Big Data Real - Time Security Analytics

Master Thesis
by **Neofytos A. Kountardas**
Supervisor Professor: **Kostas Psannis**

Personal info

My BackGround

»Master in Applied Informatics, University of Macedonia, 2015-2017
»Bachelor of Economics, University of Macedonia, 2010-2014
»Hellenic Police Academy (Officers' School), 2006-2010

My ambition®

Mentally & morally participate in establishing an equally and worldwide accessed secure cyber space, by deploying enhanced and contemporary big data tools and analytics

University ID

mai16076

Contents

When IoT, Big Data & Cloud Computing became ubiquitous…

Daunting grim thoughts emerged among Security and Privacy experts

1.

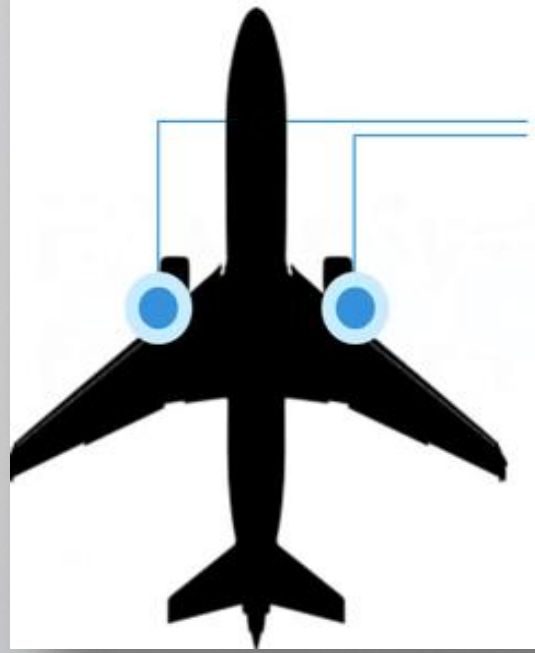
Focus on Big Data

Let’s grasp the volume and the velocity

2,499,841,200

Senior Director of Marketing, Aerospace and Defense at SAP
Thomas Pohl estimated…

2,7 Zettabytes only from US Commercial Flights?



20 TB ✖ 2 ✖ 6 ✖ 28,537 ✖ 365

20 terabytes of information per engine every hour	twin-engine Boeing 737	six-hour, cross-country flight from New York to Los Angeles	# of commercial flights in the sky in the United States on any given day.	days in a year

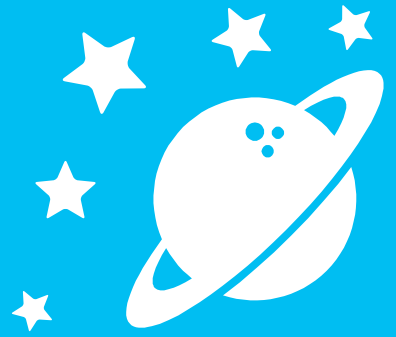
= 2,499,841,200 TB

2.

Why Security?

Security

Outdated!!!



PwC Global Economic Crime Survey 2016

CyberCrime…

Is the Second most reported economic crime after misappropriation

32% of organisations

Had already been affected by cybercrime

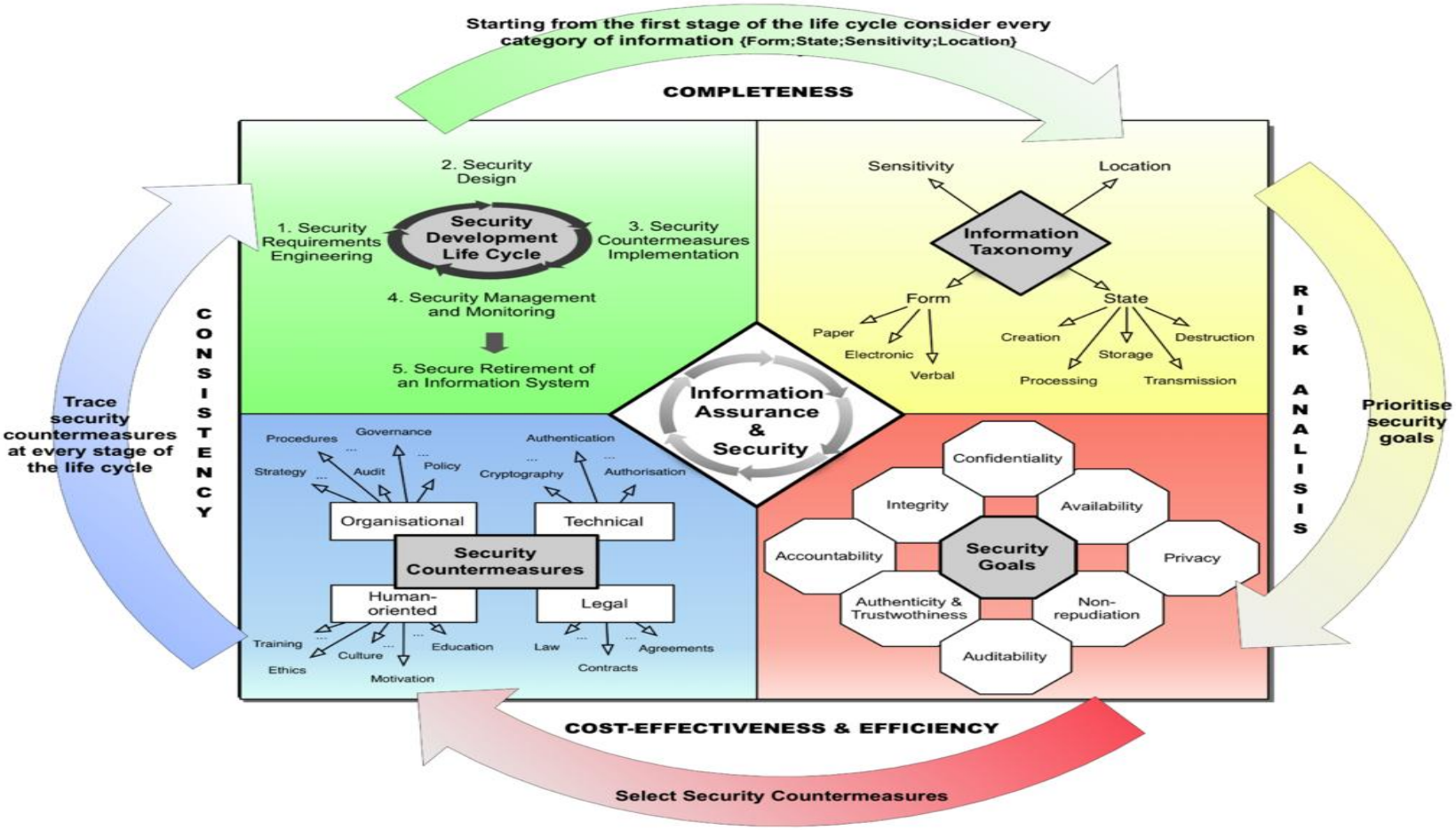
Over 1.000.000$ losses

Reported 14% of organisations (1% was over 100m$)

3.

A model to approach Security

A Reference Model of Information Assurance & Security (RMIAS)



4.

Let’s deal with big data challenges

Making Big Data processing and computing more Secure (CSA, 2016)

Infrastructure Security

-Secure computations in distributed programming frameworks

- Security best practices for non-relational data stores

Data Privacy

-Privacy preserving data mining and analytics

-Cryptographically enforced data centric security

-Granular access control

Integrity & Reactive Security infrastructure

-Secure data storage and transactions logs

-Granular audits

-Data provenance

Data Management

-End-point input validation/filtering

-Real-time security monitoring

5.

What about IoT?

