Updated 03/05/2024

Sign up

DataProc training

2 days (14 hours)

Presentation

Our DataProc training course will enable you to perform complex data manipulation for batch processing, query issuing, streaming and machine learning. DataProc is a managed Hadoop and Spark service that lets you create data clusters extremely quickly and manage them cost-effectively.

Our program will teach you data management on GCP, including big data concepts and the solutions available on Google Cloud Platform. You'll be able to use the Dataproc Cloud Dashboard to create projects.

This course will also teach you how to create and manage data clusters. Infrastructure as Code (IaC) concepts and use with Terraform will also be covered.

Like all our training courses, it will be run on the latest version of the tool: Dataproc 2.2.

Objectives

- Understanding data management with DataProc
- Use the tool to create and manage clusters
- Using DataProc in an Infrastructure as Code context

Target audience

- Data Scientists
- AI Engineers

Prerequisites

- Knowledge of Google Cloud Platform
- Knowledge of Terraform

OUR DATAPROC TRAINING PROGRAM

INTRODUCTION TO DATAPROC AND BIG DATA

- Introduction to big data concepts and solutions
- Definition of Cloud Dataproc and its advantages over traditional solutions
- Dataproc access methods
- Dashboard navigation
- Project creation and resource management in Dataproc

CLUSTER CREATION AND MANAGEMENT

- Detailed Dataproc cluster creation process
- Cluster removal and lifecycle management
- Master and worker node roles
- Cluster customization with preemptible machine types and workers
- Identity and access management, permissions and roles

INTEGRATION WITH OTHER GCP SERVICES

- Using BigQuery for interactive data analysis
- Data storage and management with Cloud Storage
- Integration with Cloud SQL and Firestore database services
- Automate workflows with Dataproc workflow templates

DATA PROCESSING

- Batch and real-time processing (streaming)
- Writing, submitting and managing Hadoop and Spark jobs
- Job and cluster monitoring, logging and debugging
- Autoscaling and cluster performance optimization

INFRASTRUCTURE AS CODE WITH TERRAFORM ON GCP

- Introduction to Infrastructure as Code and the benefits of Terraform
- Declarative management of GCP infrastructure with Terraform
- Writing and organizing Terraform code for Dataproc resources
- Best practices for modularity and reusability of Terraform code

ADVANCED FEATURES AND OPTIMIZATION

- Customize clusters with initialization actions and personalized images
- Performance optimization techniques for data processing
- High availability and fault tolerance strategies for clusters
- Tips for optimizing cluster costs over the long term

Companies concerned

This training course is aimed at both individuals and companies, large or small, wishing to train their teams in a new advanced computer technology, or to acquire specific business knowledge or modern methods.

Positioning on entry to training

Positioning at the start of training complies with Qualiopi quality criteria. As soon as registration is finalized, the learner receives a self-assessment questionnaire which enables us to assess his or her estimated level of proficiency in different types of technology, as well as his or her expectations and personal objectives for the training to come, within the limits imposed by the selected format. This questionnaire also enables us to anticipate any connection or security difficulties within the company (intra-company or virtual classroom) which could be problematic for the follow-up and smooth running of the training session.

Teaching methods

Practical course: 60% Practical, 40% Theory. Training material distributed in digital format to all participants.

Organization

The course alternates theoretical input from the trainer, supported by examples, with brainstorming sessions and group work.

Validation

At the end of the session, a multiple-choice questionnaire verifies the correct acquisition of skills.

Sanction

A certificate will be issued to each trainee who completes the course.