Case Study: Census 2000 Analysis & Dissemination

Seth Grimes
Alta Plana Corporation

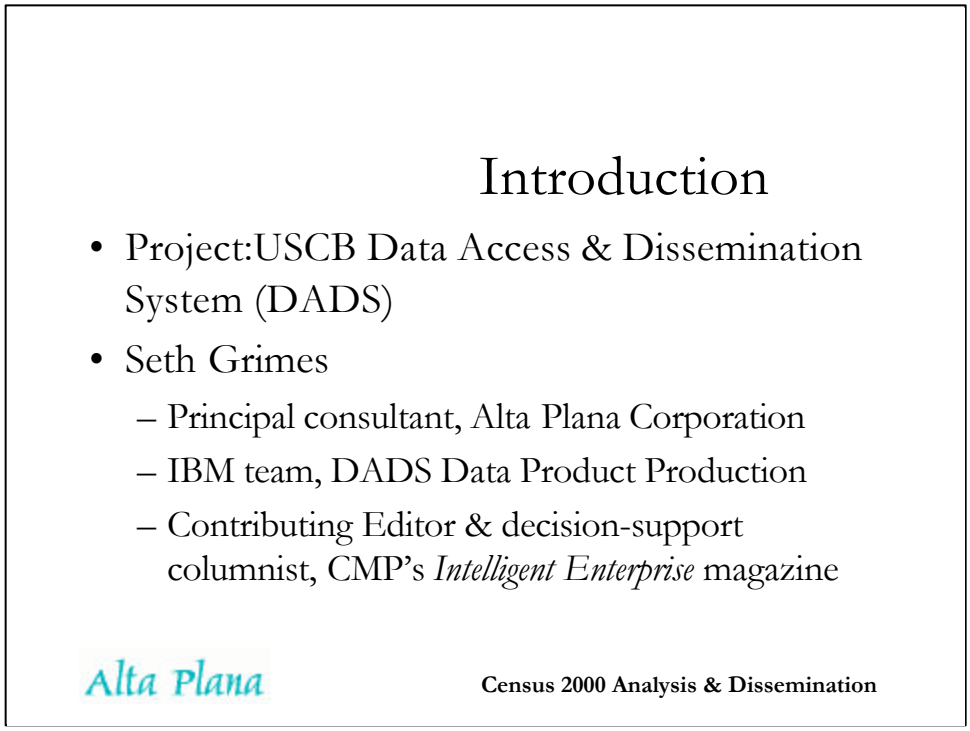
DAMA-NCR
January 8, 2002

Alta Plana

Census 2000 Analysis & Dissemination

Today's presentation is a case study.

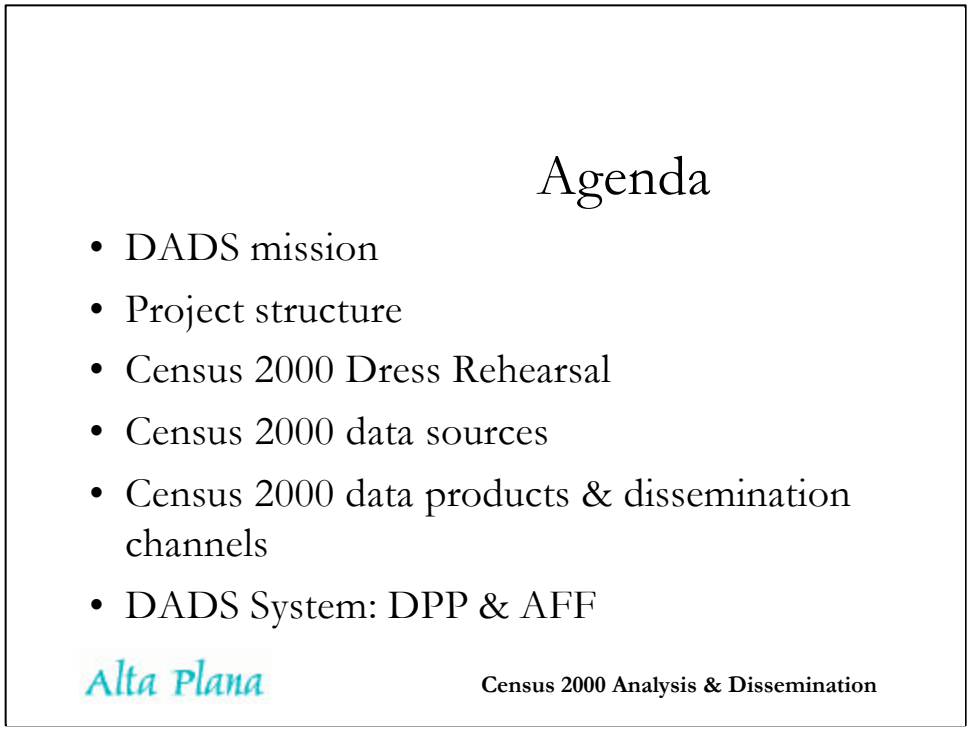
Rather than focusing on a particular technology or application area, it will describe a particular project. It will show how a technology solution was created to meet the project's requirements and will generalize a bit, suggesting areas where that solution can be applied to other, similar problems.



