Centralization and Regionalization at the National Agricultural Statistics Service

Roger Schou
National Agricultural Statistics Service
IBUC XV
Washington, DC, USA
Centralized Surveys

- IBUC XIII
  - 1 survey – Mink (April 2010)
- IBUC XIV
  - Approximately 25 surveys
- Currently
  - Approximately 55 surveys
    - Weekly, bi-weekly, monthly, semi-annual, annual, and pentennial
Blaise Data Storage at NASS

- MySQL Database
  - All surveys in one set of tables
- Generic In-Depth Storage
  - Allows for one ETL (Extract, Transfer, Load) program to copy necessary data from transactional MySQL database to the analyticl Redbrick database
  - Eight Blaise tables
- A few additional NASS-specific tables
Generic In-Depth Tables

- BLAISE_DICTIONARY
- BLAISE_ID
- BLAISE_CASE
- BLAISE_FORM
- BLAISE_KEY
- BLAISE_DATA
- BLAISE_REMARK
- BLAISE_OPEN
BLAISE_DICTIONARY

• Catalog of all the surveys
  – Each instance of a survey has a unique DMKEY (data model key)
    • Each time a data model changes, a new DMKEY is assigned
    • Can be as minimal as a data model name
    • We use the survey’s folder name as the data model name, so we programmatically change it for weekly and monthly surveys
    • External sections are created on the fly and referenced via INCLUDES in the data model code
BLAISE_ID

- Contains all block names and field names
  - All metadata about the block/field
    - Including FieldTag, DescriptionText, & ArrayIndex

- NASS Cameleon scripts that created item code related files have been replaced by VB.NET code utilizing the Blaise API

- Table is not used much by the Blaise system, but it’s a critical table for NASS
Item Codes and Varnames

• NASS developed “hash notation” for item codes and varnames for repeated blocks
  – Example 1 in the paper illustrates hash notation (page 310)

• Arrays with item code and varname coded within the block.
  – Item code mapping file gets a special coding pattern to insure the correct element is used
  – Varname has Table_Row number appended
BLAISE_CASE

- Contains the unique JOINKEY and PRIMARY KEY for each record in the survey
BLAISE_KEY

• Contains all of the Primary and Secondary Keys as defined in an instrument
• BeginStamp is part of the key for this table which is needed if versioning is active
BLAISE_FORM

• Contains the status information for each form
  – Form status
  – Error count
  – Remark count
  – Don’t Know count
  – Refusal count
BLAISE_DATA

• Contains the data for the records in the survey
BLAISE_REMARK

- Contains the remarks left on fields for a survey
BLAISE_OPEN

• Contains the answers to any OPEN type fields
NASS CASIC Tables

• CASIC_SURVEYINFO
• CASIC_FAT
• CASIC_MANAGEMENT
• CASIC_EVENT_LOG
CASIC_SURVEYINFO

- Survey-level information
  - Instrument Name
  - Folder Name
  - BOI File Name
  - Assorted indicators
  - Some Start and End Dates

- Menu system makes extensive use of this table
CASIC_FAT

• Used to control access

• Identifies:
  – State to whom the record belongs
  – Data Collection Center (DCC) Assigned
  – Estimation Center (EC) Assigned
  – Region to which the state belongs
  – Region to which the DCC belongs
  – Region to which the EC belongs
CASIC_MANAGEMENT

• Update to the paper:
  – Key of the CASIC_MANAGEMENT table is now: DMKEY, JOINKEY, BEGININSTAMP

• Contains several fields that appear in nearly all of our instrument

• Table is indexed on these fields to increase the performance of instruments

• Allows record filters to be more efficient
CASIC_MANAGEMENT

• June Area instrument is exception to NASS “standard” fields
• Similar fields, but different blocks
  – Table was not populated by our system
  – Performance ground to a halt
    • Complicated instrument
    • No indexes utilized
    • An extremely large external file
CASIC_EVENT_LOG

- Used as a debugging tool
- Tracks the activity on the CASIC Menu
- Every button click registers a “BEGIN”
- When process finishes, registers an “END”
- We can detect what processes are running, who is running them, and how long they are running
Centralized Blaise Concept

• All data in one central location
• All surveys look the same in the database
  – Only one ETL needed to copy data
• Records logically separated using record filters
• Dynamic menu reacts to user’s location and role
LAN = Local Area Network
WAN = Wide Area Network

Eastern Citrix

Workstation ➔ LAN ➔ Blaise Data Server ➔ LAN ➔ MySQL
(1 ms latency)  (1 ms latency)

Western Citrix

Workstation ➔ WAN ➔ Blaise Data Server ➔ LAN ➔ MySQL
(30 ms latency)  (1 ms latency)
Future NASS Infrastructure

Eastern & Western Citrix

<table>
<thead>
<tr>
<th>Workstation</th>
<th>LAN</th>
<th>Blaise Data Server</th>
<th>LAN</th>
<th>MySQL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1 ms latency)</td>
<td></td>
<td>(1 ms latency)</td>
<td></td>
</tr>
</tbody>
</table>

Citrix Server, Blaise Data Server, and MySQL Database all located in Kansas City

LAN = Local Area Network
Hybrid Surveys

• Temporary Workaround
  – For high profile, quick turnaround surveys where 1 missed night of calling is not acceptable

• CATI data collection is done in a decentralized (local) Blaise data set, then sent to the Central MySQL Blaise data set via Manipula
  – Remaining processes done centrally

Currently being phased out
REGIONALIZATION

- Moved from 46 state field offices to 12 regional offices
  - Nearly complete
- New set of record filters now group by region
- CASIC System works regionally whether you are sitting in the Regional Field Office or in one of the states within a region
Other Centralization

• Other NASS systems undergoing centralization efforts
  – Survey Management System
  – Tracking and Control
  – Enumerator Skills Database

• Centralizing commonly used tables
  – County lookup/validation

• Inter-database communication
  – Blaise API
• Streamlining Processes
  – Centralization
  – Regionalization
  – Colocation of Servers
• Do more with less
• Future of Blaise at NASS
  – Only CATI software
  – One of two edit systems (after sunsetting legacy system)
Questions