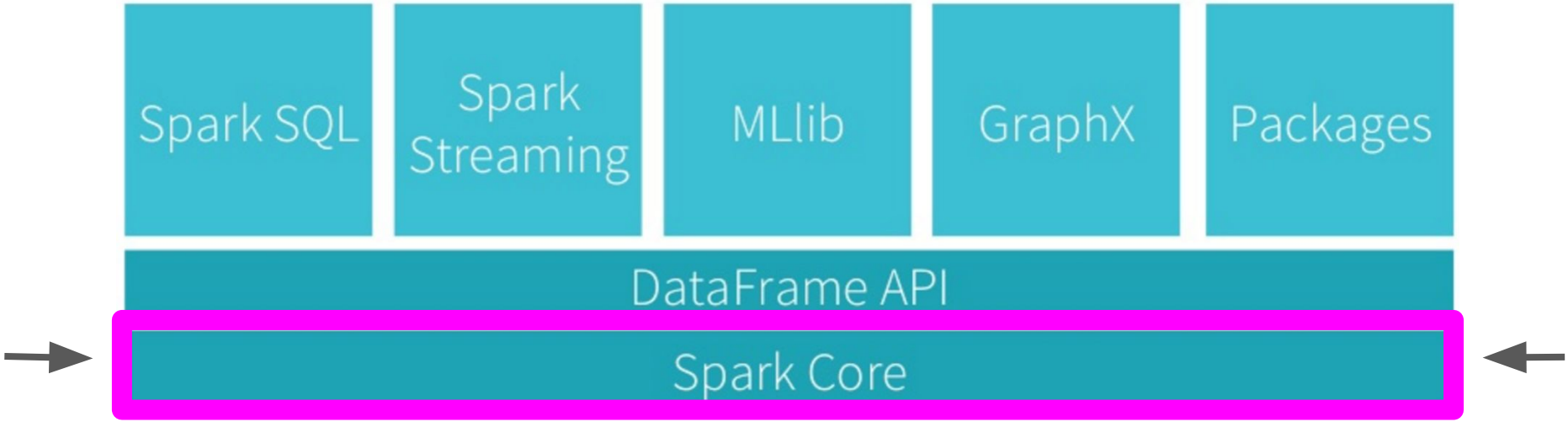
apache spark for everyone

amcasari + deb siegel

2016 May 05 - Seattle Spark Meetup

Jump Start Spark Part 2

Spark Core



How Does Spark Work?

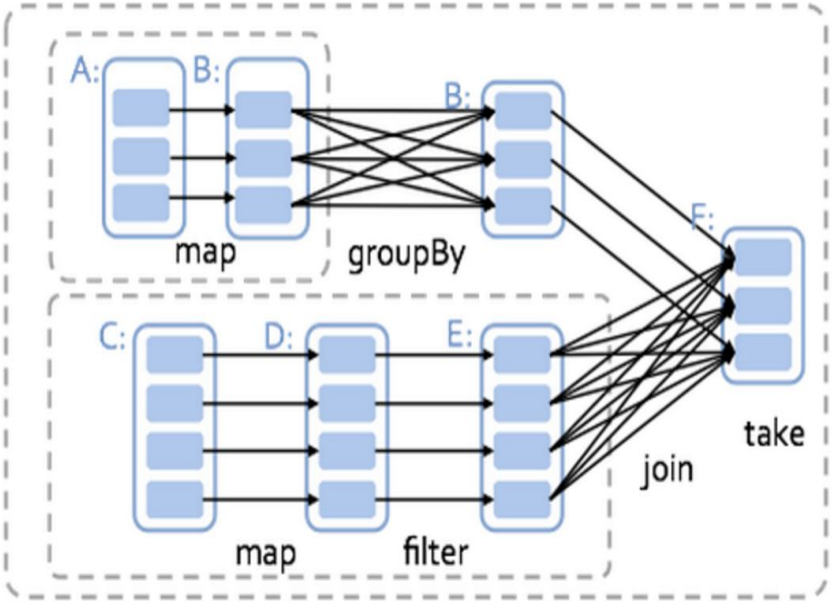
Basic Abstraction: Resilient Distributed Dataset (RDD)

RESILIENT

Persist with your chosen combination of memory, disk, serialization, replication.

