Hadoop, BigInsights, and Your Future

What you need to know to succeed
Agenda

• Big Data Overview
• Hadoop History, Basics, and Terms
• Hadoop and BigInsights
• Technical Information
• Demo

• Goal: Show something useful, quickly
Not The Goal!!!
Not Your Typical Presentation

• “Fresh” presentation style and format
• Please interrupt if you have a question
• Please answer questions, out loud, when asked 🙂
World is Changing

The resulting explosion of information creates a need for a new kind of intelligence

...to help build a Smarter Planet
Data Explosion Quiz

2 Billion Internet users by 2011

4.6 Billion Mobile Phones World Wide

1.3 Billion RFID tags in 2005

30 Billion RFID today

Twitter processes 7 terabytes of data every day

Facebook processes 10 terabytes of data every day
Big Data Overview

• There are actually two forms of Big Data…
Data in Motion
Data at Rest
Data Warehouse Challenge:

• Organize Operational Data
• Quickly Answer Known Questions
• Make Decisions from the Analytics
Big Data Challenge: Find New Insights

- Predict weather patterns to plan optimal wind turbine usage and placement
- Detect life-threatening conditions at hospitals in time to intervene
- Make risk decisions based on real-time transactional data
- Identify criminals and threats from disparate video, audio, and data feeds
Real-World Examples

**Data in Motion**
- Hospital in Toronto
- Real-time Vitals Monitoring

**Data at Rest**
- Vestas Energy
- Wind Turbine Placement
IBM Watson is a breakthrough in analytic innovation, but it is only successful because of the quality of the information from which it is working.
Big Data and Watson

**Big Data technology is used to build Watson’s knowledge base**

Watson uses the Apache Hadoop open framework to distribute the workload for loading information into memory.

- Approx. 200M pages of text (To compete on Jeopardy!)

**Watson technology offers great potential for advanced business analytics**

- POS Data
- CRM Data
- Social Media
- InfoSphere BigInsights

Distilled Insight
- Spending habits
- Social relationships
- Buying trends

Advanced search and analysis
Potential for You

• Ask yourself a few simple questions…
  • What data gets thrown away?
  • Which details are missing from DB2 or the Data Warehouse?
  • What data sources are not being used?
Covering the Three Vs

IBM’S BIG DATA APPROACH
Three Vs

• A Big Data solution must cover...
Velocity
Variety
IBM’s Big Data

• Data in Motion
  • Focuses on Velocity and Variety
• Data at Rest
  • Focuses on Volume and Variety
Data in Motion

- IBM InfoSphere Streams
- Current version is 2.0
- Established Leader
- Real-world customer examples
Why InfoSphere Streams?

• Processing of **Sensor** Data
  • GPS, Video, Environmental Sensors, etc.
• "**Data Exhaust**" processing
  • Web Server/Application Logs, Network, etc.
• Processing **High-data-rate** information
  • Financial Transactions, Call Details, On-the-Fly speech to text, real-time scoring
Data at Rest

- IBM InfoSphere BigInsights
- Hadoop-based platform
- IBM adds Enterprise-readiness
  - Removes Single Point of Failure
  - Provides Spreadsheet Interface
  - Ease of Use and Monitoring
Hadoop Basics, History, and Terminology

HADOOP OVERVIEW
Hadoop History

- **Google’s Challenge:**
  - Store and Process the Entire Internet
  - Provide Quick Search Results

- **Solution:** Required Something New
  - Divide and Conquer
  - Open Source Result: Hadoop
Terminology

- Open Source
- Committer
- Contributor
Open Source & IBM

Leveraging Open Source Innovation ...

...Committing ...

Big Data Platform

...Contributing ...

...and Giving Back

Open Source & IBM

Apache Commons
http://commons.apache.org/

PIG
ZooKeeper

Hadoop

Lucene™

HD: HBase

AVRO™

Eclipse

jaql

Unstructured Information Management Architecture
An Apache Project.
Terminology

- Open Source
- Committer
- Contributor
- XML / HTML / Java
- Java Script / JSON
- Data
  - Structured Data
  - Unstructured Data
Data Formats

