Big Data Platform

Lessons Learned in Growing a Big Data Capability for Network Defense

Who am I?

- Technical Director, Enlighten IT Consulting, a MacAulay-Brown company
- Software Engineering Consultant
- Helped found Apache Rya
- Chief Architect of DoD’s Big Data Platform
- Currently working for:
  - Defense Information Systems Agency (DISA)
  - Army Cyber Command
  - US Cyber Command
  - Center for Army Analysis
  - Air Force

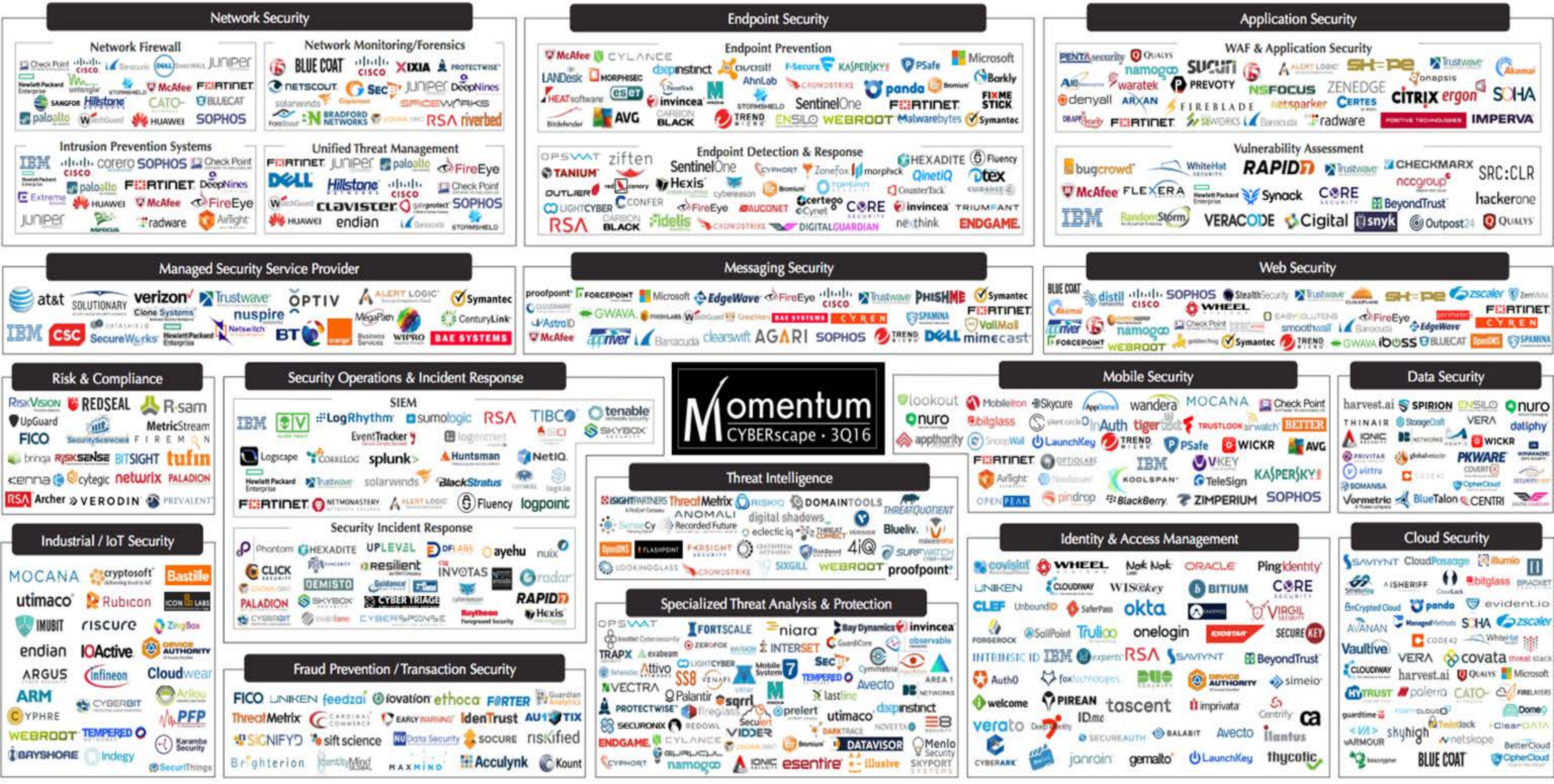
Talk Overview

- DCO Big Data Problem Space
- DoD’s Big Data Platform
- Scaling for Big Data
- Multi-Tenancy
- Lessons Learned

