

Smart Tech Korea 2022

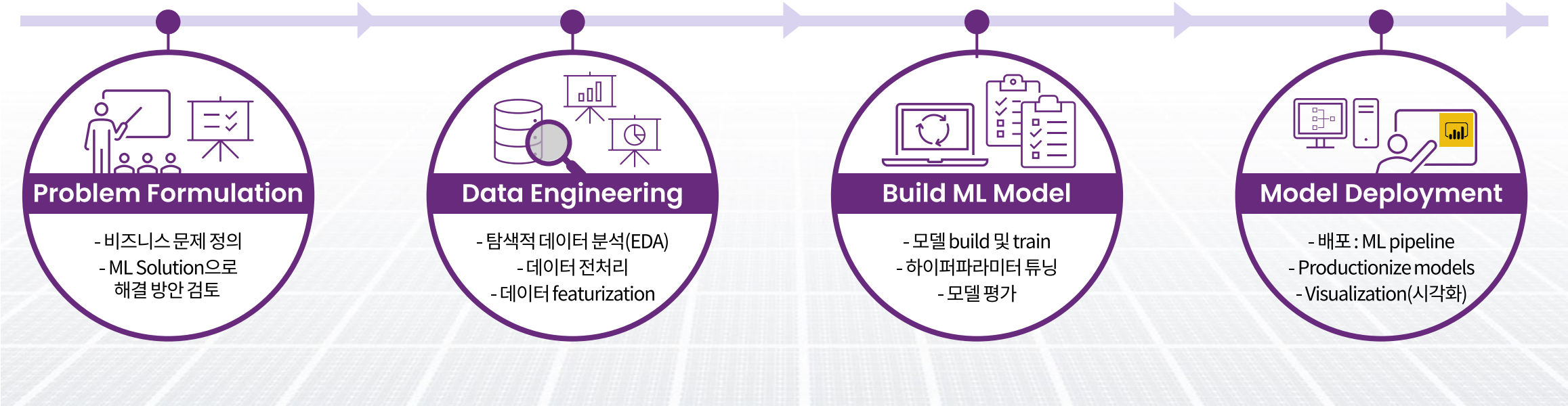
# Data Science 및 Machine Learning을 위한 데이터브릭스 활용 사례

Cloocus

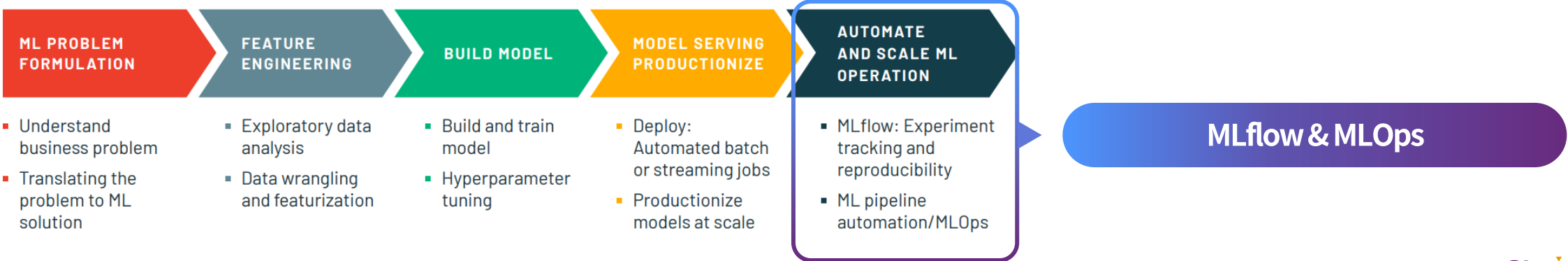
# ML pipeline workflow

일반적인 머신러닝의 워크플로우는 다음과 같습니다.

