

[Home](#) > [Blog](#) > [Data Engineering](#)

14 Essential Data Engineering Tools to Use in 2024

Learn about the top tools for containerization, infrastructure as code (IaC), workflow management, data warehousing, analytical engineering, batch processing, and data streaming.

[Contents](#)

Mar 8, 2024 · 10 min read

**Abid Ali Awan**

Certified data scientist, passionate about building ML apps, blogging on data science, and editing.

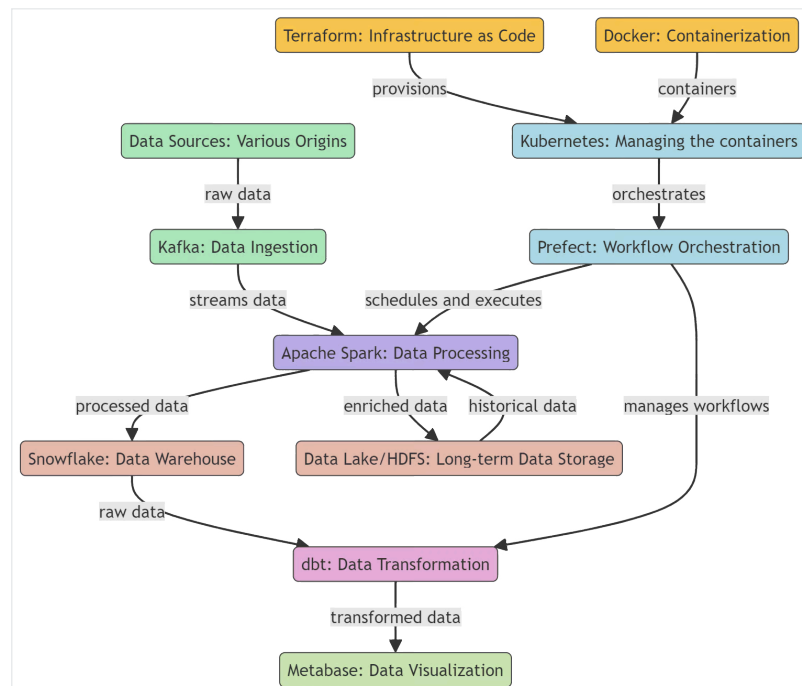
TOPICS

[Data Engineering](#)

Data engineers are responsible for creating data pipelines that can ingest, process, and deliver data to various endpoints, such as databases, data warehouses, and analytics platforms. By building and maintaining these data pipelines, data engineers enable data scientists and analysts to access real-time data for analysis and decision-making.

Modern data engineers are expected to perform even more tasks. They must also maintain and deploy data solutions, manage workflows, oversee data warehouses, transform and visualize data, and use various batch processing and streaming tools to optimize, ingest, and process different types of data.

Discover what data engineering is, how it differs from data science, its scope, and ways to learn it by reading our guide, [What is Data Engineering?](#)



In this post, we will learn about the essential tools that are popular and sometimes necessary for data engineers. These tools are used for data ingestion, processing, storage, transformation, and visualization. Additionally, we'll look into tools for containerization and workflow management.

Learn essential data engineering skills by reading our [How to Become a Data Engineer](#) blog.

Containerization Tools

Containerization tools provide a standardized way to package, distribute, and manage applications across different environments. It ensures consistency, scalability, and efficiency of data engineering workflows.

1. Docker

Docker is a popular containerization platform that is often used in data engineering to develop, ship, and run data tools and applications. It provides a lightweight, portable, and consistent way to package and deploy data tools and applications, making it an ideal choice for data engineers.

Docker can be used to create and manage containers for various data tools, such as databases, data warehouses, data processing frameworks, and data visualization tools. Check out our [Docker for Data Science](#) tutorial to learn more.

2. Kubernetes

Kubernetes is an open-source platform for automating the deployment, scaling, and management of containerized applications, including those built using Docker.

Docker is a tool that can be used to package data processing applications, databases, and analytics tools into containers. This ensures consistency in environments and isolates applications. Once the containers are created, Kubernetes steps in to manage them by handling their deployment, scaling based on workload, and ensuring high availability.

