Categories of Data Science Tools

Key Learning Objectives

- List tasks a data scientist needs to perform.
- Understand how code asset management and data asset management contribute to building models.
- Describe the role of execution and development environments in model implementation.

Data Science Task Categories

1. **Data Management**
    - Involves collecting, storing, and retrieving data securely, efficiently, and cost-effectively.
    - Data is sourced from multiple platforms like Twitter, Flipkart, sensors, media, etc.
    - The data is stored in persistent storage to ensure it is always available.
2. **Data Integration and Transformation (ETL)**
    - **ETL**: Extracting, Transforming, and Loading data.
        - **Extraction**: Data is pulled from multiple repositories like databases, data cubes, and flat files.
        - **Transformation**: The extracted data is converted to suitable formats (e.g., converting height and weight into metric units).
        - **Loading**: The transformed data is loaded into a Data Warehouse, which is used for large-scale data analysis.
3. **Data Visualization**
    - The graphical representation of data using charts, plots, maps, and animations.
    - Visualization helps decision-makers comprehend data effectively.
    - **Common Visualization Types**:
        - **Bar Chart**: Compares component sizes.
        - **Treemap**: Displays hierarchical data.
        - **Line Chart**: Shows data trends over time.
        - **Map Chart**: Visualizes data by geographical location.
4. **Model Building**

- Involves training data with machine learning algorithms to uncover patterns and make predictions.
- The system 'learns' to provide insights or decisions based on the data.
- Example: IBM Watson Machine Learning provides tools and services for model building.

5. **Model Deployment**
   - Deploying the developed model into a production environment where it becomes accessible via APIs.
   - Allows business users to access and interact with the model through third-party applications for data-driven decision-making.
   - Example: SPSS Collaboration and Deployment Services are used for deploying assets created in SPSS tools.
6. **Model Monitoring and Assessment**
   - Continuous checks are made to ensure model accuracy, fairness, and robustness.
   - Tools like **Fiddler** help track model performance.
   - Assessment metrics include:
     - F1 Score.
     - True Positive Rate.
     - Sum of Squared Errors.
   - Example: IBM Watson OpenScale monitors machine learning and deep learning models to improve prediction quality.

Supportive Data Science Tools

7. **Code Asset Management**
   - Provides a unified view to manage code assets, enabling version control and collaboration among teams.
   - Developers use centralized repositories like **GitHub** for tracking changes, fixing bugs, and updating code.
   - Key Features:
     - Versioning.
     - Collaboration.
     - Access control.
8. **Data Asset Management (DAM)**
   - Involves organizing and managing data collected from various sources.
   - DAM platforms support:

        - Versioning and collaboration.
        - Replication and backup.
        - Access rights management.
    - Enables secure storage and control over who can manage and edit the data.
9. **Development Environments (IDEs)**
    - Provide a workspace and necessary tools for coding, testing, and deployment.
    - IDEs like **IBM Watson Studio** allow for simulation and testing of code to predict real-world performance after deployment.
10. **Execution Environments**
    - Provide libraries, system resources, and tools for compiling and executing code.
    - Cloud-based execution environments (e.g., **IBM Watson Studio**) are hardware-independent, offering services for data preprocessing, model training, and deployment.
11. **Integrated Tools**
    - Platforms like **IBM Watson Studio** and **IBM Cognos Dashboard Embedded** combine all aspects of data science tooling, from data management to model deployment and monitoring.

Summary of Data Science Tasks

- **Data Management**: Collect, store, retrieve data.
- **Data Integration and Transformation**: Extract, transform, and load data.
- **Data Visualization**: Use charts and graphs to present data.
- **Model Building**: Train models using machine learning.
- **Model Deployment**: Integrate models into production.
- **Model Monitoring and Assessment**: Continuously assess model performance.

Supportive Tools

- **Data Asset Management**: Organize and manage data.
- **Code Asset Management**: Manage code versions and collaboration.
- **Development Environments**: Workspaces for coding and deployment.
- **Execution Environments**: Compile and execute code.

