Building Data Sharing Infrastructures at the State Level

Context, Stakeholders, Technology

Aaron D. Schroeder, Ph.D.
Senior Data Scientist, Social & Decision Analytics Laboratory
Virginia Tech Bioinformatics Institute

State-Level Linkage Projects

Builder of Interagency Data Partnerships & Systems

- 511 Virginia
  - Virginia State Police
  - Virginia Department of Transportation
  - Virginia Tourism Corporation
  - Virginia Tech
- Child HANDS
  - Virginia Department of Education
  - Virginia Department of Social Services
  - Virginia Department of Health
- Virginia Longitudinal Data System
  - Virginia Department of Education
  - State Council on Higher Education in Virginia
  - Virginia Employment Commission
  - Virginia Community College System

Technology facilitates, but isn't the key

Long processes of trust and partnership-building are paramount

Keys to Building Successful Data Partnerships

- Must get to a shared vision
- Must establish TRUST between all key data partners
- Technology should be designed as much as possible to work within the existing political and economic context of the deployment

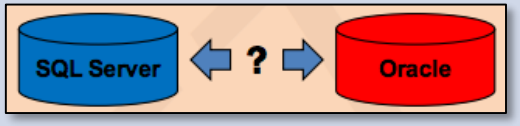
Impediments to Public Sector Data Integration

Top-Down = "Saddle on a Sow"

Impediments Common to All Data Integration Efforts

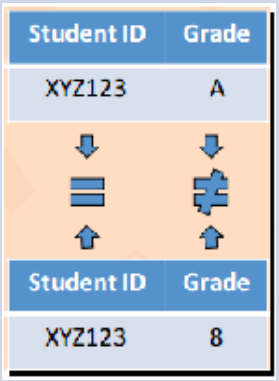
Technological Heterogeneity

- Hardware Differences
- Software Differences



Semantic Heterogeneity

Differences in meaning, Interpretation, or Intended use of data



Additional Public Sector Impediments – The "Tough Stuff"

Regulatory Heterogeneity

- Multiple Sets of Statutory Law at the Federal and State Levels (FERPA, HIPAA, State Privacy Acts)
- Multiple Interpretations of Statutory Law (Regulations) at the Federal, State and Local Levels

Authority Structure Heterogeneity

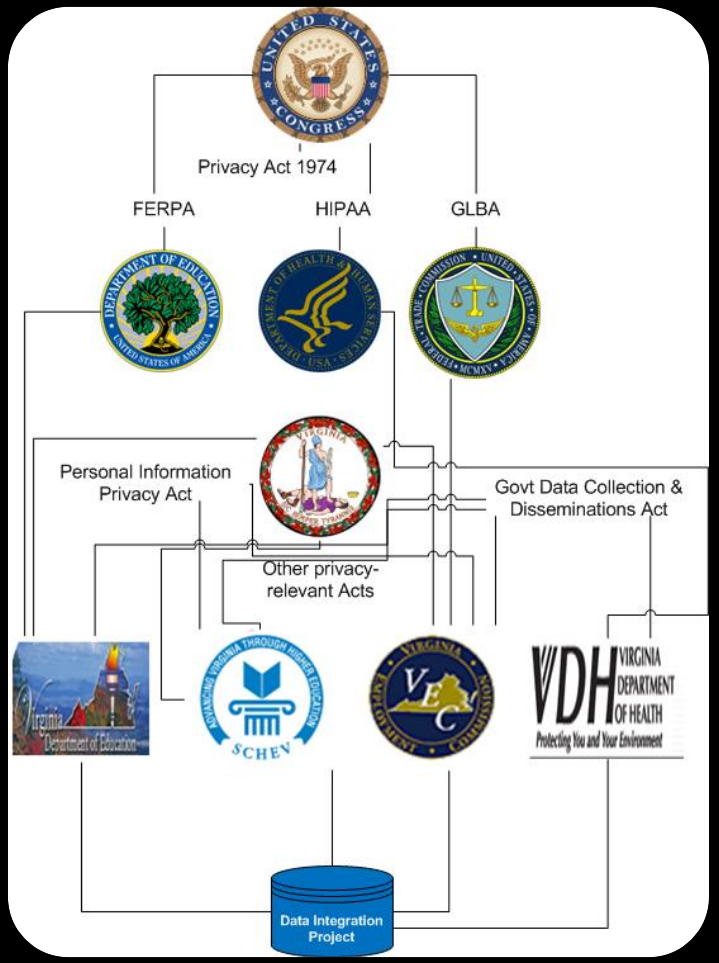
- Variability in the division and lines of authority in an organization
- Structure of authority varies from agency to agency, especially at the state level where authority is often shared with local level agencies

Example
Implementation Environment of the Virginia Longitudinal Data System

- Multiple levels of statutory law
- Multiple implementations of regulatory law at each level of statutory law
- Most conservative interpretation of regulatory law becomes de facto standard

"No one person , inside or outside a government agency, should be able to create a set of identified linked data records between partner agencies"

- Has a direct and significant effect on the potential success of the technical approach chosen – A Centralized, Hierarchical Data Warehouse will likely Fail!
- Easy to see, if you look for it!



