

# Case Study: Census 2000 Analysis & Dissemination

Seth Grimes  
Alta Plana Corporation

Capital Area Business Intelligence Network  
March 7, 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.

## Introduction

- Project:USCB Data Access & Dissemination System (DADS)
- Seth Grimes
  - Principal consultant, Alta Plana Corporation
  - IBM team, DADS Data Product Production
  - Contributing Editor & decision-support columnist, CMP's *Intelligent Enterprise* magazine

Alta Plana

Census 2000 Analysis & Dissemination

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-back 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.

## Agenda

- DADS mission
- Project structure
- Census 2000 Dress Rehearsal
- Census 2000 data sources
- Census 2000 data products & dissemination channels

*Alta Plana*

**Census 2000 Analysis & Dissemination**

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.

## DADS Mission

- Production of Census 2000 data products
- Web dissemination of Census 2000, 1990 Census, Economic Census, and ACS data
- Eventually, perhaps, other USCB data-product production and Web dissemination

*Alta Plana*

**Census 2000 Analysis & Dissemination**

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.

## Project Structure

- Integrated Project Team
  - US Bureau of the Census
  - IBM Global Services
    - ESRI
- History
- Execution

*Alta Plana*

**Census 2000 Analysis & Dissemination**

The government created two DADS prototypes in-house in the '95-'97 time frame. They decided to seek contract help after 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.”

## Census 2000 Dress Rehearsal

- April 1998 survey
- PL, HSF, and SSF data product production and dissemination
- Lessons learned

*Alta Plana*

**Census 2000 Analysis & Dissemination**

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.

## Census 2000 data sources

- Geographic data: tabulation and spatial
- Detail-file metadata
  - fields
  - value sets
  - file layout
- Detail data
  - edited
  - adjusted/unadjusted
- Product specifications

*Alta Plana*

Census 2000 Analysis & Dissemination

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

USBC calls respondent-provided information "detail data." Other synonyms are microdata and unit-record data. 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 USBC decided not to publish summary statistics based on adjusted data.

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

## Census 2000 data products

- Summary data
  - Redistricting data (PL 94-171)
  - Summary Files 1 & 2: state & national
  - Summary Files 3 & 4: state & national
- Non-DADS:
  - Population counts
  - Demographic profiles
  - Supplemental survey

*Alta Plana*

**Census 2000 Analysis & Dissemination**

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. SF4 will feature 335 CIs.

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

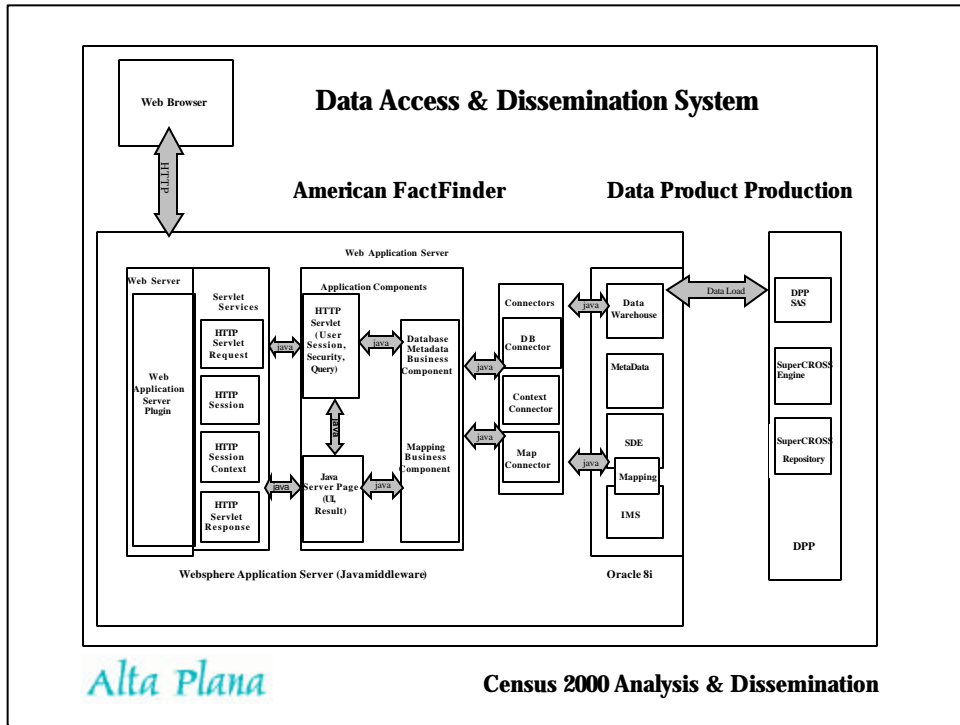
## Census 2000 dissemination

- American FactFinder: <http://factfinder.census.gov>
- CDs & DVDs (ACSD)
- ftp download
- printed reports (PHC series)

