Analyzing Pwned Passwords with Spark

Kelley Robinson

@kelleyrobinson

Developer Evangelist

twilio

twilio

twilio +

Spark: then and now

The state of passwords

Spark in action

Big Data ∩ Security

APACHE Spark™

Gary Bernhardt
@garybernhardt

Following

@KELLEYROBINSON

An update to twitter.com/garybernhardt/ …! You can now have a x1.16xlarge EC2 instance (0.976 TiB of RAM) for $3.836/hour (reserved) or $6.669/hour (on demand). Your "big data" problems now fit in RAM, on-demand, at a cost of one latte per hour. Don't overengineer! Hardware is real!

2:44 PM - 20 Apr 2018

586 Retweets **1,238** Likes

21 586 1.2K

Apache Spark Ecosystem

Spark SQL + DataFrames

Streaming

MLlib
Machine Learning

GraphX
Graph Computation

Spark Core API

Java

R

Spark Abstractions

Then RDD (Resilient Distributed Dataset)

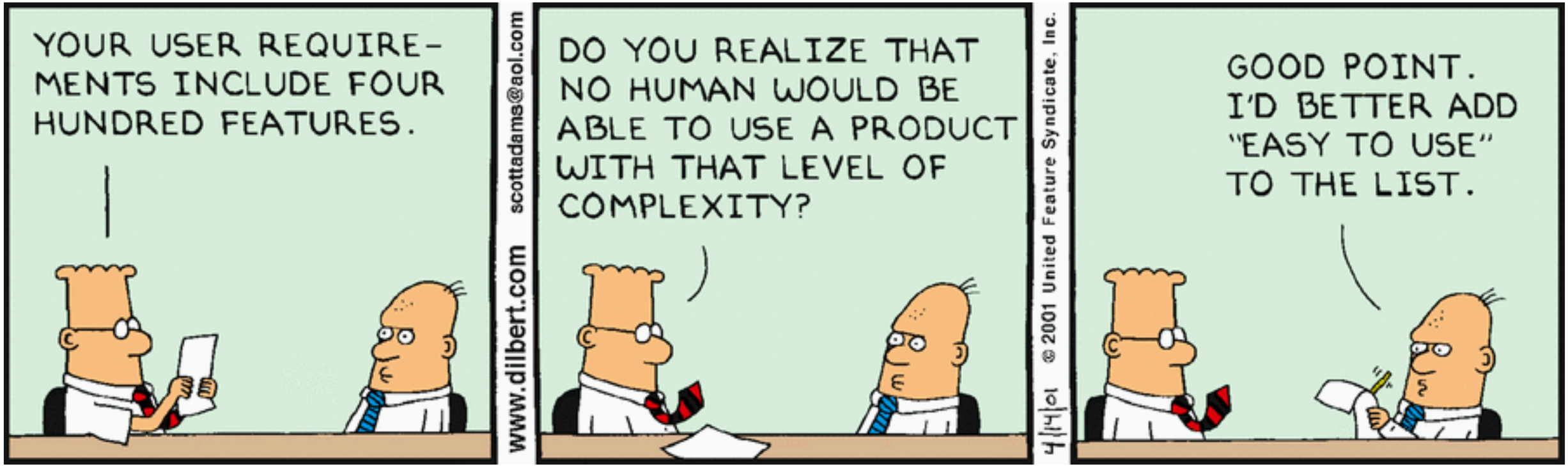
Now DataFrames / Datasets

RDDs

- Immutable & distributed collection
- Unstructured data
- Low-level transformation and control

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

https://databricks.com/blog/2016/07/14/a-tale-of-three-apache-spark-apis-rdds-dataframes-and-datasets.html


YOUR USER REQUIREMENTS INCLUDE FOUR HUNDRED FEATURES.
DO YOU REALIZE THAT NO HUMAN WOULD BE ABLE TO USE A PRODUCT WITH THAT LEVEL OF COMPLEXITY?
GOOD POINT. I'D BETTER ADD "EASY TO USE" TO THE LIST.
scottadams@aol.com
www.dilbert.com
© 2001 United Feature Syndicate, Inc.
4/14/01

