The most widely used open-source data integration and transformation tools are relational databases, NoSQL Databases, file-based tools, elastic search tools, Extract, Transform, and Load (ETL), Data Refinery and Cleansing, Apache AirFlow, KubeFlow, Apache Kafka, Apache Nifi, Apache SparkSQL, and NodeRED. These tools are used for building, deployment, monitoring, and assessment, and for code and data asset management. They must be distinguished between programming libraries and tools containing a user interface. Pixie Dust, Hue, Kibana, Apache Superset, Seldon, TensorFlow, MLeap, ModelDB, Prometheus, and IBM AI Fairness 360 are all important tools for machine learning models. Model deployment is a crucial step and should be consumable by other developers and turned into an API. Model monitoring is an important step to track prediction performance while new data arrives. Model bias against protected groups like gender or race is also important.

The IBM Adversarial Robustness 360 Toolbox detects vulnerability against adversarial attacks and leverages the model to be more robust, while the IBM AI Explainability 360 toolkit addresses the training of a simpler machine learning model to explain the responsibility of different input variables. Git is the de facto standard for code asset management, but there are other services such as GitHub, GitLab, Bitbucket, ODPi Egeria, and Kylo.

Data management tools are MySQL, PostgreSQL, MongoDB, Apache CouchDB, Apache Cassandra, Hadoop File System, Ceph, and elastic search. Data integration and transformation tools are Apache AirFlow, KubeFlow, Apache Kafka, Apache Nifi, Apache SparkSQL, and NodeRED. Data Visualization tools are Pixie Dust, Hue, Kibana, and Apache Superset. Model deployment tools are Apache PredictionIO, Seldon, Kubernetes, Redhat OpenShift, Mleap, TensorFlow service, TensorFlow lite, and TensorFlow dot JS. Model monitoring tools are ModelDB, Prometheus, IBM AI Fairness 360, IBM Adversarial Robustness 360 Toolbox, and IBM AI Explainability 360. Code asset management tools are Git, GitHub, GitLab, and Bitbucket. Data asset management tools are Apache Atlas, ODPi Egeria, and Kylo.

Open-Source Tools for Data Science are used to compare and contrast different open-source tools, and describe the relevant features. Jupyter is the most famous development environment for interactive Python programming, while Apache Zeppelin provides a similar experience. RStudio is among the oldest development environments for statistics and data science, and it exclusively runs R and all associated R libraries. RStudio unifies programming, execution, debugging, remote data access, data exploration, and visualization into one tool. Spyder is an open-source tool for data scientists that integrates code, documentation, and visualizations into a single canvas.

Apache Spark is a batch data processing engine capable of processing vast amounts of data one by one or file by file. Apache Flink is a stream-processing image with its main focus on processing real-time data streams. Ray is the latest development in the data science execution environments and has a focus on large-scale deep learning model training. KNIME originated from the University of Konstanz in 2004 and has a visual user interface with drag-and-drop capabilities. Open-source tools are essential for data science, such as KNIME and Orange.

Commercial Tools for Data Science include data management tools, data integration and transformation tools, data visualization tools, model tools for building, deployment, monitoring, and assessment, and tools for code asset management, data asset management, development environment, and fully integrated visual. In data management, most of the relevant data is stored in an Oracle Database, Microsoft SQL Server, or IBM Db2. Data integration tools include extract, transform, and load (ETL) tools, while data visualization tools use business intelligence (BI) tools. Model deployment in commercial software is tightly integrated into the model-building process. SPSS Collaboration and Deployment Services is used to deploy any type of asset created by the SPSS software tools suite, and commercial software can export models in an open format.

Data asset management, often called data governance or data lineage, is a crucial part of enterprise-grade data science. Vendors provide tools for these tasks, such as the Information Governance Catalog, which facilitates the discovery of data assets.

Watson Studio is a fully integrated development environment for data scientists, combining Jupyter Notebooks with graphical tools to maximize performance. Commercial tools support the most common tasks in data science, such as Oracle Database, Microsoft SQL Server, IBM Db2, SAP, Oracle, SAS, Talend, Microsoft, and Watson Studio Desktop. Watson Studio and Watson Open Scale are a fully integrated tool covering the data science life cycle.

