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.

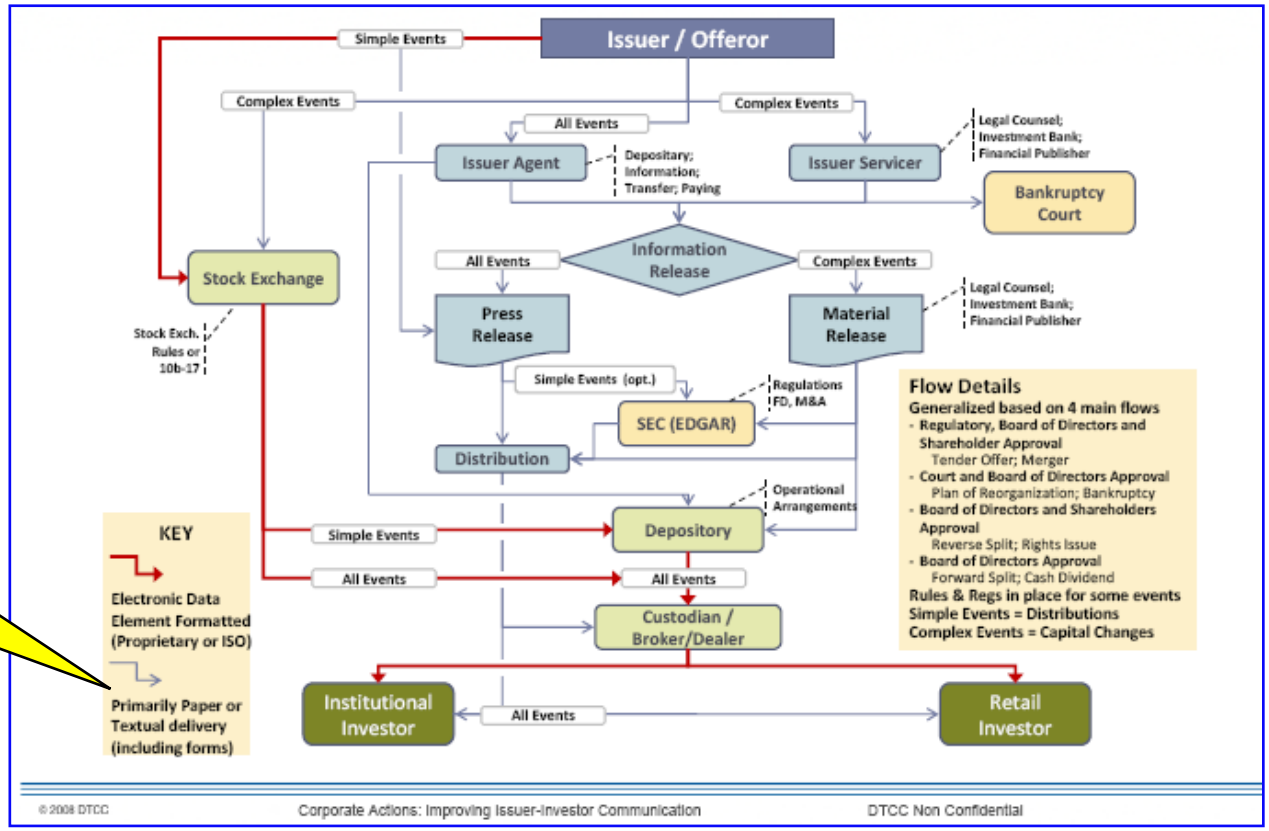
Source: http://www.treasuryandrisk.com/News/Pages/Corporate-Actions-Reporting-in-XBRL-Crosshairs.aspx

- 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

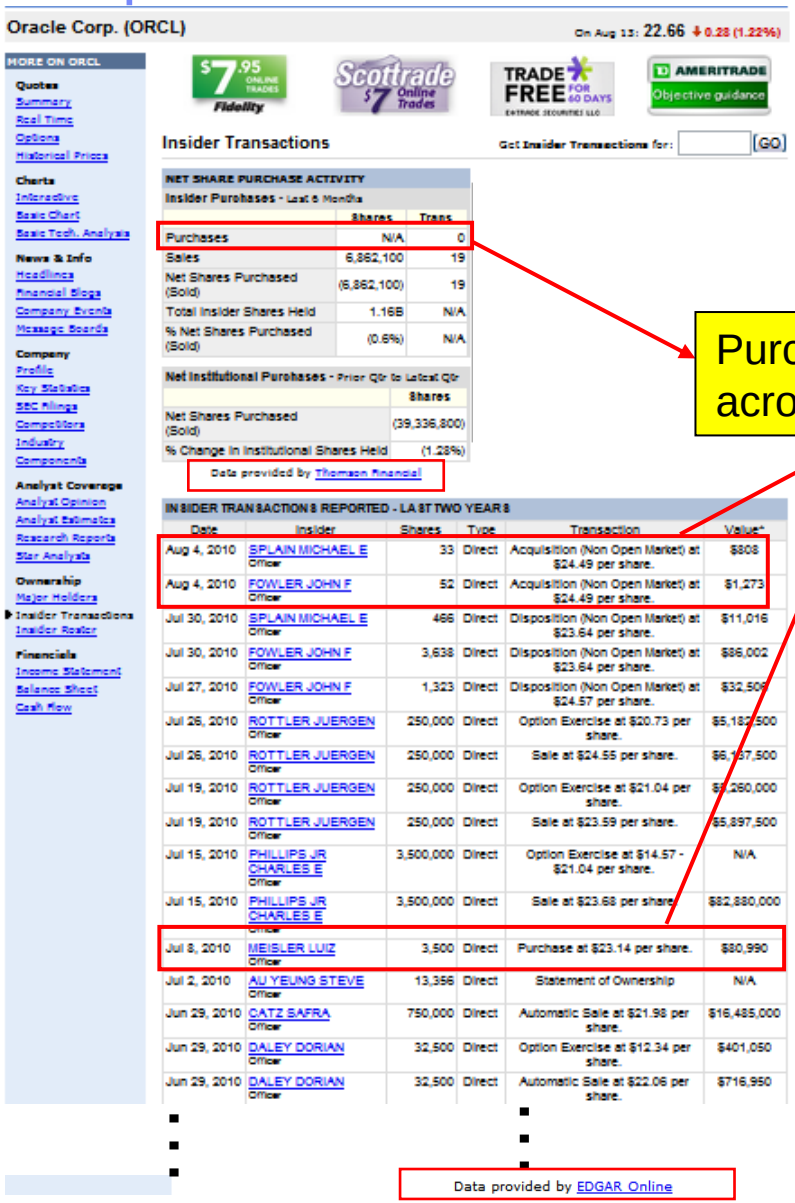
Example Scenario : Corporate Action flow in the U.S. market

