Approaches for Managing and Analyzing Unstructured Data

#1N. Veeranjaneyulu, M. Nirupama Bhat, A. Raghunath

School of Computing, Vignan's University, Guntur, India
#1veeru2006n@gmail.com

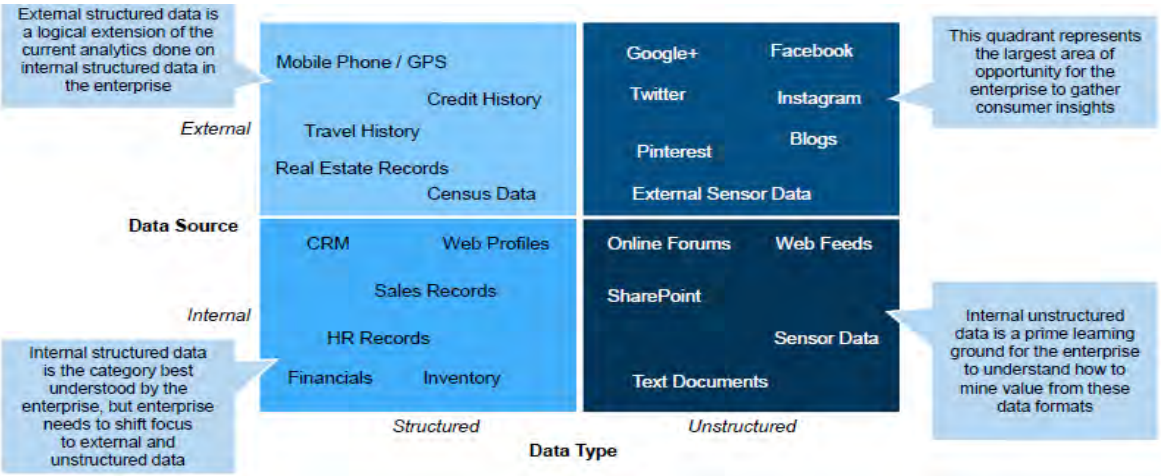
***Abstract*— Large volumes of data that will be stored and accessed in future is unstructured. The unstructured data is generated in a very fast pace and uses large storage areas. This increases the storage budget. Extracting value from this unstructured data which balances the budget is the most challenging task. Archives of interactive media, satellite and medical images, information from social network sites, legal documents, presentations and web pages from various data sources affects the data center's ability to maintain control over the unstructured data. Therefore, it is very essential to design systems to provide efficient storage, and access to these vast and continuously growing repositories of unstructured data. This can be achieved by retrieving structured information from the unstructured data. In this paper, we discuss approaches to process and manage such data. We also elaborate the architecture, technologies and applications to facilitate system design and evaluation.**

Keywords- Unstructured data, structured data, data centers, data source, design and evaluation.

I. INTRODUCTION

Big data refers to huge data sets that are of larger magnitude (volume); more diverse, including structured, semistructured, and unstructured data (variety); and arriving faster (velocity) than you or your organization has had to deal with before. This flood of data is generated by connected devices— from PCs and smart phones to sensors such as RFID readers and traffic cams. It is heterogeneous and comes in many formats, including text, document, image, video, and more. The real value of big data is in the insights it produces when analyzed— discovered patterns, derived meaning, indicators for decisions and ultimately the ability to respond to the world with greater intelligence. Big data analytics is a set of advanced technologies designed to work with large volumes of heterogeneous data. It uses sophisticated quantitative methods such as machine learning, neural networks, robotics, computational mathematics, and artificial intelligence to explore the data and to discover interrelationships and patterns.

The corporate information that is not stored in a database is generally labeled as unstructured data, can be textual or non-textual. Textual unstructured data is generated in media like email messages, PowerPoint presentations, Word documents, collaboration software and instant messages. Non-textual unstructured data is generated in media like JPEG images, MP3 audio files and Flash video files. If left unmanaged, the sheer volume of unstructured data that's generated each year within an enterprise can be costly in terms of storage. Unmanaged data can also pose a liability if information cannot be located in the event of a compliance audit or lawsuit. The information contained in unstructured data is not always easy to locate. It requires that data in both electronic and hard copy documents and other media be scanned so a search application can parse out concepts based on words used in specific contexts. This is called semantic search. It is also referred to as enterprise search.


