SafeTab-P v5.0.0 Documentation

Executive Summary

The SafeTab-P algorithm produces differentially private tables of statistics (counts) of demographic characteristics crossed by detailed races, ethnicities, and American Indian and Alaska Native tribes and villages at varying levels of geography (i.e., nation, state, county, tract, place, and American Indian, Alaska Native, and Native Hawaiian [AIANNH] areas). The data product derived from the output of SafeTab-P is known as Detailed Demographic and Housing Characteristics File A (Detailed DHC-A).

Goals

1. **Produce tables of statistics (counts)**.
2. **Satisfy differential privacy:** the algorithm used to produce the tables satisfies either pure differential privacy or zero concentrated differential privacy
3. **Low error:** the algorithm should allow users to tune input parameters to improve the error in statistics.

Problem Specification

Demographic Statistics: As per the "20200207: Proposal for 2020 Census Data Products_Race.Ethnicity v3," the Census Bureau would like to release the following statistics for each population group as part of the Detailed DHC-A product:

- T1: (T01001) Total Population by Detailed Race/Ethnicity
- T2: (T02001, T02002, T02003) Sex by Age of Population by Detailed Race/Ethnicity

Each of the tabulations must be released for a set of predefined population groups. The set of population groups (and tabulations released for each) are enumerated below.

Geographies: Statistics are released for population groups at the following geographic levels:

- USA: corresponds to national level counts where the nation is composed of the 50 states + DC
- State / PR-State: corresponds to populations groups at the state level for the 50 states + DC and Puerto Rico
- County / PR-County: corresponds to counties or county equivalents within the 50 states + DC or in Puerto Rico
- Tract / PR-Tract: corresponds to census tracts within the 50 states + DC or in Puerto Rico
- Place / PR-Place: corresponds to Census Bureau places within the 50 states + DC or in Puerto Rico
- AIANNH areas: correspond to the following areas -
    - 0001 - 4999: Federally recognized American Indian Reservations and Off-Reservation Trust Lands
    - 5000 - 5499: Hawaiian Home Lands
    - 5500 - 5599: Oklahoma Tribal Statistical Areas (OTSAs)
    - 6000 - 7999: Alaska Native Village Statistical Areas
    - 8000 - 8999: Tribal Designated Statistical Areas
    - 9000 - 9499: State recognized American Indian Reservations
    - 9500 - 9998: State Designated Tribal Statistical Areas

Note: SafeTab-P will not tabulate statistics for all AIANNH = 9999 nor for PLACE = 99999.

Race and Ethnicity Characteristic Iterations: Statistics must be released for race and ethnicity characteristic iterations for both *Alone* as well as *Alone or in any Combination*. The list of characteristic iterations is partitioned into "`Common`," "`Detailed`," and "`Other`" lists described below:

Race and Ethnicity "`Common`" Characteristic Iterations:

- These are designated in the input file 'race-characteristic-iterations.txt' with DETAILED_ONLY = False and COARSE_ONLY = False (see Appendix A)

Race and Ethnicity "Other" Characteristic Iterations:

- A list of characteristic iterations that capture "Other" detailed and intermediate race and ethnicity characteristic iterations that are not captured in the "Common" list
- These are designated in the input file 'race-characteristic-iterations.txt' with DETAILED_ONLY = False and COARSE_ONLY = True (see Appendix A)

Race and Ethnicity "Detailed" Characteristic Iterations:

- Additional detailed race and ethnicity characteristic iterations
- For instance, these include American Indian and Alaska Native groups that meet a population threshold of 100 or more nationally in the 2010 CPH-T-6 but are not in the "Common" list
- These are designated in the input file 'race-characteristic-iterations.txt' with DETAILED_ONLY = True and COARSE_ONLY = False (see Appendix A)

Depending on the geographic level and which list an iteration comes from, the statistics that need to be released are different as shown in the table below. T1/T2 indicates that SafeTab-P will tabulate either a Total Population Count (T1) *or* an Age x Sex breakout (T2). Whether T1 or T2 is directly tabulated, as well as the level of detail of the Age x Sex breakout, is chosen by the algorithm using a fraction of the privacy-loss budget. The total population for every characteristic iteration in T2 is output in T1 using post processing. The sex marginal (male count and female count) for every characteristic iteration eligible for T2 statistics is constructed from the noisy T2 sex by age measurements and output as part of T2. See the "Approach" section for a more detailed description of the algorithm and these details.