Knowledgebase

1. Best Practices

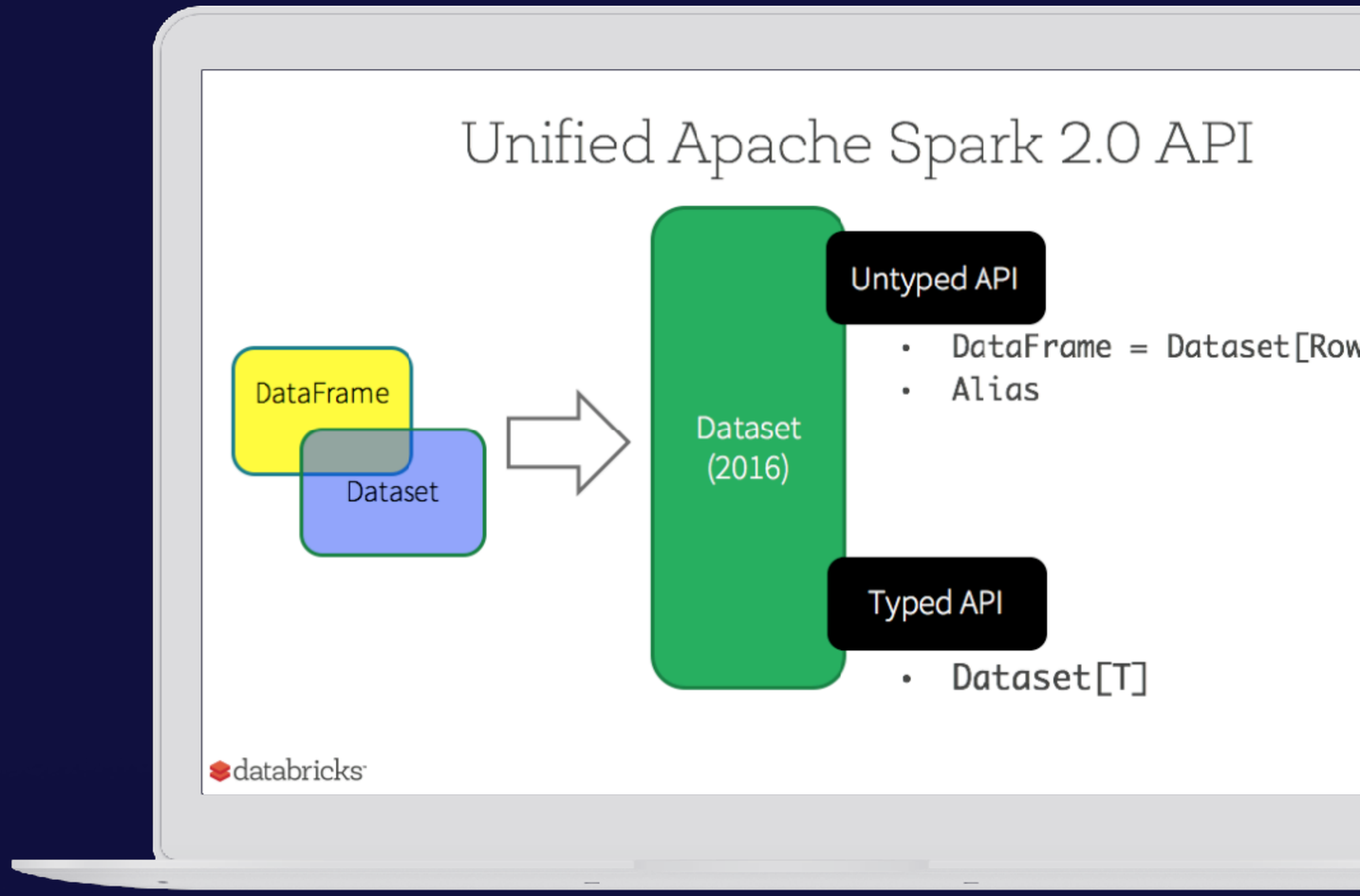
1.1. Avoid GroupByKey

Avoid GroupByKey

https://databricks.gitbooks.io/databricks-spark-knowledge-base/content/best_practices/prefer_reducebykey_over_groupbykey.html

