광고 데이터 처리 시스템 소개

박성훈 _ k.ey
카카오

01 Ad platforms & data

02 Blocks

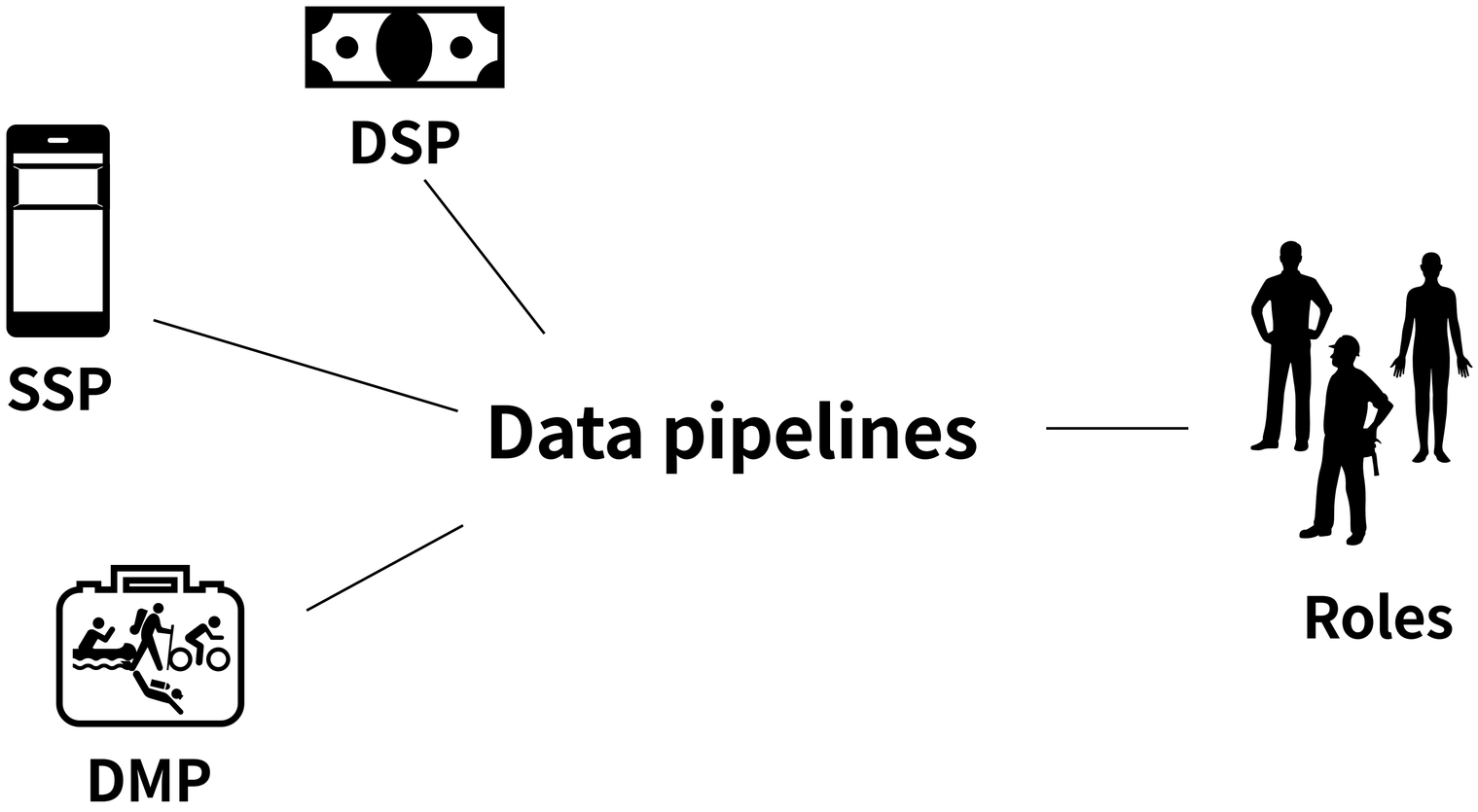
03 Data pipelines

01 Ad platforms & data

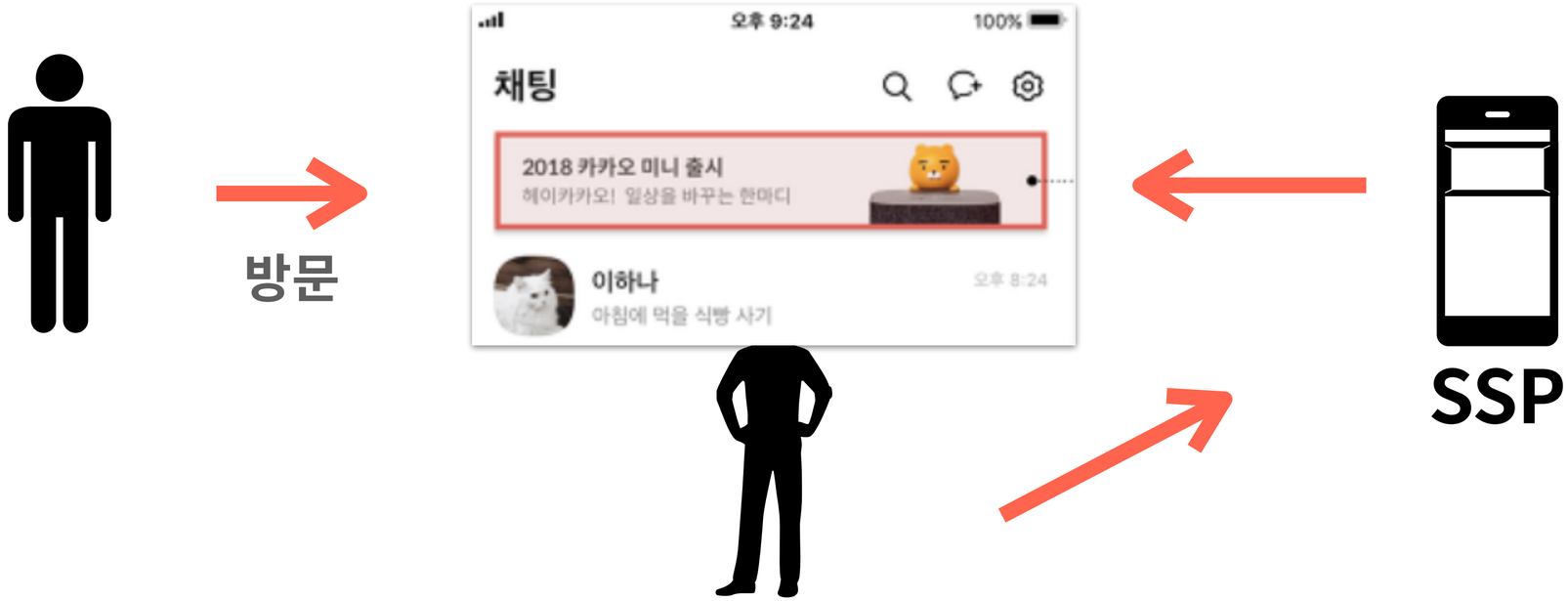
Overview

광고 + 데이터

Overview



Supply Side Platform



Demand Side Platform

Data Management Platform

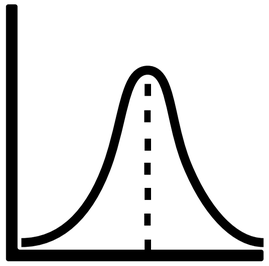
Roles

데이터 분석가

영업

기획

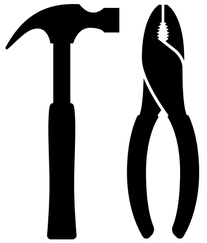
의사결정



데이터 과학자

노출모델개발

추천모델개발

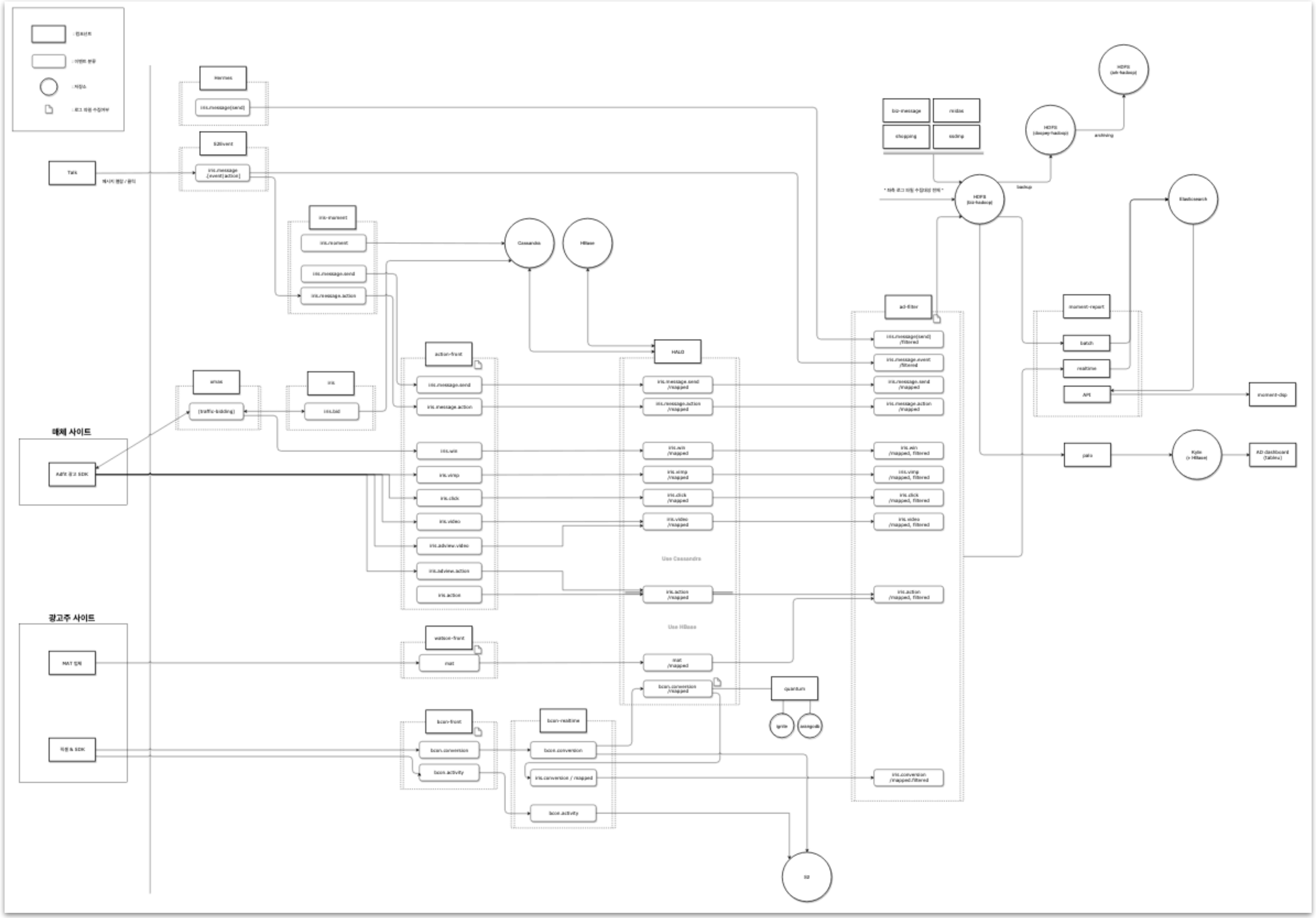


데이터 엔지니어

개발

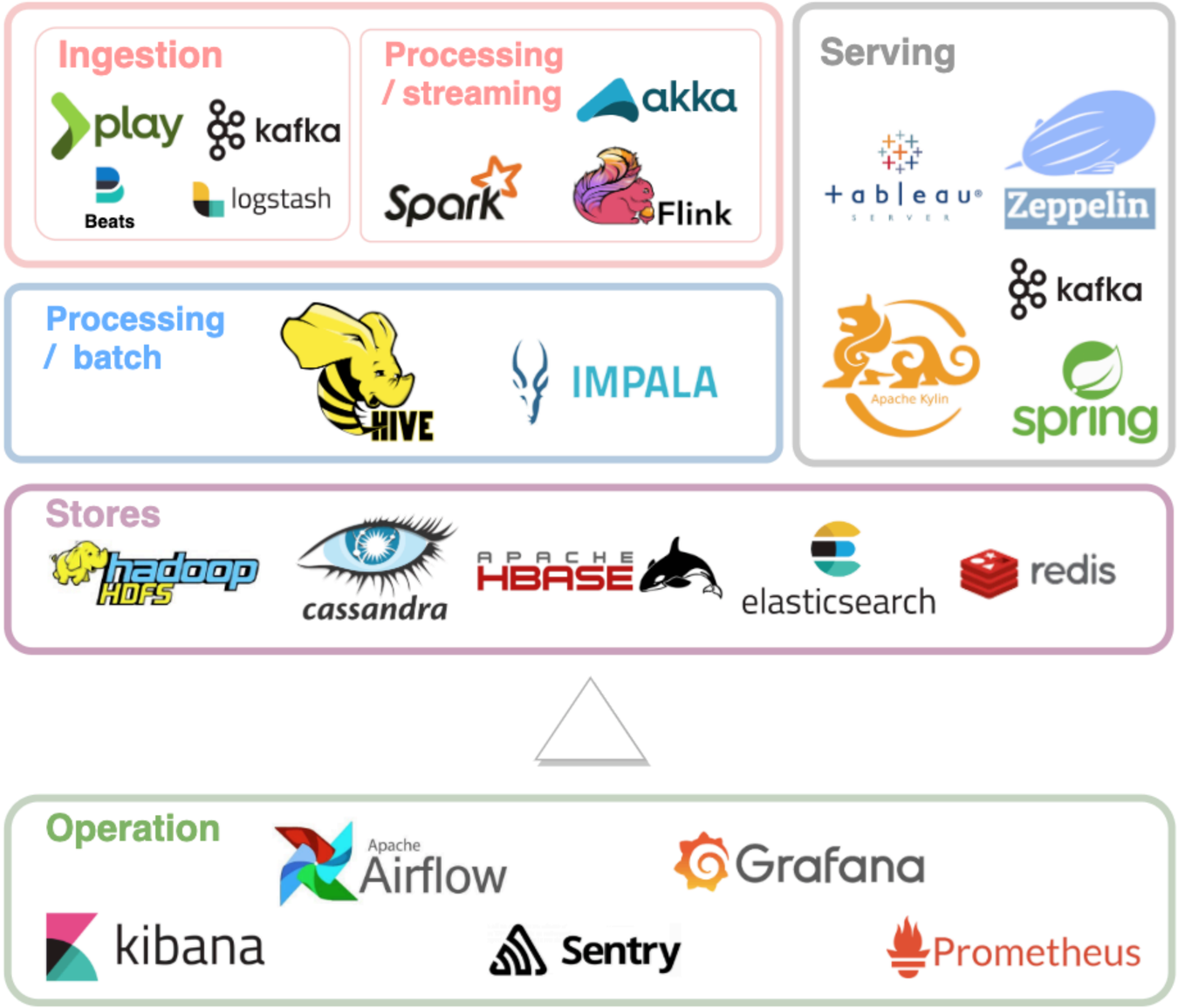
02 Blocks

Overview



매체 사이트
광고주 사이트
Talk
Hermes
action-front
HALO
Cassandra
HBase
ad-filter
HDFS
Elasticsearch
moment-report
batch
realtime
API
moment-dsp
Kylin
AD dashboard (tableu)
quantum
ignite
mongodb
Use Cassandra
Use HBase
iris.message.send
iris.message.action
iris.win
iris.vimp
iris.click
iris.video
iris.adview.video
iris.adview.action
iris.action
bcon-front
bcon-realtime
bcon.conversion
