




Databricks ISV Summit Workshop steps

1. To login navigate to <https://isv-tech-summit-apac.cloud.databricks.com> and enter your email address and password. If you don't have your password or don't remember your password, enter your email address and click Forgot Password? Link. This will send you an email with which you'll be able to reset your password. Once you login go to step 2.



Sign In to Databricks

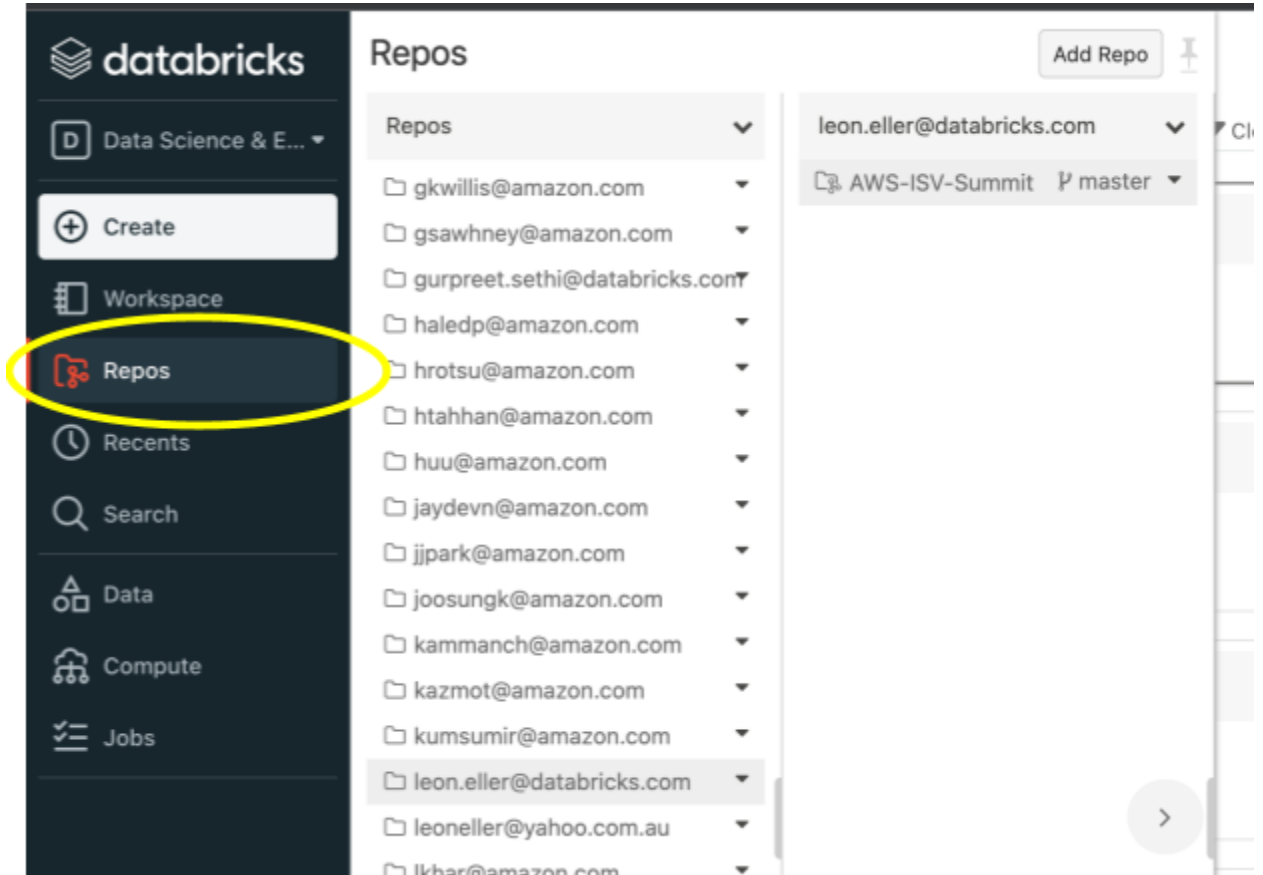




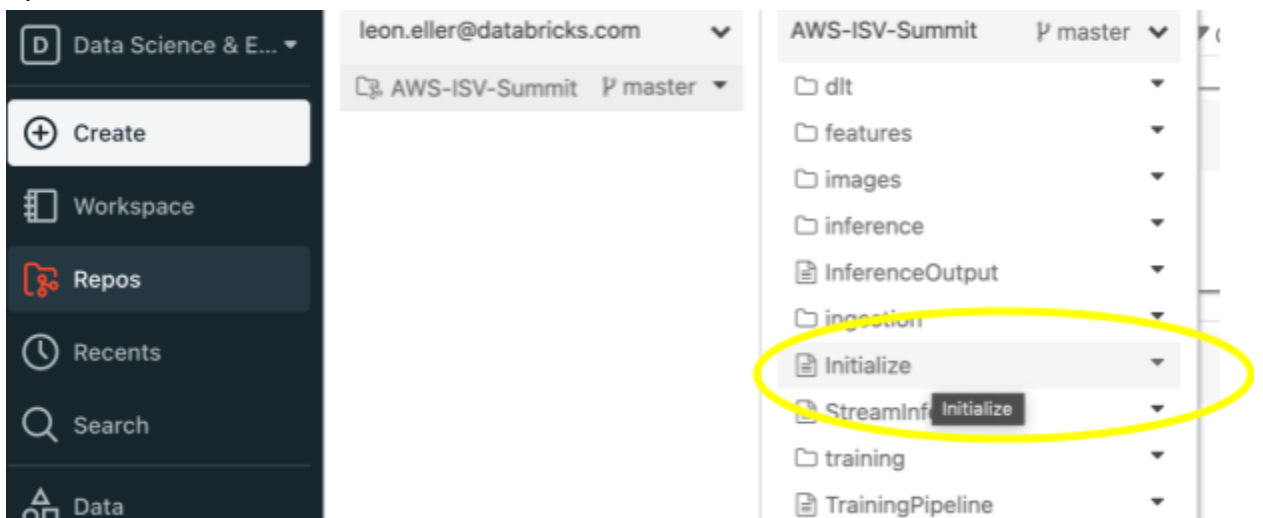
[Forgot Password?](#)

