Basic Introduction to Data Science Pipeline

Pranshu Sharma — Published On August 16, 2022 and Last Modified On September 8th, 2022

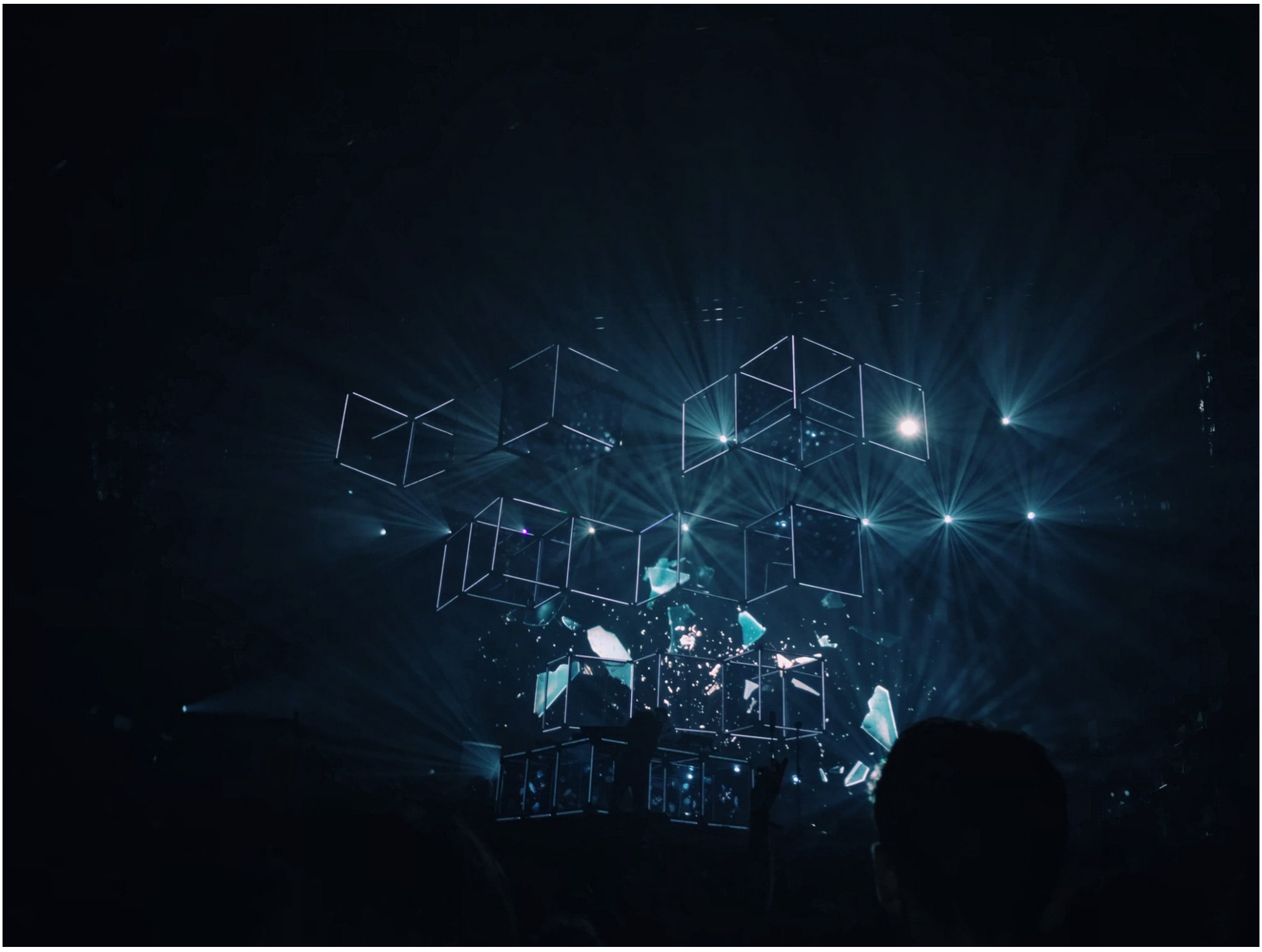
Beginner Data Engineering Data Science Data Warehouse Database MLops

This article was published as a part of the Data Science Blogathon.

