Module II:
Toolset Introduction

On ApacheSpark, NoSQL, ObjectStorage and the rest…

In this Video you will learn…

Programming Language options on ApacheSpark

- ApacheSpark itself is implemented in Scala
- ApacheSpark runs on top if the JVM
- ApacheSpark jobs can be implemented in
  - Java
  - Scala
  - Python
  - R

Scala

- Scala is native to ApacheSpark
- Complete API set is available
- Scala code normally runs fastest

Scala

Video

a0_m2_p4_scala

Java

- Java not a Data Science programming language
- Complete API set is available
- De-factor standard in Enterprise IT
- Language of “Hadoop”

R

- R is THE Data Science programming language
- Only subset of ApacheSpark API is available
- De-factor standard in academic research
- > 8000 add-on packages available
- Awesome plotting and charting capabilities
- Slow

Python

- as widely used in Data Science as R
- Only subset of ApacheSpark API is available
- nice plotting and charting but R is better here
- can get slow

Summary

	Scala	Java	R	Python
Spark API support	complete	complete	very limited	limited
ease of use	low	very low	high	very high
Speed	very high	high	very low	low
3rd party libraries	few	few	many	many

Quiz

- There exists a very famous data science library for python called PANDAS, what is the main disadvantage of that library?
  - PANDAS is running on top of Python and python is very slow
    False: Although python is slower than Scala it is still quite usable as long as your data set fits into a single node (machine)
  - PANDAS is not capable of using the ApacheSpark API
    Correct, If you use the RDD API your code implicitly is getting executed in parallel on the ApacheSpark cluster which is not the case with PANDAS
  - PANDAS is very outdated and doesn’t support the state-of-the-art data analysis methods
    False: PANDAS is the state of the art framework for data scientists using python
  - PANDAS does not support the analysis of IoT sensor data because it is meant for more general data science tasks
    False: PANDAS supports all sorts of data science tasks including sensor data analysis but only on one single not. There is no parallel processing taking place
  - PANDAS is very expensive
    False: PANDAS is open source and free of charge

The next Video covers…

Functional Programming basics