

RED HAT
SUMMIT

BOSTON, MA
JUNE 23-26, 2015

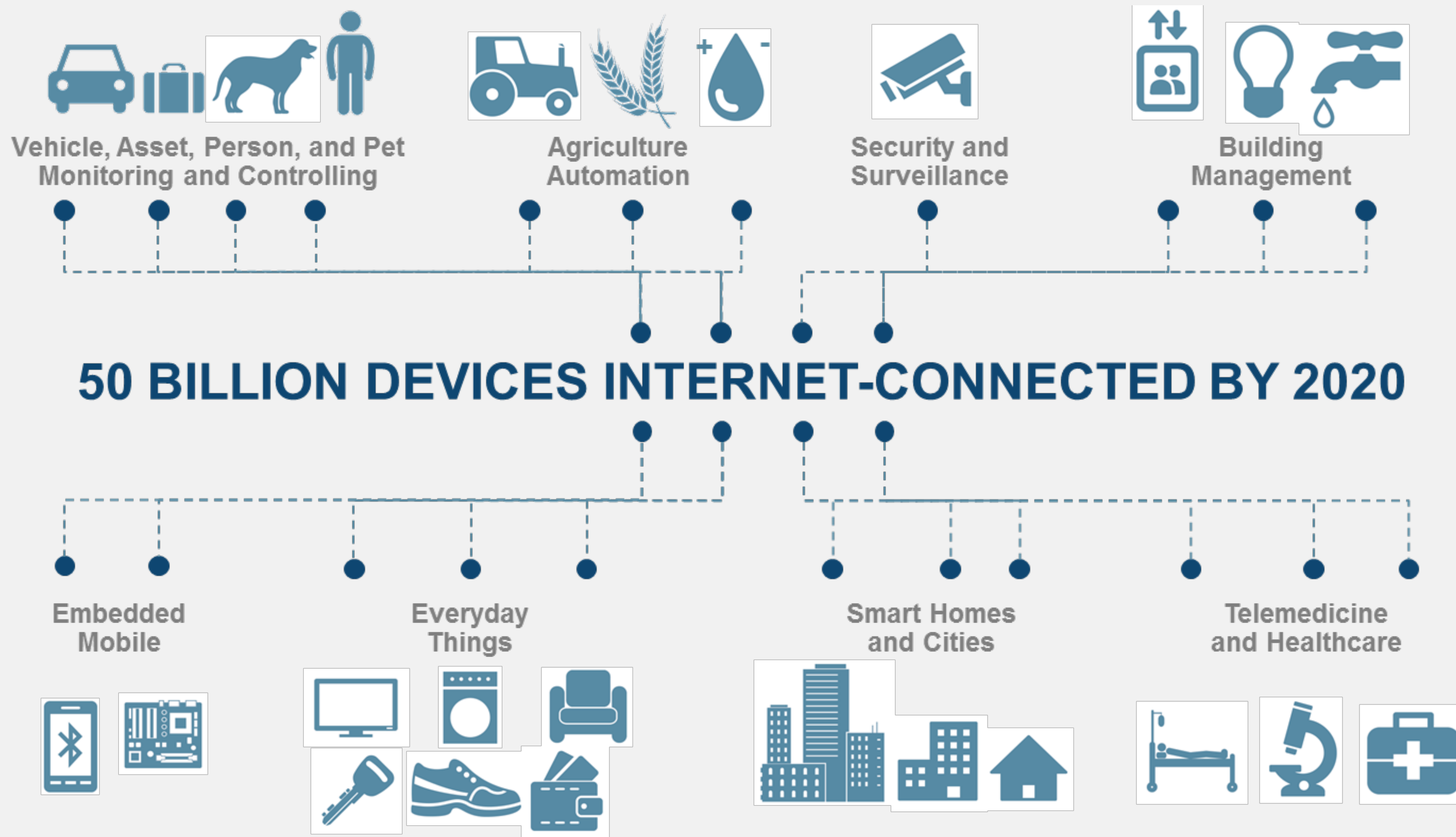
BIG DATA IN THE OPEN PRIVATE CLOUD

CSC BIG DATA PAAS + RED HAT ENTERPRISE LINUX OPENSTACK PLATFORM

Tim Gasper
Global Offerings Manager
CSC Big Data & Analytics
6/25/2015

BIG DATA IS NOT BUSINESS AS USUAL

Machine Data, Sensors, and the Internet of Things



Busting Corporate Data Silos



MetLife

The Wall

Life (3) ▾

Policy #

xxxxxxx5456

Customer Service Phone Number

800-638-5433

Lucy Merryweather

Owner

SSN

xxx-xx-xxxx

Phone

(819) 555-7702

DOB

August 10, 1973

Alternate Phone

(819) 555-2231

Address

109 West 93rd Avenue
Lincoln, OK 07882

Email

lkerry@gmail.com

Product Type

Life

Contract Status

Active

Contract Type

Whole Life

Franchise Name

New England Financial

Group Name

[No Data]

Group Number

[No Data]

Sales Agent

Tommy Topseller

Agent ID

99B 560

TCA (1) ▾

Confidence Level

High

Medium

Low

Close Transaction Details

Transaction Details

Last 60 days • 2 transactions

Self Service Address Change

Status: Completed

Source: Inbound

Channel: eService

System: BOSSLA

Received: January 20, 2012 @ 10:00 am

View Documents (2)

Change Confirmation (Mailed) • January 24, 2012 @ 4:31 pm

Change Request • January 20, 2012 @ 9:50 am

Self Service Bene Change

Status: Completed

Source: Inbound

Channel: eService

System: BOSSLA

Received: June 1, 2011 @ 11:30 am

View Documents (2)

Change Confirmation (Mailed) • June 3, 2011 @ 6:03 pm

Change Request • June 1, 2011 @ 11:20 am

Source: <https://gigaom.com/2013/05/07/with-300m-earmarked-for-tech-innovation-metlife-wants-to-remake-insurance/>

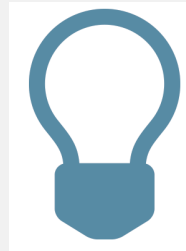
#redhat #rhsummit

The Power of Predictive Analytics



Financial Services

- Fraud detection
- Risk management
- 360° view of the customer



Utilities

- Analysis of weather impact on power generation
- Transmission monitoring
- Smart grid management



Transportation

- Real-time route optimization based on traffic and weather
- Maintenance optimization and asset tracking



Health and Life Sciences

- Epidemic early warning system
- ICU monitoring
- Remote healthcare monitoring



Retail

- 360° view of the customer
- Click-stream analysis
- Real-time promotions



Telecommunications

- CDR processing
- Churn prediction
- Geomapping/marketing
- Network monitoring



Law Enforcement

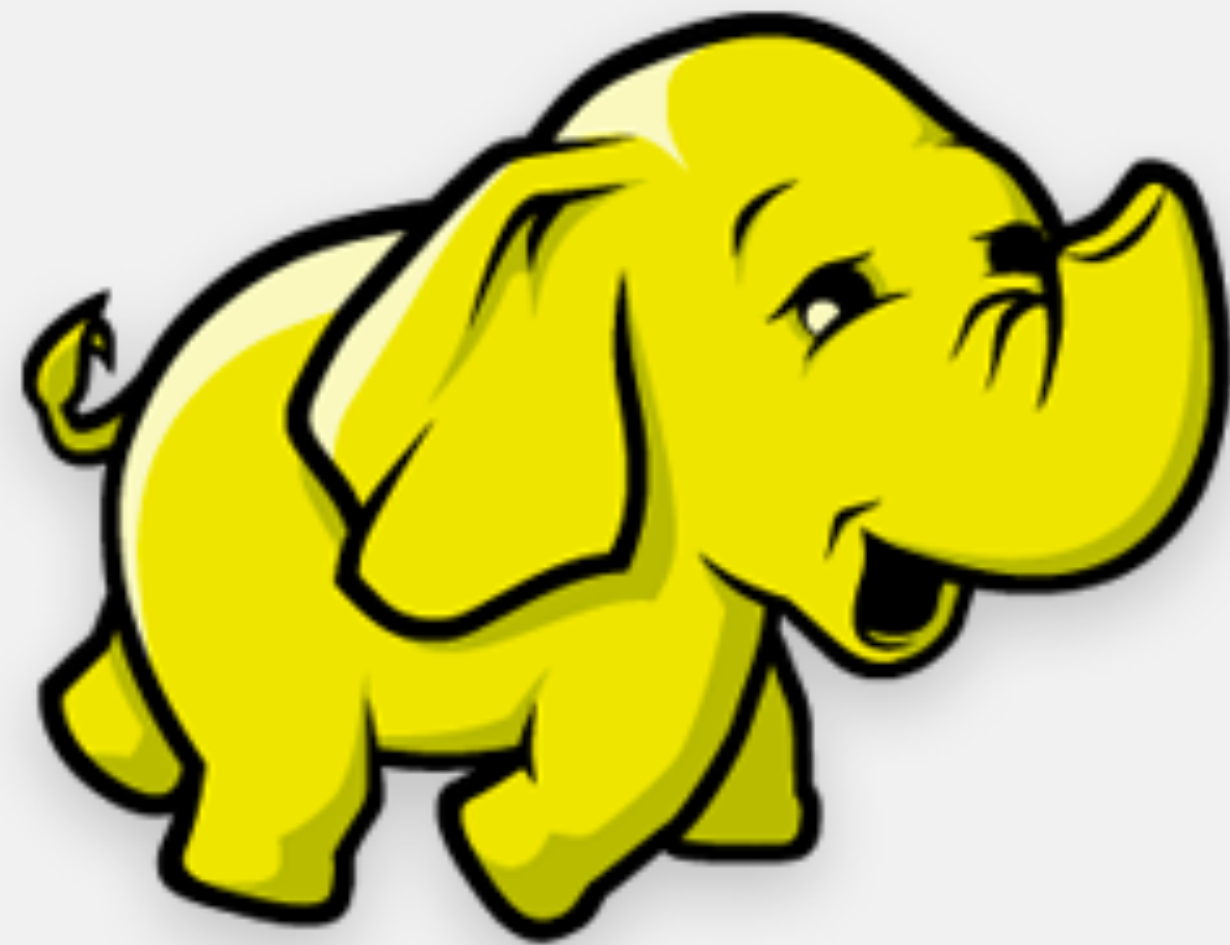
- Real-time multimodal surveillance
- Situational awareness
- Cybersecurity detection



Manufacturing

- Predictive maintenance
- Real-time parts flow monitoring
- Product configuration planning

Hadoop: The Solution?



