Taking Jupyter Notebooks and Apache Spark to the next level with PixieDust

David Taieb
Distinguished Engineer
IBM Watson Data Platform, Developer Advocacy

@DTAIEB55

WHY ARE YOU HERE?

- More companies making bet-the-business data driven decisions
  - Good news: they are drowning in Data
  - Bad news: they are drowning in Data
- Solving the Data problems of tomorrow cannot be done by data scientists alone.
- Developers are getting more involved with Data Science, moving from stovepipe applications to “data pipelines” that integrate data and analytics.

How do we blur the lines between developers and data scientists?

Let’s start with a story… we all know too well.

Disclaimer: All characters and events depicted in this story are entirely fictitious. Any similarity to actual use cases, events or persons is actually intentional.

MEET BEN

THE DEVELOPER

- Hold a master degree in computer science
- 10 year experience, 6 years with the company
- Languages of choice: Java, Node.js, HTML5/CSS3
- Data: No SQL (Cloudant, Mongo), relational
- No major experience with Big Data

“The best line of code is the one I didn't have to write!”

MEET NATASHA

THE DATA SCIENTIST

- Hold a PHD in data science
- 5 year experience, 2 years with the company
- Experienced in Python and R
- Expert in Machine Learning and Data visualization
- Software engineering is not her thing

"In God we trust. All others bring data."

— W. Edwards Deming


SURPRISE MEETING
With the VP of Development
“We have an urgent need for our marketing department to build an application that can provide real-time sentiment analysis on Twitter data.”
https://unsplash.com/search/meeting?photo=3fPXt37X6UQ
@DTAIEB55

KEY CONSTRAINTS

- You only have 6 weeks to build the application
- Target consumer is the business-focused user
  - Must be easy to use even for non technical people
- It must scale out of the box
  - I want you to look at Apache Spark

SOME LEARNING TO DO…

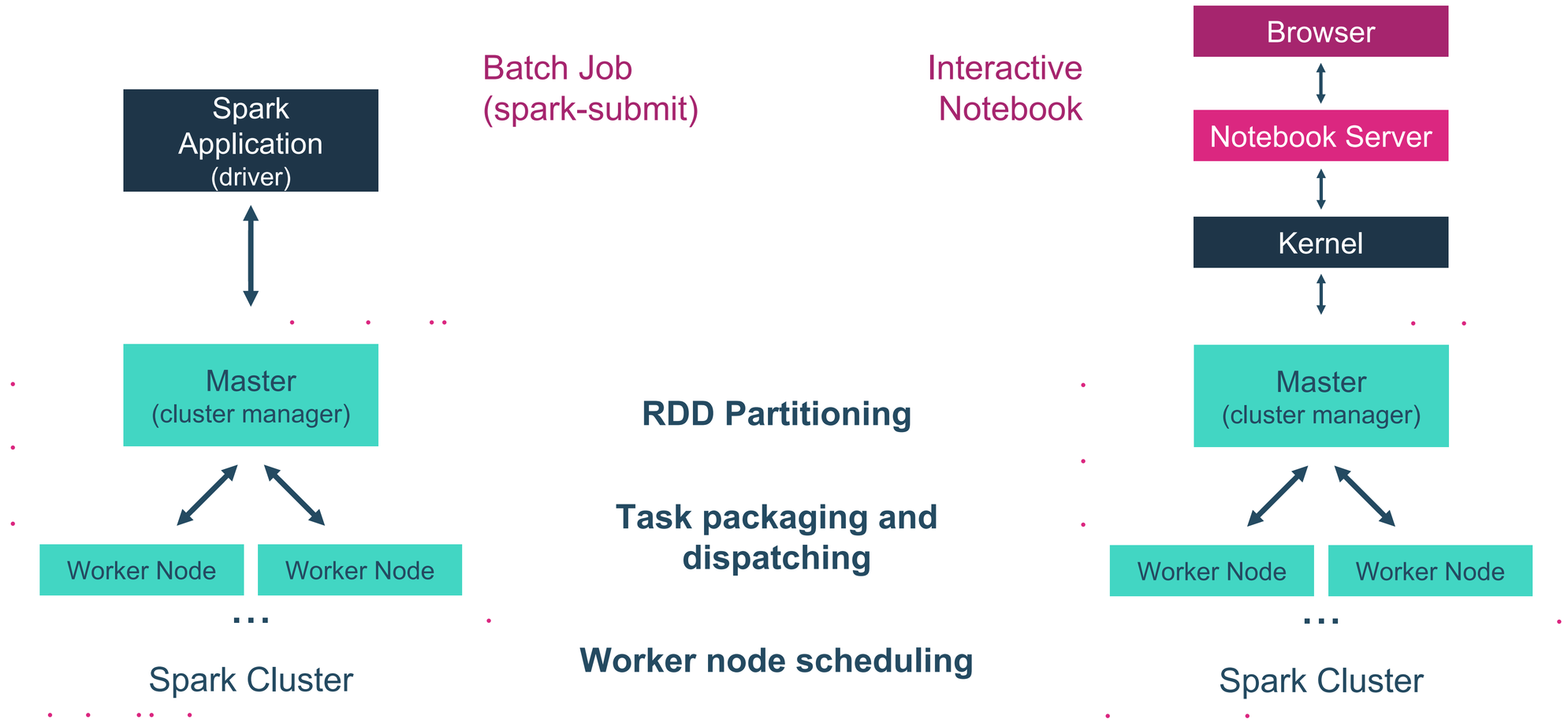
“What exactly is Apache Spark?”

— NATASHA 

Great Question Natasha!

Best way to answer it is to arrange a ticket to the Spark Summit for you to find out

CONSUMING SPARK



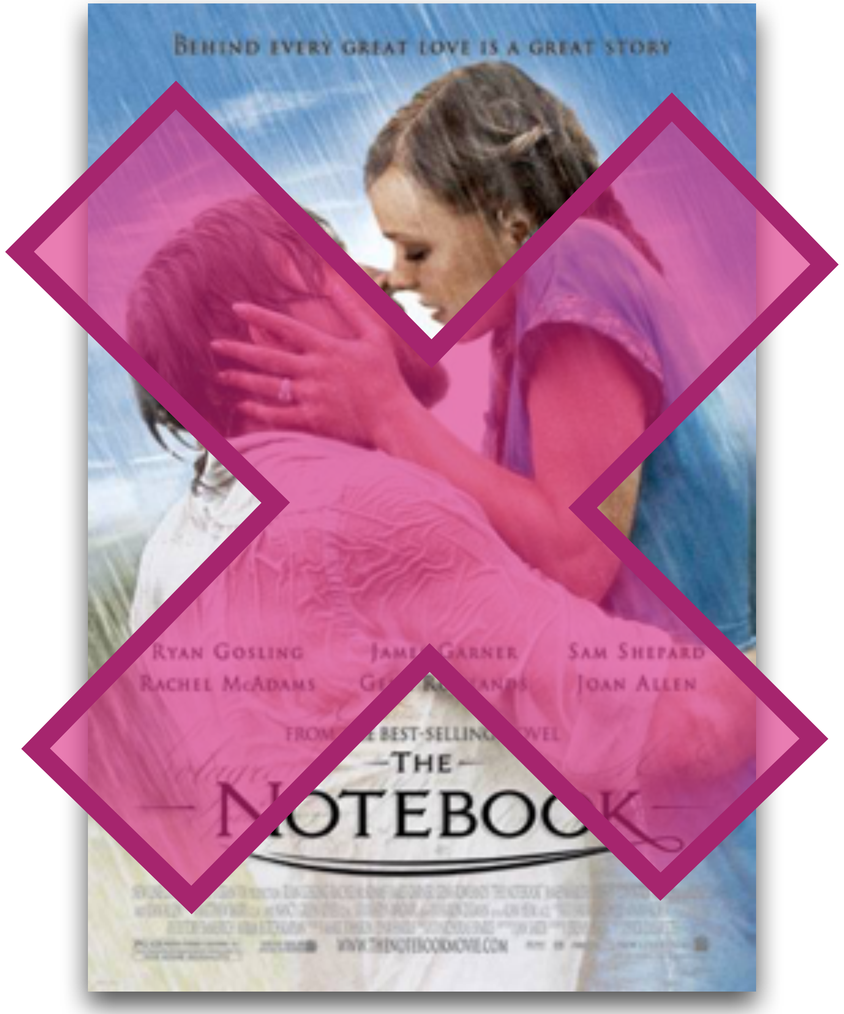
CAN WE COLLABORATE USING NOTEBOOKS?

