
AWS
re:Invent

ARC305

AWS Well-Architected Analytics Lens: X-ray view for your systems

Jongnam Lee

Analytics Lead SA, Well-Architected

Amazon Web Services

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Who should watch this session?

This session is for data architects who design analytics systems for their teams



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is the key takeaway?

The AWS Well-Architected Analytics Lens reduces the time it takes to design analytics workloads, because AWS gathered customer-proven best practices into the Analytics Lens for you.



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ARE YOU WELL-ARCHITECTED TO RUN ANALYTICS WORKLOAD?



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Well-Architected Framework



Security



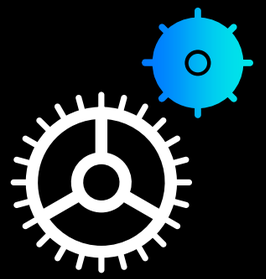
Reliability



Performance Efficiency



Cost Optimization



Operational Excellence

AWS Well-Architected Analytics Lens

- A set of questions and best practices for analytics workload review
- Clarifies the AWS Well-Architected Framework
- *Gives you an X-ray view of your analytics workload*



Security



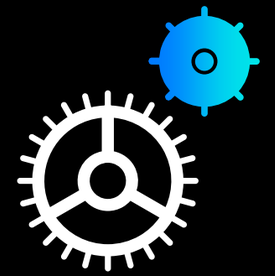
Reliability



Performance Efficiency



Cost Optimization



Operational Excellence

ANALYTICS SYSTEM IS A DATA HUB FOR OTHER SYSTEMS

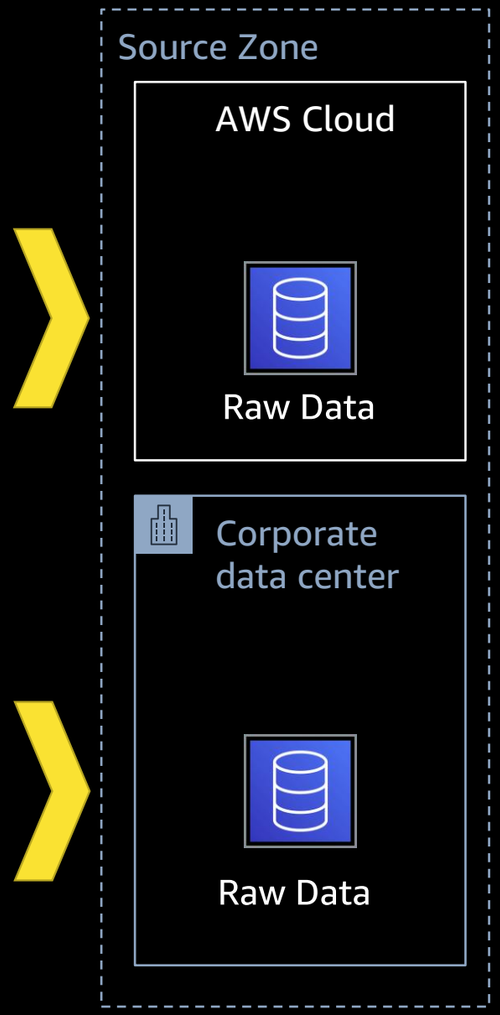


© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

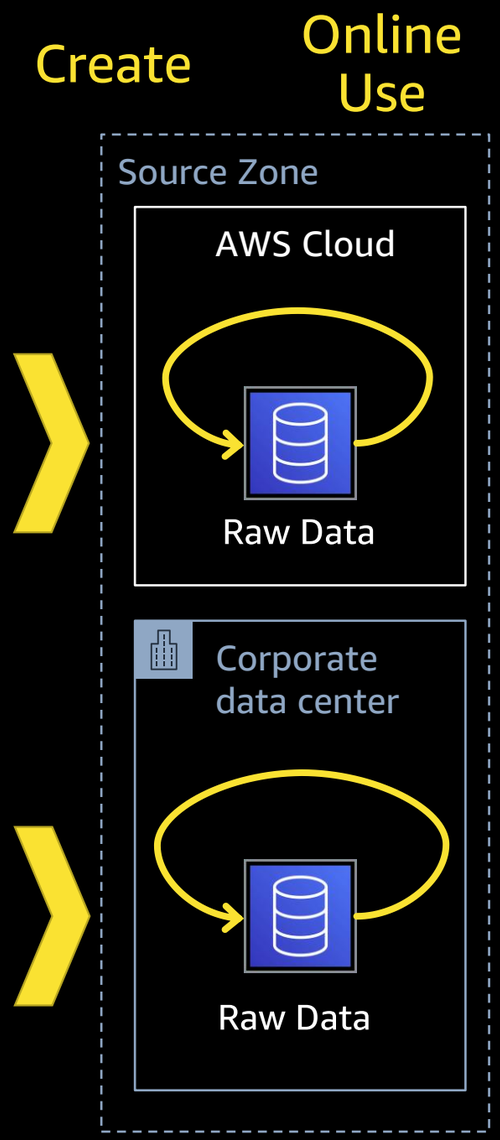
The Data Lifecycle & Analytics Workload

The Data Lifecycle & Analytics Workload

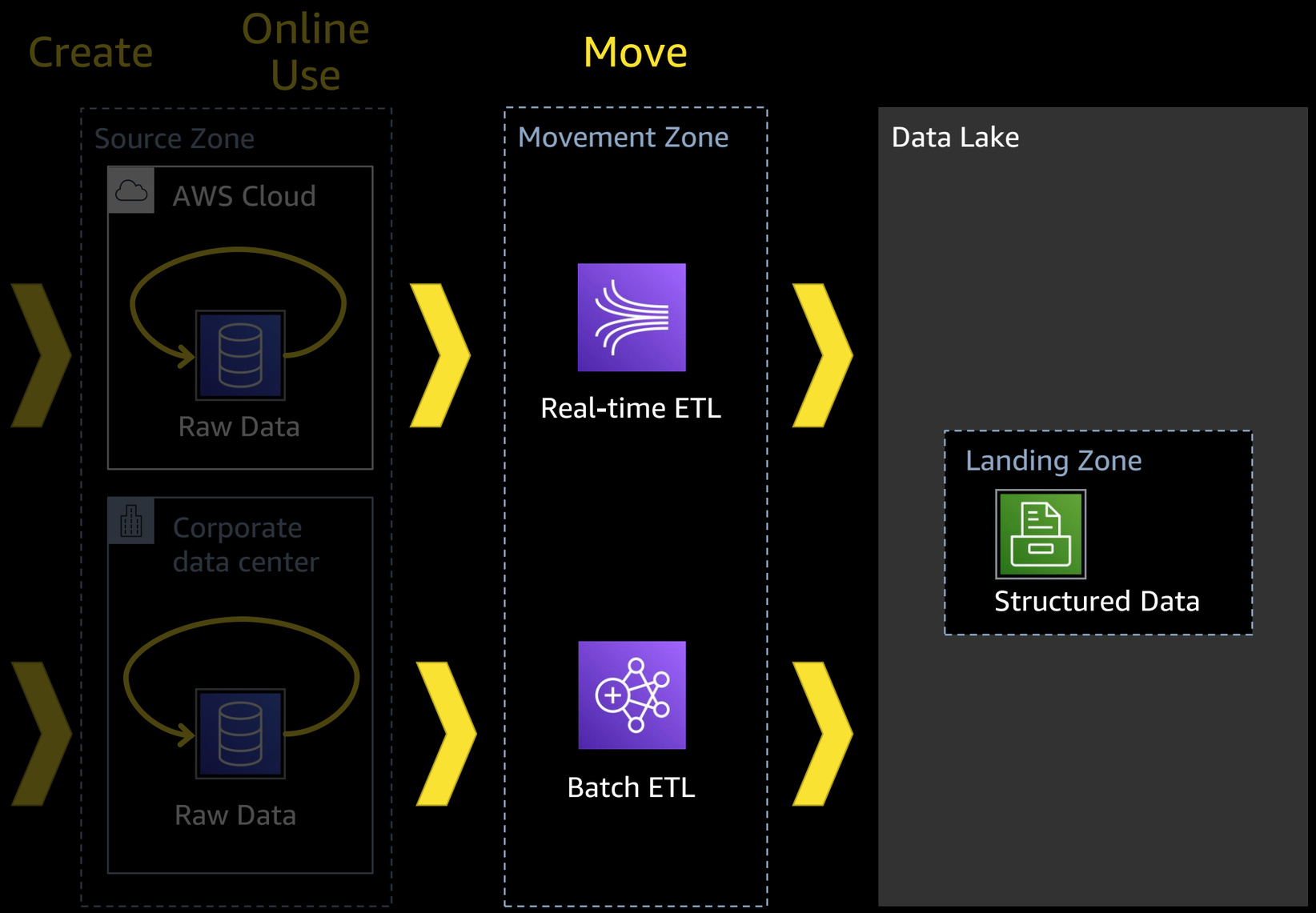
Create



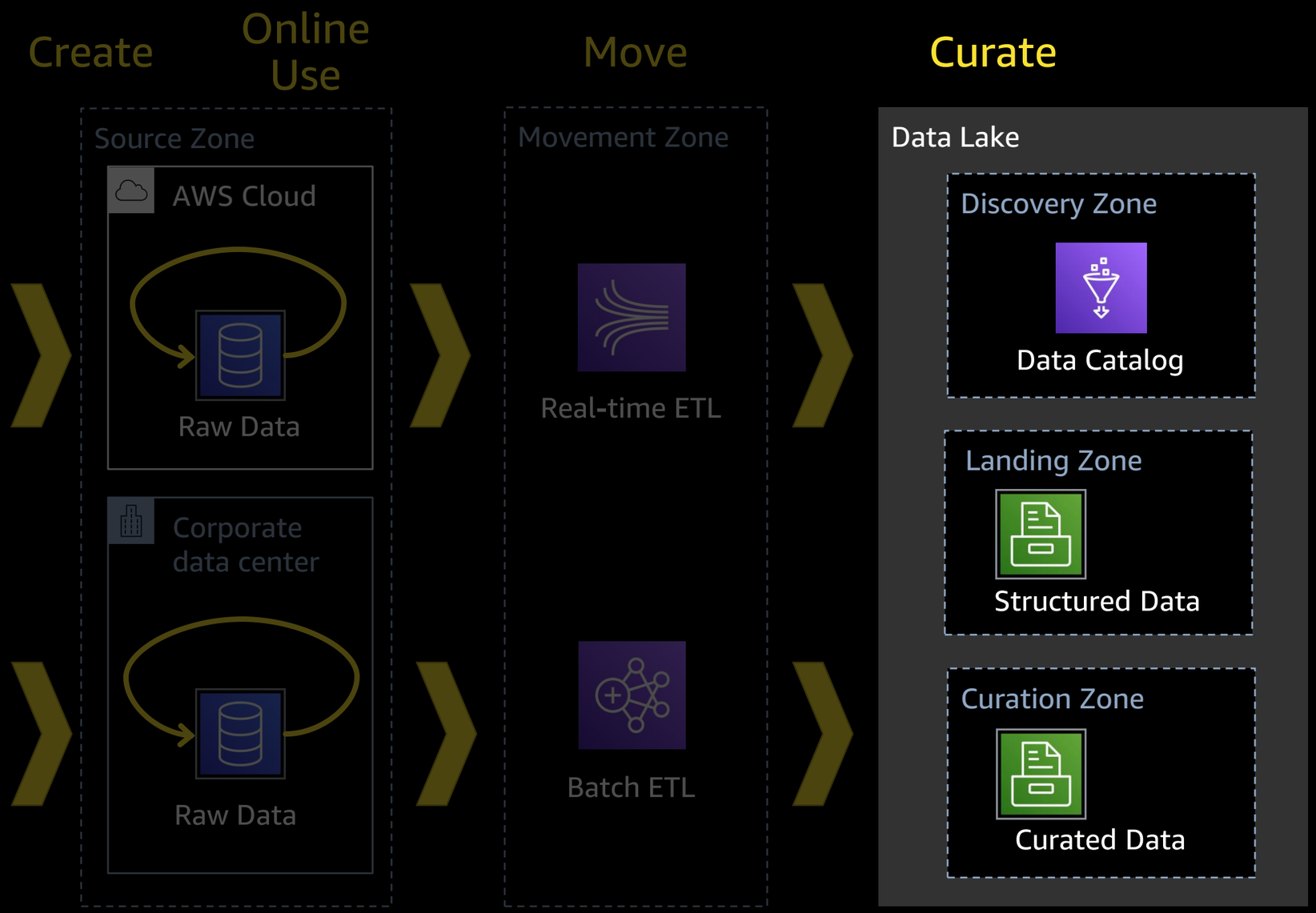
The Data Lifecycle & Analytics Workload



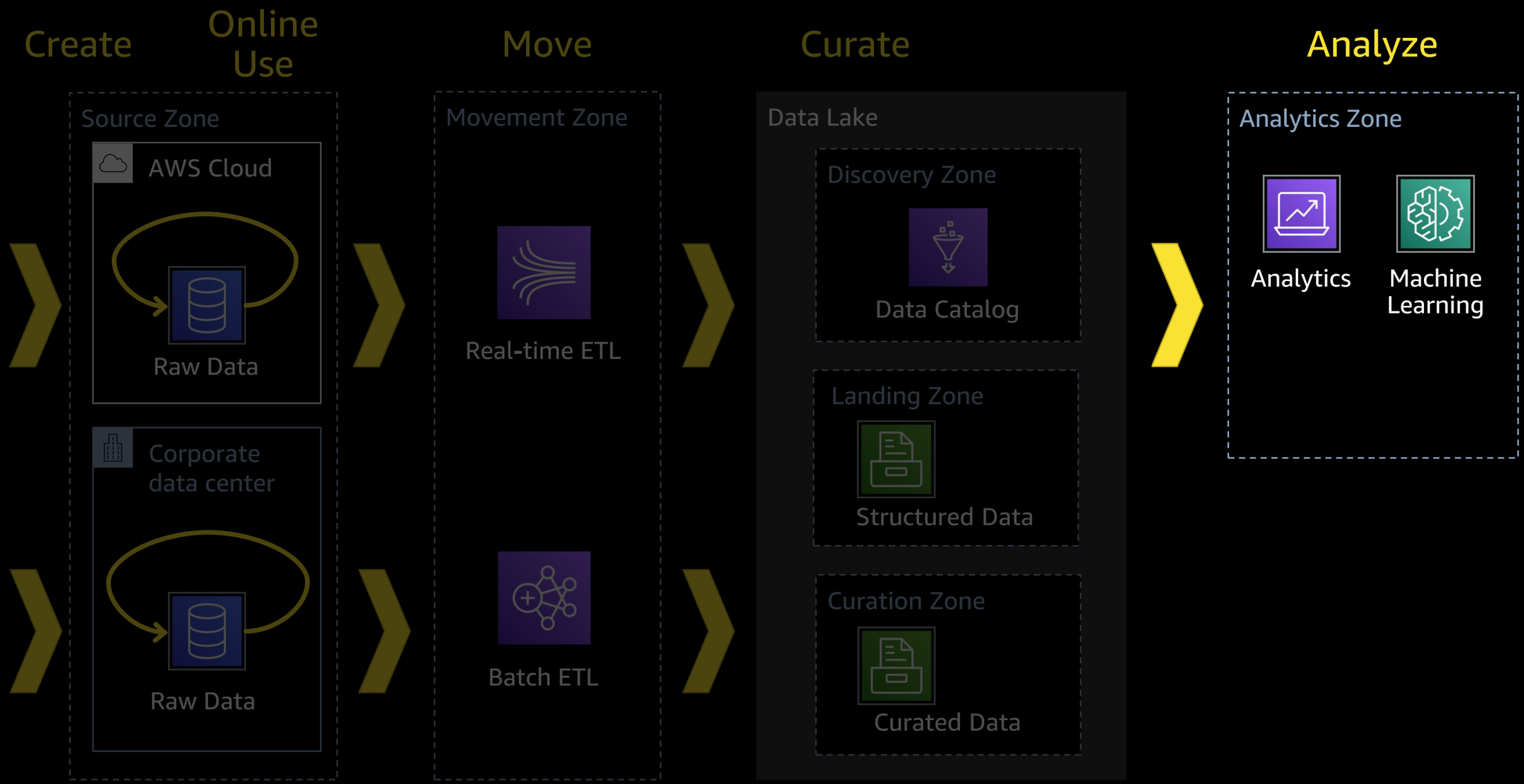
The Data Lifecycle & Analytics Workload



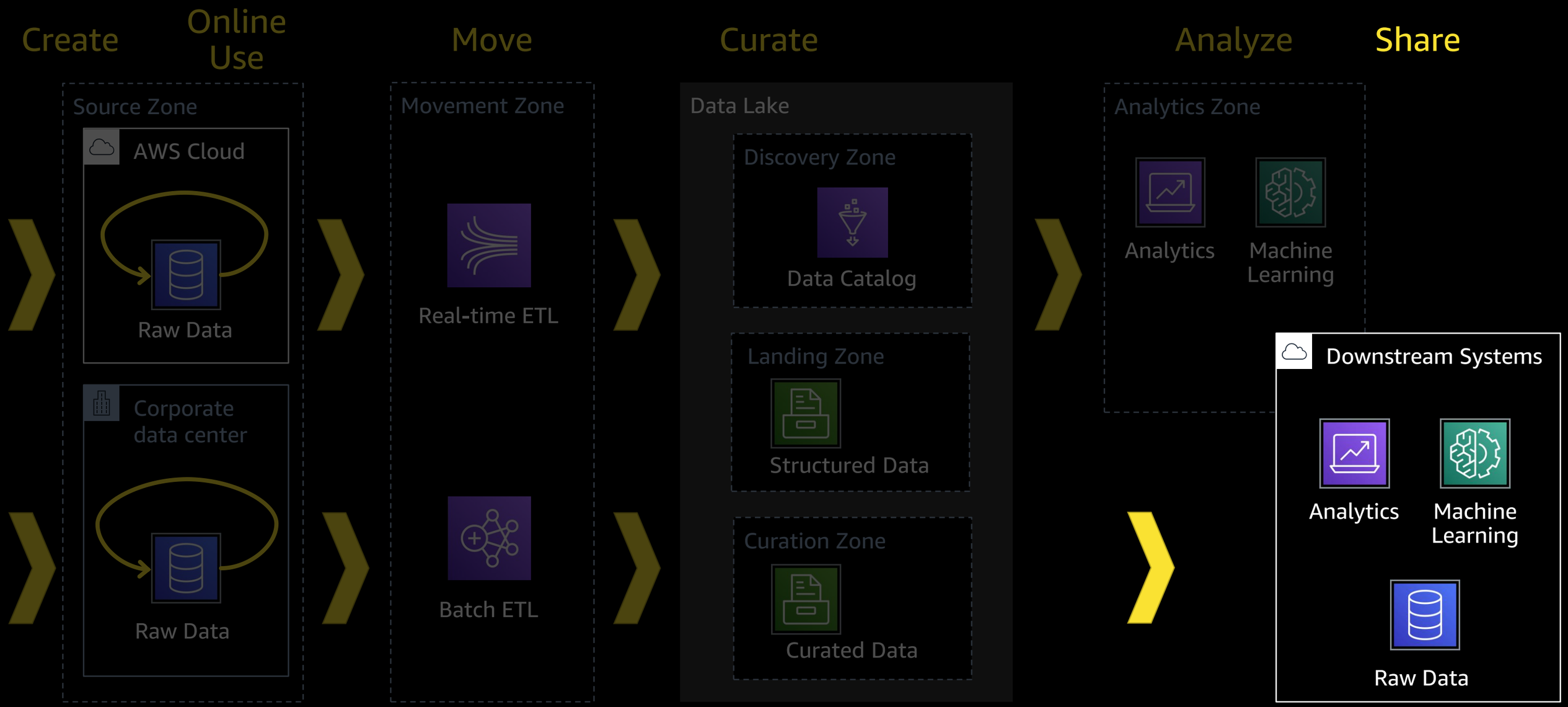
The Data Lifecycle & Analytics Workload



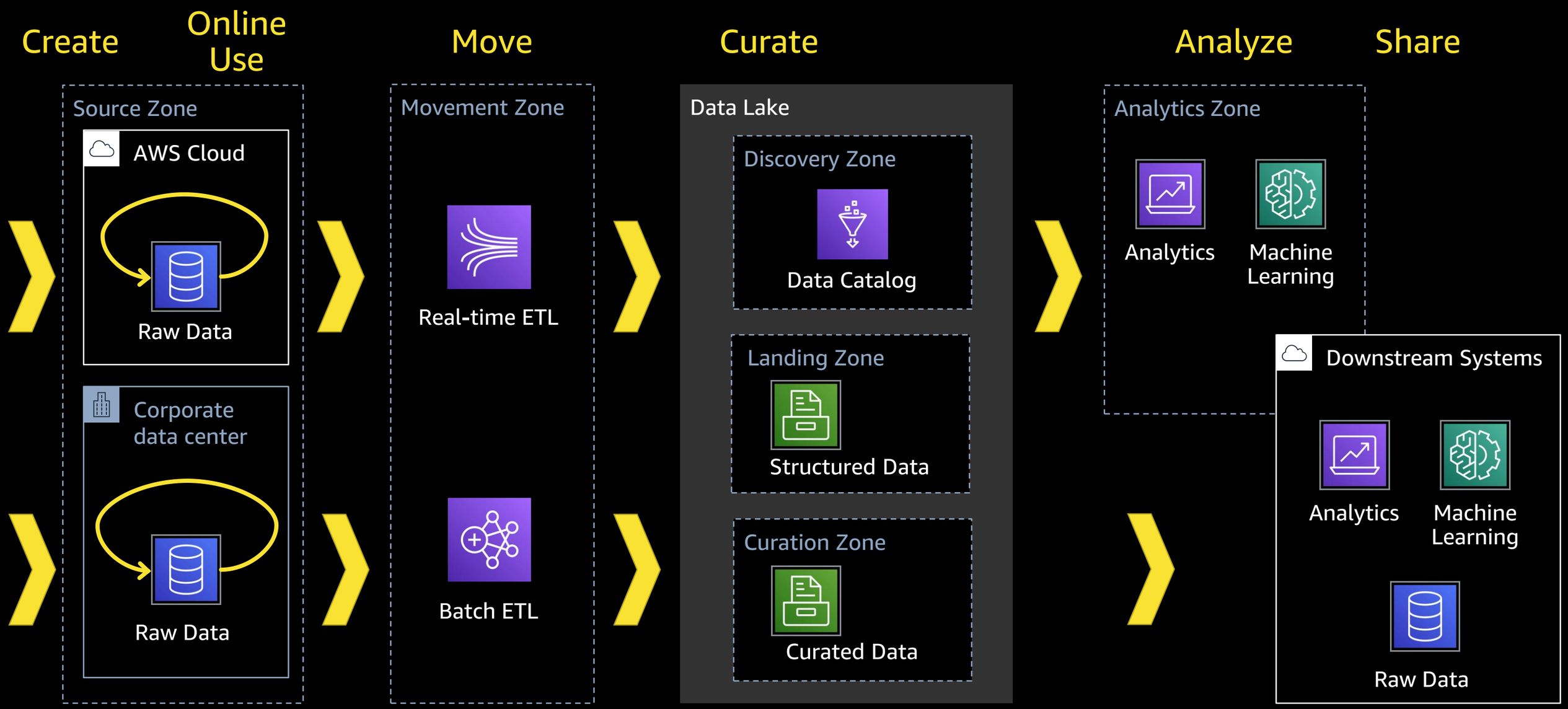
The Data Lifecycle & Analytics Workload



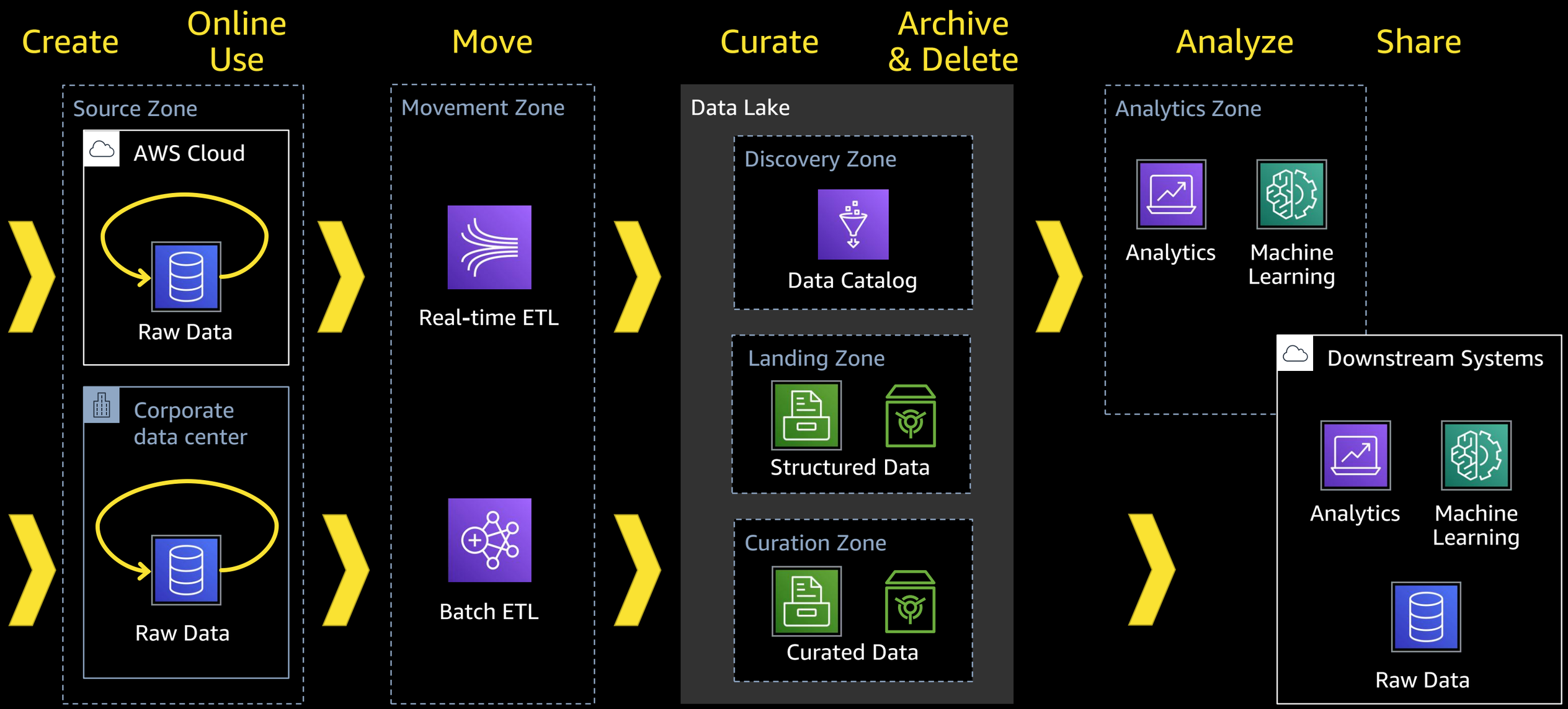
The Data Lifecycle & Analytics Workload



The Data Lifecycle & Analytics Workload



The Data Lifecycle & Analytics Workload



Analytics Workload Architecture

WHERE SHOULD WE BEGIN?

Source Systems


Amazon DynamoDB


Amazon RDS


Amazon Redshift


Amazon EMR


Amazon S3

Data Movement


Amazon Kinesis


Amazon MSK


AWS Lambda


AWS Database Migration Service


AWS Batch


Amazon EMR


AWS Glue


AWS Snowball

Data Lake


AWS Glue


AWS Lake Formation


Amazon S3

Analytic Processing


Amazon Athena


Amazon Forecast


Amazon EMR


Amazon Redshift


Amazon QuickSight


Amazon SageMaker


Amazon Kinesis Data Analytics


Amazon Elasticsearch Service

Security, Identity & Compliance


AWS IAM


AWS Directory Service


AWS KMS


AWS Secrets Manager


Amazon Macie

Management & Governance


Amazon CloudWatch


AWS CloudTrail


AWS Control Tower


AWS Cost Explorer


AWS CloudFormation


AWS Personal Health Dashboard


AWS Trusted Advisor


AWS Budgets

LET'S DESIGN AN ANALYTICS WORKLOAD FOR A TEAM



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Control/Metadata Flow

Data Flow

Source System Architect

Analytics Architect

Data Engineers

Business Analysts

Data Scientists

Source Zone



Amazon S3 Raw Data

Corporate data center



On-prem Database



Security

SECURITY: HOW TO PROTECT WORKLOADS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

1. Retain Least Privilege for Source Access

Ask yourself:

How do you manage permissions for integrating with data source systems?

You should

- *Understand permission requirements of the source systems*
- *Request least privilege to the source systems*
- *Establish emergency access process to the source and analytics systems*
- *Periodically review the privileges (Are they being utilized?)*

Control/Metadata Flow

Data Flow

Source System Architect

Analytics Architect

Data Engineers

Business Analysts

Data Scientists

Source Zone

Amazon S3 Raw Data

Corporate data center

On-prem Database

Control/Metadata Flow

Data Flow

Source System Architect

Analytics Architect

Data Engineers

Business Analysts

Data Scientists

AWS IAM Policy for Source System's S3 Bucket

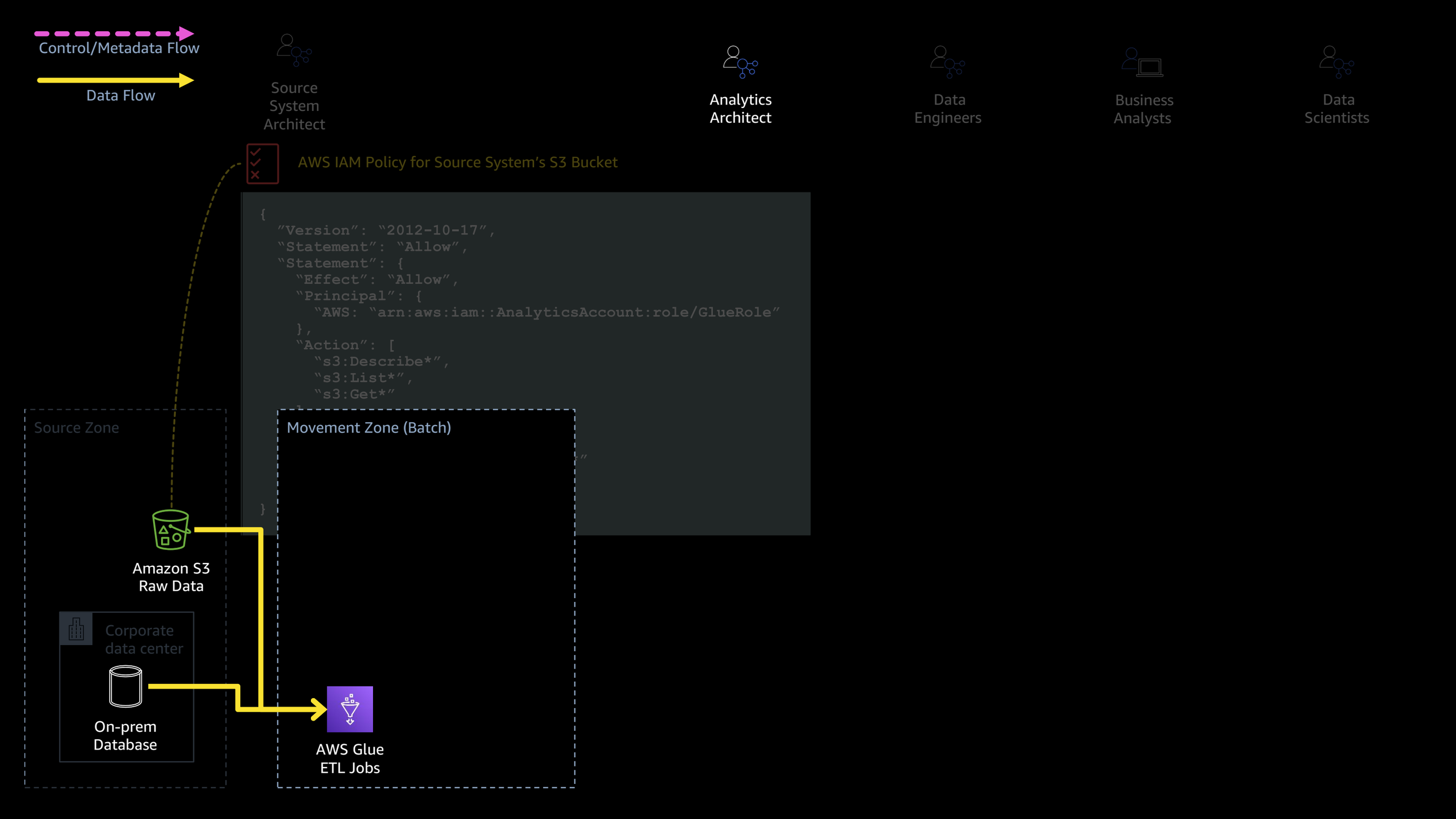
```
{
  "Version": "2012-10-17",
  "Statement": "Allow",
  "Statement": {
    "Effect": "Allow",
     "Principal": {
       "AWS: "arn:aws:iam::DataLakeAccount:role/GlueRole"
     },
    "Action": [
      "s3:Describe*",
      "s3:List*",
      "s3:Get*"
    ],
    "Resource": [
      "arn:aws:s3:::source-bucket",
      "arn:aws:s3:::source-bucket/*"
    ]
  }
}
```

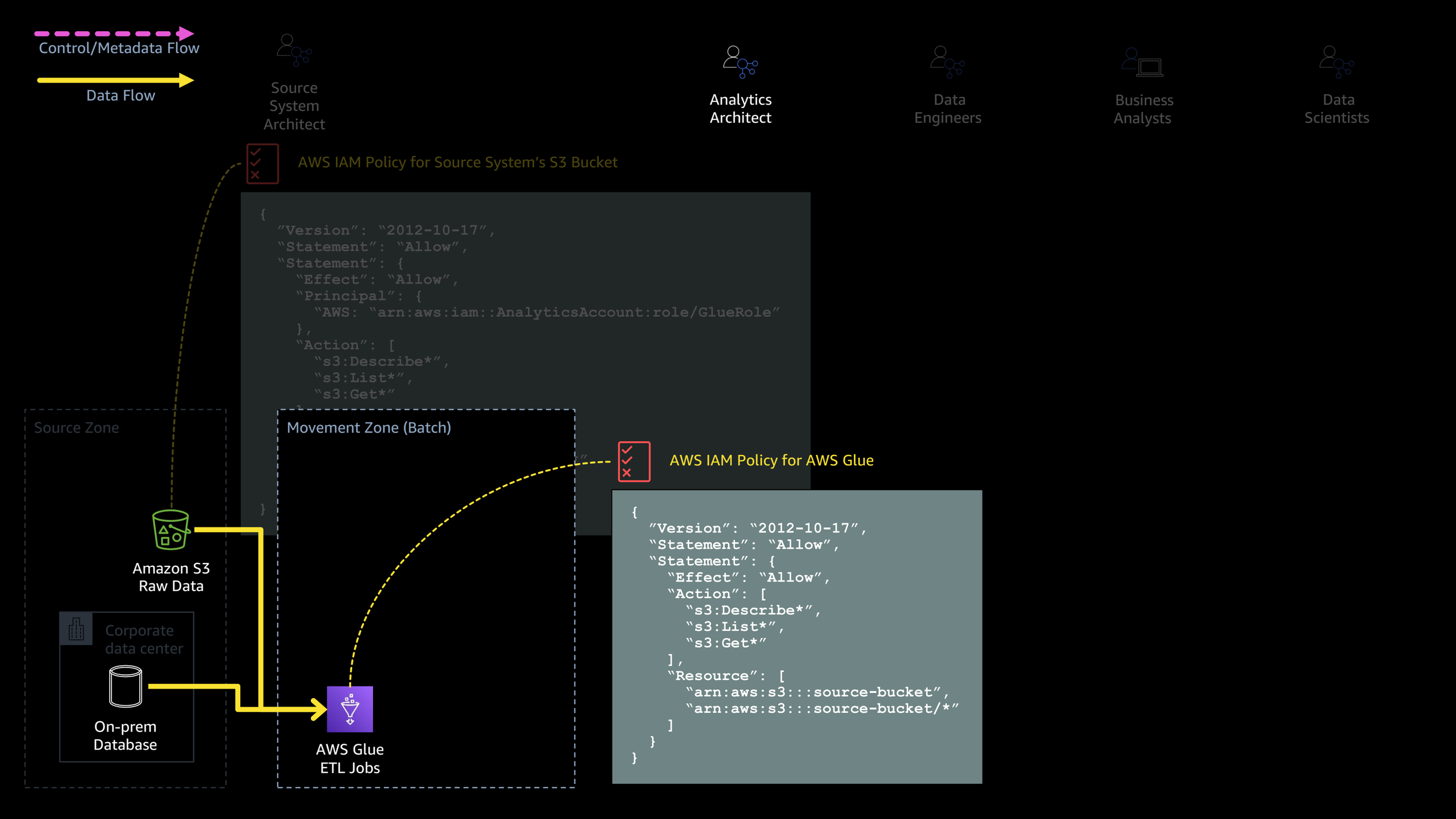
Source Zone

Amazon S3 Raw Data

Corporate data center

On-prem Database


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
AWS IAM Policy for Source System's S3 Bucket
{
"Version": "2012-10-17",
"Statement": "Allow",
"Statement": {
"Effect": "Allow",
"Principal": {
"AWS": "arn:aws:iam::AnalyticsAccount:role/GlueRole"
},
"Action": [
"s3:Describe*",
"s3:List*",
"s3:Get*"
}
Source Zone
Movement Zone (Batch)
Amazon S3 Raw Data
Corporate data center
On-prem Database
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
AWS IAM Policy for Source System's S3 Bucket
{
"Version": "2012-10-17",
"Statement": "Allow",
"Statement": {
"Effect": "Allow",
"Principal": {
"AWS": "arn:aws:iam::AnalyticsAccount:role/GlueRole"
},
"Action": [
"s3:Describe*",
"s3:List*",
"s3:Get*"
Source Zone
Movement Zone (Batch)
AWS IAM Policy for AWS Glue
{
"Version": "2012-10-17",
"Statement": "Allow",
"Statement": {
"Effect": "Allow",
"Action": [
"s3:Describe*",
"s3:List*",
"s3:Get*"
],
"Resource": [
"arn:aws:s3:::source-bucket",
"arn:aws:s3:::source-bucket/*"
]
}
}
Amazon S3 Raw Data
Corporate data center
On-prem Database
AWS Glue ETL Jobs

2. Classify Data

Ask yourself:

How do you classify data in your analytics workload?