Key Open-Source Tools for Data Science:

12. **Data Management Tools**:
    - **Relational Databases**: MySQL, PostgreSQL
    - **NoSQL Databases**: MongoDB, Apache CouchDB, Apache Cassandra
    - **File-Based Tools**: Hadoop File System, Cloud File systems like Ceph
    - **Text Storage**: Elastic search
13. **Data Integration and Transformation Tools** (ETL/ELT):
    - **Apache AirFlow** (created by Airbnb)
    - **KubeFlow** (executes pipelines on Kubernetes)
    - **Apache Kafka** (from LinkedIn)
    - **Apache Nifi** (visual editor for data flows)
    - **Apache SparkSQL** (handles massive data clusters)
    - **NodeRED** (low resource consumption, suitable for devices like Raspberry Pi)
14. **Data Visualization Tools**:
    - **Pixie Dust** (a Python library with a UI for plotting)
    - **Hue** (SQL-based visualizations)
    - **Kibana** (for data exploration, limited to Elasticsearch)
    - **Apache Superset** (data exploration and visualization)
15. **Model Deployment Tools**:
    - **Apache PredictionIO** (supports Apache Spark ML)
    - **Seldon** (supports TensorFlow, Apache SparkML, R, scikit-learn)
    - **MLeap** (deploys SparkML models)
    - **TensorFlow** (TensorFlow service, TensorFlow Lite, TensorFlow JS for web)
16. **Model Monitoring Tools**:
    - **ModelDB** (for Apache Spark ML and scikit-learn)
    - **Prometheus** (multi-purpose tool, not exclusive to ML)
    - **IBM AI Fairness 360** (detects and mitigates bias)
    - **IBM Adversarial Robustness 360 Toolbox** (detects model vulnerabilities)
    - **IBM AI Explainability 360** (improves model transparency and explainability)
17. **Code Asset Management Tools**:

- **Git** (version control)
- **GitHub**, **GitLab** (open-source, self-hosted option), **Bitbucket**

18. **Data Asset Management Tools**:
    - **Apache Atlas** (data governance and lineage)
    - **ODPi Egeria** (open APIs for data management)
    - **Kylo** (data management software platform)

These tools help handle the entire lifecycle of data science processes, from managing raw data to deploying and monitoring models.

Key Takeaways from Part 2:

1. Development Environments for Data Science:

- **Jupyter**:
    - Originally for interactive Python programming, but now supports over 100 programming languages via kernels.
    - **Jupyter Notebooks** unify documentation, code, output, shell commands, and visualizations.
    - **Jupyter Lab**: A modern, modular successor to Jupyter Notebooks with features like opening different file types and better layout control.
- **Apache Zeppelin**:
    - Inspired by Jupyter but with integrated plotting (no coding required for basic visualizations).
    - Extensible with additional libraries.
- **RStudio**:
    - Exclusive to the R environment, unifies programming, debugging, remote data access, and visualization.
    - Also integrates Python but remains most effective with R.

- **Spyder**:
    - Aimed at mimicking RStudio for Python.
    - While it lacks some features compared to RStudio, it is considered a solid Python IDE.

2. Cluster Execution Environments:

- **Apache Spark**:
    - Famous for its **batch data processing** capabilities.
    - Its key property is **linear scalability**: doubling the servers in a cluster roughly doubles its performance.
    - Used extensively across industries, including Fortune 500 companies.
- **Apache Flink**:
    - Focuses on **real-time data stream processing**.
    - Supports batch processing but is known for stream processing, making it more suitable for real-time applications.
- **Ray**:
    - A more recent development with a focus on **large-scale deep learning model training**.

3. Visual Data Science Tools (No Programming Required):

- **KNIME**:
    - Originated in 2004, this tool provides a visual interface with drag-and-drop capabilities.
    - Supports tasks like data integration, visualization, and model building.
    - Can be extended with R, Python, and connectors to Apache Spark.
- **Orange**:
    - Simpler than KNIME but easier to use for beginners.
    - Good for tasks like data visualization and basic model building but less flexible for advanced tasks.

