Data science languages

Sql: structered query language, was developed at IBM, managing data in relational data bases,

Python: open source

R: free software, statistical computing language

Visual basic

C++

Java script

Scala: the name is a mix of scalable and language

Julia

The most widely used open source data management tools are relational databases such as MySQL and PostgreSQL; NoSQL databases such as MongoDB Apache CouchDB, and Apache Cassandra; and file-based tools such as the Hadoop File System or Cloud File systems like Ceph.

“ELT” – Extract, Load, Transform“ELT”,

Open source data integration and transformation tools:

Apache AirFlow, originally created by AirBNB; KubeFlow, which enables you to execute data science pipelines on top of Kubernetes; Apache Kafka, which originated from LinkedIn;  Apache Nifi, which delivers a very nice visual editor;

Apache SparkSQL (which enables you to use ANSI SQL and scales up to compute clusters of 1000s of nodes), and NodeRED, which also provides a visual editor. NodeRED consumes so little in resources that it even runs on small devices like a Raspberry Pi

Open source data visualization tools:

Hue, which can create visualizations from SQL queries.

Kibana, a data exploration and visualization web application, is limited to Elasticsearch (the data provider).

Finally, Apache Superset is a data exploration and visualization web application.

Model deployment:

Apache PredictionIO currently only supports Apache Spark ML models for deployment, but support for all sorts of other libraries is on the roadmap.

Seldon is an interesting product since it supports nearly every framework, including TensorFlow, Apache SparkML, R, and scikit-learn. Seldon can run on top of Kubernetes and Redhat OpenShift.

Another way to deploy SparkML models is by using MLeap.

Finally, TensorFlow can serve any of its models using the TensorFlow service.

You can deploy to an embedded device like a Raspberry Pi or a smartphone using TensorFlow Lite, and even deploy to a web browser using TensorFlow dot JS.

model monitoring tools:

ModelDB is a machine model metadatabase where information about the models are stored and can be queried. It natively supports Apache Spark ML Pipelines and scikit-learn.

A generic, multi-purpose tool called Prometheus is also widely used for machine learning model monitoring, although it’s not specifically made for this purpose.

The IBM AI Fairness 360 open source toolkit  detects and mitigates against bias in machine learning models. Machine learning models, especially neural-network-based deep learning models, can be subject to adversarial attacks, where an attacker tries to fool the model with manipulated data or by manipulating the model itself

The IBM Adversarial Robustness 360 Toolbox can be used to detect vulnerability to adversarial attacks and help make the model more robust. Machine learning modes are often considered to be a black box that applies some mysterious “magic.”

The IBM AI Explainability 360 Toolkit makes the machine learning process more understandable by finding similar examples within a dataset that can be presented to a user for manual comparison. it can also illustrate training for a simpler machine learning model by explaining how different input variables affect the final decision of the model.

Code asset management version management or version control

Git is now the standard. Multiple services have emerged to support Git, with the most prominent being GitHub, which provides hosting for software development version management. The runner-up is definitely GitLab, which has the advantage of being a fully open source platform that you can host and manage yourself. Another choice is Bitbucket.

Fully integrated visual tools and platforms :

Watson Studio, together with Watson OpenScale, covers the complete development life cycle for all data science, machine learning, and AI tasks.

Microsoft Azure Machine Learning. H2O driverless al

Data management:

SaaS stands for “software as a service.” It means that the cloud provider operates the tool for you in the cloud.

Amazon Web Services DynamoDB, a NoSQL database that allows storage and retrieval of data in a key-value or a document store format.

Cloudant, which is a database-as-a-service offering. But, under the hood it is based on the open source Apache CouchDB.

Two widely used commercial data integration tools are Informatica Cloud Data Integration and IBM’s Data Refinery.

An example of a smaller company’s cloud-based data visualization tool is DataMeer. IBM offers it’s famous Cognos Business intelligence suite as cloud solution as well. IBM Data Refinery also offers data exploration and visualization functionality in Watson Studio.

Model building can be done using a service such as Watson Machine Learning. Google cloud

Libraries for data science:

Scientifics computing libraries: pandas, numpy,

Visualization libraries: matplotlib, seaborn

Machine learning and deep learning: scikit-learn, keras

Deep learning libraries: tensorflow, pytorch

Apache Spark is a general-purpose cluster-computing framework that enables you to process data using compute clusters. This means that you process data in parallel, using multiple computers simultaneously. The Spark library has similar functionality as Pandas Numpy Scikit-learn Apache Spark data processing jobs can use Python

R Scala, or SQL

Scala libraries: vegas, big dl for deep learning

R libraries: ggplot2, keras tensorflow

Application programming interfaces (API):

Rest API: representational state transfer

Data sets powering data science:

Deep learning models are implemented using popular frameworks such as TensorFlow, PyTorch,

and Keras.

Deep learning frameworks typically provide a Python API, and many support other programming

languages, such as C++ and JavaScript.

Jupiter notebook:

A Jupyter notebook is a browser-based application that allows you to create and share documents that contain code, equations, visualizations, narrative text links, and so much more.

Jupyter Notebooks originated as “iPython” originally developed for the Python Programming language. As it came to support additional languages it was renamed Jupyter which stands for: Julia - Python - R.

JupyterLab allows for: Interactive control of the notebook cells and output, real time

editing markdowns, CSV etc.

It is compatible with several file formats like CSV, JSON, PDF, Vega and more.

And is open source.

Jupyter notebooks can be used with a cloud-based service like IBM and Google Collab.

They do not require you to install anything on your local machine; they give you access to the Jupyter Notebook environment.

They do allow you to import and export notebooks using the standard IPython Notebook file format.

They support the Python language and other languages as well.

R libraries for data science:

Data manipulation: dplyr

String manipulation: stringr

Data visualization: ggplot

Machine learning: caret