DATA+AI
SUMMIT 2022

Practical Data Governance

Managing data compliance and privacy within a large scale Databricks environment

Brad Nicholas
Director, Digital Platforms, Corning

Aaron Colcord
Sr. Director, Privacera

ORGANIZED BY databricks

About us

Aaron Colcord

Privacera

Aaron is an Adaptive technical leader with 20+ years' experience in spearheading enterprise data solutions and enabling scalable, secure processes which lead to powerful insights from complex data systems.

He has spent the last couple of years working passionately inside evolving technologies such as Lakehouse, Data Mesh, the ever-evolving modern data stack, and acquiring 7 Patents in this area.

He joined Privacera because a belief in the mission and technology to advance Customers and their data management programs.

Brad Nicholas

Corning

Brad runs the software engineering team at Corning Emerging Technology responsible for digital transformation platforms including Databricks & Privacera.

The team works exclusively with open source and open core software and is responsible for the broad-scale adoption of these technologies by digital practitioners across Corning's businesses and functions worldwide.

His experience includes software engineering, product and general management roles at multiple startups and large-scale enterprises. He holds 12 networking and IIoT patents.

What is practical data governance?

Realizing global data-driven scale and value

- Material financial impact through scalable value delivery, with velocity
  - This scopes everything we do
  - Global data governance, software patterns & inner-sourcing are essential
  - Fine-grained runtime access control in the Data Lakehouse is essential as well
  - Security must be designed in, not an overlay
- Open, adaptable, enterprise-grade tech stack
  - Apache Spark, Delta Lake, MLflow
  - Apache Ranger (Data Security)

Allow domain experts to self-serve

Patterns, CI/CD & automated data governance are essential

- Professional SWEs and SREs scale data-driven transformation solutions
  - Not a core competency for many successful enterprises – their differentiating expertise is elsewhere
- The key challenge is to enable domain experts to leverage advanced analytics effectively
  - Experts must focus on the right questions and have a straightforward path to data for answers
  - They act on what they learn & move on to the next question

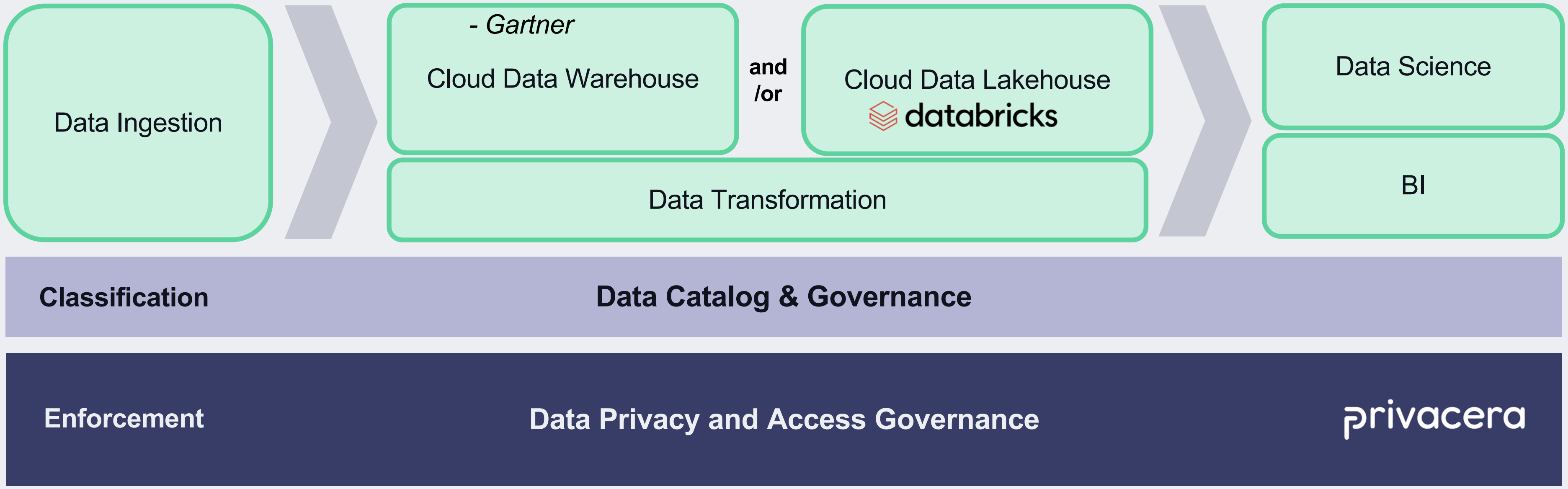
- We're done when practitioners don't need our expertise to be productive

Curated data accessibility empowers the data platform

Secure landing, enrichment, serving, versioning at scale

Data governance is the specification of decision rights and an accountability framework to ensure the appropriate behavior in the valuation, creation, consumption and control of data and analytics.

