Multi-Agency Integration of Child-Relevant Data Sets in the Commonwealth of Virginia
Application of a Privacy Protecting Federated Model

Aaron D. Schroeder, Ph.D.
Institute for Policy & Governance
Virginia Tech

Current Application of Privacy Protecting Federated Model

Project Child HANDS: Child Care Subsidy and Early Education: **H**elping **A**nalyze **N**eeded **D**ata **S**ecurely

Funded By: US Department of Health & Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation (3 years, 800K)

Project Partners:

- Virginia Tech Institute for Policy & Governance (VT-IPG)
- Virginia Tech Child Development Center for Learning and Research (VT-CDCLR)
- VA Office of Early Childhood Development (OECD)
- VA Department of Social Services (VDSS)
- VA Department of Education (VDOE)
- VA Department of Health (VDOH)
- VA Information Technology Agency (VITA)
- Virginia Early Childhood Foundation (VECF)
- Virginia Child Care Resource & Referral Network (VACCRRN)

Purpose: Build an interactive, integrated data system for early childhood across the state to guide program evaluation and policy analysis

Goals:

- Incorporate data from Departments of Social Services, Education, and Health
- Link up local with state-wide data
- Integrate data from independent organizations
- Make it user-friendly – adaptable for local policy
- Maintain data security and participant confidentiality

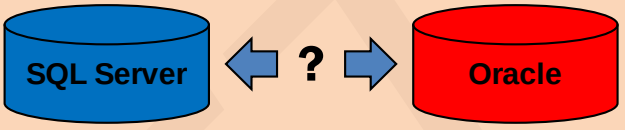
Proposed Application of Privacy Protecting Federated Model

- Currently submitting with VDOE, VITA, and VEC a proposal to US Department of Education Statewide Longitudinal Data Systems (SLDS) Grant Program
- Using the same model, will create a cross-agency federated data linking and reporting system between secondary education agencies, post-secondary education agencies, and numerous Virginia workforce development programs.

Impediments to Public Sector Multi-Agency Data Integration – "Saddle on a Sow"

Impediments Common to all Data Integration Efforts

- **Technological Heterogeneity**
  - Hardware Differences
  - Software Differences (DBMS)



- **Semantic Heterogeneity**
  - Differences in meaning, interpretation, or intended use of data
  - "Grade" in one education department data table may refer to a letter from the list "A,B,C,D,F" while in another data table the same word may refer to the year of current school attendance as in "8th Grade").