Problem Space

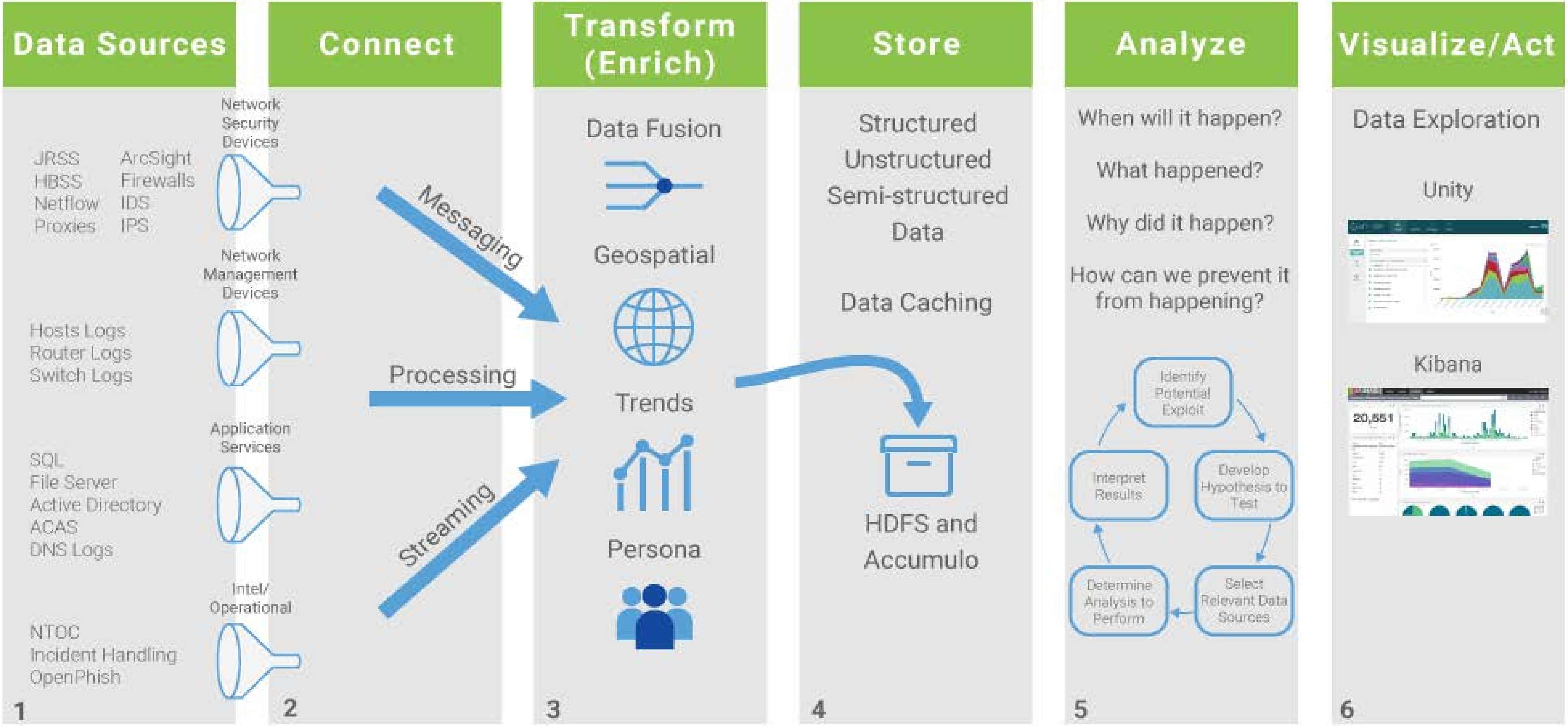
- Huge variety of DCO sensors
- Heterogeneous data formats
- No enterprise standardization on infrastructure
- Petabyte scale storage/retention/analysis requirements
- No single “out of the box” COTS, GOTS, or OSS solution by itself meets the unique DoD cyber security challenges
- Enabling collaborative investigation while eliminating redundant efforts

