Research Improving People's Lives (RIPL):

Unlocking data and science to create better policy and outcomes

RIPL works with government to help them **use data, science, and technology to improve policy and lives.**

We partner government with science experts to **provide powerful, scientific-grade insights at the speed of policy directly with policymakers** using **the best existing data and science methodologies** to solve policy problems.

We build government capacity to partner broadly around data, science, and technology so that government can **lead and support a community of fact-based policy innovation and improvement.**

Our platform, the Data-Driven Policy System (DDPS), includes the following components:

Government-owned data integration

Cloud-based secure environment; FedRAMP approved; automated logging, monitoring.

Combines anonymized data across agencies and policy domains.

Government controls user access: dual authentication, all exports must be approved.

Governance structure for data use to facilitate external and internal research partnerships.

Scientific-grade insights applied to policy

Projects ideas start with government priorities; projects are designed in collaboration with policy leaders.

Program officers and external researchers bring scientific expertise to the analysis.

Data-driven scientific insights with impact; science translates directly to measurably improve policy.

Research at the speed of policy

Projects oriented towards policy improvements and solutions to real-world problems.

High public and stakeholder engagement in research design and policy application processes.

Research communications

Policy briefs written for general audience (see *Hamilton Project paper on Fact-Based Policy*).

Data and research communications at local and national forums (see *Rhode2College*).

Open source amplifies impact

Research-and insights-generating code is open-source on GitHub for any government to adopt or adapt.

Open-source code empowers government improvement with reliable, transparent, and reproducible insights.

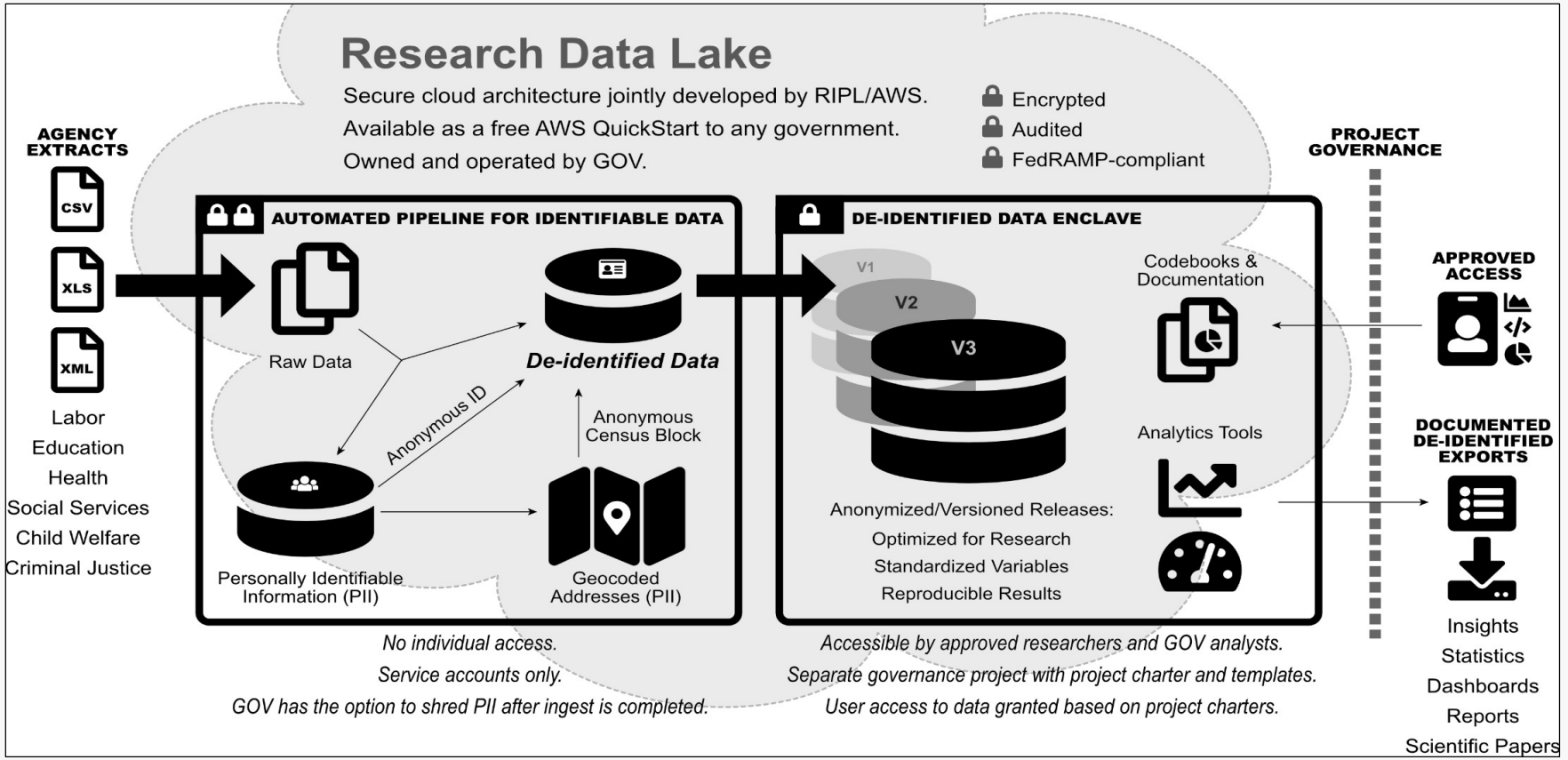
Science-based metrics and dashboards for the community to better use public services.

