Big Data and Analytics in Government

Nov 29, 2012
Mark Johnson
Director, Engineered Systems Program
Agenda

- What Big Data Is
- Government Big Data Use Cases
- Building a Complete Information Solution
- Conclusion
The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle’s products remain at the sole discretion of Oracle.
Big Data Summary
Big data: techniques and technologies that enable organizations to effectively and economically analyze all of their data.
90% Of the World’s Data Has Been Created in the Last 2 Years
And That Data is Estimated to Grow 50X By 2020
Social Data Generated Every Minute

- 695,000 Status updates
- 510,040 Comments
- 204,166,667 Emails
- 2,000,000 Search Queries
- 571 New Websites
Big Data Buzz

“Why big data is a big deal”
InfoWorld – 9/1/11

“The challenge—and opportunity—of big data”
McKinsey Quarterly—5/11

“Ten reasons why Big Data will change the travel industry”
Tnooz -8/15/11

“Keeping Afloat in a Sea of 'Big Data”
ITBusinessEdge – 9/6/11

“Getting a Handle on Big Data with Hadoop”
Businessweek-9/7/11

“The promise of Big Data”
Intelligent Utility-8/28/11
What is Big Data?
What Makes it Big Data?

**VOLUME**
Very large quantities of data

**VELOCITY**
Extremely fast streams of data

**VARIETY**
Wide range of datatype characteristics
VALUE IS HARD TO FIND
Big Data - Tools

Hadoop (Map/Reduce)
A Map/Reduce Pipeline

**INPUT**

- Hurricane (2)
- Sandy (1)
- Hurricane (1)
- Flooding (2)
- Sandy (1)
- Water (1)
- Flooding (1)

**MAP**

**SHUFFLE /SORT**

**REDUCE**

**OUTPUT**

- Flooding (3)
- Hurricane (3)
- Sandy (2)
- Water (1)
Big Data Tools Add to Existing Toolsets

New ways of processing data we couldn’t access before
Government Big Data Use Cases
Big Data in Public Sector

- Fraud Prevention
- Revenue Management
- Constituent Sentiment
- Threat Identification
- Economic Analysis
- Healthcare
- Regulatory Compliance, Licensing & Law Enforcement
- Open Government
- Maintenance & Utilities
Predictive Policing
Where is a violent crime likely to occur?

- Weather - recent and forecasted
  - Retrieved from the National Weather Service during nightly ETL processes
- Contact Cards
- 911 Calls
- Incidents
- Arrests
- Day of Week
- Date of Month
Big Data to detect Sales Tax Fraud

Potential revenue losses approach $2.8B/year in one state
Big Data for Building Audits
Targeting over-occupied buildings

Neighborhood Socioeconomic Status

Year of Construction

New York City improved audit rates from 13% to over 70%

<table>
<thead>
<tr>
<th>Borough</th>
<th>Pot Change 00-05</th>
<th>Total S.00 plus to 2005</th>
<th>Total S.00 plus to 2005</th>
<th>S.00-500,000</th>
<th>S.500,000-1M</th>
<th>S.1M+</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manhattan</td>
<td>11.1%</td>
<td>81,435</td>
<td>92,207</td>
<td>84,020</td>
<td>14,894</td>
<td>12,415</td>
<td>217,750</td>
</tr>
<tr>
<td>Bronx</td>
<td>-1.4%</td>
<td>43,029</td>
<td>46,298</td>
<td>39,635</td>
<td>6,810</td>
<td>3,150</td>
<td>53,595</td>
</tr>
<tr>
<td>Queens</td>
<td>0.7%</td>
<td>75,676</td>
<td>77,429</td>
<td>75,295</td>
<td>2,135</td>
<td>2,135</td>
<td>73,665</td>
</tr>
<tr>
<td>Suffolk</td>
<td>-6.3%</td>
<td>29,125</td>
<td>31,129</td>
<td>32,801</td>
<td>3,402</td>
<td>1,725</td>
<td>36,028</td>
</tr>
<tr>
<td>Kings</td>
<td>-6.3%</td>
<td>17,271</td>
<td>18,664</td>
<td>16,226</td>
<td>2,879</td>
<td>1,189</td>
<td>18,090</td>
</tr>
<tr>
<td>Erie</td>
<td>0.0%</td>
<td>1,870</td>
<td>1,870</td>
<td>1,870</td>
<td>1,870</td>
<td>1,870</td>
<td>1,870</td>
</tr>
<tr>
<td>Queens</td>
<td>-11.2%</td>
<td>11,374</td>
<td>12,477</td>
<td>9,697</td>
<td>2,863</td>
<td>282</td>
<td>10,642</td>
</tr>
<tr>
<td>Westchester</td>
<td>-8.9%</td>
<td>1,147</td>
<td>1,494</td>
<td>1,494</td>
<td>1,494</td>
<td>414</td>
<td>2,444</td>
</tr>
<tr>
<td>Rockland</td>
<td>-6.6%</td>
<td>1,356</td>
<td>1,520</td>
<td>1,141</td>
<td>786</td>
<td>379</td>
<td>1,604</td>
</tr>
<tr>
<td>Orange</td>
<td>-4%</td>
<td>4,708</td>
<td>4,708</td>
<td>3,720</td>
<td>582</td>
<td>246</td>
<td>4,708</td>
</tr>
<tr>
<td>Richmond</td>
<td>-10.6%</td>
<td>4,380</td>
<td>5,345</td>
<td>4,380</td>
<td>884</td>
<td>267</td>
<td>5,345</td>
</tr>
<tr>
<td>Allegany</td>
<td>-5.2%</td>
<td>3,365</td>
<td>3,497</td>
<td>2,744</td>
<td>444</td>
<td>167</td>
<td>3,497</td>
</tr>
</tbody>
</table>

