Big Data

AccuInsight+ 데이터 분석 플랫폼 서비스

SK ICT Tech Summit 2018

SK 주식회사 C&C / AccuInsight+ Unit

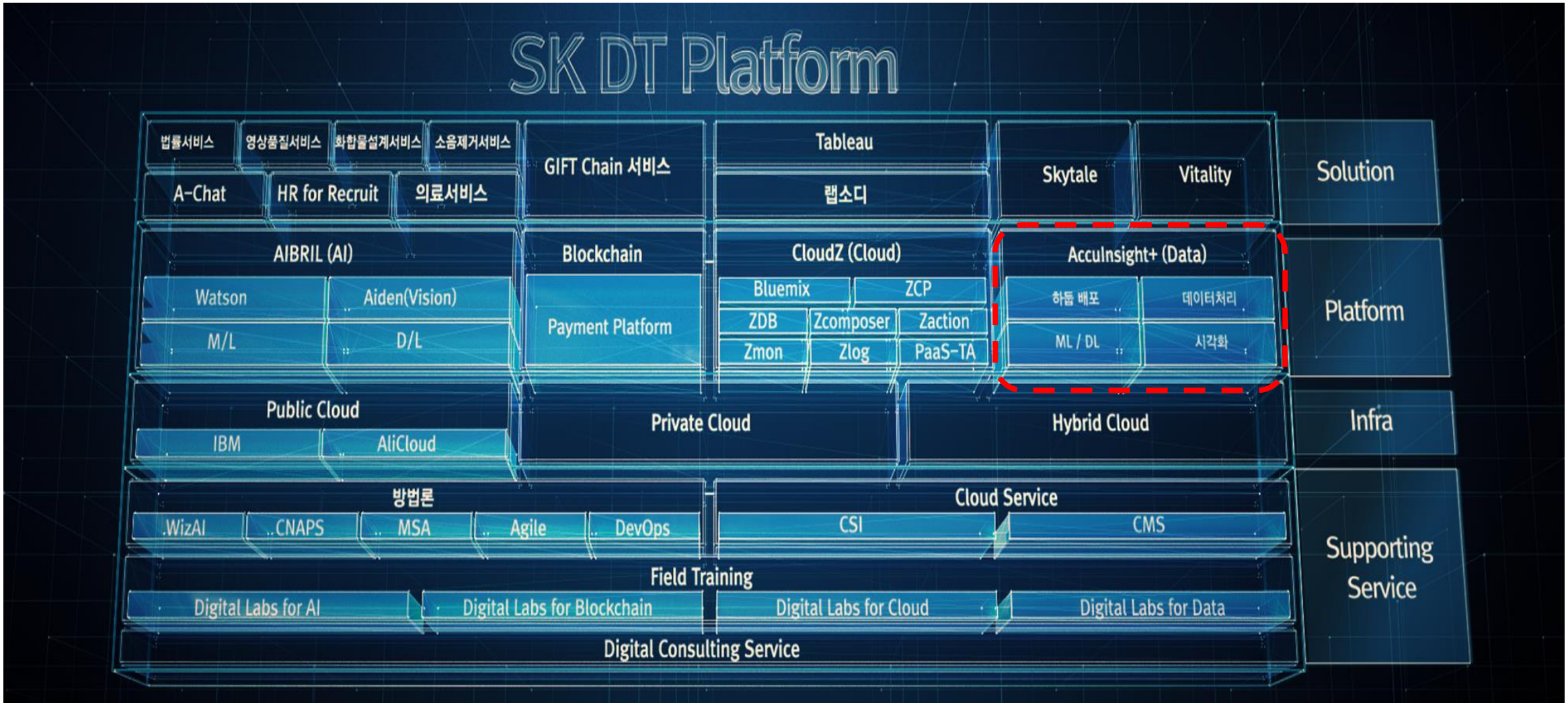
서정욱 위원

Contents

목 차

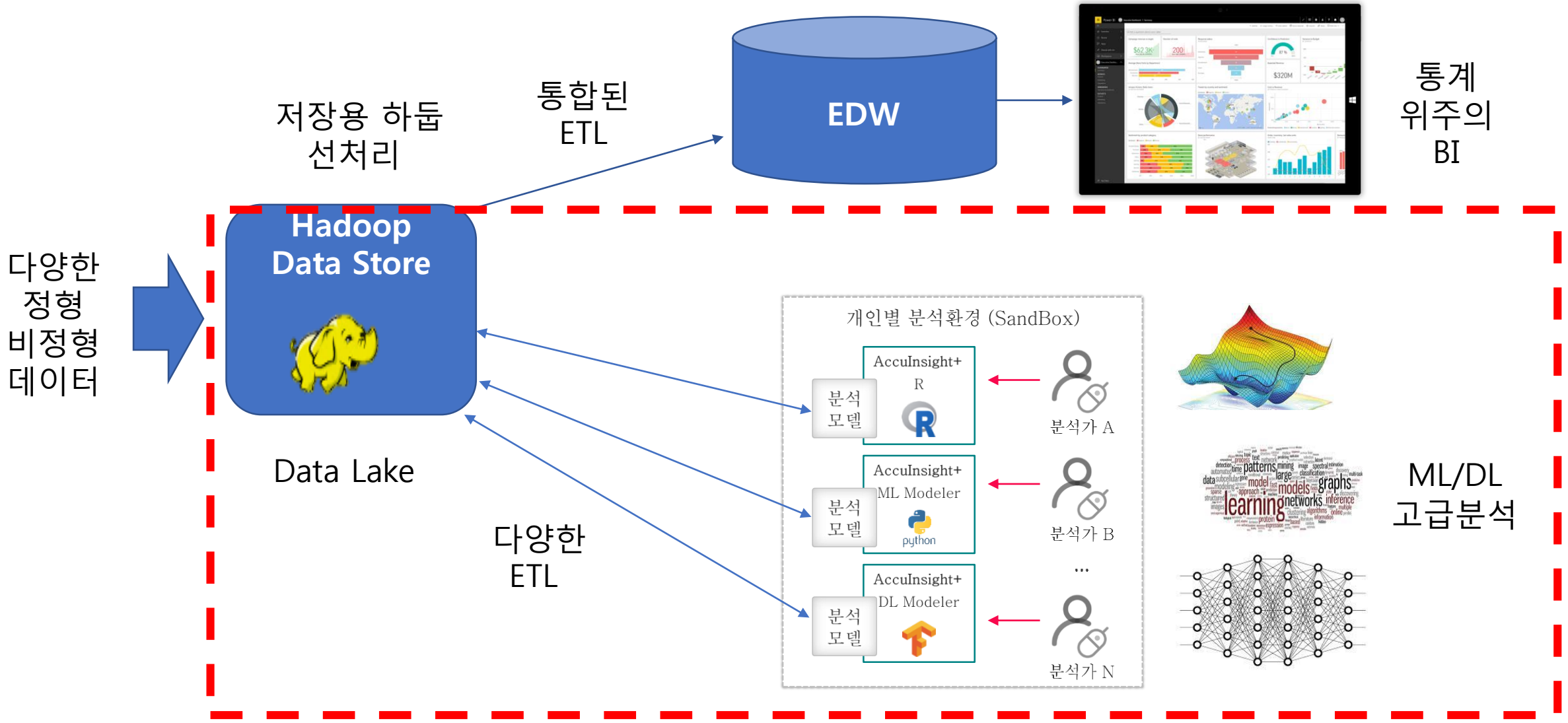
1. 서비스 개요 – SK DT Platform

SK DT Platform 전체 아키텍처



1. 서비스 개요 – Data Lake vs Data Warehouse

Data Warehouse 에서 Data Lake로 진화



1. 서비스 개요 – AccuInsight+ 아키텍처

SaaS Layer

DHP | Batch Pipeline | Realtime Pipeline | BigQL | ML Modeler | DL Modeler | Cloud Search | Data Insight

Data APIs

수집 Client | 머신 러닝 | 실시간 Streaming | Hadoop Batch | RDBMS Batch | Global W/F | 운영 관리 | Container 배포 | ...

Docker Container 관리, Multi-Tenancy, MSA, 인증/권한관리, 작업관리(배치/실시간/ML/DL)
클러스터스케일관리 로그관리(Spark/MapReduce/Tensorflow), 이력관리, 데이터관리/모델관리/모델배포관리(DL)

Infra Layer

Data Lake

Object Storage

Hadoop Cluster

RANCHER | kubernetes | R | python | APACHE Spark + python | TensorFlow | PyTorch | Keras | scikit learn | docker