How to Implement in this Environment?

- Assess the Implementation Context
  - Understanding impediments from multiple frames of reference (political, economic, organizational, technological)
- Assess and Recruit Stakeholders
  - Understanding who needs to be, and who does NOT need to be, involved
- Build a Joint-Vision including
  - The desired output of the effort
  - An understanding of who needs to be involved at the program level
  - An understanding of who needs to be involved at the technical level

Theories and Methods for my "Steps"

- Assessing the Environment or Contextual Assessment
  - Political Economy of Organizations
  - Quota Sampling
  - Snowballing
- Selecting and Building a Stakeholder Network
  - Political Economy of Organizations
  - Stakeholder Analysis
- Building a New Organization/System from the Stakeholder Network or Joint Visioning
  - Political Economy of Organizations
  - Implementation Networks
  - Techniques of Facilitation: Goal Setting, Program Development, Implementation

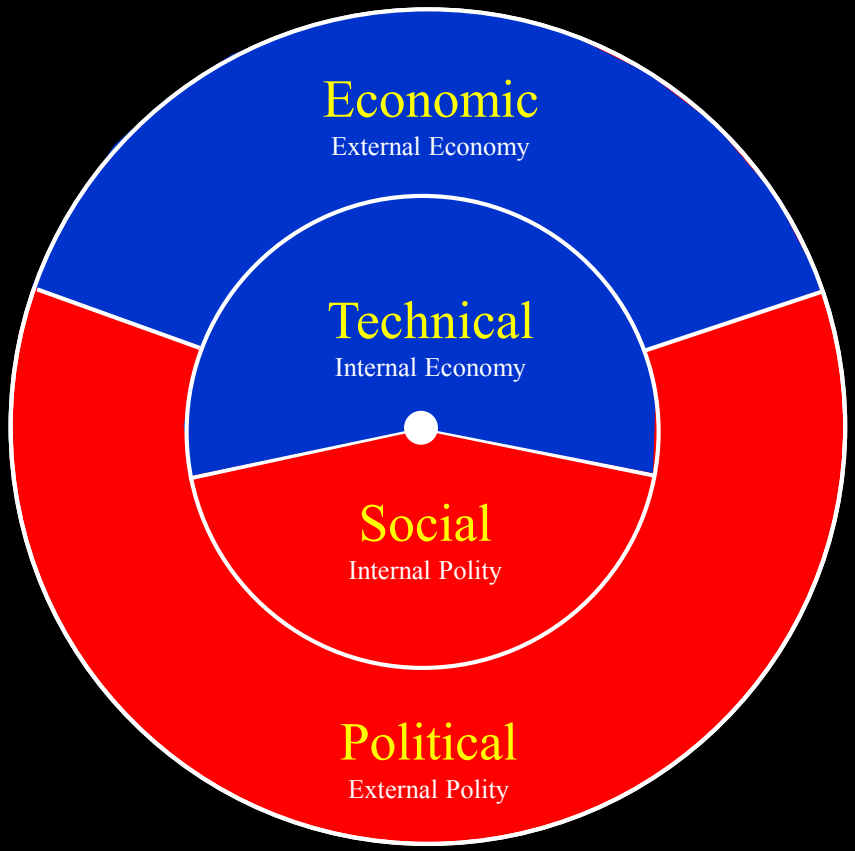
Theories and Methods for my “Steps”

- There are, of course, MANY tools/frameworks/rubrics to help you frame your potential implementation
- Many are based on some form of systems/dependency-network analysis
- I find that they miss or give short-shrift to what I have found to be the most important elements of implementation in complex multi-organizational, multi-sectorial scenarios. Namely, the Political, Economic, and Organizational dimensions – in addition to the Technical Dimension

