



the INTERNET of EVERYWHERE

how The Weather Company scales
Robbie Strickland



the INTERNET of EVERYWHERE

how The Weather Company scales
Robbie Strickland



Everywhere Defined

- 26B forecasts /day or *250,000/second*
 - vs *3.5B Google queries daily*
- 2.2 billion unique locations
- 200k personal weather stations
- 200M active mobile users
- Petabytes of data generated daily

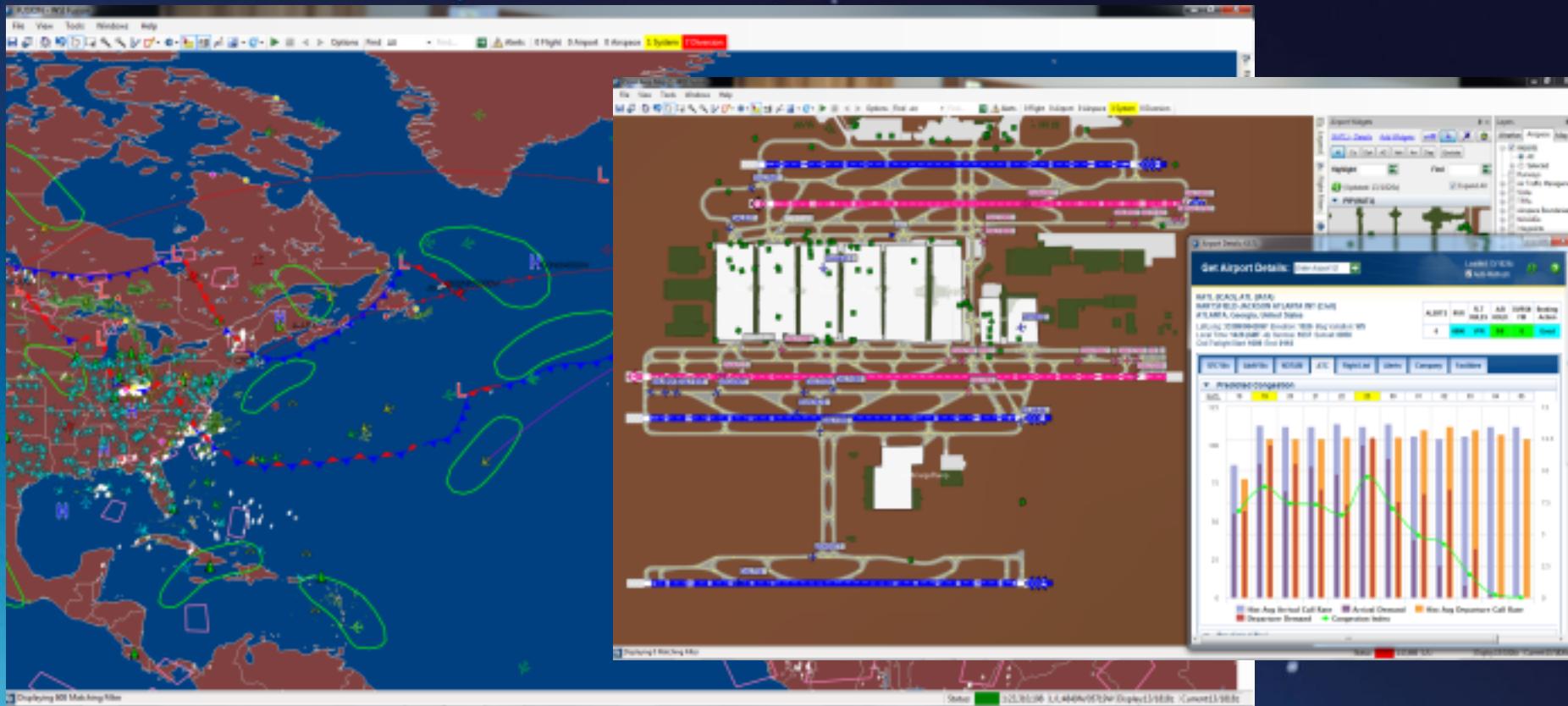
Our Brands

The
Weather
Channel

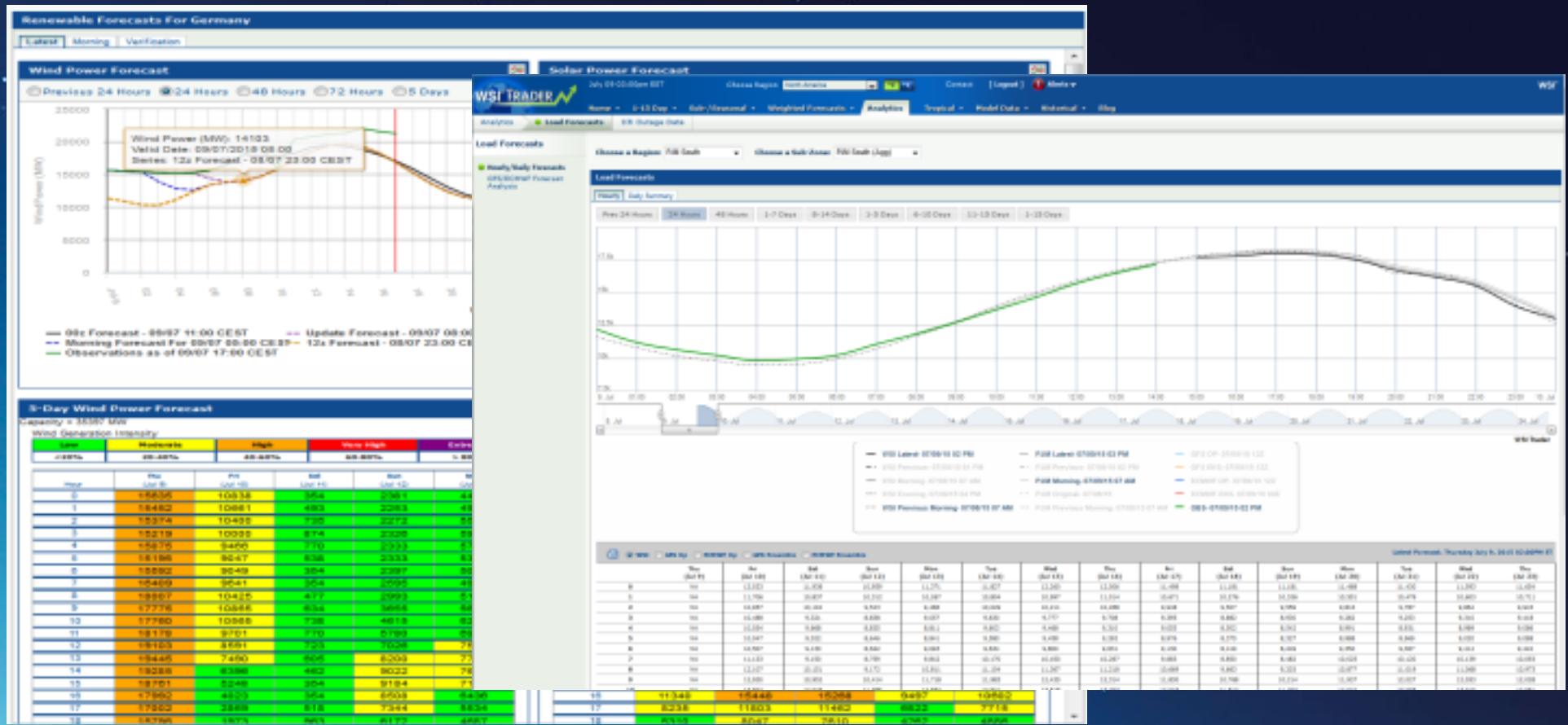


Over
30 Billion
Served

Flight Routing



Energy Trading



Insurance

WeatherFX Response

search for address...

Create Alert

Population

Settings

Feed Overview Social Replay

Damaging Winds

Flood

Hail

Ice & Snow

Severe

Tornado

Tropical Cyclone

Rain

Mix

Snow

Colorado

Kansas

Missouri

Oklahoma

New Mexico

Arkansas

Texas

Thu 11:30 am

Start: Wednesday 11:15 am

Population: 184,211

Hail Greater Than 1" Past 24 Hours

Leaflet, terms, privacy, data providers

	Current Hail	Predicted
Louisiana	Hail Greater Than 1" Past 24 Hours	
Oklahoma		
Texas	184,211	
Arkansas		
New England	>	
Mid Atlantic	>	
West Coast	>	

Weather Alerting

4:32

Monday, April 27



Real-Time Rain Alert

4:32 PM

Rain will begin around 4:49pm, continuing off and on over next half hour. The rain will be light.

11:25

Monday, April 27



The Weather now

Lightning struck at 11:25pm, 10 miles NE of your current location. Make your move to safety now.

slide to view

Decisions at Scale

101001110100101



101001110100101
010100101011001
101010101011100
000011010110010



Who Are You?



RDBMS

→ **Spark** ?

