



Modern Data Architecture with Quality Software

Natasha Nicolai, AWS Public Sector
HHS Data Analytics Leader

What State and Local Government are looking for

Improve User Experiences

Delayed benefits to Citizens, burdensome hurdles to access, lack of visibility into case status, application backlogs and overburdened case workers, lack of user-centered design.

Make Better Use of Data





Lack of visibility into program operations and errors; lack of timely reporting and program compliance; unnecessary administrative friction; program fraud, waste and abuse.

Increase Agility

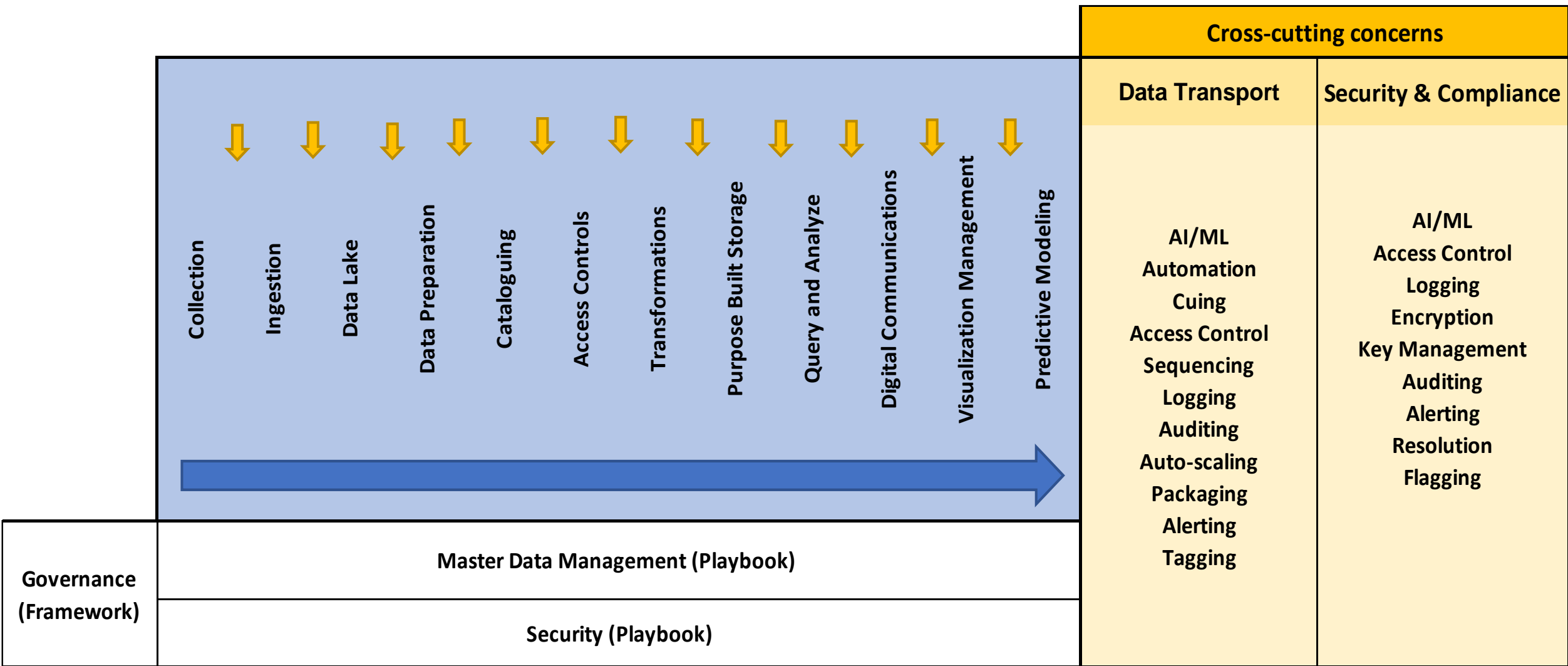
Cannot adapt to policy changes and new programs; legacy systems result in technical debt and bloated program O&M costs; cannot scale in and scale out resources.

Lower Cost

High infrastructure costs that are not aligned with usage.

