

<제목 차례>

2. 기술(제품)의 주요 내용	4
□ 주요 기술 내용	4
● 기술(제품)의 개요	4
● 기술(제품)의 특성 및 핵심기술	5
1. 데이터 통합(Data Integration) 기술	7
가. 데이터 통합 (TeraStream™)	7
1) 제품 특징 및 차별화 요소	7
2) 제품 적용기술 내역	8
3) 주요 기능	9
4) 적용사례	11
나. 실시간 변경분 데이터 처리 (DeltaStream™)	11
1) 제품 특징 및 차별화 요소	11
2) 제품 적용기술 내역	13
3) 적용사례	13
다. IoT 데이터 실시간 처리 솔루션 (TeraStream BASS™)	14
1) 제품 특징 및 차별화 요소	14
2) 제품 적용기술 내역	15
3) 주요 적용사례	16
2. 빅데이터 플랫폼(Big Data Platform) 기술	17
가. 빅데이터 저장소 (TeraONE™ HDFS)	17
1) 제품 특징 및 차별화 요소	17
3) 주요 적용사례	19
3. 인공지능(AI) 및 어날리틱스(Analytics)를 위한 분석 플랫폼 기술	20
가. 인공지능 기반 빅데이터 분석 환경 (TeraONE IDEA™)	20
1) 제품 특징 및 차별화 요소	20
2) 제품 적용기술 내역	21
3) 주요 적용사례	22
4. 데이터 패브릭(Data Fabric) 기술	23
가. 데이터 가상화 (TeraONE SuperQuery™)	23
1) 제품 특징 및 차별화 요소	23
2) 제품 적용기술 내역	24
3) 주요 적용사례	25
5. 데이터 거버넌스 (Data Governance) 기술	25
가. 데이터 거버넌스 플랫폼 (IRUDA™)	25
1) 제품 특징 및 차별화 요소	26
2) 제품 적용기술 내역	26
3) 주요 적용사례	27
나. 데이터 표준화 및 메타데이터 관리 도구 (MetaStream™)	28
1) 제품 특징 및 차별화 요소	28
2) 제품 적용기술 내역	29
3) 주요 적용사례	29
다. 데이터 품질관리 도구 (QualityStream™)	30
1) 제품 특징 및 차별화 요소	30

2) 제품 적용기술 내역	31
3) 주요 적용사례	32
라. 비즈메타관리 도구 (MetaStream For BizData™)	32
1) 제품 특징 및 차별화 요소	32
2) 제품 적용기술 내역	33
3) 주요 적용사례	34
마. 데이터 흐름관리 도구 (Q-Track™)	34
1) 제품 특징 및 차별화 요소	34
2) 제품 적용기술 내역	35
3) 주요 적용사례	36
바. 마스터데이터관리 도구 (MasterStream™)	37
1) 제품 특징 및 차별화 요소	37
2) 제품 적용기술 내역	38
3) 주요 적용사례	38
◉ 기술개발내용 및 과정	39
1. 빅데이터 플랫폼의 활용 배경	39
2. 제품의 개요	40
3. 제품의 활용	42
4. 활용 예시	43
5. 구성 제품의 개념 및 활용	45
가. 데이터 통합(Data Integration) 기술	45
1) 데이터 통합 및 빅데이터 준비 도구	45
나. 빅데이터 플랫폼(Big Data Platform) 기술	46
1) 빅데이터 통합 저장소	46
다. 인공지능(AI) 및 어날리틱스(Analytics)를 위한 분석 플랫폼 기술	48
1) 인공지능 기반 빅데이터 분석 환경 (TeraONE IDEA™)	48
라. 데이터 패브릭(Data Fabric) 기술	50
1) 데이터 가상화 (TeraONE SuperQuery™)	50
마. 데이터 거버넌스(Data Governance) 기술	51
1) AI 기반 데이터 카탈로그 서비스 (IRUDA Navigator™)	52
2) 데이터 표준화 및 메타데이터 관리 도구 (MetaStream™)	54
3) 비즈메타관리 도구 (MetaStream For BizData™)	55
4) 메타데이터 초(超)상세화 (메타데이터 확장 서비스)	56
5) 데이터 품질관리 도구 (QualityStream™)	57
6) 데이터 흐름관리 도구 (Q-Track™)	58
7) 마스터데이터관리 도구 (MasterStream™)	59
◉ 국내외 기술동향	60
1. 전통적인 Data Warehouse기술에서 빅데이터 분석 환경으로 변화	60
가. 분석 대상 데이터의 확대	60
나. 데이터 분산처리 기술의 등장과 성장	61
다. EDW와 빅데이터 분석	62
라. 빅데이터 플랫폼으로 확장	63
2. 클라우드 환경으로 변화	63
가. 빅데이터 분석에 대한 인식 변화	63
나. 빅데이터 인프라 기술의 변화	63

다. 클라우드 기반의 빅데이터 플랫폼	64
1) 구성에 대한 요구사항을 입력하고 수분 내에 자동적으로 구성하는 기술	64
2) 데이터 탐색을 지원하는 데이터 허브 탐색, 데이터 큐레이션을 통한 인사이트 도출 지원 기술	64
3) 접근 가능한 데이터에 대한 데이터 맵 생성 기술	64
4) 데이터의 특성을 인지하여 수집 방법, 저장 방법, 분석 방법에 대한 추천 기술	64
5) 데이터 맵에 기반한 분석 자동화 기술	64
3. Data Fabric 환경으로 변화	64

2. 기술(제품)의 주요 내용

□ 주요 기술 내용

◎ 기술(제품)의 개요

차세대 빅데이터 플랫폼 TeraONE™은 (주)데이터스트림즈가 20년간 개발해온 기술력이 총망라된 소프트웨어 솔루션으로 빅데이터 수집·가공·저장은 물론, 데이터 거버넌스(Data Governance), 데이터 가상화(Virtualization), 인공지능 분석 및 빅데이터 관리 기능을 제공하는 빅데이터 패브릭(Fabric) 제품입니다. 각 제품의 상품명 및 역할은 아래 표와 정리하였습니다.

표 1 (주)데이터스트림즈 제품명 및 계통도

제 품 명			제품 요약
Level1	Level2	Level3	
TeraONE™ (Standard, Professional, Fabric)	빅데이터 패브릭 기능을 제공하는 차세대 빅데이터 플랫폼		
	FACT™/TeraSORT™/TeraTDS™	초고속 데이터 추출, 소팅, 대칭형 가명화	
		TeraStream™	데이터 통합 (ETL)
		TeraStream BASS™	실시간 데이터 수집 (IoT 등)
		DeltaStream™	실시간 변경 데이터 적재 (CDC)
		TeraONE IDEA™	개인화 기반 인공지능 분석 환경
	IRUDA™	TeraONE SuperQuery™	데이터 가상화
		전사차원의 데이터 거버넌스 플랫폼	
		IRUDA Navigator™	AI 기반 데이터 카탈로그
		MetaStream™	데이터 표준 및 메타데이터 관리
	MetaStream™	MetaStream for BizData™	비즈니스 메타데이터 관리
		QualityStream™	데이터 품질 관리
		Q-Track™	데이터 흐름 관리
		MasterStream™	마스터데이터 (기준정보) 관리

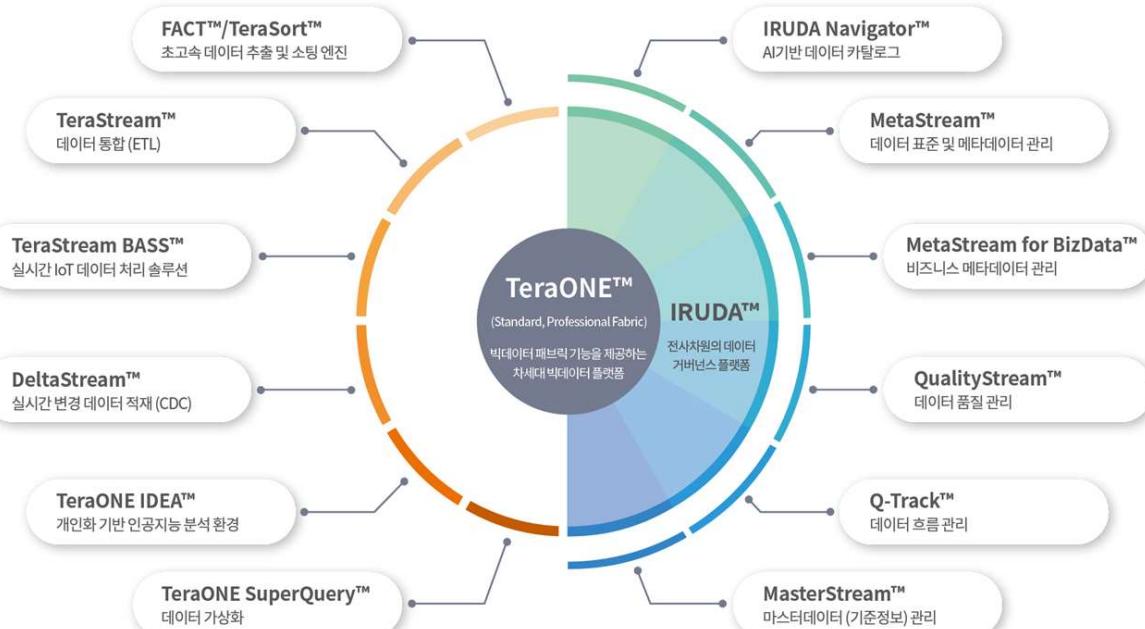


그림 1 (주)데이터스트림즈 주요 생산 제품 구성도

◎ 기술(제품)의 특성 및 핵심기술

(주)데이터스트림즈는 데이터 통합 및 관리 토클 솔루션 전문 연구개발 기업으로서, 차세대 데이터 플랫폼을 구성하는 데이터 통합, 데이터 거버넌스, 데이터 가상화 3가지 핵심 요소를 TeraONE™이라는 제품으로 통합하여 제공하고 있습니다.

- (1) 데이터 통합화는 데이터 수집/준비/저장/분석/시각화 등의 기능을 제공하는 제품입니다. 당사는 2001년 창업 시 자체 개발한 데이터 통합 솔루션인 FACT™/TeraSort™/TeraStream™을 근간으로 아파치 소프트웨어 재단에 공개된 오픈 소스인 하둡 에코 시스템을 패키징하여 공급하고 있으며, 국내외 대형 SI 회사들도 당사의 제품으로 사업을 수행할 정도의 패키징 역량을 자랑합니다. 데이터 통합화 관련 제품군은 국내 경쟁사는 없으며, 주로 클라우데라나 인포매티카와 같은 글로벌 탑 벤더와 경쟁하고 있습니다.
- (2) 데이터 가상화는 데이터 거버넌스를 기반으로 이기종의 데이터 플랫폼 또는 데이터 소스를 한군데의 거대한 빅데이터 저장소로 물리적으로 이동하여 통합하지 않고 가상화 레이어의 메모리 기술을 기반으로 원쿼리로 매우 빠르게 분석을 지원하는 제품입니다. 당사의 데이터 통합 원천기술인 FACT™, 메타데이터 관리기술인 MetaStream™, 데이터 거버넌스 종합관리 기술인 IRUDA™, 오픈소스 Apache Spark을 활용해 자체 개발한 분산 쿼리 엔진의 기술을 통합하여 개발한 제품으로 시중에 판매되고 있는 모든 종류의 데이터베이스 (Oracle, DB2, Tibero, SybaseIQ, Oracle Exadata, Altibase, MySQL, Hive, PostgreSQL, MariaDB 등)와 클라우드 플랫폼 (아마존, 구글, 마이크로소프트, KT, 네이버, NHN 등)의 데이터 소스를 하나의 데이터베이스처럼 조회하고 분석할 수 있습니다.
- (3) 데이터 거버넌스는 데이터 표준화, 데이터에 대한 설명 정보인 메타데이터 관리, 데이터 품질 관리, 데이터 흐름 관리, 기준 정보 관리의 기능을 포함하는 기업의 데이터 관리 도구입니다. 최근 빅데이터 분석이 활성화되면서, 데이터 레이크 내의 분석에 활용되는 데이터 셋에 대한 표준 관리, 품질 관리를 포함한 데이터 카탈로그 서비스의 수요가 급증하고 있습니다. 당사는 2005년부터 연구 발전 시켜온 데이터 거버넌스 기술을 기반으로 AI 기반 데이터카탈로그 분야 특허를 취득하고 관련 사업에 박차를 가하고 있습니다. (특허 제 10-2249466 호. “인공지능 추천 모델을 사용하여 추천 정보를 제공하는 데이터 카탈로그 제공 방법 및 시스템” 2021년 4월 특허등록 결정)

(주)데이터스트림즈의 보유 제품은 크게 TeraONE™으로 정의할 수 있으며, TeraONE™스탠다드 및 프로페셔널 버전은 데이터 통합화 솔루션입니다. TeraONE™패브릭 버전은 데이터 통합화 솔루션 외에도 데이터 가상화 솔루션인 TeraONE SuperQuery™와 데이터 거버넌스 솔루션인 IRUDA™까지 포함된 확장 버전입니다. 자사 기술로 데이터 패브릭 전략을 구현한 데이터 솔루션 기업으로는 당사가 유일합니다.

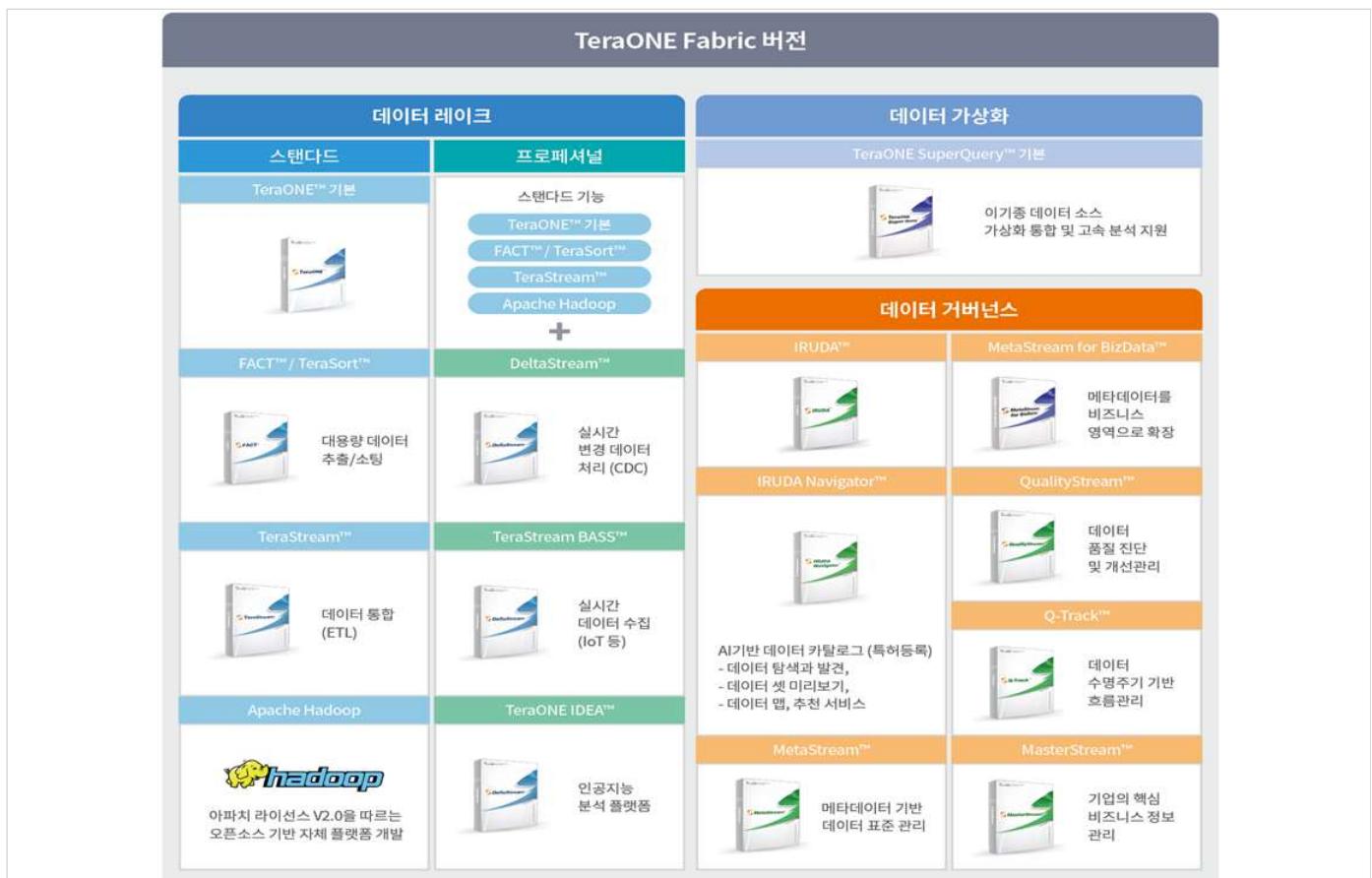


그림 2 TeraONE Fabric 버전 = 데이터 통합화 + 데이터 가상화 + 데이터 거버넌스

당사의 보유 제품군으로 차세대 데이터 플랫폼인 데이터 패브릭의 아키텍처 구성 모습은 다음과 같습니다.