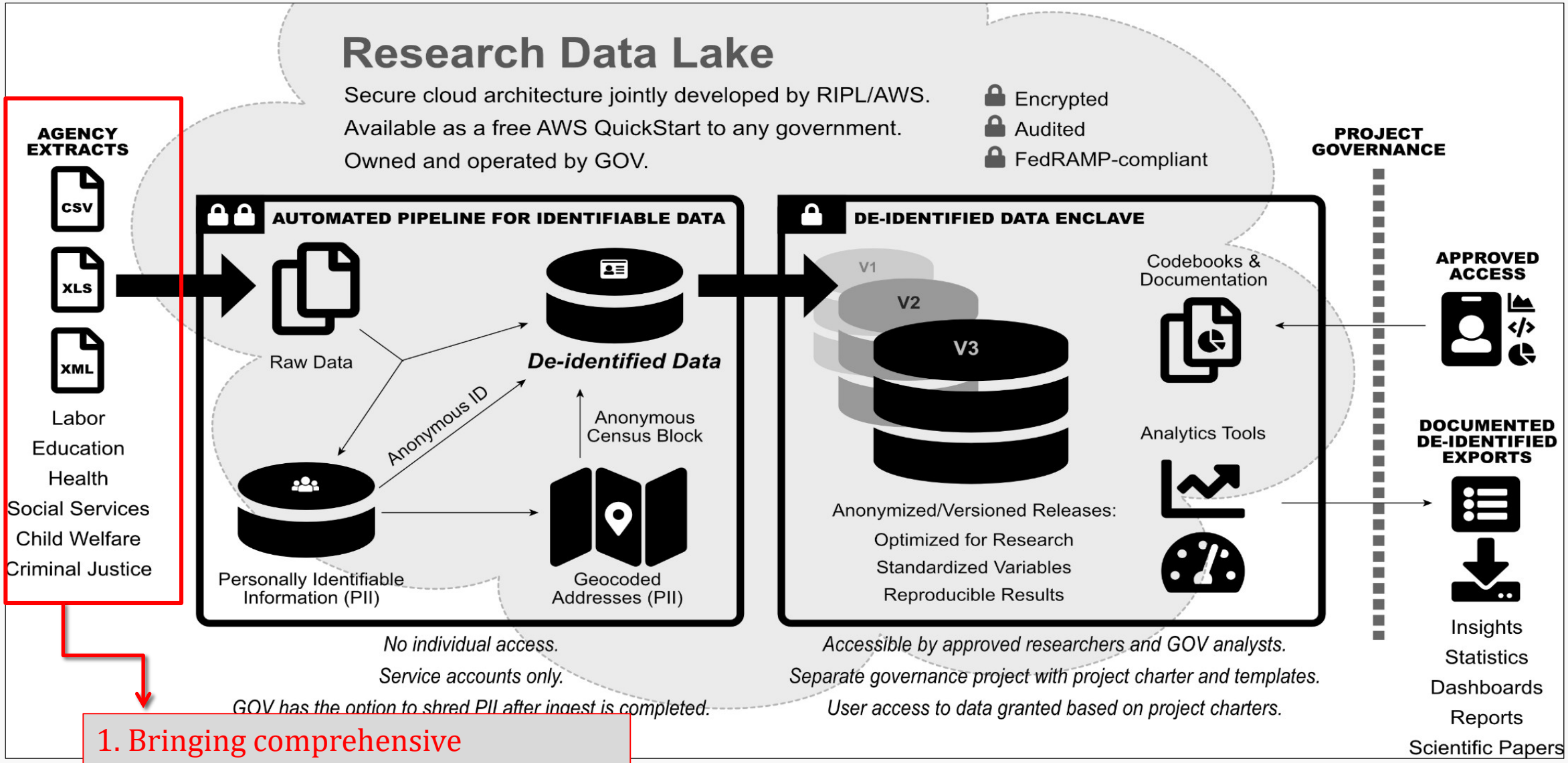
The foundation of DDPS is a Research Data Lake

What is a Research Data Lake (RDL)?

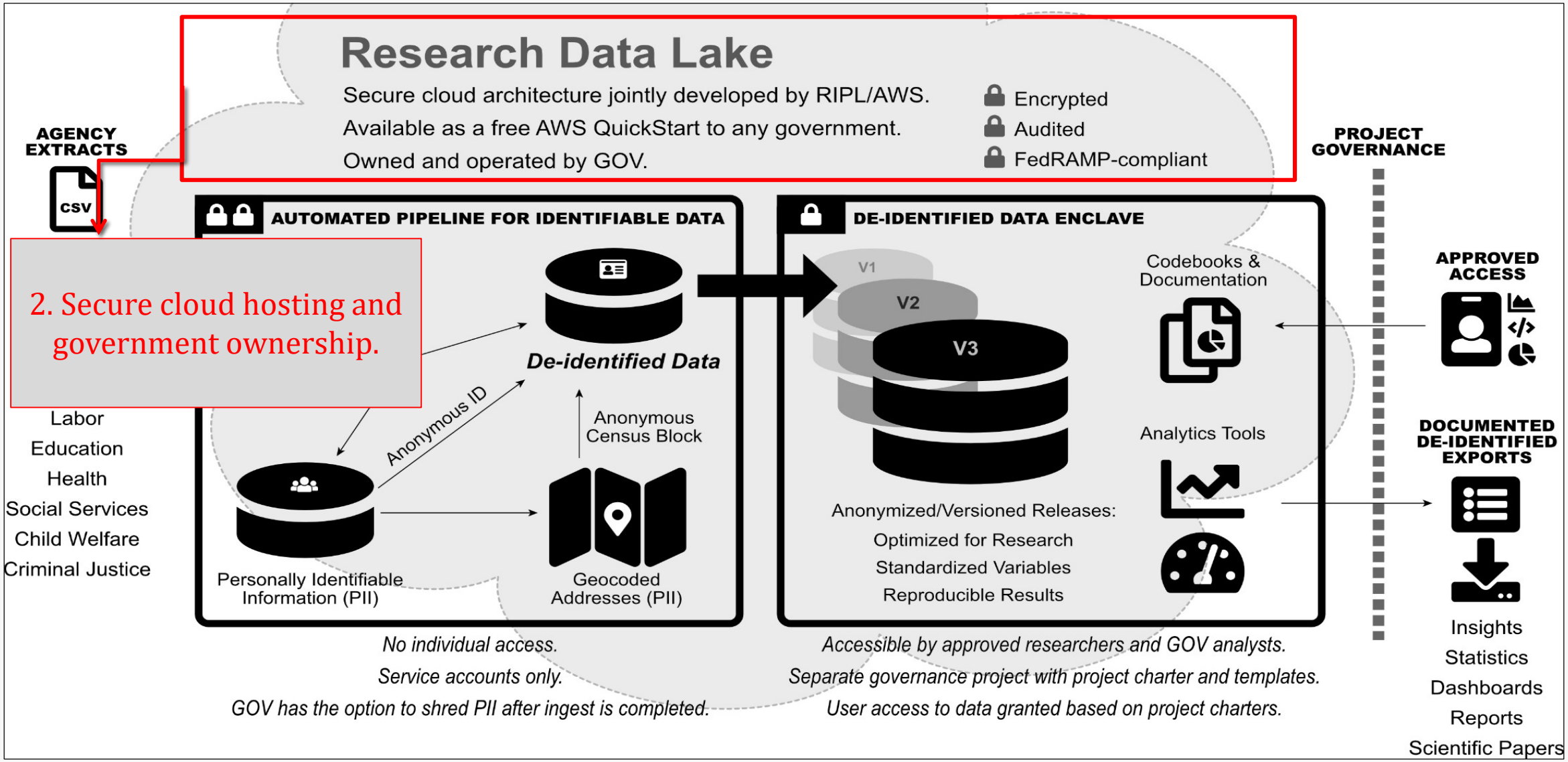
- A secure, cloud-based system that houses and integrates administrative data
- Owned/managed by government for transparency and capacity building
- Provides identity resolution and anonymization
- Automated codebooks and derived tables accelerate new projects and support consistent, robust, and reliable analysis
- Modules for “science-as-service” enable government to partner with and lead a community of evidence-based improvement
- Extensible to meet changing needs and for crisis response