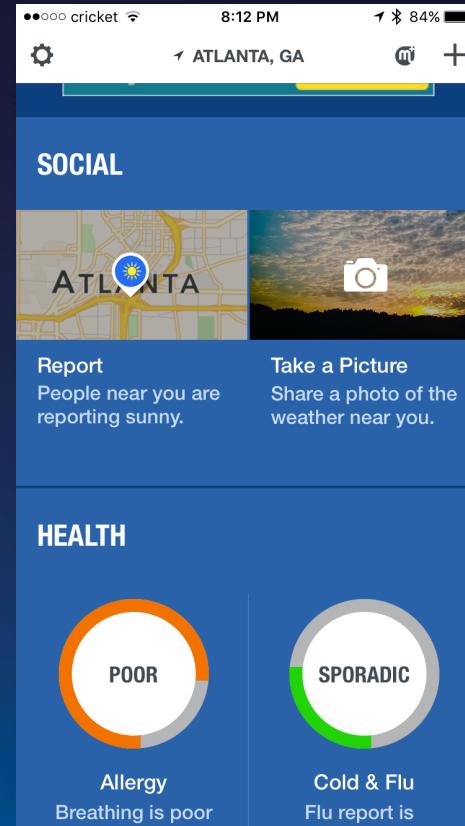
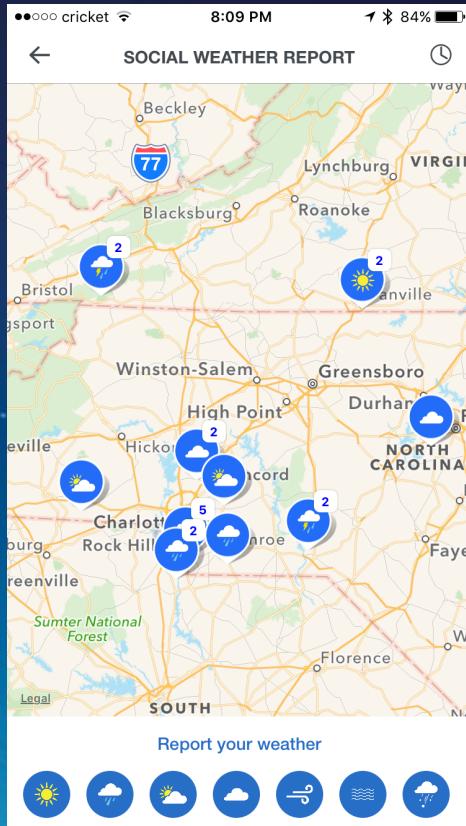
Who Are You?



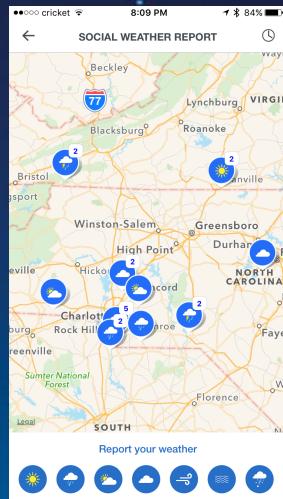
Spark
?

The word "Spark" is written in a large, white, lowercase, sans-serif font. A five-pointed star is positioned above the letter "k". Below "Spark", there is a large white question mark.

Social Weather

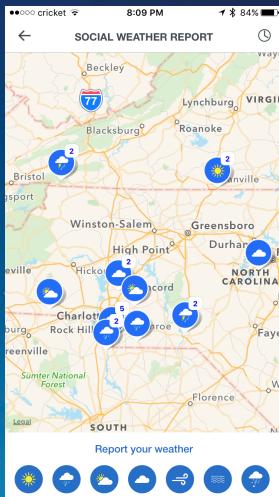


Social Weather



RDBMS

Social Weather

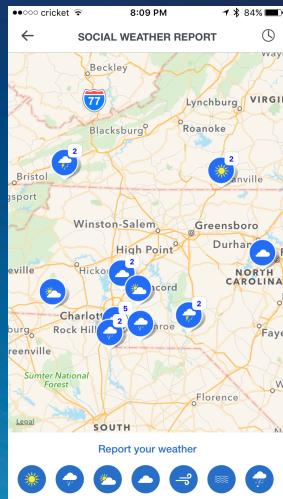


RDBMS

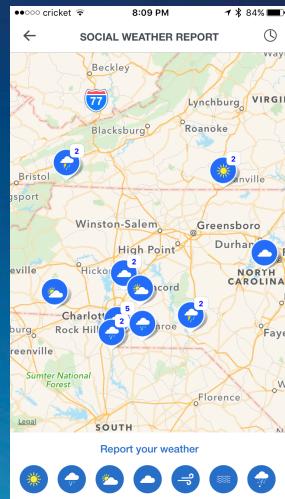


```
SELECT count(*) FROM wx_reports  
GROUP BY time / 300000 * 300000
```

