FEDCASIC 2009

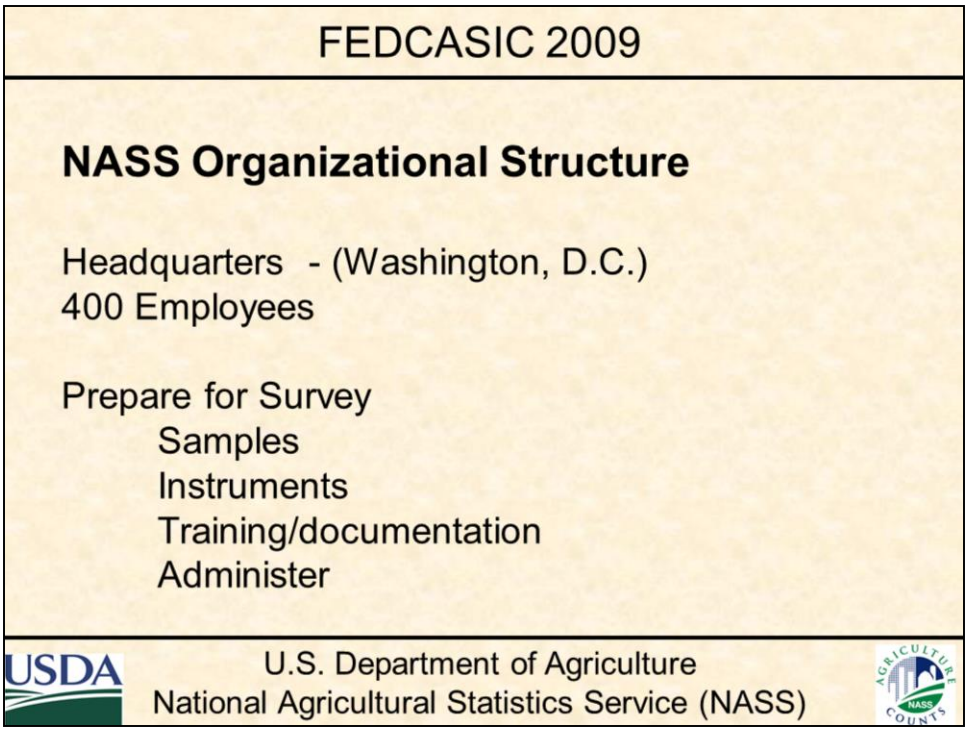
The Agricultural Census Is Not Just A Big Survey

David McDonell
March 18, 2009

U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)

Objectives

- Overview of Census Program
- Failures and Successes
- Impact on Edit System
- Applying Lessons Learned on Census to Surveys



Instruments

- Blaise program
- Analysis program
- Summary programs
- Estimation programs
- Publication programs

Estimation manual

Survey administration manual

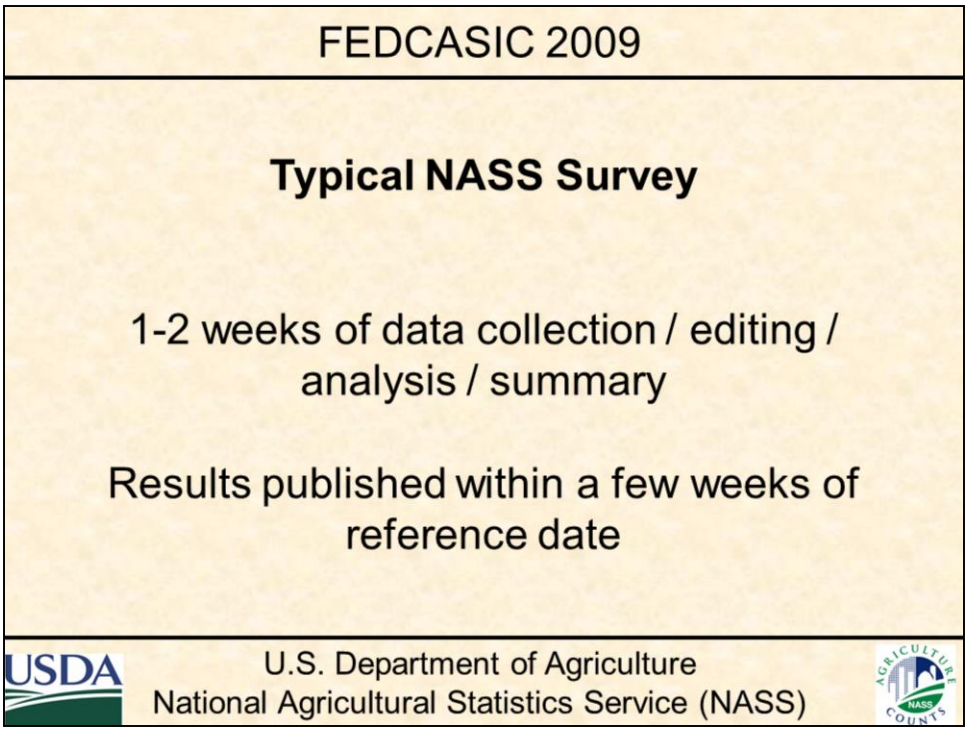
Interview's manual

NASS Organizational Structure – Continued

Field Offices (FO)

44 locations
600+ employees
2,000+ enumerators

Conduct Survey
- Collect
- Edit
- Analyze
- Set estimates for their state(s)

