aws INNOVATE

DATA EDITION

23 August, 2022

Data governance on AWS

Francis McGregor-Macdonald

Analytics Specialist Solutions Architects Leader
Amazon Web Services

Agenda

1. Data-driven and data governance

2. AWS models for data governance

3. Governing the data lifecycle

4. AWS Services

Customers want more value from their data

Growing exponentially

From new sources

Increasingly diverse

Used by many people

Analyzed by many applications

The business value of data

By making 10% more data accessible, a typical Fortune 1000 company will see a $65 million increase in net income[1]

415%
five-year ROI[1]

48%
reduced total cost of operations[2]

1. Data driven organization
2. IDC whitepaper: The business value of AWS data lakes, analytics and ML services

Many businesses want to be data driven but few have been successful

Growing exponentially

Increasingly diverse

Data lake

Spread across multiple data stores on-prem and in the cloud

Stored in multiple regions and accounts

Used by more users for more use cases

There is **more data** than people think

Data growth

90% of data worldwide was generated in last 5 years[1]

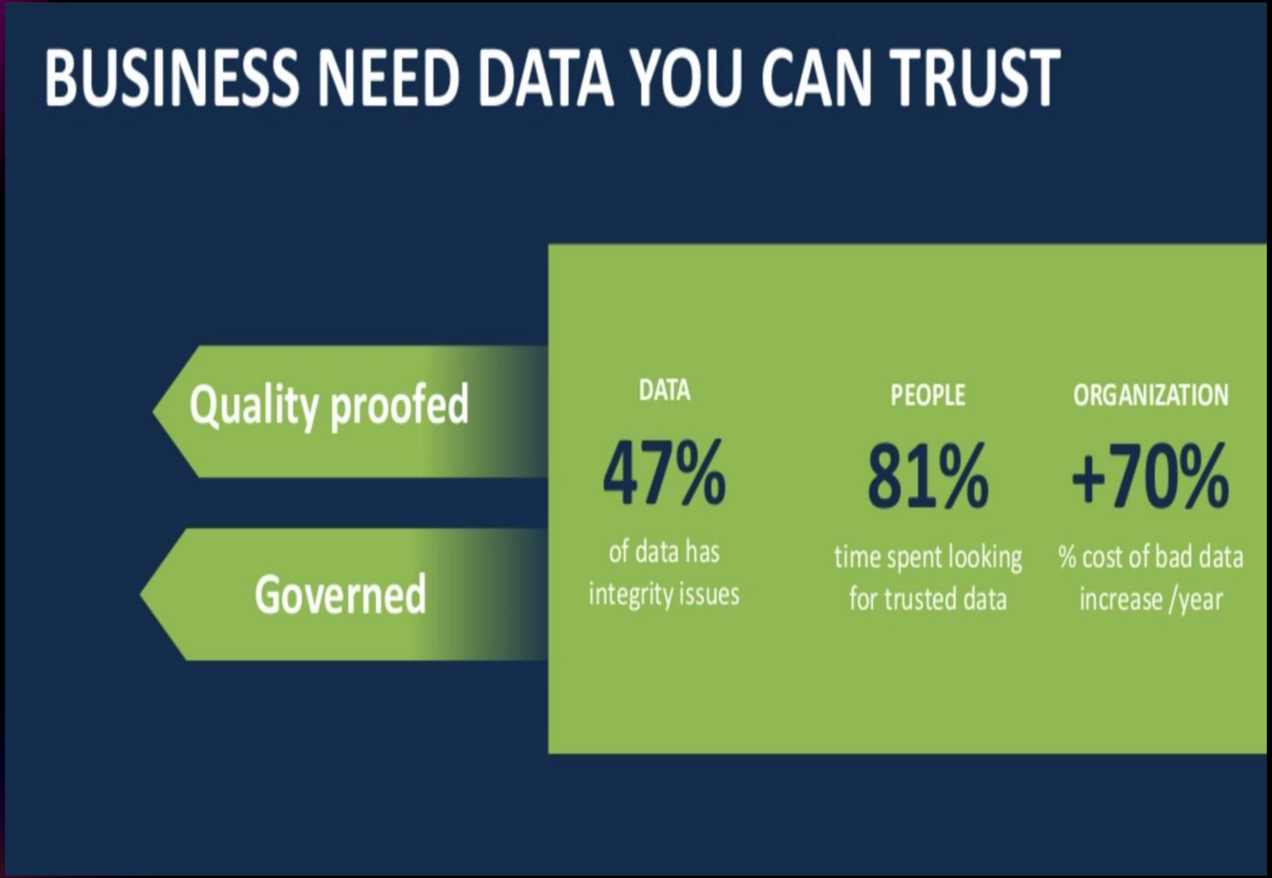
Data creation will grow to 163 zettabytes (ZB) by 2025[1]

Data platforms

Need to scale exponentially

[1] AWS Data Lake Storage Infrastructure paper

Why data governance matters



BUSINESS NEED DATA YOU CAN TRUST

Quality proofed

Governed

DATA	PEOPLE	ORGANIZATION
47%	81%	+70%
of data has integrity issues	time spent looking for trusted data	% cost of bad data increase /year

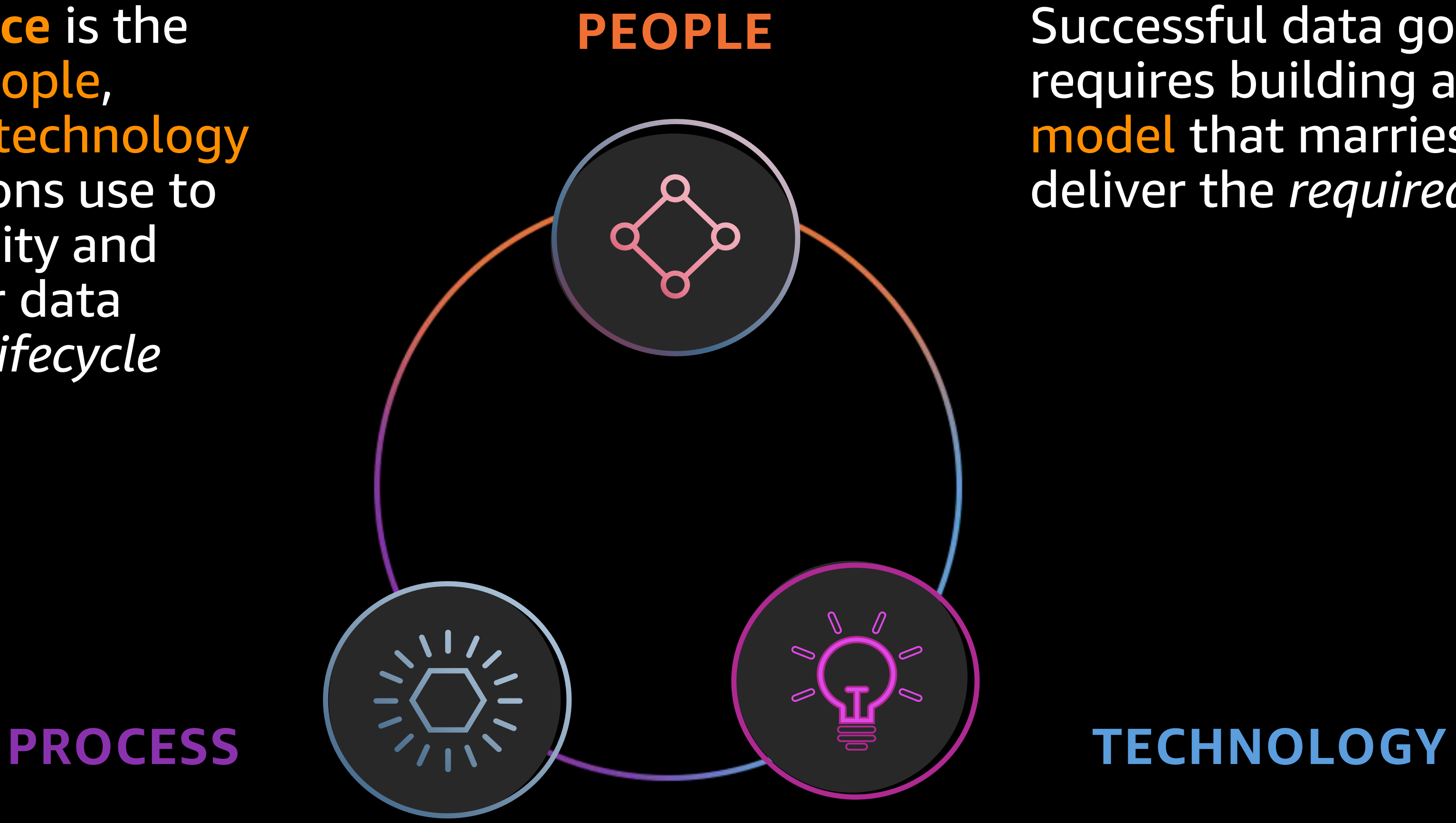
Learn more

Organizations lack data knowledge for efficient and effective data governance activities and time spent on data governance is wasted.

