tripadvisor®

Data Engineering for Data Science

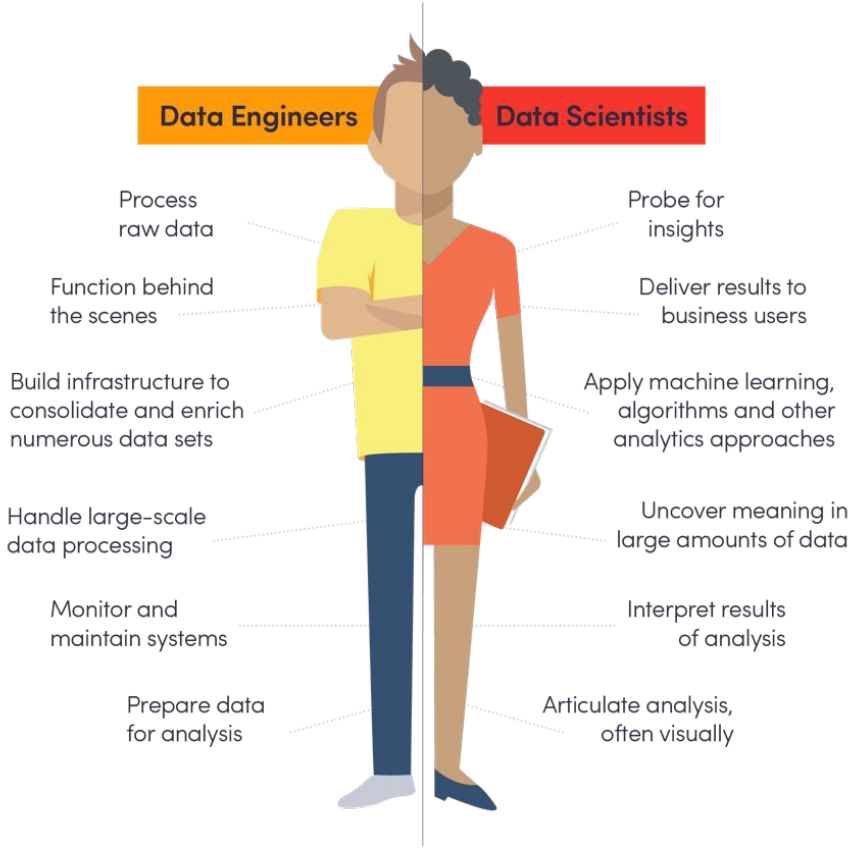
Sharon Xu
February 2020

Table of Contents

Workshop Agenda

1. Data Engineering at TripAdvisor
2. MapReduce, Hadoop, and Spark
3. Hands-on Exercise

Scientists “vs.” Engineers



Engineering ↔ Modeling

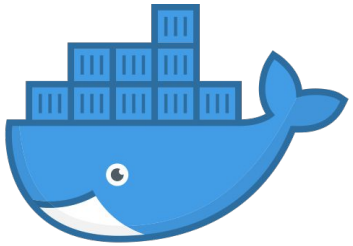
As data scientists we should…

- Understand ETLs and how features are created. Need to be aware of bias and noise that may be introduced.
- Consider real-time complexity when choosing features and building models.