GPU GPU GPU GPU GPU

1. 서비스 개요 – Reference

SK hynix

Public Service PoC
Data Lake 구축

KEB 하나은행

Big Data Platform

AIA 생명 THE REAL LIFE COMPANY

헬스케어 분석 플랫폼

KDB산업은행

Big Data Platform

KYOWON 교원

Big Data Platform

KB 국민은행

AI Platform

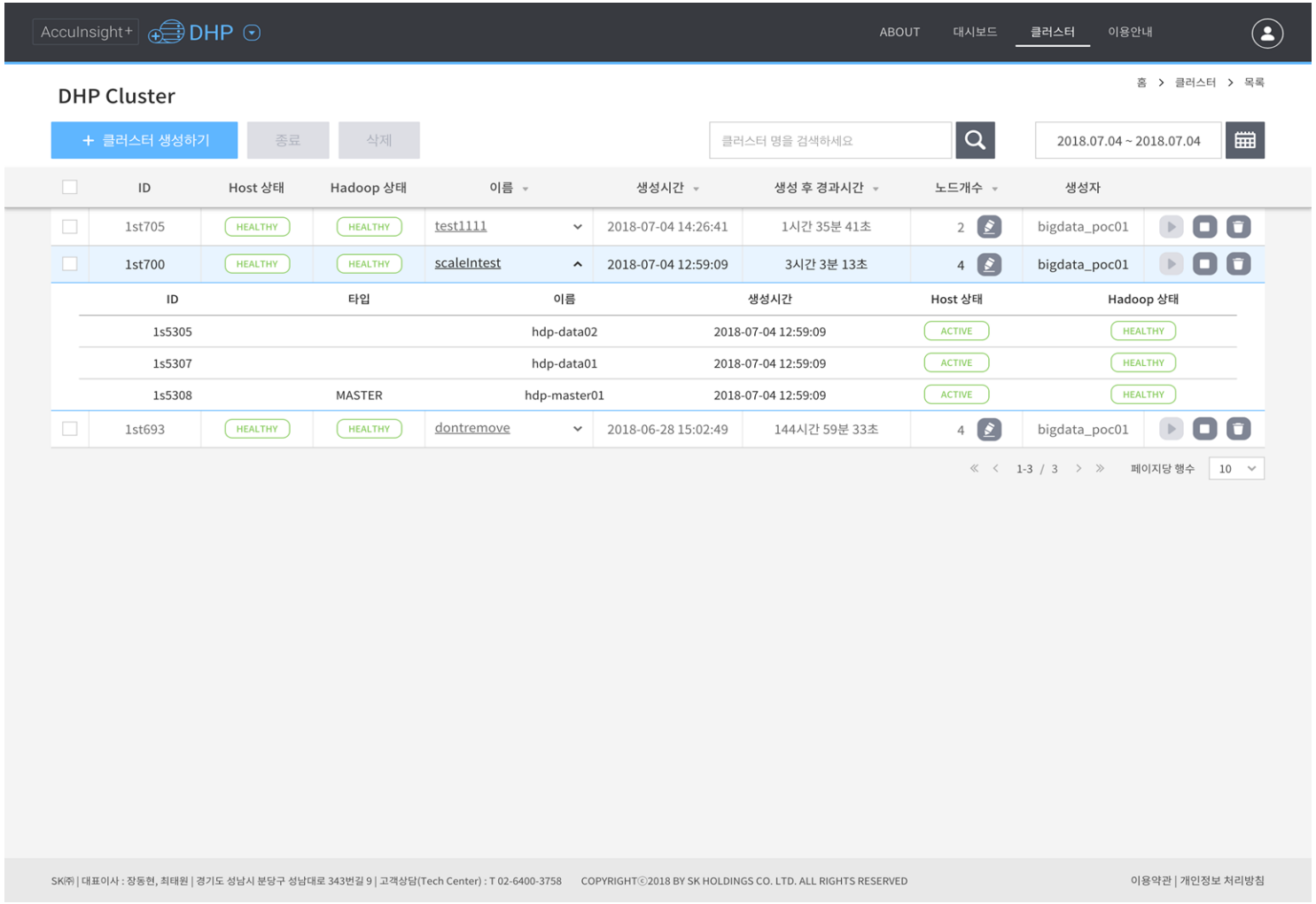
Contents

목 차

2. Dynamic Hadoop Provisioning

SK ICT Tech Summit 2018

관리형 Hadoop / Spark 클러스터 자동 배포 서비스

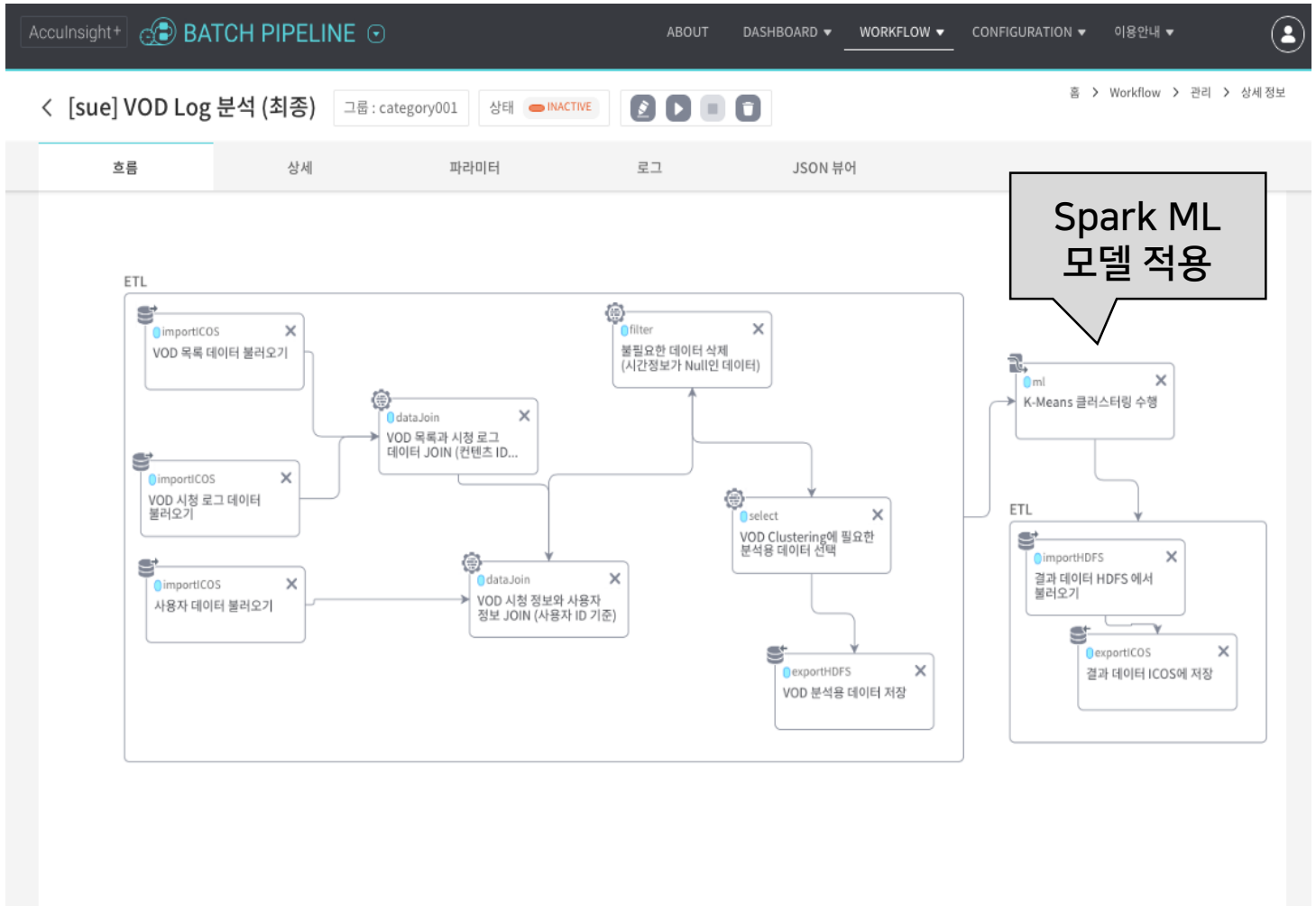


주요 기능

- Container 기반 Hadoop / Spark 클러스터 자동 배포
- Yarn Container 사이즈 조절 가능
  - Memory CPU 사이즈 선택 가능
- Data Node 추가 삭제
- Batch Pipeline, Real-Time Pipeline 및 Machine Learning(ML Modeler) 등의 분석작업에 필수 Hadoop Eco SW데이터 배포
- 자원 모니터링 기능
- 실행된 모든 Job 이력 확인 (Hive / Spark)
- Jupyter Notebook 제공 – PySpark으로 분산처리 가능
- 대시보드 기능 (모든 자원 상태 모니터링)