## ML pipeline workflow

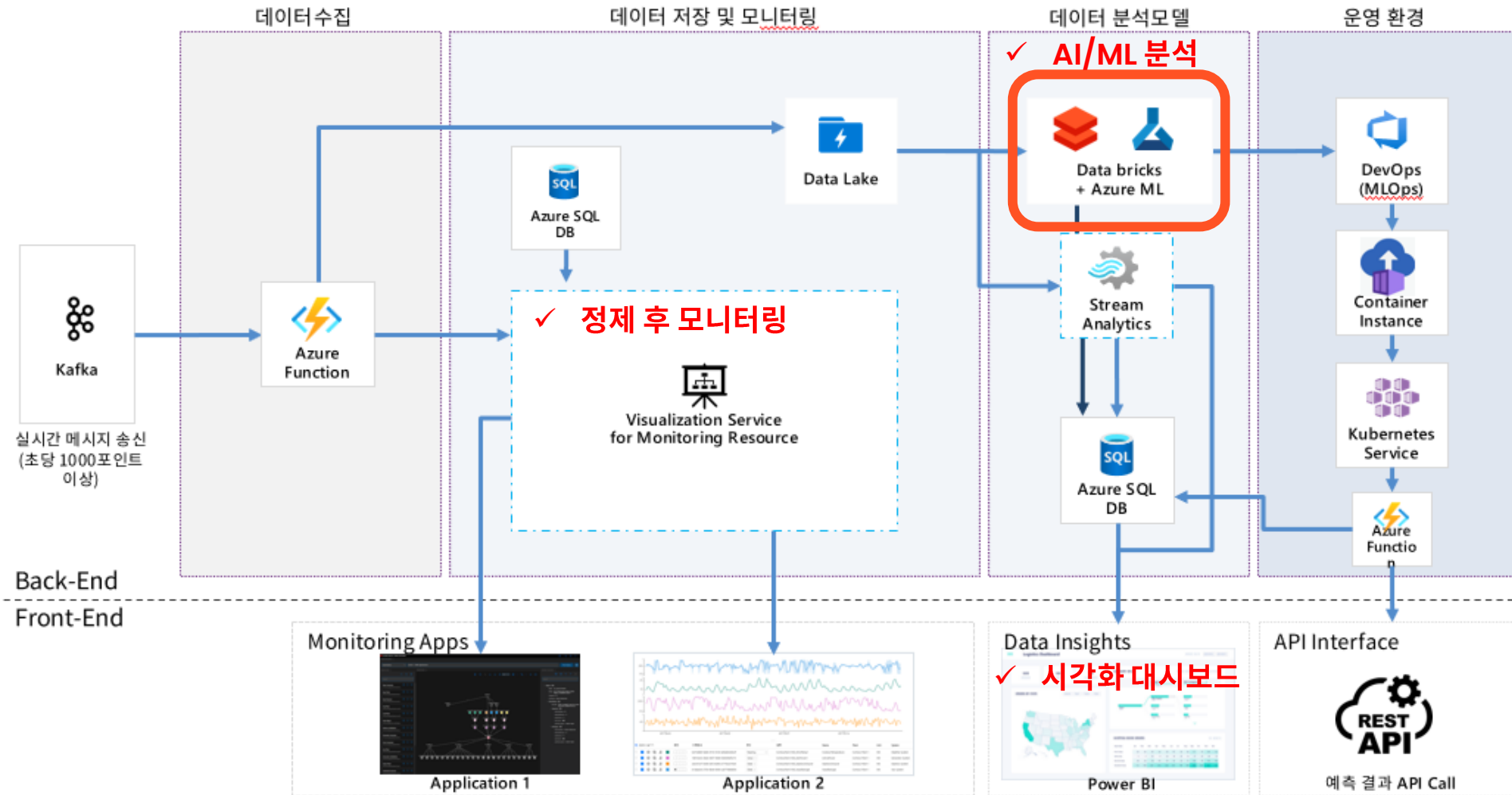


## Databricks ML pipeline workflow



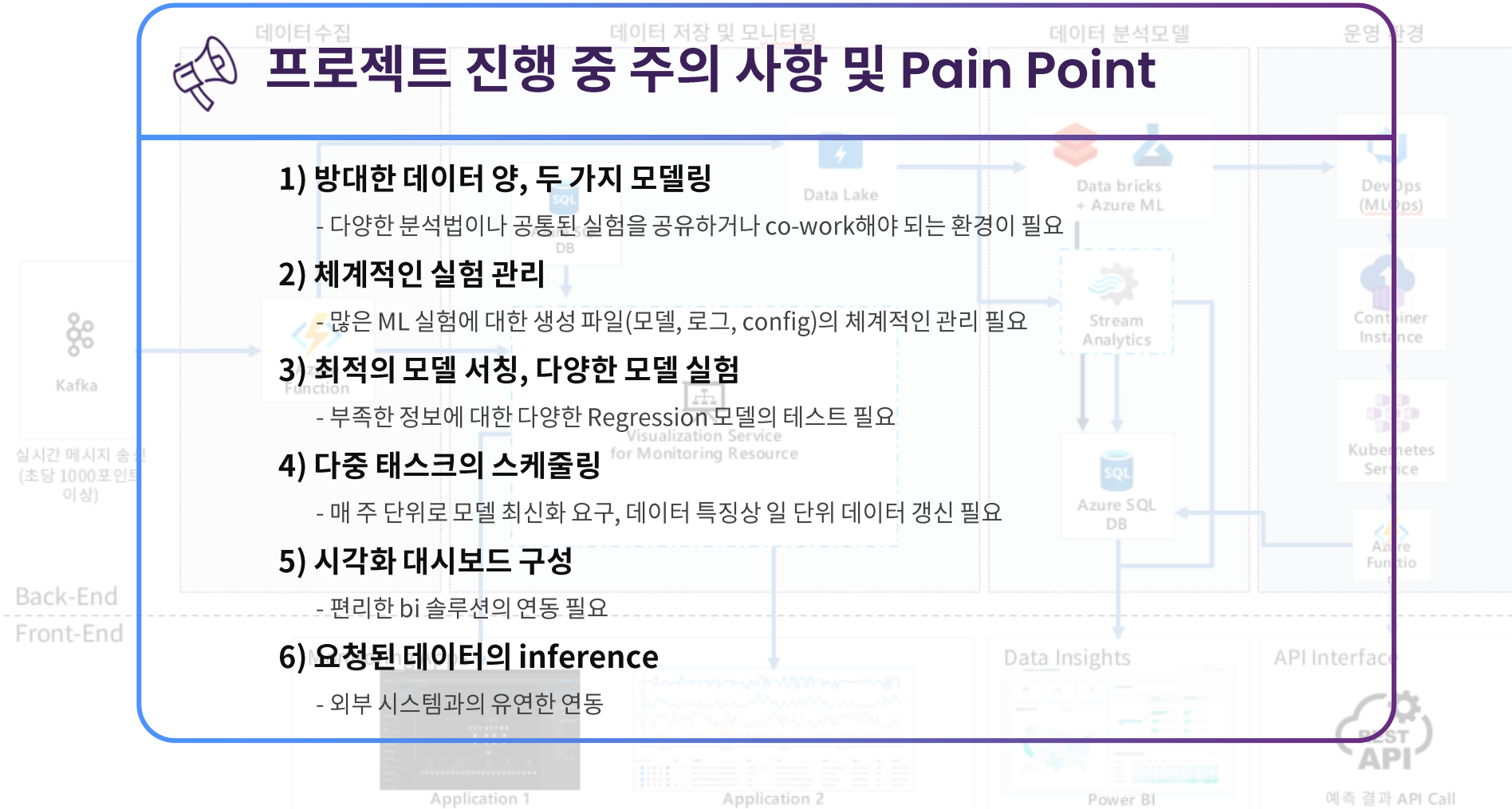
## Scenario

### 1. A사에 대한 시나리오(아키텍처)



## Scenario

### 1. A사에 대한 시나리오



# 1) 방대한 데이터 양, 두 가지 모델링

팀원들 간에 다양한 분석 결과나 공통된 실험을 공유할 수 있고, co-work 환경을 지원합니다.

## Collaborative data science with multi languages and tools



Databricks의 노트북으로 EDA, 데이터 분석, 모델 학습 등을 협업할 수 있다.

Collaboration with 2 people

The screenshot shows the Databricks workspace interface. On the left, the 'Workspace' sidebar is visible with a 'Users' list. A 'Create Notebook' dialog is open, showing a dropdown menu for 'Language' with 'R' selected. A language selection menu is also shown, listing Python, Markdown, Python (Notebook default), Scala, SQL, and R, with 'SQL' highlighted. The main notebook area displays a SQL query and its results in a table format. A red box highlights the user avatars in the top right corner, and an orange callout box provides information about the workspace.

