2010 DEMONSTRATION DATA PRODUCTS DISCLOSURE AVOIDANCE SYSTEM DESIGN SPECIFICATION

Version 1.4

November 12, 2019

This is version 1.4 of the 2010 Demonstration Data Products Disclosure Avoidance System Design Specification. Printed copies of this document may not contain the most recent updates.

This document branches off the *2018 END-TO-END TEST DISCLOSURE AVOIDANCE SYSTEM* version 1.3 developed for the 2018 End-to-End Test.

United States™ Census Bureau

Document Revision History

Version	Publication Date	Revision Description	Section	Author(s)
1.0	10/9/2018	Initial Version	All Sections	Simson Garfinkel, Joseph Cortez
1.1	10/26/2018	Updates	All Sections	Chris Rivers
1.2	03/05/2019	Updates	All Sections	Claudia Molinar
1.2.1	03/11/2019	Revised for E2E	All Sections	Simson Garfinkel Claudia Molinar Knexus Research Corp.
1.2.2	03/20/2019	Updates	All Sections	Knexus Research Corp.
1.2.3	03/21/2019	Minor Edits	All Sections	Simson Garfinkel
1.2.4	03/22/2019	Addition of Appendix	All Sections	Knexus Research Corp.
1.2.5	03/25/2019	Minor Edits	All Sections	John Maron Abowd
1.2.6	03/25/2019	Updates to Appendix	All Sections	Knexus Research Corp.
1.2.7	04/01/2019	Updates to Document	All Sections	Knexus Research Corp.
1.2.8	04/05/2019	Updates to Tables	All Sections	Knexus Research Corp.
1.2.9	04/15/2019	Update to Guide	All Sections	Knexus Research Corp.
1.3	07/09/2019	New Standalone Instr.	Section 10	Knexus Research Corp.
1.4	11/12/2019	Revised for 2010 Demonstration Products	All Sections including Section 10	Knexus Research Corp. Claudia Molinar Simson Garfinkel

Table of Contents

List of Tables

List of Figures

1. BACKGROUND

In 2020, the United States Census Bureau will conduct the 2020 Census, which aims to enumerate every person residing in the United States, covering all 50 states, the District of Columbia, and Puerto Rico. All persons alive on April 1, 2020 who reside in these places, according to residency criteria finalized in 2018, must be counted.

The Census Bureau must submit state population totals to the United States President by December 31, 2020. The United States Constitution mandates this decennial enumeration be used to determine each state's Congressional representation.

Public Law 94-171 directs the Census Bureau to provide data to the governors and legislative leadership in each of the 50 states for redistricting purposes. This product will be the first file released that will include demographic and housing characteristics about detailed geographic areas.

Decisions about the data that will be included in this file have already been set—this release contains no file design changes from the 2018 prototype version. The Redistricting Data File will be released by April 1st, 2021, within one year of Census Day.[1]

The goal of the 2010 Demonstration Data Products is to help finalize decisions regarding the remaining list of planned tables and associated geographies by April 1, 2020. This includes the products released using the 2020 Disclosure Avoidance System (most of which are included in the Demographic and Housing Demonstration File), as well as the household join and more detailed tables that will be protected using distinct differentially private code modules that are not yet part of the 2020 DAS.[2]

As part of the Census Bureau's collection activities, the Census Bureau by statute must assure that the decennial census data products meet the legal requirements of Title 13, Section 9(a)(2) of the U.S. Code, which means the published results of the census must not identify data from specific individuals; nor should data from specific individuals be reasonably inferable.

In previous decennial censuses, a variety of techniques were used to protect the confidentiality of responses, including the use of synthetic data and household swapping.[3] For the 2020 Decennial, the Census Bureau applied the latest science in developing the 2020 Disclosure Avoidance System (DAS). Following the instructions of the Data Stewardship Executive Policy Committee (DSEP), the Census Bureau implemented differential privacy (DP) as the primary methodology.

This public release of the 2010 Demonstration Products test DAS source code and the accompanying release of datasets created by applying the DAS to private data from the 2010 Decennial Census provides increased transparency of the Census Bureau's effort to adopt DP at the national scale.

[1] See *Status Update on 2020 Census Data Products Plan*, available at https://www2.census.gov/cac/sac/meetings/2019-09/status-update-2020-census-data-products-plan.pdf.

[2] See *Frequently Asked Questions for the 2010 Demonstration Data Products*, available at https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products/faqs.html.

[3] See *Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing*, Laura McKenna, October 2018, Center for Economic Studies Working Paper 18-47, US Census Bureau, available at *https://www2.census.gov/ces/wp/2018/CES-WP-18-47.pdf*.

2. OVERVIEW

Article 1 Section 2 of the U.S. Constitution directs the U.S. Government to conduct an "actual enumeration" of the population every ten years.

The Census Bureau is now engaged in conducting the 24thCensus of Population and Housing with reference date April 1, 2020 and producing public-use data products that conform to the requirements of Title 13 of the U.S. Code. The goal is to count everyone once, only once, and in the right place.[4] All residents must be counted. After the data have been collected by the Census Bureau, but before the data are tabulated to produce data products for dissemination, the confidential data will undergo *statistical disclosure limitation* so that the impact of statistical data releases on respondent privacy can be quantified and controlled.

In the 2010 Census of Population and Housing, the trade-off between accuracy and privacy protection was viewed as a technical matter to be determined by disclosure avoidance statisticians.[5] Disclosure avoidance was performed primarily using household-level record swapping and was supported by maintaining the secrecy of key disclosure avoidance parameters.

However, there is a growing recognition in the scientific community that record-level household swapping fails to provide provable privacy guarantees. There is also growing concern that it may be possible to reconstruct a significant portion of the confidential data that underlies the census data releases using a so-called *database reconstruction attack*, as outlined by Dinur and Nissim (2003), and that such reconstructed microdata could be used to successfully re-identify the respondents who provided a significant proportion of the underlying confidential data. Indeed, in 2019 the Census Bureau announced that it had performed a database reconstruction attack using just the publicly available 2010 decennial census publications and had been able to reconstruct microdata that was overwhelmingly consistent with the 2010 confidential microdata.

In order to fulfill its requirements to produce an accurate count and to protect personally identifiable information, the Disclosure Avoidance System for the 2020 Census will implement a new approach to disclosure avoidance that applies mathematically rigorous disclosure avoidance controls to provide the required Title 13 data protections for the released data. The Disclosure Avoidance System (DAS) will read the Census Edited File (CEF) and apply formally private algorithms to produce a Microdata Detail File (MDF). By design, the CEF will contain information that is protected by Title 13, while the MDF will not. The MDF will then be tabulated to create the P.L. 94-171 Redistricting data and the Demographics and Housing Characteristics (DHC) tables for Persons and Housing Units. It is these tables, produced from the 2010 CEF, that have been publicly released as part of the 2010 Demonstration Data Product release.

Thus, the DAS can be thought of as a privacy filter or barrier that allows some aspects of data to pass while preventing leaks of Title 13 data. As an important side effect, all data that are publicly released by the Census Bureau based on the 2020 Census must go through the 2020 Disclosure Avoidance System, although the modules for some tables, not part of the 2010 Demonstration Data Products have not yet been incorporated into the DAS. The data products that must be processed by the DAS include the PL94-171 redistricting data, any summary files, quality assurance reports shared outside the Census Bureau, and other kinds of statistical summaries. (The Census Bureau is not holding releases based on the 2010 data to this standard.)

[4] https://www.census.gov/programs-surveys/decennial-census/about/why.html.

[5] Note: for historical reasons, the term *disclosure avoidance* is used at the US Census Bureau to describe statistical disclosure limitation; that term will be used in the remainder of this document.

Please see Section 3 for more information on the Design Decisions made by the Data Stewardship Executive Policy Committee for the 2010 Demonstrations Data Products.

Visit the following websites for more information on the 2010 Demonstration Data Products release:

- https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products/faqs.html
- https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html
- https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/

3. DSEP DESIGN DECISIONS

While the Data Stewardship Executive Policy Committee (DSEP) is responsible for significant decisions, actions, and accomplishments of the 2020 Census Program, the Associate Director for Decennial Programs publicly documents these policies in the 2020 Census Decision Memorandum Series for the purpose of informing stakeholders, coordinating interdivisional efforts, and documenting important historical changes. This memorandum series is available at https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html.

On September 12, 2019, the DSEP was asked to make a final decision on the design parameters for the version of the DAS that would be used to create the 2010 Demonstration Data Products. When making these decisions, DSEP reviewed many supporting materials and illustrations of the accuracy privacy-loss tradeoffs of several values of the privacy-loss budget (PLB, the technical parameter "epsilon" or "ε"), as well as analysis provided by the 2020 DAS team and the Demographic Programs Directorate. Possible values of the privacy-loss budget represent privacy/accuracy trade-offs along the spectrum between perfect privacy/low accuracy (ε = 0), to perfect accuracy/low privacy (ε = ∞). All of these decisions apply exclusively to the DAS used to produce the 2010 Demonstration Data Products and do not extend to the version of the DAS that will be used for the 2020 Census itself. DSEP will decide on the DAS design parameters for the 2020 Census publications separately and decide on a value of epsilon at a later date.

DSEP approved the following design parameters for the DAS used to produce the 2010 Demonstration Data Products. To reiterate, these figures will be produced from data collected during the 2010 Census:

- The total population will be reported as enumerated (invariant) at the state level.
- The count of total housing units (not population) will be reported as enumerated (invariant) at the block level.
- The count of group quarters facilities (not population) by the seven category types used in P.L. 94-171 Table P5 will be used as enumerated (invariant) at the block level.
- The global privacy-loss budget for these publications will be an epsilon of six (ε = 6.0). Of that six:
    - Four (ε = 4.0) will be allocated to the microdata detail file that supports population tables produced in the 2010 Demonstration Data Product.
    - Two (ε = 2.0) will be allocated to the microdata detail file that supports the housing and household tables in the 2010 Demonstration Data Product.

