Welcome to Categories of Data Science Tools. After watching this video, you will be able to list the tasks that a data scientist needs to perform show how code asset management and data asset management help build models, and describe how execution and development environments implement a model. Before it can be useful, raw data must pass through various Data Science task categories, such as data management, data integration and transformation, data visualization, model building, model deployment, and model monitoring and assessment. To do these tasks, you need data asset management, code asset management, execution environments, and development environments. Let’s see how each category enables you to make the best use of raw data. Data management is the process of collecting, persisting, and retrieving data securely, efficiently, and cost-effectively. Data is collected from many sources, like Twitter, Flipkart, Media, Sensors, and more. Store collected data in persistent storage so it is available whenever you need it. Data Integration and Transformation, is the process of Extracting, Transforming, and Loading data. This is called “ETL”. Some of this data is distributed in multiple repositories. For example, a database, a data cube, and flat files. Use the Extraction process to extract data from these numerous repositories and save to a central repository like a Data Warehouse. Data Warehouses are primarily used to collect and store massive amounts of data for data analysis. Next, Data Transformation is the process of transforming the values, structure, and format of data. After extracting the data, the next step is to transform the data. In this example, height and weight data needs to be transformed to metric. And once the data is transformed, it’s time to load the data. Transformed data is loaded back to the Data Warehouse. Data visualization is the graphical representation of data and information. You can use visualization to represent data in the form of charts, plots, maps, animations, etc. And data visualization conveys data more effectively for decision-makers. It is a crucial step in the data science process. Various forms of data visualizations include a bar chart, which compares the size of each component, a treemap, which displays hierarchy data, a line chart, which plots a series of data points over time, and a map chart, which displays data by location. Map charts can also be applied to other locations like websites. Now, model building is the next step. This is where you train the data and analyze patterns with machine learning algorithms. The system ‘learns’ how to provide predictions or decisions by itself. You can then use this model to make predictions on new, unseen data. Model building can be done using a service called IBM Watson Machine Learning. It provides a full range of tools and services for building models. The next step is model deployment: the process of integrating a developed model into a production environment. In model deployment, a machine learning model is made available to third-party applications via APIs. Business users can access and interact with the data through these third-party applications. And this helps them make data-based decisions. As an example, the SPSS Collaboration and Deployment Services can be used to deploy any type of asset created by the SPSS software tools suite.

4:01 Model monitoring and assessment run continuous quality checks to ensure a model’s accuracy, fairness, and robustness. Model monitoring uses tools like Fiddler to track the performance of deployed models in a production environment. Now, model assessment uses evaluation metrics like the F1 score, true positive rate, or the sum of squared error to understand a model's performance. A well-known example is the IBM Watson Open scale, which continuously monitors deployed machine learning and deep learning models. It will improve the accuracy and quality of your predictions. So, now that we've reviewed the data science task categories, let's look at some of the tools that support them. Code asset management provides a unified view where you manage an inventory of assets. When you want to develop a model, you may need to update it, fix bugs, or improve code features incrementally. All of this requires version control. Developers use versioning to track and manage changes to a software project's code. When working on a model, teams a centralized repository where everyone can upload, edit, and manage the code files simultaneously. Collaboration allows diverse people to share and update the same project together. GitHub is a good example of a code asset management platform. It's web-based and provides sharing, collaboration, and access control features. As a data scientist, you want to properly store and organize all your images, videos, text, and other data in a central location. You also want control over who can access, edit, and manage your data. Data asset management, also called digital asset management (DAM), is the organizing and managing of important data collected from different sources. DAM is performed on a DAM platform that allows versioning and collaboration. DAM platforms also support replication, backup, and access right management for the stored data. Development Environments, also called Integrated Development Environments, or "IDEs", provide a workspace and tools to develop, implement, execute, test, and deploy source code. IDEs like IBM Watson Studio provide testing and simulation tools to emulate the real world so you can see how your code will behave after it is deployed. An execution environment has libraries to compile the source code and system resources that execute and verify the code. Cloud-based execution environments are not tied to any specific hardware or software, and offer tools like IBM Watson Studio for data preprocessing, model training, and deployment. Finally, fully integrated visual tools like IBM Watson Studio and IBM Cognos Dashboard Embedded cover all the previous tooling components, and can be used to develop deep learning and machine learning models.

7:09 In this video, you learned that the Data Science Task Categories are: Data Management, Data Integration and Transformation, Data Visualization, Model Building, Model Deployment, and Model Monitoring and Assessment. Data Science Tasks are supported by Data Asset Management, Code Asset Management, Execution Environments, and Development Environments.