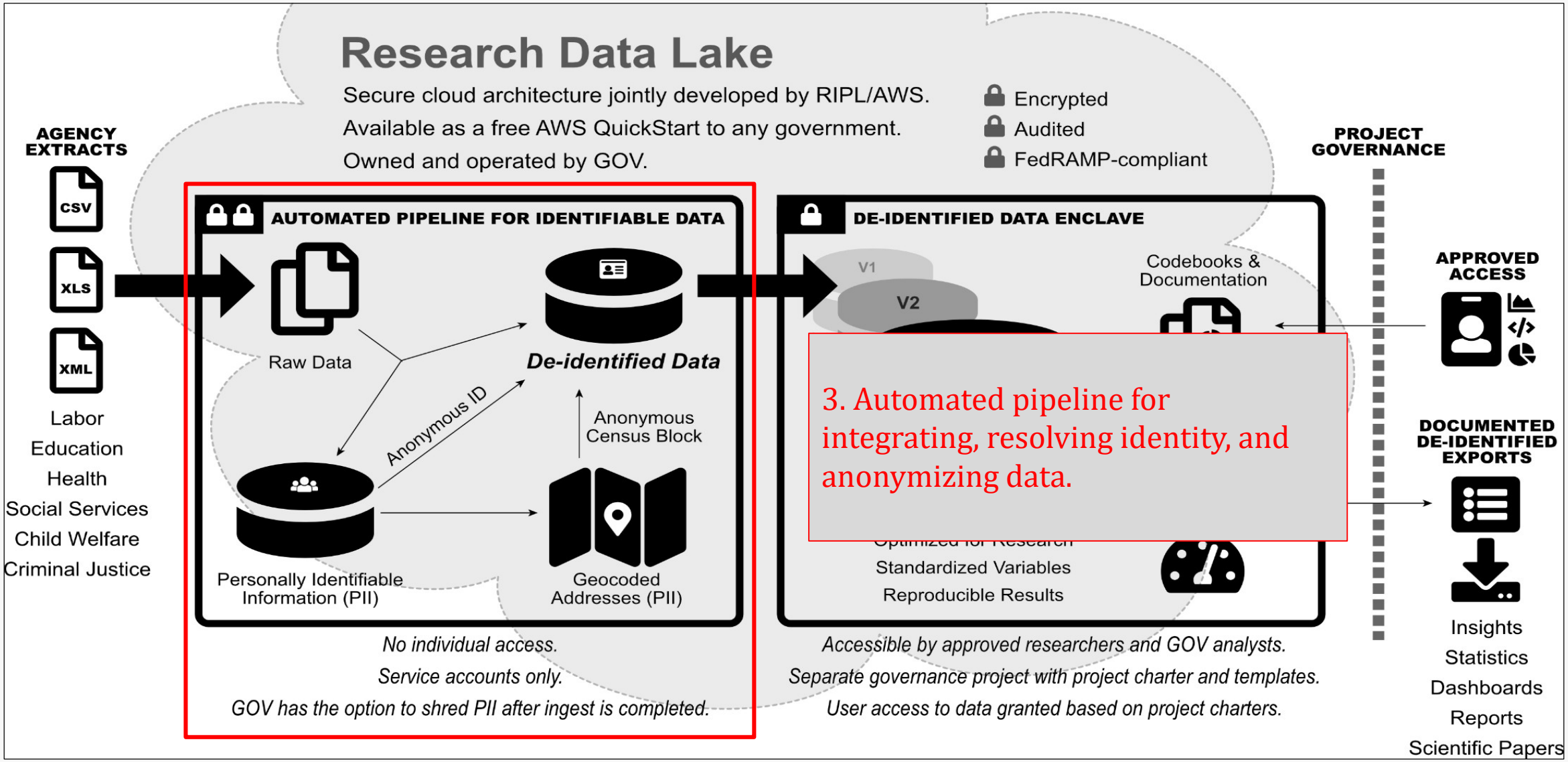
Overview of a Research Data Lake

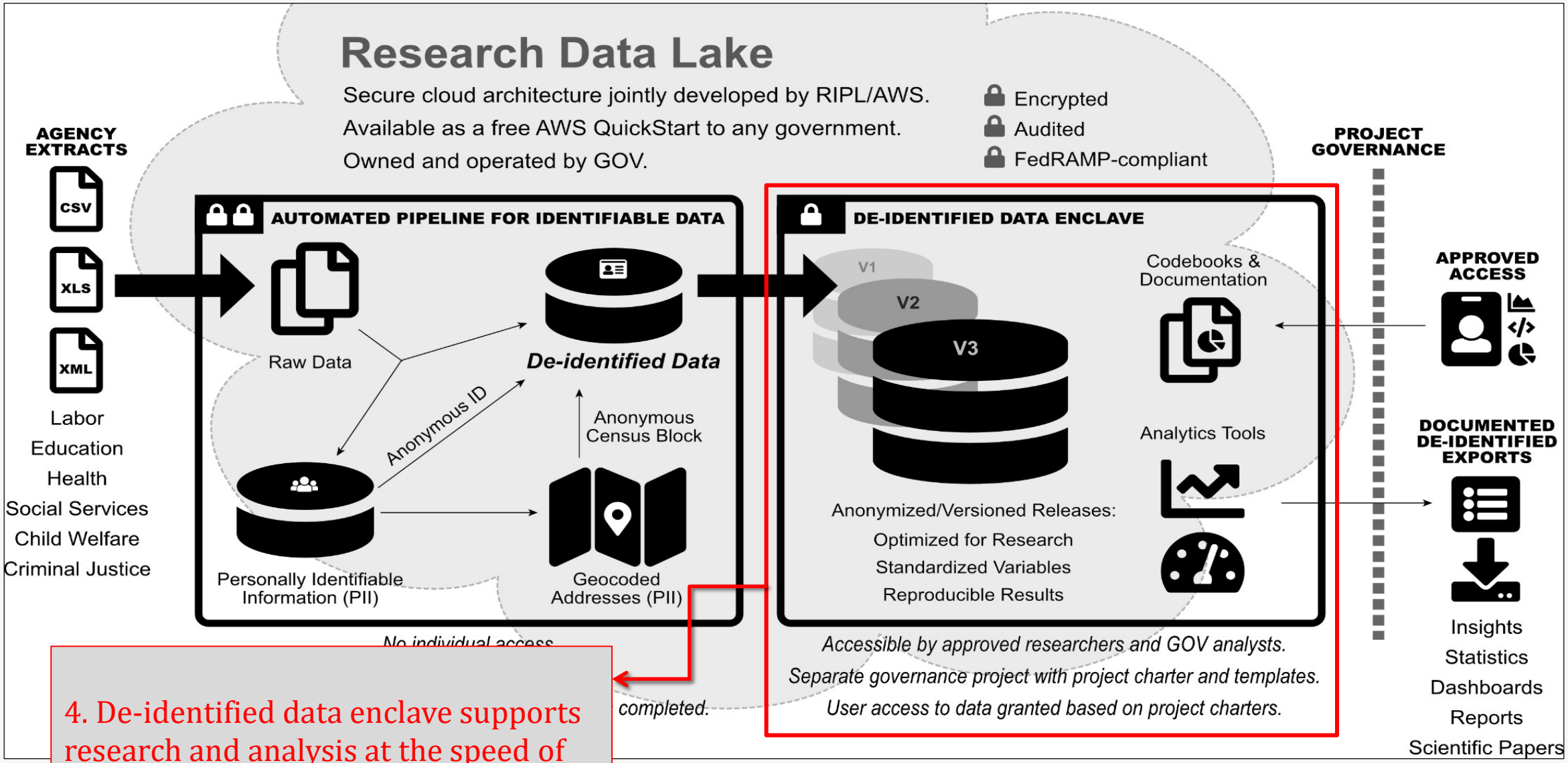




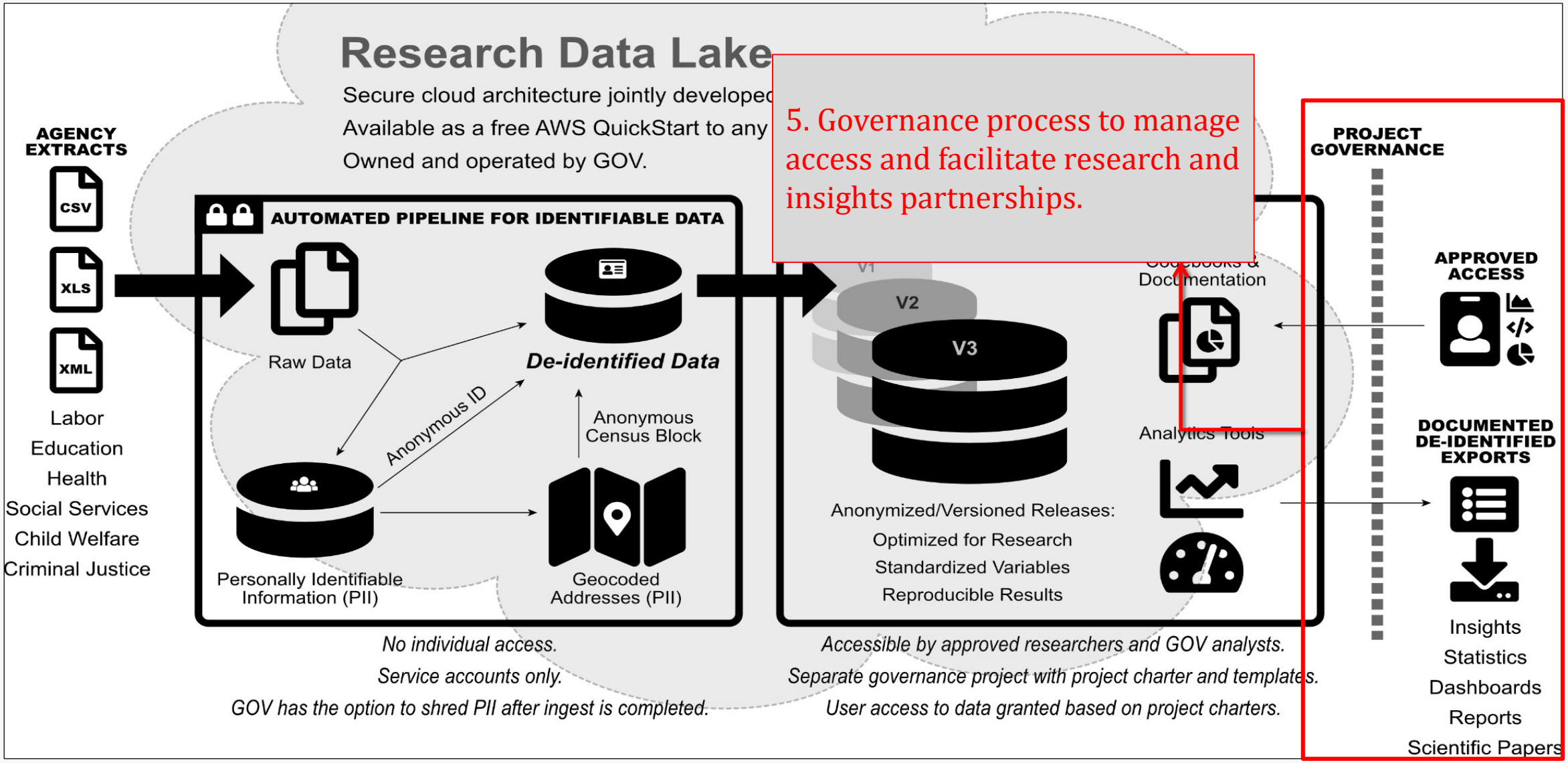
1. Bringing comprehensive administrative data into the system for 360-degree insights.


Research Data Lake
Secure cloud architecture jointly developed by RIPL/AWS.
Available as a free AWS QuickStart to any government.
Owned and operated by GOV.
Encrypted
Audited
FedRAMP-compliant
AGENCY EXTRACTS
csv
2. Secure cloud hosting and government ownership.
Labor
Education
Health
Social Services
Child Welfare
Criminal Justice
AUTOMATED PIPELINE FOR IDENTIFIABLE DATA
De-identified Data
Anonymous ID
Anonymous Census Block
Personally Identifiable Information (PII)
Geocoded Addresses (PII)
No individual access.
Service accounts only.
GOV has the option to shred PII after ingest is completed.
DE-IDENTIFIED DATA ENCLAVE
V1
V2
V3
Anonymized/Versioned Releases:
Optimized for Research
Standardized Variables
Reproducible Results
Codebooks & Documentation
Analytics Tools
Accessible by approved researchers and GOV analysts.
Separate governance project with project charter and templates.
User access to data granted based on project charters.
