H2O-3

Pavel Pscheidl

H2O.ai

Content

- What is H2O-3
  - What is H2O-3
  - Development process
  - Architecture
- Using H2O-3
  - H2O-3 interfaces
  - Quick algorithm overview
  - Live demo
- Where to go next
  - Useful resources & online tutorials to follow

What is H2O-3

- Lots of great machine-learning algorithms

Word2Vec

CoxPH DRF Deep Learning

GBM Aggregator GLRM

K-Means GLM PCA

XGBoost Stacking Naive Bayes

What is H2O-3

- Open-source
- Hackable
- Rapid development

- Scalable
- Convenient
- Fast

github.com/h2oai

Based on JVM, version >= 1.7

Well documented: docs.h2o.ai

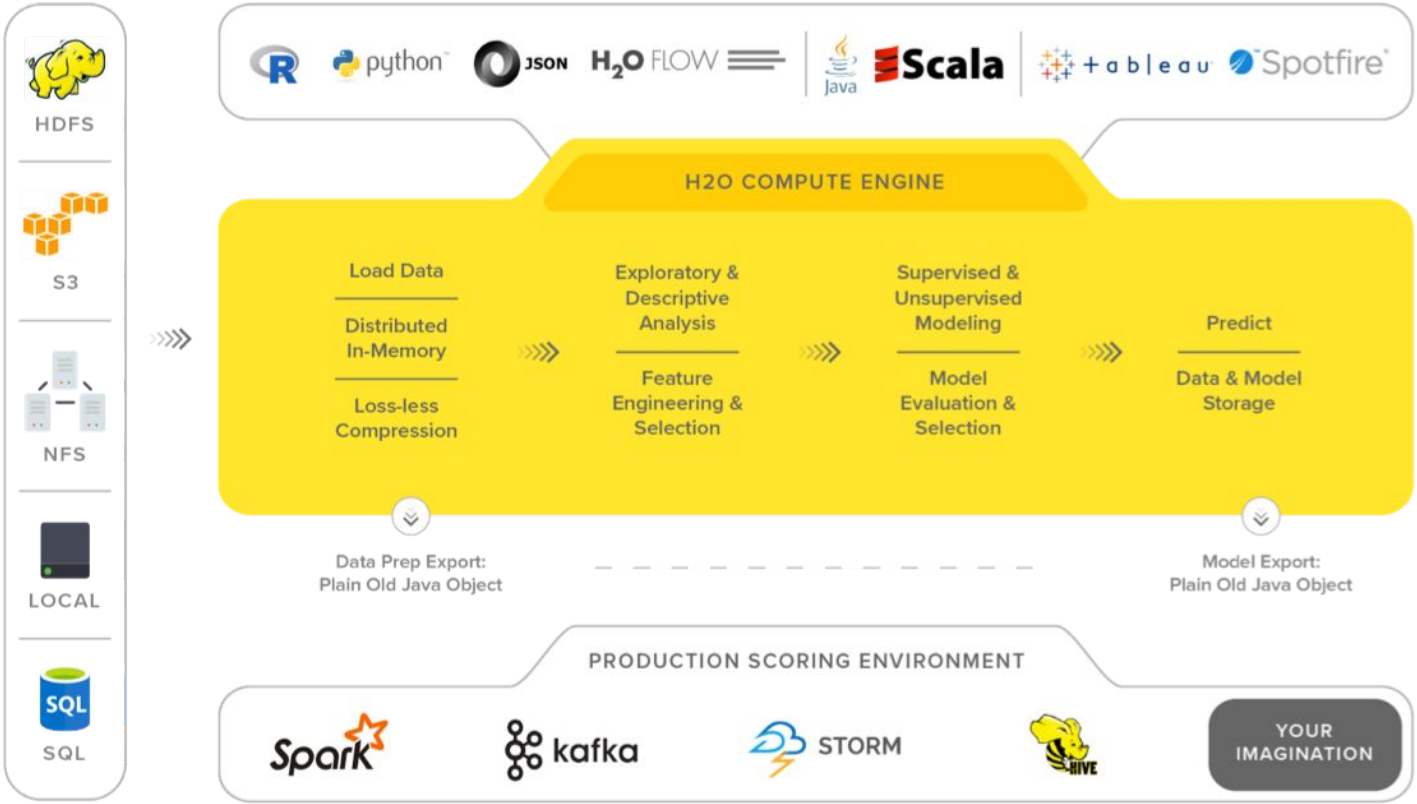
Development process

- Big release few times a year
- Multiple fix releases for each big release
- Open-source **contributions welcome !**

