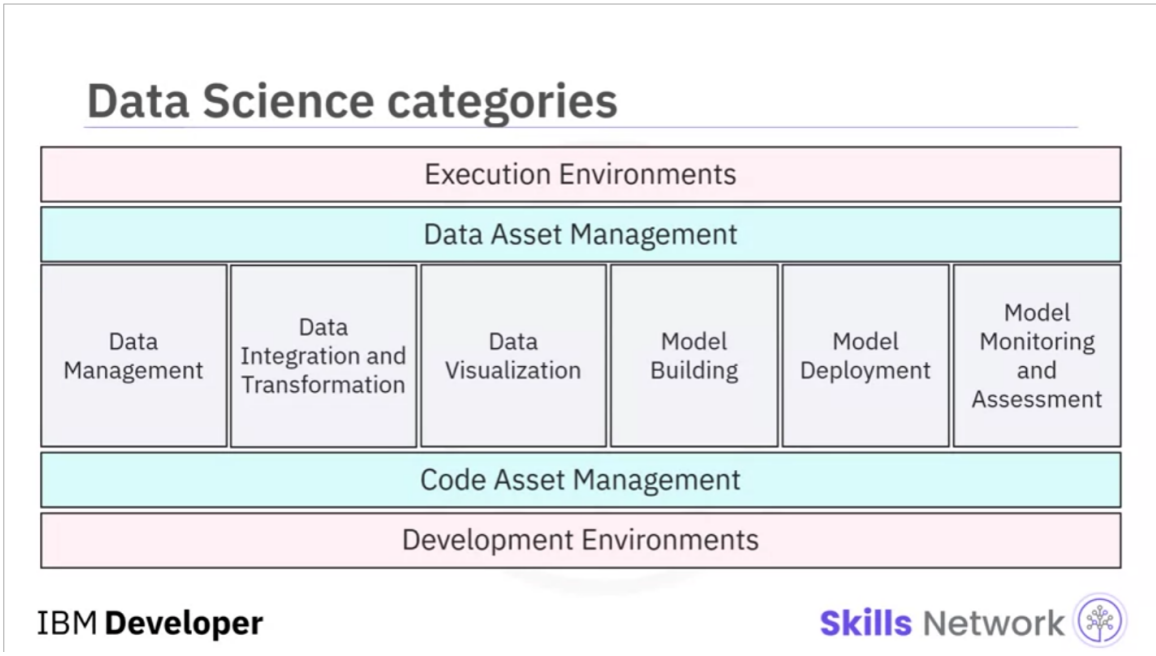
Tools for Data Science

▼ Week 1

Categories of Data Science Tools:



Open Source Tools for Data Science -part 1-:

- 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 -part 2-:

- Just use Jupyter lab.

Commercial Tools for Data Science:

- Commercial tools support the most common tasks in data science
- Data management tools are Oracle Database, Microsoft SQL Server, and IBM Db2
- Data integration tools are mainly provided by Informatica PowerCenter, IBM InfoSphere DataStage. These are followed by products from SAP, Oracle, SAS, Talend, Microsoft, and Watson Studio Desktop
- Model building tools are SPSS Modeler, and SAS enterprise miner
- SPSS Modeler is also available in Watson Studio Desktop
- Data asset management tools are provided by Informatica and IBM
- Watson Studio, together with Watson Open Scale is a fully integrated tool covering the data science life cycle

Cloud Based Tools for Data Science:

- Watson Studio and Watson OpenScale cover the complete development life cycle for all data science, machine learning, and AI tasks
- In data management, with some exceptions, there exists a software-as-a-service (SaaS) version of existing open source and commercial tools
- Two commercial data integration tools widely used are Informatica Cloud Data Integration and IBM's Data Refinery
- An example of a cloud-based data visualization tool is Datameer and IBM's Congos Business intelligence suite
- Model building can be done using a service such as Watson Machine Learning
- Amazon SageMaker Model Monitor is an example of a cloud tool to monitor deployed machine learning and deep learning models continuously