VLDS Insights

Data: Security & Technology Panel

David Ihrie, Center for Innovative Technology (Moderator)
Will Goldschmidt, Virginia Department of Education
Aaron Schroeder, Virginia Tech Institute for Policy & Governance

Principles of Good Data Programs

	Facebook/ Google	NSA	VLDS
Utility (Why)	?	✓	✓
Transparency (How, When, Where)	?	X → ?	✓
Security (Appropriate)	?	✓	✓
Boundary (What)	X	?	✓
Accountability (Who)	X	X	✓

VLDS

VLDS Insights

Data: Security & Technology Panel

David Ihrie, Center for Innovative Technology (Moderator)
Will Goldschmidt, Virginia Department of Education
Aaron Schroeder, Virginia Tech Institute for Policy & Governance

Privacy Protecting Federated query

- Two Steps:

1. Identity Resolution Process
2. Query Execution Process

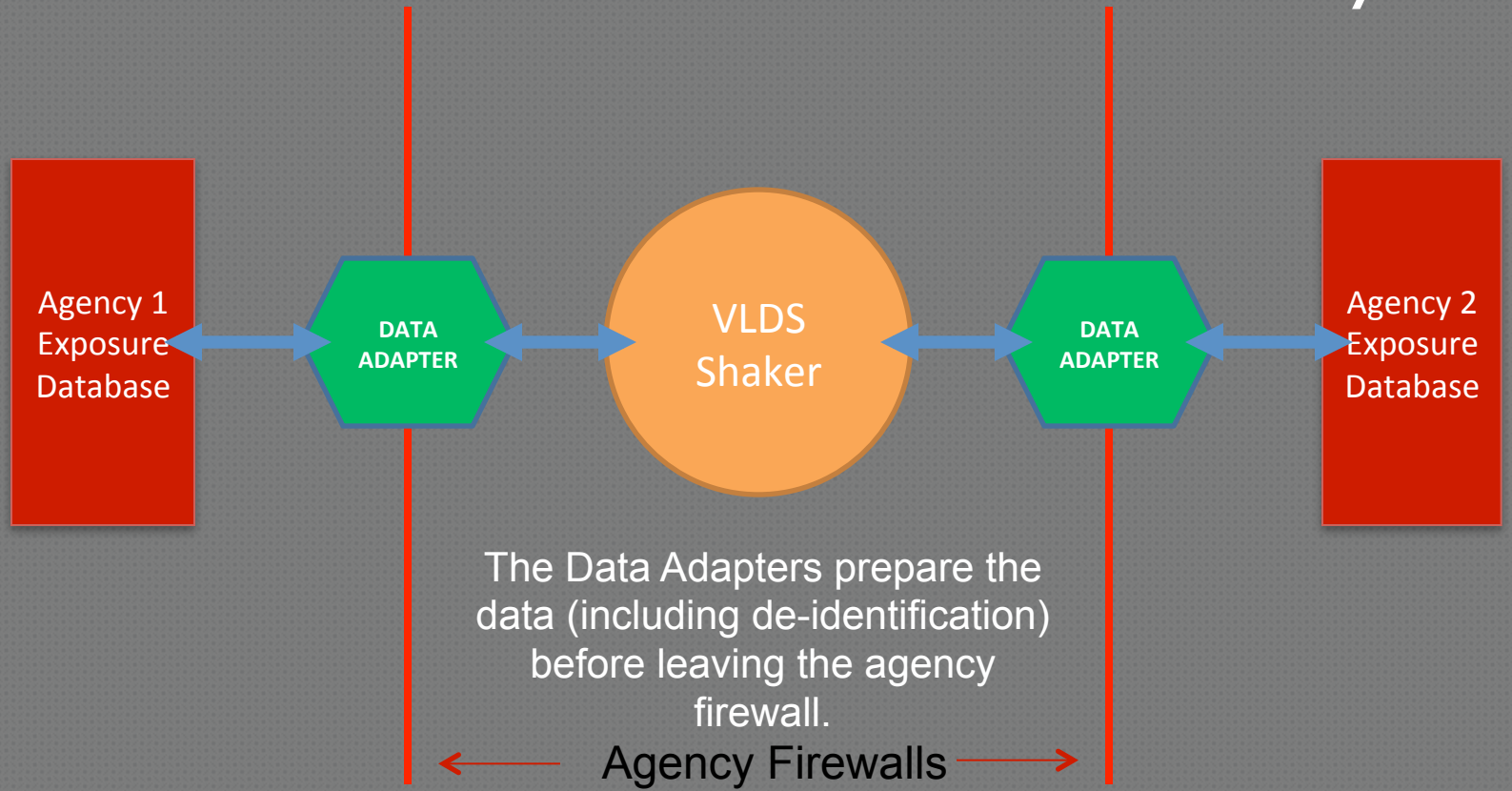
- Three Components

1. Shaker
2. Data Adapter
3. Exposure Database

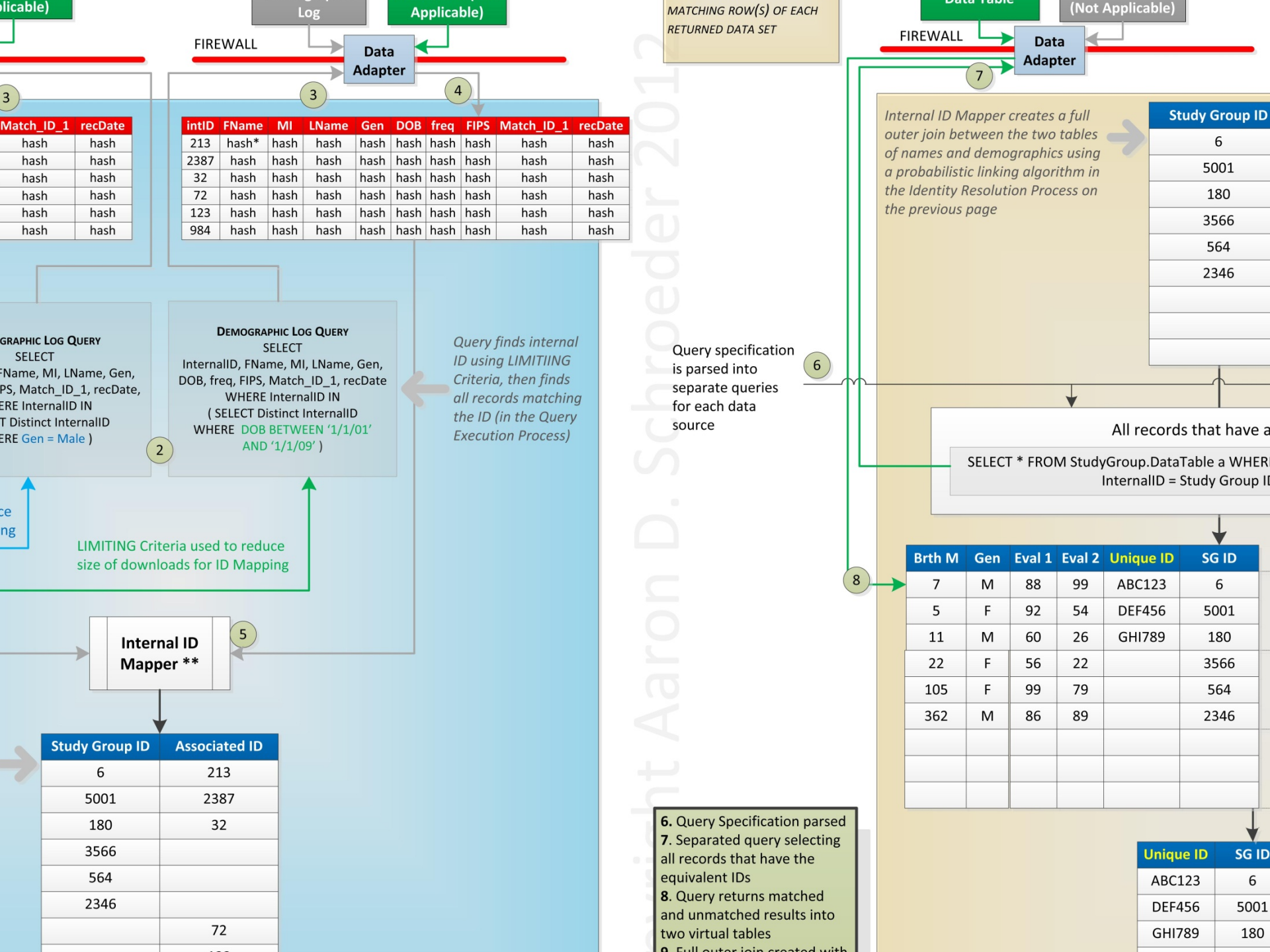
VLDS

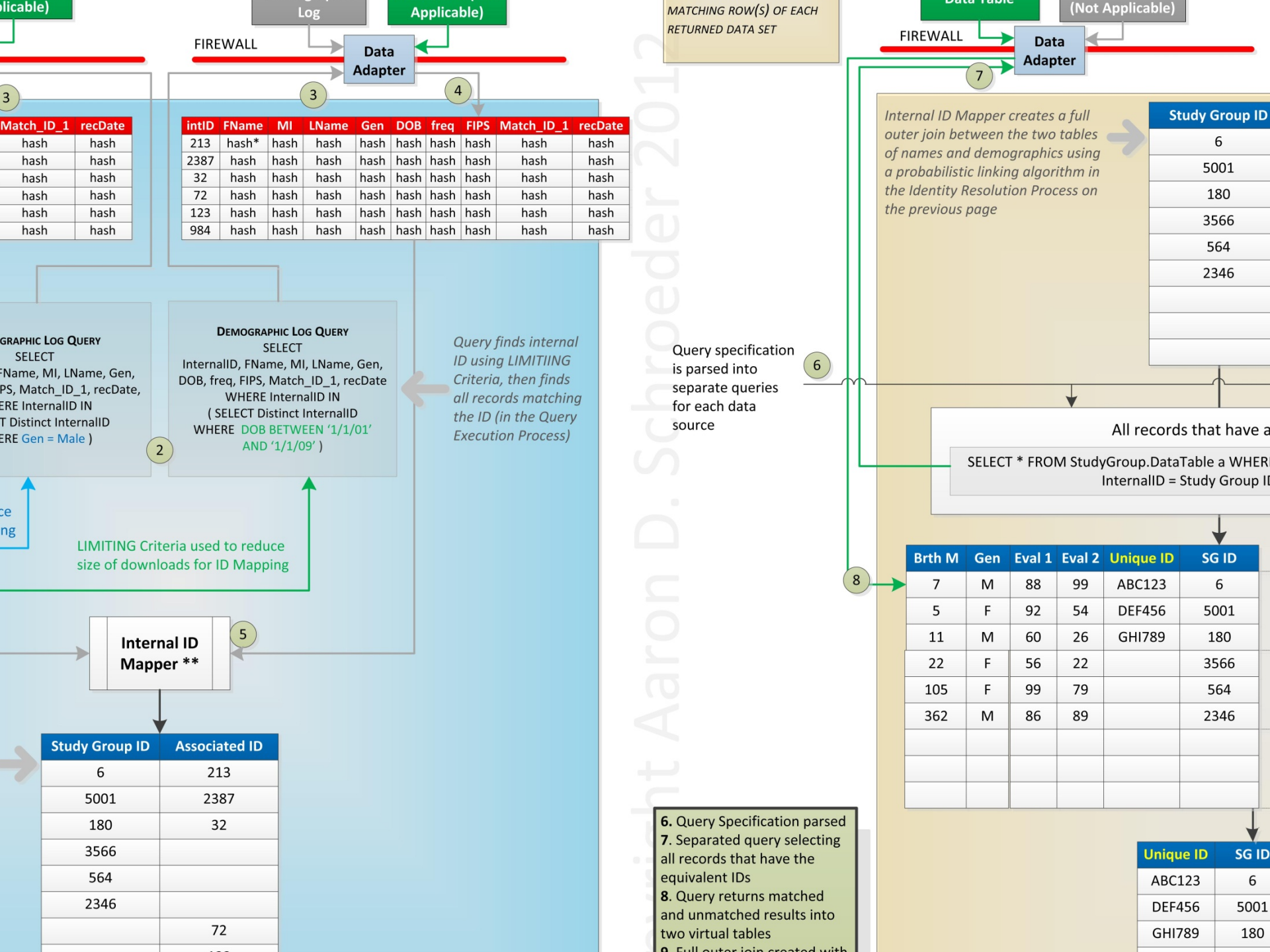
Getting Data Ready for "De-Identified Federation"

