Data Science Tools

Data Science Tools Categories

Data Must pass through the data science task stages

Data Management

Data Integration and Transformation

Data Visualization

Model Building

Model Deployment

Model Monitoring and Assessment

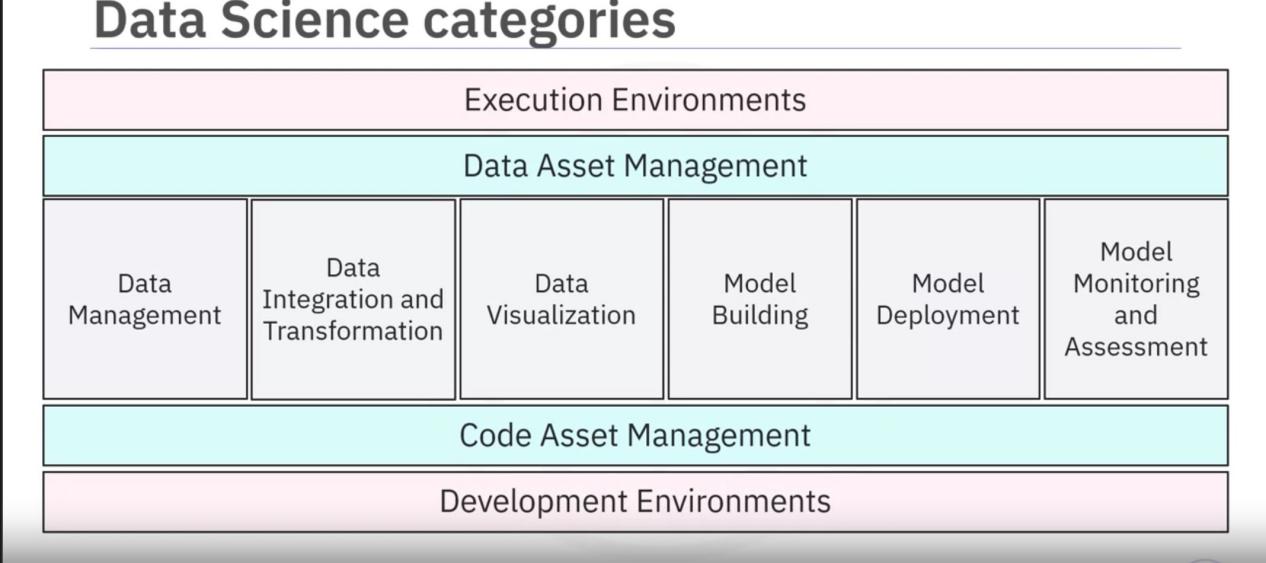
To do these we need

Data Asset Management

Code Asset Management

Execution Environments

Development Environments



Data Management:

Collecting, persisting, and retrieving data securely, efficiently, and cost-effectively.

Collected from many sources like social media, traditional media, Internet, Sensors and more and stored in persistent storage so its available when we need it.

Data Integration and Transformation:

ETL(Extract, Transform, and Load)

Data is distributed in different repositories. Like databases, data cube, Flat files.

WE extract the data , process it and save it in a central repo like in a data warehouse.

Used to Collect Massive amount of data for data analysis.

The process of transforming the values, structure and format of the data is called Data Transformation.

Example:

Converting units of a field. (From Imperial to Metric or Celcius to Fahrenheit) or converting continuous data to categorical data by grouping.

Once the transformation is finished the data is loaded back to the data warehouse.

Data Visualization:

The graphical representation of the data and information is called Data Visualization.

Charts, plots, maps, animations etc.

Conveys the info better

Examples:

Bar chart - Compares size of each component

TreeMap- shows hierarchy

line chart- plots a series of points over time.

Map Chart - data by location, can be applied to other locations like websites.

Model Building:

This is where we train the data and analyze patterns using suitable Machine learning algorithm where the system learning to provide solutions by itself.

