PHSafe v3.0.0 Documentation

Executive Summary

Tumult Labs is delivering a production-ready implementation of PHSafe, a differentially private algorithm, that produces population in households tables for the Supplemental Demographic and Housing Characteristics File (S-DHC) at a fixed number of geography levels. These tables require a join or merge of the person and unit file.

Goals

1. **Produce tables of statistics (PH1, PH2, PH3, PH4, PH5, PH6, PH7, PH8)**.
2. **Satisfy differential privacy:** the algorithm used to produce the tables satisfies zero-concentrated differential privacy and also supports pure differential privacy.
3. **Low error:** the algorithm should allow users to tune input parameters to improve the error in statistics.

Problem Specification

Demographic Statistics: The Census Bureau would like to release the following statistics for each population group (geographic entity and race iteration pair) as part of the S-DHC:

- PH1: Average Household Size by Age
- PH2: Household Type for the Population in Households
- PH3: Household Type by Relationship for the Population Under 18 Years
- PH4: Population in Families by Age
- PH5: Average Family Size by Age
- PH6: Family Type and Age for Own Children Under 18 Years
- PH7: Total Population in Occupied Housing Units by Tenure
- PH8: Average Household Size of Occupied Housing Units by Tenure

Race Characteristic Iterations:
PH1, PH3, PH4, PH5, PH7, and PH8 are iterated by the following major race and ethnicity categories:

- (A) White alone
- (B) Black or African American Alone
- (C) American Indian and Alaska Native alone
- (D) Asian alone
- (E) Native Hawaiian and Other Pacific Islander alone
- (F) Some Other Race alone
- (G) Two or More Races
- (H) Hispanic or Latino

- (I) White alone, not Hispanic or Latino

All tables are also tabulated for the (*) or unattributed category, which counts all races and ethnicities. These iterations are grouped into three “levels”:

- *
- A-G
- H, I

The PHSafe algorithm produces a single table for each set of output statistics (PH2, PH3, etc.), except for averages for which it produces numerator and denominator tables (PH1_num, PH1_denom, etc.). At publication, numerator and denominator tables are put through a modeling algorithm to produce the final averages, and tables will be split into one sub-table for each iteration (PH1A, PH1B, etc.).

Geographies: PHSafe produces statistics for a fixed set of geographic levels:

- United States
- States

Consistency Requirements:

- All count estimates are integral.
-

Selected list of potential data inconsistencies in the PHSafe outputs:

- Counts may be negative.
- Counts may not “add up”.
  - A national count may not match the sum of its corresponding state counts.
  - An unattributed count may not match the sum of its corresponding A-G race iteration counts.
- Equivalent table cells may not be consistent across tables. For example, the PH1 numerator cell for “population under 18 years old” may not match the aggregated PH3 total cell for population under 18 years old.
- Numerators and denominators may not be consistent. For example, a positive count numerator corresponding to a zero or negative count denominator.

Approach

Privacy loss is measured with respect to persons rather than households. That is, the privacy-loss parameter (epsilon / rho) quantifies the privacy risks to individuals. A collective household’s privacy risk is measured via the “group privacy” guarantee of differential privacy / zero-concentrated differential privacy. We do not consider the alternative privacy framework that assumes the household is the basic unit of privacy. Privacy loss in PHSafe is defined with the “unbounded” add/remove notion of neighboring databases. The corresponding “bounded” neighbors privacy-loss values may be obtained by multiplying the “unbounded” values by 2.

For PH2, PH3, PH4, PH6, and PH7, PHSafe directly computes estimates of the most detailed cells in the tables (roughly the deepest indentation level of each table). For PH1, PH5, and PH8, PHSafe does not directly compute averages. PHSafe estimates the numerator for PH1 of the most detailed cells in the table. For PH5 and PH8, the numerators are created from PH4 and PH7 respectively by postprocessing. For all three average tables, estimates are produced for the denominators.

The algorithm will do the following for each population group for PH1 numerator, PH2, PH3, PH4, PH6, and PH7:

1. Filter the data by restricting to records in the table's universe.
2. Join the person and unit dataframes and truncate by dropping persons so that households do not exceed a threshold (given as an input parameter for each table)
3. Generate noisy counts by adding noise drawn from a two-sided geometric distribution or discrete Gaussian distribution to the true counts for each of the most detailed table cells. (The code supports both "puredp" and "zcdp" modes).

For PH1 denominator, PH5 denominator, and PH8 denominator, step 2 above is omitted (counts are on the unit dataframe alone) but the other two steps are repeated for each population group.

A privacy budget can be assigned to each population group level (iteration level/geo level) per tabulation. If certain population groups have zero budget, then they are not tabulated. If all population groups for a table have zero budget, then the entire tabulation is skipped.

