RED WINGED BLACK BIRD Sighting PREDICTION

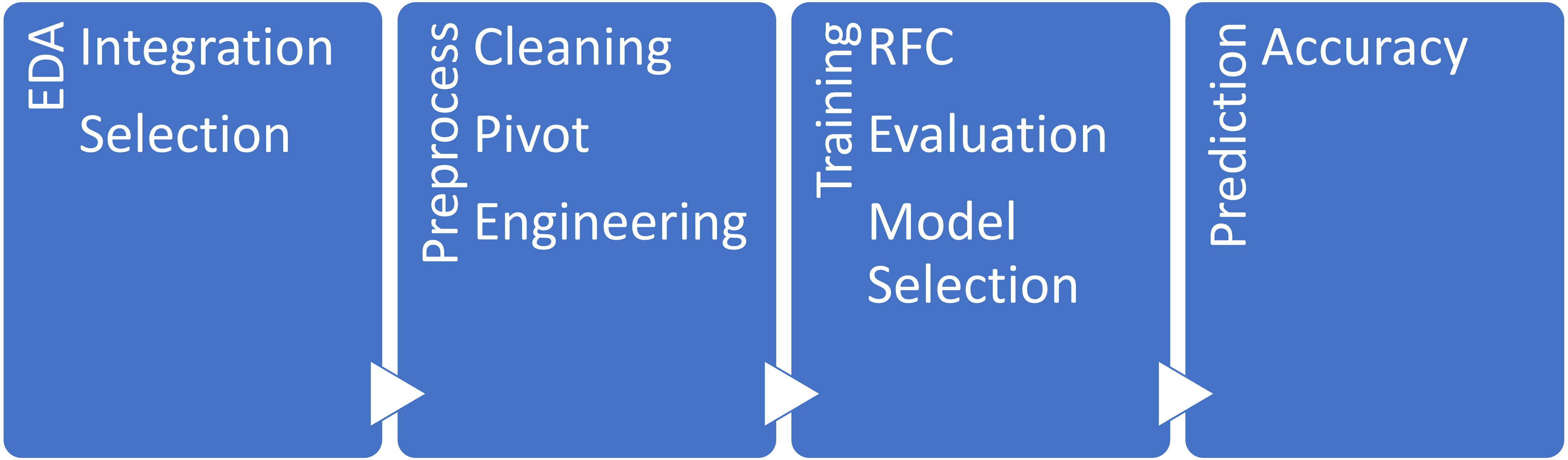
DATA MINING in Spark Mllib USING RANDOM FOREST CLASSIFIER

Ayush Singh & Gautam Vashisht
CS 6240: Parallel big data processing with MapReduce
Prof: Mirek Riedewald
Spring 2017, Northeastern University

Overview

- Predict sightings
- Accuracy on hold out set: 88%
- With minimal pre processing
- Random Forest Classifier Model is used to predict the final output