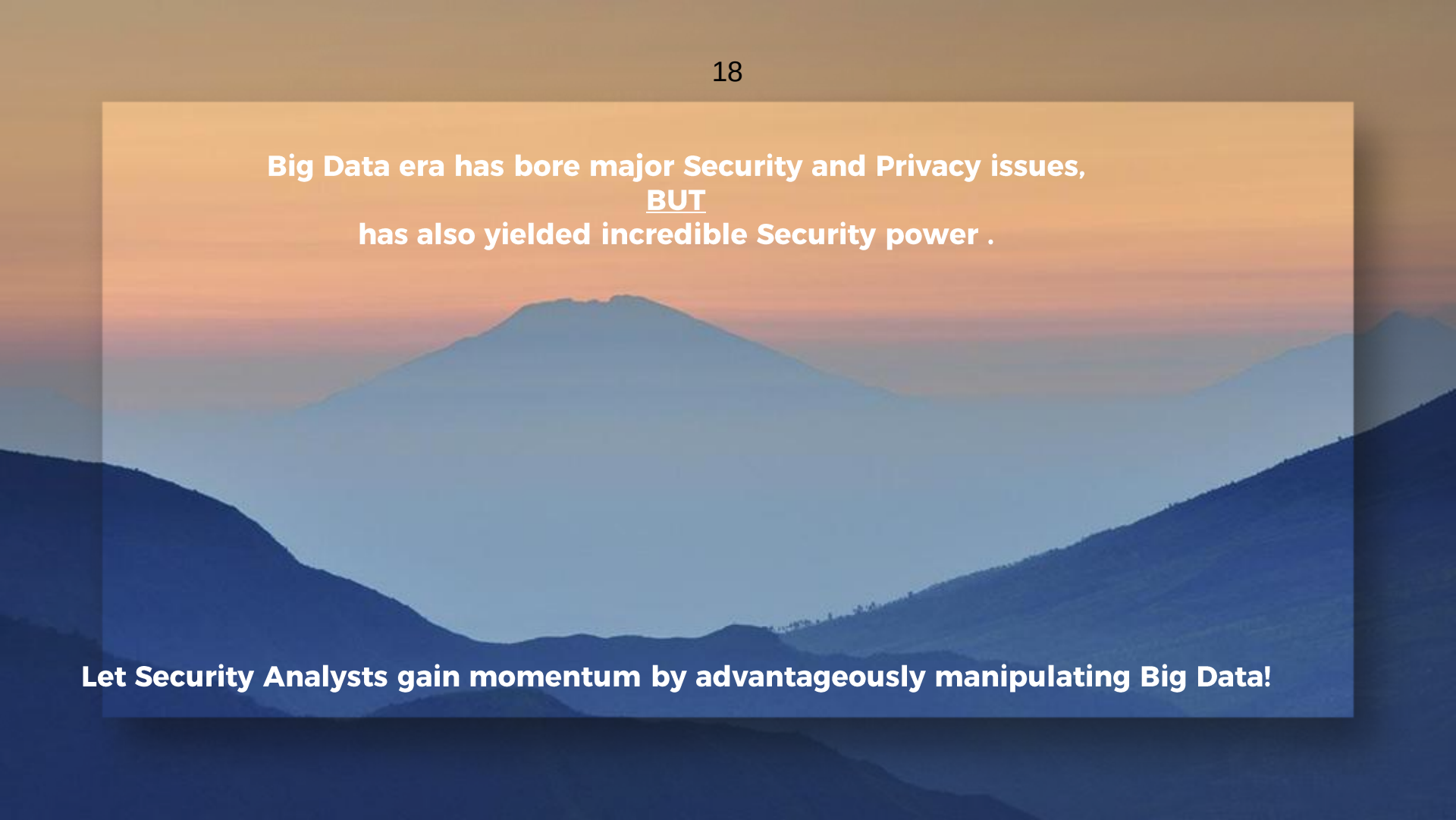
Sound and safe?

Pretty Much Every Smart Home Device You Can Think of Has Been Hacked
(L. H. Newman, 2014)

Reconsider your Privacy at home…

Not only widespread internet references, but also detailed tutorials on hacking…

- Networked light bulbs
- Nest thermostats
- Fitness trackers (FitBits)
- Toasters
- Refrigerators
- **Smart TVs** (bbc, forbes …)



Big Data era has bore major Security and Privacy issues,
BUT
has also yielded incredible Security power .

Let Security Analysts gain momentum by advantageously manipulating Big Data!

My Master Thesis' Methodology

Many Security related notions demystified

Intelligence, Business Intelligence & Big Data Analytics

Business Intelligence and Analytics (BI&A)

Firewalls and IDS/IPS

Security intelligence & Analytics(SI&A)

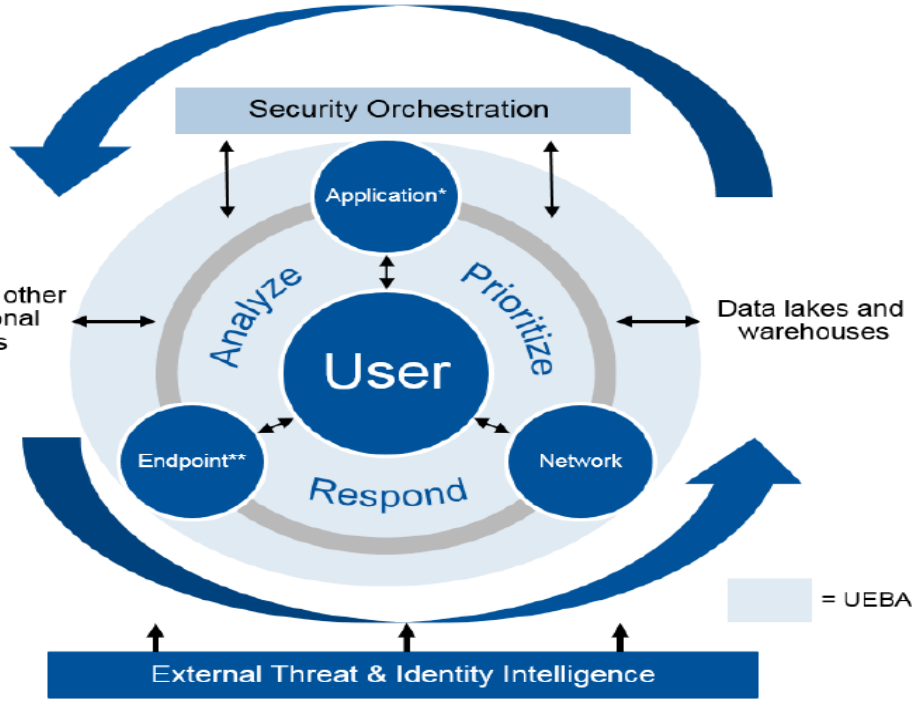
Cyber Threat Intelligence & Analytics(CTI&A)

SIEM
Security information and event management

NBAD
Network behavior anomaly detection

Network Forensics

APTs



* includes cloud, mobile and other on-premises applications
** includes managed and unmanaged endpoints

Source: Gartner (September 2015)

