GS Way							
	1. 고객 최우선	□	2. 트렌드 선도	■	3. 최고 지향 목표 설정	■	4. 데이터 중심 의사 결정	■
	5. 신속한 판단과 실행	□	6. 적극적인 소통과 협업	□	7. 비효율 개선	■	8. 기본에 충실	■

데이터레이크 1단계 구축 완료보고

목 차

2023. 4. 5

데이터플랫폼본부
데이터레이크팀

I. 사업 개요(1/4) : 추진 배경

전사 데이터를 단일 플랫폼에서 수집, 적재하고 데이터 분석 환경을 제공하여 데이터 분석 업무의 많은 시간을 소모하는 데이터 탐색 및 획득 시간을 단축하고, 분석 결과로 생성된 데이터를 일관성 있는 인터페이스로 제공하는 전사 플랫폼을 구축하고자 추진

GS리테일 데이터레이크

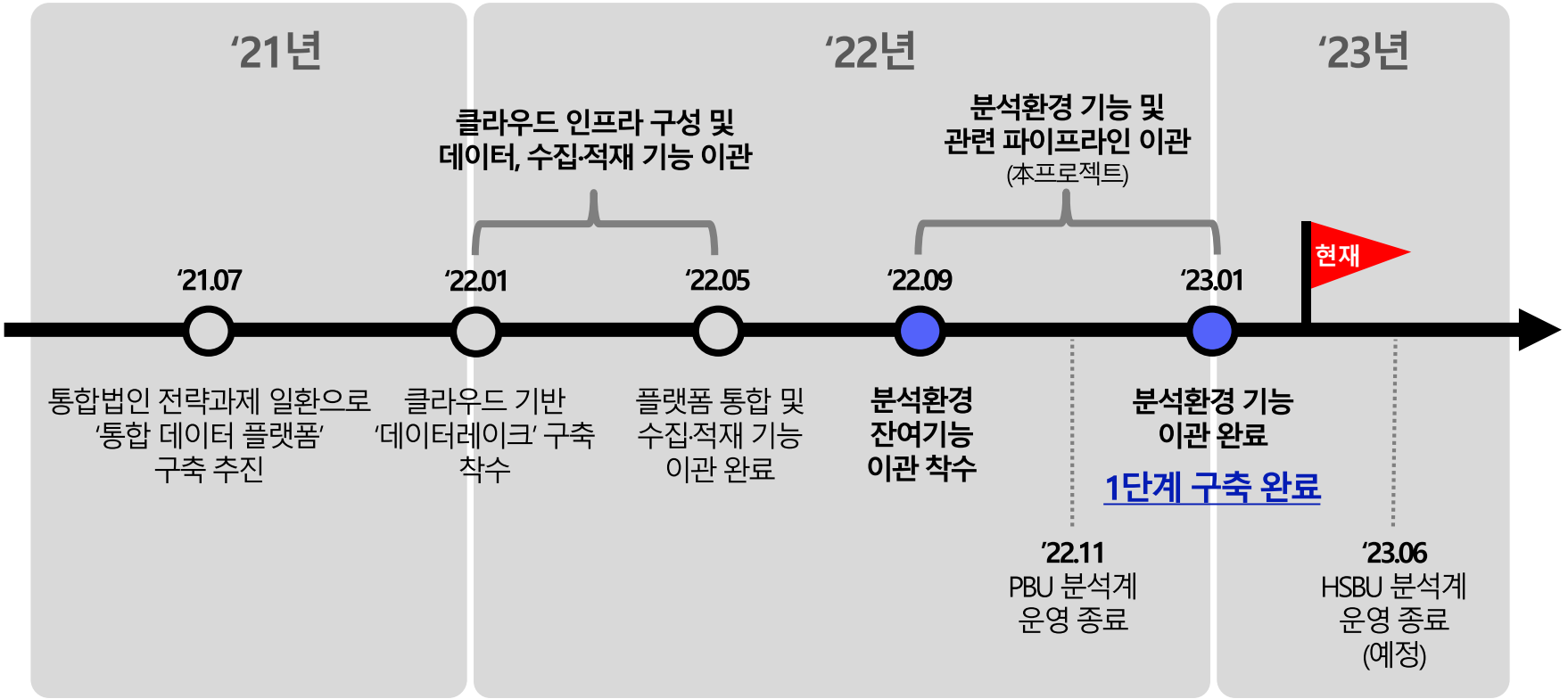
"데이터 분석계 시스템 클라우드 기반 통합 이전"

추진 목적	통합 데이터 수집/적재 인프라 구축	분석용 원천 데이터 이관 및 신규 분석환경 구성	클라우드 이전에 따른 비용절감
추진 목표 및 전략	• 클라우드 기반 통합 분석환경 구축 • 모든 유형의 데이터를 담을 수 있는 유연한 저장소 구축 • 저장 공간과 컴퓨팅 자원의 신축적인 활용		• Hadoop 라이선스 비용 절감 및 클라우드 전환을 통한 운영 비용 개선 • 서비스 확대 및 연계 서비스를 위한 아키텍처의 확장성과 신속한 개발 환경 확보
기대 효과	**1단계(완료)** · 전사 데이터 통합 수집·적재 환경 구성을 통해 향후 새로운 데이터 가치 발굴의 기반 확보 · 클라우드 기반 이관으로 인한 신속한 기능 개발 및 테스트가 가능한 환경 확보	**2단계('23)** · 데이터 활용성 제고 · MLOps 운영체계 확립	**3단계('24~)** · 분석환경 사용자 확대 · 전사 데이터 hub

