
H2O
WORLD
2017

Jakub Háva

Software Engineer

jakub@h2o.ai

Sparkling Water

PySparkling

First-time Qwiklab Account Setup

- Go to http://h2oai.qwiklab.com
- Click on “JOIN”
- Create a new account with a valid email address
- You will receive a confirmation email
  - Click on the link in the confirmation email
- Go back to http://h2oai.qwiklab.com and log in
- Go to the Catalog on the left bar
- Choose “Sparkling Water”

MEET THE MAKERS

MICHAL MALOHLAVA

Chief platform architect at H2O.ai and creator of Sparkling Water

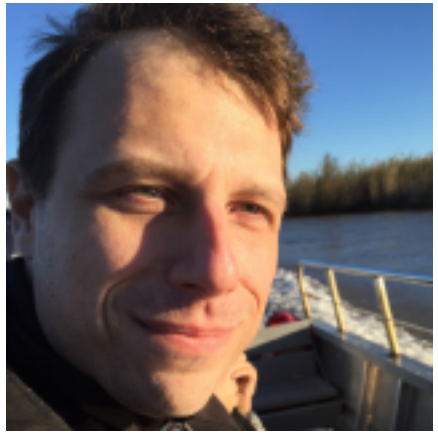
NAVDEEP GILL

Software Engineer and Data Scientist at H2O.ai, author of RSparkling.

Huge thanks to Michal for help and the guidance with the materials for this presentation

JAKUB HAVA

Software Engineer at H2O.ai at Sparkling Water project.



MICHAL KURKA

Head of H2O and senior Software Engineer at H2O.ai

H2O WORLD 2017

Sparkling Water Road Map

Feature	Q1	Q2	Q3	Q4
Continuous updates to Spark/H2O				
Stability Fixes				
Enterprise Steam Integration				
Driverless AI MOJOs Support				
Telemetry				
Pipelines Improvements				
Tighter Integration with Spark API				
More H2O algorithms exposed				

ARCHITECTURE

PySparkling

Sparkling Water Overview

- Transparent **integration** of H2O with Spark ecosystem
  - MLlib and H2O side-by-side
- Transparent **use** of H2O data structures and algorithms with Spark API
- **Platform** for building Smarter Applications
- Excels in existing Spark workflows requiring **advanced Machine Learning algorithms**

Ecosystem

Scala: Sparkling Water

Spark

```
val sc =
SparkContext.getOrCreate(...)

val df = sc.parallelize(1 to 10).toDF
```

```
val h2oContext =
H2OContext.getOrCreate(sc)

val hf = h2oContext.asH2OFrame(df)
```

Python: PySparkling Water

PySpark

```
sc = SparkContext(...)

df = sc.parallelize(range(1,11))
       .toDF("int")
```

```
h2o_context =
H2OContext.getOrCreate(sc)

hf = h2o_context.as_h2o_frame(df)
```

R: RSparkling Water

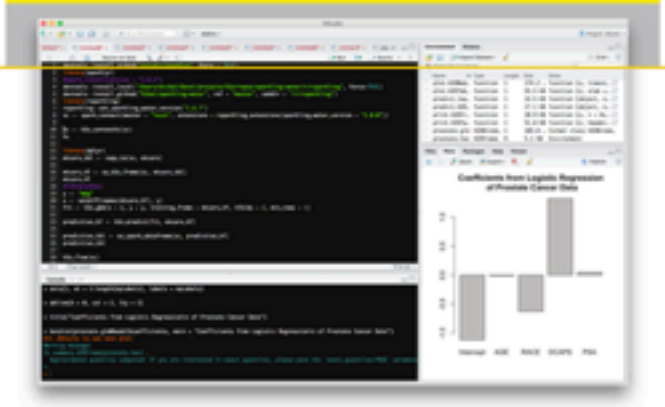
sparklyr

```
sc <- spark_connect(...)

tbl <- data_frame(c(1:10))
df <- copy_to(sc, tbl)
```

```
hc <- h2o_context(sc)

hf <- as_h2o_frame(sc, df)
```



