Running AI Workloads in Containers
Problem Statement & Landscape
Prior to Containers

Research
5% of time spent researching new attack vectors, exploit patterns and 0 day threats

AI and Rule Authoring
15% of time spent on machine learning, rule authoring and A/B testing

Performance Tuning
20% time spent tuning AI, rules and platform throughput

Platform Maintenance
60% time being spent maintaining and monitoring the platform
Key Challenges

**Complexity**
The complexity of mining for suspicious activity requires specialization beyond data engineering, forcing engineering to work with data scientists.

**Scale**
The volume and velocity of the data is enormous and requires specialized and elastic distributed frameworks.

**Durability**
The platform, rules and AI must have the ability to recover from, or prevent failure.

**Agility**
Lots of new malware and hackers across the globe mean that we need to be able to move quickly and adapt to the changing environment.

**Extensibility**
The goal is to build a platform that can be extended to uses we never even dreamed of.
Big Data Scale

**500 billion**
Daily records processed
Each day 500 billion to 1 trillion records are processed in stream and batch cleaning, enrichment and aggregate transformations

**5.5 trillion**
Records persisted in data store
More than 5.5 trillion records are persisted for machine learning, product analytics and data science studies

**100+ petabytes**
Data size
More than 100 petabytes (uncompressed) is stored at any given point in the data store
Personas

- **Detection Rules**
  - Research Science

- **Whitelist Rules**
  - Research Science

- **Process Graph Building**
  - Research Science

---

**EDR Platform**

- Data Engineering
Business Workflow
Workflow for Containerized Workloads

Researcher Workflow
10/100 Rules per day
AI & Rules

Telemetry Data Flow
40+ TB per day
Data Store

Continuous Data Mining
Detections

Data

AI & Rules

bmc Control-M
API
Workflow for Containerized Workloads

Researcher Workflow
10/100 Rules per day

Telemetry Data Flow
40+ TB per day

Data Store

AI & Rules

Detections
The research scientists follow a specific pipeline for developing, testing and promoting to production.
Silent Mode: A/B Testing
Containers and Scheduling
Inside a Container
Anomaly Detection: Clustering

PCA - 6 Clusters

PCA = 4 Clusters
Anomaly Detection: Variance Distribution

- **Normal variance**
  - Low variance between input and output.
  - Normal data points are tightly clustered.

- **Perfect match**
  - Even lower variance, indicating a perfect match between input and output.
  - Data points are even more tightly clustered.

- **Anomalies**
  - High variance between input and output.
  - Data points are spread out, indicating unusual behavior.

---

Malwarebytes

![Image of container with Python logos](image-url)
Container Lifecycle

- Step 1 – Build Docker Image
- Step 2 – Run Docker container
- Step 3 – Execute Workload
- Step 4 – Stop container
- Step 5 – Cleanup Zombie container
Container Scheduling: Control-M

- Host Groups
- Associated Hosts
- Host Settings
Container Scheduling: Control-M
Performance Statistics
# Execution Scheduling

<table>
<thead>
<tr>
<th>Latency observations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rule Push</strong></td>
</tr>
<tr>
<td>Researchers committing rules to deploying Docker containers</td>
</tr>
<tr>
<td>10-20 secs</td>
</tr>
</tbody>
</table>

* By design

** Application logic
Agent Container Statistics

Average for past 2 weeks with 1 minute periods

0 Failures!

MTTR (Mean time to recovery): 2-3 seconds
Final Thoughts
Control-M

High Availability
Using host groups and an HA configuration we are able to achieve high availability

Global Visibility
Control-M allows us to monitor and analyze at a bird’s eye view of the entire ecosystem

Enterprise Class
Control-M is an enterprise class tool with features built to future proof our platform
Successful Detection Platform

**Research**
Provide a data platform that harvests data and provides the data scientist a means for historical research

**Improve**
The platform is continuously improving as new containers are deployed with better detection engines in each new release

**Develop**
The platform makes it easy for the research team to craft any rule or AI algorithm inside the container and insure correctness in production

**Test**
The platform leverages containers and a manifest to allow certain containers to execute in silent mode for A/B testing
Thank you.