Datasets

- Structured data
- Strongly typed
- Fast



https://databricks.com/blog/2016/07/14/a-tale-of-three-apache-spark-apis-rdds-dataframes-and-datasets.html

Datasets

- Structured data
- Strongly typed
- Fast
- SQL DSLs

```
val df = spark.read.json("logs.jsonl
df.groupBy("name")
  .count()
```

Apache Spark Ecosystem

Spark SQL + DataFrames

Streaming

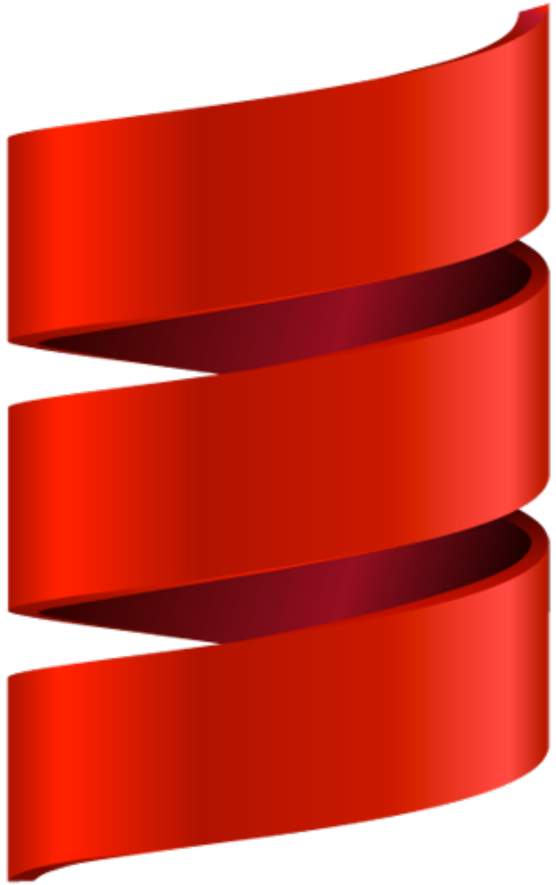
MLlib
Machine Learning

GraphX
Graph Computation

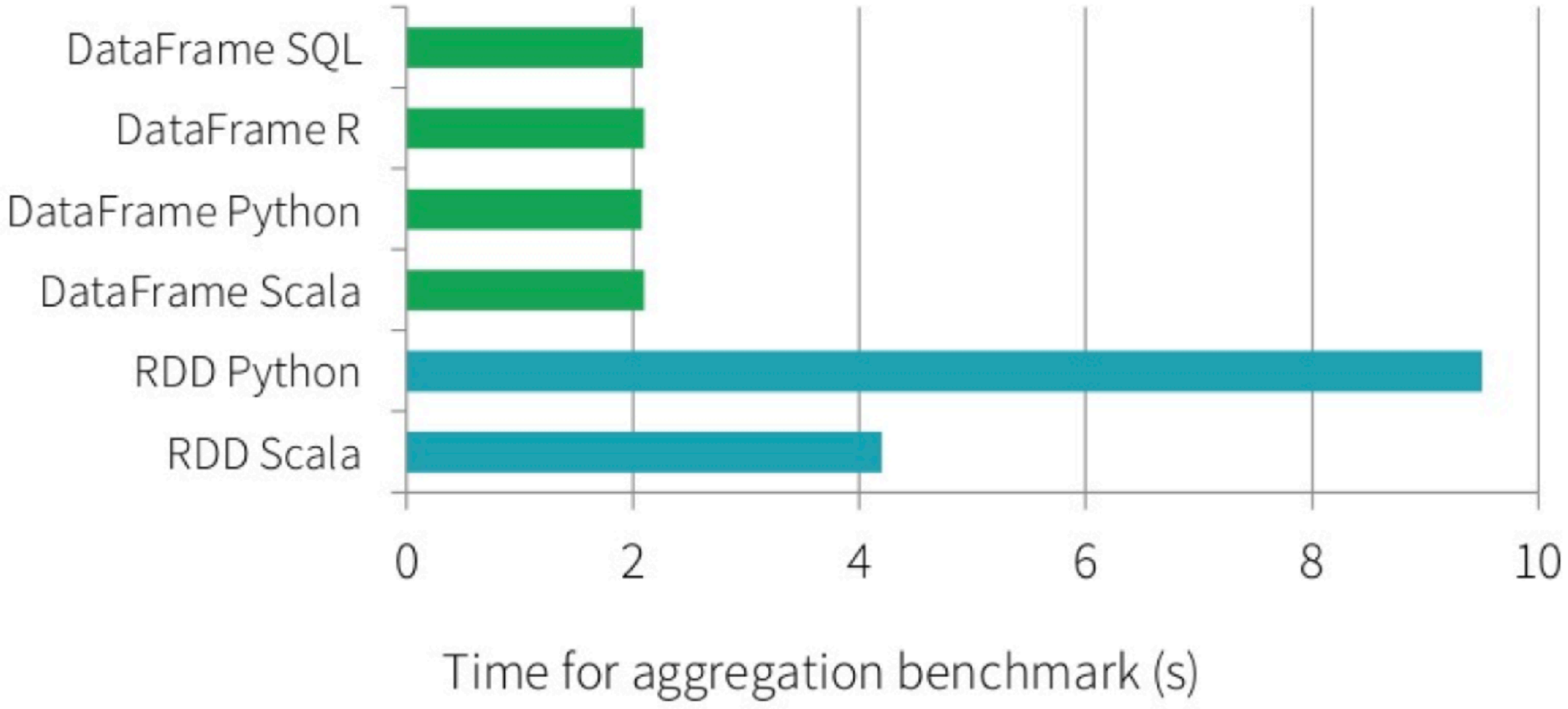
Spark Core API

Java

R



Scala has the most robust language API



https://www.slideshare.net/databricks/composable-parallel-processing-in-apache-spark-and-weld

Spark: then and now

The state of passwords

Spark in action

Big Data ∩ Security


HACKED


My
Password
123456
End

Pwned Passwords

Pwned Passwords are half a billion real world passwords previously exposed in data breaches. This exposure makes them unsuitable for ongoing use as they're at much greater risk of being used to take over other accounts. They're searchable online below as well as being downloadable for use in other online system. Read more about how HIBP protects the privacy of searched passwords.

123456 pwned?

Oh no — pwned!

This password has been seen 20,760,336 times before

This password has previously appeared in a data breach and should never be used. If you've ever used it anywhere before, change it!

Spark: then and now

The state of passwords

Spark in action

Big Data ∩ Security

```scala
case class PasswordHash(pwnedhash: String, count: Int)

val pwnedpasswords: Dataset[PasswordHash] = spark
  .read
  .text("./code/password-data/pwnedpasswords/*")
  .as[String]
  .map { line =>
    val Array(password, count) = line.trim.split(":")
    PasswordHash(password, count.toInt)
  }

pwnedpasswords.select($"pwnedhash").show()
```

```
|C9D1912B4B14C69B9...|
|C9D1916269845B485...|
|C9D194F192250C900...|
|C9D19CEAEEFE8F7D7...|
|C9D1A748AE8416A78...|
|C9D1A89B84E4F593F...|
|C9D1ACC093A47EAE1...|
```

WeRateDogs™
@dog_rates

Following

This is Kylo. Good boy levels off the h*ckin charts. 13/10 would use name in every password.

