How I passed the Google Professional Data Engineer Exam in 2020

In 8 days. Quick learner's guide for those who don't have time to read the manuals. August 2020.

Mike Shakhomirov
Aug 26 · 23 min read ★

Google Cloud Certified Professional Data Engineer

Want to get this certification? Well it is not an easy one. You'll need to do the homework. From what I read online people usually spend 2–3 month on preparation.

It's not a secret that many of us won't be using each of the Google products every day but we need to know them, right? This article is for those who don't have time to read all the manuals. I will describe what I did to get ready for this exam in 8 days.

First of all I need to say that I didn't have a clue how serious that exam really is. Exam questions were way more complex and different from any online course questions I know. So if you don't have any developer background please take your time, read the books and do the tutorials.

It took 1 hour and 35 minutes to pass the test. Every exam question was exactly the same though.

"Why did I choose to do that to myself?"

After the 3rd question I had a strong feeling that I know nothing. I was really scared as I told everyone that I was going to take this exam. A bit of advice — don't put yourself under that pressure. Don't tell your boss or your girlfriend. If you fail you will be able to take this exam again in two weeks. I think I passed only because I am a lucky guy and was wearing my lucky T-shirt that day. After I submit my final answers

I saw exam result straight away. It was a 'Pass'. My certificate came next week as well as a promo code for a hoodie.

Recommended read

There is a **May 2020** book, an Official Google Cloud Certified Professional Data Engineer Study Guide by Willey. It's $40 and it has a table of contents which gives a basic understanding of types of questions on the exam. To buy it or not it's up to you. You might want to just familiarise yourself with the contents and probably that will be enough.

Preparation

Day 1

I started with practice tests. There is a plenty of online courses for this task. I gave an overview below on some of them that helped me to get ready. Long story short, if you don't want to waste your time start with the tests.

There is a practice exam from Google which I did and failed but then I knew how questions looked exactly. It gives you the **format, level, and scope** of questions you may encounter on the certification exam.

Practice exam — Google Professional Data Engineer

Then there is a Linux Academy practice exam.

Recently they did have a migration of some courses from A Cloud Guru, which would include the Google Professional Data Engineer course! This course itself would be newer than the course that they had within their platform, but you are still able to view their Data Engineer course here Google Cloud Certified Professional Data Engineer (LA). Which was updated to the latest July 2019 exam objectives.

The first one is free! So go unlock the challenge. For the second one you will have two free attempts to practice after the registration.

And they have a handy course book (free): *Google Cloud Professional Data Engineer Exam Handbook by Linux Academy*: summary of the main concepts in scope.

Day 2

On day 2 I started to shape an idea of how to deal with case studies, what exam structure and questions are. I started to pay attention to words like economically, cost-effective, as soon as possible, etc. These types of keywords very often define the right answer because on exam you can find multiple answers that technically satisfy the requirments.

Day 3–5

I did 2 practice exams a day occasionally reading google docs related to topics I didn't know. I did that during my morning cardio in the gym while I was cycling. 30–40 minutes is more than enough to do the practice exam. Also there is a research suggesting that moderate aerobic activity improves cognitive function. I found it very useful and cardio didn't sound painful anymore. I was learning.

Day 6–8

I did two practice exams a day but now I had two browser tabs opened with previously passed practice exams. Every question I was uncertain about I checked straight away and read the docs. I think this tactics helped to polish my knowledge. Also I started to take some product specific notes and tie them to those keywords I was talking about earlier.

By day 7 I was getting at least 90% pass on practice.

Prcatice Google Certified Data Engineer exam

How to pass the exam?

There is no generic answer to that question. During the real exam I felt that I know nothing and questions seemed very difficult. However, the following strategy worked for me:

1. Do the practice tests to understand the type of questions and structure.
2. Learn product features
3. Pay attention to question keywords as very often they define the correct answer.

> Put on your lucky T-Shirt or whatever lucky thing you have. You'll need this.

4. Read the manual. It's **optional** but very useful.

Read official Google docs. At least an overview and case studies. These guides are great and have all the information you need to pass the exam. I was checking a topic from a Professional Data Engineer Study Guide by Willey and then searched for that topic in Google docs.

> Most of the exam questions were case studies, how to fix something, design a process, best practice or about machine learning.

Real exam is very machine learning heavy.