Comparison and Contrast:

- **Jupyter vs. Zeppelin**: Jupyter requires external libraries for plotting, while Zeppelin has integrated plotting. Jupyter Lab offers a more modular and versatile interface compared to the traditional Jupyter Notebooks.

- **RStudio vs. Spyder**: RStudio excels with R, integrating a wider range of functionality, while Spyder is a lightweight alternative for Python users.
- **Apache Spark vs. Apache Flink**: Spark is primarily for batch processing, while Flink specializes in real-time stream processing. Both tools support the other paradigm, but their focus differs.
- **KNIME vs. Orange**: KNIME is more flexible and extensible, while Orange is simpler and easier for beginners without requiring programming knowledge.

This part of the series emphasizes the versatility and variety of open-source tools available for data scientists, tailored to different tasks and user expertise levels.

Key Takeaways from "Commercial Tools for Data Science":

1. Data Management Tools:

- **Oracle Database**, **Microsoft SQL Server**, and **IBM Db2** are the industry-standard commercial databases for enterprise data management.
- Although open-source databases are becoming popular, these commercial products are widely used due to their reliability and the commercial support provided by vendors and partners.

2. Data Integration and Transformation Tools (ETL):

- Leading tools in this category include **Informatica PowerCenter** and **IBM InfoSphere DataStage**.
- Other significant players are **SAP**, **Oracle**, **SAS**, **Talend**, and **Microsoft**.
- **Watson Studio Desktop** offers a component called **Data Refinery**, which allows for ETL processes using a spreadsheet-like interface.
- These tools provide graphical interfaces for designing and deploying ETL pipelines, with connectors to both commercial and open-source systems.

3. Data Visualization Tools:

- For business intelligence (BI) and visualization, the most common tools are **Tableau**, **Microsoft Power BI**, and **IBM Cognos Analytics**.
- These tools focus on creating visual reports and live dashboards for business users.
- For data scientists, visualization of data relationships (e.g., relationships between columns) is supported in **Watson Studio Desktop**.

4. Model Building, Deployment, and Monitoring:

- Leading model-building tools include **SPSS Modeler** and **SAS Enterprise Miner**.
- **SPSS Modeler** is also available in **Watson Studio Desktop** and integrates with other tools from IBM.
- These tools allow exporting models in **Predictive Model Markup Language (PMML)**, a format compatible with other commercial and open-source tools.
- **Model deployment** is tightly integrated with these commercial tools, such as **SPSS Collaboration and Deployment Services** for SPSS models.
- **Model monitoring** is a relatively new field, with no widespread commercial solutions yet, making open-source tools a primary choice.

5. Code and Data Asset Management:

- For **code asset management**, open-source tools like **Git** and **GitHub** remain the industry standard.
- **Data asset management** (also known as **data governance** or **data lineage**) is vital in enterprise data science.
    - **Informatica Enterprise Data Governance** and **IBM Information Governance Catalog** are prominent tools in this space.
    - These tools allow for metadata management, assigning data stewards or owners, and tracking data lineage from the source to its transformed state.
    - They also ensure compliance with regulatory and business policies regarding data privacy and retention.

6. Fully Integrated Development Environments:

- **Watson Studio**, particularly when paired with **Watson Open Scale**, is a fully integrated development environment (IDE) for data scientists, covering all stages of the data science lifecycle.

    - Watson Studio combines tools like **Jupyter Notebooks** with graphical interfaces to optimize data science workflows.
    - It can be deployed in the cloud or in local environments using Kubernetes or RedHat OpenShift.
- **H2O Driverless AI** is another example of a commercial tool that covers the complete data science lifecycle.

Summary of Tools by Category:

- **Data Management**: Oracle Database, Microsoft SQL Server, IBM Db2.
- **Data Integration and Transformation**: Informatica PowerCenter, IBM InfoSphere DataStage, SAP, Oracle, SAS, Talend, Microsoft, Watson Studio Desktop.
- **Data Visualization**: Tableau, Microsoft Power BI, IBM Cognos Analytics, Watson Studio Desktop.
- **Model Building**: SPSS Modeler, SAS Enterprise Miner.
- **Data Asset Management**: Informatica Enterprise Data Governance, IBM Information Governance Catalog.
- **Fully Integrated Environments**: Watson Studio with Watson Open Scale, H2O Driverless AI.