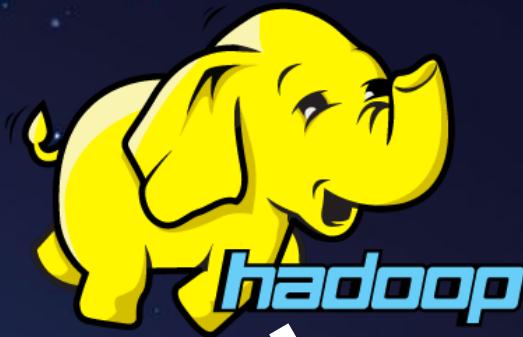
Social Weather



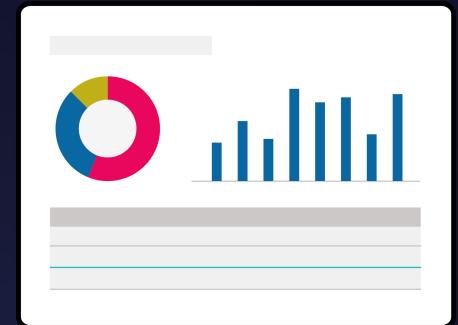
Social Weather



Sqoop

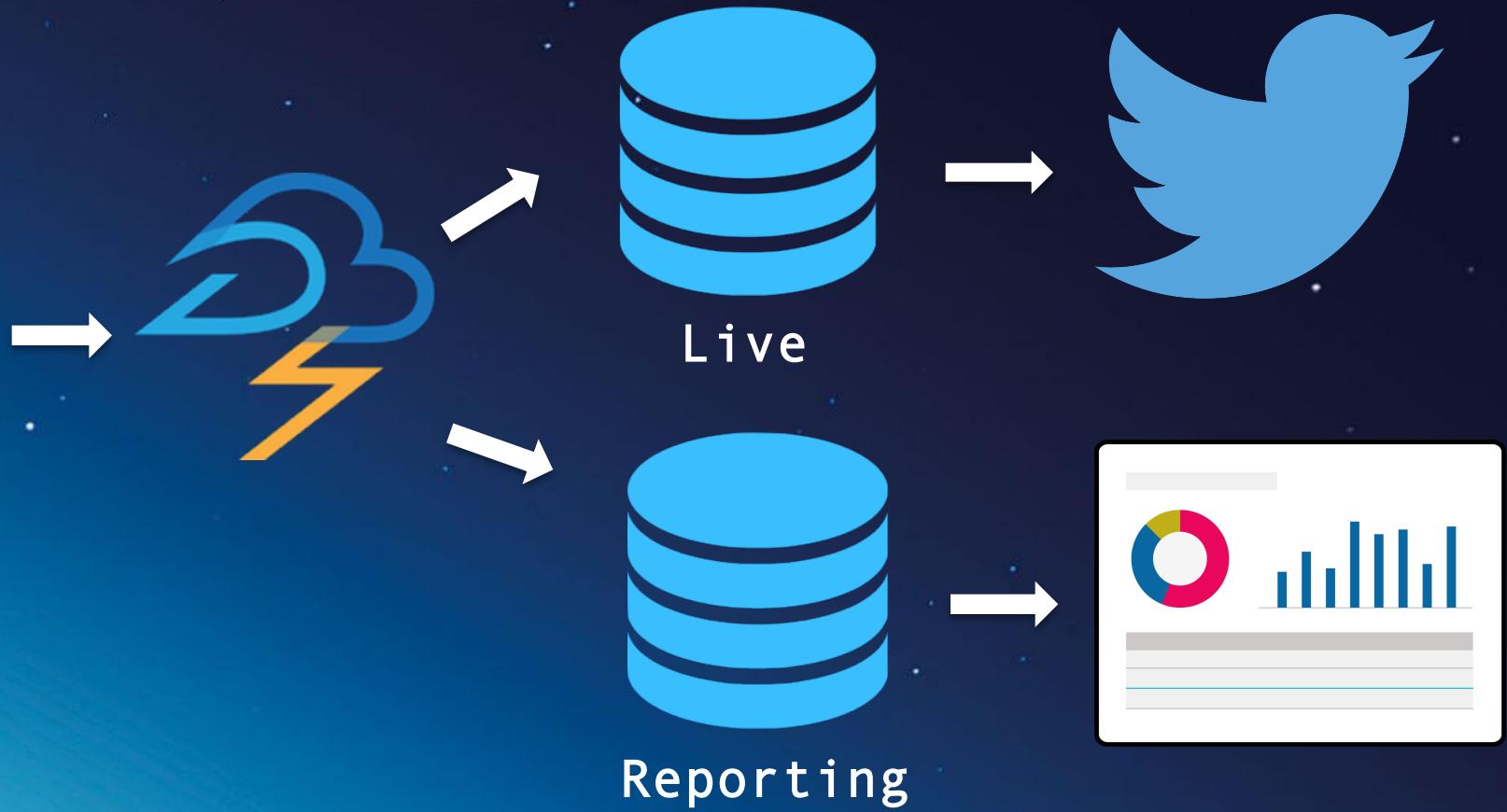
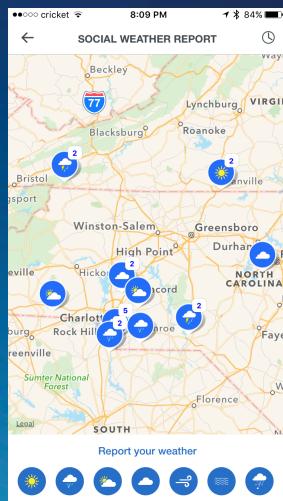


