I. Research Overview and Outcome

Background: Publicly 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, prohibitive 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.

Challenges

Data Importance: How to identify and extract important financial information? “u0097” vs “Assets”


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 / Jaql Scripts)

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?

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

Thank you Aditya, Amlk, 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.