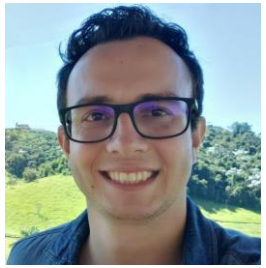
Spark – Big Data Processing

Aula 6

Semantix®

All about data

Quem sou eu?



Rodrigo Augusto Rebouças

Engenheiro de dados da Semantix
Instrutor do Semantix Academy

Contatos
rodrigo.augusto@semantix.com.br
linkedin.com/in/rodrigo-reboucas

RODRIGO.A.REBOUCAS

Semantix®
All about data

Spark application

Semantix®
All about data

Spark shell x Spark applications

- Spark shell
  - Exploração e manipulação dos dados
- Spark applications
  - Rodar os programas de forma independente
  - Jobs para ETL e streaming
  - Criação de objetos
    - SparkSession – spark (spark SQL)
    - SparkContext – sc

Rodar uma Spark application

- spark-submit --class NameList MyJarFile.jar people.json namelist/

- Opções submit
  - master: local, yarn, mesos ou spark standalone
  - jars: adicionar arquivos jar
  - py-files: lista de arquivos em .py, .zip ou .egg
  - driver-java-options: parâmetros para o driver JVM
  - deploy-mode: client ou cluster
  - driver-memory: Memória alocada para o spark driver (1G)
  - executor-memory: Memória alocada para a aplicação
  - num-executors: Número de executores para iniciar com a aplicação
  - driver-cores: Número de cores alocados para o spark driver
  - queue: Rodar na fila do Yarn
  - help

Instalação IDE

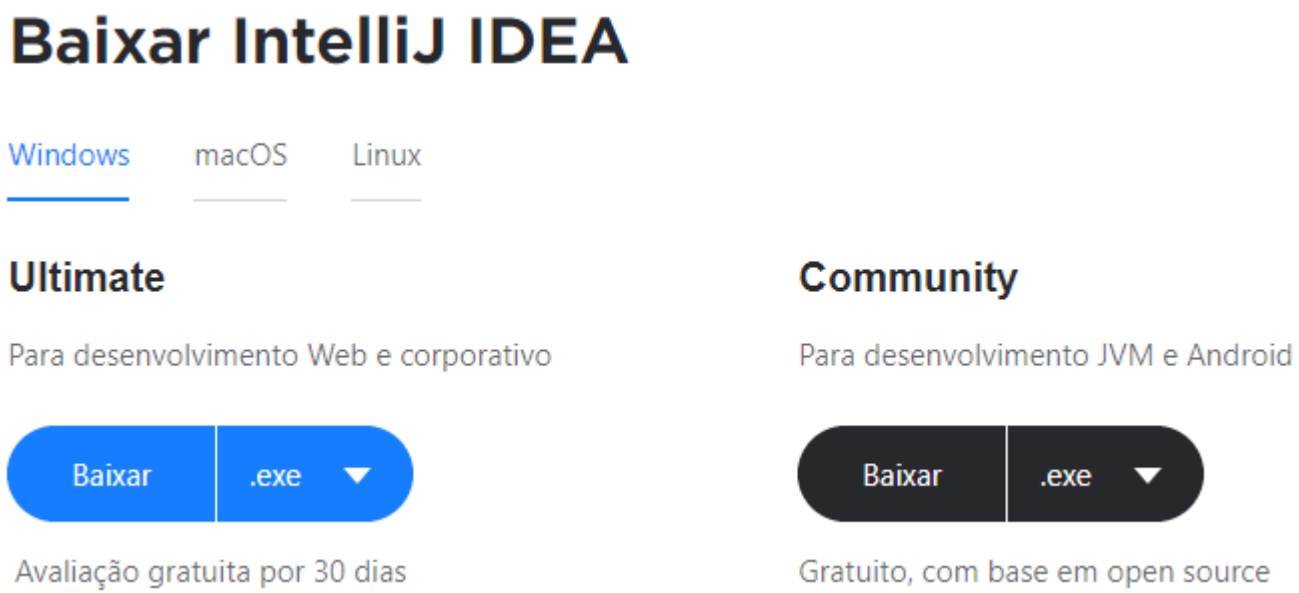
Semantix®
All about data

Intalação

- IntelliJ
  - Plugins:
    - Scala – 2.11
    - SBT (Preferível instalação separada)
  - Java 8
    - https://www.oracle.com/br/java/technologies/javase/javase8u211-later-archive-downloads.html
- PyCharm
  - Python 3

