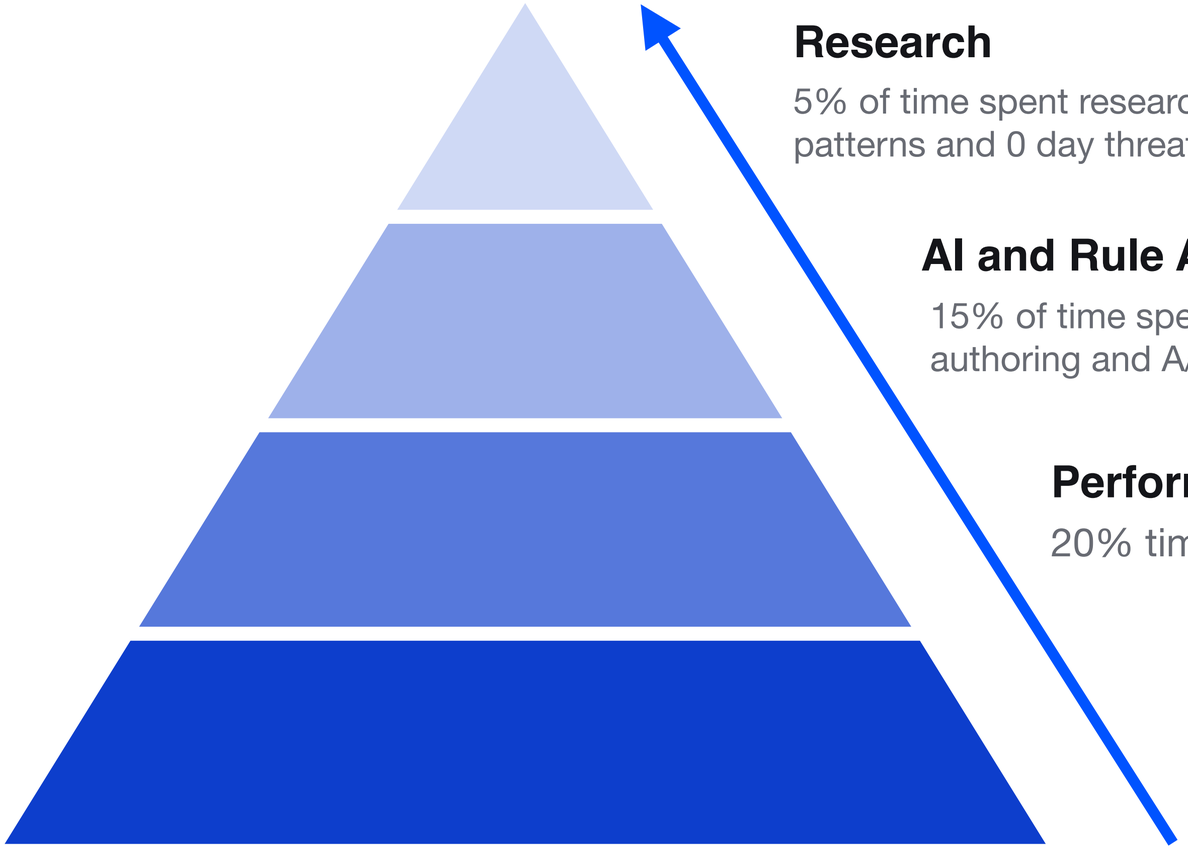
Malwarebytes
Running AI Workloads in Containers
Artwork is copyrighted property of Malwarebytes Inc.
bmc
Control-M

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

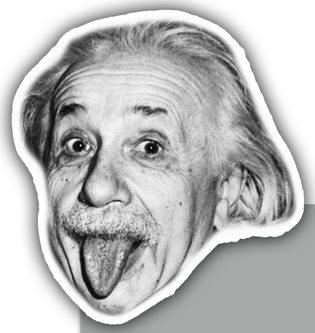
More than 5.5 trillion records are persisted for machine learning, product analytics and data science studies

100+ petabytes

Data size

More than 100 petabytes (uncompressed) is stored at any given point in the data store

