AWS re:Invent

GPSTEC303

Data Privacy and Governance in the Age of Big Data: Deploying a De-Identified Data Lake

Ryan Peterson
Principal Solutions Architect,
Data & Analytics
AWS Partner Team

Danielle Greshock
Sr. Manager, Business Applications
AWS Partner Team

AWS re:Invent

aws

Let’s look at some metrics

breachlevelindex.com, 2017 data

What problems are customers trying to solve?

- What type of data am I collecting?
- Where do I collect it?
- Where do I store it?
- Do I have the appropriate legal collection statements?
- How and when do I delete data?
- How do I secure the data?
- What responsibility do I have?
- Why do I collect the data?
- What is my legal basis for processing and using the data?
- Where is a list of all my data?
- Do I communicate with the subject I am collecting from?
- Who do I share it with?
- Who has access to my data? How do I control it?
- What are the use cases for the data? Are they permitted? Who provided permission?
- How do I find my data?

How are privacy regulations attempting to protect consumers?

Global

CSA
Cloud Security Alliance Controls

ISO 9001
Global Quality Standard

ISO 27001
Security Management Controls

ISO 27017
Cloud Specific Controls

ISO 27018
Personal Data Protection

PCI DSS Level 1
Payment Card Standards

SOC 1
Audit Controls Report

SOC 2
Security, Availability, & Confidentiality Report

SOC 3
General Controls Report

United States

CJIS
Criminal Justice Information Services

DoD SRG
DoD Data Processing

FedRAMP
Government Data Standards

FERPA
Educational Privacy Act

ISO FFIEC
Financial Institutions Regulation

FIPS
Government Security Standards

FISMA
Federal Information Security Management

GxP
Quality Guidelines and Regulations

HIPAA
Protected Health Information

CCPA
California Consumer Privacy Act of 2018

ITAR
International Arms Regulations

MPAA
Protected Media Content

NIST
National Institute of Standards and Technology

SEC Rule 17a-4(f)
Financial Data Standards

VPAT/Section 508
Accountability Standards

Asia Pacific

FISC [Japan]
Financial Industry Information Systems

IRAP [Australia]
Australian Security Standards

K-ISMS [Korea]
Korean Information Security

MTCS Tier 3 [Singapore]
Multi-Tier Cloud Security Standard

My Number Act [Japan]
Personal Information Protection

Europe

C5 [Germany]
Operational Security Attestation

Cyber Essentials Plus [UK]
Cyber Threat Protection

G-Cloud [UK]
UK Government Standards

IT-Grundschutz [Germany]
Baseline Protection Methodology

GDPR [EU]
General Data Protection Regulation

AWS re:Invent

aws

Internal challenges with development lifecycle

What challenges do I have, and what can I do?
What are my major risks for data compromise?

Situation two: External challenges with bad actors

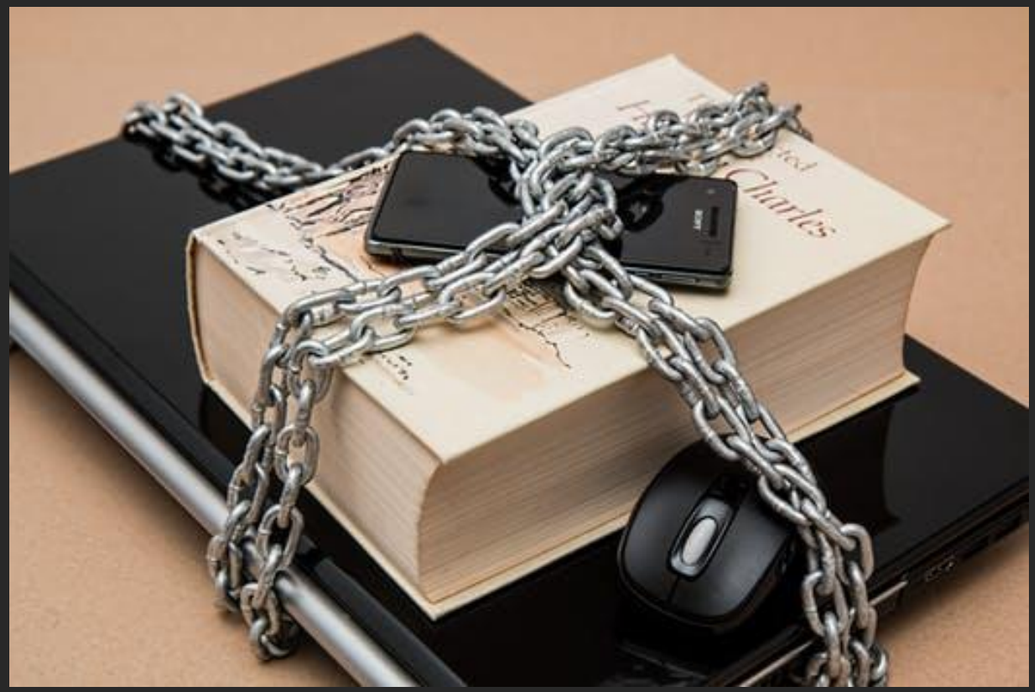
What challenges do I have, and what can I do?
What are my major risks for data compromise?

How we've handled such challenges, pre-GDPR

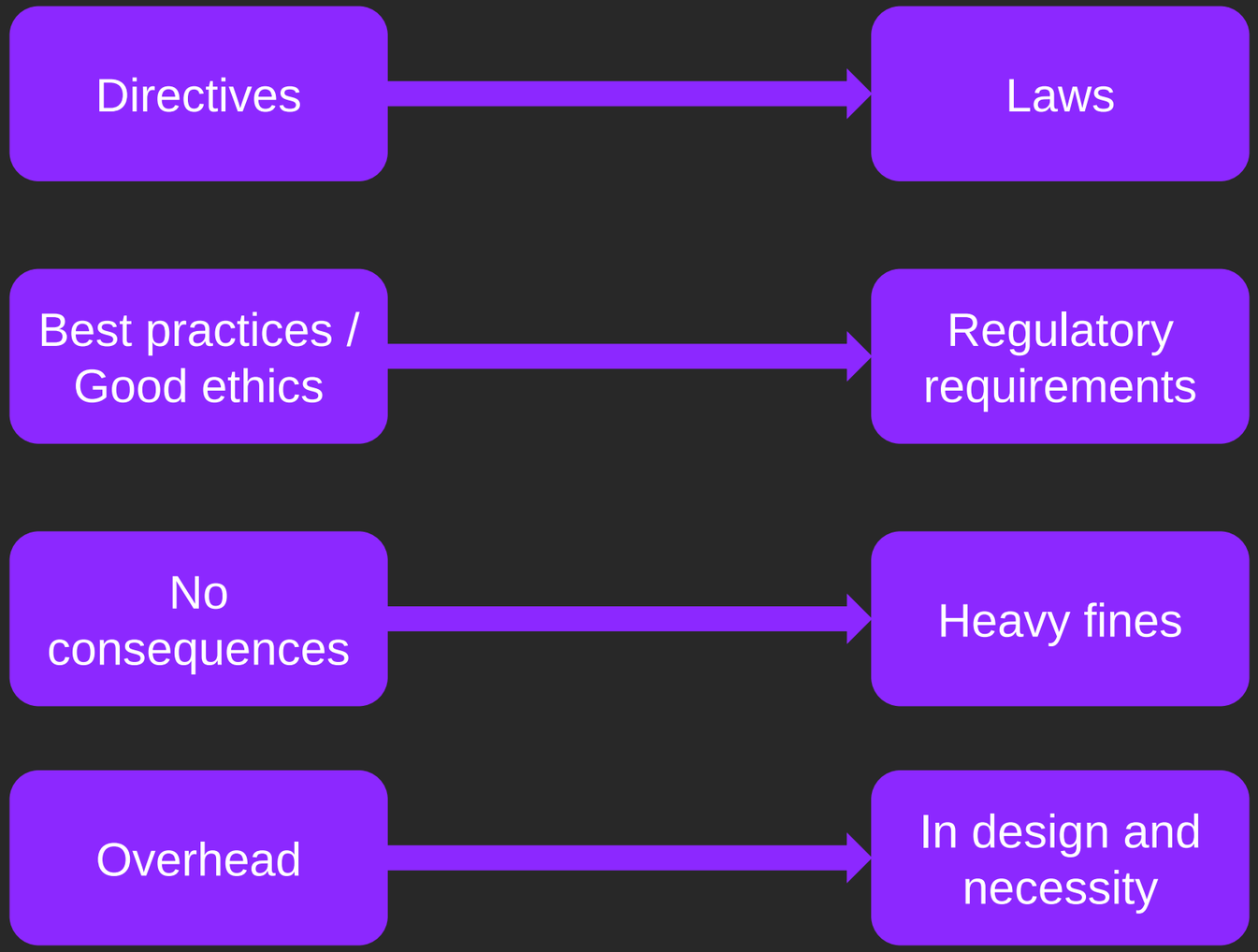
- Outsource credit card processing
- Masking personally identifiable information (PII)
- Firewall rules for distributed denial-of-service (DDoS)
- Network lockdowns
- Database encryption
- Alerting
- Environment automation (no SSH)
- Validation and sanitation of your data input and output (SQL injection / cross-site scripting). Does the data look like it's supposed to?

How does a de-identified data lake help?

