# Goal 1: Create Fat-JAR on Laptop and run JavaSparkPi on Cluster

## Fat-JAR creation:

gradle fatJarWithTests

task fatJarWithTests(type: Jar) {

baseName = project.name + '-all-with-tests'

from { configurations.testCompile.collect { it.isDirectory() ? it : zipTree(it) } }

from sourceSets.test.output

with jar

}

## Command from example working

Download Spark examples from

<https://downloads.mesosphere.com/spark/assets/spark-examples_2.10-1.4.0-SNAPSHOT.jar>

sync-to-hdfs.sh ~/Downloads/spark-examples/ /spark-examples/

dcos spark run --submit-args='-Dspark.mesos.coarse=true --driver-cores 1 --driver-memory 1024M --class org.apache.spark.examples.SparkPi hdfs://hdfs/spark-examples/spark-examples\_2.10-1.4.0-SNAPSHOT.jar 30'



Alternative: Execute mesosphere example directly:  
dcos spark run --submit-args='-Dspark.mesos.coarse=true --driver-cores 1 --driver-memory 1024M --class org.apache.spark.examples.SparkPi [https://downloads.mesosphere.com/spark/assets/spark-examples\_2.10-1.4.0-SNAPSHOT.jar 30](https://downloads.mesosphere.com/spark/assets/spark-examples_2.10-1.4.0-SNAPSHOT.jar%2030)'

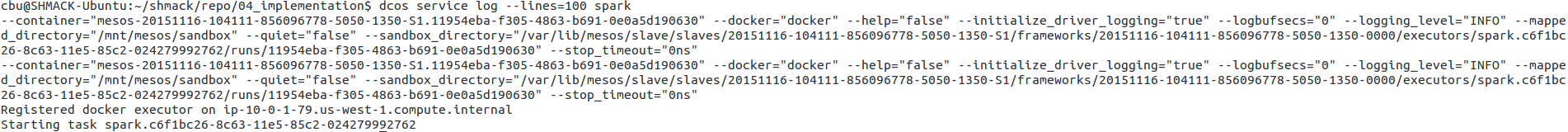
(See <https://docs.mesosphere.com/services/spark/> )

## Command from Fat-JAR not working

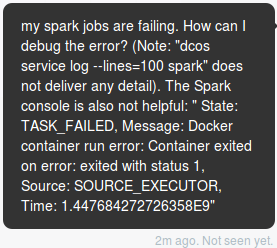
dcos spark run --submit-args='-Dspark.mesos.coarse=true --driver-cores 1 --driver-memory 1024M --class com.zuehlke.shmack.sparkjobs.tutorials.JavaSparkPi hdfs://hdfs/libs/sparkjobs-all-with-tests-1.0-SNAPSHOT.jar 30'



### Where is the “Why”? Where is the Log?

This is all ;-(   


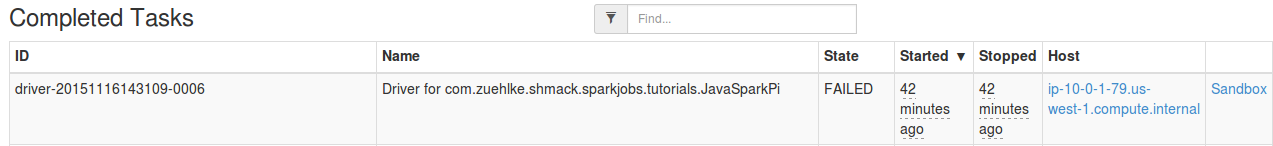
Request by chat:



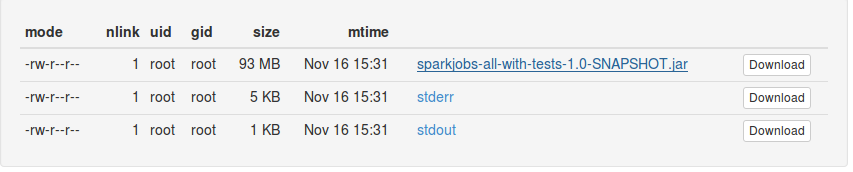
Solution:



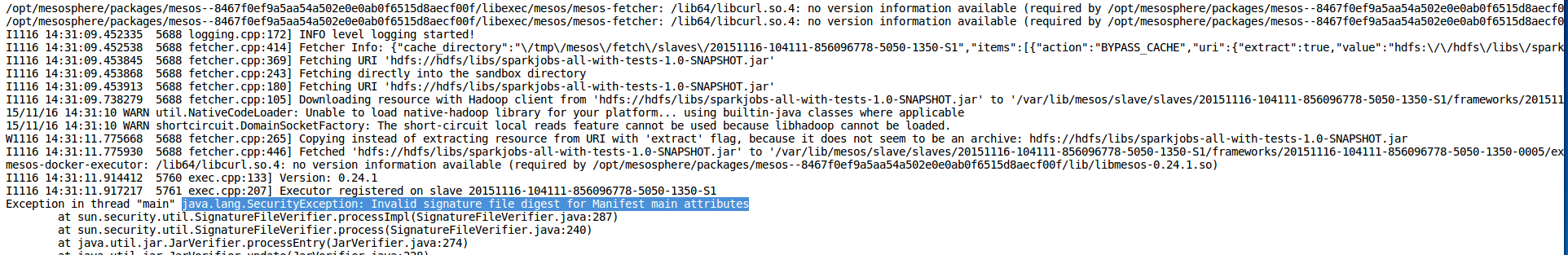
🡪



* Click on “sandbox”



Stderr shows the problem:



java.lang.SecurityException: Invalid signature file digest for Manifest main attributes

### Why is the manifest defective?

### Numerous (1022) duplicate files (but not the manifest)

unzip -l ../sparkjobs-all-with-tests-1.0-SNAPSHOT.jar | awk '{print $4;}' | sort | wc

58484 58481 3616989

unzip -l ../sparkjobs-all-with-tests-1.0-SNAPSHOT.jar | awk '{print $4;}' | sort -u | wc

57462 57461 3564247

* 58484 - 57462 = 1022

Which files:

unzip -l ../sparkjobs-all-with-tests-1.0-SNAPSHOT.jar | awk '{print $4;}' | sort | awk '{if (x[$1]) { x\_count[$1]++; print $0; if (x\_count[$1] == 1) { print x[$1] } } x[$1] = $0}'

* See <https://github.com/Zuehlke/SHMACK/blob/master/03_analysis_design/Issues/Issue-7%20Spark%20Word%20Count/duplicate-classes-in-fat-jar.txt>

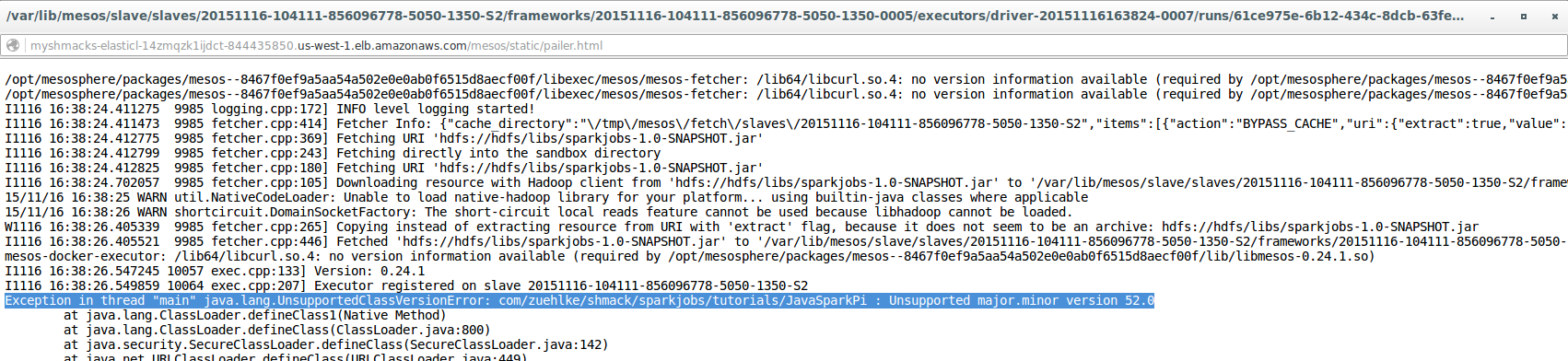
## Try “Slim-JAR”