Automatically recovered from DAG after node failures

DISTRIBUTED



DATASET

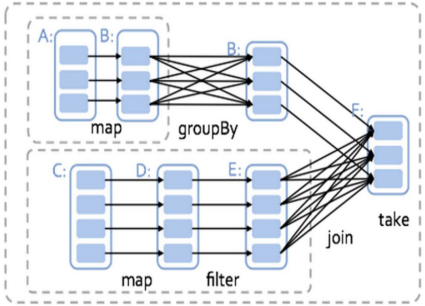
Key, Value : (“Seattle”, 23)

Case Class or Object

Vertices and Edges

Rows of a DataFrame or Table

Feature sets for iterative machine learning



How does Spark Work?

Core Functionality

TRANSFORMATIONS

NARROW

.textFile()
.map()
.filter()
.sample()
.cache()

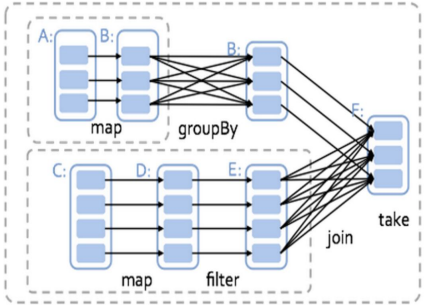
No Shuffle. Staged Together.

WIDE

.join()
.reduceByKey()
.groupByKey()
.aggregateByKey()

Shuffle. Staged Separately.

https://spark.apache.org/docs/latest/programming-guide.html#transformations



How does Spark Work?

Core Functionality

ACTIONS

TAKING

.take()
.count()
.reduce()
.collect()

Your driver needs to fit all the data you get back!

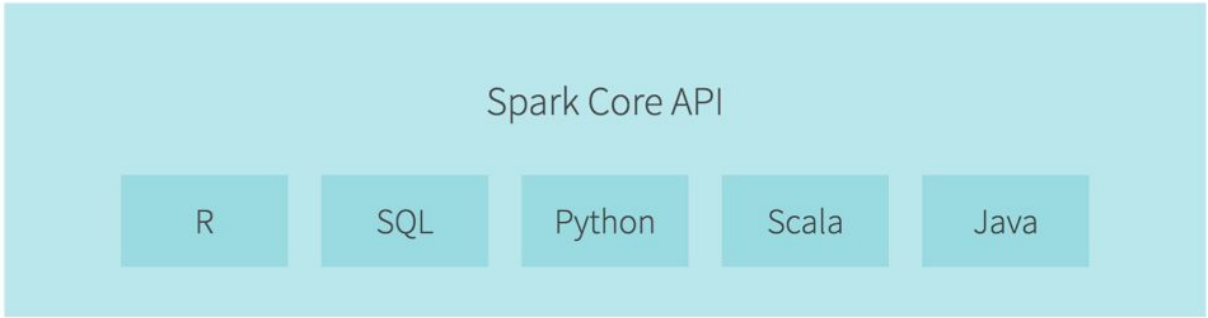
SAVING

.saveAsTextFile()
.saveAsSequenceFile()
.saveAsObjectFile()

Written as partitioned, driver does not need to fit the data.

https://spark.apache.org/docs/latest/programming-guide.html#actions

Language APIs



Scala

```
val textFile = sc.textFile("hdfs://...")
val counts = textFile.flatMap(line => line.split(" "))
             .map(word => (word, 1))
             .reduceByKey(_ + _)
```