Student ID	Grade
XYZ123	A

=

Student ID	Grade
XYZ123	8

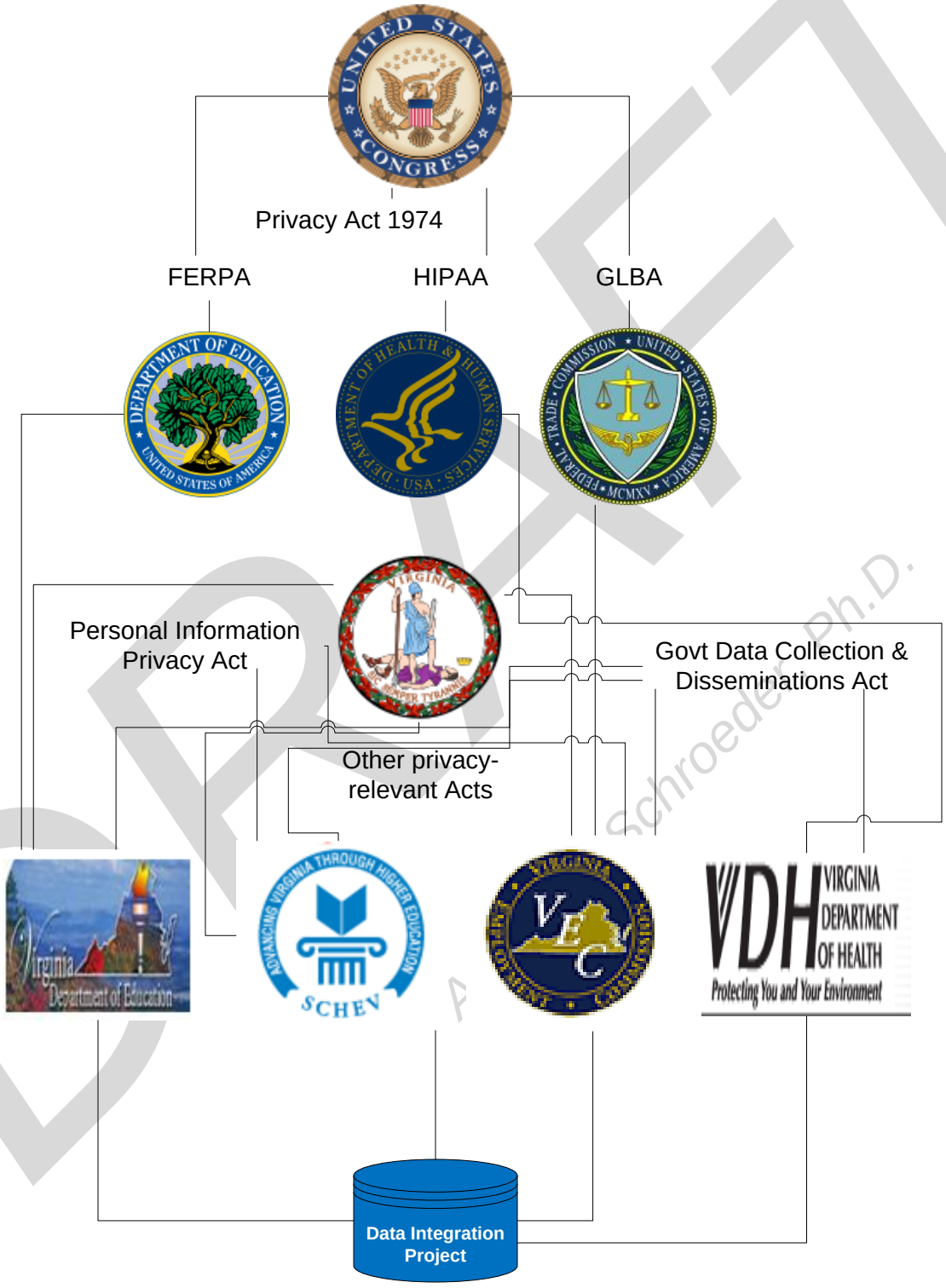
Additional Impediments in Public Sector Multi-Agency Integration Efforts

- **Regulatory Heterogeneity**
  - Multiple sets of statutory law at the federal and state levels (sometimes even local) – HIPAA, FERPA, GLBA, State Privacy Acts
  - Multiple agency interpretations of the statutory laws (regulatory law) – HHS, ED, FTC, State Agency Regulations

- **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 ("locally administered, state supervised")

Public Sector Complexity: Regulatory Heterogeneity

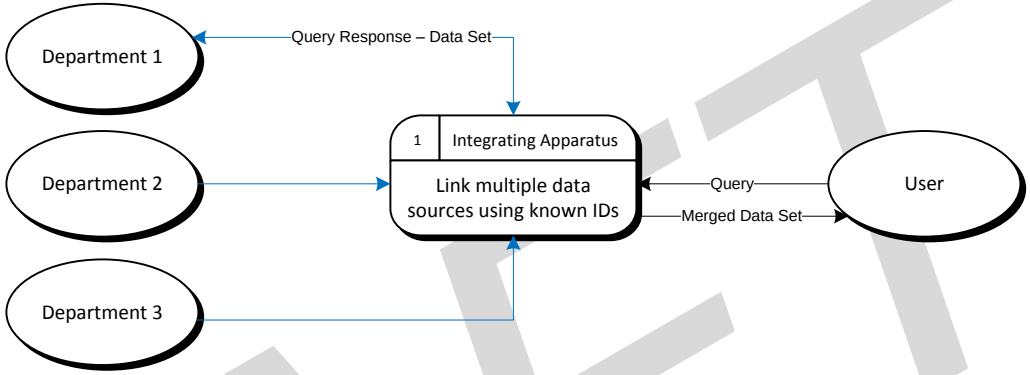
- Multiple levels of statutory law
- Multiple implementations of regulatory law at each level of statutory law
- Most conservative interpretation of regulatory law becomes de facto standard