By 2017
according to
GARTNER

200,000,000$

UEBA market Revenue

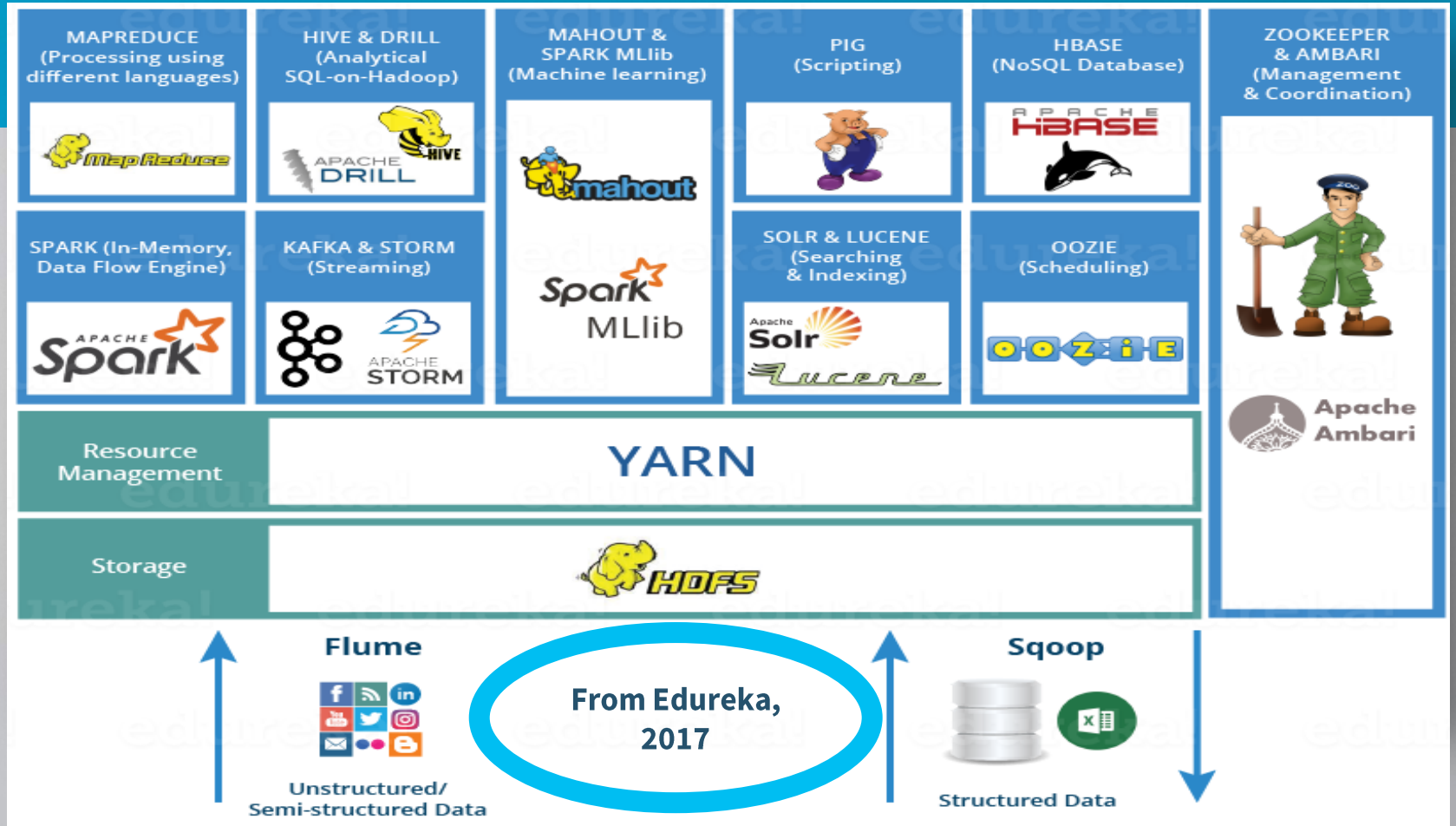
60% Cloud Access Security Brokers
25% SIEM and DLP vendors

…will opt for advanced analytics and UEBA functionality

…Deep Learning into UEBA!

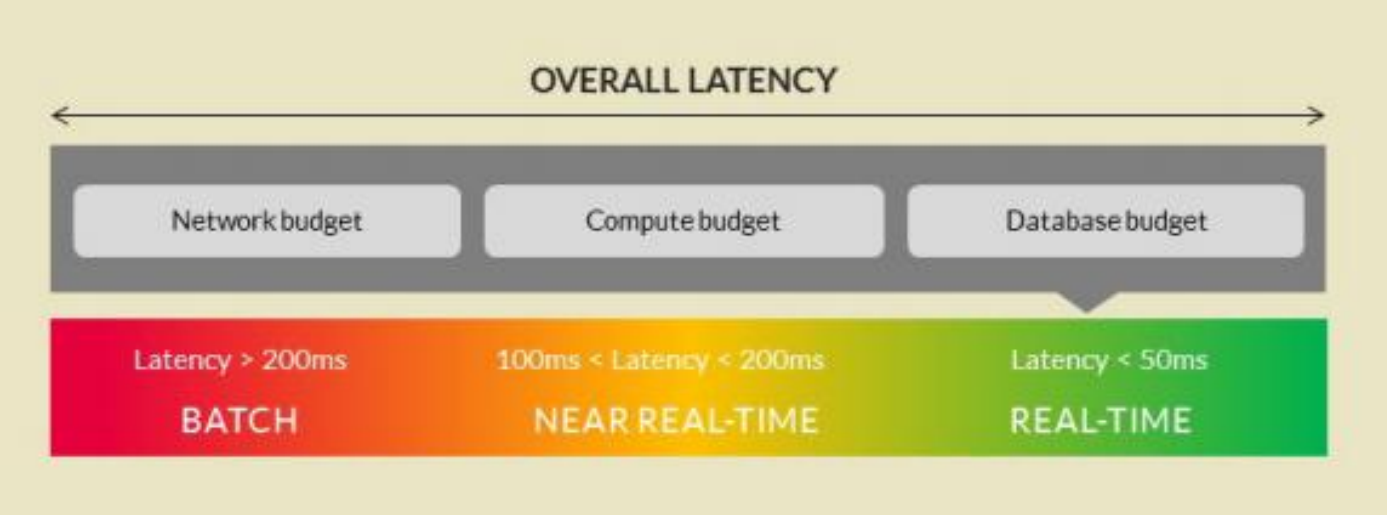
Open Source Solutions for enhanced Security

- Hadoop Ecosystem
- Batch vs Stream
- Stream Analytics


MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
Flume
Unstructured/
Semi-structured Data
From Edureka,
2017
Sqoop
Structured Data

MapReduce is fundamentally a batch processing system, and is not suitable for interactive analysis. You can't run a query and get results back in a few seconds or less. (White, 2012)

Batch vs Stream



Overall Latency (CSA, 2014)

Scalable

»Storm topologies are inherently parallel and run across a cluster of machines.

»Different parts of the topology can be scaled individually by tweaking their parallelism.

Fault Tolerant

»When a part dies, Storm will automatically restart it.

»The Storm daemons, Nimbus and the Supervisors, are stateless and fail-fast.