[Privacy Policy](#) | [Terms of Use](#)

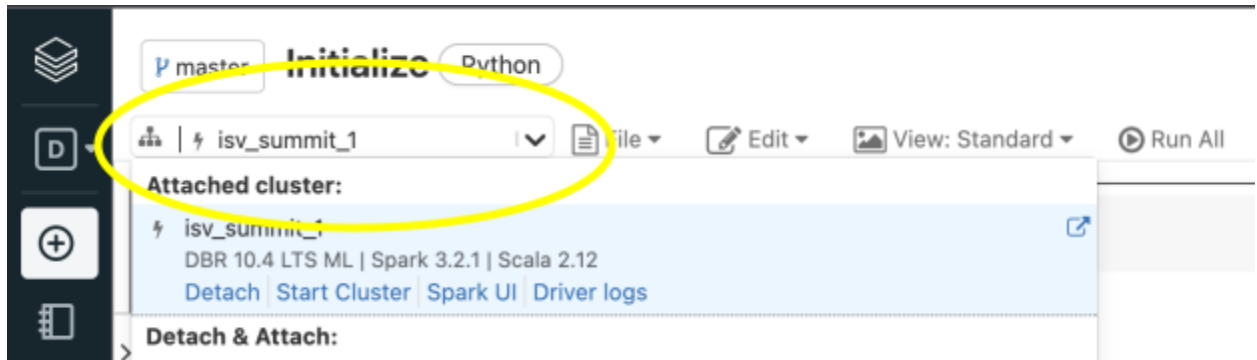
- Click on Repos tile and navigate to your Repos folder



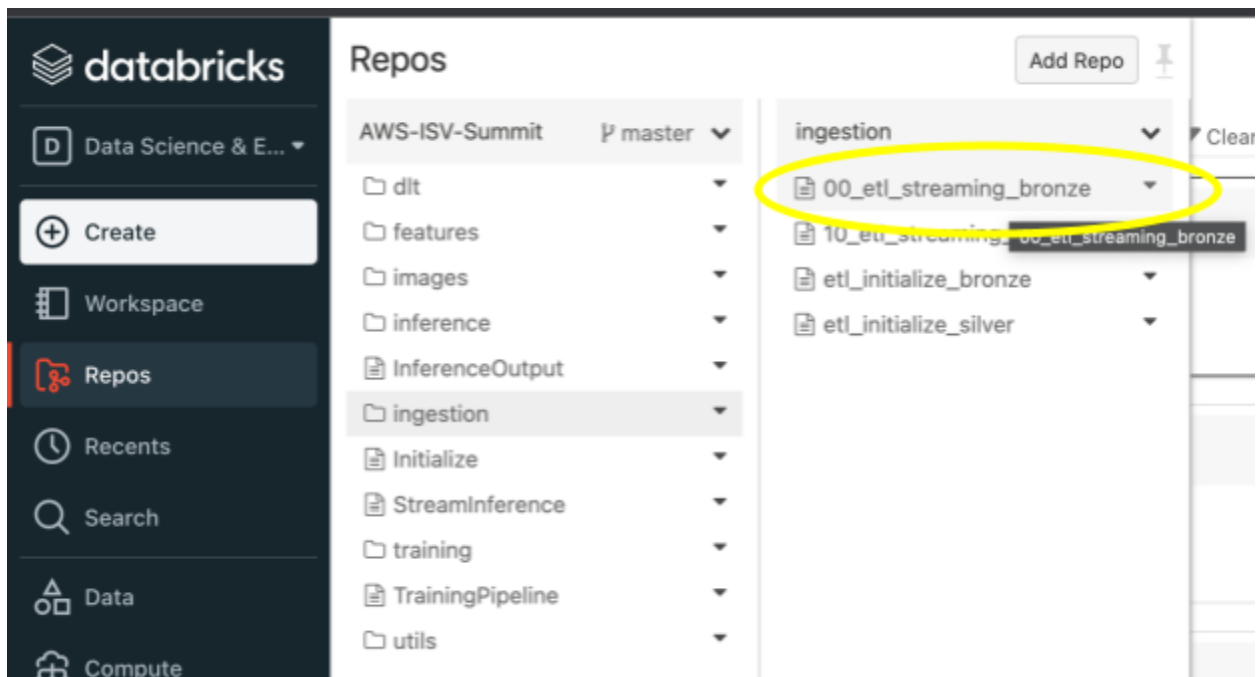
- Open a Notebook called Initialize



4. In the top left corner select a Cluster to attach your Notebook to. You should only have access to 1 Cluster called `isv_summit_1` or `isv_summit_2` etc.