The project that is the subject of this case study is the US Census Bureau's Data Access & Dissemination System, known by the acronym DADS, which is comprised of two, complementary subprojects. The Data Product Production system creates analytical data products for congressional redistricting and to support federal funding mandates. And American FactFinder is the Web dissemination channel for DPP-generated data, as well as for economic census and American Community Survey data.

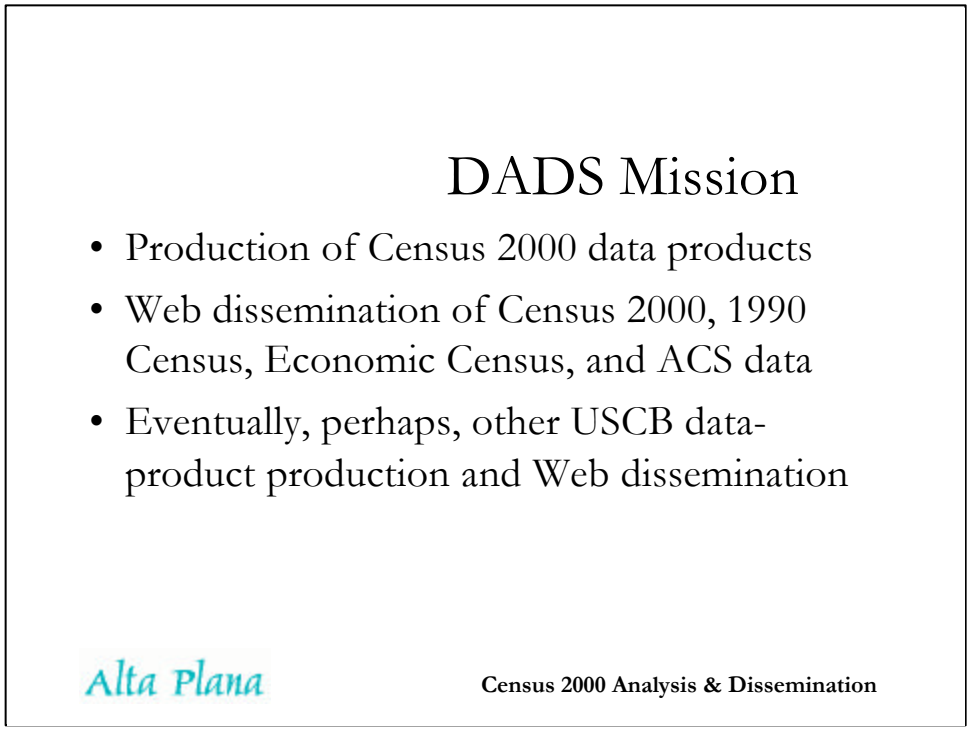
I have served as team leader for the DPP side of the project for most of the last four years, working on subcontract to IBM Global Services. This job was a natural for me: I've spent all but two years of my career working with scientific and statistical data for governmental organizations -- NASA, the US Dept. of Transportation, the OECD [Organisation for Economic Co-operation and Development],and the IMF [International Monetary Fund] -- and the balance of my work has been developing Internet services including dynamic, database-backed Websites.

I do this through a consulting company I set up in 1997, Alta Plana, and I do some writing on the side, principally for CMP's Intelligent Enterprise magazine, where I'm a contributing editor and write a decision support column.



The DADS project's mission is to produce and disseminate Census data with the highest regard for correctness, security, and timeliness. I'll explain more fully and then discuss how the government has achieved that mission.

I'm going to speak more extensively on DPP, first because it's my primary project involvement and secondly because AFF is public -- you can look at it yourself -- and the technology is much more conventional.

