Deep Dive into threat detection in the cloud with Spark and Python

David Melamed, PhD
Senior Research Engineer, CloudLock
dmelamed@cloudlock.com @dvdmelamed

PyCon Israel 2016, May 2-3 2016

Who is this guy?

CloudLock

ANSIBLE

docker

Spark

Where is he working?

CloudLock

157K	10 M	4 B
APPS	USERS	ACTIVITIES

Founded: 2011

Corporate Headquarters: Waltham, Mass. (U.S.A.)

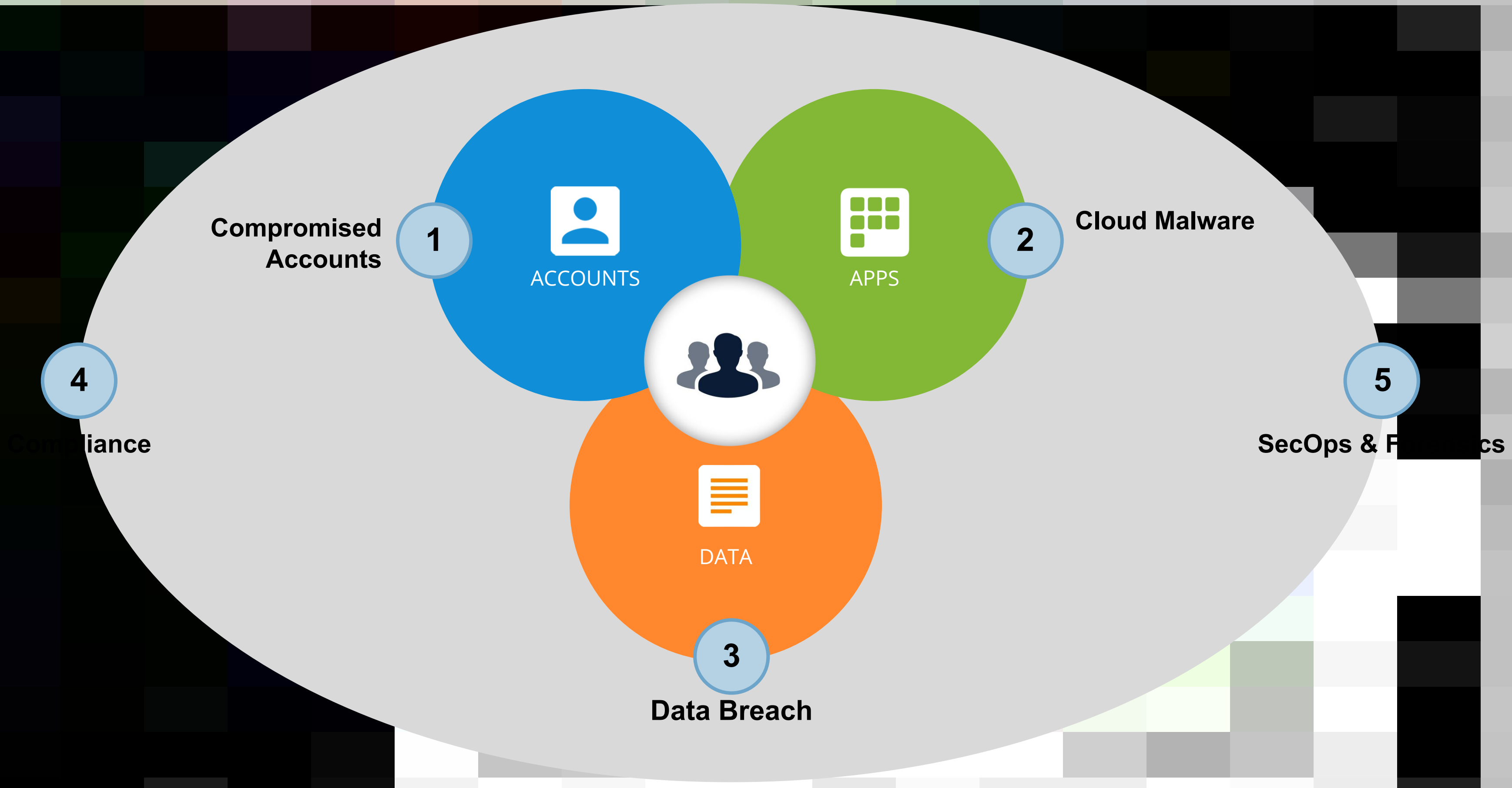
R&D Headquarters: Tel Aviv

Employees: 140 (30 in TLV)

