This Conference brought to you by

www.ttcus.com

Linkedin/Group: Technology Training Corporation

@Techtrain

Technology Training Corporation

www.ttcus.com
Optimizing Data Storage and Management for Big Data in the Federal Government

Greg Gardner, PhD, CISSP
Colonel, Infantry, U.S. Army (Retired)
Chief Architect, Defense and Intel Solutions
NetApp
Conference Agendas…
and Baseball Lineups

- No. 1, or lead off, hitter: Traditionally, lead off hitters were fast and reached base often.
- No. 2 hitter: Managers often used a hitter with exceptional bat control in the number 2 spot.
- No. 3 hitter: This spot usually is reserved for the team's best hitter, who often hits for a high average and still has power.
- No. 4, or cleanup, hitter: This batter is usually the team's biggest home-run threat.
Conference Agendas… and Baseball Lineups

- No. 7 and 8 hitters: These players often are weak hitters who are more important to the team for their defensive capabilities. The No. 7 hitter might have a little more pop in his bat that can drive in the players hitting in front of him. **AARON**

- The No. 8 batter is often light-hitting, but ideally can reach base on walks. **GREG**

- This allows the manager to use the No. 9 hitter to sacrifice, making his otherwise sure out into one that at least can advance the baserunner. **ALAN / GERARD**
All This Data and Not Even to the “Peak”

- Peak of Inflated Expectations
- Plateau of Productivity
- Slope of Enlightenment
- Trough of Disillusionment

- Technology Trigger

- VISIBILITY

- TIME

- Estimated size of the digital universe in 2020: **35 Zettabytes**
- Estimated number of Smart phones: **5 Billion**
- Number of Pieces of new content to Facebook per month: **30 Billion**
- Percentage of Unstructured data: **80%**
Analytics Platform

- Intelligent Search
- Rules Engine
- Visualization

Real Time Search Index

- N1
- N2
- N3
- N4
- …
- NN

Nodes

- Big Data Analytics
- Scale Out Platform
- Fault Tolerant
- Data Archival

Parallel Analytics Platform

Data Collectors

Hub Location

Agent
Adapters
Edge Location

Edge Location

Agent
Adapters

©2012 Zaloni, Inc. All Rights Reserved.
Coke’s Smart Dispenser Monitoring and Analytics

- Micro-dispensing Technology to serve 100+ drinks

- Cartridges store concentrated ingredients in dispenser and are RFID enabled.

- RFID chips detect supplies and call-home

- Other call home data: Daily debug, Diagnostics, Product usage

- 100,000 Freestyles in 2012
Coke Freestyle Hadoop Analytical Architecture

Dispenser
Agent

Collectors

Rules Engine
Visualization
Intelligent Search

HDFS
HBase
Lucene Index

Analytics Platform

Local Maintainer
Regional Analyst / Manager
National Sales Director
Local Data Capture for Machine Maintenance and Stockage

© 2012 Zaloni, Inc. All Rights Reserved.
Regional Metrics and Intelligent Search

Smart Dispenser Analytics Dashboard

Search: error OR failed OR severe OR (date=lastmonth)

Messages Found (109)

<table>
<thead>
<tr>
<th>Location</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlanta (32)</td>
<td>Aug 12, 2011 8:53:12 AM 2312 [IP: 10.9.10.212] [Atlanta Moes #211] [ERROR] Dispenser controller buffer overflow</td>
</tr>
<tr>
<td>Seattle (18)</td>
<td>Aug 22, 2011 11:45:41 AM 2312 [IP: 10.9.10.47] [Dallas Six Flags #92] [ERROR] Dispenser controller buffer overflow</td>
</tr>
<tr>
<td>Denver (15)</td>
<td></td>
</tr>
<tr>
<td>Washington (11)</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td></td>
</tr>
<tr>
<td>Microdispenser (68)</td>
<td></td>
</tr>
<tr>
<td>Electrical (45)</td>
<td></td>
</tr>
<tr>
<td>Environmental (39)</td>
<td></td>
</tr>
<tr>
<td>Download (21)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous (20)</td>
<td></td>
</tr>
</tbody>
</table>
National Level Metrics and Visualization

