Significance of metadata and data modeling of metadata by using Marklogic

Mrs. K Lakshmi Prasanna[1]
Department of Computer Science and Engineering
Farah Institute of Technology
Chevella, R.R. Dt –Telangana, India - 501503
lakshmi.prasanna123@gmail.com

Dr. Jangala. Sasi Kiran[2]
Department of Computer Science and Engineering
Farah Institute of Technology,
Chevella, R.R. Dt –Telangana, India - 5015031&2
sasikiranjangala@gmail.com[2]

Dr. K Sreerama Murthy[c]
Department of Information Technology
Sreenidhi institute of science and technology,
Yamanampet, Medchal Dt –Telangana, India - 501301
drsreeram1203@gmail.com

Abstract-**Metadata means data about data which illustrates, traces, and is simple to locate a resource. Metadata has been impacting almost every firm. It has become mandatory for organizations to know the data flow across business processes to take strategic decisions. But, collecting metadata across departments/business processes and putting into a commonality is very difficult by using conventional databases. We need to concentrate on the metadata managing technologies. There are data models that are designed which work on NOSQL database. Envelope pattern in Marklogic provide commonality for the metadata across processes. The data that is gathered across different processes need to be managed in a consistent way. We want to verify metadata management in banking domain. In this paper, we have ingested metadata across multiple departments in banking domain and verified the performance of search results.**

Keywords-**Metadata; Data modeling; Marklogic; envelope pattern;**

I. INTRODUCTION

A key development is bringing a data strategy that precedes metadata management. The available resources and business data can be reused at optimum extent and it can generate safety reports at regular intervals, risk compliance reports etc. Metadata generates profiles of key assets and can reuse real world data based on its impact. Content retrieval for strategic decision making is simple with metadata by classifying key elements like document type, author, document created time stamp etc. With metadata one can improve different content types alerting with a key element such as creation or submission time stamp, in this case, one need not require actual content, the metadata will suffice the need. Firms even need to submit risk approvals and need to generate multi functional reports. Technically metadata can be defined as running complicated commands by considering the strong search indexes without the need for the original data. The metadata indexes will make the data framework as a search engine where one can search over data and metadata.

As per Dataversity report on Metadata, around one third of firms have begun dealing with metadata and one fourth of firms have not thinking of any metadata strategy itself. Managing metadata through traditional RDBMS is difficult. Now, organizations use enterprise NoSQL technology and semantic data models. Metadata can be structural (e.g.,the location of actual data), breakthrough (e.g., author, indexing information etc.), or managerial (e.g., file type, technical details etc). It helps to think of metadata like the glue that harmonizes links and adds context to data.

The data that describes about other data is metadata. There are 3 varieties of metadata:

1. Breakthrough metadata – a resource which discovers/identifies elements like author, title, abstract etc.
2. Structural metadata – a resource which should contain data about page ordering for forming chapters. It has data describing versions, index information etc.

3.Managerial metadata – the resource that stores metadata about technical details, the creation time, and access information etc.

There exists data of some processes need to be used in some other processes. Even though the data might be similar, but the data structure differs and is unknown when it was built. The elements of multiple processes are ingested in a consistent manner.

II. RELATED WORK

Transportation issues are no way related to document data, separation of functionality is very much required where this is best handled by Envelope pattern[7]. Envelope pattern concentrates on delivery mechanism. Document data should not involve in validation of transport information, security problems, or package issues. The data is separated from transport. The document type which has been created for identifying the domain data and the envelope[7] is used for delivery mechanism. The business data and transport/security information is being separated in envelope pattern.

The Data Hub Framework[12] is used for staging documents. It contains the following elements:

1. *Header*s: The data about metadata[1] information
   - Data creation timestamp
   - Source origin
   - Batch identifier
   - The types of data.
   - Data verification.
   - Unified resource indicator of document

2. *Content:* This section will have business data unchanged but into valid formats like json, xml etc.

Triples[8]: The minute data entity semantics which support data model[1] by providing joins, supports RDF framework[10] format for retrieving and manipulation of data, Web Ontology language-based reasoning.

