Data analytics approaches

to assist Network & Security Operations

Usen Tulemisov, PhD

Sr. Systems Engineer, Global Service Providers

Sep 2016

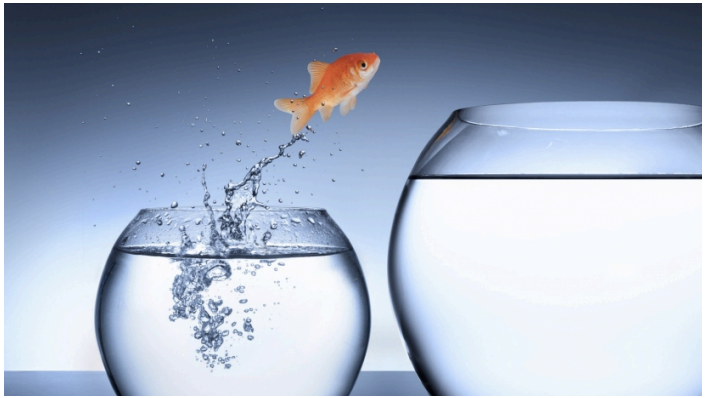
“What’s In It For Me?"

This session will help understand how Data Analytics is relevant to Service Provider Operations

Agenda	Out of scope
Why Data Analytics?: 1. Burning platform 2. Opportunities 3. Industry developments	Detailed market research data
High Level Architecture and key elements of the system (How?)	Review of all components
Key use cases (So what?) 1. Application Dependency Mapping 2. Policy Compliance and Simulation 3. Flow Search and Forensics	All possible use cases, such as: Accounting, Security anomaly detection, Proactive technical assistance and others

Why Data Analytics?

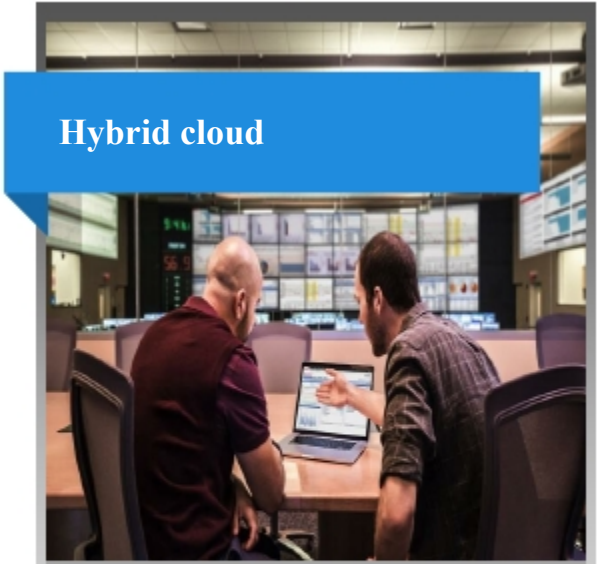
- ‘Burning platform’
  - Increasingly complex Operations & Decreasing Profitability
  - Scarce resources (Time/Talent/Lab)
  - OTT players have first mover advantage
  - NFV needs different OSS
- Economic Logic and Opportunities
  - Low cost of Data Storage (such as CEPH)
  - Data Virtualisation & OSS DB Consolidation
  - Telemetry
  - Application & Customer Centricity



Modern data centers are getting increasingly complex

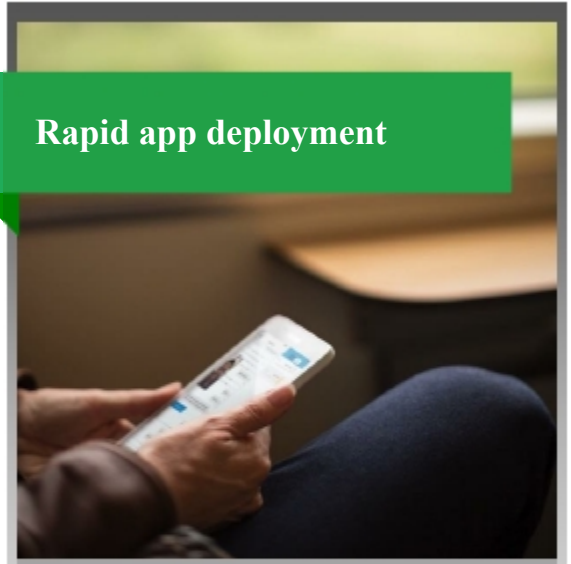
Big and fast data

- Increase in east-west traffic
- Expanded attack surface
- Open source



Hybrid cloud

- Zero trust model
- Multi cloud orchestration
- Application portability



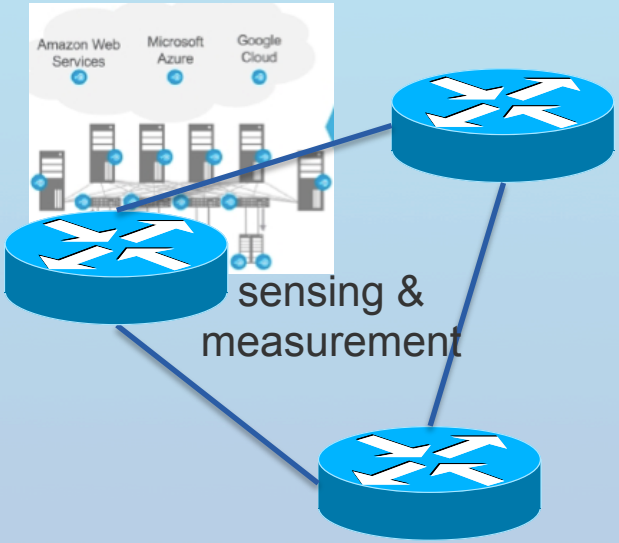
Rapid app deployment

- Continuous development
- Application mobility
- Micro services

Traditional Monitoring Is Showing Its Age

Not suited for Modern Network and Security Operations

Where Data Is Created



Incomplete

Scale Issues

SNMP
(invented in 1987)

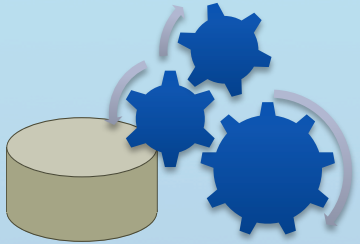
Syslog
(invented in early 1980s)

CLI

Subject to Change

Unstructured

Where Data Is Useful



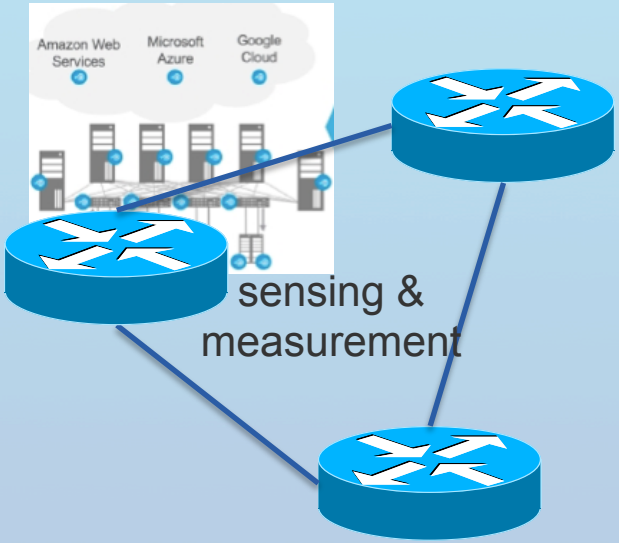
storage & analysis

Strong burden on back-end

Normalize different encodings, transports, data models, timestamps

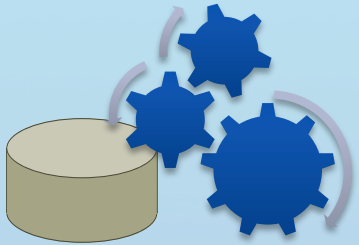
Streaming Telemetry is a game changer
Monitoring becomes a big data problem

Where Data Is Created



As Much Data
As Fast
As Useful
As Easy
As Possible

Where Data Is Useful



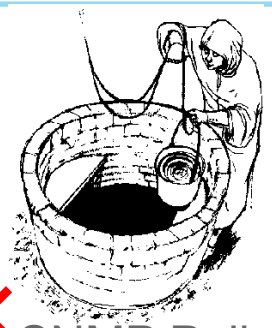
storage & analysis

Volume – Scale of Data
Velocity – Analysis of Streaming Data
Variety – Different Forms of Data

