



Official Google Cloud Certified

Professional Data Engineer Study Guide

Includes interactive online learning environment and study tools:

- 2 custom practice exams
- More than 100 electronic flashcards
- Searchable key term glossary

DAN SULLIVAN



Official Google Cloud Certified Professional Data Engineer

Study Guide



Dan Sullivan

Copyright © 2020 by John Wiley & Sons, Inc., Indianapolis, Indiana

Published simultaneously in Canada and in the United Kingdom

ISBN: 978-1-119-61843-0

ISBN: 978-1-119-61844-7 (ebk.)

ISBN: 978-1-119-61845-4 (ebk.)

Manufactured in the United States of America

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Web site is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or Web site may provide or recommendations it may make. Further, readers should be aware that Internet Web sites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services or to obtain technical support, please contact our Customer Care Department within the U.S. at (877) 762-2974, outside the U.S. at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

Library of Congress Control Number: 2020936943

TRADEMARKS: Wiley, the Wiley logo, and the Sybex logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries, and may not be used without written permission. Google Cloud and the Google Cloud logo are trademarks of Google, LLC. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

to Katherine

Acknowledgments

I have been fortunate to work again with professionals from Waterside Productions, Wiley, and Google to create this Study Guide.

Carole Jelen, vice president of Waterside Productions, and Jim Minatel, associate publisher at John Wiley & Sons, continue to lead the effort to create Google Cloud certification guides. It was a pleasure to work with Gary Schwartz, project editor, who managed the process that got us from outline to a finished manuscript. Thanks to Christine O'Connor, senior production editor, for making the last stages of book development go as smoothly as they did.

I was also fortunate to work with Valerie Parham-Thompson again. Valerie's technical review improved the clarity and accuracy of this book tremendously.

Thank you to the Google Cloud subject-matter experts who reviewed and contributed to the material in this book:

Name	Title
Damon A. Runion	Technical Curriculum Developer, Data Engineering
Julianne Cuneo	Data Analytics Specialist, Google Cloud
Geoff McGill	Customer Engineer, Data Analytics
Susan Pierce	Solutions Manager, Smart Analytics and AI
Rachel Levy	Cloud Data Specialist Lead
Dustin Williams	Data Analytics Specialist, Google Cloud
Gbenga Awodokun	Customer Engineer, Data and Marketing Analytics
Dilraj Kaur	Big Data Specialist
Rebecca Ballough	Data Analytics Manager, Google Cloud
Robert Saxby	Staff Solutions Architect
Niel Markwick	Cloud Solutions Architect
Sharon Dashet	Big Data Product Specialist
Barry Searle	Solution Specialist - Cloud Data Management
Jignesh Mehta	Customer Engineer, Cloud Data Platform and Advanced Analytics

My sons James and Nicholas were my first readers, and they helped me to get the manuscript across the finish line. This book is dedicated to Katherine, my wife and partner in so many adventures.

About the Author



Dan Sullivan is a principal engineer and software architect. He specializes in data science, machine learning, and cloud computing. Dan is the author of the *Official Google Cloud Certified Professional Architect Study Guide* (Sybex, 2019), *Official Google Cloud Certified Associate Cloud Engineer Study Guide* (Sybex, 2019), *NoSQL for Mere Mortals* (Addison-Wesley Professional, 2015), and several LinkedIn Learning courses on databases, data science, and machine learning. Dan has certifications from Google and AWS, along with a Ph.D. in genetics and computational biology from Virginia Tech.

About the Technical Editor

Valerie Parham-Thompson has experience with a variety of open source data storage technologies, including MySQL, MongoDB, and Cassandra, as well as a foundation in web development in software-as-a-service (SaaS) environments. Her work in both development and operations in startups and traditional enterprises has led to solid expertise in web-scale data storage and data delivery.

Valerie has spoken at technical conferences on topics such as database security, performance tuning, and container management. She also often speaks at local meetups and volunteer events.

Valerie holds a bachelor's degree from the Kenan Flagler Business School at UNC-Chapel Hill, has certifications in MySQL and MongoDB, and is a Google Certified Professional Cloud Architect. She currently works in the Open Source Database Cluster at Pythian, headquartered in Ottawa, Ontario.