bcon.activity
iris.conversion / mapped

Blocks


Ingestion
play
kafka
Beats
logstash
Processing / streaming
akka
Spark
Flink
Serving
+ableau SERVER
Zeppelin
kafka
Apache Kylin
spring
Processing / batch
HIVE
IMPALA
Stores
hadoop HDFS
cassandra
APACHE HBASE
elasticsearch
redis
Operation
Apache Airflow
Grafana
kibana
Sentry
Prometheus

Block ingestion + processing / streaming

Block processing / batch

Processing / batch

Block stores

Block serving

CASE: 데이터를 다각도로 빠르게 분석하면 좋겠어요

- 요구사항
  - 다양한 조합의 빠른 분석을 할 수 있으면
  - 조회하는 방식이 친숙했으면
  - BI 툴과의 연동이 쉬웠으면
  - 새로운 조합의 데이터를 쉽게 만들수 있었으면

CASE: 데이터를 다각도로 빠르게 분석하면 좋겠어요

- 요구사항
  - 다양한 조합의 빠른 분석을 할 수 있으면 👉 OLAP Cube
  - 조회하는 방식이 친숙했으면 👉 SQL
  - BI 툴과의 연동이 쉬웠으면 👉 JDBC/ODBC
  - 새로운 조합의 데이터를 쉽게 만들수 있었으면 👉 Cube wizard

CASE: 데이터를 다각도로 빠르게 분석하면 좋겠어요

- 요구사항
  - 다양한 조합의 빠른 분석을 할 수 있으면 👉 OLAP Cube
  - 조회하는 방식이 친숙했으면 👉 SQL
  - BI 툴과의 연동이 쉬웠으면 👉 JDBC/ODBC
  - 새로운 조합의 데이터를 쉽게 만들수 있었으면 👉 Cube wizard