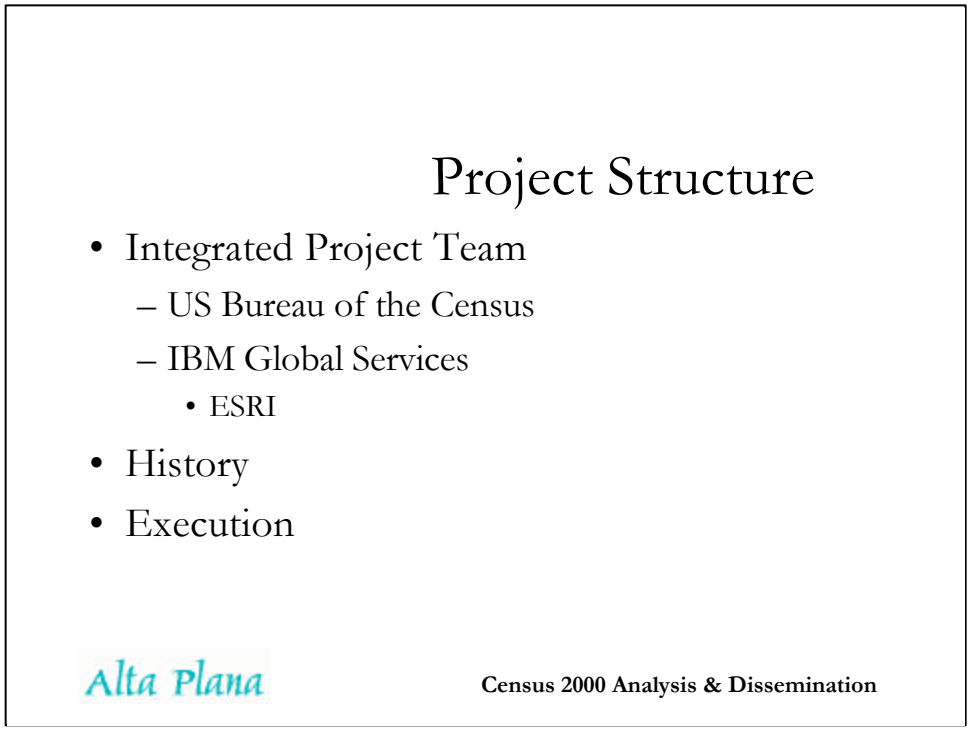


Census 2000 data products each consist of up to one thousand statistical tables with basic aggregations -- counts, sums, averages, medians, quartiles -- of a variety of demographic characteristics over a number of universes, usually persons, households, householders, and families. Tables are computed for state and national "summary areas."

Every number in every data-product table -- and every survey question and allowed response -- is designed to meet a federal government program need for data to support congressional redistricting or government programs. Forms and data products were designed with the participation of the Office of Management and Budget, the Department of Justice, and other agencies and with Congressional approval.

In addition to 1990 & 2000 population & housing census summary data, AFF also disseminates results from the every-5-year census of businesses and from the American Community Survey, a "continuous measurement instrument" that will likely replace the decennial-census long-form survey.

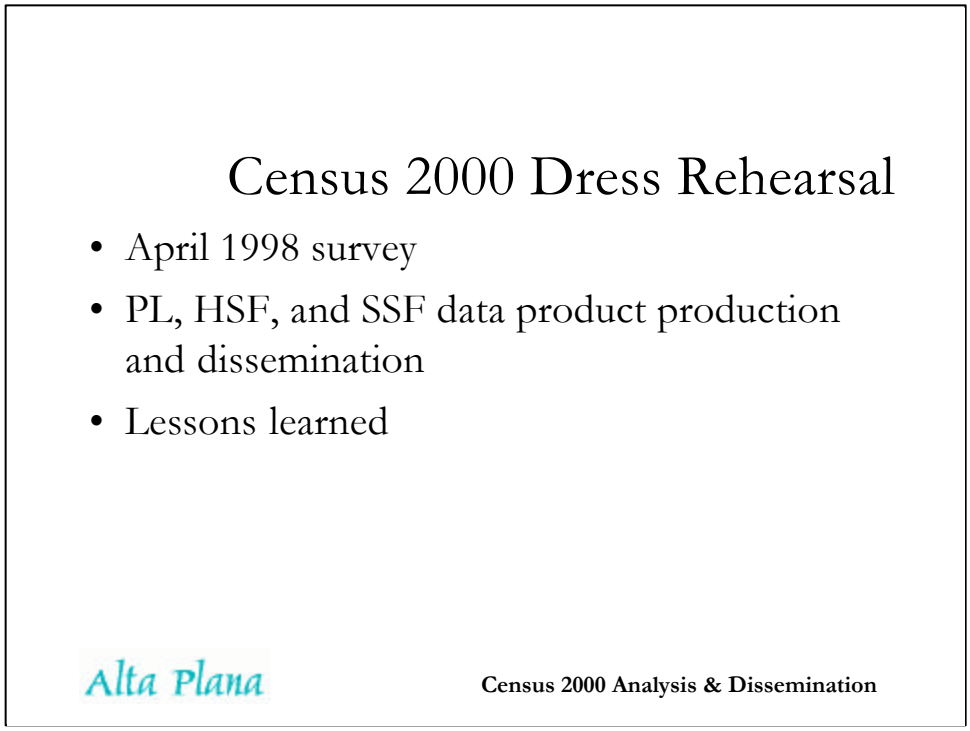
Lastly, the DADS subsystems are largely metadata driven and may therefore be adapted to analyze & disseminate other demographic data than the data for which the subsystems were originally designed.



The government created two DADS prototypes in-house in the '95-'97 time frame. They decided to seek contract helpafter evaluating the 1990-round and DADS prototype experiences and in-house resources and 2000-round schedules. Previous population censuses had been analyzed by in-house staff.

IBM Global Services came on board on October 1, 1997 and, in turn, engaged expert consultants for key technical roles. I, myself, joined the project the following April.

The project team has government and contractor employees working side-by-side in an "integrated project team."



The government conducted a dress rehearsal survey in three locations -- Sacramento, California; Columbia, South Carolina; and Menominee, Wisconsin, an American Indian Tribal area -- with subsequent processing, analysis, and dissemination that anticipated, on a limited scale, the challenges that would be faced in the full census.

The DR data products -- PL, short for Public Law 94-171, the congressional redistricting data; HSF, for Hundred Percent Summary File; and SSF, for Sample Summary File -- were close in structure, content, and coverage to the products that have since been defined for the full 2000 census.

