Back to Blog

Data Engineering

Data Products: The Definitive Guide

The lakeFS team | August 1, 2022

Table of Contents

Introduction

Have you ever heard the phrase “data is the new oil” before? Of course you have. In this age, data is becoming a valued currency of its own, as more and more businesses take advantage of and derive value from data. Using data, a company can plan ahead and create strategies for the future to come. These plans may include drastically improving the efficiency of a business’s marketing towards its customers or boosting the company’s product quality to reach a grander audience. Whether you want to improve your sales, product quality, or marketing strategies, utilizing comprehensive data about your surrounding environment and consumers is the way to go.

The applications companies build in order to collect, analyze and present data, and to derive insights from it, are data products. The term is important as it implies best practices for delivering data should apply when creating data intensive applications.

What Are Data Products?

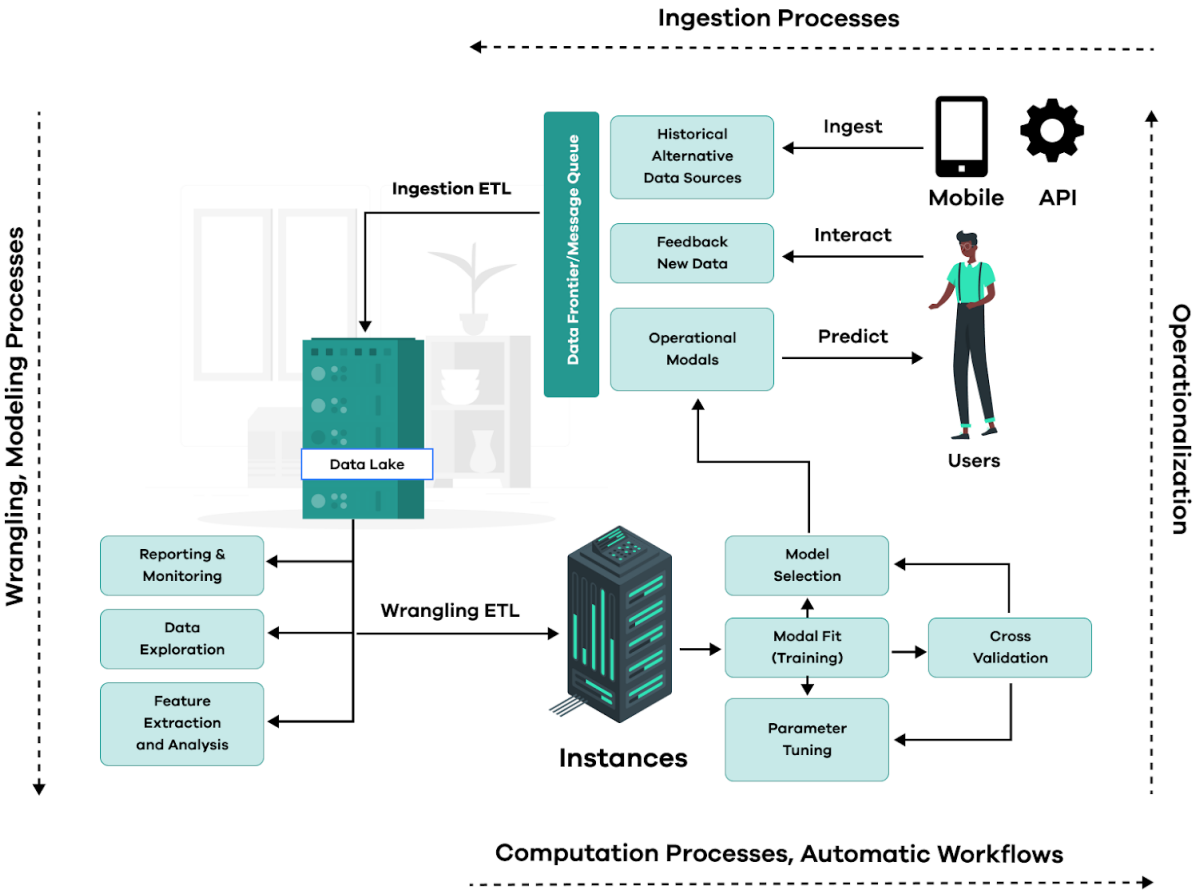

Figure: High level diagram of Data Operationalization, processes and workflows

Source

A **data product** is any tool or application that processes data and generates insights. These insights are aimed at helping businesses make better decisions for the future.