- It protects what both internal and external bad actors want—the data!
- It allows developers to focus on their goals—high-quality, tested software

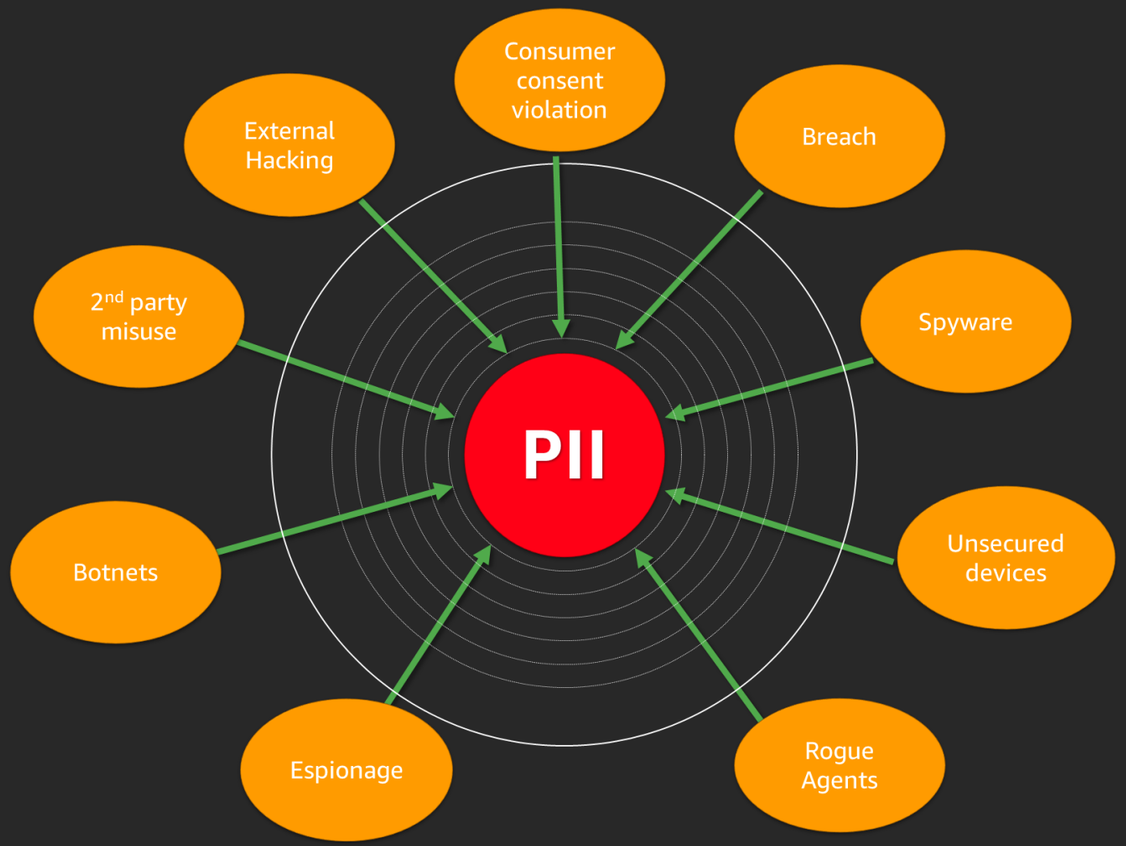


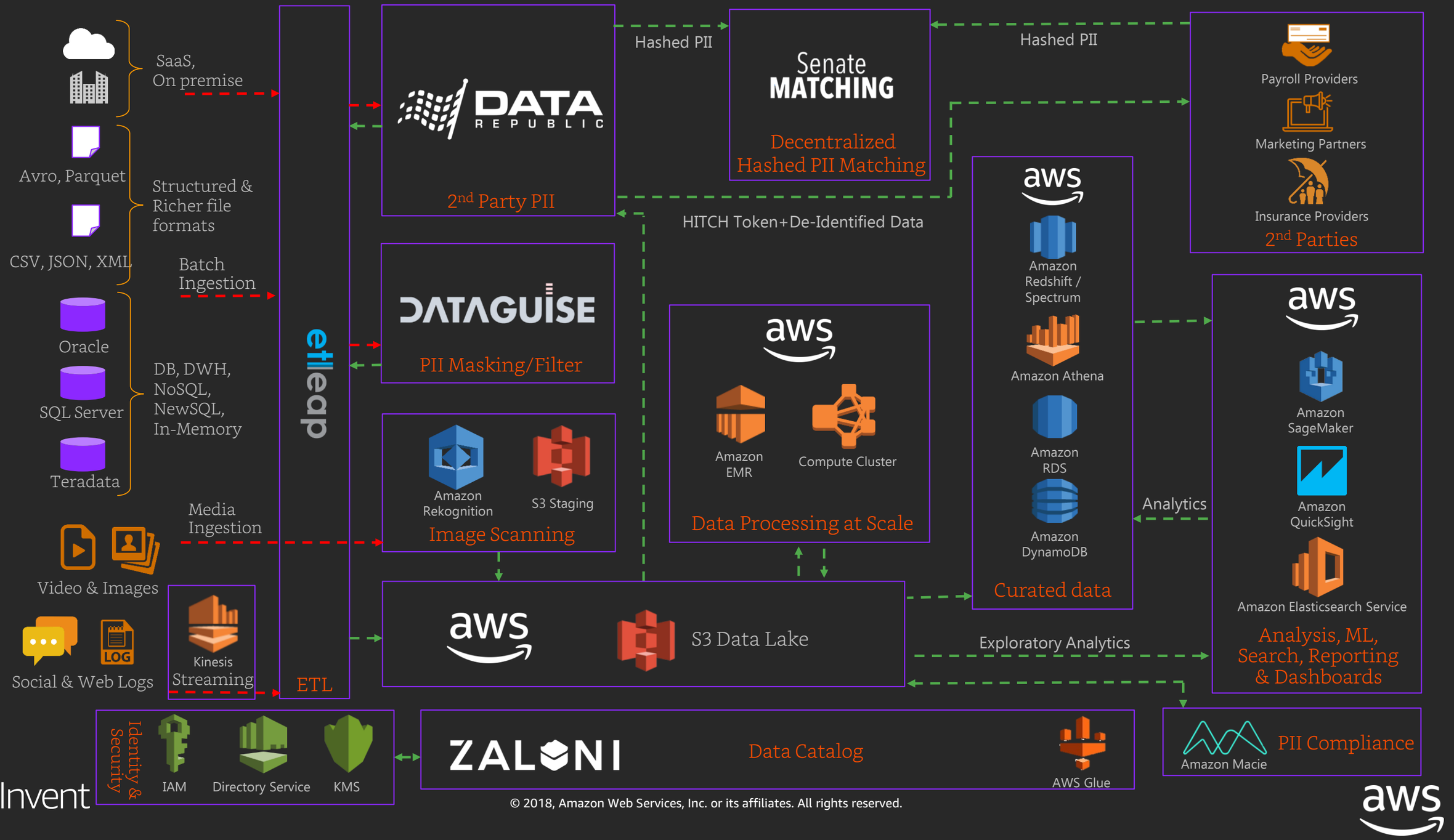
Then and now



How do we resolve PII dangers?

- Do we need to solve these individual issues?
- Is there a solution architecture that solves all PII issues?
- What best practices have you used to mitigate PII dangers?





Why catalogs are key

New business initiatives

What data is available to me?

Advanced analytics

How do I use my data for business value?

Regulatory compliance

How do I understand, document, and ensure proper usage?

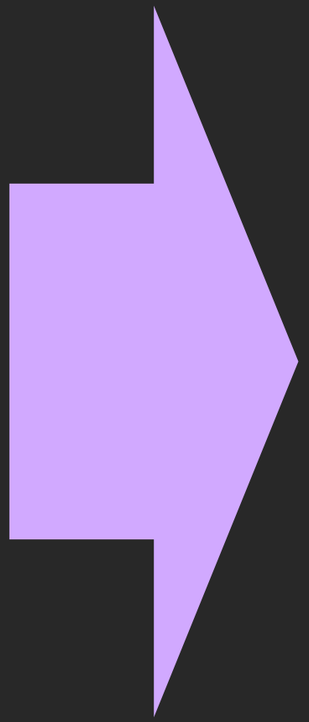
Data Catalog current & future state

Current state

Two types of data catalogs:

1. Pure data cataloging for inventorying and identification
2. Catalogs embedded in apps to make more useful

Has resulted in limited usefulness because no integration into larger enterprise information management activities



