# Data Science Tasks

## Data management

* Process of collecting, persisting, and retrieving data securely, efficiently, and cost-effectively.
* Data comes from sources eg. Twitter, Flipkart, Media, Sensors, and more.
* Store collected data in persistent storage so it is available whenever you need it.
* Open-Source Tools:
  + Relational databases – MySQL, PostgreSQL.
  + NoSQL Databases - MongoDB, Apache CouchDB, Apache Cassandra
  + File-based tools - Hadoop File System
  + Cloud File systems – Ceph
  + Elastic search tool that stores text data, including the creation of a search index for fast document retrieval.
* Commercial Tools:
  + Oracle Database
  + Microsoft SQL Server
  + IBM Db2.

## Data Integration and Transformation (a.k.a  Data Refinery and Cleansing.)

* Process of Extracting, Transforming, and Loading data (ETL).
  + **Extraction** from multiple repositories eg. database, data cube, flat files & save them to a central repository like a Data Warehouse.
    - Data Warehouses are primarily used to collect and store massive amounts of data for data analysis.
  + **Transformation** of values, structure, and format of data and loading them back to Data Warehouse
  + **visualization** the graphical representation of data and information in the form of charts, plots, maps, animations, etc. to convey data more effectively for decision makers
* Open-Source Tools:
  + Open Source Extraction & Transformation
    - Apache AirFlow - created by Airbnb
    - KubeFlow - allows the execution of data science pipelines on top of Kubernetes
    - Apache Kafka - originated from LinkedIn
    - Apache Nifi - delivers a very nice visual editor
    - Apache SparkSQL - lets you use ANSI SQL and scales up to compute clusters of thousands of nodes
    - NodeRED - low in resource consumption that it even runs on tiny devices like a Raspberry Pi.
  + Open Source Data Visualization
    - Pixie Dust - library with a user interface that facilitates plotting in Python
    - Hue - create visualizations from SQL queries
    - Kibana - a data exploration & visualization web application limited to Elasticsearch
    - Apache Superset - data exploration and visualization web application.
* Commercial Tools:
  + Support the design and deployment of ETL data processing pipelines through a graphical interface.
  + Bring along connectors to most of the commercial and open-source target information systems.
  + Leaders (according to a Gartner Magic Quadrant, Informatica):
    - Informatica PowerCenter
    - IBM InfoSphere DataStage
  + Others
    - SAP
    - Oracle
    - SAS
    - Talend
    - Microsoft products.
  + Watson Studio Desktop
    - Data Refinery enables definition and execution of data integration processes in a spreadsheet-style.
  + Data visualization tools:
    - Tableau
    - Microsoft Power BI
    - IBM Cognos Analytics.
    - Watson Studio Desktop.

## Model Building

* Process of training the data and analyze patterns with machine learning algorithms
* The system ‘learns’ the patterns and provide predictions on new, unseen data.

## Model deployment

* Process of integrating a developed model into a production environment.
* A machine learning model is made available to third-party applications via APIs.
* Business users use these APIs to make data-based decisions.
* Open-Source Tools:
  + Apache PredictionIO - supports Apache Spark ML models for deployment
  + Seldon - supports nearly every framework including, TensorFlow, Apache SparkML, R, and sklearn. Can run on top of Kubernetes and Redhat OpenShift
  + MLeap – to deploy SparkML models
  + TensorFlow service - serve any tensor flow model.
    - TensorFlow Lite – for Raspberri Pi & Smartphones
    - TensorFlow dot js – for Web Browser
* Commercial Tools:
  + SPSS Modeler
    - Also available in Watson Studio Desktop, based on the tool’s cloud version.
    - Supports exporting models as predictive model markup language (PMML), which an abundance of other commercial and open software packages can read.
  + SAS Enterprise Miner
  + SPSS Collaboration and Deployment Services
    - Used to deploy any type of asset created by the SPSS software tools suite.

## Model Monitoring & Assessment

* Runs continuous quality checks to ensure a model’s accuracy, fairness, and robustness.
  1. Model Monitoring
     + Uses tools like Fiddler to track the performance of deployed models in a production environment.
  2. Model Assessment
     + Uses evaluation metrics like the F1 score, true positive rate, or the MSE to understand a model's performance.
* Improve the accuracy and quality of your predictions through regular monitoring & assessment.
* Open-Source Tools:
  + ModelDB - A machine model metadata base where information about the models is stored and queried. Natively supports Apache Spark ML Pipelines and scikit-learn
  + Prometheus – generic but multipurpose but still used for this
  + IBM AI Fairness 360 open-source toolkit - detects and mitigates bias in machine learning models because models are subject to adversarial attacks where an attacker tries to mislead the model with manipulated data or by controlling it
  + IBM Adversarial Robustness 360 Toolbox – detects vulnerability against adversarial attacks and leverages the model to be more robust.
  + IBM AI Explainability 360 toolkit - addresses that problem by finding similar examples in a dataset to be presented to an end-user for manual comparison. It can also address the training of a simpler machine learning model to explain the responsibility of different input variables directed toward the final decision of the model.

