Sparkling Water

Why use Sparkling Water?

Spark Cluster

Leverage Spark Cluster

- Launch H2O-3 Machine Learning Engine on top of your Spark Cluster

Spark Data Munging

Leverage Spark Data Munging Functionality

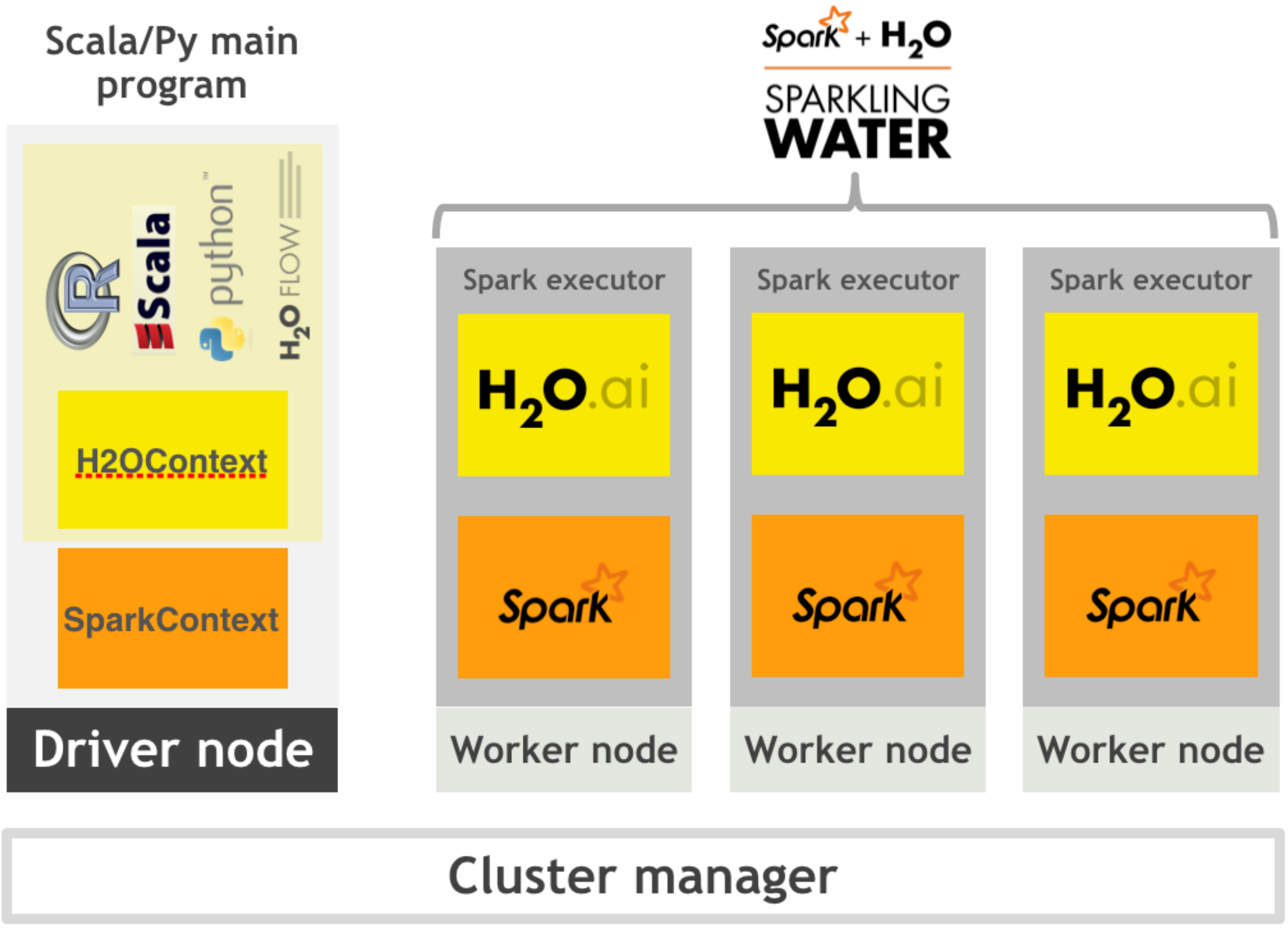
- Transparent integration of H2O with Spark ecosystem
- Transparent use of H2O data structures and algorithms with Spark API

