



The Definitive Guide to Google Vertex AI

Copyright © 2023 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the authors, nor Packt Publishing or its dealers and distributors, will be held liable for any damages caused or alleged to have been caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

Group Product Manager: Niranjan Naikwadi

Publishing Product Manager: Sanjana Gupta

Book Project Manager: Hemangi Lotlikar

Senior Editor: Gowri Rekha

Technical Editor: Rahul Limbachiya

Copy Editor: Safis Editing

Project Coordinator: Shambhavi Mishra

Proofreader: Safis Editing

Indexer: Pratik Shirodkar

Production Designer: Shankar Kalbhor

DevRel Marketing Executive (DRME): Vinishka Kalra

First published: December 2023

Production reference: 1211223

Published by

Packt Publishing Ltd.

Grosvenor House

11 St Paul's Square

Birmingham

B3 1RB, UK

ISBN 978-1-80181-526-0

www.packtpub.com

To my incredible wife, who gracefully navigates our cloud rivalry with love and patience. This book wouldn't be possible without your constant support and encouragement. Thank you for tolerating my late-night writing and even later-night snacking. Your patience is my favorite superpower.

To my parents, who still can't fully explain what I do for a living but are proud nonetheless – you're the original algorithms of my life. Thank you for programming me with constant love, support, and the occasional necessary reboot!

To my wonderful daughters, without whom I would never have really understood why so many authors joke about their kids delaying their books. Now I do. Thank you for bringing immense joy and well-needed system shutdowns to my life.

And to my colleagues, the wizards of Google Cloud, who speak fluent Python and dream in code – without you, this book would just be a collection of funny error messages.

This book is dedicated to all of you. May our models always converge, and may we all never run out of GPUs!

– Jasmeet Bhatia

To my mother, Smt. Sarita Devi, and my father, Mr. Inderpal Singh, for their sacrifices, constant love, and never-ending support. Thank you for teaching me to believe in myself, in God, and in my dreams.

To my little brother, Chakit Gill, for continuous encouragement, support, and love. Thanks for being my best friend; I am really proud of you.

To my friends and colleagues for their inspiration, motivation, and always being there for me.

And, most importantly, to all the readers – I hope this book helps you with your goals, because that's the real motivation behind writing this book and every single technical article that I share publicly on my blog.

– Kartik Chaudhary

Contributors

About the authors

Jasmeet Bhatia is a machine learning solution architect with over 18 years of industry experience, with the last 10 years focused on global-scale data analytics and machine learning solutions. In his current role at Google, he works closely with key GCP enterprise customers to provide them guidance on how to best use Google's cutting-edge machine learning products. At Google, he has also worked as part of the Area 120 incubator on building innovative data products such as Demand Signals, and he has been involved in the launch of Google products such as Time Series Insights. Before Google, he worked in similar roles at Microsoft and Deloitte.

When not immersed in technology, he loves spending time with his wife and two daughters, reading books, watching movies, and exploring the scenic trails of southern California.

He holds a bachelor's degree in electronics engineering from Jamia Millia Islamia University in India and an MBA from the University of California Los Angeles (UCLA) Anderson School of Management.

Kartik Chaudhary is an AI enthusiast, educator, and ML professional with 6+ years of industry experience. He currently works as a senior AI engineer with Google to design and architect ML solutions for Google's strategic customers, leveraging core Google products, frameworks, and AI tools. He previously worked with UHG, as a data scientist, and helped in making the healthcare system work better for everyone. Kartik has filed nine patents at the intersection of AI and healthcare.

Kartik loves sharing knowledge and runs his own blog on AI, titled *Drops of AI*.

Away from work, he loves watching anime and movies and capturing the beauty of sunsets.

I would like to thank my parents, my brother, and my friends for their constant love and support.

Chetan Apsunde is an experienced software engineer, specializing in conversational AI and machine learning with a robust nine-year IT background. He works with Google to build cloud solutions using CCAI and GenAI. He is passionate about creating intelligent, user-centric solutions at the intersection of technology and human interaction.

Surya Tripathi is a seasoned data scientist with nearly nine years of expertise in data science, analysis, and data engineering. He holds a bachelor's degree in electronics and communications engineering and a master's in applied mathematics from Liverpool John Moores University. He is proficient in cloud platforms (GCP, Azure, AWS, and IBM) and has extensive GCP experience, delivering ML solutions in CPG, healthcare, banking, and supply chain. Involved in the full data science life cycle, he excels in requirement gathering, data analysis, model development, and MLOps. With experience in both consulting and product companies, he is currently affiliated with a top consulting firm, where his primary focus areas include generative AI and demand forecasting.