		Characteristic Iterations		
		Common	Detailed	Other
Geography Levels	USA	T1/T2	T1	N/A
	STATE			
	COUNTY	T1/T2	N/A	T1/T2
	TRACT			
	PLACE			
	AIANNH			

Consistency Requirements:

- The statistics released will be integral.

Selected list of potential data inconsistencies in the SafeTab-P output:

- Statistics may contain negative values.
- Levels may not add up.
    - A state's statistics may not match the sum of its corresponding counties
    - A regional iteration code's statistics may not match the sum of its corresponding detailed iteration code.
- The SafeTab-P outputs may appear out-of-sync with thresholds due to the nature of the adaptive procedure.
    - A population group with T2 statistics may have a T1 population count below the threshold for T2 eligibility.
    - A population group with no T2 statistics may have a T1 population count above the threshold for T2 eligibility.
- Alone or in any combination iteration code statistics may be less than the corresponding alone iteration code statistics.

Approach

The T1 and T2 statistics are *linear* over the person records dataframe (See Appendix A for dataframe details). That means, every query in T1 and T2 counts the number of rows in the person records dataframe that satisfy a predicate. The Census categorizes these tables as counts over the *Population Universe.*

Tumult Labs will develop the **SafeTab-P** executable for the Population Universe (i.e., T1 and T2 of Detailed DHC-A). This will take as input configurations for a reader that creates custom records derived from the CEF-Person file and output for each population group (i.e., characteristic iteration and geographic entity) the total population and sex by age statistics.

SafeTab-P will estimate the total count (T1) for all Detailed characteristic iterations. For Common and Other characteristic iterations, SafeTab-P uses a portion of the privacy loss budget to determine whether to release a total count (T1) *or* a Sex x Age breakout (T2). If a Sex x Age breakout is selected, the level of detail of the breakout is also chosen by SafeTab-P. The total count for every characteristic iteration in T2 is output in T1 using post processing. The sex marginal (male count and female count) for every characteristic iteration eligible for T2 statistics is constructed from the noisy T2 sex by age measurements and output as part of T2. The total population for every characteristic iteration is consistent between T1 and T2.

Additionally, SafeTab-P suppresses small population groups in sub-state geographies. SafeTab-P allows users to specify a probability p such that population groups with a true T1 total count of zero will be suppressed.

SafeTab-P can be instantiated to run either to satisfy pure differential privacy (PureDP), or zero concentrated differential privacy (Rho-zCDP).

Performance

We recommend running SafeTab-P 5.0.0 on the r4.16xlarge Amazon EMR instance that comes with 64 vCPUs (2.3 GHz Intel Xeon E5-2686 v4 Processor) and 488 G memory. Observed run times are given below.

Validate Subcommand

SafeTab-P `validate` runs successfully on the full geography (GRF-C.txt) inputs, race/ethnicity inputs and Tumult generated simulated person-records.txt having 300 million records within about 30 minutes on an r4.16xlarge machine on cluster. The validate subcommand uses spark.

Execute Subcommand

Note: `execute` includes `validate` by default

safetab-p on a cluster: On an EMR 6.10 cluster with 1 master (r4.16xlarge) and 2 executor nodes (each an r4.16xlarge instance with 64 vCPUs and 488 G memory) and with the spark settings specified in `resources/spark_configs/spark_cluster_properties.conf`, we observed the SafeTab-P spark app runs successfully (with and without output validation) on the full geography (GRF-C.txt) inputs, race/ethnicity inputs and Tumult generated simulated person-records.txt having up to 300 million records including both the US execution (all 50 states and DC) and the Puerto Rico execution, within 9 hours using either the PureDP or Rho-zCDP privacy definition.

System Requirements

See safetab_p README.

Testing Plan

See safetab_p TESTPLAN.

Documentation

README - includes information and/or pointers to relevant documents on product description, system requirements, installation, testing, and CLI run command instructions.
LICENSE - software license under which SafeTab-P is distributed.

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 for Tumult Labs or by reading synthetic data in csv file format. Output Files refers to files produced by Tumult Labs intended for further use by 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

Input Dataframes

GRF-C-df

This dataframe is from the pipe separated grfc_tab20_natl.txt file formatted as per the [GRFC] schema.

