AWS BUILDERS KOREA PROGRAM SPECIAL

오픈소스 기반 매니지드 서비스를 활용한 데이터플랫폼 만들기

Lambda Architecture Data Platform

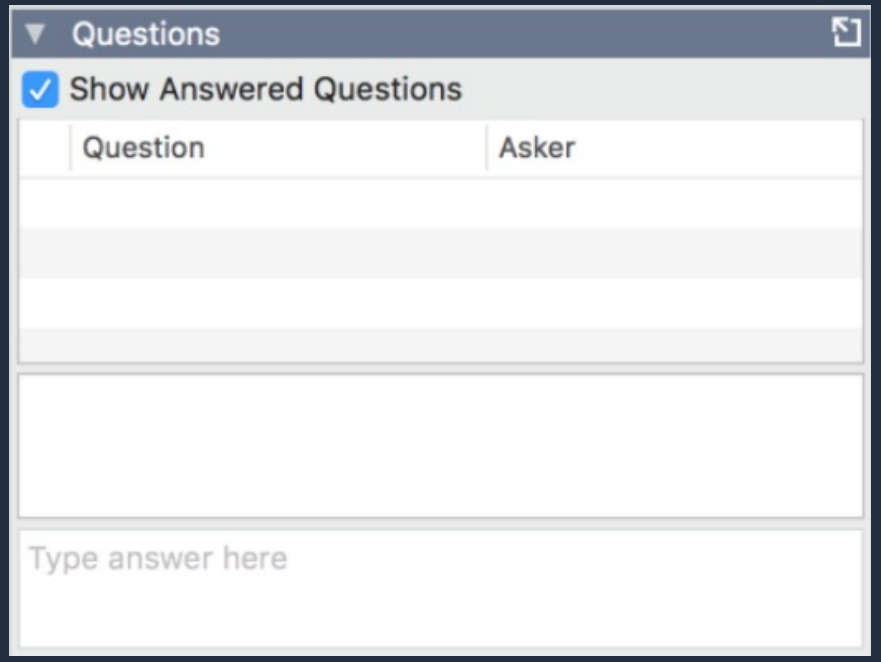
박 현 수

Solutions Architect
AWS

강연 중 질문하는 방법

AWS Builders Korea Go to Webinar **“Questions (질문) ”** 창에 자신이 질문한 내역이 표시됩니다. 본인만 답변을 받고 싶으실 경우 (비공개)라고 하고 질문해 주시면 됩니다.

질문 주신 사항에 대해서는 질문창을 통해 답변을 드립니다.



고지 사항 (Disclaimer)

본 컨텐츠는 고객의 편의를 위해 AWS 서비스 설명을 위해 온라인 세미나용으로 별도로 제작, 제공된 것입니다. 만약 AWS 사이트와 컨텐츠 상에서 차이나 불일치가 있을 경우, AWS 사이트(aws.amazon.com)가 우선합니다. 또한 AWS 사이트 상 에서 한글 번역문과 영어 원문에 차이나 불일치가 있을 경우(번역의 지체로 인한 경우 등 포함), 영어 원문이 우선합니다.

AWS는 본 컨텐츠에 포함되거나 컨텐츠를 통하여 고객에게 제공된 일체의 정보, 콘텐츠, 자료, 제품(소프트웨어 포함) 또는 서비스를 이용함으로 인하여 발생하는 여하한 종류의 손해에 대하여 어떠한 책임도 지지 아니하며, 이는 직접 손해, 간접 손해, 부수적 손해, 징벌적 손해 및 결과적 손해를 포함하되 이에 한정되지 아니합니다.

실습 시작 전 준비 사항

AWS 계정으로 시작

1. 실습 전 계정을 꼭 신청해주세요 : https://portal.aws.amazon.com/billing/signup#/start

2. AWS 계정이 없으신 경우, 행사 참여 전에 미리 AWS 계정 생성 가이드를 확인하시고 AWS 계정을 생성해 주시길 바랍니다.

   *AWS 계정 생성 가이드: https://aws.amazon.com/ko/premiumsupport/knowledge-center/create-and-activate-aws-account/

3. 검증된 호환성을 위하여 실습 시 사용할 웹 브라우저는 Mozilla Firefox 또는 Google Chrome Browser로 진행 부탁드립니다.

실습 마무리 및 설문 참여 방법

- 실습이 모두 끝난 후에는 **자원 삭제**를 잊지 마세요. 직접 준비하신 AWS 계정으로 실습을 진행하신 고객 분들의 경우, 가이드에 따라 자원 삭제를 진행하셔야 합니다. 또한, 기존에 사용하시던 자원이 있으신 고객 분들의 경우, **오늘 생성한 자원만 삭제**하는 것에 주의 부탁드립니다.

- **가이드:** (세션별 제공)

- 마지막으로 세션이 끝난 후, **GoToWebinar 창을 종료하면 설문 조사 창**이 나옵니다.
  이때, **설문 조사를 진행해 주셔야 AWS 크레딧**(1인당 $50 크레딧, 전체 세션당 1회 제공) 을 제공받으실 수 있습니다.

AWS는 고객 피드백을 기반으로 의사 결정을 수행하며 이러한 피드백은 추후에 진행할 세션 방향을 결정합니다.
더 나은 세션을 위하여 여러분들의 소중한 의견을 부탁드립니다.

감사합니다.

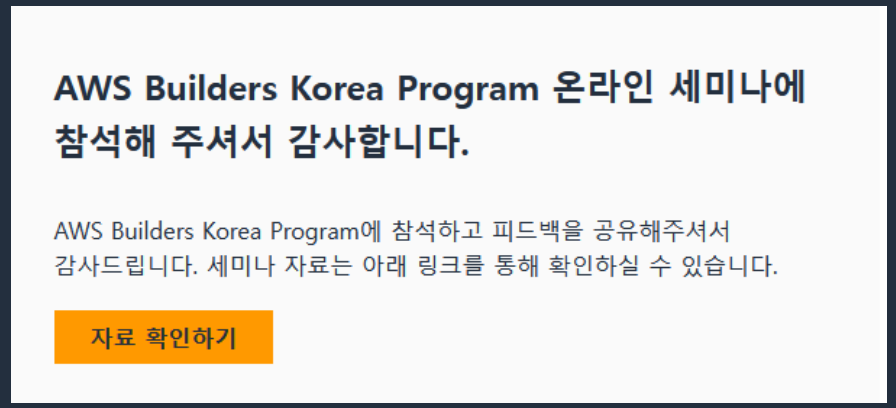
크레딧 안내

- AWS 계정으로 시작하실 경우, **금일 실습에서 발생하는 비용은 당월 과금이 되는 점** 미리 확인 부탁 드립니다.
- 웨비나 종료 후 **설문 조사에 참여해주신 분들께는 AWS 크레딧 바우처 (1인당 $50 USD 크레딧, 전체 세션당 1회 제공)**를 드립니다.
- 해당 **AWS 크레딧**은 등록하신 이메일 계정으로 **행사 종료 후 1개월 내** 발송 드릴 예정이며, 전달 받은 AWS 크레딧은 바로 사용 가능합니다.

감사 메일 & 참석 증명서

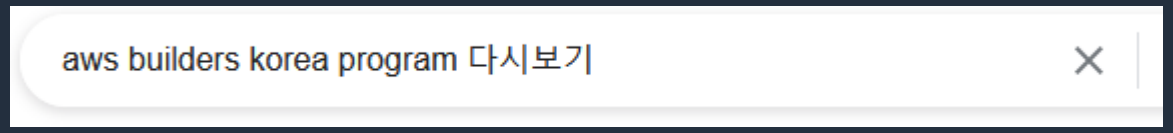
- AWS Builders Korea 세션에 참석해 주신 분들께 행사 종료 후 1개월 내 감사메일과 참석 증명서가 순차 발송됩니다.
- 등록 진행 후 참석하지 않으실 경우 별도 메일 및 증명서는 발급되지 않습니다.

