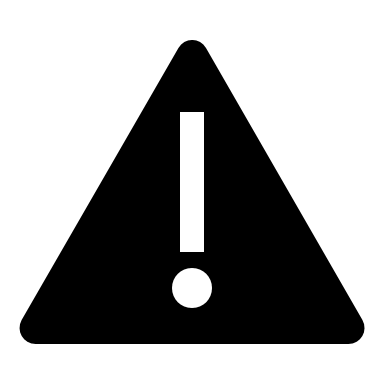
Databricks in a Day

Presenter Runbook

Make sure your class does not use IE or Edge for this course and that all of your clusters are Runtime 5.2. This is necessary for the MLFlow portion of the labs to work correctly.

# Overview

This is the presenter “script” for Insight’s one-day Databricks training workshop. In this workshop, we provide an overview of Databricks functionality though demos and hands-on labs, using a fictitious customer analytics use case as a central theme. Materials, information and support for this workshop can be found in the “DI Data & AI Practice > Workshops” Teams Channel.

The labs for this course were written in [PySpark](https://s3.amazonaws.com/assets.datacamp.com/blog_assets/PySpark_SQL_Cheat_Sheet_Python.pdf).

## New to Databricks?

If you aren’t familiar with Databricks, you can run through the labs to get more familiarized. There is also a great deal of good [documentation](https://docs.databricks.com/index.html) available online.

## Customer Analytics Use Cases

This workshop uses a customer analytics story as a central theme for the labs and demos.

*Wear it Well Retail recently conducted A/B testing to determine optimal discount amounts to increase conversion rates on their website and within their mobile application. They want to use the results of that A/B testing to productionize a new, optimized discount program that displays a different discount offer to each user as they browse the website or app.*

*In order to implement this strategy, the company needs to make available the results of their A/B test data from marketing to their data science team and develop a predictive model to identify which offer will most likely result in a conversion.*

Our students will serve in three different roles throughout the course of the day:

* Data Engineer – As data engineers, we will connect to data, transform it and load it into Delta tables for access by our analysts and data scientists. We will also connect to streaming data and make it available as a data source for our analysts.
* Marketing Analyst – As marketing analysts, we will make use of the new views created by our data engineers to explore the data and answer some basic questions about our campaign. We will also connect the data to Power BI for reporting.
* Data Scientist – Finally, as data scientists, we will build a predictive model that makes recommendations about which discount program to display to our users.

### Use Case Data Sources

* Pre-Campaign Details Web
* Pre-Campaign Details App
* Campaign Results
* Browsing Data (partition of above data, streamed via triggering one file at a time)

# Before the Workshop

* Familiarize yourself with the presenter materials outlined in the README file.
* Confirm your access to the lab environment.
* Download Power BI Desktop.
* [Mount the Data in the lab environment](#_Mount_the_Data)
* Confirm MLFlow Library is installed in your environment.
* Run through the lab in the lab environment.
* Confirm your session details: location, times, food and supporting teammates (sales, tech, etc.).
* Make sure you have a local copy of all course materials on your laptop.
* [Optional] Print out this runbook.

# The Morning of the Workshop

* Arrive on site early.
* Login to the environment and spin up the Databricks clusters.
* Check that there are sufficient seats and power for your attendees.
* Test your projector connection.
* Get the venue WiFi creds and post them for attendees.

## Mount the Data in the Lab Environment

1. Import the Admin.dbc file into your workspace.
2. Run Admin.dbc. This will mount your data.

**Note:** We are mounting to a data source that is owned by Insight. The access key presented in the mount code should not be shared with the clients.

# Recommended Schedule for Course

|  |  |
| --- | --- |
| Topic | Time |
| **Arrival and Setup** | 08:30 – 9:00 |
| **Databricks Overview and Interface Walkthrough** | 09:00 – 10:00 |
| **Break** | 10:00 – 10:15 |
| **Labs: Read, Transform, and Explore Data** | 10:15 – 12:00 |
| **Lunch** | 12:00 – 01:00 |
| **Labs: Power BI (demo) and Predictive Modeling** | 01:00 – 02:30 |
| **Break** | 02:30 – 02:45 |
| **Labs: Structured Streaming and Jobs** | 02:45 – 03:30 |
| **Closing** | 03:30 – 04:00 |

Note: Try to not fall behind. If you are running quickly, it is ok to let the class go early. However, we recommend NOT doing the Predictive Modeling lab before lunch.