With multiple sets of federal regulations, Virginia specific privacy laws, and a system of "state-supervised, locally administered public services," Virginia provides a case study in the difficulties of combining data from multiple agencies – "If it can work in Virginia, it can work anywhere"

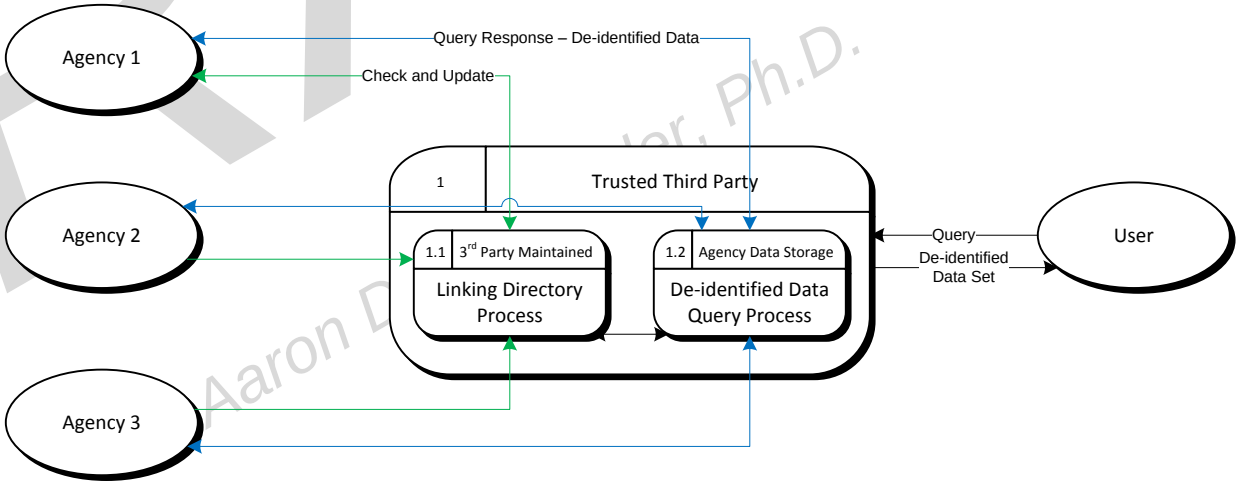
Federated Data System

- System that interacts with multiple data sources on the back-end and presents itself as a single data source on the front-end
- The key to linking up the different data sources is a central linking apparatus
- Generally a DBMS with a unique identifier-populated linking table



