빅데이터 기술동향 및 차세대 데이터 플랫폼 전략

Big Data Fabric & Federation

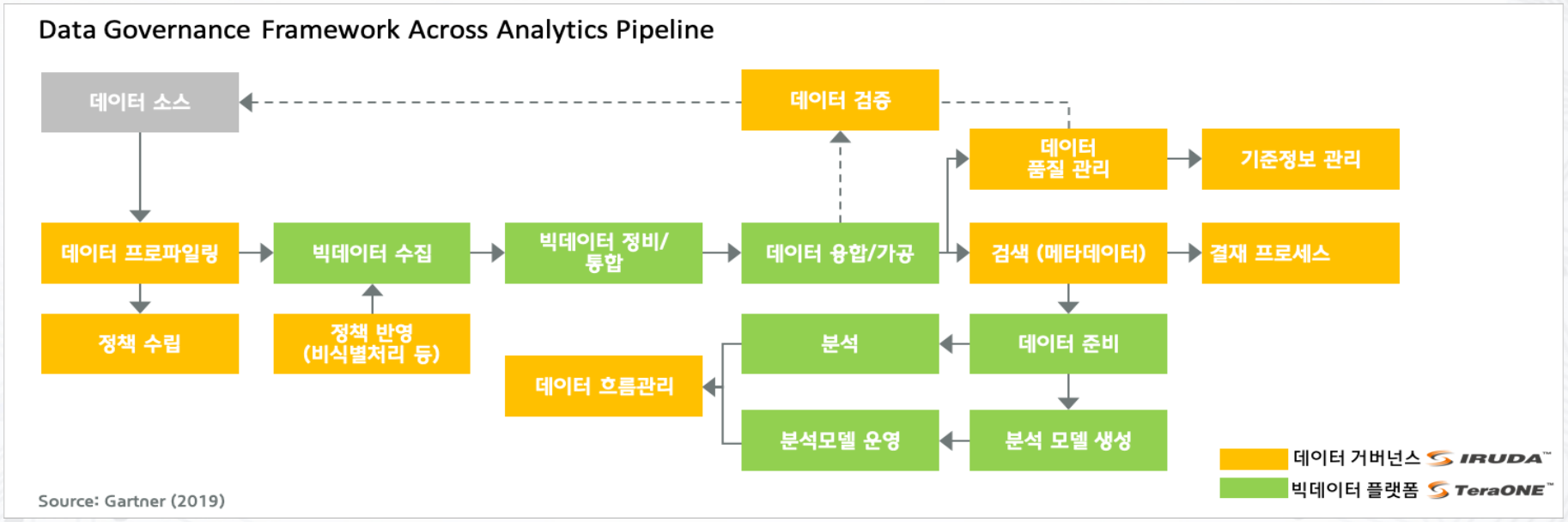
2019. 10

목차

1 데이터 거버넌스 기반의 빅데이터 역량

데이터 레이크 사업이후에 데이터 관리체계의 중요성이 다시 부각되고 있으며, 빅데이터 분석의 파이프라인에 데이터 거버넌스의 접목이 필수불가결해졌습니다. 당사는 이 분야의 솔루션, 컨설팅, 수행 역량을 모두 갖추고 있습니다.

솔루션 역량



유관 조직

- 시장 중심의 제품 개발과 제품 중심의 컨설팅 및 수행이 가능한 주요 조직 간의 유기적인 협업 체계가 경쟁력의 원천

Business Consulting 본부
빅데이터 분석 컨설팅
데이터 거버넌스 컨설팅
데이터 전략 컨설팅
마켓 센싱 및 요건 수렴

Product Service 본부
제품기획총괄
시장조사 및 제품컨설팅
품질보증 / 신기술 지원
유지보수

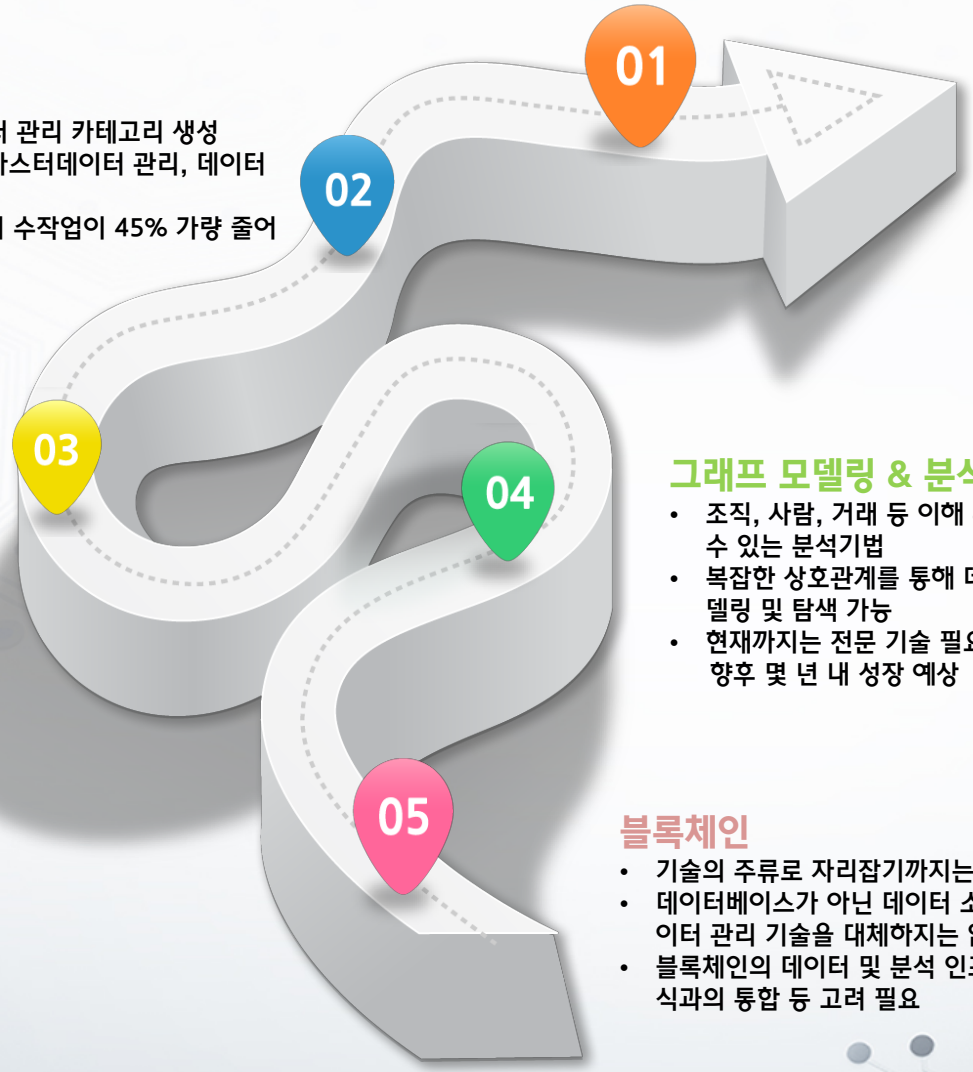
기술연구소
선행기술 연구
제품 개발
연구과제 수행
기술 지원

Technical Service 본부
당사 제품기반
프로젝트 수행

2 빅데이터 기술 트렌드

가트너가 2019년 2월 발표한 10대 데이터 · 분석 기술 트렌드 중 발췌하여 요약한 내용입니다.