Results were disseminated on American FactFinder and on CD, again as valuable prototypes of the 2000 efforts to come.

We learned many lessons from the DR exercise. I'll list some of them now and intersperse lessons learned from the 2000-census analysis in the rest of my talk:

- With a large volume of repetitive steps, it is essential to automate production & verification steps to the greatest extent possible, breaking only when essential, for external review, for instance. This means doing all production on a uniform computing platform.

- You can't do enough manual & automated verifications, of input & of output.

- Not having software-processable -- database formatted -- metadata, slows down processing and introduces risk of errors. (We got metadata in spreadsheets & formatted documents.)

DR lessons learned, cont.:

- Timely inputs -- specifications and trial data -- are essential, but you can allow for delays by planning in advance to produce your own simulated inputs and by parameterizing processing to the greatest extent possible to avoid coding business logic that may change.

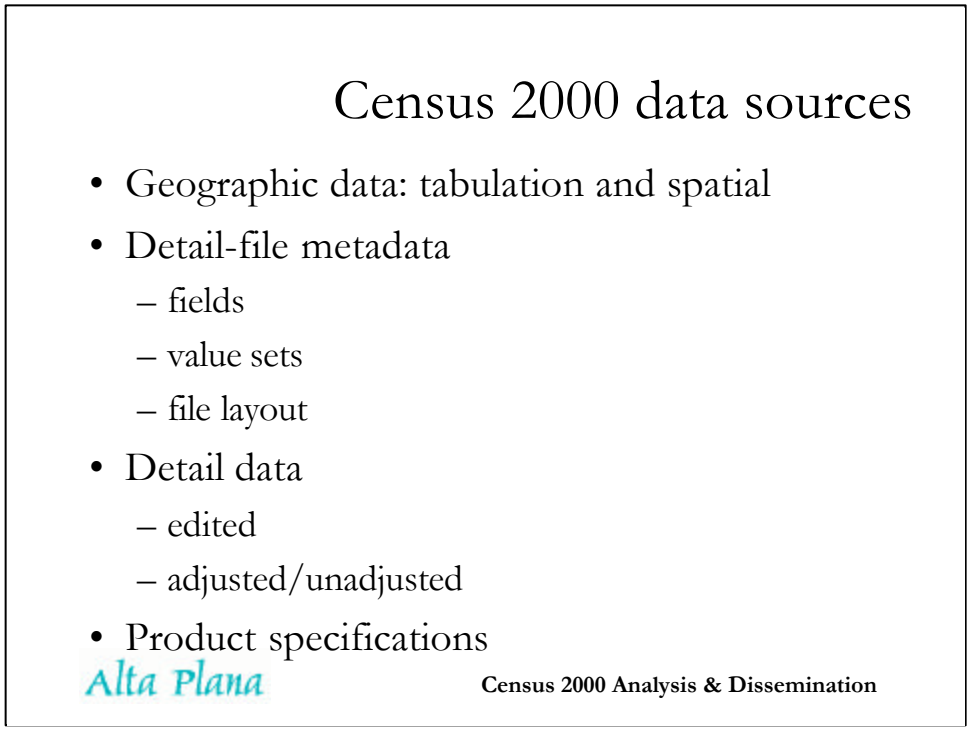
- There are three keys to optimal performance and throughput:

. Adequate hardware,

. Optimization of COTS software, and

. Good system design.

- Technical competence, for instance with programming tools, is not a substitute for subject-matter expertise and experience dealing with large-scale processing requirements.

