Cloud Innovation 2023

에이블클라우드의 국산 HCI 솔루션 '에이블스택'
그리고 Eco System 솔루션 소개 세미나

ABLESTOR
Dynamic Value Creator

ABLECLOUD
All about data & cloud

데이터 통합, 실시간 빅데이터 플랫폼 TeraONE 이해

백민호 부장

DataStreams

Ⅰ. 데이터스트림즈 소개

Ⅱ. 빅데이터 플랫폼 TeraONE 이해

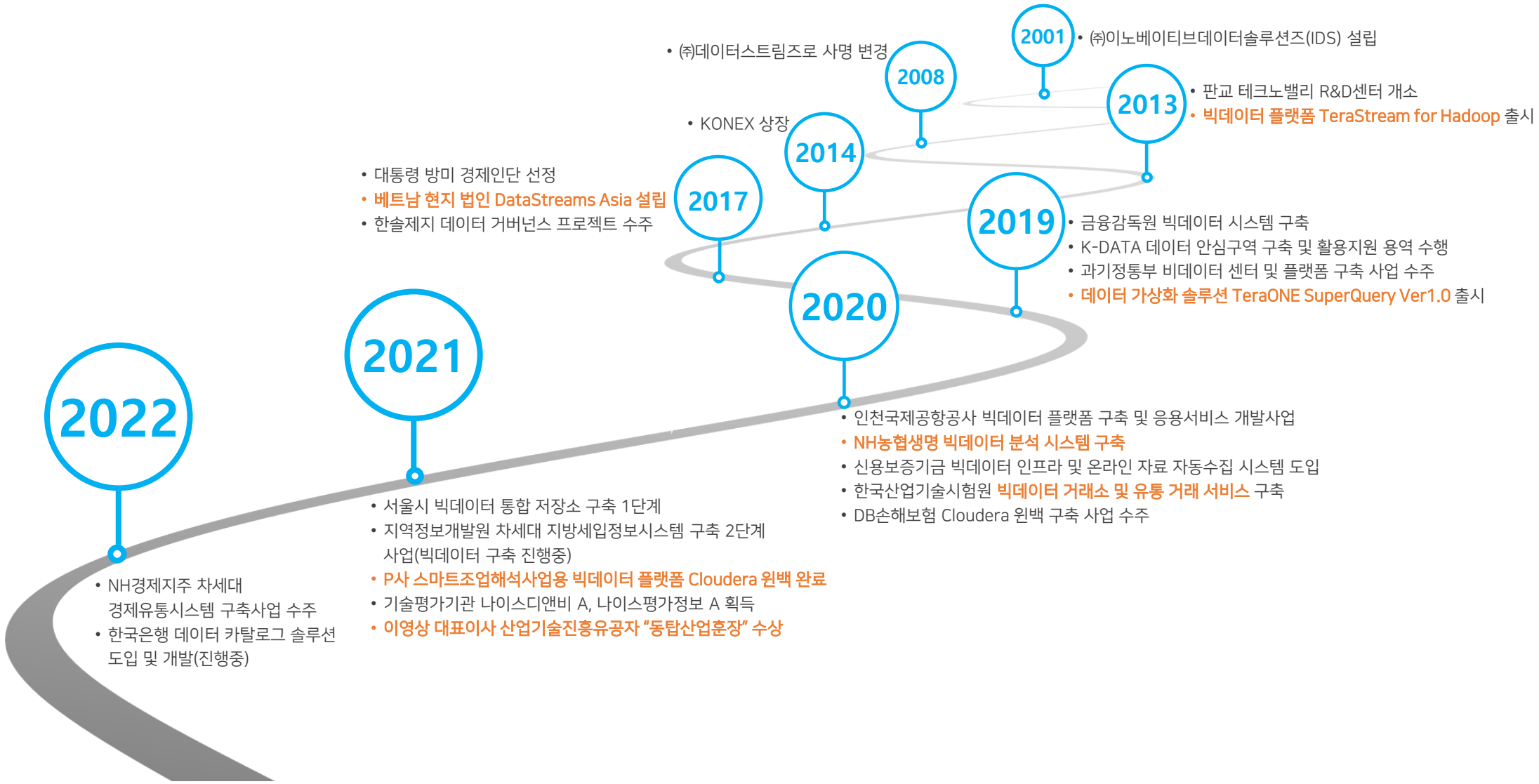
이 영 상 대표이사

- ㈜데이터스트림즈 대표이사
- 한국 PMO협회 명예회장
- 한국 소프트웨어전문기업협회 명예회장
- 대우전자/대우통신
- 한국과학기술원(KAIST) 전자공학 박사 수료
- 미국 미시간 주립대학교 전자공학 M.S.E.E.
- 경북대학교 전자공학과

- 2021년 이영상대표 산업기술진흥유공자 "동탑산업훈장" 수상 (2021. 11)
- 2020년 ICT 대한민국 Innovation Awards 특별상 수상 (한국지능형사물인터넷협회)
- 2019년 제6회 코리아빅데이터어워드 경영자부문 중기부 장관상 수상 (중소벤처기업부)
- 2017년 제4회 SW품질대상 TeraStream V4.4 우수상 수상 (한국정보통신기술협회)
- 2017년 전자정부지원사업 우수성과 기업 행정안전부 장관 표창 (행정안전부)
- 2015년 글로벌 상용 SW명품 대전 공공부문 발주자 협의회장상 (공공부문발주자협의회)
- 2014년 신SW상품대상 TeraStream™ V3.2 미래창조과학부 장관상 (미래창조과학부)
- 2014년 제1회 코리아 빅데이터 어워드 통계청장상 (통계청, 행정안전부)
- 2011년 제12회 소프트웨어산업인의 날 국무총리 표창 (지식경제부)

- 회사명 - ㈜데이터스트림즈
- 설립일 - 2001년 9월 19일
- 대표이사 - 이 영 상
- 자본금 - 21억 원
- 주요사업 - 데이터 통합, 빅데이터 플랫폼, 데이터 거버넌스 관리, 데이터 컨설팅, 데이터 구축 관련 용역, 데이터 서비스
- 임직원 수 - 194명 (2021년 9월 30일 현재)
- 소재지 - 본사: 서울시 서초구 사임당로28 청호나이스빌딩 2층, 6층
  - 연구소: 경기도 성남시 분당구 대왕판교로 670 유스페이스몰 6층
- 홈페이지 - http://datastreams.co.kr/kor/

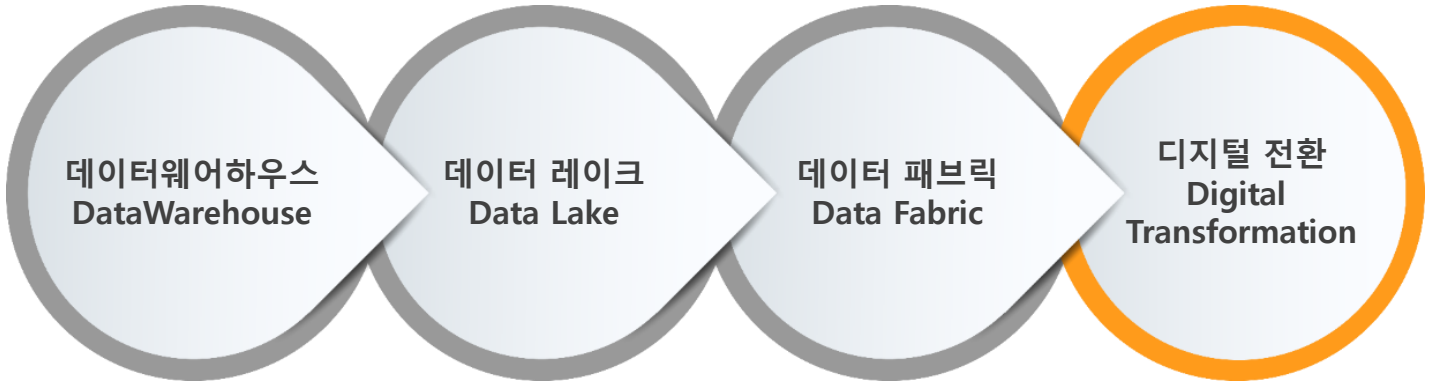
㈜데이터스트림즈는 빅데이터 패브릭 시대를 선도하며 늘 더 나은 미래와 고객의 성공을 목표로 나아갑니다.