“What exactly is a Notebook?”

— NATASHA 

— BEN 

RYAN GOSLING MOVIE?



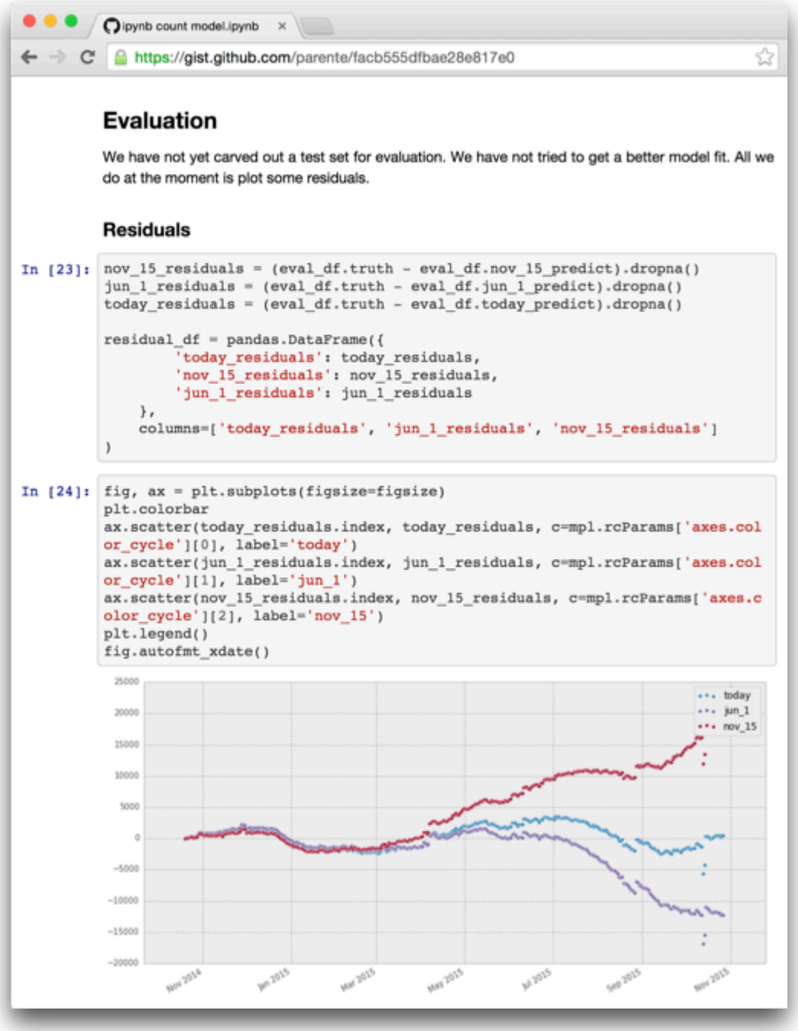
http://ia.media-imdb.com/images/M/MV5BMTk3OTM5Njg5M15BMl5BanBnXkFtZTYwMzA0ODI3._V1_SX640_SY720_.jpg

WHAT IS A NOTEBOOK?

Text Annotations

Code Data

Visualizations Widgets Output



- Web based UI for running Apache Spark console commands
- Easy, no install spark accelerator
- Best way to start working with spark
- Multiple flavors
  - Jupyter
  - Zeppelin
- Local or cloud hosted
  - IBM Data Science Experience
  - Databricks

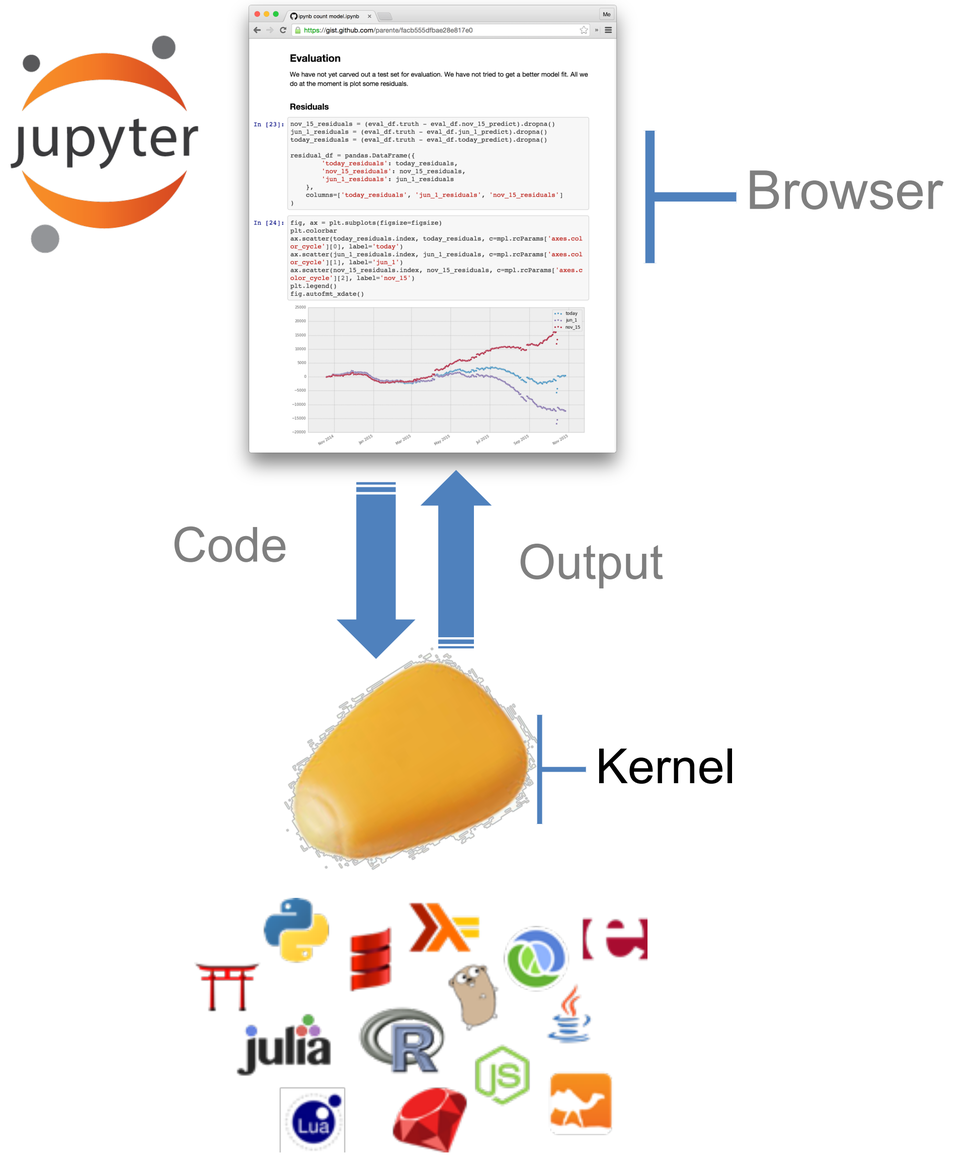
What is Jupyter?

"Open source, interactive data science and scientific computing"