Trusted by major brands:

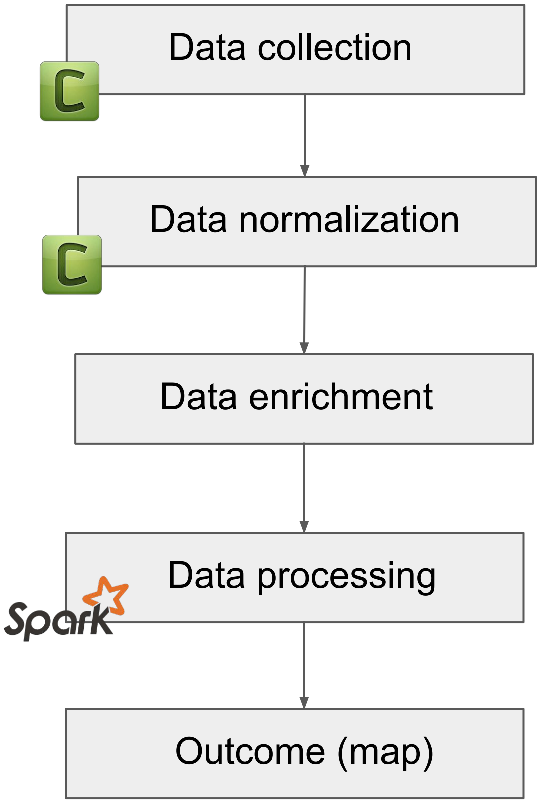
Brand	Users	Use case	Industry
GSA	20,000	Data Privacy	Federal
Woolworths (Australia's fresh food people)	250,000	PCI-DSS	Retail
BBVA	140,000	Reg Compliance	Financial Services
Akamai	5,000	Trade Secrets	Technology
Dominion Enterprises	4,300	PCI-DSS	Retail
Sigma-Aldrich	10,000	PHI/IP	Life Sciences
ASU	750,000	PII / FERPA	Education
Yahoo!	12,000	Data Privacy	High Tech
ACI Universal Payments	4,800	Reg Compliance	Financial Services
Enterprise rent-a-car	80,000	PII / PCI	Transportation
Whirlpool	27,000	Data Privacy	Manufacturing
(cloud)	72,000	Trust	Cloud Vendor