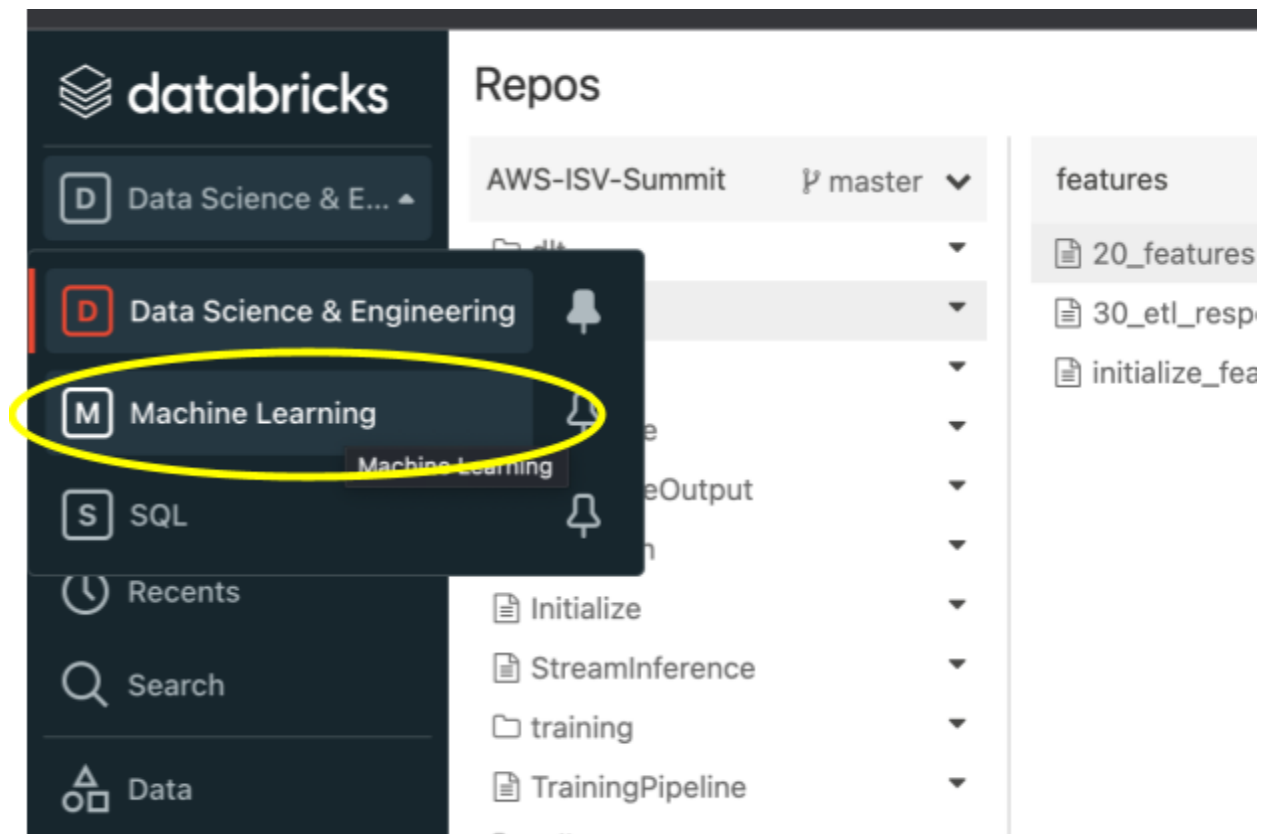


5. Also in the drop down click on Start Cluster link, if the cluster is not already running
6. Click Run All icon in top bar next to View: Standard icon. This will execute the Notebook clear and reinitialize the tables where you'll be storing the data. Wait for the Notebook to finish executing (When executing Run All button changes to Stop Execution and when finished changes back to Run All). After this step you are ready to run through the steps to train the ML model.
7. Open a Notebook in ingestion folder called `00_etl_streaming_bronze`