Main Concepts of Storm

Tuples

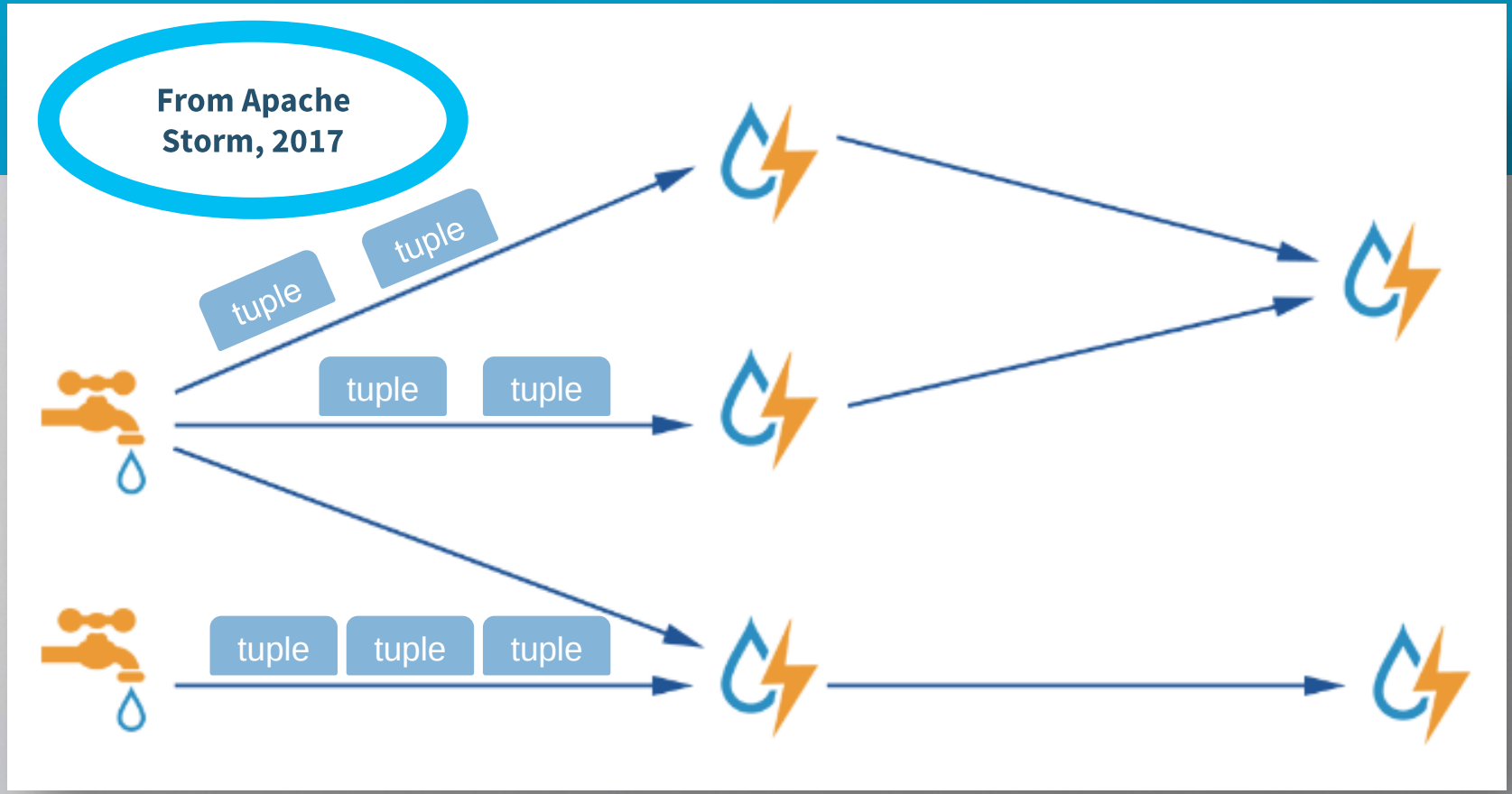
The main data structure in Storm. A tuple is a named list of values

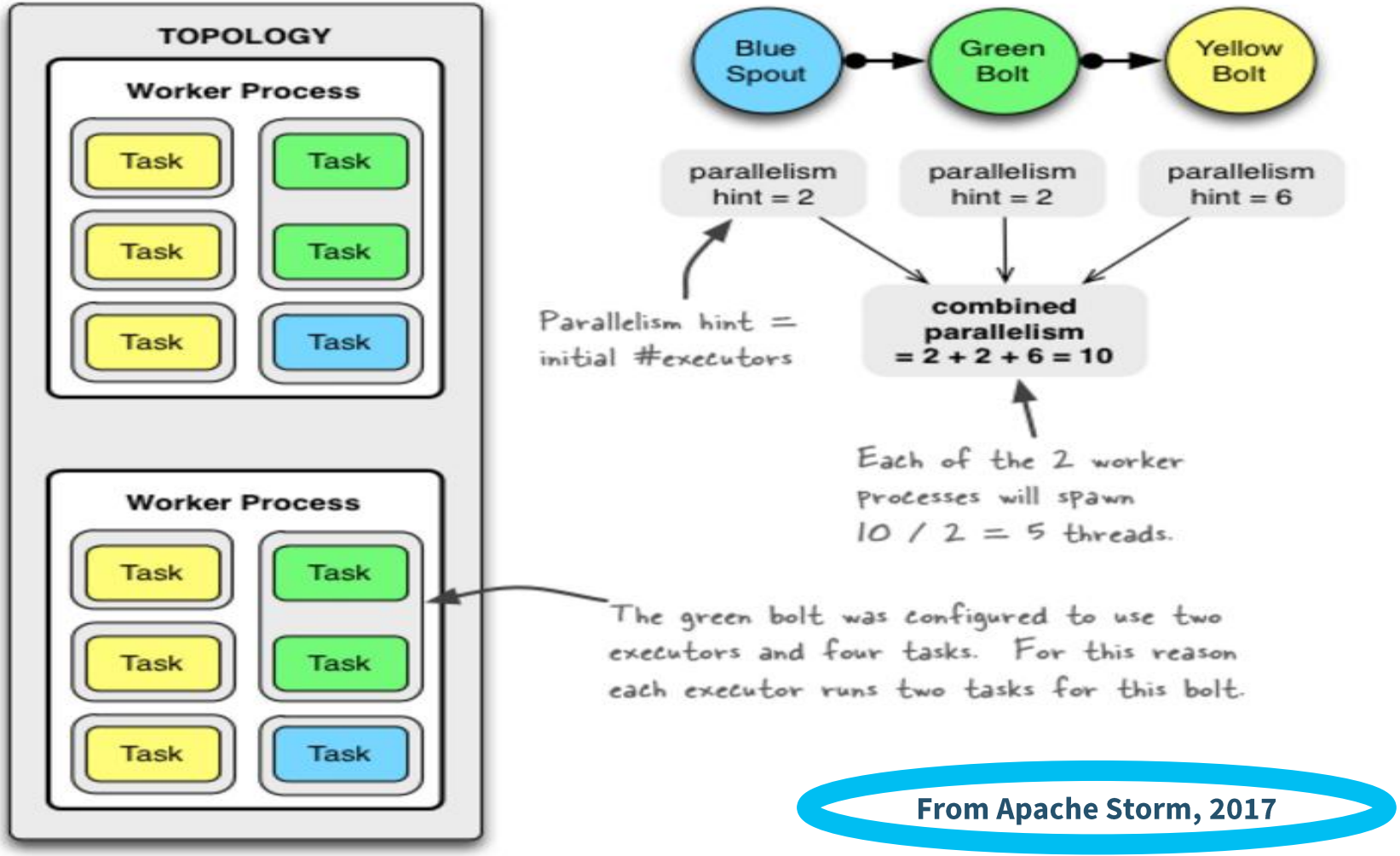
Spouts

A source of streams into topology. Usually spouts read tuples from an external source and emit them into the topology

Bolts

Processing units. Bolts processes input streams and produces new streams . filtering, functions, aggregations, joins, talking to databases, or simple stream transformations.