Version and Date

2020-01-14.v1

Column Names and Format Definitions

The schema for this file appears in Appendix A.

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

person-records-df

A representation of custom person records, derived from the CEF Person file that serves as an input to the DAS. We assume that *person-records-df* will contain exactly one row for each person in the US and PR. Domains for all columns (except HOUSEHOLDER) match the identically named columns in CEF20_PER file.

Version and Date

2022-12-02.v3

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
QAGE	Edited Age	IntegerType(3)	0-115
QSEX	Edited Sex	StringType(1)	1 = Male 2 = Female
HOUSEHOLDER	True if RELSHIP==20 in CEF20_PER, otherwise False	StringType(5)	True = Householder False = Not Householder
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
CENRACE	Represents all possible major 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
QRACE1	Edited First Race Variable	StringType(4)	1000-8999
QRACE2	Edited Second Race Variable	StringType(4)	1000-8999 or Null if no code. If QRACE2 is Null, then so are QRACE3-QRACE8
QRACE3	Edited Third Race Variable	StringType(4)	1000-8999 or Null if no code. If QRACE3 is Null, then so are QRACE4-QRACE8
QRACE4	Edited Fourth Race Variable	StringType(4)	1000-8999 or Null if no code. If QRACE4 is Null, then so are QRACE5-QRACE8
QRACE5	Edited Fifth Race Variable	StringType(4)	1000-8999 or Null if no code. If QRACE5 is Null, then so are QRACE6-QRACE8
QRACE6	Edited Sixth Race Variable	StringType(4)	1000-8999 or Null if no code. If QRACE6 is Null, then so will QRACE7-QRACE8
QRACE7	Edited Seventh Race Variable	StringType(4)	1000-8999 or Null if no code. If QRACE7 is Null, then so is QRACE8
QRACE8	Edited Eighth Race Variable	StringType(4)	1000-8999 or Null if no code.

QSPAN	Final Edited Hispanic origin variable	StringType(4)	1000-8999, 9950

Encoding
UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

Sample Records
QAGE|QSEX|HOUSEHOLDER|TABBLKST|TABBLKCOU|TABTRACTCE|TABBLK|CENRACE|QRACE1|QRACE2|QRACE3|QRACE4|QRACE5|QRACE6|QRACE7|QRACE8|QSPAN

28|1|True|08|500|000300|0060|01|1000|Null|Null|Null|Null|Null|Null|Null|1000

Assumptions
1. SafeTab-P will assume all non-null race codes appear in the first few QRACE variables. SafeTab-P will ignore races that appear after the first Null. For instance, consider the following input,
   028|1|True|01|001|000001|0001|01|1010|Null|1011|Null|Null|Null|Null|Null|1010
   SafeTab-P will ignore the second race code 1011 as it appears after a Null.
2. If there are duplicate race-codes in the QRACE variables, SafeTab-P will only consider the unique race codes. For instance, consider the following input,
   028|1|True|01|001|000001|0001|01|1010|1010|Null|Null|Null|Null|Null|Null|1010
   SafeTab-P will use 1010 as the race code even though it appears twice in QRACE1 and QRACE2.

Additional Inputs

config.json

Description

json file encoding inputs to SafeTab-P as key, value pairs. The key value pairs expected by SafeTab-P 5.0.0 are described below.

Version and Date

2022-12-02.v3

Key Value Names and Format Definitions

Key	Description	Value Format	Legal Values
max_race_codes	Maximum race codes for a single record.	Int	1-8
privacy_budget_p_level_<x>_<geo> (for x in {1,2} and geo in {usa, state, county, tract, place, aiannh, pr_state, pr_county, pr_tract, pr_place})	The privacy loss budget assigned to geo level <geo> and characteristic iteration level <x>.	Float	0.128, 5.66, etc
privacy_budget_p_stage_1_fraction	The fraction of privacy budget to be used to compute the noisy estimate of the size of a population group, used in workload selection (where applicable).	Float	.1, .25, .5, etc Must be between 0 and 1

thresholds_pghbn[p-=== “””””””””””	A list of thresholds for each combination of geography level and iteration level for safetab-p.
	
	Level 1 AIANNH counts will not be published, so these threshold values can be set to anything.
	