- Massive-scale, batch oriented analytical processing
- Ultimate scale-out flexibility to add compute and storage
- Wide variety of processing frameworks, including Spark, Search, Hive, HBase... support for Java, Python, R, etc.
- Robust add-on ecosystem, including Pentaho, Datameer, H2O, Waterline Data, etc.
- Impala and Hive provide interactive data access, but optimized for medium to high latency
- High upkeep for managing table structures and metadata
- 3+ replication of data
- Ideal DR is a second cluster with data replication meaning syncing concerns and 6+ total replication of data
- Minimal support for data transactionality
- High maintenance and management overhead

Often an important foundation and hub, but not the analytics silver bullet.

Right Tool for the Job, Often Multiple Tools for the Job



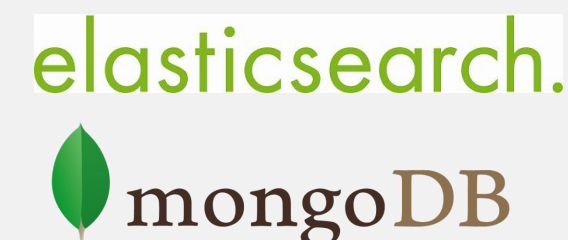
Analytics File Store

- Scale out to petabytes
- Foundation for many data processing frameworks
- Any data format
- Interactive query w/ Hive and Impala



Row-Oriented Database

- ANSI SQL compliance
- Transactional data (OLTP)
- Frequent complex joins
- Frequent row updates
- Fast, small data business intelligence support



Document Store

- Real-time data visualization
- Fast, flexible app development
- JSON/BSON data formats



Full-Text Search

- Text searches (e.g. social media, email / chat, claims, contracts, etc.)
- Data exploration, discovery, and indexing



Columnar Database

- Scale out to petabytes
- Fast time-series analysis (e.g. web activity, sensor readings, purchases, etc.)
- Fast characteristic or numerical analysis (age, ratings, churn, etc.)



In-Memory Analytics

- Fast analytical operations on data that fits in memory
- Rapid data mining
- Real-time apps with fast ad hoc analytic requirements



Advanced Analytics Tool

- Tools for executing advanced analytical techniques (statistics, forecasting, simulations, clustering, etc.)
- Can run standalone or in conjunction with in-memory analytics, columnar DB, analytics file store, etc.



Business Intelligence Tool

- Tools for combining metrics and visualizing them through reports, dashboards, and interactive visualizations
- Can run standalone or in conjunction with in-memory analytics, columnar DB, row DB, analytics file store, etc.



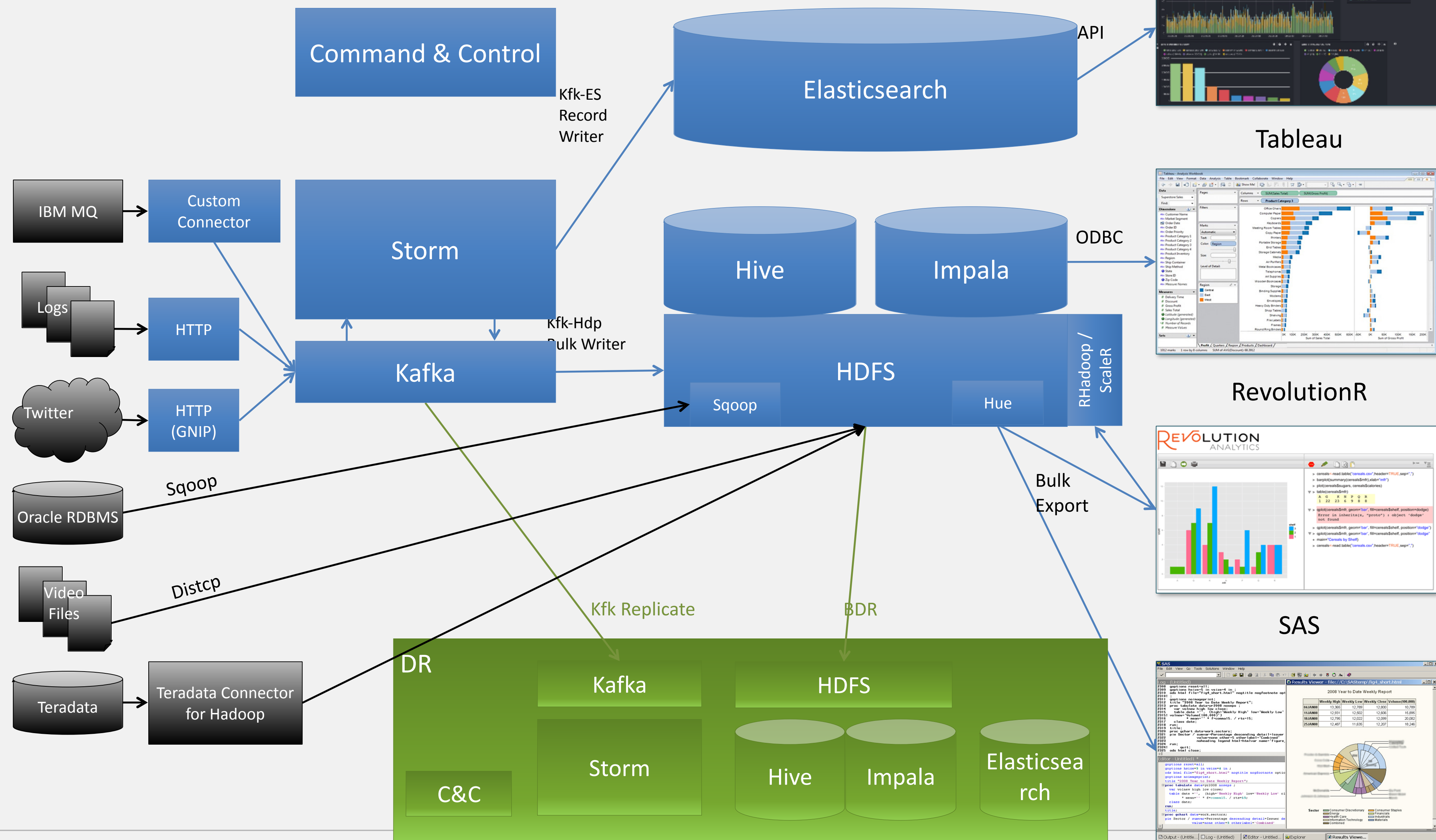
Data Integration Tool

- Tools for connecting with databases, file stores, or other systems and transferring data
- Can run standalone or in conjunction with in-memory analytics, stream processing, analytics file store, etc.



Stream Processing

- Processing event data that is constantly flowing (e.g. sensor data, web log data, user or customer activity, etc.)
- Augmentation, cleansing, transformation, and monitoring of data as it flows to other destinations



In This Landscape, Companies Consistently Struggle with the Same Four Challenges

1. Setting up and operating a big data and analytics platform

- Data complexity
- Operational complexity
- Managing varied production workloads to tight SLAs

2. Attracting, managing, and applying big data & analytics skills

3. Integrating insights into their business processes

4. Iterating quickly enough

- Limited infrastructure scalability with hard resource boundaries
- Optimized for IT purchase cycle and industry performance benchmarks, not for business ROI

Overcoming These Challenges Requires a New Approach

Complex Data

Provide application developers a platform designed to ingest, integrate, and manage data from any source and in any format

Skills Shortage

Leverage Big Data Platform as a Service to allow your talent to focus on the business application versus platform gymnastics