감사 메일 예시



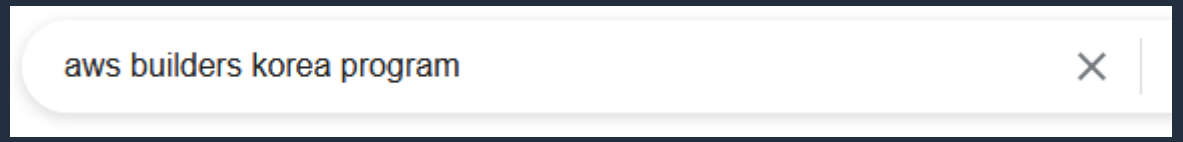
참석 증명서 예시

강연 다시보기



https://kr-resources.awscloud.com/aws-builders-korea-program

AWS Builders Korea 프로그램 정보



https://aws.amazon.com/ko/events/seminars/aws-builders/

Data Platform?

급격한 데이터 볼륨의 증가

상상하는 것 보다 훨씬 많은 데이터

Data

grows

>10x

every 5 years

Data platforms need to

live for

15

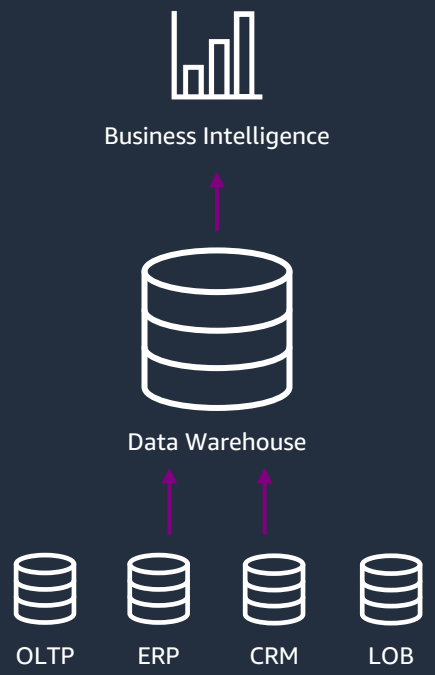
years

scale

1,000x

* IDC, Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data, Focus on the Data That's Big, April 2017.

