Spark Streaming for World Domination

(and other projects)

Win Suen

PyGotham 2017

Who am I?

I am Win Suen

Data Scientist @ AppNexus.
Likes: Hiking, chocolate, book-reading, and fighting fraud.

Contents

1. Why Spark?
2. Cool, but why use Spark Streaming?
3. Oh good, code!
    a. Build basic Spark app to stream Twitter firehose.
4. Conclusion.

What is Spark?

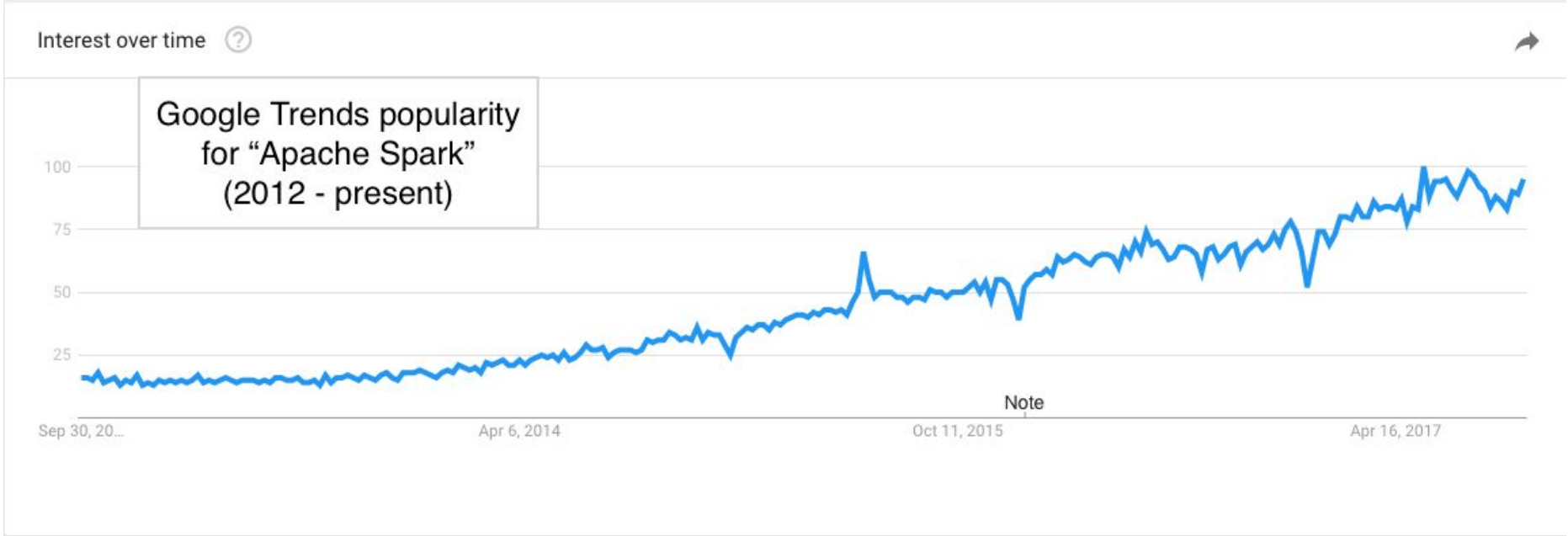
- Cluster computing platform
- Open source
- Many APIs, including PySpark

