

Jakub Háva  
[jakub@h2o.ai](mailto:jakub@h2o.ai)

# Sparkling Water 2.0

R Addicts Paris, Paris  
December 1, 2016

 Spark + H<sub>2</sub>O

---

SPARKLING  
WATER

# **Who am I**

- Finishing high-performance cluster monitoring tool for JVM based languages ( JNI, JVMTI, instrumentation )
- Finishing Master's at Charles Uni in Prague
- Software engineer at H2O.ai - Core Sparkling Water
- Tea lover ( doesn't mean I don't like beer!)

# **Distributed Sparkling Team**

- Michal - Mt. View, CA
- Kuba - Prague, CZ
- Mateusz - Tokyo, JP
- Vlad - Mt. View, CA

**H<sub>2</sub>O+Spark =  
Sparkling  
Water**

# Sparkling Water

- 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

Functionality missing in H2O can be replaced by Spark and vice versa

# Benefits



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



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

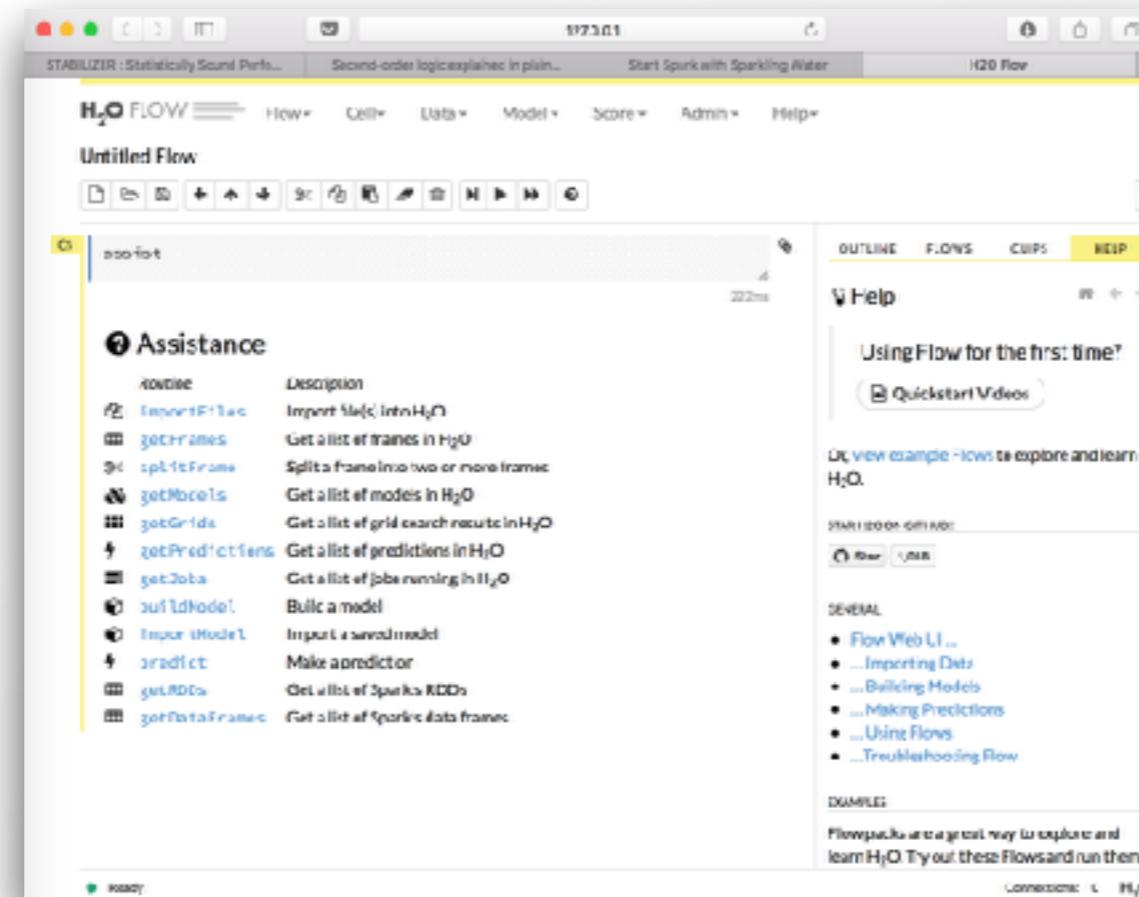
# How to use Sparkling Water?

# Start spark with Sparkling Water

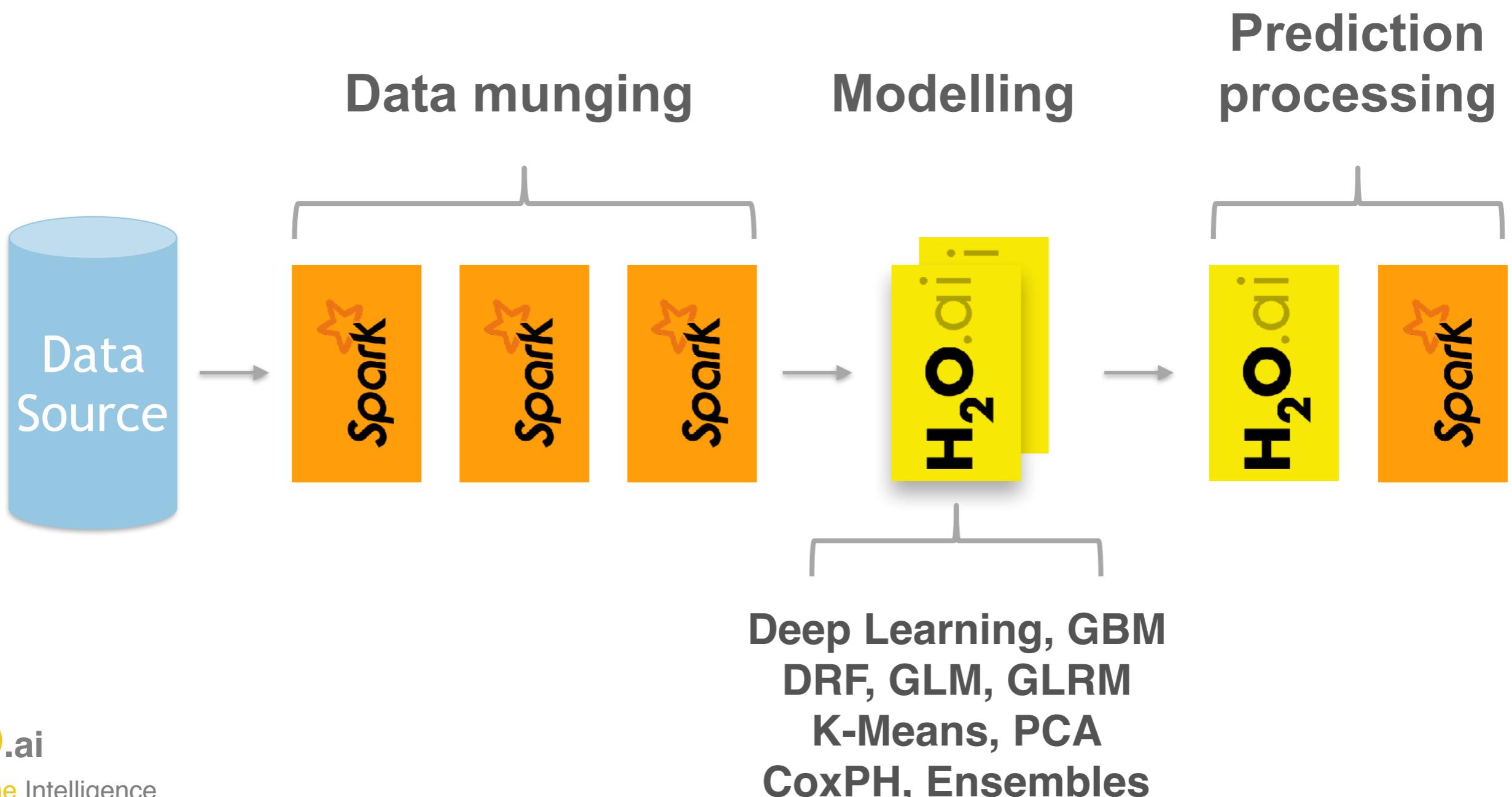
start.sh

```
1 $SPARK_HOME/bin/spark-submit \
2   --class water.SparklingWaterDriver \
3   --packages ai.h2o:sparkling-water-examples_2.10:1.6.3 \
4   --executor-memory=6g \
5   --driver-class-path scalastyle.jar /dev/null
```