# During the Workshop

## Databricks Overview

This PPT deck provides you with content that you will utilize throughout the day. The Overview section provides an introduction to Databricks and the use case the class will be developing to lay a foundation for the labs.

Before each lab, you will return to the slides to give information about the Databricks features used during that lab.

## Interface Walk Through

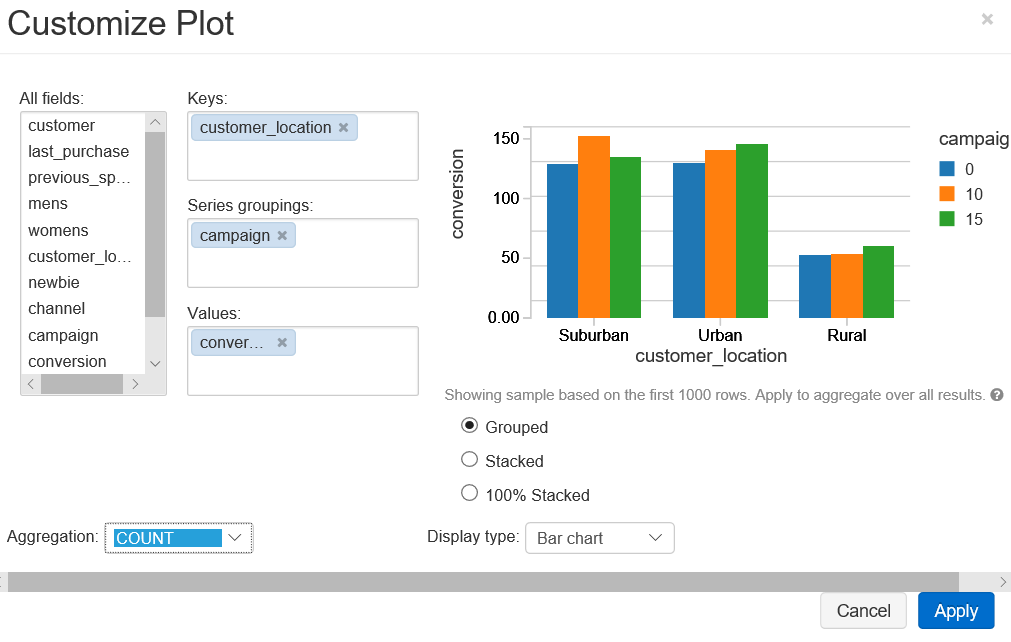
Write the password and URL for Databricks on a white board. Get everyone logged into Databricks with their user IDs. Walk them through the interface:

1. How to find their workspace on the left menu
2. Where to see Clusters and Jobs.
3. Show them a Notebook and walk them through Revision History, Clear, Run, and attaching to a cluster.
4. Have them Clone the labs folder in Shared to their personal workspace and open Lab 01. Help if people are having any trouble. Let them start their break after they have successfully cloned the folder.

## 15 min. Break (should start around 10am)

## Labs

The labs are written so that no one is required to write any code. The instructor should walk through the labs and run cells along with the students. There will be a few places throughout where students are given an opportunity to write code, using examples from previous cells, but answers are always provided. These opportunities could be skipped in the event you are running short on time.