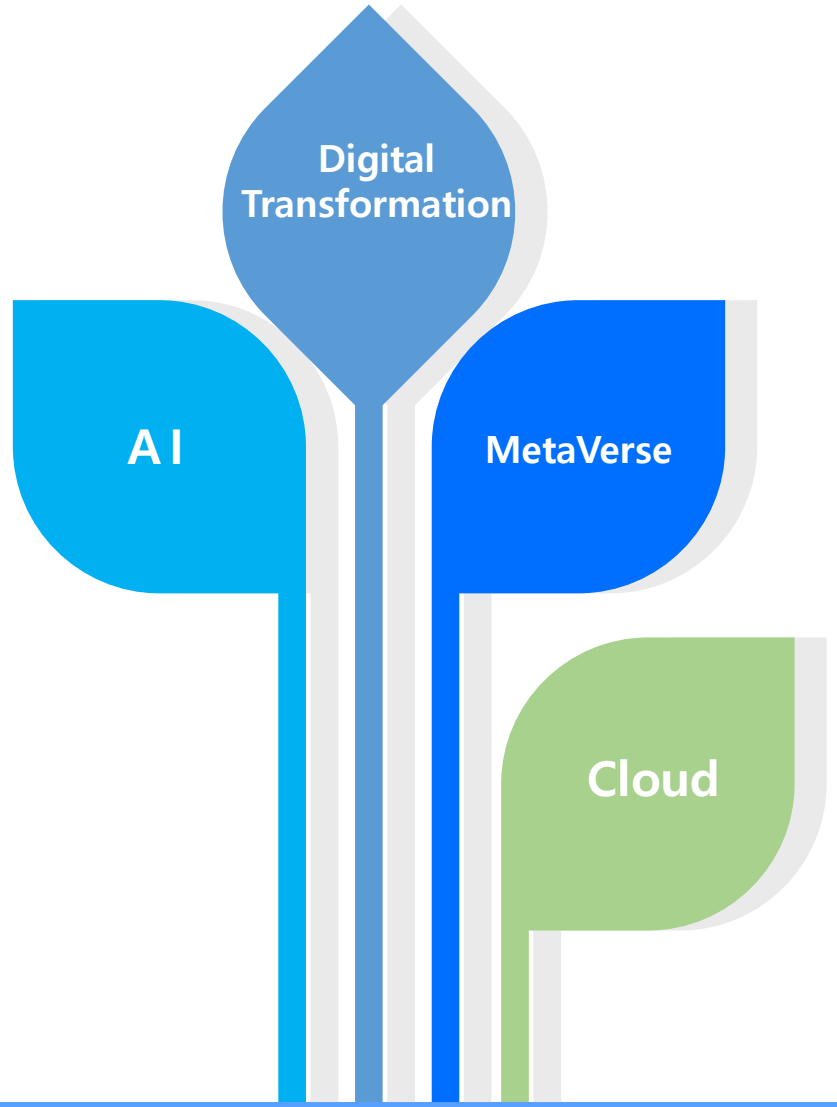


디지털 혁신을 이끌 빅데이터를 통해,

미래의 서비스를 가능하게 하는

데이터스트림즈





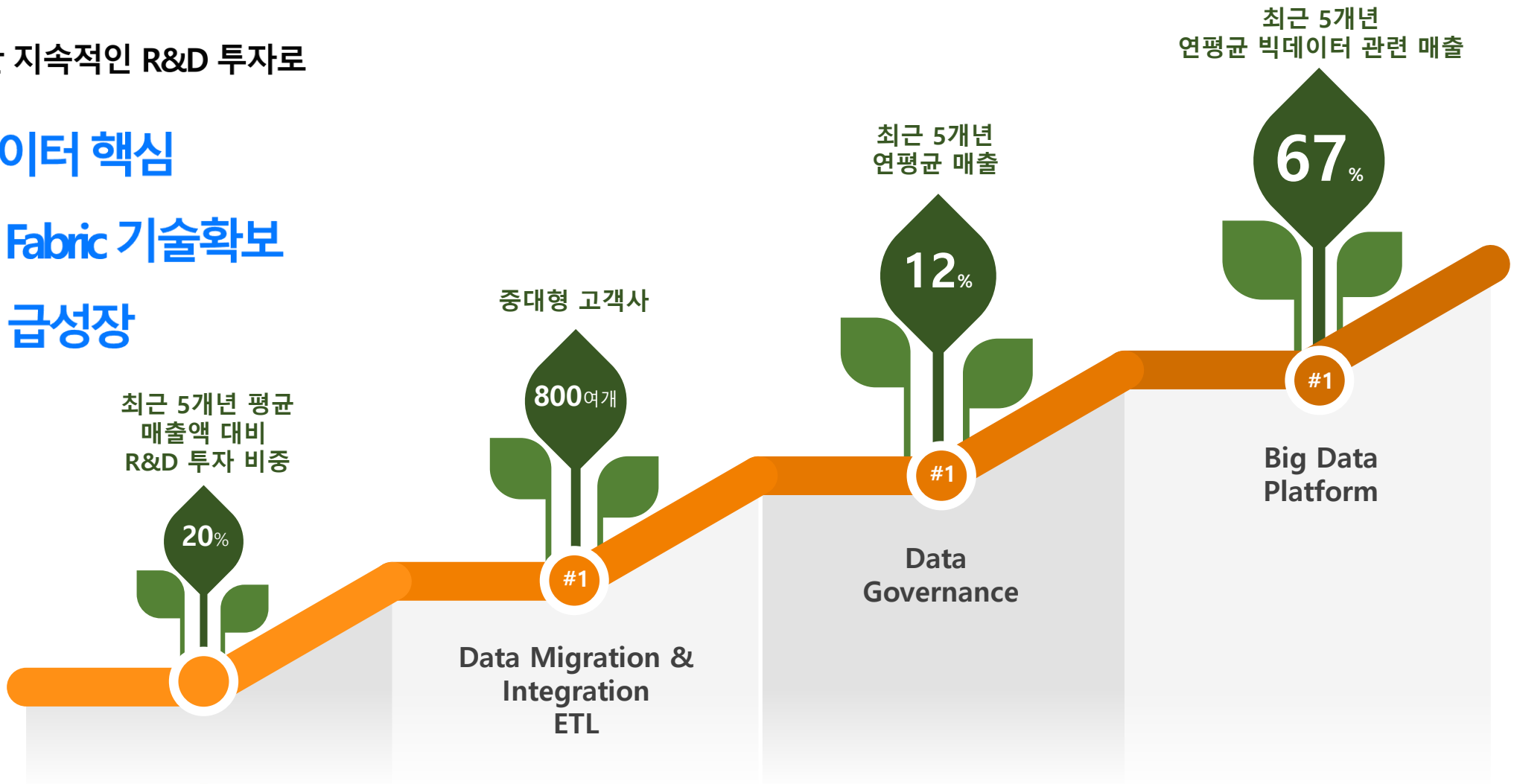
글로벌 시장에서도 찾기 힘든 22년 기술의 빅데이터 토탈 솔루션 원천기술 보유기업입니다.

22年간 지속적인 R&D 투자로

빅데이터 핵심

Data Fabric 기술확보

매출 급성장



빅데이터 및 사물인터넷 분야를 포함하여 데이터 관리 전 영역에서의 통합, 분석, 거버넌스에 대한 SW 제품과 컨설팅, 그리고 핀테크, 스마트 리테일과 같은 ICT 산업융합형 서비스를 위한 데이터 플랫폼을 제공하고 있습니다.

컨설팅 서비스

- Data Integration 컨설팅
- Data Governance 컨설팅
- Data Warehousing 컨설팅
- Big Data 및 차세대 ISP 수립
- 데이터 기반 신규 사업모델 컨설팅

데이터 웨어하우스, 고급분석

- DW,CRM,BI 통계분석 시스템
- 부정사용방지시스템 (FDS)
- 탈세/부정 사용 적발 및 대응 시스템
- 거래매매 데이터 분석 서비스
- 금융 거래 분석 서비스

데이터 거버넌스

- 데이터 거버넌스 아키텍쳐
- 데이터 거버넌스 어플라이언스
- 마스터데이터 관리 (MDM)
- 데이터 표준화 및 품질 관리

Big Data Platform

- Hadoop 기반 Big Data 플랫폼
- 초고속 Big Data 및 In-Memory 기반 실시간 분석
- IoT 데이터 수집 / 통합 / 분석
- DW, IoT, Omni-Channel Marketing Fintech 데이터 통합 기반

클라우드 데이터 통합 서비스

- 빅데이터 저장
- 클라우드 기반의 양방향 통합
- 실시간 데이터 통합 (CDC)
- 가독형 데이터 암호화
- 분산파일 및 GRID 기반 통합

데이터 통합 및 데이터 품질 전 범위를 아우르는 기술력을 보유한 국내 최고의 데이터 거버넌스 솔루션 보유 전문 기업이며, 신기술을 기반으로 하는 Big Data 등 신성장 산업 시장을 선도하고 있습니다.

데이터 수집 · 통합제품군
TeraStream™

- 고속 데이터 통합 솔루션 개발 · 공급
- ETL 처리 과정 및 실시간 모니터링
- 빅데이터 준비 도구

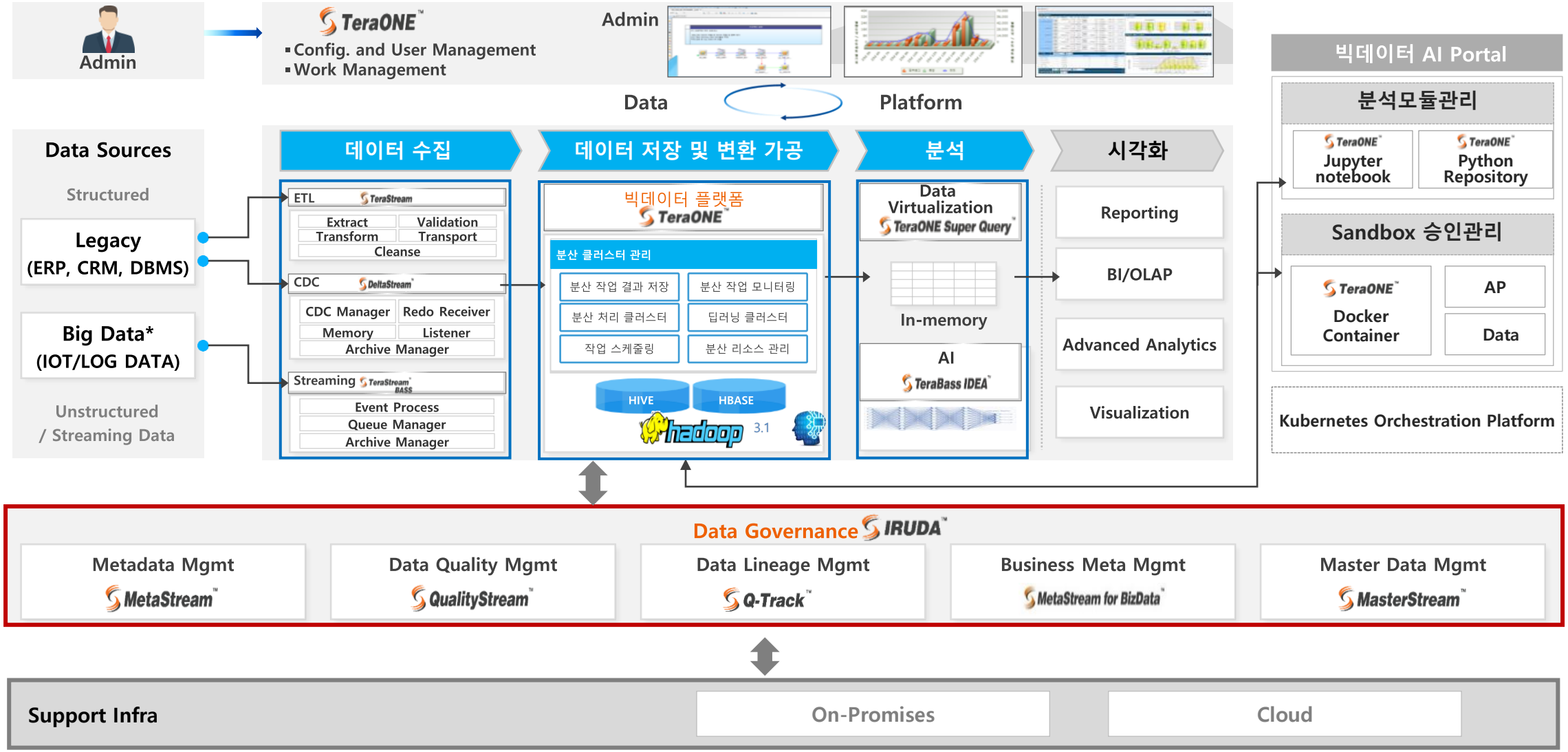
차세대 빅데이터 플랫폼 제품군
TeraONE™

- 빅데이터 패브릭 기능을 제공하는 차세대 플랫폼 개발
- 데이터 스트리밍과 수집 / 가공 / 저장 / 분석 시스템 도입
- VM 및 AI 기술 확장하여 차별화된 서비스 제공