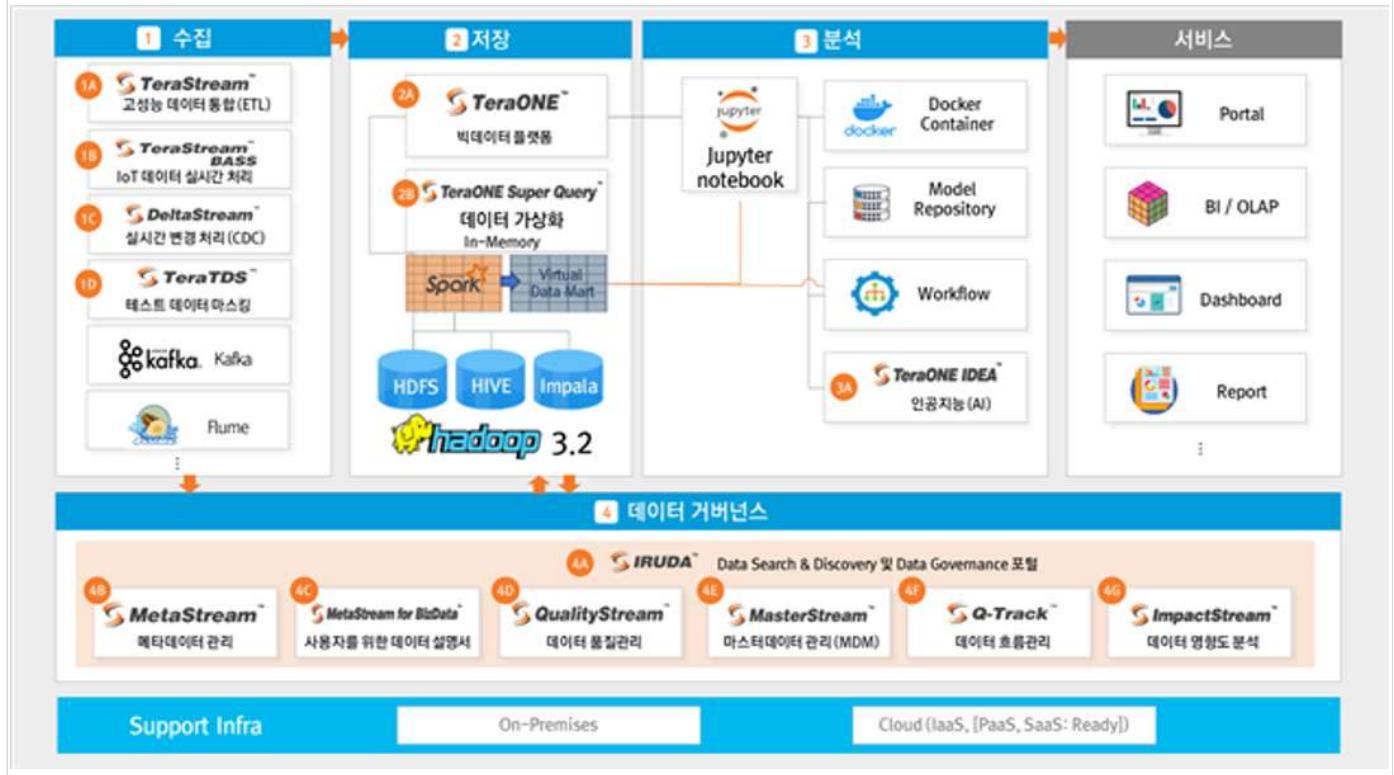


그림 3 (주)데이터스트림즈의 차세대 빅데이터 플랫폼(Data Fabric) 아키텍처

1. 데이터 통합(Data Integration) 기술

가. 데이터 통합 (TeraStream™)

1) 제품 특징 및 차별화 요소

가) 제품 개요

- 다양한 서버환경에서 원천 데이터를 빠르게 가공 처리하는 솔루션으로 ETL, Batch, 실시간 데이터 처리 연계 등 데이터 전환 업무를 수행합니다.
- 고속 추출엔진(FACT)과 Sorting 엔진을 탑재하여 데이터 가공 성능을 향상시키고, 시스템 자원을 효율적으로 이용하여 시스템 부하를 최소화합니다.

나) 개념도



그림 4 TeraStream™ 개념도

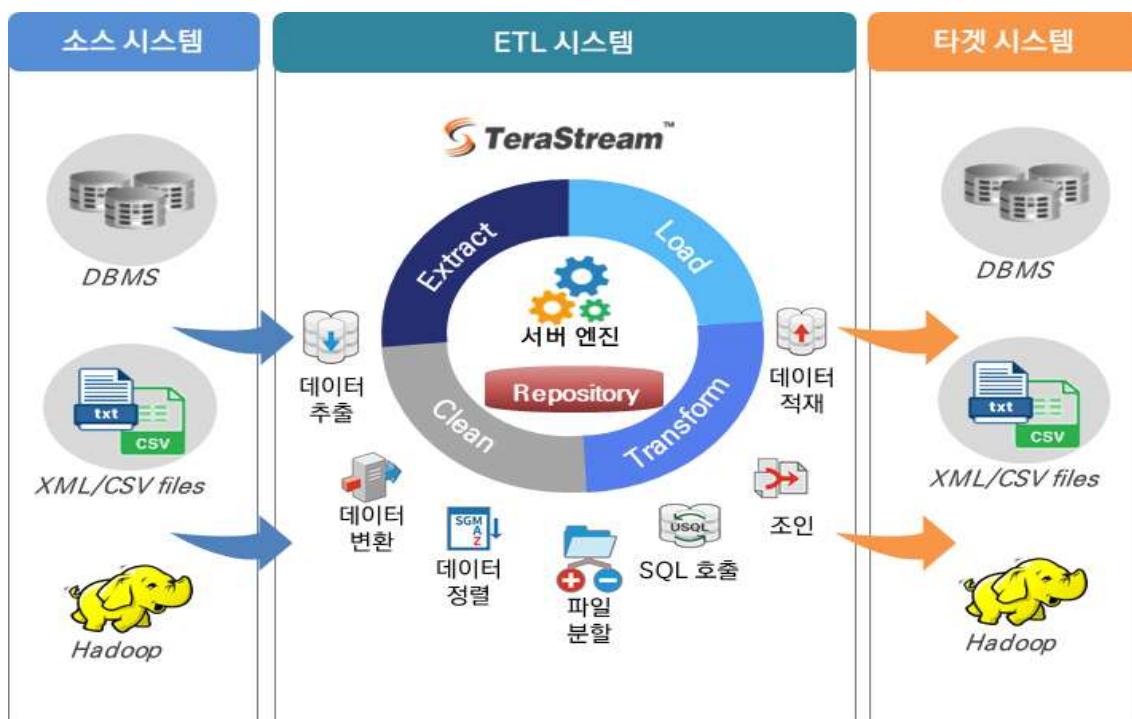


그림 5 TeraStream™ 아키텍처

다) 특장점

구분	설명
고성능	<ul style="list-style-type: none"> - 대용량 데이터 고속 추출 - File 처리 방식으로 대량의 ETL 작업을 빠르게 처리 병렬 프로세싱을 통한 데이터 가공으로 작업 속도 향상
자원 이용시 고효율성	<ul style="list-style-type: none"> - DB와 분리된 시스템 자원을 이용하여 DB부하 감소 - On-Line DB 작업시 자원에 부하를 주지 않고 Batch작업 가능 CPU, Memory, I/O의 효율적인 자원 할당 가능
높은 업무 생산성	<ul style="list-style-type: none"> - 통합 GUI 환경을 제공하여 프로그램 개발 편리 - 스크립트를 이용하여 타 프로그램에서 Sort 엔진, 고속 추출 엔진(FACT)실행 SQL개발자의 개발생산성을 높여 개발비용 절감과 개발 품질 상향 평준화
효율적 기능제공	<ul style="list-style-type: none"> - 로직 강화로, ETL외 업무 영역 사용 확대(온라인 batch등) - 준 실시간 데이터에 적용
높은 호환성	<ul style="list-style-type: none"> - TeraStream™이 설치된 서버간의 분산 스케줄링 및 통합 스케줄링 지원 - 분산된 데이터 통합 솔루션 환경에서 통합 Metadata 관리 지원
통합관리환경 제공	<ul style="list-style-type: none"> - 응용 프로그램 - 타 ETL, EAI 솔루션과 호환

라) 차별화 요소



그림 6 TeraStream™과 DBMS 처리 시간 비교에 따른 성능 우위

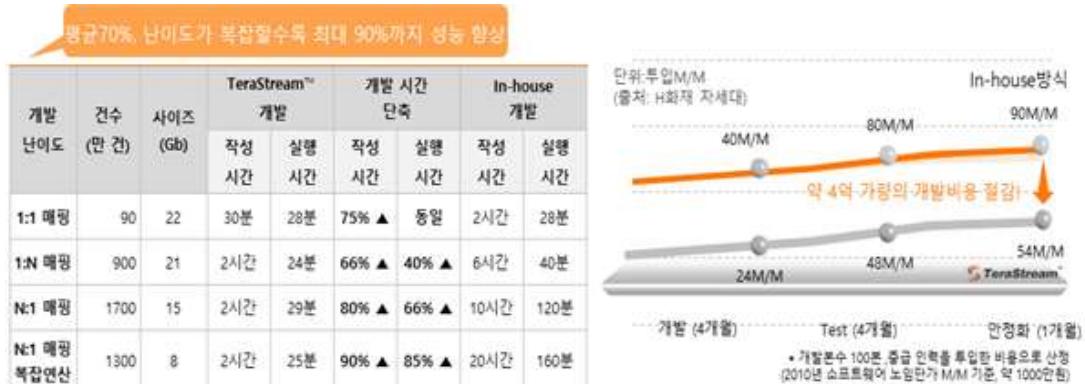


그림 7 TeraStream™ 적용시 비용절감

2) 제품 적용기술 내역

가) 고속 데이터 추출 및 저장 기술

- 다양한 원천 데이터로부터 가능한 적은 컴퓨팅 자원을 이용하여 고속으로 데이터를 추출하고 이를 파일시스템에 저장하는 기술입니다. DBMS의 경우 Low Level API와 많은 Record를 한번에 추출하는 Multi-Record 추출 기술이 포함됩니다. 데이터 저장 역시 Record 단위 저장이 아닌 Buffer에 일정 데이터를 모아 Block 단위로 데이터를 저장하여 파일시스템의 성능을 최대한 활용합니다.

나) 고속 데이터 변환 기술

- 다양한 변환 함수를 제공하면서 컴퓨팅 파워를 최대한 활용하여 고속으로 데이터 변환을 처리하는 기술입니다. 분산처리, 병렬처리 및 사용자 정의 변환을 지원하기 위한 동적 변환 엔진 생성 등의 기술과 함께 파일로부터 데이터를 읽고, 쓰는 성능을 확보하기 위한 File I/O 최적화 기술이 포함되어 있습니다.

다) 고속 데이터 소트 및 조인 기술

- 데이터의 통합을 위해 서로 다른 형태의 데이터간 통합을 지원하는 조인은 필연적으로 데이터의 소팅을 필요로 합니다. 조인은 파일 기반 처리에서 가장 컴퓨팅 파워를 많이 사용하게 되는 기술로 최소의 자원으로 최대의 성능을 내야하는 최적화 기술이 핵심입니다. 최적화 기술에는 메모리 보다 큰 데이터의 정렬을 하는 External Sort, External Sort에서 발생하는 임시 데이터의 Random File Access의 최적화 기술이 포함되며 소트를 기반으로 한 Sort Join 엔진이 파일 기반 변환의 핵심 기술입니다.

라) 고속 데이터 적재 기술

- 고속 데이터 적재를 위해서는 다양한 Target에 대한 데이터 전송 최적화 기술이 핵심입니다. DBMS의 경우 다양한 DBMS 지원 및 여러 Record를 한번에 전송하는 Bulk 적재 기술이 핵심입니다.

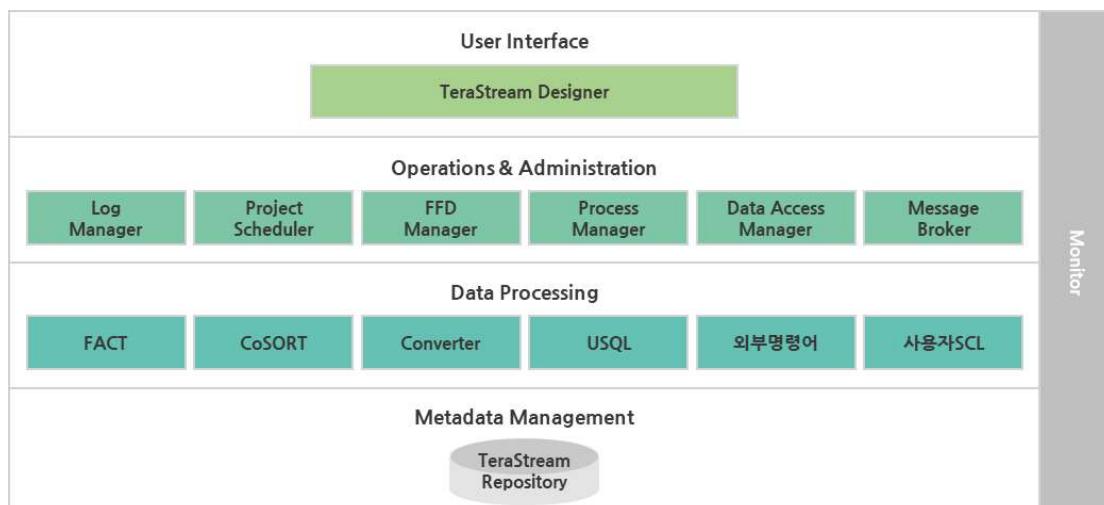


그림 8 TeraStream™ 적용기술

3) 주요 기능

대메뉴명	소메뉴명	주요 기능 설명
연계	3-Tier연계	ODBC 연동이 아닌 3-Tier 방식의 연계 지원

ETL	DBMS 추출	ORACLE, SYBASE, Teradata, Nettezza, Greenplum, Tibero, Altibase, DB2, SQLServer, MySQL, Altibase, RedShift, PostGreSQL, GoldiLocks, CuBrid, InforMix, Dynamic365, Beeline, Vector 지원 고정형/가변형 추출 지원 DBMS Table 내에 개행 문자 지원 Date Type 및 TimeStamp 형식 지원
		입/출력에 대한 파일 처리 지원 입력 파일을 출력 DBMS에 데이터 처리 지원 DBMS Table 정보를 File 형식으로 변환하여 지원 소스 DBMS Table에서 목적 DBMS Table로 변환하여 지원 데이터 변환 시 내장 함수 지원 데이터와 데이터간의 파일 처리 기능 제공 사이즈, 건수를 통한 데이터 파일 분할 기능 제공
		SQL 쿼리 기능 지원 DBMS Procedure 호출
		ORACLE, SYBASE, Teradata, Nettezza, Greenplum, Tibero, Altibase, DB2, SQLServer, MySQL, Altibase, RedShift, PostGreSQL, GoldiLocks, CuBrid, InforMix, Dynamic365, Beeline, Vector 지원 Bulk Load 적재 방식 지원 FILE에 개행 문자 생성시 처리 방안 지원 DBMS 적재 방식 시 날짜 형식 지원 DBMS에서 제공하는 적재 방식 옵션 지원 DBMS에 대한 INSERT/UPDATE/DELETE/UPSERT 기능 지원
	스케줄링	예약실행 기능 연/월/일, 시/분/초 단위의 예약실행 기능 지원
		영업일 등록 기능 영업 기준 등록 기능 지원
	모니터링	초당 처리 건수 데이터 초당 처리 건수 기능 지원
		실행/종료 시간 블록 실행/종료 시간
		수행 시간 프로젝트에 대한 수행 시간 제공
		블록 진행률 막대 그래프 형식의 작업 진행 상황 표시 기능
		이벤트 알림 프로그램에 대한 오류 이벤트 알람
		프로젝트 전체 모니터링 ETL 프로젝트에 대한 Dashboard 기능
	외부 연계	데이터 전송 서버와 서버간의 데이터 전송 기능 지원
		데이터 압축 데이터 발생시 압축 기능 제공
		외부 스케줄 연계 Control-M, JCL 등 외부 스케줄과의 연계 기능 제공
	영향도 분석	FFD 컬럼 변경시 영향도 대상 테이블 컬럼 변경시 영향도 버전 기능 FFD 컬럼 변경 시 영향도 대상 블록 리스트 제공
		추출 및 적재 대상 테이블 컬럼 변경시 영향도 리스트 제공
		신규 버전과 이전 버전 수정 상황 제공
	보안	계정 ETL 접속 계정 관리 제공 비밀번호 Network 전송 암호화 제공
		외부 보안 API연계 외부 제품과의 보안 API 연계 제공
		로그 프로젝트 실행 로그 제공 블록 실행로그 제공
	로그	이벤트 로그 프로젝트 접속 계정 로그 제공