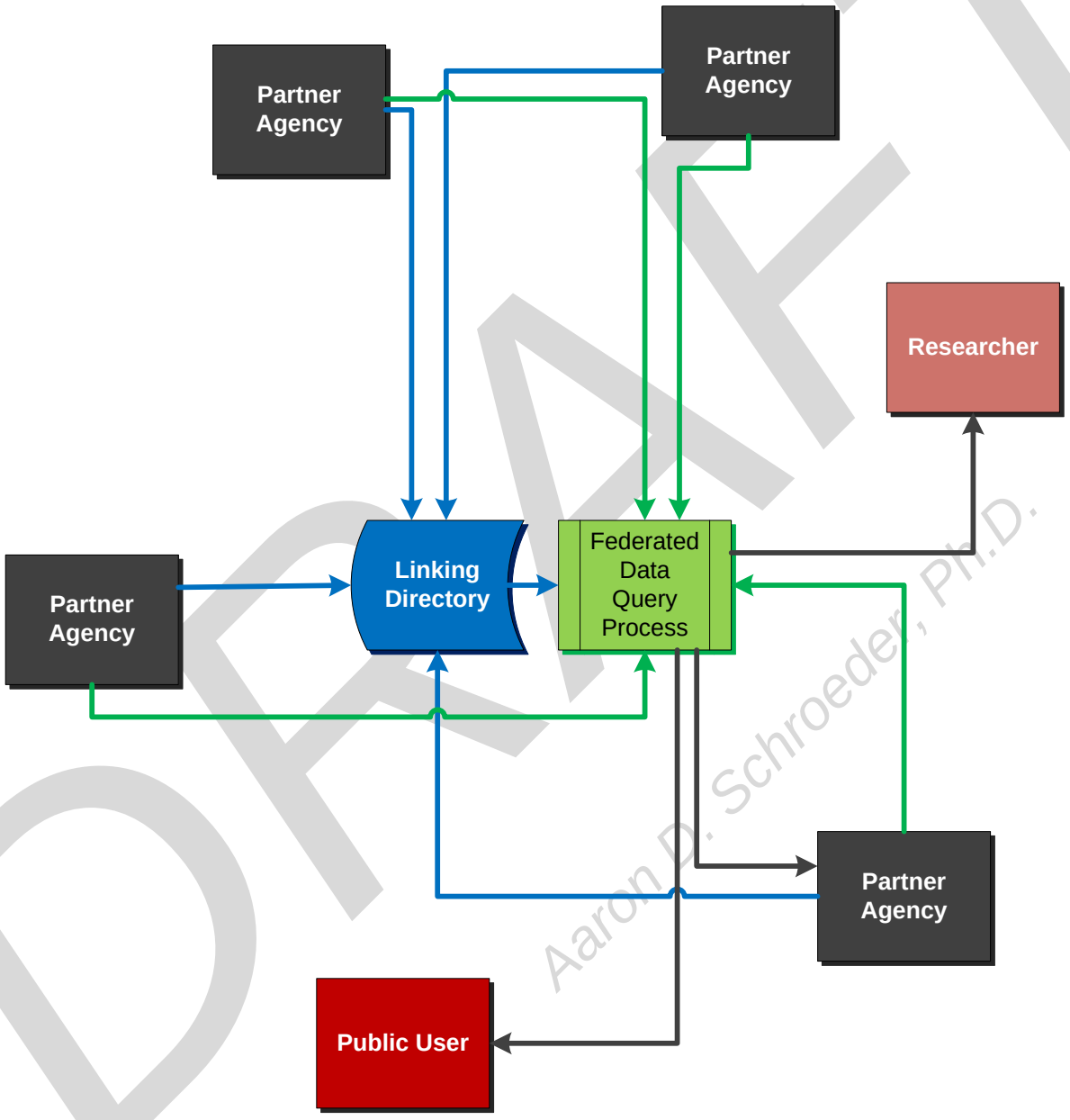
Privacy Protecting Federated Data System

- Need a system that will permit linking of data relevant to longitudinal research but does not allow personal identification of any of the individuals used in the data set
- Need to know how quality of day care effects "white males" – don't need to know how it effects Bill Jones.



Privacy Protecting Federated Model Operations - Overview

- Virginia's Interpretation - *No one person , inside or outside a government agency, should be able to create a set of identified linked data records between partner agencies*
- Crucial element, in terms of privacy protection, is the Linking Directory
- Linking Directory updating process links personal records in multiple agencies
- Linking Directory's design precludes identification of specific individuals within the resulting data set