Traditional Data Platform



전통적인 데이터 분석

관계형 데이터

데이터 로드 전에 정의된 스키마

운영 보고서 및 ad hoc 쿼리

GBs~TBs 스케일

대규모 초기 설비 투자 + $10K-$50K / TB / Year

Modern Data Platform

DW 아키텍처의 확장 또는 발전

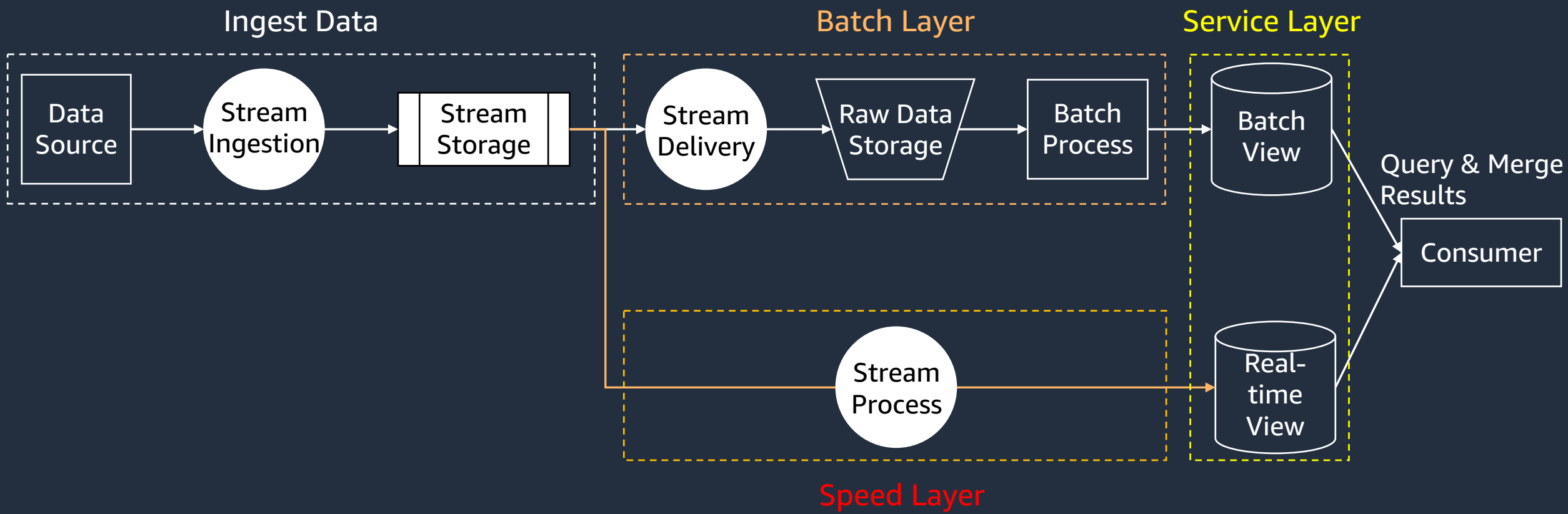
모든 데이터를 모든 형식으로 저장

내구성, 가용성 및 엑사바이트 규모

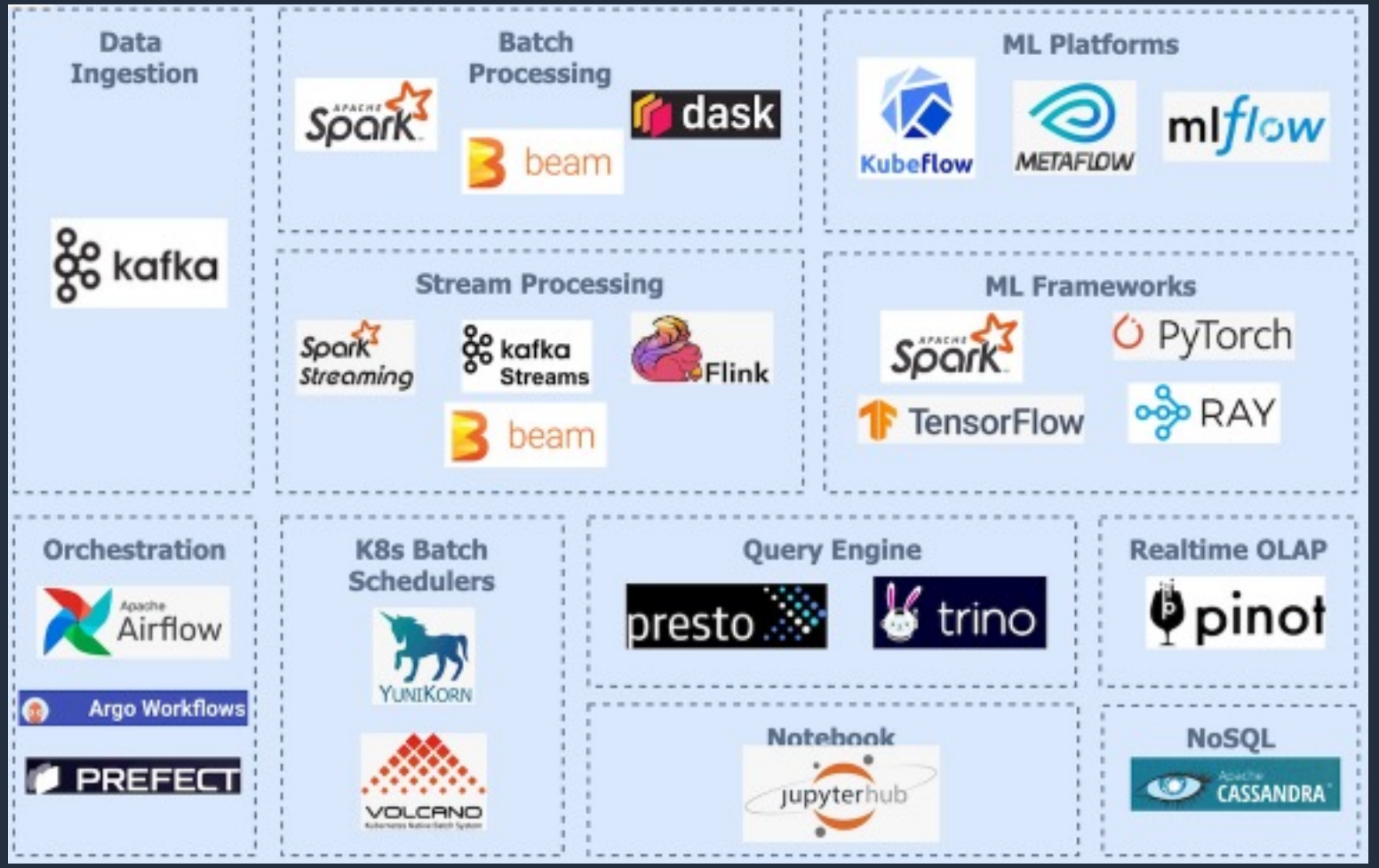
보안, 규정 준수, 감사 가능

DW에서 예측에 이르기까지 모든 유형의 분석 실행

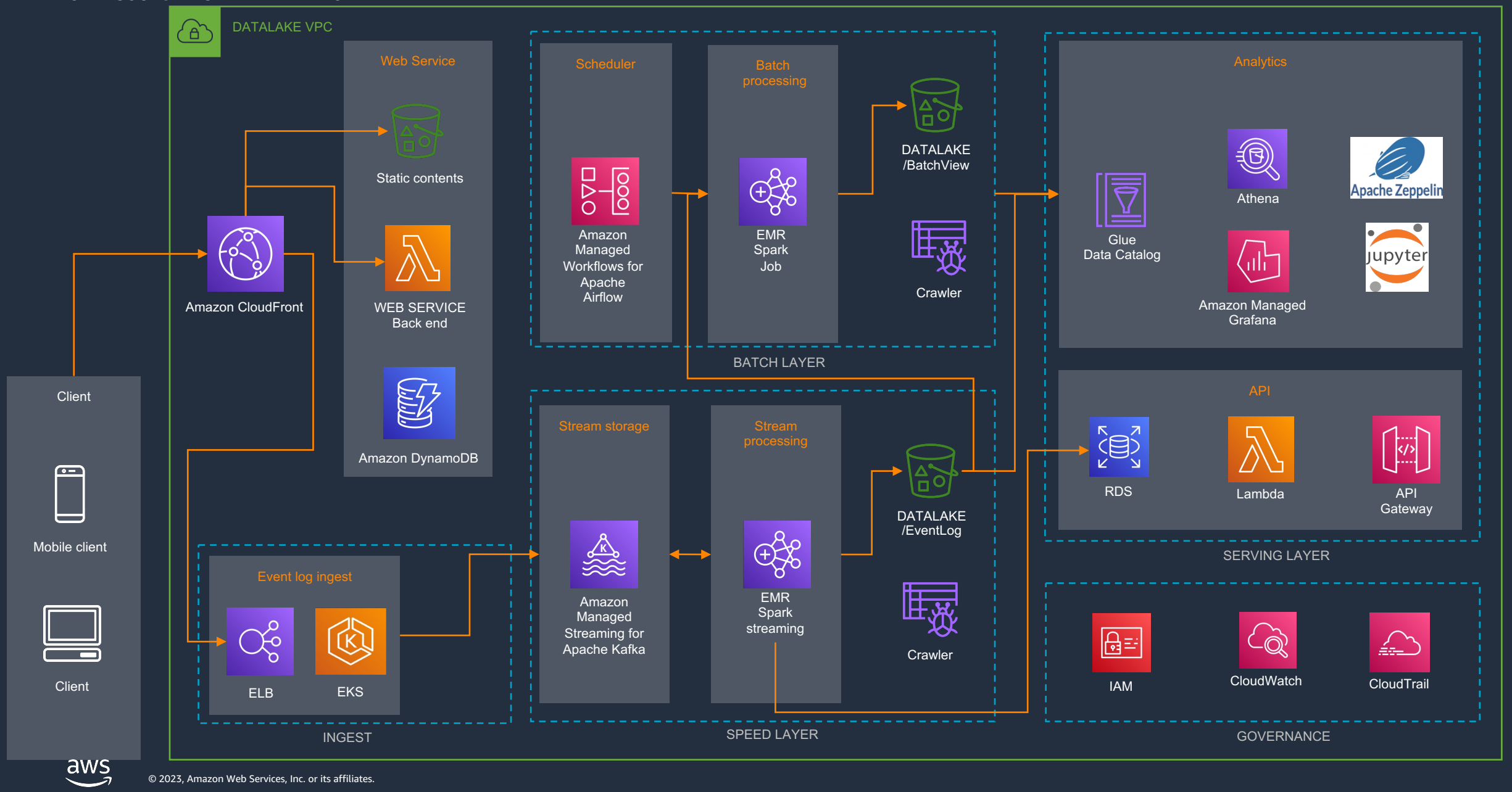
Data platform Lambda Architecture



Importance of Open Source



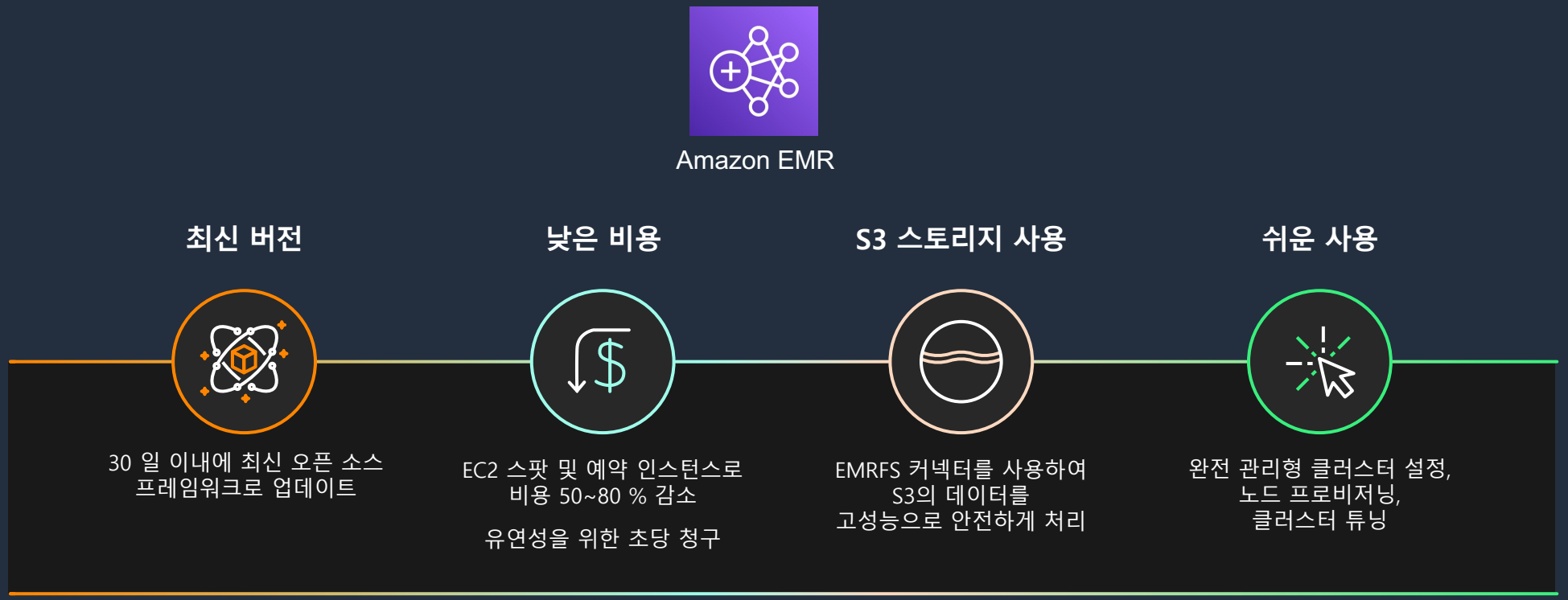