- Gartner

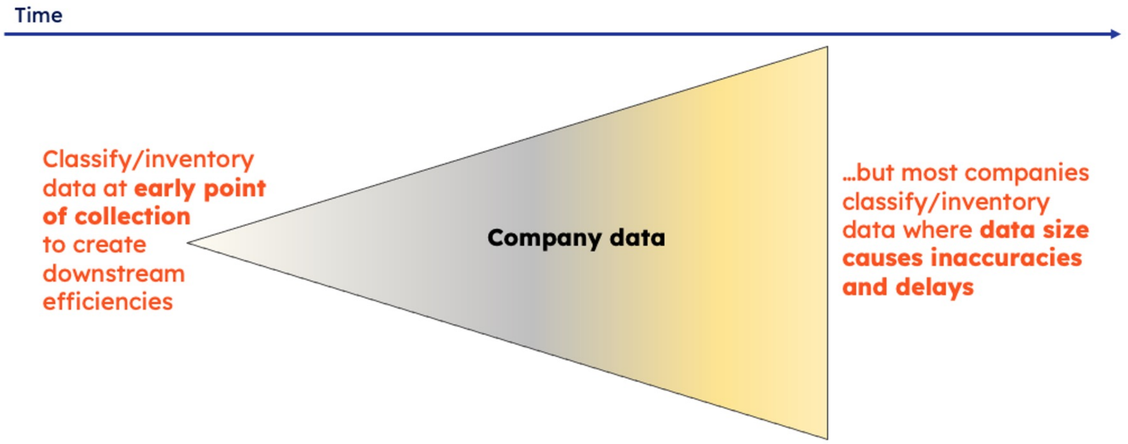


Everyone's reality is existing "brownfield"

Transformation forces a focus on the data lake / lakehouse

- Transformation data sources get processed first
- Change Data Capture decouples the source from the landing platform
- Address data governance & pipeline automation in parallel, opportunistically
- Lather, rinse, repeat

Start early, define patterns, scale through repetition



Classify/inventory data at **early point of collection** to create downstream efficiencies

…but most companies classify/inventory data where **data size causes inaccuracies and delays**

Balance governance & borderless data access

Don’t compromise on either

Drive effective data governance with Policy as Code

- Runtime data access enforcement of operational data contracts
- Agile, automated, best practice policy management
- Full compliance with legal/regulatory obligations

Proactively leverage data access control & discovery capabilities

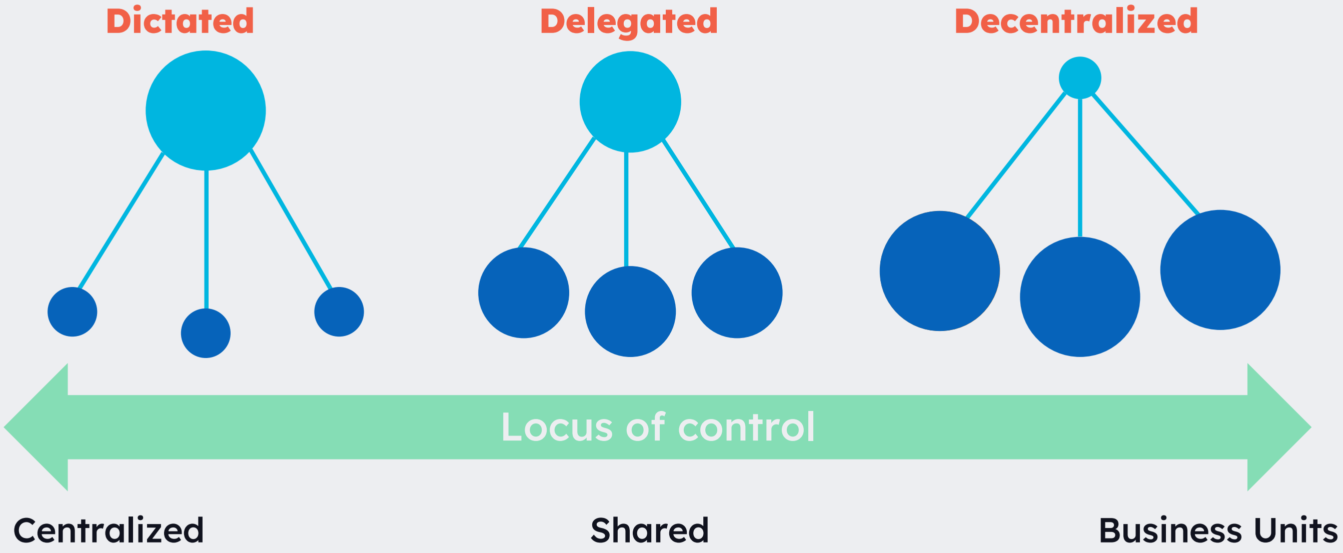
- Active tagging of technical metadata
- PolicyCTL as a gateway to automated version control