2. Batch Pipeline

Spark을 활용한 ETL과 Oozie 기반의 워크플로우 및 배치 스케줄러 기능 제공

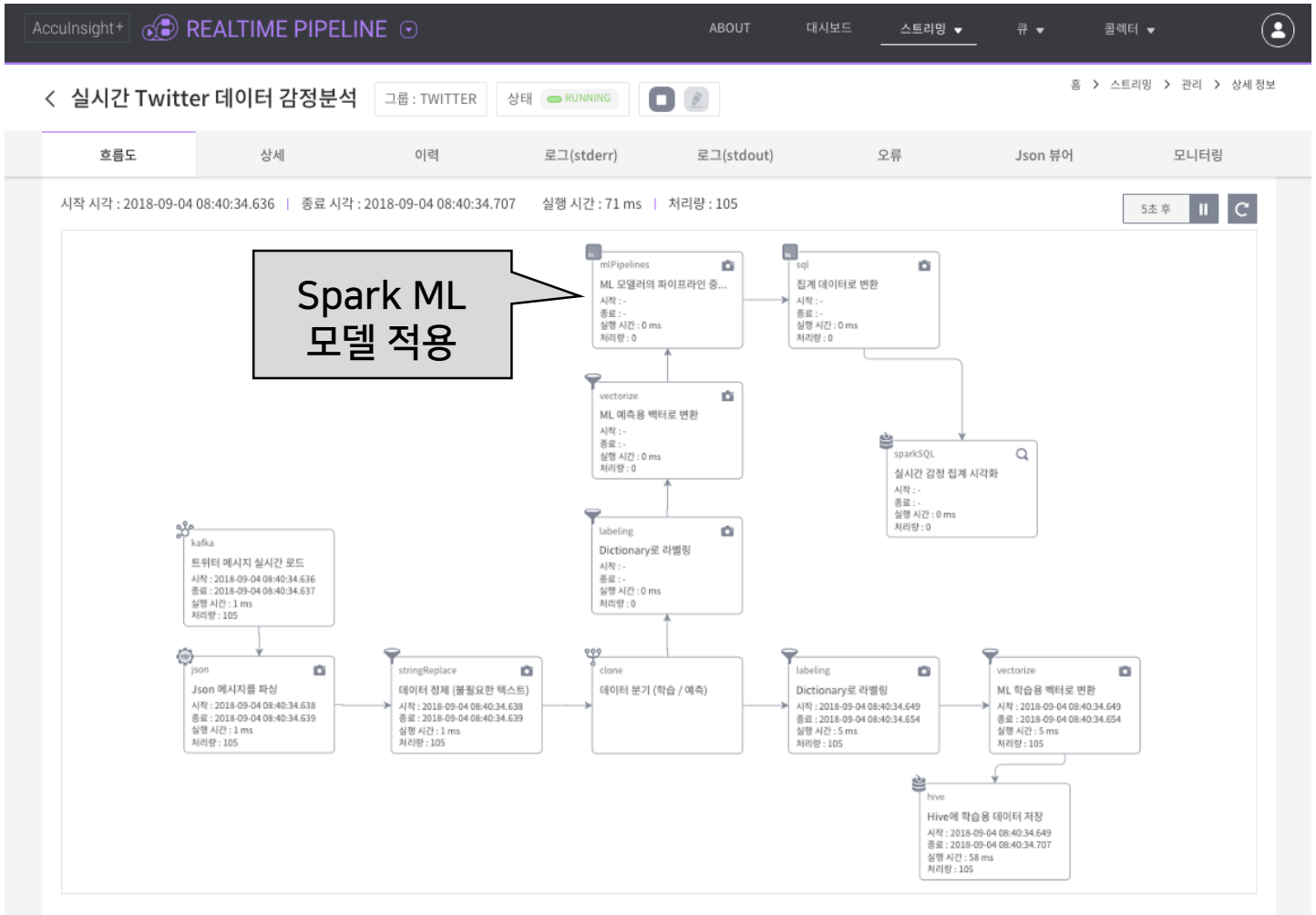


주요 기능

- Spark을 활용한 안정적인 빅데이터 ETL (PySpark)
- Oozie 기반의 워크플로우 관리
- Spark ETL + Oozie W/F 동시 사용
- Job Scheduling (시간/일/주/월배치)
- ML modeler에서 생성한 모델을 로딩하여 운영 적용
- 실시간 MapReduce/Spark 로그 확인
- 실행된 모든 Job 이력 확인 (Hive / Spark)
- On-premise 클러스터 정보 등록
- Job별 클러스터 선택하여 수행 (On-premise/SandBox)
- 대시보드 기능
    - 시간대별 Job 실행 개수
    - Job별 실행 예측 시간 제공

2. Real-Time Pipeline

Queue / Collector / Streaming 처리를 워크플로우 기반으로 UI에서 쉽게 구현 할 수 있는 서비스

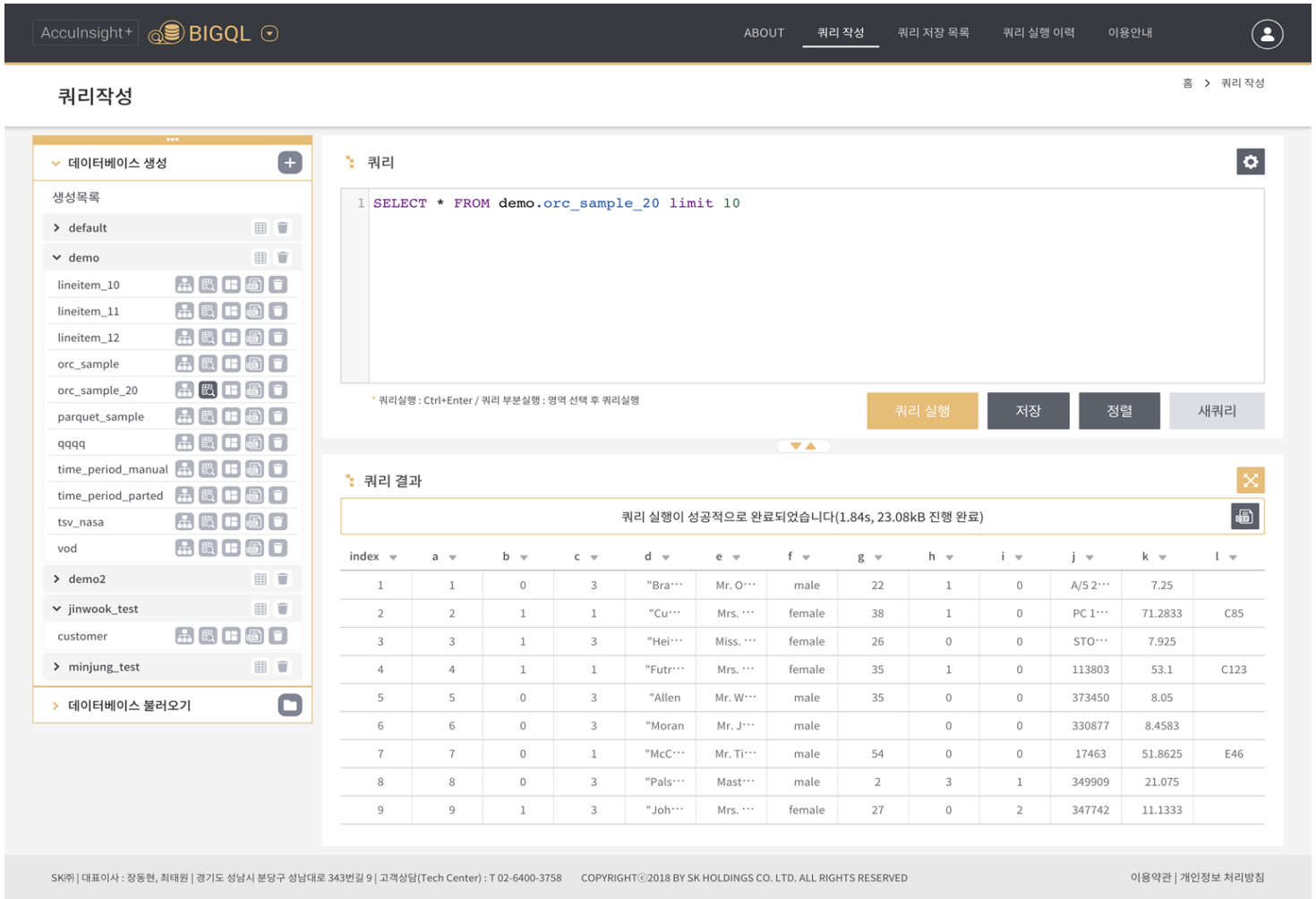


주요 기능

- 큐 클러스터 및 인스턴스(topic)들에 대한 관리
- Collector 서비스 (Flume 기반 Source Chanel Sink)
- 스트리밍 서비스
  - 다양한 경로의 스트리밍 데이터 수신
  - 자동 파싱 및 스키마 생성 (Json, CSV)
  - 노드 별 데이터 처리 현황 파악 - 데이터 스냅샷
  - 실시간 스트리밍 처리
  - Rule 엔진을 활용한 CEP
  - Spark SQL을 활용한 실시간 시각화
  - 실시간 처리량 및 처리시간 모니터링 제공
  - 실시간 로그 확인
- 대시보드 기능
  - 규 콜렉터 스트리밍 처리에 대한 현황

2. BigQL

오브젝트 스토리지나 Hive에 적재된 데이터를 쉽게 분석할 수 있는 대화형 쿼리 서비스



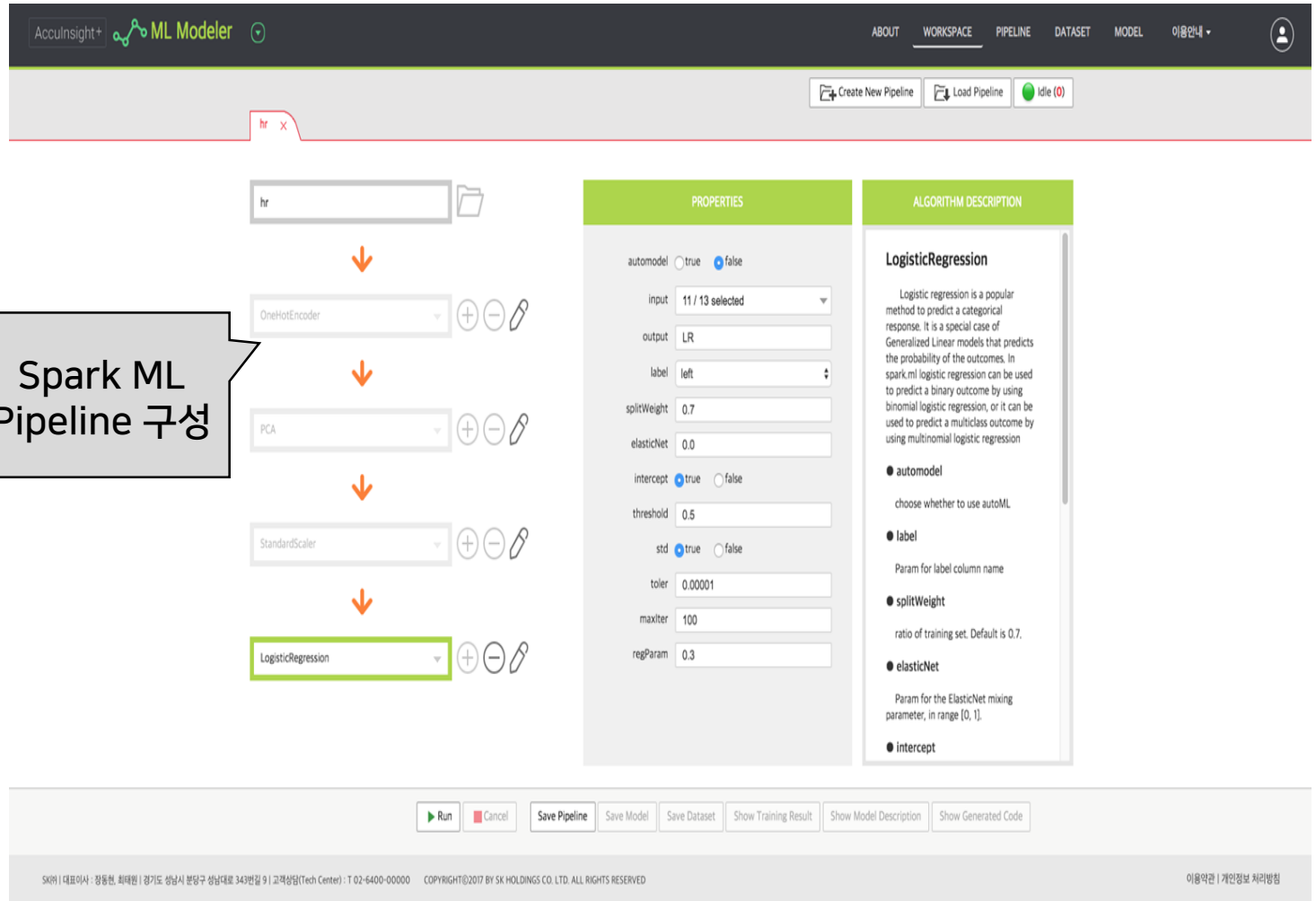
주요 기능

- 스토리지에 적재된 데이터 파일 기반으로 관리
- 표준 ANSI SQL 사용
- 다양한 파일 형식 지원
  - CSV, Apache Web log, Parque, ORC
- Public 서비스의 경우 스캔한 데이터 량에 따라 과금
- Hive 나 Object Storage에 적재된 데이터 스키마 관리 및 카탈로그 제공
- Hive / Object Storage에 적재된 데이터 간 이기종 조인 가능
- 사용한 쿼리나 쿼리 결과 저장 및 재사용
- 스토리지나 Computing 자원 분리하여 확장

