Architecting Data Platforms for Cybersecurity

Charaka Goonatilake
CTO @ Panaseer
What will we cover today?

Cyber Security is hard...

...data can provide answers...

...but data & technology is hard...

...so how should you approach building a winning data platform?
Cyber Security is Hard!

2017: The year Ransomware went mainstream

 Oops, your files have been encrypted!

What Happened to My Computer?

Your important files are encrypted.

Many of your documents, photos, videos, databases and other files are no longer accessible because they have been encrypted. Maybe you are busy looking for a way to recover your files, but do not waste your time. Nobody can recover your files without our decryption service.

Can I Recover My Files?

Sure. We guarantee that you can recover all your files safely and easily. (But you have not so enough time.)

You can try to decrypt some of your files for free. Try now by clicking <Decrypt>. If you want to decrypt all your files, you need to pay.

You only have 3 days to submit the payment. After that the price will be doubled. Also, if you don't pay in 7 days, you won't be able to recover your files forever.

How Do I Pay?

Send $300 worth of bitcoin to this address:

If you see this text, then your files are no longer able to have been encrypted. Perhaps you are busy looking for your files, but don’t waste your time. Nobody can recover without our decryption service.

Please follow the instructions:

Ooops, your important files are encrypted.
Cyber Security is Hard!

The Biggest Reported Data Breaches
Data Can Provide Answers

Identify

- Understand business context – identify assets & assess risks

Protect

- Measure & optimise effectiveness of security safeguards

Detect

- Monitor for security anomalies / events / incidents

Respond

- Investigate full impact of security incident and remediate

Proactive Prevention

- Measure & Mitigate Cyber Risk

Reactive Remediation

- Monitor & Remediate Security Incidents

Strategy

Operations
3 Biggest Challenges (aka Architectural Drivers)

Diversity in Stakeholders & Use Cases

Diversity in Data Access Patterns & SLAs

Diversity in Data
<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Questions</th>
<th>Insight Tools &amp; Interfaces</th>
</tr>
</thead>
<tbody>
<tr>
<td>Board</td>
<td>• Is the business secure?</td>
<td>• Reports on Paper / PPT</td>
</tr>
<tr>
<td></td>
<td>• Is our risk status &quot;green&quot;?</td>
<td></td>
</tr>
<tr>
<td>CISO &amp; Security Executives / Management</td>
<td>• Why is our risk out of tolerance in the UK?</td>
<td>• Dashboards on Web Apps</td>
</tr>
<tr>
<td></td>
<td>• Where should we invest our resources to reduce risk / improve security?</td>
<td></td>
</tr>
<tr>
<td>IT / Security Analysts</td>
<td>• Is this machine compromised?</td>
<td>• Search in Free-Text</td>
</tr>
<tr>
<td></td>
<td>• What else has suffered the same compromise?</td>
<td>• Query in SQL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Graph Exploration</td>
</tr>
<tr>
<td>Security Data Scientists</td>
<td>• How can we measure security effectiveness?</td>
<td>• Statistics / Summaries in SQL</td>
</tr>
<tr>
<td></td>
<td>• How can we detect anomalous behaviour on the network?</td>
<td>• ML / Clustering in Python / R</td>
</tr>
</tbody>
</table>

Increasing Complexity in Analysis Workloads

Increasing Tech Skills of Stakeholder
Challenge #2 – Access Patterns & SLAs

**Batch**
- Complex / Historical / Exploratory Analytics
- Modelling

**Interactive**
- Interactive Dashboards
- Query & Search
- Interactive REPL

**Near Real-Time Streaming**
- Continuous Risk Assessment & Threat Detection

**Real-Time**
- Automated Robotic Remediation

**Response Times**
- Hours / Minutes
- 1-10 Seconds
- Seconds
- Sub-Seconds
Challenge #3 – The Data

CYBERscape: The Cybersecurity Landscape

Challenge #3 – The Data

Volume
- Enterprise Scale
- Dozens of Data Sources
- 100s of TBs

Variety
- Formats – (Un-/Semi-)Structured Sources & Protocols
  - APIs / DBs
  - Pull & Push

Velocity
- Batch
- Streaming Real-Time

Common Characteristics of Business, IT & Security Data
Solution Space is Vast
Focus on Optimising for Agility

- Accessibility
  - Democratised / Self-Service Access

- Flexibility
  - Adaptable to Change

- Iterative & Incremental Delivery
  - User-centric & Use Case Driven

**agile**

1. able to move quickly and easily.
High-Level Data Platform Architecture

Data Platform

Data Sources
- DBs
- APIs
- Streams

Acquire Data

Data Pipelines

Data Stores

Interfaces
- Board / CxO
- Analyst
- Data Scientist

Stakeholders

Data Sources

Confidential

WWW.PANASEER.COM

2018 Panaseer Limited. Registered in England and Wales with the company registration number 09098199.
Data Flow for Cybersecurity Insights