데이터 거버넌스 제품군
IRUDA™

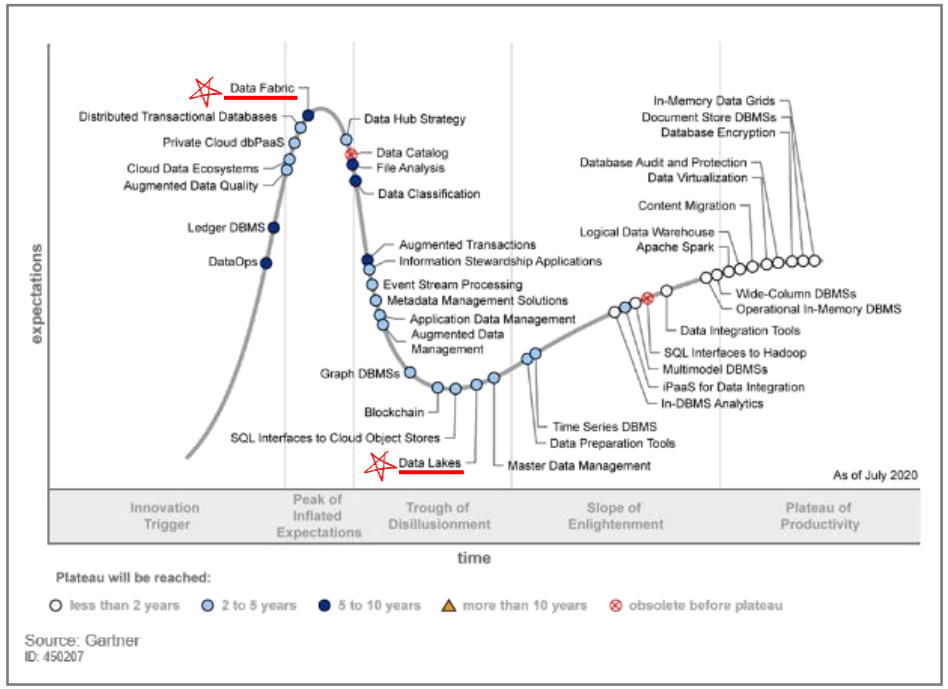
- 메타데이터 관리 및 품질관리 등 포털 거버넌스 제품 보유
- 데이터 통합 검색, 필터링 관리 솔루션

빅데이터 플랫폼과 데이터 거버넌스 플랫폼을 위한 총 15개 제품으로 구성



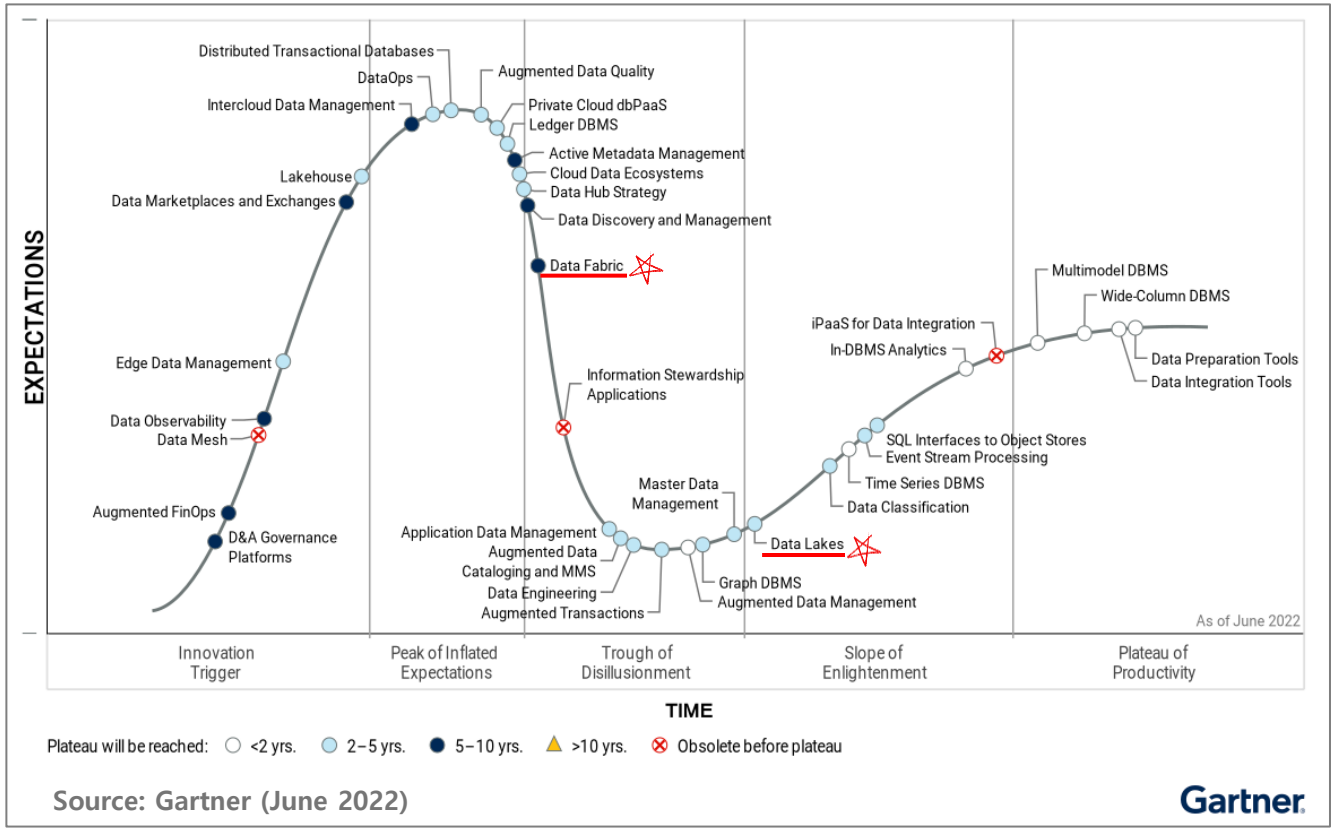
디지털 비즈니스로의 전환이 가속화되면서, 현업 주도의 지능형 데이터 관리 요구가 증가하고 있습니다. 결국 조직 내 분산되어 존재하는 다양한 데이터를 필요 시점에 빠르게 찾아 활용성을 제고하고 비즈니스 가치 창출을 강화하기 위한 새로운 데이터 플랫폼 전략이 확산 되고 있습니다.

