Getting Started & Successful with Big Data

@Pentaho #BigDataWebSeries
Your Hosts Today

Davy Nys
VP EMEA & APAC
Pentaho

Paul Brook
Cloud EMEA Program Manager
Dell

Chuck Yarbrough
Technical Solutions Marketing
Pentaho
Big Data Integration Webinar Series

Join Pentaho for a four-part technical webinar series focusing on the key elements and trends surrounding big data. Each week of the series will bring a new, content-rich webinar helping organizations find the right track to understand, recognize value and cost-effectively deploy big data analytics.

<table>
<thead>
<tr>
<th>Session</th>
<th>Topics</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1</td>
<td>Enterprise Data Warehouse Optimization with Hadoop Big Data</td>
<td>May 8, 2013</td>
</tr>
<tr>
<td>Session 2</td>
<td>Getting Started and Successful with Big Data</td>
<td>May 15, 2013</td>
</tr>
<tr>
<td>Session 3</td>
<td>Reducing the Implementation Efforts of Hadoop, NoSQL and Analytical Databases</td>
<td>May 22, 2013</td>
</tr>
<tr>
<td>Session 4</td>
<td>Reporting, Visualization and Predictive from Hadoop</td>
<td>May 29, 2013</td>
</tr>
</tbody>
</table>

Sign-up at: pentaho.com
To Understand:

- How to get a Hadoop cluster up and running
- Where Hadoop and other pieces fit into the architecture
- How you can easily get data in & out Hadoop
- How to leverage Hadoop with Pentaho
- Initial Best Practices
Complete Analytics and Visual Data Management

Data Ingestion, Manipulation & Integration

Enterprise & Ad Hoc Reporting

Data Discovery & Visualization

Predictive Analytics & Machine Learning

NoSQL Databases

Hadoop

Analytic Databases

© 2012, Pentaho. All Rights Reserved. pentaho.com. Worldwide +1 (866) 660-7555
Data Warehouse Optimization

Data Sources
- ERP
- CRM
- CDR
- Other Data

Big Data Architecture
- Data Warehouse (Master & Transactional Data)
- Analytic Data Mart(s)

Analytic Datasets
- Raw Data
- Master Data
- Parsed Data
- Analytic Datasets

Hadoop

Tape Archive
### Steps To Start with Hadoop

<table>
<thead>
<tr>
<th>Step</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td><strong>Hadoop Installation</strong></td>
</tr>
<tr>
<td>• Install locally – as Pseudo-Distributed mode</td>
<td></td>
</tr>
<tr>
<td>• Leverage tools like Dell Crowbar</td>
<td></td>
</tr>
<tr>
<td>• Cloud sandbox</td>
<td></td>
</tr>
<tr>
<td><strong>2</strong></td>
<td><strong>Pentaho Installation</strong></td>
</tr>
<tr>
<td>• Easy download &amp; installation</td>
<td></td>
</tr>
<tr>
<td>• Start on desktop</td>
<td></td>
</tr>
<tr>
<td><strong>3</strong></td>
<td><strong>Start Loading</strong></td>
</tr>
<tr>
<td>• Extract or access data from source systems</td>
<td></td>
</tr>
<tr>
<td>• Load it (in its raw form) into Hadoop</td>
<td></td>
</tr>
<tr>
<td>• Tokenize &amp; parse as required</td>
<td></td>
</tr>
<tr>
<td>• Transform &amp; enrich</td>
<td></td>
</tr>
<tr>
<td>• Load into destination</td>
<td></td>
</tr>
</tbody>
</table>
Data Architecture and Integration Challenges

Data Sources
- ERP
- CRM
- CDR
- Other Data

Big Data Architecture
- Data Warehouse (Master & Transactional Data)
- Analytic Data Mart(s)
- NOSQL

- Raw Data
- Master Data
- Parsed Data
- Analytic Datasets
Data Architecture and Integration Challenges

Data Sources
- ERP
- CRM
- CDR
- Other Data

Extract Transform Load

Big Data Architecture
- Data Warehouse (Master & Transactional Data)
- Analytic Data Mart(s)
- NOSQL

Orchestration & Integration

Analytic Datasets
- Raw Data
- Master Data
- Parsed Data
- Analytic Datasets

Extract Transform Load

ERPs, CRMs, CDRs, and Other Data are connected to an Extract Transform Load process, which in turn connects to a Data Warehouse (Master & Transactional Data). The Data Warehouse connects to Analytic Data Mart(s) and NOSQL databases. The NOSQL database connects to an Orchestration & Integration process, which includes an MR component for processing data. The Orchestration & Integration process connects back to the Extract Transform Load process, forming a closed loop for data management and analytics.
Data Architecture and Integration Challenges

Data Sources
- ERP
- CRM
- CDR
- Other Data

Big Data Architecture
- Data Warehouse (Master & Transactional Data)
- Analytic Data Mart(s)
- NOSQL

Extract Transform Load

Orchestration & Integration
- Raw Data
- Master Data
- Parsed Data
- Analytic Datasets

MR

hadoop
Fast and easy way to deploy Hadoop clusters with Dell
Fast and easy way to deploy Hadoop clusters with Dell
Well we are ready, but how will the Hardware Team know how to size and design the Hadoop cluster?