I. 사업 개요(2/4) : Timeline

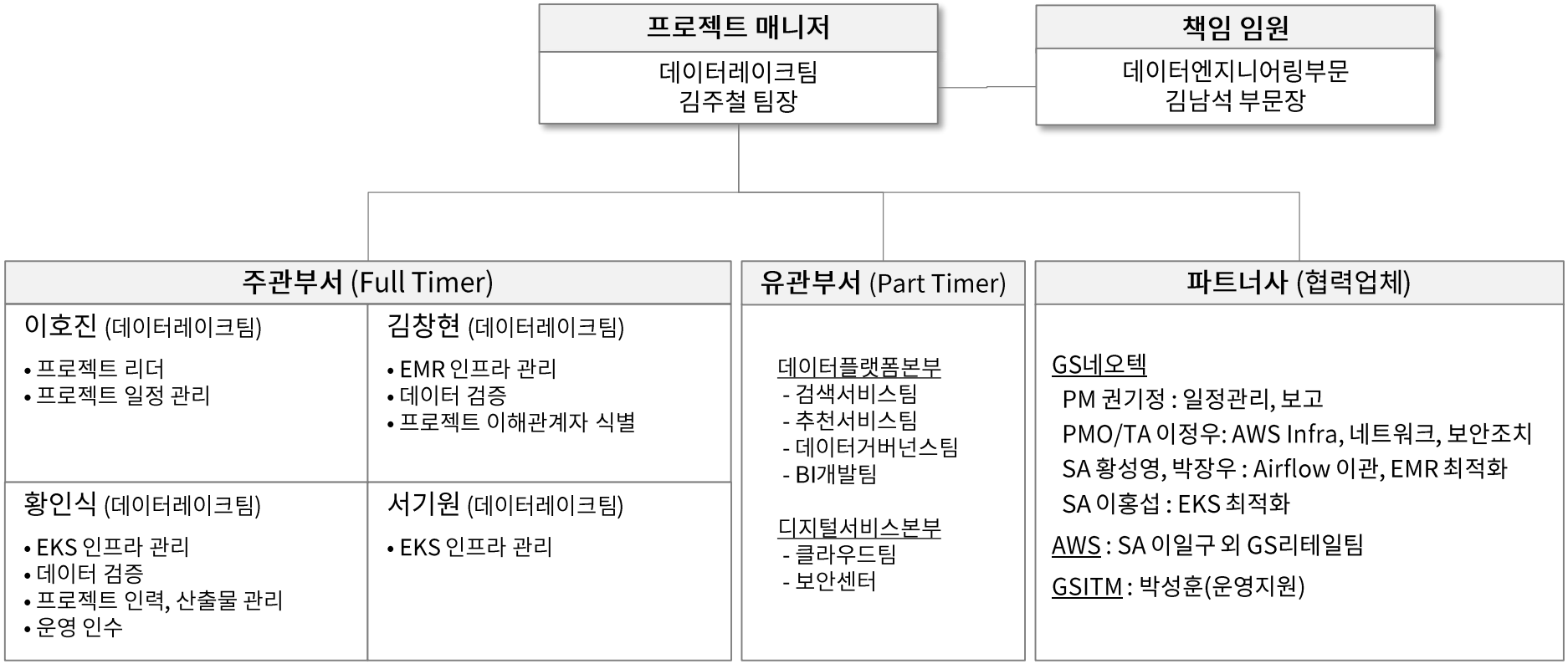
PBU와 HSBU의 분석계 환경을 통합하기 위한 '통합 데이터 플랫폼 구축' 과제에서 출발하여, 데이터레이크 1단계(도입) 구축 수행을 완료함 (AS-IS 분석계 시스템 통합 및 클라우드 기반 이전)

- '22년 5월 PBU 분석계 시스템의 모든 데이터와 기능 이관을 완료하고, 11월 운영 종료
- '22년 9월 HSBU 분석계 시스템의 분석환경 기능 이관 착수, '23년 1월 데이터레이크 이관 완료



I. 사업 개요(3/4) : 프로젝트 조직도

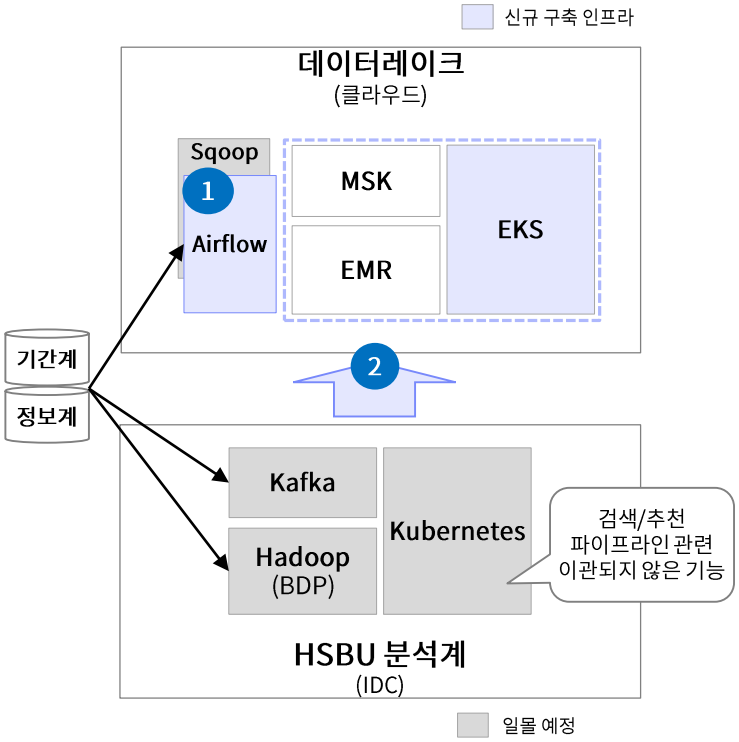
- **주관부서** : 데이터레이크팀
- **파트너사** : 주관 GS네오텍, AWS, GSITM
- **유관부서** : 클라우드팀, 보안센터, 데이터플랫폼본부



I. 사업 개요(4/4) : 파트너사 프로젝트 수행 내역

파트너사는 AS-IS 시스템 분석과 클라우드 기반 PoC를 거쳐, AWS 기반 환경을 구축하고 양 BU의 기존 분석계 시스템의 적재 데이터와 데이터 수집·적재 기능 통합 이관을 수행하였음 (~'22.05)

이후 HSBU 분석계의 잔여기능 이관을 위한 프로젝트에서 검색/추천 데이터 파이프라인 이관을 수행하였음



1 데이터 수집·적재 기술 고도화

◆ **현재 구축된 ETL 기술 전환** (Sqoop → Airflow)

	AS-IS (Sqoop)	TO-BE (Airflow w/Spark)
구현	웹UI 통해 작업. 생산성 낮음 기술은 성숙하나 러닝커브 높음	코드(Python) 기반. 높은 생산성 & 유지보수성 직관적이고 이해하기 쉬운 개발 방식
운영	구현 가능한 워크플로우의 복잡도가 제한적	비교적 유연한 워크플로우 구성 지원
비용	EMR 기반. 현재 클러스터 규모 유지 필요	EMR 독립적. 비용절감 포인트 발굴에 유리

2 검색/추천 파이프라인 클라우드 이관

◆ **On-premise 파이프라인 이관** ('22.5 데이터 수집·적재 기능 구축 범위에서 제외된 기능)
- HSBU 분석계 인프라 : Hadoop, Kubernetes, Kafka
- On-premise 인프라 연내 일몰 예정에 따른 클라우드 전환
- 파이프라인 및 기능 약 70개