A. Harmonized Documents

- The key elements which act as headers would be used for indexing, can make faceting on those elements, and sometimes can be generated as views.
- Date elements can even be used as header elements but should be in compliance with standard date formats.
- The data which is not in original can be stored in header section.
- Original data section – is the section where the given data is stored and will be stored in specific format if required.

Example:{“envelope”:{{“header”:{{“firstName”:”Lakshmi”},{“lastName”:”Prasanna”}, {“pincode”:”500072”}}, {”triples”:{}}, {“content”:{{“businessKey”:”value”}, {“businessKey2”:”value2”, {“businessKey3”:”value3”}}}}

Envelope[7] has multiple sections like header, triples[8], and content sections. The content section can be exposed via REST end points.

We use envelope pattern to store canonical forms of data.

The transformations in Marklogic[10] – Business entities can export from current system, imported to Marklogic as-is, transform progressively, and relate the documents.

III. EXPERIMENTAL RESULTS

We have verified the metadata data modeling in banking domain and verified the ingestion and search responses, where strategic decisions can be taken at faster pace. Below are the details of the structure being used.

A. Indexing structure

MarkLogic uses a distinct parent-child index to identify field hierarchy. The indexing is a fast phrase index where it uses parent-child names as keys.

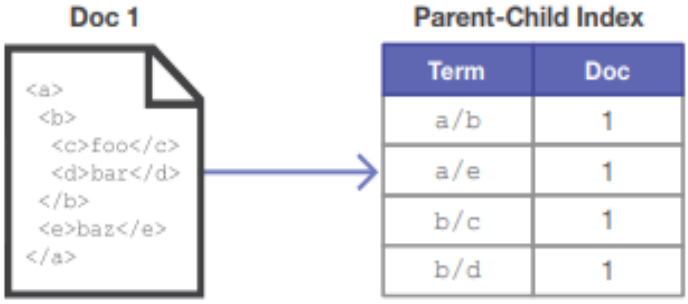

Figure 1. Enabling search based on doc structure by indexing on parent-child relation

B. Data ingestion

Data ingestion can be used for ingesting data from multiple departments wherein it can be saved into a common repository. The data may vary, but common fields can be stored in standard repository.

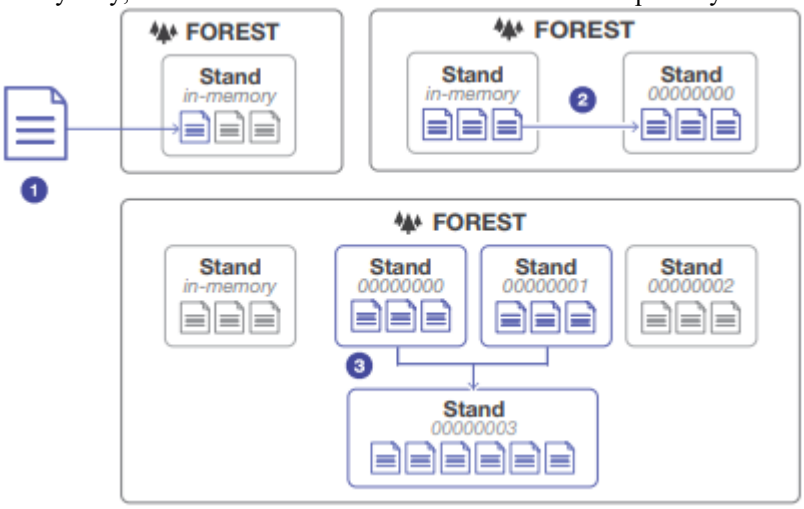

Fig 2. Document ingestion initially to in-memory stand, then to on-disk stand, and then to accumulation step.

The document search can happen wherein indexing is being used while searching. The common terms are being stored as keys through which indexing happens on key terms and range indexes can be created for fast retrieval.

