

Jakub Háva
jakub.hava@h2o.ai

The Next Generation of Machine Learning on Apache Spark

H2O.ai, Prague
Oct 16, 2018

Spark[★] + H₂O

SPARKLING
WATER

Who are we?

- **Kuba**
 - Senior Software engineer at H2O.ai - Code Owner - Sparkling Water
 - Master's at Charles University (CZ)
 - Implemented high-performance cluster monitoring tool for JVM based languages (JNI, JVMTI, instrumentation)
- **Michal**
 - VP of Engineering at H2O.ai
 - Author of Sparkling Water
 - Ph.D at Charles University (CZ), PostDoc at Purdue (US)

H₂O Products

H₂O.ai

In-Memory, Distributed
Machine Learning Algorithms
with H2O Flow GUI



H2O AI Open Source Engine
Integration with Spark

H₂O4GPU

Lightning Fast machine
learning on GPUs

DRIVERLESSAI

Automatic feature
engineering, machine learning
and interpretability

Steam

Secure multi-tenant H2O clusters

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
- Excels in existing Spark workflows requiring advanced Machine Learning algorithms
- Deployment tool for Driverless AI MOJOs

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

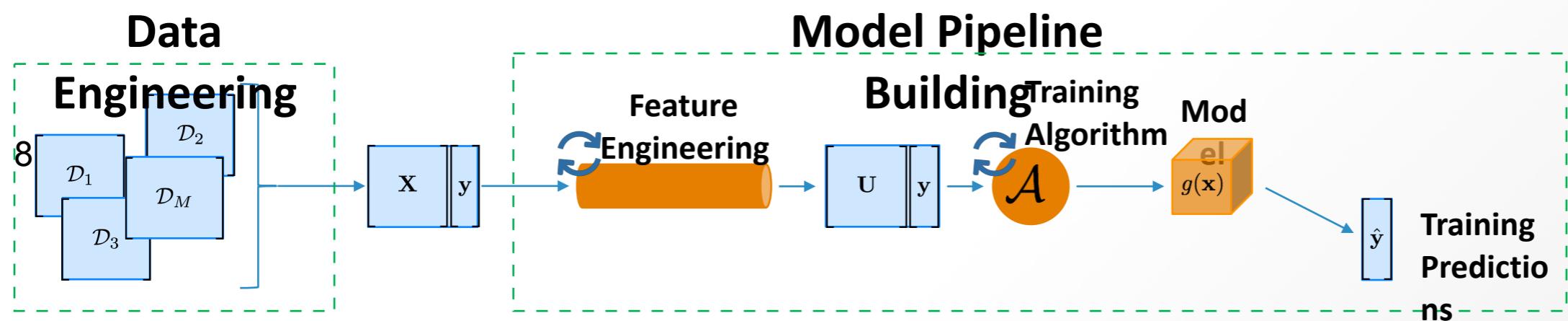
Benefits



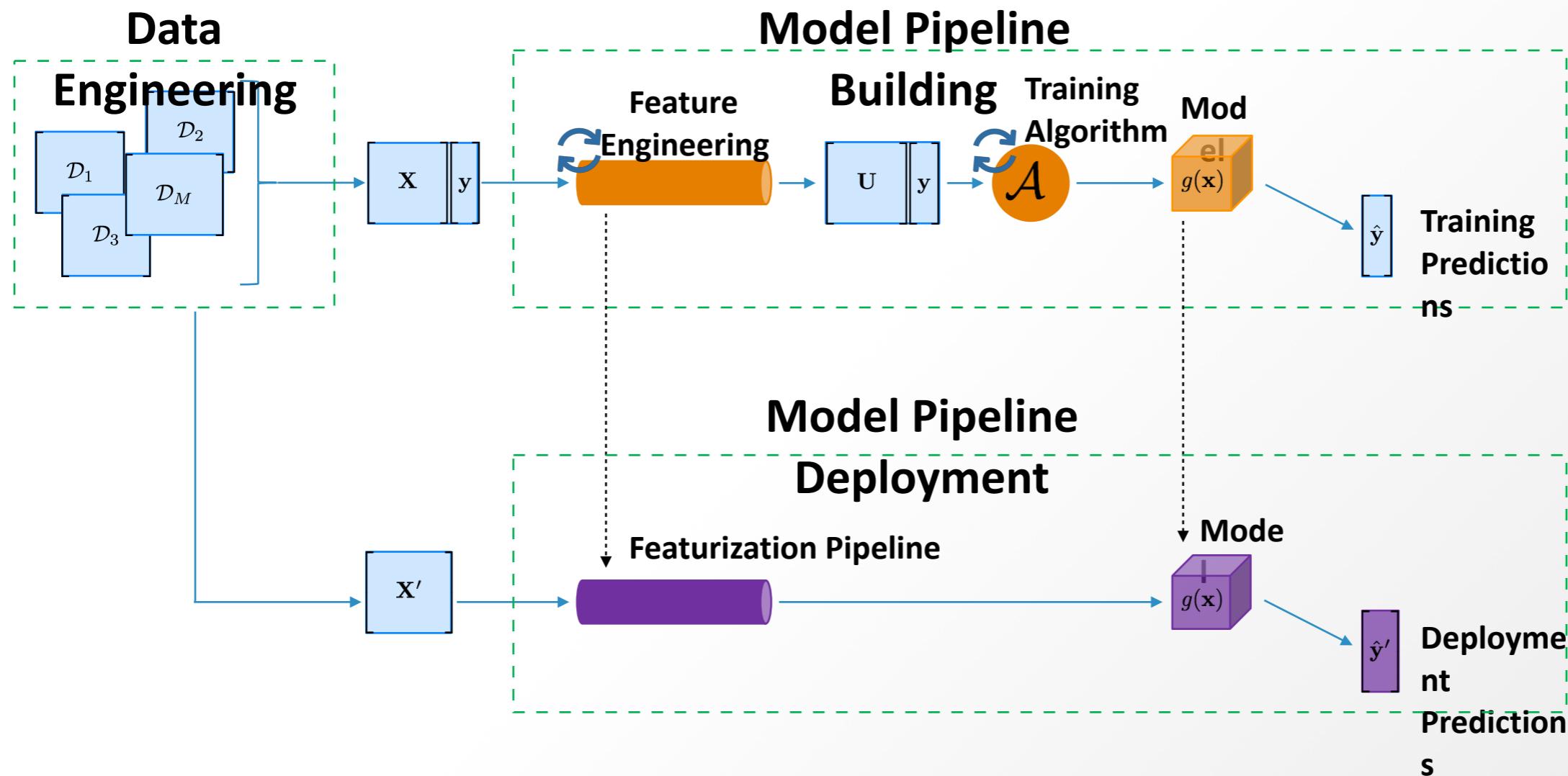
- Additional algorithms
 - NLP
- Powerful data munging
- ML Pipelines
- Advanced algorithms
 - speed v. accuracy
 - advanced parameters
- Fully distributed and parallelised
- Graphical environment
- R/Python interface

How it fits to
ML Life-
Cycle ?