You should

- *Know the source data owners*
- *Understand data classifications in the source systems*
- *Implement data protection & lifecycle controls in analytics workload*
- *Demand downstream systems to honor the data classifications*

Control/Metadata Flow

Data Flow

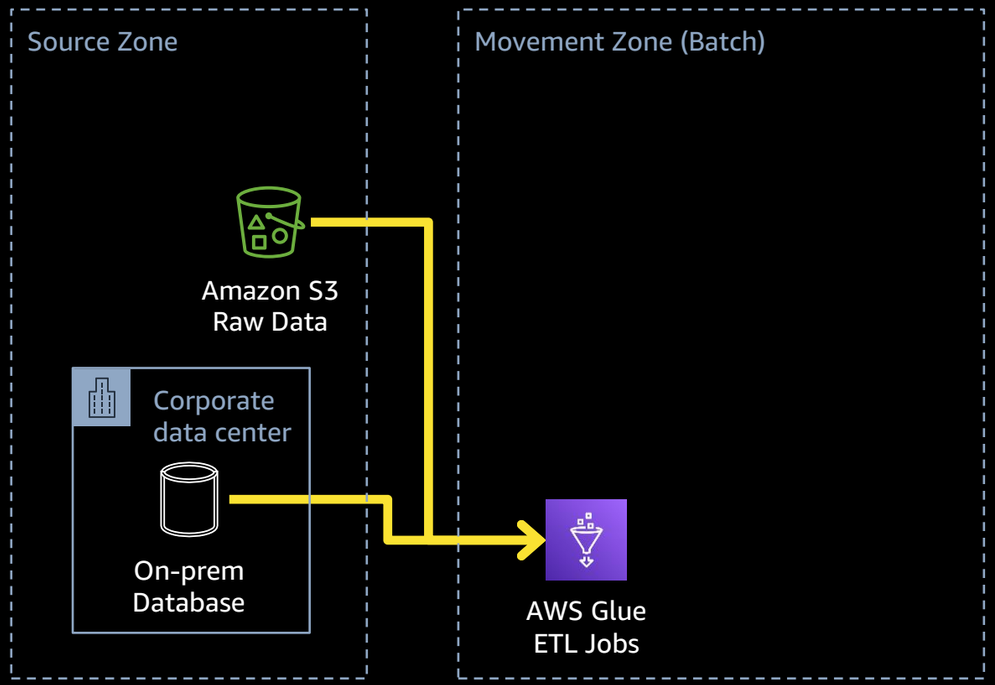
Source System Architect

Analytics Architect

Data Engineers

Business Analysts

Data Scientists



Control/Metadata Flow

Data Flow

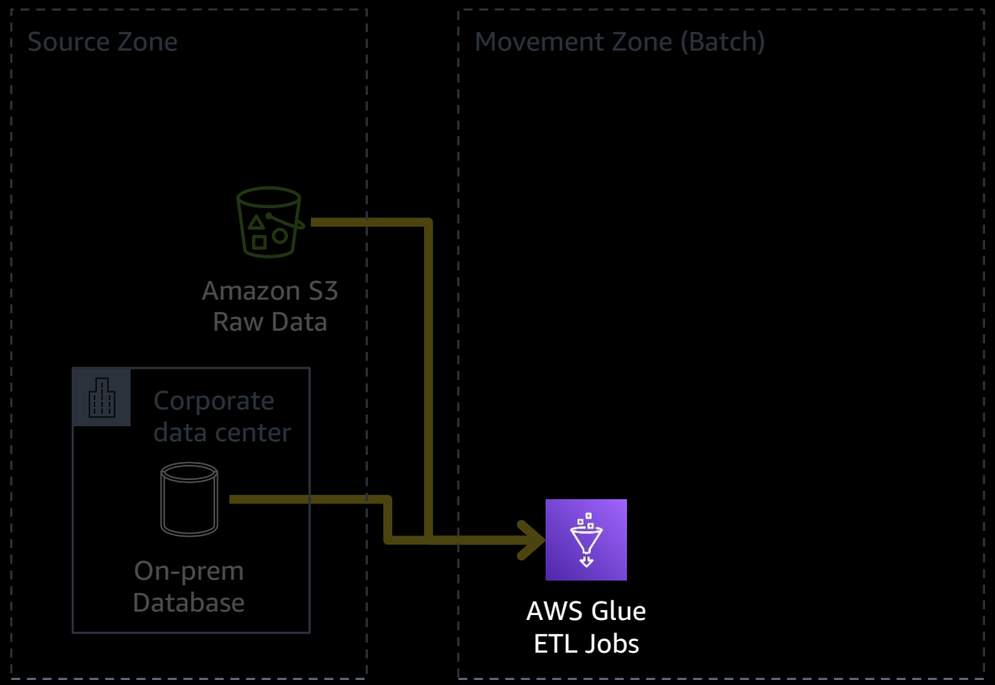
Source System Architect

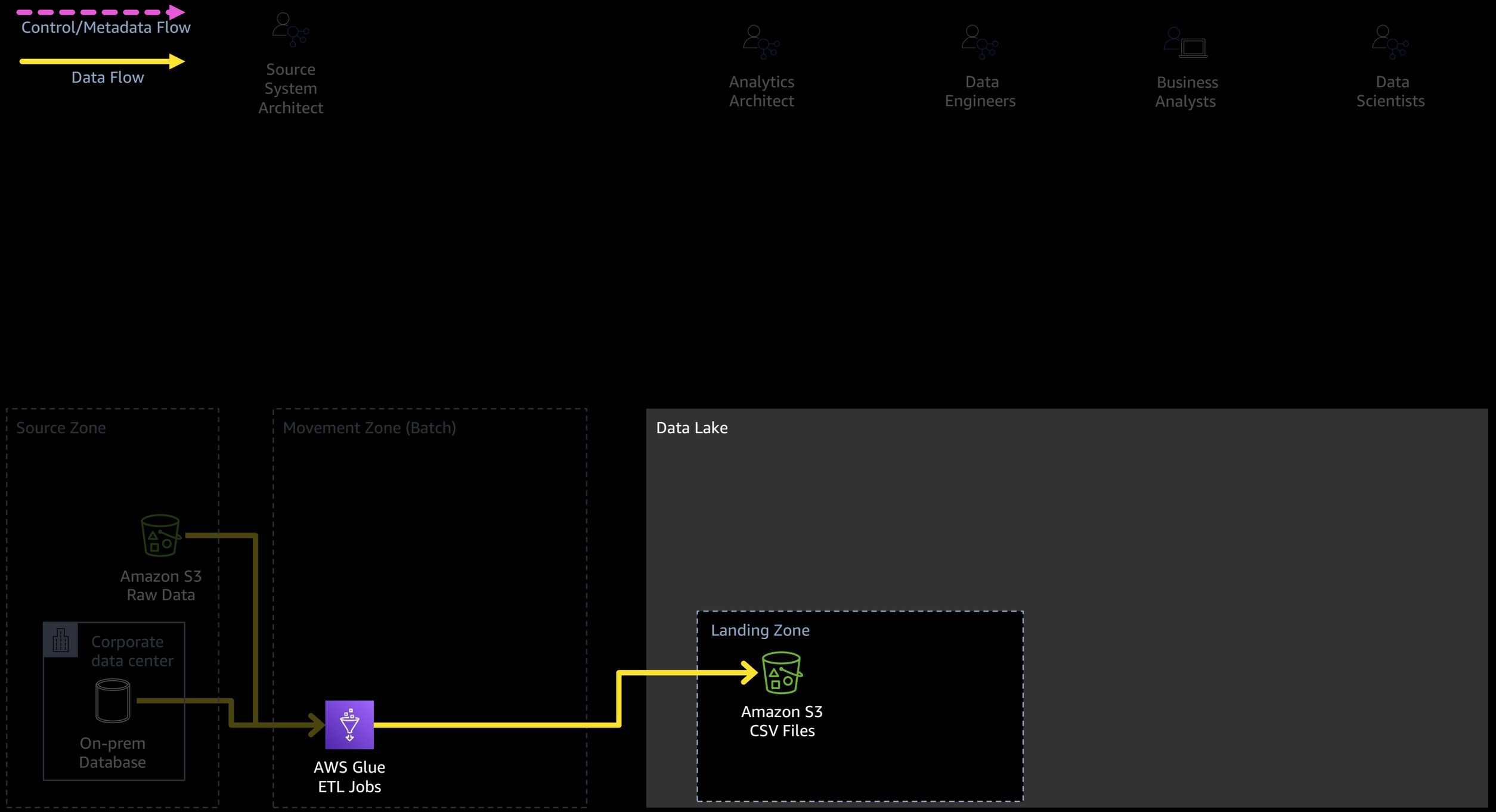
Analytics Architect

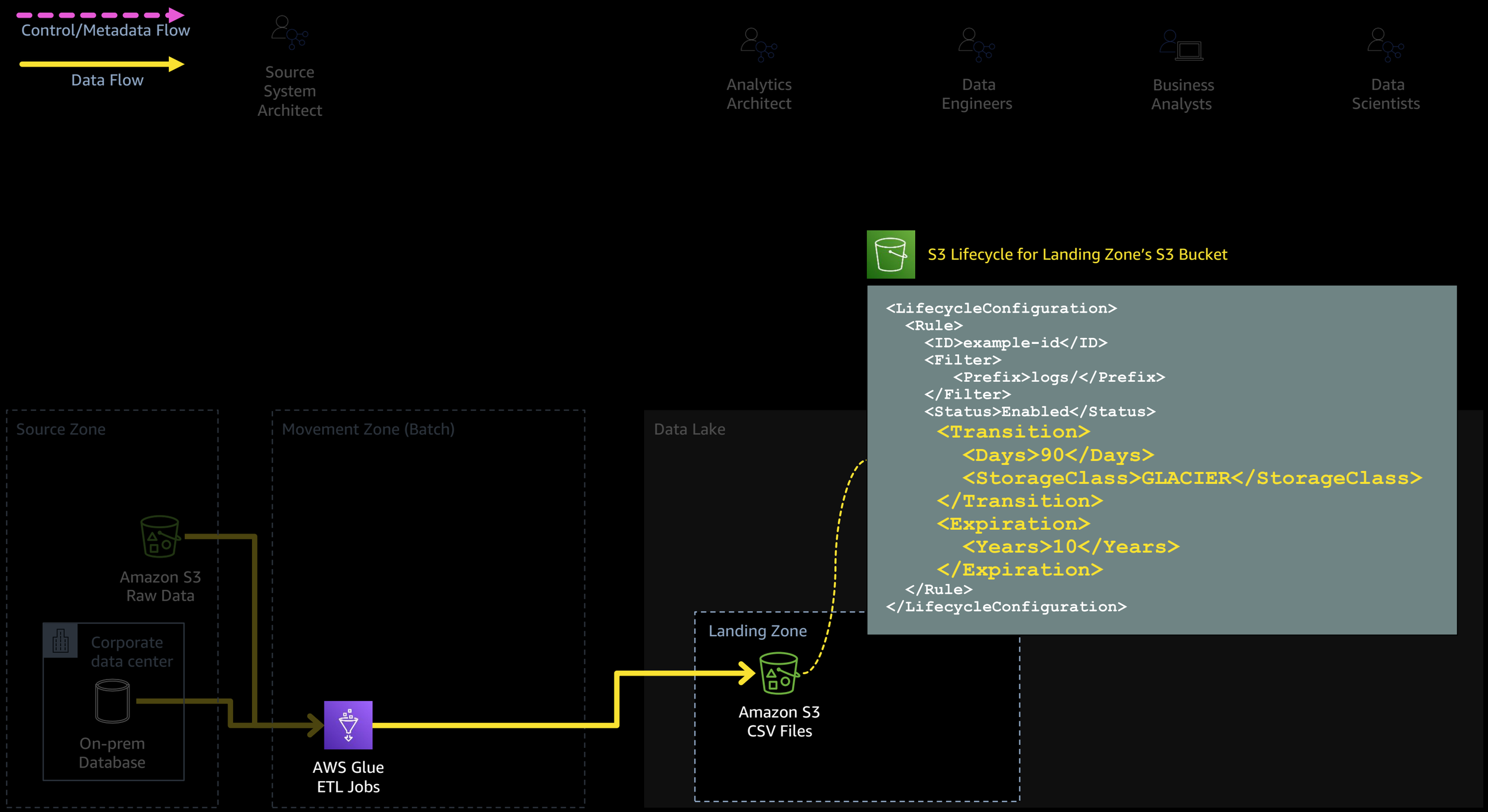
Data Engineers

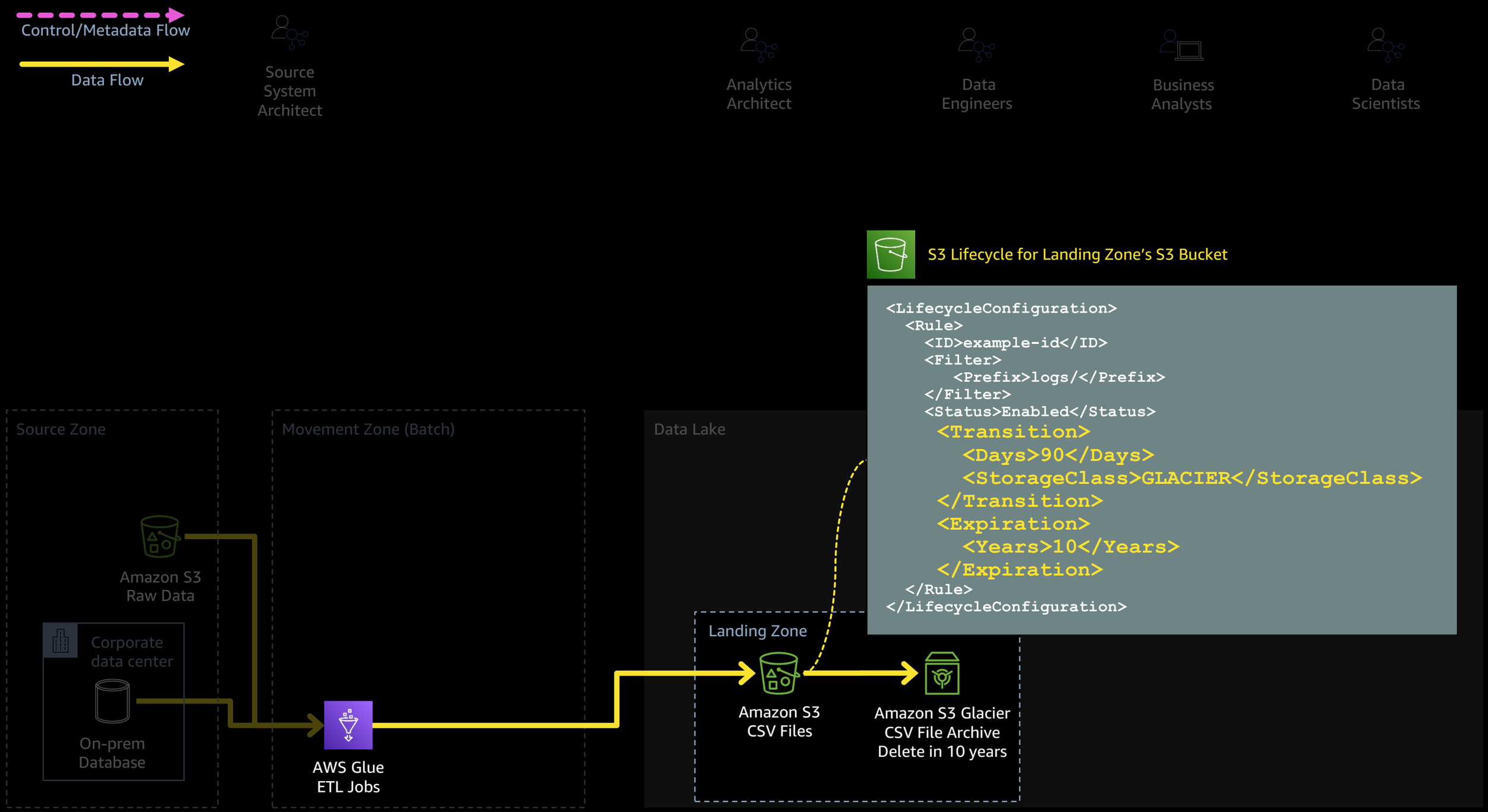
Business Analysts

Data Scientists




Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Source Zone
Movement Zone (Batch)
Data Lake
Amazon S3 Raw Data
Corporate data center
On-prem Database
AWS Glue ETL Jobs
Landing Zone
Amazon S3 CSV Files


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
S3 Lifecycle for Landing Zone's S3 Bucket
<LifecycleConfiguration>
<Rule>
<ID>example-id</ID>
<Filter>
<Prefix>logs/</Prefix>
</Filter>
<Status>Enabled</Status>
<Transition>
<Days>90</Days>
<StorageClass>GLACIER</StorageClass>
</Transition>
<Expiration>
<Years>10</Years>
</Expiration>
</Rule>
</LifecycleConfiguration>
Source Zone
Movement Zone (Batch)
Data Lake
Landing Zone
Amazon S3 Raw Data
Corporate data center
On-prem Database
AWS Glue ETL Jobs
Amazon S3 CSV Files


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
S3 Lifecycle for Landing Zone's S3 Bucket
<LifecycleConfiguration>
<Rule>
<ID>example-id</ID>
<Filter>
<Prefix>logs/</Prefix>
</Filter>
<Status>Enabled</Status>
<Transition>
<Days>90</Days>
<StorageClass>GLACIER</StorageClass>
</Transition>
<Expiration>
<Years>10</Years>
</Expiration>
</Rule>
</LifecycleConfiguration>
Source Zone
Movement Zone (Batch)
Data Lake
Amazon S3 Raw Data
Corporate data center
On-prem Database
AWS Glue ETL Jobs
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years

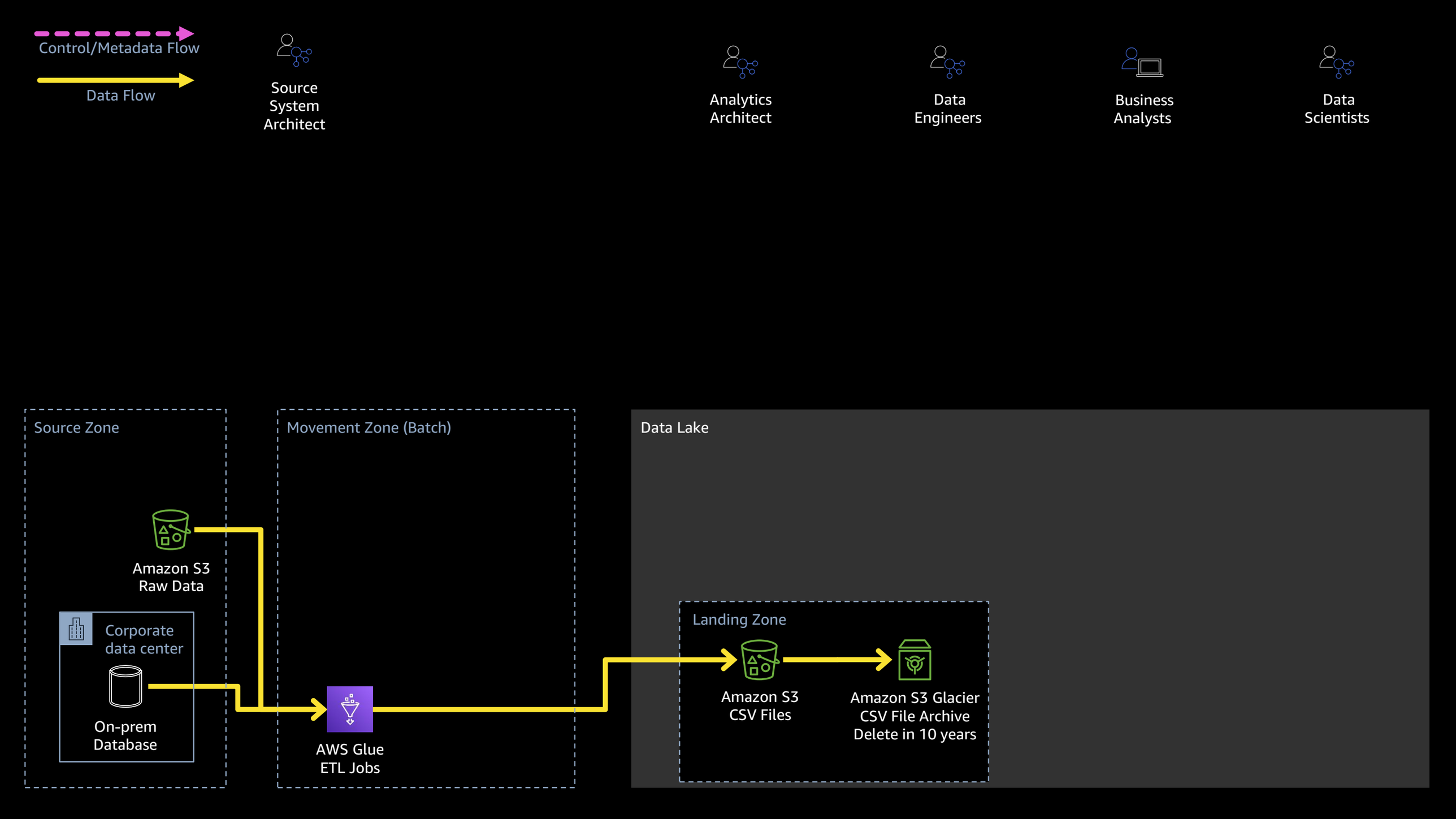
3. Grant Least Privilege to Analytics Users

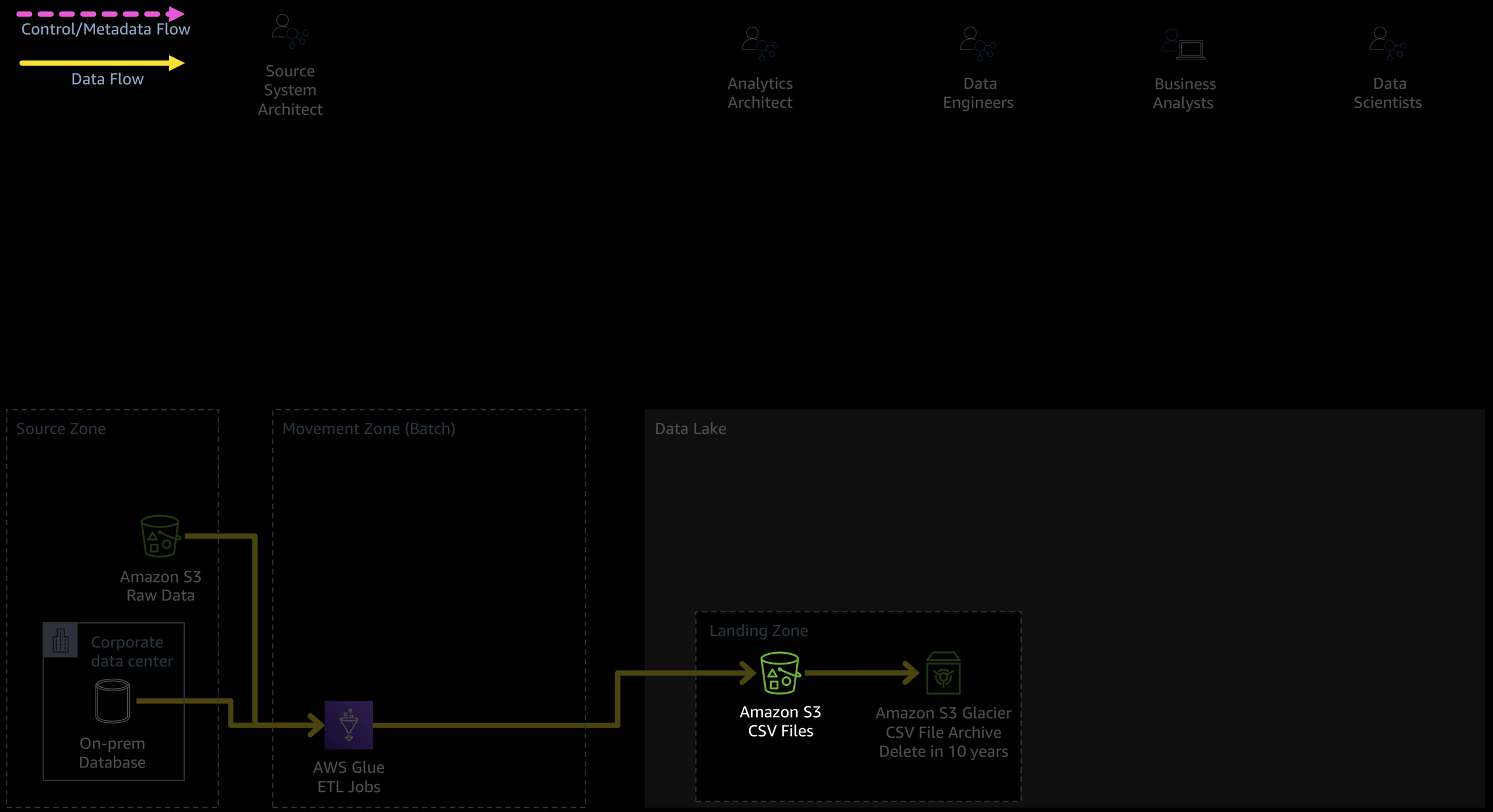
Ask yourself:

How do you manage permissions to the analytics workloads for people and machines?

You should

- *Implement permission controls for each attack surface*
- *Implement AWS account-level permission guardrails*


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Source Zone
Movement Zone (Batch)
Data Lake
Amazon S3 Raw Data
Corporate data center
On-prem Database
AWS Glue ETL Jobs
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Source Zone
Movement Zone (Batch)
Data Lake
Amazon S3 Raw Data
Corporate data center
On-prem Database
AWS Glue ETL Jobs
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years

Control/Metadata Flow

Data Flow

Source System Architect

Analytics Architect

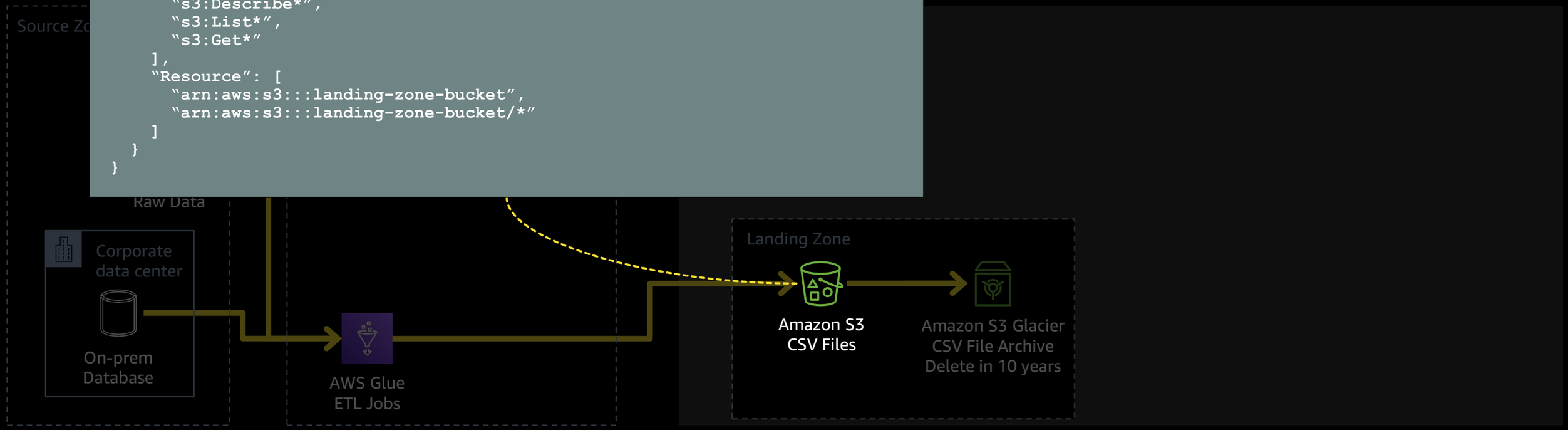
Data Engineers

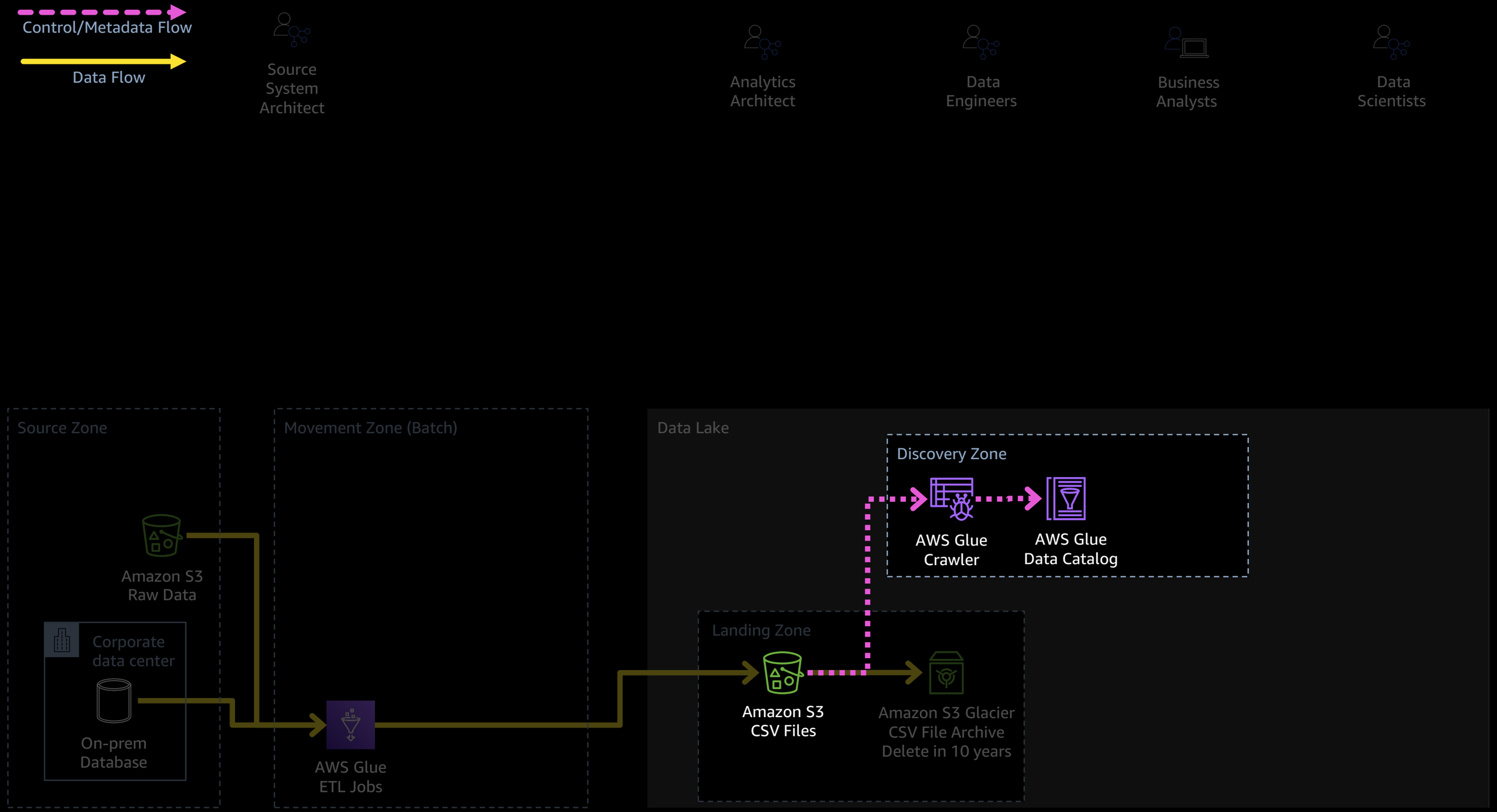
Business Analysts

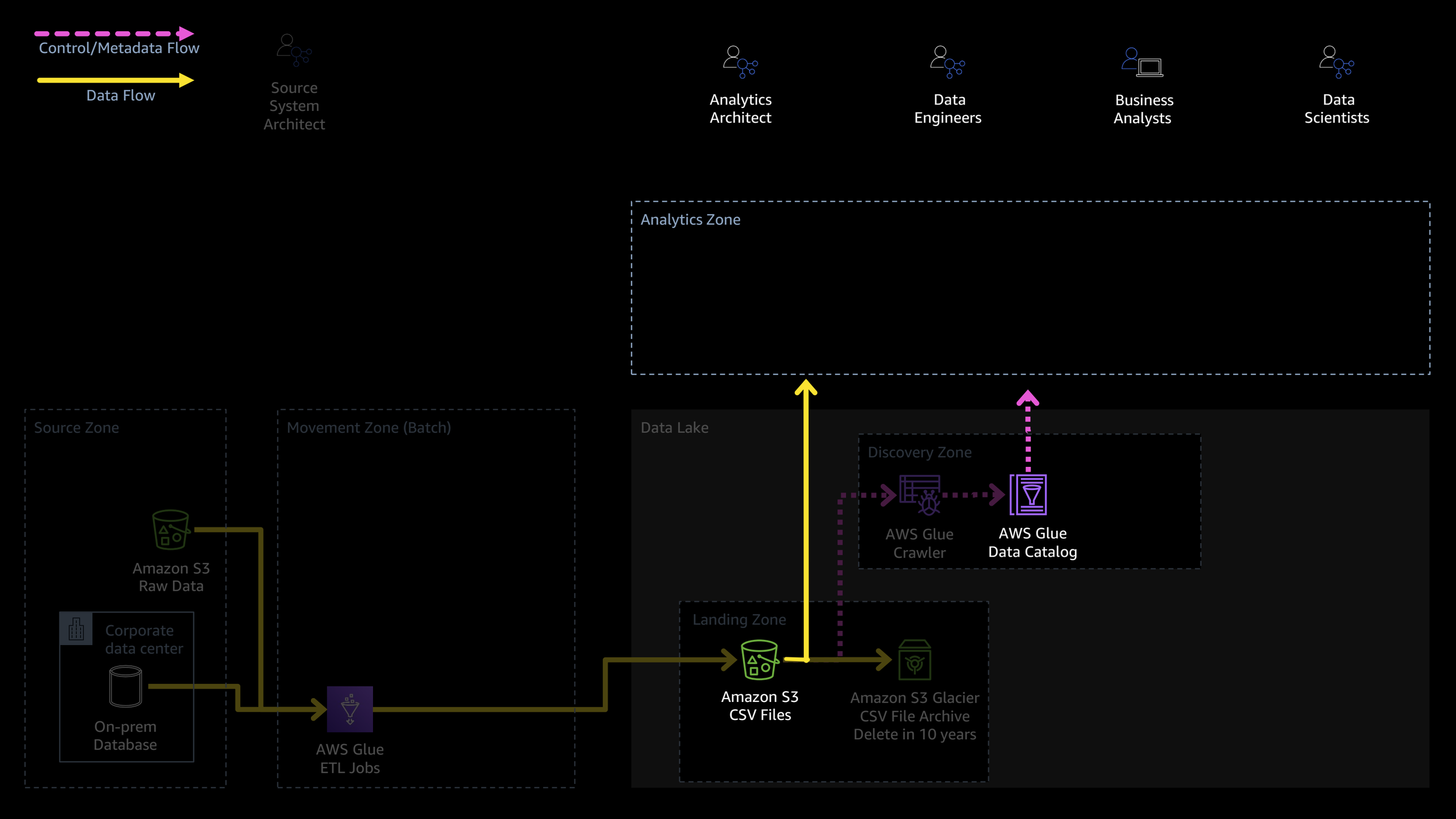
Data Scientists

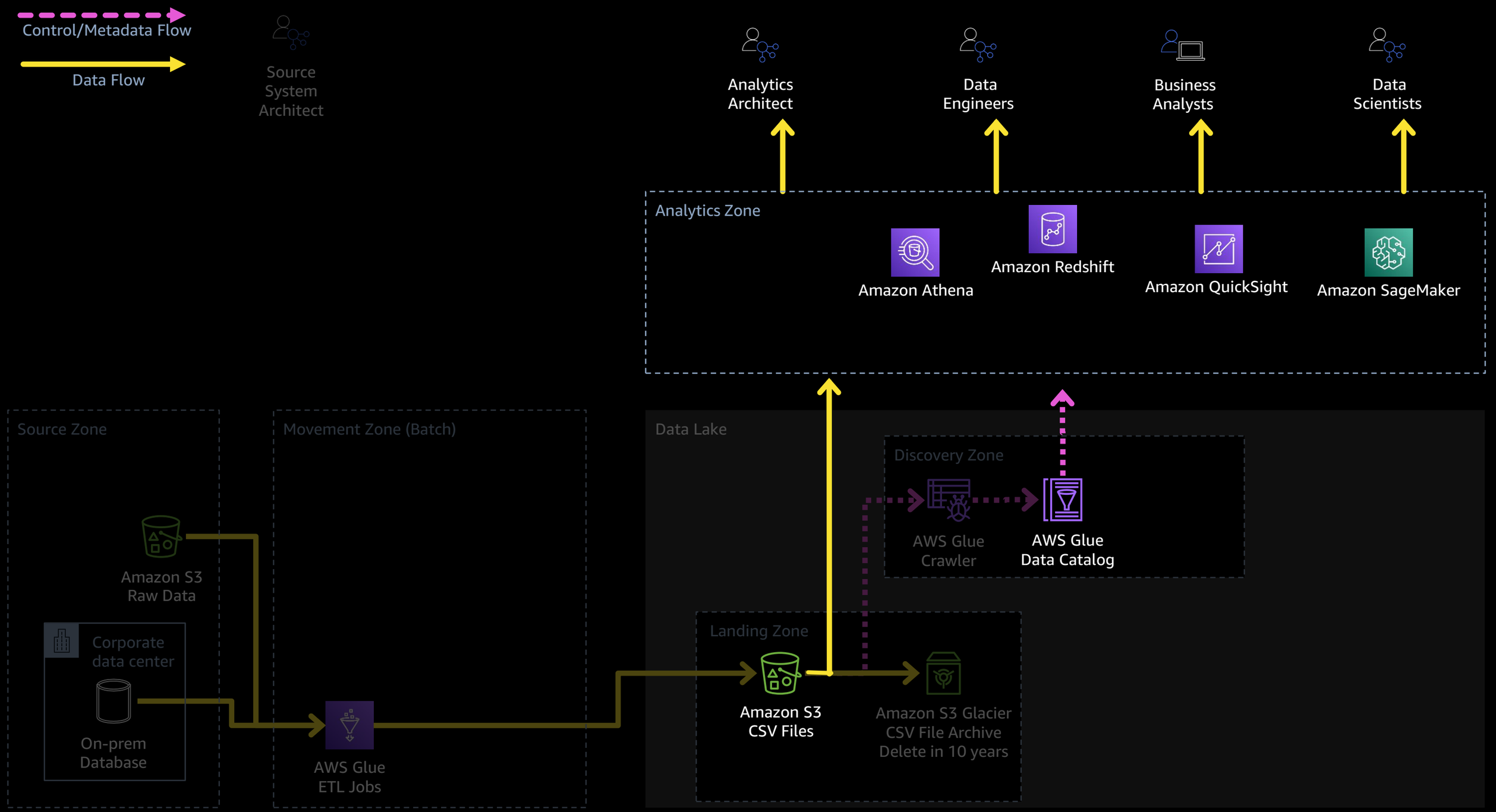
AWS IAM Policy for Landing Zone's S3 Bucket

```
{
  "Version": "2012-10-17",
  "Statement": "Allow",
  "Statement": {
    "Effect": "Allow",
     "Principal": {
       "AWS: "arn:aws:iam::AnalyticsAccount:role/ReadOnly"
     },
    "Action": [
      "s3:Describe*",
      "s3:List*",
      "s3:Get*"
    ],
    "Resource": [
      "arn:aws:s3:::landing-zone-bucket",
      "arn:aws:s3:::landing-zone-bucket/*"
    ]
  }
}
```




Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
Corporate data center
On-prem Database
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
Amazon S3 Raw Data
Corporate data center
On-prem Database
AWS Glue ETL Jobs
AWS Glue Crawler
AWS Glue Data Catalog
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
Corporate data center
On-prem Database
AWS Glue ETL Jobs
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years



Reliability

RELIABILITY: HOW TO BUILD RESILIENT ANALYTIC PIPELINES

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

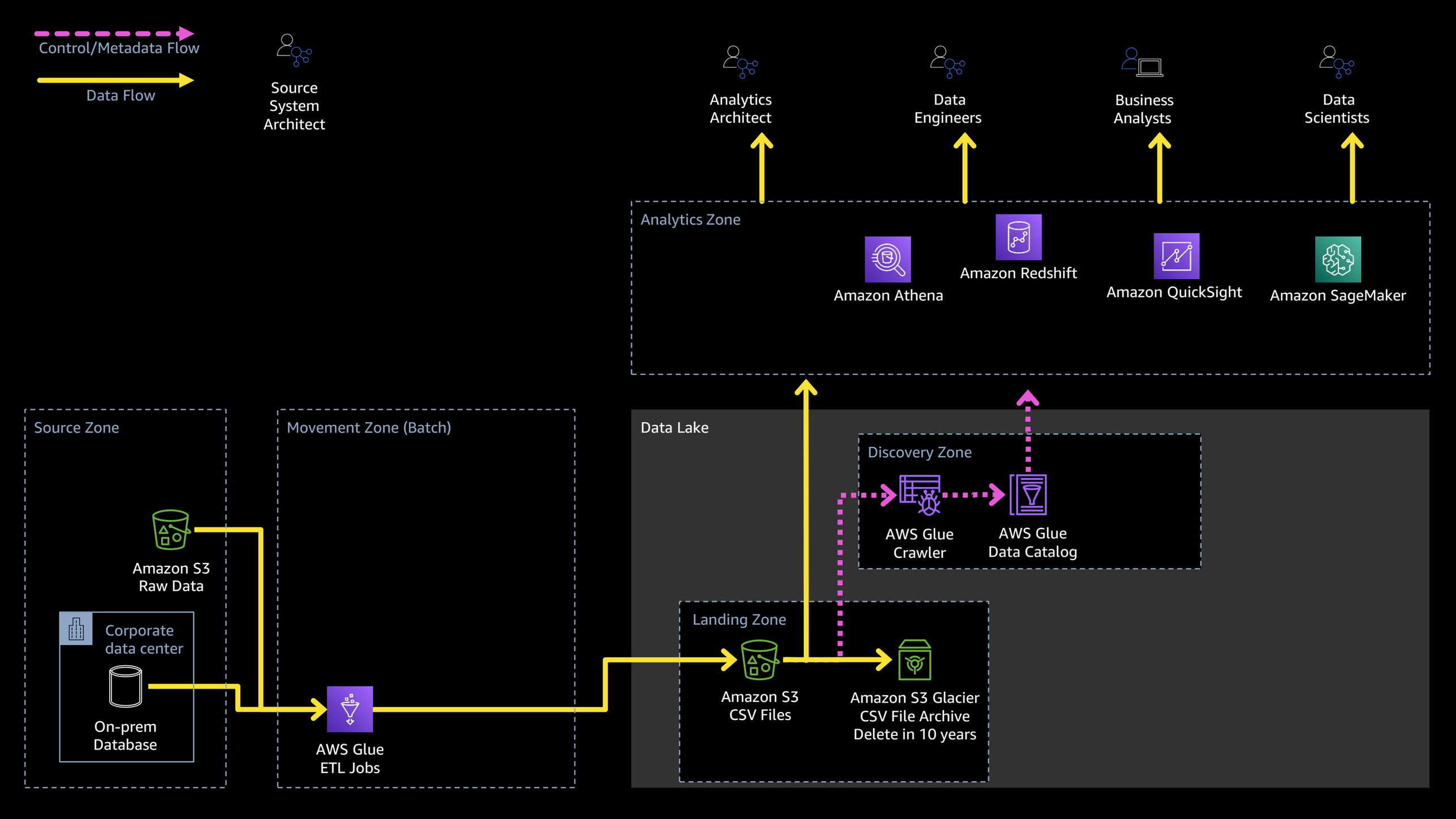
4. Design Resiliency for Analytics Pipelines

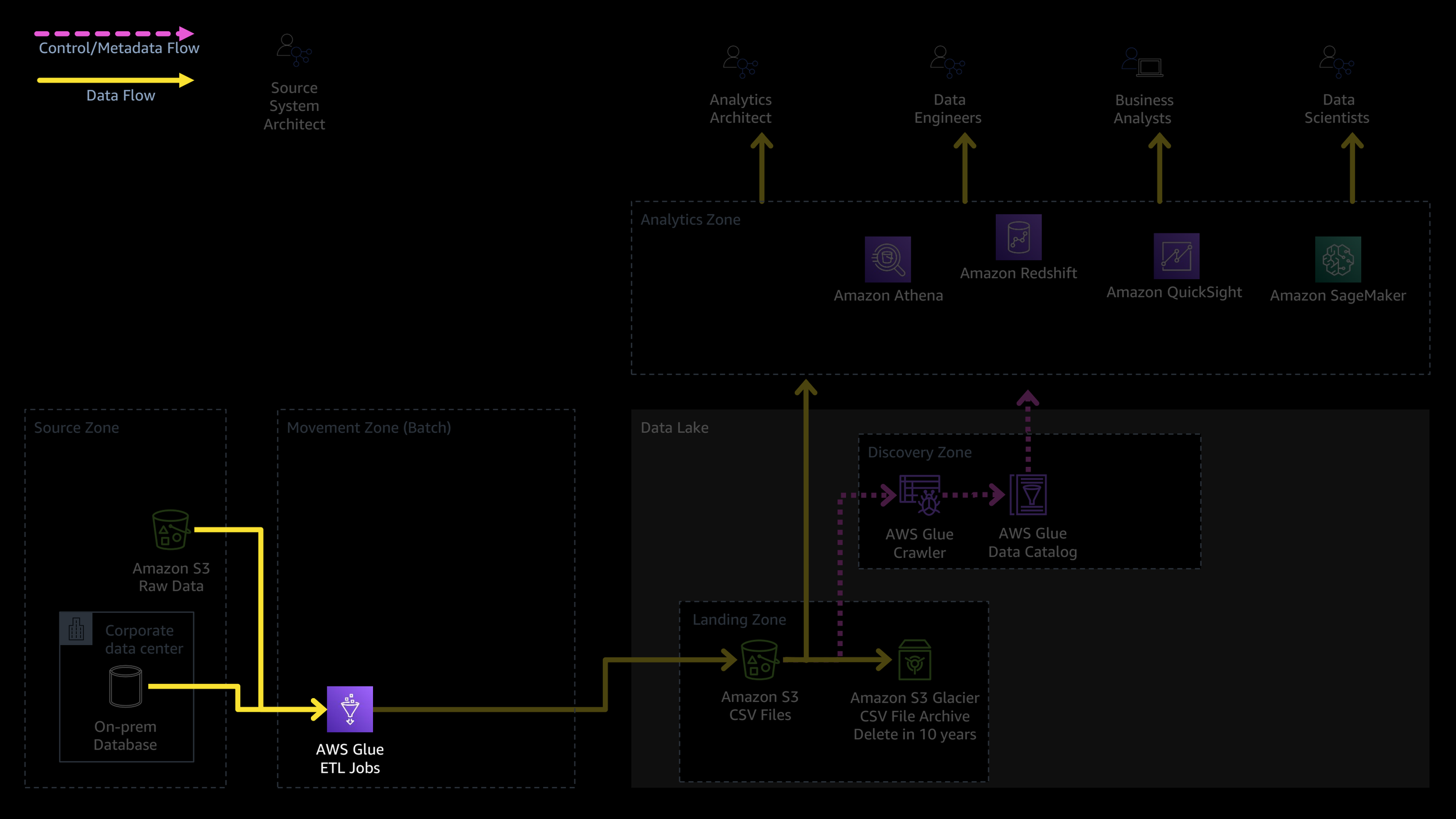
Ask yourself:

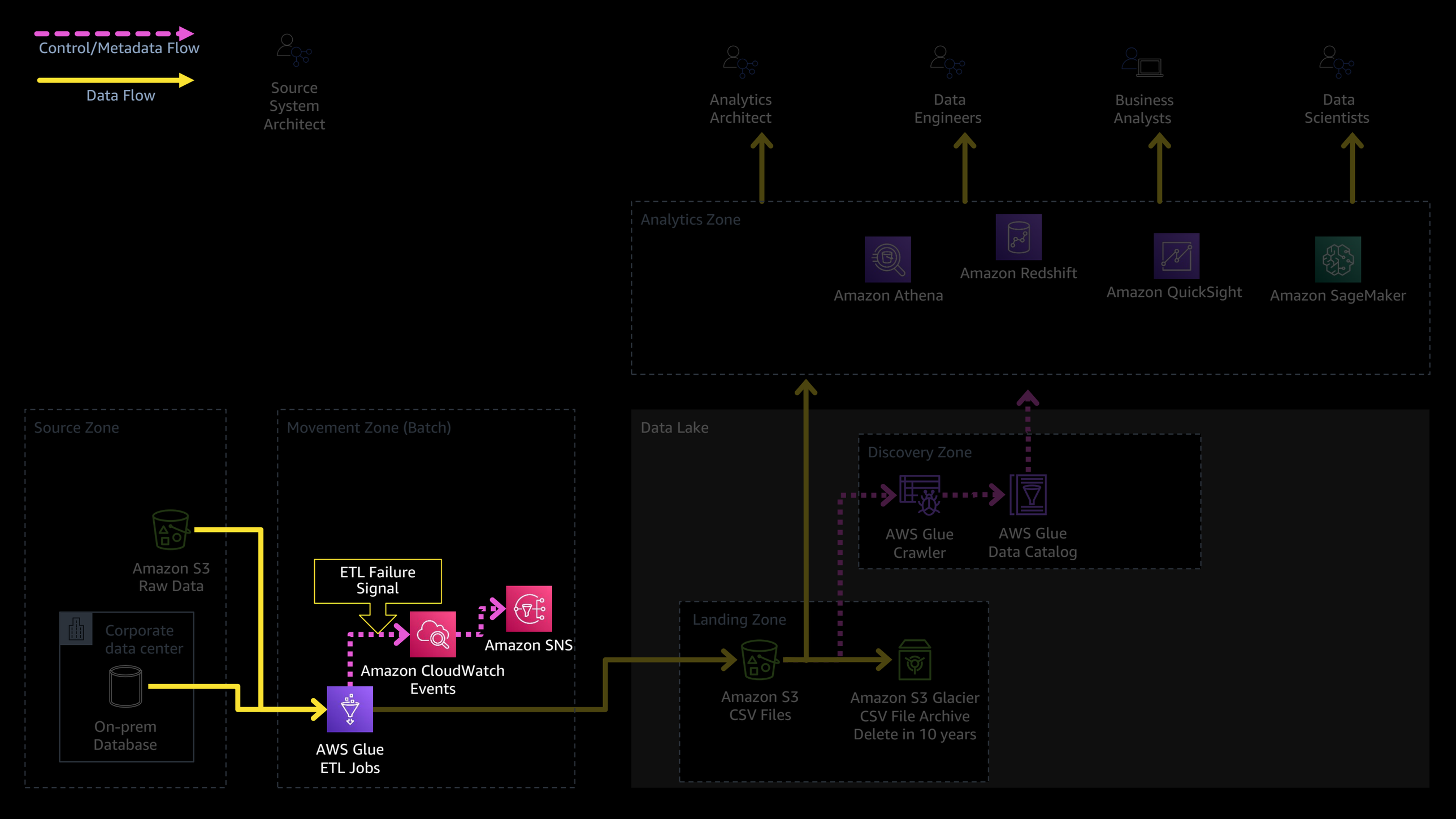
How do you design analytics workloads to withstand ETL job failures?

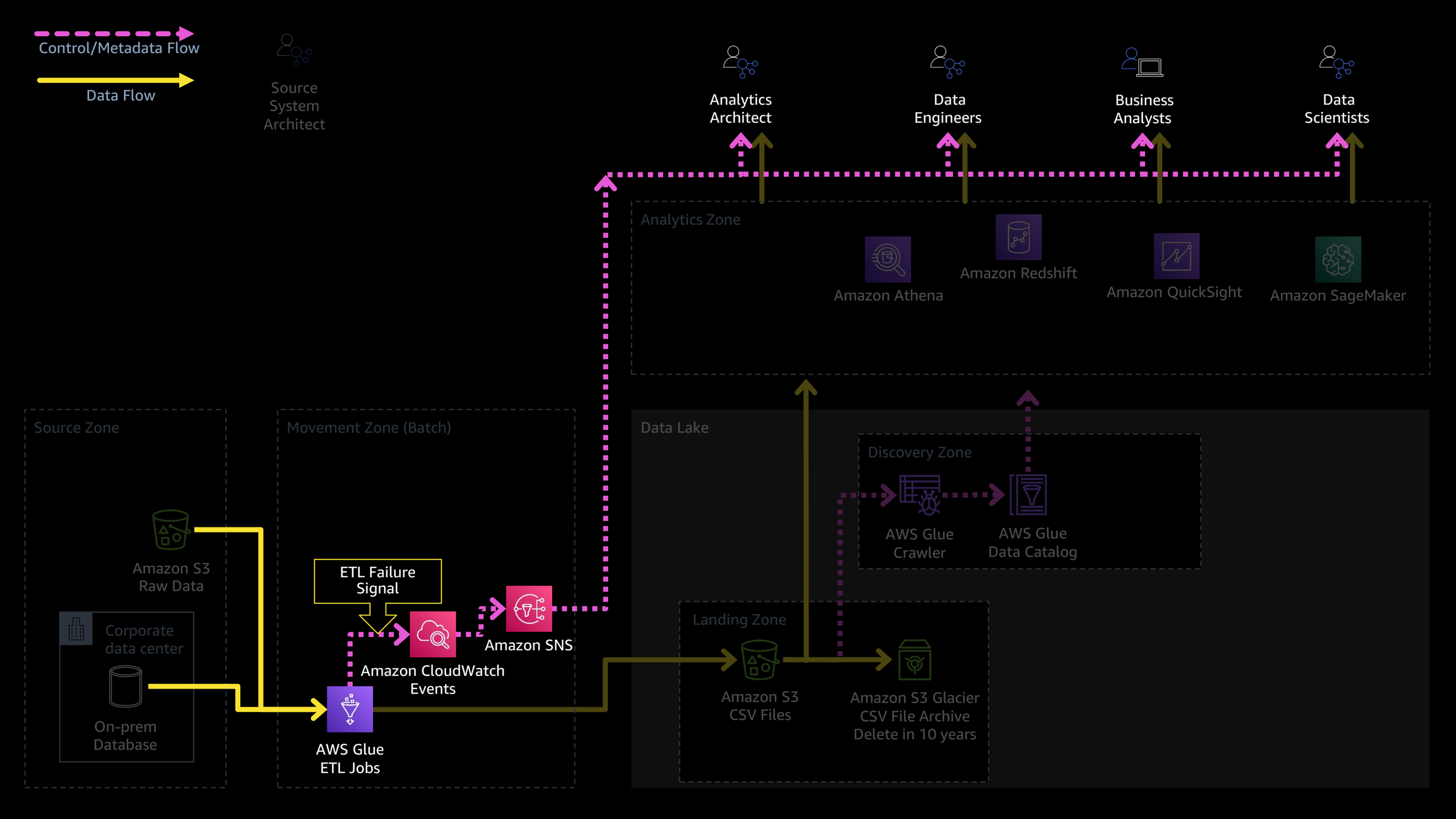
You should

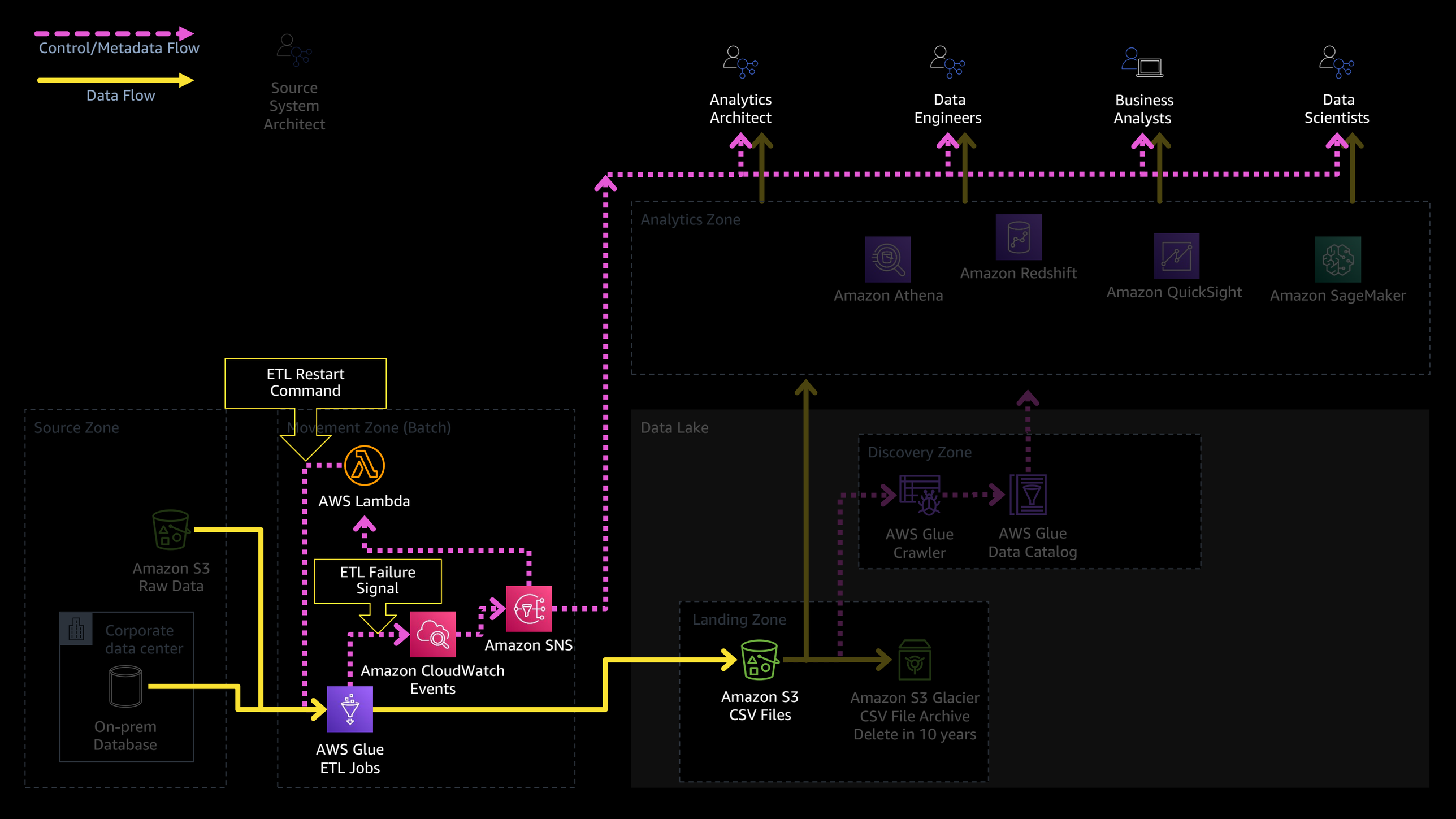
- *Understand business requirements of ETL jobs.*
- *Monitor analytics systems to detect ETL job failures.*
- *Notify source, analytics, and downstream workloads about ETL failures.*
- *Automate recovery of ETL job failures.*


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
Corporate data center
On-prem Database
AWS Glue ETL Jobs
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
Corporate data center
On-prem Database
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
Corporate data center
On-prem Database
ETL Failure Signal
Amazon SNS
Amazon CloudWatch Events
AWS Glue ETL Jobs
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
ETL Failure Signal
Amazon SNS
Amazon CloudWatch Events
Corporate data center
On-prem Database
AWS Glue ETL Jobs
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
ETL Restart Command
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
ETL Failure Signal
Amazon SNS
Corporate data center
Amazon CloudWatch Events
Landing Zone
On-prem Database
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years
AWS Glue ETL Jobs

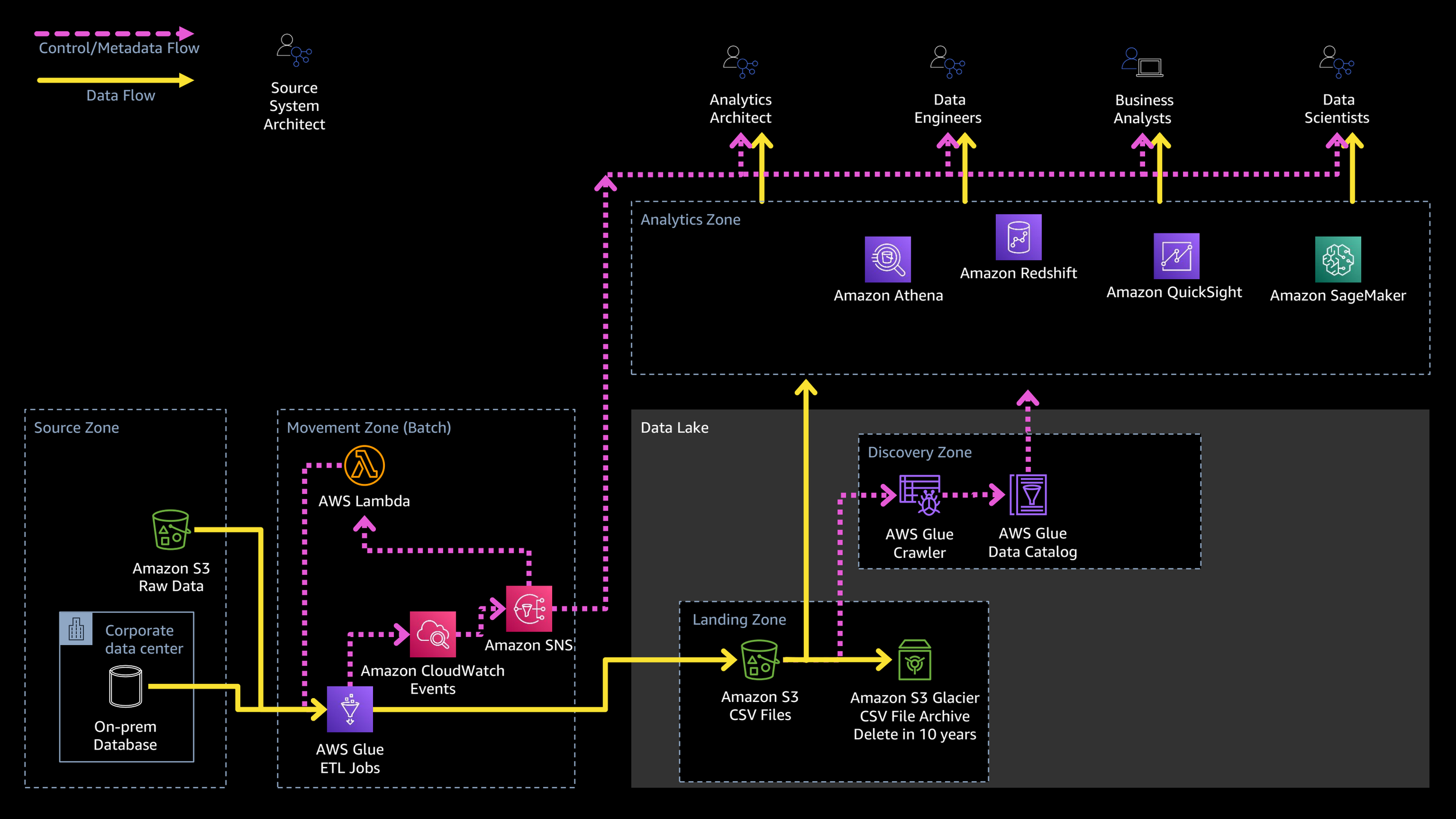
5. Govern Data Changes

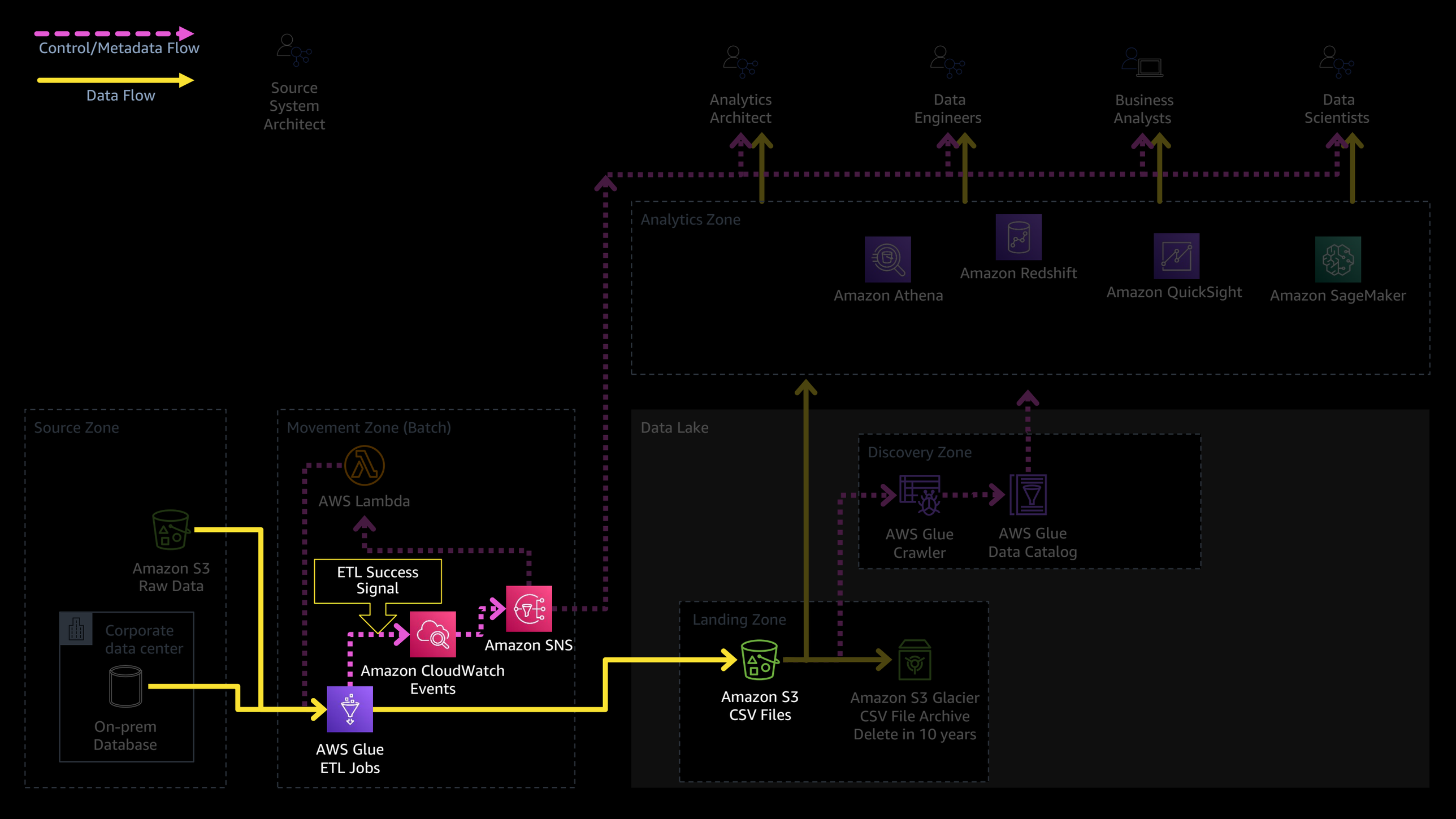
Ask yourself:

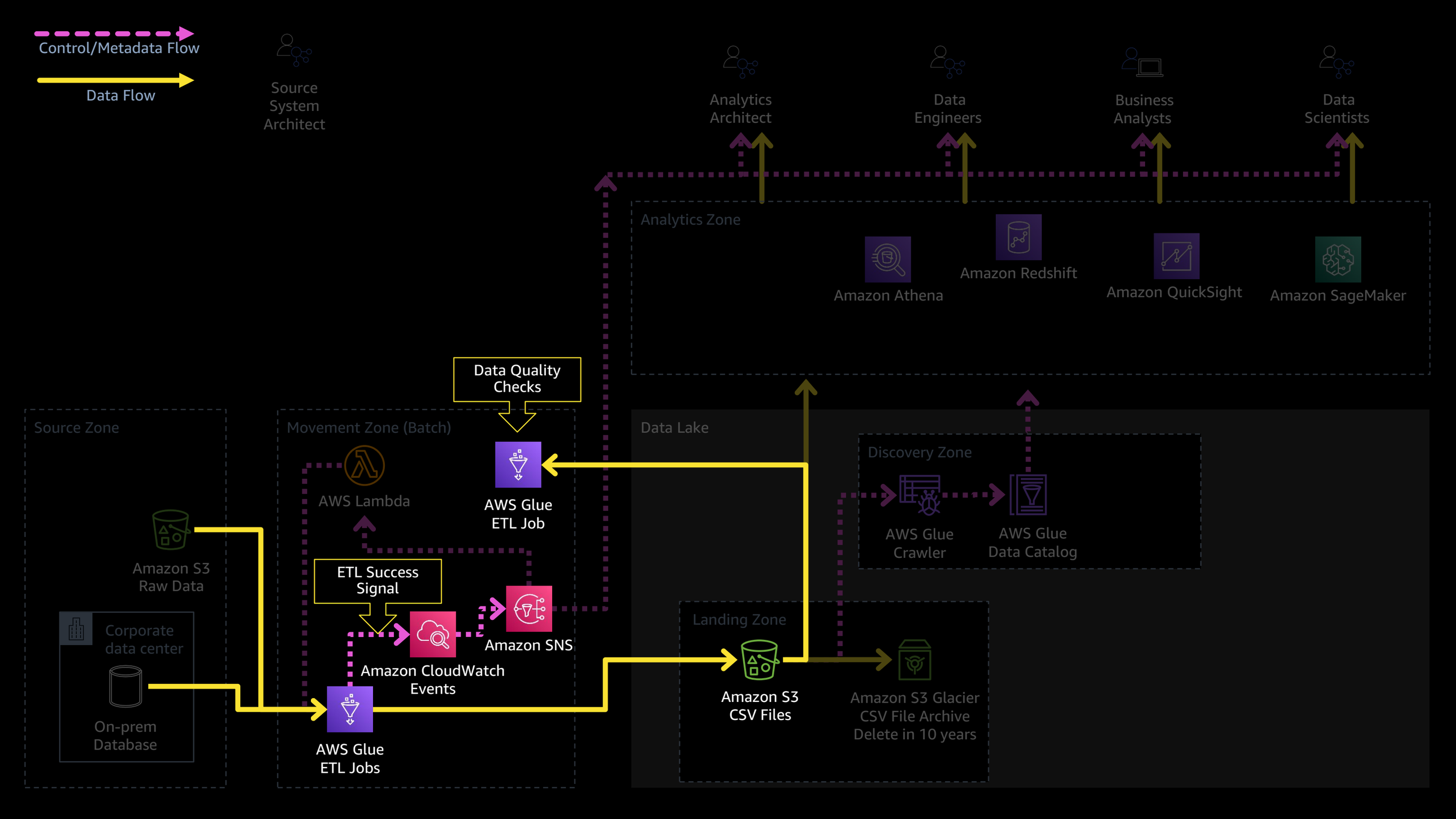
How do you govern data changes?

