Duration: 20 mins

Objective

This reading provides a summary of key open-source tools for Data Science covered in the Part 1 and Part 2 videos of this course.

They are broadly classified as -

- **Data Management Tools** - Facilitates the storage, organization, and retrieval of data. Includes Relational Databases, NoSQL Databases, and Big Data platforms.
- **Data Integration and Transformation Tools** - Streamlines data pipelines and automate data processing workflows. Task of data integration and transformation in the classic data warehousing world is to Extract, Transform, and Load (ETL).
- **Data Visualization Tools**- Provides graphical representation of data and assist with communicating insights.
- **Model Deployment, Monitoring and Assessment Tools**- Supports the building, deploying, monitoring, and evaluation of data and machine learning models.
- **Data Asset Management Tools**- Organizes and manages data, enforce access controls, and ensure asset backups.
- **Code Development and Execution Tools** - ProvideS environments for developing, testing, and deploying code, offering computational resources to execute it.
- **Code Asset Management Tools** - Enables the storage and management of code, track changes, and support collaborative development.

Data Management Tools

MySQL

- Popular open source **relational database management system** (RDBMS)
- Uses structured query language (SQL) to manage and store data.
- Common uses:
  - Web applications
  - Data warehousing
  - E-commerce

PostgreSQL

- Powerful and open source **relational database management system** (RDBMS)
- Emphasizes extensibility and SQL compliance.
- Offers advanced features such as:
  - Support for JSON
  - Full-text search
  - Spatial data

Apache CouchDB

- Document-oriented **NoSQL** database
- Uses JSON to store data
- Highly scalable
- Fault-tolerant
- Easy to use

MongoDB

- Document-oriented **NoSQL** database
- Stores data in a flexible JSON
- Provides:
  - Scalability
  - High availability
  - Data distribution
- Suitable for modern web applications that handle large volumes of unstructured data

Apache Cassandra

- Highly scalable, distributed Document-oriented **NoSQL** database
- Can handle large amounts of structured and unstructured data across many commodity servers.
- Offers:
  - High availability
  - Fault tolerance
  - Tunable consistency levels
- Suitable for mission-critical applications

Hadoop Distributed File System (HDFS)

- Designed to work with large datasets like Apache Hadoop in a distributed computing environment
- High-throughput data processing by splitting files into blocks (default 128MB), and these blocks are distributed across multiple DataNodes
- Data is replicated across different DataNodes ensuring high availability and fault tolerance
- Scalable and efficient

Ceph

- Free, open source software-defined storage platform suitable for hybrid cloud environments
- Designed for modern data centers
- Provides highly scalable, unified storage system that can be used for object storage (like AWS S3), block storage (like virtual disks for VMs), and file storage (like NFS) under one unified system
- High performance, availability and reliability

Elasticsearch

- Primarily a distributed RESTful search engine and analytics tool
- Based on the Lucene library.
- Full-text search, real-time data analytics
- Highly scalable
- Easy to use
- Powerful querying capabilities
- Real-time data indexing for fast document retrieval.

Data Integration and Transformation Tools

Apache Airflow

- Open-source platform for programmatically authoring, scheduling, and monitoring workflows
- Created originally by Airbnb
- Allows users to define and execute complex workflows
- Support for:
  - Task dependencies
  - Parallelism
  - Error handling

Kubeflow

- An open-source machine learning toolkit that allows execution of data science pipelines on top of Kubernetes.
- Provides a platform for building, deploying, and managing end-to-end machine learning workflows at scale
- Support for:
  - Distributed training
  - Model serving
  - Hyperparameter tuning

Apache Kafka

- Distributed streaming platform that allows applications to publish, process, and subscribe to streams of records in real-time
- Created originally from LinkedIn.
- It is scalable, fault-tolerant, and high-throughput
- Suitable for building mission-critical, data-intensive applications

Apache NiFi

- An open-source data integration platform that allows users to automate the flow of data between systems
- Provides a web-based user interface for designing and managing data flows
- Support for:
  - Data routing
  - Transformation
  - Enrichment
  - Among other capabilities

Apache Spark SQL

- A module in the Spark ecosystem that provides a programming interface for working with structured data using:
  - SQL
  - Data frames
  - Datasets
- Supports a wide range of data sources and provides optimized performance for complex data processing tasks.

Node-RED

- An open-source visual programming tool for wiring together hardware devices, APIs, and online services
- Allows users to create event-driven flows of messages
- low in resource consumption that it even runs on tiny devices like a Raspberry Pi.
- Support for:
  - Data transformation
  - Filtering
  - Aggregation

Data Visualization Tools

PixieDust

- Open-source library for creating interactive, exploratory data visualizations in Python and Jupyter notebooks
- Provides a range of built-in visualizations and data connectors
- Support for customization and extensibility through third-party libraries

Hue

- Open-source web interface for analyzing and visualizing large datasets in Apache Hadoop
- Offers a user-friendly experience for exploring data and creating visualizations
- No need for programming skills; can create visualizations from SQL queries

Kibana

- Open-source data visualization tool that allows users to interact with their data through a web-based interface
- Commonly used with Elasticsearch to analyze and visualize large datasets

Apache Superset

- A modern, enterprise-ready business intelligence web application that makes it easy to visualize and explore large datasets
- Offers a rich set of data visualization options, including:
  - Charts
  - Tables
  - Maps
  - Geospatial analysis
  - Real-time data processing

Model Deployment Tools

Apache PredictionIO

- Open-source machine learning server built on a scalable and distributed infrastructure
- Allows developers to quickly build, evaluate, and deploy predictive engines for various use cases such as:
  - Recommendation
  - Classification
  - Clustering

