[VNSGU JOURNAL OF SCIENCE AND TECHNOLOGY] Vol.4. No. 1, July, 2015
113 - 118,ISSN : 0975-5446

Comparison of Unstructured Data Handling Mechanisms in Oracle 11g and MySQL 5.6

DESAI Vaibhav D.
SDJ International College, Surat.
vaibhav.sdjic@gmail.com

VAIWALA Vimal A.
SDJ International College, Surat.
vimal.sdjic@gmail.com

Dr. JOSHI Hiren D.
Dr. Babasaheb Ambedkar Open University (BAOU) Ahmedabad.
hiren.joshi@baou.edu.in

Abstract

The proper functioning of companies, firms and other organizations have to properly manage, understand, analyze and effectively use the large amounts of data **and** unstructured information. The structured information is generally referred to as database. The other large amount of data which are not structured by nature, which can be from various sources like business transactions, social media, web content, sensors and machine output and XML documents. The conventional structured data can be easily managed, processes and queried using various database management systems. All the firms are having some amount of data which are not fully structured. The unstructured data generally has some logical relevance with the structured data. This forces the conventional DBMS to incorporate the functionality to handle unstructured data along with structured data. In this paper we will discuss unstructured database, how unstructured data is handled in two leading databases Oracle 11g & MySQL 5.6 and comparison both the databases how they handle unstructured data.

Keywords : Unstructured, database, Oracle 11g, MySQL 5.6

1. Introduction

In the era of relational database management system, the term data generally refers to the information lying in rows and columns format in database objects called tables. The structured data could easily be queried and processed since they are following some rules defined by relational model. Data other then structured data are called unstructured data. In other words, we can say the data that does not reside in rows and columns are unstructured data. It is exactly opposite to structured data and bares all the opposite characteristics of structured data processing. Unstructured data often includes text and multimedia content. Following are some example of

unstructured data: e-mail texts and graphics, text files, word processing documents, multimedia files, presentation files, web pages and many other kinds of business documents. The important point to be noted is that the unstructured data may have an internal structure, but they definitely do not follow the structure of native database. Hence they can never fit in structured database directly.

It is not so easy to automate the transformation of unstructured data to structured data. Had it been easy and accurate, the unstructured data could become intelligent. The main problem with unstructured data is they are for humans. And humans do not understand data in strictly DBMS format. One prime example of unstructured data is Email. For a busy corporate manager it may be somewhat useful as they can be arranged by date, size or time. If these emails could be grouped by subject and content, it would have been a great help for the manager by having structured data in form of email. Only email subject do not suffice the criteria for the emails being structured, as it is subjective and option how one writes email subjects. Unstructured data is percentage wise increasing over structured data in enterprises. As shown in the figure 1-a, Nearly 90% of the data today in enterprise is estimated to be unstructured since last 10 years.[1]

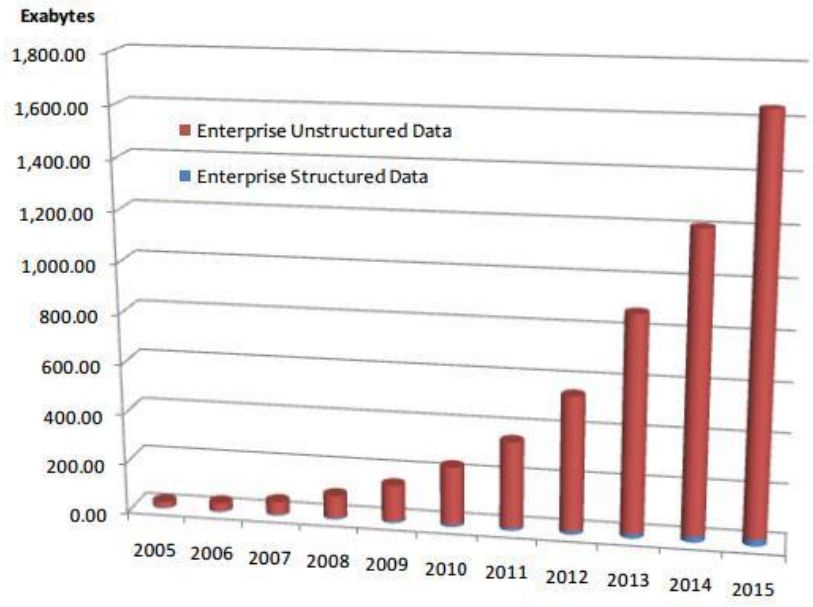

Figure 1-a

Unstructured data could be one of the data sources in a heterogeneous database. Typically, heterogeneous database is a database having different databases at different sites in different formats. If managing and transforming unstructured data to structured data format is a challenge, unstructured data in a heterogeneous database is even bigger challenge. In this paper we are going to explore how unstructured data is handled within various popular databases, namely Oracle 11g and MySQL 5.6. We are not discussing the unstructured data as a part of heterogeneous database. Let us first discuss the advantages of unstructured data maintenance inside a database.[2]

- **Robust Administration, Tuning and Management:** Any data stored inside the database has some default advantages like administration, fine tuning and overall management of data. Unstructured data is also not an exception to this. It is really having above advantages over independent existence.
- **Simplicity of Application Development:** Now a days all the databases have at least some facility to manage the unstructured data. E.g. Oracle has a support for a specific

type of content includes SQL language extensions, PL/SQL and JAVA APIs, Xpath and Xquery (in the case of XML) and, in many cases, JSP Tag Libraries, as well as algorithms that perform common or valuable operations through built in operators.