Introduction

The Data science pipeline is the procedure and equipment used to compile raw data from many sources, evaluate it, and display the findings in a clear and concise manner. Businesses use the method to get answers to certain business queries and produce insights that can be used for various business-related planning.

Due to the ever-growing complexity and volume of enterprise data, as well as its crucial role in decision-making and long-term planning, organizations are investing in the Data science pipeline-related technologies necessary to extract useful business insights from their data assets in order to use for planning and other business approaches



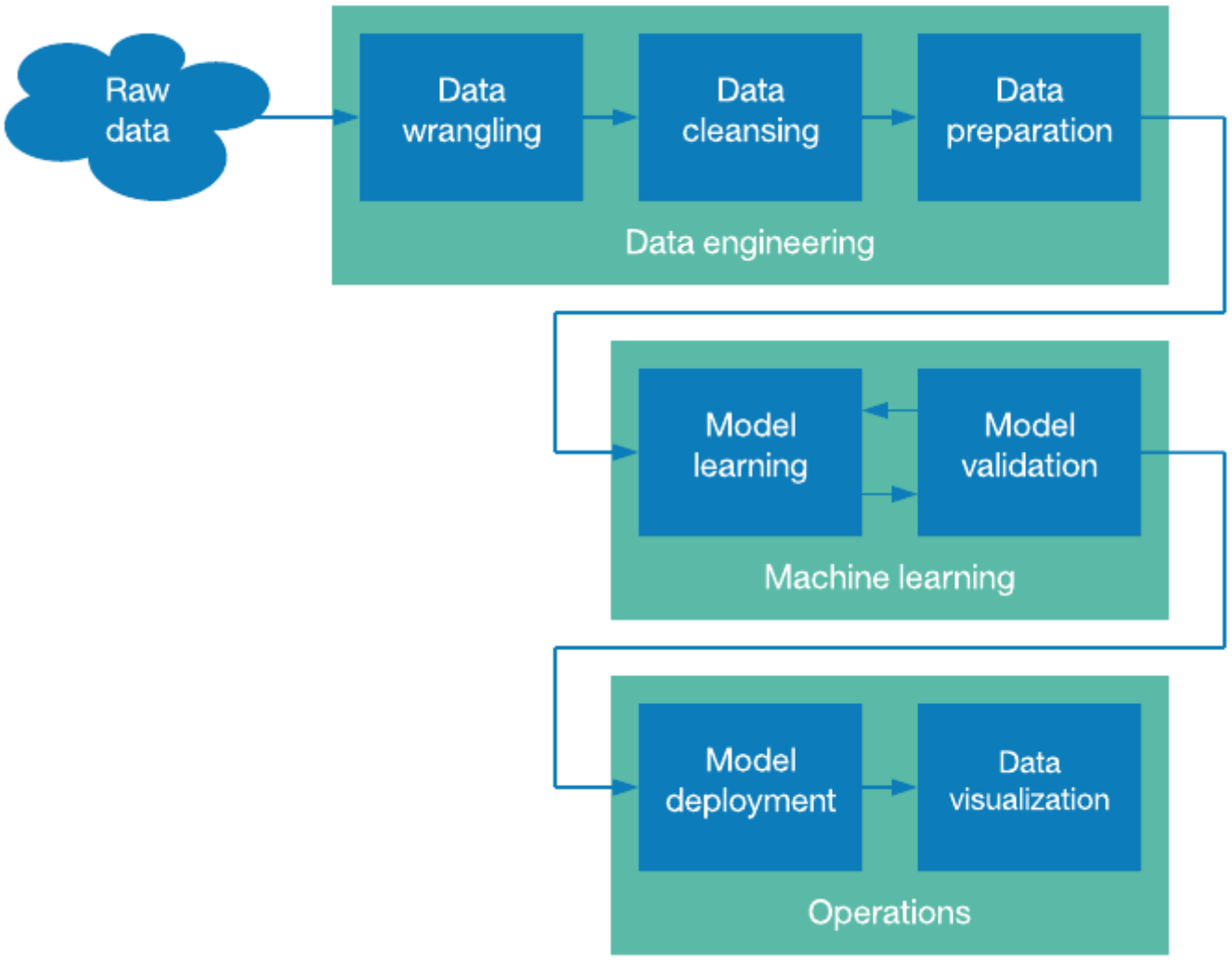
Source: Unsplash

What is meant by a data science pipeline?