M/R



Reporting

Scaling with Spark



Easing the Transition

101001110100101



101001110100101
010100101011001
101010101011100
000011010110010



Easing the Transition

101001110100101



101001110100101
010100101011001
101010101011100
000011010110010

Easing the Transition

101001110100101



101001110100101
010100101011001
101010101011100
000011010110010

101001110100101
010100101011001
101010101011100
000011010110010



10100,11101,00101
01010,01010,11001
10101,01010,11100
00001,10101,...

Easing the Transition

101001110100101

101001110100101

010100101011001

101010101011100

000011010110010



101001110100101

010100101011001

101010101011100

00011010110010

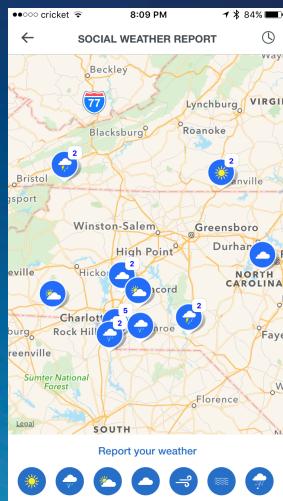
0100,11101,00101

01010,01010,11001

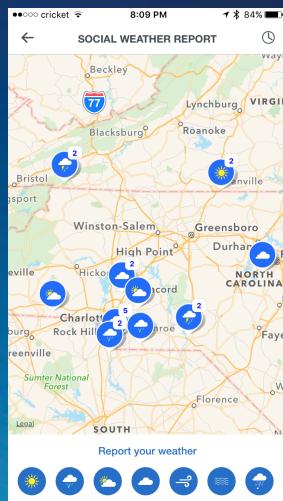
10101,01010,11100

00001,10101,...

Scaling with Spark



Scaling with Spark



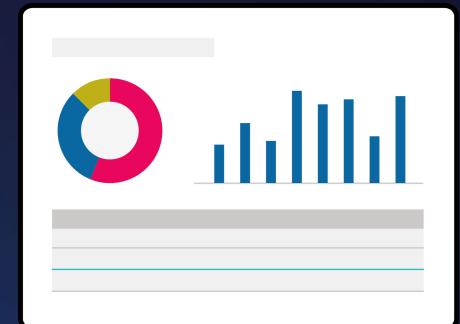
→ **Spark** *



Live



Reporting



Batch Aggregation

```
val wx_reports = // load data from database

val sql = new org.apache.spark.sql.SQLContext(sc)
import sql.implicits._

wx_reports.toDF.registerTempTable("wx_reports")

val counts = sql("select count(*) from wx_reports group by
timestamp / 300000 * 300000")
```

Streaming Aggregation

```
val wx_reports = // load from streaming source

wx_reports.foreachRDD { rdd =>
  val sql = SQLContext.getOrCreate(rdd.sparkContext)
  import sql.implicits._
  rdd.toDF.registerTempTable("wx_reports")
  val count = sql("select count(*) from wx_reports")
}
```

Data Science Roles



Data Scientist



Data Engineer

Data Science Roles



Data Scientist

Machine learning expert



Data Engineer

Data Science Roles



Data Scientist

Machine learning expert



Data Engineer

Scalable algorithms expert

Data Science Roles



Data Scientist

Builds pipelines that
work on her laptop



Data Engineer

Data Science Roles



Data Scientist



Data Engineer

Rewrites her pipelines
to scale better

Collaborative Data Science

Jupyter Spark - Scala Test Last Checkpoint: 12/07/2015 (autosaved)

File Edit View Insert Cell Kernel Help

Cell Toolbar: None

Control Panel Logout

Spark 1.4.1 (Scala 2.10.4)

In []:

```
import com.datastax.spark.connector._  
import org.apache.spark._  
  
case class LUE(userid: String, timestamp:  
  
val sqlContext = kernel.sqlContext  
import sqlContext.implicits._  
sc.cassandraTable[LUE]("prod_analytics_e
```

In []:

```
sqlContext.sql("select * from lues limit
```

In []:

```
%sparkr  
sqlContext <- sparkRSQl.init(sc)  
df <- sql(sqlContext, "select * from lue  
showDF(df)
```