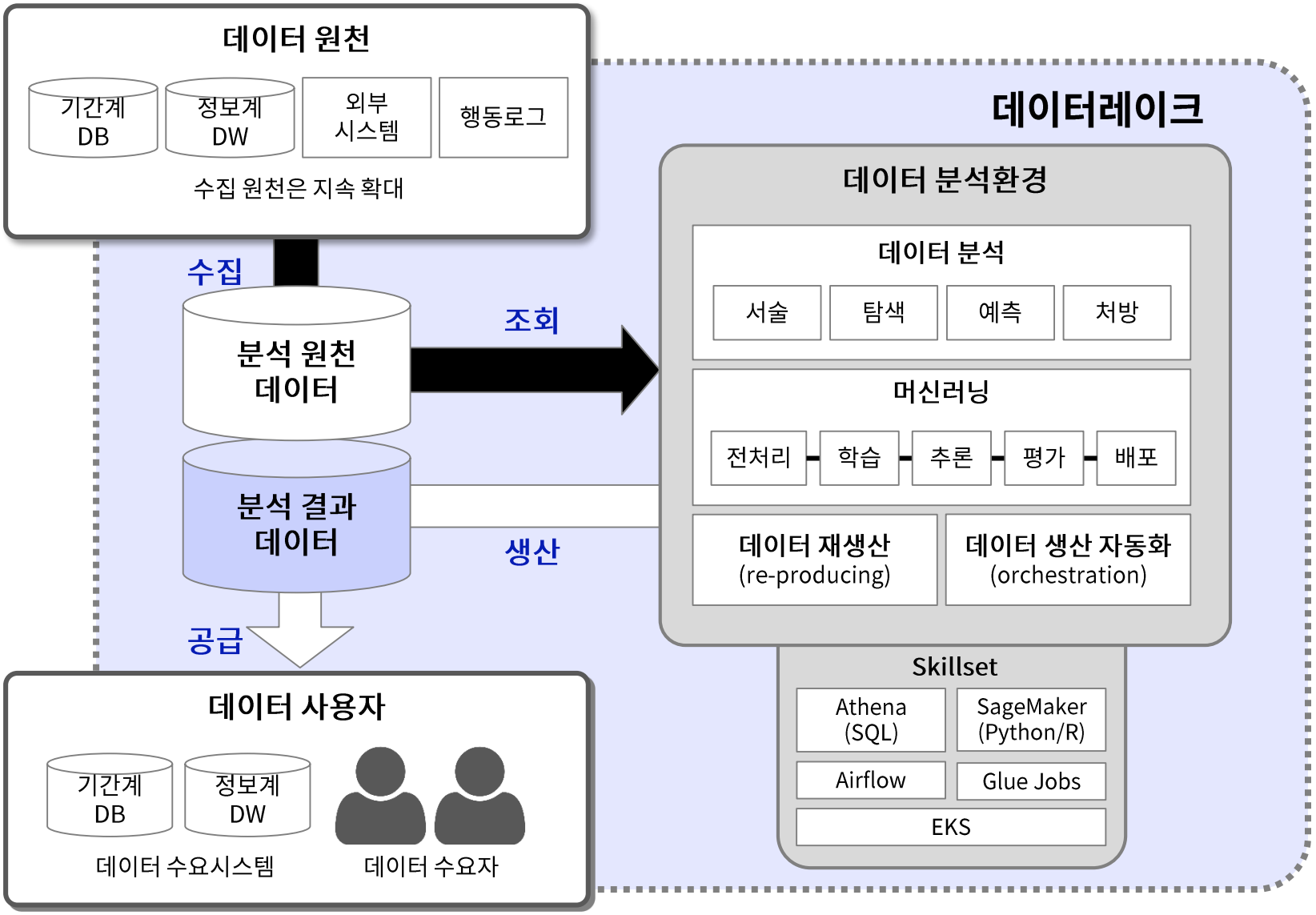
데이터 ETL (20개)
- HDFS to S3
- API (bytedance, glue)
- Kafka to HDFS
  → MSK to S3
- 상품 데이터 이관

검색 (5개)
- 인덱싱
- 검색DB 데이터 이관
- 자동완성 기능
- Spark, Hive, MySQL, Oracle
- Kafka, HDFS, Git, Scala

추천 (40개)
- HDFS to Oracle
  → HDFS to 추천DB
- Docker, S3, Spark, Presto
- Hive, Python, Oracle, Git

데이터 분석 (6개)
- 분석 과제의 결과물 이관
- Docker, R, Spark
- Hive, Oracle, Git

II. 수행 내용(1/5) : 데이터레이크 기반 분석환경



II. 수행 내용(2/5) : 분석 결과 데이터 생산·활용 현황

1. 데이터플랫폼본부 內 데이터 분석/개발 과제

- 편) FF 수요예측 고도화
- 편) 점포 유형화 고도화
- 편) 신규점포 매출 예측
- 편/수) 고객 VOC 분석 고도화
- 전사 고객 데이터 표준화
- 상품 데이터 표준화 및 속성 확대
- 데이터거버넌스 **1**단계 구축 완료**, 2**단계 예정

2. 他**시스템 대상 원천 데이터 제공**

- 상품 검색 & 추천 서비스(상품정보, 행동로그 등) : GSShop, 프레시몰
- PBU 마케팅 캠페인 시스템
- GSShop GIP 대시보드
- 트렌드분석플랫폼 (TrendLens)

II. 수행 내용(3/5) : 데이터레이크 적재 데이터

데이터레이크 적재 데이터는 분석에 활용하기 위한 **분석 원천 데이터**와 분석 사용자가 생성한 **분석 결과 데이터**로 구분할 수 있음

1. **분석 원천 데이터**는 데이터 수요자의 요청에 따라 기간계와 정보계 데이터를 선별하여 복제본을 수집·적재하고 있음 (적재 대상은 요청에 따라 지속적으로 확대)
2. **분석 결과 데이터**는 데이터 사용자 유형(사람 또는 시스템)에 따라 활용 가능한 형태로 정제되고, 적합한 인터페이스를 통해 제공되고 있음

[데이터레이크 內 분석용 원천 데이터 적재 현황] (23.3.3 기준)

구분	원천	테이블수	용량(GB)	설명
내부 시스템	PBU 기간계[1]	346	3,158.9	PBU 운영DB 데이터
	편)점포경영	90	2,457.6	
	편)가맹지원	40	383.5	
	편)나만의냉장고	49	35.9	
	수)통합운영	29	83.0	
	수)캠페인	10	141.9	
	수)오프라인앱	38	15.0	
	통합MD운영	63	27.5	
	통합서비스앱	3	0.9	
	해피콜	18	5.9	
	네트웍스TMS	6	7.8	
	HSBU 기간계[1]	11	16,007.8	HSBU 운영DB 데이터
	GSSHOP	8	15,974.4	WCS로그
	WEBDB	3	33.4	
내부 시스템	프레시몰 기간계[1]	298	574.5	프레시몰 운영DB 데이터
	e커머스BOS	213	332.6	
	e커머스본부	45	12.5	
	e커머스TMS	21	11.7	
	e커머스WMS	16	29.8	
	WCS로그	3	187.9	
	PBU 정보계	575	52,956.3	PBU DW 데이터
	HSBU 정보계	365	5,327.6	HSBU DW 데이터
	검색/추천	266	570.4	검색/추천용 상품 데이터
행동로그	Amplitude	3	247.6	모바일앱[2] 고객 행동로그
	Airbridge	9	20,172.8	
합계		1,873	99,015.9	(약 96.7 TB)

1) (AS-IS) 기간계DB 직접 접근, (TO-BE) 공용 ODS 영역 확보하여 기간계 영향도 억제
2) 우리동네GS, GS프레시몰, GSSHOP