Kylin > overview

eBay / Apache project

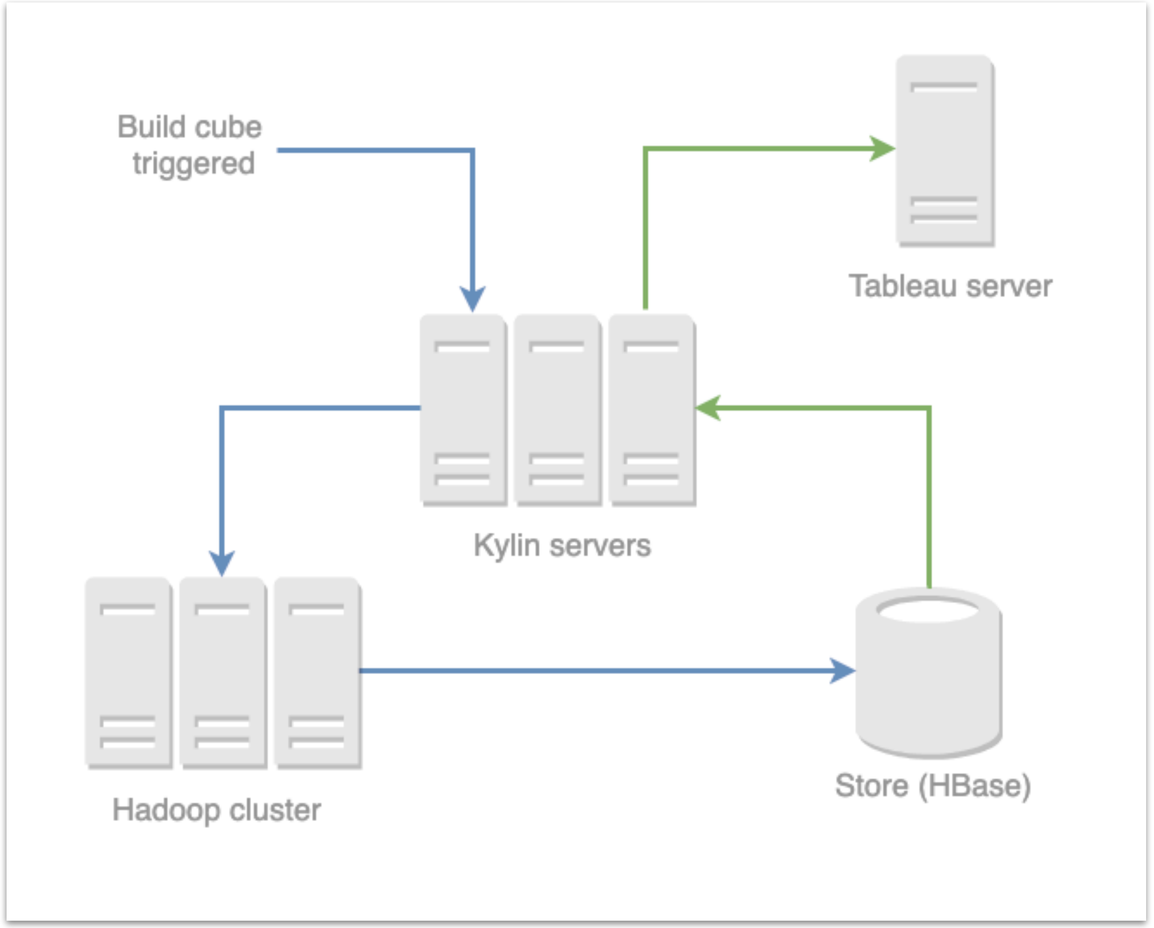
ANSI SQL

MOLAP Cube

Integration with BI Tools

Management Web UI

Kylin > our architecture



Block operation

CASE: 배치 작업의 운영이 쉬웠으면 좋겠어요

- 요구사항
  - 작업간 의존성을 잘 관리할 수 있었으면
  - 재시도/재실행 이 쉬웠으면
  - 필요한 기능을 손쉽게 추가할 수 있었으면
  - 처리 상태를 한눈에 볼 수 있었으면

CASE: 배치 작업의 운영이 쉬웠으면 좋겠어요

- 요구사항
  - 작업간 의존성을 잘 관리할 수 있었으면 👉 Task dependency manage
  - 재시도/재실행 이 쉬웠으면 👉 Task retry/re-run
  - 필요한 기능을 손쉽게 추가할 수 있었으면 👉 Plugin
  - 처리 상태를 한눈에 볼 수 있었으면 👉 Graph/Tree view

CASE: 배치 작업의 운영이 쉬웠으면 좋겠어요

- 요구사항
  - 작업간 의존성을 잘 관리할 수 있었으면 👉 Task dependency manage
  - 재시도/재실행 이 쉬웠으면 👉 Task retry/re-run
  - 필요한 기능을 손쉽게 추가할 수 있었으면 👉 Plugin
  - 처리 상태를 한눈에 볼 수 있었으면 👉 Graph/Tree view

Airflow > overview

Airbnb / Apache incubator project

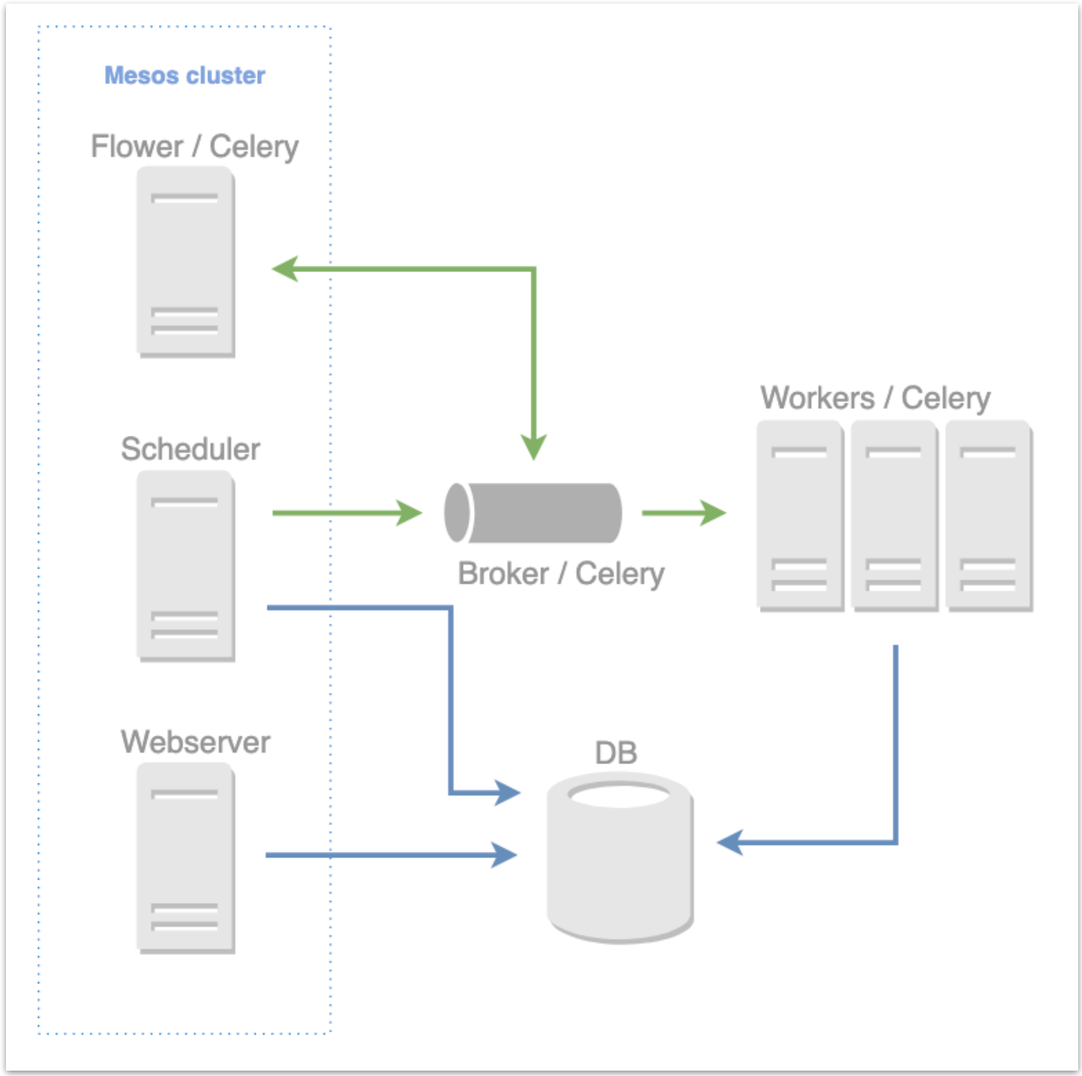
Workflow management tool

Define DAG with python code

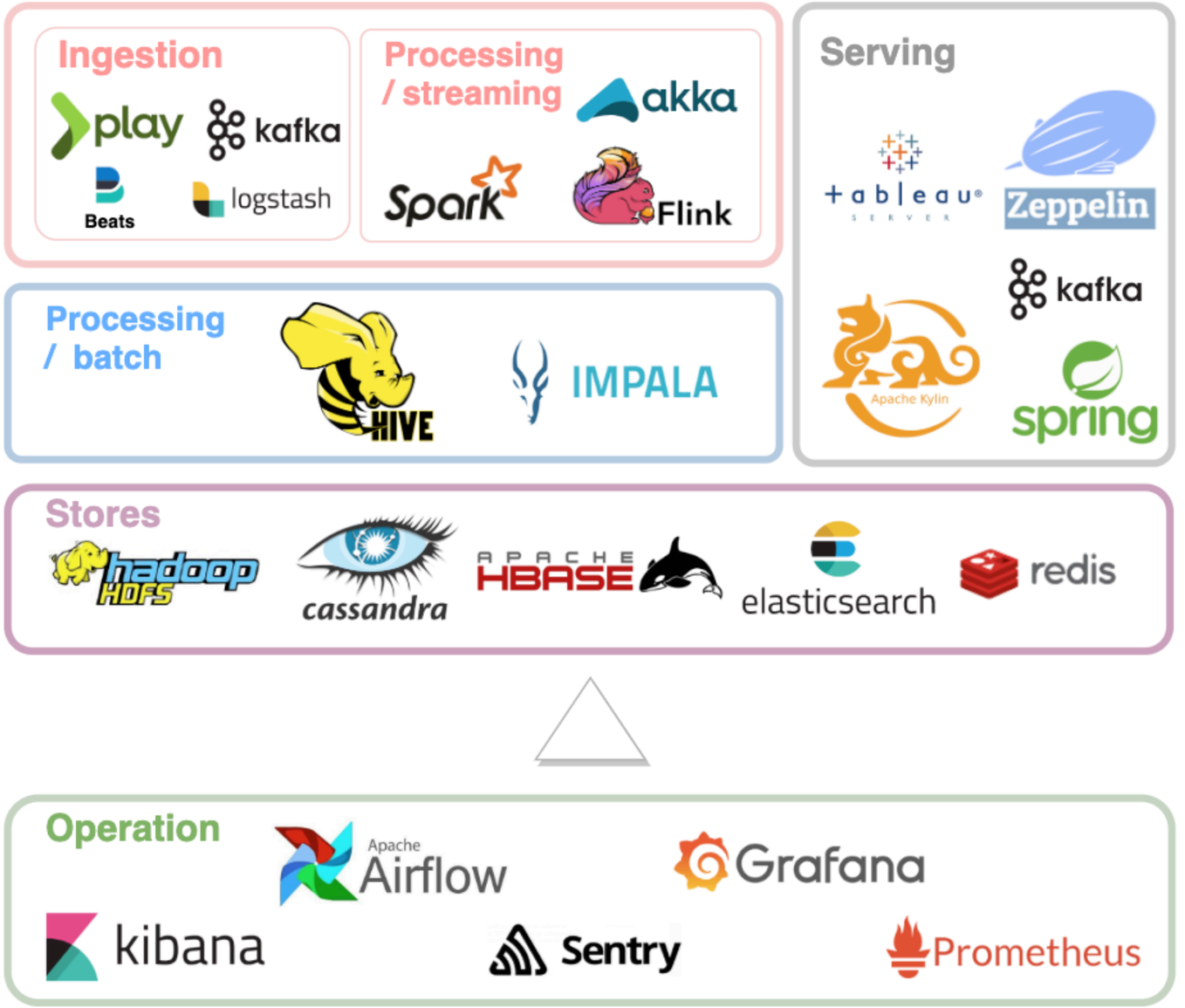
Distributed task run

Management Web UI

Airflow > our architecture



Blocks


Ingestion
play
kafka
Beats
logstash
Processing / streaming
akka
Spark
Flink
Serving
+ableau SERVER
Zeppelin
kafka
Apache Kylin
spring
Processing / batch
HIVE
IMPALA
Stores
hadoop HDFS
cassandra
APACHE HBASE
elasticsearch
redis
Operation
Apache Airflow
Grafana
kibana
Sentry
Prometheus