4) 적용사례

사업명 : 국민은행 EDW 통합 DM 구축(1, 2 차 프로젝트)	
주요 요구사항	<ul style="list-style-type: none"> - M/F 및 IMS HDS의 컨버전 기능 요구 - 시계열 컬럼 부재 상황에서 변경분 처리 - 대용량 데이터의 배치시간내 처리(원천기준 : 일 10TB) - 용량 다일 파일의 병렬 처리
수행 내용	<ul style="list-style-type: none"> - 메인 프레임 환경 데이터를 UNIX 환경 데이터로 전환 - 25TB의 데이터를 18 시간에 처리 - 데이터의 한글 변환 등 다양한 데이터 변환 가공 - 계정 서버에서 신ODW 서버로 ETL <ul style="list-style-type: none"> • Daily 변경분 데이터 약 200GB에 대해 TeraStream FACT 엔진을 이용하여 약 1.5 시간 내에 추출 - 정보계 시스템에서의 ETL 및 Batch 업무 수행 내역 <ul style="list-style-type: none"> • 수신, 여신, 외국환, 통합, 공토정보계 업무 • OLAP Mart 구축 - 정보계 Batch를 작업 기준 목표시간(6 시간) 내에 적재 <ul style="list-style-type: none"> • 주요 작업을 목표시간 내에 완료하여 요구 부서에 제공 OLAP Mart 구성
도입 효과	<ul style="list-style-type: none"> - 정보계 업무의 다양한 데이터베이스(IMS HDB, HOST DB2, Oracle, DB2, UDB 등)를 TeraStream을 이용하여 처리 다양한 비즈니스 로직이 적용되어 있는 대용량 데이터(일 EBCDIC 4TB)의 일 Batch 목표시간(2.5 시간) 만족
시스템 구성도	<p>The diagram illustrates the data flow from various mainframe and server databases (IMS HDB, IBM M/F, HDB, DB2, Server RDB) through Informover and TS(FACT) to Flat Files, then via ETL to Sybase ASIQ's A-SOR, and finally to the EDW's 영역DM and 통합DM. The flow is categorized by file processing type (Flat File) and database query (DB QUERY). The Sybase ASIQ section includes A-SOR and DM components.</p>

나. 실시간 변경분 데이터 처리 (DeltaStream™)

1) 제품 특징 및 차별화 요소

가) 제품의 개요

- CDC(Change Data Capture) 방식을 적용하여 실시간으로 데이터를 처리하는 솔루션으로, DBMS의 데이터 변경 정보(Transaction Log)를 추출하여 목표(Target)시스템에 전송합니다.

나) 개념도

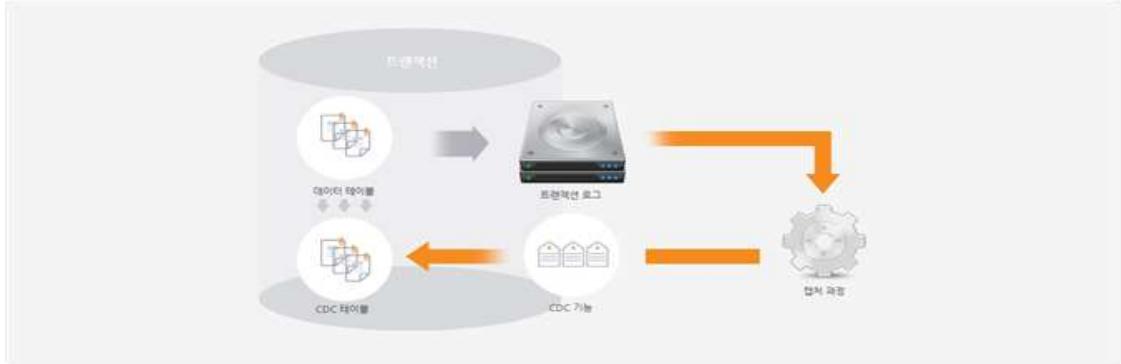


그림 9 DeltaStream™ 개념도

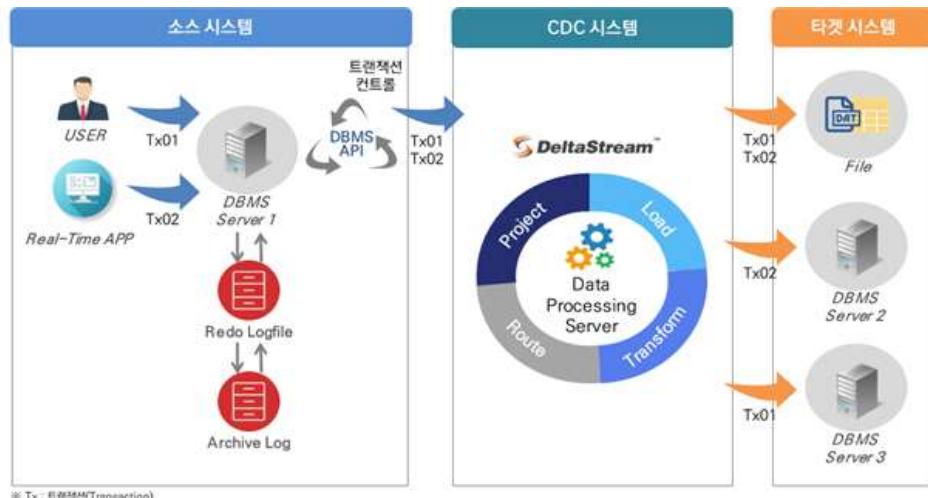


그림 10 DeltaStream™ 아키텍처

다) 특장점

구분	설명
고성능	<ul style="list-style-type: none"> - 병렬 처리 엔진을 통한 고속 변경 데이터 추출 제공 - 고속 추출 모듈[FACT]을 통한 초기 데이터 이행 지원(SnapShot)하고, 변경 데이터는 SAMFILE 변환 DML 정보의 Audit 로깅 기능을 제공하여 해당 트랜잭션의 고속 동작 변환 지원
다양한 트랜잭션 추출 모드 지원	<ul style="list-style-type: none"> - 2가지 모드 (세부적으로 3가지 방법)의 DBMS의 트랜잭션 로그 Data 추출 방법 지원 <ul style="list-style-type: none"> • API를 이용하는 방법: DBMS가 제공하는 Log 추출 API를 이용하여 트랜잭션 변경사항을 추출 • 로그를 이용한 방법: Remote Agent 이용(DeltaStream의 Remote 추출 Agent를 이용), Log Sync Agent 이용(DeltaStream의 Remote 추출 Agent를 이용하여 DBMS Log를 DeltaStream 서버로 복제하여 트랜잭션 변경 분 추출)
사용자 편의성 극대화	<ul style="list-style-type: none"> - GUI를 기반의 Designer Tool을 적용하여 개발의 편의성 제공 - 사용자 웹 모니터링 기능을 제공하고 장애 발생 시 이벤트 알림(SMS, SMTP 연동) - ETL 연동 위한 Data 변환 기능 제공

업무 효율성 지원

- 추출 기반의 라이선스 정책으로 타겟 서버 종설 시에도 추가 비용 없음
- 일/월 변경 데이터 처리 건수 모니터링
 - 소스 테이블과 타겟 테이블 간의 Data Validation
 - 실시간 데이터 연동 이력관리 제공

2) 제품 적용기술 내역

가) Transaction Log 분석 기술

- 변경 데이터를 인지하고 이를 변화 전 데이터와 변화 후 데이터를 추출하여 변경내용을 파악하고 이를 변경 SQL로 변환하는 기술로 Transaction Log가 일반 Text 데이터가 아니라 DBMS가 스스로 데이터를 복구하기 위한 데이터이므로 이를 해석하여 정합성을 확보하는 것이 가장 어렵고 중요한 기술입니다.

나) Log 동기화 기술

- 원천 데이터 시스템에서 변경데이터를 추출하는 것은 원천 데이터 시스템에 부하를 주기 때문에 별도의 분석서버로 로그의 변화된 내용만 전송 동기화하는 기술이 필요하며 이는 매우 짧은 시간 대량의 데이터도 동기화할 수 있어야 하며 데이터 누락과 중복이 허용되지 않아 핵심 기술입니다.

다) 변경 데이터 실시간 적재 기술

- 변경 데이터를 추출한 후 이를 타겟 데이터 저장소로 적재하는 기술로 데이터 통합 기술과 같이 많은 데이터를 동시에 전송하는 기술이 포함되며 동시에 변경 데이터는 데이터 적용의 순서가 매우 중요하여 데이터 순서보장 기술도 함께 적용됩니다.

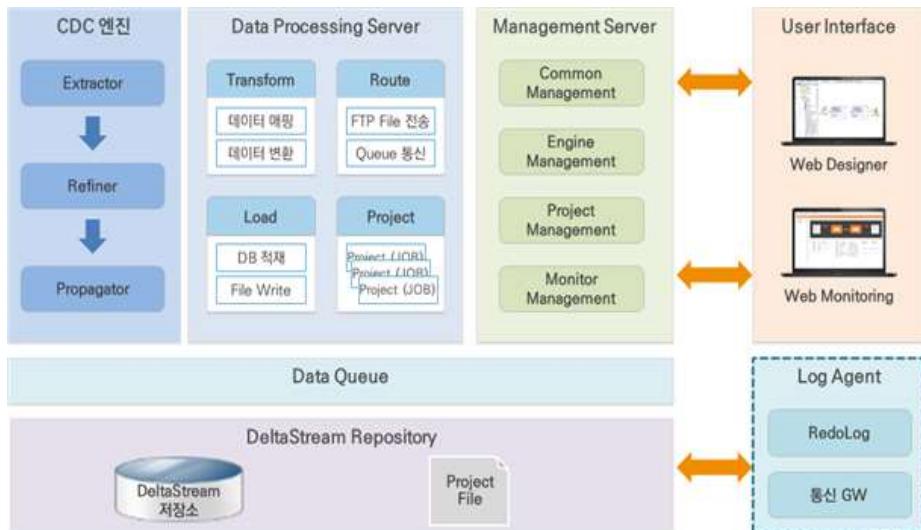


그림 11 DeltaStream™ 적용기술

3) 적용사례

사업명 : 신한은행 eCRM 구축

사업명 : 신한은행 eCRM 구축	
주요 요구사항	<ul style="list-style-type: none"> - 전자금융시스템 <ul style="list-style-type: none"> • 4node Oracle RAC으로 대량의 거래 트랜잭션이 발생하는 시스템 • 30 분 주기로 아카이브 로그 백업 후 삭제 - CRM 서비스를 위해 전자금융시스템의 트랜잭션 데이터가 필요 - 기존 시스템에 영향이 없어야 함
수 행 내 용	<ul style="list-style-type: none"> - Log 추출 방식 업그레이드 적용

	<ul style="list-style-type: none"> • DBMS Redo Log 파일을 실시간으로 델타스트림 서버로 이관하여 직접 추출하는 방식 적용 - 초기 적재 : 총 57 개 테이블의 234GB 데이터를 90 분 이내에 적재 완료 - CDC 동기화 : 60 개 테이블 CDC 동기화 - 델타스트림 적용 관련 <ul style="list-style-type: none"> • 초기 적재 이후 동기화 안정시 CPU 사용량은 10% 정도 • 델타스트림 서버의 가용성을 높이기 위해 HA(Active-Standby)로 구성
도입 효과	<ul style="list-style-type: none"> - 전자금융시스템에서 CRM시스템으로 실시간 데이터 전송 - 전자금융시스템은 대량 트랜잭션이 발생하는 서버로 아카이브 로그 백업주기 이전 동기화 - 안정적인 데이터 동기화 가능
시스템 구성도	<p>The diagram illustrates the system architecture for data warehousing. It shows data flowing from various sources (IMS HDB, IBM M/F, HDB/DB2, Server RDB) through Informover and TeraStream ETL to Sybase ASIQ's A-SOR and Synapse DM. The data is stored in Flat Files and accessed via DB QUERY.</p>

다. IoT 데이터 실시간 처리 솔루션 (TeraStream BASS™)

1) 제품 특징 및 차별화 요소

가) 제품의 개요

- TeraStream BASS는 다양한 장비의 로그 데이터, 각종 센서 데이터, 관계형 DBMS의 정형 데이터를 고속 수집 저장하고, 수집된 데이터를 실시간 분석정보를 제공하여 의사결정을 지원하는 플랫폼

나) 개념도

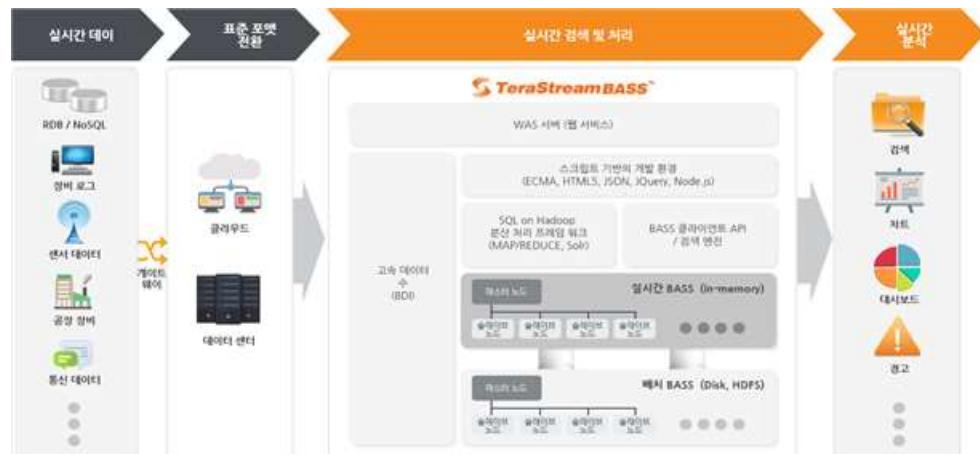


그림 12 TeraStream BASS™ 개념도

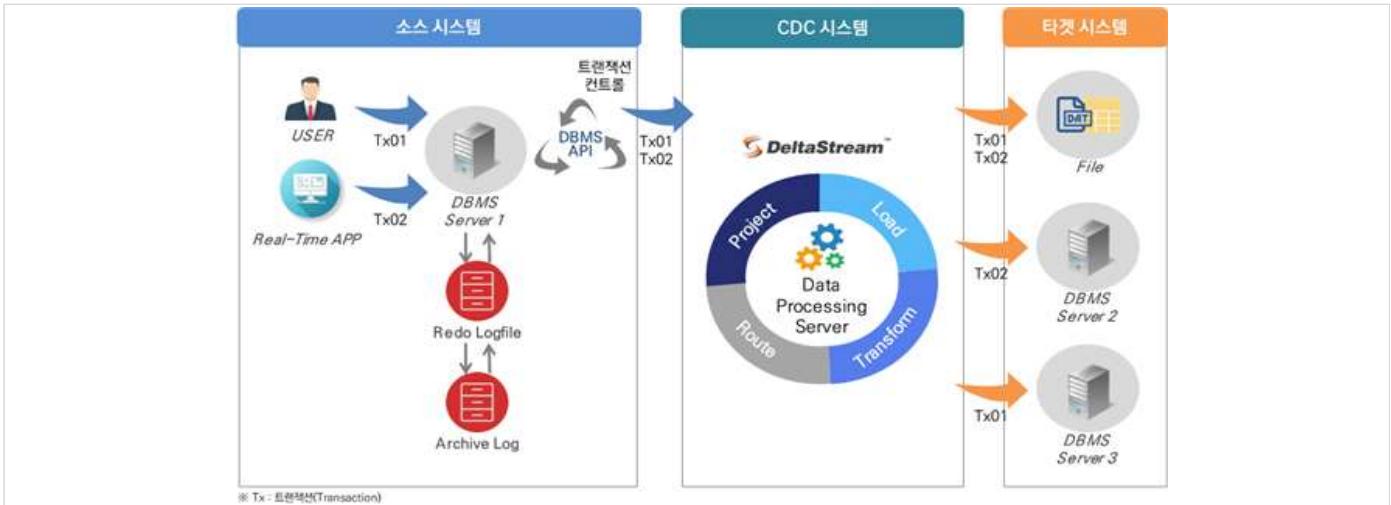


그림 15 TeraStream BASS™ 아키텍처

다) 특장점

구분	설명
인메모리 실시간 데이터 처리 기능 제공	<ul style="list-style-type: none"> - 메모리 기반으로 마스터 노드와 데이터 노드를 구성하여 고속 인덱싱 및 분산 저장 기능 구현 - 마스터 노드의 HA 구성 이중화로 안정성 제공 - Fail Over를 위한 메모리 복제 기술 - 자체 인덱싱 기술을 적용하여 고속검색
표준 Agent API 제공	Data의 특성에 따라 관리하는 통신 방식이나 인터페이스가 다양하지만, BASS의 표준 Agent 기술은 적은 비용으로 종류에 무관하게 데이터 수집 가능
Memory와 HDFS 인덱싱	Memory 및 HDFS 데이터를 초고속 검색(2. 핵심기술 참조)
Memory 버퍼링 기능	<ul style="list-style-type: none"> - Thread based connection pool 방식과 연계된 Memory 버퍼링 기능 제공 <ul style="list-style-type: none"> • 10 ~ 15Gbps의 속도로 수집 및 저장
데이터 연계 Hybrid 기술 적용	<ul style="list-style-type: none"> - Memory & HDFS 데이터의 연계 Hybrid 기술을 적용 <ul style="list-style-type: none"> • Memory 데이터와 HDFS 데이터의 유연한 검색 및 분석 가능
메모리 복제 기술 적용	<ul style="list-style-type: none"> - 메모리 복제 기술을 적용하여 시스템 장애 발생하여도 작업의 유연한 대처 가능

2) 제품 적용기술 내역

가) 분산 인메모리 저장소 기술

- 여러 서버로 구성된 Cluster의 메모리를 하나의 저장소로 구성하고 서버의 장애 발생 시에도 데이터 유실이 발생하지 않도록 유지 관리하는 기술입니다.

