CS 698: Special Topics in Big Data

Chapter 3. Overview of Big Data Analytics

Chase Wu
New Jersey Institute of Technology
Oak Ridge National Laboratory

Some of the slides have been provided through the courtesy of Dr. Ching-Yung Lin at Columbia University
Big Data is going to impact us soon!

Call Bruce Willis!!
Big Data Market


http://wikibon.org/wiki/w/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017
## Big Data Market

### 2013 Worldwide Big Data Revenue by Vendor ($US millions)

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Big Data Revenue</th>
<th>Total Revenue</th>
<th>Big Data Revenue as % of Total Revenue</th>
<th>% Big Data Hardware Revenue</th>
<th>% Big Data Software Revenue</th>
<th>% Big Data Services Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>$1,368</td>
<td>$99,751</td>
<td>1%</td>
<td>31%</td>
<td>27%</td>
<td>42%</td>
</tr>
<tr>
<td>HP</td>
<td>$869</td>
<td>$114,100</td>
<td>1%</td>
<td>42%</td>
<td>14%</td>
<td>44%</td>
</tr>
<tr>
<td>Dell</td>
<td>$652</td>
<td>$54,550</td>
<td>1%</td>
<td>85%</td>
<td>0%</td>
<td>15%</td>
</tr>
<tr>
<td>SAP</td>
<td>$545</td>
<td>$22,900</td>
<td>2%</td>
<td>0%</td>
<td>76%</td>
<td>24%</td>
</tr>
<tr>
<td>Teradata</td>
<td>$518</td>
<td>$2,665</td>
<td>19%</td>
<td>36%</td>
<td>30%</td>
<td>34%</td>
</tr>
<tr>
<td>Oracle</td>
<td>$491</td>
<td>$37,552</td>
<td>1%</td>
<td>28%</td>
<td>37%</td>
<td>36%</td>
</tr>
<tr>
<td>SAS Institute</td>
<td>$480</td>
<td>$3,020</td>
<td>16%</td>
<td>0%</td>
<td>68%</td>
<td>32%</td>
</tr>
<tr>
<td>Palantir</td>
<td>$418</td>
<td>$418</td>
<td>100%</td>
<td>0%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Accenture</td>
<td>$415</td>
<td>$30,606</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>PWC</td>
<td>$312</td>
<td>$32,580</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Deloitte</td>
<td>$305</td>
<td>$33,050</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Pivotal</td>
<td>$300</td>
<td>$300</td>
<td>100%</td>
<td>15%</td>
<td>50%</td>
<td>35%</td>
</tr>
<tr>
<td>Cisco Systems</td>
<td>$295</td>
<td>$50,200</td>
<td>1%</td>
<td>72%</td>
<td>12%</td>
<td>16%</td>
</tr>
</tbody>
</table>

Big Data Revenue by Sub-Type, 2013

(in $US millions)
(n=$18,814)

- Professional Services: $6,148
- Compute: $3,647
- Storage: $3,085
- Applications: $1,691
- SQL: $1,306
- Cloud: $1,192
- Infrastructure Software: $830
- Networking: $417
- NoSQL: $290
NoSQL DB ==> Distributed DB, Document-Oriented DB, Graph NoSQL DB, and In-Memory NoSQL DB.

“It is not uncommon for an enterprise IT organization to support multiple NoSQL DBs alongside legacy RDBMSs. Indeed, there are single applications that often deploy two or more NoSQL solutions, e.g., pairing a document-oriented DB with a graph DB for an analytics solution.” [Dec 2013]