(증강분석, 증강데이터관리, 지속적 지능화, 설명가능한 AI, 그래프, 데이터 패브릭, NLP 및 대화형 분석, 상용 AI와 머신러닝, 블록체인, 퍼시스턴트 메모리 서버)



증강 데이터 관리

- ML과 AI를 활용해 기업의 데이터 관리 카테고리 생성
- 데이터 품질, 메타데이터 관리, 마스터데이터 관리, 데이터 통합 등이 포함됨
- 2022년 말에 이르면 데이터 관리 수작업이 45% 가량 줄어들 것으로 전망

증강분석 & 대화형 분석

- ML과 AI를 통해 분석 컨텐츠가 개발, 소비, 공유되는 방식인 증강분석이 차세대 혁신의 물결
- 2020년까지 분석 쿼리의 50%가 검색, 자연어처리, 음성을 통해 생성될 것임

빅데이터 패브릭

- 분산 데이터 환경에서 마찰없는 엑세스와 데이터 공유 가능
- 사일로화 된 저장소를 탈피해 일간된 단일 데이터 관리 프레임웍 구축

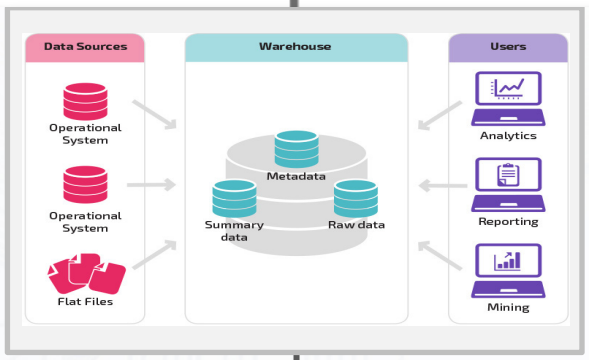
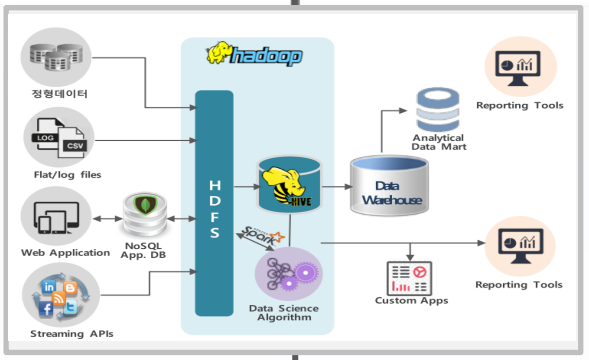
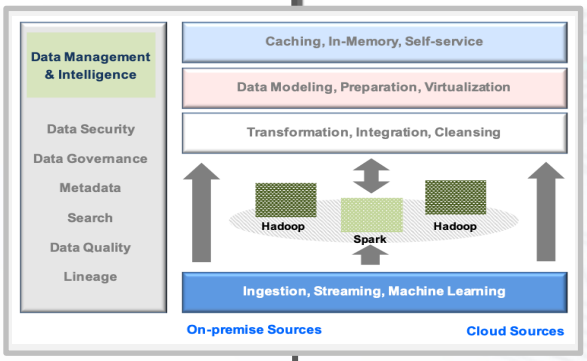
그래프 모델링 & 분석

- 조직, 사람, 거래 등 이해 주체 간 관계를 탐색할 수 있는 분석기법
- 복잡한 상호관계를 통해 데이터를 효율적으로 모델링 및 탐색 가능
- 현재까지는 전문 기술 필요해서 제한적이나 향후 몇 년 내 성장 예상

