Building a Scalable Machine Learning Pipeline

Dan Dixey & Chris Harris

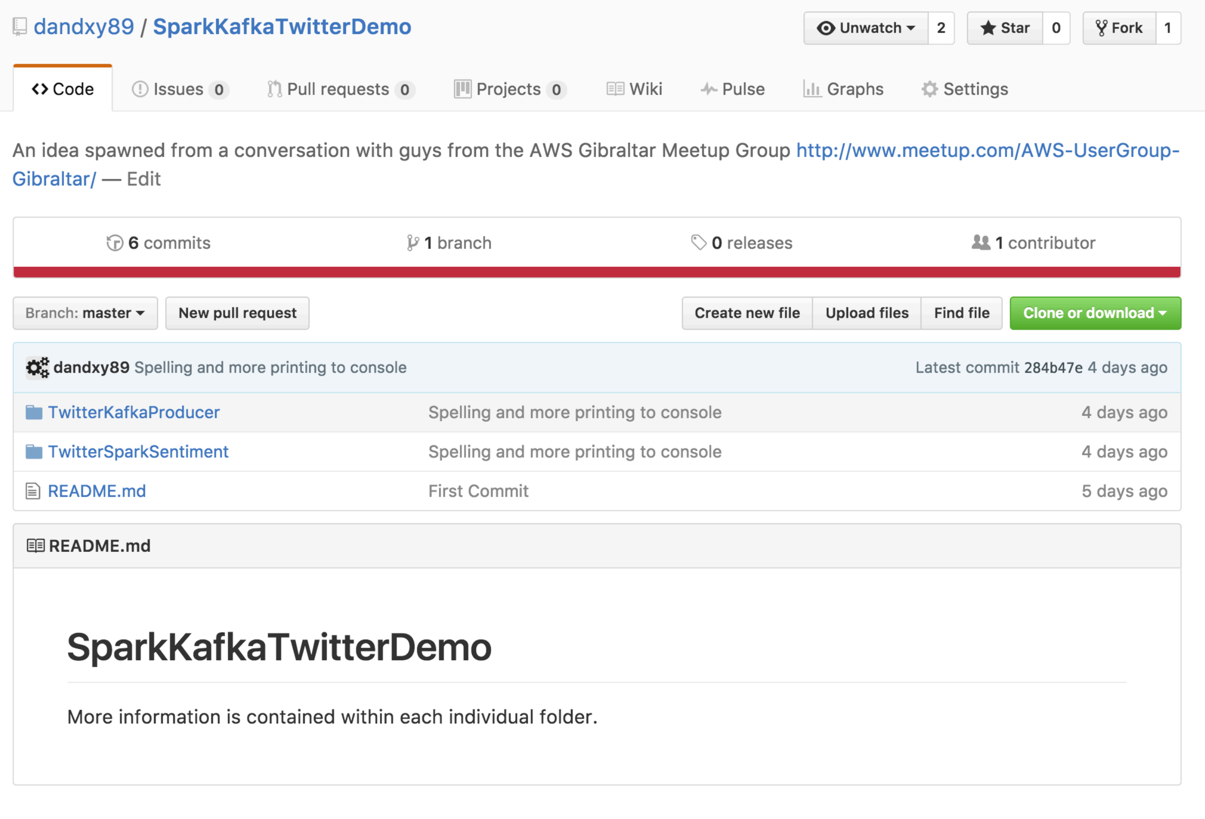
amazon web services AWS User Group Gibraltar

Why?

- Based on the discussion at the last meet up group these these themes were discussed:
  - Machine Learning
  - Natural Language Processing
  - “Big Data”
  - And others of course…

What have you done?

- Setup a mini-project that can be used to:
  - learn and demonstrate the power of the AWS ecosystem
  - gain some exposure to Machine Learning
  - play with cool technologies
  - And most importantly, promote the AWS Gibraltar Meetup group



An idea spawned from a conversation with guys from the AWS Gibraltar Meetup Group http://www.meetup.com/AWS-UserGroup-Gibraltar/ — Edit

