The Met Office’s Logical Data Store

Bruce Wright, John Ward, Malcolm Field, Met Office, United Kingdom

Background

Data are the lifeblood of the Met Office. However, over time, the ‘organic’, un-governed growth of data management solutions has led complex, incoherent and (often) undocumented IT systems. This ‘information silo’ approach has resulted in a number of organisational issues:

- Inconsistent, locally processed data;
- Proliferation of data & data access mechanisms;
- Poor access to ‘enterprise’ information assets.

The business drivers for change are:

- Cost efficiency;
- Improved agility in developing new solutions;
- Improved consistency of information.

The concept of a Logical Data Store (LDS) is the means by which these issues will be addressed.

Logical Data Store concept

The concept of the Logical Data Store (LDS) is that of a single, logical repository for all core (shared) enterprise meteorological information, delivering information that is:

- Consistent;
- Uniquely identifiable;
- ‘Spatially-enabled’ (facilitating spatial manipulation and querying);
- Accessible through a set of common interfaces
- Managed in a standard way

The LDS forms a key part of the future information architecture (figure 1), ingesting data from and serving data to all the other components.

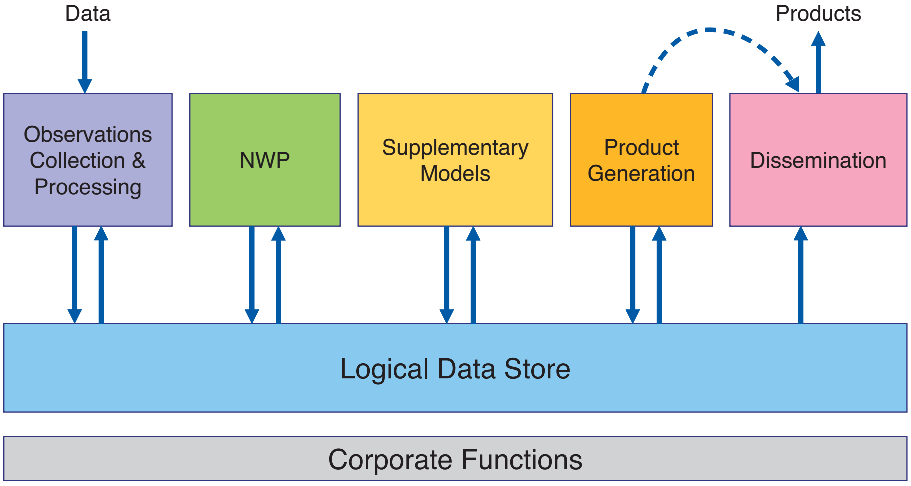

Fig. 1 High-level Information Architecture

The LDS is composed of a small number of high-level components which can be characterised by their interaction with each other and other external entities (figure 2). Key logical components:

- Information Lifecycle Management administers the data though it’s lifecycle, from ingestion to deletion, on behalf of the data owner;
- Data Discovery Portal provides the facility to search for information;
- Public Interface provides the access to the information in a standard way;

- Shared Data Model provides a set standard solutions for the internal management of the data;
- Catalogue Records hold metadata, both static and a limited amount harvested from the data.

The latter two are only exposed through the first three interfaces and can be regarded as ‘under the covers’.