Data Corner

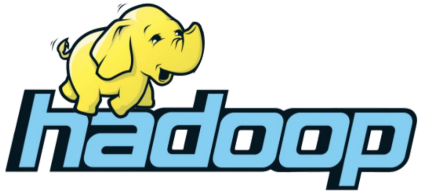
Data Engineering at TripAdvisor

The Tools



docker

kubernetes



python
SOFTWARE FOUNDATION

Java

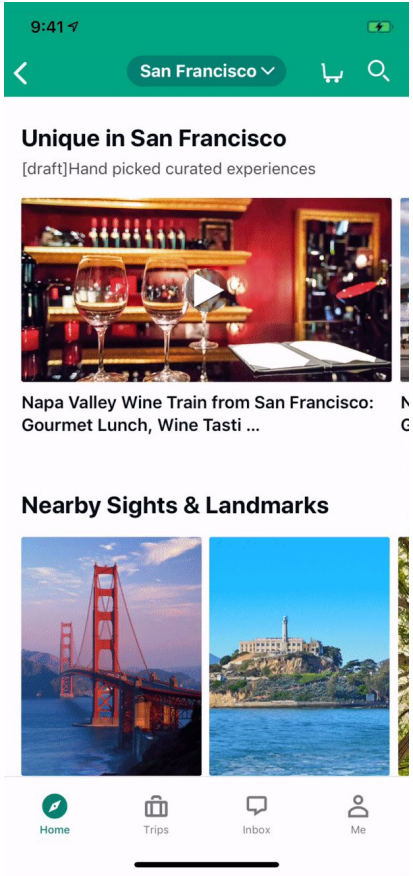
PostgreSQL

tripadvisor

What We Work On

A diverse set of applications

- Learn to rank models
- Recommender systems
- Machine vision
- Search results ranking
- Causal inference
- Review fraud detection
- Bot detection
- Booking / content models
- Automated collection generation
- Keyphrase generation
- Location classification



Blog: https://www.tripadvisor.com/engineering/tag/machine-learning

Questions?

Spark: A Hands-On Exercise

Setup

1. Clone the repository

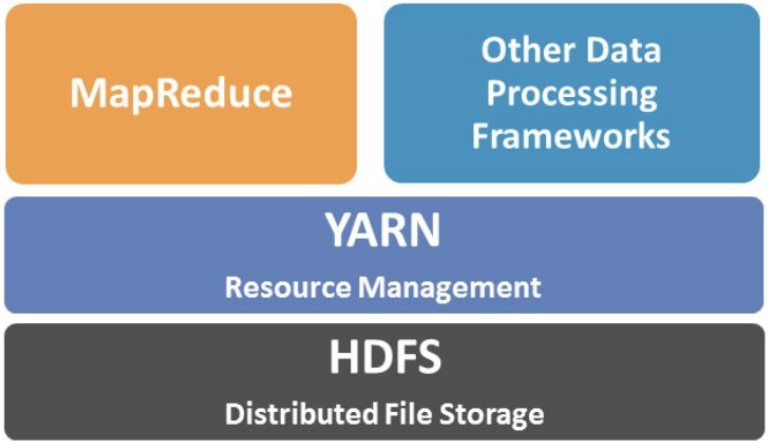
   ```
   git clone https://github.com/sharonxu/data-engineering-tufts.git
   ```

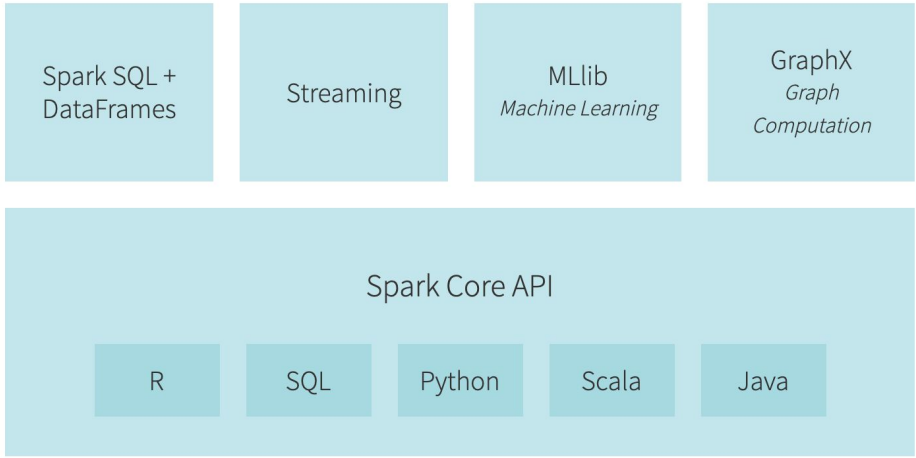
2. Open the Colab Notebook： Data Exploration with Spark

3. Upload data.zip to the notebook

4. Run cells up to “Step 1 - What inputs do you have?”

MapReduce, Hadoop, and Spark





Time to code!

What Next?

Resources and Next Steps

Data Engineering

- Learn SQL (SQLZOO)
- Big Data Analysis with Apache Spark
- Play around with an API (like Spotify's)

Data Science

- RecSys Challenge
- Google's Dataset Search
- Kaggle

Cool Colab Notebooks

- Intro Notebook with ML Resources
- Language Models from N-grams to GPT2 (Harvard ComputeFest 2020)
- Solutions for our PySpark Exercise

Thank You!

Questions?

LinkedIn: **https://www.linkedin.com/in/sharon6**

Github: **https://github.com/sharonxu**