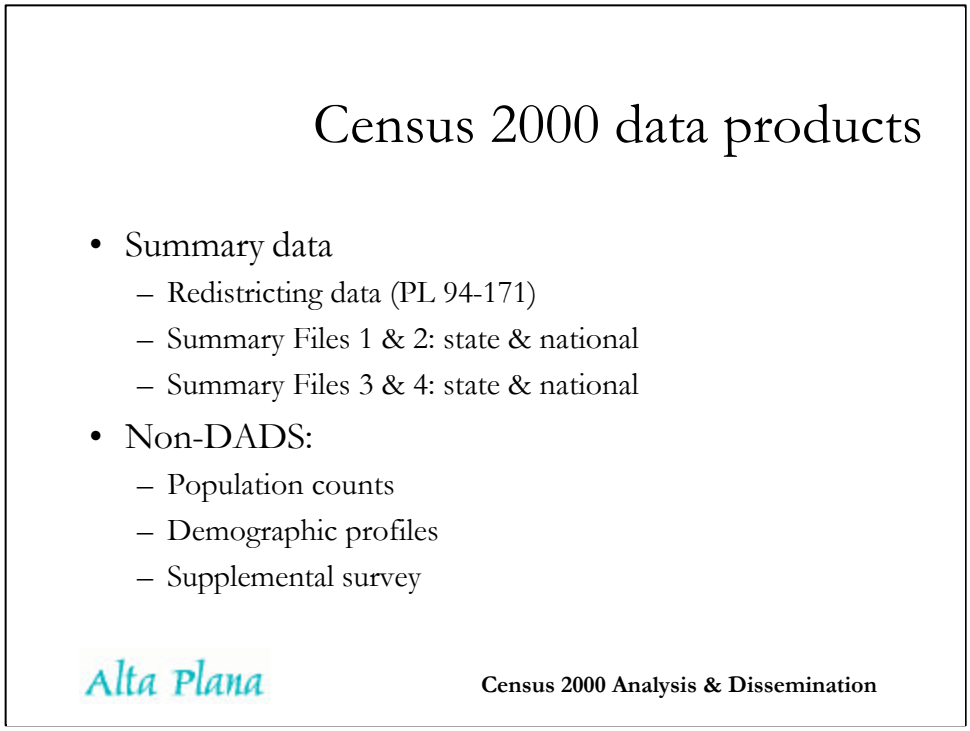


The census survey was conducted using a comprehensive Master Address File (MAF) and the Topologically Integrated Geographic Encoding and Reference (TIGER) system. For TIGER, the Census Bureau (USCB) further defines tabulation (as opposed to collection) geographies and produces spatial data suitable for mapping applications.

USCB calls respondent-provided information "detail data." Other synonyms are microdata and unit-record data. The data comes in two forms, 100% and Sample. Survey-form responses go through a painstaking editing process that cleanses data, corrects certain classes of errors, and ensures confidentiality. Data were also statistically adjusted to correct gross survey errors, namely undercounting and double counting, but USCB decided not to publish summary statistics based on adjusted data.

In the 2000-census round, as in DR, we received metadata in spreadsheets and formatted documents, which forced us to do extensive manual preparation.

Product specifications include the values to be computed, the geographic coverage, and statistical algorithm details.



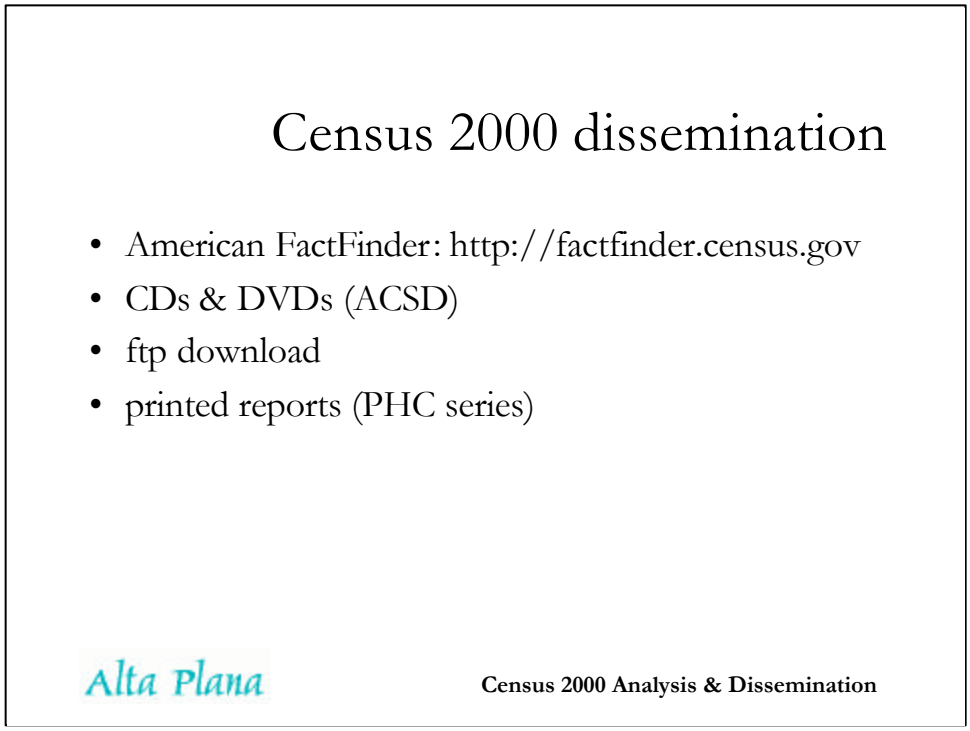
The Congressional Redistricting data product draws from the 100% survey questions about race, age, and Hispanic or Latino origin.

Summary Files 1 and 2 more extensively report statistics based on the 100% questions, SF2 with an innovation: analysis over 250 Characteristic Iterations of race, ancestry, and ethnicity.