In []:

```
%sql select * from lues limit 10
```

In []:

```
%lsmagic
```

In []:

```
%showtypes
```

In []:

```
%%html  
<h1>Hello</h1>  
<p>World</p>
```

In []:

```
%adddeps com.databricks spark-csv_2.10 1
```

In []:

```
import org.apache.spark._  
val sqlContext = kernel.sqlContext  
import sqlContext.implicits._
```

Zeppelin Notebook Interpreter Connected

Verifying Akamai Aggregation

val fn = "s3a://twc-prod-akamai-log-analysis/cleansed.akamai.parquet/file=dsx_prod/logDate=201512??"
sqlContext.read.parquet(fn).registerTempTable("dsxakamai")
fn: String = s3a://twc-prod-akamai-log-analysis/cleansed.akamai.parquet/file=dsx_prod/logDate=201512??
Took 89 seconds.

FINISHED

%sql
select logDate, count(*) as bounces
from dsxakamai
where httpVerb = 'POST'
and requestURI like '/dsx.weather.com/subs/bounce%'
group by logDate
order by logDate

FINISHED

201,137.00
150,000.00
100,000.00
50,000.00
0.00
20151201 20151204 20151206 20151208 20151210 20151212 20151214 20151216 20151218 20151220 20151222 20151224 20151226 20151228 20151231