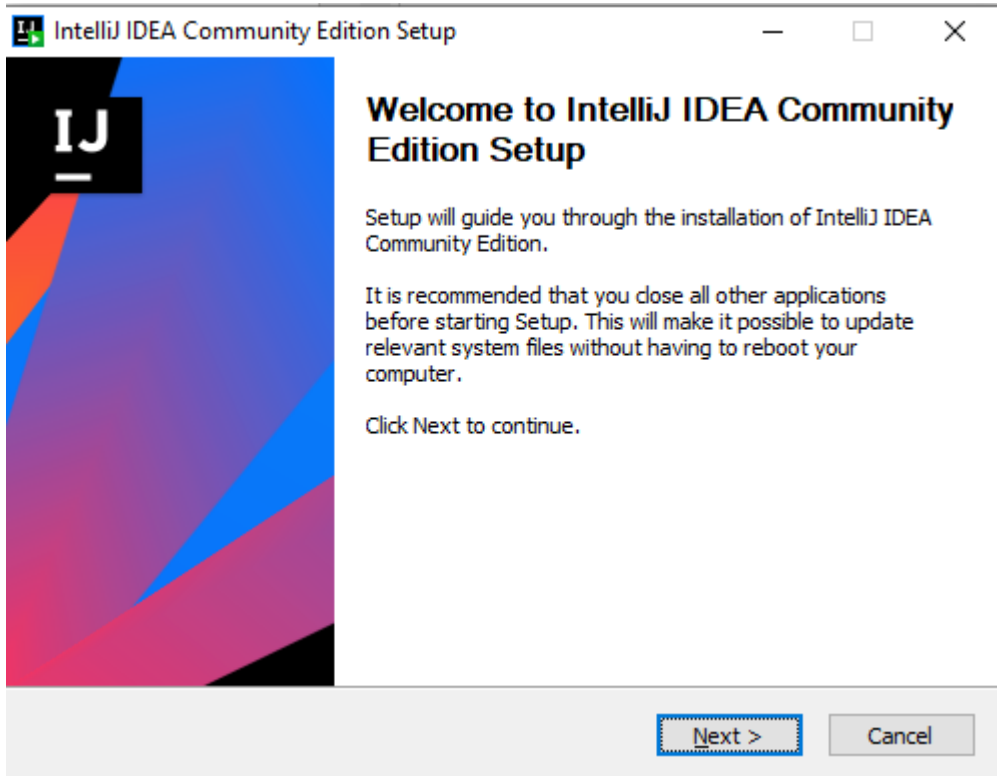
IntelliJ

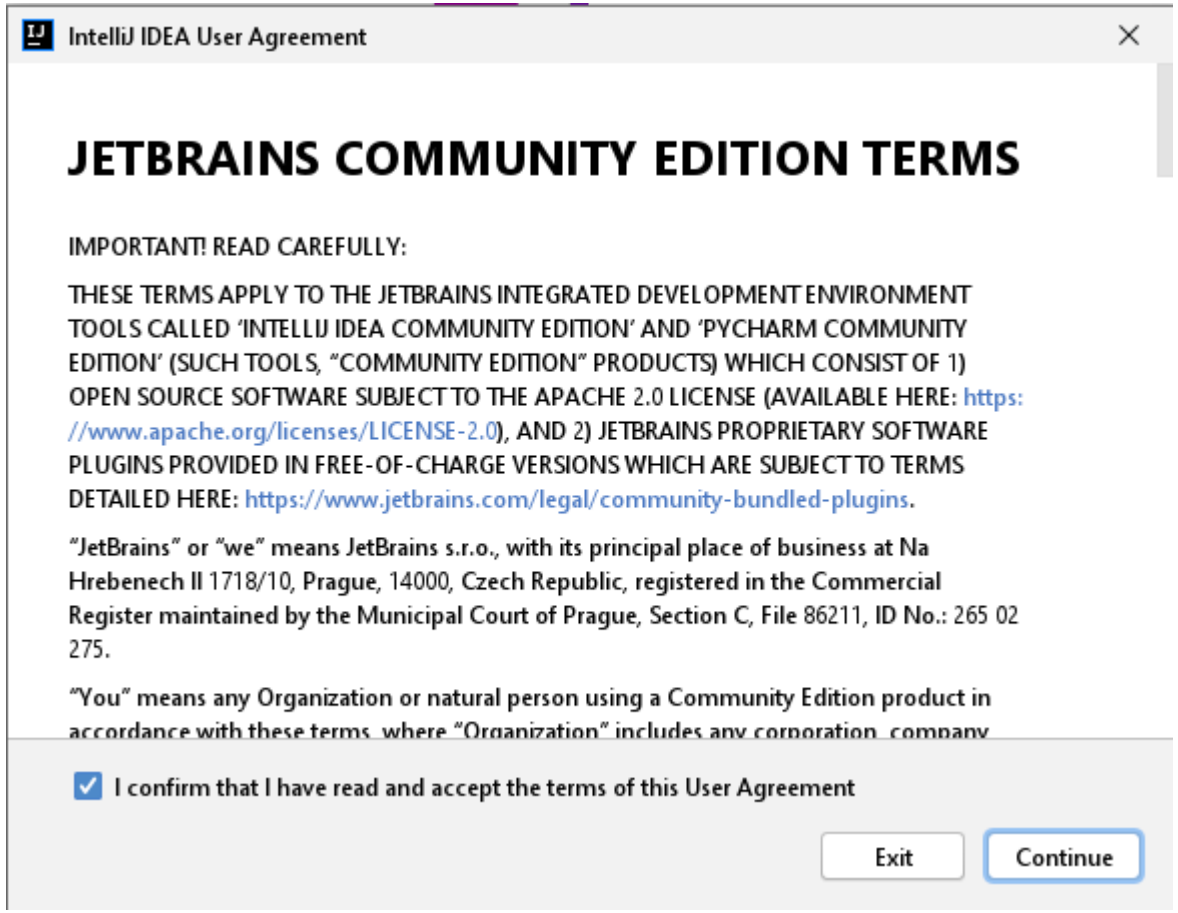
- https://www.jetbrains.com/pt-br/idea/



IntelliJ

- Espaço em disco
  - 1,6 GB

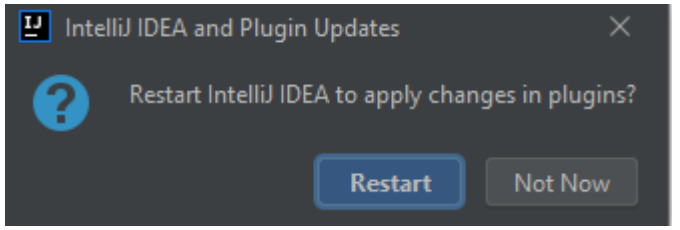


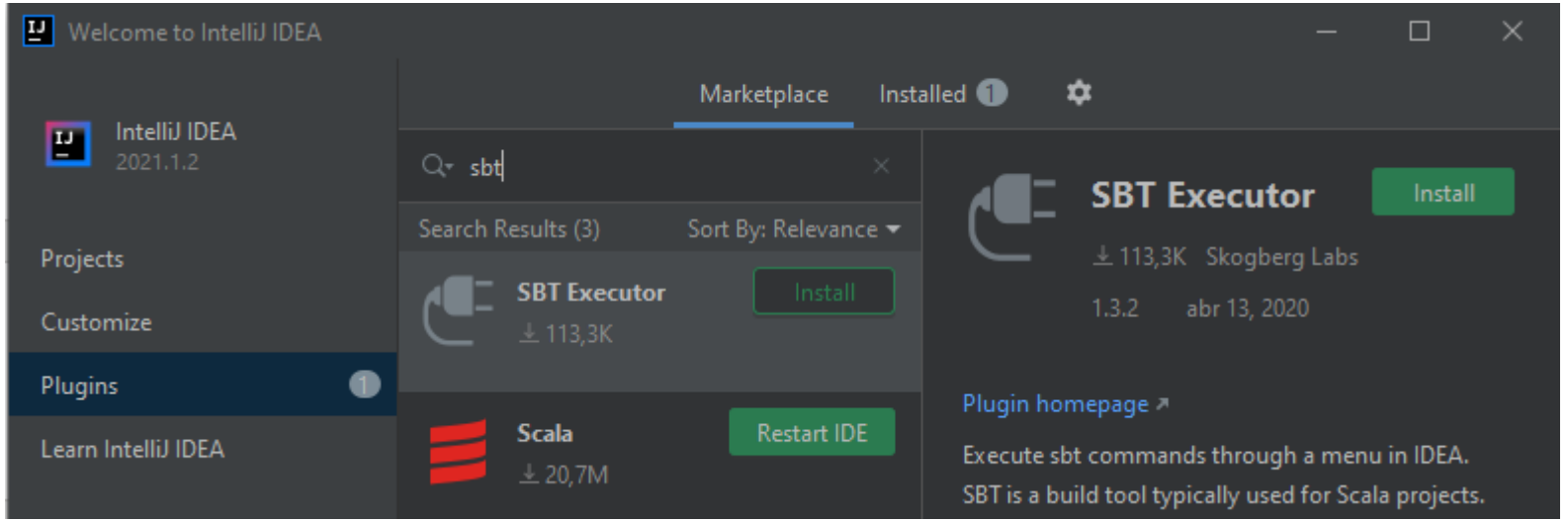


Semantix®
All about data

IntelliJ

- Plugins
  - Scala
  - SBT Executor





Semantix®
All about data

Python

- Windows
  - python --version
- Linux
  - python3 –version

- Instalação
  - https://www.python.org/

- Anaconda (Opcional)

