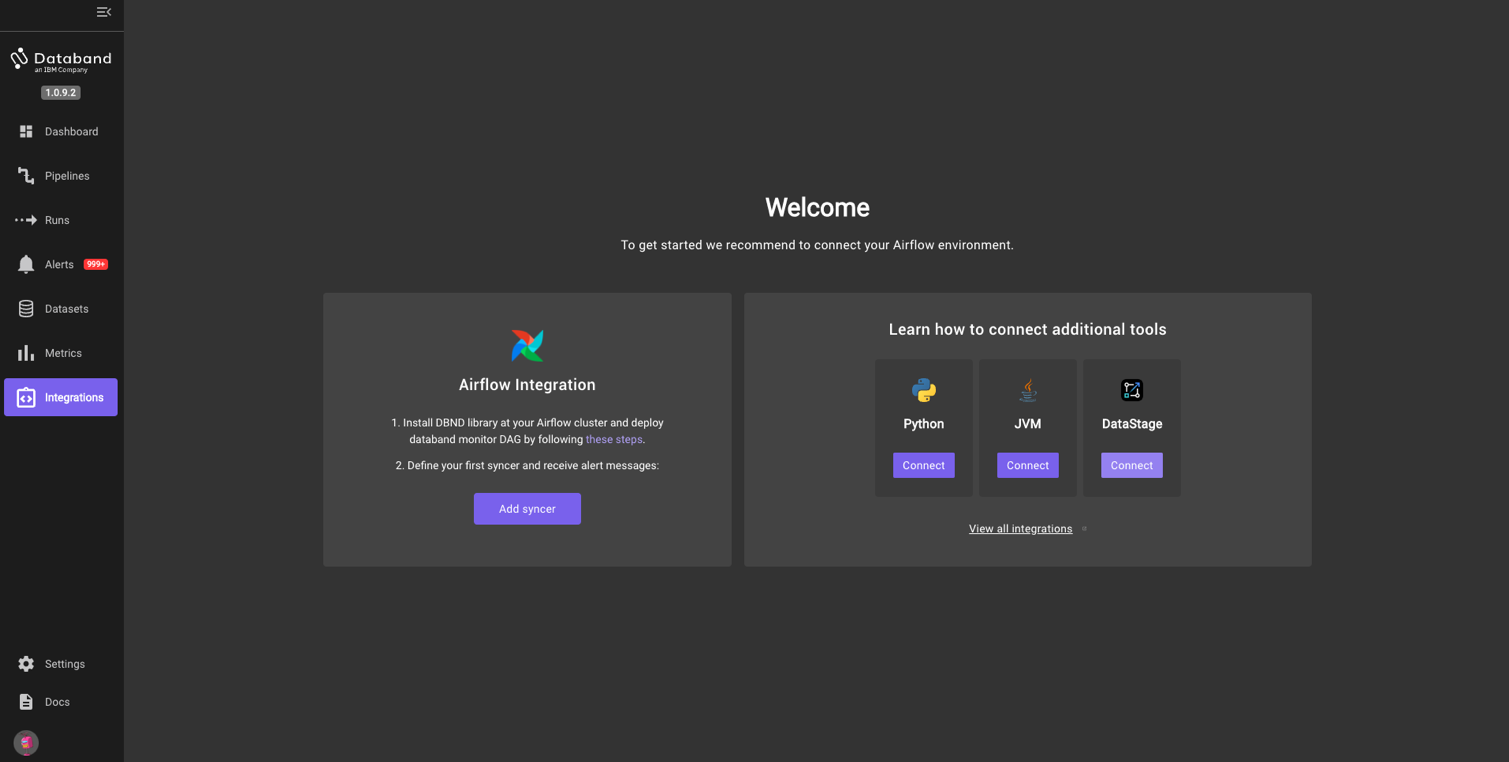




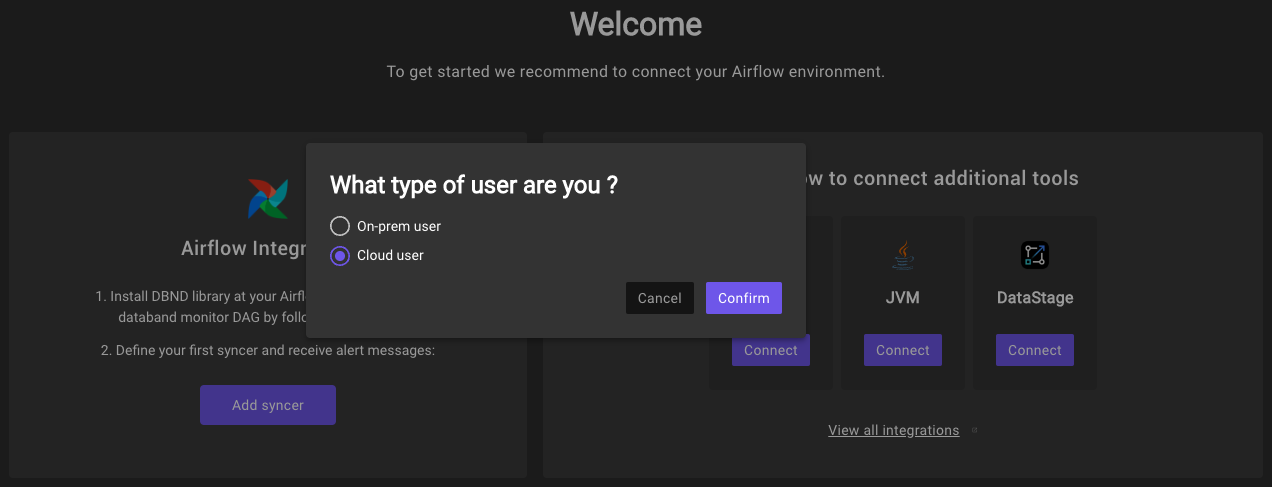
1. Open a new web browser tab and navigate to the Databand environment. Log in using the credentials you were given after signup.



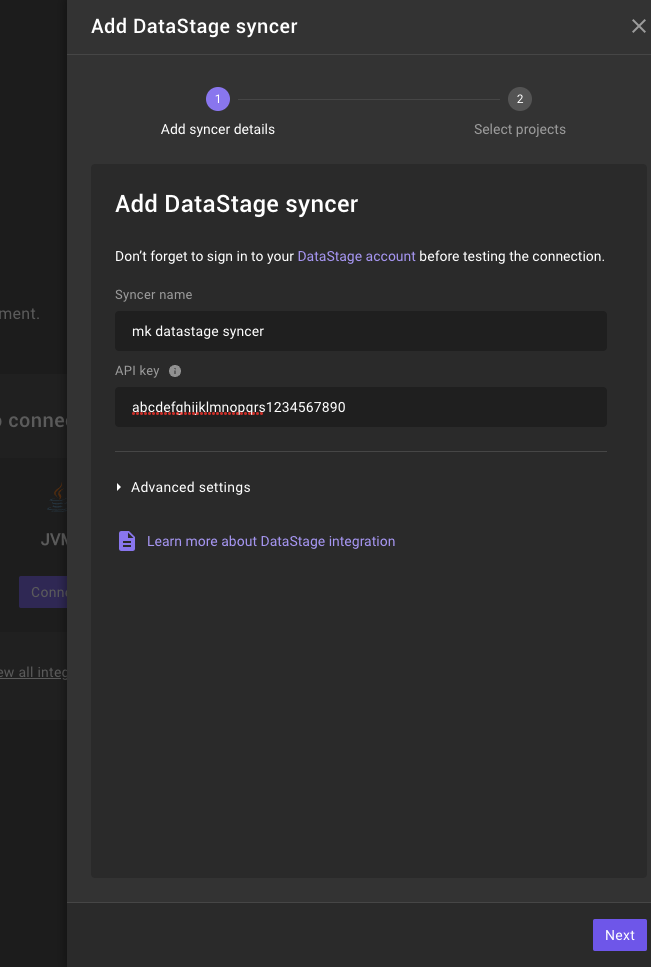
1. We will now create our DataStage Syncer within Databand. A syncer will "sync” or integrate your DataStage environment with your Databand environment. To do this, click the “Integrations” button on the lefthand menu. Then, click the purple “Connect” button under DataStage.



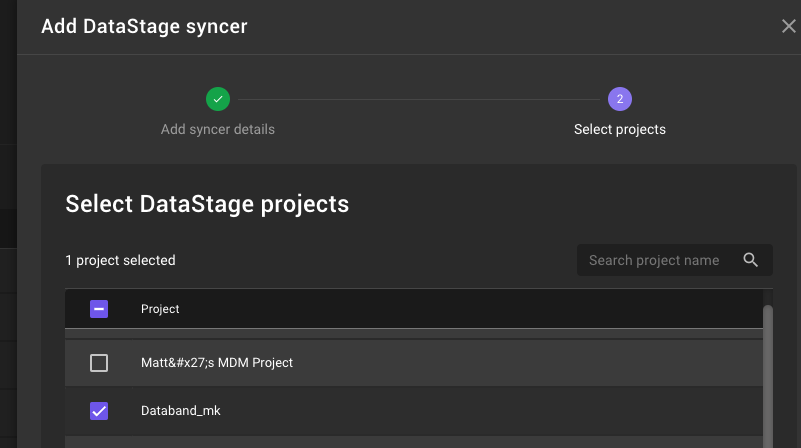
1. On the next prompt, select “Cloud user” and click “Confirm”.



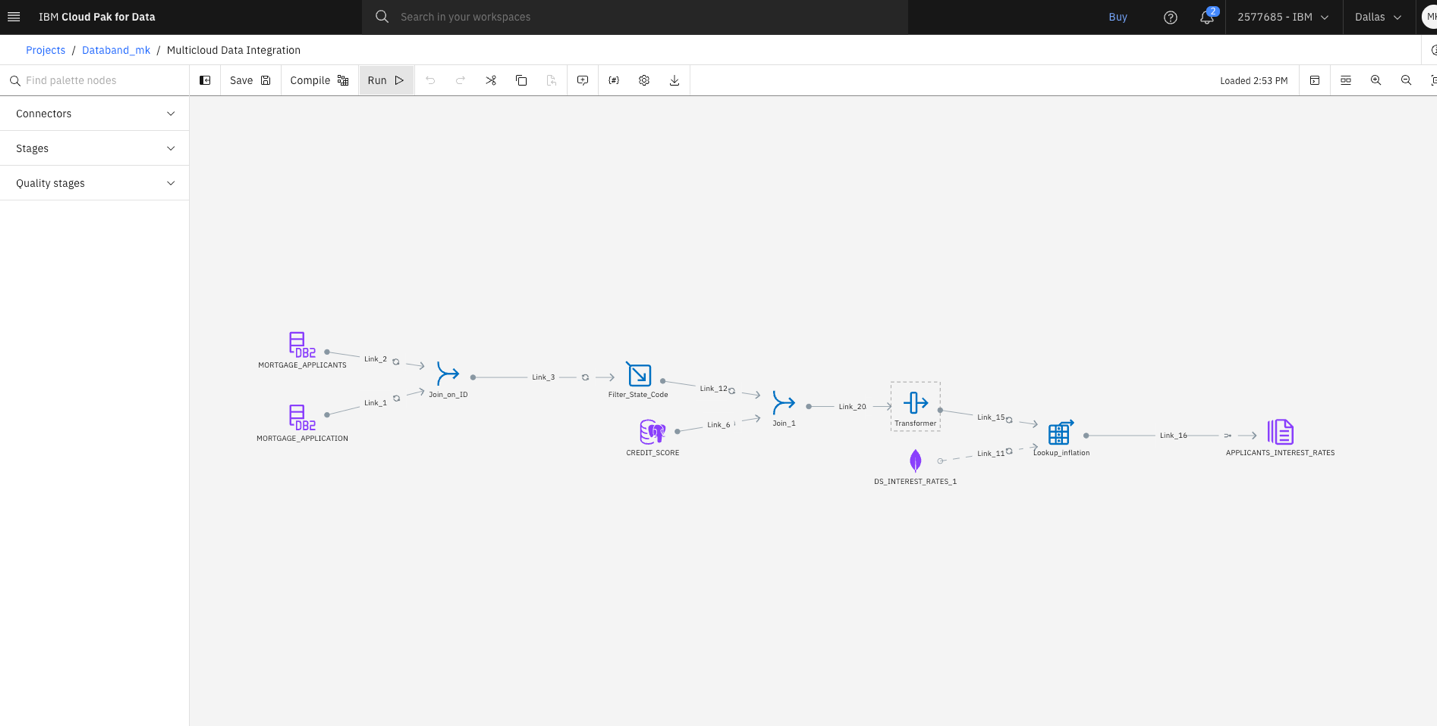
1. Create a unique syncer name (yourinitials datastage syncer) and paste your API key that you saved during step 9 in to the “API Key” field. Then click “Next”