View the Memorandum 2019.25: 2010 Demonstration Data Products – Design Parameters and Global Privacy-Loss Budget for more information on these decisions.

4. DAS OPERATIONAL OVERVIEW

In production, the 2020 DAS operations team launches a pre-defined Amazon Web Services (AWS) Elastic Map Reduce (EMR) cluster. This cluster includes a bootstrap script that installs pre-defined DAS code and associated software on each of the member nodes.

AWS EMR supports three kinds of nodes: a single MASTER node, CORE nodes that are used for both data storage and for computation, and TASK nodes that are only used for computation. DAS requires a single MASTER node and at least two CORE nodes, although in practice substantially more CORE nodes are required.

The data transferred (file at rest) to the storage bucket attached to these nodes are encrypted using the AWS Key Management Service (KMS) utilizing 256-bit Advanced Encryption Standard (AES) encryption. Once the data are transferred to the specified S3 bucket, the master node of the cluster reads and executes the DAS code. The DAS creates an internal representation of the estimated 330 million persons and 140 million households in the U.S. in 2020. This population is arranged in a multi-dimensional national histogram (MDNH). Measurements (statistical summaries of the MDNH) are taken at each geographical level. Statistical noise is added to each measurement taken from the MDNH, privatizing the assemblage of data. Finally, the results are post-processed to create a consistent set of microdata that can be used for tabulation.

5. DAS INFRASTRUCTURE SPECIFICATION

AWS Cluster Instantiation and Bootstrap Configuration

The DAS cluster infrastructure is a managed service provided by the Technical Integration (TI) program. The TI controls the cluster configuration, which includes the AWS machine type, number of nodes for each cluster, node types for both head and core nodes, and the amount and configuration of allocated AWS Elastic Block Store (EBS) attached to each cluster node. The cluster specification used during the production of the 2010 Demonstration Data Products is given in Table 1.

Table 1: 2010 Demonstration Data Products Cluster Specification
Technical Infrastructure 2020 AWS Cloud Environment

AWS Cluster Node Type	AWS EMR Cluster Role	Node Count	vCPU (Cores)	RAM (GB)	EBS Storage (GB)
MASTER	m4.16xlarge	1	64	256	1000
CORE	r5.24xlarge	24	96	768	1000

After the TI instantiates the cluster, a sequence of bootstrap scripts installs and configures the software necessary for DAS to execute. The first bootstrap script installs the AWS Amazon Linux (Center for Internet Security (CIS) Secure Baseline) hardened image and installs the necessary end protection, security, and monitoring tools.

After the first bootstrap script completes, a second bootstrap script installs the necessary DAS components and software tools. This bootstrap also configures the necessary cluster permissions, license configurations for the Gurobi Optimizer, and other DAS-specific node configuration settings. This completes the cluster instantiation step.

The completion of the bootstrap on every node causes the DAS Step 1 script to execute on the master node. For the creation of the 2010 Demonstration Products, DAS subject-matter experts initiated the DAS Core Application Framework manually. In the final 2020 production run of the DAS system, this process will be automated.

Acquiring and Verifying the CEF

When the DAS starts, the CEF has been pre-positioned in an appropriate S3 bucket at a pre-specified location. The DAS verifies that the CEF is present and properly formatted.

For the 2010 Demonstration Data Products release, the 2010 CEF was used.

DAS Core Application

DAS next executes the TopDown Algorithm (TDA) using the approach outlined below. This algorithm utilizes the Gurobi optimizer in parallel to generate microdata maximally consistent with a set of noisy (formally private) measurements.

First, at the national level (a single geographic unit), approximately 400,000 differentially private, noisy summary query measurements are taken, on which some pre-specified proportion of the global privacy budget is expended (e.g., 1/6 of total privacy budget might be spent, if the global budget is evenly split between the geographic levels). See the Appendix for the allocation of the PLB in the 2010 Demonstration Data Products.

- These summary queries are then post-processed (primarily through the solution of large-scale linear programming, quadratic programming, and mixed-integer linear/quadratic programming models constructed and solved with the Gurobi Optimizer) to generate a national-level histogram that is informationally equivalent to microdata.

- The synthetic individual and household records generated at the national level are then allocated to the 52 state-and-state-equivalent geographies using a second formally private algorithm, which again involves taking differentially private measurements (this time returning noisy counts at the state level) and generating histograms (i.e., microdata) of persons and households (but now at the state level) consistent with known invariants.
- In analogous fashion, the individuals and households in each state are then allocated to the 3,143 counties, and then to the 73,057 census tracts, and then to 217,740 block groups, and finally to the 11,078,297 blocks.
- The "fan-out problem" occurs when the set of data structures that have to be held in memory is large due to the number of child geographic units associated with a parent geographic unit. To address the fan-out problem for counties with a large number of tracts, a synthetic geographic level, "Tract Group", was introduced as an intermediary between counties and tracts. Counties are represented with 5-digit GEOIDs and tracts are represented with 11-digit GEOIDs. Tract Groups are then represented with 9-digit GEOIDs, where the first 4 digits of the tract component (representing the original tract geographies that were then split as tracts evolve over time) are appended to the county GEOID.
- The taking of formally private measurements at each geographic level in order to make informed microdata-generation and allocation decisions consumes some portion of the privacy-loss budget.
- With each allocation, the DAS ensures that several variables will be "invariant" — that is, that the tabulations of the synthetic data exactly match the tabulations of the CEF. For the 2010 Census and the 2010 Demonstration Products, the invariants are listed in Table 2:
    - ○ **C1:** Total population (invariant at the state level).
    - ○ **C2:** Number of housing units (invariant at the block level).
    - ○ **C5:** Number of group quarters facilities' by group quarters facilities' type (invariant at the block level).
- For other variables, the DAS will attempt to make the allocation of synthetic individuals match as closely as possible the actual tabulations of these variables in the CEF within the constraints allowed by the privacy-loss budgets. This goal is achieved by expending the privacy budget as carefully as is possible when taking formally private measurements at each geographic level, and then using mathematical optimization to generate microdata/allocations that closely match the noisy measurements taken.

Table 2: 2010 Census and 2010 Demonstration Data Products Invariants*

Invariant	Definition	2010 Geographic Level
C1	Total population	state
C2	Voting-age population	removed
C3	Total Number of housing units	block
C4	Number of occupied and vacant housing units	removed
C5	Number and type of group quarters facilities'	block

*Note: these are different from the 2018 End-to-End Census Test.

END OF PROCESSING

After DAS processing, the MDF file is then written back to the specified S3 bucket, again encrypted at rest using AWS KMS. The file is then transferred to other decennial systems for further processing.

MICRODATA FILE SPECIFICATIONS

This section of the document outlines the Decennial Census Management Division (DCMD), Center for Enterprise Dissemination - Disclosure Avoidance (CED-DA), and Decennial Information Technology Division (DITD) specifications to create the Microdata Detail File. This specification contains the Record Layouts for the two sections of the MDF.

Table 3: Production Input

Data Title	Data File Name
2010 Census Unit File Information	$DAS_S3ROOT/title13_input_data/table12a_20190705/ $DAS_S3ROOT/title13_input_data/table10_20190610
2010 Census Person File Information	s3://uscb-decennial-ite-das/title13_input_data/table1a_20190709/ s3://uscb-decennial-ite-das/title13_input_data/table10/

Table 4: Production Output

Data Title	Data File Name
Unit File Information	s3://uscb-decennial-ite-das/DHC_DemonstrationProduct_Fixed/DHCH/
Person File Information	s3://uscb-decennial-ite-das/users/lecle301/DemonstrationProducts_Sept2019_fixedTotals/full_person

Note that the two output files are both pickle files. A conversion program was used to convert the pickle files into the MDF_PERSON and MDF_UNIT files. For more on this conversion program, see "Output Conversion Program" at the bottom of section 5.

Table 5: Glossary and Conventions used in Record Layouts

Terminology	Definition
CENHISP	A recode of the eight edited Hispanic origin codes into 2 values representing Hispanic and not Hispanic.
CENRACE	A recode of the eight edited race codes into a single 2-digit code representing one of 63 race group categories.
CHAR(#)	A fixed-width field of # characters long. **CHAR is used for numbers if the numbers are not used for mathematical operations. CHAR is used for zero-filled numbers.**
Disclosure Avoidance (DA)	Items noted with Disclosure Avoidance (DA) have undergone disclosure avoidance in accordance with DSEP policy.
FINAL_POP	Final Population Count from the CUF – includes count imputation.
Linkage Variable	A variable that links between two tables.
INT(#)	An Integer up to # characters wide. Not zero-filled.
Not Reported	Items noted as Not Reported in the 2010 Demonstration Data Products MDF represent data that might be included in the 2020 MDF but are not present in the 2010 Demonstration Data Products MDF due to policy or procedural reasons. They are indicated with the notation "Not Reported"
Pipe delimited	A "pipe-delimited" file is a text file in Unicode UTF-8 encoding in which each field is separated by the Unicode Character "VERTICAL LINE" (U+007C) (e.g., " | ") also known as the "pipe" character from its use in Unix pipelines.
QAGE	Edited Age as defined in the Edits and Characteristics Imputation Specification.
QRACEX	Edited Race Groups as defined in the Edits and Characteristics Imputation Specification.
QREL	Edited Relationships as defined in the Edits and Characteristics Imputation Specification.
QSEX	Edited Sex as defined in the Edits and Characteristics Imputation Specification.
Recode	A recode is a new variable that is created by combining or collapsing the value categories of an existing variable
Protected recode	A protected recode is a new variable created from **existing protected** variables.
Redundant; Remove	Items noted as Redundant exactly replicate other items already in the MDF.
RTYPE	Record Type
TEN	Edited Tenure

Microdata File Record Layouts

Microdata Detail File (MDF) Person Data Notes:

1. Data will be pipe-delimited.
2. Habitable geographies without persons will have disclosure avoidance applied; as a result, zero population blocks can gain population under the 2010 Demonstration Data Products invariants. This is a change from 2018 ETE.
3. It is not possible to perform meaningful joins between this table and MDF.Unit.
4. This table links to GRF-C using the TABBLKST, TABBLKCOU, TABTRACTCE, TABBLKGRPCE, and TABBLK linkage variables.

