databricks

Databricks

The DATA and AI Company

Lakehouse Workshop

클라우드 데이터 분석 플랫폼의 새로운 표준

Seungdon Choi | Senior Solutions Architect | 2022.9.28

databricks

온라인 워크샵 안내

- 본 온라인 워크샵은 **구글 크롬 브라우저**에 최적화 되어 있습니다.
- 사용하고 계시는 PC의 **스피커 설정**을 확인해 주세요.
- 브라우저 하단의 **아이콘**을 클릭해 창활성화 선택이 가능합니다.

Q&A 창에 질문 작성하여 주시면 답변 드리겠습니다.

Contact Us 창을 활성화 하시면 메일로 질의사항을 보내실 수 있습니다.

- 아래 메일을 통해서 문의사항 보내주시면 답변 드리겠습니다.
  koreamarketing@databricks.com

databricks

퀴즈 안내

- 워크샵 중간에 진행되는 **POP QUIZ** 세션에 참여하세요!
- 질문에 대한 답을 브라우저 하단의 **아이콘 Q&A** 를 클릭하셔서 답변을 기입하셔서 전달 주시면 선착순으로 소정의 기념품을 전달 드립니다.
- Q&A 창에 답변을 주실 때 **성함, 전화번호**를 꼭 기입해주시기 바랍니다 :)

databricks

설문 참여 이벤트

설문에 참여해주신 분들께 추첨을 통하여
스타벅스 커피 쿠폰을 보내 드립니다.

*경품은 등록하신 휴대전화로 연락 및 발송 드릴 예정입니다.
반드시 정확하게 입력하시기 바랍니다.

databricks

databricks

Databricks

The DATA and AI Company

Lakehouse Workshop

클라우드 데이터 분석 플랫폼의 새로운 표준

Seungdon Choi | Senior Solutions Architect | 2022.9.28

Agenda

- Databricks Intro/Architecture
- Start with Databricks Community Edition
- Delta Lake
- Data Engineering
- Databricks SQL
- Summary & Q/A

Databricks Introduction

데이터브릭스 소개

레이크하우스

모든 데이터, 분석 및 AI 워크로드를 위한
통합 클라우드 분석 플랫폼

Customers

7000+
across the globe

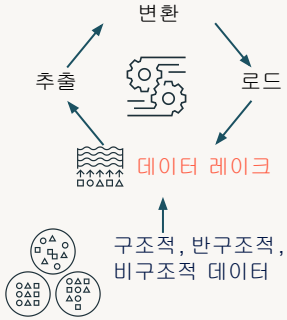
Original creators of:

Apache Spark™ Delta Lake mlflow™

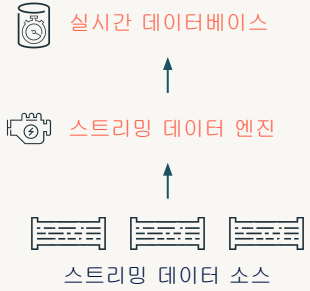
오늘날 대부분 기업은 데이터로 고전 중

데이터 웨어하우징

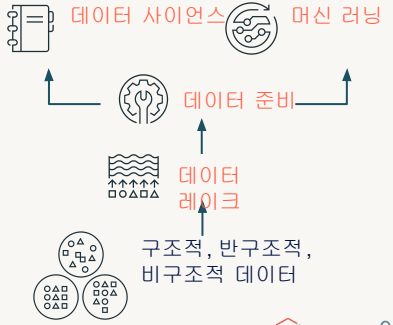
데이터 엔지니어링



스트리밍



데이터 사이언스와 머신 러닝



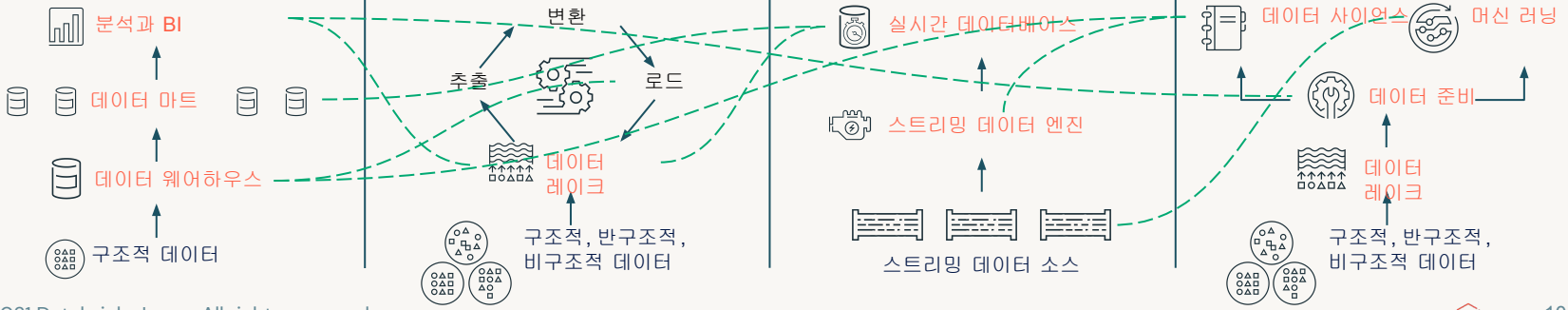
사일로화된 기술 스택 때문에 점점 더 복잡해지는 데이터 아키텍처

오늘날 대부분 기업은 데이터로 고전 중

데이터 웨어하우징		데이터 엔지니어링		스트리밍		데이터 사이언스와 머신 러닝	
Amazon Redshift	Teradata	Hadoop	Apache Airflow	Apache Kafka	Apache Spark	Jupyter	Amazon SageMaker
Azure Synapse	Google BigQuery	Amazon EMR	Apache Spark	Apache Flink	Amazon Kinesis	Azure ML Studio	MatLAB
Snowflake	IBM Db2	Google Dataproc	Cloudera	Azure Stream Analytics	Google Dataflow	Domino Data Labs	SAS
SAP	Oracle Autonomous Data Warehouse			Tibco Spotfire	Confluent	TensorFlow	PyTorch

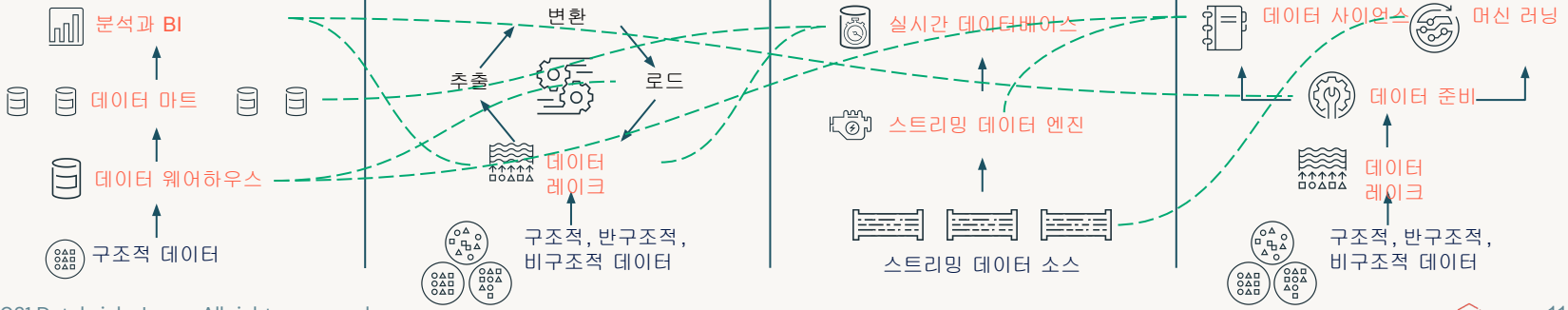
분절된 시스템과 비호환 데이터 형식 때문에 통합이 어려움

사일로화된 기술 스택 때문에 점점 더 복잡해지는 데이터 아키텍처

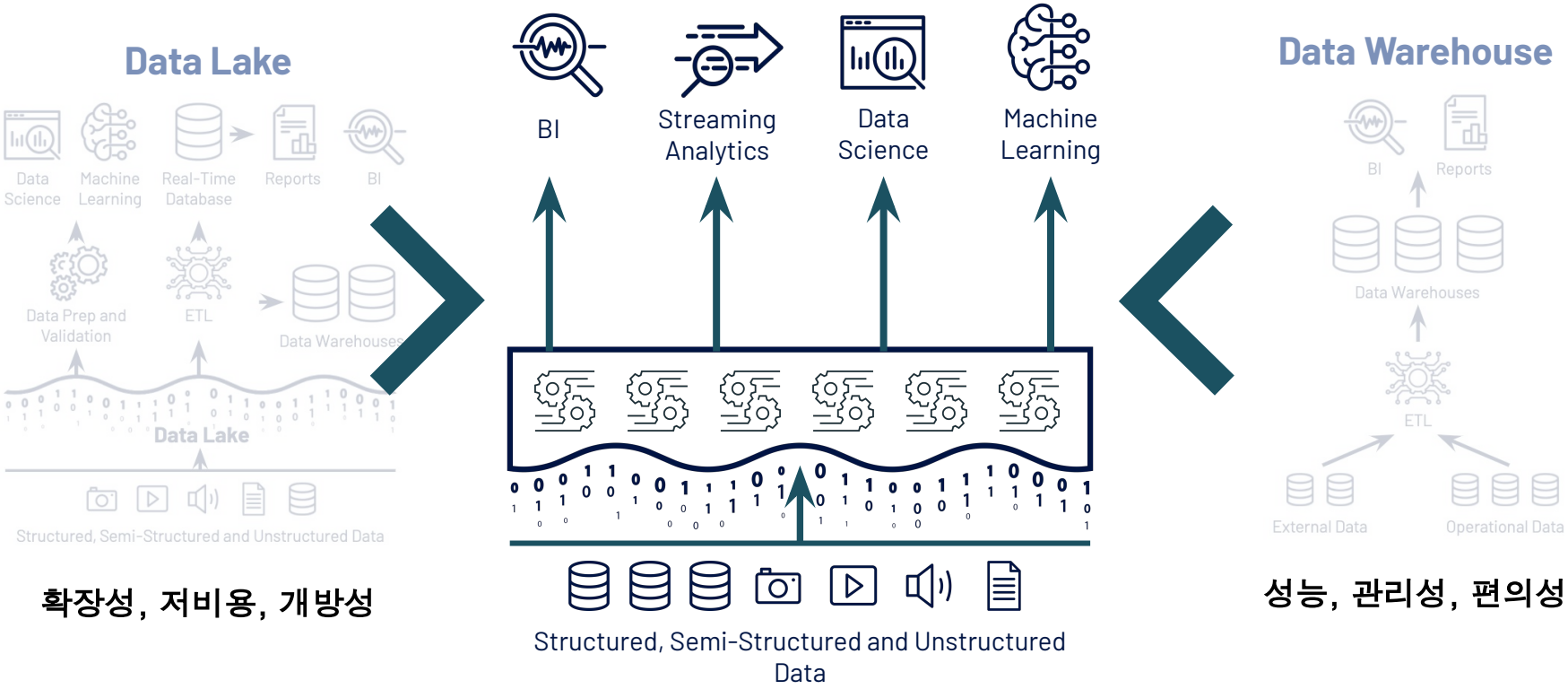


오늘날 대부분 기업은 데이터로 고전 중

데이터 웨어하우징	데이터 엔지니어링	스트리밍	데이터 사이언스와 머신 러닝
사일로화된 데이터 팀의 생산성 감소			
데이터 애널리스트	데이터 엔지니어	데이터 엔지니어	데이터 사이언티스트
분절된 시스템과 비호환 데이터 형식 때문에 통합이 어려움			
Amazon Redshift	Hadoop	Apache Kafka	Jupyter
Azure Synapse	Amazon EMR	Apache Flink	Azure ML Studio
Snowflake	Google Dataproc	Azure Stream Analytics	Domino Data Labs
SAP	Apache Airflow	Tibco Spotfire	TensorFlow
Teradata	Apache Spark	Apache Spark	Amazon SageMaker
Google BigQuery	Cloudera	Amazon Kinesis	MatLAB
IBM Db2		Google Dataflow	SAS
Oracle Autonomous Data Warehouse		Confluent	PyTorch
사일로화된 기술 스택 때문에 점점 더 복잡해지는 데이터 아키텍처			



Data Warehouse, Data Lake의 장점을 하나로?



