Big Data architektúrák és elemző módszerek Gyakorlat

Gombos Gergő

Elérhetőségek

- Gyak.vez: Dr. Gombos Gergő
- honlap: http://ggombos.web.elte.hu
- email: ggombos@inf.elte.hu
- szoba: D. 2-503

Tematika

- Architektúra ismeretek:
  - Hadoop / MapReduce
  - HDFS
  - Spark
- Elemző módszerek
  - Python (pandas, sklearn, numpy)
  - Adatvizualizáció (matplotlib)
  - SparkML vs. SKLearn

Követelmények

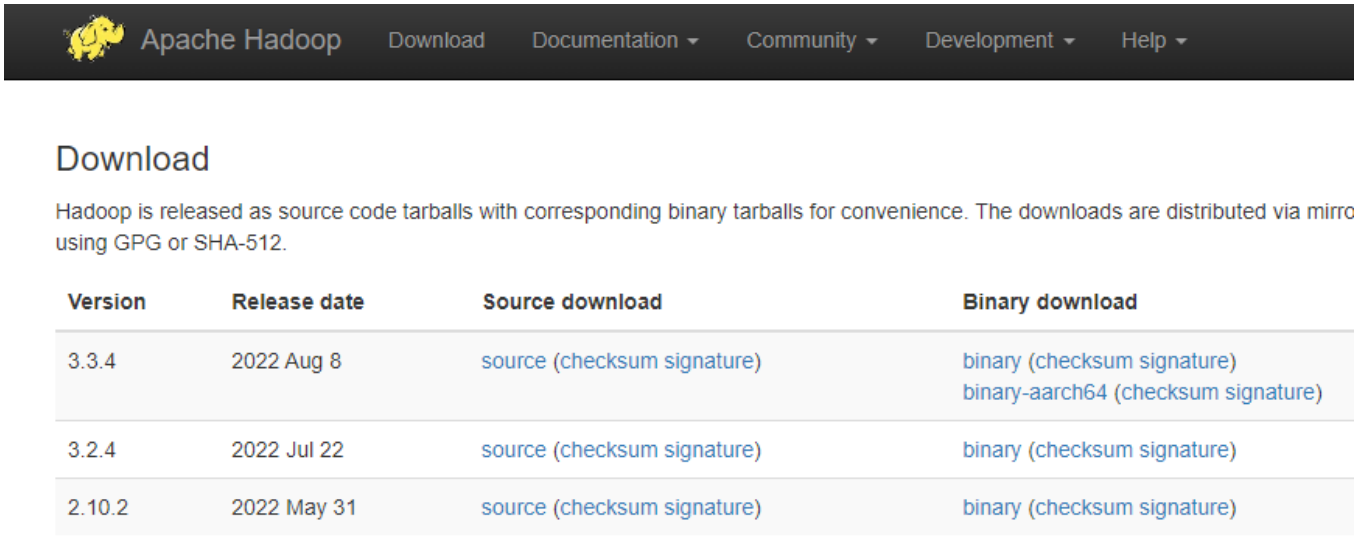
- Hadoop / MapReduce beadandó
- Spark ZH
  - Spark Batch adatelemzés, adatfeldolgozás
- Python adatelemzés ZH
  - Pandas, SKLearn, Adatvizualizáció

Hadoop

- Nyílt forráskódú keretrendszer
- Lehetőséget ad nagy adathalmazok feldolgozására számítógép klaszterek használatával
- Fő részei
  - Hadoop Distributed File System (HDFS) elosztott fájlrendszer
  - MapReduce programozási modell
  - YARN erőforrás menedzser