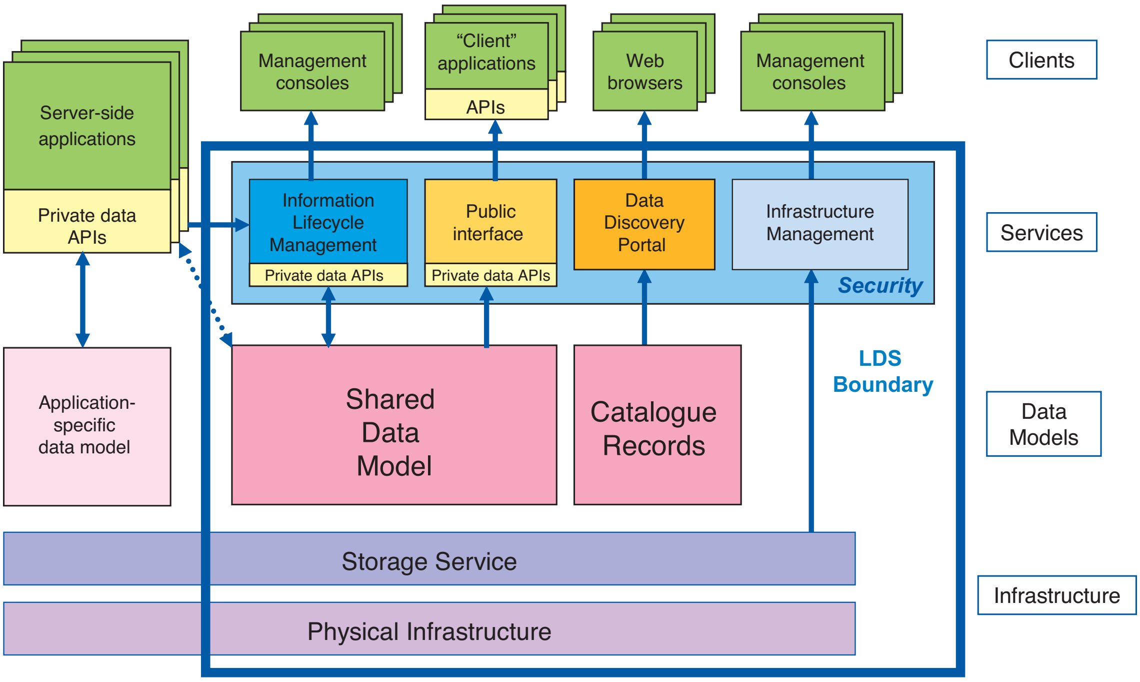

Fig. 2 LDS component diagram

LDS Public Interface

The ‘Public Interface’ is the key to the LDS concept[1]; it should:

- Hide the complexity of:
  - Databases & archives;
  - Formats & codes;
  - Interfaces to different data types.
- De-couple the client application from the data store.
- Provide:
  - A single way to access all data in the LDS;
  - Using a standard request (metadata).

As such it should provide controlled, secure read access to the LDS, through a set of well-defined, common interfaces, using a range of selection criteria, and delivering both data and products (and metadata). The capability of getting a resource can be expressed in terms of:

- Select Dataset (a set of data , rather than a physical file (e.g. an NWP model run, all the UK climate observations);
- Subset:
  - Geospatial extent (e.g. latitude-longitude box);
  - Temporal extent (e.g. validity time);
  - Parameter / attribute (e.g. screen temperature)
  - Domain restriction (e.g. temperature > 28degC)

[1] Note: In the development of the LDS, it is also the intention to:
- Rationalise and consolidate data stores;
- Take advantage of new data management technologies.

- Process:
  - Transform (i.e. isomorphic (reversible) process, such as temperature conversion from Kelvin to degrees Celsius);
  - Derive (i.e. non-isomorphic (irreversible) process, such as computation of dewpoint temperature from temperature, humidity and pressure);
  - Re-project (i.e. computation of the co-ordinates of the data points of features in a new projection given the co-ordinates in an initial projection - re-projection of raster (gridded) data usually also implies a re-gridding)
  - Interpolate/re-grid;
  - Filter (i.e. sub-sample data to, for example, reduce the resolution by a factor).
- Format (Change the data into a particular format different from that of the underlying data, this includes the transformation of the data to specific precision, number format and packing (or compression), the combination of the appropriate data and the derivation of the required metadata. There should be a restricted range of forms for the returned data, including: PP/Fieldsfile, netCDF, GRIB, BUFR, XML, CSV)
- Return Information.

The solution will need to support access from a range of clients, including a web client, Java, C and FORTRAN.

The Public Interface will make use of web services, as they:

- Use an HTTP Transport for messages (like web pages), providing:
  - A high level interoperability;
  - Clear and simple client-server interaction.
- Use XML as a standard form for the request and response, facilitating:
  - Self-describing data;
  - Implementation of metadata standards;
  - Use a standard schema (e.g. GML).

In particular, it is intended to use OGC[2]-compliant web services, such as the Web Feature Service (WFS), which characterises the data as a set of (real world) 'features' and make use of Geography Markup Language (GML). The possible risks are the poor performance (especially for voluminous data), and difficulty in getting the new technology to work.

Work in already completed or in progress

Some work has already taken place on:

- Consistent use of Oracle RDBMS to hold a range of data types;
- Standard Java interfaces to the database;
- Web Services with XML for data exchange;
- Draft Met Office Metadata Standard building on the ISO191xx, WMO standards and CF Convention;
- Standard components for deriving best climatological observation values from our archive.

Currently, various other initiatives are underway:

1. *An operational lightning location database (figure 3 shows the demonstrator), which will:*

- Store direct to Oracle Database (data and products);
- Provide Web Services interface (initially using bespoke XML, but probably ultimately as a Web Feature Service).