5:23 PM - 18 Apr 2018

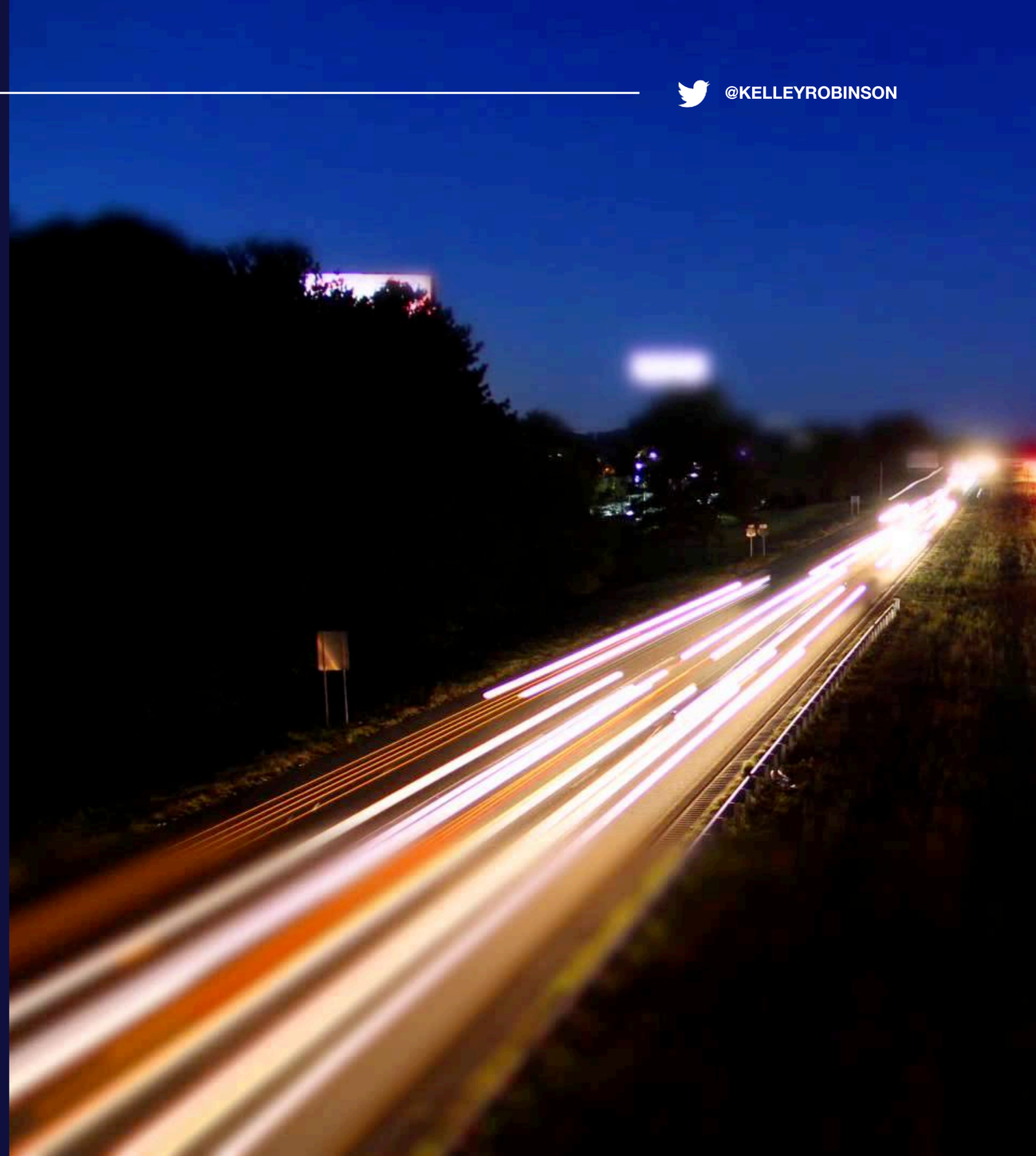
17,085 Retweets **109,408** Likes

486 17K 109K

https://twitter.com/dog_rates/status/986762231290490881

Benefits

- Fast
- Flexible
- Good for exploration
- Proven for large systems



Challenges

- Opaque error messages
- Operationalizing
- Documentation

http://heather.miller.am/blog/launching-a-spark-cluster-part-1.html

```
mapped: org.apache.spark.sql.DataFrame = [plainpassword: string, hashed: stri
joined: org.apache.spark.sql.DataFrame = [plainpassword: string, pwnedhash: s
org.apache.spark.SparkException: Job aborted due to stage failure: Task 16 in
e: Lost task 16.0 in stage 383.0 (TID 6083, localhost, executor driver): scal
 class [Ljava.lang.String;)
        at $anonfun$1.apply(<console>:58)
        at $anonfun$1.apply(<console>:57)
        at org.apache.spark.sql.catalyst.expressions.GeneratedClass$Generated
        at org.apache.spark.sql.execution.BufferedRowIterator.hasNext(Buffere
        at org.apache.spark.sql.execution.WholeStageCodegenExec$$anonfun$8$$a
        at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:408)
        at org.apache.spark.shuffle.sort.BypassMergeSortShuffleWriter.write(B
        at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.s
        at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.s
        at org.apache.spark.scheduler.Task.run(Task.scala:99)