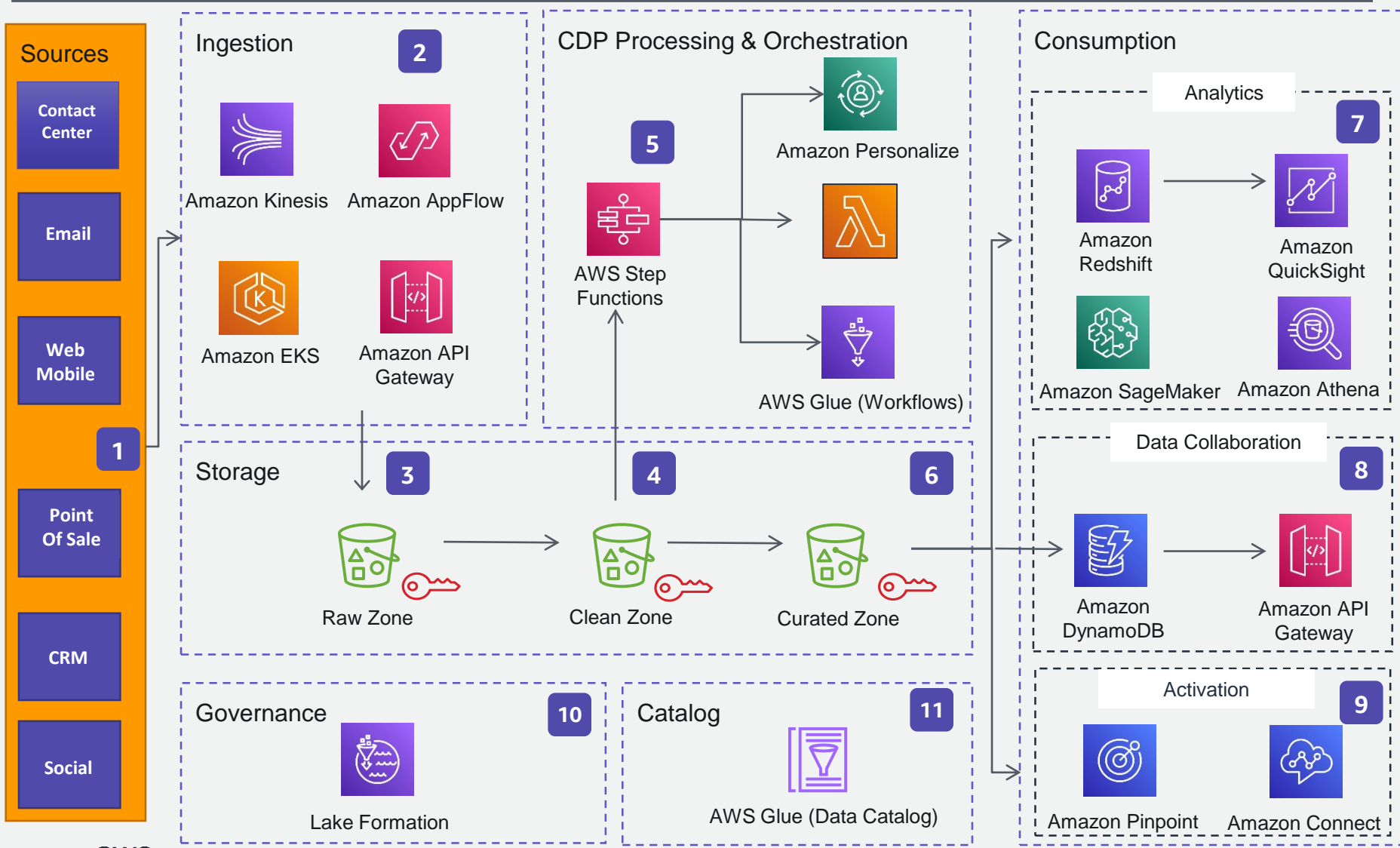
Data Lifecycle		
Capture	Curate	Infer
<p>How is the data collected?</p>  <p>Where are the data sources and how are they ingested? Do they come into a central system? Are controls federated?</p>  <p>Can clients access and contribute directly?</p> <p>Team: Central and Disbursed Program + IT</p>	<p>How is the data manipulated?</p>  <p>Datasets and databases should be intentional. How can one leverage purpose built opportunities?</p> <p>Are the right access tools and controls in place?</p> <p>Team: Data + IT + Program</p>	 <p>What is the data narrative?</p> <p>What are the critical insights and priority questions?</p> <p>Who needs answers to the questions and does IT meet those needs?</p> <p>Team: Data + Program</p>