Follow Valerie's contributions to technical blogs on Twitter at [dataindataout](https://twitter.com/dataindataout).

CONTENTS

[Cover](#)

[Acknowledgments](#)

[About the Author](#)

[About the Technical Editor](#)

[Introduction](#)

[Assessment Test](#)

[Answers to Assessment Test](#)

[Chapter 1 Selecting Appropriate Storage Technologies](#)

[From Business Requirements to Storage Systems](#)

[Technical Aspects of Data: Volume, Velocity,
Variation, Access, and Security](#)

[Types of Structure: Structured, Semi-Structured,
and Unstructured](#)

[Schema Design Considerations](#)

[Exam Essentials](#)

[Review Questions](#)

[Chapter 2 Building and Operationalizing Storage
Systems](#)

[Cloud SQL](#)

[Cloud Spanner](#)

[Cloud Bigtable](#)

[Cloud Firestore](#)

[BigQuery](#)

[Cloud Memorystore](#)

[Cloud Storage](#)

[Unmanaged Databases](#)

[Exam Essentials](#)

[Review Questions](#)

[Chapter 3 Designing Data Pipelines](#)

[Overview of Data Pipelines](#)

[GCP Pipeline Components](#)

[Migrating Hadoop and Spark to GCP](#)

[Exam Essentials](#)

[Review Questions](#)

[Chapter 4 Designing a Data Processing Solution](#)

[Designing Infrastructure](#)

[Designing for Distributed Processing](#)

[Migrating a Data Warehouse](#)

[Exam Essentials](#)

[Review Questions](#)

[Chapter 5 Building and Operationalizing Processing Infrastructure](#)

[Provisioning and Adjusting Processing Resources](#)

[Monitoring Processing Resources](#)

[Exam Essentials](#)

[Review Questions](#)

[Chapter 6 Designing for Security and Compliance](#)

[Identity and Access Management with Cloud IAM](#)

[Using IAM with Storage and Processing Services](#)

[Data Security](#)

[Ensuring Privacy with the Data Loss Prevention API](#)

[Legal Compliance](#)

[Exam Essentials](#)

[Review Questions](#)

[Chapter 7 Designing Databases for Reliability, Scalability, and Availability](#)

[Designing Cloud Bigtable Databases for Scalability and Reliability](#)

[Designing Cloud Spanner Databases for Scalability and Reliability](#)

[Designing BigQuery Databases for Data Warehousing](#)

[Exam Essentials](#)

[Review Questions](#)

[Chapter 8 Understanding Data Operations for Flexibility and Portability](#)

[Cataloging and Discovery with Data Catalog](#)

[Data Preprocessing with Dataprep](#)

[Visualizing with Data Studio](#)

[Exploring Data with Cloud Datalab](#)

[Orchestrating Workflows with Cloud Composer](#)

[Exam Essentials](#)

[Review Questions](#)

[Chapter 9 Deploying Machine Learning Pipelines](#)

[Structure of ML Pipelines](#)

[GCP Options for Deploying Machine Learning Pipeline](#)

[Exam Essentials](#)

[Review Questions](#)

[Chapter 10 Choosing Training and Serving Infrastructure](#)

[Hardware Accelerators](#)

[Distributed and Single Machine Infrastructure](#)

[Edge Computing with GCP](#)

[Exam Essentials](#)

[Review Questions](#)

[Chapter 11 Measuring, Monitoring, and Troubleshooting Machine Learning Models](#)

[Three Types of Machine Learning Algorithms](#)

[Deep Learning](#)

[Engineering Machine Learning Models](#)

[Common Sources of Error in Machine Learning Models](#)

[Exam Essentials](#)

[Review Questions](#)

[Chapter 12 Leveraging Prebuilt Models as a Service](#)

[Sight](#)

[Conversation](#)

[Language](#)

[Structured Data](#)

[Exam Essentials](#)

[Review Questions](#)

[Appendix Answers to Review Questions](#)

[Chapter 1: Selecting Appropriate Storage Technologies](#)

[Chapter 2: Building and Operationalizing Storage Systems](#)

[Chapter 3: Designing Data Pipelines](#)

[Chapter 4: Designing a Data Processing Solution](#)

[Chapter 5: Building and Operationalizing Processing Infrastructure](#)