Understanding Implementation
The Political Economic Framework

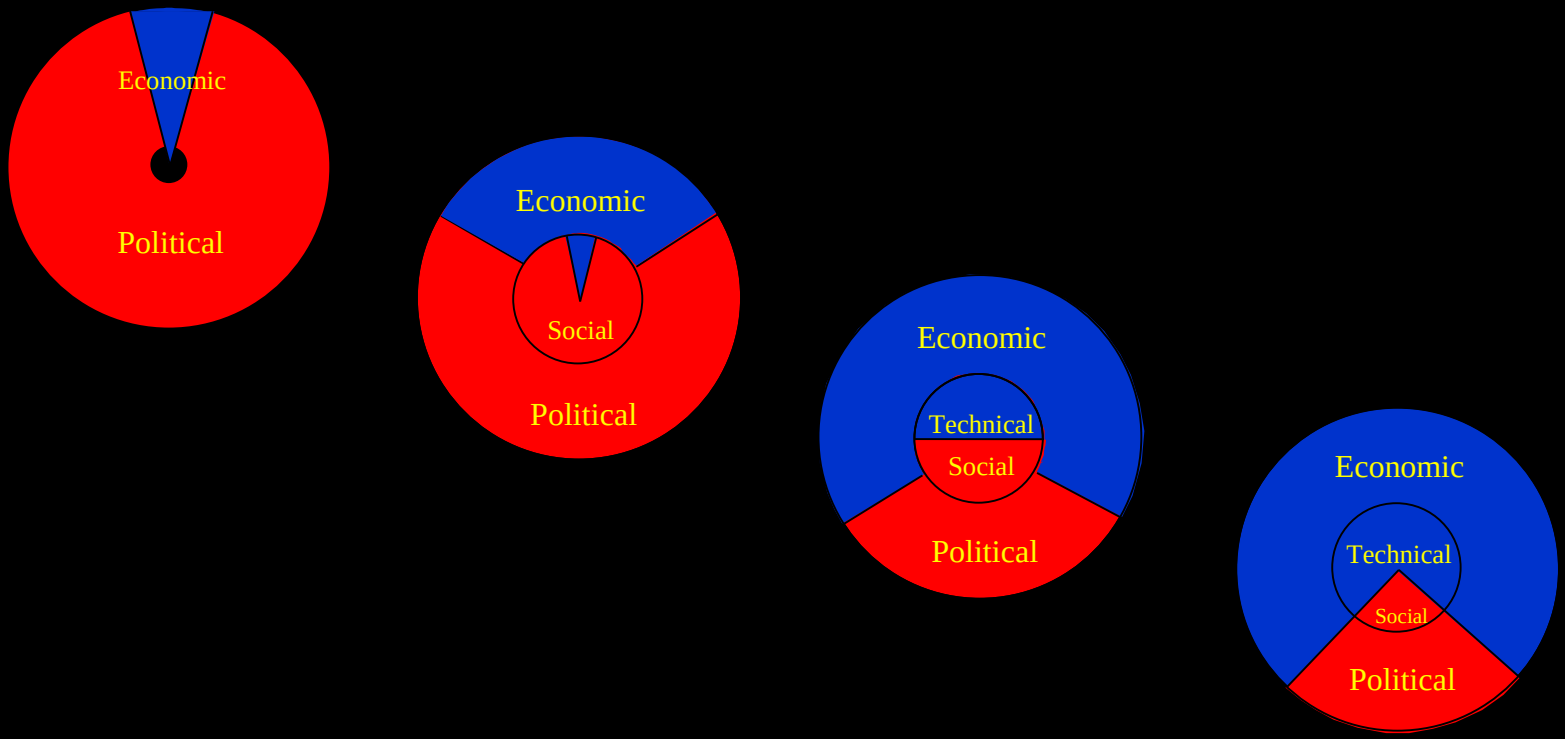
- **Political Environment**
  - Who likes us?
  - Level of surveillance by external actors; External actors understanding of org. goals; Match between statutory charge and political environment; Level which external control mechanisms dictate internal resource allocation; Level of external support & influence available to org. from larger network
- **Economic Environment**
  - Show me the money!
  - Level of demand for outputs (products); Availability of resource inputs (personnel, $$, technical resources); Recipients of outputs (citizens, customers?); Amount received for output ($$, power, prestige, fuzzy feeling?); Level of competition
- **Social / Organizational System**
  - Sempre Fi!
  - Organization mission; Organization goals; Dominant norms and values; Measurement and analysis of job performance; Recruitment system(s); Incentive System(s)
- **Technical / Functional System**
  - Which budget do we pay for the 100-base-T upgrade with?
  - The "production system"; Primary system functions; Required functional positions; Required functional responsibilities; Technological requirements; Budget and budgeting system; Purchasing & accounting system

The Four Political Economic Dimensions
(Re-envisioned, Re-named, Dynamized)



Network Implementation as Political Economy

Where you want to go



The Evolution of 511 Virginia

Goal Setting Network

Original Stakeholders
ITS Director, Virginia Department of Transportation (VDOT)
President, Virginia Tourism Corporation (VTC)
Associate Planner, Lord Fairfax Planning District Commission (LFPDC)
Vice President, SHENTEL Telephone Corp. (SHENTEL)
Dir. Tech Policy & Deployment, Center for Transportation Research (CTR)

Additional Stakeholders Added After Iteration
EDS Dir., Virginia State Police (VSP)
Dir. Public Affairs, Shenandoah National Park
Dir. Shenandoah Valley Travel Association (SVTA)

Program Development Network

Original Program Level Representatives
Policy Analyst, ITS Department, VDOT
Dir. Shenandoah.Com, SHENTEL
Sr. Transport Research Fellow, CTR
Special Projects Dir., VTC
Dir. Tech Policy & Deployment, CTR
Research Associate, CTR

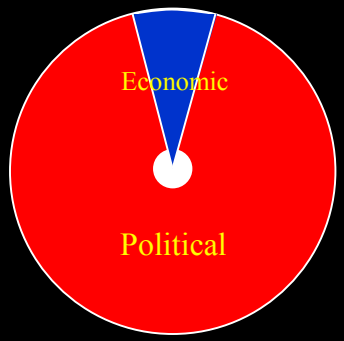
Additional Representatives Added After Iteration
Dir. Emergency Operations Center (EOC), VDOT
Dir. Shenandoah Valley Travel Association (SVTA)
Dir. Virginia.Org, VTC/VT

