

# Hands on Introduction to Sparkling Water



# Agenda



# Today's Talk

## Sparkling Water Installation

---

- Installing Sparkling Water
- Installing PySparkling

## Introduction to Sparkling Water

---

- Typical Uses
- How it Works?

## Hands on Demo

---

- Our Use Case
- Importing data into H2O
- Converting Spark Dataframe to H2O Frame and vice versa
- Data Cleaning and Feature Engineering
- Using Flow with Sparkling Water

# Sparkling Water Installation



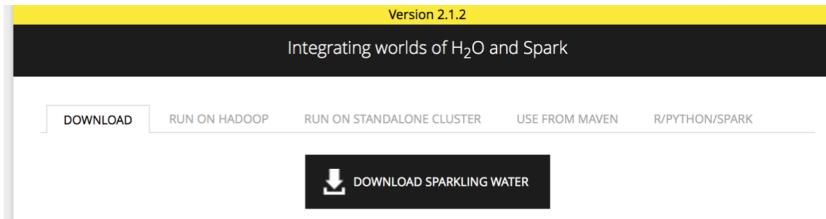
# Sparkling Water Prerequisites

To Launch Sparkling Water the prerequisites are:

- Java 7+
- Spark 1.6+

# Installing Sparkling Water

## Download Sparkling Water



## Export Environment Variables

```
export SPARK_HOME="/path/to/spark/installation"  
# To launch a local Spark cluster with 3 worker nodes with 2 cores and 1g per node.  
export MASTER="local[*]"
```

## Start Sparkling Water

```
cd ~/Downloads  
unzip sparkling-water-2.3.5.zip  
cd sparkling-water-2.3.5  
bin/sparkling-shell --conf "spark.executor.memory=1g"
```

# Installing PySparkling

Prerequisite: Python 2.7

## Install Dependencies

```
pip install requests
pip install tabulate
pip install six
pip install future
```

## Start PySparkling

```
cd ~/Downloads
unzip sparkling-water-2.3.5.zip
cd sparkling-water-2.3.5
# Run PySparkling
bin/pysparkling
# Run Ipython Notebook
PYSPARK_DRIVER_PYTHON="ipython" PYSPARK_DRIVER_PYTHON_OPTS="notebook" bin/pysparkling
```