II. 수행 내용(4/5) : 데이터레이크 인프라 기술

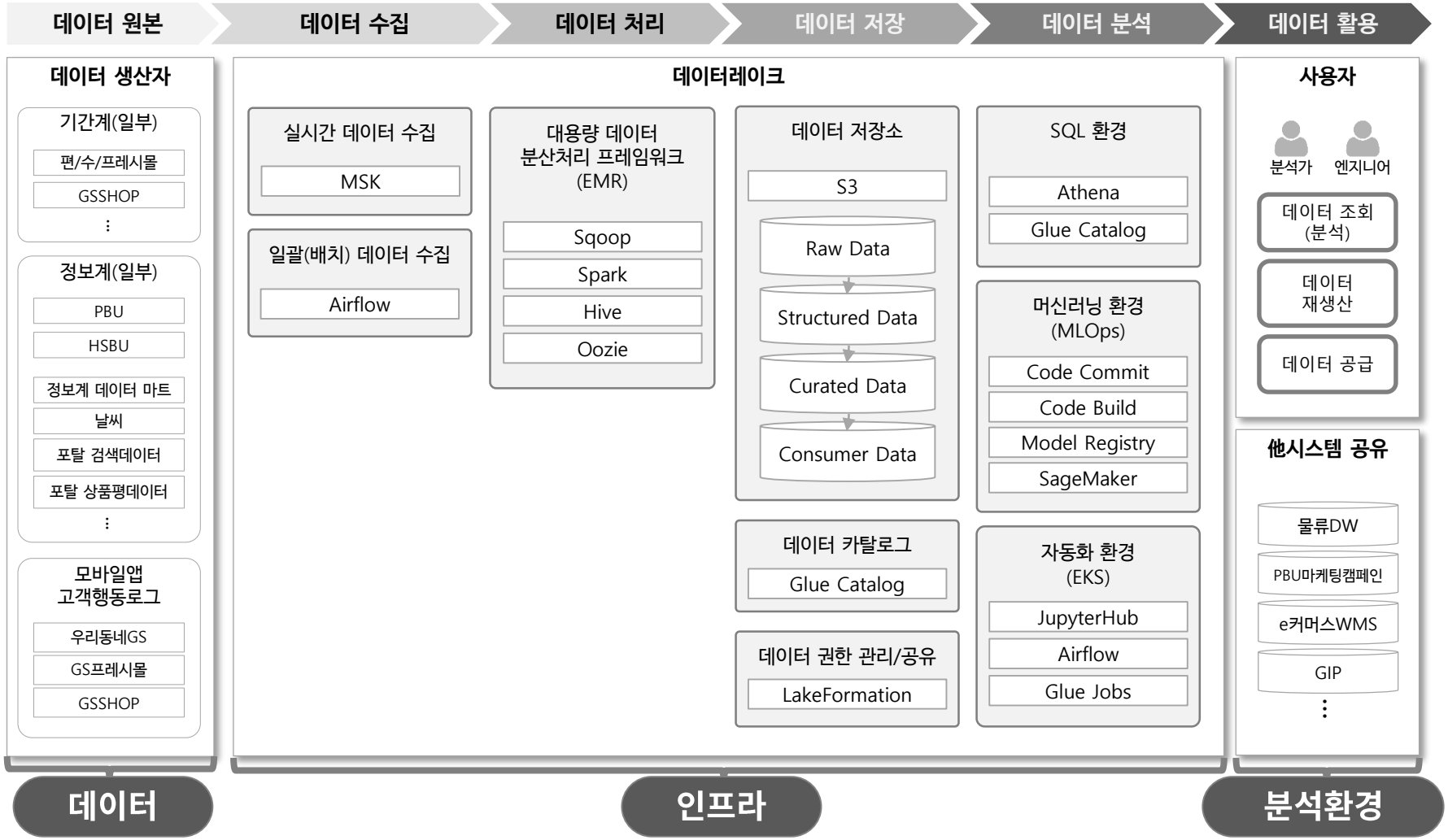
인프라는 데이터레이크 시스템을 구성하는 기술 요소이며, 그 기능에 따라 크게 4가지로 구분할 수 있음

구분	기술	설명
데이터 수집(획득)	Sqoop	대용량 데이터의 주기적인 수집 (배치)
	MSK[1]	실시간 데이터의 수집 (스트리밍)
	Oozie	데이터 수집 기능의 스케줄링과 모니터링을 담당하는 오케스트레이션(orchestration=관리/관제) 도구
	Airflow	
데이터 처리(가공)	Spark	대용량 데이터를 분산 처리하기 위한 프레임워크[2]
	EMR[3]	대용량 데이터를 분산 처리하는 플랫폼 SW
	EKS[4]	다양한 SW 구동을 간단하게 사용할 수 있는 환경을 제공하는 플랫폼 SW
데이터 저장, 공유	S3	무한 용량의 클라우드 저장소 (데이터레이크의 모든 데이터는 S3 파일로 저장)
	Glue Catalog	S3 데이터를 데이터베이스 형태로 관리하는 카탈로그 서비스
	LakeFormation	AWS에 저장한 데이터의 접근 권한을 관리하고, 데이터 공유 기능을 제공하는 서비스
데이터 분석	Athena	S3 데이터를 SQL을 통해 분석하는 기능 제공
	SageMaker	AWS에서 python을 실행할 수 있는 환경 제공

1) MSK(Managed Service of Kafka) - AWS에서 제공하는 Kafka 서비스
2) 프레임워크(framework) - 개발자가 기능을 구현할 수 있는 기반을 제공하는 소프트웨어
3) EMR(Elastic MapReduce) - AWS에서 제공하는 Hadoop 서비스
4) EKS(Elastic Kubernetes Service) - AWS에서 제공하는 Kubernetes 서비스

II. 수행 내용(5/5) : GS리테일 데이터레이크 구성도

분석 원천 데이터 수집(파이프라인) 및 기본적인 분석환경(SQL, 머신러닝) 등 분석계의 기반 기능을 클라우드에 구축하여, 향후 데이터의 재생산-공급 기능 고도화 및 분석환경 개선을 빠르게 수행할 수 있는 발판을 마련하였음