Performance

We recommend running PHSafe on an EMR cluster with 1 master node and 2 executor nodes of the r4.16xlarge AWS instance type, which comes with 64 vCPUs (2.3 GHz Intel Xeon E5-2686 v4 Processor) and 488 G memory. Our recommended spark settings are specified in `resources/spark_configs/spark_cluster_properties.conf`. We ran PHSafe in "private" mode with input and output validation enabled, using Tumult generated simulated dataframes with 300 million person records and 150 million unit records, using a Rho zCDP privacy definition. The run completed successfully within 1 hour and 20 minutes.

System Requirements

See PHSafe README.

Testing Plan

See PHSafe TESTPLAN.

Documentation

README - includes information on the relevant packages within the PHSafe Repository. It also includes instructions for downloading sample data.
phsafe/README.md - includes installation instructions, hardware and software requirements, instructions for use, and known warnings.
phsafe/LICENSE - software license under which PHSafe is distributed.
phsafe/TESTPLAN.md - includes instructions to test if PHSafe and the required Tumult packages are installed correctly.

Input/Output Specification

See Appendix A

Appendix A: File Specifications

This appendix provides details on the formats for the input and output to be used in the 2020 Census Disclosure Avoidance System (DAS) activities supported by Tumult Labs. Input Dataframes refers to python spark dataframe objects created by Census DAS reader programs or by reading synthetic data in csv file format. Output Files refers to files produced by PHSafe intended for further use by the Census Bureau.

A note of notation:

DataType	Description
StringType(n)	A string with up to n characters
StringType	A string without a character limit
IntegerType(n)	A number with up to n digits
IntegerType	A number without a digit limit
FloatType	A decimal valued number with no limit on digits.

Input Dataframes

person_df

Representation of custom person records derived from the Census Edited File (CEF) Person file that is input to DAS. We assume that *person_df* will contain exactly one row for each person in the United States and Puerto Rico.

Version and Date

2022-12-02.v1.0.0

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
RTYPE	Record Type	StringType(1)	3 = Person in Housing Unit 5 = Person in GQ
MAFID	Foreign key to Unit Table. Master Address File ID.	IntegerType(9)	100000001-899999999
QAGE	Edited Age	IntegerType(3)	0-115
CENHISP	A recode of the edited Hispanic origin variable	IntegerType(1)	1 = Not Hispanic 2 = Hispanic

	(QSPAN) into two values representing Hispanic and not Hispanic		
CENRACE	A recode of edited race codes (QRACE1-QRACE8) into a single 2-digit code representing all of the possible race categories	StringType(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 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
RELSHIP	Final Edited Relationship to householder	StringType(2)	20 = Householder 21 = Opposite-sex husband/wife/spouse 22 = Opposite-sex unmarried partner 23 = Same-sex husband/wife/spouse 24 = Same-sex unmarried partner 25 = Biological son or daughter 26 = Adopted son or daughter 27 = Stepson or stepdaughter 28 = Brother or sister 29 = Father or mother 30 = Grandchild 31 = Parent-in-law 32 = Son-in-law or daughter-in-law 33 = Other relative 34 = Roommate or housemate 35 = Foster child

			36 = Other nonrelative 37 = Institutional GQ Person 38 = Non-institutional GQ Person

Encoding

UTF-8

Sample Records

```
RTYPE|MAFID|QAGE|CENHISP|CENRACE|RELSHIP

3|100000001|24|2|08|20
```

unit_df

Representation of custom unit records derived from the CEF Unit file that is input to DAS. We assume that *unit_df* will contain exactly one row for each unit (housing unit or group quarters (GQ)) in the United States and Puerto Rico.

Version and Date

2022-12-02.v1.0.0

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
RTYPE	Record Type	StringType(1)	2 = Housing Unit 4 = GQ
MAFID	Primary key. Master Address File ID.	IntegerType(9)	100000001-899999999
FINAL_POP	Final Population Count	IntegerType(5)	0-99999
NPF	Number of people in families	IntegerType(2) (Same as NPF on Review CEF20_UNIT)	0, 2-97
HHSPAN	Hispanic householder	IntegerType(1) (Same as HHSPAN in CEF20_UNIT, except 0 instead of whitespace for not in universe NIU)	0 = GQ or vacant 1 = Not Hispanic 2 = Hispanic
HHRACE	Edited CENRACE of Householder	StringType(2) (Same as HHRACE in CEF20_UNIT, except 00 instead of whitespace for NIU)	00 = GQ or vacant 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 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
TEN	Edited Tenure	StringType(1)	0 = Not in Universe (Vacant or GQ) 1 = Owned with a mortgage 2 = Owned free and clear 3 = Rented 4 = Occupied without payment of rent