External structured data is a logical extension of the current analytics done on internal structured data in the enterprise
External
Mobile Phone / GPS
Credit History
Travel History
Real Estate Records
Census Data
Google+
Facebook
Twitter
Instagram
Blogs
Pinterest
External Sensor Data
This quadrant represents the largest area of opportunity for the enterprise to gather consumer insights
Data Source
CRM
Web Profiles
Sales Records
HR Records
Financials
Inventory
Internal
Online Forums
Web Feeds
SharePoint
Sensor Data
Text Documents
Internal structured data is the category best understood by the enterprise, but enterprise needs to shift focus to external and unstructured data
Internal unstructured data is a prime learning ground for the enterprise to understand how to mine value from these data formats
Structured
Unstructured
Data Type

We know unstructured data is one without a defined data model or cannot be easily usable by a computer program. In a structured document, certain information always appears in the same location on the page. For example, in an employment application the applicant’s name always appear in the same box in the same place on the document. In contrast, an unstructured document has the opposite characteristics – information can appear in unexpected places on the document.

Value of Unstructured Data: a.. Business Value b.. Better information c.. Timely information d.. Relevant Information e. Greater business impact and f. More information is available to store, manage and modeled.

In customer-facing businesses, the information contained in unstructured data can be analyzed to improve customer relationship management and relationship marketing. In social media applications like Twitter and Face book go mainstream, the growth of unstructured data is expected to far outpace the growth of structured data. According to the "IDC Enterprise Disk Storage Consumption Model" report released in Fall 2009, while transactional data is projected to grow at a compound annual growth rate (CAGR) of 21.8%, it's far outpaced by a 61.7% CAGR prediction for unstructured data.

Challenges in unstructured data:

A. Solution Maturity:

- Limited number of large implementation of Big Data solutions exist in the enterprise
- Most of the enterprise implementations are in pilot stages

B. Organization Limitations:

- Talent – Lack of truly skilled professionals on the types of data, and its appropriate use
- Culture – Organizations have not yet fully realized the implication of Big Data on business modeling and insights, and IT architecture and execution

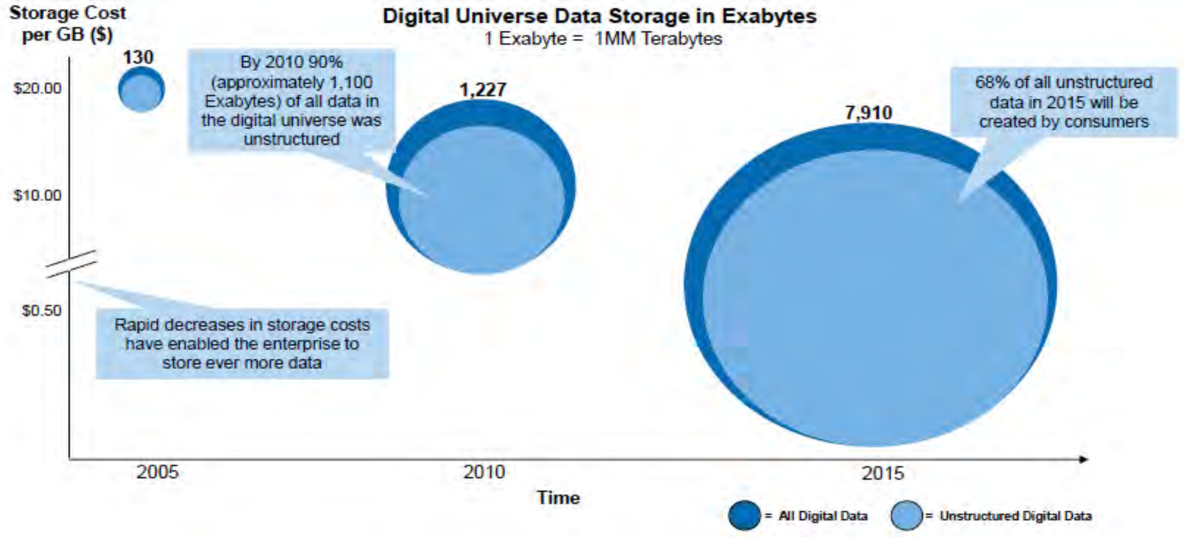
C. Privacy/trust concerns:

- As more data is available, acceptable use of personal data becomes of greater concern to customers
- Amount of data is increasing faster than organizations can properly secure the data

D. Technology:

- Velocity, volume, and variety of data show no signs of slowing which will require new technology solutions

Unstructured consumer data, called Big Data, represents majority of growth in data volume, up 56% CAGR since 2005



II. PROCESSING UNSTRUCTURED DATA

Unstructured data processing is therefore a very important emerging class of applications. There are a number of unstructured data processing applications that are already in use today. These applications include text searches (exact and approximate searches) [2], content-based searches of image, video, and audio files [3], and data fusion. Although some of these applications are used in relatively niche domains (e.g., geo-spatial data fusion is used in urban planning and forestry), the core methods used in these applications are expected to become commonplace across a wider range of applications in the future. For example, Content-Based Image Retrieval (CBIR), which is very promising and I/O intensive, is now used in the field of medicine for querying biomedical digital libraries.

The growing demand for unstructured data management has already started creating a market for hardware appliances that are specifically designed for searching and processing unstructured data. Given this shift in momentum towards unstructured data processing, it is imperative to develop benchmarks that can aid in the design and evaluation of future systems that would have to run these applications efficiently. A common feature across several unstructured data processing applications are that they are very I/O intensive, and therefore the place heavy demand on the storage system to deliver high performance. Moreover, unstructured data processing workloads exhibit a variety of unique data access patterns that are not sufficiently captured by traditional server I/O workloads, such as, the TPC benchmarks. Therefore, in order to design storage systems for this important emerging class of applications, we need a benchmark suite that can capture their processing and I/O characteristics. A benchmark suite with four work loads: Edge detection, Proximity search, Data Scanning and Data fusion for unstructured data was discussed in [16].

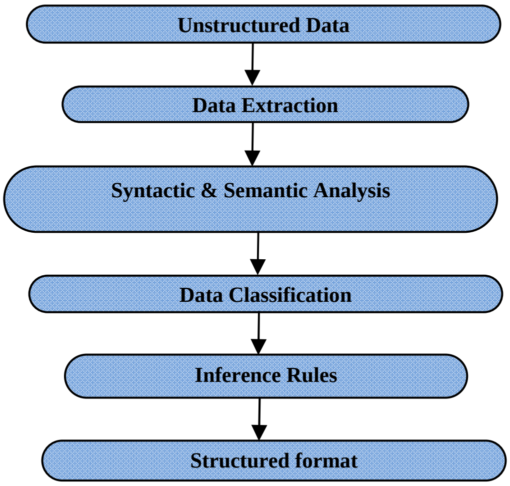
Unstructured Data Management

To manage unstructured data, information from various sources has to be extracted, organized, characterised, analyze the data, data mining, classification of data, text mining and modelling of the processed data.

- Extract Information
- Feature extraction
- Organized the facts
- Text mining
- Modelling and defined the structure of processed data.

For managing unstructured data in web pages for database using XML:

It’s hard to find a tool that deals the unstructured data which can be stored, retrieve data extracted into structured database. The following steps to be carried out to get the output into actionable form from unstructured data.



Unstructured Data: Unstructured data to be analyzed is considered as input either a web page or a document. *Data Extraction:* Data extraction is a process of retrieving and capturing the data from one medium to another medium. Medium can be web pages, documents, database, and stack of information. Web pages are typically considered unstructured data though web pages are defined by HTML, which has rich structure. This is because web pages also contains lot of static text, links and references to external, images, XML files, animations and databases. Therefore extract and categorized information out of data. A wrapper access HTML document and exports it into structured format XML or data relations. Maintaining the Integrity of the Specifications.

Target the extraction: Extraction target can be a relation of ‘k’ tuples, where k is number of attributes in a record or object.