1. Create a DataFrame
   1. Overview
      1. The goal of the first lab is to show the class how to read data from mount and create a [DataFrame](https://docs.databricks.com/spark/latest/dataframes-datasets/index.html). They will learn about how schemas are defined in this lab, inferring and creating a user defined schema (this will come up again in the streaming exercise near the end of the day).
   2. Read Mounted Data
      1. Remind them that the data for our course has already been mounted for them.
      2. Show them the example code for how to mount data from blob storage.
      3. Have them run the code to view what data is available in the directory.
   3. Create a DataFrame
      1. Have them create a DataFrame by running the code. Talk them through how the lines of the code using the comments.
   4. Inferred Schemas
      1. Discuss how easy it is to infer a schema, simply by adding one more line of code. Look at how long it takes to run the code with the inferred schema. Discuss how user defined schemas are faster and that speed can be important at scale.
   5. User Defined Schemas
      1. Create the schema object and discuss how we are specifying data types and column names. After you run the code to create a DataFrame with a user defined schema, have them look at how much faster it is.
2. Transform and Load Data
   1. Overview:
      1. The goal of this lab is to show users how Databricks can be utilized for common transformations done during ETL/ELT. They will learn how to do joins, merges, cleanse data, drop fields, and other common transformations. They will also be introduced to [Databricks Delta](https://docs.databricks.com/delta/delta-intro.html) by creating a Delta table that will be used throughout the remainder of the labs.
   2. Read Data
      1. For this lab, they will access three different files and merge and join them together. Remind them that we are working from mount and have them run the code to view the directory and create a DataFrame from the first of file we’ll be working with.
      2. At cmd 10, the students have a choice between writing the code themselves to create DataFrames for the other two files, using the variable names specified and the code above as a guide, or they simply run cmd 11, which contains the answers. Give them 5 minutes to try writing the code themselves. Walk around the room and see if anyone needs help. The instructor can simply run cmd 11 to keep moving through the course.
   3. Review Data
      1. Have them print the schema and note if there are any discrepancies between our two campaign details files. There is a data type discrepancy on one field.
      2. Have them run the code to cast dm\_campaign as a string.
   4. Merge Data
      1. Now they will run code to merge the two data frames and then do counts to ensure that they didn’t lose any records. You can not that these common operations are both easy and fast inside of Databricks.
   5. Join Data
      1. Run the cells throughout this section, walking them through what the code is doing. They will identify a join field, join the data, then cast two variables as decimals. Again, just note how these common transformations are easy and performant, even on much larger data sets.
   6. Remove Duplicates
      1. When the web and app details tables were merged, some duplicate records were created. Run through the code to look for duplicates and then remove them.
   7. Resolve Discrepancies
      1. In this exercise, we’re going to show them how to use One Click Visualizations to identify issues with the data.
      2. Have them display the full table and walk them through creating a visualization. Have them set the plot options as follows:  
         
      3. Explain that the web and app teams had different mechanisms for referring to Suburban.
      4. Have them run the code to resolve and visualize the results.
   8. Drop Fields
      1. For this exercise, explain that there is an underutilized field that can be dropped from the DataFrame and have them run the code to identify and drop it.
   9. Create Delta Tables
      1. Only the instructor needs to run the code to create and write to the delta table (cmd 49). But don’t worry, if another person in the class also runs it, it won’t break anything!
      2. Discuss how the Delta table will provide access to everyone simply by having on person run it.
3. Reading, Exploring, Manipulating Data
   1. Overview:
   2. For this lab, our students are now acting as marketing analysts, exploring the data table created in the previous exercise. Remind them that this is the first time the analysts have seen all of this data joined together. Read Data from Delta Tables
      1. Have them run the code to read and display the data from the Delta tables. Remind them this is the transformed code created in the previous exercise.
   3. Aggregate by Spend
      1. For this part of the exercise, they will explore answering a few basic questions, then be given some time to explore on their own. You can skip the self-exploration time if you are running behind.
   4. One-Click Visualizations
      1. Have them run the cell to display all data. Give them some time (5 min) to explore the data as they see fit. Toward the end, show them the answer to the question of whether there is any variation of features that results in a lesser discount leading to higher conversion rates by choosing the following plot options:  
         Ask whether anyone else found anything interesting.

**Now it's time for lunch! If you are running ahead, you could do the Power BI Demo before lunch, but definitely save the Machine Learning labs for after!**

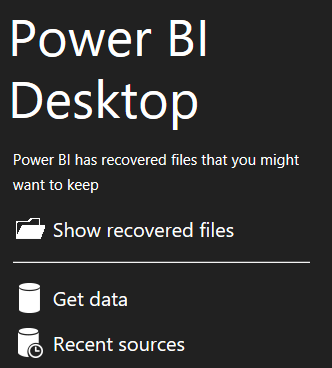
1. Power BI
   1. Overview:
      1. This will be a demo rather than a lab. You can share your screen and walk them through the process of connecting Power BI to a cluster to develop reports from Delta tables that use the power of distributed computing for increased performance.
   2. Connect PBI to Databricks
   3. Introduce PowerBI to Databricks Connection
      1. Define Power BI for anyone who has not used it:

Microsoft Power BI is a business analytics service that provides interactive visualizations with self-service business intelligence capabilities, enabling end users to create reports and dashboards by themselves.