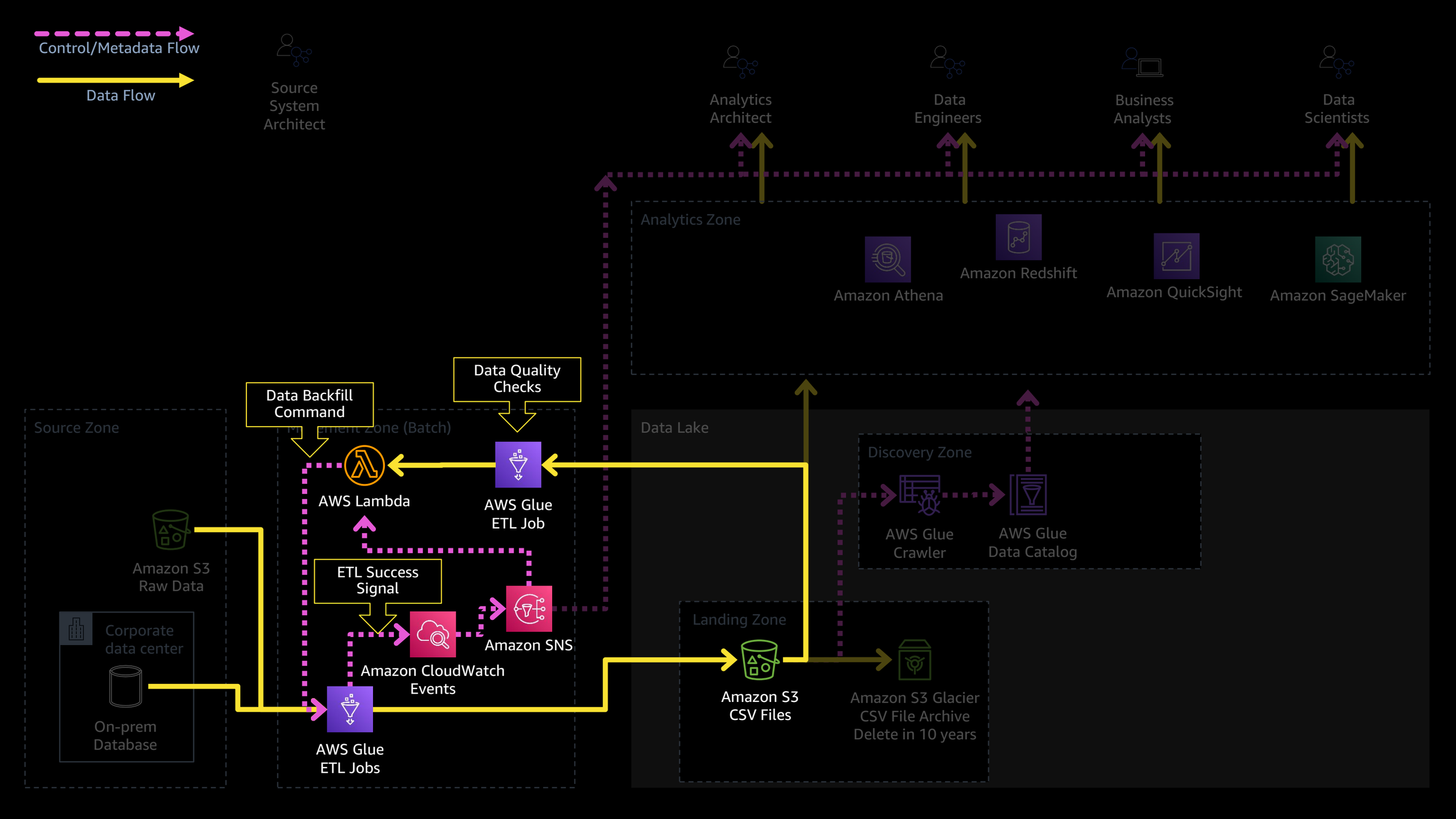
You should

- *Understand business requirements such as delivery SLAs*
- *Monitor data quality anomalies*
- *Define standard operating procedures (SOPs) such as data backfilling*
- *Automate so data changes propagate to downstream workloads*


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
ETL Success Signal
Amazon SNS
Amazon CloudWatch Events
Corporate data center
On-prem Database
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Data Quality Checks
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
ETL Success Signal
Amazon SNS
Corporate data center
Amazon CloudWatch Events
Landing Zone
On-prem Database
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Data Quality Checks
Data Backfill Command
Source Zone
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
ETL Success Signal
Amazon SNS
Corporate data center
Amazon CloudWatch Events
Landing Zone
On-prem Database
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years
AWS Glue ETL Jobs



Performance Efficiency

PERFORMANCE EFFICIENCY: HOW TO OPTIMIZE PERFORMANCE

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

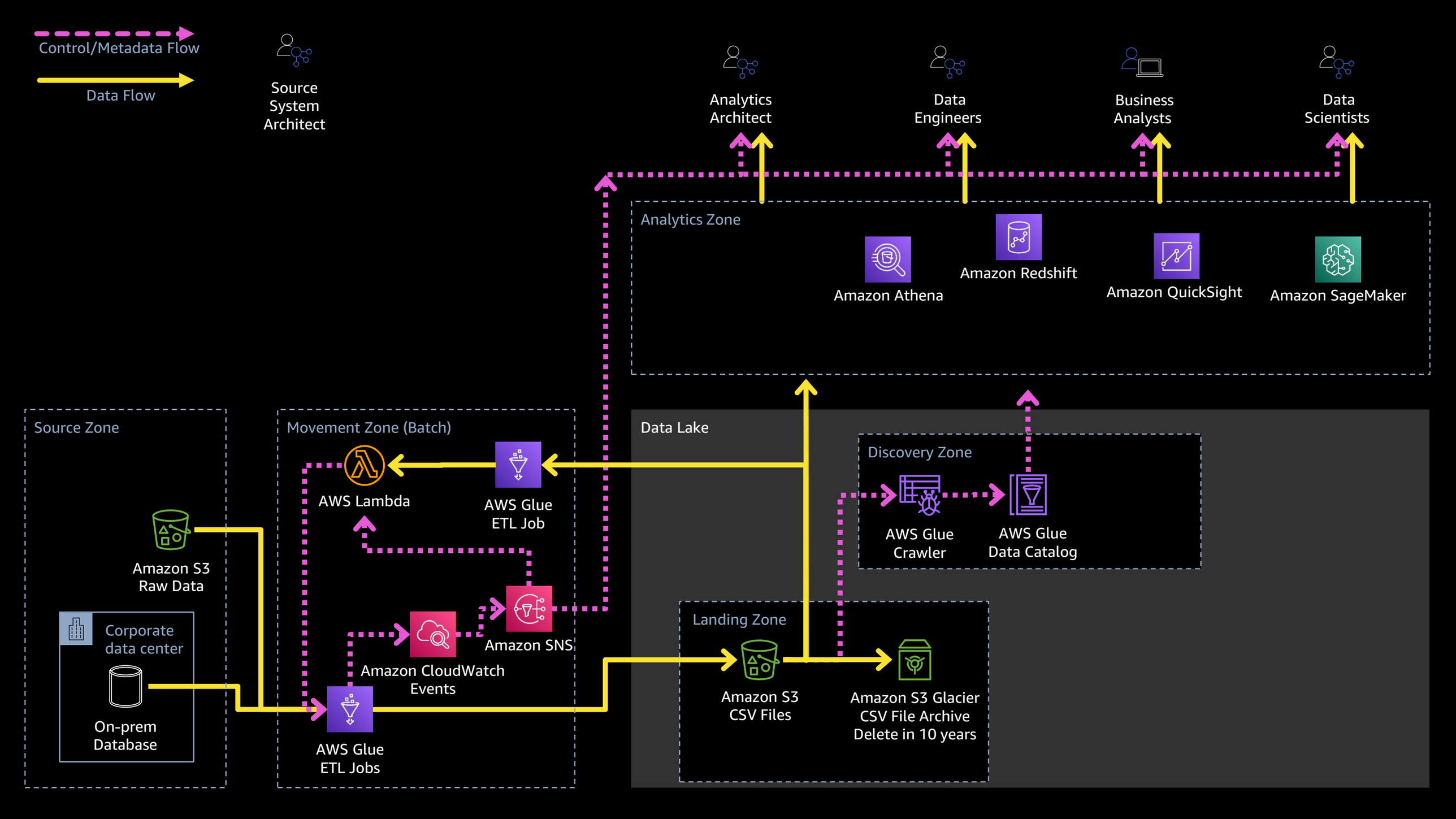
6. Choose Optimal File Format and Partitioning

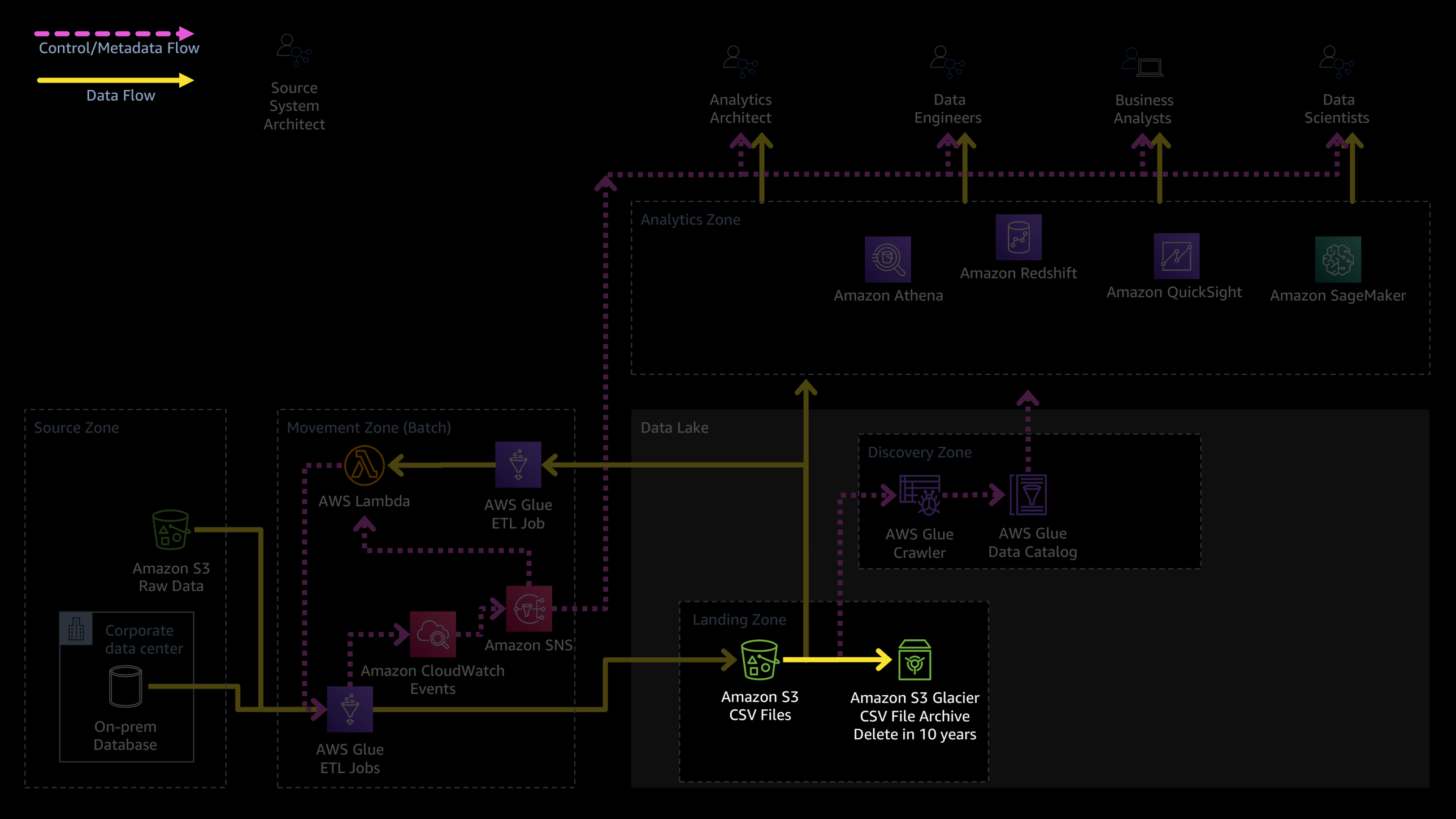
Ask yourself

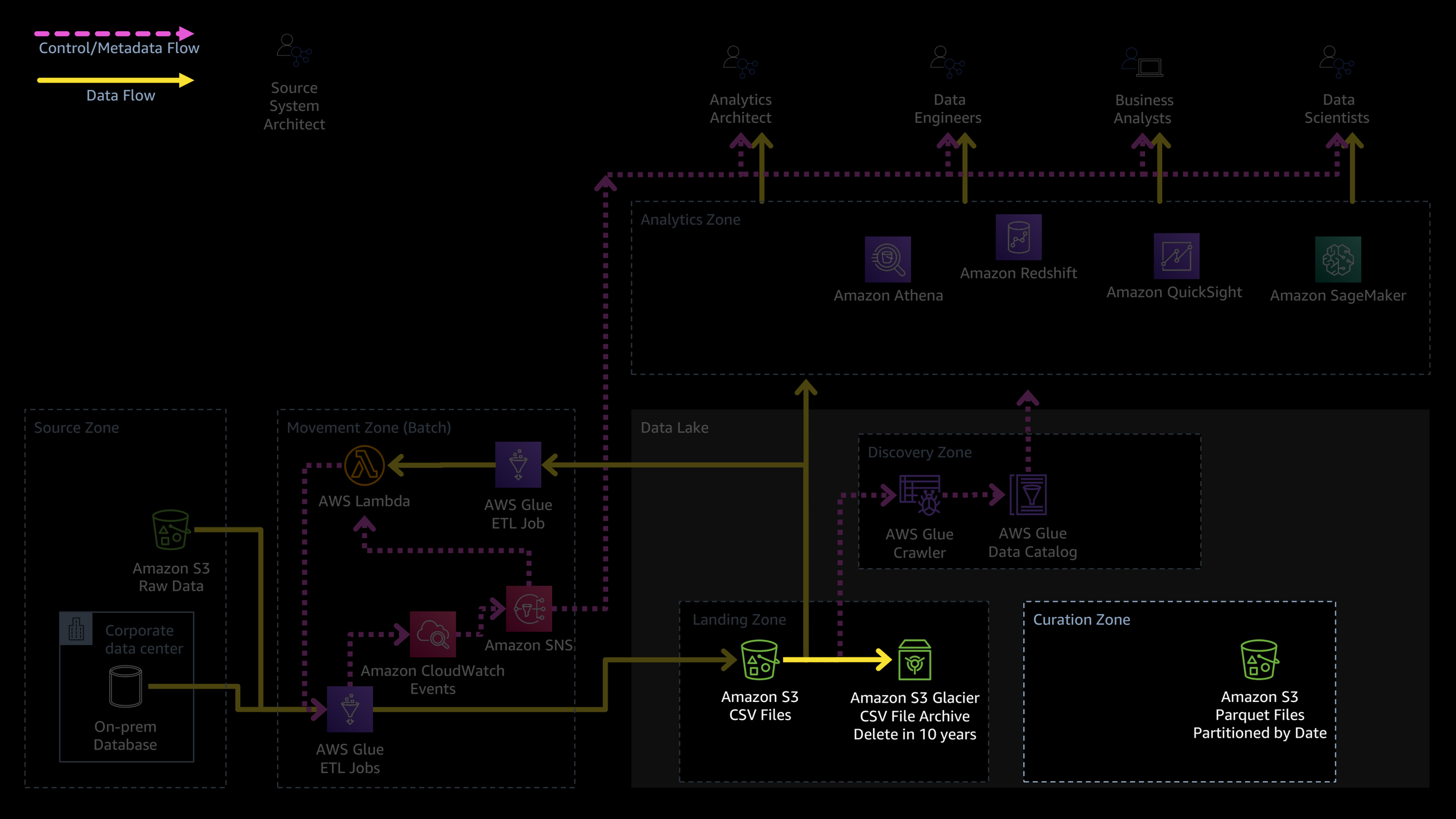
How do you select the best-performing file formats and partitioning?

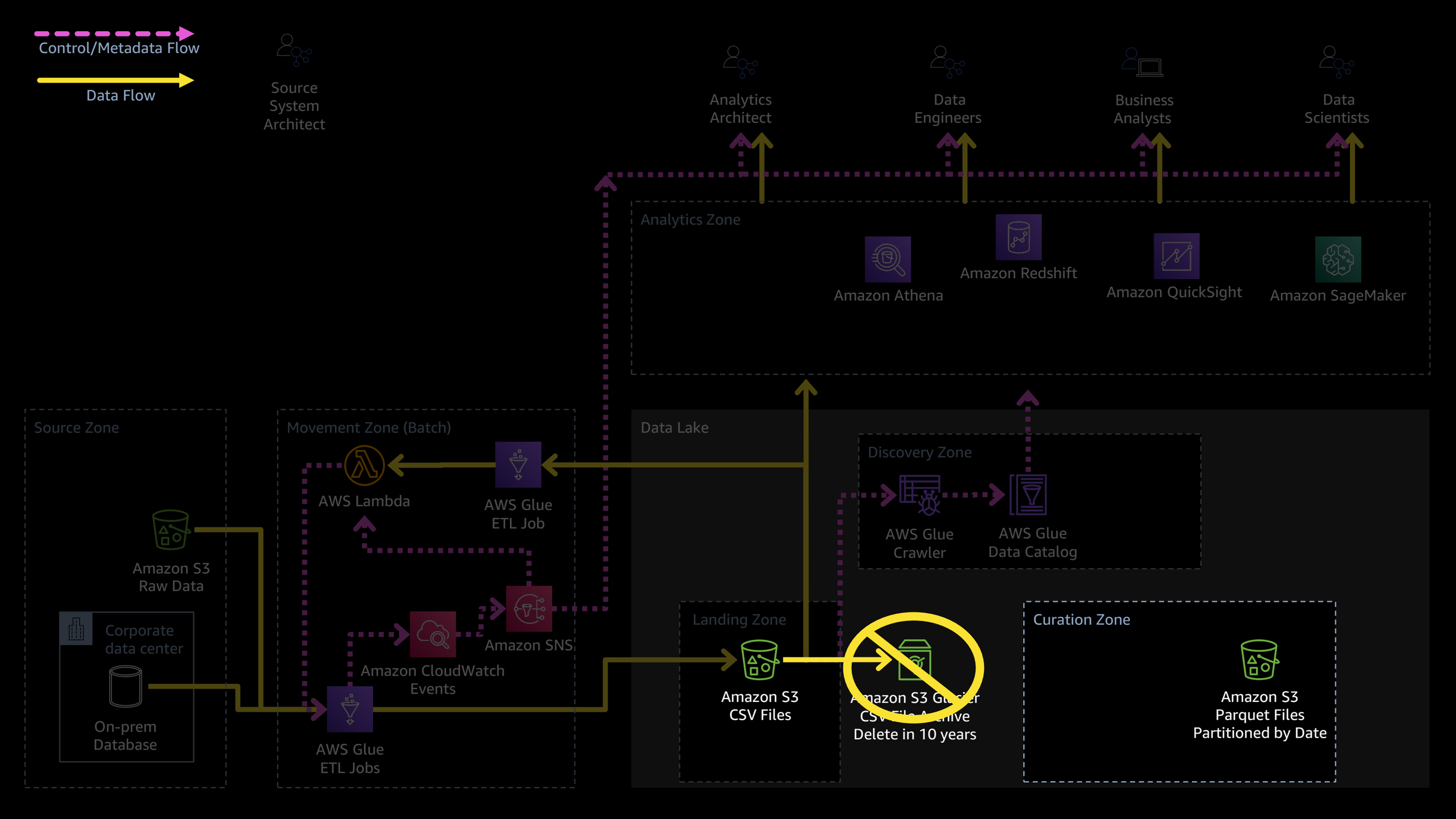
You should

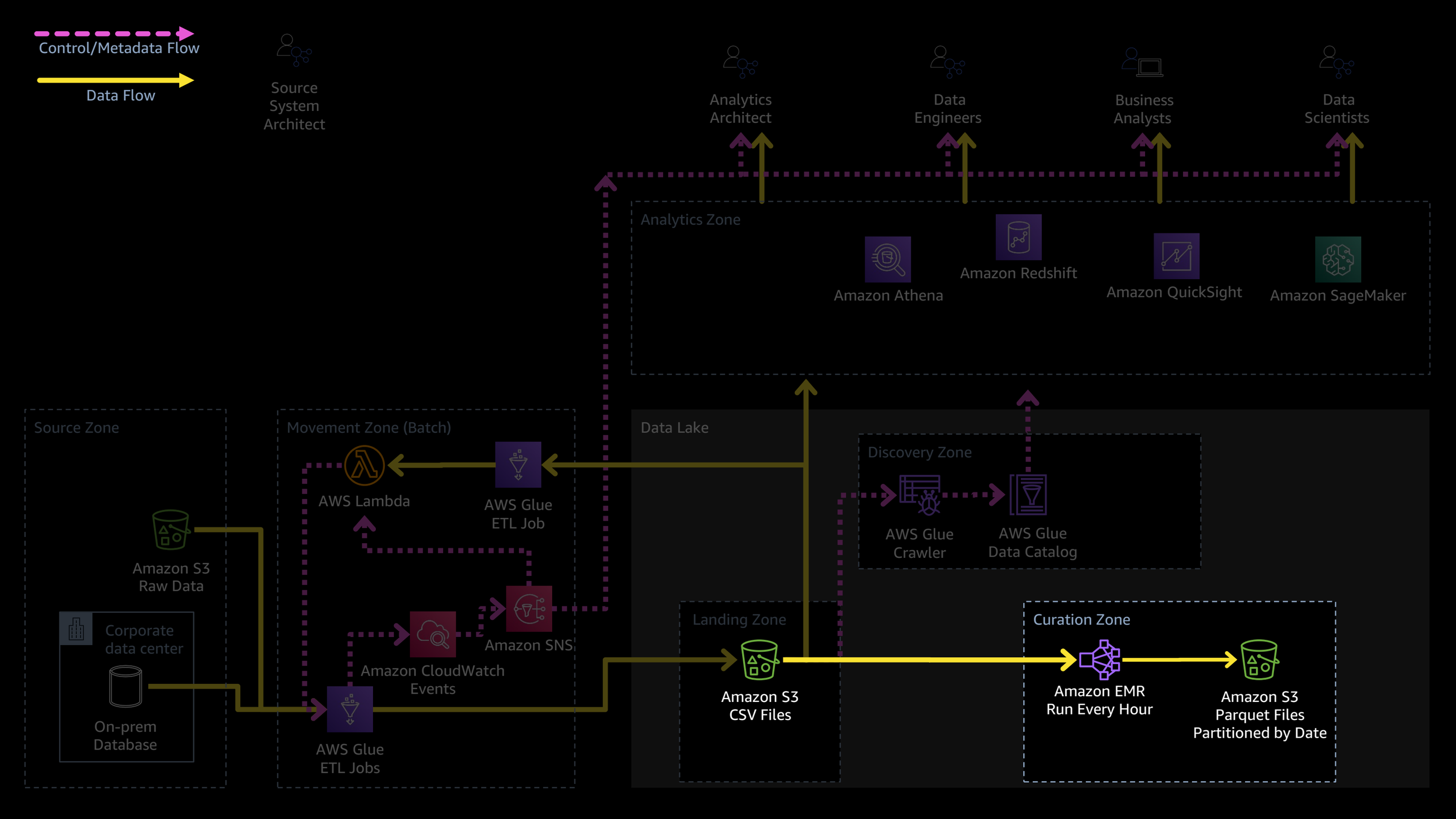
- *Select format based on data update patterns (append-only vs. in-place update)*
- *Select format based on data read patterns (`SELECT … WHERE … GROUP BY …`)*
- *Use file compression to improve file IO efficiency*
- *Partition the data to avoid unnecessary file reads*

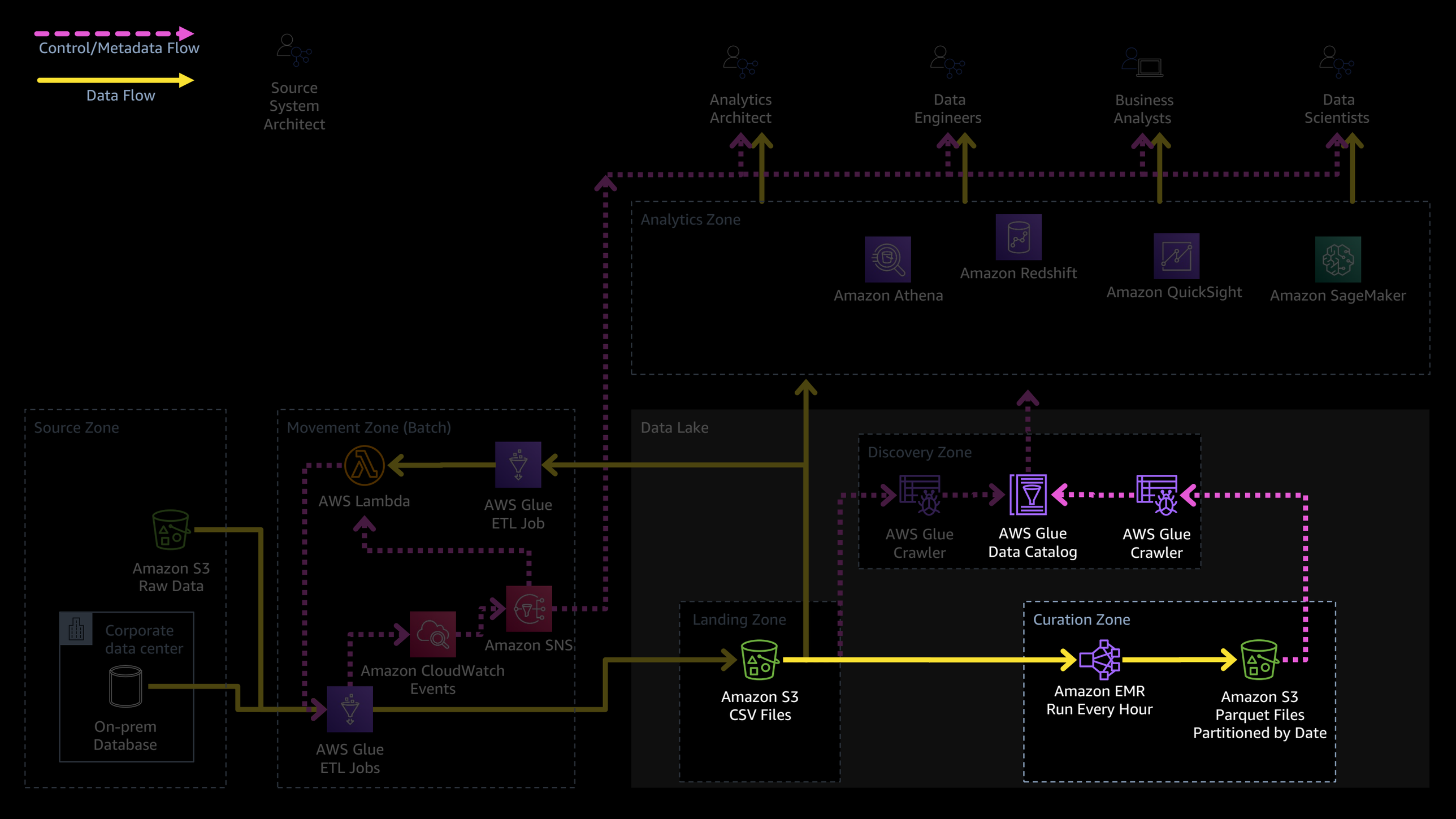

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years
AWS Glue ETL Jobs

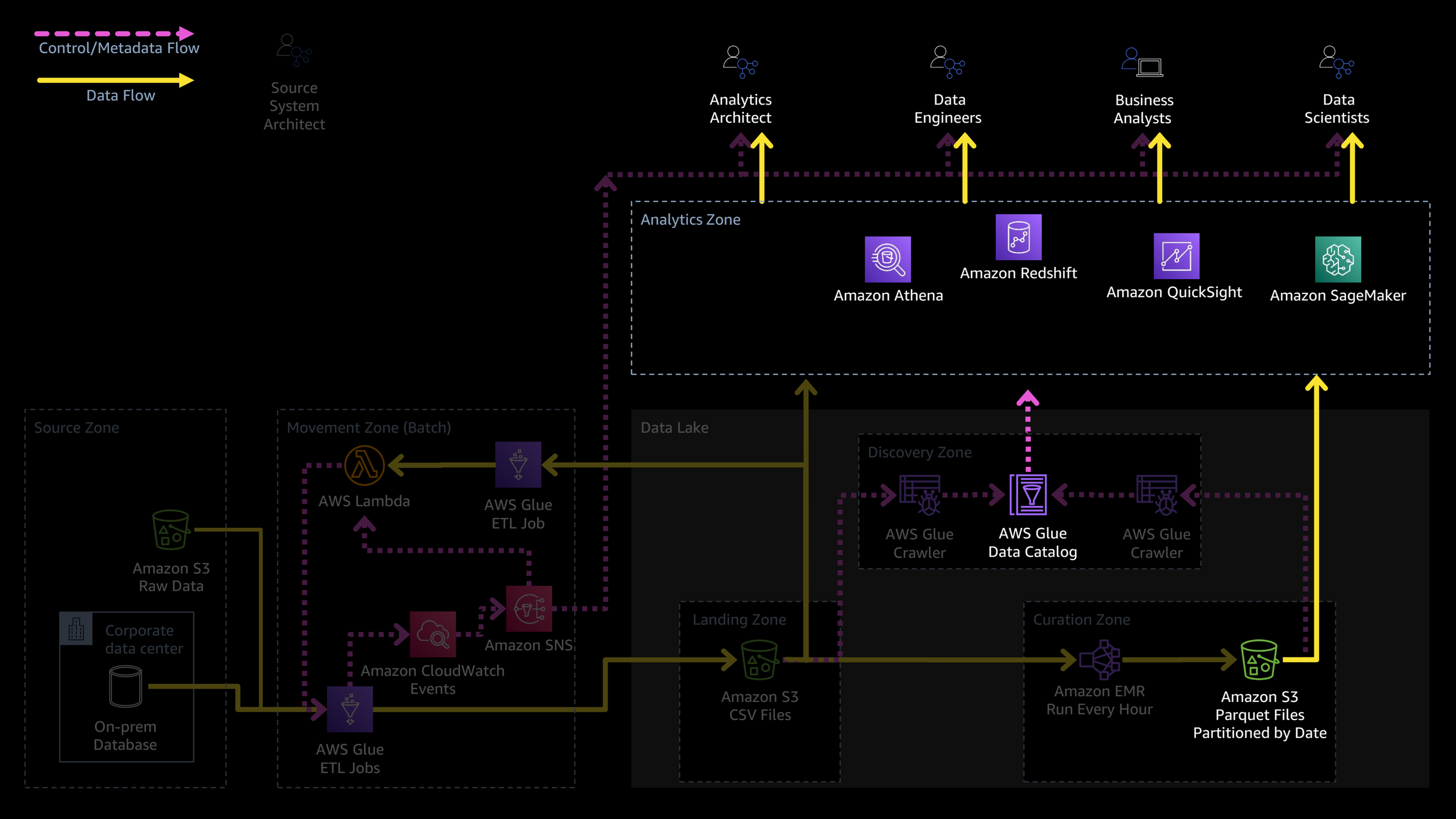

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years
AWS Glue ETL Jobs

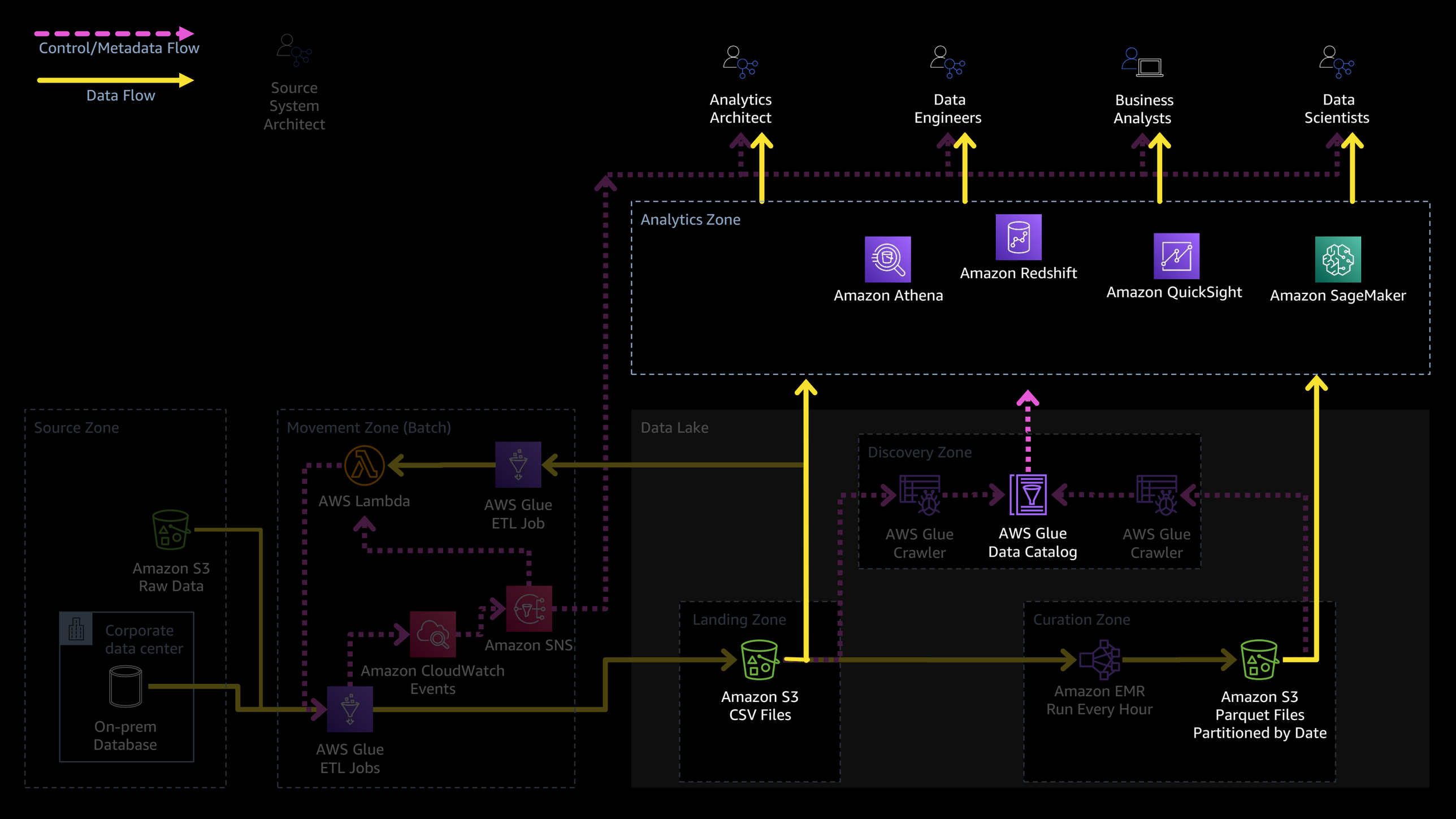

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
AWS Glue ETL Jobs
Landing Zone
Amazon S3 CSV Files
Amazon S3 Glacier CSV File Archive Delete in 10 years
Curation Zone
Amazon S3 Parquet Files Partitioned by Date


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Amazon S3 CSV Files
Curation Zone
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs
Delete in 10 years


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs

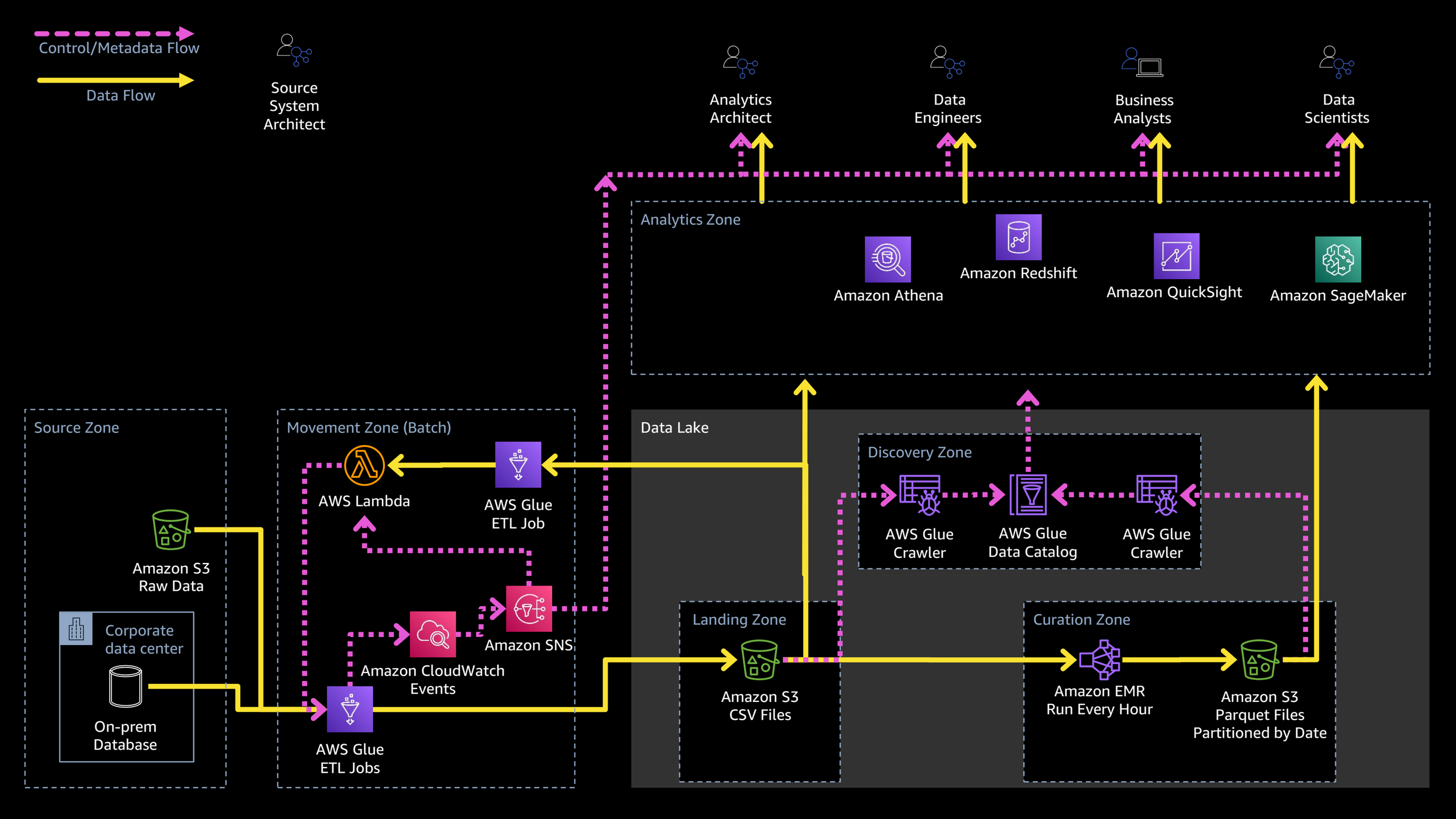
7. Choose the Optimal Storage Solution

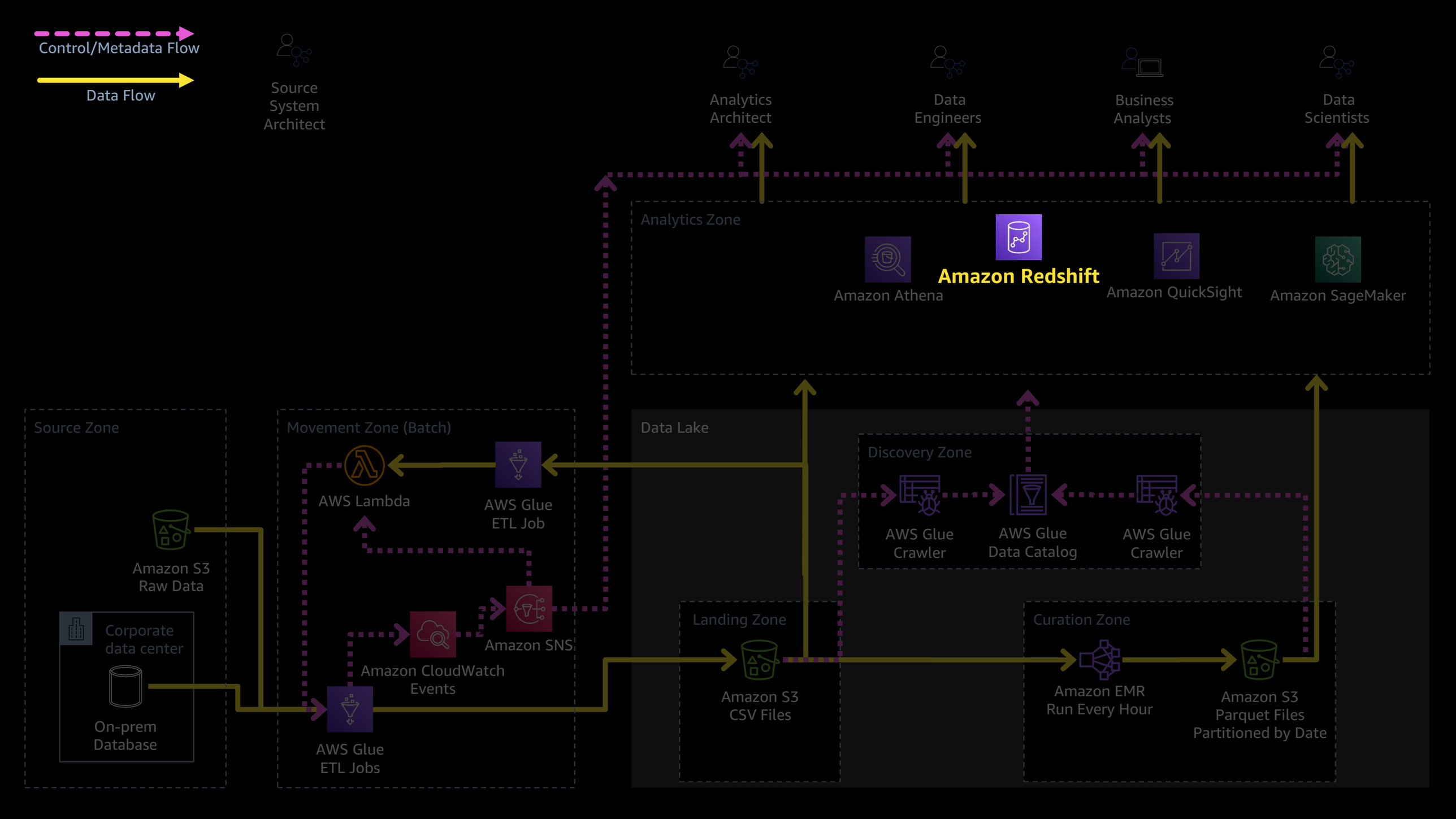
Ask yourself:

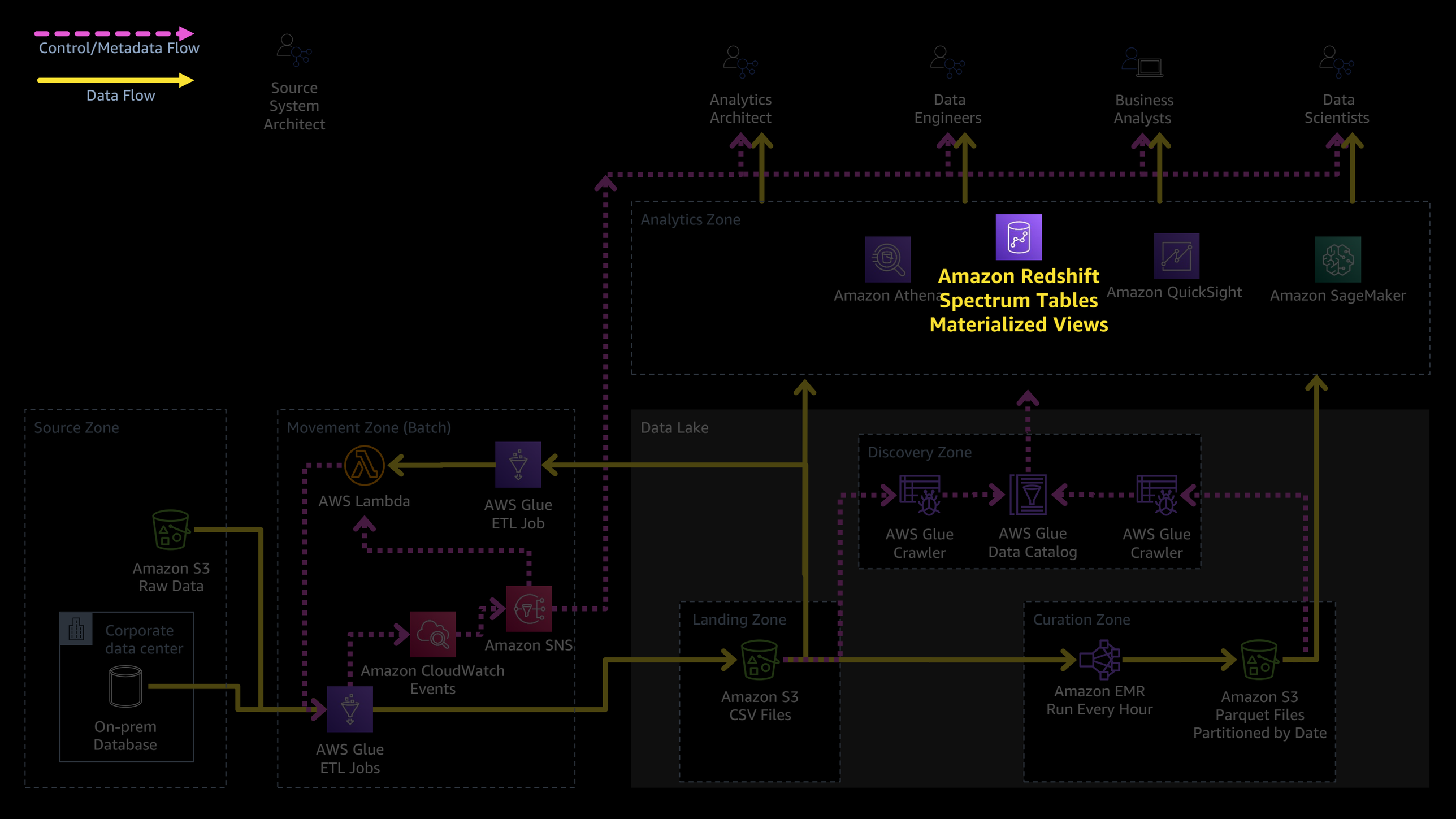
How do you select the best storage options for your workload?

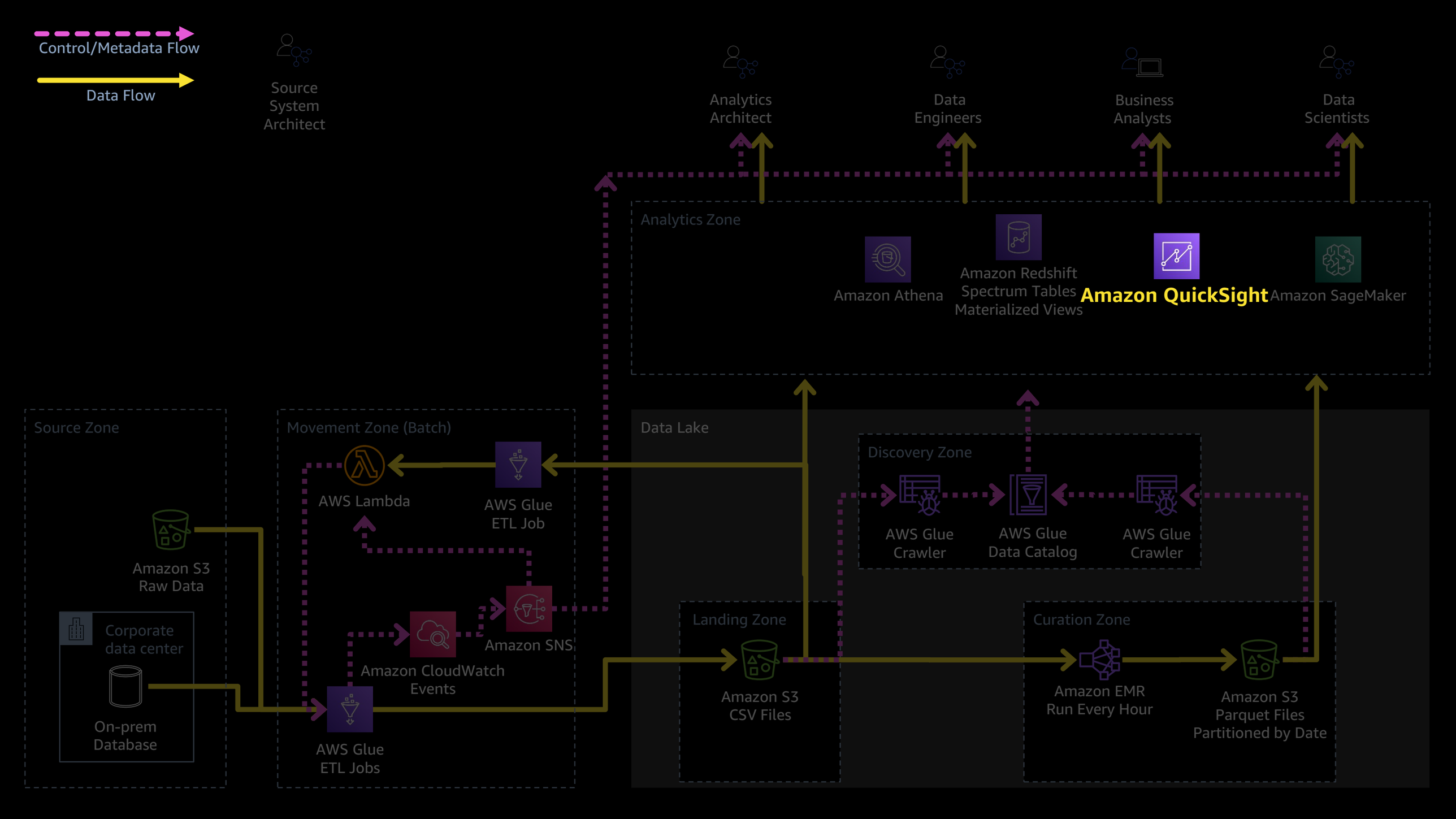
You should

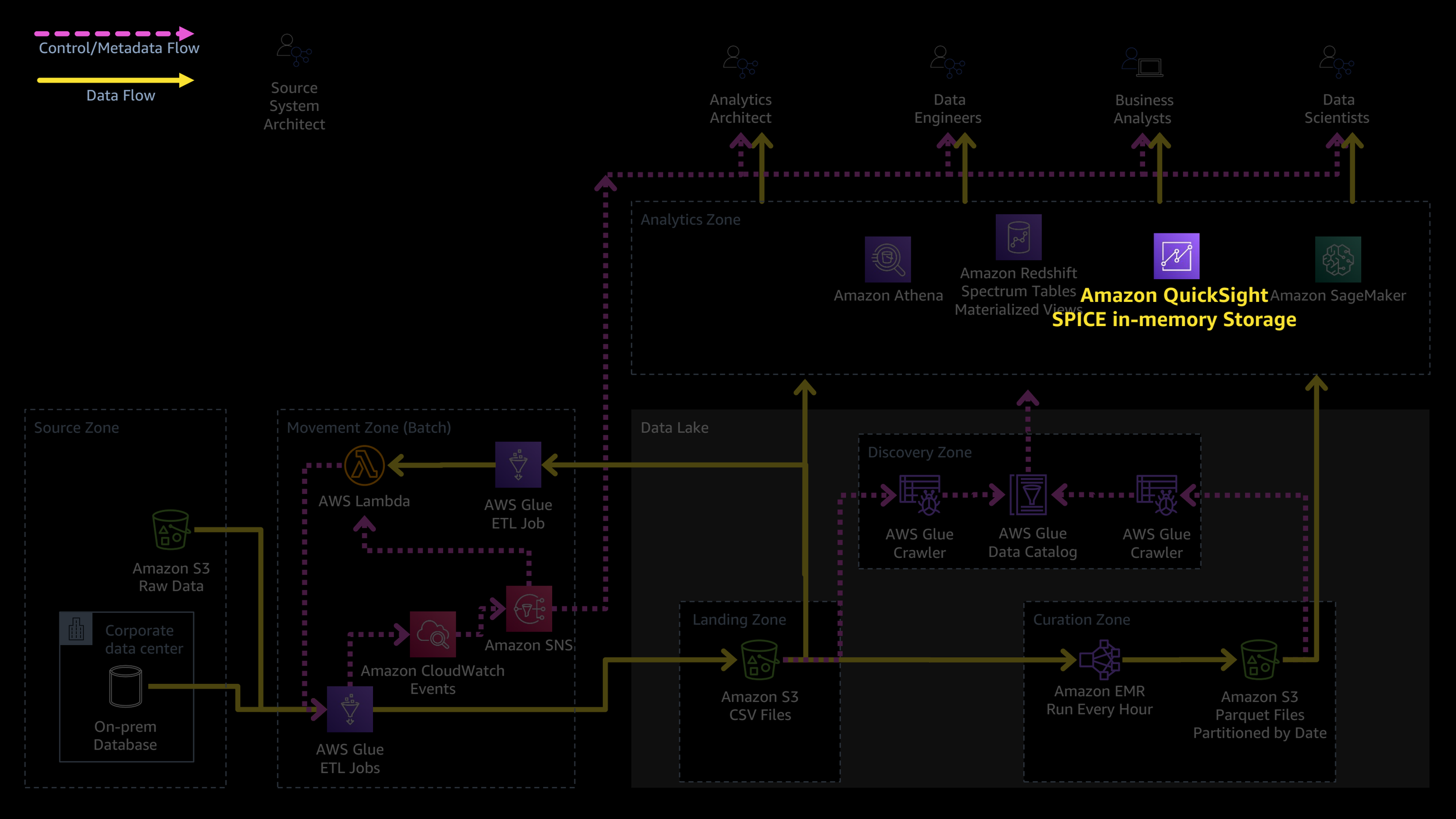
- *Understand storage IO demands and characteristics for each workload component*
- *Evaluate available configuration options*
- *Make decisions based on access patterns, data growth, and metrics*

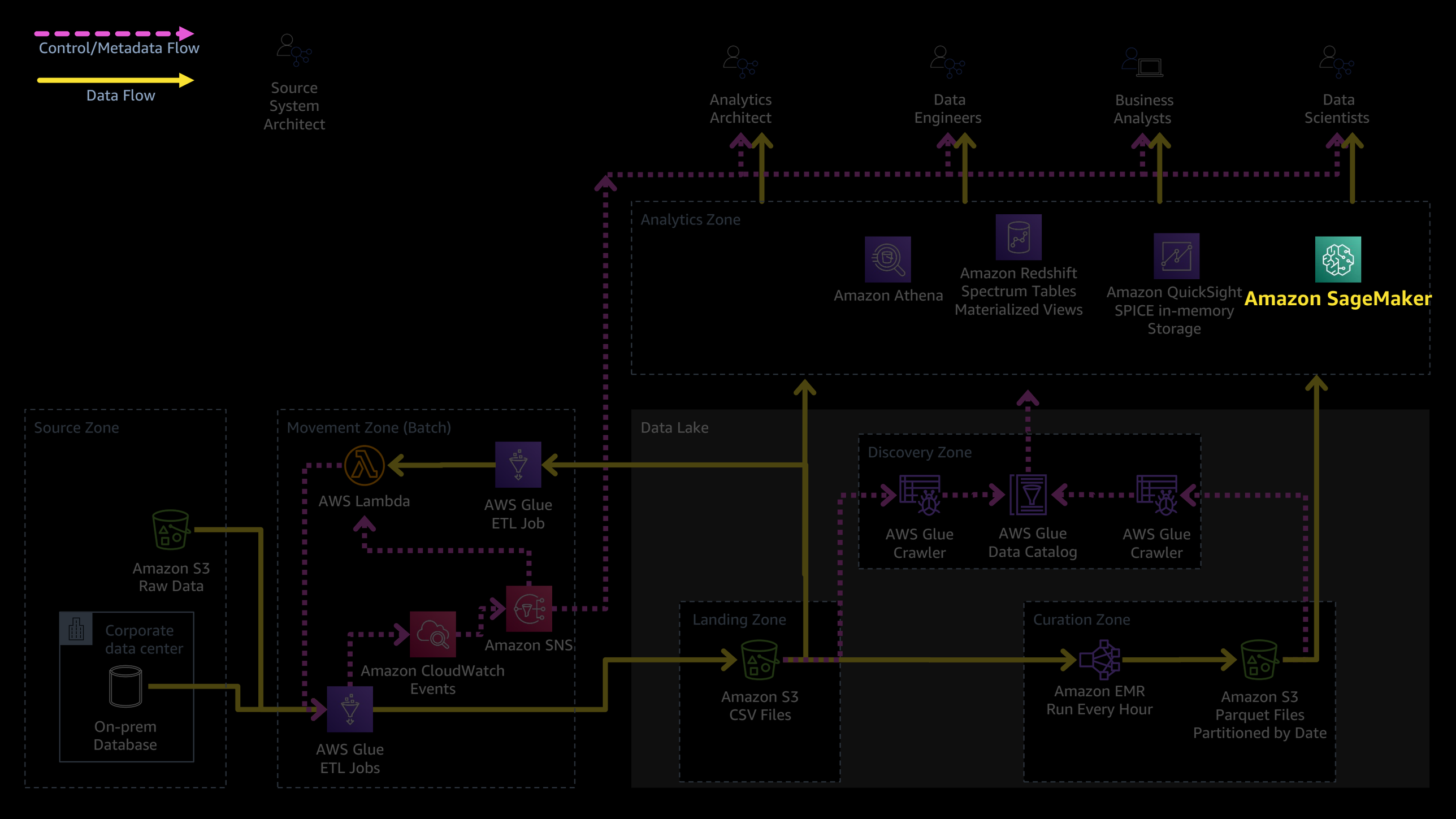

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs

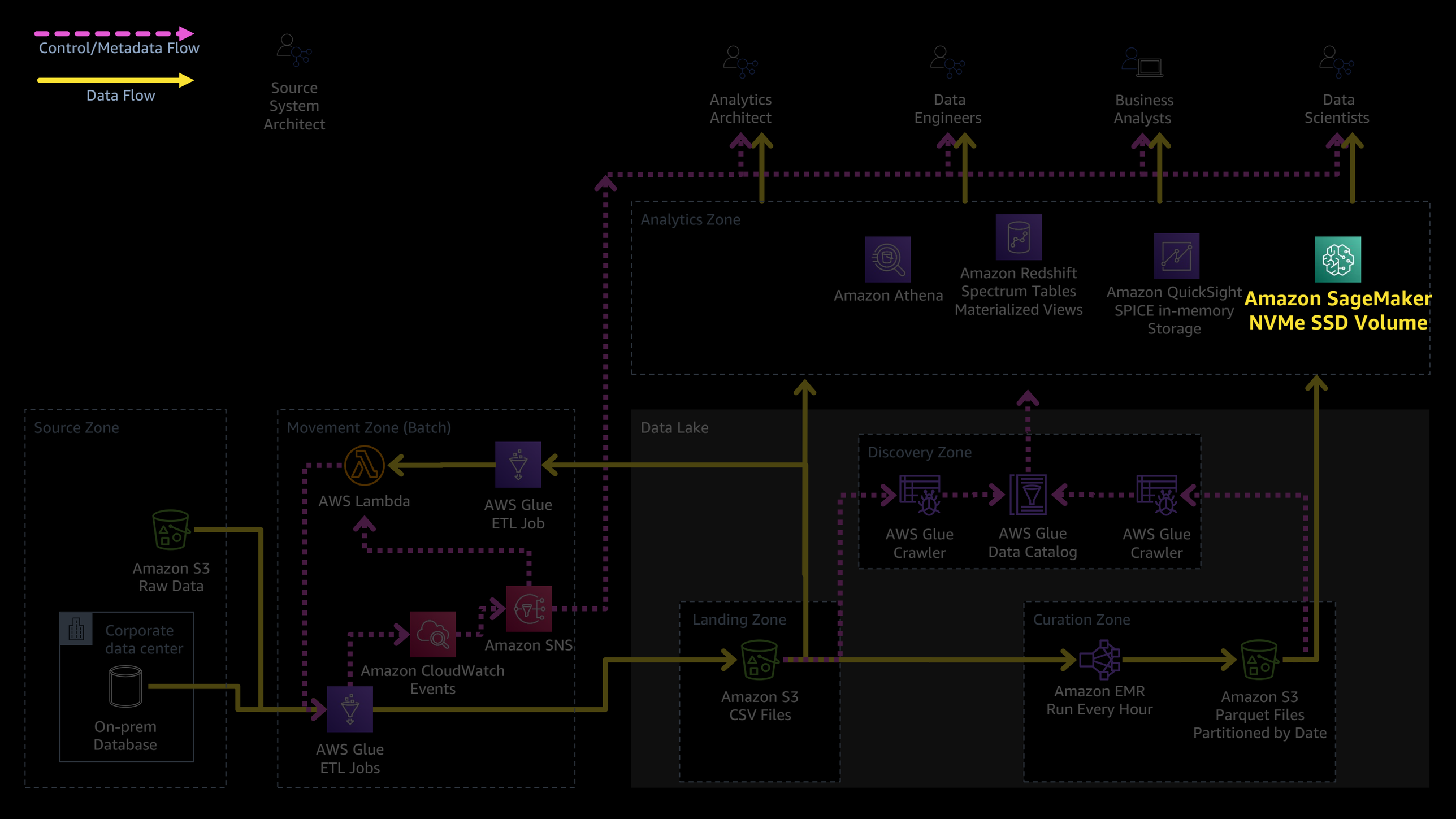

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views
Amazon QuickSight
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views
Amazon QuickSight
SPICE in-memory Storage
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs

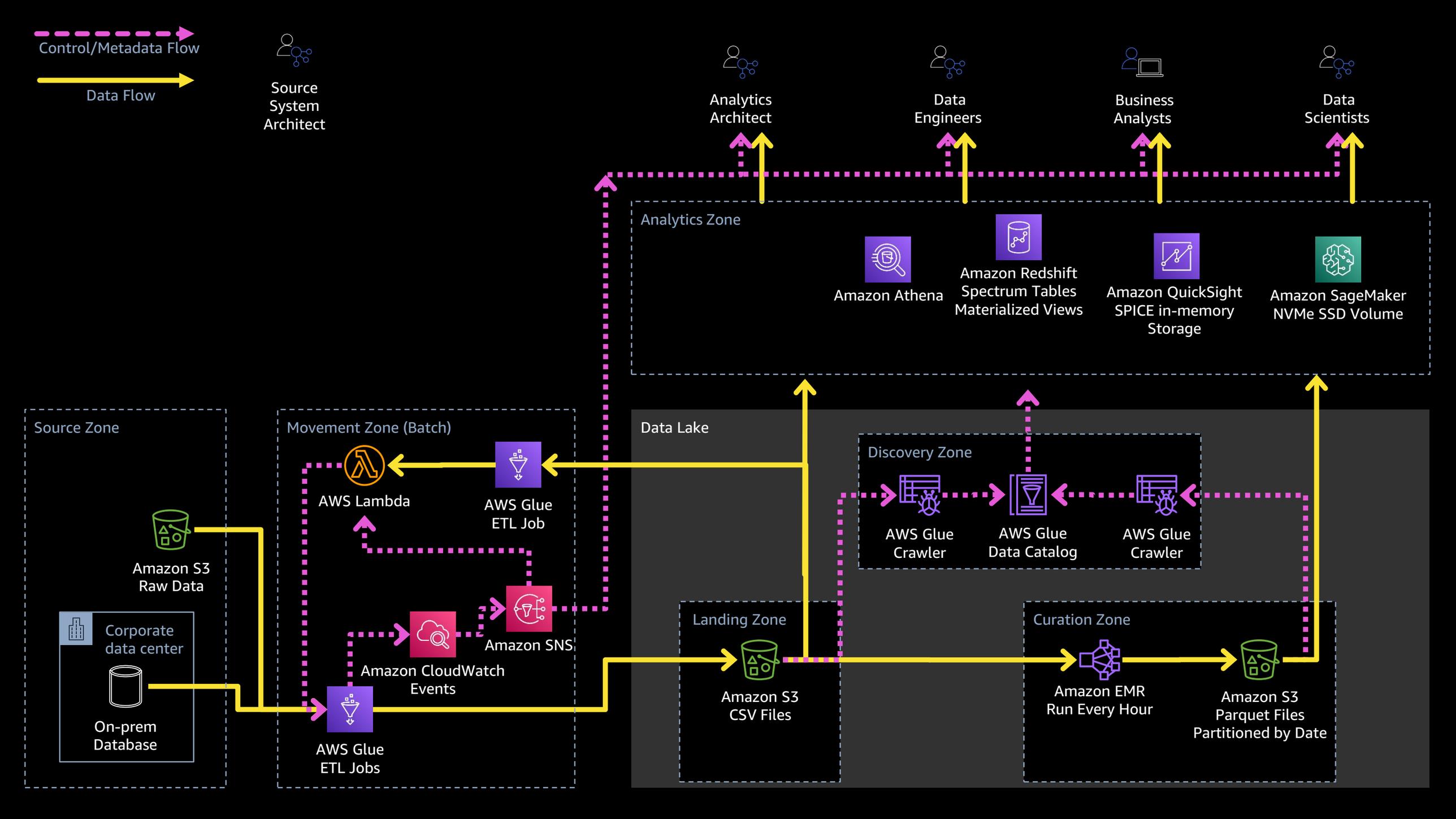
8. Choose the Optimal Compute Solution

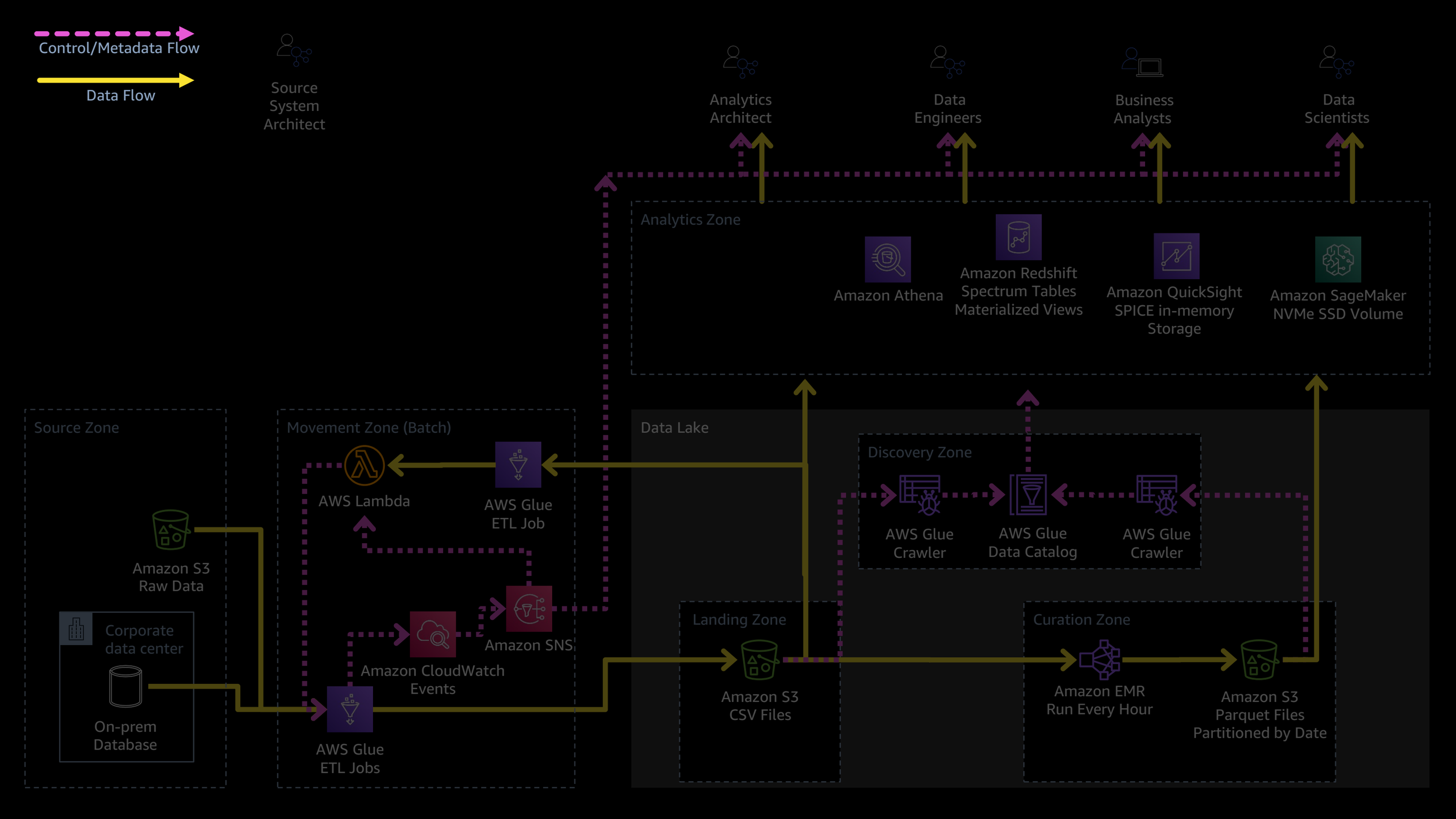
Ask yourself:

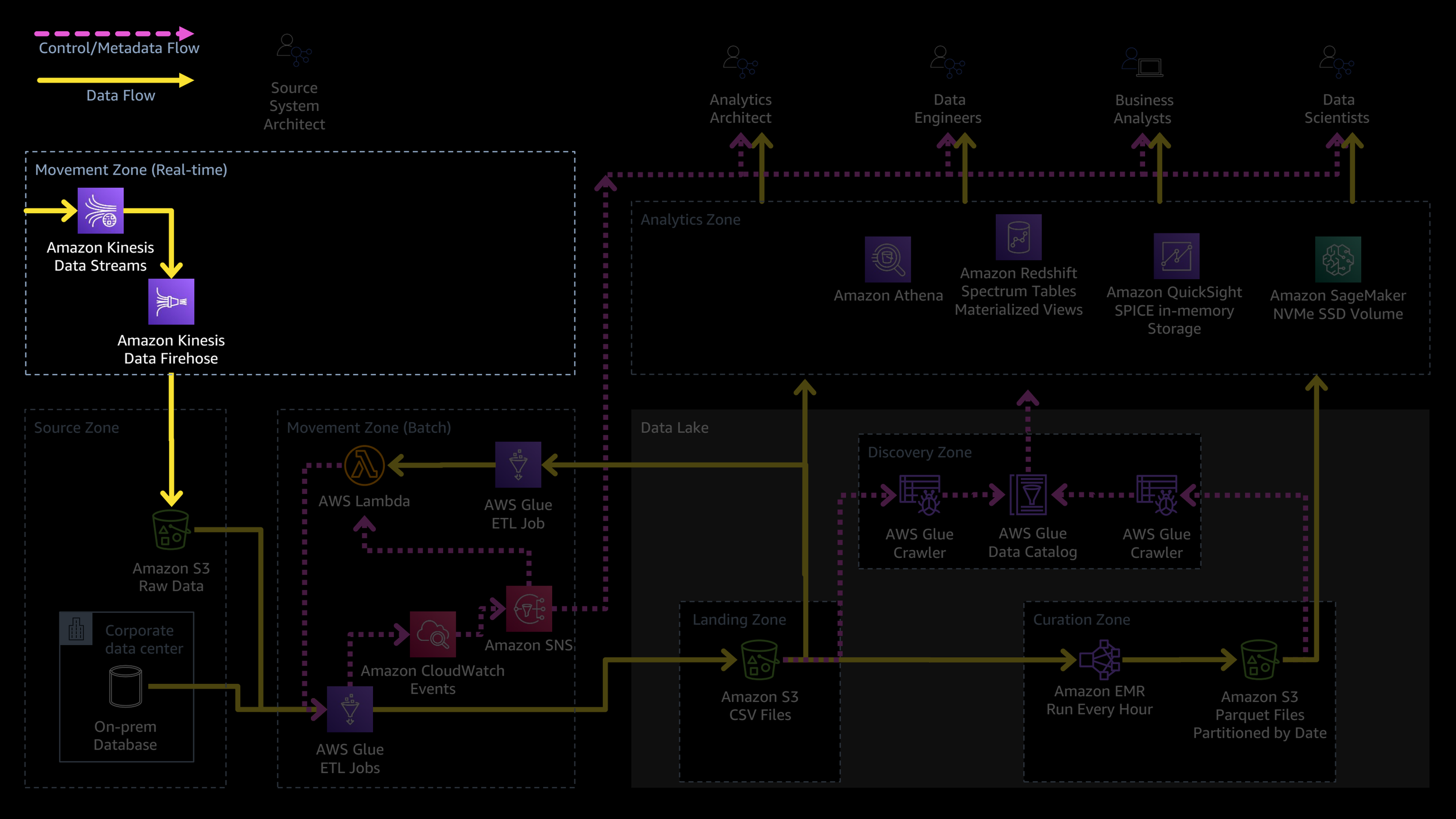
How do you select the best compute options for your analytics workload?