0xdata.atlassian.net

github.com/h2oai

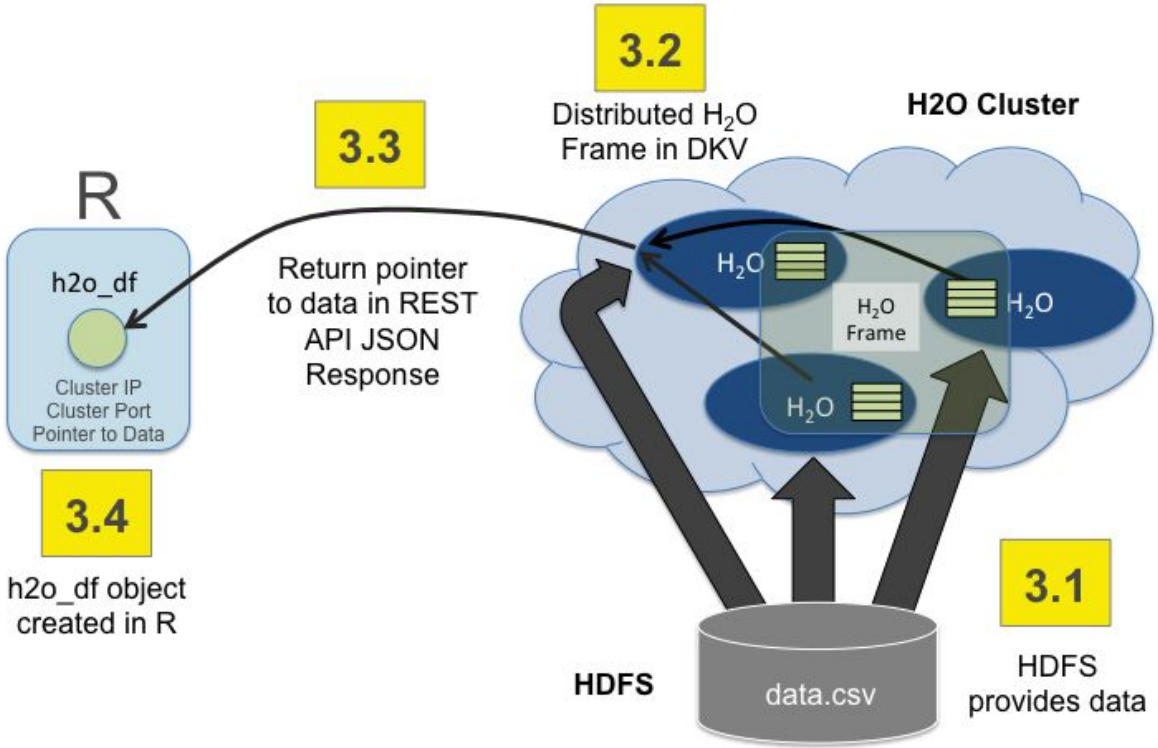
H2O-3 Architecture



H2O-3 Architecture

- From one node to big clusters
- Automatic load-balancing
- Automatic data distribution

H2O-3 Architecture



H2O.ai

Data loading

- Load from many resources
  - CSV, ARFF, SQL, Parquet, ORC, Avro …
- Cloud-ready
  - S3, Google Cloud Storage (OS), HDFS...
- Data type guessing
  - Automatic data-type recognition
  - Best compression for given data-type (nothing is lost !)
- Filter & transform data
  - Skip columns, delete rows, create new features, merge data
  - Missing data imputation
  - Rapids expressions

Creating a model

- Choose an algorithm
  - Both supervised & unsupervised learning supported
  - GBM, GLM, XGBoost … x K-Means, PCA, Aggregator …
- Tune hyper-parameters
  - H2O-3 is about control
- Validate the model
  - Easy-to-do x-validation
- Checkpointing
  - Further develop existing model
- Share the model with other data scientists

Exporting the model

- Export & share with other data scientists ...
- … or create a self-contained Java model
- Pluggable into existing systems
- Easy collaboration between data scientists & developers

POJO/MOJO

How to obtain H2O-3

- Both Python & R download automatically
- Downloadable via h2o.ai/download

How to use H2O-3

- Python
- R
- H2O Flow
- Java API (for experts)

Python

- Pandas
- Scikit
- Python Package Index (PIP)

R

- Same capabilities as Python
- Respects R environment

Algorithm overview

Supervised

- Cox Proportional Hazards (CoxPH)
- Deep Learning (Neural Networks)
- Distributed Random Forest (DRF)
- Generalized Linear Model (GLM)
- Gradient Boosting Machine (GBM)
- Naive Bayes Classifier
- Stacked Ensembles
- XGBoost

AutoML available

Unsupervised:

- Aggregator
- Generalized Low Rank Models (GLRM)
- K-Means Clustering
- Principal Component Analysis (PCA)

Other: Word2Vec

Python

- Both Python 2 & 3 supported (>= 2.7)
- pip install h2o
- import h2o;h2o.init()

LIVE DEMO

H2O.ai

R

- Available in CRAN
- packages.install(“h2o”)
- require(“h2o”);h2o.init()

LIVE DEMO

H2O-Flow

- Web-based interface
- Great for quick prototyping
- Work can be continued in Python/R

LIVE DEMO

Where to go next ?

- DZone - Machine Learning With H2O — Hands-On Guide for Data Scientists
- H2O Docs- http://docs.h2o.ai/
- Twitter- twitter.com/h2oai
- YouTube - youtube.com/user/0xdata
- H2O World - h2oworld.h2o.ai

H2O Book: Practical Machine Learning with H2O