Big Data and Machine Learning Problem

Telemetry: Key Principles

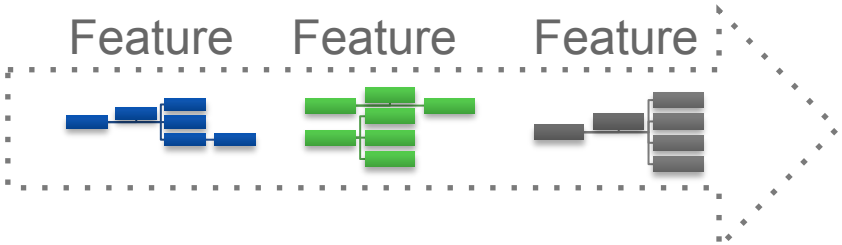
Push Not Pull

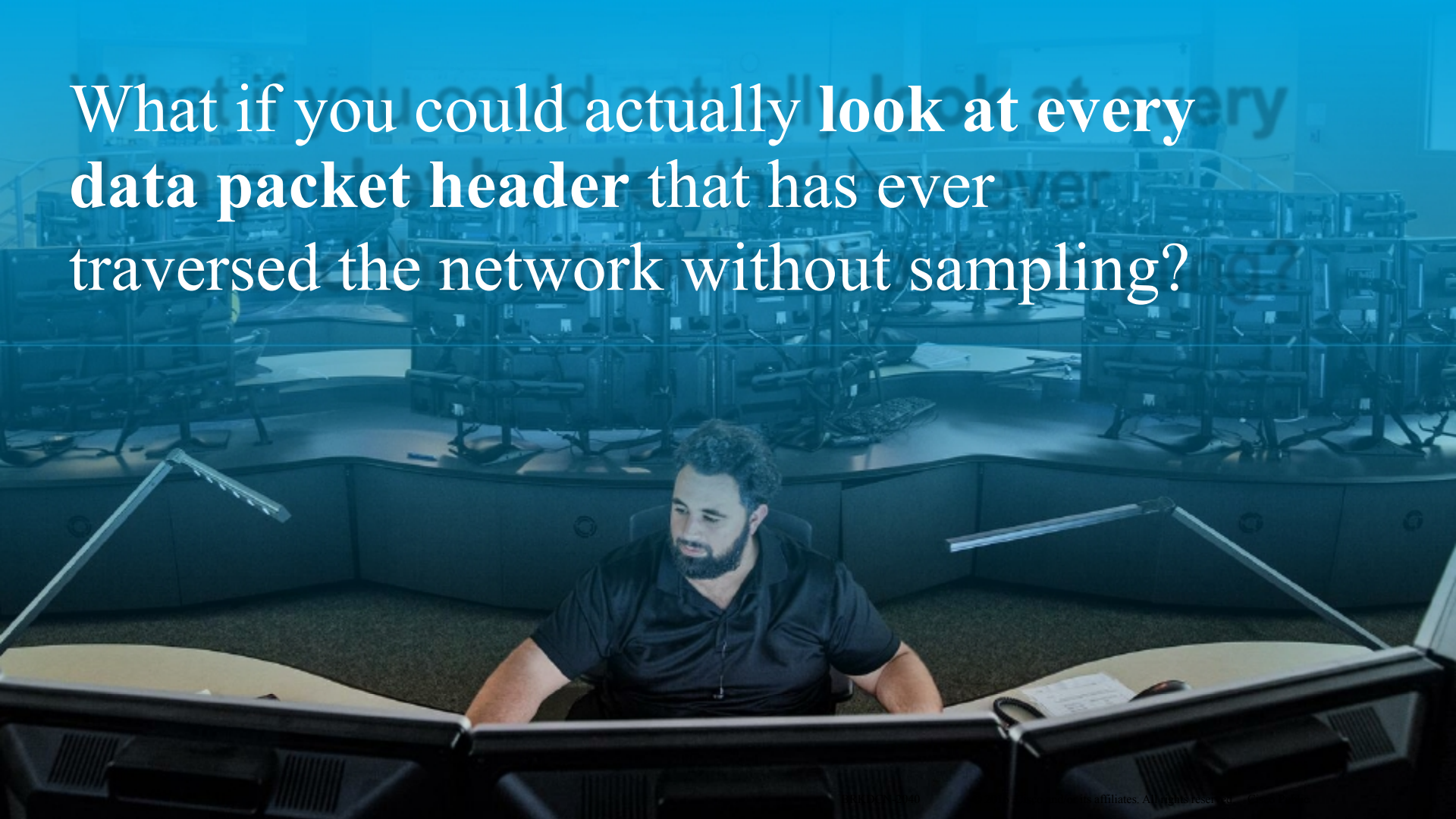


✗ SNMP Pull

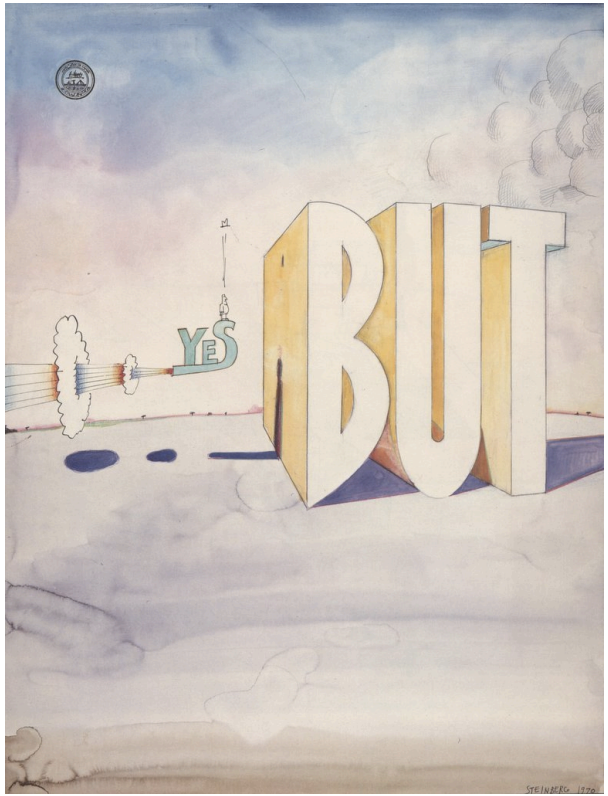
✔ Telemetry Push

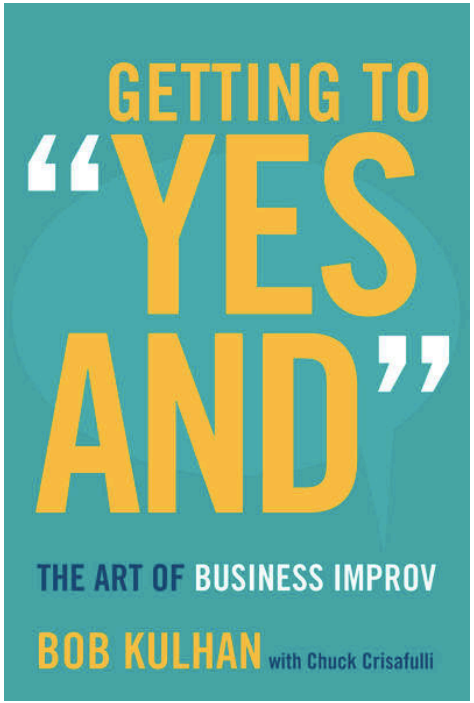
Data Model (YANG)




What if you could actually look at every
data packet header that has ever
traversed the network without sampling?

Culture



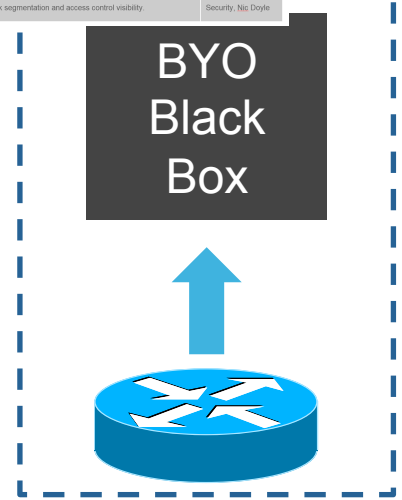


* - http://tumblr.austinkleon.com/post/111960448351

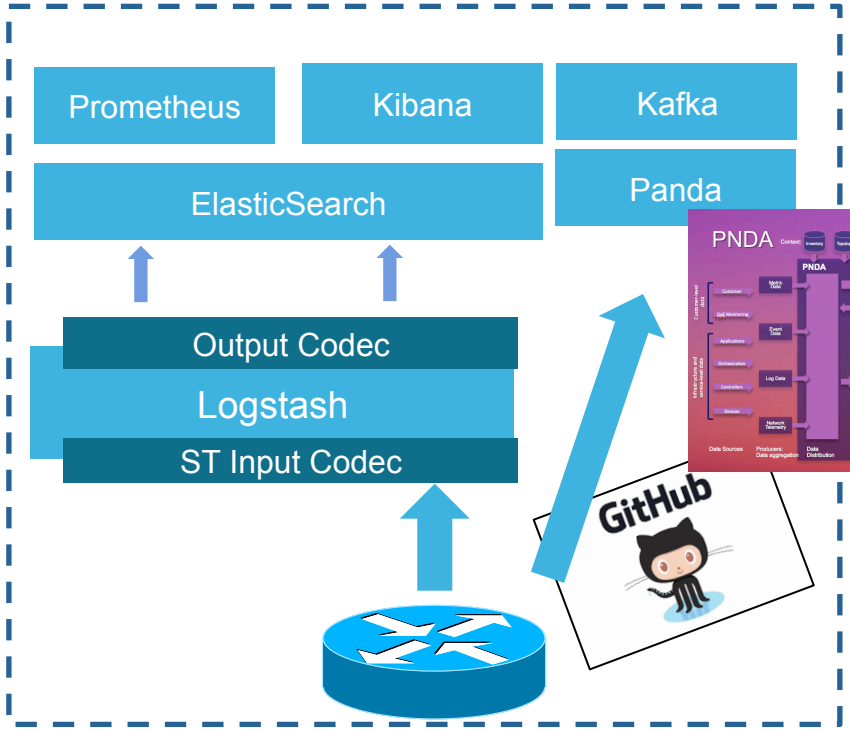
Different Customers, Different Models

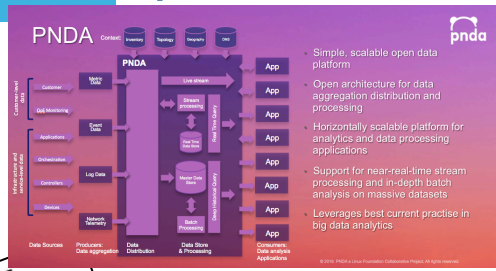
Current Analytics Projects Underway

PROJECT	DESCRIPTION	STAKEHOLDER
Tesseract	Network analytics enablement platform, with a goal of simplifying network analytics by hiding complexities of network, infra, data collection & providing a rich network model centric analytical pipeline	CSG, Rizal Tamsil
ROBOT	A platform that brings together telemetry, big data, machine learning and Model Driven API(s) to truly re-define SP operations by addressing their main OPEX pain-points	CSG, Sanjay Lal
PANDA	Open big data platform initially targeted at OSS layer intelligence; platform is general purpose and can apply to various use cases	CTO, Nick Hall
Spectre	SaaS based full stack monitoring & insights platform giving customers a way to discover how apps are performing, predict how they will perform, and automate ways to cost effectively improve them	Cloud, Thomas Wyatt
Video	New project kicked off to build real-time predictive & security analytics offering for SP Video; looking to build on top of an existing big data platform	SPVSS, Jonathon Beavon
Zeus	SaaS offer for IT admins, application operators, and developers who want operational insights from their telemetry data, specialize in log management initially	Cloud, Kip Compton
Tetration	DC analytics enablement platform; near real-time analytics system capable of watching thousands of servers and the network at wire speed	INSBU
Magellan	Integrated Threat Defense platform focused on network segmentation and access control visibility.	Security, Nic Doyle



Custom





Open Source, Customizable

Gartner categorizes this approach as Algorithmic IT Operations (ITOP) – By 2018 more than 25% of customers will be using this technology