Fast

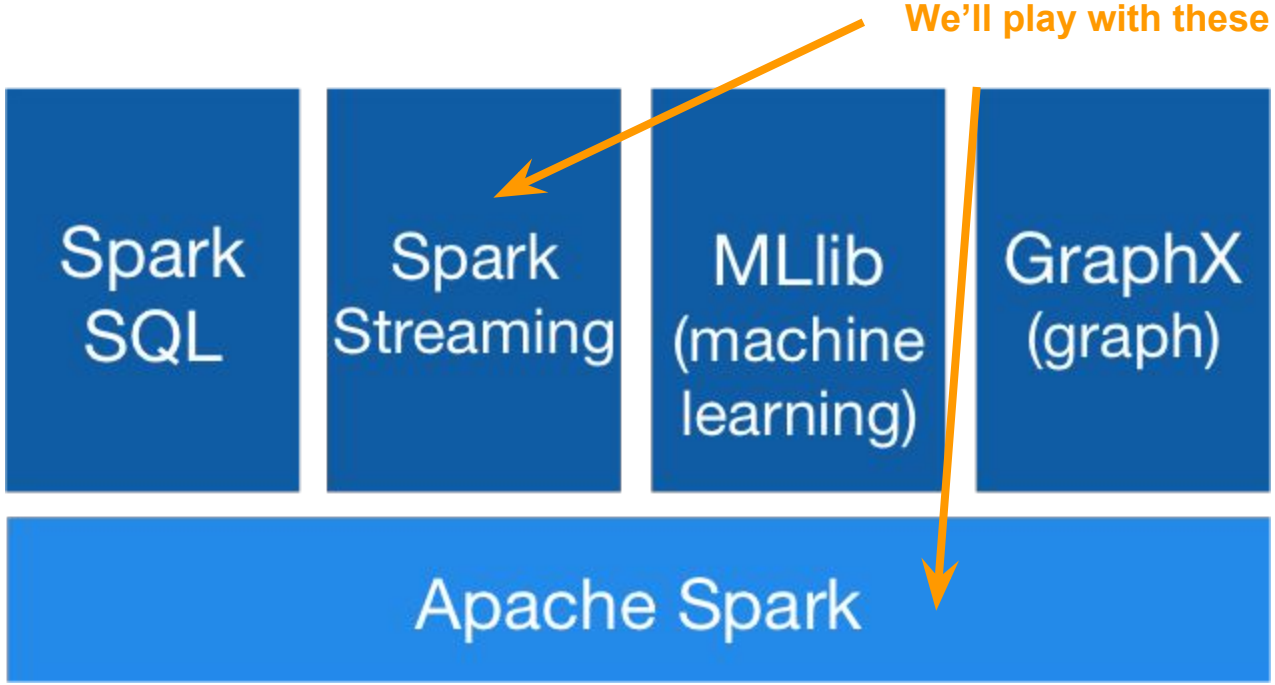
Powerful

General purpose

Spark: one of the cool kids



Spark Libraries

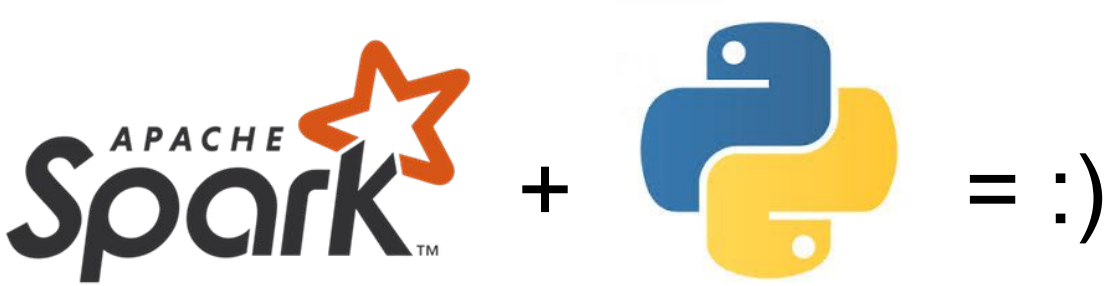


Why streaming?

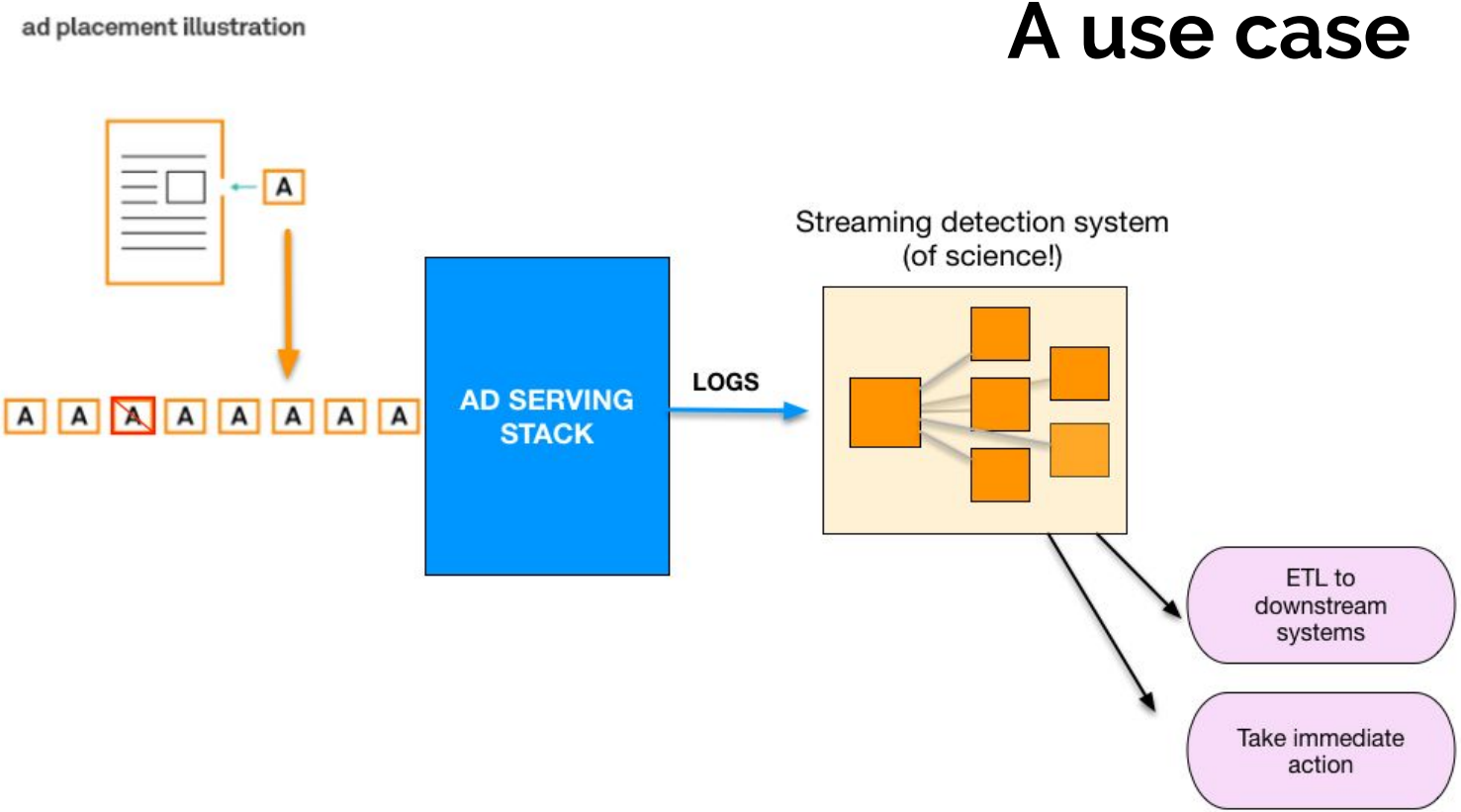
- Spark Streaming features
  - Parallelism in data input
  - Parallelism in data processing
  - Fault tolerance
  - Powerful libraries
  - Near real-time processing

Why streaming?

