Feasibility Study

# Overview

As we’ve discussed in the initial section, the advent of data warehousing coupled with the increasing importance of unstructured data have led to a greater need for processing and interpreting data outside of the standard structured formats. Because over 70 percent of all actual data in business and enterprise are classified as unstructured, the necessity for advancements or even simple adaptations in current technology to account for this data is increasing every year. When it comes to data warehousing, the transition to unstructured data is not as big a leap as for other technologies like AI or voice recognition, but still requires some changes in order to match the lack of categorization present. And when it comes to using graph technology to represent this data (especially unstructured data), a study of what is viable and whether or not our proposal is realistic/achievable in the time span provided could be useful. We will also assess the current status of previous research in this field and determine how much can be incorporated or whether adjustments will need to be made to the proposal.

# Description of Concepts and Processes

The primary consideration of this proposal is the usage and integration of visualization techniques that lie outside of the scope of any previous courses at RIT and personal past experiences. Until now all DW discussion in previous RTI courses has revolved around structured data, and ISTE-724, while thorough, lies outside of both particular SQL and handling of unstructured formats (these lie more in the scope of 610 or specific domains of study). Among these will be the need to not only translate familiar structured data formats in SQL data warehouses to the unstructured equivalent, but also determine the optimal way to store and represent the topical data in a graphical form. The crux of the feasibility lies in whether or not the unstructured data framework established using SQL and the data mart method can be “translated” over into graphical representation or worked with via noSQL or other similar extensions of SQL.

While there are no considerable changes to the format or schema of existing structured layout as of the writing of this proposal, some fitting and adjustments will likely have to be made in the transition, particularly when it comes to naming or type, of which the latter will certainly be more difficult, but with type conversions becoming less of an issue in newer versions of SQL. In particular, since NoSQL and SQL have very different frameworks for data this transition into document/graph form will take some work and adjustment. This will be addressed alongside the above concerns in section 4.

# Technology Considerations

Above we mention that the technology that we will be using will include MySQL Workbench, Tableau, and even possibly Weka, all software that is covered in the courses of ISTE 724, 782, and 600 respectively. Other less used technologies at RIT like Pentaho, SAS, or Neo4j are also covered, which may need a bit of familiarization in normal circumstances but are not needed due to extensive prior work with these in data warehousing and relevant research. The primary consideration would lie in the compatibility of SQL and graphical technology, and whether or not there is even a venue for such a possibility or if the project would simply lead to a dead end.

Luckily, despite the differences in setup, there are previous studies in using NoSQL tech and representation for standard Entity-Relation models and diagrams. MongoDB, while demonstrably better at complex queries but worse when aggregate functions are present, can allow for a transfer, but much like Google Translate and other translation services can never fully 100 percent capture the original essence of the sentence, it also can’t migrate all of the nuance and details of the original schema. So when it comes to a direct transfer, while there are a few options ranging from directly rewriting to initial refactoring to a simple host of the schema, we will have to strike a balance between the detail preserved and the amount of time spent on certain specifics that may not make a big difference in the end. This kind of mapping via SQL server can be investigated later in the course of the project, but it is certainly doable.

One more technological consideration to keep in mind is the size of the data. While loading and transferring up to millions of data entries shouldn’t be a big concern in SQL nowadays, certain factors like foreign constraints and indexing can slow loading down to a crawl (as evinced in 724 and 612). Ample time must be left to ensure that if changes or fixes need to happen in later stages, they will not affect the deadline or end product.

# Addressing Concerns and Predictions

The combination of the unstructured topic model alongside the more innovative graphical representation is unique but also not unfounded in similar areas of work in the past decade and beyond. Multiple comparison works have highlighted the differences in specific aspects of both and what conversion between the two would entail. Many RDBMS have tried to address the inconsistency in most noSQL data stores, and ultimately the two options that are available are either migration or direct virtualization using only SQL while at the same time incorporating the noSQL elements of graphs or documents via equivalent independent technologies developed.

Usage of graph and nodes equivalent to topic and weights can translate from topic terms and weights very well, and as with previous discussions the primary concern would be to represent these weights in a visual manner on the graph or document image, along with how to best store these results. This concern is slightly less mitigated if we are using noSQL, and the particulars can be worked out in another exploration if it ends up lying outside the time scope of this study. Potential methods like sentence topic coloring and bigrams can also take elements from NoSQL but are not entirely reliant on them. And for topic modeling in particular, when it comes to representation in graphs more standard visuals like bar graphs or line plots should also be considered and tested comparatively.

# Analysis and Considerations

With the evidence presented, it is recommended that while we should proceed with this proposal, but also advise caution and flexibility when it comes to execution. The findings show that the framework exists, and the difficulty will lie in whether or not the initial goals are too ambitious or vague to consider completed at the end of the timeframe allotted. The potential benefits are not only relevant to database representation but also could help in making connections between noSQL (or at least elements of it) and standard SQL closer and easier to access.

We must also consider that everything relies on the data warehouse structure, and not a simple RDB or entity diagram. The star/snowflake schema must be preserved when conversion occurs, and all the standards and conventions of regular data warehouse development must also be followed. If these rules are kept, the rest is simply a matter of preserving original data integrity and everything else is freeform and flexible in terms of getting the results.

# Works Cited

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Cattell, R. (2011). Scalable SQL and NoSQL data stores. *Acm Sigmod Record*, *39*(4), 12-27.

Chevalier, M., El Malki, M., Kopliku, A., Teste, O., & Tournier, R. (2015). Implementing multidimensional data warehouses into NoSQL.

DePaolo, C. A., & Wilkinson, K. (2014). Get your head into the clouds: Using word clouds for analyzing qualitative assessment data. *TechTrends*, *58*(3), 38-44.

Lawrence, R. (2014, March). Integration and virtualization of relational SQL and NoSQL systems including MySQL and MongoDB. In *2014 International Conference on Computational Science and Computational Intelligence* (Vol. 1, pp. 285-290). IEEE.

Liu, Y., & Vitolo, T. M. (2013, June). Graph data warehouse: Steps to integrating graph databases into the traditional conceptual structure of a data warehouse. In *2013 IEEE International Congress on Big Data* (pp. 433-434). IEEE.

Rocha, L., Vale, F., Cirilo, E., Barbosa, D., & Mourão, F. (2015). A framework for migrating relational datasets to NoSQL. *Procedia Computer Science*, *51*, 2593-2602.

Tessem, B., Bjørnestad, S., Chen, W., & Nyre, L. (2015). Word cloud visualisation of locative information. *Journal of location Based services*, *9*(4), 254-272.

Venkatraman, S., Fahd, K., Kaspi, S., & Venkatraman, R. (2016). SQL versus NoSQL movement with big data analytics. *International Journal of Information Technology and Computer Science*, *8*(12), 59-66.