Typical questions.

Ensuring solution quality.

Case studies and Best practice questions. There will be a lot of them.

Example 1: *You are monitoring GCP Operations (formerly Stackdriver) metrics which show that your Bigtable instance's storage utilization is approaching 70% per node. What do you do?*

> ***Answer**: **Add additional nodes to the cluster to increase storage processing capacity.** Even though Cloud Bigtable table data is stored in Google Colossus, a cluster needs to be sized appropriately so that nodes have enough resources to process the total storage in use. When instance storage utilization reaches 70% per node, additional nodes should be added.*

Read: Quotas & limits | Cloud Bigtable Documentation | Google Cloud

Example 2: Your organization has just recently started using Google Cloud. Everyone in the company has access to all datasets in BigQuery, using it as they see fit without documenting their use cases. You need to implement a formal security policy, but need to first determine what everyone has been doing in BigQuery. What is your first step to do so?

> ***Answer**: Use **Stackdriver Logging** to review data access. Stackdriver Logging will record the audit logs of jobs and queries of each individual user's actions. Query slots won't work because they measure BigQuery performance and resource usage, but gives no visibility to individual user activity. You will not be able to view user activity via billing records. IAM policies are applied to datasets, but not individual tables inside each dataset. Furthermore, IAM policies show who has permissions to resources, but not their activity.*

Read: BigQuery documentation | Google Cloud

Example 3: Your security team have decided that your Dataproc cluster must be isolated from the public internet and not have any public IP addresses. How can you achieve this?

> ***Answer**: Using the — no-address flag will prevent public IPs from being assigned to nodes in a Cloud Dataproc cluster. However, Private Google Access is still required for the subnet to access certain GCP APIs.*

Read: Dataproc Cluster Network Configuration | Dataproc Documentation

Explore what Google recommends as best practice

- BigQuery: https://cloud.google.com/bigquery/docs/best-practices
- Stackdriver and Logging: https://cloud.google.com/products/operations
- BigTable: https://cloud.google.com/bigtable/docs/performance
- IAM and security: https://cloud.google.com/iam/docs/concepts
- Cloud Storage: https://cloud.google.com/storage/docs/best-practices

After all I would recommend to read overviews of all **database products** as there will be a lot of questions about them: https://cloud.google.com/products/databases

Designing data processing systems

Example: *A customer has a 400GB MySQL database running in a datacentre. What would be the best approach for migrating this database to GCP?*

> ***Answer****:* Create a Cloud SQL for MySQL 2nd generation instance and migrate the data. *For a MySQL database of this size, a Cloud SQL for MySQL instance would be the recommended approach. Using Compute Engine adds additional operational overhead. Postgres and Spanner would not be suitable migration hosts for a MySQL database.*

Recommended read: Migration from MySQL to Cloud SQL | Solutions | Google Cloud

Choosing Google Database products

Choosing Google Database products

> ***Example****: Your database is 500 GB in size. The data is semi-structured and does not need full atomicity. You need to process transactions in a point-of-sale application on Google Cloud Platform? You need to account for exponential user growth, but you do not want to deal with managing your infrastructure overhead?*
>
> *Use* ***Datastore***
>
> ***Example****: Data is more than 1 Tb and low latency required (also you probably don't care about costs):*
>
> *Use* ***BigTable***
>
> *Low latency* ***not required*** *and/or need to run* ***ANSI SQL*** *analytics and do it* ***economically****? Need to easily load data from* ***CSV and JSON*** *for later inspection with SQL?*
>
> *Use* ***BigQuery****. Cloud Datastore supports JSON and SQL-like queries but cannot easily ingest CSV files. Cloud SQL can read from CSV but not easily convert from JSON. Cloud Bigtable does not support SQL-like queries.*
>
> *You are designing a* ***relational*** *data repository on Google Cloud to grow as needed. The data will be* ***transactionally consistent*** *and added from* ***any location in the world****. You want to monitor and adjust node count for input traffic, which can spike unpredictably.*
>
> *Use* ***Cloud Spanner***
>
> *You need* ***strongly consistent transactions****? Data less than 500 Gb? The data* ***does not need to be streaming*** *or real-time?*
>
> *Use* ***Cloud SQL***

Pay attention to:

High Availability and Performances and things like **failover and read replicas**.