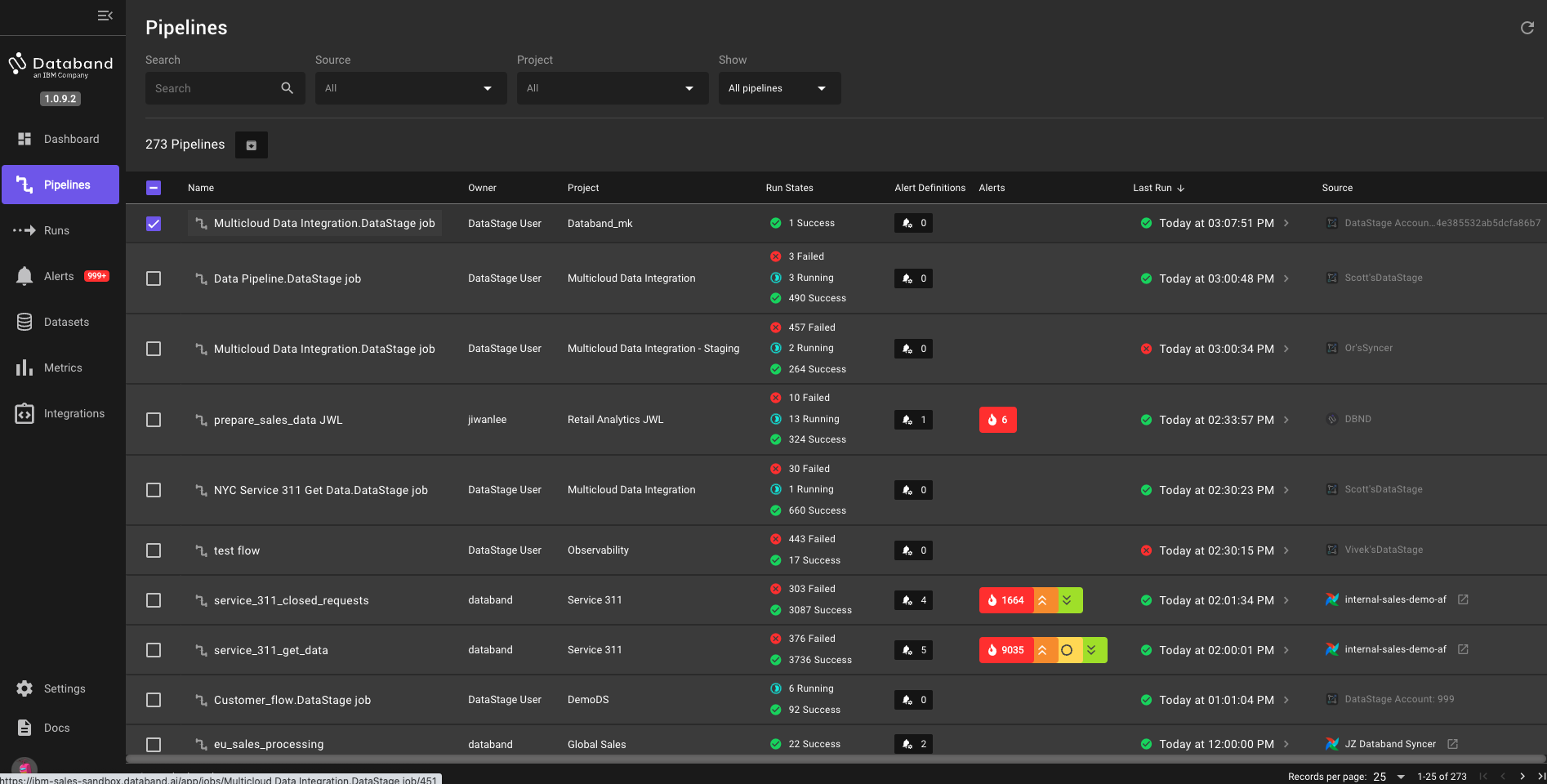
1. Select the “Databand\_yourinitials” project that you created at the beginning of this lab. Then click “Save”.



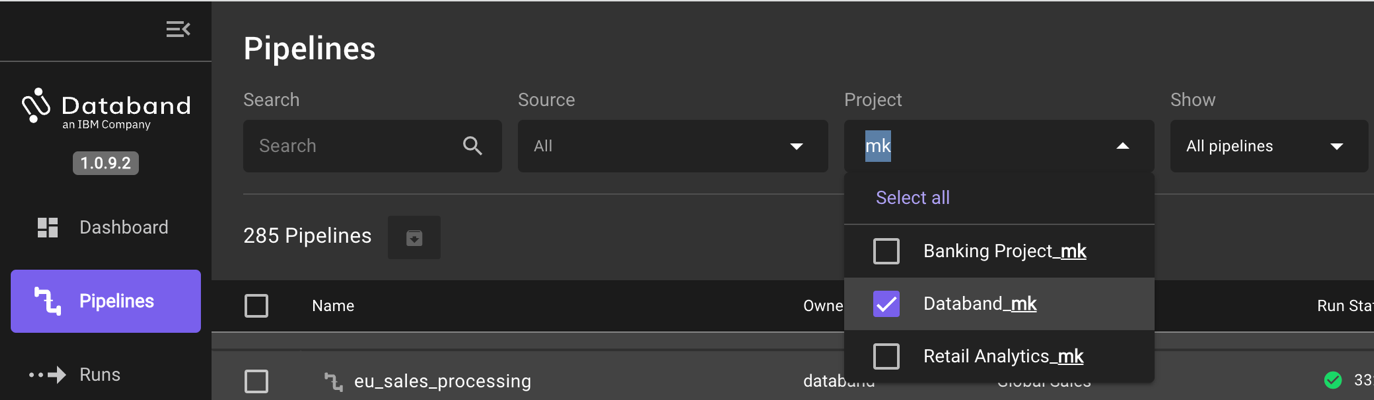
1. We have successfully synced our Cloud Pak for Data as a Service project with our Multicloud Data Integration flow, with Databand.
2. Next, return to the CP4DaaS tab, open the Multicloud Data Integration flow, and run click the “Run” button.



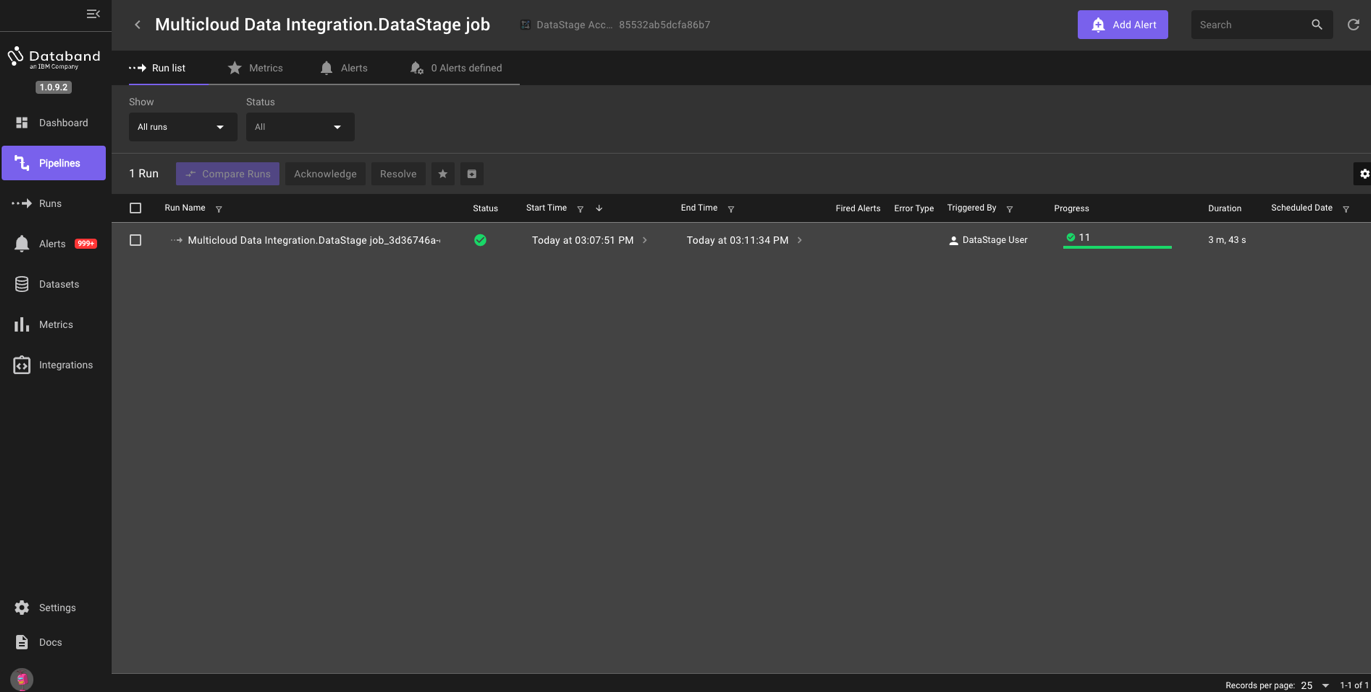
1. The job may take a few minutes to run. Upon completion, you will see a green “Run successful with warnings” banner. Once you see this, navigate back to the Databand environment. On the lefthand menu, navigate to the “Pipelines” tab. Within the first few pipelines on this page will be your specific ETL pipeline. You can identify your specific DataStage job / ETL pipeline by looking at the “Project” column of this page, which will display the project name of your DataStage environment. Click on the “Name” column of your specific pipeline (I.e. Databand\_mk)



QUICK TIP – You can search quickly for your specific DataStage job but clicking the “Project” dropdown and typing in your CP4DaaS project name in the search bar. See the below screenshot for an example of this search function.