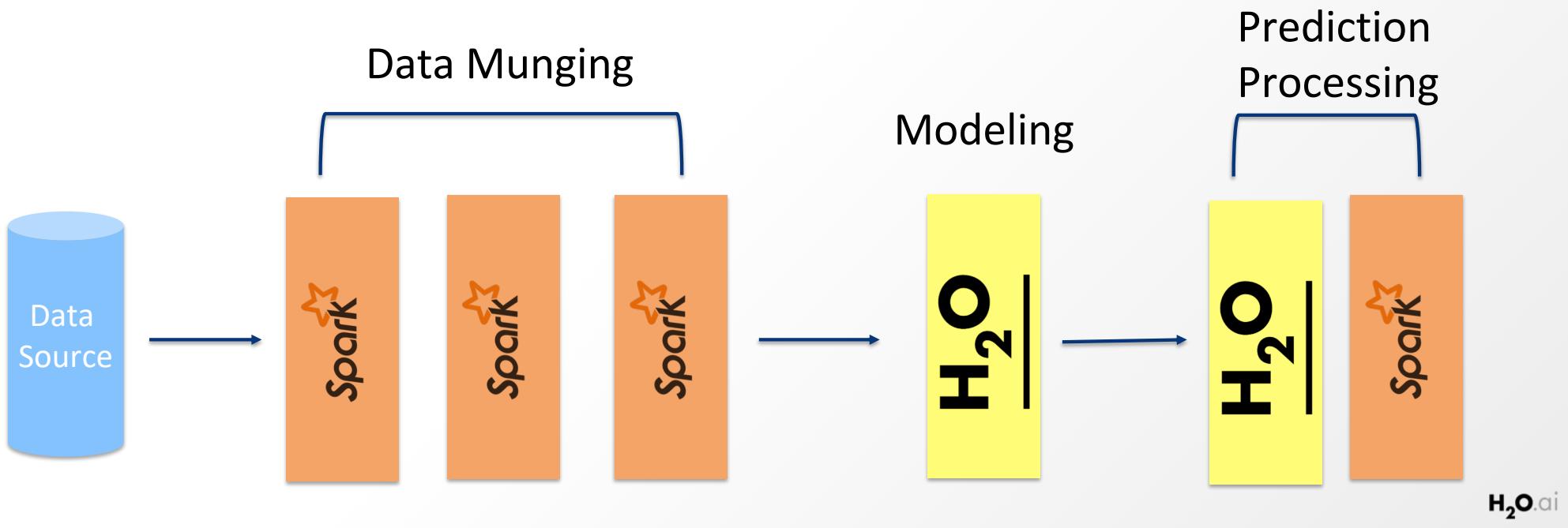
# Introduction to Sparkling Water



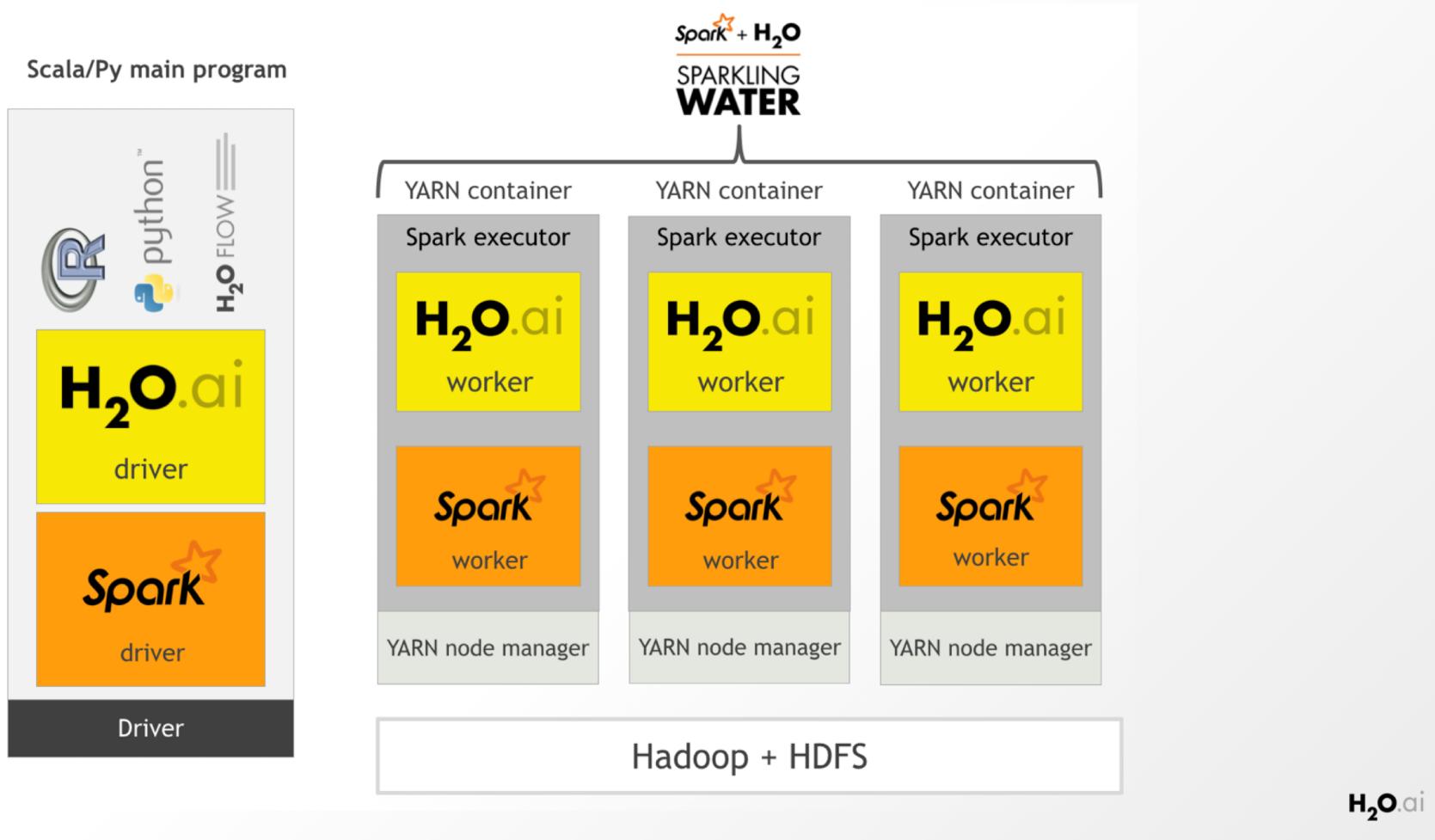
# What is Sparkling Water?

