Partnership for International Research and Education
A Global Living Laboratory for Cyberinfrastructure Application Enablement

Financial Table Analysis for Midas

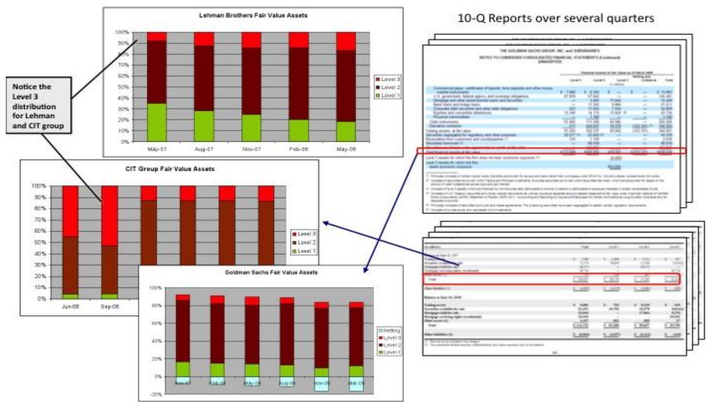
Student: Melita Jaric, PhD student, Florida International University
Research Advisor: Professor Naphtali Rishe, Florida International University
CI-PIRE Partner Advisors: Dr. Howard Ho, IBM Research Almaden, Dr. Ullas Nambiar, IBM IRL India

National Science Foundation

I. Research Overview and Outcome

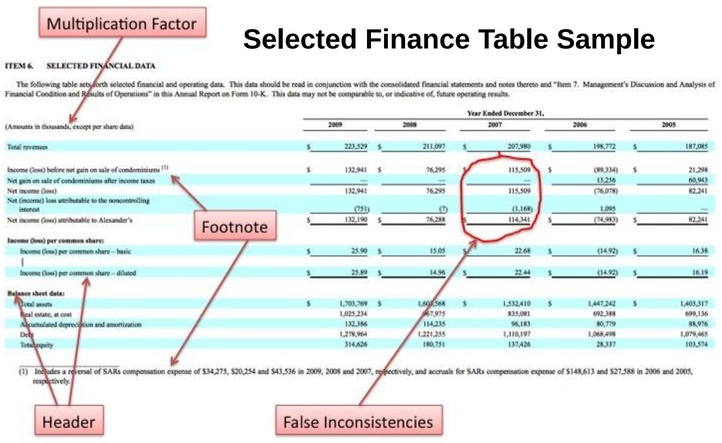
Background: Publically traded companies are required to periodically (quarterly/yearly) file accurate state of their financial activities with Security Exchange Commissioner (SEC). Most of the filing is in Text, with pertinent financial activity reported in few non-standardized tables.
Currently financial analysts are manually extracting data from these filings. This is inefficient, prone to human error, prohibitively time consuming to scale or expand and hence limited in providing accurate financial state of a company and its position within its sector or with regards to its competitors.

Motivation: Hence, what is needed is an automated and scalable extraction and cleansing of financial entities, their resolution and linkage across multiple sources to enable computation of key financial metrics, comparison and correlation of financial ratios over time, per company or industry sector, trend prediction from historical data, and uncovering of non-obvious relationships between financial entities or industry sectors.



Solution Overview: The goal of Financial Table Analysis for Midas is to convert semi-structured consolidated financial tables' data into clean, normalized, structured temporal data that can be accurately queried, aggregated, processed and analyzed. Since this is the first attempt at discovering data mapping rules and dependencies, data flow and data analysis are interdependent. However, system must be designed so that in future, data can be updated incrementally, and data mapping and fusion stage will precede and be independent of data analysis stage.
Scalable architecture leveraging Cloud technology: Running Jaql queries on Json files using the Hadoop MapReduce environment.