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



Oracle Corp. (ORCL)

On Aug 13: 22.66 ↓0.28 (1.22%)

MORE ON ORCL

Quotes
Summary
Real Time
Options
Historical Prices

Charts
Interactive
Basic Chart
Basic Tech. Analysis

News & Info
Headlines
Financial Blogs
Company Events
Message Boards

Company
Profile
Key Statistics
SEC Filings
Competitors
Industry
Components

Analyst Coverage
Analyst Opinion
Analyst Estimates
Research Reports
Star Analysts

Ownership
Major Holders
Insider Transactions
Insider Roster

Financials
Income Statement
Balance Sheet
Cash Flow

Insider Transactions

Get Insider Transactions for: [GO]

NET SHARE PURCHASE ACTIVITY

Insider Purchases - Last 6 Months

	Shares	Trans
Purchases	N/A	0
Sales	6,862,100	19
Net Shares Purchased (Sold)	(6,862,100)	19
Total Insider Shares Held	1.16B	N/A
% Net Shares Purchased (Sold)	(0.6%)	N/A

Net Institutional Purchases - Prior Qtr to Latest Qtr

	Shares
Net Shares Purchased (Sold)	(39,336,800)
% Change in Institutional Shares Held	(1.28%)

Data provided by Thomson Financial

INSIDER TRANSACTIONS REPORTED - LAST TWO YEARS

Date	Insider	Shares	Type	Transaction	Value*
Aug 4, 2010	SPLAIN MICHAEL E Officer	33	Direct	Acquisition (Non Open Market) at $24.49 per share.	$808
Aug 4, 2010	FOWLER JOHN F Officer	52	Direct	Acquisition (Non Open Market) at $24.49 per share.	$1,273
Jul 30, 2010	SPLAIN MICHAEL E Officer	466	Direct	Disposition (Non Open Market) at $23.64 per share.	$11,016
Jul 30, 2010	FOWLER JOHN F Officer	3,638	Direct	Disposition (Non Open Market) at $23.64 per share.	$86,002
Jul 27, 2010	FOWLER JOHN F Officer	1,323	Direct	Disposition (Non Open Market) at $24.57 per share.	$32,508
Jul 26, 2010	ROTTLER JUERGEN Officer	250,000	Direct	Option Exercise at $20.73 per share.	$5,182,500
Jul 26, 2010	ROTTLER JUERGEN Officer	250,000	Direct	Sale at $24.55 per share.	$6,137,500
Jul 19, 2010	ROTTLER JUERGEN Officer	250,000	Direct	Option Exercise at $21.04 per share.	$5,260,000
Jul 19, 2010	ROTTLER JUERGEN Officer	250,000	Direct	Sale at $23.59 per share.	$5,897,500
Jul 15, 2010	PHILLIPS JR CHARLES E Officer	3,500,000	Direct	Option Exercise at $14.57 - $21.04 per share.	N/A
Jul 15, 2010	PHILLIPS JR CHARLES E Officer	3,500,000	Direct	Sale at $23.68 per share.	$82,880,000
Jul 8, 2010	MEISLER LUIZ Officer	3,500	Direct	Purchase at $23.14 per share.	$80,990
Jul 2, 2010	AU YEUNG STEVE Officer	13,356	Direct	Statement of Ownership	N/A
Jun 29, 2010	CATZ SAFRA Officer	750,000	Direct	Automatic Sale at $21.98 per share.	$16,485,000
Jun 29, 2010	DALEY DORIAN Officer	32,500	Direct	Option Exercise at $12.34 per share.	$401,050
Jun 29, 2010	DALEY DORIAN Officer	32,500	Direct	Automatic Sale at $22.06 per share.	$716,950