	{
	<geo>_<x>: [at least 3 non-decreasing integers]
	for geo in {usa, state, county, tract, place, aiannh, pr_state, pr_county, pr_tract, pr_place} and x in {1,2}
	}
		{
		Str: List[int]
		}
			See example config below

zero_suppression_chance	The probability that a sub-state population group t1 count with a true count of zero is suppressed. We calculate thresholds based on the amount of added noise such that a	Float	.99, .999, etc. Must be between 0 and

	zero total + noise has a chance greater than this of being suppressed. Larger zero_suppression_chances will result in larger thresholds. The parameter can be set to 0.0 or omitted from the config file to disable suppression.		1, 0 is allowed (but 1 is not).
allow_negative_counts	Whether SafeTab-P should apply nonnegative postprocessing.	Bool	{true, false}
run_us	When true, run safetab-p for 50 states + DC	Bool	{true, false}
run_pr	When true, run safetab-p for Puerto Rico. Only records corresponding to “72” will be tabulated for the PR run (if run_pr=true).	Bool	{true, false}
reader	The reader being used. This reader reads the person, and geo files and filters them based on which states are being used. Which states are being used is based on state_filter_us for US runs and is just “72” for PR runs.	Str	{“csv”, “cef”}
state_filter_us	A list of states from the 50 states + DC to include in the US run. PR should not be included in the list. If run_us=true, only records corresponding to these states will be tabulated for the US run. If run_us=false, state_filter_us is ignored.	List[str]	See example config below
privacy_defn	The privacy definition being used, either Pure DP (“puredp”) or Rho zCDP (“zcdp”). Determines how privacy budgets are interpreted.	Str	{“puredp”, “zcdp”}

Encoding

UTF-8

Delimiter Character

JSON uses structured key-value pairs so does not have delimiters.

Comment Character

JSON does not support comment characters.

Sample

```
{
  "max_race_codes": 8,
  "privacy_budget_p_level_1_usa": 0.54,
  "privacy_budget_p_level_2_usa": 5.4,
  "privacy_budget_p_level_1_state": 0.54,
  "privacy_budget_p_level_2_state": 5.4,
  "privacy_budget_p_level_1_county": 0.54,
  "privacy_budget_p_level_2_county": 2.43,
```

```
  "privacy_budget_p_level_1_tract": 0.54,
  "privacy_budget_p_level_2_tract": 2.43,
  "privacy_budget_p_level_1_place": 0.54,
  "privacy_budget_p_level_2_place": 2.43,
  "privacy_budget_p_level_1_aiannh": 0,
  "privacy_budget_p_level_2_aiannh": 2.43,
  "privacy_budget_p_level_1_pr_state": 0.54,
  "privacy_budget_p_level_2_pr_state": 5.4,
  "privacy_budget_p_level_1_pr_county": 0.54,
  "privacy_budget_p_level_2_pr_county": 2.43,
  "privacy_budget_p_level_1_pr_county": 0.54,
  "privacy_budget_p_level_2_pr_county": 2.43,
  "privacy_budget_p_level_1_pr_county": 0.54,
  "privacy_budget_p_level_2_pr_county": 2.43,
  "privacy_budget_p_stage_1_fraction": 0.1,
  "thresholds_p": {
      "(USA, 1)": [5000, 20000, 150000],
      "(USA, 2)": [500, 1000, 7000],
      "(STATE, 1)": [5000, 20000, 150000],
      "(STATE, 2)": [500, 1000, 7000],
      "(COUNTY, 1)": [5000, 20000, 150000],
      "(COUNTY, 2)": [1000, 5000, 20000],
      "(TRACT, 1)": [5000, 20000, 150000],
      "(TRACT, 2)": [1000, 5000, 20000],
      "(PLACE, 1)": [5000, 20000, 150000],
      "(PLACE, 2)": [1000, 5000, 20000],
      "(AIANNH, 1)": [5000, 20000, 150000],
      "(AIANNH, 2)": [1000, 5000, 20000],
      "(PR-STATE, 1)": [5000, 20000, 150000],
      "(PR-STATE, 2)": [500, 1000, 7000],
      "(PR-COUNTY, 1)": [5000, 20000, 150000],
      "(PR-COUNTY, 2)": [1000, 5000, 20000]
      "(PR-TRACT, 1)": [5000, 20000, 150000],
      "(PR-TRACT, 2)": [1000, 5000, 20000],
      "(PR-PLACE, 1)": [5000, 20000, 150000],
      "(PR-PLACE, 2)": [1000, 5000, 20000]
  },
  "zero_suppression_chance": 0.9999,
  "allow_negative_counts": true,
  "run_us": true,
  "run_pr": true,
  "reader": "cef",
  "state_filter_us": ["55", "30", "46", "11", "48", "36", "23", "51", "06",
"28",
       "27", "42", "56", "01", "21", "20", "08", "13", "45", "12", "05",
       "38", "22", "37", "19", "31", "16", "41", "24", "02", "17", "09",
       "54", "29", "47", "40", "18", "25", "04", "53", "26", "39",
       "35", "10", "49", "34", "32", "50", "44", "15", "33"],
  "privacy_defn": "zcdp"
}
```