I would like to acknowledge my wife, Dhivya, and my kids, Sanjana and Saisasthik, for being a constant source of support and encouragement throughout this book-reviewing journey.

Lakshmanan Sethu Sankaranarayanan is an award-winning AI/ML cloud industry leader in the fields of data, AI, and ML. He helps enterprise customers to migrate to Google Cloud, Azure, and AWS. He serves on the Technical Advisory Board for AI/ML solutions at Packt, and he is also the Technical Editor for Packt and O'Reilly. He has been honored with three LinkedIn TopVoice awards for his contributions to AI/ML and cloud computing. He earned four Microsoft Most Valuable Player awards for his outstanding contribution to the cloud community.

I would like to thank my parents, my brother, and my friends for their constant love and support.

Gopala Dhar has worked with and implemented state-of-the-art technology in the field of AI to solve real-world business use cases at scale. He has four published patents to his name, ranging from the field of software design to hardware manufacturing, including embedded systems. His latest stint is at Google as an AI engineer. His areas of expertise include ML, ML system design, reinforcement learning, and, most recently, generative AI. He shares what he learns frequently through blog posts and open source contributions. He has won several awards from various academic as well as professional institutions, including the Indian Institute of Technology in Mumbai, the Indian Institute of Management in Bangalore, Texas Instruments, and Google.

Surya Tripathi is a seasoned data scientist with nearly nine years of expertise in data science, analysis, and data engineering. He holds a bachelor's degree in electronics and communications engineering and a master's in applied mathematics from Liverpool John Moores University. He is proficient in cloud platforms (GCP, Azure, AWS, and IBM) and has extensive GCP experience, delivering ML solutions in CPG, healthcare, banking, and supply chain. Involved in the full data science life cycle, he excels in requirement gathering, data analysis, model development, and MLOps. With experience in both consulting and product companies, he is currently affiliated with a top consulting firm, where his primary focus areas include generative AI and demand forecasting.

Lakshmanan Sethu Sankaranarayanan is an award-winning AI/ML cloud industry leader in the fields of data, AI, and ML. He helps enterprise customers to migrate to Google Cloud, Azure, and AWS. He serves on the Technical Advisory Board for AI/ML solutions at Packt, and he is also the Technical Editor for Packt and O'Reilly. He has been honored with three LinkedIn TopVoice awards for his contributions to AI/ML and cloud computing. He earned four Microsoft Most Valuable Player awards for his outstanding contribution to the cloud community.

I would like to thank my parents, my brother, and my friends for their constant love and support.

Chetan Apsunde is an experienced software engineer, specializing in conversational AI and machine learning with a robust nine-year IT background. He works with Google to build cloud solutions using CCAI and GenAI. He is passionate about creating intelligent, user-centric solutions at the intersection of technology and human interaction.

Surya Tripathi is a seasoned data scientist with nearly nine years of expertise in data science, analysis, and data engineering. He holds a bachelor's degree in electronics and communications engineering and a master's in applied mathematics from Liverpool John Moores University. He is proficient in cloud platforms (GCP, Azure, AWS, and IBM) and has extensive GCP experience, delivering ML solutions in CPG, healthcare, banking, and supply chain. Involved in the full data science life cycle, he excels in requirement gathering, data analysis, model development, and MLOps. With experience in both consulting and product companies, he is currently affiliated with a top consulting firm, where his primary focus areas include generative AI and demand forecasting.

Gopala Dhar has worked with and implemented state-of-the-art technology in the field of AI to solve real-world business use cases at scale. He has four published patents to his name, ranging from the field of software design to hardware manufacturing, including embedded systems. His latest stint is at Google as an AI engineer. His areas of expertise include ML, ML system design, reinforcement learning, and, most recently, generative AI. He shares what he learns frequently through blog posts and open source contributions. He has won several awards from various academic as well as professional institutions, including the Indian Institute of Technology in Mumbai, the Indian Institute of Management in Bangalore, Texas Instruments, and Google.

Surya Tripathi is a seasoned data scientist with nearly nine years of expertise in data science, analysis, and data engineering. He holds a bachelor's degree in electronics and communications engineering and a master's in applied mathematics from Liverpool John Moores University. He is proficient in cloud platforms (GCP, Azure, AWS, and IBM) and has extensive GCP experience, delivering ML solutions in CPG, healthcare, banking, and supply chain. Involved in the full data science life cycle, he excels in requirement gathering, data analysis, model development, and MLOps. With experience in both consulting and product companies, he is currently affiliated with a top consulting firm, where his primary focus areas include generative AI and demand forecasting.