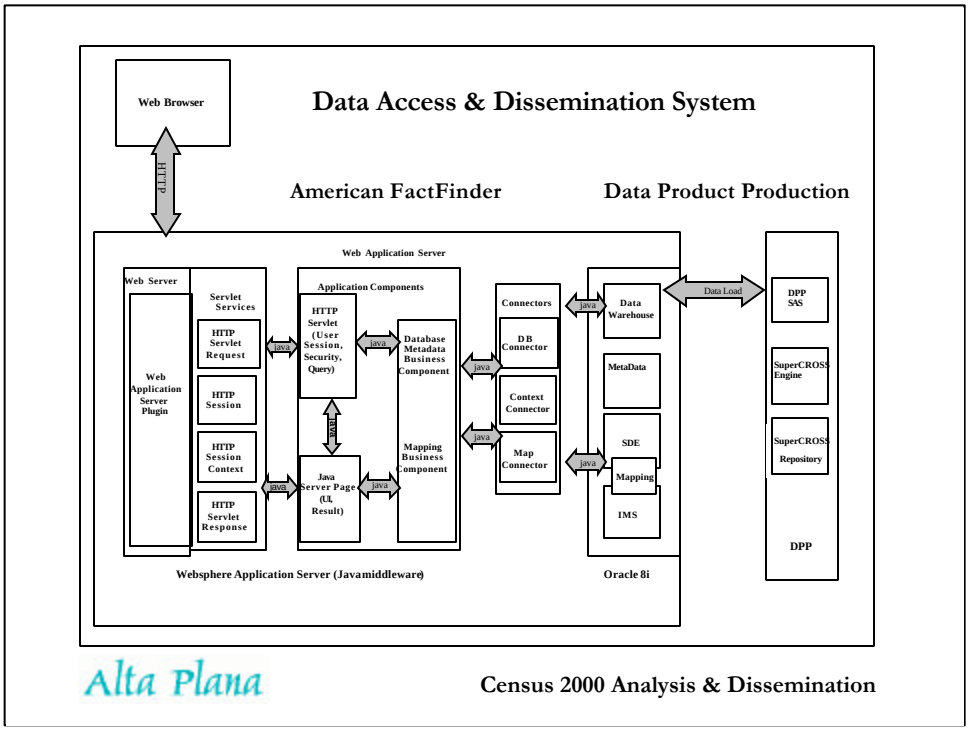
Summary Files 3 and 4 are based on the 1-in-6 Sample survey. SF3 has about 16,000 data values in 813 tables computed for about a million geographic areas. SF4 will feature 336 CIs but "only" 300 or so tables and half as many geographic areas.

Subsequent products will cover, for instance, the 108th Congressional Districts and School Districts.

The large volume of data to be computed led us to clever, efficient strategies, for example, computing in advance values on which "thresholds" would be based and using those values to avoid tabulating for geographic areas whose results we knew would be suppressed because the values did not meet the thresholds.



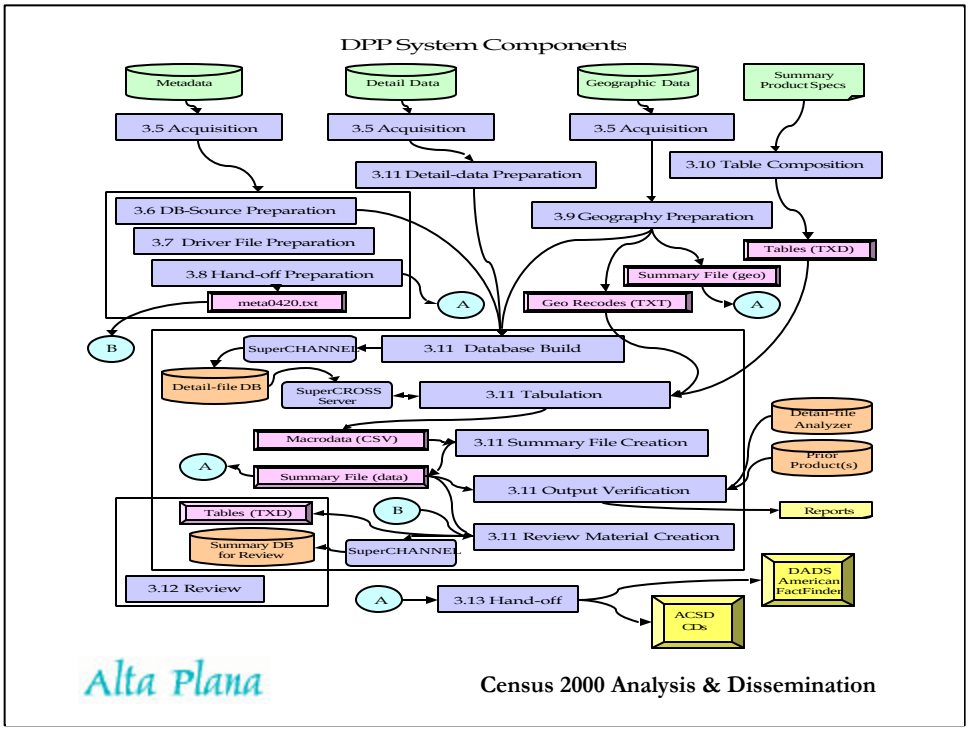
Internet dissemination of population census results is new and takes two principal forms: AFF and ftp download. AFF is most suitable for casual, ad hoc use, while those with more intensive and extensive needs will want to download datasets for analysis.



I've included this DADS architecture slide, which actually isn't completely up to date, to show the principal software components and their relationships.