확장성, 저비용, 개방성

성능, 관리성, 편의성

데이터
레이크

레이크하우스

데이터, 분석과 AI 워크로드를
모두 통합하는 단 하나의 플랫폼

데이터
웨어하우스

데이터 레이크

DELTA LAKE

데이터 관리와 거버넌스를
데이터 레이크로 가져오는
개방적인 방식

트랜잭션 안정성 강화

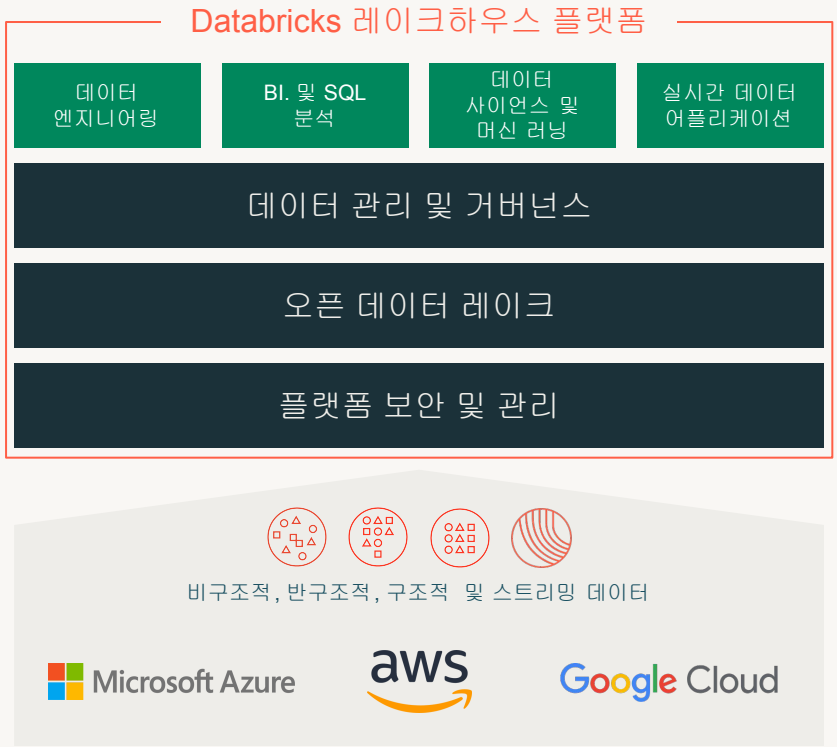
인덱싱을 통해 데이터 처리속도 48배 개선

세밀한 ACL을 사용한 대규모 데이터
거버넌스

데이터 웨어하우스

Databricks 레이크하우스 플랫폼

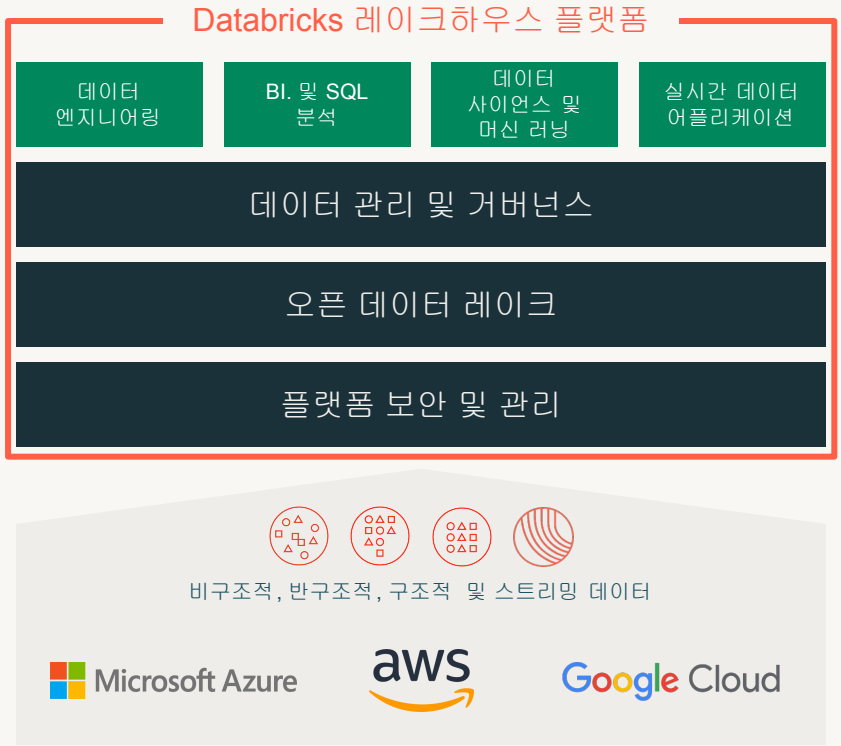
- ✓ 단순성
- ✓ 개방성
- ✓ 협업



Databricks 레이크하우스 플랫폼

단순성

데이터 사용 사례 전체를 통틀어 데이터, 분석과 AI를 단 하나의 공용 플랫폼에 통합



Databricks 레이크하우스 플랫폼

개방성

오픈 소스, 표준 및 형식으로 데이터 에코시스템 통합

세계에서 가장 성공적인 오픈 소스 데이터 프로젝트의 혁신 기술을 바탕으로 빌드

3천만 건 이상

월별 다운로드 횟수

Databricks 레이크하우스 플랫폼

개방성

오픈 소스, 표준 및 형식으로
데이터 에코시스템 통합

450+

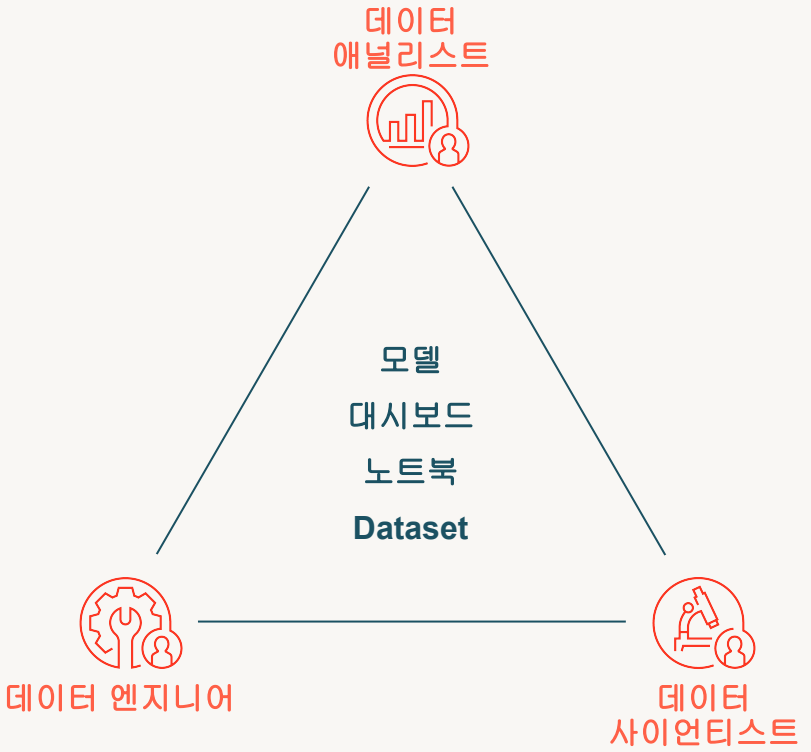
데이터 영역의 파트너

Databricks 레이크하우스 플랫폼

협업

데이터 팀을 통합,

전체 데이터 및 AI
워크플로에서 공동 작업



REGENERON

ABN·AMRO

일반적인 BI 툴을 사용해 자사 데이터 레이크에 보관되어 있는 페타바이트급 Dataset에 빠른 속도로 쿼리 실행, 데이터를 배분하고 비용 절약.

레이크하우스 아키텍처로 데이터 웨어하우징, BI와 ML을 Delta Lake에 통합하고 **70여 건의 usecase**에 IoT와 스트리밍 활용

예측기반 유지보수와 인벤토리 관리를 통해 **수백만 달러 절약 가능**

페타바이트급 규모의 유전체학, 임상 데이터를 분석하여 새로운 치료법 개발 속도 가속.

협업 방식을 개선하고 파이프라인 속도를 빠르게 하여 데이터 준비 기간을 **기존 3주에서 2일로 단축**

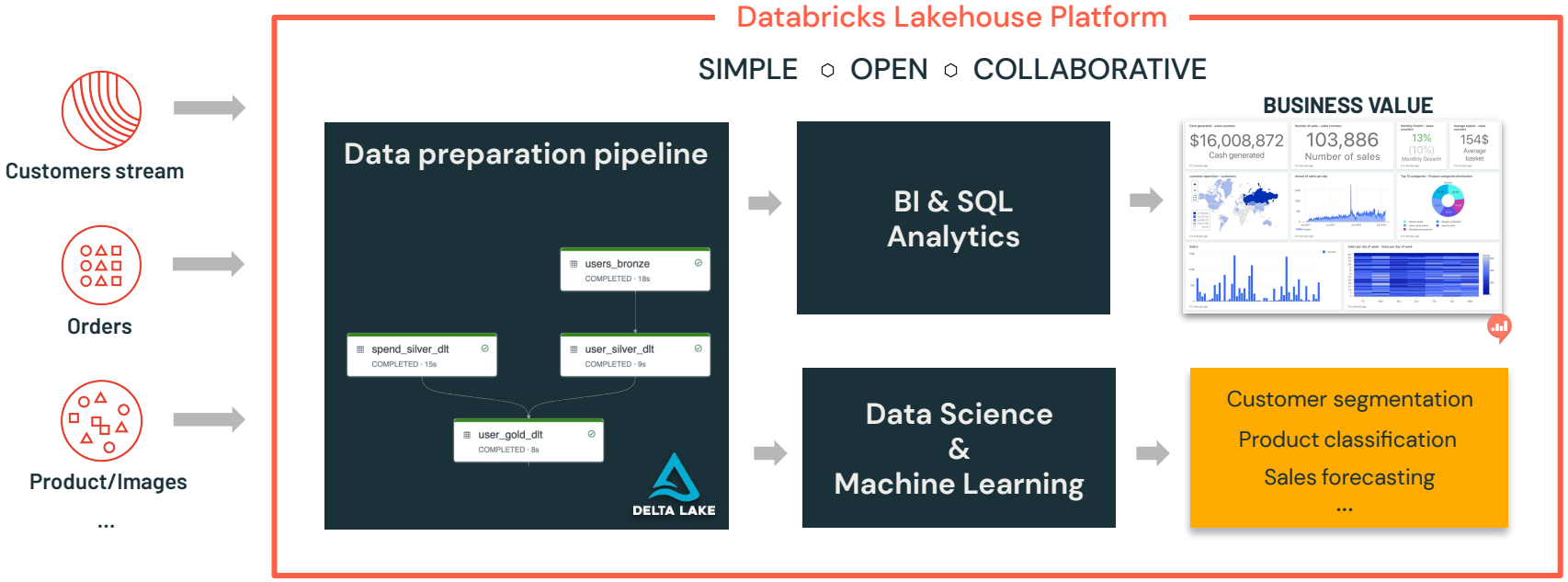
자사 Dataset에 쿼리 성능 **600배 향상**

데이터와 AI를 활용하여 대량의 뱅킹 정보 구동, 매달 새로운 GTM 역량 도입 가능.

레이크하우스 아키텍처를 사용해 서로 다른 150종의 소스에서 수집한 수백 TB 규모의 데이터를 하나의 플랫폼에 통합

100여 종의 ML 모델을 이전보다 10배 빨리 출시, 사기 행위 탐지부터 고객 서비스, 마케팅과 물류에 이르기까지 50가지 usecase 지원, 100여 가지 usecase 적용 예정

Building a Lakehouse for DE/BI/ML



aws Microsoft Azure Google Cloud

Databricks Architecture and Services

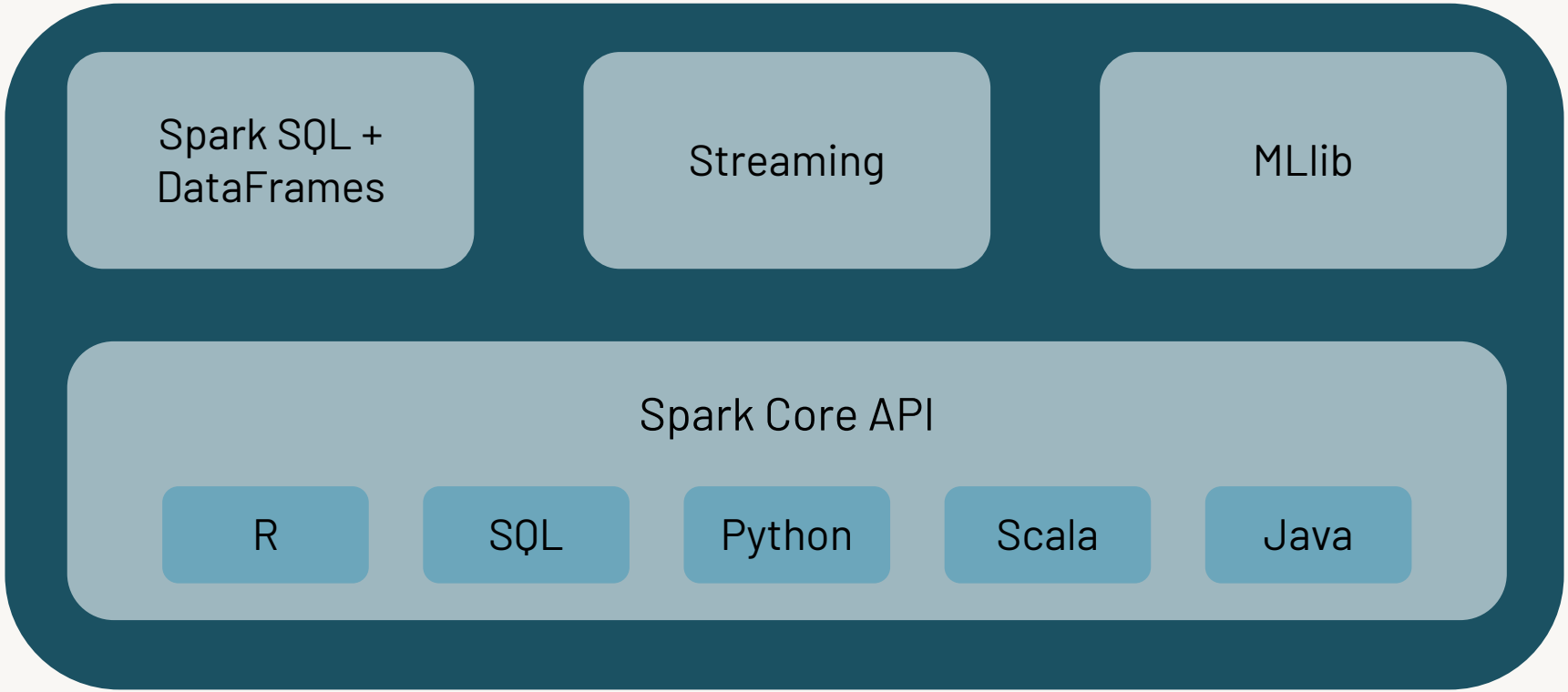
- 빅데이터 프로세싱의 사실상 시장 표준 통합 분석 엔진
- 데이터 프로세싱 영역에서 가장 큰 오픈 소스 프로젝트
- 데이터브릭스의 창립자들이 만든 기술

