Midas: Scalable Entity Integration for Unstructured Data Sources

Rajasekar Krishnamurthy
IBM Research - Almaden
Why Entity Integration for Unstructured Data Sources?

Challenges in Scalable Entity Integration

Midas Financial Insights Demo
Entity View of the World

- **Data is prevalent**
  - **Business Data:**
    - Company filings to regulatory bodies
    - Security market (e.g., stock, fund, option) trading data
    - News articles, analyst reports, …
  - **Government Data:**
    - US federal government spending data, earmarks data
    - Congress data (voting, members, …)

- **Users and applications prefer an entity view of the underlying data**
  - Entities (Companies, People, Securities, …)
  - Relationships (Employment, Investment, Ownership, …)
  - Events (Mergers, Acquisitions, Bankruptcy, Appointment, …)
Sample questions posed over the entity view

Questions posed over Business Data

- Which public companies currently share one or more board member?
- Which high-level federal government officials moved between federal government and industry recently?
- How has Berkshire Hathaway’s investment profile changed recently?
- How has bank lending to small businesses changed over time?
- Which companies do business together a lot (e.g., banks making joint loans to other large institutions)?

Questions posed over Government Data

- What is spending by each government department for each geographic region?
- How many (or total value) earmarks in 2009 were solely sponsored by Republican (or Democrat) congress members?
- Who are the Top k congress members with the most number of earmarks tied to the Department of Defense?
Why do we need entity integration?

- A significant portion of business data is in unstructured format
- Financial service firms use manual methods to analyze regulatory files, news articles etc.
  - Error-prone, cost ineffective, not scalable

Errors made while manually converting corporate actions information into electronic format annually cost financial services firms from $400 million to $900 million a year, according to U.K. consulting firm Oxera. However, a company’s responsibility ends once it releases its announcement and makes any required filings. “The issuer is obviously concerned if there is an error in the redistribution of that information, but the fact is, any redistribution is not the issuer’s responsibility,” Morgan said, adding that company press releases are not regulated so companies won’t be required to put them in XBRL.

Source: http://www.treasuryandrisk.com/News/Pages/Corporate-Actions-Reporting-in-XBRL-Crosshairs.aspx

- Users have to “search” data sources to obtain answers
  - Hard to answer questions that need to
    - combine facts from multiple places of the same data source or multiple data sources, or
    - aggregate all data with certain properties
  - For each question asked, manual post-processing of related facts is needed
Business Data is exchanged frequently in unstructured format

Example Scenario: Corporate Action flow in the U.S. market

Unstructured data (text/html/paper formats) are predominant

Source: XBRL Pacific Rim Workshop, 2009
http://xbrl.us/events/Documents/PacificRim/CorporateActions_Hands.pdf
Complications in understanding multiple “related” facts

Purchase numbers and number of transactions differ across aggregate report and individual transactions list

- Aggregate Data and individual transactions list provided by different data providers!!
- Possible semantic differences on what is a purchase across the data providers
- Understanding multiple facts can be complicated even on a single document!!

Source: http://finance.yahoo.com/q/it?s=ORCL+Insider+Transactions
Aug 13, 2010
The Value of Entity Integration

- Obtain an entity view of the world
- Entities, relationships and events are represented as structured objects
- Can answer complex questions over the Integrated Entity Data:
  - Which public companies currently share one or more board member?
  - Which high-level federal government officials moved between federal government and industry recently?
  - How has Berkshire Hathaway’s investment profile changed recently?
Example: Government and Corporate Positions

Questions we can answer by extracting the employment history of key financial officers:

- What are the (past) government positions held by directors of different companies?
- What is the employment history of key government officers?
- How significant is the interlock between companies receiving TARP funding and government officers?

Technical Challenges:

- Extraction of current/past positions from biographies and appointment/resignation of special officers
- Entity resolution of person and companies mentioned in the biographies (and other locations)
- Fusion of employment history,
- Selection of government positions from the employment records

Examples:

- Robert E. Rubin: Former Secretary of the Treasury and Former officer in both Citigroup and Goldman Sachs.
- Arthur Levitt: Former Chairman of the SEC and officer in AIG.

