Jakub Háva
jakub.hava@h2o.ai

The Next Generation of Machine Learning on Apache Spark

H2O.ai, Prague
Oct 16, 2018

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)

H_2O Products

H_2O.ai

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

Spark + H_2O SPARKLING WATER

H2O AI Open Source Engine Integration with Spark

H_2O4GPU

Lightning Fast machine learning on GPUs

DRIVERLESSAI

Automatic feature engineering, machine learning and interpretability

Steam

Secure multi-tenant H2O clusters

H2O+Spark = Sparkling Water

H_2O.ai

Machine Intelligence

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

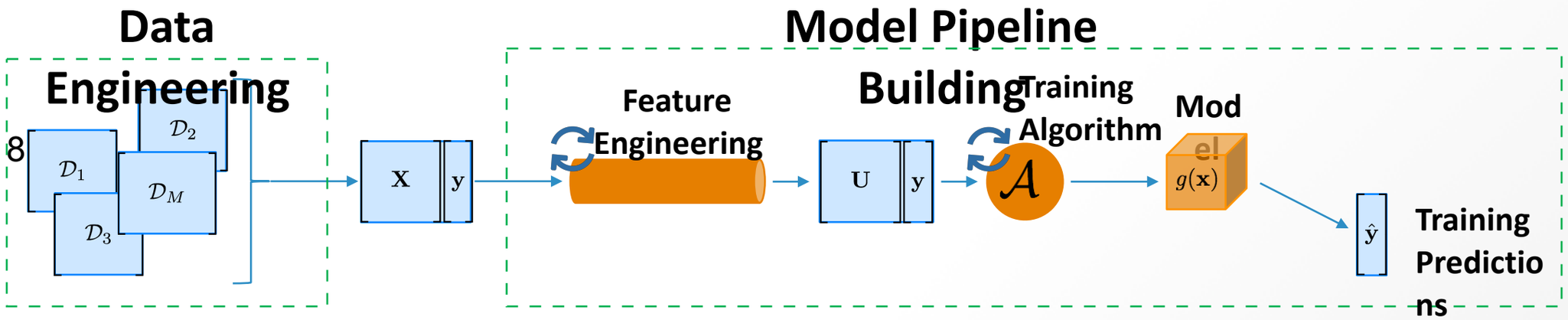
Benefits

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

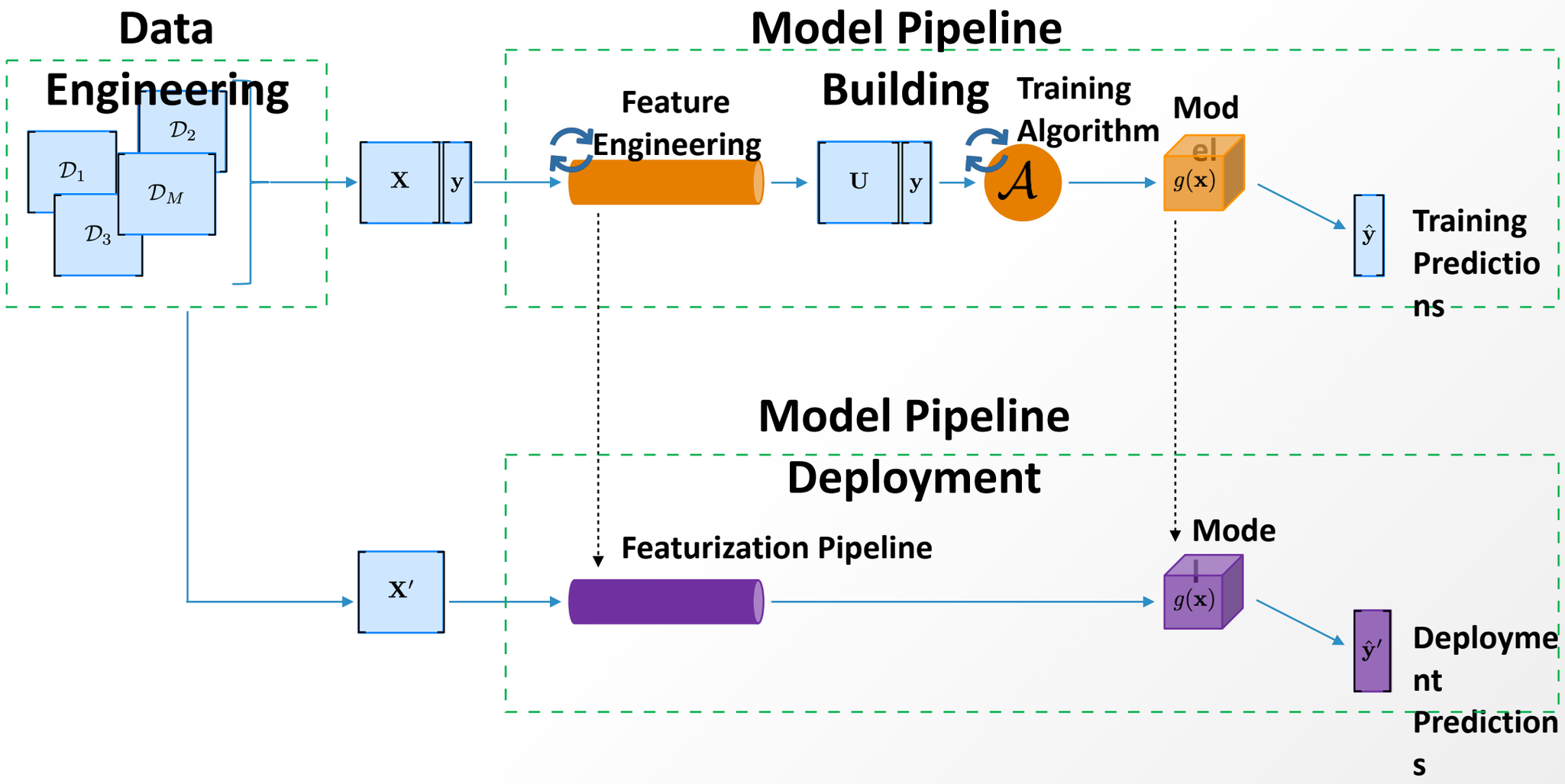
- Advanced algorithms
  - speed v. accuracy
  - advanced parameters
- Fully distributed and parallelised
- Graphical environment
- R/Python interface

How it fits to ML Life-Cycle ?