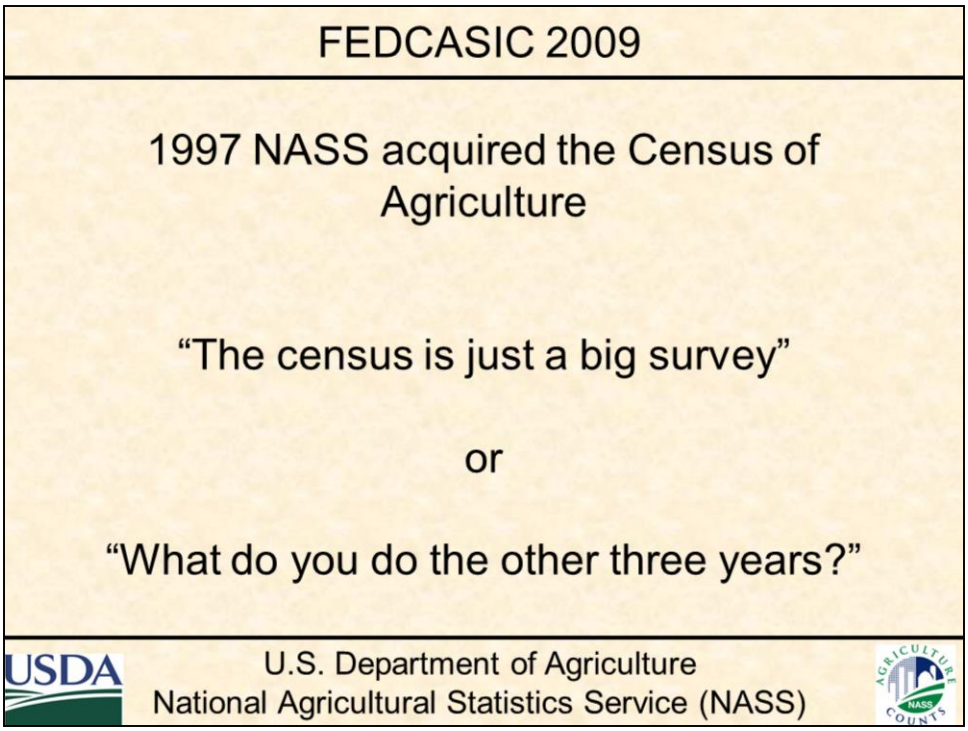


Able to touch and or review every record in a survey through editing and analytical review