After all there are a lot of BigTable questions.

Pay attention to:

Development and Production instances, Disk Types (HDD vs. SSD).

BigTable Performance Example: Your organization will be deploying a new fleet of IoT devices, and writes to your Bigtable instance are expected to peak at 50,000 queries per second. You have optimized your row key design and need to design a cluster that can meet this demand. What do you do?

> ***Answer****: An optimized Bigtable instance with a well-designed row key schema can theoretically support up to 10,000 write queries per second per node, so 5 nodes are required.*

Read: Understanding Cloud Bigtable performance

BigTable Performance Example: You are asked to investigate a Bigtable instance that is performing poorly. Each row in the table represents a record from an IoT device and contains 128 different metrics in their own column, each metric containing a 32-bit integer. How could you modify the design to improve performance?

> ***Answer****: Large numbers of cells in a row can cause poor performance in Cloud Bigtable. When the data itself is so small, as in this scenario, it would be more efficient to simply retrieve all of the metrics from a single cell, and use delimiters inside the cell to separate the data. Row versioning would compound the problem by creating the most new entries along the least efficient dimension of the table, and HDD disks will always slow things down.*

Read: Understanding Cloud Bigtable performance

BigTable Performance Example: Your production Bigtable instance is currently using four nodes. Due to the increased size of your table, you need to add additional nodes to offer better performance. How should you accomplish this without the risk of data loss?

> ***Answer****: Edit instance details and increase the number of nodes. Save your changes. Data will re-distribute with no downtime.You can add/remove nodes to Bigtable with no downtime necessary.*

Read: Overview of Cloud Bigtable | Cloud Bigtable Documentation

BigTable Performance Example: You currently have a Bigtable instance you've been using for development running a development instance type, using HDDs for storage. You are ready to upgrade your development instance to a production instance for increased performance. You also want to upgrade your storage to SSDs as you need maximum performance for your instance. What should you do?

> ***Answer****: you cannot change the disk type on an existing Bigtable instance, you will need to export/import your Bigtable data into a new instance with the different storage type. You will need to export to Cloud Storage then back to Bigtable again.*

BigTable Performance Example: Your customer uses a Bigtable instance that contains 2 replicating clusters for regional disaster recovery. Table transactions from the application are required to be strongly consistent. How can you guarantee that for this configuration?

> ***Answer****: Determine one cluster as the master, and use an application profile that specifies single-cluster routing. By default, Cloud Bigtable is eventually consistent. To guarantee strong consistency you must limit queries to a single cluster in an instance by using an application profile.*

Read: Overview of Replication | Cloud Bigtable Documentation | Google Cloud

Read: Overview of Cloud Bigtable | Cloud Bigtable Documentation

BigTable Performance Example: What will happen to your data in a Bigtable instance if a node goes down?

> ***Answer****: Nothing, as the storage is separated from the node compute. Rebuilding from RAID is not a valid Bigtable function. Storage and compute are separate, so a node going down may affect performance, but not data integrity; nodes only store pointers to storage as metadata.*

Read: Overview of Cloud Bigtable | Cloud Bigtable Documentation

BigTable Performance Example: You are monitoring GCP Operations (formerly Stackdriver) metrics which show that your Bigtable instance's storage utilization is approaching 70% per node. What do you do?

> ***Answer****: Add additional nodes to the cluster to increase storage processing capacity. Even though Cloud Bigtable tablet data is stored in Google Colossus, a cluster needs to be sized appropriately so that nodes have enough resources to process the total storage in use. When instance storage utilization reaches 70% per node, additional nodes should be added.*

Read: Quotas & limits | Cloud Bigtable Documentation | Google Cloud

BigTable Performance Example: Which of these is NOT a valid reason to choose an HDD storage type over SSD in a Bigtable instance?

> ***Answer****: Bigtable can integrate with Cloud Storage regardless of the type of disk in use by the instance. The other reasons are valid for choosing HDD as an outlying case, but in general SSD disks are preferred as HDD disks will cause a significant drop in performance.*

Read: Overview of Cloud Bigtable | Cloud Bigtable Documentation

Relational Database questions

Pay attention to:

Replicas, availability and migration guides.

Example: You are designing a relational data repository on Google Cloud to grow as needed. The data will be transactionally consistent and added from any location in the world. You want to monitor and adjust node count for input traffic, which can spike unpredictably. What should you do?

