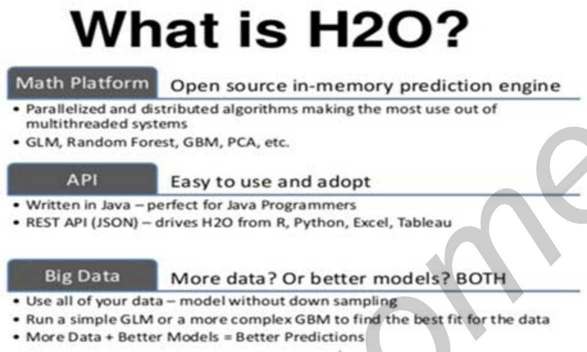
DATA ANALYTICS USING H2O

F. MOHAMMED
Utech LLC location – Madison

D. BALTUNOU
Utech LLC location – Madison

C. DZIK
Utech LLC location – Madison

Utech LLC, Madison, MS USA

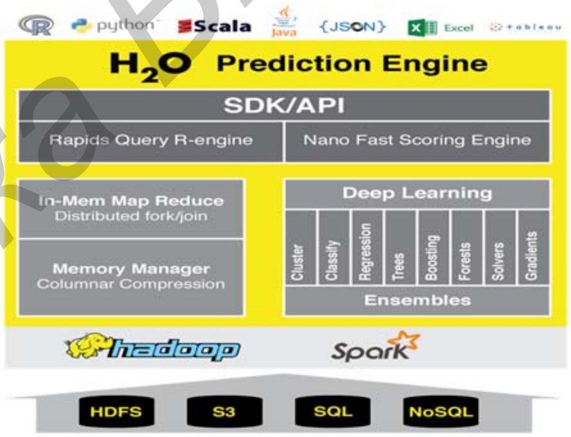


What is H2O?

Math Platform — Open source in-memory prediction engine
- Parallelized and distributed algorithms making the most use out of multithreaded systems
- GLM, Random Forest, GBM, PCA, etc.

API — Easy to use and adopt
- Written in Java – perfect for Java Programmers
- REST API (JSON) – drives H2O from R, Python, Excel, Tableau

Big Data — More data? Or better models? BOTH
- Use all of your data – model without down sampling
- Run a simple GLM or a more complex GBM to find the best fit for the data
- More Data + Better Models = Better Predictions



H_2O Prediction Engine

SDK/API

Rapids Query R-engine | Nano Fast Scoring Engine

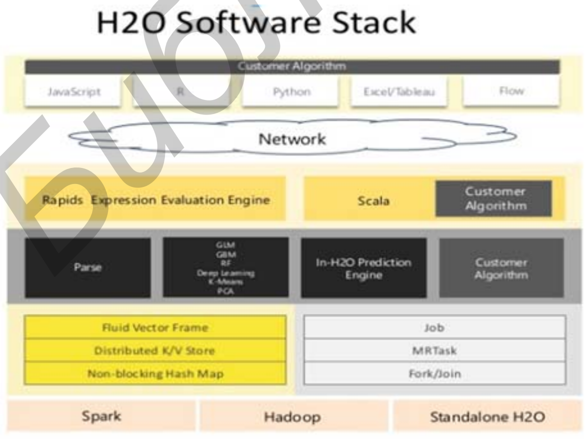
In-Mem Map Reduce
Distributed fork/join

Memory Manager
Columnar Compression

Deep Learning
Cluster, Classify, Regression, Trees, Boosting, Forests, Solvers, Gradients
Ensembles

hadoop | Spark

HDFS | S3 | SQL | NoSQL



H2O Software Stack

Customer Algorithm
JavaScript | R | Python | Excel/Tableau | Flow

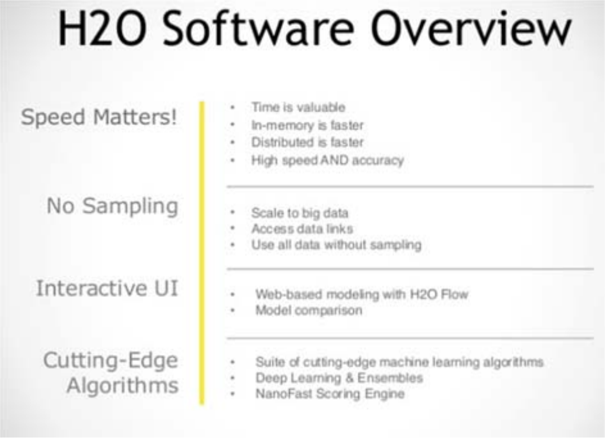
Network

Rapids Expression Evaluation Engine | Scala | Customer Algorithm

Parse | GLM, GBM, RF, Deep Learning, K-Means, PCA | In-H2O Prediction Engine | Customer Algorithm

Fluid Vector Frame | Job
Distributed K/V Store | MRTask
Non-blocking Hash Map | Fork/Join