Data governance is no longer optional for enterprise organizations. They are finally realizing the value of data as an asset that needs to be protected, managed, and maintained to increase asset value.

Data governance on AWS

- **Data Governance** is the collection of people, processes, and technology that organizations use to ensure the quality and security of their data *throughout its lifecycle*



Successful data governance requires building an operating model that marries all three to deliver the *required capabilities*

Focusing on business outcomes

Customer experience

Built a customer engagement service using a Modern Data Architecture to serve over eight million developers working with 190k+ businesses in 100+ countries

Twilio

Real-time insights to give tens of millions of users personalized streaming recommendations

Disney+

Increased the use of self-service analytics platform by over 40% for daily active fans—sharing richer information in near real-time

OneFootball

Personalizes searches for better customer experience and gets fewer returns due to improved sizing recommendations

Zappos

Agility and innovation

Accelerates zero-carbon transition with automated energy predictions and maximized wind farm energy production

ENGIE

Helps drive better insights needed to make key race-time decisions, giving a technological edge over competitors

Toyota Racing Development

With Amazon Managed Streaming for Apache Kafka, the company is able to experiment with big changes safely with little risk

New Relic

Built a sophisticated infectious disease tracker in four months for retirement community residents and employees

Erickson Living

Cost optimization

Manages over 150 PB of data at $5 per terabyte of data scanned

FINRA

Shifting to AWS saves more than $2 million annually in data storage costs

INVISTA

AWS Analytics reduced operational costs by over 30% while freeing software engineers of low-value work

Pinterest

Amazon EMR as its core ML platform allows for more accurate ML models 80% faster at an 80% lower cost

Eightfold.ai

Performance and scale

Moved to a Modern Data Architecture to ingest 70 billion records per day, and now runs Amazon Redshift queries 32% faster

Nasdaq

Scalability and cost efficiency during a global pandemic with 20x increase in ventilator production while reducing first-pass inspection failures by 60%

Vyaire Medical

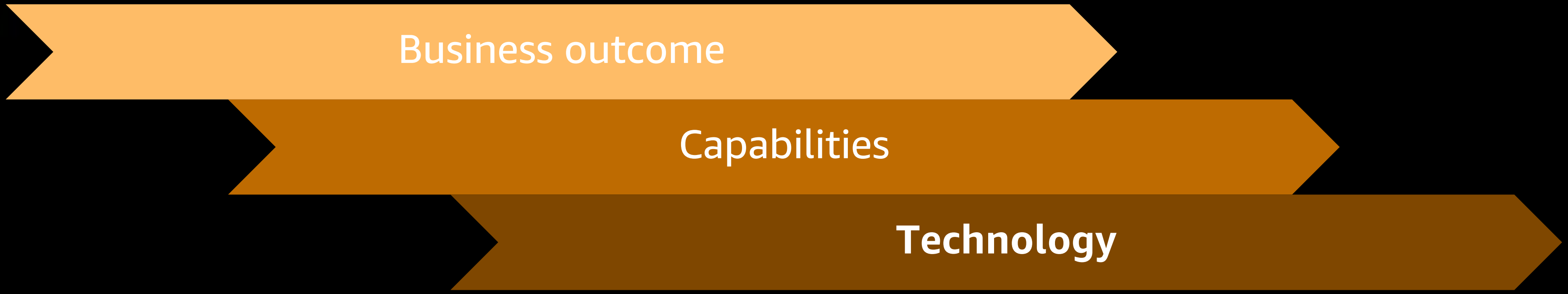
Scaled ingestion to six billion documents per day using Amazon OpenSearch Service (successor to Amazon Elasticsearch Service)