HHT	Household/Family Type	StringType(1)	0 = NIU (GQ or Vacant Housing Unit) 1 = Married couple household 2 = Other family household: Male householder 3 = Other family household: Female householder 4 = Nonfamily household: Male householder, living alone 5 = Nonfamily household: Male householder, not living alone 6 = Nonfamily household: Female householder, living alone 7 = Nonfamily household: Female householder, not living alone
HHT2	Household/Family Type (Includes Cohabiting)	StringType(2)	00 = NIU (GQ or Vacant Housing Unit) 01 = Married couple household: With own children <18 02 = Married couple household: Without own children <18 03 = Cohabiting couple household: With own children < 18 04 = Cohabiting couple household: Without own children < 18 05 = Female householder, no spouse/partner present: Living alone 06 = Female householder, no spouse/partner present: With own children < 18 07 = Female householder, no spouse/partner present: With relatives, without own children <18 08 = Female householder, no spouse/partner present: Only nonrelatives present 09 = Male householder, no spouse/partner present: Living alone 10 = Male householder, no spouse/partner present: With own children < 18 11 = Male householder, no spouse/partner present: With relatives, without own children <18 12 = Male householder, no spouse/partner present: Only nonrelatives present
CPLT	Couple Type	StringType(1)	0 = NIU (GQ or Vacant Housing Unit)

			1 = Opposite-sex husband/wife/spouse household 2 = Same-sex husband/wife/spouse household 3 = Opposite-sex unmarried partner household 4 = Same-sex unmarried partner household 5 = All other households

Encoding

UTF-8

Sample Records

RTYPE|MAFID|FINAL_POP|NPF|HHSPAN|HHRACE|TEN|HHT|HHT2|CPLT

2|100000001|5|2|1|01|1|1|01|1

geo_df

Representation of custom geography lookup table derived from the Census Edited File (CEF) Unit file and the Geography Reference File Code (GRFC) that is input to DAS. We assume that *geo_df* will contain exactly one row for each unit (housing unit or GQ) in the United States and Puerto Rico.

Version and Date

2022-12-02.v1.0.0

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
RTYPE	Record Type	StringType(1)	2 = Housing Unit 4 = GQ
MAFID	Foreign key to Unit Table. Master Address File ID.	IntegerType(9)	100000001-899999999
TABBLKST	State code	StringType(2)	01–02, 04–06, 08–13, 15–42, 44–51, 53–56, 60, 66, 69, 72, 78
TABBLKCOU	County Code	StringType(3)	001–840
TABTRACTCE	Census Tract Code	StringType(6)	000100–998999
TABBLK	Block Code	StringType(4)	0001–9999
TABBLKGRPCE	Block Group Code	StringType(1)	0-9
REGIONCE	Region	StringType(1)	1-4, 9
DIVISIONCE	Division	StringType(1)	0-9
PLACEFP	Place	StringType(5)	00001-89999, 99999
AIANNHCE	AIANNH (Census)	StringType(4)	0001–9998; 9999

Encoding

UTF-8

Sample Records

RTYPE|MAFID|TABBLKST|TABBLKCOU|TABTRACTCE|TABBLK|AIANNHCE

2|100000001|08|500|000300|0060|0001

Additional Inputs

config.json

Description

json file encoding inputs as key, value pairs.

We are not aware of a standard for rounding privacy loss budget values but recommend rounding to 3-4 significant digits (not including zeros). For example, 0.12345 becomes 0.1235 whereas 0.0012345 becomes 0.001235.

Version and Date

2023-09-01.v2.0.1

Key Value Names and Format Definitions

Key	Description	Value Format	Legal Values

privacy_budget	The privacy loss budget assigned to geo level <geo> and characteristic iteration level <iteration_level> per tabulation. If all <geo>_<iteration_level> budget is set to 0, then that entire tabulation is skipped.
	{
	table : {
	<geo>_<iteration_level>: (for geo in {usa, state} and iteration_level in {"A-G", "H,I", "*"})
	}
	(for table in {PH1_num, PH1_denom, PH2, PH3, PH4, PH5_denom, PH6, PH7, PH8_denom}
	}
		Map[string, Map[string, Float]]	See Sample Records below

tau	Truncation threshold to limit the max persons per household. PH1_num, PH2, PH3, PH4, PH6, PH7 are the keys to specify threshold per tabulation.	Map[string, Int]	See Sample Records below

state_filter	A list of 2-digit FIPS codes for the 50 continental states + DC to include in the US run OR [“72”] for the PR run. Only records that correspond to blocks in included states will be tabulated. Use values from TABBLKST in GRF-C.txt.	List[string]	e.g., ["37", "45"] to include North and South Carolina.
reader	Key that indicates the reader being used. csv: Tumult’s CSV reader cef: MITRE’s CEF reader	string	{“csv”, “cef”}
privacy_defn	The privacy definition being used, either Pure DP (“puredp”) or Rho zCDP (“zcdp”). Determines how privacy budgets are interpreted.	string	{“puredp” , “zcdp”}