Desired state

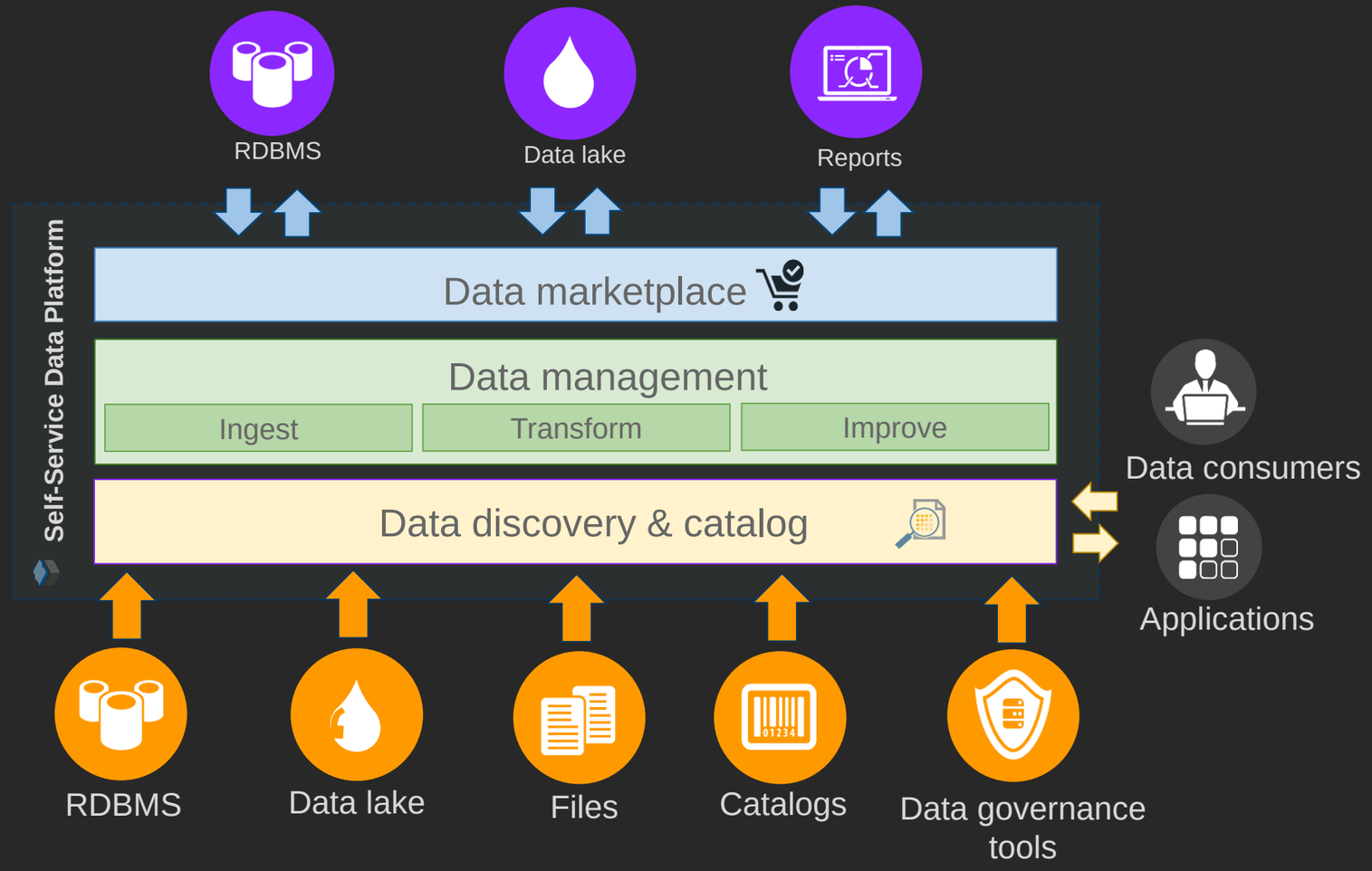
Data marketplace with an actionable data catalog:

- Find data
- Transform data
- Provision data
- Maintain data governance and catalog currency

Actionable Data Catalog via a self-service Data Platform

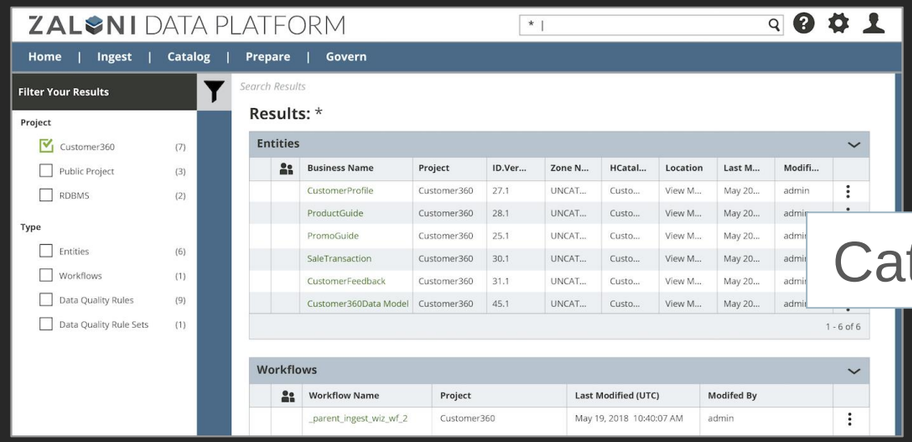
Actionable data catalog

- Enables a self-service Data Platform
- Increases productivity of data producers and consumers
- Governance and catalog currency are infused throughout the process

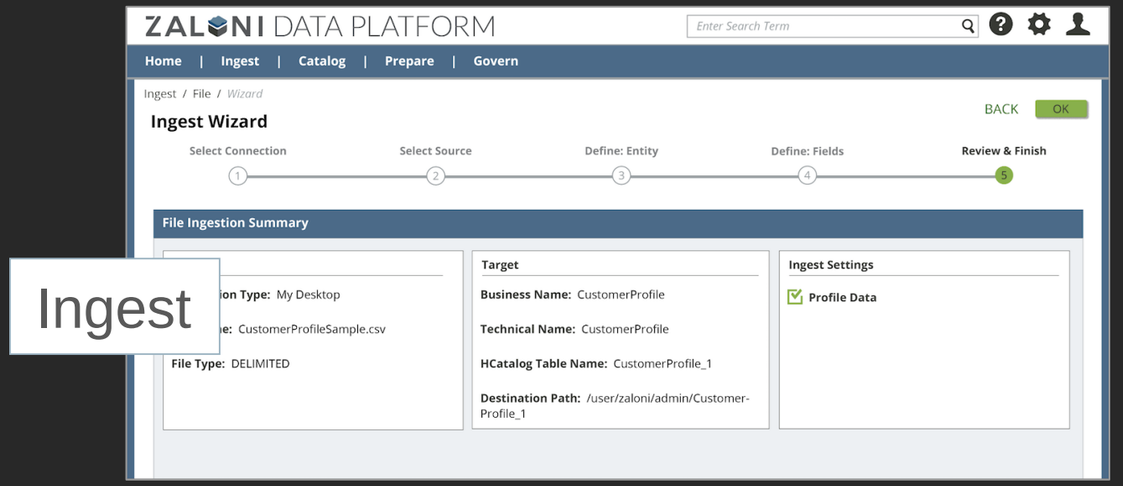


Self-Service Data Platform: Discover, catalog & ingest

- Catalog everything
  - Relational database management system (RDBMS), Data Lakes, Automated Data Inventory
- Leverage enterprise definitions & standards
- Annotate & customize for business need
- Empower SME's to discover new data



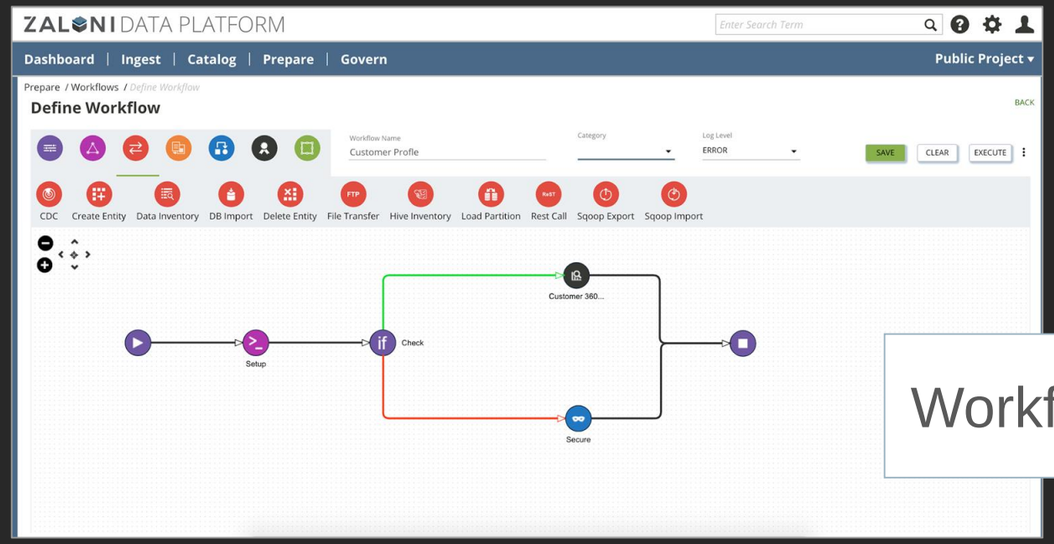
Catalog



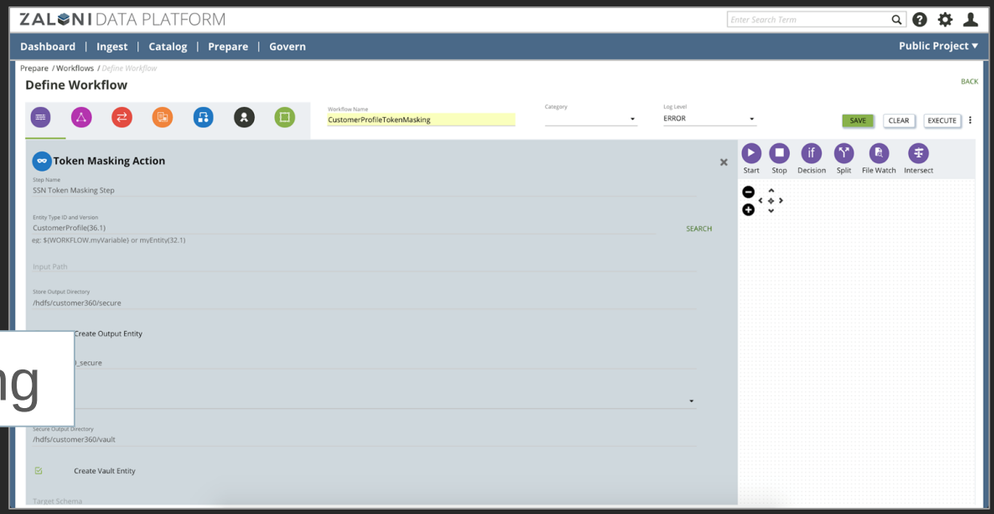
Ingest

Self-Service Data Platform: Prepare & manage

- Enrich from catalog
  - Leverage data profile & quality findings
  - Build recipes to cleanse, join, and aggregate
- Schedule, Manage, Operationalize
  - Scale out
  - Govern based on DM policies



