**Using Spark SQL on Databricks**

# **Introduction to Module 3**

Hey, welcome to another module. So at this point you're probably wondering when do I actually get to work with data. And you guessed it, it's in this module and we're pretty excited about it. So in the previous module we talked all about the big data landscape.

We talked about Apache spark. We mentioned Spark SQL and now we want you to do some hands-on practice. The first thing we're going to do is get you familiar with and get you set up with Databricks Community Edition. Now Databricks Community Edition is what you'll be using to complete all of the hands on components of this module. It's free and it will be a great way to introduce you to Databricks as we work with it throughout this module. So by the end of this module you'll be able to successfully navigate the Databricks Community Edition user interface to complete your course assignments. You'll be able to use Databricks to manipulate data by running basic queries with Spark SQL. And finally, you've apply what you've learned in the module to complete a hands-on lab assignment. And in this lab assignment you'll be performing exploratory data analysis in Databricks. So when you're ready, go ahead and click on to the next video.

Welcome to this module.

# **Signing up for Databricks Community Edition**

Hey, I'm so glad you made it here.

We're about to have some fun.

First, we're going to get you signed up on

Databricks Community Edition so you can start

using Spark, power your queries.

Databricks Community Edition is

a free version of the Databricks platform.

As I mentioned earlier,

all of us work for Databricks,

which is a unified platform for

massive scale data engineering,

collaborative data science,

full cycle machine learning and business analytics.

In this class, we'll be using Databricks to access Spark.

Without Databricks, Spark requires instead of

a management that is beyond the scope of this course.

Databricks Community Edition includes

a managed Spark service that

will allow you to jump straight in

and start working with Spark and using SQL on Spark

without having to worry about

managing computational resources.

By the end of this video, you'll be up and running with a Databricks Community Edition account. Before we get started, note that if you are already signed up for community edition, you can skip this process and move onto preparing your workspace environment, and two, if you already have a work issued Databricks account and access to a company workspace, we still recommend you sign up for a community edition account for the work in this course. To create a new Community Edition account will follow these steps. First, go to the "Try Databricks" sign-up page and enter all of the required information. We'll be using community edition for this class. So click the "Get Started" button on the right. You should get an e-mail in the next few minutes

asking you to verify your e-mail address. Click the link in that e-mail to set a password for your new account. Make sure it includes letters, numbers, and symbols. Okay, great. Now you're set up to participate in the hands-on activities for this course. Next, we'll give you a quick tour of the workspace.

# **Preparing your workspace**

All right. Now that you've set up the Databricks Community Edition, let's review a couple of features that will help you as you work through this course. By the end of this video, you should feel comfortable preparing the workspace to run queries. Let's start with the Databricks user interface. Anytime you log in, you'll first be brought to this workspace homepage. Notice the workspace sidebar on the left. You'll use the sidebar as your primary navigational tool, as you work through the course content. In addition to the sidebar, the home screen provides some quick links that you may find useful. For example, if you have a few minutes to spare, please feel free to go through the QuickStart tutorial to familiarize yourself with the workspace functionality. Aside from using the sidebar to navigate, you'll need to know how to start a cluster. A cluster is the set of computational resources that will power your queries. All work on Databricks must be powered by a cluster. To do this, simply click on the Clusters tab in the sidebar, and then click ''New Cluster''. Then we're going to select a name for the cluster. Note that in a shared workspace, creating a meaningful name for your cluster is important so that you can easily find and use it. In Community Edition, you'll be the only one working in this workspace, unless you invite another user. Once you've selected a name, you'll select the Databricks runtime version 7.2 from the drop-down menu. The Databricks runtime is a highly optimized version of Apache Spark. That means when you run Spark in Databricks, you're getting all the benefits of working with Spark plus some extra features and optimizations added by Databricks. Click "Create Cluster" to start your cluster. It will take a few minutes for your cluster to start. Once it does, you'll be ready to start running queries. Keep in mind that clusters automatically shut down after two hours if they're not being used. So if this happens to you, don't worry too much about it. You can always restart the cluster if it appears on the cluster menu or you can create a new one. For now, feel free to move on to the next video in this course while your cluster finishes starting up.

# **Course Materials**

This is the link to all of the course materials. Follow the instructions in the video Working with Notebooks" (coming up next) to add these files to your Databricks Community Edition workspace.

Please note that DBC files cannot be opened directly on your computer.

<https://files.training.databricks.com/courses/moocs/SQLDA/Lessons.dbc>

# **Working with notebooks**

All of the work that you will perform in the databricks workspace will be done in a notebook. A notebook is a series of cells where you execute commands in a number of languages actually, but we'll be using SQL.

In this video, we'll explore Databricks Notebooks and the ways that you'll work with them in this course. By the end of this video, you'll be able to explain how to use Databricks Notebooks to execute SQL queries.

