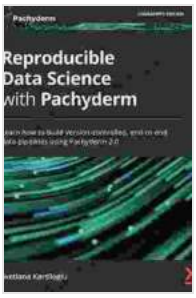
Build Version Controlled End-to-End Data Pipelines Using Pachyderm

Data pipelines are essential for modern data-driven organizations. They enable you to automate the movement and processing of data between different systems, ensuring that your data is always up-to-date, accurate, and accessible.

However, building and maintaining data pipelines can be a complex and time-consuming process. Traditional approaches often involve manually scripting each step of the pipeline, which can lead to errors and inconsistencies. Additionally, it can be difficult to track changes to the pipeline over time, making it challenging to troubleshoot issues or roll back changes.



Reproducible Data Science with Pachyderm: Learn how to build version-controlled, end-to-end data pipelines using Pachyderm 2.0 by Svetlana Karslioglu

★★★★★ 5 out of 5

Language	: English
File size	: 11815 KB
Text-to-Speech	: Enabled
Screen Reader	: Supported
Enhanced typesetting	: Enabled
Print length	: 364 pages
Paperback	: 200 pages
Item Weight	: 11.2 ounces
Dimensions	: 5.5 x 0.5 x 8.5 inches

Pachyderm is a new open-source platform that makes it easy to build and maintain version controlled end-to-end data pipelines. Pachyderm provides a unified platform for data ingestion, processing, storage, and serving, and it uses a Git-like version control system to track changes to the pipeline over time.

In this book, you will learn how to use Pachyderm to build version controlled end-to-end data pipelines. You will cover the following topics:

- to Pachyderm
- Building a simple data pipeline
- Versioning and managing data pipelines
- Scaling and securing data pipelines
- Advanced topics in Pachyderm

This book is for data engineers, data scientists, and anyone else who wants to learn how to build and maintain robust and reliable data pipelines.

Table of Contents

Pachyderm is a new open-source platform that makes it easy to build and maintain version controlled end-to-end data pipelines. Pachyderm provides a unified platform for data ingestion, processing, storage, and serving, and it uses a Git-like version control system to track changes to the pipeline over time.

Pachyderm is designed to address the challenges of building and maintaining data pipelines in a modern data-driven organization. Traditional approaches to data pipeline development often involve manually scripting each step of the pipeline, which can lead to errors and inconsistencies. Additionally, it can be difficult to track changes to the pipeline over time, making it challenging to troubleshoot issues or roll back changes.

Pachyderm solves these problems by providing a unified platform for data pipeline development and management. Pachyderm's Git-like version control system makes it easy to track changes to the pipeline over time, and its declarative pipeline definition language makes it easy to define and manage complex data pipelines.

Building a Simple Data Pipeline

In this section, you will learn how to build a simple data pipeline using Pachyderm. We will start by creating a new Pachyderm repository and then we will add a data source, a data processor, and a data sink to the pipeline.

1. Create a new Pachyderm repository
2. Add a data source to the pipeline
3. Add a data processor to the pipeline
4. Add a data sink to the pipeline

5. Run the pipeline

Versioning and Managing Data Pipelines

One of the most important features of Pachyderm is its Git-like version control system. This makes it easy to track changes to the pipeline over time, and to roll back changes if necessary.

To version a data pipeline, simply commit the changes to the pipeline's Git repository. Pachyderm will automatically track the changes and create a new version of the pipeline. You can then view the history of the pipeline and roll back to any previous version if necessary.

In addition to version control, Pachyderm also provides a number of other features for managing data pipelines. These features include:

- Pipeline branching and merging
- Pipeline testing and validation
- Pipeline deployment and monitoring

Scaling and Securing Data Pipelines

As your data pipelines grow in complexity, you will need to scale them to meet the demands of your organization. Pachyderm provides a number of features for scaling data pipelines, including:

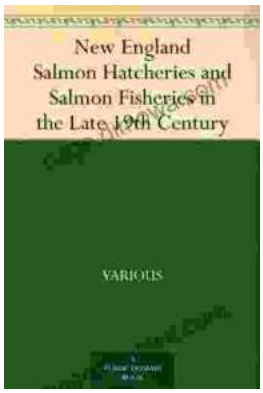
- Horizontal scaling
- Vertical scaling
- Elastic scaling

In addition to scaling, you will also need to secure your data pipelines to protect them from unauthorized access. P

Reproducible Data Science with Pachyderm: Learn how to build version-controlled, end-to-end data pipelines using Pachyderm 2.0 by Svetlana Karslioglu

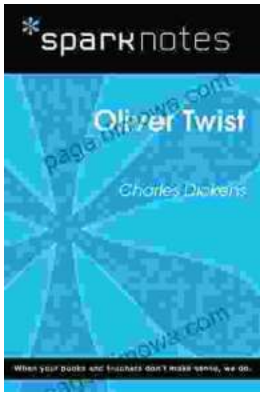
★★★★★ 5 out of 5

Language	: English
File size	: 11815 KB
Text-to-Speech	: Enabled
Screen Reader	: Supported
Enhanced typesetting	: Enabled
Print length	: 364 pages
Paperback	: 200 pages
Item Weight	: 11.2 ounces
Dimensions	: 5.5 x 0.5 x 8.5 inches



Unveiling the Legacy of New England Salmon Hatcheries and Salmon Fisheries in the Late 19th Century

Journey back in time to the late 19th century, a period marked by significant advancements in the field of fisheries management and aquaculture. New...



Embark on a Literary Adventure with Oliver Twist: A Comprehensive SparkNotes Guide

Unveiling the Complex World of Oliver Twist: A Captivating Journey In the shadowy labyrinth of 19th-century London, a young orphan named Oliver Twist embarks on a...