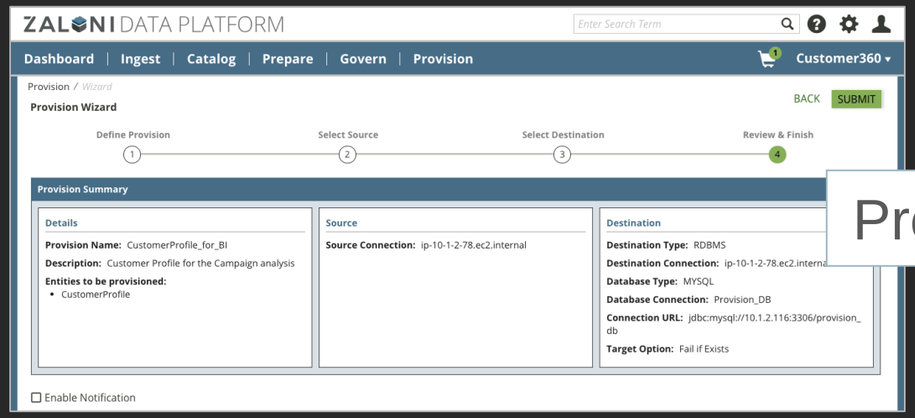
Workflow designer



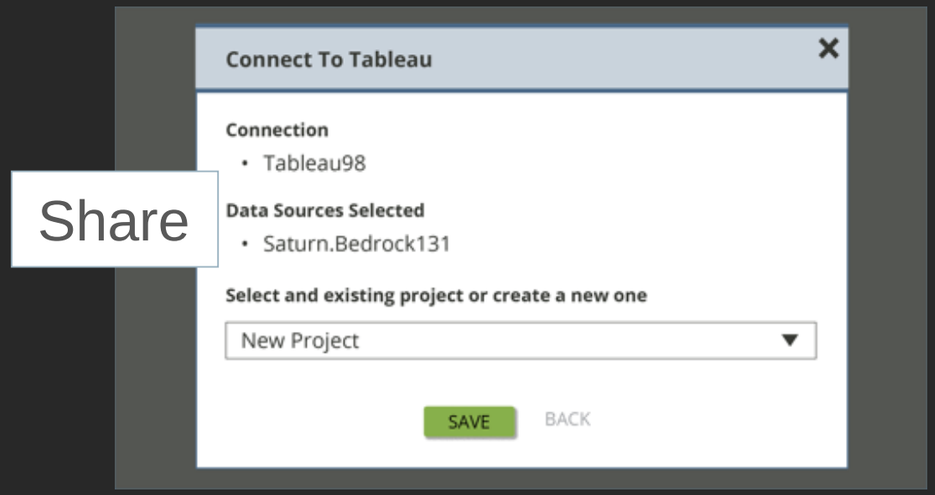
Token masking

Self-Service Data Platform: Marketplace

- Collaborate
  - Workspaces & rankings
  - Annotate and customize
- Share
  - Provision to existing enterprise data warehouses (EDW) or data marts
  - Deliver to data science data lake or zone
  - Enable in data visualization or BI tools (Tableau)