We can use this to predict things

Model Deployment:

The process of integrating a model into a production environment.

Model Deployment uses a APIs to enable data-based decisions to which third party applications can connect and iteract with.

Model Monitoring and Model Assessment:

Track deployed model

Performs quality checks on the model to ensure accuracy,fairness and robust ness.

Uses tools like ‘fiddler’ to mesasure the accuracy

Uses metrics like F1 Score , True positive rate, Sum of Squared error to understand the peformance of the model.

IBM Watson Open Scale is an open source monitoring and assessment tool that continuosly monitor the deployed ML and Deep Leaning models helping you to improve the prediction accuracy of your model.

Tools for the different Data Science task categories:

Code Asset Management:

It is a unified view where you manage an inventory of assets. When working with data science you may want to improve it, perform bug fix or improve code features incrementally. So we need to have version control tools

Version Control tracks and manages the code changes to the software projects

Teams use a centra repo where each member can perform changes to the code simultaneously.

Collaboration allows different people to share and update the same project together allowing faster development.

Github.

Data Asset Management:

As Data Scientist you need to maintain all the code, files, images, etc. in central location and also you need to to have control on who can access and edit the data,

Data Asset Management is a platform for organizing and managing the data which has versioning and colaboration.

It might also have features like Replication, backup, and access right management.

Development Environment:

Also called Integrated Development Environment(IDEs) provide a workspace and tools to work on source code.

Helps us to perform following tasks

Develop

Implement

Execute`

Test

Deploy

Execution Environment:

Has libraries for the code compiling and system resources to execute and verify the code.

Cloud tools

IBM Watson Studio

IBM Cognos Dashboard Embedded

These are fully- integrated visual tools cover all tooling components and can be used to develop Deep learning and ML models.

Open Source Tools for Data Science

rdbms--mysql, postgres

nosql-- mogodb, couchdb, cassandra

filebased tools - hadop hdfs, ceph, elasticseach(helps to store files, seach and create indexes for fast retrieval).

Data Integration and Transformation:

ELT or ELT

Data Reinery and Cleansing

Apache Airflow(originally developed by Air bnb)

Kubeflow - allows execution of data science pipelines on Kubernetes

Apache Kafka -Originated from linked in

Apache nifi - provide a nice visueal editor

Apache Spark SQl - Lets you see sql and helps you create a 1000s of cluster nodes

Node-red -- visual editor and low resource consumption.

Data Visualization:

2 types:

programming

UI

PixieDust library and Ui,

Hue from sql

Kibana - data exploration and visualization web tool limited to elastic search or data provider.

Acache Superset - data exploration and visualization web tool

Model Deployment:

Once we have generated a model that performs preditions, we have to turn it into API so other devs will be able to access the model.

Apache PredictionIO - supports only Apache spark Ml models for deployment. Support for other libraries is on the way.

Seldon - supports all libraries like tensorflow, apache spark ML, R and scikitlearn. It can run on Kubernetes and Redhat Openshift

Mleap - another way to run Ml model

TensorFlow Serving - can run and serve any Tensorflow model using a tensor flow service. It can be In any embeded device like raspberry pi and can be deployed to web using Tensorflow.js

Model Monitoring and Assessment:

Tools to keep track of the ML model’s prediction performance to maintain oudated models.

ModelDb - ML machine model meta database where information is stored and queried. It natively supports the Apache Spark pipeline and scikitlearn.

Prometheus - No specifically made for ML monitoring but it still is being used.

Model monitoring is not only done by model accuracy but also model biases like gender or race is also important.

IBM AIfairness 360 Open Source Toolkit measures and mitigates biases in ML models

These models especially Neural network and deep learning models can be subjected to adverserial attacks where the attacker could mislead the model by providing manipulated data that favors one gender or race.

IBM Adversarial Robustness 360 Toolbox detects vulnarability of a model agains the adverserial attacks and helps us to make the model more robust.