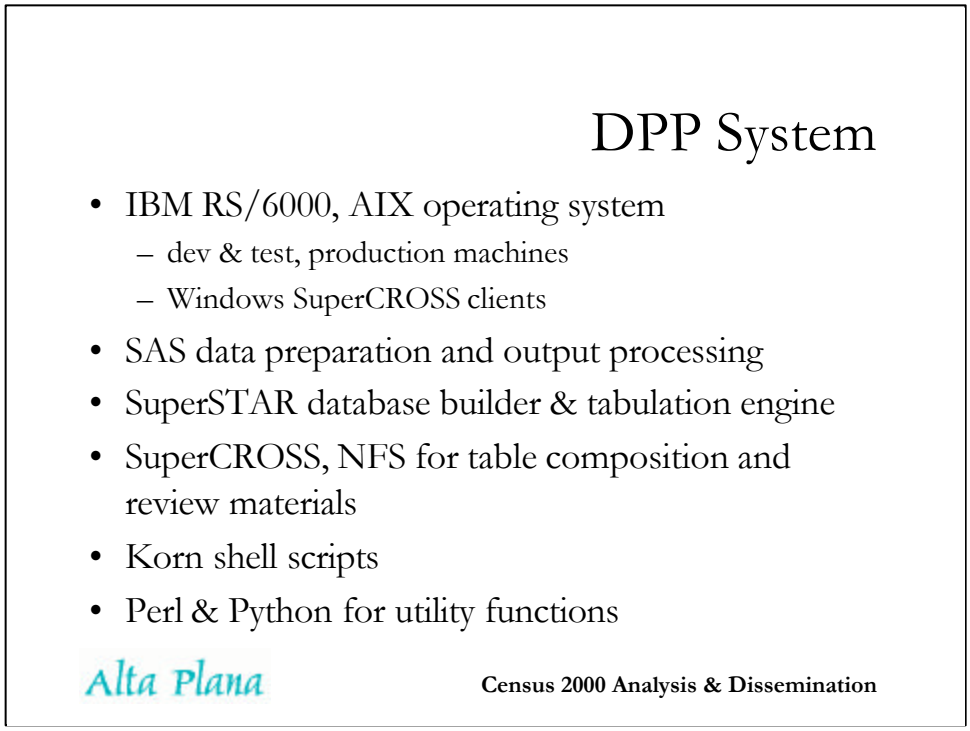
Initial design was done in accordance with IBM's WSSDM methodology. Staffing early on include business & technical architects who, as the outcome of "joint application development" (JAD) sessions, created initial functional and "non-functional" requirements documentation.

Another lesson learned: The functional requirements were expressed as UML-style "use cases," mainly because the use-case approach was chose for AFF. In fact, it is very suitable for object-oriented, event-driven systems like AFF but unlike DPP, which is highly procedural and involves few "actors" and "scenarios."



And this DPP -- Data Product Production -- slide shows key flows through the DPP subsystem. Steps in the central box, which run from database build through presentation of review materials, are automated in a single script called Tab. Two other scripts, for geographic data processing and for results hand-off, with Tab automate the vast majority of the production processes.

Ignore the numbers in the diagram: they correspond to sections in the document from which the diagram was taken.

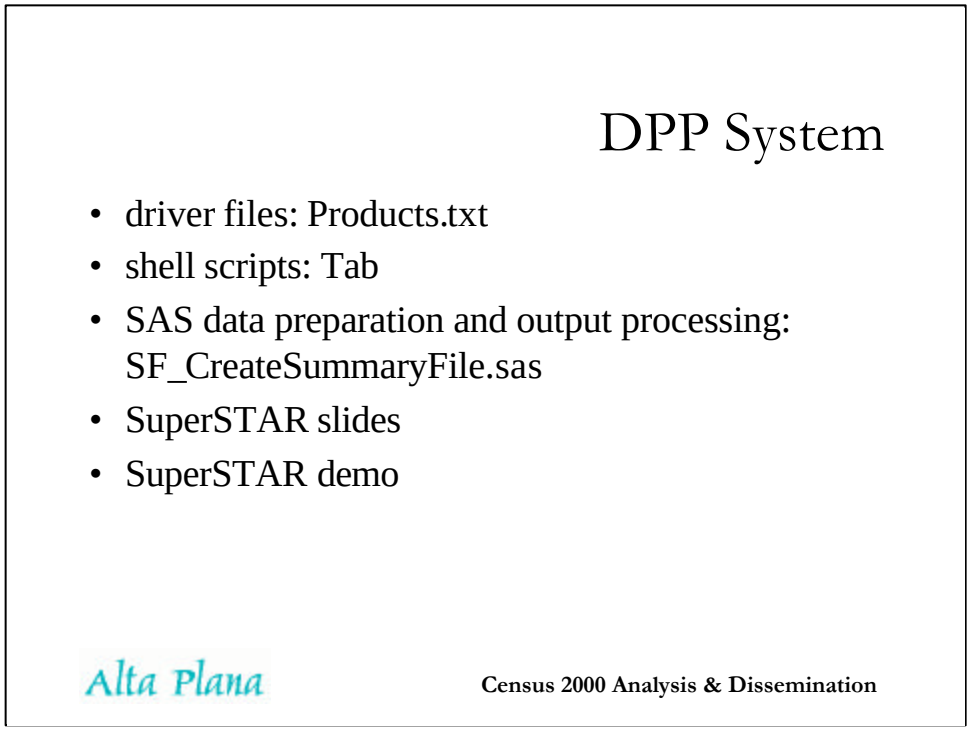


The DPP computing platform was decided early in the project and took shape over about a year in the first of seven major DPP system releases. The first major decision was to use the SuperSTAR analytical suite from Space-Time Research (STR) of Melbourne, Australia, for the core tabulation component. IBM evaluated the suite's SuperCROSS GUI and tabulation engine for performance and usability against SAS and Oracle, for both of which USCB has a site license.