- **High availability:** Databases and data are so important and critical for any type of application, that almost all popular databases ensure high availability i.e. 24x7 uptime. This directly ensures the availability of unstructured data also, if they are part of a database.
- **Scalable Architecture:** Many database features like indexing, partitioning, triggers are applicable to semi structured data also. This can be a big advantage of the mechanism having unstructured data inside a database.
- **Security:** Continuing the above logic of all major database related prime advantages are applicable to unstructured data also, the unstructured data could be the level best secure that the database provides to structured data.

2. How Oracle 11g Manages Unstructured data

Oracle has come up with a new and high performance LOB (large object) called SecureFiles. The maximum size of a LOB in Oracle 11g is between 8TB to 128TB based on parameter DB_BLOCK_SIZE. (Maximum size: 4GB-1 * DB_BLOCK_SIZE)[7]. This ensures retrieval of unstructured data equal or speedier, than its equivalent file system independent existence. The optimized algorithms with SecureFiles make it up to 10x faster than older LOBs. With SecureFiles, unstructured data can be part of a database transaction, thereby freeing the application from the complexity of guaranteeing atomicity, read consistency and other backup and recovery procedures. SourceFiles extends Transparent Data Encryption (TDE) capability for securing unstructured data in Oracle 11g. Older applications require no changes to upgrade to TDE enabled LOB's from older versions of LOB's. The algorithms of SourceFiles take care of it. SecureFiles supports the following encryption algorithms:[3]

- 3DES168: Triple Data Encryption Standard with a 168-bit key size.
- AES128: Advanced Encryption Standard with a 128 bit key size.
- AES192: Advanced Encryption Standard with a 192-bit key size. (Default)
- AES256: Advanced Encryption Standard with a 256-bit key size.

On sizing front, SourceFiles supports features such as de-duplication and compression. Oracle automatically detects multiple, identical SecureFiles data and stores only one copy, thereby saving storage space. This is entirely transparent to the developer. This removal of duplication significantly improves performance also. For compression various industry standard algorithms are followed by SourceFiles. The compression level can be configured at database level and the user can set the suitable compression level, which is best suited to his/her storage and performance criteria.

Any database management system includes support of data types, storage and index structure, which allows user to write meaningful queries and fetch desired structured data. In a similar manner, the programmer needs these basic features for unstructured data also. Oracle has provided some new data types like DICOM (Digital Imaging and Communications in Medicine) data from release 11g. It has also improved in some pre-existing old data types such as XML, Text, Spatial and multimedia.[3]

i. *Oracle XML Db*: Oracle XML DB is a high-performance, native XML storage and retrieval technology that is delivered with all versions of Oracle Database. It provides full

support for all of the key XML standards, including XML, Namespaces, DOM, XQuery, SQL/XML and XSLT. New features in Oracle Database 11g offer improved performance and scalability. To address non-schema based XML in an optimal manner, Oracle Database 11g introduces a new Binary XML storage option and new XML Indexing capabilities that deliver high performance insert, update and query operations.

ii. *Oracle Text*: Oracle text has been introduced to integrate text searching, retrieval and management of text inside Oracle database. The number of supported partitions has been dramatically increased; in Oracle Database 10g, the maximum number of partitions that could be used was 9999; in Oracle Database 11g the limit for text index partitions is now the same as the limit for table partitions – 220 - 1, or 1,048,575.[3]

iii. *Oracle Spatial*: Oracle Locator is a feature which is built in feature of Oracle from Orale 10g onwards. It enables any business application to directly incorporate location information and realize competitive advantages. Oracle supports both vector and raster data. It supports 3D data, topology and network models. From on, Oracle 11g has started supporting OpenGIS Web Services standards: Web Map Service (WMS), Web Feature Service – Transactions (WFS-T), Web Catalog Services (CS-W), and Open Location Services (OpenLS).[3]

iv. *Multimedia*: Oracle multimedia is a feature is a feature which allows user to store, manage, and retrieve images, audio, video, or other media data in the database. From Oracle 11g, the size limit of individual multimedia object within BLOB has been increased to at lease 8 terabytes. (i.e. between 8 to 128 terabytes). In addition to this, Oracle multimedia is enhanced to support images that contain upto two billion pixels.

v. *Oracle DICOM:* DICOM (Digital Imaging and Communications in Medicine) is a format related to medical images. DICOM is a image format that is used in computerized axial tomography (CAT) and magnetic resonance imaging (MRI) scan images.[4] Oracle DICOM satisfies all the standards of DICOM standards committee. It has some basic features like to store and retrieve DICOM images, validate an image as DICOM or not, convert non-DICOM image to DICOM.