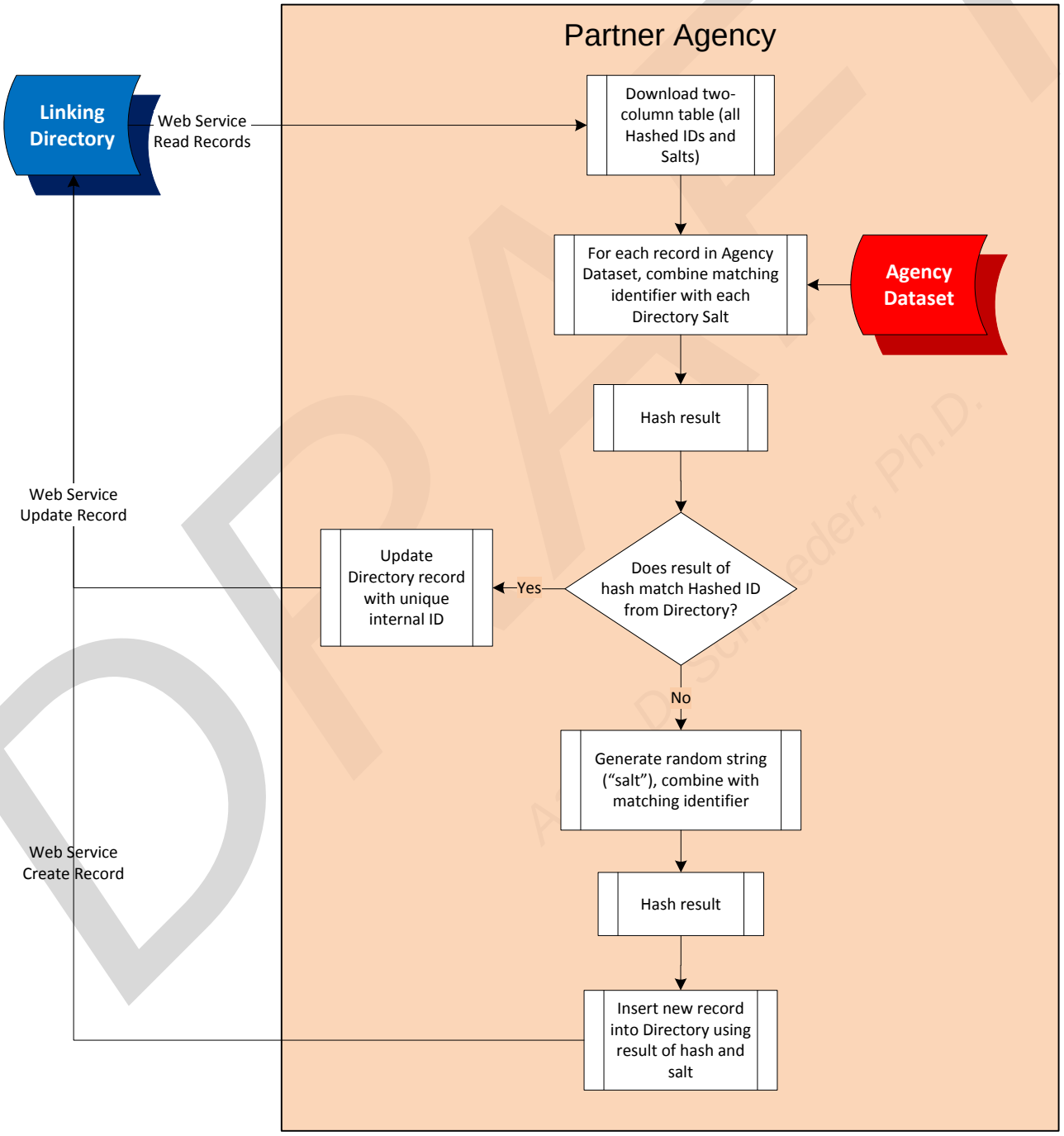
Individual partner agencies keep the LDS Directory updated on a periodic basis

Individual partner agencies provide de-identified data to the Federated Data Query Process on request via a web service

Queries can be submitted from the individual partner agencies, researchers, and the public. An authentication and authorization scheme is used to regulate levels of data access.