블록체인

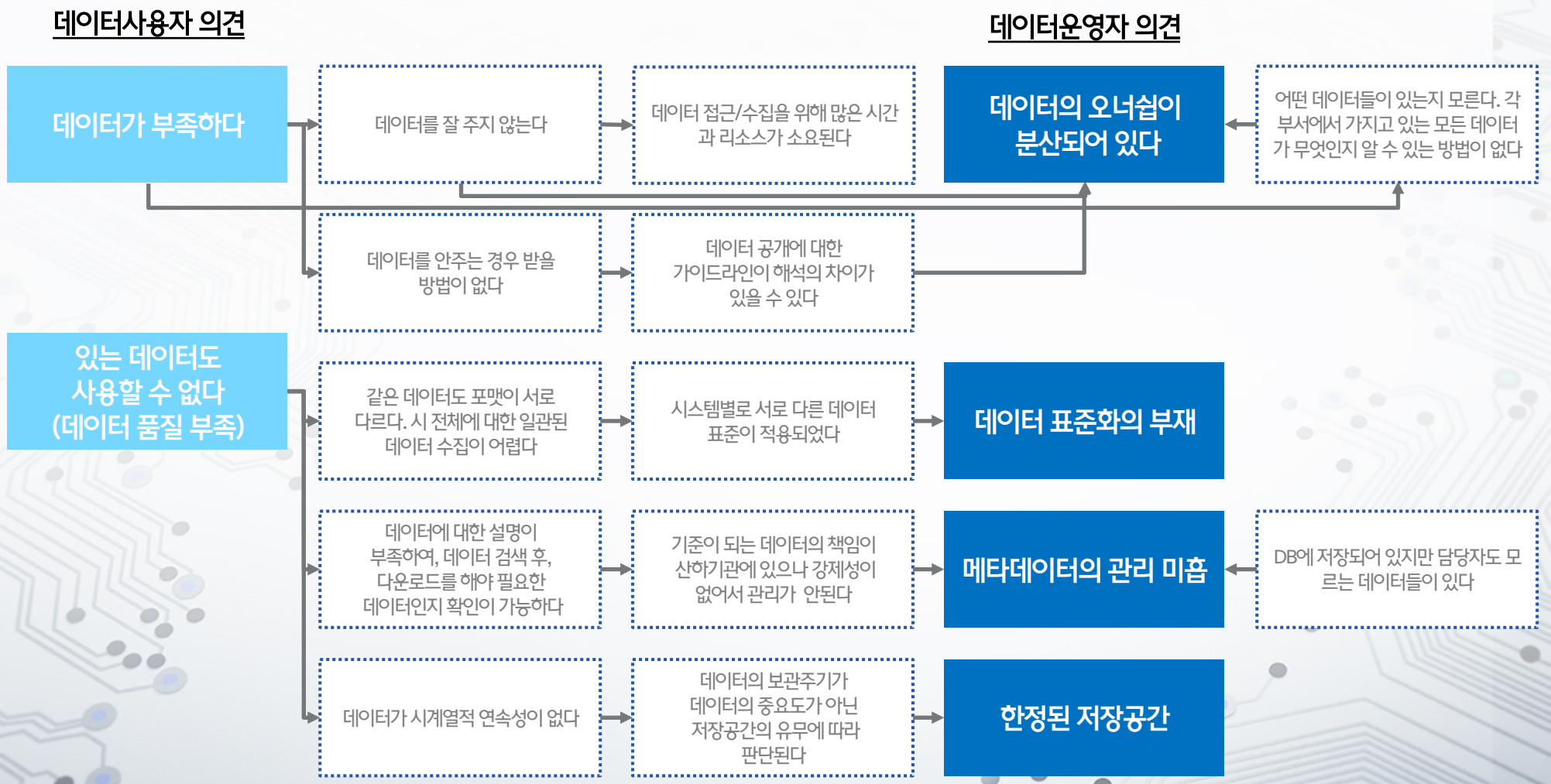
- 기술의 주류로 자리잡기까지는 수년 소요 예상
- 데이터베이스가 아닌 데이터 소스이며, 기존 데이터 관리 기술을 대체하지는 않을 것임
- 블록체인의 데이터 및 분석 인프라와의 기존 방식과의 통합 등 고려 필요

3 빅데이터 플랫폼 발전 동향 – 데이터 레이크에서 빅데이터 패브릭으로

	데이터웨어하우스	데이터 레이크	빅데이터 패브릭
참조 아키텍처			
주요 특징	• 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
제약 사항	• 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

3 빅데이터 플랫폼 발전 동향 – 빅데이터 활용관점의 데이터 거버넌스

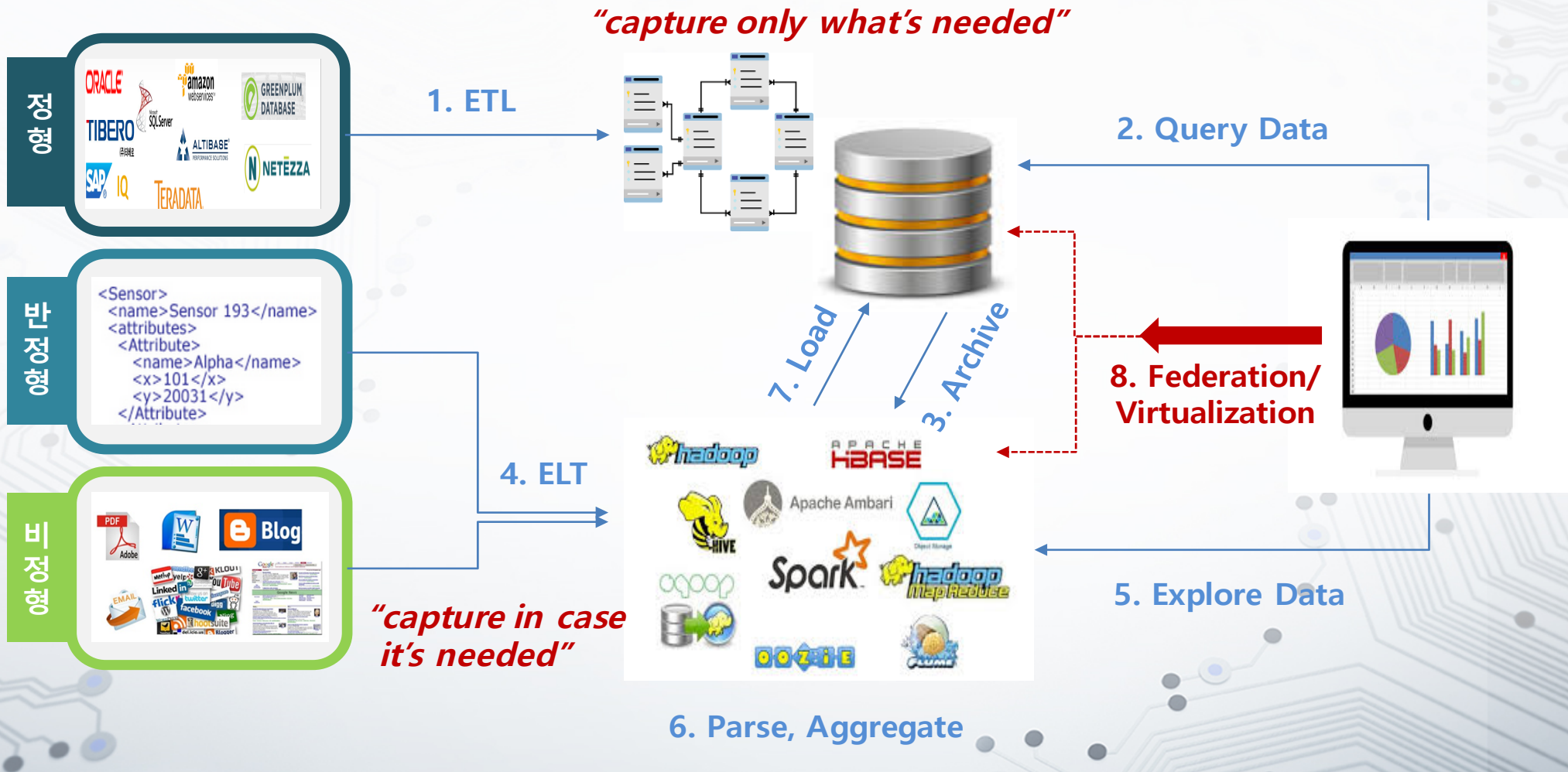
빅데이터 도입이 확대되면서 분석가들이 활용할 데이터가 부족하고, 이마저도 각 부문에서 별도 관리하거나 데이터에 대한 총괄 오너십이 부족한 현상이 많이 나타나면서, 빅데이터를 제대로 하기 위한 거버넌스 요구가 증대되고 있습니다.