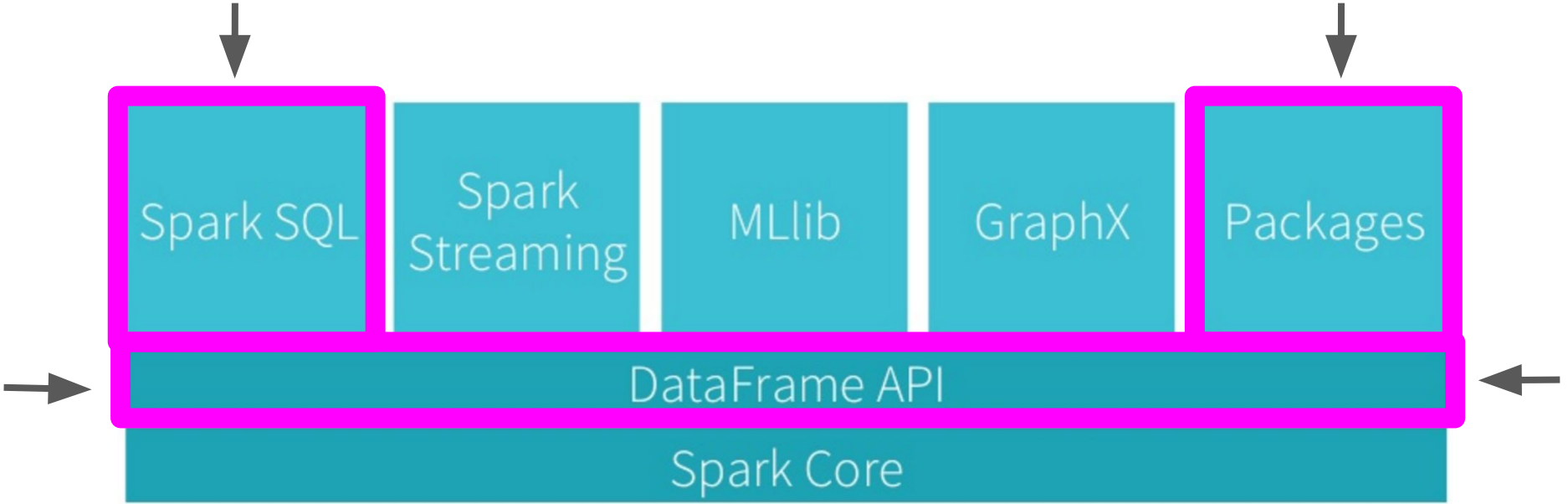
python

```
text_file = sc.textFile("hdfs://...")
counts = text_file.flatMap(lambda line: line.split(" ")) \
         .map(lambda word: (word, 1)) \
         .reduceByKey(lambda a, b: a + b)
```

SQL Catalyst Optimizer. Flattens speed differences. Helps enable language mixing.

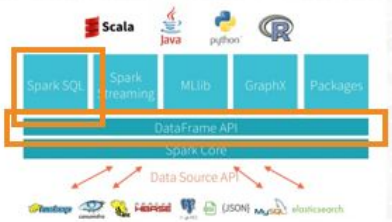
Jump Start Spark Part 3

DataFrames, SparkSQL, spark-packages



what is spark?

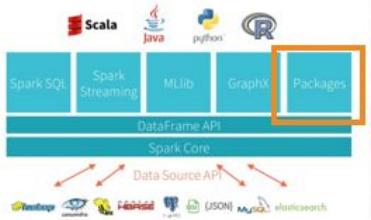
- Spark SQL allows you to query structured data in Spark programs either using SQL or DataFrames API
- can be used in applications + iterative workflows from a shell or notebook
- DataFrames API conceptually similar to a table in a relational database or data frame in R/Python
- preserves schema of original data for many file formats, including Parquet
- highly optimized, distributed collection of data
- Datasets: experimental interface (Scala + Java)



courtesy databricks

what is spark?