빠르고

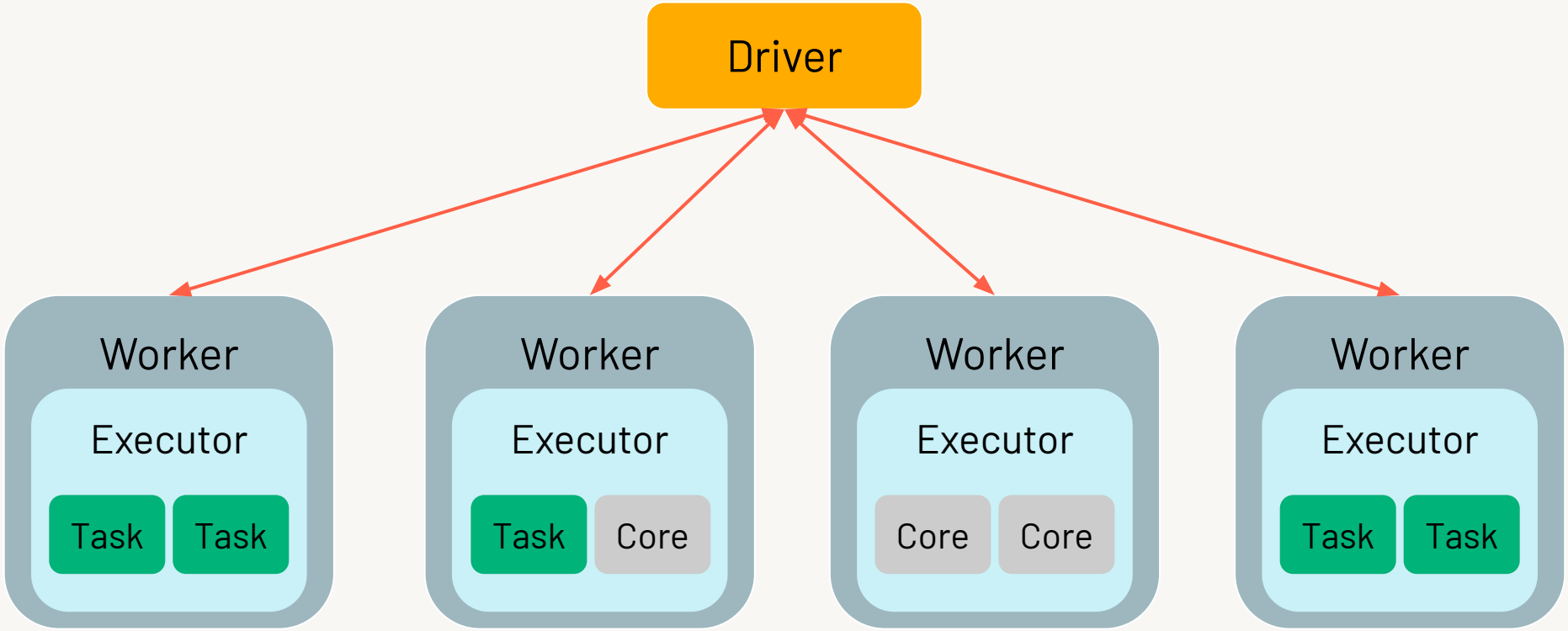
사용이 쉽고

다용도

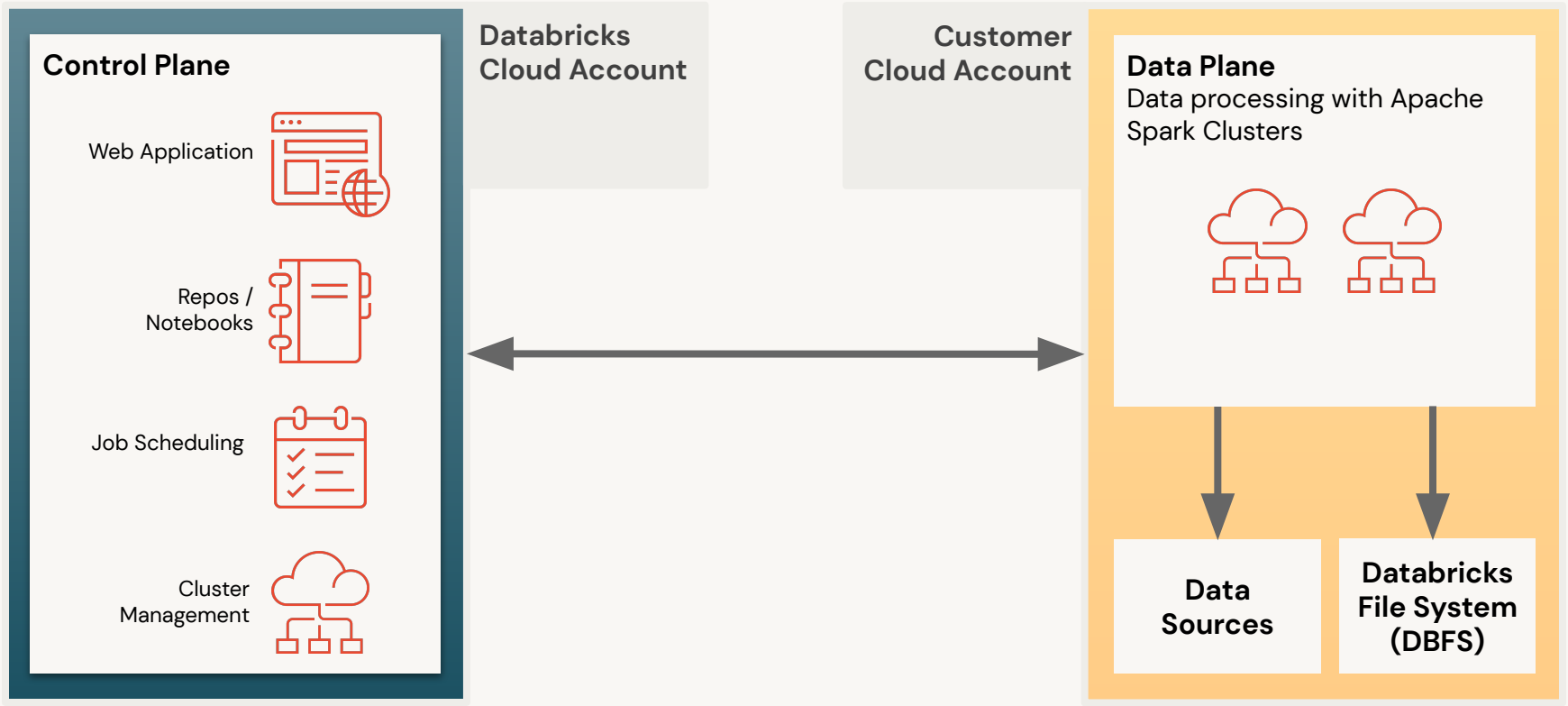
Spark API



Spark Cluster



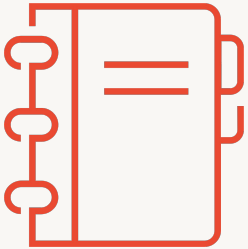
Databricks Architecture



Databricks Services

Control Plane in Databricks
고객 계정, 데이터셋 및 클러스터 관리 기능 제공

Databricks Web Application

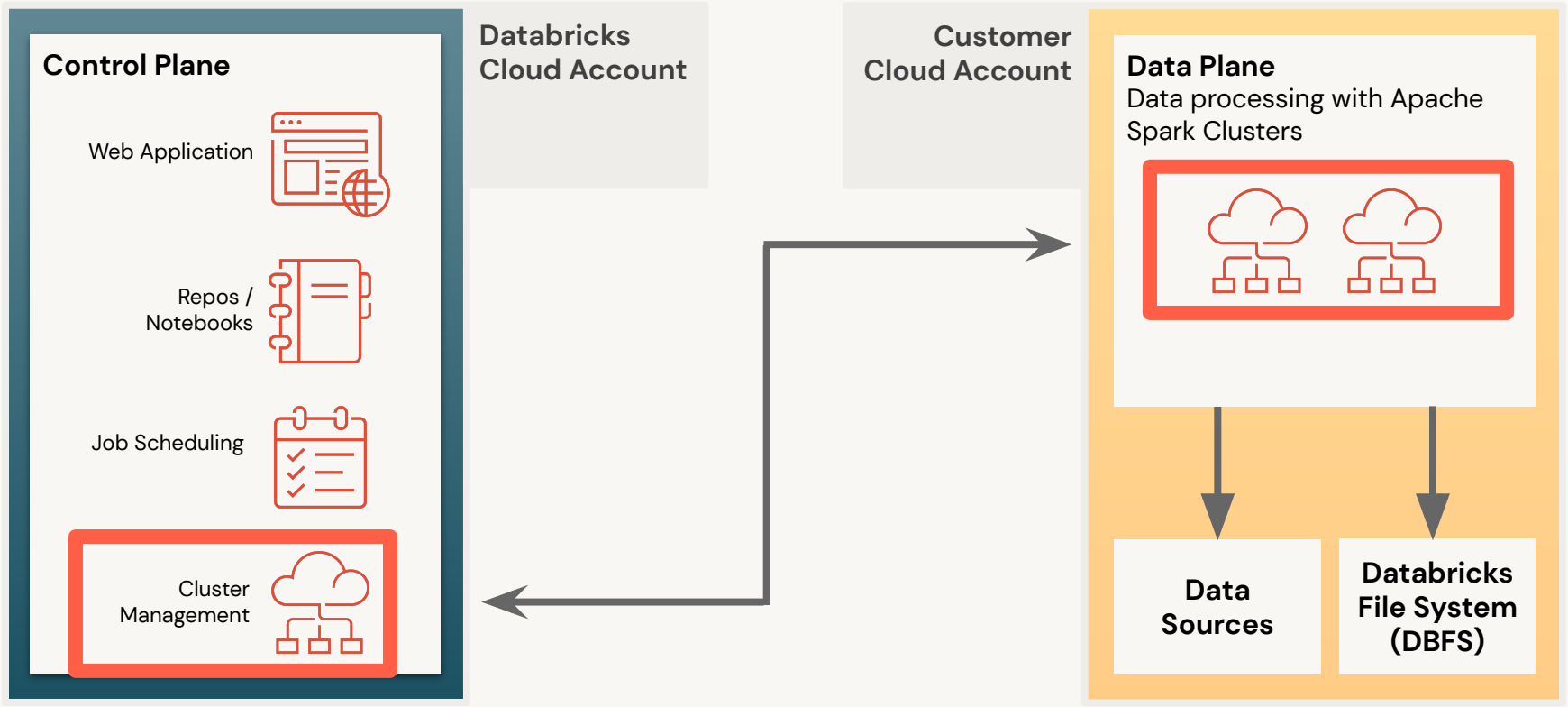


Repos / Notebooks

Jobs

Cluster Management

Clusters



Clusters

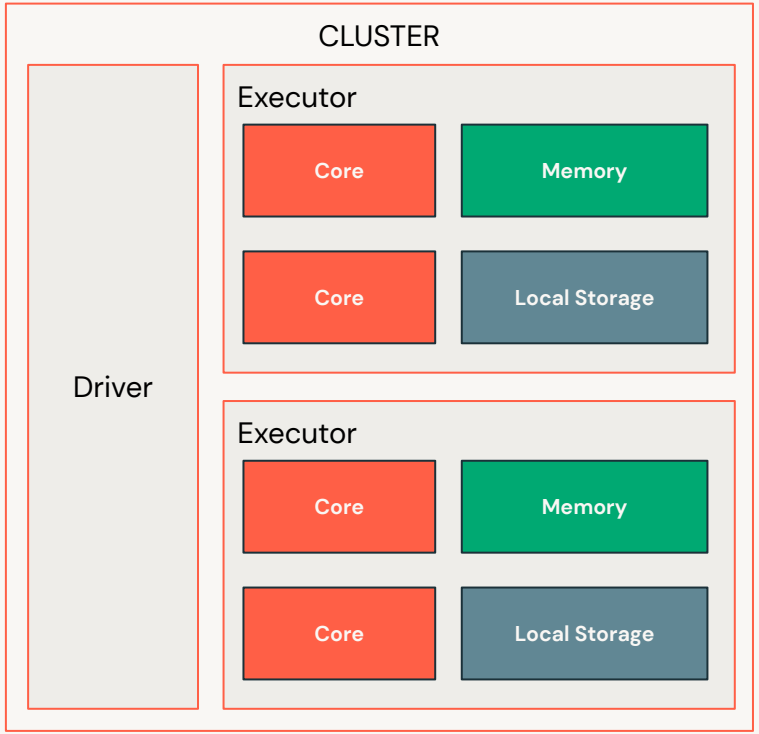
Overview

고객사의 클라우드계정내에서
동작하는 스파크 클러스터

1개 이상의 VM 인스턴스(들)로 구성

Driver
executor 노드의 작업 코디네이션

Executors Spark job 을 구성하는
task들을 수행



Cluster Types

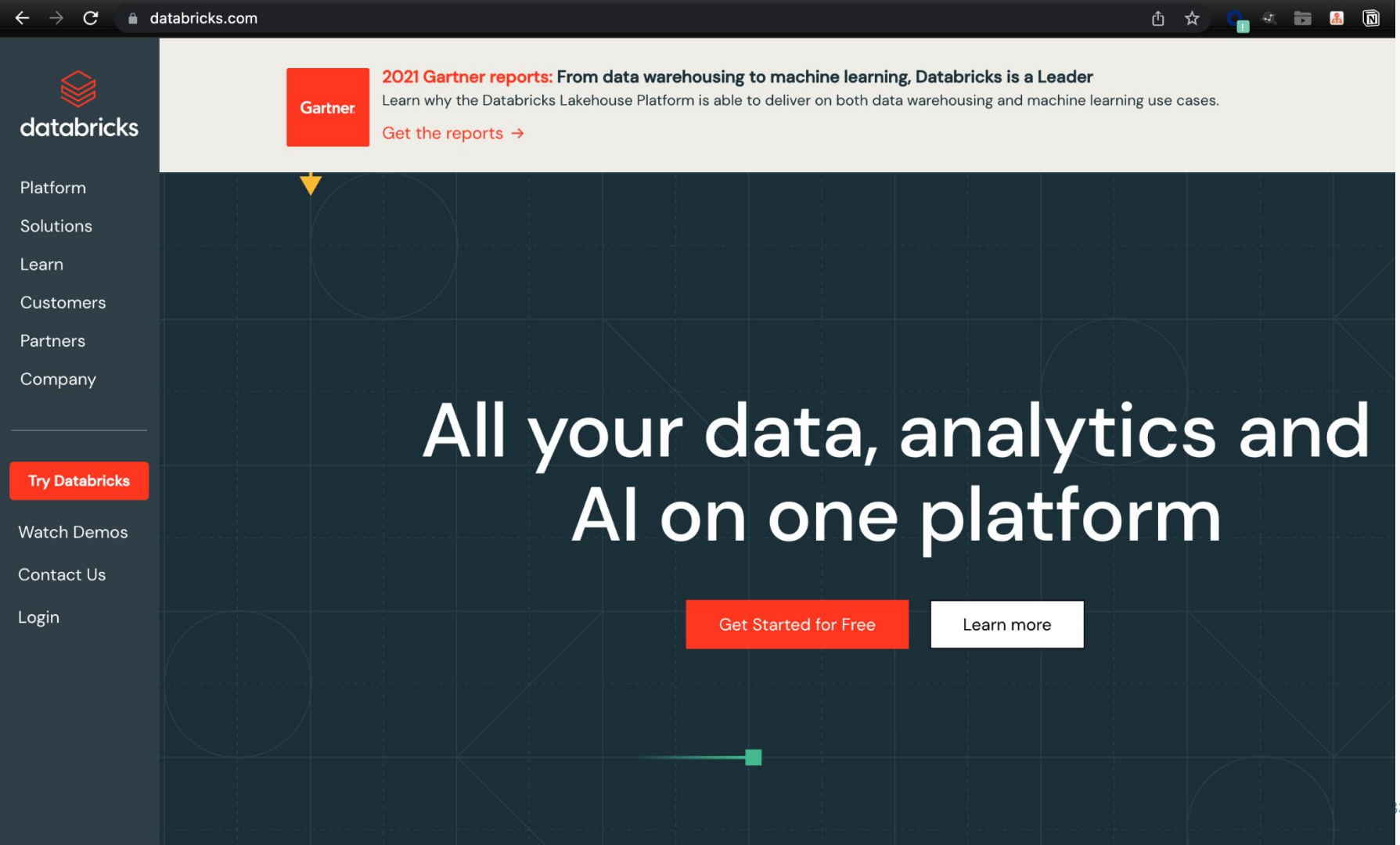
Type	All-Purpose Cluster	Jobs Cluster
Persistence	Persistent cluster	Ephemeral cluster created for job, terminated on completion
Workload	Interactive (data analytics)	Automated (data engineering)
Use	Analysis and ad-hoc work	Production and repeated workloads
Benefits	Collaborate interactive analysis, Ability to restart cluster	Isolation and debugging, Scheduled runs
Cost	$$$	$

Demo
데이터브릭스 훑어보기

databricks

How to start

Databricks Community Edition
으로 사용 시작해 보기


databricks.com
databricks
Platform
Solutions
Learn
Customers
Partners
Company
Try Databricks
Watch Demos
Contact Us
Login
Gartner
2021 Gartner reports: From data warehousing to machine learning, Databricks is a Leader
Learn why the Databricks Lakehouse Platform is able to deliver on both data warehousing and machine learning use cases.
Get the reports →
All your data, analytics and AI on one platform
Get Started for Free
Learn more

Try Databricks for free

An open and unified data analytics platform for data engineering, data science, machine learning, and analytics. From the original creators of Apache Spark™, Delta lake, MLflow, and Koalas.

aws Microsoft Azure Google Cloud

Databricks trial:

- Collaborative environment for data teams to build solutions together.
- Interactive notebooks to use Apache Spark™, SQL, Python, Scala, Delta Lake, MLflow, TensorFlow, Keras, Scikit-learn and more.
- Available as a 14-day full trial in your own cloud, or as a lightweight trial hosted by Databricks.