# Data Science Tasks

## Code asset management

* provides a unified view where you manage an inventory of assets
* Use version control to update, debug, and improve code features incrementally of a model.
* Developers use versioning to track and manage changes to a software project’s code.
* While working on a model, teams need a centralized repository where everyone can upload, edit, and manage the code files simultaneously.
* Data scientist want to store and organize all images, videos, text, and other data in a central location besides controlling over who can access, edit, and manage your data.
* Open-Source Tools:
  + Git – also known as version management or version control.
    - GitHub – most prominent
    - GitLab - entirely open source and can be hosted and managed on your own.
  + Bitbucket.

## Digital/Data Asset Management (DAM) a.k.a Data Governance / Data Lineage

* Organizing and managing of important data collected from different sources.
* Performed on a DAM platform that allows versioning and collaboration. They also support replication, backup, and access right management for the stored data.
* Open-Source Tools:
  + 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 - open-source data management software platform with extensive support for data asset management tasks.
* Commercial Tools
  + Informatica Enterprise Data Governance and IBM.
    - Covers functions like a data dictionary, which facilitates the discovery of data assets.
    - Each data asset is assigned to a data steward or the data owner, which is then responsible for that data asset and can be contacted. Then, data lineage is covered, allowing tracking back the transformation steps in creating the data assets. The 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 retention.
  + Watson Studio
    - Fully integrated development environment for data scientists.
    - Most people consume it through the cloud but desktop version also available.
    - Combines Jupyter Notebooks with graphical tools to maximize the performance of data scientists.
    - Together with Watson Open Scale, is a fully integrated tool covering the data science life cycle
    - Can be deployed in a local data center, on top of Kubernetes / RedHat OpenShift.
* H2O Driverless AI, which covers the complete data science life cycle.

## Integrated Development Environments (IDE)

* Provide a workspace and tools to develop, implement, execute, test, and deploy source code.
* Open-Source Tools:
  + Jupyter
    - a tool for interactive Python programming
    - supports more than a hundred different programming languages through “kernels.”
    - Unifies documentation, code, output from the code, shell commands, and visualizations in a single document.
  + Jupyter lab
    - Next version of Jupyter Notebooks, will replace Jupyter Notebooks in the long term
    - Better than Jupyter in terms of ability to open different types of files including Jupyter Notebooks, data, and terminals, and then arrange them on the canvas.
  + Apache Zeppelin
    - Inspired by Jupyter Notebooks and provides a similar experience but with integrated plotting capabilities.
  + RStudio
    - among the oldest development environments for statistics and data science.
    - Has origins in the year 2011.
    - It exclusively runs R and all associated R libraries but Python development is also possible.
    - Tightly integrated into the Jupyter tool and provides optimal user experience.
    - Unifies programming, execution, debugging, remote data access, data exploration, and visualization into one tool.
  + Spyder
    - Tries to mimic RStudio using Python language but not much functionality
    - Integrates code, documentation, and visualizations, among others, into a single canvas.

## Cluster Execution environment

* Has libraries to compile the source code and system resources that execute and verify the code.
* Exists because sometimes your data doesn’t fit into a single computer’s storage or main memory capacity.
* Open-Source Tools:
  + Apache Spark
    - Most active Apache projects that are used across all industries, including many Fortune 500 companies.
    - Has linear scalability, where doubling the number of servers in a cluster roughly double its performance.
    - Has batch data processing engine capable of processing vast amounts of data one by one or file by file.
  + Apache Flink
    - Developed after Apache Spark continued to gain market share
    - Stream-processing image focusing on processing real-time data streams.
  + Ray
    - Has a clear focus on large-scale deep learning model training.

## Fully Integrated and Visual Tools

* No programming language required
* The tools support a subset of important tasks that include data integration and transformation, data visualization, and model building.
* Open-Source Tools:
  + KNIME
    - Originated from the University of Konstanz in 2004.
    - Has a visual user interface with drag-and-drop capabilities.
    - Has built-in visualization capabilities
    - Can be extended by programming in R and Python and even has connectors to Apache Spark.
  + Orange
    - less flexible than KNIME but is easier to use.