3 빅데이터 플랫폼 발전 동향 – 다양한 이기종 데이터 소스 연계 분석

하둡기반의 데이터 레이크를 구축한 고객사들은 최근 기존 데이터웨어하우스의 가치를 유지하면서 효율적인 빅데이터 플랫폼 구축을 고민하고 있습니다. 관계형 데이터베이스와 분산 파일시스템의 강점을 살린 최적화 방안을 요구합니다.

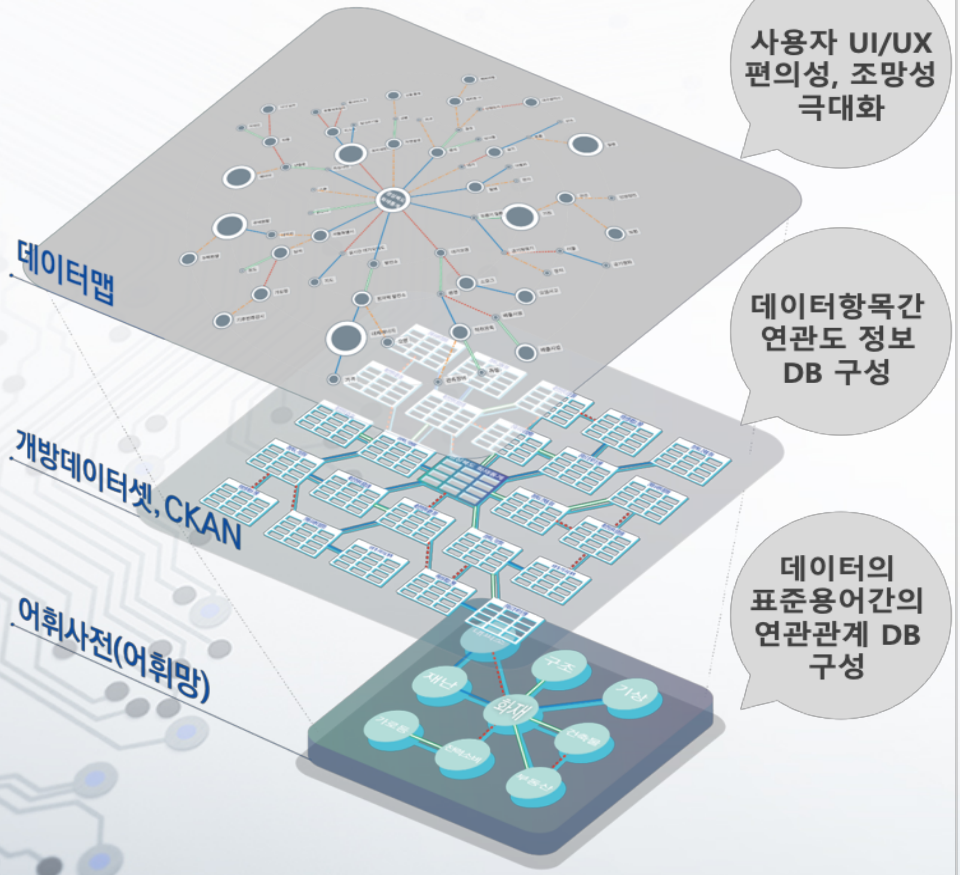
➡ **RDB, Hadoop에 상관없이 하나의 Query로 데이터를 조회/분석하는 기능 요구**



3 빅데이터 플랫폼 발전 동향 – 분석하고자 하는 데이터의 용이한 사용

데이터의 분류, 관리, 검색 등을 용이하게 할 수 있도록 검색엔진 기반으로 데이터맵을 제공하여 사용자는 활용하고자 하는 데이터의 정보를 쉽게 파악할 수 있습니다.

메타데이터사전 기반 데이터맵 설계



데이터맵을 통한 데이터 자산화

데이터맵 엔진 개발

데이터 연관도 분석	연관도 분석결과 DB 저장	데이터맵 UI 반영
		
벡터 및 데이터사전 기반 유사도 계산	**메타데이터DB**	**데이터맵**
• 데이터(셋)의 메타데이터에 대한 벡터화를 수행하여 연관도를 도출	• 산림메타데이터DB에 계산된 '연관도' 및 '연관데이터명'을 입력 (또는 갱신)	• 국가데이터맵 UI에 3가지 유형의 데이터 연관도 반영

데이터맵 활용

메타데이터 관리시스템 | 타 플랫폼 연계 | 빅데이터 공통기반

3 빅데이터 플랫폼 발전 동향 – 빅데이터 마켓플레이스

데이터 거버넌스 기반의 빅데이터 플랫폼을 통해 수집/가공/융합/분석된 데이터는 빅데이터 마켓플레이스를 통해 거래/유통할 수 있는 상품의 개념으로 발전하고 있습니다.