Basic ML Lifecycle



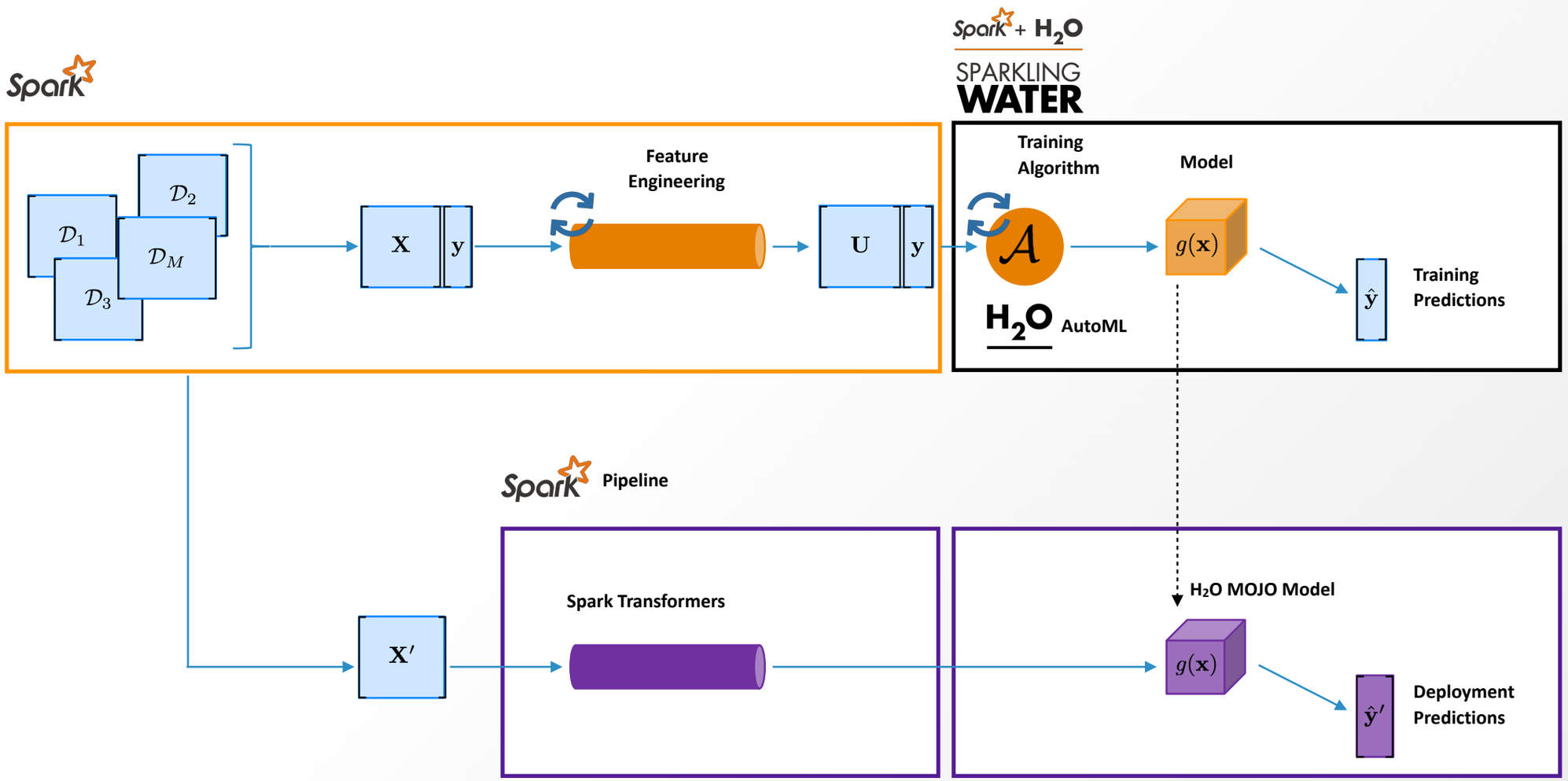
Basic ML Lifecycle



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

Basic ML Lifecycle



How to use Sparkling Water?