Table 6: MDF.Person

#	Name	Label	Type	Values	Recode
1	**SCHEMA_TYPE_CODE**	Schema Type Code	CHAR(3)	MPD	
2	**SCHEMA_BUILD_ID**	Schema Build ID	CHAR(5)	1.1.0	
3	**TABBLKST**	2010 Tabulation State (FIPS)	CHAR(2)	01-02 04-06 08-13 15-42 44-51 53-56 72	

#	Name	Label	Type	Values	Recode
4	**TABBLKCOU**	2010 Tabulation County (FIPS)	CHAR(3)	001-840	
5	**TABTRACTCE**	2010 Tabulation Census Tract	CHAR(6)	000100-998999	
6	**TABBLKGRPCE**	2010 Census Block Group	CHAR(1)	0-9	
7	**TABBLK**	2010 Block Number	CHAR(4)	0001-9999	
8	**EPNUM**	Privacy Edited Person Number	INT(9)	999999999 = Not reported	
9	**RTYPE**	Record Type	CHAR(1)	3 = Person in housing unit	
				5 = Person in group quarters facilities'	
10	**GQTYPE**	Group Quarters Facilities Type'	CHAR(3)	000 = NIU	
				101 = Federal detention centers; Federal prisons; State prisons; Local jails and other municipal confinement facilities; Local jails and other municipal confinement facilities; Correctional residential facilities; Military disciplinary barracks and jails	
				201 = Group homes for juveniles (non-correctional); Residential treatment centers for juveniles (non-correctional); Correctional facilities intended for juveniles	
				301 = Nursing facilities/skilled-nursing facilities	
				401 = Mental (psychiatric) hospitals and psychiatric units in other hospitals; Hospitals with patients who have no usual home elsewhere; In-patient hospice facilities; Military treatment facilities with assigned patients; Residential schools for people with disabilities	
				501 = College/university student housing	
				601 = Military quarters; Military ships	
				701 = Emergency and transitional shelters (with sleeping facilities) for people experiencing homelessness; Group homes intended for adults; Residential treatment centers for adults; Maritime/merchant vessels; Workers' group living quarters and job corps centers; Other non-institutional facilities (GQ types 702, 704, 706, 903, 904)	
11	**RELSHIP**	Edited Relationship	CHAR(2)	99 = Not Reported	
12	**QSEX**	Edited Sex	CHAR(1)	1 = Male	
				2 = Female	
13	**QAGE**	Edited Age	INT(3)	0-115	
14	**CENHISP**	Hispanic Origin	CHAR(1)	1 = Not Hispanic	
				2 = Hispanic	
15	**CENRACE**	Census Race	CHAR(2)	01 = White alone	
				02 = Black alone	
				03 = AIAN alone	
				04 = Asian alone	
				05 = NHPI alone	
				06 = SOR alone	
				07 = White; Black	
				08 = White; AIAN	
				09 = White; Asian	
				10 = White; NHPI	
				11 = White; SOR	
				12 = Black; AIAN	

#	Name	Label	Type	Values	Recode
				13 = Black; Asian	
				14 = Black; NHPI	
				15 = Black; SOR	
				16 = AIAN; Asian	
				17 = AIAN; NHPI	
				18 = AIAN; SOR	
				19 = Asian; NHPI	
				20 = Asian; SOR	
				21 = NHPI; SOR	
				22 = White; Black; AIAN	
				23 = White; Black; Asian	
				24 = White; Black; NHPI	
				25 = White; Black; SOR	
				26 = White; AIAN; Asian	
				27 = White; AIAN; NHPI	
				28 = White; AIAN; SOR	
				29 = White; Asian; NHPI	
				30 = White; Asian; SOR	
				31 = White; NHPI; SOR	
				32 = Black; AIAN; Asian	
				33 = Black; AIAN; NHPI	
				34 = Black; AIAN; SOR	
				35 = Black; Asian; NHPI	
				36 = Black; Asian; SOR	
				37 = Black; NHPI; SOR	
				38 = AIAN; Asian; NHPI	
				39 = AIAN; Asian; SOR	
				40 = AIAN; NHPI; SOR	
				41 = Asian; NHPI; SOR	
				42 = White; Black; AIAN; Asian	
				43 = White; Black; AIAN; NHPI	
				44 = White; Black; AIAN; SOR	
				45 = White; Black; Asian; NHPI	
				46 = White; Black; Asian; SOR	
				47 = White; Black; NHPI; SOR	
				48 = White; AIAN; Asian; NHPI	
				49 = White; AIAN; Asian; SOR	
				50 = White; AIAN; NHPI; SOR	
				51 = White; Asian; NHPI; SOR	
				52 = Black; AIAN; Asian; NHPI	
				53 = Black; AIAN; Asian; SOR	
				54 = Black; AIAN; NHPI; SOR	
				55 = Black; Asian; NHPI; SOR	
				56 = AIAN; Asian; NHPI; SOR	
				57 = White; Black; AIAN; Asian; NHPI	
				58 = White; Black; AIAN; Asian; SOR	
				59 = White; Black; AIAN; NHPI; SOR	
				60 = White; Black; Asian; NHPI; SOR	
				61 = White; AIAN; Asian; NHPI; SOR	
				62 = Black; AIAN; Asian; NHPI; SOR	
				63 = White; Black; AIAN; Asian; NHPI; SOR	
16	**LIVE_ALONE**	Person Living Alone	CHAR(1)	9 = Not Reported	

Microdata Detail File (MDF) Unit Data Notes:

1. Data will be pipe-delimited.
2. Geographies without units will not have disclosure avoidance applied; nor will they be present in the MDF.Unit file.
3. It is not possible to perform meaningful joins between this table and MDF.Person.
4. This table links to GRF-C using the TABBLKST, TABBLKCOU, TABTRACTCE, TABBLKGRPCE, and TABBLK linkage variables.
5. Throughout this section of the specification, “for the householder” indicates “RELSHIP = 20”.

Table 7: MDF.Unit

#	Name	Label	Type	Values	Recode
1	**SCHEMA_TYPE_CODE**	Schema Type Code	CHAR(3)	MUD	
2	**SCHEMA_BUILD_ID**	Schema Build ID	CHAR(5)	1.1.0	
3	**TABBLKST**	2010 Tabulation State (FIPS)	CHAR(2)	01-02	
				04-06	
				08-13	
				15-42	
				44-51	
				53-56	
				72	
4	**TABBLKCOU**	2010 Tabulation County (FIPS)	CHAR(3)	001-840	
5	**TABTRACTCE**	2010 Tabulation Census Tract	CHAR(6)	000100-998999	
6	**TABBLKGRPCE**	2010 Census Block Group	CHAR(1)	0-9	
7	**TABBLK**	2010 Block Number	CHAR(4)	0001-9999	
8	**RTYPE**	Record Type	CHAR(1)	2 = Housing unit	
				4 = Group quarters facilities'	
9	**GQTYPE**	Group Quarters Facilities' Type	CHAR(3)	000 = NIU	
				101 = Federal detention centers	
				102 = Federal prisons	
				103 = State prisons	
				104 = Local jails and other municipal confinement facilities	
				105 = Correctional residential facilities	
				106 = Military disciplinary barracks and jails	
				201 = Group homes for juveniles (non-correctional)	
				202 = Residential treatment centers for juveniles (non-correctional)	
				203 = Correctional facilities intended for juveniles	
				301 = Nursing facilities/skilled nursing facilities	
				401 = Mental (psychiatric) hospitals and psychiatric units in other hospitals	
				402 = Hospitals with patients who have no usual home elsewhere	
				403 = In-patient hospice facilities	
				404 = Military treatment facilities with assigned patients	
				405 = Residential schools for people with disabilities	
				501 = College/university student housing	
				601 = Military quarters	
				602 = Military ships	
				701 = Emergency and transitional shelters (with sleeping facilities) for people experiencing homelessness	
				801 = Group homes intended for adults	
				802 = Residential treatment centers for adults	
				900 = Maritime/merchant vessels	
				901 = Workers' group living quarters and job corps centers	
				702 = Soup Kitchens	
				704 = Regularly scheduled Mobile Food Vans	
				706 = Targeted Non-Sheltered Outdoor Locations	