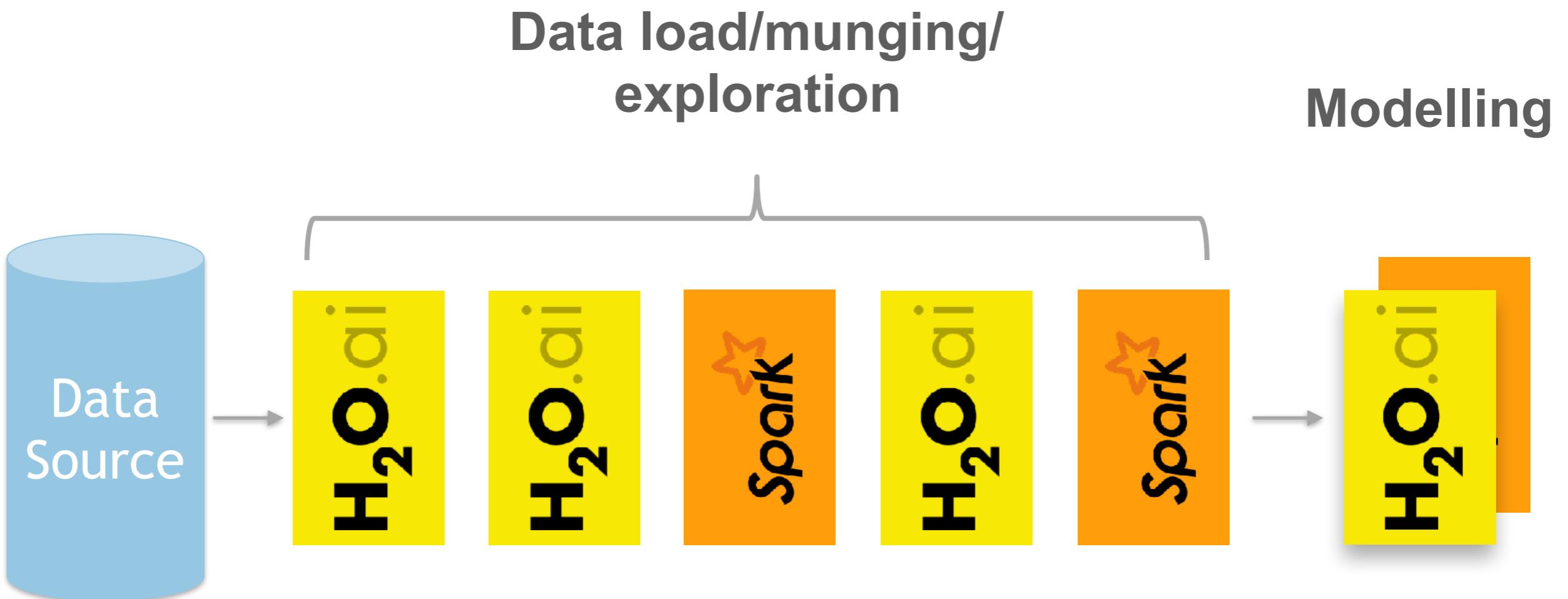
Raw



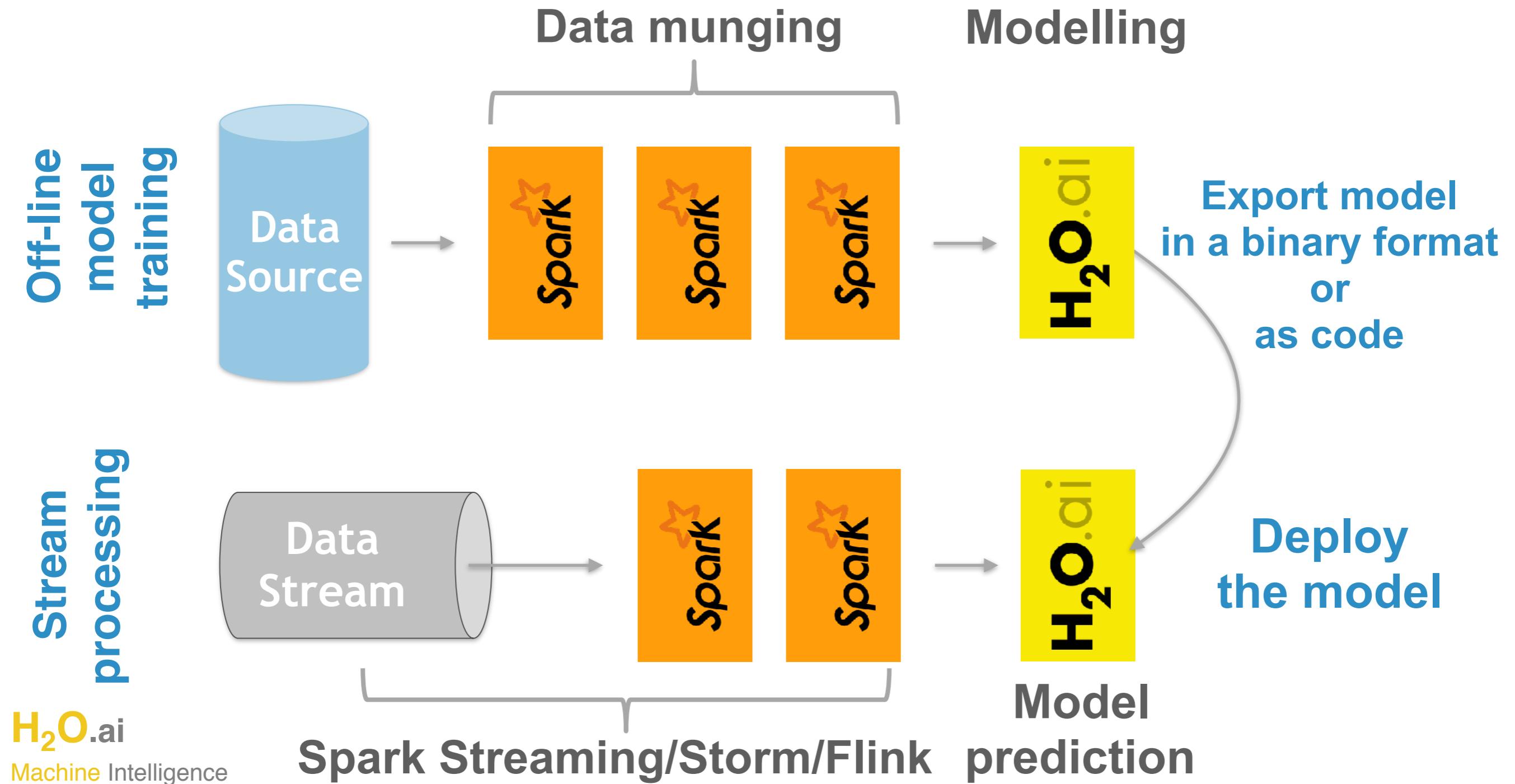
# Model Building



# Data Munging

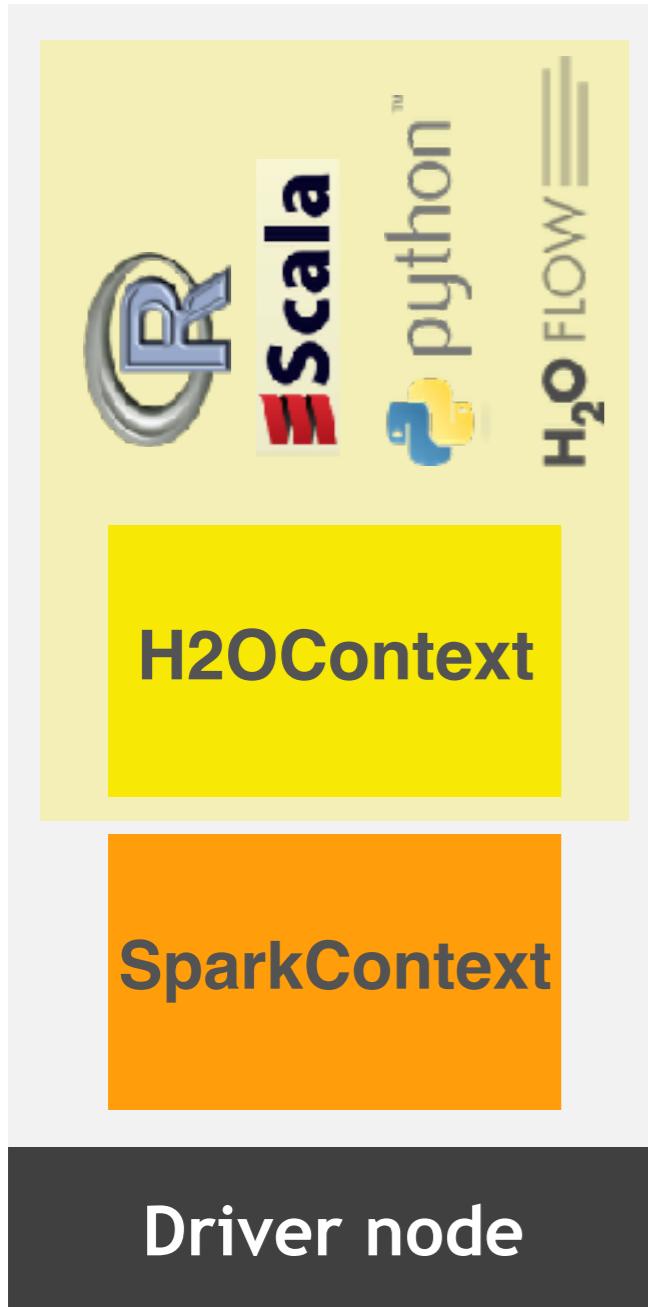


# Stream processing



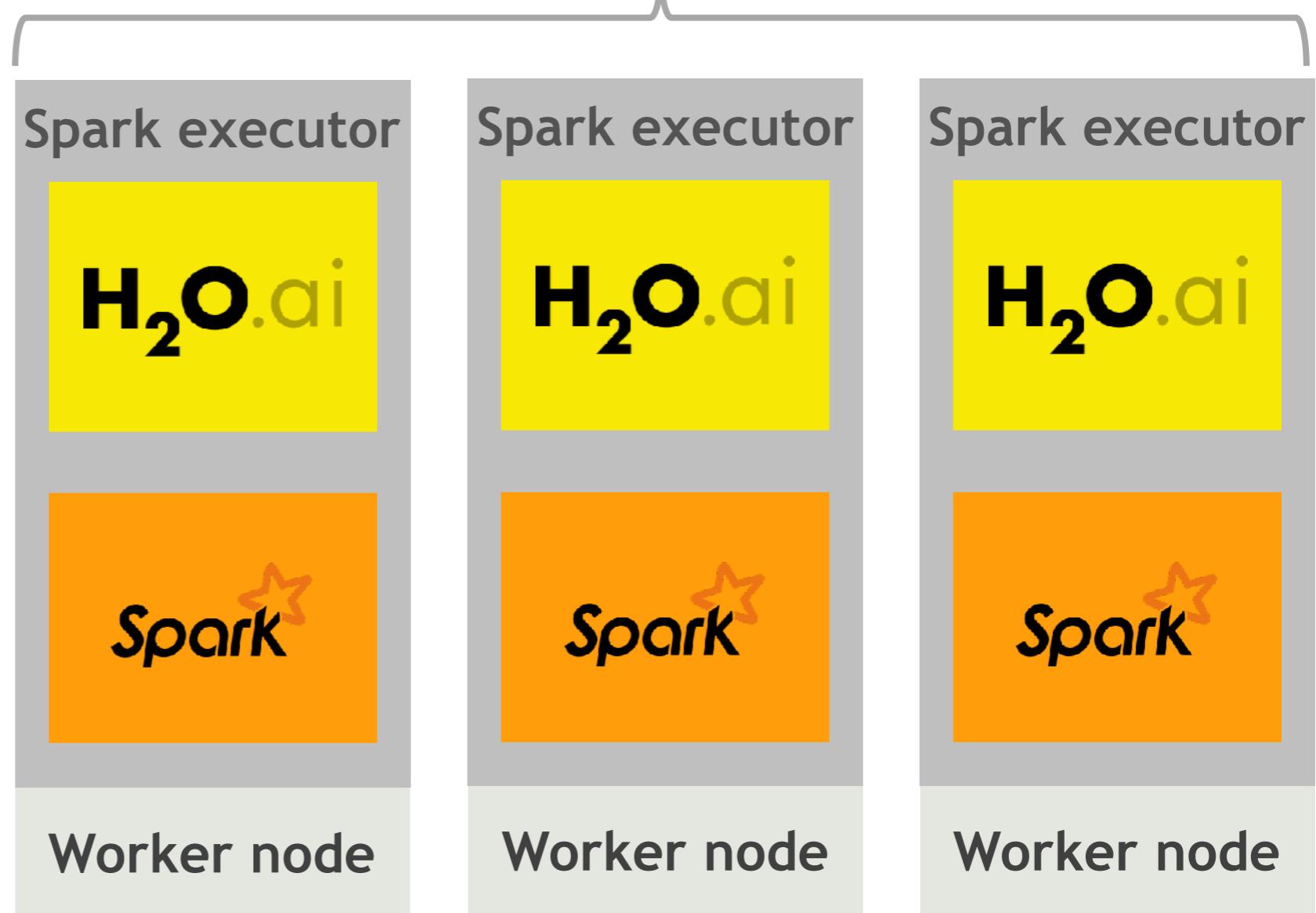
**What is  
inside?**