PROJECT GOVERNANCE
APPROVED ACCESS
DOCUMENTED DE-IDENTIFIED EXPORTS
Insights
Statistics
Dashboards
Reports
Scientific Papers


Research Data Lake
Secure cloud architecture jointly developed by RIPL/AWS.
Available as a free AWS QuickStart to any government.
Owned and operated by GOV.
Encrypted
Audited
FedRAMP-compliant
AGENCY EXTRACTS
CSV
XLS
XML
Labor
Education
Health
Social Services
Child Welfare
Criminal Justice
AUTOMATED PIPELINE FOR IDENTIFIABLE DATA
Raw Data
De-identified Data
Anonymous ID
Anonymous Census Block
Personally Identifiable Information (PII)
Geocoded Addresses (PII)
No individual access.
Service accounts only.
GOV has the option to shred PII after ingest is completed.
DE-IDENTIFIED DATA ENCLAVE
V1
V2
Codebooks & Documentation
Standardized Variables
Reproducible Results
Accessible by approved researchers and GOV analysts.
Separate governance project with project charter and templates.
User access to data granted based on project charters.
PROJECT GOVERNANCE
APPROVED ACCESS
DOCUMENTED DE-IDENTIFIED EXPORTS
Insights
Statistics
Dashboards
Reports
Scientific Papers
3. Automated pipeline for integrating, resolving identity, and anonymizing data.