Can be done with a relatively small number of statisticians

Most survey are inventory and production type

Not a lot of demographic and financial expertise

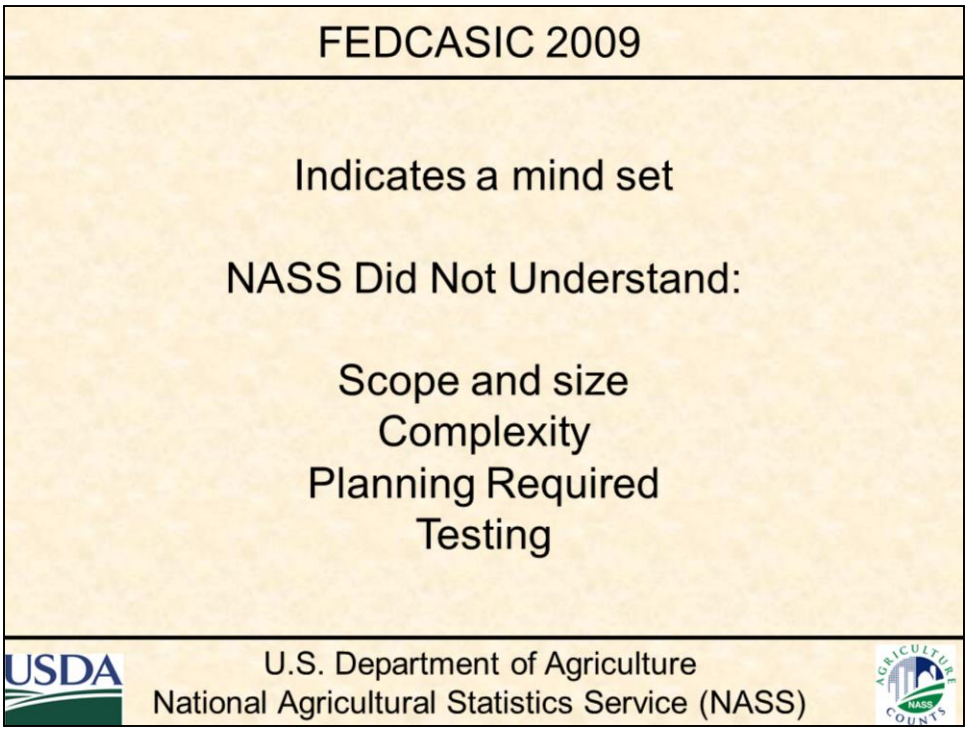


All the Farmers in the US - 3,000,000+ mailout

Takes 22 months from beginning to end

64 FTE's from Census

Quotes from upper management about the Census



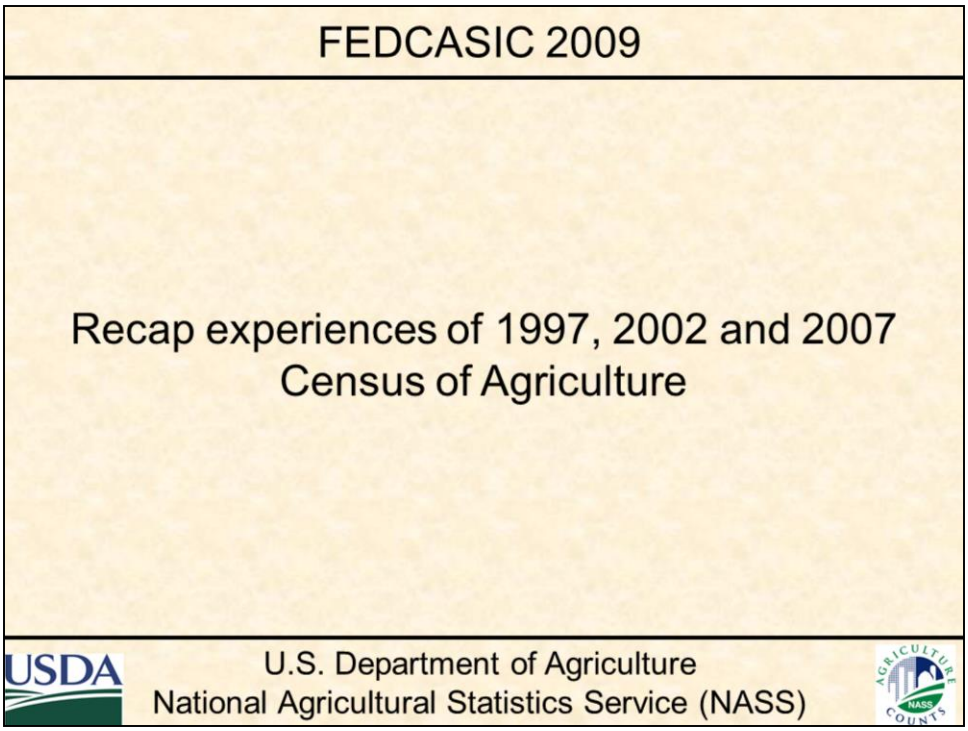
Management very naïve about the process and size

Had never conducted a survey with a mail list of 3,000,000 records before

Nor published a report with this much data

Have not had to plan for or understand the issues with this size survey

Did not address testing strategies or time requirements



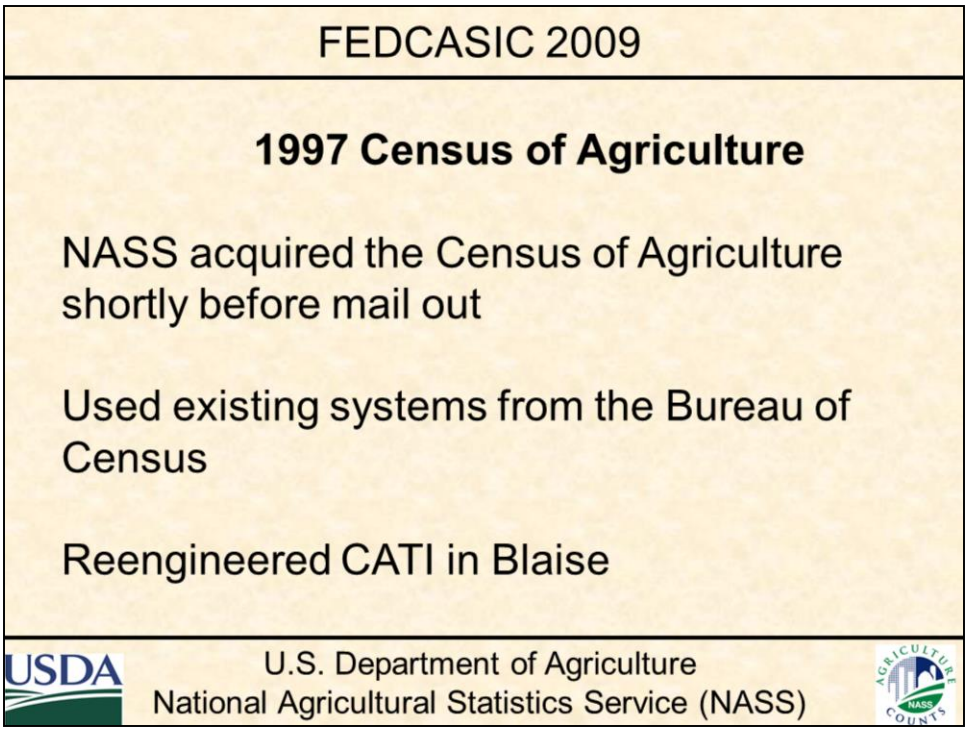
Management very naïve about the process and size