Encoding

UTF-8

Sample Records

```
  "privacy_budget": {
          "PH1_num": {
                "usa_A-G": 0.33,
                "usa_H,I": 0.33,
                "usa_*": 0.33,
                "state_A-G": 0.33,
                "state_H,I": 0.33,
                "state_*": 0.33,
        },
        "PH1_denom": {
                "usa_A-G": 0.33,
                "usa_H,I": 0.33,
                "usa_*": 0.33,
                "state_A-G": 0.33,
                "state_H,I": 0.33,
                "state_*": 0.33,
        },
        "PH2": {
                "usa_*": 1.0,
                "state_*": 1.0,
        },
        "PH3":  {
                "usa_A-G": 0.5,
                "usa_H,I": 0.5,
                "usa_*": 0.5,
                "state_A-G": 0.5,
                "state_H,I": 0.5,
                "state_*": 0.5,
        },
        "PH4":  {
                "usa_A-G": 0.5,
                "usa_H,I": 0.5,
                "usa_*": 0.5,
                "state_A-G": 0.5,
                "state_H,I": 0.5,
                "state_*": 0.5,
        },
        "PH5_denom": {
                "usa_A-G": 0.5,
                "usa_H,I": 0.5,
                "usa_*": 0.5,
                "state_A-G": 0.5,
                "state_H,I": 0.5,
                "state_*": 0.5,
        },
        "PH6": {
                "usa_*": 1.0,
                "state_*": 1.0,
        },
        "PH7": {
                "usa_A-G": 0.5,
                "usa_H,I": 0.5,
                "usa_*": 0.5,
                "state_A-G": 0.5,
                "state_H,I": 0.5,
                "state_*": 0.5,
        },
        "PH8_denom": {
                "usa_A-G": 0.5,
                "usa_H,I": 0.5,
                "usa_*": 0.5,
                "state_A-G": 0.5,
                "state_H,I": 0.5,
                "state_*": 0.5,
        }
  },
  "tau": {
          "PH1_num": 5,
          "PH2": 10,
          "PH3": 5,
          "PH4": 5,
          "PH6": 5,
          "PH7": 5
  },
  "state_filter": ["08", "04", "35", "49"],
  "reader": "cef",
  "privacy_defn": "zcdp"
}
```

reader_config

Description

The CEF Reader program reads the setup input parameters from an .ini file. The INI configuration file consists of sections, each led by a [section] header, followed by key/value entries separated by “= “ string. The following is a sample of CEF Reader configuration file.

```
[paths]
cef_year = 2010
per_dir = <US Person CEF Path>
unit_dir = <US Units CEF Path>
per_dir_pr = <PR Persons CEF Path>
unit_dir_pr = <PR Units CEF Path>
grfc_dir = <GRFC Path>
per_file_format = CEF20_PER_%%s.txt
unit_file_format = CEF20_UNIT_%%s.txt
geo_file_format = grfc_tab20_%%s.txt
```

Output Files

The output from each tabulation is saved in its own directory (including separate directories for numerators and denominators), formatted as a pipe-delimited csv file. Tabulations which were assigned zero privacy loss budget in the configuration are not tabulated, and no output is written for them. The output directory names are listed below. The output directories contain estimates for all the geographies and race iterations at which the given table is being published.

For a detailed explanation of how each table is defined, including which values are in- and out-of-universe for each table, refer to the Census Bureau's technical documentation for S-DHC.

PH1_denom

Description

Number of units by region and iteration (of the householder). Each row contains a specification of attributes and the count for those attributes. When the PHSafe algorithm is run, the PH1_denom/ folder will contain exactly one csv file with the prefix part-00000-XXXXX.

Version and Date

2023-09-01.v2.0.1

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	Geocode corresponding to one of USA or STATE.	StringType(2)	1, 06, 44, etc
REGION_TYPE	Name of geography level	StringType(5)	{USA, STATE}

ITERATION_CODE	Characteristic iteration code	StringType(1)	* = Unattributed
			A = White alone
			B = Black or African American alone
			C = American Indian and Alaska Native alone
			D = Asian alone
			E = Native Hawaiian and Other Pacific Islander alone
			F = Some Other Race alone
			G = Two or More Races
			H = Hispanic or Latino

			I = White alone, not Hispanic or Latino
PH1_DENOM_DATA_CELL	Follows PH1 table shell data cell values. The PH1 denominator only includes the total cell.	IntegerType(1)	1 = Total
COUNT	Number of units corresponding to the given attributes.	IntegerType	12, -3, 670, etc
NOISE_DISTRIBUTION	"Discrete Gaussian" if the privacy definition used is "zcdp". "Two-Sided Geometric"if the privacy definition used is "puredp".	StringType	{"Discrete Gaussian", "Two-Sided Geometric"}
VARIANCE	Measure of dispersion	FloatType	31.67, 412.889, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

REGION_ID|REGION_TYPE|ITERATION_CODE|PH1_DENOM_DATA_CELL|COUNT|NOISE_DISTRIBUTION|VARIANCE

1|USA|*|1|12345|Discrete Gaussian|200.78

01|STATE|A|2|13245|Discrete Gaussian|300.14

PH1_num

Description

Contains counts of people living in households by age (over/under 18), region, and iteration (of the householder). Each row contains a specification of attributes, and the count for those attributes. When the PHSafe algorithm is run, the PH1_num/ folder will contain exactly one csv file with the prefix part-00000-XXXXX.