중국

- 세계 최초로 민관 합작형태의 데이터 거래소 설립을 통해 공공데이터의 가공 및 판매, 민간데이터 거래 등 공공 주도

미국

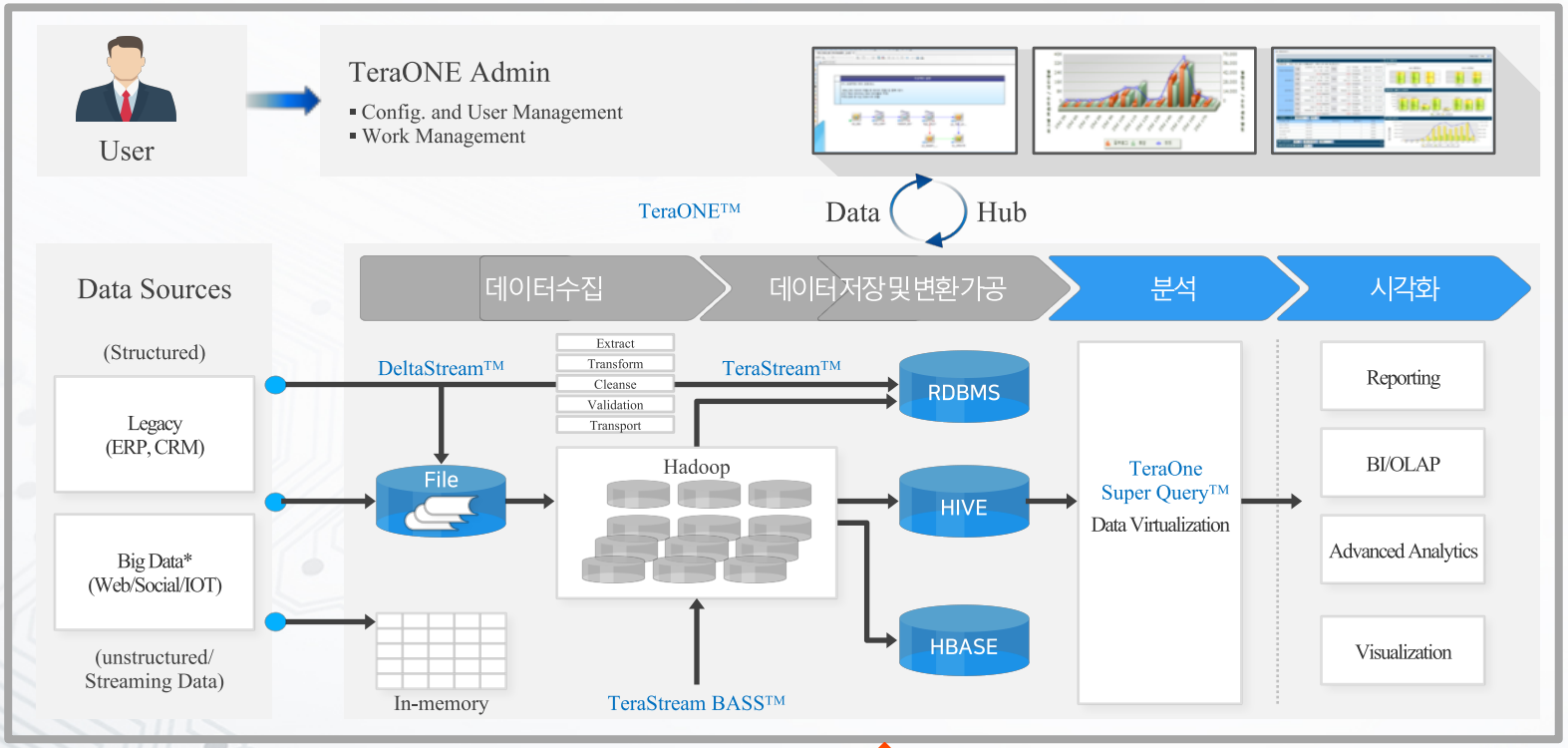
- 개인정보 활용이 비교적 용이한 환경하에서 세계 최대 규모의 데이터 브로커 시장 (약 1,500억불) 등 민간 기업 위주로 데이터 거래 활발

한국

- 정부주도 대규모 데이터 사업 시행 (바우처 사업, 빅데이터 플랫폼 사업 등), 빅데이터 일자리 창출
- 각종 규제 완화 검토 중

4 차세대 데이터 플랫폼 전략 – 빅데이터 패브릭 TeraONE Fabric™

당사 빅데이터 패브릭 플랫폼 TeraONE Fabric™은 빅데이터 플랫폼 TeraONE™, 데이터 거버넌스 플랫폼 IRUDA™, 데이터 가상화 플랫폼 TeraONE Super Query™로 구성되어 있습니다.



주요특징

01 자사 ETL은 파일 분산병렬 처리기술 원천을 기반으로 하둡으로 손쉬운 확장 가능

02 TeraONE은 기존 RDBMS 및 하둡 동시 지원 가능

03 빅데이터 메타기능 제공 및 전사 메타와 용이한 연계

TeraONE Fabric™

Data Governance
IRUDA™

Metadata Mgmt	Quality Mgmt.	Data Lineage	Master Mgmt
MetaStream™	QualityStream™	Q-Track™	MasterStream™

4 차세대 데이터 플랫폼 전략 - 빅데이터 패브릭 3대 구성요소 ① 빅데이터 플랫폼 TeraONE™

40여개의 빅데이터 플랫폼 구축 및 기술지원 경험으로 고객들이 가장 많이 활용하는 오픈소스 중심으로 자사 제품과 패킹하여 최적화 아키텍처를 제공합니다.

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, HBASE , RDB등에 유연하게 적재 가능합니다.

탑재된 오프소스 버전

Ambari HDP 2.7.0
Hadoop / Mapreduce : 3.1
Spark : 2.3
Hive : 3.0
Flume : 1.5.2
Hbase : 1.1.2
Zeppelin : 0.7.3
R / R-studio : 3.6.x (최신버전)
Python : 3.6.x

업계정상급 경쟁제품군에서 가장 최신 버전인 빅데이터 플랫폼 TeraONE

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	SQL HIVE		In-Memory SPARK
Transform TeraStream for Hadoop Engine		Scheduling TeraStream Designer		Data Visualization Zeppline
	NoSQL HBASE	SQL On Hadoop HIVE	Analyzing R, SAS, PYTHON	

MetaData Management (Hcatalog)

Kerberos (Network Authentication Protocol)

Storage : HDFS(Hadoop Distributed File System)

추가 옵션(타사제품 포함)

Sandbox	SNS/Web Crawler	OLAP/Visualization	Data Virtualization TeraONE Super Query
MetaData Management MetaStream	Data Quality QualityStream	Data Linage Q-TRACK	CDC DeltaStream