- –Formerly IPython
- –Large, open, growing community and ecosystem

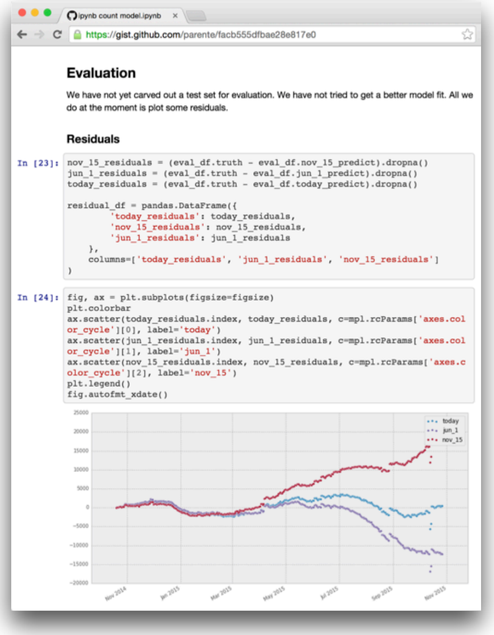
Very popular

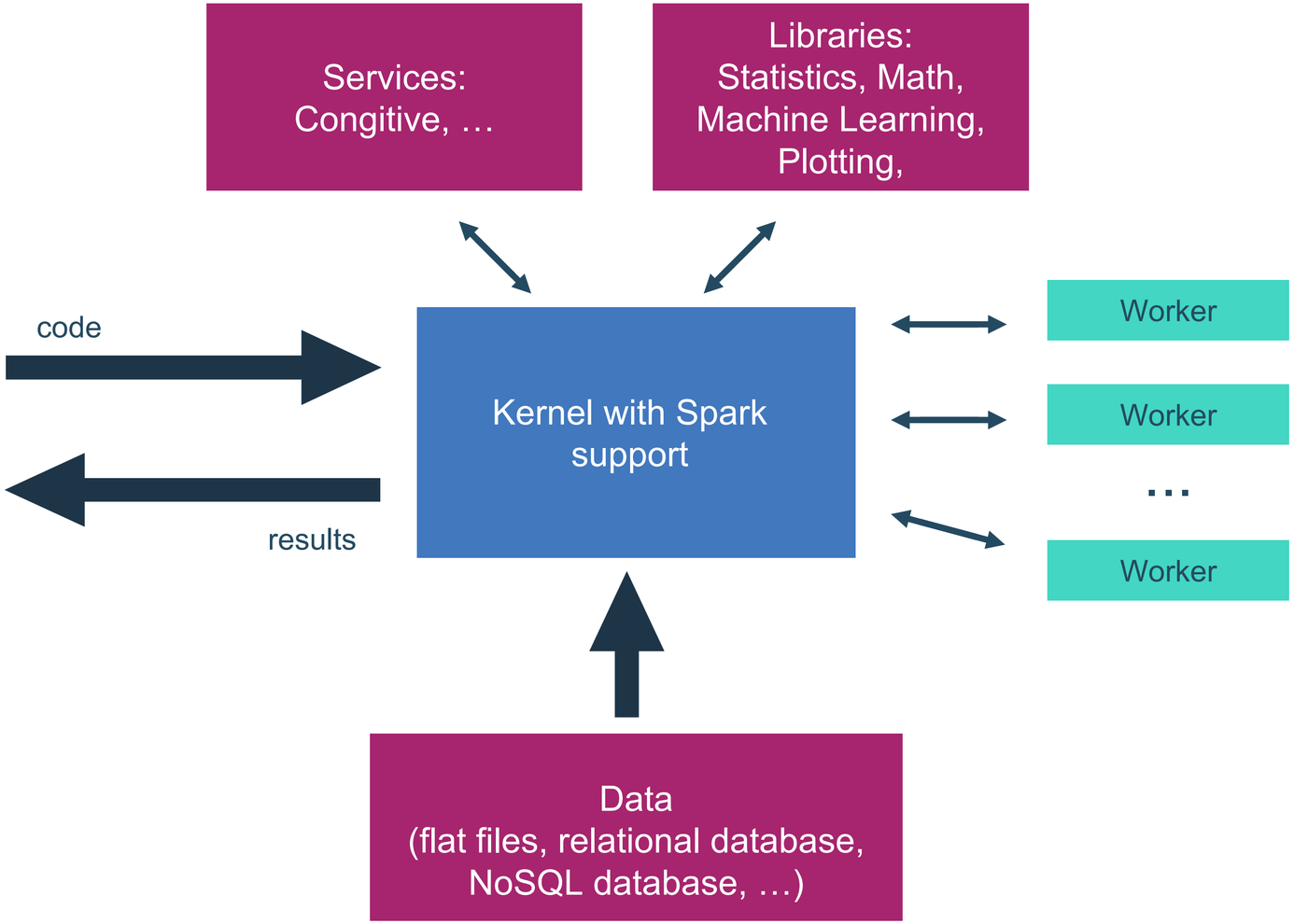
- –"~2 million users for IPython" [1]
- –$6m in funding in 2015 [3]
- –200 contributors to notebook subproject alone [4]
- –275,000 public notebooks on GitHub [2]



https://www.bluetrack.com/uploads/items_images/kernel-of-corn-stress-balls1_thumb.jpg?r=1

BIG DATA ANALYSIS





NOTEBOOKS ARE POWERFUL TOOLS FOR DATA SCIENTISTS

"But they seem complicated for developers like me"

— BEN 

ENTER PIXIEDUST

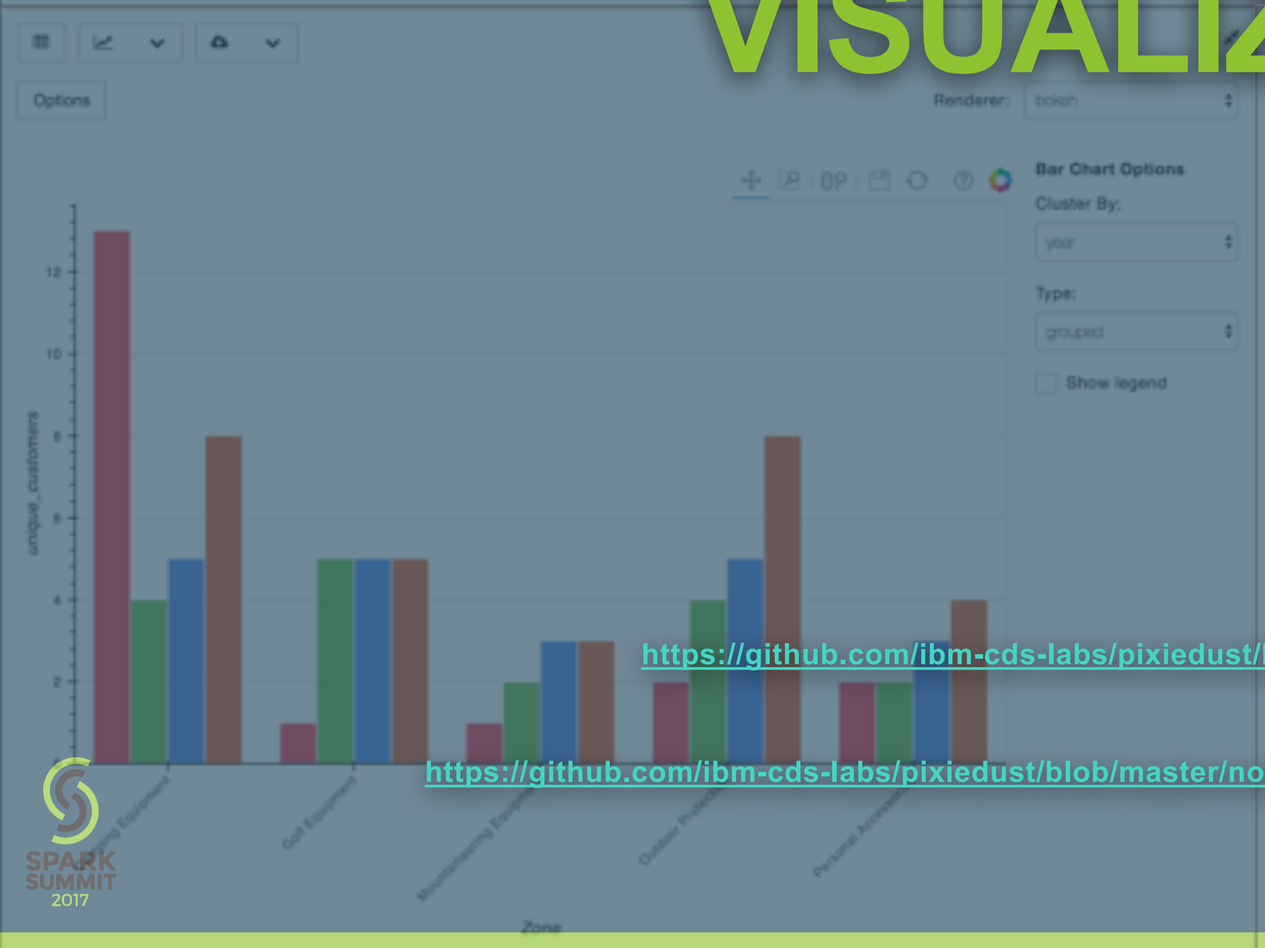
Open Source Python helper library for Jupyter Notebooks