Hype Cycle for Data Management, 2020



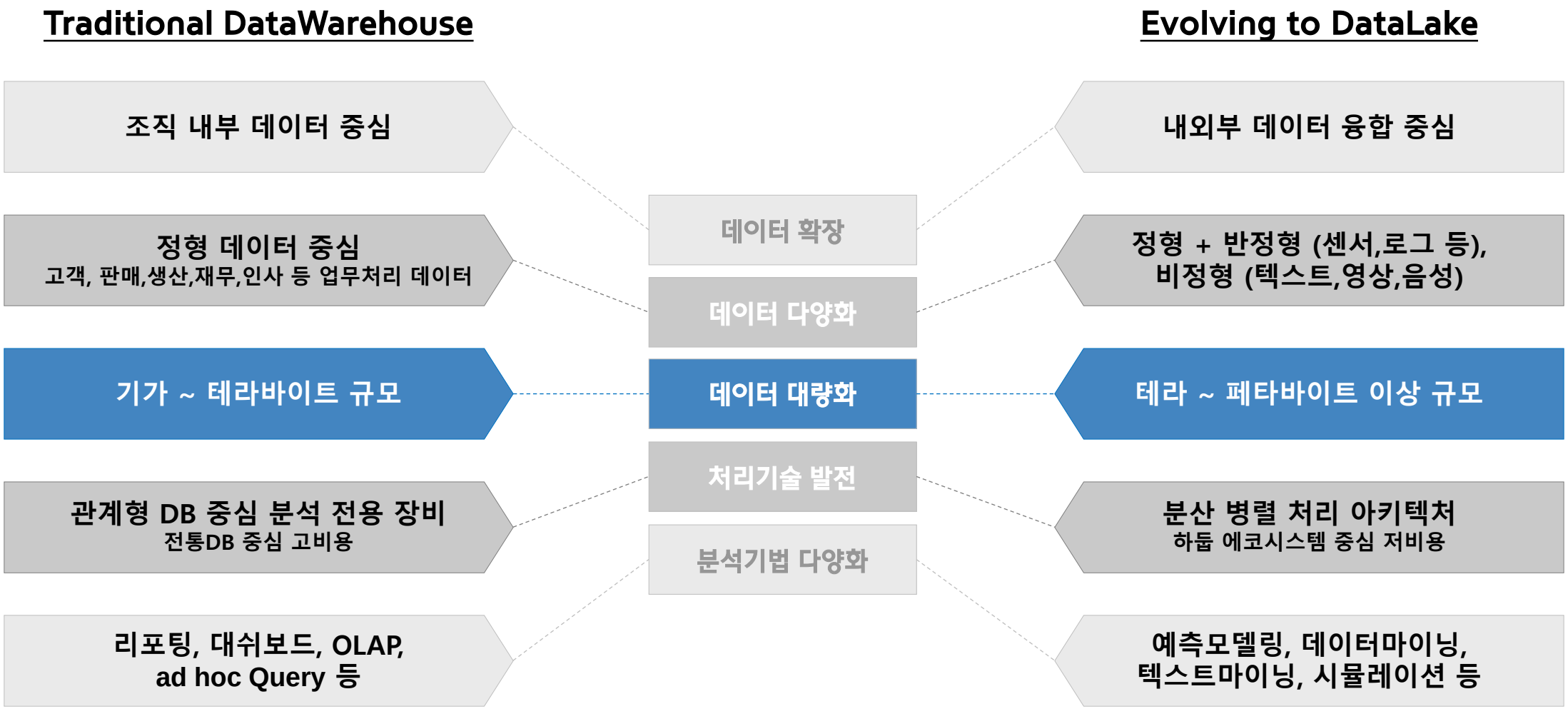
Source: Gartner
ID: 450207

Hype Cycle for Data Management, 2022

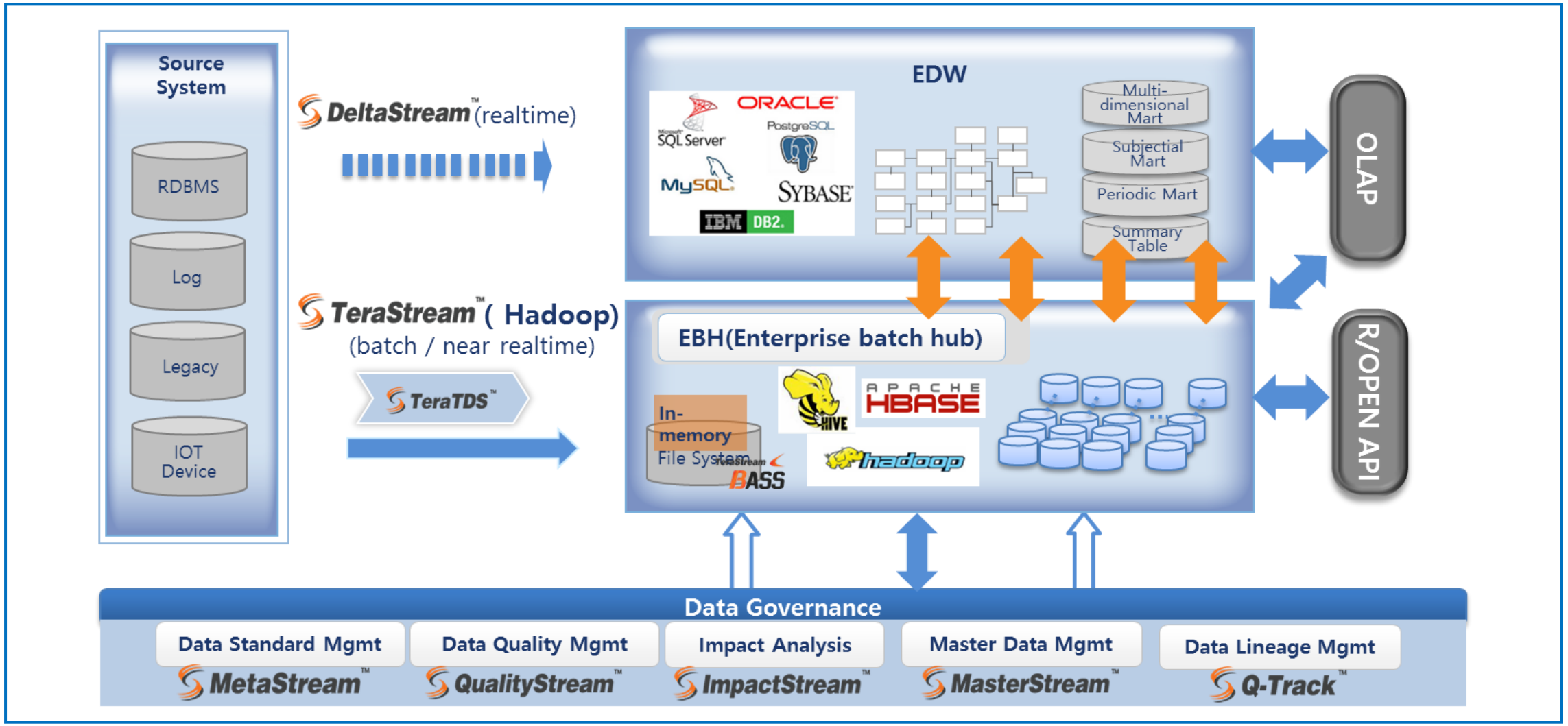


Source: Gartner (June 2022)

하둡 기반의 빅데이터 플랫폼인 데이터레이크는 저렴한 비용으로 모든 유형의 데이터를 빠르게 처리할 수 있는 유연한 분산 병렬 아키텍처를 제공함으로써, 고비용 데이터 통합 아키텍처인 DW구성을 효율화할 수 있습니다.



빅데이터와 RDBMS를 동시 처리할 수 있는 플랫폼 아키텍처 구현으로 전통적인 SQL 및 ETL, CDC , OLAP, 거버넌스 기술을 활용하여 운영 환경 구현합니다.



빅데이터 플랫폼 TeraONE은 Apache Hadoop 오픈소스를 최적화하여 안정화시킨 아키텍처를 구성하여 제공하며, 가상화 기술 및 사용자 분석 포털을 통해 분석플랫폼을 구축을 지원합니다.

자사 제품 | 타사제품 | 오픈 소스 | 추가옵션

기본모듈

TeraONE Admin

- Resource Monitoring Ambari
- Hadoop Data Browser
- User Access Control Ranger
- ETL Job Monitoring
- Cluster Manager Ambari

- Batch/Bulk DataExtraction FACT
- 비정형 데이터 수집 Flume / Logstash
- In-Memory SPARK
- AI Portal Supervisual Portal (옵션)

- Transform TeraStream for Hadoop Engine
- Scheduling TeraStream Designer
- Data Virtualization TeraONE SuperQuery (옵션)

- NoSQL
- SQL On Hadoop HIVE + Tez
- Jupyter Notebook + PYTHON

MetaData Management (Hcatalog)

Kerberos (Network Authentication Protocol)

Storage : HDFS(Hadoop Distributed File System)

추가 옵션(타사제품 포함)

- IoT TeraStream BASS
- SNS/Web Crawler
- BI / Visualization
- ELK Stack
- MetaData MetaStream
- DQ QualityStream
- Linage Q-TRACK
- CDC DeltaStream

특장점

- 고객들이 가장 많이 사용하는 오픈소스 중심으로 자사제품과 패킹한 최적 아키텍처 제공
- 국내 최고의 ETL인 TeraStream이 데이터 수집 허브로 내장되어 내외부의 모든 유형의 데이터 (정형/반정형/비정형)를 HDFS, Hive, Kudu, HBASE , RDB 등에 유연하게 적재 가능합니다.

탑재된 주요 오프소스

Hadoop / Mapreduce : 3.1.1
Ambari HDP 2.6.5
Spark2 : 2.3
Hive : 3.0
Apache Impala 3.2.0
Apache Kudu 1.10.0
Tez 0.92
Flume : 1.5.2
Hbase : 1.1.2
Zeppelin : 0.7.3
Python : 3.6.x

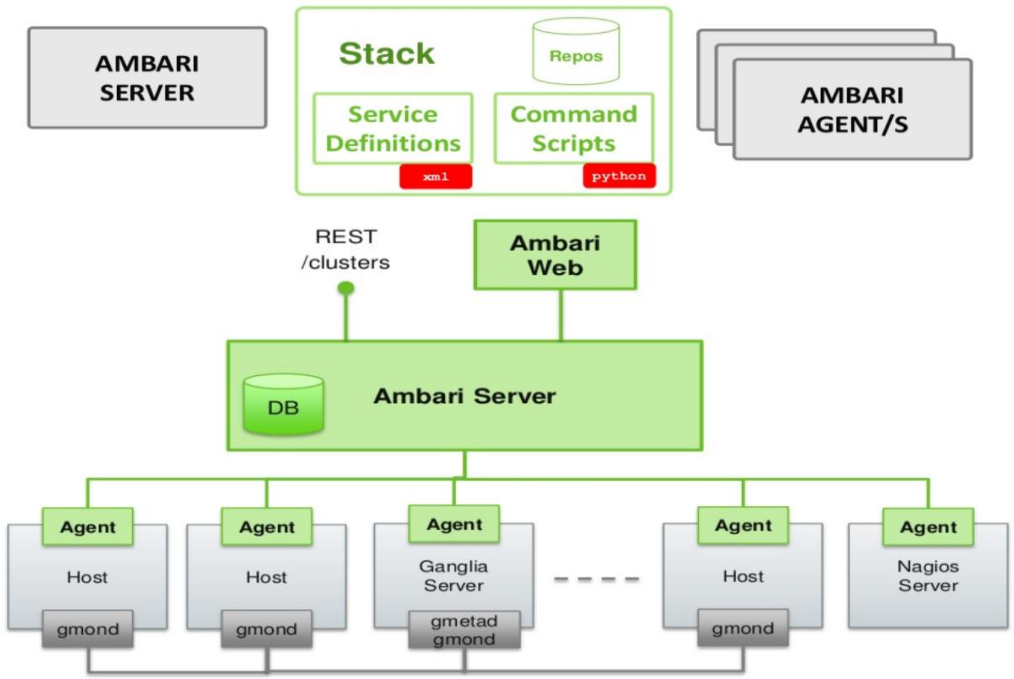
데이터 통합 원천 기술을 기반으로 고객들이 많이 사용하는 오픈소스를 자사 제품으로 유지보수 가능하도록 자체 패키징하였으며, 전담조직이 효율적이고 신속한 기술지원을 제공합니다.

구분	구성 항목	기능 설명	기본 설치	비고
데이터 수집	▪ TeraStream for Hadoop	▪ 대용량 데이터 고속 추출, RDBMS 및 Hadoop 동시 적재, 워크플로우 스케쥴러	O	
	▪ DeltaStream	▪ 실시간 변동 데이터 추출	X	별매
	▪ TeraStream for BASS	▪ IoT 센서데이터 등 스트리밍 데이터 추출,적재, 분석,시각화	X	별매
	▪ Kafka	▪ 분산환경에 특화된 메시지 큐	O	
데이터 저장/처리	▪ Hadoop (HDFS, YARN, MR)	▪ 대량의 자료를 처리할 수 있는 분산 소프트웨어 프레임워크	O	
	▪ PostgreSQL	▪ 플랫폼레파지토리 (SuperQuery, MetaStream 리포지토리 공유 가능)	O	
	▪ Hbase	▪ 대용량 NoSQL 데이터 베이스	X	옵션
	▪ Hive	▪ 하둡 내 데이터 웨어하우스 구성, 데이터 요약, 질의 및 분석	O	
	▪ Beeline	▪ SQL Line 기반의 하이브서버2(hiveserver2) 내 쿼리 실행 도구	O	
	▪ Spark	▪ 대용량 데이터 분석 처리 엔진으로 실시간 프로세싱 지원	O	
	▪ Tez	▪ MapReduce(MR)를 대체하여 인메모리를 통해 처리 성능을 높이는 엔진	O	
데이터 분석 시각화	▪ Jupyter Notebook (Python)	▪ Machine Learning/Deep Learning 구현 가능한 파이썬 기반 분석 환경	X	별매
	▪ R	▪ 고급 분석 환경 (라이선스 이슈로 고객사 주도 사용, 설치 지원만 가능)	X	라이선스 이슈
	▪ ELK	▪ ElasticSearch(검색)/LogStash(로그데이터수집)/Kibana(시각화) 스택	X	별매
	▪ Superset	▪ 데이터 시각화	X	별매
	▪ Supervisual Portal	▪ 빅데이터 사용자 포털	X	별매
운영 관리	▪ Ambari	▪ Hadoop Cluster 구성, 관리, 모니터링	O	
	▪ Kerberos/Ranger	▪ 보안/인증/접근관리, LDAP과 연계한 관리 가능	O	
	▪ Zookeeper	▪ 분산처리 환경 코디네이터	O	
	▪ Oozie	▪ Hadoop job 워크플로우 스케줄러	X	옵션

국내 경쟁사 대부분은 Hortonworks 제공 HDP를 패키징하여 사업을 수행하고 있습니다. 당사는 순수 오픈소스인 Apache Hadoop 배포판을 자체 패키징하므로 오픈소스 라이선스 이슈에서 자유롭습니다.