Stored data can then be sold or consumed by users internally, or by customer organizations which then process the data as needed. Take the Netflix movie recommendations algorithm as a real-life data product example, Depending on the data previously stored regarding a given viewer's most liked and disliked movies or shows, further suggestions of similar entertainment can be presented for the viewer.

Examples of Data Products

Figure: Lifecycle of a data team: Monitor, Experiment, Implement, Deploy and so on.

Source

When in possession of highly valuable data, companies can use certain techniques to process such data further depending on their desired goal. This processing allows for the extraction of all the necessary customer information for major future uses. While data products have multiple categorizations, they are most commonly characterized by how they process data and what type of service they provide.

A) Breakdown of Data Services :

Raw data

This is the most basic form of data collected into the system. If the data is raw, it means none of the data has been processed or used yet. Companies can still process this data to extract more value from it.

Example of raw data: All the purchases at a store for a given period of time.

Derived data

This is a more customized version of raw data, where further steps are taken to make the raw data more understandable, like calculating the average or sum of a given attribute.

Example of derived data: Calculating the average purchase of a customer.

Algorithms

The data is processed by a dedicated algorithm. The purpose of an algorithm is to provide a final end result for data customers. Some of these algorithms can require running a machine learning model on the given data. From this category rise the following two data products – decision support and automated decision making.

Decision support

While decision support algorithms are capable of doing a good part of the processing, the digested answer is returned to the user to make inferences from processed data. That is, this requires further human intervention to work.

Example of decision support: GPS Navigating application.

Automated decision-making

This can be thought of as a do-it-all algorithm. These kinds of data products will give a final result without the need for external help from the user of the service.

Example of automated decision-making: A self driving car.

B) Types of Services Data Products Provide:

Automated decision making or data enhanced data products

This category includes data products that help make business decisions without human intervention. Algorithms that recommend products or services depending on the

customer's previous likes and dislikes come here. Some examples of such services are recommending shows to a Netflix viewer or recommending products on Amazon.

By studying the customer's older interest patterns, more customized product suggestions can be offered to each unique customer. It goes without saying that using such tactics bumps up future sales.

Data as a Service Products

Data as a service products offer a service that customers can subscribe to by connecting to a given API. These kinds of services are usually created by running different machine learning models on the given data. Again, it is worth noting that such services are usually sold to customers to be embedded in their applications and websites.

Some examples include weather applications or geography data services. They can provide information like atmospheric pressure, weather conditions, visibility distance, relative humidity, and precipitation in different units and in different parts of the world. They can also give information about geographic boundaries and maps.

Data as Insights

In the final category of data products, data does not generate revenue on its own, but it is used to improve business insights and in turn the sales of a product. In this category, the data is not viewed by customers.

Data insight products allow companies to optimize processing to improve performance; uncover new markets, products, or services to add new sources of revenue; better balance risk vs. reward to reduce loss, and deepen the understanding of customers to increase loyalty and lifetime value. One example of a data insight service is Facebook collecting data on customers for future offers.

Facebook can use its stored data to gain insights about your lifestyle. Customized offers can be sent to each unique customer depending on these insights.

Data Product Build Strategy

In this part, we will explain **Zhamak's Principles for Data Products** and **data product architectures**.

1. Zhamak's Principles for Data Products

So, who is Zhamak Dehghani? Zhamak Dehghani is the esteemed founder of the concept of data mesh. She defines data mesh as "a sociotechnical approach to share, access and manage analytical data in complex and large-scale environments—within or across organizations".

Zhamak states that to define a product as a data product, the data product being used should check off the main principles in terms of the data being used. These principles are extremely important for data products to be eligible for general use. Some of these principles as stated by Zhamak include:

Discoverability: The discoverability of a data product is of utmost importance. Information such as the data owner, source of origin, location, and quality metrics are necessary for a discoverable product. Note that such information can improve the data's trustworthiness and security.