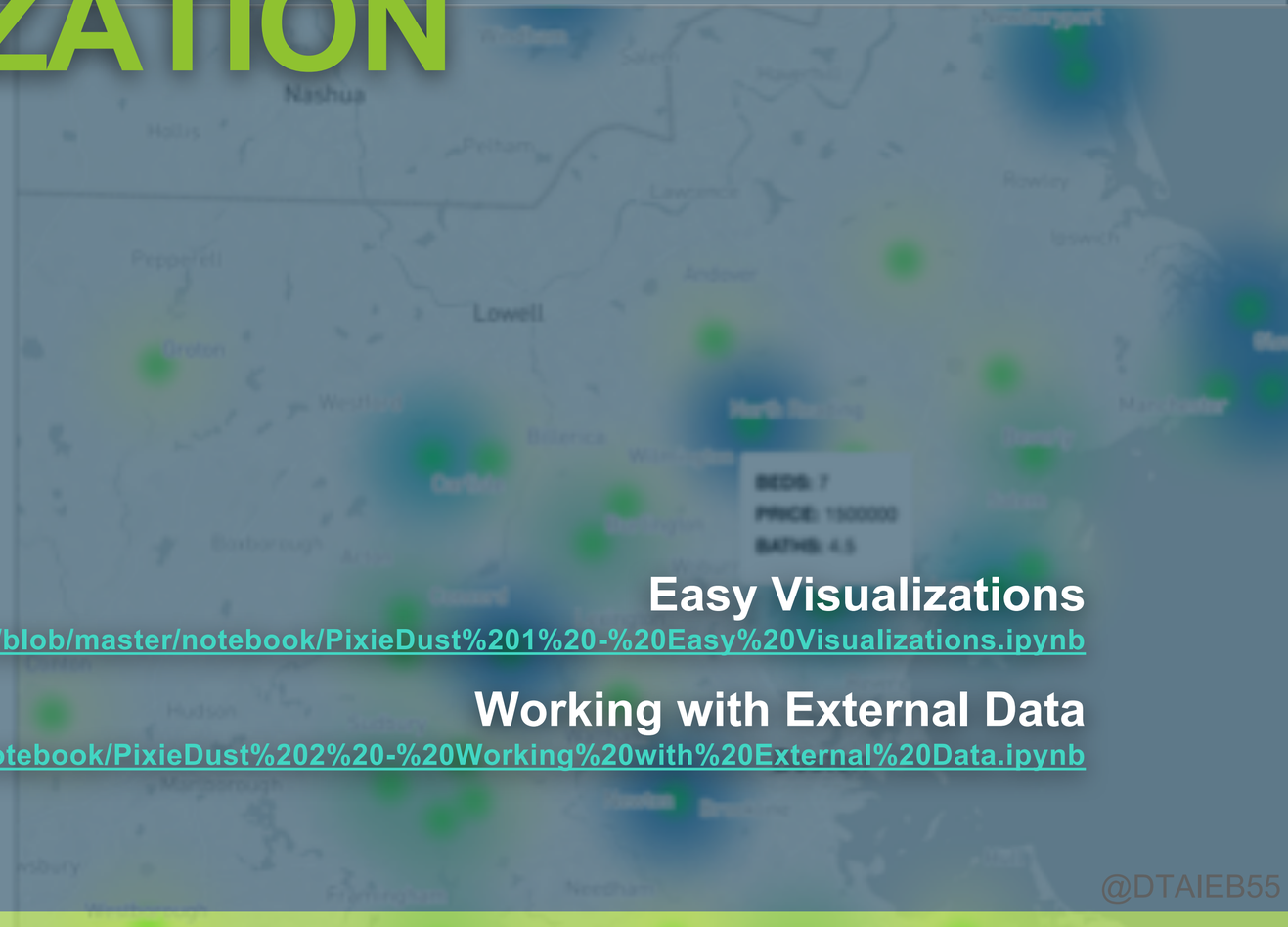
- Visualize data (e.g., Table, Charts, Map, etc)
- Data Management with PixieApps
- Download/export data (e.g., File, Cloudant, etc.)
- Use Scala directly in a Python notebook
- Install Spark packages into Python notebook
- Spark job progress monitor
- Extensible

https://github.com/ibm-cds-labs/pixiedust

```
import pixiedust
d1 = sqlContext.createDataFrame(
[(2010, 'Camping Equipment', 3, 200),(2010, 'Camping Equipment', 10, 200),(2010, 'Golf Equipment', 1, 240),
 (2010, 'Mountaineering Equipment', 1, 348),(2010, 'Outdoor Protection',2,200),(2010, 'Personal Accessories', 2, 200),
 (2011, 'Camping Equipment', 4, 489),(2011, 'Golf Equipment', 5, 234),(2011, 'Mountaineering Equipment',2, 123),
 (2011, 'Outdoor Protection', 4, 654),(2011, 'Personal Accessories', 2, 234),(2012, 'Camping Equipment', 5, 876),
 (2012, 'Golf Equipment', 5, 200),(2012, 'Mountaineering Equipment', 3, 156),(2012, 'Outdoor Protection', 5, 200),
 (2012, 'Personal Accessories', 3, 65),(2013, 'Camping Equipment', 8, 177),(2013, 'Golf Equipment', 5, 434),
 (2013, 'Mountaineering Equipment', 2, 274),(2013, 'Outdoor Protection', 4, 218),(2013, 'Personal Accessories',4, 200)],
["year","zone","unique_customers","revenue"])
display(d1)
```

DEMO: PIXIEDUST DATA VISUALIZATION





Easy Visualizations

https://github.com/ibm-cds-labs/pixiedust/blob/master/notebook/PixieDust%201%20-%20Easy%20Visualizations.ipynb

Working with External Data

https://github.com/ibm-cds-labs/pixiedust/blob/master/notebook/PixieDust%202%20-%20Working%20with%20External%20Data.ipynb

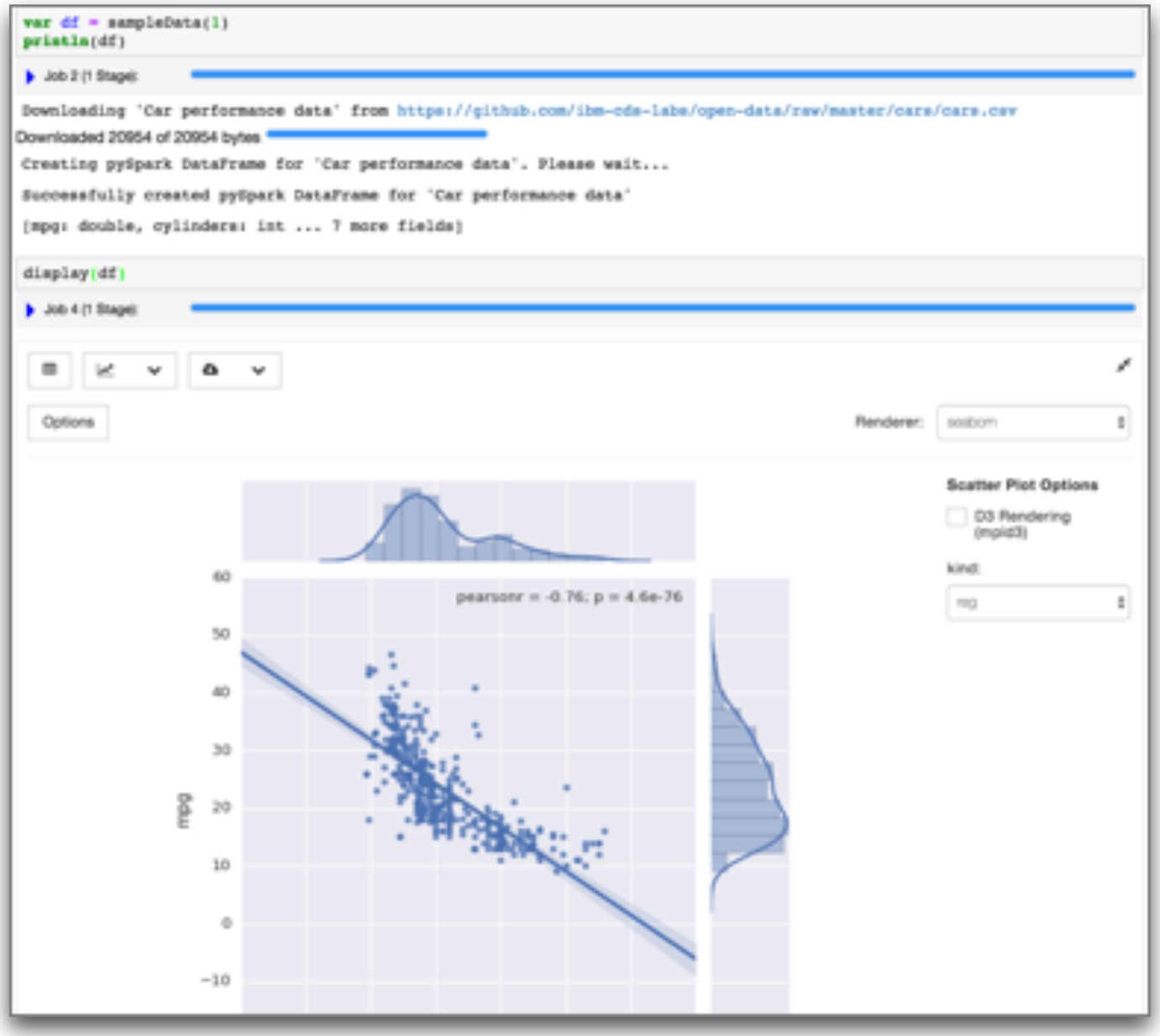
I AM OK TO USE PYTHON

"But I am really more comfortable with Scala"

— BEN 

SCALA NOTEBOOKS

PixieDust also works with Scala Notebooks



Same PixieDust Scala APIs as in Python

WHAT ABOUT THE LINE OF BUSINESS USER?