I don’t know... and it may take a long time to build the Hadoop cluster.

Time is a critical factor, we need to get this project moving.
Reduce risk & increase flexibility with Dell

Reduce time to Cluster Sizing, Design & Deployment

Faster time to productive operations

Optimize and adapt for your needs

Deliver the best return on investment

Dell.com/Crowbar
Dell | Hadoop Solution

Proven solutions

Excels at supporting complex big data analyses across large collections of structured and unstructured data

- **Hadoop handles a variety** of workloads, including search, log processing, data warehousing, recommendation systems and video/image analysis

- Work on the **most modern scale-out architectures** using a clean-sheet design data framework

- **Without vendor lock-in**

---

“Dell … was one of the first of the hardware vendors to grasp the fact that cloud is about provisioning services, not about the hardware.”

Maxwell Cooter, Cloud Pro

---

Proven components

- Apache Hadoop software
- Crowbar software framework with a Hadoop barclamp
- PowerEdge C8000 Series, C6220, R720, R720XD
- Force10 or PowerConnect switches
- Reference Architecture
- Deployment Guide
- Joint Service and Support
- Partner Ecosystem
Crowbar Software Framework
A modular, open source framework

Crowbar
• Accelerates multi-node deployments
• Simplifies maintenance
• Streamlines ongoing updates

Built with DevOps
• Provides an operational model for managing big data clusters and cloud

Field-proven technologies
• Build on locally deployed Chef Server
• Raw servers to full cluster in <2 hours
• Hardened with more than a year of deployments

Apache 2 open source
• Multi-apps (Hadoop & OpenStack)
• Multi-OS (Ubuntu, RHEL, CentOS, SUSE)

NOT limited to Dell hardware
Deploy a Hadoop cluster in ~2 hours
Evolve to meet your needs over time with built in DevOps

Crowbar software framework

Use Crowbar to:
• Automate the deployment and configuration of a Hadoop cluster
• Quickly provision bare-metal servers from box to cluster with minimal intervention
• Maintain, upgrade and evolve your Hadoop cluster over time
• Leverage an open source framework backed by a growing global developer ecosystem

Reduce software licensing fees
100%

Reduce development time
4-6 mo.
Crowbar dashboard provides visibility
Leverage developer expertise worldwide

**Download** the open source software: https://github.com/dellcloudedge/crowbar

**Participate in an active community** http://lists.us.dell.com/mailman/listinfo/crowbar

**Get resources** on the Wiki: https://github.com/dellcloudedge/crowbar/wiki

**Visit** Dell.com/Crowbar, Crowbar@Dell.com
pentaho.com/download

- Install on a local desktop – no need for a cluster
- “Managed Code” no additional installations
- Pentaho will write to the Hadoop Distributed Cache for execution
3
Start Loading

Loading into HDFS & HIVE
- Hadoop Copy Files
- Specify source files / destination

Loading into HBASE
- Zookeeper host & port
- Specify HBASE Mapping
Demo
Maximize Performance

Parallel execution as MapReduce in the Hadoop cluster.

As much as 15x faster than hand-written code.
### Additional Best Practices

| Leverage Hadoop | Don’t do database lookups inside a Mapper/Reducer – bring the data set into HDFS  
|                 | Don’t transfer data between two clustering technologies – network overload |
| Don’t Boil the Ocean | Start with a small data set and validate logic & performance outside the cluster  
|                     | Gradually increase volumes and fine tune the application, cluster, data stores & network |
| It’s AND... AND | Leverage the various technologies available  
|                 | A combination of easy to use tools, powerful scripting and custom coding provides the best mix |
Q & A
Big Data Integration Webinar Series

Join Pentaho for a four-part technical webinar series focusing on the key elements and trends surrounding big data. Each week of the series will bring a new, content-rich webinar helping organizations find the right track to understand, recognize value and cost-effectively deploy big data analytics.

<table>
<thead>
<tr>
<th>Session</th>
<th>Topics</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1</td>
<td>Enterprise Data Warehouse Optimization with Hadoop Big Data</td>
<td>May 8, 2013</td>
</tr>
<tr>
<td>Session 2</td>
<td>Getting Started and Successful with Big Data</td>
<td>May 15, 2013</td>
</tr>
<tr>
<td>Session 3</td>
<td>Reducing the Implementation Efforts of Hadoop, NoSQL and Analytical Databases</td>
<td>May 22, 2013</td>
</tr>
<tr>
<td>Session 4</td>
<td>Reporting, Visualization and Predictive from Hadoop</td>
<td>May 29, 2013</td>
</tr>
</tbody>
</table>

Reducing the Implementation Efforts of Hadoop, NoSQL and Analytical Databases

It’s easy to put a working script together as part of an R&D project, but it’s not cost effective to maintain it throughout an ever building stream of user change requests, system and product updates. Watch the third webinar in the series to learn how choosing the right technologies and tools can provide you the agility and flexibility to transform big data without coding.

**Date/Time:**
Wednesday, May 22, 2013
8 am PT / 11 am ET / 16:00 GMT

Contact Us or Sign-up at: pentaho.com