spark

with coursier

and ammonite

Alexandre Archambault
Criteo
`github.com/alexarchambault`
`@alxarchambault`

coursier

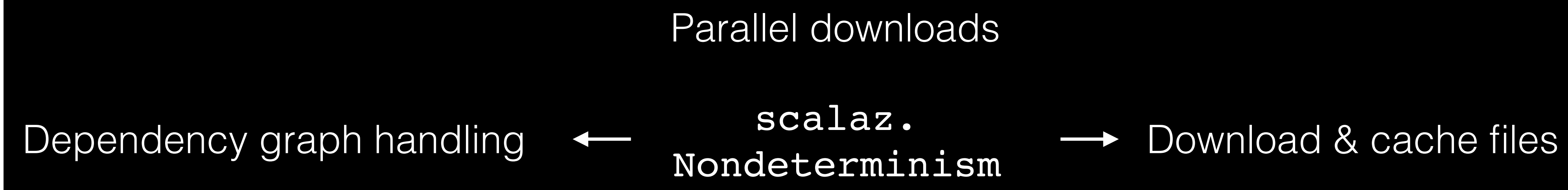
`http://get-coursier.io`

Library to manage dependencies

CLI tool

```
$ coursier fetch io.circe:circe-generic_2.11:0.9.0

$ coursier launch com.lihaoyi:ammonite_2.12.4:1.0.3
```



Parallel downloads

Dependency graph handling ← `scalaz.Nondeterminism` → Download & cache files

Why use coursier to handle spark jobs?

Develop spark job

- Grab spark distribution (contains spark JARs, …)
- Put your job in an assembly (merge JARs…)

```
$ sbt my-job/assembly
$ spark-submit \
    --master yarn-client \
    --executor-memory 4g \
    --num-executors 50 \
    my-job-assembly.jar \
      …
```

Run spark jobs

Many jobs

- as many assemblies
    - takes time to generate, load on CI
    - not that fit for Nexus servers
- spark distribution(s)
    - if automated, ad hoc scripts for that

Run spark jobs with coursier

```
$ sbt my-job/publish

$ coursier spark-submit \
    com.pany:my-job_2.11:0.x.y \
    -- \
      --master yarn \
      --executor-memory 4g \
      --num-executors 50 \
      -- \
        …
```

- Fetch JARs of my-job and its dependencies
- Find spark version
- Fetch JARs of spark
- Calls spark-submit in the spark JARs, passes it the job JARs + spark options

No spark distributions, no assemblies, all JARs automatically downloaded and cached

Run spark jobs with coursier

Limitations

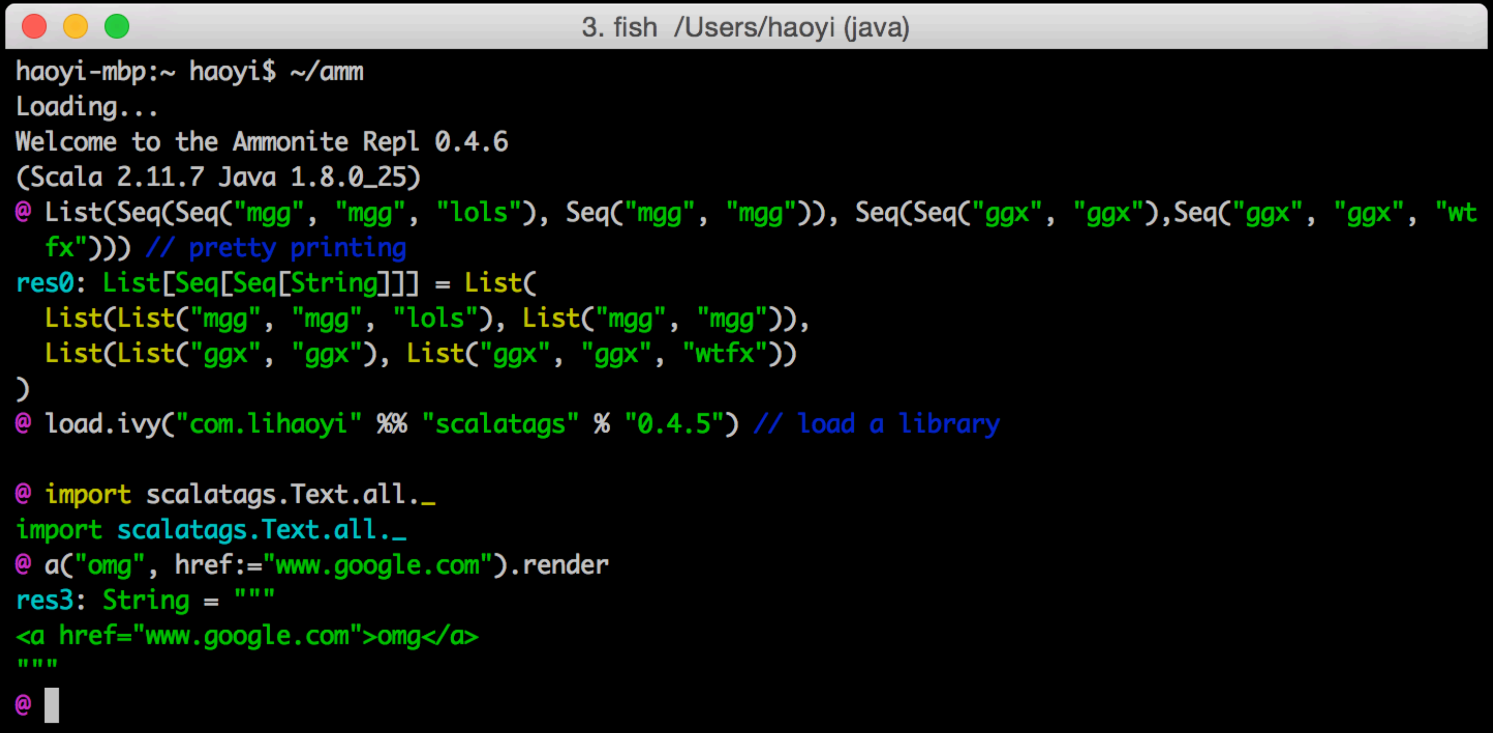
- only used on YARN clusters for now
- Mainly CLI tool, no clean API