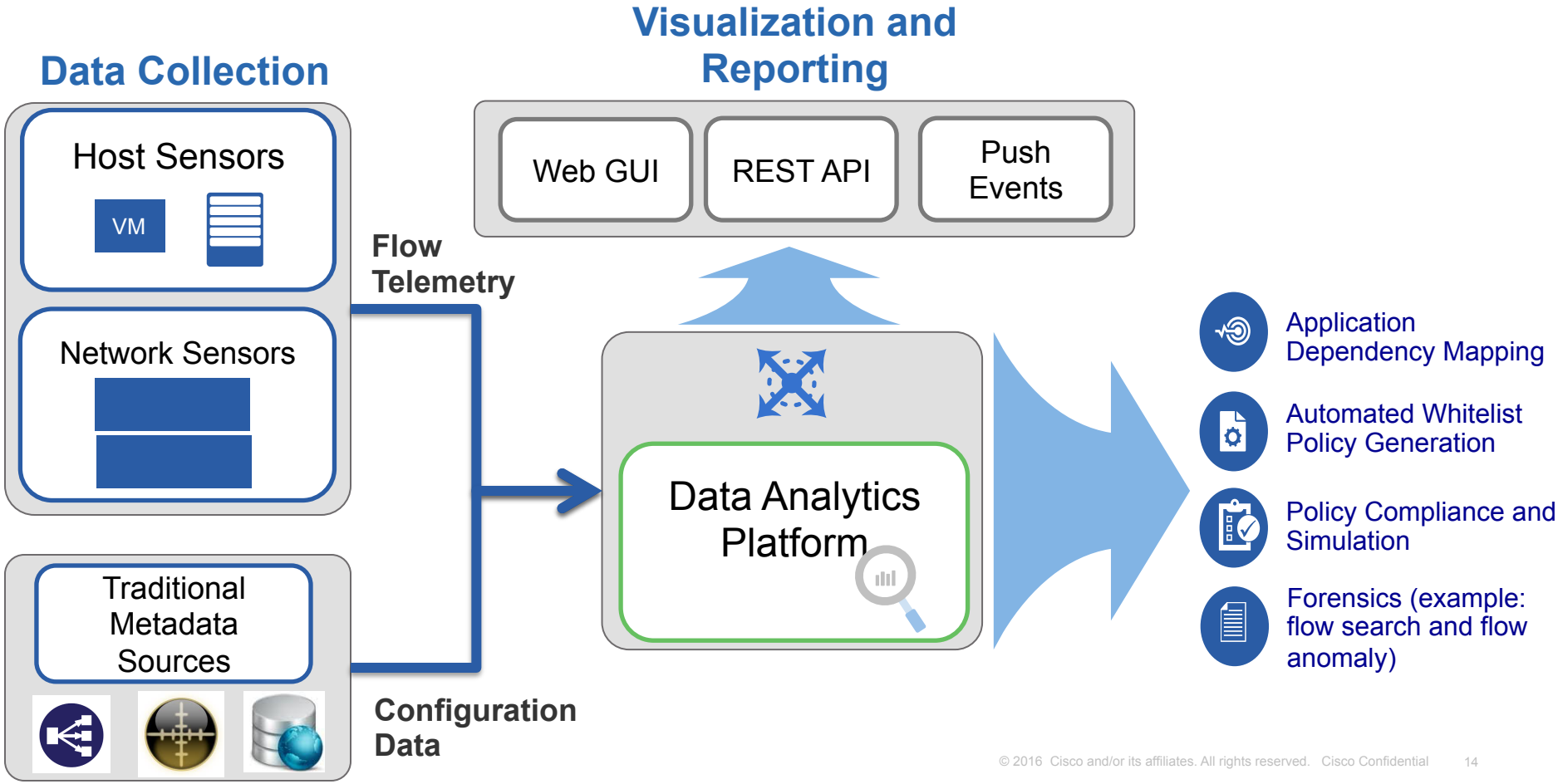
Gartner, IT Operations Market Analysis

Most data is not used currently…
The data that are used today are mostly for anomaly detection and control, not optimisation and prediction, which provide greatest value.

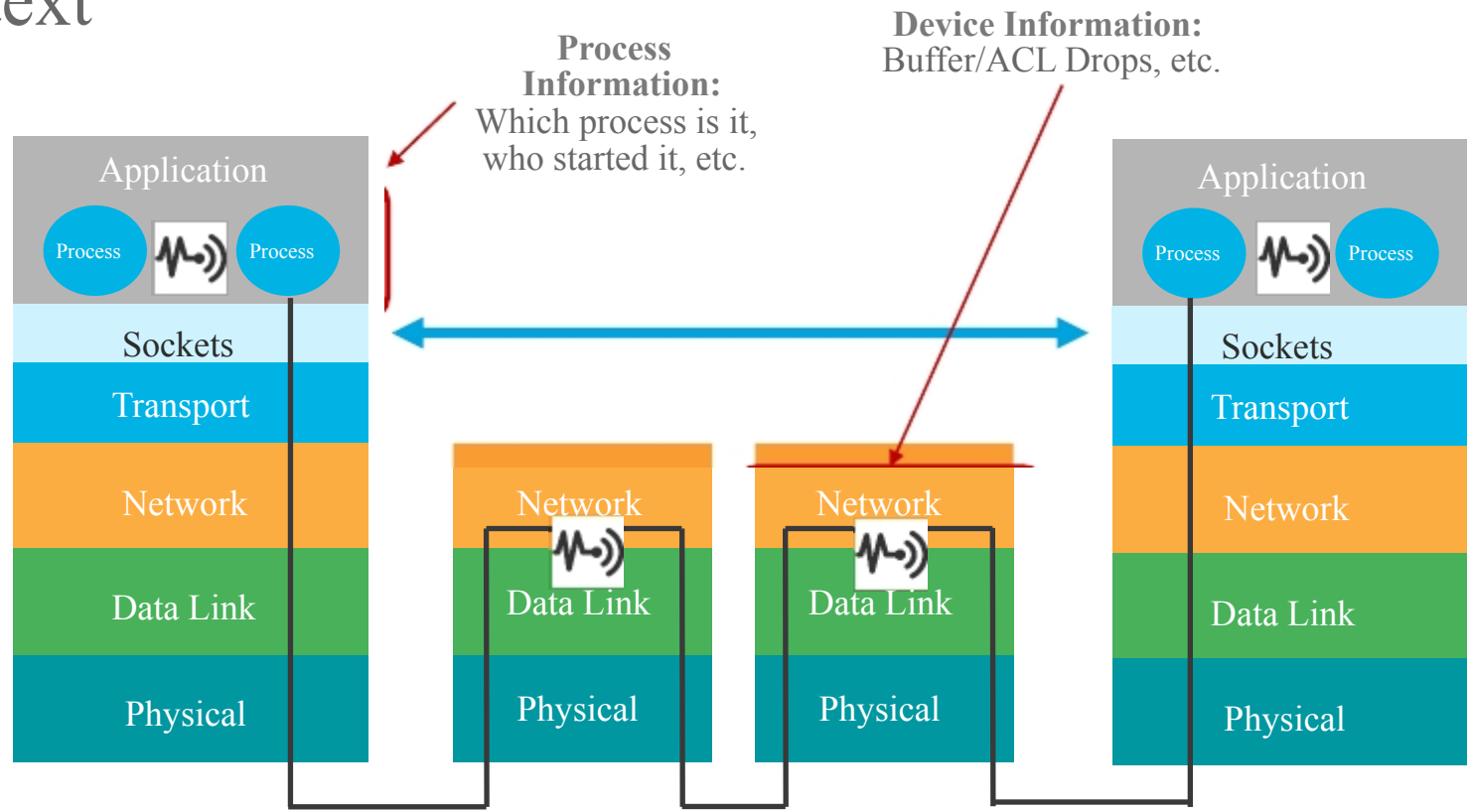
McKinsey, Mapping the Value Beyond the Hype