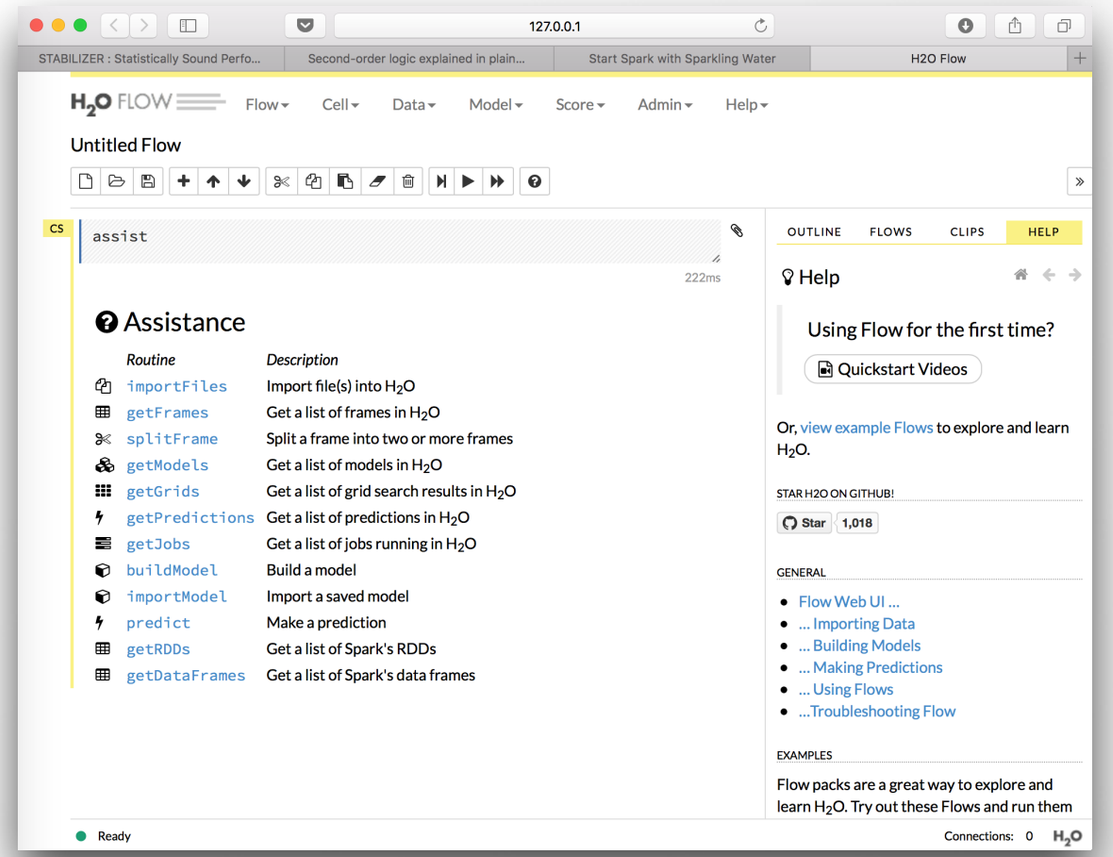
H_2O.ai

Machine Intelligence

Start spark with Sparkling Water

start.sh — Raw

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