cd ~/shmack/repo/04\_implementation/sparkjobs

gradle jar

sync-to-hdfs.sh ~/shmack/repo/04\_implementation/sparkjobs/build/libs /libs

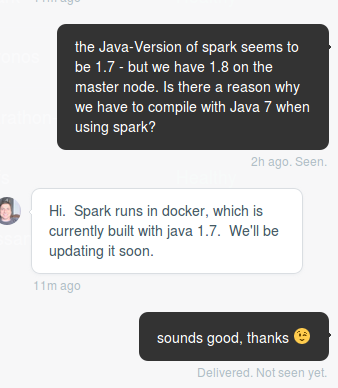
dcos spark run --submit-args='-Dspark.mesos.coarse=true --driver-cores 1 --driver-memory 1024M --class com.zuehlke.shmack.sparkjobs.tutorials.JavaSparkPi hdfs://hdfs/libs/sparkjobs-1.0-SNAPSHOT.jar 30'



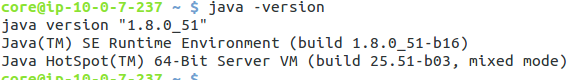
* Need to downgrade to Java 7

## Downgrade to Java 7

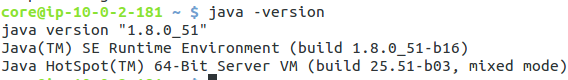
Interesting: Spark seems to have only Java 1.7 but the Master EC2 instance has Java 1.8…



Master:

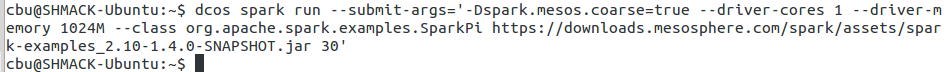


Slave 0:



## Try to run Jobs again --> failing

Even mesosphere reference Jobs are failing 🡪 no Job ID returned, nothing in Spark UI:



### Ask mesosphere…

Just submitted the Demo-spark-Job to my cluster.

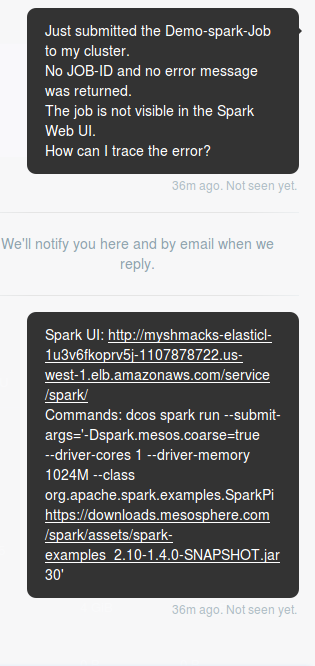
No JOB-ID and no error message was returned.

The job is not visible in the Spark Web UI.

How can I trace the error?

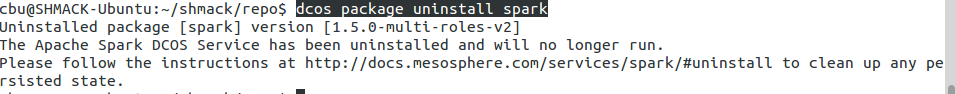
Spark UI: http://myshmacks-elasticl-1u3v6fkoprv5j-1107878722.us-west-1.elb.amazonaws.com/service/spark/