High Level Architecture

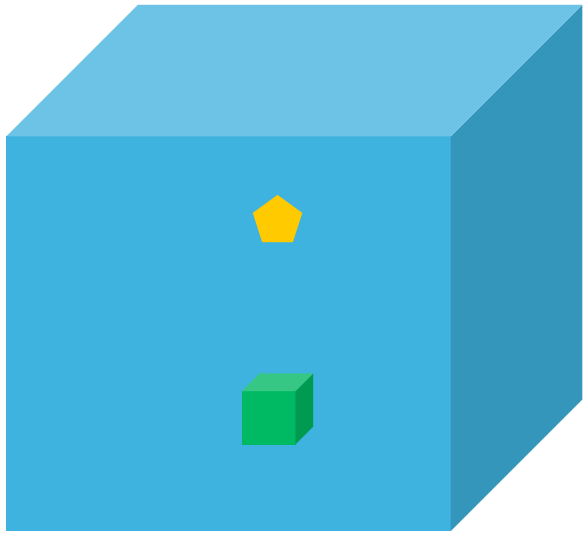
Functional Architecture - Overview



What does the Sensor Collect Context

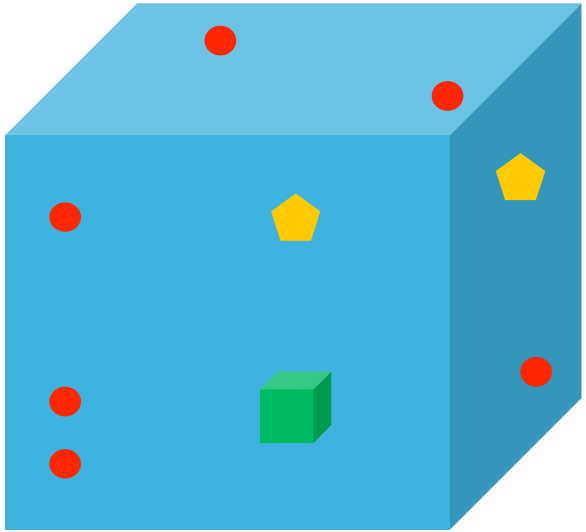


Why Multiple Sensors?
Example monitoring temperature in a room



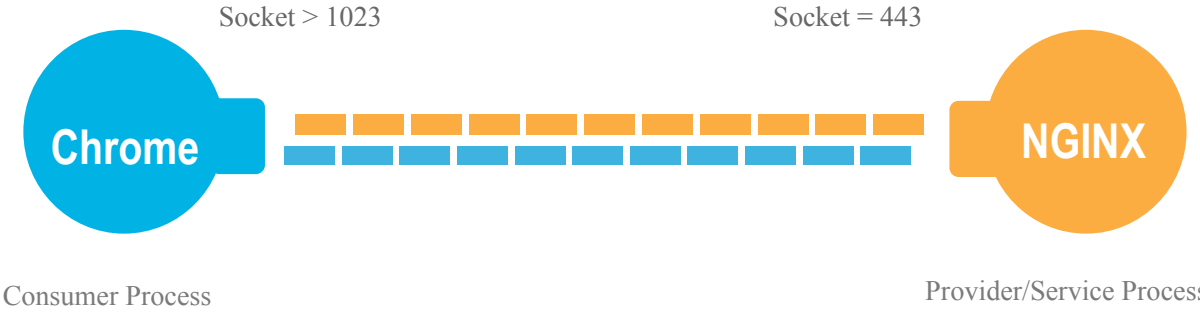
Lamp Sensor

Heater



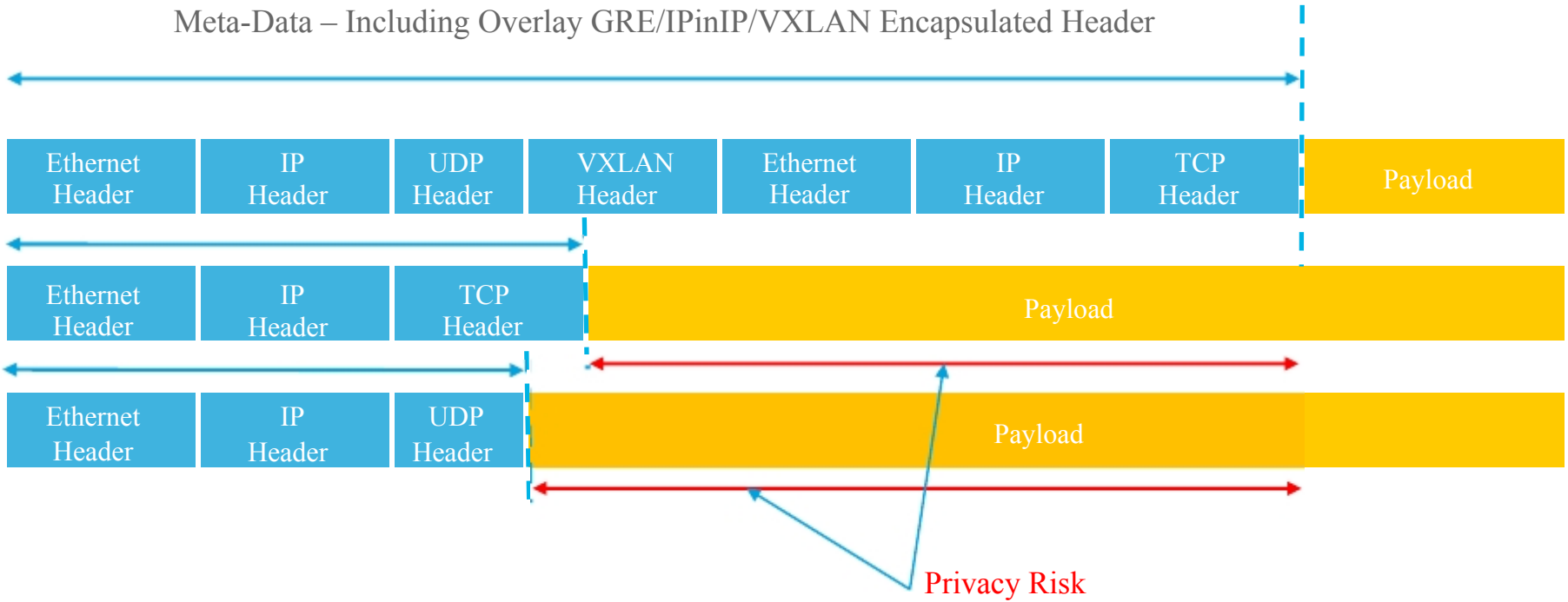
Plug Sensor

Looking Beyond Connectivity
Application Processes and Sockets



- Application developers implement business logic as code that runs as processes and threads
- TCP/IP which forms a foundation of the Internet was designed to allow these application processes to interact via sockets
- Application logic can be viewed on one level as the interaction between a group of processes and their associated sockets
- Understanding the inter-process communication and mapping that directly to the infrastructure provides a direct correlation between the application and the infrastructure

Collects the Meta-Data not the Packet



Machine Learning

Cognitive Computing - Finding and remembering all the relationships between data, querying the matrix of relationships (Watson)

Machine Learning - Remember what has happened before and then look at new data coming in that context to try and find patterns, build up a body of knowledge and then use that data to make a decision based on the new data. Can machines remember and apply what they remember to new data

Deep Learning - Not trying to maintain data and relationships over time but analyze that data through better representations and create model to learn these representations from large scale unlabeled data. Succession analysis

Machine Learning

The programmers construction of algorithms that can learn from and make predictions on data (as opposed to static programming instructions).

7:00 am = 65 degrees
8:00 am = 75 degrees
9:00 am = 85 degrees

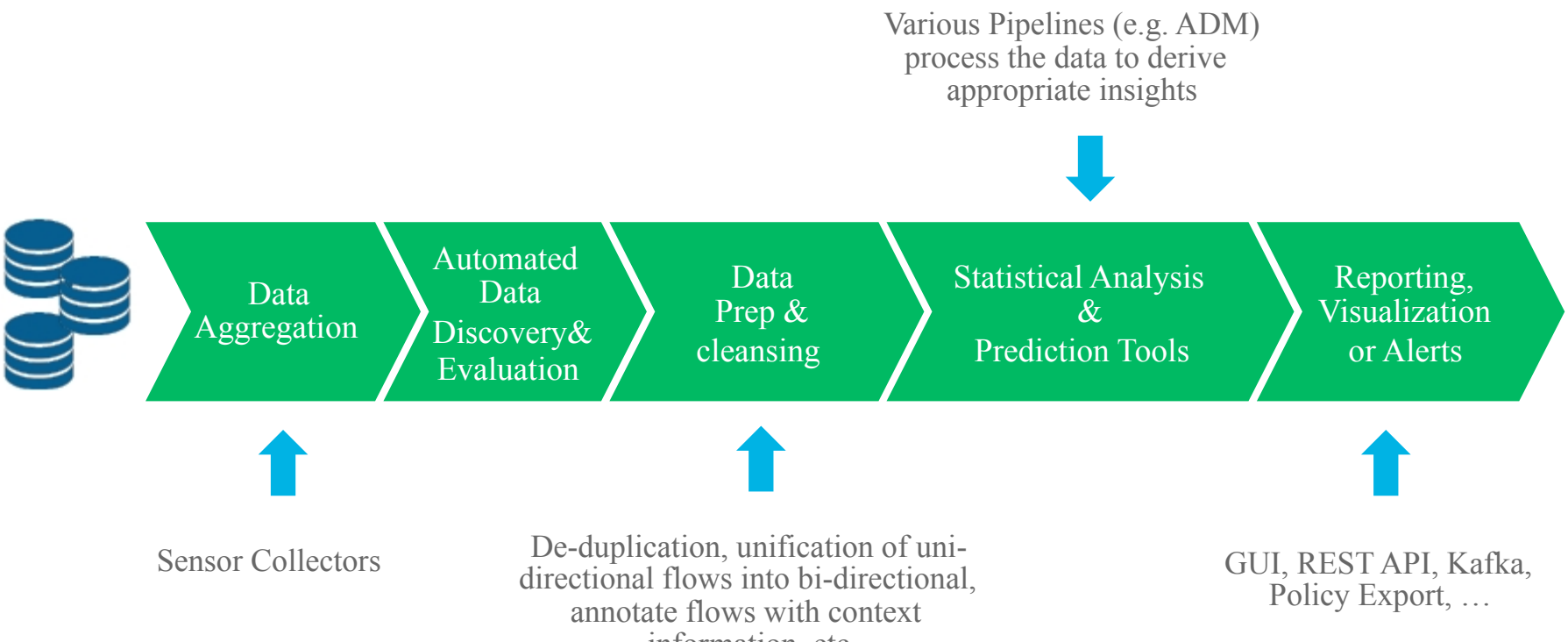
77.5 degrees

How warm will it be at 8:30 am tomorrow?

Supervised learning: Linear regression , Logistics regression, SVMs

Unsupervised learning: K-means, PCA, Anomaly detection

Standard Data Analytics Pipeline
Data Analysis



Appealing Use Cases

Network Team
- Application dependency and visibility
- Flow search and exploration
- Application Latency Information (without Time Synch)

Security Group
- White list policy recommendation
- Policy simulation
- Policy compliance
- Anomaly detection

Server/Application administrator teams
- Application dependency mapping
- End point behavior deviations

Application Dependency Mapping

Why should I understand dependencies?

Identify a single point of failure that should be replicated

Find all the parts of a service that should be migrated together (to the cloud)

Replace infrastructure components of an undocumented application

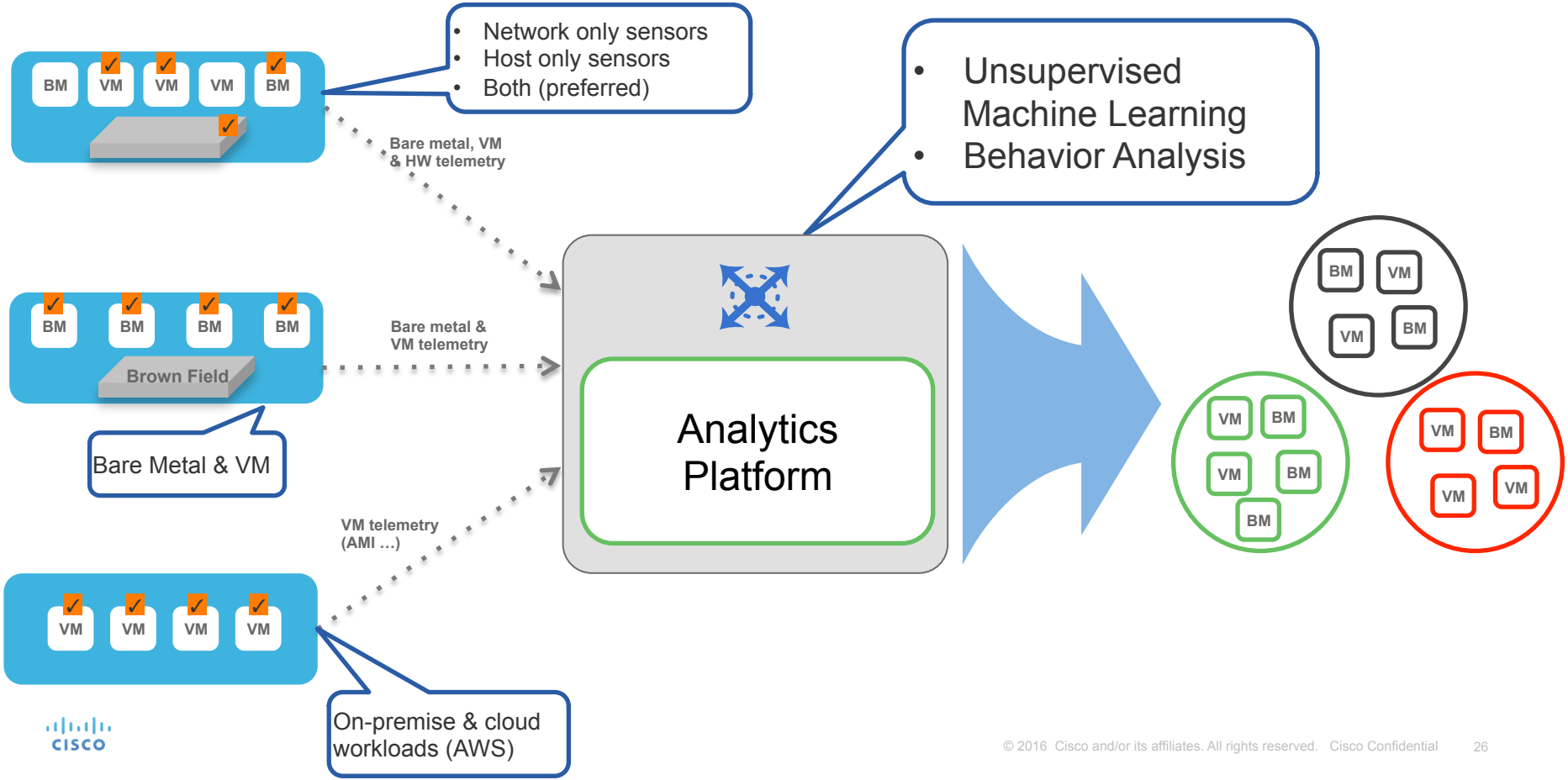
SDN application profiles, end point groups, and contracts based on applications