Serve curated, ‘Borderless’ data where it enables business value

- Collapse time to data leverage
- Design apps for effective governance (e.g, avoid open fields, PII honey pots)
- Enable Virtual Boundaries that can shift and adapt

Democratization vs Governance

Avoid sacrificing productivity



Common pitfalls of data governance

5 Challenges that lead to Failure

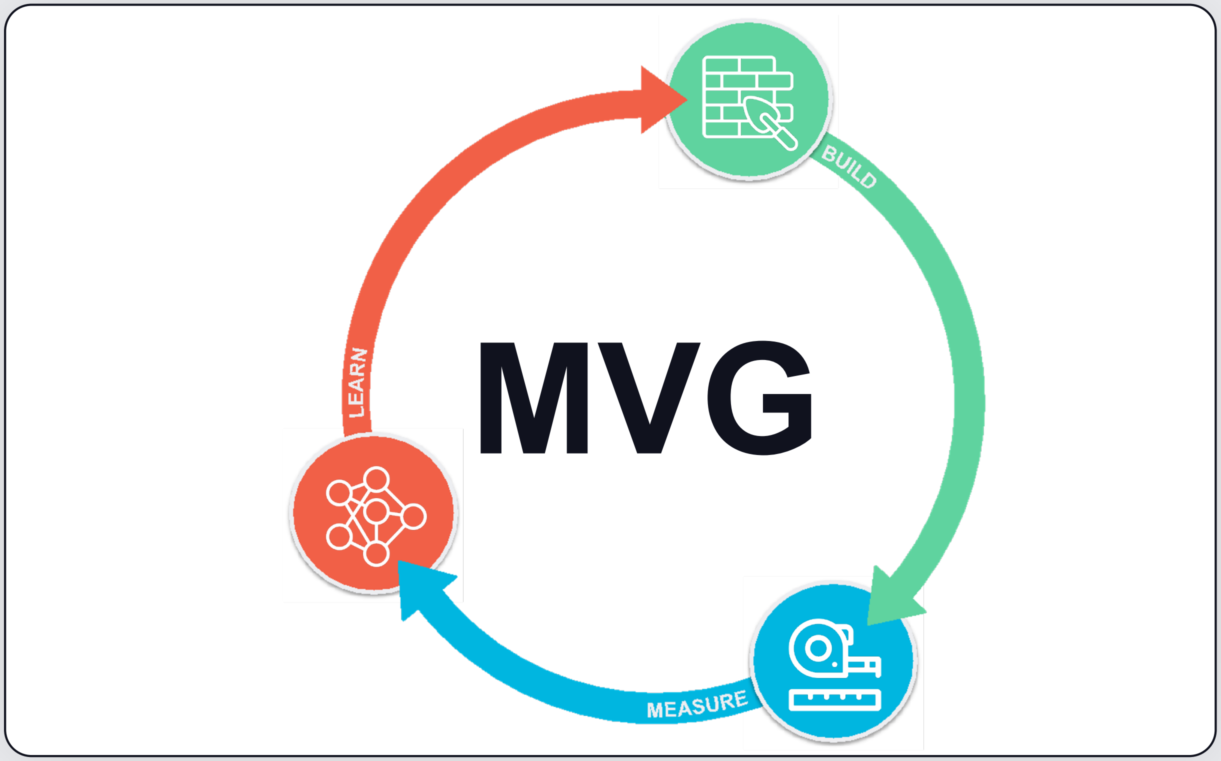
- **Connecting data governance to business outcomes**
- Difficulty defining and adopting consistent data governance processes and policies
- Failure to define common enterprise data definitions
- **Inconsistent approach to data across projects**
- Difficulty defining and sustaining a path to target state (competency to implement)