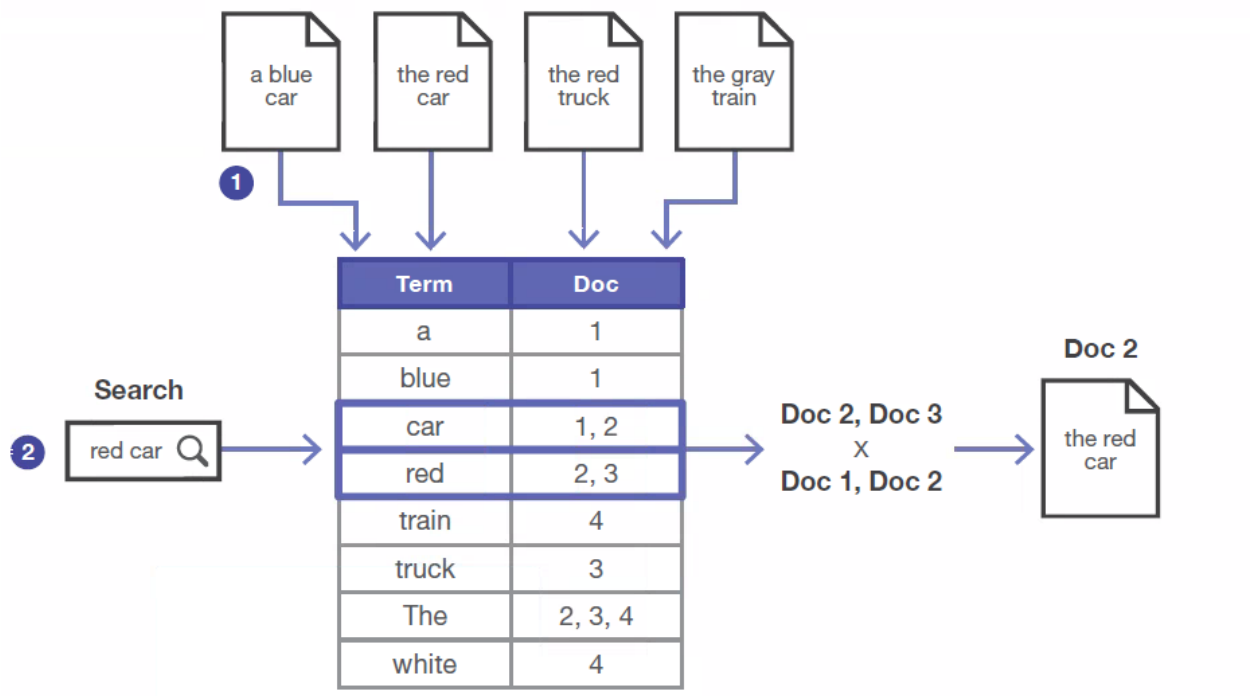

Figure 3. Invert indexing for doc terms search.

C. *Search performance*

Fast phrase searches will incorporate two-word terms into an inverted index.

FAST PHRASE SEARCHES

Term	Doc
a	1
a blue	1
blue	1, 2
blue car	2, 3
…	…

WORD POSITIONS

Term	Doc:Pos
a	1:1
blue	1:2
car	1:3, 2:3
red	2:2, 3:2
…	…

Figure 4. Word positions by fast phrase searches where indexing with term location.

D. *Performance of ingestion and search results:*

The below figure depicts the performance of ingestion and search results.