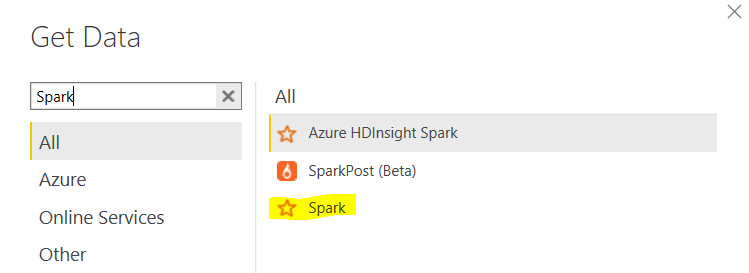
* + 1. Explain to attendees why they might want to connect PowerBI directly to the Spark cluster and access data in delta tables:

Connecting to Databricks improves performance. Using the cluster speeds up operations by offloading data processing to Databricks. This is useful for large datasets or near real-time analytics.

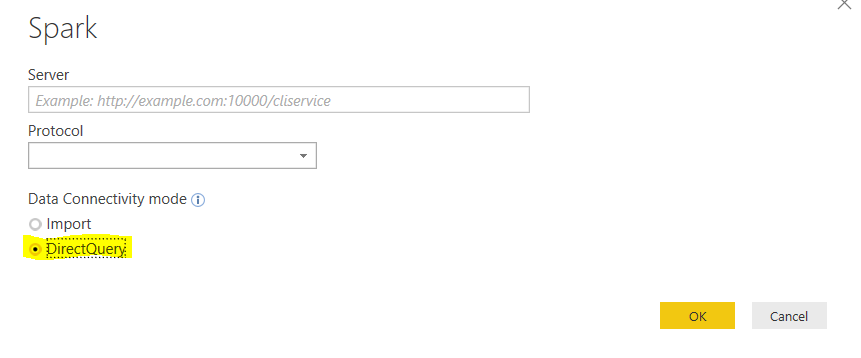
* 1. Get Started with PowerBI
     1. Open PowerBI Desktop
     2. Select Get Data from the welcome menu



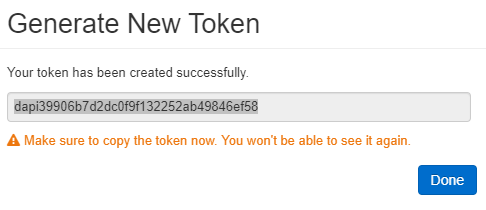
* + 1. In the search field, type “Spark”. Then select Spark > Connect.



* + 1. Before typing anything in the fields, select DirectQuery radio button.

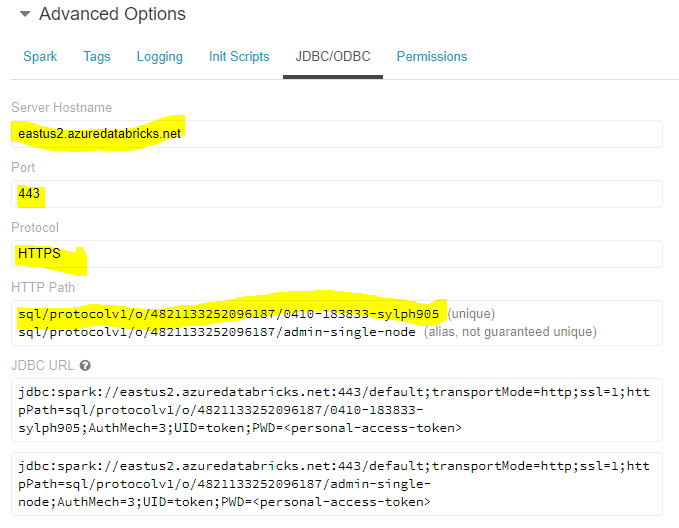


* 1. Collect Cluster Connection Info
     1. Navigate to Databricks >  (aka your user account).
     2. Under Access Tokens, select Generate New Token.
     3. Type ‘Power BI’ in the Comment, then select Generate.
     4. You should see a token screen that looks like the one below. Use Ctrl+C to copy this token.