> ***Answer****: Use Cloud Spanner for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.*

Example: Your customer is looking to move a 2TB MySQL database to GCP. Their business requires an uptime SLA exceeding 99.95%. How can you achieve this?

> ***Answer****: Migrate the database to a Cloud SQL for MySQL high availability configuration with a standby instance in a secondary zone. Cloud SQL's standard SLA is 99.95%. Uptime in excess of this can be achieved by using a high availability*

> *configuration with a failover instance in a secondary availability zone. Failover replicas are not a feature — read replicas are. Cold-spares are inefficient as they will not be automatically switched to, unlike the proper HA configuration. Compute Engine options are not required to achieve the required SLA.*

Read: Cloud SQL Service Level Agreement (SLA) | Cloud SQL Documentation
Overview of the high availability configuration | Cloud SQL for MySQL

A lot of questions about Pub/Sub, Kafka and windowing.

Pay attention to:

Kafka Mirroring, Differences between these two.

Pub/Sub

Pub/Sub handles the need to scale exponentially with traffic coming from around the globe. **Apache Kafka** will not be able to handle an exponential growth in users globally as well as Pub/Sub.

Cloud Pub/Sub guarantees to deliver messages at least once to every subscriber. As multiple systems need to be notified of every order, you should create one topic and use multiple subscribers. Order of delivery is not guaranteed by Pub/Sub so attach a timestamp in the publishing system if possible.

Read: https://cloud.google.com/pubsub/architecture

Example: Your company's Kafka server cluster has been unable to scale to the demands of their data ingest needs. Streaming data ingest comes from locations all around the world. How can they migrate this functionality to Google Cloud to be able to scale for future growth?

> ***Answer**: Create a single Pub/Sub topic. Configure endpoints to publish to the Pub/Sub topic, and configure Cloud Dataflow to subscribe to the same topic to process messages as they come in.*

Apache Kafka and Google Cloud Pub/Sub

Some of the contenders for Big Data messaging systems are Apache Kafka, Google Cloud Pub/Sub, and Amazon Kinesis (not…

www.jesse-anderson.com

Security, Encryption and key management

Example: Your organization has a security policy which mandates that the security department must own and manage encryption keys for data stored in Cloud Storage buckets. Analysts and developers need to store data that is encrypted with these keys without access to the keys themselves. How can this be achieved?

> ***Answer**: Use Cloud KMS for the security team to manage their own encryption keys in a dedicated project. Grant the Cloud KMS CryptoKey Encrypter/Decrypter role for the keys to the Cloud Storage service accounts in the other projects. Cloud KMS allows you to create and manage your own encryption keys which can then be used by service accounts in other projects. Developers in those projects can then access services without any access to the underlying keys. A staging area is not required, neither is any other manual intervention by the security team.*

Read: Using customer-managed encryption keys | Cloud Storage | Google Cloud

Comparison of encryption methods. Source: Coursera

Building and operationalizing data processing systems

Dataproc

Pay attention to:

HDFS vs. Google Cloud Storage for Dataproc workloads.

Best practice: Dataproc clusters better be job specific. Use cloud storage if you need scaling because HDFS won't scale well and needs custom settings. Also Google recommends using Cloud Storage instead of HDFS as it is much more cost effective especially when jobs aren't running.

Read: https://cloud.google.com/solutions/migration/hadoop/hadoop-gcp-migration-jobs

Dataflow

Pay attention to:

`PCollection` branching, Flatten and Joins, transformations and sliding windows.

- **Flatten** — You can use the `Flatten` transform in the Beam SDKs to merge multiple `PCollection`s of the **same type**.
- **Join** — You can use the `CoGroupByKey` transform in the Beam SDK to perform a relational join between two `PCollection`s. The `PCollection`s must be keyed (i.e. they must be collections of key/value pairs) and they must use the same key type.

Read: https://beam.apache.org/documentation/pipelines/design-your-pipeline/

Windowing:

Beam Programming Guide

Apache Beam is an open source, unified model and set of language-specific SDKs for defining and executing data…

beam.apache.org

Example: You are writing a streaming Cloud Dataflow pipeline that transforms user activity updates before writing them to a time-series database. While continually transforming each element as it arrives, you also need to depend on some additional data at run-time to create the transformation. How can you achieve this?