Stages of the Data Lifecycle



Guidance for Customer Data Platform on AWS

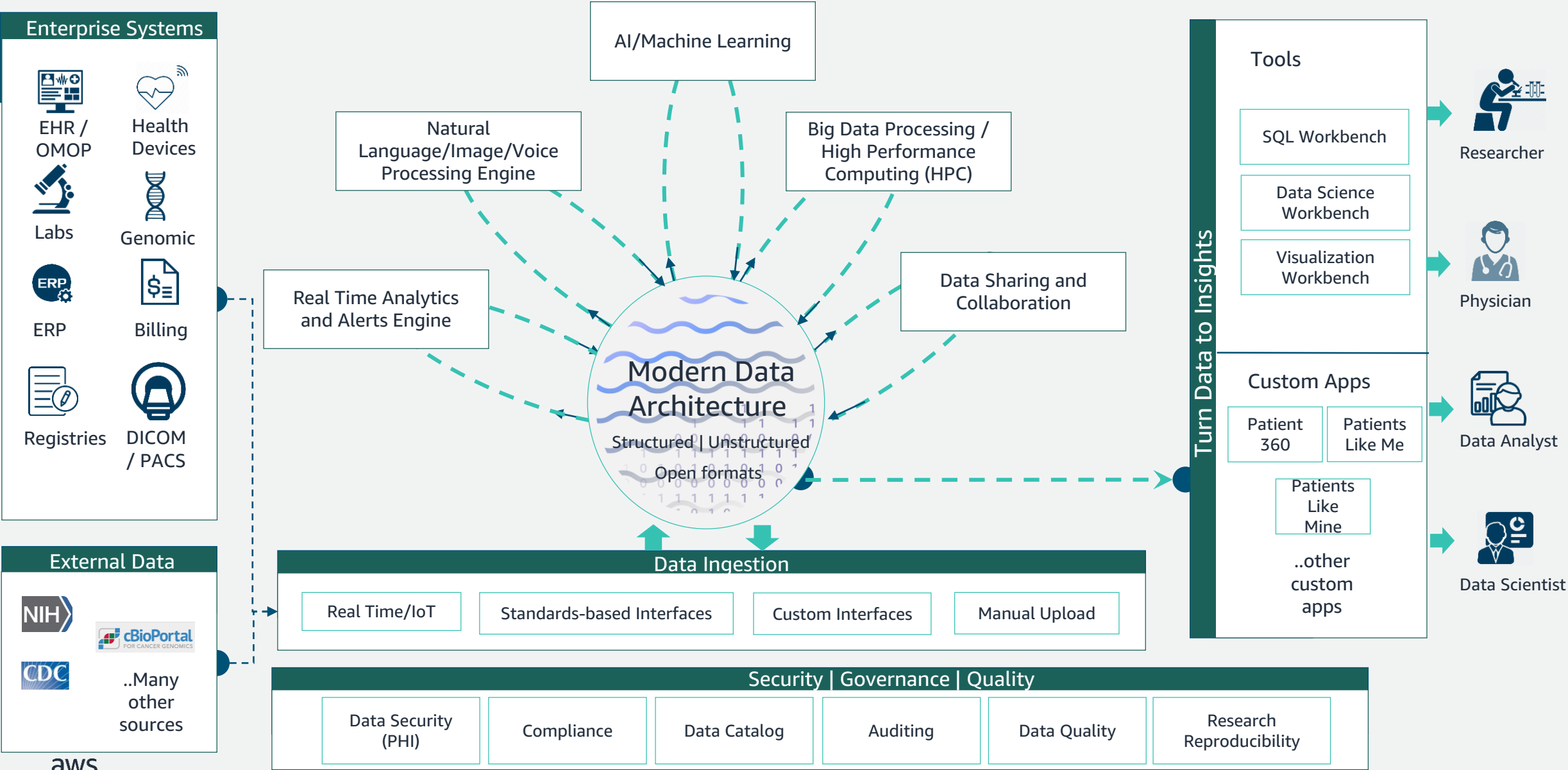
This guidance provides a reference architecture showing best practices in the building of a customer data platform covering data ingestion, identity resolution, segmentation, analysis and activation.



- 1 Source systems including customer interactions, clickstreams, call center logs etc.
- 2 Ingesting data across customer touchpoints into marketing CDP data lake using variety of protocols
- 3 Ingested data in its original, immutable format on S3 Raw Zone bucket
- 4 Transforming raw data into efficient data formats such as Parquet or Avro into Clean Zone S3 bucket
- 5 CDP processing and pipeline orchestration using purpose-built data processing components and transformation libraries
- 6 Curated zone contains data ready for consumption post CDP processing organized by subject areas, segments and profiles
- 7 Analytics layer natively integrated with Curated zone for analytics, dashboards, ad hoc reporting, and ML purposes
- 8 Aggregate customer data across platforms and publish customer APIs for consumption
- 9 Activate multiple customer channels such as mobile push, voice, and email for targeted marketing communications
- 10 Enforce fine-grained access controls on catalog tables, columns and rows on data lake
- 11 Manage business and technical metadata with versioning at scale



Modern Health Data Platform on AWS



Data quality in software development:

How do you sanitize your inputs?