Problem Space


Momentum CYBERscape · 3Q16
Network Security
Network Firewall
Check Point, Cisco, Barracuda, Dell, Juniper, Hewlett Packard Enterprise, untangle, McAfee, Fortinet, Sangfor, Hillstone, Cato, Bluecat, paloalto, WatchGuard, Huawei, Sophos
Network Monitoring/Forensics
Blue Coat, Cisco, Ixia, Protectwise, Netscout, Sec, Juniper, DeepNines, solarwinds, Gigamon, SpiceWorks, ForeScout, Bradford Networks, RSA, riverbed
Intrusion Prevention Systems
IBM, Cisco, corero, Sophos, Check Point, Hewlett Packard Enterprise, paloalto, Fortinet, DeepNines, Extreme, Huawei, McAfee, FireEye, Juniper, radware, Airlight
Unified Threat Management
Fortinet, Juniper, paloalto, FireEye, Dell, Hillstone, Cisco, Check Point, WatchGuard, clavister, gateprotect, Sophos, Huawei, endian, Barracuda, Stormshield
Endpoint Security
Endpoint Prevention
McAfee, Cylance, LANDesk, Morphisec, deepinstinct, avast, F-Secure, Kaspersky, PSafe, Microsoft, HEAT software, eset, AhnLab, CrowdStrike, panda, Bromium, Barkly, invincea, Stormshield, SentinelOne, Fortinet, FixMeStick, Bitdefender, AVG, Carbon Black, Trend Micro, enSilo, Webroot, Malwarebytes, Symantec
Endpoint Detection & Response
OPSWAT, ziften, Tanium, SentinelOne, Cyphort, ZoneFox, morphick, Hexadite, Fluency, Outlier, Hexis, cybereason, Bromium, QintetiQ, Dtex, LightCyber, Confer, FireEye, Aujas, Counter Tack, Guidance, certego, Cynet, Core Security, invincea, Triumfant, RSA, Carbon Black, Fidelis, CrowdStrike, Digital Guardian, nexthink, Endgame
Application Security
WAF & Application Security
Penta Security, Qualys, namogoo, Sucuri, Alert Logic, Shape, Trustwave, Akamai, A10, waratek, Prevoty, onapsis, denyall, Arxan, NSFocus, Zenedge, Citrix, ergon, Soha, Fireblade, netsparker, Certes, Fortinet, Signal Sciences, Barracuda, radware, Positive Technologies, Imperva
Vulnerability Assessment
bugcrowd, WhiteHat Security, Rapid7, Trustwave, Checkmarx, nccgroup, SRC:CLR, McAfee, Flexera, Hewlett Packard Enterprise, Synack, Core Security, BeyondTrust, hackerone, IBM, RandomStorm, Veracode, Cigital, snyk, Outpost24, Qualys
Managed Security Service Provider
at&t, Solutionary, verizon, Trustwave, Optiv, Alert Logic, Symantec, Clone Systems, nuspire, MegaPath, CenturyLink, IBM, CSC, Datashield, SecureWorks, Hewlett Packard Enterprise, Netswitch, BT, orange, Business Services, Wipro, BAE Systems
Messaging Security
proofpoint, Forcepoint, Microsoft, EdgeWave, FireEye, Cisco, Trustwave, PhishMe, Symantec, Cloudmark, GWAVA, Fishlabs, WatchGuard, GreatHorn, BAE Systems, Cyren, Spamina, Fortinet, AstroID, VadeSecure, McAfee, appriver, Barracuda, clearswift, Agari, Sophos, Trend Micro, Dell, mimecast
Web Security
Blue Coat, distil networks, Cisco, Sophos, StealthSecurity, Trustwave, Cloudflare, Shape, zscaler, Zenedge, Akamai, Wheel, EasySolutions, Fortinet, appriver, namogoo, Check Point, smoothwall, Barracuda, FireEye, EdgeWave, Cyren, Forcepoint, Webroot, Symantec, Trend Micro, GWAVA, iboss, Bluecat, OpenDNS, Spamina
Risk & Compliance
RiskVision, RedSeal, R-sam, UpGuard, MetricStream, FICO, SecurityScorecard, Firemon, bringa, RiskSense, BitSight, tufin, kenna, cytegic, netwrix, Paladion, RSA Archer, Verodin, Prevalent
Security Operations & Incident Response
SIEM