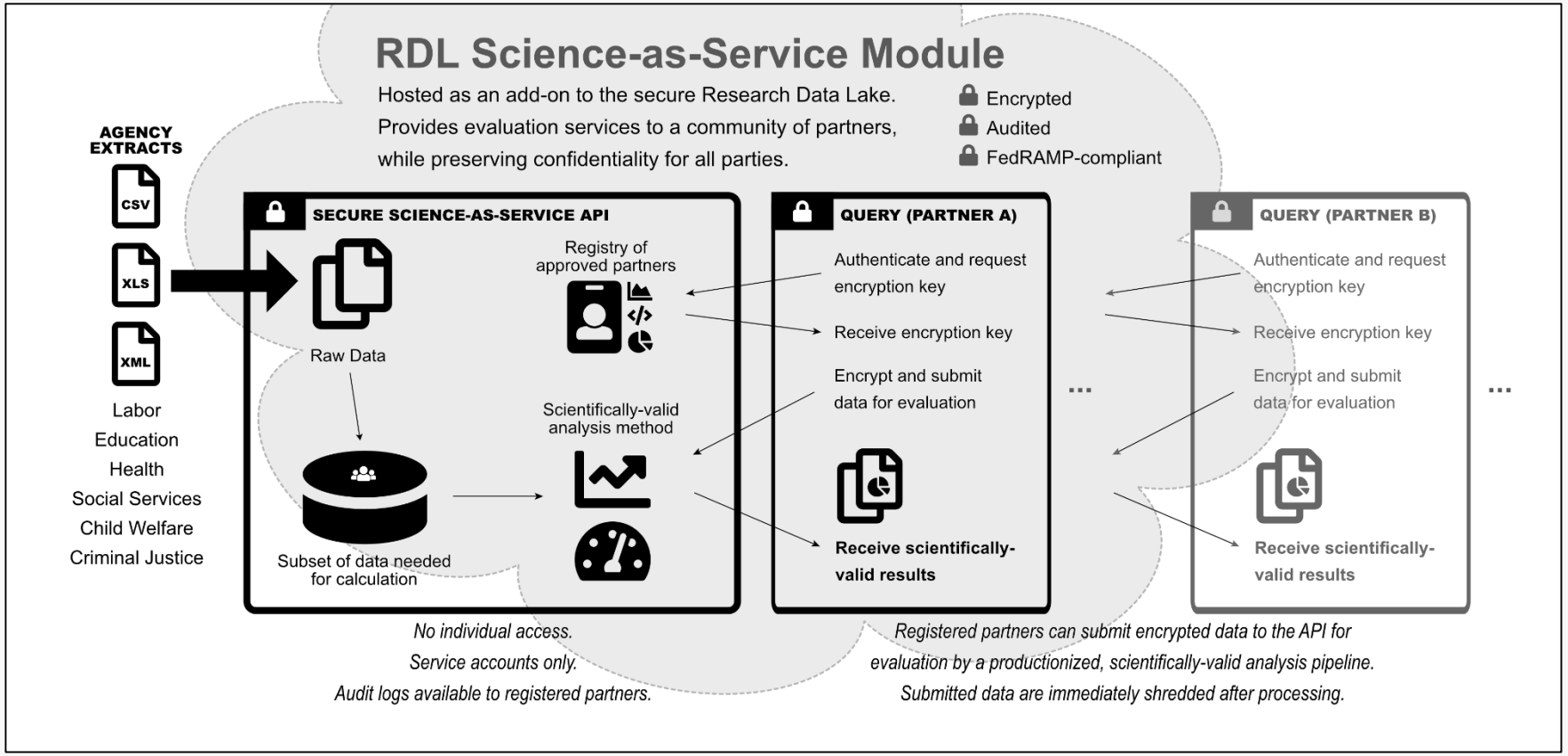


4. De-identified data enclave supports research and analysis at the speed of policy with scalable computing.


Research Data Lake
Secure cloud architecture jointly developed
Available as a free AWS QuickStart to any
Owned and operated by GOV.
5. Governance process to manage access and facilitate research and insights partnerships.
AGENCY EXTRACTS
CSV
XLS
XML
Labor
Education
Health
Social Services
Child Welfare
Criminal Justice
AUTOMATED PIPELINE FOR IDENTIFIABLE DATA
Raw Data
De-identified Data
Anonymous ID
Anonymous Census Block
Personally Identifiable Information (PII)
Geocoded Addresses (PII)
No individual access.
Service accounts only.
GOV has the option to shred PII after ingest is completed.
V1
V2
V3
Anonymized/Versioned Releases:
Optimized for Research
Standardized Variables
Reproducible Results
Codebooks & Documentation
Analytics Tools
Accessible by approved researchers and GOV analysts.
Separate governance project with project charter and templates.
User access to data granted based on project charters.
PROJECT GOVERNANCE
APPROVED ACCESS
DOCUMENTED DE-IDENTIFIED EXPORTS
Insights
Statistics
Dashboards
Reports
Scientific Papers

“Science-as-service” module enables data-driven program improvement with community partners

- Provides a secure way for external programs to evaluate outcomes with government administrative data.
- Partners query scientifically-valid results through an Application Programming Interface (API).
- All parties maintain confidentiality of data.
- Turns scientific results into sustainable impact.

