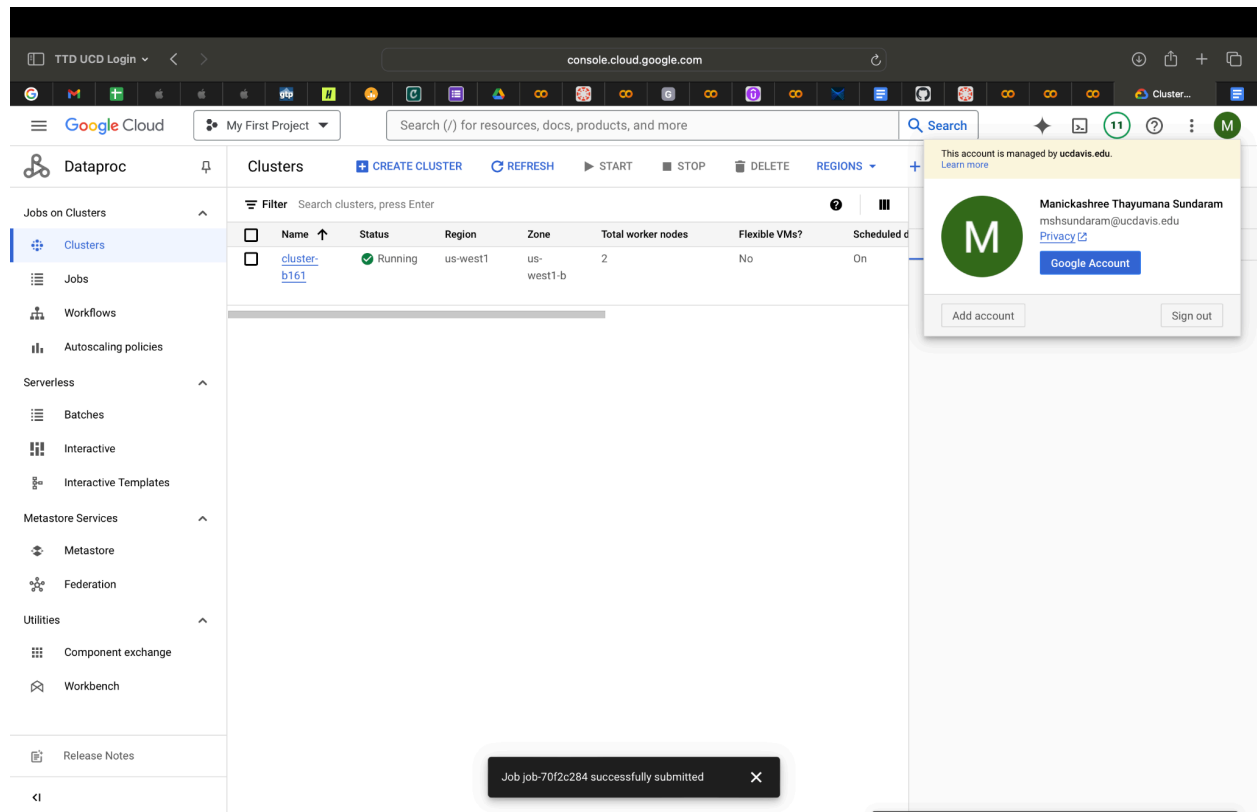


# Homework 2: Getting Started with Spark and Climate Change

## Cluster in GCP Steps



The screenshot displays the Google Cloud Platform (GCP) console interface. The left sidebar shows the navigation menu with categories like 'Jobs on Clusters', 'Serverless', 'Metastore Services', and 'Utilities'. The main content area is titled 'Clusters' and features a table of active clusters. A single cluster, 'cluster-b161', is shown with a 'Running' status. Below the table, a dark notification box indicates that a job has been successfully submitted. On the right, a user profile overlay for 'Manickashree Thayumana Sundaram' is visible, showing the user's email and a 'Sign out' button.

Name	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled d
cluster-b161	Running	us-west1	us-west1-b	2	No	On

Job job-70f2c284 successfully submitted

## Steps

1. Setting up your Google Cloud account.
2. Define necessary IAM roles to users

The screenshot shows the Google Cloud IAM & Admin console for project "My First Project". The left sidebar lists various IAM & Admin tools. The main content area displays the "Permissions for project 'My First Project'" page, which includes tabs for "VIEW BY PRINCIPALS" and "VIEW BY ROLES". A table lists the current permissions, with a filter bar at the top. A user profile overlay is visible in the top right corner, showing the user's name, email, and a "Sign out" button.

Type	Principal	Name	Role	Security insights
	694204297719-compute@developer.gserviceaccount.com	Compute Engine default service account	Compute Engine Service Agent Compute Storage Admin Connect Gateway Admin Dataplex Storage Data Owner Dataproc Administrator Dataproc Hub Agent Dataproc Metastore Admin Dataproc Service Agent Dataproc Worker Editor Environment and Storage Object Administrator Storage Admin Storage Object Admin Storage Object Creator	

### 3. Enabling Google Cloud Dataproc API(Cloud Dataproc API, Dataproc Metastore API, Cloud Resource Manager API)

The screenshot shows the Google Cloud console page for the "Cloud Dataproc API". The page includes a "MANAGE" button, a "TRY THIS API" button, and a status indicator showing "API Enabled". The "Overview" section provides details about the API, including its type, last product update, category, and service name. The "Additional details" section lists the API's type, last product update, category, and service name.