OPEN SOURCE BASED DATA PLATFORM



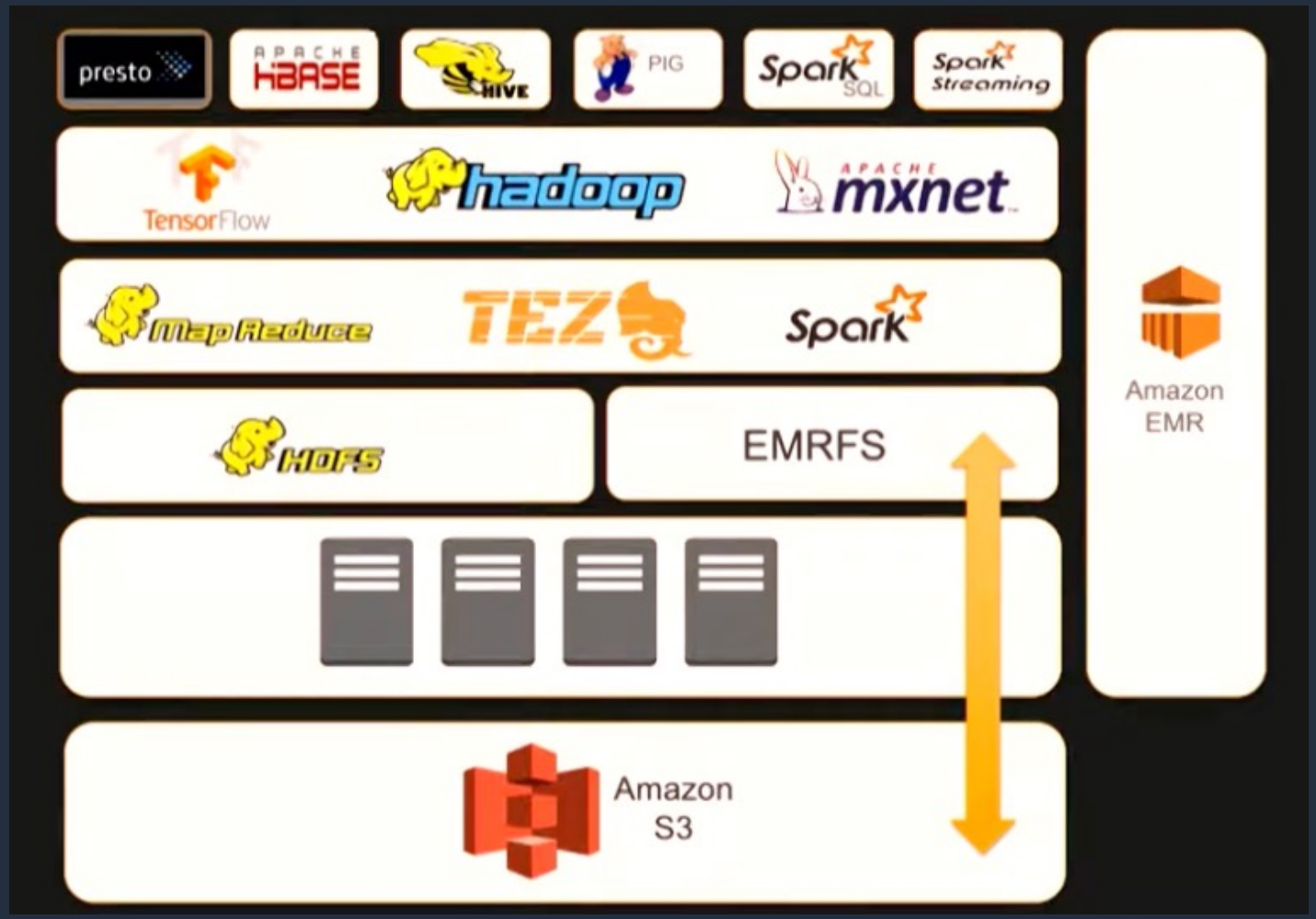
Speed Layer

Speed layer - EMR

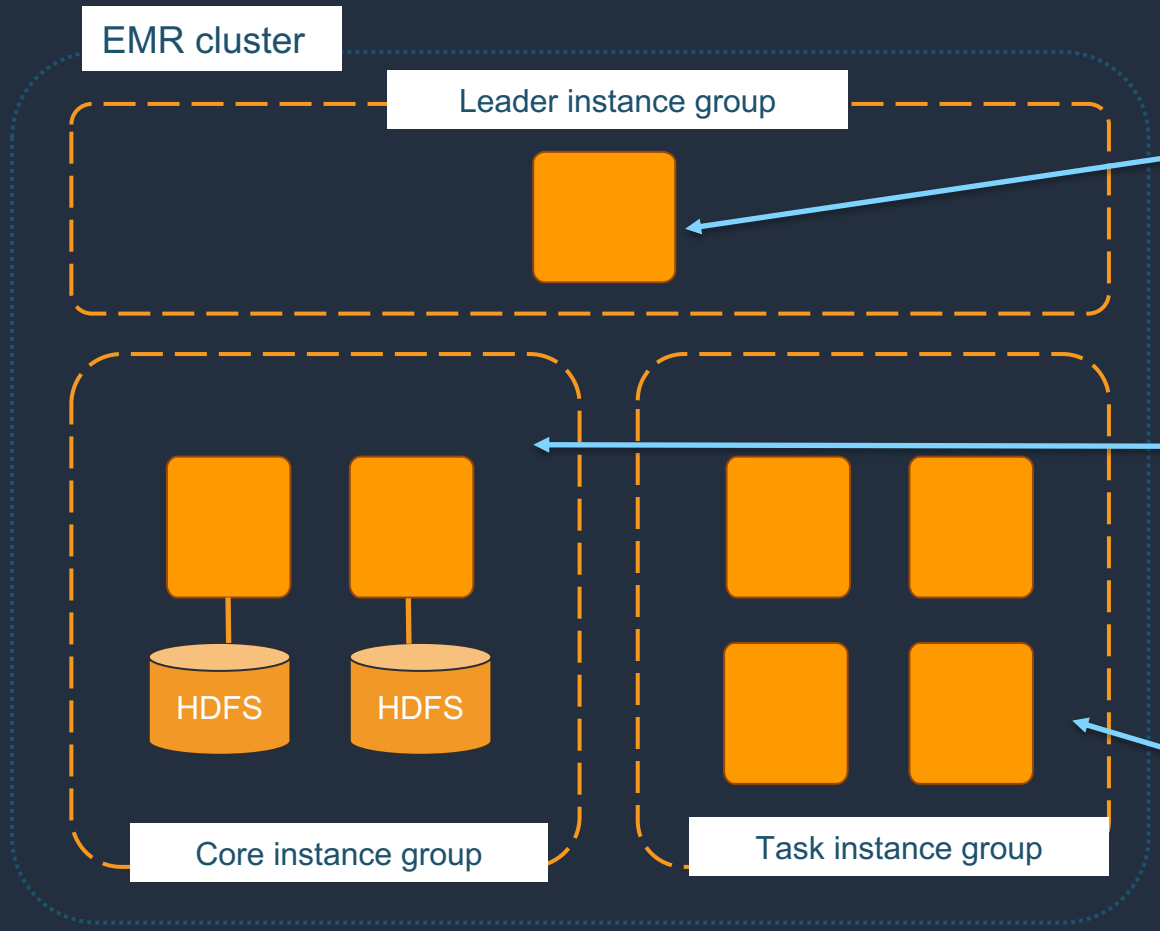
- AWS에서 Hadoop Cluster를 쉽게 올려, Spark, Hadoop, Hive, Presto, HBase 및 기타 빅 데이터 앱을 쉽게 실행



EMR 을 통한 Hadoop Ecosystem의 오픈소스 활용



EMR Cluster Architecture



Leader Node must keep running
(**Leader node (multi leader nodes),** NameNode (HDFS), ResourceManager (YARN), Cluster Manager (Spark), Coordinator (Presto), Act as gateway)