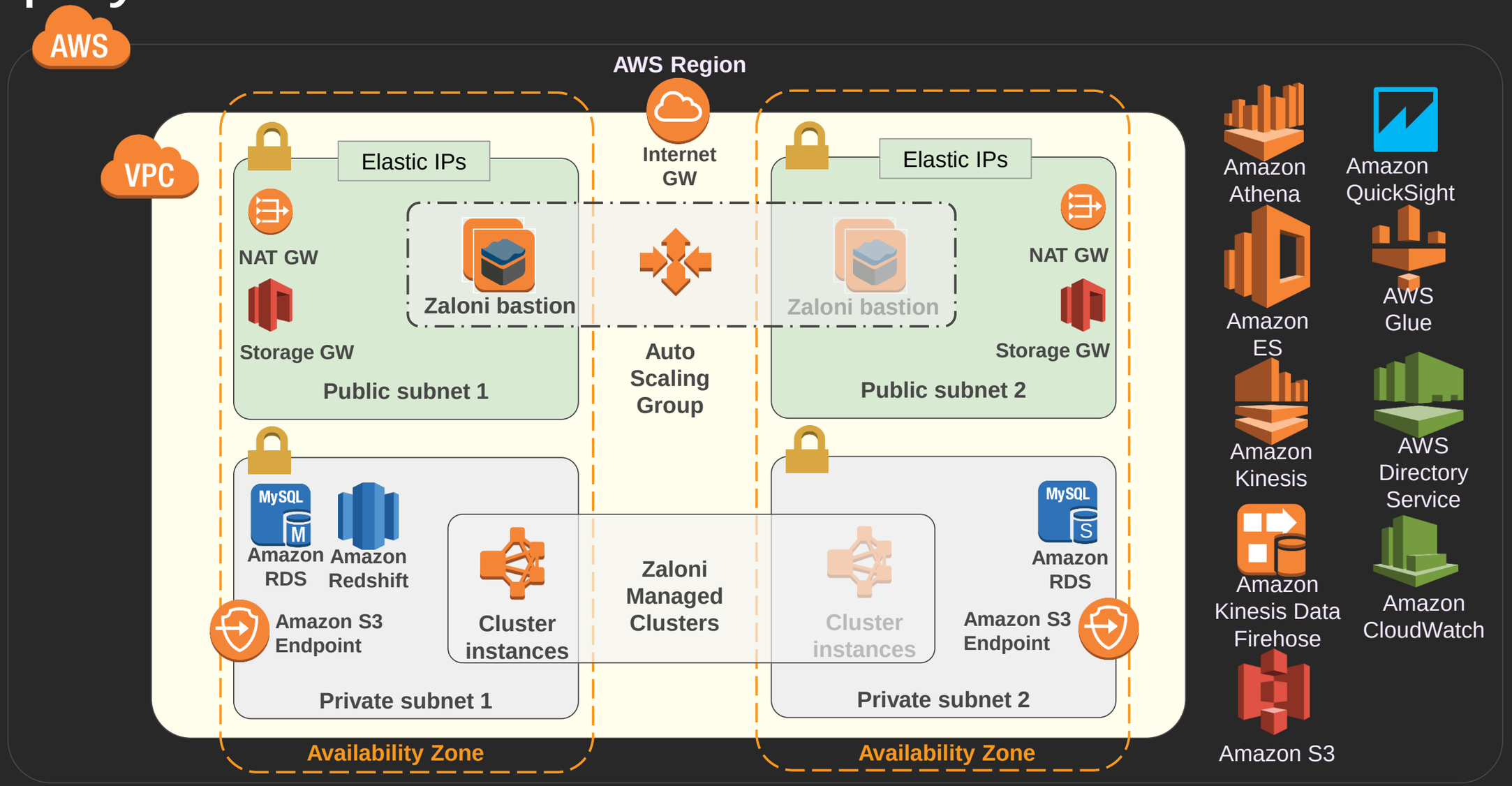
Provision



Share

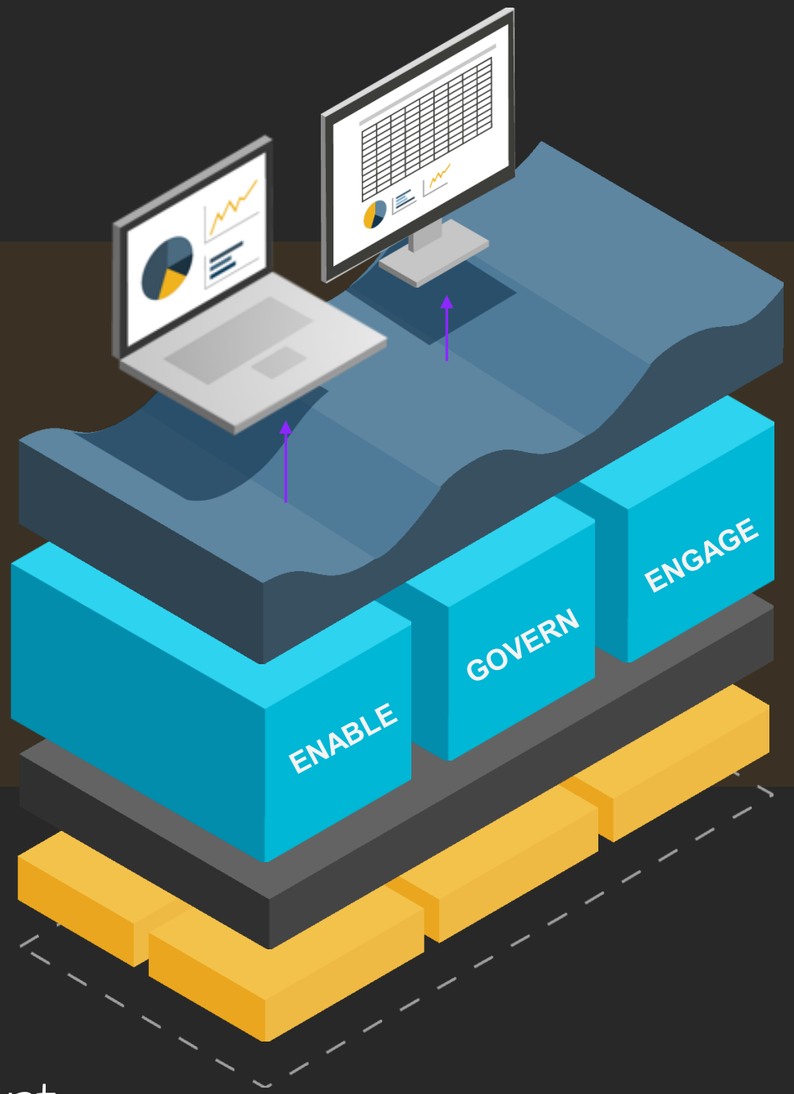
aws

Deployment architecture for AWS data lake with ZDP



aws

Zaloni's Integrated Self-Service Data Platform (ZDP) offerings



Enable

- Batch ingestion
- Streaming ingestion
- Metadata capture
- Auto discovery

Govern

- Data quality
- Data lineage
- Data mastering
- Data privacy/security
- Data enrichment
- Data lifecycle management

Engage

- Discovery catalog
- Self-service ingestion
- Self-service preparation

AWS re:Invent

aws

Dataguise DgSecure capabilities

Detect

Find and report the exact quantity and location of sensitive data in structured, unstructured and semi-structured content

Protect

Remediate your sensitive data exposure by masking and / or encryption it at the element level

Monitor

Track how and where sensitive data is being accessed through a 360° dashboard

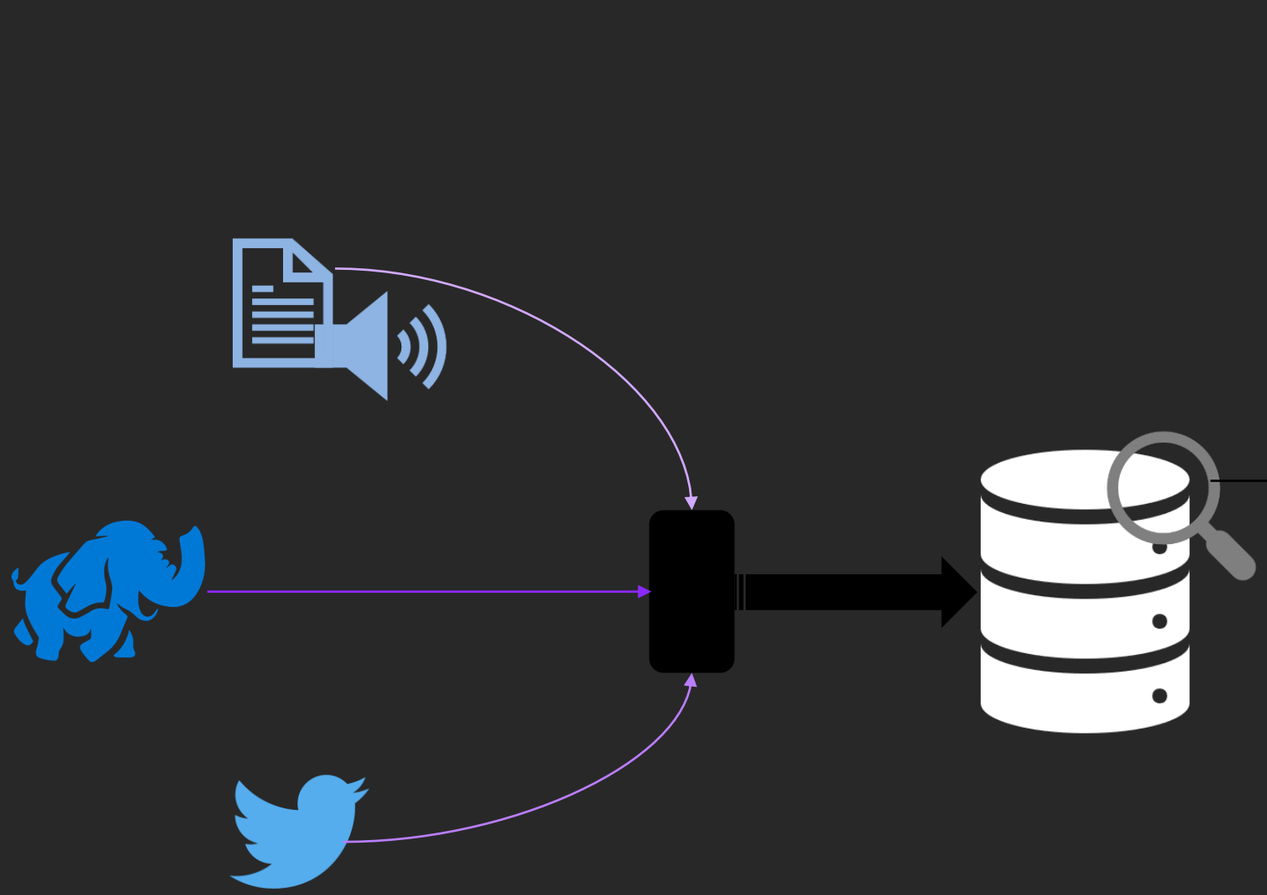
Right of access

Upon request, precisely find and report all records and other information of a specific individual

Right to erasure

Upon request, delete all records and other information of a specific individual

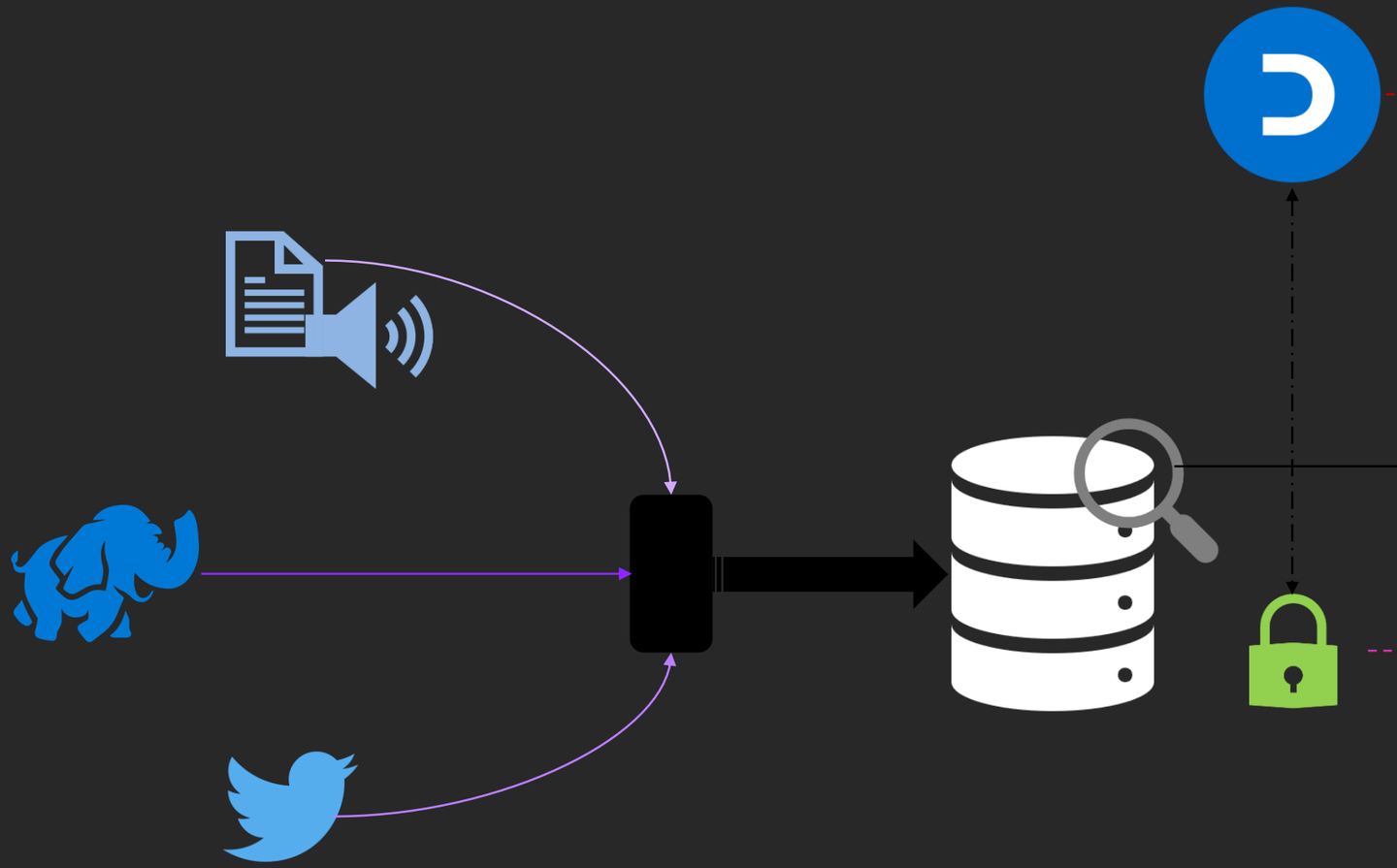
Validating the knowns & finding the unknowns— Structured and semi-structured data



