BOAZ' ENGINEERING 22년 학기 세션

데이터 파이프라인 (Data Pipeline)

19기 송우석

CONTENTS

Data Pipeline이란?

효율을 위한 작업

시작

데이터 파이프라인의 시작은 왜, 어디에서, 어떻게 데이터를 수집할 것인가에서 부터 시작한다.

목표

데이터의 흐름에서 실패 지점을 없애고 장애를 최소화

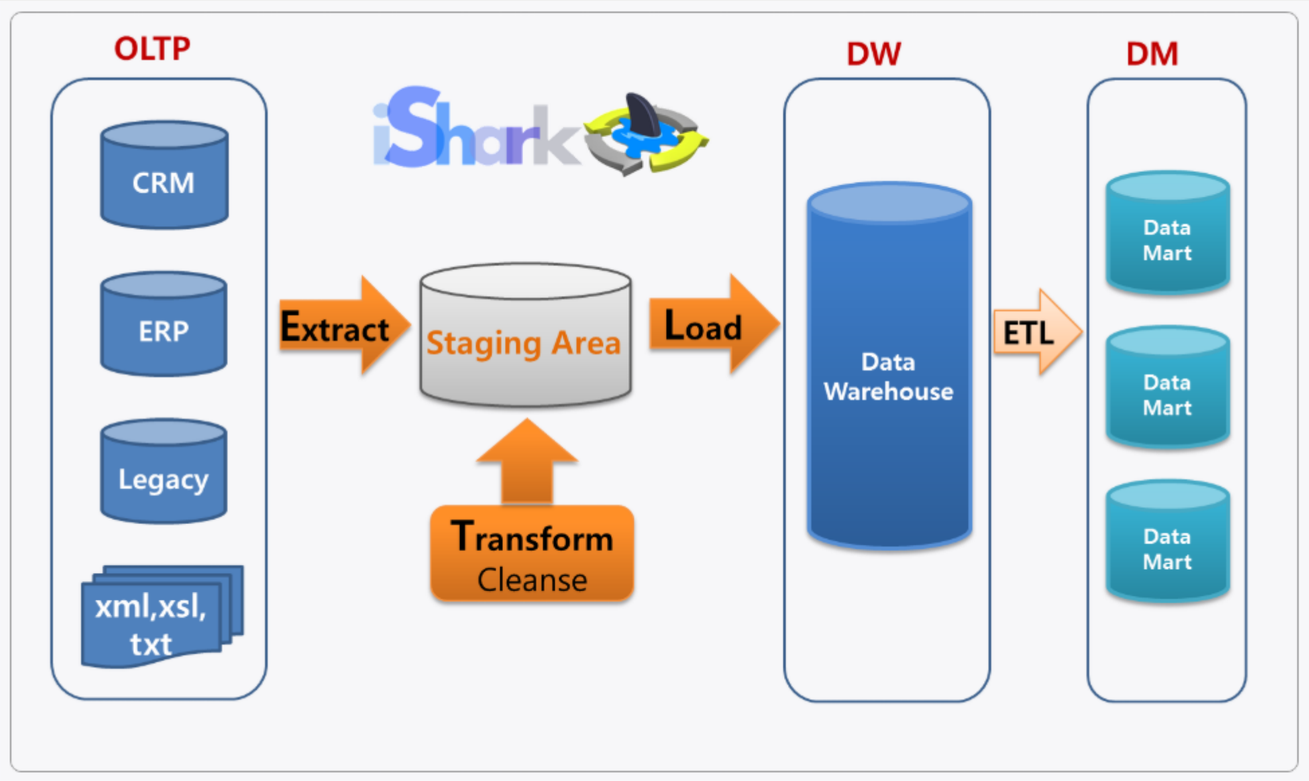