2. Machine Learning Modeler

SK ICT Tech Summit 2018

대용량 데이터를 코딩 없이 UI를 통해 머신러닝을 구현 할 수 있는 서비스

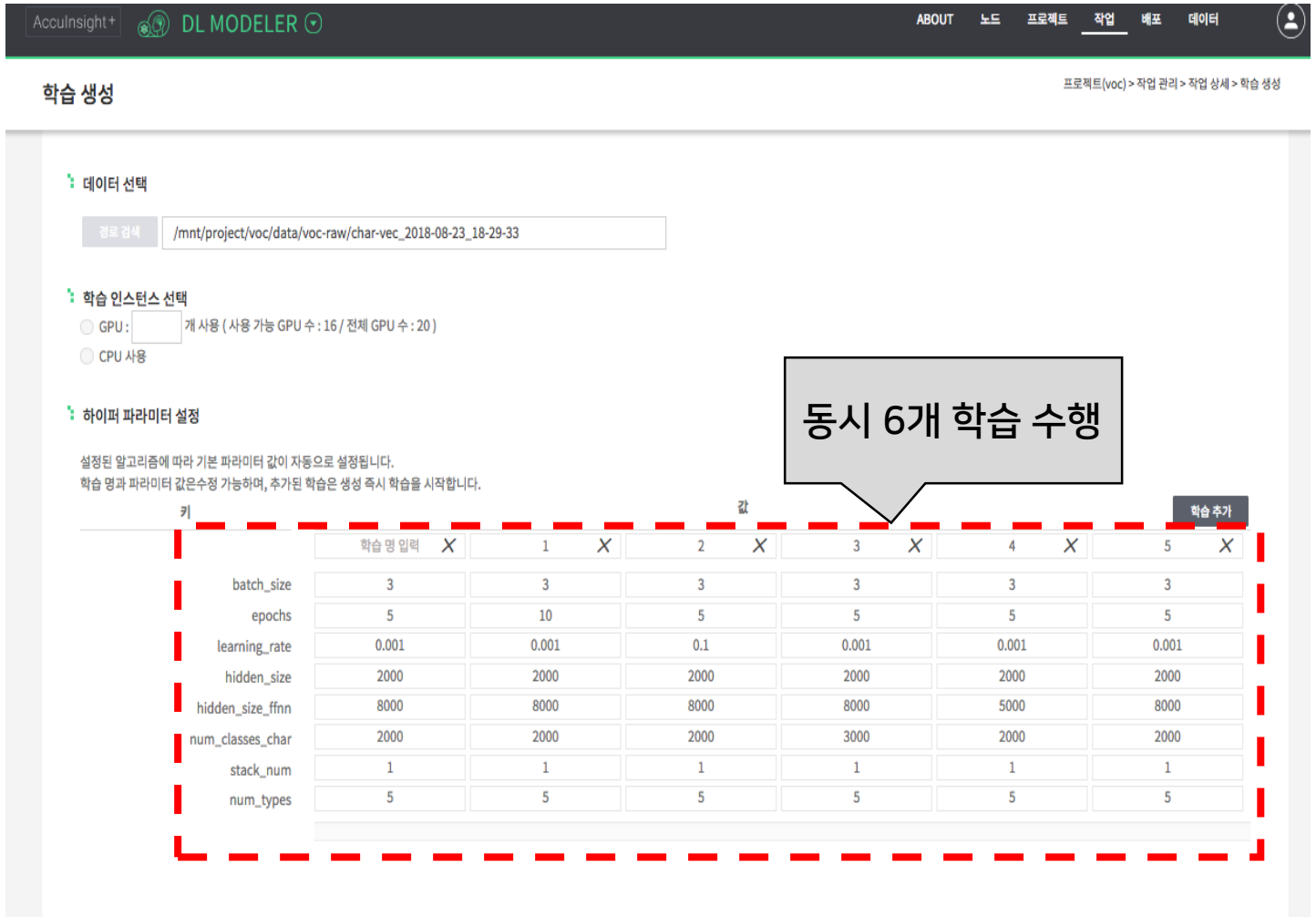


주요 기능

- R에서 학습할 수 없는 대용량 데이터 머신러닝 가능
- 코딩없이 쉽게 Spark ML의 Pipeline 구성하여 실행
  - 분산 병렬 pre-processing과 algorithm의 자유로운 조합hyper parameter 설정
  - 12개 선처리 / 19 개 알고리즘 제공
- Auto ML 기능
  - classification 알고리즘의 최적 hyper parameter 자동 탐색
- Data / Pipeline / Model 관리
- 학습 결과, 모델의 시각화
  - ROC curve, feature importance, tree view, 등 시각화된 학습 모델의 성능 분석

2. Deep Learning Modeler

딥러닝 분석 Life cycle을 완전 관리형으로 제공

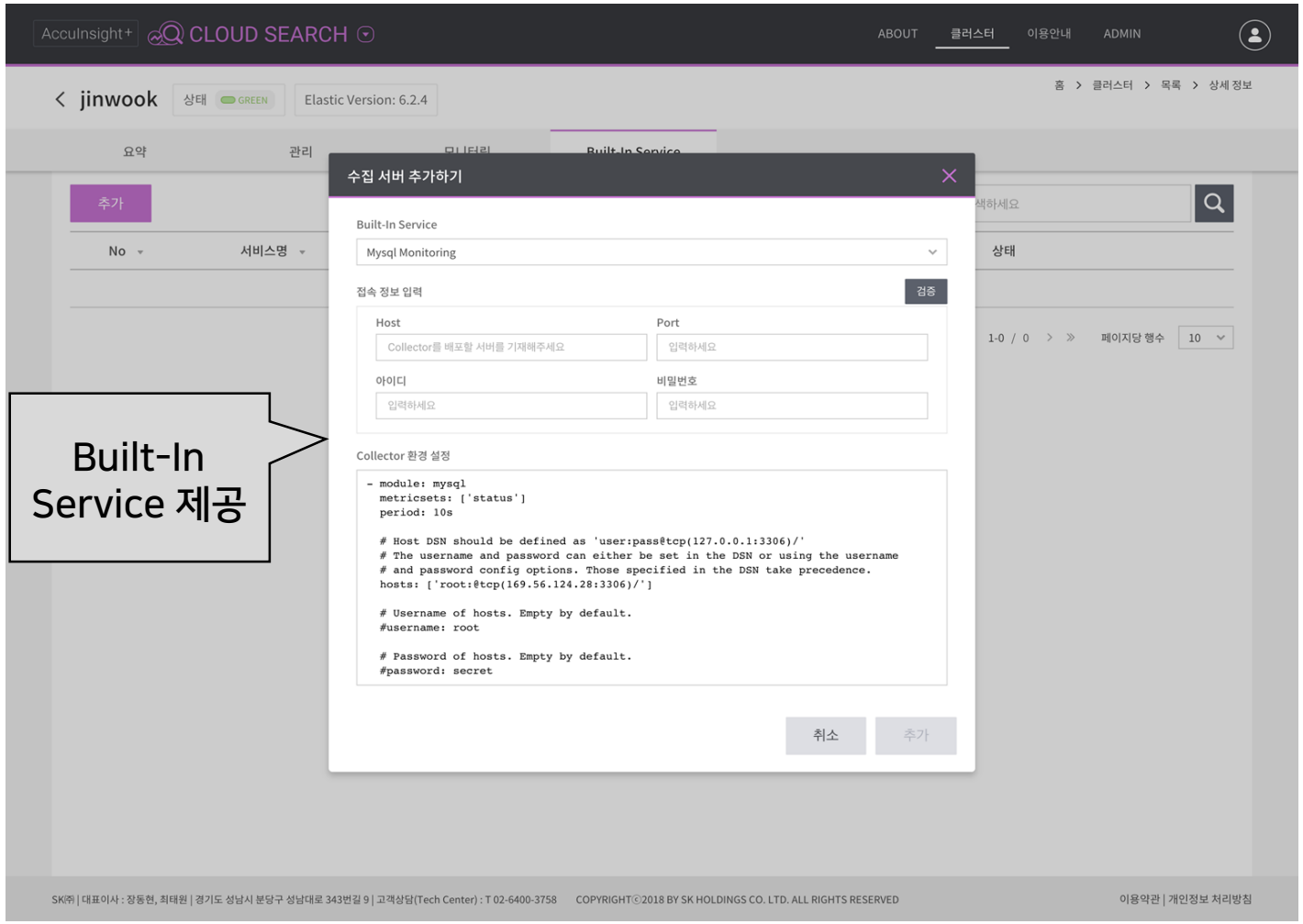


주요 기능

- 분석 주제별 프로젝트 관리
- GPU 서버 관리
- Built-In 알고리즘과 Jupyter Notebook 제공
- 동시에 여러 학습 수행
  - 다양한 Hyper Parameter 세팅하여 동시 학습하여 최적의 모델 신속히 찾음
- One-Click 배포
  - 최적의 모델 One-Click으로 API 형태로 배포
    원천 / 예측 데이터 관리
- 비정형 데이터 pre-processing 알고리즘 제공
  - 이미지 Augmentation, 문장분리, 형태소분석, Word2Vec
  - 한글특화 사전 제공
- 작업 및 학습 history 추적 (hyper param등)