Pearson

Had the tools to support a 101% increase in language learners

Duolingo

Work backwards from the outcome

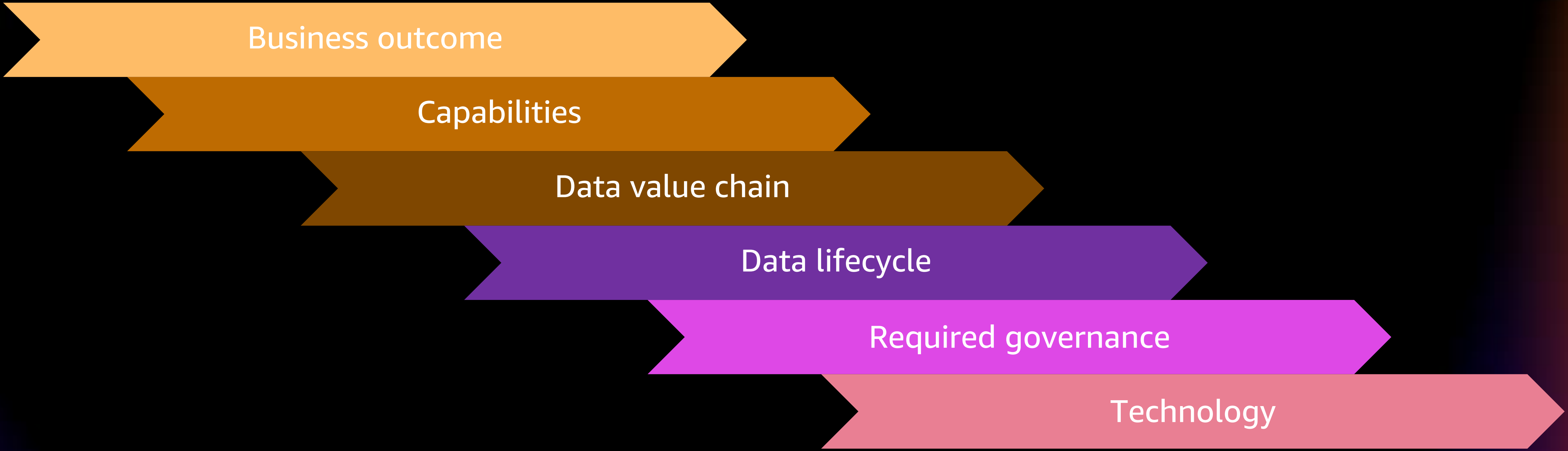


Business outcome

Capabilities

Technology

Work backwards to governance



Business outcome

Capabilities

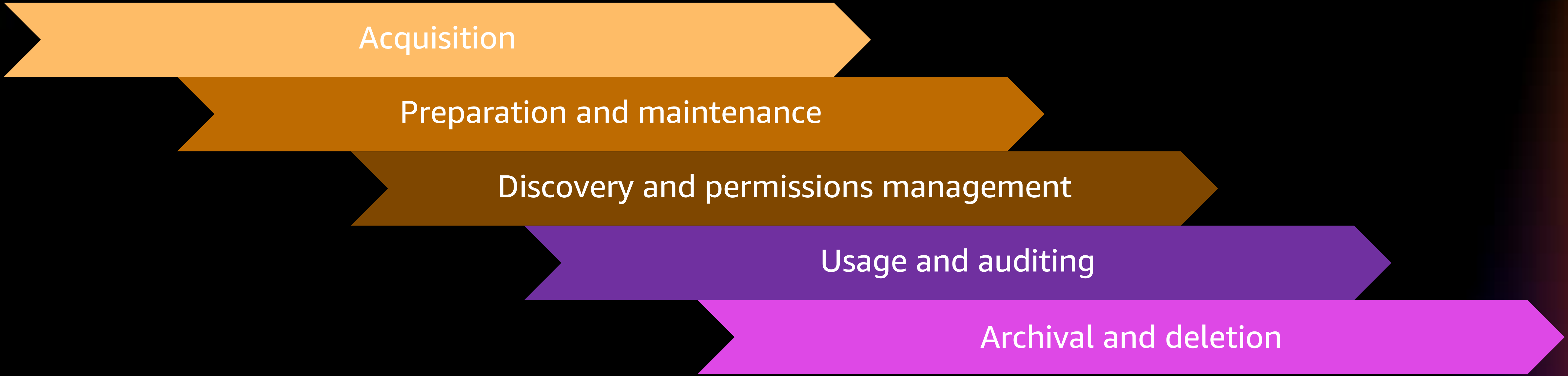
Data value chain

Data lifecycle

Required governance

Technology

Data lifecycle



Acquisition

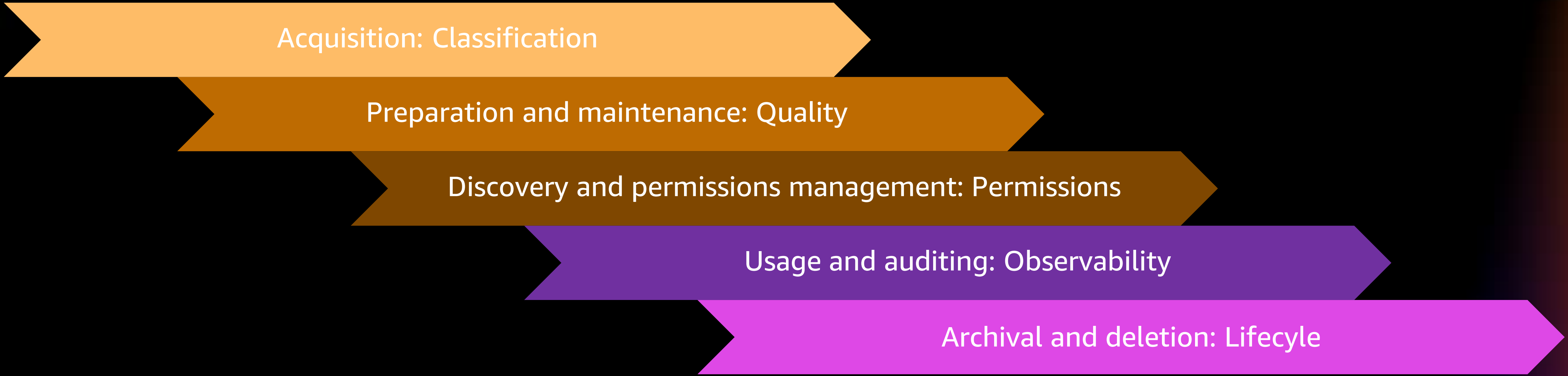
Preparation and maintenance

Discovery and permissions management

Usage and auditing

Archival and deletion

Data lifecycle: Capabilities



Acquisition: Classification

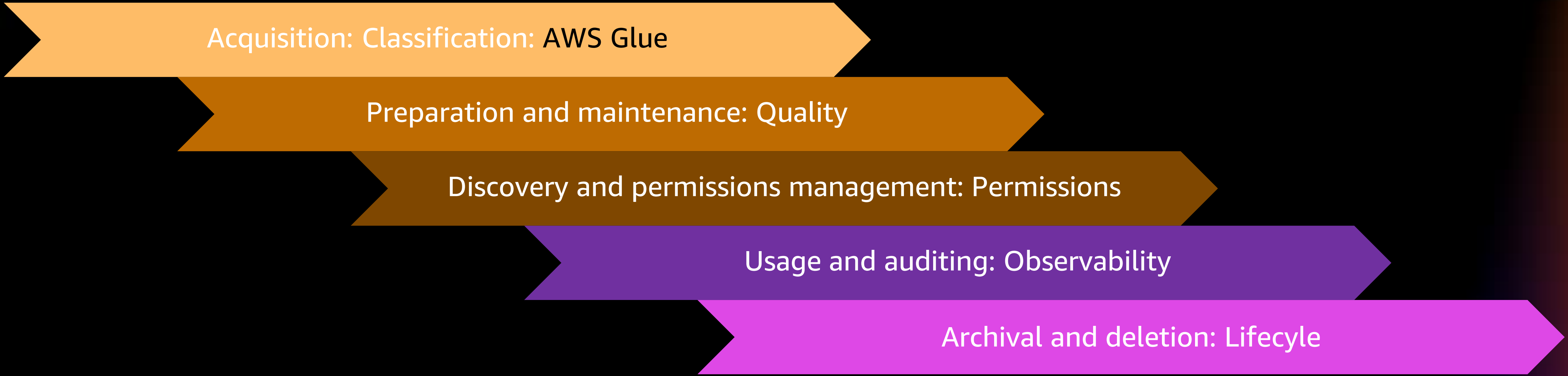
Preparation and maintenance: Quality

Discovery and permissions management: Permissions

Usage and auditing: Observability

Archival and deletion: Lifecyle

Data lifecycle: Capabilities: Technology



Acquisition: Classification: AWS Glue

Preparation and maintenance: Quality

Discovery and permissions management: Permissions

Usage and auditing: Observability

Archival and deletion: Lifecyle

AWS Glue: Serverless data integration

Connect & transform

Glue connectors

Built-in data transforms

Easy to migrate

Centralized & Unified Data Governance

Glue Data Catalog

Glue Crawlers

Lake Formation

User Productivity and Data Ops

Variety of Persona

Productivity tools

Data ops tools

Serverless Execution Engine

Scalable Execution engine

Flexible, cost-effective options

Execution monitoring

AWS Glue PII Detection and remediation

THREE SIMPLE STEPS

1 Type of Scan

- Full Scan
- Sample Scan

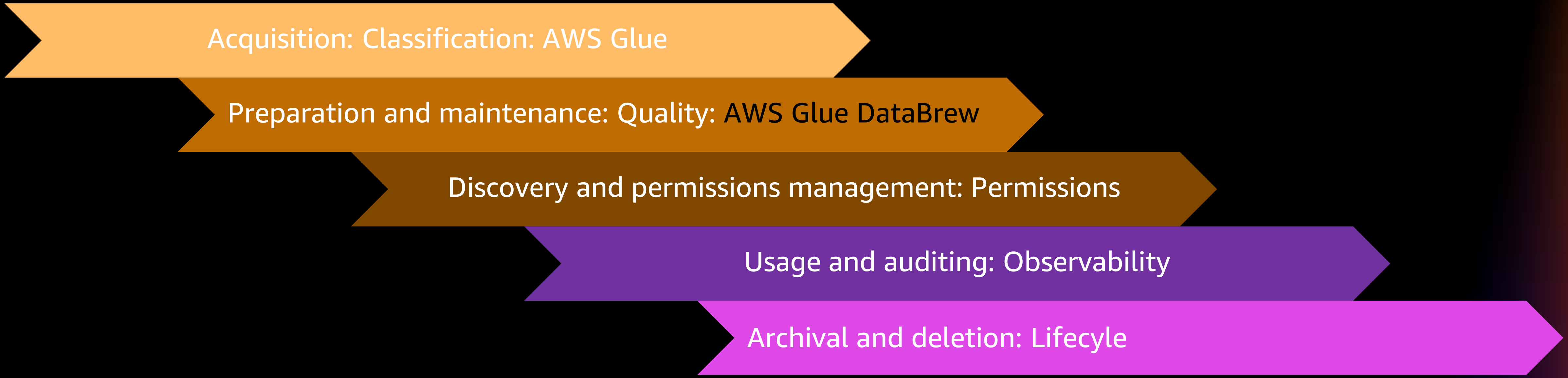
2 Entities to detect

- Built-in Entities (e.g. SSN, passport)
- Custom Entities

3 Remediation ✓

- Store results
- Redact/mask results

Data lifecycle: Capabilities: Technology



Acquisition: Classification: AWS Glue

Preparation and maintenance: Quality: AWS Glue DataBrew

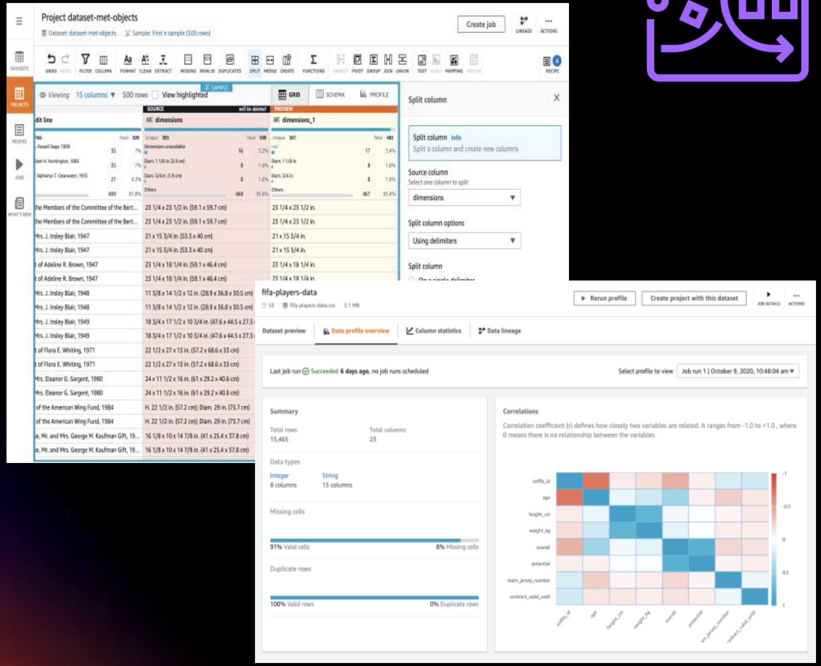
Discovery and permissions management: Permissions

Usage and auditing: Observability

Archival and deletion: Lifecyle

AWS Glue DataBrew

Visual data profiling and preparation



Project dataset-met-objects

Dataset: dataset-met-objects Sample: First n sample (500 rows)

Create job

Viewing 15 columns ▼ 500 rows □ View highlighted

dit line	dimensions	dimensions_1
the Members of the Committee of the Bert...	23 1/4 x 23 1/2 in. (59.1 x 59.7 cm)	23 1/4 x 23 1/2 in.
