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

- Vehicle, Asset, Person, and Pet Monitoring and Controlling
- Agriculture Automation
- Security and Surveillance
- Building Management

50 BILLION DEVICES INTERNET-CONNECTED BY 2020

- Embedded Mobile
- Everyday Things
- Smart Homes and Cities
- Telemedicine and Healthcare
Busting Corporate Data Silos

Source: https://gigaom.com/2013/05/07/with-300m-earmarked-for-tech-innovation-metlife-wants-to-remake-insurance/
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)
- Fast, small data business intelligence support

**Document Store**
- Full-text search
- Text searches (e.g. social media, email/chat, claims, contracts, etc.)
- Data exploration, discovery, and indexing

**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.

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

**Business Intelligence 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.

**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.)

**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

- Powerful Integration
  - Any Data Source, Real-Time to Batch
- Most Trusted Security Capabilities
- World-Class Managed Operations and Expert Services
- Cloud-Enabled Scalable Distributed

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

8 – 12 Week Sprint
Agile Scientific Approach to Measurable Business Improvement

Data Exploration & Transformation

Business Discovery

Data Modeling & Algorithm Development

Insight Operationalization

Change Management

Use Case Prioritization & Roadmap

Data Inventory Identification & Coordination

InsightLab

Inputs

Outputs

#redhat #rhsummit
PRODUCTION

Push Code w/ Git

Test w/ Jenkins

Push to Production

User Queries

LOCAL

Sample Data, Partial/Full Flows, or DR Replication

Performance Window

DEV / DR

VM/Sandbox or "local node" environment or "direct-dev" on BDPaaS

PRODUCTION Data Flows

C & C

Storm

Hive

Impala

Elasticsearch

Kafka

HDFS
**PRODUCTION**

- Kafka
- Storm
- HDFS
- Hive
- Impala
- Elasticsearch

- 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**

- Kafka
- Storm
- HDFS
- Hive
- Impala
- Elasticsearch

- 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
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

CSC Big Data Platform as a Service

ENTERPRISE GRADE SECURITY
FLEXIBLE DEPLOYMENT OPTIONS

General Compute
Dell 730

Storage-Heavy Compute
Dell 730xd

Management Compute
Dell 630
Red Hat Enterprise Linux OpenStack Platform + Dell Reference Architecture

Source: Red Hat + Dell Reference Architecture White Paper
ToR Switches providing in rack, inter rack and uplink communication

Dedicated physical servers 10GbE connectivity between servers for data and 1GbE for management