1. Collect
   - Data & Protocol Integration
   - Push / Pull
   - Parse

2. Normalise
   - Consolidate
   - Clean
   - Standardise
   - Enrich
   - Data Modelling

3. Resolve
   - Resolve real-world entities (e.g. devices)

4. Analyse
   - Correlate
   - Summarise
   - Derive Insights

5. Persist
   - Storage based on Data, Access Patterns & SLAs

6. Present
   - Interfaces & tools to consume insights
Collect

- Acquire Data
- Data Pipelines
- Data Stores
- Interfaces

- Collect
- Normalise
- Resolve
- Analyse
- Persist
- Present

- Data & Protocol Integration
- Push / Pull
- Parse
Misadventures with Flume
NiFi – Enabling Rapid Data On-Boarding

From Flume to NiFi

From code to GUI-driven config
Data Pipelines

Acquire Data

Data Pipelines

Data Stores

Interfaces

Collect

• Consolidate
• Clean
• Standardise
• Enrich
• Data Modelling

Normalise

• Resolve real-world entities (e.g. devices)

Resolve

• Correlate
• Summarise

• Derive Insights

Persist

• Storage based on Access Patterns

Present

• Interfaces & tools to consume insights
Spark: A Game Changer

Unified Workloads
Batch / Streaming / Interactive

Democratised Access
DataFrames with SQL / Python / R / Java / Scala

Performance
Enables Iterative / Incremental
Accelerates Batch

Rich Libraries
SQL Functions / MLlib / GraphX
Configurable ETL Pipelines

- Timestamp Standardisation
- Enrichment
- Mapping to Standard Data Model
- Clean & Standardise
Entity Resolution

Data Source-Centric View → Entity-Centric View

Entity Resolution Engine

Applications
Devices
Vulnerabilities
Users
Threats
Store / Persist

- Acquire Data
- Data Pipelines
- Data Stores
- Interfaces

Collect • Normalise • Resolve • Analyse • Persist • Present

- Storage based on Data, Access Patterns & SLAs
Polyglot Persistence

Sadly, no such thing as...

One Database to Rule Them All...
Storage – Compromises

- **Speed at Scale**: Fast but constrained analysis on high volumes
- **Scale with Flexibility**: Any workload on any data but slow
- **Speed with Flexibility**: Fast & complex analysis on data subsets
- **Flexibility**
Storage Choices – Horses for Courses

Relational

Graph

Key-Value

Document

Columnar

MySQL

PostgreSQL

neo4j

JanusGraph

redis

MEMCACHED

mongoDB

Solr
Present / Consume

Acquire Data

Data Pipelines

Data Stores

Interfaces

Collect -> Normalise -> Resolve -> Analyse -> Persist -> Present

• Interfaces & tools to consume insights
<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Insight Tools &amp; Interfaces</th>
</tr>
</thead>
<tbody>
<tr>
<td>Board</td>
<td>• Reports on Paper / PPT</td>
</tr>
<tr>
<td>CISO &amp; Security Executives / Management</td>
<td>• Dashboards on Web Apps</td>
</tr>
<tr>
<td>IT / Security Analysts</td>
<td>• Search in Free-Text</td>
</tr>
<tr>
<td></td>
<td>• Query in SQL</td>
</tr>
<tr>
<td></td>
<td>• Graph Exploration</td>
</tr>
<tr>
<td>Security Data Scientists</td>
<td>• Statistics / Summaries in SQL</td>
</tr>
<tr>
<td></td>
<td>• ML / Clustering in Python / R</td>
</tr>
</tbody>
</table>
Putting it All Together – A Panaseer Example

DATA SOURCES

INGEST
- Data Collector
- Streaming Data Queue
- nifi
- kafka

PROCESS
- SQL
- Scala
- python
- Java
- Spark
- YARN: Distributed Resource Manager & Job Scheduler

STORE
- Application Data Stores
  - Random, Real-Time Read/Write Store
  - Analytic Data Store
  - Apache Phoenix
  - Apache HBase
  - Hive
  - Parquet

PRESENT
- REST over HTTPS
- Web App Server
- Single-Page Application In Web Browser

EXTENSIBLE INTERFACES
- HUE
- Tableau
- Jupyter
- Zeppelin

APIs
- Distributed Execution Engines
- REST over HTTPS

Base (Underlayer) File System

Collect

Putting it All Together – A Panaseer Example
Key Takeaways

Data Platform

Data Sources
- DBs
- APIs
- Streams

Acquire Data

Data Pipelines

Data Stores

Interfaces

Stakeholders
- Board / CxO
- Analyst
- Data Scientist
Thank You!

Get in touch:
charaka.goonatilake@panaseer.com
linkedin.com/in/charakag/
@charakag