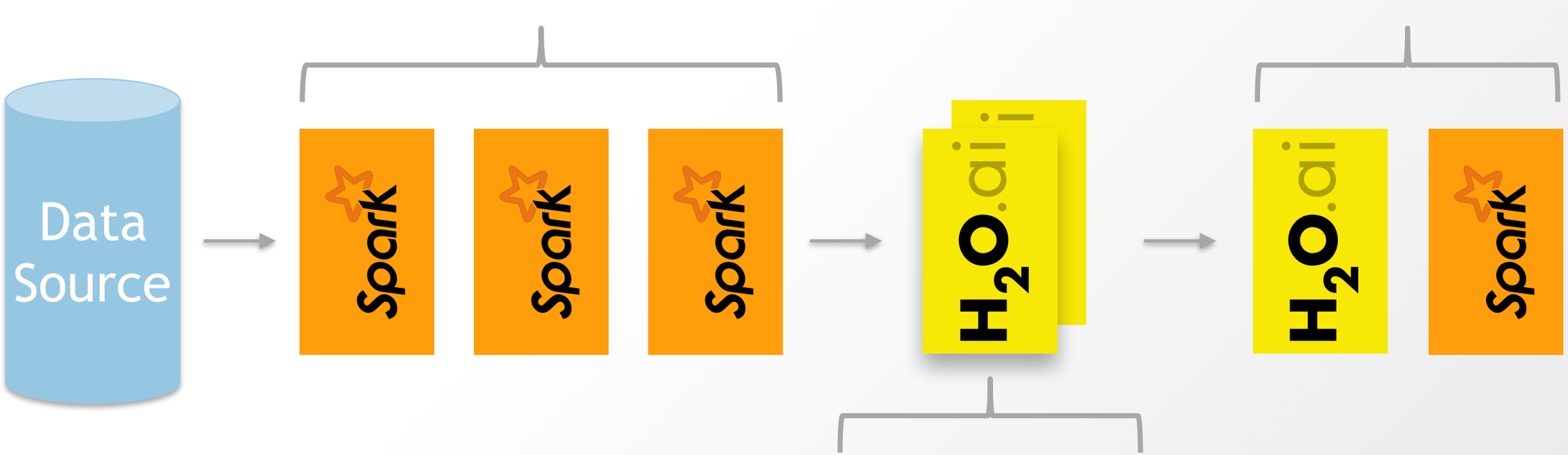


Model Building

Data munging

Modelling

Prediction processing

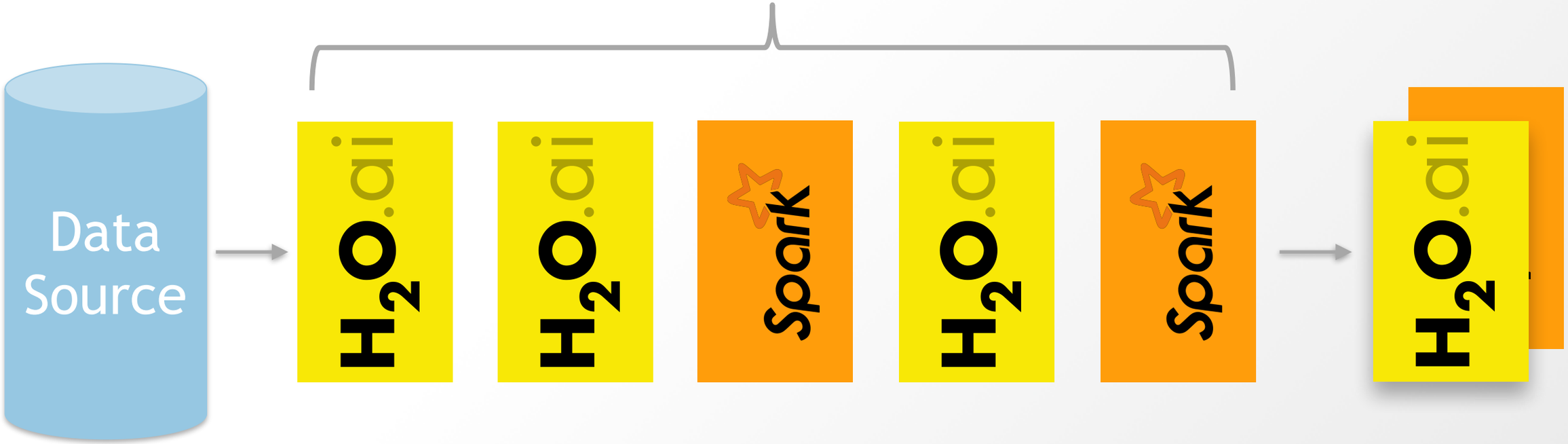


Deep Learning, GBM