the Members of the Committee of the Bert...	23 1/4 x 23 1/2 in. (59.1 x 59.7 cm)	23 1/4 x 23 1/2 in.
Mrs. J. Insley Blair, 1947	21 x 15 3/4 in. (53.3 x 40 cm)	21 x 15 3/4 in.
Mrs. J. Insley Blair, 1947	21 x 15 3/4 in. (53.3 x 40 cm)	21 x 15 3/4 in.
t of Adeline R. Brown, 1947	23 1/4 x 18 1/4 in. (59.1 x 46.4 cm)	23 1/4 x 18 1/4 in.
t of Adeline R. Brown, 1947	23 1/4 x 18 1/4 in. (59.1 x 46.4 cm)	23 1/4 x 18 1/4 in.
Mrs. J. Insley Blair, 1948	11 3/8 x 14 1/2 x 12 in. (28.9 x 36.8 x 30.5 cm)	
Mrs. J. Insley Blair, 1948	11 3/8 x 14 1/2 x 12 in. (28.9 x 36.8 x 30.5 cm)	
Mrs. J. Insley Blair, 1949	18 3/4 x 17 1/2 x 10 3/4 in. (47.6 x 44.5 x 27.3	
Mrs. J. Insley Blair, 1949	18 3/4 x 17 1/2 x 10 3/4 in. (47.6 x 44.5 x 27.3	
t of Flora E. Whiting, 1971	22 1/2 x 27 x 13 in. (57.2 x 68.6 x 33 cm)	
t of Flora E. Whiting, 1971	22 1/2 x 27 x 13 in. (57.2 x 68.6 x 33 cm)	
Mrs. Eleanor G. Sargent, 1980	24 x 11 1/2 x 16 in. (61 x 29.2 x 40.6 cm)	
Mrs. Eleanor G. Sargent, 1980	24 x 11 1/2 x 16 in. (61 x 29.2 x 40.6 cm)	
of the American Wing Fund, 1984	H. 22 1/2 in. (57.2 cm); Diam. 29 in. (73.7 cm)	
of the American Wing Fund, 1984	H. 22 1/2 in. (57.2 cm); Diam. 29 in. (73.7 cm)	
e, Mr. and Mrs. George M. Kaufman Gift, 19...	16 1/8 x 10 x 14 7/8 in. (41 x 25.4 x 37.8 cm)	
e, Mr. and Mrs. George M. Kaufman Gift, 19...	16 1/8 x 10 x 14 7/8 in. (41 x 25.4 x 37.8 cm)	

Split column

Split column Info
Split a column and create new columns

Source column
Select one column to split
dimensions ▼

Split column options
Using delimiters ▼

Split column
○ On a single delimiter

fifa-players-data

S3 fifa-players-data.csv 3.1 MB

▶ Rerun profile Create project with this dataset

Dataset preview | Data profile overview | Column statistics | Data lineage

Last job run ✓ Succeeded **6 days ago**, no job runs scheduled

Select profile to view: Job run 1 | October 9, 2020, 10:48:04 am ▼

Summary

Total rows: 15,465
Total columns: 23

Data types
Integer: 8 columns
String: 15 columns

Missing cells
91% Valid cells 8% Missing cells

Duplicate rows
100% Valid rows 0% Duplicate rows

Correlations

Correlation coefficient (r) defines how closely two variables are related. It ranges from -1.0 to +1.0 , where 0 means there is no relationship between the variables.

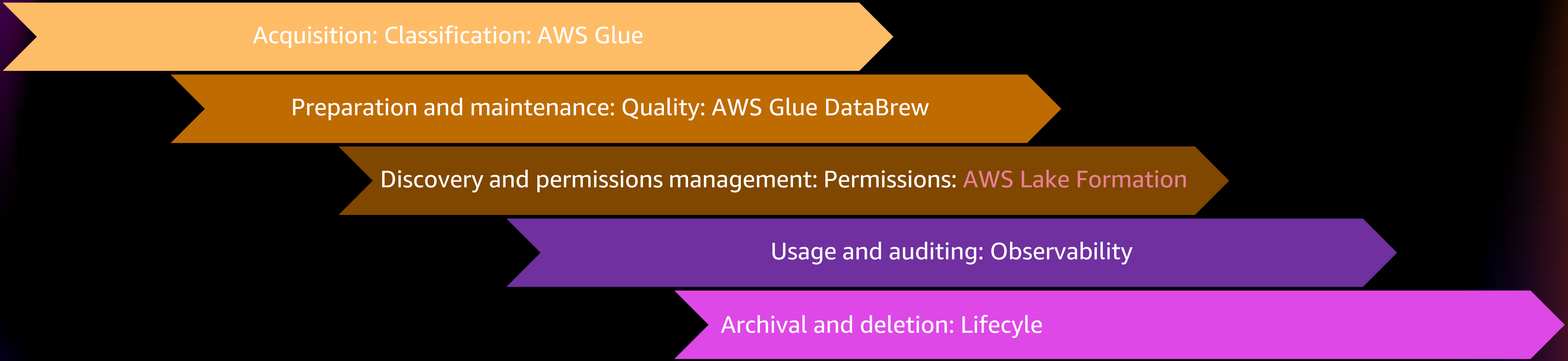
Clean and normalize data with a visual interface