Why This Approach Is Different

- App insight derived based on actual communication
  - Automated grouping of similar endpoints in a cluster
  - Keep your App insight up-to-date based on application evolution
- Flexibility of using hardware or software sensors

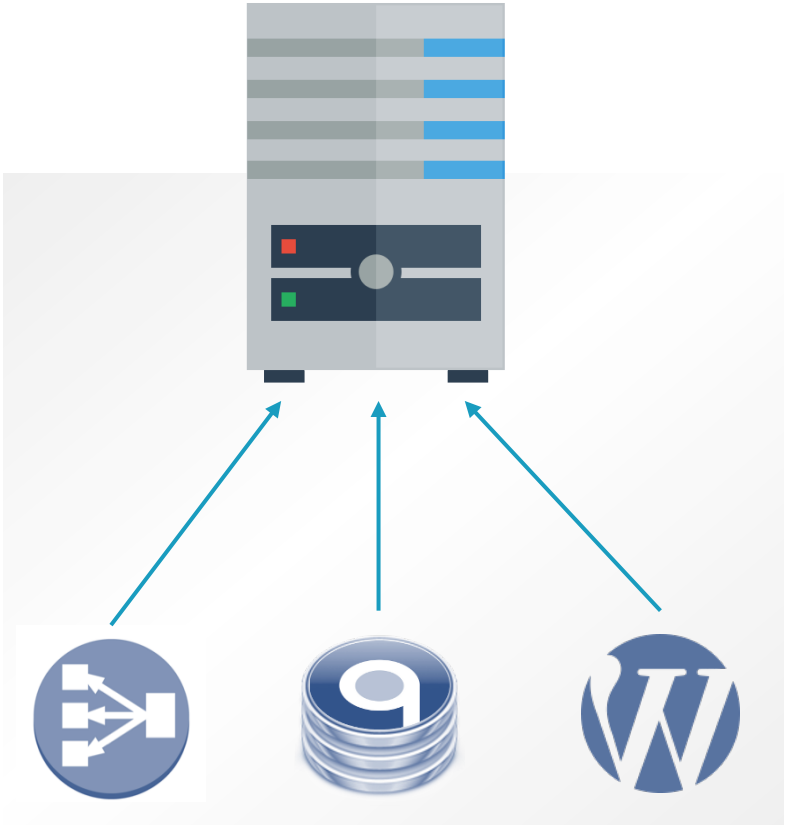
* - image url http://www.mediacore.co.nz/web-applications/web-development

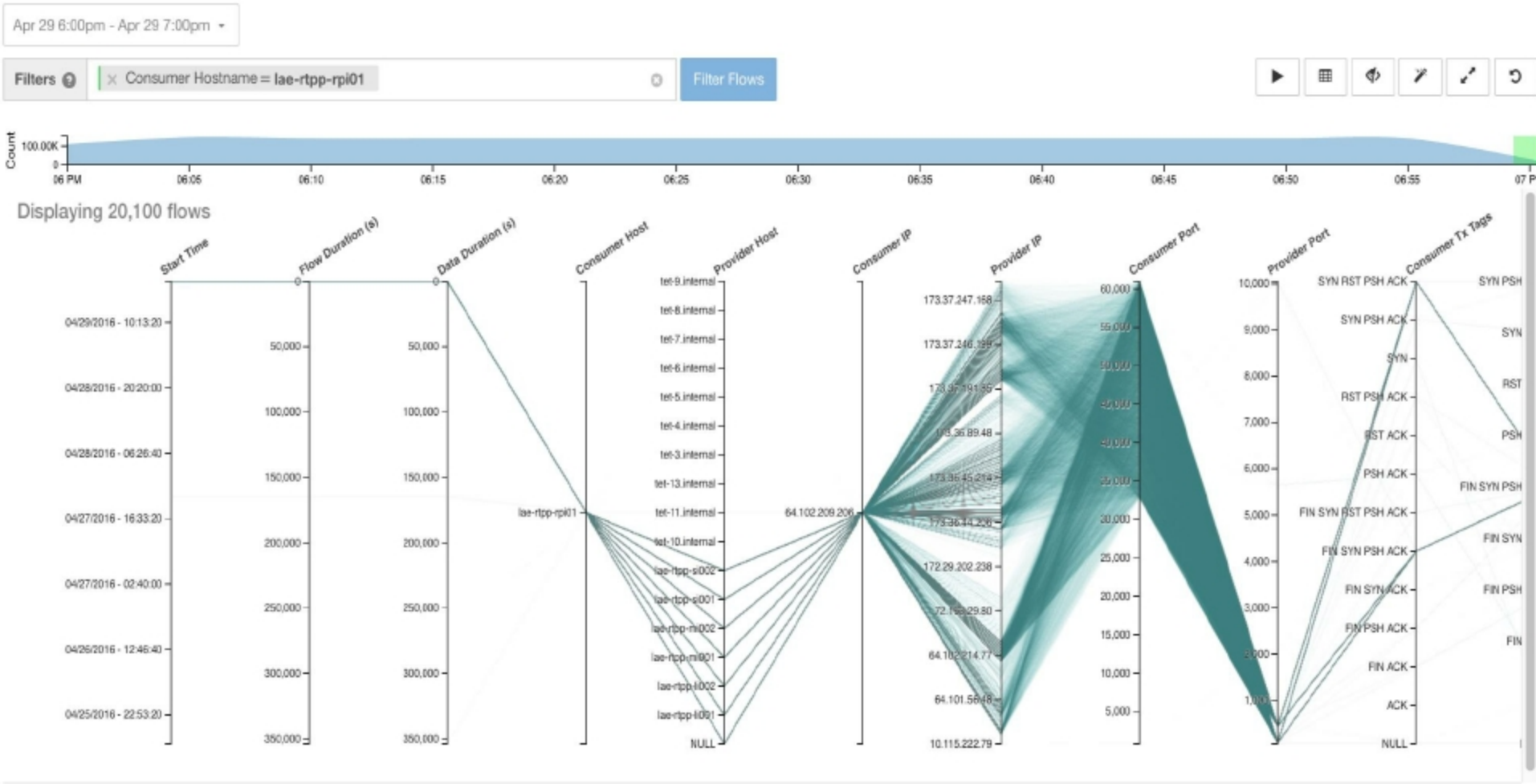
Application Discovery and End Point Grouping



ADM Process

- Every flow from every application is collected and stored
- User selects a time range of flows to perform ADM on (current or historical)
- Side Information such as load balancer configurations, DNS records, and route tags are uploaded



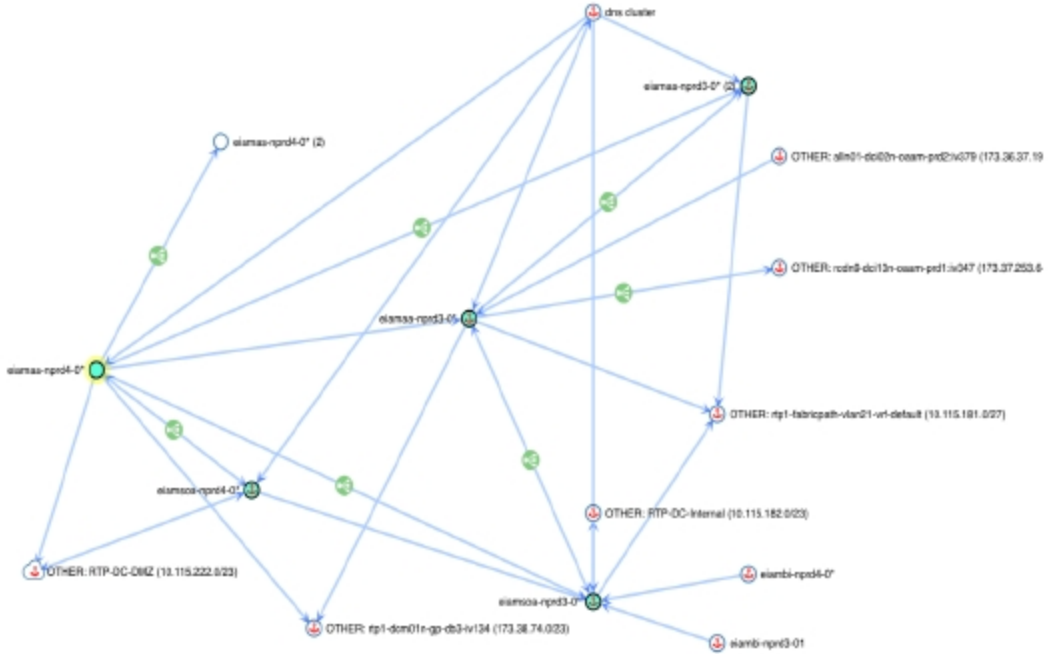


ADM Process (Contd..)

- Mapping algorithm applies unsupervised machine learning to detect clusters
- Operator can modify clustering results based on additional “off network” intelligence
- Accurate clusters are confirmed and remembered for any subsequent repeat

10 Zones | 615 Policies | 2 Applications | 12 Clusters | 11K Conversations | 40 Endpoints

Application View / nprd app



Cluster: eiamaa-nprd4-0*

Name	eiamaa-nprd4-0*
Description	
Confidence	0.862

Endpoints (6)

eiamaa-nprd4-01
173.38.90.144
eiamaa-nprd4-02
173.38.90.135
eiamaa-nprd4-03
173.38.90.136
eiamaa-nprd4-04
173.38.90.137
eiamaa-nprd4-05
173.38.90.138
eiamaa-nprd4-06
173.38.90.139

Neighbors (40)

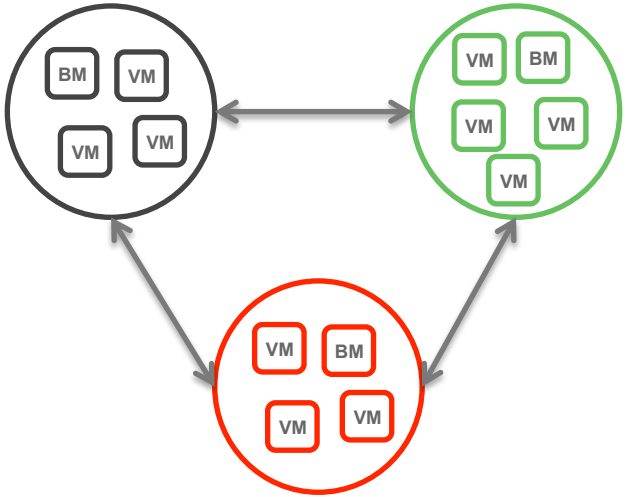
Subnets (2)

