Community Experience Distilled

Data Lake Development with Big Data

Explore architectural approaches to building Data Lakes that ingest, index, manage, and analyze massive amounts of data using Big Data technologies

Pradeep Pasupuleti
Beulah Salome Purra

[PACKT] PUBLISHING

Data Lake Development with Big Data

Explore architectural approaches to building Data Lakes that ingest, index, manage, and analyze massive amounts of data using Big Data technologies

Pradeep Pasupuleti

Beulah Salome Purra

[PACKT] PUBLISHING

BIRMINGHAM - MUMBAI

Data Lake Development with Big Data

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, and its dealers and distributors will be held liable for any damages caused or alleged to be 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.

First published: November 2015

Production reference: 1241115

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78588-808-3

www.packtpub.com

Credits

Authors
Pradeep Pasupuleti
Beulah Salome Purra

Reviewer
Dr. Kornel Amadeusz Skałkowski

Commissioning Editor
Priya Singh

Acquisition Editor
Ruchita Bhansali

Content Development Editor
Rohit Kumar Singh

Technical Editor
Saurabh Malhotra

Copy Editor
Trishya Hajare

Project Coordinator
Izzat Contractor

Proofreader
Safis Editing

Indexer
Hemangini Bari

Graphics
Jason Monteiro
Kirk D'Penha

Production Coordinator
Shantanu N. Zagade

Cover Work
Shantanu N. Zagade

About the Authors

Pradeep Pasupuleti has 18 years of experience in architecting and developing distributed and real-time data-driven systems. He constantly explores ways to use the power and promise of advanced analytics-driven platforms to solve the problems of the common man. He founded Datatma, a consulting firm, with a mission to humanize Big Data analytics, putting it to use to solve simple problems that serve a higher purpose.

He architected robust Big Data-enabled automated learning engines that enterprises regularly use in production in order to save time, money, and the lives of humans.

He built solid interdisciplinary data science teams that bridged the gap between theory and practice, thus, creating compelling data products. His primary focus is always to ensure his customers are delighted by assisting and addressing their business problems through data products that use Big Data technologies and algorithms. He consistently demonstrated thought leadership by solving high-dimensional data problems and getting phenomenal results.

He has performed strategic leadership roles in technology consulting, advising Fortune 100 companies on Big Data strategy and creating Big Data Centers of Excellence.

He has worked on use cases such as enterprise Data Lake, fraud detection, patient re-admission prediction, student performance prediction, claims optimization sentiment mining, cloud infrastructure SLA violation prediction, data leakage prevention, and mainframe offloaded ETL on Hadoop.

In the book *Pig Design Patterns*, *Packt Publishing*, he has compiled his learning and experiences from the challenges involved in building Hadoop-driven data products such as data ingest, data cleaning and validating, data transformation, dimensionality reduction, and many other interesting Big Data war stories.

Out of his office hours, he enjoys running marathons, exploring archeological sites, finding patterns in unrelated data sources, mentoring start-ups, and budding researchers.

He can be reached at `Pasupuleti.pradeepkumar@gmail.com` and `https://in.linkedin.com/in/pradeeppasupuleti`.

Acknowledgement

This book is dedicated to the loving memory of my mother, Smt. Sumathy; without her never-failing encouragement and everlasting love I would have never been half as good.

First and foremost, I have to thank my father, Sri. Prabhakar Pasupuleti, who never ceases to be a constant source of inspiration, a ray of hope, humility and strength, and whose support and guidance have given me the courage to chase my dreams.

I should also express my deep sense of gratitude to each of my family members, Sushma, Sresht, and Samvruth, who stood by me at every moment through very tough times and enabled me to complete this book.

I would like to sincerely thank all my teachers who were instrumental in shaping me. Among them, I would like to thank Usha Madam, Vittal Rao Sir, Gopal Krishna Sir, and Brindavan Sir for their stellar role in improving me.

I would also like to thank all my friends for their understanding in many ways. Their friendship makes my life a wonderful experience. I cannot list all the names here, but you are always on my mind.

Special thanks to the team at Packt for their contribution to this book.

Finally, I would like to thank my team, Salome et.al, that has placed immense faith in the power of Big Data analytics and built cutting edge data products.

Thank you lord, for always being there for me.

Beulah Salome Purra has over 11 years of experience and she specializes in building highly scalable distributed systems. She has worked extensively on architecting multiple large-scale Big Data solutions for Fortune 100 companies. Her core expertise lies in working on Big Data Analytics. In her current role at ATMECS, her focus is on building robust and scalable data products that extract value from huge data assets.

She can be reached at `https://www.linkedin.com/in/beulahsalomep`.

I am grateful to my parents, Rathnam and Padma, who have constantly encouraged and supported me throughout. I would like to thank my husband, Pratap, for his help on this book, his patience, love, and support; my brothers, Joel and Michael, for all their support.

I would like to profusely thank Pradeep Pasupuleti for mentoring me; working with him has been an enriching experience. I can't thank him enough for his constant encouragement, guidance, support, and for providing me an opportunity to work with him on this book.

Special thanks to David Hawke, Sanjay Singh, and Ravi Velagapudi—the leadership team at ATMECS—for their encouragement and support while I was writing this book.

Thanks to the editors and reviewers at Packt for all their effort in making this book better.