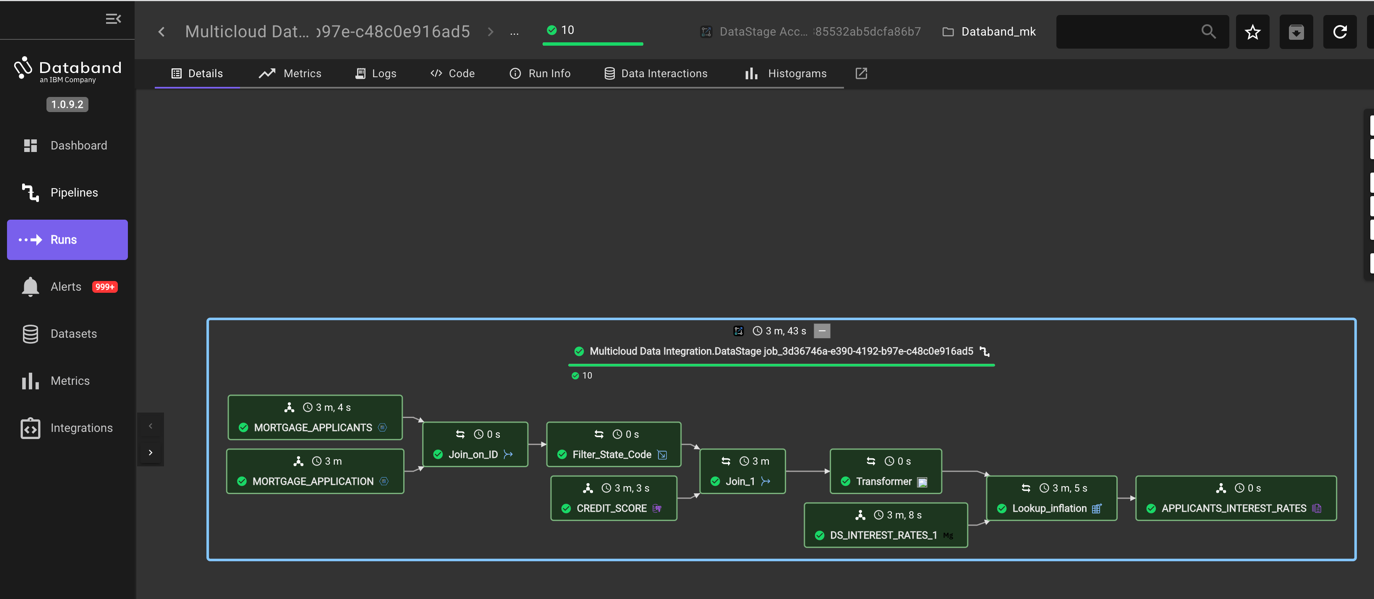
1. The new screen that pops up will be the “Run list” of each of the ETL pipeline (DataStage job) runs. This page displays the sequential list of runs for that specific DataStage job, the status of those runs, start and end time, alerts, errors, the number of successful/failed tasks, and the duration of those tasks. We have only run this once, so only one run will show. However, as we continue to run our job throughout this lab, we will see this page fill up with each sequential run.



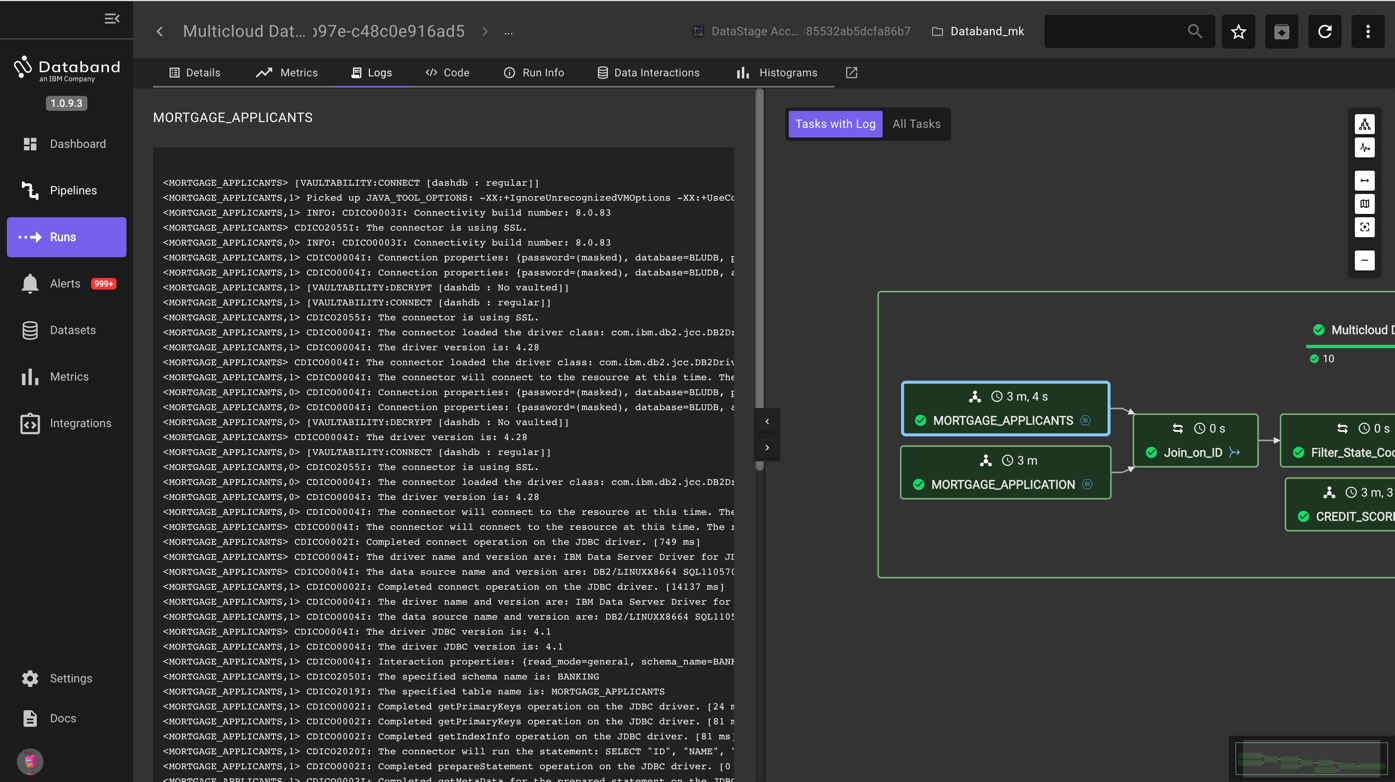
1. On this same screen, click on the “Run name” of this specific job run.



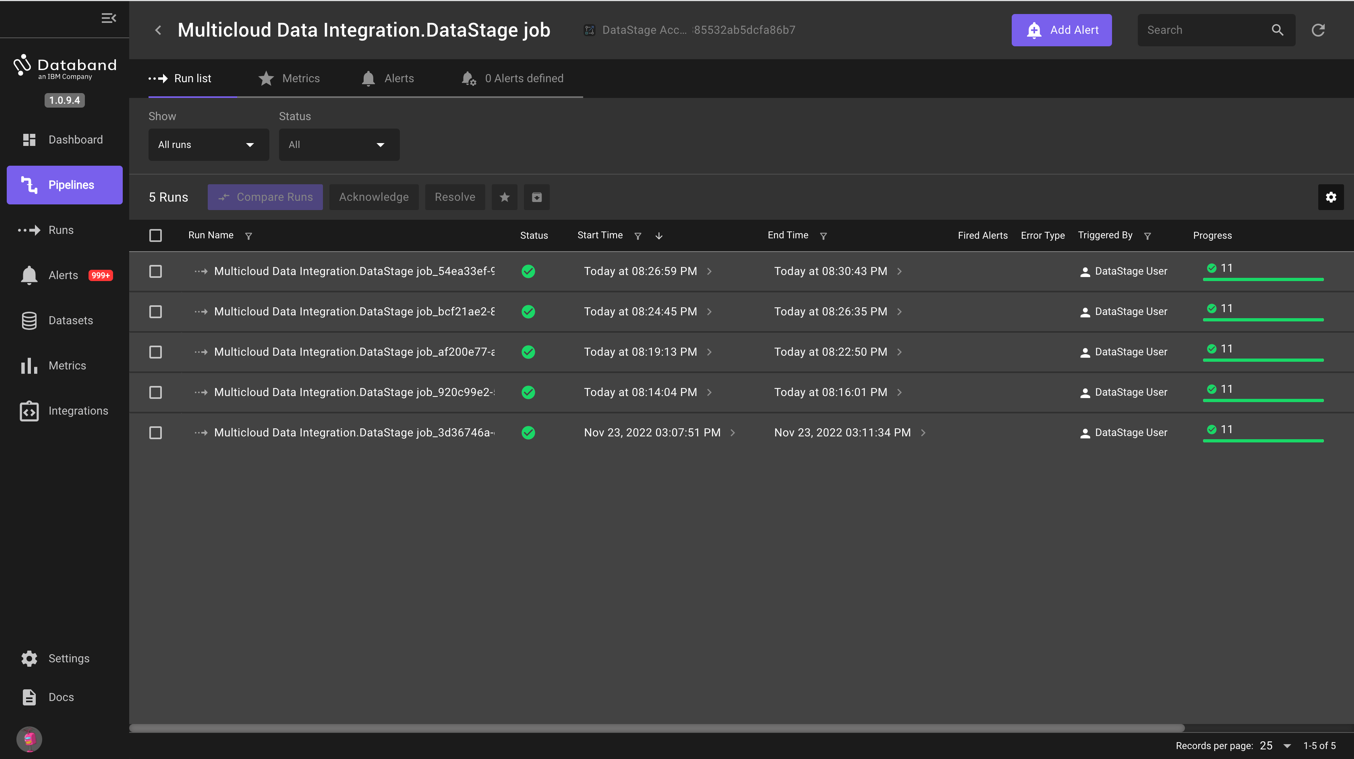
1. Click and hold the arrow pointing to the left in the middle of the screen, dragging the window as far left as possible so your screen looks like the screenshot below. You may have to drag and zoom the screen to center the job.



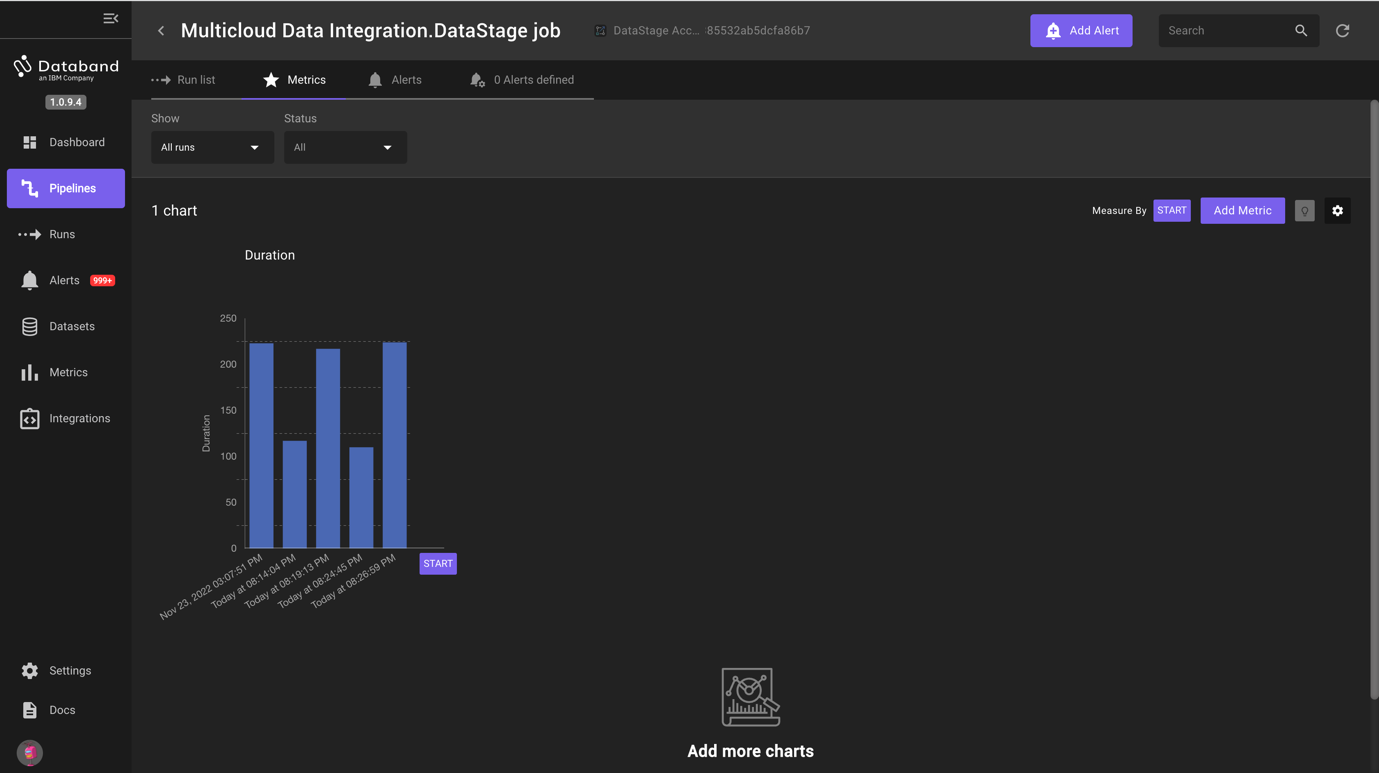
1. Here we can see the graphical representation of the DataStage job. We see that each stage is green, and thus ran successfully. Each individual stage contains the name of that stage, and a timestamp of how long it took each stage to run. Additionally, the top of this graphical representation shows the total time it took this job to run.
2. NOTE – these stages can run in parallel, and thus, the sum total of the time of each individual stage shown is likely much larger than the total time to run the job.
3. Next, click on the “MORTGAGE\_APPLICANTS” stage, and then click on the “Logs” tab on the top menu. Drag the arrow pointing to the RIGHT over to view the log for the selected stage. This view will show us the logs associated with that specific stage. Feel free to click on other stages to view those logs as well.



