

Jakub Háva
jakub@h2o.ai

Sparkling Water 2.0: The next generation of machine learning on Apache Spark

Spark Summit Europe, Brussels
October 26, 2016

Spark  + H₂O

**SPARKLING
WATER**

Who am I

- Finishing high-performance cluster monitoring tool for JVM based languages (instrumentation, JNI, JVMTI)
- Finishing Master's at Charles Uni in Prague
- Core engineer in Sparkling Water team in H2O.ai
- 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

H2O.ai

- Open Source AI Platform
- H2O, Steam, Sparkling Water, DeepWater
- Core algorithms written in high-perf Java
- Bindings for R/Python/Java/Scala/REST API
- Tries to make AI simple

**H2O+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

Spark 

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

H₂O.ai

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

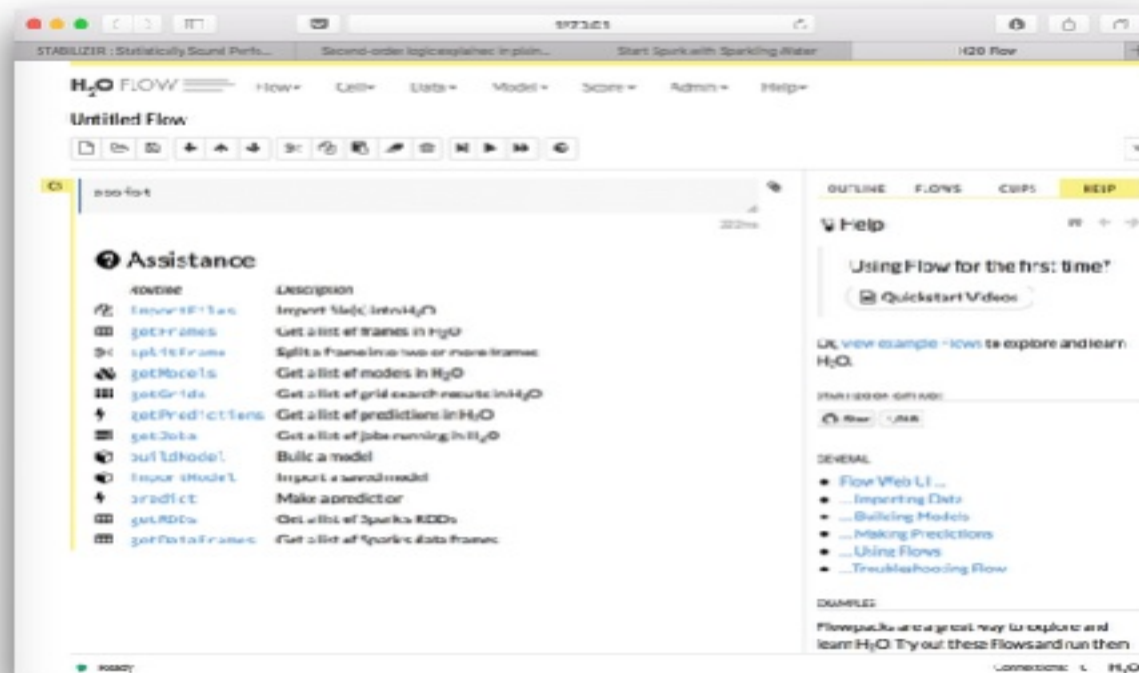
How to use Sparkling Water?