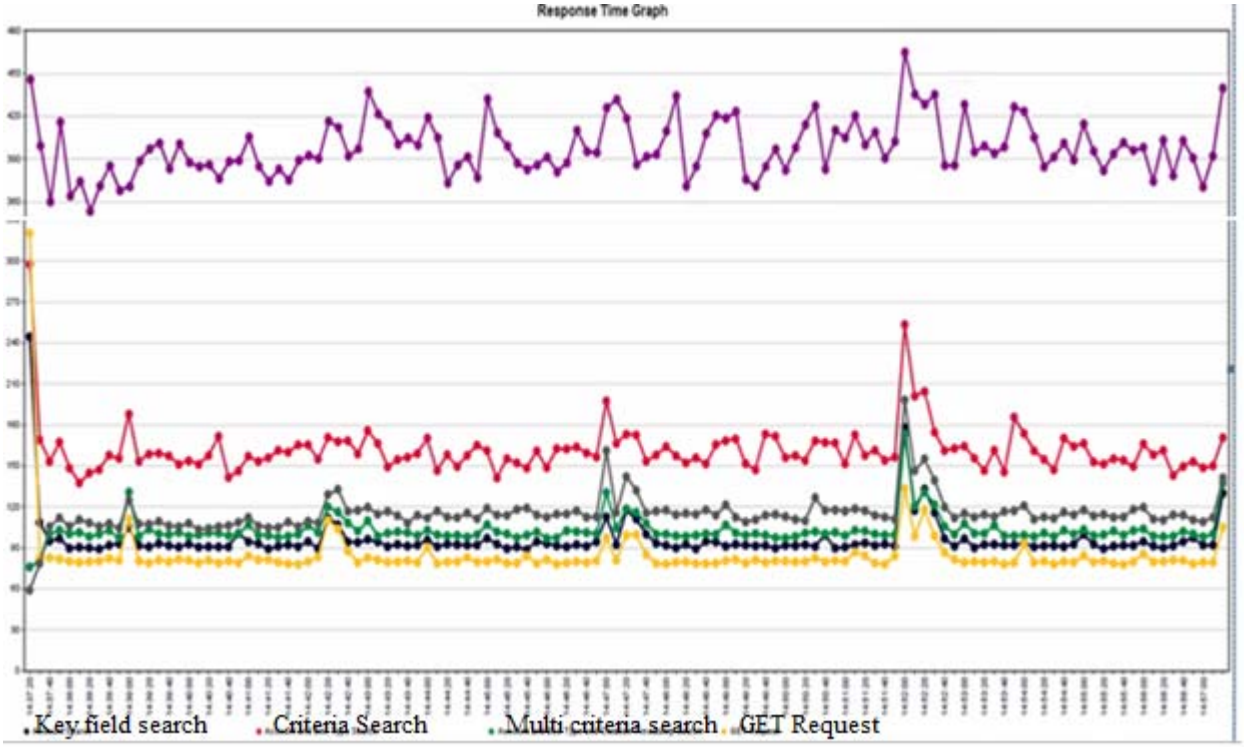

Figure 5. Response graph showing ingestion and search performance

IV. CONCLUSION

In this paper, we tried to understand the importance of metadata and data modeling of metadata by using envelope pattern in Marklogic. We applied these data models in banking domain for different processes like ingestion and search across multiple departments. We have identified, departments can retrieve data and ingest data at a faster pace.

V. REFERENCES

[1] Alink, W. "XIRAF: An XML-IR Approach to Digital Forensics." Master's Thesis, University of Twente, 2005. http://www-db.informatik.uni-tuebingen.de/files/research/pathfinder/publications/alink.pdf

[2] Alink, W., R.A.F. Bhoedjang, P.A. Boncz, and A.P. de Vries. "XIRAF – XML-Based Indexing and Querying for Digital Forensics." Digital Investigation (2006): S50-S58.

[3] Aven, Pete. "A Final 'Word': Part 6 in a series on MarkLogic Server and Office 2007." Mark Logic TechBlog. January 22, 2008. http://xqzone.marklogic.com/columns/smallchanges/"

[4] Aven, Pete. "Running (a.k.a. -ing) with Word: Part 4 in a series on MarkLogic Server and Office 2007." Mark Logic TechBlog. December 18, 2007. http://developer.marklogic.com/columns/smallchanges/2007-12-18.xqy.

[5] Byers, Simon. "Information Leakage Caused by Hidden Data in Published Documents." IEEE Security and Privacy 2, no. 2 (2004): 23-27.

[6] Menoti, David. "Segmentation of Postal Envelopes for Address Block Location: an approach based on feature selection in wavelet space". January 10, 2003. IEEE 10.1109/ICDAR.

[7] S, Decker."The Semantic Web: the roles of XML and RDF", IEEE Internet Computing Volume: 4, Issue: 5, Sep/Oct 2000.

[8] Aven, Pete. "Excel-ing with XQuery: Part 2 in a series on MarkLogic Server and Office 2007." Mark Logic TechBlog. December 4, 2007. http://xqzone.marklogic.com/columns/smallchanges/2007-12-04.xqy.

[9] Ding, Ying. "The Research on data semantic description framework using RDF/XML". 2005 First International Conference on Semantics, Knowledge and Grid, 10.1109/SKG.2005.127.

[10] Aven, Pete. "A Final 'Word': Part 6 in a series on MarkLogic Server and Office 2007."