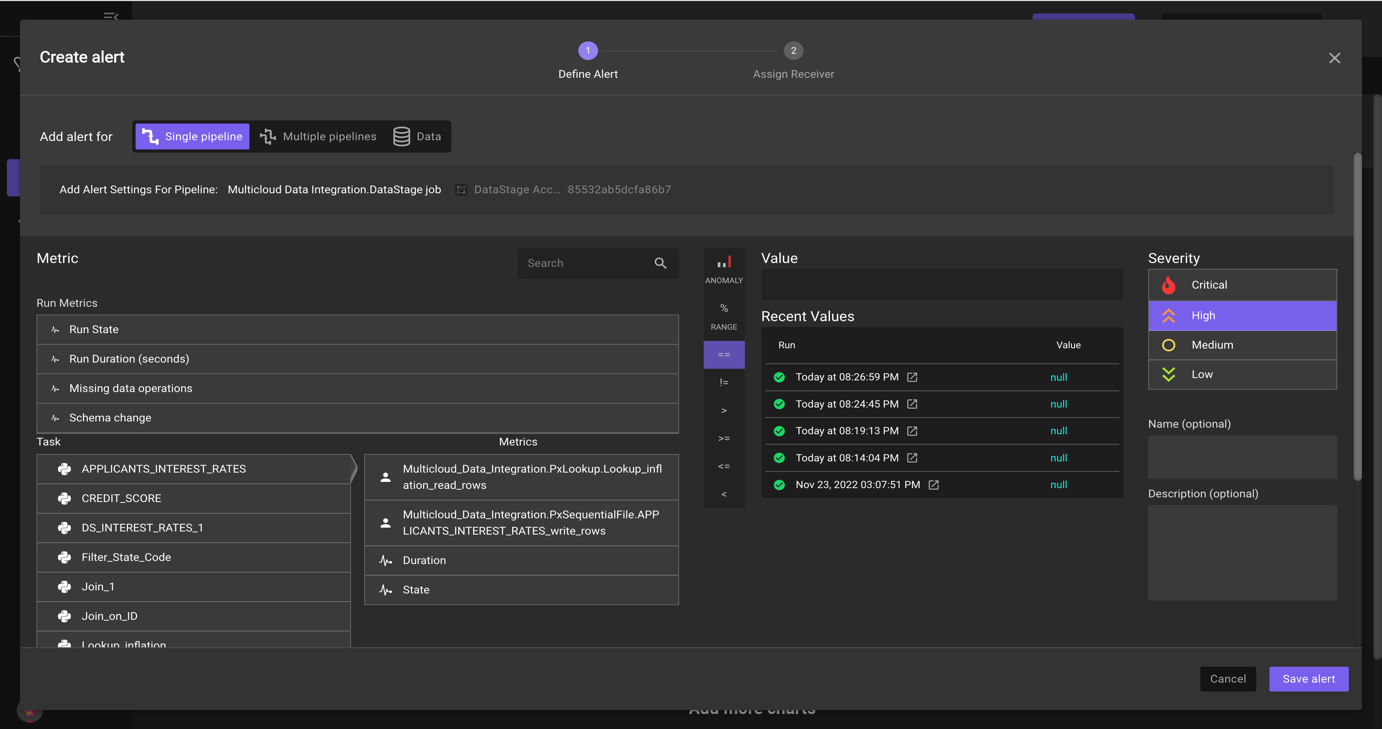
1. At this point, we have looked through some of the “step-through” functionality that Databand brings to observing our DataStage job. Switch back to your DataStage environment and run the job 4 more times to generate more metrics and set some baselines for our Databand environment. This will make sense as we continue our lab and will simulate what a customer environment would look like where a job would run many times.
2. After running this Multicloud Data Integration job four more times (for a total of five times), go back to your Databand environment. Go to the pipelines tab on the lefthand menu, find your pipeline, and drill into the pipeline, which will show the “Run list” tab on the top menu.



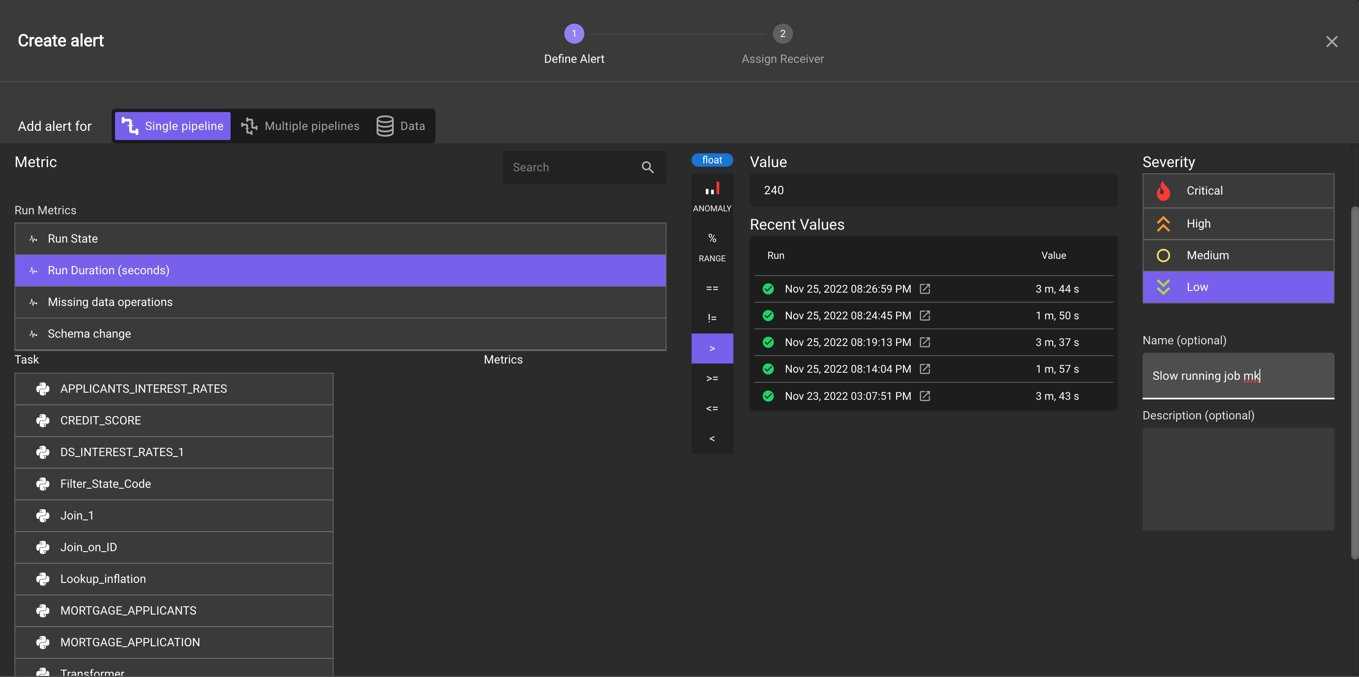
1. Click the “Metrics” tab on the top menu. The default metric shown is the “Duration” of each run. You can observe metrics on other areas of the DataStage job around rows being written to and read from each stage. This is outside the scope of this lab.



1. We will now create our first alert. To begin, click the purple “Add Alert” button near the top right corner of your screen.



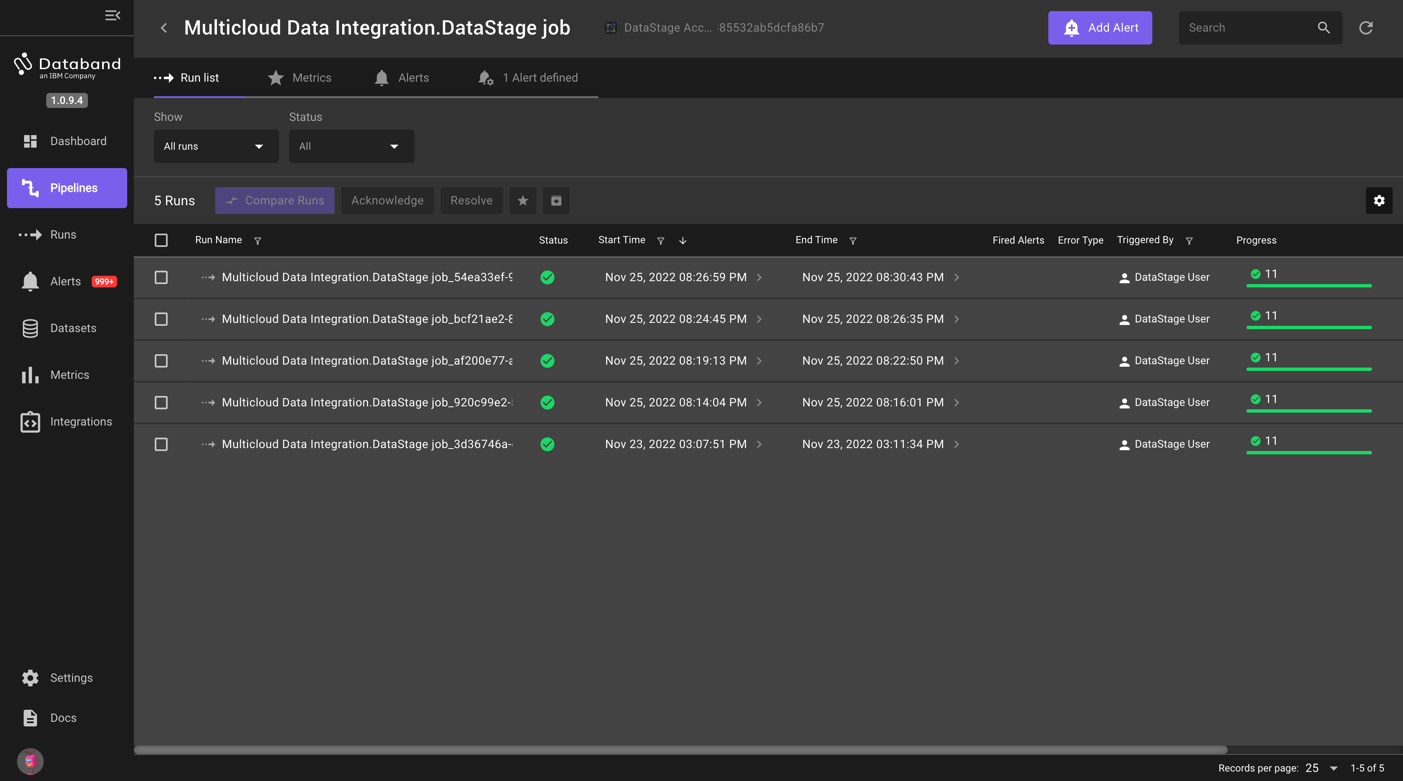
1. The first step in creating our alert is to create the “Alert definition”. This is the logic behind your alert. Look at all of the alert possibilities we can create within Databand. We can create an alert based on Run Metrics of our DataStage job such as successful or failure, run duration, missing data operations, and schema changes. We can even alert on metrics of each individual stage within our job (shown under the “Tasks” designation).
2. Since our jobs take around 2 or 3 minutes to run, we will create an alert if our job takes greater than 4 minutes. To create this alert, click the “Run Duration (seconds)”.
3. Next, click the “>” symbol to designate the greater than logic. Note the other operators listed here, including “Percentage Range” and “Anomaly”.
4. Enter “240” in the value box to designate the time of 4 minutes.
5. Click on the “Low” box listed under severity. This will alert the assigned individual group on how important this specific alert is. Since this is due to a job running slightly slower than normal, we will mark this as “Low severity”.
6. Lastly, give this alert a name titled “Slow running job *yourinitials*” (I.e. Slow running job mk).
7. Before clicking save, make sure the logic of your alert definition matches the screenshot below.