# Scala/Py main program

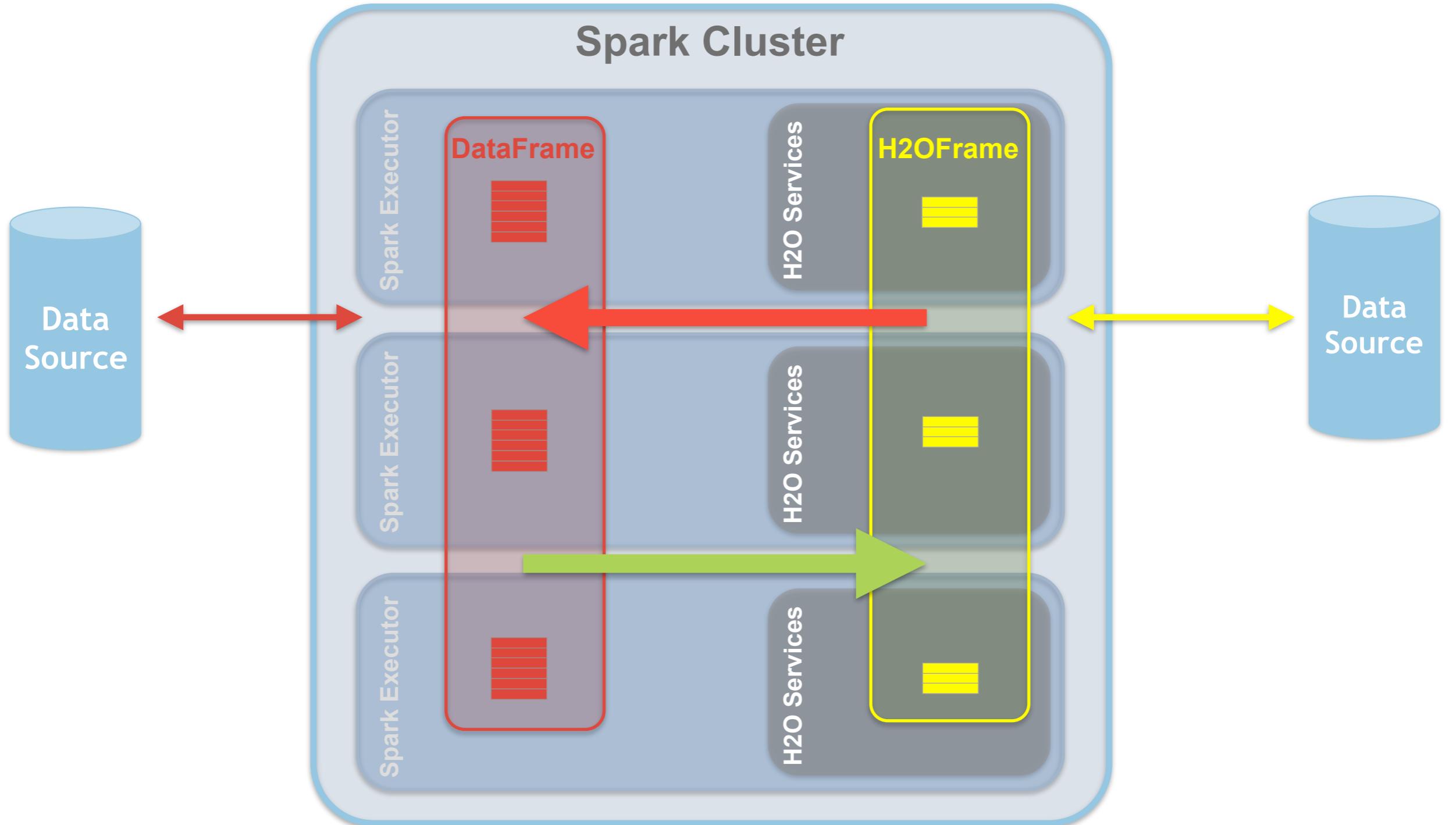


# Spark + H<sub>2</sub>O

## SPARKLING WATER



Cluster



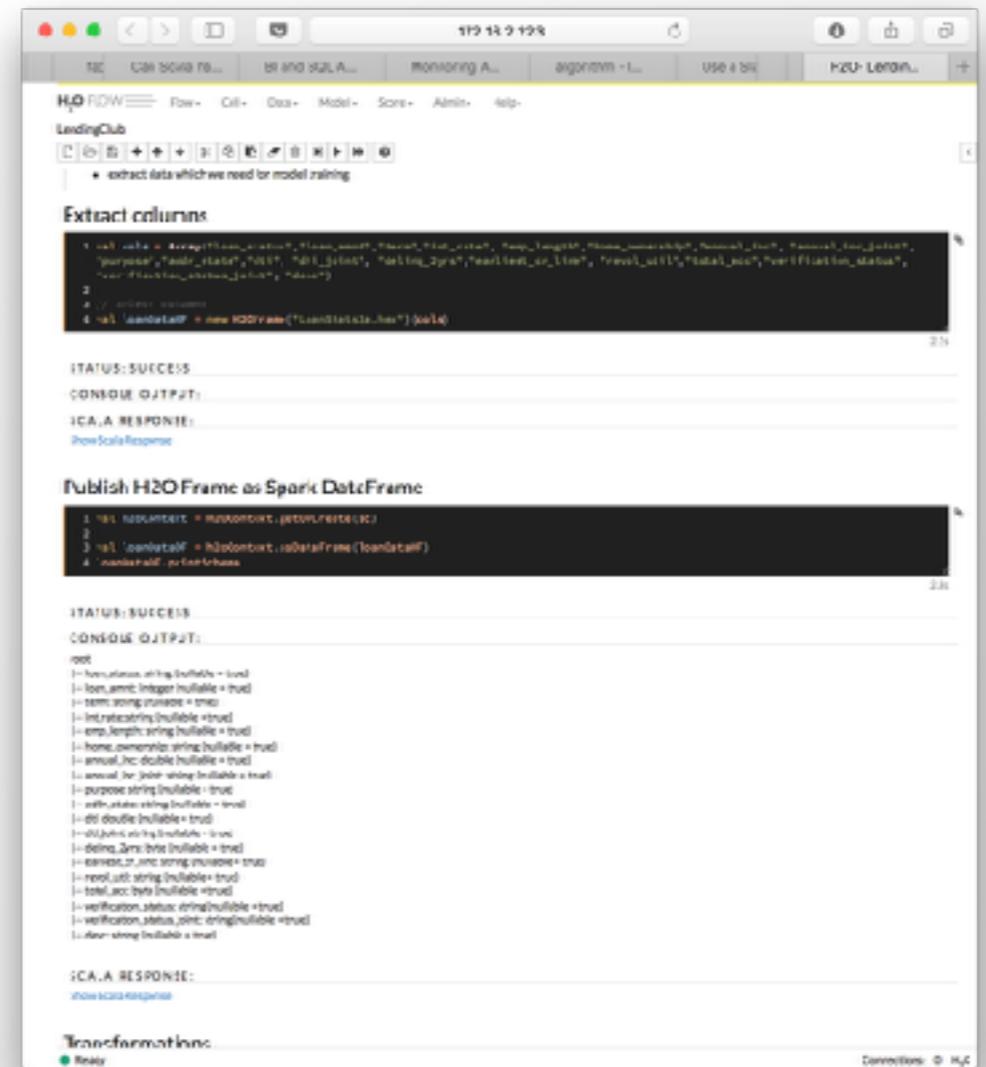
← **h2oContext.asDataFrame**

→ **h2oContext.asH2OFrame**

**New features,  
finally!**

# Scala code in H2O Flow

- New type of cell