race_and_ethnicity_codes.txt

Description

Universe of detailed race and ethnicity codes.

Version and Date

2022-12-02.v5

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
RACE_ETH_CODE	Numeric race or ethnicity code	StringType(4)	1000-9999
RACE_ETH_NAME	Common name of race or ethnicity	StringType	See example records below

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

RACE_ETH_CODE|RACE_ETH_NAME

1000|White (Checkbox)

1001|White

1002|White American (By Checkbox)

1003|White American (By Write-in)

1004|White (Edit Generated)

1010|Albanian

1015|Alsatian

1020|Andorran

race-characteristic-iterations.txt

Description

Universe of race characteristic iterations that includes the code associated with the iteration, the name of the iteration, whether the iteration is Alone (if ALONE=True) or Alone or in any Combination (if ALONE=False). Additionally, this file indicates whether the iteration should be tabulated *only* as a total at the national and state level, or *only* in T1 below the state level and T2. Note iterations that should be tabulated everywhere will have DETAILED_ONLY=False and COARSE_ONLY=False.

Assumptions

- Iteration codes in this file are distinct from iteration codes in ethnicity-characteristic-iterations.txt.
- Rows with DETAILED_ONLY=True and COARSE_ONLY=True are not permitted.

Version and Date

2022-12-02.v7

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
ITERATION_CODE	Numeric iteration code	StringType(4)	1000-9999
ITERATION_NAME	Common name of iteration (including Alone/Alone or in any Combination)	StringType	See examples below
LEVEL	Level of iteration in the iteration hierarchy	StringType(1)	0 = major races (not be tabulated by SafeTab-P) 1 = regional race groups 2 = detailed race groups
ALONE	True if Alone, False if Alone or in any Combination	StringType(5)	{“True”, “False”}
DETAILED_ONLY	True if this iteration should be tabulated only with the detailed iterations, i.e. **only** the total at the state and national level should be computed. Null if this is a level 0 iteration (and thus will not be tabulated).	StringType(5)	{“True”, “False”, “Null”}
COARSE_ONLY	True if this iteration should be tabulated only with the coarse iterations, i.e. **only** the tabulations for T2 (all levels) and T1 (below	StringType(5)	{“True”, “False”, “Null”}

	state level) should be computed. Null if this is a level 0 iteration (and thus will not be tabulated).		

Encoding
UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

Sample Records
ITERATION_CODE|*ITERATION_NAME*|*LEVEL*|*ALONE*|*DETAILED_ONLY*|*COARSE_ONLY*

1001|White alone|0|True|Null|Null

1002|European alone|1|True|False|False

1003|Albanian alone|2|True|False|False

1004|Alsatian alone|2|True|True|False

…

1070|Other European alone (All geos)|2|True|False|True

…

1060|European alone or in any combination|1|False|False|False

1061|Albanian alone or in any combination|2|False|False|False

Note: `h_r`, or the number of race characteristic iterations a race group is contained in, will equal the number of levels specified in the input file only when race characteristic iterations at each level correspond to ranges of distinct race codes. When this is not true, SafeTab-P will compute the correct `h_r` and raise the following validation error:

```
The maximum value in the LEVEL fields of 'race-characteristic-iterations.txt'
is 2, but the computed height is 3
```

ethnicity-characteristic-iterations.txt

Description

Universe of ethnicity characteristic iterations that includes the code associated with the iteration, the name of the iteration. Additionally, this file indicates whether the iteration should be tabulated *only* as a total at the national and state level, or *only* in T1 below the state level and T2. Note that iterations that should be tabulated everywhere will have DETAILED_ONLY=False and COARSE_ONLY=False.