자사 제품 | 타사제품 | 오픈 소스

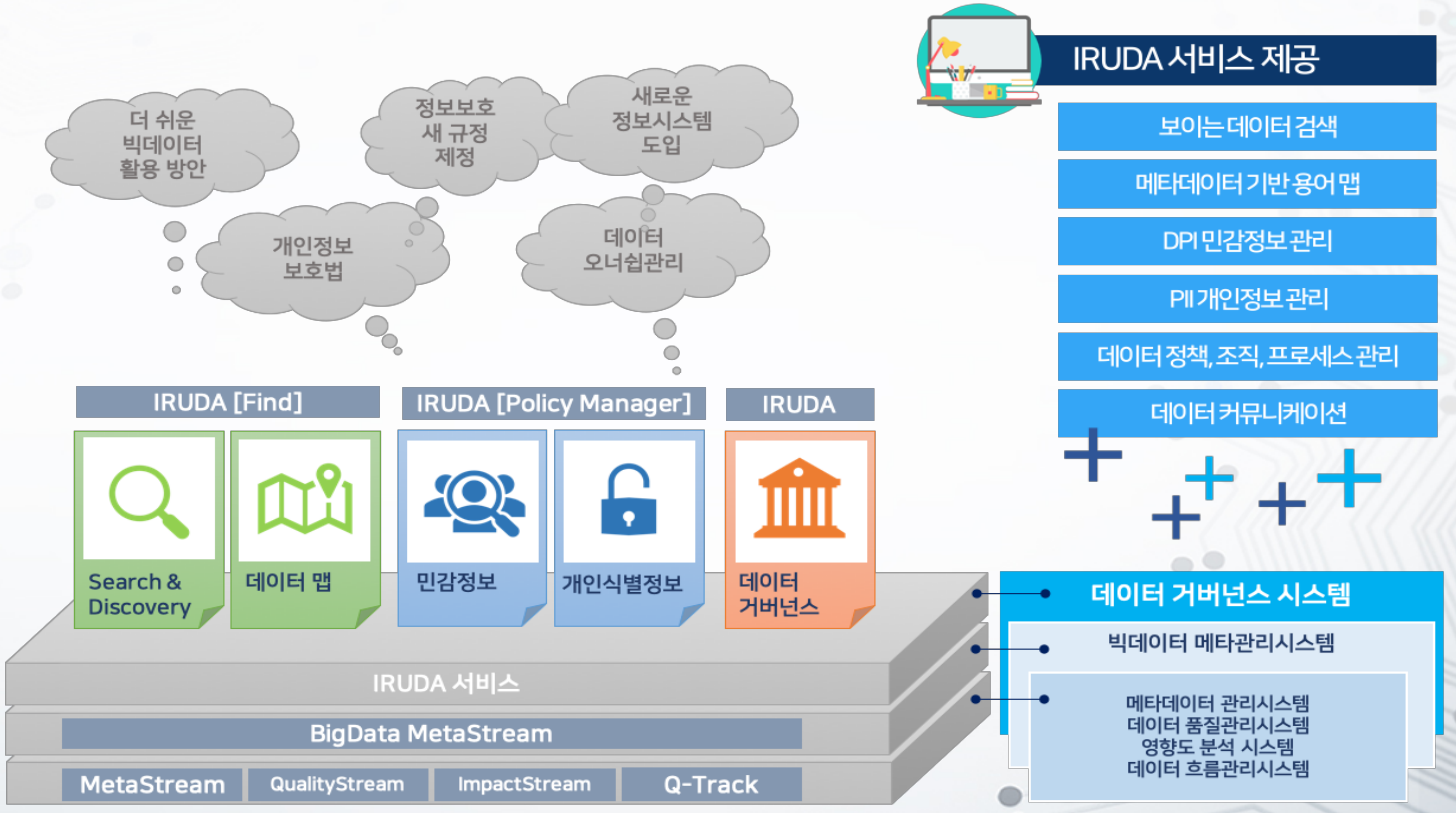
4 차세대 데이터 플랫폼 전략 - 빅데이터 패브릭 3대 구성요소 ② 데이터 거버넌스 IRUDA™

데이터 거버넌스 플랫폼은 데이터 분석과 활용관점에서 데이터를 자산화해 체계적으로 관리하고 데이터의 활용을 극대화할 수 있는 전반적인 거버넌스 체계를 지원합니다.

특장점

- 전사 메타데이터와 연계한 빅데이터 메타 관리시스템 구축을 통해 향후 데이터 거버넌스 관리 체계 구축하도록 확장성을 보장함
- 데이터 거버넌스 플랫폼 IRUDA를 통해서 향후 데이터 관련 서비스 제공
- 보다 쉽게 빅데이터를 활용하기 위한 검색 서비스 제공
- 개인정보보호법과 같이 데이터 정책의 변경 관리 및 대응 체계 구축
- 정보시스템 추가 시 데이터 관리의 전사적인 프로세스 및 조직 체계 가이드 제공

IRUDA의 데이터 거버넌스 서비스 확장



*DPI: Data Protection & Inspection, *PII: Personally identifiable information

4 차세대 데이터 플랫폼 전략 - 빅데이터 패브릭 3대 구성요소 ③ 데이터 가상화 TeraONE Super Query™

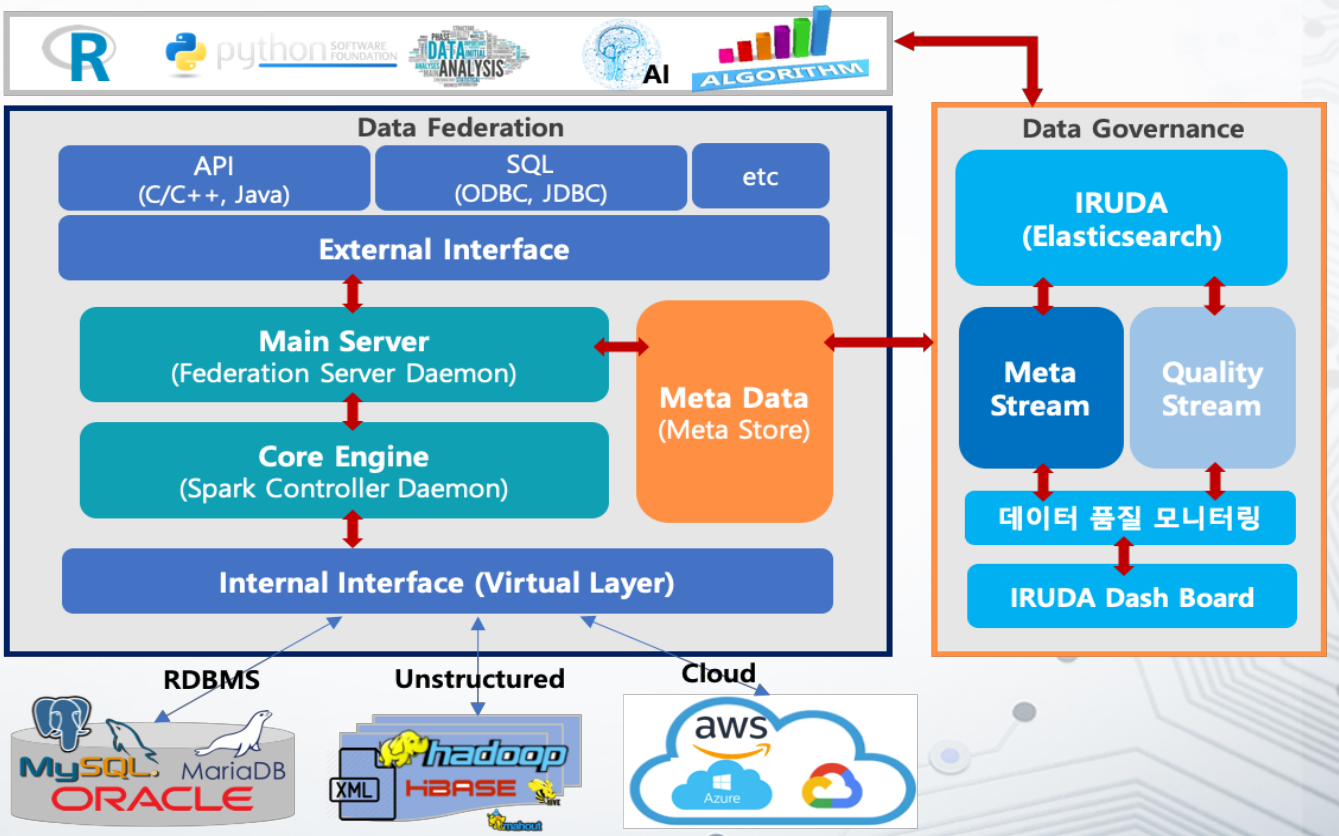
2019년 9월 출시한 데이터 가상화 솔루션은 자체 제품과 Apache Spark을 패킹하여 특허 출원하였으며, TPC-H 벤치마크 테스트 결과 상용DBMS나 기존 Hadoop보다 4배~11배 성능이 개선되었습니다.