[11] Chauha, Hitesh." Error Handling Framework for Data Lakes".

[12] Aven, Pete. "Office Logic." Mark Logic TechBlog. November 27, 2007. http://xqzone.marklogic.com/columns/smallchanges/2007-11-27.xqy.

[13] Aven, Pete. "Running (a.k.a. <w:r>-ing) with Word: Part 4 in a series on MarkLogic Server and Office 2007." Mark Logic TechBlog. December 18, 2007.

[14] http://developer.marklogic.com/columns/smallchanges/2007-12-18.xqy.

[15] Balkestein, Marjan, and Heiko Tjalsma. "The ADA Approach: Retro-Archiving Data in an

[16] Academic Environment." Archival Science 7, no. 1 (2007): 89-105.

[17] Born, Günter. The File Formats Handbook. London: International Thomson Computer Press,

[18] Brezinski, Dominique, and Tom Killalea. "Guidelines for Evidence Collection an Archiving." Request for Comments 3227. 2002. http://www.ietf.org/rfc/rfc3227.txt

[19] Byers, Simon. "Information Leakage Caused by Hidden Data in Published Documents." IEEE Security and Privacy 2, no. 2 (2004): 23-27.

[20] Caloyannides, Michael A. "Digital 'Evidence' Is Often Evidence of Nothing." In Digital Crime and Forensic Science in Cyberspace, edited by Panagiotis Kanellis, 334-39. Hershey, PA: Idea Group, 2006.

[21] Caloyannides, Michael A., Michael A. Caloyannides. Privacy Protection and Computer

[22] Forensics. 2nd ed, Artech House Computer Security Series. Boston: Artech House, 2004. [Seeespecially: 8-22, 32-44.]

[23] Carrier, Brian. File System Forensic Analysis. Boston, MA: Addison-Wesley, 2005. [Seeespecially: "Computer Foundations" (17-45), "Hard Disk Data Acquisition" (47-66), and "FileSystem Analysis (173-210).]

[24] Carrier, Brian D. "A Hypothesis-Based Approach to Digital Forensic Investigations." Doctoral Dissertation, Purdue University, 2006.

[25] Casey, Eoghan. "Error, Uncertainty, and Loss in Digital Evidence." International Journal of Digital Evidence 1, no. 2 (2002). [See especially the discussions of clock offsets and log files.]

[26] Chaski, Carole. "The Keyboard Dilemma and Authorship Identification." In Advances in Digital Forensics III: IFIP International Conference on Digital Forensics, National Center for Forensic Science, Orlando, Florida, January 28-January 31, 2007, edited by Philip Craiger and Sujeet Shenoi. New York, NY: Springer, 2007.

[27] Cohen, Tyler, and Amber Schroader. Alternate Data Storage Forensics. Burlington, MA: Syngress, 2007.

AUTHORS PROFILE

Mrs. K Lakshmi Prasanna is a Student of M.Tech CSE from Farah Institute of Technology, Chevella, RR Dist – Telangana. Her research interests include Data Mining, Cloud Computing

Dr. J. Sasi Kiran Graduated in B.Tech [EIE] from JNTU Hyd. He received Masters Degree in M.Tech Computers and Communications from Bharath University, Chennai and M.Tech Computer Science and Engineering[CSE] from JNTUH Hyderabad. He received Ph.D degree in Computer Science from University of Mysore, Mysore. At Present he is working as Professor in CSE and Principal in Farah Institute of Technology, Chevella, R.R. Dist Telangana State, India. His research interests include Image Processing, Cloud Computing and Network Security. He has published several research papers till now in various National, International Conferences, Proceedings and Journals.

Dr. K. Sreerama murthy Graduated in B.Tech [CSIT] from JNTU Hyd. He received Masters Degree in M.Tech Software Engineering,JNTUH Hyderabad. He received Ph.D. degree in Computer Science and system engineering from Andhra University Visakhapatnam. At Present, he is working as Associate Professor in IT Department, Sreenidhi Institute of science and technology, Yamanmpet Hyderabad, R.R. Dist., Telangana State, India. His research interests include Data Mining ,IDS, Big-Data analytics , Cloud Computing and Security. He has published several research papers till now in various National, International journals and Conferences, Proceedings .