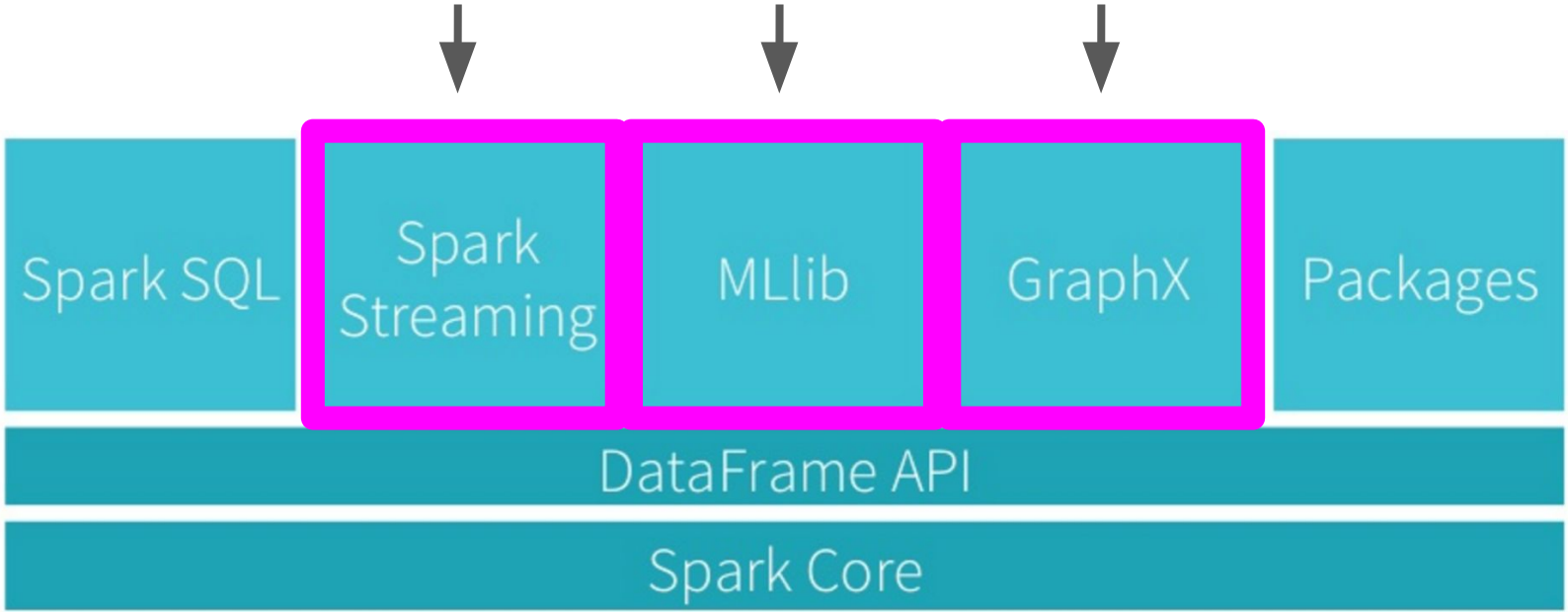
- spark-packages is a hosted module resource center for packages developed by the Spark community
- extends functionality + integration options for current Spark releases
- examples: spark-csv, spark-testing-base



courtesy databricks

Jump Start Spark Part 4

Modules



Modules

MLLIB

OPERATIONS: BASIC STATS, FEATURE TRANSFORMATIONS, ML
DATA: RDD

Local vectors (sparse or dense)
LabeledPoint (a local vector for supervised learning**)**
Distributed Matrix (backed by an RDD of its rows, where each row is a local vector)

val features:**RDD[Vector]**

val clusters: **KMeansModel** = **KMeans.train**(features, numClusters, numIterations)
val predictionRDD: **RDD[Int]** = **clusters.predict**(features)