250+ built-in transformations without writing code

Profile data to understand data patterns and anomalies

Work on large datasets at scale

Data lifecycle: Capabilities: Technology



Acquisition: Classification: AWS Glue

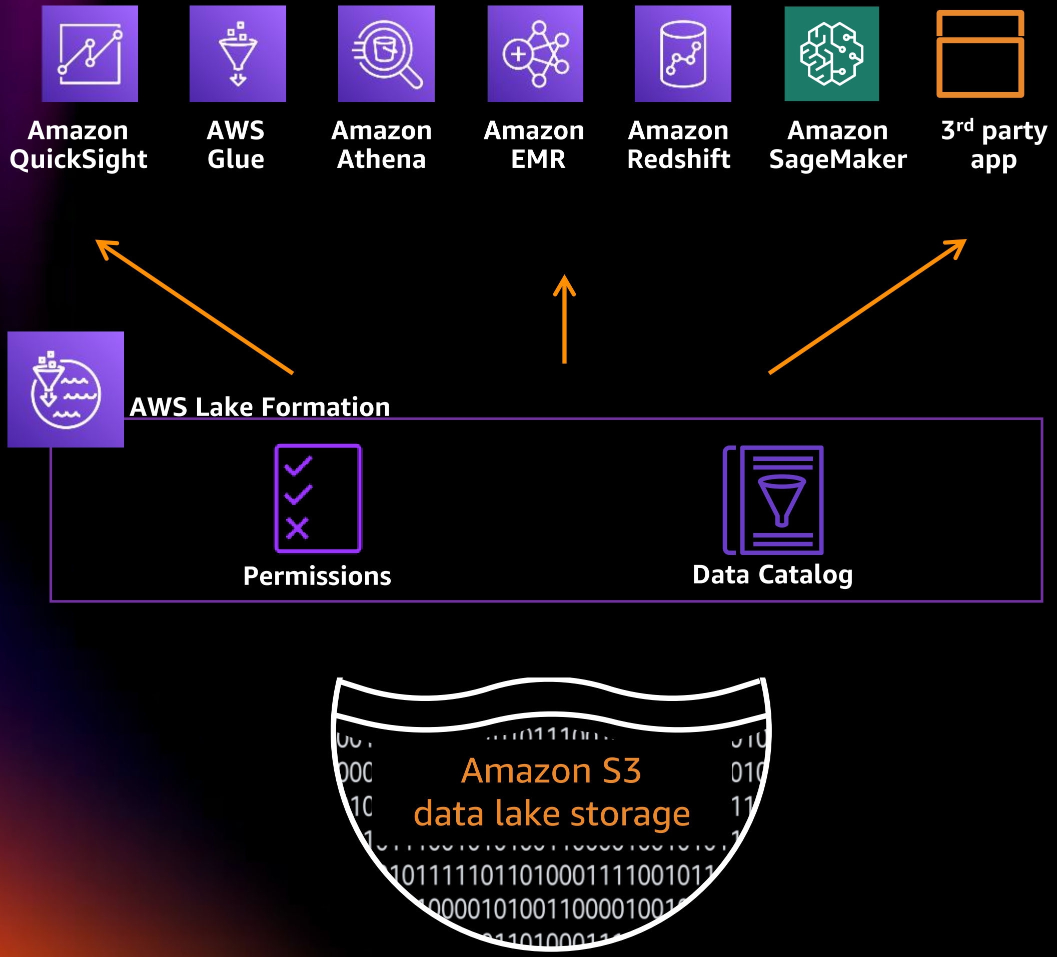
Preparation and maintenance: Quality: AWS Glue DataBrew

Discovery and permissions management: Permissions: AWS Lake Formation

Usage and auditing: Observability

Archival and deletion: Lifecyle

AWS Lake Formation permissions model



Database-style fine-grained permissions

Fine-grained permissions on catalog resources

S3 access managed by permission on resources

LF-Tag based access control (LF-TBAC) to scale

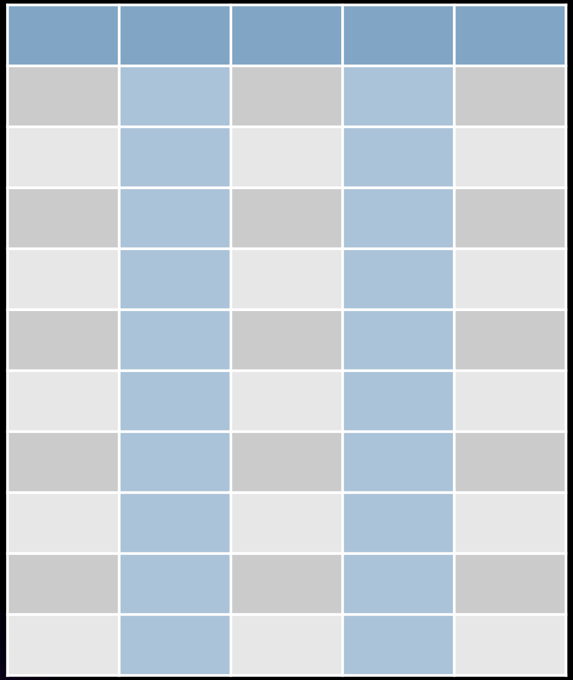
Integrated with services and tools

Easy to audit permissions and access

NEW!

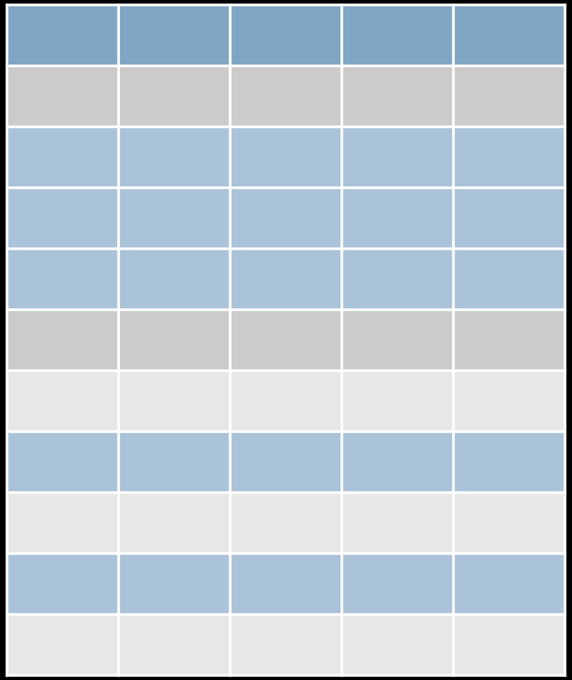
AWS Lake Formation permissions on:

Columns



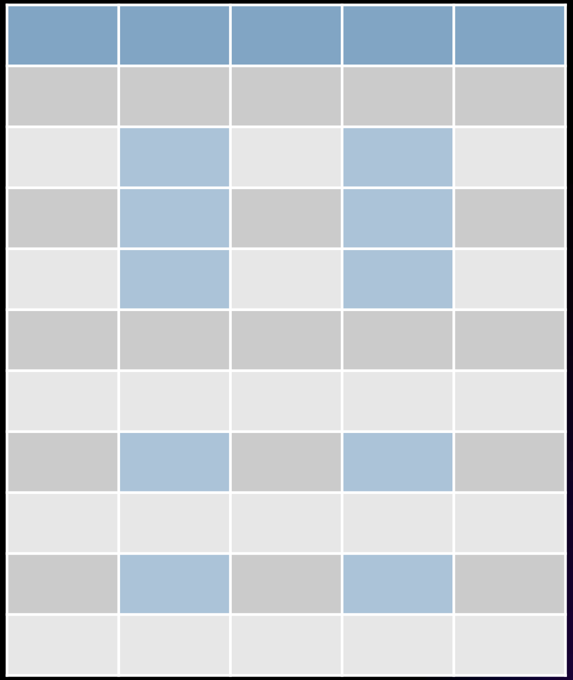
Specify include or exclude list

Rows



Specify row filter with PartiQL

Cells



Combine column and row filters

Customers are efficiently securing data lakes

freee

“We found AWS Lake Formation easy to use to build and secure our data lake. Without AWS Lake Formation, we would have to make constant access policy updates to Amazon S3 when we added more users and data . . . With the adoption of AWS Lake Formation, we are able to . . . reduce Amazon S3 policy edits by over 90% . . ."

