Data Science IBM Notes

César García

November 6, 2024

1 Categorices Tools

1. Data Management: Collection
2. Data integration and transformation Extract, transform and Load (ETL)
3. Data visualization:
4. Model Building
5. Model deployment
6. Model Monitoring and Assessment

2 Tools

Data Managment

Relational Data bases

1. MySQL
2. ProsgreSQL

No Relational Data bases

1. MongoDB (NoSQL stores data in JSON)
2. Apache Couch MongoDB
3. Apache Cassandra
4. Hadoops HDFS
5. Ceph

Data integration and transformation

1. Apache Ariflow (Arib&b)
2. Kafka
3. Jubeflow
4. Squrk MySQL
5. Note-RED (low in resource consumption)

Data visualization

1. PixieDust
2. Hue

3. Kibana
4. Superset

Model deployment

1. PredictionIO
2. Seldon
3. mleap
4. TensorFlowServing (embeding devices, like RasperryPi)

Model monitoring and Assestment

1. ModelDB (supports apacha sparks and Skilearn)
2. Prometheus
3. IBM IA Fariness 360 (bias in machine learning models)
4. IBM Adeverarial robutsness 360 Toolbox
5. IBM AI explainbaility 360

Code asset Managment
Also known as version control

1. Git
2. GitHub
3. GitLab
4. Bitbucket

Data governance or data lineage

1. ApacheAtlas
2. ODPI-EGERIA
3. Kylo

3 Open-source tools

Development Enviroments

1. JupyterNotebooks
2. JupyeterLab (next version) (modular and ability of open different files)
3. R Studio (statistics and DataScience)
4. Apache Zeppelin
5. Syder

Execution Envarionment

1. Apache Spark (linear scalability)
2. Apache Flink (Stream processing image, real time data streams)
3. Ray, focos in real scale deep learning model training

Fully Integrationd Visual Tools

1. KNIME
2. IBM Developer

4 Comercial tools for data science

1. Oracle Database
2. SQL server
3. IBM DB2

Commersioal supportes delivered by software vendors, influential parthers and support networks
ETL

1. Informatica
2. InfoSphereStage
3. Talend
4. WatsonStudioDesktop

DataVisulaization

1. Tableau
2. PowerBI
3. IBM Congnos Analytics

Model Building

1. SPSS
2. SAS
3. IBM watson studio

Model monitoring
Open source is the best, like Git
WatsonStudioDesktop fully integrated tool, covering all task discosed previously.

5 Cloud Base Tools

For fully integrates visual tools and platforms. Large-scale execution of data science workflows happens in compute clusters:

1. Composed of multiple server machines
2. WatsonStudio, machineLearning and al AI task
3. Azure
4. H2O.ai

Data management

1. AWS
2. Cloudant
3. CouchDB
4. IMB DB2

Data visualization
Examples of data visualization. 3D bar graph, Hierarchical edge bundling, classic bar chart, 2D scatter plot, tree map, pie chart, word Cloud.

6 Open Data Sets and Sources

Government Data

1. https://www.data.gov/
2. https://www.census.gov/data.html
3. https://data.gov.uk/
4. https://www.opendatanetwork.com/
5. https://data.un.org/

Financial Data Sources:

1. https://data.worldbank.org/
2. https://www.globalfinancialdata.com/
3. https://comtrade.un.org/
4. https://www.nber.org/
5. https://fred.stlouisfed.org/

Crime Data:

1. https://www.fbi.gov/services/cjis/ucr
2. https://www.icpsr.umich.edu/icpsrweb/content/NACJD/index.html
3. https://www.drugabuse.gov/related-topics/trends-statistics
4. https://www.unodc.org/unodc/en/data-and-analysis/

Health Data:

1. https://www.who.int/gho/database/en/
2. https://www.fda.gov/Food/default.htm
3. https://seer.cancer.gov/faststats/selections.php?series=cancer
4. https://www.opensciencedatacloud.org/
5. https://pds.nasa.gov/
6. https://earthdata.nasa.gov/
7. https://www.sgim.org/communities/research/dataset-compendium/public-datasets-topic-grid

Academic and Business Data:

1. https://scholar.google.com/
2. https://nces.ed.gov/
3. https://www.glassdoor.com/research/
4. https://www.yelp.com/dataset

Other General Data:

1. https://www.kaggle.com/datasets
2. https://www.reddit.com/r/datasets/