http://spark.apache.org/docs/latest/mllib-guide.html

Modules

ML

OPERATIONS: PIPELINES, GRID SEARCH, CROSS VALIDATION

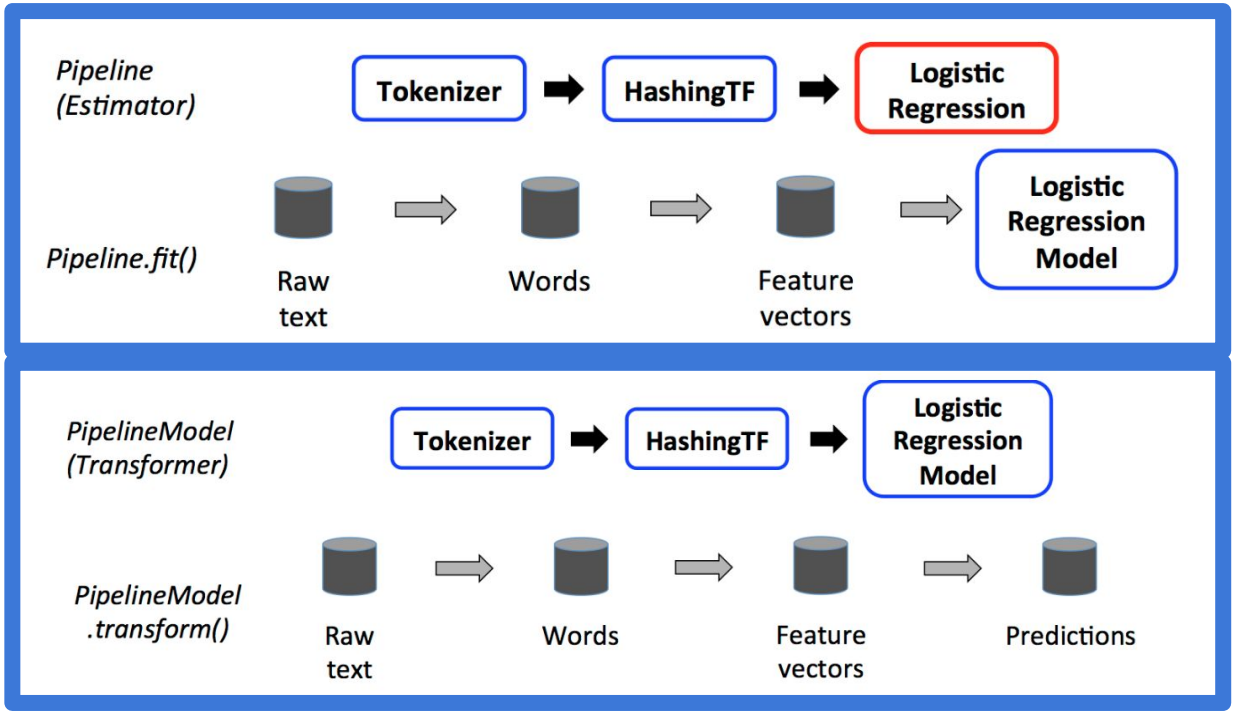
DATA : DATAFRAMES

build a pipeline

pipeline.fit() → **model**

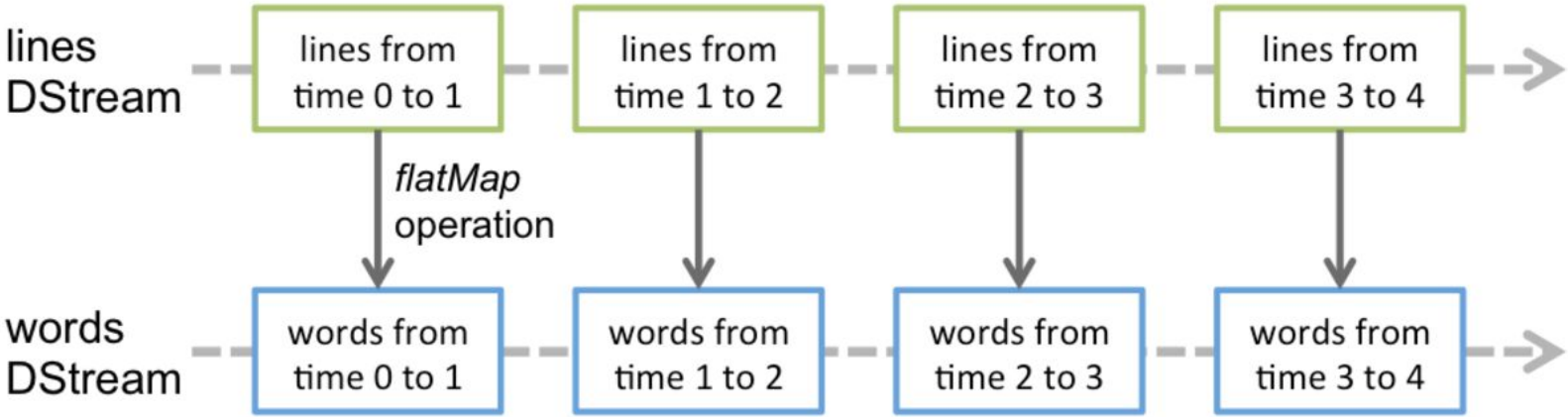
go through same pipeline

model.transform → **predictions**



Modules

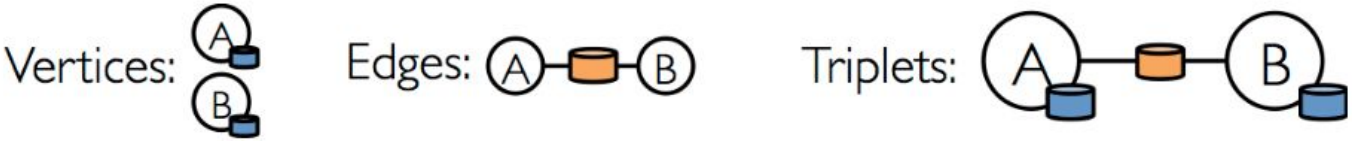
STREAMING



http://spark.apache.org/docs/latest/streaming-programming-guide.html

Modules

GRAPHX




```
class Graph[VD, ED] {
  val vertices: VertexRDD[VD]
  val edges: EdgeRDD[ED]
}
```

built in Graph operators!

http://spark.apache.org/docs/latest/graphx-programming-guide.html

resources

Learning Spark (book) (Holden Karau)
http://shop.oreilly.com/product/0636920028512

How to Tune your Spark Jobs: (Sandy Ryza
http://blog.cloudera.com/blog/2015/03/how-to-tune-your-apache-spark-jobs-part-2/
http://blog.cloudera.com/blog/2015/03/how-to-tune-your-apache-spark-jobs-part-1/

How to translate from mapreduce to apache spark (Sean Owen)
http://blog.cloudera.com/blog/2014/09/how-to-translate-from-mapreduce-to-apache-spark/