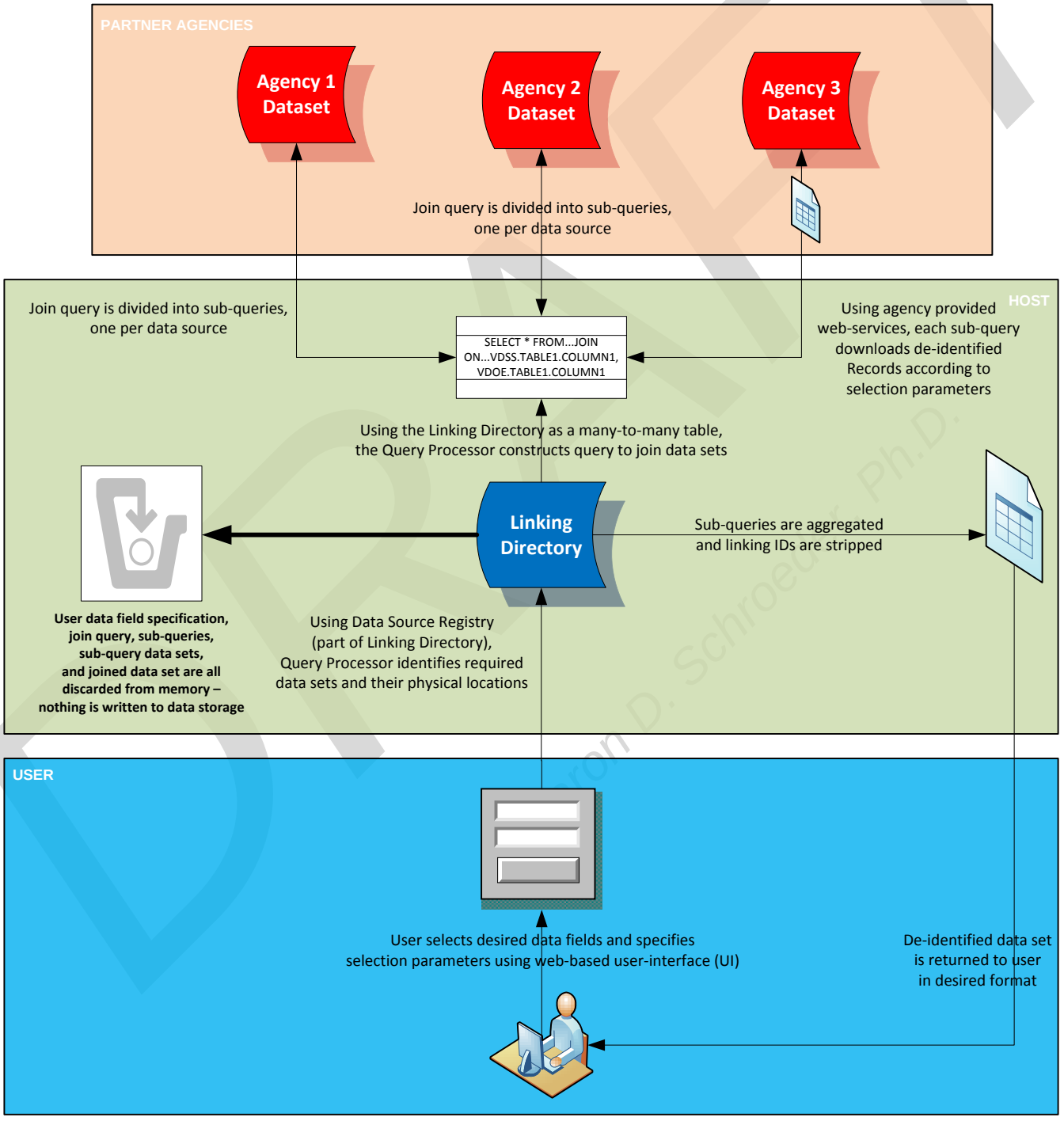
Linking Directory Update Process
(no un-hashed ID ever leaves the agency)

- Linking Directory contains a record for each individual served by the public agencies participating in the federated system
- Linking Directory contains only a one-way encrypted hash of the unique identifier(s) that can be used by multiple agencies for update purposes
- Linking Directory contains entry for each internal unique id used by each system for which the individual has an entry
- Internal unique ids are devoid of personally identifying information