IBM, Alien Vault, LogRhythm, sumologic, RSA, TIBCO, tenable, EventTracker, logentries, SkyBox, Logscape, Correlog, splunk, Huntsman, NetIQ, Hewlett Packard Enterprise, Trustwave, solarwinds, BlackStratus, Fortinet, Netmonastery, Alert Logic, Fluency, logpoint
Security Incident Response
Phantom, Hexadite, Uplevel, DFLabs, ayehu, nuix, Click Security, Resilient, Invotas, radar, Paladion, Demisto, Guidance, Skybox, Cyber Triage, RAPID7, Cybersponse, Raytheon, Hexis
Threat Intelligence
iSight Partners, ThreatMetrix, RiskIQ, DomainTools, ThreatQuotient, Anomali, digital shadows, SenseCy, Recorded Future, eclectic iq, ThreatConnect, Blueliv, OpenDNS, Flashpoint, Farsight, Centripetal Networks, 4iQ, SurfWatch, LookingGlass, CrowdStrike, SixGill, Webroot, proofpoint
Specialized Threat Analysis & Protection
OPSWAT, Fortscale, niara, Bay Dynamics, invincea, ZeroFox, InterSet, GuardiCore, observable, TrapX, exabeam, LightCyber, Mobile System 7, SecBI, Attivo, SS8, Venafi, Tempered Networks, Cymmetria, Area 1, Vectra, Palantir, sqrrl, Avecto, Protectwise, fireglass, lastline, Securonix, Redowl, Vidder, prelert, utimaco, deepinstinct, Endgame, Cylance, Darktrace, Datavisor, Cyphort, namogoo, Ionic, esentire, illusive, Menlo Security, Skyport Systems
Mobile Security
lookout, MobileIron, Skycure, AppDome, wandera, Mocana, Check Point, nuro, bitglass, InAuth, tiger text, TrustLook, zimperium, Better, appthority, SnoopWall, LaunchKey, Trend Micro, PSafe, Wickr, AVG, Fortinet, Optio Labs, IBM, VKey, Airlight, Kaspersky, Koolspan, TeleSign, OpenPeak, pindrop, BlackBerry, Zimperium, Sophos
Data Security
harvest.ai, Spirion, enSilo, nuro, Thinair, StorageCraft, Vera, dataiphy, Ionic, Wickr, Privitar, globalvelocity, PKWARE, virtru, Code42, CipherCloud, Somansa, Vormetric, BlueTalon, Centri
Identity & Access Management
covisint, Wheel, Nok Nok, Oracle, PingIdentity, Uniken, Cloudway, WISeKey, Bitium, Core Security, Clef, UnboundID, SaferPass, okta, Virgil Security, Forgerock, SailPoint, Trulioo, onelogin, Exostar, SecureKey, Intrinsic ID, IBM, experts, RSA, Saviynt, BeyondTrust, Auth0, foxtechnologies, Duo Security, Device Authority, simeio, welcome, Pirean, IDme, tascent, imprivata, Centrify, ca, verato, SecureAuth, Balabit, Avecto, iLantus, CyberArk, janrain, gemalto, LaunchKey, thycotic
Industrial / IoT Security
Mocana, cryptosoft, Bastille, utimaco, Rubicon, Icon Labs, IMUBIT, riscure, ZingBox, endian, IOActive, Device Authority, Argus, Infineon, Cloudwear, ARM, CyberBit, Cyphre, PFP, Webroot, Tempered Networks, Karamba Security, Bayshore, Indegy, SecuriThings
Fraud Prevention / Transaction Security
FICO, Uniken, feedzai, iovation, ethoca, Forter, Guardian Analytics, ThreatMetrix, Cardinal Commerce, Early Warning, IdenTrust, Au10tix, Signifyd, sift science, Data Security, Socure, riskified, Brighterion, IdentityMind Global, Maxmind, Acculynk, Kount
Cloud Security
Saviynt, CloudPassage, illumio, Skyhigh, iSheriff, bitglass, Bracket, Encrypted Cloud, panda, evident.io, Avanan, ManagedMethods, Soha, zscaler, Vaultive, Code42, WhiteHat, Cloudway, Vera, covata, harvest.ai, Qualys, Microsoft, HyTrust, palerra, Cato, FireLayers, guardtime, Dome9, Twistlock, ClearDATA, Armour, skyhigh, netskope, BetterCloud, Blue Coat, CipherCloud

What is the BDP?