Spark Pipelines

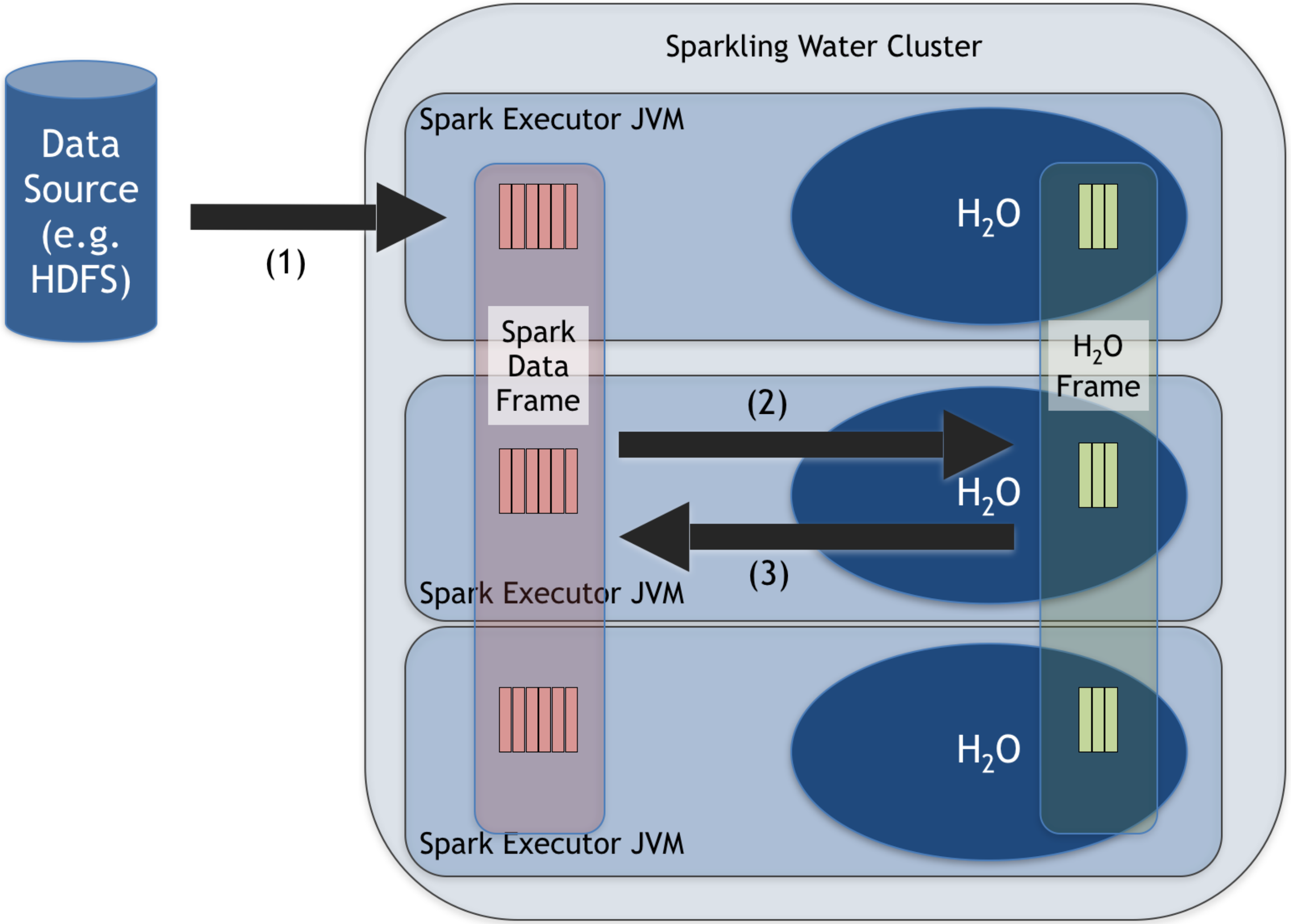
Leverage Spark Pipelines

- Incorporate H2O-3 models into existing Spark pipelines
- Excels in existing Spark workflows requiring advanced Machine Learning algorithms