Additional Data Nodes dynamically added for increase capacity and processing

LaunchPad

LaunchPad + N
LaunchPad

1. Power Cabling
2. R720XD Compute/Storage Server
3. R720XD Compute/Storage Server
4. R720XD Compute/Storage Server
5. R720XD Compute/Storage Server
6. R720XD Compute/Storage Server
7. R720XD Compute/Storage Server
8. R720XD Compute/Storage Server
9. R720XD Compute/Storage Server
10. R720XD Compute/Storage Server
11. R720XD Compute/Storage Server
12. R720XD Compute/Storage Server
13. R720XD Compute/Storage Server
14. R720 General Purpose Compute
15. R720 General Purpose Compute
16. R720 General Purpose Compute
17. R720 General Purpose Compute
18. R720 General Purpose Compute
19. R620 Admin Server
20. R620 Network Server
21. R620 Admin Server
22. Access Switch
23. Data Switch 2
24. Data Switch 1
25. Firewall
26. Management Switch
27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42.
<table>
<thead>
<tr>
<th>Component</th>
<th>Main Stack</th>
<th>Mini Stack</th>
<th>POC Stack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workload</td>
<td>Full production deployment, client deployed, cluster separation by physical separation of workload, i.e. Production Cluster would have its own Data nodes etc</td>
<td>Local applications limitations - will deploy a deduplicated stack into the client's network. Clusters would logically segregated from each other if required.</td>
<td>Non-Production - deployed for POC engagement, prove and test data etc.</td>
</tr>
<tr>
<td>Min usable storage</td>
<td>4/2TB</td>
<td>10/26TB (1TB)</td>
<td>1 - 3TB (1TB)</td>
</tr>
<tr>
<td>Firewall</td>
<td>Cisco ASA 5520-5</td>
<td>N/A - Assumed will connect into clients existing network</td>
<td>N/A - Assumed will connect into clients existing network</td>
</tr>
<tr>
<td>Network Aggregation</td>
<td>Brocade VDX 5700 (Optional if client does not have a suitable aggregation layer)</td>
<td>N/A - Assumed will connect into clients existing network</td>
<td>N/A - Assumed will connect into clients existing network</td>
</tr>
<tr>
<td>Switch – Data</td>
<td>2 x Brocade ICX 6850 (1GBE)</td>
<td>1 x Brocade IXL 6810 (1GGE)</td>
<td>N/A - Assumed will connect into clients existing network</td>
</tr>
<tr>
<td>Switch – Management</td>
<td>1 x Brocade IXL 6810 (1GGE)</td>
<td>1 x Brocade IXL 6810 (1GGE)</td>
<td>1 x Brocade IXL 6810 (1GGE) covers all connectivity</td>
</tr>
<tr>
<td>Purpose</td>
<td>General Purpose Compute Server</td>
<td>General Purpose Compute Server</td>
<td>General Purpose + Data Nodes Compute Server</td>
</tr>
<tr>
<td>Virtualization</td>
<td>Yes - OpenStack (Multi workload)</td>
<td>Yes - OpenStack (Segregation of all applications and data nodes / tenants would be segregated as well)</td>
<td>Yes - OpenStack</td>
</tr>
<tr>
<td>Applications Hosted</td>
<td>control (dashboard, mongodb, vcd, edge, splunk...), streaming (storm,kafka,listeners), queries (elasticsearch, perhaps some hbase components), hadoop (namenode, jobtracker), any aux instances (tableaux et al)</td>
<td>control (dashboard, mongodb, vcd, edge, splunk...), streaming (storm,kafka,listeners), queries (elasticsearch, hbase), hadoop (namenode, jobtracker, datanode/tasktracker), any aux instances (tableaux et al)</td>
<td>control (dashboard, mongodb, vcd, edge, splunk...), streaming (storm,kafka,listeners), queries (elasticsearch, hbase), hadoop (namenode, jobtracker, datanode/tasktracker), any aux instances (tableaux et al)</td>
</tr>
<tr>
<td>Minimum</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Maximum</td>
<td>128 GB</td>
<td>128 GB</td>
<td>128GB</td>
</tr>
<tr>
<td>Provider</td>
<td>Dell</td>
<td>Dell</td>
<td>Dell</td>
</tr>
<tr>
<td>Model</td>
<td>R730</td>
<td>R730</td>
<td>R730</td>
</tr>
<tr>
<td>CPU</td>
<td>Dual 18 Core CPU Intel® Xeon® E5-2695 v2 2.40GHz</td>
<td>Dual 18 Core CPU Intel® Xeon® E5-2695 v2 2.40GHz</td>
<td>Dual 18 Core CPU Intel® Xeon® E5-2695 v2 2.40GHz</td>
</tr>
<tr>
<td>RAM</td>
<td>192GB</td>
<td>192GB</td>
<td>192GB</td>
</tr>
<tr>
<td>Network</td>
<td>2 * 10Gb Nics / 2 * 1GB Nics</td>
<td>2 * 10Gb Nics / 2 * 1GB Nics</td>
<td>2 * 10Gb Nics / 2 * 1GB Nics</td>
</tr>
<tr>
<td>Storage</td>
<td>16TB (8x2TB,3.5&quot;;SATA,7.2k Near line SAS)</td>
<td>16TB (8x2TB,3.5&quot;;SATA,7.2k Near line SAS)</td>
<td>16TB (8x2TB,3.5&quot;;SATA,7.2k Near line SAS)</td>