        at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:2
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecut
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecu
```

Mastering Apache Spark (2.3.0)

Welcome to **Mastering Apache Spark** gitbook! I'm very excited to hav

I'm Jacek Laskowski, an independent consultant, software developer

The missing Spark documentation

https://jaceklaskowski.gitbooks.io/mastering-apache-spark/content/

Spark: then and now

The state of passwords

Spark in action

Big Data ∩ Security


LETS JUST COLLECT MILLIONS OF SOCIAL SECURITY NUMBERS INTO THIS DATABASE HERE
AAAAAAND WE'VE BEEN HACKED

@KELLEYROBINSON

Eric™ @Korni22 · Apr 6

Well, what if your infrastructure gets breached and everyone's password is published in plaintext to the whole wide world?

5 89 1.8K

T-Mobile Austria

@tmobileat

Follow

Replying to @Korni22 @c_pellegrino and 2 others

@Korni22 What if this doesn't happen because our security is amazingly good? ^Käthe

2:27 AM - 6 Apr 2018

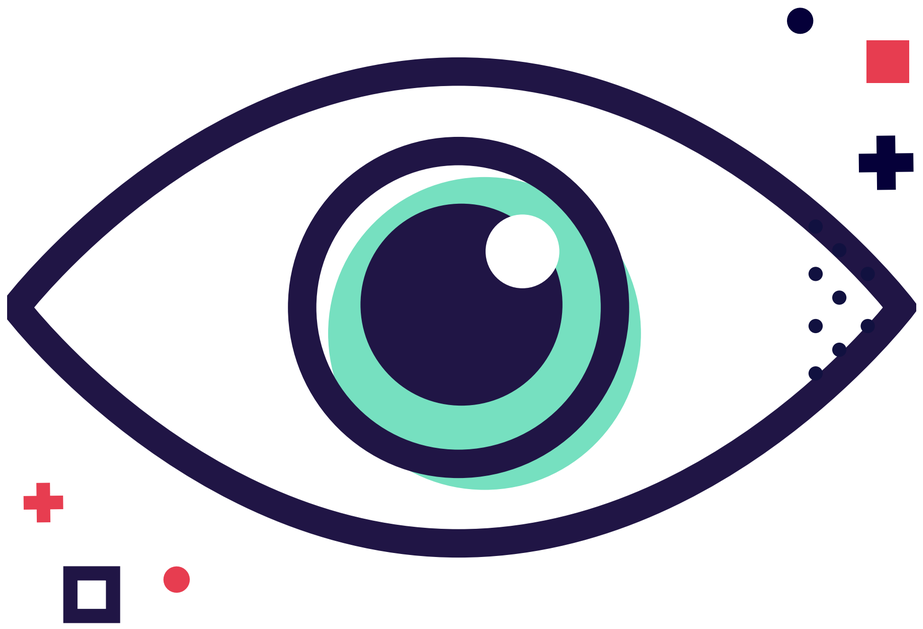
704 Retweets **915** Likes

461 704 915

LastPass

1Password


twilio
THANK YOU!
@kelleyrobinson



Spark Resources

- Apache Spark
- Jacek's Spark Documentation
- Zeppelin
- RDDs vs. Datasets
- Running Spark on a Cluster

Security Resources

- Pwned Passwords
- Reverse SHA1 hashes
- LastPass and 1Password
- 2FA Guides

Zeppelin Notebook Job Search your Notes admin

Analyzing pwned passwords