ammonite

`http://ammonite.io`

More user-friendly scala REPL, by @li_haoyi

Pretty-printing, scripting, add dependencies on-the-fly, …



```
haoyi-mbp:~ haoyi$ ~/amm
Loading...
Welcome to the Ammonite Repl 0.4.6
(Scala 2.11.7 Java 1.8.0_25)
@ List(Seq(Seq("mgg", "mgg", "lols"), Seq("mgg", "mgg")), Seq(Seq("ggx", "ggx"),Seq("ggx", "ggx", "wt
  fx"))) // pretty printing
res0: List[Seq[Seq[String]]] = List(
  List(List("mgg", "mgg", "lols"), List("mgg", "mgg")),
  List(List("ggx", "ggx"), List("ggx", "ggx", "wtfx"))
)
@ load.ivy("com.lihaoyi" %% "scalatags" % "0.4.5") // load a library

@ import scalatags.Text.all._
import scalatags.Text.all._
@ a("omg", href:="www.google.com").render
res3: String = """
<a href="www.google.com">omg</a>
"""
@ 
```

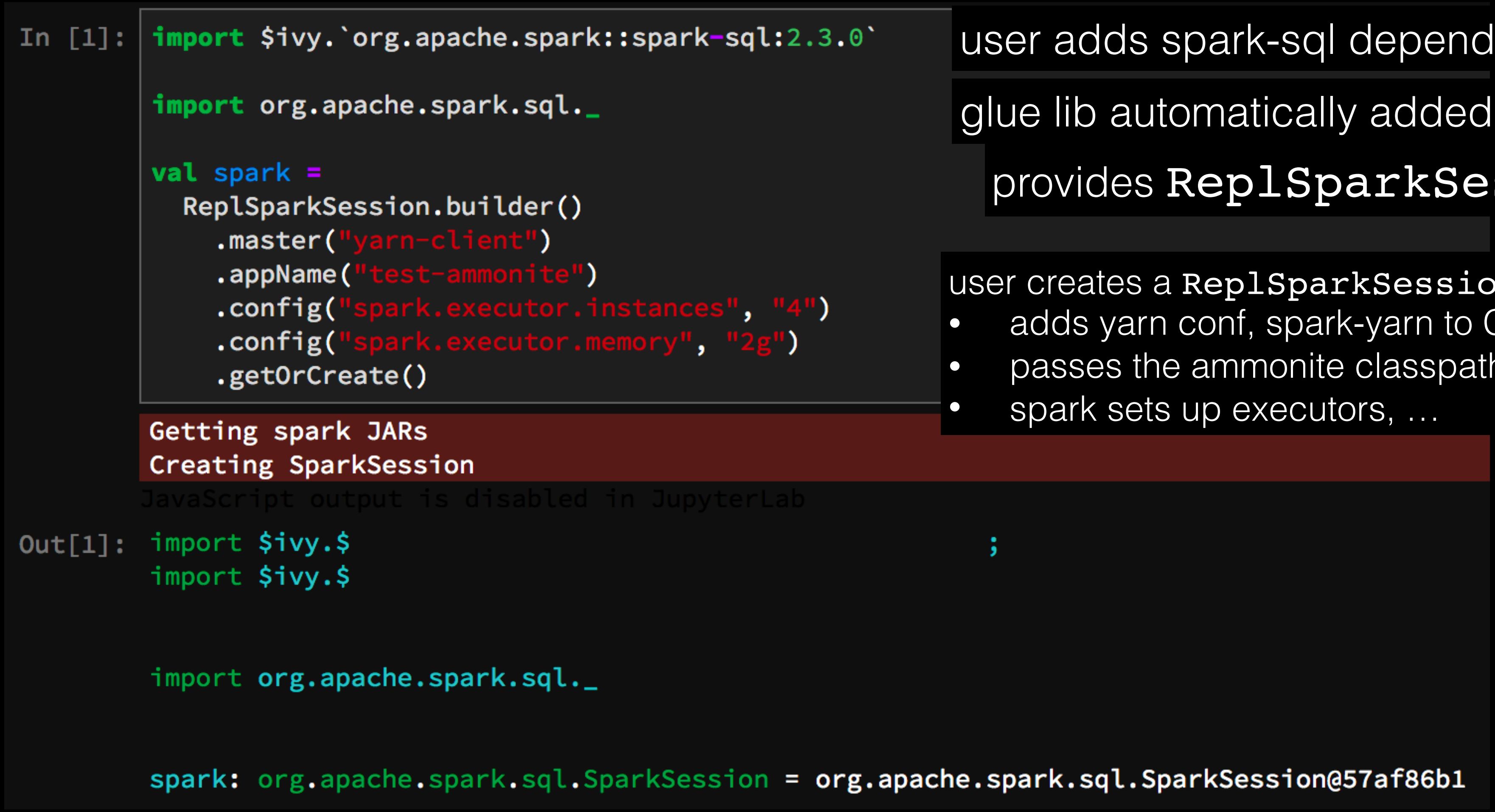
Spark with Ammonite

Status:

- Java serialization ✅ (#736, equivalent of `-Yrepl-class-based`)
- glue code, to pass the session classpath to spark ❌
- needs more careful classpath handling by Ammonite ❌

Shown here: unpublished things, still relying on bits of ammonium (`github.com/alexarchambault/ammonium`)

Ammonite session



```
In [1]: import $ivy.`org.apache.spark::spark-sql:2.3.0`

        import org.apache.spark.sql._

        val spark =
          ReplSparkSession.builder()
            .master("yarn-client")
            .appName("test-ammonite")
            .config("spark.executor.instances", "4")