Start spark with Sparkling Water

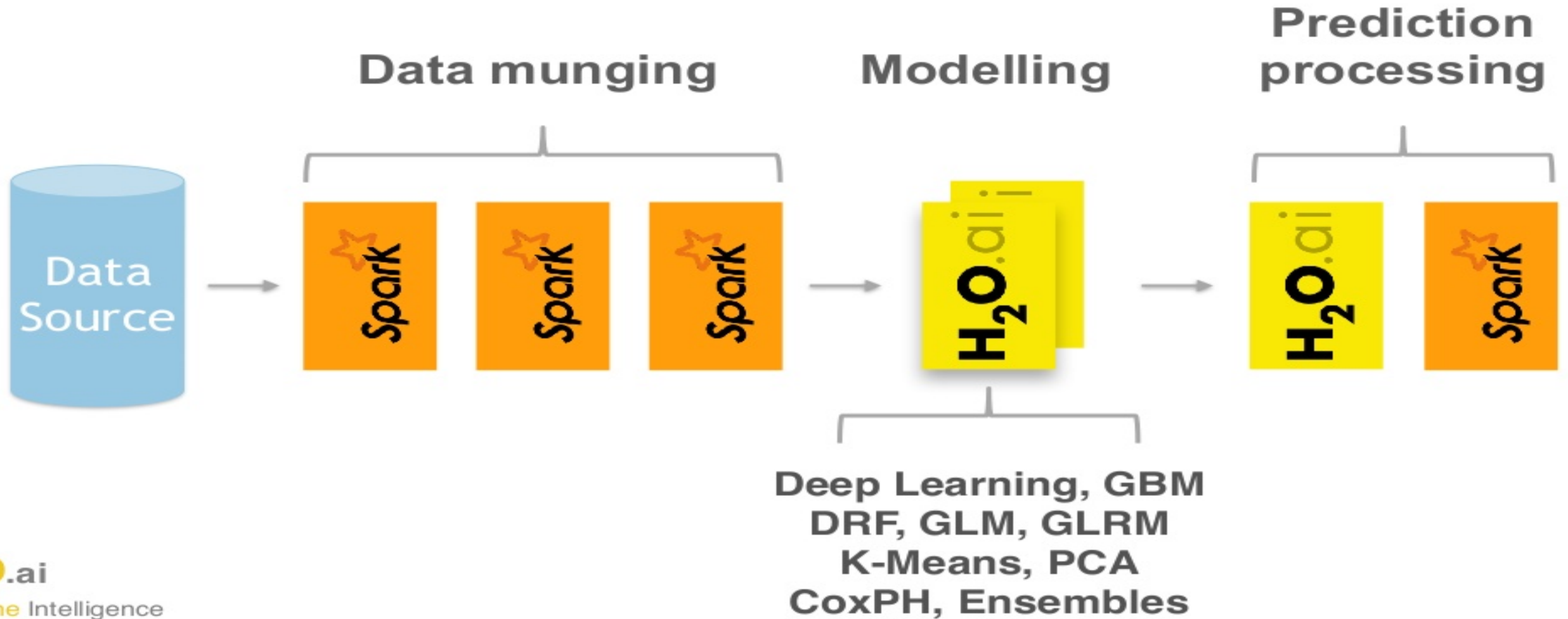
 start.sh

Raw

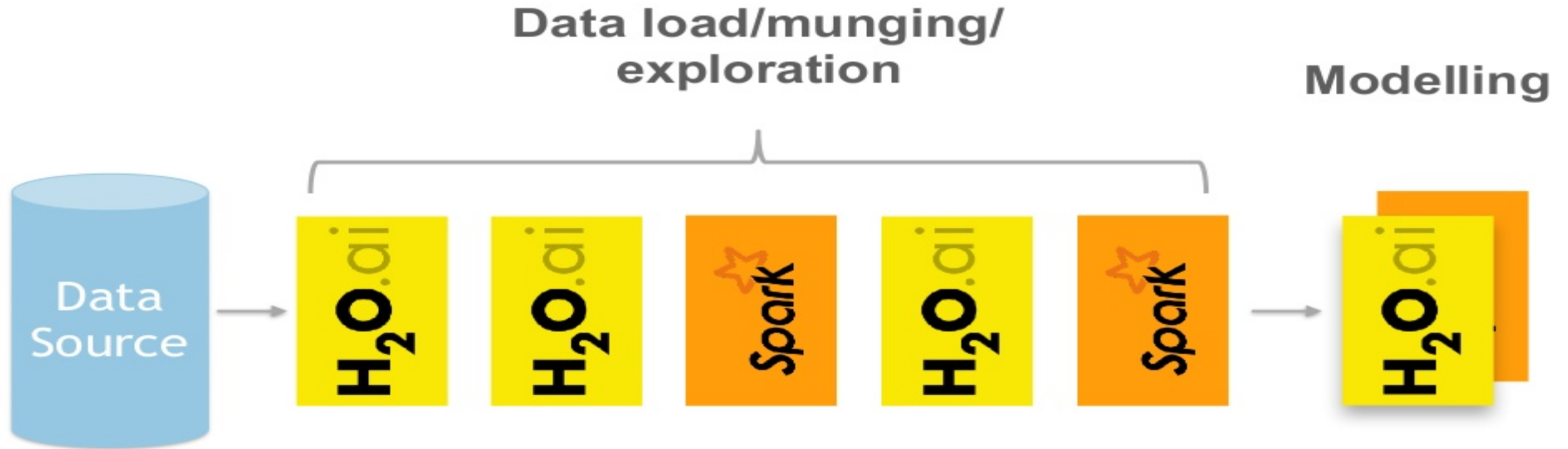
```
1 $SPARK_HOME/bin/spark-submit \  
2 --class water.SparklingWaterDriver \  
3 --packages ai.h2o:sparkling-water-examples_2.10:1.6.3 \  
4 --executor-memory=6g \  
5 --driver-class-path scalastyle.jar /dev/null
```