나) 초고속 데이터 전송 기술

- 노드당 200만건의 초고속 데이터 수집을 위해서는 수집엔진의 전송속도와 메모리 저장소에 저장하는 두 기술이 모두 초고속 처리 속도를 확보하는 것이 중요하며 이 기술이 초고속 저장의 핵심입니다.

다) 실시간 검색을 위한 독자 색인 기술

- 연관 데이터를 다시 검색하지 않고 검색의 최대 성능을 확보하기 위해 기존의 DBMS에서 많이 활용되는 B+ Tree에 복합 색인인 경우 두 색인이 서로 연관관계를 저장하여 검색 성능을 확보합니다.

다. 이는 독자 기술이며 특허 확보 기술입니다.



그림 14 TeraStream BASS™ 적용기술

3) 주요 적용사례

사업명 : 한국도로공사 시설물정보 분석 시스템 구축	
주요 요구사항	<ul style="list-style-type: none"> 시설물 관리 센터의 경우, 매년 유지관리 비용은 증가하고 있지만, 수집된 센서 정보 분석기능은 부재 센서 장비 공급 업체 대부분이 영세하여 납품된 센서 데이터의 표준화 및 통신 표준화 필요
수행내용	<ul style="list-style-type: none"> 데이터 표준화 및 통신 규약 표준화 작업. 센서 → PLC → 중계기 → 통신 서버 간의 센서 데이터 기술 기준 정의 <ul style="list-style-type: none"> PLC 데이터 통신을 위한 데이터 표준화 및 통신 규약 표준화 IoT 플랫폼구축 <ul style="list-style-type: none"> 다양한 센서 데이터 고속 수집 /저장/분석을 위한 플랫폼 구축 관련 시스템(관제, 통합 상황, 빅데이터 등)에 수집 데이터 연계 재난 상황 예방 관리를 위한 이벤트 프로세싱 기술 적용 스트리밍 데이터 분석, 실시간 대시보드, 쿼리 분석 등
도입효과	<ul style="list-style-type: none"> 공사 터널, 공사 현장, 염료 등 다양한 센서 관리 통합 PLC 및 통신 서버 등의 데이터 표준 및 통신 규약 표준화 실시간 센서 데이터 수집 및 모니터링을 통해 재난/재해 예방
시스템 구성도	

2. 빅데이터 플랫폼(Big Data Platform) 기술

가. 빅데이터 저장소 (TeraONE™ HDFS)

1) 제품 특징 및 차별화 요소

가) 제품의 개요

- 데이터 거버넌스 기반의 빅데이터 플랫폼 아키텍처로, 데이터의 품질관리, 메타데이터 관리, 마스터 데이터 관리를 비롯하여 실시간 빅데이터 분석과 시각화, IoT 데이터 분석 등을 가능하도록 구현한 솔루션입니다.
- 기업 내부의 대용량 정형 데이터,内外부 정형·비정형 데이터, 실시간·대용량 배치 데이터 등 다양한 형식의 빅데이터 처리 등을 통합 수용할 수 있습니다.
- 특히 데이터 통합 및 저장, 처리, 분석 과정을 하나로 연결한 통합 분석이 가능하기 때문에 기업 내부에 적용하면 기존 데이터웨어하우스(DW)의 분석을 넘어 빅데이터 시스템으로 확장이 가능해 운영비용을 최소화할 수 있습니다.

나) 개념도



그림 15 TeraONE™ 개념도

다) 특장점

구분	설명
분산 메모리 기반 실시간 데이터 통합 아키텍처 구현	<ul style="list-style-type: none"> - 빅데이터 처리 운영 인프라 복잡성 해소 - 다양한 데이터를 쉽고 빠르게 수집 및 저장하는 분산 메모리 저장 기술을 적용 - 초기 도입 비용을 최소화한 고성능/고효율의 데이터 통합 아키텍처 구현 (외산 S사 솔루션 대비 4~7 배 빠른 실시간 데이터 수집 성능 제공)
통합 데이터 거버넌스 체계 확립	<ul style="list-style-type: none"> - TeraONE™의 거버넌스 체계를 통하여 수집된 데이터(실시간/배치, 정형/비정형)의 품질 개선 - 품질 개선된 양질의 데이터를 확보하여 분석의 신뢰성 향상
통합 GUI 적용으로 쉬운 사용과 운영 환경 제공	<ul style="list-style-type: none"> - 사용자 친화적인 GUI 환경으로 개발/운영 담당자에게 높은 생산성과 운영 효율성 제공 - GUI를 통한 편리한 Platform의 통합 운영 및 모니터링 기능 제공
분석 데이터 증가에 따른 확장 용이성 제공	<ul style="list-style-type: none"> - 데이터는 Memory 클러스터와 HDFS에 저장으로 높은 성능과 Scale-Out이 우수한 노드 확장성 제공

2) 제품 적용기술 내역

가) 빅데이터 통합 관리 기술

- 1000개 이상의 대규모 서버로 구성된 클러스터에서 동적으로 서버의 추가, 소프트웨어의 추가 설치 및 HA, 장애 및 이상 관리 등의 통합 관리 기술입니다.

나) 빅데이터 서버 관리 기술

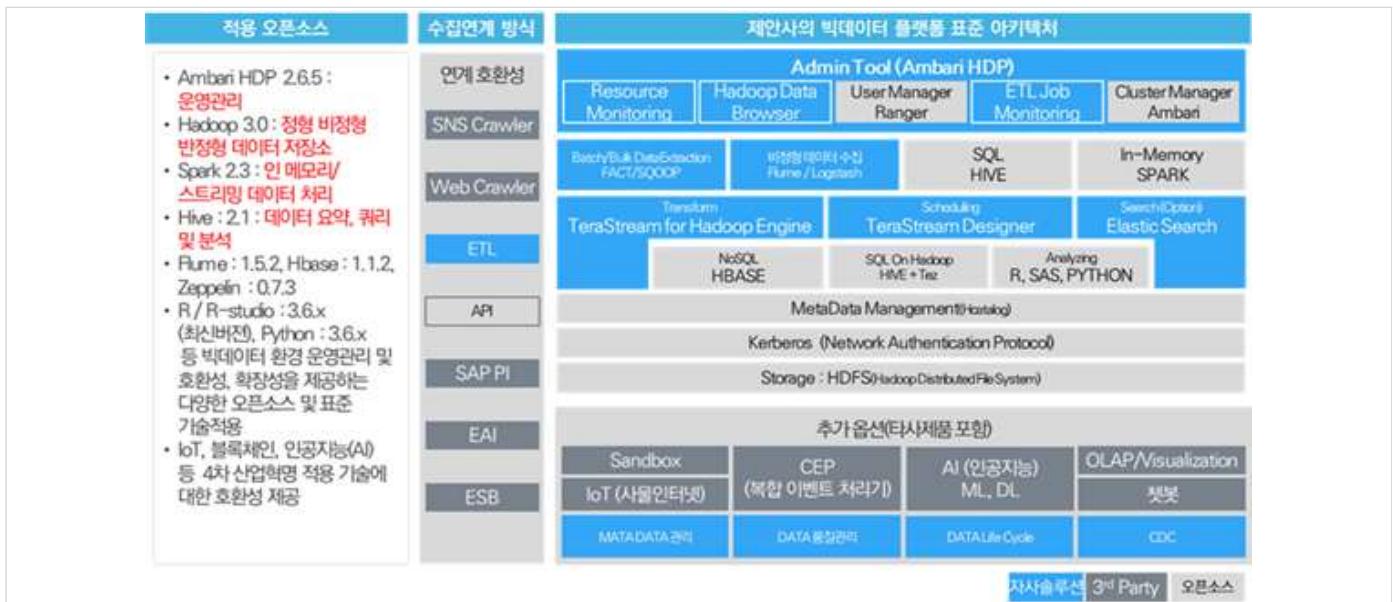
- 각 서버의 장애나 이상에 대한 관리, 서버에 대한 모니터링을 하기 위한 기술이며 서버의 이상이 발생하기 전 사용자가 사전 인지하여 대응할 수 있도록 하는 장애 사전 인지 기술이 포함되어 있습니다.

다) 빅데이터 소프트웨어 관리 기술

- 30여개 이상의 다양한 소프트웨어의 상호 연관성, 의존성에 대해 자동으로 설치, 삭제, 관리를 하는 기술로 당사의 소프트웨어 뿐만 아니라 다양한 오픈소스의 관리 자동화 기술이 포함되어 있습니다.

라) 빅데이터 사용자 및 권한 관리 기술

- 시스템에 접근하는 사용자에 대해 식별하고 데이터의 접근에 대해 통제하는 기술로 데이터보호의 핵심 기술입니다. 30여개 이상의 소프트웨어와 연계하여 포함된 모든 소프트웨어가 단일 사용자 관리 체계, 단일 권한 관리 체계로 통합되도록 하여 통일된 권한 관리 체계를 유지하고 데이터의 접근을 기록하여 향후 감사 등 사후 관리를 지원하는 기술입니다.



3. 인공지능(AI) 및 애널리틱스(Analytics)를 위한 분석 플랫폼 기술

가. 인공지능 기반 빅데이터 분석 환경 (TeraONE IDEA™)

1) 제품 특징 및 차별화 요소

가) 제품 개요

- AI 분석환경인 Jupyter Notebook 기반으로 머신러닝 및 딥러닝 알고리즘을 빌트인하여 제공하고 있습니다. (Scikit-Learn, Theano, Tensorflow, Keras 등)
- Python 자체 패키지와 표준 라이브러리 기반으로 수치해석 및 통계, 리포팅, 텍스트 분석, 머신러닝 및 딥러닝 분석이 가능하고 ELK(Elastic Search, Logstash, Kibana), Superset 기반의 시각화 모듈도 제공합니다.

나) 개념도



그림 17 TeraONE IDEA™ 개념도



그림 18 TeraONE IDEA™ - 시각화 이미지 예시

다) 특장점

구분	설명
기본 분석환경 제공	<ul style="list-style-type: none"> - Python 기본 분석 도구 및 분석 환경 제공 (Jupyter Notebook) - 분석 인프라에 대한 모니터링 및 관리 기능 제공 - 기본 분석 라이브러리 내장, 추가 필요 라이브러리 설치 지원
분석 패키지 제공	<ul style="list-style-type: none"> - 수치 해석 및 통계 패키지 <ul style="list-style-type: none"> • Numpy : 다차원 배열과 이에 대한 수학적 연산 • Scipy : 선형대수, 희소행렬, 신호 및 이미지 처리, 최적화 알고리즘 • Pandas : 빅데이터를 위한 수치연산, 데이터 프레임 및 시계열 처리 - 텍스트 분석 패키지 <ul style="list-style-type: none"> • Gensim : 병렬 분산처리 문자 집합 분석, 주제 모델링, 텍스트 벡터 - 기계학습 및 딥러닝 <ul style="list-style-type: none"> • Scikit-Learn : 학습 데이터 전처리, 지도 및 비지도 학습, 모델 선택 및 검증 • Theano : 빌딩 블록형 딥러닝 모델 • Tensorflow : 다차원 데이터 배열 기반 딥러닝 모델 • Keras : 고수준의 딥러닝 모델 - 리포팅 <ul style="list-style-type: none"> • Matplotlib : 배열로부터 Plot을 생성하고 대화형 시작화 지원
정형 데이터 시작화	<ul style="list-style-type: none"> - Apache Superset을 기반으로 정형 데이터 시작화 제공 <ul style="list-style-type: none"> • 사용자 친화적 UI, 데이터 탐색, 시작화의 높은 반응속도 • 대시보드 작성 및 공유
반정형/비정형 데이터 시작화	<ul style="list-style-type: none"> - ELK (ElasticSearch/Logstash/Kibana) 기반으로 검색 기반 데이터 탐색 및 시작화 <ul style="list-style-type: none"> • 고속 Searching 기반의 데이터 탐색 및 시작화 • 반정형/비정형 데이터의 시계열 분석 및 위치 분석 • 검색엔진의 연관성 기능, 그래프 탐색을 활용한 관계 분석

2) 제품 적용기술 내역

가) 분석 환경 구성 기술

- Cloud Native 기반의 컨테이너 가상화 기술을 이용하여 사용자가 요구하는 분석환경을 동적으로 프로비저닝 하는 기술. 사용자의 요구사항에 따라 가상환경을 구성하고 필요한 소프트웨어를 자동으로 구성하여 함께 프로비저닝하는 기술이 포함

나) 분석 환경 운영/관리 기술

- 분석 환경이 한 명의 개인을 위한 것이 아닌 다수의 팀, 다수의 사용자가 사용하는 환경을 독립된 환경으로 제공하는 기술로 구성된 분석환경을 효율적으로 운영하고 관리하는 기술입니다. 특히 한정된 자원을 여러 분석 환경에서 공유하여야 하므로 제한된 자원의 배분, 스토리지 자원의 관리, 컨테이너 및 독립된 공간의 보안, 사용자 인지 및 권한 등 분석환경의 운영에 관련된 기술이 포함됩니다.

다) API 동적 운영 관리 기술

- 사용하는 분석 환경이 운영 서비스로 활용될 경우 API를 이용한 데이터 제공 및 분석 결과 전달에 활용되므로 제공되는 API에 대한 자동 수집, API Gateway 기술을 활용한 API 자동 문서 생성, 사용자 API 활용 등 API의 통합 운영, 사용에 관한 기술입니다.