Personas



Detection Rules
Research Science

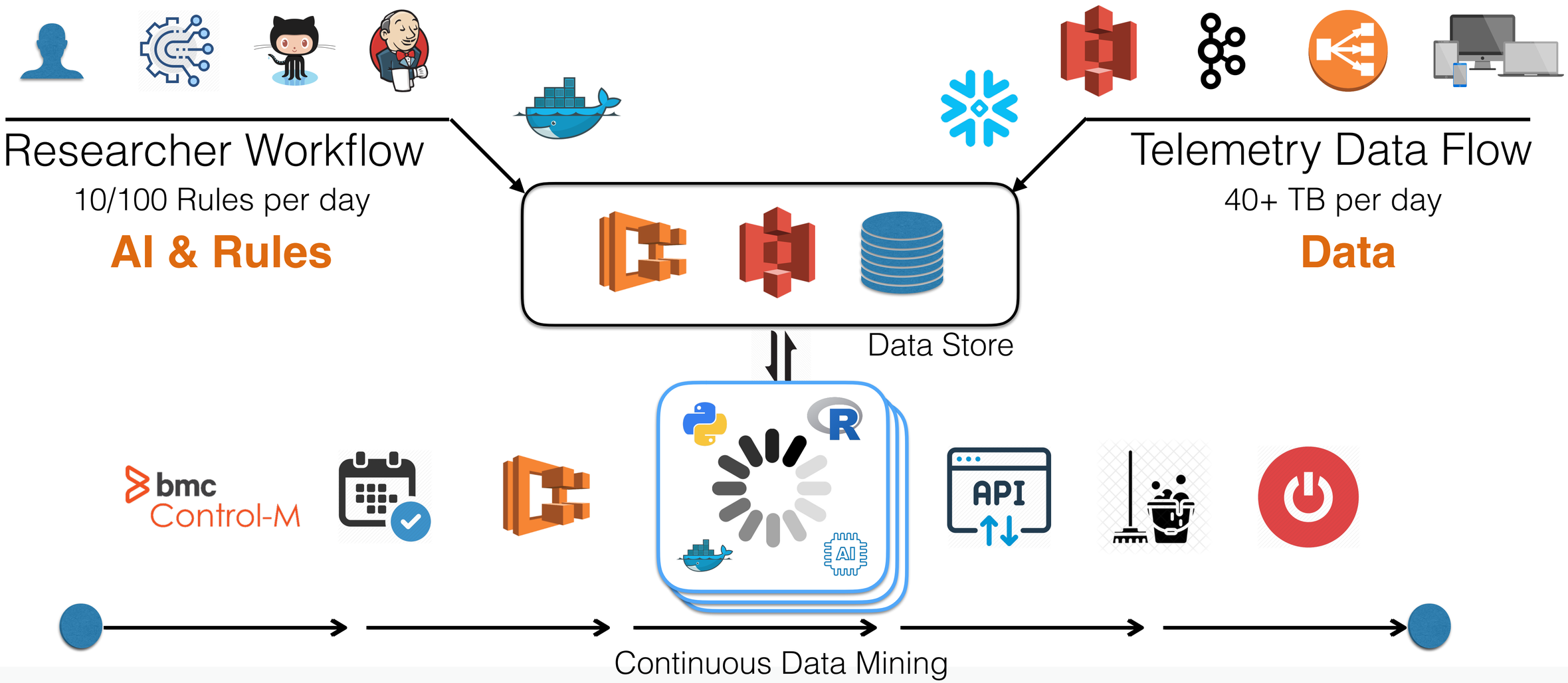
Whitelist Rules
Research Science

Process Graph Building
Research Science

EDR Platform
Data Engineering

Business Workflow

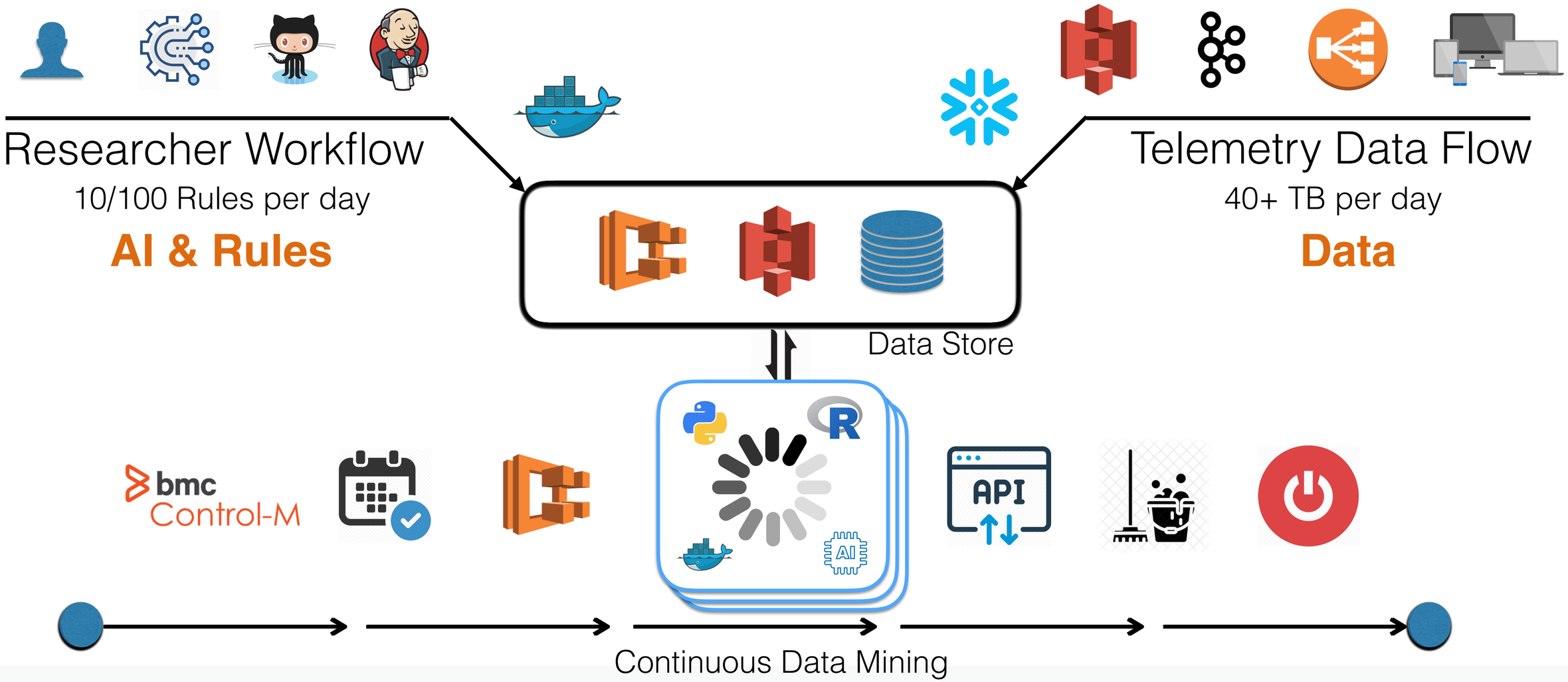
Workflow for Containerized Workloads