"Expressing everything in code is nice but LOB users will not be able to linearly run large number of cells"

— NATASHA

Enter PixieApps

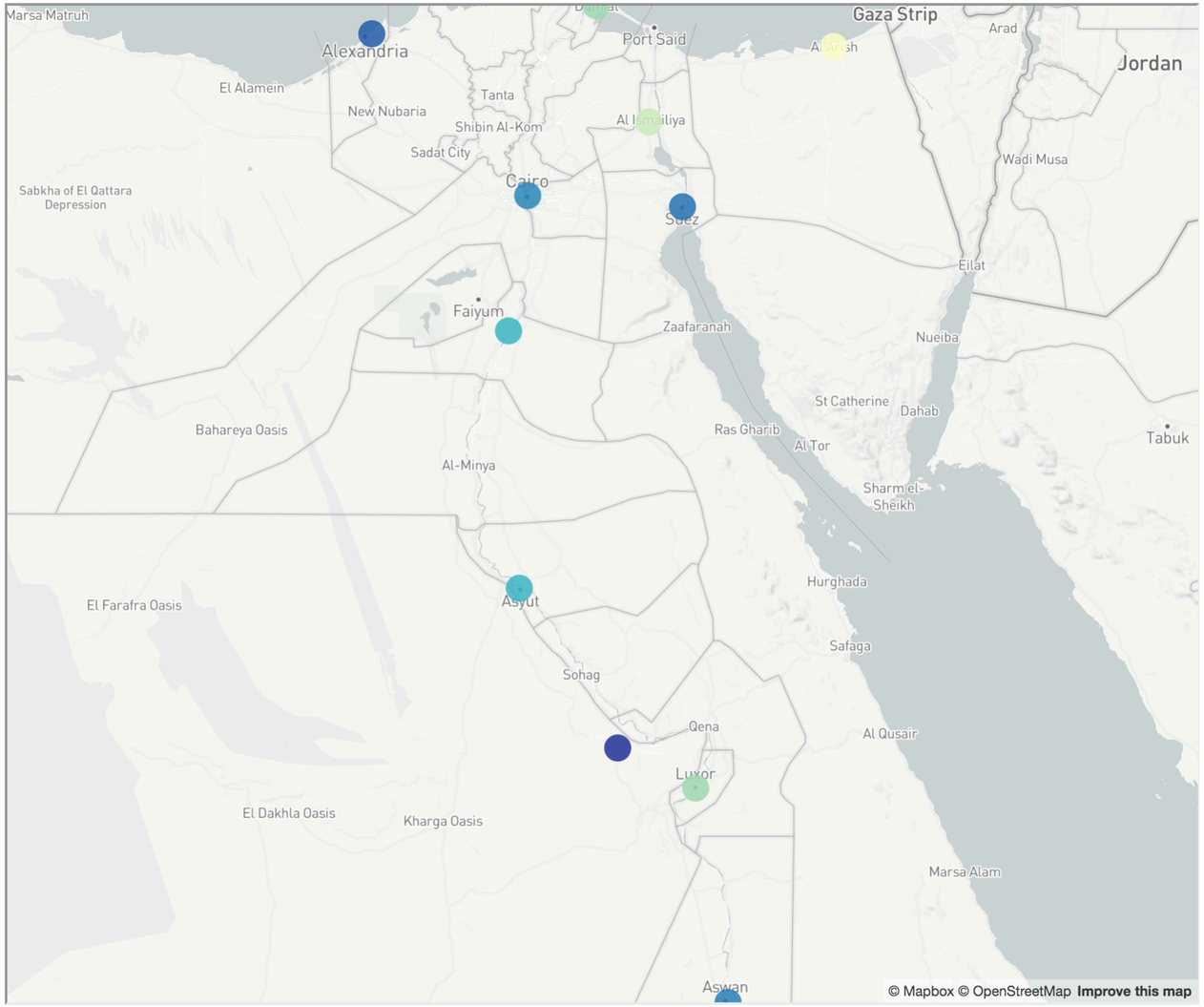
- PixieApps are Python classes that let you write UI for your analytics
- Easy to build: mostly HTML and CSS with some custom attributes (micro-format style)
- Leverage PixieDust Display visualization for charting
- With PixieApps you can:
  - Create different html views with routes to invoke them
  - Invoke Python Scripts from user interactions
  - Run in the notebook cell output or in a Dialog
  - and much more…
- Use cases:
  - Dashboards
  - Data Browsers
  - Data Pipeline Management

Demo: PeaceTech GroundTruth Global Dashboard

Pixiedust: groundTruth global

Disruption Dashboard | GroundTruth Explorer

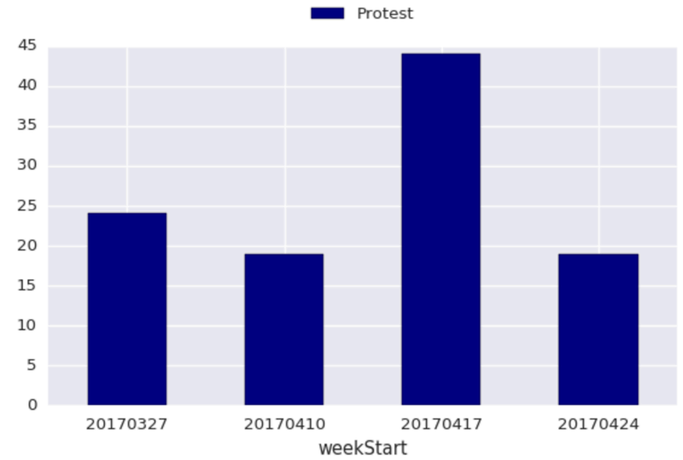
Egypt | 03/27/2017 | 04/30/2017 | Go



Search table

Showing 5 of 5

dispText	url
Outrage has broken out following the President of Egypt ratifying amendments that give him increasing power over the courts of the country. The Judges Club has stated that there will be an increase of protests this week following this ruling, and that some strikes will be held to protest the what is known as "authoritarian" ruling.	https://www.alaraby.co.uk/english/news/2017/04/30/egypts-judges-up-in-arms-after-sisi-passes-authoritarian-judicial-law
Protests have been on the rise following the Egyptian courts decision to overturn the ruling that blocked the transfer of two Red Sea islands to Saudi Arabia	http://www.defenceweb.co.za/index.php?option=com_content&view=article&id=47368:egyptian-court-overturns-block-on-red-sea-islands-transfer&catid=56:diplomacy-a-peace&Itemid=111
Protests are still increasing as the outrage from the attack by ISIS on the Alexandria Saint Mark's Cathedral moves into the next week. Palestinians and Egyptians together are protesting this deadly attack, and condeming it. Protestors want an end to the attacks on Christians, and want	http://normangeestar.net/2017/04/23/palestinians-and-egyptians-protest-against-egypt-attacks/



SPARK SUMMIT 2017

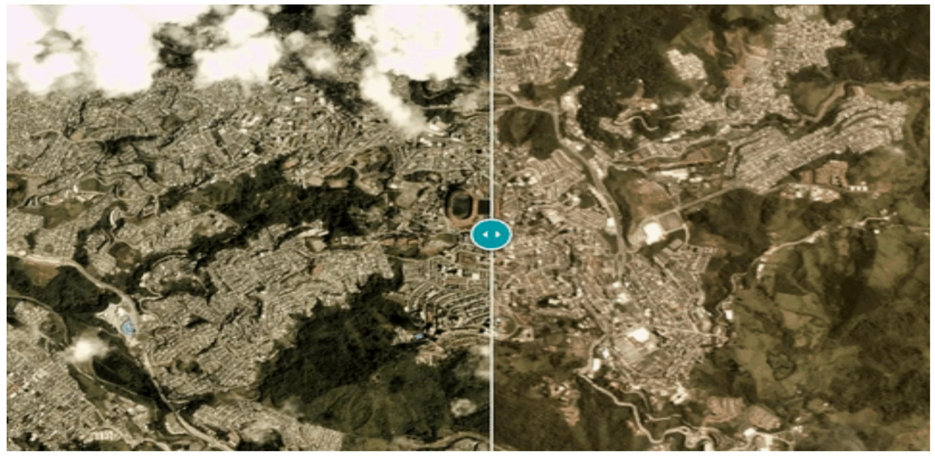
Demo: PeaceTech GroundTruth Global Dashboard

