# Development Environment Setup

1. Running Spark on local computer
2. Standalone spark – one master one slave local setup
3. Spark Cluster

# Programming Bindings

1. Python
2. Java

# Abilities

|  |  |  |  |
| --- | --- | --- | --- |
| Getting data to spark | File Source | csv  json\*  parquet\*  ORC  Text??  Hive Table  S3 buckets – AWS, Azure, GCP, MinIO\* | library compatibility  optimum parsers |
| Reading from OLTP | MySQL  MariaDB  PostgreSQL\*  MS-SQL  Oracle  No-SQL???? – couchbase\* | connection management/pooling  library compatibility |
| Reading data from In-Memory Stores | Redis\*  memcache |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Data processing | Basic RDD and data-frame processing |  |  |
| Spark-APIs | Sort  Join  Filter  Group-By  Order-By |  |
| Spark-SQL | Sort  Join  Filter  Group-By  Order-By |  |
| Spark-ML | Linear Regression\*  Logistic Regression  Decision Tree  Random Forest  K-mean  Collaborative Filtering  NLP\* | Kaggle-datasets |

|  |  |  |  |
| --- | --- | --- | --- |
| Output of Spark | File-output | Compression technique  Bzip2, gzip\*, lz4, snappy\*, deflate  File-Formats/Destination\*  S3, Parquet, CSV |  |
| Database-Output | OLTP  NoSQL  In-Memory |  |
|  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Solution Configuration | Properties | JSON\*(  YAML  Files *(local, cluster, hdfs??)* |  |
| Environment |  |  |
| Secrets | Keys, Passwords |  |
| Runtime configuration issuances |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Streaming | Spark-Streaming |  |  |
| Structured Streaming? |  |  |
|  |  |  |

# Advanced Topic for developers

Memory management

Data caching in spark

Dynamic Resource Allocation

# Performance Tunning

# References

* Basic course
  + <https://www.udemy.com/course/spark-and-python-for-big-data-with-pyspark/v> (ML)
  + <https://www.udemy.com/course/taming-big-data-with-apache-spark-hands-on/> (RDD)
* Advanced course
  + <https://www.udemy.com/course/spark-sql-and-pyspark3/> (Passing Parameters)
  + <https://www.udemy.com/course/best-hands-on-big-data-practices-and-use-cases-using-pyspark/> (Tuning and log)
  + <https://intel.udemy.com/course/apache-spark-3-beyond-basics/> (Advanced)
* YouTube vid
  + Setting up standalone single node on local computer - <https://www.youtube.com/watch?v=CvOkr-fKuKc>
* Medium Articles
  + Setting up a cluster - <https://medium.com/codex/setup-a-spark-cluster-step-by-step-in-10-minutes-922c06f8e2b1>
* Documentation
  + Understanding standalone arch - <https://spark.apache.org/docs/latest/spark-standalone.html>
  + Spark Submit - <https://spark.apache.org/docs/latest/submitting-applications.html>

# Task Level 1

* Scheduling spark jobs – airflow and shell
* Writing a listener to monitor spark execution