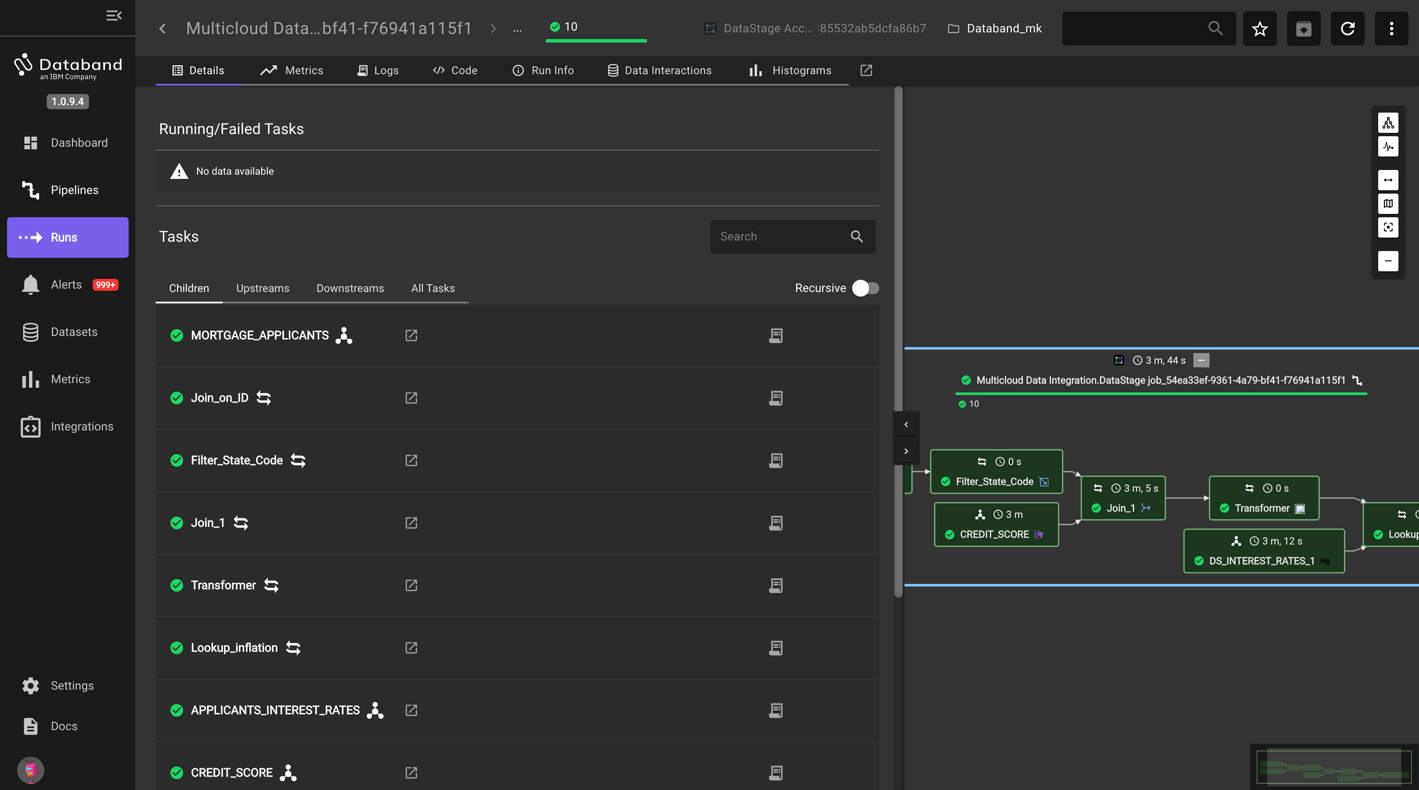
1. Click the save button to create this alert. The next screen allows you to assign this alert to a receiver, which is a user or group of users that will be notified of this alert through Slack, email, or PagerDuty. We will look more at this in the next portion of the lab. We will keep this alert within Databand. Scroll down and click the “Done” button.
2. We can see here how helpful such alerting can be for monitoring the success, failure, and overall performance of our DataStage jobs.

For the final portion of this lab, we will view the data interactions of our job and create an alert based on those interactions. **This is one of the key value-adds of integrating Databand with your DataStage environment**, as we can now alert users in near-real-time on many custom failures, job changes, delays, and much more.

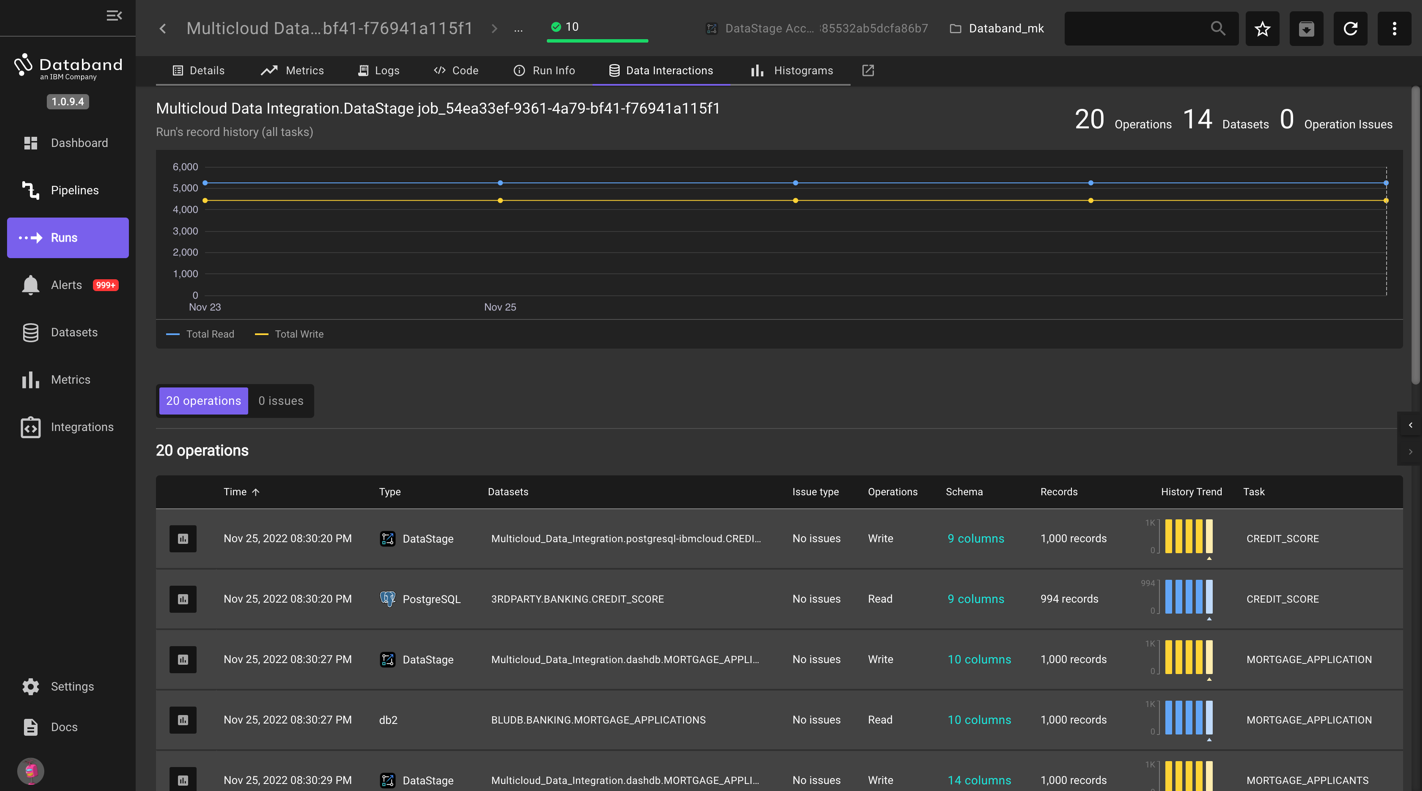
1. Navigate to the run list page, which is shown in the screenshot below.



1. Click into the top “Run Name”. You should see a screen like the one below. You may have to move the screen using the left or right arrows.

