Tools for Data Science

Contents

1 Languages of Data Science **2**
- 1.1 What is Data Science: . . . 2
- 1.2 Roles of Data Scientist: . . . 2
- 1.3 Python: . . . 2

2 Data Science Tools **3**
- 2.1 Categories of Data Science Tools: . . . 3
- 2.2 Data Management: . . . 3
- 2.3 Data Integration and Transformation (Extract, Transform, Load (ELT): . . . 3
- 2.4 Data Visualisation: . . . 4
- 2.5 Model Deployment: (SPSS SAS) . . . 4
- 2.6 Model Monitoring and Assessment: . . . 4
- 2.7 Data Asset Management: . . . 4
- 2.8 Fully Integrated Tool: . . . 4

3 Packages, APIs, Datasets and Models **4**
- 3.1 Python Libraries . . . 4
  - 3.1.1 Scientific Computing Libraries in Python: . . . 4
  - 3.1.2 Visualisation Libraries: . . . 5
  - 3.1.3 High Level Machine Learning and Deep Learning Libraries: 5
- 3.2 Machine Learning: . . . 5
  - 3.2.1 Supervised: . . . 5
  - 3.2.2 Un-Supervised: . . . 5
  - 3.2.3 Reinforcement: . . . 5

4 GitHub **6**

1 Languages of Data Science

1.1 What is Data Science:

- **Recommended:** Python, R, Sql
- **Can be used:** : Scala, Java, C++, Julia, JS, PHP, Go, Ruby, VB

1.2 Roles of Data Scientist:

- Business Analyst
- Database Engineer.
- Research Scientist.
- Product Manager.
- Statistician

1.3 Python:

- 80% data professionals use it worldwide.
- Heavily used in data science, AI, ML, web dev, IOT, etc.
- General Purpose Language.
- Large Standard Library.
- **Libraries to use for data science:** Pandas, Numpy, Scipy, Matplotlib.
- **Tools:** Pytorch, Tensorflow, Keras, Scikit-Learn.

2 Data Science Tools

2.1 Categories of Data Science Tools:

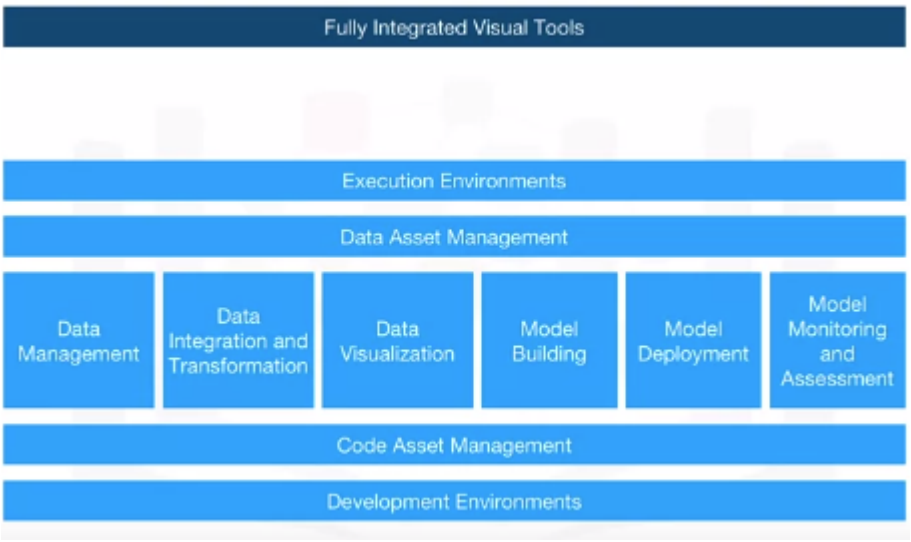

Figure 1: Categories of Data Science Tools

2.2 Data Management:

- **Tools:** Oracle, Mysql, IBM DBZ.
- **RDBMS:** Mysql, Postgre Sql.
- **No Sql:** MongoDB, Apache CouchDB and Cassandra.
- **File Systems:** Hardoop HDFS (EPH).
- **Store and Retrieve data for docs:** Elastic Search.

2.3 Data Integration and Transformation (Extract, Transform, Load (ELT):

- **Data Integration:** Apache Airflow.
- **Enable to execute Data Science Pipeline:** Kubeflow.
- Apache Kafla, Apache Nifi, Spark Sql, Node-Red.
- Talented Infomatica, IBM Watson, IBM infosphere.

2.4 Data Visualisation:

- **Creates Data Visualisation from Sql:** HUE.
- **Creates Data Visualisation from Elastic Search:** Kibana
- **Data Visualisation:** Apache Superset.

2.5 Model Deployment: (SPSS SAS)

- PredictionIo
- Seldon
- mleap
- TensorFLow service

2.6 Model Monitoring and Assessment:

- Model DB
- Prometheus
- AI Explainability 360

2.7 Data Asset Management:

- Apache
- Atlas
- Kylo
- Egeria

2.8 Fully Integrated Tool:

- Knime

3 Packages, APIs, Datasets and Models

3.1 Python Libraries

3.1.1 Scientific Computing Libraries in Python:

- Numpy
- Pandas

3.1.2 Visualisation Libraries:

- Matplotlib
- Seaborn

3.1.3 High Level Machine Learning and Deep Learning Libraries:

- Scikit Learn
- Keras
- Pytorch
- Tensor Flow

3.2 Machine Learning:

Machine Learning Models identifies patterns in data. Model requires training before predictions.

3.2.1 Supervised:

1. Data Label - Correct Output
2. Creates Relations
3. Regression
4. Classification

3.2.2 Un-Supervised:

1. Data Label
2. Tries to identify relations
3. Clustering
4. Anomaly Detection

3.2.3 Reinforcement:

1. Environment
2. Tries learn on itself
3. Games

4 GitHub

1. Free and Opensource software.
2. Distributed version control system.
3. Most common version control system.
4. Can also version control images, docs, etc.
5. **UI for Git:** GitHub (Widely Used), Gitlab, Bit Bucket.
6. **SSH Protocol:** A method for secure remote login from one computer to another.
7. **Repository:** The folders of your project that are set up for version control.
8. **Pull Request:** The process you use to request that someone reviews and approves your changes before they become final.
9. **Working Directory:** A directory on your file system, including it's file and sub directories that is associated with a git repo.
10. **Important Commands:** git init, git add, git status, git commit, git reset, git log, git branch, git checkout, git merge.