Spark | Hadoop | Standalone H2O



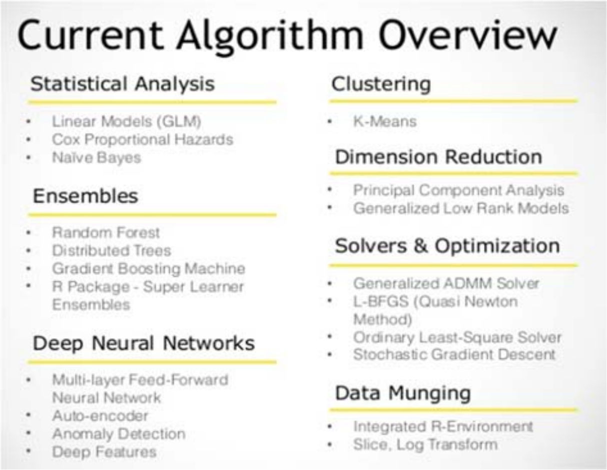
H2O Software Overview

Speed Matters!
- Time is valuable
- In-memory is faster
- Distributed is faster
- High speed AND accuracy

No Sampling
- Scale to big data
- Access data links
- Use all data without sampling

Interactive UI
- Web-based modeling with H2O Flow
- Model comparison

Cutting-Edge Algorithms
- Suite of cutting-edge machine learning algorithms
- Deep Learning & Ensembles
- NanoFast Scoring Engine



Current Algorithm Overview

Statistical Analysis

- Linear Models (GLM)
- Cox Proportional Hazards
- Naïve Bayes

Ensembles

- Random Forest
- Distributed Trees
- Gradient Boosting Machine
- R Package - Super Learner Ensembles

Deep Neural Networks

- Multi-layer Feed-Forward Neural Network
- Auto-encoder
- Anomaly Detection
- Deep Features

Clustering

- K-Means

Dimension Reduction

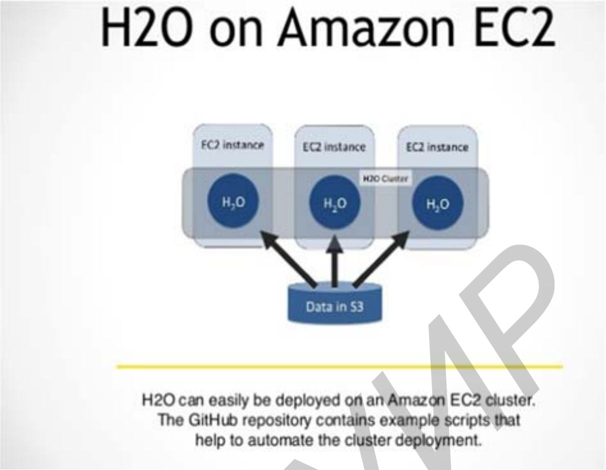
- Principal Component Analysis
- Generalized Low Rank Models

Solvers & Optimization

- Generalized ADMM Solver
- L-BFGS (Quasi Newton Method)
- Ordinary Least-Square Solver
- Stochastic Gradient Descent

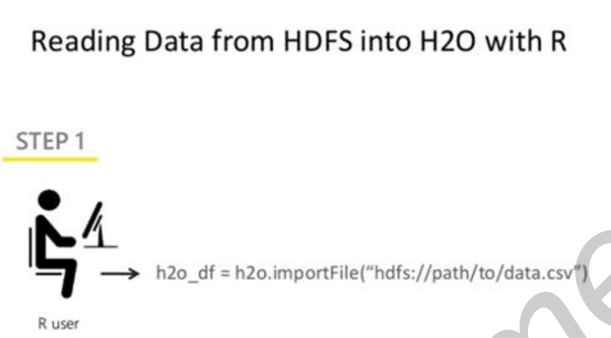
Data Munging

- Integrated R-Environment
- Slice, Log Transform



H2O on Amazon EC2

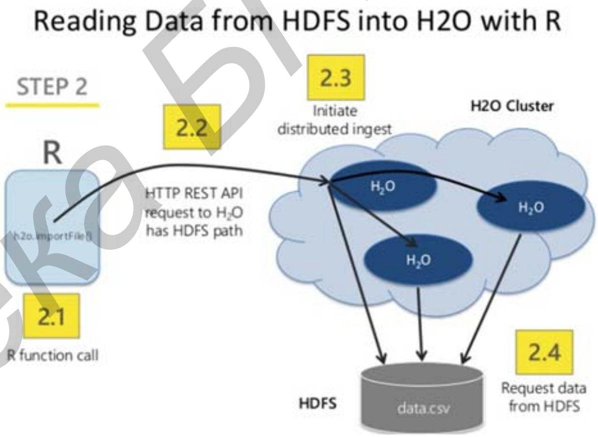
H2O can easily be deployed on an Amazon EC2 cluster. The GitHub repository contains example scripts that help to automate the cluster deployment.



