

Open Source Tools for Data Science

Welcome to **Open Source Tools for Data Science**! Data scientists use open source apps to take advantage of their secure and stable features. In this activity, you will explore eight types of open source tools and discover their features and uses.

Objectives

- Identify common open source tools for data science

Player controls

> Next

< Back



Accessibility

Learn more

Open Source Tools for Data Science



Start



Recap



Congratulations! You have successfully learned how to identify open source tools that data scientists use. You explored tools for data management, data integration and transformation, data visualization, model deployment, model monitoring and assessment, code assets, and data assets.

Next steps

Continue to identify and compare open source tools for data science.

Data management tools

Select each of the eight following tools to learn more. Then, select **Next**.

- ✓ MySQL
- PostgreSQL
- cassandra
- elasticsearch
- ceph
- CouchDB
- mongoDB
- Hadoop

MySQL is a popular open source relational database management system (RDBMS) that uses structured query language (SQL) to manage and store data. Its common use is for web applications, data warehousing, and e-commerce.

Data management tools

Select each of the eight following tools to learn more. Then, select **Next**.

- ✓

MySQL
- ✓

PostgreSQL
- cassandra
- elasticsearch
- ceph
- CouchDB
- mongoDB
- Hadoop

PostgreSQL is a powerful and open source relational database management system that emphasizes extensibility and SQL compliance. It offers advanced features such as support for JSON, full-text search, and spatial data.

Data management tools

Select each of the eight following tools to learn more. Then, select **Next**.

- ✓

MySQL
- ✓

PostgreSQL
- ✓

cassandra
- elasticsearch
- ceph
- CouchDB
- mongoDB
- Hadoop

Apache Cassandra is a highly scalable, distributed NoSQL database that can handle large amounts of structured and unstructured data across many commodity servers. It offers high availability, fault tolerance, and tunable consistency levels, making it suitable for mission-critical applications.

Data management tools

Select each of the eight following tools to learn more. Then, select **Next**.

- ✓ MySQL
- PostgreSQL
- cassandra
- elasticsearch
- ceph
- CouchDB
- mongoDB
- Hadoop

Elasticsearch is a distributed, RESTful search and analytics engine based on the Lucene library. It is highly scalable and easy to use, with powerful querying capabilities and real-time data indexing.

Data management tools

Select each of the eight following tools to learn more. Then, select **Next**.

- ✓ MySQL

✓ PostgreSQL

✓ cassandra

✓ elasticsearch

✓ ceph

CouchDB

mongoDB

Hadoop

Ceph is a free, open source software-defined storage platform designed for modern data centers. It provides scalable object, block, and file storage under one unified system, with high availability, reliability, and performance.

Data management tools

Select each of the eight following tools to learn more. Then, select **Next**.

- ✓ MySQL
- ✓ PostgreSQL
- ✓ cassandra
- ✓ elasticsearch
- ✓ ceph
- ✓ CouchDB
- mongodb
- Hadoop

CouchDB is a NoSQL document-oriented database that uses JSON to store data. It is highly scalable, fault-tolerant, and easy to use.

Data management tools

Select each of the eight following tools to learn more. Then, select **Next**.

- ✓ MySQL
- ✓ PostgreSQL
- ✓ cassandra
- ✓ elasticsearch
- ✓ ceph
- ✓ CouchDB
- ✓ mongoDB
- Hadoop

MongoDB is a document-oriented NoSQL database that stores data in a flexible JSON. It provides scalability, high availability, and data distribution, making it suitable for modern web applications that handle large volumes of unstructured data.

Data management tools

Select each of the eight following tools to learn more. Then, select **Next**.

- ✓ MySQL
- ✓ PostgreSQL
- ✓ cassandra
- ✓ elasticsearch
- ✓ ceph
- ✓ CouchDB
- ✓ mongoDB
- ✓ Hadoop

Hadoop HDFS (Hadoop Distributed File System) is a distributed file system that provides high-throughput access to application data. It is fault-tolerant, scalable, and efficient, making it suitable for storing and processing large datasets in a distributed computing environment.

Operating system tools

Select each of the six following tools to learn more. Then, select **Next**.

✓ Jupyter

Pycharm

RStudio

Spyder

Microsoft Visual Studio

Anaconda Navigator

The **Jupyter IDE**, an open-source effort, supports Julia, Python, and R development with Jupyter Notebook, JuypyterLab, and JupyterHub. Users can create and share documents containing live code, equations, visualizations, and narrative text. JuypyterLab includes customized notebook organization. JupyterHub extends all these capabilities to the enterprise.