● Stacked ○ Stream ○ Expanded

bounces

The Analytics OS

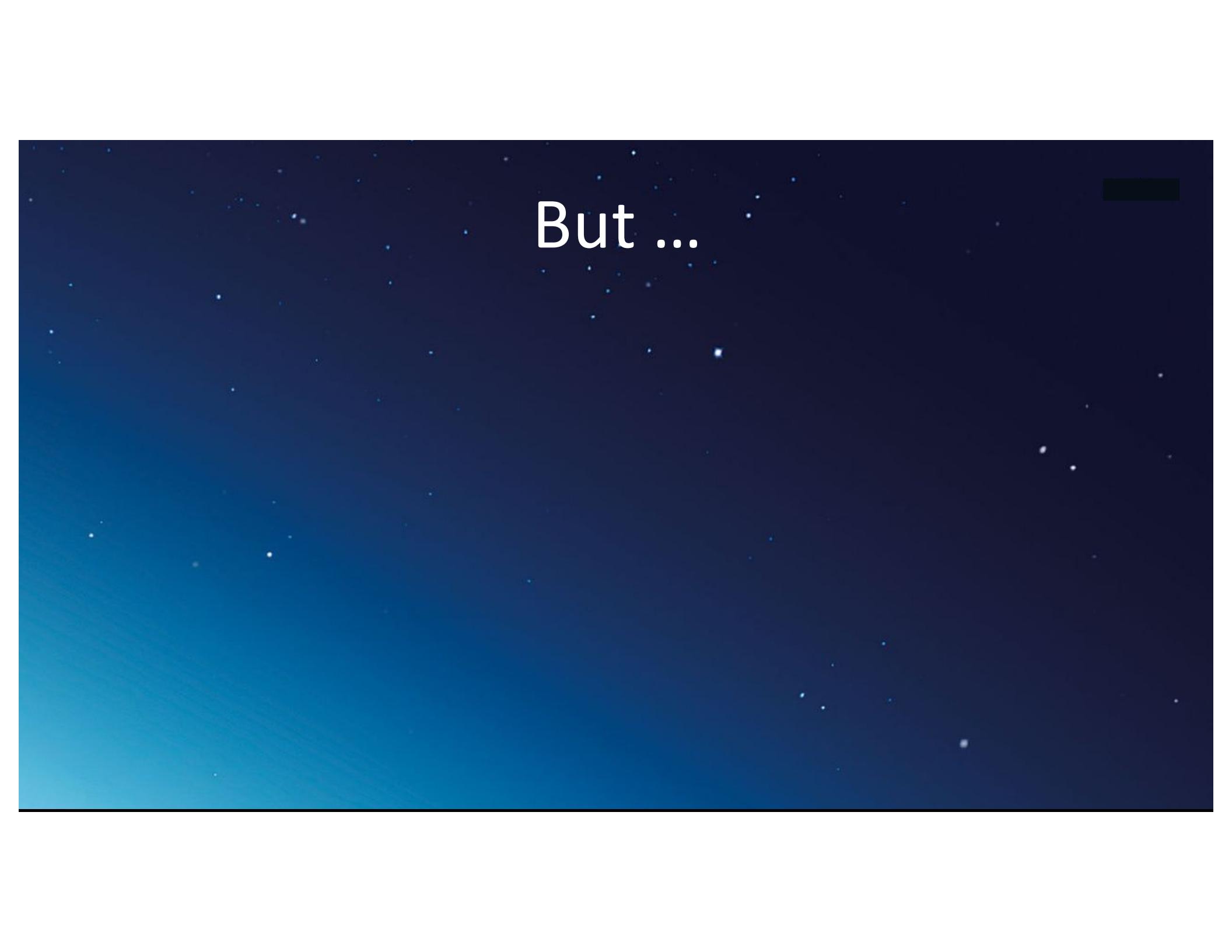


Notebooks

Stream
Analytics

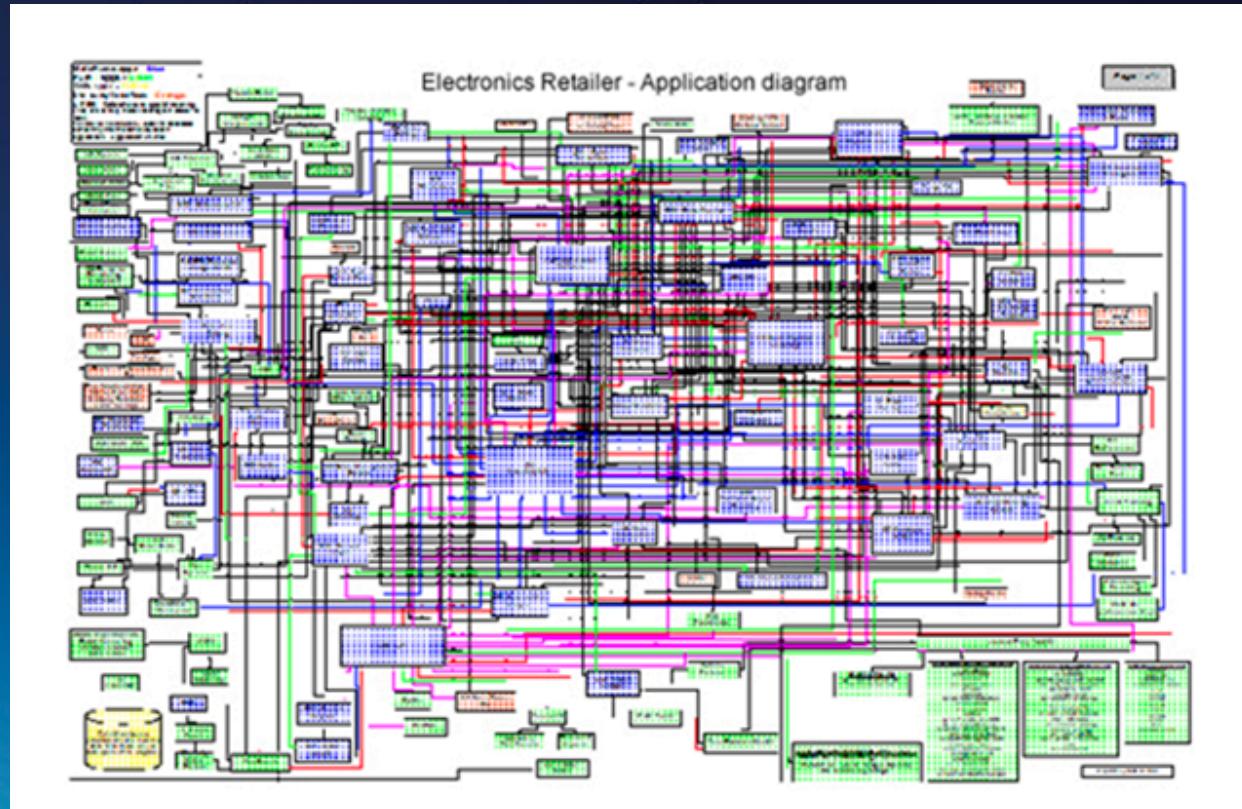
Batch
Analytics

Spark 

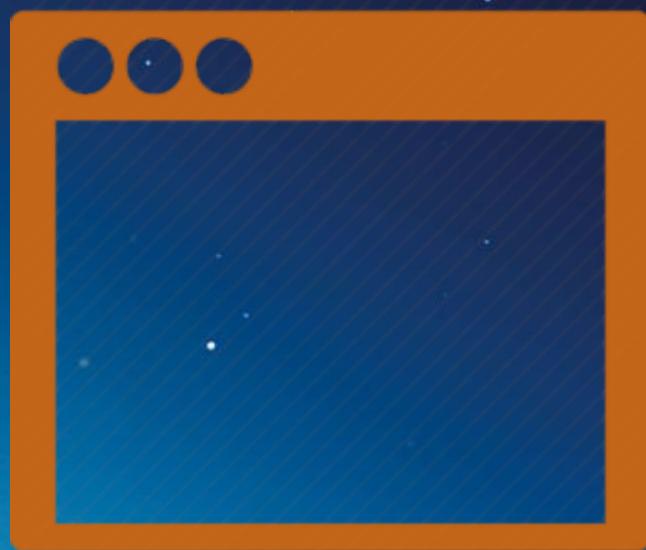


But ...