Operational Implementation Network

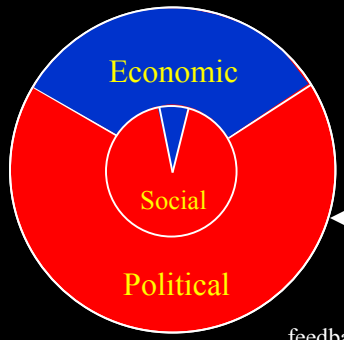
Original Operational Implementation Network Staff
Dir. Tech Policy & Deployment, CTR
Research Associate, CTR
Ops. Mgr. EOC, VDOT
Dir. Shenandoah.Com, SHENTEL
Sr. Transportation Research Fellow, CTR
Systems/Database Programmer, CTR
Systems/Database Programmer, EOC, VDOT
Systems/Database Programmer, VT Outreach/VTC

Additional Staff Added After Iteration
Research Associate, CTR
Data Analyst 1, TS, SHENTEL
Commission Sales Staff, TS, SHENTEL
Market Analyst, CTR
Marketing Dir., TravelShenandoah.Com (TS), SHENTEL
Data Analyst 2, TS, SHENTEL
Data Analyst, CTR
Systems/Database Programmer, SVTA

To Start: *No "Organization" to speak of. Only a loosely configured political environment. Many thoughts on what to do, but little, if any, mobilization of resources.*

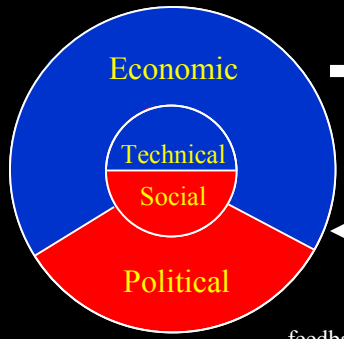


Result of Goal Setting
As stakeholders are brought together to discuss the possible implementation of a new system, ideas about what this means to each stakeholder begin to coalesce. An Idea about what this new system/organization might look like, and who would be responsible for it begins to form (the internal structure begins to form). This coming together of ideas allows the preliminary commitment of resources to begin (an economy begins to form).



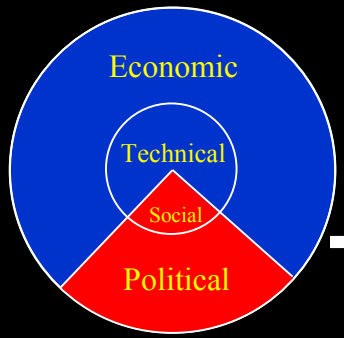
feedback

Result of Program Development
After organizational commitment is secured, departmental responsibilities are assigned. The economic viability of the new organization is more secure, and the technical side of the new organization begins to grow.