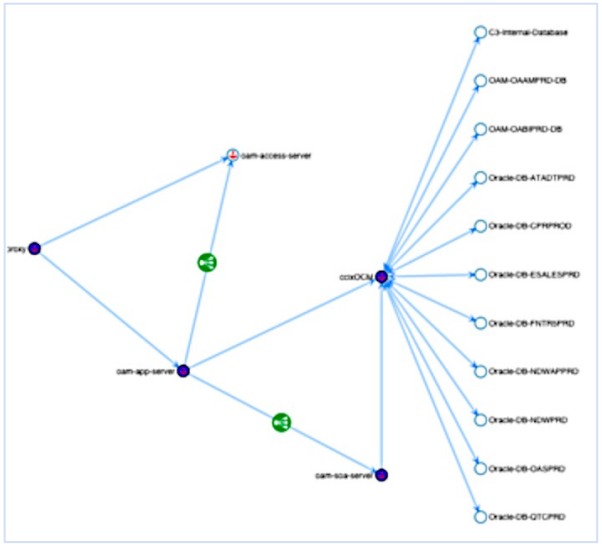
rtp1-dcm02n-oaam-stg:iv756 173.38.90.128/26
RTP-DC-Internal 173.38.88.0/22

Provides (39)

Consumes (203)

Create application dependency map & policy

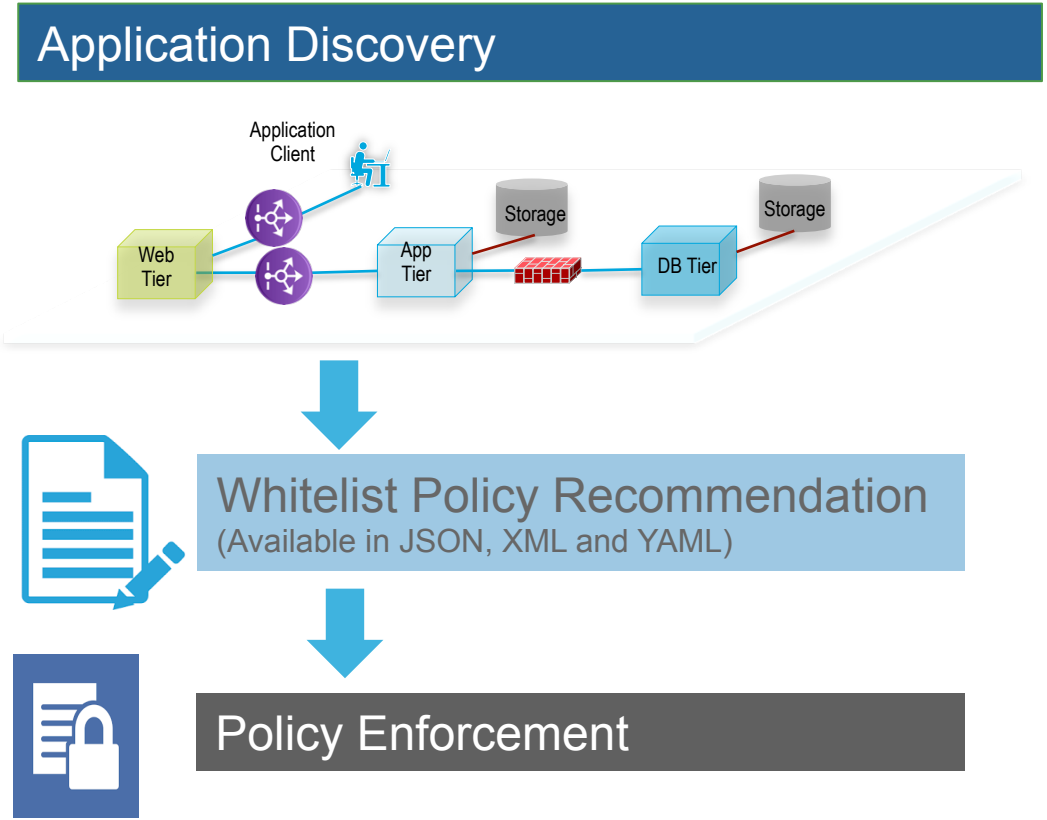




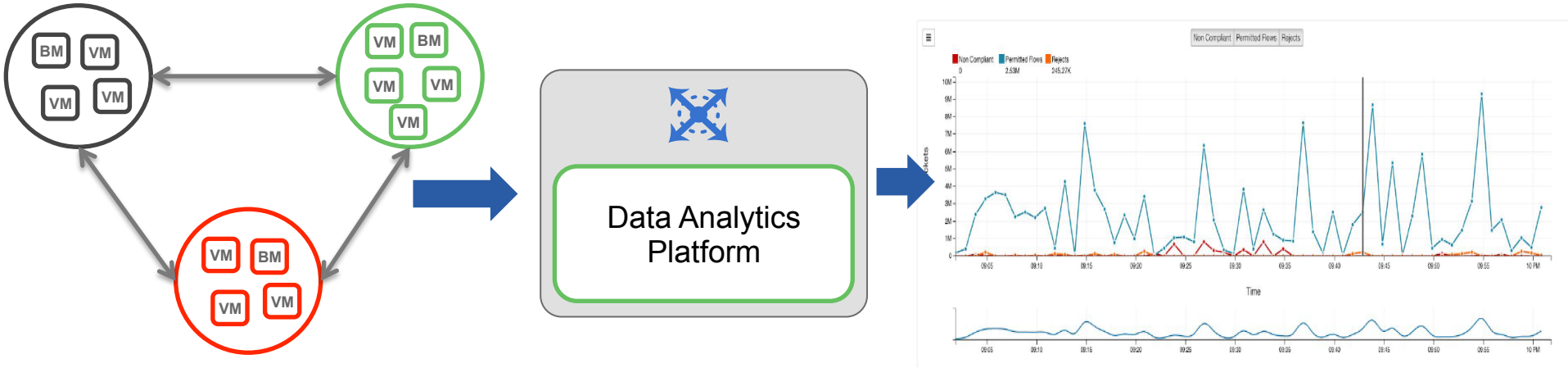
- Eliminate malicious flows
- Create white-list policy based on behavioral analysis
- Manage policy lifecycle

Policy Simulation and Compliance

Whitelist Policy Recommendation and Enforcement

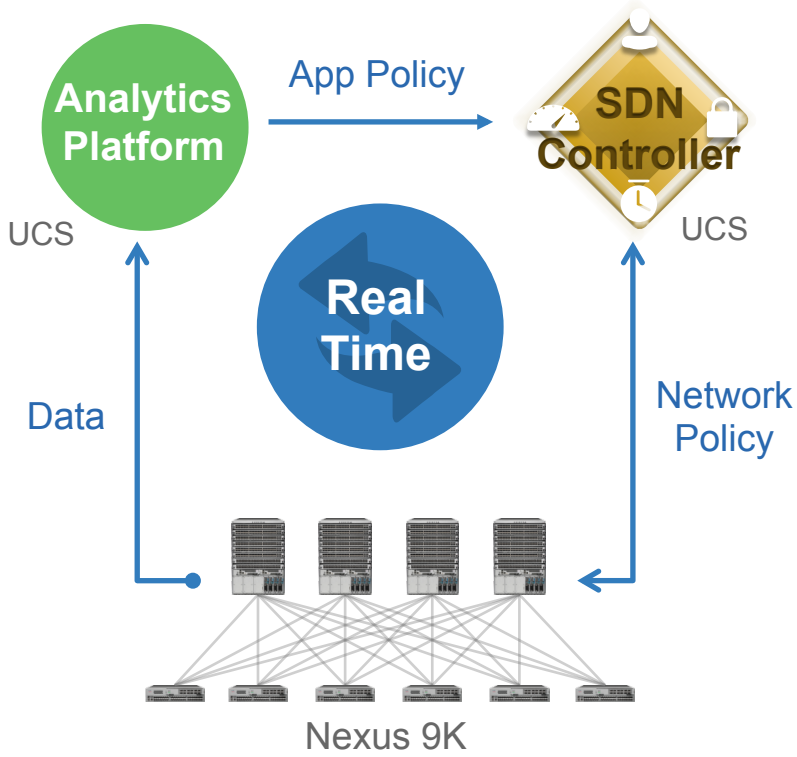


Real time and historical policy simulation



- Policy impact assessment in real time
- Try before apply
- Policy lifecycle management

Get To Zero-Trust Model with SDN Model



Application Policy Recommendation

Import Policy using SDN Toolkit

Automatic creation of End Point Groupss and Contracts

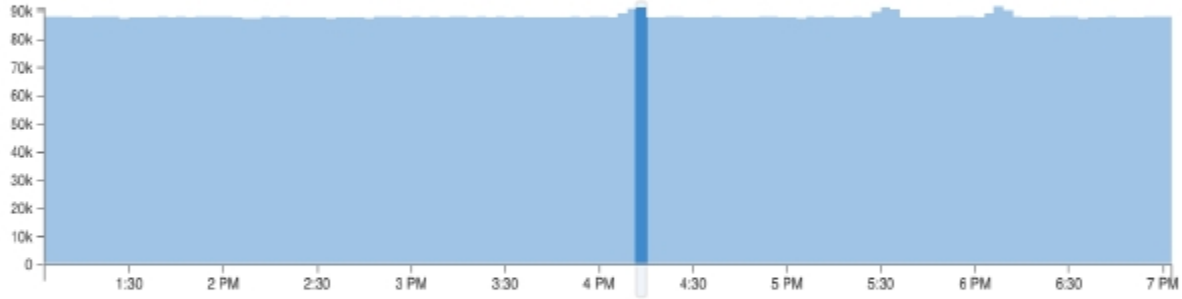
Flow Search and Forensics

Select time range
Apr 29 1:03pm - Apr 29 7:03pm

74,823,632,159 total observations
Showing flow observations

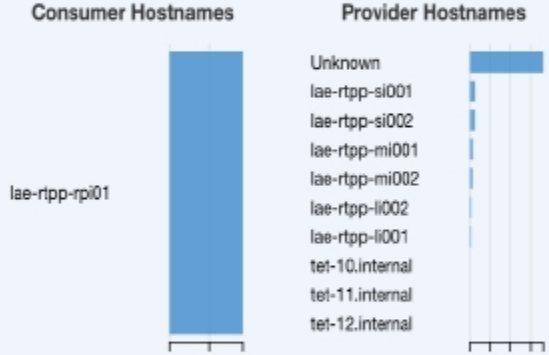
Filters | Consumer Hostname = lae-rtpp-rpi01 | Filter Flows