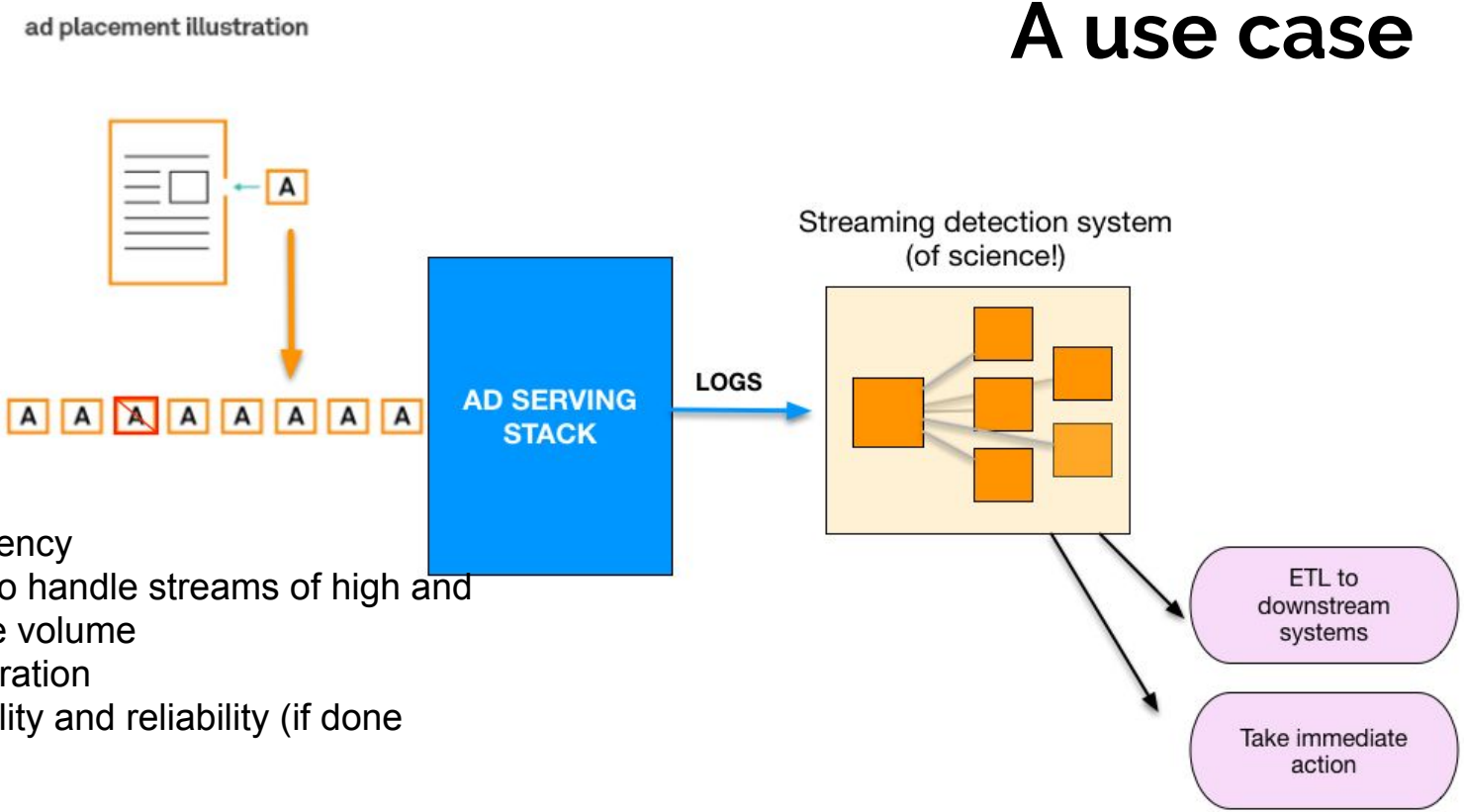
- Python features
  - Ease of use and code sharing
  - Fast prototyping
  - Ability to integrate existing code



A use case


ad placement illustration
A
A A A A A A A A
AD SERVING STACK
LOGS
Streaming detection system (of science!)
ETL to downstream systems
Take immediate action