Version and Date

2023-09-01.v2.0.1

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	Geocode corresponding to one of USA or STATE.	StringType(2)	1, 06, 44, etc
REGION_TYPE	Name of geography level	StringType(5)	{USA, STATE}

ITERATION_CODE	Characteristic iteration code	StringType(1)	* = Unattributed
			A = White alone
			B = Black or African American alone
			C = American Indian and Alaska Native alone
			D = Asian alone
			E = Native Hawaiian and Other Pacific Islander alone
			F = Some Other Race alone
			G = Two or More Races
			H = Hispanic or Latino
			I = White alone, not Hispanic or Latino

PH1_NUM_DATA_CELL	Follows PH1 table shell data cell values. The PH1 numerator only includes the interior cells.	IntegerType(1)	2 = under 18 years 3 = 18 years and over
COUNT	Number of people in households corresponding to the given attributes	IntegerType	12, -3, 670, etc
NOISE_DISTRIBUTION	"Discrete Gaussian" if the privacy definition used is "zcdp".	StringType	{"Discrete Gaussian", "Two-Sided Geometric"}

	“Two-Sided Geometric”if the privacy definition used is “puredp”.		
VARIANCE	Measure of dispersion	FloatType	31.67, 412.889, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

REGION_ID|REGION_TYPE|ITERATION_CODE|PH1_NUM_DATA_CELL |COUNT|NOISE_DISTRIBUTION|

VARIANCE

1|USA|A|1|345|Discrete Gaussian|100.11

01|STATE|*|2|245|Discrete Gaussian|200.22

PH2

Description

Contains the number of people in households by household/couple type and region. Each row contains a specification of attributes, and the count for those attributes. When the PHSafe algorithm is run, the PH2/ folder will contain exactly one csv file with the prefix part-00000-XXXXX.

Version and Date

2023-09-01.v2.0.1

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	Geocode corresponding to one of USA or STATE.	StringType(2)	1, 06, 44, etc
REGION_TYPE	Name of geography level	StringType(5)	{USA, STATE}
PH2_DATA_CELL	Data Cell number from PH2 table shell. The output only includes the interior cells.	IntegerType(2)	3 = Opposite-sex married couple 4 = Same-sex married couple 6 = Opposite-sex cohabiting couple 7 = Same-sex cohabiting couple 9 = Male householder, no spouse or partner present: Living alone 10 = Male householder, no spouse or partner present: Living with others 12 = Female householder, no spouse or partner present: Living alone 13 = Female householder, no spouse or partner present: Living with others
COUNT	Number of people corresponding to the given attributes.	IntegerType	12, -3, 670, etc
NOISE_DISTRIBUTION	“Discrete Gaussian” if the privacy definition used is “zcdp”. “Two-Sided Geometric”if the	{“Discrete Gaussian”, “Two-Sided Geometric”}	{“Discrete Gaussian”, “Two-Sided Geometric”}

	privacy definition used is “puredp”.		
VARIANCE	Measure of dispersion	FloatType	31.67, 412.889, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

REGION_ID|REGION_TYPE|PH2_DATA_CELL|COUNT|NOISE_DISTRIBUTION|VARIANCE

1|USA|1|12345|Discrete Gaussian|100.111

01|STATE|2|13245|Discrete Gaussian|200.222

PH3

Description

Contains the number of people in households by relationship, region, and iteration (of the household member) for population under 18 years of age. Each row contains a specification of attributes, and the count for those attributes. When the PHSafe algorithm is run, the PH3/ folder will contain exactly one csv file with the prefix part-00000-XXXXX.