DRF, GLM, GLRM
K-Means, PCA
CoxPH, Ensembles

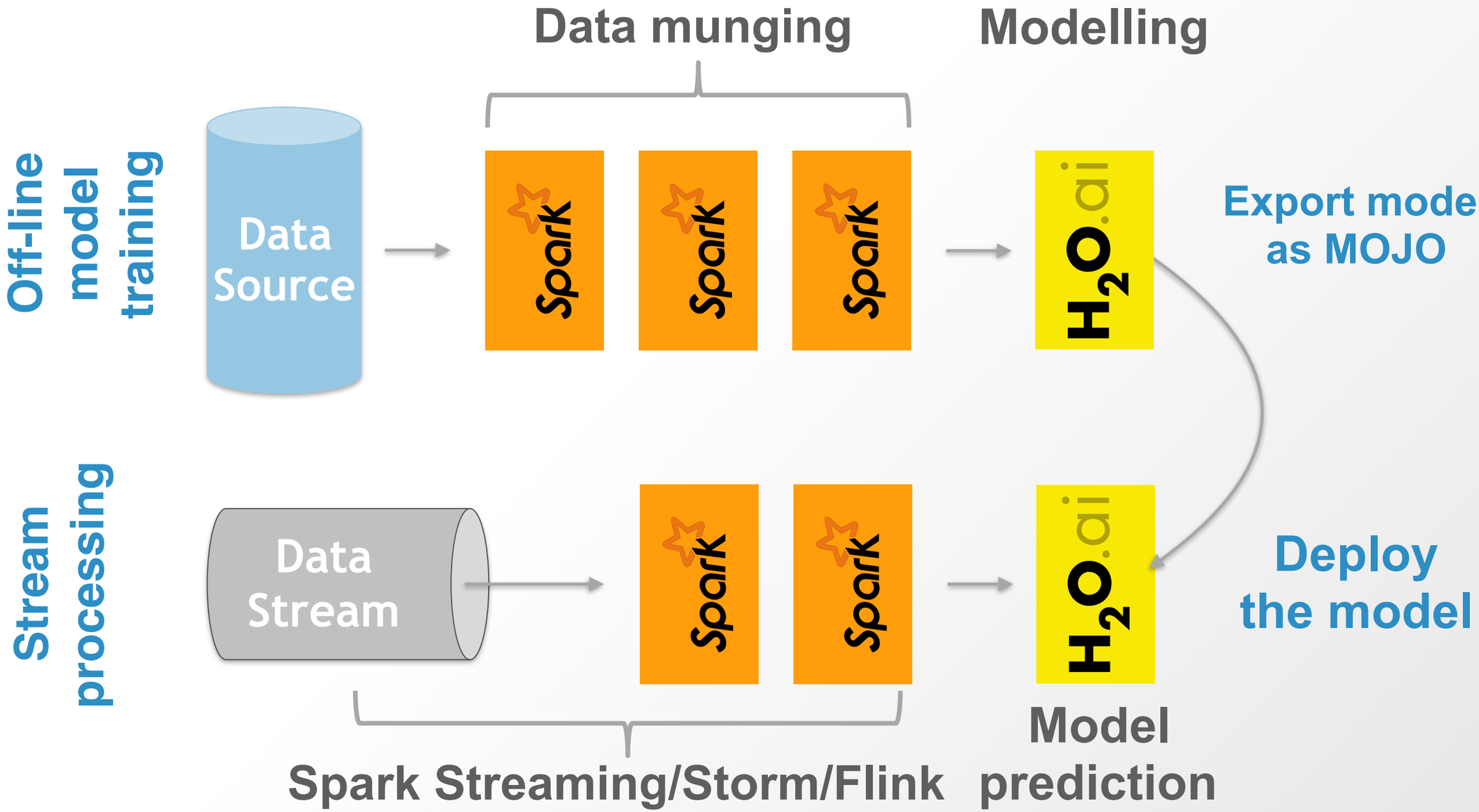
Data Munging

Data load/munging/
exploration

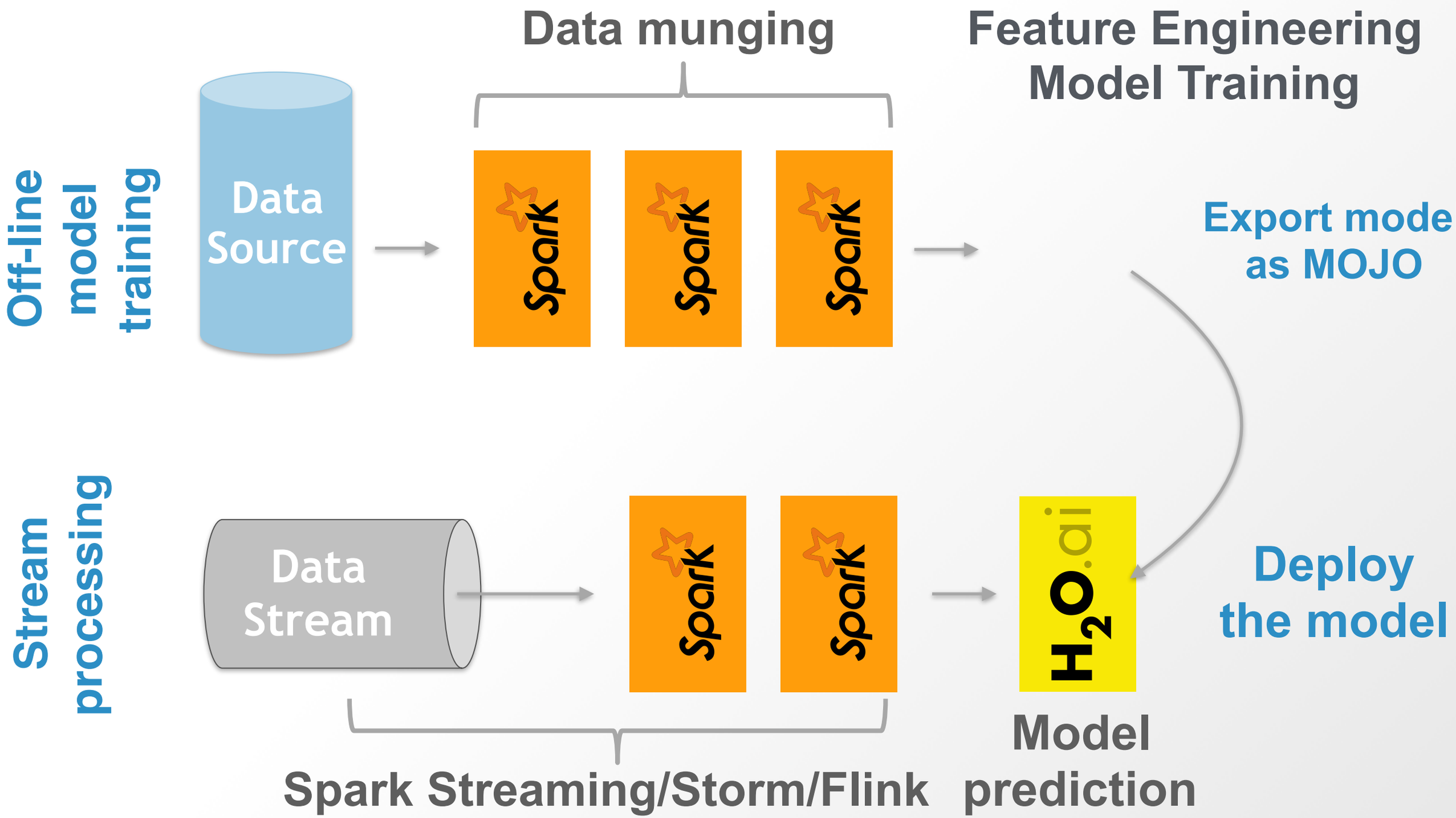