Email *ID* *Name, SSN, State*

Email	Customer ID	Transcript
csalazar@example.com	19664	Just talked to Carlos Salazar
mary@example.com	23423	Mary's SSN is 000000000
mateo@example.com	99644	Mateo is moving to Nevada
NA	02945	It is expected to rain tomorrow

Validating the knowns & finding the unknowns—Structured and semi-structured

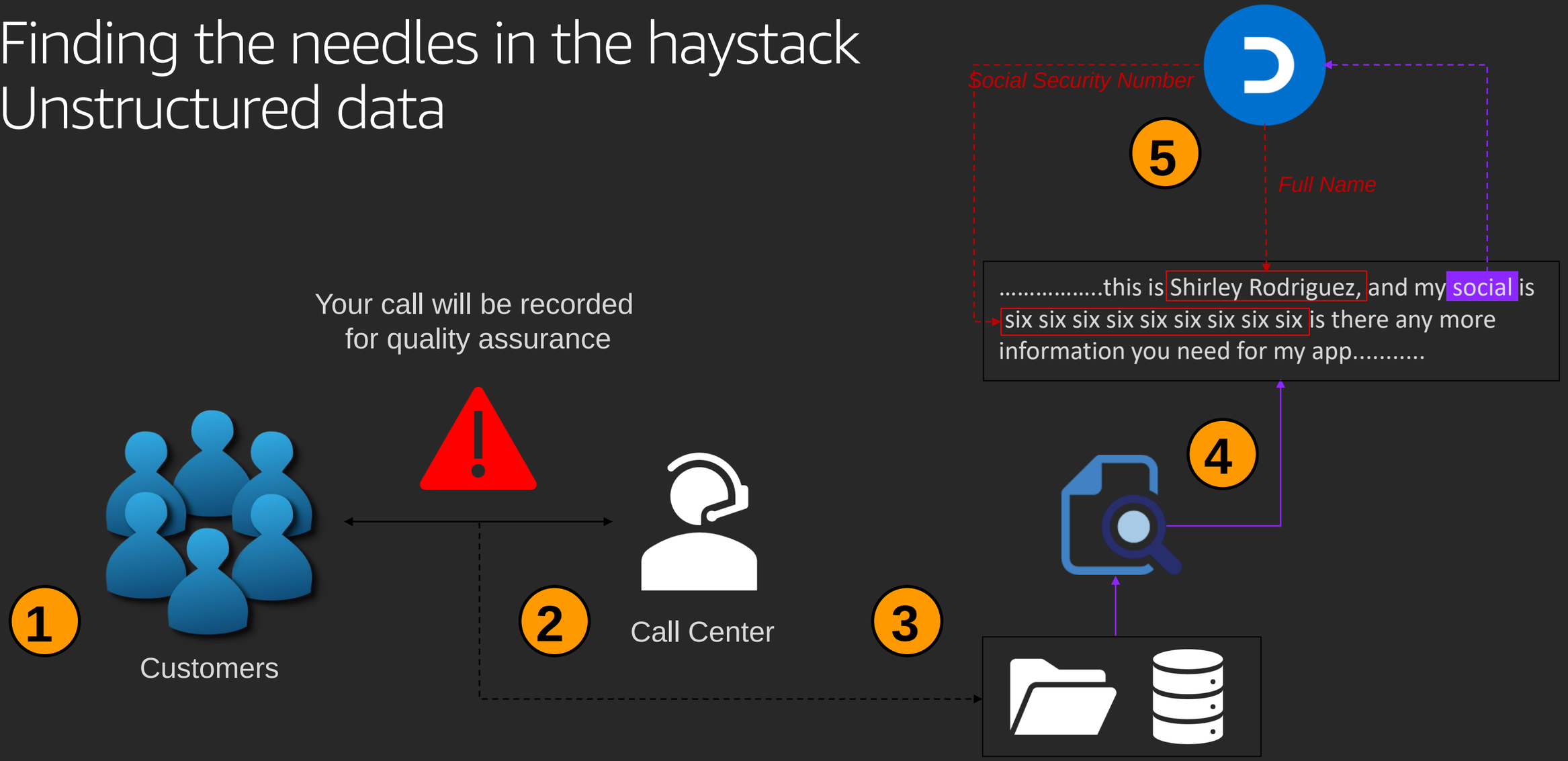


Email *ID* *Name, SSN, State*

Email	Customer ID	Transcript
csalazar@example.com	19664	Just talked to Carlos Salazar
mary@example.com	23423	Mary's SSN is 000000000
mateo@example.com	99644	Mateo is moving to Nevada
NA	02945	It is expected to rain tomorrow

Email	Customer ID	Transcript
4t34gttt	7462391	Just talked to Jane Roe
44e5325	1239474	Jorge's SSN is 666666666
0we&yrw	9983487	Sofia is moving to Texas
NA	3344325	It is expected to rain tomorrow

Finding the needles in the haystack
Unstructured data



AWS re:Invent

aws

Senate matching

Match datasets without PII ever leaving your organization

Private by design

PII never leaves your organization’s firewall
Future-proof compliance

The end of the PII honeypot

Decentralized matching occurs on hash fragments. Protect PII at all times.

Remain in control

Matching subject to strict governance and licensing workflows. Audit all matching.

AWS re:Invent

aws

The decentralized matching process

Senate matching protects customer PII for data custodians:

1. **Tokenization:** Anonymizes PII by replacing it with random tokens when datasets are uploaded to the contributor node

2. **Hashing:** Protects PII by hashing (one-way, non reversible encryption)—original PII is not stored

3. **Slicing:** Distributes hash fragments across a distributed computing network

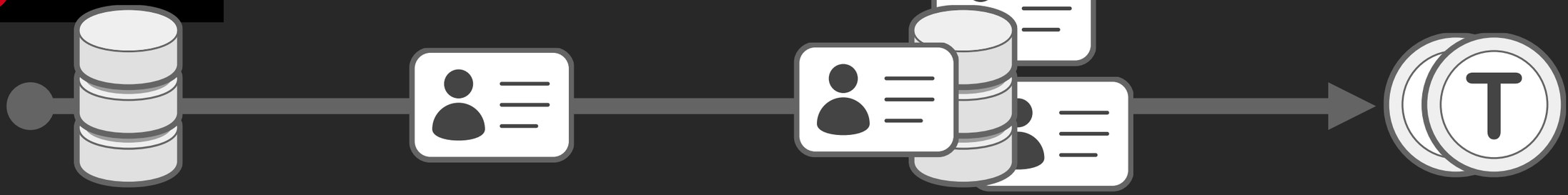
4. **Matching:** Provides a matching service to link identities across different contributor nodes, without disclosing data or putting it at risk of misuse

5. **Governance:** Performs these matches according to data exchange governance within Data Republic's Senate Platform

6. **Auditable:** Functions in a transparent and verifiable way so that analysts and custodians can build trust in the system, the data, and with each other

Comparing methods for PII management

The old way

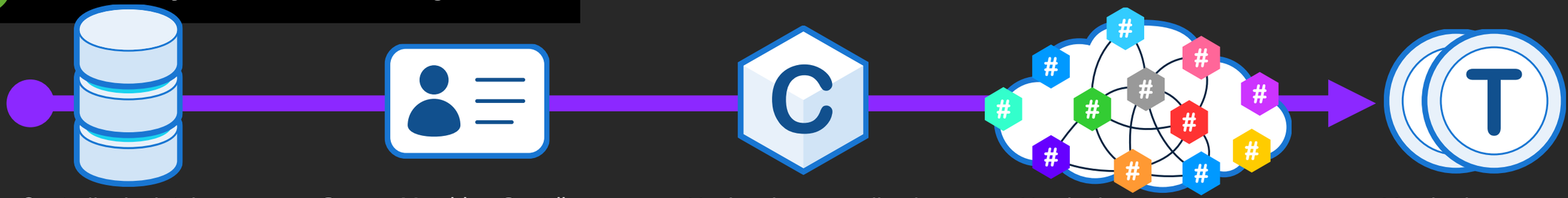


Custodian's database (CRM or EDW) contains customer PII data that must be protected

The central data store is able to match customer records from multiple contributors

Matched customer tokens are used by an analyst to link datasets

The new way with senate matching



Custodian's database (CRM or EDW) contains customer PII data that must be protected

