**Categories of data science tools**

**Data management** is the process of persisting and retrieving data.

**Data integration and transformation**, often referred to as extract, transform and load, or ETL is the process of retrieving data from remote data management. Transforming data and loading it into a local data management system is also part of Data integration and transformation.

**Data visualization** is part of an initial data exploration process, as well as being part of a final deliverable.

**Model building** is the process of creating a machine learning or deep learning model using an appropriate algorithm with a lot of data.

**Model deployment** makes such a machine learning or deep learning model available to third party applications.

**Model monitoring and assessment** ensures continuous performance quality checks on the deployed models. These checks are for accuracy, fairness and adversarial robustness.

**Code asset management** uses versioning and other collaborative features to facilitate teamwork.

**Data asset management** brings the same versioning and collaborative components to data. Data asset management also supports replication, backup, and environments, commonly known as integrated development environments, or “IDEs”, are tools that help the data scientist to implement, execute, test, and deploy their work.

**Execution environments** are tools where data preprocessing, model training, and deployment take place.

**Open Source Tools for Data Science**

Relational Data Bases

Open source data management tools are relational data bases such as MySQL and PosgreSQL.

No SQL databases such as MongoDB Apache CouchDB, and Apache Cassandra.

File based-Tools such as the Haddop File System or Cloud File systems like Ceph.

Elastic search is mainly used for storing text data and creating a search index for fast document retrieval.

The task of data integration and transformation in the classic data warehousing world called ETL, which stands for extract, transform, and load.

Data Refinery and cleansing.

**Open source data integration and transformation tools**:

Apache Airflow, originally created by AirBNB; Kubeflow, which enables you to execute data sicence 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 so small devices like a Raspberry Pi.

**Open source data visualization tools.**

Programming libraries you need to code differ from tools that contain a user interface.

Hue, which can create visualizations from SQL queries.

Kibana, a data exploration and visualization web application, is limited to Elastic search (the data provider). Finally, Apache Superset is a data exploration and visualization web application. Model deployment is extremely important. Once you have created a machine learning model capable of predict some key aspects of the future, you should make that model consumable by other developers and turn it into an API. Apache prediction IO 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 kubernrtes and Redhat OpenShift. Aother 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.

ModelDB is a machine model metadata base 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. Model performance is not exclusive measured through accuracy. Model bias against protected groups like gender or race is also important. The IBM AI Fairness 360 open source toolkit does exactly this. It detects and mitigates against bias in machine learning models. Machine learning models, especially neural-network-based deep learning models, can be subjected 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 Explain-ability 360 toolkit makes the machine learning process more understandable by finding similar examples within a database that can be presented to a user for manual comparison. The IBM AI Explain-ability 360 toolkit can also illustrate training for a simpler machine learning model by explaining how different input variables affect the final decision of the model. Options for code asset management tools have been greatly simplified: for code asset management – also referred to as version management of 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. Data asset management, also known as data governance or data lineage, is another crucial part of enterprise grade data science. Data has to be versioned and annotated with metadata. Apache Atlas is a tool that supports this task. Another interesting project, ODPi Egeria, is managed through the linux Foundation and is an open ecosystem. It offers a set of open APIs, Types, and interchange protocols that metadata repositories used to share and exchange data. Finally Kylo is an open source data lake management software that provides extensive support for a wide range of data asset management tasks.