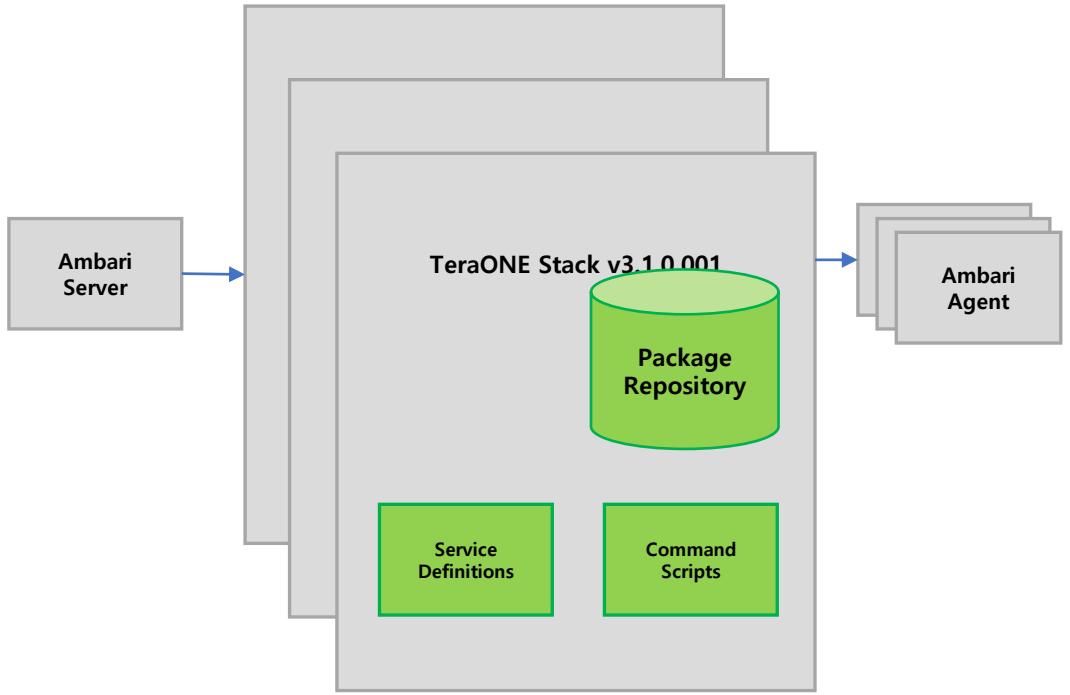
암바리 패키징 구조

- 패키지들을 내부 혹은 원격 공용 Repository 에 저장
- 컴포넌트를 설치할 때 Service Definition으로 Ambari의 Interface를 확정하고, Command Script로 설치 및 모니터링, 제어 절차 결정



테라원 암바리 패키징 구조

- Apache Ambari는 항상 검증된 최신 소스 적용
- 내부 Stack을 독자적으로 정의하여 패키징
  - 최신성 유지 가능하며, 로드맵은 6개월 단위 검증하여 수립



DataStream Data Integration Platform

Scalability	▪ 병렬 분산 저장 방식을 통한 데이터 확장성 제공
Usability	▪ 사용 용이성을 통한 개발의 편의성 제공
Performance	▪ Local System, Hadoop 및 DBMS를 통한 최적의 성능 제공
Easy	▪ 최소 비용을 통하여 데이터 가공 및 분석을 편리하게 제공
Reliability	▪ Data Governance를 통한 데이터 신뢰성 확보

Scalability, Speed, Reliability and Management at Low TCO

데이터 사이언티스들의 데이터 활용성 강화를 통한 비즈니스 가치 창출을 위해 다음과 같은 추가적인 기능들을 요구하고 있음. 하둡 기반의 데이터레이크에 모든 유형의 데이터를 물리적으로 이동하여 통합해오던 기존 데이터레이크 중심의 벤더들은 미흡한 데이터 가버넌스와 가상화 기술에 대한 도전을 받고 있습니다.

데이터 거버넌스 기능 강화

- 하이브 기반 데이터 표준/품질 관리
- 데이터 활용 Audit Log 관리
- 데이터 흐름 관리

머신러닝/딥러닝 워크벤치 확장

- 자체적인 워크벤치 제공 강화
- 상용 솔루션과의 유기적인 연계 가능
- 데이터 레이크 분석 성능 고도화

메타데이터 기반으로 이기종 데이터 소스를 탐색하여 메모리 캐쉬에서 가상 통합/분석 기능 제공

분석 워크벤치와 가상화 기술을 연계하여 컨테이너 기반으로 개인 맞춤형 최적 분석 인프라 제공

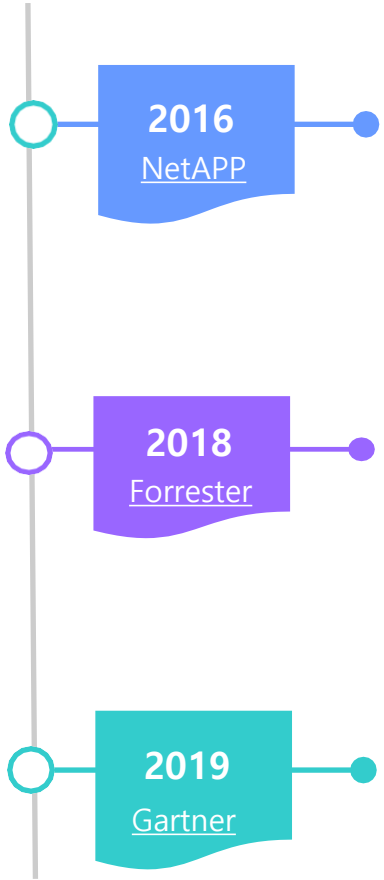
AI 데이터 카탈로그 서비스

- 데이터 거버넌스 기반 스마트 검색기능
- 활성 메타데이터 기반 데이터 학습
- 분석가를 위한 데이터 맵 및 추천 서비스

컨테이너 기반 맞춤형 인프라

- 쿠버네티스 기반 컨테이너 서비스
- 개인별 맞춤형 분석 환경 제공
- 프로비저닝과 오토 스케일링 지원

데이터 패브릭은 거버넌스 기반으로 데이터 활용성을 제고하고 비즈니스 가치 창출을 강화하기 위한 데이터 플랫폼의 새로운 디자인 컨셉이자 아키텍처 전략입니다.



On-premise와 다양한 Cloud상의 데이터 셋을 하나의 관리체계 하에서 일관성 있게 처리할 수 있는 데이터 아키텍처 개념에서 확대

(2016, NetApp Data Fabric Architecture Fundamentals)

빅데이터 패브릭을 Hadoop, EDW, 이기종 DBMS 상의 모든 유형의 데이터를 일관된 관리체계 하에서 프로세싱할 수 있는 유연한 아키텍처로 정의

Exdata, SybaseIQ, Vertica..

Oracle, SQLServer, PostgreSQL..

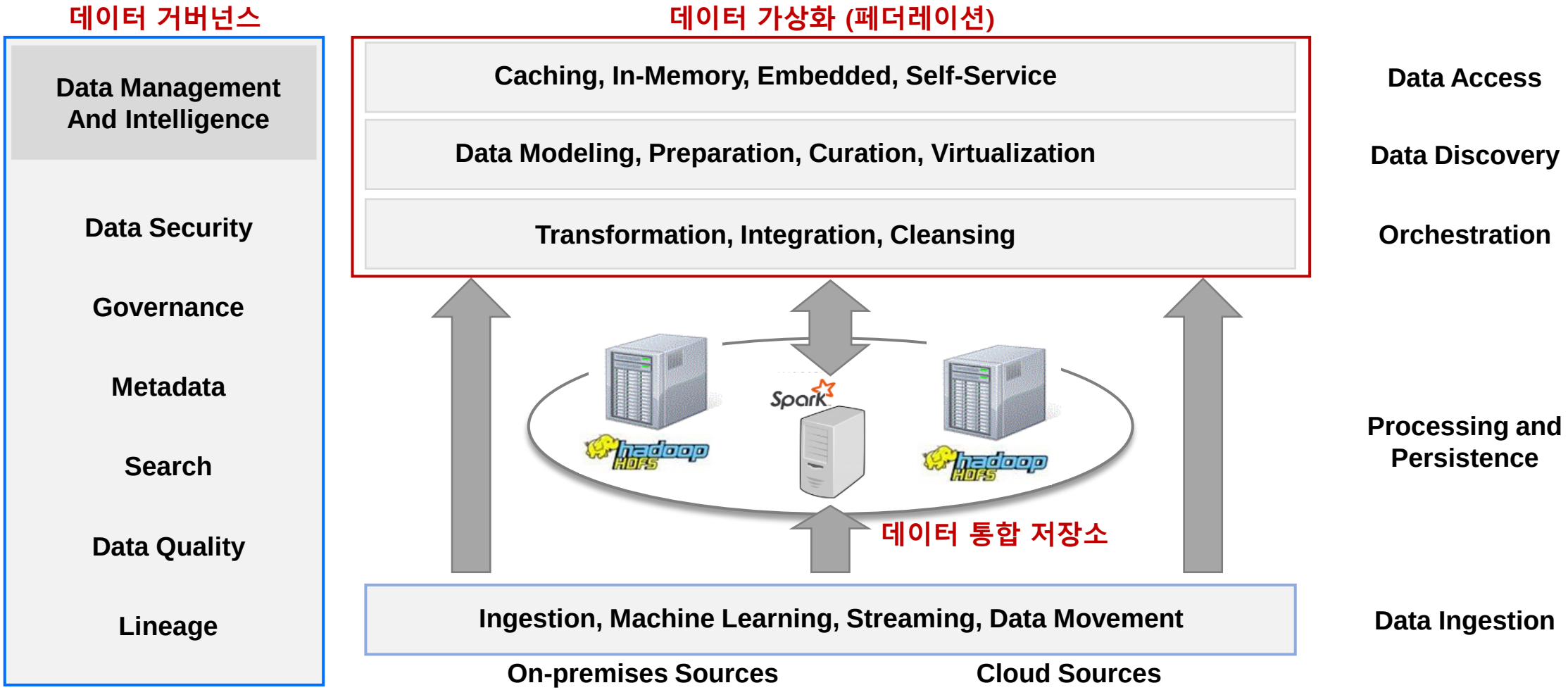
(2018.6, The Forrester Wave™ : Big Data Fabric, Q2 2018)

