Creating the Ultimate Data Scientist’s Cyber Playground

Building a Multi-Petabyte Analytic Infrastructure for Cyber Defense

Lee Blum
Big Data Architect
Verint Systems
About Me

Lee Blum
@theLeeBlum

+ Big Data Architect, Verint Systems
+ 15 Years in
  - Big Data
  - Cyber Defense
  - IP Networks
  - Storage

Lee.Blum@Verint.com
What We’re Up Against

The Equifax Hack Has the Hallmarks of State-Sponsored Pros

Investigations into the massive breach aren’t complete, but the intruders used techniques that have been linked to nation-state hackers in the past.
Our Use Case

- Government Ministries and Agencies
- Agencies’ Public Web Pages and On-line Services
- Network Management
- IT Services – Network and Endpoints
- Cyber Protection
- Government ISP
Cyber Crash Course
Corporate Network and the Cyber Kill Chain

1. Define Goals & Target
2. Gather Intelligence
3. Weaponize & Infiltrate
4. Call Home
5. Move Laterally
6. Exfiltration & Damage

DMZ:
- WWW
- Mail
- Honeypot
- FTP

Private LAN:
- DBs
- RADIUS
- Client EP
- Directory

Internet
Firewall

Confidential and proprietary information of Verint Systems Inc. © All rights reserved worldwide
Cyber Defense - Sensors

- Define Goals & Target
- Gather Intelligence
- Weaponize & Infiltrate
- Call Home
- Move Laterally
- Exfiltration & Damage

sensor types:
- C&C
- DMZ
- WWW
- AV
- Mail
- Honeypot
- FTP
- Firewall
- Private LAN
- DBs
- RADIUS
- EP
- Directory
- Client EP
- Logs
- Internet
- Firewall
Multiple Organization Protection

- Internet Service Provider
- Cyber Defense System
- Organization 1
- Alerts & Metadata
- Internet
- Organization 2
- ... Organization N
# How Much Data?

## For 100Gbps network and 100k Endpoints

**Structured**
- Alerts
  - C&C: 1 billion, 0.5TB
  - AV: 1 billion, 0.5TB
  - EP: 8 billion, 4TB
  - LM: 1 billion, 0.5TB
  - NetFlows: 4 trillion, 300TB

**Unstructured**
- Logs: 800 billion, 350TB
- Alert Files & Raw: 10 billion, 100TB
- Network raw packets (3 days): 3TB

---

**650k Events p/s | 90 days | 5 Trillion events | 4 PB**
Alert Fatigue

+ Example: Sensor analyzing TCP connections
+ ZERO False Negatives
+ Only one in every 100K False Positives
+ One in every 10 million connections is malicious
+ Analysts process 100 false alerts per each true one!
Multi-Modal Data Fusion

Attack is a process, creating a distinctive trail of evidence.

The defender has a partial view - observing few, not all of the steps.

Alerting over each phase alone will create a flood of false alarms.

A True Event is a Statistical Correlation of Events from Many Noisy Sensors.
Meet Our Personas!

I want to browse alerts and find something interesting in seconds of interactivity.

Mark  
CSOC Operator  
(Tier 1)

Rajesh  
Cyber Analyst  
(Tier 2)

Amy  
Data Scientist  
(Tier 3)

Violet  
DevOps

Linda  
Director
Meet Our Personas!

Mark
CSOC Operator
(Tier 1)

Rajesh
Cyber Analyst
(Tier 2)

Amy
Data Scientist
(Tier 3)

Violet
DevOps

Linda
Director

I want to make complex queries over long retention periods, mostly metadata, in dozens of seconds to minutes.
I want to run batch operations on structured and unstructured data in a notebook environment. Can take hours per job.
Meet Our Personas!

Mark
CSOC Operator
(Tier 1)

Rajesh
Cyber Analyst
(Tier 2)

Amy
Data Scientist
(Tier 3)

Violet
DevOps

Linda
Director

I want to create a **single cluster and infrastructure** to manage all applications and share the same tools such as logs and HM.
Meet Our Personas!