Detections

Malwarebytes

Workflow for Containerized Workloads



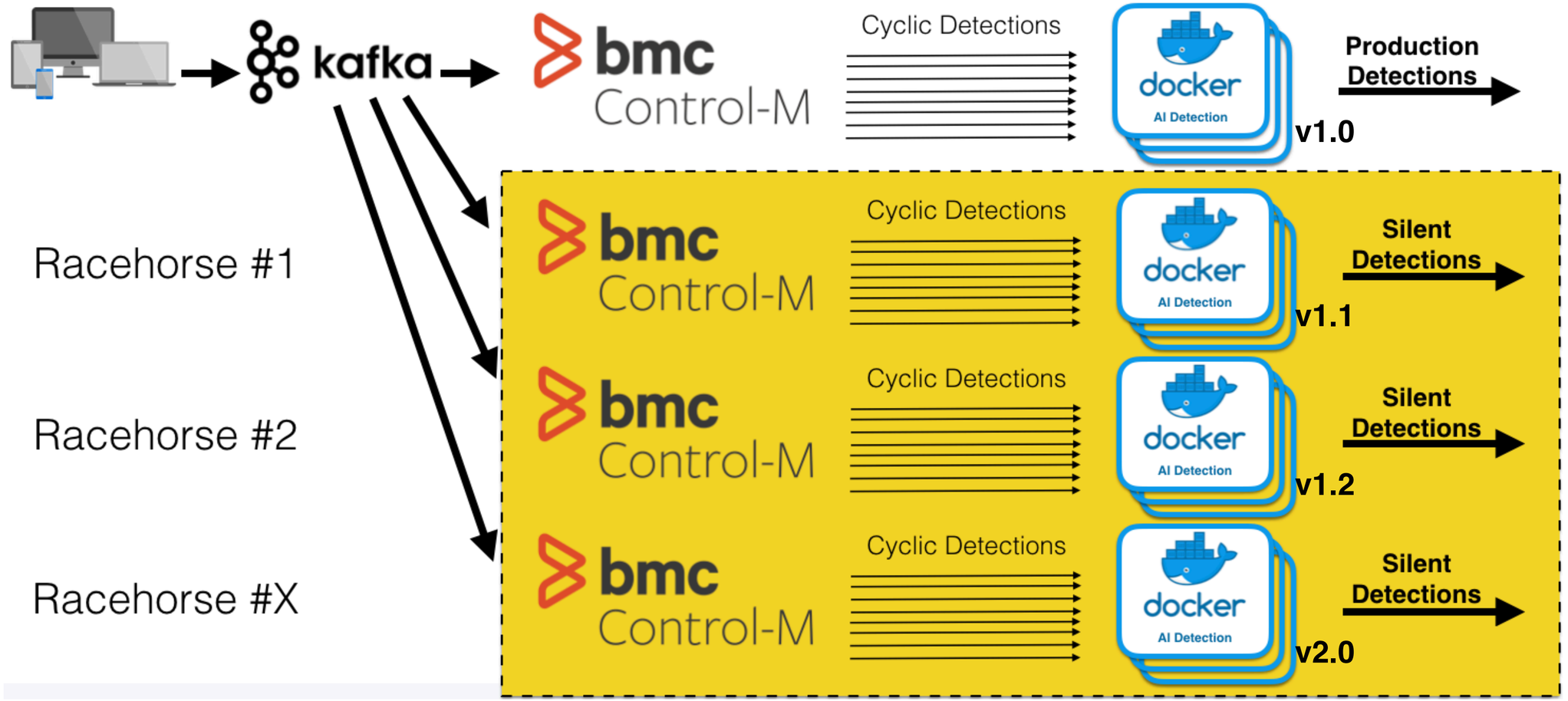
Detections

Malwarebytes

Research Team Workflow

The research scientists follow a specific pipeline for developing, testing and promoting to production