<table>
<thead>
<tr>
<th>Service</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Data XaaS Revenue</td>
<td>$1.71</td>
<td>$2.43</td>
<td>$2.87</td>
<td>$3.19</td>
</tr>
<tr>
<td>Big Data Professional Services Revenue</td>
<td>$9.24</td>
<td>$12.31</td>
<td>$14.06</td>
<td>$15.30</td>
</tr>
<tr>
<td>Big Data Application (Analytic and Transactional) Revenue</td>
<td>$3.24</td>
<td>$4.94</td>
<td>$6.05</td>
<td>$6.89</td>
</tr>
<tr>
<td>Big Data NoSQL Database Revenue</td>
<td>$0.73</td>
<td>$1.14</td>
<td>$1.41</td>
<td>$1.62</td>
</tr>
<tr>
<td>Big Data SQL Database Revenue</td>
<td>$2.00</td>
<td>$2.48</td>
<td>$2.74</td>
<td>$2.91</td>
</tr>
<tr>
<td>Big Data Infrastructure Revenue</td>
<td>$0.67</td>
<td>$0.93</td>
<td>$1.08</td>
<td>$1.19</td>
</tr>
<tr>
<td>Big Data Networking Revenue</td>
<td>$0.67</td>
<td>$0.89</td>
<td>$1.02</td>
<td>$1.11</td>
</tr>
<tr>
<td>Big Data Storage Revenue</td>
<td>$4.39</td>
<td>$5.85</td>
<td>$6.68</td>
<td>$7.27</td>
</tr>
<tr>
<td>Big Data Compute Revenue</td>
<td>$5.23</td>
<td>$6.70</td>
<td>$7.50</td>
<td>$8.06</td>
</tr>
<tr>
<td>Total Big Data Revenue</td>
<td>$27.9</td>
<td>$37.7</td>
<td>$43.4</td>
<td>$47.5</td>
</tr>
</tbody>
</table>
5 Key Big Data Use Case Categories – IBM’s Perspective

Big Data Exploration
Find, visualize, and understand all big data to improve decision making

Enhanced 360° View of the Customer
Extend existing customer views (MDM, CRM, etc.) by incorporating additional internal and external information sources

Security/Intelligence Extension
Lower risk, detect fraud and monitor cyber security in real-time

Operations Analysis
Analyze a variety of machine data for improved business results

Data Warehouse Augmentation
Integrate big data and data warehouse capabilities to increase operational efficiency
Techniques towards Big Data

• Massive Parallelism
• Huge Data Volumes Storage
• Data Distribution
• **High-Speed Networks**
• **High-Performance Computing**
• Task and Thread Management
• **Data Mining and Analytics**
• Data Retrieval
• Machine Learning
• **Data Visualization**

⇒ Techniques exist for years to decades. Why did Big Data become hot now?
Why Big Data now?

- More data are being collected and stored
- Open source code
- Commodity hardware
Definition and Characteristics of Big Data

"Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

-- Gartner, Inc.

which was derived from:

“While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: volumes, velocity and variety. In 2001/02, IT organizations much compile a variety of approaches to have at their disposal for dealing each.” – Doug Laney
What made Big Data needed?

“Big Data Analytics”, David Loshin, 2013
### Contrasting Approaches in Adopting High-Performance Capabilities

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Typical Scenario</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Application development</strong></td>
<td>Applications that take advantage of massive parallelism developed by specialized developers skilled in high-performance computing, performance optimization, and code tuning</td>
<td>A simplified application execution model encompassing a distributed file system, application programming model, distributed database, and program scheduling is packaged within Hadoop, an open source framework for reliable, scalable, distributed, and parallel computing</td>
</tr>
<tr>
<td><strong>Platform</strong></td>
<td>Uses high-cost massively parallel processing (MPP) computers, utilizing high-bandwidth networks, and massive I/O devices</td>
<td>Innovative methods of creating scalable and yet elastic virtualized platforms take advantage of clusters of commodity hardware components (either cycle harvesting from local resources or through cloud-based utility computing services) coupled with open source tools and technology</td>
</tr>
<tr>
<td><strong>Data management</strong></td>
<td>Limited to file-based or relational database management systems (RDBMS) using standard row-oriented data layouts</td>
<td>Alternate models for data management (often referred to as NoSQL or “Not Only SQL”) provide a variety of methods for managing information to best suit specific business process needs, such as in-memory data management (for rapid access), columnar layouts to speed query response, and graph databases (for social network analytics)</td>
</tr>
<tr>
<td><strong>Resources</strong></td>
<td>Requires large capital investment in purchasing high-end hardware to be installed and managed in-house</td>
<td>The ability to deploy systems like Hadoop on virtualized platforms allows small and medium businesses to utilize cloud-based environments that, from both a cost accounting and a practical perspective, are much friendlier to the bottom line</td>
</tr>
</tbody>
</table>