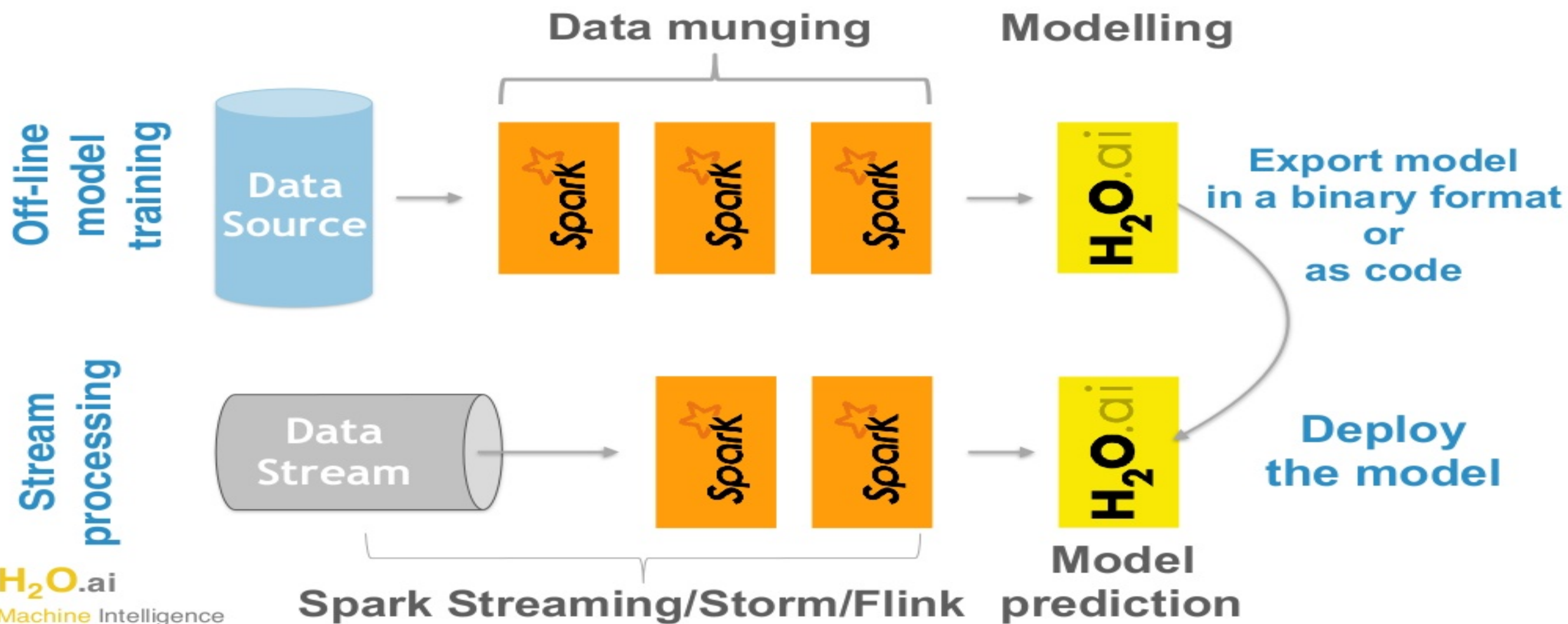
Model Building



Data Munging

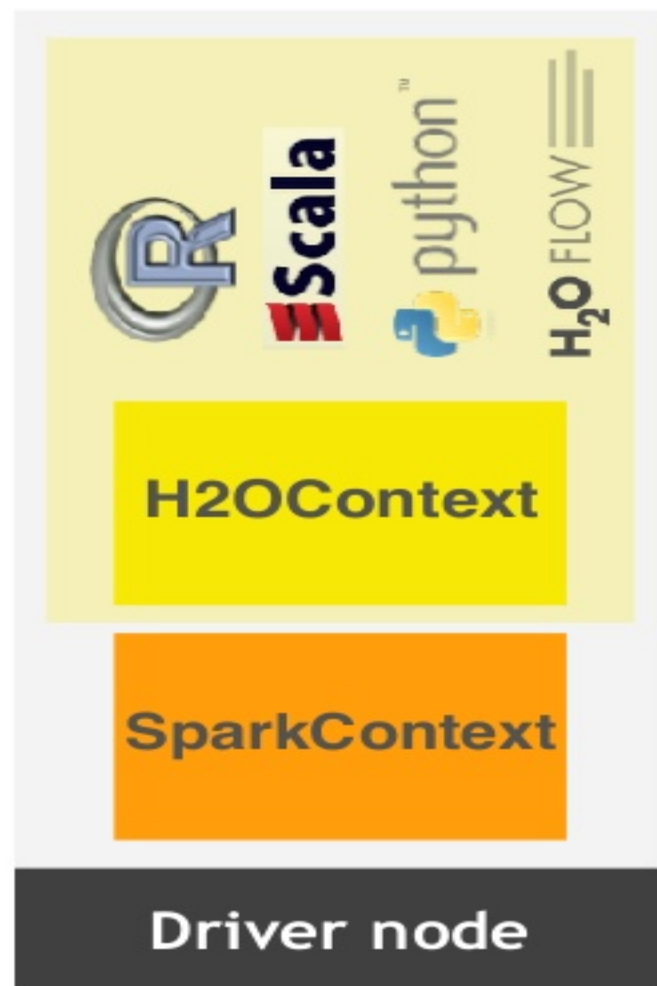