- How De-Identification Works in VLDS System



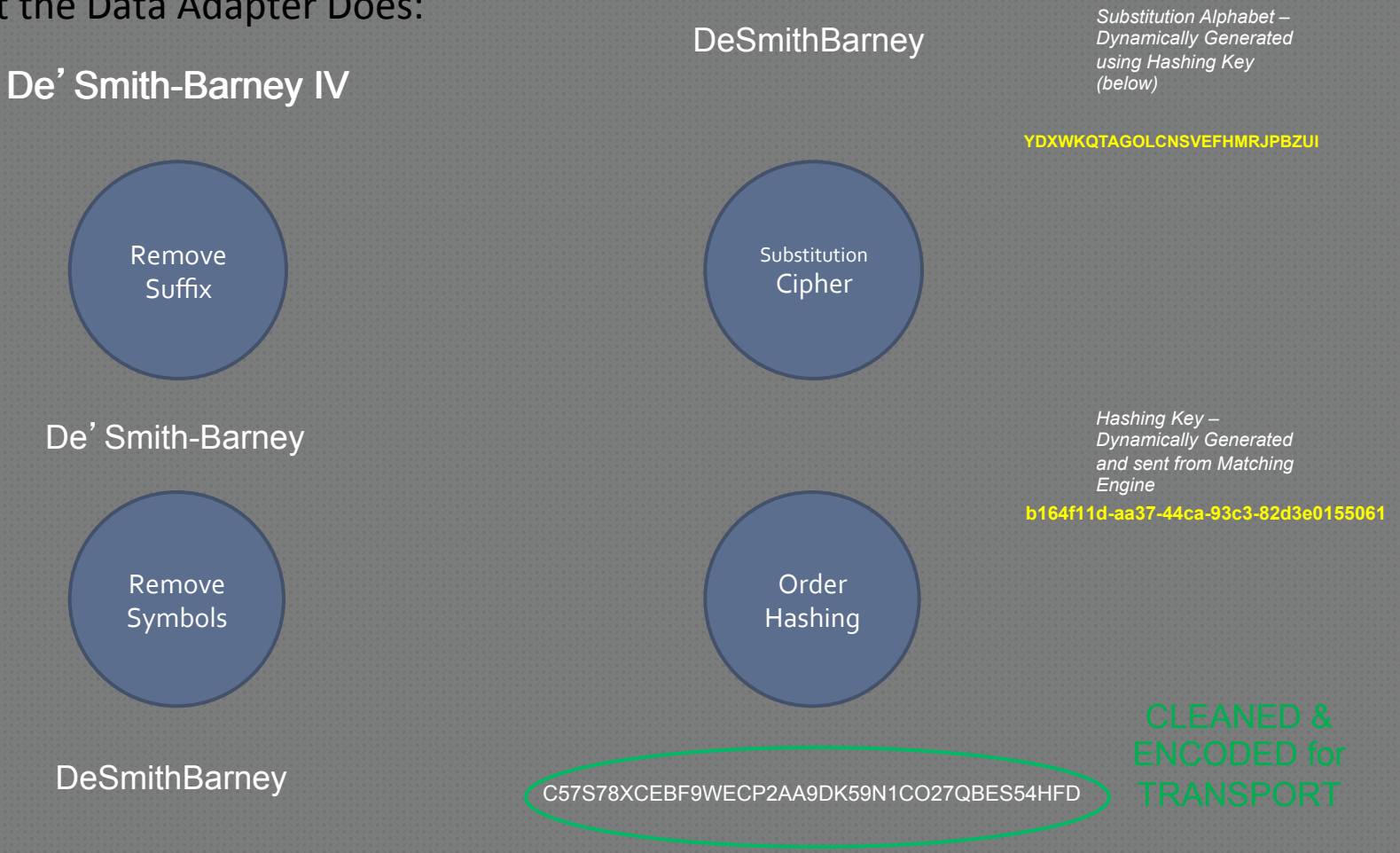
VLDS


Applicable)
Log
Applicable)
FIREWALL
Data Adapter
3
3
4
Match_ID_1 recDate
hash hash
hash hash
hash hash
hash hash
hash hash
hash hash
intID FName MI LName Gen DOB freq FIPS Match_ID_1 recDate
213 hash* hash hash hash hash hash hash hash hash
2387 hash hash hash hash hash hash hash hash hash
32 hash hash hash hash hash hash hash hash hash
72 hash hash hash hash hash hash hash hash hash
123 hash hash hash hash hash hash hash hash hash
984 hash hash hash hash hash hash hash hash hash
GRAPHIC LOG QUERY
SELECT
FName, MI, LName, Gen,
PS, Match_ID_1, recDate,
ERE InternalID IN
T Distinct InternalID
ERE Gen = Male)
2
DEMOGRAPHIC LOG QUERY
SELECT
InternalID, FName, MI, LName, Gen,
DOB, freq, FIPS, Match_ID_1, recDate
WHERE InternalID IN
(SELECT Distinct InternalID
WHERE DOB BETWEEN '1/1/01'
AND '1/1/09')
Query finds internal ID using LIMITIING Criteria, then finds all records matching the ID (in the Query Execution Process)
LIMITING Criteria used to reduce size of downloads for ID Mapping
Internal ID Mapper **
5
Study Group ID Associated ID
6 213
5001 2387
180 32
3566
564
2346
72
MATCHING ROW(S) OF EACH RETURNED DATA SET
Data Table
(Not Applicable)
FIREWALL
Data Adapter
7
Internal ID Mapper creates a full outer join between the two tables of names and demographics using a probabilistic linking algorithm in the Identity Resolution Process on the previous page
Study Group ID
6
5001
180
3566
564
2346
Query specification is parsed into separate queries for each data source
6
All records that have a
SELECT * FROM StudyGroup.DataTable a WHER
InternalID = Study Group ID
8
Brth M Gen Eval 1 Eval 2 Unique ID SG ID
7 M 88 99 ABC123 6
5 F 92 54 DEF456 5001
11 M 60 26 GHI789 180
22 F 56 22 3566
105 F 99 79 564
362 M 86 89 2346
Unique ID SG ID
ABC123 6
DEF456 5001
GHI789 180
6. Query Specification parsed
7. Separated query selecting all records that have the equivalent IDs
8. Query returns matched and unmatched results into two virtual tables
9. Full outer join created with
Copyright Aaron D. Schroeder 2012