About the Reviewer

Dr. Kornel Amadeusz Skałkowski has a solid academic and industrial background. For more than 5 years, he worked as an assistant at AGH University of Science and Technology in Krakow. In 2015, he obtained his PhD. in the subject of machine learning-based adaptation of the SOA systems. He has cooperated with several companies on various projects concerning intelligent systems, machine learning, and Big Data. Currently, he works as a Big Data developer for SAP SE.

He is the co-author of 19 papers concerning software engineering, SOA systems, and machine learning. He also works as a reviewer for the American Journal of Software Engineering and Applications. He has participated in numerous European and national scientific projects. His research interests include machine learning, Big Data, and software engineering.

I would like to kindly thank my family, relatives, and friends, for their endless patience and support during the reviewing of this book.

Table of Contents

Preface

The book *Data Lake Development with Big Data* is a practical guide to help you learn the essential architectural approaches to design and build Data Lakes. It walks you through the various components of Data Lakes, such as data intake, management, consumption, and governance with a specific focus on practical implementation scenarios.

Data Lake is a highly scalable data platform for better search, analytical processing, and cheaper storage of huge volumes of any structured data acquired from disparate sources.

Traditional Data Management systems are constrained by data silos, upfront data modeling, rigid data structures, and schema-based write approaches while storing and processing data. This hampers the holistic analysis of data residing in multiple silos and excludes unstructured data sources from analysis. The data is generally modeled to answer known business questions.

With Data Lake, there are no more data silos; all the data can be utilized to get a coherent view that can power a new generation of data-aware analytics applications. With Data Lake, you don't have to know all the business questions in advance, as the data can be modeled later using the schema-less approach and it is possible to ask complex far-reaching questions on all the data at any time to find out hidden patterns and complex relationships in the data.

After reading this book, you will be able to address the shortcoming of traditional data systems through the best practices highlighted in this book for building Data Lake. You will understand the complete lifecycle of architecting/building Data Lake with Big Data technologies such as Hadoop, Storm, Spark, and Splunk. You will gain a comprehensive knowledge of various stages in Data Lake such as data intake, data management, and data consumption with focus on the practical use cases at each stage. You will benefit from the book's detailed coverage of data governance, data security, data lineage tracking, metadata management, data provisioning, and consumption.

As Data Lake is such an advanced complex topic, we are honored and excited to author the first book of its kind in the world. However, at the same time, as the topic being so vast and as there is no one-size-fits-all kind of Data Lake architecture, it is very challenging to appeal to a wide audience footprint. As it is a mini series book, which limits the page count, it is extremely difficult to cover every topic in detail without breaking the ceiling. Given these constraints, we have taken a reader-centric approach in writing this book because the broader understanding of the overall concept of Data Lake is far more important than the in-depth understanding of all the technologies and architectural possibilities that go into building Data Lake.

Using this guiding principle, we refrained from the in-depth coverage of any single topic, because we could not possibly do justice to it. At the same time we made efforts to organize chapters to mimick the sequential flow of data in a typical organization so that it is intuitive for the reader to quickly grasp the concepts of Data Lake from an organizational data flow perspective. In order to make the abstract concepts relatable to the real world, we have followed a use case-based approach where practical implementation scenarios of each key Data Lake component are explained. This we believe will help the reader quickly understand the architectural implications of various Big Data technologies that are used for building these components.

What this book covers

Chapter 1, *The Need for Data Lake*, helps you understand what Data Lake is, its architecture and key components, and the business contexts where Data Lake can be successfully deployed. You will also learn the limitations of the traditional data architectures and how Data Lake addresses some of these inadequacies and provides significant benefits.

Chapter 2, *Data Intake*, helps you understand the Intake Tier in detail where we will explore the process of obtaining huge volumes of data into Data Lake. You will learn the technology perspective of the various External Data Sources and Hadoop-based data transfer mechanisms to pull or push data into Data Lake.

Chapter 3, *Data Integration, Quality, and Enrichment*, explores the processes that are performed on vast quantities of data in the Management Tier. You will get a deeper understanding of the key technology aspects and components such as profiling, validation, integration, cleansing, standardization, and enrichment using Hadoop ecosystem components.

Chapter 4, *Data Discovery and Consumption*, helps you understand how data can be discovered, packaged, and provisioned, for it to be consumed by the downstream systems. You will learn the key technology aspects, architectural guidance and tools for data discovery, and data provisioning functionalities.

Chapter 5, *Data Governance*, explores the details, need, and utility of data governance in a Data Lake environment. You will learn how to deal with metadata management, lineage tracking, data lifecycle management to govern the usability, security, integrity, and availability of the data through the data governance processes applied on the data in Data Lake. This chapter also explores how the current Data Lake can evolve in a futuristic setting.

What you need for this book

As this book covers only the architectural details and acts as a guide for decision-making, we have not provided any code examples. Hence, there is no explicit software prerequisite.

Who this book is for

Data Lake Development with Big Data is intended for architects and senior managers who are responsible for building a strategy around their current data architecture, helping them identify the need for Data Lake implementation in an organizational business context.

Good knowledge on master data management, information lifecycle management, data governance, data product design, data engineering, systems architecture, and experience on Big Data technologies such as Hadoop, Spark, Splunk, and Storm is necessary.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "We can include other contexts through the use of the `include` directive."

New terms and **important words** are shown in bold. Words that you see on the screen, for example, in menus or dialog boxes, appear in the text like this: "Clicking the **Next** button moves you to the next screen."

Warnings or important notes appear in a box like this.

Tips and tricks appear like this.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply e-mail `feedback@packtpub.com`, and mention the book's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at `www.packtpub.com/authors`.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.