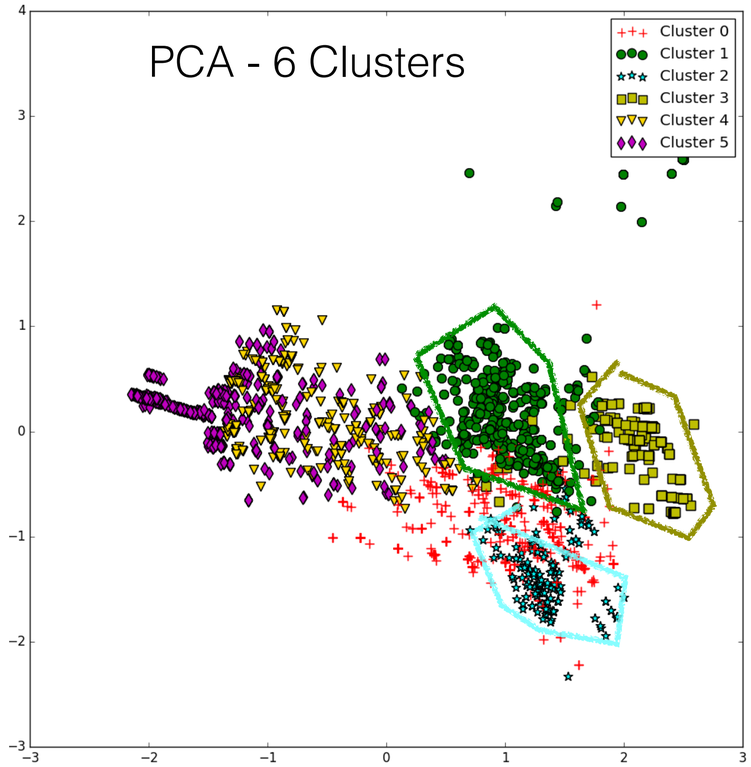
Silent Mode: A/B Testing

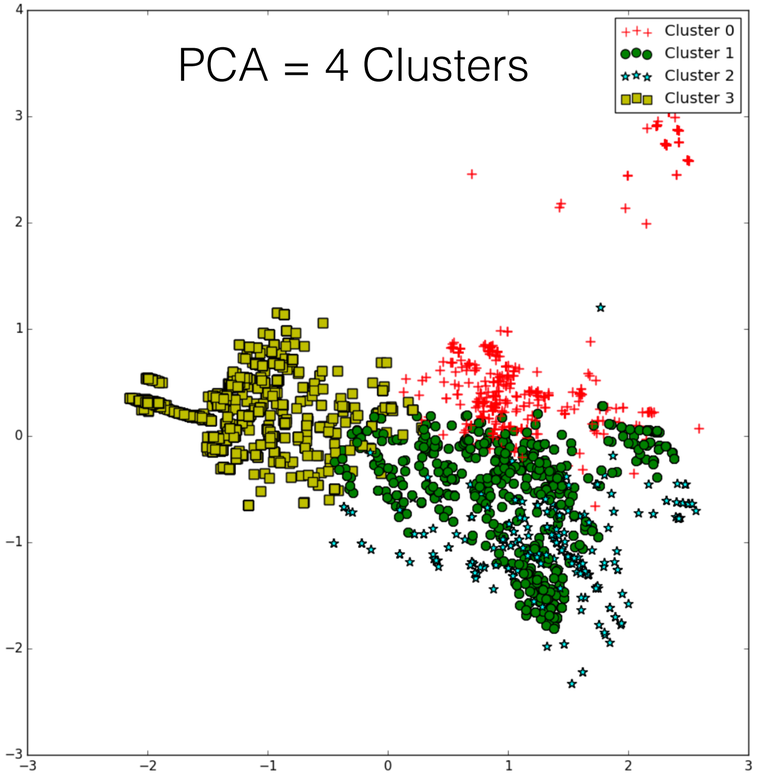


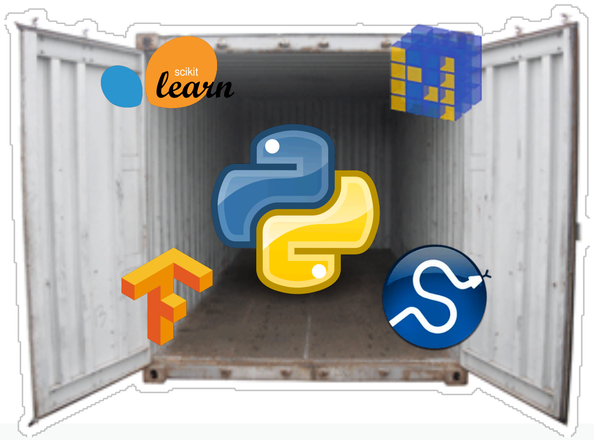
Containers and Scheduling

Inside a Container

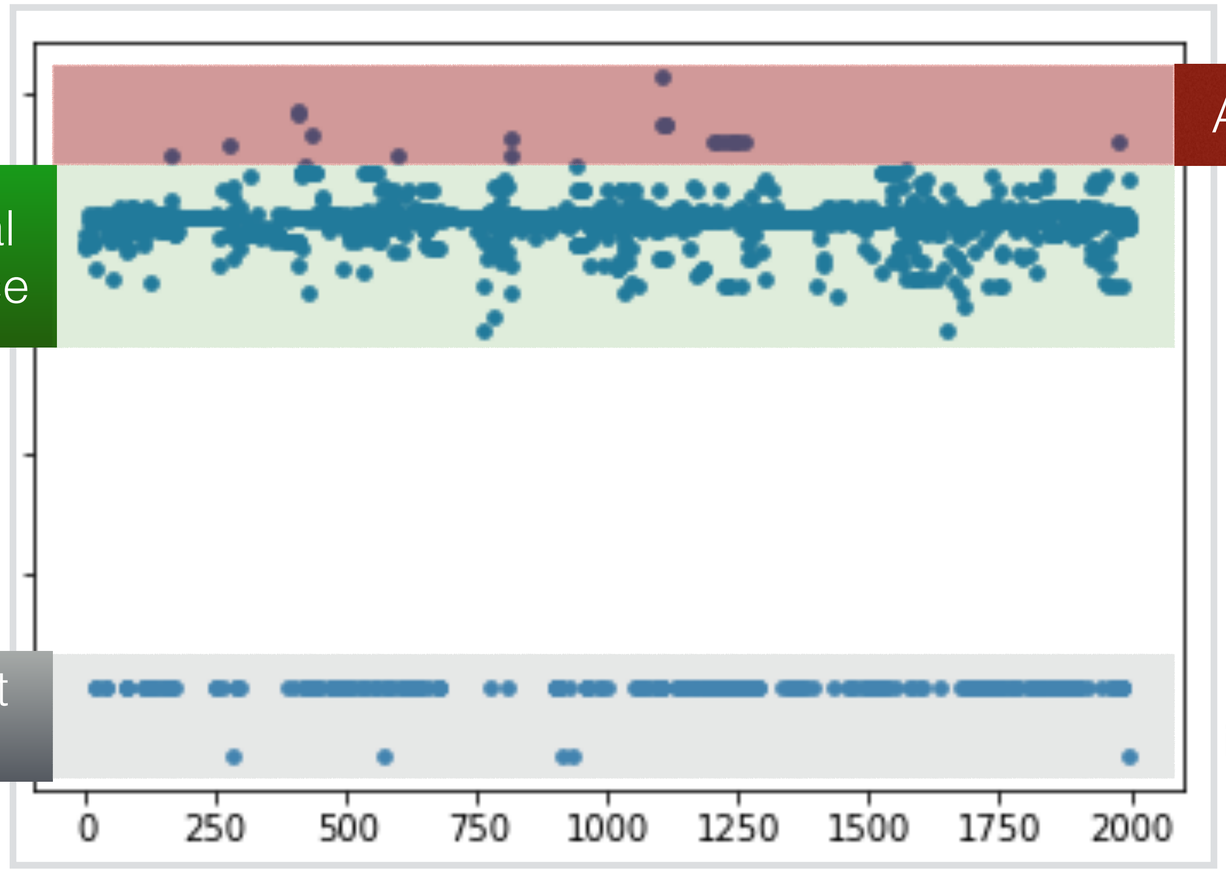
Anomaly Detection: Clustering





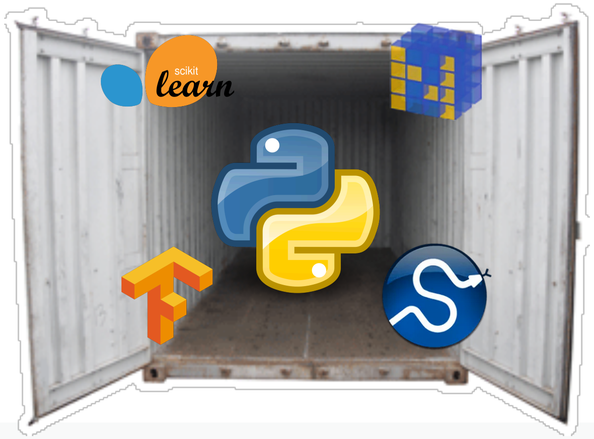


Anomaly Detection: Variance Distribution



High variance between input and output.

Low variance between input and output.