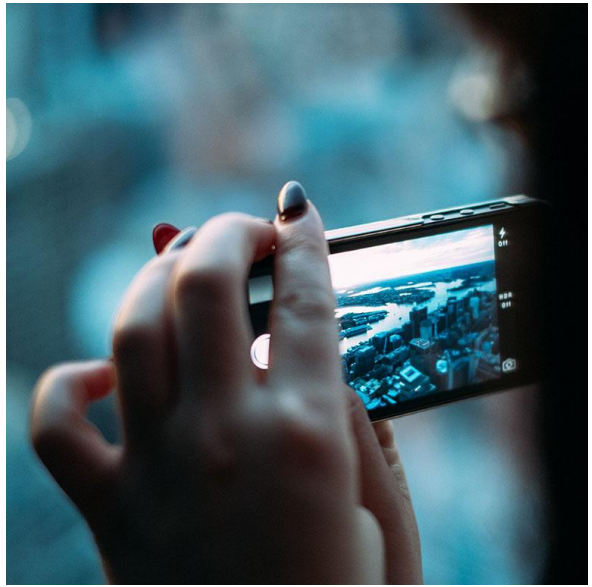

From Apache Storm, 2017
tuple
tuple
tuple
tuple
tuple
tuple
tuple



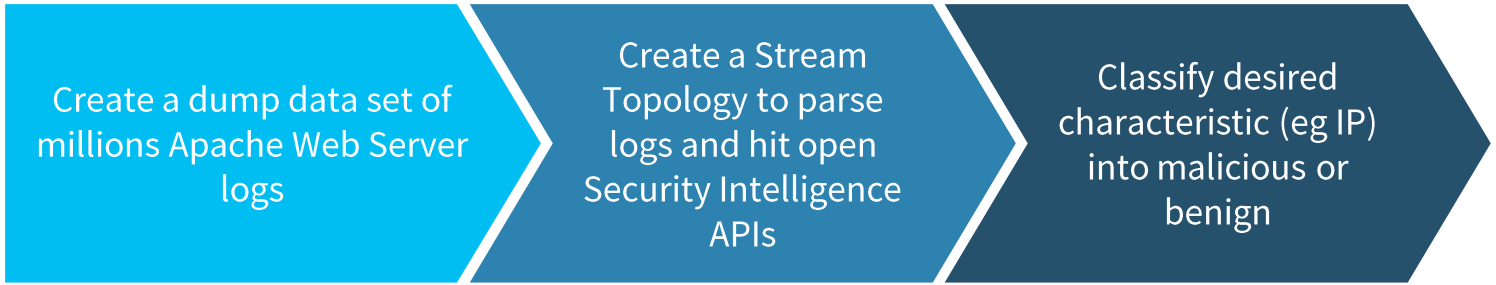
From Apache Storm, 2017

I set a clear goal

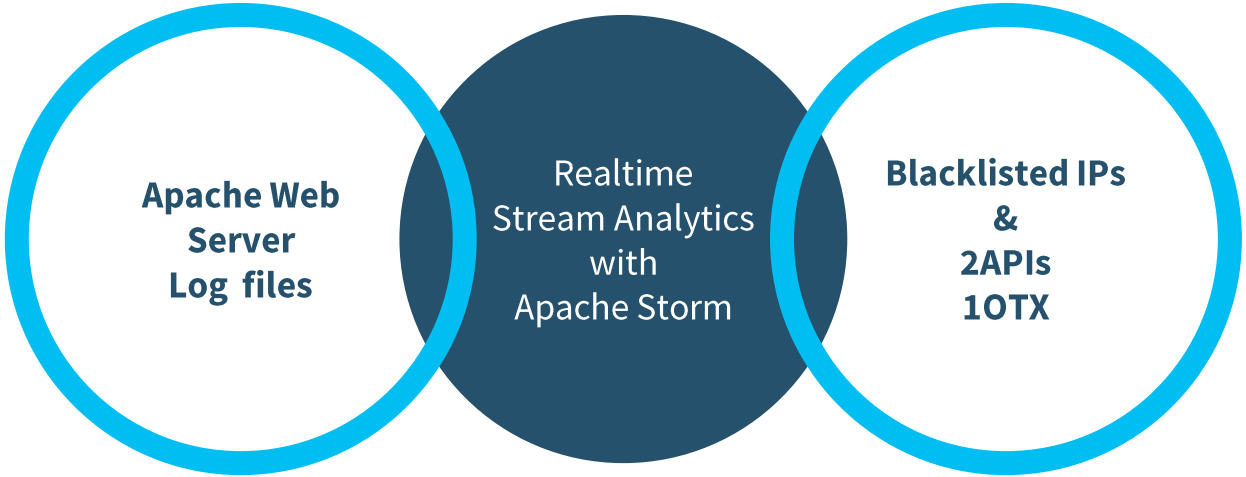
Accomplish a generally security enhanced solution utilizing a novel real-time stream processing tool combined with automatically updated cyber security intelligence.



My Timeline


Create a dump data set of millions Apache Web Server logs
Create a Stream Topology to parse logs and hit open Security Intelligence APIs
Classify desired characteristic (eg IP) into malicious or benign