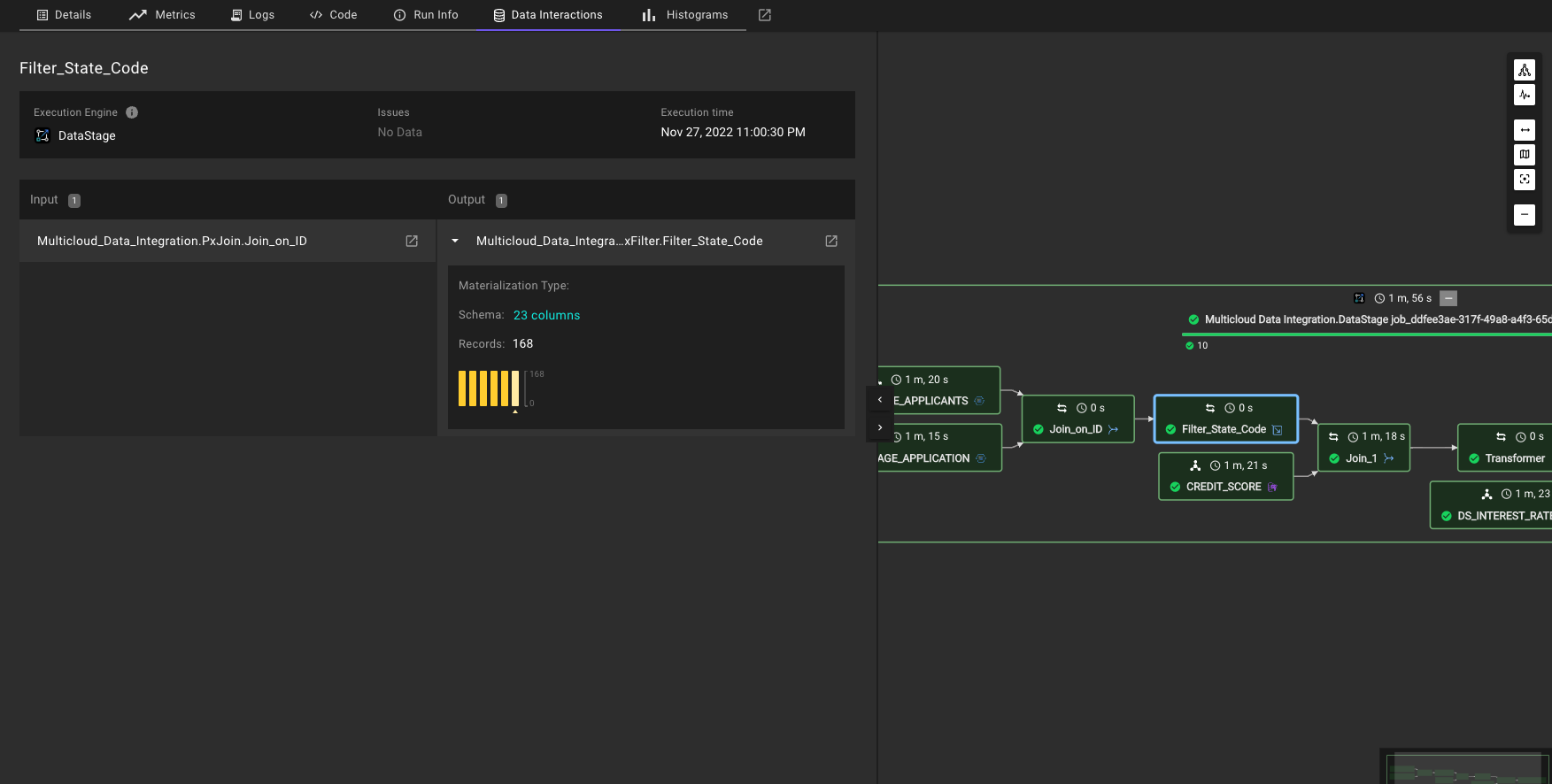


1. Click on the “Data Interactions” tab in the top menu.

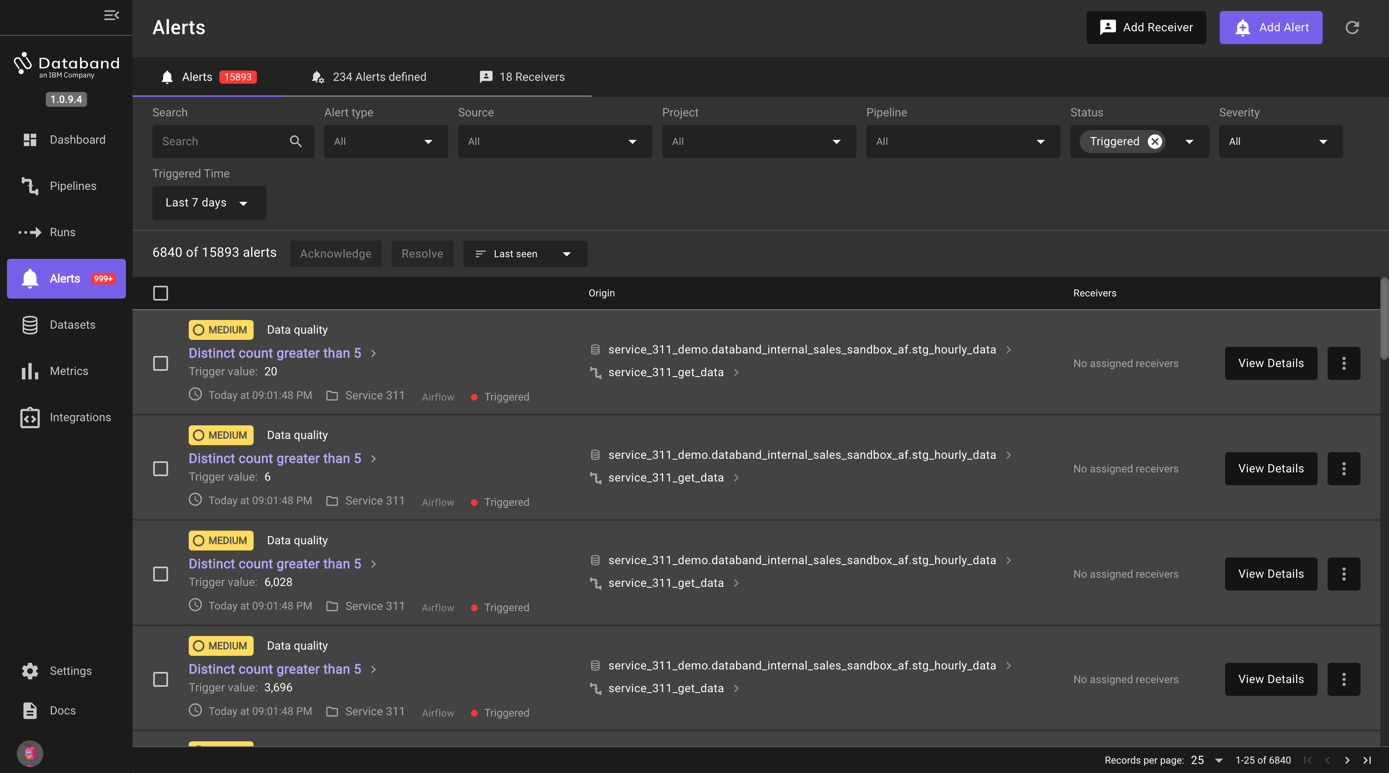


1. Here we can see the inputs and outputs (reads/writes) of the records and of the columns in each respective stage. We can see the source type, associated datasets, any issues that may have come up, information on the schema and records, and the associated stage. The sum total of these records is represented in the chart at the top of the screen. The chart at the top titled “Runs record history (all tasks)” as well as the column titled “History Trend” gives the user a view of the job’s historical performance.

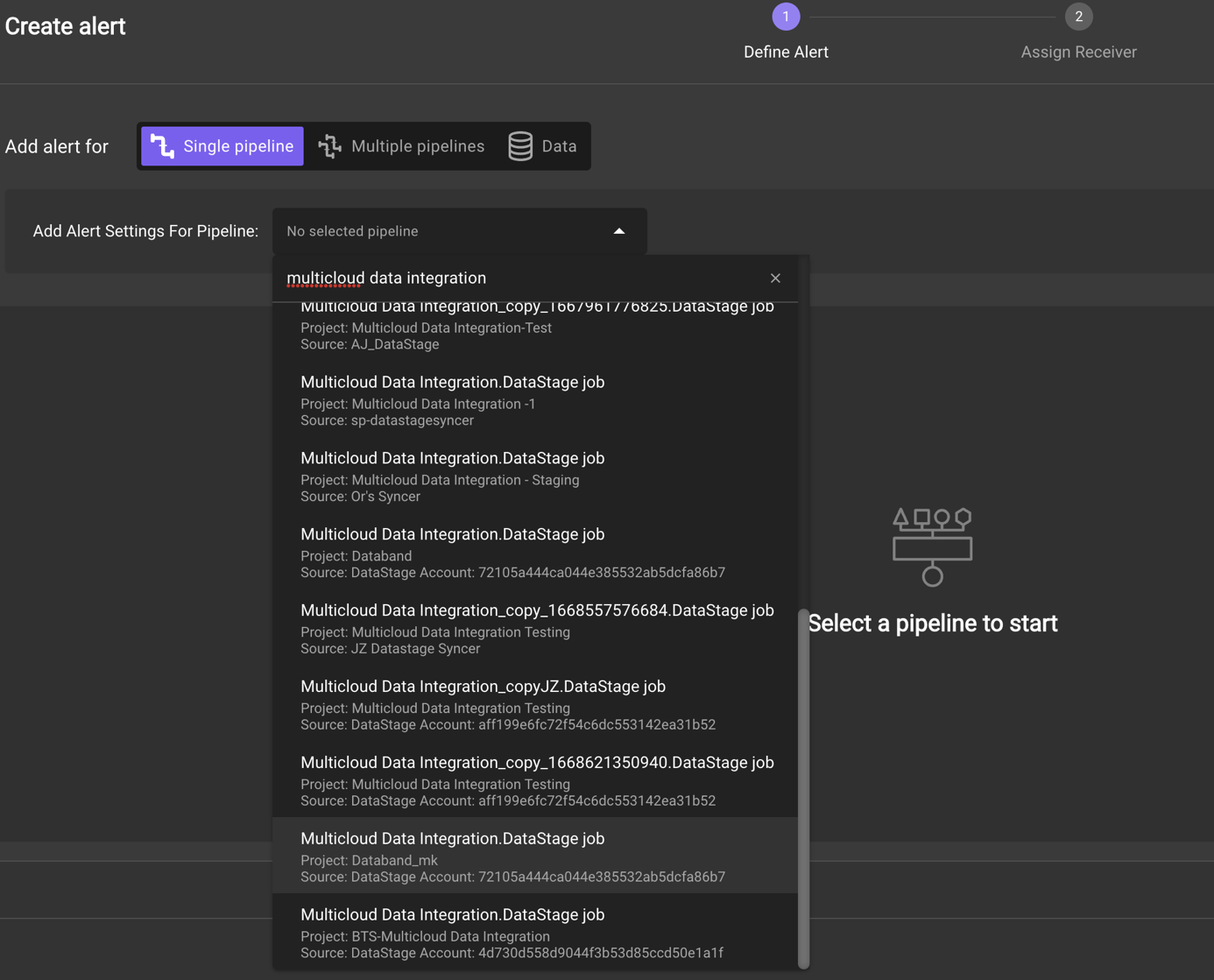
NOTE – you can also view the “Data Interactions” tab for each individual stage, such as in the screenshot below. By clicking on the specific transformation, you can see information such as the schema, the row count, and the historical trend of that row count.

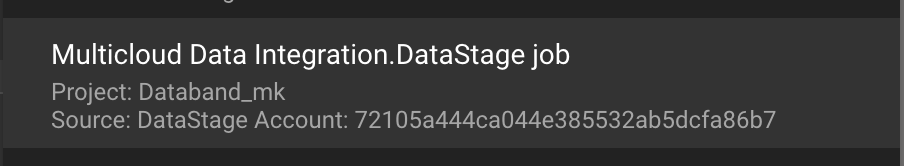


1. We will now create an alert around a schema change. Click the “Alerts” tab on the left-hand menu. This is where all Databand alerts are shown. If you wish, look around this page to understand what information is shown to the user.

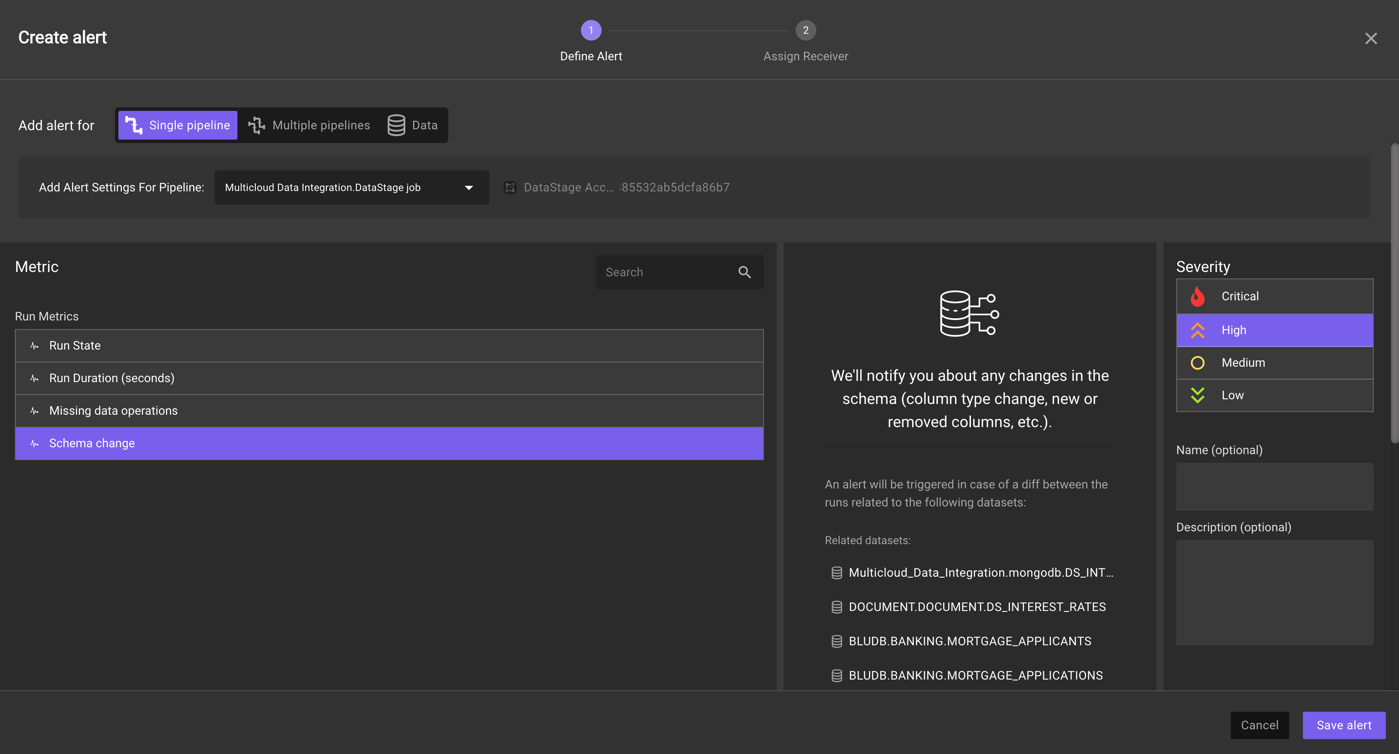


1. Click the purple “Add Alert” button in the top right corner of your screen. Here we can see that we can create an alert on one pipeline (DataStage job), multiple pipelines, or on data quality. We are going to create an alert to monitor for a schema change and set the receiver to be Slack.
2. First, we will specify our pipeline and create our alert definition. Click on the “no selected pipeline” dropdown, type in “Multicloud data integration” and search the listed pipelines for the one that shows your specific project. In the second screenshot below, you can see an example of where the “Project” is shown.