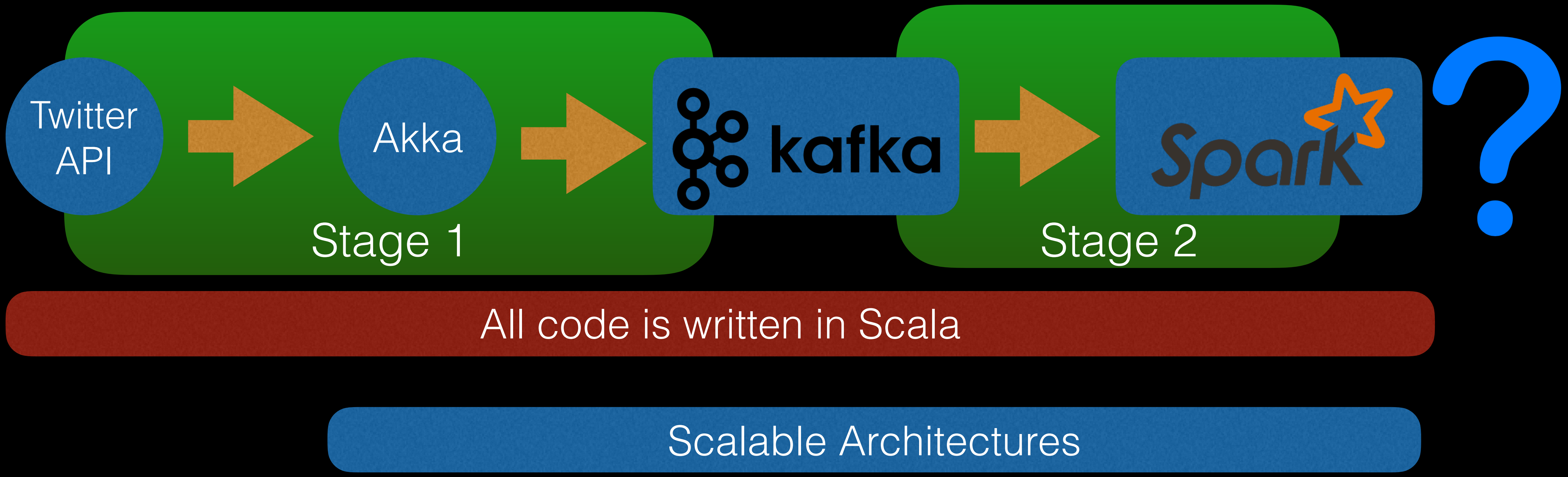
dandxy89 Spelling and more printing to console — Latest commit 284b47e 4 days ago

TwitterKafkaProducer	Spelling and more printing to console	4 days ago
TwitterSparkSentiment	Spelling and more printing to console	4 days ago
README.md	First Commit	5 days ago

SparkKafkaTwitterDemo

More information is contained within each individual folder.

The master plan.



Twitter API

Akka

kafka

Spark

?