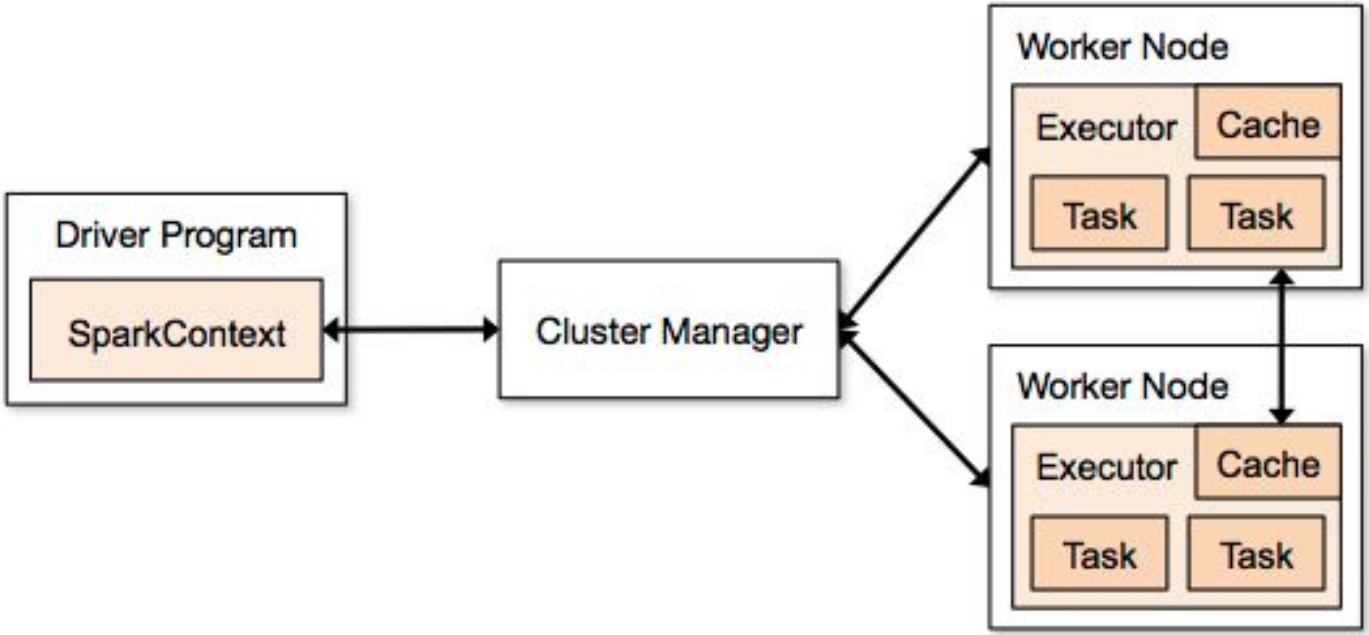
A use case



WANTS

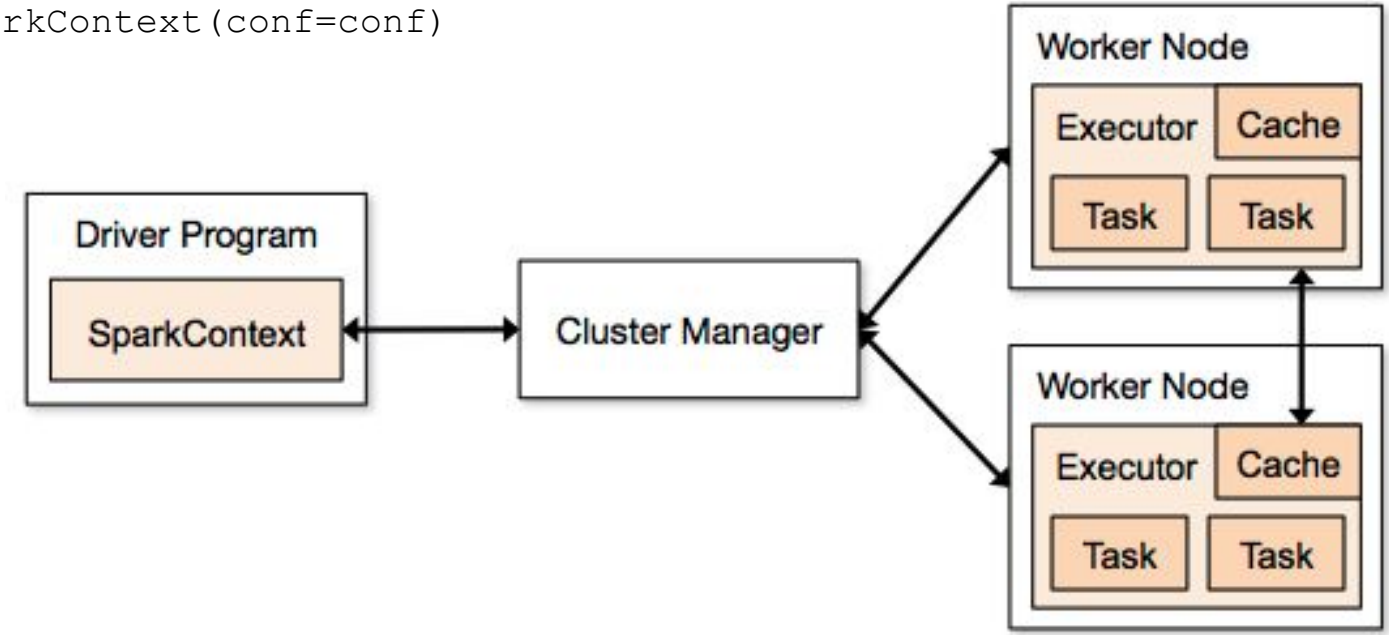
- Low latency
- Ability to handle streams of high and variable volume
- Fast iteration
- Scalability and reliability (if done right)

