## Production Machine Learning Systems

Professional Machine Learning Engineer Certification Learning Path navigate_next Production Machine Learning Systems navigate_next Designing Adaptable ML Systems

# Introduction to TensorFlow Data Validation

2 hours Free

## Overview

This lab is an introduction to TensorFlow Data Validation (TFDV), a key component of TensorFlow Extended. This lab serves as a foundation for understanding the features of TFDV and how it can help you understand, validate, and monitor your data.

TFDV can be used for generating schemas and statistics about the distribution of every feature in the dataset. Such information is useful for comparing multiple datasets (e.g., training vs inference datasets) and reporting.

Statistical differences in the features distribution TFDV also offers visualization capabilities for comparing datasets based on the Google PAIR Facets project.

### Learning objectives

You learn how to:

- Review TFDV methods.
- Generate statistics.
- Visualize statistics.
- Infer a schema.
- Update a schema.

## Setup

For each lab, you get a new Google Cloud project and set of resources for a fixed time at no cost.

1. Sign in to Qwiklabs using an **incognito window**.
2. Note the lab's access time (for example, `1:15:00`), and make sure you can finish within that time.
   There is no pause feature. You can restart if needed, but you have to start at the beginning.
3. When ready, click **Start lab**.
4. Note your lab credentials (**Username** and **Password**). You will use them to sign in to the Google Cloud Console.
5. Click **Open Google Console**.
6. Click **Use another account** and copy/paste credentials for **this** lab into the prompts.
   If you use other credentials, you'll receive errors or **incur charges**.
7. Accept the terms and skip the recovery resource page.

**Note:** Do not click **End Lab** unless you have finished the lab or want to restart it. This clears your work and removes the project.

# Task 1. Set up your environment

### Enable the Recommended APIs

1. In the Google Cloud Console, on the **Navigation menu**, click **Vertex AI**.
2. Click **Enable All Recommended API**.

# Task 2. Launch Vertex AI notebooks

1. In the Google Cloud Console, on the **Navigation Menu**, click **Vertex AI > Workbench**. Select **User-Managed Notebooks**.
2. On the Notebook instances page, click **Create Notebook**.
3. In the **Create instance** dialog, confirm the name of the deep learning VM, if you don't want to change the region and zone, leave it to default.
4. Click **Environment**, and select **TensorFlow Enterprise 2.6 (with LTS and Intel MKL-DNN/MKL)** from the dropdown.
5. Click **Machine type**, and select **E2 standard > e2-standard-2**.
6. Leave all settings as they are and then click **Create**. The new VM will take 2-3 minutes to start.
7. Click **Open JupyterLab**.
   A JupyterLab window will open in a new tab.

# Task 3. Clone course repo within your Vertex AI notebooks instance

To clone the training-data-analyst notebook in your JupyterLab instance:

1. In JupyterLab, to open a new terminal, click the **Terminal** icon.
2. At the command-line prompt, run the following command:

   ```
   git clone https://github.com/GoogleCloudPlatform/training-data-analyst
   ```

3. To confirm that you have cloned the repository, double-click on the training-data-analyst directory and ensure that you can see its contents.
   The files for all the Jupyter notebook-based labs throughout this course are available in this directory.

# Task 4. Introduction to TensorFlow data validation

1. In the notebook interface, navigate to **training-data-analyst > courses > machine_learning > deepdive2 > production_ml > labs**, and open **tfdv_basic_spending.ipynb**.
2. In the notebook interface, click **Edit > Clear All Outputs**.

Carefully read through the notebook instructions and fill in lines marked with #TODO where you need to complete the code.

Tip: To run the current cell, click the cell and press SHIFT+ENTER. Other cell commands are listed in the notebook UI under **Run**.

- Hints may also be provided for the tasks to guide you along. Highlight the text to read the hints (they are in white text).

- If you need more help, look at the complete solution at **training-data-analyst > courses > machine_learning > deepdive2 > production_ml > solutions** and open **tfdv_basic_spending.ipynb**.

# End your lab

When you have completed your lab, click **End Lab**. Qwiklabs removes the resources you've used and cleans the account for you.

You will be given an opportunity to rate the lab experience. Select the applicable number of stars, type a comment, and then click **Submit**.

The number of stars indicates the following:

- 1 star = Very dissatisfied
- 2 stars = Dissatisfied
- 3 stars = Neutral
- 4 stars = Satisfied
- 5 stars = Very satisfied

You can close the dialog box if you don't want to provide feedback.

For feedback, suggestions, or corrections, please use the **Support** tab.