Basic ML Lifecycle



Basic ML Lifecycle



#ML4SAIS

Example Implementation

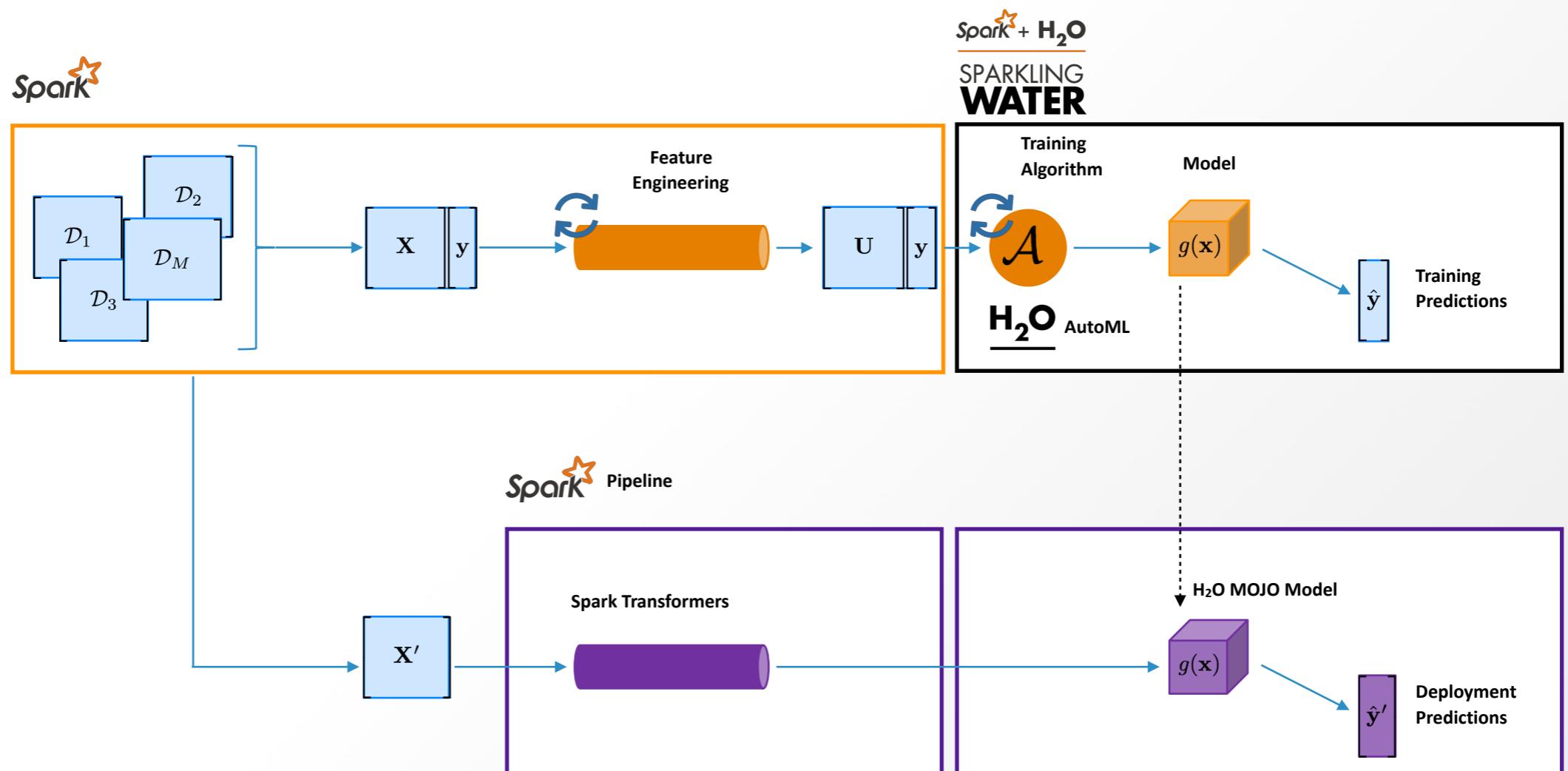
Model Building

Model Deployment

Data Engineering	Feature Engineering	Training Algorithm	Deployment Pipeline	Model
Spark	H2O	Spark	H2O MOJO	
Spark	H2O Driverless AI		Spark	H2O Driverless AI MOJO

#ML4SAIS

Basic ML Lifecycle



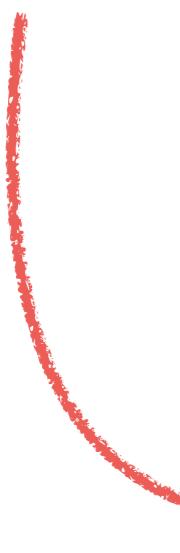
How to use Sparkling Water?