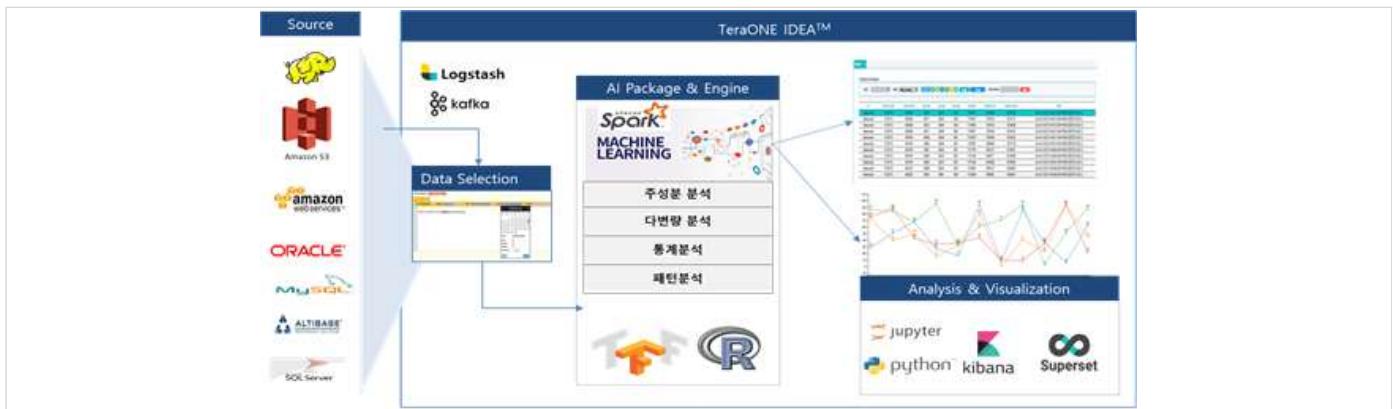


그림 19 TeraONE IDEA™ 기술

3) 주요 적용사례

사업명 : 인공지능을 활용한 케이블레인시스템 구축	
주요 요구사항	<ul style="list-style-type: none"> - 인공지능 기술을 활용하여 경륜·경정 업무 전반의 효율성 향상 - 국민적 관심이 집중되고 있는 기술 활용을 통한 홍보 효과 극대화 - 인공지능 기술을 경정 편성업무에 우선 적용하여 기술적 타당성 검증
수행내용	<ul style="list-style-type: none"> - 인공지능기술 활용을 위한 기반 구축 <ul style="list-style-type: none"> • 경정 편성자동화 이후 인공지능을 확산 적용할 수 있는 기반 구축 - 경정 편성자동화 시스템 구축 <ul style="list-style-type: none"> • 딥러닝을 활용한 모델 구축 • 예상 배당률 산정 모델 구축 • 데이터 분석시스템 - 내부 전문가 양성 <ul style="list-style-type: none"> • 시스템 유지보수 및 인공지능 기술 사내 확산 가능한 내부 전문가 양성
도입효과	<ul style="list-style-type: none"> - 공단의 인공지능 기술 도입을 위한 기반 마련 - 편성업무의 생산성 향상 - 담당자 교체 시 일정 수준 이상의 편성 품질 보장 - 인공지능 기술 활용을 통한 홍보효과
시스템 구성도	<p>The flowchart details the system configuration process:</p> <ul style="list-style-type: none"> 선수, 보트, 모터 확장: Includes 72명의 선수, 16경기 72명 선수 확장, 선수, 보트, 모터. 편성 담당자 조건 입력: Includes 72명의 선수, 16경기 72명 선수 확장, 선수/보트/모터 기술 및 고급화. 예측(통계분석, 딥러닝): Includes 선수 기량 고급화 설정, 선수+보트 고급화 설정, 선수+모터 고급화 설정, 선수+선수 고급화 설정, 선수+보트 고급화 설정, 선수+모터 고급화 설정. 필수요건 필터링: Includes 고급 필터, 고급 필터 조건 설정, 고급 필터 조건 설정, 고급 필터 조건 설정, 고급 필터 조건 설정. 출주표 생성: Includes 고급 필터 조건 설정, 고급 필터 조건 설정, 고급 필터 조건 설정, 고급 필터 조건 설정. 도입방법: Includes 자동화, 자동화, 자동화, 자동화, 자동화, 자동화. 경쟁률 선수 조합 생성: Includes 조합형 필터링 (필수조건), 가능 출주표 생성 (모든 조건), 조합형 필터링 (필수조건), 가능 출주표 생성 (모든 조건), 조합형 필터링 (필수조건), 가능 출주표 생성 (모든 조건). 수정된 출주표 생성: Includes 조합형 필터링 (필수조건), 가능 출주표 생성 (모든 조건), 조합형 필터링 (필수조건), 가능 출주표 생성 (모든 조건). <p>Legend: <ul style="list-style-type: none"> → 기획수립/편성 → 자동화 </p>

4. 데이터 패브릭(Data Fabric) 기술

가. 데이터 가상화 (TeraONE SuperQuery™)

1) 제품 특징 및 차별화 요소

가) 제품의 개요

- 다양한 이기종 원천 데이터 소스 커넥션 정보 설정 가능하여, 등록된 데이터 소스에 대해서는 이기종이더라도 하나의 데이터 소스처럼 One Query로 분석할 수 있는 데이터 가상화 솔루션입니다.
- 기존의 데이터웨어하우스와는 달리 물리적 또는 논리적 데이터 모델 없이도 데이터의 물리적 이동을 최소화하여 통합하는 신기술을 적용하였습니다.

나) 개념도

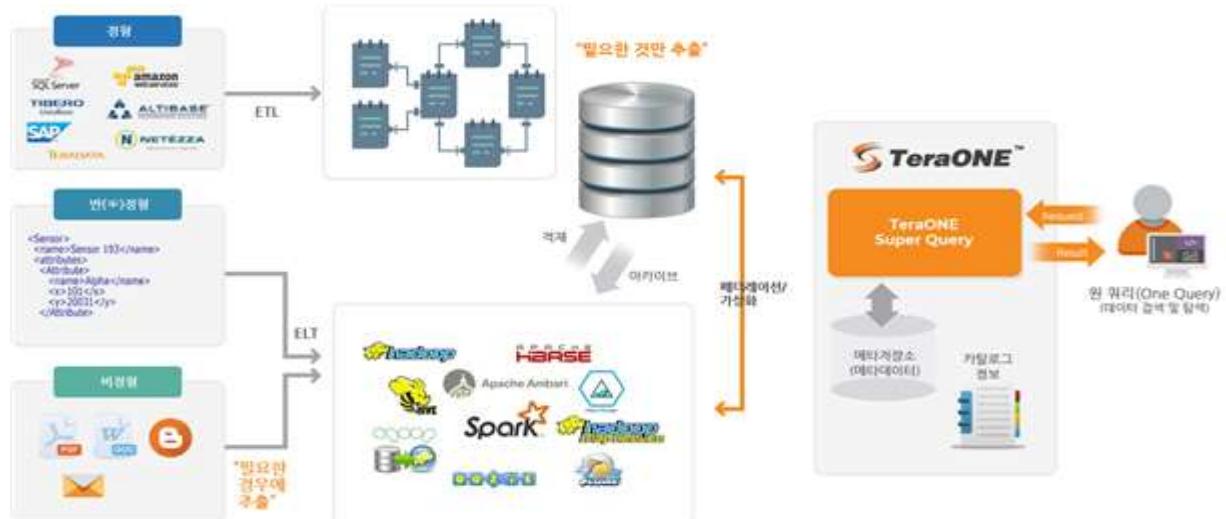


그림 20 TeraONE SuperQuery™ 개념도



그림 21 TeraONE SuperQuery™ - 가상화 화면 예시

다) 특장점

구분	설명
이기종 DBMS 가상화	서로 다른 DBMS 데이터 뿐 아니라 HDFS, Hve 등 백데이터 및 클라우드 데이터 소스를 하나의 공통 DB내에 있는 것처럼 가상화해 처리함으로써 물리적인 데이터 통합 과정 생략
빅 데이터 연계	대용량 하둡(Hadoop) 데이터를 RDBMS 데이터와 통합

데이터 통합 자동화	분석기가 실행한 SQL를 자체 분석하여 자동으로 소스 DB로부터 데이터를 추출, 통합하는 기능을 제공
다양한 분석환경 지원	JDBC 인터페이스를 이용해 일반 웹 프로그램 및 OLAP/BI/시각화 도구와 연계하여 SQL 쿼리 또는 분석 결과 화면 즉시 구현 가능
고속 SQL 기능 지원	인메모리 분산 처리 기능으로 대상 데이터를 약 5배 빠른 고성능 SQL 쿼리 성능을 보장
쿼리 대상 데이터 신뢰도 수준 제공	- 쿼리 대상 소스 데이터에 대한 표준, 품질 수준을 사전에 측정해 쿼리 질의 결과에 대한 데이터 신뢰도 수준을 제공

2) 제품 적용기술 내역

가) 분산 인메모리 기반 데이터 처리 기술

- 데이터 가상화는 사용자의 요구에 즉시 필요한 데이터를 통합하여 제공하는 역할을 수행하게 되는데 이때 필요한 데이터를 통합하기도 하지만 이미 사용한 데이터를 캐시에 보관하여 즉시 제공하는 것이 성능향상에 매우 중요하며 이를 분산 인메모리를 활용하여 성능을 극대화하는 기술입니다.

나) 이중 캐시 관리 기술

- 분산 인메모리 캐시와 함께 분산 파일시스템을 2차 캐시로 활용하기 때문에 이중 캐시에 대한 관리 시스템과 사용자의 행위를 예측한 캐시 관리 기술이 성능에 매우 큰 영향을 주는 중요한 기술입니다. 이 기술에는 사용하기 전에 미리 데이터를 가져오는 캐싱 스케줄링 기능을 포함합니다.

다) 분석 쿼리 분산 처리 기술

- 분산 인메모리 캐싱 데이터에 대한 분석 및 조회 쿼리를 고속으로 처리하기 위한 분산 처리 기술입니다.

라) 분석 쿼리 분산 처리 기술

- 이기종의 원천 데이터에 대한 데이터 통합 자동화 기술입니다. 다양한 데이터에 대해 하나의 데이터 저장소처럼 처리하기 위해 데이터 통합이 필수적이며 이를 자동화하는 것이 데이터 가상화의 핵심기술입니다.

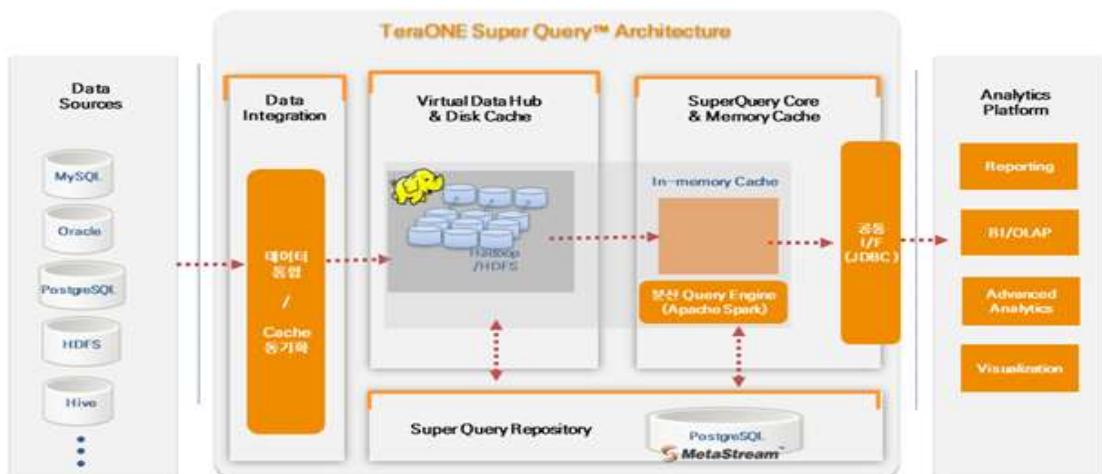


그림 22 TeraONE SuperQuery™ 적용기술

3) 주요 적용사례

사업명 : 근로복지공단 재정추계 시뮬레이션

주요 요구사항	<ul style="list-style-type: none"> - Oracle, SybaseIQ, Altibase DB 데이터를 소스로 분석 - R, 시각화 도구로 분석 모델 Query를 TeraONE SuperQuery로 Join 조회 - 사람의 개입 없이 이기종 소스데이터를 바로 가공하여 결과 제공
수행 내용	적용근로자수 추계, 재해자 수 추계, 수입/지출 추계, 시나리오에 따른 수지차 분석, 급여항목별 상관 분석 등 매매 재정추계 분석
도입 효과	<ul style="list-style-type: none"> - 기존 데이터웨어하우스 구축 대비 투입 공수 20% 절감, 개발 생산성 36% 향상, 분석 성능 평균 10 배 이상 개선
시스템 구성도	<p>The diagram illustrates the architecture of the TeraONE Super Query platform. On the left, multiple data sources (Oracle, SybaseIQ, Altibase DB) are connected to an 'Augmented ETL' process. This ETL process feeds into a 'Cache (Memory Disk)' component, which is represented by a cluster of blue cylinders. The 'Cache' then connects to a central 'TeraONE Super Query' box. Within the Super Query box, there are 'Augmented ETL' and 'MetaStore' components. To the right of the Super Query box, a 'JDBC' interface is shown. On the far right, there are two sections: 'Potal' and '활용대상'. The 'Potal' section contains icons for various reports and dashboards, such as '적용근로자 수', '사용자 수', and '수입 추계'. The '활용대상' section contains icons for different user groups, including '보통근로자 및 근무시간 분석', '기초통계분석', '시계열분석', and '회귀분석 등'.</p>

5. 데이터 거버넌스 (Data Governance) 기술

가. 데이터 거버넌스 플랫폼 (IRUDA™)

1) 제품 특징 및 차별화 요소

가) 제품의 개요

- 누구나 쉽게 필요한 데이터에 적기에 접근하여 분석하고 업무에 활용할 수 있도록 데이터를 자산화하기 위한 “보이는 데이터 거버넌스” 서비스 제공 플랫폼입니다.
- 기업 내외부에서 수집되는 다양한 데이터별로 오너쉽과 저장 위치, 경로 정의 및 분석에 활용하기 위해 참조하는 정보를 제공합니다.
- 사용자 중심의 데이터 거버넌스 포털 제공으로 직관적인 키워드 검색을 통해 데이터를 쉽게 탐색할 수 있습니다.

나) 개념도

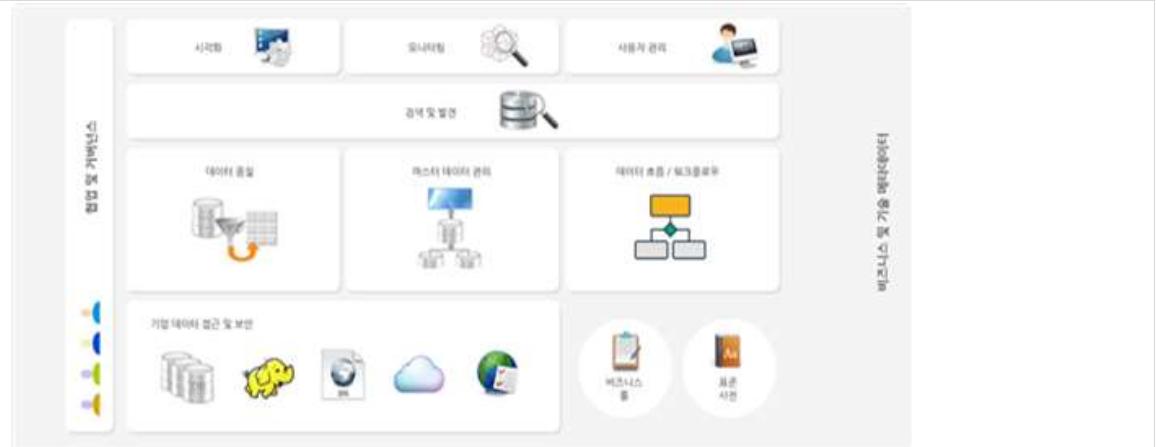


그림 23 IRUDA™ 개념도

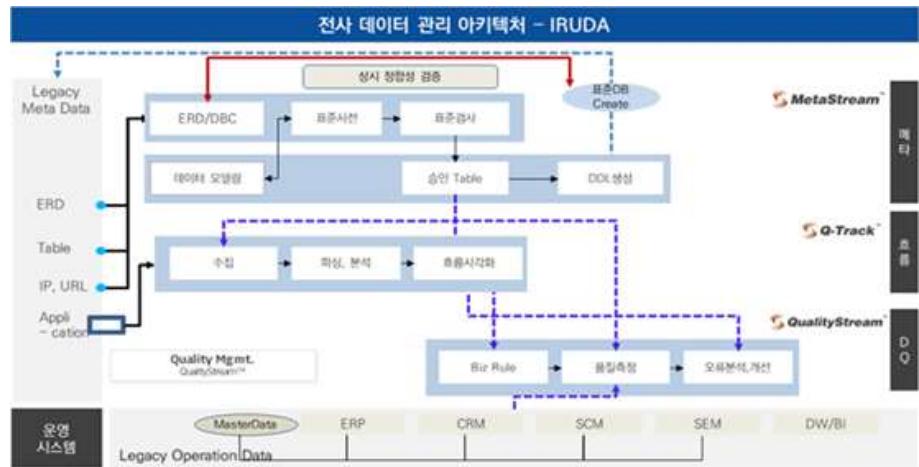


그림 24 IRUDA™ 아키텍처

다) 특장점

구분	설명
데이터 통합 검색	- 전사 데이터 거버넌스 관점으로 데이터를 통합 검색할 수 있음 - 메타데이터 매펑 정보를 시각화하여 제공 - 메타데이터, 품질관리, 기준정보관리, 데이터 흐름관리 기능을 종합적으로 구성하여 제공
데이터 거버넌스 서비스 제공	- 메타데이터, 품질관리, 기준정보관리, 데이터 흐름관리 기능을 종합적으로 구성하여 제공 - 를 엔진 기반 데이터 거버넌스 정책 관리 서비스 제공 - 데이터 맵 제공으로 원하는 데이터를 손쉽게 찾아갈 수 있음

2) 제품 적용기술 내역

가) 데이터 카탈로그 기술

- 데이터 통합 검색을 위한 기반 기술로 데이터의 설명, 품질, 위치 등의 데이터 특성을 미리 색인, 저장하여 통합 검색을 활용하여 데이터의 정보를 제공하는 기술입니다. 데이터 정보의 통합 수집, 색인, 검색 최적화 기술이 포함되어 있습니다.

