

B03.2

Delivering OS MasterMap National Coverage

Don Boyack, Applications Development Team Leader DEFRA and Nigel Benham, Independent GIS Database Designer/Consultant

Abstract

When the Ordnance Survey decided to switch from Land-Line to OS MasterMap the repercussions for users was more significant than simply changing the file format. This paper will focus on the impact that OS MasterMap has on a business using National Coverage.

The paper will focus on key decisions that managers will have to make, e.g,

- Choosing a Relational Database to store OS MasterMap
- Management and manipulation of OS MasterMap within GIS software.
- I.T. infrastructure and network considerations.

The paper will then focus on the use of Oracle Spatial as a Relational Database solution for OS MasterMap storage. Particular attention will be paid to the issues faced by DEFRA (Department of Environment, Food and Rural Affairs).

Introduction

Management of spatial data due to its complex nature has, in the past, been in the domain of GIS vendors. These formats, such as shp and tab, are commonly referred to as flat-files because the data (spatial features and attributes) stored within are structured in a way that makes the file inflexible. For example, multi user access to the same file is virtually impossible. Access to individual features within the file is not possible, therefore forcing the GIS to open the entire file, this can be time consuming and an inefficient use of computing resources.

Recent improvements to flat file formats has seen an attempt to make them user friendly spatial indices allow relevant portions of a file to be opened in a GIS. However, they still lag behind relational database management system when it comes to data management in a scalable, multi-user environment.

The growth in the use of geographical information has prompted database vendors, such as Oracle and Microsoft to extend their relational database software to cope with the complex nature of spatial data. By doing so Oracle have been able to capitalise of years of development in the data management field and take spatial data management away from GIS vendors and bringing it into the mainstream.

In November 2001 the Ordnance Survey (OS GB) launched a new digital mapping product called OS MasterMap. Unlike previous spatial data formats, the data is supplied in GML (Geography Markup Language) and is updated on a 'change-only' basis. To gain maximum benefit from this product and to maintain and update it efficiently, many users are to relational databases as a storage mechanism.

This paper will outlines a 'generic', strategic approach that can be taken when attempting to store large volumes of OS MasterMap data in an Oracle database.

Data Storage

As mentioned above, traditional flat file storage is not suitable for maintaining large volumes of OS MasterMap data due to the format of the data (GML) and the detail of attribution. The storage and maintenance of OS MasterMap is best facilitated by using a relational database, more specifically a 'spatially aware' database system such as Oracle Spatial, which is a part of Oracle's Location-Based Services for 9i.

National coverage of OS MasterMap consists of over 400 million features, the prospect of managing this level of data can prove daunting even to the most die-hard of database administrators. Even breaking the data down into geographic areas or regions can still mean contending with millions of features.

General issues users need to be aware of include;

1. Data Storage
2. Data Conversion
3. Load Strategy
4. Data Model Design Issues

Storage

To support the demand for maintainable data structures, businesses are moving away from flat file storage to the relational database management system (RDBMS) for their GIS requirements. Advantages of the RDBMS over the flat file include:

- Single centrally managed data storage
- Multi-user access
- Scalability - Proven ability to store large volumes of data (Terabytes)
- Ability to store and update data according to your business requirements – Storage Model
- SQL (Structured Query Language)- SQL is the industry standard interface to a relational database

Feature level maintenance of OS MasterMap data is only available if the data is stored in a relational database. Database companies such as Oracle have responded well to this shift by extending their database schemas to allow GIS data to be stored and manipulated as if it was ordinary data. Oracle has invested heavily in extending its SQL data manipulation language to allow server side GIS computations and analysis to be carried out on the data. For example Oracle Spatial supports the following spatial operands:

- Contains
- Coveredby
- Covers
- Disjoin
- Equal
- Inside
- Overlap
- Touch

Oracle have also developed the RTree Index which allows spatial data to be indexed (stored and retrieved according to a particular RTree algorithm). This allows data to be retrieved quickly. Oracle Spatial is capable

of carrying out GIS analysis on volumes of data that a standard desktop GIS could not readily cope with. This adds a new exciting dimension to GIS analysis and introduces new flexibility to business models.

Oracle Partitions

Oracle9i Enterprise Edition supports table partitioning for large databases.

National cover of OS MasterMap requires approximately 500mb of disk storage. Oracle recommends that this quantity of spatial data should be stored in a single partitioned table.

Partitioned tables allow very large databases to be broken into smaller, manageable pieces called partitions. Each partition can be managed (e.g. deleted, split, merged, indexed etc) independently of other partitions, thus providing a storage model that can be more effectively tuned for availability and performance.

When a spatial index is built on a partitioned table, the index maps itself to the partition structures of the table. For each partition, a corresponding index partition is created. The partitions are transparent to the end user.

The structure of the partitions is controlled by the DBA and can therefore be based on the business model. For example, a partition can be a geographical area such as a town or OS national grid tile.

The key to fast retrieval of spatial data from a partitioned table is to ensure that during a geographic query, only those partitions that contain the target geographic area are searched. One way to guarantee this is to implement a storage model that is based upon spatially organised partitions. In order to organise the data into partitions each feature must have a partition key associated with it. If the GIS being used retrieves data from a particular partition it can do it by either:

- using a spatial query – return all data within the following spatial extent
- using partition key – return all data where partition key = x (relational search as opposed to a spatial one)

Designing a Partitioned Storage Model

Under the old NTF tile system the Great Britain is geographically divided into 100km squares. The same tessellation process can be used when building a storage model for OS MasterMap. A table can be partitioned to mimic the National Grid. This means there should be one table partition for each 100km National Grid square e.g. one partition for 'TQ', one partition for 'TL' and so on. There also needs to be 'UK or GB' partition(s) for features with spatial extents that cross more than one tile.