Provides

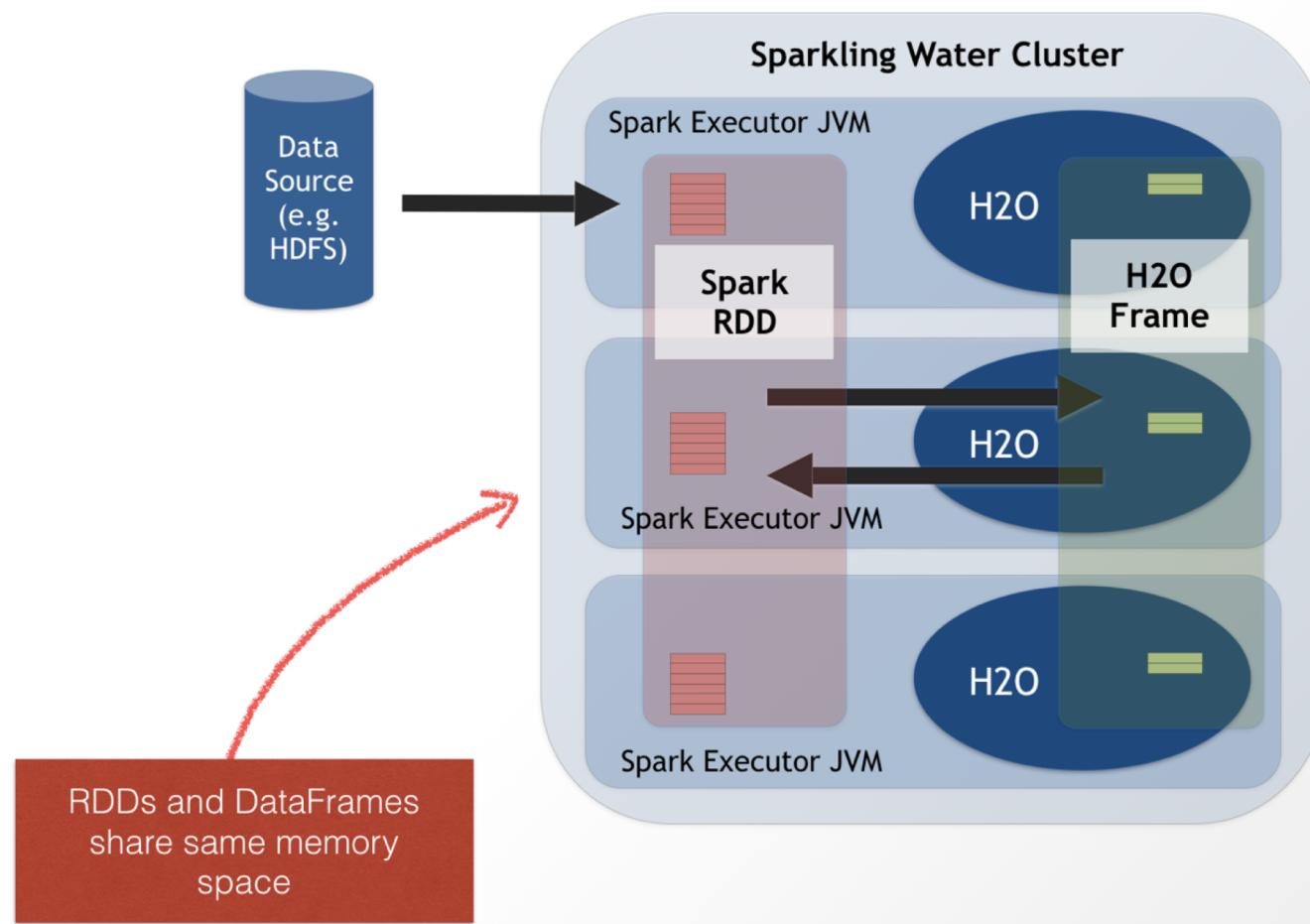
- Transparent integration of H2O with Spark ecosystem
- Transparent use of H2O data structures and algorithms with Spark API
- Seamlessly toggle between Spark Dataframes and H2O Frames



# How Does it Work?



# How Does it Work?



# Demo



# Use Case

## End Goal

Predict whether an arrest will be made for a given crime.