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

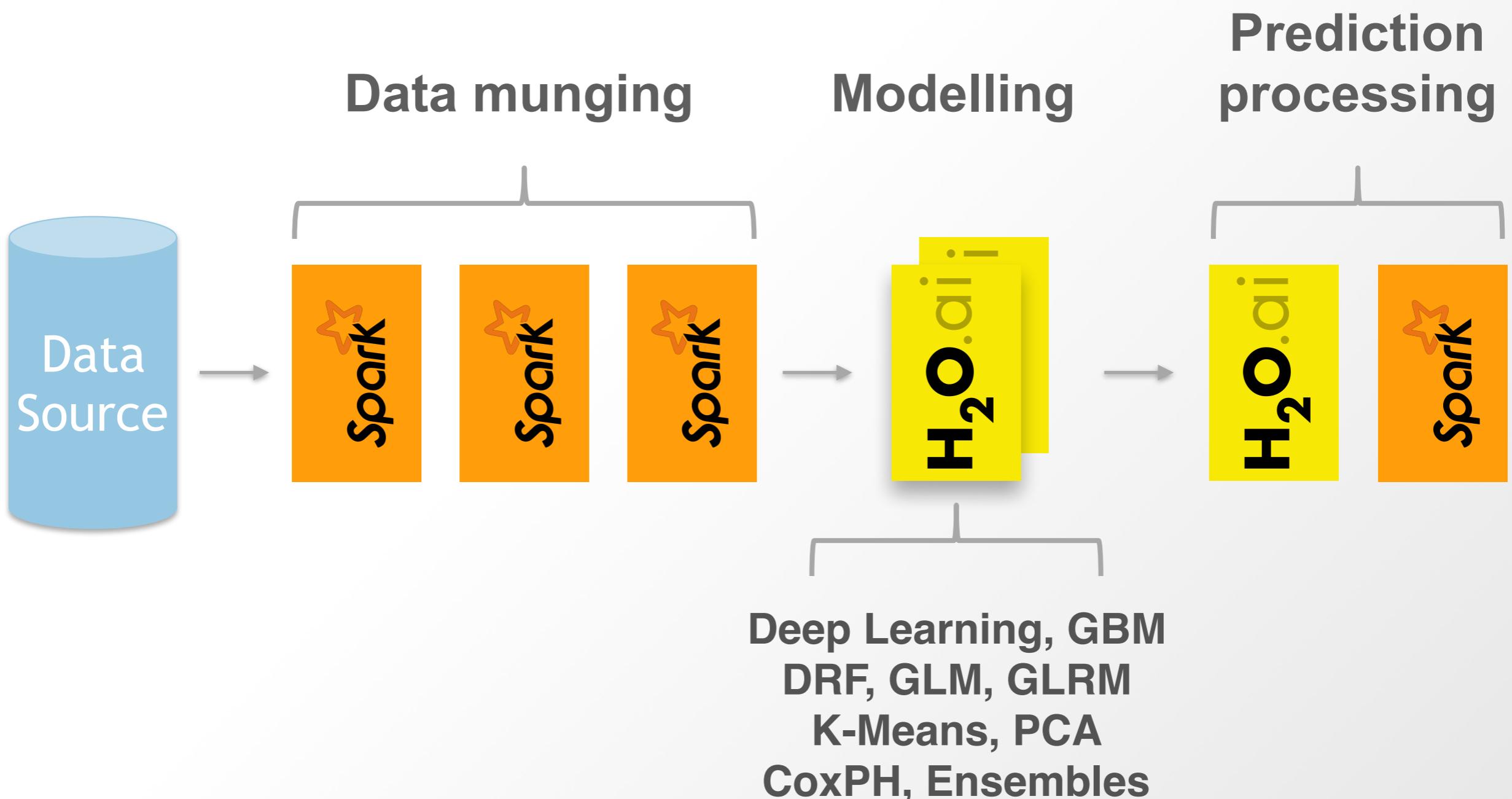


The screenshot shows the H2O Flow web application running at 127.0.0.1. The main title bar says "Start Spark with Sparkling Water". The interface has a toolbar with various icons for file operations like Open, Save, and Print. Below the toolbar is a navigation menu with tabs: OUTLINE, FLOWS, CLIPS, and HELP (which is currently selected). The left sidebar is titled "assit" and contains a section titled "Assistance" with a table of H2O routines and their descriptions:

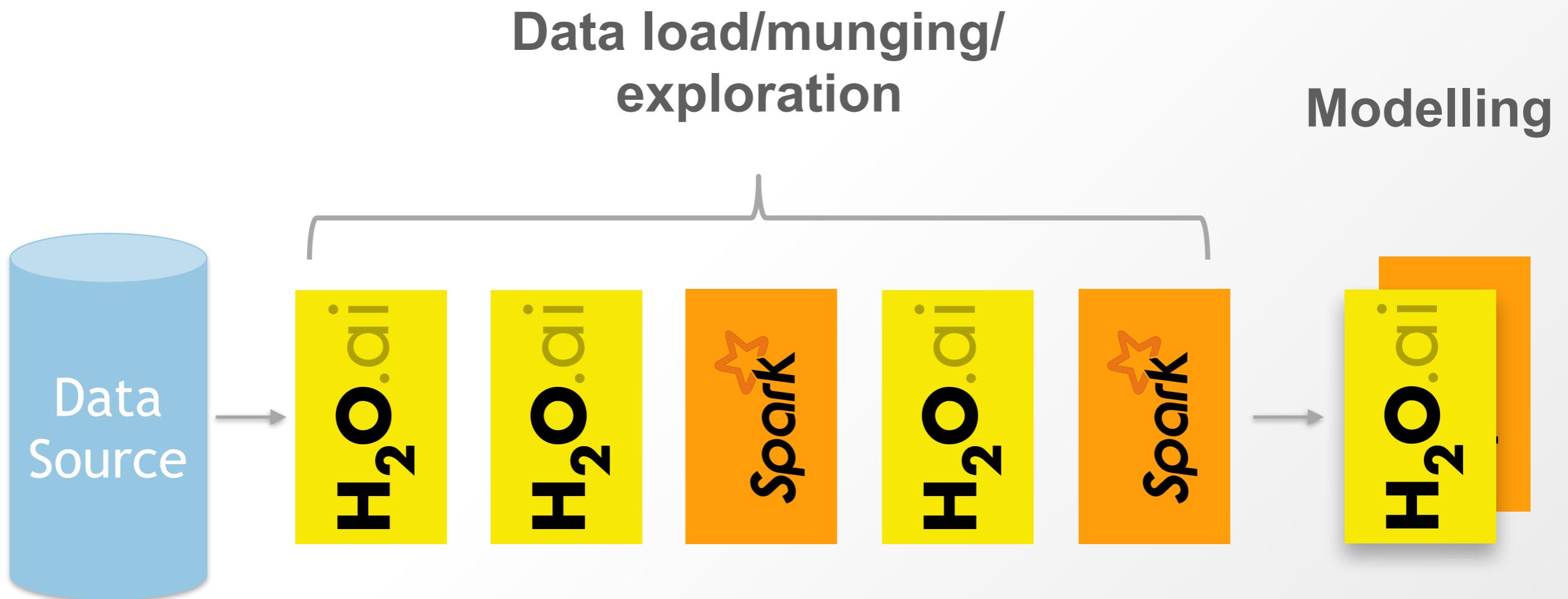
Routine	Description
importFiles	Import file(s) into H2O
getFrames	Get a list of frames in H2O
splitFrame	Split a frame into two or more frames
getModels	Get a list of models in H2O
getGrids	Get a list of grid search results in H2O
getPredictions	Get a list of predictions in H2O
getJobs	Get a list of jobs running in H2O
buildModel	Build a model
importModel	Import a saved model
predict	Make a prediction
getRDDs	Get a list of Spark's RDDs
getDataFrames	Get a list of Spark's data frames

The right side of the interface has sections for "Help", "Using Flow for the first time?", "Quickstart Videos", "view example Flows", "STAR H2O ON GITHUB!", and "GENERAL" and "EXAMPLES" sections with links to various H2O documentation pages. At the bottom, there are status indicators: "Ready" and "Connections: 0".