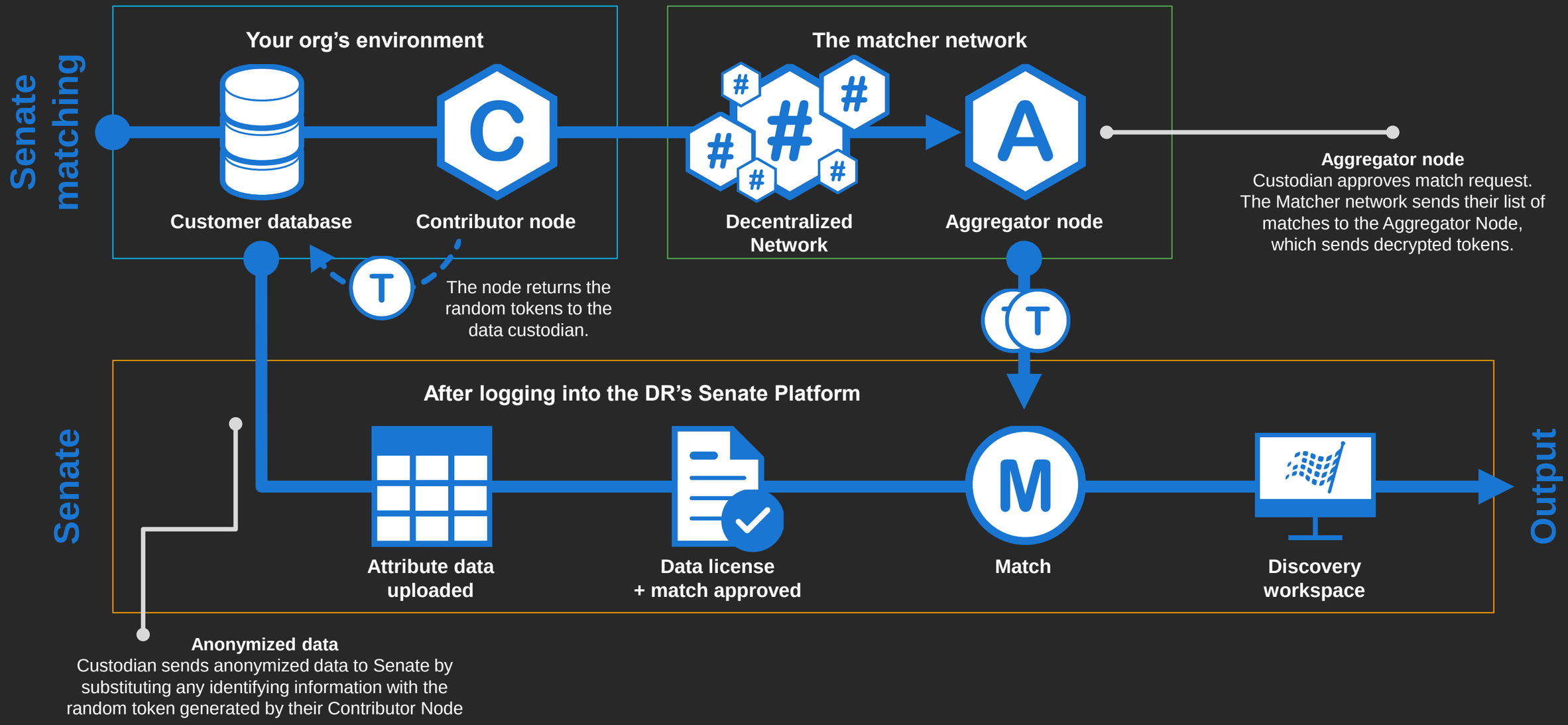
Senate Matching Contributor Node hashes the PII so that the original text is unrecoverable

Hash values are sliced into parts and stored in a distributed network

No single company can match or re-identify customer data

Matched customer tokens are used by an analyst to link datasets

Senate matching process

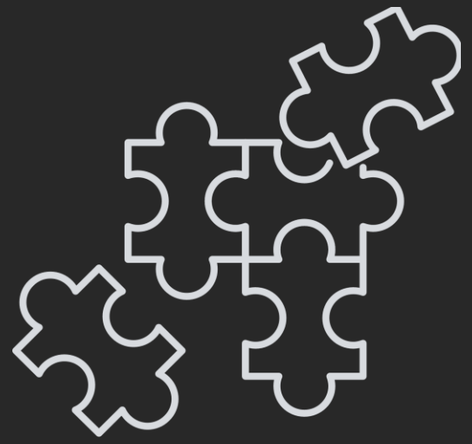


The problem

The old way
Bringing data from sources like databases into data lakes and data warehouses to perform analytics

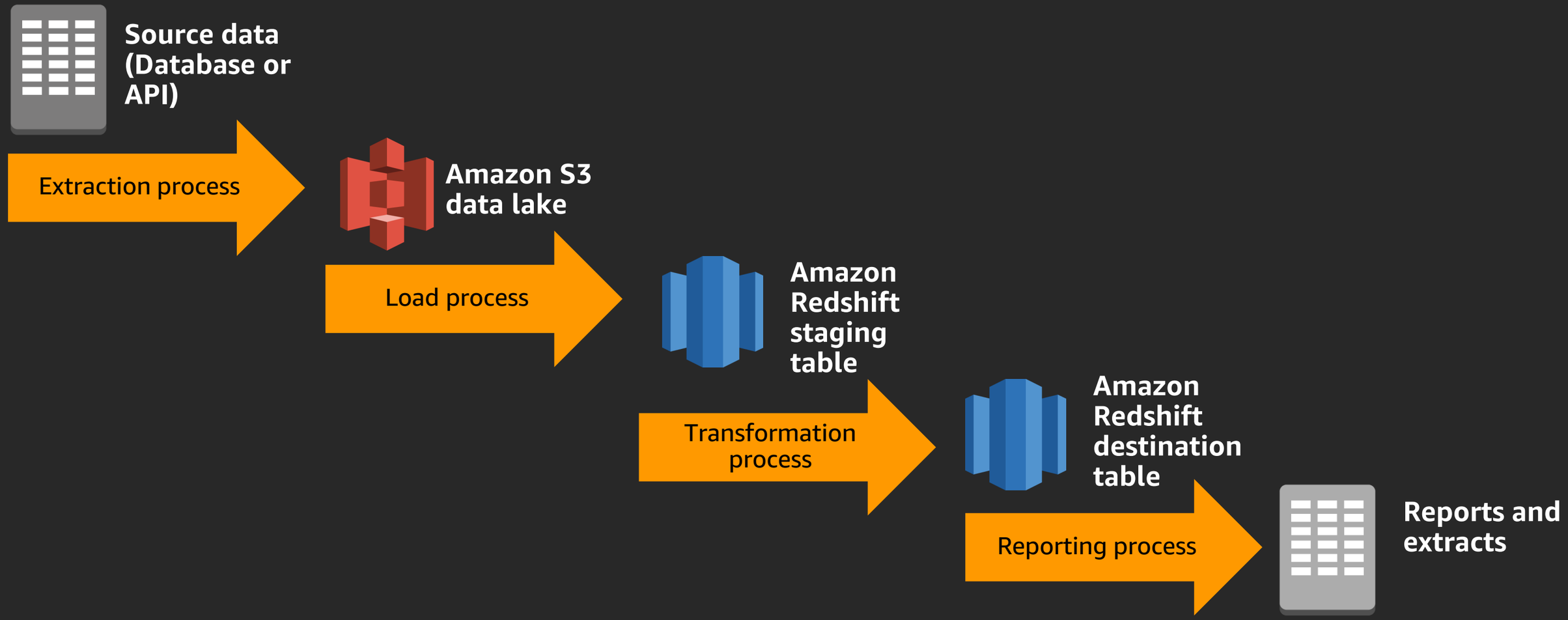
Today
There are many operational systems and sources of data

Cloud-based repositories
Modern data lakes and analytics tools are all based in the cloud

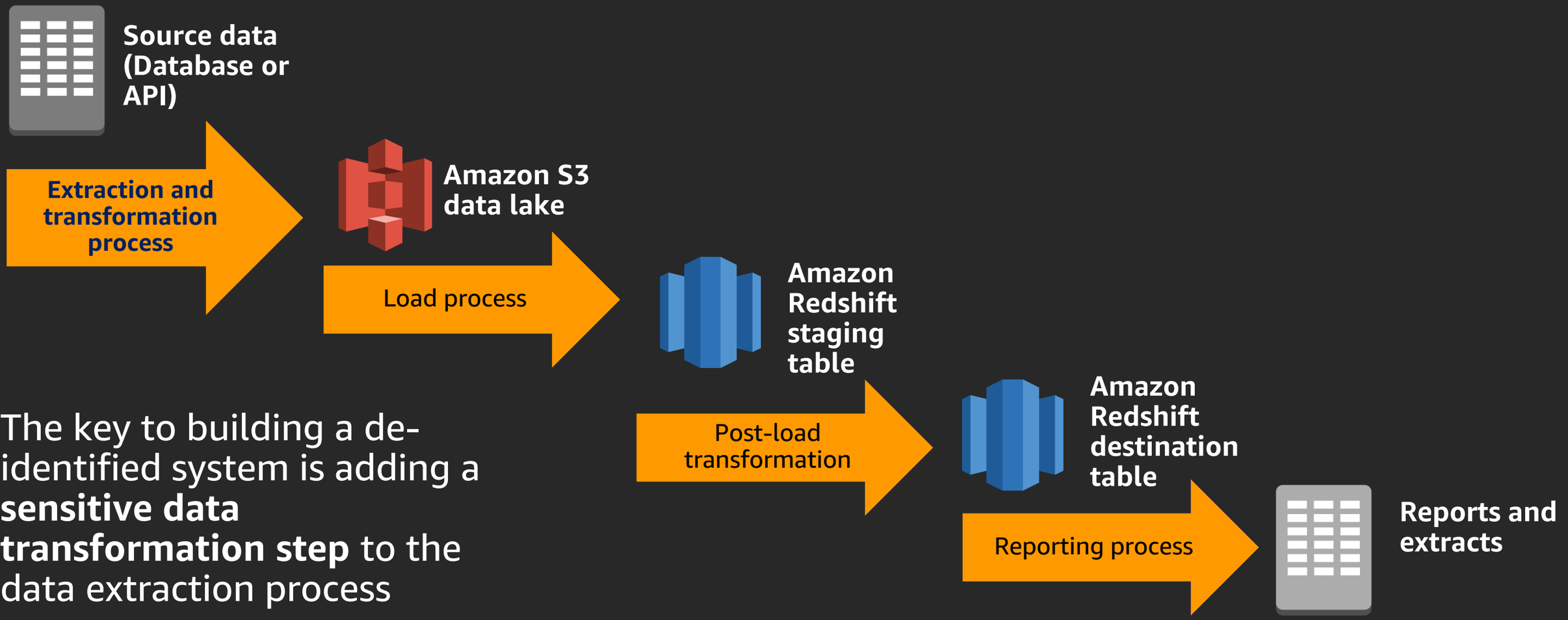


ETL needs to adapt
Risk is increased by compounding the number of locations where data is stored