“Big Data Analytics”, David Loshin, 2013
Comparison of Data Analytics and Computing Ecosystems

Application Level
- Mahout, R, and Applications
- Hive
- Pig
- Sqoop
- Flume
- Storm
- Hbase BigTable (key-value store)
- HDFS (Hadoop File System)

Middleware and Management
- Zookeeper (coordination)
- Cloud Services (e.g., AWS)
- Virtual Machines and Cloud Services (optional)
- MPI/OpenMP + Accelerator Tools
- Numerical Libraries
- Lustre (Parallel File System)
- Batch Scheduler (such as SLURM)
- System Monitoring Tools

System Software
- Linux OS variant

Cluster Hardware
- Ethernet Switches
- Local Node Storage
- Commodity X86 Racks
- Infiniband + Ethernet Switches
- SAN + Local Node Storage
- X86 Racks + GPUs or Accelerators

Data Analytics Ecosystem

Computational Science Ecosystem

Applications and Community Codes
- FORTRAN, C, C++, and IDEs
- Domain-specific Libraries
- Performance and Debugging (such as PAPI)
Chapter 1: Market and Business Drivers for Big Data Analysis
Chapter 2: Business Problems Suited to Big Data Analytics
Chapter 3: Achieving Organizational Alignment for Big Data Analytics
Chapter 4: Developing a Strategy for Integrating Big Data Analytics into the Enterprise
Chapter 5: Data Governance for Big Data Analytics: Considerations for Data Policies and Processes
Chapter 6: Introduction to High-Performance Appliances for Big Data Management
Chapter 7: Big Data Tools and Techniques
Chapter 8: Developing Big Data Applications
Chapter 9: NoSQL Data Management for Big Data
Chapter 10: Using Graph Analytics for Big Data
Chapter 11: Developing the Big Data Roadmap
Key Computing Resources for Big Data

- Processing capability: CPU, processor, or node.
- Memory
- Storage
- Network

“Big Data Analytics”, David Loshin, 2013
The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

http://hadoop.apache.org
Hadoop-related Apache Projects: Hadoop Ecosystem

- **Ambari™**: A web-based tool for provisioning, managing, and monitoring Hadoop clusters. It also provides a dashboard for viewing cluster health and ability to view MapReduce, Pig and Hive applications visually.
- **Avro™**: A data serialization system.
- **Cassandra™**: A scalable multi-master database with no single points of failure.
- **Chukwa™**: A data collection system for managing large distributed systems.
- **HBase™**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive™**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout™**: A scalable machine learning and data mining library.
- **Pig™**: A high-level data-flow language and execution framework for parallel computation.
- **Spark™**: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **Tez™**: A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases.
- **Zookeeper™**: A high-performance coordination service for distributed applications.
Hadoop Distributed File System (HDFS)

- HDFS is a java-based file system that provides the scalable, fault-tolerant, cost-efficient storage for big data
  - The file content is split into large blocks (typically 128 megabytes), each of which is independently replicated at multiple DataNodes
  - The NameNode maintains the namespace tree (in RAM) and the mapping of blocks to DataNodes

[Diagram of HDFS Architecture]

http://hortonworks.com/hadoop/hdfs/
MapReduce example

The overall MapReduce word count process

- Input
  - Deer Bear River
  - Car Car River
  - Deer Car Bear

- Splitting
  - Deer
  - Bear
  - River

- Mapping
  - Deer, 1
  - Bear, 1
  - River, 1
  - Car, 1
  - Car, 1
  - Car, 1
  - Deer, 1
  - Deer, 1
  - Deer, 1

- Shuffling
  - Bear, 1
  - Bear, 1
  - Car, 1
  - Car, 1
  - Car, 1
  - Deer, 1
  - Deer, 1
  - Deer, 1
  - River, 1
  - River, 1

- Reducing
  - Bear, 2
  - Car, 3
  - Deer, 2
  - River, 2

- Final result
  - Bear, 2
  - Car, 3
  - Deer, 2
  - River, 2

http://www.alex-hanna.com
MapReduce Data Flow

Map phase

Reduce phase

Apache Spark

Building on top of HDFS

Speed
Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing.

Ease of Use
Write applications quickly in Java, Scala or Python.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it interactively from the Scala and Python shells.