In conclusion, commercial tools support a wide array of tasks in data science, providing enterprise-grade functionality and support across data management, integration, visualization, model building, and governance.

Key Takeaways from "Cloud Based Tools for Data Science":

1. Fully Integrated Cloud Platforms:

- **Watson Studio** and **Watson OpenScale** offer a complete development life cycle for data science, machine learning, and AI tasks, utilizing cloud-based compute clusters.
- **Microsoft Azure Machine Learning** also provides full cloud-hosted support for data science tasks, from development to deployment.
- **H2O Driverless AI** is another tool, which, while downloadable, can be easily deployed on cloud platforms.

2. Data Management in the Cloud:

- Many commercial tools have software-as-a-service (**SaaS**) versions that operate in the cloud, relieving users of operational responsibilities like backups and updates.
- **Amazon DynamoDB** is an example of a NoSQL database offered exclusively as a cloud service, using a key-value or document store format like JSON.
- **IBM Cloudant** is a database-as-a-service offering built on **Apache CouchDB**, which enables seamless migration to CouchDB servers.
- **IBM Db2 as a Service** is a commercial database provided as a SaaS cloud service, automating operational tasks for users.

3. Cloud-Based Data Integration and Transformation Tools:

- Commercial data integration tools such as **Informatica Cloud Data Integration** and **IBM's Data Refinery** allow data engineers and data scientists to transform large datasets in a cloud environment.
- **Data Refinery** is integrated with Watson Studio and offers a user-friendly, spreadsheet-like interface for ETL (Extract, Transform, Load) operations.

4. Cloud-Based Data Visualization Tools:

- **Datameer** is an example of a smaller company providing cloud-based data visualization solutions.
- **IBM Cognos Business Intelligence Suite** is available as a cloud solution, along with visualization functionalities within **Watson Studio**'s **Data Refinery**.

- Watson Studio supports various data visualizations, such as 3D bar charts, hierarchical edge bundling, scatter plots with heat maps, and word clouds.

5. Model Building and Deployment:

- Model building services like **Watson Machine Learning** use open-source libraries to train and build models in the cloud.
- **Google AI Platform Training** is a similar cloud-based service offered by Google for machine learning tasks.
- For model deployment, **SPSS Collaboration and Deployment Services** allow for the deployment of any model built within the SPSS suite.
- Many cloud platforms, including Watson Machine Learning, allow models to be deployed via **REST interfaces**, making them accessible to end users and other applications.

6. Model Monitoring:

- **Amazon SageMaker Model Monitor** is a cloud-based tool that continuously monitors machine learning and deep learning models.
- **Watson OpenScale** provides similar functionality, monitoring deployed models and ensuring they perform as expected over time.

Summary of Tools and Services:

- **Fully Integrated Cloud Platforms**: Watson Studio, Watson OpenScale, Microsoft Azure Machine Learning, H2O Driverless AI.
- **Data Management (SaaS)**: Amazon DynamoDB, IBM Cloudant, IBM Db2 as a Service.
- **Data Integration & Transformation**: Informatica Cloud Data Integration, IBM's Data Refinery.
- **Data Visualization**: Datameer, IBM Cognos, Watson Studio Data Refinery.
- **Model Building**: Watson Machine Learning, Google AI Platform Training.
- **Model Deployment**: SPSS Collaboration and Deployment Services, Watson Machine Learning REST interfaces.
- **Model Monitoring**: Amazon SageMaker Model Monitor, Watson OpenScale.

In conclusion, cloud-based tools offer comprehensive support for data science tasks, including data management, integration, visualization, model building, deployment, and

monitoring, with the advantage of scaling, automation, and integration across tasks within cloud environments.