Data Warehouse and Analytical Tools

Development and Implementation
Who we are:

The Information Centre provides Social Security related statistical and management information outputs to meet the needs of internal and external reporting requirements.

Originally - a statistical data processing & analysis Branch, serving mainly policy advisers, Ministers & Parliament. Success of ASSIST has given us a broader remit for Departmental information.
What we have done:

Transformed the way we operate by building and implementing a modern data collection, processing and analysis system that handles many diverse, complex and large datasets.

Impact on DSS:

Respond quickly to change and deliver information that meets new and evolving needs.
Introduction

Why?
Rich source of data on clients, staff & business
Chimney/silo based - individual benefit/pension
Difficult to exploit through old IT systems

Key drivers from Government:
Modernisation - Look across client groups
Link information on performance & resources to outcomes
Introduction

What did we need?

Extract data from a wide range of source systems
Bring together into a single data source
Be confident about its quality
Present it in a format to suit the business
Give flexibility to present and analyse it
e.g. by time period and spatial unit
DSS Information Centre

Information Delivery Model

Top Level
(PSA/SDA)

Information Level
(Analysed and Assured Data - information to manage the business)

Source Data Level

Information Datasets

- Corporate Services
- Children and Families
- Working Age
- Pensioners

Geographical Splits

Departmental Reports

Time

Finance
Performance Monitoring
HR
Fraud and Accuracy
Research
National Statistics
Original development using PFI failed - different approach

Generic solution - software & database

Use best tools for each job & integrate together

Those with knowledge of data build the warehouse

Users can customise the system for each Project

IT staff not needed for core business

Phased approach with achievements at each stage

Full user control & involvement
Integrated information system

- IC staff run the Data Warehouse
- Minimal support from IT staff
- Analysts have direct access to data
- Outputs are created for wider user community
The Development

- The Problem
- Business Requirement
- Business Solution
- Technical Solution
- Development Method
- Implementation
- Delivered System
The Problem

Where did we start from?

- Outdated technology;
- No data warehousing (Tape Library);
- Inflexible and lengthy time scales;
- Individual dataset processing on multiple systems;
- Reliance on teams of programmers & specifications
- Analysts/producers had little control over processing;
- Cross-matching and ad hoc reporting difficult;
The Problem

Numerous Data Sources

Data
Data
Data
Data
Data
Data
Data
Data
Data
Data

Programming Teams

Programs
Programs
Programs
Programs
Programs
Programs
Programs
Programs
Programs
Programs

Shared Service

Mainframe

Tape Library

Standard Outputs

Output
Output
Output
Output
Output
Output
Output
Output
Output
Output
Business Requirement

- Develop a generic system to support business;
- Data warehousing providing on-line access;
- Enable easy cross-matching/longitudinal;
- Enable time scales to be shortened;
- Provide Users with full control over processing;
- Allow flexible and low cost change;
- Eliminate/reduce technical support requirement.
Numerous Data Sources

Any format

Analytical Services Statistical Information System (ASSIST)

Data Collection
Validation and Cleaning

Data Repository
Clean Datasets

Publication Area
Hardcopy and electronic

Extract, transform & quality assure

Info/analysis layer

Dissemination layer

Outputs
ASSIST Technical Solution

Newcastle System
3 x Alphaserver

Validation and Cleaning
- Visual Basic
- SQL
- ORACLE

Data Repository (Newcastle)
- SAS
- ORACLE

Information Centre (inc. website)
- SAS/PDF/Excel
- Webserver

London System
1 x Alphaserver

Data Repository (London)
- SAS
- ORACLE

250 Networked Users
- Newcastle
- Leeds
- London

2 TB data
3 TB current capacity
40 GB per month
Expanding
Development Method

- JAD/RAD techniques with high User involvement;
- Dynamic Systems Development Method (DSDM);
- Development Team co-located with key User area;
- Feasibility Study, Functional Prototype, Final System;
- High use made of prototyping;
- User managed development - prioritise & agree.
Development and Implementation

Feasibility Stage
- "Proof of Concept"
  - Test Data
  - Show it can work

Functional Prototype
- "Working Model"
  - First "Live" Data
  - Use it for actual business

Final System
- "Full System"
  - Phased take-on of "Live" Data
  - Roll out in controlled way
Delivered System

Analytical Services Statistical Information System (ASSIST)