Generality
Combine SQL, streaming, and complex analytics.

Spark powers a stack of high-level tools including Spark SQL, MLlib for machine learning, GraphX, and Spark Streaming. You can combine these frameworks seamlessly in the same application.
NoSQL Database

- Key-Value Store
- Document Store
- Tabular Store
- Object Database
- Graph Database
  - Property graphs
  - Resource Description Framework (RDF) graphs
Key-Value Store

Example Data Represented in a Key–Value Store

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>“BMW”</td>
<td>{“1-Series”, “3-Series”, “5-Series”, “5-Series GT”, “7-Series”, “X3”, “X5”, “X6”, “Z4”}</td>
</tr>
<tr>
<td>“Buick”</td>
<td>{“Enclave”, “LaCrosse”, “Lucerne”, “Regal”}</td>
</tr>
</tbody>
</table>

- Get(key), which returns the value associated with the provided key.
- Put(key, value), which associates the value with the key.
- Multi-get(key1, key2,.., keyN), which returns the list of values associated with the list of keys.
- Delete(key), which removes the entry for the key from the data store.
The following diagram highlights the components of a MongoDB insert operation:

```javascript
db.users.insert({
    name: "sue",
    age: 26,
    status: "A"
});
```

The components of a MongoDB insert operation.

The following diagram shows the same query in SQL:

```
INSERT INTO users (name, age, status) VALUES ("sue", 26, "A")
```

The components of a SQL INSERT statement.

**Relational data model**
Highly-structured table organization with rigidly-defined data formats and record structure.

**Document data model**
Collection of complex documents with arbitrary, nested data formats and varying "record" format.
Graph Data

• Property Graphs
• RDF (Resource Description Framework: Triplestore) Graphs
What is the fundamental challenge for RDB on Linked Data?

In Relational DB, relationships are distributed. It takes a long time to JOIN to retrieve a graph from data.

Native Graph DB stores nodes and relationships directly. It makes retrieval efficient.

Retrieving multi-step relationships is a 'graph traversal' problem.

Cited “Graph Database” O’liey 2013.
Preliminary datastore comparison for Recommendation & Visualization

People who bought this also bought that...

Recommendation == 2-hop traversal & ranking

Visualization == 4-hop traversal & rankings
Google Trends on Relational vs Graph Databases (8/4/2014)

Trends of search interest on **Graph Database** and **Relational Database**, real-time from Google (Google Trend normalizes Y-axis to the highest value in a chart to 100%).

**Comparison of relative amounts of searches on Relational Database and Graph Database:**


View full report in Google Trends

View full report in Google Trends
How to Visualize Huge Static Graph

76425 species

Tree of Life by Dr. Yifan Hu

14.8 million tweets

The information diffusion graph of the death of Osama bin Laden by Gilad Lotan

500 million users

Facebook friendship graph by Paul Butler

**Challenging Task:**

Squeezing millions and even billions of records into million pixels (1600 X 1200 ≈ 2 million pixels)
Visualization Key Challenges

Visual clutter
How can we encode the information intuitively?

Performance issues
How can we render the huge datasets in real time with rich interactions?

Cognition
How can users understand the visual representation when the information is overwhelming?
Platform Dependent Graphical Models

Homogeneous multicore processors
Intel Xeon E5335 (Clovertown)
AMD Opteron 2347 (Barcelona)
Netezza (FPGA, multicore)

Homogeneous manycore processors
Sun UltraSPARC T2 (Niagara 2), GPGPU

Heterogeneous multicore processors
Cell Broadband Engine
Clusters
HPCC, DataStar, BlueGene, etc.
Graph Workload Types

- **Type 1**: Computations on graph structures / topologies
  - Example → converting Bayesian network into junction tree, graph traversal (BFS/DFS), etc.
  - Characteristics → Poor locality, irregular memory access, limited numeric operations

- **Type 2**: Computations on graphs with rich properties
  - Example → Belief propagation: diffuse information through a graph using statistical models
  - Characteristics
    - Locality and memory access pattern depend on vertex models
    - Typically a lot of numeric operations
    - Hybrid workload