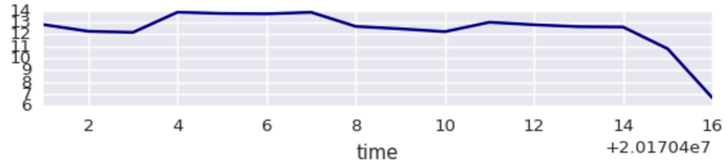
Pixiedust: groundTruth global

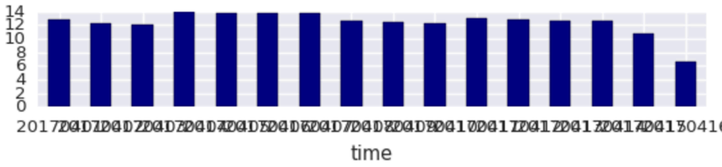
Disruption Dashboard | GroundTruth Explorer

Egypt ▾ 03/27/2017 04/30/2017 Go

- Civil Unrest
- Environment
- Infrastructure
- Food
- Crime
- Social Issues







Search table

Showing 5 of 5

dispText	url
Outrage has broken out following the President of Egypt ratifying amendments that give him increasing power over the courts of the country. The Judges Club has stated that there will be an increase of protests this week following this ruling, and that some strikes will be held to protest the what is known as "authoritarian" ruling.	https://www.alaraby.co.uk/english/news/2017/04/30/egypts-judges-up-in-arms-after-sisi-passes-authoritarian-judicial-law
Protests have been on the rise following the Egyptian courts decision to overturn the ruling that blocked the transfer of two Red Sea islands to Saudi Arabia	http://www.defenceweb.co.za/index.php?option=com_content&view=article&id=47368:egyptian-court-overturns-block-on-red-sea-islands-transfer&catid=56:diplomacy-a-peace&Itemid=111
Protests are still increasing as the outrage from the attack by ISIS on the Alexandria	

Search table

Showing 5 of 5

hashtag	url
#Venezuela	https://twitter.com/TITORODRIGUEZZ/status/851277765118537729?ref_src=twsrc%5Etfw&ref_url=https%3A%2F%2F
#Caracas	https://twitter.com/VOANoticias/status/856666681137123328?ref_src=twsrc%5Etfw&ref_url=https%3A%2F%2F
#Ahora	https://twitter.com/LaNocheNTN24/status/857822686097022980?ref_src=twsrc%5Etfw&ref_url=https%3A%2F%2F
#VIDEO	https://twitter.com/AndrewsAbreu/status/850875727243350016?ref_src=twsrc%5Etfw&ref_url=https%3A%2F%2F
#10Abr	https://twitter.com/EfectoCocuyo/status/851479662882684932?ref_src=twsrc%5Etfw&ref_url=https%3A%2F%2F

Thousands of people have gathered in Bogota and Medellin to protest against corruption and the Santos administration more broadly. In Bogota, at least 20,000 people joined protests in Bolivar Square. The protests have been organized by former president and current Senator Alvaro Uribe and have received widespread support among conservative groups. More protests are planned for the next few weeks. In addition, farmers in rural areas like Narino have begun protesting against the government's efforts to eradicate coca, claiming that the government is not holding up its end of the bargain for these programs. On April 24th, women also took to the streets of Bogota to bring awareness to the increasing assault and murder against women that goes unnoticed in the country.

ISIS attacks at the Coptic Christian

Demo: Data Browser for Cloudant/CouchDB

```
In [3]: from pixiedust.apps.cloudantBrowser import *

c = CloudantBrowser()
c.run()
```

Select a cloudant connection: local Go

Back

flight-metadata

All Documents
Query
Design Documents
Views
_design/flightMetadata
US Airports
airports
airlines
airlines by Name

View (flightMetadata/US Airports)

```
f4f2c5ee32a9328500ffc78e5a82272d {
  "city": "Bay Springs"
  "countryCode": "US"
  "countryName": "United States"
  ...
}
f4f2c5ee32a9328500ffc78e5a822acd {
  "city": "Bridgeton"
  "countryCode": "US"
  "countryName": "United States"
  ...
}
f4f2c5ee32a9328500ffc78e5a823247 {
  "city": "Livingston"
  "countryCode": "US"
  "countryName": "United States"
  ...
}
f4f2c5ee32a9328500ffc78e5a82371e {
  "city": "Mc Kenzie Bridge"
  "countryCode": "US"
  "countryName": "United States"
  ...
}
f4f2c5ee32a9328500ffc78e5a82400c {
  "city": "Colorado Springs"
  "countryCode": "US"
  "countryName": "United States"
  ...
}
```

Next 1 to 5 of 5584

Generate DataFrame

Back

Databases

_replicator
_users
aaaa
auth_users
baseline-20170418$180058
baseline-20170418-175615
baseline-20170421$113733
couchapp
dataframe-20170414-144716
dataframe-20170414-155453
dataframe-20170414-163626
dataframe-20170414-163940
dataframe-20170414-164242
dataframe-20170417-115929
dataframe-20170417-120602
david2
demo_demotable
egypt_training
enotes
flight-metadata

SPARK SUMMIT 2017

WHAT DOES IT TAKE TO BUILD A PIXIEAPP?

"Do I need to learn yet another framework?"

— BEN 

PIXIEAPP HELLO WORLD

```
from pixiedust.display.app import *

@PixieApp
class HelloWorldPixieApp:

    @route()
    def main(self):
        return"""
            <input pd_options="clicked=true" type="button" value="Click Me">
        """

    @route(clicked="true")
    def _clicked(self):
        return """
            <input pd_options="clicked=false" type="button" value="You Clicked, Now Go back">
        """

#run the app
HelloWorldPixieApp().run(runInDialog='false')
```

Import app package to start things off

Simple annotation to tell PixieDust it's an app

Define default route (no args)

Method will return the view's html fragment

set option clicked to true when button is pushed, so that (route) is loaded next

Define a new route that triggers when option clicked is set to true

Allows Jinja2 template macros

Import app package to start things off

PIXIEAPP HELLO WORLD WITH DATA

```
from pixiedust.display.app import *

@PixieApp
class HelloWorldPixieAppWithData:

    @route()
    def main(self):
        return"""
        <div class="row">
            <div class="col-sm-2">
                <input pd_options="handlerId=dataframe"
                       pd_entity
                       pd_target="target{{prefix}}"
                       type="button" value="Preview Data">
            </div>
            <div class="col-sm-10" id="target{{prefix}}"/>
        </div>
        """

#Create dataframe
df = SQLContext(sc).createDataFrame(
[(2010, 'Camping Equipment', 3, 200),(2010, 'Camping Equipment', 10, 200),(2010, 'Golf Equipment', 1, 240),
 (2010, 'Mountaineering Equipment', 1, 348),(2010, 'Outdoor Protection',2,200),(2010, 'Personal Accessories', 2, 200),
 (2011, 'Camping Equipment', 4, 489),(2011, 'Golf Equipment', 5, 234),(2011, 'Mountaineering Equipment',2, 123),
 (2011, 'Outdoor Protection', 4, 654),(2011, 'Personal Accessories', 2, 234),(2012, 'Camping Equipment', 5, 876),
 (2012, 'Golf Equipment', 5, 200),(2012, 'Mountaineering Equipment', 3, 156),(2012, 'Outdoor Protection', 5, 200),
 (2012, 'Personal Accessories', 3, 345),(2013, 'Camping Equipment', 8, 987),(2013, 'Golf Equipment', 5, 434),
 (2013, 'Mountaineering Equipment', 3, 278),(2013, 'Outdoor Protection', 8, 134),(2013,'Personal Accessories',4, 200)],
["year","zone","unique_customers", "revenue"])

#run the app
HelloWorldPixieAppWithData().run(df, runInDialog='false')
```

