**IBM Data Science Coursera Certificate – Tools for Data Science (Course 2)**

**Week 1**

**Notes:**

**Categories of Data Science Tools**

A picture containing text, screenshot, font, number

Description automatically generated

**Open-Source Tools for Data Science – Part 1**

Objective of this part given below:

1. After this video, I will be able to list the open source data management tools
2. I will be able to list open source integration and transformation tools
3. List the open source data visualization tools
4. List the model tools for building, deployment, monitoring, and assessment

**Open Source Data Management Tools:**

* Open source data management tools are relational databases like
  + Most widely used **relational databases** like MySQL, PostgreSQL
  + Also, there are **NoSQL** Databases like MongoDB, Apache CouchDB, and Apache Cassandra. In addition, there are **file-based tools** like the **Hadoop File System** or Cloud File systems like Ceph. You also have an **Elasticsearch** tool that stores text data, including the creation of a search index for fast document retrieval.

**Open Source Data Integration and Transformation Tools:**

* Data scientists often propose **Extract, Load, Transform (ELT)** as data is dumped somewhere, and the data engineer or data scientist handles the transformation of the data.
* Another term for this process emerged as **Data Refinery and Cleansing**.
* Tools used for data integration and transformation are
  + Apache Airflow (created by AirBnB)
  + Kubeflow
  + Apache Kafka
  + Apache nifi
  + Apache Spark SQL
  + Node-RED

**Open Source Data Visualization Tools:**

* Supported by programming libraries where you need to use code
* Contain a user interface
* Open Source Tools used for data visualization are
  + PixieDust - is a library but has a user interface that facilitates plotting in python.
  + Hue - visualization tool that users SQL queries to plots the graphs
  + Kibana -
  + Apache Superset -

**Open Source Tools For Model Deployment:**

* Model deployment is a crucial step. Once you’ve created a machine learning model capable of predicting some critical aspects of the future, you should make it consumable by other developers and turn it into an API.
* Tools used for Model Deployment are
  + **Apache Prediction IO** – Apache Prediction IO currently only supports Apache Spark ML models for deployment, but support for all libraries is on the roadmap.
  + **Seldon** - is an interesting product since it supports nearly every framework including TensorFlow, Apache SparkML, R, and scikit learn. Interestingly, it can run on top of Kubernetes and Redhat OpenShift.
  + **MLeap** - Another way to deploy SparkML models is MLeap. Finally,
  + **TensorFlow -** TensorFlow can serve any TensorFlow model using the TensorFlow service. It can be an embedded device like a Raspberry Pi or smartphone using TensorFlow lite and deployed to a web browser using TensorFlow.JS

**Open Source Tools for Model Monitoring and Assessment**

* Tools to keep track of machine learning models’ prediction performance to maintain outdated models.
* Tools used for data integration and transformation are
  + **ModelDB** - a system that manages ML models and ModelDB is a machine model metadata base where information about the models is stored and queried. It natively supports Apache Spark ML Pipelines and scikit-learn.
  + **Prometheus** - A generic, multi-purpose tool. Although it is not specifically made for machine learning model monitoring, it is used for this purpose as well.

**Open Source Tools For Code Asset Management**

* Tools for code asset management, also known as version management or version control
* Open Source Tools used for code asset management
  + **Git**
  + **GitHub**
  + **GitLab**
  + **Bigbucket**

**Open Source Tools For Data Asset Management**

* Tools for data asset management, also known as data governance or data lineage
* Is a crucial part of enterprise-grade data science.
* Data has to be versioned and annotated with metadata.
* Open Source Tools used for management asset management
  + **Apache Atlas -**
  + **ODPI – EGERIA :** managed through the Linux Foundation, is an open ecosystem offers a set of open APIs, types, and interchange protocols that metadata repositories use to share and exchange data
  + **Kylo -** is an open-source data management software platform, with extensive support for data asset management tasks

**Summary :**

* Data management tools are MySQL, PostgreSQL, MongoDB, Apache CouchDB, Apache Cassandra, Hadoop File System, Ceph, and Elasticsearch.
* Data integration and transformation tools are Apache AirFlow, KubeFlow, Apache Kafka, Apache Nifi, Apache Spark SQL, 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.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.
* And finally, data asset management tools are Apache Atlas, ODPi Egeria, and Kylo.

**Open-Source Tools for Data Science – Part 2**

**Development Environment:**

**Jupyter:** Currently, the most famous development environment data scientists are using is “Jupyter,” which emerged as a tool for interactive Python programming. Jupyter now supports more than a hundred different programming languages through “kernels.” This encapsulates the execution environment for the different programming languages. A key property of Jupyter Notebooks is to unify documentation, code, output from the code, shell commands, and visualizations in a single document.

**Jupyter Lab:** Jupyter lab is the next version of Jupyter Notebooks, and in the long term will replace Jupyter Notebooks. The abundance of architectural changes makes Jupyter more modern and modular. From a user’s perspective, the main difference between Jupyter Lab and Jupyter Notebooks is the ability to open different types of files, including Jupyter Notebooks, data, and terminals, and then arrange them on the canvas.