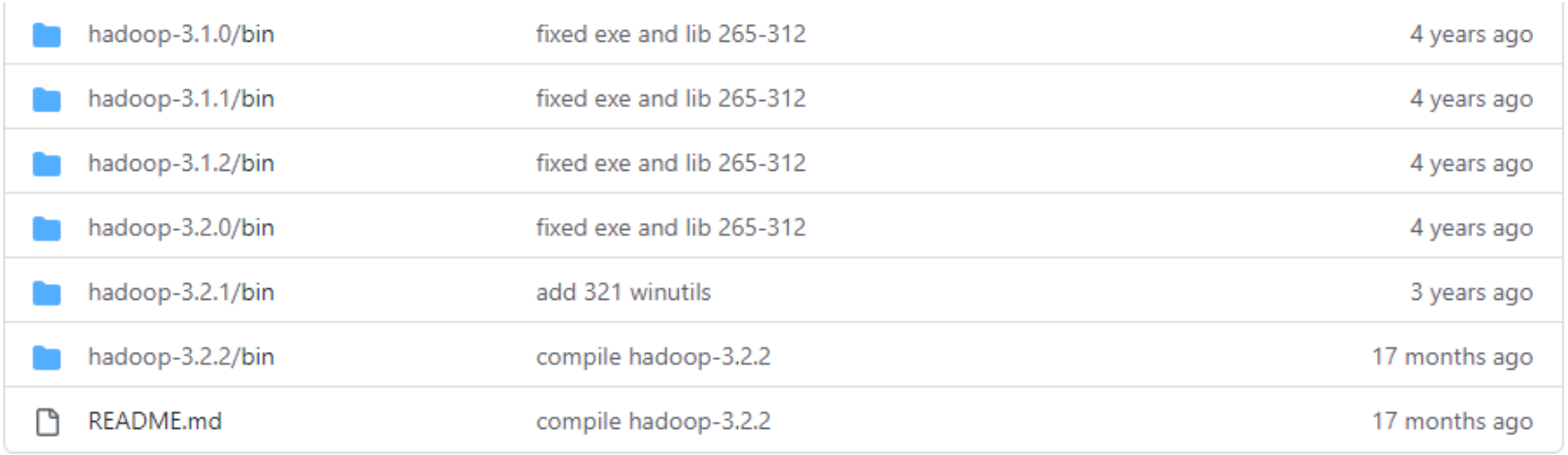
Hadoop használata (1)

- Java telepítése (remélhetőleg már van): https://www.oracle.com/java/technologies/downloads/#java8-windows
- Hadoop bináris letöltése és kicsomagolása: https://hadoop.apache.org/releases.html



Hadoop használata (2)

- Winutils fájlok letöltése és bemásolása a hadoop-3.2.4\bin mappába: https://github.com/cdarlint/winutils
  - wintils.exe
  - hadoop.dll



hadoop-3.1.0/bin	fixed exe and lib 265-312	4 years ago
hadoop-3.1.1/bin	fixed exe and lib 265-312	4 years ago
hadoop-3.1.2/bin	fixed exe and lib 265-312	4 years ago
hadoop-3.2.0/bin	fixed exe and lib 265-312	4 years ago
hadoop-3.2.1/bin	add 321 winutils	3 years ago
hadoop-3.2.2/bin	compile hadoop-3.2.2	17 months ago
README.md	compile hadoop-3.2.2	17 months ago