- A cloud-based distributed architecture for ingesting and storing large datasets, building analytics, and visualizing the results.
- Allows critical decisions to be made based on rich and broad data.
- Developed around open source and unclassified components while leveraging community tech transfer from other DoD entities.
- DISA-controlled software baseline
- RMF accredited with current Authority To Operate in multiple organizations
- 99% open source, specifically integrated to meet DoD’s needs



1 Data Sources

Network Security Devices

JRSS
HBSS
Netflow
Proxies

ArcSight
Firewalls
IDS
IPS

Network Management Devices

Hosts Logs
Router Logs
Switch Logs

Application Services

SQL
File Server
Active Directory
ACAS
DNS Logs

Intel/ Operational

NTOC
Incident Handling
OpenPhish

2 Connect

Messaging

Processing

Streaming

3 Transform (Enrich)

Data Fusion

Geospatial

Trends

Persona

4 Store

Structured
Unstructured
Semi-structured
Data

Data Caching

HDFS and Accumulo

5 Analyze

When will it happen?

What happened?

Why did it happen?

How can we prevent it from happening?

Identify Potential Exploit → Develop Hypothesis to Test → Select Relevant Data Sources → Determine Analysis to Perform → Interpret Results → Identify Potential Exploit

6 Visualize/Act

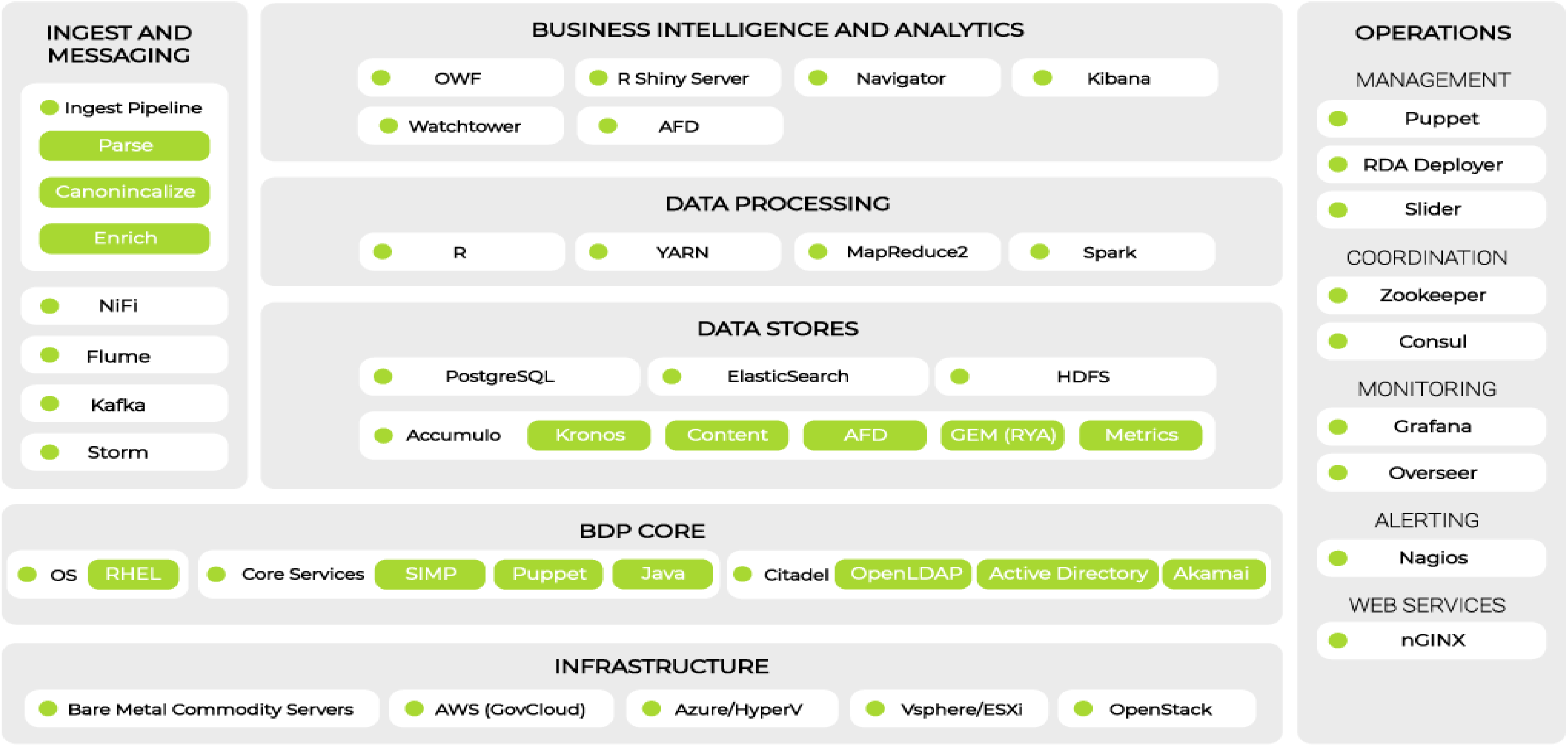
Data Exploration

Unity

Kibana

BIG DATA PLATFORM

Big Data Platform Technology Stack



INGEST AND MESSAGING

- Ingest Pipeline
  - Parse
  - Canonincalize
  - Enrich
- NiFi
- Flume
- Kafka
- Storm

BUSINESS INTELLIGENCE AND ANALYTICS

- OWF
- R Shiny Server
- Navigator
- Kibana
- Watchtower
- AFD

DATA PROCESSING

- R
- YARN
- MapReduce2
- Spark

DATA STORES

- PostgreSQL
- ElasticSearch
- HDFS
- Accumulo: Kronos, Content, AFD, GEM (RYA), Metrics

BDP CORE

- OS: RHEL
- Core Services: SIMP, Puppet, Java
- Citadel: OpenLDAP, Active Directory, Akamai

INFRASTRUCTURE

- Bare Metal Commodity Servers
- AWS (GovCloud)
- Azure/HyperV
- Vsphere/ESXi
- OpenStack