All right, so just to get started, I'm going to click on Workspace and I'm going to right click here and I'll create a notebook. I'm going to name that Demo because that's what this is. My default language is going to be SQL. You could also choose Python, Scala or R if you know those languages. Not for this course but in general, and then I'll click Create. Okay, so that has attached to my running cluster. You may have noticed in the previous menu that cluster was selected so that automatically attached. If you come into a notebook like this and you find that you're not attached to a cluster, you can use this menu to attach to any running clusters you have permission to. So here we have an example of a blank databricks notebook. It contains one cell and I'm going to use this space to write and execute my first SQL command.

Now I have some data preloaded on this workspace, so I'm going to use that for this demo. It looks like I could use the table bikeshare.

So I'm going to just write SELECT \* to pull that information up. [SOUND] And then I'm going to run this command. You have a couple of options for running. You can use this button on the right side of the cell.

Or if you're less of a mouse user, you could use the Control and Enter button to run a command or also Shift and Enter. All right, so we can see it ran that SELECT F statement and the return is this table. If you click on this table, you can see you can scroll within the actual table. There too many columns to display. Beyond that, we'll work with all of these buttons a little bit later, but just to make you aware this is how you get to databricks built in visualization tools. And then this is if you want to download a CSV to your local machine, you can do it with this button.

It is, it may be limited to a certain number of records though. Okay then, let's also go through each of these on the right side of the cell. We've got our run button that we just used and then we got this button which is will add any table or visualization to a dashboard. We'll be doing that later on in this module.

We've got this formatting cell which is very handy. I'm personally really like this format SQL command, and then we've also got some other options for you to explore. If you ever don't like a thing that you've done you can also use ordinary keyboard shortcuts within the cell. So I hit command V, I'm going to mark here. It undoes that formatting.

As you work on your course assignments, you won't be working with a blank notebook like this one. Instead you'll import course notebooks that we've already prepared for you. We'll walk through how to use those in the next video.

# **Using course materials**

In this video, we're going to show you how to work with the readings in the labs. As you move through this course. In course era we've provided you with the link in the course materials reading in module 3. The link ends with the extension dot DBC. This is a collection of notebooks that you can easilly important to Databricks. This link is a downloadable file which contains all of the hands on activities and last. For this course you don't need to download it, however were simply going to use the link to import courseware to your community addition account. Let's walk through that process now.

If you click on the link it will automatically download DC files or computer. That's totally fine. You could also import that way instead. When I'd like you to do is just copy it so that we can import without downloading. Once you get that link back in community edition. We're going to click on the workspace tab. And then just right click anywhere on this whitespace. At that point you'll click import. And we're going to choose the option to import via URL, so I'll paste that here and then. Like import. And there it is. This folder contains all the hands on lessons for this course. If we look inside the folder we can see that it separated into modules and then we've also got a solutions folder. If you click inside any module. You'll see a collection of notebooks I'm going to start by opening this 3.2 basic basic queries notebook. An notice that when I open that notebook immediately it is not attached to a cluster. If your notebook is not attached to a cluster, it won't run. If I try to run something. I'll get this message that let's me know I'm not attached to a cluster an I should be so you can just click attach and run if you want that to be a general setting, you can of course click this. I'm going to attach that and run and it does that. The other thing you could do is your cluster will appear right up here and you can attach via this menu. All right? So this is an example of one of the pre populated notebooks that you're going to be working with. The first notebooks in any hands-on module are interactive reading. They contain some explanations and context for the SQL commands that you're working with. When you're working through these notebooks, read the cell, run the code and observe what happens. The first thing that you want to do in any of these notebooks is run. This includes classroom setup cell. I said before that we would set you up in community edition and you wouldn't have

to worry about connecting to storage or database or anything like that. Actually, we're going to handle that for you, and this is a script that just runs in the background and make some necessary connections so you can get data into your workspace. I'm going to click run cell. You don't have to pay too much attention to this output. All this is telling me is that you are now connected to our training data. All right, so this is a good example of what your explanations might look like as you work through, you will be able to click on some links to explore documentation associated with some of these commands. We of course encourage you to read through the documentation as it will provide some background for some of these commands. These explanations are usually followed by some cell that contains code. When you run that. You'll get some kind of output. This is just the table creation command, so the only output here is OK.

As you move on, you'll be querying tables, so we just created this table, People10M.

And if I want to view the table, I can run that select star to view all the rows and columns.

And then the cell that contains code, please feel free to go in and explorer what you can do. If you want to test out some new commands. This is obviously not super new.

If we want to, instead of selecting all, just select this first name column from this chart you can adjust any of these cells to show that if anything happens in your exploration and you somehow break it. So nothing, nothing works anymore and you're getting some kind of error message. You can always rollback it changes. There's a version history in this notebook. And you can look at previous versions if you want to roll all the way back to the beginning. Very first, imported the notebook and restore that revision.