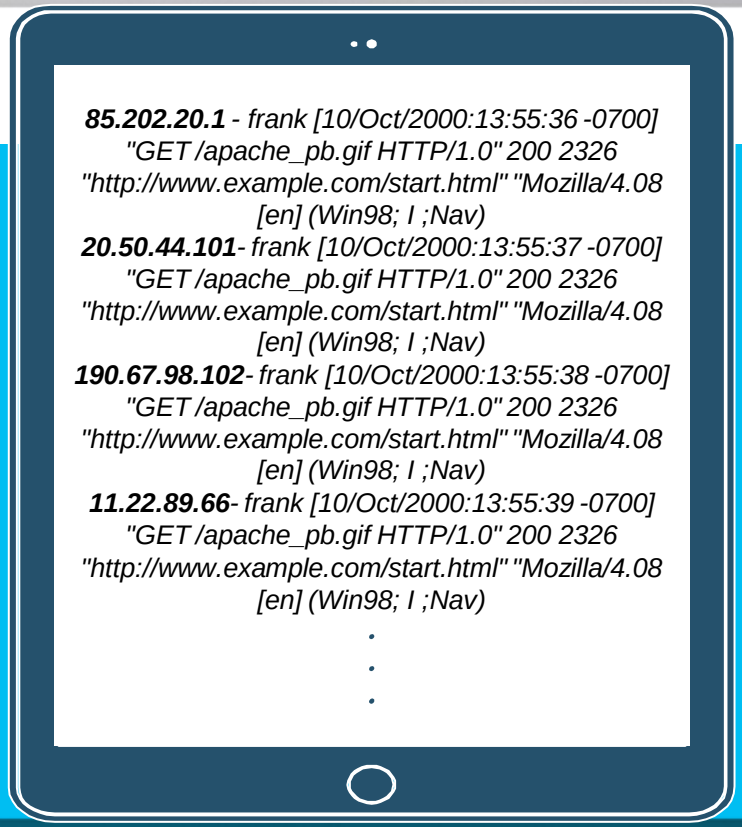
Real-time Big data Security Analytics



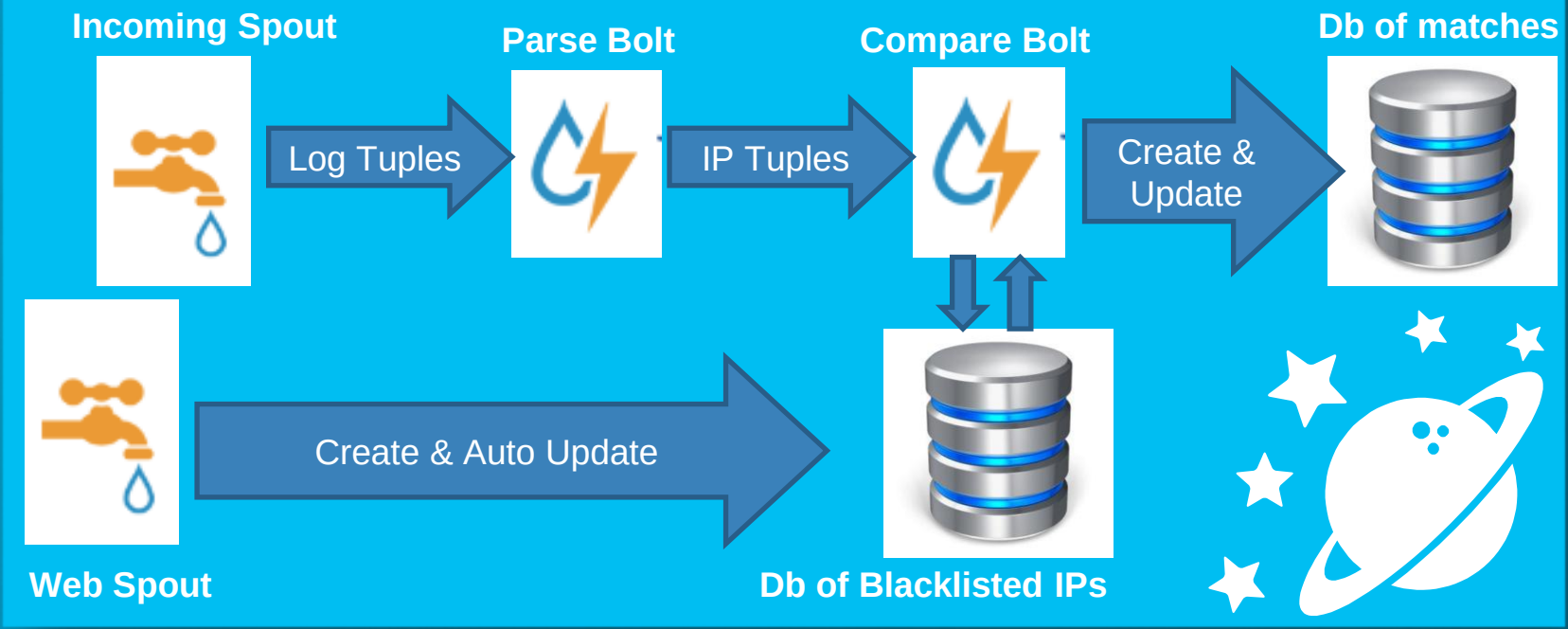
50.000.000 records

Apache Web Server Log File

We Created a tiny Java program to rapidly generate dump Apache Web Server-like log files



My Topology


Incoming Spout
Parse Bolt
Compare Bolt
Db of matches
Log Tuples
IP Tuples
Create & Update
Create & Auto Update
Web Spout
Db of Blacklisted IPs

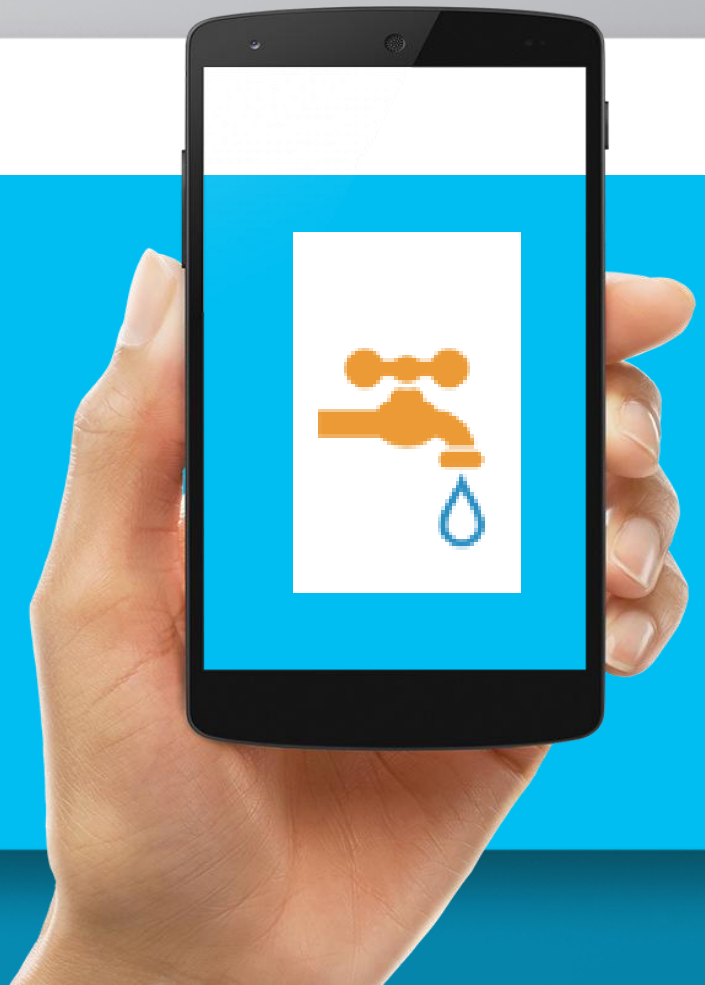
Web Spout

Hits:

2 open APIs &

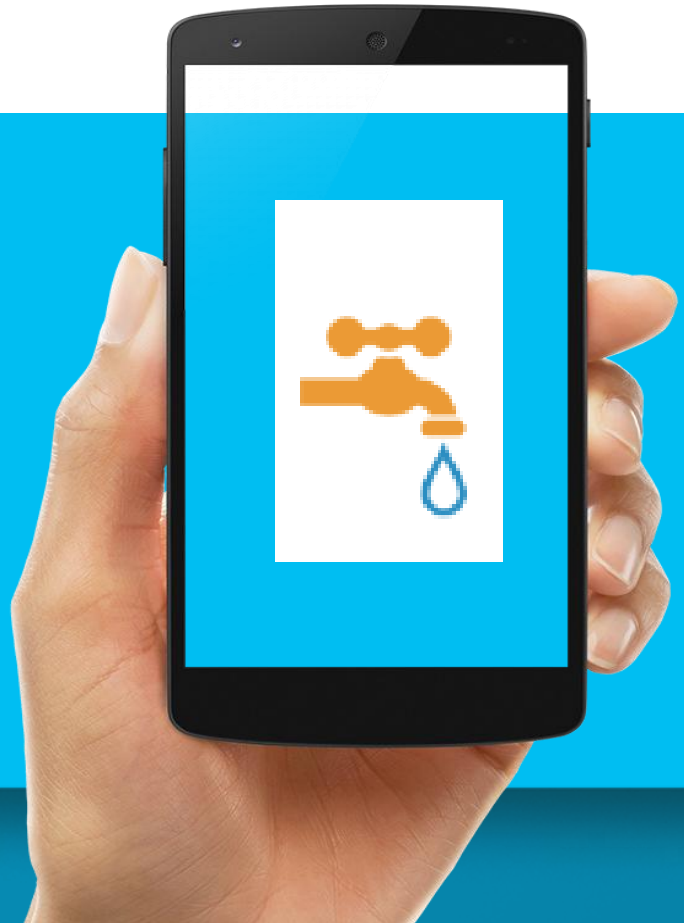
1 OTX Alien Vault pulse.

Downloads, creates & auto updates blacklisted IPs DB



Incoming Spout

Convert Apache log records into tuples and bring them into our topology…



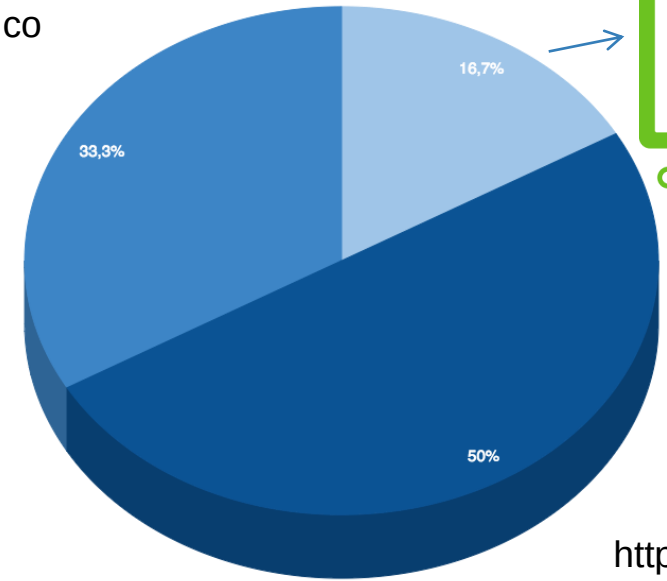
Parse Bolt

```
Static String
ValidIpAddressRegex =
"(?:(?:25[0-5]|2[0-4][0-
9]|[01]?[0-9][0-9]?)\\.){3}(?:25[0-
5]|2[0-4][0-9]|[01]?[0-9][0-9]?)";
```

Compare Bolt

Seeks the incoming IP tuple
within DB of blacklisted IPs.
If found … updates DB of matches
and may drop further packets

http://www.ipspamlist.com/public_feeds.csv"



OPEN THREAT EXCHANGE

40

https://myip.ms/files/blacklist/general/latest_blacklist.txt"

Constructed DB of Blacklisted IPs

Storm provides submitting a topology...