2. Cloud Search

ELK 스택을 검색엔진으로 배포 관리해주고 Built-In 알고리즘으로 쉽게 다양한 로그 분석해 주는 서비스



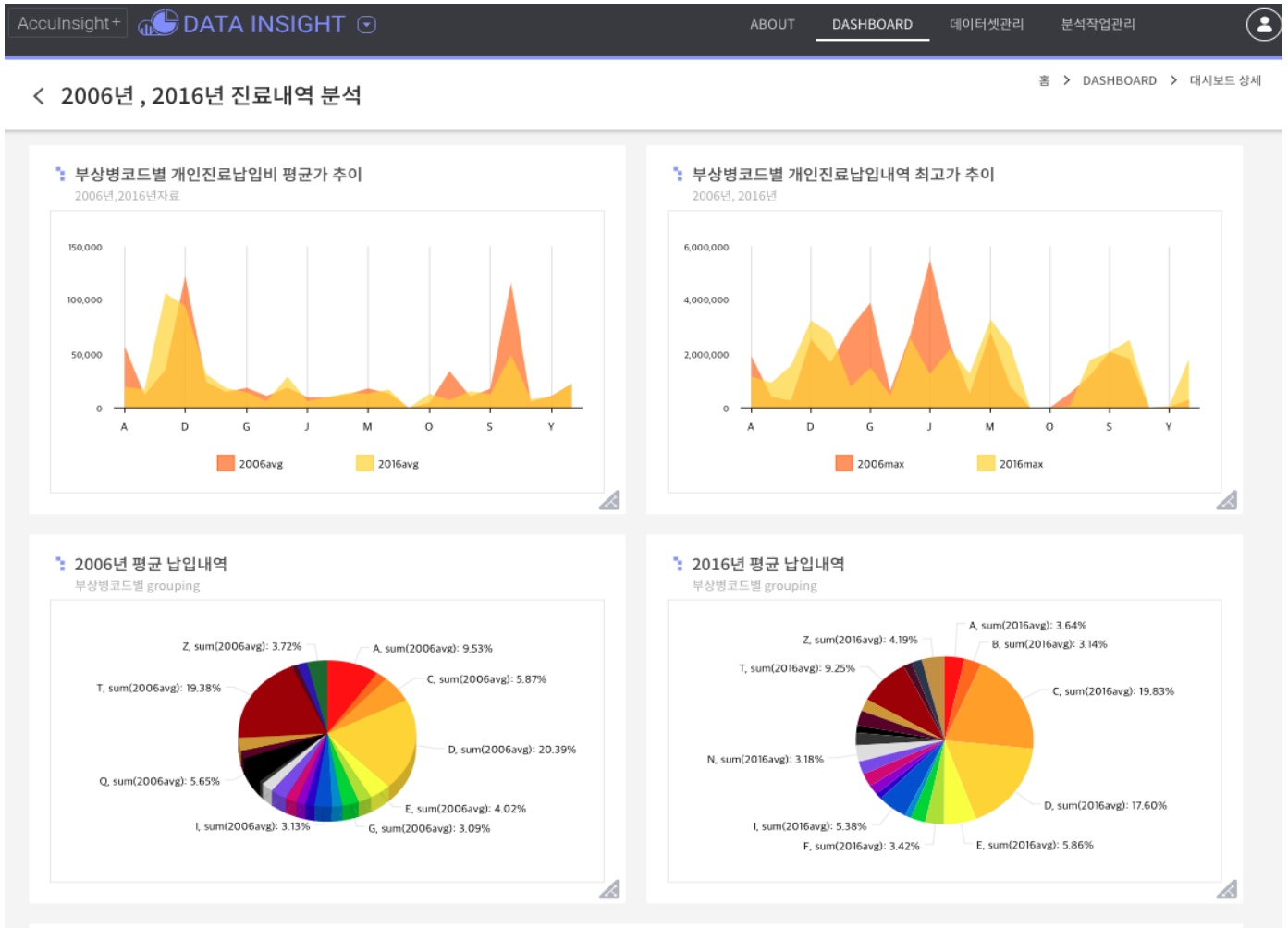
주요 기능

- 간단한 ELK Stack 배포 및 확장
  - Scale-out 지원
  - Resource Type별 클러스터 생성
  - ES/Kibana end-Point 제공
- 완전 관리형 서비스
  - Cluster health Check 지원
  - Node Log 조회 지원
  - 사용자가 직접 접속 가능한 Node 단위 CLI 지원
- Built-in Service
  - 사용자 목적에 맞는 Built-in 서비스를 제공
  - MySQL, System, Apache Web, Kafka
  - Collector 배포/설정 및 Kibana Dashboard 설정까지 One-click으로 구현

2. Data Insight

SK ICT Tech Summit 2018

다양한 데이터를 연결하여 시각화 할 수 있는 BI 서비스



주요 기능

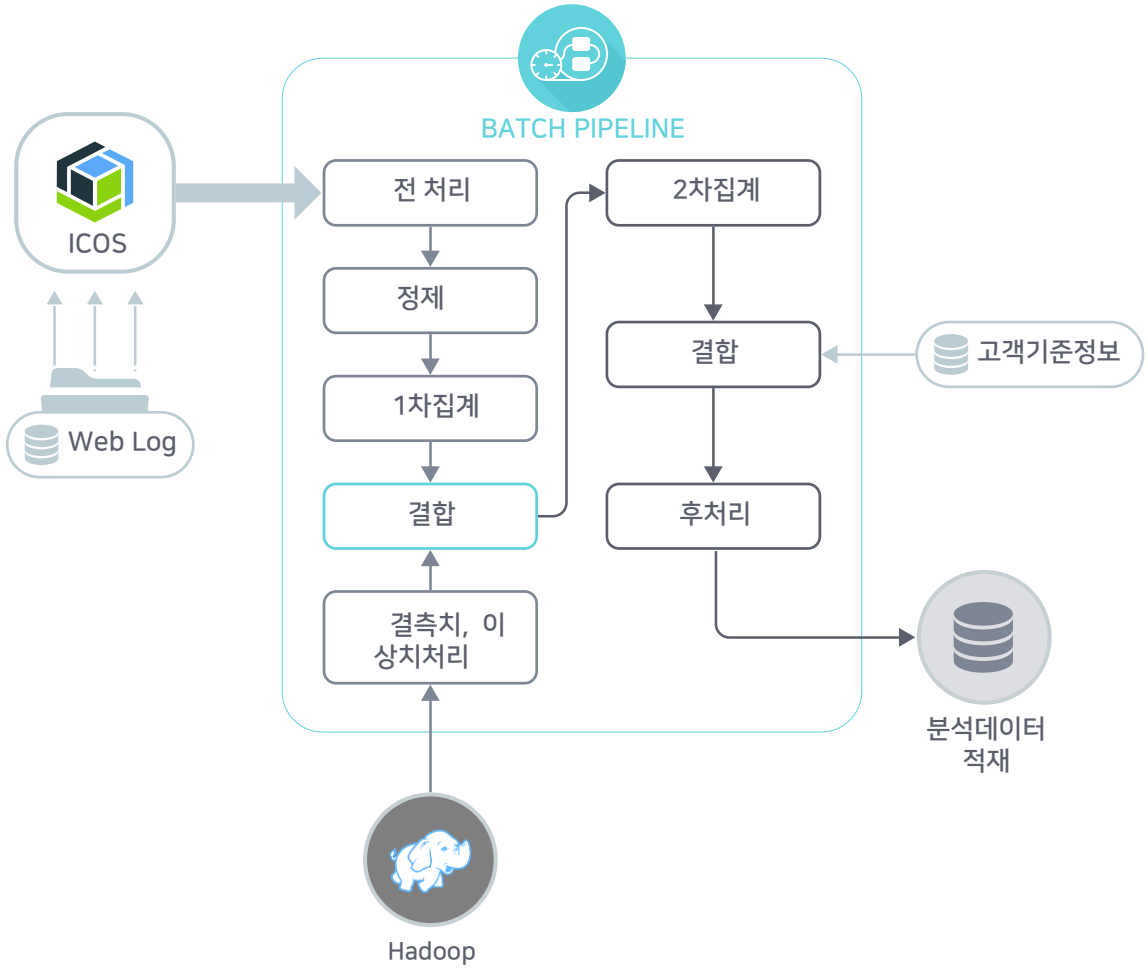
- 데이터 수집을 위한 다양한 리소스 어댑터 제공
  - 여러 인프라에 분산되어 있는 데이터를 하나의 공간에 가져 올 수 있게 여러 리소스 어댑터 제공
  - 로컬/Hive/AWS RDS/MySQL/Maria DB/S3
- 데이터 셋, 분석 작업 관리 기능
  - 데이터 상세 편집, 스키마 편집
- 차트 관리 및 편집 기능
  - 수집된 데이터를 별도 가공없이 바로 다양한 시각화 자료 생성
- 대쉬보드 구현
  - 분석 완료된 작업들을 대쉬보드 형태로 구현