Scale and Speed

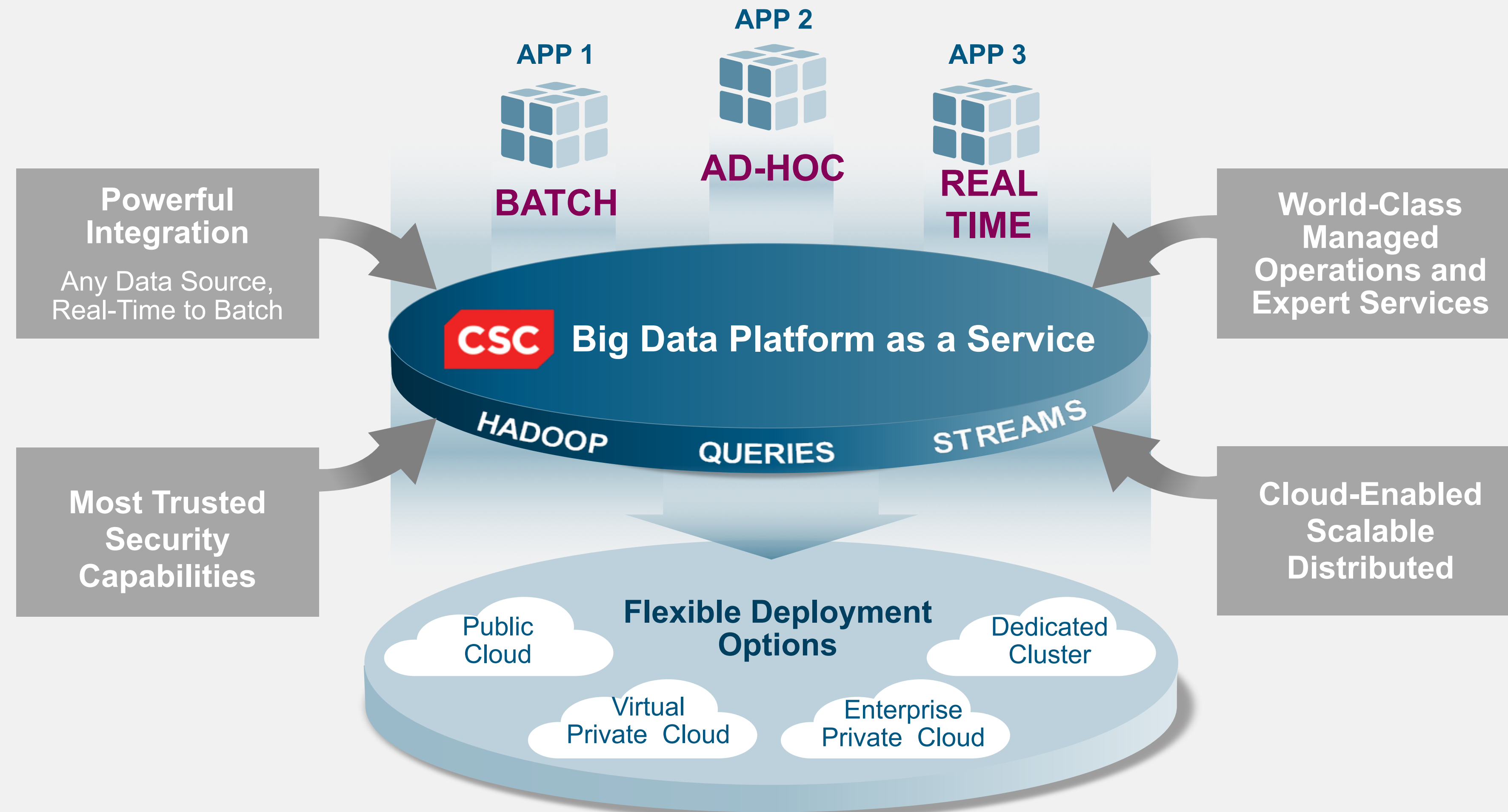
Benefit from the technologies, proven scale, and **cloud operating principles** that underpin Facebook, Twitter, Yahoo, and other businesses that rely on big data

ROI Expectations

Access an environment that can be deployed and operational in days, not months -- and changed in minutes, not weeks

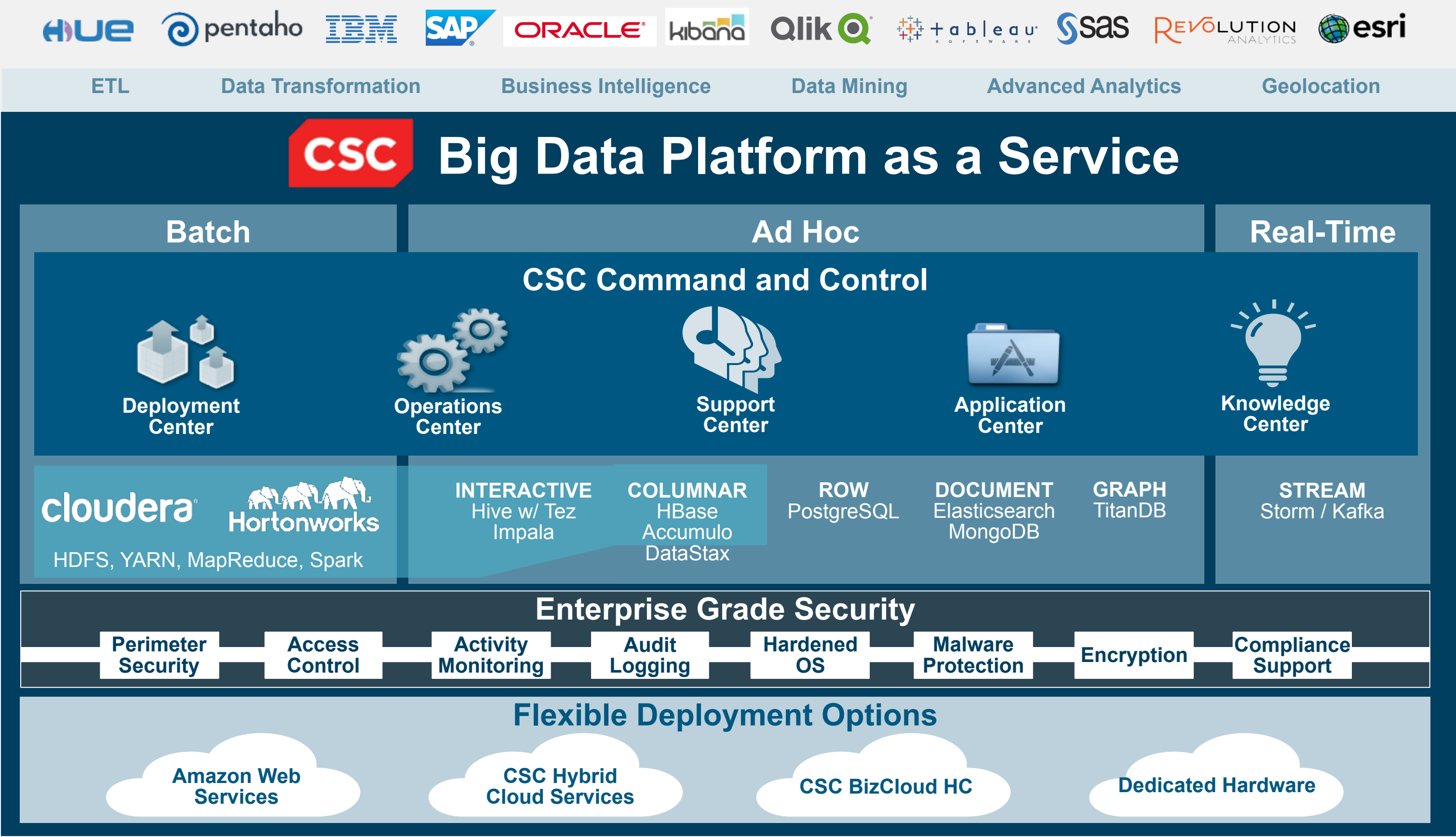
CSC BIG DATA PAAS

Big Data Platform That Enables Insights from Your Data in Less Than 30 Days



Agile Application Development Environment that is Scalable, Sustaining, Self Healing

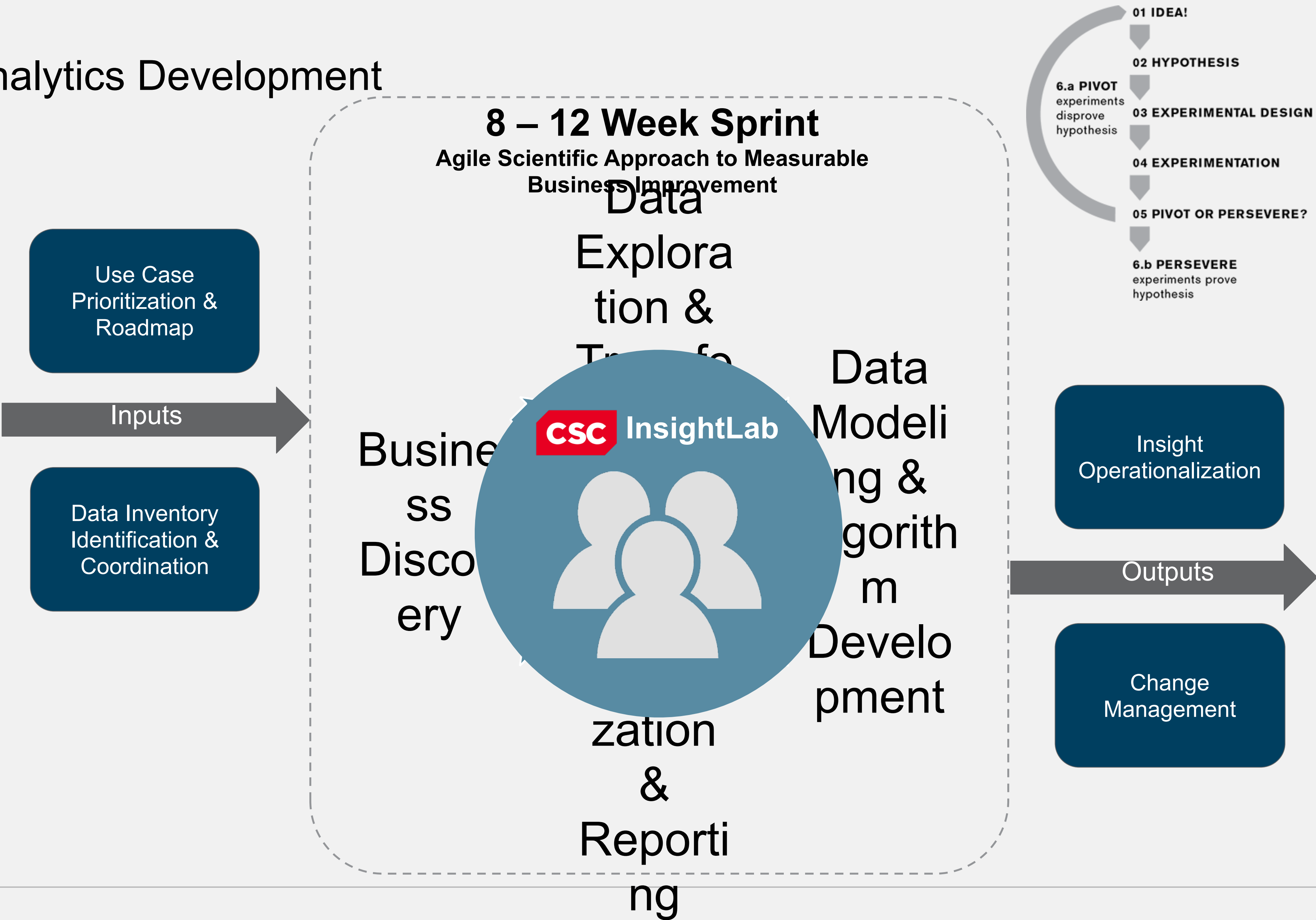
Fully Integrated and Managed Big Data Platform-as-a-Service