Learn more about [Containerization using Docker and Kubernetes](#) in a separate article.

Infrastructure as Code Tools

Infrastructure as Code (IaC) streamlines the deployment and maintenance of cloud infrastructure by utilizing general-purpose programming languages or YAML configurations. This approach fosters the creation of consistent, repeatable, and automated environments, facilitating smoother transitions across development, testing, and production phases.

3. Terraform

Terraform is an open-source infrastructure as code (IaC) tool created by HashiCorp. It enables data engineers to define and deploy data infrastructure, such as databases and data pipelines, in a consistent and reliable manner using a declarative configuration language, which describes the desired end state of the infrastructure rather than the steps needed to reach that state.

Terraform supports version control, resource management via code, team collaboration, and integration with various tools and platforms.

4. Pulumi

Pulumi is an open-source infrastructure as a code tool that allows developers to create, deploy, and manage cloud infrastructure using general-purpose programming languages such as JavaScript, TypeScript, Java, Python, Go, and C#. It supports a wide range of cloud providers, including AWS, Azure, GCP, and Kubernetes.

The Pulumi framework provides a downloadable Command-Line interface (CLI), Software development kit (SDK), and Deployment engine to deliver a robust platform for provisioning, updating, and managing cloud infrastructure.

Become a Data Engineer

Become a data engineer through advanced Python learning

Start Learning for Free

Workflow Orchestration Tools

Workflow orchestration tools automate and manage the execution of complex data processing workflows, ensuring tasks are run in the correct order while managing dependencies.

5. Prefect

Prefect is an open-source workflow orchestration tool for modern data workflows and ETL (extract, transform, load) processes. It helps data engineers and scientists automate and

manage complex data pipelines, ensuring data flows smoothly from source to destination with reliability and efficiency.

Prefect offers a hybrid execution model that merges the advantages of cloud-based management with the security and control of local execution. Its user-friendly UI and rich API make it easy to monitor and troubleshoot data workflows.

6. Luigi

Luigi is an open-source Python package that helps you build complex data pipelines of long-running batch jobs. It was developed by Spotify to handle dependency resolution, workflow management, visualization, handling failures, and command line integration.

Luigi is designed to manage various tasks, including data processing, data validation, and data aggregation, and it can be used to build simple and sophisticated data workflows. Luigi can be integrated with various tools and platforms, such as Apache Hadoop and Apache Spark, allowing users to create data pipelines to process and analyze large volumes of data.

Data Warehouse Tools

Data warehouses offer cloud-based solutions that are highly scalable for storing, querying, and managing large datasets.

7. Snowflake

Snowflake is a cloud-based data warehouse that enables the storage, processing, and analytical querying of large volumes of data. It is based on a unique architecture that separates storage and compute, allowing them to scale independently.

Snowflake can dynamically adjust the amount of computing resources based on the demand. This ensures that queries are processed in a timely, efficient, and cost-effective manner. It is compatible with major cloud providers, such as **AWS, GCP, and Azure**.

Check out our **Introduction to Snowflake course** to explore this tool in more detail.

8. PostgreSQL

PostgreSQL is a powerful open-source relational database management system (RDBMS) that can also be used as a data warehouse. As a data warehouse, PostgreSQL provides a centralized repository for storing, managing, and analyzing large volumes of structured data from various sources.

PostgreSQL offers features such as partitioning, indexing, and parallel query execution that enable it to handle complex queries and large data sets efficiently.

Remember, a PostgreSQL data warehouse is a local solution that may not scale as well as some fully managed solutions. It requires more manual administration and maintenance compared to these solutions.

Learn more in our **Beginner's Guide to PostgreSQL**.

Analytics Engineering Tools

Analytics engineering tools streamline the transformation, testing, and documentation of data in the data warehouse.

9. dbt

dbt (data build tool) is an open-source command-line tool and framework designed to facilitate data transformation workflow and modeling in a data warehouse environment. It supports all major data warehouses, including Redshift, BigQuery, Snowflakes, and PostgreSQL.

dbt can be accessed through dbt Core or dbt Cloud. The dbt Cloud offers a web-based user interface, a dbt Cloud-powered CLI, a hosted environment, an in-app job scheduler, and integrations with other tools.