DATA

Data

서비스 데이터

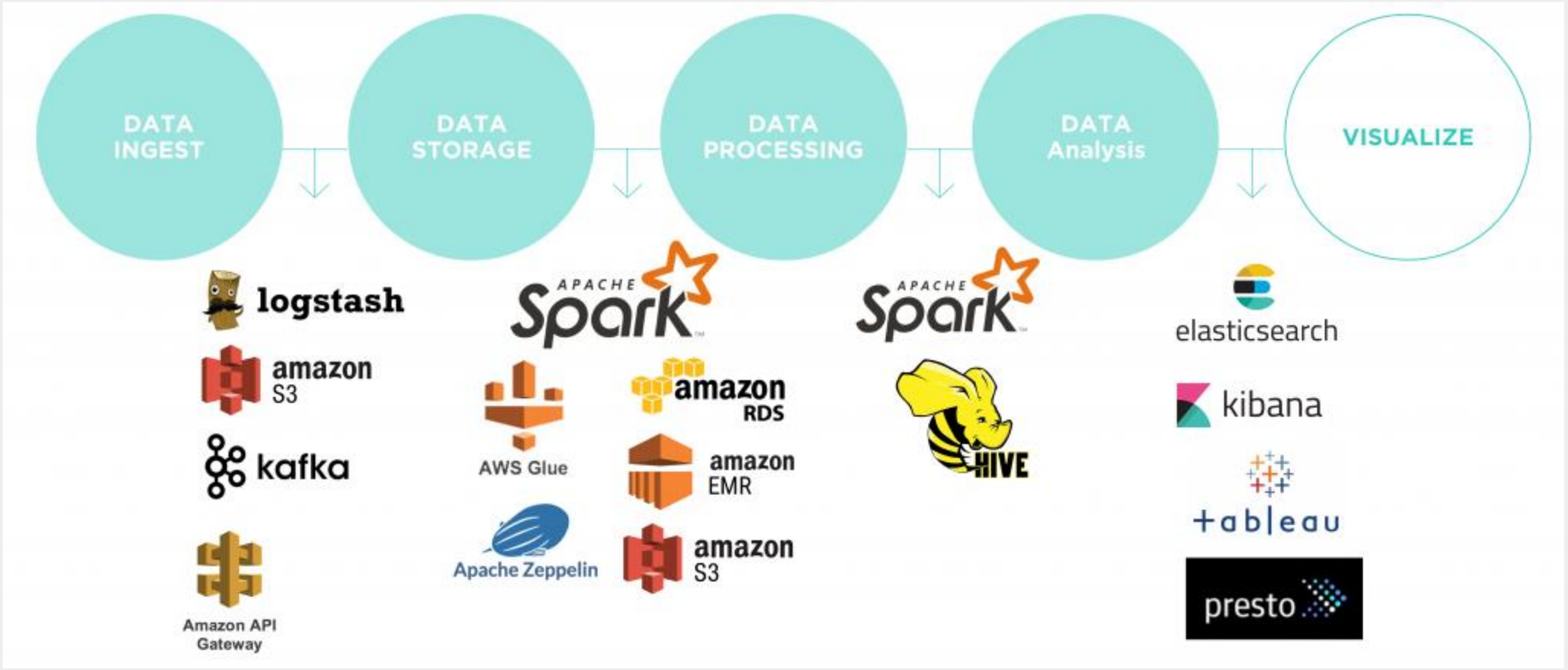
로그 데이터

ARCHITECTURE

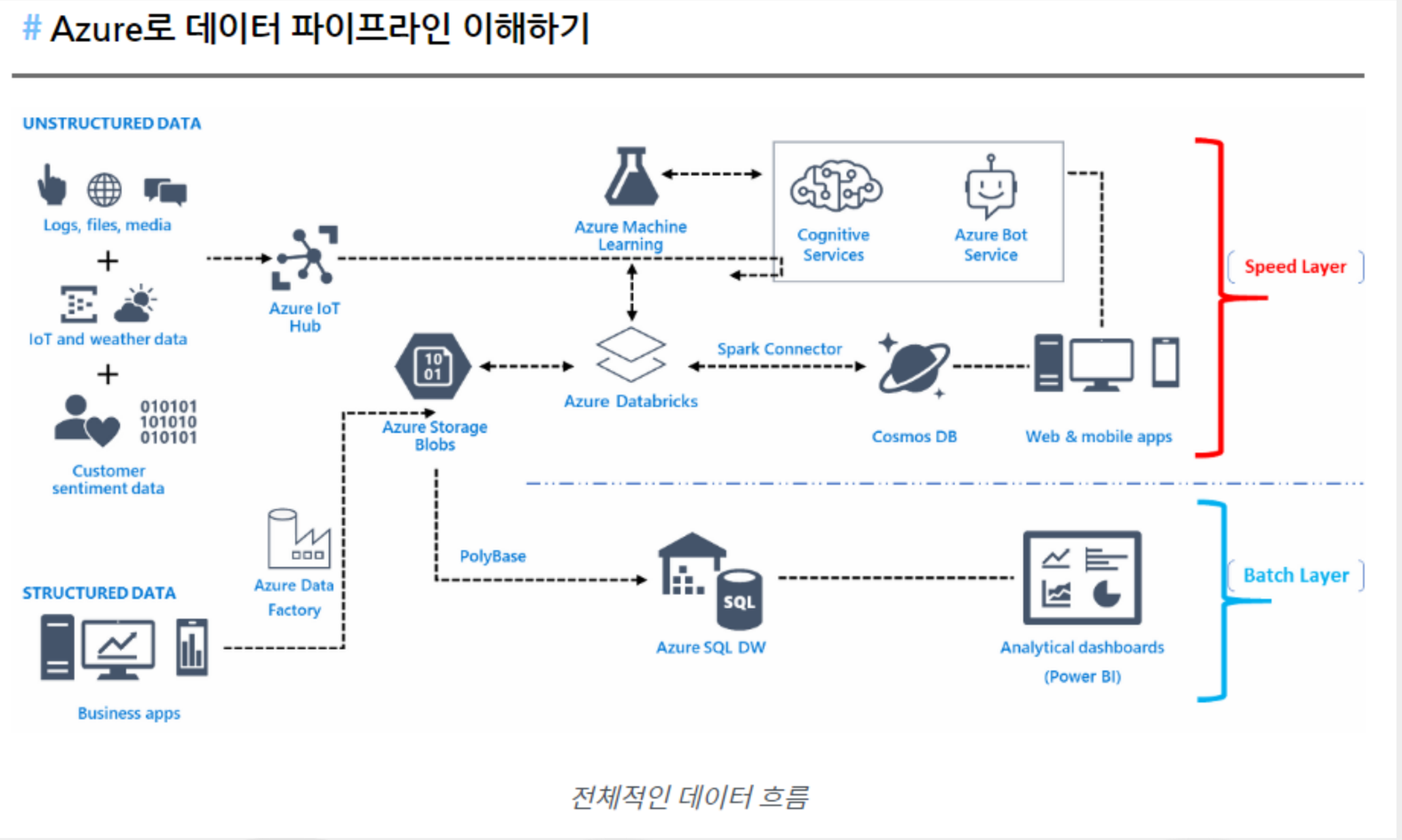


Architecture
데이터 생성
데이터 수집
데이터 가공 후 저장(ETL)
데이터 시각화

각 단계별 서비스


DATA INGEST
DATA STORAGE
DATA PROCESSING
DATA Analysis
VISUALIZE
logstash
amazon S3
kafka
Amazon API Gateway
APACHE Spark
AWS Glue
amazon RDS
amazon EMR
Apache Zeppelin
amazon S3
APACHE Spark
HIVE
elasticsearch
kibana
+ableau
presto