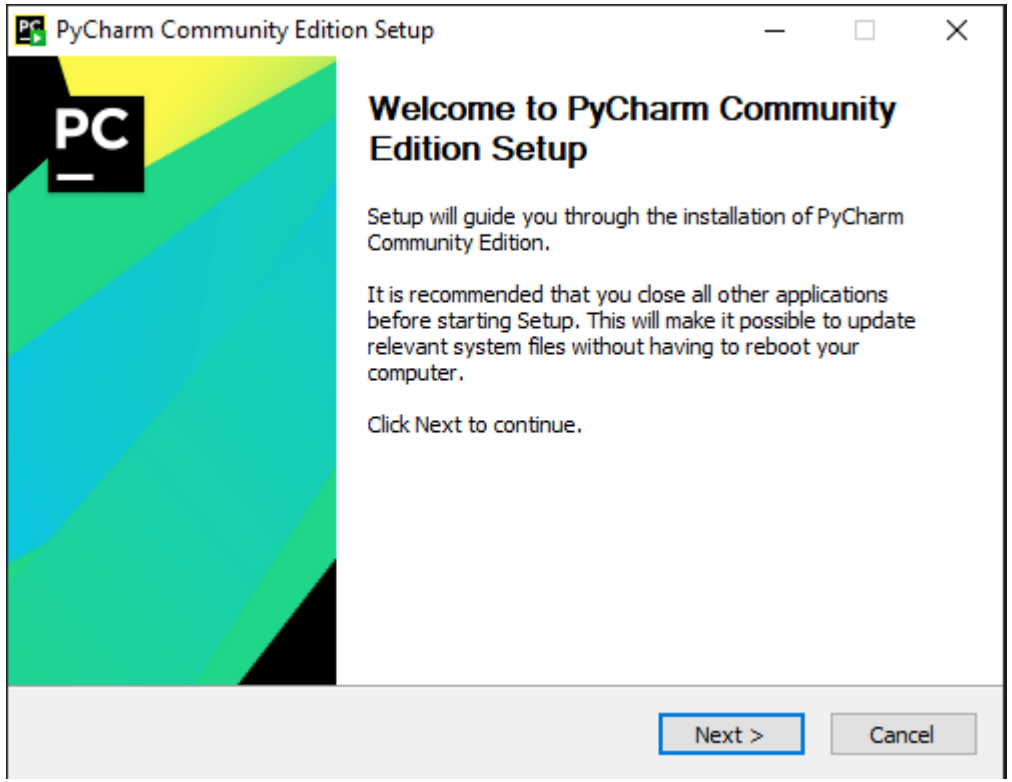
Semantix®
All about data

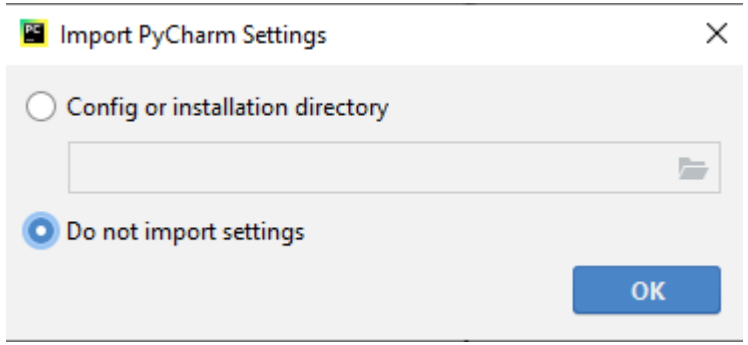
PyCharm

- https://www.jetbrains.com/pt-br/pycharm/

PyCharm

- Espaço em disco
  - 1 GB





Semantix®
All about data

Build Spark

Semantix®
All about data

Build Spark

- IntelliJ
  - sbt package
    - https://www.scala-sbt.org/sbt-native-packager/gettingstarted.html#setup
- Pycharm
  - pip install
    - https://packaging.python.org/tutorials/installing-packages/
- https://spark.apache.org/docs/latest/building-spark.html

Semantix®

All about data

contato@semantix.com.br

www.semantix.com.br