Malwarebytes

Container Lifecycle

Control-M Server
HA configuration

Control-M Agent Host Groups

Step 1 – Build Docker Image

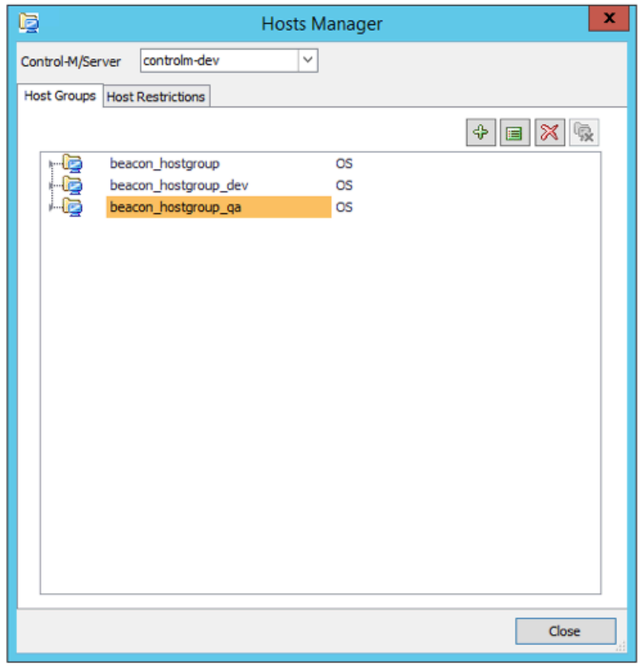
Step 2 – Run Docker container

Step 3 – Execute Workload

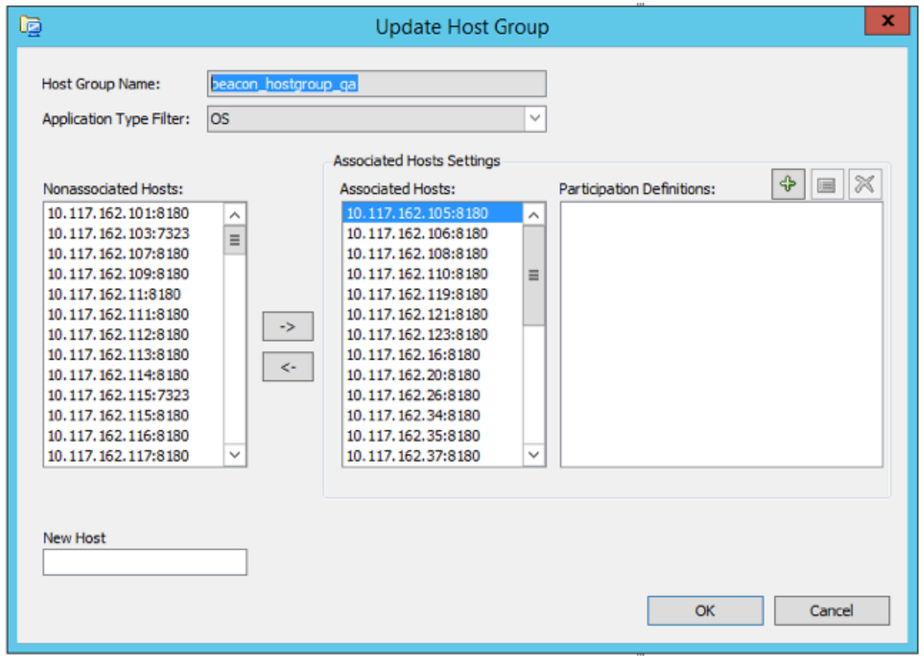
Step 4 – Stop container

Step 5 – Cleanup Zombie container

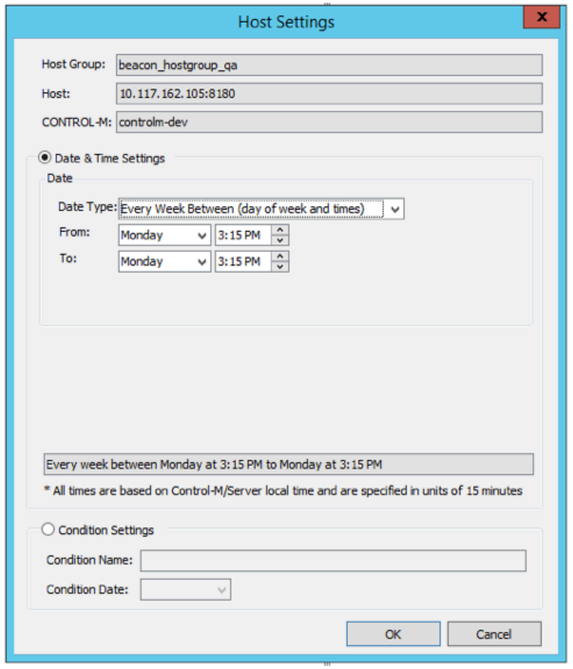
Container Scheduling: Control-M



Host Groups



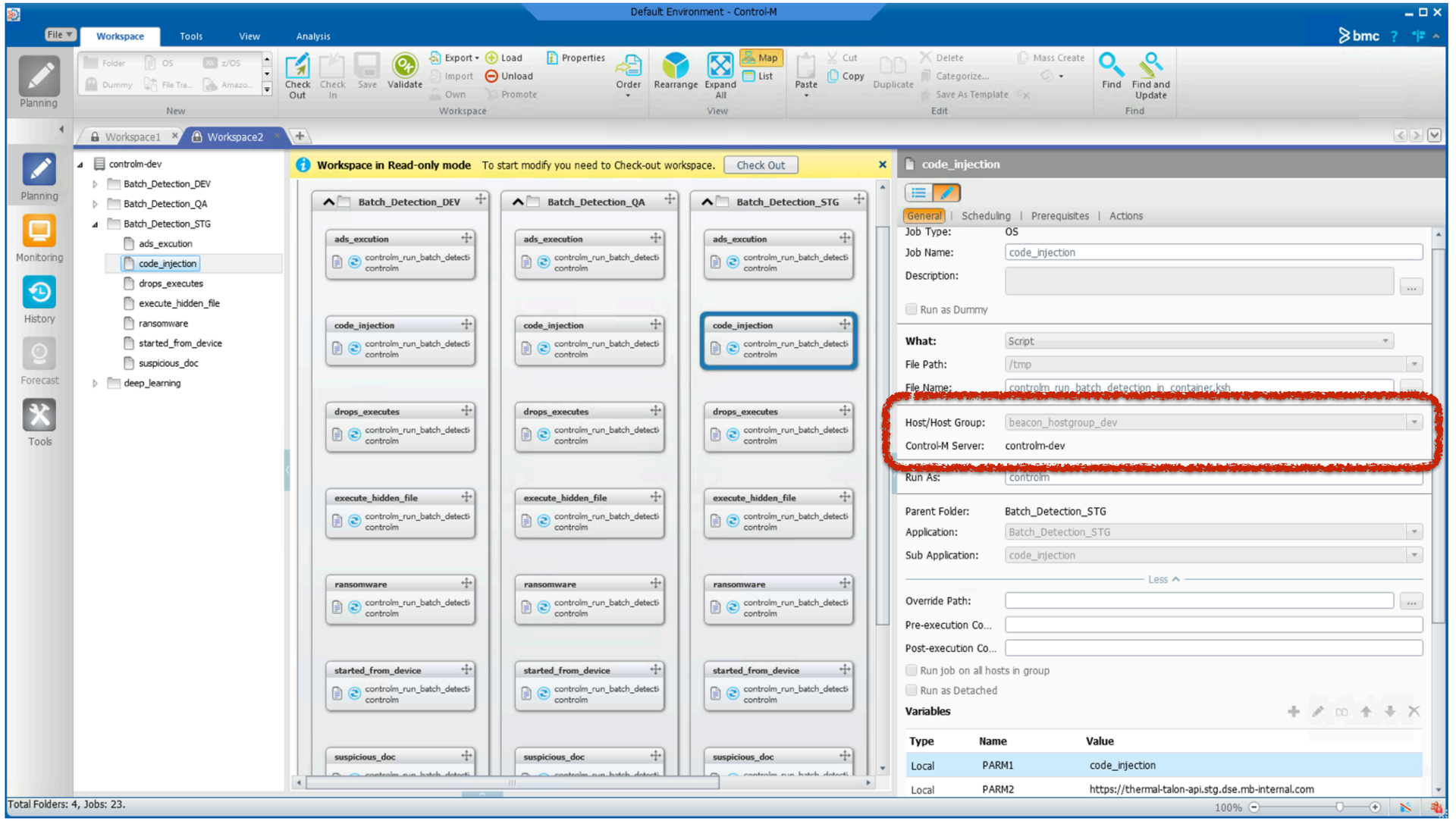
Associated Hosts



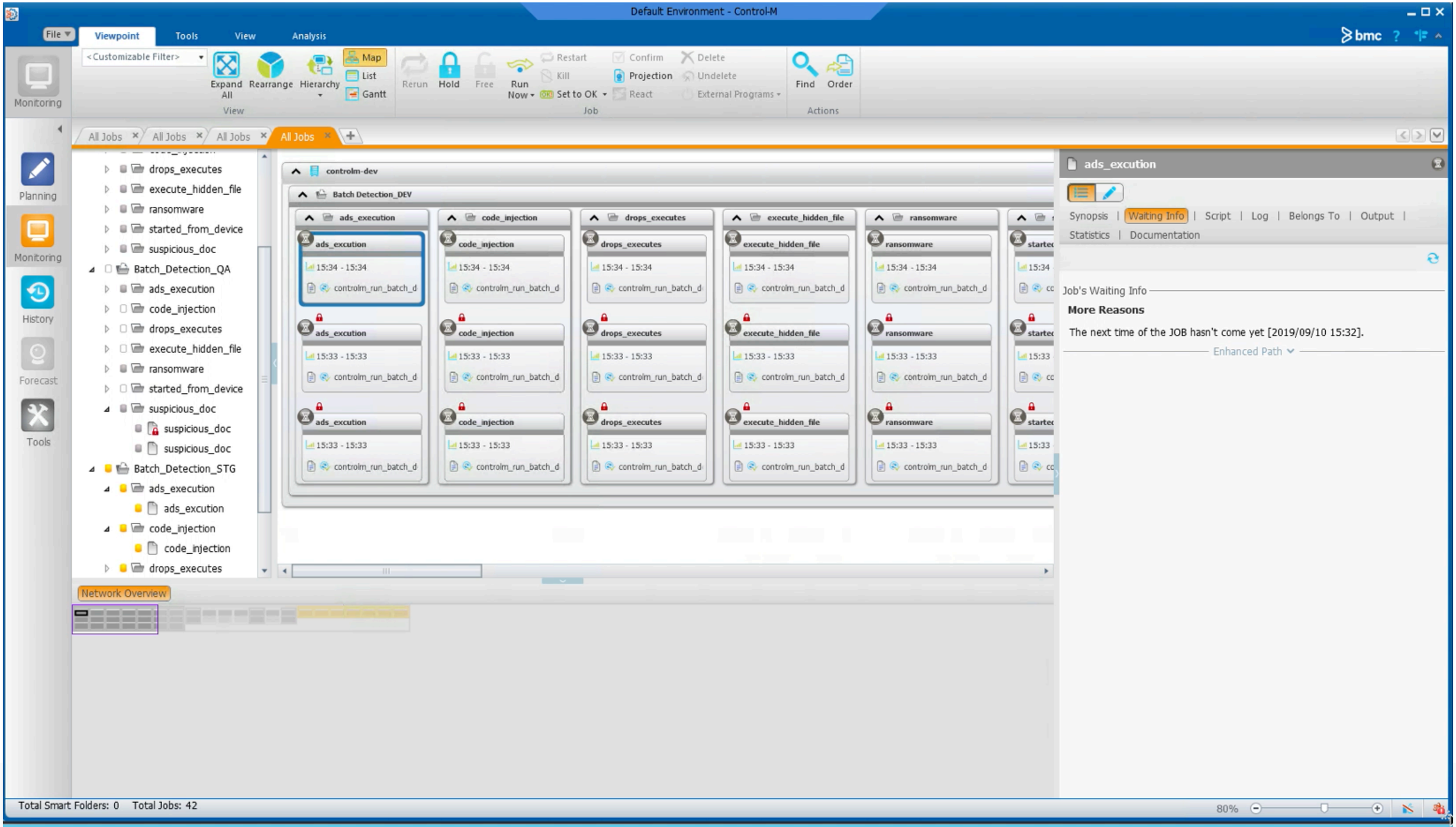
Host Settings

Malwarebytes

Container Scheduling: Control-M



Malwarebytes


Default Environment - Control-M
File
Viewpoint
Tools
View
Analysis
bmc
Monitoring
<Customizable Filter>
Expand All
Rearrange
Hierarchy
Map
List
Gantt
View
Rerun
Hold
Free
Run Now
Restart
Kill
Set to OK
Confirm
Projection
React