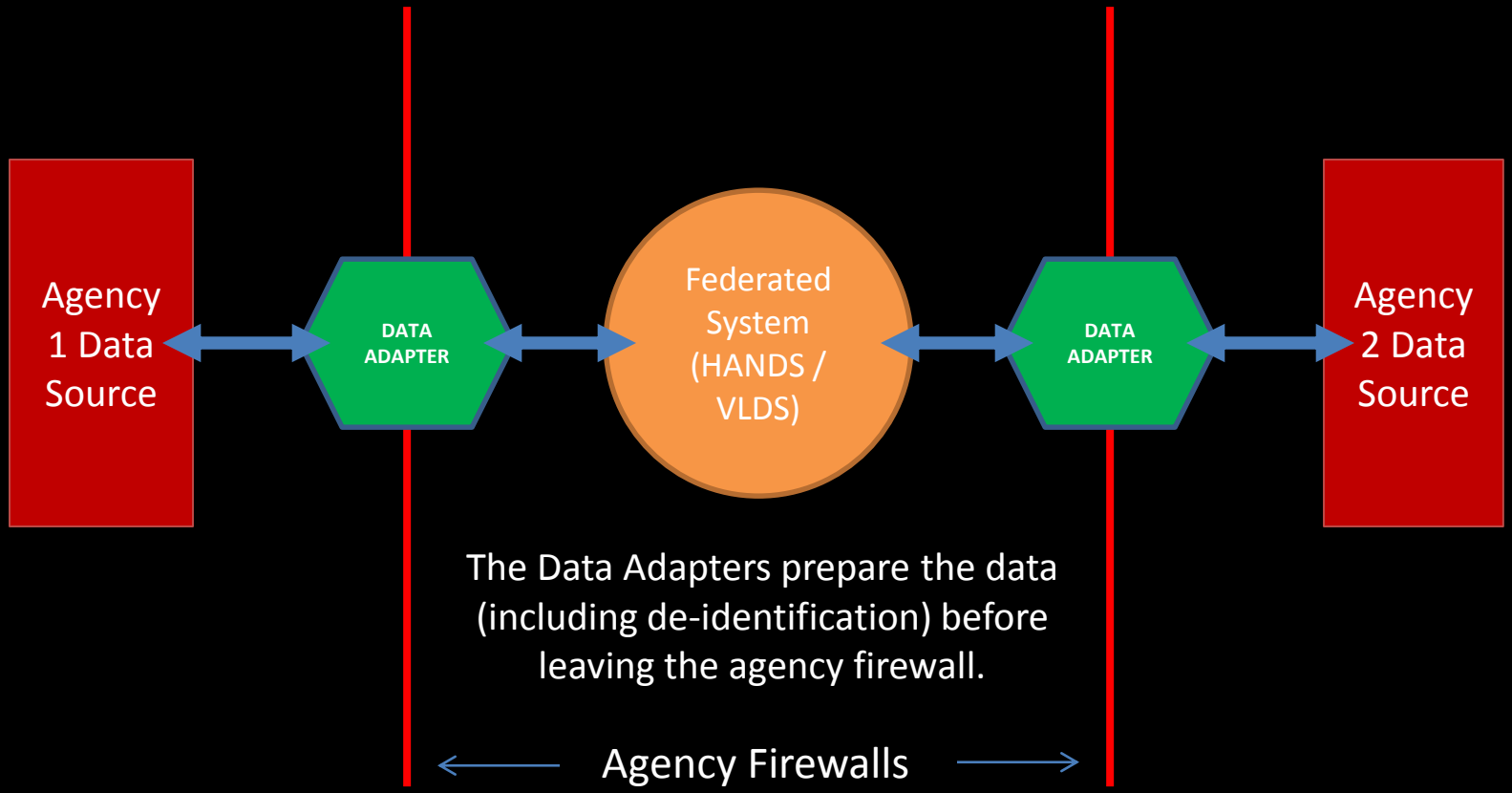
feedback

Result of Operational Implementation
The ideal result of the operational implementation stage is a socio-technical system (internal PE) that is functioning as a stable production system in balance with its political economic environment.