Ml models are often considered as blackbox. The AI Explainability 360 to find similar models and helps comparing the models to the end user. Can also help to train a simpler model to explain the responsibility of diiffernt input variables directed toward the final decision making of the model.

Code asset management tool:

Git

Github

GutLab

BitBucket

Data Asset Management:

Managing huge data is very important task.

Data has to be versioned and annotated with Meta data.

Apache Atlas

Odpi Egeria - Managed through Linux Foundation. open ecosystem with open apis with types and interchange protocols that meta data repos use to store and share data.

Kylo - Open Source

Dev Env

Jupyter Notebook-

Supports hundreds of programming languages using kernels.

Encapsulates the execution enviroment for the different programming languages.

docs, code, shellscript, visualizations

JupiterLab - New version. Has new architecture. allows to open different type of files and arrange them as per our needs.

Apache Zapplin - Inspired from Juputer. Main differnt integrated plotting. In jupyter you need to use external libraries and write code to plot. But in zeppling you dont need to write code to build plots.but you can extend the capabilities by using additional libraries.

RStudio: Runs R and R libraries,

Unifies

Programming

Execution

Debugging

Remote data access

Data exploration

Visualization

in one tool.

Spyder:

Mimics RStudio to bring it to python work.

Apache Spark:

Sometimes the data might not fit in personal computers. So cluster execution is needed. Apache Spark is the most used in the industry for it. It provides linear scalability. doubling number of servers means you double the server performance. Its a batch data processing engine capable of processing large data file by file

Apacke Flink:

Similar to Apache Spark . This performs the processing by stream processing image.

Real- time data stream.

Spark is major choice for most usecases.

riselab Ray:

Enables large scale deep learning model training

Open Source and Fully Integrated tool (UI based and No programming needed):

Knime - has drag and drop, builtin visualization, Can be extended by R or python and has connectors to apache spark.

Orange - less flexible but easier to use.

Commercial Tools:

Data Management -

Oracle Databses

Sql Server

Ibm Db2

Data Integration and transformation:

Informatica Power center

IMB InfoShpere DataStage

SAP

Oracle

SAS

talend

Provide GUI for Data Science pipelines.

IBM Watson Studion Desktop - Includes data refinary.Enable definition and data Execution process in Spreadsheet stye.

Data Visualization:

Tableau

PowerBi

IBM Cognos Analytics

Model Building

SPSS Modeler

sas

IBM Whatson Desktop- includes SPSS Modeler.

Model Deployment:

IBM SPSS Colaboration and Deployment Services

Comercial Software can export the models into open models.

Model Monitoring:

New.

No new tools

Data Asset Management:

Includes Data governance

Data versioning and annotation with meta data

Data dictionary.

Data lineage

Data privacy and retendion

Informatica Data governance

IBM InfoSphere Information Governance Catalog.

Dev Env:

Whatson Studion - fully integrated dev env.

cloud and desktop versions available.

Combine Jupyter notebooks wutg Graphical tools.

Fully Integrated Visual tools

Whatson Studio+ IBM Scale.

Can be deployed on local data center over kubernetes or rehat openshift.

H2O driverless AI.

Cloud Based:

FullY integrated:

Composedof multiple server machines.

Watson Studio

Azure ML

H2O driver less ai-- one-click deployment, No maintainance.

Data Managements:

Saas

Bakup Data and install updates

Aws dynamodb.

cloudant based on open source Couchdb. - can be easily migrated to other couchdb service.

IBM db2 service in Whatson studio

Data Integration and transformation:

Informatica cloud data integration

IBM Data Refinery. on IBM Watson

Data Visualization:

Datameer.

IBM Cognos Analytics.

Model building

IBM Whatson ML

Google Cloud Training.

Model Deployment

SPSS services.

Ibm Whatson ML in REST API

Amazon SageMaker Model Monitor.

Whatson OpenScale.