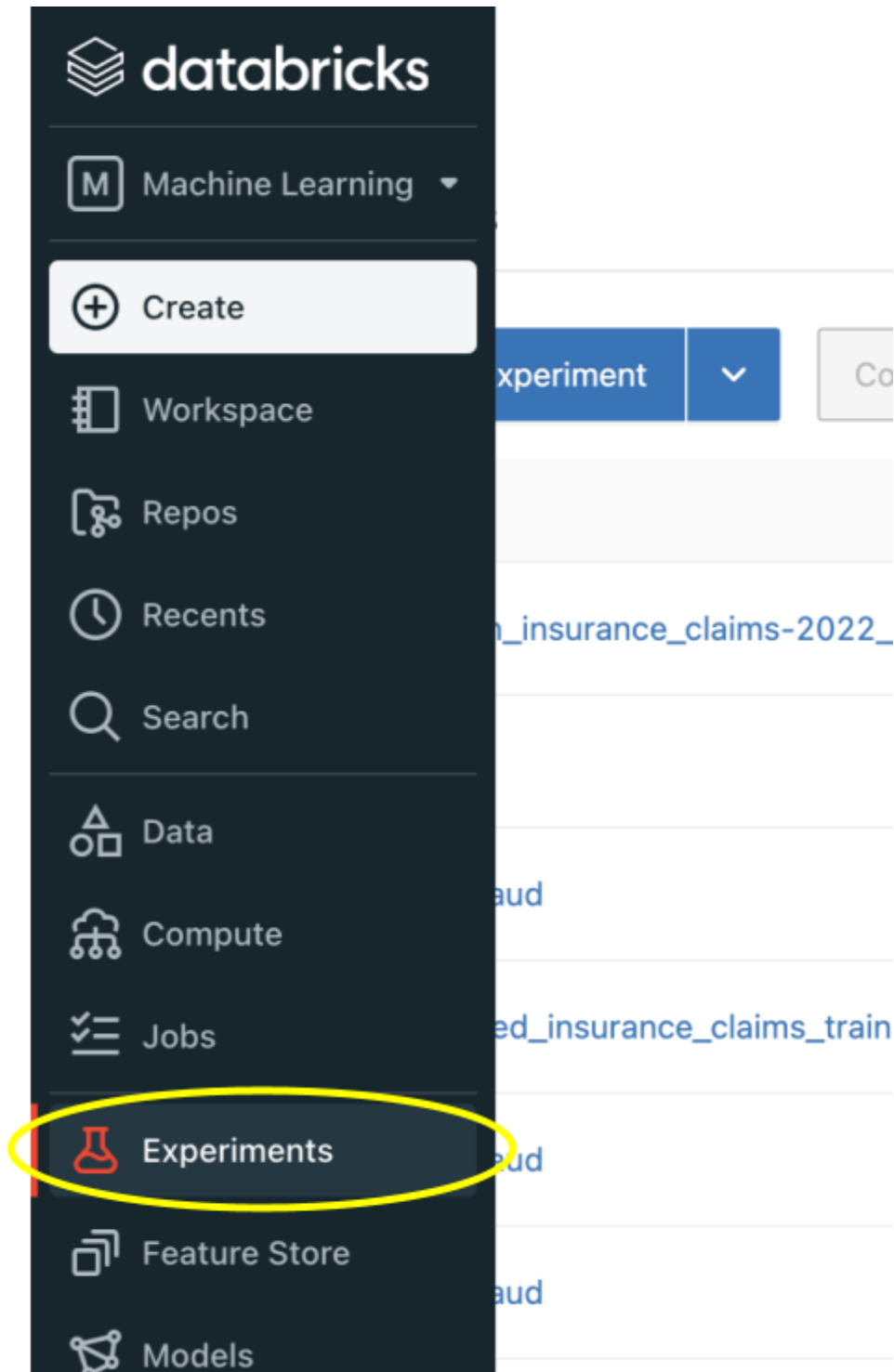


8. If a Cluster is not already attached to the Notebook attach a Cluster following the same steps as Step 3 above. Now click Run All to run the Notebook. This will ingest Insurance Claims details from a CSV file and create a Bronze Delta table
9. Open ingestion/10_etl_streaming_silver Notebook, attach a Cluster and execute it with Run All. This will ingest from Bronze table and create a Silver table that merges the data into the table

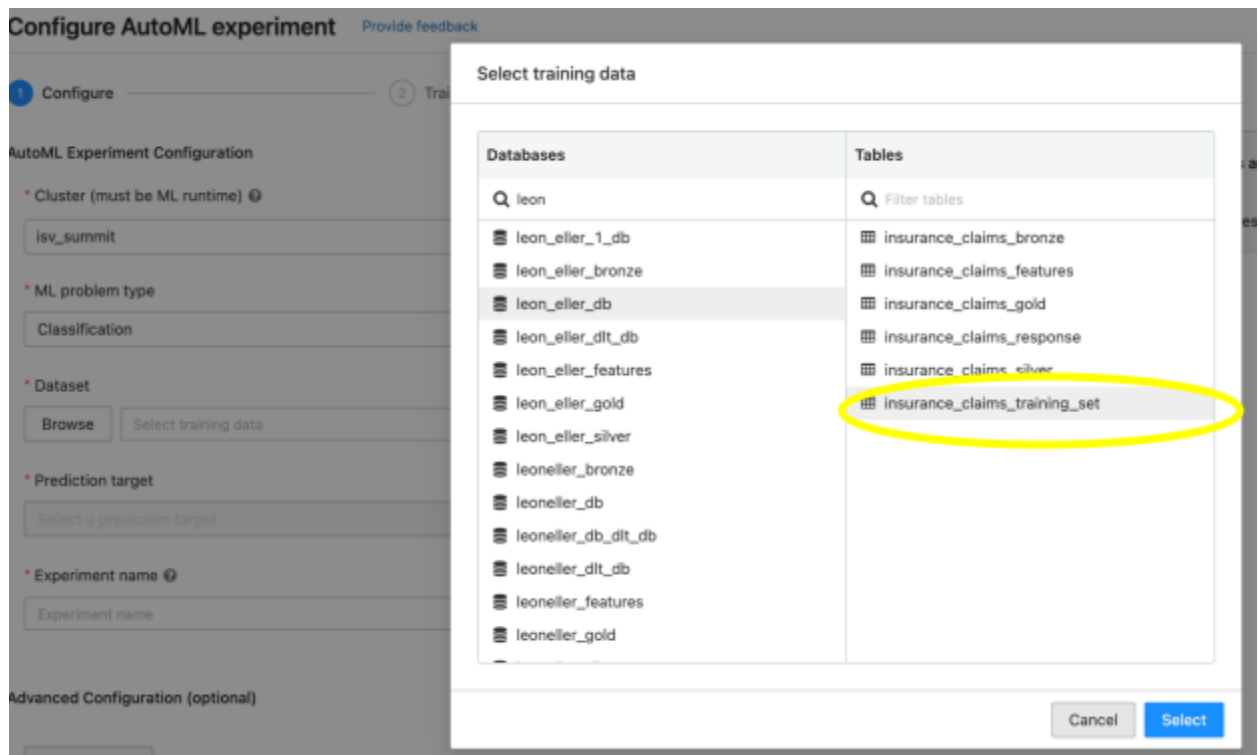
10. Open features/20_features Notebook attach a Cluster and Run All. This will do Feature related transformation from Silver table and register the Features in Feature Store
11. Open features/30_etl_response Notebook, attach a cluster and Run All. This will use join know data whether a claim is fraud or not to the Features resulting in fraud_detected column of 1 or 0. This is used to then train the model. It's separated out because during inference runs this is the column we are trying to predict and is not part of the Features table.
12. You now have the Features and are ready to do ML model training. We will use Databricks AutoML to demonstrate generated ML training Notebooks by AutoML that determines the best model and produces a Notebook that can subsequently be used to enhance and train the model.
13. Switch to Machine Learning view by clicking Machine Learning from the tile menu on the left