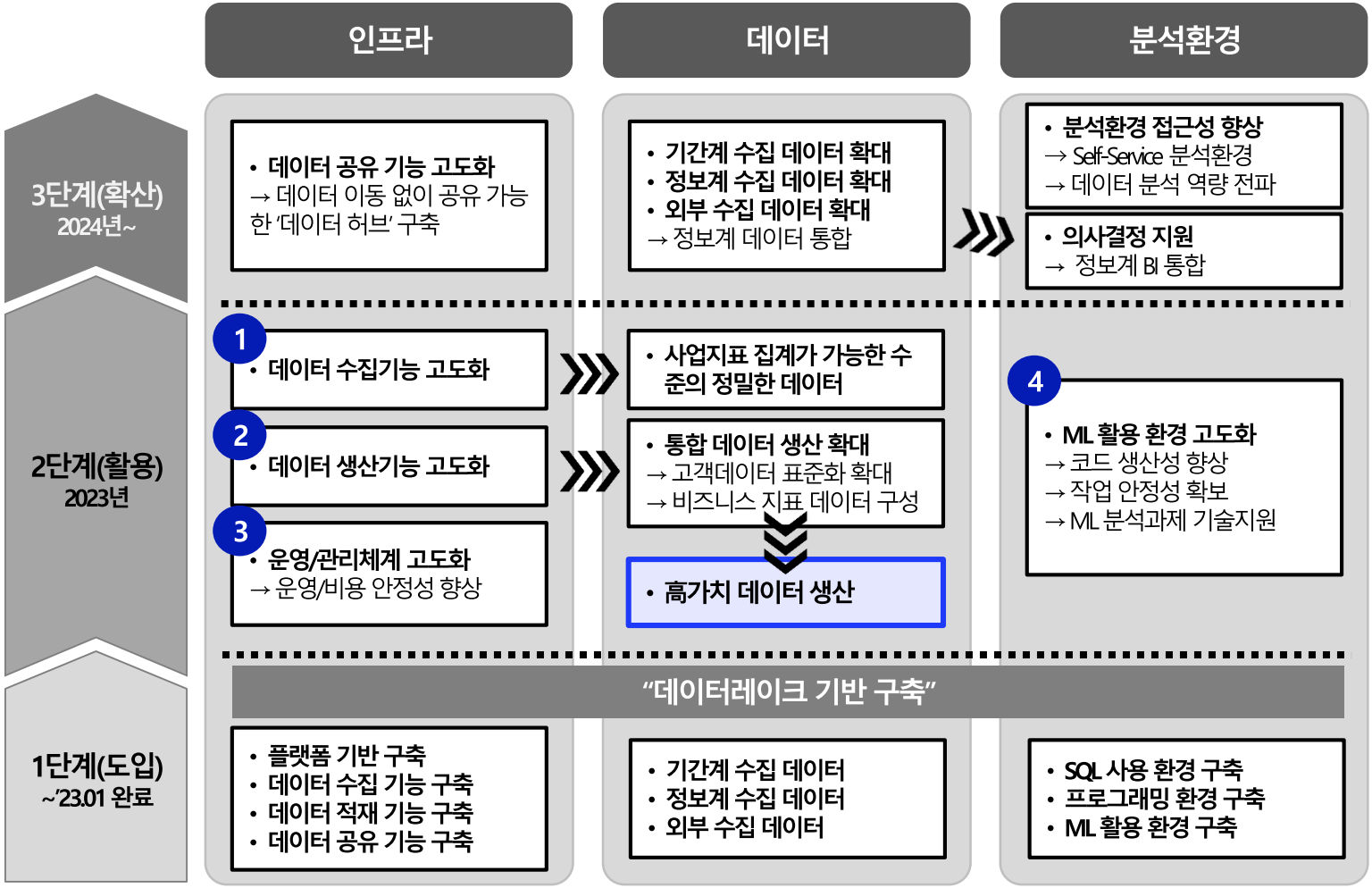


1. **전사의 모든 데이터에 접근할 수 있는 통합 데이터 저장소 및 분석환경의 기반을 구축하여, 전사 통합 관점의 새로운 데이터 개발에 활용할 수 있는 시스템을 확보**
   - ➢ PBU와 HSBU의 기간계 DB, 정보계 DW에서 분석 원천 데이터를 획득할 수 있는 기반을 갖추어, 분석 수요에 따라 필요한 데이터의 복제본을 수집하고 데이터레이크에 적재하여 분석에 활용 가능
   - ➢ 향후 전사 통합 view의 새로운 데이터를 개발하여 수요처에 공급할 수 있는 기술적 기반을 확보함
   - ➢ 데이터플랫폼본부에서 수행하는 데이터 분석·개발·거버넌스 활동의 기반 환경 제공

2. **클라우드 기반으로 분석계의 모든 기능 이관을 완료하여, 향후 플랫폼 인프라 개선 과정에서 적은 인력으로 빠르게 기술 테스트 및 도입을 실행할 수 있는 환경적 이점을 확보**
   - ➢ 분석계를 구성하는 IDC 서버 46ea(/총 58ea)의 클라우드 전환 및 일몰 완료
   - ➢ IDC 기반 서버 관리에 소요되는 운영 인력이 대폭 감소하여, 데이터레이크 구현에 필요한 기술 역량 내재화에 더 많은 인력을 투입할 수 있게 되었음
   - ➢ 향후 데이터레이크 기반 분석환경 고도화를 내부 인력으로 자체 수행할 수 있는 기반 환경 조성

IV. 향후 계획(1/2)

데이터레이크 2단계(활용) 구축에서는 데이터 수집·생산 기능 고도화를 통해 High Value 데이터를 생산할 수 있는 환경을 갖추어, 데이터 플랫폼으로써의 활용도를 높이는 활동을 수행하고자 함



※비즈니스 지표 데이터 : 22년 데이터거버넌스 구축 시 정의한 비즈니스 용어 기반의 주요 지표 기준 데이터 구성

Ⅳ. 향후 계획(2/2)

1 데이터 수집기능 고도화를 통한 적재 데이터의 정확도 향상

- 현재 분석계 데이터는 단순 현황 분석과 추이를 감지할 수 있는 수준의 정확도
- 원천 시스템에서 변경된 데이터를 정밀하게 감지하고 수집하여 데이터 정합성 증대
  (필요시 기간계 DBA와 협의하여 기간계 영향도 발생 억제, 최소화)

2 데이터 생산기능 고도화를 통한 신규 데이터 생성 환경 확보

- 원천 데이터를 활용하여 새로운 가치를 갖는 데이터를 생산하는 기능을 고도화
- 고객데이터 표준화 확대 및 데이터 거버넌스 1단계 추진으로 정의된 비즈니스 지표 데이터 구성