나) 데이터 거버넌스 통합 관제 기술

- 데이터 거버넌스 플랫폼을 관리하는 기술로 통합된 사용자, 권한, 서비스 접근 제어를 관리하는 기술로 데이터 거버넌스 정책 관리 기술이 포함되어 있습니다.

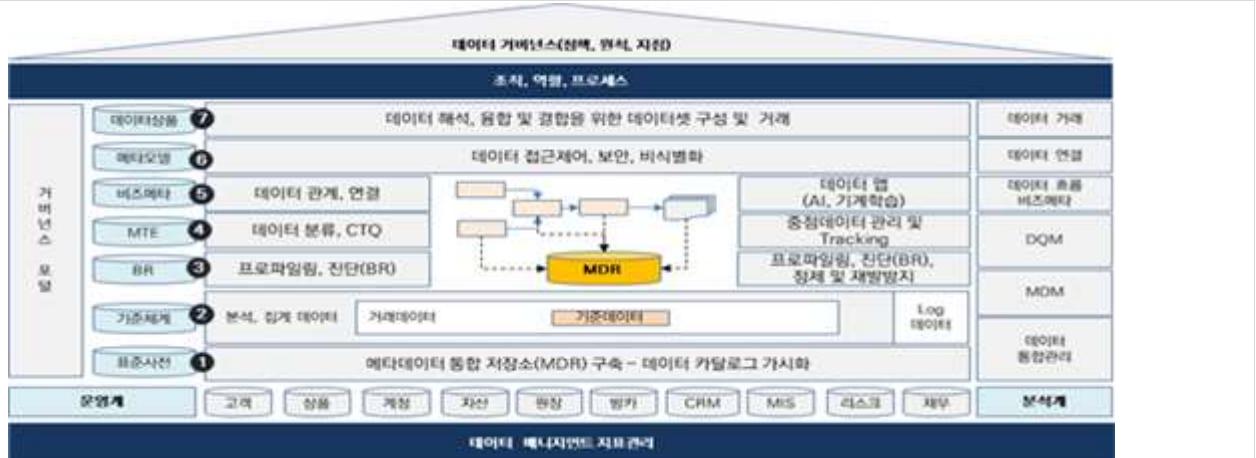


그림 25 데이터 거버넌스 플랫폼 기술

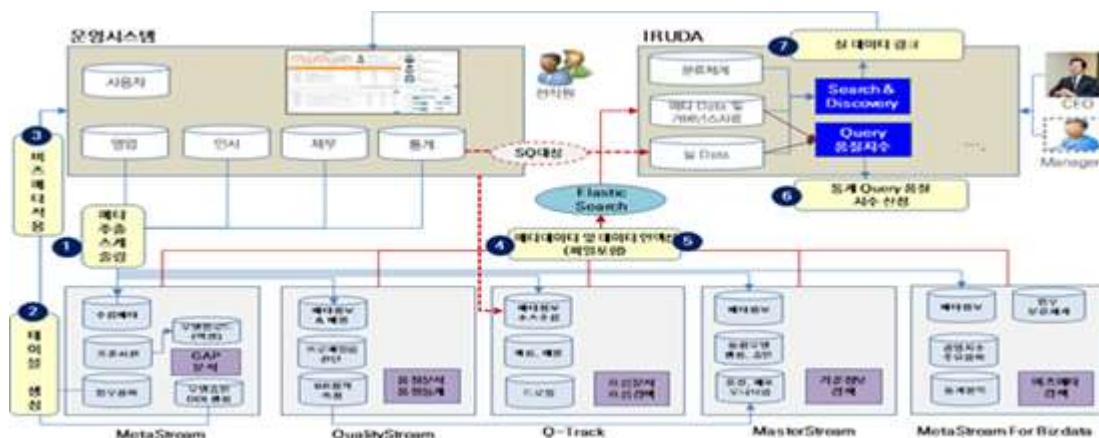
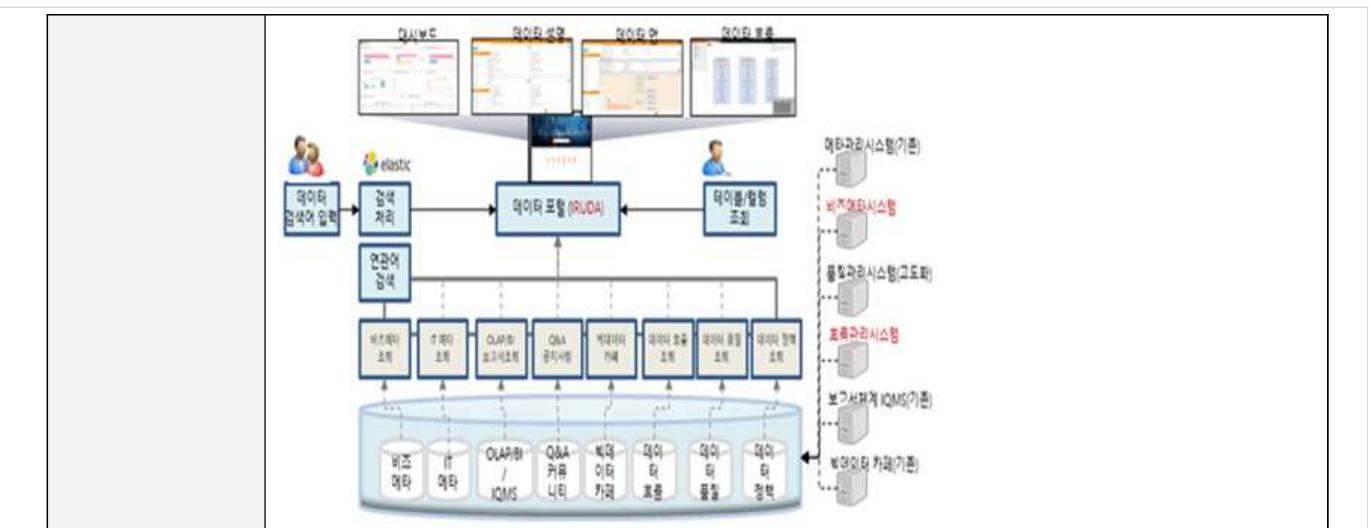


그림 26 데이터 거버넌스 운영 모델

3) 주요 적용사례

사업명 : KB국민은행 전사 거버넌스 구축	
주요 요구사항	<ul style="list-style-type: none"> 데이터 거버넌스 및 데이터 포털 구축으로 여러 곳에 산재하는 데이터를 전사 관점에서 표준 기반으로 통합하고 고품질의 데이터를 확보하여 누구나 쉽게 빅데이터 분석에 활용할 수 있어야 함 데이터 사용의 접근성, 편의성, 효율성 확대 및 데이터 활용에 따른 리스크 예방방안 마련 중장기적으로 경영진의 데이터 기반 의사결정 지원 기반으로 활용
수행 내용	<ul style="list-style-type: none"> ElasticSearch 엔진 기반으로 데이터 통합 검색 기능 제공 데이터 인덱싱, 우선순위, 유사도 검색 결과 제공하고 데이터 맵을 구성하여 보이는 데이터 관리 기능 데이터 표준, 데이터 품질, 데이터 흐름, 비즈메타 사용량, 데이터 값, 데이터 활용 등의 대쉬보드 제공
도입 효과	<ul style="list-style-type: none"> 데이터 자산화 및 자산측정 기반 마련 신규 데이터 확보 체계 수립 및 데이터 활용성 제고 목적별 데이터 관리 기준 마련 및 데이터 접근성 개선 데이터 오남용 모니터링을 위한 정보 제공 및 데이터 거버넌스 가이드라인 정립
시스템 구성도	



나. 데이터 표준화 및 메타데이터 관리 도구 (MetaStream™)

1) 제품 특징 및 차별화 요소

가) 제품의 개요

- Data의 Data, 즉 메타데이터를 관리하는 솔루션으로, 기업내 다양한 시스템에 분산된 메타정보를 추출/통합하여 표준화 및 관리합니다.
- 통합 메타데이터로 생성 위치/변경 상태를 관리하고, 변경관리 프로세스로 전사 메타데이터를 관리합니다.

나) 개념도



그림 34 MetaStream™ 개념도

다) 특장점

구분	설명
표준/비표준 용어관리	<ul style="list-style-type: none"> - 표준 용어외 별도로 비표준용어관리 기능을 제공 - 논리명, 물리명을 시스템 특성에 맞게 적용할 수 있도록 제공 - 표준과 마찬가지로 모델승인을 통하여 메타 레파지토리를 통해 관리할 수 있도록 제공 - 도메인(도메인유형-소분류-도메인) 하위 구조하에 하나의 용어는 하나의 도메인을 정의할 수 있도록 하여 표준관리에 충실
데이터 표준화관리	<ul style="list-style-type: none"> - 표준화를 위하여 단어, 용어(컬럼), 도메인 유사어 및 공통코드에 대한 조회 및 표준 체크 기능을 제공 - 표준 변경에 따른 변경 요청 및 승인 관리기능 제공
지원 환경	<ul style="list-style-type: none"> - 모델 뷰어: ERWin에서 제공하는 모델 레파지토리 정보를 침조하여 모델뷰어 쪽 표기방식으로 가능 제공 - 멀티 레파지토리 지원 : 오라클, 티베로 지원 - Web기반 UI 제공

2) 제품 적용기술 내역

가) 데이터 표준화 기술

- 데이터의 설계 단계에서부터 실제 데이터 저장소 생성까지 표준화된 용어로 관리할 수 있는 체계와 프로세스를 제공하는 기술입니다.

나) 메타데이터 추출 및 공유 기술

- 다양한 데이터 저장소의 메타데이터를 수집하여 표준준수 감시 및 표준화된 체계를 통한 시스템 간에 공유합니다.

다) 모델 관리 기술

- 모델링 도구와 연계되어 데이터 설계 단계에서 표준을 지켜 데이터가 설계되도록 관리하는 기술과 설계된 데이터를 실제 생성되는 단계까지 관리 될 수 있도록 하는 기술입니다.

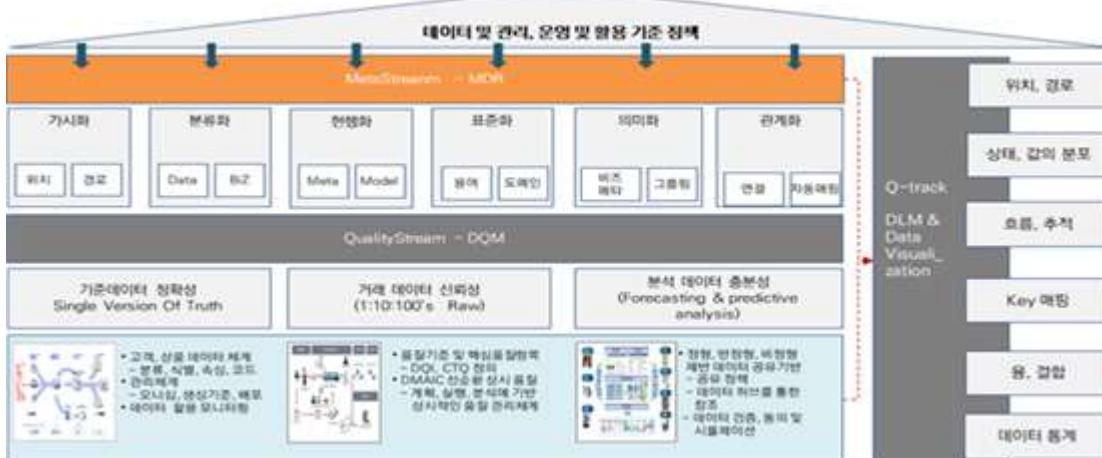


그림 28 표준화 및 메타데이터 관리 기술

3) 주요 적용사례

사업명 : 한국거래소(KRX) 메타데이터 관리시스템 구축	
주요 요구사항	<ul style="list-style-type: none"> - 전사 데이터 표준 지침 필요 - 전사 데이터 관리 조직 및 운영 프로세스 필요 - 데이터 표준 및 운영지원을 위한 메타 통합 레파지토리 구축 및 관리 기능 필요
수행내용	<ul style="list-style-type: none"> - 1 단계 데이터 표준화 관리 <ul style="list-style-type: none"> • 기간 : 2007년 10월 ~ 2008년 01월 • 내용 : 표준화 현행 자료수집 및 업무 분석, • 데이터 표준화 및 관리체계 수립 - 2 단계 전사 메타데이터 관리 시스템 <ul style="list-style-type: none"> • 기간 : 2007년 10월 ~ 2008년 06월 • 내용 : 데이터 표준화 시스템 가동과 추가 요건 도출 - 연계시스템(ETL, OLAP, JOB, Appl 품질)연동 - 전사적 메타데이터관리 시스템 개발, 이행, 가동
도입효과	<ul style="list-style-type: none"> - KRX의 전사 메타데이터관리 시스템 구축 과제 수행에 앞서 '데이터 표준화'를 선행 과제로 수행 - 데이터 표준화 결과를 중심으로 차세대 시스템의 분석단계부터 적용하고 데이터 검증을 받으며 구축
시스템 구성도	<p>The diagram illustrates the system architecture for data management. It starts with 'Application Source 정보' (Application Source Information) which feeds into the '영향도분석 영역' (Impact Analysis Area). This area includes 'Application 형상관리' (Application Shape Management) and '영향도 제공' (Impact Provision). The '영향도분석 영역' also provides data to the 'Data Quality Management' section. The 'Data Quality Management' section contains 'Broker/Parser' (주술), 'Data Quality Set' (표준 쿼리 세트, 표준 매핑 세트, 표준 용어 사용), and '표준 용어' (표준 용어) and '표준 코드' (표준 코드). This section also receives data from the '영향도분석 영역'. Finally, the 'Data Quality Management' section feeds into the 'Integrated Data Management' section. The 'Integrated Data Management' section includes 'DA조직 구성' (DA Organization Configuration), '관리 프로세스 정의' (Management Process Definition), '데이터 표준화 지침' (Data Standardization Guidelines), '데이터 관리 정책' (Data Management Policies), and '역할과 책임' (Role and Responsibility).</p>

다. 데이터 품질관리 도구 (QualityStream™)

1) 제품 특징 및 차별화 요소

가) 제품의 개요

- 분석 대상 데이터에 접근하여 품질진단, 결과도출 및 분석을 통해 지속적인 데이터 품질을 유지 향상 확보하는 솔루션입니다.
- 분석대상 데이터베이스에 대한 프로파일링을 수행하여 품질 수준 분석 후 관리 대상을 등록하여 지속적인 품질 정비 프로세스 수행합니다.