Operating system tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓

Jupyter

Pycharm
- ✓

RStudio

Spyder
- Microsoft Visual Studio

Anaconda Navigator

Developers can use **RStudio**, a free and open-source IDE, built to manage and execute R code. RStudio works on all platforms and includes version control and project management capabilities.

Operating system tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓

Jupyter

Pycharm
- ✓

RStudio

Spyder
- ✓

Microsoft Visual Studio

Anaconda Navigator

Microsoft Visual Studio is an IDE that supports a variety of programming languages, including C, C++, C++/CLI, Visual Basic .NET, C#, F#, JavaScript, TypeScript, XML, XSLT, HTML, and CSS. Visual Studio supports Python, Ruby, Node.js, and M and other languages using plug-ins.

Operating system tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓

Jupyter
- ✓

Pycharm
- ✓

RStudio
- Spyder
- ✓

Microsoft Visual Studio
- Anaconda Navigator

Pycharm, primarily a subscription-based IDE environment, offers 16+ additional tools for coding assistance, testing, and web development. Pycharm supports scientific development with IPython integration and Matplotlib and NumPy support. PyCharm also offers a free community-based, open-source IDE with limited capabilities.

Operating system tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓

Jupyter
- ✓

Pycharm
- ✓

RStudio
- ✓

Spyder
- ✓

Microsoft Visual Studio
- Anaconda Navigator

Spyder is a free, open-source Python-based IDE designed by and for scientists, engineers, and data analysts. This IDE features a unique combination of comprehensive development tools for advanced editing, analysis, debugging, profiling, and visualization capabilities.

Operating system tools

Select each of the six following tools to learn more. Then, select **Next**.

✓ Jupyter

✓ Pycharm

✓ RStudio

✓ Spyder

✓ Microsoft Visual Studio

✓ Anaconda Navigator

Anaconda Navigator is an open-source GUI-based Navigator that supports Python development and integrates with Eclipse and PyDev, IDLE, IntelliJ, Microsoft Visual Studio Code (VS Code), Ninja IDE, PyCharm, Python for Visual Studio Code, Python Tools for Visual Studio (PTVS), Spyder, Sublime Text and Wing IDE.

Data integration and transformation tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓ Spark SQL
- Kubeflow
- Node-RED
- Airflow
- nifi
- kafka

Apache Spark SQL is a module in the Spark ecosystem that provides a programming interface for working with structured data using SQL, data frames, and datasets. It supports a wide range of data sources and provides optimized performance for complex data processing tasks.

Data integration and transformation tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓ Spark SQL
- ✓ Kubeflow
- Node-RED
- Airflow
- nifi
- kafka

Kubeflow is an open source machine learning toolkit built on top of Kubernetes. It provides a platform for building, deploying, and managing end-to-end machine learning workflows at scale, with support for distributed training, model serving, and hyperparameter tuning.

Data integration and transformation tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓ Spark SQL
- ✓ Kubeflow
- ✓ Node-RED
- Airflow
- nifi
- kafka

Node-RED is an open source visual programming tool for wiring together hardware devices, APIs, and online services. It allows users to create event-driven flows of messages, with support for data transformation, filtering, and aggregation.

Data integration and transformation tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓ Spark SQL
- ✓ Kubeflow
- ✓ Node-RED
- ✓ Airflow
- nifi
- kafka

Apache Airflow is an open source platform for programmatically authoring, scheduling, and monitoring work flows. It allows users to define and execute complex work flows, with support for task dependencies, parallelism, and error handling.

Data integration and transformation tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓ Spark SQL
- ✓ Kubeflow
- ✓ Node-RED
- ✓ Airflow
- ✓ nifi
- kafka

Apache NiFi is an open source data integration platform that allows users to automate the flow of data between systems. It provides a web-based user interface for designing and managing data flows, with support for data routing, transformation, and enrichment, among other capabilities.

Data integration and transformation tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓ Spark SQL
- ✓ Kubeflow
- ✓ Node-RED
- ✓ Airflow
- ✓ nifi
- ✓ kafka

Apache Kafka is a distributed streaming platform that allows applications to publish, process, and subscribe to streams of records in real-time. It is scalable, fault-tolerant, and high-throughput, making it suitable for building mission-critical, data-intensive applications.

Data visualization tools

Tableau and Power BI are used for data visualization, but data scientists use other tools. Select the following four tools to learn more. Then, select **Next**.