> ***Answer****: Side inputs are useful if your ParDo needs to inject additional data when processing each element in the input PCollection, but the additional data needs to be determined at runtime (and not hard-coded). A combine would not achieve the same outcome, and using an external shell script is unnecessary and inefficient.*

Read: Beam Programming Guide

Example: You need to design a pipeline that can ingest batch data from your organization's application metrics as well as your user database, then join the data using a common key before outputting to BigQuery. What is the most efficient way to go about this?

> ***Answer****: Create a Cloud Dataflow pipeline and join the two PCollections using CoGroupByKey transform on the common key. CoGroupByKey performs a relational join of two or more key/value PCollections that have the same key type; "common key" is the magic clue in this question.*

Read: Beam Programming Guide

Example : *You are setting up multiple MySQL databases on Compute Engine. You need to collect logs from your MySQL applications for audit purposes. How should you approach this?*

> ***Answer****: Install the Stackdriver Logging agent on your database instances and configure the fluentd plugin to read and export your MySQL logs into Stackdriver Logging. The* ***Stackdriver Logging*** *agent requires the fluentd plugin to be configured to read logs from your database application. Not Stackdriver Monitoring, not Cloud Composer. Cloud Composer is used for managing workflows, not logging. Stackdriver Monitoring is useful for measuring performance metrics and alerts, but not for logs.*

Read: About the Logging agent | Cloud Logging | Google Cloud

Example : You want to make changes to a Cloud Dataflow pipeline that is currently in production, which reads data from Cloud Storage and writes the output back to Cloud Storage. What is the easiest and safest way to test changes while in development?

> ***Answer****: Use a DirectRunner to test-run the pipeline using local compute power, and a staging storage bucket. Using a DirectRunner configuration with a staging storage bucket is the quickest and easiest way of testing a new pipeline, without risking changes to a pipeline that is currently in production.*

Read: Direct Runner

Example : You are asked to investigate a Bigtable instance that is performing poorly. Each row in the table represents a record from an IoT device and contains 128 different metrics in their own column, each metric containing a 32-bit integer. How could you modify the design to improve performance?

> ***Answer****: Store the metrics in a single column by using delimiters. Make sure the cluster is using SSD disks. Large numbers of cells in a row can cause poor performance in Cloud Bigtable. When the data itself is so small, as in this scenario, it would be more efficient to simply retrieve all of the metrics from a single cell, and use delimiters inside the cell to separate the data. Row versioning would compound the problem by creating the most new entries along the least efficient dimension of the table, and HDD disks will always slow things down.*

Read: Understanding Cloud Bigtable performance

Operationalizing machine learning models

I started with Google docs straight away as I was already familiar with ML basic concepts but I think official Google ML crash course is really useful for exam purposes. 100% I should have started with this one first. And it has a lot of reading content as well as videos.

Introduction to Machine Learning | Machine Learning Crash Course

This module introduces Machine Learning (ML). Estimated Time: 3 minutes Learning Objectives Recognize the practical…

developers.google.com

And then there is a quiz in each section so you can check your understanding:

Dynamic (Online) Inference

This exam is really machine learning heavy.

Example questions:

Example L1/L2 regularization : You are attempting to train a TensorFlow model but you are aware that some of your input features will have no significant impact on a prediction. What technique can you employ to discourage model complexity?

> ***Answer: L2 regularization*** *is more relevant when all features have relatively equal weights/influence, which is not the case here. Hyperparameters deal with learning rate, which is not relevant for this question.* ***L1 regularization*** *is able to reduce the weights of less important features to zero or near zero.*

Read: Regularization for Simplicity | Machine Learning Crash Course

Example: What is over/underfitting and how to fix it. You are training a facial detection machine learning model. Your model is suffering from overfitting your training data. What steps can you take to solve this problem?