Have you codified data policy into your APIs, and is that policy-driven so that you can change policy without code changes?

How do you think about separating compute from data?

Are data storage decisions intentionally made and aligned with purpose?

How are you thinking about the security of your application and how it operates with data?

Customers want more value from their data



Growing
Exponentially



From new
sources



Increasingly
diverse



Used by
many people



Analyzed by many
applications

Effective Data Management Strategies Reflect

It must be high quality data: complete, clean, contextualized, and normalized

It must be wisely stored, secured, organized

It must be powered sufficiently

It must be tied to core business initiatives and outcomes

It must be checked for data accuracy

Can be centralized or federated to enable functions across organization

Modern Data Architecture Core Characteristics



Durability and Availability

Replicate data across regions and availability zones to ensure your data is available globally with 99.999999999% durability and 99.99%+ availability



Security

Protect data with advanced encryption, fine grain access control (IAM), encryption key management (KMS), logging (CloudWatch / CloudTrail), and sensitive data discovery (Macie)



Object Level Controls

Fine-grain, object level control allows tagging of valuable data for replication and tiered storage, saving money, and increasing performance



Flexibility

Storing all data in one data platform avoids data silos and the cost of moving data around



Operation Data Store

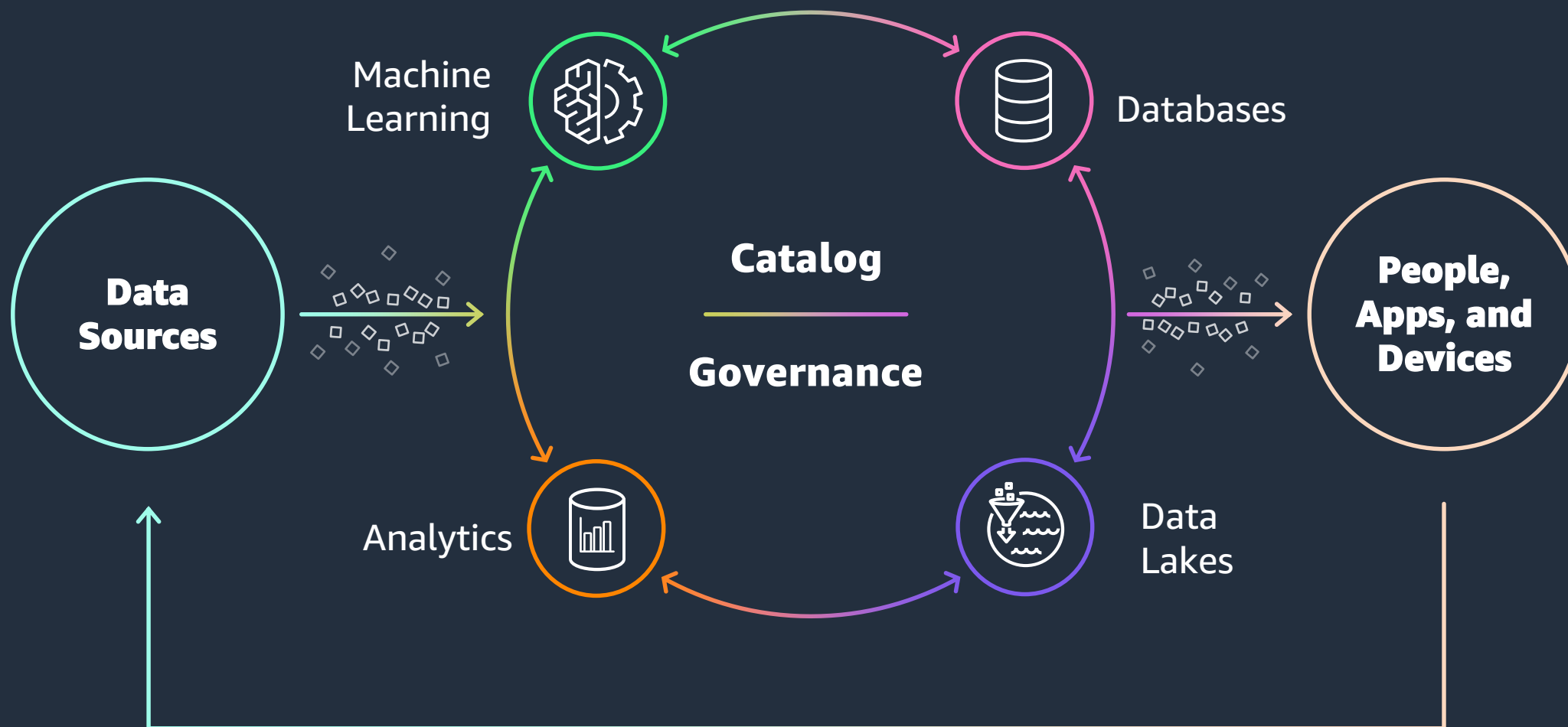
Creating an Operational Data Store (ODS) to access structured frequently used data for real time insights with off the shelf API