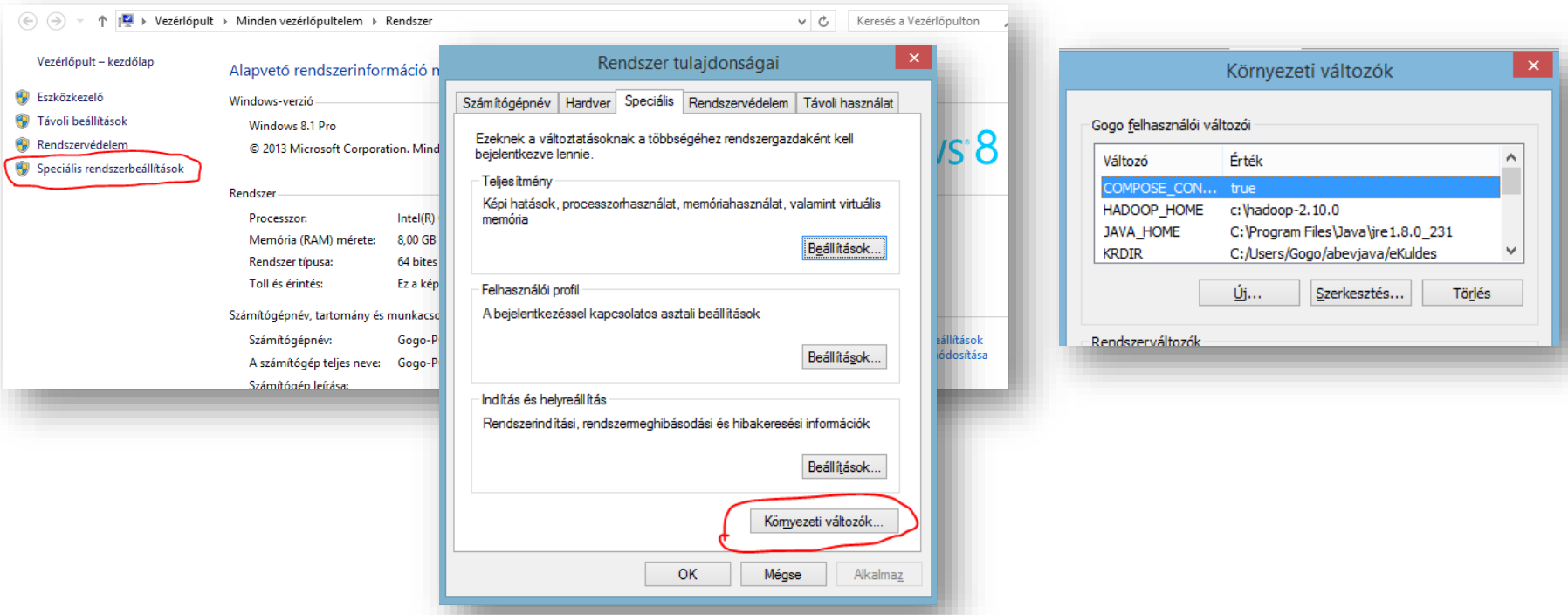
Hadoop használata (3)

- Környezeti változók beállítása
  - HADOOP_HOME – a kicsomagolt hadoop mappa
  - Hozzáadás a Path-hez: %HADOOP_HOME%\bin

Hadoop telepítése - Windows

- Környezeti változók beállítása

HADOOP_HOME	**-->**	**C:\hadoop-2.10.0**
PATH	**-->**	**%HADOOP_HOME%\bin, %HADOOP_HOME%\sbin**