Reading Data from HDFS into H2O with R

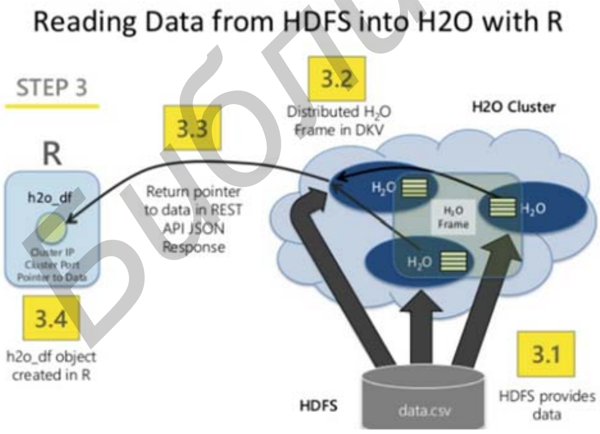
STEP 1

```
h2o_df = h2o.importFile("hdfs://path/to/data.csv")
```



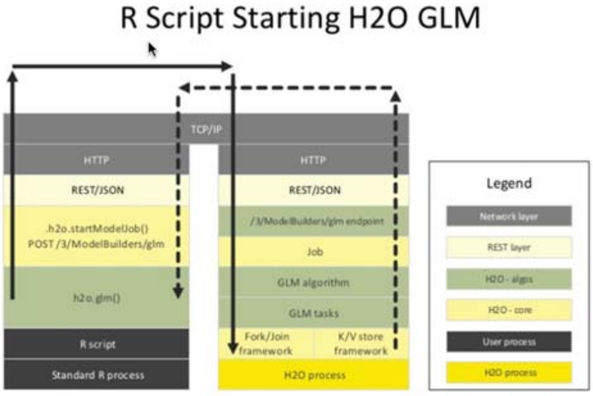
Reading Data from HDFS into H2O with R

STEP 2

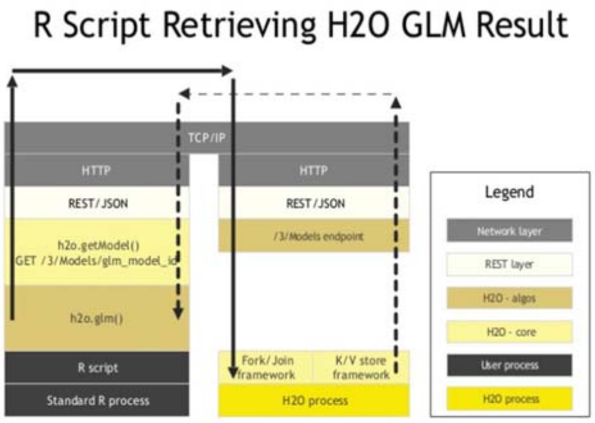


Reading Data from HDFS into H2O with R

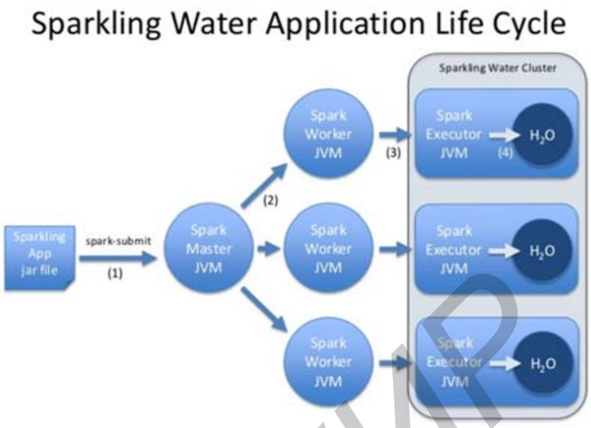
STEP 3



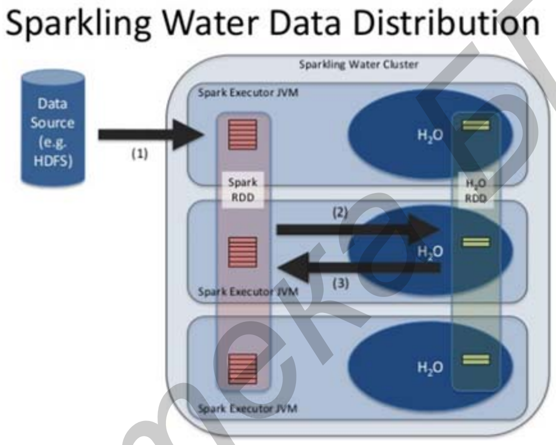
R Script Starting H2O GLM



R Script Retrieving H2O GLM Result



Sparkling Water Application Life Cycle



Sparkling Water Data Distribution