✓ PixieDust

Hue

Kibana

Superset

PixieDust is an open source library for creating interactive, exploratory data visualizations in Python and Jupyter notebooks. It provides a range of built-in visualizations and data connectors, with support for customization and extensibility through third-party libraries.

Data visualization tools

Tableau and Power BI are used for data visualization, but data scientists use other tools. Select the following four tools to learn more. Then, select **Next**.

✓ PixieDust

✓ Kibana

Hue

Superset

Kibana is an open source data visualization tool that allows users to interact with their data through a web-based interface. It is commonly used with Elasticsearch to analyze and visualize large datasets.

Data visualization tools

Tableau and Power BI are used for data visualization, but data scientists use other tools. Select the following four tools to learn more. Then, select **Next**.

- ✓ PixieDust
- ✓ Kibana
- ✓ Hue
- Superset

Hue is an open source web interface for analyzing and visualizing large datasets in Apache Hadoop. It offers a user-friendly experience for exploring data and creating visualizations without the need for programming skills.

Data visualization tools

Tableau and Power BI are used for data visualization, but data scientists use other tools. Select the following four tools to learn more. Then, select **Next**.

- ✓ PixieDust
- ✓ Kibana
- ✓ Hue
- ✓ Superset

Apache Superset is a modern, enterprise-ready business intelligence web application that makes it easy to visualize and explore large datasets. It offers a rich set of data visualization options, including charts, tables, and maps, as well as advanced features such as geospatial analysis and real-time data processing.

Model deployment tools

Select each of the eight following tools to learn more. Then, select **Next**.

PredictionIO

Seldon

Kubernetes

OpenShift

MLeap

TensorFlow
Serving

TensorFlow
Lite

TensorFlow.js

Apache PredictionIO is an open source machine learning server built on a scalable and distributed infrastructure. It allows developers to quickly build, evaluate, and deploy predictive engines for various use cases such as recommendation, classification, and clustering.

Model deployment tools

Select each of the eight following tools to learn more. Then, select **Next**.

PredictionIO

Seldon

Kubernetes

OpenShift

MLeap

TensorFlow
Serving

TensorFlow
Lite

TensorFlow.js

Kubernetes is an open source platform for container orchestration. It automatically launches, scales, and manages containerized applications. Offering features like automatic scaling, self-healing, and load balancing, Kubernetes enables the management and orchestration of containers across numerous hosts.

Model deployment tools

Select each of the eight following tools to learn more. Then, select **Next**.

PredictionIO

Seldon

Kubernetes

OpenShift

MLeap

TensorFlow
Serving

TensorFlow
Lite

TensorFlow.js

MLeap is an open source library for serializing and deserializing **learning** models in a cross-platform file. It gives users the ability to export models from different machine learning libraries and frameworks, such as Spark, scikit-learn, and TensorFlow, and implement them in high-throughput, low-latency production environments.

Model deployment tools

Select each of the eight following tools to learn more. Then, select **Next**.

PredictionIO

Kubernetes

MLeap

TensorFlow Lite

Seldon

OpenShift

TensorFlow Serving

TensorFlow.js

TensorFlow Lite is an open source tool for running machine learning models on mobile and embedded devices. To allow effective inference on mobile and embedded platforms, TensorFlow Lite supports a variety of hardware accelerators such as CPUs, GPUs, and custom ASICs.

Model deployment tools

Select each of the eight following tools to learn more. Then, select **Next**.

PredictionIO

Kubernetes

MLeap

TensorFlow Lite

Seldon

OpenShift

TensorFlow Serving

TensorFlow.js

Apache Seldon is an open source platform for deploying and managing machine learning models on Kubernetes. It provides a way to serve models at scale, automate model deployment workflows, and monitor the performance of deployed models in real-time.

Model deployment tools

Select each of the eight following tools to learn more. Then, select **Next**.

PredictionIO ✓

Seldon

Kubernetes ✓

OpenShift

MLeap

TensorFlow Serving

TensorFlow Lite

TensorFlow.js

Red Hat OpenShift is a container application framework based on Kubernetes. With characteristics like automation, scalability, and security, it offers a method for creating, deploying, and managing containerized applications.

Model deployment tools

Select each of the eight following tools to learn more. Then, select **Next**.