특장점

- 인메모리 방식의 Spark 엔진을 적용하여 다양한 소스의 데이터에 빠르고 유연하게 접근하여 연산 처리 (표준 SQL 인터페이스)
- DB엔진 기반의 경쟁사 제품과 다르게 논리모델 사전 정의 없이 SQL 실행계획 최적화 가능하여 성능 보장
- 표준 데이터베이스 인터페이스를 활용하여 기존 애플리케이션 변경 최소화
- 모든 데이터 소스의 메타정보 관리 기반으로 하나의 데이터베이스 처럼 분석가능
- 데이터 소스의 품질측정 지수 제공 등 당사 IRUDA와 연계하여 강력한 데이터 활용 기반 제공

데이터거버넌스 기반 이기종 데이터 통합 및 연계 분석

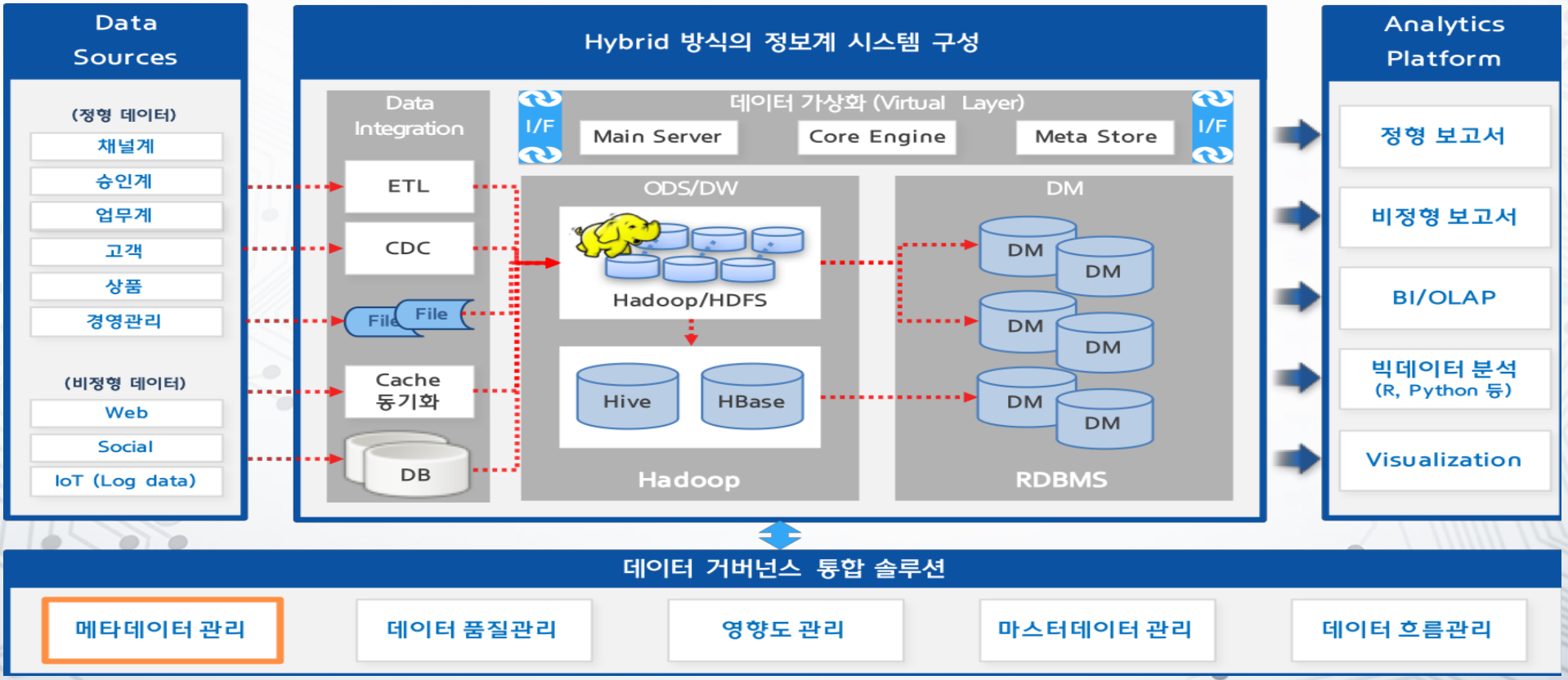
RDBMS, 비정형 데이터 소스, 클라우드 데이터 소스 등 모든 이기종의 데이터소스를
메타데이터로 접근하여 활용 가능한 가상화 기술 기반의 분석지원 플랫폼