* + 1. Open Notepad, then paste the token here.
    2. Navigate back to the Databricks window.
    3. Go to Clusters > (select your cluster) > Advanced Options > JDBC/ODBC
    4. Using your Notepad, copy and paste the highlighted values below into this format:

https://eastus2.azuredatabricks.net:443/sql/protocolv1/o/4821133252096187/0410-183833-sylph905

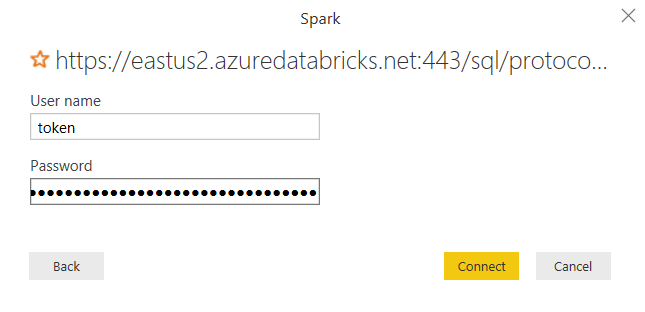


If you have trouble: <https://docs.azuredatabricks.net/user-guide/bi/power-bi.html>

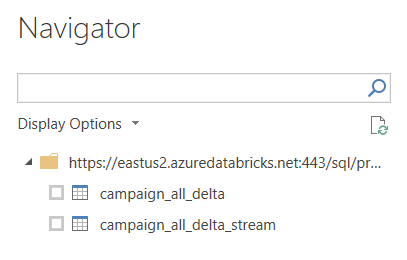
* 1. Connect Power BI to the Cluster
     1. In order for this to work, make sure your cluster is running.
     2. Navigate back to Power BI, where you have the Spark connection window open.
     3. Enter the connection info and select OK.



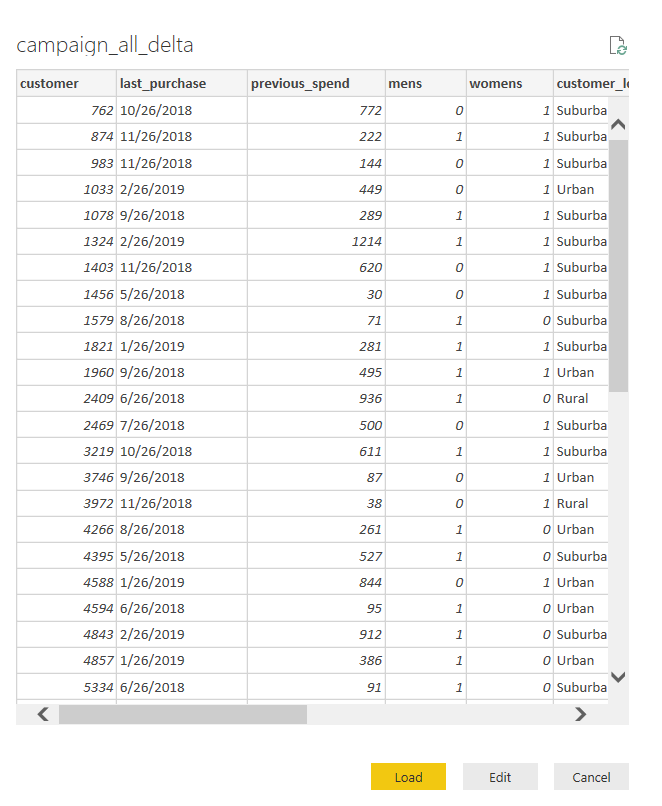
* + 1. On the next prompt, use ‘token’ for the username, and refer to your Notepad for the token password. Then click Connect.