CyberSecurity in the cloud - the big picture

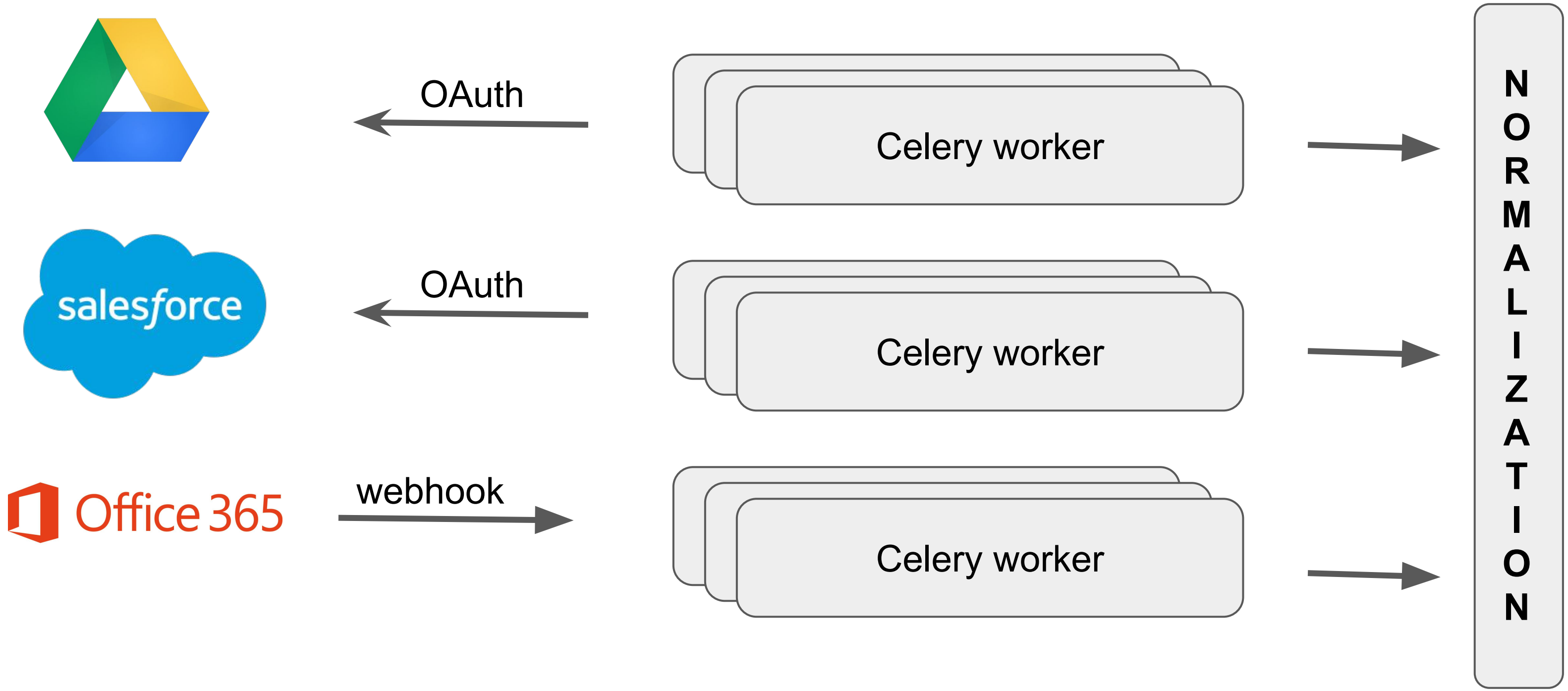


CyberSecurity in the cloud - multiple providers

Data Pipeline



Data Collection



Celery - tips when dealing with scale

- If you don’t use the celery task result, use `@task(ignore_result=True)`
- `max_child_tasks_per_node` is important to determine the number of tasks before killing the interpreter (impacts cached data)
- Limitation: Celery can only work with 1 RabbitMQ
- Celery heartbeats
- Batching mode (experimental)

Data Cleaning, Normalization and Enrichment

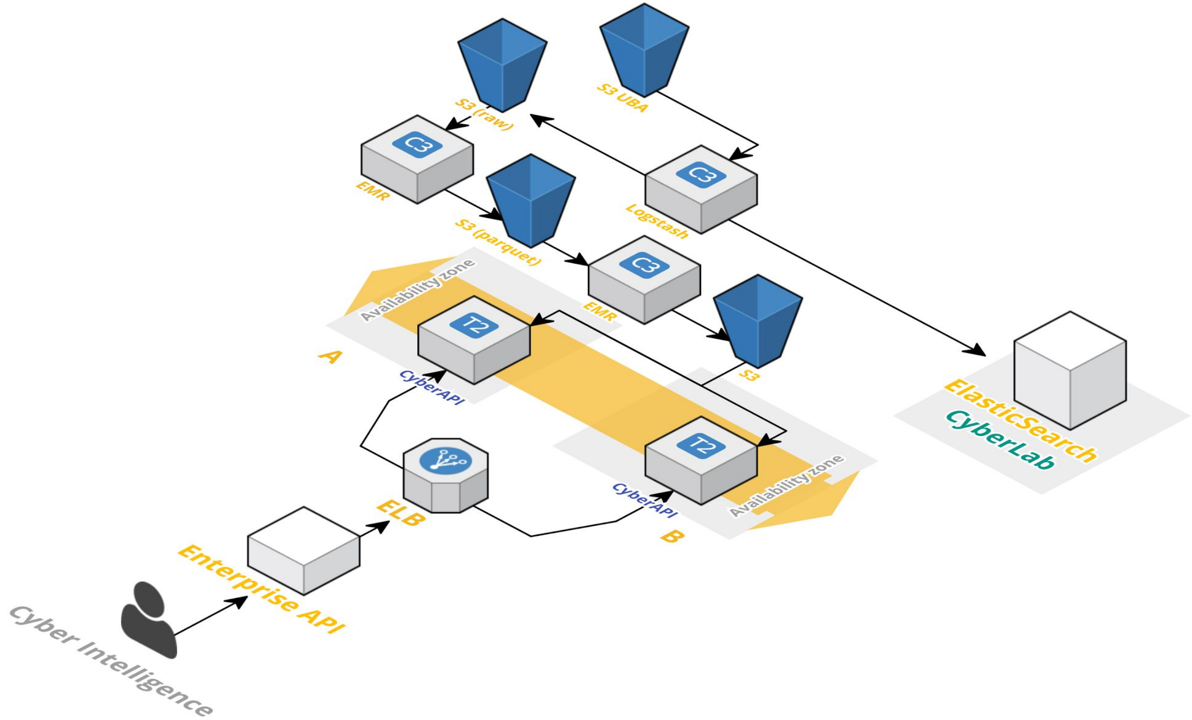
- Multiple sources: fields mapping
- Schema enforcement: use versioning
- Data enrichment for IP address: geolocation data, IP Reputation data
- Other possible enrichment: early sensitive activity detection
- Use *gevent* since we do a lot of I/O