Stream processing

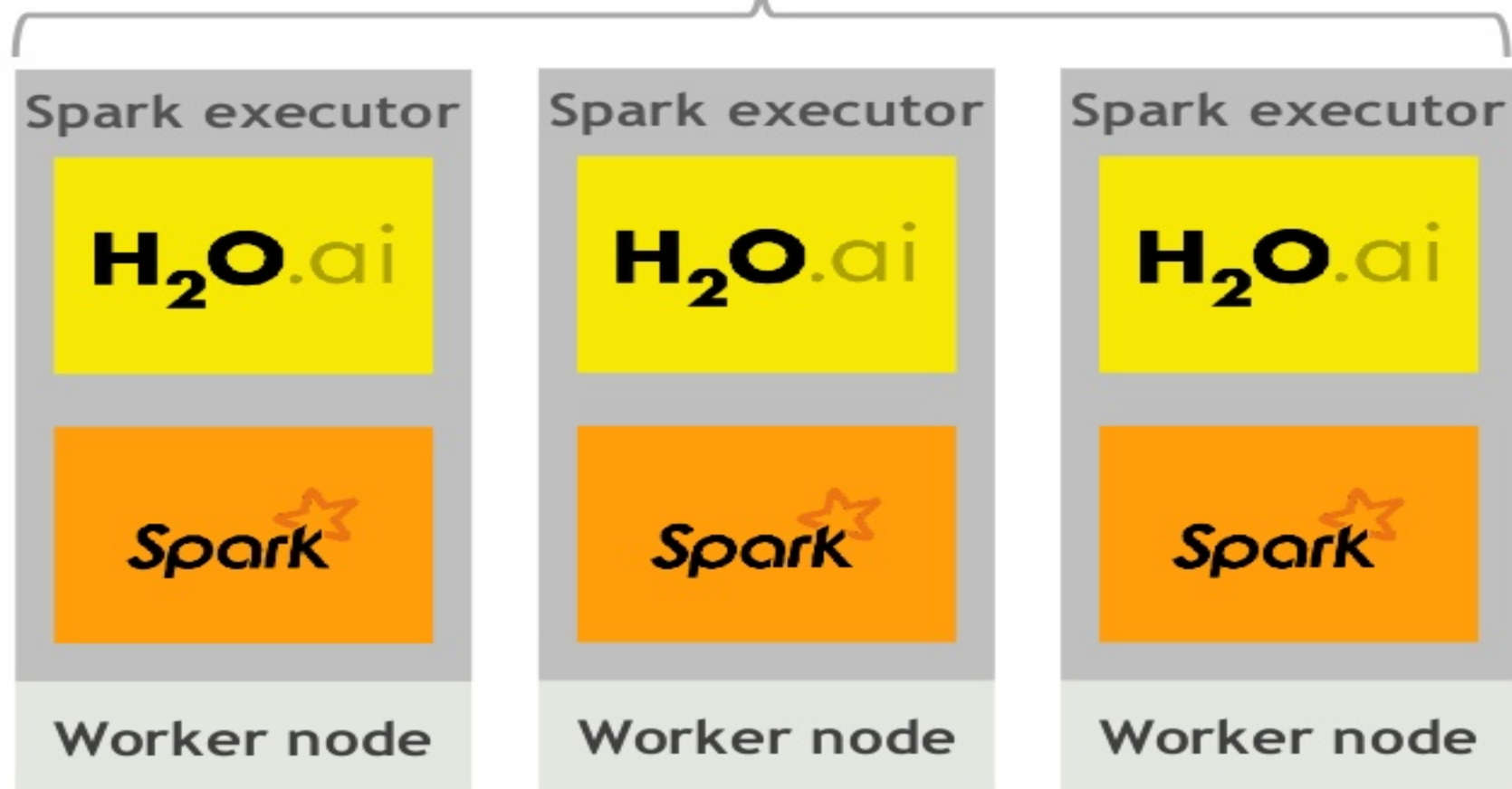


What is inside?

