NPFL102 – Data Intensive Computing

- learn how to easily process terabytes of data using hundreds or thousands of computer cores
- start using **SGE** (originally Sun Grid Engine, now Oracle Grid Engine or Son of Grid Engine)
- get to know **MapReduce** programming model and Apache **Spark** framework – second generation framework for distributed execution of MapReduce and more complex paradigms with Python, Scala, Java and R APIs
- develop and debug simple and complex algorithms on our educational Spark cluster
- run machine learning algorithms in distributed fashion using Spark and **MLlib**
- advanced topics according to interest
  - distributed graph processing, OpenMPI, ...

http://ufal.mff.cuni.cz/course/npfl102

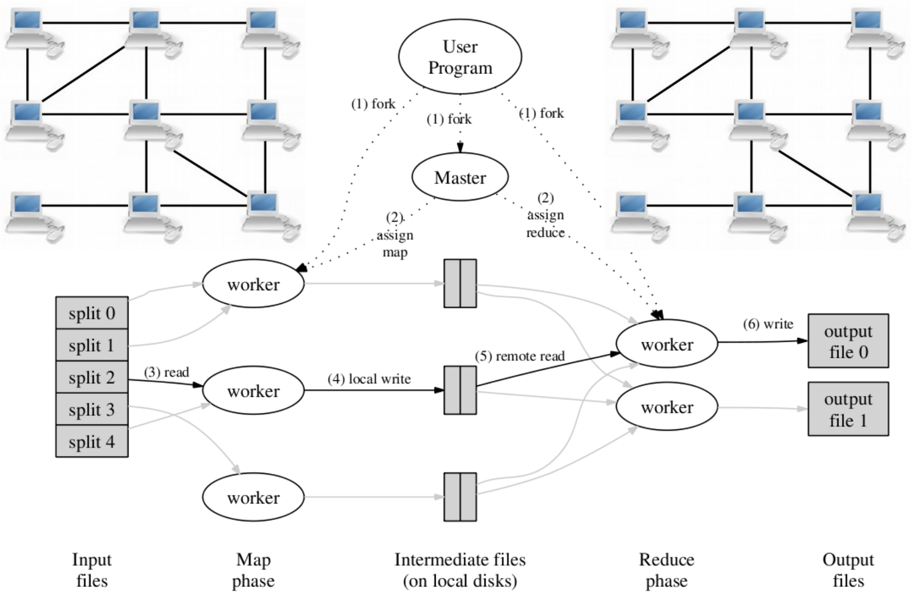
```
text_file = spark.textFile("hdfs://...")

text_file.flatMap(lambda line: line.split())
         .map(lambda word: (word, 1))
         .reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

Monday **15:40** in **SU1**

First lesson **February 29th**


User Program
(1) fork
(1) fork
(1) fork
Master
(2) assign map
(2) assign reduce
worker
worker
worker
worker
worker
split 0
split 1
split 2
split 3
split 4
(3) read
(4) local write
(5) remote read
(6) write
output file 0
output file 1
Input files
Map phase
Intermediate files (on local disks)
Reduce phase
Output files