**Apache Zeppelin:** Apache Zeppelin was inspired by Jupyter Notebooks and provides a similar experience. One key differentiator is the integrated plotting capability. In Jupyter Notebooks, you are required to use external libraries and in Zeppelin, plotting doesn’t require coding. You can also extend the capabilities by using additional libraries.

**RStudio:** In the R environment, Python development is possible. R is tightly integrated into the Jupyter tool and provides optimal user experience. RStudio unifies programming, execution, debugging, remote data access, data exploration, and visualization into one tool.

**Spyder:** Spyder tries to mimic the behavior of RStudio to bring its functionality to the Python world. Although not at par with the functionality of RStudio, data scientists consider it as an alternative.

**Execution Environment:**

**Apache Spark:** It provides the cluster execution environment

The extensively famous Apache Spark is among the most active Apache projects that are used across all industries, including many Fortune 500 companies. The key property of Apache Spark is linear scalability. This means that if you double the number of servers in a cluster, you’ll roughly double its performance.

**Apache Flink:** It is a stream-processing image with its main focus on processing real-time streams.

*The key difference between Apache Spark and Apache Flink is that Apache Spark is a batch data processing engine, capable of processing vast amounts of data one by one or file by file. Whereas Apache Flink is a stream-processing image with its main focus on processing real-time data streams. Meaning that process the infinite data in real-time.*

**RiseLab Ray:** Ray is one of the latest developments in the data science execution environments and has a clear focus on large-scale deep learning model training.

**Programming Free Open-Source Tools**

**KNIME:**

* drag and drop capabilities,
* Built-in visualization capabilities
* Can be extended by programming in R and Python
* Has connectors to Apache Spark

**Orange:** It is one of the programming free open-source tools which is less flexible than KNIME but easier to use.

**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 and IBM InfoSphere DataStage. These are followed by products from SAP, Oracle, SAS, Talend, Microsoft, and IBM Watson Studio Desktop.

**Data Visualization tools** are Tableau, PowerBI, and IBM Cognos analytics.

**Model building tools** are SPSS Modeler, and SAS enterprise miner. SPSS Modeler is also available in Watson Studio Desktop.

Informatica and IBM provide **data asset management tools**. Data Asset Management Functions includes:

* Data governance
* Data versioned and annotated
* Data Dictionary
* Data Lineage
* Data Privacy and retention

And finally, Watson Studio, together with Watson Open Scale is a fully integrated tool covering the data science life cycle.

H2O.ai is also a fully integrated tool for data science life cycle and it is not deployed by cloud based service provider. It needs to be downloaded. Simply, It is not cloud-based tool.

**Cloud Based Tools for Data Science**

Fully Integrated Visual Tools categories:   
  
Large-scale execution of data science workflows happens in compute clusters:

1. Watson Studio and Watson OpenScale cover the complete development life cycle for all data science, machine learning, and AI tasks.
2. Microsoft Azure Machine Learning is a cloud-hosted tool offering complete development life cycle for data science, machine learning, and AI tasks.
3. H2O.ai is a product you can download and install. It is one click deployment for the common cloud service providers. This cloud provider doesn’t do operations and maintenance. Also, it needs to be downloaded to work with.

**Data Management Cloud-Based Tools**

* AWS Amazon DynamoDB (No SQL required)
* IBM DB2
* Cloudant
* Apache CouchDB Relax

**Data Integration and Transformation Cloud-Based Tools**

Commercial data integration tools that include extract, transform, and load (ETL) tools and extract, load, and transform (ELT) tools. It means the transformation steps are not done by a data integration team but are pushed toward the domain of the data scientist or data engineer. Two commercial data integration tools widely used are **Informatica Cloud Data Integration** and **IBM’s Data Refinery.**

* Informatica
* IBM Data Refinery

**Cloud Based Data Visualization Tools**

In the market of cloud based software, almost every major vendor has cloud based data visualization tool. For a smaller company offering a cloud-based data visualization is Datameer.

* Datameer
* IBM Cognos Analytics
* And much more

**Cloud Based Model Building Tools**

Model building can be done using a service. One example of a service is **Watson Machine Learning**. Watson Machine Learning can train and build models using various open-source libraries.

**Google Cloud** has a similar service on their cloud called **AI Platform Training**. Every cloud provider has a solution for this task.

Model deployment in commercial software is **usually tightly integrated into the model-building process**.

**Cloud Based Model Monitoring & Assessment**

Amazon SageMaker Model Monitor is an example of a cloud tool to monitor deployed machine learning and deep learning models continuously. Every major cloud provider has similar tooling. Another tool for model monitoring is Watson OpenScale. Everything marked in green can be done using Watson Studio and OpenScale.

**Summary:**

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 Cognos 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.