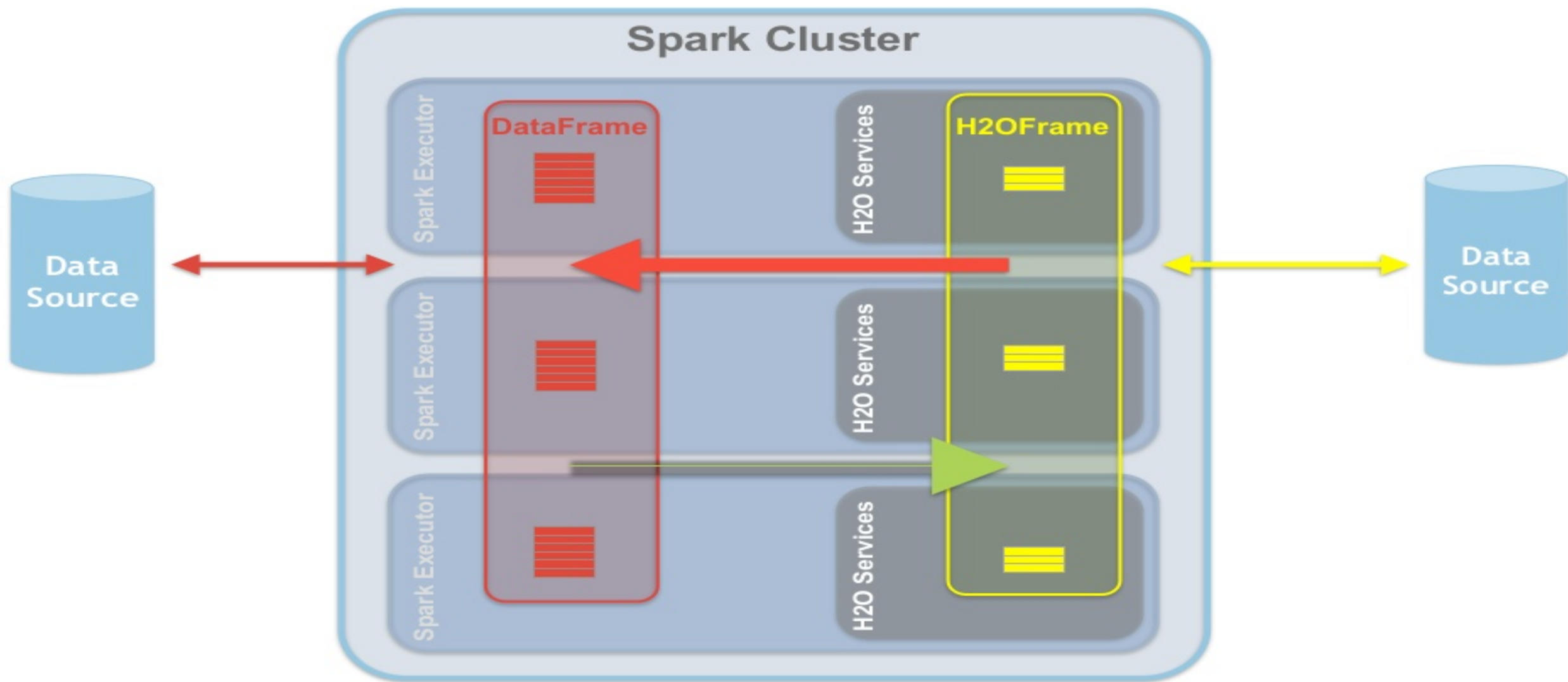
Scala/Py main
program



Spark + H₂O
SPARKLING
WATER



Cluster

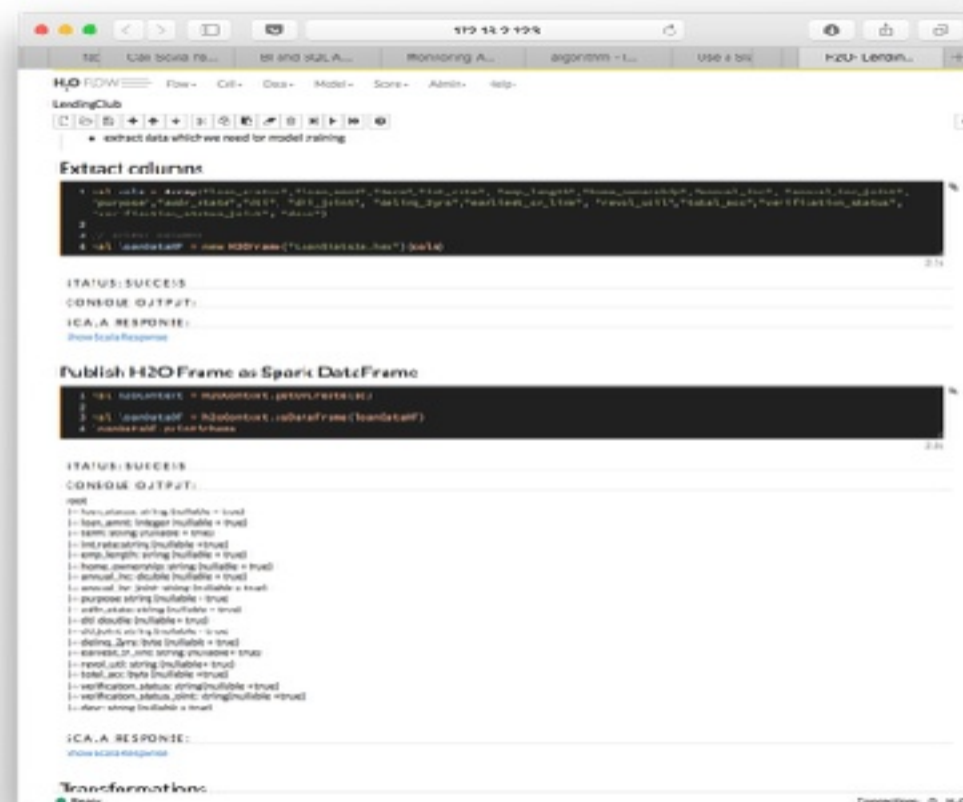


← **h2oContext.asDataFrame**
→ **h2oContext.asH2OFrame**

**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
- For example: Support Vector Machines (SVM)

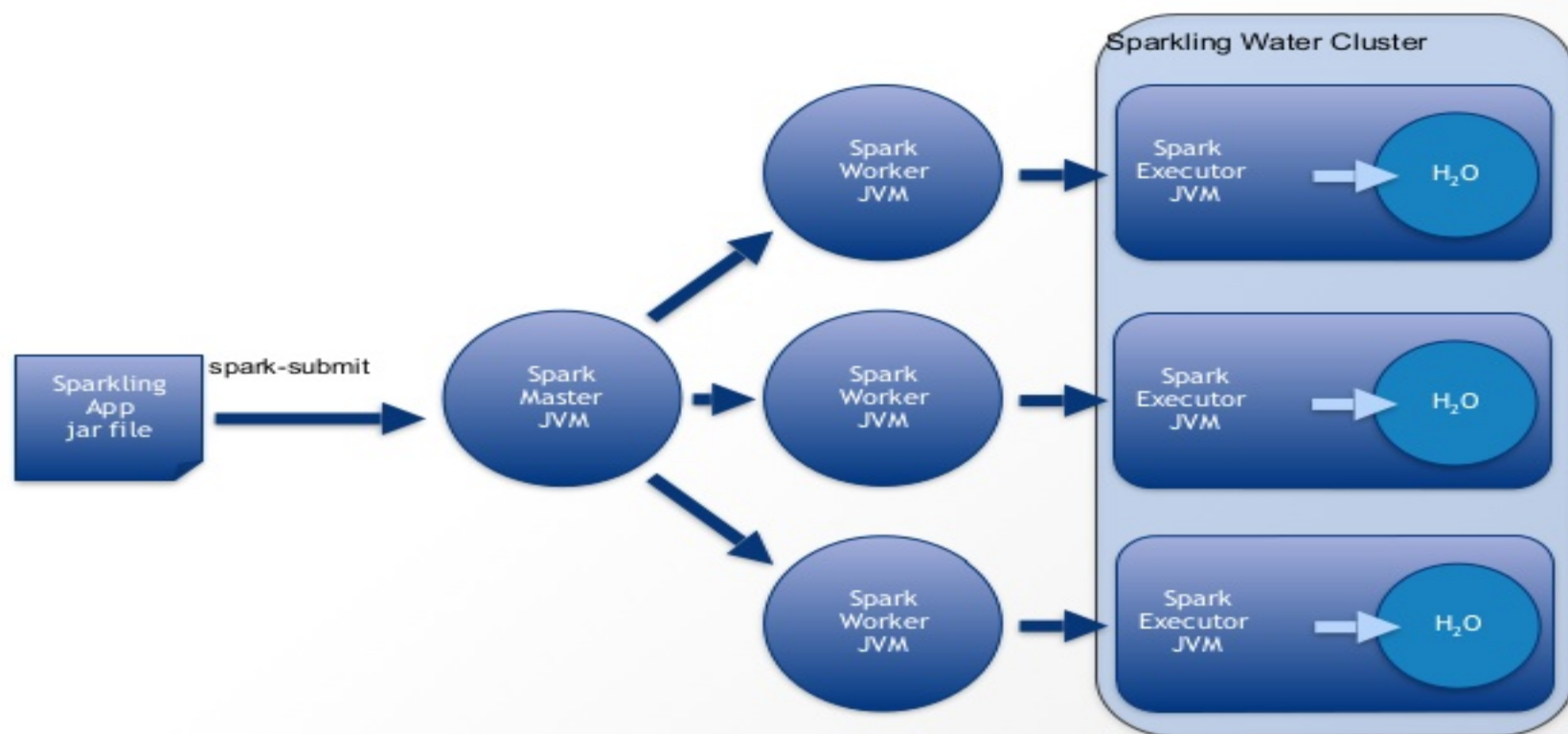
PySparkling made easy