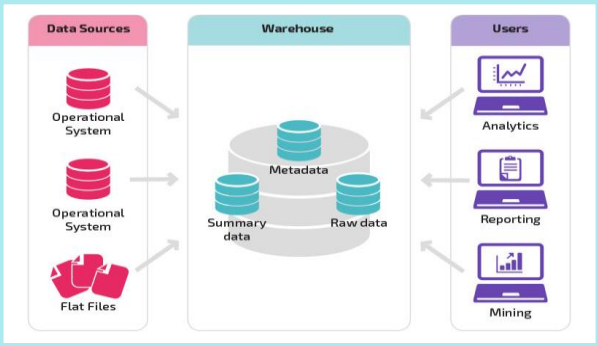
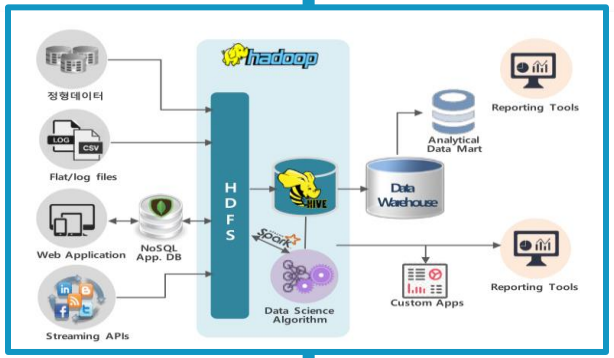
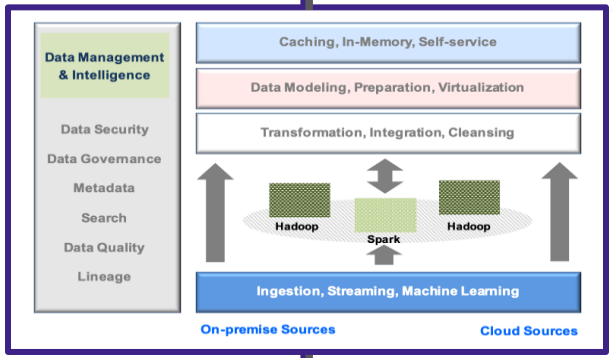
데이터 패브릭을 다양한 데이터 플랫폼에서 제공되는 데이터 관리 및 통합 기술 을 수용할 수 있는 아키텍처 설계 및 서비스의 조합을 최적화한 개념으로 정의

(2019.12, Gartner : Data Fabrics Add Augmented Intelligence to Modernize Your Data Integration)

빅데이터 시장이 성숙함에 따라 데이터의 품질 관리가 중요한 이슈가 되었으며, 일관된 데이터 관리체계 하에 양질의 데이터를 이기종 데이터 소스에 상관없이 분석 및 활용할 수 있는 빅데이터 패브릭이 각광을 받고 있으며, 데이터 거버넌스, 데이터 통합 저장소, 데이터 가상화가 핵심 구성 요소입니다.

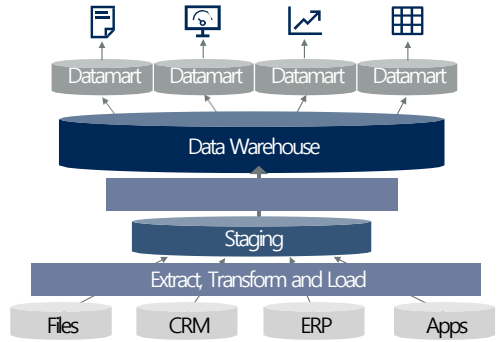
[Forrester Research의 Big Data Fabric Reference Architecture]



	데이터웨어하우스	데이터 레이크	빅데이터 패브릭
참조 아키텍처			
주요특징	• 고가의 Unix 서버 + RDBMS • 내부 정형 데이터 중심 • Schema on Write • 테라바이트 규모 • 리포트, 대쉬보드, OLAP 분석	• 저가의 Linux 서버 + Hadoop/NoSQL • 내외부 정형/반정형/비정형 데이터 • Schema on Read • 페타바이트 규모 이상 • 머신러닝, 딥러닝, 시각화	• 데이터웨어하우스와 데이터레이크 포괄 • 빅데이터 패브릭 구성요소 - 데이터 거버넌스 - 데이터 웨어하우스/데이터 레이크 - 데이터 가상화
제약사항	• 정형 데이터 중심의 고비용 아키텍처가 오픈소스 벤더들의 마케팅 타겟이 됨 • 센서/텍스트/이미지 등 빅데이터 요건 수용 어려움	• 데이터원본을 HDFS에 저장 후 처리하여 RDB 데이터 처리가 복잡함 • 수집된 빅데이터의 잠재적 가치와 영구 보관에 따른 낭비 사이 균형점 도출 필요	• 이기종 데이터를 하나의 물리적 저장소에 모두 통합하지 않고 가상 통합 • 빅데이터를 통해 비즈니스 가치창출을 위해 데이터 거버넌스가 반드시 필요

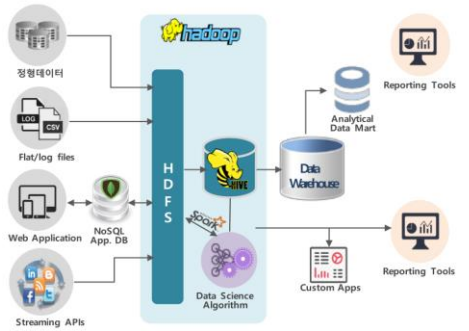
전통적인 DW

- 고가의 어플라이언스 장비 / Unix / RDBMS (Oracle Exadata, Sybase-IQ, Vertica…)
- 내부 정형데이터 중심으로 사전 데이터 모델 정의
- 테라바이트 규모
- 비정형데이터 및 대용량 데이터 처리에 한계



하둡기반 데이터레이크

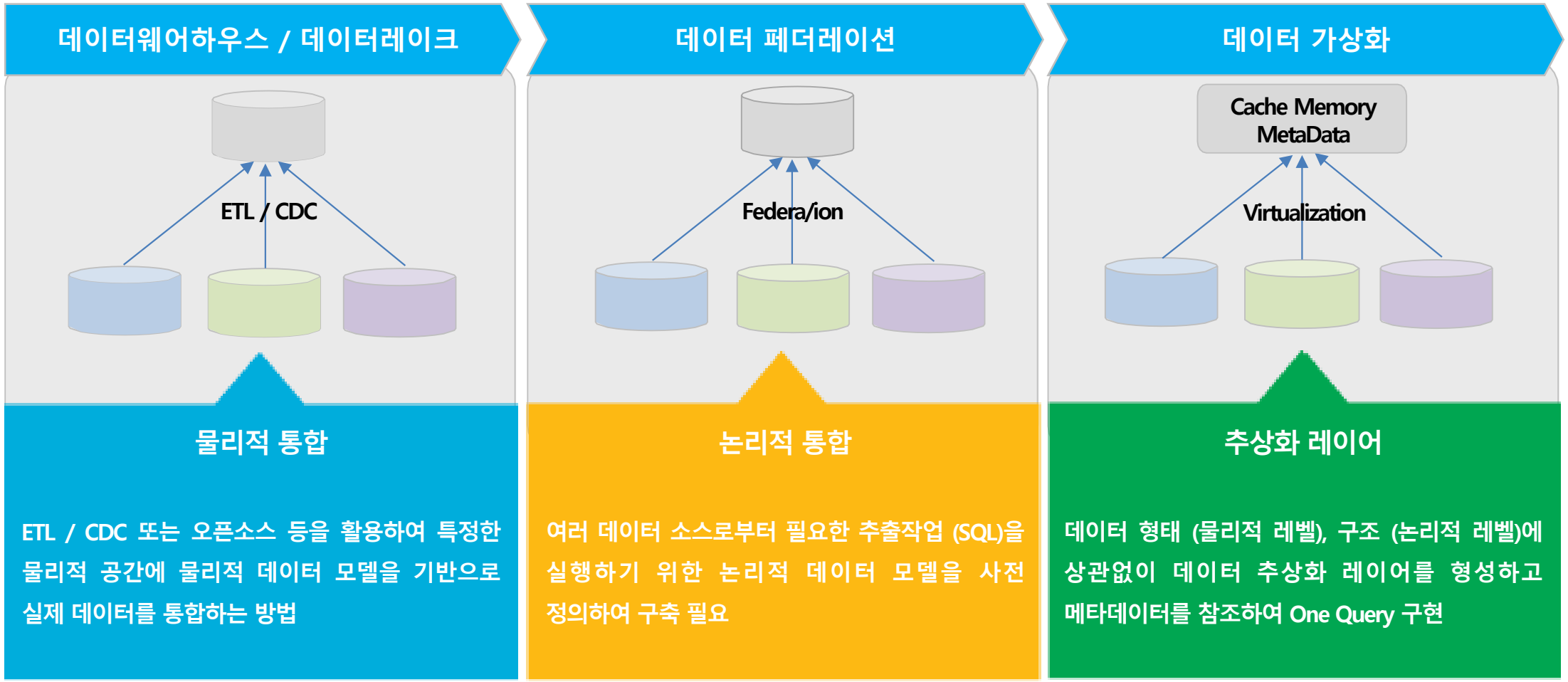
- 저가의 리눅스 장비 / 오픈소스 하둡기반 파일 분산 병렬 처리
- 모든 유형의 데이터를 사전 모델 정의 없이 적재
- 페타바이트 규모
- 오픈소스 유지보수 복잡성으로 비용 증가