Advanced Spark Features (**Matei Zaharia))**
http://ampcamp.berkeley.edu/wp-content/uploads/2012/06/matei-zaharia-amp-camp-2012-advanced-spark.pdf

Tips for writing better spark jobs (Holden Karau and Vida Ha)
http://www.slideshare.net/databricks/strata-sj-everyday-im-shuffling-tips-for-writing-better-spark-programs

RDD function examples
http://homepage.cs.latrobe.edu.au/zhe/ZhenHeSparkRDDAPIExamples.html

A deeper understanding of Spark's internals (**Aaron Davidson)**
https://www.youtube.com/watch?v=dmL0N3qfSc8

Tuning and debugging in Apache Spark (Patrick Wendell, Feb 2015)
https://www.youtube.com/watch?v=kkOG_aJ9KjQ

GraphX (Text Rank) (Paco Nathan)
http://www.slideshare.net/pacoid/microservices-containers-and-machine-learning-50862677

Pyspark (Holden Karau)
http://www.slideshare.net/hkarau/a-really-really-fast-introduction-to-py-spark-lightning-fast-cluster-computing-with-python-1

you are not alone…

NEVER HAVE I FELT SO CLOSE TO ANOTHER SOUL

AND YET SO HELPLESSLY ALONE

AS WHEN I GOOGLE AN ERROR

AND THERE'S ONE RESULT

A THREAD BY SOMEONE WITH THE SAME PROBLEM

AND NO ANSWER

LAST POSTED TO IN 2003

courtesy xkcd