Agile Machine Learning with Scalding and scikit-learn

QCON San Francisco

Thursday, Nov 6, 2014

By Sasha Ovsankin

Summary

Use the tool that serves your customer best

About Me

- Graduated from Moscow University, Dept. of Physics, Chair of Mathematics
- Work on Messaging Experience at LinkedIn
- Interested in Big Data, Machine Learning, Software Architecture

  – https://linkedin.com/in/sashao
  – https://twitter.com/SashaO

The Plan

- Acquire data using Scalding
- Explore data using scikit-learn
- Think and reflect

Why Agile

- Data is Big
- Clusters are big
- Processes are Slow
- But there is a hope: you don’t always need all the data
- Solution: sample data

Scalding

https://github.com/SashaOv/MLTalk

git clone https://github.com/SashaOv/MLTalk.git

iPython

Scikit-learn

Bibliography

- Programming MapReduce with Scalding
- https://www.packtpub.com/big-data-and-business-intelligence/programming-mapreduce-scalding
- IPython Interactive Computing and Visualization Cookbook https://www.packtpub.com/big-data-and-business-intelligence/ipython-interactive-computing-and-visualization-cookbook

Summary

- Use the tool that serves your customer best
- Be agile