Applicable)
Log
Applicable)
FIREWALL
Data Adapter
3
3
4
Match_ID_1 recDate
hash hash
hash hash
hash hash
hash hash
hash hash
hash hash
intID FName MI LName Gen DOB freq FIPS Match_ID_1 recDate
213 hash* hash hash hash hash hash hash hash hash
2387 hash hash hash hash hash hash hash hash hash
32 hash hash hash hash hash hash hash hash hash
72 hash hash hash hash hash hash hash hash hash
123 hash hash hash hash hash hash hash hash hash
984 hash hash hash hash hash hash hash hash hash
GRAPHIC LOG QUERY
SELECT
FName, MI, LName, Gen,
PS, Match_ID_1, recDate,
ERE InternalID IN
T Distinct InternalID
ERE Gen = Male)
2
DEMOGRAPHIC LOG QUERY
SELECT
InternalID, FName, MI, LName, Gen,
DOB, freq, FIPS, Match_ID_1, recDate
WHERE InternalID IN
(SELECT Distinct InternalID
WHERE DOB BETWEEN '1/1/01'
AND '1/1/09')
Query finds internal ID using LIMITIING Criteria, then finds all records matching the ID (in the Query Execution Process)
LIMITING Criteria used to reduce size of downloads for ID Mapping
Internal ID Mapper **
5
Study Group ID Associated ID
6 213
5001 2387
180 32
3566
564
2346
72
MATCHING ROW(S) OF EACH RETURNED DATA SET
Data Table
(Not Applicable)
FIREWALL
Data Adapter
7
Internal ID Mapper creates a full outer join between the two tables of names and demographics using a probabilistic linking algorithm in the Identity Resolution Process on the previous page
Study Group ID
6
5001
180
3566
564
2346
Query specification is parsed into separate queries for each data source
6
All records that have a
SELECT * FROM StudyGroup.DataTable a WHER
InternalID = Study Group ID
8
Brth M Gen Eval 1 Eval 2 Unique ID SG ID
7 M 88 99 ABC123 6
5 F 92 54 DEF456 5001
11 M 60 26 GHI789 180
22 F 56 22 3566
105 F 99 79 564
362 M 86 89 2346
Unique ID SG ID
ABC123 6
DEF456 5001
GHI789 180
6. Query Specification parsed
7. Separated query selecting all records that have the equivalent IDs
8. Query returns matched and unmatched results into two virtual tables
9. Full outer join created with
Copyright Aaron D. Schroeder 2012

Getting Data Ready for "De-Identified Federation"

- What the Data Adapter Does:



VLDS

Cleaned and Encoded Matching Data (Internal ID, First and Last Name)

INTERNAL_ID_HASHED	FIRST_NAME	LAST_NAME
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	X11SDAK3EF86
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	X11SDAK3EF86
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	X11SDAK3EF86
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	X11SDAK3EF86
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	X11SDAK3EF86
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	X11SDAK3EF86
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	X11SDAK3EF86
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E
044AA90CE74E2ED3B6B0B0CFE93F8ED263B73050	F11BDAE3EM86AE8	J11EDAV3E

INTERNAL_ID is the same

LAST_NAME is NOT

Many agencies DO NOT have an Index of unique individuals. There can be many representations of that individual.

What do we do?

Statistical Log Analysis and Reduction

We dynamically build a new "virtual" record made up of "most likely" demographics

VLDS

Probabilistic Linkage Process (Creating a Linking Directory)

(After we have a unique person index for each agency dataset)

- Blocking
- m and u Parameter Calculation
- Matching-Column Weight Calculations
- Match Scoring
- Linkage Determination and addition to Linking Directory

- Linkage Determination – A Cutoff score needs to be set for each blocked comparison, below which a link is not accepted as a real "link"
- The best method of establishing this cutoff is for the system operator to work with a content-area expert to determine the peculiarities of data for that content-area
- In some data sets in may be very unlikely that a birthdate was entered incorrectly, while in another, it may happen very regularly – a computer can not automatically know this
- Once these cutoffs are set, they don't need to be changed unless something drastic occurs to change the nature of the dataset