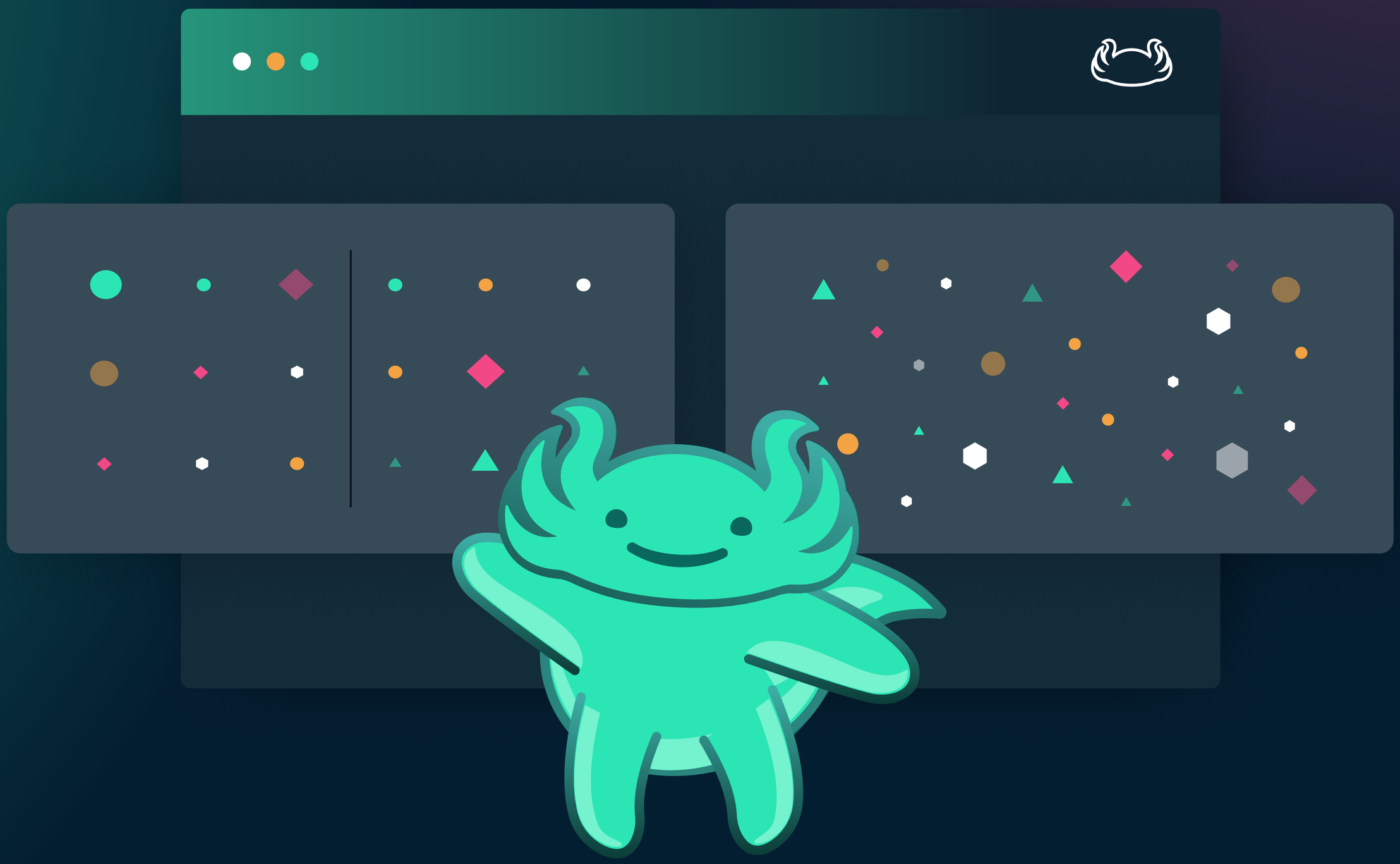
lakeFS

Managing Structured and Unstructured Data with lakeFS



lakeFS

Introduction

As many organizations move to managing combined structured and unstructured data, they face a challenge with their master data or the format of their data storage.