예시 - AZURE

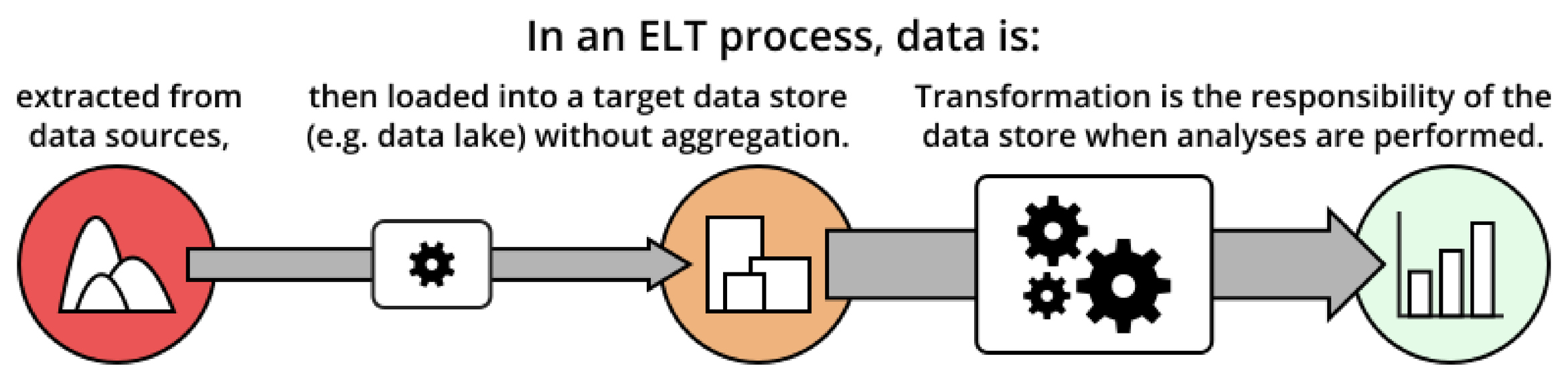


전체적인 데이터 흐름

ETL

추출(Extract)	변환(Transform)	적재(Load)
원본 데이터베이스 또는 데이터 소스에서 소스 데이터를 가져오는 것	대상 데이터 시스템 및 해당 시스템의 나머지 데이터와 통합할 수 있도록 정보의 구조를 변경하는 과정	정보를 데이터 스토리지 시스템에 보관하는 과정

ETL vs ELT



ETL 장점

1. 효율적이고 안정적
2. 비용이 휴율적
3. 정보 보호 규정 준수에 적합

ELT 장점

1. 빠른 속도
2. 유지 관리의 번거로움 감소
3. 신속한 로드

Extract

의미

추출은 데이터를 분석, AI/ML 등에 유용하게 만들어 주는 첫단계.

방식

1. 변경 알림 기반 데이터 추출
2. 증분 데이터 추출
3. 전체 데이터 추출

중요성

1. 의사 결정시 정보에 입각
2. 가치가 높은 활동에 집중
3. 오류 최소화
4. 생산성 향상 .

추출 유형

1. 구조화되지 않은 데이터
2. 구조화된 데이터

운영데이터
고객데이터
재무데이터

Transform

의미

여러 데이터 소스로부터 추출한 raw data를 분석을 위한 형태로 가공하는 과정

예시

1. 구조화되지 않은 데이터 구조화
2. 데이터 필터링
3. 데이터 유효성 검사
4. 중복 레코드 제거

장점

1. 데이터 가치 향상
2. 데이터 품질 향상
3. 데이터 조직 및 관리 개선
4. 쿼리 및 데이터 검색 가속화

Load

의미

분석에 알맞은 형태로 데이터 가공이 완료된 뒤 데이터 웨어하우스(DW)에 적재하는 과정

데이터 레이크

조직에서 수집한 정형·반정형·비정형 데이터를 원시 형태(raw data)로 저장하는 단일한 데이터 저장소

데이터 웨어하우스

조직 전체의 여러 소스로부터 데이터를 저장하고 처리하여 중요한 비즈니스 분석, 보고서 및 대시보드에 사용할 수 있는 의사결정 지원 시스템

Data Lake와 Data Warehouse

Data Lake 특징

1. 데이터를 저장하기 전에는 이를 정제하지 않아도 됨
2. 정형·반정형·비정형 데이터를 모두 저장 가능
3. 미리 정의된 목적이 없는 데이터를 저장
4. 즉시 데이터를 수집 가능
5. 환경설정이 유연