Version and Date

2023-09-01.v2.0.1

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	Geocode corresponding to one of USA or STATE.	StringType(2)	1, 06, 44, etc
REGION_TYPE	Name of geography level	StringType(5)	{USA, STATE}

ITERATION_CODE	Characteristic iteration code	StringType(1)	* = Unattributed
			A = White alone
			B = Black or African American alone
			C = American Indian and Alaska Native alone
			D = Asian alone
			E = Native Hawaiian and Other Pacific Islander alone
			F = Some Other Race alone
			G = Two or More Races
			H = Hispanic or Latino
			I = White alone, not Hispanic or Latino

PH3_DATA_CELL	Data cell number from PH3 table shell. The output only includes the interior cells.	IntegerType(2)	2 = Householder, spouse, unmarried partner, or nonrelative 4 = Own child: In married couple family 5 = Own child: In cohabiting couple family 6 = Own child: In male householder, no spouse or partner present family

			7 = Own child: In female householder, no spouse or partner present family 9 = Grandchild 10 = Other relatives
COUNT	Number of people corresponding to the given attributes.	IntegerType	12, -3, 670, etc
NOISE_DISTRIBUTION	“Discrete Gaussian” if the privacy definition used is “zcdp”. “Two-Sided Geometric”if the privacy definition used is “puredp”.	StringType	{“Discrete Gaussian”, “Two-Sided Geometric”}
VARIANCE	Measure of dispersion	FloatType	31.67, 412.889, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

```
REGION_ID|REGION_TYPE|ITERATION_CODE|PH3_DATA_CELL|COUNT|NOISE_DISTRIBUTION|VARIANC
E

1|USA|B|2|12345|Discrete Gaussian|501

01|STATE|I|10|13245|Discrete Gaussian|11.38
```

PH4

Description

Contains the count of people in families by age (over/under 18), region, and iteration (of the householder). Each row contains a specification of attributes, and the count for those attributes. When the PHSafe algorithm is run, the PH4/ folder will contain exactly one csv file with the prefix part-00000-XXXXX.

Version and Date

2023-09-01.v2.0.1

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	Geocode corresponding to one of USA or STATE.	StringType(2)	1, 06, 44, etc
REGION_TYPE	Name of geography level	StringType(5)	{USA, STATE}

ITERATION_CODE	Characteristic iteration code.	StringType(1)	* = Unattributed
			A = White alone
			B = Black or African American alone
			C = American Indian and Alaska Native alone
			D = Asian alone
			E = Native Hawaiian and Other Pacific Islander alone
			F = Some Other Race alone
			G = Two or More Races
			H = Hispanic or Latino
			I = White alone, not Hispanic or Latino

PH4_DATA_CELL	Data cell number from PH4 table shell. The output only includes the interior cells.	IntegerType(1)	2 = under 18 years 3 = 18 years and over
COUNT	Number of people corresponding to the given attributes.	IntegerType	12, -3, 670, etc
NOISE_DISTRIBUTION	“Discrete Gaussian” if the privacy definition used is “zcdp”. “Two-Sided Geometric”if the	StringType	{“Discrete Gaussian”, “Two-Sided Geometric”}

	privacy definition used is “puredp”.		
VARIANCE	Measure of dispersion	FloatType	31.67, 412.889, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

REGION_ID|REGION_TYPE|ITERATION_CODE|PH4_DATA_CELL|COUNT|NOISE_DISTRIBUTION|VARIANCE

1|USA|C|2|12345|Discrete Gaussian|3.1415

01|STATE|*|3|13245|Discrete Gaussian|2.78

PH5_num

Description

Contains number of people in families by age (over/under 18), region, and iteration (of the householder). Each row contains a specification of attributes, and the count for those attributes. Present only if at least one PH4 geo/race iteration is assigned a positive budget. When the PHSafe algorithm is run, the PH5_num/ folder will contain exactly one csv file with the prefix part-00000-XXXXX.

Version and Date

2023-09-01.v2.0.1

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	Geocode corresponding to one of USA or STATE.	StringType(2)	1, 06, 44, etc
REGION_TYPE	Name of geography level	StringType(5)	{USA, STATE}

ITERATION_CODE	Characteristic iteration code	StringType(1)	* = Unattributed
			A = White alone
			B = Black or African American alone
			C = American Indian and Alaska Native alone
			D = Asian alone
			E = Native Hawaiian and Other Pacific Islander alone
			F = Some Other Race alone
			G = Two or More Races
			H = Hispanic or Latino
			I = White alone, not Hispanic or Latino

PH5_NUM_DATA_CELL	Follows PH5 table shell data cell values. The PH5 numerator only includes the interior cells.	IntegerType(1)	2 = under 18 years 3 = 18 years and over
COUNT	Number of people in families corresponding to the given attributes.	IntegerType	12, -3, 670, etc

NOISE_DISTRIBUTION	“Discrete Gaussian” if the privacy definition used is “zcdp”. “Two-Sided Geometric”if the privacy definition used is “puredp”.	StringType	{“Discrete Gaussian”, “Two-Sided Geometric”}
VARIANCE	Measure of dispersion	FloatType	31.67, 412.889, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

REGION_ID|REGION_TYPE|ITERATION_CODE|PH5_NUM_DATA_CELL|COUNT|NOISE_DISTRIBUTION|

VARIANCE

1|USA|*|1|12345|Discrete Gaussian|11.2

01|STATE|A|2|13245|Discrete Gaussian|35.8

PH5_denom

Description

Contains count of family households by region and iteration (of the householder). Each row contains a specification of attributes, and the count for those attributes. When the PHSafe algorithm is run, the PH5_denom/ folder will contain exactly one csv file with the prefix part-00000-XXXXX.