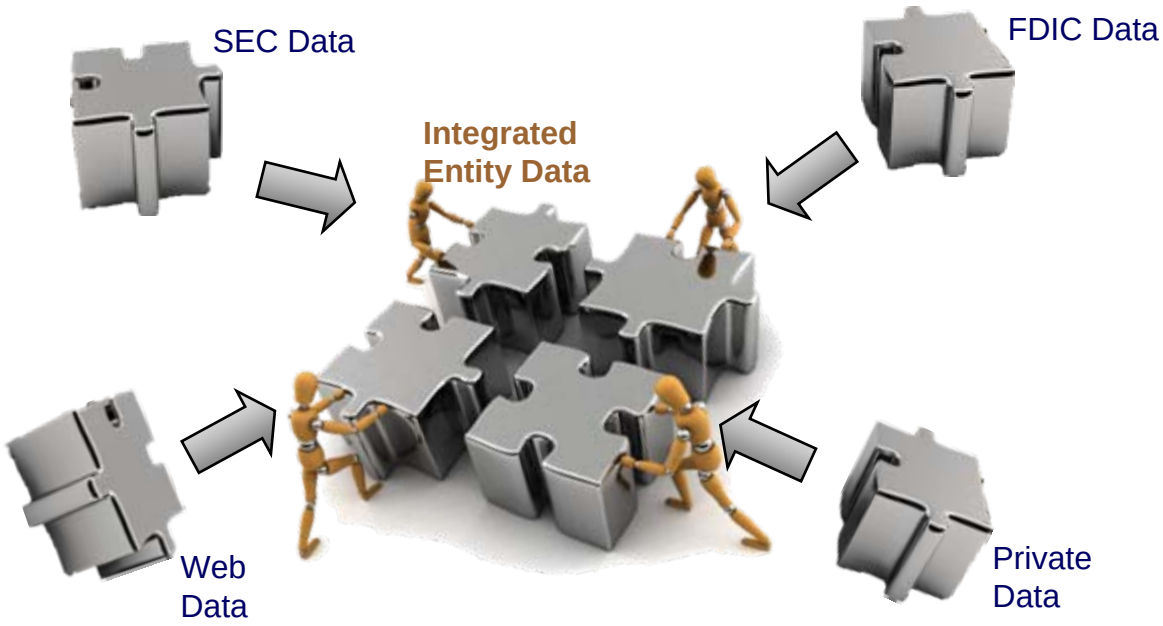
Data provided by EDGAR Online

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?

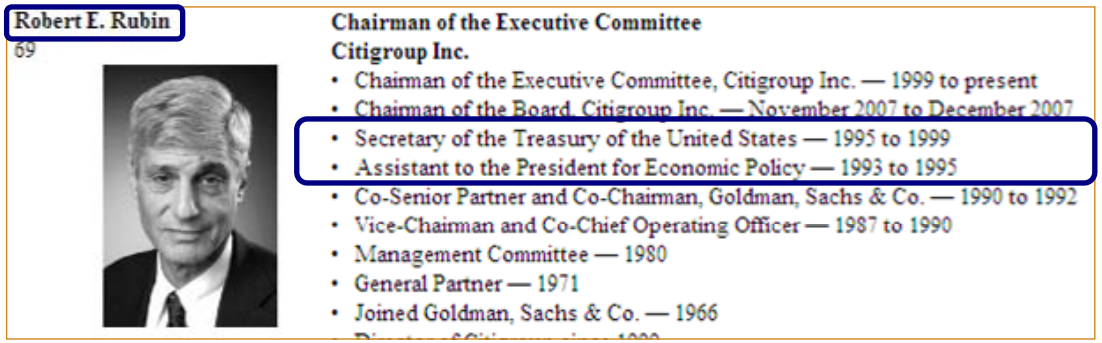
Technical Challenges:

- 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
69

Chairman of the Executive Committee
Citigroup Inc.

- Chairman of the Executive Committee, Citigroup Inc. — 1999 to present
- Chairman of the Board, Citigroup Inc. — November 2007 to December 2007
- Secretary of the Treasury of the United States — 1995 to 1999
- Assistant to the President for Economic Policy — 1993 to 1995
- Co-Senior Partner and Co-Chairman, Goldman, Sachs & Co. — 1990 to 1992
- Vice-Chairman and Co-Chief Operating Officer — 1987 to 1990
- Management Committee — 1980
- General Partner — 1971
- Joined Goldman, Sachs & Co. — 1966