Filtered | Flow Observations



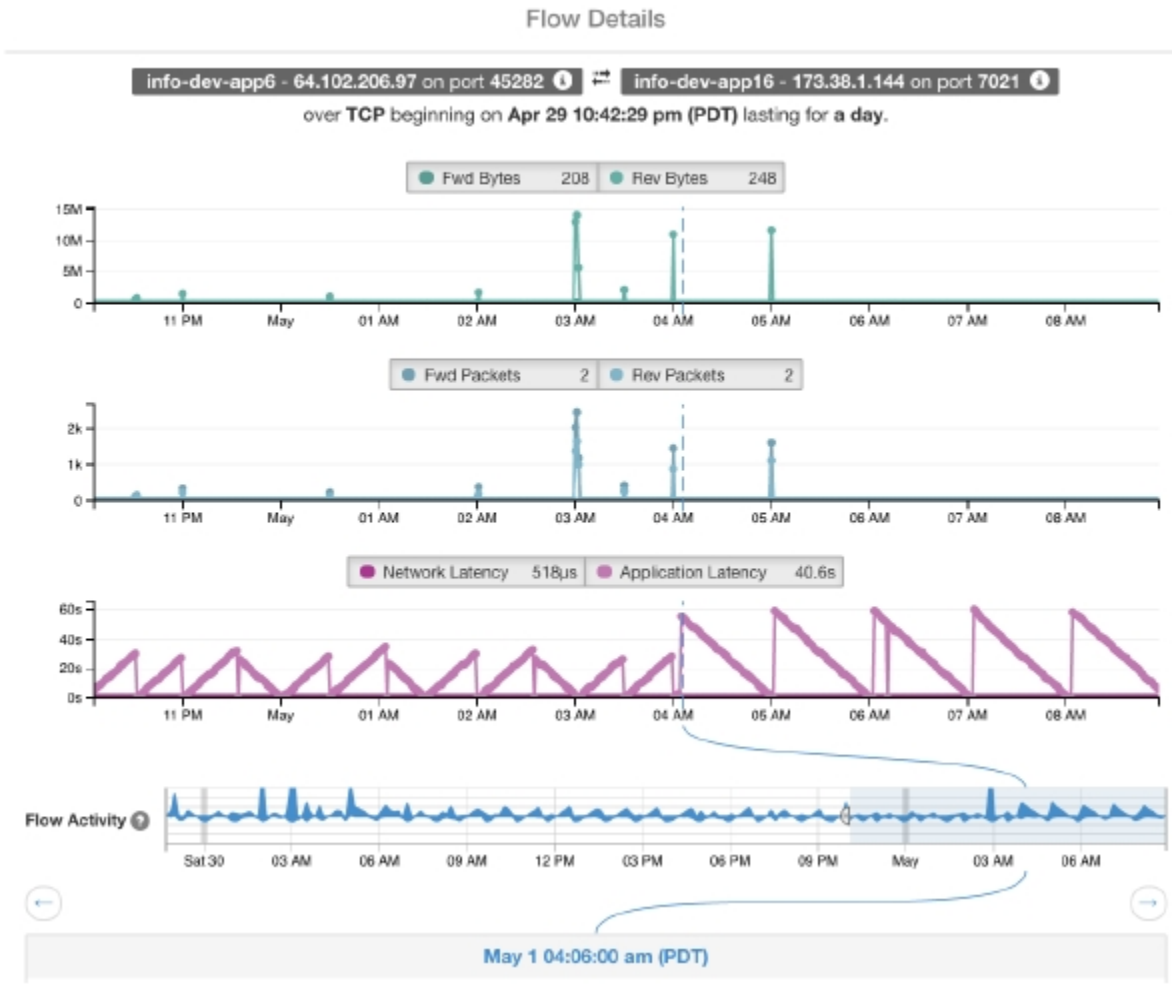
Current selection: **Apr 29 4:12pm** to **Apr 29 4:15pm**

Top Hostnames contributing to the selected **flow observations**.



Found **90,918** flow observations (185ms), showing 100 Load more

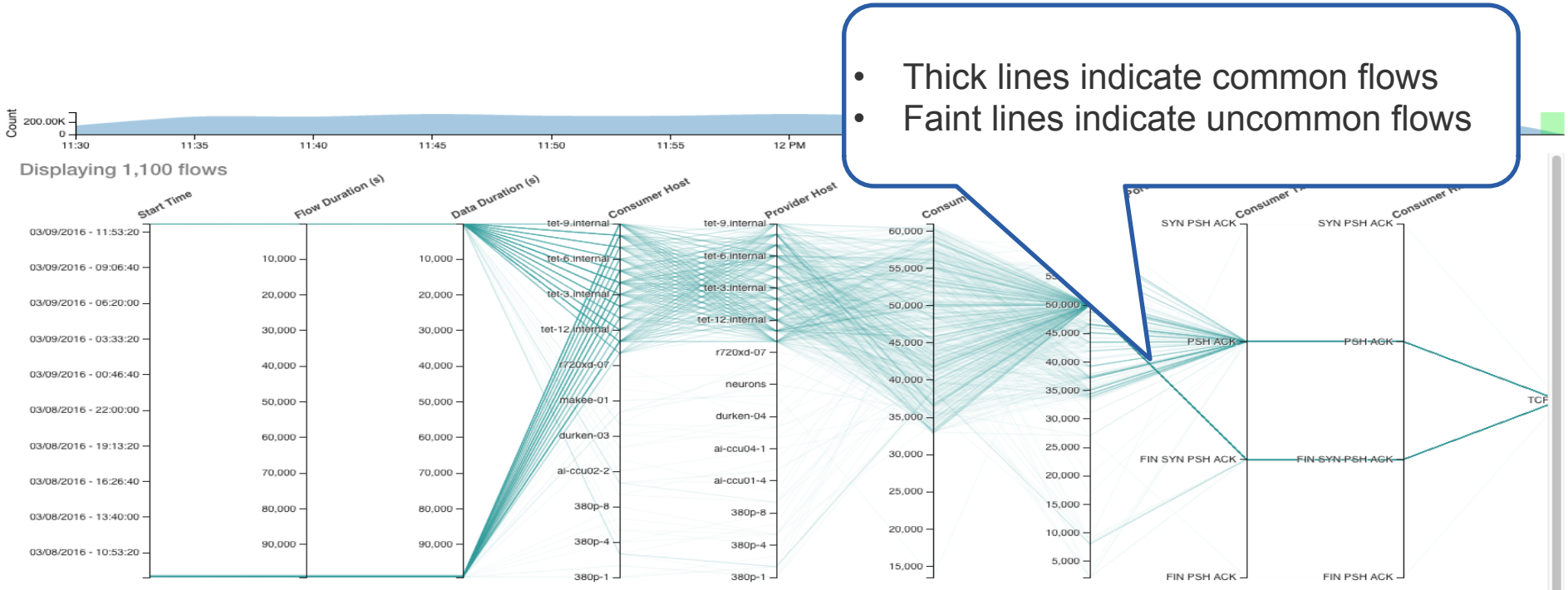
Timestamp	Consumer Hostname	Consumer Address	Consumer Port	Provider Hostname	Provider Address	Provider Port	Protocol	Flow Type	Flow Start Time	Consum
Apr 29 4:12:00pm	lae-rtpp-rpi01	64.102.209.206	50214	Unknown	64.101.56.249	443 (HTTPS)	TCP	IPv4	Apr 29 4:12:22pm	
Apr 29 4:12:00pm	lae-rtpp-rpi01	64.102.209.206	49317	Unknown	173.37.246.41	443 (HTTPS)	TCP	IPv4	Apr 29 4:12:34pm	
Apr 29 4:12:00pm	lae-rtpp-rpi01	64.102.209.206	60067	Unknown	173.37.246.96	80 (HTTP)	TCP	IPv4	Apr 29 4:12:39pm	



May 1 04:06:00 am (PDT)		
	Consumer	**Provider**
Flags	PSH ACK	PSH ACK
Byte Count	208 (3,575,500 so far)	248 (160,829,080 so far)
Packet Count	2 (28,820 so far)	2 (25,760 so far)
Application latency	54.4s	
Network latency	509µs	
Consumer TX Policies	PERMITTED:INTRA_EPG_FLOW_DEFAULT:policy-TCP:*:0-65535~Internet~Internet:cont-TCP:*:0-65535~Internet~Internet:ALLOW:0-65535:0-65535:ANY#PERMITTED:INTRA_EPG_FLOW_DEFAULT:policy-TCP:*:0-65535~Internet~Internet:cont-TCP:*:0-65535~Internet~Internet:ALLOW:0-65535:0-65535:ANY	
Consumer RX Policies	PERMITTED:INTRA_EPG_FLOW_DEFAULT:policy-TCP:*:0-65535~Internet~Internet:cont-TCP:*:0-65535~Internet~Internet:ALLOW:0-65535:0-65535:ANY#PERMITTED:INTRA_EPG_FLOW_DEFAULT:policy-TCP:*:0-65535~Internet~Internet:cont-TCP:*:0-65535~Internet~Internet:ALLOW:0-65535:0-65535:ANY	

Visual Query with Flow Exploration

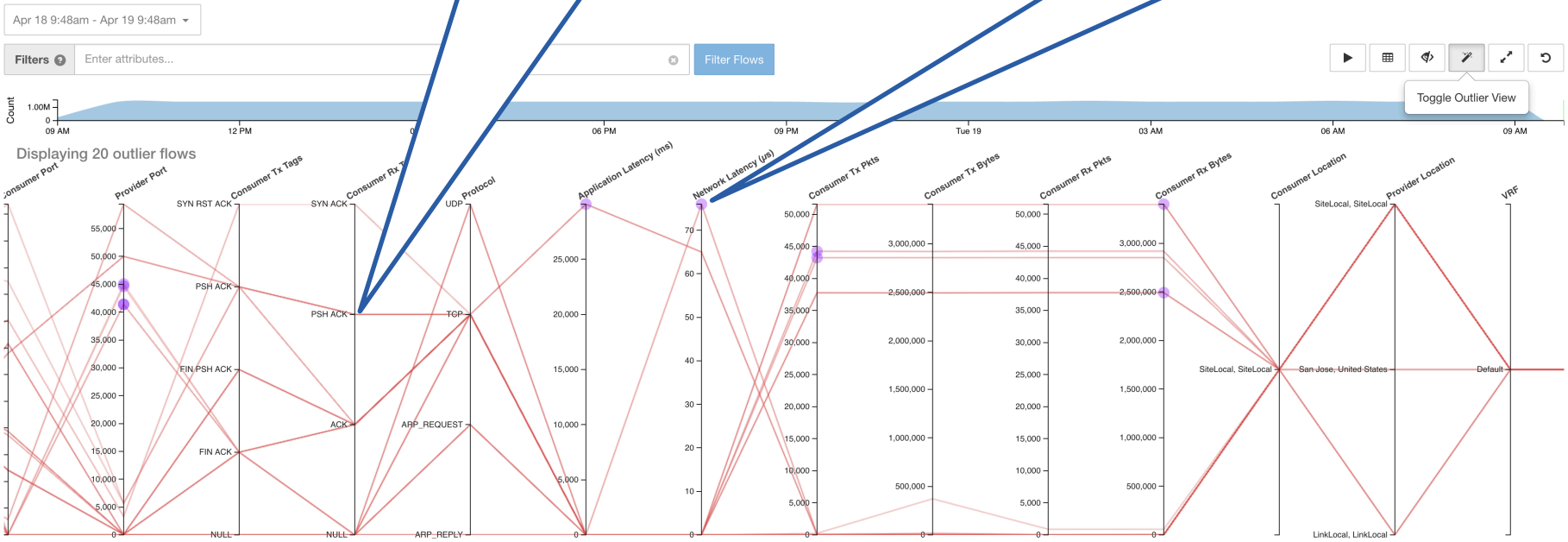
- Replay flow details like a DVR
- Information mapped across 25 different dimensions



