Big Data – Finding the Needle in the Haystack
October 2013
Ron Marshall
Agenda

• Brief Overview of Big Data/Hadoop Technology
• Overview of a Big Data POC
Big Data Technology Overview

Velocity

Variety

Volume
Apache Hadoop is an open-source system to reliably store and process a LOT of information across many commodity computers.
What is Special about Hadoop

• Separates distributed system fault-forgiving code from application logic
• Lets you interact with the cluster as a entity
• Scales Linearly with data size of analysis complexity
## Why Not Just use RDBMS

<table>
<thead>
<tr>
<th></th>
<th>Traditional RDBMS</th>
<th>MapReduce</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Size</strong></td>
<td>Gigabytes/Terabytes</td>
<td>Petabytes</td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>Interactive and Batch</td>
<td>Batch</td>
</tr>
<tr>
<td><strong>Updates</strong></td>
<td>Read and Write Many Times</td>
<td>Write Once, Read Many Times</td>
</tr>
<tr>
<td><strong>Structure</strong></td>
<td>Static Schema</td>
<td>Dynamic Schema</td>
</tr>
<tr>
<td><strong>Integrity</strong></td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Scaling</strong></td>
<td>Nonlinear</td>
<td>Linear</td>
</tr>
</tbody>
</table>
The Hadoop Ecosystem

- **ELT Tools**
  - Pig (Data Flow)

- **BI Reporting**
  - Hive (SQL)

- **RDBMS**
  - Sqoop

- **Zookeeper (Coordinator)**
  - MapReduce (Job Scheduling/Execution System)
  - Hbase (Column DB)

- **AVRO (Serialization)**
  - HDFS
  - Hadoop Distributed File System
# The Hadoop Ecosystem

## Layer | Examples
---|---
Distributed File System | Hadoop File System (HDFA), Google File System (GFS)
Distributed Processing Frameworks | Google’s MapReduce
Application Frameworks | Hadoop
Distributed Databases | Google’s BigTable, Apache Hbase, Cassandra, Mongo, CouchDB, Impala, NoSQL
Data Warehouse Infrastructure | Hive, Sqoop
Query Processing Language | Pig
Scientific Languages | Mahout, R
Configuration Management | Zookeeper
Workflow | Oozie
Two Core Components

Hadoop provides: a reliable shared storage and analysis system. The storage is provided by HDFS, and analysis by MapReduce.
Hadoop

- The Hadoop Distributed File System (HDFS) is the primary distributed storage used by Hadoop applications.
- A HDFS cluster primarily consists of a NameNode that manages the file system metadata and DataNodes that store the actual data.
Mapreduce

The overall flow of a MapReduce operation
Cluster Nodes

Master Nodes (1 each)
- Name Node: Metadata server and database
- Secondary Name Node: Periodic Name Node File Merge
- Job Tracker: Scheduler

Slave Nodes (1-Many)
- Data Nodes Task Tracker
- Data Nodes Task Trackers

(Block Storage and Task Execution)

Oracle Big Data Appliance
The POC

Data Mart – Customer Analytics
Oracle Exadata V2
The POC – The Problem

Data Mart – Customer Analytics
The POC – The Problem

Up to 22 Hours to load 1.4 Billion Rows of Web-log data

Data Mart – Customer Analytics
The POC – The Problem

Data Mart – Customer Analytics
The POC – The Solution

Data Mart – Customer Analytics

40 GBs Infiniband

Analysis and Reports
The POC – The Solution

1. Oracle Loader for Hadoop
2. Sqoop
3. Oracle Expose External Table

Data Mart – Customer Analytics
The POC – The Solution

Data Mart – Customer Analytics

40 GBs Infiniband

Analysis and Reports
The Results

- Entire POC completed in 17 working days
  - Including installation and completing the use-cases
  - Live connection of Infiniband (no service Interruption)
- Processing time for main use-case reduced from 20+ hours down to 2 hours
  - 1.4 Billion rows of URL data map/reduced and loaded into the Exadata
- Oracle Loader for Hadoop up to 5 times faster than Sqoop
A Discovery

- The loading process during the POC was being throttled by Database Resource Controls
- After some tuning and removal of the resource controls the run time was to 4 Min.
Security

• Originally is was for a large amount of Public Web Data
• The systems are standard Linux and Commodity Hw
• LDAP
• Kerberos
• Network Encryption
• Pluggable Authentication
• Enforced HDFS File permissions
Use Cases

• Risk (from Overnight Batch to Hourly)
• Real Time Churn Analysis (Customer Interactions)
• Outside Patterns (Project to scan pictures of the sky to determine if there is significant changes)
• Disaster Response based on Trends (Social Media Chatter)
• Fraud Detection
• Traffic Flow
Conclusion

• Hadoop is a Powerful Tool
• It Is Early but Evolution is Rapid
• Once we Monetize/Privatize Data there is a Responsibility to Protect it, but also sometimes Budget
References

• Hadoop – The Definitive Guide (Tom White)
  • ISBN 978-0-596-52197-4