**Open source Tools for Data Science 2**

Jupyter first emerged as a tool for interactive Python programming; it now supports more than a hundred different programming languages through “Kernels”. Kernels shouldn’t be confused with operating system Kernels. Jupyter Kernels are encapsulating the different interactive interpreters for the different programming languages. A key property of jupyter Notebooks is the ability to unify documentation, code, and output from code, shell commands, and visualizations into a single document. Jupyter Lab is the next generation of jupyter Notebooks and in the long term, will actually replace Jupyter notebooks. The architectural changes being introduced in jupyter lab makes jupyter more modern and modular. From a user’s perspective, the main difference introduced by jupyter Lab is the ability to open different types of files, including Jupyter Notebooks, data, and terminals. You can then arrange these files on the canvas. Although Apache Zeppelin has been fully implemented, it’s inspired by Jupyter Notebooks and provides a similar experience. One Key differentiator is the integrated plotting capability. In Jupyter Notebooks, you are required to use external libraries in Apache Zeppelin, and plotting doesn’t require coding. You can also extend these capabilities by using additional libraries. RStudio is one of the oldest development environments for statistics and data science, having been introduced in 2011. It exclusively runs R and all associated R libraries. However, Python development is possible and E is therefore tightly integrated into this tool to provide an optimal user experience. RStudio unifies programing, execution, debugging, remote data access, data exploration, and visualization into a single tool. Spyder tries to mimic the behavior of RStudio to bring its functionality to the Python World. Although Spyder does not have the same level of functionality as RStudio, data scientists do consider it an alternative. But in the Python world, Jupyter is used more frequently. Sometimes your data doesn’t fit into a single computer’s storage or main memory capacity. That’s where cluster execution come in. The well-known cluster-computing framework Apache Spark is among the most active Apache projects and is used across all industries, including in many Fortune 500 companies. The Key property of Apache Spark is linear scalability. This means, if you double the number of servers in a cluster, you will also roughly double its performance. After Apache Spark began to gain market share, Apache Flink was created. The key difference between Apache Spark and Apache Flink is that Apache Spark is a batch data processing engine, capable of processing huge amounts of data file by file. Apache Flink, on the other hand, is a stream processing image, with its main focus on processing paradigms, Apache Spark is usually the choice in most use cases. One of the latest developments in the data science execution environments is called “Ray”, which has a clear focus on large-scale deep learning model training.

Open Source tools for data scientists that are fully integrated and visual. With these tools, no programming knowledge is necessary. Most important tasks are supported by these tools; this task include data integration, transformation, data visualization, and model building.

KNIME originated at the University of Konstanz in 2004. KNIME has a visual user interface with drag-and-drop capabilities. It also has built-in visualization capabilities. Knime can be extended by programming in R and Python, and has connectors to Apache Apache Spark. Another example of this group of tools is Orange. It’s less flexible than KNIME, but easier to use.

**Commercial Tools for Data Science**

In data management, most of an enterprise’s relevant data is stored in an Oracle Data base, Microsoft SQL Server, or IBM Db2. According to a Gartner Magic Quadrant, Informatica Power Center and IBM Info Sphere Data Stage are the leaders, followed by products from SAP, Oracle, SAS, Talend, and Microsoft. These tools support design and deployment of ETL data processing pipelines through a graphical interface. They provide connectors to most of the commercial and open source target information systems.

Watson studio Desktop includes a component called Data Refinery, which enables the defining and execution of data integration processes in a spreadsheet style.

Commercial environment, data visualization are utilizing business intelligence “BI”, tools.

Their main focus is to create visually attractive and easy-to-understand reports and live dashboards.

Commercial Examples

Tableau

Microsoft Power BI

IBM Cognos Analytics

How can different columns in a table relate to each other?

This type of functionality is contained in Watson Studio Desktop.

If you want to build a machine learning model using a commercial tool, you should consider using a data mining product.

SPSS Modeler and SAS Enterprise Miner.

A version of SPSS Modeler is also available in Watson Studio Desktop, based on the cloud version of the tool.

Commercial Software, model deployment is tightly integrated in the model building process.

Commercial Software can also export models in an open format. Example, SPSS Modeler supports the exporting of models as Predictive Model Markup language, or PMML, which can be read by many other commercial and open software packages. Model monitoring is a new discipline and there are currently no relevant commercial tools available. As a result, open source is the first choice. The same is true for code asset management. Open Source with Git and Github is the effective standard. Data asset management, often called data governance or data lineage, is a crucial part of enterprise grade data science. Data must be versioned and annotated using metadata. Vendors, including informatica Enterprise Data governance and IBM, provide tools for these specific tasks. The IBM Info sphere information Governance catalog covers functions like data dictionary, which facilitates discovery of data assets. Each data asset is assigned to a data steward – The data owner. The data owner is responsible for the data asset and can be contacted. Data lineage is also covered; this enables a user to track back through the transformation steps followed in creating the data assets. The data lineage also includes a reference to the actual source data. Rules and polices can be added to reflect complex regulatory and business requirements for data privacy and retention. Watson studio is a fully integrated development environment for data scientists. It’s usually consumed through the cloud. Watson studio desktop combines Jupyter Notebooks with graphical tools to maximize data scientists’ performance. Watson studio, together with Watson open scale, is a fully integrated tool covering the full data science life cycle.

**Cloud Based Tools for Data Science**

Cloud products follow the trend of having multiple tasks integrated in tools.

Microsoft Azure (machine learning)

H2O Driverless AI

Saas stands for “Software as a service” it means that the cloud provider operates the tool for you in the cloud.

Example, the cloud provider operates thee product by backing up your data and configuration and installing updates.

Only available as a cloud product, sometimes it’s only available from a single