#	Name	Label	Type	Values	Recode
				903 = Living Quarters for Victims of Natural Disasters	
				904 = Religious Group Quarters Facilities' and Domestic Violence Shelters	
10	**TEN**	Tenure	CHAR(1)	0 = NIU 9 = Occupied	
11	**VACS**	Vacancy Status	CHAR(1)	0 = NIU 9 = Vacant	
12	**HHSIZE**	Population Count	INT(5)	0 (vacant or NIU),1,2,3,4,5,6,7+	
13	**HHT**	Household/Family Type	CHAR(1)	0=NIU	(RTYPE= 2 and HHSIZE =0) or RTYPE =4
				1 = Married couple household	RTYPE = 2 and HHSIZE > 1 and (RELSHIP = 21, 23 for one person in housing unit)
				2 = Other family household: Male householder	RTYPE = 2 and HHSIZE > 1 and (RELSHIP ≠ 21, 23 for any person in housing unit) and (RELSHIP = 25-33 for one or more people in housing unit) and (QSEX = 1 for householder)
				3 = Other family household: Female householder	RTYPE = 2 and HHSIZE > 1 and (RELSHIP ≠ 21, 23 for any person in housing unit) and (RELSHIP = 25-33 for one or more people in housing unit) and (QSEX = 2 for householder)
				4 = Nonfamily household: Male householder, living alone	RTYPE = 2 and HHSIZE = 1 and (QSEX = 1 for householder)
				5 = Nonfamily household: Male householder, not living alone	RTYPE = 2 and HHSIZE ≥ 2 and (RELSHIP = 22, 24, 34-36 for all people in housing unit besides householder) and (QSEX = 1 for householder)
				6 = Nonfamily household: Female householder, living alone	RTYPE = 2 and HHSIZE = 1 and (QSEX = 2 for householder)
				7 = Nonfamily household: Female householder, not living alone	RTYPE = 2 and HHSIZE ≥ 2 and (RELSHIP = 22, 24, 34-36 for all persons in housing unit besides householder) and (QSEX = 2 for householder)
14	**HHT2**	Household/Family Type (Includes Cohabitating)	CHAR(2)	00 = NIU	(RTYPE = 2 and HHSIZE = 0) or RTYPE = 4
				01 = Married couple household: With own children < 18	RTYPE = 2 and HHSIZE > 1 and (RELSHIP = 21, 23 for one person in housing unit) and (RELSHIP = 25-27 and QAGE < 18 for at least one person in housing unit)
				02 = Married couple household: No own children < 18	RTYPE = 2 and HHSIZE > 1 and (RELSHIP = 21,

#	Name	Label	Type	Values	Recode
					23 for one person in housing unit) and (RELSHIP = 25-27 and QAGE < 18 for no person in housing unit)
				03 = Cohabiting couple household: With own children < 18	RTYPE = 2 and HHSIZE > 1 and (RELSHIP = 22, 24 for one person in housing unit) and (RELSHIP = 25-27 and QAGE < 18 for at least one person in housing unit)
				04 = Cohabiting couple household: No own children < 18	RTYPE = 2 and HHSIZE > 1 and (RELSHIP = 22, 24 for one person in housing unit) and (RELSHIP = 25-27 and QAGE < 18 for no person in housing unit)
				05 = Female householder, no spouse/partner present: Living alone	RTYPE = 2 and HHSIZE = 1 and (QSEX = 2 for householder)
				06 = Female householder, no spouse/partner present: With own children < 18	RTYPE = 2 and HHSIZE > 1 and (QSEX = 2 for householder) and (RELSHIP = 21-24 for no person in housing unit) and (RELSHIP = 25-27 and QAGE < 18 for at least one person in housing unit)
				07 = Female householder, no spouse/partner present: With relatives, no own children < 18	RTYPE = 2 and HHSIZE > 1 and (QSEX = 2 for householder) and (RELSHIP = 21-24 for no person in housing unit) and (RELSHIP = 25-33 for at least one person in housing unit) and (RELSHIP = 25-27 and QAGE <= 17 for no person in the housing unit)
				08 = Female householder, no spouse/partner present: Only nonrelatives present	RTYPE = 2 and HHSIZE > 1 and (QSEX = 2 for householder) and (RELSHIP = 21-33 for no person in housing unit) and (RELSHIP = 34-36 for at least one person in the housing unit)
				09 = Male householder, no spouse/partner present: Living alone	RTYPE = 2 and HHSIZE = 1 and (QSEX = 1 for householder)
				10 = Male householder, no spouse/partner present: With own children < 18	RTYPE = 2 and HHSIZE > 1 and (QSEX = 1 for householder) and (RELSHIP = 21-24 for no person in housing unit) and (RELSHIP = 25-27 and QAGE < 18 for at least one person in housing unit)

#	Name	Label	Type	Values	Recode
				11 = Male householder, no spouse/partner present: With relatives, no own children < 18	RTYPE = 2 and HHSIZE > 1 and (QSEX = 1 for householder) and (RELSHIP = 21-24 for no person in housing unit) and (RELSHIP = 25-27 and QAGE <= 17 for no person in housing unit) and (RELSHIP = 25-33 for at least one person in housing unit)
				12 = Male householder, no spouse/partner present: Only nonrelatives present	RTYPE = 2 and HHSIZE > 1 and (QSEX = 1 for householder) and (RELSHIP = 21-33 for no person in housing unit) and (RELSHIP = 34-36 for at least one person in the housing unit)
15	**CPLT**	Couple Type	CHAR(1)	0 = NIU	(RTYPE = 2 and HHSIZE ≤ 1) or RTYPE = 4
				1 = Opposite-sex husband/wife/spouse household	RTYPE = 2 and HHSIZE > 1 and (RELSHIP = 21 for one person in housing unit)
				2 = Same-sex husband/wife/spouse household	RTYPE = 2 and HHSIZE > 1 and (RELSHIP = 23 for one person in housing unit)
				3 = Opposite-sex unmarried partner household	RTYPE = 2 and HHSIZE > 1 and (RELSHIP = 22 for one person in housing unit)
				4 = Same-sex unmarried partner household	RTYPE = 2 and HHSIZE > 1 and (RELSHIP = 24 for one person in housing unit)
				5 = All other households	RYPTE = 2 and HHSIZE > 1 and RELSHIP in (21, 22, 23, 24) for no persons in housing unit
16	**UPART**	Presence and Type of Unmarried Partner Household	CHAR(1)	0 = NIU	(RTYPE = 2 and HHSIZE = 0) or RTYPE = 4
				1 = Male householder and male partner	RTYPE = 2 and HHSIZE > 1 and QSEX = 1 for householder and (RELSHIP = 24 and QSEX = 1 for one person in housing unit)
				2 = Male householder and female partner	RTYPE = 2 and HHSIZE > 1 and QSEX = 1 for householder and (RELSHIP = 22 and QSEX = 2 for one person in housing unit)
				3 = Female householder and female partner	RTYPE = 2 and HHSIZE > 1 and QSEX = 2 for householder and (RELSHIP = 24 and QSEX = 2 for one person in housing unit)
				4 = Female householder and male partner	RTYPE = 2 and HHSIZE > 1 and QSEX = 2 for

#	Name	Label	Type	Values	Recode
					householder and (RELSHIP = 22 and QSEX = 1 for one person in housing unit)
				5 = All other households	RTYPE = 2 and HHSIZE ≥ 1 and RELSHIP = 22, 24 for no persons in housing unit
17	**MULTG**	Multigenerational Household	CHAR(1)	0 = NIU	(RTYPE = 2 and HHSIZE <= 2) or RTYPE = 4
				1 = Not a multigenerational household	RTYPE = 2 and HHSIZE>= 3 and [RELSHIP = 25-27 for no person in the housing unit OR (RELSHIP = 25-27 for at least one person in the housing unit, and RELSHIP = 29-31 for no person in the housing unit)
				2 = Yes, a multigenerational household	RTYPE = 2 and HHSIZE ≥ 3 and (RELSHIP = 25-27 for at least one person in housing unit) and [(RELSHIP = 30 for at least one person in housing unit) or (RELSHIP = 29, 31 for at least one person in housing unit)]
18	**HHLDRAGE**	Age of Householder	CHAR(1)	0 = NIU	(RYPTE = 2 and HHSIZE = 0) or RTYPE = 4
				1=Householder 15 to 24 years	QAGE = 15-24 for householder
				2=Householder 25 to 34 years	QAGE = 25-34 for householder
				3=Householder 35 to 44 years	QAGE = 35-44 for householder
				4=Householder 45 to 54 years	QAGE = 45-54 for householder
				5=Householder 55 to 59 years	QAGE = 55-59 for householder
				6=Householder 60 to 64 years	QAGE = 60-64 for householder
				7=Householder 65 to 74 years	QAGE = 65-74 for householder
				8=Householder 75 to 84 years	QAGE = 75-84 for householder
				9=Householder 85 years and over	QAGE = 85-115 for householder
19	**HHSPAN**	Hispanic Householder	CHAR(1)	0 = NIU	(RYPTE = 2 and HHSIZE = 0) or RTYPE = 4
				1 = Not Hispanic	RTYPE = 2 and HHSIZE > 0 and CENHISP = 1 for householder
				2 = Hispanic	RTYPE = 2 and HHSIZE > 0 and CENHISP = 2 for householder
20	**HHRACE**	Race of Householder	CHAR(2)	00 = NIU	(RTYPE = 2 and HHSIZE = 0) or RTYPE = 4
				01 = White alone	RTYPE = 2 and HHSIZE > 0 and CENRACE = 01 for householder