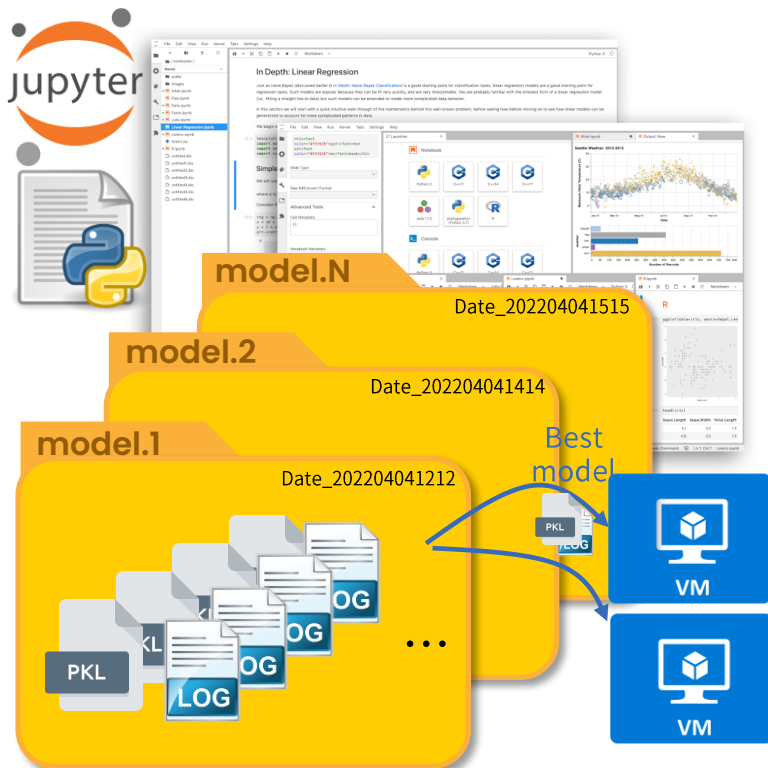
loan_status	int_rate	revol_util	issue_d	earliest_cr_line	emp_length	verification_status	total_pymnt	loan_amnt	grade	annual_inc	dti	addr_state	term
Fully Paid	6.39	55	Apr-2015	Dec-1999	8	Not Verified	8962.76	8400	A	120000	14.33	NY	36 mor
Fully Paid	6.39	15.4	Mar-2015	Sep-1989	7	Not Verified	12659.08	12000	A	75000	4.66	NY	36 mor
Fully Paid	6.39	37.5	Apr-2015	Jan-2005	10	Verified	9670.3837188229	9000	A	93000	16.22	TX	36 mor
Fully Paid	6.39	91.4	May-2015	Aug-1997	7	Verified	6765.3000000055	6400	A	115000	27.47	CA	36 mor
Fully Paid	6.39	83.2	Apr-2015	Dec-1974	10	Verified	21151.09	20000	A	300000	10.19	SD	36 mor

**Workspace:** One central place to store and share notebooks, experiments, and projects backed with role-based access control.

많은 ML 실험에 대한 생성 파일(모델, 로그, config 등)을 체계적으로 관리 할 수 있습니다.

### Efficient end-to-end ML Pipeline

Databricks는 mlflow를 기본적으로 제공하여 쉽게 ML의 lifecycle을 관리할 수 있습니다.



mlflow on databricks

#### Tracking

Record and query experiments: code, data, config, results

#### Projects

Packaging format for reproducible runs on any platform

#### Models

General format for sending models to diverse deploy tools



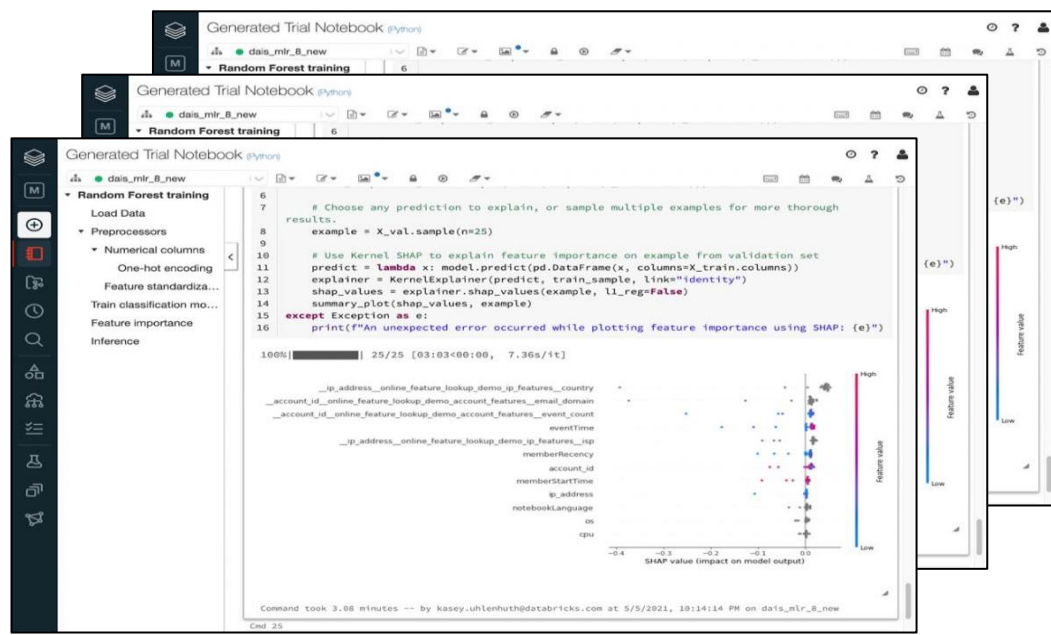
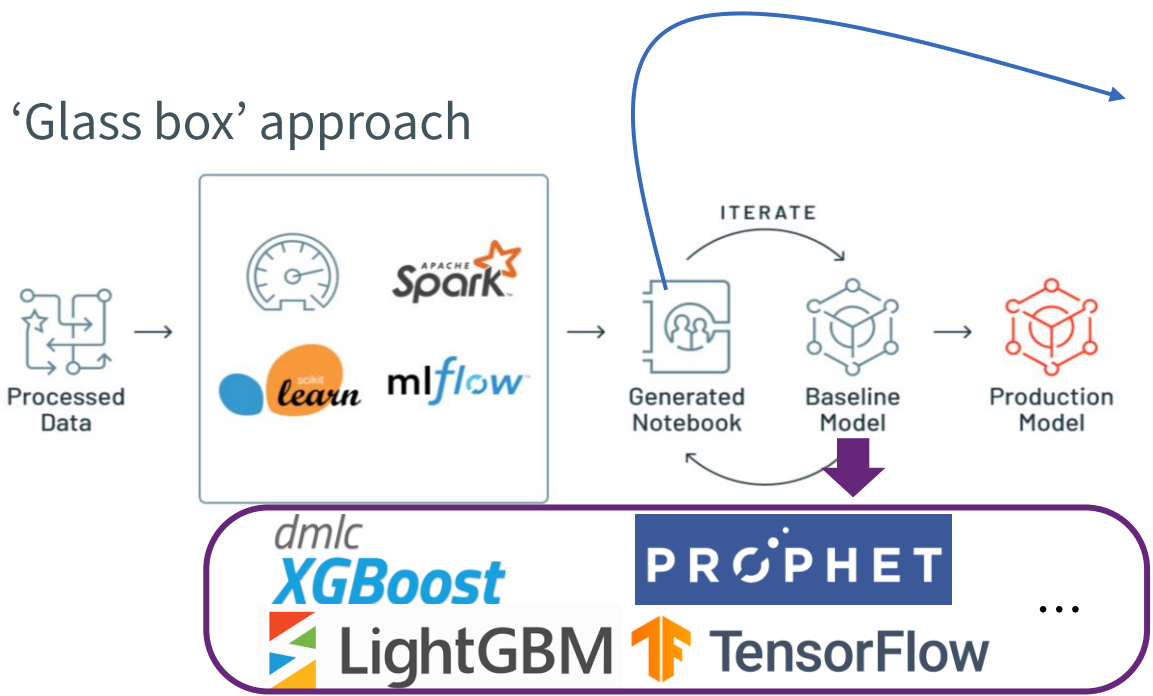
### 3) 최적의 모델 서칭, 다양한 모델 실험