Used by:

Virgin hyperloop edmunds nielsen

RIOT GAMES REGENERON Grab

Please tell us about yourself

First Name: *

seungdon

Last Name: *

Choi

Company *

Company Email *

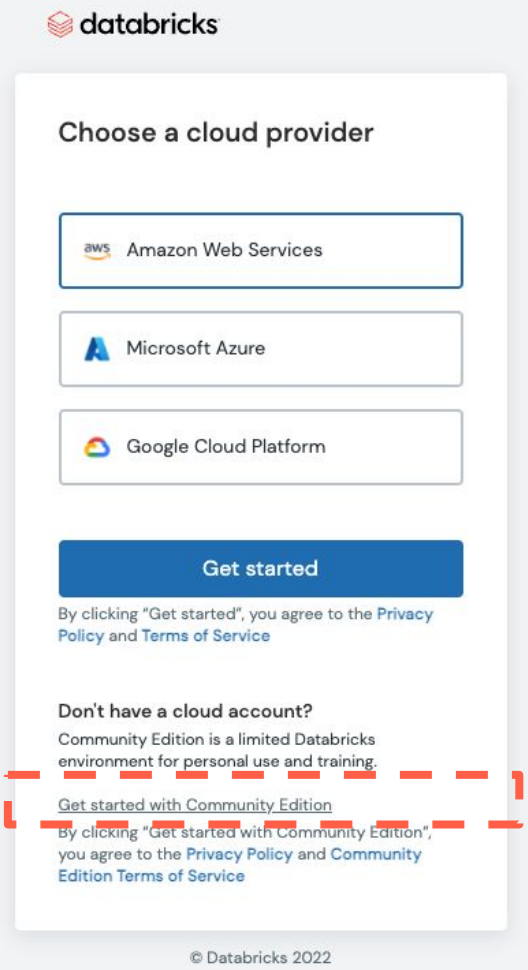
Title *

Phone Number

Keep me informed with occasional updates about Databricks and related open source products

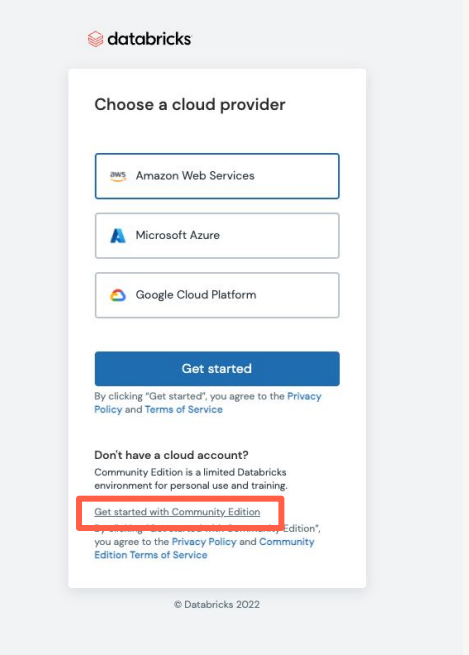
By Clicking "Get Started For Free", you agree to the Privacy Policy.

GET STARTED FOR FREE

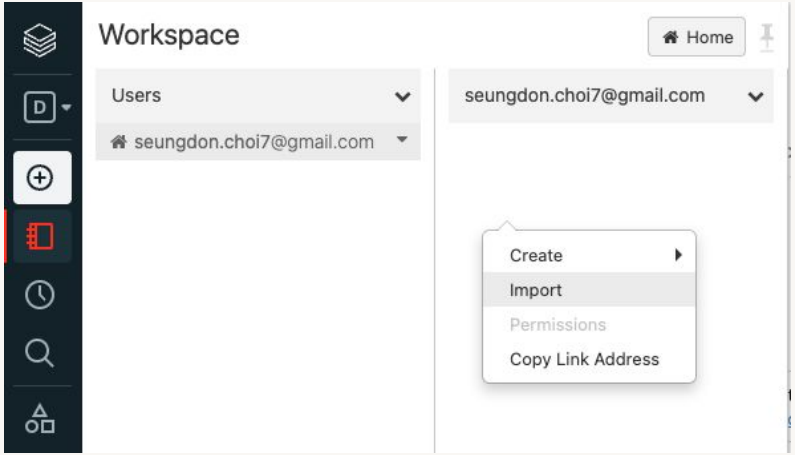


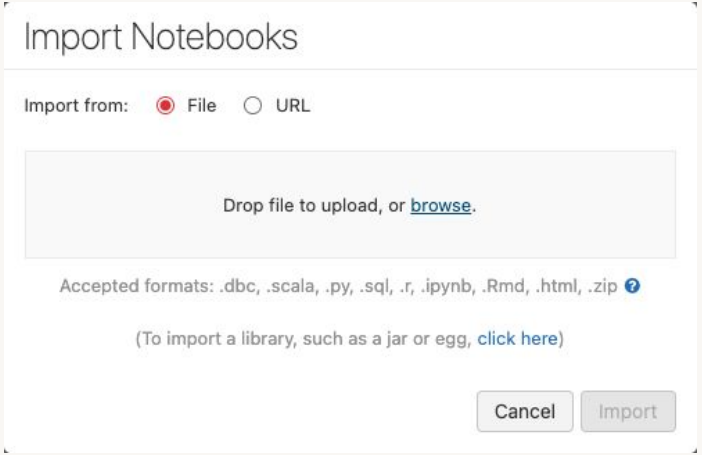
Community Edition

- **완전 무료**(databricks 및 cloud 비용 무료)
- 기능 제한
  - Single Node 클러스터 Only
  - Job Cluster 사용 불가
  - Databricks SQL 사용 불가
  - 권한제어등…
- 전체기능 가능한 Databricks Edition으로 업그레이드 가능
- https://community.cloud.databricks.com/



Import workshop notebooks





https://tinyurl.com/db-lakehouse-webinar

Lakehouse_KR.dbc 파일 import

Lab(10min)

Community Edition 에 등록하기

Launch Cluster

Import your first Notebook

1. Play with Databricks Notebook

databricks

What is Delta Lake?

Delta Lake 는 기존의 스토리지 시스템 위에 데이터 레이크 하우스를 구축할 수 있게 도와주는 오픈소스 프로젝트입니다.

Delta Lake brings ACID to object storage

- **A**tomicity
- **C**onsistency
- **I**solation
- **D**urability

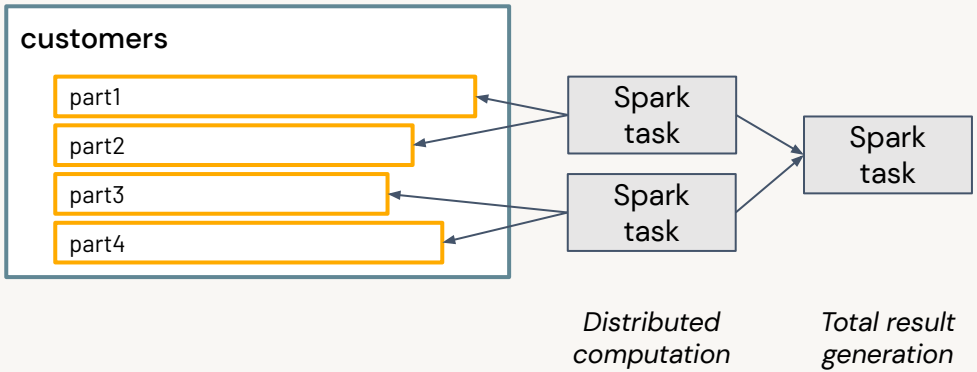
ACID 지원으로 해결되는 많은 문제들

1. 단일 테이블에 여러 트렌젝션이 동시 Write 처리가 힘듬
2. 기존 데이터의 수정이 힘듬
3. 중간에 실패한 Job - Corrupt Data
4. 실시간 운영 힘듬(Batch중심)
5. 과거 데이터 스냅샷 유지 비용 증가

Before Delta : What does Parquet look like?

```
%fs ls /tmp/bernhard/loan_by_state_delta
```

path
dbfs:/tmp/bernhard/loan_by_state_delta/part-00000-cd80fd12-457b-4058-a904-3628c6e90a57-c000.snappy.parquet
dbfs:/tmp/bernhard/loan_by_state_delta/part-00004-69b2735a-18d1-474e-8318-ecd13ca410a7-c000.snappy.parquet
dbfs:/tmp/bernhard/loan_by_state_delta/part-00009-05545d85-f051-4a37-852a-18298fb4f3cd-c000.snappy.parquet
dbfs:/tmp/bernhard/loan_by_state_delta/part-00010-3fa88ba6-61cc-45cd-8c0d-e0949d1bbcce-c000.snappy.parquet



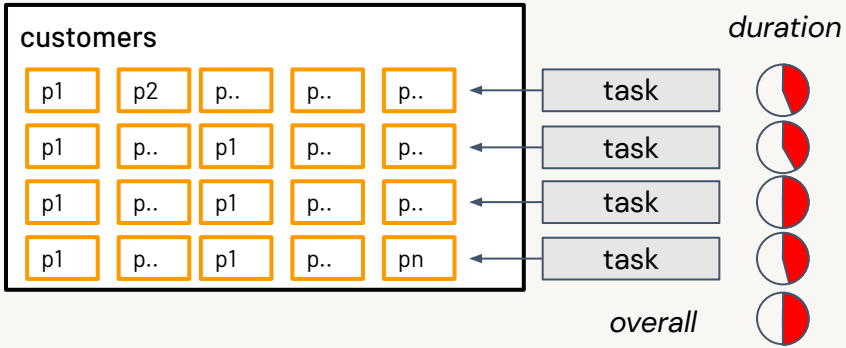
Parquet data

- Folder 형식 파티션 구분
- 여러 part file 들을 포함
- 분산 테스크에서 read 에 최적화되게 설계

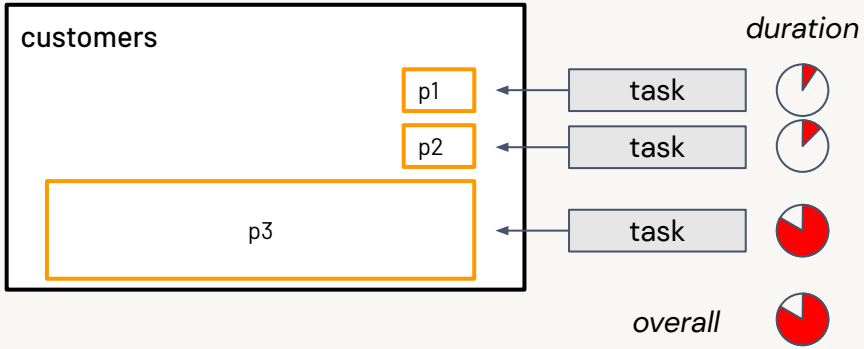
Note: Parquets 형식은 data Append만 지원!

Parquet 파일 처리시 문제점

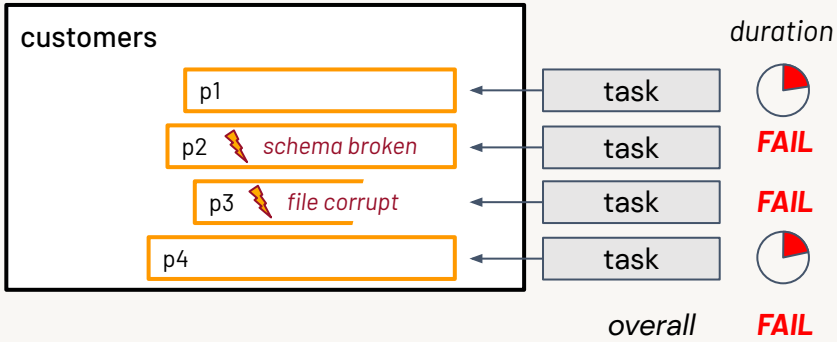
Small file problem



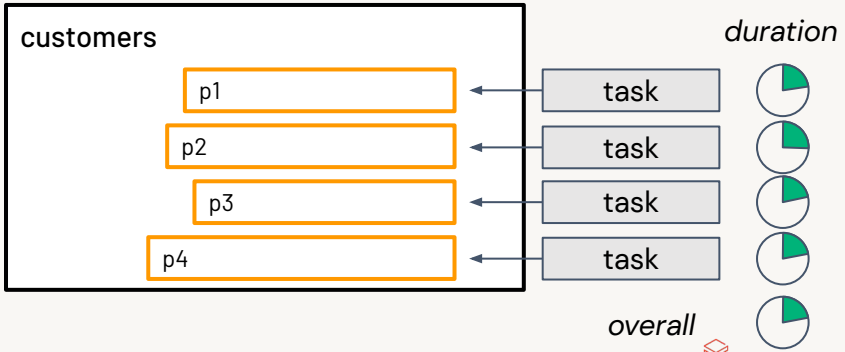
Big file (data skew) problem



Corrupt data

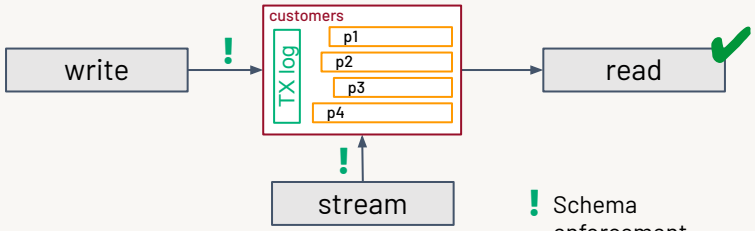


Goal

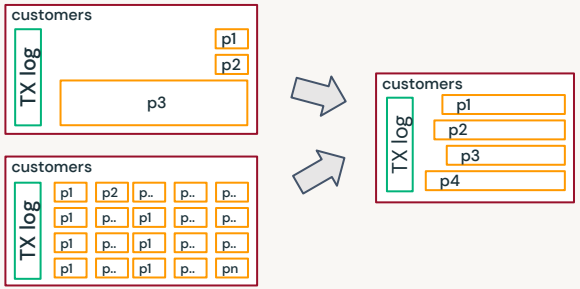


Delta 가 가져오는 안정성과 성능 향상 기능

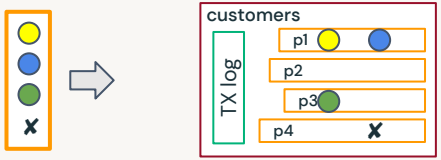
Consistency
(never read broken, unfinished or wrong data)



Optimizations on the fly
(no need to have a complex pipeline)

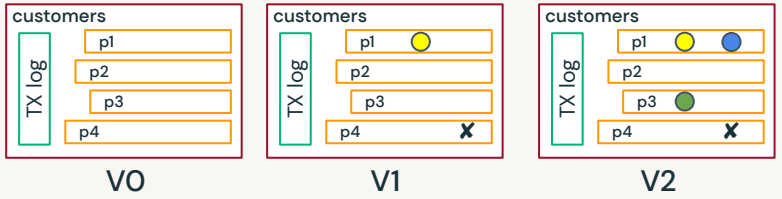


Direct updates and deletes
(no need to have a complex pipeline)



- GDPR
- Change Data Capture (CDC & SCD)

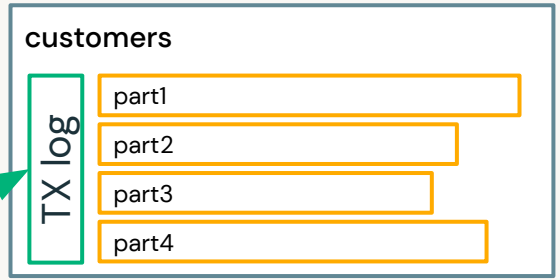
Time travel
(implicit snapshots)