Lakshmanan Sethu Sankaranarayanan is an award-winning AI/ML cloud industry leader in the fields of data, AI, and ML. He helps enterprise customers to migrate to Google Cloud, Azure, and AWS. He serves on the Technical Advisory Board for AI/ML solutions at Packt, and he is also the Technical Editor for Packt and O'Reilly. He has been honored with three LinkedIn TopVoice awards for his contributions to AI/ML and cloud computing. He earned four Microsoft Most Valuable Player awards for his outstanding contribution to the cloud community.

I would like to thank my parents, my brother, and my friends for their constant love and support.

Chetan Apsunde is an experienced software engineer, specializing in conversational AI and machine learning with a robust nine-year IT background. He works with Google to build cloud solutions using CCAI and GenAI. He is passionate about creating intelligent, user-centric solutions at the intersection of technology and human interaction.

Surya Tripathi is a seasoned data scientist with nearly nine years of expertise in data science, analysis, and data engineering. He holds a bachelor's degree in electronics and communications engineering and a master's in applied mathematics from Liverpool John Moores University. He is proficient in cloud platforms (GCP, Azure, AWS, and IBM) and has extensive GCP experience, delivering ML solutions in CPG, healthcare, banking, and supply chain. Involved in the full data science life cycle, he excels in requirement gathering, data analysis, model development, and MLOps. With experience in both consulting and product companies, he is currently affiliated with a top consulting firm, where his primary focus areas include generative AI and demand forecasting.

Gopala Dhar has worked with and implemented state-of-the-art technology in the field of AI to solve real-world business use cases at scale. He has four published patents to his name, ranging from the field of software design to hardware manufacturing, including embedded systems. His latest stint is at Google as an AI engineer. His areas of expertise include ML, ML system design, reinforcement learning, and, most recently, generative AI. He shares what he learns frequently through blog posts and open source contributions. He has won several awards from various academic as well as professional institutions, including the Indian Institute of Technology in Mumbai, the Indian Institute of Management in Bangalore, Texas Instruments, and Google.

Surya Tripathi is a seasoned data scientist with nearly nine years of expertise in data science, analysis, and data engineering. He holds a bachelor's degree in electronics and communications engineering and a master's in applied mathematics from Liverpool John Moores University. He is proficient in cloud platforms (GCP, Azure, AWS, and IBM) and has extensive GCP experience, delivering ML solutions in CPG, healthcare, banking, and supply chain. Involved in the full data science life cycle, he excels in requirement gathering, data analysis, model development, and MLOps. With experience in both consulting and product companies, he is currently affiliated with a top consulting firm, where his primary focus areas include generative AI and demand forecasting.

Lakshmanan Sethu Sankaranarayanan is an award-winning AI/ML cloud industry leader in the fields of data, AI, and ML. He helps enterprise customers to migrate to Google Cloud, Azure, and AWS. He serves on the Technical Advisory Board for AI/ML solutions at Packt, and he is also the Technical Editor for Packt and O'Reilly. He has been honored with three LinkedIn TopVoice awards for his contributions to AI/ML and cloud computing. He earned four Microsoft Most Valuable Player awards for his outstanding contribution to the cloud community.

I would like to thank my parents, my brother, and my friends for their constant love and support.

Chetan Apsunde is an experienced software engineer, specializing in conversational AI and machine learning with a robust nine-year IT background. He works with Google to build cloud solutions using CCAI and GenAI. He is passionate about creating intelligent, user-centric solutions at the intersection of technology and human interaction.

Surya Tripathi is a seasoned data scientist with nearly nine years of expertise in data science, analysis, and data engineering. He holds a bachelor's degree in electronics and communications engineering and a master's in applied mathematics from Liverpool John Moores University. He is proficient in cloud platforms (GCP, Azure, AWS, and IBM) and has extensive GCP experience, delivering ML solutions in CPG, healthcare, banking, and supply chain. Involved in the full data science life cycle, he excels in requirement gathering, data analysis, model development, and MLOps. With experience in both consulting and product companies, he is currently affiliated with a top consulting firm, where his primary focus areas include generative AI and demand forecasting.

Gopala Dhar has worked with and implemented state-of-the-art technology in the field of AI to solve real-world business use cases at scale. He has four published patents to his name, ranging from the field of software design to hardware manufacturing, including embedded systems. His latest stint is at Google as an AI engineer. His areas of expertise include ML, ML system design, reinforcement learning, and, most recently, generative AI. He shares what he learns frequently through blog posts and open source contributions. He has won several awards from various academic as well as professional institutions, including the Indian Institute of Technology in Mumbai, the Indian Institute of Management in Bangalore, Texas Instruments, and Google.