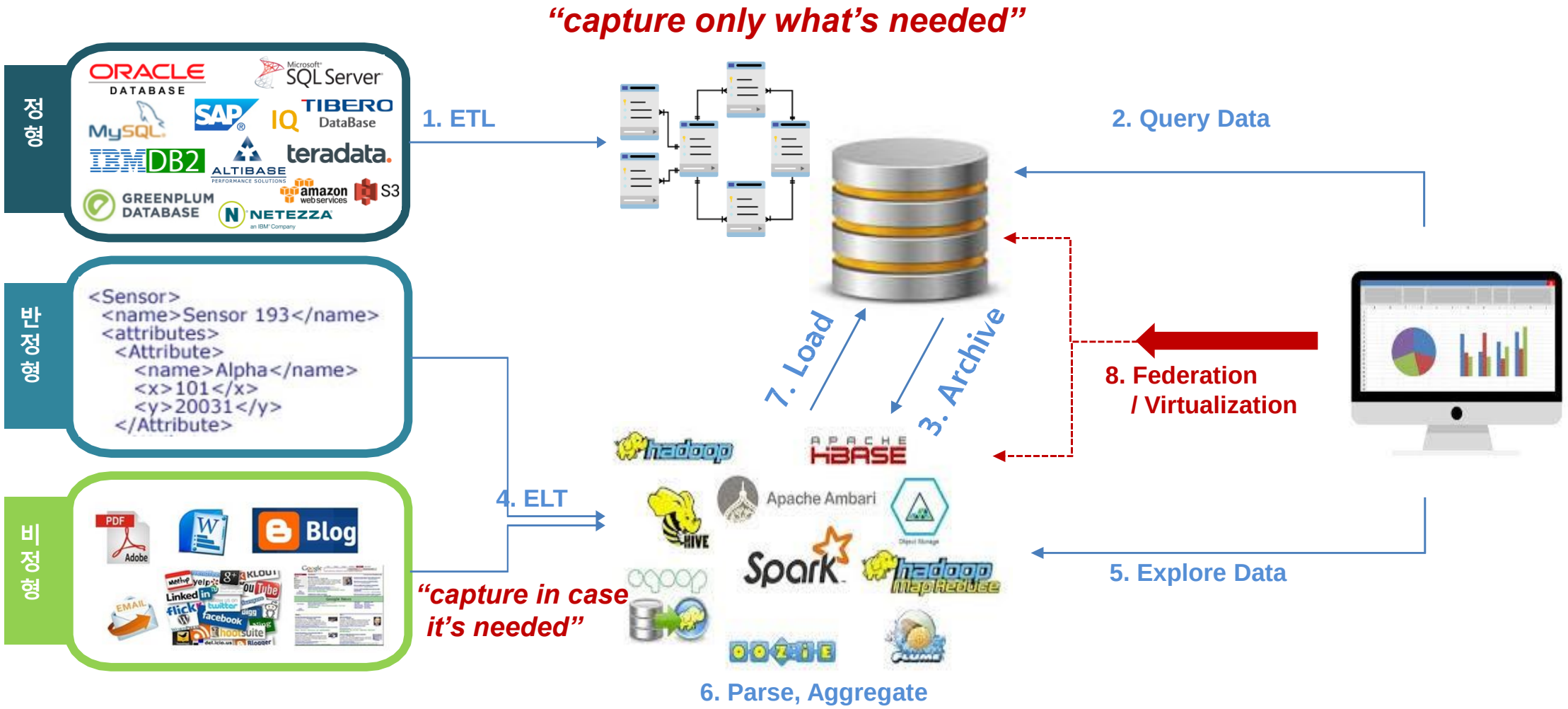
하이브리드 DW

- 기존 DW와 빅데이터 플랫폼 장점 중심으로 최적화 구성
- DW는 정형데이터 저장용 허브 또는 데이터마트용으로 가볍게 구성
- 대용량 데이터, 비정형데이터는 레이크 중심으로 구성
- 데이터 가상화를 통해 이기종 데이터소스의 물리적 적재없이 분석

데이터 통합 기술은 궁극적으로는 논리적 모델없이 추상화 레이어를 통해 이기종 데이터 소스에 관계없이 통합하는 데이터 가상화를 지향하고 있으며, 물리적인 데이터 이동 및 데이터 마트 생성을 최소화할 수 있음.

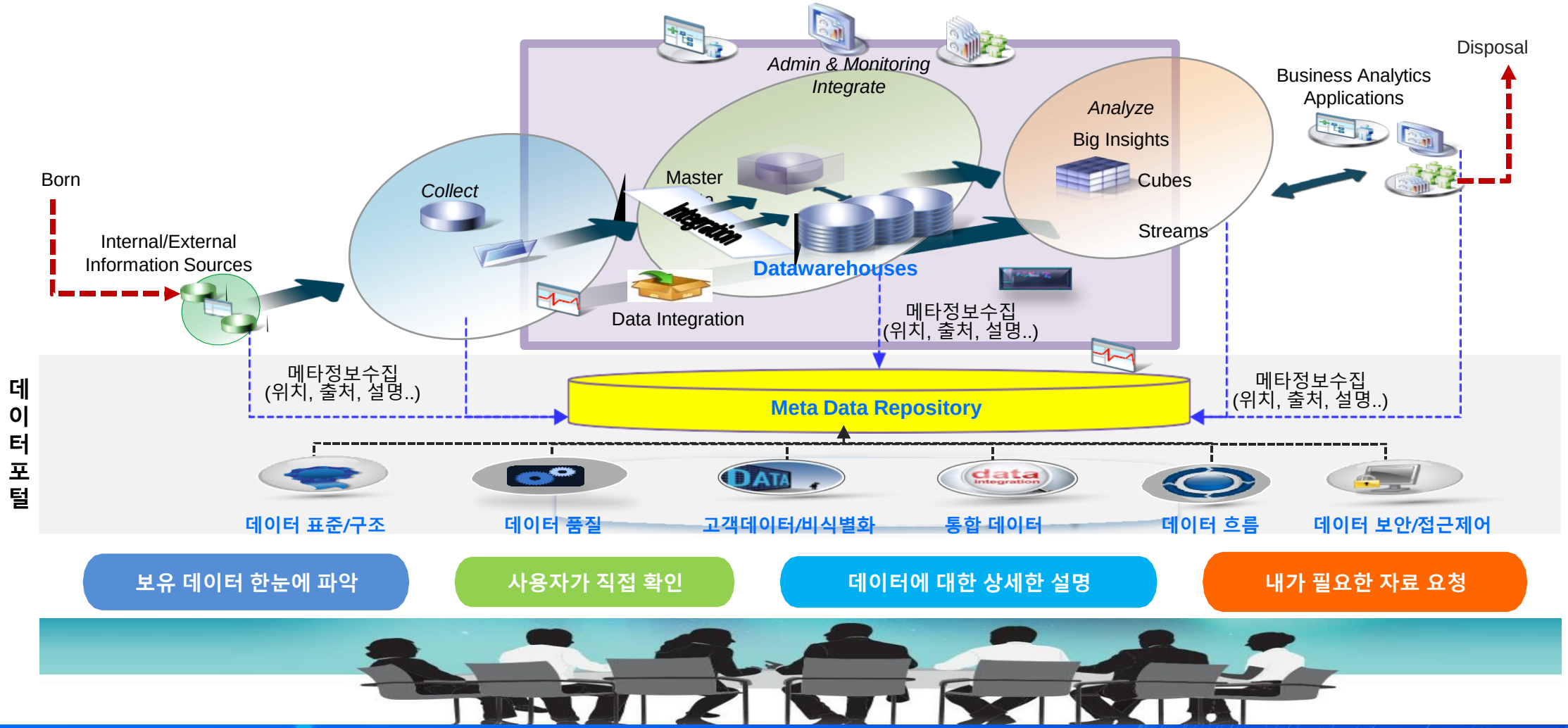


하둡기반의 데이터레이크를 구축한 고객사들은 최근 기존 데이터웨어하우스의 가치를 유지하면서 효율적인 빅데이터 플랫폼 구축을 고민하고 있습니다. 관계형 데이터베이스와 분산 파일시스템의 강점을 살린 최적화 방안을 요구합니다. (RDB, Hadoop에 상관없이 하나의 Query로 데이터를 조회/분석하는 기능 요구)

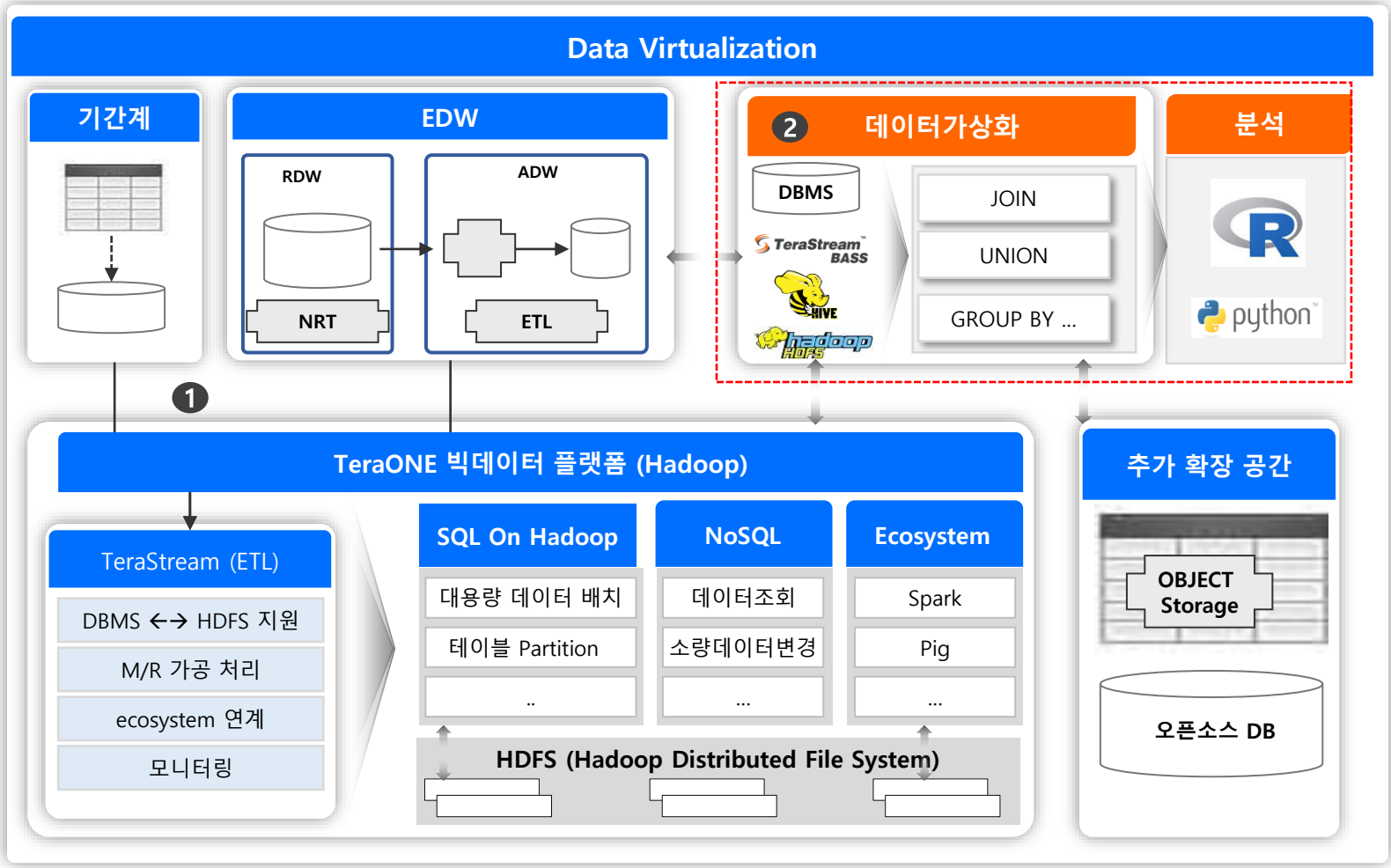


빅데이터 도입이 확대되면서 분석가들이 활용할 데이터가 부족하고, 각 부문에서 별도 관리하거나 데이터에 대한 총괄 오너십이 부족한 현상이 많이 나타나면서, 빅데이터를 제대로 관리하기 위한 거버넌스 요구가 증대되고 있습니다.

