Data Engineering on Google Cloud

Varighed: 4 Days   Kursus Kode: GO5975   Leveringsmetode: Virtuel deltagelse

Beskrivelse:

This four-day instructor-led class gives participants a hands-on introduction to designing and building data processing systems on the Google Cloud. Through a combination of presentations, demonstrations, and hands-on labs, participants will learn how to design data processing systems, build end-to-end data pipelines, analyze data, and perform machine learning. The course covers structured, unstructured and streaming data.

Målgruppe:

This class is intended for experienced developers who are responsible for managing large data transformations, including Data extraction, loading, transformation, cleaning and validation
Design of pipelines and architectures for data processing
Creation and maintenance of statistical and machine learning models
Consultation of data sets, display of consultation results and creation of reports

Agenda:

■ This course teaches participants the following skills:

■ Design and build data processing systems on the Google Cloud

■ Process batch and streaming data by implementing auto-scaling data channels in the cloud data stream.

■ Gain business insight from large datasets with Google BigQuery

■ Train, assess and predict using self-learning models using Tensorflow and Cloud ML

■ Leverage unstructured data via Spark and ML APIs in Cloud Dataproc

■ Allow a snapshot from data transmission

Forudsætninger:

To get the most out of this course, participants should have

■ Completed Google Cloud Basics: Great Machine and Data Learning course OR have equivalent experience
■ Basic knowledge of the most common query language, such as SQL
■ Experience in data modeling, extraction, transformation, loading activities
■ Application development using a common programming language such as Python
Familiarity with machine learning and/or statistics

Indhold:

Leveraging unstructured data with Cloud Dataproc on the Google Cloud.

Module 1: Overview of Google Cloud Dataproc

Creation and management of clusters.

Use of customized machine types and preferential work nodes.

Cluster scaling and deletion.

Lab: Creating Hadoop clusters with Google Cloud Dataproc.

Module 2: Running Dataproc jobs

Doing Pig and Hive jobs.

Separating storage and computing.

Lab: Running Hadoop and Spark Jobs with Dataproc.

Lab: Submitting and supervising jobs.

Module 3: Dataproc integration with Google Cloud

Customize the cluster with initialization actions.

BigQuery support.

Lab: Taking advantage of Google Cloud Platform services.

Module 4: Making sense of unstructured data with Google's self-learning APIs

Google Machine Learning APIs.

Common cases of LD use.

Invoking ML APIs.

Lab: Complex queries.

Performance and price.

Module 6: Data pipelines without server and with automatic scaling with Dataflow

The beam programming model.

Data piping in Beam Python.

Data piping in Beam Java.

Lab: Writing a Dataflow pipe.

Scalable processing of Big Data with Beam.

Lab: MapReduce in Dataflow.

Adding additional data.

Lab: Lateral inputs.

Flow data management.

GCP reference architecture.

Serverless machine learning with TensorFlow on the Google cloud platform

Module 7: Introduction to Machine Learning

What is machine learning (ML).

Effective LD: concepts, types.

LD data sets: generalization.

Lab: Exploring and creating ML datasets

Module 8: Building ML models with Tensorflow

Lab: Run an ML model locally and in the cloud.

Module 10: Feature engineering

Creating good features.

Transforming the inputs.

Synthetic characteristics.

Preprocessing with Cloud ML.

Laboratory: Characteristics engineering.

Creating resilient streaming systems on the Google Cloud Platform

Module 11: Architecture of stream analysis pipes

Real-time Data Processing: Challenges.

Handling variable data volumes.

Handling of unsorted or late data.

Lab: Streamline design.

Module 12: Ingesting Variable Volumes

What is Cloud Pub/Sub?

How it works: Themes and subscriptions.

Lab: Simulator.

Module 13: Implementing streaming pipes

Challenges in processing flows.

Late data management: watermarks, triggers, buildup.

Lab: Adding Machine Learning Capabilities to Large Data Analysis.

Serverless data analysis with Google BigQuery and Cloud Dataflo

Module 5: Serverless Data Analysis with BigQuery

What is BigQuery.

Queries and functions.

Lab: Writing queries in BigQuery.

Loading data into BigQuery.

Exporting data from BigQuery.

Lab: Loading and exporting data.

Nested and repeated fields.

Querying multiple tables.

Introduction to TensorFlow.

Lab: Using tf.learn.

TensorFlow graphs and loops + laboratory.

Lab: Using low-level TensorFlow + early stop.

Monitoring LD training.

Lab: TensorFlow training charts and graphs.

Module 9: Scaling ML models with CloudML

Why Cloud ML?

Packing a TensorFlow model.

Training from start to finish.

Lab: Streaming data processing pipeline for real-time traffic data.

Module 14: Streaming analysis and dashboards

Streaming analysis: from data to decisions.

Streaming data consultation with BigQuery.

What is Google Data Studio?

Lab: Creating a real-time control panel to visualize the processed data.

Module 15: High performance and low latency with Bigtable

What is Cloud Spanner?

Design of the Bigtable scheme.

Ingesting in Bigtable.

■ Lab: streaming to Bigtable.

Flere Informationer:

For yderligere informationer eller booking af kursus, kontakt os på tlf.nr.: 44 88 18 00

training@globalknowledge.dk

www.globalknowledge.com/da-dk/

Global Knowledge, Stamholmen 110, 2650 Hvidovre