Had never conducted a survey with a mail list of 3,000,000 records before

Nor published a report with this much data

Have not had to plan for or understand the issues with this size survey

Did not address testing strategies or time requirements

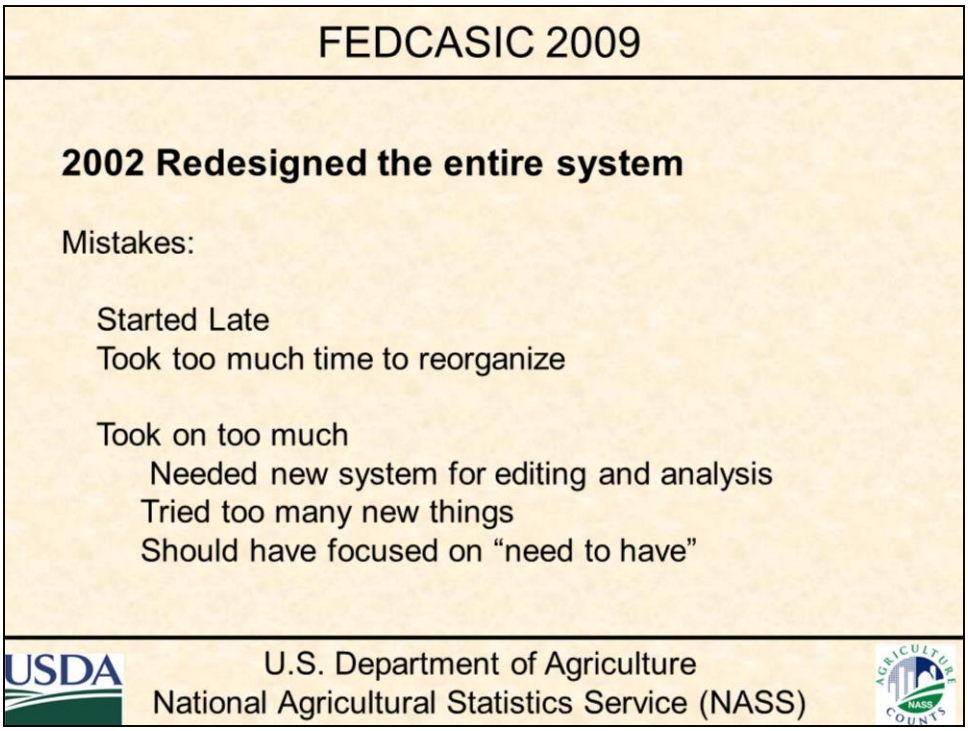


Officially acquired in Feb 1997. Mail out 12 1997

Developed for the 1987 Census, improved in 1992, used the same system in 1997

The CATI process did not work in 1992. So NASS wrote their own Blaise instruments.

These worked well and have been used since with slight modifications



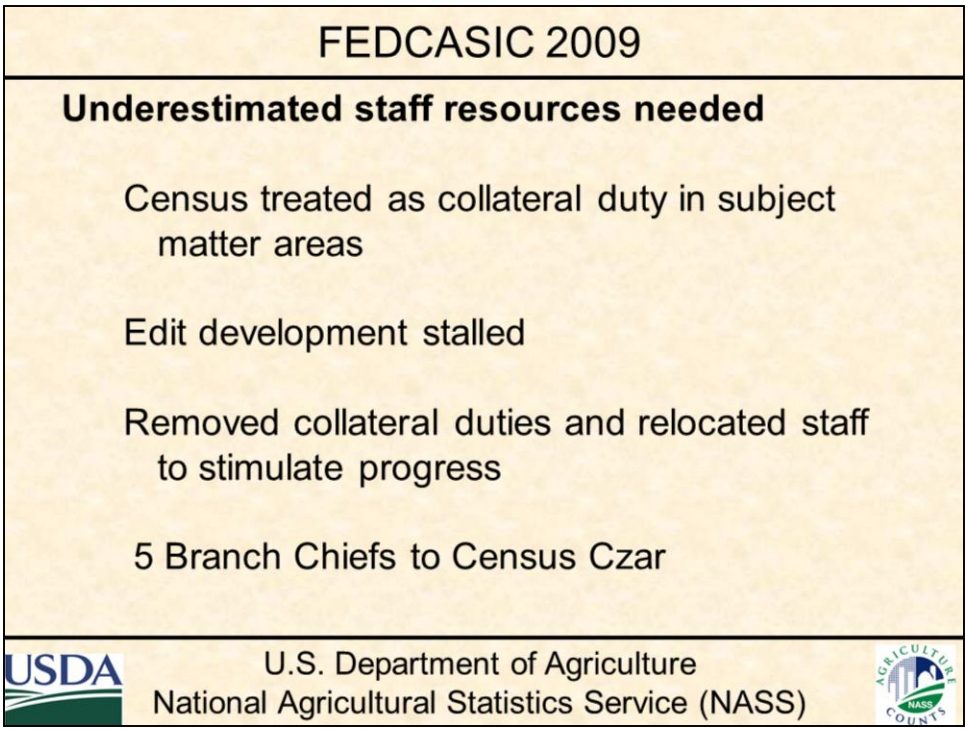
Re-organized and started working on the new process in early 2000.

Did not have enough knowledge of the process to know what would take the most time

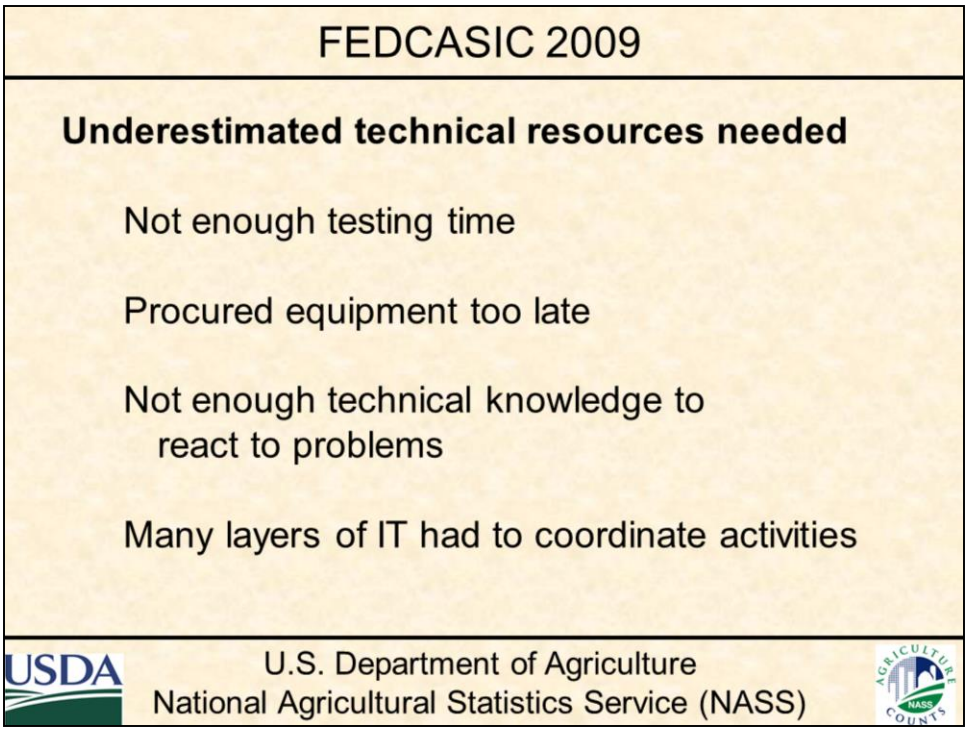
Tried to develop every process from scratch.