다양한 Regression 모델을 빠르게 테스트 할 수 있게 AutoML notebook을 제공합니다.

#### A "Glass box" approach to AutoML

Databricks의 AutoML은 실행된 결과에 대한 수정 가능한 노트북을 제공하여 쉽고 빠르게 baseline 모델을 튜닝 할 수 있습니다.

'Glass box' approach

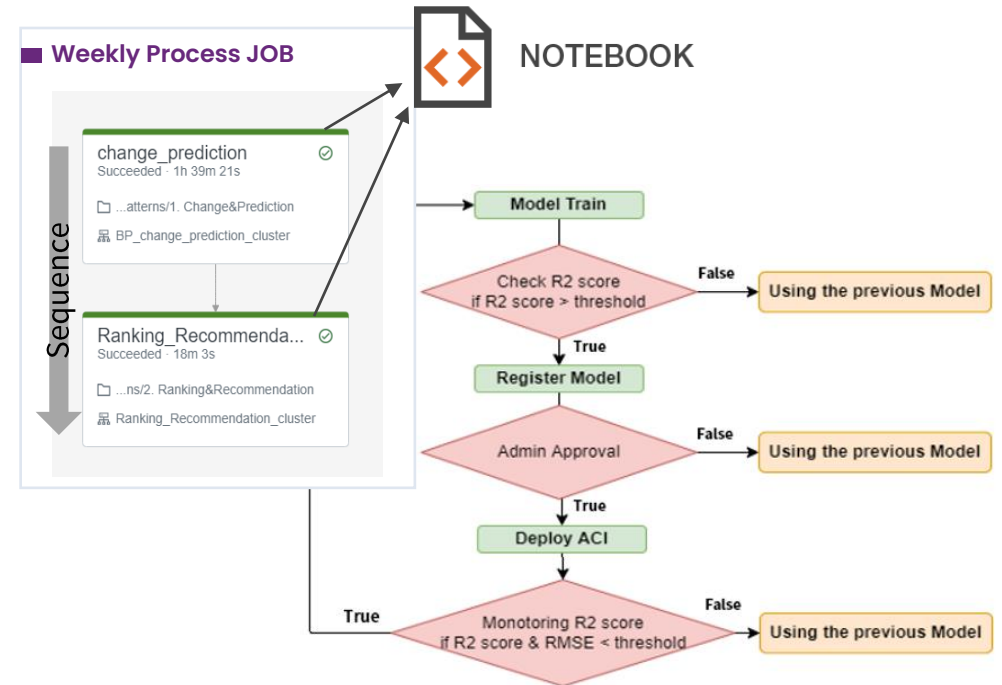
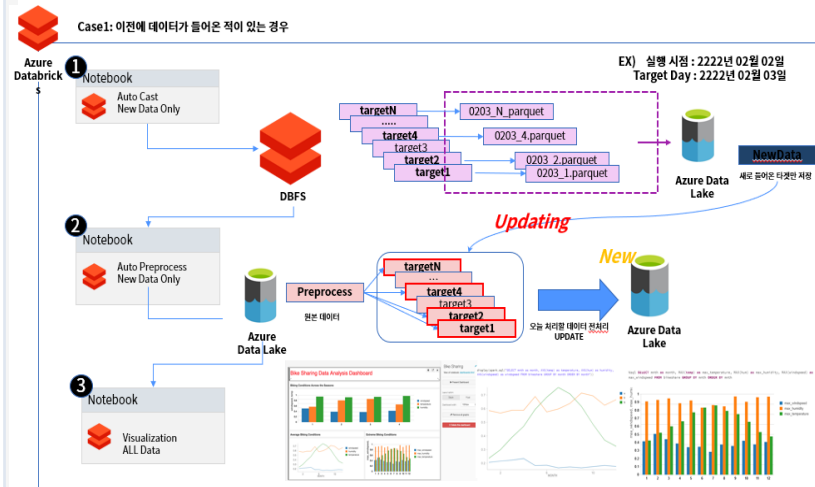
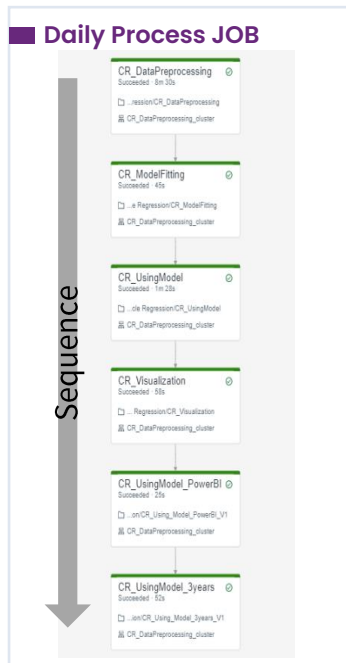


Generated notebooks

주 단위, 일 단위의 다양한 작업 스케줄링을 쉽게 구현하고 관리할 수 있습니다.