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

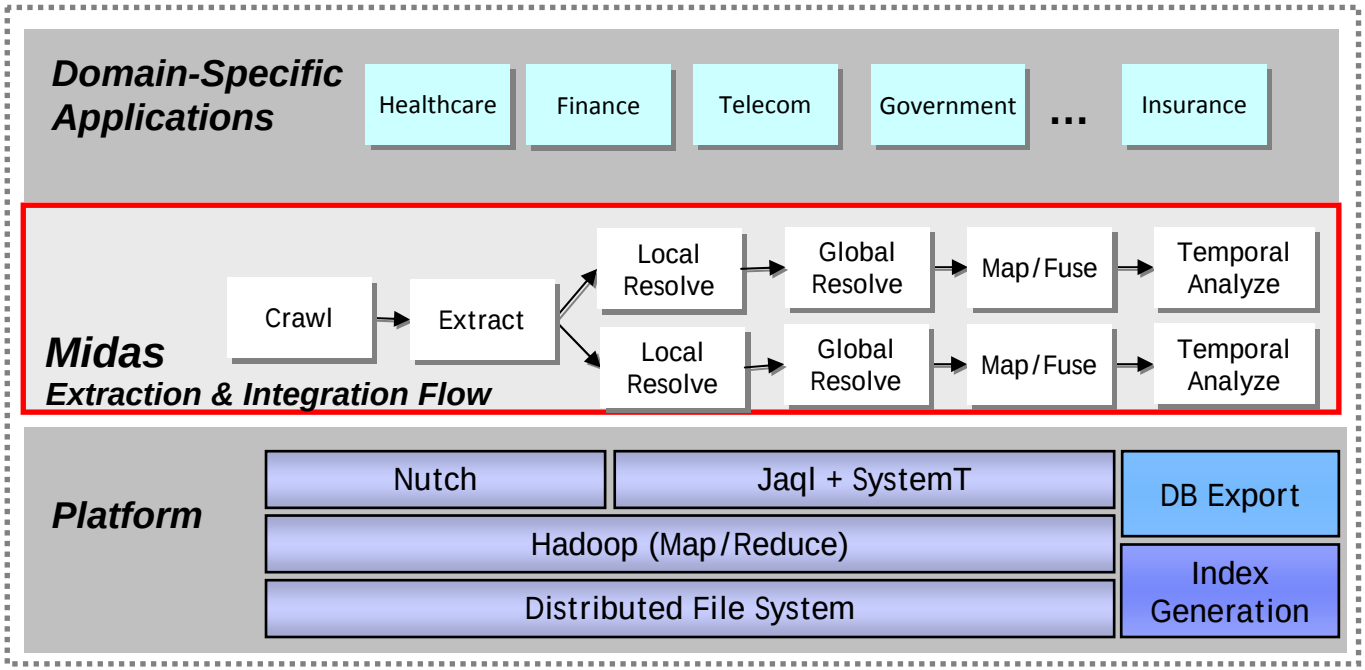
Midas Architecture: A Detailed View

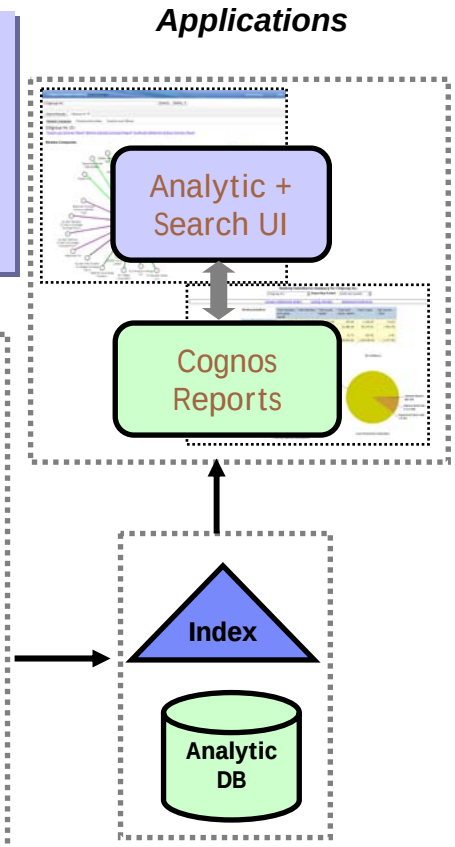
Core Extraction & Integration Technology:

- 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





Use case : Midas Financial Insight
Entity Integration Over Financial Data

Other Data

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

- 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

Public Data

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

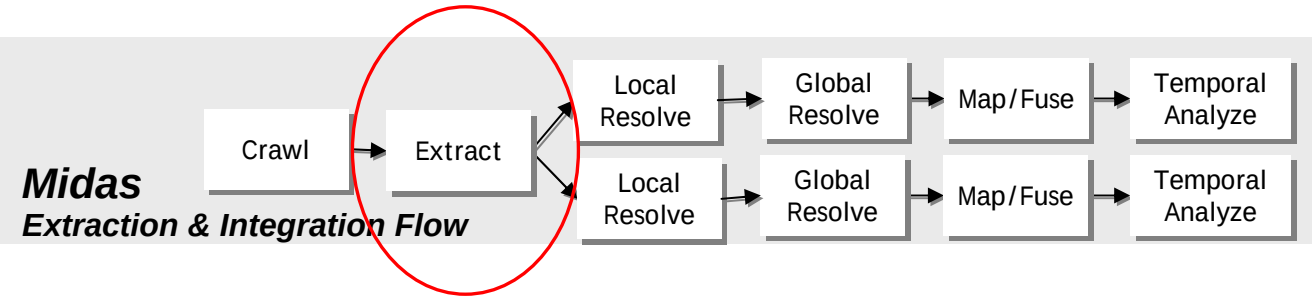
Midas Financial Insight

Controlled Data

- Specific data requests between transaction parties

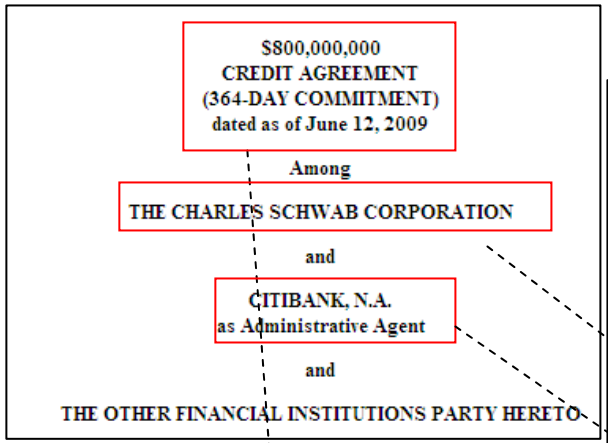
- Loan officers
- Credit Committees
- Regulatory analyst
- Analyst for financial data services
- Investment Banker
- Individual investor