Smart Dispenser Analytics Dashboard

Categories
- Location
  - Atlanta (32)
  - Los Angeles (29)
  - Seattle (18)
  - Denver (15)
  - Washington (11)
- Type
  - Microdispenser (68)
  - Electrical (45)
  - Environmental (39)
  - Download (21)
  - Miscellaneous (20)

Results
- Aug 12, 2011 8:53:12 AM 2312 [IP: 10.9.10.212] [Atlanta Moes #211] [ERROR] Dispenser controller buffer overflow
  - Dispenser: 201002910 | sourcetype: microdispenser | sourcefile: /logs/08-12-2011-201002910
- Aug 19, 2011 3:22:12 PM 2822 [IP: 10.9.8.22] [Seattle SEATAC #23] [FAILED] Power supply fan failure - Primary
- Aug 22, 2011 11:45:41 AM 2312 [IP: 10.9.10.47] [Dallas Six Flags #92] [ERROR] Dispenser controller buffer overflow
  - Dispenser: 210920199 | sourcetype: microdispenser | sourcefile: /logs/08-12-2011-210920199
Same model for Military Conditions Based Maintenance (CBM)??
Local Data Capture / Retrieval of Maintenance and Logistical information
The Goal: Net-centric Logistics Environment

Improved Knowledge of Situation
Improved Readiness
Improved Materiel Solutions
Reduced Life-Cycle Costs

AILA
Common Operations
Common Systems Views
Common Processes
Common Data Standards

Common Information Management Service (CIMS) – Software *
Net-Centric Information

Implementation and End-to-End Integration
Joint Integration

Concept/Doctrine/Policy Foundation

Data
Condition Based Maintenance (CBM)

**PREVENTIVE**
- Reactive Maintenance
- Time Based Inspection/Overhaul

**INDICATORS**
- Digital Source Collector Installation
- Knowledge Development
- Fault Diagnosis
- Remaining Useful Life Calculation
- Inspection Targeting

**DIAGNOSTICS**
- Proactive Maintenance
- ‘On Condition’ Inspection/Overhaul

**PROGNOSTICS**

**ON-CONDITION**

The Purpose of Army Maintenance is to Generate Combat Power.

**CBM Program Objectives:**
- Decrease Maintenance Burden on the Soldier
- Increase Platform Availability and Readiness
- Enhance Safety
- Reduce Operations & Support (O&S) Costs

**Key CBM Enablers:**
- Digital Source Collectors
- Flight Line Diagnostics
- Data Fusion/Analysis

© 2011 NetApp, Inc. All rights reserved.

This is NOT industry best practice…
The SmartSignal Difference
Multiple Sensor, Dynamic Empirical Analysis in Real Time

Single-sensor equipment protection: traditional thresholds & rules

Machine Trip
Alarm / Trip

SmartSignal — Early Detection
Monitors all signals simultaneously

Axial position
Gas pressure Δ
Speed
Bearing Temp

Early Warning
Time Difference
Alarm Hi
Dynamic Band

Normal operation
Few false alarms

Early stages of damage
SmartSignal alerts before traditional alarms
What’s The Problem?

- Incentives
- Culture
Big Data ABCs

Analytics
Analytics for extremely large datasets

Gain Insight

Big Data

Keep Everything

Go Fast

Content
Secure boundless data storage

Bandwidth
Performance for data intensive workloads
Big Data Market Requirements

Analytics
- Distributed compute farms with standard software
- Simpler systems, faster to deploy
- Optimized balance between compute & storage
- Reliable data (metadata)

Bandwidth
- Very high streaming bandwidth
- Dense configurations - GB/s/Rack Unit
- Large data sets (containers)
- Simpler system configurations

Content
- Boundaryless containers (10’s PBs)
- Simple access to dynamic datasets
- New access technology (CDMI) with legacy application interfaces
- Highly efficient, self-managing systems
## Big Content Solutions

<table>
<thead>
<tr>
<th>Solution</th>
<th>Workloads/Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>File Services</td>
<td>- Multiapplication/client workloads</td>
</tr>
<tr>
<td></td>
<td>- Frequent read/write access</td>
</tr>
<tr>
<td>Enterprise Content Repositories</td>
<td>- Infinite container</td>
</tr>
<tr>
<td></td>
<td>- Active archive</td>
</tr>
<tr>
<td>Distributed Content Repositories</td>
<td>- Large, multisite repository</td>
</tr>
<tr>
<td></td>
<td>- Policy-based data management</td>
</tr>
</tbody>
</table>
What Is Apache Hadoop?