나) 개념도

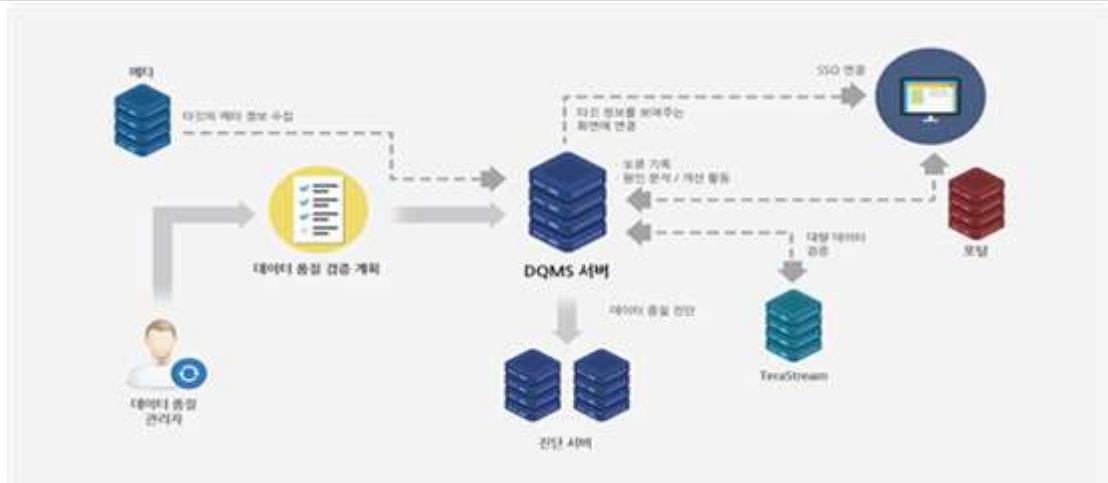


그림 29 QualityStream™ 개념도

다) 특장점

구분	설명
데이터 품질 관리 프로세스 지원 기능 제공	<ul style="list-style-type: none"> - 최신의 DBC 및 모델정보를 기반으로 데이터 품질측정 기준(DQI, CTQ) 적용 <ul style="list-style-type: none"> • 품질 측정대상, 항목 및 측정방법에 따라 품질측정 • 품질 측정 결과를 관리하고 상시 모니터링 수행 • 부서별, BR(업무규칙)별, 단위업무별 품질지수 측정 및 추이분석
비즈니스 룰 가중치 관리	<ul style="list-style-type: none"> - BR 가중치를 관리하여 품질 지수 산정시 오류율에 대한 레벨 반영 <ul style="list-style-type: none"> • 기업에서 중요하게 관리하는 데이터 항목에 대한 집중 품질 관리
개선활동 관리 및 모니터링	<ul style="list-style-type: none"> - BR측정 오류에 대하여 개선계획을 수립하고, 정제개선 진행을 수행한 후, 품질 재측정을 통해 개선활동 결과에 대한 진척관리를 지원 <ul style="list-style-type: none"> • 진행상황을 개선 회차별로 비교하여 모니터링 가능
통합 품질현황 및 통계분석	<ul style="list-style-type: none"> - 솔루션 내 차트그리드 컴포넌트를 활용하여 다양한 차트기능 제공 - 포털솔루션과 연계하여 구축한 다양한 사례 확보(특허청, 도로공사, 삼성화재 등)
품질 검증 성능	<ul style="list-style-type: none"> - 자사의 ETL솔루션(TeraStream™)과 연동하여 DQ-ODS영역으로 데이터를 이관 후 품질 검증 수행
메타시스템과 연동	<ul style="list-style-type: none"> - 자사 메타데이터 관리 솔루션(MetaStream™)과 연동하여 DBC정보를 실시간 공유 - 도메인 기반의 프로파일링 설정을 수작업 입력 없이 솔루션 연계 지원

2) 제품 적용기술 내역

가) 데이터 품질 관리 프로세스 관리 기술

- 데이터 품질 지표의 관리부터 데이터 품질 모니터링까지 데이터 품질 관리 전체 Cycle에 대한 프로세스 관리 기술입니다.

나) 데이터 품질 모니터링 기술

- 데이터 품질 지표에 따라 지속적인 데이터 품질을 모니터링하고 모니터링 결과를 활용하여 품질 개선 활동을 연결될 수 있도록 지원하는 기술입니다.

다) 빅데이터 품질 진단 기술

- 빅데이터 저장소의 데이터도 동일한 기준과 방식으로 품질 관리 프로세스에서 관리합니다.

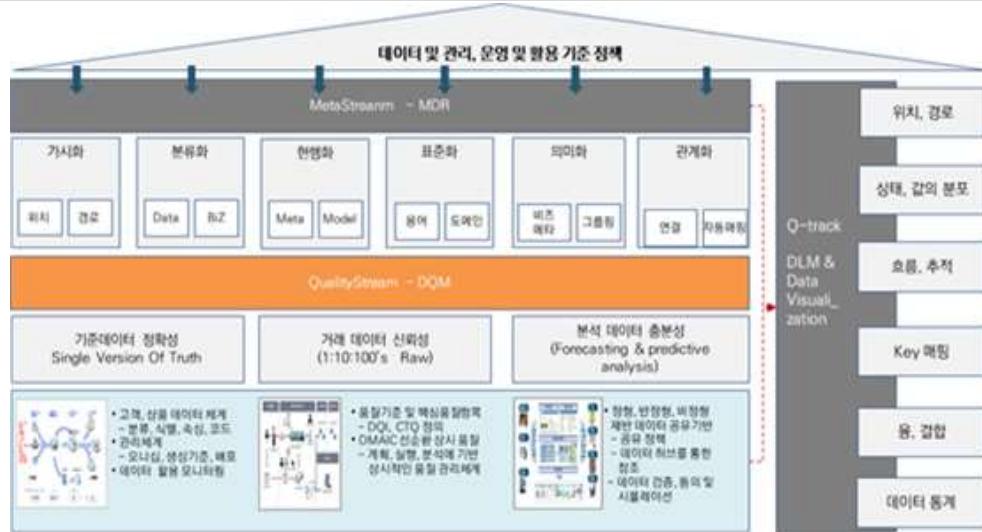


그림 38 데이터 품질관리 기술

3) 주요 적용사례

사업명 : 특허청 데이터 품질 상시감시체계 구축	
주요 요구사항	<ul style="list-style-type: none"> - 시스템의 복잡도 증가 및 다양한 형태의 자료를 처리에 따른 오류증가 - 고품질 데이터 제공 요구 - 대국민 서비스 증가에 따라 외부 서비스 데이터에 대한 검증 필요
수 행 내 용	<ul style="list-style-type: none"> - 통합 레파지토리를 중심으로 한 데이터품질 상시 감시 시스템 구축 <ul style="list-style-type: none"> • 메타, 품질, 영향도 시스템 구축 - 데이터 거버넌스 체계 수립 <ul style="list-style-type: none"> • 데이터관리정책(데이터표준, 데이터품질) 수립 • 데이터관리프로세스 수립 • 데이터관리 조직
도 입 효 과	<ul style="list-style-type: none"> - 대국민 서비스 데이터에 대한 신뢰성 확보 - 다양한 국제교류 및 도약을 위한 체계적인 데이터관리로 데이터 품질 국제간 경쟁 우위 확보 - 국내 및 국제 환경 변화로 인한 다양한 고품질 데이터 제공 체계 구축
시스템 구성도	

라. 비즈메타관리 도구 (MetaStream For BizData™)

1) 제품 특징 및 차별화 요소

가) 제품의 개요

- 부서, 업무기능별 작성하는 보고서 및 통계자료에 표출되는 항목 중 용어에 대한 이해도 차이로 값이 서로 상이한 경우가 있습니다. 누구나 같은 의미로 업무 용어를 이해하고 같은 산출결과를 도출하기 위해 비즈니스 용어를 메타화하여 관리하는 메타 데이터 관리 솔루션입니다.
- 전사 관점의 비즈니스 메타(경영지수/계수, 핵심업무 용어 등)를 사용자 관점에서 표준 정의, 합의 과정을 거쳐 표준으로 관리합니다.

나) 개념도

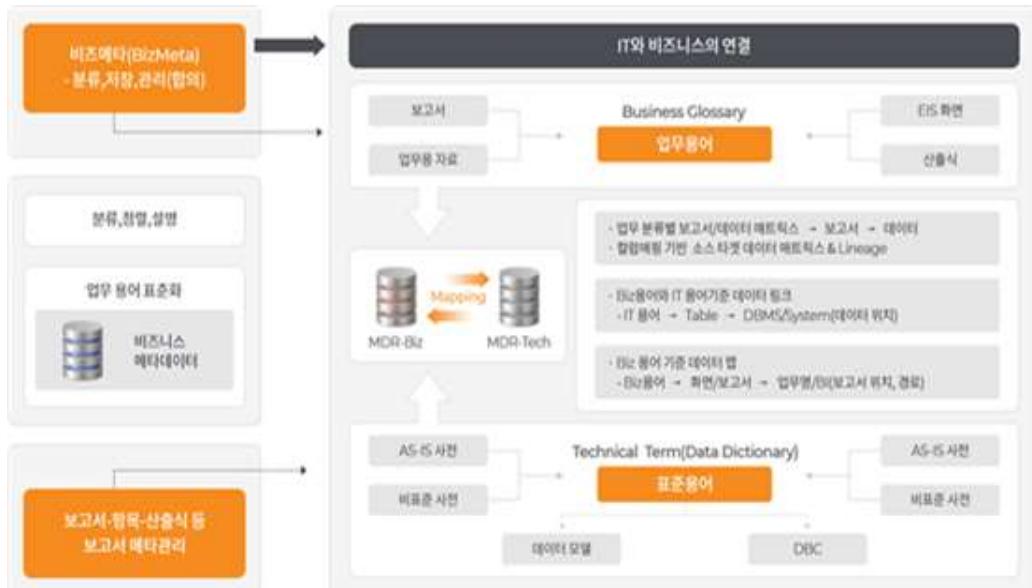


그림 40 MetaStream for BizData™ 개념도

다) 특장점

구분	설명
현업 업무 담당자 관점에서 데이터 가시화 관리	<ul style="list-style-type: none"> - 업무에서 사용하는 용어의 형식, 값을 표준화하여 부서간 상호 운용성 확보 - 데이터를 주가 만들고, 어떻게 사용되며, 어디로 이동되는지 투명한 관리 가능 - 통합, 공유 기반의 의사결정 지원이 가능하도록 지원
데이터 활용성 제고	<ul style="list-style-type: none"> - 데이터 질과 양을 겸비하여 AI 및 고품질 데이터 기반 빅데이터 분석 역량 확보 - 모든 사용자가 모든 데이터를 활용할 수 있게 한다는 데에 초점을 두고 현업 사용자가 데이터에 접근하는 프로세스 개선 관리 가능 - 같은 시간에 더 다양한 데이터에 접근할 수 있도록 지원 - IT 메타데이터와 연계하여 현업과 IT의 원활한 소통 지원

2) 제품 적용기술 내역

가) 비즈니스 메타데이터 분류체계 관리 기술

- 비즈니스 형태나 관리 데이터의 형태에 따라 동적으로 분류체계를 설정하고 이에 따라 비즈니스 메타를 관리하는 기술입니다.

나) 비즈니스 메타데이터 동적 관리 화면 생성 기술

- 다양한 형태의 데이터에 대한 통일된 뷰를 제공하기 위한 동적 관리 화면 생성 기술입니다.

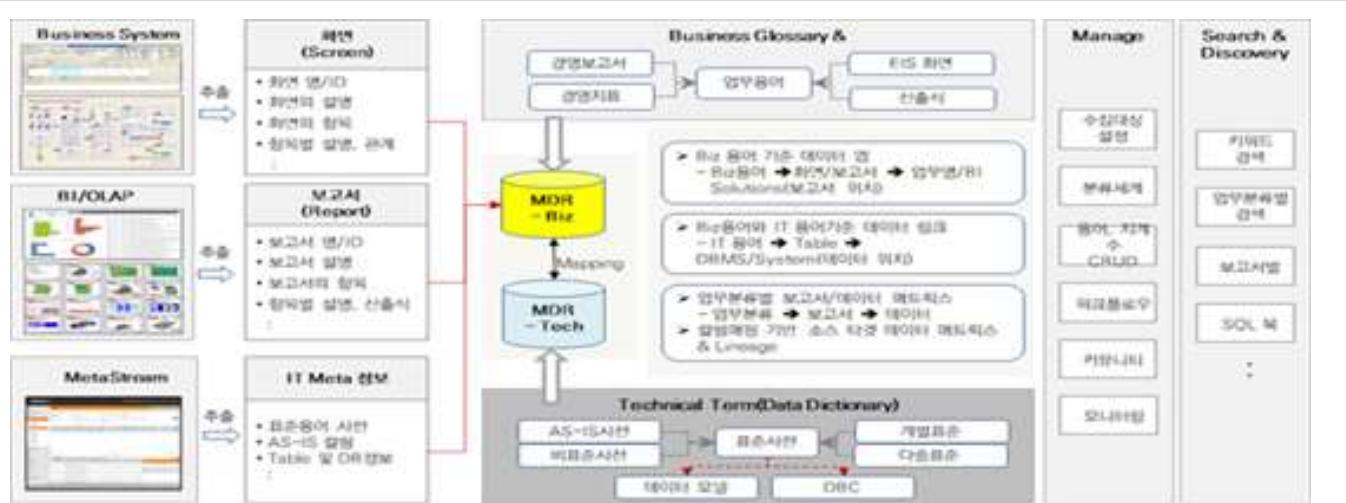


그림 41 비즈메타 관리 기술

3) 주요 적용사례

사업명 : 한국투자증권 비즈메타 기반 데이터 포털 구축	
주요 요구사항	<ul style="list-style-type: none"> - 협업 사용자, 관리자 및 데이터 분석가 입장에서 활용하는 데이터에 대한 설명 제공 - 각종 보고서, 화면, 장표에서 표현되는 업무 용어, 용어별 산출식, 산출근거, 보고서 작성자 등 필요 항목을 관리하고 다양한 형태로 시각화 - 업무용어 및 분류체계를 정의하고 표준계수, KPI 산출식, 산출 근거, SQL 복을 작성할 수 있도록 하며 보고서 분류 검색과 데이터 링크를 통한 상세 조회 가능
수행 내용	<ul style="list-style-type: none"> - 회사에서 작성되는 화면 정보 및 DB구성정보, 정형 보고서, 비즈니스 메타데이터, IT 메타데이터, 및 BI/OLAP 보고서 속성 정보 수집하여 DB를 구축하고 관련정보에 대한 관리기능 구현
도입 효과	<ul style="list-style-type: none"> - 협업과 IT간 소통채널 역할로 데이터 연계 및 매핑을 통한 검색서비스 포털로 제공 - 회사 내외부 중요한 업무 용어가 의미하는 내용에 대한 설명 산출식, 재활용 정합성 확보 - IT 메타데이터와 연계하여 정보계 시스템, 경영보고용으로 시간을 많이 소모하는 대상과 과거에 어떤 기준으로 보고서를 출력했었는지 상시 참조 가능 - 누구나 같은 의미로 이해하고 같은 방법으로 산출하여 어느 부서에서나 동일한 결과를 도출함
시스템 구성도	<p>The diagram illustrates the Data Portal architecture. At the center is the 'Data Portal (Data sharing and reuse)' box. It is connected to various data sources on the left: 'Reporting' (with icons for 경영표준계수, OLAP/BI, KPI/Reporting, and 기준정보 목록), 'Search & Discovery' (with icons for 키워드 검색, 업무 분류별 검색, Self BI/SQL 복, 외부 데이터 검색, and 대시보드/My Page), and 'Report sharing and IT integration' (with icons for 보고서 및 용어 확보, 표준업무용어 정의 및 술언, and 보고서와 IT 연결). On the right, the portal connects to external users via 'Related Data Link' (Customer, EIS 화면 Link, BI/OLAP Link, IT Meta Link, and 보고서 위치 설정, 내용 찾기, and 풀려찾기/분석요청), and to internal users via 'Community' (공지사항, 게시판, and 데이터 카페). Three user profiles are shown on the right: '데이터 사용자 (한글 운영)', '데이터 관리자 (시스템 운영)', and '데이터 분석가 (CRM)'.</p>

마. 데이터 흐름관리 도구 (Q-Track™)

1) 제품 특징 및 차별화 요소

가) 제품의 개요

- 운영 시스템으로부터 데이터웨어하우스, 정보성 단위업무 시스템에 이르는 데이터의 생성, 가공, 활용에 이르는 흐름을 시각화 기능을 이용하여 관리할 수 있는 솔루션입니다.
- 직관적이고, 정확하며, 의미 있는 데이터 흐름 정보를 Single View로서 모니터링 및 관제할 수 있는 환경 제공
- 데이터 흐름 분석 및 변경 영향 분석 기술을 활용하여 데이터분석에 필수적인 핵심개념 파악 및 분석값의 추적 및 운영 관리성을 강화할 수 있음