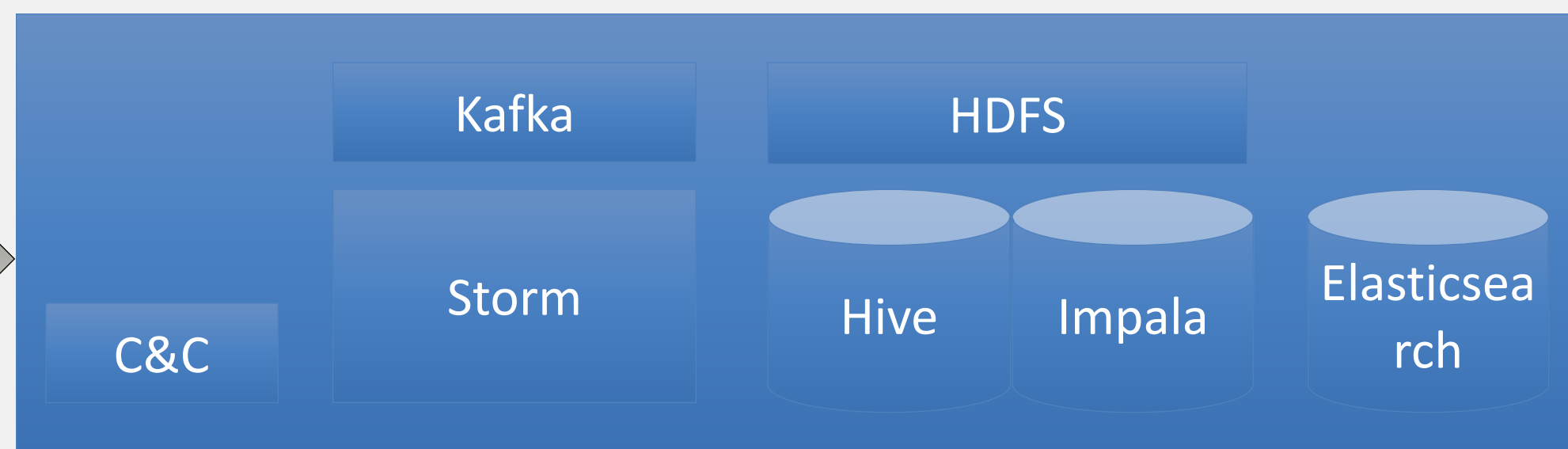
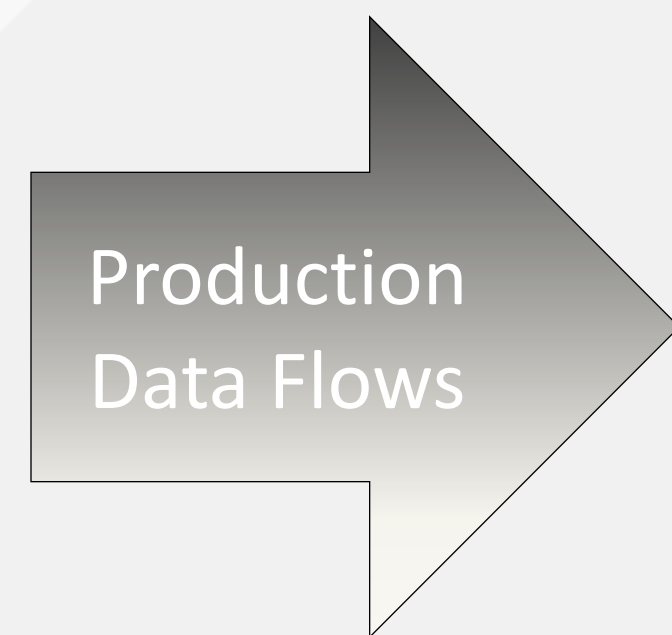
Making it faster, easier, and far less costly to develop and deploy big data applications

AN AGILE BIG DATA PLATFORM APPROACH

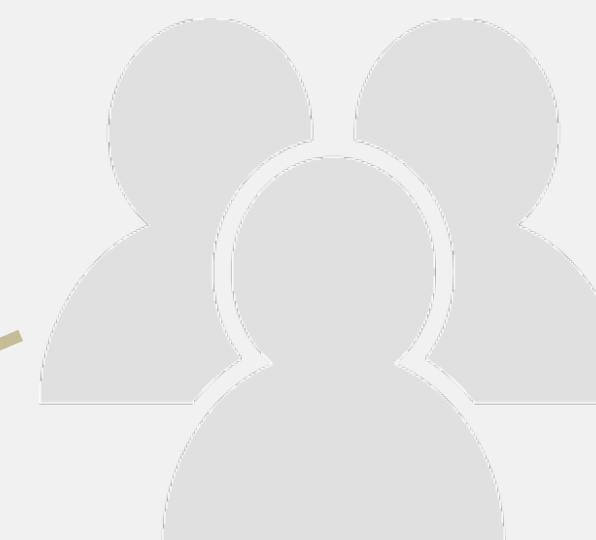
Rapid Analytics Development



PRODUCTION



User
Queries



LOCAL



VM/Sandbox

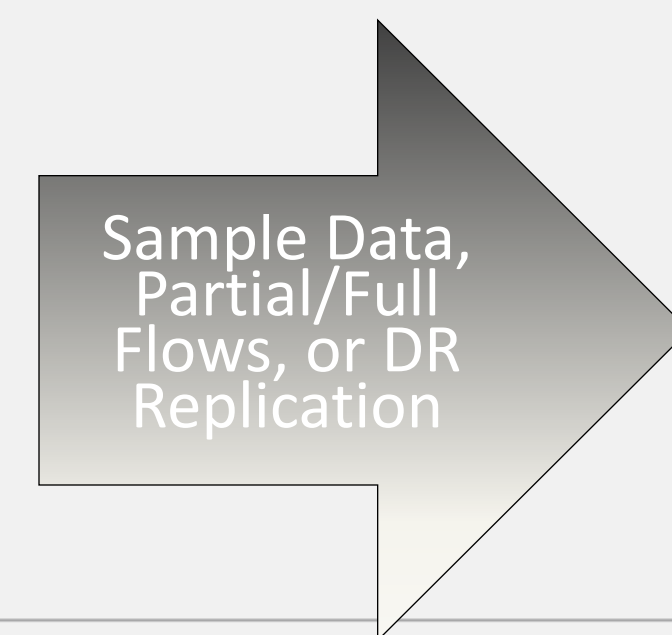
or "local node" environment
or "direct-dev" on BDPaaS

Maintenance Window
Push to Production

Push Code
w/ Git

Test w/
Jenkins

DEV / DR

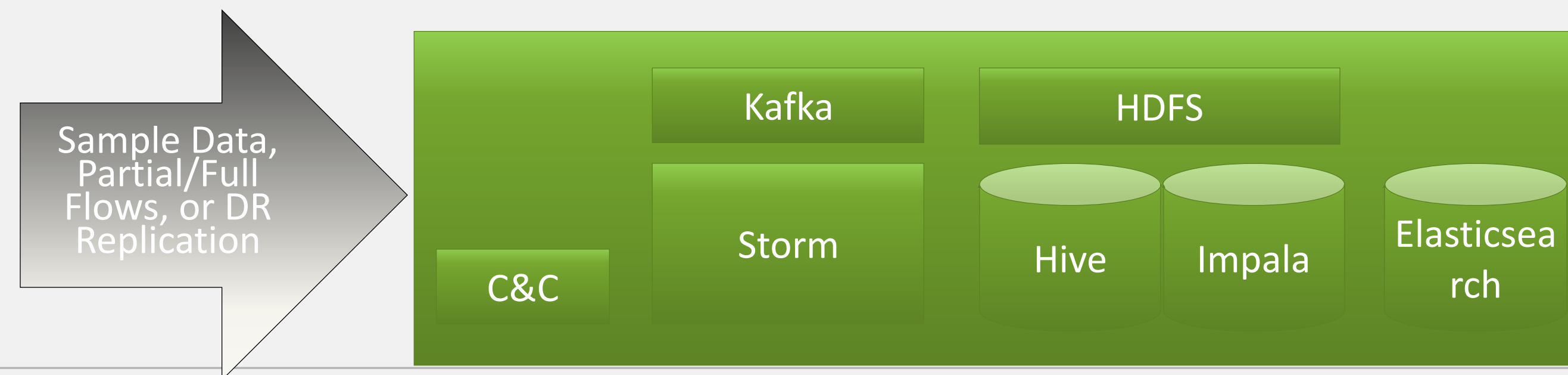


PRODUCTION



- ADD OR REMOVE NODES
- RECONFIGURE NODES
- RECONFIGURE OVERALL CLUSTER
- ADD OR REMOVE CLUSTERS
- SCALE UP OR SCALE DOWN CPU, RAM, DISK
- ADD OR REMOVE ENVIRONMENTS