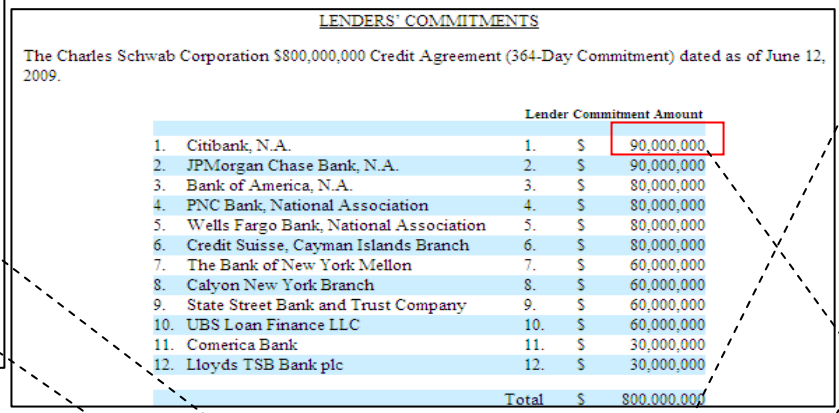
❑ Information Extraction



Example: Extraction of loan information data

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

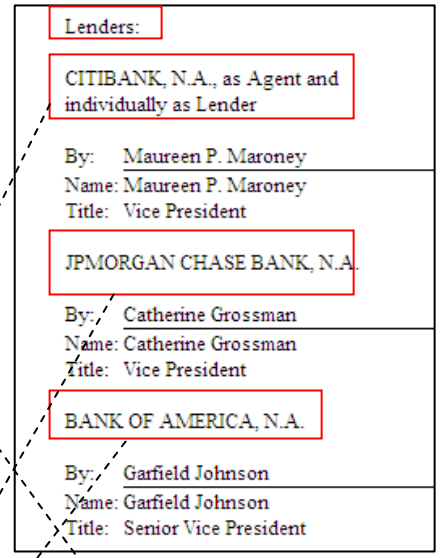




LENDERS' COMMITMENTS

The Charles Schwab Corporation $800,000,000 Credit Agreement (364-Day Commitment) dated as of June 12, 2009.

		Lender Commitment Amount
1.	Citibank, N.A.	1. $ 90,000,000
2.	JPMorgan Chase Bank, N.A.	2. $ 90,000,000
3.	Bank of America, N.A.	3. $ 80,000,000
4.	PNC Bank, National Association	4. $ 80,000,000
5.	Wells Fargo Bank, National Association	5. $ 80,000,000
6.	Credit Suisse, Cayman Islands Branch	6. $ 80,000,000
7.	The Bank of New York Mellon	7. $ 60,000,000
8.	Calyon New York Branch	8. $ 60,000,000
9.	State Street Bank and Trust Company	9. $ 60,000,000
10.	UBS Loan Finance LLC	10. $ 60,000,000
11.	Comerica Bank	11. $ 30,000,000
12.	Lloyds TSB Bank plc	12. $ 30,000,000
		Total $ 800,000,000



Id	Agreement Name	Date	Total Amount
1	Credit Agreement	June 12, 2009	$800,000,000
...			

Loan Information

Id	Company	Role	Commitment
1	Charles Schwab Corporation	Borrower	
1	Citibank, N.A.	Administrative Agent	
1	Citibank, N.A.	Lender	$90,000,000
1	JPMorgan Chase Bank, N.A.	Lender	$90,000,000
1	Bank of America, N.A.	Lender	$80,000,000
...			

Loan Company Information

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

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)

Who Is James Dimon?

position

history

committee membership

Sincerely,

James Dimon
Chairman and Chief Executive Off

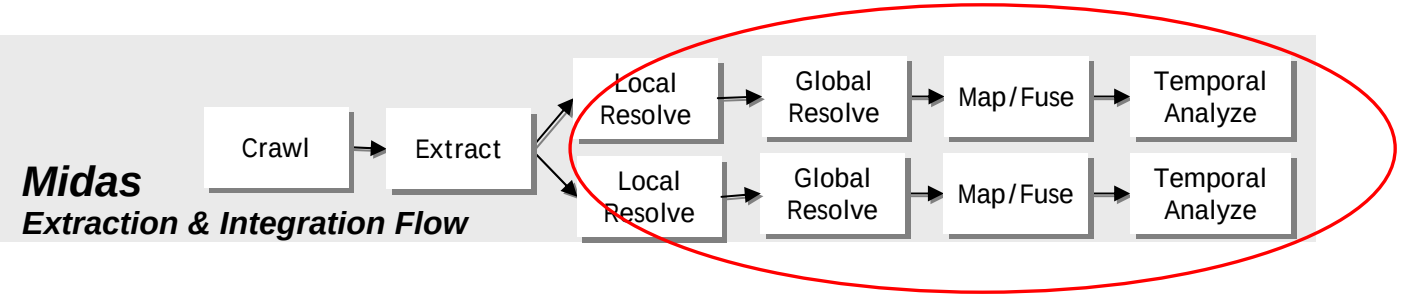
Director	Audit	Compensation & Management Development	Corporate Governance & Nominating	Public Responsibility	Risk Policy
Crandall C. Bowles	Member				
Stephen B. Burke		Member	Member		
David M. Cote				Member	Member
James S. Crown				Member	Chair
James Dimon					
Ellen V. Futter				Member	Member

Name	Grant date	Approval date	Stock awards: Number of shares of stock or units (#) (1)	Option awards: Number of securities underlying options (#)	Option awards: Exercise price ($/Sh)	Grant date fair value ($)
James Dimon	1/22/2008	1/15/2008	364,048			$14,500,000
	1/22/2008 (3)	1/15/2008		2,000,000 (3)	$39.830	19,868,000 (3)
Michael J. Cavanagh (4)	1/22/2008	1/15/2008	94,151			3,750,000
	1/22/2008	1/15/2008		300,000	39.830	2,980,200
	1/30/2008	N/A		54,271	47.835	360,577
	1/30/2008	N/A		10,625	47.835	91,821
Frank J. Bisignano	1/22/2008	1/15/2008	94,151			3,750,000
	1/22/2008	1/15/2008		300,000	39.830	2,980,200

James Dimon, 53, Chairman and Chief Executive Officer of JPMorgan Chase. Director since 2000.