#	Name	Label	Type	Values	Recode
				02 = Black alone	RTYPE = 2 and HHSIZE > 0 and CENRACE = 02 for householder
				03 = AIAN alone	RTYPE = 2 and HHSIZE > 0 and CENRACE = 03 for householder
				04 = Asian alone	RTYPE = 2 and HHSIZE > 0 and CENRACE = 04 for householder
				05 = NHPI alone	RTYPE = 2 and HHSIZE > 0 and CENRACE = 05 for householder
				06 = SOR alone	RTYPE = 2 and HHSIZE > 0 and CENRACE = 06 for householder
				07 = Two or more races	RTYPE = 2 and HHSIZE > 0 and CENRACE = 07-63 for householder
21	**PAOC**	Presence and Age of Own Children Under 18	CHAR(1)	0 = NIU	(RTYPE = 2 and HHSIZE = 0) or RTYPE = 4
				1 = With own children under 6 years only	RTYPE = 2 and HHSIZE > 1 and (RELSHIP = 25-27 and QAGE < 6 for at least one person in housing unit) and (RELSHIP = 25-27 and QAGE = 6-17 for no person in housing unit)
				2 = With own children 6-17 years only	RTYPE = 2 and HHSIZE > 1 and (RELSHIP = 25-27 and QAGE = 6-17 for at least one person in housing unit) and (RELSHIP = 25-27 and QAGE < 6 for no person in housing unit)
				3 = With own children under 6 years and 6-17 years	RTYPE = 2 and HHSIZE > 1 and (RELSHIP = 25-27 and QAGE < 6 for at least one person in housing unit) and (RELSHIP = 25-27 and QAGE = 6-17 for at least one person in housing unit)
				4 = No own children under 18	RTYPE = 2 and HHSIZE = 1 or (HHSIZE > 1 and (RELSHIP = 25-27 and QAGE < 18 for no persons in housing unit)
22	**P18**	Presence of People Under 18 Years in Household	CHAR(1)	9 = Not reported	
23	**P60**	Presence of People 60 Years and Over in Household	CHAR(1)	0 = NIU	(RTYPE = 2 and HHSIZE = 0) or (RTYPE = 2 and QAGE < 60 for all persons in housing unit) or RTYPE = 4
				1=With one or more people 60 years and over in household	Where RTYPE = 2 and HHSIZE > 0 and QAGE ≥ 60 for at least one person in housing unit

#	Name	Label	Type	Values	Recode
24	**P65**	Presence of People 65 Years and Over in Household	CHAR(1)	0 = NIU	(RTYPE = 2 and HHSIZE = 0) or (RTYPE = 2 and QAGE < 65 for all persons in housing unit) or RTYPE = 4
				1=With on e or more people 65 years and over in household	Where RTYPE = 2 and HHSIZE > 0 and QAGE ≥ 65 for at least one person in housing unit
25	**P75**	Presence of People 75 Years and Over in Household	CHAR(1)	0 = NIU	(RTYPE = 2 and HHSIZE = 0) or (RTYPE = 2 and QAGE < 75 for at all persons in housing unit) or RTYPE = 4
				1=With one or more people 75 years and over in household	Where RTYPE = 2 and HHSIZE > 0 and QAGE ≥ 75 for at least one person in housing unit
26	**PAC**	Presence and Age of Children Under 18	CHAR(1)	9 = Not Reported	
27	**HHSEX**	Sex of Householder	CHAR(1)	0 = NIU	(RTYPE = 2 and HHSIZE = 0) or RTYPE = 4
				1 = Male householder	QSEX = 1 for householder
				2 = Female householder	QSEX = 2 for householder

Output Conversion Program

The conversion program takes as input the result from a Core DAS application execution in the format of pickled data files and produces the same results as output in microdata detail file (MDF) format. The conversion program uses Spark to read in the pickled results, and makes use of existing DAS Writer components to write the results out in MDF format. The program was used during the production of the 2010 Demonstration Data Products to convert the results of the DHC-P and DHC-H from their pickled format into MDF format. The program was run once for the DHC-H pickled file and once for the DHC-P pickled file, and the outputs were written to S3 buckets relative to the location of the pickled files.

Using this program after the Core DAS execution currently produces metadata headers on the microdata files, which incorrectly indicate that the conversion programs were responsible for producing the data files. This will be corrected in a future version.

6. DESIGN PLAN SPECIFICATIONS

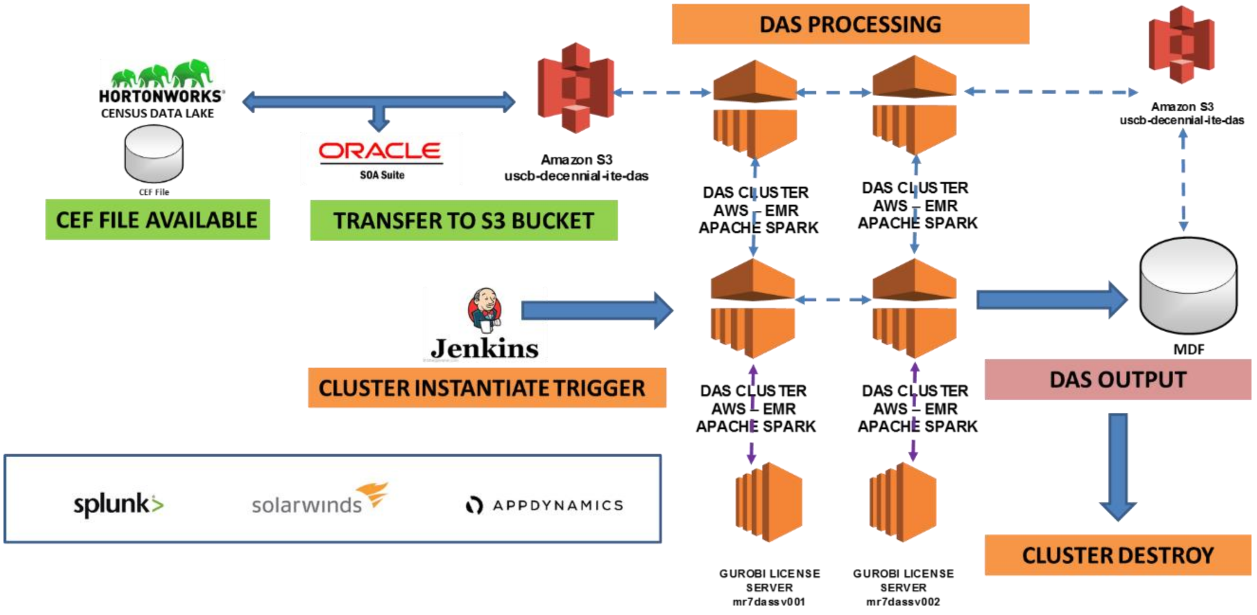
The DAS design plan can be described by the following components:

DAS Cluster Infrastructure

DAS utilizes AWS GovCloud and the following AWS components:

- AWS Elastic Map Reduce (EMR) Cluster installed with Apache Spark.
- AWS Simple Storage Service (S3).
- AWS Simple Notification Service (SNS).
- AWS Elastic Block Store (EBS).
- AWS Elastic Compute Cloud (EC2).

Figure 1: DAS Cluster Infrastructure



DAS ⇔ S3 Interface Framework

DAS receives the Census Edited File (CEF) from S3 and then writes back the results as a Microdata Detail File (MDF).

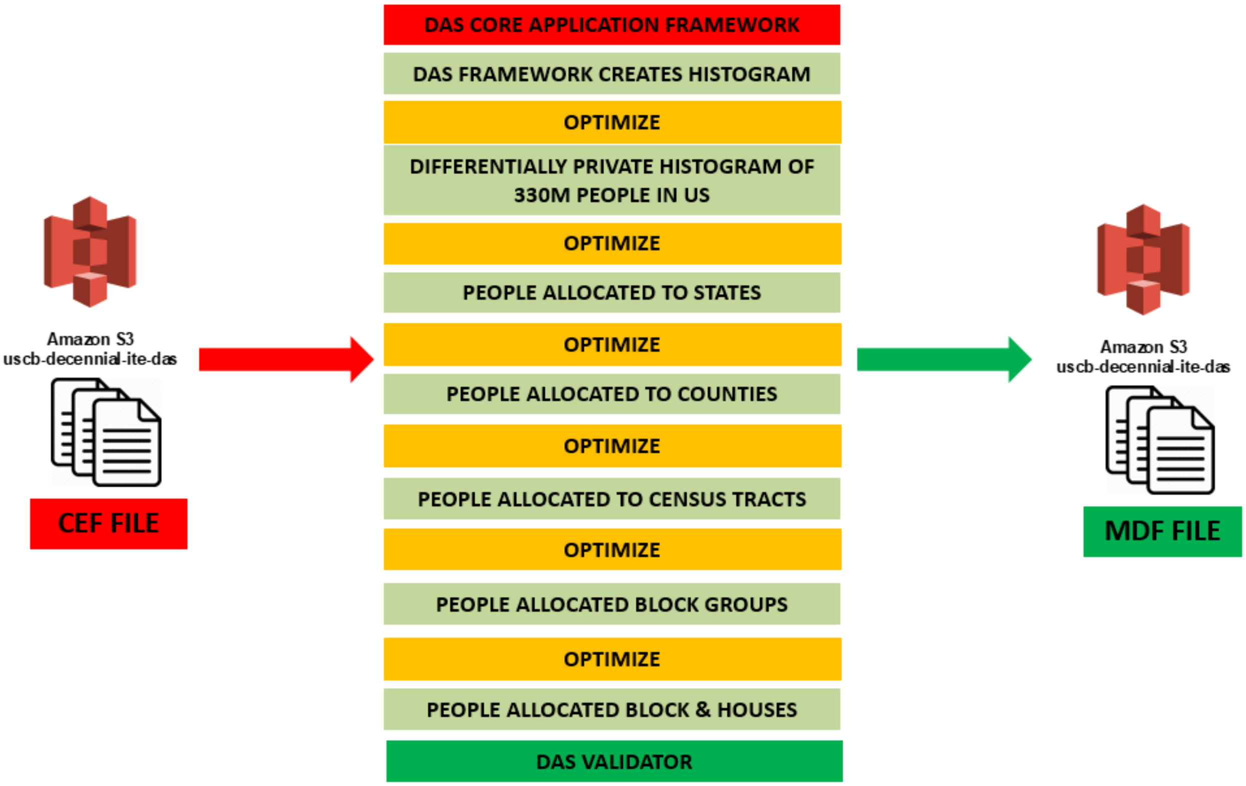
DAS Core Application Framework

The Disclosure Avoidance System (DAS) applies privacy controls to microdata in the data flow from the Census Edited File (CEF) to the Microdata Detail File (MDF). The privacy controls ensure that there is no direct mapping between individual records in the CEF to individual records in the MDF and regulate the privacy loss implied by the production of the MDF.