Sparkitecture



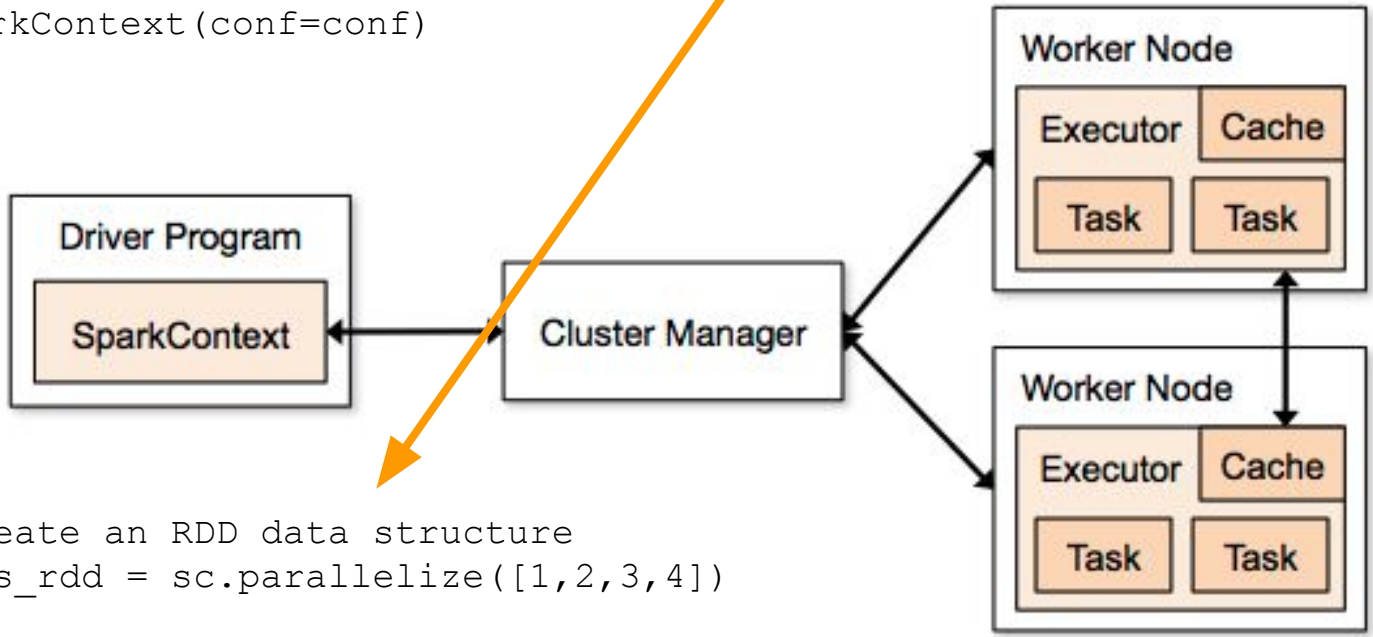
```
import pyspark as ps
from pyspark import SparkConf
from pyspark import SparkContext

conf = SparkConf().setMaster(master).setAppName("foo")
sc = SparkContext(conf=conf)
```

This creates a SparkContext. You need this.



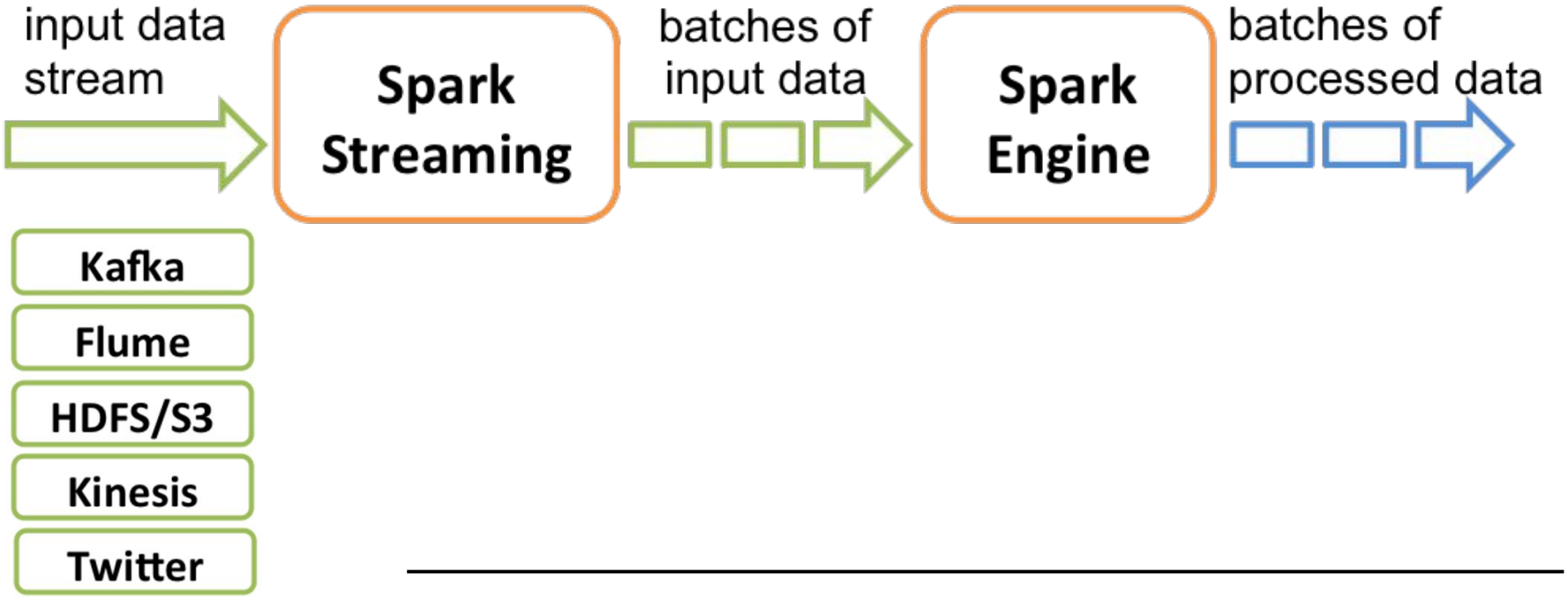
```
import pyspark as ps
from pyspark import SparkConf
from pyspark import SparkContext

conf = SparkConf().setMaster(master).setAppName("foo")
sc = SparkContext(conf=conf)
```

