www.cloudskillsboost.google /games/4111/labs/26484

Create a Streaming Data Lake on Cloud Storage: Challenge Lab | Google Cloud Skills Boost

Qwiklabs : 6-7 minutes

ARC110

Overview

In a challenge lab you're given a scenario and a set of tasks. Instead of following step-by-step instructions, you will use the skills learned from the labs in the quest to figure out how to complete the tasks on your own! An automated scoring system (shown on this page) will provide feedback on whether you have completed your tasks correctly.

When you take a challenge lab, you will not be taught new Google Cloud concepts. You are expected to extend your learned skills, like changing default values and reading and researching error messages to fix your own mistakes.

To score 100% you must successfully complete all tasks within the time period!

Setup

Before you click the Start Lab button

Read these instructions. Labs are timed and you cannot pause them. The timer, which starts when you click **Start Lab**, shows how long Google Cloud resources will be made available to you.

This hands-on lab lets you do the lab activities yourself in a real cloud environment, not in a simulation or demo environment. It does so by giving you new, temporary credentials that you use to sign in and access Google Cloud for the duration of the lab.

To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).

Note: Use an Incognito or private browser window to run this lab. This prevents any conflicts between your personal account and the Student account, which may cause extra charges incurred to your personal account.

- Time to complete the lab---remember, once you start, you cannot pause a lab.

Note: If you already have your own personal Google Cloud account or project, do not use it for this lab to avoid extra charges to your account.

Challenge scenario

You are just starting your junior data engineer role. So far you have been helping teams create and manage data using Pub/Sub, Dataflow, and Cloud Storage.

You are expected to have the skills and knowledge for these tasks.

Your challenge

You are asked to help a newly formed development team with some of their initial work on a live messages streaming project. You have been asked to assist the team with a simulation of streaming live messages into Cloud Storage using Pub/Sub and Dataflow; you receive the following request to complete the following tasks:

- Use the command line to create up a Pub/Sub topic.
- Use the command line to create a Cloud Scheduler job to publish messages to Pub/Sub on a regular interval.
- Use the command line to create a Cloud Storage bucket as the output destination for a Dataflow job.
- Use the command line to create and run a Dataflow job to stream data from a Pub/Sub topic to a Cloud Storage bucket, then check the output files in Cloud Storage bucket.

Some standards you should follow:

- Ensure that any needed APIs (such as Dataflow) are successfully enabled.
- Create all resources in the region, unless otherwise directed.
- Complete the challenge lab in cloud shell instead of console, unless otherwise directed.

Each task is described in detail below, good luck!

Task 1. Create a Pub/Sub topic

- Use the command line to create a Pub/Sub topic called .

Click *Check my progress* to verify the objective. Create a Pub/Sub topic

Task 2. Create a Cloud Scheduler job

1. Use the command line to create an App Engine app for your project.
2. Use the command line to create a Cloud Scheduler job in this project to publish messages at one-minute intervals to the Pub/Sub topic in task 1. Message body: .
3. Use the command line to start the scheduler job.

Click *Check my progress* to verify the objective. Create a Cloud Scheduler job

Task 3. Create a Cloud Storage bucket

- Use the command line to create a Cloud Storage bucket with the following bucket name:

Click *Check my progress* to verify the objective. Create a Cloud Storage bucket

Task 4. Run a Dataflow pipeline to stream data from a Pub/Sub topic to Cloud Storage