- 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

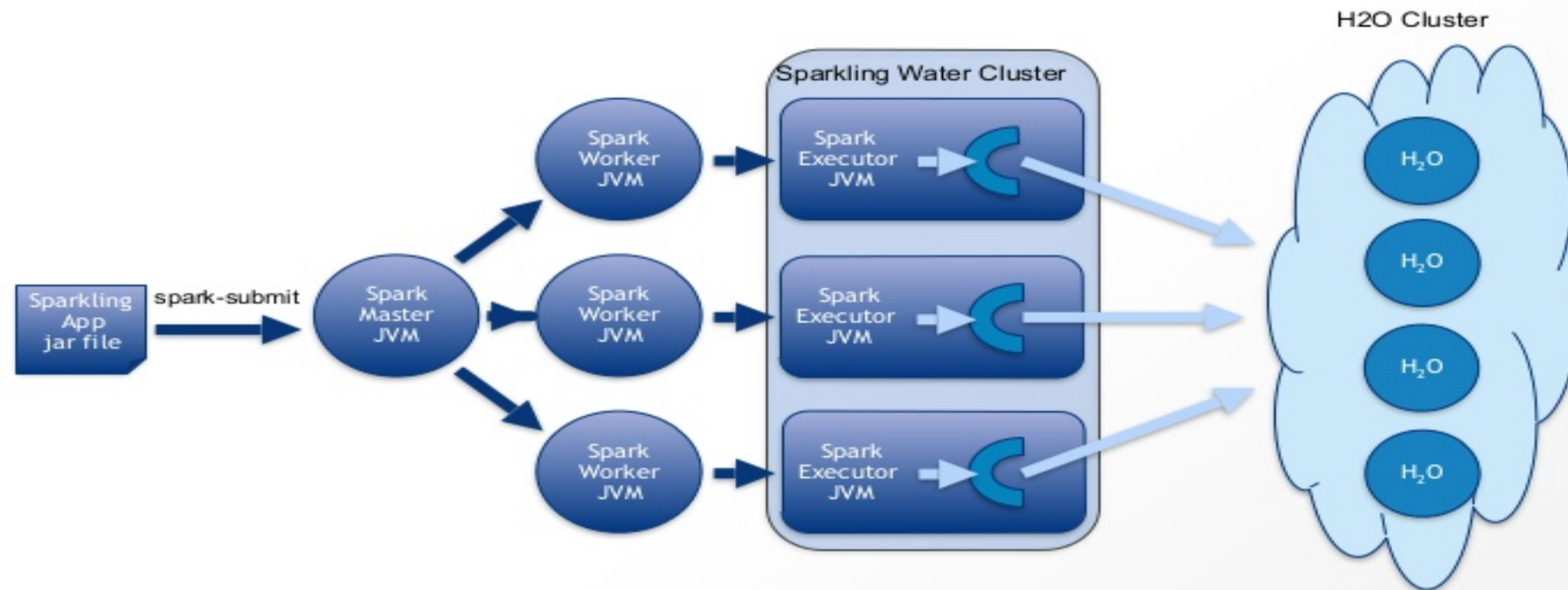
Sparkling Water high-availability

- New solution
- About to be integrated soon
- 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

Sparkling Water Internal Backend



Sparkling Water External Backend



And others!

- Support for Datasets
- RSparkling (Sparkling Water for R)
- Zeppelin notebook support
- Integration with TensorFlow, MXNet, Caffe (H2O DeepWater)
- Support for high cardinality (billions) joins
- A Lots of bug fixes..

Coming features

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

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

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

Write me at jakub@h2o.ai