A data science pipeline is a process collection that transforms raw data into useful solutions to business issues. Pipelines for data science streamline data movement from source to destination, allowing you to make better business decisions.

Basic Introduction to Data Science Pipeline

results comprehensibly. Companies use the method to answer particular business issues and derive actionable insights from real-world data. In simple terms, a data science pipeline is a sequence of operations that converts raw data from diverse sources into a comprehensible format so that it may be stored and analyzed.



Source: IBM Developer

Why is the data science pipeline noteworthy?

The data science pipeline is the key to extracting insights from ever-larger and more complicated information. Teams must depend on a process that disintegrates datasets and offers meaningful insights in real-time as the amount of data available to enterprises continues to grow.

- The data science pipeline makes data analysis and handling of large chunks of data easier
- Smooth management of various tasks like collecting data from several teams, cleansing it, and displaying it in a readily understandable format.
- It enables you and your team to make data-driven decisions quickly.
- We can bypass the time-consuming and error-prone procedure of traditional data collection.
- It allows consumers to explore deeper into data at a more granular level.

Working of a Data Science Pipeline

Having precise queries is critical before pushing raw data through the pipeline. This allows users to concentrate on the relevant facts to gain the necessary insights.

There are various steps in the data science pipeline, including

1. **Obtaining information**

   This is where data is collected and processed from internal, external, and third-party sources into a useful format (XML,

Basic Introduction to Data Science Pipeline

2. **Data cleansing**

   This is the process's most time-consuming step. Anomalies in data, such as duplicated parameters, missing values, or pointless data must be cleaned before a data visualization can be created.

   Data cleansing can be classified into two types:

   a) Examining data to look for errors, missing numbers, or entries that have been corrupted.

   b)Cleaning data entails filling in gaps, correcting errors, deleting duplicates, and discarding obsolete records or data.

3. **Data exploration and modeling**

   After the data has been completely cleaned, data visualization tools and charts can be utilized to detect patterns and values. This is where artificial intelligence (AI) techniques come into play. You can detect patterns and apply specific rules to data or models using classification accuracy, confusion matrix, logarithmic loss, etc.

4. **Data interpretation**

   This stage aims to uncover and link insights with your data findings. You can then use charts, dashboards, or reports/presentations to present your results to corporate leaders or coworkers.

5. **Revision of the information**

   It's critical to reassess your model regularly as your business requirements evolve and new data becomes available.