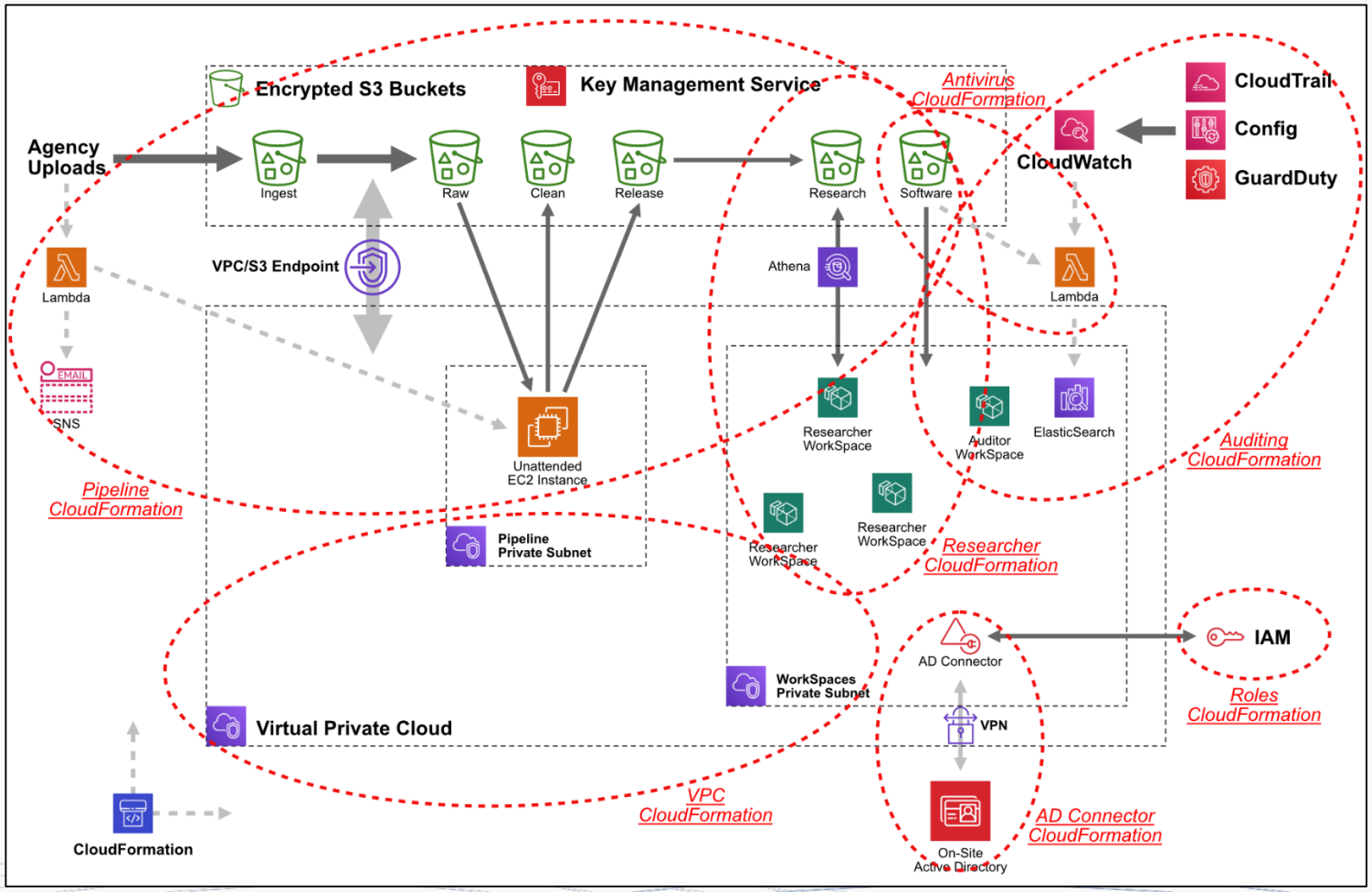


An RDL can be deployed securely in the cloud

- Deployable in Amazon Web Services (AWS) **by any government using freely-available CloudFormation templates.**
- The cloud infrastructure is **FedRAMP-approved** and designed to be **FISMA, HIPAA, and FERPA-compliant.**
- Data **never leave** government custody.
- Government **always controls who accesses data**, down to the field and variable level.
- Best-in-class firewall, encryption, and auditing tools **keep data safe.**
- **Pay-as-you-go model** means that government pays only for computing time and storage **it actually uses.**

Modular architecture meets the needs of any government for an efficient deployment

- The architecture is scalable and divided into seven modules (red dashed ovals).
- Each module is freely available as a customizable AWS CloudFormation template.
- This modular approach meets the needs of a wide range of state and local government partners, wherever they may be in terms of data warehousing and integration.



An RDL transforms government data to build the key features needed to measure success

- Our software helps Government analysts standardize their data and create variables that researchers can **begin using immediately.**
- It then generates **automated codebooks** for each dataset. Codebooks contain descriptive statistics that help researchers easily **understand data at a glance**.
- Codebooks are **automatically updated and monitored** for data quality issues and contain descriptive statistics.
- Supports research with modern, scalable computing. Partnerships with science teams like RIPL **seamlessly enable predictive modeling, causal analysis, and forecasting** which can then flow directly into improved public policy.

Example of a codebook for individual-level wage data.

Codebook for WageIndividual.csv

9 variables x **19,122,070** observations

Index	Variable	Type	%Missing	#Distinct					
0	**SIRAD_ID**	Numeric	NONE	*1,123,873*	**Min**	**25th**	**50th**	**75th**	**Max**
					1	330765	661969	993701	1325275
1	**YYQ**	Numeric	NONE	*40*	**Min**	**25th**	**50th**	**75th**	**Max**
					101	123	151	173	194
2	**WAGES**	Numeric	NONE	*149,498*	**Min**	**25th**	**50th**	**75th**	**Max**
					0	3744	8263	14570	Suppressed
3	**NAICS4**	Numeric	3.90%	*305*	**Min**	**25th**	**50th**	**75th**	**Max**
					1111	4482	5613	6231	9261
4	**NAICS2**	Numeric	3.90%	*24*	**Min**	**25th**	**50th**	**75th**	**Max**
					11	44	56	62	92
5	**INDUSTRY**	Categorical	NONE	*19*	**Example Values**			**Count**	**Frequency**
					Health			3,144,524	*16.4%*
					Retail			1,985,287	*10.4%*
					...				
					Utilities			50,638	*0.3%*
					Agriculture/Mining			44,846	*0.2%*
6	**EMPLOYER_SIZE**	Numeric	NONE	*2,174*	**Min**	**25th**	**50th**	**75th**	**Max**
					1	28	134	739	Suppressed
7	**EMPLOYERS**	Numeric	NONE	*13*	**Min**	**25th**	**50th**	**75th**	**Max**
					1	1	1	1	13
8	**INDUSTRIES**	Numeric	NONE	*6*	**Min**	**25th**	**50th**	**75th**	**Max**
					1	1	1	1	6

An RDL lets policymakers immediately begin deriving research insights

- Anonymized datasets are transformed into **derived tables** that turn operational data into variables needed for research.
- Derived tables contain **standardized variables** most frequently used for research and insights.

GLOBAL_ID	MONTH	YR	AGE	RACE	SEX	MEDICAID_ENROLLED	MEDICAID_PAYMENTS	SNAP_ENROLLED	SNAP_PAYMENTS	UI_EXHAUSTED
123456	201810	2018	28	Hispanic	F	0	(null)	1	204	0
123456	201811	2018	28	Hispanic	F	0	(null)	1	204	0
123456	201812	2018	28	Hispanic	F	1	129.85	1	204	0
123456	201901	2019	28	Hispanic	F	1	72.45	1	204	0
123456	201902	2019	28	Hispanic	F	1	(null)	1	232	1
123456	201903	2019	28	Hispanic	F	1	512.67	1	232	1
123456	201904	2019	29	Hispanic	F	1	28	1	232	1
123456	201905	2019	29	Hispanic	F	1	50.1	1	232	1

Research Data Lake Master Table: Important health, social services, and labor variables are joined on a person-month level (example uses synthetic data).

Health, labor, and social services records are joined in a **master table.**

GLOBAL_ID	MONTH	YR	AGE	RACE	SEX	MEDICAID_ENROLLED	MEDICAID_PAYMENTS	SNAP_ENROLLED	SNAP_PAYMENTS	UI_EXHAUSTED
123456	201810	2018	28	Hispanic	F	0	(null)	1	204	0
123456	201811	2018	28	Hispanic	F	0	(null)	1	204	0
123456	201812	2018	28	Hispanic	F	1	129.85	1	204	0
123456	201901	2019	28	Hispanic	F	1	72.45	1	204	0
123456	201902	2019	28	Hispanic	F	1	(null)	1	232	1
123456	201903	2019	28	Hispanic	F	1	512.67	1	232	1
123456	201904	2019	29	Hispanic	F	1	28	1	232	1
123456	201905	2019	29	Hispanic	F	1	50.1	1	232	1