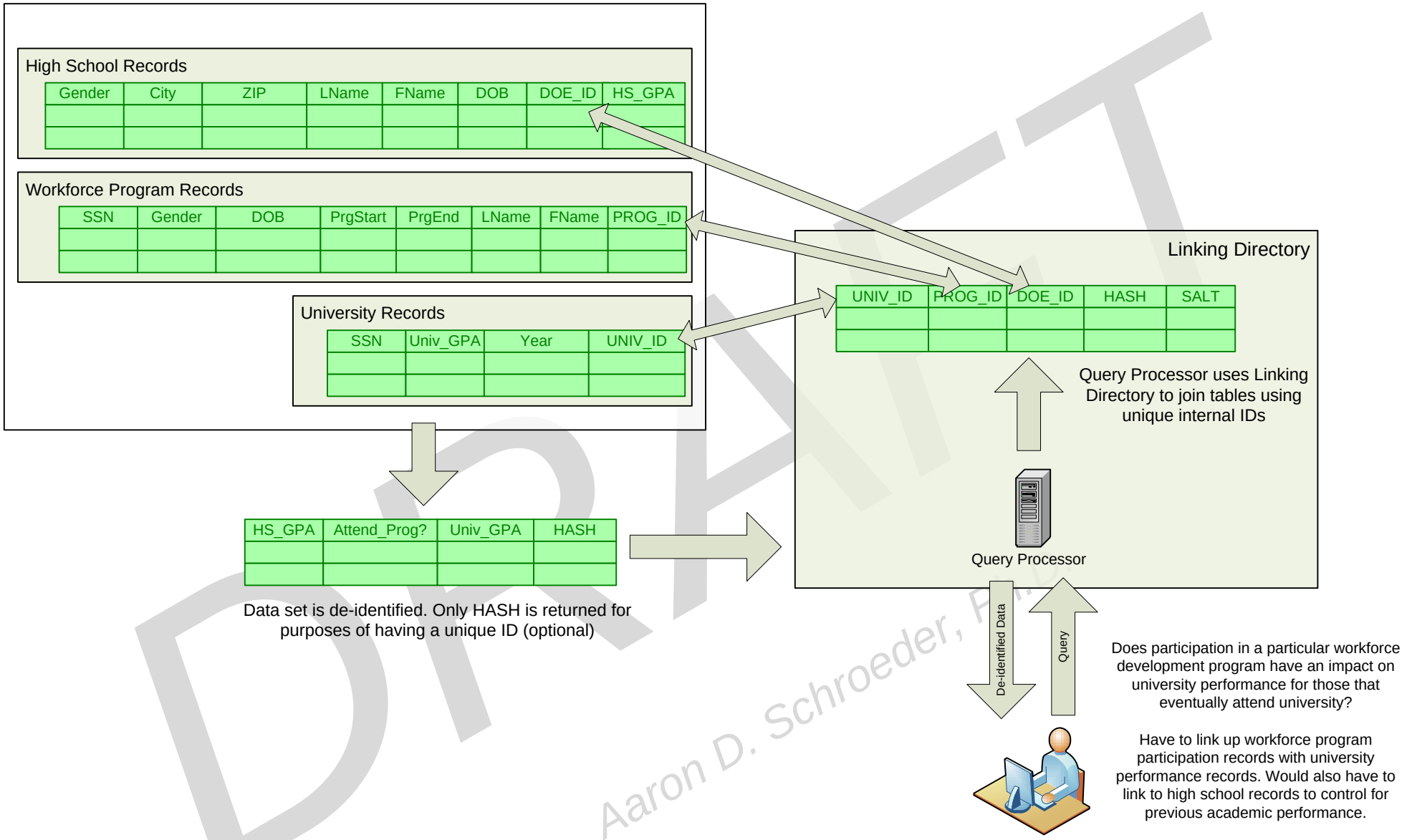


De-Identified Data Query Process

- Uses Linking Directory to anonymously join data records at the individual level from multiple data sources
- Uses the internal identifiers used by each source
- No resulting data sets are permanently stored.