Following the application of the privacy controls, the microdata in the MDF is ready for tabulation.

The CEF contains other private information that is not destined for the MDF to support other Census Bureau business processes. Gurobi optimization software is used by DAS for mathematical optimization (typically to generate or allocate microdata from noisy measurements).

Figure 2: DAS Core Application Framework


DAS CORE APPLICATION FRAMEWORK
DAS FRAMEWORK CREATES HISTOGRAM
OPTIMIZE
DIFFERENTIALLY PRIVATE HISTOGRAM OF 330M PEOPLE IN US
OPTIMIZE
PEOPLE ALLOCATED TO STATES
OPTIMIZE
PEOPLE ALLOCATED TO COUNTIES
OPTIMIZE
PEOPLE ALLOCATED TO CENSUS TRACTS
OPTIMIZE
PEOPLE ALLOCATED BLOCK GROUPS
OPTIMIZE
PEOPLE ALLOCATED BLOCK & HOUSES
DAS VALIDATOR
Amazon S3
uscb-decennial-ite-das
CEF FILE
Amazon S3
uscb-decennial-ite-das
MDF FILE

7. SYSTEM ARCHITECTURE

The section below provides an overview of the DAS System Architecture.

Business Architecture

DAS is solving a particular data-processing problem for the Census Bureau, that is, to apply disclosure avoidance and privacy controls to uphold Title 13-mandated confidentiality protections to the data published from the 2020 Census. DAS is neither integrated into nor solving any Census Bureau operational, business, financial, or transactional functions.

Application Architecture (Front End)

DAS is designed and built without user access. DAS employs a single DAS operations account to run DAS from start to completion. No front-end user interface is built into DAS.

Information Architecture (Data)

DAS is designed and architected without any user data input, data transactions, or data storage requirements. DAS will not employ a database engine in order to complete the system's intended use case.

Data-at-Rest Security Controls:

Cluster Node(s) Attached Elastic Block Storage (EBS)

When an ***encrypted*** Amazon EBS volume is attached to a supported Amazon Elastic Map Reduce (EMR) instance, data stored at rest on the volume, disk I/O, and snapshots created from the volume are all encrypted. Amazon EBS encryption uses AWS Key Management Service (AWS KMS) customer master keys (CMKs) when creating encrypted volumes and any snapshots created from them. The encryption occurs on the servers that host Amazon Elastic Map Reduce (EMR) cluster node members. When the DAS EMR cluster is instantiated, an ***encrypted EBS volume*** is automatically attached to all the node members; as a result, the following types of data are encrypted:

- Data at rest inside the volume.
- All data moving between the volume and the instance.
- All snapshots created from the volume.
- All volumes created from those snapshots.

S3 Bucket Data Encryption

The TI 2020 Cloud supports AWS S3 bucket integration for AWS EMR. Every cluster that requires S3 storage will be assigned a specific S3 storage bucket, restricted to each project's cluster node members and configured to encrypt any file stored to the assigned bucket. Amazon S3 encryption provides a way to set the encryption behavior for an S3 bucket. DAS AWS EMR sets encryption on a bucket so that all objects are encrypted when they are stored in the bucket. The objects are encrypted using server-side encryption AWS KMS-managed keys (SSE-KMS).

Data-in-Flight Security Controls:

Clusters instantiated within the TI 2020 Cloud will be installed with TLS certificates for node-to-node communications for EMR-specific task execution.

Baseline Management

Architecture, infrastructure, and code baseline management will adhere to the TI-2020 Program Change Management Plan.

8. CLOUD BASED SECURITY CONSIDERATIONS

Table 8: Cloud Based Security Considerations

Consideration	Response
Uptime expectation from the Business Owner of the System in the Cloud:	100%
Cloud Based SLA for Security Monitoring:	100%
Cloud Based System Level High Availability:	YES
Cloud Based Site Level Disaster Recovery for the System:	Provided TI 2020 Cloud Capability
State whether the Cloud Vendor is FEDRAMP Certified:	YES
Data Retention Requirements:	YES

9. RELIABILITY, MAINTAINABILITY, AND AVAILABILITY CONSIDERATIONS

The design of the system is such that there is a single MASTER node and multiple WORKER and/or CORE nodes. If a WORKER/CORE node fails, the EMR system will restart that load and schedule work on the failed node to be re-computed on a new node. However, there are no provisions in EMR for a failed MASTER node. If the MASTER node fails, the system will need to be manually restarted.

The current design has minimal built-in checks. If the system fails during execution, all work will need to be redone from the beginning. The DAS development team will be adding check-pointing to the system at a later point. When check-pointing is added, the system will note which phase of the TDA executed last, and it will restart execution at that point.

The system has a growing number of self-tests that are executed using the Python "py.test" framework. These tests will check both the code and the execution environment. The "py.test" will be run by the DAS prior to the start of the TDA so that failures can be rapidly detected and diagnosed.

DAS is developing a framework for recording and alerting on out-of-memory or out-of-storage conditions.

DAS is based on Intel Corporation's Anaconda Python Distribution using Python Version 3.6.8.

The 2010 Demonstration Data Products were produced with Gurobi version 7.5; version 8.1 will be used for the 2020 Census.

DAS assumes that the Gurobi license manager will be available. If the license manager is not available, or if a license is not available, the client code will retry until it is. A RETRY LIMIT is not currently specified.

PERFORMANCE ENGINEERING CONSIDERATIONS

DAS system performance depends on the following considerations:

- File transfer performance from S3 to EMR.
- Performance of EMR distributing Python tasks.
- Performance of the Java-Python gateway.
- Performance of the Gurobi optimizer.
- File transfer performance from DAS back to S3.

PERFORMANCE METRICS

The DAS developers have developed a system for capturing the utilization of memory and CPU resources and matching them to individual runs of the DAS. The DAS developers will perform this by capturing per-process and per-CPU usage every 5 seconds using DFXML, aggregating the results on the DAS EMR MASTER node, and then transferring those results to S3 storage bucket. Separately, each use of the Gurobi optimizer captivates CPU usage, CPU load, memory usage, and other process information. This is all used to monitor system performance and tune the use of Gurobi during development.

SYSTEM PLATFORM AND DESIGN

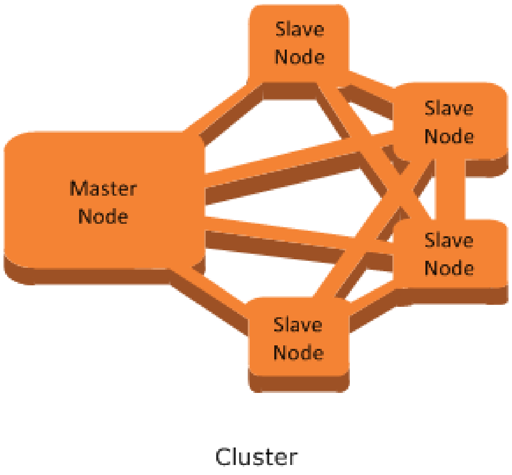
DAS utilizes Amazon EMR, which is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data.

The central component of Amazon EMR is the cluster. A cluster is a collection of Amazon Elastic Compute Cloud (Amazon EC2) instances. Each instance in the cluster is called a node. Each node has a role within the cluster, referred to as the node type. Amazon EMR also installs different software components on each node type, giving each node a role in a distributed application like Apache Hadoop.

The node types in Amazon EMR are as follows:

- **Master node**: A node that manages the cluster by running software components to coordinate the distribution of data and tasks among other nodes—collectively referred to as slave nodes—for processing. The master node tracks the status of tasks and monitors the health of the cluster.
- **Core node**: A slave node with software components that run tasks and store data in the Hadoop Distributed File System (HDFS)[6] on the cluster.

Figure 3: EMR Cluster Diagram



[6] DAS currently does not use HDFS.

10. API INTERFACE BUSINESS PURPOSE

The business purpose for this data exchange is to support privacy controls to microdata in the data flow from the Census Edited File (CEF) to the Microdata Detail File (MDF). The CDL platform will serve as a storage mechanism for the Disclosure Avoidance System. The Disclosure Avoidance System (DAS) to CDL interface is necessary in order for DAS to successfully complete its processing and write non-Title 13 data back to the Census Data Lake.

Interface Responsibilities

The interface includes a one directional flow of data from the Disclosure Avoidance System to the CDL. The DAS is responsible for providing the MDF to the CDL to allow for microdata tabulation. This interface will be executed using the SOA Managed File Transfer (MFT) process. The MDF file is structured as 2 tables for each of the 50 states, DC, and PR.

The 2010 Demonstration Products MDF Delivery Consists of Two Pipe-Delimited ASCII Text Files:

- MDF_PER.txt.
- MDF_UNIT.txt.