An open-source software framework that:

- Supports data-intensive distributed applications
- Enables applications to work with thousands of nodes and petabytes of data
- Runs on a collection of commodity, shared-nothing servers
- Has two key components:
  - HDFS. Hadoop Distributed File System
  - MapReduce. Programming model for processing and generating large datasets
Hadoop
Target Markets

- **Full Motion Video (FMV)**
  - Government agencies building multi-petabyte video capture/playback solutions from satellites, drones, etc.

- **Media Content Management (MCM)**
  - Broadcast and cable networks video w/ ingest/playback active repositories

- **Seismic Processing Solution (SPS)**
  - Upstream Energy customers building Seismic Processing systems

- **Open Solution for Hadoop (OSH)**
  - Orgs. building large Hadoop clusters requiring high bandwidth and density

- **HPC Solution for Lustre (Lustre)**
  - Government and research institutes building HPC solutions
Vertical Applications

Retail
- Sales reports
- Target marketing
- New product rollout
- Customer loyalty
- Customer buying habits
- Inventory predictions
- New customer acquisition

Financial
- Sales reports
- Fraud detection
- Targeted marketing
- Customer data
- Compliance

Enterprise
- Sales reports
- Customer analysis
- Cost trends
- Quality control

Government
- Event monitoring
- Terrorism monitoring
- Data mining

Internet
- E-commerce
- Web sites
- Data mining
- Customer data
Enterprise-Class Features with Hadoop

- **Dynamic Disk Pools**: 6x faster rebuild times (minutes, not days) and continuous high performance during drive rebuild
- **SSD Cache and Hybrid Storage**: Expedited access to “hot” data through automated, real-time caching to SSD; mix and match SSDs and HDDs
- **Enhanced Snapshot Copies**: More precise recovery point objectives and faster recovery
- **Thin Provisioning**: Improve storage utilization up to 35% and eliminate overprovisioning
- **Enterprise Replication**: Cost-effective enterprise-class disaster recovery of data with FC and IP replication
- **Encrypted Drive Support**: Extended security enhancements for compliance and regulations
Use the ESG Report
Validated for the Enterprise

- Improved the performance of the Hadoop cluster by 62%
- Completed jobs under drive failure more than 200% faster
- Delivered linear performance scalability as data nodes, sets expanded

The Open Solution for Hadoop improves capacity and performance efficiency and recoverability compared to a server-based DAS deployment.

- ESG, 2012
Use Case Example: NetApp Auto Support

Phone home data representing information about the status NetApp storage controllers

- Correlate disk latency (hot) with disk type
  - 24 billion records
  - 4 weeks to run query
  - Hadoop implementation 10.5 hours

- Bug detection through pattern matching
  - 240 billion records – Too large to run
  - Hadoop implementation 18 hours
NIF: Big Science Needs Big Data

With a private cloud, NIF was able to:

• Deliver nonstop, 24/7 availability for critical data
• Cut planned downtime by 60 hours per year to maximize facility availability
• Reduce latency by 97% to provide data to scientists within 15 minutes of an experiment
• Enable secure multi-tenancy to protect sensitive scientific and government data
State-of-the-Science in HPC Today

**LLNL Sequoia** - designed to be the fastest supercomputer and storage combination on the planet

- 16 Max / 20 Peak PetaFLOPS \( (20 \times 10^{15}) \)
- 1.6 Million Cores
- 1.6 PB main memory
- > 1.3 Terabyte/sec using:
  - 55 PB of usable storage
  - Lustre Cluster Parallel File System
  - QDR (40 Gigabit) Infiniband
- ~8 MW of Power
- Simulations for
  - Nuclear Weapons Viability
  - Counter Terrorism
  - Energy Security
  - Climate Change


Big Data in HR: Some Fundamental Principles

- Transparency
- Privacy
- “Do No Harm”
- Validity and Verification
- Security
Big Data: Transforming the Design Philosophy of Future Internet
Remember:
With Big Data, it’s all about…

- Incentives
- Culture
Thank you