Delta Tables

Transaction Log Provides Metadata Layer for Consistency

Data in Parquet




```
%fs ls /tmp/bernhard/loan_by_state_delta
```

path
dbfs:/tmp/bernhard/loan_by_state_delta/_delta_log/
dbfs:/tmp/bernhard/loan_by_state_delta/part-00000-cd80fd12-457b-4058-a904-3628c6e90a57-c000.snappy.parquet
dbfs:/tmp/bernhard/loan_by_state_delta/part-00004-69b2735a-18d1-474e-8318-ecd13ca410a7-c000.snappy.parquet
dbfs:/tmp/bernhard/loan_by_state_delta/part-00009-05545d85-f051-4a37-852a-18298fb4f3cd-c000.snappy.parquet
dbfs:/tmp/bernhard/loan_by_state_delta/part-00010-3fa88ba6-61cc-45cd-8c0d-e0949d1bbcce-c000.snappy.parquet

https://databricks.com/blog/2019/08/21/diving-into-delta-lake-unpacking-the-transaction-log.html

Transaction Log / Metadata

Paquet Checkpoint

JSON Transaction

```
%fs ls /mnt/datalake/delta_workshop_data/lending_club_accepted_2007-2018/_delta_log
```

	path	name	size
3	dbfs:/mnt/datalake/delta_workshop_data/lending_club_accepted_2007-2018/_delta_log/.s3-optimization-2	.s3-optimization-2	0
4	dbfs:/mnt/datalake/delta_workshop_data/lending_club_accepted_2007-2018/_delta_log/00000000000000000000.checkpoint.parquet	00000000000000000000.checkpoint.parquet	28009
5	dbfs:/mnt/datalake/delta_workshop_data/lending_club_accepted_2007-2018/_delta_log/00000000000000000000.json	00000000000000000000.json	46221
6	dbfs:/mnt/datalake/delta_workshop_data/lending_club_accepted_2007-2018/_delta_log/00000000000000000001.crc	00000000000000000001.crc	94
7	dbfs:/mnt/datalake/delta_workshop_data/lending_club_accepted_2007-2018/_delta_log/00000000000000000001.json	00000000000000000001.json	5522
8	dbfs:/mnt/datalake/delta_workshop_data/lending_club_accepted_2007-2018/_delta_log/_last_checkpoint	_last_checkpoint	24

```
dfParquetTransaction = spark.read.parquet('/mnt/datalake/delta_workshop_data/lending_club_accepted_2007-2018/_delta_log/00000000000000000000.checkpoint.parquet')
display(dfParquetTransaction)
```

▸ (2) Spark Jobs

▸ dfParquetTransaction: pyspark.sql.dataframe.DataFrame = [txn: struct, add: struct ... 4 more fields]

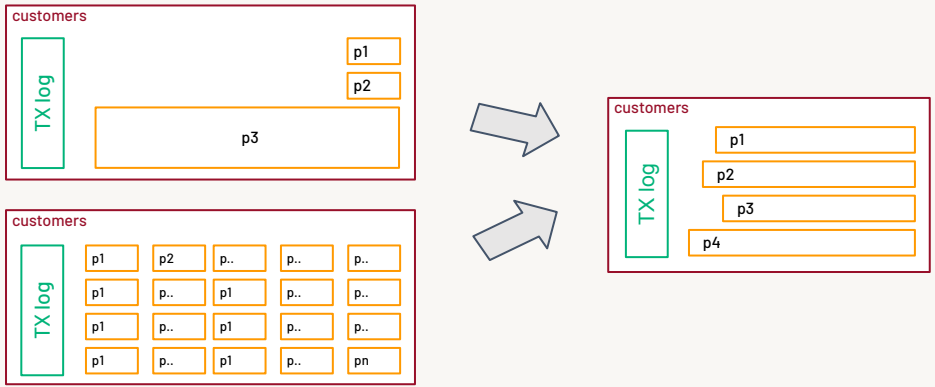
	txn	add	remove	metaData	protocol	commitInfo
1	null	null	null	null	▸ {"minReaderVersion": 1, "minWriterVersion": 2}	null
	null	null	null	▸ {"id": "2739cf3a-a95b-4a56-9bd3-4dc7b7f19a84", "name": null, "description": null, "format": {"provider": "parquet", "options": {}}, "schemaString": "{\"type\":\"struct\",\"fields\":[{\"name\":\"id\",\"type\":\"string\",\"nullable\":true,\"metadata\":{}},{\"name\":\"member_id\",\"type\":\"string\",\"nullable\":true,\"metadata\":{}},{\"name\":\"loan_amnt\",\"type\":\"double\",\"nullable\":true,\"metadata\":{}},{\"name\":\"funded_amnt\",\"type\":\"double\",\"nullable\":true,\"metadata\":{}},{\"name\":\"funded_amnt_inv\",\"type\":\"double\",\"nullable\":true,\"metadata\":{}},{\"name\":\"term\",\"type\":\"string\",\"nullable\":true,\"metadata\":{}},	null	null

Scalable Metadata

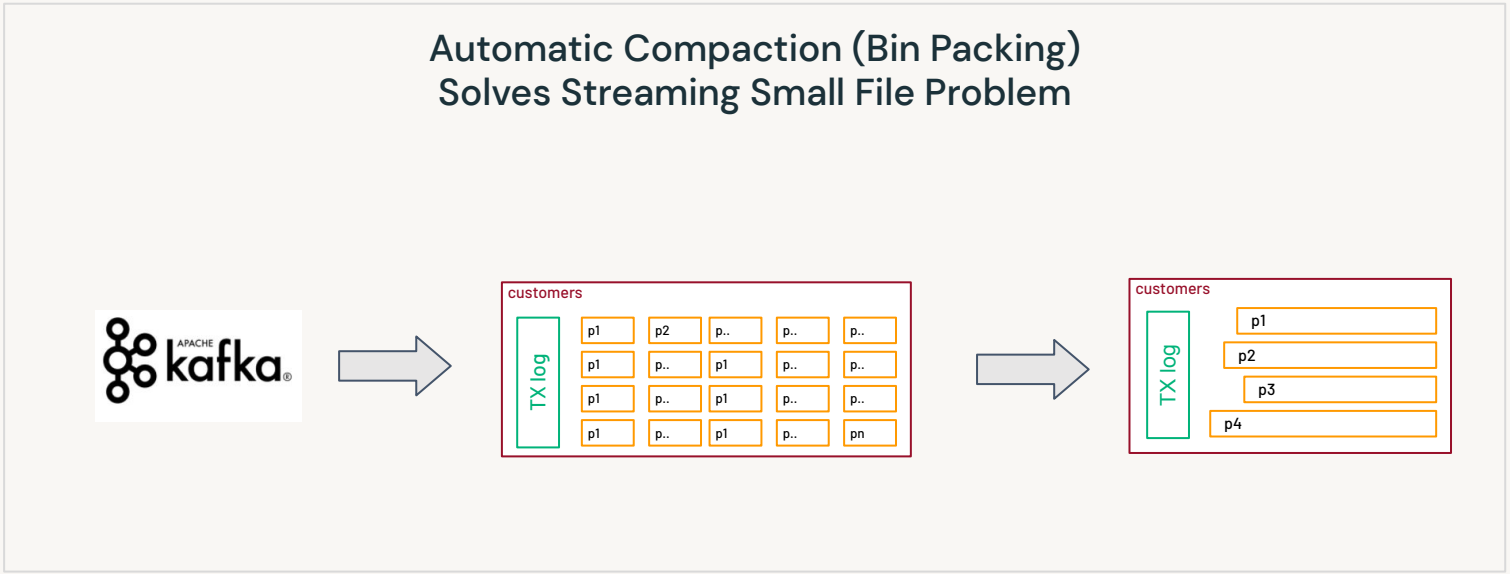
Delta Engine – Optimize

```
OPTIMIZE events
WHERE date >= current_timestamp() - INTERVAL 1 day
ZORDER BY (eventType)
```

Compaction with Optimize Command



Delta Engine – Auto Optimize



Data skipping을 이용한 비약적 쿼리 성능 향상

- 많은 DBMS 나 Big Data 시스템에서 활용되는 I/O 프루닝 기술
- Idea: file level 로 min&max등의 통계정보 활용으로 불필요한 파일 스캔을 방지
- Example:

SELECT * FROM table WHERE col = 5

SELECT file_name FROM index
WHERE col_min <= 5 AND col_max >= 5

file_name	col_min	col_max
file1	6	8
file2	3	10
file3	1	4

Delta performance features

- **Optimize**
  Command that bin-packs files to the right size.
- **Auto-optimize**
  Small/large files compacted to enable data lake applications experience great consistent performance and scalability.
- **Scalable writes**
  Fine-grained conflict resolution allowing multiple writers to succeed
- **Data skipping**
  Improves read performance by only reading subsets of the files.
- **Z-Order**
  Clusters files in a way that enables data skipping for multi-dimensional filters.
- **Bloom filters**
  Improves read performance by only reading subsets of the files that have data matching users filters.
- **Caching**
  Automatically caches input Delta (and Parquet) tables, improving read throughput by 2X to 10X
- **Skewed join**
  Supports joining two datasets with severe data skew, a problem with a lot of real-world datasets. Needed at scale.
- **Range join (time-series data)**
  Supports joining two datasets based on overlapping ranges, such as time series analysis.

Lakehouse

One platform for every use case

Structured transactional layer

Data Lake for all your data

Demo

2. Play with Delta

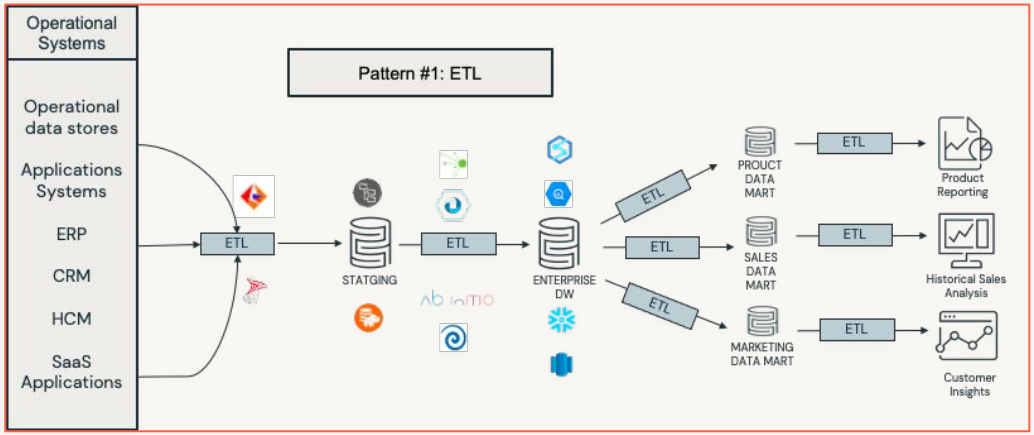
Delta Lake Summary

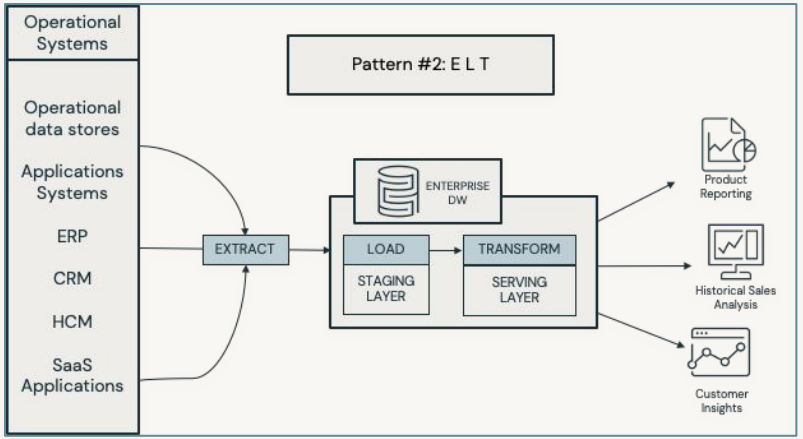
- Lakehouse 아키텍처의 핵심 컴포넌트
- ACID 지원으로 데이터 일관성/품질 보장
- Robust data store
- Apache Spark 으로 처리하기 최상의 궁합

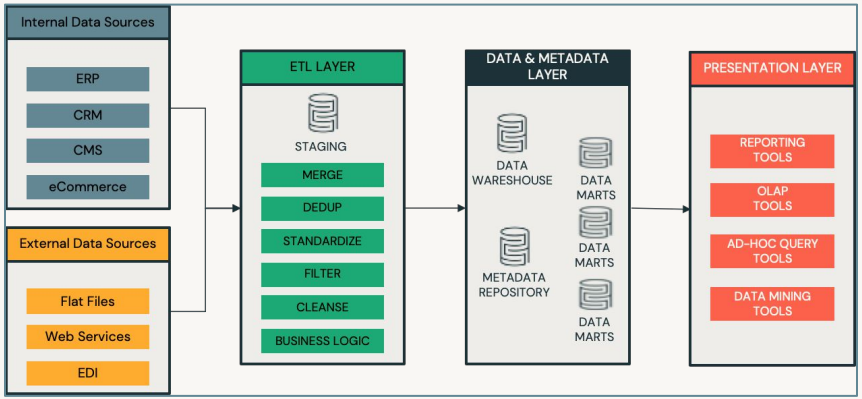
Data Engineering

Raw Data to Insight!