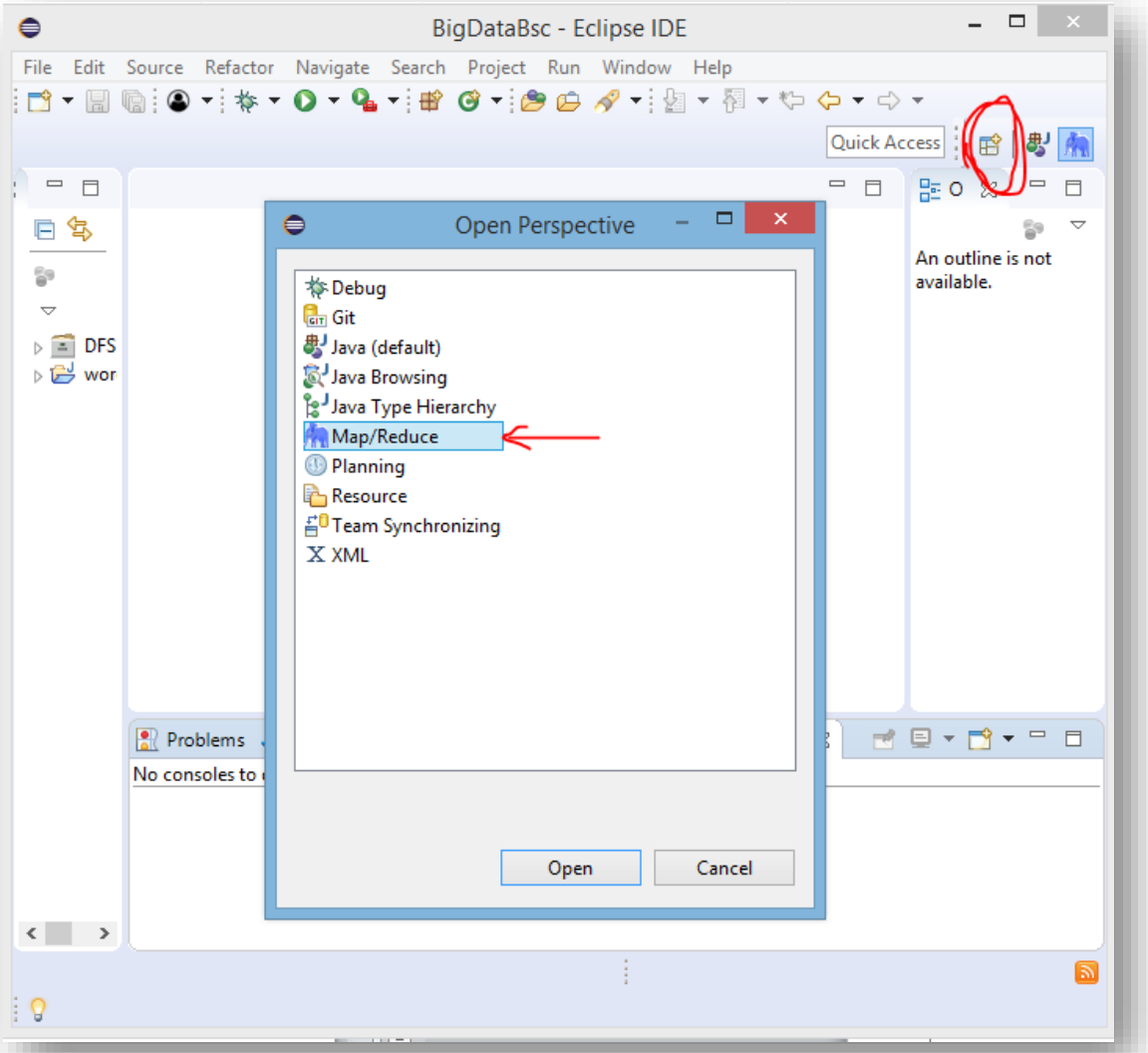
Fejlesztő környezet kialakítása

1. Szükséges: Eclipse letöltése
   - Eddig minden verzióval működött
   - https://www.eclipse.org/downloads/

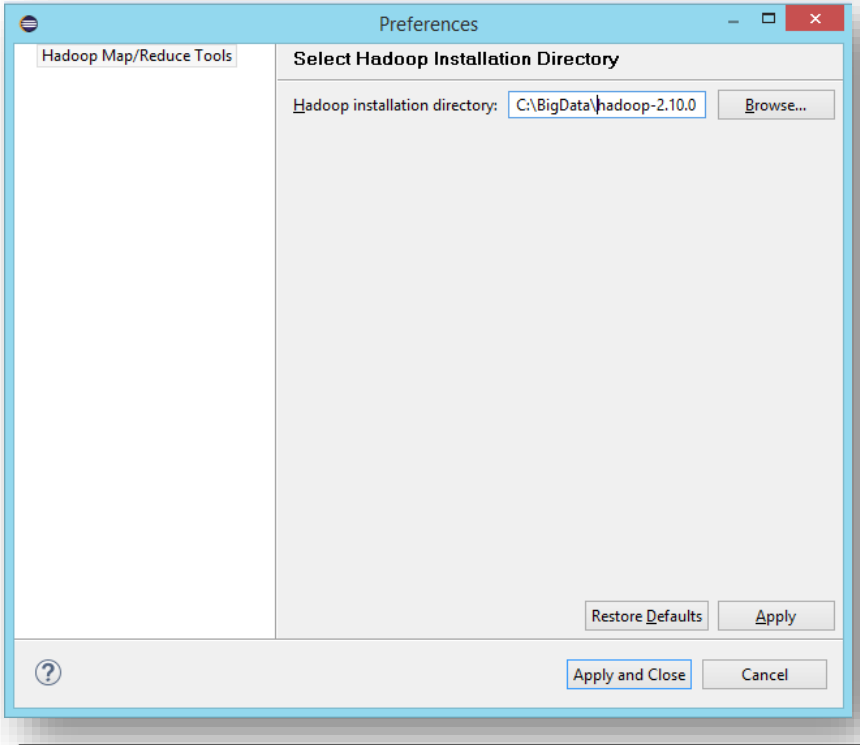
2. hadoop-eclipse-plugin-2.6.0.jar
   - bemásolni a eclipse/dropins mappába, régebbi eclipseknél a eclipse/plugins mappába
   - http://ggombos.web.elte.hu/oktatas/BigDataArchitekturaEsElemzo/GY/gyak1/