Stage 1

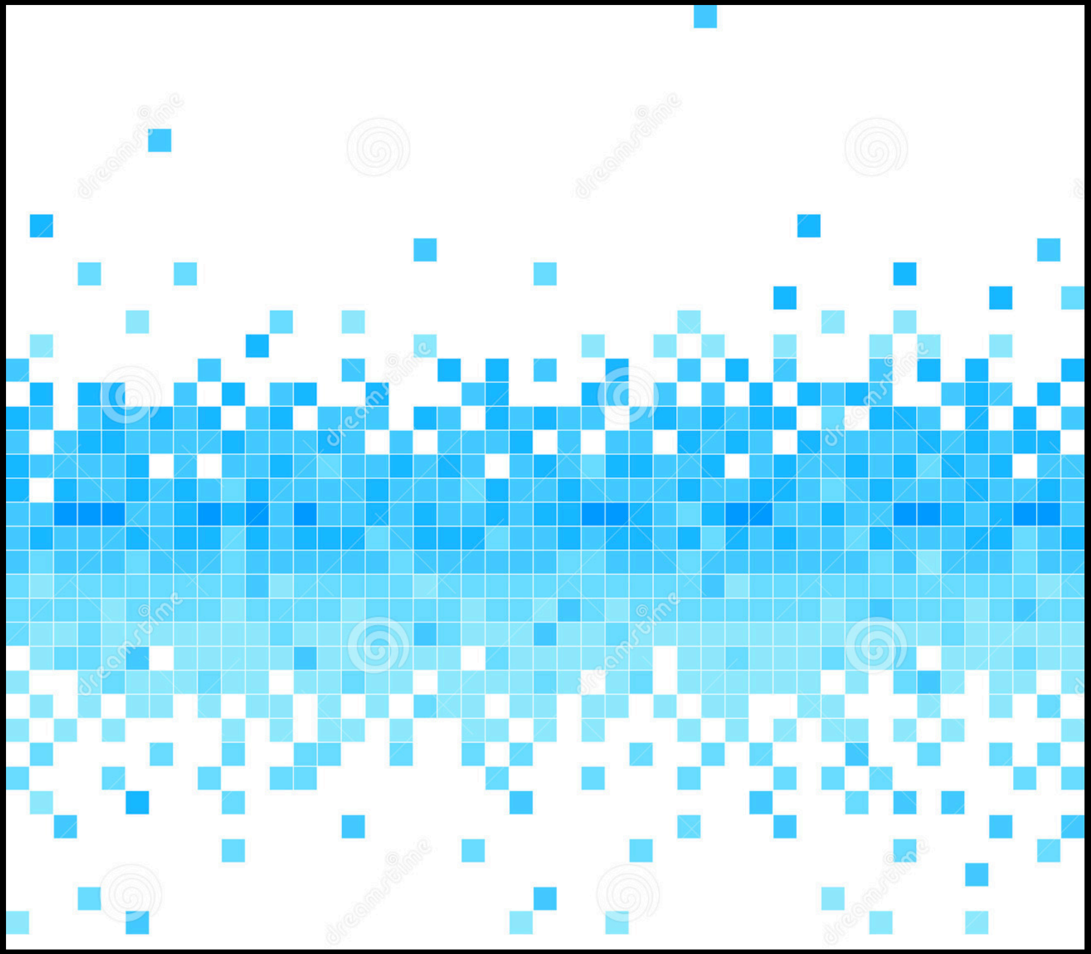
Stage 2

All code is written in Scala

Scalable Architectures

Ok, so what does it do?

- Stage 1:
  - Access the *free* twitter API to make use of streaming, high volume data
  - Apply some simple parsing to extract information for each twitter message
  - Push the data as a JSON into a Kafka topic



Here is the interesting part…

- Stage 2:
  - Apache Spark listens to the topic
  - Extracts the tweet message
  - Applies SENTIMENT ANALYSIS to it using a trained model and a well known JAVA NLP project (CoreNLP)

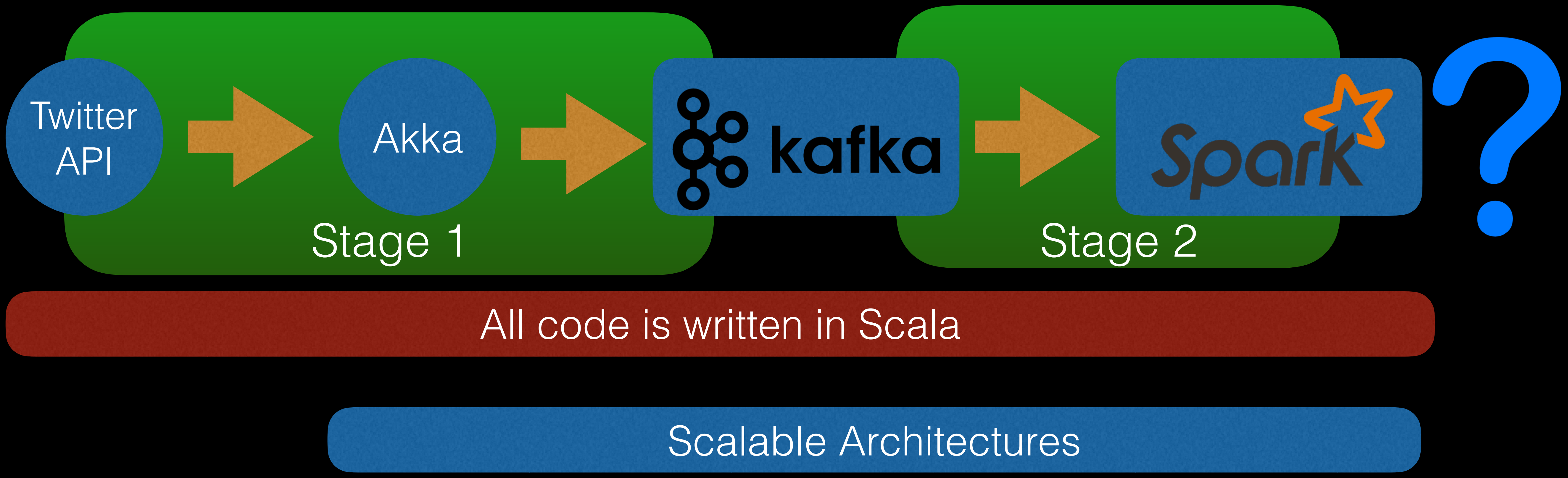
And…

- The custom sentiment model is trained by YOU using freely available sources
    - actually its a requirement, otherwise it won’t run…
- And this is all ready to be run
    - right now.

To summarise!

- So far the ground work as been laid out for an interesting project to explore “Big Data”, ML and often most importantly AWS
- A seemingly end-less (almost) project that you can immediately start working on!

Whats left?



Twitter API

Akka

kafka

Spark

Stage 1

Stage 2

All code is written in Scala

Scalable Architectures

- Both stages can be placed into Dockers
  - EC2 Container Service
- Each element can be replaced by an AWS equivalent
  - Kafka => SQS
  - Akka => Kinesis
  - Spark => Elastic Map Reduced
  - UI   => Chris?

“Once I get on a puzzle, I can't get off.”

– Richard P Feynman

Let’s stay in touch

Dan Dixey

Data Scientist @ QbizUK

dan.dixey@qbizuk.com

part of the QGroup