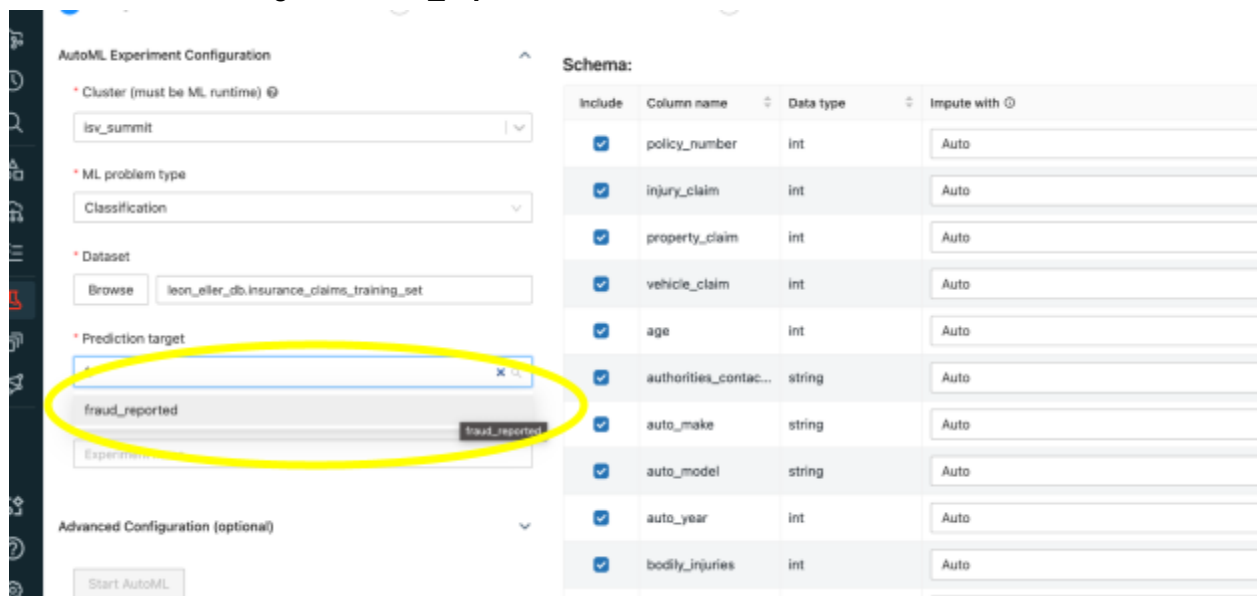
14. Click on Experiments and Create AutoML Experiment



15. Select your Cluster, Model Problem Type is Classification, click Browse to select Feature set. Select your database, it's name should be the first part of your email address you used to login, before the @ sign. Then choose insurance_claims_training_set as Table.



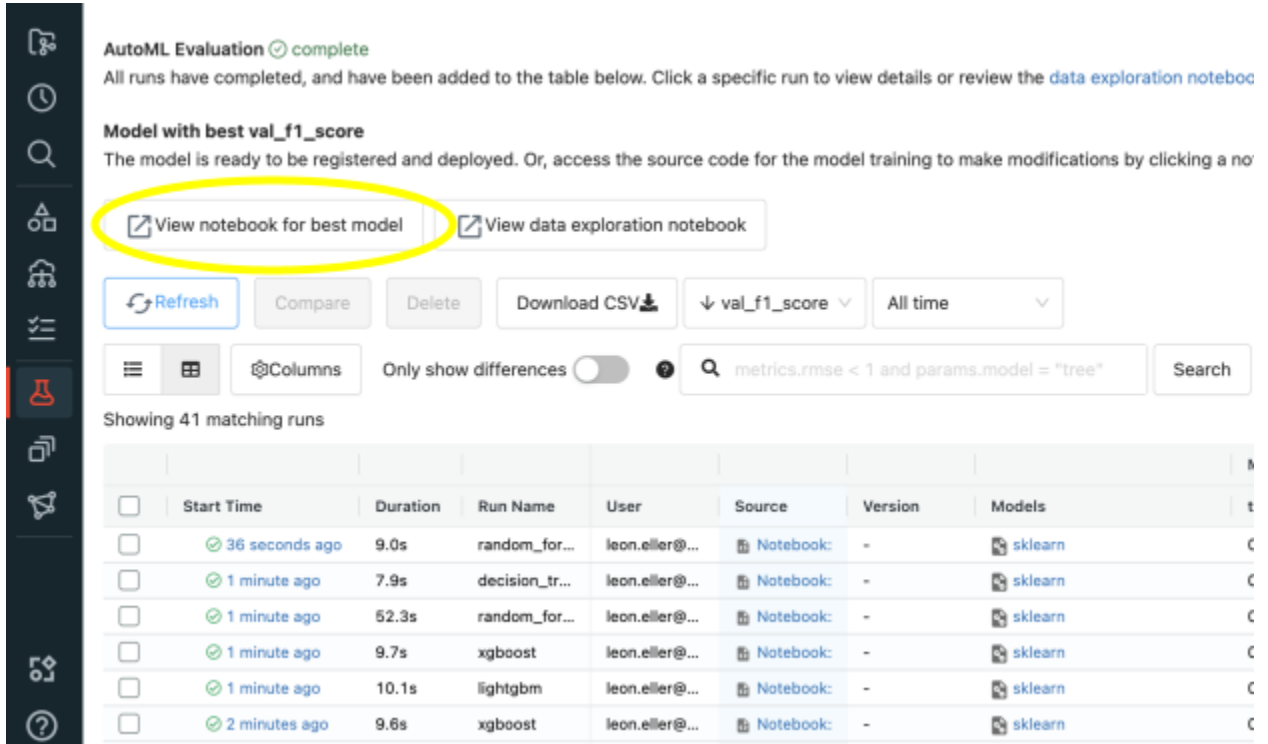
This is the training Feature set that will be used by AutoML model to fit models over.
Select Prediction Target of **fraud_reported**



Click Start AutoML.