Citigroup’s Proxy Statement filed on April 22 2008

Robert E. Rubin

AIG’s Proxy Statement filed on April 05 2006

In July 2005, Arthur Levitt, a former Chairman of the Securities and Exchange Commission, was named a special advisor to the Board of Directors and the Committee. Mr. Levitt has worked closely with the Committee advising on corporate governance initiatives and possible independent director nominees.
How to Bridge the Gap?
Multiple Raw Unstructured Datasets ➔ Consolidated Entities?

Core Technology Requirements for Understanding Unstructured Data

- **Information Extraction**
  - Information is present in multiple formats (e.g., Text, XML, HTML)
  - Extract entities, events, relationships from unstructured documents
  - Unstructured data ➔ Structured data

- **Entity Integration**
  - Resolving mentions to same real-world entity across filings
  - Normalize and cleanse extracted values
  - Aggregate related facts extracted from multiple filings

- **Scalable Architecture**
  - Millions of documents of varying size and format
  - New documents arrive daily

Above challenges to handle unstructured data sources are complementary to issues discussed in earlier talk by Prof. Felix Naumann on “Web Data Integration”
Why Entity Integration for Unstructured Data Sources?

Challenges in Scalable Entity Integration

Midas Financial Insights Demo
Midas Architecture: A Detailed View

**Core Extraction & Integration Technology:**
- Developed over 5 years in IBM Research
- Deployed and validated in multiple IBM products

**Platform:**
- Integrate core technology with Hadoop
- Drive large volume of data through extract and integrate stages
- Refresh incrementally and continuously

**Domain-Specific Applications**
- Healthcare
- Finance
- Telecom
- Government
- Insurance

**Midas Extraction & Integration Flow**
- Crawl
- Extract
- Local Resolve
- Global Resolve
- Map/Fuse
- Temporal Analyze

**Platform**
- Nutch
- Jaql + SystemT
- Hadoop (Map/Reduce)
- Distributed File System
- DB Export
- Index Generation

**Applications**
- Analytic + Search UI
- Cognos Reports

**Index**
- Analytic DB
Use case: Midas Financial Insight
Entity Integration Over Financial Data

Other Data
- Financial Information Providers
  - News
  - Internet – blogs, online discussions, boards

Public Data
- FDIC Call Data Records
- SEC Filings
- OTS Thrift Financial Records

Controlled Data
- Specific data requests between transaction parties

Midas Financial Insight

- Extraction and cleansing of financial data and linking information across multiple sources
- Uncovering non-obvious relationships between organizations
- Computation of key financial metrics using data extracted from multiple sources of public data

- Loan officers
- Credit Committees
- Regulatory analyst
- Analyst for financial data services
- Investment Banker
- Individual investor

Internet – blogs, online discussions, boards

SEC Filings

Use case: Midas Financial Insight
Entity Integration Over Financial Data
Information Extraction

Midas
Extraction & Integration Flow

- Crawl
- Extract
- Local Resolve
- Global Resolve
- Map/Fuse
- Temporal Analyze
Example: Extraction of loan information data

Extract and cleanse information from headers, tables, main content, and signatures.

Loan Information

<table>
<thead>
<tr>
<th>Id</th>
<th>Agreement Name</th>
<th>Date</th>
<th>Total Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Credit Agreement</td>
<td>June 12, 2009</td>
<td>$800,000,000</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Documents filed by Charles Schwab Corporation On Aug 6, 2009

Loan Company Information

<table>
<thead>
<tr>
<th>Id</th>
<th>Company</th>
<th>Role</th>
<th>Commitment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Citibank, N.A.</td>
<td>Administrative Agent</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Citibank, N.A.</td>
<td>Lender</td>
<td>$90,000,000</td>
</tr>
<tr>
<td>1</td>
<td>JPMorgan Chase Bank, N.A.</td>
<td>Lender</td>
<td>$90,000,000</td>
</tr>
<tr>
<td>1</td>
<td>Bank of America, N.A.</td>
<td>Lender</td>
<td>$80,000,000</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: Extraction of person information across documents

Do these filings refer to the same person?
- variability in the person’s name
- lack of a key identifier
- supporting attributes vary depending on the context (form type)

Sincerely,

James Dimon,
Chairman and Chief Executive Officer of JPMorgan Chase. Director since 2000.