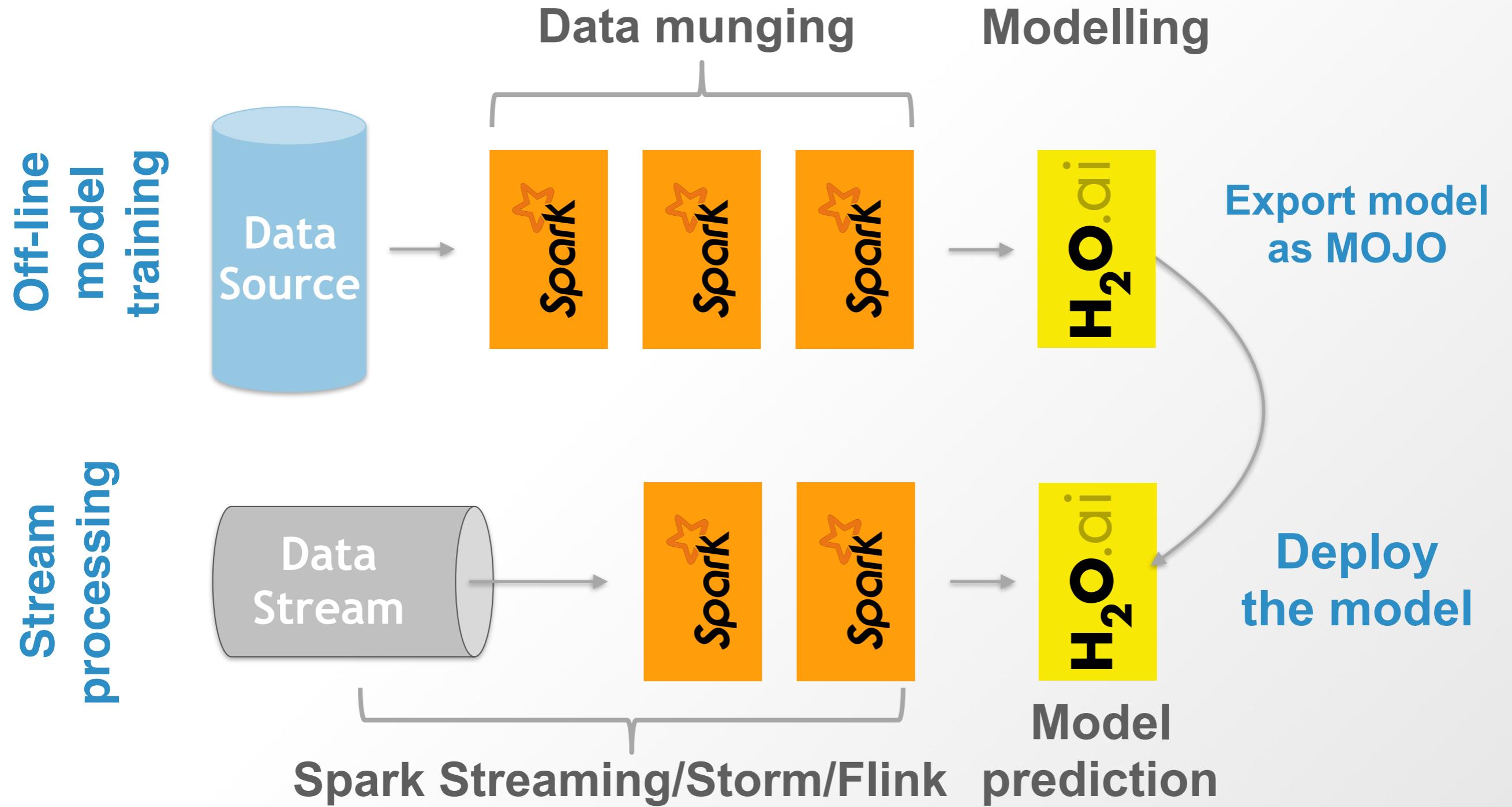
Model Building



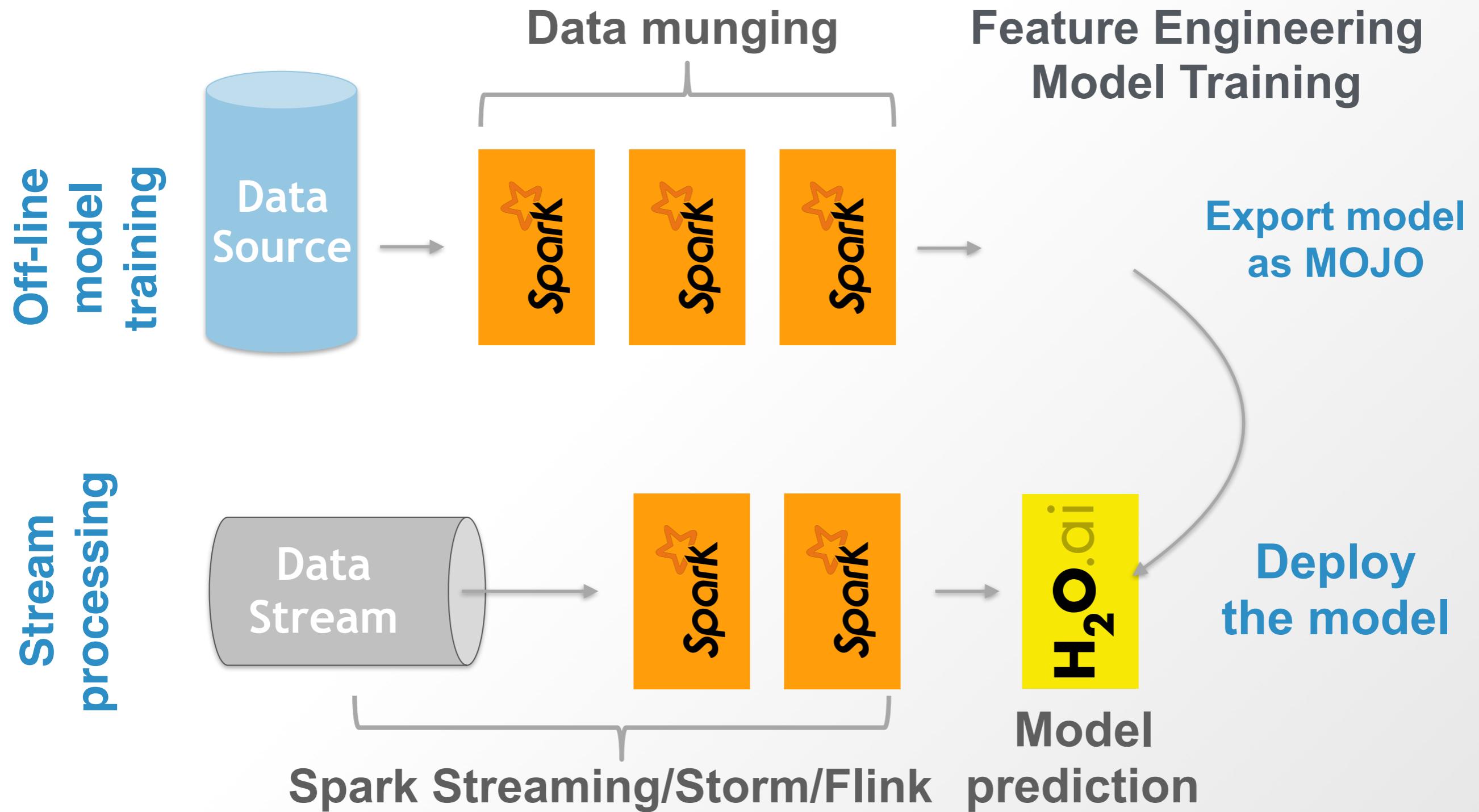
Data Munging



Stream Processing



Stream Processing 2



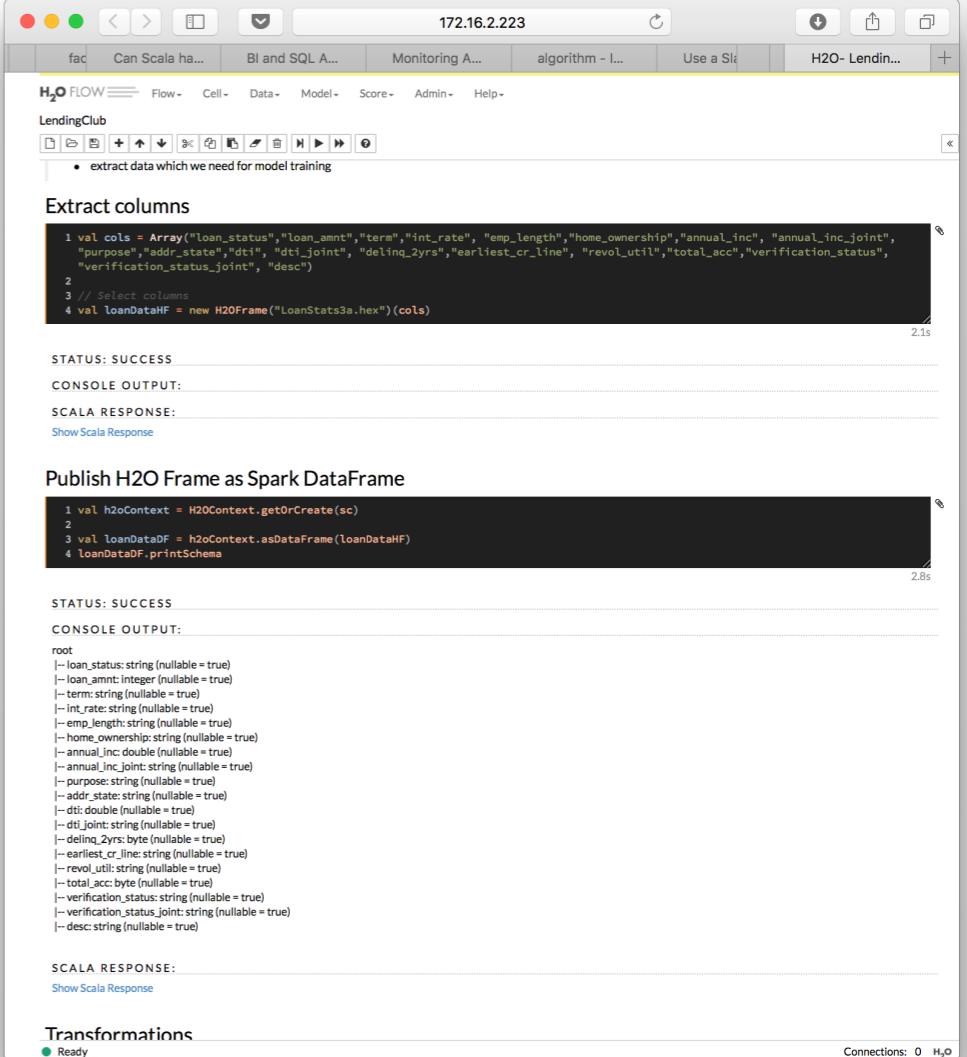
Scoring

- POJO
 - Plain Old Java Object
- MOJO
 - Model Object Optimised
- MOJO Pipeline
 - MOJO from DAI with additional future transformations
- No runtime dependency on H2O/Driverless AI frameworks

Features Overview

Scala code in H2O Flow

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



The screenshot shows the H2O Flow interface running on a Mac OS X system. The window title is "172.16.2.223". The main area displays two code snippets in a Scala code editor.

Extract columns:

```