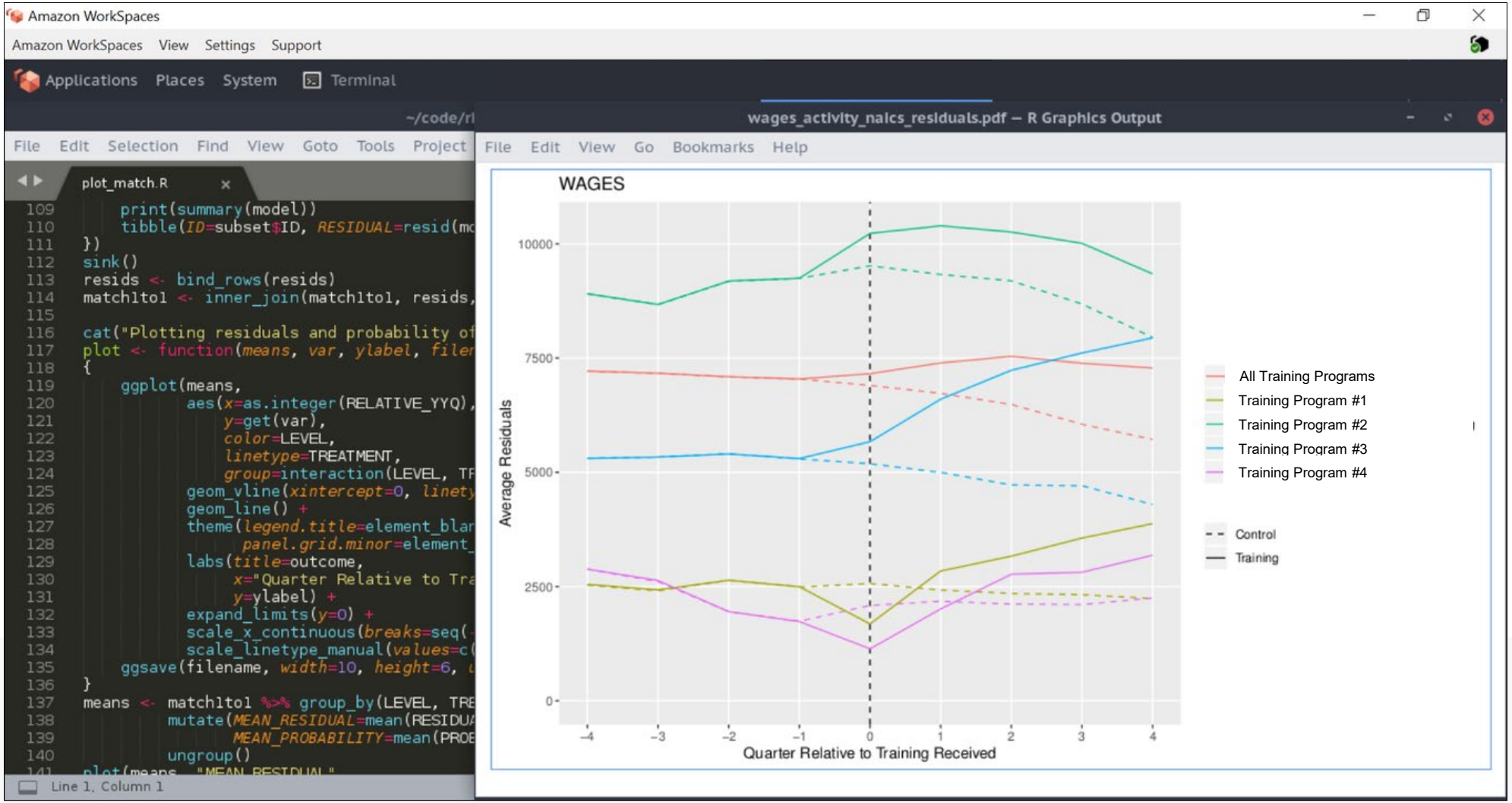
The same individual, represented by an anonymous ID, can be joined to education test data…

GLOBAL_ID	RECORD_ID	BEG_YEAR	SCH_CODE	GRADE	PSAT_SCORE	SAT_SCORE	FEMALE	WHITE	BLACK	HISPANIC	TWO_ETHNIC	AM_IND
123456	666000	2008	43210	10	1420	(null)	1	0	0	1	0	0
123456	666001	2009	43210	11	(null)	1540	1	0	0	1	0	0
123456	666002	2010	43300	12	(null)	(null)	1	0	0	1	0	0

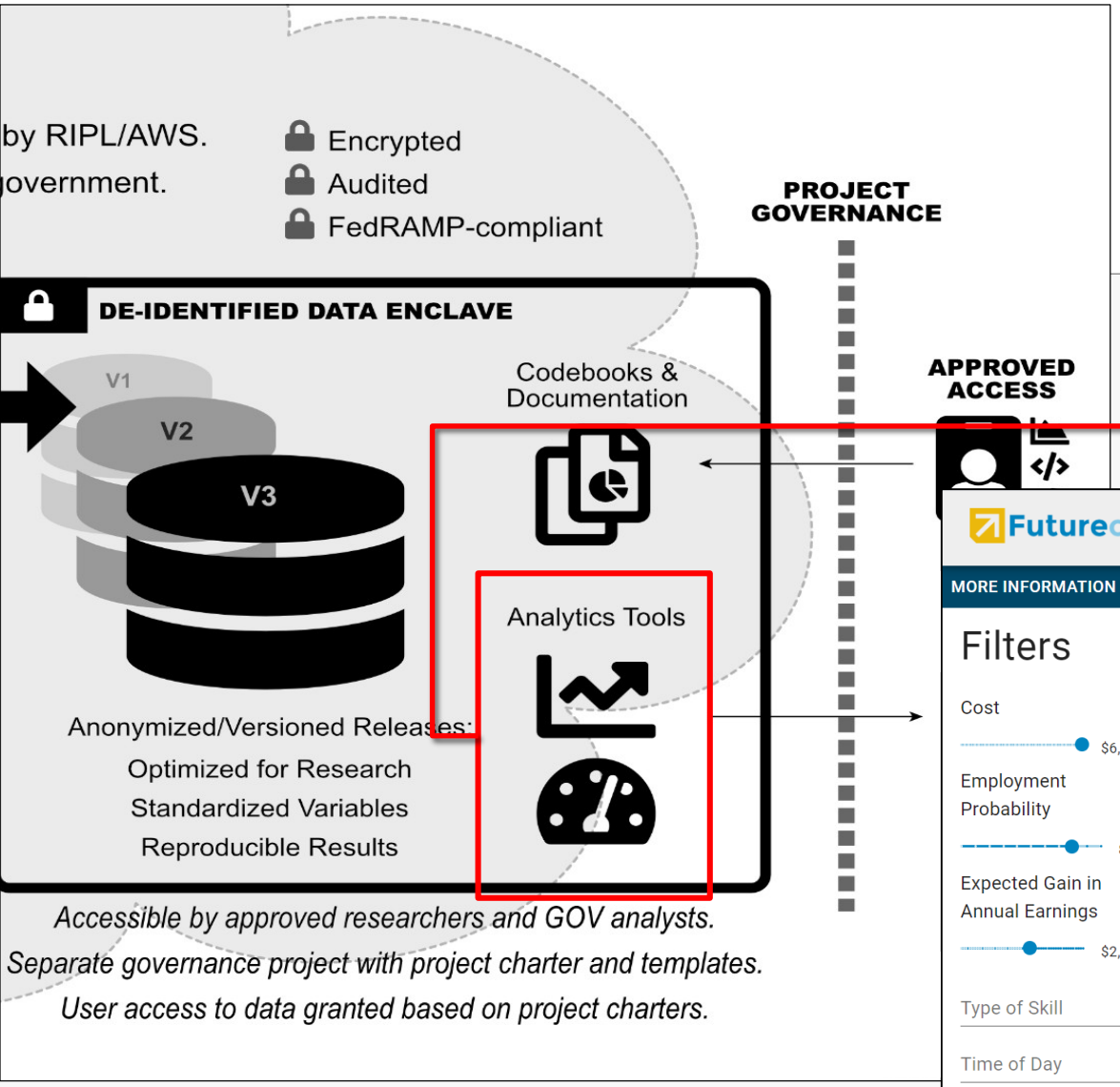
…and prior school enrollment on a person-month level.