After creating Spark Context, define data input into Resilient Distributed Dataset (RDD).

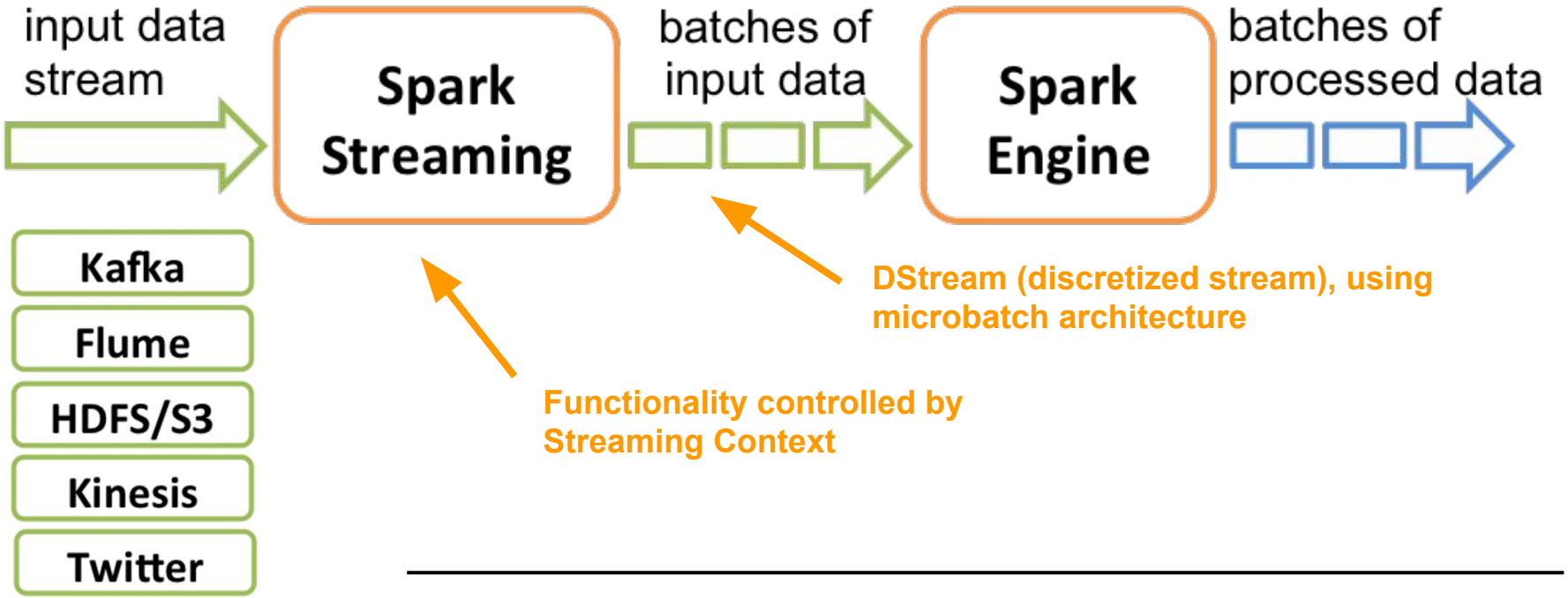



```
#create an RDD data structure
wins_rdd = sc.parallelize([1,2,3,4])
```

