# Big Data Architectures for Machine Learning and Data Mining

## Short Talk Topics

Michael Völske[1]

April 28, 2016

[1] michael.voelske@uni-weimar.de

# Outline

# Overview

1. Apache Spark
2. Apache Spark: MLLib
3. Apache Flink
4. $H_2O$
5. Apache Mahout
6. DeepLearning4J
7. Petuum

# Apache Spark

- Starting point: spark.apache.org

## Talk Ideas

- Installation and setup on YARN cluster
- Spark basic concepts, e.g. what is an RDD, how does it work?
- Advantages and drawbacks vs MapReduce
- Interactive data analysis with `(py)spark-shell`
- Some interesting features, e.g. DataFrames, SQL, GraphX, Streaming...

- Languages: Java, Python, Scala

# Apache Spark: MLLib

- Starting point: spark.apache.org/mllib

## Talk Ideas

- Basic data types: LabeledPoint, local and distributed matrix etc.
- `spark.ml` "pipeline" API
- Available algorithms and what they are for

- Languages: Java, Python, Scala

# Apache Flink

- Starting point: flink.apache.org

## Talk Ideas

- Installation and setup on YARN cluster
- DataSet and DataStream APIs
- FlinkML Machine Learning library
- Gelly graph library
- Comparison to Apache Spark

- Languages: Java, Scala

# $H_2O$

- Starting point: www.h2o.ai/docs

## Talk Ideas

- Overview of available learning algorithms
- "Flow" web interface
- Basic Python API
- "Sparkling Water" – integration with Spark

- Languages: Python, R, Scala

# Apache Mahout

- Starting point: mahout.apache.org

## Talk Ideas

- Available algorithms
- Old MapReduce-based implementations vs Mahout-Samsara
- Distributed Linear Algebra
- How to run Mahout programs on Spark/Flink
- Mahout's interactive Spark shell

- Languages: Java, Scala

# DeepLearning4J

- Starting point: deeplearning4j.org

## Talk Ideas

- Installation and setup on YARN cluster
- NLP Framework for text analytics
- Running on Spark
- A simple learning example (e.g. Logistic Regression)

- Languages: Java, Scala

# Petuum

- Starting point: petuum.github.io

## Talk ideas

- Components of the Petuum framework
- Available machine learning algorithms
- Running Petuum programs on YARN
- A simple learning example

- Languages: Java, C++

# For All Talks

- Find additional material, maybe look at the code (usually on GitHub)
- Answer the questions, "what is it?" and "what can it do?"
- Focus on running on a YARN cluster
- Learn to actually use the system for some basic tasks
- Show some (simple) code examples
- The "talk ideas" are just a suggestion; the main goal is a good, informative talk

# Outlook

- Until next week:
  - Collect material
  - Decide what you plan to present
  - Prepare an outline for your talk

# Outlook

- Until next week:
  - Collect material
  - Decide what you plan to present
  - Prepare an outline for your talk
- Next week:
  - Coordinate with other presenters regarding collaboration/overlaps.
  - Show me what you have planned, discuss any problems/open questions.

# Outlook

- Until next week:
  - Collect material
  - Decide what you plan to present
  - Prepare an outline for your talk
- Next week:
  - Coordinate with other presenters regarding collaboration/overlaps.
  - Show me what you have planned, discuss any problems/open questions.
- In two weeks: (holiday)
- In three weeks:
  - 3-4 short talks.
  - Final presentation topics and organization
- In four weeks:
  - The remaining short talks