Minimal Viable Data Governance

The paradox is real

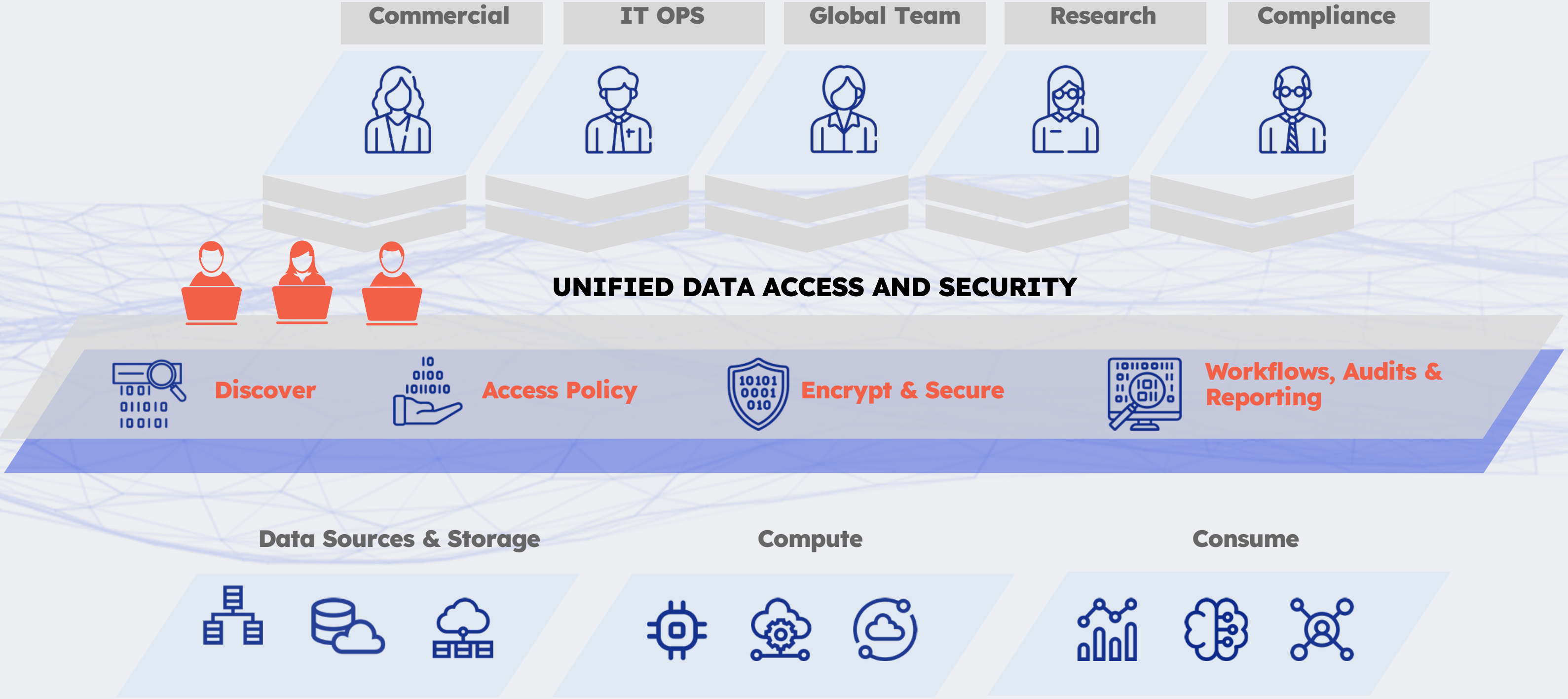
Business value is…

- Data contract is the spirit of MVG
- Most ‘data products’ have no defined value.
- Without quantified value, what is the justification?
- ‘Rogue’ business groups
- Usually the focus is on the symptoms, not the cure.