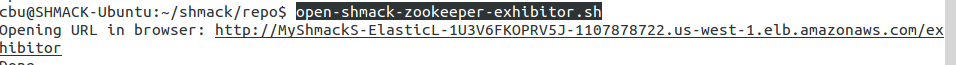
Commands: dcos spark run --submit-args='-Dspark.mesos.coarse=true --driver-cores 1 --driver-memory 1024M --class org.apache.spark.examples.SparkPi [https://downloads.mesosphere.com/spark/assets/spark-examples\_2.10-1.4.0-SNAPSHOT.jar 30](https://downloads.mesosphere.com/spark/assets/spark-examples_2.10-1.4.0-SNAPSHOT.jar%2030)'

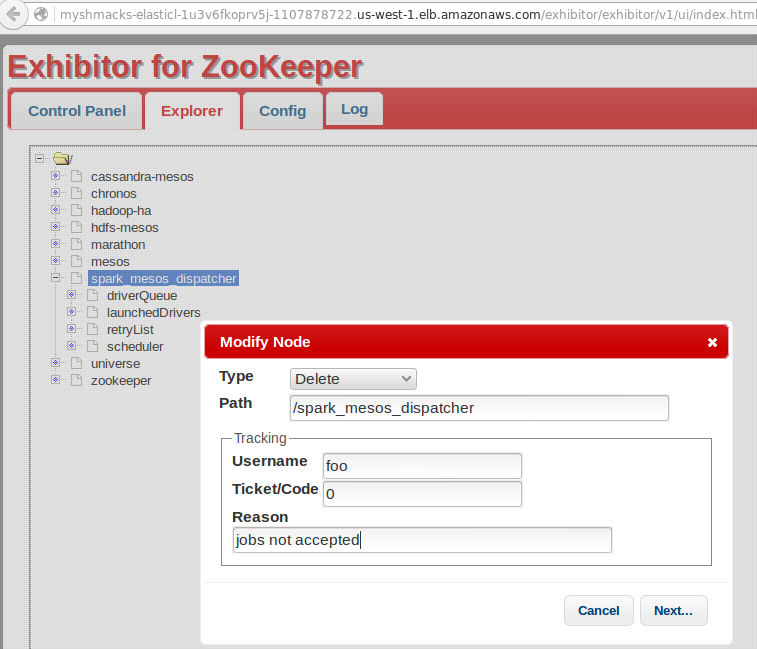


### Try to fix: Re-Install Spark as Service…

<http://docs.mesosphere.com/services/spark/#uninstall>







* OK

Install Spark again:

dcos package install spark

### Still not better



* Maybe downgrade to Java 7 could habe been the problem… ?
* Reboot VM
* No change. Still no JobID

### Re-Create cluster 🡪 Now it works:



## Now it’s working ;-)

dcos spark run --submit-args='-Dspark.mesos.coarse=true --driver-cores 1 --driver-memory 1024M --class com.zuehlke.shmack.sparkjobs.tutorials.JavaSparkPi hdfs://hdfs/libs/sparkjobs-1.0-SNAPSHOT.jar 30'

Run job succeeded. Submission id: driver-20151117135313-0008

hadoop fs -copyToLocal hdfs://hdfs/tmp/JavaSparkResult /tmp/JavaSparkResult

more /tmp/JavaSparkResult



BUT: we do not want to have a main in the productive code, which is used for testcases only.

Next Steps:

* Make a fat jar with testcases
* Avoid spark-base to be in fat jar ( as we do not want to upload 92 MB for a simple test…)

compile - Compile classpath for source set 'main'.