03 Data pipelines

Data pipelines

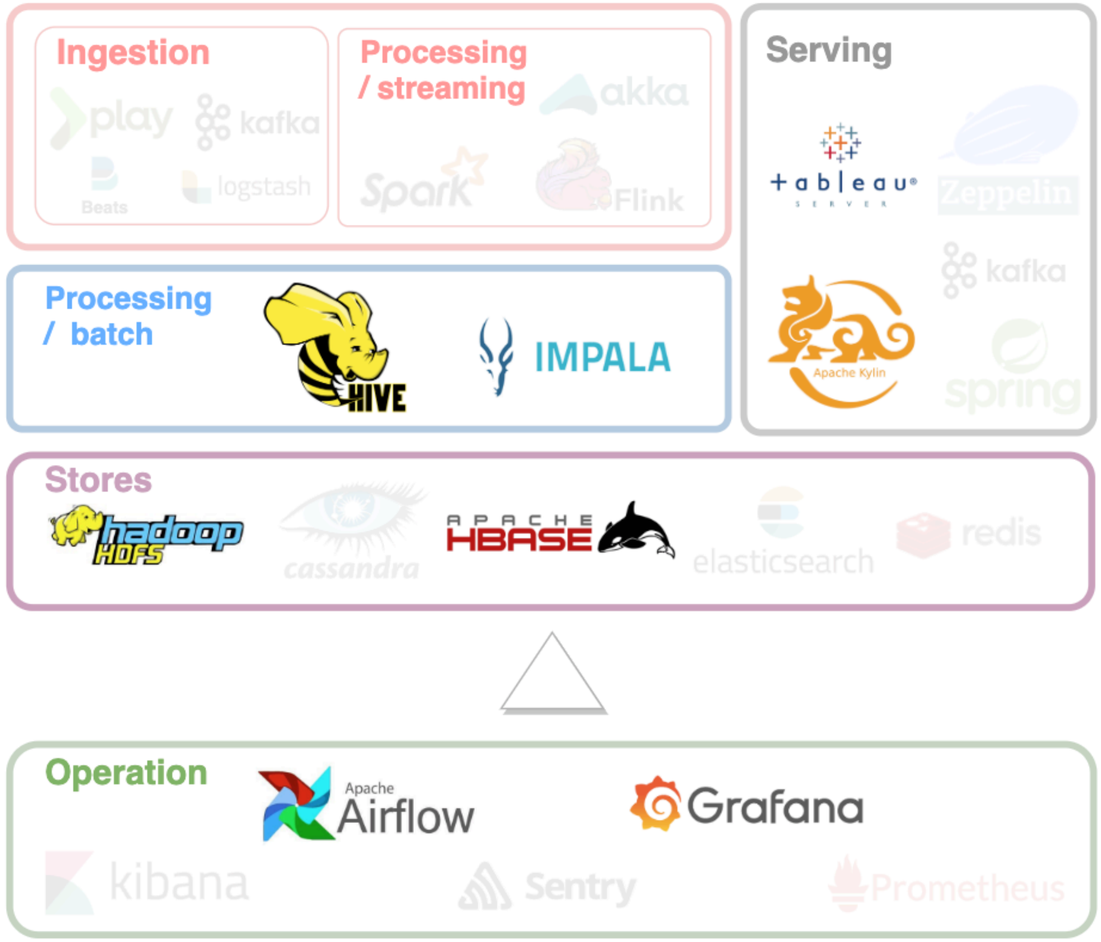
Data pipeline OLAP 데이터

Data pipeline 이벤트 연결

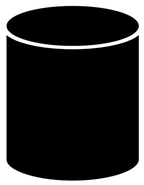
Data pipeline 보고서

...