Specify Display options for visualization

Entity binding: use the df passed by user

Allows binding of any entity created by the app

Display the output in the specified target

Placeholder div for displaying data

Pass data to the app

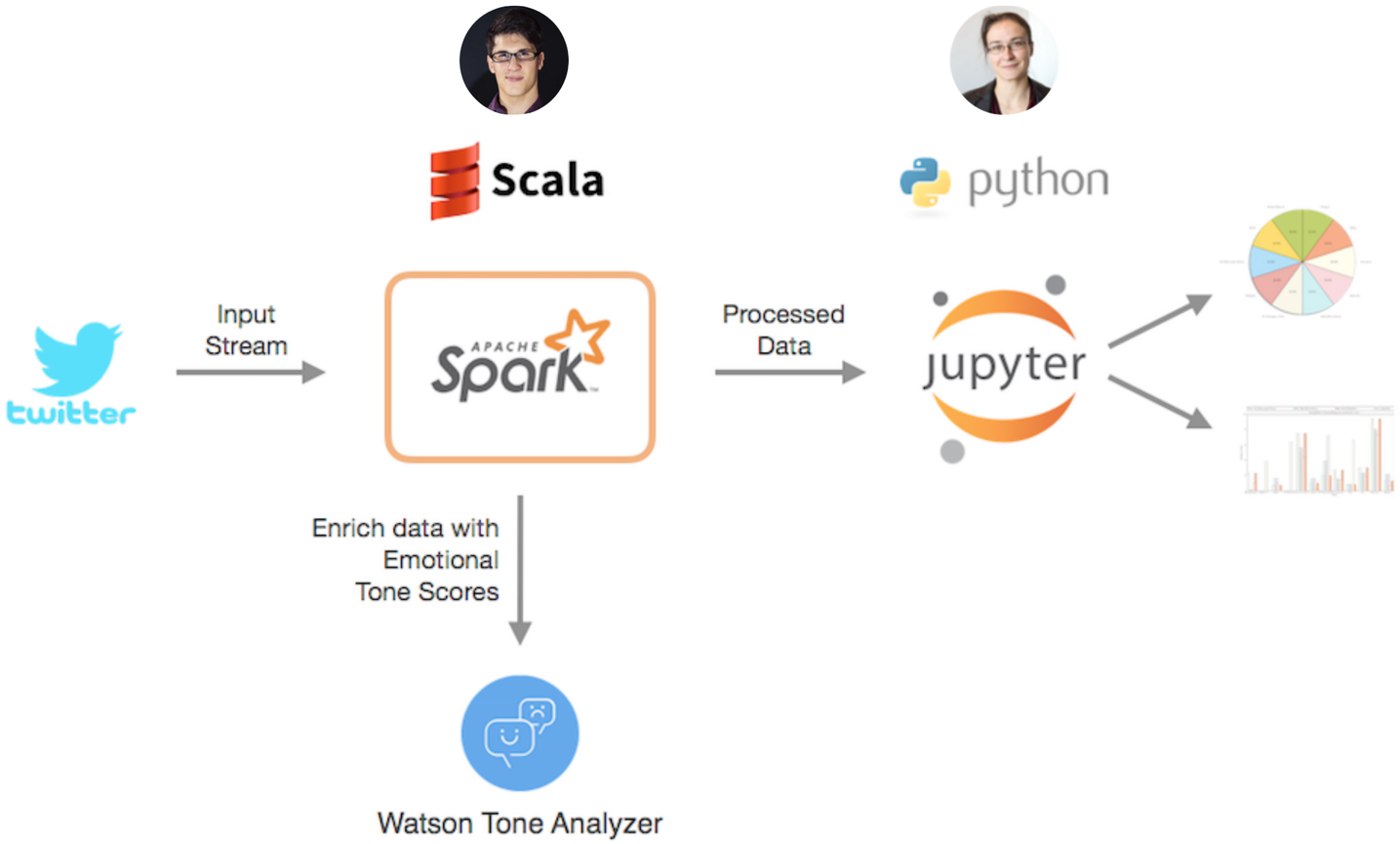
Want to learn more about PixieApp?

See PixieApp documentation here:

https://ibm-cds-labs.github.io/pixiedust/pixieapps.html

OK, I'M SOLD…

LET'S AGREE ON THE ARCHITECTURE



BEN and NATASHA

START BRAINSTORMING

- I'll work on data acquisition from Twitter and enrichment with sentiment analysis scores using Spark Streaming
- I'll implement the code in Scala and I'll create a jar library for Natasha.

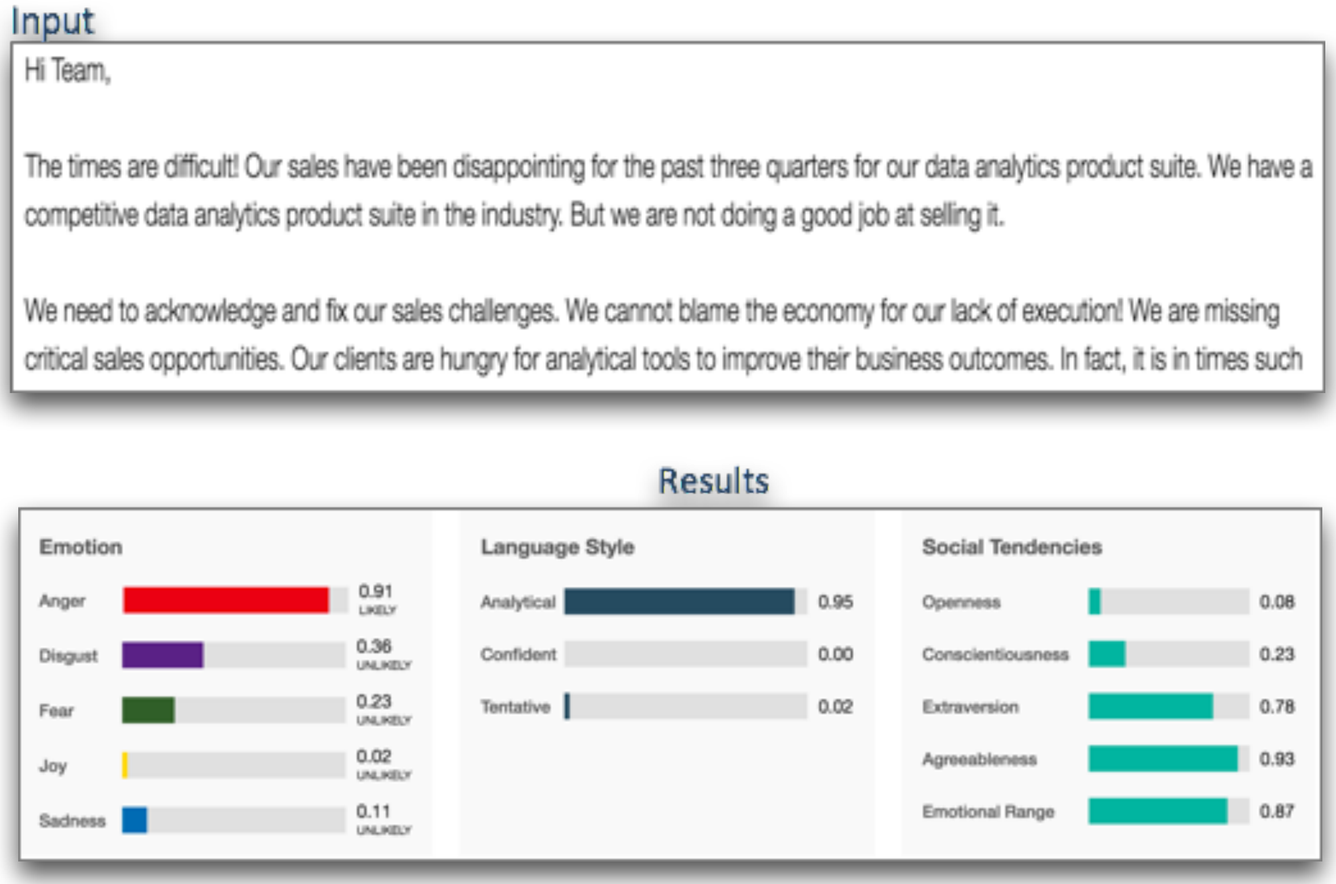
- I'll use the jar from Ben to create Spark dataframes with the tweets enriched with the Tone analyzer scores
- I'll perform the data exploration and analysis using the PixieDust display() API

We'll work together to write a PixieApp Dashboard with Real-Time visualization for the Line of Business User