Contents

목 차

3. Use Case - 빅데이터 ETL

다양한 대용량 데이터를 안정적으로 ETL



1. 데이터 가져오기

- ICOS - Web Access Log, HDFS – 고객 통화정보, RDBMS - 고객 기준정보 적재

2. 데이터 변환하기

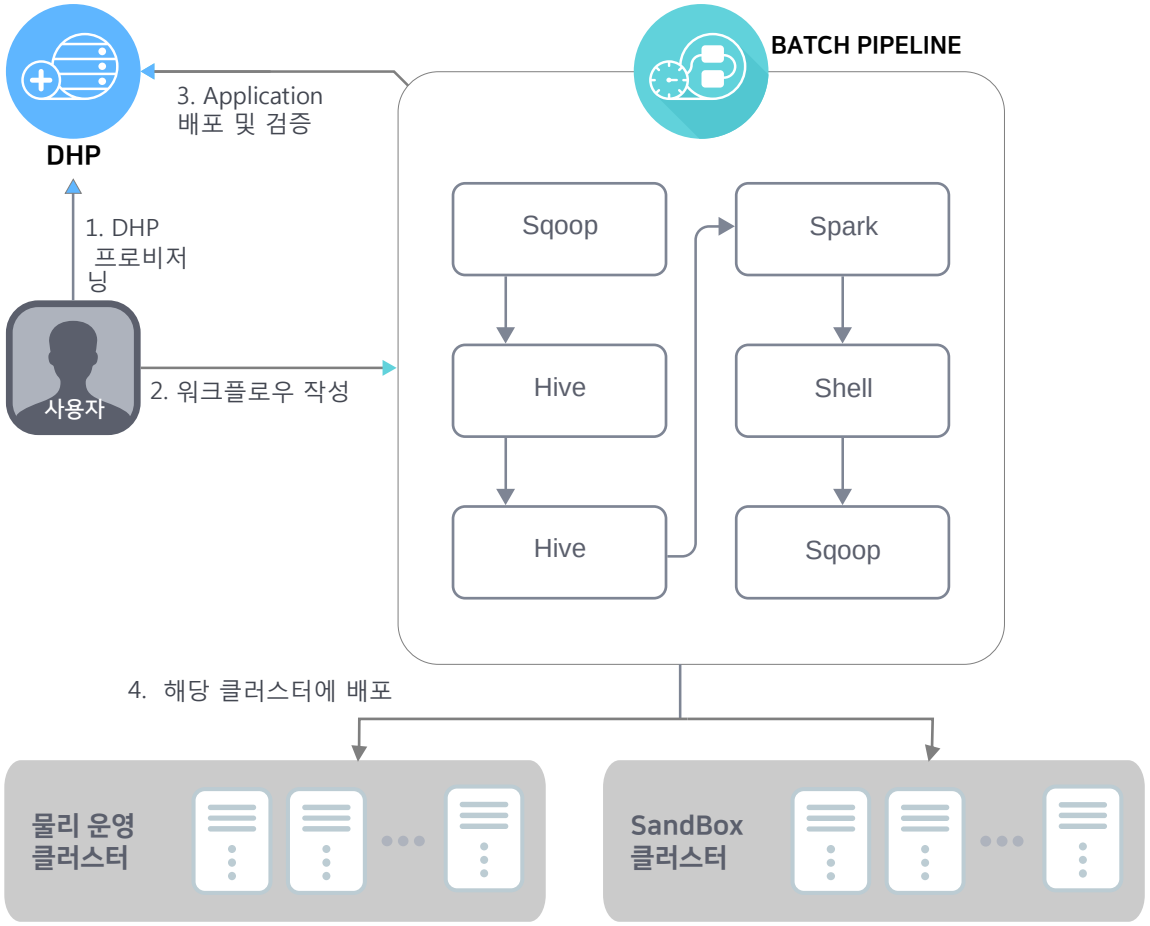
- 하나의 데이터 스토리지에 모을 필요 없이 서비스 상에서 조인, 필터링, 집계 작업 후 분석 데이터 ICOS/RDB/HDFS에 적재

3. 대용량 데이터 안정적인 ETL

- 대용량 데이터를 Spark으로 안정으로 분산 처리하여 타겟 저장소에 안정적으로 신속하게 적재 (ex. SandBox)

3. Use Case - 멀티 Hadoop 클러스터 활용

배치 작업 별 클러스터 선택이 가능하여 개발 클러스터에서 개발 검증 후 운영 클러스터로 쉽게 배포



1. 탄력적인 클러스터 운영

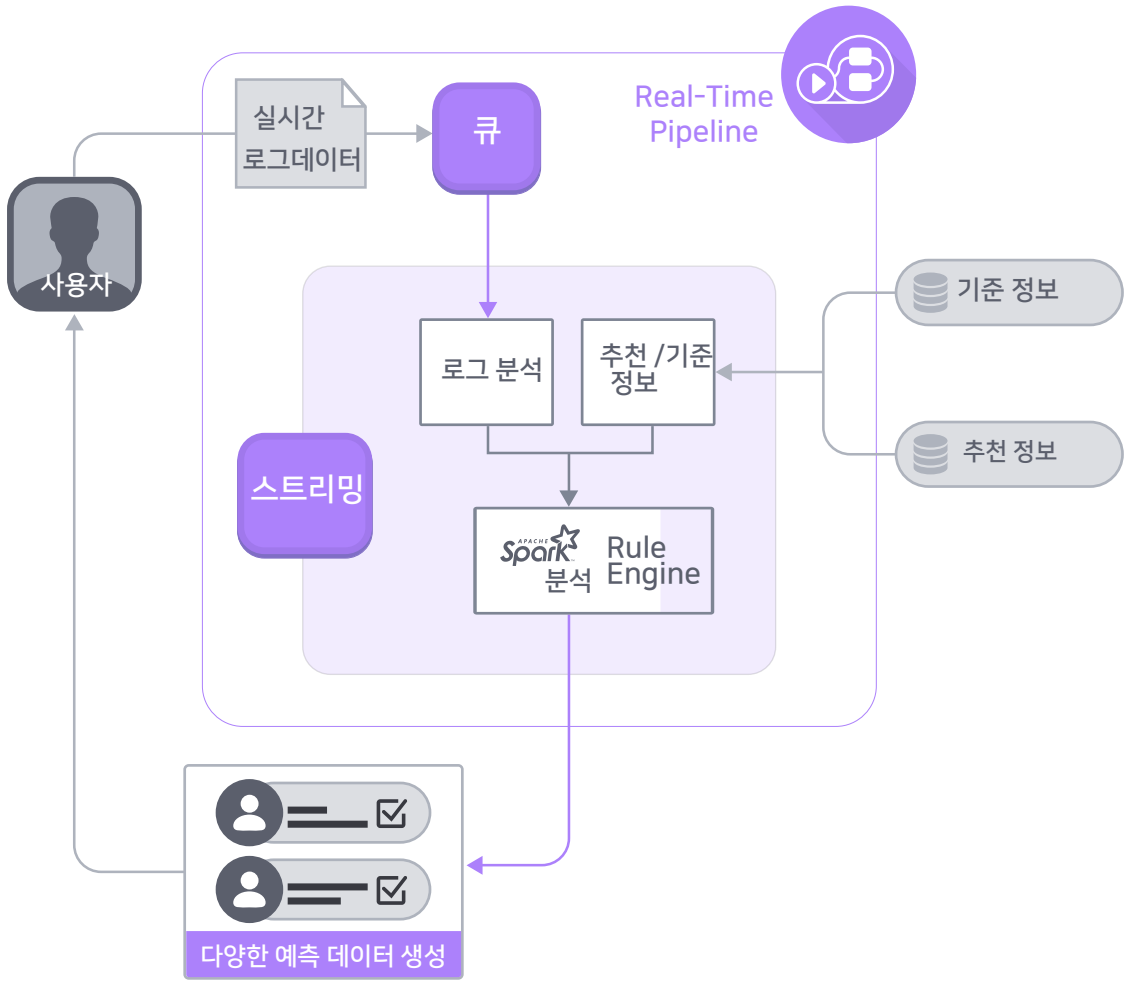
- Hadoop Cluster 운영 중에 개발환경을 구성하기 어려움
- DHP로 개발용 SandBox 생성하여 배치 작업 개발
- 배치 작업 검증 후 운영 클러스터에 개발된 배지 작업 배포 -> SandBox 삭제

2. 효율적인 배치 작업 및 인프라 관리

- 운영 환경에 영향을 주지 않고 별도의 개발 클러스터 운영
- 배치 작업 별 클러스터 선택하여 실행 가능
- SandBox를 Computing 노드로 활용

3. Use Case - 이벤트기반실시간 CEP

실시간 로그 데이터와 기준정보 및 분석된 데이터 결합하여 다양한 실시간 비즈니스 구현



1. 이벤트기반 마케팅

- 사용자의 결제 위치 기반 제휴사 정보 추천
- 사용자의 위치 및 소비 패턴정보를 활용하여 추천

2. 실시간 침입 탐지

- 보안장비에서 발생되는 로그를 침입 탐지 Rule에 적용하여 실시간 침입 탐지

3. 금융사/통신사 Fraud Detection

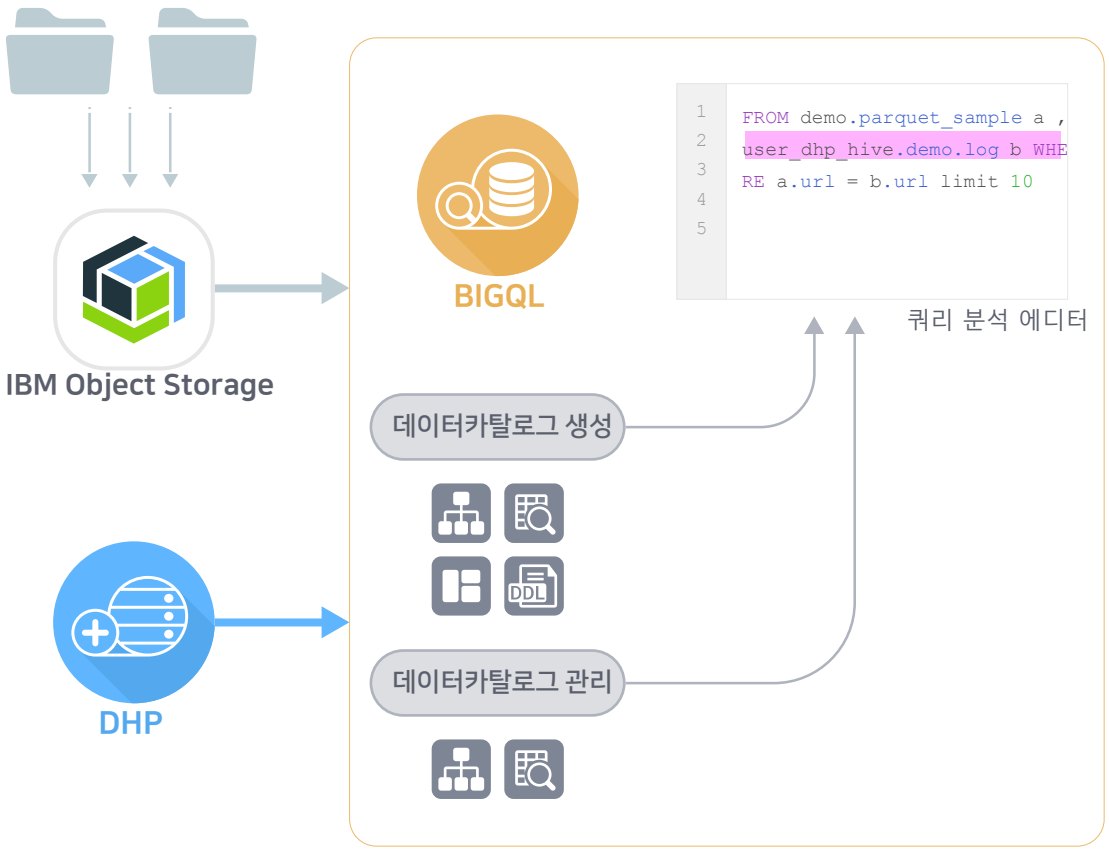
- 사용자의 소비 패턴이나 위치정보와 연계하여 이상거래 감지
- 사용자 결제 내역과 기기 정보를 결합하여 비정상 단말을 통한 결제 시도 감지