Assumptions

- Iteration codes in this file are distinct from iteration codes in race-characteristic-iterations.txt.
- Rows with DETAILED_ONLY=True and COARSE_ONLY=True are not permitted.

Version and Date

2022-12-02.v7

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
ITERATION_CODE	Numeric iteration code	StringType(4)	1000-9999
ITERATION_NAME	Common name of iteration	StringType	See examples below
LEVEL	Level of iteration in the iteration hierarchy:	StringType(1)	0 = major races (not be tabulated by SafeTab-P) 1 = regional race groups 2 = detailed race groups
DETAILED_ONLY	True if this iteration should be tabulated only with the detailed iterations, i.e. **only** the total at the state and national level should be computed. Null if this is a level 0 iteration (and thus will not be tabulated).	StringType(5)	{“True”, “False”, “Null”}
COARSE_ONLY	True if this iteration should be tabulated only with the coarse iterations, i.e. **only** the tabulations for T2 (all levels) and T1 (below state level) should be computed. Null if this is a level 0 iteration (and thus will not be tabulated).	StringType(5)	{“True”, “False”, “Null”}

Encoding
UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

Sample Records
ITERATION_CODE|ITERATION_NAME|LEVEL|DETAILED_ONLY|COARSE_ONLY

3008|Hispanic or Latino (of any race)|0|Null|Null

3009|Mexican|1|False|False

3010|Central American|1|False|False

3011|Costa Rican|2|False|False

…

3039|Other Hispanic, Latino, or Spanish responses, not specified (National)|2|True|False

3040|Other Hispanic or Latino, not specified (All geos)|2|False|True

Note: h_e, or the number of ethnicity characteristic iteration a race group is contained in, will equal the number of levels specified in the input file only when ethnicity characteristic iterations at each level correspond to ranges of distinct race codes. When this is not true, SafeTab-P will compute the correct h_e and raise the following validation error:

```
The maximum value in the LEVEL fields of
'ethnicity-characteristic-iterations.txt' is 2, but the computed height is 3
```

race-and-ethinicty-code-to-iteration.txt

Description

Mapping of characteristic iterations to detailed race and ethnicity codes. There will be one row for every iteration_code that a race_eth_code is associated with. Tumult provides a script, `convert_short_form.py`, within SafeTab-Utils to generate this file from the race-and-ethnicity-code-to-iteration-short-form.txt file.

Version and Date

2022-12-02.v5

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal values
ITERATION_CODE	numeric code of characteristic iteration	StringType(4)	1000-9999
RACE_ETH_CODE	numeric race or ethnicity code	StringType(4)	1000-9999

Encoding

UTF-8

Delimiter Character

vertical bar (|)

Comment Character

Not supported.

Sample Records

ITERATION_CODE|RACE_ETH_CODE

1003|1010

1003|1011

1003|1012

1003|1013

1003|1014

reader_config

Description

CEF Reader program reads its input parameters from a .ini file. The INI configuration file consists of sections, each beginning with a [section] header, followed by key/value entries separated by "=". Following is a sample of a CEF Reader configuration file.

Sample Section

```
[paths]
cef_year = 2020
per_dir = s3://v-s3-das-prod-data-412241963457-us-gov-west-1/mft/cdl-to-das/PER/
unit_dir = s3://v-s3-das-prod-data-412241963457-us-gov-west-1/mft/cdl-to-das/UNIT/
per_dir_pr = s3://v-s3-das-prod-data-412241963457-us-gov-west-1/mft/cdl-to-das/PER/
unit_dir_pr = s3://v-s3-das-prod-data-412241963457-us-gov-west-1/mft/cdl-to-das/UNIT/
grfc_dir = s3://v-s3-das-common-drps-412241963457-us-gov-west-1/2020/production/grfc/
per_file_format = CEF20_PER_%%s.txt
unit_file_format = CEF20_UNIT_%%s.txt
geo_file_format = grfc_tab20_%%s.txt
```

Output Files

Output is saved in directories with multiple pipe-delimited csv files. The output directory names are listed below.