Unified Data Access Governance

A single location for Data Access and Security Governance



DATA+AI SUMMIT 2022

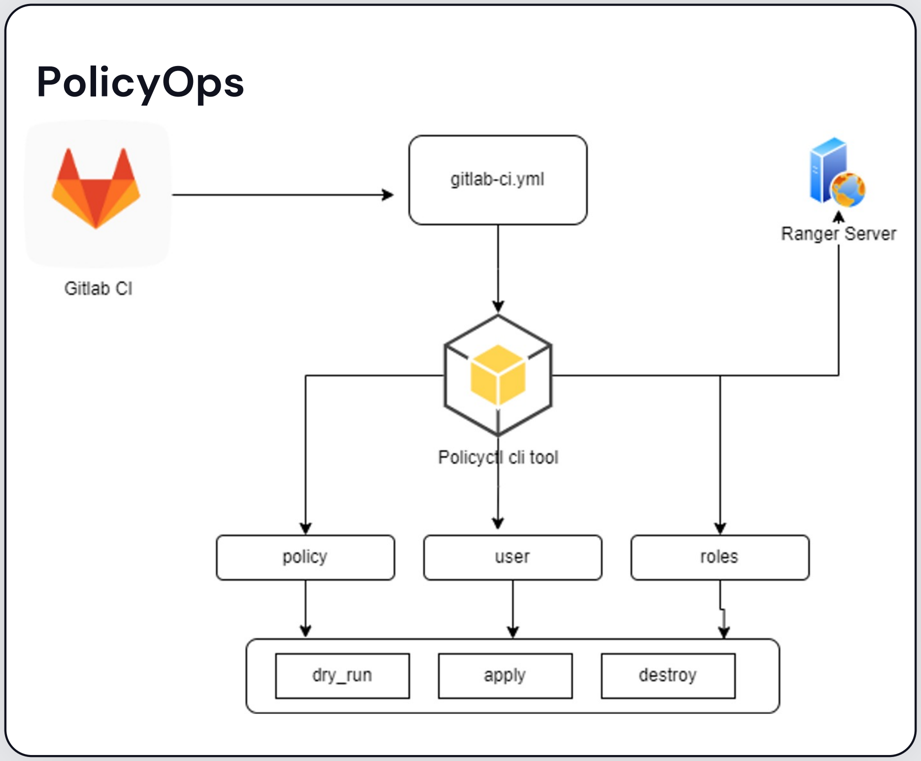
What is PolicyOps?

Automating for Scale

Open Content

Policy Ops is CI/CD tool for Apache Ranger policy management. It provide flexibility to store policy, user, roles etc., inside version control so any changes will be controlled using git and it will be easy to maintain and deploy. Policy ops is based on a cli tool called policyctl which is developed by **Privacera** for its customer to make changes in the Apache Ranger policies using cli commands.

- YAML based properties for easy understanding
- More Controlled and Error Free Changes.
- Easy command-line step execution.



Scaling techniques

Automating access control through PolicyOps

Policy as Code

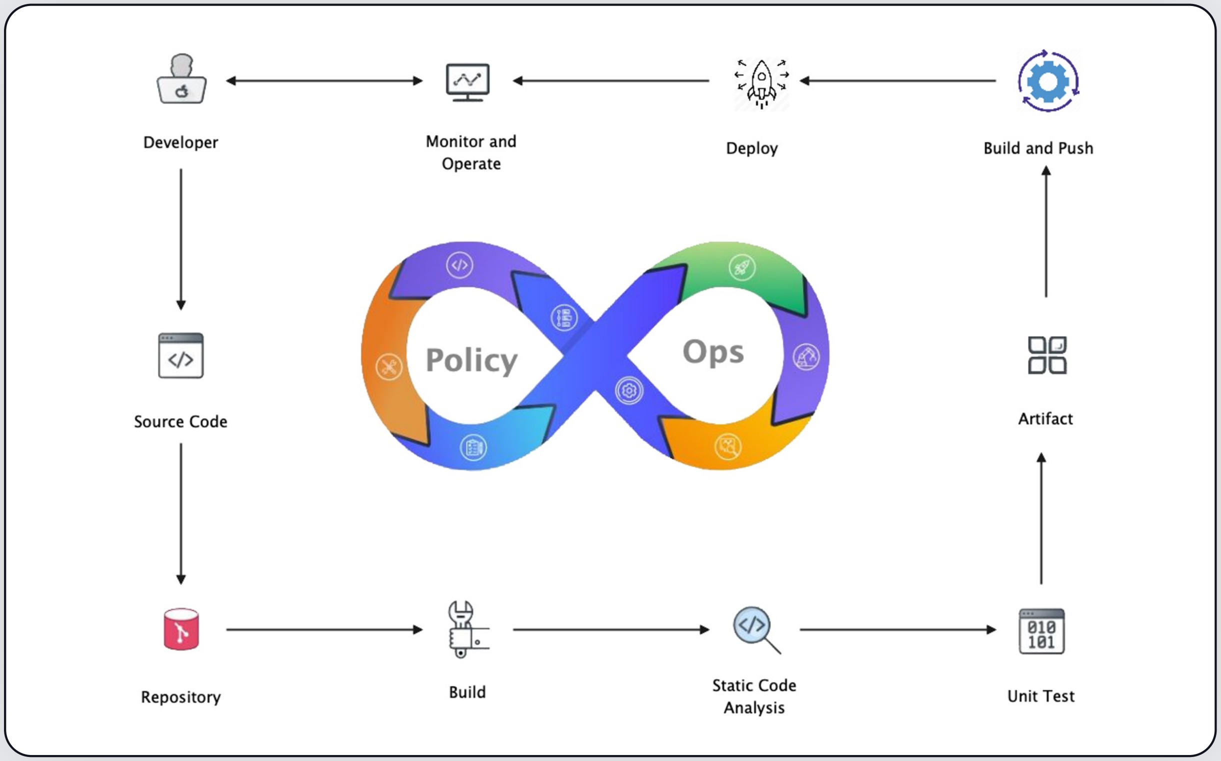
- Roles
- Data contracts
- Test datasets