> ***Simple answer:*** *To fix* ***underfitting*** *(for example, when the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set.) increase the complexity of your model (introduce an additional layer or increase the size of vocabularies).*
> *A model* ***overfits*** *when predicts training data well but performs poor on the validation set. To fix* ***overfitting: reduce the number of features*** *(****Regularization****),* ***add more data*** *to increase the variety of samples and better generalize your model****,****use* ***Dropout layers*** *or* ***reduce*** *the network's capacity by removing layers or reducing the number of elements in the hidden layers.*
> ***Increasing your regularization parameters*** *also allows you to reduce 'noise' in your model to reduce overfitting.*

Read: https://developers.google.com/machine-learning/crash-course/generalization/peril-of-overfitting and here: Machine learning workflow | AI Platform | Google Cloud

- **Synthetic features**

Simple answer:

> *A* ***feature*** *not present among the input features, but created from one or more of them. Kinds of synthetic features include: — **Bucketing** a continuous feature into range bins. -Multiplying (or dividing) one feature value by other feature value(s) or by itself. — Creating a* ***feature cross :*** *A* ***synthetic feature*** *formed by crossing (taking a Cartesian product of) individual binary features. Feature crosses help represent nonlinear relationships.*

Example Google Machine Learning APIs : You are developing an application that will only recognize and tag specific business to business product logos in images. You do not have an extensive background working with machine learning models, but need to get your application working. What is the current best method to accomplish this task?

> ***Answer****: Use the AutoML Vision service to train a custom model using the Vision API. Cloud Vision API can recognize common logos, but would struggle to find specific business logos on its own. The best option is AutoML, which allows you to take the pre-trained Vision API and apply it to custom images. Creating a custom ML model from scratch would be time-consuming and is not necessary when you can build on existing models.*

Read: Cloud Auto ML Vision

Example: You need to quickly add functionality to your application that will allow it to process uploaded user images, extract any text contained within them and perform sentiment analysis on the text. What should you do?

> ***Answer****: Call the Cloud Vision API for Optical Character Recognition (OCR) then call the Natural Language API for sentiment analysis. Cloud Vision API for OCR is*

> *the quickest way to extract from user uploaded images. The Natural Language API already has a built-in model for sentiment analysis.*

Read: Detect text in images | Cloud Vision API | Google Cloud Analyzing Sentiment | Cloud Natural Language API | Google Cloud

Example: You're developing a mobile application that allows a food processing business to detect fruit that has gone bad. Staff at warehouses will use mobile devices to take pictures of fruit to determine whether it should be discarded. Which GCP services could you use to accomplish this?

> ***Answer**: Train an AutoML Vision model using labeled images of fruit that has gone bad. Use AutoML Vision Edge in ML Kit to deploy the custom model to mobile devices using ML Kit client libraries. The AutoML Vision service is the quickest way to train a custom classification model using image data, and the Vision Edge in ML Kit will allow the model to be deployed to Android and IOS devices. Images must be labelled in order for the model to be trained. There is no need to use Kubernetes Engine or App Engine.*

Read: AutoML Vision API Tutorial | Cloud AutoML Vision | Google Cloud

Example: You have a large number of images that you wish to process through a custom AutoML Vision model. Time is not a factor, but cost is. Which approach should you take?

> ***Answer**: Make an asynchronous prediction request for the entire batch of images using the batchPredict method. Batch prediction often offers a lower cost per inference and higher throughput than synchronous (online) prediction. However, batch prediction produces a long-running operation (LRO), meaning that results are only available once the LRO has completed.*

Read: Making batch predictions | Cloud AutoML Vision | Google Cloud

ML products

This is a full list and very easy to find all docs. Just skim through Overview and use cases sections.

Vision AI use cases

Speech to text API best practices: https://cloud.google.com/speech-to-text/docs/best-practices

Cloud Vision API: https://cloud.google.com/vision/docs/labels

Source: Linux Academy

Natural Language API

Example question: You wish to build an AutoML Natural Language model for classifying some documents with user-defined labels. How can you ensure you are providing quality training data for the model?

> ***Answer**: Ensure you provide at minimum 10 training documents per label, but ideally 100 times more documents for the most common label than for the least common label. To achieve the best results when preparing training data for AutoML Natural Language classification models, the minimum number of documents per label is 10, and ideally you should have at least 100 times more documents for the most common label than for the least common label.*

Read: Preparing your training data | AutoML Natural Language | Google Cloud

- Google Cloud AI Platform
- Google Cloud TPUs
- Google Glossary of ML terms

Also I would recommend this guide for best practice:

Rules of Machine Learning: | ML Universal Guides | Google Developers

This document is intended to help those with a basic knowledge of machine learning get the benefit of Google's…

developers.google.com

Online practice exams

Google Certified Professional Data Engineer from Linux Academy

Rank: 4/5 - Probably the best one. Still won't cover all aspects you may face during exam.