Delete
Undelete
External Programs
Job
Find
Order
Actions
Planning
Monitoring
History
Forecast
Tools
All Jobs
All Jobs
All Jobs
All Jobs
drops_executes
execute_hidden_file
ransomware
started_from_device
suspicious_doc
Batch_Detection_QA
ads_execution
code_injection
drops_executes
execute_hidden_file
ransomware
started_from_device
suspicious_doc
suspicious_doc
suspicious_doc
Batch_Detection_STG
ads_execution
ads_excution
code_injection
code_injection
drops_executes
controlm-dev
Batch Detection_DEV
ads_execution
code_injection
drops_executes
execute_hidden_file
ransomware
ads_excution
15:34 - 15:34
controlm_run_batch_d
15:33 - 15:33
Network Overview
ads_excution
Synopsis
Waiting Info
Script
Log
Belongs To
Output
Statistics
Documentation
Job's Waiting Info
More Reasons
The next time of the JOB hasn't come yet [2019/09/10 15:32].
Enhanced Path
Total Smart Folders: 0
Total Jobs: 42
80%
3:33 PM
9/10/2019

Performance Statistics

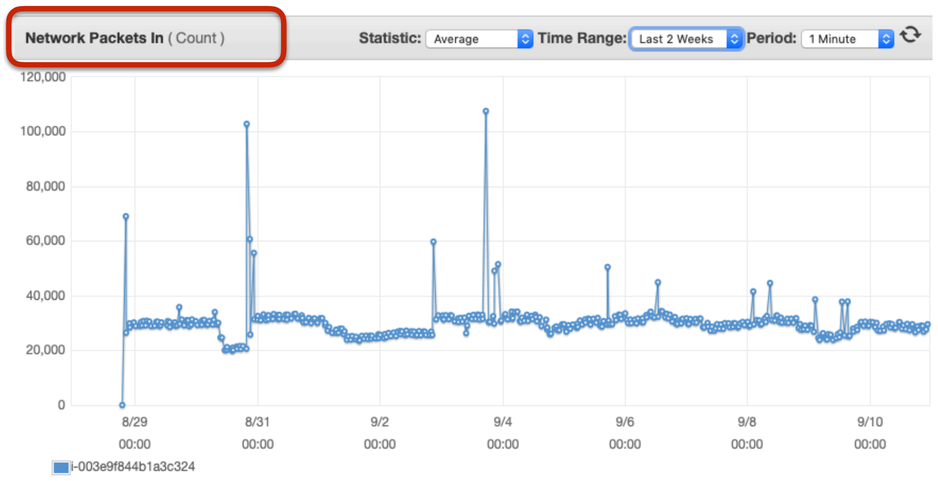
Execution Scheduling

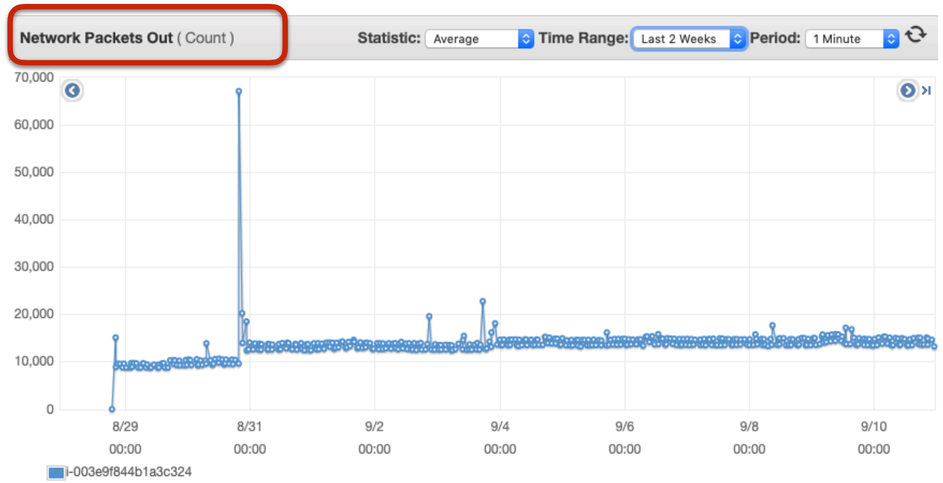
Latency observations						
Rule Push	**Data availability**	**Control-M signal**	**Docker Run**	**Return Control**	**Using Resultset**	**Termination**
Researchers committing rules to deploying Docker containers	From Endpoint to Data Store via Kafka Cluster	Kicking off workloads in containers from Control-M Server	Executing workloads in Containers	Returning control post execution to calling env	Calling subsequent processes with computed result in container	Validation & Terminating container post execution
10-20 secs	**30 secs***	**300-400 ms**	**2-3 secs***	**200-300 ms**	**200-300 ms**	**300-500 ms**