Resources & Access Methods

- Notebooks
- Jobs
- Files
- Buckets

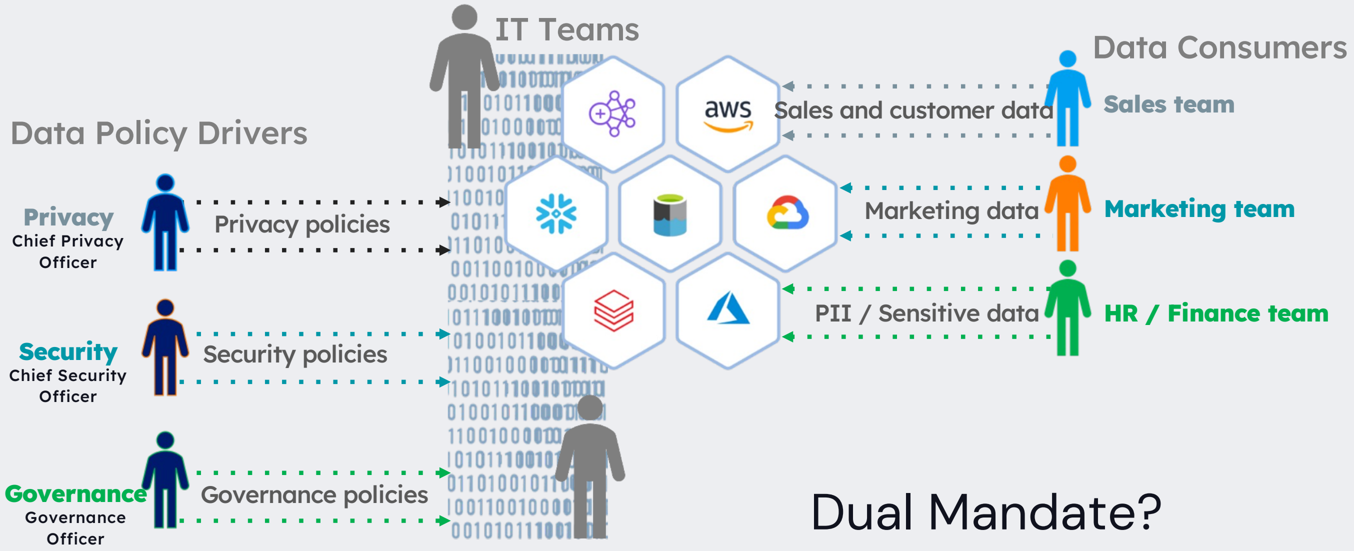
Terraform

- Plan/Apply is your friend



Our demo scenario

One of our enterprise stakeholders now has a concern



Dual Mandate?