## Orchestrate Multiple Tasks with Databricks Jobs

Databricks는 job 기능을 통해 원하는 task를 주기적으로 관리할 수 있습니다.



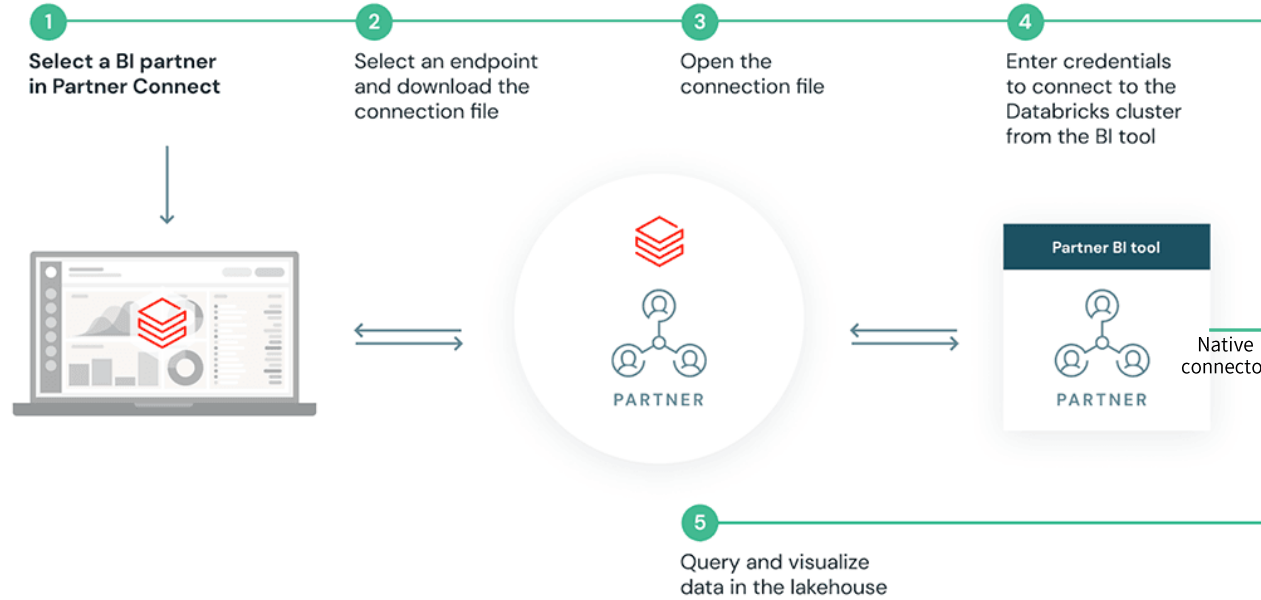


Databricks를 통해 편리하게 bi 솔루션의 연동을 할 수 있습니다.

## Navigate tables and views to start visualizing data

Databricks의 native connector를 통해 시각화 대시보드를 쉽게 구현할 수 있습니다.

- ☑ 거버넌스 수립
- ☑ 데이터 기반 의사결정



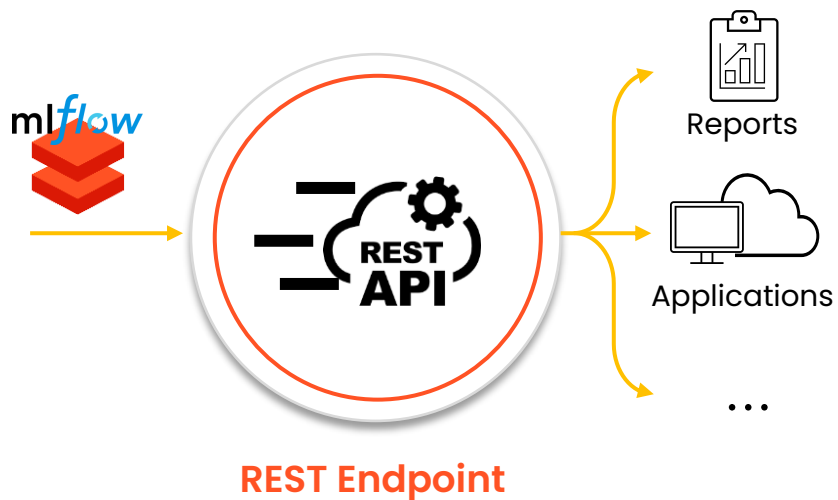
BI 솔루션들



Databricks는 외부 시스템과의 유연한 연동이 가능합니다.

### Model Serving on Databricks

Databricks는 등록된 모델에 대한 REST API endpoint를 제공하여 모델을 원하는 서비스에 배포할 수 있습니다.



### Inference with simple request

Model Versions    Model Events    Cluster Settings

Model Versions

Version 1 **활성화 상태** ● Ready

Version 1  
Model URL: [https://adb-2732377089277677.17.azuredatabricks.net/model/test\\_model/1/invocations](https://adb-2732377089277677.17.azuredatabricks.net/model/test_model/1/invocations)

Call the model

Browser    Curl    Python

Request **?**

```
{
  "columns": ["sepal length (cm)", "sepal width (cm)", "petal length (cm)", "petal width (cm)"],
  "data": [[5.2, 3.4, 1.3, 0.3], [6.0, 3.1, 5.0, 1.81]]
}
```

Response **?**

```
[0.2]
```

Send Request    Show Example

이렇게 프로젝트를 진행하며 발생하는 다양한 상황과 pain points를 Databricks를 통해 쉽게 해결할 수 있습니다.