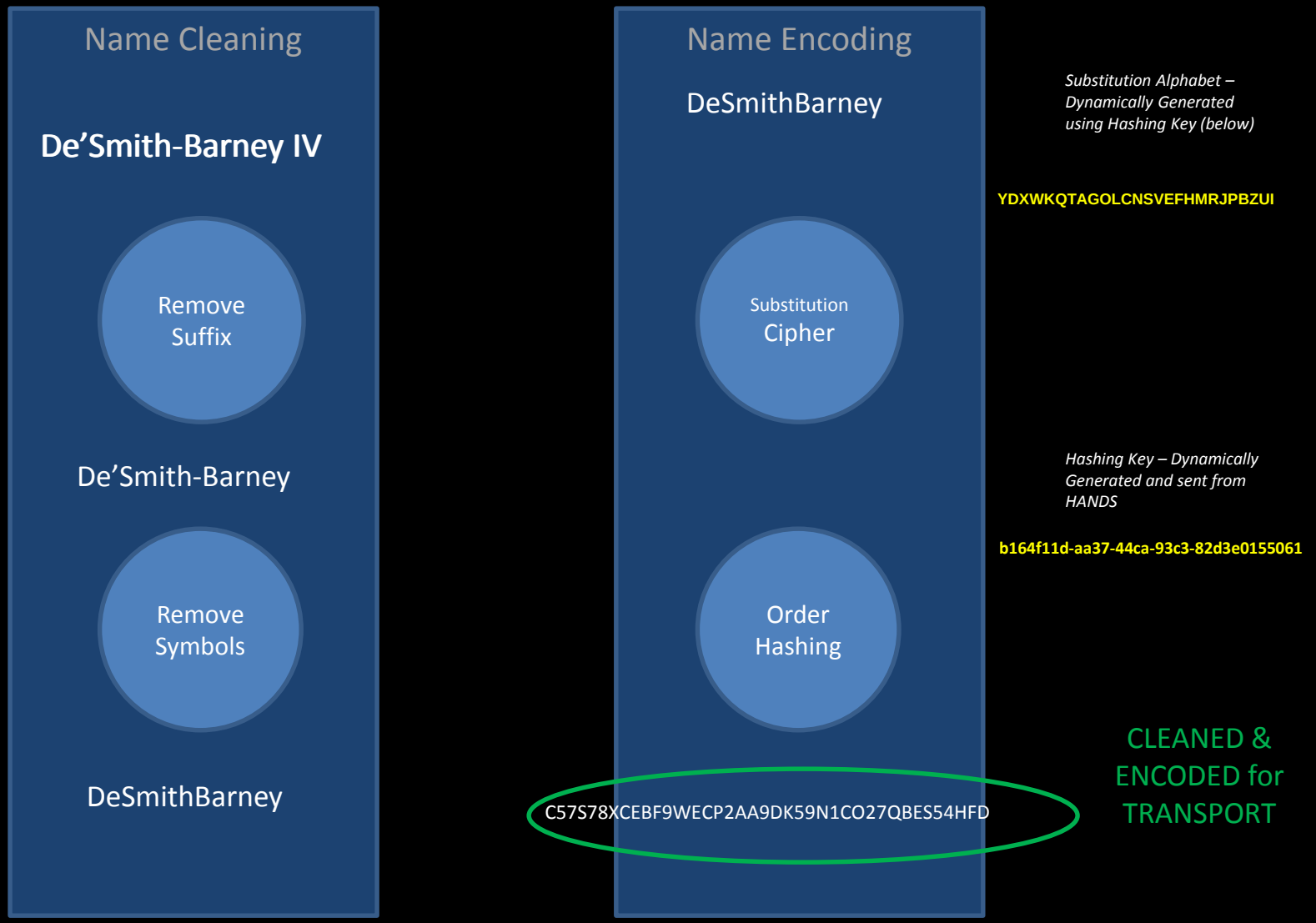


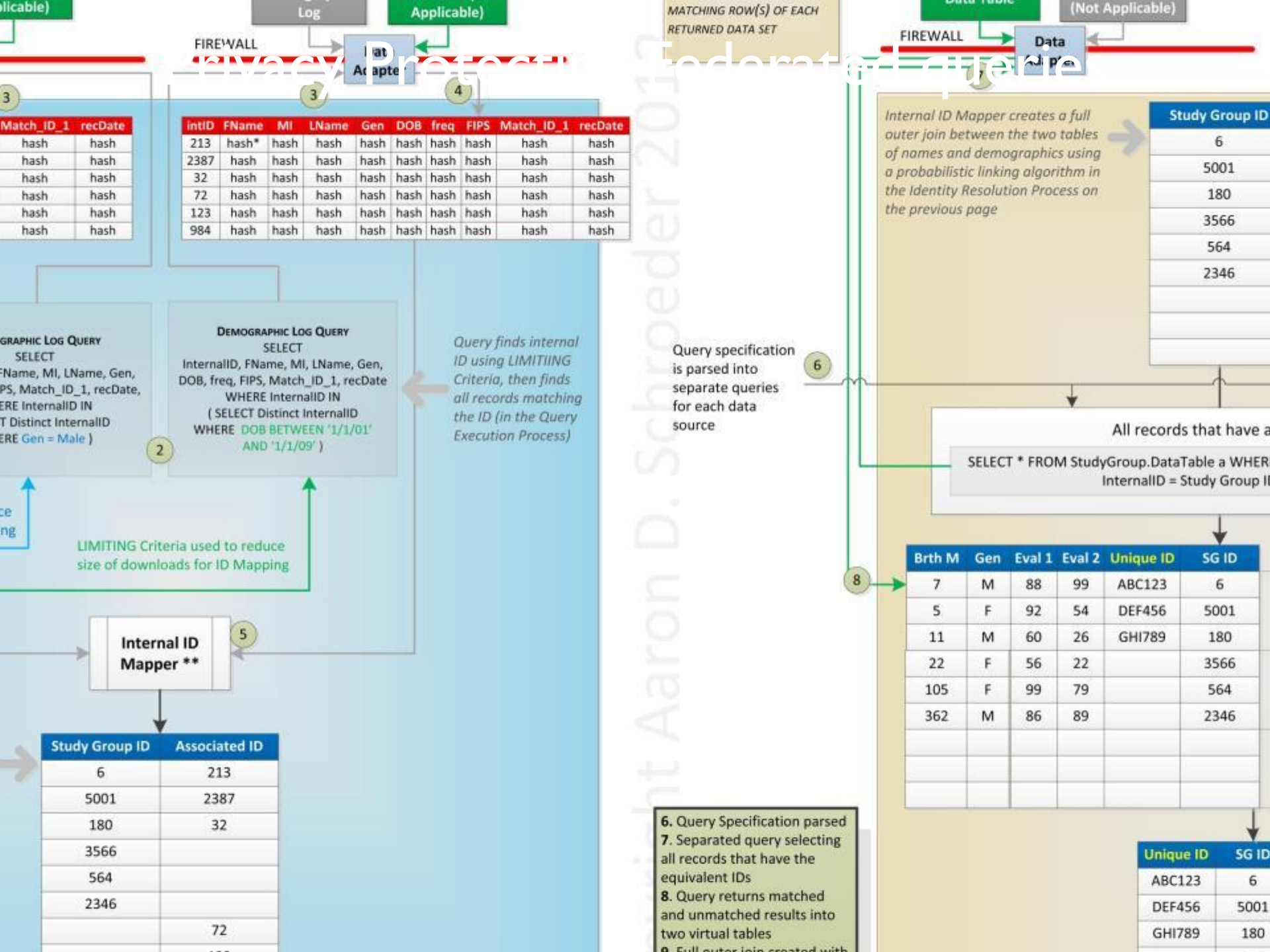
The Technology

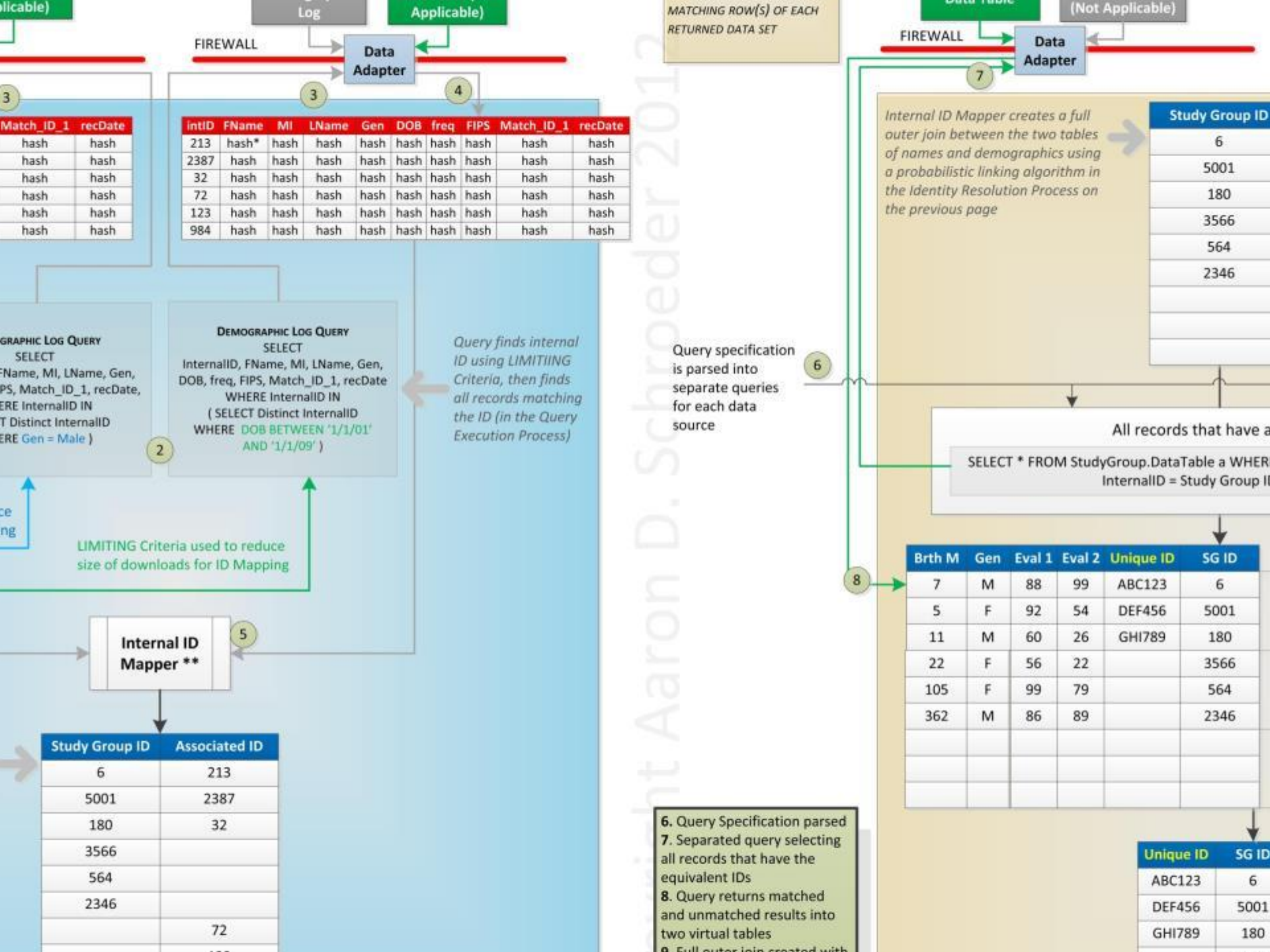
Virginia – Protecting Partner Data



What the Data Adapter Does




Applicable)
Log
Applicable)
FIREWALL
MATCHING ROW(S) OF EACH RETURNED DATA SET
(Not Applicable)
FIREWALL
Data
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
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
GRAPHIC LOG QUERY
SELECT
FName, MI, LName, Gen,
PS, Match_ID_1, recDate,
ERE InternalID IN
T Distinct InternalID
ERE Gen = Male)
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')
2
Query finds internal ID using LIMITIING Criteria, then finds all records matching the ID (in the Query Execution Process)
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 I
LIMITING Criteria used to reduce size of downloads for ID Mapping
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
Internal ID Mapper **
5
Study Group ID Associated ID
6 213
5001 2387
180 32
3566
564
2346
72
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