How data normally flows …

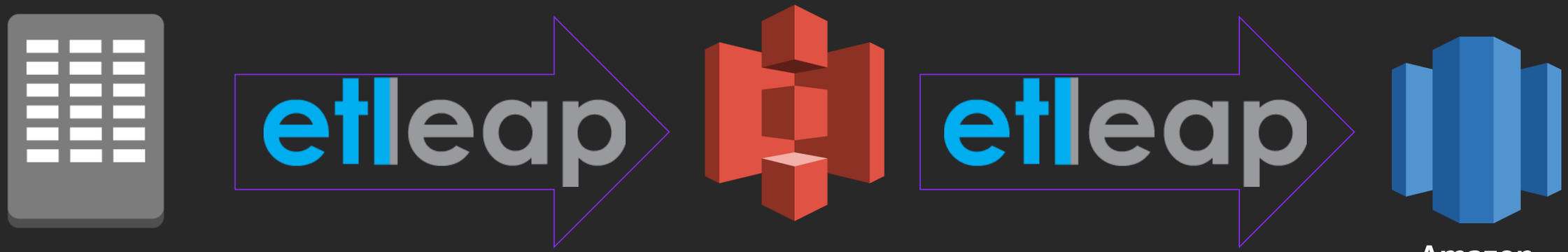


The solution: Transforming sensitive data



The key to building a de-identified system is adding a **sensitive data transformation step** to the data extraction process

How customers use Etleap



Source data (Database or API)

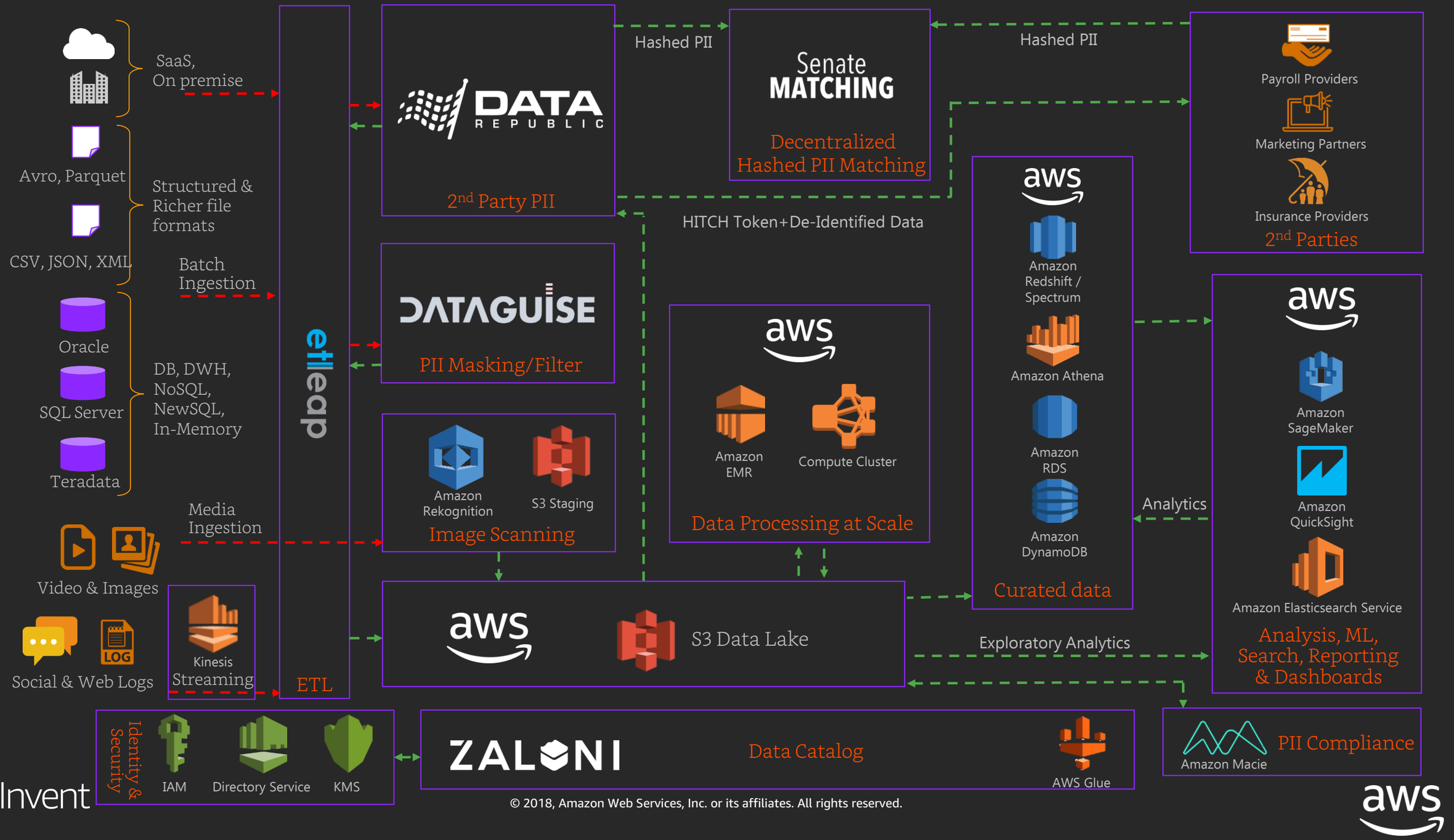
Extraction, sensitive data filtering, milestoning, and transformation

Amazon S3

de-identified data lake

Data curation

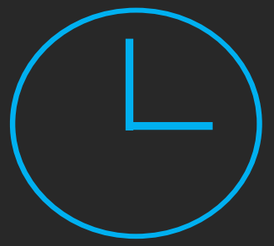
Amazon Redshift data warehouse



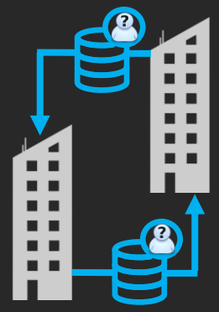
Key takeaways

Detailed catalog of locations where sensitive data resides

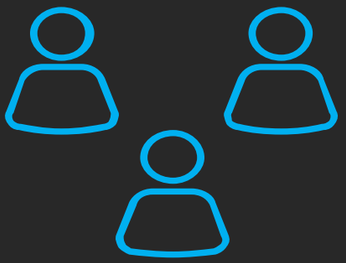
Methodology to audit, assess, and then handle PII



Significantly accelerated time to value—months to minutes



Risk of 2nd party misuse reduced, enables data monetization



Improved consumer, employee, and shareholder trust and revenue

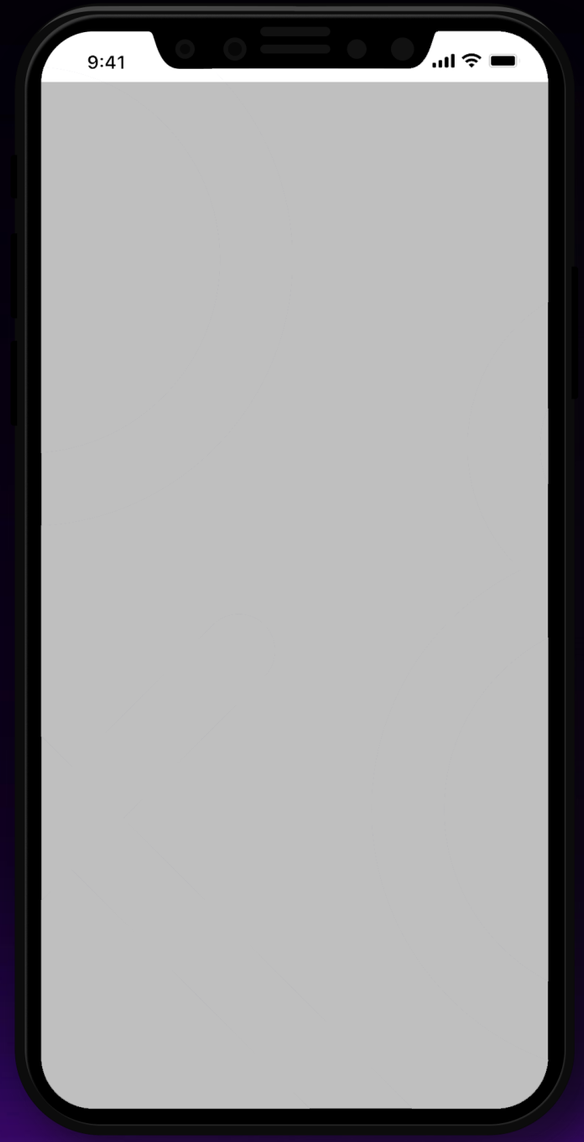
Thank you!

Ryan Peterson
ryapet@amazon.com

AWS re:Invent

aws

Please complete the session survey in the mobile app.



AWS re:Invent

aws