4. 생산/제조 분야 Fault Detection Control

- 장비 로그 데이터 분석하여 설비 이상 감지

3. Use Case – 이기종 데이터 조인

BigQL을 활용하여 Object Storage와 Hive에 적재된 이기종 데이터 관리



1. DHP 데이터 분석 환경 제공

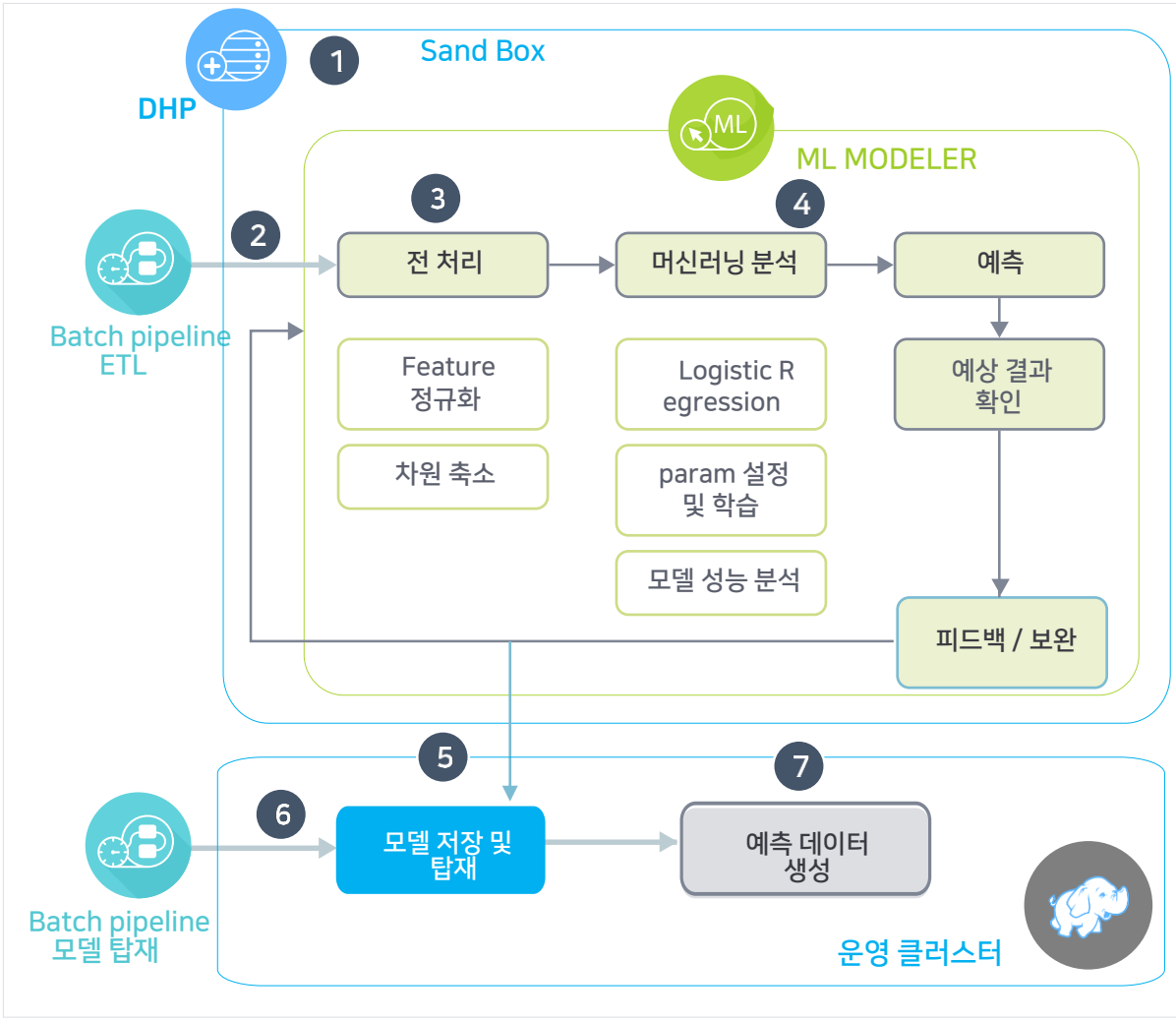
- DHP 사용자인 경우 자동으로 카탈로그 추가
- DHP 메타저장소에 연결하여 BIGQL 데이터와 조인 분석 환경 제공

2. 데이터카탈로그 관리 기능 제공

- UI를 통한 데이터카탈로그 관리 기능
- 데이터 미리보기, 스키마 보기 기능 제공

3. Use Case - Sand Box를 활용한 ML 모델 생성 및 운영 배포

ML 모델을 활용한 다양한 예측 데이터 적용 (가격예측/수요예측/양불판정/개인화추천)

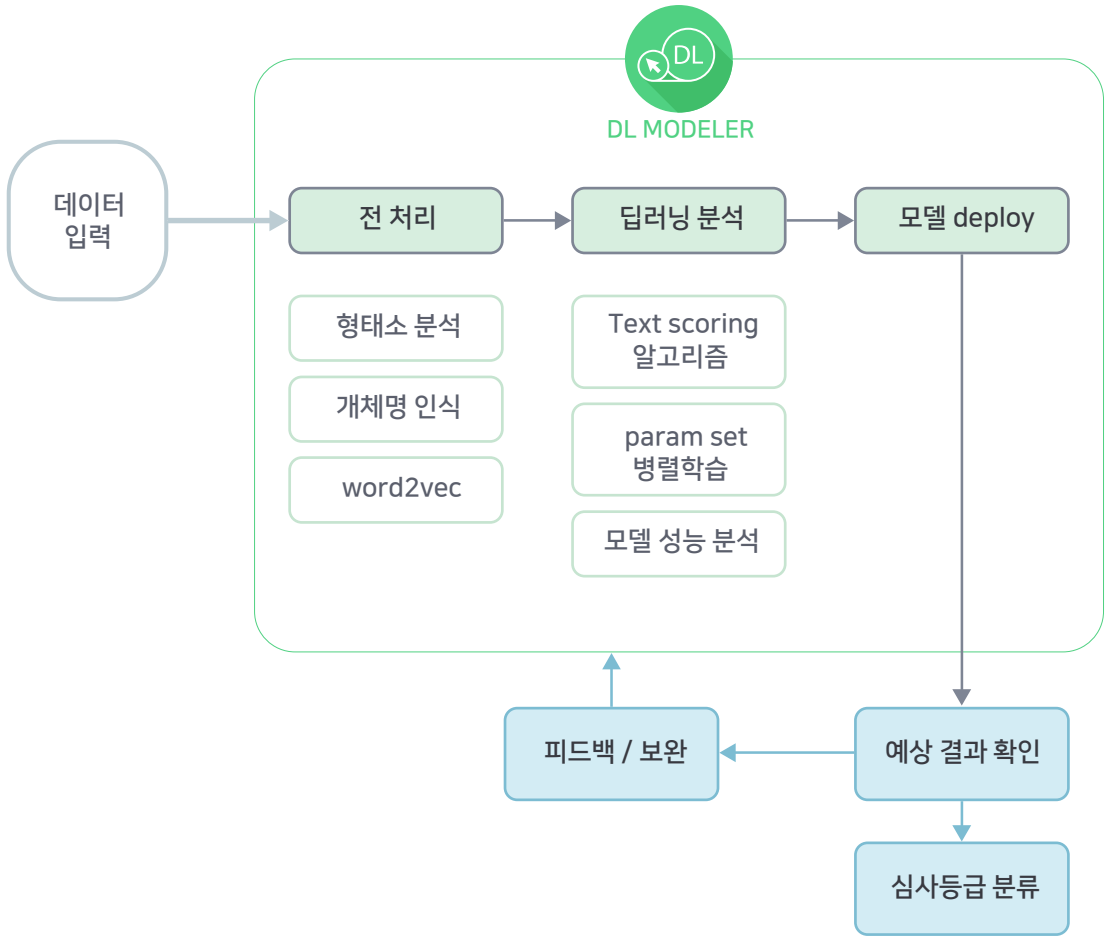


1. DHP를 활용한 SandBox 배포
2. Batch pipeline 서비스에서 ETL로 데이터 가공, HDFS 저장
3. ML Modeler에서 HDFS에서 ETL된 데이터 로딩
4. Workspace canvas에서 파이프라인 디자인