Cloud Based Tools for Data Science describe how commercial cloud tools support data science tasks, and how integration provides the ability to use the same tools for multiple tasks. The fully integrated visual tools category includes Watson Studio and Watson OpenScale, Microsoft Azure Machine Learning, and H2O Driverless AI. The cloud provider operates the tool for you in the cloud, backing up data and configuring and installing updates. Software-as-a-service (SaaS) versions of existing open source and commercial tools exist. Some proprietary tools are only available from a single cloud provider, such as Amazon Web Services DynamoDB and Cloudant. Commercial data integration tools include extract, transform, and load (ETL) tools and extract, load, and transform (ELT) tools.

Two commercial data integration tools widely used are Informatica Cloud Data Integration and IBM's Data Refinery, which allow transforming large amounts of raw data into consumable, quality information in a spreadsheet-like user interface. Datameer is a smaller company offering a cloud-based data visualization tool, and IBM offers its Cognos Business intelligence suite as a cloud solution.

Watson Studio provides visualizations to better understand data, such as a 3D bar chart, hierarchical edge bundling, 2D scatter plots, tree maps, word clouds, and AI Platform Training. Model building can be done using a service, such as Watson Machine Learning, which can train and build models using open-source libraries. Model deployment is integrated into the model-building process. SPSS Collaboration and Deployment Services can be used to deploy any asset created by the SPSS software tools suite, and commercial software can export models in an open format. Watson Studio and Watson OpenScale cover the complete development life cycle for all data science, machine learning, and AI tasks, and two commercial data integration tools are Informatica Cloud Data Integration and IBM's Data Refinery. Model building can be done using a service such as Watson Machine Learning, and Amazon SageMaker Model Monitor is an example of a cloud tool to monitor deployed machine-learning and deep learning models continuously.

Module 1 Summary

Congratulations! You have completed this module. At this point in the course, you know:

* The Data Science Task Categories are:
  + Data Management
  + Data Integration and Transformation
  + Data Visualization
  + Model Building
  + Model Deployment
  + Model Monitoring and Assessment
* Data Science Tasks support the following:
  + Data Asset Management
  + Code Asset Management
  + Execution Environments
  + Development Environments
* MySQL, PostgreSQL, MongoDB, Apache CouchDB, Apache Cassandra, Hadoop File System, Ceph, and elastic search are data management tools.
* Data integration and transformation tools are Apache AirFlow, KubeFlow, Apache Kafka, Apache Nifi, Apache SparkSQL, and NodeRED.
* Data Visualization tools are Pixie Dust, Hue, Kibana, and Apache Superset.
* Model deployment tools are Apache PredictionIO, Seldon, Kubernetes, Redhat OpenShift, Mleap, TensorFlow service, TensorFlow lite, and TensorFlow dot JS.
* Model monitoring tools are ModelDB, Prometheus, IBM AI Fairness 360, IBM Adversarial Robustness 360 Toolbox, and IBM AI Explainability 360.
* Code asset management tools are Git, GitHub, GitLab, and Bitbucket.
* Data asset management tools are Apache Atlas, ODPi Egeria, and Kylo.
* Data management tools are Oracle Database, Microsoft SQL Server, and IBM Db2.
* Data integration tools are provided by Informatica PowerCenter, and IBM Infosphere DataStage.
* Other products are from SAP, Oracle, SAS, Talend, Microsoft, and Watson Studio Desktop.
* Model building tools are SPSS Modeler, and SAS enterprise miner.
* Informatica and IBM provide data asset management tools.
* Watson Studio and Watson Open Scale are fully integrated tool covering the data science life cycle.
* In data management, with some exceptions, a software-as-a-service (SaaS) version of existing open-source and commercial tools exists.
* Two commercial data integration tools widely used are Informatica Cloud Data Integration and IBM’s Data Refinery.
* Two examples of cloud-based data visualization tools are Datameer and IBM’s Cognos Business Intelligence Suite.
* Watson Machine Learning can build models.