ML/AI

Once your data is in an AWS data platform, automate data transport and security functions, and pull business insights faster and more efficiently with ML/AI

Modern data strategy in action



Put data **to work**



Make better
decisions



Improve
efficiencies



Respond
faster



Uncover
opportunities

AWS Data Pillars Applied



Scalable data
lakes



Purpose-built
for performance
and cost



Serverless and
easy to use



Unified data
access, security,
and governance



Built-in machine
learning



Scalable data lakes

Broadest portfolio
of analytics tools

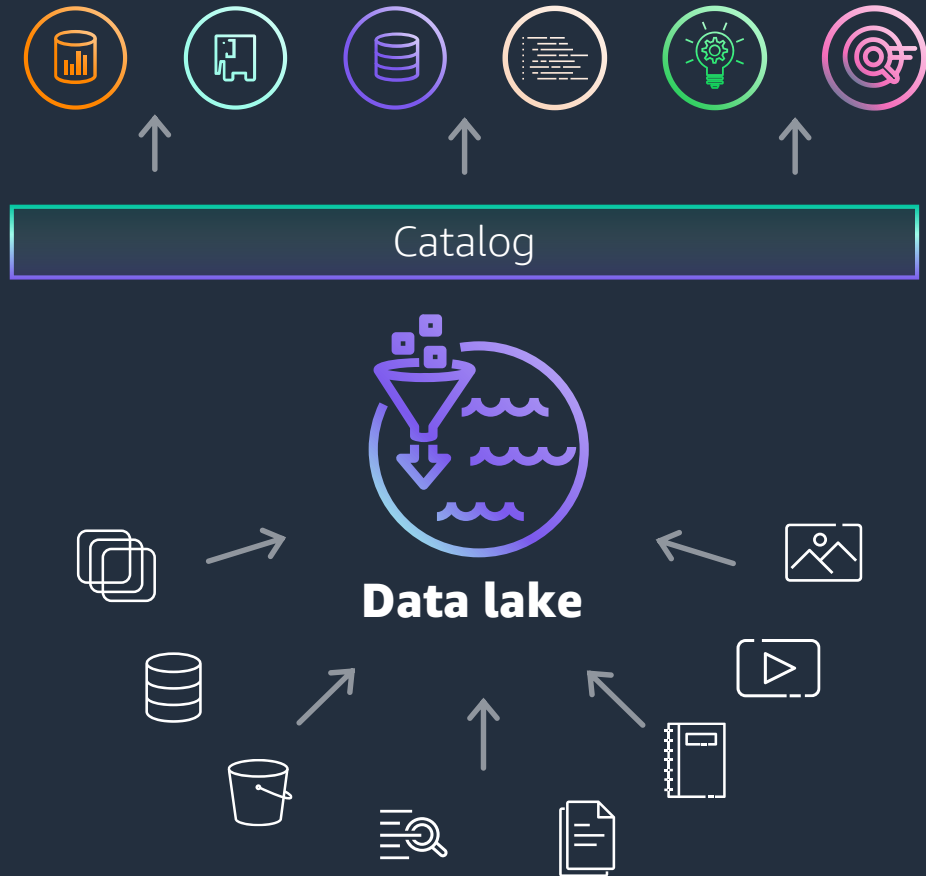


**Amazon
S3**

Unmatched durability,
availability, and scalability

Built to store and retrieve
any amount of data

The benefits of data lakes



Store all your data in open formats

Cost-effectively scale storage to exabytes

Decouple storage from compute

Choice of analytical and ML engines

Process data in place



Purpose-built for
**performance
and cost**



AMAZON
REDSHIFT

Data
warehousing



AMAZON
ATHENA

Interactive
query with SQL



AMAZON
EMR

Big data
processing



AMAZON
OPENSEARCH
SERVICE

Log and search
analytics



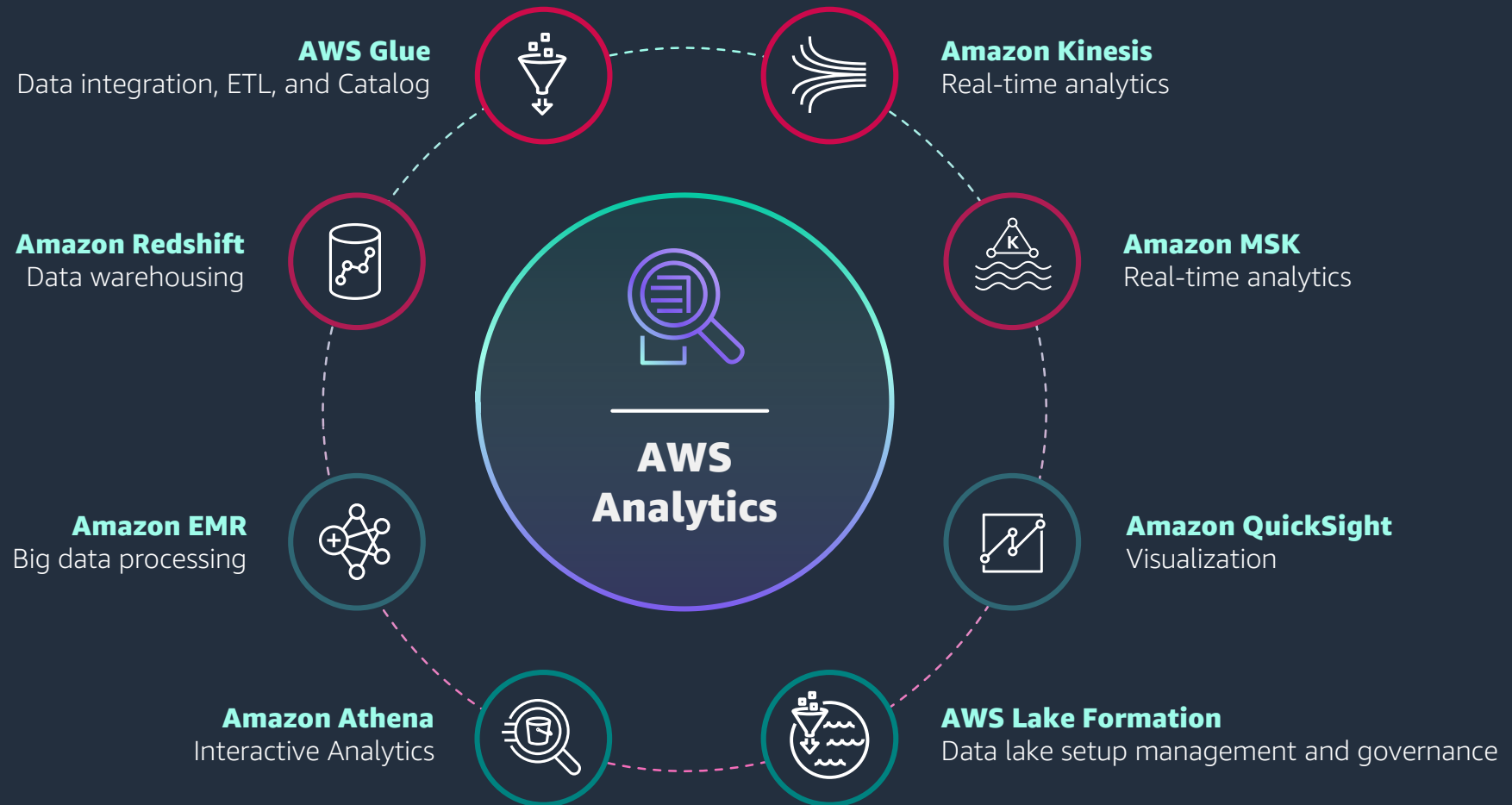
AMAZON
KINESIS & MSK

Real-time
analytics



Serverless
and easy to use

Serverless options for data analytics in the cloud





Unified data access, security, and governance

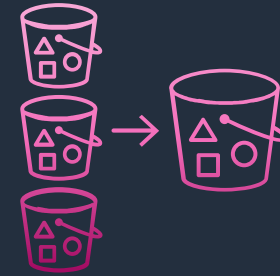
Challenges of building and securing modern data lakes



Support updates
and deletes



Row-level
Fine-grained
Secure sharing



Automatic storage
optimization

Break down data silos



Extract,
transform, load



Visual data
preparation



Data
replication



Data warehouse
to/from data lake



Federated
query



ML **Integration**

AWS brings ML closer to data



Databases

+



Data warehouses
+ data lakes

+



Business
intelligence tools

AMAZON
AURORA ML



AMAZON
NEPTUNE ML



AMAZON
REDSHIFT ML



AMAZON
ATHENA ML



AMAZON
QUICKSIGHT
ML



Modern data strategy on AWS



AWS security, identity, and compliance solutions



Identity and access management

AWS Identity and Access Management (IAM)