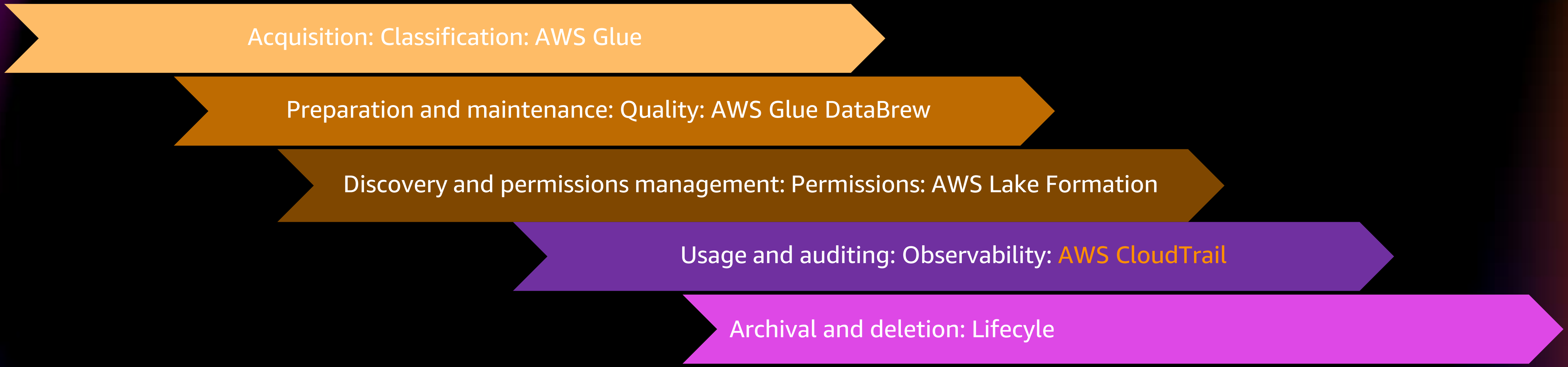
Hisatoshi Imaoka
Tech. Lead Data Infrastructure
freee K.K.

FANDUEL

“AWS Lake Formation enables us to create a secure Data Lake with fine-grained controls on our users’ personal information. Our Analysts can now deliver much needed insights, much faster without compromising the governance and security policies."

Damian Grech
Data Engineering, Sr. Manager
FanDuel

Data lifecycle: Capabilities: Technology



Acquisition: Classification: AWS Glue

Preparation and maintenance: Quality: AWS Glue DataBrew

Discovery and permissions management: Permissions: AWS Lake Formation

Usage and auditing: Observability: AWS CloudTrail

Archival and deletion: Lifecyle

AWS CloudTrail provides visibility and audit logs for AWS

- Track user and resource activity across your AWS Data Lake and resources for governance and auditing.
- Identify and respond to unusual usage based on automated analysis.

Capture

Record activity as AWS CloudTrail events

Store

Retain events logs in secure S3 bucket

Act

Trigger actions when important events are detected

Review

Analyze findings or recent and historical activity

AWS CloudTrail can help customers in a variety of ways

Compliance Aid

Visibility into Activity

Anomaly Detection

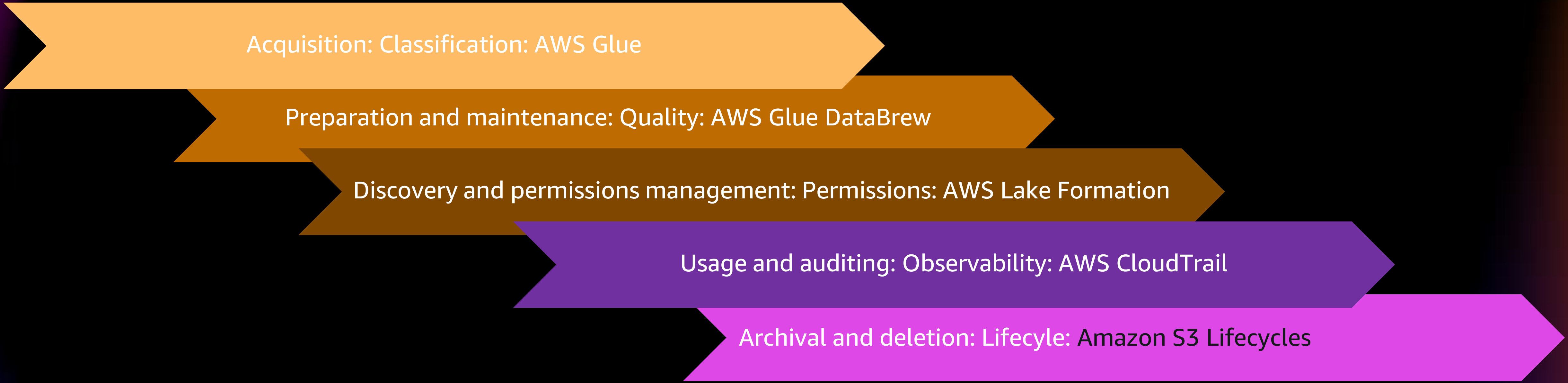
Detect Data Exfiltration

Automate Security Analysis

Analyze Permissions

Detect Unusual Activity

Data lifecycle: Capabilities: Technology



Acquisition: Classification: AWS Glue

Preparation and maintenance: Quality: AWS Glue DataBrew

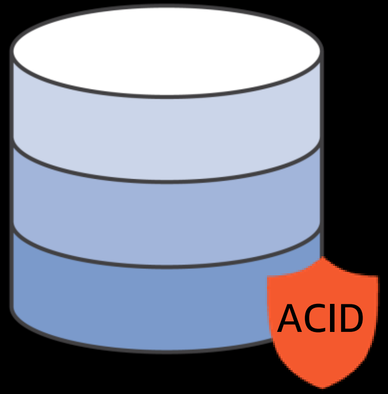
Discovery and permissions management: Permissions: AWS Lake Formation

Usage and auditing: Observability: AWS CloudTrail

Archival and deletion: Lifecyle: Amazon S3 Lifecycles

AWS Lake Formation Governed Tables

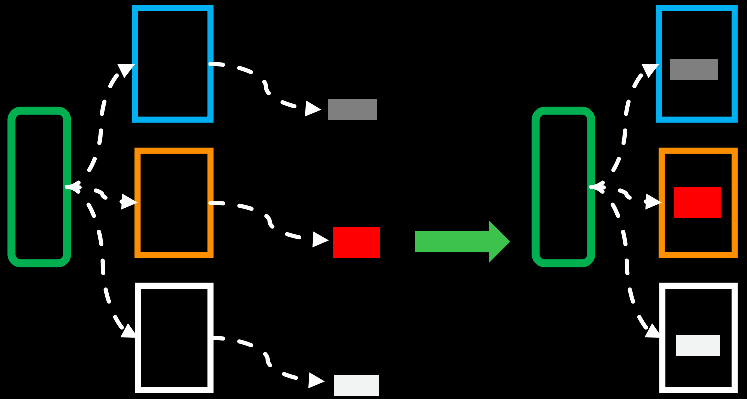
Simplify data ingestion and data management



ACID transactions

Consistent across tasks
Insert, update, delete
Converge batch & real-time

Reliable



Storage optimization

Auto-compact small files
Push-down filters
Reduce data scan

Performant

Time travel

Data history
Reproduce experiments
Audit changed data

Versioned

Transactions simplify development…

““. . . Transactional ETL processes are an important part of how we ensure data integrity and . . . required additional development time and complexity. We’re excited about AWS Lake Formation Transactions’ ability to simplify our ETL and reduce the overall effort needed to produce trustworthy data in our data lake."