※ 데이터 수집·생산 기능 고도화를 통해 데이터레이크 내에서 새로운 가치를 갖는 데이터를 생산할 수 있는 환경을 확보, 데이터 활용성을 증대

3 운영/관리체계 고도화를 통한 플랫폼 운영 및 비용 안정성 향상

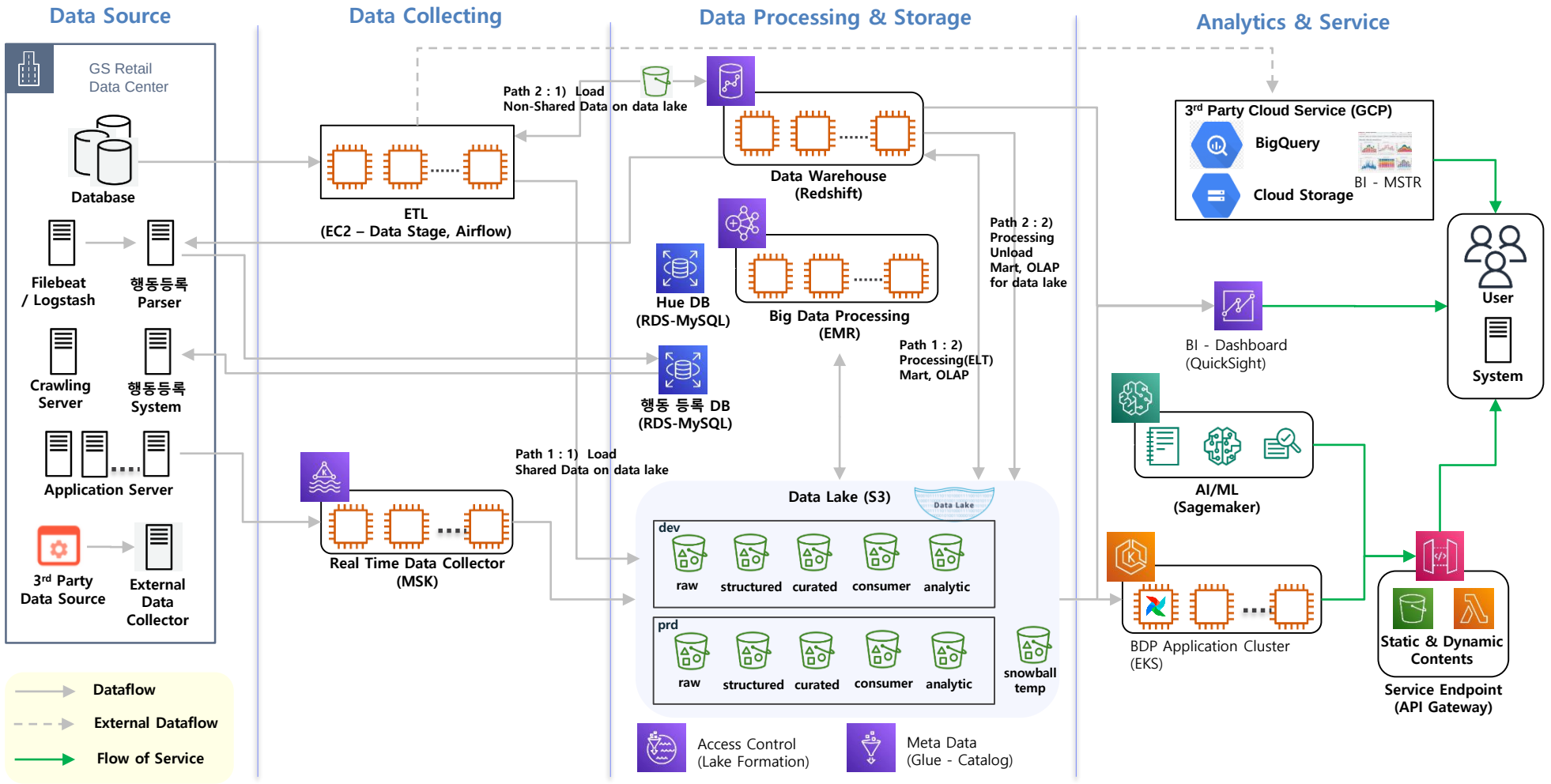
- 클라우드 자원 사용을 통제하고 효율적인 사용자 관리가 가능한 수준의 관리체계 수립

4 데이터 분석환경 고도화를 통한 데이터 분석가의 작업 생산성 향상

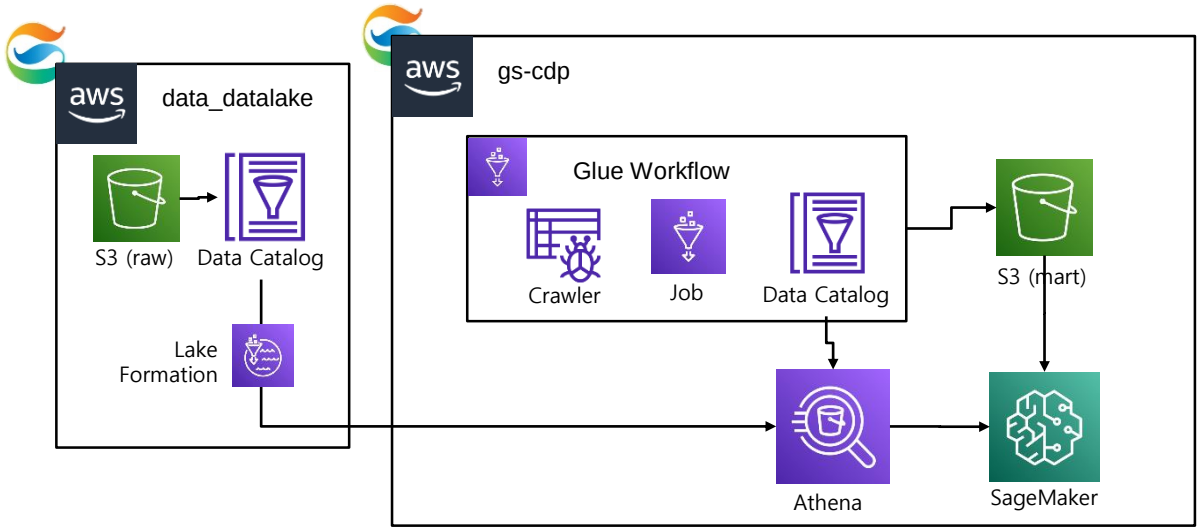
- MLOps 환경 고도화를 통해 사용성 및 안정성을 향상
- 분석가 대상으로 작업 가이드와 기술지원을 제공하여 향후 지속적으로 머신러닝을 활용한 분석 과제 수행 지원

End Of Document

별첨1. 데이터레이크 아키텍처



별첨2. 분석환경 데이터 파이프라인



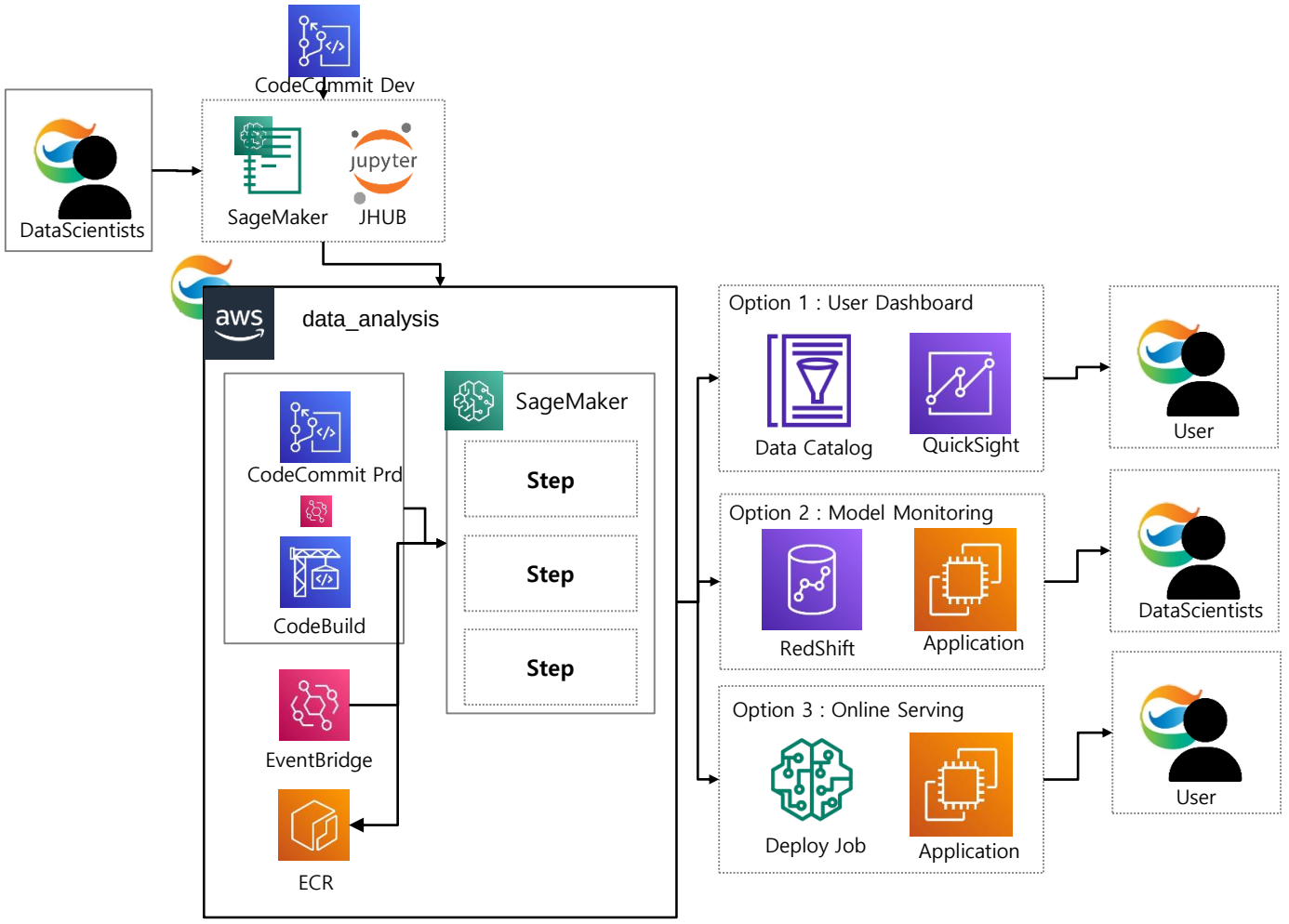
✓ **원본 데이터**
- data_datalake : S3, Data Catalog