- **Type 3**: Computations on dynamic graphs
  - Example → streaming graph clustering, incremental k-core, etc.
  - Characteristics
    - Poor locality, irregular memory access
    - Operations to update a model (e.g., cluster, sub-graph)
    - Hybrid workload
Large-scale graph benchmark – Graph 500

**November 2013**

<table>
<thead>
<tr>
<th>No.</th>
<th>Rank</th>
<th>Machine</th>
<th>Installation Site</th>
<th>Number of nodes</th>
<th>Number of cores</th>
<th>Problem scale</th>
<th>GTEPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>DOE/NNSA/LLNL Sequoia (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)</td>
<td>Lawrence Livermore National Laboratory</td>
<td>65536</td>
<td>1048576</td>
<td>40</td>
<td>15363</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>DOE/SC/Argonne National Laboratory Mira (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)</td>
<td>Argonne National Laboratory</td>
<td>49152</td>
<td>786432</td>
<td>40</td>
<td>14328</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>JUQUEEN (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)</td>
<td>Forschungszentrum Juelich (FZJ)</td>
<td>16384</td>
<td>262144</td>
<td>38</td>
<td>5848</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>K computer (Fujitsu - Custom supercomputer)</td>
<td>RIKEN Advanced Institute for Computational Science (AICS)</td>
<td>65536</td>
<td>524288</td>
<td>40</td>
<td>5524.12</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>Fermi (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)</td>
<td>CINECA</td>
<td>8192</td>
<td>131072</td>
<td>37</td>
<td>2567</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>Tianhe-2 (MilkyWay-2) (National University of Defense Technology - MPP)</td>
<td>Changsha, China</td>
<td>8192</td>
<td>196608</td>
<td>36</td>
<td>2061.48</td>
</tr>
</tbody>
</table>

IBM BlueGene or P775: 24 out of Top 30, except #4, #14, #22: Fujitsu #6: Tianhe #15: Cray #24: HP
Big Data Analytics Example Use Cases

1. Social Network Analysis
2. Recommendation
3. Commerce
4. Financial Analysis
5. Social Media Monitoring
6. Telco Customer Analysis
7. Watson
8. Data Exploration and Visualization
9. Personalized Search
10. Anomaly Detection (Espionage, Sabotage, etc.)
11. Fraud Detection
12. Cybersecurity
13. Sensor Monitoring (Smarter another Planet)
14. Cellular Network Monitoring
15. Cloud Monitoring
17. Traffic Navigation
18. Image and Video Semantic Understanding
19. Genomic Medicine
20. Brain Network Analysis
21. Data Curation
22. Near Earth Object Analysis
Use Case 1: Social Network Analysis in Enterprise for Productivity

**Production Live System used by IBM GBS since 2009 – verified ~$100M contribution**

- 15,000 contributors in 76 countries; 92,000 annual unique IBM users
- 25,000,000+ emails & SameTime messages (incl. Content features)
- 1,000,000+ Learning clicks; 14M KnowledgeView, SalesOne, ..., access data
- 1,000,000+ Lotus Connections (blogs, file sharing, bookmark) data
- 200,000 people’s consulting project & earning data

- On BusinessWeek four times, including being the Top Story of Week, April 2009
- Help IBM earned the 2012 Most Admired Knowledge Enterprise Award
- Wharton School study: $7,010 gain per user per year using the tool
- In 2012, contributing about 1/3 of GBS Practitioner Portal $228.5 million savings and benefits
- APQC (WW leader in Knowledge Practice) April 2013:
  “The Industry Leader and Best Practice in Expertise Location”

**Dynamic networks of 400,000+ IBMers:**
- Shortest Paths
- Social Capital
- Bridges
- Hubs
- Expertise Search
- Graph Search
- Graph Recomm.
Use Case 2: Recommendation

Hello, Ching Yung Lin. We have recommendations for you. (If you’re not Ching Yung Lin, click here.)

Recommended for you

- **Spikes** [Reprint] Paperback by Fred Rieke
  (Why is this recommended to me?)

- **Spiking Neuron Models** Paperback by Wulfram Gerstner
  (Why is this recommended to me?)

- **Methods in Neuronal Modeling - 2nd Edition** Hardcover by Christof Koch
  (Why is this recommended to me?)