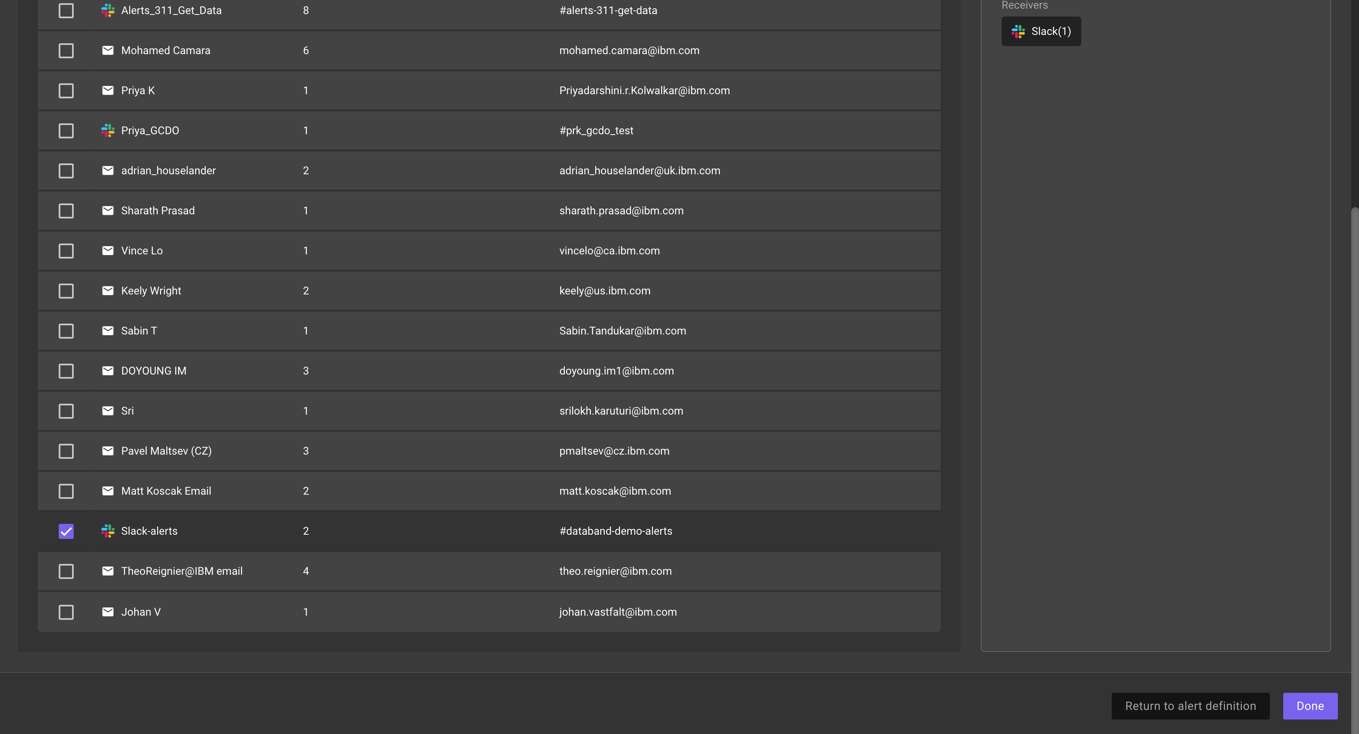




1. After selecting your specific job, click on the “Schema change” run metric and select “High” as the severity. Make sure your screen matches the one shown below, then save the alert.

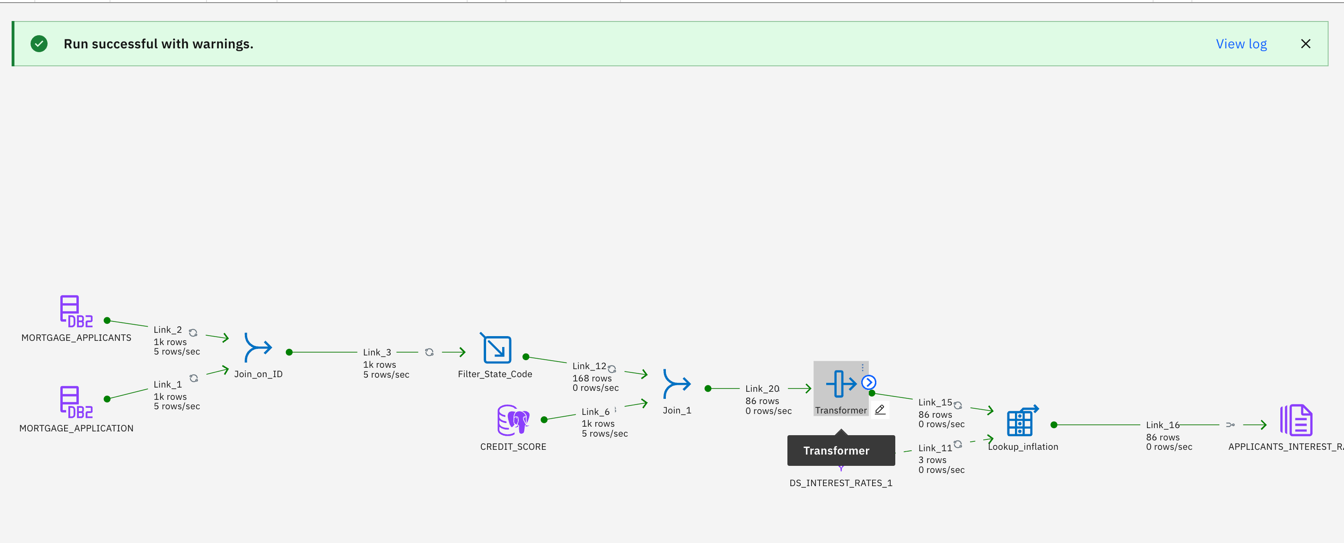


1. On the receivers page, select the “Slack-alerts” that is associated with the “#databand-demo-alerts" slack channel. Then click “Done”.

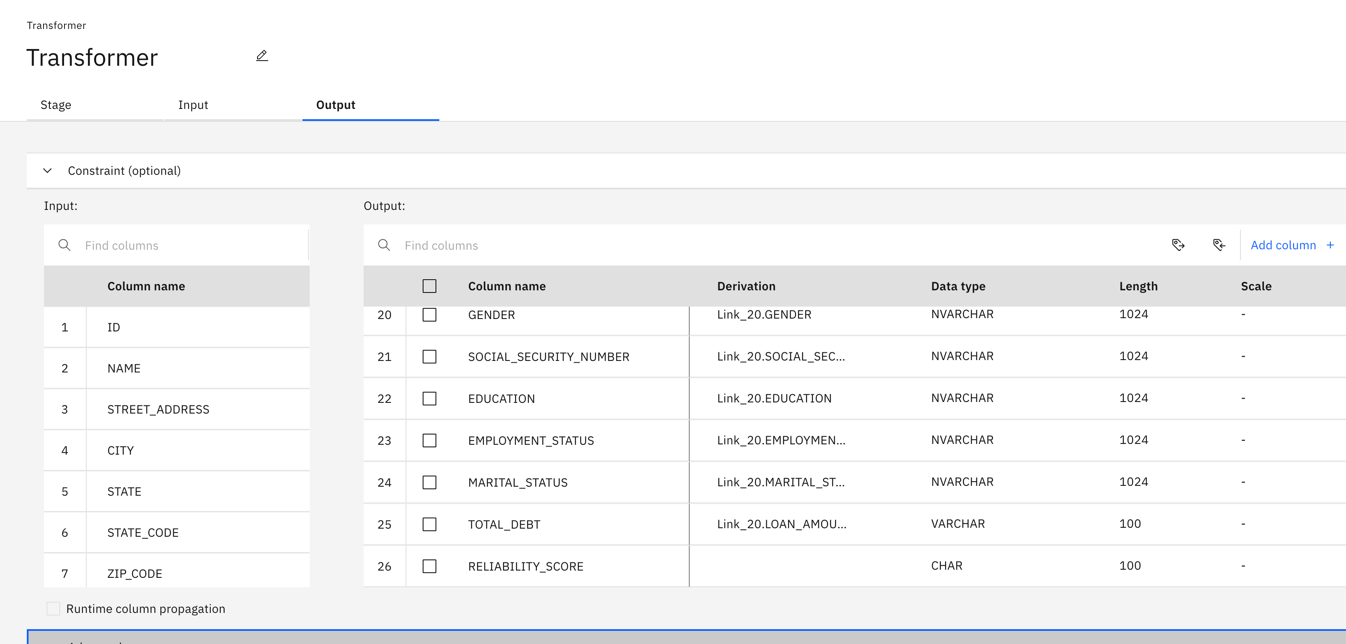


*NOTE – if you have not already joined the #databand-demo-alerts slack channel, do this now. This is located in the Data&AI workspace.*

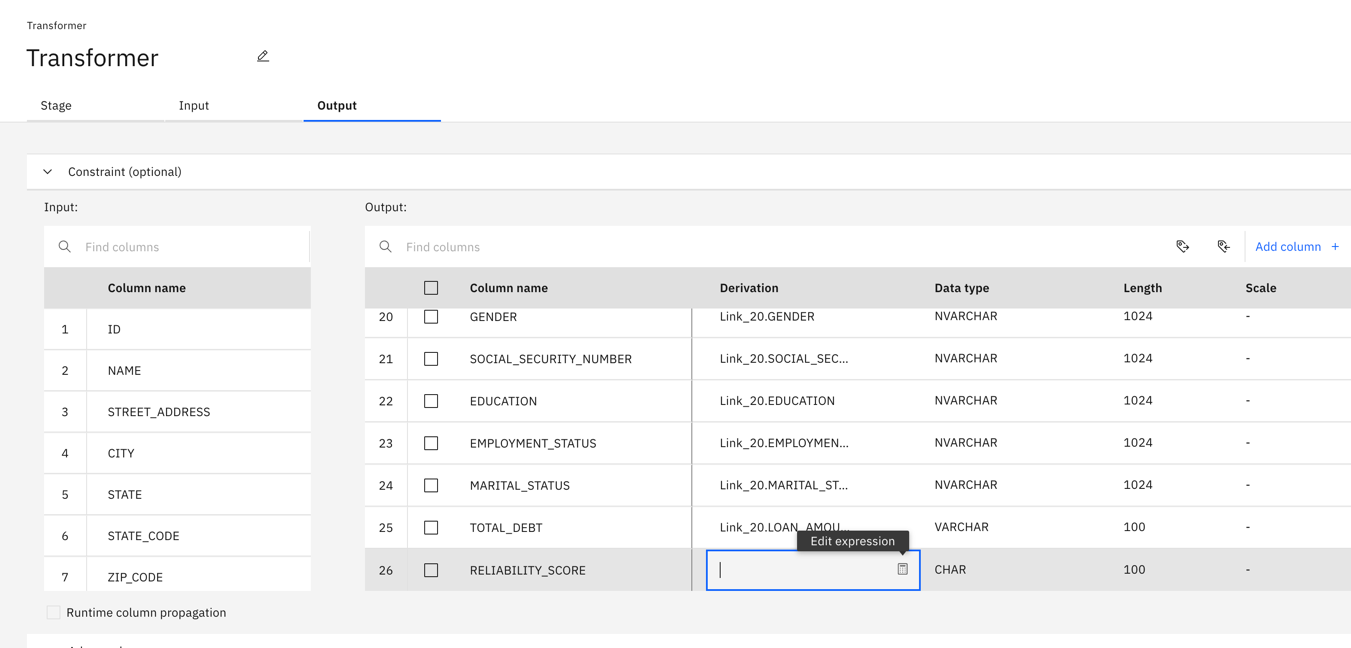
1. Switch back to your DataStage environment, where we will purposefully introduce a schema change. Click on the “Transformer” stage to open the settings.