## 클라우드와 함께하는 AI / ML 서비스

### 1) 방대한 데이터 양, 두 가지 모델

- 다양한 분석법이나 공통된 실험을 공유하거나 co-work해야 되는 환경이 필요

### 2) 체계적인 실험 관리

- 많은 ML 실험에 대한 생성 파일(모델, 로그, config)의 체계적인 관리 필요

### 3) 최적의 모델 서칭, 다양한 모델 실험

- 부족한 데이터에 대한 다양한 Regression 모델의 테스트 필요

### 4) 다중 태스크의 스케줄링

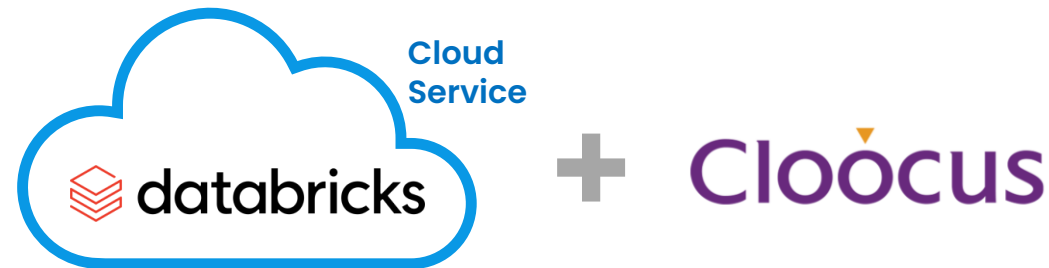
- 매 주 단위로 모델 최신화 요구, 데이터 특징상 일 단위 데이터 갱신 필요

### 5) 시각화 대시보드 구성

- 편리한 bi 솔루션의 연동 필요

### 6) 요청된 데이터의 inference

- 외부 시스템과의 유연한 연동



# ■ Demo

1. Azure Databricks **AutoML**
2. Azure Databricks **MLflow**



Smart Tech Korea 2022

# 빅데이터 분석을 위한 데이터 레이크하우스 파이프라인

Cloócus

ELT?

ETL PIPELINE?

Data Lake?

Data Warehouse?

Real Time DATA?

데이터 관리 방법?

시각화 해야 하는데..

Lakehouse?

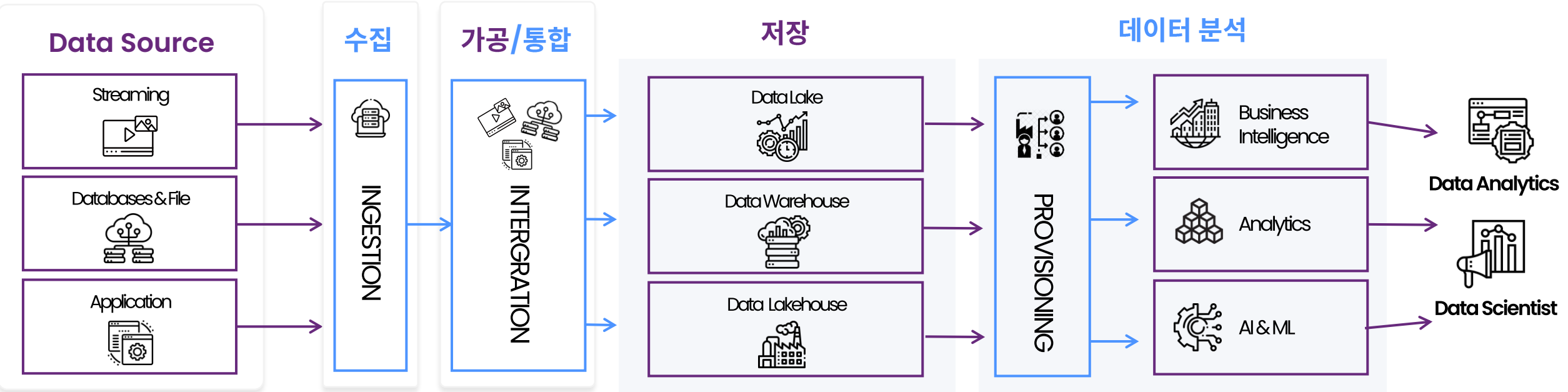
데이터는 가지고 있는데..

AI & ML..





## 데이터 전략



## MANAGEMENT

Metadata | Quality | Governance | Privacy | Protection | Master Data

## DATAOPS

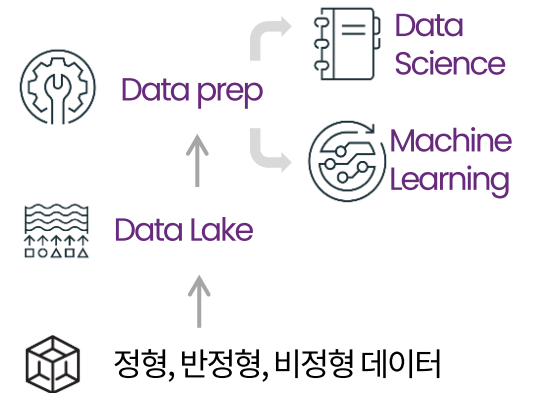
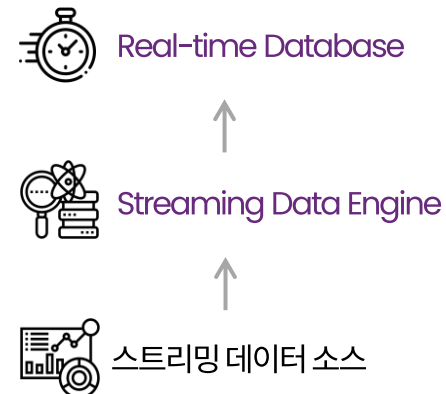
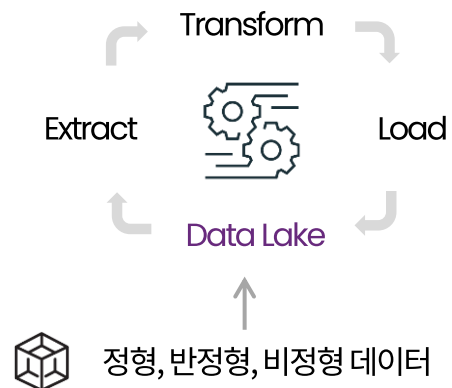
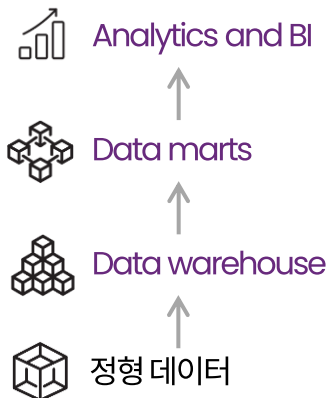
## 데이터 웨어하우징