Spark Streaming



Spark Streaming



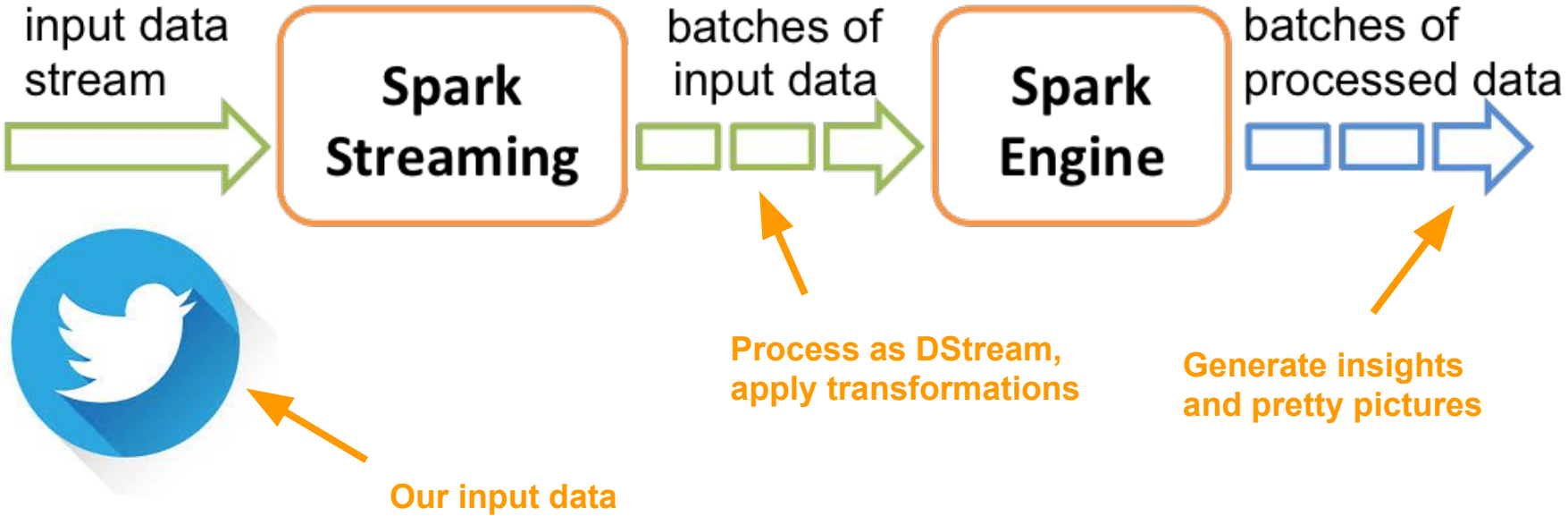
```
#plain ol’ Spark app
from pyspark import SparkContext
sc = SparkContext("local[2]", "winAwesomeApp")

#data input (in the form of RDDs)
myData = sc.parallelize([1,2,3,4,5,6])

#stuff to do
myCount = myData.count()
print myCount
```

```
#Spark Streaming
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
sc = SparkContext("local[2]", "winAwesomeStreamingApp")
ssc = StreamingContext(sc, 60)

#data input (in the form of DStreams)
lines = ssc.socketTextStream("localhost", 9999)

#stuff to do
myCount = lines.count()
print myCount

#start streaming context
ssc.start()
```

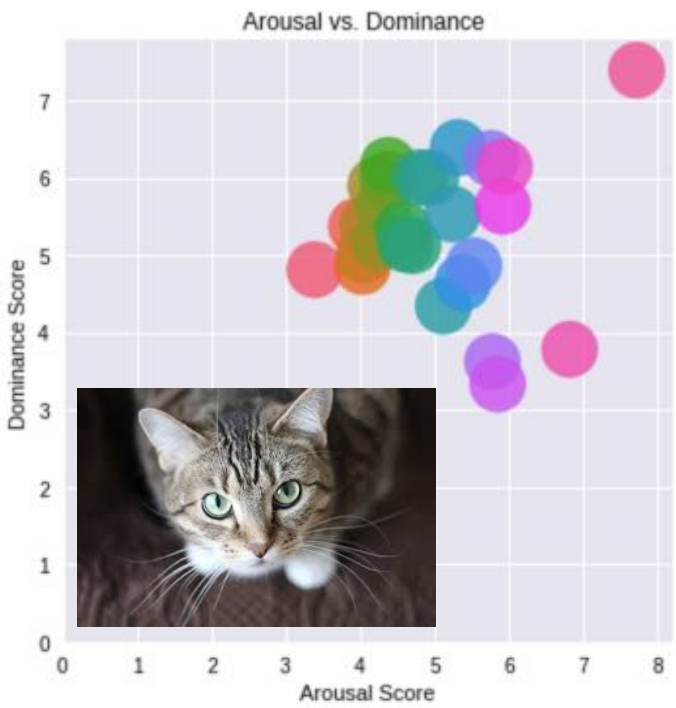
Demo!

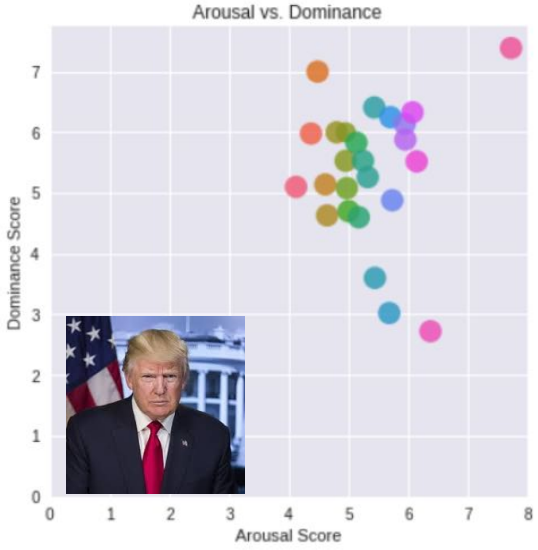

input data stream
Spark Streaming
batches of input data
Spark Engine
batches of processed data

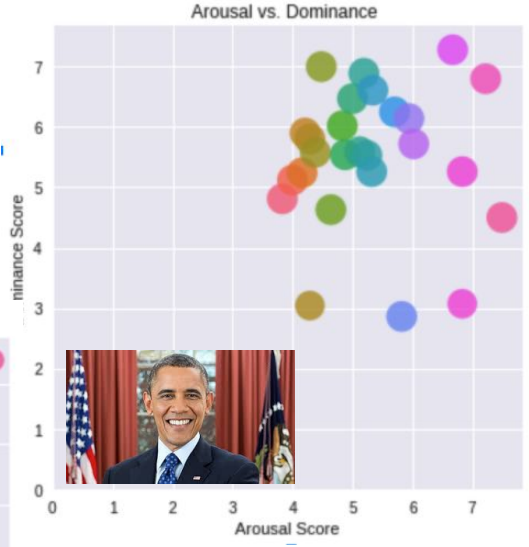
Demo!