  - Access Spark from Flow UI
  - Experimenting made easy



# **H2O Frame as Spark's Datasource**

- Use native Spark API to load and save data
- Spark can optimise the queries when loading data from H2O Frame
- Use of Spark query optimiser

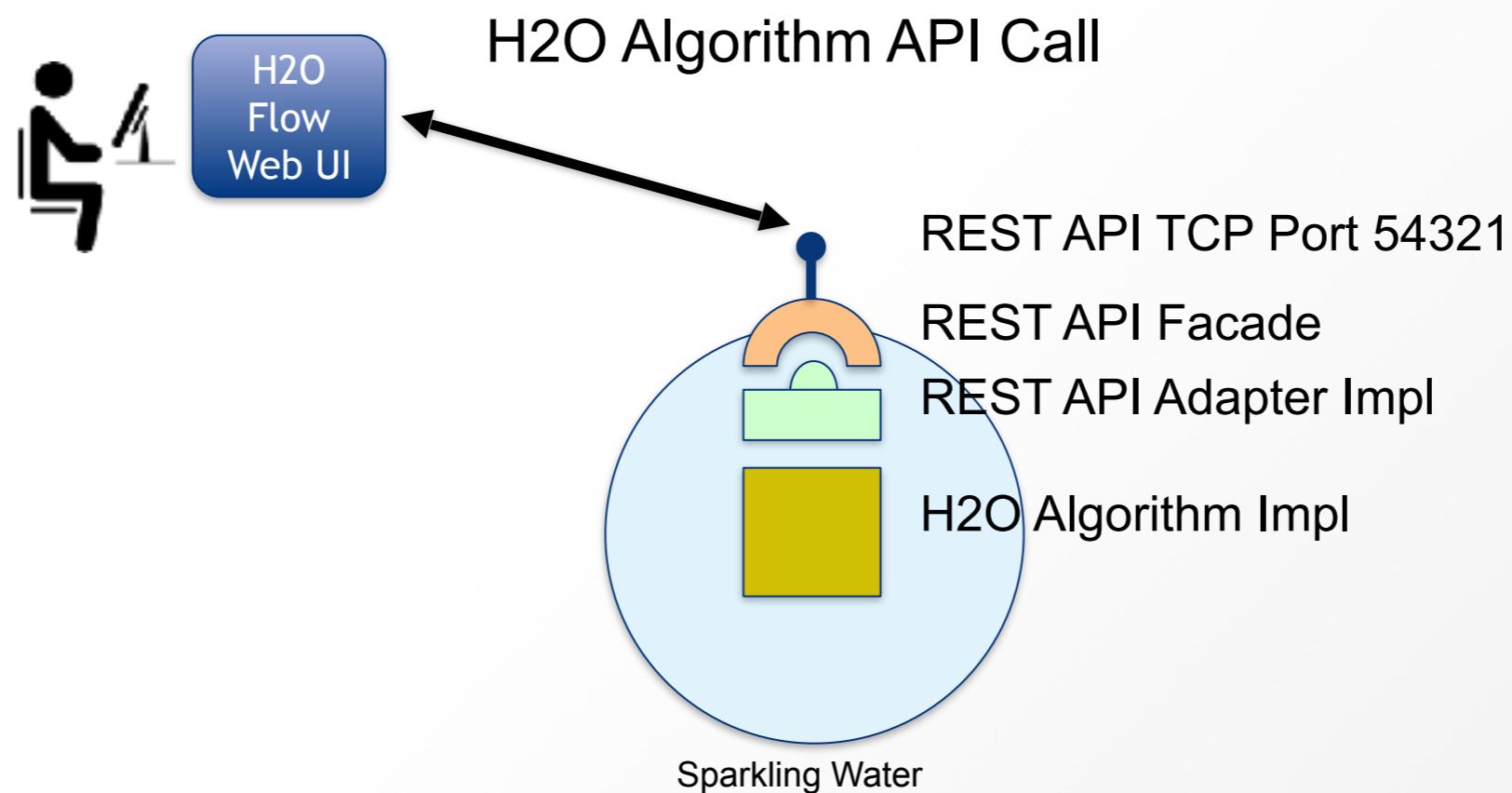
# Machine learning pipelines

- Wrap our algorithms as Transformers and Estimators
- Support for embedding them into Spark ML Pipelines
- Can serialise fitted/unfitted pipelines
- Unified API => Arguments are set in the same way for Spark and H2O Models