16. You should see output that provides feedback about what models AutoML is trying to use for training to determine the best model based on the metrics you are targeting. By default it is F1 Score.
17. AutoML will execute until it runs all specified number of experiments, runs out of specified time, or the targeted metric has stopped converging and is not improving any

more. At that point it will also enable View Notebook for Best Model button and you can navigate to the Notebook with Model Training code



AutoML Evaluation complete

All runs have completed, and have been added to the table below. Click a specific run to view details or review the [data exploration notebook](#)

Model with best val_f1_score

The model is ready to be registered and deployed. Or, access the source code for the model training to make modifications by clicking a notebook

[View notebook for best model](#) [View data exploration notebook](#)

[Refresh](#) [Compare](#) [Delete](#) [Download CSV](#) [val_f1_score](#) [All time](#)

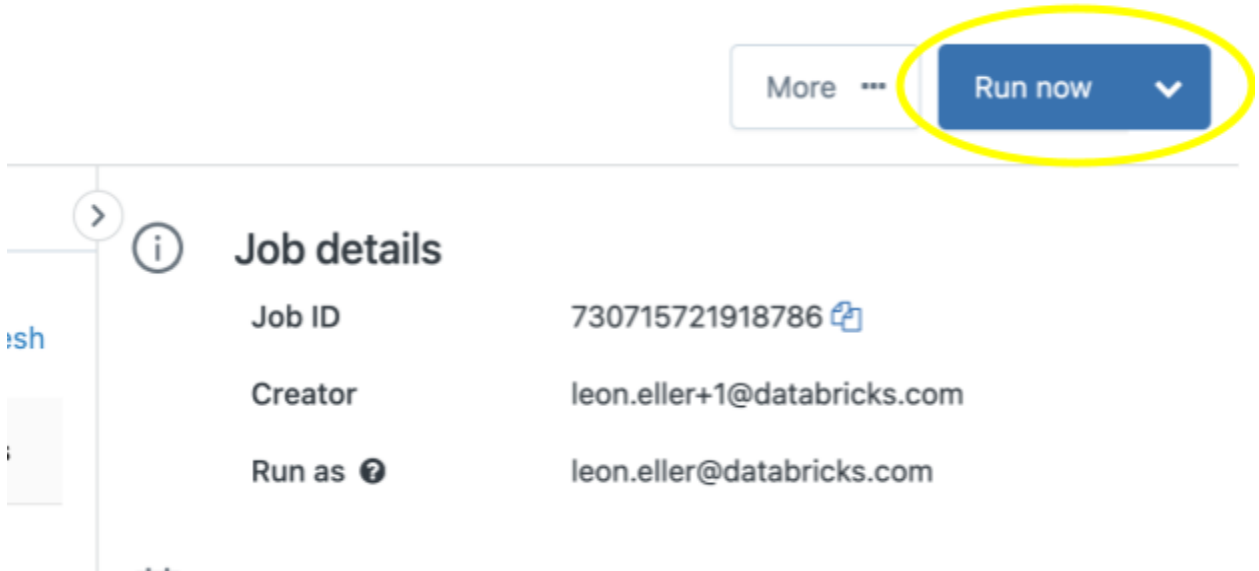
[Columns](#) Only show differences ☐ [Search](#)

Showing 41 matching runs

	Start Time	Duration	Run Name	User	Source	Version	Models
<input type="checkbox"/>	36 seconds ago	9.0s	random_for...	leon.eller@...	Notebook:	-	sklearn
<input type="checkbox"/>	1 minute ago	7.9s	decision_tr...	leon.eller@...	Notebook:	-	sklearn
<input type="checkbox"/>	1 minute ago	52.3s	random_for...	leon.eller@...	Notebook:	-	sklearn
<input type="checkbox"/>	1 minute ago	9.7s	xgboost	leon.eller@...	Notebook:	-	sklearn
<input type="checkbox"/>	1 minute ago	10.1s	lightgbm	leon.eller@...	Notebook:	-	sklearn
<input type="checkbox"/>	2 minutes ago	9.6s	xgboost	leon.eller@...	Notebook:	-	sklearn

18. There is already a Notebook previously created by AutoML in training/40_baseline_model Notebook. Open that Notebook and click Run All. This will execute and register a model in the Model Registry. You can now use this model for inference.
19. Click on Jobs tile and click on your Job to open it. You should see a Job that you can run to perform inference on a streaming pipeline.