Use Case

Copyright © 2012, Oracle and/or its affiliates. All rights reserved.
Big Data in Healthcare

Find relationship between gene to cancer interaction

- Cross-referenced the relationships between 17000 genes and five major cancer types across 20 million medical publication abstracts.
- Cross-referenced genes from 60 Million patients and miRNA for a simulated 900 Million population.
- Understanding additional layers of the pathways these genes operate in and the drugs that target them is expected to help researchers in their work.
Policy Implications of Big Data in Government

• What information can/should the government collect & aggregate?
• How is that information (and the resultant data) protected?
• How are errors identified & corrected?
• How is data collected by private companies accessed?
• Which government agencies can use the data?
• Etc…
Building a Unified Information Management Solution
BIG DATA LIFECYCLE

NEW REQUIREMENTS

DECIDE

ACQUIRE

ANALYZE

ORGANIZE

New Tools for Acquiring & Organizing Information

Easy Integration of New Infrastructure

In-memory Processing

Advanced Analytics
Footprint in Most Organizations Today

Historic Source of Truth

Data Warehouse / Data Marts

ERP, CRM & Other Transactional Apps

Reporting, Query and Analysis Tools

Business Intelligence Tools
Goal: Make an Informed Recommendation

Structured Sources
- Constituent Analytics
  - Constituent History
- Job Analytics
  - Position Inventory

Unstructured Sources
- Constituent Behavior
  - Sentiment & Influence
  - Channel Impact
  - Job Placement

Real-Time Recommendations
- Analytics
- Search Promotions
Using all Data to Understand Constituents

Data Warehouse

Website Logs & Data

NoSQL DB

Employment Site
Discovering Valuable Data

Knowledge Discovery Engine

Unstructured
Semi-structured

Data Warehouse

Structured

High Volume Distributed File System

Website Logs & Data

NoSQL DB

Copyright © 2012, Oracle and/or its affiliates. All rights reserved.
Information Discovery Engine

Overview

Searc

Guided

Navigatio

Analytics

Copyright © 2012, Oracle and/or its affiliates. All rights reserved.
Information Discovery

Case Study

![Interaction Detail]

### Sentiment Trend

- # Negative, # Positive, # Neutral by created date hour range

```
created at: 2012-05-20T15:59:45.000Z
author username: PapaMarkyT
type: twitter
content: RT @OccupyBtown: Just saw two guys with bloody faces after riot police move in #NoNATO
link: http://twitter.com/PapaMarkyT/statuses/204330597865369600
klout score: 10
links url:
```

![Sentiment Trend Graph]

created date hour range
Valuable Data Found – Now Store it Securely

Knowledge Discovery Engine → Data Warehouse
Discoveries

High Volume Distributed File System

MapReduce code separates valued data, then sent to via specialized adapters to Data Warehouse

Website Logs & Data → NoSQL DB
Deploy Widely Available Reports & Analytics

Persistent Data Store for All Data of Value + In-DB Analytics

Knowledge Discovery Engine → Data Warehouse

Data Warehouse → BI Tools and Dashboards

Enterprise-class for reporting & analysis

High Volume Distributed File System

MapReduce code separates valued data, then sent to via specialized adapters to Data Warehouse

Website Logs & Data

NoSQL DB

Oracle
Make Well-Tuned Real-Time Recommendations

Persistent Data Store for All Data of Value + In-DB Analytics

Knowledge Discovery Engine → Data Warehouse

High Volume Distributed File System

MapReduce code separates valued data, then sent to via specialized adapters to Data Warehouse

BI Tools and Dashboards

Real-Time Analytics and Recommendations

Location & User Profile

Recommend

Website Logs & Data

NoSQL DB

Copyright © 2012, Oracle and/or its affiliates. All rights reserved.
Summary
How do I Start?

• Think Big and Start Small
  – Find a question or requirement that your organization has been wrestling with responding to

• Establish a clear scope
  – Try for a quick win

• Use tools to flatten the initial learning curve
  – Setting up a Hadoop cluster is a specialized skill set

• Scale up as necessary
Hardware and Software
Engineered to Work Together