Modelling



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

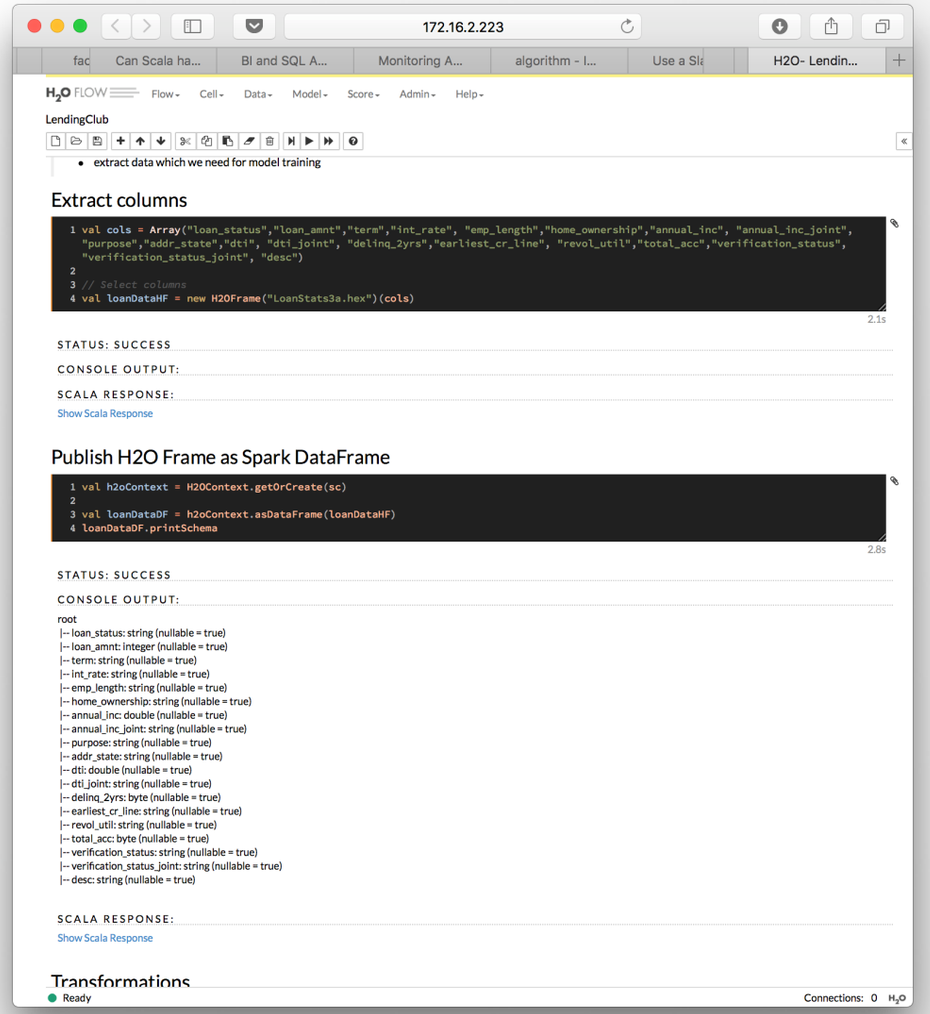
Overview

H_2O.ai

Machine Intelligence

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
- 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

H2O.ai

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
PS: We are hiring!