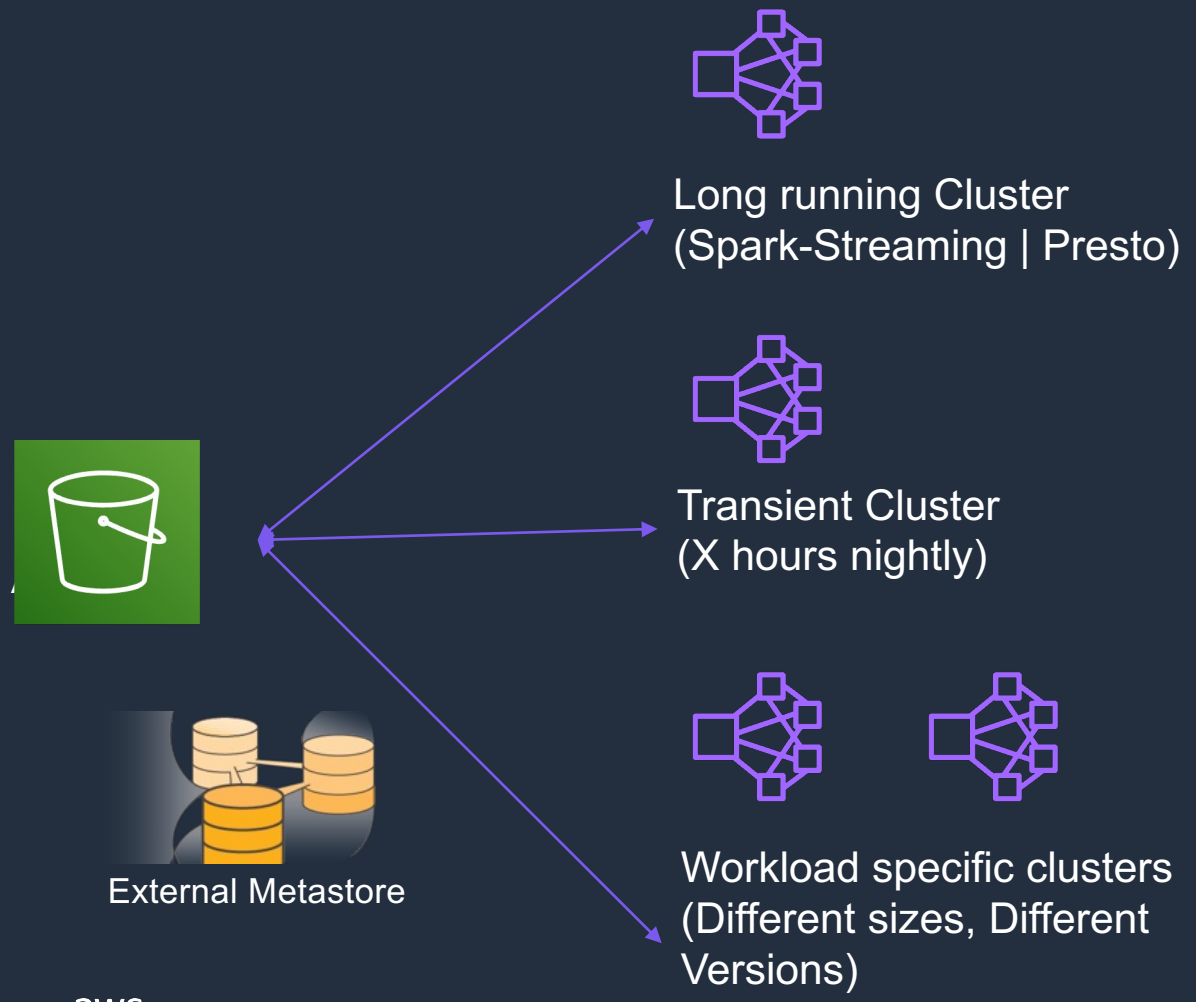
Core nodes can be added and removed gracefully
(**DataNode (HDFS), NodeManger (YARN), Worker (Spark,Presto**)

Cluster can tolerate loss of task nodes.
(**NodeManger (YARN) Worker (Spark, Presto)**)

Spot Instance 는 EMR 인스턴스 플릿에 적합합니다.

- ✓ 노드는 온디맨드 인스턴스와 스팟 인스턴스를 혼합하여 구성할 수 있습니다.
- ✓ 가장 낮은 가격으로 가장 높은 용량의 인스턴스를 찾습니다.
- ✓ 작업 노드의 스팟 인스턴스가 회수되면 플릿의 다른 인스턴스가 이를 대체합니다.

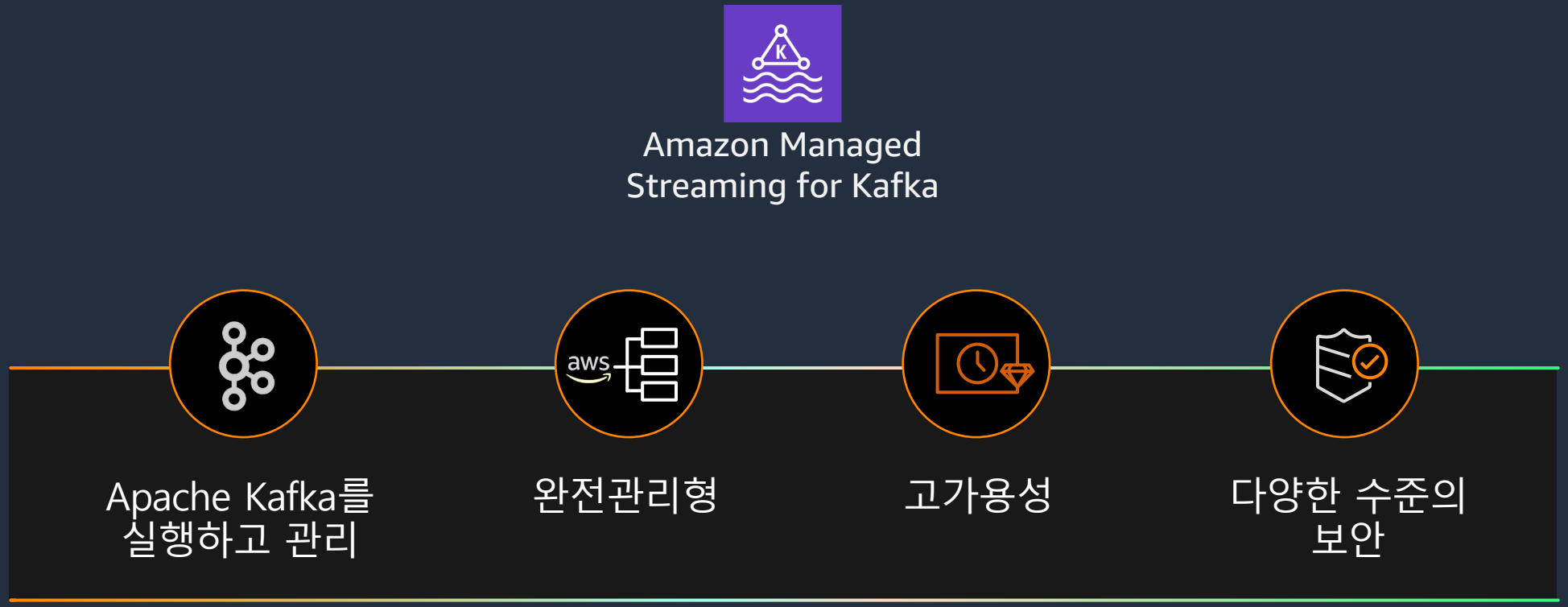
EMRFS 를 통해 Amazon S3를 영구 데이터 스토어로 활용



- EMR 을 통한 Hadoop Ecosystem의 오픈소스 활용(Spark, Hive, Presto, etc.)
- Storage 와 compute 의 디커플링
  - 스토리지 유지를 위해 컴퓨트 노드를 실행해두지 않아도 된다.(unlike HDFS)
  - Amazon EC2 스팟 인스턴스로 Amazon EMR 클러스터를 실행할 수 있다.
  - 이기종 분석 클러스터와 서비스들이 동일한 데이터를 사용할 수 있습니다
- 99.999999999% durability