Syntactic & Semantic Analysis: For syntactic analysis, structure is determined by generating a parse tree by classifying sentence into subjects, verb phrase (verb, object). Similarly semantic analysis finds synonyms.

Data classification: Data classification is to categorize data based on required models like object oriented model or ER model. There are many algorithms to classify in data mining like 'K-nearest neighbour (KNN)' algorithm. Some more algorithms include Bayesian algorithm and concept vector based (CVB) algorithm to classify words in documents. 'Page rank algorithm' uses search ranking technique based on hyperlinks on the web.

Inference rules and Representation into structured format: Inference rules can be employed to draw conclusions of the classified data by preserving the semantic property. XML is used to store and transport the data. The classified data is stored in the form of data tables or XML is used to store the data based on the requirement of the desired action planned from the unstructured data.

III. RELY ON ARCHITECTURE

Cloud computing and service oriented architecture are the supporting enablers for Big data analytics. Big data analytics offers the promise of providing valuable insights that can create competitive advantage, spark new innovations, and drive increased revenues. As a delivery model for IT services, cloud computing has the potential to enhance business agility and productivity while enabling greater efficiencies and reducing costs. Both cloud computing and SOA continue to evolve. SOA lays the foundation for cloud using the reusable, plug and play services of SOA it is layered in the design of cloud environment providing the better architecture. Storage of big data to address and derive meaningful analytics that respond to real business needs is a big challenge. Efficient and agile cloud environments and cloud providers are offering these services. Cloud computing offers a cost-effective way to support big data technologies and the advanced analytics applications that can drive business value.

With the potential for so much data to reveal insights that can boost competitiveness, companies must find new approaches to processing, managing, and analyzing their data—whether it's structured data typically found in traditional relational database management systems (RDBMSs) or more varied, unstructured formats. Plus, combining diverse data sources and types has the potential to uncover some of the most interesting unexplored patterns and relationships.

Real time supports predictive analytics. Predictive analytics enables organizations to move to a future-oriented view of what's ahead and offers organizations some of the most exciting opportunities for driving value from big data. Real time data provides the prospect for fast, accurate, and flexible predictive analytics that quickly adapt to changing business conditions. The faster the data is analyzed, the more timely the results are produced, and has greater its predictive value. The scope of big data analytics continues to expand. Early interest in big data analytics focused primarily on business and social data sources, such as e-mail, videos, tweets, Face book posts, reviews, and Web behavior. The scope of interest in big data analytics is growing to include data from intelligent systems, such as in-vehicle infotainment, kiosks, smart meters, and many others, and device sensors at the edge of networks—some of the largest-volume, fastest-streaming, and most complex big data. Ubiquitous connectivity and the growth of sensors and intelligent systems have opened up a whole new storehouse of valuable information. Interest in applying big data analytics to data from sensors and intelligent systems continues to increase as businesses seek to gain faster, richer insight more cost-effectively than in the past, enhance machine-based decision making, and personalize customer experiences.

Organizations continue to store more and more data in cloud environments, which represent an immense, valuable source of information to mine. Clouds also offer business users scalable resources on demand. Combining the latest Intel Xeon processor-based servers and storage, along with Intel SSDs and Intel 10 GbE networking resources used in cloud environments, with big data processing tools like Intel Distribution for Apache Hadoop software provides the high-performance compute power needed to analyze vast amounts of data efficiently and cost-effectively. Running Hadoop in virtualized environments continues to evolve and mature with initiatives like VMware's open-source project Serengeti, among others.

Capability Enabler for Big Data

Mobility:

Use of mobile devices for real-time analytics with "Anytime, anywhere" connectivity provides a full range of functions on mobile channels, build infrastructure to support apps and use cloud-sourced services. As enterprise applications are accessed by multiple customer devices, the user interface must be separated from back-end. Migration to web-centric tools that reduce dependency on specific devices and platforms.

High-end Analytics:

This paradigm needs a shift in enterprise information management technology by including non-relational databases , new data distribution architecture, in-memory processing, machine learning-based analytics, full population analytics etc.,

Data Management:

The big data management can be made by bringing fundamental changes in enterprise data architecture, governing enterprise data warehouse and analytical data structure enhancements. This accommodates unstructured data from multiple external providers. This is possible by the integration of relational and non-relational structures. New distributed technologies such as Hadoop and MapReduce processes large volumes of unstructured data resulting its usage in broad range of business functions.