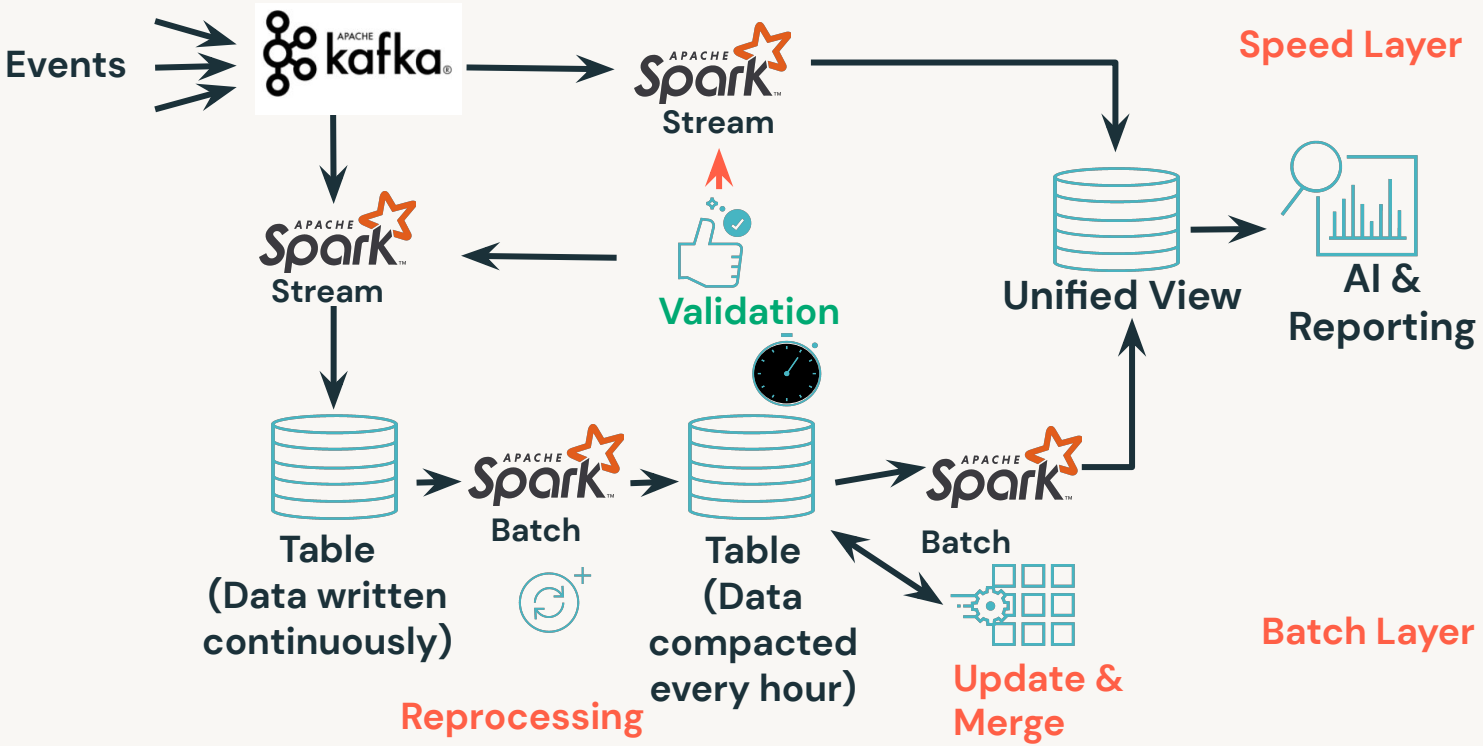
Old: ETL/ELT on Data Warehouse



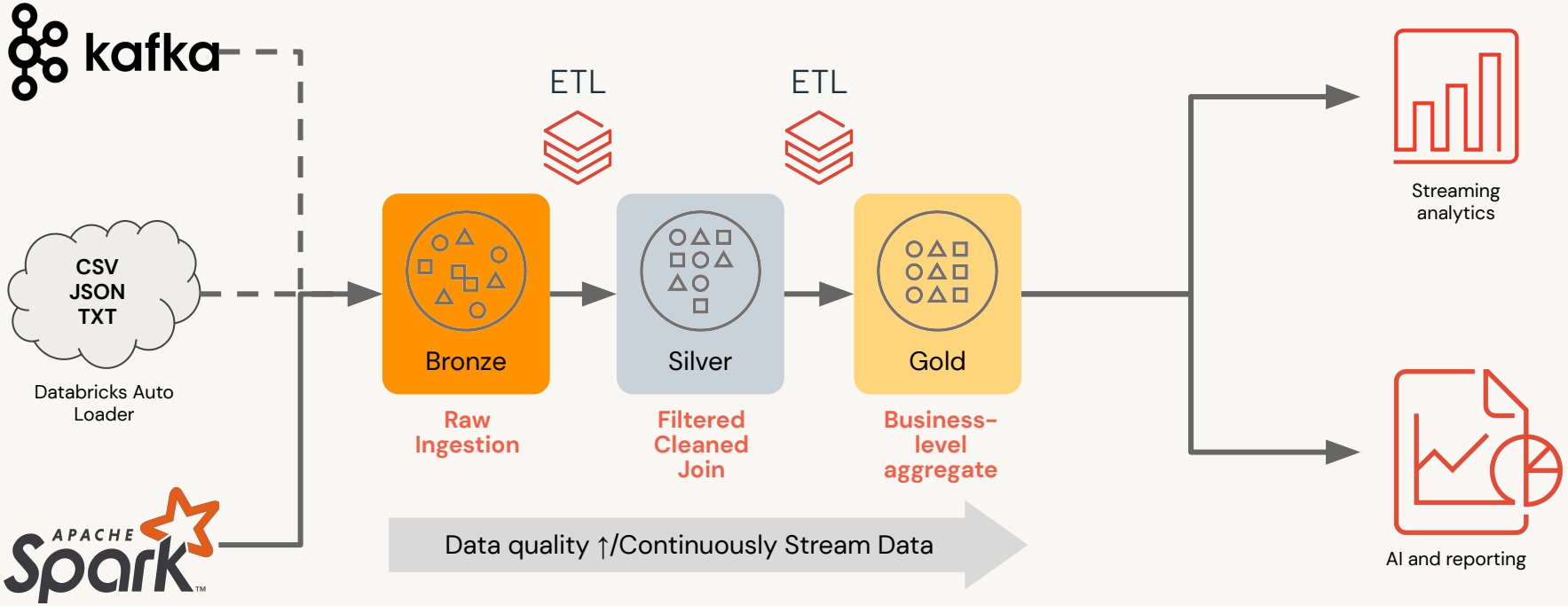




Old: Lambda Architecture



Multi-Hop in the Lakehouse

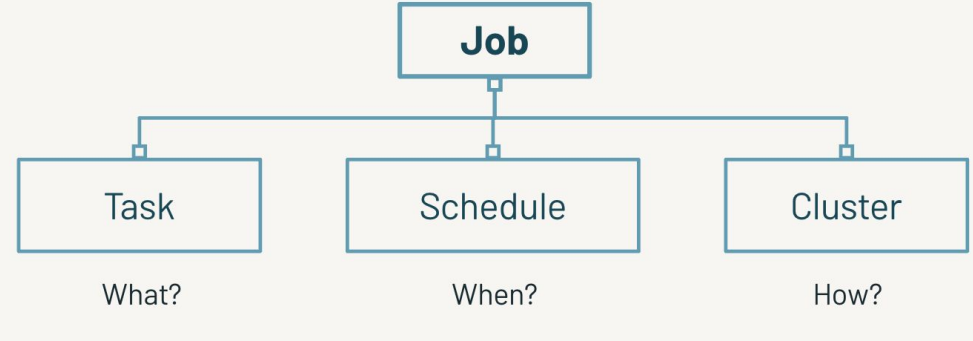


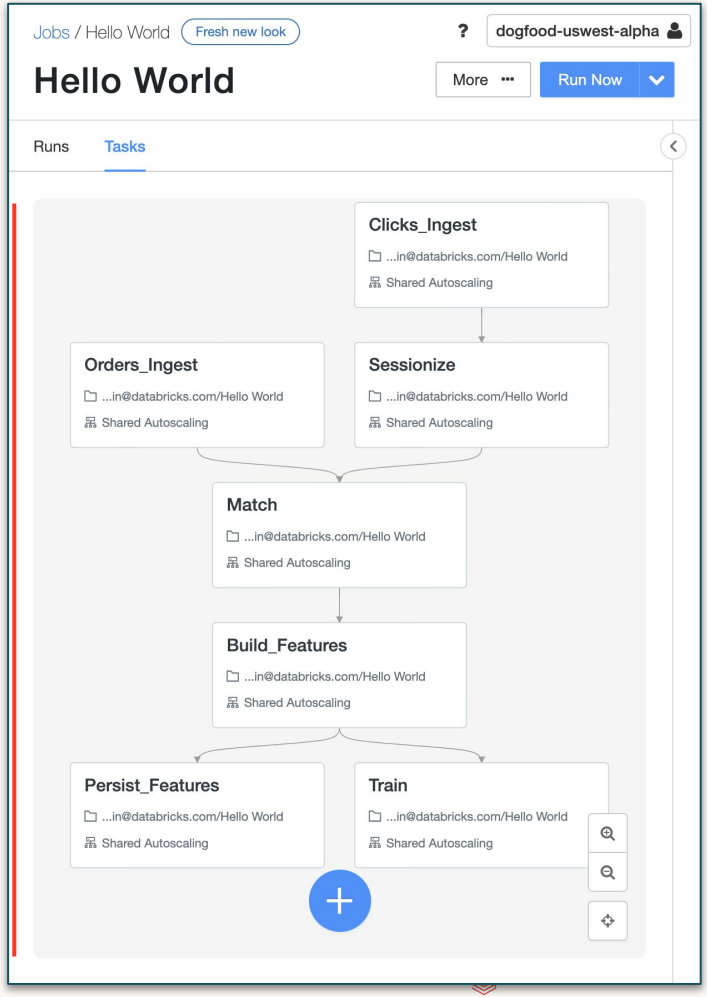
Databricks Workflow

Job

Multi Task Job에 대한 Orchestration/Schedule 기능

DAG of tasks





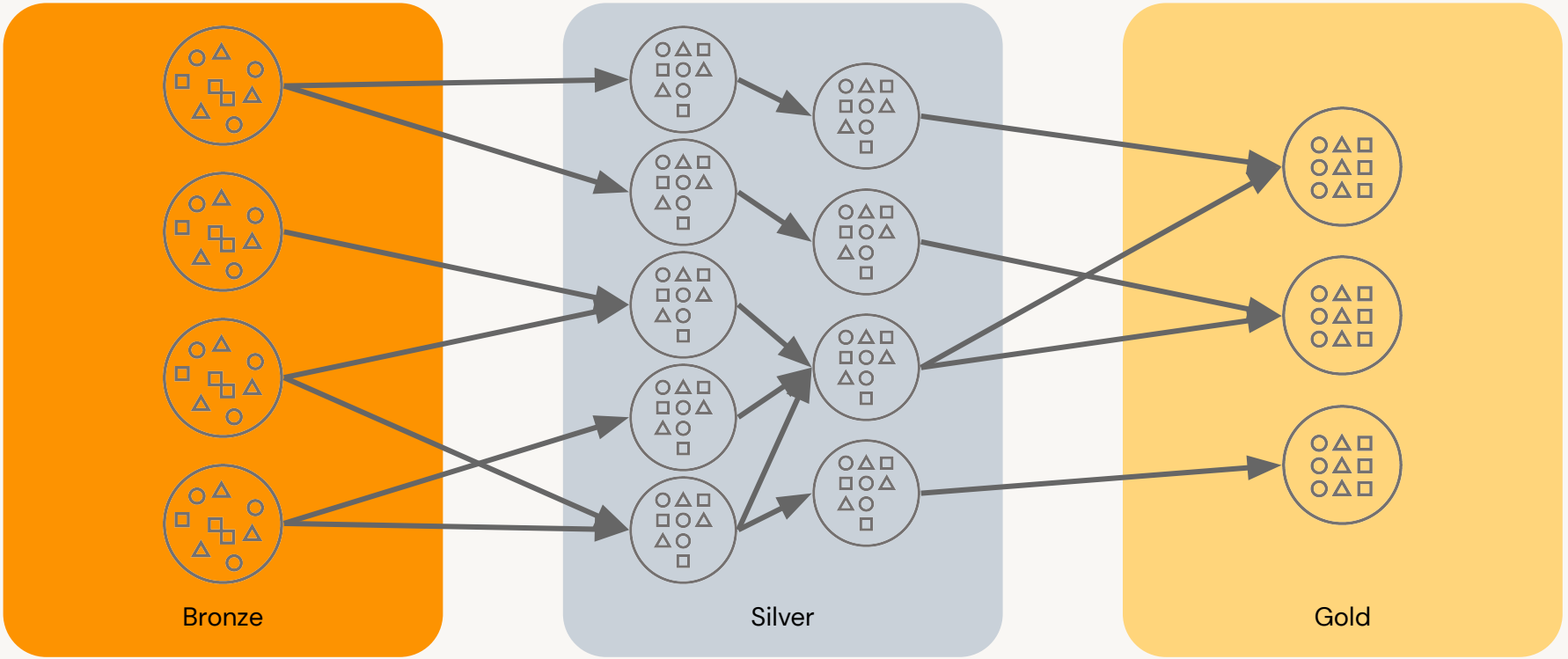
Demo

Databricks Workflow
3. Incremental Multihop datapipeline

Delta Live Table

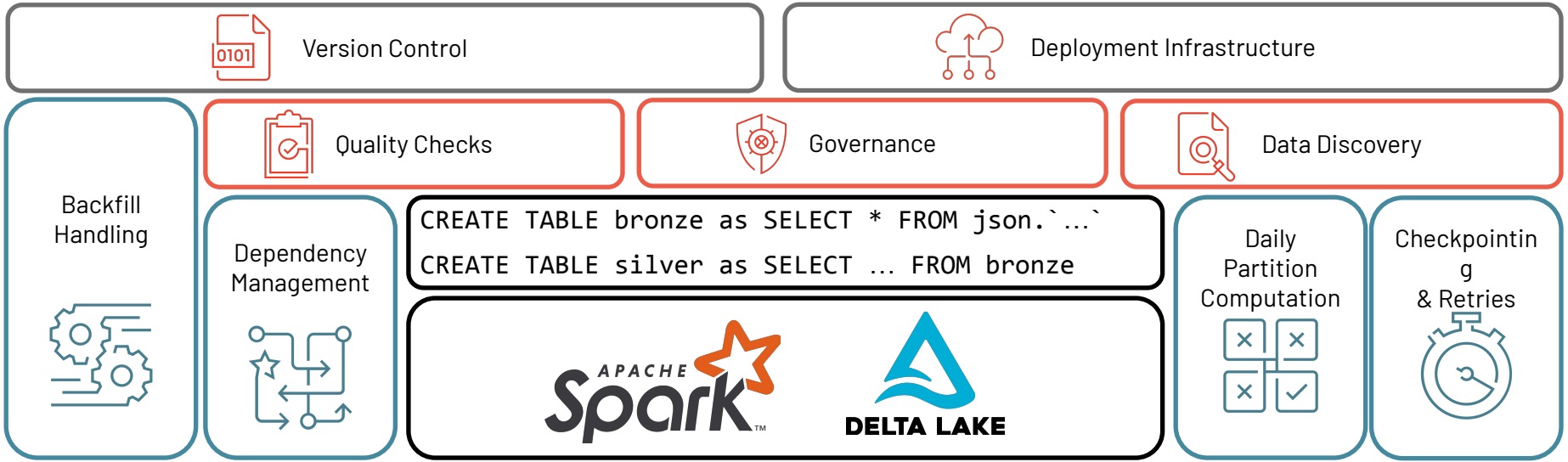
ETL Service made by Databricks

The Reality is Not so Simple



Going From Query to Production

ETL Pipeline을 만들기 위해서 고려해야할 여러가지 사항들



Introducing Delta Live Tables

Delta Lake 상에서 손쉽게 ETL 파이프라인을 생성

Operate with agility

배치/스트리밍 데이터 파이프 라인 구성을 테이블 선언만으로 손쉽게

Trust your data

빌트인 품질 관리 기능 제공

데이터 중복방지,Checkpoint 관리등 자동화

Scale with reliability

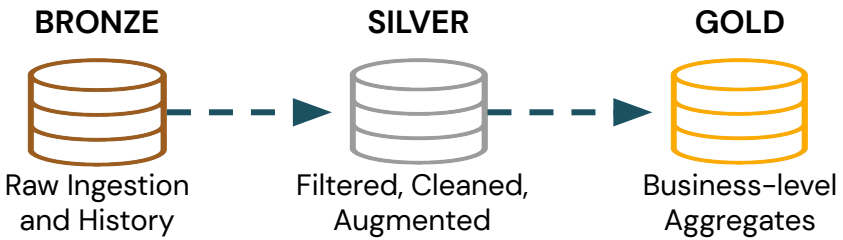
ETL 인프라 관리 자동화

Streaming 워크로드에도 Autoscaling기능 제공

DLT Demo
https://github.com/databricks/delta-live-tables-notebooks

Data Engineering workloads Summary

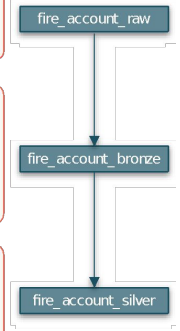
- Databricks Workflow를 이용한 Data Orchestration
- Delta Live Tables 을 이용한 전체 ETL 파이프라인 관리 자동화
- Delta Lake Multihop 아키텍처로 data engineering 작업 단순화