Data Lake 한계

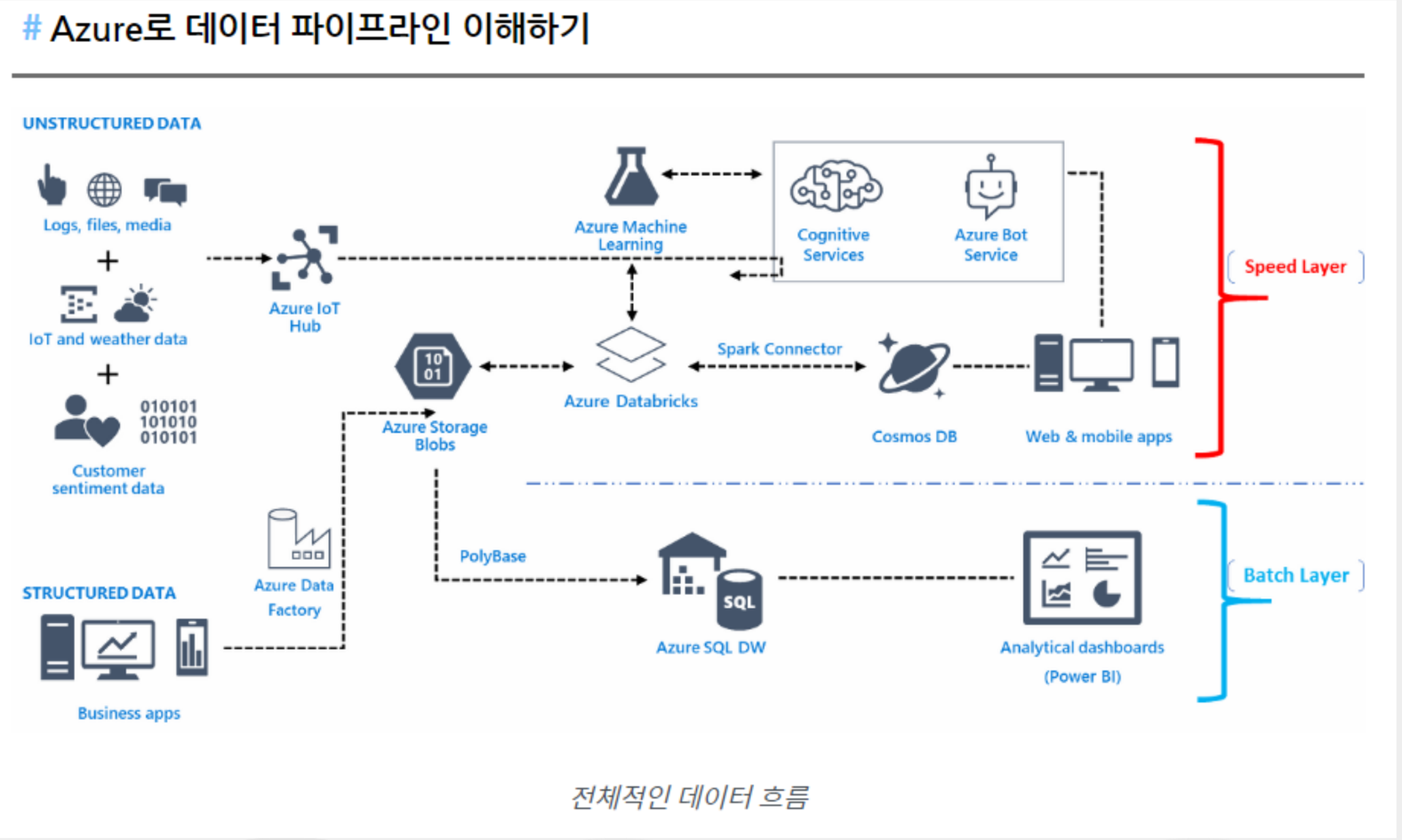
1. 데이터 늪이 될 수 있음
2. 보안과 액세스 제어 문제
3. 성능 저하의 가능성