DEV / DR



- ADD OR REMOVE NODES
- RECONFIGURE NODES
- RECONFIGURE OVERALL CLUSTER
- ADD OR REMOVE CLUSTERS
- SCALE UP OR SCALE DOWN CPU, RAM, DISK
- ADD OR REMOVE ENVIRONMENTS

VPN

IAM

CloudFront

AMI Service

Kibana

Command & Control

Elasticsearch

API

CloudFormation

Tableau

IBM MQ

Custom Connector

Storm

C4-Instances

RDS

I2-Instances

Hive

Impala

ODBC

M3-Instances

RevolutionR

Storage Gateway

Logs

HTTP

Twitter

HTTP (GNIP)

Kafka

Kfk-Hdp Bulk Writer

Sqoop

HDFS

Hue

RHadoop / Scaler

Oracle RDBMS

Sqoop

C3-Instances

D2-Instances

Bulk Export

Direct Connect

Video Files

Distcp

EBS Volumes

R3-Instances

Teradata

Teradata Connector for Hadoop

S3

DR

Glacier

Kafka

Storm

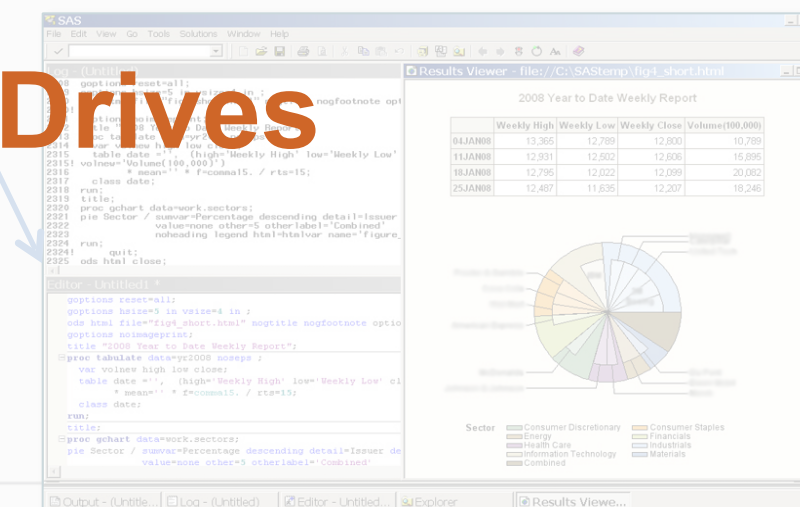
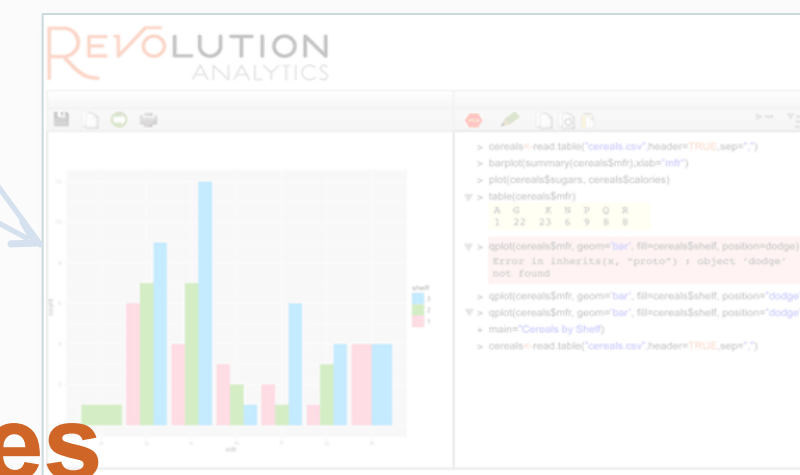
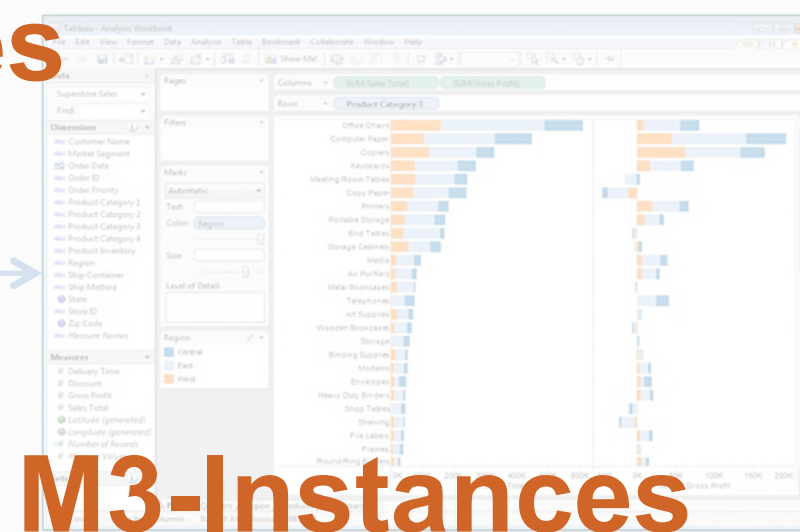
Hive

Impala

Elasticsearch

Ephemeral Local Drives

SAS



Why Private Cloud

Private

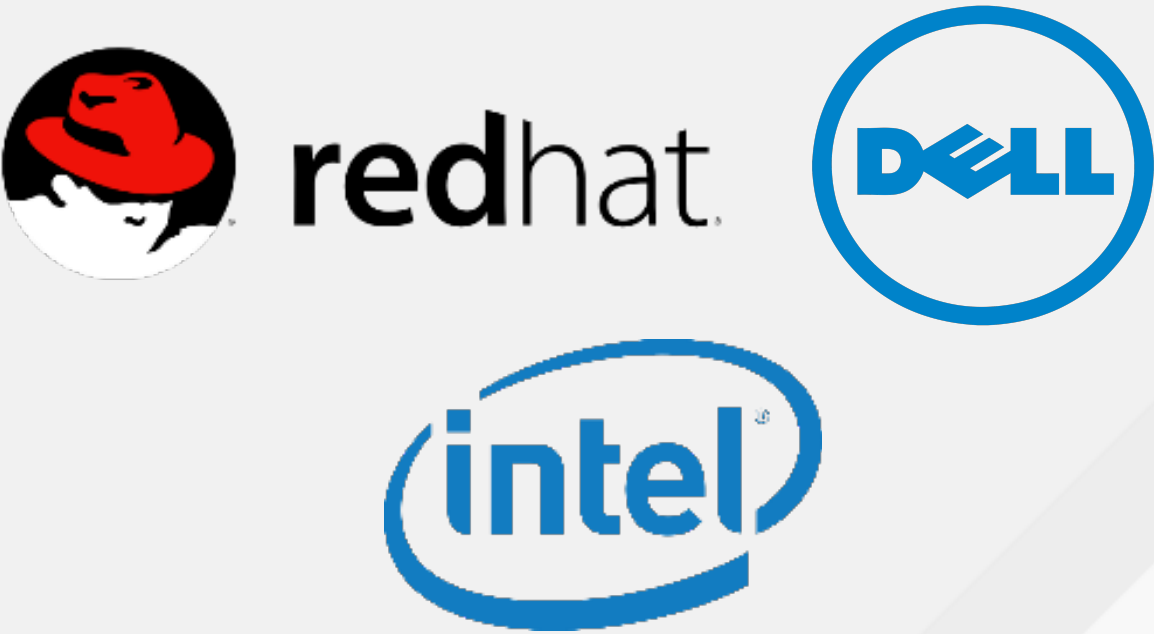
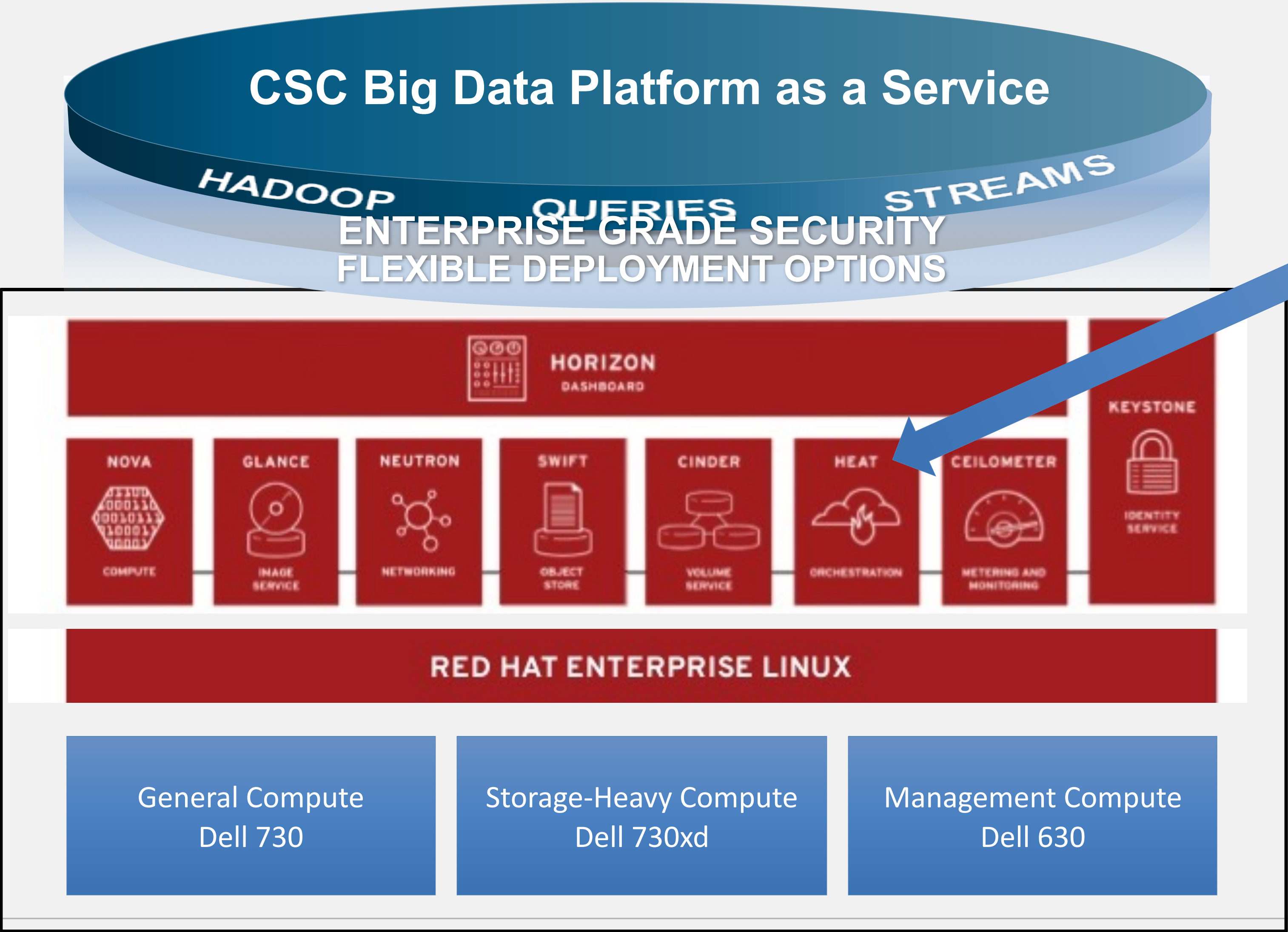
- Security & Compliance
- Ownership
- Internal Network Benefits
- High Scale Cost Savings