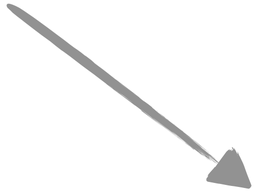
Data pipeline OLAP 데이터

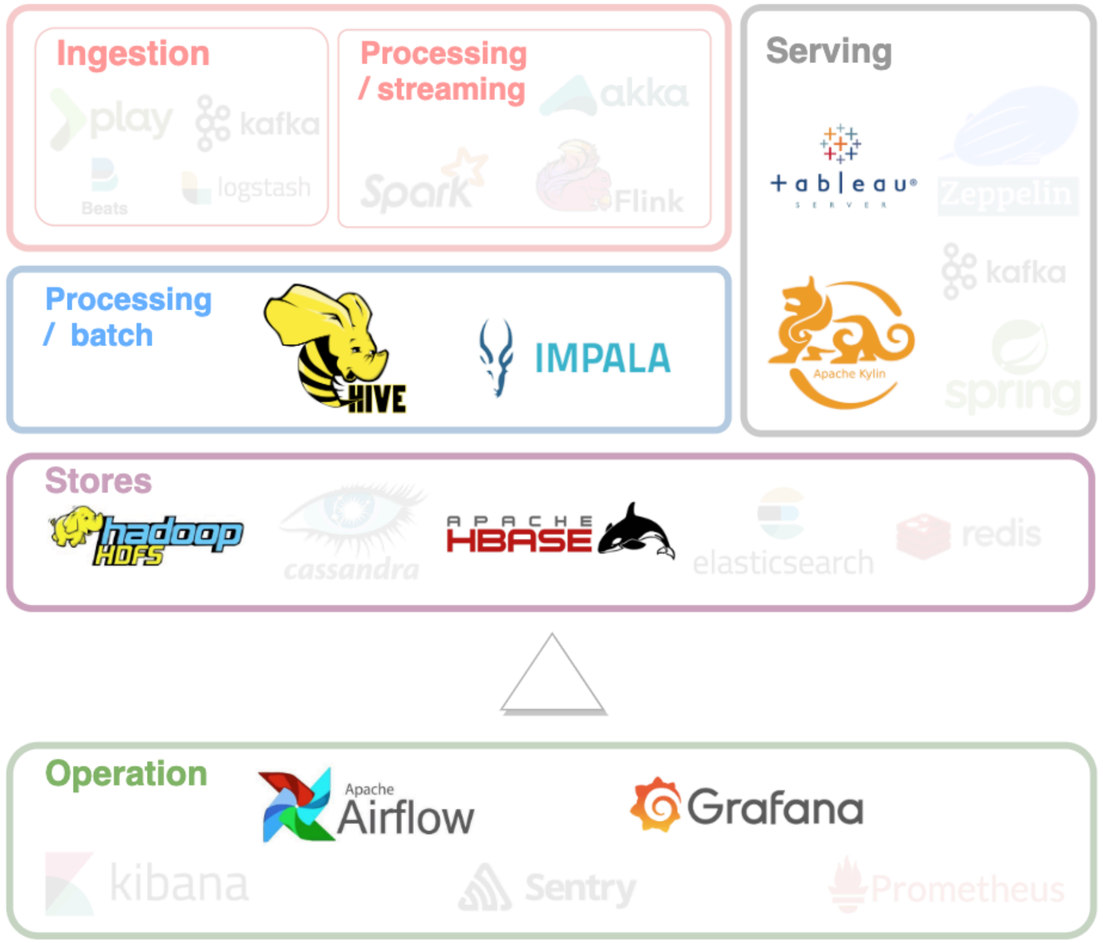


Data pipeline OLAP 데이터

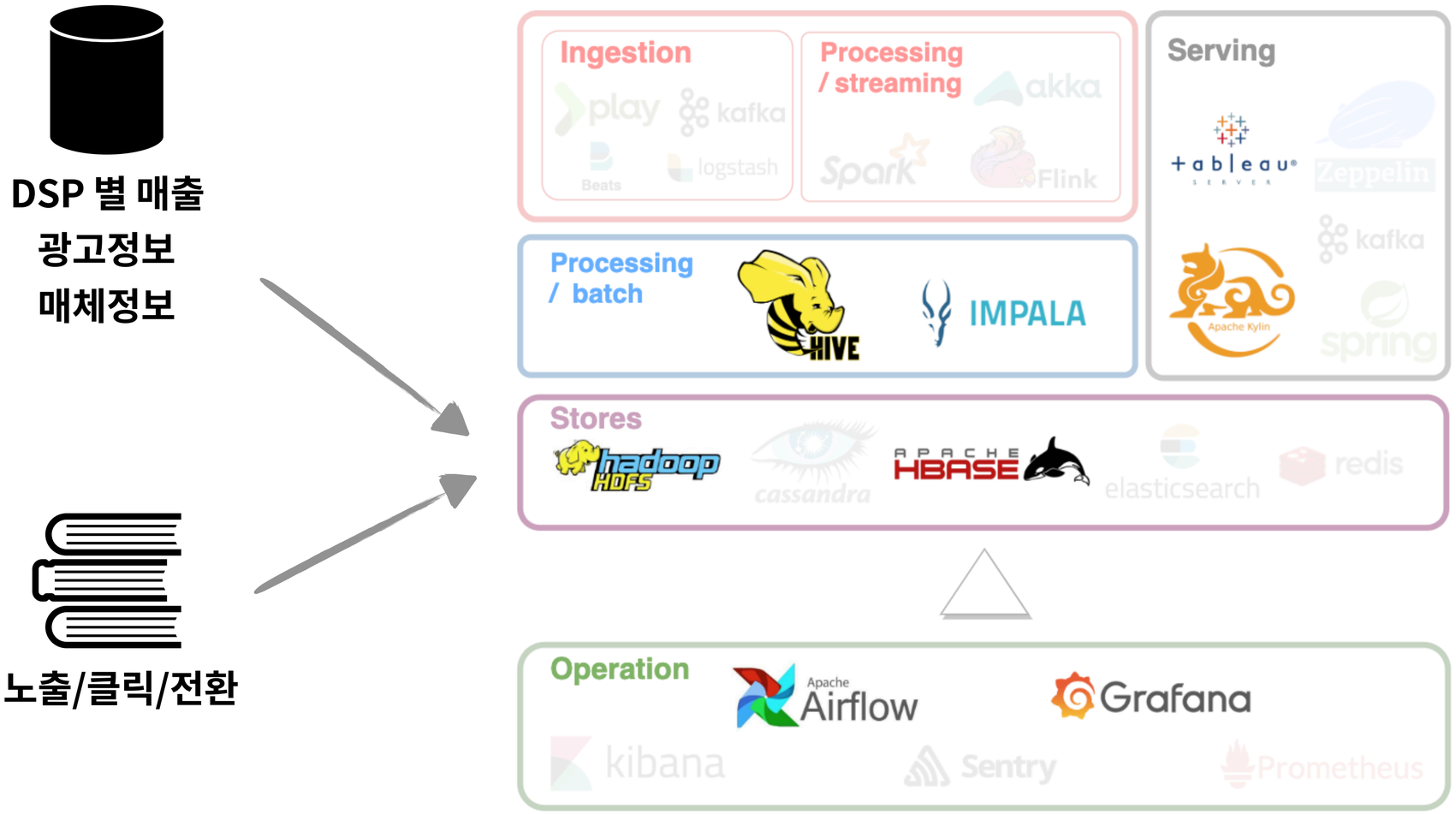


DSP 별 매출
광고정보
매체정보

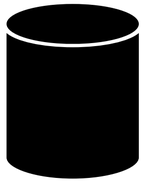




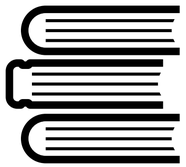
Data pipeline OLAP 데이터



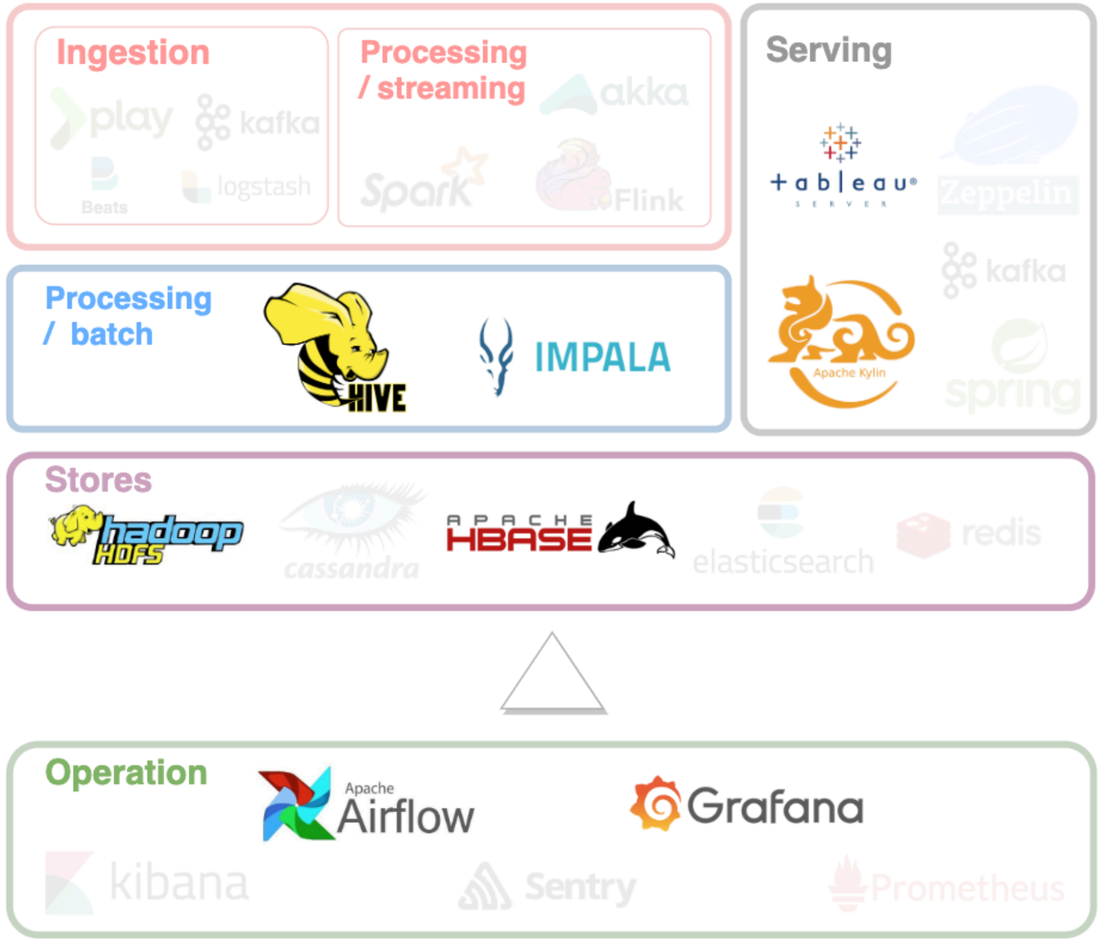
Data pipeline OLAP 데이터



DSP 별 매출
광고정보
매체정보