Mark Peck



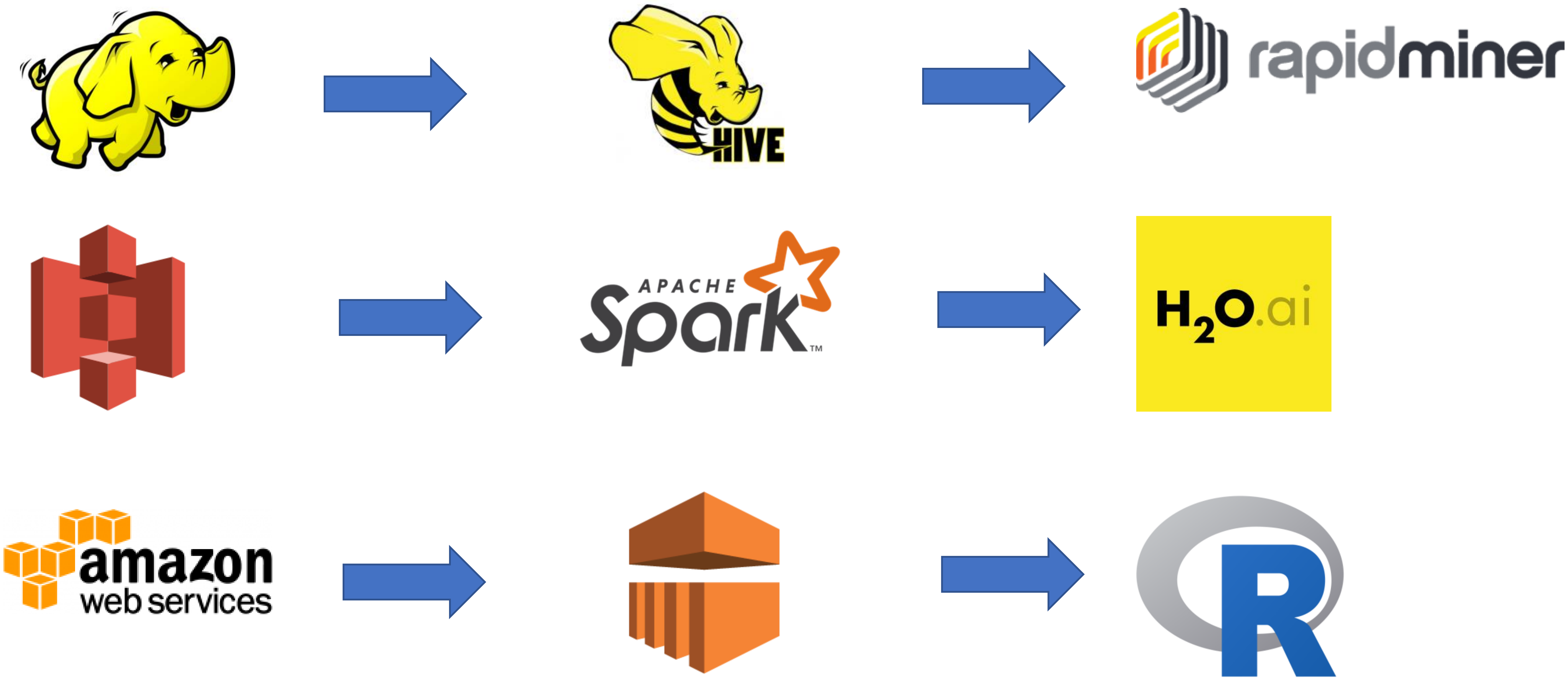
EDA
- Integration
- Selection

Preprocess
- Cleaning
- Pivot
- Engineering

Training
- RFC
- Evaluation
- Model Selection

Prediction
- Accuracy

Exploratory Data Analysis: Tools



Feature Engineering and Findings

- Stratified sampling correlative variable evaluation
- Imputed with minimal values to not disturb distribution
- Binning and Discretization on continuous data
- Pivot on First Order statistics
- Integrate Weather API to fill remaining GIS values
- Replacing by mean reduced accuracy
- Generalized low-rank modelling for compressing

Random Forest Classifier

- Why Spark?
- Random Forest is best suited for parallelism
- Can handle large variance
- Ensemble of randomly built trees
- Parallel Implementation in spark mllib

Evaluation and Selection

- Increase trees only if accuracy increases
- Increasing depth highly correlated with performance
- Cross Validation traded off due to OOB error of RFC
- Optimal Parameters: 50 trees, 20 depth
- Maximum number of bins = 1024

Different AWS Instance Run

AWS Instance	Prediction on data Set	Time	Output	Test Error
m4.2xlarge with 16 nodes	On split test set from labeled data.	6.20 hour and still running	Aborted the program as it was exceeding expected time limit.	NA
M4.2xlarge with 11 nodes	On unlabeled test set	3.48 hour and still running	Aborted the program as it was exceeding expected time limit.	NA
r4.2xlarge	On split test set from labeled data.	2.28 hour	Successful	**0.1261512360**
r4.2xlarge	On unlabeled test set	1.18 hour	Successful	NA

Conclusion

- Resources vs Performance is a tradeoff
- Occam’s Razor still stands true
- Big data is scary

Thank You