Locally

No nimbus,
Zookeeper or
supervisor needed,
due to the testing
nature of local
submitting.
Not horizontically
scalable.
Time limited.

Regularly

Via Nimbus,
Zookeeper and
Supervisor.
Storm UI updated.
Scalable & Reliable.
Has to be killed.
Ready for Cluster
deployment.

a.

Local submitter

No user interface…

Storm Topology - Our Results!

From totally **17327031 in 159050 seconds** processed Apache log files converted into Storm tuples,
we created secondary tuples fetching only the corresponding IP address targetting our Web Server.
Among that amount of tuples we swiftly distinguished totally **14** potentially malicious attacks!!!
In order to construct our unique and automatically updated Blacklist Database,
we chose **2 open APIs** of well-known blacklisted IPs feeds and **1 OTX Alien Vault** pulse!
Our blacklist database was configured to be regularly updated and during our testings
was updated totally **0** times (updates were set to occur every hour - 3600 seconds).
The number of blacklisted IPs populated into our Blacklist database was on average **9062**.
Finally our Map of malicious IPs to their corresponding attempts to establish connection with our system was registered into a HashMap.
Here we present the created HashMap too:
{109.121.167.229=1, 191.250.51.107=1, 131.161.9.253=1, 190.9.57.159=1, 85.157.119.115=1, 217.92.86.109=1, 201.52.218.97=1, 117.178.160.241=1, 74.120.91.148=1, 176.193.42.29=1, 185.118.154.54=1, 203.91.112.43=1, 63.243.252.179=1, 69.210.226.66=1}

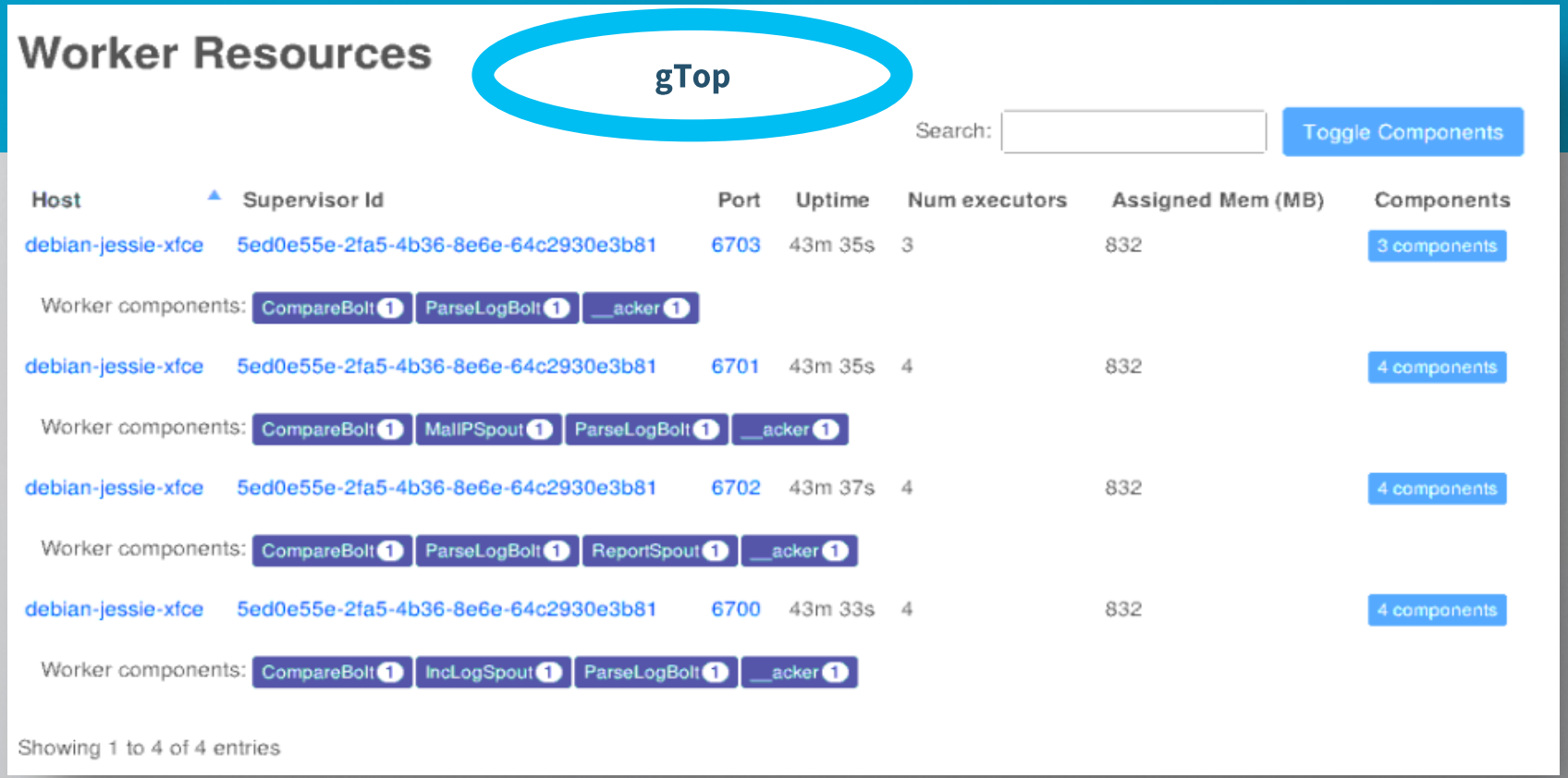
b.