Next Gen Data Processing:

It needs specialized database management systems with integrated data architecture to manage relational data with new data structures. These are combined into enterprise architectures. The in-memory processing is coupled with event and trigger based processing. These new database constructs needs support of data, query and analysis architecture. Integrated data architecture is necessary to manage relational data with new data structures.

Supporting Enablers for Big Data

Cloud technology is one of the several enablers of high end analytics for handling big data. It provides on-demand, standardized infrastructure, platform, processes and services. It also provides external on-demand services (such as analytical computing). It sources applications to manage data intensive analytics, with minimal impact to internal systems. These technologies also support mobility and complex analytics.

Service Oriented Architecture:

It is a service-based technology architecture supporting the big data capability by providing re-usable, plug-and-play services. The SOA implementation approach must be refined to provide outputs of Big Data processing as services. The enterprise technology strategy must be refined to ensure service based interoperability.

IV. Conclusion

The enterprise's success is dependent on the analysis of the information it collects and manages, using the appropriate technology. The successful execution of big data road map requires an operating model that incorporates the right skills, practices, processes and funding patterns. The winning capability and strategies can be enabled by understanding the value of data. The technology experts must know how to collect, organize, structure and store data for use. The usage experts or data scientists bridges the business and technology areas. They must know how to architect big data, use the data, create required insights and embed the insights in operational business processes. This enables the enterprise's business capabilities and provides competitive advantage.

V. References

[1] K. Albayraktaroglu, A. Jaleel, X.Wu, M. Franklin, B. Jacob, C.-W. Tseng, and D. Yeung. BioBench: A Benchmark Suite for Bioinformatics Applications. In Proceedings of the International Symposium on Performance Analysis of Systems and Software (ISPASS), pages 2–9, March 2005.

[2] R. Chamberlain, M. Franklin, and R. Indeck. Exploiting Reconfigurability for Text Search. In Proceedings of the Workshopon High Performance Embedded Computing (HPEC), September 2006.

[3] A. Smeulders, M.Worring, S. Santini, A. Gupta, and R. Jain. Content-based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1349–1380, December 2000.

[4] G. Fountain and S. Drager. High performance real-time fusion architecture. In Proceedings of the Fifth International Conference on Information Fusion, pages 1478–1485, 2002.

[5] Mansuri I.R. Sarawagi S. "Integrating Unstructured Data into Relational Databases" Data Engineering. ICDE '06. Proceedings of the 22nd International Conference, IIT Bombay 2006.

[6] David Alfred Ostrowski. IEEE International Conference on Semantic Computing "A Framework for the Classification of Unstructured Data." Berkeley, CA, USA 2009.

[7] Rao R. "From unstructured data to actionable form" appeared in IT professional, ieee.org computer society." Inxight, Sunnyvale, CA, USA.

[8] http://searchbusinessanalytics.techtarget.com/feature/Managing-unstructured-data-in-the-organization.

[9] Maluf D.A. Tran, P .B "Managing unstructured data with structured legacy systems" , Aerospace conference 2008, IEEE.

[10] Unstructured Data in http://en.wikipedia.org/wiki/Unstructured_data

[11] Seth Grimes. "is unstructured data merely modelled" published in Intelligent Information week journal. 2005.

[12] Robert Malone. “Structuring unstructured data” published in Forbes magazine, USA. 04-may-2007.

[13] Ayaz Ahmed Shariff K, Leveraging Unstructured Data into Intelligent information – Analysis & Evaluation, 2011 International Conference on Information and Network Technology IACSIT Press, Singapore.

[14] http://www.information management.com/issues/20030201/6287-1.html

[15] Ramesh Nair, Andy Narayanan, “Benfitting from Big data Leveraging Unstructured data Capabilities for Completitive Advantage”, Booz & Company Inc. 2012.

[16] Clinton Wills Smullen, “A Benchmark Suite for Unstructured Data Processing”.