Virtualized w/ Cloud Management

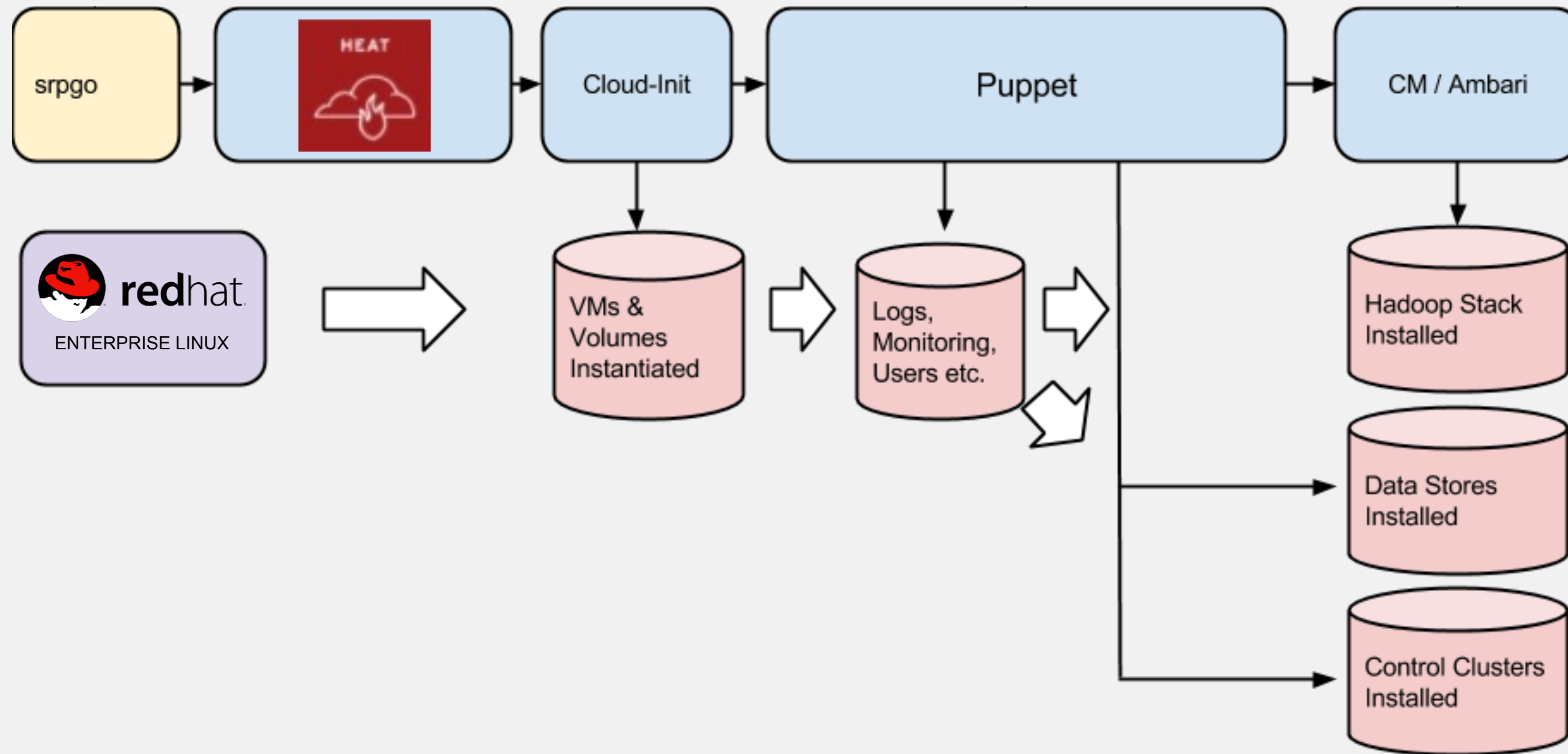
- Higher Resource Efficiency for Increase Savings
- Significantly Greater Workload and Resource Flexibility
- More compatible with software-defined-everything approach
- Shared Services (image service, identity management, object storage, block storage, telemetry, etc.)
- High Scale Cost Efficiency
- Hybrid Cloud Compatibility