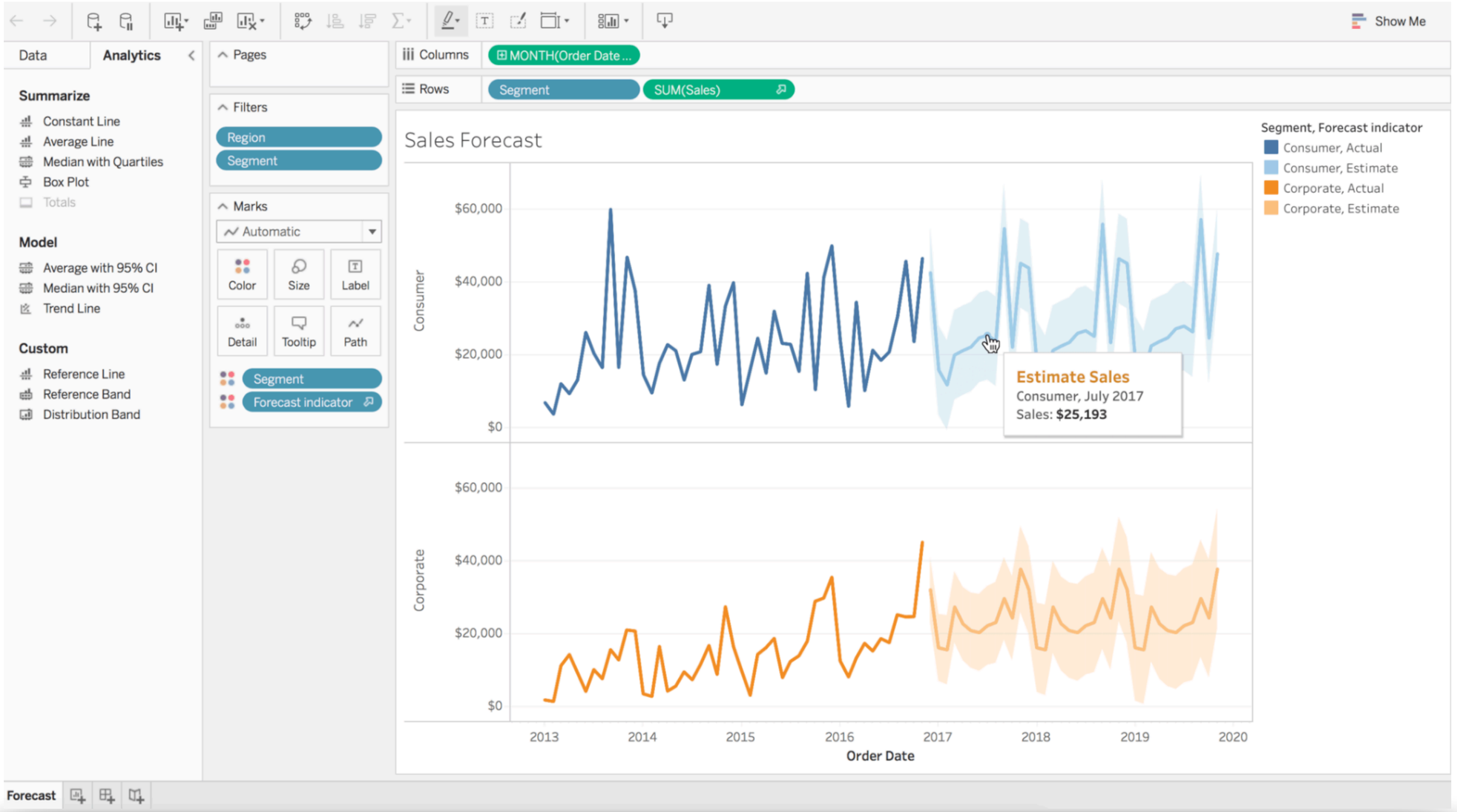


노출/클릭/전환



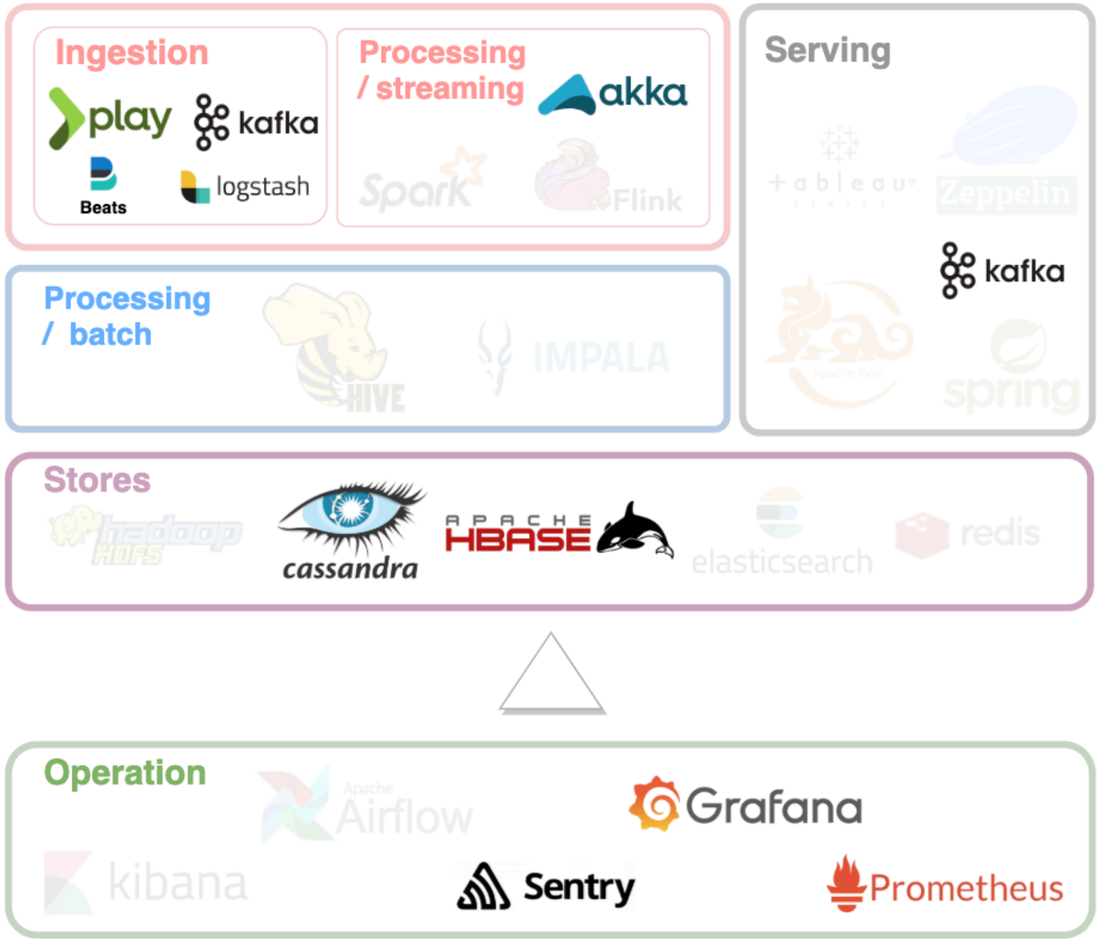
데이터 분석가

Data pipeline OLAP 데이터

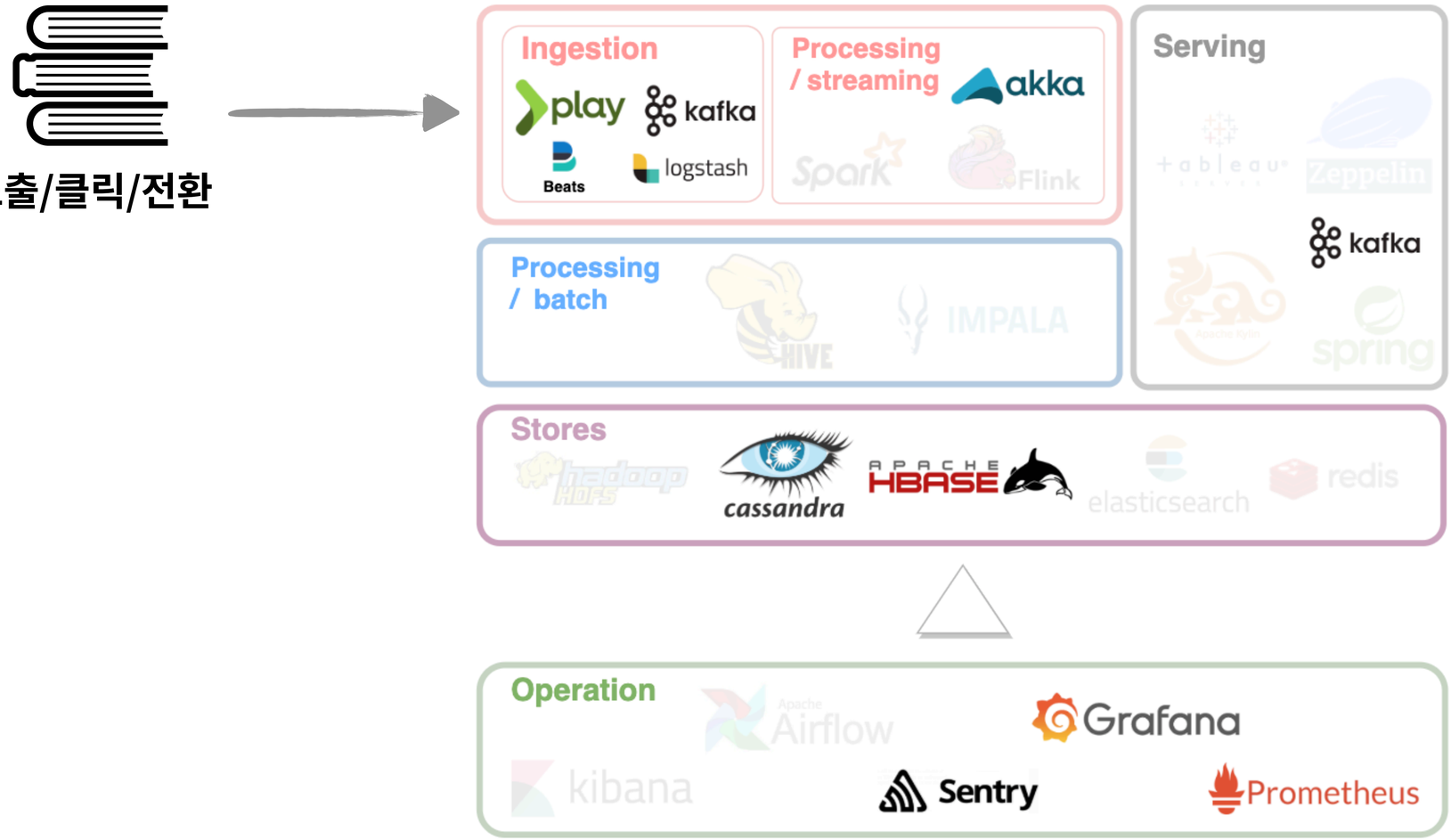


출처: https://www.tableau.com/ko-kr/products/server

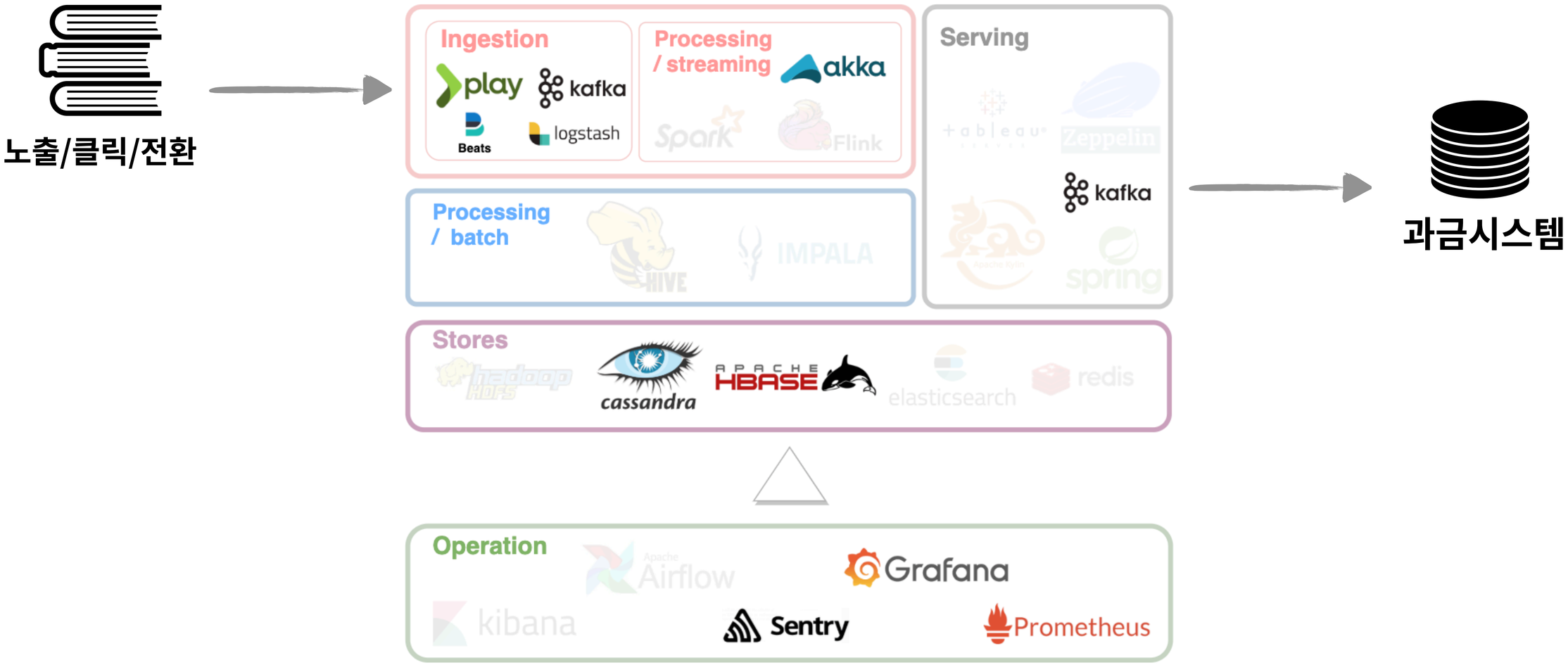
Data pipeline 이벤트 연결



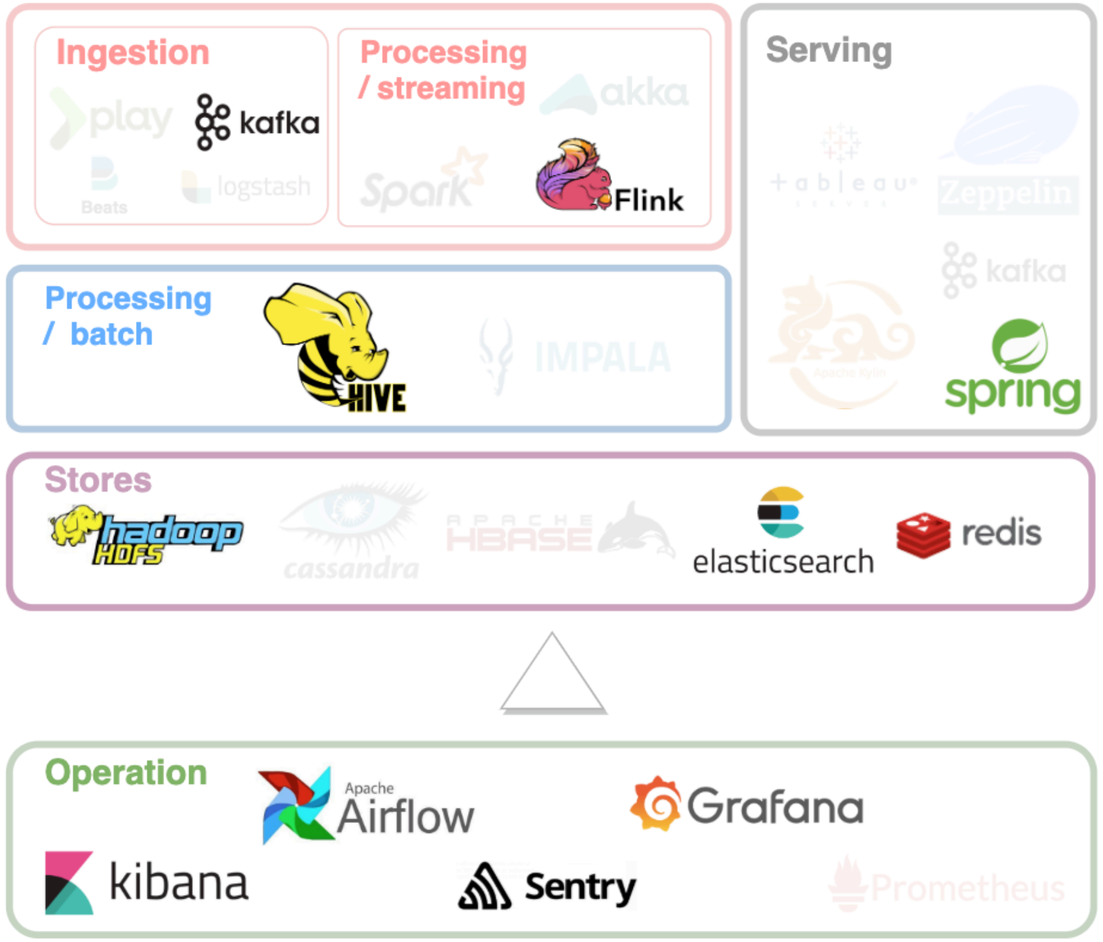
Data pipeline 이벤트 연결



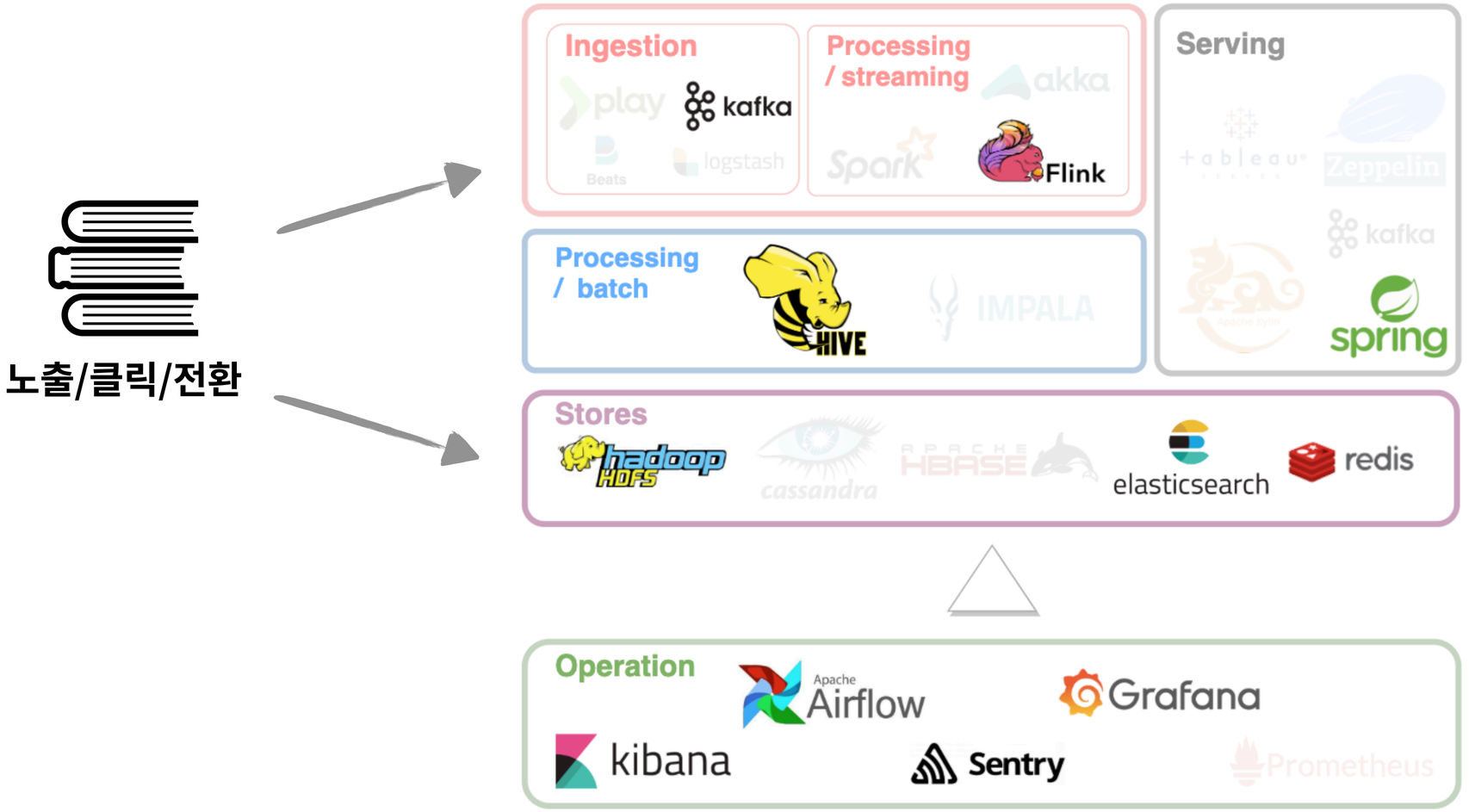