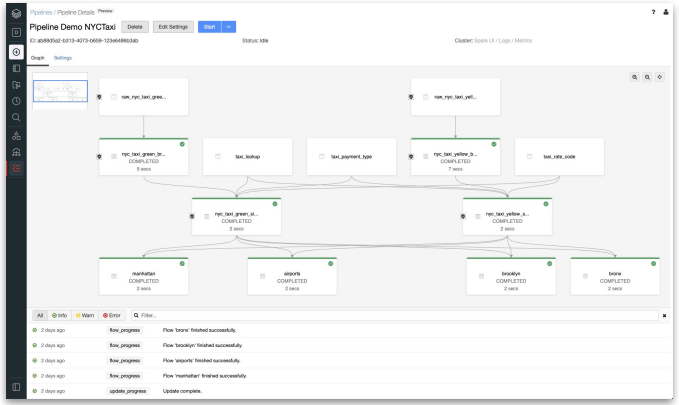



```
/* Create a temp view on the accounts table */
CREATE INCREMENTAL LIVE VIEW fire_account_raw AS
SELECT * FROM cloud_files("/data", "csv");
```

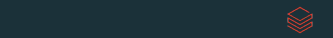
```
/* Stage 1: Bronze Table drop invalid rows */
CREATE INCREMENTAL LIVE TABLE fire_account_bronze AS
COMMENT "Bronze table with valid account ids"
SELECT parse(date), … FROM LIVE.fire_account_raw ...
```

```
/* Stage 2:Send rows to Silver, run validation rules */
CREATE INCREMENTAL LIVE TABLE fire_account_silver AS
COMMENT "Silver Accounts table with validation checks"
SELECT sum(revenue) FROM LIVE.fire_account_bronze ...
```



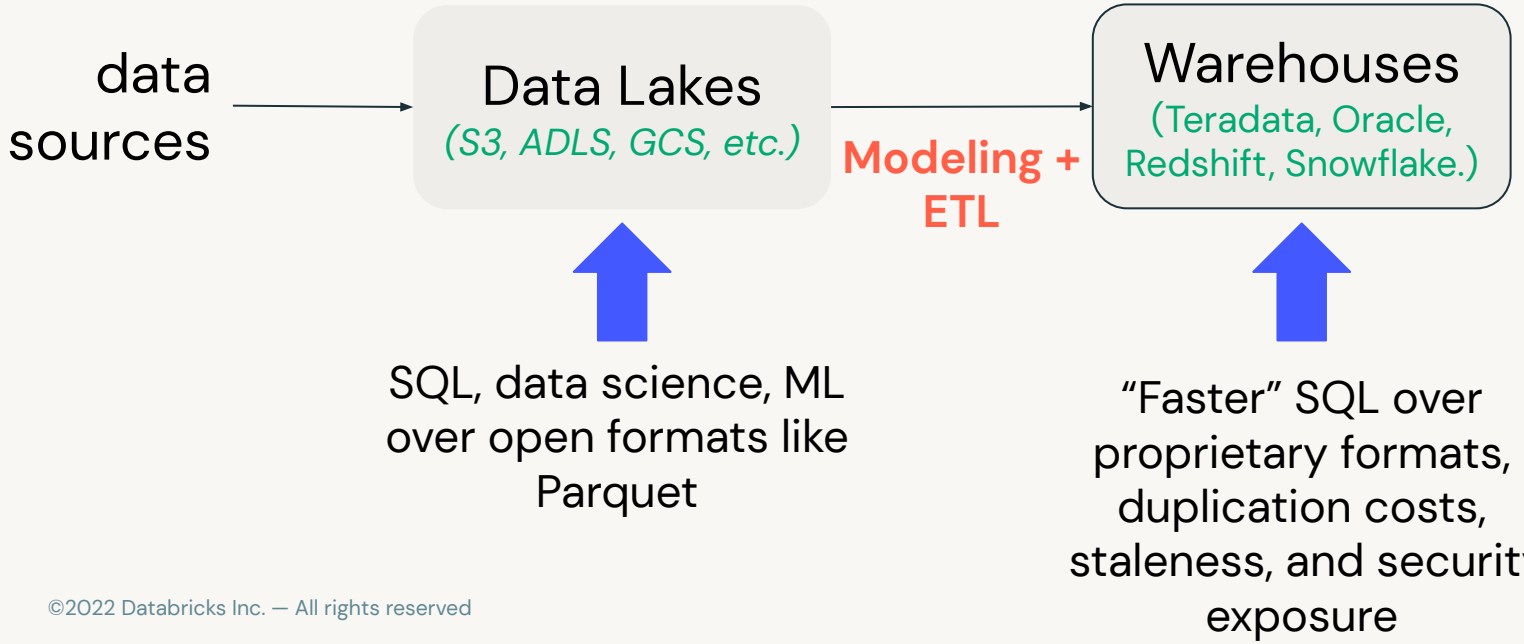


Databricks SQL



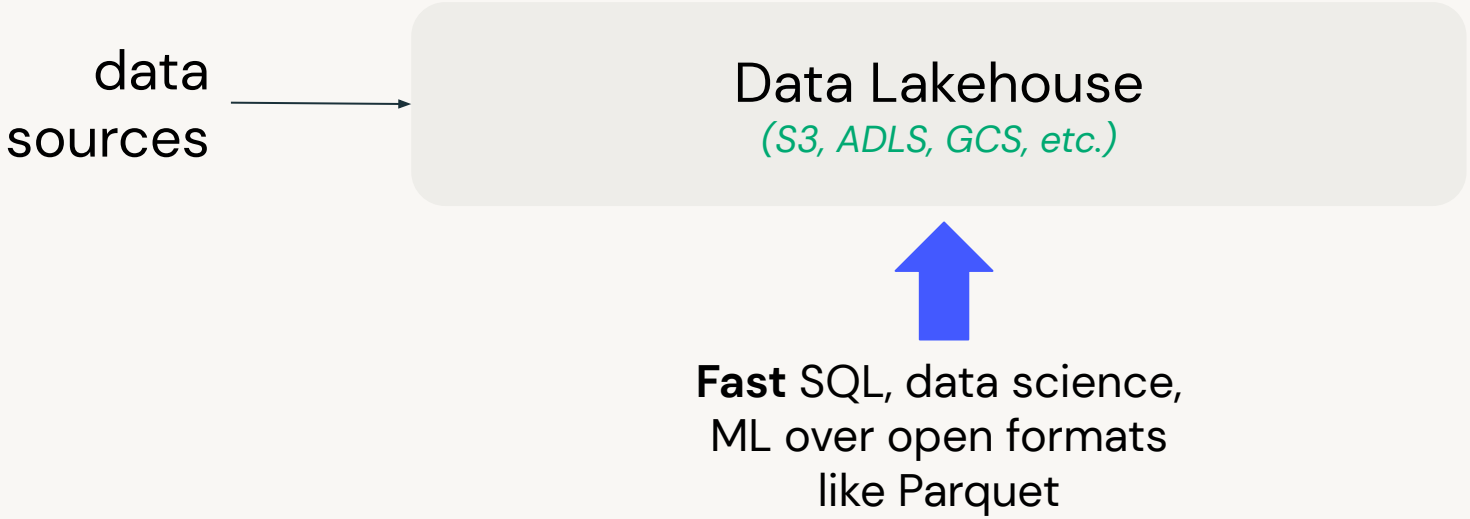
Old Way: 원본은 Data Lake에, BI는 DW에서

중복 데이터로 인한 보안/품질 이슈, 실시간성 결여, 비용 중복 투자

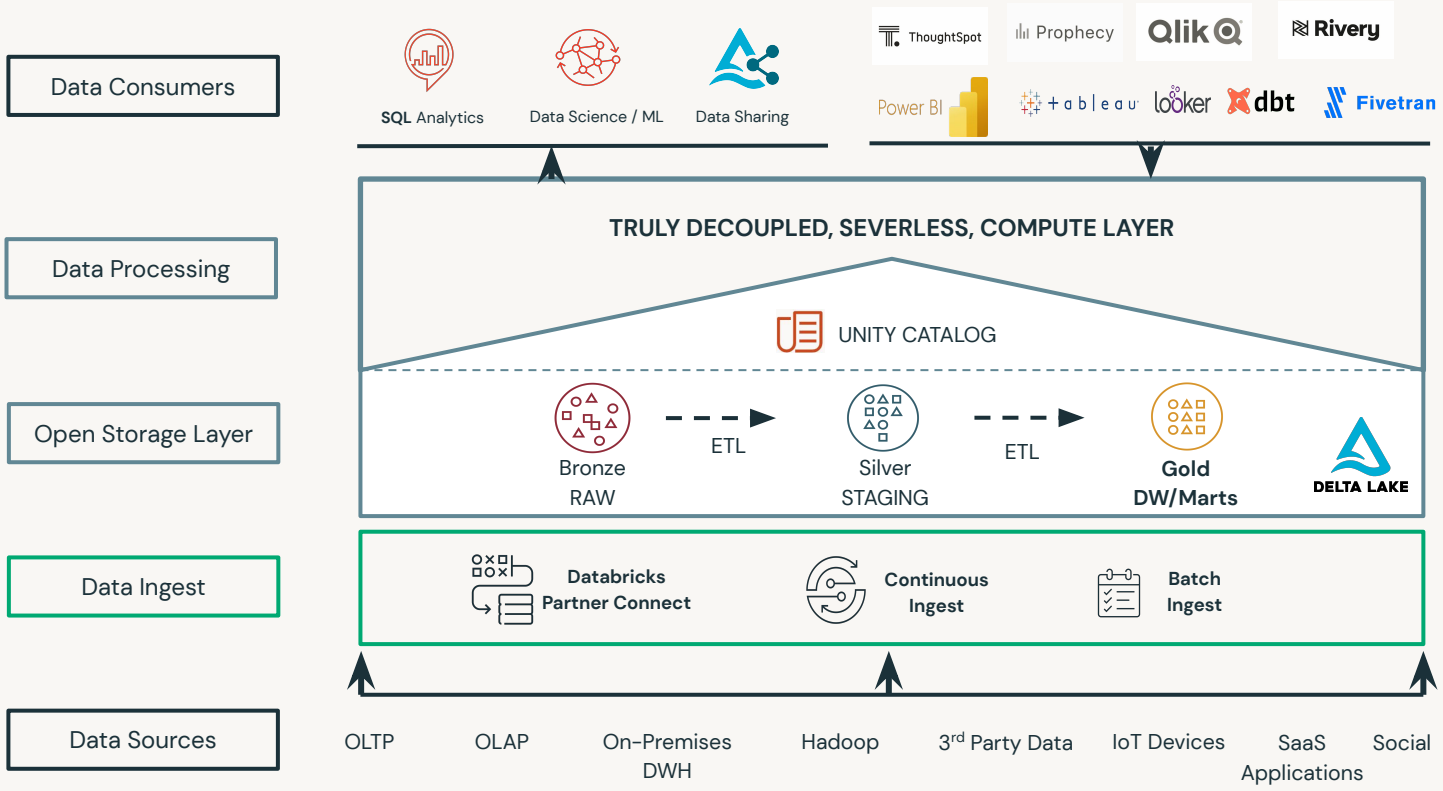


Lakehouse: Unifying Data Lakes and Warehouses

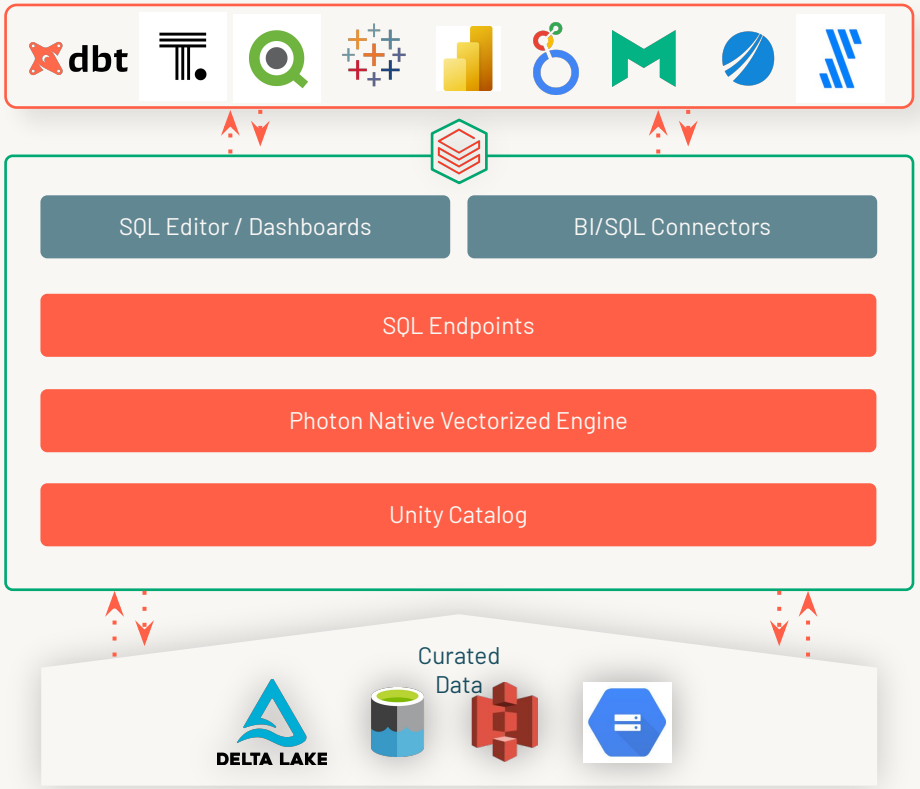
open formats 과 **저비용의 Cloud Storage** 를 활용해서도
상용 데이터웨어하우스와 같거나 더 나은 쿼리 성능을 제공할 수 있다면?