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

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

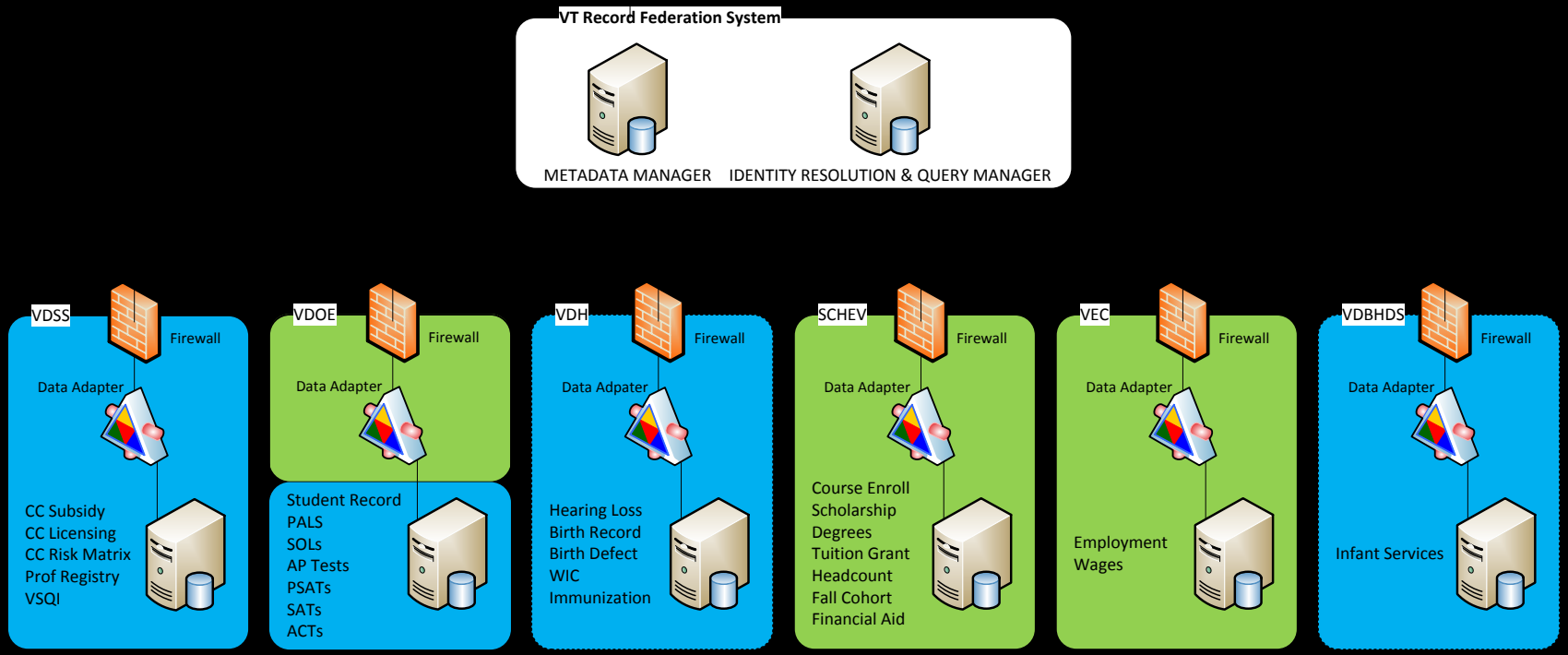
Linking Technology Supports Multiple Systems

Child HANDS

Early Childhood System focused on the relationship between child care subsidy, child care quality, and early school performance

Virginia Longitudinal Data System (VLDS)

Connecting K-12 to Higher Education and Workforce Data focused on the relationship between K-12 preparation and higher education performance and/or entrance into the workforce

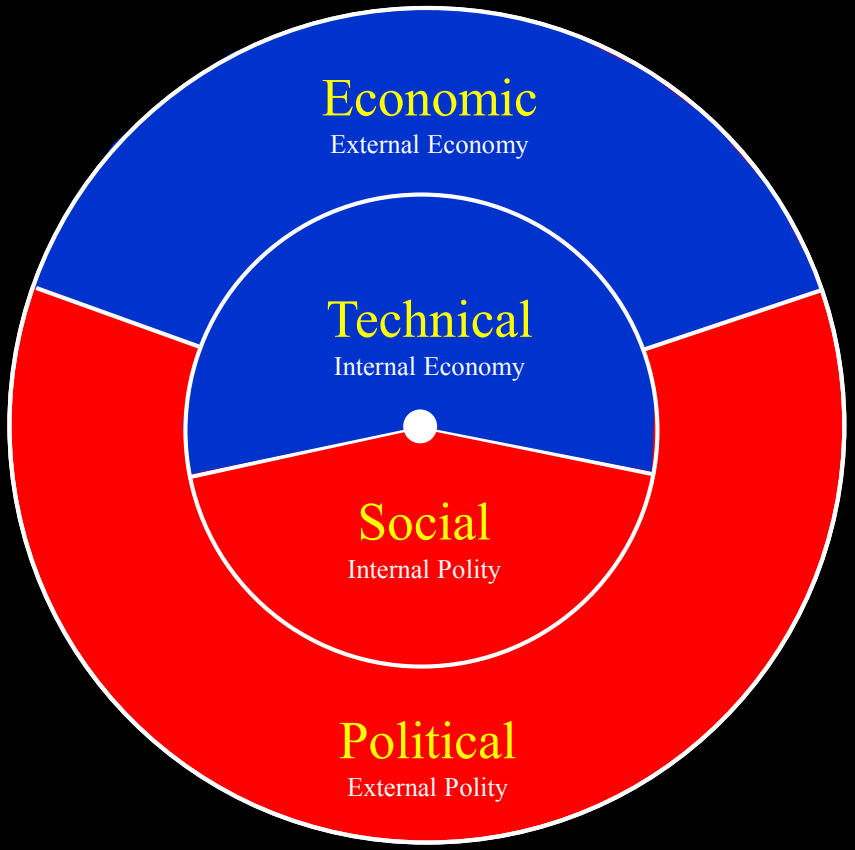


Inter-organizational, multi-sectorial, project timing Rule of Thumb

- 75%
  - Building trust and the attendant political and economic support necessary for implementation to be allowed to succeed
- 25%
  - Building, Testing and Deploying the technology

(if you find most of your initial time is spent on the technology, you should be concerned)

Thank You!


Economic
External Economy
Technical
Internal Economy
Social
Internal Polity
Political
External Polity