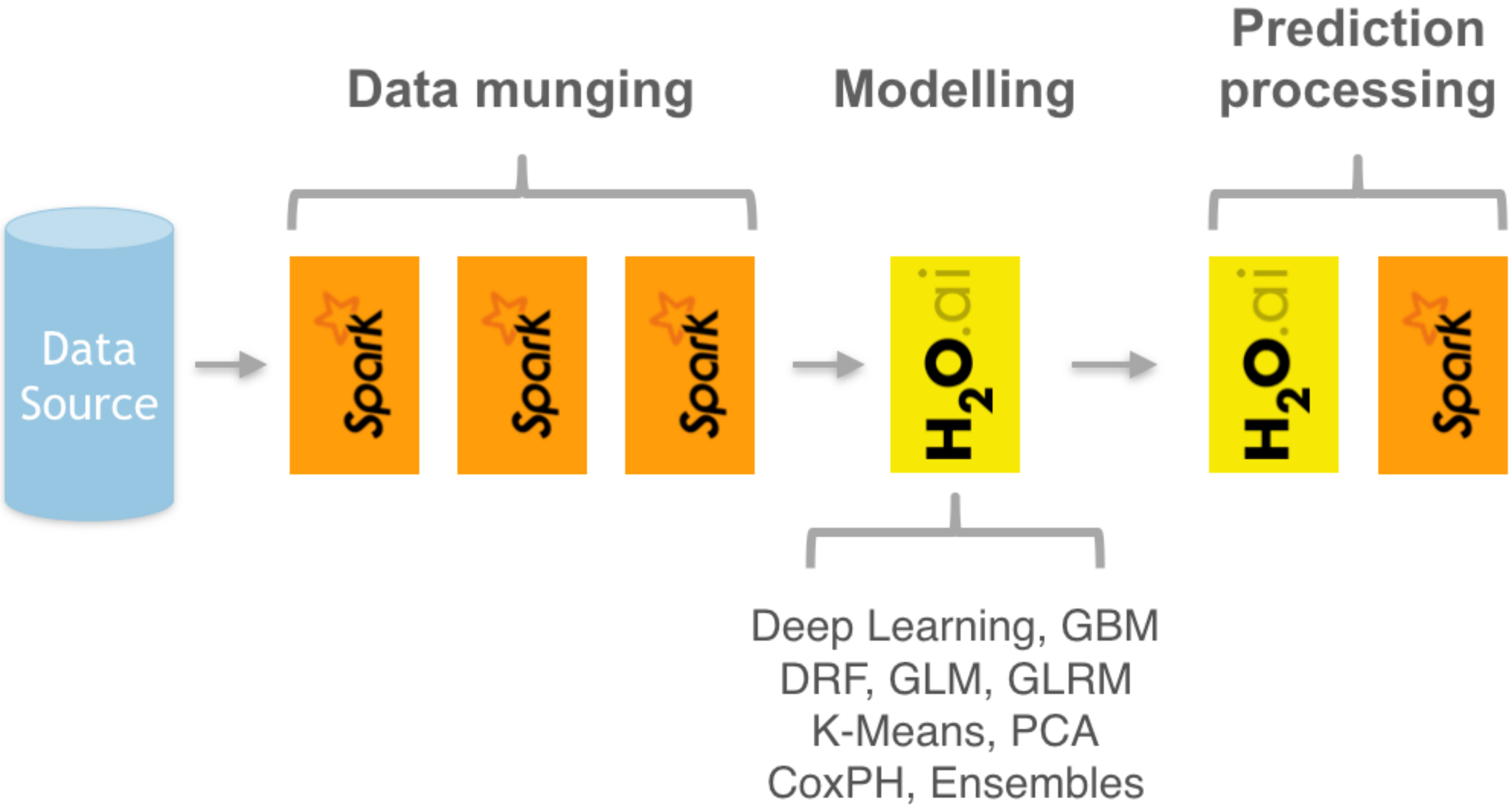
High Level Architecture



Sparkling Water Data Management



Use Case



Data Source

Data munging

Modelling

Prediction processing

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

Sparkling Water Data Conversion Functions

- Converting an H2OFrame into an RDD[T]

```
def asRDD[A <: Product: TypeTag: ClassTag](fr : H2OFrame) : RDD[A]
```

- Converting an H2OFrame into a DataFrame

```
def asDataFrame(fr : H2OFrame)(implicit sqlContext: SQLContext) : DataFrame
```

- Converting an RDD[T] into an H2OFrame

```
def asH2OFrame[A <: Product : TypeTag](rdd : RDD[A], frameName: Option[String]) : H2OFrame
```

- Converting a DataFrame into an H2OFrame

```
def asH2OFrame(rdd : DataFrame, frameName: Option[String]) : H2OFrame
```

Sparkling Water Hands On

The Data

- Dow Jones Industrial Average data from 2006-2017
- 93,612 rows

Forecasting for these groups

Time Unit

What we want to predict

Additional Attributes

Name	Date	Volume	Open	High	Low	Close
AABA	2006-01-03	24,232,729	$39.69	$41.22	$38.79	$40.91
AABA	2006-01-04	20,553,479	$41.22	$41.90	$40.77	$40.97
AABA	2006-01-05	12,829,610	$40.98	$41.73	$40.85	$41.53
AABA	2006-01-06	29,422,828	$42.88	$43.57	$42.80	$43.21

The Data

Original Data

Name	Date	Volume	Open	High	Low	Close
AABA	2006-01-03	24,232,729	$39.69	$41.22	$38.79	$40.91
AABA	2006-01-04	20,553,479	$41.22	$41.90	$40.77	$40.97
AABA	2006-01-05	12,829,610	$40.98	$41.73	$40.85	$41.53
AABA	2006-01-06	29,422,828	$42.88	$43.57	$42.80	$43.21

Formulated for Machine Learning

Name	Close Yesterday	Volume 2 Days Ago	Volume Yesterday	Volume (Target)
AABA	NA	NA	NA	24,232,729
AABA	$40.91	NA	24,232,729	20,553,479
AABA	$41.53	24,232,729	20,553,479	12,829,610
AABA	$43.21	20,553,479	12,829,610	29,422,828

Performance Comparison

Data	Mean Absolute Percent Error
Original Data	134%
Data with Lags	21%