Cyber Intelligence

Data encryption & issues we encountered

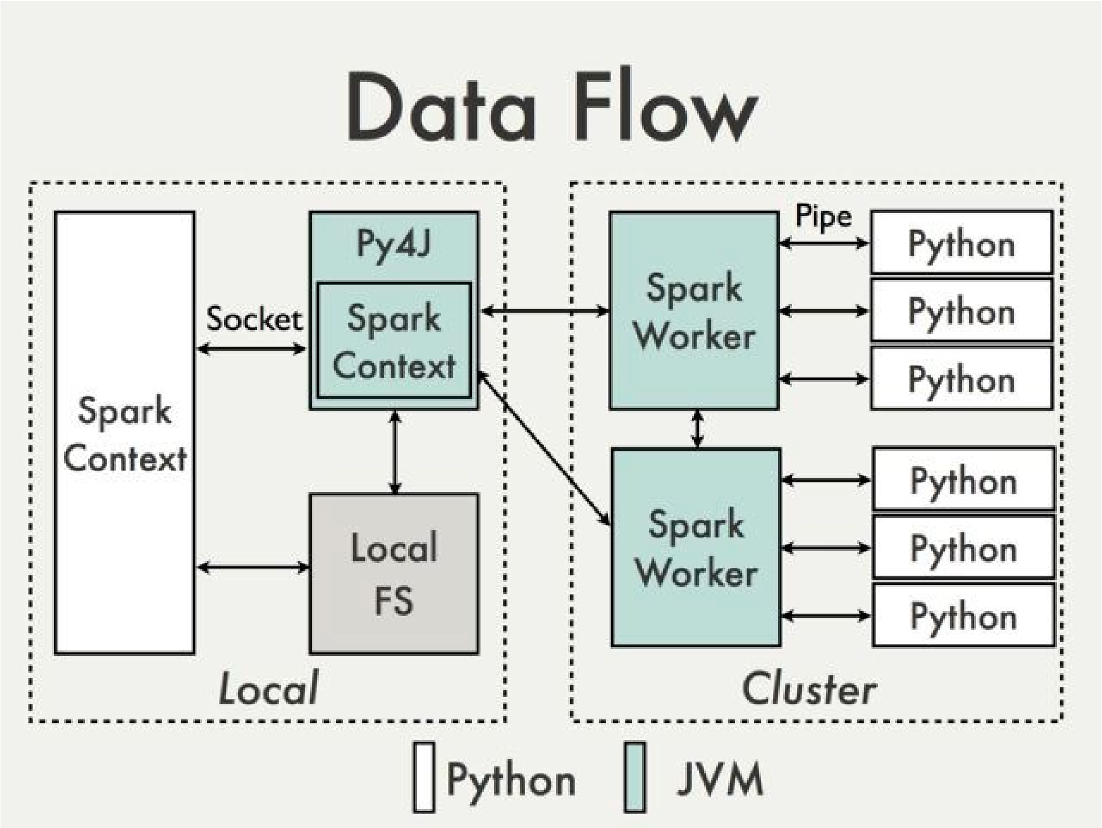
- S3 input (official) - API call for each bucket object
- S3 input (official) - latency between files to process
- S3 input (non official) - only initial latency but stops when no new file
- Encryption filter - no multi-core support