1 val cols = Array("loan_status","loan_amnt","term","int_rate", "emp_length","home_ownership","annual_inc", "annual_inc_joint",
2 "purpose","addr_state","dti","dti_joint", "delinq_2yrs","earliest_cr_line", "revol_util","total_acc","verification_status",
3 // Select columns
4 val loanDataHF = new H2OFrame("LoanStats3a.hex")(cols)
```

PUBLISH H2O FRAME AS SPARK DATAFRAME:

```
1 val h2oContext = H2OContext.getOrCreate(sc)
2
3 val loanDataDF = h2oContext.asDataFrame(loanDataHF)
4 loanDataDF.printSchema
```

Both code blocks result in a **STATUS: SUCCESS** message. The second code block also shows the **SCALA RESPONSE** which lists all the columns and their types for the DataFrame.

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
- Integration with Mojo pipelines

PySparkling

- PySparkling is on 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
- Correctly specifies dependency on PySpark

RSparkling

- Sparkling Water for R
- Based on SparklyR package
- Independent on Spark and Sparkling Water version

And others!

- Support for Datasets
- Zeppelin notebook support
- XGBoost Support in Distributed Mode
- Support for high-cardinality fast joins
- Secure Communication - SSL
- Support for Sparse Data conversions
- External Cluster Stabilisation
- Enterprise security support - LDAP, Kerberos
- Driverless AI Pipeline deployment availability

Coming features

- Integration with Steam
 - Deploy on remises
 - Deploy to Cloud

More Info

- 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](https://twitter.com/h2oai)

PS: We are hiring!



www:says-it.com/unclesam/