The Real World (Enterprise Version)



The Real World (Startup Version)



Application

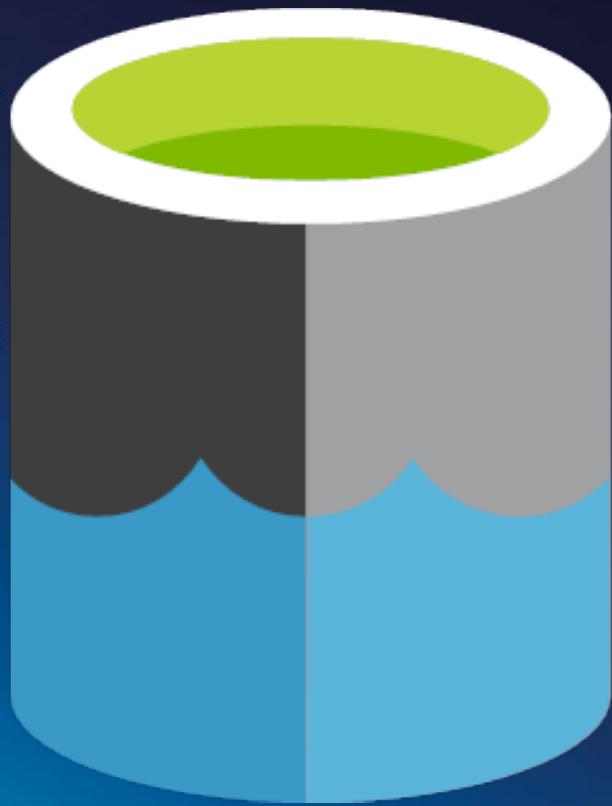


MySQL

Step 1: Pick a Problem to Solve



Step 2: Build a Data Lake





Step 3: Set up Spark

- Direct download
- Hadoop distribution
(Hortonworks, Cloudera, etc)
- Managed service (Elastic
MapReduce, Databricks,
BlueMix, etc)

Step 4: Start Collecting Data

- Options:
 - Sqoop to move RDBMS tables
 - Flume/FluentD to move logs
 - Import from Spark-supported data sources
 - Using Spark Streaming attached to a queue
 - ...

Step 5: Use a Notebook

Jupyter Spark - Scala Test Last Checkpoint: 12/07/2015 (autosaved)

File Edit View Insert Cell Kernel Help

Control Panel Logout

Spark 1.4.1 (Scala 2.10.4)

In []:

```
import com.datastax.spark.connector._  
import org.apache.spark._  
  
case class LUE(userid: String, timestamp:  
  
val sqlContext = kernel.sqlContext  
import sqlContext.implicits._  
sc.cassandraTable[LUE]("prod_analytics_e
```

In []:

```
sqlContext.sql("select * from lues limit
```

In []:

```
%sparkr  
sqlContext <- sparkRSQl.init(sc)  
df <- sql(sqlContext, "select * from lue  
showDF(df)
```

In []:

```
%sql select * from lues limit 10
```

In []:

```
%lsmagic
```

In []:

```
%showtypes
```

In []:

```
%%html  
<h1>Hello</h1>  
<p>World</p>
```

In []:

```
%adddeps com.databricks spark-csv_2.10 1
```

In []:

```
import org.apache.spark._  
val sqlContext = kernel.sqlContext  
import sqlContext.implicits._
```

Zeppelin Notebook Interpreter Connected

Verifying Akamai Aggregation

val fn = "s3a://twc-prod-akamai-log-analysis/cleansed.akamai.parquet/file=dsx_prod/logDate=201512??"
sqlContext.read.parquet(fn).registerTempTable("dsxakamai")
fn: String = s3a://twc-prod-akamai-log-analysis/cleansed.akamai.parquet/file=dsx_prod/logDate=201512??
Took 89 seconds.

FINISHED

%sql
select logDate, count(*) as bounces
from dsxakamai
where httpVerb = 'POST'
and requestURI like '/dsx.weather.com/subs/bounce%'
group by logDate
order by logDate

FINISHED

201,137.00
150,000.00
100,000.00
50,000.00
0.00
20151201 20151204 20151206 20151208 20151210 20151212 20151214 20151216 20151218 20151220 20151222 20151224 20151226 20151228 20151231

● Stacked ○ Stream ○ Expanded

bounces

Final Thoughts



Thank You!



Robbie Strickland
@rs_atl

(we're hiring!)