Data Warehousing on Databricks



Databricks SQL 구성 요소



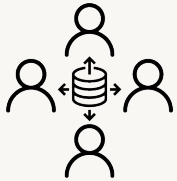
세계적 수준의 성능 및 Data Lake 가성비 동시 제공

- 고객이 선택한 BI 도구로 최신 데이터에 대한 분석 지원
- 모든 쿼리에 빠르고 예측 가능한 성능
  - 최적화된 ODBC/JDBC 드라이버
  - 향상된 큐잉 및 부하 분산
- C++로 작성된 고성능 스파크 실행 엔진 제공(Photon)
- 시스템 관리 편의성 및 상세한 거버넌스 제공
- 개방형, 신뢰성 있는 데이터 레이크를 데이터 기반으로 활용

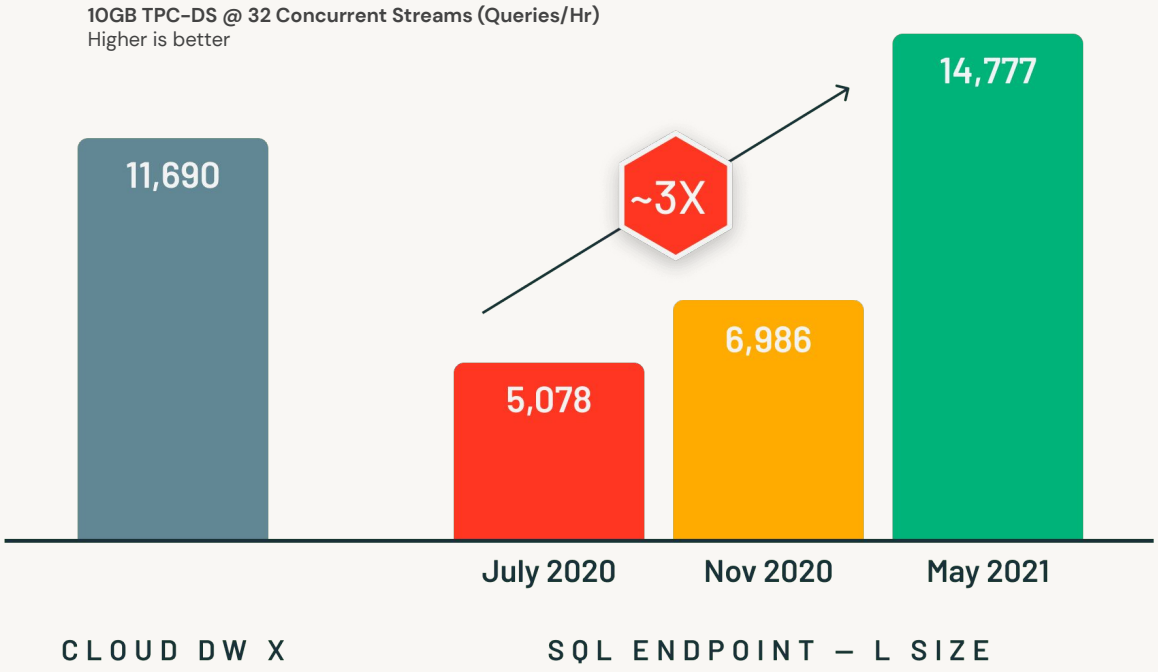
모든 쿼리에 빠르고 예측 가능한 성능

대형 쿼리 뿐만 아니라 작은 쿼리 실행 성능도 개선

높은 동시 접속



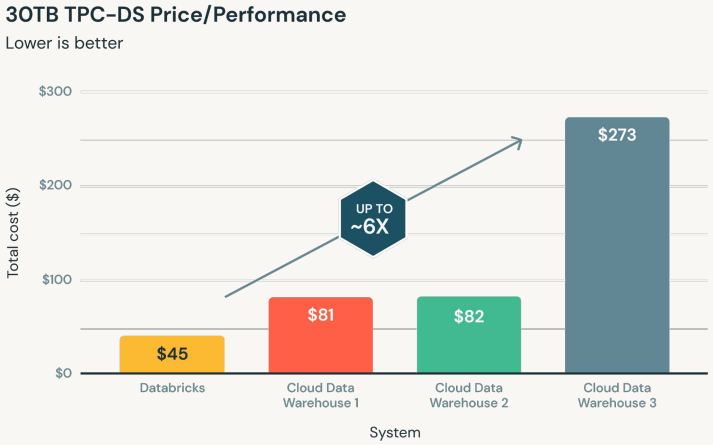
시스템 오버헤드를
줄이고 QpH를 3배 향상

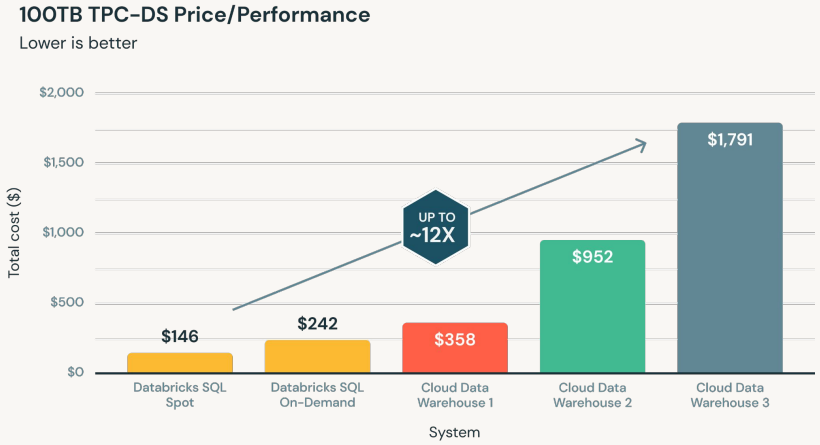


높은 가격 대비 성능 제공

빠른 분석 성능

레거시 클라우드 데이터 웨어하우스보다 최대 12배 더 나은 가격 대비 성능으로 최신의 모든 데이터를 쿼리하고 분석





Source: Performance Benchmark with Barcelona Supercomputing Center

Databricks SQL : New World Record(100TB TPC-DS)

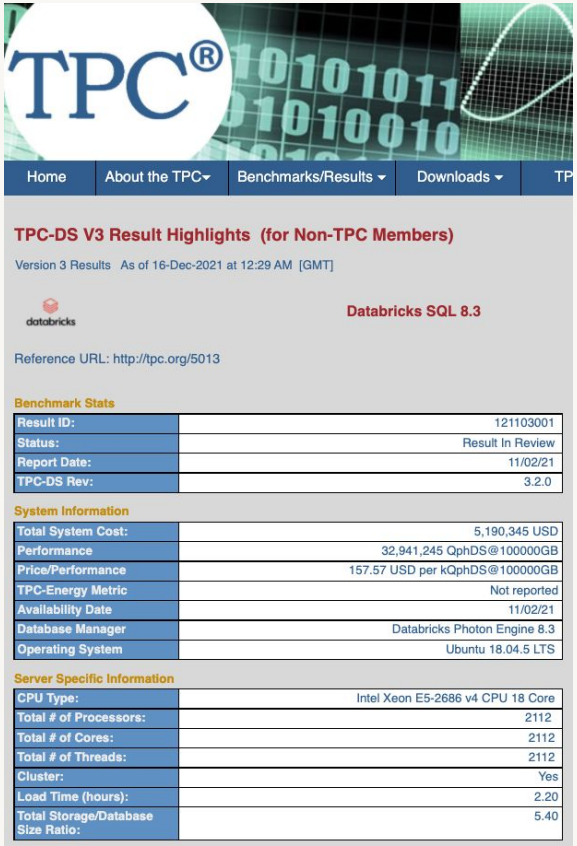
기존 최고 성능 대비 2.2배 높은 결과

Databricks Sets Official Data Warehousing Performance Record

by **Reynold Xin** and **Mostafa Mokhtar**
Posted in **COMPANY BLOG** | **November 2, 2021**

Today, we are proud to announce that **Databricks SQL** has set a **new world record in 100TB TPC-DS**, the gold standard performance benchmark for data warehousing. **Databricks SQL outperformed the previous record by 2.2x**. Unlike most other benchmark news, this result has been formally audited and reviewed by the TPC council.

These results were corroborated by research from **Barcelona Supercomputing Center**, which frequently runs TPC-DS on popular data warehouses. **Their latest research benchmarked Databricks and Snowflake, and found that Databricks was 2.7x faster and 12x better in terms of price performance**. This result validated the thesis that data warehouses such as Snowflake become prohibitively expensive as data size increases in production.



TPC®

Home | About the TPC | Benchmarks/Results | Downloads | TP

TPC-DS V3 Result Highlights (for Non-TPC Members)

Version 3 Results As of 16-Dec-2021 at 12:29 AM [GMT]

databricks

Databricks SQL 8.3

Reference URL: http://tpc.org/5013

Benchmark Stats

Result ID:	121103001
Status:	Result In Review
Report Date:	11/02/21
TPC-DS Rev:	3.2.0

System Information

Total System Cost:	5,190,345 USD
Performance	32,941,245 QphDS@100000GB
Price/Performance	157.57 USD per kQphDS@100000GB
TPC-Energy Metric	Not reported
Availability Date	11/02/21
Database Manager	Databricks Photon Engine 8.3
Operating System	Ubuntu 18.04.5 LTS

Server Specific Information

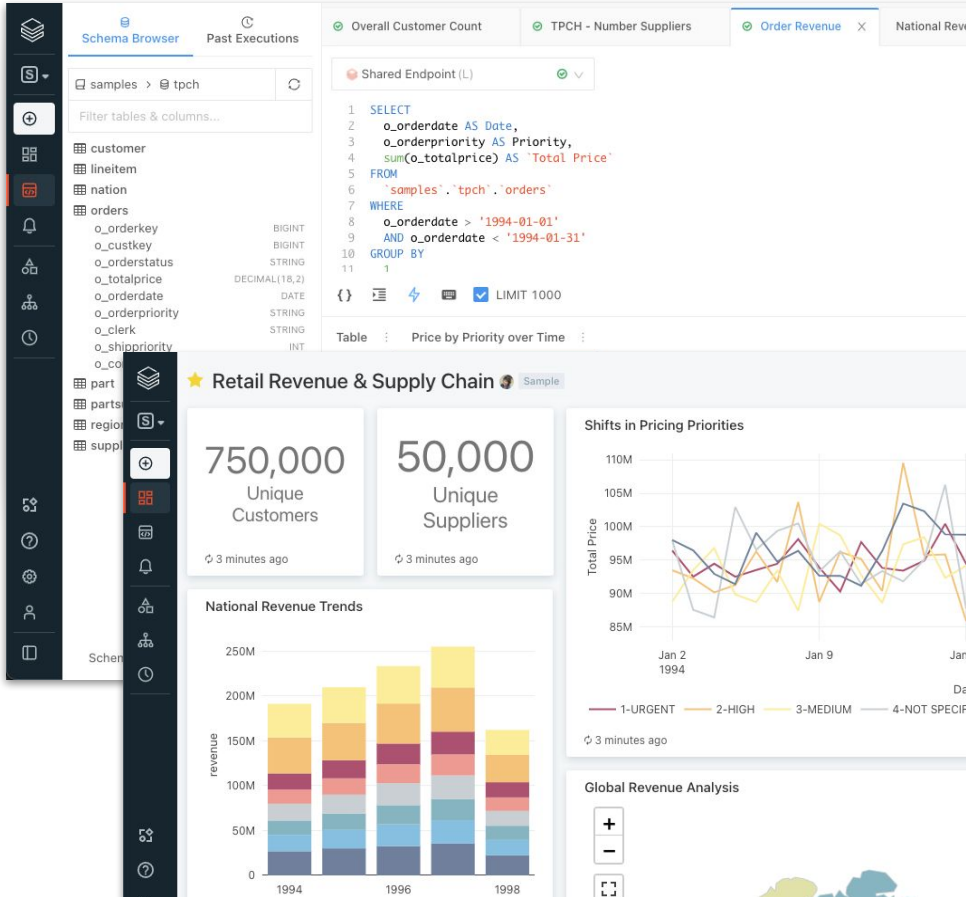
CPU Type:	Intel Xeon E5-2686 v4 CPU 18 Core
Total # of Processors:	2112
Total # of Cores:	2112
Total # of Threads:	2112
Cluster:	Yes
Load Time (hours):	2.20
Total Storage/Database Size Ratio:	5.40

Demo

Databricks SQL Endpoint
SQL editor
Visualization Dashboard
Alert

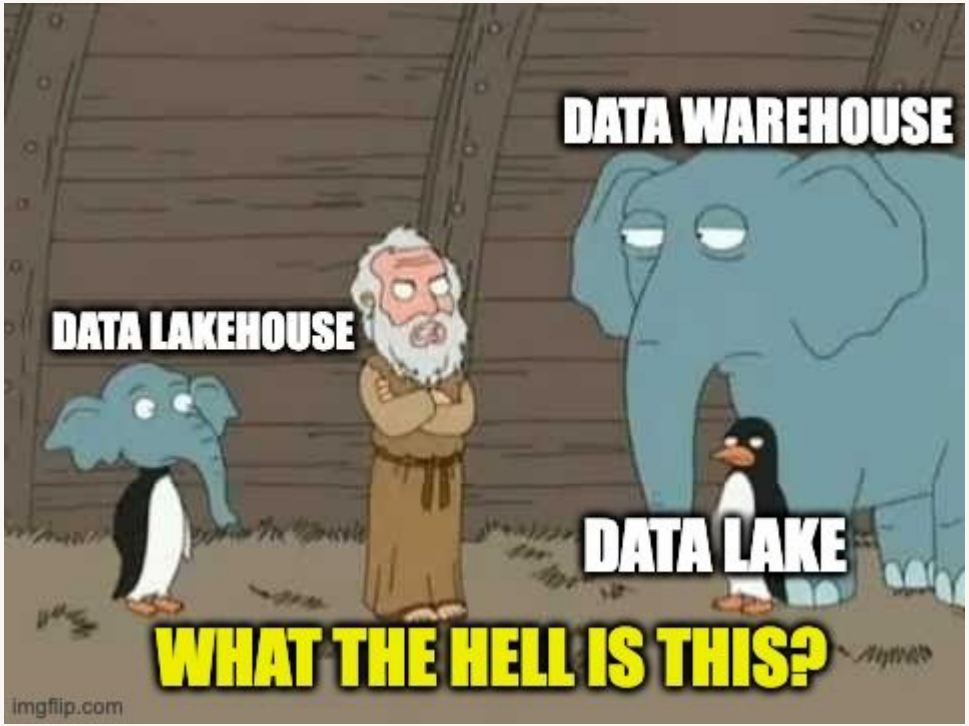
SQL workloads on Databricks Summary

- Delta Lake 상에서 BI와 SQL workload 처리를 위한 고성능 + 동시성 향상 쿼리 엔진
- 분석가들을 위한 Native SQL 인터페이스 지원
- DL → DW 이중저장 구조 제거로 TCO 절감/분석 파이프라인 단순화
- Built-in dashboard 또는 기존 BI툴들을 통해 빠른 인사이트 확보



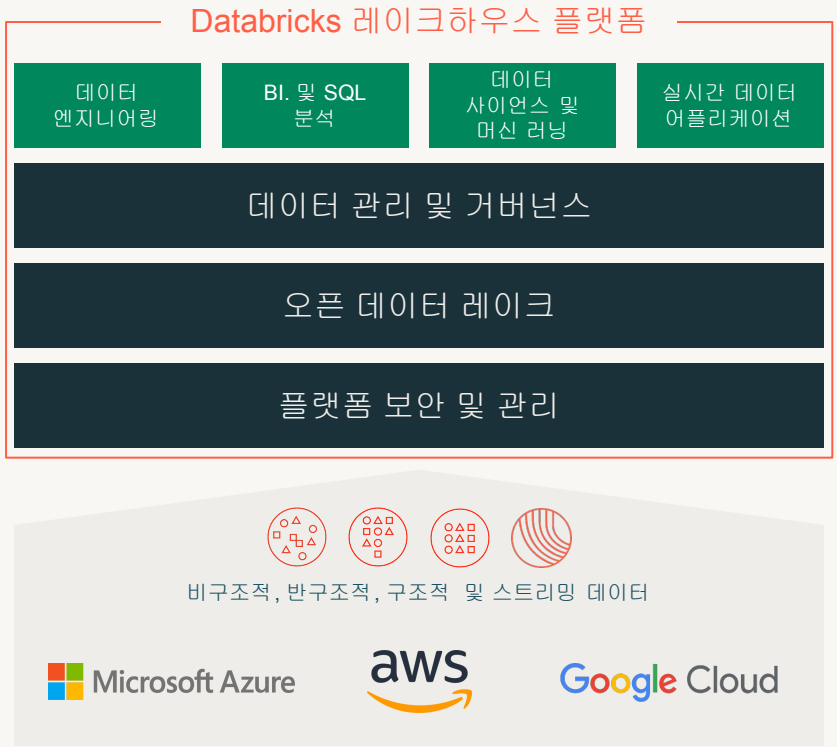
databricks

Summary



Databricks 레이크하우스 플랫폼

- ✓ 단순성
- ✓ 개방성
- ✓ 협업



Databricks Lakehouse

Resources

- 한국어 웹사이트 오픈! https://databricks.com/kr/
- Databricks Documentation (http://docs.databricks.com)
- Databricks Academy(https://academy.databricks.com)
- Solutions Accelerator(https://databricks.com/kr/solutions)
- Help Center(http://help.databricks.com)
- Today's Deck & Notebooks(https://tinyurl.com/db-lakehouse-webinar)


databricks