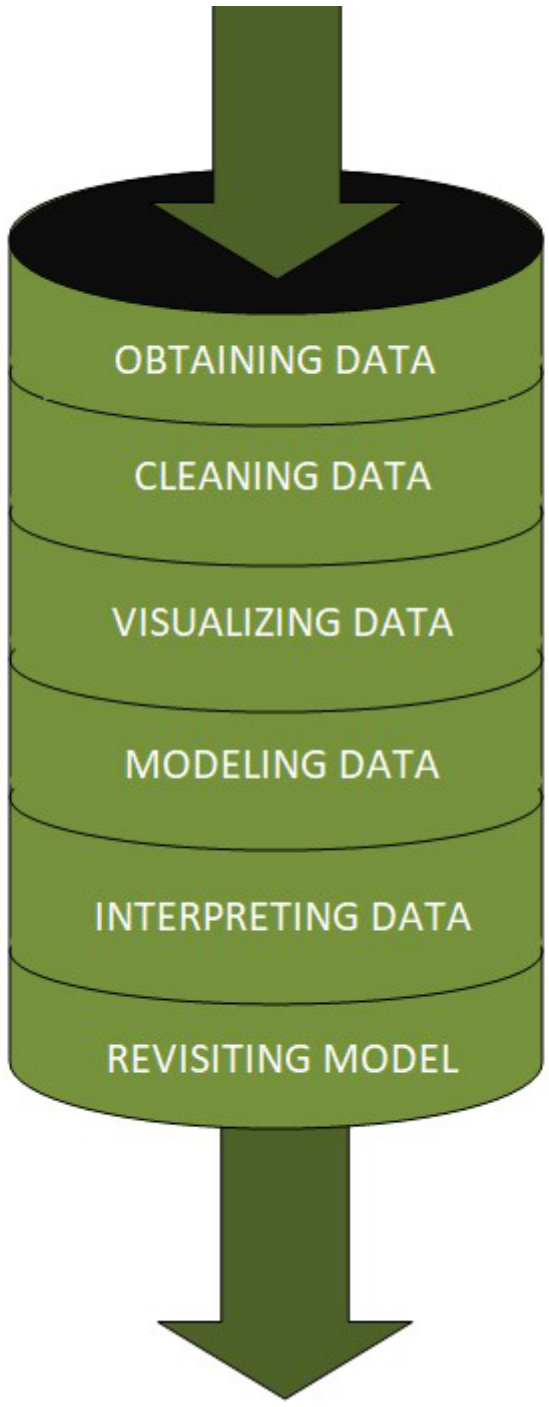
Stages of Data Science Pipeline

The following are the main steps in a data science pipeline:

1. Data Engineering consists of the collection, cleansing, and preparation
2. Computer-assisted learning consists of collection model learning and model validation
3. The result, which consists of collection model deployment and data visualization

However, establishing the business challenges, you need the data to solve, and ***the data science methodology is the first step in building a data science pipeline***. Formulate the questions you need to be answered, and machine learning and other techniques will offer you answers you can use.

Basic Introduction to Data Science Pipeline



Source: Geeksforgeeks.org

The following are the steps in a data science pipeline:

- Data collection includes identifying data sources and extracting data from those sources into formats that may be used.
- ETL (Extraction, Transformation, and Loading) may be used in data preparation.
- Machine learning is deployed to detect patterns and apply the rules to data using algorithms, which are subsequently validated on sample data in data modeling and validation of the model.
- Model deployment, which entails deploying the model to both old and new data
- Reviewing and upgrading the model in response to shifting business needs

Benefits

Following are the benefits of Data Science Pipelines

1. **The pattern that can be replicated**

   Individual pipes are patterns in a larger architecture that may be recycled and reused for new data flows when data processing is viewed as a network of pipelines.
2. **Integration of new data sources takes less time.**

   Having a common concept and techniques for how data should pass through analytics systems makes it simpler to plan for integrating new data sources and minimizes the time and expense of integrating them.
3. **Data quality assurance**

   Understanding data streams as pipelines that need to be regulated and useful to end-users increases data quality and minimizes the chances of pipeline breakdowns going undiscovered.

Basic Introduction to Data Science Pipeline

With repetitive patterns and consistent knowledge of tools and architectures, security is baked in from the start. Good security procedures can easily apply to new dataflows or data sources.