Disclosure Avoidance - CDL Data

The MDF was delivered from the DAS to the CDL in the form of pipe-delimited ASCII text files. DAS delivered the pipe-delimited ASCII files (2 tables for each state) to S3 using Spark. These “files” will appear as directories containing multiple parts. Spark convention is to designate parts as part-0001, part-0002, etc. The program *s3cat.py* located in *das_decennial/programs* directory is then used to combine these files into a single file that could be moved back to the CDL.

Boundary Data Structure

The MDF was delivered from the DAS to the CDL and tabulation in the form of pipe-delimited ASCII Text files.

Table 9: Boundary Data Structure Details

Data Structure Unique Identifier	Name	Data Structure Characterization	File Name	File Size (bytes)	Exchange Format
MDF.Person	Microdata Detail File (MDF) Person Data	*Disclosure Avoidance Applied*	*MDF_PER.txt*	19,957,230,224	*Pipe-delimited ASCII*
MDF.Unit	Microdata Detail File (MDF) Unit Data	*Disclosure Avoidance Applied*	*MDF_UNIT.txt*	10,013,138,363	*Pipe-delimited ASCII*

For Boundary Data Structure, see MDF Specs in the Architecture section.

Performance Considerations

The 2010 Demonstration Data Products MDF delivery consisted of two pipe-delimited ASCII text files: MDF_UNIT.txt and MDF_PER.txt. The two text files were approximately 10GB and 20GB in size.

Attempts to transfer the file using the Census Bureau's SOA system failed, so these files were transferred using the Unix "scp" command.

11. ACRONYMS

Table 10: Acronyms

Acronym	Meaning
ACSO	American Community Survey Office
AES	Advanced Encryption Standard
AIAN	American Indian or Alaska Native
AMI	Amazon Machine Image
AWS	Amazon Web Services
CED-DA	Center for Enterprise Dissemination - Disclosure Avoidance
CEF	Census Edited File. CEF processes the CUF, the tabulation geography table, and produces tables of edited data
CDL	Census Data Lake
CIS	Center for Internet Security
CUF	Census Unedited File, consisting of tables that are an input to the CEF
CMK	Customer Master Key
DAS	Disclosure Avoidance System
DCMD	Decennial Census Management Division
DITD	Decennial Information Technology Division
DSEPC	Data Stewardship Executive Policy Committee
DP	Differential Privacy
E2E	End-to-End
EBS	Elastic Block Store
EC2	Elastic Compute Cloud
EMR	Elastic Map Reduce
GQ	Group Quarters Facilities'
HDFS	Hadoop Distributed File System
IPT	Integrated Project Team
KMS	Key Management Service
MDF	Microdata Detail File
MDNH	multi-dimensional national histogram
MFT	Managed File Transfer
NHPI	Native Hawaiian or Other Pacific Islander
NIU	Not In Universe
PLB	Privacy-Loss Budget
POP	Population Division
RPO	Response Processing Operation
TDA	TopDown Algorithm
TEA	Type of Enumeration Area
TI	Technical Integration
TSD	Technical Specification Document
S3	Simple Storage Service
SEHSD	Social, Economic, Housing and Statistics Division
SSE	Server-Side Encryption
SNS	Simple Notification Service
SOR	Some Other Race or Ethnicity

12. APPENDIX

2010 Demonstration Product Privacy-Loss Budget Allocations

The privacy-loss budget (PLB) expended on a tabulation at a target geographic level is not the only source of accuracy on that tabulation, and the expected error in such a tabulation can't be reasonably estimated just by considering the variance of the subtotal PLB expended on it. The reason is that, in addition to using the direct estimate of a tabulation value, the TDA also "borrows strength" from related tabulations at higher geographic levels and uses information in them to improve the accuracy of tabulations at lower geographic levels. However, there is no "closed-form" formula for expressing this borrowing of strength; it must be examined empirically.

A total PLB of 6.0 was used to generate the 2010 Demonstration Data Products, split into subtotals of 4.0 and 2.0 for the person records and housing records, respectively. These subtotal PLBs were then allocated among competing tabulations and geographic levels as follows:

Proportion of Total PLB Assigned to Each Geographic Level

(identical for person & housing records)

Geographic Level	PLB
Nation	0.2
State	0.2
County	0.12
Tract Group	0.12
Tract	0.12
Block Group	0.12
Block	0.12

PLB Expenditure on Tabulations Supporting Person Records

Proportion of Per-Geographic-Level PLB Subtotal Assigned to Sets of Tabulations

Tabulation	**PLB**
Detailed Tabulations (fully saturated contingency table)	0.1
HHGQ (8 numbers per geographic unit)	0.2
Voting Age * Hispanic * CENRACE * Citizen (2 * 2 * 63 * 2 numbers per geographic unit)	0.5
Age * Sex (116 * 2 numbers per geographic unit)	0.05
Bucketed 4-Year Ages * Sex (29 * 2 numbers per geographic unit)	0.05
Bucketed 16-Year Ages * Sex (8 * 2 numbers per geographic unit)	0.05
Bucketed 64-Year Ages * Sex (2 * 2 numbers per geographic unit)	0.05

PLB Expenditure on Tabulations Supporting Housing Records

Proportion of Per-Geographic-Level PLB Subtotal Assigned To Sets of Tabulations

Tabulation	**PLB**
Detailed Tabulations (fully saturated contingency table)	0.2
Hisp * Race * Size * HHType (2 * 7 * 8 * 24 numbers per geographic unit)	0.25
HHSex * Hisp * Race * HHType (2 * 2 * 7 * 24 numbers per geographic unit)	0.25

Hisp * Race * Multi (2 * 7 * 2 numbers per geographic unit)	0.1
HHSex * HHType * Elderly (2 * 24 * 8 numbers per geographic unit)	0.1
HHSex * HHAge * HHType (2 * 9 * 24 numbers per geographic unit)	0.1

Each tabulation over which PLB was expended corresponded to a marginal on a set of variables, where each variable itself was a recode of variable(s) present in the Census Edited File. To understand these variables in more detail, the user should refer to the Python files Schema_DHCP_HHGQ.py (for persons records) and Schema_Household2010.py (for housing records) (as well as their base classes, which are located in other files in the same directory) located in *programs/schema/schemas* in the forthcoming official code release.

A handful of brief comments on specific variables and syntax may aid the reader:

- Asterisks refer to taking all possible combinations of the levels of two (or more) variables.
- A citizenship variable appears in the person record workload because we developed the 2020 DAS in anticipation of the question appearing on the 2020 Census. No actual citizenship data were used to produce the 2010 Demonstration Data Products (the variable was imputed using a crude model for testing purposes). The citizenship variable will be removed from the version of the code base used to generate the production PL94-171 and DHC data products.
- HHGQ is an 8-level variable: the first level indicates a person that resides in a housing unit, while the remaining levels indicate the person resides in one of the major (first-digit) group quarters facility types.
- The CENRACE variable is a 63-level variable, each level of which refers to a non-empty combination of major OMB race categories (see Schema_DHCP_HHGQ.py for clarification).
- The housing workload adopts a convenient short-hand for variable names: Hisp refers to whether the householder is of Hispanic ethnicity (or not), Race to whether the householder reported being a member of one of the six major OMB race categories alone or being of two or more races, and so on. The interested user should consult Schema_Household2010.py, scrolling down to find detailed variable definitions, if there is doubt about the meaning of a variable.
- "HH" in the housing tabulations refers to the householder: for example, householder sex, householder age, and so on.
- "Tract group" is not an official Census Bureau geographic level. It was introduced into the TDA to produce the demonstration data products because adding additional geographic levels helps to control the fan-out (number of child geographic units) per geographic unit, which is a key determinant of both algorithm complexity (memory usage and total run-time) as well as of the privacy loss-accuracy tradeoff.
- "Bucketed" age dimensions refer to nonoverlapping ages. The full list is available in the official code release, but the 64-year-age buckets tabulation provides a simple example:
  - *"Under 64 years" : list(range(0, 64))*
  - *"64 to 115 years" : list(range(64, 116))*
- Note that 115 was the maximum age allowed according to 2010 Decennial Census edit specifications. Values of ages outside the set {0,1,…,115} were set to null values and replaced using statistical imputation.
- The choice of tabulations on which to expend PLB, and in what amount, was motivated by on-going empirical experimentation internally at the Census Bureau and discussion/iteration with internal stakeholders. In reviewing the assigned PLBs, it is notable that typically more PLB was assigned to tabulations that contain more individual values, with the exception of the "detailed" tabulation, which is present both to ensure that the true data will be recovered if the PLB is

increased toward infinity, and to capture relationships between variables not otherwise directly measured.