Speed layer - Amazon Managed Streaming for Apache Kafka

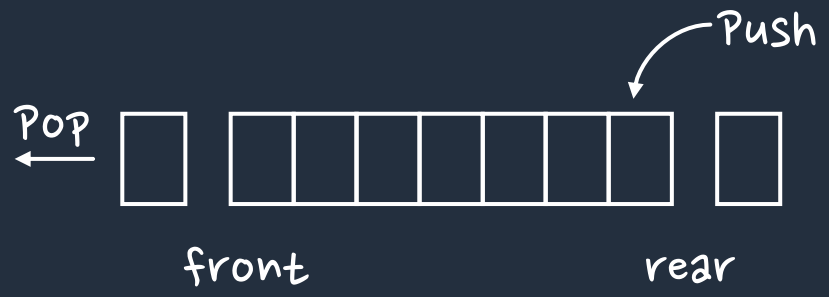
안전한 완전관리형 고가용성 Apache Kafka 서비스



Speed layer - Amazon Managed Streaming for Apache Kafka

안전한 완전관리형 고가용성 Apache Kafka 서비스

- Distributed Queue



#Queue

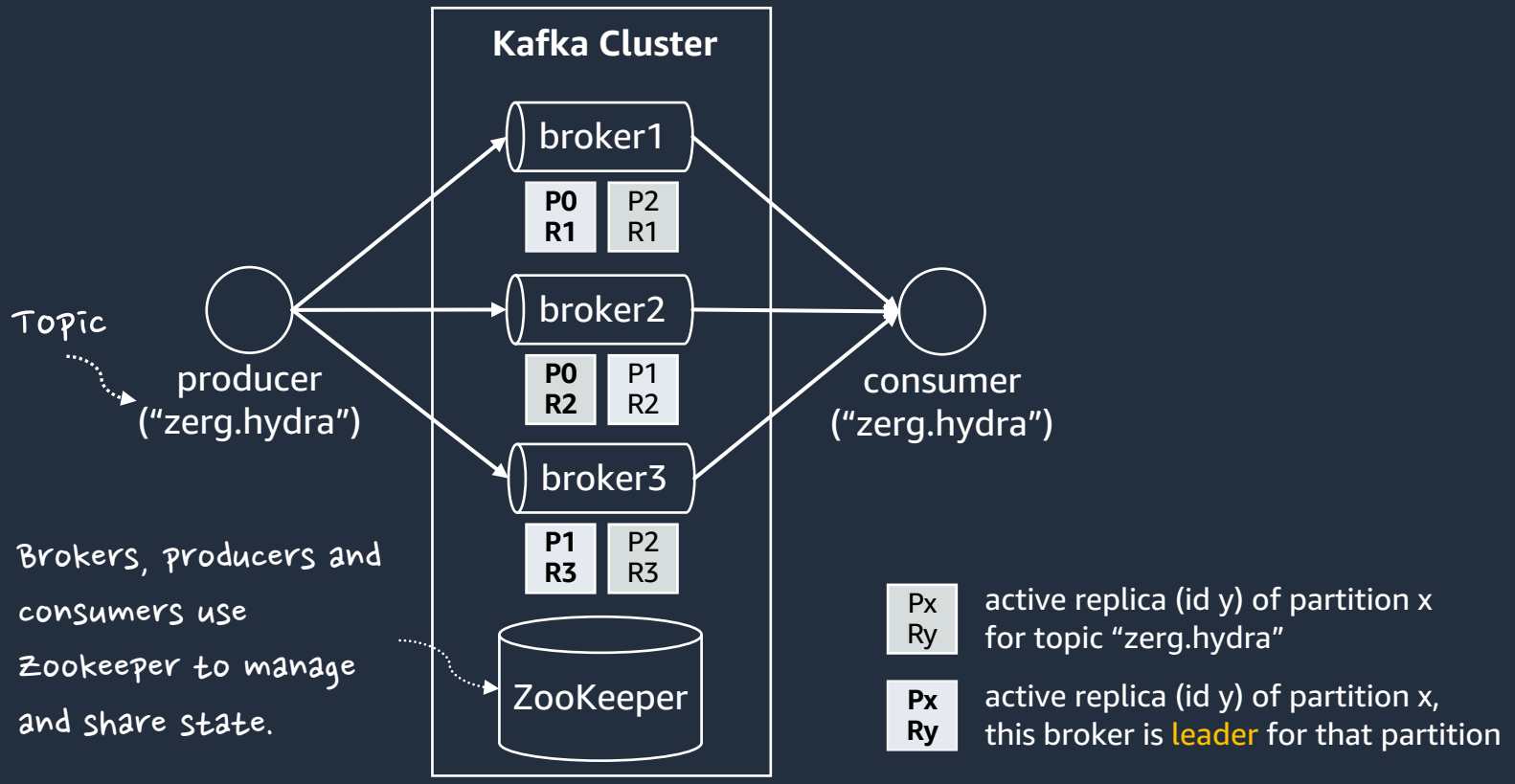
#Distributed

- Stream Storage

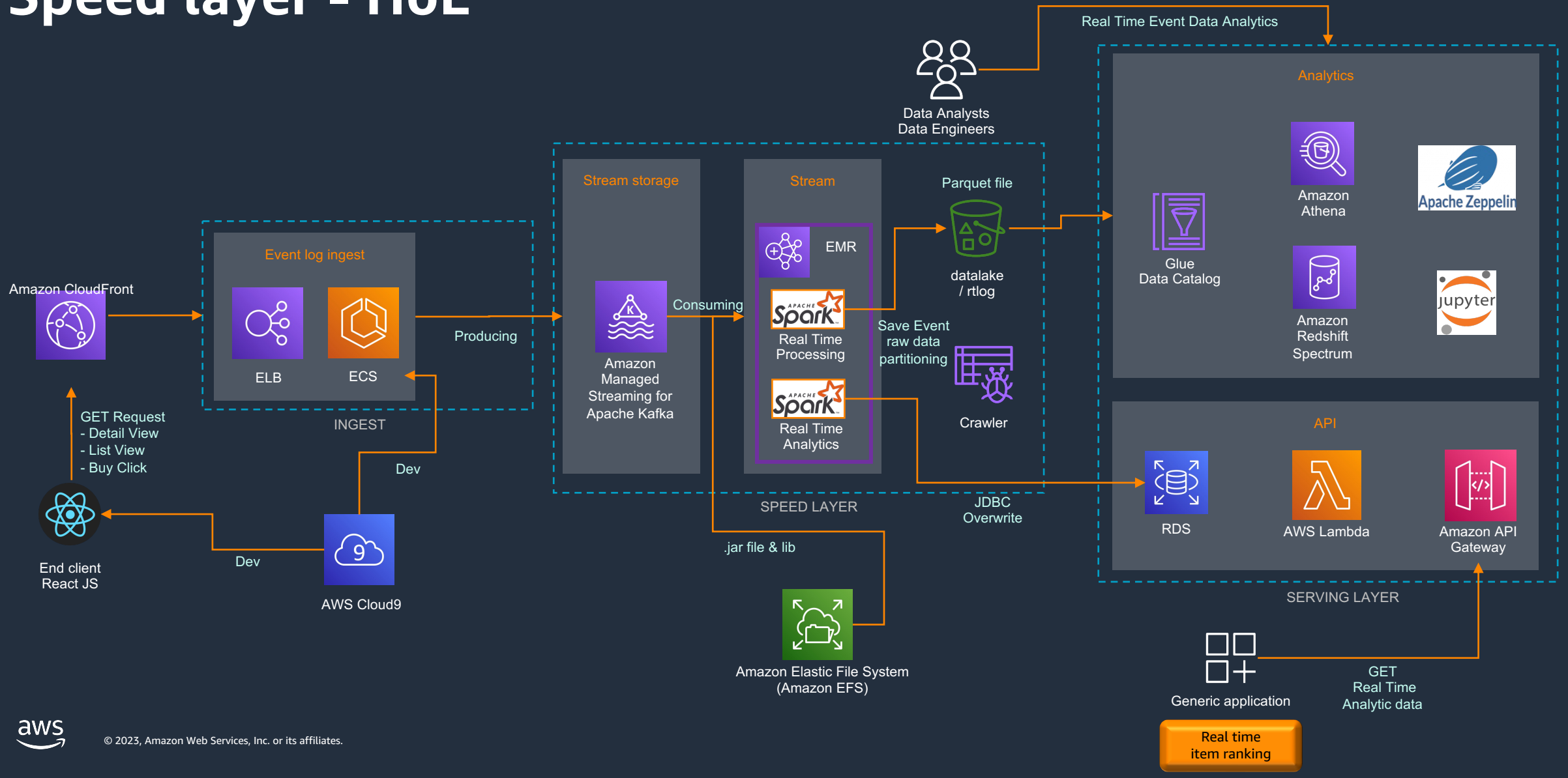
#Storage

Speed layer - Amazon Managed Streaming for Apache Kafka