Rob Hruska
Engineering Director
Hudl

“PowerBuy decided to *forego traditional database-based architecture* in favor of a data lake using *AWS Lake Formation Governed Tables*. Governed Tables make it easy to insert, update and delete data for all of our PowerBuy products using highly scalable ACID transactions. With Governed Tables we can *release new products quicker* like PowerBuy AI and PowerBuy Dashboard *without worrying about scaling…*”

Thu Truong
CTO and Co-Founder
PowerBuy

Amazon S3 is the best place to build data lakes

durability, availability, scalability

security, compliance, audit

get data in

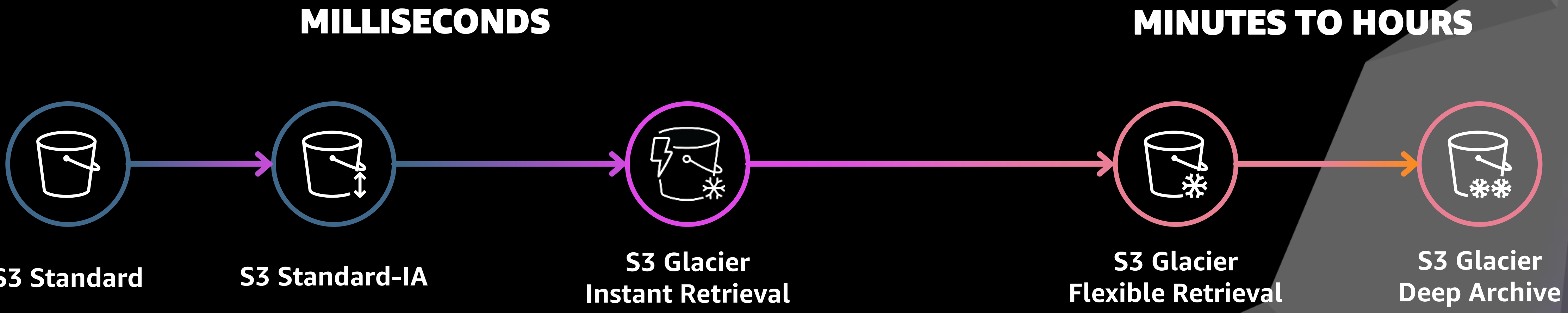
Cost-optimized

in place

Cold storage and instant archive retrieval

broadest set of analytics and ML

Amazon S3 and S3 Glacier storage classes



MILLISECONDS

MINUTES TO HOURS

S3 Standard

S3 Standard-IA

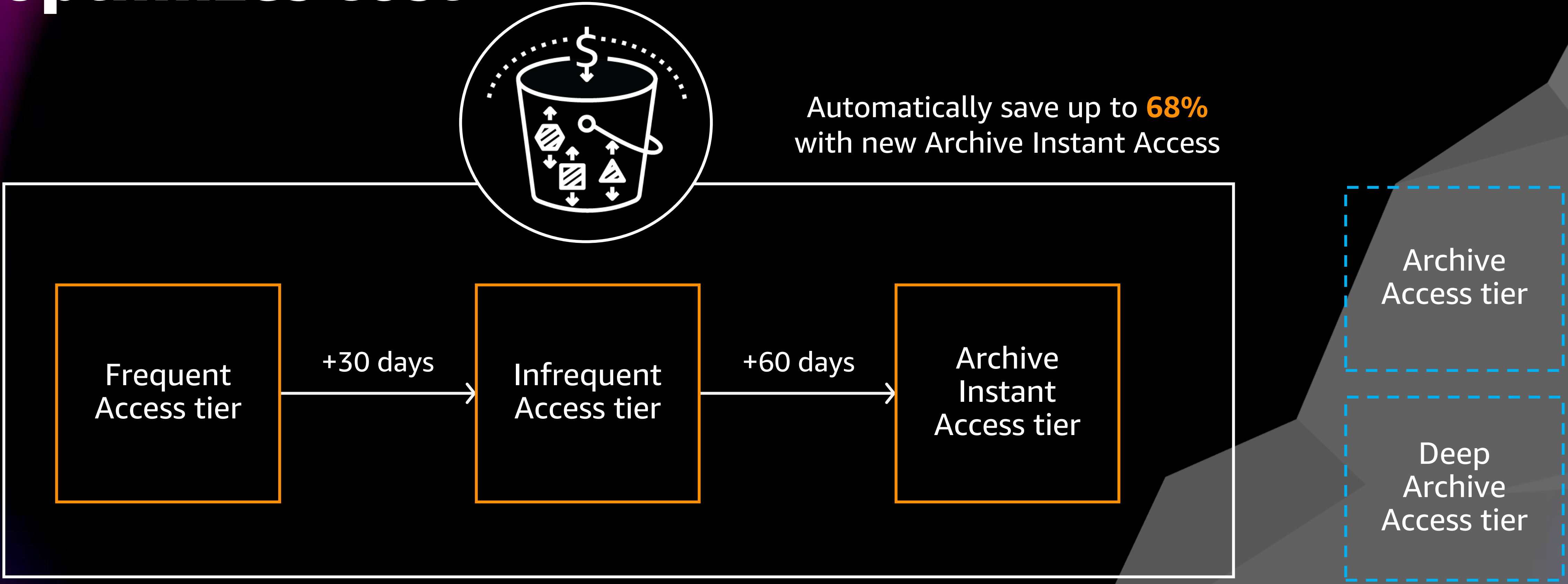
S3 Glacier Instant Retrieval

S3 Glacier Flexible Retrieval

S3 Glacier Deep Archive

Cost optimized

Amazon S3 Intelligent-Tiering automatically optimizes cost



Automatically save up to **68%**
with new Archive Instant Access

Frequent Access tier

+30 days

Infrequent Access tier

+60 days

Archive Instant Access tier

Archive Access tier

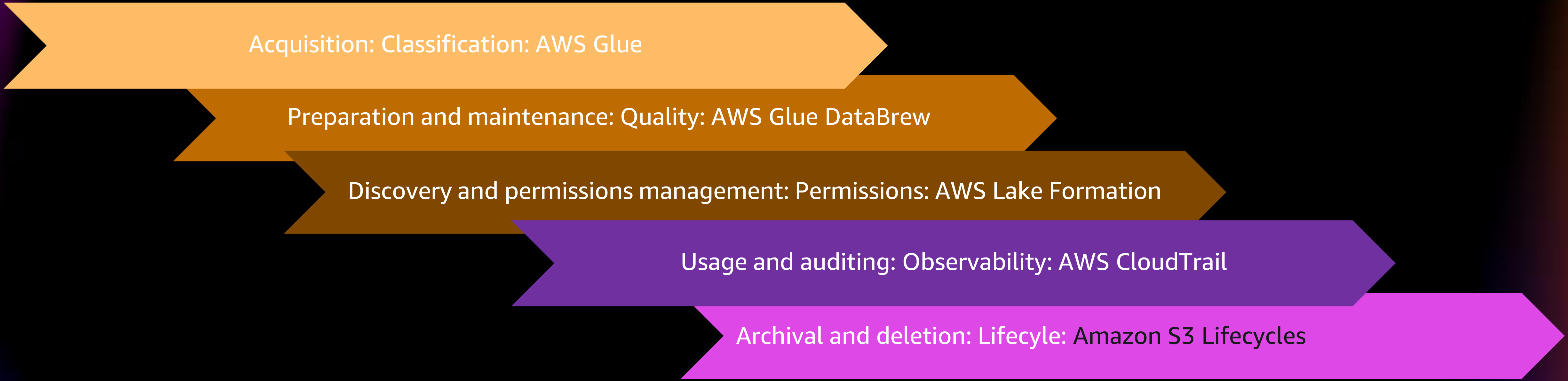
Deep Archive Access tier

Milliseconds access (automatic)