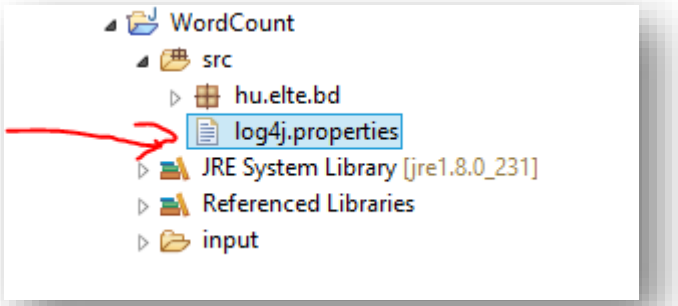
3. Eclipse elindítása

Nézet beállítása



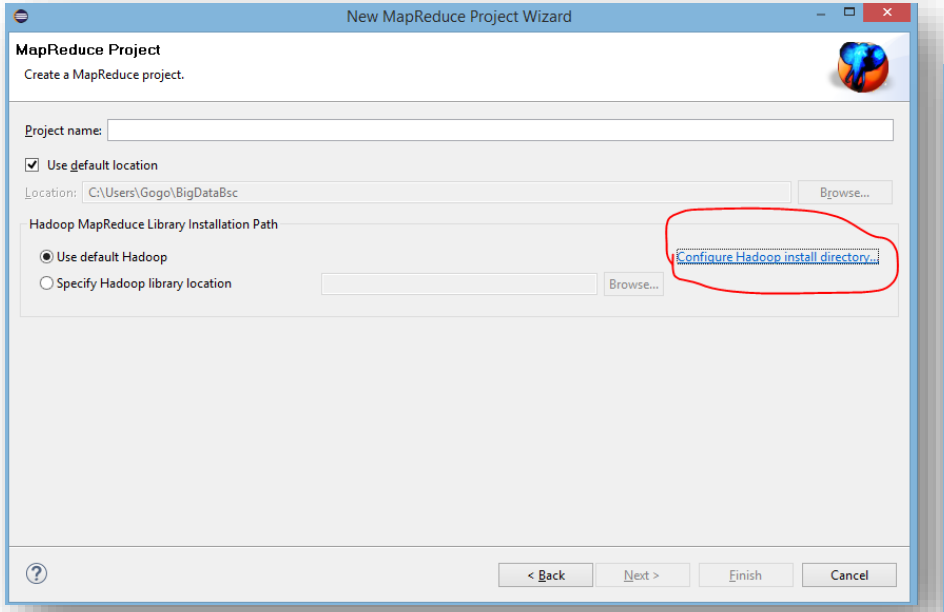
Futtatás Eclipse-ből

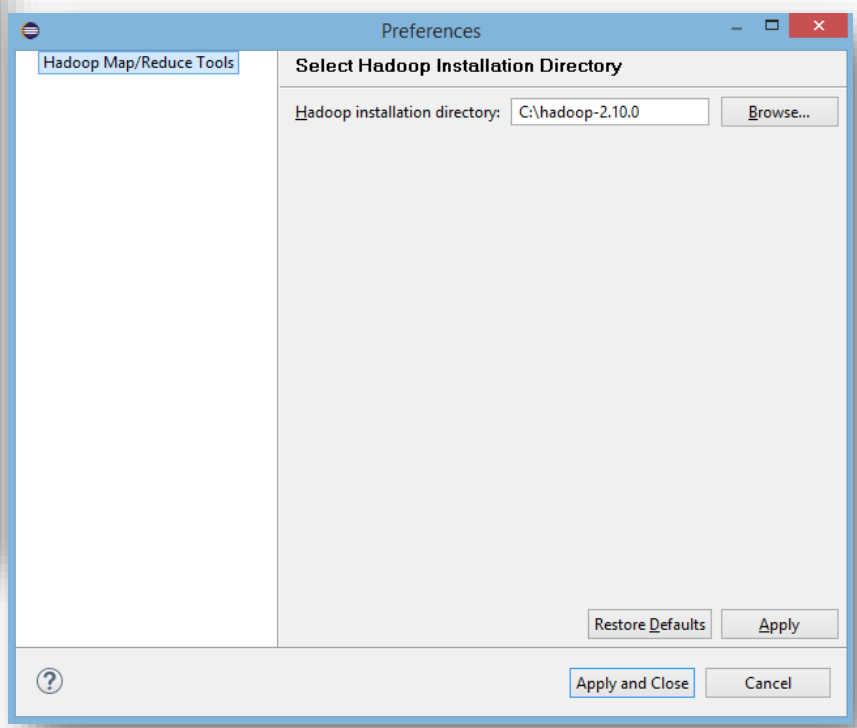





```
log4j.rootLogger=info,stdout
log4j.appender.stdout=org.apache.log4j.ConsoleAppender
log4j.appender.stdout.layout=org.apache.log4j.PatternLayout
# Pattern to output the caller's file name and line number.
log4j.appender.stdout.layout.ConversionPattern=%5p [%t] (%c:%L) %d{yyyy-MM-dd HH:mm:ss,SSS} ---- %m%n
```