Introduced scanning, nearest neighbor imputation, disclosure, database processing, UNIX processing,

Did not have a way of prioritizing the system requirements



Staff were asked to do their regular job in addition to the Census work
Operational activities always take priority to new stuff that is years away.
Project plans and milestones were created
Almost all milestones were missed
Teams became frustrated
Outside consultants were called in
One person was chosen to be the czar.



Did not have much if any testing time

New UNIX machine arrived in November 2002. Became operational Feb 14, 2003.

No time to configure optimally before it was needed for production.

Learned on the fly.

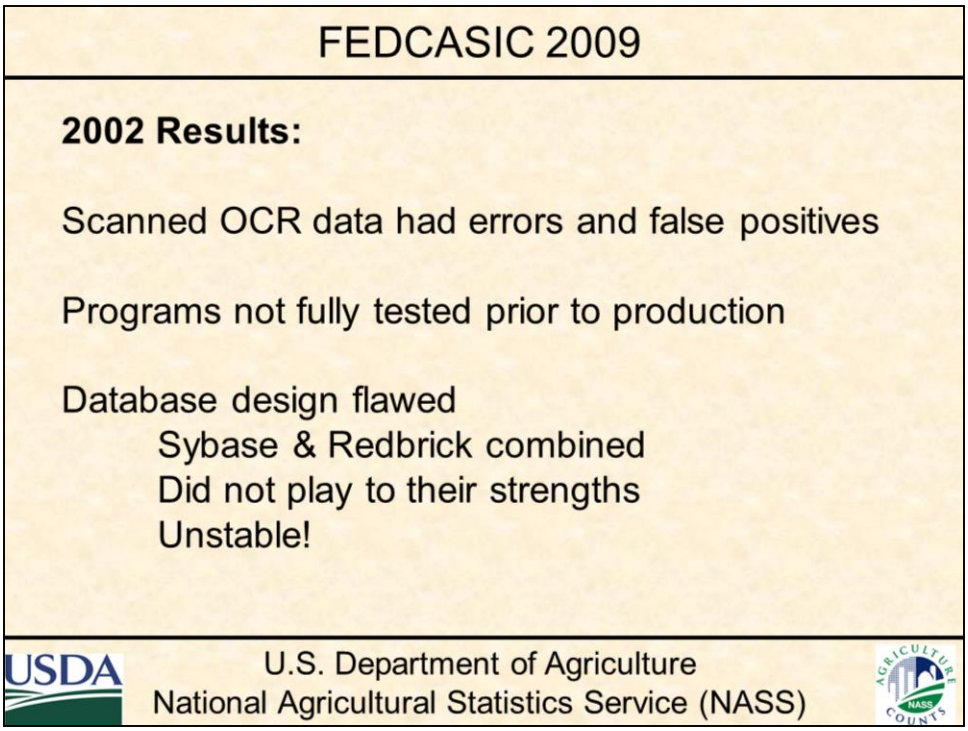
Had the wrong DASD configuration

Did not have the proper tools and knowledge to monitor the box.

Difficulty working across infrastructure teams to identify possible problems.

No stress testing was performed at all.

Had to coordinate with UNIX, Communications, Database, Security, LAN, Developers, Business Users



Respondents would draw line through the page indicating no response. However, the scanners would inturpret as a 1, 7 , 11 based on where the line hit the hot zones

Reran large amounts of data because of data integrity problems. Would move from one process to another and find out flags or values were not set properly. Would need to reprocess.

Had the 2 databases in the same partition. Did not play well together.

We would get POSIX locks. Forced us to reboot the databases once a week. The increased locks caused the databases to get slower and slower.

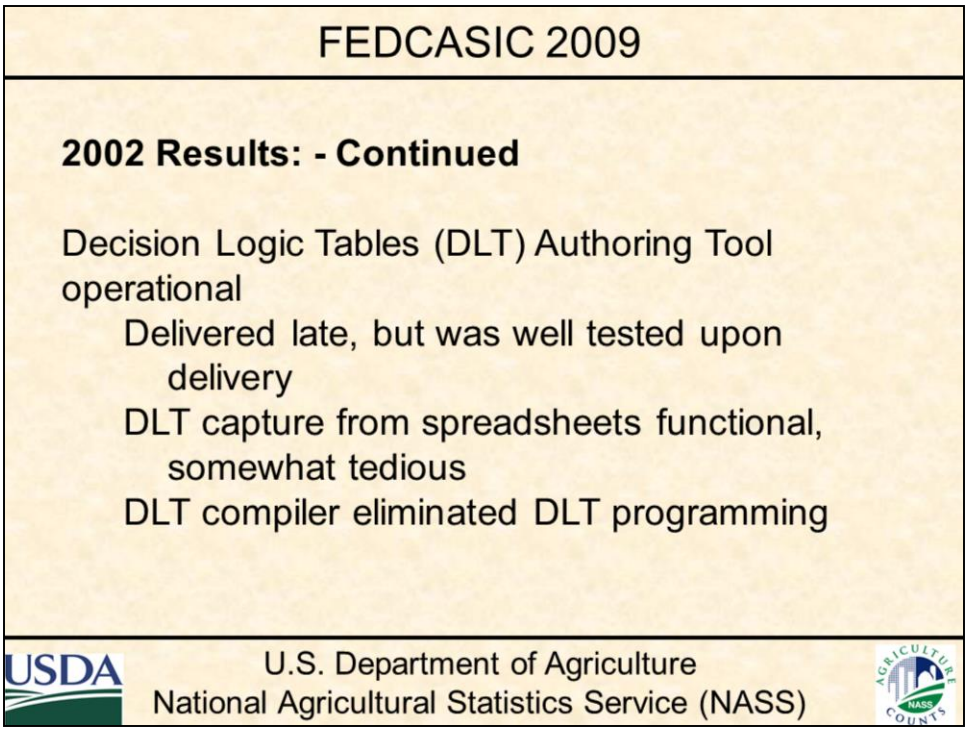
Didn't know that we needed to reorganize the databases on a regular basis when large amounts of data are being loaded.