AWS IAM Identity Center (successor to AWS SSO)

AWS Organizations

AWS Directory Service

Amazon Cognito

AWS Resource Access Manager



Detective controls

AWS Security Hub

Amazon GuardDuty

Amazon Inspector

Amazon CloudWatch

AWS Config

AWS CloudTrail

VPC Flow Logs

AWS IoT Device Defender



Infrastructure protection

AWS Firewall Manager

AWS Network Firewall

AWS Shield

AWS WAF

Amazon VPC

AWS PrivateLink

AWS Systems Manager



Data protection

Amazon Macie

AWS Key Management Service (KMS)

AWS CloudHSM

AWS Certificate Manager

AWS Secrets Manager

AWS VPN

Server-Side Encryption



Incident response

Amazon Detective

Amazon EventBridge

AWS Backup

AWS Security Hub

AWS Elastic Disaster Recovery



Compliance

AWS Artifact

AWS Audit Manager

Resilience

The ability for workloads to respond to and quickly recover from failures.

The mental model

High Availability

Resistance to common failures through design and operational mechanisms



Core services, design goals to meet availability goals

Continuity of Operations

Returning to operations within specific targets for more rare but highly impactful failures



Backup & Recovery, Data Bunkering, Managed RPO/RTO

Continuous Resilience

← CI/CD, Code Refinement, Operational Testing, Observability/Monitoring →

A culture built around **resilience**

Our service design and deployment, operational model, and mechanisms help maintain resilience of the cloud



Service Ownership Model

Incentivizes continuous improvement of operations



Operational Readiness Reviews (ORR)

Ensures compliance to best practices prior to a service launch



Safe, Continuous Deployment

Minimizes impact on production caused by faulty deployments



Correction of Error (CoE) Processes

Helps teams understand root cause & prevents reoccurrence

Enabling your **resilience** in the cloud

We offer the most comprehensive set of best practices tooling, services, and guidance to enable your success.



Defining & Measuring Resilience Goals

One size does not fit all: Set goals at the workload level, not at the organization level



Identifying & Mitigating Risks

De-risk your architecture: Understand current resilience posture and fix high risk issues



Continuous Code Refinement

Stop issues before they start: Identify and resolve code issues before deployment



Continuous Integration/Continuous Deployment

Automate as much as possible: Remove opportunity for manual errors during deployment



Continuous Testing

Expect the unexpected: Simulate real-world failures to see how your teams and systems react