Google Certified Professional Data Engineer from Linux Academy

Google Certified Professional Data Engineer Results

Coursera Google Cloud Professional Data Engineer course

Preparing for the Google Cloud Professional Data Engineer Exam

Offered by Google Cloud. From the course: "The best way to prepare for the exam is to be competent in the skills…

www.coursera.org

It has a good resources section with **pdfs** you could use for your exam preparation. It has 7 days free trial as well. Final practice exam includes 25 questions of which only 4 I found somewhat different from **Linux Academy**.

Coursera GCP Professional Data Engineer — Exam pdf and resources

Cloud Academy course

Latest update here was on **Jul 10 2020.** Unfortunately you can't take a practice exam for free here but they have a nice mobile app and lecture transcripts. So I did that one while I was cycling in the gym.

Data Engineer – Professional Certification Preparation for Google – Cloud Academy

This learning path is designed to help you prepare for the Google Certified Professional Data Engineer exam. Even if…

cloudacademy.com

Just click register for 7-Day Free trial.

Cloud Academy — GCP Data Engineer Exam (Not available in Free version)

Final tips

1. Decide if you really need this certification. Exam preparation is a huge commitment.
2. Don't tell anyone.
3. Learn more about machine learning. There will be a lot of them.
4. Pay attention to ML products features
5. Real exam questions are more complex then the ones you'll face during practice tests.

Recommended read:

A study guide by Ivam Luz : https://docs.google.com/spreadsheets/d/1LUtqhOEjUMySCfn3zj8Arhzcmazr3vrPzy7VzJwIshE/edit?usp=sharing&source=post_page-----bb6a0812a1b1----------

Hacker's Guide to Fixing Underfitting and Overfitting Models

TL;DR Learn how to handle underfitting and overfitting models using TensorFlow 2, Keras and scikit-learn. Understand…

www.curiousily.com

How I Passed the Google Cloud Professional Data Engineer Certification Exam

Without the recommended 3-years hands-on experience

towardsdatascience.com

ml874/Data-Engineering-on-GCP-Cheatsheet

This cheatsheet is currently a 9-page reference Data Engineering on the Google Cloud Platform. It covers the data…

github.com

Google Cloud Professional Data Engineer Certification — 2020 Mini-Guide

Making this as a mini-guide for people taking the exam in 2020. Can be difficult to find up to date guides on the…

medium.com

Machine Learning Crash Course | Google Developers

Learn and apply fundamental machine learning concepts with the Crash Course, get real-world experience with the…

developers.google.com

Official Google Cloud Certified Professional Data Engineer Study Guide

The proven Study Guide that prepares you for this new Google Cloud exam TheGoogle Cloud Certified Professional Data…

www.wiley.com

Taking Google Cloud Professional Data Engineer Certification in 2020

Google Cloud Professional Data Engineer Certification is a technical attestation that proves the ability to know and…

medium.com

How to pass the Google Cloud Professional Data Engineer exam

Without the recommended 3-year industry experience

towardsdatascience.com

Notes from my Google Cloud Professional Data Engineer Exam

Immediately after the exam I do a memory dump as notes. Hence it is also quite unordered. This is a sanitized list that…

medium.com

The Ultimate Hack to passing Google Cloud

So, why the Google Cloud Professional Data Engineer Certified exam?

medium.com

Google Cloud — Data Engineer Exam Study Guide

This is a 12-page exam Study Guide hopes to cover all keys points for GCP Data Engineer Certification Exam

medium.com

How to Prepare for Google Cloud Certified Professional Data Engineer Exam and Pass It On…

Thinking about new year's resolutions? How about adding Google Cloud certificate to your career portfolio? Data…

medium.com

Google Cloud Certification: Preparation and Prerequisites

Google Cloud Platform (GCP) has evolved from being a niche player to a serious competitor to Amazon Web Services and…

cloudacademy.com

Professional Data Engineer Practice Exam | Google Certified Professional

The Data Engineer practice exam will familiarize you with types of questions you may encounter on the certification…

cloud.google.com

Data Data Science Exam Certification Google Cloud Platform