You can grasp the fundamentals of dbt with our **Introduction to dbt course**.

10. Metabase

Metabase is an open-source business intelligence (BI) and analytics tool that enables users to create and share interactive dashboards and analytics reports. It's designed to be user-friendly, allowing non-technical users to query data, visualize results, and gain insights without needing to know SQL or other query languages.

It offers easy setup, support for various data sources, a simple user interface, collaboration features, customizable notifications, and robust security for data exploration, analysis, and sharing.

Data analytics and dashboarding are part of data science. Learn about the differences between a Data Scientist and a Data Engineer by reading this article: [Data Scientist vs Data Engineer](#).

Batch Processing Tools

These data engineer tools enable efficient processing of large data volumes in batches, running complex computational tasks, data analysis, and machine learning applications across distributed computing environments.

11. Apache Spark

[Apache Spark](#) is a powerful open-source distributed computing framework designed for large-scale data processing and analysis. While it is commonly known for its ability to handle real-time streaming data, Spark also excels in batch processing, making it a valuable tool in data engineering workflows.

Apache Spark features Resilient Distributed Datasets (RDDs), rich APIs for various programming languages, data processing across multiple nodes in a cluster, and seamless integration with other tools. It is highly scalable and fast, making it ideal for batch processing in data engineering tasks.

12. Apache Hadoop

[Apache Hadoop](#) is a popular open-source framework for distributed storage and processing of large datasets. At the core of the Hadoop ecosystem are two key components: the Hadoop Distributed File System (HDFS) for storage and the MapReduce programming model for processing.

Apache Hadoop is a powerful and scalable tool for data engineers, offering cost-effective storage, fault tolerance, distributed processing capabilities, and seamless integration with other data processing tools.

Streaming Tools

Streaming tools provide a powerful way of building real-time data pipelines, enabling the continuous ingestion, processing, and analysis of streaming data.

13. Apache Kafka

[Apache Kafka](#) is a distributed event streaming platform designed for high-performance, real-time data processing and streamlining of large-scale data pipelines. It is used for building real-time data pipelines, streaming analytics, data integration, and mission-critical applications.

Kafka is a system that can handle a large amount of data with low latency. It stores data in a distributed and fault-tolerant manner, ensuring that data remains available even if hardware failures or network issues occur.

Kafka is highly scalable and supports multiple subscribers. It also integrates well with different data processing tools and frameworks, such as Apache Spark, Apache Flink, and Apache Storm.

14. Apache Flink

[Apache Flink](#) is an open-source platform for distributed stream and batch processing. It can process data streams in real-time, making it a popular choice for building streaming data pipelines and real-time analytics applications.

Flink is a data processing tool that provides fast and efficient real-time and batch data processing capabilities. It supports several APIs (including Java, Scala, and Python), enables seamless integration with other data processing tools, and offers efficient state management. Therefore, it is a popular choice for real-time analytics, fraud detection, and IoT applications due to its ability to process high-throughput data with low latency.

Conclusion

That concludes our list, but it doesn't have to be the end of your data engineering journey. Enroll in the [Data Engineer skill track](#) to learn about Python, SQL, database design, cloud computing, data cleaning, and visualization. Once you complete the track, you can take the [Data Engineer Career Certification](#) exam and become a certified professional Data Engineer.

Data engineers play a critical role in building and maintaining the data pipelines that feed analytics and decision-making across organizations. As data volumes and complexity

continue to grow exponentially, data engineers must leverage the right tools to ingest, process, store, and deliver quality data efficiently.

We've covered the top 14 data engineering tools for containerization, infrastructure provisioning, workflow automation, data warehousing, analytics engineering, batch processing, and real-time streaming.

If you're new to data engineering, start by learning Docker, Kubernetes, Terraform, Prefect, Snowflake, dbt, Apache Spark, Apache Kafka, and more.

