**DATA SCIENCE TOOLS IBM**

In this module, you will get an overview of the programming languages commonly used, including Python, R, Scala, and SQL. You’ll be introduced to the open source and commercial data science tools available. You’ll also learn about the packages, APIs, data sets and models frequently used by Data Scientists.

**Learning Objectives**

* Cite popular open source, commercial, and cloud-based tools used by data scientists.
* Explain the function of an API and list some common API-related terms.
* Discuss the characteristics of a dataset and the ways data can be structured.
* Identify  some of the libraries used in data science and the types of functionalities a library can provide.
* Compare and contrast between machine learning, deep learning and their use cases.
* Discuss the different classes of machine learning models.
* Summarize the advantages of using a pre-trained model as opposed to training a model from scratch.
* Identify the languages, tools, and data used by data scientists.
* Access and explore data sets in the Data Asset Exchange (DAX).
* Examine deep learning models on the Model Asset Exchange (MAX) and interact with an image-detection deep learning model.
* We previously covered open source tools for data science.
* Now, let’s look at the commercial options you’ll find in many enterprise projects.
* Let’s revisit our overview of different tool categories.
* In data management, most of an enterprise’s relevant data is stored in an
* Oracle Database, Microsoft SQL Server, or IBM Db2.

***We previously covered open source tools for data science.***

***Now, let’s look at the commercial options you’ll find in many enterprise projects.***

***Let’s revisit our overview of different tool categories.***

***In data management, most of an enterprise’s relevant data is stored in an***

***Oracle Database, Microsoft SQL Server, or IBM Db2.***

**Although open source databases are gaining popularity, those three data management products are still considered the industry-standard. They won’t disappear in the near future. It’s not just about functionality. Data is at the heart of every organization, and the availability of commercial supports plays a major role. Commercial supports are delivered directly from software vendors, influential partners, and support networks. When we focus on commercial data integration tools, we’re talking about “extract, transform, and load,” or “ETL” tools. According to a Gartner Magic Quadrant, Informatica Powercenter and IBM InfoSphere**

**DataStage are the leaders, followed by products from SAP, Oracle, SAS, Talend, and Microsoft.** Data is at the heart of every organization, and the availability of commercial supports

plays a major role. Commercial supports are delivered directly from software vendors, influential partners, and support networkS. When we focus on commercial data integration tools, we’re talking about “extract, transform, and load,” or “ETL” tools. According to a Gartner Magic Quadrant, Informatica Powercenter and IBM InfoSphere DataStage 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 also provide connectors to most of the commercial and open source target information systems. Finally, Watson Studio Desktop includes a component called Data Refinery, which enables the defining and execution of data integration processes in a spreadsheet style.In the commercial environment, data visualizations are utilizing business intelligence, or “BI”, tools. According to a Gartner Magic Quadrant, Informatica Powercenter and IBM InfoSphere DataStage 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 also provide connectors to most of the commercial and open source target information systems. Finally, Watson Studio Desktop includes a component called Data Refinery, which enables the defining and execution of data integration processes in a spreadsheet style. In the commercial environment, data visualizations are utilizing business intelligence, or “BI”, tools. Their main focus is to create visually attractive and easy-to-understand reports and live dashboards. The most prominent commercial examples are: Tableau, Microsoft Power BI, and IBM Cognos Analytics. Another type of visualization targets data scientists rather than regular users. A sample problem might be “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. The most prominent of these types of products are: SPSS Modeler and SAS Enterprise Miner. In addition, A version of SPSS Modeler is also available in Watson Studio Desktop, based on the cloud version of the tool. We’ll talk more about cloud-based tools in the next video. In commercial software, model deployment is tightly integrated in the model building process. This diagram shows an example of the SPSS Collaboration and Deployment Services which are used to deploy any type of asset created by the SPSS software tools suite. Other vendors use the same type of process. 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, 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 InfoSphere Information Governance Catalog covers functions like data dictionary, which facilitaThe data lineage also includes a reference to the actual source data. Rules and policies 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, and we’ll cover more about it in a later lesson. There is also a desktop version available. 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 and all the tasks we’ve discussed previously. We’ll talk more about both in the next lesson. but just keep in mind that they can be deployed in a local data center on top of Kubernetes or RedHat OpenShift. Another example of a fully integrated commercial tool is H2O Driverless AI, which covers the complete data science life cycle. In this lesson, you’ve learned how most common data science tasks are supported by commercial tools. In the next video, we’ll discover data science tools that are available exclusively on the cloud.tes discovery of data assets.**