5. **Build in stages**

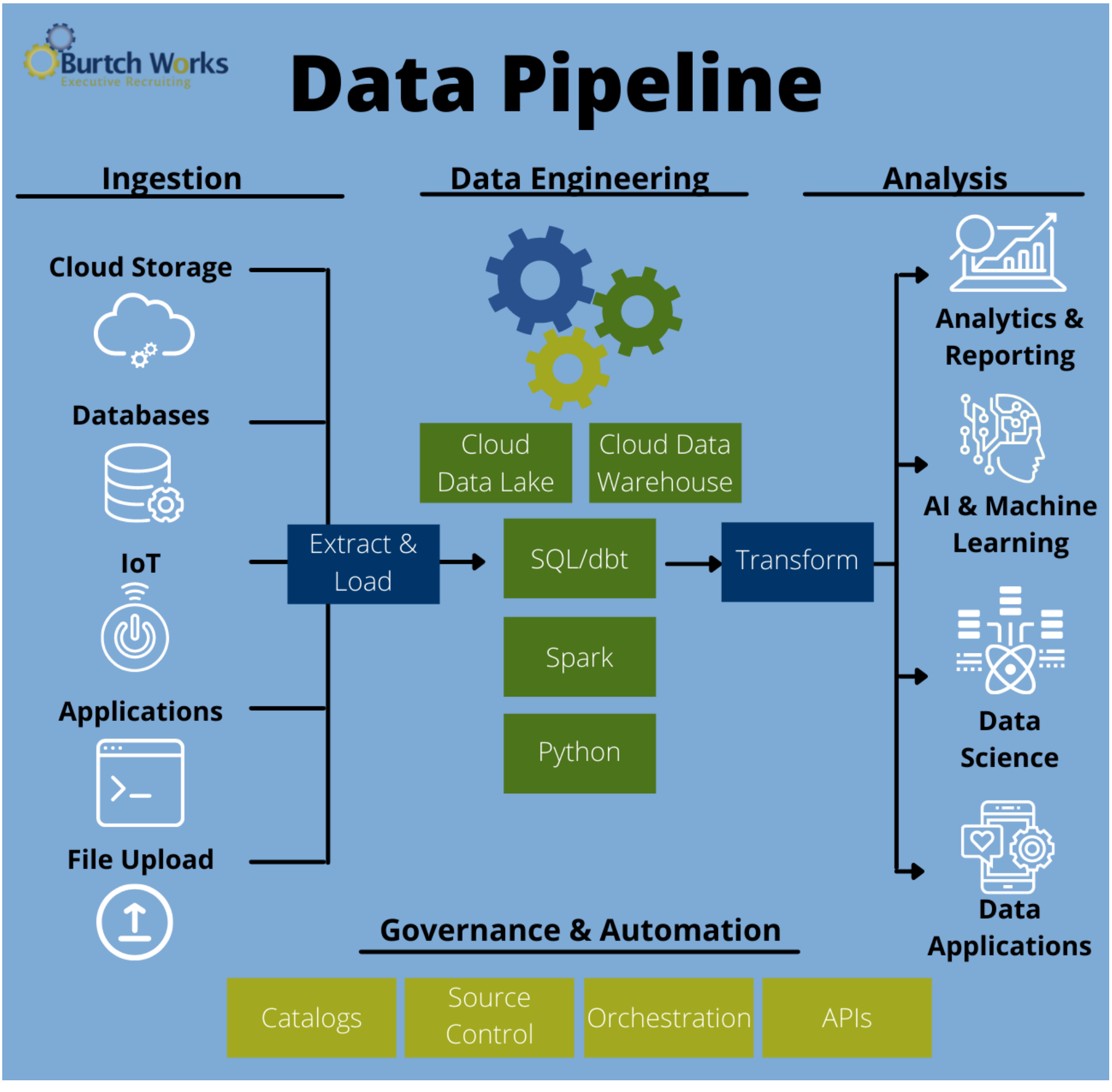
   When you think of your dataflows as pipelines, you can scale them up gradually. You can get started early and achieve benefits immediately by starting with a modest controllable segment from a data source to a user.

6. **6 Agility and flexibility**

   Pipelines give a structure for responding dynamically to modifications in the sources or the needs of your data users. Extensible, modular, and reusable Data Pipelines are a bigger topic in Data Engineering that is very significant.

Features

A well-designed end-to-end data science pipeline can find, collect, manage, analyze, model, and transform data to uncover possibilities and create cost-effective business operations.



Source: Burtch Works

Basic Introduction to Data Science Pipeline

The finest data science pipelines contain the following features to accomplish this:

- Data processing that is both continuous and expandable
- Elasticity and agility afforded by the cloud
- Access to data on a large scale and the capacity to self-serve
- Disaster recovery and high availability

How do various industries use the data science pipeline?

Regardless of the industry, the data science pipeline is beneficial to teams. The following are some instances of how different teams have used the process:

1 Risk analysis: Risk analysis is a method financial institutions use to make sense of enormous amounts of unstructured data to determine where potential hazards from rivals, the market, or consumers are located and how they might be avoided.

Organizations have also used Domo's (a software company) DSML tools and model findings for proactive risk mitigation and planning. Medical experts make use of data science to help them conduct research. Machine learning algorithms are used in one study to aid in the research of how to increase picture quality in MRIs and x-rays.