OPERATIONS

MANAGEMENT

- Puppet
- RDA Deployer
- Slider

COORDINATION

- Zookeeper
- Consul

MONITORING

- Grafana
- Overseer

ALERTING

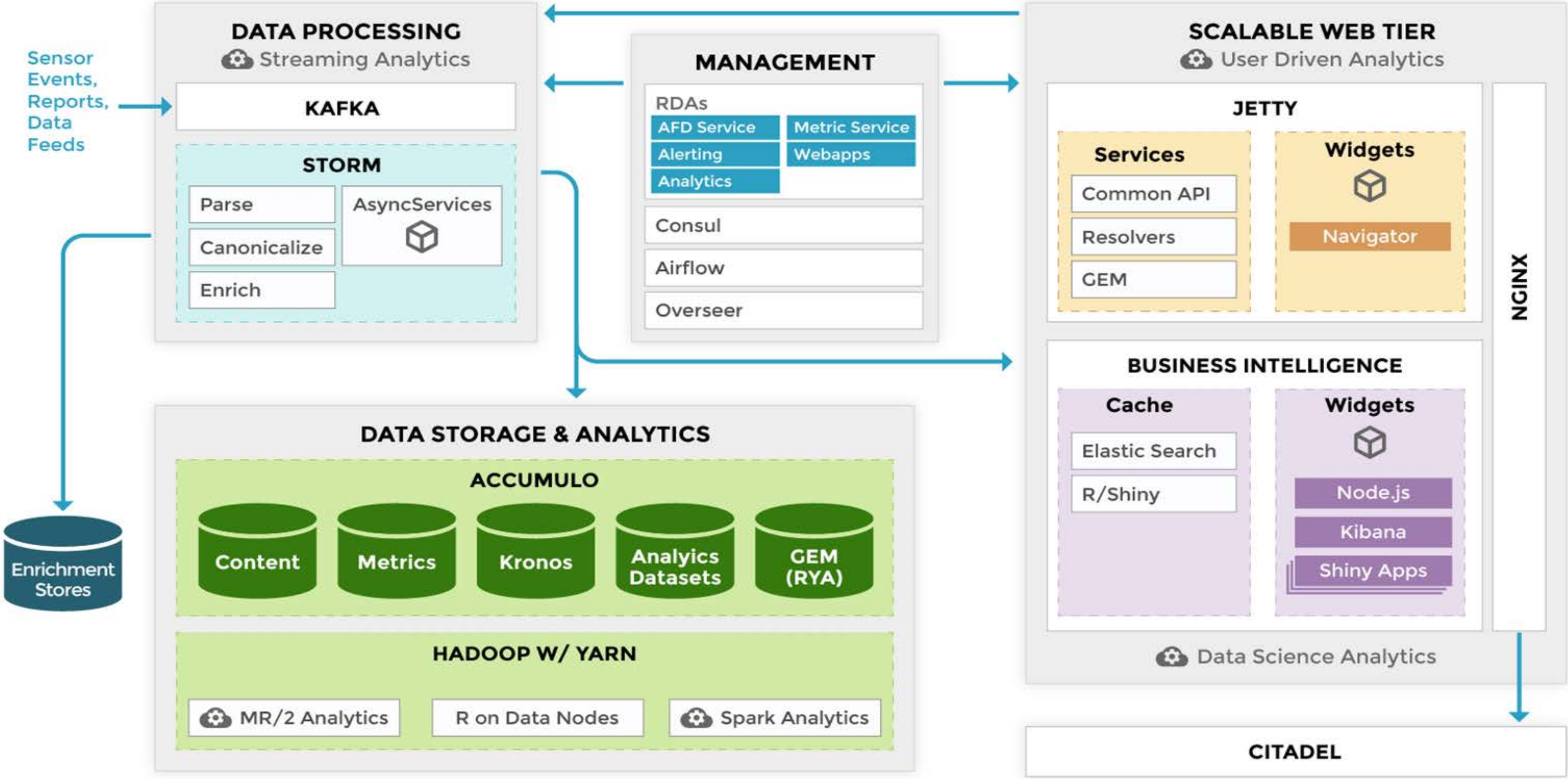
- Nagios

WEB SERVICES

- nGINX

BIG DATA PLATFORM 3.0 ARCHITECTURE

Analytics · Capability Deployments



Sensor Events, Reports, Data Feeds

DATA PROCESSING

Streaming Analytics

KAFKA

STORM

- Parse
- Canonicalize
- Enrich
- AsyncServices

Enrichment Stores

MANAGEMENT

RDAs

- AFD Service
- Metric Service
- Alerting
- Webapps
- Analytics

Consul

Airflow

Overseer

SCALABLE WEB TIER

User Driven Analytics

JETTY

Services

- Common API
- Resolvers
- GEM

Widgets

- Navigator

NGINX

BUSINESS INTELLIGENCE

Cache

- Elastic Search
- R/Shiny

Widgets

- Node.js
- Kibana
- Shiny Apps

Data Science Analytics

CITADEL

DATA STORAGE & ANALYTICS

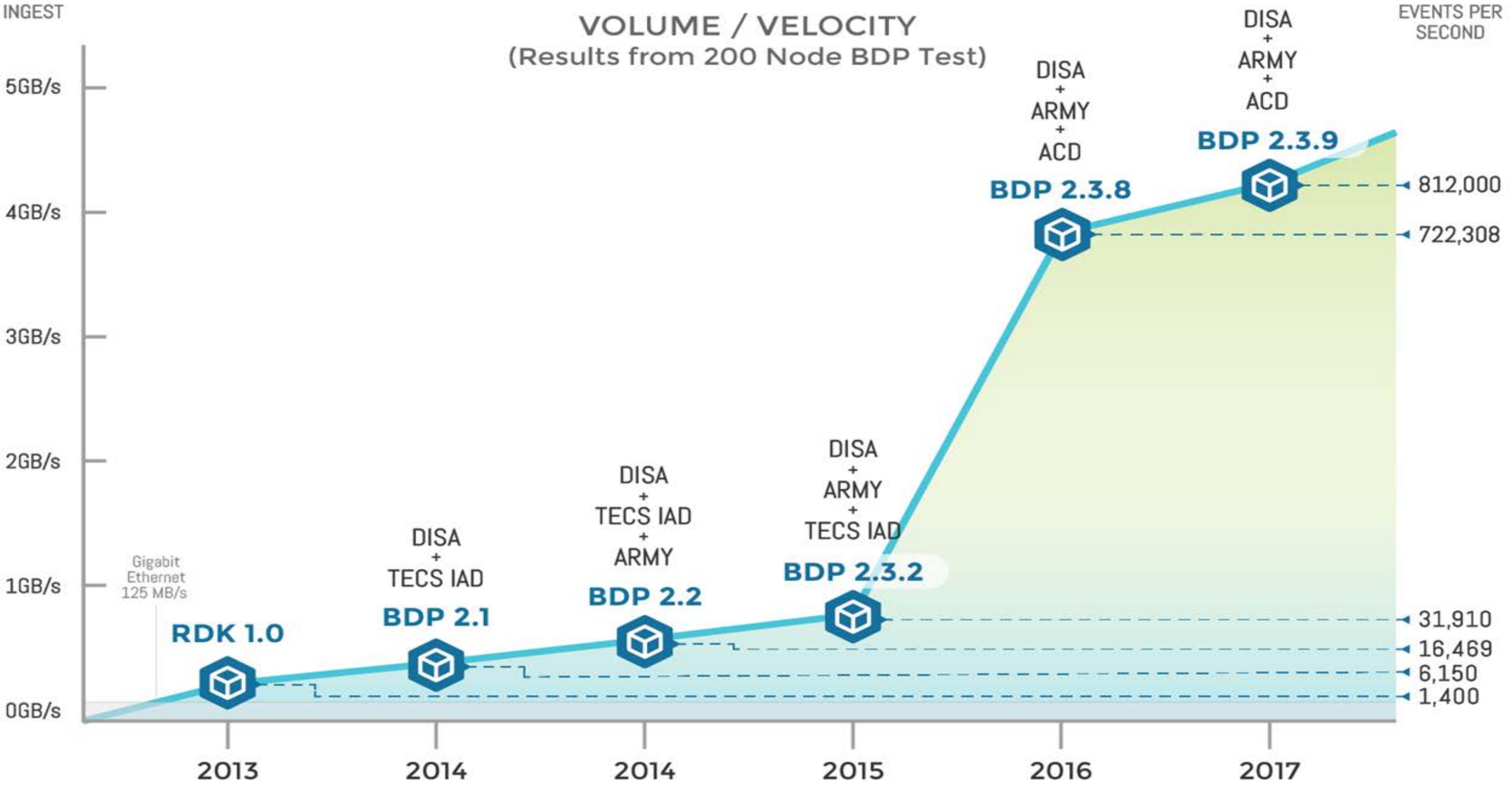
ACCUMULO

- Content
- Metrics
- Kronos
- Analyics Datasets
- GEM (RYA)

HADOOP W/ YARN

- MR/2 Analytics
- R on Data Nodes
- Spark Analytics