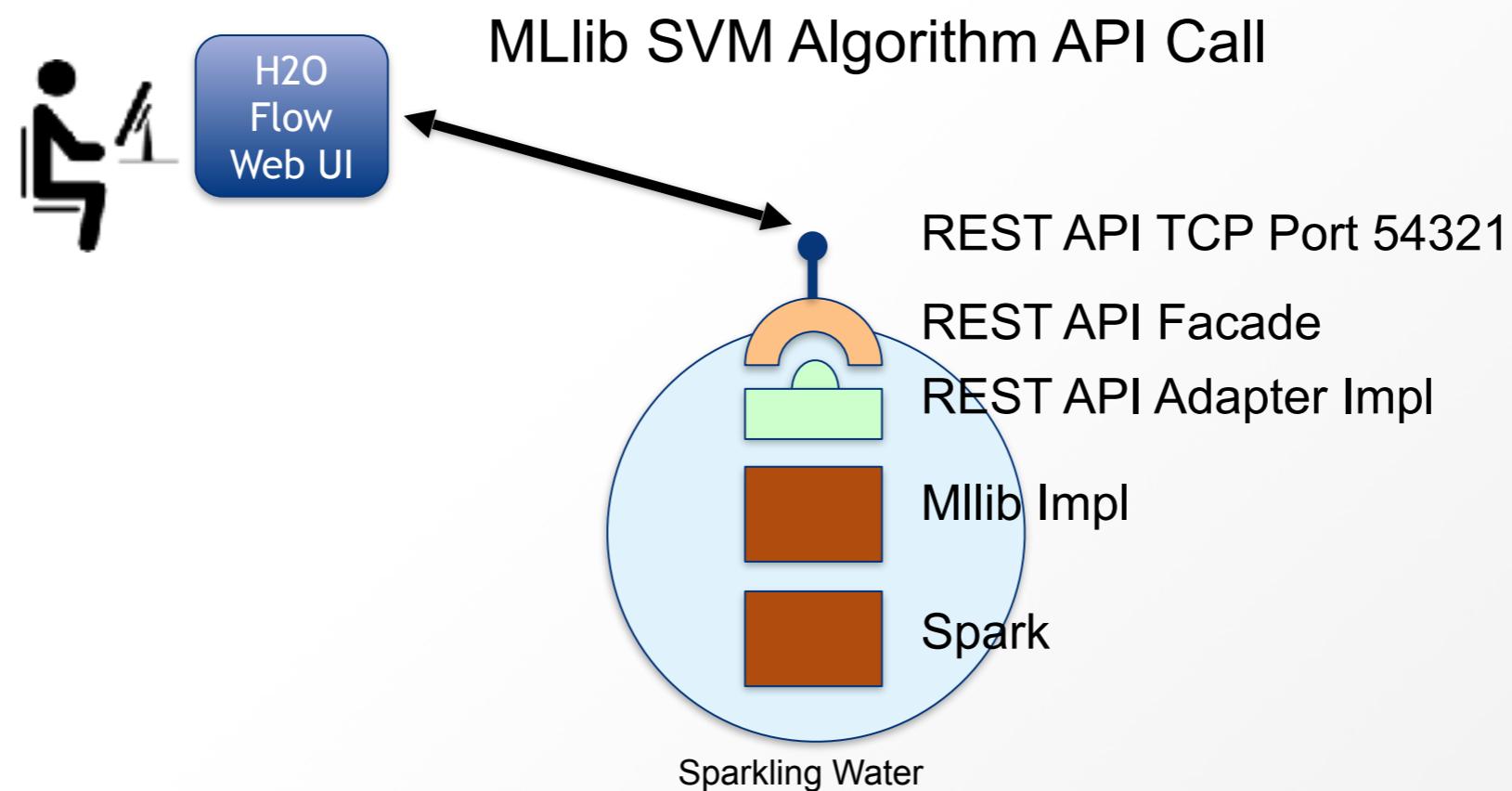
# **MLlib Algorithms in Flow UI**

- Can examine them in H2O Flow
- Can generate POJO out of them

# H2O Algo from Flow



# Pure MLlib Algo from Flow



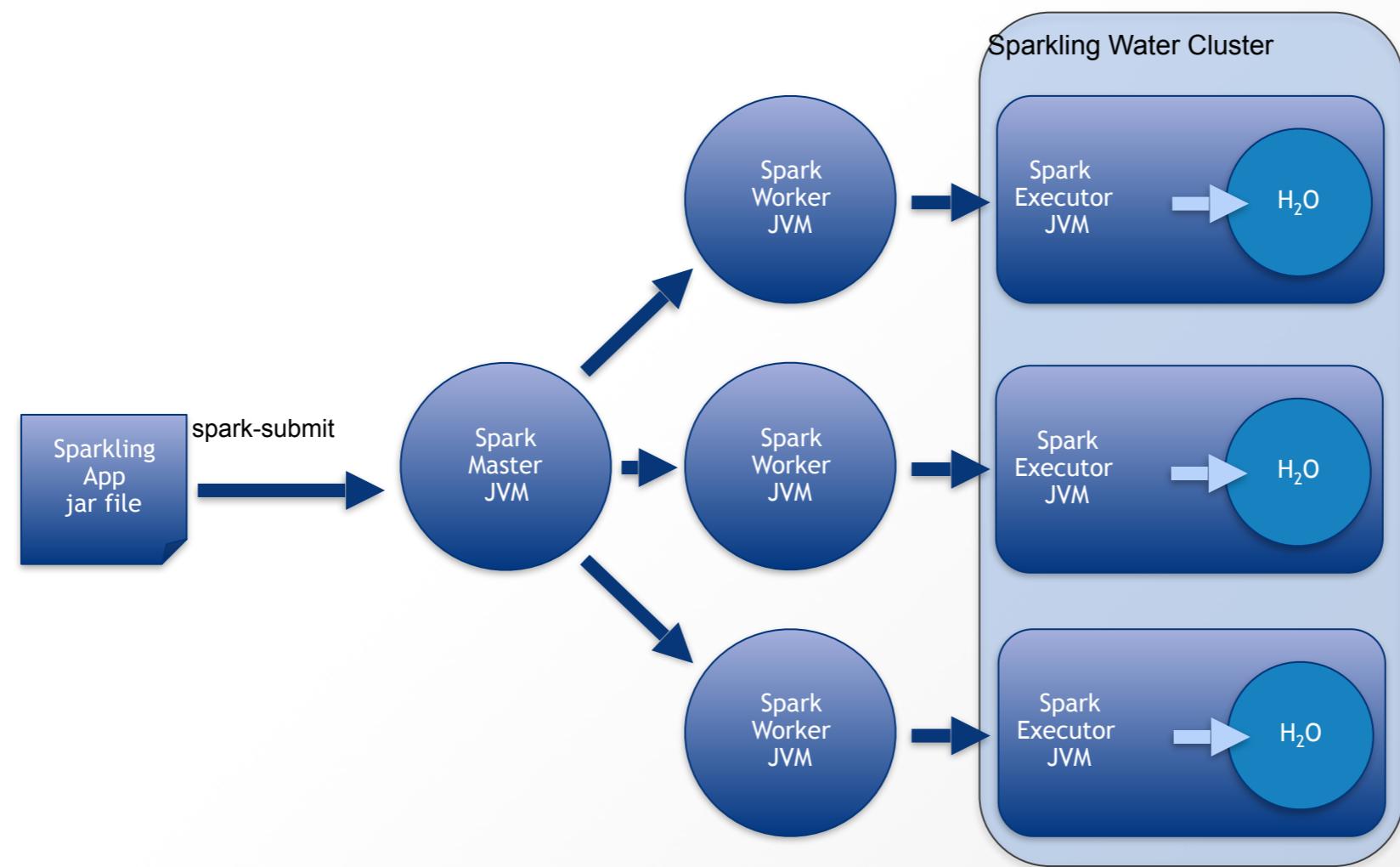
# **PySparkling made easy**

- PySparkling now in PyPi
- Contains all H2O and Sparkling Water dependencies, no need to worry about them
- Just add in on your Python path and that's it

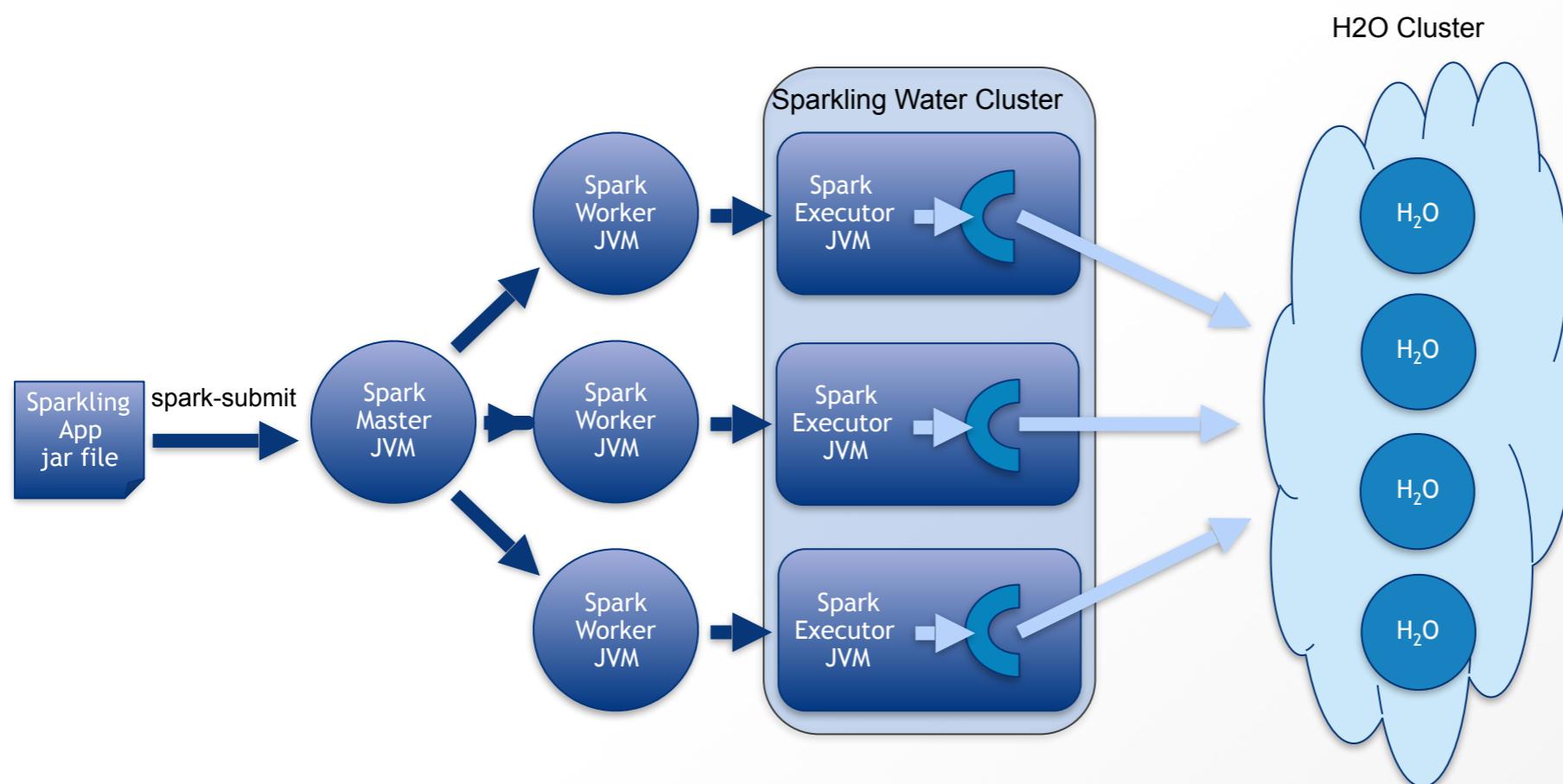
# **Sparkling Water Stable**

- New solution
- Not yet in master, in testing phase
- Sparkling Water is using external H2O cluster instead of starting H2O in each executor
- Spark executors can come and go and H2O won't be affected
- Start H2O cluster on YARN automatically

# Sparkling Water Internal Backend



# Sparkling Water External Backend



# RSparkling

- Built on top of sparklyr - R interface for Spark
- Ability to call both Spark and H2O methods
- Let's see a demo at the end!

## **And others!**

- Support for Datasets
- Zeppelin notebook support
- Integration with TensorFlow ( via DeepWater)
- Support for high-cardinality fast joins
- Bug fixes..