4 차세대 데이터 플랫폼 전략 - 아키텍처 제안

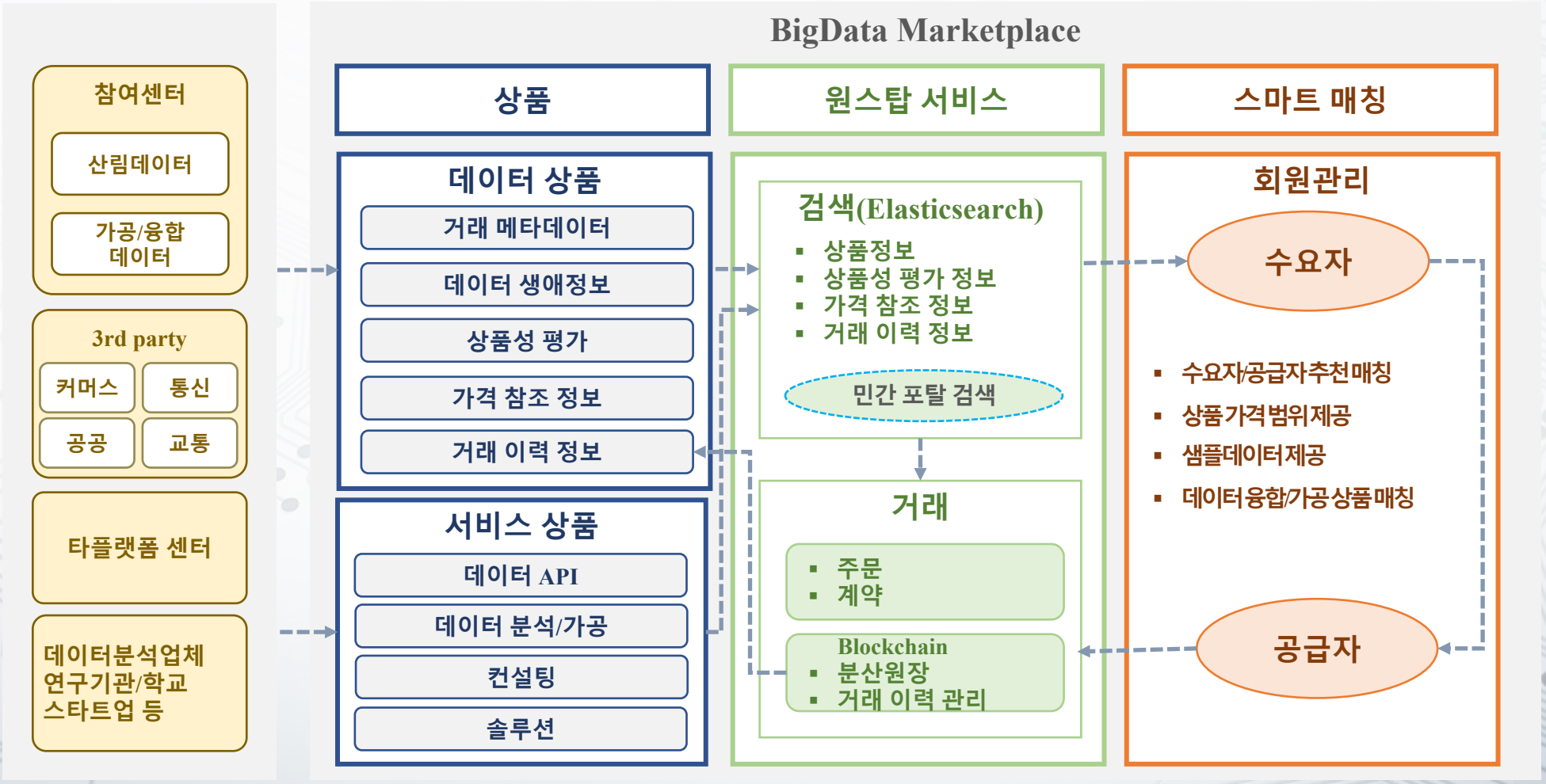
원천 소스에서 정형/비정형 데이터를 추출하여 하둡 기반의 ODS에 적재 후, RDBMS로 데이터 마트를 구축 할 경우 상대적으로 저장비용이 저렴합니다. 데이터 가상화 레이어를 구성하여 ODS/DW, DM 등과 연계 및 분석이 가능합니다.

ODS/DW-Hadoop / DM-RDBMS : 데이터 가상화 기반 Hybrid 정보계 구성



4 차세대 데이터 플랫폼 전략 – IRUDA for Bigdata Marketplace™

데이터 거버넌스 및 블록체인 기반으로 데이터의 원본 훼손 방지, 소유자 인증, 거래이력 추적 및 실 거래를 위한 정산 프로세스가 구현되어 데이터 거래의 활성화를 제고합니다.



5 데이터스트림즈의 빅데이터 기술 혁신

데이터스트림즈 기술연구소의 기술 혁신 노력은 특허와 인증 그리고 우수 논문 선정 등의 결과로 이어지고 있습니다.

특허

SparkSQL기반의 데이터 페더레이션 저장 장치 (2018년 1월 31일 출원)

- **SparkSQL 기반의 '빅 데이터의 실시간 저장 및 검색 시스템' 으로 '17년 11월 특허 출원하여 '19년 4월 등록**
  - ☛ IoT 솔루션 TeraStream for BASS 기반 기술 (도로공사/철도공사에서 품질 입증)
- **SparkSQL에 대한 지속적인 연구결과가 'SparkSQL기반의 데이터 페더레이션 저장장치'의 특허 출원으로 이어짐**
  - ☛ 데이터 가상화 솔루션 TeraONE Super Query 기반 기술 '18년 1월 특허 출원 후 시제품 개발, '19년 9월 출시함
  - ☛ 핵심 아이디어 : 자사솔루션 (메타관리 MetaStream과 고속 데이터 추출 FACT) + 오픈소스 (고성능 분산 컴퓨터 프레임웍 Apache Spark)
  - ☛ 주요 이슈 해결
    - MetaStream을 통한 가상화 레이어와 물리테이블 매핑
    - Fact를 통한 Local/HDFS 등 다양한 원천 수집 가능
    - Apache Spark의 고성능 기술과 결합하여 단일 쿼리 완성

논문

빅데이터와 딥러닝을 이용한 도메인 자동 추천 시스템에 관한 연구 (2019년 10월 게재)

- **선행연구에서 제시되었던 도메인 자동판별 기법이 아닌, 도메인 자동 추천 시스템의 모델을 딥러닝과 빅데이터 관점에서 제시함**
  - ☛ 순환신경망, 합성곱 신경망 등 딥러닝으로 학습 시킨 후 순위시스템 적용으로 확률이 높은 Top 5 자동 추천
  - ☛ MetaStream 차기 버전에 반영하여 도메인 관리 공수 절감 및 데이터 품질 향상 개선 기대

  (2019년도 한국인터넷정보학회 추계학술발표대회 논문집 제20권2호, 우수논문으로 채택)

- **MDMM(Multi Database Meta-Information Management) 연구중이며, 11월 15일 논문 투고 목표**
  - ☛ 딥러닝을 활용한 다중의 표준어 사전 관리 알고리즘 연구

감사합니다.