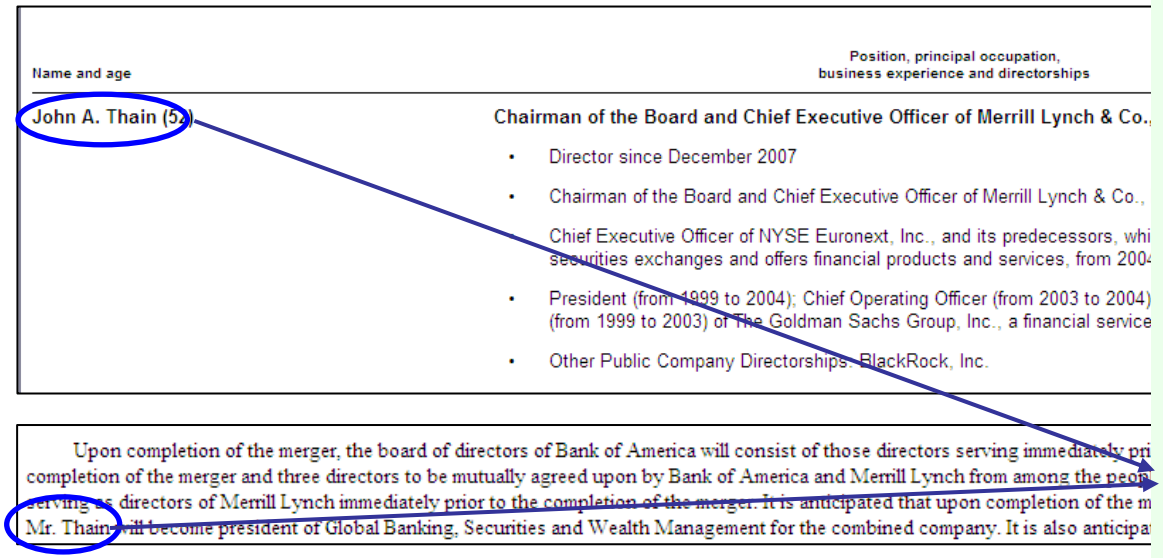
Mr. Dimon became Chairman of the Board on December 31, 2006, and has been Chief Executive Officer and President since December 31, 2005. He had been President and Chief Operating Officer since JPMorgan Chase's merger with Bank One Corporation in July 2004. At Bank One he had been Chairman and Chief Executive 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.



Name and age	Position, principal occupation, business experience and directorships
John A. Thain (52)	Chairman of the Board and Chief Executive Officer of Merrill Lynch & Co., • Director since December 2007 • Chairman of the Board and Chief Executive Officer of Merrill Lynch & Co., Chief Executive Officer of NYSE Euronext, Inc., and its predecessors, whi securities exchanges and offers financial products and services, from 200 • President (from 1999 to 2004); Chief Operating Officer (from 2003 to 2004) (from 1999 to 2003) of The Goldman Sachs Group, Inc., a financial service • Other Public Company Directorships: BlackRock, Inc.

Upon completion of the merger, the board of directors of Bank of America will consist of those directors serving immediately pri
completion of the merger and three directors to be mutually agreed upon by Bank of America and Merrill Lynch from among the peop
serving as directors of Merrill Lynch immediately prior to the completion of the merger. It is anticipated that upon completion of the m
Mr. Thain will become president of Global Banking, Securities and Wealth Management for the combined company. It is also anticipat

List of insiders for Bank of America & Merrill Lynch

…
BANKS, 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.
…

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

Entity Resolution: Context Is Important

Kansas City Life Insurance Proxy Statement (DEF 14A)

Mr. Bixby has been a Director of the Company since 1985.

Mr. Bixby is President, CEO and Chairman of the Board. He was elected Assistant Secretary in 1979; Assistant Vice President in 1982; Vice President in 1984; Senior Vice

Mr. Bixby has been a Director of the Company since 1996.

Mr. Bixby is Vice Chairman of the Board. He was elected Assistant Vice President of the Company in 1985; Vice President, Marketing in 1990; Vice President Marketing Operations in 1992 and President of Old American, a subsidiary, in 1996. Mr. Bixby is the brother of R. Philip Bixby and the cousin of Nancy Bixby Hudson. He also serves as a Director of Sunset Life, Old American and Sunset

Ms. Hudson has been a Director of the Company since 1996.

Ms. Hudson is an investor, and is the cousin of R. Philip Bixby and Walter E. Bixby. She also serves as a Director of Sunset Life and Old American, subsidiaries.