데이터 거버넌스 기반 데이터 통합 역량 요구 (보이는 데이터 거버넌스)



데이터 가상화 기능을 통해 이기종의 DBMS 및 Hadoop의 데이터를 사용자의 ETL 개발 과정 없이도 하나의 쿼리로 분석 가능한 환경을 제공합니다.

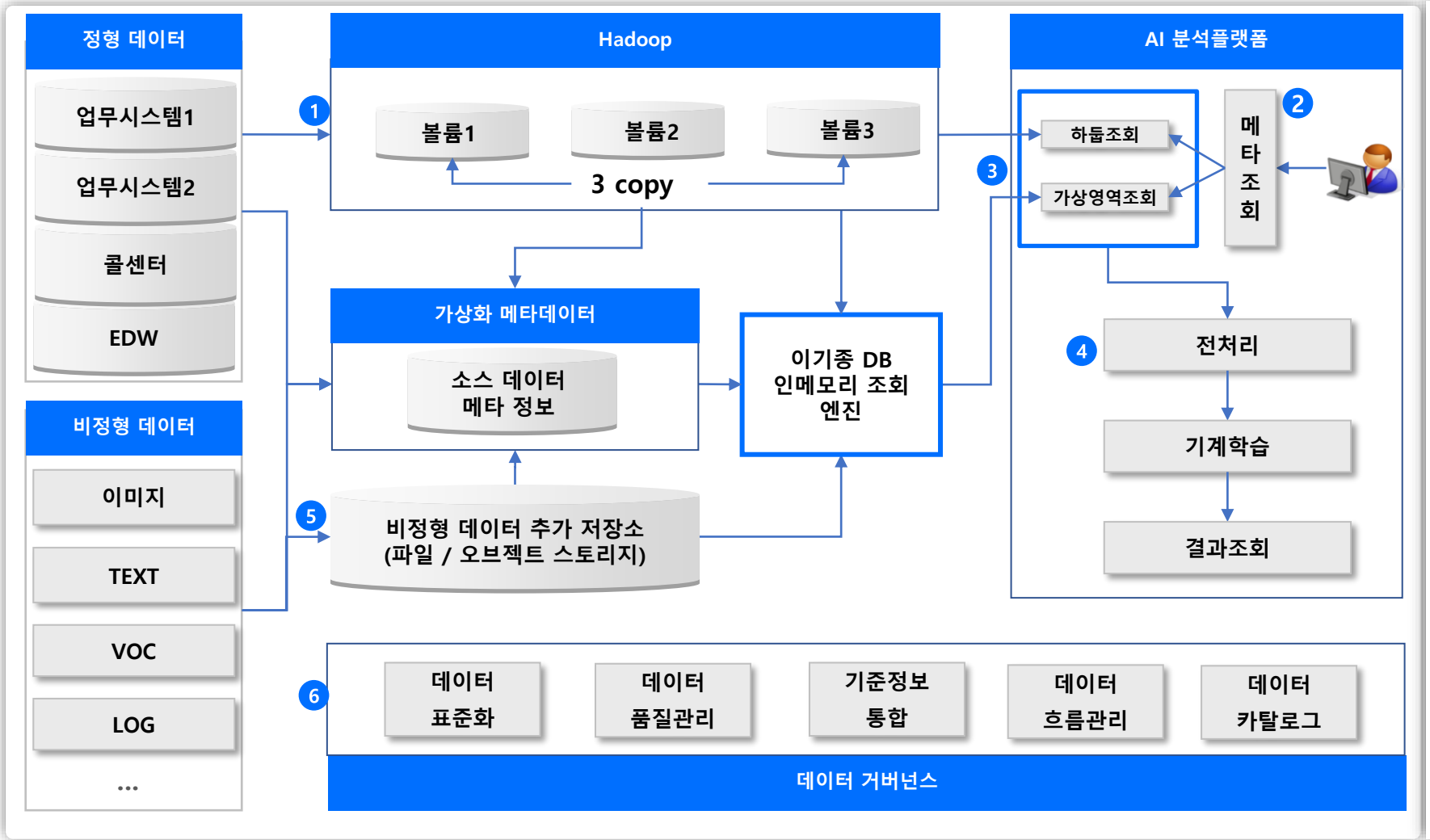


구현 방안

1 기간계 및 EDW의 데이터이동
- TeraONE ETL을 통하여 추출
- MapReduce엔진으로 hadoop에 분산 적재

2 데이터 가상화
- 이기종 DB를 하나의 가상 DB connection 만으로 모든 ANSI-SQL 동작
- 가상 DB connection을 위한 JDBC Drive 제공
- 기존 DBMS와 HDFS 및 HIVE 간 Join 데이터 조회
- Join 분석을 위한 별도의 ETL 개발이 필요 없음

하둡에 전체 저장/관리하기 어려운 대용량 이미지 및 텍스트 파일 등은 별도 스토리지에 보관하고, 데이터 가상화의 솔루션의 메타정보를 활용하여 분석에 활용하고자 하는 모든 데이터 소스에 접근할 수 있습니다.



구현 방안

1 메타데이터 수집
- 분석대상의 메타 데이터 수집

2 AI학습 대상 데이터 검색
- 분석대상 정보를 검색 하여 데이터 레이크 존재여부 확인

3 AI학습 대상 데이터 획득
- Hadoop에 있는 데이터는 Hive SQL을 이용하고 그외 데이터는 SuperQuery SQL을 이용하여 AI분석용 데이터 획득

4 AI 분석
- 획득된 데이터를 전처리한 후 AI 학습 결과 도출

5 비정형 데이터 실시간 수집
- 향후 비정형 실시간 데이터도 수집 분석하여 활용 가능

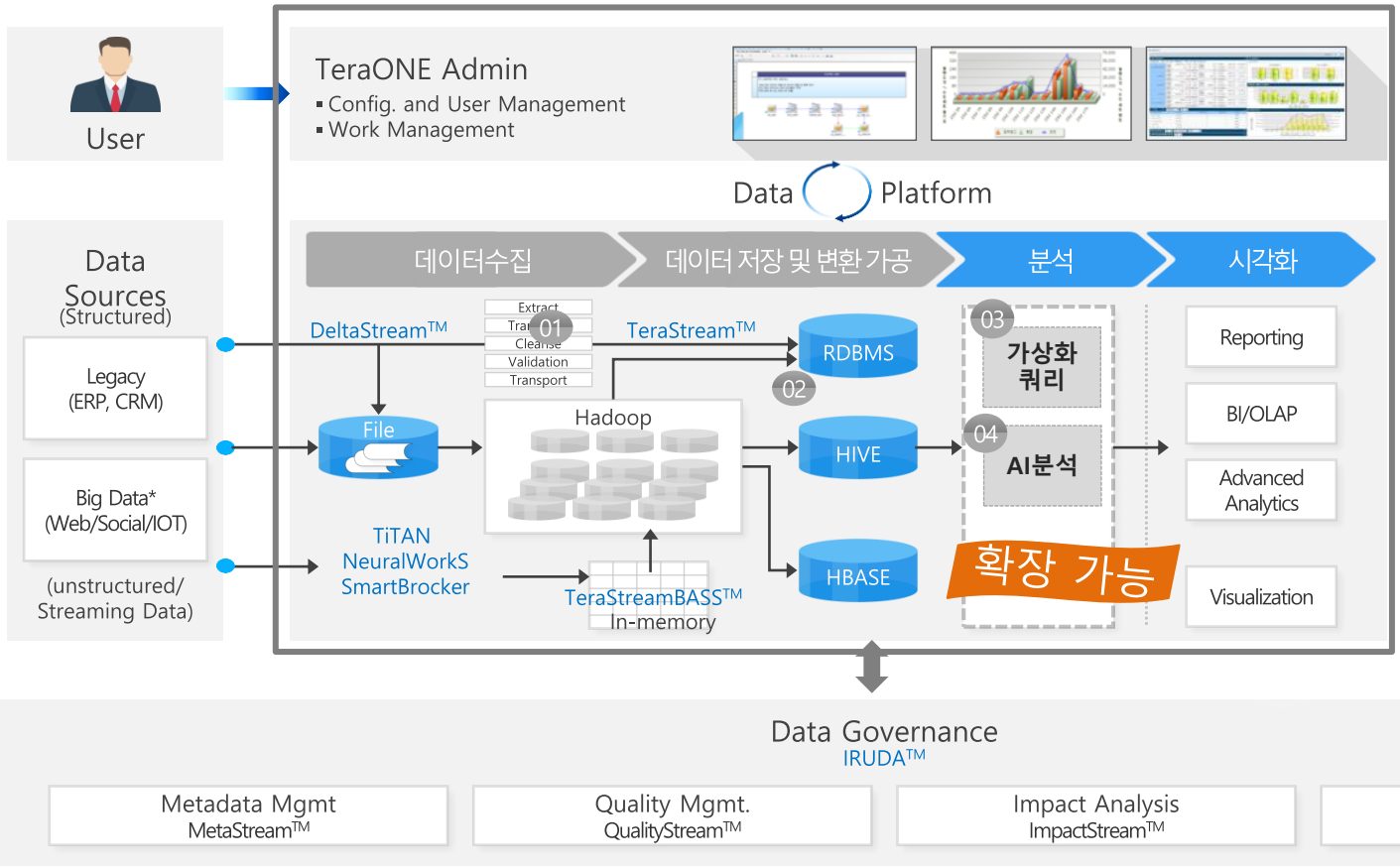
6 데이터 거버넌스 확대
- 향후 데이터거버넌스 기능 단계적 확대 가능

당사 빅데이터 플랫폼은 아파치 하둡 기반으로 자사 강점인 ETL 및 인메모리 분산 기술을 적용한 순수 국산 상용 빅데이터 플랫폼입니다.

특장점

- TeraONE은 Apache hadoop 배포판을 기반으로 자사 강점인 ETL 및 인메모리 분산 기술을 적용한 순수 국산 빅데이터 플랫폼 입니다.
- 국립암센터, 데이터산업진흥원, 임업진흥원, 주택도시보증공사, 한국철도공사, 금융감독원, 금융결제원, 서울교통공사, 농촌진흥청, 마사회 등 다수의 기관에서 사용중인 안정적인 빅데이터 플랫폼입니다.

업계정상급 경쟁제품군에서 가장 최신 버전인 빅데이터 플랫폼 TeraONE



주요특징

01 자사 ETL은 파일 분산 병렬 처리기술 원천을 기반으로 하둡으로 손쉬운 확장 가능

02 TeraONE은 기존 RDBMS 및 하둡 동시 지원 가능

03 데이터 가상화 분석 확장 가능(SuperQuery)

04 AI분석 확장가능(IDEA)

Thank You

ABLESTOR
Dynamic Value Creator

ABLECLOUD
All about data & cloud