Mark  
CSOC Operator  
(Tier 1)

Rajesh  
Cyber Analyst  
(Tier 2)

Amy  
Data Scientist  
(Tier 3)

Violet  
DevOps

Linda  
Director

Do it all on a single cluster and framework in a minimal footprint. All on-prem and do not have the IT to manage huge clusters.
# High Level Architecture

## Applications

| Enrichments & Correlations |
| Entity Analytics |
| Rule Engine |
| Encrypted Event Analytics |

| Large Scale Repository |
| DNS, IP, Cookie, etc. |
| CEP |
| SparkML |

## Research

### Data Exploration

- Enrichments & Correlations
- Entity Analytics
- Rule Engine
- Encrypted Event Analytics

## Tools

- Logs
- Statistics
- Monitoring

## Devops

- Installation
- Cluster Management
- Patching
- Upgrade
- InfoSec

## Infrastructure

- Spark
- Hadoop
- YARN
- Ambari

## Alerts

- C&C, AV, EP, LM – MD and Files
- Network Forensics – NetFlows, Raw NW
- Logs – Text (Semi-structured)
Enrichments and Correlations

Static/Dynamic Join Enricher

- Events DF (field1..n, <Avro BLOB>)
- Dynamic Enrichment Data

Static Iterative Enricher

- Iterate records on engines
- Merge enrichments to Avro

ID Correlator

- Structured Streaming
- Generate unified events
- Based on state

Events (Avro)

Dynamic Enrichment Data
Entity Analytics

### Entity

<table>
<thead>
<tr>
<th>Host</th>
<th>URL</th>
<th>IP</th>
<th>User</th>
</tr>
</thead>
</table>

### Aggregation

<table>
<thead>
<tr>
<th>Hourly</th>
<th>Morning, Evening</th>
<th>Daily &amp; Day in Week</th>
<th>Monthly &amp; Month in Year</th>
</tr>
</thead>
</table>

### Patterns

<table>
<thead>
<tr>
<th>Daily Peak</th>
<th>Traffic Profile</th>
<th>Destination Geo Location</th>
<th>Accessed EP by time of day</th>
</tr>
</thead>
</table>

### Characteristics

<table>
<thead>
<tr>
<th>Employee Endpoint</th>
<th>News website</th>
<th>C&amp;C Server</th>
<th>SysAdmin</th>
</tr>
</thead>
</table>

### Anomalies

<table>
<thead>
<tr>
<th>Surf out of hours</th>
<th>Irregular Protocol</th>
<th>Irregular Upload/Download</th>
<th>Irregular login to Endpoint</th>
</tr>
</thead>
</table>
Entity Analytics – Timely Aggregation

**Streaming**
15 minute aggregations

- 00:00 - 00:15
- 00:15 - 00:30
- ...

**Batch**
Hierarchical Time Based Aggregations

- 00:00 – 01:00
- 01:00 – 02:00
- ...

- Daily, Day 1 of month
- Day 2
- ...

- Weekly, Sunday
- Monday
- ...

- Monthly, January
- ...

Time
Parquet as our Standard

+ Experience with Data Warehouses – Too expensive, Not enough Analytics
+ Considered alternative open source solutions
+ Favored Parquet due to:
  − Faster ingestion
  − Better scan rate for analytics
  − Better compression (write once, column based)
  − But - Slower random access
    − Solved in our Large Scale Repository
# Large Scale Repository

<table>
<thead>
<tr>
<th>Rapid Ingestion</th>
<th>Interactive Queries</th>
<th>Low Footprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spark Streaming</td>
<td>Interactive Queries</td>
<td>Low Footprint</td>
</tr>
<tr>
<td>Parquet</td>
<td></td>
<td>Parquet</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Long Retention</th>
<th>Batch Analytics</th>
<th>Backward Compatible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parquet</td>
<td>Batch Analytics</td>
<td>Backward Compatible</td>
</tr>
<tr>
<td>Spark SQL</td>
<td>Spark SQL</td>
<td></td>
</tr>
<tr>
<td>Spark ML</td>
<td>Spark SQL</td>
<td></td>
</tr>
</tbody>
</table>

Confidential and proprietary information of Verint Systems Inc. © All rights reserved worldwide
Large Scale Repository

**HTTP_index.parquet**

| Block 1 Index   | Time Range1, IP_BF, Port BF, URL BF, Body BF, ...
| Block 2 Index   | Time Range2, IP_BF, Port BF, URL BF, Body BF, ...
| ...             |
| Block N Index   | Time RangeN, IP_BF, Port BF, URL BF, Body BF, ...