안전한 완전관리형 고가용성 Apache Kafka 서비스



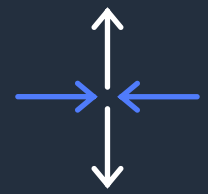
Speed layer - HoL



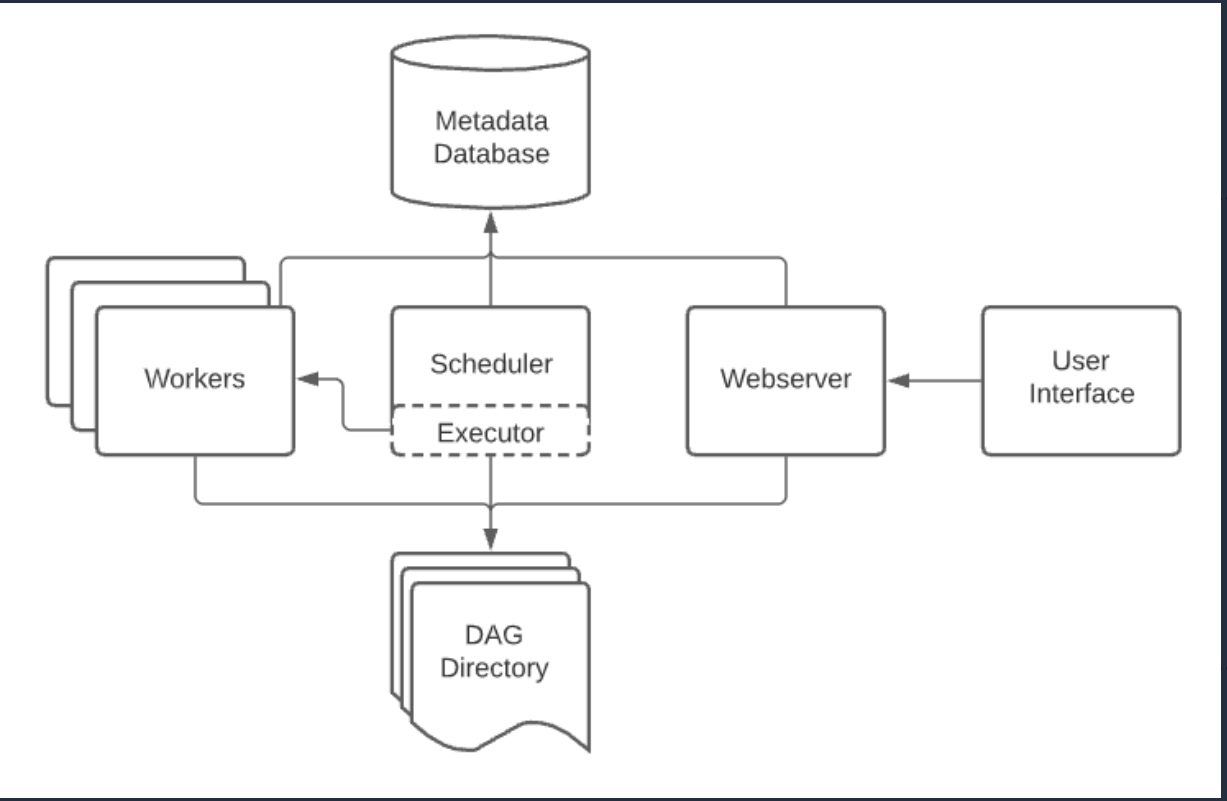
Batch Layer

Amazon MWAA

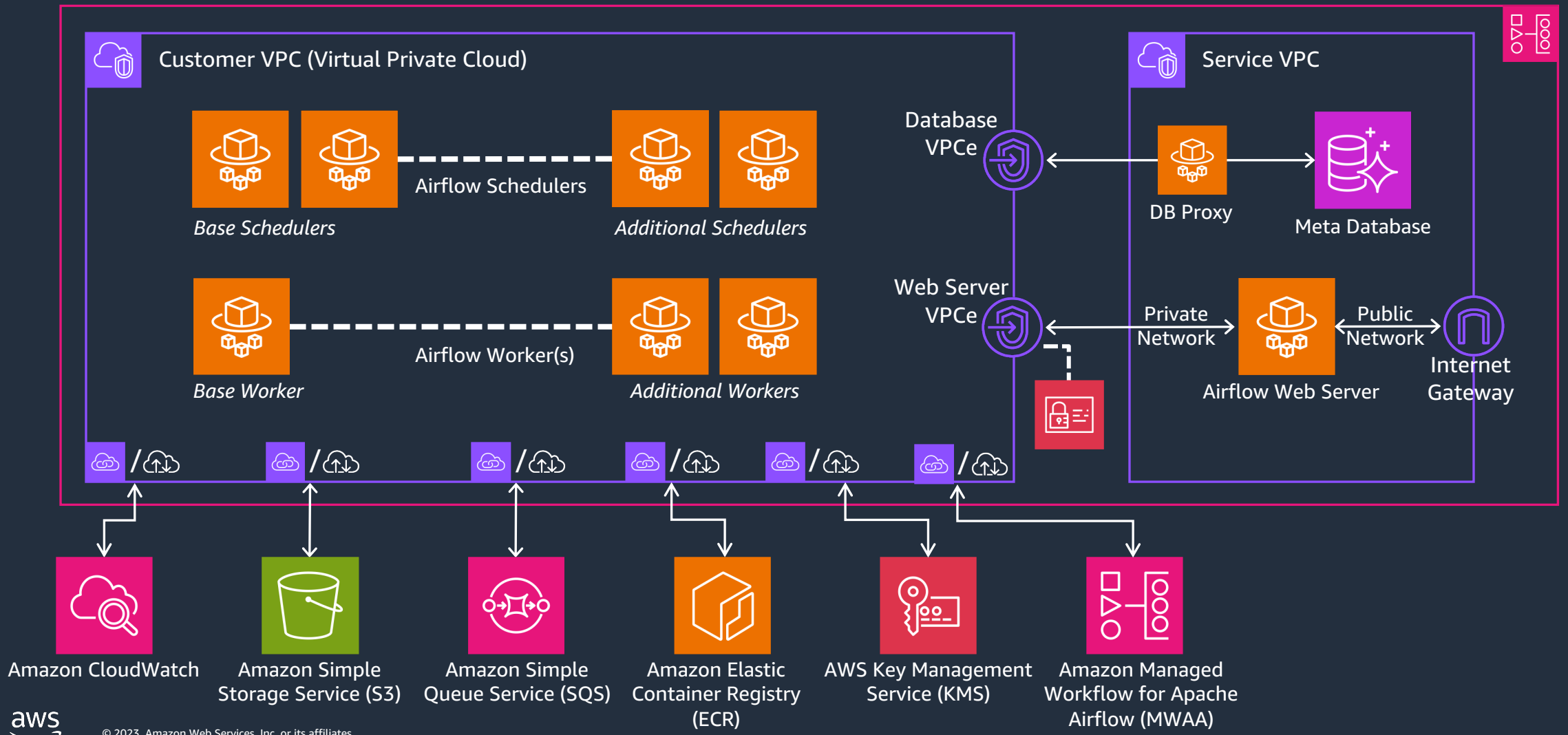
- **Deployments and Operations**
  - Easy to Set Up and Maintain
- **Availability and Sizing**
  - Multi-AZ for HA with Airflow on ECS Fargate
- **Scaling**
  - Auto Scaling and Celery Executor
- **Security**
  - IAM and VPC