            .config("spark.executor.memory", "2g")
            .getOrCreate()

        Getting spark JARs
        Creating SparkSession
        JavaScript output is disabled in JupyterLab

Out[1]: import $ivy.$                                              ;
        import $ivy.$


        import org.apache.spark.sql._


        spark: org.apache.spark.sql.SparkSession = org.apache.spark.sql.SparkSession@57af86b1
```

user adds spark-sql dependency

glue lib automatically added,
provides **`ReplSparkSession`**

user creates a **`ReplSparkSession`**

- adds yarn conf, spark-yarn to CP
- passes the ammonite classpath to spark
- spark sets up executors, …

No spark distributions

```
$ coursier spark-submit \
    com.pany:my-job_2.11:0.x.y \
    -- \
      --master yarn \
      --executor-memory 4g \
      --num-executors 50 \
      -- \
        …
```

```
$ amm
@ import $ivy.`org.apache.spark::spark-sql:2.3.0`
  import org.apache.spark.sql._
  val spark = ReplSparkSession.builder().
    appName("test").
    master("yarn-client")
    config("spark.executor.instances", "50").
    config("spark.executor.memory", "2g").
    getOrCreate()
```

spark with coursier and ammonite

Needs *you*

- `coursier spark-submit` soon in its own repository
- all of that really only tuned for YARN clusters
- last PRs for spark in Ammonite soon, come tell your opinion, contribute, …

Questions?