**HTTP.parquet**

<table>
<thead>
<tr>
<th>Block 1</th>
<th>Block 2</th>
<th>Block 3</th>
<th>Block 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>Block N</td>
</tr>
</tbody>
</table>

**SELECT * FROM HTTP.parquet WHERE TimeRange=[1, 2] AND IP=1.2.3.4**
Large Scale Repository - Query – Filter by Time Range

```
SELECT * FROM HTTP.parquet WHERE TimeRange=[1, 2] AND IP=1.2.3.4
```
Large Scale Repository - Query – Filter by Fields Bloom Filter

HTTP_index.parquet

| Block 1 Index | Time Range1, IP_BF1.mightContain(1.2.3.4)=FALSE |
| Block 2 Index | Time Range2, IP_BF2.mightContain(1.2.3.4)=TRUE |

HTTP.parquet

| Block 2 |

SELECT * FROM HTTP.parquet WHERE TimeRange=[1, 2] AND IP=1.2.3.4
Large Scale Repository - Query Remaining Blocks (Partitioned by)

```
SELECT * FROM HTTP.parquet WHERE Block=2 AND TimeRange=[1, 2] AND IP=1.2.3.4
```

**HTTP_index.parquet**

- Block 2 Index
- Time Range2,
  IP_BF2.mightContain(1.2.3.4)=TRUE

**HTTP.parquet**

- 25/5 10:01 1.1.1.1 80 http... AVDDX3
- 25/5 10:11 1.1.1.1 80 http... BFDS4
- 25/5 10:22 2.3.3.3 80 http... NGV4N
- 25/5 10:32 1.2.3.4 80 http... HFBCV2
- 25/5 10:35 5.5.5.5 80 http... NVDF1
Large Scale Repository - Block ID Smart Partitioning on Ingestion

• Block contains 1M records

• Data is partitioned by Day+Hour and Block ID in HDFS.

+ Example – SSL Connection Block ID:
  • BatchTS_ClientIPHash[0-9]_CN[0 = Common, 1-9 Rare]
Large Scale Repository – Results

- 100Gbps traffic
- 30 workers
- Bloom filter returns 100 blocks (<0.5%) 25% are FP
- 17,280 Blocks in Day (1M Records per block)
- Total Query time 5-30 seconds
- Smart partitioning accelerates queries by up to 10x
- Result Set Exploration is immediate
Big Data Pipeline

- Alerts
- Logs
- Metadata
- NetFlows

- Streaming Enrichments, Correlations, Aggregations, Indexing

- Batch Processing (Data Scientists)

- Persistency (Entities, Events)

- Batch Correlations & Aggregations

- Orchestrator

- Spark

- Query (Spark Thrift Server)

- Result Set Exploration

- Parquet

- Inadoop HDFS

- Spark

- Data Scientists

- Result Set Exploration
Research to Production

Connecting the dots
Amy’s Playground

Train Models, Test Detections

SQL / ML

Entities

Alerts

NetFlows

Logs

Raw Network Packets

HTTP / SSL Connections

Amy
Data Scientist
(Tier 3)
The new model is integrated in the system to generate new alerts.

A group of suspicious alerts forms an incident.

Research of stored MD and Raw generates new model for detection on Streaming / Batch job.

Incident analysis may suggest a deeper pattern.
Summary

+ Transformation from Collection to Analytics

+ Democratization of Big Data
  - Rapid R&D
  - Single platform
  - High performance
  - Elasticity

+ Research to Production Cycle
Thank you