SuperSTAR is well suited for analysis of demographic statistics and is used by a number of governmental statistical agencies worldwide. (And it's time for me to restate that my company has a business relationship with STR as a reseller and consultant, an outgrowth of our DADS SuperSTAR use.)

We chose SAS for major system components and the whole system is automated with shell scripts. If we were starting from scratch -- or "refactoring" -- we'd probably choose a more robust scripting tool.

The production machine through SF2 processing has been a 24-processor IBM RS/6000 S85 with 96 GB memory and a 5.6 TB disk array. For subsequent processing, additional machines and storage are being added to the mix.



A key design point is the DPP system's use of what we call "driver files": sets of parameters that describe data sources, data products, transformation steps, and so on. The driver files and use of environment variables allow us to avoid hard-coding business logic in scripts and programs.

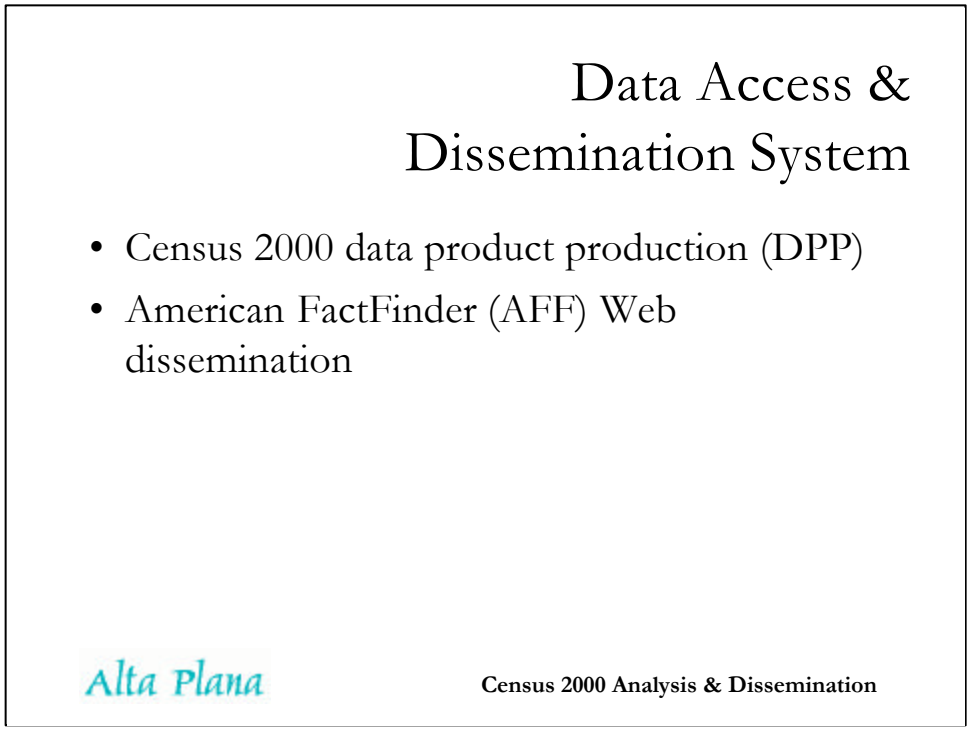
To get a taste of how the system is implemented, we'll look at one of the system's key driver files and then look quickly at a piece of shell script and SAS program.

Then I'll tell you more about the SuperSTAR suite and demonstrate the GUI.

(Readers of the speaker's notes: please contact me for further information on the code and SuperSTAR software.)

American FactFinder System

- IBM RS/6000 SP, AIX operating system
  - internal, staging, public
- Oracle DW
  - Oracle parallel server
  - ISO compliant metadata schema
  - Perl ETL code
- IBM WebSphere
  - Java servlets
- ESRI spatial data engine



This concludes my presentation. You'll find a copy at altaplana.com/pp, and you might also check out some of the references listed on the next slide.

References

"Making Sense of the Census," govtech.net/magazine/story.phtml?id=3030000000002819

Census 2000 product overview, census.gov/dmd/www/products.html

American FactFinder, factfinder.census.gov

"Automating a Data Driven Dissemination System on the Internet," Sandra Rowland

Census 2000 summary data download, ftp2.census.gov/census_2000/datasets

Space-Time Research, http://www.str.com.au

Lastly, here's contact information if you think up any questions later or want to find out more about Alta Plana or have any interesting business or collaboration ideas.

Case Study: Census 2000 Analysis & Dissemination

Seth Grimes
Alta Plana Corporation

DAMA-NCR
January 8, 2002

Alta Plana

Census 2000 Analysis & Dissemination