GCP PROFESSIONAL DATA ENGINEERING LEARNING PATH

# PREPARING FOR PROFESSIONAL DATA ENGINEERING JOURNEY

## Introduction to PDE certification:

### Course Structure:

1. Introduction
2. Designing data processing system.
3. Ingesting and processing data
4. Storing data
5. Preparing and using data for analysis.
6. Maintaining and automating data workloads.
7. Your next steps.

### Role of PDE:

A Professional Data Engineer makes data usable and valuable for others by collecting, transforming, and publishing data.

This individual evaluates and selects products and services to meet business and regulatory requirements.

A Professional Data Engineer creates and manages robust data processing systems. This includes the ability to design, build, deploy, monitor, maintain, and secure data processing workloads.

In this course, you put yourself in the role of a Professional Data Engineer working for Cymbal Retail, a fictional company. Cymbal Retail is amongst the largest B2C retailers in the world serving customers in 52 countries and territories.

The founders started with brick and mortar stores and had the foresight to continuously invest in technologies over many decades. Cymbal Retail has seen significant changes in purchasing behavior, with accelerated growth in online purchases in recent years even as they continue to grow their brick and mortar presence.

Cymbal Retail has multiple data centers of their own across the world.

Cymbal Retail is in the process of transitioning these data centers to being entirely on Google Cloud, with a significant portion of their data and applications already moved.

When migrating and operating, as a Professional Data Engineer you will need to ensure that Cymbal Retail complies with all local regulations.

You will also need knowledge of security around data and the tools available for compliance.

Customer delight has been the guiding purpose for Cymbal Retail through generations.

As a Professional Data Engineer, you must be able to integrate products and solutions available within Google Cloud with technologies employed by the acquired company.

You have four key technology areas of focus that drive your short term technology strategy:

* Rapidity in the entire supply chain.
* Accuracy of analytics and predictions.
* Centralized control and security.
* controlled costs.

Analysts should have the tools to ask questions about the data.

It is one of your responsibilities as a Professional Data Engineer to make data available to all parts of the organization that need it.

Data should be available quickly and in easily consumable ways.

As you continue through this course, you will continue to explore the role of a Professional Data Engineer at Cymbal Retail.

We use this scenario to illustrate the types of considerations and tasks that correspond to each section of the exam guide.

### Certification value and benefits

**Why become a Google Cloud Certified Professional Data Engineer?**

Certification value has skyrocketed. Becoming Google certified gives you industry recognition.

It validates your technical expertise and can be the starting point to for the next phase of your career.

You may be curious about what differentiates a “professional” cloud certification from an “associate” level one.

The professional level certification expects the exam taker to know how to evaluate case studies and design solutions to meet business requirements—in addition to knowing about technical requirements—for customer implementation.

### Certification process

As you explore the role of the Professional Data Engineer at Cymbal Healthcare in this course, you also explore different sections of the exam guide which forms the basis for the certification exam.

In the following modules, you take diagnostic questions to assess your knowledge of each section of the exam guide. The exam guide for the Professional Data Engineer is divided into five sections, each containing one or more objectives. We focus on where you can find resources at the section objective level.

You can find the exam guide on the certification page: https://cloud.google.com/learn/certification/guides/data-engineer The certification page will always have the latest available version of the exam guide.

It's important to note that there are separate teams developing the Professional Data Engineer exam questions and developing the courses and exam preparation materials. The course developers and instructors don’t know what questions will be on your certification exam. The goal of the course is to determine what you know and what you don’t know to help you prepare a study plan and get ready for the Professional Data Engineer job role and exam.

Later, when you take the exam, you will demonstrate whether you have the skills and knowledge required to earn the certification.

Throughout this course, you are pointed to specific resources and documentation that can help you fill the gaps you identify through the diagnostic questions.

Let’s go over the types of resources you may want to include in your study plan.

Google provides resources to help you develop your skills and experience with Google Cloud products and services. The learning path for this certification includes online courses, online practice labs and quests, and practice questions.

The courses recommended for the Professional Data Engineer certification can be taken on demand or as instructor-led courses. Google Cloud Big Data and Machine Learning Fundamentals is available in both formats.