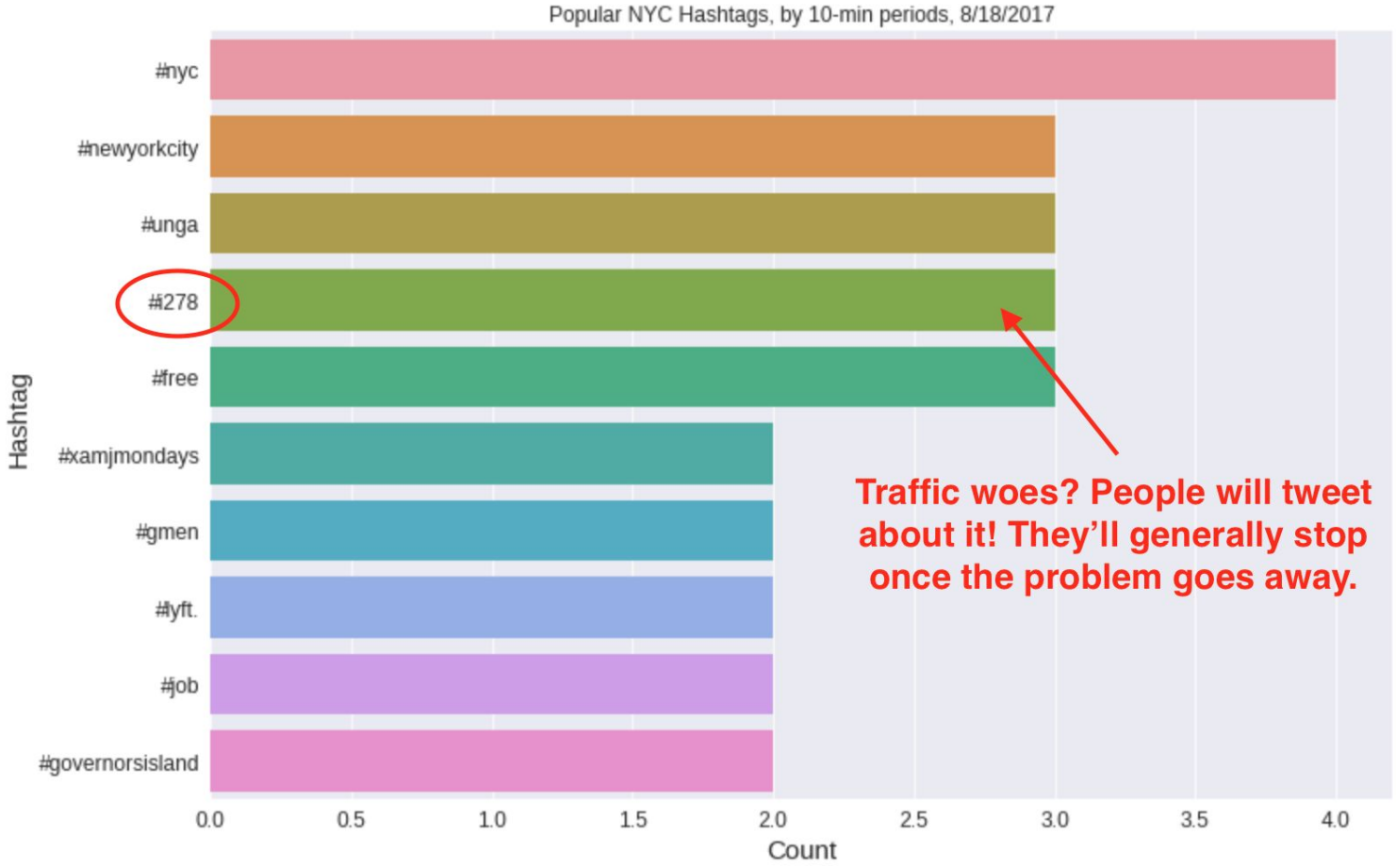

input data stream
Spark Streaming
batches of input data
Spark Engine
batches of processed data
Process as DStream, apply transformations
Generate insights and pretty pictures
Our input data

Demo!


Arousal vs. Dominance
Dominance Score
Arousal Score
0
1
2
3
4
5
6
7
8


Arousal vs. Dominance
Dominance Score
Arousal Score
0
1
2
3
4
5
6
7
8


Arousal vs. Dominance
Arousal Score
0
1
2
3
4
5
6
7


Popular NYC Hashtags, by 10-min periods, 8/18/2017
#nyc
#newyorkcity
#unga
#i278
#free
#xamjmondays
#gmen
#lyft.
#job
#governorsisland
Hashtag
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Count
Traffic woes? People will tweet about it! They'll generally stop once the problem goes away.

Recap

1. Why Spark is so cool.
2. What are Spark Streaming superpowers.
3. Learn to use superpowers.

Thank you!