[2] Open Geospatial Consortium; see: http://www.opengeospatial.org/

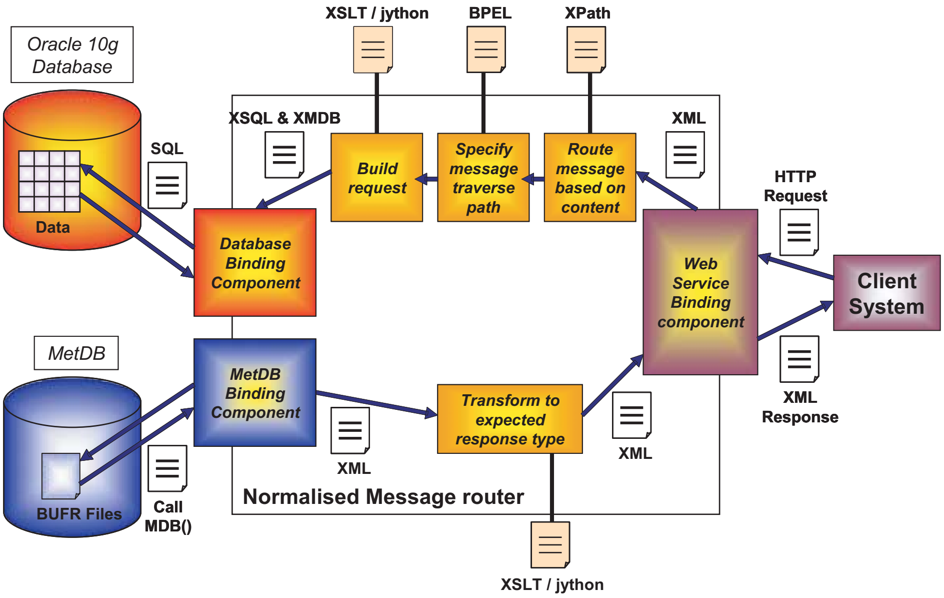

Fig. 3 Lightning location Web Service demonstrator

2. *Investigations into the use of the Oracle 10g Database cluster functionality, with a 3-node Dell cluster using Oracle RAC (Real Application Cluster) (figure 4).*

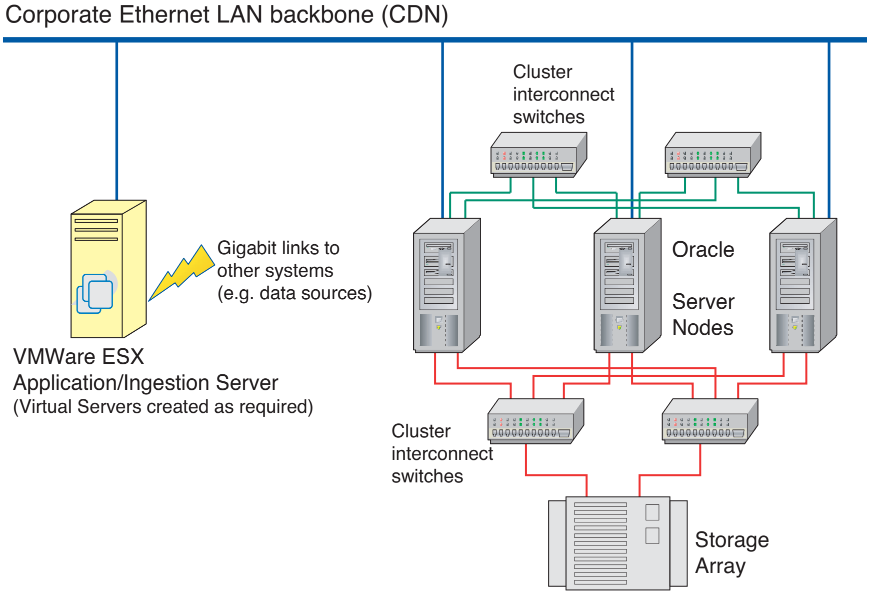
Corporate Ethernet LAN backbone (CDN)
Cluster interconnect switches
Gigabit links to other systems (e.g. data sources)
Oracle
Server Nodes
VMWare ESX Application/Ingestion Server (Virtual Servers created as required)
Cluster interconnect switches
Storage Array

Fig. 4 Hardware architecture for the Oracle database cluster

3. *A proof of concept exercise looking at the use of an Oracle RDBMS the data management and provision for the NWP process, will investigate:*

- Extraction of observations directly to supercomputer from database;
- Storing NWP forecast output direct from supercomputer into database;
- Providing application data access using current (FORTRAN-based) APIs.

Approaches to be adopted

The Logical Data Store will not be implemented using a 'Big Bang' approach, as this is very high risk, requires a huge amount of effort and it would take along time before the benefit was achieved.

Instead, the approach will be 'Piecemeal', focussing on specific data types to:

- Prove the approach (technical solution);
- Demonstrate end-to-end capability;
- Address existing problems;
- Provide 'quick wins'.

However, the work will form a part of a longer term plan.

Use will also be made of two other approaches:

- 'Wrappering' to interface to existing data stores (for the present):
  - Where migration costs are high (e.g. the archive);
  - To provide simple migration paths to use LDS.
- 'Proxy' interfaces, i.e. provide access to the LDS via existing (legacy) interfaces for widely used interfaces:
  - To allow partial/gradual data migration;
  - To allow gradual application migration.

Other parallel activities

These are a number of external activities that will influence the development of the Logical Data Store; two are:

- SIMDAT (EU co-funded project to promote use of GRID technology), which will:
  - Involve collaboration with ECMWF, DWD, MeteoFrance & EUMETSAT to deliver a 'meteorological scenario' for the Future Weather Information System;
  - Be developing a catalogue for managing distributed data.
- DEWS – Demonstrating Environmental Web Services (a DTI co-funded collaborative project):
  - Using leading edge technology in real scenarios for health & marine;
  - With academic (BADC, ESSC) & commercial (Lost Wax, BMT, Intel) input.

Contact Information

For further information, please contact either:

Bruce Wright (bruce.wright@metoffice.gov.uk); or:

John Ward (john.c.ward@metoffice.gov.uk).