- Data Collection
- Validation and Cleaning

Data Repository
- Clean Datasets

Publication Area
- Hardcopy and electronic

Outputs
Delivered System
Data Collection

Data Collection

Validation and Cleaning

- Loads Source Data
- Validation and Derivation
- Validation Analysis
- Manual Cleaning
Extracting & transforming data

- *Describe source data to system*
- *Decide how you want it to be formatted and stored*
- *Map the source to the desired layout*
- *Standardise data -*
- *eg inconsistent code values, field formats*

Relational data set in a suitable / consistent format
Delivered System
Data Collection

Generic functionality allowing source data to be described to the system - as per record description.
Delivered System
Data Collection

SAS naming for data warehouse - fields can be dropped if not required.

Source and derived data now in ORACLE format.
Delivered System
Data Collection

Source data mapped to ORACLE database and put into relational format
Delivered system
Data collection

Validating and correcting the data

- *Rules based system - user defined & created*
- *Check consistency, credibility & accuracy of data*
- *Compare with other data sets / previous values*
- *On-line changes and status reports*
- *Automatic and manual changes - individual records*
- *Create new data sets from various source data sets*

Move data set to repository when owner is content
Validation and Derivation rules are created & documented
Delivered System Rules

Conditions can be defined.
Delivered System Validation Analysis

Error analysis provides information on how clean the data is - continually updated.

Spot errors with source data

Analyze type of error
Delivered System
Manual Cleaning

Individual records can be viewed and corrected - on-line application of rules.

Clean all or part of dataset

Customised screen
Delivered System
Data Repository

Data Repository

Clean Datasets

Online Access
Search and Find
Merge and Sub-setting
Reporting
Delivered system
Data repository

Analysing the data - SAS based solution

• **Warehouse with a series of relational data sets**
• **Meta data and search facilities**
• **Front-end tool to select and merge data (high speed)**
• **Longitudinal & cross-sectional merge**
• **View output dataset**
• **Front end tool to create reports and export to Excel**

High quality reports without having to write code
Delivered System
Data Repository

Access to data through hierarchical structure

Basic and Advanced search facilities on metadata

Point & click
Delivered System
Data Repository

Selected datasets
to merge

Data items
to include

Merge, rename & view features

Generates SAS code
Delivered System
Data Repository

User interface allowing simple tabulation - generates SAS code

Also use base SAS code
User interface for 2 dimensional reporting and export to EXCEL

Builds table, ie row & column groups & totals

Apply conditions

Select variables

Counts & stats
Delivered System Dissemination

Analytical Services Statistical Information System (ASSIST)

Data Collection

Data Validation and Cleaning

Data Repository

Clean Datasets

Publication Area

Hardcopy and electronic

Outputs
Delivered System Dissemination

Hardcopy Publications

Publication Area

Hardcopy and electronic

Electronic Applications

Intranet - Internet
Delivered system
Dissemination

Disseminating the output

• Approach to suit limitations of DSS infrastructure
• Website - PDF documents & Excel tables to download
• MDDBs & applications to send to customers via CD
• Excel sheets via E-mail
• Hard copy publications

Still need to use web applications

Catering for a diverse user base
Delivered System Dissemination
Delivered System Dissemination

SAS MDDB Application
Delivered System

Analytical Services Statistical Information System (ASSIST)

Data Collection

Validation and Cleaning

Data Repository

Clean Datasets

Publication Area

Hardcopy and electronic

Outputs
Impact

- Major improvements in timeliness of data;
- Opened up access to data to analysts (forecasting);
- Process & merge data easily (JSA);
- Created new types of analyses (CBA);
- Take in very large volumes of data (ONE/LA);
- Take in data from all DSS systems;
- Produce a single set of management information.
Summary

- Full end-User involvement in development;
- Single Generic System;
- End-User in control;
- Minimal technical intervention in processing;
- Flexible and easy to change;
- Easy access to data;
- Simple reporting functionality;
- User buy-in and significant business improvement.
David Frazer or Nigel Sowerby
DSS Information Centre
Room B2613
DSS Longbenton
Newcastle upon Tyne
NE98 1YX
United Kingdom

Telephone : 0191- 2257799
E-Mail : d.frazer@ms03.dss.gsi.gov.uk
: n.sowerby@ms01.dss.gsi.gov.uk