The instructor-led course Data Engineering on Google Cloud is equivalent to the following series of on demand courses:

Modernizing Data Lakes and Data Warehouses with Google Cloud.

Building Batch Data Pipelines on Google Cloud.

Building Resilient Streaming Analytics Systems on Google Cloud.

and Smart Analytics, Machine Learning, and AI on Google Cloud.

The instructor-led course Serverless Data Processing with Dataflow is available as an on demand 3-course series:

Serverless Data Processing with Dataflow: Foundations.

Serverless Data Processing with Dataflow: Develop Pipelines.

and Serverless Data Processing with Dataflow: Operations.

The skill badges provide hands-on experience working in Google Cloud. Skill badges are learning paths made up of labs that give you practice with Google Cloud services or solutions.

## Designing data processing system

Let’s start by discussing how a Professional Data Engineer performs this role at Cymbal Retail.

**Migrating data from private data centers to Google cloud for Cymbal Retail:**

Cymbal Retail’s existing private data centers are going to be decommissioned in the next two years. You need to help migrate existing data and data processing systems to Google Cloud. Similar migrations have to also be conducted in acquired companies with their own private data centers.

As a Professional Data Engineer, your role involves ensuring that there is a uniform approach to managing and securing the data. Controlling access at a granular level based on employee role is critical to complying with regional laws and industry regulations.

You need to understand and deploy options to control the security of data via encryption. Whether during migration or in the course of regular operations, certain data has to be captured and stored, but never revealed to other parts of the organization, for example, analysts.

You need to employ tools to redact information, and you need be able to extend these tools as required to meet future requirements.

You will be required to move data reliably from external sources into Google Cloud.

Sometimes, the data movement might be done over many days. In other cases, data needs to be quickly moved as and when it arrives in other data sources for up to date analytics. As data is moved into Google Cloud, it will also need to be assessed for quality and cataloged for easy discoverability by consumers of the data. Cymbal Retail’s data is used simultaneously by different divisions to manage their business—finance, marketing, purchasing, sales, compliance, legal, store operations, and more. However, all of these groups do not need access to all the data. You should ensure fine-grained control on the data which is strictly on a need-to-know basis.

## Ingesting and processing data

In this module you’ll explore considerations for ingesting and processing data, which corresponds to section 2 of the Professional Data Engineer Exam Guide.

Let’s start by discussing how a Professional Data Engineer performs this role at Cymbal Retail. Next, you’ll assess your skills in this section through 10 diagnostic questions. Then you’ll review these questions. Based on the areas you need to learn more about, you’ll identify resources to include in your study plan.

Let’s begin by exploring the role of a Professional Data Engineer in handling data ingestion and data processing for Cymbal Retail, which includes using the various services provided by Google Cloud to support data ingestion and data processing.

Cymbal Retail receives data from multiple sources, both internal and external. As the business has grown, the volume of ingested data has also increased exponentially. Processing the data has become increasingly complex and costly. In Cymbal Retail’s current on-premises data centers, Spark and Hadoop jobs are executed on pre-configured, static infrastructure. Part of your role involves determining how to lift-and-shift these jobs to Google Cloud. You have to design the architecture of the data ingestion and processing. Some data can be directly loaded into data warehouses using an extract and load approach, while others might be transformed before being uploaded into the data warehouse.

Building, deploying, and operating effective flexible data pipelines for all the stages of data processing is a primary expectation from you as a Professional Data Engineer. You need to identify and deploy the right approach between EL, ETL, or ELT and choose the right Google Cloud tools for the job. Cymbal Retail’s customers want features that require the increased use of real-time data. In this regard, tools like open source Apache Beam and the hosted Dataflow are important skills for a data professional.

Your knowledge of ways to apply different types of windowing for various use cases will provide the right approach to analyze streaming data. Your role also requires you to optimize all data ingestions and data processings tasks. Your optimizations should bring considerable savings on effort and cost, while improving availability and responsiveness. As the volume of data and scale of processing increases, Cymbal Retail does not want the latency, effort, or cost to increase linearly, or worse, exponentially. Your early design decisions on automation and orchestration could reduce effort later on.

## Storing Data