Domo's (a software company) Natural Language Processing and DSML have been used successfully by companies outside the medical field to predict how specific actions affect the customer experience. This allows people to anticipate dangers and maintain a favorable experience.

2 Forecasting: Data science pipelines are used by the transportation industry to estimate the impact of development or other road projects on traffic. This also aids experts in formulating effective solutions.

Domo's(a software company) DSML solutions have also shown to forecast future product demand for other business teams effectively. The platform includes multivariate time series modeling at the SKU level, allowing them to appropriately plan across the supply chain and beyond.

What will the future data science pipeline look like?

The data science pipeline is essential to extracting insights from ever-larger and more detailed information. Organizations must depend on a methodology that disintegrates datasets and offers meaningful insights in real-time as the amount of available data to enterprises continues to grow.

The data science pipeline's agility and speed will only improve as new technology arrives. The method would become smarter, more agile, and more flexible, allowing teams to dig a little deeper into data than ever before.

Conclusion

So in this article, we studied Data Science Pipelines. Some of the key takeaways are:

- Working of data science pipelines.
- Various stages in data science pipelines.
- Various features of data science pipelines
- Real-Time usage by industries

Data science isn't about working with various machine learning algorithms; it's about creating solutions using them. It's also critical to ensure that your pipeline is strong from beginning to end and that you identify specific business problems to provide precise solutions

Basic Introduction to Data Science Pipeline

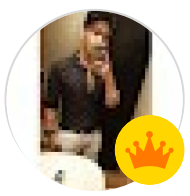
My name is Pranshu Sharma, and I am a Data Science Enthusiast. Thank you so much for taking your precious time to read this blog. Feel free to point out any mistake(I'm a learner, after all) and provide respective feedback or leave a comment.

Feedback:Email: pranshu453@gmail.com

The media shown in this article is not owned by Analytics Vidhya and is used at the Author's discretion.

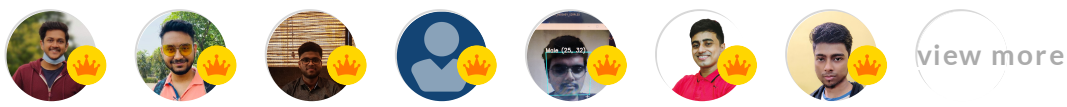
blogathon data science Data Science pipelines machine learning

About the Author



Pranshu Sharma

Our Top Authors



Download

Analytics Vidhya App for the Latest blog/Article

Previous Post

Different Ways of Loading Data using Python

Next Post

Database Normalization | A Step-by-Step Guide with Examples

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment

Name*

Email*

Basic Introduction to Data Science Pipeline

Top Resources

H1B Visa Data Analysis: Unveiling Patterns of H1B Visa Approval

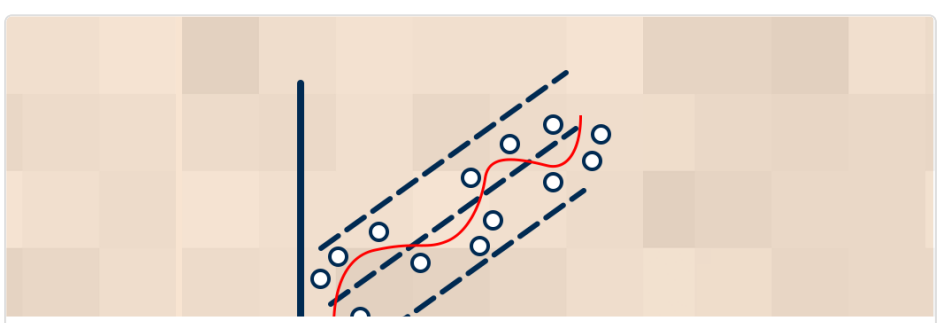
Hareeharan Elangovan - JUN 22, 2023

Building Language Models: A Step-by-Step BERT Implementation Guide

Kajal Kumari - JUN 29, 2023

Understand Random Forest Algorithms With Examples (Updated 2023)

Sruthi E R - JUN 17, 2021



Everything you need to Know about Linear Regression!

KAVITA MALI - OCT 04, 2021