WATSON TONE ANALYZER

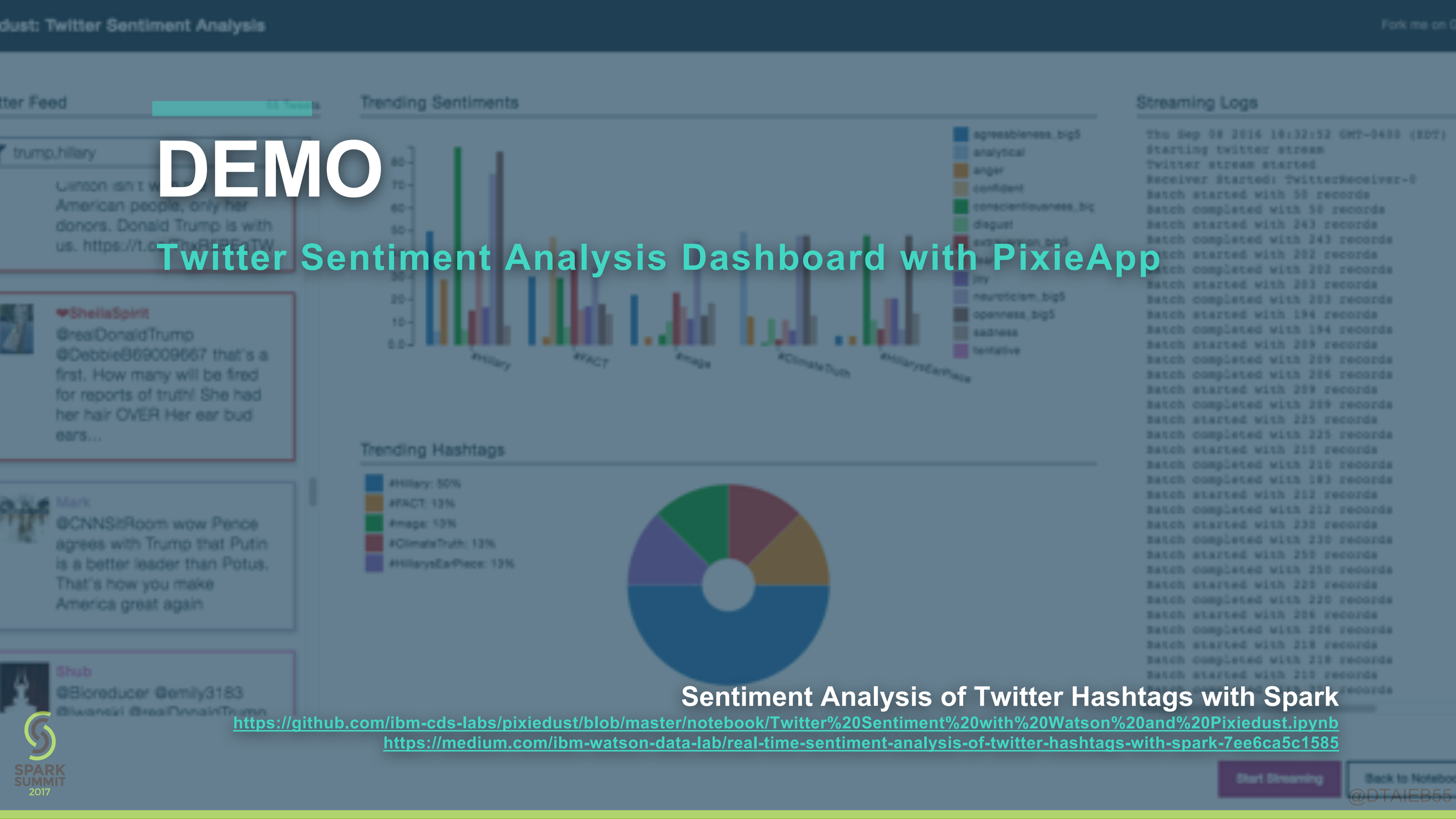
http://www.ibm.com/watson/developercloud/tone-analyzer.html

- Uses linguistic analysis to detect 3 types of tones
  - Emotion
  - Social Tendencies
  - Language Styles
- Available as a cloud service on IBM Bluemix



DEMO

Twitter Sentiment Analysis Dashboard with PixieApp



Sentiment Analysis of Twitter Hashtags with Spark

https://github.com/ibm-cds-labs/pixiedust/blob/master/notebook/Twitter%20Sentiment%20with%20Watson%20and%20Pixiedust.ipynb

https://medium.com/ibm-watson-data-lab/real-time-sentiment-analysis-of-twitter-hashtags-with-spark-7ee6ca5c1585

SPARK SUMMIT 2017

MEETING WITH THE VP

https://unsplash.com/search/meeting?photo=3fPXt37X6UQ

What's next for PixieDust

- Support Visualization for Streaming data
  - Start with Structured Streaming, MessageHub/Kafka and IBM Streams
- Ability to publish/embed PixieApps into Web Application (Nodejs to begin with)
- PixieDust visualization enhancements
  - Custom colors
  - Custom GeoJSON layers for maps
  - Sorting/filtering
  - More renderers: Brunel, ArcGIS, etc.
  - …
- Ability to run Node.js code to load and visualize data
- Support for Jupyter Labs and Jupyter Hub

As always…

We look forward for your feedback
and pull requests on GitHub

https://github.com/ibm-cds-labs/pixiedust

CONCLUSION

- Solving the Data problems of tomorrow cannot be done by data scientists alone.
- Notebooks, considered by most to be the domain of data scientists, can help break down traditional silos and help team of all types who are working on data problems

[1] Not just for data scientists

Try it for yourself today:

- IBM Data Science Experience
  http://datascience.ibm.com/
- Locally using PixieDust automated installer
  https://ibm-cds-labs.github.io/pixiedust/install.html

RESOURCES

- https://github.com/ibm-cds-labs/pixiedust
- https://ibm-cds-labs.github.io/pixiedust
- https://medium.com/ibm-watson-data-lab/i-am-not-a-data-scientist-efe7ca6ceba2
- https://spark.apache.org
- https://www.ibm.com/us-en/marketplace/spark-as-a-service
- http://datascience.ibm.com
- https://www.ibm.com/watson/developercloud/tone-analyzer.html
- https://medium.com/ibm-watson-data-lab/real-time-sentiment-analysis-of-twitter-hashtags-with-spark-7ee6ca5c1585
- https://ibm.biz/pixiedustvis
- https://ibm.biz/pixiedustlab

If you want to learn more

Register now!
seti.org/ML4SETI

IBM galvanize skymind

Hackathon
Galvanize, San Francisco

June 10-11, 2017

Online Code Challenge

June/July, 2017

The New Builders Podcast Episode 35: **Data Science for All—From Mundane to Magic with PixieDust**

Featured Guest

David Taieb
Distinguished Engineer,
IBM Watson Data Platform

Free Workshop and Demo

Data Science Boot Camp at Spark Summit

Presented by the
IBM Watson Data Platform Team

Wednesday, June 7, 2017, 3:00-5:00pm

Galvanize San Francisco – SoMa
Speakeasy Room
44 Tehama St., San Francisco
(8 minute walk from the Moscone Center)

Want to be more efficient with Apache Spark™ and Jupyter Notebooks? Try a sprinkling of PixieDust, a new open source library for data manipulation and visualization.

In this two-hour workshop, IBM Distinguished Engineer David Taieb will walk through how to use PixieDust to build charts and maps to discover insights.

Whatever your skill level with Spark and Notebooks, you'll be able to participate. Bring your laptop!

Refreshments will be served.

SPARK SUMMIT 2017

Questions


thank you
merci
gracias
danke
謝謝
obrigado
go raibh maith agat
dziękuję
bedankt
sukriya
kop khun krap
terima kasih
teşekkür ederim
ngiyabonga
спасибо
рахмат
Баярлалаа
dank je
tapadh leat
dakujem
arigatô
grazie
takk
taiku
sagolun
hvala
mauruuru
dhanyavad
kiitos
dankie
faafetai lava
vinaka
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
спасибі
barka
welalin
tack
spas
mersi
kia ora
köszönöm
nanni
nandri
bayarlalaa
gràçie
enkosi
akun
dankon
ačiū
djiere dieuf
tau
mochchakkeram
obrigada
asante
manana
tenki
хвала
murakoze
chokrane
mamnun
дякую
chnorakaloutioun
gratias ago
gràcies
sulpáy
sobodi
dëkuji
mèsi
didi madloba
kam sah hamnida
najis tuke
rahmat
ありがとう
tanemirt
rahmet
dhanyavadagalu
diolch
shukriya
merci
мерси
trugarez
ευχαριστώ
xièxie
감사합니다
তোমাকে ধন্যবাদ