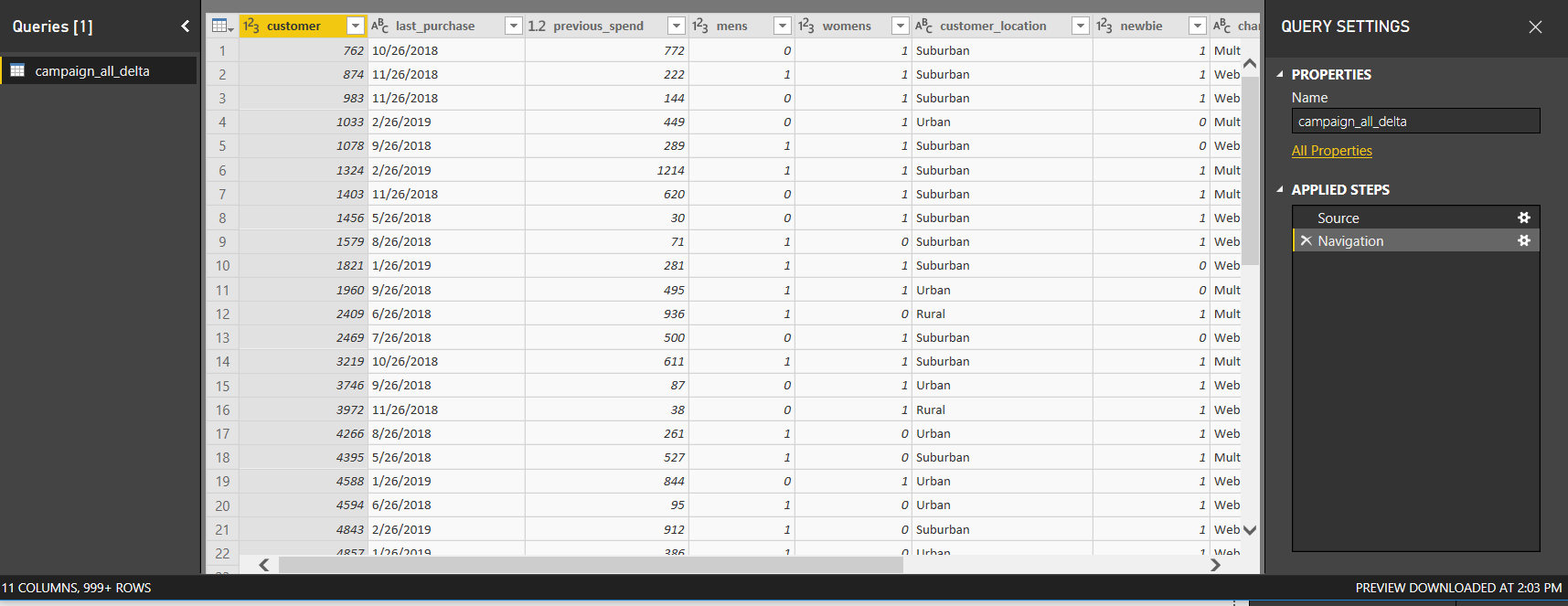
* + 1. You should soon see the delta table listed under the Navigation pane.



* + 1. Select campaign\_all\_delta, then Load.

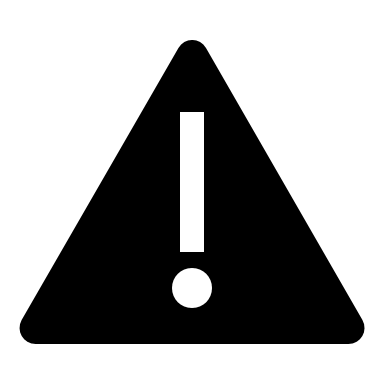


* + 1. If you click on Edit Queries in the header menu, you should see the delta table data there.



* 1. Closing

Explain to attendees that we will stop at building the connection, since dashboarding is out-of-scope for the training. However, this method will connect PowerBI dashboards to the Spark cluster such that any updates to the delta table will be automatically reflected in PowerBI.

1. Machine Learning  
   Check in with your class before starting this exercise. If you don’t have any data scientists or attendees with strong interest in machine learning, this lab may lose them. In that case, skip ahead to Step c below. Have everyone update the code as outlined in the instructions, then Run All for this Notebook. Continue on with the instructions after Step C to walk them through the benefits of MLFlow to enable a data science team to experiment and iterate on model development in a transparent and organized way.
   1. Introduce the Lab
      1. In this lab, the attendees will create a predictive model. This predictive model will produce a probability of conversion for each customer.
      2. Screen for questions about the use case.
2. Train & Test the Machine Learning Model
   1. Guide attendees through 1. Load Required Libraries and 2. Read Data from Delta Tables.
   2. When starting 3. Feature Engineering, instruct attendees on how to think about the large block of code:

It’s not important that the attendees understand the code listed for each part of feature engineering. Just know that this step is essential for preparing data to be fed into a machine learning algorithm. Under other machine learning libraries, these steps requires a lot more code. However, Spark MLlib keeps these steps relatively short and easy to reproduce.

* 1. When running step 4. Train the Predictive Model, instruct attendees:

When we train a model, we set parameters (or instructions) that tell the model how to learn patterns about the data. For every machine learning problem, there is a “sweet spot” for these parameters that result in the best performance. We cycle through training, evaluating, and tuning these parameters over and over again until we are satisfied with performance.