   Min max Scaler로 데이터 정규화 PCA로 차원 축소 및 outlier제거 Logistic regression으로 양불 판별 모델 생성
5. 운영 Hadoop Cluster에 최적 모델 저장
6. 저장한 모델을 Batch Pipeline에 탑재하여 예측 데이터 생성 하는 배치 작업 생성
7. 배치 작업 스케줄링

3. Use Case - 딥러닝을 활용한 비정형 데이터 분석

다양한 비정형 데이터를 선처리 후 딥러닝 분석 적용 (보험자동심사등급분류/VOC 데이터 분석)



1. 고객으로부터 조사된 설문 Text 데이터 입력

2. DL 데이터셋 전처리
   - 형태소 분석
   - 개체명 인식
   - Word2Vec

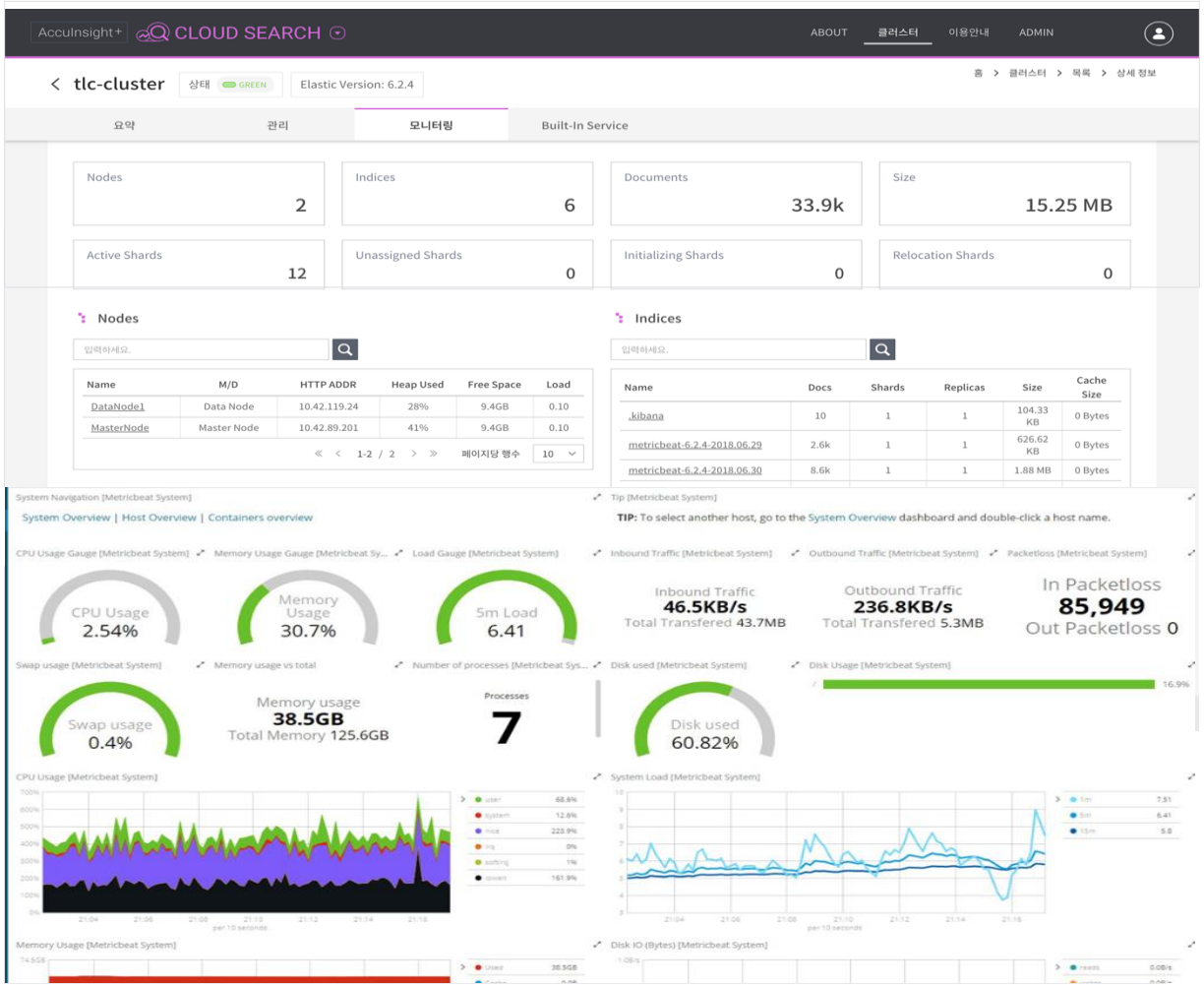
3. 작업 생성 및 학습
   - Basic 작업 생성
   - Text classification 알고리즘선택 학습 생성 후 GPU instance, hyper parameter 설정 학습 및 결과 확인

4. 모델 배포

5. 예측을 통한 최종 심사등급 분류

3. Use Case - 데이터센터의 서버 상태 모니터링

Cloud Search를 활용하여 대규모 서버에서 발생하는 로그를 쉽고 빠르게 모니터링



1. 빠르고 간편한 로그 수집

- 한번의 클릭으로 대규모 서버의 로그를 빠르게 수집 및저장

2. 실시간 서버 자원 사용량 시각화

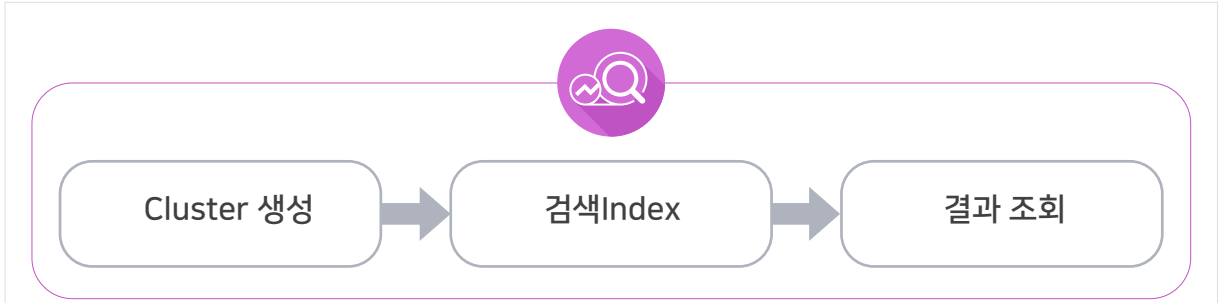
- 서버의 자원 사용량 (CPU, Memory, Disk I/O, Memory I/O)
- 실시간 조회 많이 사용되는 프로세스 Top 10 조회

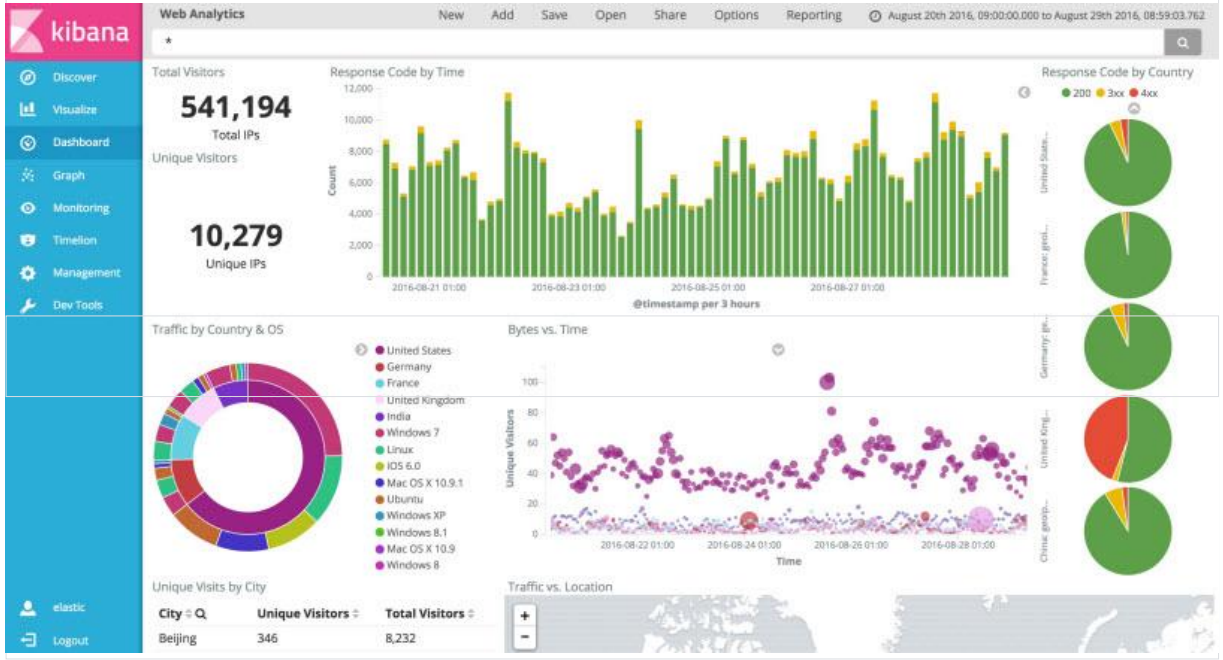
3. 원클릭 모니터링 환경 구성

- 사용자가 Web UI에서 One-Click으로 대규모 서버를 모니터링하는 시스템 구성 가능
- 모니터링 대상(서버 IP)만 지정하면, 수집 Agent설치 및 시각화까지 자동화 가능

3. Use Case - 검색엔진

온라인 쇼핑몰의 대용량 검색 및 분석 서비스에 활용





1. 상품 데이터 검색

- 1억 건 이상의 상품 데이터 기반의 빠른 검색

2. 고객 행위 지표 수집

- 고객이 검색하거나, 장바구니에 담은 상품등의 행태 지표를 수집
- 실시간 수집된 지표를 색인화 하여 빠르게 고객의 패턴을 분석 가능

3. 사용자 구매 후기 분석

- 형태소 분석을 통한 사용자 평가분석
- 단어의 감성(긍정, 부정)을 패턴을 이용하여 제품의 감성평가 분석

Contents

목 차

데모 시연 !

감사합니다 !