+--- org.apache.spark:spark-core\_2.11:1.5.1  
| +--- org.apache.avro:avro-mapred:1.7.7  
| | +--- org.apache.avro:avro-ipc:1.7.7  
| | … (several screen pages…)

* Successful ;-)

## Also Working with new Medium-Sized-Jar

cd ~/shmack/repo/04\_implementation/sparkjobs ; gradle clean fatJarWithTests

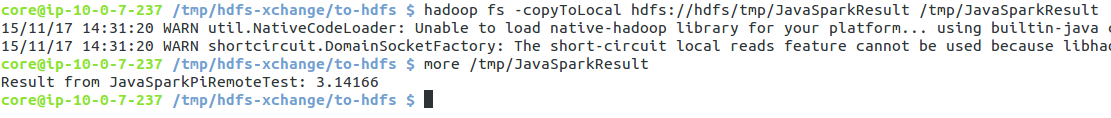
sync-to-hdfs.sh ~/shmack/repo/04\_implementation/sparkjobs/build/libs /sparkjobs-tests-libs

submit-spark-job.sh -Dspark.mesos.coarse=true --driver-cores 1 --driver-memory 1024M --class com.zuehlke.shmack.sparkjobs.tutorials.JavaSparkPiRemoteTest hdfs://hdfs/sparkjobs-tests-libs/sparkjobs-all-with-tests-1.0-SNAPSHOT.jar 30

rm /tmp/JavaSparkResult

hadoop fs -copyToLocal hdfs://hdfs/tmp/JavaSparkResult /tmp/JavaSparkResult

more /tmp/JavaSparkResult

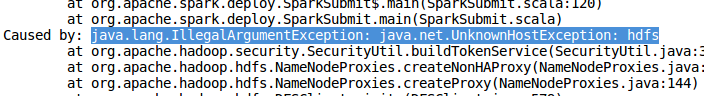


# Goal 2: Run WordCount on Cluster

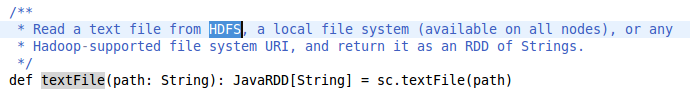
cd ~/shmack/repo/04\_implementation/sparkjobs ; gradle clean fatJarWithTests

sync-to-hdfs.sh ~/shmack/repo/04\_implementation/sparkjobs/build/libs /sparkjobs-tests-libs

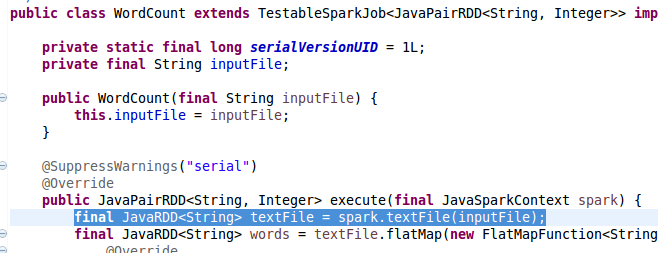
submit-spark-job.sh -Dspark.mesos.coarse=true --driver-cores 1 --driver-memory 1024M --class com.zuehlke.shmack.sparkjobs.wordcount.WordCountRemoteTest hdfs://hdfs/sparkjobs-tests-libs/sparkjobs-all-with-tests-1.0-SNAPSHOT.jar 30



Not nice:



Should work here:



We may be not alone with this problem (likely related to mesos):

<http://apache-spark-user-list.1001560.n3.nabble.com/Inconsistent-behavior-when-running-spark-on-top-of-tachyon-on-top-of-HDFS-HA-td1544.html>

## Prepare error report for mesosphere

* See Hostname-Not-Found-in-Spark-Job.docx in folder of this document.

# Additional Ideas

## Can we automate the navigation to the Logs based on the JobID… ?

Submission id: driver-20151117150232-0013

<http://myshmacks-elasticl-oa79a9ed8vqn-1620895950.us-west-1.elb.amazonaws.com/mesos/#/slaves/20151117-130536-3976658954-5050-1540-S3/browse?path=%2Fvar%2Flib%2Fmesos%2Fslave%2Fslaves%2F20151117-130536-3976658954-5050-1540-S3%2Fframeworks%2F20151117-130536-3976658954-5050-1540-0005%2Fexecutors%2Fdriver-20151117150232-0013%2Fruns%2F45b22b73-66e3-4418-9da9-cf60c2ea217c>

* Not so easy… skip this.

Small improvement:

