AWS PARTNER CERTIFICATION READINESS

Content Review Session

Week 4 – Domain 3

Data Operations and Support

OPTIONAL AWS Skill Builder Subscription

The Skill Builder subscription provides access to official AWS Certification practice exams, self-paced digital training content including open-ended challenges, self-paced labs, and game-based learning. **Please note, the Skill Builder subscription is not required for this Accelerator program.**

Free digital training

LINK HERE

Special features include:

- 600+ digital courses
- Learning plans
- 10 Practice Question Sets
- *AWS Cloud Quest (Foundational)*

Individual subscription

LINK HERE

Everything in free digital training, plus:

- AWS Cloud Quest (Intermediate - Advanced)
- AWS Certification Official Practice Exams
- Enhanced Exam Prep Courses
- Unlimited access to 1000+ hands-on labs
- AWS Jam Journeys (lab-based challenges)
- AWS Digital Classroom (Annual only)

Individual subscriptions are priced at $29 USD per month (*Flexibility to cancel anytime*) or $449 USD per year.

Access ***65*** Data Engineer - Associate Practice Exam Questions with feedback on your answer choices

Today's Learning Outcomes

During this session, we will cover:

- Automate data processing by using AWS services
- Analyze data by using AWS services
- Maintain and monitor data pipelines
- Ensure data quality

AWS PARTNER CERTIFICATION READINESS

Domain 3: Data Operations and Support

Automate data processing by using AWS services

Automate data processing by using AWS services

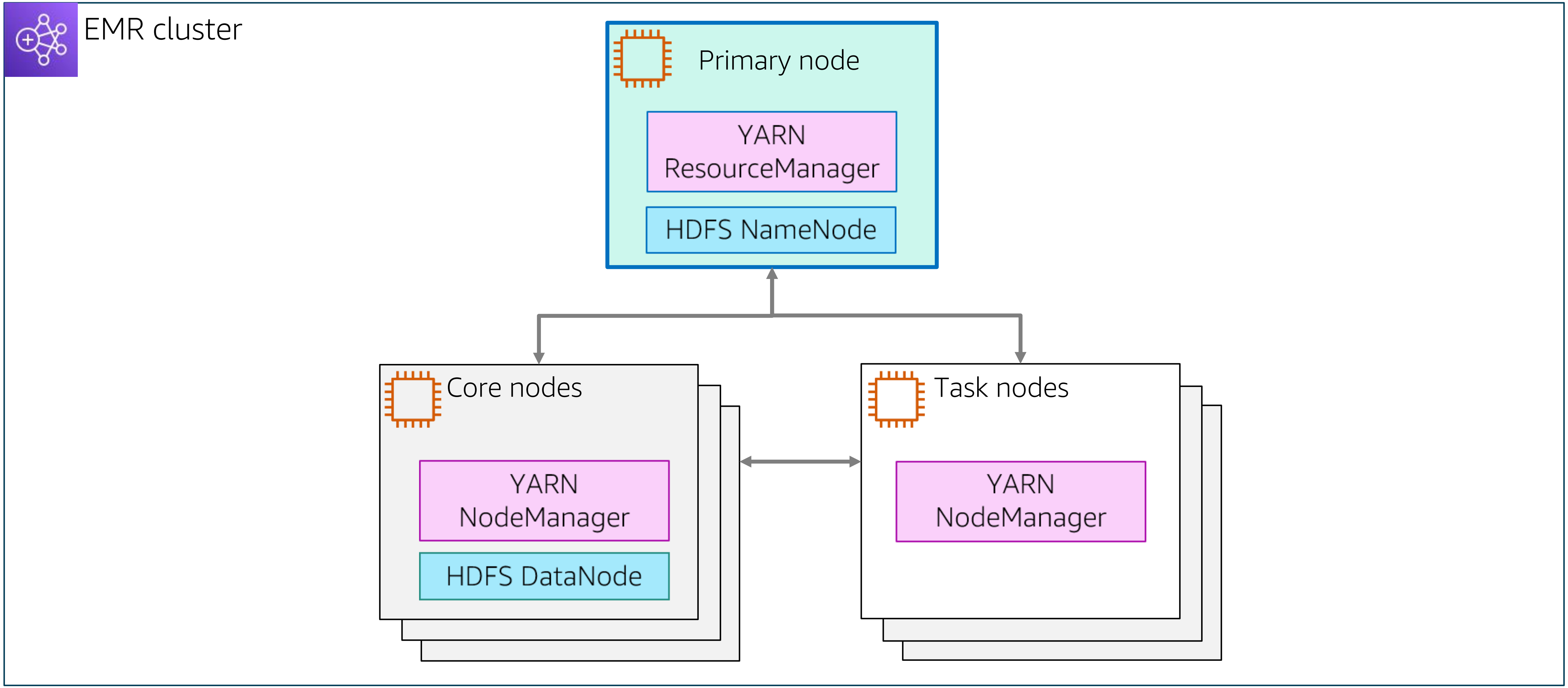
Knowledge of:

- How to maintain and troubleshoot data processing for repeatable business outcomes
- API calls for data processing
- Which services accept scripting (for example, Amazon EMR, Amazon Redshift, AWS Glue)

Skills in:

- Orchestrating data pipelines (for example, Amazon MWAA, Step Functions)
- Troubleshooting Amazon managed workflows
- Calling SDKs to access Amazon features from code
- Using the features of AWS services to process data (for example, Amazon EMR, Amazon Redshift, AWS Glue)
- Consuming and maintaining data APIs
- Preparing data transformation (for example, AWS Glue DataBrew)
- Querying data (for example, Amazon Athena)
- Using Lambda to automate data processing
- Managing events and schedulers (for example, EventBridge)

Amazon Elastic Map Reduce (EMR)



EMR cluster

Primary node

YARN ResourceManager

HDFS NameNode

Core nodes

YARN NodeManager

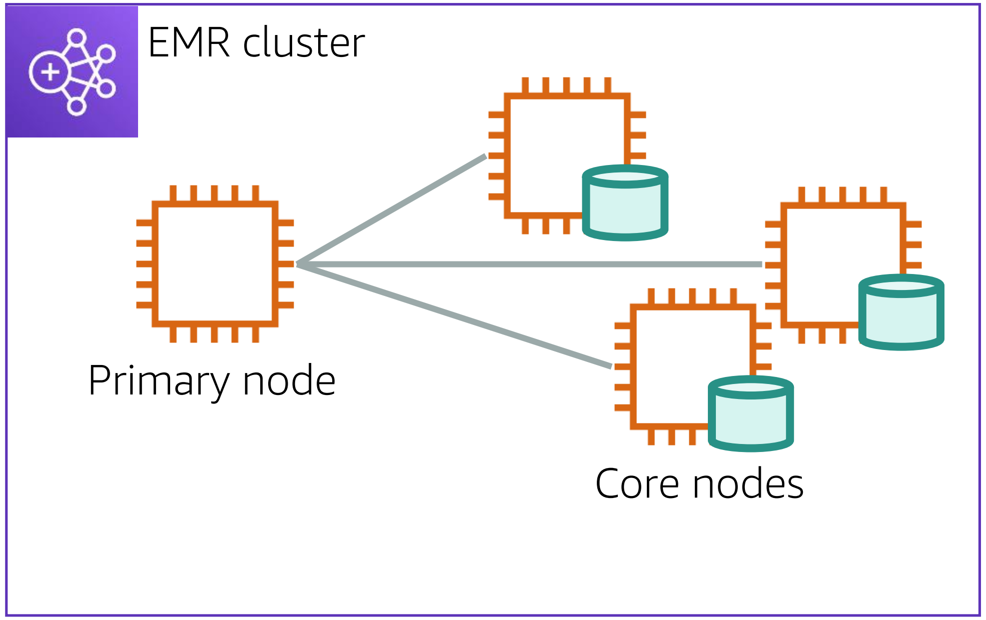
HDFS DataNode

Task nodes

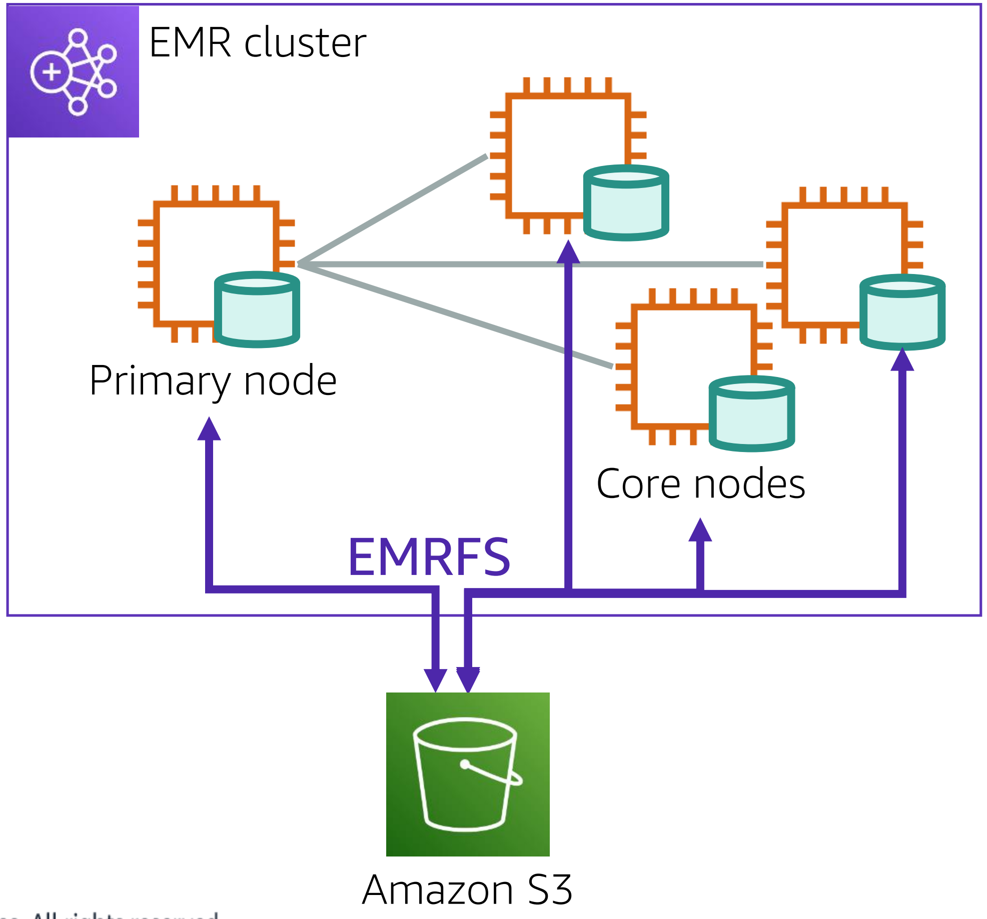
YARN NodeManager

Amazon EMR data storage options

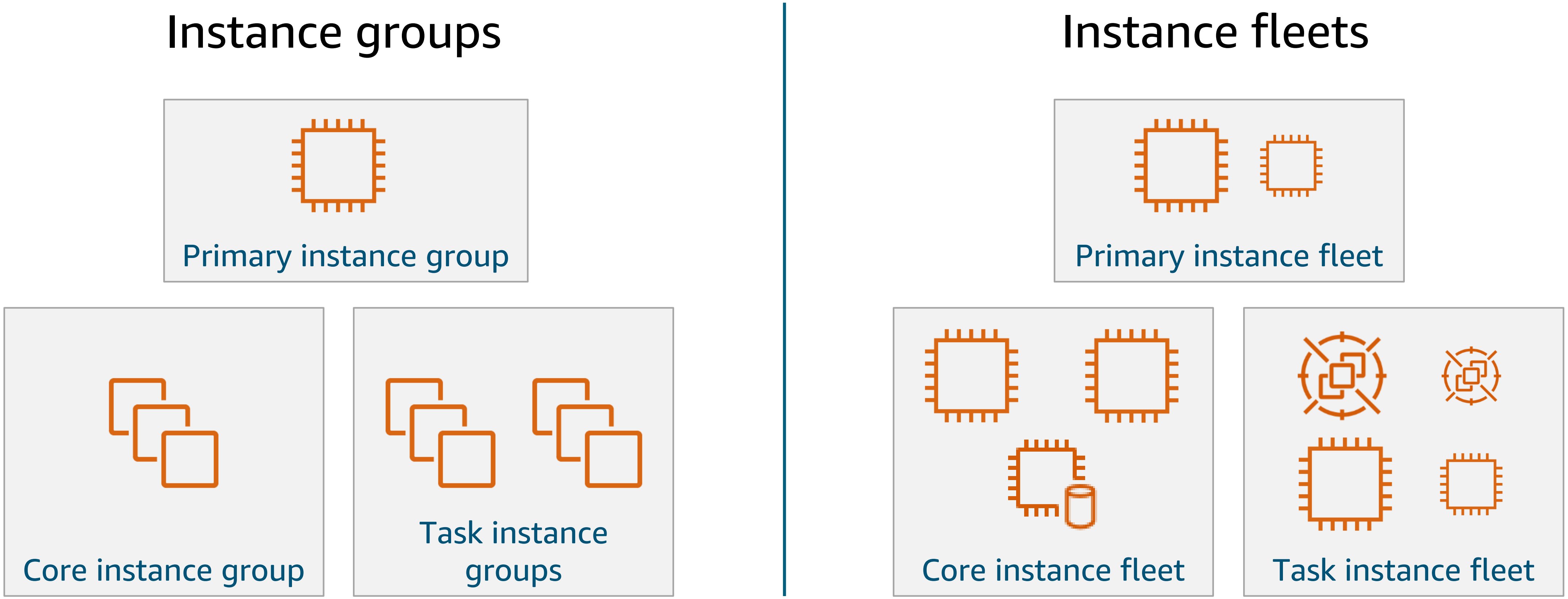
HDFS



EMRFS



Node configuration options



Instance groups

Primary instance group

Core instance group

Task instance groups

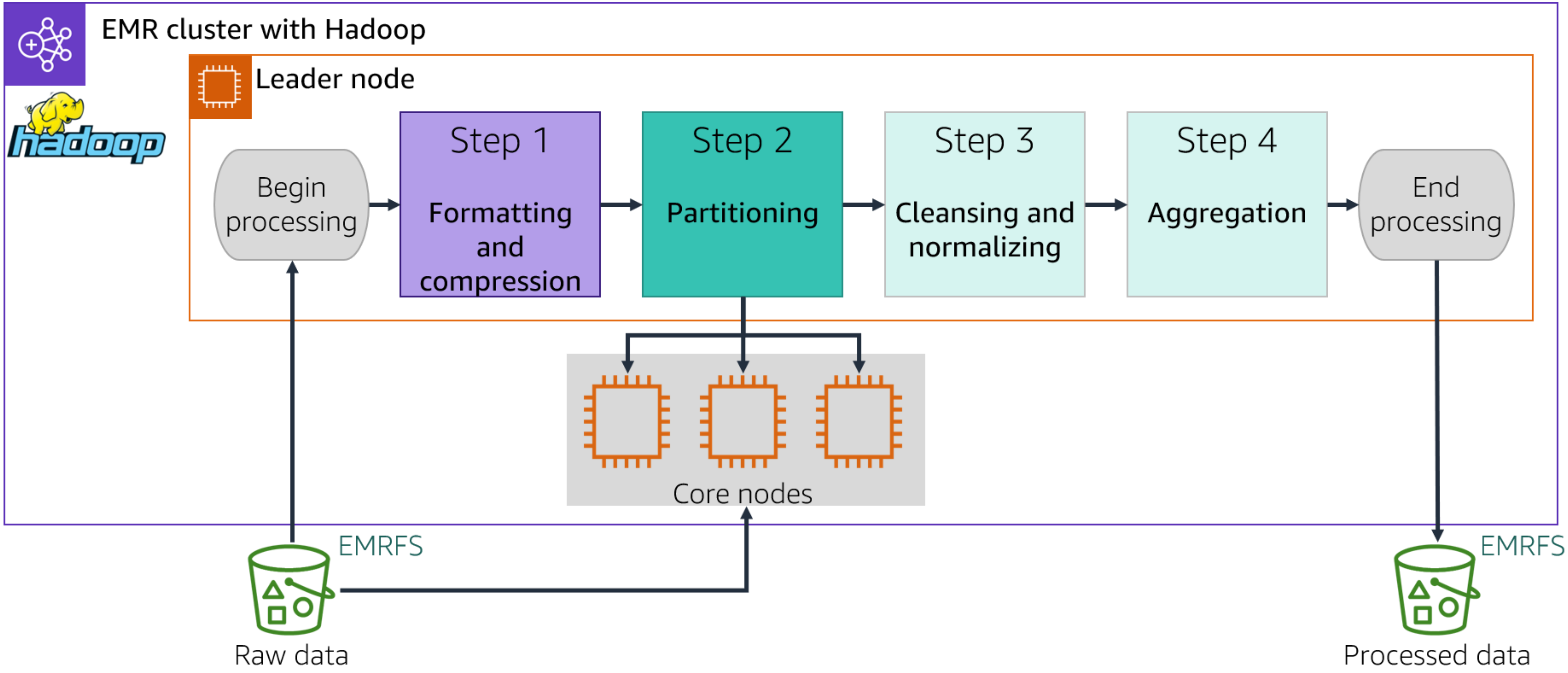
Instance fleets

Primary instance fleet

Core instance fleet

Task instance fleet

Transforming data in Amazon EMR



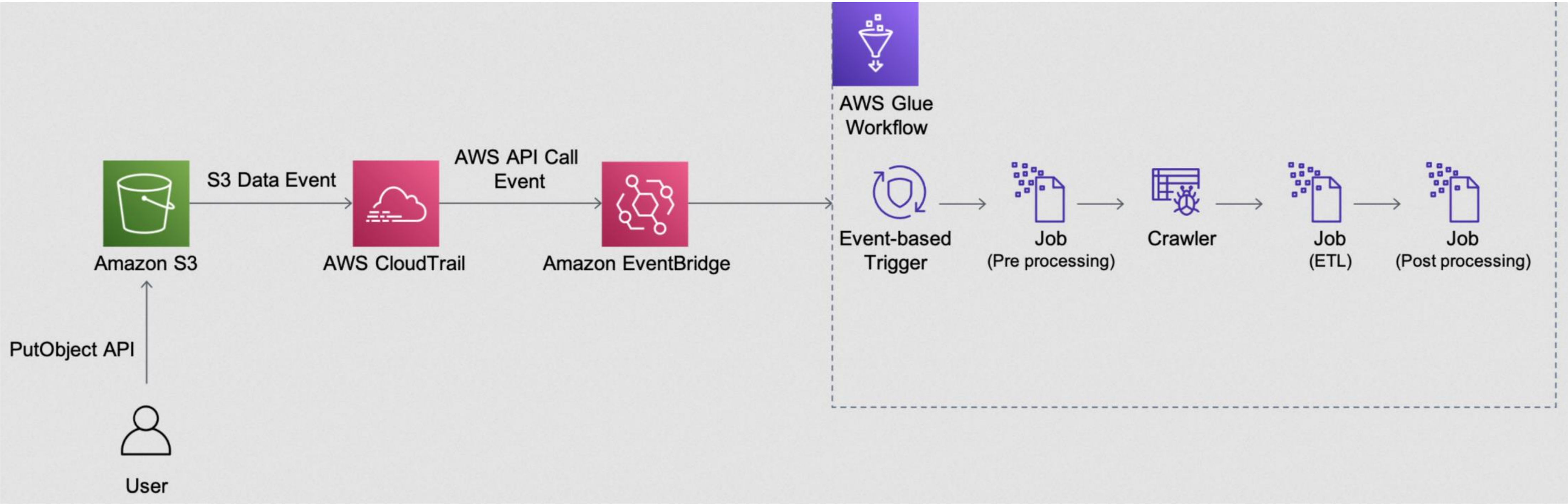
Considerations in choosing Spark on Amazon EMR or AWS Glue ETL

Consideration	Spark on Amazon EMR	AWS Glue ETL
Responsibility model	Fully managed	Serverless
Degree of control	More flexibility	More automation
Data model	Schema before load	Schema on read
Scalability	Policy-based or managed scaling	Specify max workers and concurrency
Pricing	Stable cost for consistent workloads	Pay for use based on needs

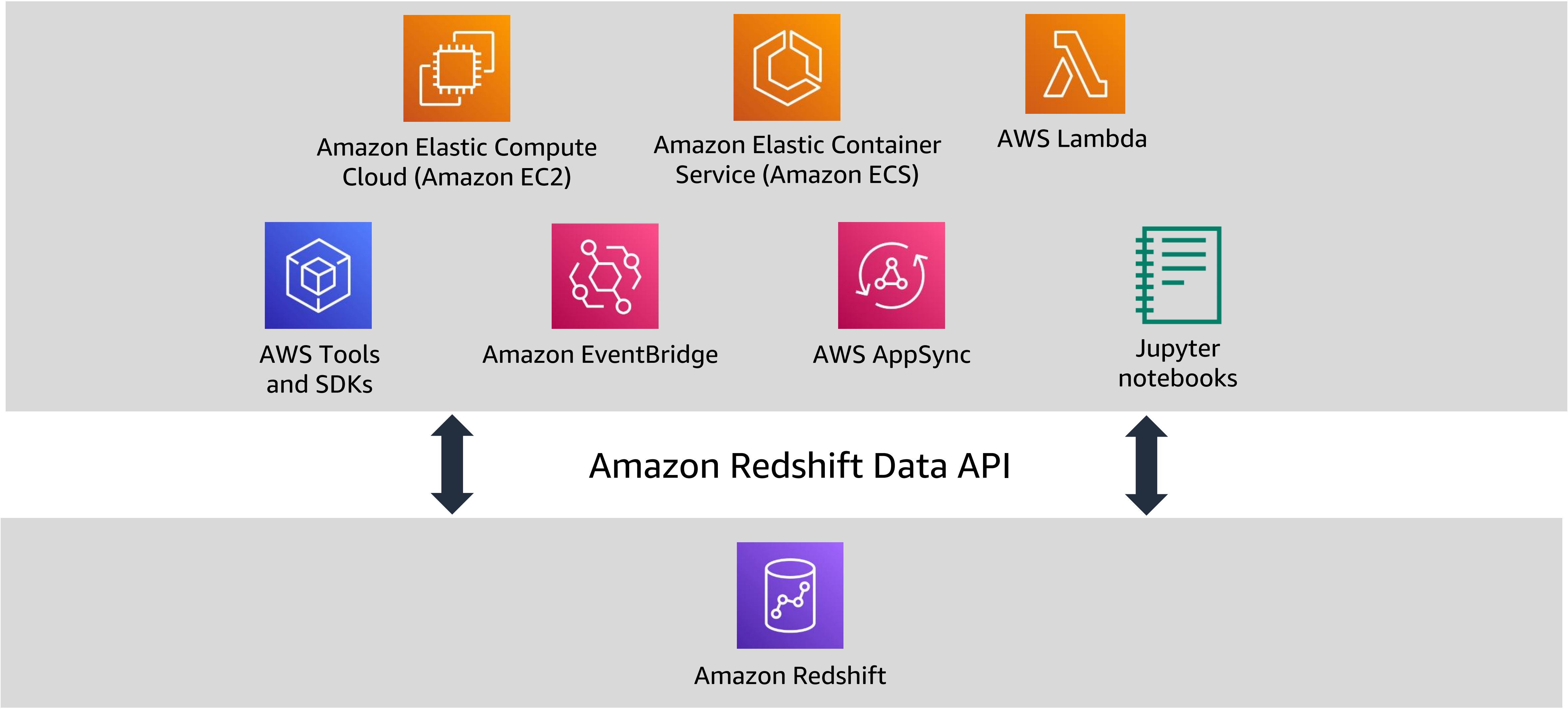
Amazon EventBridge

- EventBridge is a serverless service for building event-driven, loosely-coupled applications by routing events between sources and targets

- Event buses receive events from many sources and deliver to multiple targets, with optional event transformation

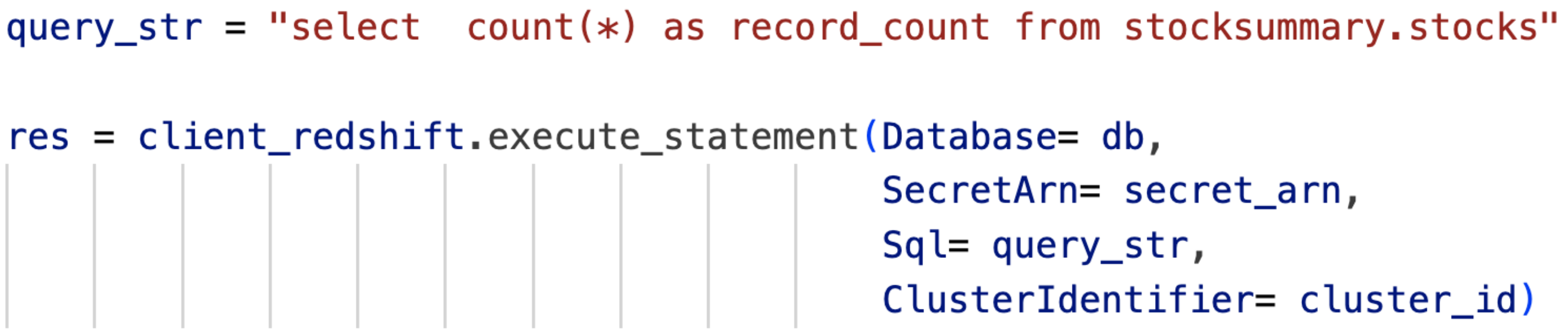


Redshift Data API



Redshift Data API code example (Python)

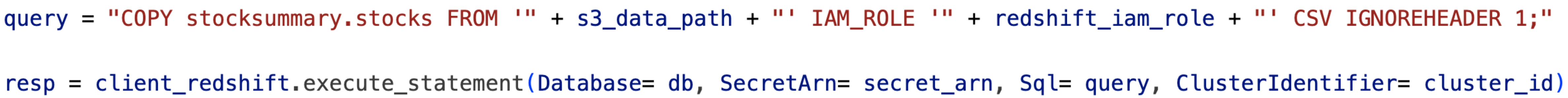
Using the Redshift Data API to run SQL:



```python
query_str = "select  count(*) as record_count from stocksummary.stocks"

res = client_redshift.execute_statement(Database= db,
                                        SecretArn= secret_arn,
                                        Sql= query_str,
                                        ClusterIdentifier= cluster_id)
```

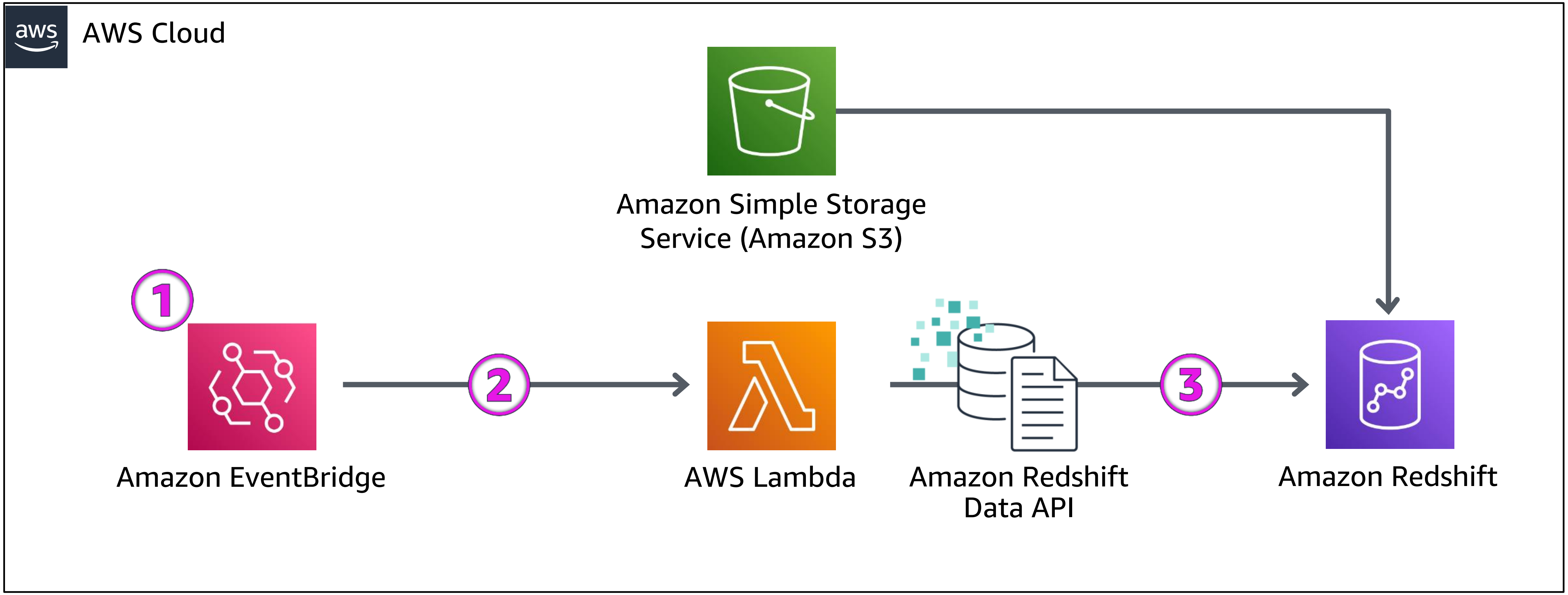
Using the COPY command and the Redshift Data API:



```python
query = "COPY stocksummary.stocks FROM '" + s3_data_path + "' IAM_ROLE '" + redshift_iam_role + "' CSV IGNOREHEADER 1;"

resp = client_redshift.execute_statement(Database= db, SecretArn= secret_arn, Sql= query, ClusterIdentifier= cluster_id)
```

Architecting an event-driven architecture using the Amazon Redshift Data API



AWS PARTNER CERTIFICATION READINESS

Domain 3: Data Operations and Support

Analyze data by using AWS services

Analyze data by using AWS services

Knowledge of:

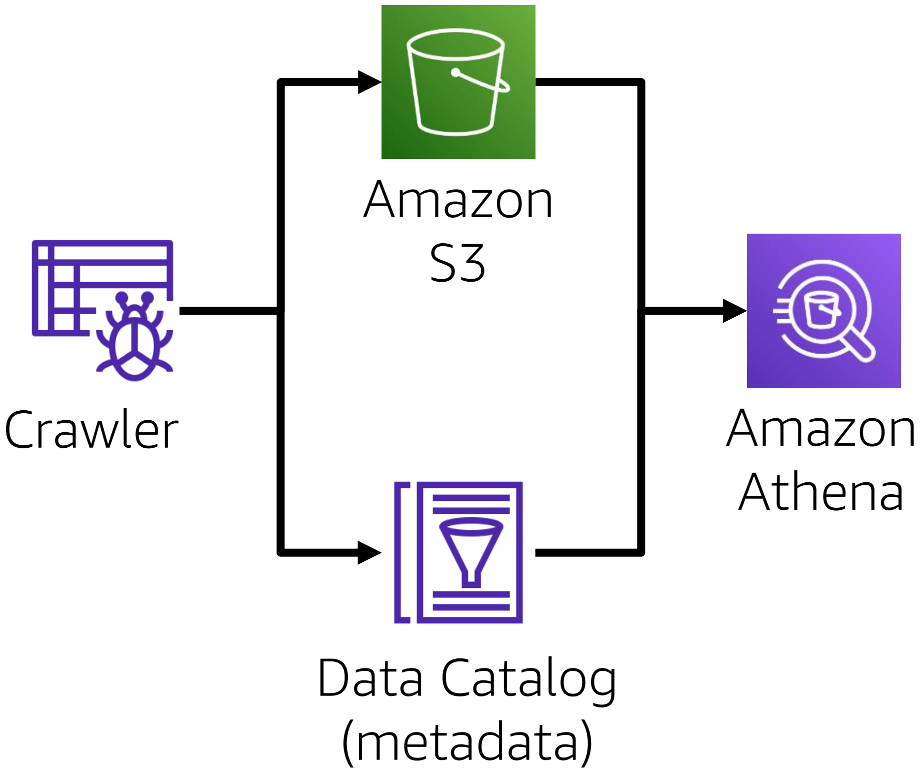
- Tradeoffs between provisioned services and serverless services
- SQL queries (for example, SELECT statements with multiple qualifiers or JOIN clauses)
- How to visualize data for analysis
- When and how to apply cleansing techniques
- Data aggregation, rolling average, grouping, and pivoting

Skills in:

- Visualizing data by using AWS services and tools (for example, AWS Glue DataBrew, Amazon QuickSight)
- Verifying and cleaning data (for example, Lambda, Athena, QuickSight, Jupyter Notebooks, Amazon SageMaker Data Wrangler)
- Using Athena to query data or to create views
- Using Athena notebooks that use Apache Spark to explore data

Athena

Serverless query engine

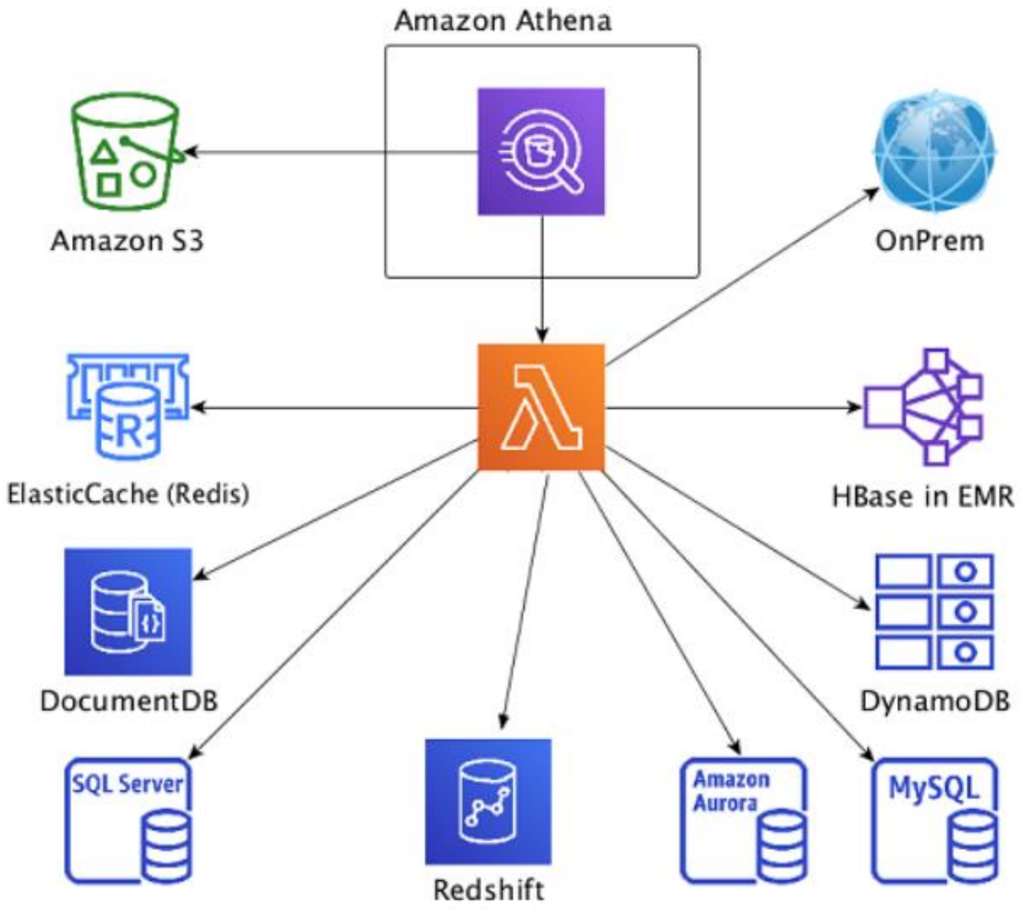


Benefits

- Query Amazon S3 data directly or the AWS Glue Data Catalog
- Use SQL based on Presto
- Supports CSV, JSON, ORC, Avro, Parquet
- Can integrate with Amazon QuickSight
- Support for federated query

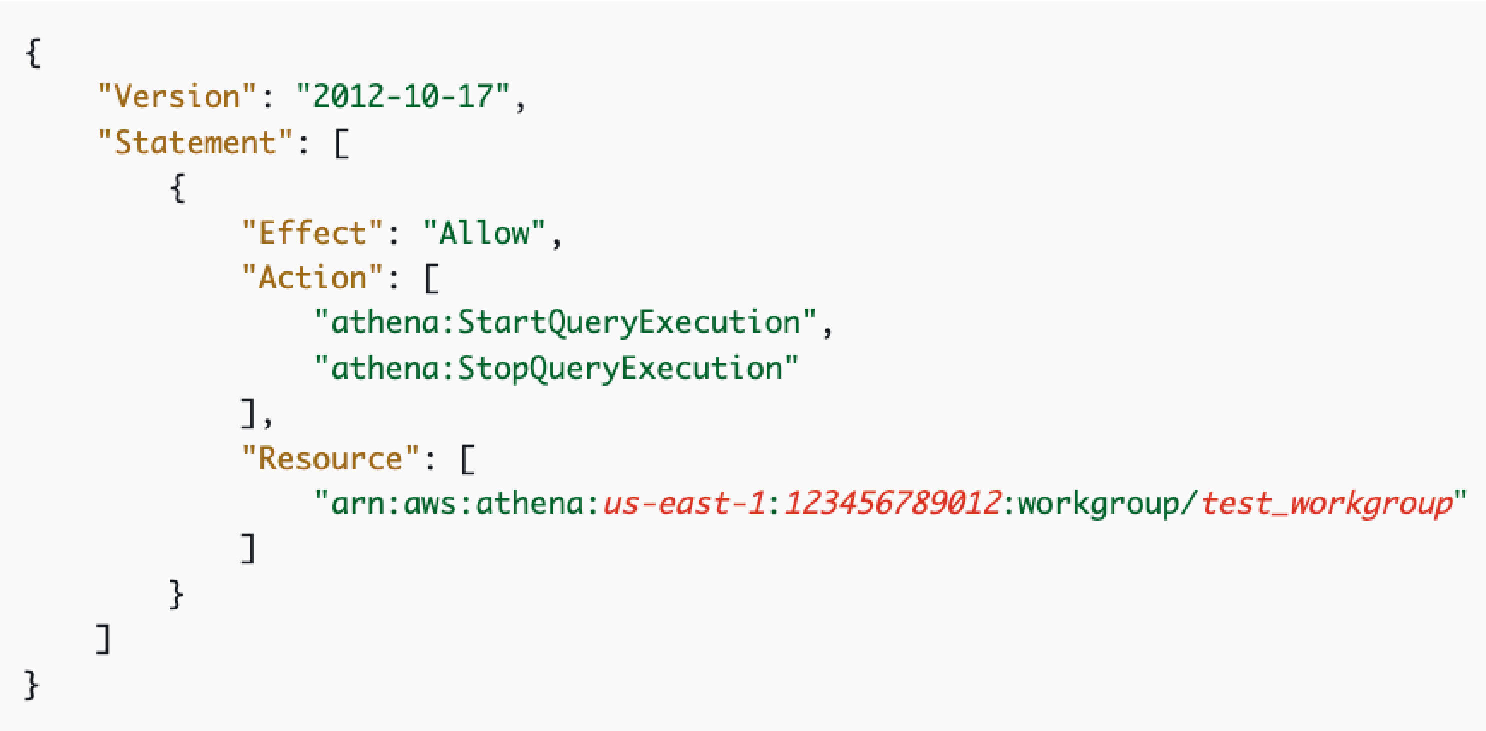
Athena Federated Query

- Enables federated querying across multiple data sources with SQL
- Utilizes Lambda connectors for different data sources
- Allows building custom connectors for proprietary data sources



Athena workgroups

- Isolate workloads.
- Control user access.
- Manage query usage, costs, and data limits
- A workgroup is an IAM resource managed by Athena.

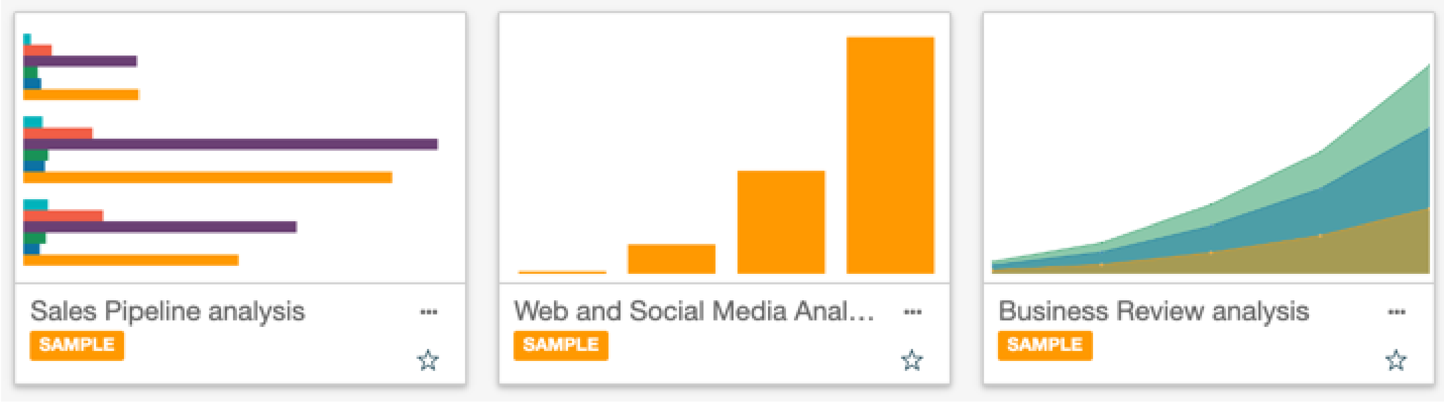


```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "athena:StartQueryExecution",
                "athena:StopQueryExecution"
            ],
            "Resource": [
                "arn:aws:athena:us-east-1:123456789012:workgroup/test_workgroup"
            ]
        }
    ]
}
```

QuickSight

- Scalable, serverless business intelligence service
- Embeddable into your existing applications
- Dynamic, machine learning (ML) powered insights

QuickSight



Accessing data with QuickSight

SPICE

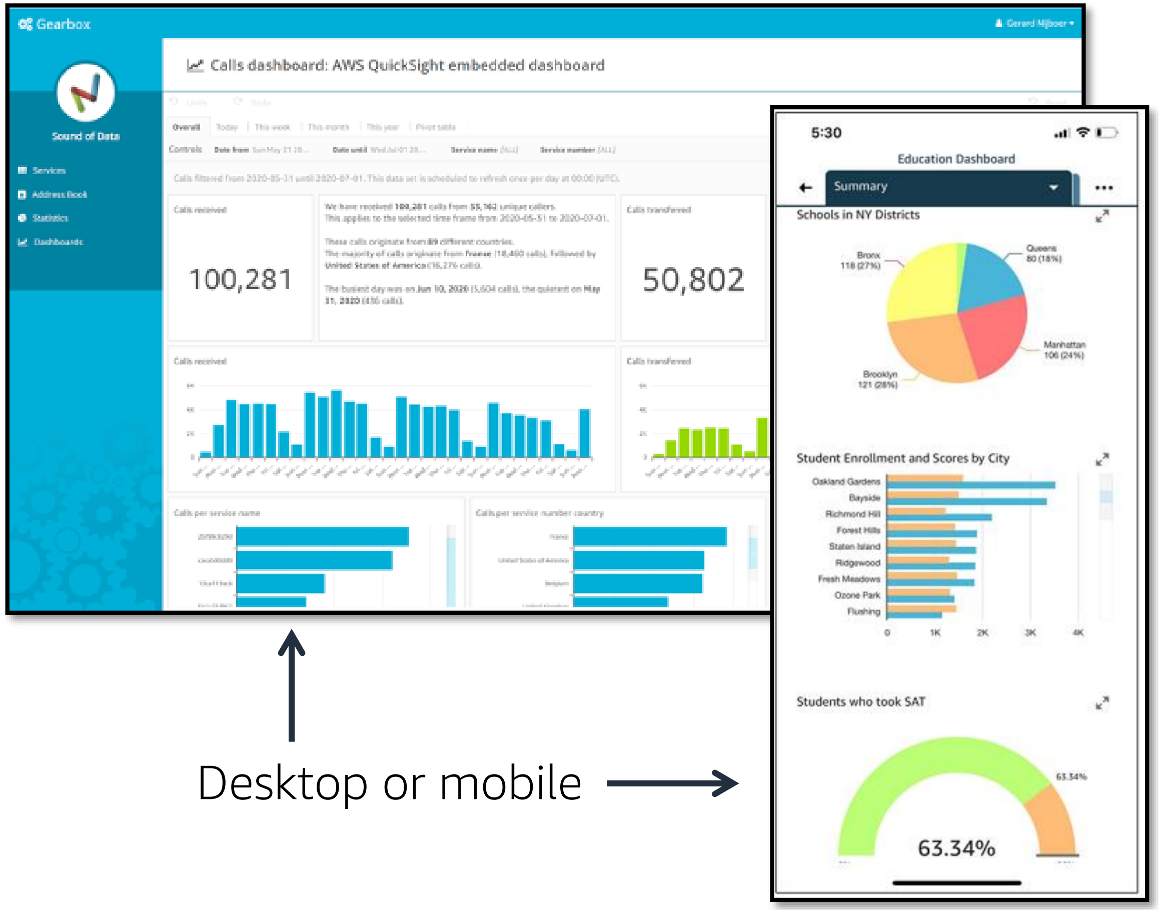
Super-fast, Parallel, In-memory Calculation Engine

- Faster processing
- Reduced wait time vs. direct queries
- Reduced cost through reuse

QuickSight embedded analytics

Embed visualizations in any application for desktop or mobile applications.

1. Create a dashboard
2. Apply permissions
3. Authenticate your app server
4. Embed via JavaScript SDK



Calls dashboard: AWS QuickSight embedded dashboard

Calls filtered from 2020-05-31 until 2020-07-01. This data set is scheduled to refresh once per day at 00:00 (UTC).

We have received 100,281 calls from 55,162 unique callers.
This applies to the selected time frame from 2020-05-31 to 2020-07-01.

These calls originate from 89 different countries.
The majority of calls originate from France (18,460 calls), followed by United States of America (16,276 calls).

The busiest day was on Jun 10, 2020 (5,604 calls), the quietest on May 31, 2020 (436 calls).

Desktop or mobile

AWS PARTNER CERTIFICATION READINESS

Domain 3: Data Operations and Support

Maintain and monitor data pipelines

Maintain and monitor data pipelines

Knowledge of:

- How to log application data
- Best practices for performance tuning
- How to log access to AWS services
- Amazon Macie, AWS CloudTrail, and Amazon CloudWatch

Skills in:

- Extracting logs for audits
- Deploying logging and monitoring solutions to facilitate auditing and traceability
- Using notifications during monitoring to send alerts
- Troubleshooting performance issues
- Using CloudTrail to track API calls
- Troubleshooting and maintaining pipelines (for example, AWS Glue, Amazon EMR)
- Using Amazon CloudWatch Logs to log application data (with a focus on configuration and automation)
- Analyzing logs with AWS services (for example, Athena, Amazon EMR, Amazon OpenSearch Service, CloudWatch Logs Insights, big data application logs)

Data pipelines on AWS

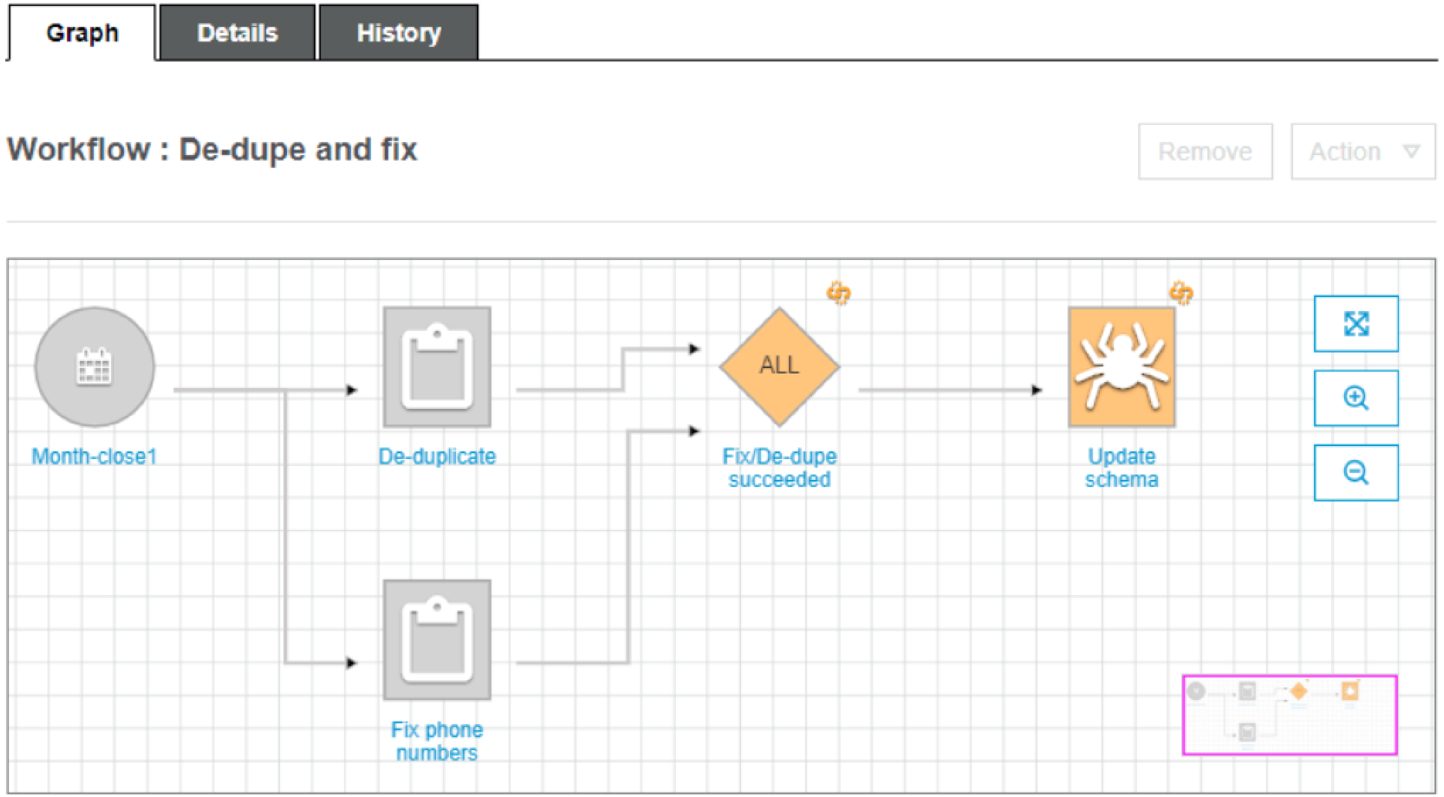
An efficient and well-designed data integration pipeline is critical for making the data available and trusted amongst the analytics consumers.

Here are some considerations to review when designing data pipelines:

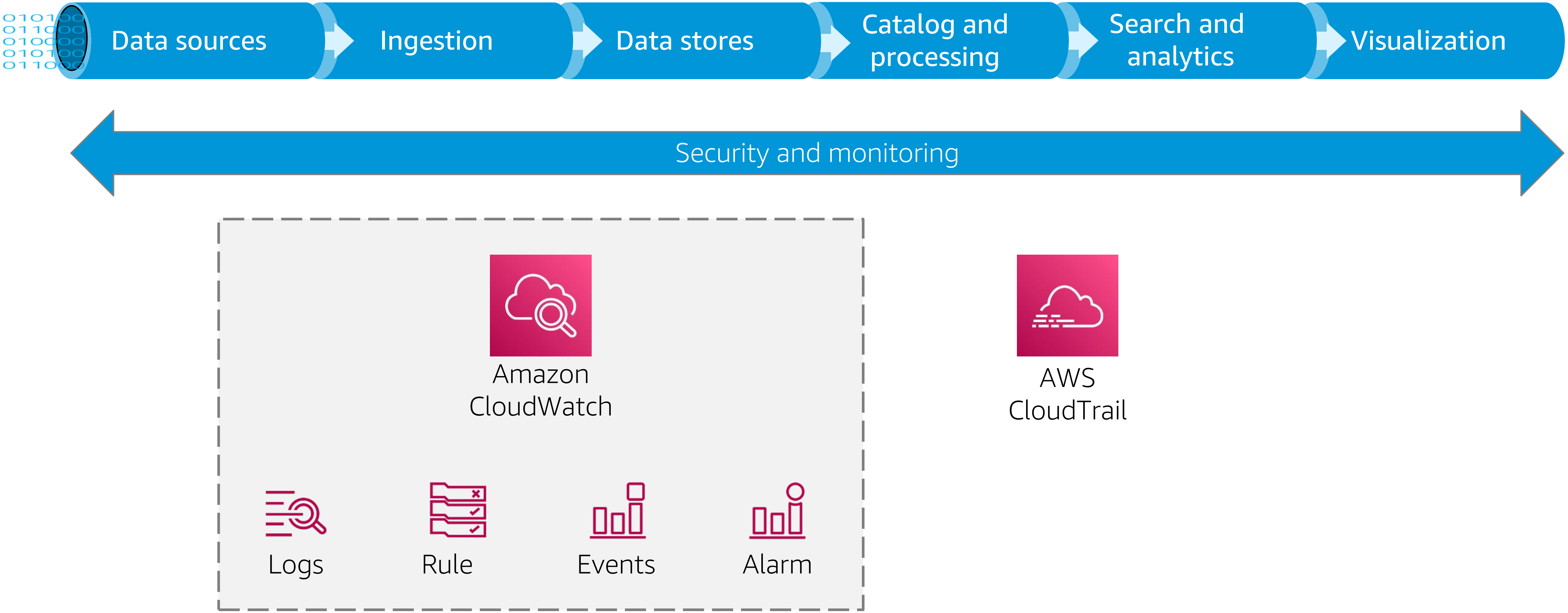
Factor	AWS Glue Workflow	AWS Step Function	Amazon Managed Workflow for Apache Airflow (MWAA)
Use case	Suitable when your pipeline consists of mostly AWS Glue jobs and crawlers.	Suitable when there is a need to integrate with different services, including AWS Lambda, SSM, and so on.	Compatible with open-source Airflow and suitable when you want to reuse existing Airflow assets.
Infrastructure	Serverless	Serverless	Managed service
Building a data pipeline	Build a data pipeline using an AWS Glue job written in Python or /Scala and crawlers.	Build a data pipeline using the Step Functions console. Possible to integrate with non-supported services using Lambda.	Workflows are created as DAGs, which are defined within a Python file that defines the DAG's structure as code.

AWS Glue Workflow

- Create/visualize complex ETL workflows in AWS Glue
- Workflow triggers:
  - Schedules
  - On-demand
  - EventBridge events
- Visual graph of execution progress and component dependencies



Monitoring a Data Lake



AWS PARTNER CERTIFICATION READINESS

Domain 3: Data Operations and Support

Ensure data quality

Ensure data quality

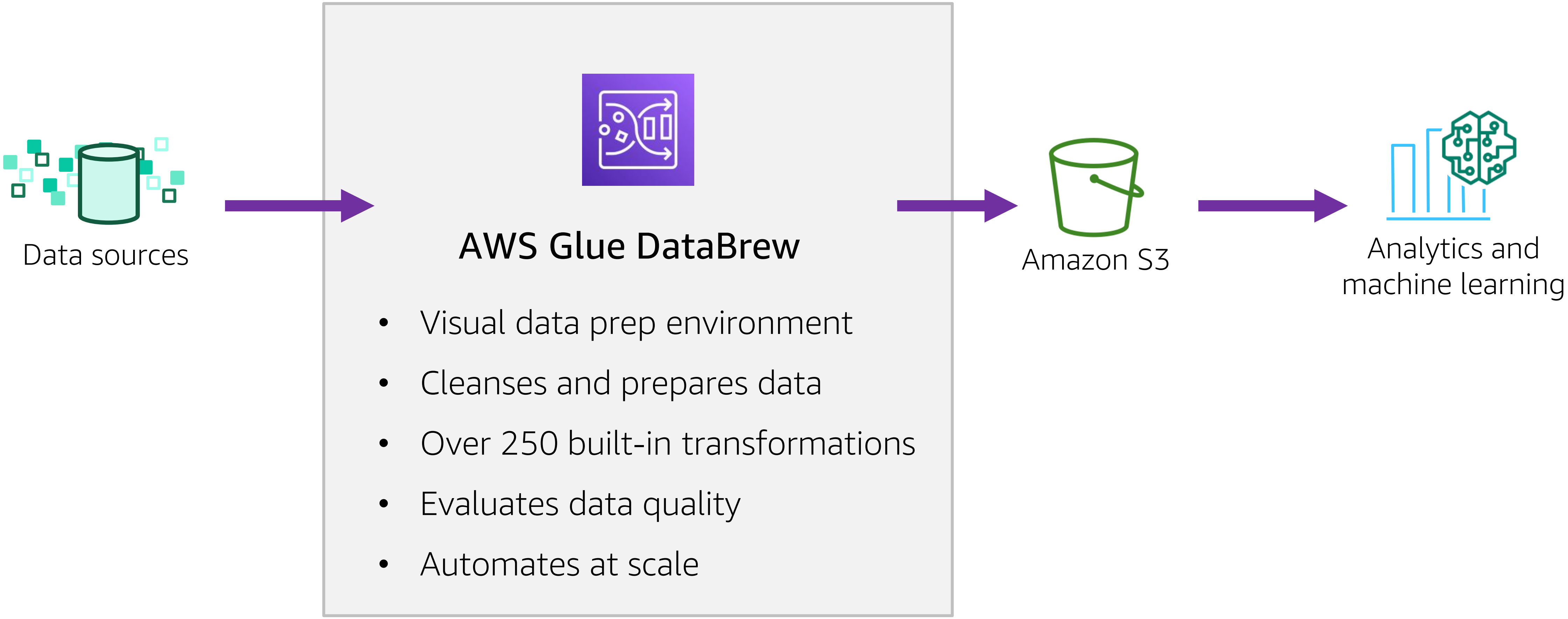
Knowledge of:

- Data sampling techniques
- How to implement data skew mechanisms
- Data validation (data completeness, consistency, accuracy, and integrity)
- Data profiling

Skills in:

- Running data quality checks while processing the data (for example, checking for empty fields)
- Defining data quality rules (for example, AWS Glue DataBrew)
- Investigating data consistency (for example, AWS Glue DataBrew)

Glue Databrew



Data sources

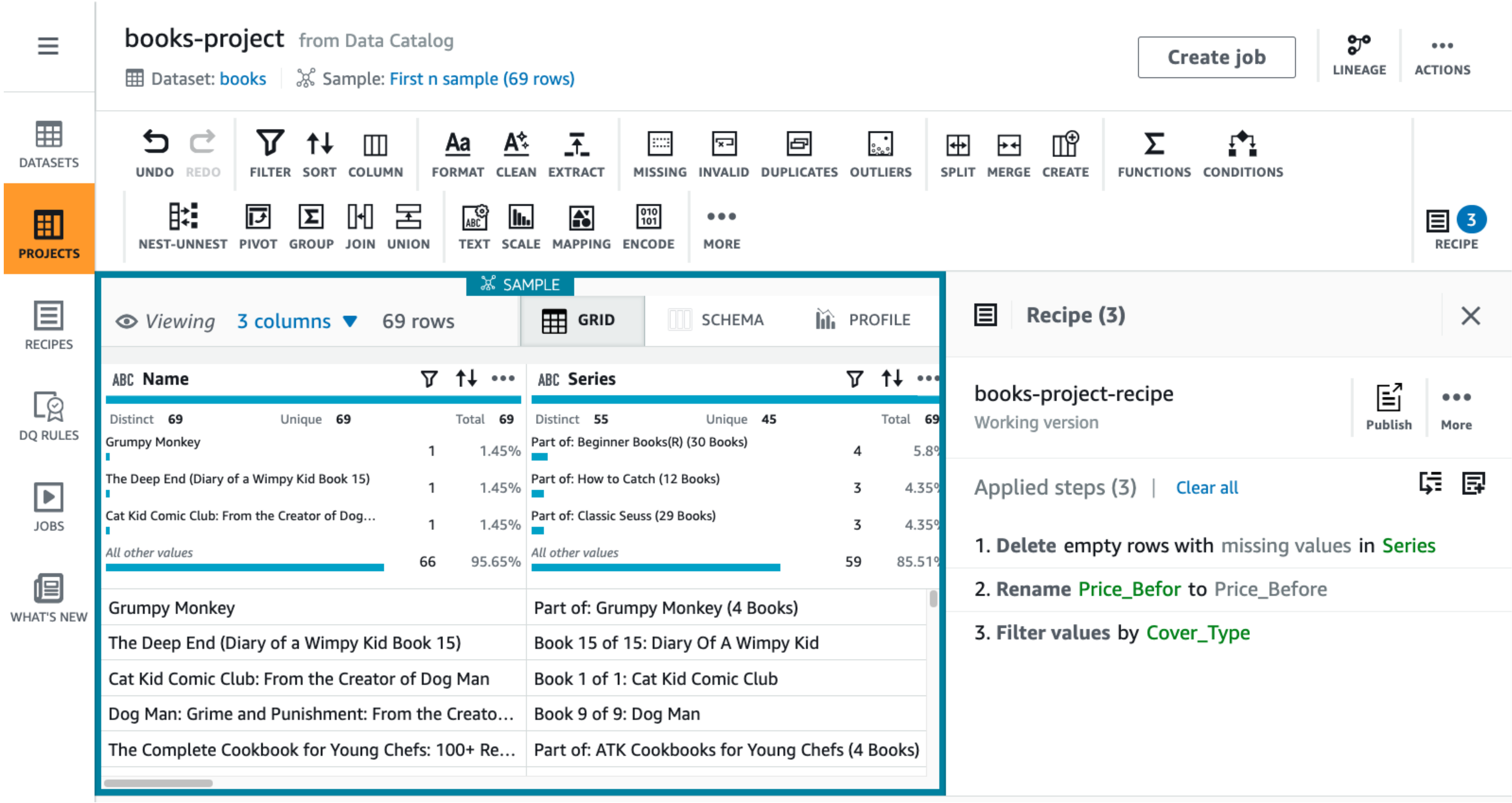
AWS Glue DataBrew

- Visual data prep environment
- Cleanses and prepares data
- Over 250 built-in transformations
- Evaluates data quality
- Automates at scale

Amazon S3

Analytics and machine learning

Glue Databrew



books-project from Data Catalog

Dataset: books | Sample: First n sample (69 rows)

Viewing 3 columns ▼ 69 rows

ABC Name			ABC Series		
Distinct 69 Unique 69		Total 69	Distinct 55 Unique 45		Total 69
Grumpy Monkey	1	1.45%	Part of: Beginner Books(R) (30 Books)	4	5.8%
The Deep End (Diary of a Wimpy Kid Book 15)	1	1.45%	Part of: How to Catch (12 Books)	3	4.35%
Cat Kid Comic Club: From the Creator of Dog...	1	1.45%	Part of: Classic Seuss (29 Books)	3	4.35%
All other values	66	95.65%	*All other values*	59	85.51%

Name	Series
Grumpy Monkey	Part of: Grumpy Monkey (4 Books)
The Deep End (Diary of a Wimpy Kid Book 15)	Book 15 of 15: Diary Of A Wimpy Kid
Cat Kid Comic Club: From the Creator of Dog Man	Book 1 of 1: Cat Kid Comic Club
Dog Man: Grime and Punishment: From the Creato...	Book 9 of 9: Dog Man
The Complete Cookbook for Young Chefs: 100+ Re...	Part of: ATK Cookbooks for Young Chefs (4 Books)

Recipe (3)

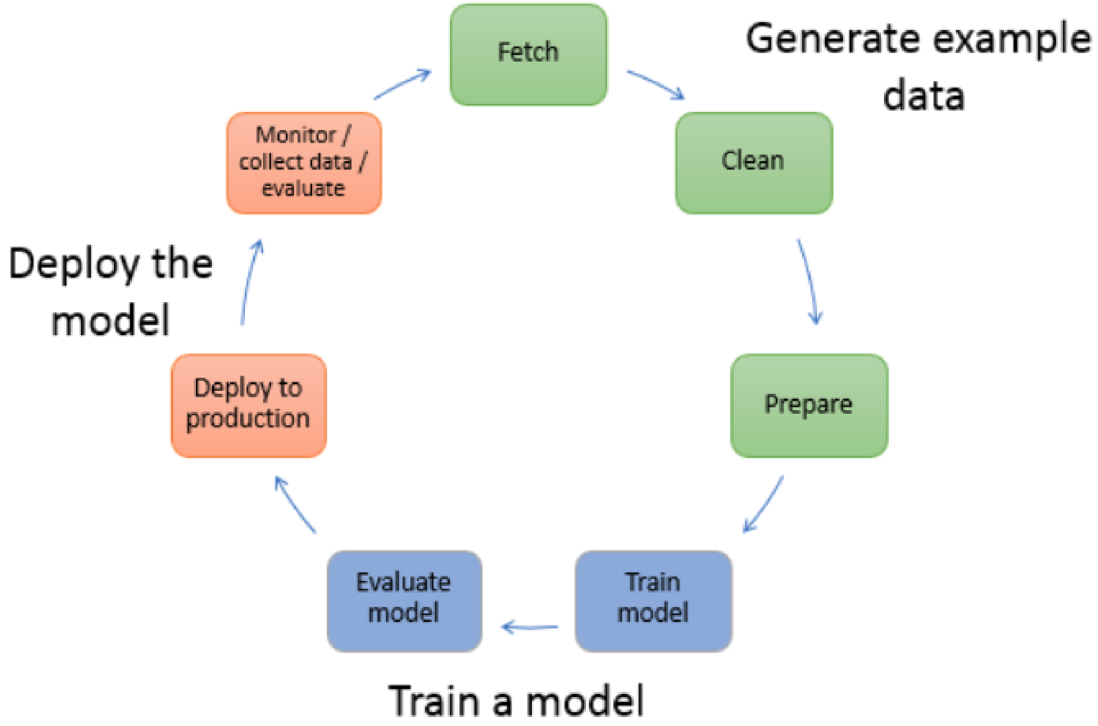
books-project-recipe
Working version

Applied steps (3) | Clear all

1. **Delete** empty rows with missing values in Series
2. **Rename** Price_Befor to Price_Before
3. **Filter values** by Cover_Type

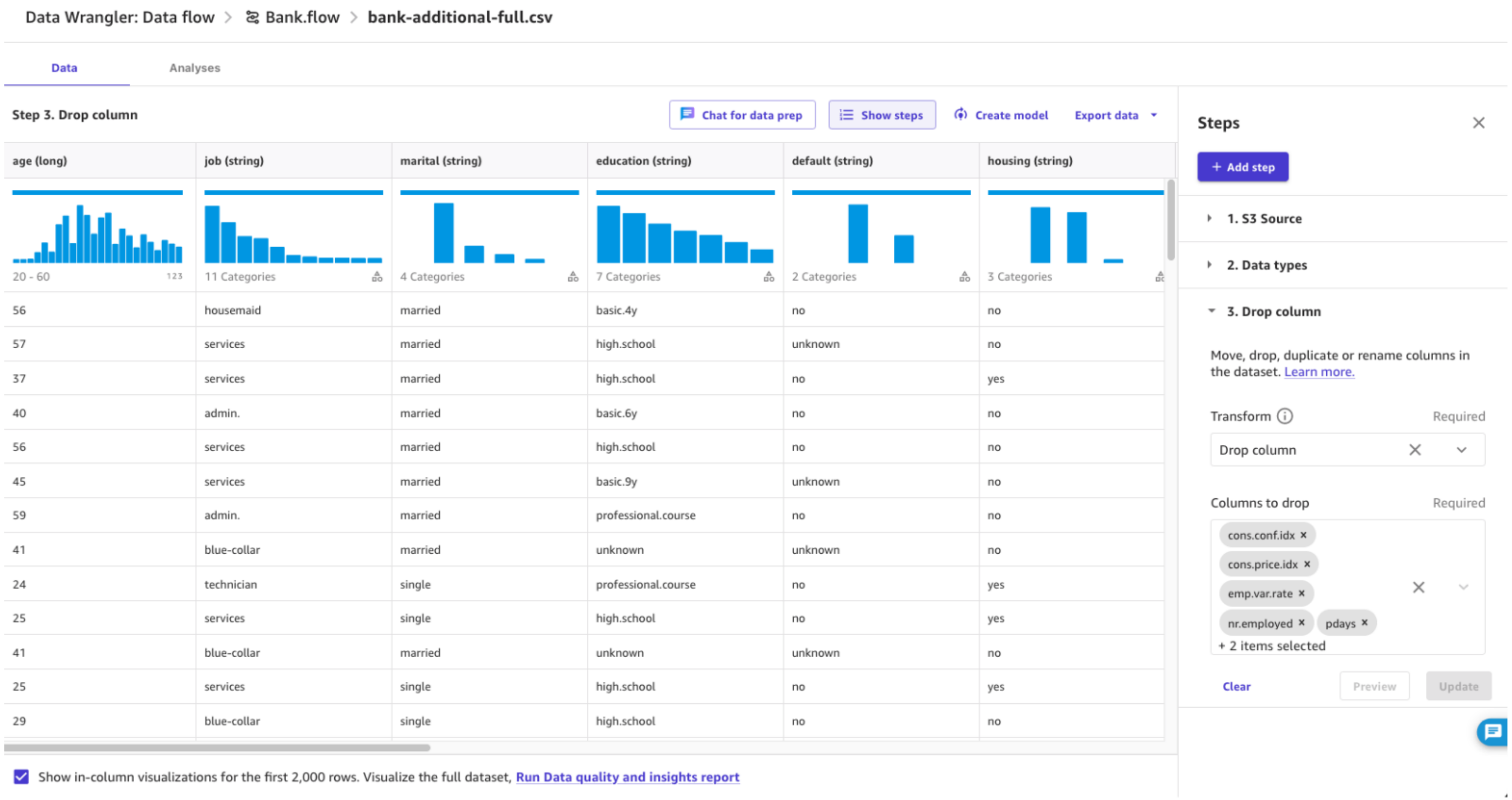
SageMaker

- Fully managed ML service
- End-to-end model development
- Built-in algorithms/models, auto-tuning, hosting, and more



SageMaker Data Wrangler

- Prepare and featurize data with Data Wrangler flows involving minimal coding.
- Analyze data quality, visualize features, and perform quick modeling.
- Export workflows to integrate with SageMaker pipelines, Feature Store, or custom scripts.



Data Wrangler: Data flow > Bank.flow > bank-additional-full.csv

Step 3. Drop column

age (long)	job (string)	marital (string)	education (string)	default (string)	housing (string)
20 - 60 123	11 Categories	4 Categories	7 Categories	2 Categories	3 Categories
56	housemaid	married	basic.4y	no	no
57	services	married	high.school	unknown	no
37	services	married	high.school	no	yes
40	admin.	married	basic.6y	no	no
56	services	married	high.school	no	no
45	services	married	basic.9y	unknown	no
59	admin.	married	professional.course	no	no
41	blue-collar	married	unknown	unknown	no
24	technician	single	professional.course	no	yes
25	services	single	high.school	no	yes
41	blue-collar	married	unknown	unknown	no
25	services	single	high.school	no	yes
29	blue-collar	single	high.school	no	no

☑ Show in-column visualizations for the first 2,000 rows. Visualize the full dataset, Run Data quality and insights report

Steps

1. S3 Source
2. Data types
3. Drop column

Move, drop, duplicate or rename columns in the dataset. Learn more.

Transform: Drop column (Required)

Columns to drop (Required): cons.conf.idx, cons.price.idx, emp.var.rate, nr.employed, pdays, + 2 items selected

Thank you for attending this session