Version and Date

2023-09-01.v2.0.1

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	Geocode corresponding to one of USA or STATE.	StringType(2)	1, 06, 44, etc
REGION_TYPE	Name of geography level	StringType(5)	{USA, STATE}

ITERATION_CODE	Characteristic iteration code	StringType(1)	* = Unattributed
			A = White alone
			B = Black or African American alone
			C = American Indian and Alaska Native alone
			D = Asian alone
			E = Native Hawaiian and Other Pacific Islander alone
			F = Some Other Race alone
			G = Two or More Races
			H = Hispanic or Latino
			I = White alone, not Hispanic or Latino

PH5_DENOM_DATA_CELL	Follows PH5 table shell data cell values. The PH5 denominator only includes the total cell.	IntegerType(1)	1 = Total
COUNT	Number of family households corresponding to the given attributes.	IntegerType	12, -3, 670, etc
NOISE_DISTRIBUTION	“Discrete Gaussian” if the privacy definition used is “zcdp”.	StringType	{“Discrete Gaussian”, “Two-Sided Geometric”}

	“Two-Sided Geometric”if the privacy definition used is “puredp”.		
VARIANCE	Measure of dispersion	FloatType	31.67, 412.889, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

REGION_ID|REGION_TYPE|ITERATION_CODE|PH5_DENOM_DATA_CELL|COUNT|NOISE_DISTRIBUTION|VARIANCE

1|USA|*|1|12345|Discrete Gaussian|99.111

01|STATE|A|2|13245|Discrete Gaussian|17.1

PH6

Description

Contains the number of householders' children under 18 by family type, age, and region. Each row contains a specification of attributes, and the count for those attributes. When the PHSafe algorithm is run, the PH6/ folder will contain exactly one csv file with the prefix part-00000-XXXXX.

Version and Date

2023-09-01.v2.0.1

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	Geocode corresponding to one of USA or STATE.	StringType(2)	1, 06, 44, etc
REGION_TYPE	Name of geography level	StringType(5)	{USA, STATE}
PH6_DATA_CELL	Data cell number from PH6 table shell. The output only includes the interior cells.	IntegerType(2)	3 = In married couple families: Under 4 years 4 = In married couple families: 4 and 5 years 5 = In married couple families: 6 to 11 years 6 = In married couple families: 12 to 17 years 8 = In cohabiting couple families: Under 4 years 9 = In cohabiting couple families: 4 and 5 years 10 = In cohabiting couple families: 6 to 11 years 11 = In cohabiting couple families: 12 to 17 years 13 = In male householder, no spouse or partner present family: Under 4 years 14 = In male householder, no spouse or partner present family: 4 and 5 years 15 = In male householder, no spouse or partner present family: 6 to 11 years 16 = In male householder, no spouse

			or partner present family: 12 to 17 years 18 = In female householder, no spouse or partner present family: Under 4 years 19 = In female householder, no spouse or partner present family: 4 and 5 years 20 = In female householder, no spouse or partner present family: 6 to 11 years 21 = In female householder, no spouse or partner present family: 12 to 17 years
COUNT	Number of people corresponding to the given attributes.	IntegerType	12, -3, 670, etc
NOISE_DISTRIBUTION	“Discrete Gaussian” if the privacy definition used is “zcdp”. “Two-Sided Geometric”if the privacy definition used is “puredp”.	{“Discrete Gaussian”, “Two-Sided Geometric”}	{“Discrete Gaussian”, “Two-Sided Geometric”}
VARIANCE	Measure of dispersion	FloatType	31.67, 412.889, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

REGION_ID|REGION_TYPE|PH6_DATA_CELL|COUNT|NOISE_DISTRIBUTION|VARIANCE

1|USA|3|12345|Discrete Gaussian|200.78

01|STATE|6|13245|Discrete Gaussian|200.78

PH7

Description

Contains the number of people in occupied housing units by tenure, region, and iteration (of the householder). Each row contains a specification of attributes, and the count for those attributes. When the PHSafe algorithm is run, the PH7/ folder will contain exactly one csv file with the prefix part-00000-XXXXX.