Scaling for Volume and Velocity



Multi Tenancy (Learning to share)

- HDFS / Accumulo (Storage)
- Analytics
  - Spark
  - Streaming- Kafka/Storm
  - RShiny
- Web Applications
  - Jetty
  - NodeJS
- Microservices
  - Spring/Java/NodeJS
- Ingest

Lesson Learned:
It’s all about the data

- Don’t underestimate the difficulty of collecting and sharing data
- End user analytic questions have to drive data priorities
- You can’t wait to start collecting data until you need to use it
- *Just enough* normalization will allow unplanned correlations to emerge
- Data from many vantage points increases the value (but analysts need to understand the vantage point of each)

Lesson Learned:
Use commercial cloud infrastructure

- It lets your engineering teams focus on your problems not on infrastructure
- It provides “just in time” capacity that reduces costs in the long run
- It has a refresh rate that is much more frequent than traditional in-house data centers
- It reduces barriers for data transport and acquisition

Lesson Learned:
Standardize your platform early, but evolve it

- Organizations can share security accreditation
- Shared data structures will encourage correlations
- Be willing to change and evolve, without reinventing everything every time
- Create and document APIs that encourage reuse
- Leverage a community to share costs

Lesson Learned:
Analytics need to scale

- Need to run on commodity hardware (if you can fit all your data into memory, you don't have big data)
- Need to be parallelizable
- Need to handle preemption (half your job may be killed at any moment to make way for higher priority tasks)
- Need to be secure (can't open ports, store passwords; need to handle data security controls)

Lesson Learned:
You need to optimize your load

- Use batch ingest
- Cache data near the web tier
- Adjust the allocation of resources to your mission (YARN is great, but it needs to be managed)
- Test with real world datasets (size and variety)
- Understand the computational costs of your analytics before deploying them

Questions?