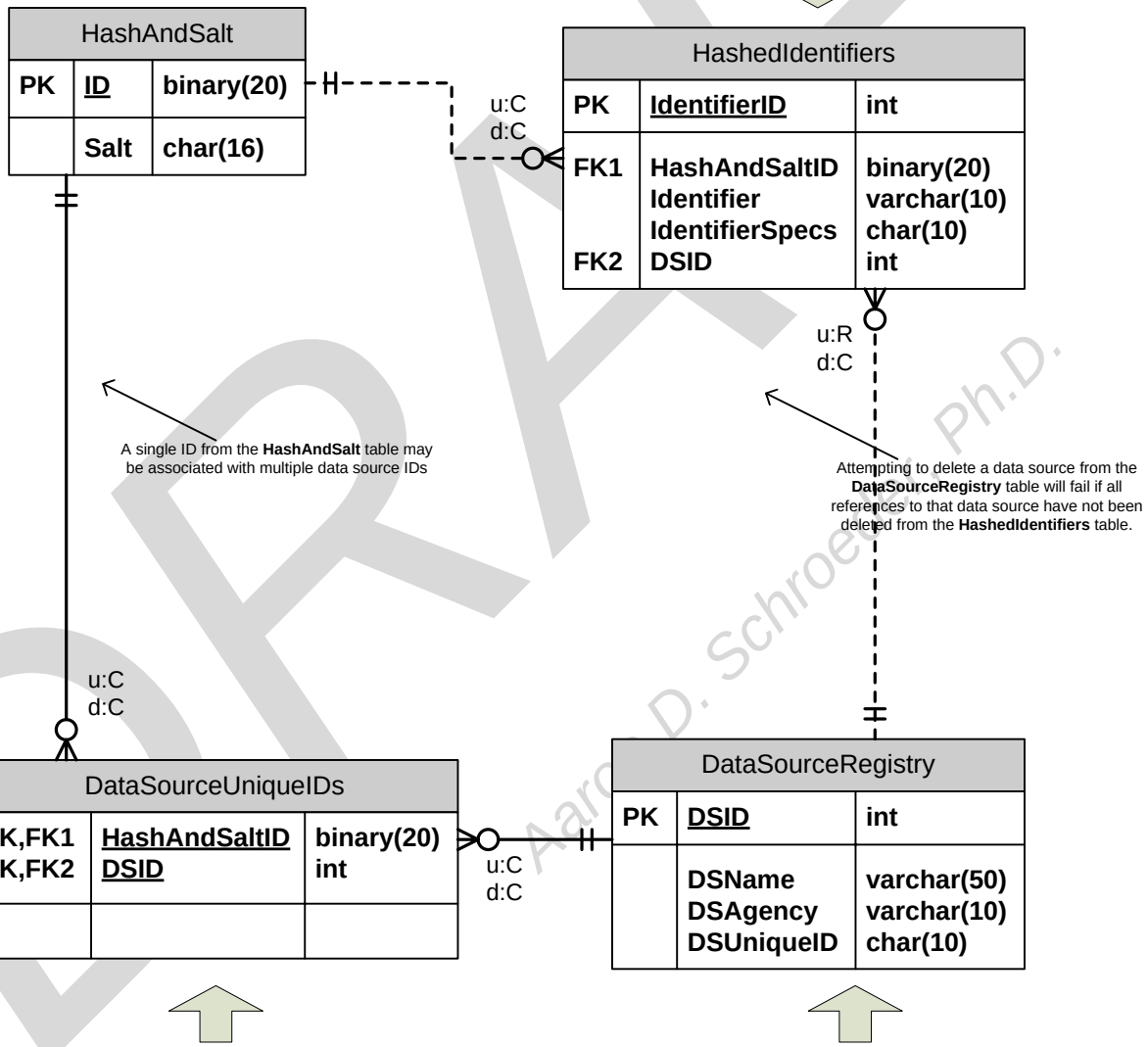
Query Example - Linking Secondary, Post-Secondary and Workforce Data Sets
(Linking Directory is shown as a simple flat table for conceptual clarity)



Linking Directory Data Model
(Linking mechanism between agency data sources – including meta-data repository)

The HashAndSalt table stores both the hashed ID created by one of the partner agencies and the Salt used in the hashed ID's creation. Storing the Salt allows other partner agencies to determine if an entry already exists for a client, or if a new entry needs to be created. This is accomplished by combining the SALT with the same identifier types (as specified in HashedIdentifiers) used to create the original hash.

The HashAndSaltID can be constructed from the concatenation of multiple identifiers (e.g. if SSN is not available, some combination of Age, Birth Date, School District, etc., may be required). The **HashedIdentifiers** table stores all identifiers used to construct each individual HashAndSaltID



Many-to-many table linking the **HashAndSalt** table with the **DataSourceRegistry** table.

Each agency data source gets a single entry in the **DataSourceRegistry** table and includes the internal unique ID used by that data source (e.g. a StudentID used in a state DOE data system).