Exploring Wikipedia with Spark

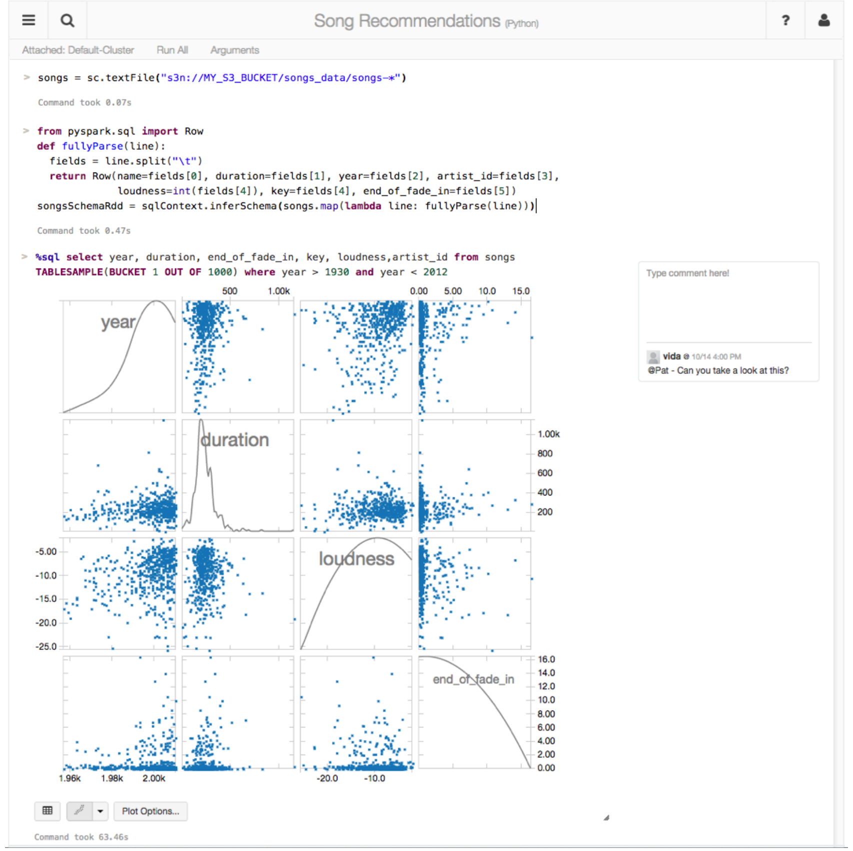
databricks

databrickstraining

databricks™

making big data simple

- Founded in late 2013
- by the creators of Apache Spark
- Original team from UC Berkeley AMPLab
- Raised $47 Million in 2 rounds
- ~65 employees
- We're hiring!
- Level 2/3 support partnerships with
  - Hortonworks
  - MapR
  - DataStax



```
> songs = sc.textFile("s3n://MY_S3_BUCKET/songs_data/songs-*")

  Command took 0.07s

> from pyspark.sql import Row
  def fullyParse(line):
    fields = line.split("\t")
    return Row(name=fields[0], duration=fields[1], year=fields[2], artist_id=fields[3],
               loudness=int(fields[4]), key=fields[4], end_of_fade_in=fields[5])
  songsSchemaRdd = sqlContext.inferSchema(songs.map(lambda line: fullyParse(line)))

  Command took 0.47s

> %sql select year, duration, end_of_fade_in, key, loudness,artist_id from songs
  TABLESAMPLE(BUCKET 1 OUT OF 1000) where year > 1930 and year < 2012
```

vida @ 10/14 4:00 PM
@Pat - Can you take a look at this?

```
Command took 63.46s
```

Databricks Cloud:
"A unified platform for building Big Data pipelines – from ETL to Exploration and Dashboards, to Advanced Analytics and Data Products."

The Databricks team contributed more than 75% of the code added to Spark in the past year

Instructor: Adam Breindel

LinkedIn: https://www.linkedin.com/in/adbreind
Email: adam@databricks.com

- 15+ years building systems for startups and large enterprises
- 8+ years teaching front- and back-end technology
- Fun big data projects…
  - Streaming neural net + decision tree fraud scoring (debit cards)
  - Realtime & offline analytics for banking
  - Music synchronization and licensing for networked jukeboxes
- Industries
  - Finance
  - Travel
  - Media / Entertainment

Today's Objectives

- Learn just enough to build simple prototypes/POCs with Spark
- Fast paced, high level overview of all major Spark components
- Hands on with Spark's programming APIs *(DataFrame/SQL, RDD, Datasets)*
- Overview of Spark architecture: Core, Streaming, Standalone Mode, DAG
- Mix of beginner + advanced topics
- Not all slides/labs covered *(reference & homework material)*
- Lots of ideas, code & datasets to play around with after class

Schedule

9:00 a.m. – Welcome, Login

Data Analysis (DA) and Query Execution (DE) - pagecounts and pageviews datasets

10:45 – 11 a.m. Coffee Break

Clickstream Aalytics (DA), Infrastructure (DE) - clickstream and pagecounts datasets

12 p.m. – 1 p.m. Lunch

RDD, Dataset, Storage/Memory (DE) – pagecounts
ETL / Graph Analysis – clickstream