Batch layer - MWAA Airflow architecture



Batch layer - MWAA Architecture



Batch layer - MWAA dag

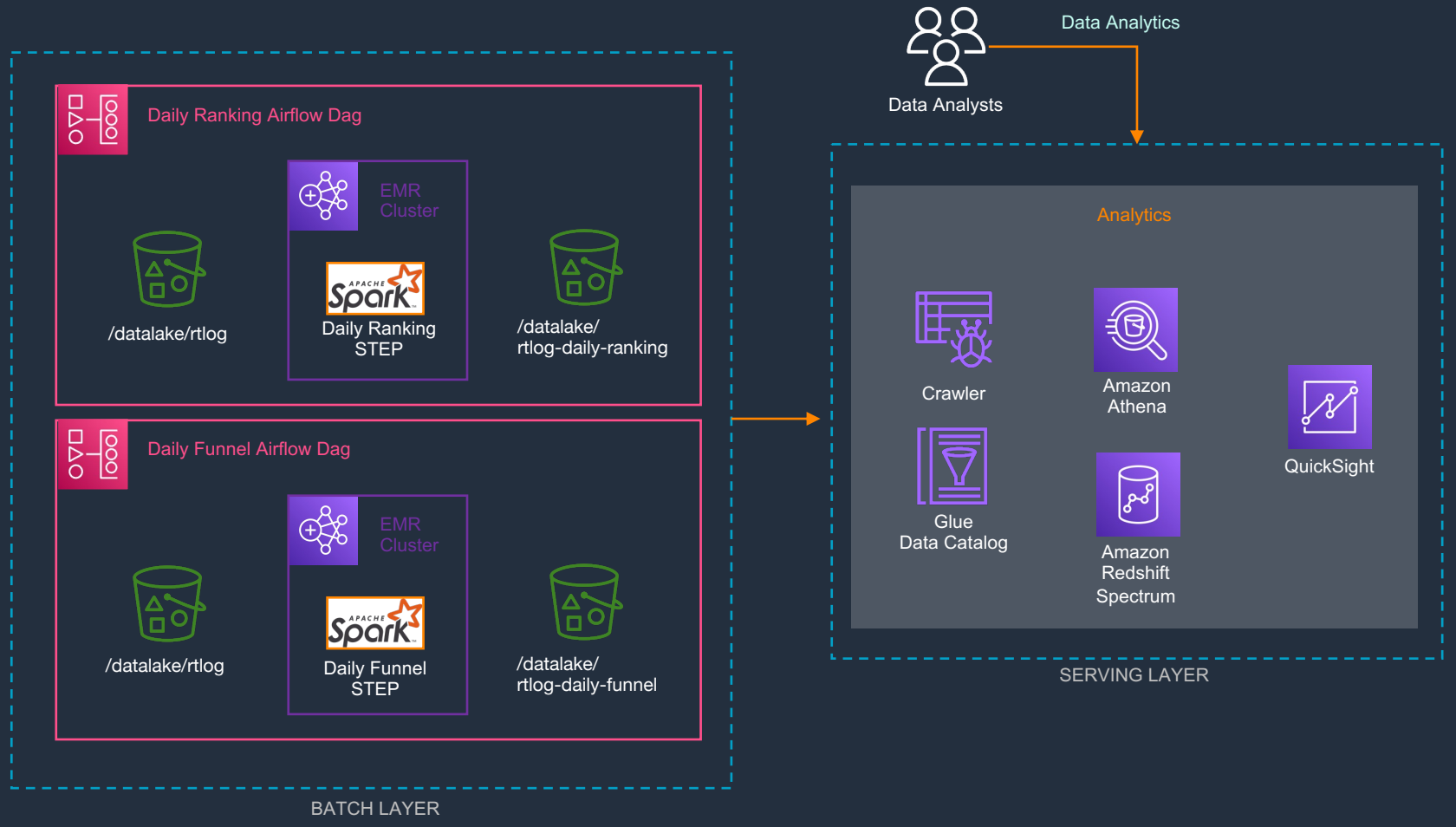
```
dag = DAG(
    'daily-ranking',
    default_args=default_args,
    dagrun_timeout=timedelta(hours=2),
    schedule_interval='0 3 * * *'
)
```

```
cluster_creator = EmrCreateJobFlowOperator(
    task_id='create_emr_cluster',
    job_flow_overrides=JOB_FLOW_OVERRIDES,
    aws_conn_id='aws_default',
    emr_conn_id='emr_default',
    dag=dag
)
# create emr cluster
```

```
step_checker = EmrStepSensor(
    task_id='watch_step',
    job_flow_id="{{ task_instance.xcom_pull('create_emr_cluster', key='return_value') }}",
    step_id="{{ task_instance.xcom_pull('add_steps', key='return_value')[0] }}",
    aws_conn_id='aws_default',
    dag=dag
)
```

```
cluster_creator >> step_adder >> step_checker >> cluster_remover
#dag 실행
```

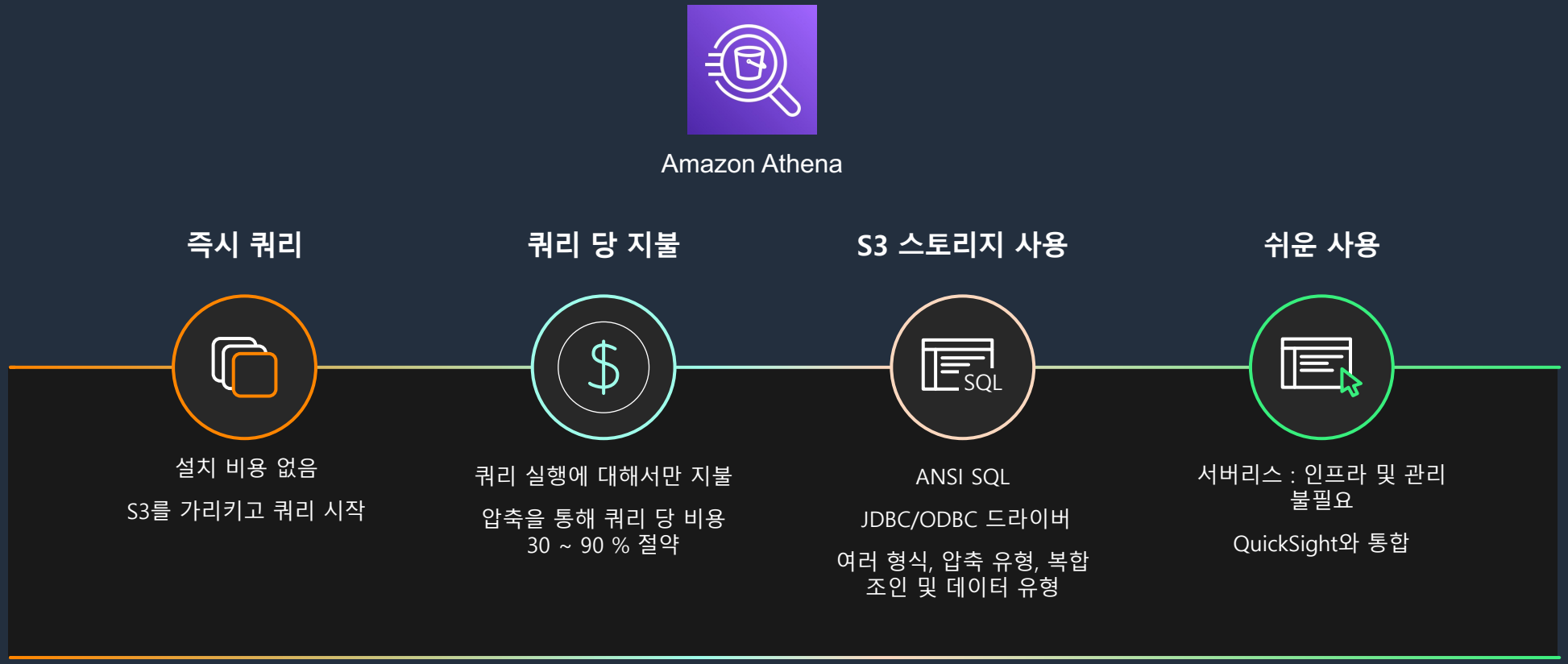
Batch Layer - HoL



Serving Layer

Serving layer- Athena

- 서버리스, 대화식 분석 서비스



Amazon Athena 의 특징 : Serverless Presto

SERVERLESS

PAY PER QUERY

OPEN AND FLEXIBLE

EASY

- ✓ 스토리지와 컴퓨팅 노드 분리
- ✓ Query를 위해 Data Loading / ETL 불필요, S3에서 직접 Query 실행
- ✓ Serverless : 인프라 관리 불필요, 자동 확장, Warm Compute Pools (Multi-AZ)
- ✓ 스캔된 데이터 만큼 과금
- ✓ 보안 : IAM을 통한 인증 / 암호화 : 테이블, Query문, Write Output
- ✓ AWS Glue 데이터 카탈로그와 통합

Serving layer- Athena

Runs standard SQL

- Presto ANSI SQL 지원
- standard data formats 지원
  - CSV
  - Apache Weblogs
  - JSON
  - Parquet
  - ORC
- 복잡한 쿼리 사용 가능
  - Large Joins
  - Window functions
  - Arrays

Serving layer- Athena

Fast Performance for Large Data Sets

- 신속한 adhoc 쿼리 처리
- 쿼리 병렬 실행
- 자동 스케일링
- Amazon Athena Federated Query

Serving layer- Athena

Partitions

- 테이블 내 아무 칼럼이나 지정 가능
- 쿼리 데이터 스캔량을 제한하여 비용 최적화
- Common practice
    - 일, 주, 월 별 파티셔닝
- Example
    - PARTITIONED BY (year STRING, month STRING, day STRING)
    - **df.write.partitionBy(”month").save(”s3://PATH")**

Thank you!