## 데이터 엔지니어링

## 스트리밍

## DS & ML

### 사일로화된 기술 스택 때문에 점점 더 복잡해지는 데이터 아키텍처



## 데이터 웨어하우징

## 데이터 엔지니어링

## 스트리밍

## DS & ML

### 연결되지 않은 시스템 및 독점적 데이터 포맷은 통합 난이도를 높임

Amazon Redshift  
Azure Synapse  
Snowflake  
SAP

Teradata  
Google BigQuery  
IBM Db2  
Oracle Autonomous  
Data Warehouse

Hadoop  
Amazon EMR  
Google Dataproc

Apache Airflow  
Apache Spark  
Cloudera

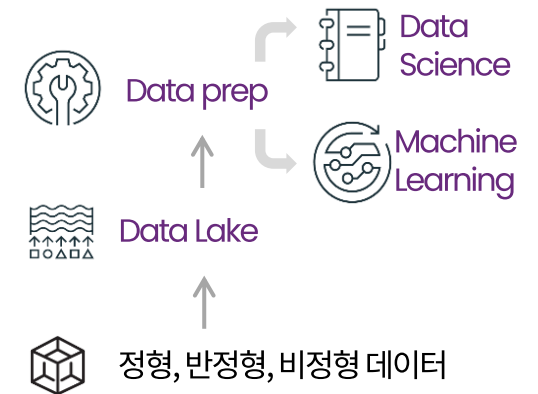
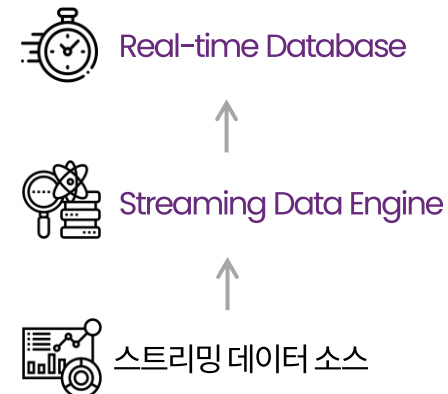
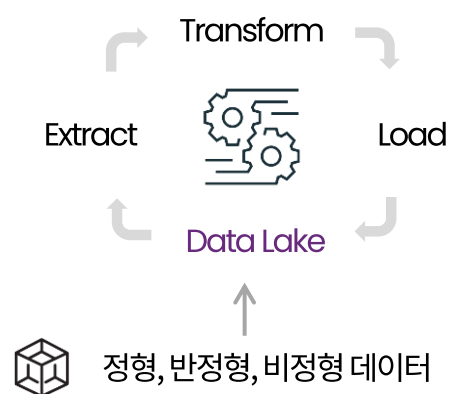
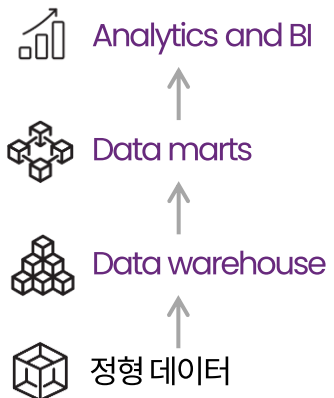
Apache Kafka  
Apache Flink  
Azure Stream Analytics  
Tibco Spotfire

Apache Spark  
Amazon Kinesis  
Google Dataflow  
Confluent

Jupyter  
Azure ML Studio  
Domino Data Labs  
TensorFlow

Amazon SageMaker  
MatLAB  
SAS  
PyTorch

### 사일로화된 기술 스택 때문에 점점 더 복잡해지는 데이터 아키텍처



## 데이터 웨어하우징

## 데이터 엔지니어링

## 스트리밍

## DS & ML

사일로화된 데이터 팀의 생산성 감소, 협업 난이도 상승

### Data Analytics



### Data Engineer



### Data Engineer



### Data Scientist



연결되지 않은 시스템 및 독점적 데이터 포맷은 통합 난이도를 높임

Amazon Redshift  
Azure Synapse  
Snowflake  
SAP

Teradata  
Google BigQuery  
IBM Db2  
Oracle Autonomous Data Warehouse

Hadoop  
Amazon EMR  
Google Dataproc

Apache Airflow  
Apache Spark  
Cloudera

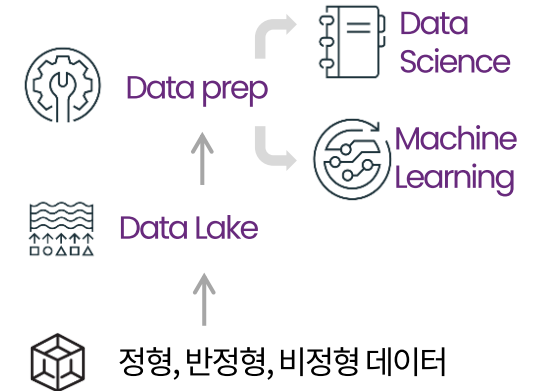
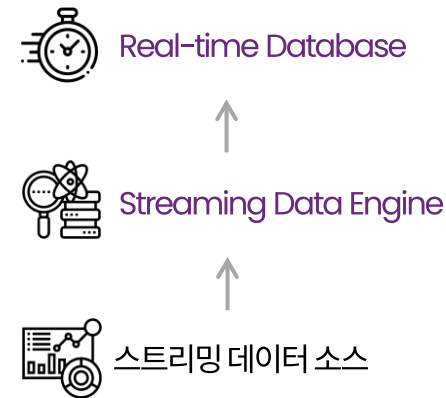
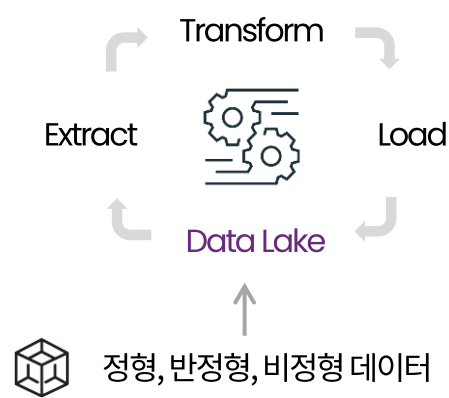
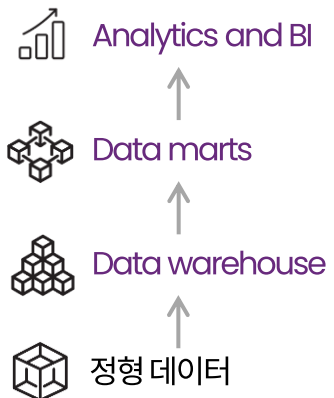
Apache Kafka  
Apache Flink  
Azure Stream Analytics  
Tibco Spotfire