[Chapter 6: Designing for Security and Compliance](#)

[Chapter 7: Designing Databases for Reliability, Scalability, and Availability](#)

[Chapter 8: Understanding Data Operations for Flexibility and Portability](#)

[Chapter 9: Deploying Machine Learning Pipelines](#)

[Chapter 10: Choosing Training and Serving Infrastructure](#)

[Chapter 11: Measuring, Monitoring, and Troubleshooting Machine Learning Models](#)

[Chapter 12: Leveraging Prebuilt Models as a Service](#)

[Index](#)

[End User License Agreement](#)

List of Tables

Chapter 1

[Table 1.1](#)

[Table 1.2](#)

[Table 1.3](#)

[Table 1.4](#)

[Table 1.5](#)

[Table 1.6](#)

[Table 1.7](#)

Chapter 9

[Table 9.1](#)

Chapter 11

[Table 11.1](#)

[Table 11.2](#)

[Table 11.3](#)

List of Illustrations

Chapter 1

[Figure 1.1 Choosing a storage technology in GCP](#)

[Figure 1.2 Example graph of friends](#)

Chapter 2

[Figure 2.1 Basic Cloud SQL configuration](#)

[Figure 2.2 Optional configuration parameters in Cloud SQL](#)

[Figure 2.3 Configuring Cloud Spanner](#)

[Figure 2.4 Configuring a Bigtable cluster](#)

[Figure 2.5 Cost of a three-node Bigtable production cluster](#)

[Figure 2.6 BigQuery interactive interface with sample query](#)

Chapter 3

[Figure 3.1 A simple directed graph](#)

[Figure 3.2 A simple cyclic graph](#)

[Figure 3.3 An example ingestion stage of a data pipeline](#)

[Figure 3.4 Data pipeline with transformations](#)

[Figure 3.5 Example pipeline DAG with storage](#)

[Figure 3.6 Complete data pipeline from ingestion to analysis](#)

[Figure 3.7 A stream with sliding and tumbling three window](#)

[Figure 3.8 Data pipeline with both a hot path and a cold path](#)

[Figure 3.9 Creating a Cloud Dataflow job in the console using a template](#)

[Figure 3.10 Specifying parameters for the Word Count Template](#)

Chapter 4

[Figure 4.1 Summary of compute option features](#)

[Figure 4.2 Edge computing brings some computation outside the cloud and closer to where th...](#)

[Figure 4.3 A hub-and-spoke message broker pattern](#)

[Figure 4.4 Simple asynchronous message processing](#)

Chapter 5

[Figure 5.1 A basic VM instance provisioning form in the cloud console](#)

[Figure 5.2 Form for creating an instance template](#)

[Figure 5.3 Form for creating a managed instance group](#)

[Figure 5.4 Cloud console user interface for creating a Kubernetes cluster](#)

[Figure 5.5 Form to create a Bigtable instance](#)

[Figure 5.6 A cloud console form for creating a Cloud Dataproc cluster](#)

[Figure 5.7 A example log listing in Stackdriver Logging](#)

Chapter 6

[Figure 6.1 Create Role form in the cloud console](#)

[Figure 6.2 Selecting permissions from predefined roles](#)

[Figure 6.3 An example of a redacted image generated by the Data Loss Prevention API](#)

Chapter 7

[Figure 7.1 Cloud Bigtable uses a cluster of VMs and the Colossus filesystem for storing, a...](#)

[Figure 7.2 An example heatmap generated by Cloud Bigtable Key Visualizer](#)

[Figure 7.3 Specifying multiple clusters when creating a Cloud Bigtable instance](#)

[Figure 7.4 Cloud Spanner distributes splits across servers to avoid hotspots.](#)

[Figure 7.5 An example query execution plan](#)

Chapter 8

[Figure 8.1 Users can search and browse data assets from the Data Catalog overview page.](#)

[Figure 8.2 Example Data Catalog tag template](#)

[Figure 8.3 Cloud Dataprep shows statistics about the distribution of data in attributes.](#)

[Figure 8.4 An example report showing Google Analytics data](#)

[Figure 8.5 An example Cloud Datalab notebook](#)

Chapter 9

[Figure 9.1 ML pipelines are usually executed in cycles.](#)