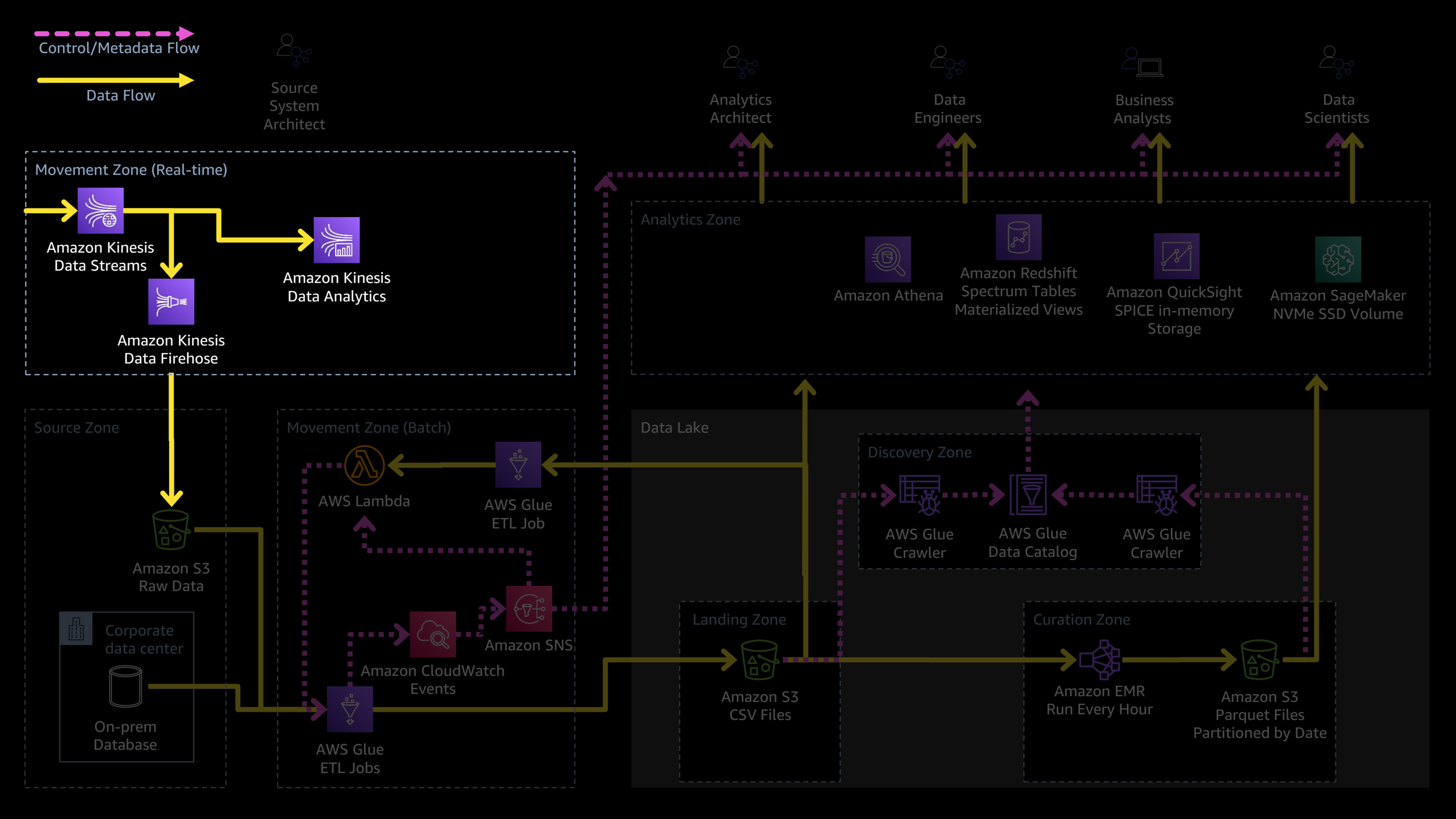
You should

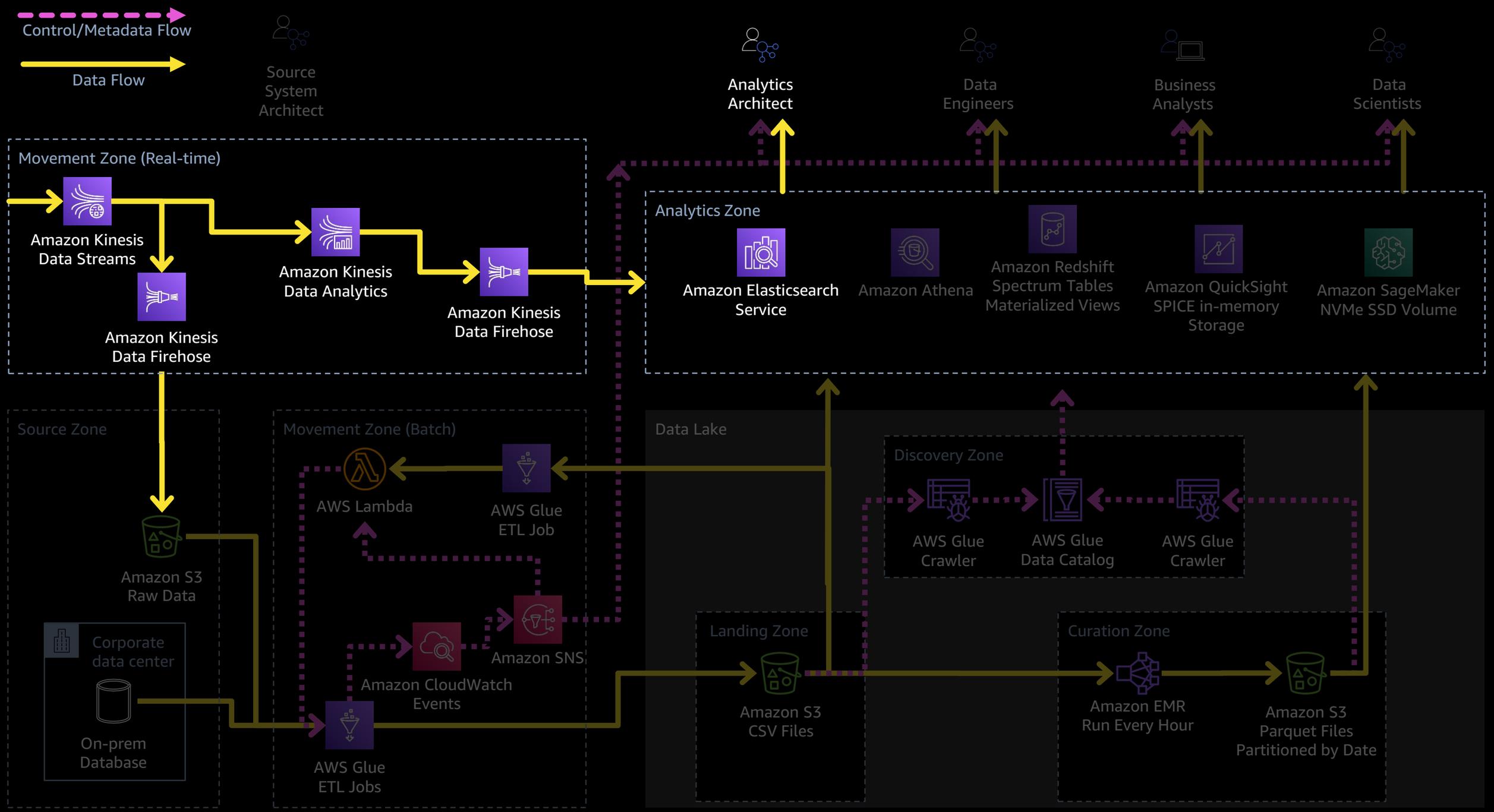
- *Follow the business requirements for ETL and analytics processing*
- *Evaluate the available ETL & analytics options*
- *Use the available elasticity of resources*
- *Re-evaluate compute needs based on metrics*
- *Use AWS-managed services for less operational burden*


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Analytics Zone
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs



Cost
Optimization

COST OPTIMIZATION:
HOW TO GOVERN COSTS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

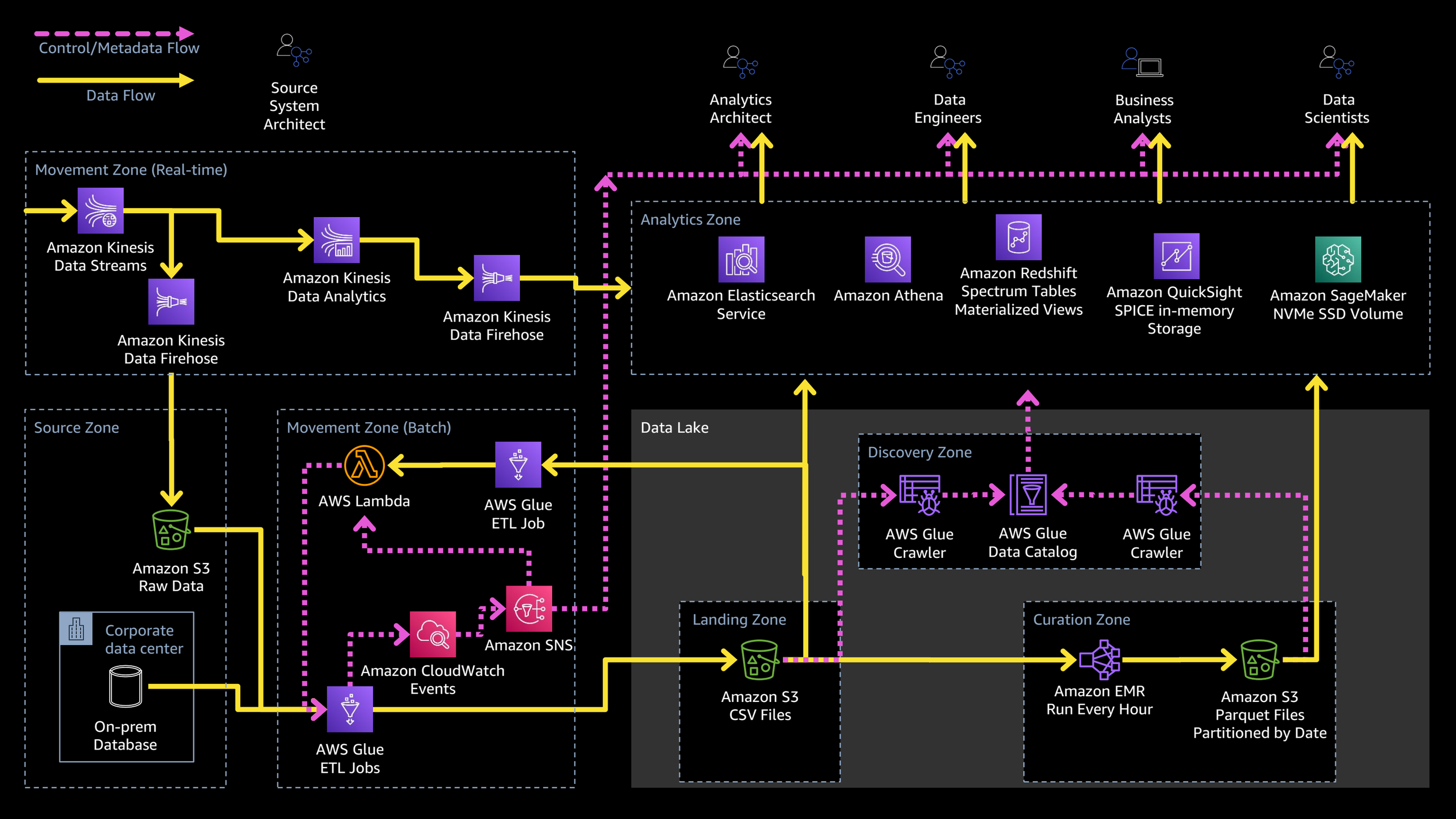
9. Delete Out-of-Use Data

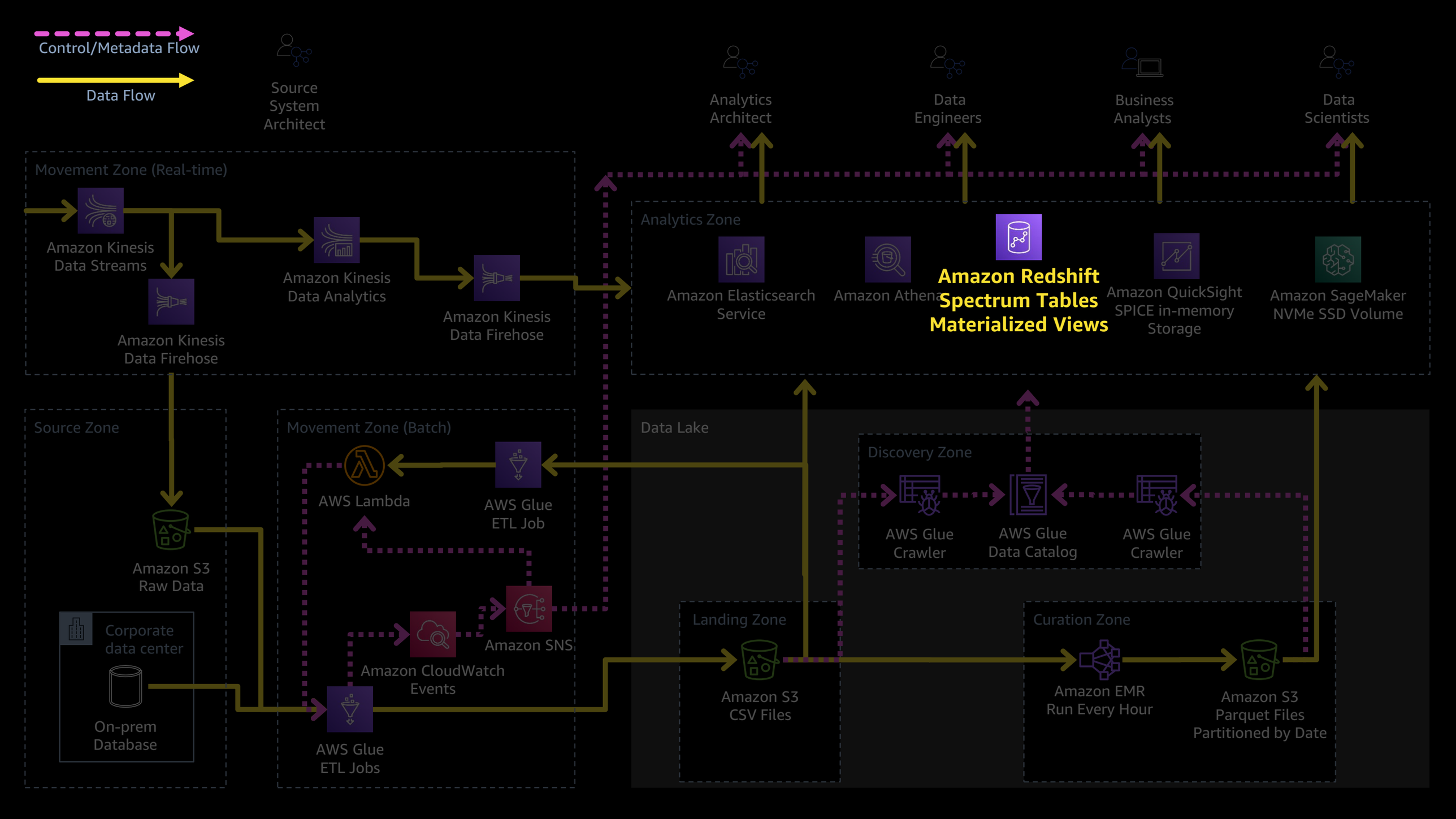
Ask yourself:

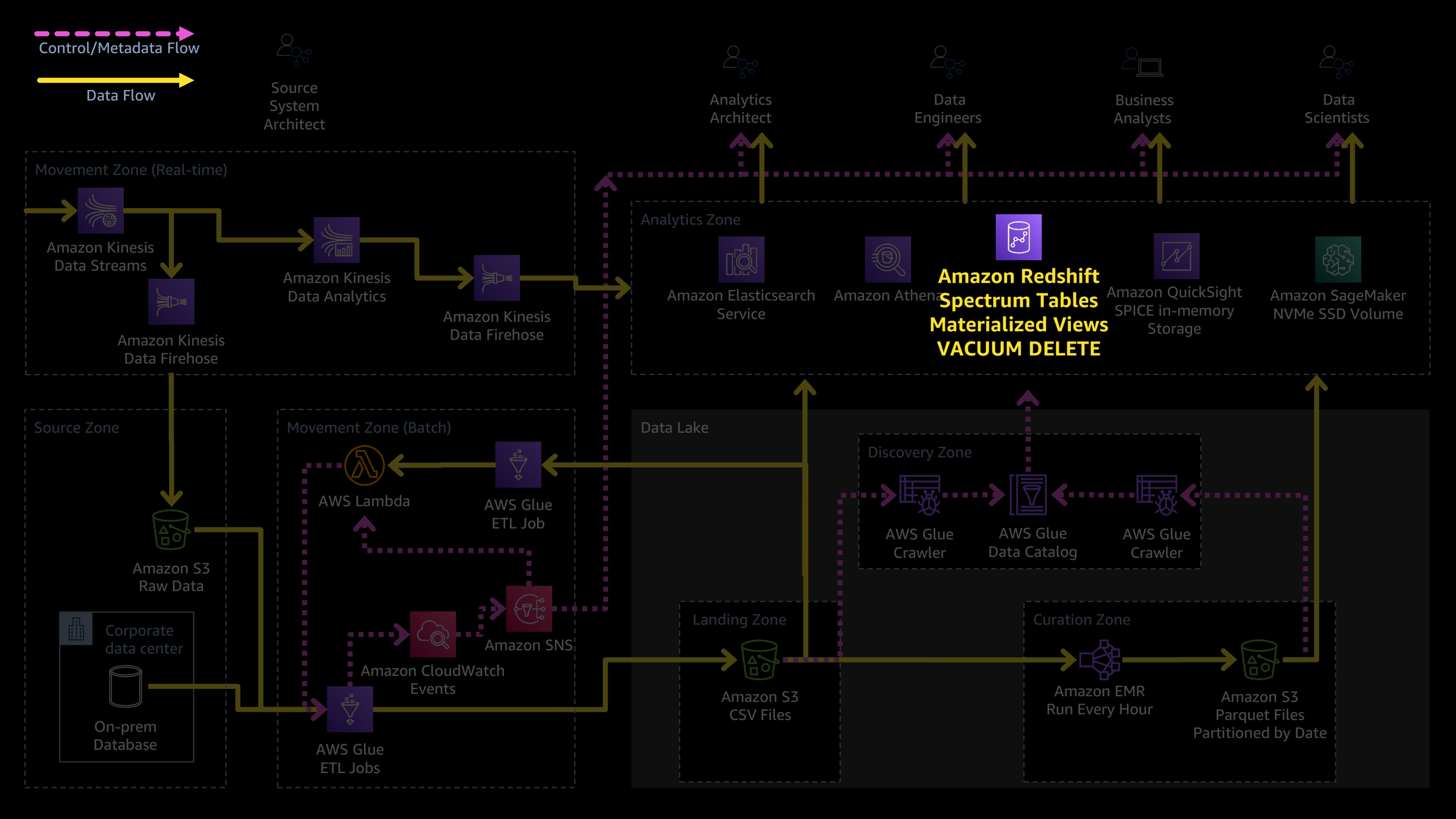
How do you decommission data from your analytics workload?

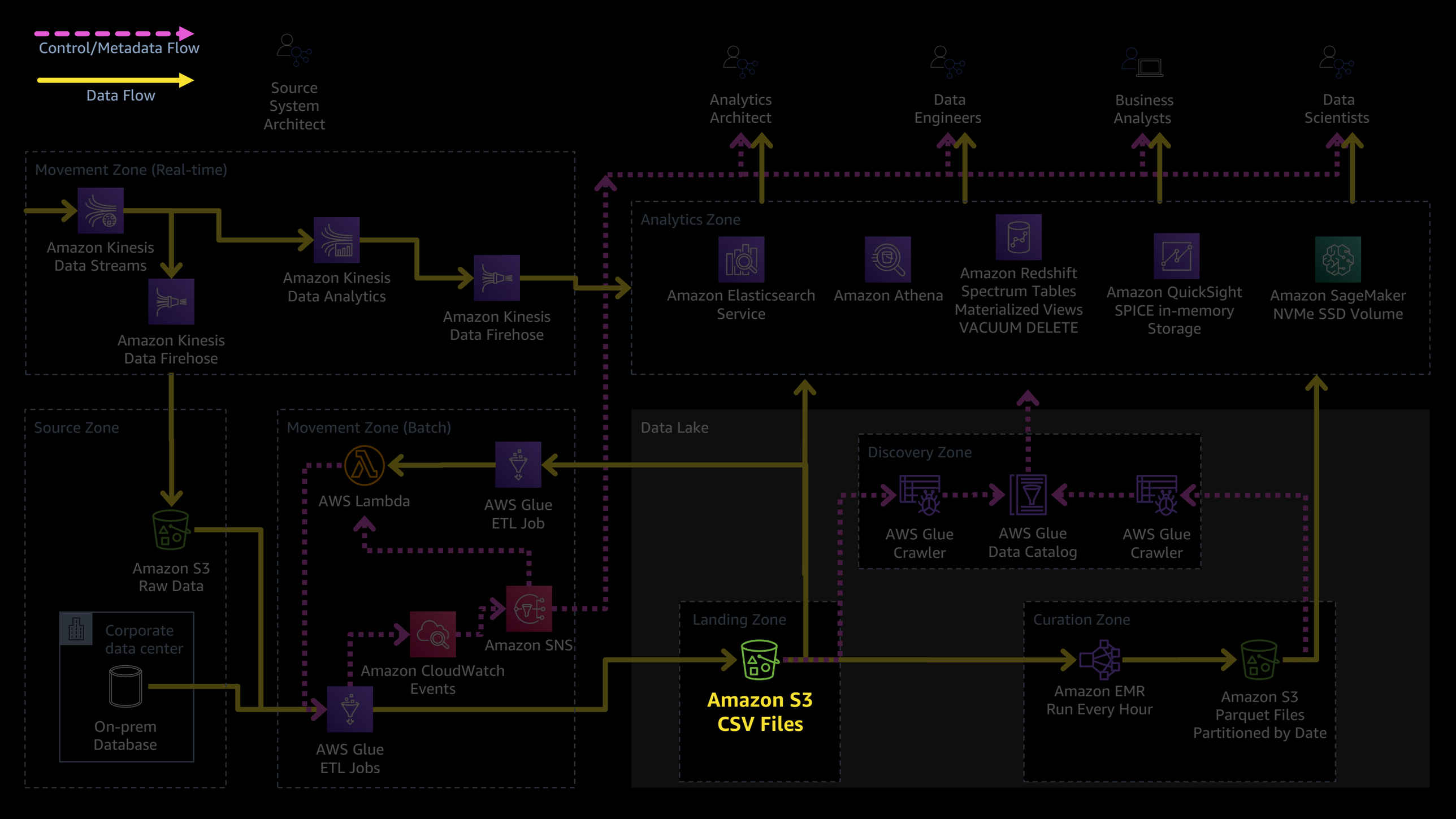
You should

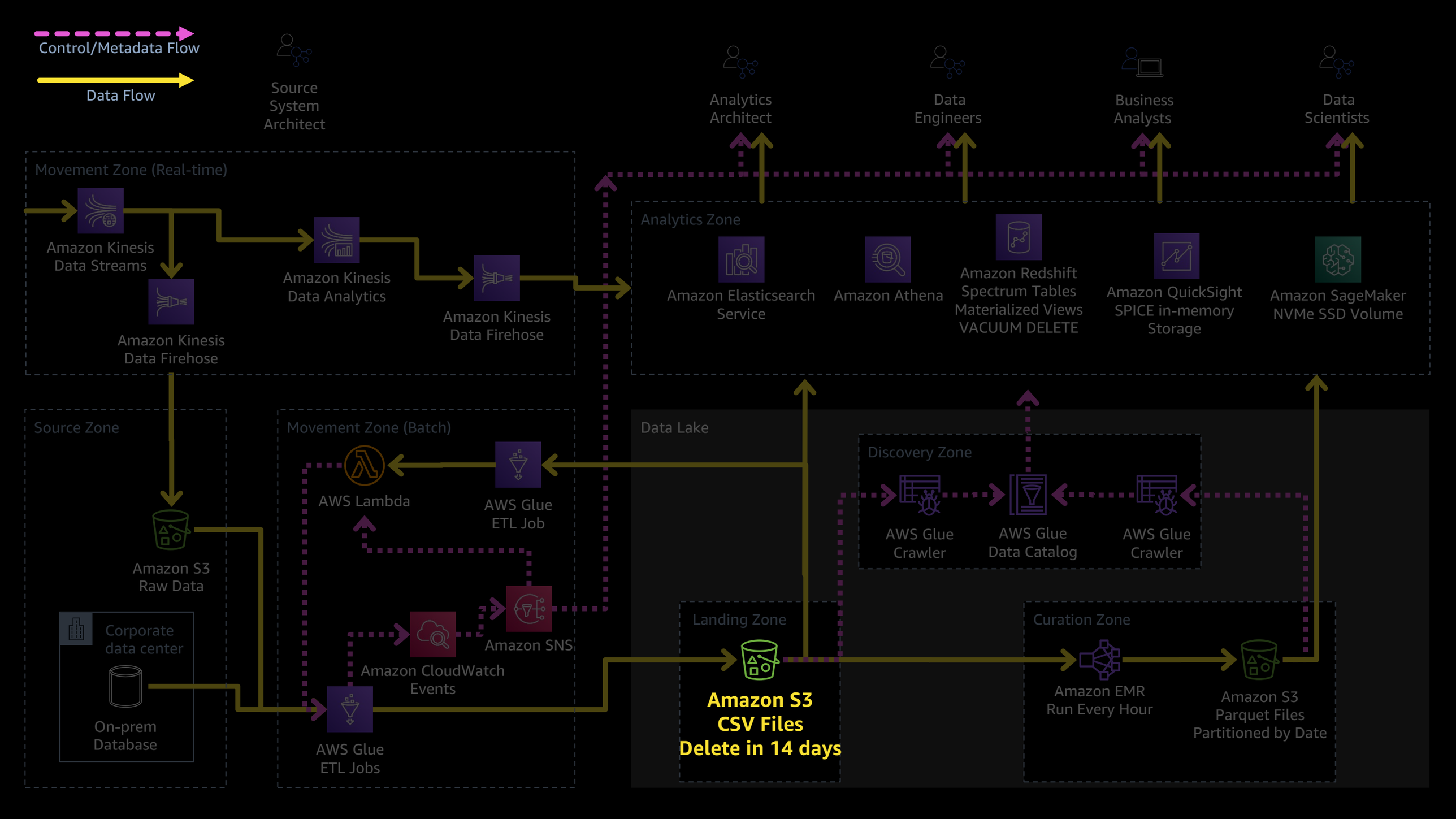
- *Follow the data retention requirements of the business*
- *Delete data that is out of the retention period*
- *Delete intermittent-processing data that can be removed promptly*
- *Automate data deletion*
- *Notify people and machines who depend on the data when data changes or delays occur*

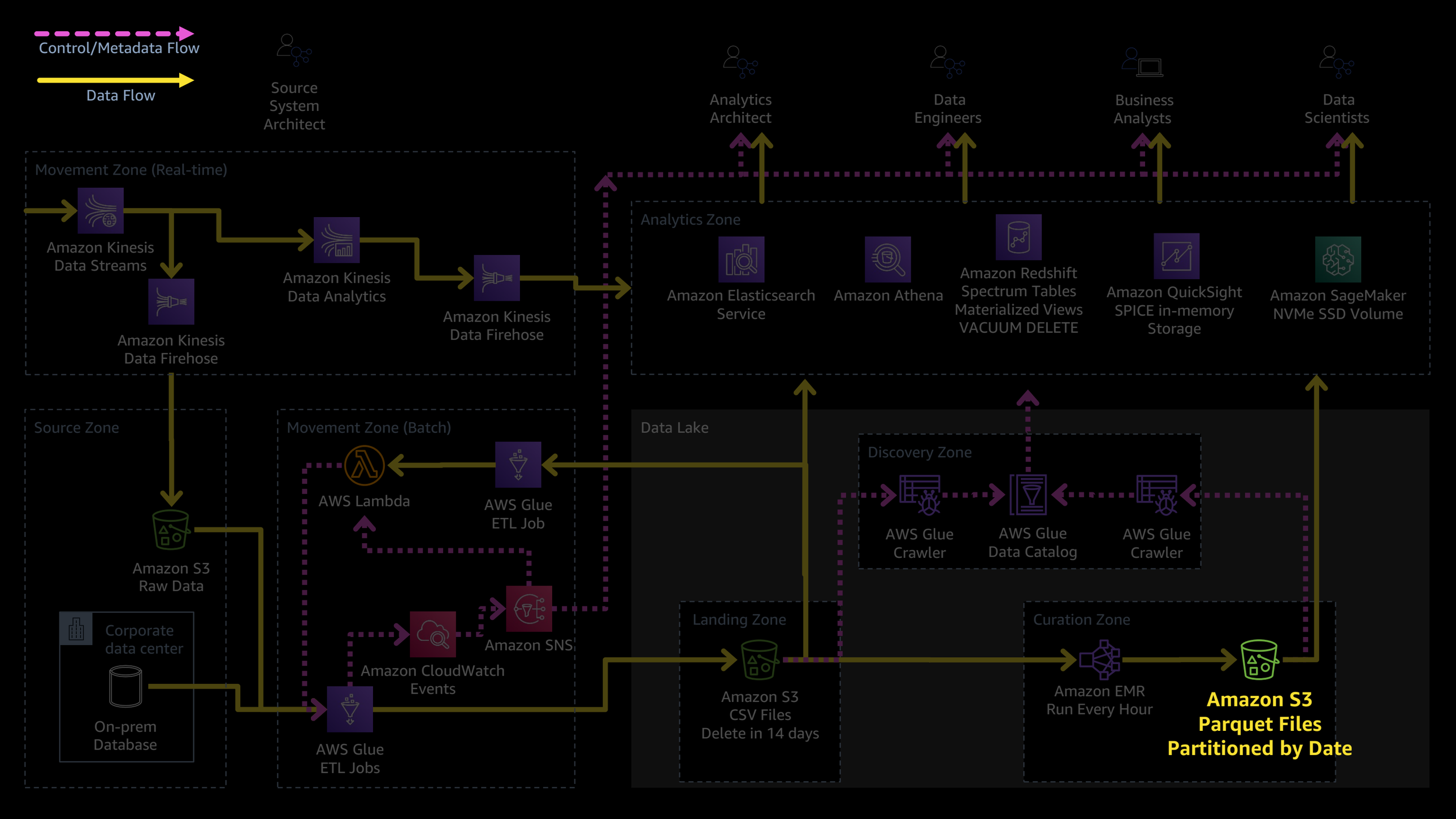

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs

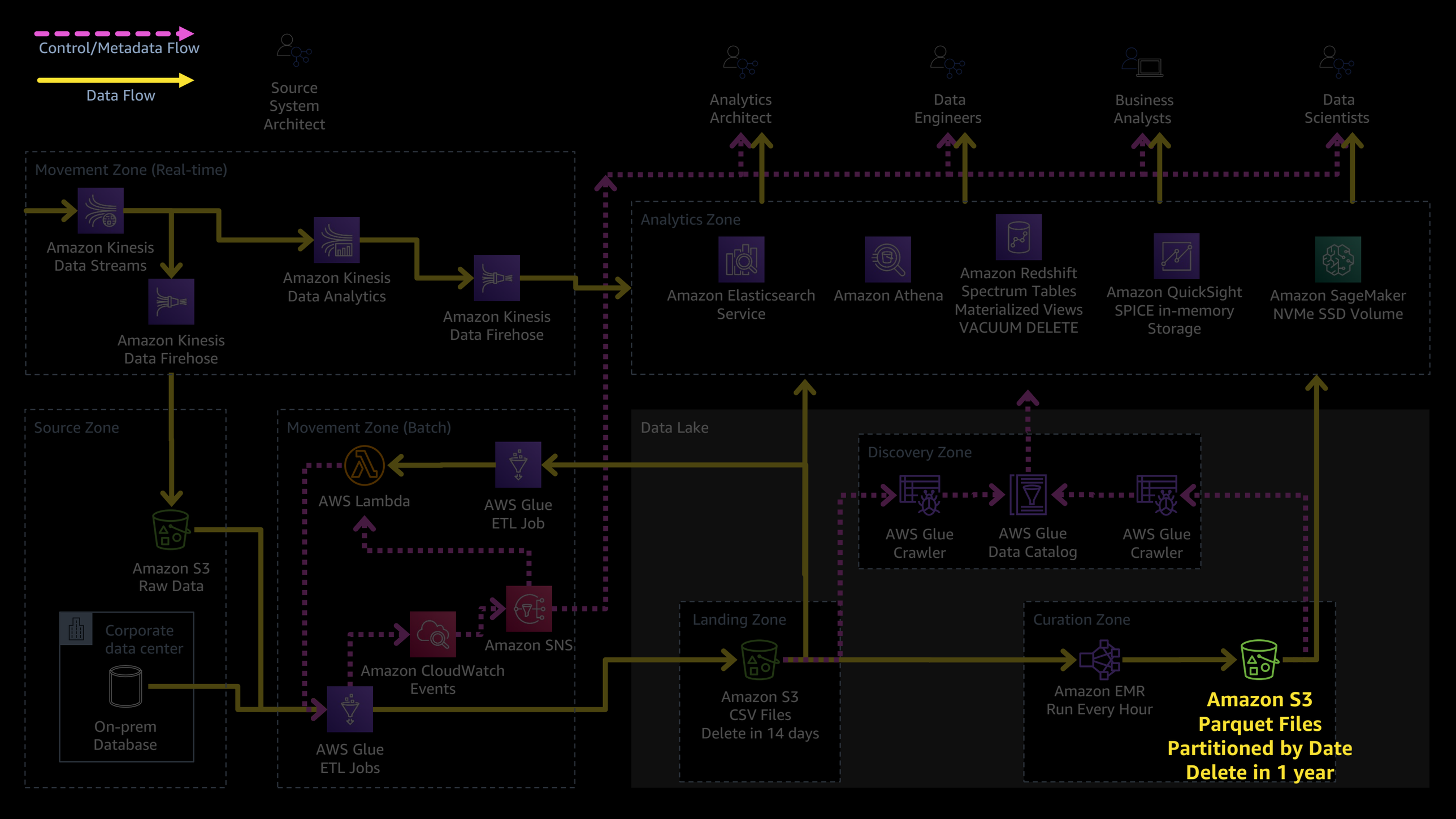

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs

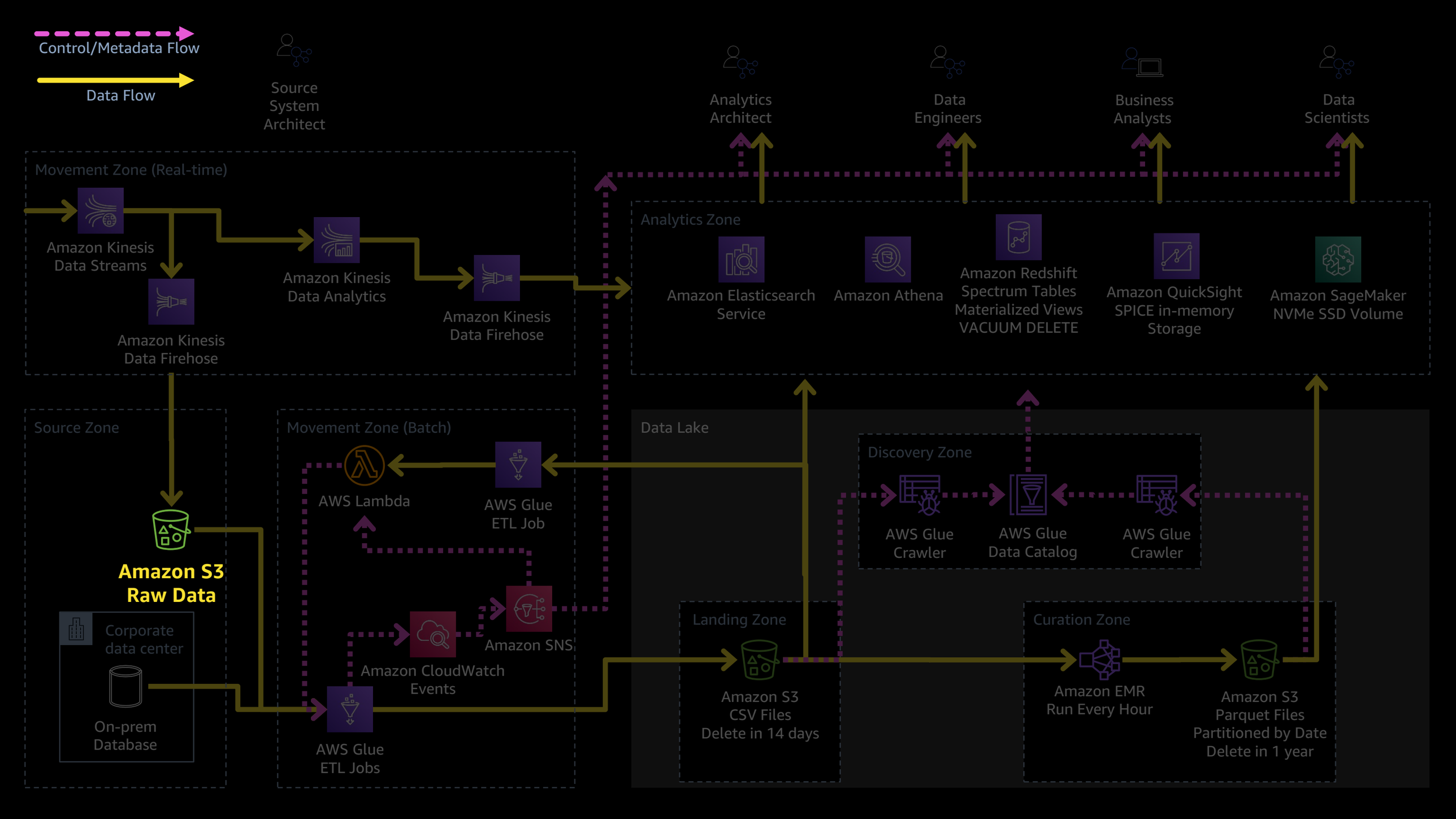

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs

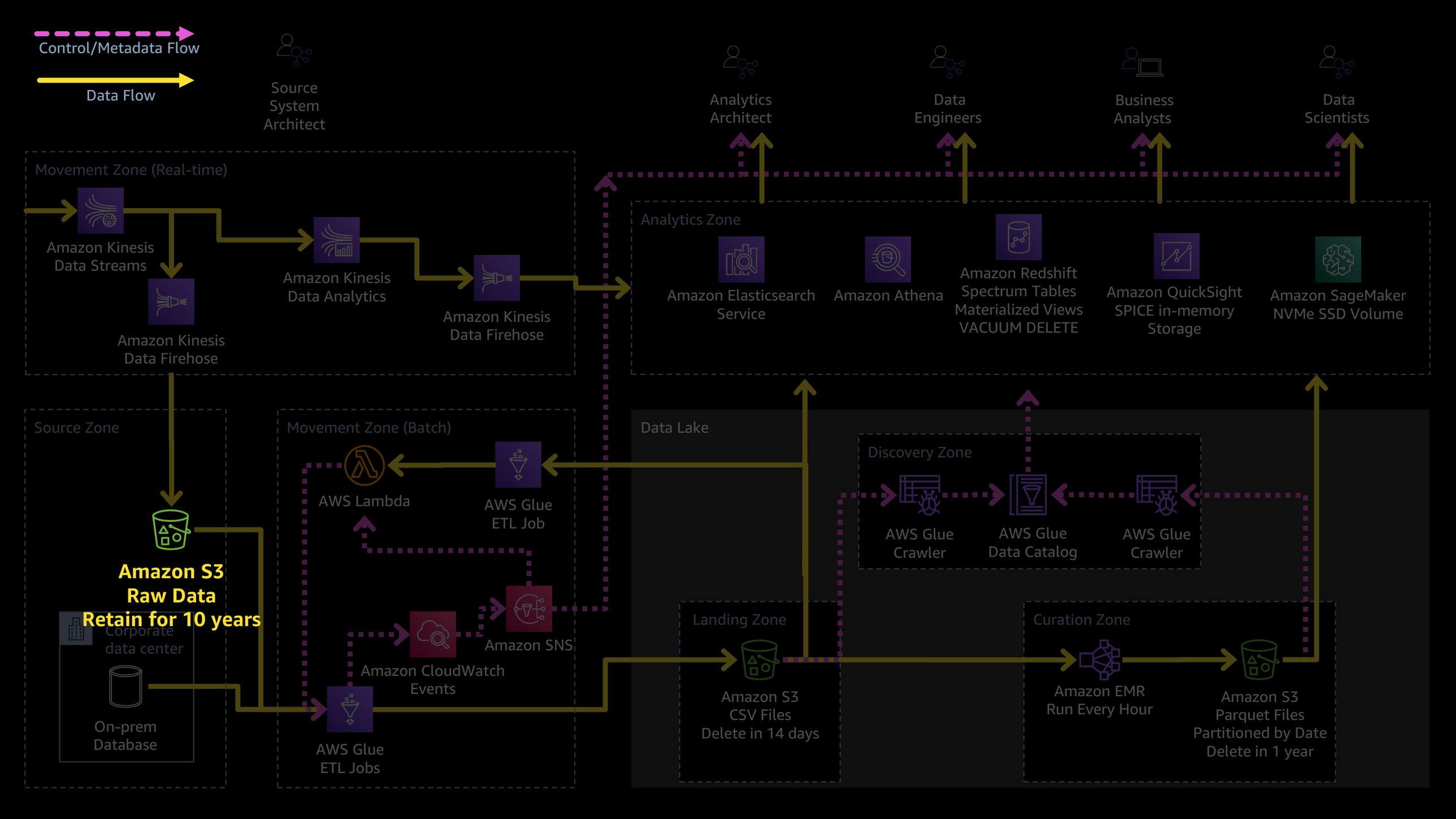

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views VACUUM DELETE
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views VACUUM DELETE
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
Amazon S3 Raw Data
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
AWS Glue ETL Jobs
Amazon S3 CSV Files Delete in 14 days
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views VACUUM DELETE
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files Delete in 14 days
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views VACUUM DELETE
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files Delete in 14 days
Amazon EMR Run Every Hour
AWS Glue ETL Jobs
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views VACUUM DELETE
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
AWS Glue ETL Jobs
Amazon S3 CSV Files Delete in 14 days
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views VACUUM DELETE
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
AWS Glue ETL Jobs
Landing Zone
Amazon S3 CSV Files Delete in 14 days
Curation Zone
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year

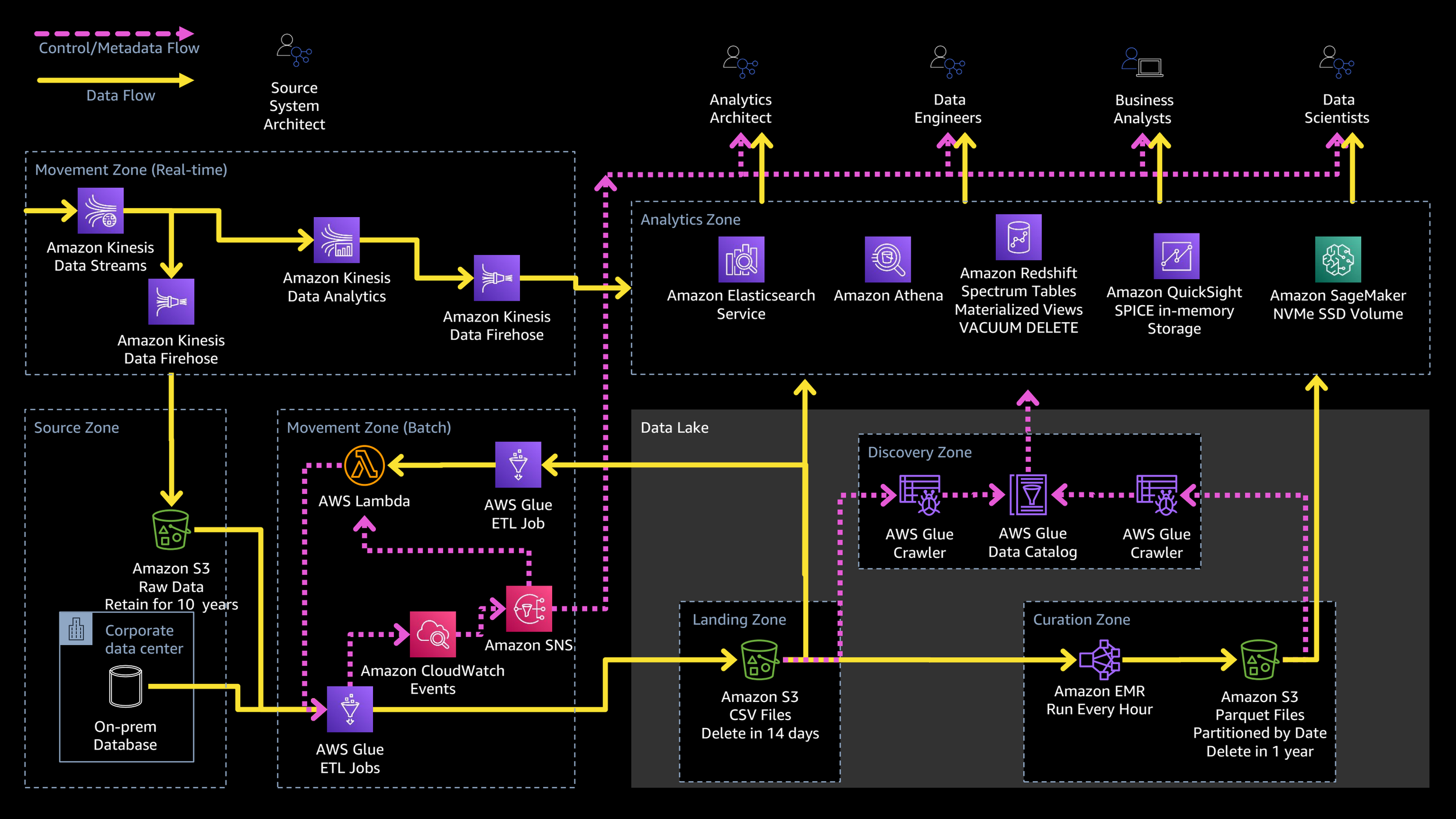
10. Choose the Lowest-Cost Storage Solution

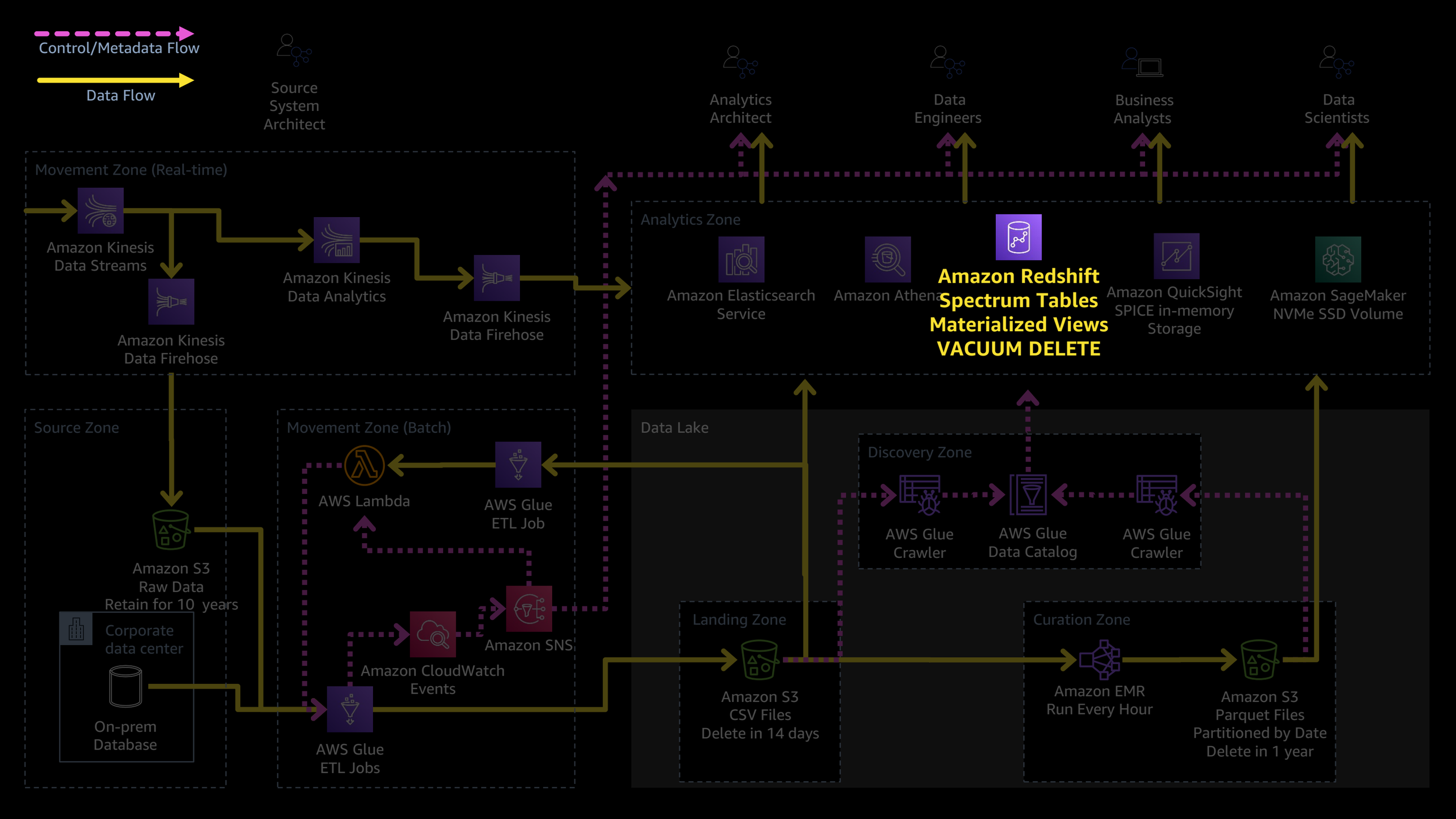
Ask yourself:

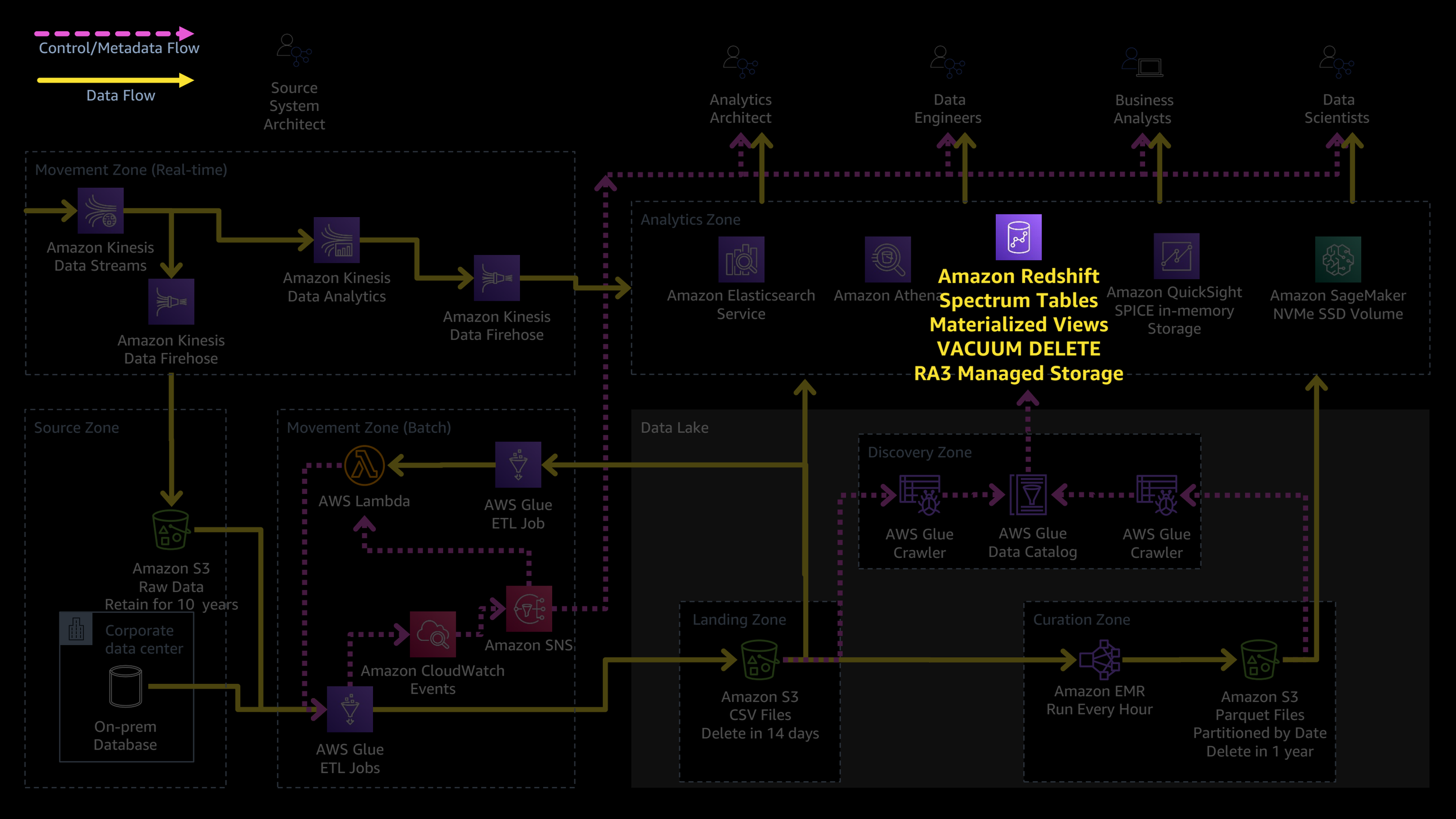
How do you select the lowest-cost storage for your analytics workload?

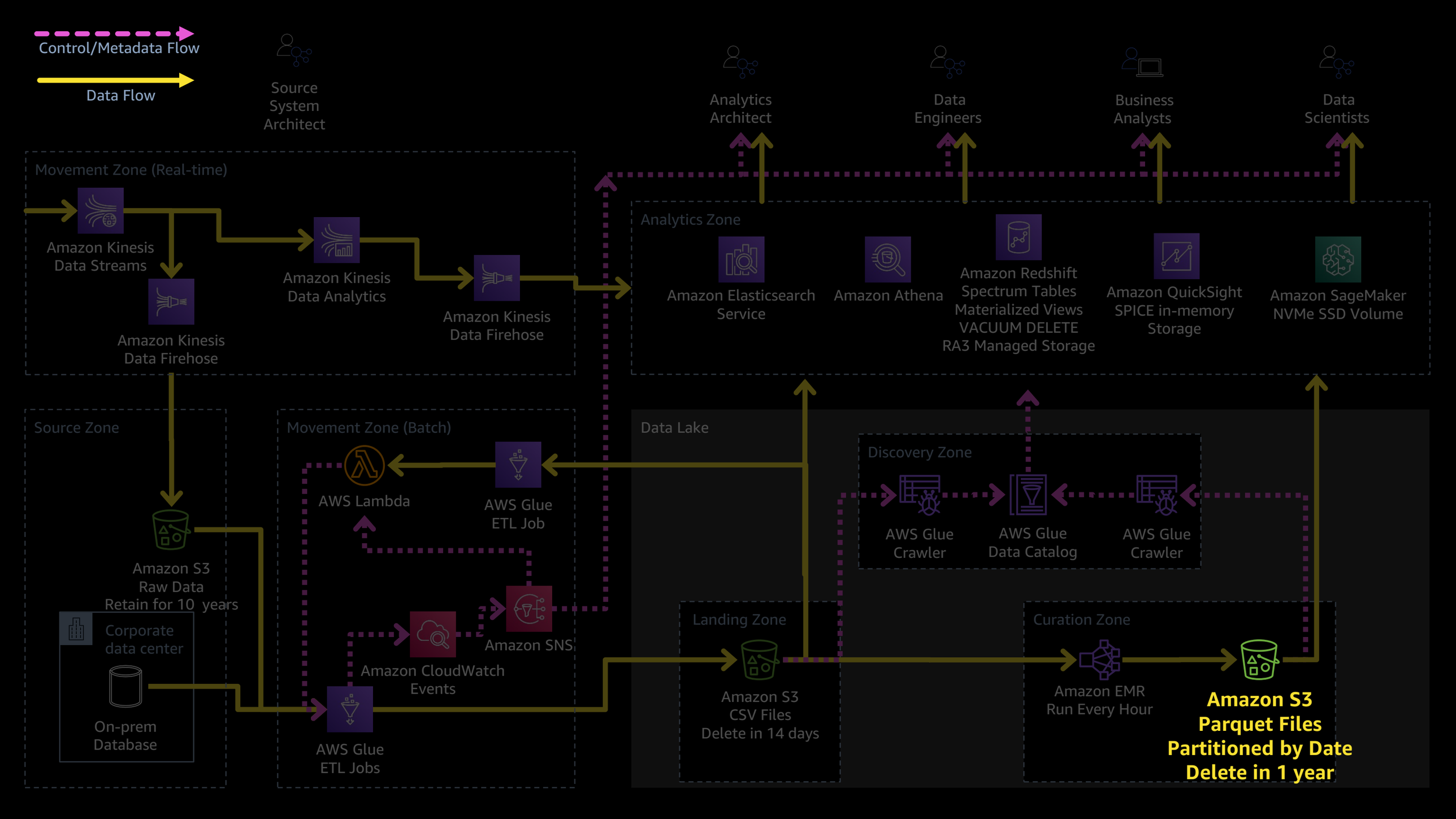
You should

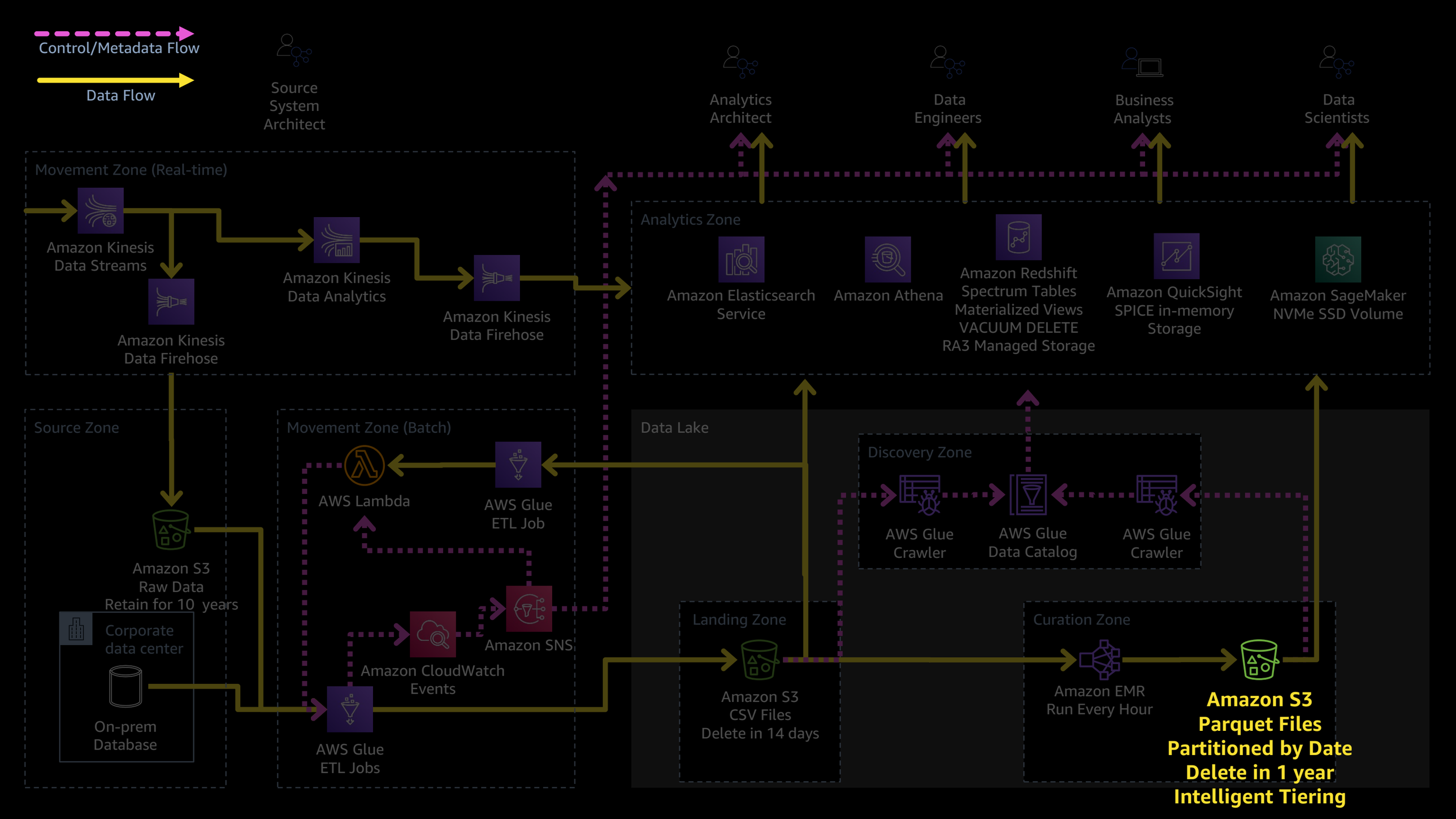
- *Decouple storage from compute*
- *Gather data read and write patterns*
- *Perform cost modeling*
- *Automate metrics-based storage configuration changes*

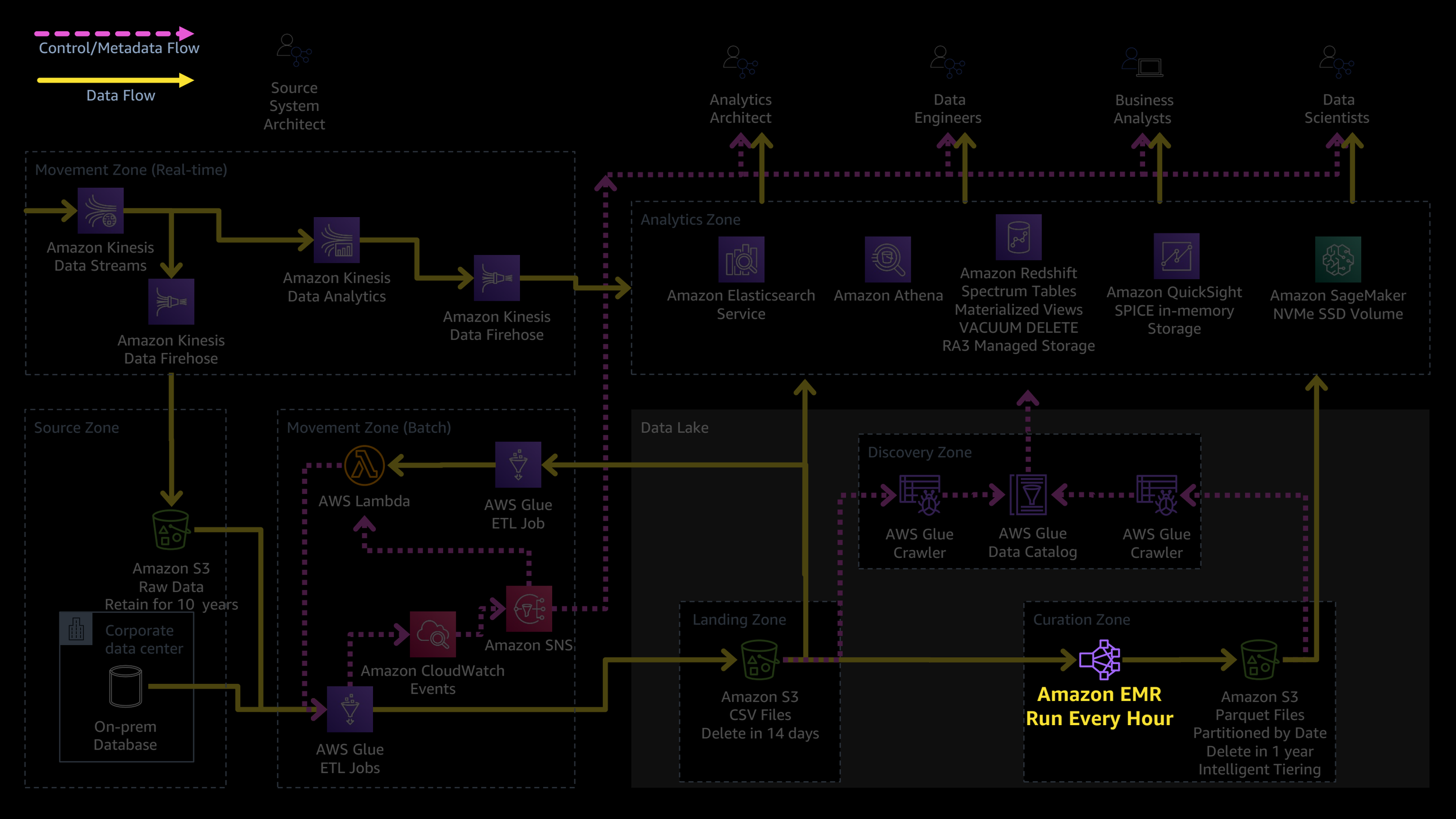

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views VACUUM DELETE
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
AWS Glue ETL Jobs
Amazon S3 CSV Files Delete in 14 days
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year

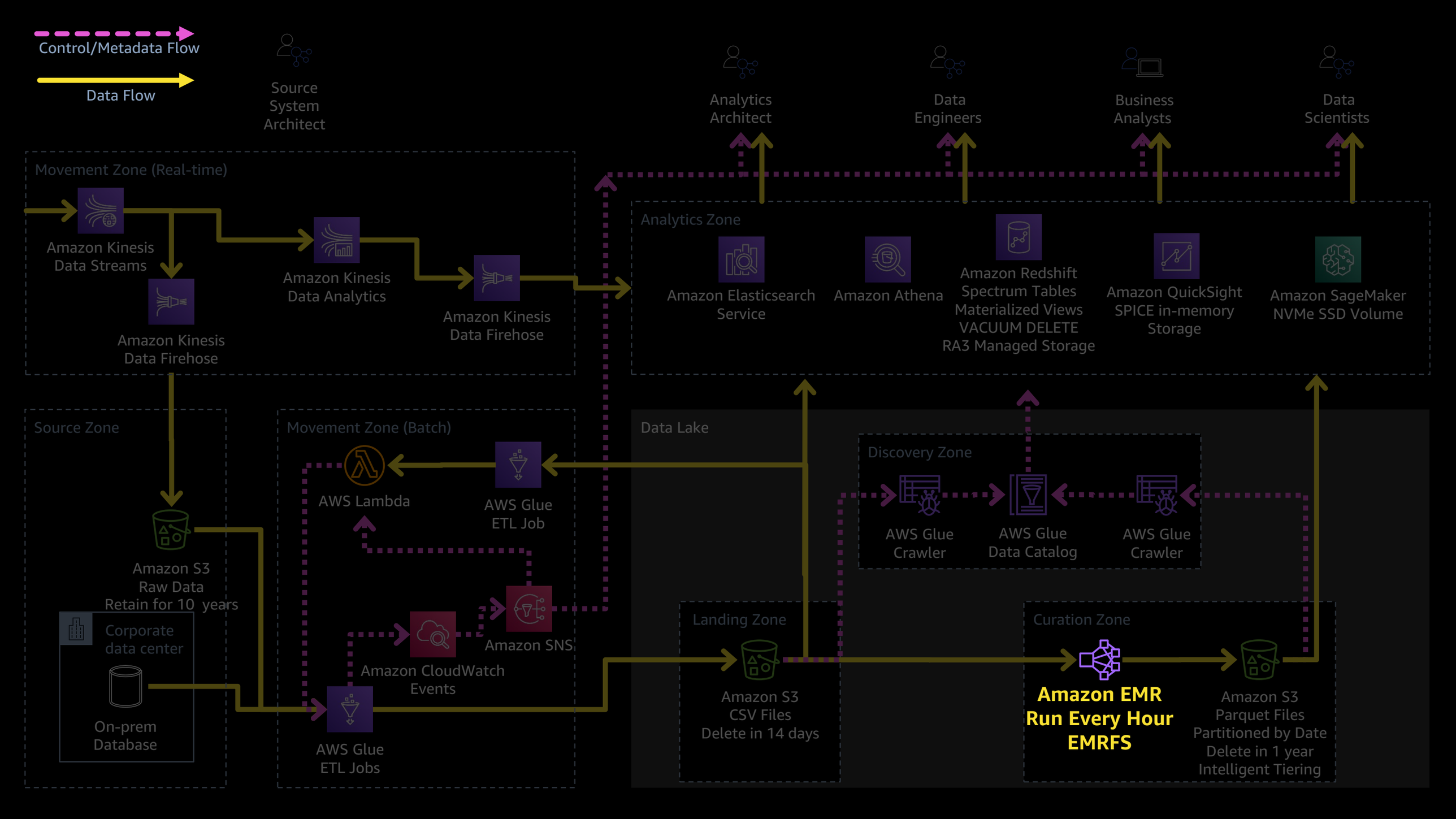

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files Delete in 14 days
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
AWS Glue ETL Jobs
Amazon S3 CSV Files Delete in 14 days
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views VACUUM DELETE RA3 Managed Storage
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3 CSV Files Delete in 14 days
Amazon EMR Run Every Hour
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Amazon QuickSight
SPICE in-memory Storage
Amazon SageMaker
NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3
Raw Data
Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3
CSV Files
Delete in 14 days
Amazon EMR
Run Every Hour
AWS Glue ETL Jobs
Amazon S3
Parquet Files
Partitioned by Date
Delete in 1 year
Intelligent Tiering


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Amazon QuickSight
SPICE in-memory Storage
Amazon SageMaker
NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3
Raw Data
Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
AWS Glue ETL Jobs
Amazon S3
CSV Files
Delete in 14 days
Amazon EMR
Run Every Hour
Amazon S3
Parquet Files
Partitioned by Date
Delete in 1 year
Intelligent Tiering


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift Spectrum Tables Materialized Views VACUUM DELETE RA3 Managed Storage
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
AWS Glue ETL Jobs
Amazon S3 CSV Files Delete in 14 days
Amazon EMR Run Every Hour EMRFS
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year Intelligent Tiering

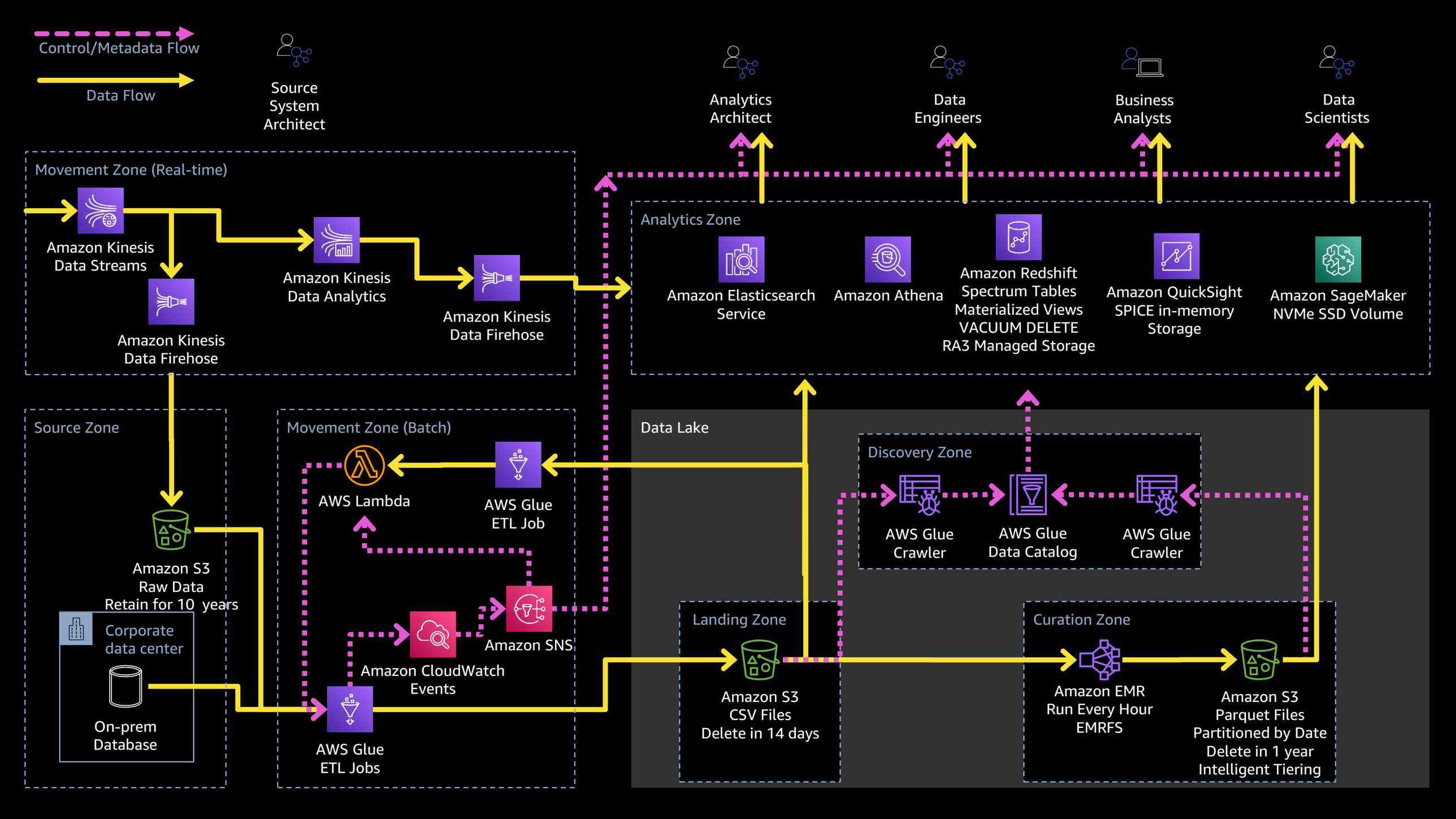
11. Choose the Lowest-Cost Compute Solution

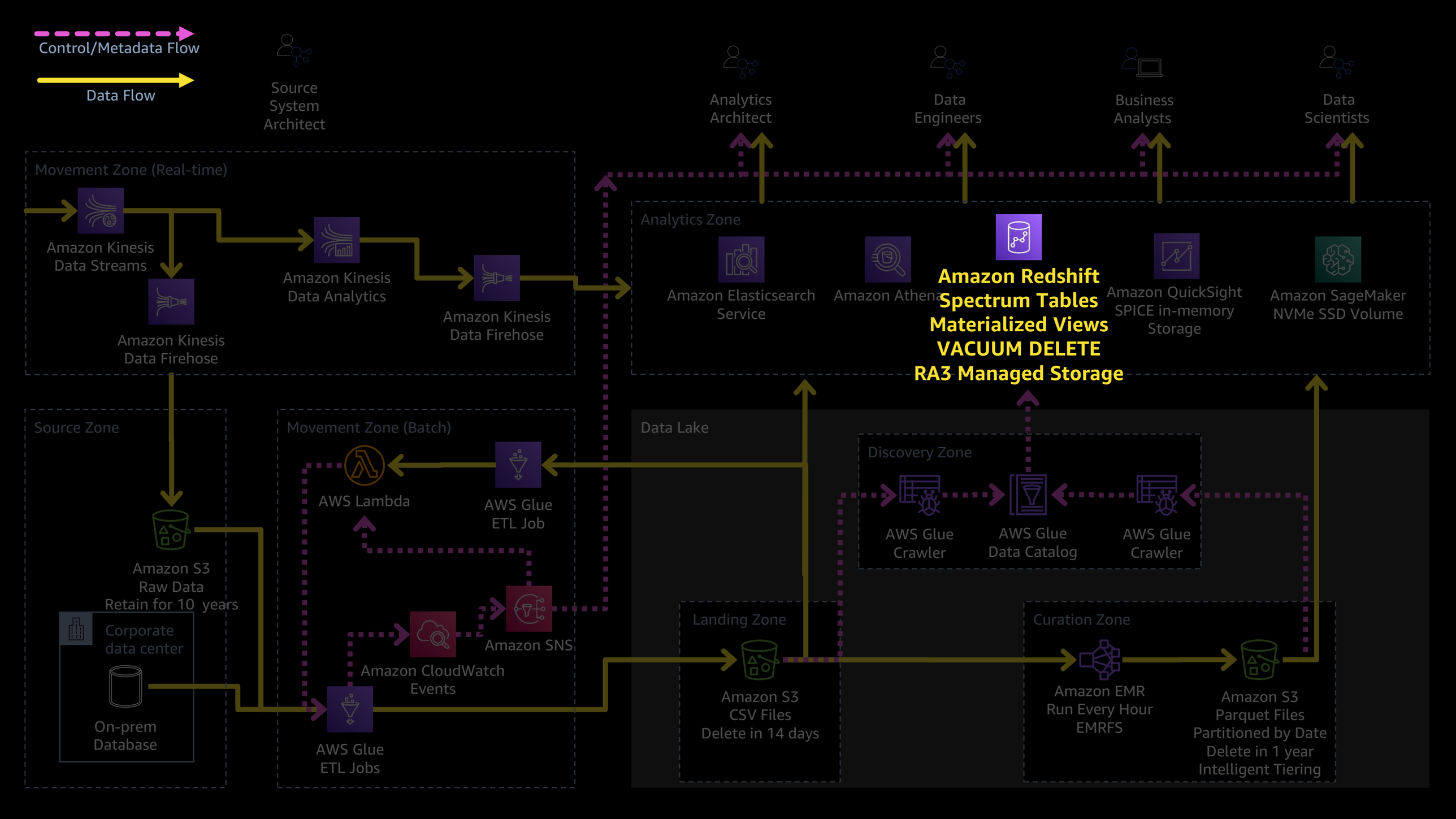
Ask yourself:

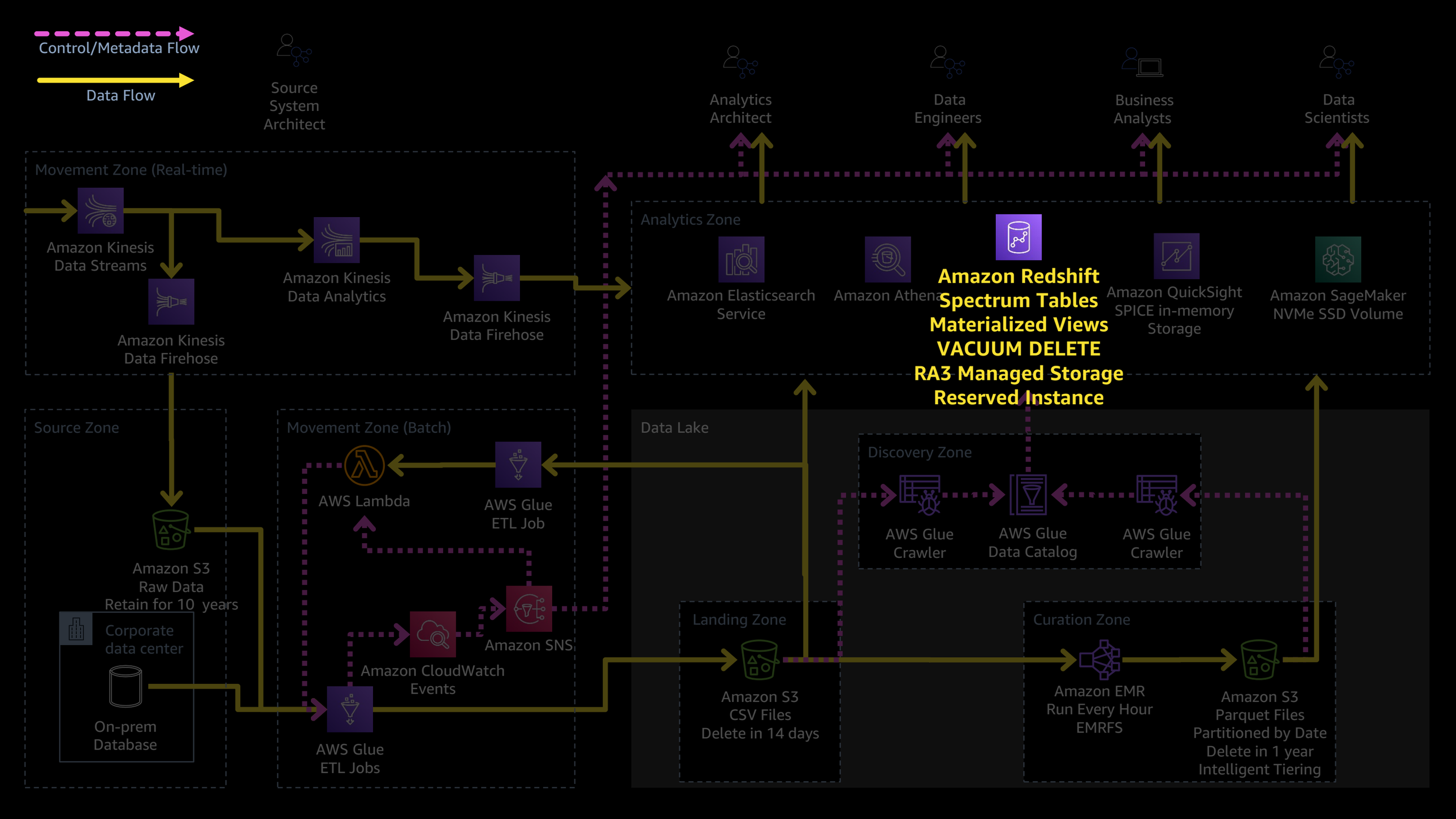
How do you select the lowest-cost compute for your analytics workload?