Example XML

```html
<html>
<head>
  ...
</head>
<body>
  <h1>XML Example</h1>
  <p>Example XML Document</p>
  <pre>
&lt;?xml version="1.0" encoding="ISO-8859-1"?&gt;
&lt;!-- Edited by XMLSpy --&gt;
&lt;CATALOG&gt;
  &lt;CD&gt;
    &lt;TITLE&gt;Hide your heart&lt;/TITLE&gt;
    &lt;ARTIST&gt;Bonnie Tyler&lt;/ARTIST&gt;
    &lt;COUNTRY&gt;UK&lt;/COUNTRY&gt;
    &lt;COMPANY&gt;CBS Records&lt;/COMPANY&gt;
    &lt;PRICE&gt;9.90&lt;/PRICE&gt;
    &lt;YEAR&gt;1988&lt;/YEAR&gt;
  &lt;/CD&gt;
  &lt;CD&gt;
    &lt;TITLE&gt;Still got the blues&lt;/TITLE&gt;
    &lt;ARTIST&gt;Gary Moore&lt;/ARTIST&gt;
    &lt;COUNTRY&gt;UK&lt;/COUNTRY&gt;
    &lt;COMPANY&gt;Virgin records&lt;/COMPANY&gt;
    &lt;PRICE&gt;10.20&lt;/PRICE&gt;
    &lt;YEAR&gt;1990&lt;/YEAR&gt;
  &lt;/CD&gt;
&lt;/CATALOG&gt;
</pre>
</body>
</html>
```

Example JSON

```json
[
  { "CATALOG" : [
    { "CD" : [
      { "Title":"Hide your Heart",
        "Artist":"Bonnie Tyler",
        "Country":"UK",
        "Company":"CBS Records",
        "Price":9.90,
        "Year":1988 },
      { "Title":"Still got the blues",
        "Artist":"Gary Moore",
        "Country":"UK",
        "Company":"Virgin Records",
        "Price":10.20,
        "Year":1990 }
    ]
  ]
]
```
Terminology

- Open Source
- Committer
- Contributor
- XML / HTML / Java
- Java Script / JSON
- Data
  - Structured Data
  - Unstructured Data
- Parallel Processing
- Big Data
- Data in Motion
- Data at Rest
- Enterprise-Ready
- Hadoop
  - HDFS
  - Map/Reduce
What is Hadoop?

- Free Open Source framework
- Batch Oriented and Read-Intensive
- Contains Two Main Components
  - HDFS
  - Map / Reduce
HDFS

- Hadoop Distributed File System

Diagram:
- Application
  - POSIX API
    - Local File System
  - HDFS API
    - HDFS Interface
- Name Node
  - Data Node
    - Data Node
    - Data Node
Map / Reduce

**Map**

- Input split into pieces
- Worker nodes process individual pieces in parallel
- Each worker node stores its result in its local file system where a reducer is able to access it

**Reduce**

- Data is aggregated (‘reduced” from the map steps) by worker nodes
- Multiple reduce tasks can parallelize the aggregation
- Optional Combiner Function can be run to reduce data transfers
Hardware

• Central Processing Unit
• Memory
• Disk
Relational Databases

- Powerful Processors
- Fewer Total nodes/boxes in a cluster
- Large Quantity of Reliable Disks
- More-Expensive Hardware
- Approach
  - Spread *Data* Across Nodes
  - Support SQL

**GOAL:** Answer Known Questions
Hadoop

• Fewer Powerful Processors *per box*
• More nodes/boxes within a cluster
• Smaller Quantity of Disks *per box*
• Commodity/Less expensive Hardware
• Parallel Processing Approach
  • Spread *processing* across nodes
  • Support Map/Reduce

**GOAL:** Answer Unknown Questions
Comparison

- Relational
- Hadoop
Map / Reduce Example

File
- File Part 1
- File Part 2
- File Part 3
- File Part 4

Split

Map

LOCAL

Shuffle

Reduce

part1

part2

part3
Big Difference = Schema

- Relational
  Schema on Load
  Data

- Hadoop
  Schema on Run
  Data
Read-Time Quiz

• Max Read Speed of 7,200 RPM Drive?
  • 80 MB/sec

• How long will it take to read 1TB
  • 1 disk = 3.4 hours
  • 10 disks = 20 mins
  • 100 disks = 2 mins
  • 1,000 disks = 12 secs
Typical Node

- CPU
- 2 quad-core 2-2.5GHz
- 16-24 GB ECC RAM
- Disk = 4 x 1TB SATA disks
- Network = Gigabit Ethernet
IBM’s Enterprise-Ready, Big Data Solution

HADOOP AND BIGINSIGHTS
Hadoop Software

- Hadoop’s Additional Components
  - Languages
    - Java, Pig, Hive, JAQL
  - Other Components
    - Zookeeper, Lucene, Hbase, Oozie
- Open Source Software Downloaded and Installed separately
BigInsights

Enterprise Edition

- GPFS-SNC Native Support
- Spreadsheet-style data exploration
- Job and Workflow Management
- Productivity and Efficiency Improvements
- Integration with InfoSphere Warehouse
- Integration with Netezza
- Integration with DB2
- Large Scale Indexing
- Text Analytics
- Machine Learning
- Tiered Terabyte Pricing

Basic Edition

Free Download, Easy Installation
24x7 Web Support, 10TB Limit
Paid Support Option
Why IBM Enterprise?