결론

Data Lake는 Data warehous의 보완재이며 대체재는 아님

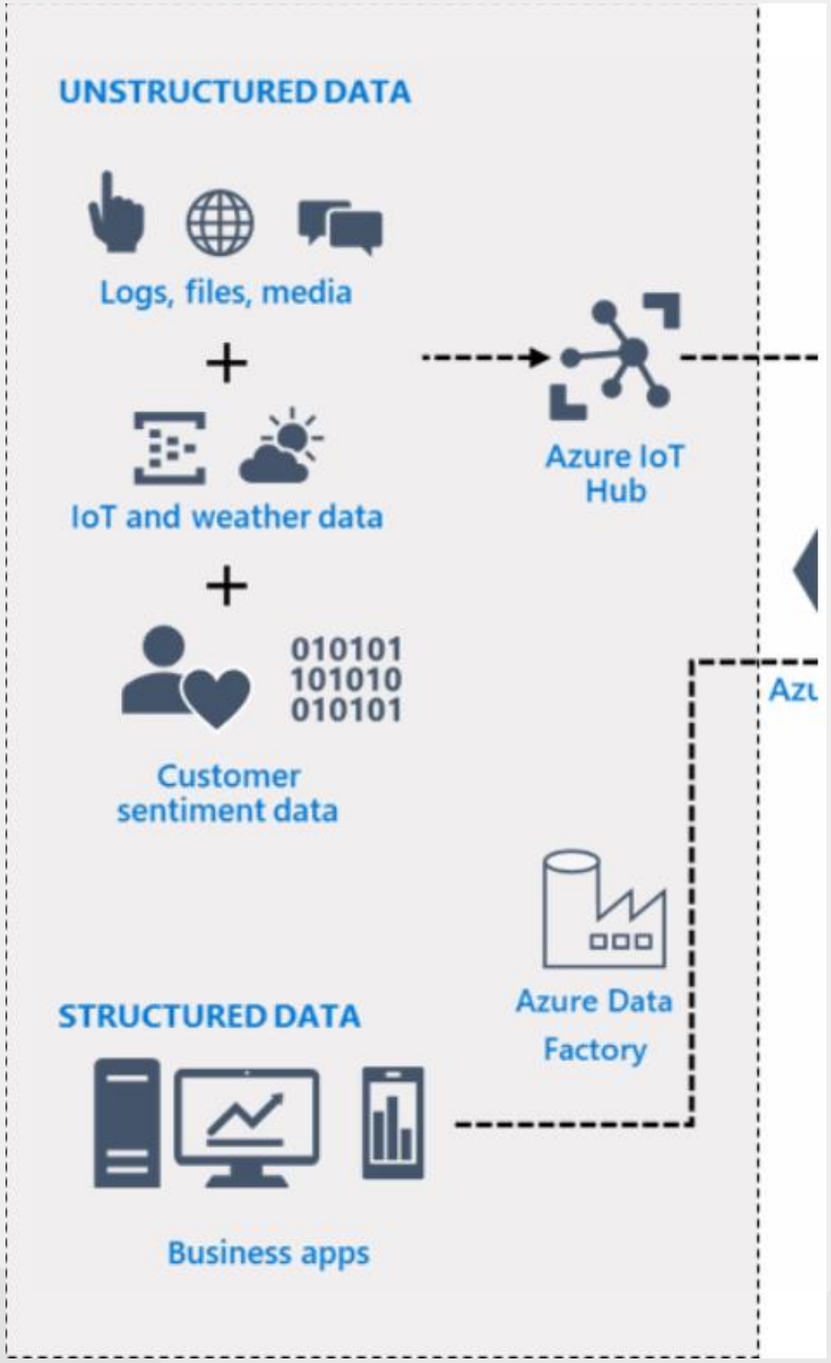
=> 각 사 필요에 맞는 걸 골라 사용하는 게 좋다!