See more Recommendations
Use Case 3: Recommendation for Commerce

Comparing to Collaborative Filtering (CF) + Similar People

- **Precision:** IF is 91% better, TIF is 108% better
- **Recall:** IF is 87% better, TIF is 113% better

Tests:
- 1 month
- 586 new docs
- 1,170 users
Use Case 4: Graph Analytics for Financial Analysis

**Goal:** Injecting Network Graph Effects for Financial Analysis. Estimating company performance considering correlated companies, network properties and evolutions, causal parameter analysis, etc.

- IBM 2003
- IBM 2009

**Data Source:**
- Relationships among 7594 companies, data mining from NYT 1981 ~ 2009

**Network feature:**
- s (current year network feature),
- t (temporal network feature),
- d (delta value of network feature)

**Financial feature:**
- p (historical profits and revenues)

**Targets:** 20 Fortune companies’ normalized profits

**Goal:** Learn from previous 5 years, and predict next year

**Model:** Support Vector Regression (RBF kernel)

Profit prediction by joint network and financial analysis outperforms network-only by 130% and financial-only by 33%.
Use Case 5: Social Media Monitoring

IBM CIO monitoring categories

Monitoring filter

Live Tweets, Sentiment, Keywords

Dynamic Graphs

Zooming / Panning

Real-Time Translation, Locations, Top Retweets

Total Tweets: 231
Positive: 37 / 15%
Negative: 194 / 85%

Select CIO Category(-ies): EXECDB, BLADE, HRTEGRITY, IBM, Security/Analysis, BLY, WATSON

or Word: Egypt

Ching-Yung Lin | Search www.ibm.com

System G SMISC Social Media Monitoring

Home | Live | Forensics

Research Projects | People | News

Monitoring filter

Language: Arabic

Page dimensions: 720.0x540.0
Use Case 6: Customer Social Analysis for Telco

**Goal:** Extract customer social network behaviors to enable Call Detail Records (CDRs) data monetization for Telco.

- **Applications based on the extracted social profiles**
  - Personalized advertisement (beyond the scope of traditional campaign in Telco)
  - High value customer identification and targeting
  - Viral marketing campaign

- **Approach**
  - Construct social graphs from CDRs based on {caller, callee, call time, call duration}
  - Extract customer social features (e.g. influence, communities, etc.) from the constructed social graph as customer social profiles
  - Build analytics applications (e.g. personalized advertisement) based on the extracted customer social profiles

PoCs with Chinese and Indian Telecomm companies
Use Case 7: Graph Analytics and Visualization for Watson

Query

Matches

(a)  (b)  (c)  (d)

Graph MATCHING

Graph Communities

headache  chill  migraine
high fever  stomachache

cancer  heart disease  HIV  diabetes  kidney disorder

42
User Case 8: Visualization for Navigation and Exploration

Cluster based huge graph visualization

Query based huge graph visualization

2-depth ego network of node id: 8:11
Use Case 9: Graph Search

existing search engine

query

Info-Socio networks

Graph analysis

query context

Graph analysis

re-ranking

index

ranking

re-ranking

Improved search results

Interest / social network based content recommendations

Practitioner Portal

Search criteria

Use "", AND or NOT for better results (default in phrases is AND), e.g. "HR" AND "Human Resource"

Top search terms, pages and tags

Search keywords: social business

18,877 results found

IBM Social Business Adoption QuickStart (U.S. English) - Proposal Insert [in Proposal and Presentation Accelerator (PPA)]

Drive the successful launch and adoption of social business software throughout your organization with a structured engagement comprised of assessments, planning and design consultation, onsite workshops, and team- and skills-building activities.

Sales Support Information (SSI)
Use Case 10: Anomaly Detection at Multiple Scales

Based on President Executive Order 13587

**Goal:** System for Detecting and Predicting Abnormal Behaviors in Organization, through large-scale social network & cognitive analytics and data mining, to decrease insider threats such as espionage, sabotage, colleague-shooting, suicide, etc.