# **Coming features**

- Support for more MLlib algorithms in Flow
- Python cell in the H2O Flow
- Secure Communication - SSL
- Integration with Steam
- ...

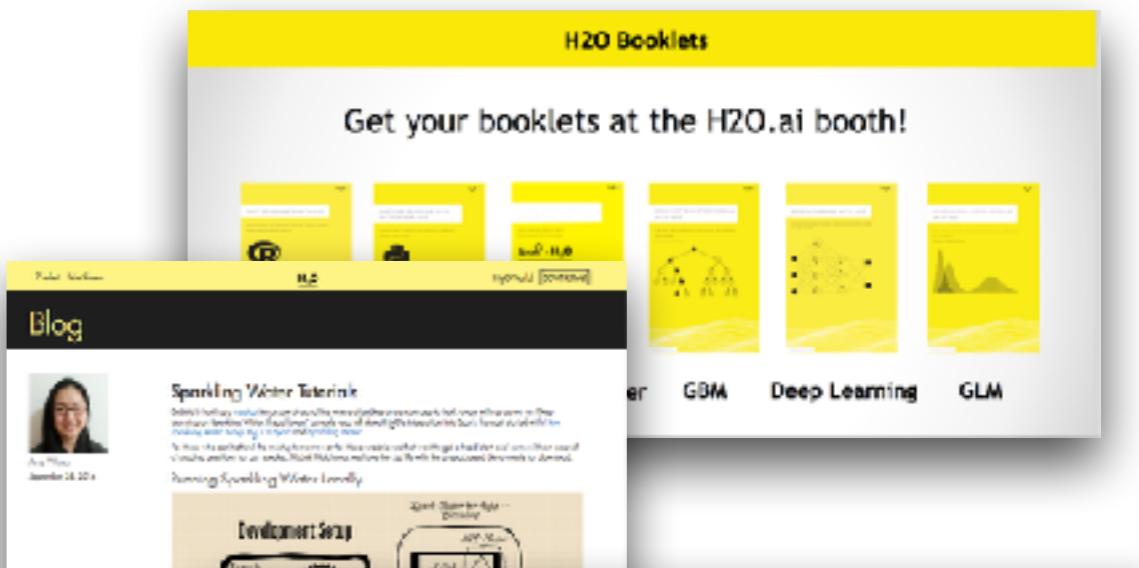
# Demo Time!

[https://github.com/h2oai/sparkling-water/  
blob/master/r/examples/nycflights13.R](https://github.com/h2oai/sparkling-water/blob/master/r/examples/nycflights13.R)

# More info

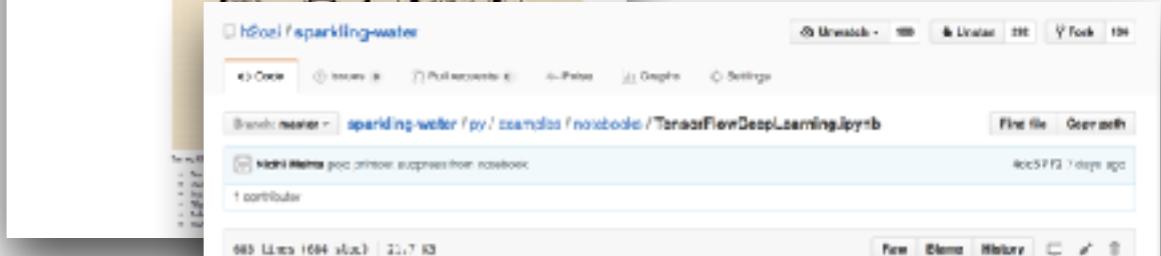
Checkout **H2O.ai** Training Books

<http://h2o.ai/resources>



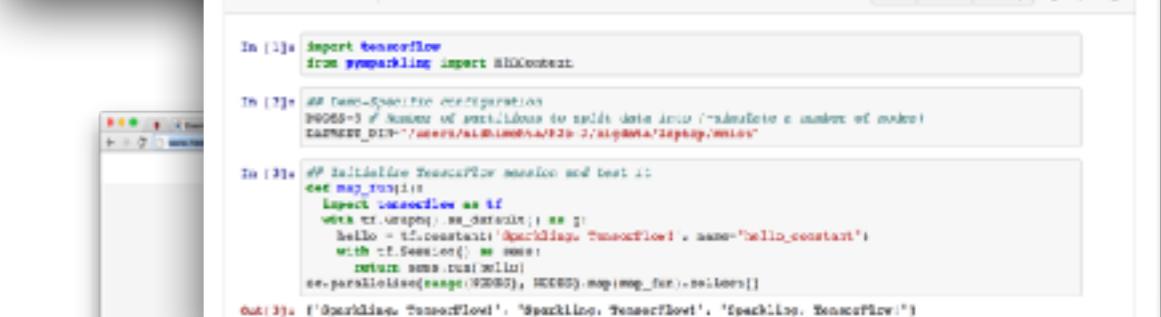
Checkout **H2O.ai** Blog

<http://h2o.ai/blog/>



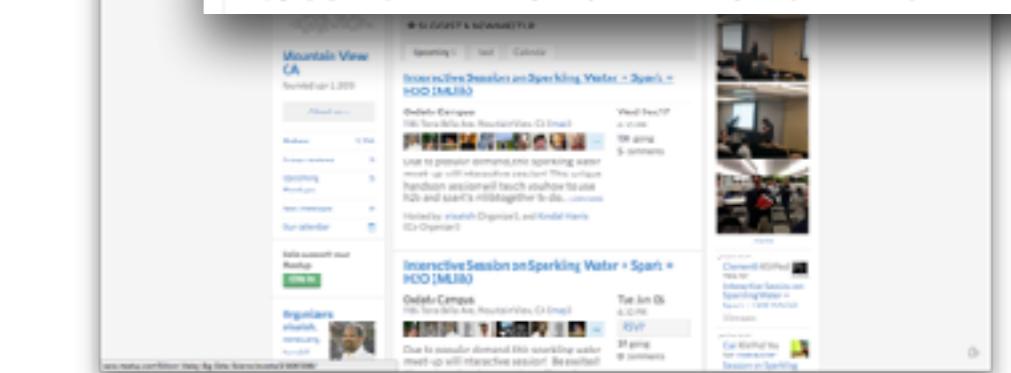
Checkout **H2O.ai** Youtube Channel

<https://www.youtube.com/user/0xdata>



Checkout GitHub

<https://github.com/h2oai/sparkling-water>



# Thank you!

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

Learn more at [h2o.ai](https://h2o.ai)

Follow us at [@h2oai](https://twitter.com/h2oai)