Data must hold value themselves: The stored data should be valuable to other organizations. It is better to not collect any data that will not be sold or used for future profits because storing such data comes with its own costs of storing and maintaining.

Trustworthy and trustful: As businesses rely on data to make significant decisions, the data needs to be highly trustworthy.

It goes without saying that any data collected should be checked for its trustworthiness as a source to be used. So, how can you check for trustworthiness of your data? Continuous checks on your datasets will be required. These checks will examine the data quality and accuracy for future use.

Understandable data: It is necessary to understand how data is organized and structured to be easily used. Understanding such syntax will allow you to extract the required data from a given database table with no need for further outside support or help.

Accessible: As most of the data gathered is more than likely going to be processed in one way or another, how such data can be accessed by different users is of utmost importance.

Different data formats will be required for different data usages and processing. For example, data that is used to run a machine or deep learning models will require different formats than data required to build a graph to extract common trends or patterns among users.

Governed security access to the product: Access to a data product must be governed by a certain entity as the given data must not be accessible to everyone.

Addressable: It is as simple as it seems—each data product requires a unique address for users to easily identify and use it.

Data Product Architecture

As you can imagine, the main components of any data product include the code, the valuable data stored with any additional data describing it, and the infrastructure of the product.

Here are the components of a data product:

Code: Code includes the code for data pipeline consuming, APIs that provide data access and explanation for the provided schema, and the code responsible for enforcing data access control, compliance, and provenance.

Data and metadata: This is the most important part of our structure. Here, the actual data sets that are going to be used are stored. The metadata will be provided here as well. The

metadata should include, documentation, semantic and syntax declaration, and quality metrics.

Infrastructure: It is used to run and deploy the data product while storing the required data.

Why Think of Your Internal Data Operations as a Product?

1. Product Management Methodology

Regardless of the delivery methodology you use or the agile development flavor you prefer, every product has a product manager who is responsible for gathering the needs of the customers, defining the functionality to deliver, and prioritizing the work for the development teams.

Having such a function for your data products is critical. When those products are part of the business of the company, that is more common, but when the data product is for internal use of insights to improve the business, that is seldom done. Data engineering teams are working without a guiding hand from the business stakeholders, so requirements are of low quality, and prioritization is often missing. Once you treat your internal data operations as a product, you assign a product manager.

2. Engineering Best Practices

Data engineering teams should be able to leverage application development best practices. To do so, they need the tools and processes that allow them CI/CD for the data products that include their code (devops practices cover that), but also their data and their infrastructure should be part of the CI/CD process. The data should be versioned controlled just like the code, and the infrastructure should be easily structured using popular technologies such as K8s. The world of data is moving in this direction, allowing more and more data technologies on K8s, and version control engines for data.

3. Measure and improve

Product delivery is measured to ensure velocity of development teams, Feature implementation vs. planning, and quality measured by bug fixes. With data products all

those measures still apply, but data quality, in all its aspects, must also be monitored, and SLA agreements should include not only availability of the interface, but also freshness of the data.

Future of Data Products

With the total amount of data created and consumed reaching ***64.2 zettabytes*** in the year 2020 and with a future estimate of **50,000–500,000 Zettabytes** of data being generated by the year 2050, bigger and more diverse data sets will need to be created in the near future. These data sets will allow users to identify patterns in the data set more easily, and continuously improve data products.

Of course, we can not forget to mention the new doors that artificial intelligence and data science branches are opening for new data products. As different organizations are starting to acknowledge the merits of rising technologies like artificial intelligence, machine learning, and deep learning, companies such as Amazon, Google, Alibaba, and Microsoft, are investing a huge amount of capital into their own artificial intelligence departments to improve such concepts even further.

Conclusion

Some of the main concepts of any data product include its types, examples, principles, components, and much more. With data changing the way we look at products and services, who knows what new things can occur in the coming year or two.

Now, you should have a better understanding of what data products are just in case you plan on creating, buying, or working with such concepts in the near future.

Subscribe

Talk to a lakeFS engineer

Email*

Let's talk

Cloud

Docs

Blog

Community

Careers

Contact Us