“Enterprise Information Leakage Impacted economy and jobs” Feb 2013

“What's emerged is a multibillion dollar detective industry” npr Jan 10, 2013

**Infrastructure + ~ 70 Analytics**
Use Case 11: Fraud Detection for Bank

Ponzi scheme Detection

Normal:
1. Clique-like
2. Two-way links

Attacker: Near-Star

Network Info Flow

Ego Net Features
Use Case 12: Detecting Cyber Attacks

Detecting DoS attack

(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.
Use Case 13: Smarter *another* Planet

**Goal:** Atmospheric Radiation Measurement (ARM) climate research facility provides 24x7 *continuous field observations* of cloud, aerosol and radiative processes. *Graphical models* can automate the validation with improvement efficiency and performance.

**Approach:** BN is built to represent the dependence among sensors and replicated across timesteps. BN parameters are learned from over 15 *years* of ARM climate data to support distributed climate sensor validation. Inference validates sensors in the connected instruments.

**Bayesian Network**
* 3 timesteps  * 63 variables
* 3.9 avg states  * 4.0 avg indegree
* 16,858 CPT entries

**Junction Tree**
* 67 cliques
* 873,064 PT entries in cliques
Use Case 14: Cellular Network Analytics in Telco Operation

**Goal:** Efficiently and uniquely identify *internal* state of Cellular/Telco networks (e.g., performance and load of network elements/links) using probes between monitors placed at selected network elements & endhosts

- Applied Graph Analytics to telco network analytics based on CDRs (call detail records): estimate traffic load on CSP network with low monitoring overhead
  - CDRs, already collected for billing purposes, contain information about voice/data calls
  - Traditional NMS* and EMS** typically lack of end-to-end visibility and topology across vendors
  - Employ graph algorithms to analyze network elements which are not reported by the usage data from CDR information

**Approach**
- Cellular network comprises a hierarchy of network elements
- Map CDR onto network topology and infer load on each network element using graph analysis
- Estimate network load and localize potential problems
**Use Case 15: Monitoring Large Cloud**

**Goal:** Monitoring technology that can track the time-varying state (e.g., causality relationships between KPIs) of a large Cloud when the processing power of monitoring system cannot keep up with the scale of the system & the rate of change

- **Causality relationships (e.g., Granger causality) are crucial in performance monitoring & root cause analysis**
- **Challenge:** easy to test pairwise relationship, but hard to test multi-variate relationship (e.g., a large number of KPIs)

**Our approach:** Probabilistic monitoring via sampling & estimation

- **Select KPI pairs (sampling) → Test link existence → Estimate unsampled links based on history → Overall graph**
Use Case 16: Code Life Cycle Improvement

Advantages of working directly with graph DB for graph applications

- Smaller and simpler code
- Flexible schema → easy schema evolution
- Code is easier and faster to write, debug and manage
- Code and Data is easier to transfer and maintain
Use Case 17: Smart Navigation Utilizing Real-time Road Information

**Goal:** Enable unprecedented level of accuracy in **traffic scheduling** (for a fleet of transportation vehicles) and navigation of individual cars utilizing the **dynamic real-time information** of changing road condition and predictive analysis on the data.

- Dynamic graph algorithms implemented in System G provide **highly efficient graph query computation** (e.g. shortest path computation) on time-varying graphs (order of magnitudes improvement over existing solutions).

- High-throughput **real-time predictive analytics** on graph makes it possible to estimate the future traffic condition on the route to make sure that the decision taken now is optimal overall.

**Our approach:** Querying over dynamic graph + predictive analytics on graph properties.
Use Case 18: Graph Analysis for Image and Video Analysis
Use Case 19: Graph Matching for Genomic Medicine

- Ongoing discussions

Figure 1: Since the Human Genome Project, various projects have started to reveal the mysteries of genomes and the $1000 Genome is almost reality.
Use Case 20: Data Curation for Enterprise Data Management

Prior Collaborative Use

Semantic Knowledge

Extracted Metadata

Supervised Curation
Use Case 21: Understanding Brain Network
Use Case 22: Planet Security

• Big Data on Large-Scale Sky Monitoring

Dangers from space
Learn about the threat to Earth from asteroids & comets and how the Pan-STARRS project is designed to help detect these NEOs. Learn more...

1,400,000,000 pixels
Pan-STARRS has the world's largest digital cameras. Read about them here...

The PS1 Prototype
PS1 goes operational and begins science mission
PS1 Science Consortium formed...
PS1SC Blog
PS1 image gallery
Questions?