CSC BIG DATA PAAS POWERED BY RED HAT ENTERPRISE LINUX OPENSTACK PLATFORM

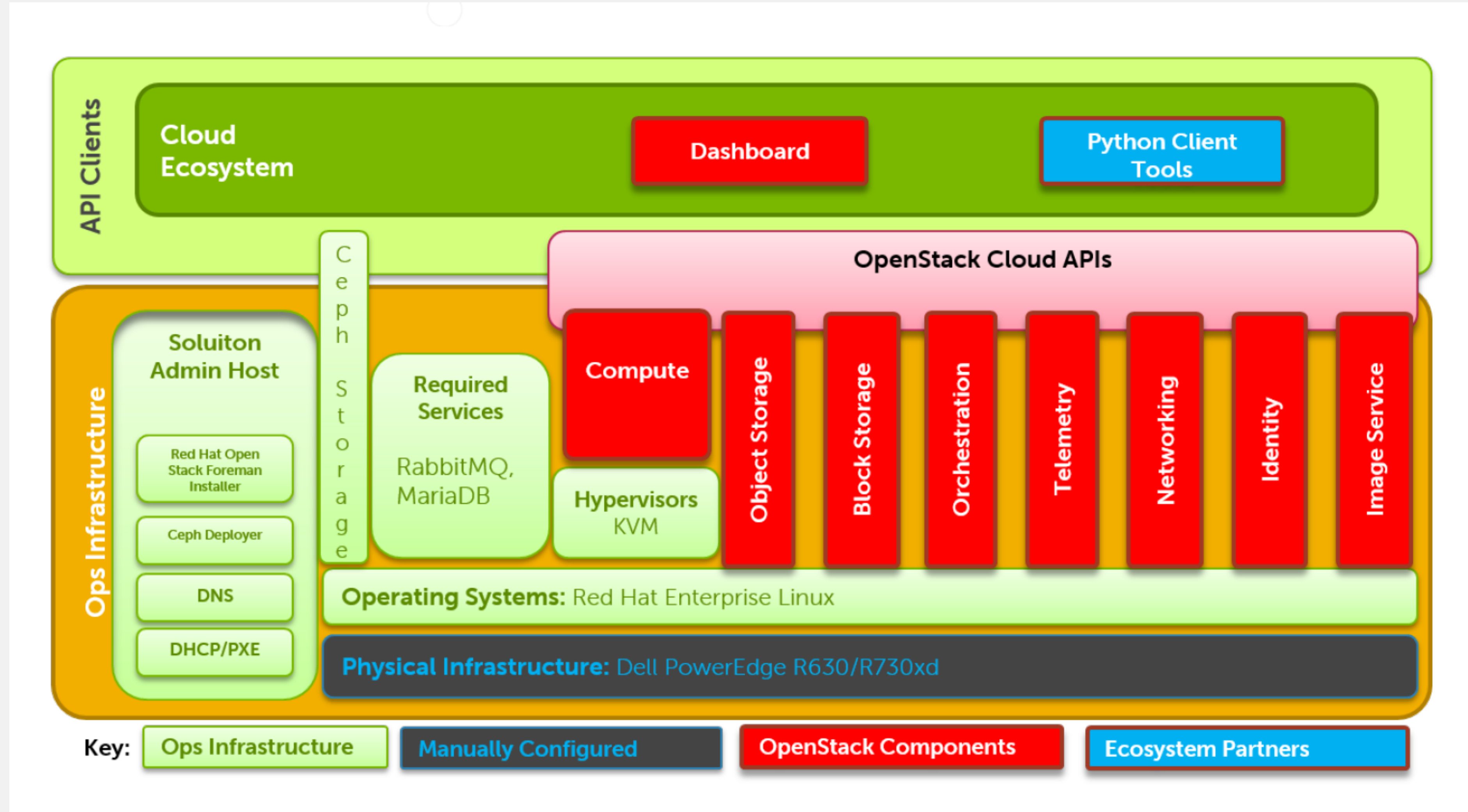
BDPaaS on Red Hat OpenStack Platform

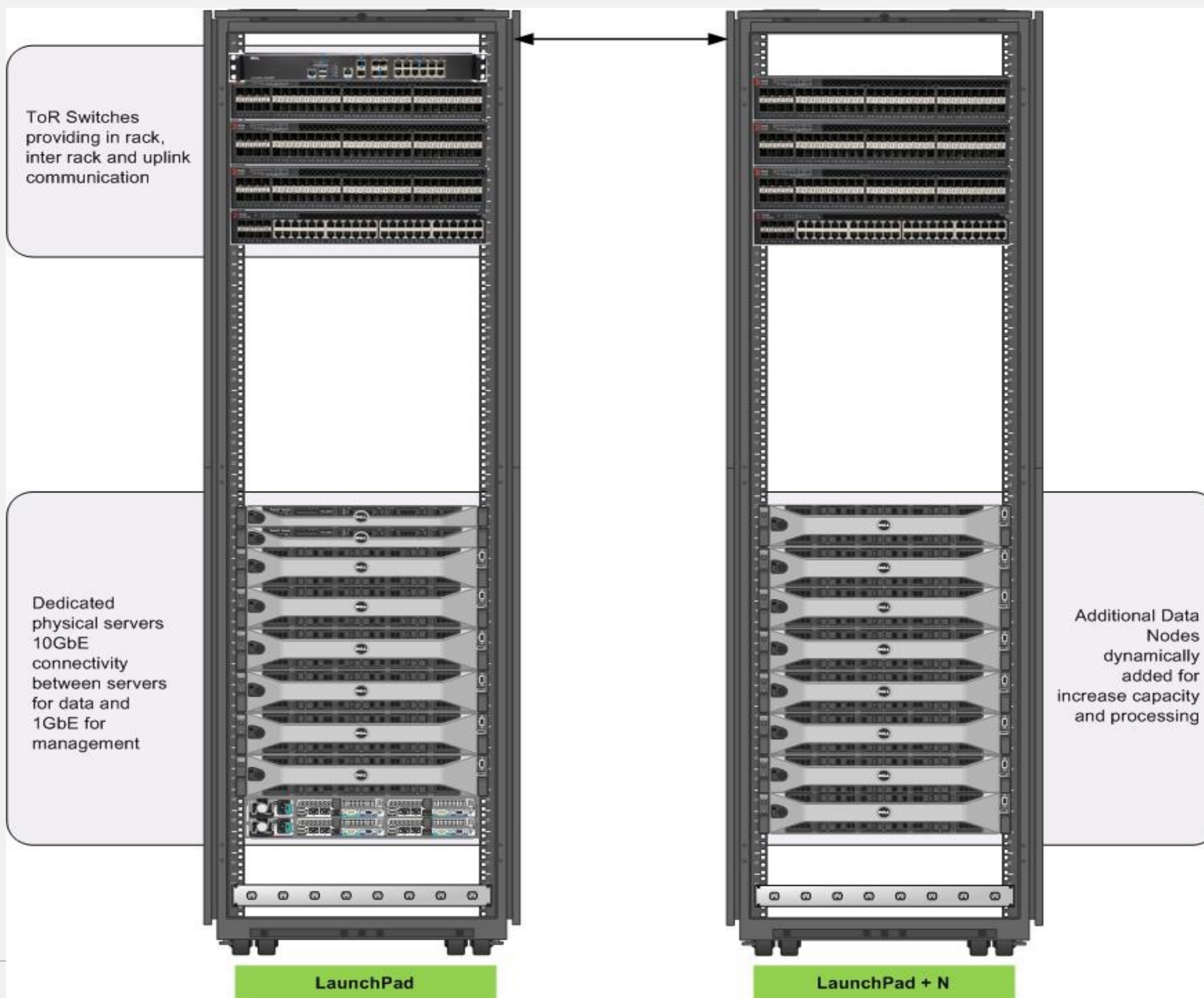


BDPaaS DevOps Automation



Red Hat Enterprise Linux OpenStack Platform + Dell Reference Architecture





LaunchPad

U	Rack1
42	Firewall
41	Cable Mgmt
40	Data Switch 1
39	Cable Mgmt
38	Data Switch 2
37	Cable Mgmt
36	Access Switch
35	Cable Mgmt
34	Management Switch
33	
32	
31	
30	
29	
28	
27	
26	
25	
24	
23	R620 Network Server
22	R620 Admin Server
21	R620 Admin Server
20	R720 General Purpose Compute
19	
18	R720 General Purpose Compute
17	
16	R720 General Purpose Compute
15	
14	R720XD Compute/Storage Server
13	
12	R720XD Compute/Storage Server
11	
10	R720XD Compute/Storage Server
9	
8	R720XD Compute/Storage Server
7	
6	R720XD Compute/Storage Server
5	
4	R720XD Compute/Storage Server
3	
2	Power Cabling
1	Power Cabling

Component	Main Stack	Mini Stack	POC Stack
Workload	Full production deployment, client dedicated, cluster separation by physical separation of workload, i.e Production Cluster would have its own Data nodes etc	Small production, limited sizing, will deploy a dedicated stack into the client's network, Clusters would logically segregated from each other if required.	Non- Production - deployed for POC engagement, prove and test data etc.
Min usable storage	42TB+	10- 28TB (TBC)	1 - 3TB (TBC)
Firewall	Cisco ASA 5525-X	N/A - Assumed will connect into clients existing network	N/A - Assumed will connect into clients existing network
Network Aggregation	Brocade VDX 8770 (Optional if client does not have a suitable aggregation layer)	N/A - Assumed will connect into clients existing network	N/A - Assumed will connect into clients existing network
Switch – Data	2 x Brocade ICX 6650 (10GbE)	1 x Brocade ICX 6650 (10GbE)	N/A - Assumed will connect into clients existing network
Switch – Management	1 x Brocade ICX 6610 (1GbE)	1 x Brocade ICX 6610 (1GbE)	1 x Brocade ICX 6610 (1GbE) (covers all connectivity)
Purpose	General Purpose Compute Server	General Purpose + Data Node Compute Server	General Purpose + Data Nodes Compute Server
Virtualization	Yes - OpenStack (Multi workload)	Yes - OpenStack (Segregation of all applications and data nodes / tenants would be segregated as well)	Yes - OpenStack
Applications Hosted	control (dashboard, mongodb, vcd, edge, splunk...), streaming (storm,kafka,listeners), queries (elasticsearch, maybe some hbase nodes), hadoop control (namenode, jobtracker), any aux instances (tableaux et al)	control (dashboard, mongodb, vcd, edge, splunk...), streaming (storm,kafka,listeners), queries (elasticsearch, hbase), hadoop (namenode, jobtracker, datanode/tasktracker), any aux instances (tableaux et al)	control (dashboard, mongodb, vcd, edge, splunk..), streaming (storm,kafka,listeners), queries (elasticsearch, hbase), hadoop (namenode, jobtracker, datanode/tasktracker), any aux instances (tableaux et al)
Minimum	3	2	1
Maximum	TBC	2 (grow Data Compute Servers for more capacity)	1
Provider	Dell	Dell	Dell
Model	R730	R730	R730
CPU	Dual 18 Core CPU Intel® Xeon® E5-2695 v2 2.40GHz	Dual 18 Core CPU Intel® Xeon® E5-2695 v2 2.40GHz	Dual 18 Core CPU Intel® Xeon® E5-2695 v2 2.40GHz
RAM	192GB	192GB	192GB
Network	2 * 10Gb Nics / 2 * 1GB Nics	2 * 10Gb Nics / 2 * 1GB Nics	2 * 1GB Nics
Storage	16TB (8x2TB,3.5",SATA,7.2k Near line SAS)	16TB (8x2TB,3.5",SATA,7.2k Near line SAS)	16TB (8x2TB,3.5",SATA,7.2k Near line SAS)
Data Node Sizing	NA	2nd Chassis Handles Data Nodes = 48/4.5 = 10TB Usable	NA
Purpose	Data Compute Server	Data Compute Server	Data Compute Server
Virtualization	Yes - OpenStack 1:1 Ratio	Yes - OpenStack - Logically separate customers	NA – Covered in General Purpose + Data Nodes Compute Server
Applications Hosted	hadoop datanode/tasktrackers, perhaps some hbase components	hadoop datanode/tasktrackers, perhaps some hbase components	
Minimum	6 Data Nodes (50TB * 6 = 300TB)	2 Data Nodes (50TB * 2 = 100TB)	
Maximum	Add Data Nodes	4 Data Nodes (50TB * 4 = 200TB)	
Raw to Usable Ratio	4.5:1	4.5:1	
Provider	Dell	Dell	
Model	R730xd	R730xd	
CPU	Dual 8 Core CPU	Dual / 8 Core CPU	
RAM	128GB	128GB	
Network	2 * 10Gb Ncs / 2 * 1GB Nics	2 * 10Gb Ncs / 2 * 1GB Nics	
Storage	50TB (12x4TB,3.5",SATA,7.2k +2x 1TB 2.5” Near Line SAS)	50TB (12x4TB,3.5",SATA,7.2k +2x 1TB 2.5” Near Line SAS)	
Purpose	Management Server (Network / Admin)	Management Server (Network / Admin)	Management Server (Network / Admin)
Virtualization	NA	NA	NA
Applications Hosted	OpenStack control: OpenStack-api, OpenStack-dashboard, glance (api, registry), keystone, nova (api, cert, conductor, console-auth, novncproxy, scheduler), openvswitch, neutron (server,metadata,l3,dhcp), and potentially swift (proxy, account, container, object)	OpenStack control: OpenStack-api, OpenStack-dashboard, glance (api, registry), keystone, nova (api, cert, conductor, console-auth, novncproxy, scheduler), openvswitch, neutron (server,metadata,l3,dhcp), and potentially swift (proxy, account, container, object)	OpenStack control: OpenStack-api, OpenStack-dashboard, glance (api, registry), keystone, nova (api, cert, conductor, console-auth, novncproxy, scheduler), openvswitch, neutron (server,metadata,l3,dhcp), and potentially swift (proxy, account, container, object)
Minimum	3	2	2
Maximum	Add Based on requirements and HA option	2	2
Provider	Dell	Dell	Dell
Model	R630	R630	R630
CPU	Dual 8 Core CPU	Dual 8 Core CPU	Dual 8 Core CPU
RAM	128 GB	128 GB	128 GB
Network	2 * 10Gb Ncs / 2 * 1GB Nics	2 * 10Gb Ncs / 2 * 1GB Nics	2 * 10Gb Ncs / 2 * 1GB Nics
Storage	9.6TB RAW 10K RPM SAS 6Gbps 2.5in	9.6TB RAW 10K RPM SAS 6Gbps 2.5in	9.6TB RAW 10K RPM SAS 6Gbps 2.5in