Batch data processing & serving architecture overview

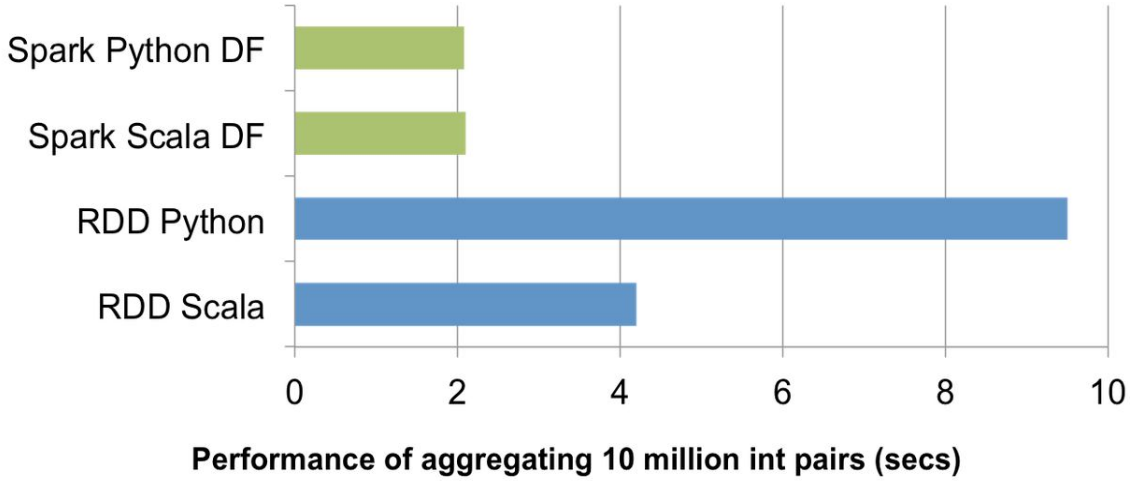


POWERED BY CLOUDCRAFT.CO

Spark Python and JVM



Spark Performance - Python vs. Scala



Source: Databricks

Parquet file format

- Open-source format
- Supported natively by Spark
- Loads only the columns needed
- Data is compressed
- Supports nested data structure
- Schema merge support
- Implementation in Java (MapReduce) and C++

```
message root {
  optional binary @timestamp (UTF8);
  optional binary @version (UTF8);
  optional binary asset_asset_id (UTF8);
...
  optional binary principal_user_user_email (UTF8);
  optional binary principal_user_vendor_id (UTF8);
  optional binary timestamp (UTF8);
  optional int64 version;
}
```

```
{
   "title": "Lord of the Rings",
   "author": "J. R. R. Tolkien",
   "year": 1937
},
{
   "title": "The Hitchhiker's Guide to the Galaxy",
   "author": "Douglas Adams",
   "year": 1979
},
{
   "title": "A Game of Thrones",
   "author": "George R.R. Martin",
   "year": 1996
}
```

Row-based storage

title1 | author1 | year1 | title2 | author2 | year2 | title3 | author3 | year3

Columna-based storage

title1 | title2 | title3 | author1 | author2 | author3 | year1 | year2 | year3

Spark / S3 - tips when dealing with scale

- Prefer lzio over gzip for input files
- S3 over HDFS adds latency
- Use the Kryo object serializer over Java if possible
- Use data partitioning
- Disable summary metadata for better performance
- Use parquet direct committer

EMR / Spark on YARN tips

- Configuration file on S3
- Make sure to turn on dynamic resource allocation or use EMR Release 4.4+
- Bootstrap file on S3, installs required packages
- Use DevPI to store packages
- Use EMR “steps” for easy access to the logs

Steps

Filter: All steps | Filter steps ... | 1 step (all loaded)

ID	Name	Status	Start time (UTC+3)	Elapsed time	Log files	Actions
s-2E3TUYE6YI14E	event_partition	Running	2016-05-03 09:08 (UTC+3)	8 minutes	controller | syslog* | stderr | stdout	Debugging not configured

Spark SQL example

```
df = sqlContext.read.parquet("s3://bucket-name/*.gz")
df.registerTempTable("events")
num_events = df.sql("SELECT COUNT(*) FROM events").show()
print "Num events: {}".format(num_events)
```