GLOBAL_ID	RECORD_ID	DIST_CODE	SCH_CODE	DOB	RACE	ENROLL_DATE	BEG_YEAR	GRADE	IEP	LEP	FREE_LUNCH	HOMELESS
123456	999444	34	43210	11-May-92	Hispanic	24-Sep-08	2008	10	0	1	1	1
123456	999445	34	43210	11-May-92	Hispanic	30-Aug-09	2009	11	0	1	1	0
123456	999446	34	43300	11-May-92	Hispanic	1-Sep-10	2010	12	0	1	1	0

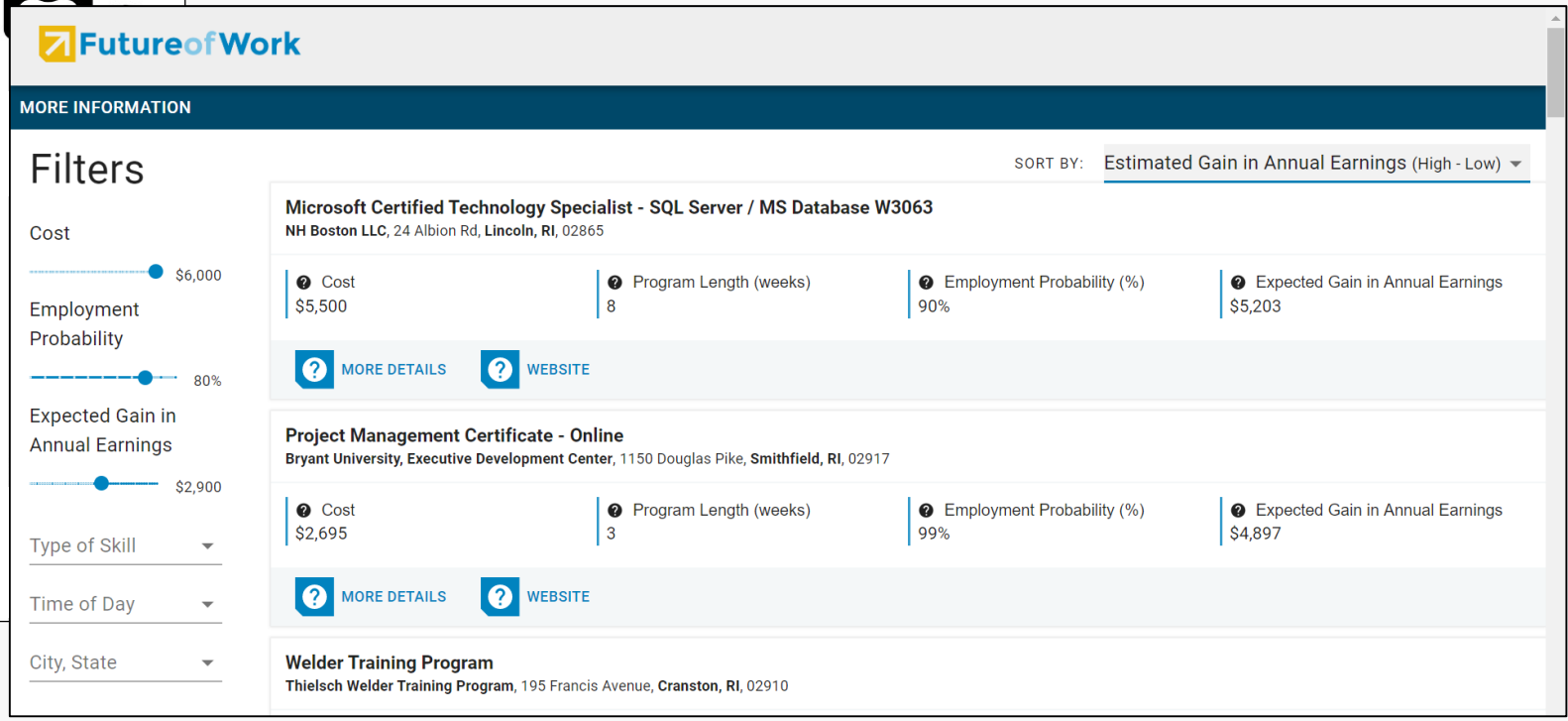
An RDL enables scientific-grade insights



Screenshot of a causal analysis in an RDL, showing average regression-adjusted wages received by workers in labor training programs in the quarters before and after receiving training in a matched treatment-control framework.



Causal results are published through an API to empower workers with information for wise reskilling decisions.



Screenshot of a prototype web/mobile interface for sorting training programs on expected gain in earnings, employment probability, and cost.

RDL users can join a dynamic learning community of like-minded policymakers, minimizing the need to "reinvent the wheel"

- Shared and open-source knowledge base **for disseminating best practices, analysis code, predictive models, dashboards and applications** across the DDPS platform.
- Implement **off-the-shelf analyses** and **test the impact of policy interventions**.
- **Extend existing solutions** to fit your needs.
- **Share your work with** others to replicate.
- Government always decides **what to share and when.**

Example RDL extension: Pivoting to support emergency UI claims in 10 days

- Rhode Island was able to reach out to RIPL **in crisis of anticipated failure of their legacy Unemployment Insurance (UI) system** under the incoming deluge of COVID-19 pandemic UI claims.

- RIPL and Amazon Web Services quickly partnered to extend Rhode Island's RDL and spin-up an emergency UI claims system in 10 days, which seamlessly scaled to handle the deluge, **making RI the first state in the nation to pay new pandemic UI claims authorized under the CARES Act.**

54,000 claims accepted, processed and paid in days

- Rhode Island **saved thousands of hours of work for over-taxed team members** by automating the validation of 54,000 claims against tax records through a "science-as-service" add-on to the RDL in partnership with the Rhode Island Department of Revenue.
- Possible thanks to **government ownership of an existing RDL cloud solution**, and the partnership and capacity building that RIPL DDPS fosters.
  - Rhode Island was able to spin-up a new, secure PUA claims system in days which could handle claims at any volume and avoid legacy-system failure. (https://www.governing.com/events/webinars/How-Rhode-Island-Rapidly-Scaled-to-Handle-Unemployment-Claims-127726.html)
- **Cloud capacity** allowed Rhode Island to quickly add additional services, such as call center automation and an email and text-messaging platform to serve clients in need with speed and efficiency.
  - Skilled DLT Team members could then focus quality time with cases that needed careful attention

Unlock data and science to create better policy and outcomes

RIPL's DDPS platform and Research Data Lake solution:

- Kick-starts a paradigm shift inside government towards evidence-based policy using administrative data.
- Provides government with the science and technical capacity to tackle their biggest policy challenges, respond quickly in crisis, and reach future goals.
- Supports a community, inside and outside of government, of continuous improvement and positive impact.