Did not utilize the server side processing. Ran most things from the client.

Didn't allocate UNIX machine resources effectively.

Databases were unavailable about 40% of the time.

Batch processes would run into the daytime and slow down the day time processes.



DLT tool was 6 months late. This short changed the DLT authors from being able to effectively test their logic.

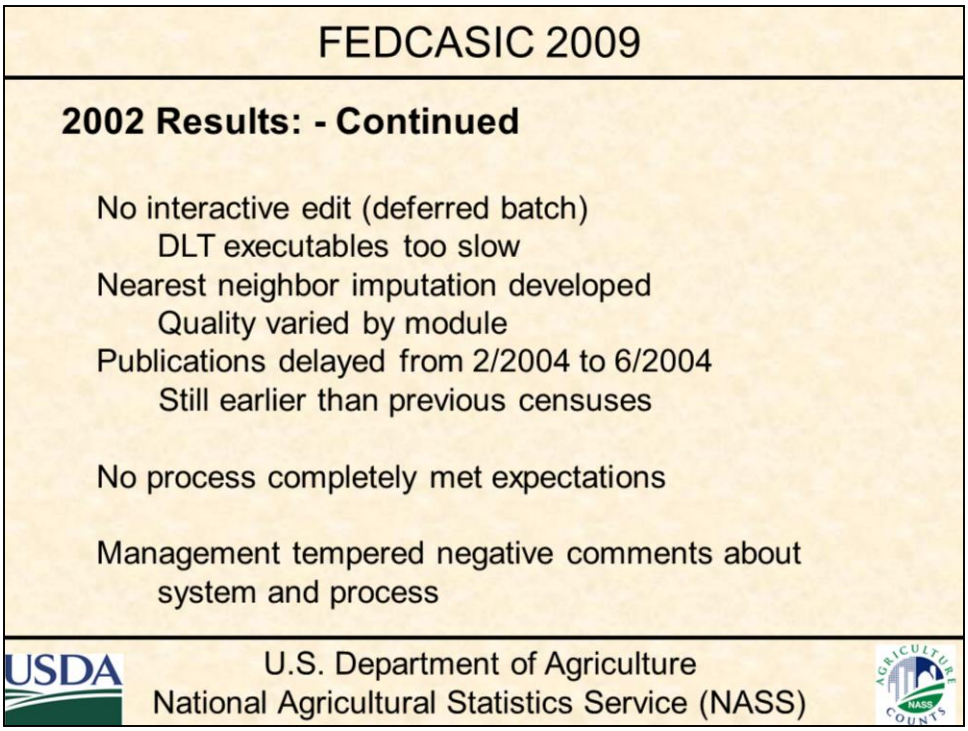
Transferring the DLT rules from the spreadsheets was very time consuming.

However once the data were in the database, these programs executed correctly.

DLT performance was not optimal

Data Review would sometimes have the “white screen of death”

Data was lost. Database tables did not stay in synch



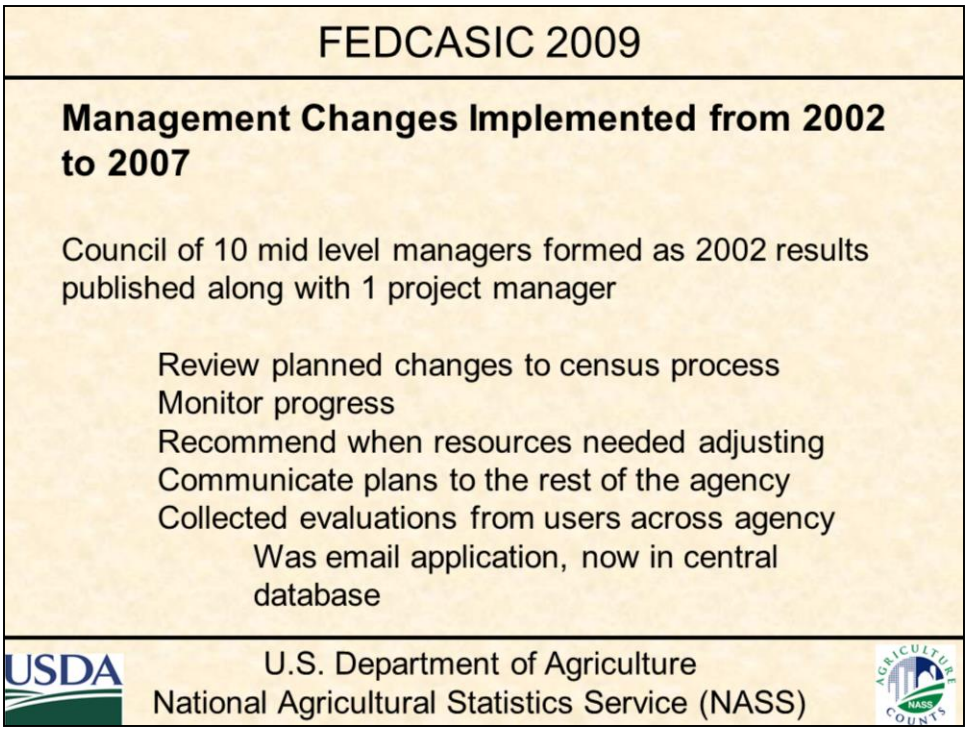
All records were edited in batch mode. Very little control on prioritizing edited or non edited records. Corrected records could take up to three days to be re-edited. It could take several editing sessions to get a record clean. Very frustrating for the end users.