Usage-Based Auto Insurance

Big Data Telematics in the Open Private Cloud

Global Insurance Company

A leading mutual insurance company with more than 18,000 agents servicing 81 million customers in the US and Canada. **CSC delivered a cost-effective, big data platform to support high volume telematics data. Solution accelerated time to market and met critical business timelines.**

Challenge

- Highly visible board level initiative to close gap with other carriers already offering products in market
- Required a platform to support a high volume telematics based analytic application
- Ability to meet time to market requirements of the business
- Internal technology skills gap
- Current technology platforms were neither capable of supporting, nor cost effective

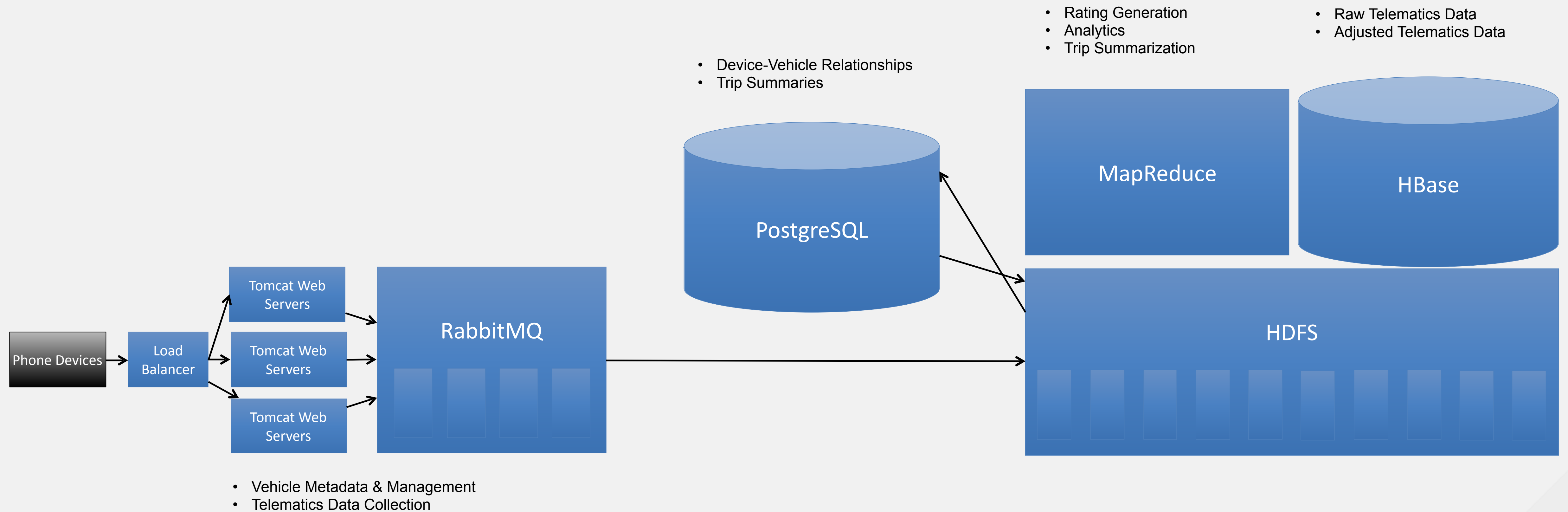
Solution

- Robust, Secure and Integrated Platform to enable Telematics Application leveraging best of breed Big Data & Analytics Cyber Security and Infrastructure platform/ technologies
- Hosted Big Data Ingestion Engine
- Highly Secure Hosted Network
- Scaled for High Volume Telemetric Data (6 million customers)
- High Availability supported by 2 Data Centers for Business Continuity
- World Secure Class Security & Monitoring Solutions
- 24X7 support

Results

- Highly Available Platform: Architected and managed to business SLA requirements
- Speed to Market – CSC platform enabled the customer to accelerate time to market and meet critical business timelines
- Allows for expansion and rapid scale up as the program is deployed and adopted across the 52 states
- Limited initial outlay to get up and running quickly

Usage-Based Auto Insurance Architecture



U	Rack1
42	Firewall
41	Cable Mgmt
40	Data Switch 1
39	Cable Mgmt
38	Data Switch 2
37	Cable Mgmt
36	Access Switch
35	Cable Mgmt
34	Management Switch
33	
32	
31	
30	
29	
28	
27	
26	
25	
24	
23	R620 Network Server
22	R620 Admin Server
21	R620 Admin Server
20	R720 General Purpose Compute
19	R720 General Purpose Compute
18	R720 General Purpose Compute
17	R720 General Purpose Compute
16	R720 General Purpose Compute
15	
14	R720XD Compute/Storage Server
13	R720XD Compute/Storage Server
12	R720XD Compute/Storage Server
11	R720XD Compute/Storage Server
10	R720XD Compute/Storage Server
9	R720XD Compute/Storage Server
8	R720XD Compute/Storage Server
7	R720XD Compute/Storage Server
6	R720XD Compute/Storage Server
5	R720XD Compute/Storage Server
4	R720XD Compute/Storage Server
3	
2	Power Cabling
1	Power Cabling

PROD

OpenStack Mgmt
OpenStack Mgmt
OpenStack Mgmt

CSC C&C
PostgreSQL
RabbitMQ
Hadoop NN
Hadoop RM

Hadoop DNs

U	Rack1
42	Firewall
41	Cable Mgmt
40	Data Switch 1
39	Cable Mgmt
38	Data Switch 2
37	Cable Mgmt
36	Access Switch
35	Cable Mgmt
34	Management Switch
33	
32	
31	
30	
29	
28	
27	
26	
25	
24	
23	
22	
21	
20	
19	
18	
17	
16	
15	
14	R720 General Purpose Compute
13	R720 General Purpose Compute
12	R720 General Purpose Compute
11	R720 General Purpose Compute
10	R720 General Purpose Compute
9	
8	R720XD Compute/Storage Server
7	R720XD Compute/Storage Server
6	R720XD Compute/Storage Server
5	R720XD Compute/Storage Server
4	R720XD Compute/Storage Server
3	
2	Power Cabling
1	Power Cabling

QA

CSC C&C
PostgreSQL
RabbitMQ
Hadoop NN
Hadoop RM

Hadoop DNs

U	Rack1
42	Firewall
41	Cable Mgmt
40	Data Switch 1
39	Cable Mgmt
38	Data Switch 2
37	Cable Mgmt
36	Access Switch
35	Cable Mgmt
34	Management Switch
33	
32	
31	
30	
29	
28	
27	
26	
25	
24	
23	R620 Network Server
22	R620 Admin Server
21	R620 Admin Server
20	R720 General Purpose Compute
19	R720 General Purpose Compute
18	R720 General Purpose Compute
17	R720 General Purpose Compute
16	R720 General Purpose Compute
15	
14	R720XD Compute/Storage Server
13	R720XD Compute/Storage Server
12	R720XD Compute/Storage Server
11	R720XD Compute/Storage Server
10	R720XD Compute/Storage Server
9	R720XD Compute/Storage Server
8	R720XD Compute/Storage Server
7	R720XD Compute/Storage Server
6	R720XD Compute/Storage Server
5	R720XD Compute/Storage Server
4	R720XD Compute/Storage Server
3	
2	Power Cabling
1	Power Cabling

DR

OpenStack Mgmt
OpenStack Mgmt
OpenStack Mgmt

CSC C&C
PostgreSQL
RabbitMQ
Hadoop NN
Hadoop RM

Hadoop DNs

FOR MORE INFO: csc.com/bigdata

RED HAT **SUMMIT**

**LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.**