Apache Spark  
Amazon Kinesis  
Google Dataflow  
Confluent

Jupyter  
Azure ML Studio  
Domino Data Labs  
TensorFlow

Amazon SageMaker  
MatLAB  
SAS  
PyTorch

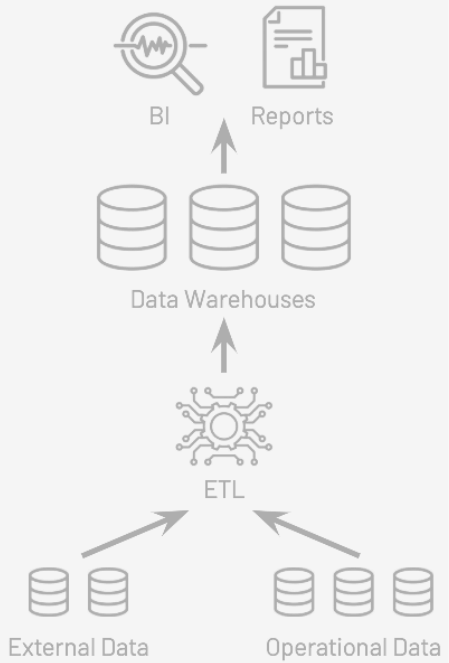
사일로화된 기술 스택 때문에 점점 더 복잡해지는 데이터 아키텍처



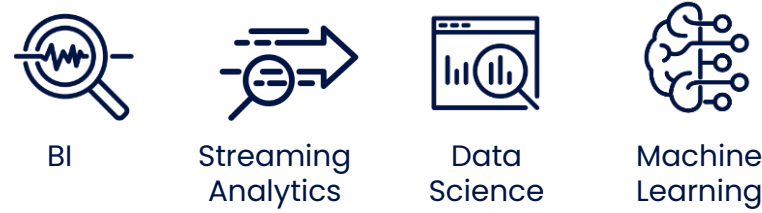


## Data LakeHouse

### Data Warehouse

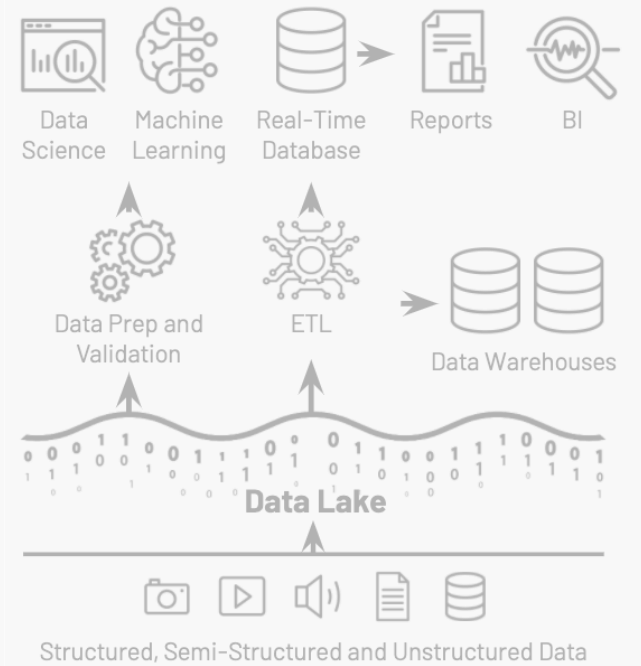


성능, 관리성, 편의성



Structured, Semi-Structured and Unstructured Data

### Data Lake



확장성, 낮은 비용, 개방성

Delta Live Tables understands and coordinates data flow between your queries



SQL, Python, R, Scala 등 Spark Engine 기반의 언어를 사용하여 ETL PIPELINE을 생성 및 관리

### Delta Live Tables Pipeline

```
CREATE LIVE TABLE raw_data as SELECT * FROM cloud_files("/data","json")
```

```
CREATE LIVE TABLE clean_data as SELECT ... FROM LIVE.raw_data
```

```
@dlt.table  
def score_records():  
    return read("clean_data").map(lambda d: model_score(d))
```



### Delta Live Tables SQL Pipeline

2/25/2022, 9:25:02 AM - Completed

**Data Quality**

77.9% (61,752,707) Written  
22.1% (17,506,422) Dropped

Name	Action	Fail %	Failed Records
valid_trip_distance	DROP	22.1%	17484277
valid_passenger_count	DROP	0.2%	176524

Expectations: All, Failures Only

- 4 minutes ago - flow\_progress - Flow 'tbi\_gold\_taxi\_for\_analysis' has COMPLETED.
- 4 minutes ago - flow\_progress - Flow 'tbi\_silver\_taxi\_payments' has COMPLETED.
- 4 minutes ago - flow\_progress - Flow 'tbi\_silver\_taxi\_rates' has COMPLETED.
- 4 minutes ago - flow\_progress - Flow 'tbi\_gold\_union\_taxi' has COMPLETED.
- 4 minutes ago - update\_progress - Update 1c1ad5 is COMPLETED.

각 Flow 별 데이터 의존성을 확립할 수 있음 (ex: Apache Airflow)



# ■ Demo

- Delta Live Table Demo



The background features a dark blue and purple color scheme with a grid pattern. A large, stylized 'AI' is centered in the middle. Surrounding it are various icons: a brain, a gear, a network diagram, and a molecular structure. A thick purple arrow points from the top right towards the 'AI' text.

**Thank you!**

**Cloócus**