Former AWS engineer convicted over hack that cost Capital One $270m

Solution Ingredients

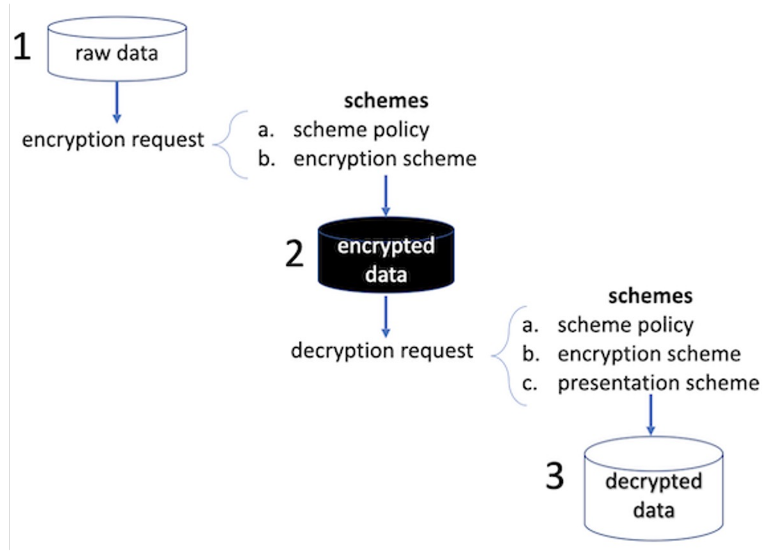
Encryption Scheme: Defines how to encrypt the Data. What users can decrypt the data

Presentation Scheme: Can Keep data sensitive hidden based on User

Keys can be stored externally

Encryption can occur automatically in a workflow

Schemes



1 raw data

encryption request

schemes

a. scheme policy
b. encryption scheme

2 encrypted data

decryption request

schemes

a. scheme policy
b. encryption scheme
c. presentation scheme

3 decrypted data

Demo

PolicyOps

Demo

PolicyOps

Clone existing ranger server policy , user , role etc. this will download existing policies from the server and put it inside a git folder.

```
policyctl package create --name path/to/my/package
```

Test what are the changes made inside yaml file . We run command in –dry_run mode which don’t apply the changes but give a glimpse of the changes.

```
policyctl package apply --name path/to/my/package -dry_run=TRUE
```

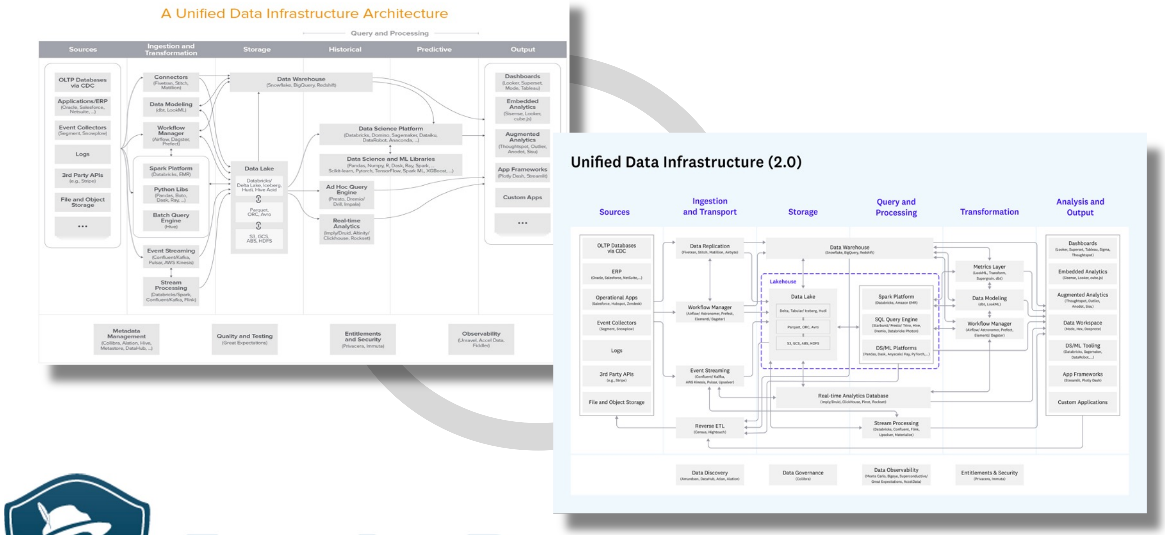
Apply changes for a single policy.

```
policyctl policy apply --name path/to/my/package/policy.yaml
```

Open-source

- Ability to adapt to your business
- Scalable
- Interoperable
- No Vendor Lock-In