H2O WORLD 2017

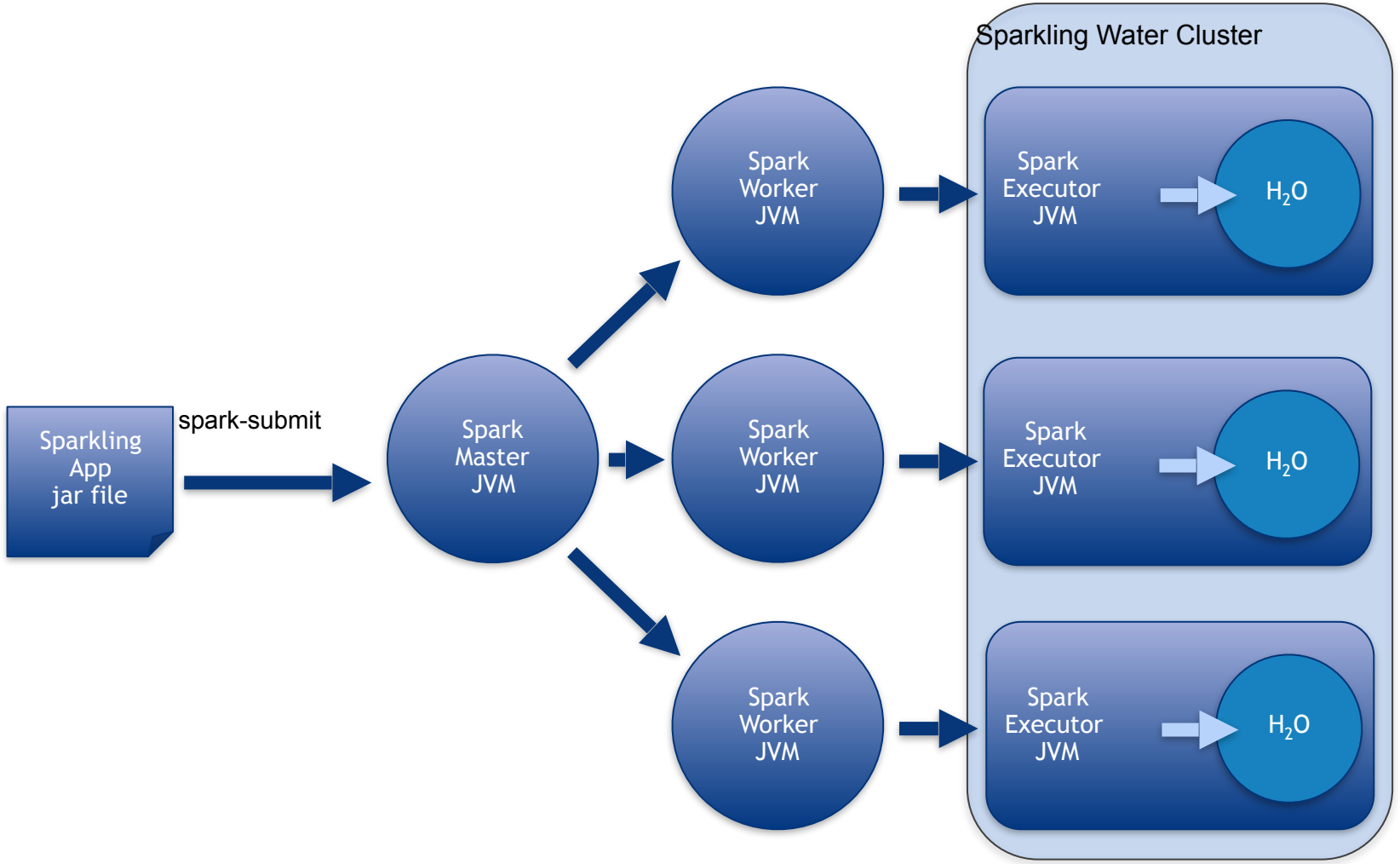
Benefits

- Additional algorithms
  - NLP
- Powerful data munging
  - SQL
- ML Pipelines

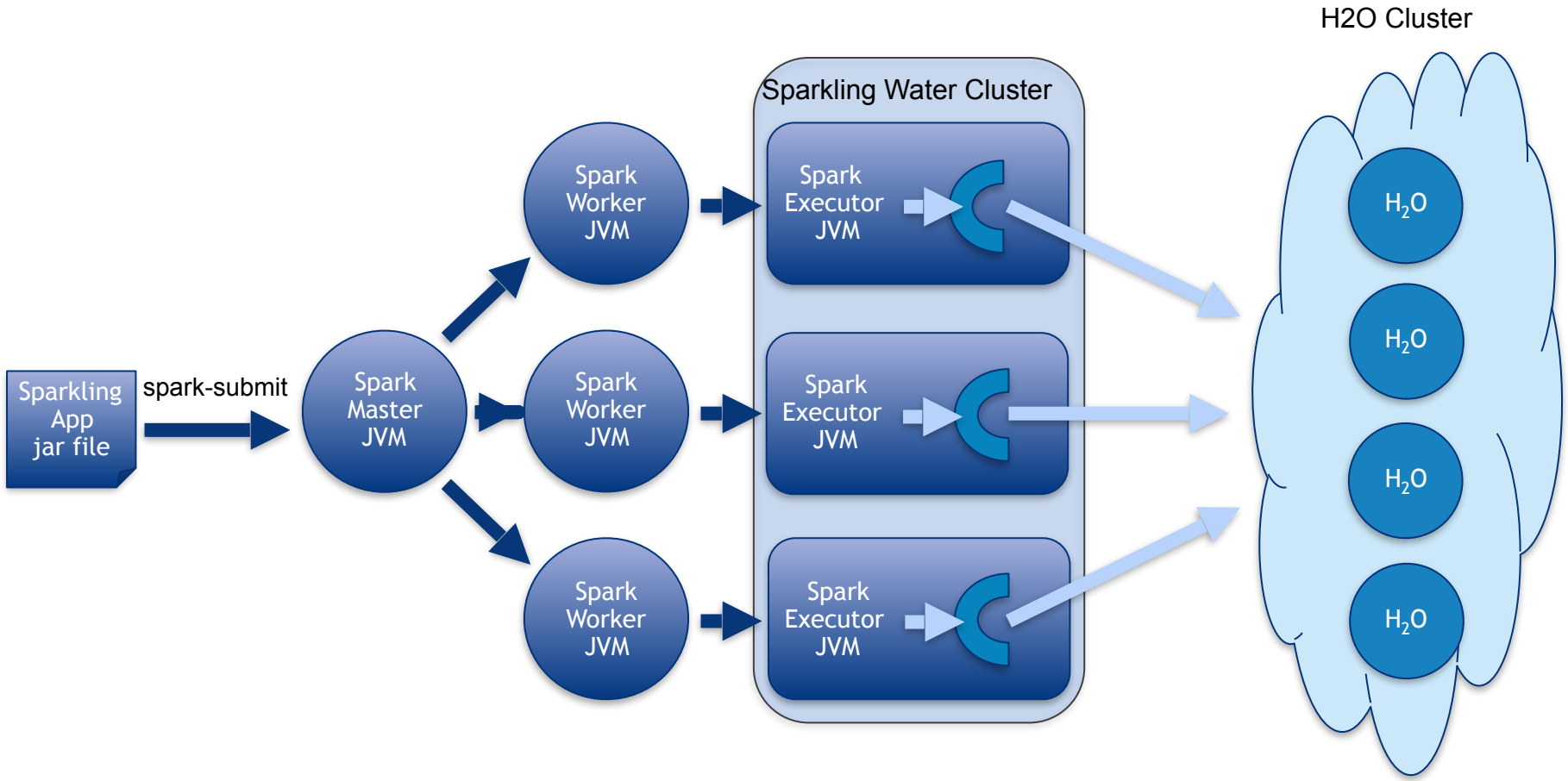
- Advanced algorithms
  - speed v. accuracy
  - advanced parameters
- Fully distributed and parallelised
- Graphical environment
- R/Python interface

H2O WORLD 2017

Internal Backend



External Backend



DEMO TIME!

The topic

- The goal of this demo is to train the pipeline in PySpark
- The resulted pipeline will be exported into language independent format
- The stored pipeline will be deployed in Scala as part of Streaming App

Resources

- Documentation: http://docs.h2o.ai
- Tutorials: https://github.com/h2oai/h2o-tutorials
- Slidedecks: https://github.com/h2oai/h2o-meetups
- Videos: https://www.youtube.com/user/0xdata
- Events & Meetups: http://h2o.ai/events
- Stack Overflow: https://stackoverflow.com/tags/sparkling-water
- Google Group: https://tinyurl.com/h2ostream
- Gitter: http://gitter.im/h2oai/sparkling-water

Thank you!

Sparkling Water is
open-source
ML application platform
combining
power of Spark and H2O

Learn more at h2o.ai

Follow us at @h2oai

H2O WORLD 2017