Open source tools:

Data Management: 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. Finally, Elasticsearch is mainly used for storing text data and creating a search index for fast document retrieval.

Data integration and transformation: 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.

Data Visualization:  A similar approach uses 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 Monitor: 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.

Code asset management: also referred to as 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. Data asset management: also known as data governance or data lineage 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 use to share and exchange data. Finally, Kylo is an open source data lake management software platform that provides extensive support for a wide range of data asset management tasks.

development environments: Jupyter first emerged as a tool for interactive Python programming; it now supports more than a hundred different programming languages through “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, output from the code, shell commands, and visualizations into a single document. JupyterLab is the next generation of Jupyter Notebooks and in the long term, will actually replace Jupyter Notebooks. RStudio is one of the oldest development environments for statistics and data science. It exclusively runs R and all associated R libraries. However, Python development is possible and R is therefore tightly integrated into this tool to provide an optimal user experience. RStudio unifies programming, execution, debugging, remote data access, data exploration, and

visualization into a single tool. Spyder tries to mimic the behaviour of RStudio to bring its functionality to the Python world.

Execution Environment: 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 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 real-time data streams. Fully Integrated visual tools: KNIME has a visual user interface with drag-and-drop capabilities. It also has built-in visualization capabilities. Knime can be be extended by programming in R and Python, and has connectors to Apache Spark. Another example of this group of tools is Orange.

Commercial tools:

Data Management:  Oracle Database, Microsoft SQL Server, or IBM Db2.

Data Integration:  Informatica Powercenter and IBM InfoSphere DataStage are

the leaders, followed by products from SAP, Oracle, SAS, Talend, and Microsoft.

Data visualization: Tableau, Microsoft Power BI, and IBM Cognos Analytics.

Model Building: SPSS Modeler and SAS Enterprise Miner.

Model Deployment: SPSS Collaboration and Deployment Services. Commercial software can also export models in an open format. For example, SPSS Modeler supports the exporting of models as Predictive Model Markup Language, or PMML.

Data asset management: Informatica Enterprise Data Governance and IBM. The IBM InfoSphere Information Governance Catalog covers functions like data dictionary, which facilitates discovery of data assets.

Development Environment: Watson Studio. Watson Studio, together with Watson Open Scale, is a fully integrated tool covering the full data science life cycle. they can be deployed in a local data center on top of Kubernetes or RedHat OpenShift. H2O Driverless AI, which covers the complete data science life cycle.

Cloud Based Tools

Watson Studio, together with Watson OpenScale, covers the complete development life cycle for all data science, machine learning, and AI tasks. Another example is Microsoft Azure Machine Learning. another example is H2O Driverless AI.

Data Management: Amazon Web Services DynamoDB, a NoSQL database that allows

storage and retrieval of data in a key-value or a document store format. The most prominent document data structure is JSON (pronounced “jay-sun”). Another flavour of such a service is Cloudant, which is a database-as-a-service offering. But, under the hood it is based on the open source Apache CouchDB. IBM offers Db2 as a service as well.

Data Integration:  Informatica Cloud Data Integration and IBM’s Data Refinery.

Data Visualization: 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: Watson Machine Learning. Google has a similar service on their cloud called AI Platform Training.

Model Deployment: SPSS Collaboration and Deployment Services. SPSS Modeler supports exporting models as Predictive Model Markup Language, or “PMML,” which can be read by numerous other commercial and open software packages. Watson Machine Learning can also be used to deploy a model and make it available to consumers using a REST interface. Amazon SageMaker Model Monitor is an example of a cloud tool that continuously monitors deployed machine learning and deep learning models.