PredictionIO ✓

Seldon

Kubernetes ✓

OpenShift

MLeap ✓

TensorFlow Serving

TensorFlow Lite

TensorFlow.js

TensorFlow Serving is an open source utility that serves machine learning models in real-world settings. It supports both HTTP and gRPC interfaces for serving predictions and provides for the high scalability and low latency deployment and management of TensorFlow models.

Model deployment tools

Select each of the eight following tools to learn more. Then, select **Next**.

PredictionIO ✓

Seldon

Kubernetes ✓

OpenShift

MLeap ✓

TensorFlow Serving

TensorFlow Lite ✓

TensorFlow.js

TensorFlow.js is an open source library for building and deploying machine learning models in JavaScript. It allows you to train and execute models directly in the browser or on Node.js, and it supports a wide range of model architectures, including neural networks, decision trees, and k-nearest neighbors.

Model monitoring and assessment tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓ AI Fairness 360
- AI Explainability 360
- Adversarial Robustness 360
- Prometheus
- ModelDB

IBM AI Fairness 360 is an open source toolkit for detecting and mitigating bias in machine learning models. It provides a way to measure the fairness and bias of models, as well as a set of algorithms for mitigating bias and creating fairer models.

Model monitoring and assessment tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓ AI Fairness 360
- ✓ AI Explainability 360
- Adversarial Robustness 360
- Prometheus
- ModelDB

IBM AI Explainability 360 is an open source toolkit for explaining the behavior and decisions of machine learning models. It provides a way to measure the explainability and interpretability of models, as well as a set of algorithms for generating explanations and visualizations of model behavior.

Model monitoring and assessment tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓ AI Fairness 360
- ✓ AI Explainability 360
- ✓ Adversarial Robustness 360
- Prometheus
- ModelDB

The **IBM Adversarial Robustness 360 Toolbox** is a free and open source library for protecting machine learning models from adversarial attacks. It includes a method for measuring model robustness and vulnerability, as well as a set of algorithms for improving model robustness and detecting adversarial examples.

Model monitoring and assessment tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓ AI Fairness 360
- ✓ AI Explainability 360
- ✓ Adversarial Robustness 360
- ✓ Prometheus
- ModelDB

Prometheus is a freely available monitoring system that collects and stores metrics in real-time from different sources. It allows you to visualize and set alerts on the health and performance of systems and apps, and it supports a variety of data gathering methods, such as HTTP endpoints, exporters, and agents.

Model monitoring and assessment tools

Select each of the six following tools to learn more. Then, select **Next**.

- ✓ AI Fairness 360
- ✓ AI Explainability 360
- ✓ Adversarial Robustness 360
- ✓ Prometheus
- ✓ ModelDB

ModelDB is an open source platform for managing machine learning models and experiments. It provides a way to track and reproduce experiments, version models, and collaborate with team members.

Code asset tools

Select each of the four following tools to learn more. Then, select **Next**.

✓

Git

GitLab

GitHub

Bitbucket

Git is an open source version control system for tracking changes in code and collaboration among developers. It provides a way to manage and organize code changes, collaborate on code development, and maintain a history of code revisions.

Code asset tools

Select each of the four following tools to learn more. Then, select **Next**.

✓

Git

✓

GitLab

GitHub

Bitbucket

GitLab is a web-based Git repository manager that provides a complete DevOps platform for source code management, continuous integration and deployment, and monitoring. It enables teams to collaborate on code development, automate build and deployment processes, and track metrics and performance across the entire software development lifecycle.

Code asset tools

Select each of the four following tools to learn more. Then, select **Next**.

✓

Git

✓

GitLab

✓

GitHub

Bitbucket

GitHub is a web-based Git repository hosting service that provides a platform for developers to collaborate on code and manage software projects. It enables users to create, fork, and contribute to open source projects, track changes in code, and manage issues and pull requests.

Code asset tools

Select each of the four following tools to learn more. Then, select **Next**.

- ✓ Git
- ✓ GitLab
- ✓ GitHub
- ✓ Bitbucket

Bitbucket from Atlassian is a web-based Git repository hosting service. It provides a platform for developers to collaborate on code and manage software projects, with features like pull requests, code review, and branch permissions.

Data asset tools

Select the following three tools to learn more. Then, select **Next**.

✓ Egeria

Kylo

Atlas

ODPi Egeria is an open source metadata management framework that provides a standard way to manage and share metadata across different platforms and tools. It enables organizations to discover, govern, and use metadata across the entire data lifecycle, from ingestion to analysis.

Data asset tools

Select the following three tools to learn more. Then, select **Next**.