Who Is James Dimon?

<table>
<thead>
<tr>
<th>Committee</th>
<th>Member</th>
<th>Member</th>
<th>Member</th>
<th>Member</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compensation &amp; Management Development</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corporate Governance &amp; Nominating</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public Responsibilities</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk Policy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Grant date</th>
<th>Approval date</th>
<th>Stock options (in)</th>
<th>Option awards</th>
</tr>
</thead>
<tbody>
<tr>
<td>James Dimon</td>
<td>1/22/2008</td>
<td>1/15/2008</td>
<td>364,048</td>
<td>$14,500,000</td>
</tr>
<tr>
<td></td>
<td>1/22/2008</td>
<td>1/15/2008</td>
<td>2,000,000</td>
<td>$39,830</td>
</tr>
<tr>
<td></td>
<td>2,000,000</td>
<td>19,888,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Michael J. Cavanagh (+)</td>
<td>1/22/2008</td>
<td>1/15/2008</td>
<td>94,151</td>
<td>3,750,000</td>
</tr>
<tr>
<td></td>
<td>1/22/2008</td>
<td>1/15/2008</td>
<td>300,000</td>
<td>39,830</td>
</tr>
<tr>
<td></td>
<td>1/30/2008</td>
<td>N/A</td>
<td>54,271</td>
<td>47,833</td>
</tr>
<tr>
<td></td>
<td>1/30/2008</td>
<td>N/A</td>
<td>10,623</td>
<td>47,833</td>
</tr>
<tr>
<td>Frank J. Bisignano</td>
<td>1/22/2008</td>
<td>1/15/2008</td>
<td>94,151</td>
<td>3,750,000</td>
</tr>
<tr>
<td></td>
<td>1/22/2008</td>
<td>1/15/2008</td>
<td>300,000</td>
<td>39,830</td>
</tr>
<tr>
<td></td>
<td>2,980,200</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mr. Dimon became Chairman of the Board on December 31, 2006, and has been Chief Executive Officer and President since December 31, 2005. He had been President and Chief Operating Officer since JPMorgan Chase’s merger with Bank One Corporation in July 2004. At Bank One he had been Chairman and Chief Executive Officer since March 2000. Mr. Dimon is a graduate of Tufts University and received an MBA from Harvard Business School. He is a director of The College Fund UNCF and serves on the Board of Directors of The Federal Reserve Bank of New York, The National Center on Addiction and Substance Abuse, Harvard Business School and Catalyst. He is on the Board of Trustees of New York University School of Medicine.
Entity Integration
Resolving Person Names: An Example

1. Build an authoritative list of insider names for each company based on insider filings
2. Compare extracted name references from other filings to entries in the list and merge data to the closest match.

<table>
<thead>
<tr>
<th>Name and age</th>
<th>Position, principal occupation, business experience and directorships</th>
</tr>
</thead>
<tbody>
<tr>
<td>John A. Thain (52)</td>
<td>Chairman of the Board and Chief Executive Officer of Merrill Lynch &amp; Co., Director since December 2007, Chairman of the Board and Chief Executive Officer of Merrill Lynch &amp; Co., Chief Executive Officer of NYSE Euronext, Inc., and its predecessors, which operates securities exchanges and offers financial products and services, from 2001 to 2007, President (from 1998 to 2004), Chief Operating Officer (from 2003 to 2004) (from 1999 to 2003) of The Goldman Sachs Group, Inc., a financial services company, Other Public Company Directorships: BlackRock, Inc.</td>
</tr>
</tbody>
</table>

However, not all real-world cases are as simple...