In this article, we'll cover:

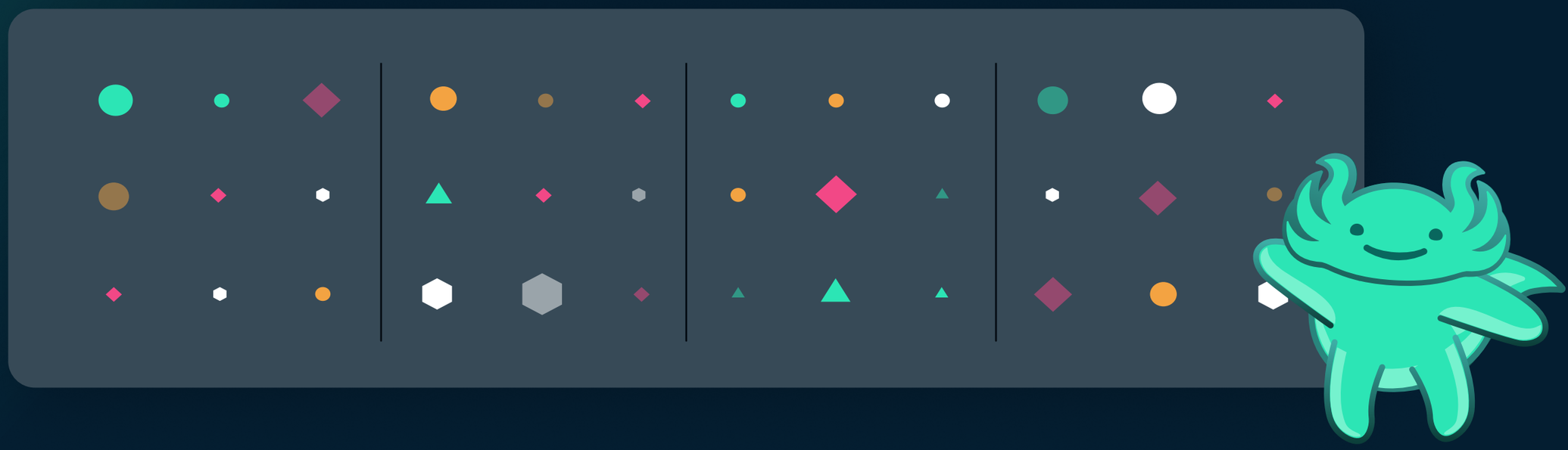
- The difference between structured and unstructured data
- Common use cases
- Common challenges
- Benefits of combining structured and unstructured data

Read the full article here

lakeFS

What is Structured Data?

Structured data refers to data that is organized into a specific format or structure, making it easy to store, access, and analyze. This type of data is typically stored in a database or data warehouse and has a clear and defined schema or set of rules for how the data is organized. Structured data can be easily transformed into numerical data that can be fairly easily used to train and evaluate machine learning models.



lakeFS

Examples of Structured Data

- Tabular: data that is organized in rows and columns, such as a spreadsheet or a database table
- Relational: data that is organized in tables with relationships between them
- Time series: data that is collected over time and organized in chronological order
- Graph: data that is represented in a graph or network structure, with nodes and edges connecting different data points
- Spatial: this is data related to geographic locations, such as addresses, GPS coordinates, and maps

lakeFS

What is Unstructured Data?

Unstructured data refers to any data that does not have a predefined data model or organizational structure. It does not conform to a specific data model or schema, and it typically includes data that is not easily searchable or analyzed using traditional data processing techniques.

Unstructured data can be more challenging to manage and analyze compared to structured data, which follows a specific format and schema. However, unstructured data can also provide nearly unlimited opportunities for companies to develop advanced insights and algorithms

lakeFS

Examples of Unstructured Data

- Textual: such as emails, social media posts, and documents in various formats like PDF, Word and HTML
- Multimedia: such as images, audio and video files
- Sensor: such as data from IoT devices, GPS, LiDAR, and RFID
- Streaming: such as data from social media feeds, online news sources and web logs
- Graph: such as data from social networks, recommendation systems, and web page links

lakeFS

Structured Data Use Cases

Structured data is typically used to store data that can be organized into a well-defined schema or format.
Here are some examples:

- Customer information such as name, address, email, phone number, etc.
- Product information such as product name, SKU, price, description, and other attributes.
- Inventory data such as product stock levels, warehouse locations, and other inventory details.
- Log data such as server logs, application logs, and other system logs.
- Sensor data such as temperature, humidity, pressure, and other environmental data.

lakeFS

Unstructured Data Use Cases

There is a variety of use cases that require a broad use of unstructured data for algorithms and analytics development.
Some examples include:

- Customer sentiment analysis such as social media posts, product reviews, and NLP techniques.
- Fraud detection such as transaction records, email communication, web logs, etc.
- Content recommendation such as user preferences, browser history and social media activity.
- Image and video analysis such as computer vision techniques for object detection, people detection, etc.
- Medical research such as medical records, research papers, and clinical trial data.

lakeFS

Challenges in Managing Structured and Unstructured Data Concurrently

Combining structured and unstructured data can be a challenge for organizations due to the inherent differences between the two types of data.

The most common challenges include:

1. **Appropriate master data**
   Creating appropriate master data for structured and unstructured data is challenging due to the lack of structure in unstructured data.

lakeFS

Challenges in Managing Structured and Unstructured Data Concurrently

2. Metadata extraction
   The challenge of metadata extraction lies in accurately and efficiently "wrapping" unstructured data properties to integrate it with structured data, and storing and managing this metadata effectively.

lakeFS

Challenges in Managing Structured and Unstructured Data Concurrently

3. **Data transformations**
   Data transformations present a challenge as unstructured data often requires content-based analysis, necessitating reliable and accurate tools to transform and extract data properties for proper classification.

4. **BI tools adaptation**
   Due to the complexity of extracting and accurately managing appropriate master data and metadata, adapting BI tools to support both structured and unstructured data poses a great challenge.

lakeFS

Bridging the Gap for Structured and Unstructured Data with lakeFS

lakeFS is a data version control system for data lakes that can help better manage both structured and unstructured data, by providing a united platform for managing different data types of data stored in a data lake and enabling users to work with these data types using familiar Git-like commands and workflows.

These are some examples of how to combine your structured and unstructured data management using lakeFS

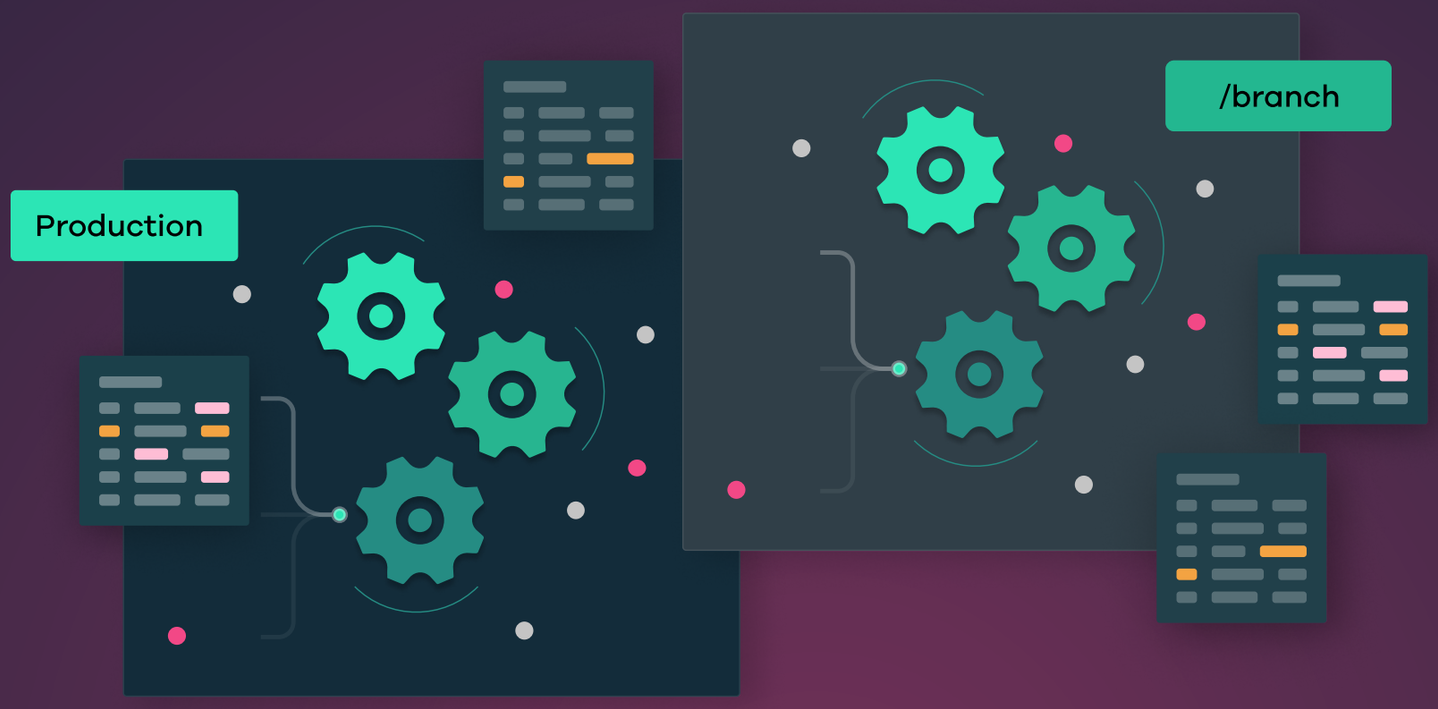
- Data versioning
- Data lineage
- Metadata management
- Data governance