Used a new nearest neighbor imputation strategy. This worked okay in the production and inventory sections of the questionnaire.

Didn't perform very well in the economic and demographic areas.

Every process experienced problems of one kind or another

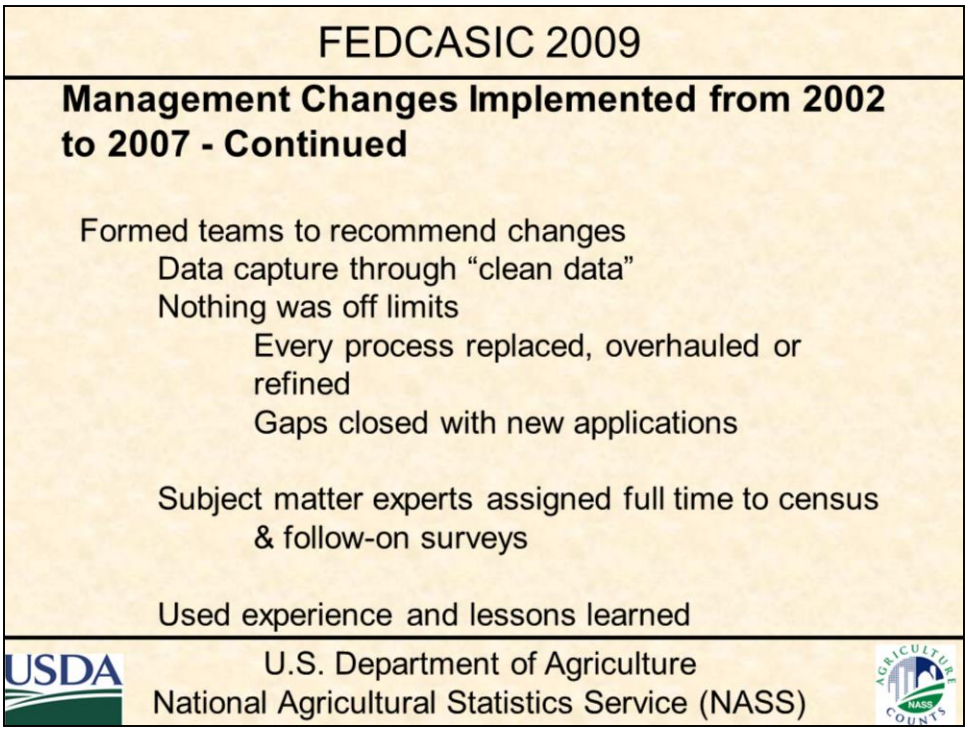
Delayed the release of the Volume 1 publication from February to June. Still better than the Bureau of Census had done in the past.



The Administrator felt the process needed more attention so 10 Branch Chiefs were put in charge.

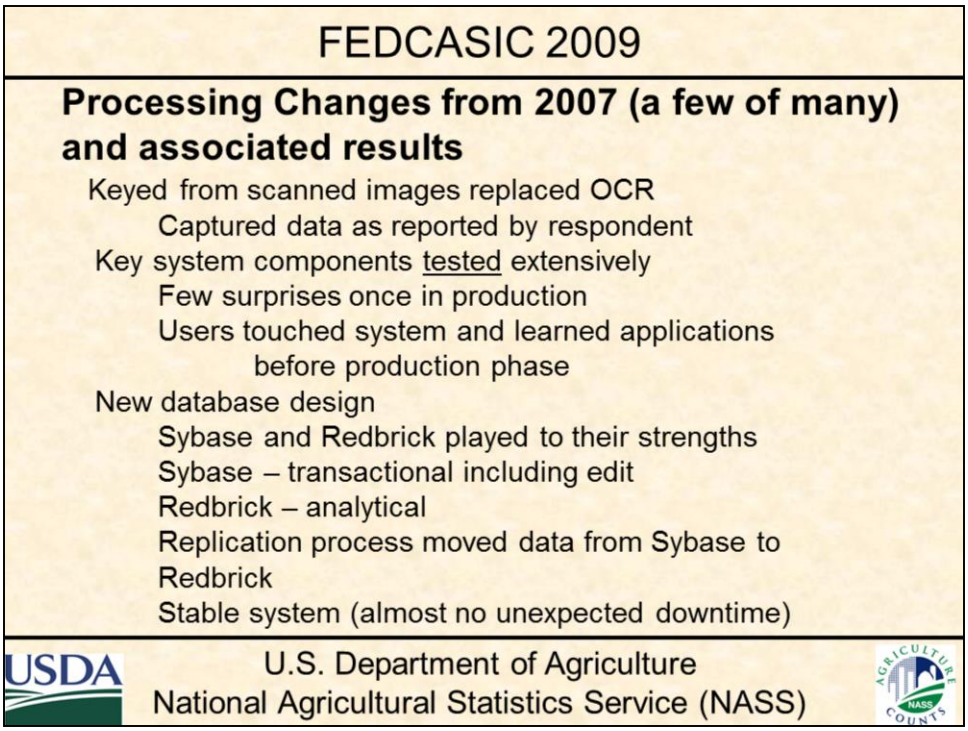
NASS had a much better, clearer idea of what it would take to run the program now.