✓ **데이터 공유**
- Lake Formation
- Glue Crawler

✓ **데이터 파이프라인 (마트 생성)**
- 마트 쿼리 : Glue Job
- 데이터 저장 : Data Catalog, S3
- 워크플로우 : Glue Workflow

✓ **데이터 활용**
- 데이터 조회 : Athena
- 데이터 분석 : SageMaker

별첨3. MLOps 아키텍처



✓ **분석 환경**
- Jupyter Notebook : SageMaker, Jhub
- MLOps : SageMaker Pipeline

✓ **코드 관리**
- 코드 관리 : CodeCommit
- 코드 빌드 : CodeBuild

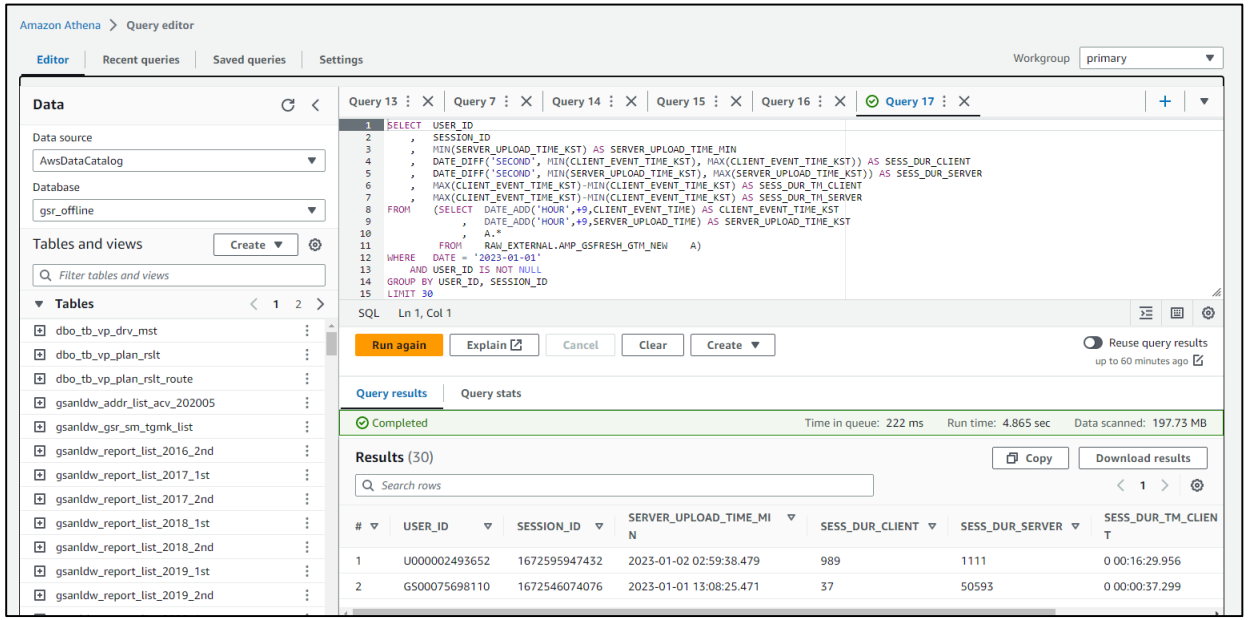
✓ **배치 스케줄링** : EventBridge

✓ **도커 이미지** : ECR

✓ **데이터 저장**
- 베스트 모델, 학습에 사용한 데이터, 하이퍼 파라미터, 인퍼런싱 결과, 모델 결과, 전처리 모델, 학습 된 모델 : S3
- 하이퍼파라미터, 인퍼런싱 결과 : Data Catalog, RedShift
- 학습 된 모델 : Model Registry

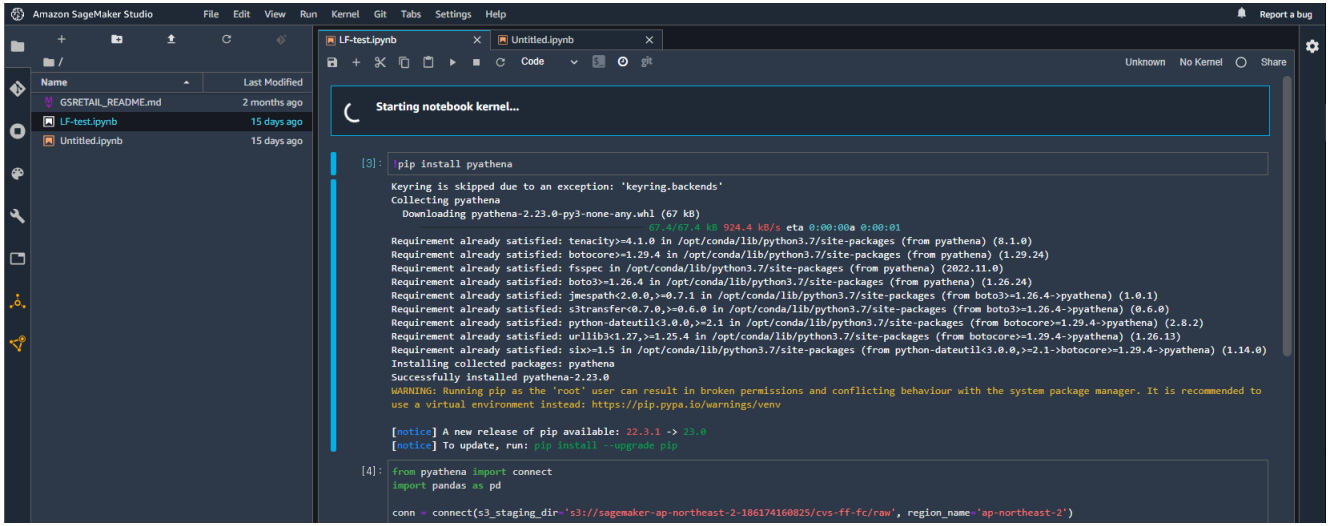
✓ **최종 결과 옵션**
- 시각화 : QuickSight
- 모델 모니터링 : Model Registry, Data Catalog, 어플리케이션
- 온라인 서빙 : EndPoint, 어플리케이션