Continuous Observability

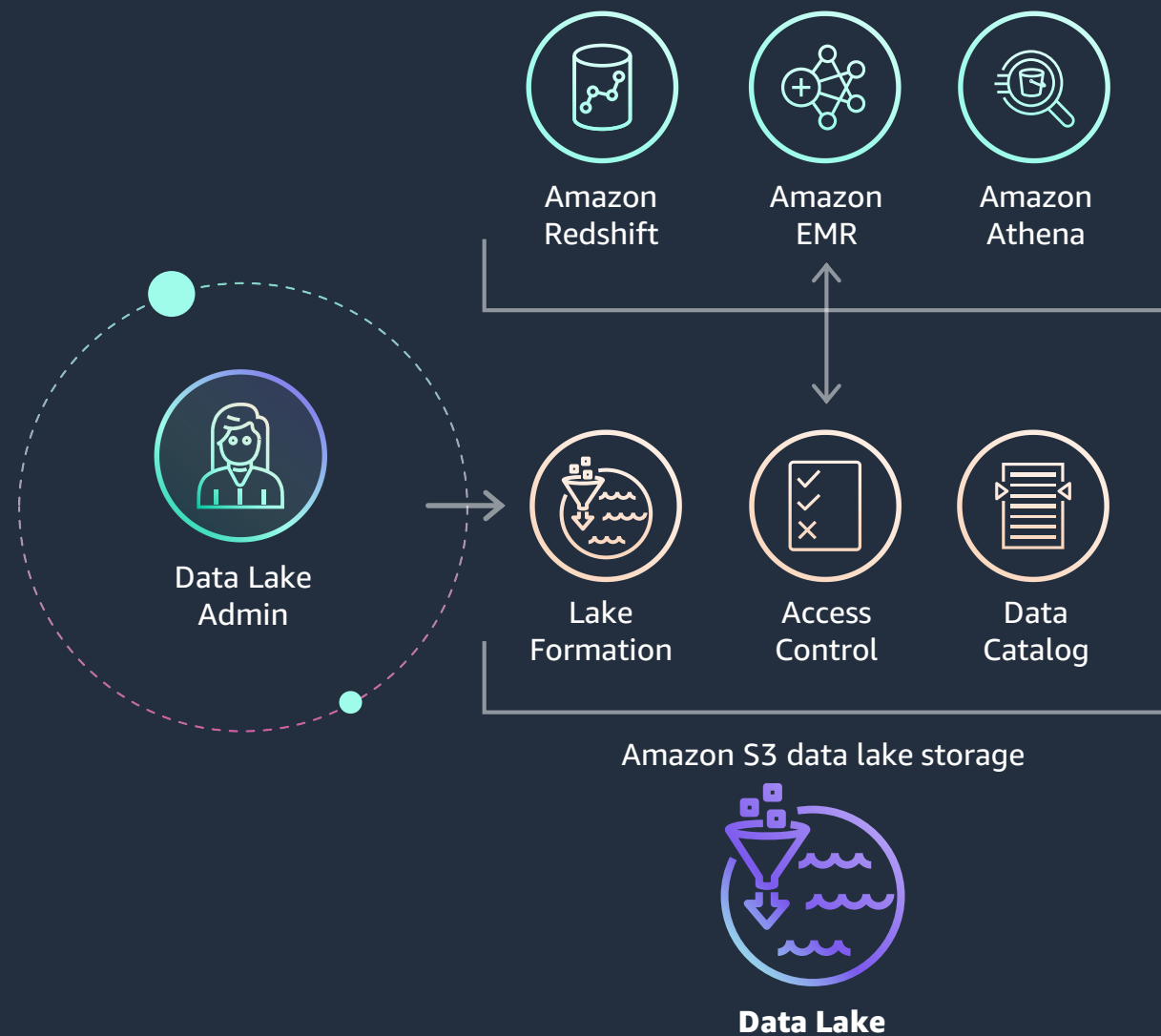
If you can't see it, you can't fix it: Monitor key business metrics using observability practices



Recovering Quickly

Failures are inevitable, but preparation helps: Proactively implement strategies like replication, redundancy, and backups

Simplify security management with **AWS Lake Formation**



AWS Lake Formation

BUILD SECURE DATA LAKES

Portfolio of integrated analytics tools



Amazon Athena



Amazon QuickSight



Amazon Redshift



AWS Glue



Amazon SageMaker



Amazon EMR

Lake Formation

Simplified ingest and cleaning



AWS Glue



Blueprints



ML Transform

Reliable and optimized data lakes



Acid Transactions



Storage Optimization



Catalog



Permissions

Amazon S3



Cost effective, durable data lake storage with global replication capabilities

AWS Glue: Key Capabilities

SERVERLESS DATA INTEGRATION SERVICE

Scalable Data Integration Engine



Built-in data transforms



Execution engine



Monitor

Centralized and Unified Data Governance



Glue data catalog



Glue crawlers



Lake formation

Connect and Ingest Data



Glue connectors



Glue connector marketplace



Variety of interfaces

User Productivity and Data Ops



Persona specific tools



Productivity tools



Data ops tools



Thank you!

Natasha Nicolai

niconat@amazon.com

Randy Staton

statonra@amazon.com

Purpose-built databases



Relational

Referential integrity, ACID transactions, schema-on-write



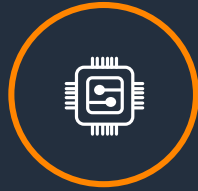
Key-value

High throughput, Low latency reads and writes, endless scale



Document

Store documents and quickly access querying on any attribute



In-memory

Query by key with microsecond latency



Graph

Quickly and easily create and navigate relationships between data



Time-series

Collect, store, and process data sequenced by time



Ledger

Complete, immutable, and verifiable history of all changes to application data



Wide Column

Scalable, highly available, and managed Apache Cassandra-compatible service

AWS Service(s)



Aurora RDS



DynamoDB



DocumentDB



ElastiCache



Neptune



Timestream



QLDB



Keyspaces
Managed Cassandra

Common Use Cases

Lift and shift, ERP, CRM, finance

Real-time bidding, shopping cart, social, product catalog, customer preferences

Content management, personalization, mobile

Leaderboards, real-time analytics, caching

Fraud detection, social networking, recommendation engine

IoT applications, event tracking

Systems of record, supply chain, health care, registrations, financial

Build low-latency applications, leverage open source, migrate Cassandra to the cloud