**Cloud Dataproc API**  
 Google Enterprise API  
 Manages Hadoop-based clusters and jobs on Google Cloud Platform.

**MANAGE** **TRY THIS API** **API Enabled**

**Overview**  
 Manages Hadoop-based clusters and jobs on Google Cloud Platform.

**Additional details**  
 Type: [SaaS & APIs](#)  
 Last product update: 7/21/22  
 Category: [Google Enterprise APIs](#)  
 Service name: dataproc.googleapis.com

**Tutorials and documentation**  
[Learn more](#)

Free trial status: \$300.00 credit and 91 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Google Cloud My First Project

Product details

### Dataproc Metastore API

Google Enterprise API

Fully managed, OSS-native technical metadata management.

MANAGE TRY THIS API [API Enabled](#)

OVERVIEW PRICING DOCUMENTATION RELATED PRODUCTS

#### Overview

Dataproc Metastore is a fully managed, highly available, auto-scaled, auto-healing, OSS-native metastore service that greatly simplifies technical metadata management. Dataproc Metastore service is based on Apache Hive metastore and serves as a critical component towards enterprise data lakes.

[Learn more](#)

#### Additional details

Type: [SaaS & APIs](#)  
Last product update: 4/29/22  
Category: [Big data](#), [Google Enterprise APIs](#)  
Service name: metastore.googleapis.com

#### Pricing

Product	Price
Dataproc Metastore Service Enterprise Unit	USD 3.42

Free trial status: \$300.00 credit and 91 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Google Cloud My First Project

Product details

### Cloud Healthcare API

Google Enterprise API

Store and access healthcare data on Google Cloud Platform.

MANAGE TRY THIS API [API Enabled](#)

OVERVIEW PRICING DOCUMENTATION RELATED PRODUCTS

#### Overview

The Cloud Healthcare API bridges the gap between care systems and applications built on Google Cloud. By supporting standards-driven data formats and protocols of existing healthcare technologies, the Cloud Healthcare API connects your data to advanced Google Cloud capabilities, including data processing with Cloud Dataproc, scalable analytics with BigQuery, and machine learning with Cloud ML Engine, while also simplifying application development and device integration. The Cloud Healthcare API accelerates digital transformation for organizations with existing clinical systems and enables new entrants to easily integrate with care networks.

[Learn more](#)

#### Additional details

Type: [SaaS & APIs](#)  
Last product update: 4/29/22  
Category: [Healthcare](#), [Big data](#), [Google Enterprise APIs](#)  
Service name: healthcare.googleapis.com

#### Pricing

4. Create a Dataproc Cluster by providing the necessary information for your cluster such as name, region and zone. Proceed to deploy the cluster created.
5. Create a bucket and upload all the required files(CO2 emissions per capita per country.csv and GlobalLandTemperatures\_GlobalLandTemperaturesByCountry.csv)

6. Upload the PySpark script to Google Cloud Storage bucket and copy the URI of it.
7. Now in the cluster details, select 'Submit job' and choose PySpark job type. Then click on submit to run the job.
8. Monitor the job, in case of failure, rectify the error and re-run the job by uploading the latest file.
9. Finally after running the job, clean up the resource.

**Reflect on your everyday life activities. What can you personally do to make a positive change for the environment? Write a short paragraph with your thoughts.**

There are many practical ways in which I contribute positively to the environment in my daily life. The first and most important step is to reduce waste. To achieve this, I start by minimizing the use of single-use plastics and opting for reusable alternatives such as water bottles, shopping bags, and food containers. This can significantly reduce the amount of waste that ends up in landfills.

Recycling and composting can also play an essential role in decreasing the amount of waste that ends up in landfills. By separating my trash into recyclables, compostables, and regular waste, I can ensure that items are disposed of in the most environmentally friendly way possible.

Conserving energy is another vital area that I can focus on. By using energy-efficient appliances, I can save money on my electricity bills and reduce my carbon footprint. Turning off lights and electronics when not in use, maximizing natural light, and choosing public transportation, carpooling, or biking instead of driving can also help reduce greenhouse gas emissions.

Lastly, I can enhance energy conservation in my home by switching to energy-efficient appliances and light bulbs and being mindful of turning off lights and electronics when not in use. By integrating these practices into my daily routine, I can help foster a more sustainable and environmentally friendly world for myself and future generations.

**Ideate 3 ideas about new sources of information that can underpin new companies, and type a short paragraph describing your ideas (bullet points are accepted)**

### **1. Language Learning Progression Analytics Platform:**

The concept is to create a platform that uses detailed user data from language learning apps to provide analytics on user progression, retention rates, and learning styles. The data source would be anonymized user data, similar to the Duolingo dataset, focusing on metrics such as recall accuracy, session activity, and learning preferences. The platform could serve educational institutions and language learning app developers by offering insights into effective curriculum development and personalized learning pathways.

### **2. Adaptive Learning Content Generator:**

**Idea:** To develop a service that creates personalized learning material by analyzing user performance data from different educational applications.

**Data Source:** The service will use datasets containing user performance metrics, learning speeds, and content interaction details to tailor educational content dynamically.

**Service Offered:** This service will be beneficial for online educational platforms seeking to optimize their content for better engagement and learning outcomes. It will adapt in real-time to cater to the user's needs.

### **3. Multilingual User Experience Optimization Tool:**

**Concept:** Create a tool that assists app developers in comprehending how diverse languages and cultural contexts influence user interaction and app usability.

**Data Source:** The data for the tool would be collected from apps like Duolingo, by analyzing user interface language preferences, performance metrics across various linguistic backgrounds, and user feedback.

**Service Offered:** The tool would be ideal for developers aiming to localize their apps effectively across multiple regions. It would ensure optimal user experience by analyzing and predicting language-based usage patterns.