예시 - Azure

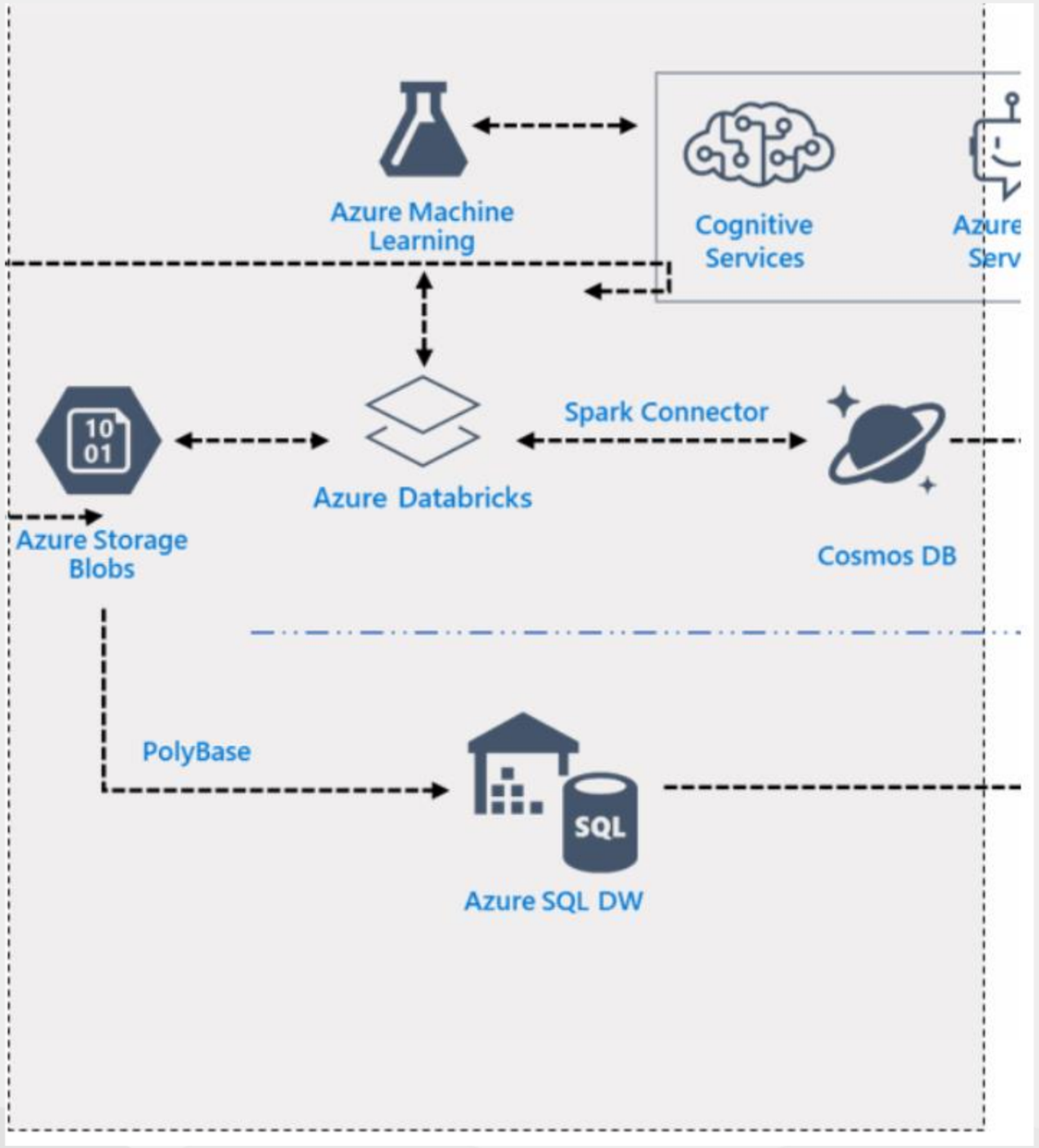


전체적인 데이터 흐름

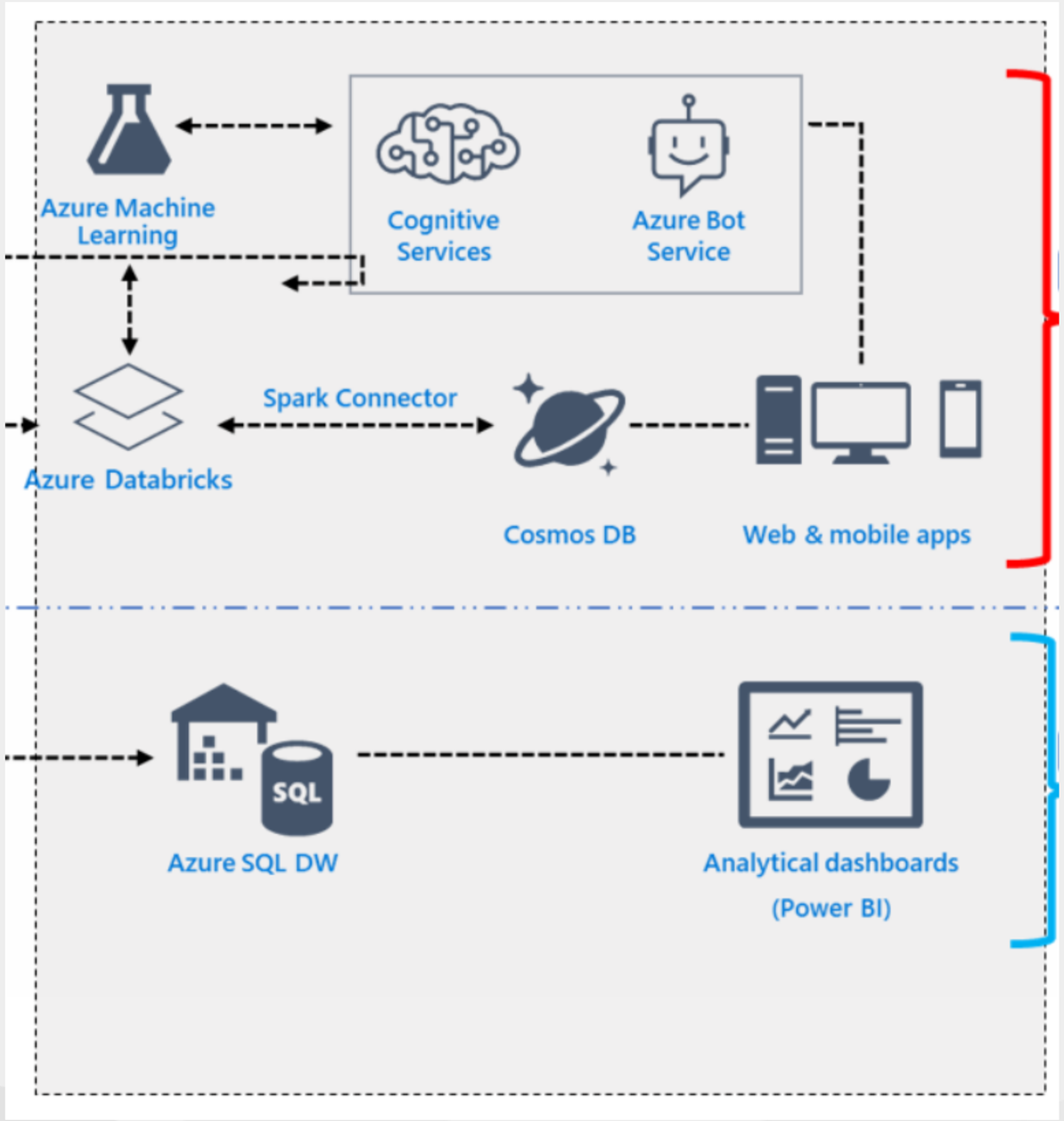
데이터 수집


UNSTRUCTURED DATA
Logs, files, media
+
IoT and weather data
+
010101
101010
010101
Customer
sentiment data
Azure IoT
Hub
Azu
Azure Data
Factory
STRUCTURED DATA
Business apps

데이터 저장 및 처리


Azure Machine Learning
Cognitive Services
Azure Serv
10 01
Azure Databricks
Spark Connector
Azure Storage Blobs
Cosmos DB
PolyBase
SQL
Azure SQL DW

데이터 저장 및 처리


Azure Machine Learning
Cognitive Services
Azure Bot Service
Spark Connector
Azure Databricks
Cosmos DB
Web & mobile apps
SQL
Azure SQL DW
Analytical dashboards
(Power BI)

감사합니다