There we go, brand new fresh network alright, so these are the interactive readings when you finish these interactive readings, you will have an end of module quiz on the content that's here. Now let's talk about the labs. They're a little bit different. The last notebook in each module is a lab which contains exercises that you will complete. Test your skills for these. You'll answer quiz questions based on the results of your lab work. So for example. You'll see a direction like this. You've got some code to fill in here, and when you get the result from that code, you're going to. Go into car Sarah. An answer that question we recommend keeping your notebook open with Corsair a side by side. So you'll be logged into cursor, of course, and you can have these open right

next to each other so that you can find the answer, and you can write it in. OK, that's pretty much it. When you reach course assignments involving hands-on work will direct you to start up a cluster an navigate to the appropriate notebook next. Well, get to running queries.

# **Basic queries with Spark SQL on Databricks**

# **Basic queries with Spark SQL reading introduction**

My role in this course is to help prepare you for all the hands-on activities. Whenever you see me, you'll know that a hands-on activity is coming up next. You can find all they hand-on activities for this course in the DBC that you were asked to upload to your workspace. They will all come in one of two formats and both of them will acquire work and the Databricks notebooks. One hands-on activity is the interactive reading. An interactive reading is where the notebook will demonstrate how to use various tools and commands to access, manipulate, and present your data. The second interactive activity is a lab. Lab they're always the final activity in any module. They are meant to assess your learning and practice your skills that you've been developing. A quick note, both of these activities will always be followed by a quiz. Be sure to leave your notebook open so that you can use it to answer the questions in the quiz. Now that you're clear on the types of activities that you'll find in this course, let's get ready to start running some queries. Good luck.

# **Basic Queries reading activity**

Login to your Databricks Community Edition account. If you haven’t already imported the course dbc files, review the video, "Working with Notebooks" (in Module 3) )to learn how to do that.

Locate your SQLDA folder. Open Module 3, then click on "3.2 Basic Queries" to open your notebook.

 As you read through the notebook, run the corresponding commands to see how they work and how results are displayed. Feel free to explore the space by modifying commands, adding new cells, and/or writing new queries.

**Data Visualization**

# **Data Visualization on Databricks reading introduction**

Hi again. Great job exploring basic SQL commands in the last hands-on activity.If you're already familiar with SQL, you probably didn't come across anything brand new.

In fact, hopefully you're able to see how easy it is to use basic SQL syntax to access Big Data via Spark. At this point, we're simply hoping to get you more acquainted with using notebooks on Databricks and exploring some of the built-in functionality that you'll find useful in your future work. Speaking of built-in functionality, next, we're going to see how we can take your queries and turn them into visualizations right from your Databricks notebooks. Have fun.

# **Data Visualization reading activity**

Login to your Databricks Community Edition account. If you haven’t already imported the course dbc files, review the video, "Working with Notebooks" (in Module 3) to learn how to do that.

Locate your SQLDA folder. Open Module 3, then click on "3.3 Data Visualization" to open your notebook.

 As you read through the notebook, run the corresponding commands to see how they work and how results are displayed. Feel free to explore the space by modifying commands, adding new cells, and/or writing new queries.

# **Data visualization tools**

You may navigate through the transcript using tab. To save a note for a section of text press CTRL + S. To expand your selection you may use CTRL + arrow key. You may contract your selection using shift + CTRL + arrow key. For screen readers that are incompatible with using arrow keys for shortcuts, you can replace them with the H J K L keys. Some screen readers may require using CTRL in conjunction with the alt key

In addition to being a place where you can execute a Spark SQL queries, Databricks notebooks also include built-in visualization tools that you can use to share insights with your team. As you view this video, you may want to have your Module 3 Notebook 3.3 data visualization with Databricks notebook, open so that you can follow along with the instructions for this visualization. In this video, we'll show you some of the ways you may visualize and present data in a Databricks notebook. In the previous reading, you work through this notebook and read the associated content. Recall that in this notebook we're investigating ratings data about a collection of movies. Scroll down to command 12, visualize the data. When we run the query, a table of results shows up underneath. For any table, you can use the built-in visualization tools to make a chart. Click on the chart icon under the table. Some chart will appear right away. Click on the plot options to customize the chart. For this chart, let's use the month as the key and the average rating as the values, then we'll specify that we want to show this as a line chart. Notice that you can specify the y-axis range, will keep the default for this chart or change to a logarithmic scale, if we have a very wide range of values, we might want to do that, and set global color consistency. This last setting, global color consistency is useful if you have many charts being presented altogether. It allows you to keep the same colors associated with a certain key.

Since we only have one chart here it's unnecessarily. Okay. Great job. Feel free to explore some of the other available chart options for how you might want to visualize this data

# **Exploratory Data Analysis lab introduction**

It's time to put your skills to the test. Now that you've had some experience working with Spark SQL and Databricks, you're ready to take on your first challenge. Ultimately, we want to be able to produce a report which will show which countries are generating the most online orders. In this lab, you'll be performing a series of steps that will lead to the creation of this dashboard. Good luck.