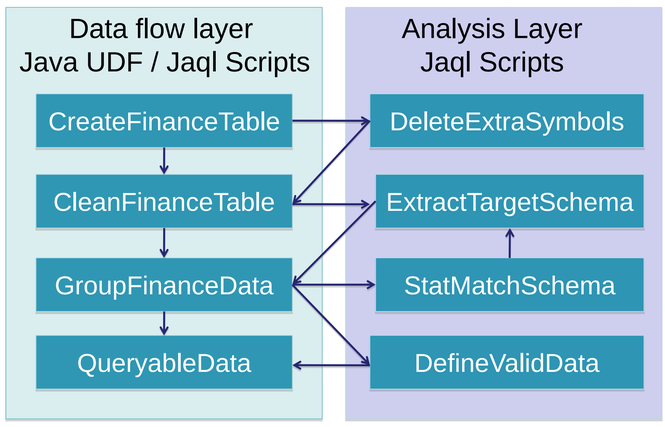
Selected Finance Table Sample



Challenges

Data Importance: How to identify and extract important financial information? "\u0097" vs "Assets"
Financial Metric Mapping: Discover correct domain patterns: "Net income (1)" -> " Net Income (loss)", "Gains" -> "Net Income (loss)"
Semantic Extraction through Record Mapping: Determine information in the string that is important and the one that provides auxiliary information
Information Propagation in jaql (must use Java UDF) Metric Name Completion, Multiplication Factor Identification
Temporal Analysis: The same information is stored in multiple filings. Make sure it is consistent, and if not alert user to a probable cause
Matching xbrl schema: How to match to a target schema when the original data does not have "similar" entry?

Data Flow and Data Analysis Interdependency



Results: Sample Queryable Data

Company Name	Financial Metric	Year	Value	Validity	Factor
PRICE T ROWE GROUP INC	Net Income (loss)	2004	337,000,000	valid	valid
PRICE T ROWE GROUP INC	Net Income (loss)	2005	530,000,000	valid	valid
CITIZENS INC	Net Income (loss)	2004	7,732,000	valid	likely
AMERICANWEST BANCORPORATION	Net Income (loss)	2008	-192,360,000	valid	valid
AMERICANWEST BANCORPORATION	Net Income (loss)	2007	8,538,000	valid	valid
FIRST AMERICAN CORP	Net Income (loss)	2004	345,847	Within 3% error	Not available

Target Schema Matching and Statistics

Jaql query: $nameMatch = join $flat, $targetSchema where $flat.tableInfo.name.final == $targetSchema into {target: $targetSchema, orig: strcat($flat.tableInfo.name.orig,$flat.tableInfo.name.basefootnote) }
-> group by $t = $.target into { target: $t, matches: distinct($[*].orig) };
$nameMatch -> write(hdfs($outDir+"nameMatch"+$tableName+".txt", $txtout));

Output for one element:

```
"matches":
[
    "earnings from discontinued operations",
    "net gain from discontinued operations",
    "net gain from discontinued operations",
    "net income from discontinued operations net",
    "net earnings from discontinued operations",
    "net earnings from discontinued operations",
    "net income from discontinued operations",
    "income from discontinued operations",
    "total income from discontinued operations",
    "net income from discontinued operations",
    "earnings from discontinued operations",
    "income from discontinued operations",
    "income from discontinued operations net",
    "gain from discontinued operations",
    "gain from discontinued operations",
    "income from discontinued operations net"
],
"target": "Net Income (loss) from Discontinued Operations"
```

Jaql query: $cntPath = count(distinct($flat -> transform($.fileInfo.Path)));
$cntPath; // 3603, 3739, 2605, 5343
$fileMatch = join $flat, $targetSchema where $flat.tableInfo.name.final == $targetSchema into {target: $targetSchema, path: $flat.fileInfo.Path}
-> group by $t = $.target into {target: $t, path: $[*].path}
-> transform { $.target, path: distinct($.path), cnt: count(distinct($.path))}
-> sort by {$.cnt desc};

Output for few target entities:

```
[
 {  "cnt": 3600,  "percentage": 96,  "target": "Net Income (loss)"}
 {  "cnt": 2837,  "percentage": 75,  "target": "Total Interest Expense"}
 {  "cnt": 2482,  "percentage": 66,  "target": "Net Interest Income (loss)"}
 {  "cnt": 1885,  "percentage": 50,  "target": "Total Provision for Loan Income (loss)"}
 {  "cnt": 3136,  "percentage": 87,  "target": "Total Assets"}
 {  "cnt": 3081,  "percentage": 85,  "target": "Total Liabilities"}
 {  "cnt": 2707,  "percentage": 75,  "target": "Total Stockholder Equity"}
 {  "cnt": 2644,  "percentage": 73,  "target": "Cash and Cash Equivalents"}
 ...
]
```

II. India - International Experience: Marvelously / shockingly diverse, bewitchingly enchanting, breathtakingly beautiful, deeply spiritual

**Thank you Aditya, Amik, Shyam and Vivek for your open minded and genuine attitude and countless conversations about history, social structure, politics, leisure, future…
I traveled half away across the globe to meet life long friends who ensured that I truly experienced, enjoyed and appreciated their country.**

III. Acknowledgement

The material presented in this poster is based upon the work supported by the National Science Foundation under Grant No. OISE-0730065. This work was also supported by the US Department of Education under P200A090061 and by IBM. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.