Regular submitter

Storm User Interface…

Parallelism of gTop

4 Worker Processes	Incoming Spout	ParseBolt	CompareBolt
Capacity	-	0.064	0.733
Executors/Threads	1	4	4
Tasks	1	4	4

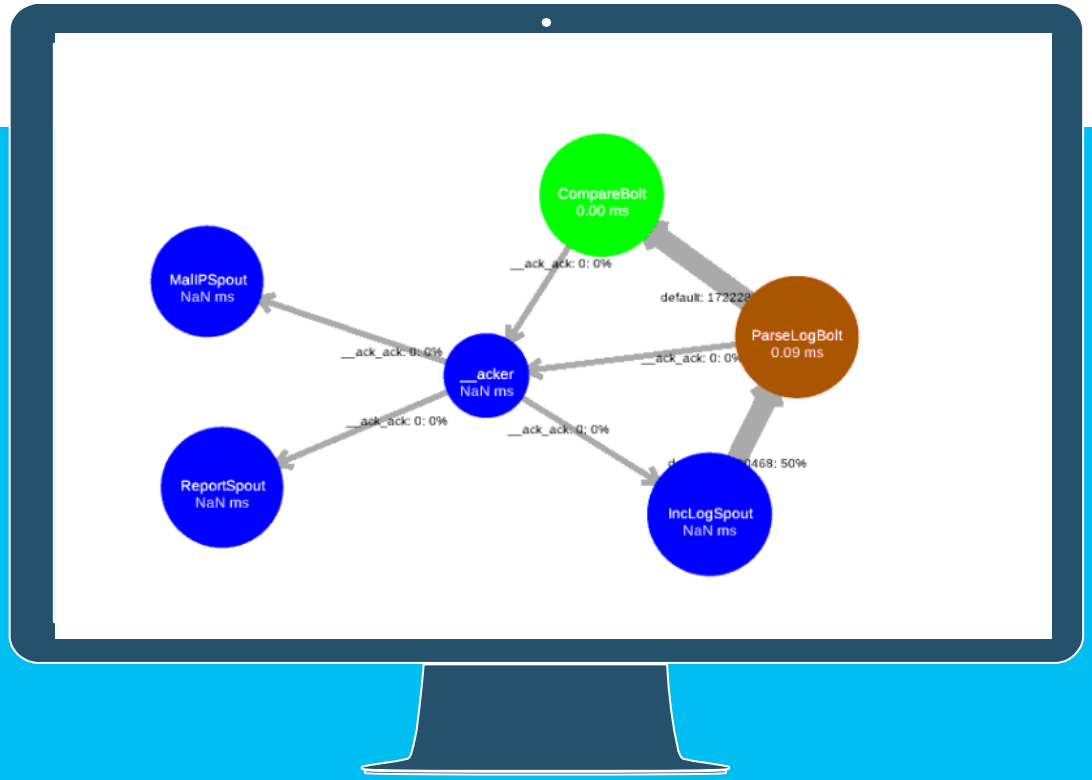

Worker Resources
gTop
Search:
Toggle Components
Host
Supervisor Id
Port
Uptime
Num executors
Assigned Mem (MB)
Components
debian-jessie-xfce
5ed0e55e-2fa5-4b36-8e6e-64c2930e3b81
6703
43m 35s
3
832
3 components
Worker components:
CompareBolt 1
ParseLogBolt 1
__acker 1
debian-jessie-xfce
5ed0e55e-2fa5-4b36-8e6e-64c2930e3b81
6701
43m 35s
4
832
4 components
Worker components:
CompareBolt 1
MailPSpout 1
ParseLogBolt 1
__acker 1
debian-jessie-xfce
5ed0e55e-2fa5-4b36-8e6e-64c2930e3b81
6702
43m 37s
4
832
4 components
Worker components:
CompareBolt 1
ParseLogBolt 1
ReportSpout 1
__acker 1
debian-jessie-xfce
5ed0e55e-2fa5-4b36-8e6e-64c2930e3b81
6700
43m 33s
4
832
4 components
Worker components:
CompareBolt 1
IncLogSpout 1
ParseLogBolt 1
__acker 1
Showing 1 to 4 of 4 entries

gTop

3 Spouts

2 Bolts

1 acker



Optimize

Strip down gTop & increase threads

i) Report Spout NOT NEEDED
ii) Triple Threads for Parse Bolt

Parallelism of hTop

4 Worker Processes	Incoming Spout	ParseBolt	CompareBolt
Capacity	-	0.010	0.437
Executors/Threads	1	4	12
Tasks	1	4	12

Worker Resources

hTop

Search: Toggle Components

Host	Supervisor Id	Port	Uptime	Num executors	Assigned Mem (MB)	Components
debian-jessie-xfce	5ed0e55e-2fa5-4b36-8e6e-64c2930e3b81	6707	5m 18s	5	832	3 components
Worker components:	CompareBolt 1 · ParseLogBolt 3 · __acker 1					
debian-jessie-xfce	5ed0e55e-2fa5-4b36-8e6e-64c2930e3b81	6705	5m 16s	6	832	4 components
Worker components:	CompareBolt 1 · MailPSpout 1 · ParseLogBolt 3 · __acker 1					
debian-jessie-xfce	5ed0e55e-2fa5-4b36-8e6e-64c2930e3b81	6706	5m 15s	5	832	3 components
Worker components:	CompareBolt 1 · ParseLogBolt 3 · __acker 1					
debian-jessie-xfce	5ed0e55e-2fa5-4b36-8e6e-64c2930e3b81	6704	5m 16s	6	832	4 components
Worker components:	CompareBolt 1 · IncLogSpout 1 · ParseLogBolt 3 · __acker 1					

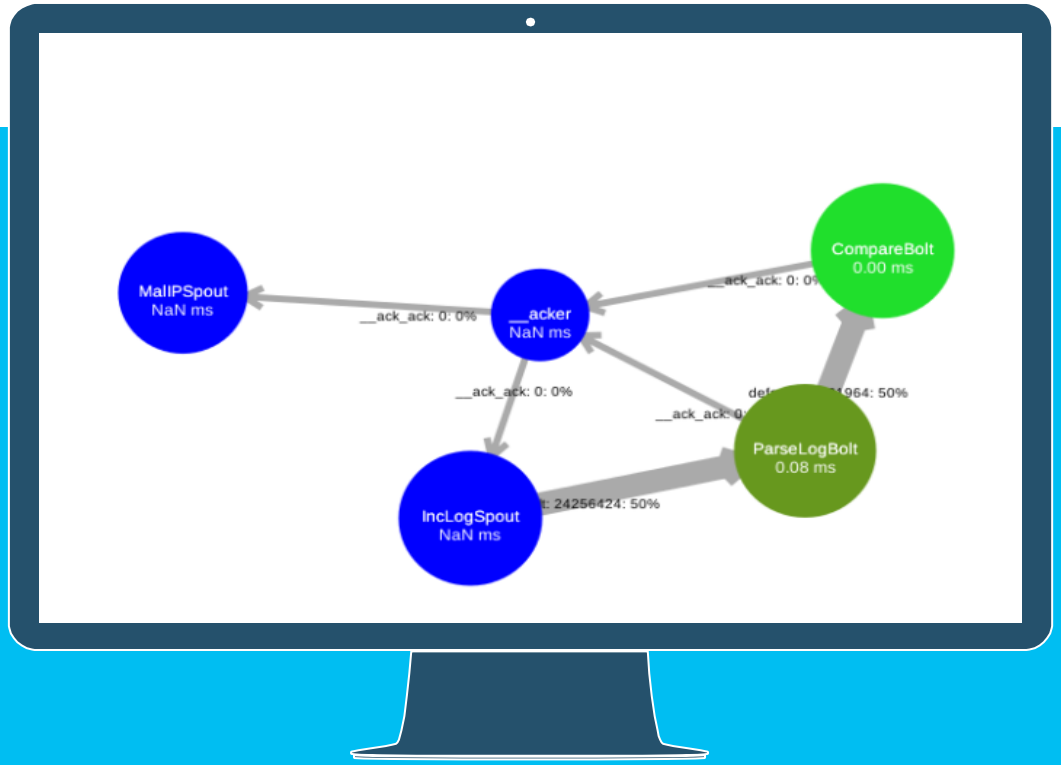
Showing 1 to 4 of 4 entries

hTop

2 Spouts

2 Bolts

1 acker



Results

	Local Submitter	gTop	hTop
Tuples/sec	108.975	33.289	36.137
Avg. RAM (MB)	9196	6646	4639
Avg. CPU cores	18	14	4

Local is Local

Constantly failed after a while

NOT scalable
NOT reliable

Example Applications of my Storm deployment

Blocking Blacklisted IPs – Analysis of an attacker

Distinguish malicious from benign IPs & avoid establishing a TCP connection with potentially malicious users

DNS traffic analysis – DNS poisoning

Distinguish malicious IPs and/or domains from benign IPs and/or domains & avoid connections to potentially malicious resources (C&C Servers)

Distributed port scanning detection

Retain number of distinct SYN packets without a corresponding ACK packet (SYN scanning type) or FIN packets without previous corresponding SYN, ACK packets (FIN scanning type).

P.S.

Personal Challenges

More than ever…

Many steps to take

Let's review some personal challenges

Security Notions

To clarify contemporary security notions is far from easy. Many overlapping definitions and misusings of marketers…

Programming Languages

Many languages appear significant similarities, while basic concepts remail the same. However the details are exhausting.

Resources

Testing your ideas need resources that may exceed your personal belongings. Cloud Servers are needed to underpin any ambitious experiment.

Cryptography

Studying Security is useless, unless cryptography is present. From symmetric keys to PKI and Certificate authorities, knowledge to grasp is enormous

OS

Utilizing Security countermeasures is considered infeasible unless you have a great command of some Linux distros.

Interaction

Remote terminals emerge many difficulties (configurations, networks issues, OS issues). Patience is not enough.

Any questions?

You can find me at:

- **mai16076@uom.edu.gr**