t1

Description

A folder containing one or more csv files. Each file contains total counts for a subset of the population groups. Each file contains a header with attribute names, followed by rows that specify a geography, characteristic iteration, and count. USA is the 50 states + DC, STATE is a state or territory, and COUNTY is a county or county equivalent. TRACT is a census tract and PLACE is a Census Bureau place. PR-STATE, PR-COUNTY, PR-TRACT, and PR-PLACE are equivalent concepts in Puerto Rico. AIANNH is an American Indian, Alaska Native, and Native Hawaiian area. When the safetab-p algorithm is run with both run_us and run_pr set to True, the t1/ folder will contain 2 csv files with the prefix part-00000-XXXXX - one with US rows and the other with PR rows. If either run_us or run_pr is set to True, the t1/folder will contain only 1 csv file with corresponding rows. The csv files will have a format as described below.

Note: Noisy measurements are produced for all geographic entities in schema regardless of whether or not those entities contain housing units or group quarters facilities.

Version and Date

2022-12-02.v3

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	Geocode corresponding to one of STATE, COUNTY, AIANNH area, etc	StringType(11)	1, 44, 44007, etc
REGION_TYPE	Name of geography level	StringType(9)	{USA, AIANNH, STATE, COUNTY, TRACT, PLACE, PR-STATE, PR-COUNTY, PR-TRACT, PR-PLACE}
ITERATION_CODE	Characteristic iteration code	IntegerType(4)	1000-9999
COUNT	Number of people corresponding to the given geography and characteristic iteration	IntegerType	12, 52, 670, etc

Encoding

UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

Sample Records
REGION_ID|REGION_TYPE|ITERATION_CODE|COUNT

1|USA|1279|12345

01|STATE|1279|13245

01123|COUNTY|1280|5543

01001000001|TRACT|1114|4

1199999|PLACE|1181|2

4444|AIANNH|1001|1138

t2

Description

A folder containing one or more csv files. Each file contains age and sex statistics. Each file contains a header with attribute names, followed by rows containing a specification of attributes, and the count for those attributes. The value * matches any attribute value. USA is the 50 states + DC, STATE is a state or territory, and COUNTY is a county or county equivalent. TRACT is a census tract and PLACE is a Census Bureau place. PR-STATE, PR-COUNTY, PR-TRACT, and PR-PLACE are equivalent concepts in Puerto Rico. AIANNH is an American Indian, Alaska Native, and Native Hawaiian area. When the safetab-p algorithm is run with both run_us and run_pr set to True, the t2/ folder will contain 2 csv files with the prefix part-00000-XXXXX - one with US rows and the other with PR rows. If either run_us or run_pr is set to True, the t2/folder will contain only 1 csv file with corresponding rows. The csv files will have a format as described below.

Note: Noisy measurements are produced for all geographic entities in schema regardless of whether or not those entities contain housing units or group quarters facilities.

Version and Date

2022-12-02.v3

Column Names and Format Definitions

Column Name	Description	Format Specification	Legal Values
REGION_ID	entity code corresponding to one of STATE, COUNTY, AIANNH area, etc	StringType(11)	1, 44, 44007, etc
REGION_TYPE	Name of geography level	StringType(9)	{USA, AIANNH, STATE, COUNTY, TRACT, PLACE, PR-STATE, PR-COUNTY, PR-TRACT, PR-PLACE}
ITERATION_CODE	Characteristic iteration code	IntegerType(4)	1000-9999
AGESTART	Beginning of age range	IntegerType(3) or StringType(1)	* or 0-115
AGEEND	End of age range	IntegerType(3) or StringType(1)	* or 0-115
SEX	Sex	StringType(1)	1 = Male 2 = Female
COUNT	Number of people corresponding to given attribute values	IntegerType	2, -52, 670, etc

Encoding

UTF-8

Delimiter Character
vertical bar (|)

Comment Character
Not supported.

Sample Records
REGION_ID|REGION_TYPE|ITERATION_CODE|SEX|AGESTART|AGEEND|COUNT

1|USA|1279|1|30|49|12345

01|STATE|1279|2|25|29|13245

01123|COUNTY|1280|1|25|29|5543

01002000001|TRACT|1003|1|45|54|2

1199999|PLACE|1294|2|67|69|6

4444|AIANNH|1138|1|*|*|56