List of insiders for Bank of America & Merrill Lynch

... BANS, KEITH T. BRAMBLE, FRANK P. SR. COLBERT VIRGID GIFFORD, CHARLES K. HAMMONDS, BRUCE L. HANCE, JAMES H. JR. LEWIS, KENNETH D. MONTAG, THOMAS K. MOYNIHAN, BRIAN T. PRUEHER, JOSEPH W. ROSSOTTI, CHARLES O. SARLES, H. JAY SLOAN, O. TEMPLE JR. TILLMAN, ROBERT L THAIN, JOHN A. ...
How do we match the partial names with the corresponding correct directors?

- We need to use additional attributes like position, gender and time period.
- We capture these “matching” semantics as “rules”. For example,

> “IF the names partially match AND the dates of the position match, THEN link the extracted data to the known director.”
Mapping, Temporal Analysis and Fusion: Creating Person Entities

Sample data records for John Thain extracted from various sources

Transactions by John Thain

<table>
<thead>
<tr>
<th>personName</th>
<th>Thain John A.</th>
</tr>
</thead>
<tbody>
<tr>
<td>cik</td>
<td>0001090355</td>
</tr>
<tr>
<td>filingDate</td>
<td>2008-01-24</td>
</tr>
<tr>
<td>reportingDate</td>
<td>2008-01-24</td>
</tr>
<tr>
<td>issuer</td>
<td>BlackRock Inc.</td>
</tr>
<tr>
<td>isOfficer</td>
<td>false</td>
</tr>
<tr>
<td>isDirector</td>
<td>true</td>
</tr>
<tr>
<td>fillingType</td>
<td>3</td>
</tr>
</tbody>
</table>

Appointment announcement

<table>
<thead>
<tr>
<th>personName</th>
<th>John Thain</th>
</tr>
</thead>
<tbody>
<tr>
<td>cik</td>
<td>0001090355</td>
</tr>
<tr>
<td>appointmentDate</td>
<td>2008-01-16</td>
</tr>
<tr>
<td>filer</td>
<td>BlackRock Inc.</td>
</tr>
<tr>
<td>appointedAs</td>
<td>Director</td>
</tr>
<tr>
<td>filingType</td>
<td>8-K</td>
</tr>
</tbody>
</table>

Committee membership

<table>
<thead>
<tr>
<th>personName</th>
<th>John Thain</th>
</tr>
</thead>
<tbody>
<tr>
<td>cik</td>
<td>0001090355</td>
</tr>
<tr>
<td>reportingDate</td>
<td>2008-06-30</td>
</tr>
<tr>
<td>company</td>
<td>Merrill Lynch</td>
</tr>
<tr>
<td>title</td>
<td>Chairman and CEO</td>
</tr>
<tr>
<td>member</td>
<td>Audit Committee</td>
</tr>
<tr>
<td>filingType</td>
<td>DEF14A</td>
</tr>
</tbody>
</table>

Need to combine data into a desired structure

- Data extracted varies in structure and values!

Entity Integration

Person Entity

<table>
<thead>
<tr>
<th>cik</th>
<th>0001090355</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>John A. Thain</td>
</tr>
<tr>
<td>company</td>
<td>Bank of America</td>
</tr>
<tr>
<td>employmentHistory:</td>
<td></td>
</tr>
<tr>
<td>{ Merrill Lynch, CEO, 2008-06-30, ... }</td>
<td></td>
</tr>
<tr>
<td>{ Black Rock, Director, 2008-01-16, ... }</td>
<td></td>
</tr>
<tr>
<td>{ Black Rock, Director, 2008-01-24, ... }</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Mapping, Temporal Analysis and Fusion: Computing Current Holdings

owner_cik: 0001179111,
owner_name: "John Deutch"

recent holdings:

{ directOrIndirectOwnership: "D",
  reportingDate: "2008-10-22",
  securityTitle: "Common Stock",
  shares: 70865.9,
  type: "nonDerivative"
}

{ directOrIndirectOwnership: "I",
  natureOfOwnership: "Deferred Shares – Compensation Plan for Non-Employee Directors",
  reportingDate: "2008-10-22",
  securityTitle: "Common Stock",
  shares: 8971,
  type: "nonDerivative"
}

{ directOrIndirectOwnership: "D",
  reportingDate: "2008-11-13",
  securityTitle: "Common Stock",
  shares: 70865.9,
  type: "nonDerivative"
}

{ directOrIndirectOwnership: "I",
  natureOfOwnership: "See Footnote (1).",
  reportingDate: "2008-10-22",
  securityTitle: "Common Stock",
  shares: 8971,
  type: "nonDerivative"
}

…

current_holdings_by_insider:

{ owner_cik: 0001179111,
  owner_name: "John Deutch"
  holdings:

  { directOrIndirectOwnership: "D",
    mostRecentDate: "2008-11-13",
    securityTitle: "Common Stock",
    shares: 70865.9,
    type: "nonDerivative"
  }

  { directOrIndirectOwnership: "I",
    natureOfOwnership: "Deferred Shares – Compensation Plan for Non-Employee Directors"
    mostRecentDate: "2009-07-24",
    securityTitle: "Common Stock",
    shares: 9227.1,
    type: "nonDerivative"
  }

  …

– some of the key information identifying the type of holding may be in a footnote

Must recognize when we have the same type of holding and then take the most recent value.
Midas Architecture: A Detailed View

**Core Extraction & Integration Technology:**
- Developed over 5 years in IBM Research
- Deployed and validated in multiple IBM products

**Platform:**
- Integrate core technology with Hadoop
- Drive large volume of data through extract and integrate stages
- Refresh incrementally and continuously

**Domain-Specific Applications**
- Healthcare
- Finance
- Telecom
- Government
- Insurance

**Midas Extraction & Integration Flow**
- Crawl
- Extract
- Local Resolve
  - Global Resolve
  - Map/Fuse
  - Temporal Analyze

**Platform**
- Nutch
- Jaql + SystemT
- Hadoop (Map/Reduce)
- Distributed File System
- DB Export
- Index Generation

**Applications**
- Analytic + Search UI
- Cognos Reports
- Index
- Analytic DB
Handling Scalability

- **Scalability Challenges**
  - Large document corpora
    - Millions of documents of different formats and document types
    - Documents vary in size (10KB – 10MB each)
    - New documents available daily
  - Maintaining a complex analysis pipeline
    - Some document types require specialized analysis
    - New analysis needs to be incorporated incrementally
    - Semi-structured results
  - Process data updates incrementally
    - Some analysis stages support incremental updates, while other stages may need to run over the entire data.
  - Tolerance to errors
    - A failure when processing a document should not be fatal to the overall flow

- **Scalable Platform on Cloud Infrastructure**
  - **Jaql**: *Declarative language for expressing transformations over semi-structured data*
  - **SystemT**: *High-performance declarative rule-based information extraction system*
- Why Entity Integration for Unstructured Data Sources?
- Challenges in Scalable Entity Integration
- Midas Financial Insights Demo
Midas: Financial Insight
Scale of current running prototype

Over 1 Million documents

Filing timeline

2005 2010

Filings of Financial Companies

Extract Integrate

Person

Company

Over 32000 key officials in financial companies

Over 2200 financial companies

FDIC Call Data

SEC Filings
Summary: Research Challenges

- Information Extraction from Text
  - Extracting entities, events, relationships from text and html documents

- Entity Integration
  - Resolving mentions to same real-world entity across filings
  - Normalize and cleanse extracted values
  - Aggregate related facts extracted from multiple filings

- Scalable architecture leveraging Cloud technology
  - Complex analysis over millions of documents in a scalable manner

- Tooling & Programmability
  - Enabling easier definition, deployment and customization of Entity Integration Flows