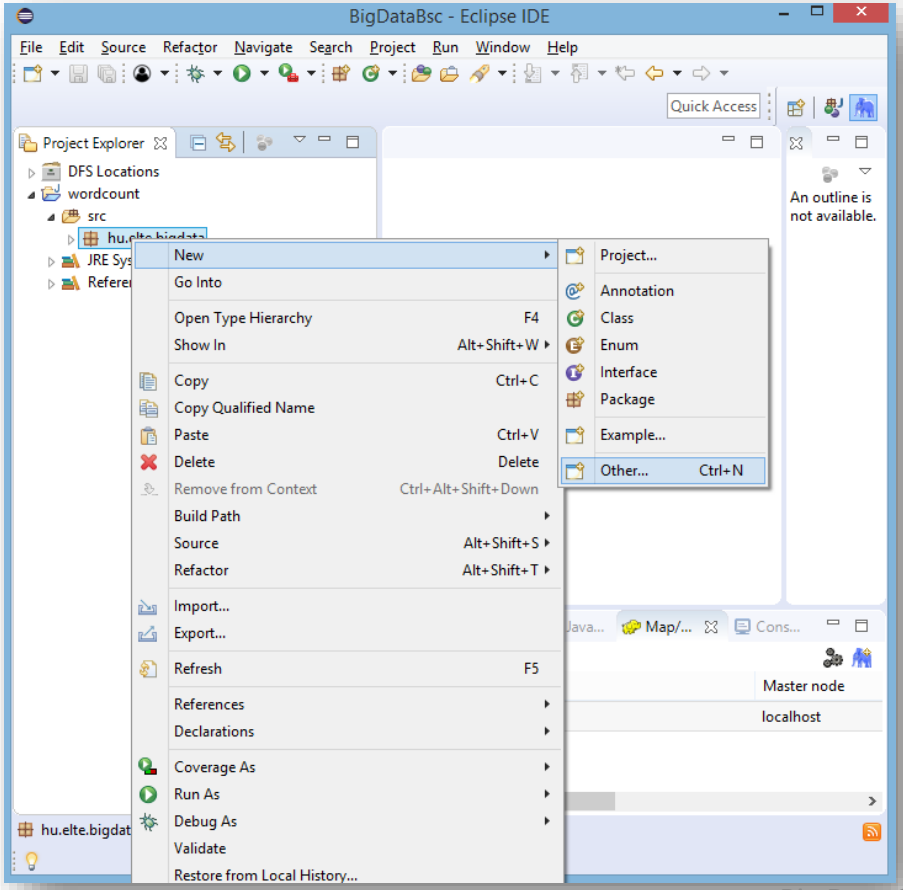
MapReduce projekt létrehozása

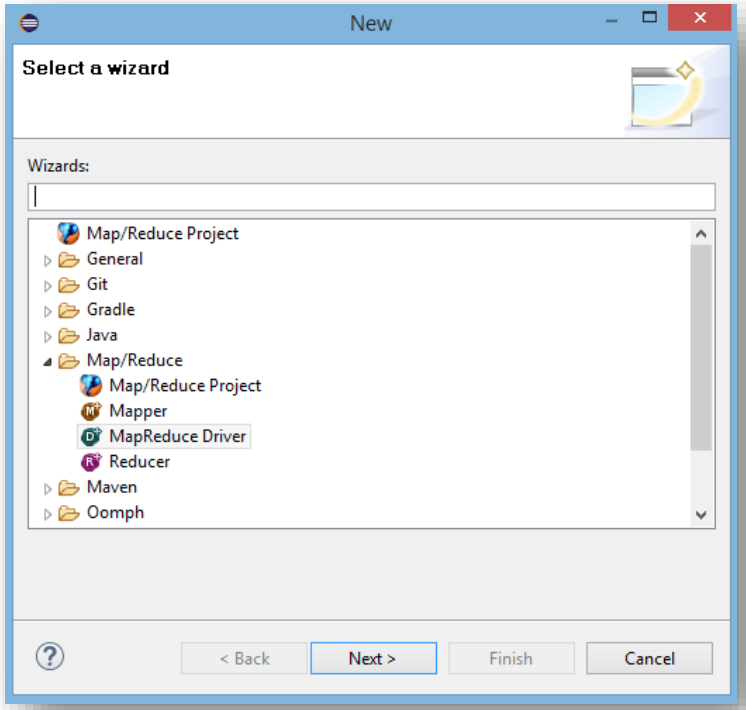




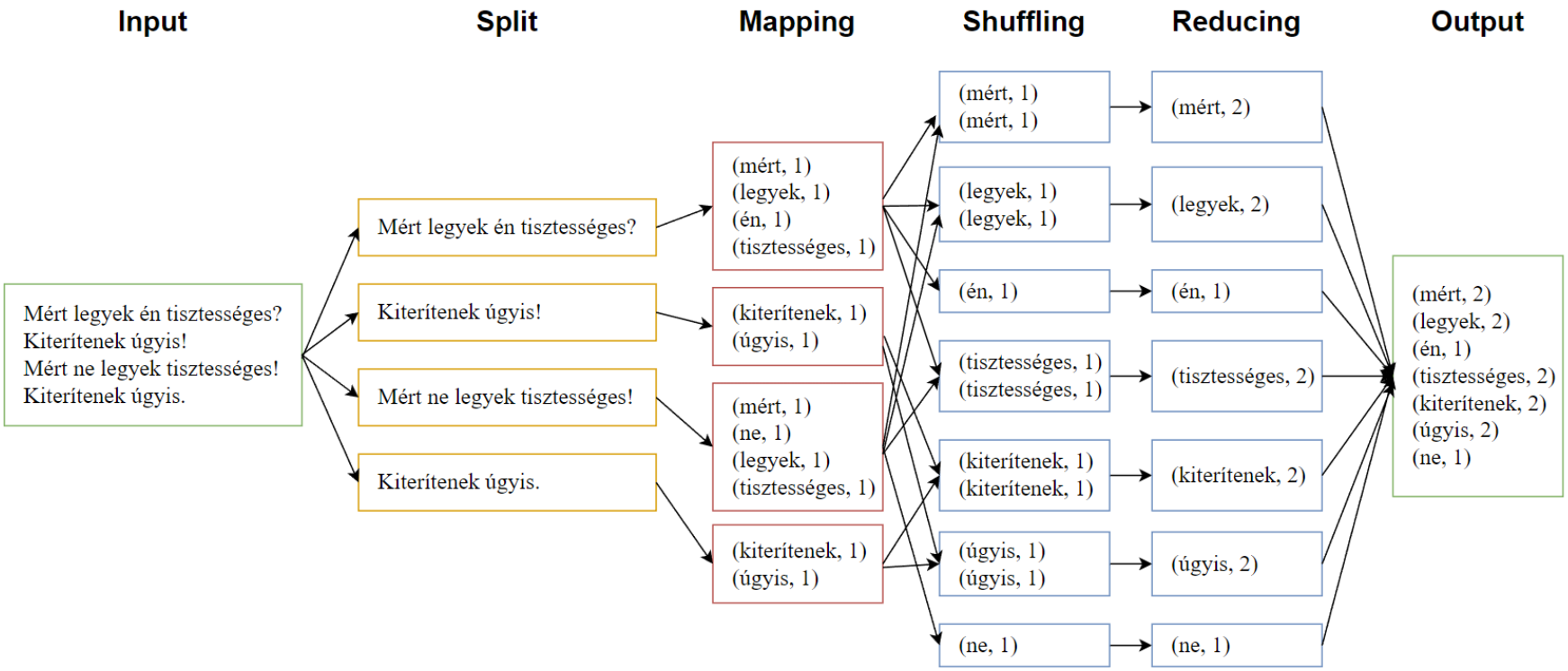
MapReduce osztályok

- FONTOS!!! Ne Mapper és Reducer-nek nevezzétek az osztályokat, mert a Hadoopnak van saját Mapper és Reducer osztályat és azt fogja használni!





Word Count


Input
Split
Mapping
Shuffling
Reducing
Output
Mért legyek én tisztességes?
Kiterítenek úgyis!
Mért ne legyek tisztességes!
Kiterítenek úgyis.
Mért legyek én tisztességes?
Kiterítenek úgyis!
Mért ne legyek tisztességes!
Kiterítenek úgyis.
(mért, 1)
(legyek, 1)
(én, 1)
(tisztességes, 1)
(kiterítenek, 1)
(úgyis, 1)
(mért, 1)
(ne, 1)
(legyek, 1)
(tisztességes, 1)
(kiterítenek, 1)
(úgyis, 1)
(mért, 1)
(mért, 1)
(legyek, 1)
(legyek, 1)
(én, 1)
(tisztességes, 1)
(tisztességes, 1)
(kiterítenek, 1)
(kiterítenek, 1)
(úgyis, 1)
(úgyis, 1)
(ne, 1)
(mért, 2)
(legyek, 2)
(én, 1)
(tisztességes, 2)
(kiterítenek, 2)
(úgyis, 2)
(ne, 1)
(mért, 2)
(legyek, 2)
(én, 1)
(tisztességes, 2)
(kiterítenek, 2)
(úgyis, 2)
(ne, 1)

Köszönöm a Figyelmet!