Once the table has been set up the Database Administrator (DBA) will then have to load the OS MasterMap data into the relevant partitions. This can be a little tricky because the OS MasterMap data is not supplied in chunks that relate to the national grid 100km squares.

DBA's will have to write database procedures that compute the location of every spatial feature in the database and assigns it a partition key value which associates it with its relevant table partition. For example, all OS MasterMap features that are geographically located in the national grid system under tile 'TQ' should be assigned a partition key value of 'TQ'. OS MasterMap features can then be moved to their relevant table partitions by virtue of their partition key.

Data should be loaded in manageable volumes into a temporary table before the partition key is computed. Once computed the data can be transferred to the partitioned table.

Faster queries

Spatially organised partitions can radically improve database performance. A geographic query performed within tile 'TQ' would only need to search the spatial index associated with partition 'TQ' and the UK partition. All other partitions would be quickly skipped.

Data Loading

The Ordnance Survey has adopted GML2 (Geography Mark-up Language) as the standard for distributing OS MasterMap to its customers. The data is supplied in a compressed format called Gzip which compresses the data by a ratio of 20:1 making the data easier to transfer.

The GIS industry has responded by developing tools for loading OS MasterMap data into database such as Oracle (Spatial and Locator).

Cadcorp has developed a wizard based tool for loading OS MasterMap to either Microsoft Access 2000, SQL Server 2000 or Oracle 9i. The wizard provides a step by step approach for bulk loading OS MasterMap into either of these database. The loader has been designed to be flexible for use with different business models and therefore offers users the ability to filter data by theme, attribute and geometric type. The loader automatically removes duplicate features during a data load process. The Cadorp SIS - OS MasterMap Loader also allows supports 'Change Only Updates' files.

Controlling Data Storage

OS MasterMap contains a large amount of information. Users wanting to reduce their total storage requirements can do so by eliminating data they do not need. Most OS Mastermap loaders offer filtering mechanisms to do this.

Spatial Indexing

To gain maximum performance from the data loaded, a spatial index needs to be created. The business model needs to allow for this task due to the time taken to create an index and its storage requirements. Oracle supports 2 types of indexes: Quad-Tree and R-Tree indexes. R-Tree indexes are the preferred index because they have better performance and are easier to create. Points to remember are ;

Storage space

The index occupies approximately 20% of the disk space used by its corresponding table. However, during index creation additional disk space is required for associated index computation tables (these are removed after index creation). The total amount of disk space required is 40% of the size of the corresponding data table.

Index Creation Time

The following figures are based on OS MasterMap data for Somerset :

Number of features:	8,388,179
Total Gzip files:	515 mb
Oracle Spatial table:	3,728 mb (7 times the gzip files)
Spatial Index :	729 mb (20% of the oracle spatial table size)
Spatial Index + computations :	1,492 mb (40% of the oracle spatial table size)

Although spatial indexes in Oracle 9i are self-maintaining, when appending data to a database table, Oracle recommend that the index be removed before the load starts.

Design Issues

Disk space

Oracle will require tablespace to manage the data, indexes and growth due to Change Only Updates being applied (approx 2-3% per annum). In addition, you will require large amounts of disk space during index creation (depending on partition sizes). Space allocation will also depend largely on how table schemas are defined. However, for full national coverage, including all themes and attribution, a minimum of 500 Gbytes of disk storage is recommended.

Partition tuning

Features are not evenly distributed across all the national grid tiles. For example, tile 'TQ', covering London, is densely populated when compared to rural areas. The storage model may need to be refined to

improve performance when using data contained in this particular partition. This partition may have to be tessellated into sub partitions.

Change only updates

The frequency at which 'Change Only Updates' are applied will impact the business. Frequent updates result in more down time for the database, especially when recreating spatial indexes.

Temporal Datasets

By adding a load date to the data initially loaded and then to all subsequent 'Change Only Updates', it will become possible for the data model to incorporate temporal data management. The date stamps can be used to create database Views of data any given point in time.

Data Management Issues

Features are uniquely identified by the Topographic Identifier (TOID) coupled with the version number. It is imperative that the data model uses both these values to identify unique features. The 'Change Only Update' systems are user driven, this will mean that businesses may update their data holding at different frequencies and at different times. Users may have different versions of the same spatial feature in their respective databases. In order to avoid any confusion features should be identified by both TOID and Version numbers.

Interoperability

Oracle stores data in a published format and most leading GIS systems are able to read and write spatial data to Oracle. OpenGIS Conformance is one marque which users should also begin to familiarize themselves with in this context.

Summary

At the time of writing this paper Oracle was the only relational database vendor actively involved with the GIS community in building scalable OS MasterMap data storage solutions.

Building a data model for national cover of OS MasterMap requires a thorough understanding of Oracle's spatial capabilities. GIS users should be aware of the impact the move from traditional 'flat file' to intelligent GIS use as the database comes of age within the GIS community.

Don Boyack

Applications Development Team Leader

Nigel Benham,

Independent GIS Database Designer/Consultant nigel.benham@defra.gsi.gov.uk