별첨4. 데이터레이크 내 분석환경 사용기술(1/2)



Athena

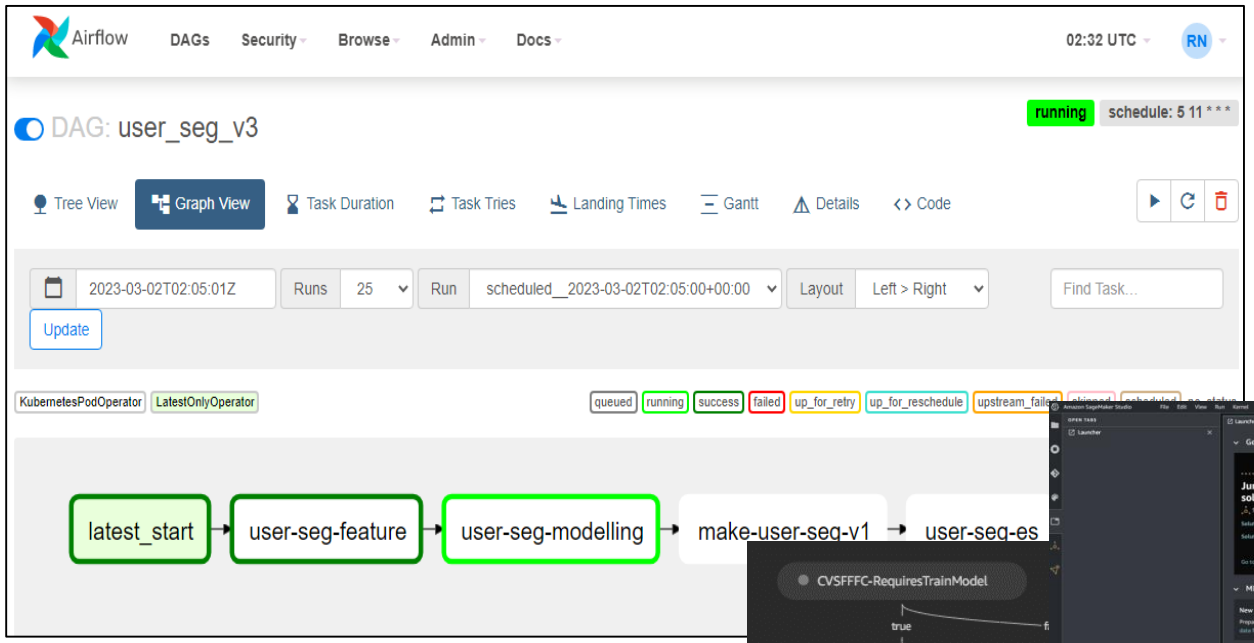
S3에 파일 형태로 저장된 데이터를 SQL을 사용해 조회하거나, 사용자가 원하는 형태로 데이터를 가공, 저장하는 기능을 제공



SageMaker

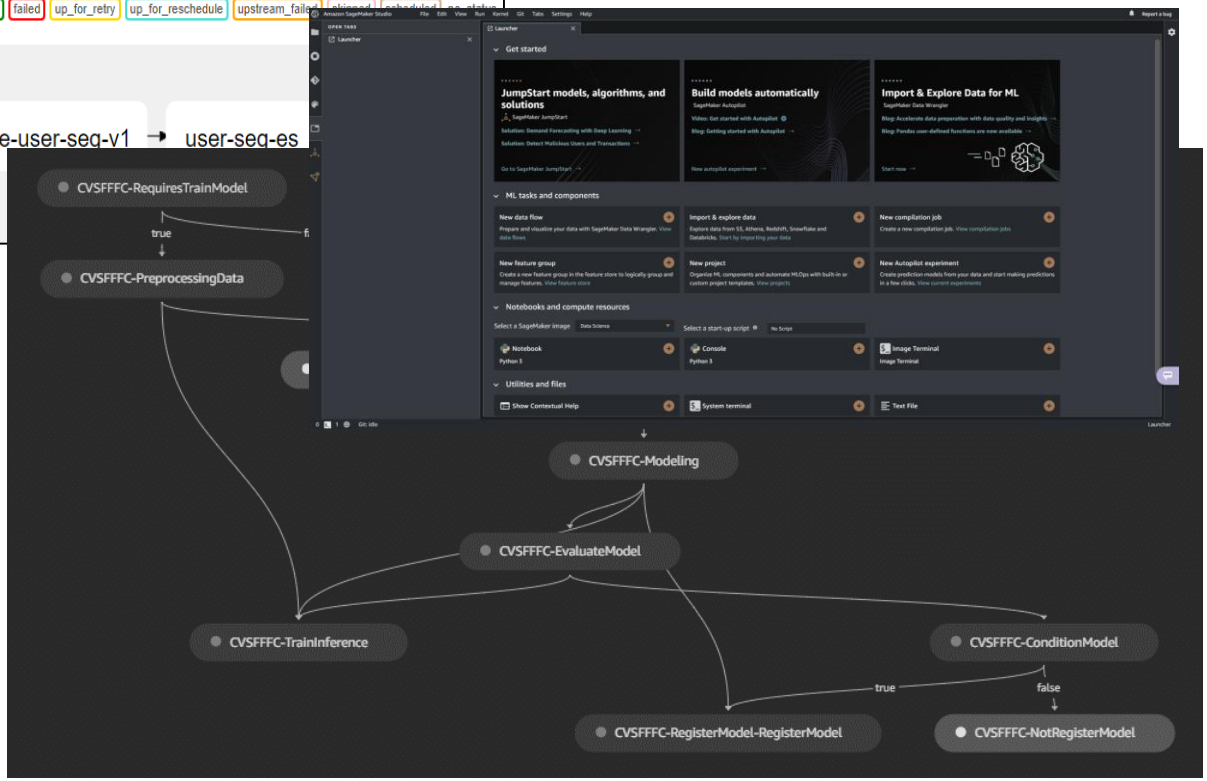
AWS 기반으로 python을 통한 데이터 분석을 수행할 수 있는 실행 환경 및 데이터레이크 인프라 내부 데이터 연계 기능 제공

별첨4. 데이터레이크 내 분석환경 사용기술(2/2)



Airflow

데이터 분석 사용자의 작업을 workflow 형태로 구성하여 정해진 일정에 따라 자동으로 실행하고 그 결과를 모니터링할 수 있는 환경을 제공



MLOps on SageMaker

모델 설계-테스트-추론-결과배포에 이르는 머신러닝의 모든 단계를 자동화하는 환경을 제공하며, 현재 머신러닝의 모든 사이클을 실제 운영 환경으로 구성 및 테스트를 진행중