* 1. Run through step 5. Evaluate the Predictive Model.

Call out that the performance improves after tuning parameters.

1. Use MLflow to Capture a Model Run
   1. When running step 6. Capture the Predictive Model in an Experiment:
      1. Ask attendees to change the # in the parquet file path to their own:
   2. Give attendees an explanation of MLflow experiments:

In MLflow, you have to set up an experiment. Within this experiment, you will create runs, which are going to be different models that you have trained and tested. For each run, you specify parameters and metrics to capture. You should also save trained models so you can access them later for production use. Today, we will explore creating MLflow experiments and runs, but we won’t be saving our trained models. However, the code for saving a model is included and commented out.

* 1. When running 6.2 Record a Run, briefly point out each chunk of code in the cell:
     1. Re-define parameters
     2. Re-train the model
     3. Evaluate the model
     4. Log parameters and evaluation metrics for the run
     5. End the model run

1. Explore the MLflow UI
   1. After the code for 6.2 has finished for everyone, instruct attendees to navigate to the experiment: Workspace > (their user directory) > marketing
   2. In the experiment UI, hover over each of the tracked items for the two experiments that have been created between steps 5 and 6. Instruct attendees:

Once again, tuning parameters improved performance. Data scientist run 10s to 100s of models. This is where the MLflow UI adds value – you have a central place to manage and retrieve your best experiments. You can easily understand what worked and what didn’t.

* 1. Experiment on your Own
     1. Navigate back to the 04 Machine Learning lab, step 6.3 Repeat with Your Own Parameters.
     2. Instruct attendees to tune parameters again in the 6.2 experiment run. Re-run the cell with new parameters that the attendee selects.

1. Streaming
   1. Overview
      1. In this lab, we are moving forward in time to after our predictive model along with business logic was operationalized to present offers to users in real-time on the website. We are going to put our data engineering hats back on to connect to data being processed from the web that provides us with information about whether those discounts are leading to conversions and how much customers are spending.
   2. Connecting to Streaming Data
      1. Explain that when creating a Data Frame from streaming data, you must specify the schema. A schema cannot be inferred because at the time of object creation, there may be only a few rows of data, which is not enough for Spark to make an inference about data types.
      2. Have them run the code to specify the schema and create the DataFrame.
   3. View Streaming Data
      1. Now they need to be reminded that this new streaming data source is just like any other Data Frame and can have similar operations performed on it. Have them run the code to view the DataFrame.
   4. Transform Streaming Data
      1. Now have them run the code that aggregates spend by campaign. They will see the plot results updating in real-time. Have them visualize it to really get the full effect.
   5. Write into Delta Table
      1. Lastly, we will write their streaming DataFrame into a Delta table. Only the instructor should run the code in cmd 12. They will be able to access your table from their notebooks.
      2. Have them run the code to view the table and count the records. They can run the code to count the records multiple times to see it increment over time.
   6. Turn off Your Streams
      1. Make sure everyone turns off the streams before moving on. They can do this by running cmd 15.
2. Jobs  
   Please run this as a demo. The students are welcome to follow along with the code in Notebook 06, but **don’t have them Run a job**.
   1. Create a Job
      1. Explain and do not run ’06 Create and Run Job’ code.
      2. From ’06 Create and Run Job’, select Jobs on left hand menu.
      3. Select + Create Job
      4. Build the Job
         1. Name: ‘DBIAD Job’
         2. Select Notebook: ’06 Create and Run Job’
         3. Mention that you can include dependent libraries (not required for this lab)
         4. Cluster: Edit > Cluster Type > Existing Cluster > Select Cluster > Confirm
         5. Schedule: Edit > show attendees the options > Cancel
         6. Expand Advanced:
            1. Alerts > show attendees that you can list an email address > Cancel
            2. Mention the rest of the options
   2. Run a Job
      1. Run Now (**instructor only**!)
      2. Explain to attendees the job duration and Pending/Succeeded/Failed status
   3. Show Output
      1. After you have a successful job run, open ’07 View Job Output’
   4. Everyone can run this notebook. Everyone should see a table with 3 rows.

## Closing The Day

Quick walk through of what we did today and how all the pieces fit together as well as how the disparate pieces might be used independently. Share the slide that shows where they can get more information on Databricks. If your Sales and Alliance teammates are there, allow them to close with any calls to action, or follow up on any items they have requested you address in closing.