Spark for Machine Learning

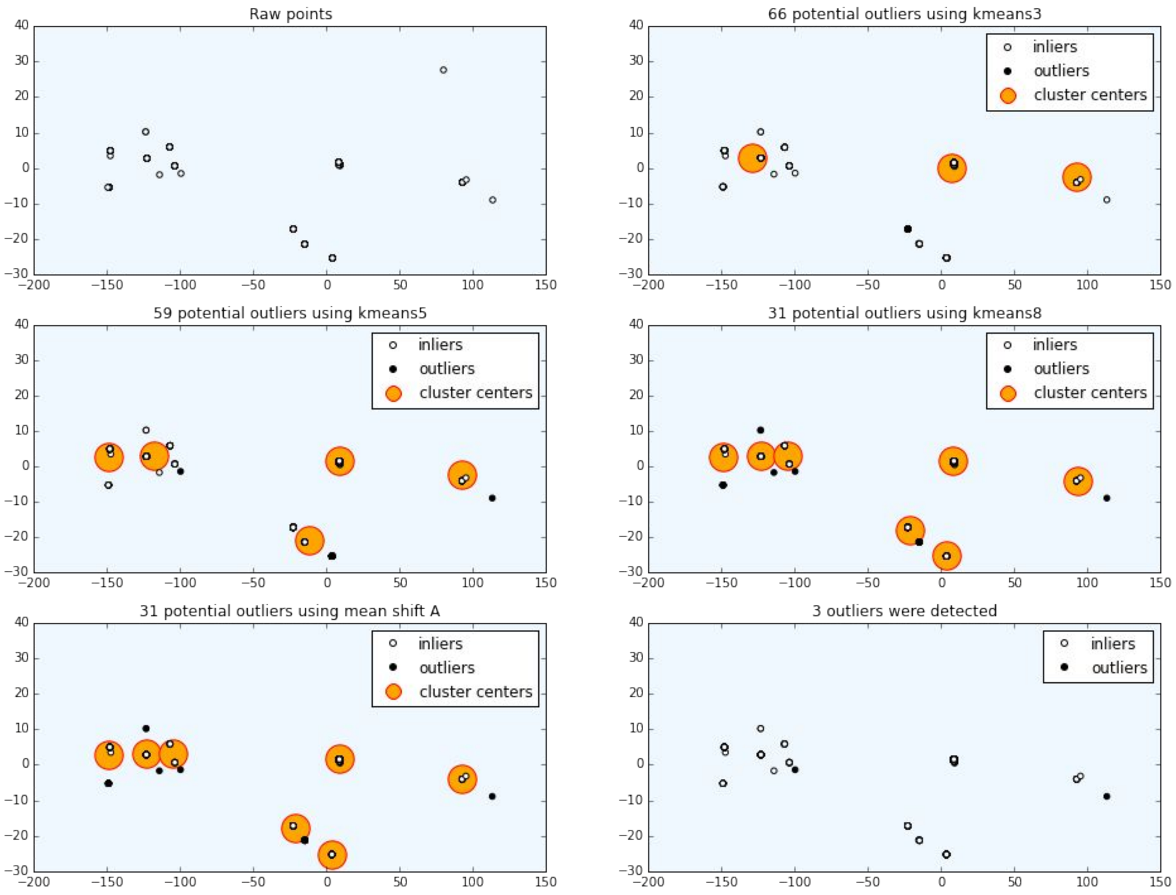
- Spark MLLib (works with RDDs)
  - Binary classification
    - SVM
    - Logistic Regression
  - Linear regression
  - Clustering
    - k-means
  - Collaborative Filtering
  - Gradient descent

- Spark ML (works with DF)

Spark MLLib - K-means example

```
from pyspark.mllib.clustering import KMeans
from numpy import array
from math import sqrt

# Load and parse the data
data = sc.textFile("kmeans_data.txt")
parsedData = data.map(lambda line: array([float(x) for x in line.split(' ')]))

# Build the model (cluster the data)
clusters = KMeans.train(parsedData, 2, maxIterations=10,
        runs=30, initialization_mode="random")

# Evaluate clustering by computing Within Set Sum of Squared Errors
def error(point):
    center = clusters.centers[clusters.predict(point)]
    return sqrt(sum([x**2 for x in (point - center)]))

WSSSE = parsedData.map(lambda point: error(point)).reduce(lambda x, y: x + y)
print("Within Set Sum of Squared Error = " + str(WSSSE))
```

Spark MLLib - Geo based Anomaly Detection



- 5K employee company
- Running an ensemble of clustering algorithms
- Using proprietary distance calculation to locate outliers
- Filter out noise
- Discovered access from 2 abnormal locations (Nigeria and Vietnam) within 4M events for this organization in a specific month

Microservice using Docker

- AWS ECR to store the Docker images
- AWS ECS for the containers (supports auto-scaling and ELB)
- Ansible to deploy

Ansible at Scale

Monday, May 9, 2016

http://www.meetup.com/Ansible-Israel/events/230313714/

Cloud Threat map

Search:

Threat	Platform	Date	City	State	Country	IP Address	Description
Location Anomaly	Google	February 28th 2016, 15:05 pm	Abu Dhabi		United Arab Emirates		Anomalous activity from Abu Dhabi (United Arab Emirates)
Location Anomaly	Google	February 28th 2016, 15:03 pm	Buenos Aires		Argentina		2 Anomalous activities from Buenos Aires (Argentina)
Location Anomaly	Google	February 28th 2016, 15:03 pm			India		17 Anomalous activities from India