20. Click Run Now at the top right



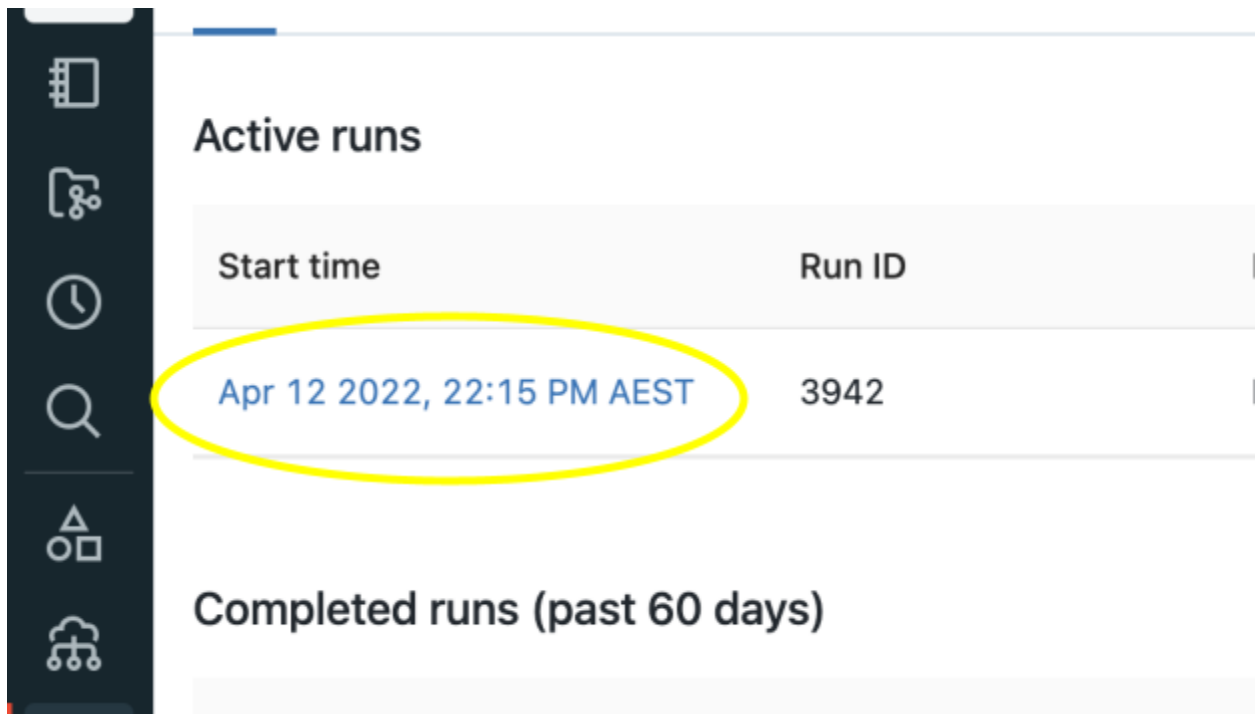
More ... **Run now** ▼

Job details

Job ID	730715721918786
Creator	leon.eller+1@databricks.com
Run as	leon.eller@databricks.com

The Job will start running. The first thing it will do is provision EC2 instances from AWS to form a cluster to run the workload.

21. Click on a link in Active Runs to navigate to the view that shows the state of the running Job

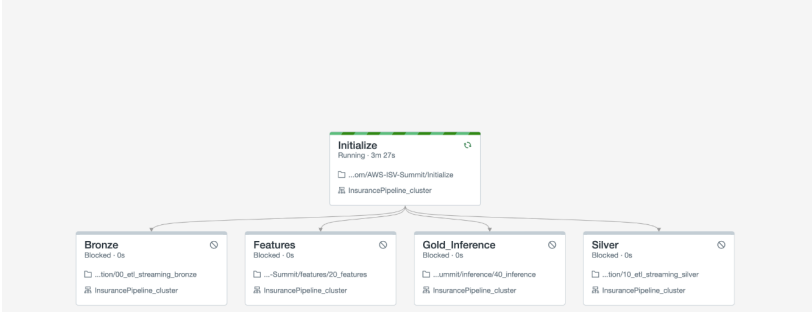


Active runs

Start time	Run ID
Apr 12 2022, 22:15 PM AEST	3942

Completed runs (past 60 days)

22. You should now see status of the running tasks



Initialize
Running - 3m 27s
...om/AWS-ISV-Summit/Initialize
InsurancePipeline_cluster

Bronze
Blocked - 0s
...bron00_el_streaming_bronze
InsurancePipeline_cluster

Features
Blocked - 0s
...Summit/features/00_features
InsurancePipeline_cluster

Gold_Inference
Blocked - 0s
...Summit/inference/40_inference
InsurancePipeline_cluster

Silver
Blocked - 0s
...silver00_el_streaming_silver
InsurancePipeline_cluster

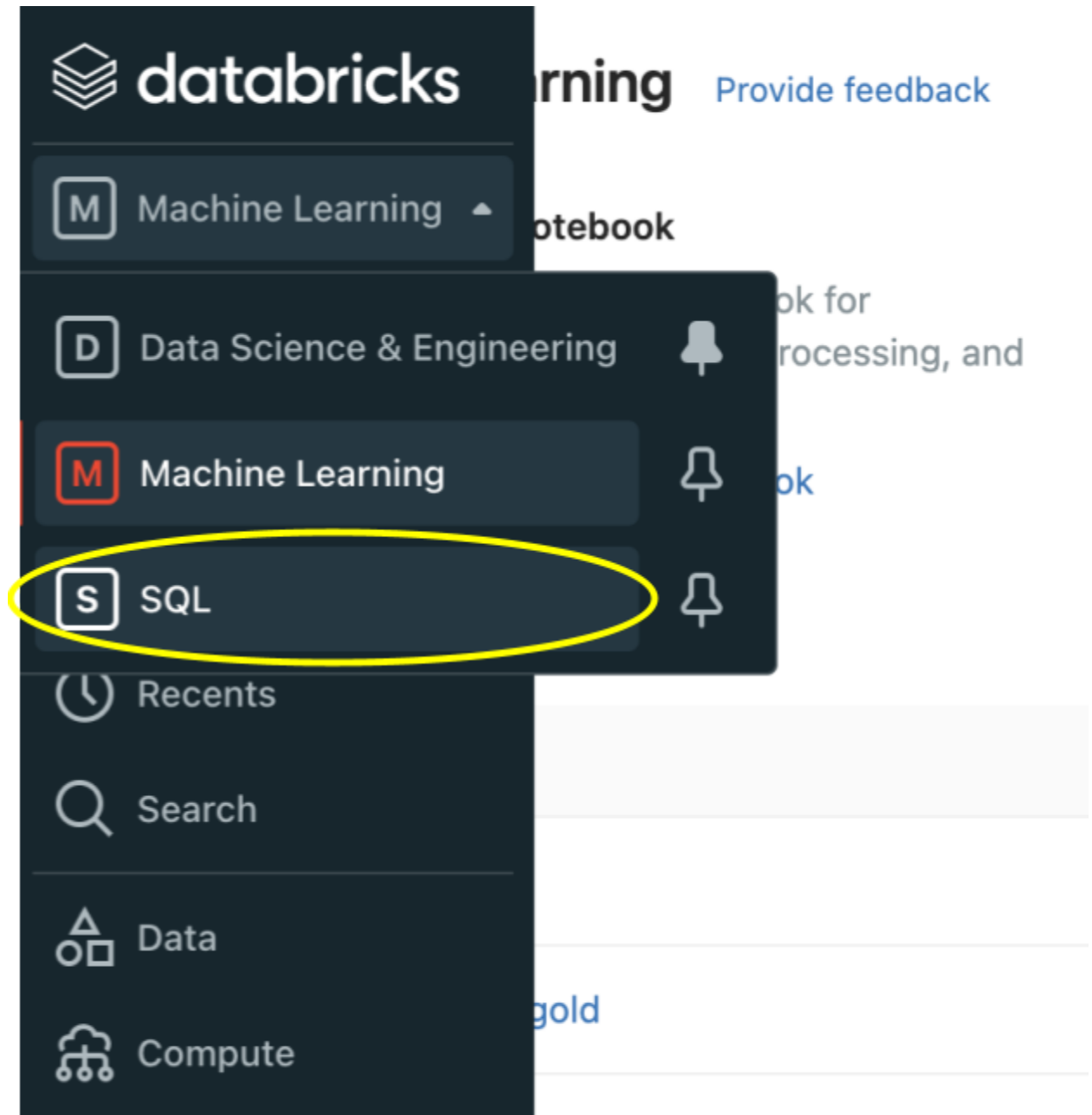
Started 2022-04-12 22:15:59 AEST
Duration 3m 27s
Status [🔄 Running - Cancel](#)