After gaining proficiency in these tools, you should review [The Top 21 Data Engineering Interview Questions, Answers, and Examples](#) to prepare for your next career move.

Get certified in your dream Data Engineer role

Our certification programs help you stand out and prove your skills are job-ready to potential employers.

[Get Your Certification](#)



AUTHOR

Abid Ali Awan



As a certified data scientist, I am passionate about leveraging cutting-edge technology to create innovative machine learning applications. With a strong background in speech recognition, data analysis and reporting, MLOps, conversational AI, and NLP, I have honed my skills in developing intelligent systems that can make a real impact. In addition to my technical expertise, I am also a skilled communicator with a talent for distilling complex concepts into clear and concise language. As a result, I have become a sought-after blogger on data science, sharing my insights and experiences with a growing community of fellow data professionals. Currently, I am focusing on content creation and editing, working with large language models to develop powerful and engaging content that can help businesses and individuals alike make the most of their data.

TOPICS

Data Engineering

Start Your Data Engineer Journey Today!

📌 TRACK

Data Engineer

🕒 40hrs hr

Gain in-demand skills to efficiently ingest, clean, manage data, and schedule and monitor pipelines, setting you apart in the data engineering field.

[See Details →](#)

[Start Course](#)

[See More →](#)

Related

BLOG

Top 10 Data Science Tools To Use in 2024





BLOG

Top 3 Trends in Data
Infrastructure for 2021



BLOG

5 Essential Data Engineering
Skills

[See More →](#)

Grow your data skills with DataCamp for Mobile

Make progress on the go with our mobile courses and daily 5-minute coding challenges.



LEARN

[Learn Python](#)

[Learn R](#)

[Learn AI](#)

[Learn SQL](#)

[Learn Power BI](#)

[Learn Tableau](#)

[Learn Data Engineering](#)

[Assessments](#)

[Career Tracks](#)

[Skill Tracks](#)

[Courses](#)

[Data Science Roadmap](#)

DATA COURSES

[Python Courses](#)

[R Courses](#)

[SQL Courses](#)

[Power BI Courses](#)

[Tableau Courses](#)

[Alteryx Courses](#)

[Azure Courses](#)

[Google Sheets Courses](#)

[AI Courses](#)

[Data Analysis Courses](#)

[Data Visualization Courses](#)

[Machine Learning Courses](#)

[Data Engineering Courses](#)

[Probability & Statistics Courses](#)

DATALAB

[Get Started](#)

[Pricing](#)

[Security](#)

[Documentation](#)

CERTIFICATION

[Certifications](#)

[Data Scientist](#)

[Data Analyst](#)

[Data Engineer](#)

[SQL Associate](#)

[Power BI Data Analyst](#)

[Tableau Certified Data Analyst](#)

[Azure Fundamentals](#)

[AI Fundamentals](#)

RESOURCES

[Resource Center](#)

[Upcoming Events](#)

[Blog](#)

[Code-Alongs](#)

[Tutorials](#)

[Docs](#)

[Open Source](#)

[RDocumentation](#)

[Course Editor](#)

[Book a Demo with DataCamp for Business](#)

[Data Portfolio](#)

[Portfolio Leaderboard](#)

PLANS

[Pricing](#)

[For Business](#)

[For Universities](#)

[Discounts, Promos & Sales](#)

[DataCamp Donates](#)

FOR BUSINESS

[Business Pricing](#)

[Teams Plan](#)

[Data & AI Unlimited Plan](#)

[Customer Stories](#)

[Partner Program](#)

ABOUT

[About Us](#)

[Learner Stories](#)

[Careers](#)

[Become an Instructor](#)

[Press](#)

[Leadership](#)

[Contact Us](#)

[DataCamp Español](#)

[DataCamp Português](#)

[DataCamp Deutsch](#)

[DataCamp Français](#)

SUPPORT

[Help Center](#)

[Become an Affiliate](#)



[Privacy Policy](#)

[Cookie Notice](#)

[Do Not Sell My Personal Information](#)

[Accessibility](#)

[Security](#)

[Terms of Use](#)

© 2024 DataCamp, Inc. All Rights Reserved.