Surya Tripathi is a seasoned data scientist with nearly nine years of expertise in data science, analysis, and data engineering. He holds a bachelor's degree in electronics and communications engineering and a master's in applied mathematics from Liverpool John Moores University. He is proficient in cloud platforms (GCP, Azure, AWS, and IBM) and has extensive GCP experience, delivering ML solutions in CPG, healthcare, banking, and supply chain. Involved in the full data science life cycle, he excels in requirement gathering, data analysis, model development, and MLOps. With experience in both consulting and product companies, he is currently affiliated with a top consulting firm, where his primary focus areas include generative AI and demand forecasting.

Lakshmanan Sethu Sankaranarayanan is an award-winning AI/ML cloud industry leader in the fields of data, AI, and ML. He helps enterprise customers to migrate to Google Cloud, Azure, and AWS. He serves on the Technical Advisory Board for AI/ML solutions at Packt, and he is also the Technical Editor for Packt and O'Reilly. He has been honored with three LinkedIn TopVoice awards for his contributions to AI/ML and cloud computing. He earned four Microsoft Most Valuable Player awards for his outstanding contribution to the cloud community.

I would like to thank my parents, my brother, and my friends for their constant love and support.

Chetan Apsunde is an experienced software engineer, specializing in conversational AI and machine learning with a robust nine-year IT background. He works with Google to build cloud solutions using CCAI and GenAI. He is passionate about creating intelligent, user-centric solutions at the intersection of technology and human interaction.

Surya Tripathi is a seasoned data scientist with nearly nine years of expertise in data science, analysis, and data engineering. He holds a bachelor's degree in electronics and communications engineering and a master's in applied mathematics from Liverpool John Moores University. He is proficient in cloud platforms (GCP, Azure, AWS, and IBM) and has extensive GCP experience, delivering ML solutions in CPG, healthcare, banking, and supply chain. Involved in the full data science life cycle, he excels in requirement gathering, data analysis, model development, and MLOps. With experience in both consulting and product companies, he is currently affiliated with a top consulting firm, where his primary focus areas include generative AI and demand forecasting.

Gopala Dhar has worked with and implemented state-of-the-art technology in the field of AI to solve real-world business use cases at scale. He has four published patents to his name, ranging from the field of software design to hardware manufacturing, including embedded systems. His latest stint is at Google as an AI engineer. His areas of expertise include ML, ML system design, reinforcement learning, and, most recently, generative AI. He shares what he learns frequently through blog posts and open source contributions. He has won several awards from various academic as well as professional institutions, including the Indian Institute of Technology in Mumbai, the Indian Institute of Management in Bangalore, Texas Instruments, and Google.

Surya Tripathi is a seasoned data scientist with nearly nine years of expertise in data science, analysis, and data engineering. He holds a bachelor's degree in electronics and communications engineering and a master's in applied mathematics from Liverpool John Moores University. He is proficient in cloud platforms (GCP, Azure, AWS, and IBM) and has extensive GCP experience, delivering ML solutions in CPG, healthcare, banking, and supply chain. Involved in the full data science life cycle, he excels in requirement gathering, data analysis, model development, and MLOps. With experience in both consulting and product companies, he is currently affiliated with a top consulting firm, where his primary focus areas include generative AI and demand forecasting.

Lakshmanan Sethu Sankaranarayanan is an award-winning AI/ML cloud industry leader in the fields of data, AI, and ML. He helps enterprise customers to migrate to Google Cloud, Azure, and AWS. He serves on the Technical Advisory Board for AI/ML solutions at Packt, and he is also the Technical Editor for Packt and O'Reilly. He has been honored with three LinkedIn TopVoice awards for his contributions to AI/ML and cloud computing. He earned four Microsoft Most Valuable Player awards for his outstanding contribution to the cloud community.

I would like to thank my parents, my brother, and my friends for their constant love and support.

Chetan Apsunde is an experienced software engineer, specializing in conversational AI and machine learning with a robust nine-year IT background. He works with Google to build cloud solutions using CCAI and GenAI. He is passionate about creating intelligent, user-centric solutions at the intersection of technology and human interaction.

Surya Tripathi is a seasoned data scientist with nearly nine years of expertise in data science, analysis, and data engineering. He holds a bachelor's degree in electronics and communications engineering and a master's in applied mathematics from Liverpool John Moores University. He is proficient in cloud platforms (GCP, Azure, AWS, and IBM) and has extensive GCP experience, delivering ML solutions in CPG, healthcare, banking, and supply chain. Involved in the full data science life cycle, he excels in requirement gathering, data analysis, model development, and MLOps. With experience in both consulting and product companies, he is currently affiliated with a top consulting