Minutes to hours (optional)

Cost optimized

Data lifecycle: Capabilities: Technology



Acquisition: Classification: AWS Glue

Preparation and maintenance: Quality: AWS Glue DataBrew

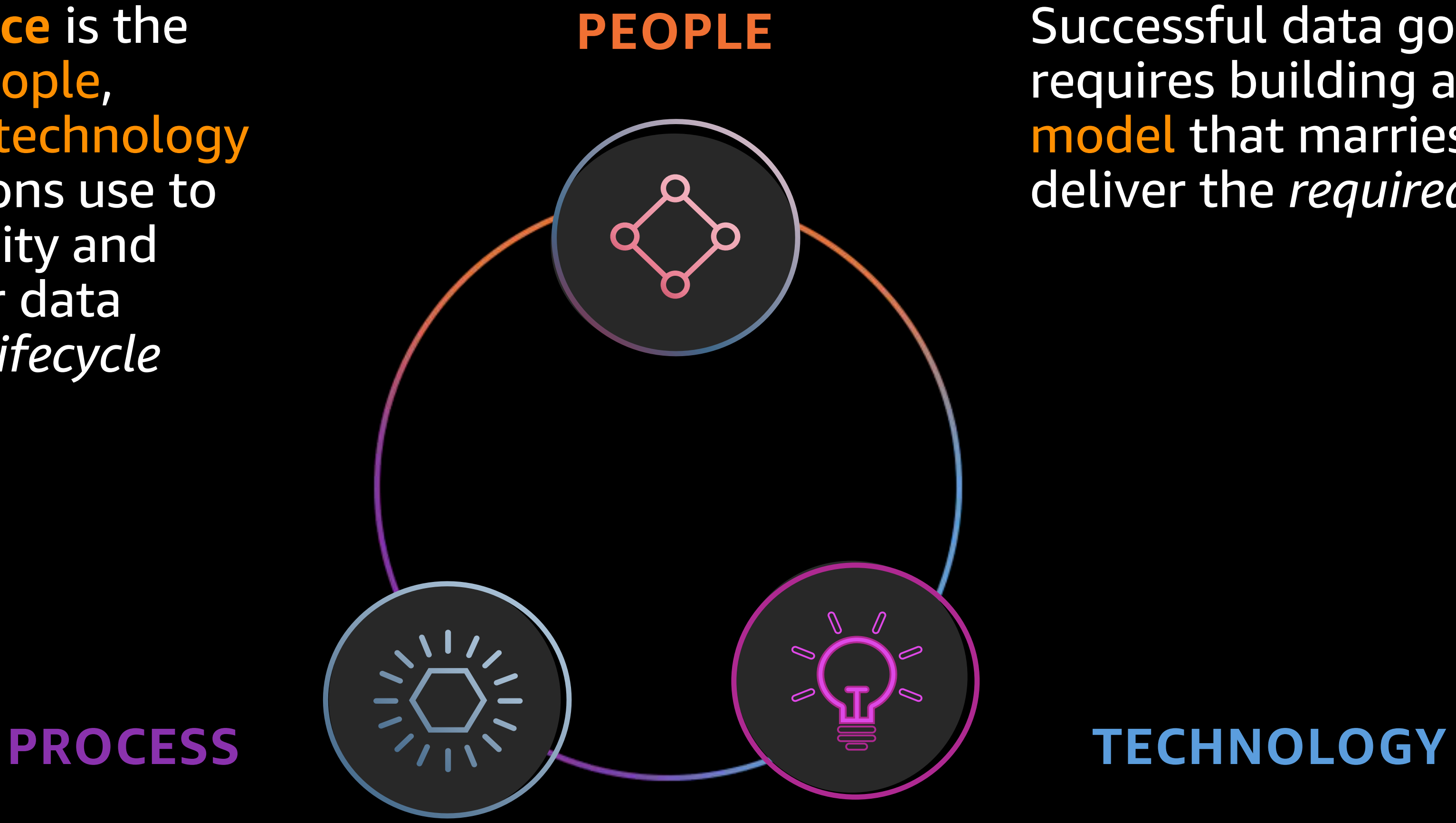
Discovery and permissions management: Permissions: AWS Lake Formation

Usage and auditing: Observability: AWS CloudTrail

Archival and deletion: Lifecyle: Amazon S3 Lifecycles

Data governance on AWS

- **Data Governance** is the collection of people, processes, and technology that organizations use to ensure the quality and security of their data *throughout its lifecycle*



Successful data governance requires building an operating model that marries all three to deliver the *required capabilities*

Visit the AWS Data resource hub

A modern data strategy can help you manage, act on, and react to your data so you can make better decisions, respond faster, and uncover new opportunities. Dive deeper with these resources today.

- Harness data to reinvent your organization
- In unpredictable times, a data strategy is key
- Make data a strategic asset
- Rewiring your culture to be data-driven
- Put your data to work with a modern analytics approach
- … and more!

https://tinyurl.com/data-hub-aws

Visit resource hub

AWS Training and Certification for Data and Analytics

AWS Data & Analytics FREE Training Resources

Discover how to harness data, one of the world's most valuable resources, and innovate at scale.

https://bit.ly/3Ntlhy7

AWS Data Analytics Learning Plan

This learning plan expose you to the fastest way to get answers from all your data to all your users. It can also help prepare you for the AWS Certified Data Analytics - Specialty certification exam.

https://bit.ly/3wBVjD1

AWS Certified Data Analytics - Specialty

Earning AWS Certified Data Analytics – Specialty validates expertise in using AWS data lakes and analytics services.

https://go.aws/3lwF0RR

Thank you for attending AWS Innovate – Data Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.

- aws-apj-marketing@amazon.com
- twitter.com/AWSCloud
- facebook.com/AmazonWebServices
- youtube.com/user/AmazonWebServices
- slideshare.net/AmazonWebServices
- twitch.tv/aws

Thank you!

Title only layout

What is data governance?

- Database admins (DBAs)
- DevOps engineers
- LOB knowledge workers
- Product managers
- IT operations
- IT security and governance
- VP/director analytics

- Architects
- Application developers
- Business intelligence (BI) analysts
- CxO
- Data engineers, operations
- Data modelers
- Data scientists
- Data warehouse admins

ENGIE builds the Common Data Hub on AWS, accelerates zero-carbon transition

Challenge

ENGIE's decentralized global customer base had accumulated lots of data, and it required a smarter, unique approach and solution to align its initiatives and to efficiently provide data across its global business units.

Solution

ENGIE built its Common Data Hub data lake on AWS, enabling the company's business units to collect and analyze data to support a data-driven strategy and to lead the zero-carbon transition.

Result

- Collected 95 TB of data across 351 projects
- Automated energy predictions
- Maximized wind farm energy production

Central area to minimize or eliminate gradient backgrounds on content slides
(gradients are fine on non-content slides (opening, closing, title, section, demo, etc.)

“

”

Recap/summary

Other resources

Color palette feedback

EXAMPLES

Current palette

										(hyperlinks)
0	255	40	242	254	222	240	229	233	64	191
0	255	40	244	143	72	130	230	127	104	112
0	255	40	244	1	230	112	255	147	138	213

Recommended palette

										(hyperlinks)
0	255	6	242	254	222	74	21	105	40	181
0	255	0	244	143	72	201	163	20	73	145
0	255	73	244	1	230	209	99	225	189	253

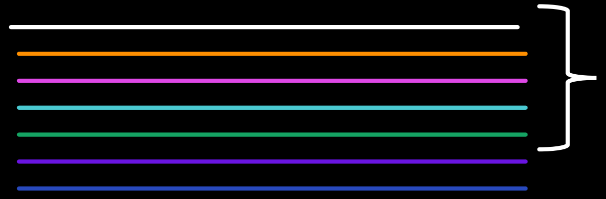
Or whatever color is predominant in the slide background if a solidish color is selected for content slides

Text
Text
Text
Text
Text

Usable on this background color

Text
Text
Hyperlink

Usable on this background color



Usable on this background color