[Figure 9.2 An example of a dataset that can be classified with a linear model](#)

[Figure 9.3 An example of a dataset that cannot be classified with a linear model](#)

[Figure 9.4 Neural networks have hyperparameters to specify the number of layers and the nu...](#)

[Figure 9.5 Random forest models have hyperparameters specifying the maximum number of tree...](#)

[Figure 9.6 The dashed line is a linear model that does not overfit, and the solid line fit...](#)

[Figure 9.7 A confusion matrix for a classifier making 100 predictions \(\$n = 100\$ \).](#)

Chapter 10

[Figure 10.1 CDNs distribute content to servers around the globe.](#)

[Figure 10.2 Edge computing can be used to process data locally and then send summary data t...](#)

[Figure 10.3 Reference architecture for Cloud IoT](#)

Chapter 11

[Figure 11.1 An example of how SVMs can be used to predict discrete values](#)

[Figure 11.2 An example decision tree to predict the type of an animal](#)

[Figure 11.3 An example logistic regression function using a sigmoid curve](#)

[Figure 11.4 A simple linear regression example](#)

[Figure 11.5 An example deep learning network](#)

[Figure 11.6 Examples of a \(a\) sigmoid function, \(b\) hyperbolic tangent function, and \(c\) re...](#)

Figure 11.7 The ROC is the curved line, and the AUC is the area under that curve line.

Introduction

The Google Cloud Certified Professional Data Engineer exam tests your ability to design, deploy, monitor, and adapt services and infrastructure for data-driven decision-making. The four primary areas of focus in this exam are as follows:

- Designing data processing systems
- Building and operationalizing data processing systems
- Operationalizing machine learning models
- Ensuring solution quality

Designing data processing systems involves selecting storage technologies, including relational, analytical, document, and wide-column databases, such as Cloud SQL, BigQuery, Cloud Firestore, and Cloud Bigtable, respectively. You will also be tested on designing pipelines using services such as Cloud Dataflow, Cloud Dataproc, Cloud Pub/Sub, and Cloud Composer. The exam will test your ability to design distributed systems that may include hybrid clouds, message brokers, middleware, and serverless functions. Expect to see questions on migrating data warehouses from on-premises infrastructure to the cloud.

The building and operationalizing data processing systems parts of the exam will test your ability to support storage systems, pipelines, and infrastructure in a production environment. This will include using managed services for storage as well as batch and stream processing. It will also cover common operations such as data ingestion, data cleansing, transformation, and integrating data with other

sources. As a data engineer, you are expected to understand how to provision resources, monitor pipelines, and test distributed systems.

Machine learning is an increasingly important topic. This exam will test your knowledge of prebuilt machine learning models available in GCP as well as the ability to deploy machine learning pipelines with custom-built models. You can expect to see questions about machine learning service APIs and data ingestion, as well as training and evaluating models. The exam uses machine learning terminology, so it is important to understand the nomenclature, especially terms such as model, supervised and unsupervised learning, regression, classification, and evaluation metrics.

The fourth domain of knowledge covered in the exam is ensuring solution quality, which includes security, scalability, efficiency, and reliability. Expect questions on ensuring privacy with data loss prevention techniques, encryption, identity, and access management, as well as ones about compliance with major regulations. The exam also tests a data engineer's ability to monitor pipelines with Stackdriver, improve data models, and scale resources as needed. You may also encounter questions that assess your ability to design portable solutions and plan for future business requirements.

In your day-to-day experience with GCP, you may spend more time working on some data engineering tasks than others. This is expected. It does, however, mean that you should be aware of the exam topics about which you may be less familiar. Machine learning questions can be especially challenging to data engineers who work primarily on ingestion and storage systems. Similarly, those who spend a majority of their time developing machine learning models may need to invest more time studying

schema modeling for NoSQL databases and designing fault-tolerant distributed systems.

What Does This Book Cover?

This book covers the topics outlined in the Google Cloud Professional Data Engineer exam guide available here:

cloud.google.com/certification/guides/data-engineer

Chapter 1: Selecting Appropriate Storage

Technologies This chapter covers selecting appropriate storage technologies, including mapping business requirements to storage systems; understanding the distinction between structured, semi-structured, and unstructured data models; and designing schemas for relational and NoSQL databases. By the end of the chapter, you should understand the various criteria that data engineers consider when choosing a storage technology.

