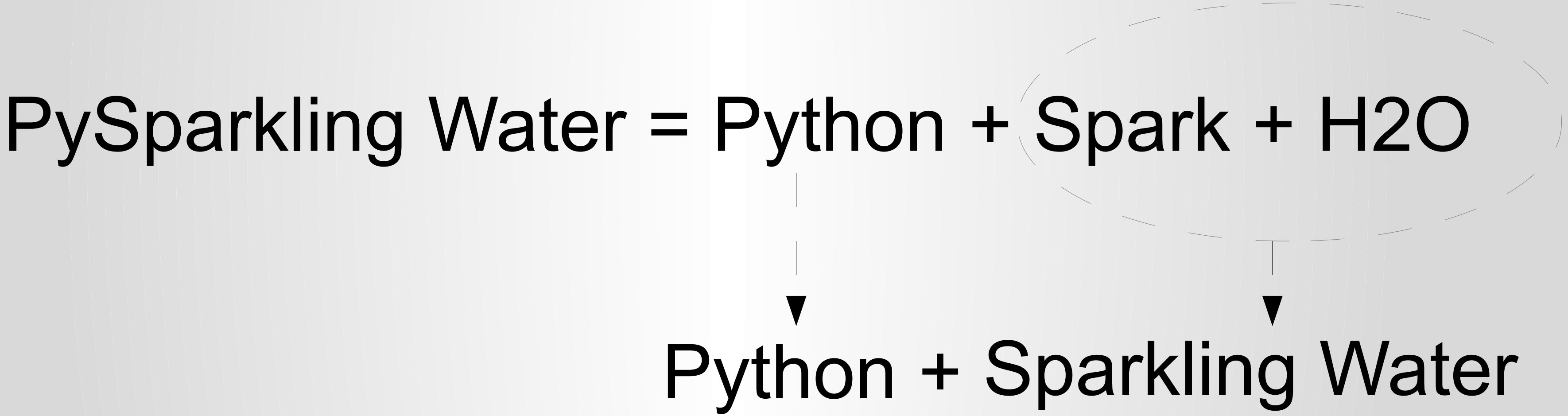
Hands On : PySparkling Water

- By Nidhi Mehta

H_2O WORLD 2015

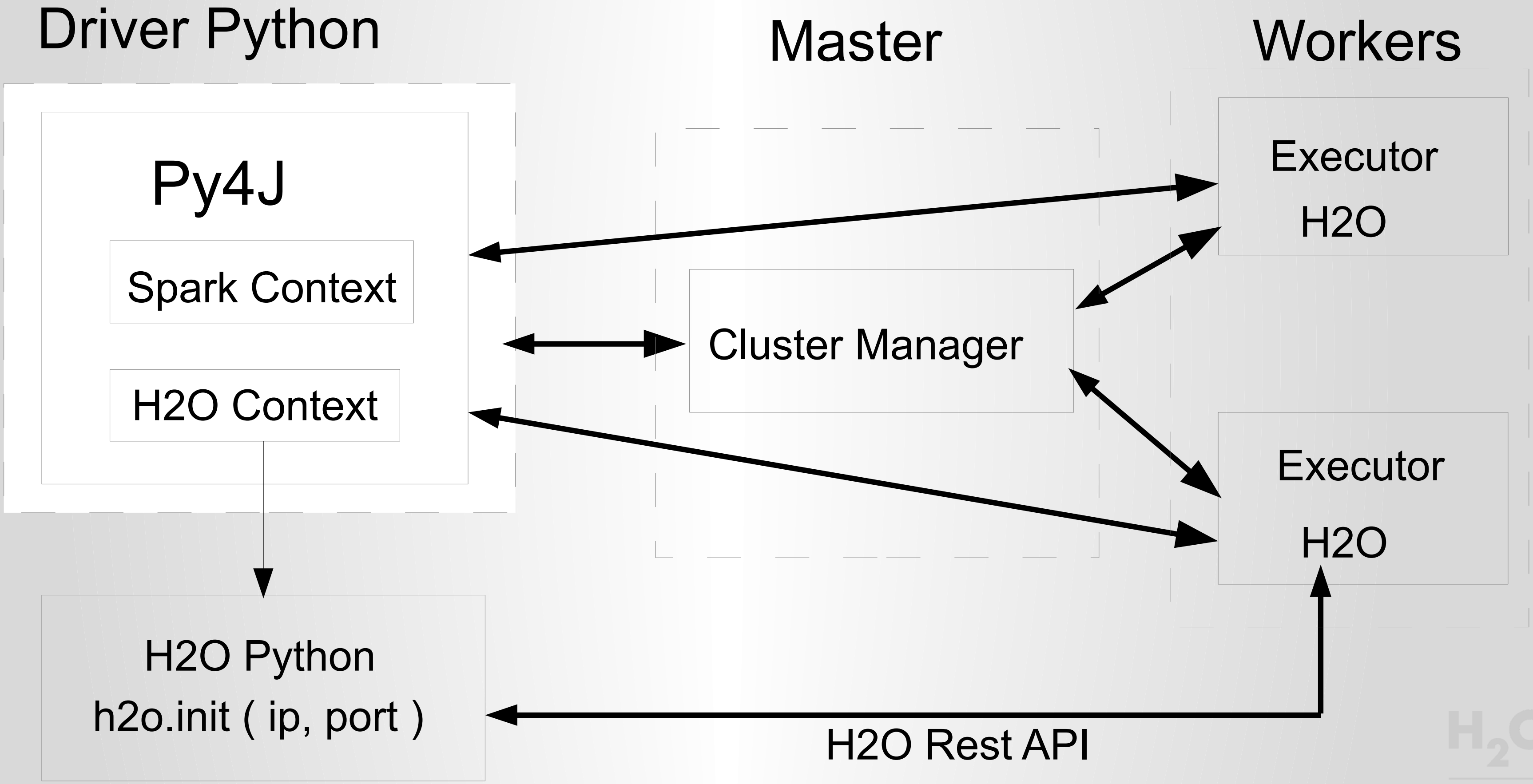
What is PySparkling Water



PySparkling Water = Python + Spark + H2O

Python + Sparkling Water

PySparkling Architecture



Demo Workflow

Aim: Build a model to predict Arrest for Chicago crime dataset

- Import Chicago Crime Dataset
- Combine Crime data with Census and Weather data.
- Build a model to predict whether an arrest was made
- Predict on a test dataset

Pre Requisites to run the demo

- Install Spark-1.5.1

- Install and Build Sparkling Water-1.5.6

(./gradlew build -x check)

- Install H2O-3.6.0.3

- Install H2O-python

(sudo pip install h2o-3.6.0.3-py2.py3-none-any.whl)

Command to Start/Access PySparking Water Cluster

1)

Set spark environment by specifying SPARK_HOME

```
export SPARK_HOME = Path_to_Spark_dir
```

2)

- To run from Python notebook-

```
IPYTHON_OPTS="notebook" Path_to_Sparkling_dir/bin/pysparkling
```

- To run from regular Python shell

```
Path_to_Sparkling_dir/bin/pysparkling
```

Let's Run the Demo!

Why use PySparkling

- Automatic Parallelization and less lines of code
- Much Faster on big data - uses H2O's rest API calls to connect to H2O Cluster

Thank You

What do these stickers mean?

I have Sparkling Water Installed

H_2O
INSTALLED

I have H2O installed

I have Python installed

DATA SETS
DOWNLOADED

I have the H2O World data sets

Pick up stickers or get install help at the information booth