Version and Date

2023-09-01.v2.0.1

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	Geocode corresponding to one of USA or STATE.	StringType(2)	1, 06, 44, etc
REGION_TYPE	Name of geography level	StringType(5)	{USA, STATE}

ITERATION_CODE	Characteristic iteration code	StringType(1)	* = Unattributed
			A = White alone
			B = Black or African American alone
			C = American Indian and Alaska Native alone
			D = Asian alone
			E = Native Hawaiian and Other Pacific Islander alone
			F = Some Other Race alone
			G = Two or More Races
			H = Hispanic or Latino
			I = White alone, not Hispanic or Latino

PH7_DATA_CELL	Data cell number from PH7 table shell. The output only includes the interior cells.	IntegerType(1)	2 = Owned with a mortgage or a loan 3 = Owned free and clear 4 = Renter occupied
COUNT	Number of people corresponding to the given attributes.	IntegerType	12, -3, 670, etc
NOISE_DISTRIBUTION	“Discrete Gaussian” if the privacy definition used is “zcdp”. “Two-Sided	StringType	{“Discrete Gaussian”, “Two-Sided Geometric”}

	Geometric"if the privacy definition used is "puredp".		
VARIANCE	Measure of dispersion	FloatType	31.67, 412.889, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

REGION_ID|REGION_TYPE|ITERATION_CODE|PH7_DATA_CELL|COUNT|NOISE_DISTRIBUTION|

VARIANCE

1|USA|A|1|12345|Discrete Gaussian|200.78

01|STATE|B|2|13245|Discrete Gaussian|200.78909

PH8_num

Description

Contains the sum of household size by tenure, region, and iteration (of the householder). Each row contains a specification of attributes and the sum for those attributes. Present only if at least one PH7 geo/race iteration is assigned a budget. When the PHSafe algorithm is run, the p12_num/ folder will contain exactly one csv file with the prefix part-00000-XXXXX.

Version and Date

2023-09-01.v2.0.1

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	Geocode corresponding to one of USA or STATE.	StringType(2)	1, 06, 44, etc
REGION_TYPE	Name of geography level	StringType(5)	{USA, STATE}

ITERATION_CODE	Characteristic iteration code	StringType(1)	* = Unattributed
			A = White alone
			B = Black or African American alone
			C = American Indian and Alaska Native alone
			D = Asian alone
			E = Native Hawaiian and Other Pacific Islander alone
			F = Some Other Race alone
			G = Two or More Races
			H = Hispanic or Latino
			I = White alone, not Hispanic or Latino

PH8_NUM_DATA_CELL	Data cell number from PH8 table shell. The output only includes the interior cells.	IntegerType(1)	2 = owner occupied 3 = renter occupied
COUNT	Number of people in households corresponding to the given attributes.	IntegerType	12, -3, 670, etc
NOISE_DISTRIBUTION	"Discrete Gaussian" if the privacy definition used is "zcdp".	StringType	{"Discrete Gaussian", "Two-Sided Geometric"}

	“Two-Sided Geometric”if the privacy definition used is “puredp”.		
VARIANCE	Measure of dispersion	FloatType	31.67, 412.889, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

REGION_ID|REGION_TYPE|ITERATION_CODE|PH8_DENOM_DATA_CELL|COUNT|NOISE_DISTRIBUTION|VARIANCE

1|USA|H|2|12345|Discrete Gaussian|200.78

01|STATE|D|3|13245|Discrete Gaussian|42

PH8_denom

Description

Contains total counts of households by tenure, region, and iteration (of the householder). Each row contains a specification of attributes, and the count for those attributes. When the PHSafe algorithm is run, the p12_denom/ folder will contain exactly one csv file with the prefix part-00000-XXXXX.

Version and Date

2023-09-01.v2.0.1

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	Geocode corresponding to one of USA or STATE.	StringType(2)	1, 06, 44, etc
REGION_TYPE	Name of geography level	StringType(5)	{USA, STATE}

ITERATION_CODE	Characteristic iteration code	StringType(1)	* = Unattributed
			A = White alone
			B = Black or African American alone
			C = American Indian and Alaska Native alone
			D = Asian alone
			E = Native Hawaiian and Other Pacific Islander alone
			F = Some Other Race alone
			G = Two or More Races
			H = Hispanic or Latino
			I = White alone, not Hispanic or Latino

PH8_DENOM_DATA_CELL	Data cell number from PH8 table shell. The output only includes the interior cells.	IntegerType(1)	2 = owner occupied 3 = renter occupied
COUNT	Number of housing units corresponding to the given attributes.	IntegerType	12, -3, 670, etc
NOISE_DISTRIBUTION	"Discrete Gaussian" if the privacy definition used is "zcdp". "Two-Sided Geometric"if the	StringType	{"Discrete Gaussian", "Two-Sided Geometric"}

	privacy definition used is “puredp”.		
VARIANCE	Measure of dispersion	FloatType	31.67, 412.889, etc

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

REGION_ID|REGION_TYPE|ITERATION_CODE|PH8_DENOM_DATA_CELL|COUNT|NOISE_DISTRIBUTION|VARIANCE

1|USA|H|2|12345|Discrete Gaussian|501.1138

01|STATE|D|3|13245|Discrete Gaussian|999.9999