lakeFS

Data Versioning with lakeFS

lakeFS allows users to version their data, making it easier to track changes, collaborate on data projects, and ensure data integrity.

This is particularly useful for unstructured data, which may change frequently and requires careful versioning to maintain consistency and accuracy.

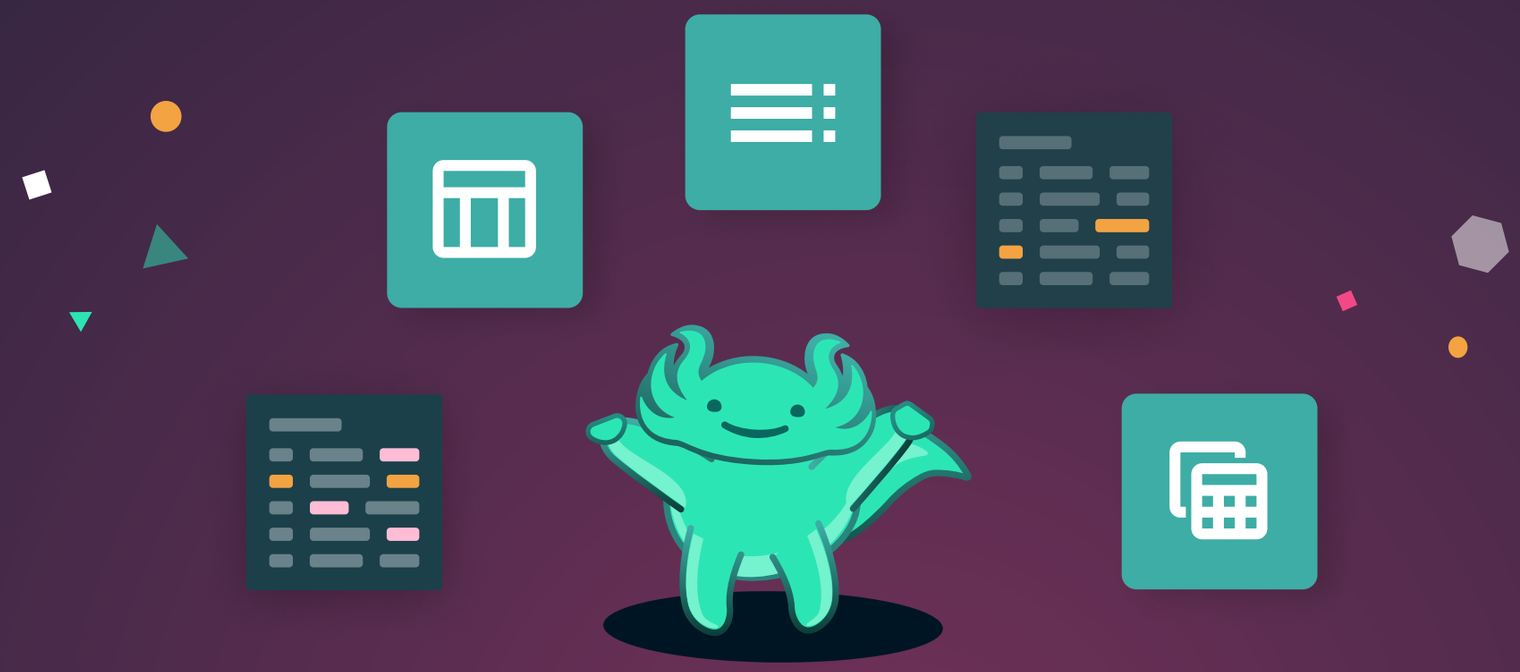


lakeFS

Data Lineage with lakeFS

lakeFS allows users to version their data, making it easier to track changes, collaborate on data projects, and ensure data integrity.

This is particularly useful for unstructured data, which may change frequently and requires careful versioning to maintain consistency and accuracy.

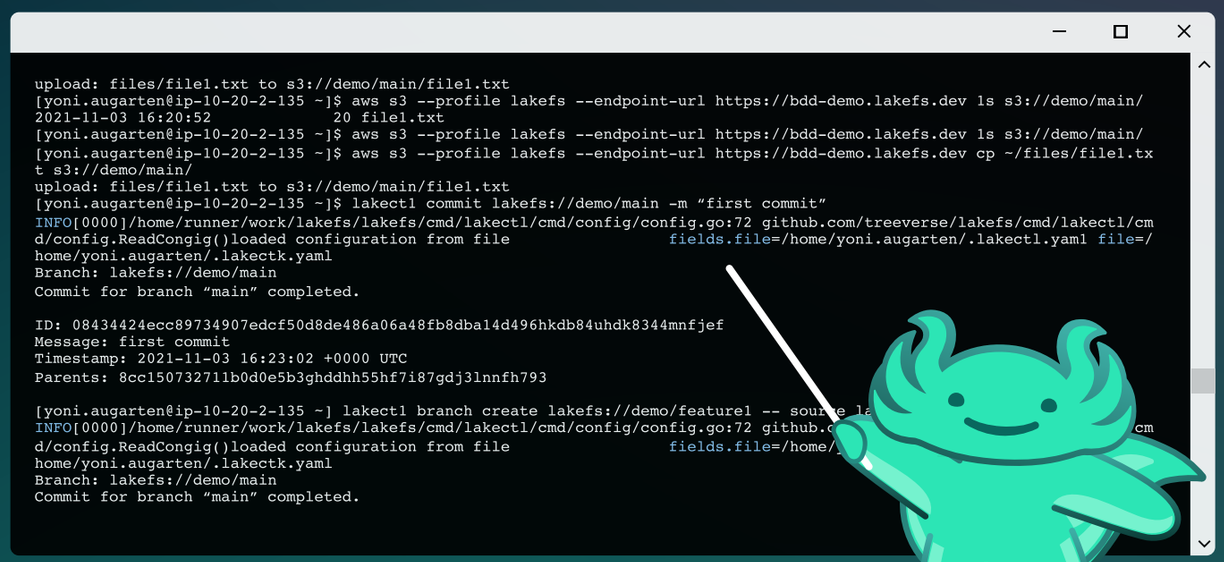


lakeFS

Metadata Management with lakeFS

lakeFS allows users to version their data, making it easier to track changes, collaborate on data projects, and ensure data integrity.

This is particularly useful for unstructured data, which may change frequently and requires careful versioning to maintain consistency and accuracy.



lakeFS

Data Governance with lakeFS

lakeFS provides a centralized platform for managing data governance policies, such as access control, auditing, and compliance.

This can help ensure that data is properly secured, tracked, and managed, regardless of its format or location.

lakeFS

Summary

While there are many challenges to combining structured and unstructured data management, with proper handling and tools, this can be accomplished.

To learn more about how lakeFS can help your organization manage the two types of data, visit lakefs.io