</tr>
<tr>
<td>Data Node Sizing</td>
<td>NA</td>
<td>2nd Chassis Handles Data Nodes = 48/4.5/10TB Usable</td>
<td>NA</td>
</tr>
<tr>
<td>Purpose</td>
<td>Data Compute Server</td>
<td>Data Compute server</td>
<td>Data Compute Server</td>
</tr>
<tr>
<td>Virtualization</td>
<td>Yes - OpenStack - 4:1 Ratio</td>
<td>Yes - OpenStack - Logically separate customers</td>
<td>NA - Covered in General Purpose + Data Nodes Compute Server</td>
</tr>
<tr>
<td>Applications Hosted</td>
<td>hadoop datanode/tasktrackers, perhaps some hbase components</td>
<td>hadoop datanode/tasktrackers, perhaps some hbase components</td>
<td>NA - Covered in General Purpose + Data Nodes Compute Server</td>
</tr>
<tr>
<td>Minimum</td>
<td>6 Data Nodes (50TB * 6 = 300TB)</td>
<td>2 Data Nodes (50TB * 2 = 100TB)</td>
<td>6 Data Nodes (50TB * 6 = 300TB)</td>
</tr>
<tr>
<td>Maximum</td>
<td>4 Data Nodes (50TB * 4 = 200TB)</td>
<td>4 Data Nodes (50TB * 4 = 200TB)</td>
<td>4 Data Nodes (50TB * 4 = 200TB)</td>
</tr>
<tr>
<td>Raw to Usable Ratio</td>
<td>4.5:1</td>
<td>4.5:1</td>
<td>4.5:1</td>
</tr>
<tr>
<td>Provider</td>
<td>Dell</td>
<td>Dell</td>
<td>Dell</td>
</tr>
<tr>
<td>Model</td>
<td>R730xd</td>
<td>R730xd</td>
<td>R730xd</td>
</tr>
<tr>
<td>CPU</td>
<td>Dual 8 Core CPU</td>
<td>Dual / 8 Core CPU</td>
<td>Dual / 8 Core CPU</td>
</tr>
<tr>
<td>RAM</td>
<td>128GB</td>
<td>128GB</td>
<td>128GB</td>
</tr>
<tr>
<td>Network</td>
<td>2 * 10Gb Nics / 2 * 1GB Nics</td>
<td>2 * 10Gb Nics / 2 * 1GB Nics</td>
<td>2 * 10Gb Nics / 2 * 1GB Nics</td>
</tr>
<tr>
<td>Storage</td>
<td>50TB (12x4TB,3.5&quot;;SATA,7.2k +2x 1TB 2.5&quot; Near Line SAS)</td>
<td>50TB (12x4TB,3.5&quot;;SATA,7.2k +2x 1TB 2.5&quot; Near Line SAS)</td>
<td>50TB (12x4TB,3.5&quot;;SATA,7.2k +2x 1TB 2.5&quot; Near Line SAS)</td>
</tr>
<tr>
<td>Purpose</td>
<td>Management Server (Network / Admin)</td>
<td>Management Server (Network / Admin)</td>
<td>Management Server (Network / Admin)</td>
</tr>
<tr>
<td>Virtualization</td>
<td>NA</td>
<td>NA</td>
<td>NA - Covered in General Purpose + Data Nodes Compute Server</td>
</tr>
<tr>
<td>Applications Hosted</td>
<td>OpenStack control: Openstack-api, Openstack-dashboard, glance (api, registry), keystone, nova (api), cert, conductor, console-auth, nolnpfoxpy, scheduler, openvswitch, neutron (server, meltdown,3.13,hcpp), and potentially swift (proxy, account, container, object)</td>
<td>OpenStack control: Openstack-api, Openstack-dashboard, glance (api, registry), keystone, nova (api), cert, conductor, console-auth, nolnpfoxpy, scheduler, openvswitch, neutron (server, meltdown,3.13,hcpp), and potentially swift (proxy, account, container, object)</td>
<td>OpenStack control: Openstack-api, Openstack-dashboard, glance (api, registry), keystone, nova (api), cert, conductor, console-auth, nolnpfoxpy, scheduler, openvswitch, neutron (server, meltdown,3.13,hcpp), and potentially swift (proxy, account, container, object)</td>
</tr>
<tr>
<td>Minimum</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Maximum</td>
<td>Add based on requirements and HA option</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Provider</td>
<td>Dell</td>
<td>Dell</td>
<td>Dell</td>
</tr>
<tr>
<td>Model</td>
<td>R630</td>
<td>R630</td>
<td>R630</td>
</tr>
<tr>
<td>CPU</td>
<td>Dual 8 Core CPU</td>
<td>Dual 8 Core CPU</td>
<td>Dual 8 Core CPU</td>
</tr>
<tr>
<td>RAM</td>
<td>128 GB</td>
<td>128 GB</td>
<td>128 GB</td>
</tr>
<tr>
<td>Network</td>
<td>2 * 10Gb Nics / 2 * 1GB Nics</td>
<td>2 * 10Gb Nics / 2 * 1GB Nics</td>
<td>2 * 10Gb Nics / 2 * 1GB Nics</td>
</tr>
<tr>
<td>Storage</td>
<td>9.6TB RAW 10K RPM SAS 6Gbps 2.5in</td>
<td>9.6TB RAW 10K RPM SAS 6Gbps 2.5in</td>
<td>9.6TB RAW 10K RPM SAS 6Gbps 2.5in</td>
</tr>
</tbody>
</table>
Usage-Based Auto Insurance
Big Data Telematics in the Open Private Cloud

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

- Vehicle Metadata & Management
- Telematics Data Collection
- Device-Vehicle Relationships
- Trip Summaries
- Rating Generation
- Analytics
- Trip Summarization
- Raw Telematics Data
- Adjusted Telematics Data

Phone Devices

Load Balancer

Tomcat Web Servers

Tomcat Web Servers

Tomcat Web Servers

RabbitMQ

PostgreSQL

MapReduce

HBase

HDFS

#redhat #rhsummit
LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.