List of insiders for Kansas Life Insurance

…
BIXBY, R. PHILIP
BIXBY, WALTER E.
BLESSING, WILLIAM BIXBY
BRAUDE, MICHAEL
COZAD, JOHN C.
HUDSON, NANCY BIXBY
KNAPP, TRACY W.
COZAD, JOHN C.
…

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

Basic extracted information

cik: **0001090355**
name: **John A. Thain**
company: **Bank of America**
…

Transactions by John Thain

personName: **Thain John A.**
cik: **0001090355**
filingDate: **2008-01-24**
reportingDate: **2008-01-24**
issuer: **BlackRock Inc.**
isOfficer: **false**
isDirector: **true**
fillingType: **3**
…

Appointment announcement

personName: **John Thain**
cik: **0001090355**
appointmentDate: **2008-01-16**
filer: **BlackRock Inc.**
appointedAs: **Director**
filingType: **8-K**
…

Committee membership

personName: **John Thain**
cik: **0001090355**
reportingDate: **2008-06-30**
company: **Merrill Lynch**
title: **Chairman and CEO**
member: **Audit Committee**
filingType: **DEF14A**
…

Need to combine data into a desired structure

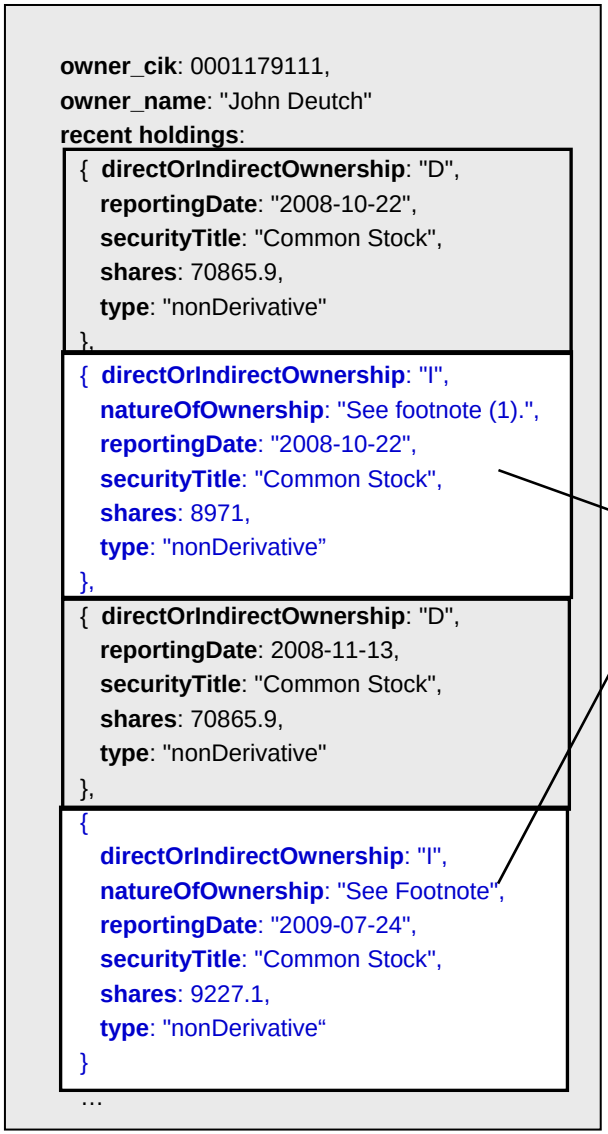
- Data extracted varies in structure and values!

Entity Integration

Person Entity

cik: **0001090355**
name: **John A. Thain**
company: **Bank of America**
employmentHistory: [
{ **Merrill Lynch**, **CEO**, 2008-06-30, …}
{ **Black Rock**, **Director**, 2008-01-16, … }
{ **Black Rock**, **Director**, 2008-01-24, …}
…
]
…

Mapping, Temporal Analysis and Fusion : Computing Current Holdings



Must recognize when we have the *same type of holding* and then take the most *recent* value.