Chapter 2: Building and Operationalizing Storage

Systems This chapter discusses how to deploy storage systems and perform data management operations, such as importing and exporting data, configuring access controls, and doing performance tuning. The services included in this chapter are as follows: Cloud SQL, Cloud Spanner, Cloud Bigtable, Cloud Firestore, BigQuery, Cloud Memorystore, and Cloud Storage. The chapter also includes a discussion of working with unmanaged databases, understanding storage costs and performance, and performing data lifecycle management.

Chapter 3: Designing Data Pipelines This chapter describes high-level design patterns, along with some variations on those patterns, for data pipelines. It also reviews how GCP services like Cloud Dataflow, Cloud Dataproc, Cloud Pub/Sub, and Cloud Composer are used to implement data pipelines. It also covers migrating data pipelines from an on-premises Hadoop cluster to GCP.

Chapter 4: Designing a Data Processing Solution In this chapter, you learn about designing infrastructure for data engineering and machine learning, including how to do several tasks, such as choosing an appropriate compute service for your use case; designing for scalability, reliability, availability, and maintainability; using hybrid and edge computing architecture patterns and processing models; and migrating a data warehouse from on-premises data centers to GCP.

Chapter 5: Building and Operationalizing Processing Infrastructure This chapter discusses managed processing resources, including those offered by App Engine, Cloud Functions, and Cloud Dataflow. The chapter also includes a discussion of how to use Stackdriver Metrics, Stackdriver Logging, and Stackdriver Trace to monitor processing infrastructure.

Chapter 6: Designing for Security and Compliance This chapter introduces several key topics of security and compliance, including identity and access management, data security, encryption and key management, data loss prevention, and compliance.

Chapter 7: Designing Databases for Reliability, Scalability, and Availability This chapter provides information on designing for reliability, scalability, and availability of three GCP databases: Cloud Bigtable, Cloud Spanner, and Cloud BigQuery. It also covers how to apply best practices for designing schemas, querying data, and taking advantage of the physical design properties of each database.

Chapter 8: Understanding Data Operations for Flexibility and Portability This chapter describes how to use the Data Catalog, a metadata management service supporting the discovery and management of data in Google Cloud. It also introduces Cloud Dataprep, a

preprocessing tool for transforming and enriching data, as well as Data Studio for visualizing data and Cloud Datalab for interactive exploration and scripting.

Chapter 9: Deploying Machine Learning Pipelines

Machine learning pipelines include several stages that begin with data ingestion and preparation and then perform data segregation followed by model training and evaluation. GCP provides multiple ways to implement machine learning pipelines. This chapter describes how to deploy ML pipelines using general-purpose computing resources, such as Compute Engine and Kubernetes Engine. Managed services, such as Cloud Dataflow and Cloud Dataproc, are also available, as well as specialized machine learning services, such as AI Platform, formerly known as Cloud ML.

Chapter 10: Choosing Training and Serving

Infrastructure This chapter focuses on choosing the appropriate training and serving infrastructure for your needs when serverless or specialized AI services are not a good fit for your requirements. It discusses distributed and single-machine infrastructure, the use of edge computing for serving machine learning models, and the use of hardware accelerators.

Chapter 11: Measuring, Monitoring, and Troubleshooting Machine Learning Models This chapter focuses on key concepts in machine learning, including machine learning terminology and core concepts and common sources of error in machine learning. Machine learning is a broad discipline with many areas of specialization. This chapter provides you with a high-level overview to help you pass the Professional Data Engineer exam, but it is not a substitute for learning machine learning from resources designed for that purpose.

Chapter 12: Leveraging Prebuilt ML Models as a Service This chapter describes Google Cloud Platform options for using pretrained machine learning models to help developers build and deploy intelligent services quickly. The services are broadly grouped into sight, conversation, language, and structured data. These services are available through APIs or through Cloud AutoML services.

Interactive Online Learning Environment and TestBank

Learning the material in the *Official Google Cloud Certified Professional Engineer Study Guide* is an important part of preparing for the Professional Data Engineer certification exam, but we also provide additional tools to help you prepare. The online TestBank will help you understand the types of questions that will appear on the certification exam.