- GPFS-SNC as an option to HDFS
- Analytics (Big Sheets, Text Analysis)
- Single Install/Monitoring = Ease of Use
- BigInsights Version 1.3 Enhancements
HDFS

- Hadoop Distributed File System
- Single Point of Failure
Big Sheets
# Single Install

## Summary

<table>
<thead>
<tr>
<th>Setting</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Install type</td>
<td>Standalone install</td>
</tr>
<tr>
<td>Configure SSH</td>
<td>Yes</td>
</tr>
<tr>
<td>BigInsights Admin Group ID</td>
<td>biadmin</td>
</tr>
<tr>
<td>BigInsights Admin User ID</td>
<td>biadmin</td>
</tr>
<tr>
<td>BigInsights installation directory</td>
<td>/opt/ibm/biginsights</td>
</tr>
<tr>
<td>BigInsights data/log directory</td>
<td>/var/ibm/biginsights</td>
</tr>
<tr>
<td>BigInsights console security</td>
<td>Install the BigInsights console with no user authentication</td>
</tr>
<tr>
<td>BigInsights console port</td>
<td>8080</td>
</tr>
<tr>
<td>Configure Jaql UDF server</td>
<td>Yes</td>
</tr>
<tr>
<td>Jaql UDF server port</td>
<td>8200</td>
</tr>
<tr>
<td>Derby port</td>
<td>1528</td>
</tr>
<tr>
<td>BigInsights orchestrator port</td>
<td>8888</td>
</tr>
<tr>
<td>Cache directory</td>
<td>/hadoop/mapred/local</td>
</tr>
<tr>
<td>Log directory</td>
<td>/var/ibm/biginsights/hadoop/logs</td>
</tr>
<tr>
<td>Map/Reduce system directory</td>
<td>/hadoop/mapred/system</td>
</tr>
<tr>
<td>Install HDFS</td>
<td>Yes</td>
</tr>
<tr>
<td>Shared POSIX File System Root</td>
<td></td>
</tr>
<tr>
<td>NameNode port</td>
<td>9000</td>
</tr>
<tr>
<td>NameNode HTTP port</td>
<td>50070</td>
</tr>
<tr>
<td>Name directory</td>
<td>/hadoop/hdfs/name</td>
</tr>
<tr>
<td>JobTracker port</td>
<td>9001</td>
</tr>
</tbody>
</table>
Monitor and Maintain
Version 1.3 Highlights

• “Apps” – Adhoc queries, Boardreader, etc.
• Eclipse Plug-ins – JAQL, Hive with SQL, MapReduce, Text Analytics
• Performance Improvements
  • Adaptive MapReduce
  • IBM’s LZO-like Compression
Background to the Demonstration

DEMO SETUP
Demo Background

• Data Source: Twitter
• Capture Tweets
• Analyze Tweet Quantity
• Analyze Tweet Content
Demo Software

• 64-bit Red Hat Enterprise Linux 6
• BigInsights EE v1.3
• InfoSphere Warehouse v9.7.3
• Cognos 10.1
Demo Flow

• Perform a Twitter Search
• Store tweets in BigInsights / Visualize Results in a Tag Cloud and Bar Chart
• Extract Summary data into InfoSphere Data Warehouse (View with Cognos)
• Analyze Trends in Tweet Quantity
• Use Big Sheets to view Tweet Text
Twitter Search
Demonstration of BigInsights with InfoSphere Warehouse, Cognos, and Big Sheets

DEMO
Last Quiz

• Big Data is…

• Data in **Motion**

• Data at **Rest**

• What are the Three Vs?

  • **Volume**, **Velocity**, and **Variety**

• Why IBM’s Enterprise Edition

  • **GPFS-SNC**, **Analytics**, **Ease of Use**
Go Try It Today

• Download BigInsights Basic Edition
  http://www.ibm.com/software/data/infosphere/biginsights/basic.html

• Review Twitter Search
  https://dev.twitter.com/docs/api/1/get/search

• Read Analyzing Big Data with JAQL
Resources

• Big Data University – Free Training
  http://bigdatauniversity.com/

• Powered by Hadoop
  http://wiki.apache.org/hadoop/PoweredBy

• IBM makes Big data easy for the Little Guy

• Understanding Big Data – Free PDF Book
Contact Information

• IBM Technical Sales
  Michael Nobles  mnobles@us.ibm.com