Rapidly adapt and Evolve


A Unified Data Infrastructure Architecture
Query and Processing
Sources
Ingestion and Transformation
Storage
Historical
Predictive
Output
OLTP Databases via CDC
Applications/ERP (Oracle, Salesforce, Netsuite, ...)
Event Collectors (Segment, Snowplow)
Logs
3rd Party APIs (e.g., Stripe)
File and Object Storage
...
Connectors (Fivetran, Stitch, Matillion)
Data Modeling (dbt, LookML)
Workflow Manager (Airflow, Dagster, Prefect)
Spark Platform (Databricks, EMR)
Python Libs (Pandas, Boto, Dask, Ray, ...)
Batch Query Engine (Hive)
Event Streaming (Confluent/Kafka, Pulsar, AWS Kinesis)
Stream Processing (Databricks/Spark, Confluent/Kafka, Flink)
Data Warehouse (Snowflake, BigQuery, Redshift)
Data Lake
Databricks/ Delta Lake, Iceberg, Hudi, Hive Acid
Parquet, ORC, Avro
S3, GCS, ABS, HDFS
Data Science Platform (Databricks, Domino, Sagemaker, Dataiku, DataRobot, Anaconda, ...)
Data Science and ML Libraries (Pandas, Numpy, R, Dask, Ray, Spark, ... Scikit-learn, Pytorch, TensorFlow, Spark ML, XGBoost, ...)
Ad Hoc Query Engine (Presto, Dremio/ Drill, Impala)
Real-time Analytics (Imply/Druid, Altinity/ Clickhouse, Rockset)
Dashboards (Looker, Superset, Mode, Tableau)
Embedded Analytics (Sisense, Looker, cube.js)
Augmented Analytics (Thoughtspot, Outlier, Anodot, Sisu)
App Frameworks (Plotly Dash, Streamlit)
Custom Apps
...
Metadata Management (Collibra, Alation, Hive, Metastore, DataHub, ...)
Quality and Testing (Great Expectations)
Entitlements and Security (Privacera, Immuta)
Observability (Unravel, Accel Data, Fiddler)
Unified Data Infrastructure (2.0)
Sources
Ingestion and Transport
Storage
Query and Processing
Transformation
Analysis and Output
OLTP Databases via CDC
ERP (Oracle, Salesforce, NetSuite,...)
Operational Apps (Salesforce, Hubspot, Zendesk)
Event Collectors (Segment, Snowplow)
Logs
3rd Party APIs (e.g., Stripe)
File and Object Storage
Data Replication (Fivetran, Stitch, Matillion, Airbyte)
Workflow Manager (Airflow/ Astronomer, Prefect, Elementl/ Dagster)
Event Streaming (Confluent/ Kafka, AWS Kinesis, Pulsar, Upsolver)
Reverse ETL (Census, Hightouch)
Data Warehouse (Snowflake, BigQuery, Redshift)
Lakehouse
Data Lake
Delta, Tabular/ Iceberg, Hudi
Parquet, ORC, Avro
S3, GCS, ABS, HDFS
Spark Platform (Databricks, Amazon EMR)
SQL Query Engine (Starburst/ Presto/ Trino, Hive, Dremio, Databricks Photon)
DS/ML Platforms (Pandas, Dask, Anyscale/ Ray, PyTorch,...)
Real-time Analytics Database (Imply/Druid, ClickHouse, Pinot, Rockset)
Stream Processing (Databricks, Confluent, Flink, Upsolver, Materialize)
Metrics Layer (LookML, Transform, Supergrain, dbt)
Data Modeling (dbt, LookML)
Workflow Manager (Airflow/ Astronomer, Prefect, Elementl/ Dagster)
Dashboards (Looker, Superset, Tableau, Sigma, Thoughtspot)
Embedded Analytics (Sisense, Looker, cube.js)
Augmented Analytics (Thoughtspot, Outlier, Anodot, Sisu)
Data Workspace (Mode, Hex, Deepnote)
DS/ML Tooling (Databricks, Sagemaker, DataRobot,...)
App Frameworks (Streamlit, Plotly Dash)
Custom Applications
Data Discovery (Amundsen, DataHub, Atlan, Alation)
Data Governance (Collibra)
Data Observability (Monte Carlo, Bigeye, Superconductive/ Great Expectations, AccelData)
Entitlements & Security (Privacera, Immuta)

Apache Ranger

Andreessen horowitz (https://future.com/emerging-architectures-modern-data-infrastructure/)

DATA+AI
SUMMIT 2022

Thank you