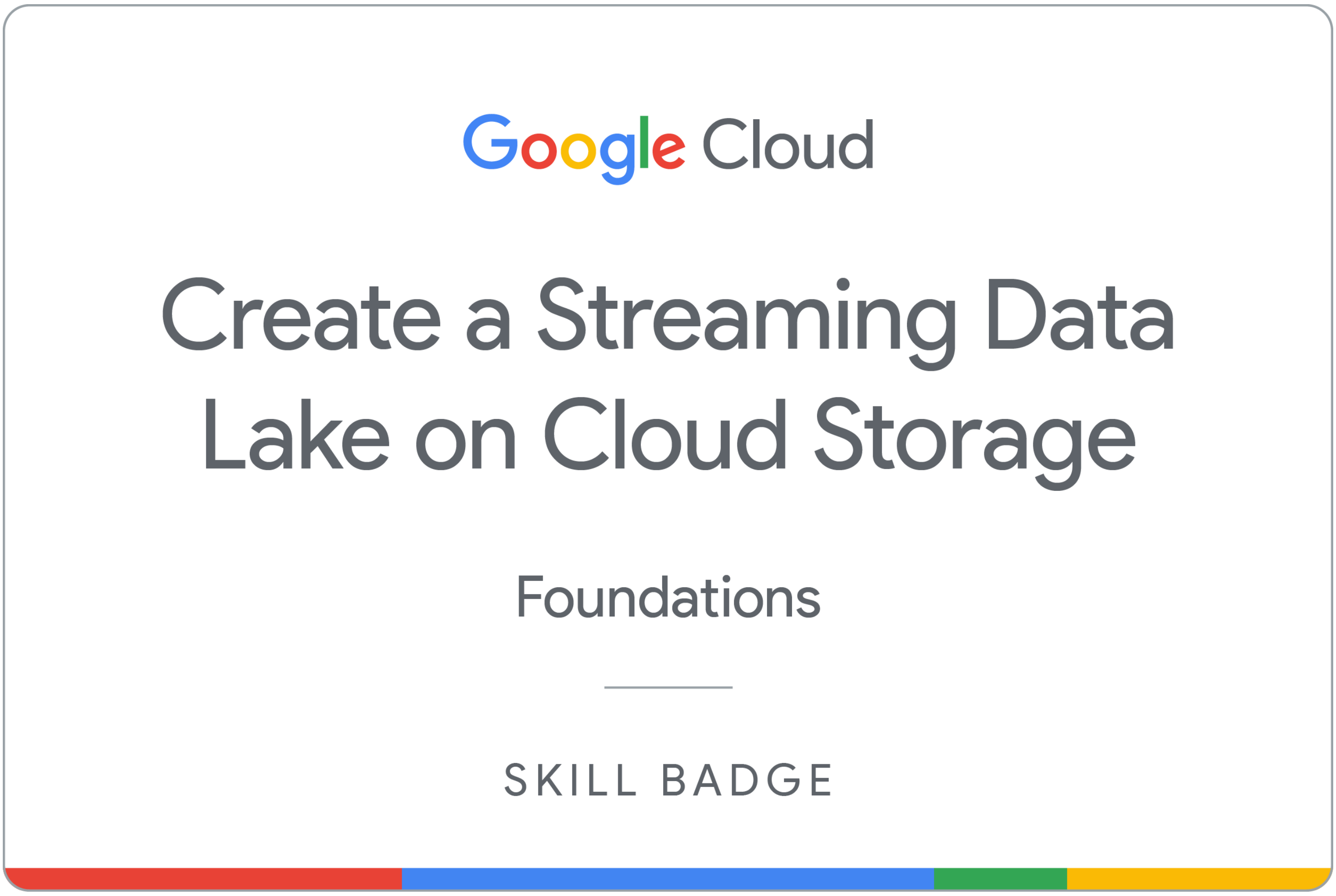
1. Use the command line to create and run a Dataflow job to stream data from a Pub/Sub topic to a Cloud Storage bucket.

- Use Java or Python script as your choice. Sample code available on GitHub pages: java-docs-samples, python-docs-samples.
- Use the Pub/Sub topic that you created in a task 1.
- Use the Cloud Storage bucket that you created in task 3 as the output location.
- Group messages based on a fixed time window of 2 minutes.

2. Use the command line to check which files have been written out in Cloud Storage.

Click *Check my progress* to verify the objective. Run a Dataflow pipeline to stream data from a Pub/Sub topic to Cloud Storage

Congratulations!



Earn your next skill badge

This self-paced lab is part of the Create a Streaming Data Lake on Cloud Storage skill badge quest. Completing this skill badge quest earns you the badge above, to recognize your achievement. Share your badge on your resume and social platforms, and announce your accomplishment using #GoogleCloudBadge.

Google Cloud training and certification

...helps you make the most of Google Cloud technologies. Our classes include technical skills and best practices to help you get up to speed quickly and continue your learning journey. We offer fundamental to advanced level training, with on-demand, live, and virtual options to suit your busy schedule. Certifications help you validate and prove your skill and expertise in Google Cloud technologies.

Manual Last Updated May 05, 2023

Lab Last Tested May 05, 2023