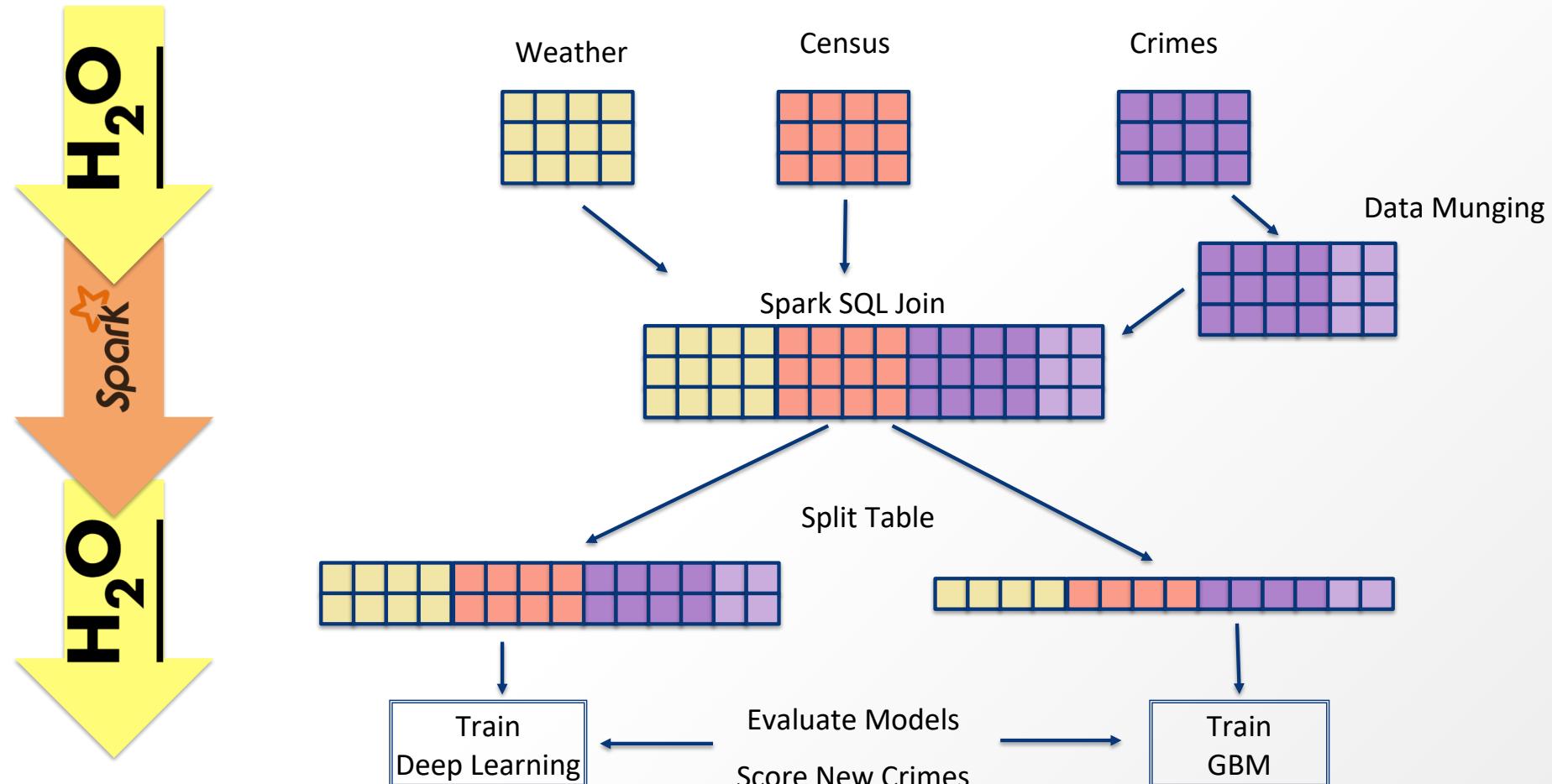
## Why?

We can use this model to improve public safety by predicting whether an arrest will be made and also by analyzing the factors that contribute to a high probability of arrest.

## How?

Join crime data with external data like weather and socioeconomic factors to train a model that will predict arrest.

# Use Case



# Resources

- Data: `sparkling-water-2.3.5/examples/smalldata`
- Scala Script: `ChicagoCrimeDemo.scala`
- Python Script: `ChicagoCrimeDemo.ipynb`

# Questions?

