**Delta Lake with Spark SQL**

### Building and maintaining Delta tables

# **Introduction to the module**

Welcome to another module. It's a little bit sad, but you're almost done with the course. Well, it's probably more set for us because we've had a really great time with you. You're probably excited, you're almost done almost that certificate though, in a previous module, you learned all about Delta. Deltabricks, Delta Engine. You learned about the lake house and now we really want to give you some hands-on practice with all of those things. In this module, we said some examples so that you can see how you can use and create delta tables to work with files. You'll also get to experiment with the optimizations built into Delta Engine on Databricks. By the end of this module, you'll be able to use Databricks to access and modify records in the Delta table. Programmatically create Delta tables, optimize performance with Delta Engine utilities. Execute queries in diagnose problems with your company's single source of truth, and update and delete records associated with the user. Welcome to this module. Good luck.

# **Intro to Using Delta reading**

In this lesson, you'll now get started using Delta Lake. As you read in the previous lesson, Delta architecture creates a pipeline that serves up data in the form of a goal table that constitute ready for analysis. In the upcoming activity, we'll work through a simplified version and see what that looks like, so that you can begin creating, converting, and modifying Delta tables. The data that we'll be using this time is from a fictional company called Moveo, a popular fitness tracker supplier. We'll be exploring heart rate data that would have been collected over a period of months. So, your job here is to demonstrate how you can convert existing data and workloads to Delta tables so that you can take advantage of optimizations available via the Delta engine.

# **8.1 Using Delta**

Login to your Databricks Community Edition account. If you haven’t already imported the course dbc files, review the video, "Working with Notebooks" to learn how to do that. Locate your SQLDA folder. Open Module 8, then click on "8.1 Using Delta" 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.

# **Managing Records in a Delta table**

For Moveo, the fitness tracking company who is relying on our analysis, collecting a high volume of complete and accurate data is imperative to making smart business decisions. Fortunately, in this particular business, users have agreed to give direct access to their personal information by wearing monitors that report data directly back to the company. In the previous reading, we discovered some errors in the data. Next, we'll be working with Spark SQL and Delta to correct those errors.

# **8.2 Managing records**

Login to your Databricks Community Edition account. If you haven’t already imported the course dbc files, review the video, "Working with Notebooks" to learn how to do that. Locate your SQLDA folder. Open Module 8, then click on "8.2 Managing Records" 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.

# **Delta Engine Optimization**

# **Delta Engine Optimization Introduction**

The advantages of Spark and Delta engine are especially clear when we're working with large quantities of data. Throughout this course, we provided sample data, so we can give you practice working with different data formats and structures. However, due to logistical constraints, the data were usually working with is actually rather small. So in this reading, we're going to take a break from our work with movieo, and we're going to begin exploring data about US-based flight schedules, which is available through our databricks datasets. This data is a CSV file with over 4 million rows and 29 columns.

# **8.3 Optimizing Delta**

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 8, then click on "8.3 Optimizing Delta" 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.

**Delta Lake Lab**

# **Delta Lake Lab Introduction**

In the next lab, we're going to put your delta skills to a test. You're going to practice leaning and repairing data by moving it from raw to a gold level table. Then you're going to create a dashboard to present your analysis.

# **Delta Lab**

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 8, then click on "8.4 Lab: Delta lab" to open your notebook.  As you work through the exercises, note the bold questions. When prompted, answer the corresponding questions in the quiz that follows this reading prompt.