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

Entity Integration

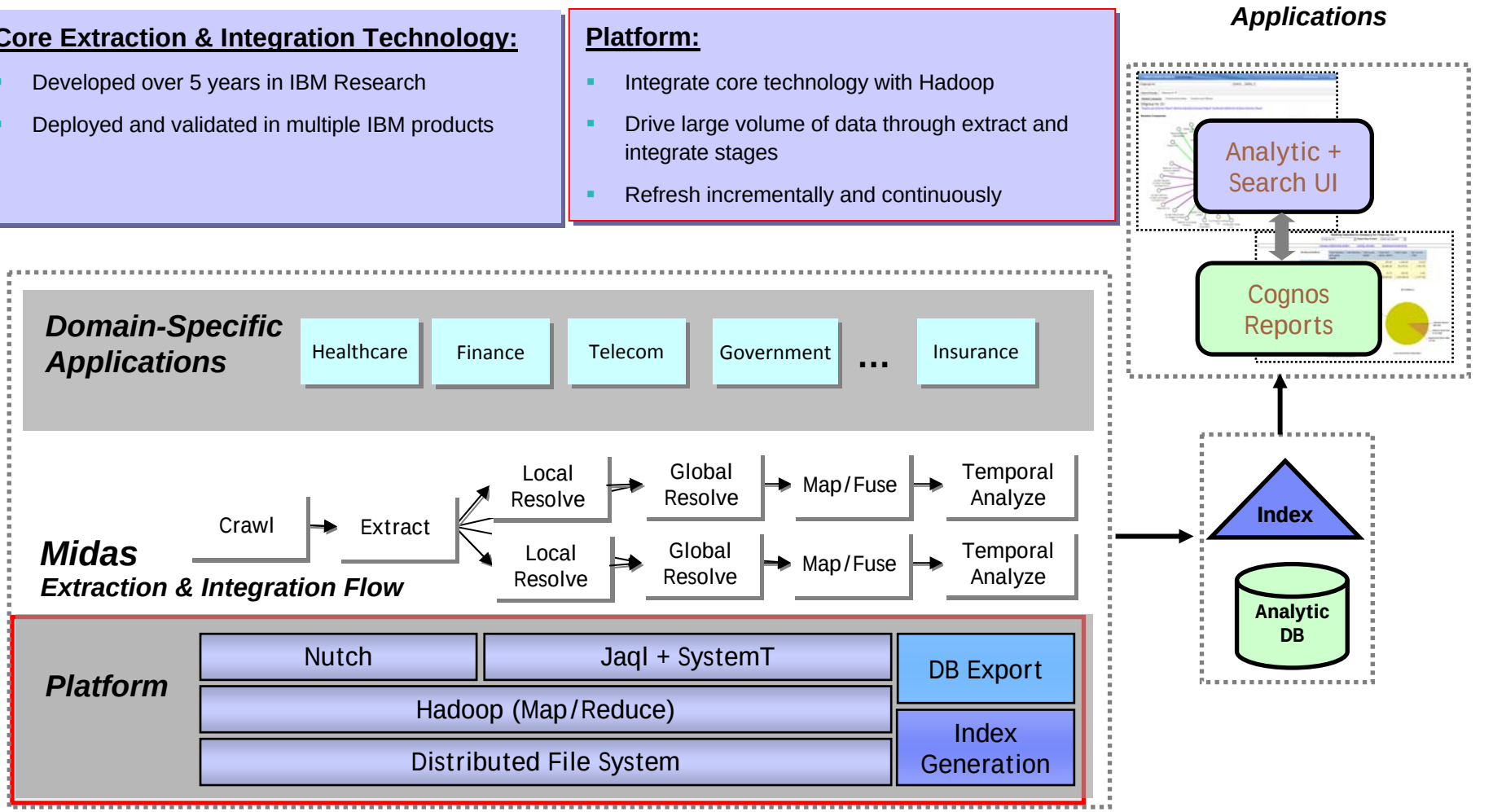
Midas Architecture: A Detailed View

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



Handling Scalability

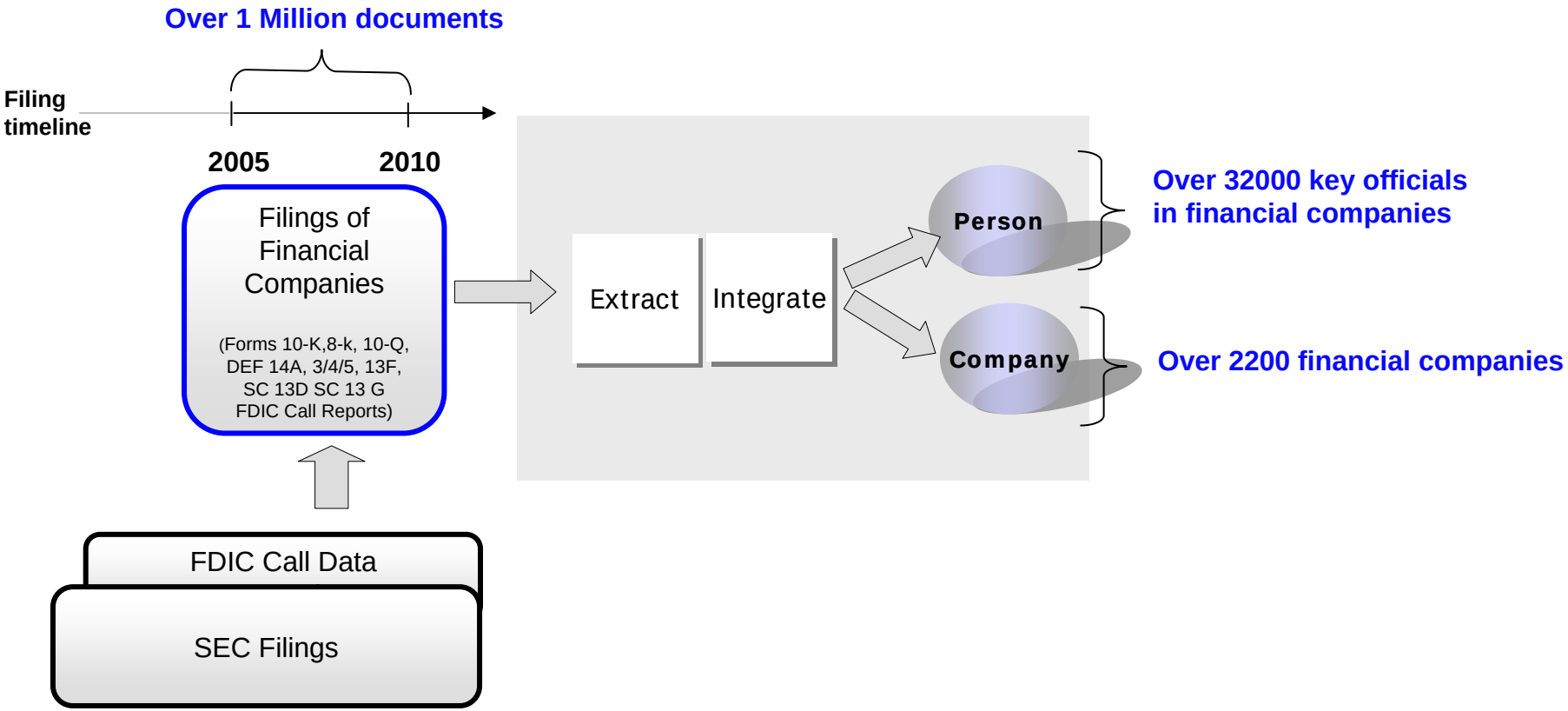
- Scalability Challenges
  - Large document corpora
    - Millions of documents of different formats and document types
    - Documents vary in size (10KB – 10MB each)
    - New documents available daily
  - Maintaining a complex analysis pipeline
    - Some document types require specialized analysis
    - 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 : D*eclarative 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



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