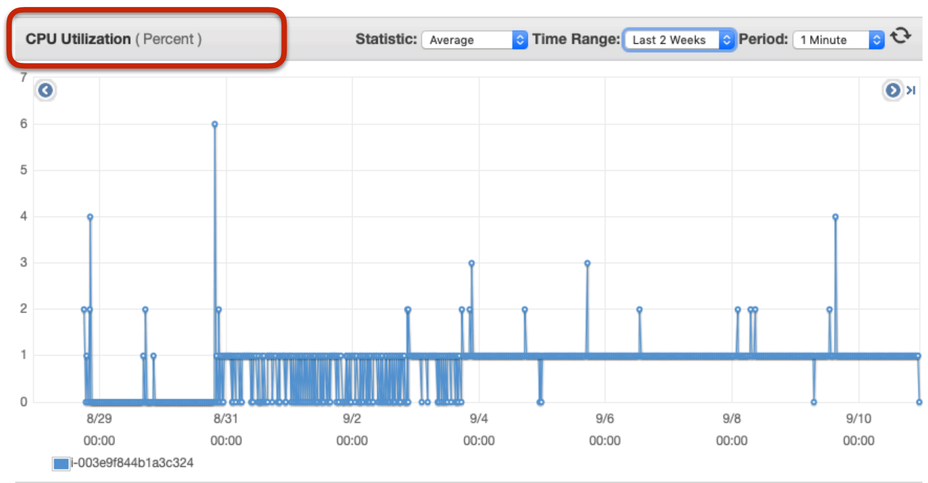
* By design
** Application logic

Agent Container Statistics

Average for past **2 weeks** with **1 minute** periods



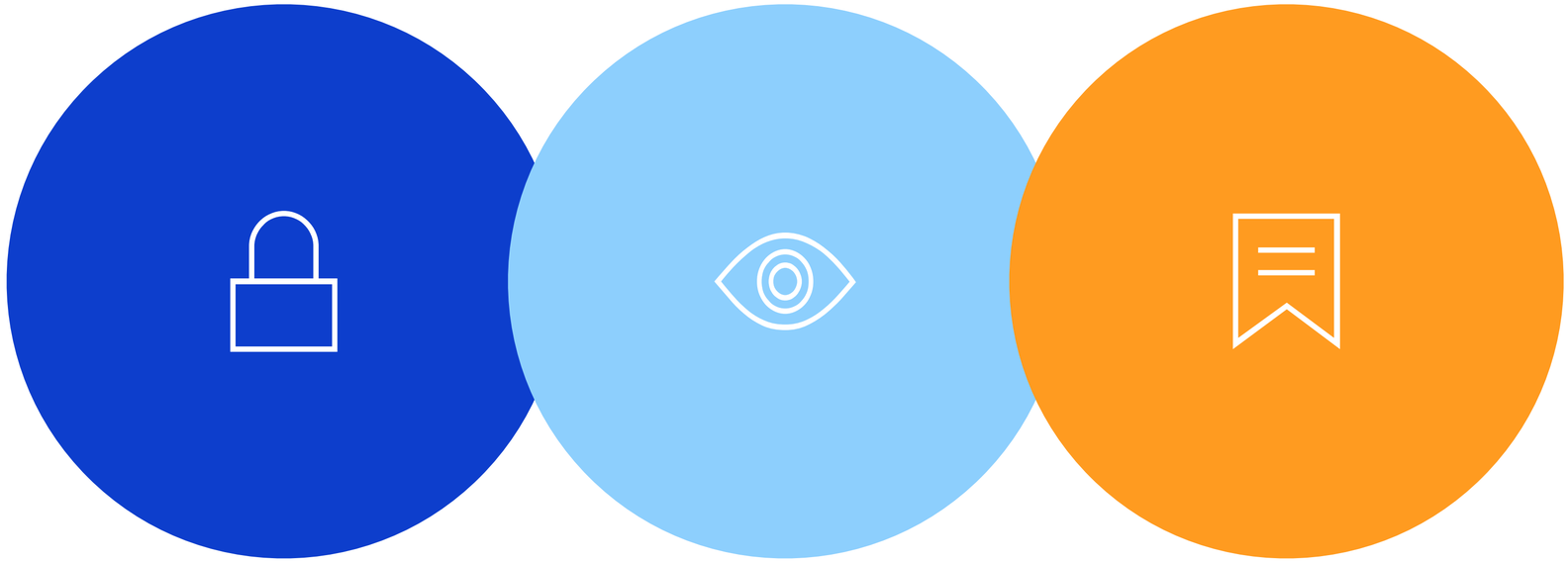




Malwarebytes

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

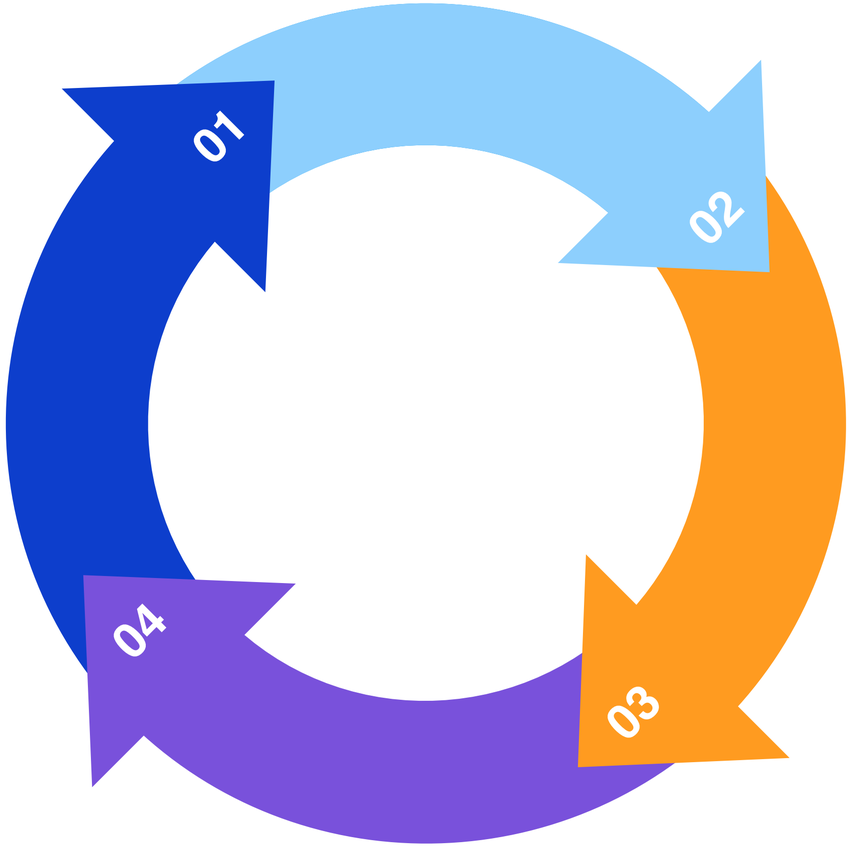
Enterprise Class

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



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

Improve

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

Test

The platform leverages containers and a manifest to allow certain containers to execute in silent mode for A/B testing

Thank you.