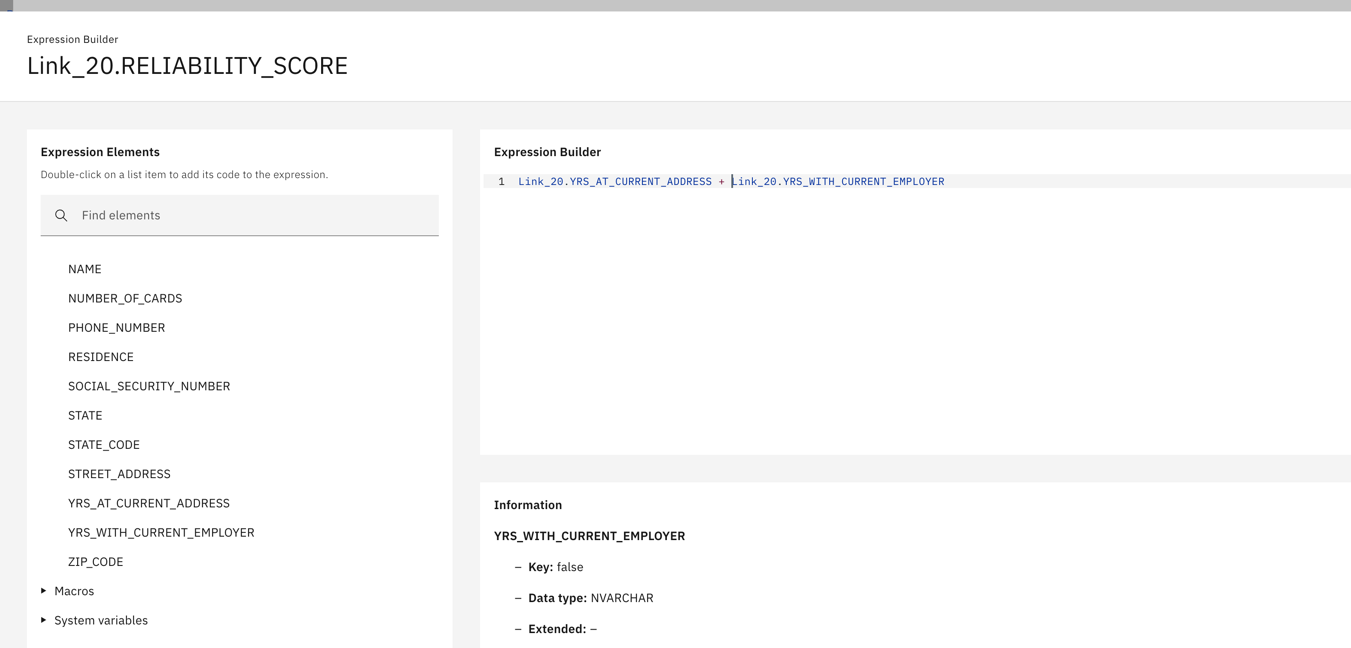
1. Go to the output tab, and select the “Add column” button. Name your new column “RELIABILITY\_SCORE”



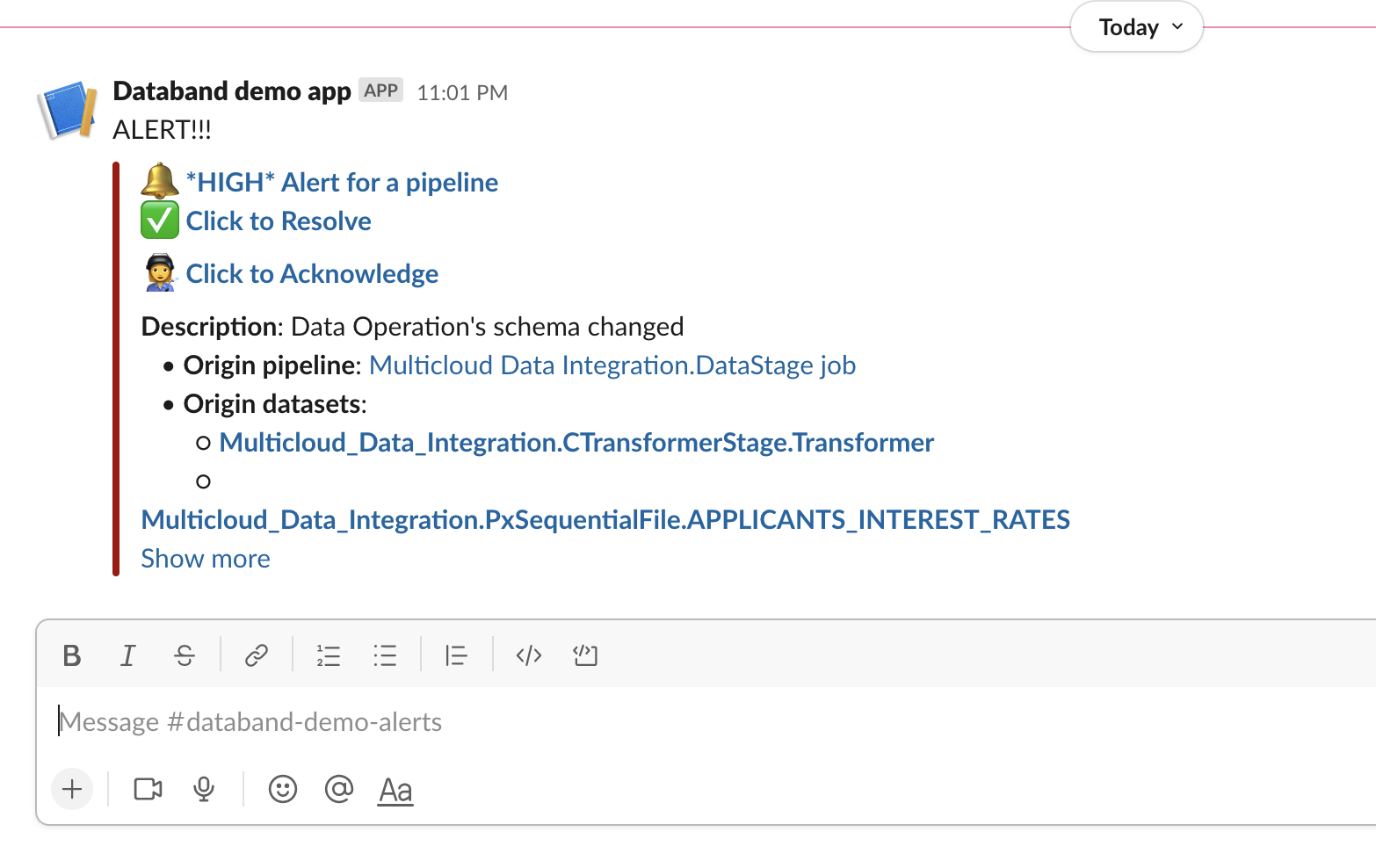
1. Next, click the edit symbol in the derivation column, then click the calculator symbol to edit the expression.



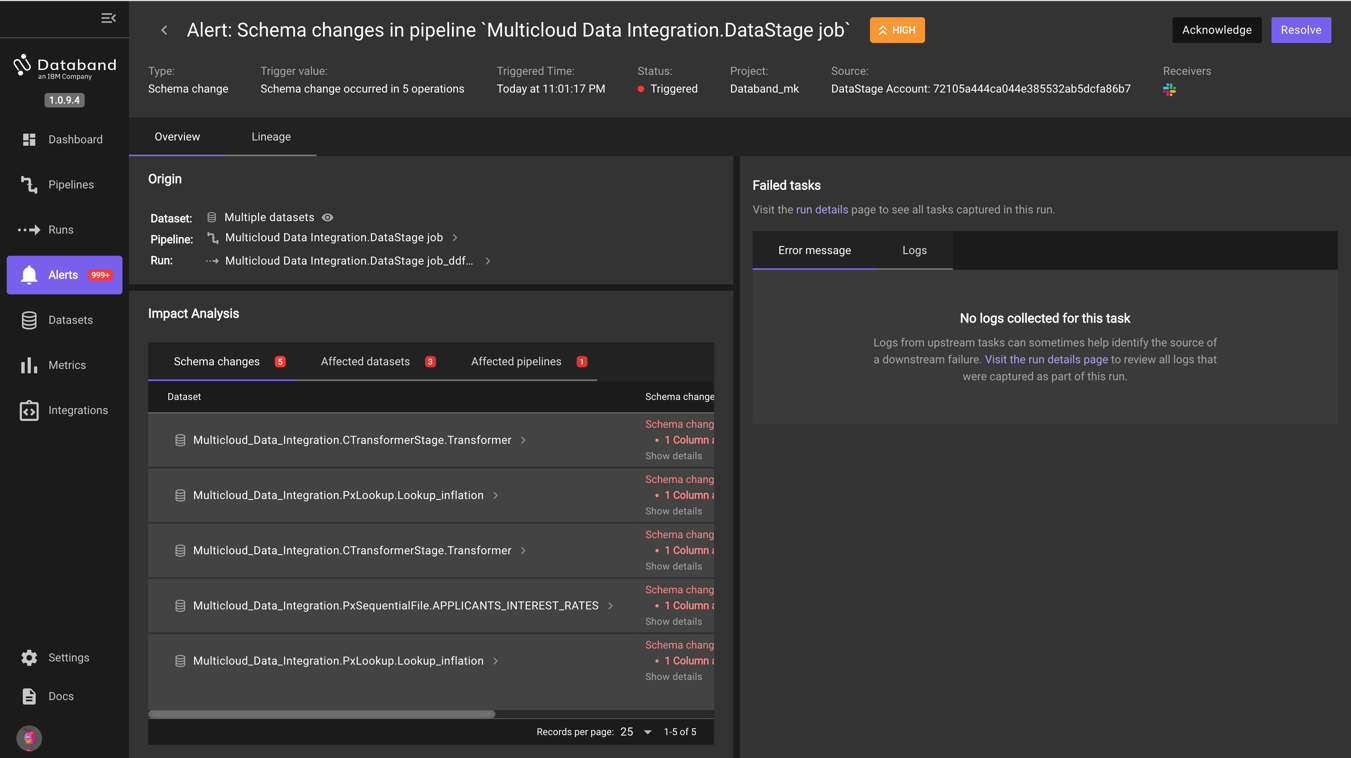
1. For the expression, we will use the sum of the “YRS\_AT\_CURRENT\_ADDRESS” column and the “YRS\_WITH\_CURRENT\_EMPLOYER” column. To create this expression, locate the two columns under the “Input columns” drop down. Double click “YRS\_AT\_CURRENT\_ADDRESS”, insert a plus symbol (+) and then double click “YRS\_WITH\_CURRENT\_EMPLOYER”. Your expression should look like the screenshot below.



1. Click “Apply and return” and then “Save and return” to save the changes you just made. Run the job and wait for a few minutes until the job completes and the alert is triggered. When the alert is triggered, you will see a message in the #databand-demo-alerts slack channel similar to the following screenshot.



1. Click on the “\*HIGH\* Alert for a pipeline” in the slack message. This is the top message next to the bell icon. This opens up the following screen.



1. Here we can see the impact analysis of this alert. Specifically, we see what schema changes happened in our job, what datasets were affected, and the pipelines (DataStage Jobs) that were affected. We can see this information graphically as well in the Lineage tab.

**This concludes the DataStage + Databand Hands on Lab. Please reach out to Matt Koscak (**[**matt.koscak@ibm.com**](mailto:matt.koscak@ibm.com)**, or on slack) with any questions, comments, concerns, or customer situations you run into.**