The sample tests in the TestBank include all the questions in each chapter as well as the questions from the assessment test. In addition, there are two practice exams with 50 questions each. You can use these tests to evaluate your understanding and identify areas that may require additional study.

The flashcards in the TestBank will push the limits of what you should know for the certification exam. Over 100 questions are provided in digital format. Each flashcard has one question and one correct answer.

The online glossary is a searchable list of key terms introduced in this Study Guide that you should know for the Professional Data Engineer certification exam.

To start using these to study for the Google Cloud Certified Professional Data Engineer exam, go to www.wiley.com/go/sybextestprep and register your book to receive your unique PIN. Once you have the PIN, return to www.wiley.com/go/sybextestprep, find your book, and click Register, or log in and follow the link to register a new account or add this book to an existing account.

Additional Resources

People learn in different ways. For some, a book is an ideal way to study, whereas other learners may find video and audio resources a more efficient way to study. A combination of resources may be the best option for many of us. In addition to this Study Guide, here are some other resources that can help you prepare for the Google Cloud Professional Data Engineer exam:

The Professional Data Engineer Certification Exam Guide:

<https://cloud.google.com/certification/guides/data-engineer/>

Exam FAQs: <https://cloud.google.com/certification/faqs/>

Google's Assessment Exam:

<https://cloud.google.com/certification/practice-exam/data-engineer>

Google Cloud Platform documentation:

<https://cloud.google.com/docs/>

Coursera's on-demand courses in "Architecting with Google Cloud Platform Specialization" and "Data Engineering with Google Cloud" are both relevant to data engineering:

www.coursera.org/specializations/gcp-architecture

<https://www.coursera.org/professional-certificates/gcp-data-engineering>

QwikLabs Hands-on Labs:

<https://google.qwiklabs.com/quests/25>

Linux Academy Google Cloud Certified Professional Data Engineer video course:

<https://linuxacademy.com/course/google-cloud-data-engineer/>

The best way to prepare for the exam is to perform the tasks of a data engineer and work with the Google Cloud Platform.



Exam objectives are subject to change at any time without prior notice and at Google's sole discretion. Please visit the Google Cloud Professional Data Engineer website (<https://cloud.google.com/certification/data-engineer>) for the most current listing of exam objectives.

Objective Map

Objective	Chapter
Section 1: Designing data processing system	
1.1 Selecting the appropriate storage technologies	1
1.2 Designing data pipelines	2, 3
1.3 Designing a data processing solution	4
1.4 Migrating data warehousing and data processing	4
Section 2: Building and operationalizing data processing systems	
2.1 Building and operationalizing storage systems	2
2.2 Building and operationalizing pipelines	3
2.3 Building and operationalizing infrastructure	5
Section 3: Operationalizing machine learning models	
3.1 Leveraging prebuilt ML models as a service	12
3.2 Deploying an ML pipeline	9
3.3 Choosing the appropriate training and serving infrastructure	10
3.4 Measuring, monitoring, and troubleshooting machine learning models	11
Section 4: Ensuring solution quality	
4.1 Designing for security and compliance	6
4.2 Ensuring scalability and efficiency	7
4.3 Ensuring reliability and fidelity	8
4.4 Ensuring flexibility and portability	8

Assessment Test

1. You are migrating your machine learning operations to GCP and want to take advantage of managed services. You have been managing a Spark cluster because you use the MLlib library extensively. Which GCP managed service would you use?
 - A. Cloud Dataprep
 - B. Cloud Dataproc
 - C. Cloud Dataflow
 - D. Cloud Pub/Sub
2. Your team is designing a database to store product catalog information. They have determined that you need to use a database that supports flexible schemas and transactions. What service would you expect to use?
 - A. Cloud SQL
 - B. Cloud BigQuery
 - C. Cloud Firestore
 - D. Cloud Storage
3. Your company has been losing market share because competitors are attracting your customers with a more personalized experience on their e-commerce platforms, including providing recommendations for products that might be of interest to them. The CEO has stated that your company will provide equivalent services within 90 days. What GCP service would you use to help meet this objective?
 - A. Cloud Bigtable