2:15 p.m. – 2:30 p.m. Soda Break

NLP / Machine Learning

4:00 p.m. – 4:15 p.m. Coffee Break

Spark Streaming

Files and Resources

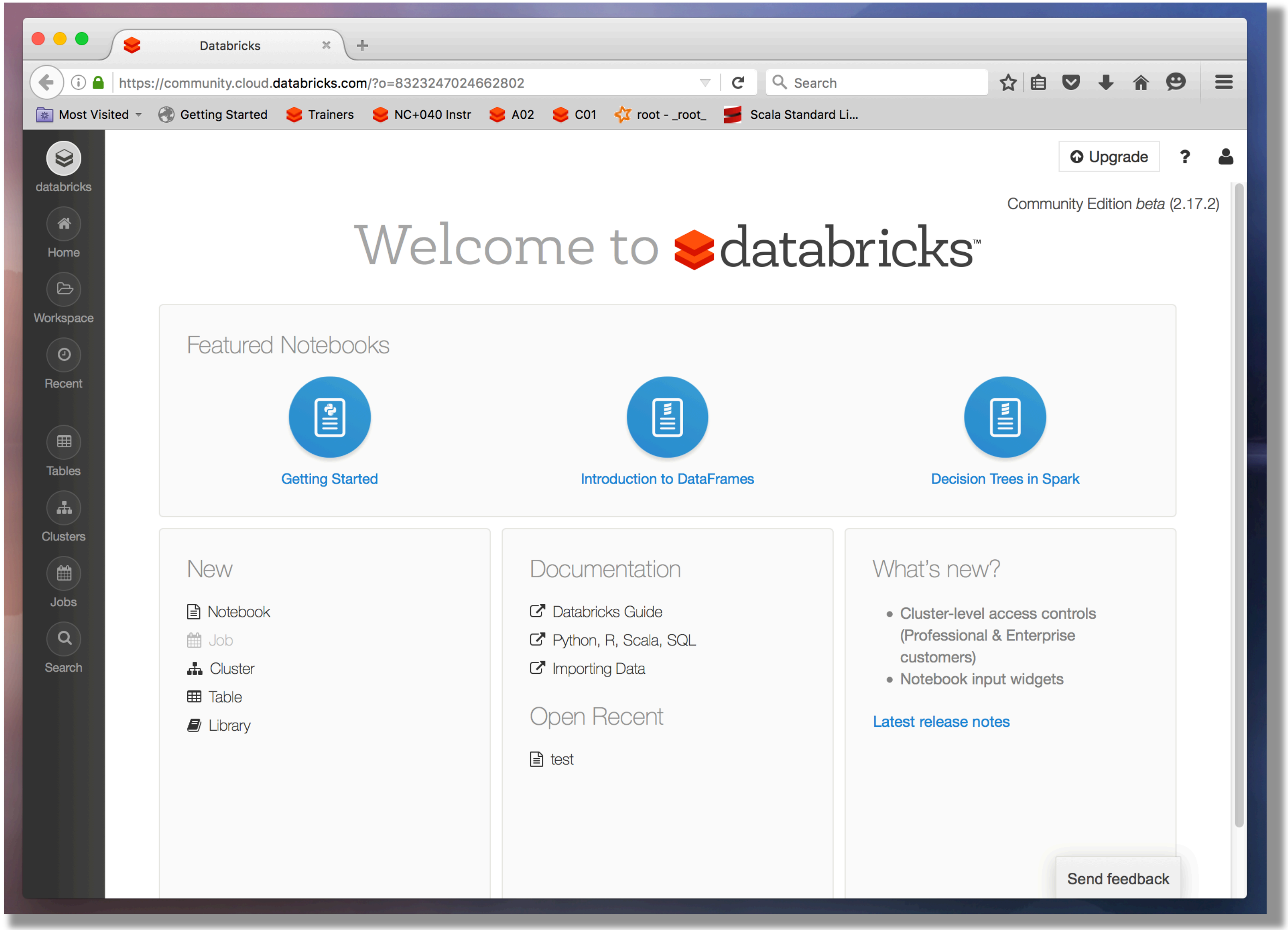
Documents

- Slides and files available at http://tinyurl.com/Spark-Wiki-Files
- Class Notes: http://tinyurl.com/Spark-Wiki-TW3

Databricks

- Databricks login at: https://ssw2016.cloud.databricks.com
  - Username distributed by email prior to class
  - Password is `BigDataSimple#01`
  - Please let us know ASAP if there is a problem
- Use a laptop with **Firefox** or **Chrome**
  - Internet Explorer / MS Edge not supported

Go Ahead and Log in to Databricks!



Community Edition *beta* (2.17.2)

Welcome to databricks™

Featured Notebooks

Getting Started

Introduction to DataFrames

Decision Trees in Spark

New

- Notebook
- Job
- Cluster
- Table
- Library

Documentation

- Databricks Guide
- Python, R, Scala, SQL
- Importing Data

Open Recent

- test

What's new?

- Cluster-level access controls (Professional & Enterprise customers)
- Notebook input widgets

Latest release notes

Send feedback

databricks™

End of Intro

databricks™