2010 Demonstration Product Configuration

The configuration files used during the production of the 2010 Demonstration Data Products is included below. The `[geodict]` section specifies the geographical levels to be included in the execution of the DAS. The `[constraints]` section specifies the invariants and constraints to be used during the DAS execution as well as the geographical levels at which to apply each invariant and constraint. The `[budgets]` section specifies the privacy loss budget as well as how the budget is to be applied to the geolevels and histograms during the DAS execution.

```
=========================Configuration File for DHCH=========================

# Main file for National DHCH runs with manual topdown

[DEFAULT]
# root specifies the root location for all files; testdir specifies ???;
# mode specifies ???
# For the demo, the root in the current directory
include=../default.ini

[logging]
logfilename: DAS
loglevel: INFO
logfolder: logs

[ENVIRONMENT]
DAS_FRAMEWORK_VERSION: 0.0.1
GRB_ISV_NAME: Census
GRB_APP_NAME: DAS
GRB_Env3: 0
GRB_Env4:

[python]

[gurobi]
# Pick a gurobi version!  Note that $PYTHON_VERSION is automatically set by the DAS runtime.
gurobi_path=/mnt/apps5/gurobi752/linux64/lib/${PYTHON_VERSION}_utf32/
#gurobi_path=/mnt/apps5/gurobi810/linux64/lib/${PYTHON_VERSION}_utf32/


# Record the stats, which has them sent by syslog to the MASTER node
record_gurobi_stats: True
record_CPU_stats: True
record_VM_stats: True

# Do not save to S3, which takes a lot of time
save_stats: False
print_gurobi_stats: False


[geodict]
#smallest to largest (no spaces)
geolevel_names: Block,Block_Group,Tract,Tract_Group,County,State,National
#(largest geocode length to smallest, put 1 for national level) (no spaces)
geolevel_leng: 16,12,11,9,5,2,1

[setup]
setup: programs.das_setup.setup

# Spark config stuff
```

```
spark.name: DAS_NAT_DEMONSTRATION_PRODUCT
#local[6] tells spark to run locally with 6 threads
#spark.master: local[9]
#Error , only writes to log if there is an error (INFO, DEBUG, ERROR)
spark.loglevel: ERROR

[reader]
INCLUDE=Reader/unit.ini
Household.path: $DAS_S3ROOT/title13_input_data/table12a_20190705/
Unit.path: $DAS_S3ROOT/title13_input_data/table10_20190610/


comment: ?
numReaderPartitions: 5000
readerPartitionLen: 12
validate_input_data_constraints: False

[engine]
engine: programs.engine.topdown_engine.TopdownEngine

# should we delete the true data after making DP measurments (1 for True or 0 for False)
delete_raw: 1
save_noisy: 0
reload_noisy: 0
check_budget: off

[schema]
schema: Household2010

[budget]
epsilon_budget_total: 2
bounded_dp_multiplier: 2

#budget in topdown order (e.g., National, State, .... , Block)
geolevel_budget_prop: 0.2,0.2,0.12,0.12,0.12,.12,0.12

# detailed query proportion of budget (a float between 0 and 1)
detailedprop: .2

# start with no queries
DPqueries: hisp * hhrace * size * hhtype, hhsex * hisp * hhrace * hhtype, hisp *
             hhrace * multi, hhsex * hhtype * elderly, hhsex * hhage * hhtype
queriesprop: .25, .25, .1, .1, .1

[constraints]
#the invariants created, (no spaces)
theInvariants.Block: tot,gqhh_vect

#these are the info to build cenquery.constraint objects
theConstraints.Block:
             total,no_vacant,living_alone,size2,size3,size4,size2plus_notalone,not_mul
             tigen,hh_elderly,age_child

[writer]
#writer: programs.writer.mdf2020writer.MDF2020HouseholdWriter
writer: programs.writer.pickled_block_data_writer.PickledBlockDataWriter
keep_attrs: geocode, syn, _invar
# Where the data gets written:
output_path: s3://uscb-decennial-ite-das/DHC_DemonstrationProduct_Fixed/DHCH/
num_parts: 30000
# Save the output:
produce_flag: 1

# delete existing file (if one) 0 or 1
```

```
overwrite_flag: 1

# upload the logfile to the dashboard:
upload_logfile: 1

classification_level: CUI//CENS
output_datafile_name: MDF_UNIT
write_metadata: 0
s3cat: 0
s3cat_suffix: .txt
s3cat_verbose: 0

[validator]
validator: programs.stub_validator.validator
results_fname: /mnt/tmp/WNS_results

[assessment]

[takedown]
takedown: programs.takedown.takedown
delete_output: 0

[experiment]
experiment: programs.experiment.experiment.experiment
run_experiment_flag: 0

[error_metrics]
error_metrics: programs.metrics.error_metrics_stub.ErrorMetricsStub

=======================================================================================

==========================Configuration File for DHCP==========================
[default]
include=../default.ini

[constraints]
theinvariants.block = gqhh_vect, gqhh_tot
theinvariants.state = tot
theconstraints.block = hhgq_total_lb, hhgq_total_ub, hhgq1_lessthan15,
             hhgq2_greaterthan25, hhgq3_lessthan20, hhgq5_lt16gt65, hhgq6_lt17gt65
theconstraints.state = total, hhgq_total_lb, hhgq_total_ub
minimalschema = hhgq

[validator]
validator = programs.stub_validator.validator
results_fname = /mnt/tmp/lecle_results

[error_metrics]
error_metrics = programs.metrics.error_metrics_stub.ErrorMetricsStub

[writer]
writer = programs.writer.pickled_block_data_writer.PickledBlockDataWriter
keep_attrs = geocode, syn
produce_flag = 1
overwrite_flag = 1
num_parts = 30000
stats_dir = $DAS_S3ROOT/rpc/upload
include = Writer/default.ini
output_path = s3://uscb-decennial-ite-
             das/users/lecle301/DemonstrationProducts_Sept2019_fixedTotals/full_person/
output_datafile_name = persons
write_metadata = 0
s3cat = 0
```

```
s3cat_suffix = .txt
s3cat_verbose = 1

[takedown]
takedown = programs.takedown.takedown
delete_output = 0

[geodict]
geolevel_names = Block,Block_Group,Tract,Tract_Group,County,State,Nation
geolevel_leng = 16,12,11,9,5,2,1

[engine]
engine = programs.engine.topdown_engine.TopdownEngine
delete_raw = 0
save_noisy = 0
reload_noisy = 0
check_budget = off

[logging]
logfilename = DAS
loglevel = INFO
logfolder = logs

[reader]
comment = ?
reader = programs.reader.table_reader.DASDecennialReader
tables = Person Unit
privacy_table = Person
constraint_tables = Unit
person.class = programs.reader.sql_spar_table.SQLSparseHistogramTable
unit.class = programs.reader.sql_spar_table.SQLSparseHistogramTable
person.path = s3://uscb-decennial-ite-das/title13_input_data/table1a_20190709/
unit.path = s3://uscb-decennial-ite-das/title13_input_data/table10/
delimiter = \t
header = True
person.variables = MAFID age geocode white black aian asian nhopi other hispanic sex
             citizen relation
unit.variables = MAFID geocode gqtype
linkage = geocode
geocode.type = str
geocode.legal = 0000000000000000-9999999999999999
mafid.type = str
mafid.legal = 000000000-999999999
sex.type = int
sex.legal = 0,1
age.type = int
age.legal = 0-115
hispanic.type = int
hispanic.legal = 0,1
white.type = int
white.legal = 0,1
black.type = int
black.legal = 0,1
aian.type = int
aian.legal = 0,1
asian.type = int
asian.legal = 0,1
nhopi.type = int
nhopi.legal = 0,1
other.type = int
other.legal = 0,1
citizen.type = int
citizen.legal = 0,1
relation.type = int
relation.legal = 0-42
ten.type = int
ten.legal = 0-3
gqtype.type = str
gqtype.legal = 000-999
vacs.type = int
vacs.legal = 0-7
person.recoder = programs.reader.e2e_recoder.DHCP_HHGQ_recoder
person.recode_variables = cenrace hhgq
cenrace = white black aian asian nhopi other
hhgq = relation
cenrace.type = int
cenrace.legal = 0-62
hhgq.type = int
hhgq.legal = 0-7
unit.recoder = programs.reader.hh_recoder.Table10RecoderSimple
unit.recode_variables = hhgqinv
hhgqinv = gqtype
hhgqinv.type = int
hhgqinv.legal = 0-7
person.geography = geocode
person.histogram = hhgq sex age hispanic cenrace citizen
numreaderpartitions = 5000
readerpartitionlen = 12
unit.geography = geocode
unit.histogram = hhgqinv
include = Reader/unit_simple.ini
validate_input_data_constraints = False

[setup]
setup = programs.das_setup.setup
spark.name = DAS_DEMO_PRODUCT_DHCP
spark.loglevel = ERROR

[assessment]


[gurobi]
include = default.ini
gurobi_lic = /apps/gurobi752/gurobi_client.lic
gurobi_logfile_name = gurobi.log
outputflag = 1
optimalitytol = 1e-9
barconvtol = 1e-8
barqcpconvtol = 0
bariterlimit = 1000
feasibilitytol = 1e-9
threads = 60
presolve = -1
numericfocus = 3
python_presolve = 1
record_gurobi_stats = False
record_cpu_stats = False
record_vm_stats = False
save_stats = False
print_gurobi_stats = False

[budget]
epsilon_budget_total = 4.0
# Bounded DP multiplier, aka "elementary" sensitivity used for each individual DP
             query
bounded_dp_multiplier = 2.0

geolevel_budget_prop = 0.2, 0.2, 0.12, 0.12, 0.12, 0.12, 0.12
```

```
detailedprop = 0.1
dpqueries = hhgq, votingage * hispanic * cenrace * citizen, age * sex, ageGroups4 *
            sex, ageGroups16 * sex, ageGroups64 * sex
queriesprop = .2, .5, 0.05, 0.05, 0.05, 0.05

[schema]
schema = DHCP_HHGQ

[ENVIRONMENT]
das_framework_version = 0.0.1
grb_isv_name = Census
grb_app_name = DAS
grb_env3 = 0
grb_env4 =
```

===

NOTE: A citizenship variable appears in the person record workload because we developed the 2020 DAS in anticipation of the question appearing on the 2020 Census. No actual citizenship data were used to produce the 2010 Demonstration Data Products (the variable was imputed using a crude model for testing purposes). The citizenship variable will be removed from the version of the code base used to generate the production PL94-171 and DHC data products.