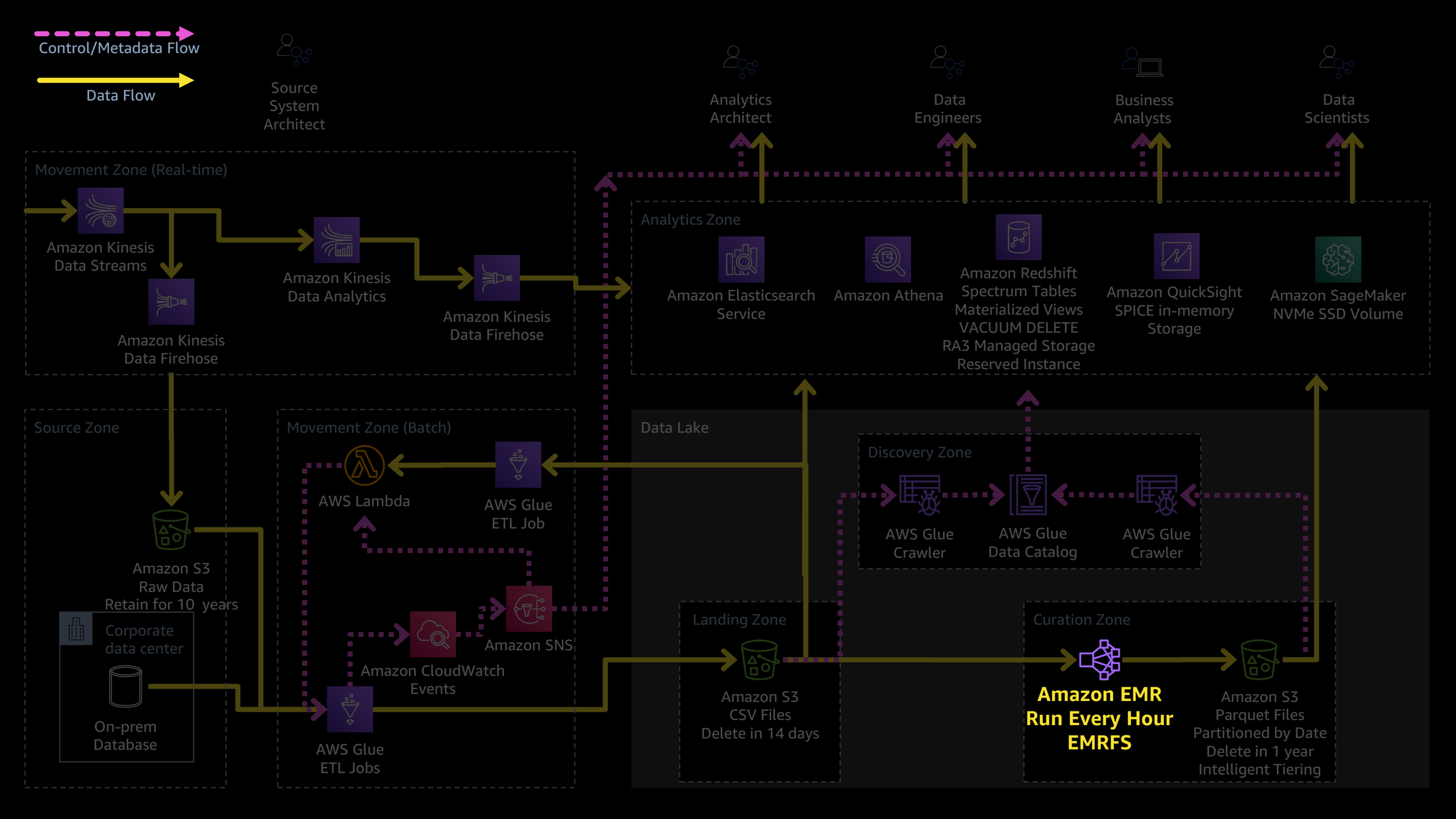
You should

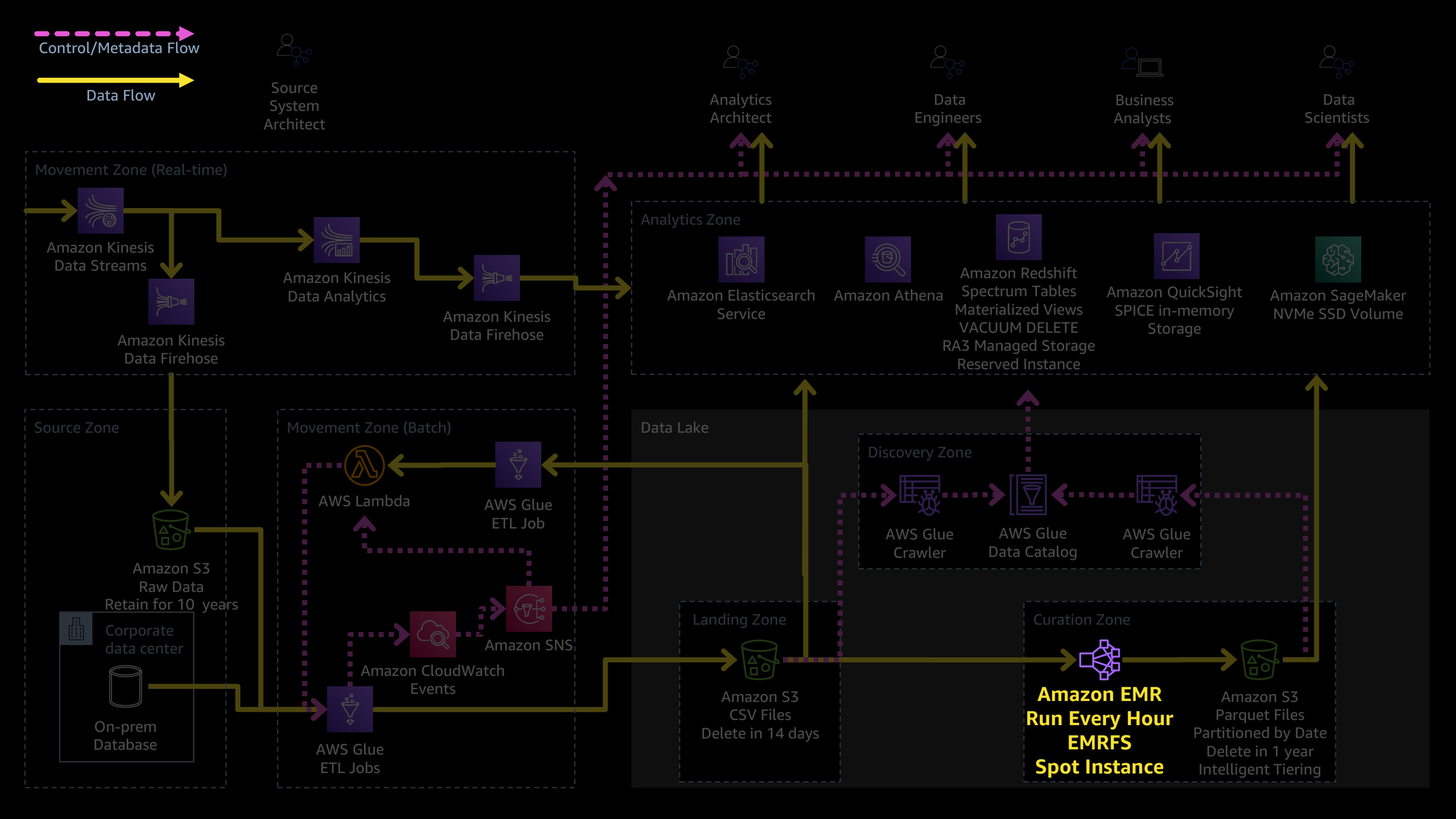
- *Perform cost modeling*
- *Automate metrics-driven resource type and size configurations*
- *Perform pricing model analysis*
- *Use the right pricing models such as On-demand, Reserved, or Spot Instances*

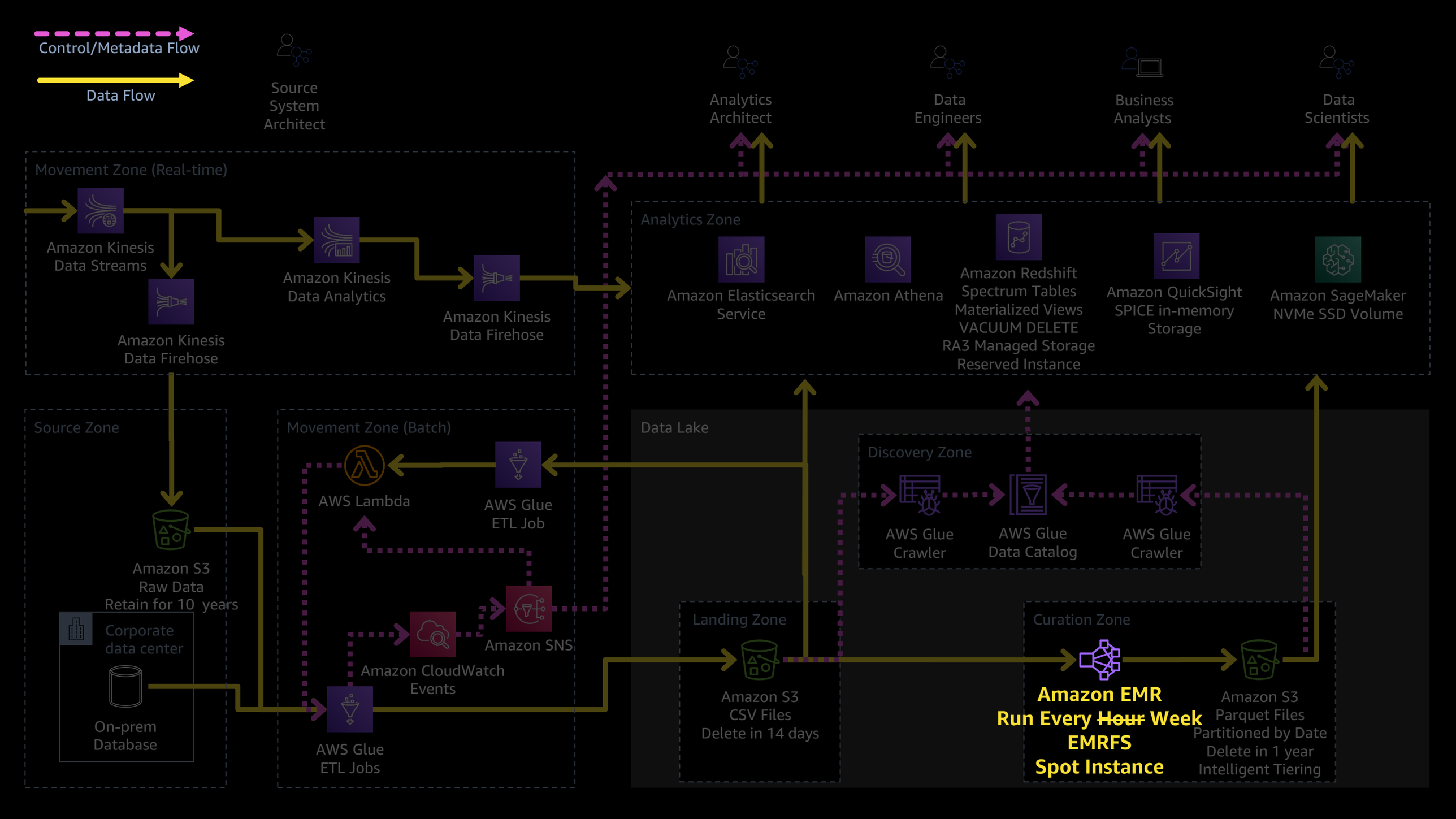

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Amazon QuickSight
SPICE in-memory Storage
Amazon SageMaker
NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3
Raw Data
Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
AWS Glue ETL Jobs
Landing Zone
Amazon S3
CSV Files
Delete in 14 days
Curation Zone
Amazon EMR
Run Every Hour
EMRFS
Amazon S3
Parquet Files
Partitioned by Date
Delete in 1 year
Intelligent Tiering

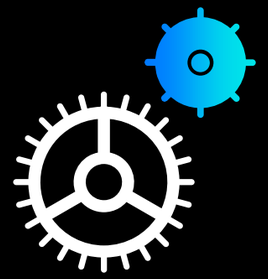

Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
AWS Glue ETL Jobs
Amazon S3 CSV Files Delete in 14 days
Amazon EMR Run Every Hour EMRFS
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year Intelligent Tiering


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Reserved Instance
Amazon QuickSight SPICE in-memory Storage
Amazon SageMaker NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
AWS Glue ETL Jobs
Landing Zone
Amazon S3 CSV Files Delete in 14 days
Curation Zone
Amazon EMR Run Every Hour EMRFS
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year Intelligent Tiering


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Reserved Instance
Amazon QuickSight
SPICE in-memory Storage
Amazon SageMaker
NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3
Raw Data
Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3
CSV Files
Delete in 14 days
Amazon EMR
Run Every Hour
EMRFS
Amazon S3
Parquet Files
Partitioned by Date
Delete in 1 year
Intelligent Tiering
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Reserved Instance
Amazon QuickSight
SPICE in-memory Storage
Amazon SageMaker
NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
AWS Glue ETL Jobs
Amazon S3 CSV Files Delete in 14 days
Amazon EMR
Run Every Hour
EMRFS
Spot Instance
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year Intelligent Tiering


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Reserved Instance
Amazon QuickSight
SPICE in-memory Storage
Amazon SageMaker
NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
AWS Glue ETL Jobs
Amazon S3 CSV Files Delete in 14 days
Amazon EMR
Run Every ~~Hour~~ Week
EMRFS
Spot Instance
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year Intelligent Tiering



Operational Excellence

OPERATIONAL EXCELLENCE: HOW TO MAINTAIN GOOD HEALTH

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

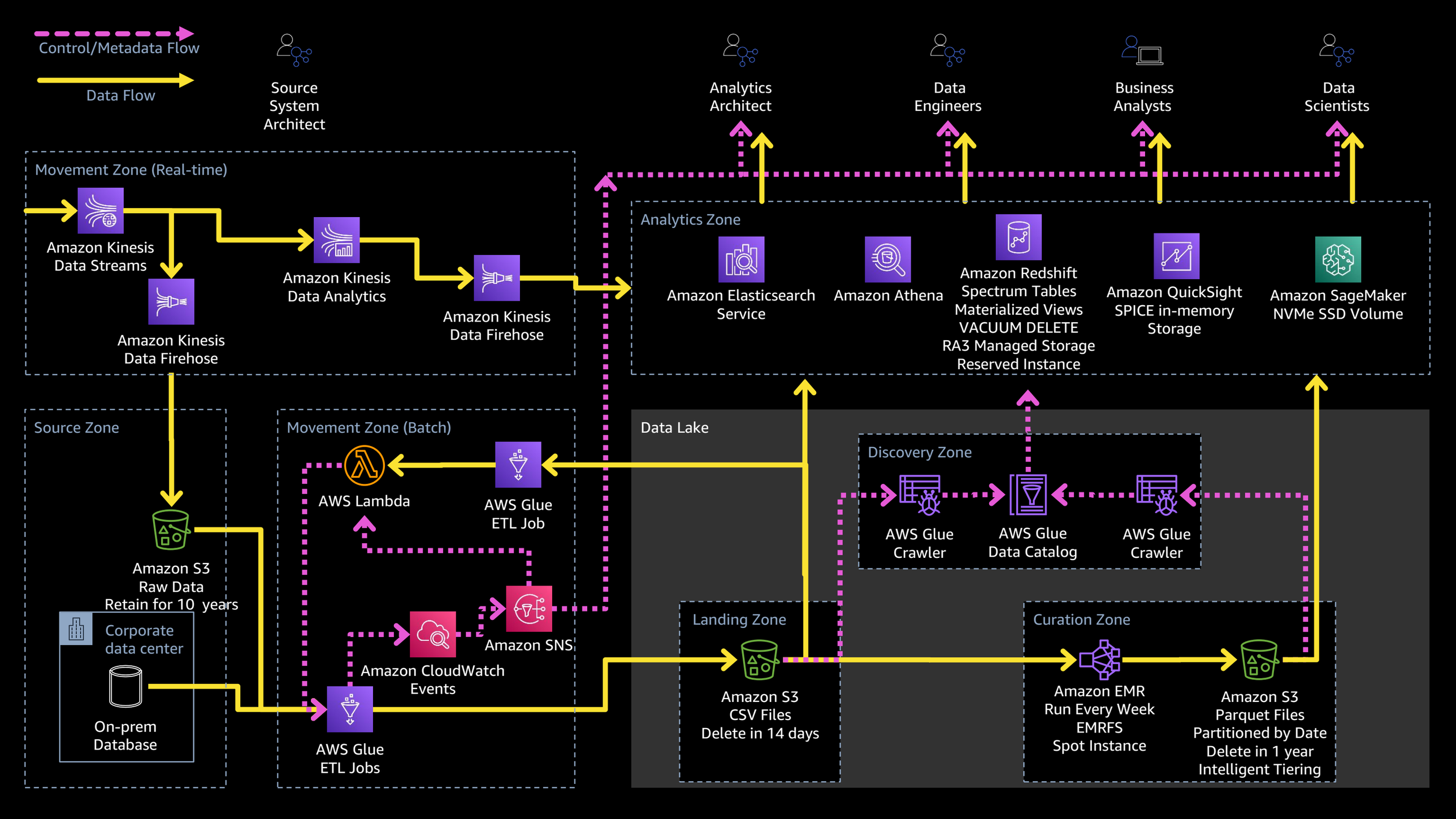
12. Monitor the Health of Analytics Pipelines

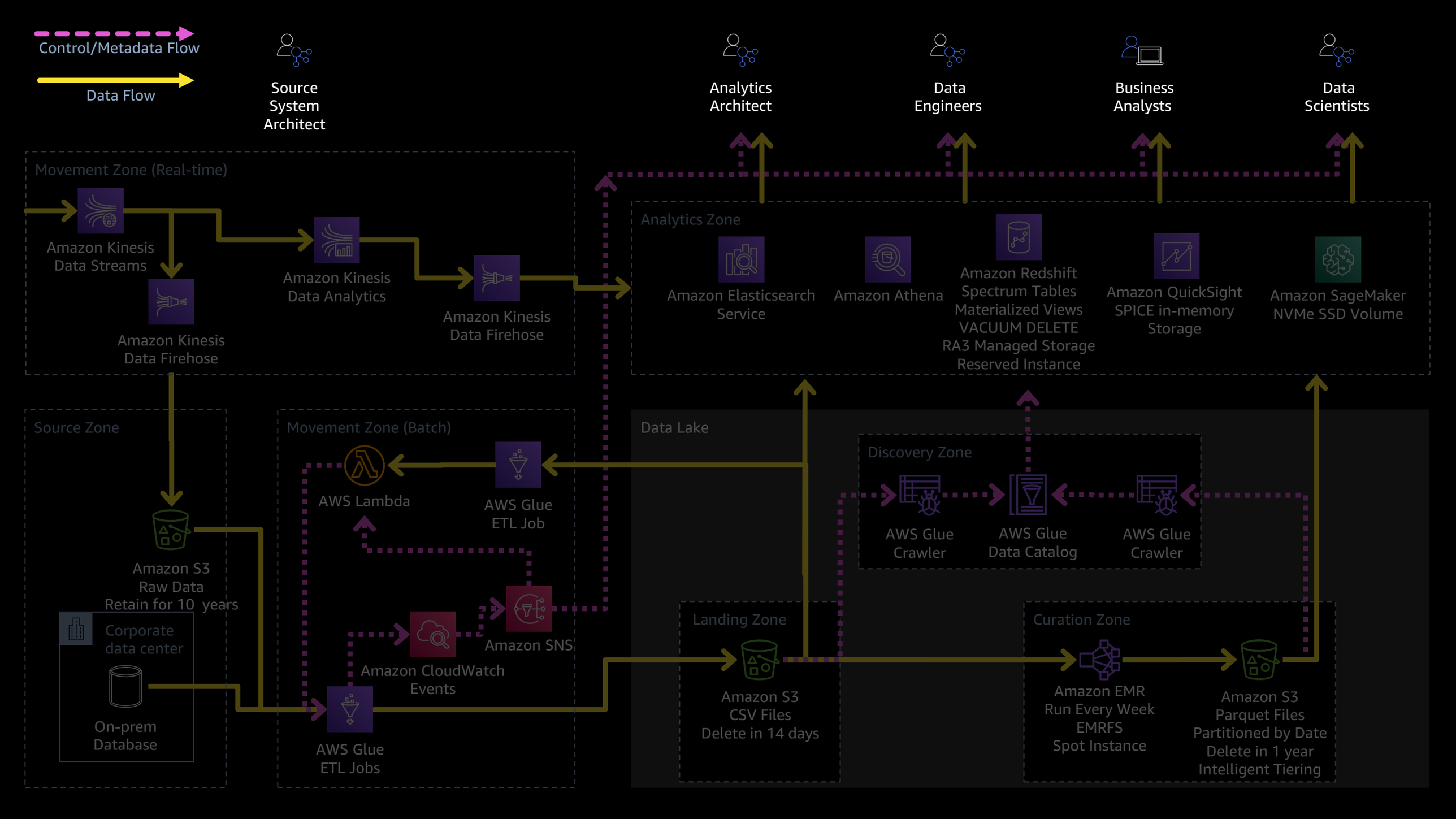
Ask yourself:

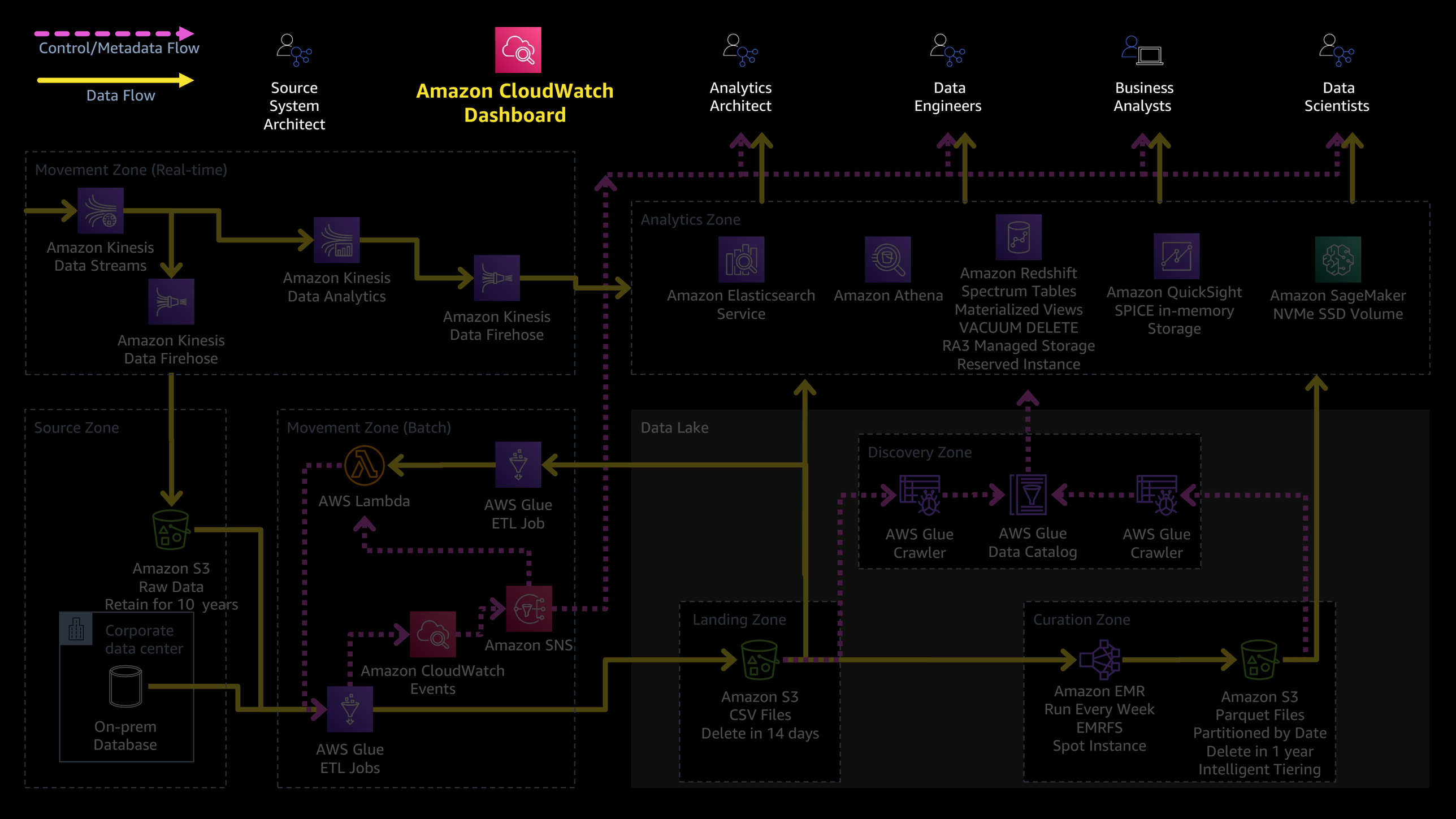
How do you measure the health of your analytics workload?

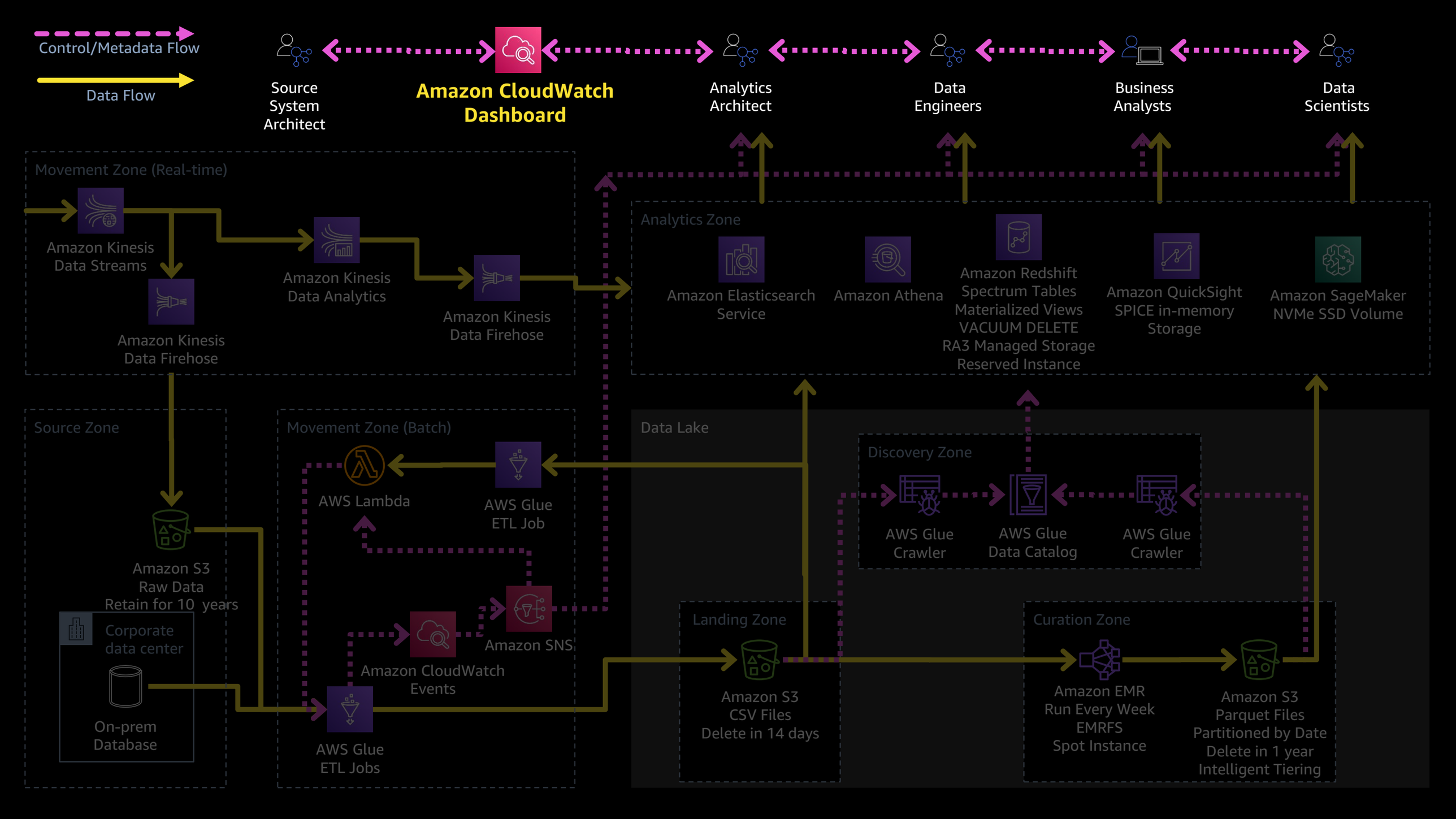
You should

- *Monitor the planned arrival of recent data from the source systems*
- *Validate the data quality of source systems before moving to analytics*
- *Define and measure the execution schedules of recurring analytics jobs*
- *Define and measure the runtime durations of analytics jobs*


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Reserved Instance
Amazon QuickSight
SPICE in-memory Storage
Amazon SageMaker
NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
AWS Lambda
AWS Glue ETL Job
Discovery Zone
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3
Raw Data
Retain for 10 years
Corporate data center
On-prem Database
Amazon CloudWatch Events
Amazon SNS
AWS Glue ETL Jobs
Landing Zone
Amazon S3
CSV Files
Delete in 14 days
Curation Zone
Amazon EMR
Run Every Week
EMRFS
Spot Instance
Amazon S3
Parquet Files
Partitioned by Date
Delete in 1 year
Intelligent Tiering


Control/Metadata Flow
Data Flow
Source System Architect
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Reserved Instance
Amazon QuickSight
SPICE in-memory Storage
Amazon SageMaker
NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3
Raw Data
Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
Amazon S3
CSV Files
Delete in 14 days
Amazon EMR
Run Every Week
EMRFS
Spot Instance
Amazon S3
Parquet Files
Partitioned by Date
Delete in 1 year
Intelligent Tiering
AWS Glue ETL Jobs


Control/Metadata Flow
Data Flow
Source System Architect
Amazon CloudWatch Dashboard
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Reserved Instance
Amazon QuickSight
SPICE in-memory Storage
Amazon SageMaker
NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3
Raw Data
Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
AWS Glue ETL Jobs
Amazon S3
CSV Files
Delete in 14 days
Amazon EMR
Run Every Week
EMRFS
Spot Instance
Amazon S3
Parquet Files
Partitioned by Date
Delete in 1 year
Intelligent Tiering


Control/Metadata Flow
Data Flow
Source System Architect
Amazon CloudWatch Dashboard
Analytics Architect
Data Engineers
Business Analysts
Data Scientists
Movement Zone (Real-time)
Amazon Kinesis Data Streams
Amazon Kinesis Data Analytics
Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Analytics Zone
Amazon Elasticsearch Service
Amazon Athena
Amazon Redshift
Spectrum Tables
Materialized Views
VACUUM DELETE
RA3 Managed Storage
Reserved Instance
Amazon QuickSight
SPICE in-memory Storage
Amazon SageMaker
NVMe SSD Volume
Source Zone
Movement Zone (Batch)
Data Lake
Discovery Zone
AWS Lambda
AWS Glue ETL Job
AWS Glue Crawler
AWS Glue Data Catalog
AWS Glue Crawler
Amazon S3 Raw Data Retain for 10 years
Corporate data center
On-prem Database
Amazon SNS
Amazon CloudWatch Events
Landing Zone
Curation Zone
AWS Glue ETL Jobs
Amazon S3 CSV Files Delete in 14 days
Amazon EMR Run Every Week EMRFS Spot Instance
Amazon S3 Parquet Files Partitioned by Date Delete in 1 year Intelligent Tiering

TAKEAWAY: USE ANALYTICS LENS FOR A WELL-ARCHITECTED DESIGN FOR YOUR WORKLOAD



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Analytics Lens Design Principles



Security

Retain Least Privilege for Source Access
Classify Data
Grant Least Privilege to Analytics Users



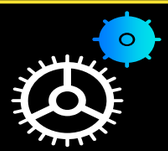
Cost Optimization

Delete Out-of-Use Data
Choose Lowest-Cost Storage Solution
Choose Lowest-Cost Compute Solution



Reliability

Design Resiliency for Analytics Pipelines
Govern Data Changes



Operational Excellence

Monitor the Health of Analytics Pipelines



Performance Efficiency

Choose Optimal File Format and Partitioning
Choose Optimal Storage Solution
Choose Optimal Compute Solution

Wrapping up

AWS Well-Architected Analytics Lens

Tailored best practices for Analytics workload in the AWS Cloud

docs.aws.amazon.com/wellarchitected/latest/analytics-lens



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Architecture resources

AWS Well-Architected Framework

The official best practices for architecting in the AWS Cloud

aws.amazon.com/architecture/well-architected

AWS Well-Architected Labs

Hands-on labs to help you learn, measure, and build using architectural best practices

wellarchitectedlabs.com

AWS Architecture Center

Official AWS repository for all architecture resources

aws.amazon.com/architecture

AWS Solutions Library

Vetted reference implementations and well-architected patterns

aws.amazon.com/solutions



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Jongnam Lee

jongnam@amazon.com
Twitter: @jongnam

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.