✓ Egeria

✓ Kylo

Atlas

Kylo is an open source data lake management platform designed to simplify the process of ingesting, preparing, and analyzing data at scale. It provides a user-friendly interface for managing data workflows and supports data sources and data preparation tools.

Data asset tools

Select the following three tools to learn more. Then, select **Next**.

✓ Egeria


✓ Kylo

✓ Atlas

Apache Atlas is an open source metadata management and governance framework for Hadoop ecosystems. It provides a way to discover, classify, and manage metadata across different data platforms, including Hadoop HDFS, Hive, and HBase.

Choosing an open source tool

Your assignment is to identify an open source platform for programmatically authoring, scheduling, and managing workflows. Select one of the following tools to perform this task.



Nifi

Kafka

Airflow

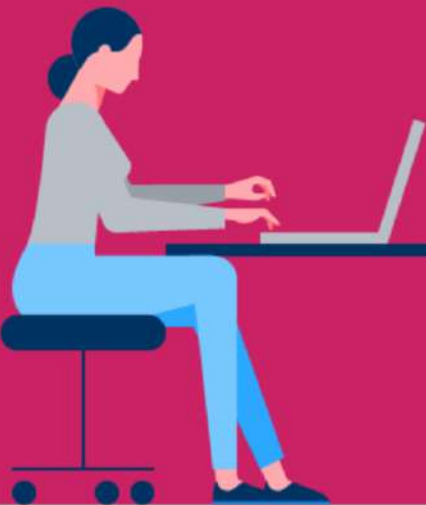
Spark SQL

Correct!

Apache Airflow is an open source platform for programmatically authoring, scheduling, and monitoring work flows.

Choosing an open source tool

Your assignment is to identify a tool that is suitable for web applications that handle large volumes of unstructured data. It should be a document oriented NOSQL database that stores data in JSON. Select the tool to perform this task.



Hadoop

MySQL

PostgreSQL

MongoDB

Correct!

MongoDB is a document-oriented NoSQL database that stores data in a flexible JSON.

Choosing an open source tool

Your assignment is to identify the open source tool that will allow your database services to scale across many commodity servers. **Select** one of the following four tools to perform the task.



Bitbucket

Atlas

cassandra

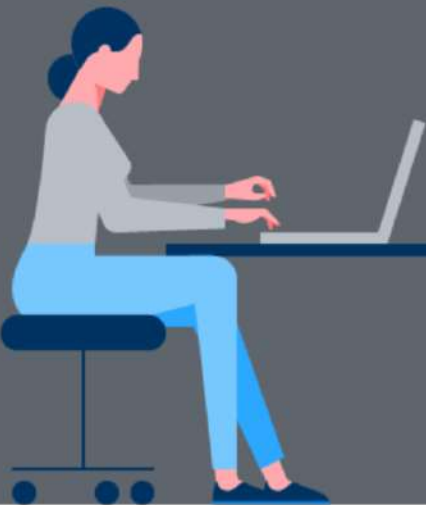
PixieDust

Correct!

Apache Cassandra is a highly scalable, distributed NoSQL database that handles large amounts of structured and unstructured data across servers.

Choosing an open source tool

Your assignment is to identify a tool which automatically launches, scales, and manages containerized applications. Select one of the following four tools to perform this task.



Kubernetes

Seldon

MLeap

TensorFlow.js

Correct!

Kubernetes is an open source platform for container orchestration. It automatically launches, scales, and manages containerized applications.

Choosing an open source tool

Your assignment is to identify a tool that provides an open source web interface for analyzing and visualizing large datasets in Apache Hadoop. Select the tool to perform this task.



Kibana

PixieDust

Hue

Superset

Correct!

Hue is an open source web interface for analyzing and visualizing large datasets in Apache Hadoop.

Choosing an open source tool

Your assignment is to identify a tool which automatically launches, scales, and manages containerized applications. Select one of the following four tools to perform this task.



ModelDB

Prometheus

AI Fairness 360

Adversarial
Robustness 360

Correct!

ModelDB is an open source platform for managing machine learning models and experiments.

Choosing an open source tool

Your assignment is to identify the IDE tool that supports Python development and integrates with multiple IDEs. Select one of the following four tools to perform this task.



Spyder

RStudio

Anaconda
Navigator

Microsoft Visual
Studio

Correct!

Anaconda Navigator is an open source GUI-based Navigator that supports Python development and integrates with IDEs.