3. How MySQL 5.6 Manages Unstructured Data

MySQL 5.6 supports unstructured data by having BLOB data type. A BLOB is a binary large object that can hold a variable amount of data. BLOB values are treated as binary strings (byte strings). It does have some variants of BLOB like, TINYBLOB, BLOB, MEDIUMBLOB, and LONGBLOB.[5] With MySQL 5.6, user can define various storage engines while creating database objects like table. Some of the storage engines are MyISAM, InnoDB, HEAP etc. MySQL 5.6 uses InnoDB as default storage engine. BLOB's are used to store any binary format unstructured data like multimedia data like images, videos and sound files. For text format unstructured data, MySQL 5.6 is having four basic data types: TINYTEXT, TEXT, MEDIUMTEXT, and LONGTEXT[5]. MySQL 5.6.4 and higher versions are having FULLTEXT search capabilities in InnoDB engine.[6] The internal representation of a MySQL 5.6 table has a maximum row size of 65,535 bytes, even if the storage engine is capable of supporting larger rows. This figure excludes BLOB or TEXT columns, which contribute only 9 to 12 bytes toward this size. For BLOB and TEXT data, the information is stored internally in a different area of memory than the row buffer. Different storage engines handle the allocation and storage of this data in different ways. According to InnoDB specifications, it can store an object of size maximum to 64TB.[6] Now according to the InnoDB engine specifications, LONGBLOB and LONGTEXT columns must be less than 4GB, and the total row length, including BLOB and TEXT columns, must be less than 4GB. This means that MySQL 5.6 BLOB or TEXT object can be of 4GB of maximum size. If the size of the unstructured data is higher than this, the only way

out is divide the data into pieces and store in the database. This is really impractical generally for binary unstructured data like multimedia data.

The known limitation with LONGTEXT is: If a TEXT column is indexed, index entry comparisons are space-padded at the end. This means that, if the index requires unique values, duplicate-key errors will occur for values that differ only in the number of trailing spaces. For example, if a table contains 'a', an attempt to store 'a ' causes a duplicate-key error. [6] There are no specific products available to manage unstructured data in MySQL 5.6.

4. Conclusion

Above 90% of the data are in unstructured format in enterprises. The unstructured data is having important content and it has high amount of relevance with the structured data, which is stored generally in relational database management system. Looking at the quantum and relevance of unstructured data, it is highly desired to keep the unstructured data in the database management system to gain following advantages:

1. Inclusion of unstructured data with structured data in various database operations like storing, querying, analyzing.
2. Basic features of database management system like; availability, scalability, reliability provides the same advantages to unstructured data also.
3. Indexing and compression of unstructured data could be automized, which helps in performance gain in various operations on unstructured data.

Following are the basic differences in the way Oracle 11g and MySQL 5.6 handles unstructured data.

Table 4.1

Feature	Oracle 11g	MySQL 5.6
Exclusive LOB to handle unstructured data	Yes, SourceFiles	No
Maximum size of unstructured data unit	8 TB to 128 TB	4GB
Compression of unstructured LOB's	Yes	No
Encryption of unstructured LOB's	Yes	No
Support to a medical specific image format DICOM	Yes	No

Based on above comparison, it is very much clear that Oracle 11g is having much more advanced mechanisms to handle unstructured data. Oracle 11g is the pioneer in large object technology and they have a definite advantage over other databases in terms of performance and storage capacity. Oracle 11g is having some advantages over MySQL 5.6 like

,

- Higher storage capability to handle unstructured data
- Feature specifically developed for unstructured data (SourceFiles)
- Unstructured data compression and encryption mechanisms available

References:

[1] Ashish Nadkarni, Natalya Yezhkova, "Structured Versus Unstructured Data: The Balance of Power Continues to Shift", IDC (Industry Development and Models) Mar 2014 (Doc # 247106)

[2] Marcelle Kratochvil, "Managing Multimedia and Unstructured Data in Oracle Database" , Packt publishing.

[3] "Managing Unstructured Data with Oracle Database 11g" – Oracle white paper (www.oracle.com/us/products/database/options/spatial/039950.pdf)

[4] http://www.dicomlibrary.com/about/

[5] http://dev.mysql.com/doc/refman/5.6/en/blob.html, MySQL 5.6 documentation on data types

[6] http://dev.mysql.com/doc/, "MySQL 5.6 Enterprise Edition Reference Manual"

[7]https://docs.oracle.com/cd/B28359_01/server.111/b28320/limits001.htm#i287903, "Datatype Limits"