Basically, they made sure resources were available and had teams working as early as possible.



The NASS staff that survived the 2002 program stayed with it for 2007. The experience provided much better understanding of what needed to be enhanced.

Time lines and milestones were much more realistic and able to be met.



NPC has a system that only presented data to be keyed in marked cells and the area around that cell.

So there was no paper handling and the keyer is able to interpret the whole area while keying.

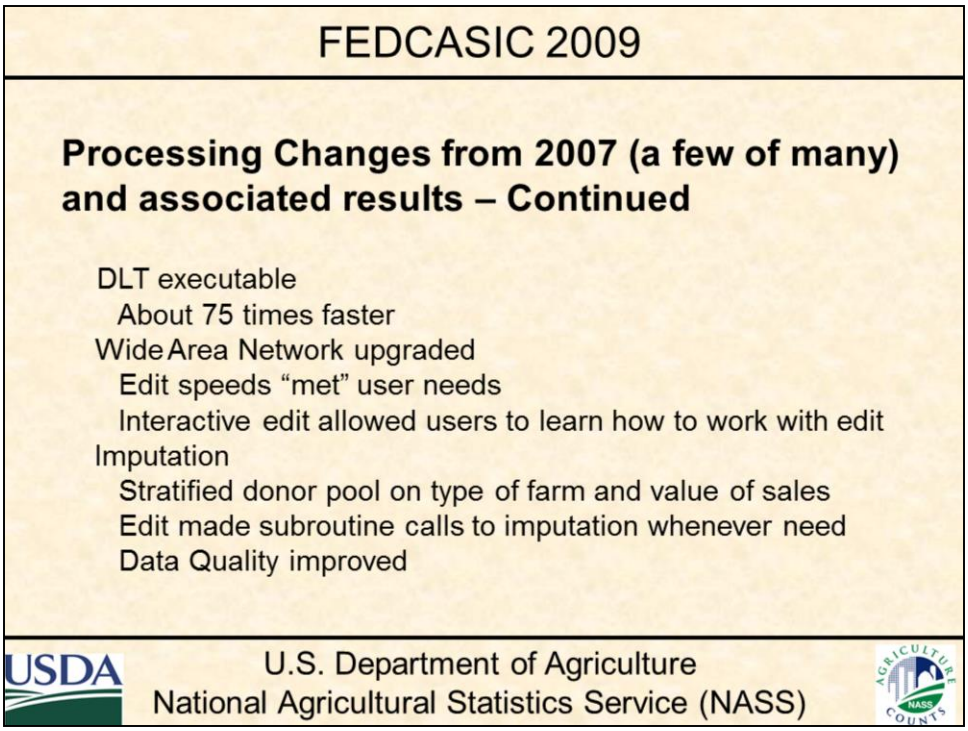
Had national training sessions with predefined test cases. Database was resettalbe.

Put each database in its own partition on the UNIX box.

Limited user to doing adhoc queries to the Redbrick database.

Interactive work and batch work ran concurrently.

Only had 3 unexpected down times in 12 months.



In 2002, we learned that we didn't always get the best candidate record. The selection process needed to be refined so that we would find the distance of similar operations of similar size. We created a 2 level selection process. To accomplish this.

Processing Changes from 2007 (a few of many) and associated results – Continued

- Upgraded UNIX box and storage devices
- Tested, tested, and tested some more
  - Every process improved
- Service Level Agreement
  - Set expected number of records to handle daily
  - Database response and reliability
  - Benchmark processes
- Quality control measures
  - Evaluate edit/imputation and data relationships
- EDR
  - Sizing and integration
- Macro aggregation – twice daily
  - Speed and timing

FEDCASIC 2009

2007 Results:

Published report on time in February 2009

For 2012 only need tuning

Ripple Impact of Census Effort

- Other systems delayed or postponed
- Census was first enterprise level editing/analysis system
- Survey application run on FO LANs
- Others on decades old mainframe systems

U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)

2007 Results – Continued

Push to move other distributed applications to enterprise level

- Survey Management (FoxPro)
- CATI (Blaise)
- Departmental initiative to secure servers in central locations
- Support for distributed application very difficult

2007 Results: – Continued

- Team formed to decide data flow at new enterprise level
  - All survey processes involved
    - Frame maintenance
    - Sample selection
    - Instrument designers
    - Survey management (FO)
    - Data collection
    - Editing
    - Survey administration (HQ)
    - Analysis and summary
  - Output will be new business rules
    - Will be fed into database/application design process

Future Direction of "Census" PRISM Processing System

- Design considering survey needs
- Tuning applications to meet survey needs
- Team of end users to recommend specs
- First census follow-on survey being developed

Lessons Learned and Best Practices

Don't underestimate amount of communication required
Within teams & across various units

Manage requirements and scope creep

Leverage your experience & lessons learned

Have well defined milestones

Systems take longer to develop than planned

Lessons Learned and Best Practices

Incorporate quality control tools into the system

You can't do enough testing

Establish service level agreements

Define and manage expectations