Data pipeline 이벤트 연결



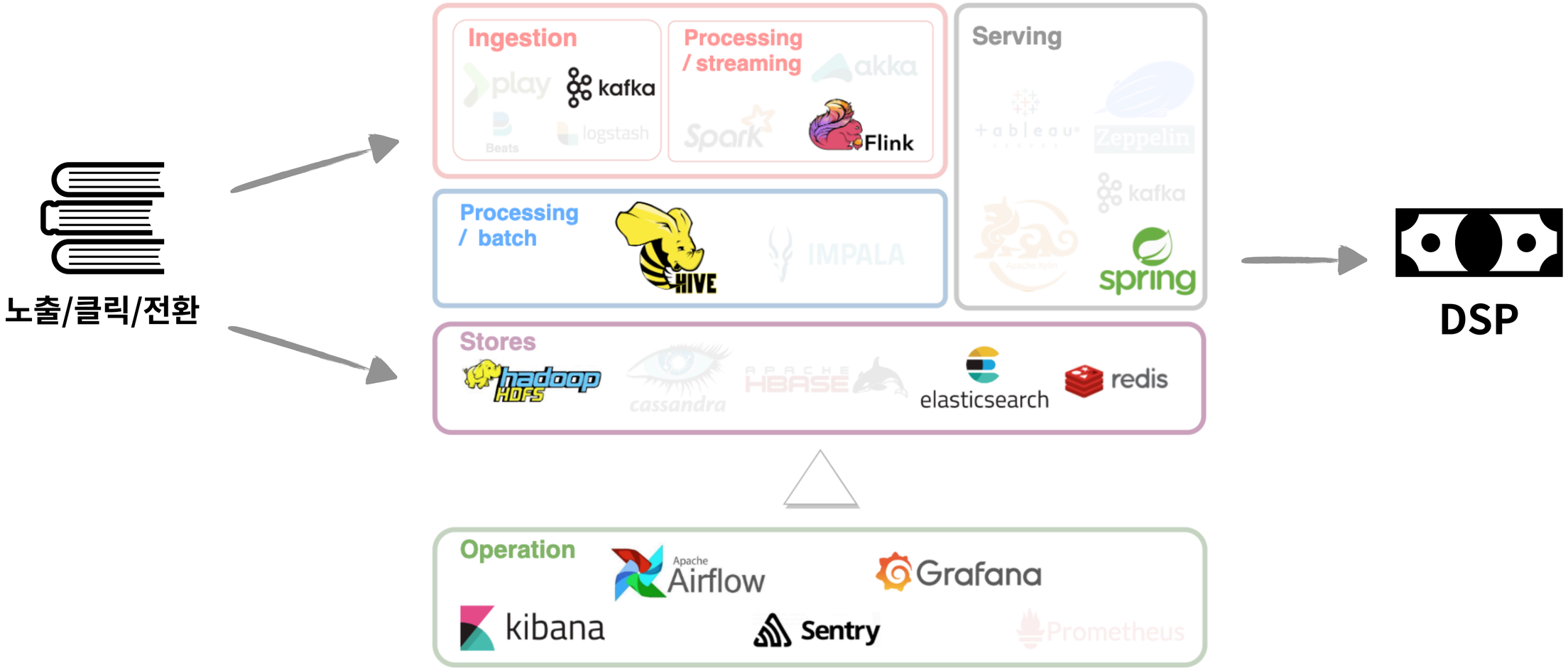
Data pipeline 보고서



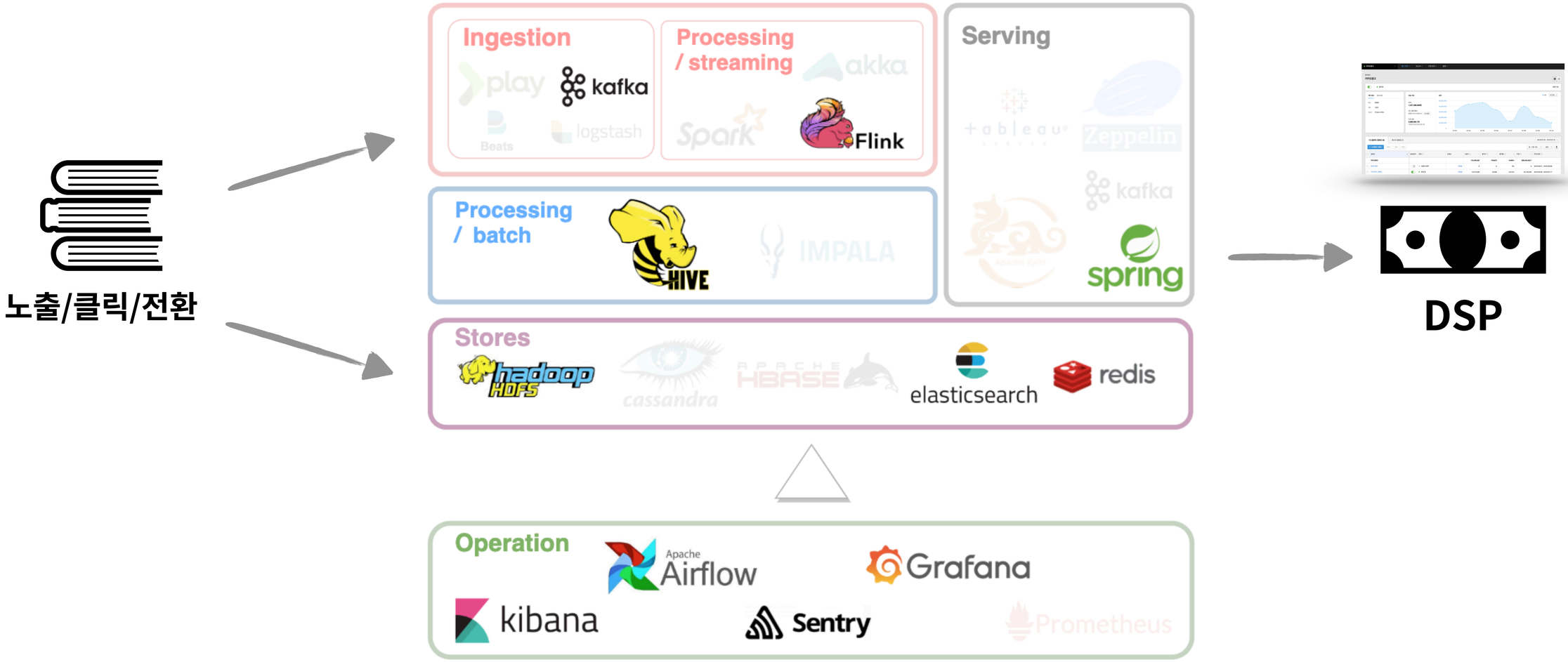
Data pipeline 보고서



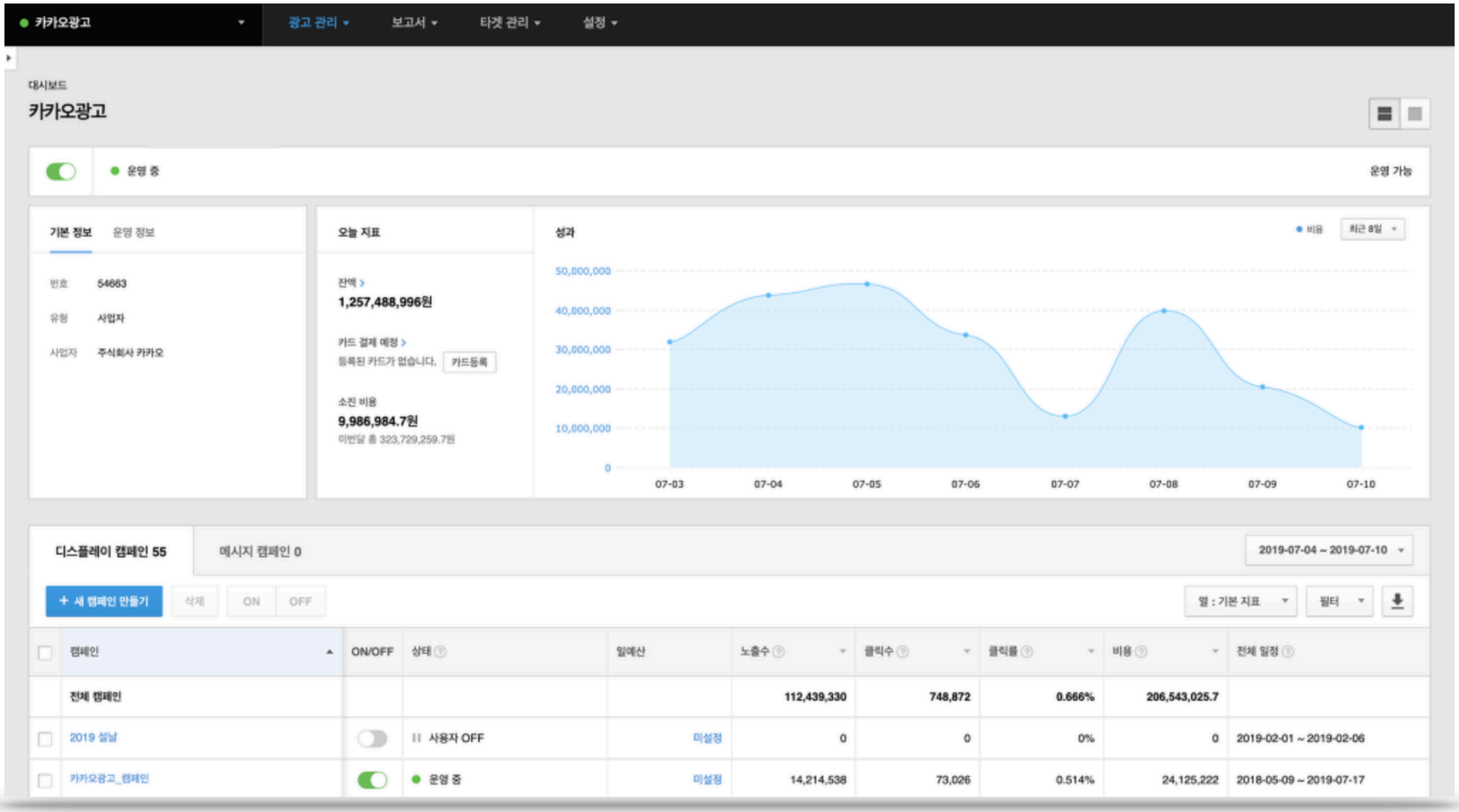
Data pipeline 보고서



Data pipeline 보고서



Data pipeline 보고서



이미지출처: https://ad.kakao.com/contents/resource/38

얼마나?

일일 처리 이벤트 수	111 억 (129k/s)
일일 수집 로그 사이즈	59 TB
보관중인 로그	14 PB
Batch 작업	90
Streaming 작업	30

얼마나?

Batch 처리 클러스터	189 nodes / 1 cluster
Streaming 처리 클러스터	144 nodes / 4 cluster
Store 클러스터	142 nodes / 8 cluster (+12PB)
Collect/serving 서버	136 nodes / 6 cluster

마치며

- Data scheme & lineage manager
- Streaming job manager
- Generic reporting platform
- Inter-cluster failover / jobs
- Airflow on k8s
- Operating bots

목적: 더 나은 데이터 사용환경을 만들기 위해

더 빠른 데이터 수집/처리/제공

더 안정적인 시스템 운영

데이터 관계 정보 제공

감사합니다

Q & A