Outliers

- Switch on Outlier view to highlight uncommon flows

- Outlier dimension is highlighted with purple circle

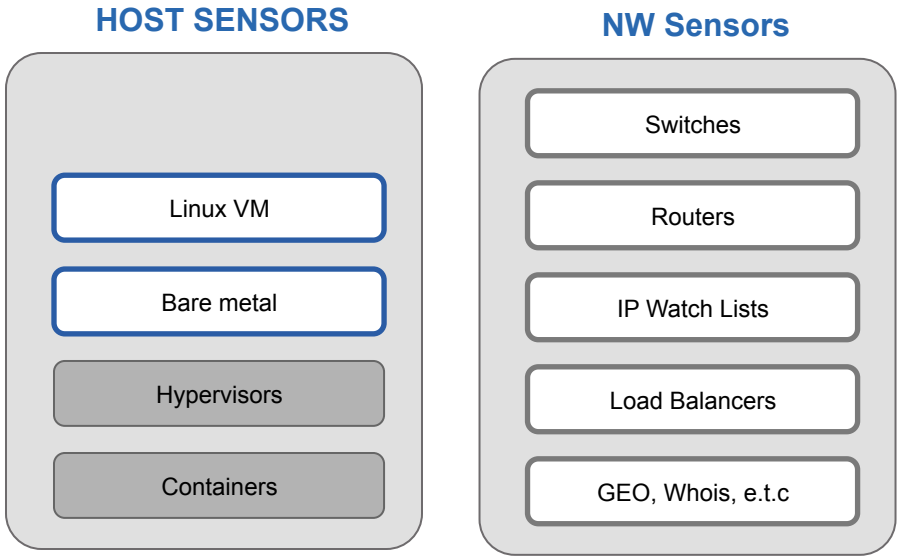


Monetizing Data Analytics – Key Segments

- Data Analytics appeals to all segments in general
- Key Segments of interest are
  - Health Care
  - Financials
  - Public sector
  - Large Enterprises

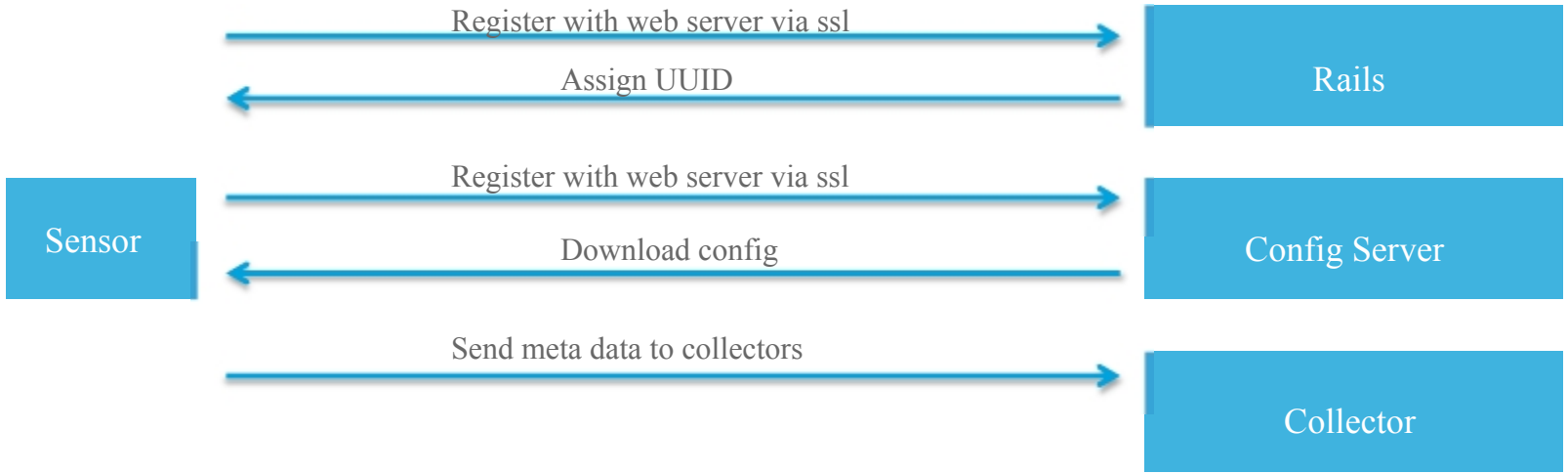
Thank you

Pervasive Sensors

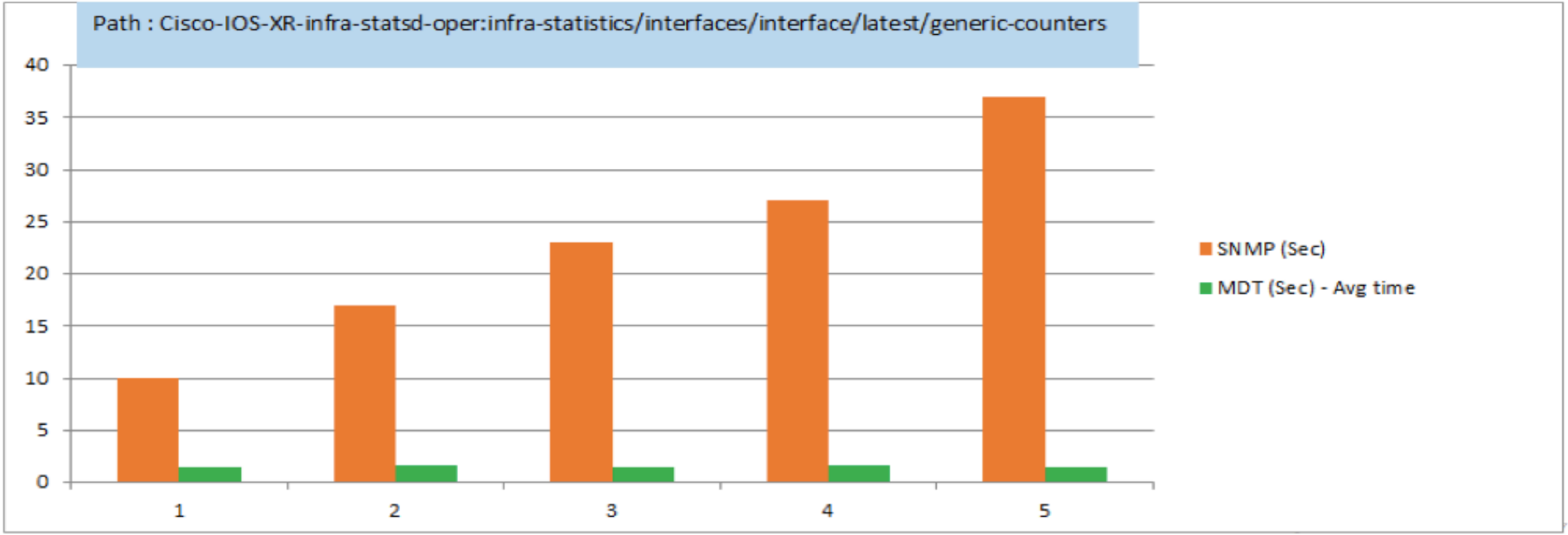


- ✓ **Low CPU Overhead (SLA enforced)**
- ✓ **Low Network Overhead (SLA enforced)**
- ✓ **Highly Secure (Code Signed, Authenticated)**
- ✓ **Every flow (No sampling), NO PAYLOAD**

How Sensor Communicate with the Cluster the First Time?



Lesson Learned: It's Not Hard to Beat SNMP



- 10 second poll / push
- 3 pollers / telemetry receivers
- 30 minute measurement intervals

- 288 100Gig E Interfaces (Line Rate)
- SNMP: IF-MIB (query by row)

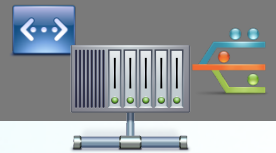
And there is more…

Pervasive flow telemetry that supports multiple datacenter infrastructure and at scale

Look for anomalous data patterns in hardware and generate events to collector

Calculate detailed network and application latency information even without time synchronization

Detailed flow performance and accounting tracked through the entire life of flow



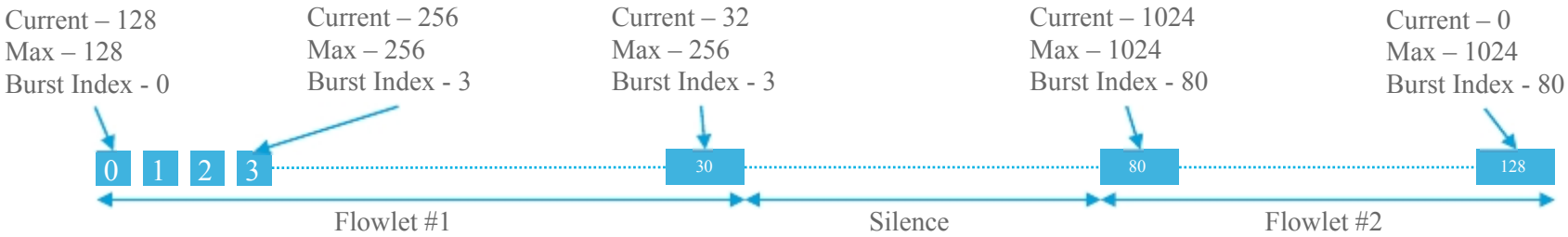
Sensor Data
Accumulated Flow Information

- Bytes, Packet Count
- IP options present
- IP length error
- DF bit set
- Fragment seen
- Last TTL
- Accumulated TCP flags
- Last ACK / SEQ
- Sampled Packet length
- Sampled Packet ID

Sensor Data Burst

- Measure the “ burstiness ” of a flow
  - Current Burst
  - Max Burst
  - Burst Index
  - Flowlets

- Burst are measured in 32k interval
- Each export period is divided by 128
- Flowlets are activity after a silence period (configurable)



Max Burst occurred at 62.5ms with a value of 1024 and 2 flowlets

Sensor Data Anomaly List

- TTL changed
- IP reserved flags are not 0
- DF bit has changed
- Ping of death
- Fragment is too small to contain L4 header (TCP, UDP and SCTP)
- TCP SYN and FIN are set
- TCP SYN and RST are set
- TCP FIN, PSH and URG are set
- TCP flags are zero'd
- TCP SYN with data
- TCP FIN with no ACK
- TCP RST with no ACK
- TCP SYN, FIN, RST and ACK zero'd
- URG set but no URG pointer
- URG pointer with no URG flag
- TCP seq outside the expected range
- TCP seq is less than expected (rexmit)