Kubernetes

- Open-source platform for container orchestration
- Automatically launches, scales, and manages containerized applications
- Offering features like:
  - Automatic scaling
  - Self-healing
  - Load balancing
- Enables the management and orchestration of containers across numerous hosts

Apache Seldon

- Open-source platform for deploying and managing machine learning models on Kubernetes
- Provides a way to:
  - Serve models at scale
  - Automate model deployment workflows
  - Monitor the performance of deployed models in real-time

MLeap

- Open-source library for serializing and deserializing learning models in a cross-platform file
- Gives users the ability to export models from different machine learning libraries and frameworks, such as:
  - Spark
  - Scikit-learn
  - TensorFlow
- Implements them in high-throughput, low-latency production environments

TensorFlow Lite

- Open-source tool for running machine learning models on mobile and embedded devices
- Allows effective inference on mobile and embedded platforms
- Supports a variety of hardware accelerators such as:
  - CPUs
  - GPUs
  - Custom ASICs

Red Hat OpenShift

- Container application framework based on Kubernetes
- With characteristics like automation, scalability, and security
- Offers a method for creating, deploying, and managing containerized applications

TensorFlow Serving

- Open-source utility that serves machine learning models in real-world settings
- Supports both HTTP and gRPC interfaces for serving predictions
- Provides high scalability and low latency deployment and management of TensorFlow models

TensorFlow.js

- Open-source library for building and deploying machine learning models in JavaScript
- Allows you to train and execute models directly in the browser or on Node.js
- Supports a wide range of model architectures, including neural networks, decision trees, and k-nearest neighbors

Model Monitoring and Assessment Tools

ModelDB

- Open-source platform for managing machine learning models and experiments
- Provides a way to track and reproduce experiments, version models, and collaborate with team members

Prometheus

- Freely available monitoring system that collects and stores metrics in real-time from different sources
- Allows you to visualize and set alerts on the health and performance of systems and apps
- Supports a variety of data gathering methods, such as HTTP endpoints, exporters, and agents

IBM AI Fairness 360

- Open-source toolkit for detecting and mitigating bias in machine learning models
- Provides a way to measure the fairness and bias of models, as well as a set of algorithms for mitigating bias and creating fairer models

IBM AI Explainability 360

- Open-source toolkit for explaining the behavior and decisions of machine learning models
- Provides a way to measure the explainability and interpretability of models, as well as a set of algorithms for generating explanations and visualizations of model behavior

IBM Adversarial Robustness 360 Toolbox

- Free and open-source library for protecting machine learning models from adversarial attacks
- Includes a method for measuring model robustness and vulnerability
- Includes a set of algorithms for improving model robustness and detecting adversarial examples

Code Development and Execution Tools

Jupyter IDE

- Open-source effort
- Supports:
  - Julia
  - Python
  - R development with Jupyter Notebook
  - JupyterLab
  - JupyterHub
- Create and share documents containing:
  - Live code
  - Equations
  - Visualizations
  - Narrative text
- JupyterLab includes:
  - Customized notebook organization
- JupyterHub extends all these capabilities to the enterprise

RStudio

- For developers
- Free and open-source IDE
- Built to manage and execute R code
- Works on all platforms
- Includes:
  - Version control
  - Project management capabilities

Microsoft Visual Studio

- An IDE that supports a variety of programming languages, including:
  - C
  - C++
  - C++/CLI
  - Visual Basic.NET

    - C#
    - F#
    - JavaScript
    - TypeScript
    - XML
    - XSLT
    - HTML
    - CSS
- Using plug-ins, supports:
    - Python
    - Ruby
    - Node.js
    - M
    - Other languages

PyCharm

- Primarily a subscription-based IDE environment
- Offers 16+ additional tools for coding assistance, testing, and web development
- Supports scientific development with IPython integration and Matplotlib and NumPy support
- Also offers a free community-based, open-source IDE with limited capabilities

Spyder

- Free, open-source Python-based IDE designed by and for scientists, engineers, and data analysts
- Features a unique combination of comprehensive development tools for:
    - Advanced editing
    - Analysis
    - Debugging
    - Profiling
    - Visualization capabilities

Anaconda Navigator

- Open-source GUI-based Navigator that supports Python development and integrates with:
    - Eclipse and PyDev
    - IDLE
    - IntelliJ
    - Microsoft Visual Studio Code (VS Code)
    - Ninja IDE
    - PyCharm
    - Python for Visual Studio Code
    - Python Tools for Visual Studio (PTVS)
    - Spyder
    - Sublime Text
    - Wing IDE

Code Asset Management Tools

Git

- Open-source version control system for tracking changes in code and collaboration among developers
- Provides a way to manage and organize code changes, collaborate on code development, and maintain a history of code revisions

GitLab

- Web-based Git repository manager
- Provides a complete DevOps platform for:
    - Source code management
    - Continuous integration and deployment
    - Monitoring
- Enables teams to collaborate on:
    - Code development
    - Automate build and deployment processes
    - Track metrics and performance across the entire software development lifecycle

GitHub

- Web-based Git repository hosting service that provides a platform for developers to collaborate on code and manage software projects
- Enables users to:
    - Create, fork, and contribute to open source projects
    - Track changes in code
    - Manage issues
    - Pull requests

Bitbucket from Atlassian

- Web-based Git repository hosting service

- Provides a platform for developers to collaborate on code and manage software projects, with features like:
  - Pull requests
  - Code review
  - Branch permissions

Author(s)

Shilpa Giridhar