Clusters
InsurancePipeline_cluster
Driver: c4.xlarge, Workers: i3.xlarge, 2-8 workers, On-Demand and Spot, fall back to On-Demand, 10.4 LTS ML (includes Apache Spark 3.2.1, Scala 2.12), ap-southeast-2a

[View cluster](#) [Spark UI](#) [Logs](#) [Metrics](#)

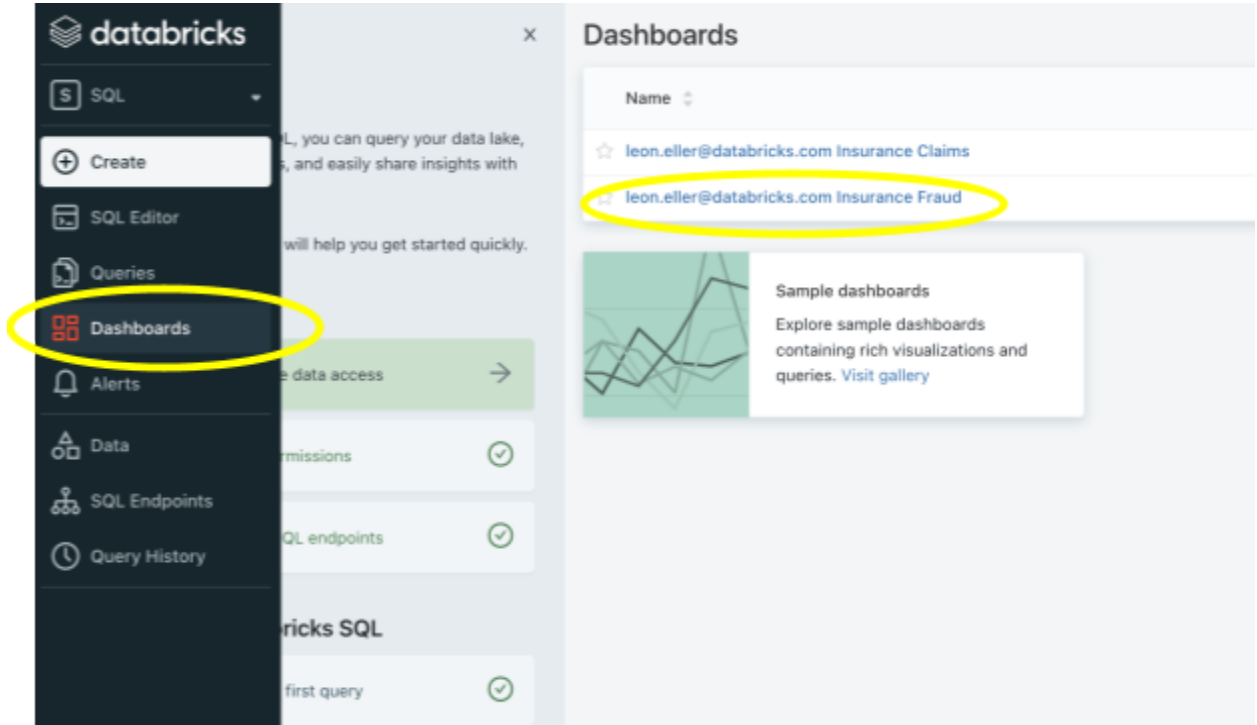
23. Once the Job starts running it continues executing because this is a streaming pipeline that will continuously ingest data as it lands in staging location on S3. Once Initialize Task completes it should be possible to see the output of inference in a table by opening Databricks SQL UI and navigating to the Dashboard with metrics. Because the Job will continue running indefinitely, please make sure you cancel it after you finished with it.

24. Select SQL from the tile on the left panel



This will take you to Databricks SQL UI

25. Click on Dashboards



And click on "... Insurance Fraud" link in Dashboards, where the first part of the dashboard name is your email address. This will take you to the Dashboard which is visualizing the Gold table you just processed with the Job you ran.

26. If you would like to see the queries behind the dashboard, click on the Queries tile from the tiles on the left and open and look at the queries that are there. Feel free to try different queries and run them.