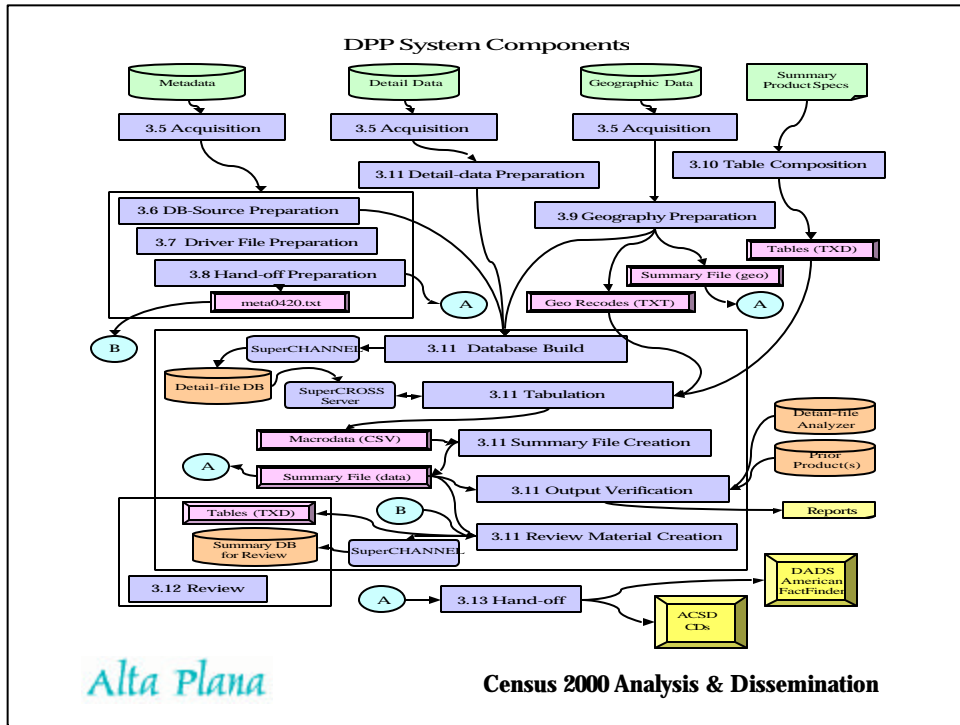
*Alta Plana*

**Census 2000 Analysis & Dissemination**

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.



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.

## DPP System

- IBM RS/6000, AIX operating system
  - dev & test, production machines
  - Windows SuperCROSS clients
- SAS data preparation and output processing
- SuperSTAR database builder & tabulation engine
- SuperCROSS, NFS for table composition and review materials
- Korn shell scripts
- Perl & Python for utility functions

*Alta Plana*

**Census 2000 Analysis & Dissemination**

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 USBC 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.

## DPP System

- driver files: Products.txt
- shell scripts: Tab
- SAS data preparation and output processing:  
SF\_CreateSummaryFile.sas
- SuperSTAR slides
- SuperSTAR demo

*Alta Plana*

**Census 2000 Analysis & Dissemination**

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.

## 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

*Alta Plana*

**Census 2000 Analysis & Dissemination**

## Data Access & Dissemination System

- Census 2000 data product production (DPP)
- American FactFinder (AFF) Web dissemination

*Alta Plana*

**Census 2000 Analysis & Dissemination**

This concludes my presentation. You'll find a copy at [altaplana.com/pp](http://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](http://govtech.net/magazine/story.phtml?id=3030000000002819)
- Census 2000 product overview,  
[census.gov/dmd/www/products.html](http://census.gov/dmd/www/products.html)
- American FactFinder, [factfinder.census.gov](http://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](http://ftp2.census.gov/census_2000/datasets)
- Space-Time Research, <http://www.str.com.au>

*Alta Plana*

**Census 2000 Analysis & Dissemination**

## Contact

Seth Grimes

Alta Plana, <http://altaplana.com>

1-301-873-8225

[grimes@altaplana.com](mailto:grimes@altaplana.com)

*Alta Plana*

**Census 2000 Analysis & Dissemination**

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

Capital Area Business Intelligence Network  
March 7, 2002

*Alta Plana*

**Census 2000 Analysis & Dissemination**