나) 개념도

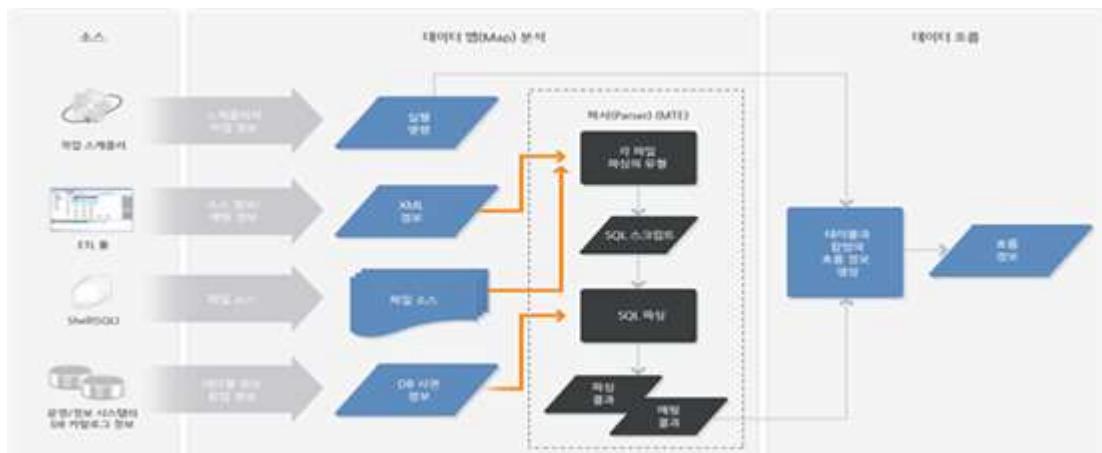


그림 43 Q-Track™ 개념도

다) 특장점

구분	설명
Web 기반의 운영/관리 환경 제공	<ul style="list-style-type: none"> - Web기반의 시스템 및 시스템 구성도 도식화 기능 - 데이터 흐름 정보의 분석/생성 자동화 기능 - 데이터 흐름 분석을 위한 작업 스케줄링 기능 - 프로그램 사용/미사용 관리 기능 - 테이블/컬럼/파일/프로그램별 연관 집계표 관리 기능
분석 기능 제공	<ul style="list-style-type: none"> - ETL 도구의 데이터 가공 정보 분석 기능 - 데이터 처리(추출/가공/적재), 전송에 사용되는 각종 프로그램과 도구 매핑 정보 분석 기능 - 파일 형태의 데이터를 포함한 테이블/컬럼 간의 매핑 분석 기능
각종 통합운영 관리 기능 제공	<ul style="list-style-type: none"> - 데이터 맵 형태의 데이터 간 관계정보 조회 기능 - 테이블 흐름정보 조회 기능 - 컬럼 매핑 및 흐름정보 조회 기능 - ETL 도구의 프로젝트 별 테이블 컬럼 흐름 정보 조회 기능 - 데이터 흐름을 생성하는 프로그램 조회기능 - 프로그램 테이블 컬럼 기준의 통합 검색 기능 - 데이터 흐름 정보에 대한 집계표, 통계정보 리포팅 기능 - 데이터 흐름을 통한 테이블/컬럼/프로그램의 정합성 검증 관리

2) 제품 적용기술 내역

가) Application 분석 기술 (Parser)

- 다양한 프로그래밍 언어로 구현된 Application을 분석 데이터 흐름을 추출하는 기술

나) Application 관계 분석 기술

- Application 간의 호출관계 선후 관계를 추출하여 데이터 흐름의 순수를 확정하는 기술

다) 데이터 흐름 분석 기술

- Application 분석 결과를 통합하여 전사 데이터 흐름을 생성하는 기술

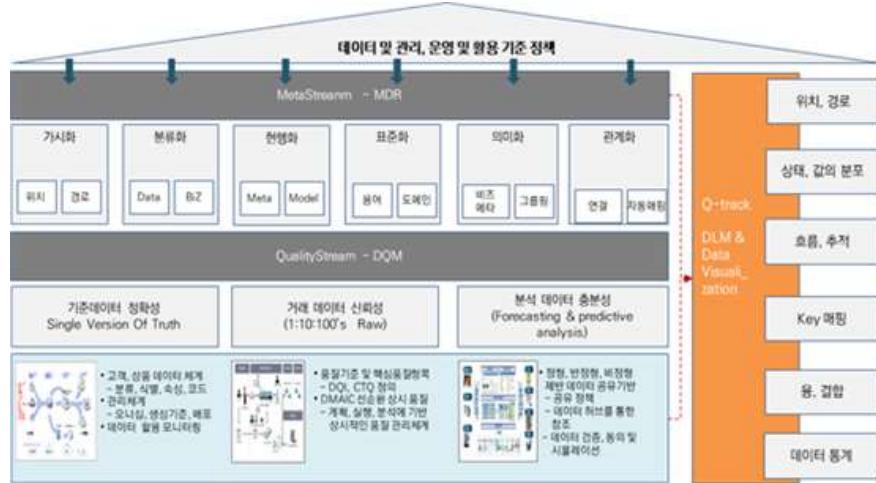


그림 44 데이터 흐름관리 적용기술

3) 주요 적용사례

사업명 : 특허청 데이터 품질 상시감시체계 구축	
주요 요구사항	<ul style="list-style-type: none"> - 정보계에서 잘못된 정보가 출력이 될 때 원천(채널, 계정계 등)에서 발생한 오류를 확인하기 어려움 - 정보계 마트 구성을 위해 개발된 단위 프로그램, SQL 그리고 테이블, 컬럼 접근 정보를 파악하기 위해 모든 소스 프로그램을 확인해야 함에 따른 생산성 저하 - 자체 생성이 아닌 외부 시스템으로부터 데이터 항목별로 전달(I/F)된 데이터의 경로를 찾고 데이터 컬럼을 추적할 수 있는 도구 필요
수행 내용	<ul style="list-style-type: none"> - 데이터 흐름을 분석하기 위한 대상시스템 정보(IP, user)와 대상 프로그램 Language 정보를 설정 - 설정된 대상시스템 프로그램, SQL 파서를 등록하고 각 분석 대상 소스정보와 매핑 - 형상관리 솔루션으로부터 주기적으로 소스 변경정보를 반영하고, 소스별 파서를 실행하여 ETL job 단위로 테이블 및 컬럼의 소스, 타겟정보를 식별하여 Lineage 형태로 Draw
도입 효과	<ul style="list-style-type: none"> - ETL 데이터 흐름을 단위 프로그램이 아닌 관련 프로그램 전체를 연결하여 시각적으로 표현하여 오류가 발생된 지점 전/후의 데이터 처리 흐름을 쉽게 파악 - 테이블, 컬럼을 사용하는 프로그램/SQL 바로 찾을 수 있으며, 생성 시점에 만들어진 데이터가 OLAP 등 통계분석 보고서에 어떻게 변환 규칙이 반영되어 나타나는지 추적 용이
시스템 구성도	<p>The diagram illustrates the system architecture for data quality monitoring. It shows the flow of data from various sources (e.g., 계정계, 단위업무) through a framework (e.g., Framework, 미증설여, EAI, ETL) to a central monitoring system (e.g., BAT, SOR). The monitoring system then feeds into reporting databases (e.g., DM, 경영정보) and other business units. The flow is managed by a series of interfaces (e.g., CDC, ETL, EAI) and connectors (e.g., 소통-1, 소통-2, 소통-3, 소통-4).</p>

바. 마스터데이터관리 도구 (MasterStream™)

1) 제품 특징 및 차별화 요소

가) 제품의 개요

- 기업 내 핵심이 되는 비즈니스 데이터 즉, 마스터 데이터를 전사적으로 일관성 있게 생성/관리하고, 각 업무 프로세스 흐름에 맞춰 동일한 데이터를 지속적으로 유지하기 위한 솔루션
- 중앙 집중형 또는 절충형 방식으로 취합/생성/검증 동시 배포
- 레ガ시(Legacy) 시스템에서 생성된 데이터를 통합하여 각 응용 시스템에서 참조하기 전 업무 규칙에 의한 검증 이후 동일한 시점에서 활용하도록 동기화

나) 개념도



그림 35 MasterStream™ 개념도

다) 특장점

구분	설명
Apache Hbase를 이용한 master data 저장소 구축	<ul style="list-style-type: none">- Master data의 변경 내용 추적 및 데이터베이스 레벨에서 특정 시점으로의 rollback이 가능 제공- MDM에서 빈번하게 일어나는 operation인 upsert가 RDBMS대비 고속 수행- 분산 데이터베이스로 Scale-out을 통한 확장 및 HA 구성 용이
Apache Spark을 이용한 인메모리 분산 처리	<ul style="list-style-type: none">- 데이터 분석과 처리를 인메모리 기술인 Spark으로 고속 처리- 처리 성능은 Node Scale-out으로 향상 가능
ETL(Extract Transform Load) tool을 이용한 수집과 배포	<ul style="list-style-type: none">- 검증된 자사 ETL 제품인 TeraStream™과 연계하여 다양한 원천 시스템으로부터 데이터를 추출하여 통합 및 타겟 시스템으로 배포- 자사의 CDC(Change Data Capture) 제품인 델타스트림(DeltaStream™)을 이용하여 실시간으로 데이터를 수집 및 배포 가능
Web 기반의 사용자 인터페이스 제공	<ul style="list-style-type: none">- Web 기반 서비스로 웹브라우저를 통해 관리/모니터링 가능

2) 제품 적용기술 내역

가) 마스터 데이터 통합 기술

- 다양한 시스템에 산재되어 있는 데이터를 통합하여 마스터 데이터를 생성하는 기술

나) 마스터 데이터 관리 서비스 생성 기술

- 설계된 마스터 데이터의 구조에 따라 관리 및 조회 화면을 동적으로 생성, 데이터의 특성에 따라 관리할 수 있도록 제공하는 기술

다) 마스터 데이터 업무 프로세스 생성 기술

- 마스터 데이터 관리 체계에 따라 결재선 설정, 프로세스 설정 등을 사용자가 할 수 있도록 하여 Workflow를 동적으로 생성 하는 기술

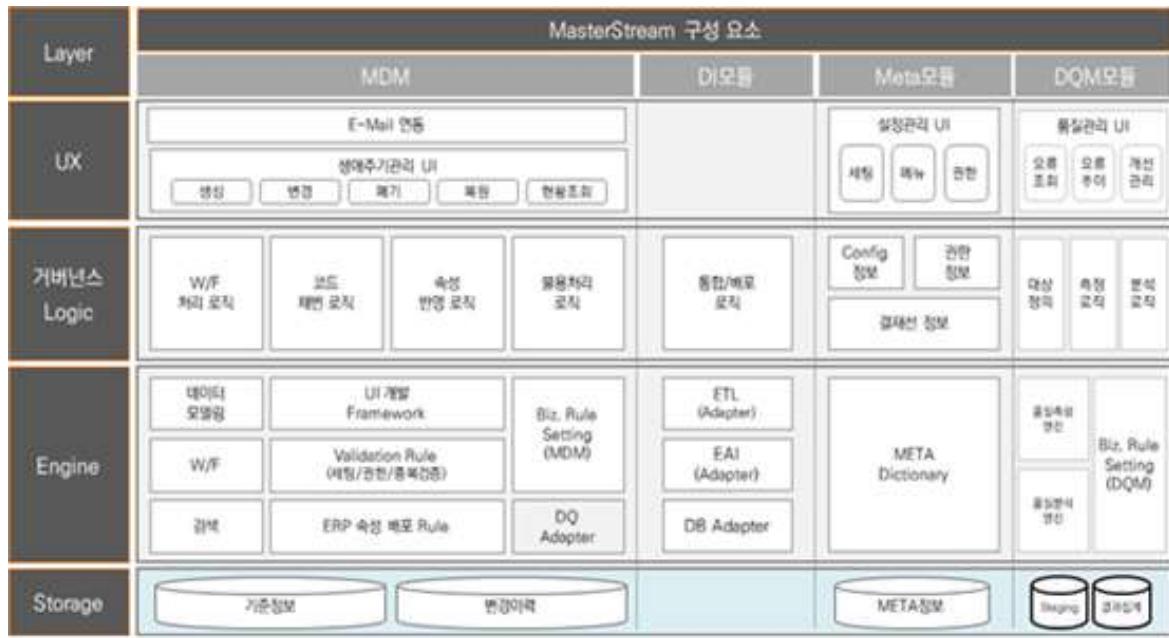
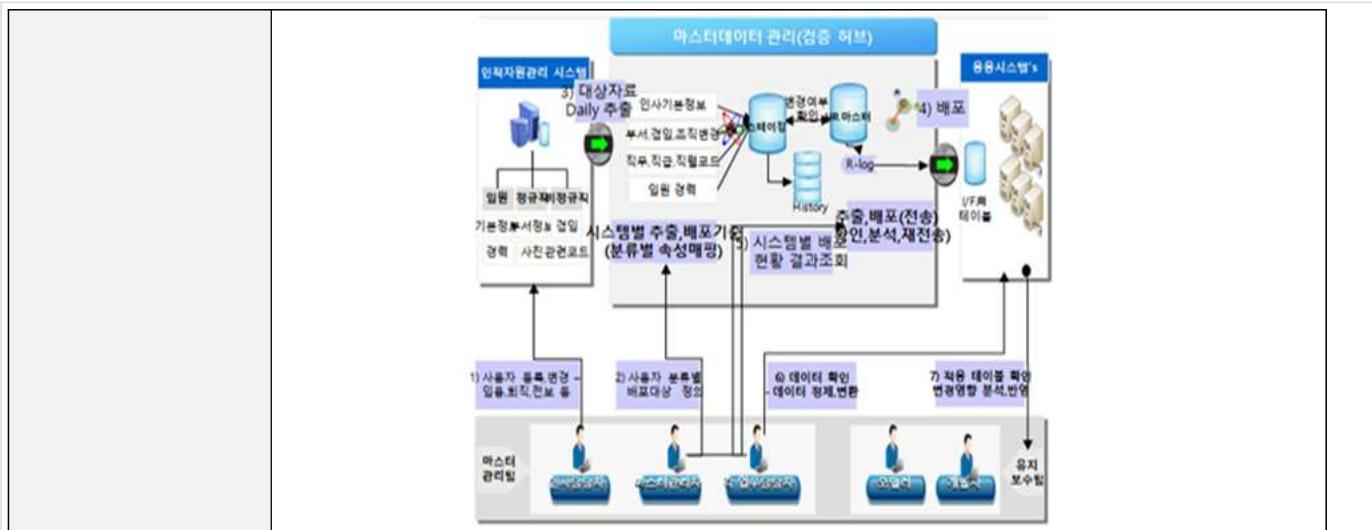


그림 36 마스터데이터 관리 기술

3) 주요 적용사례

사업명 : 국회사무처 인적자원MDM구축	
주요 요구사항	<ul style="list-style-type: none"> - HR시스템에 등록된 사용자(국회의원, 의원실 종사자, 사무처직원, 계약직원, 협력사 직원 등) 임면 정보가 적시에 배포·공유되지 않음 - HR시스템내 사용자 등록 이후 검증과정이 미흡하여 기본정보는 존재하지만 부속정보(발령, 겸직 등)가 실정보와 불일치하는 문제점 발생
수행내용	<ul style="list-style-type: none"> - HR 시스템에서 사용자 정보 신규·변경, 조직의 개편·전보, 파견, 겸무 등의 인사발령 발생시 변경 데이터를 식별함 - 검증된 룰에 의한 재직자 추출을 거쳐 사용자 그룹별로 정의된 매핑 기준에 맞추어 데이터群과 데이터 속성을 분리 - 각 시스템별로 주기에 따라 배포하고 로그를 기록하여 쉽게 추적할 수 있도록 관리
도입효과	<ul style="list-style-type: none"> - 사무처 인사시스템에서 생성·변경(퇴사포함)된 정보가 정해진 규칙에 의해 검증되고 검증된 데이터는 적시에 관련 시스템(14 개)에 배포되어 동기화 - 각 시스템별 적용 시점이 동일하고 정보의 시차가 없어졌으며, 겸직 등 근무자 정보의 정확도가 향상되었음
시스템 구성도	



● 기술개발내용 및 과정

1. 빅데이터 플랫폼의 활용 배경

4차 산업혁명 시대에 테슬라, 우버, 에어비엔비와 같이 디지털 기술을 활용한 새로운 비즈니스 모델로 시장을 파괴하는 데이터 중심의 혁신 기업이 등장하면서 기업이나 기관들의 디지털 전환 (Digital Transformation)을 통한 혁신이 가속화되고 있습니다. 디지털 전환의 핵심은 전통적인 디지털 컴퓨터, 네트워크 및 모바일 등 ICT 기술을 배경으로 IoT, 빅데이터, 클라우드, AI, 블록체인 등 최신 디지털 기술로 확장, 산업의 경계를 초월하여 융복합 비즈니스를 통해 새로운 부가가치를 창출하는 것입니다. 기존의 전산화, 자동화는 비즈니스 모델 변화 없이 기술 중심의 기업가치 향상을 제고하는 것인 반면, 디지털 전환은 전통적인 비즈니스 영업을 뛰어 넘어 신규 융합 비즈니스 모델을 창출하고 이를 기반으로 기업가치를 극대화하는 것으로 빅데이터와 AI가 가장 핵심적인 역할을 하게 됩니다.

이러한 디지털 전환의 핵심 요소가 디지털 데이터가 될 것이며 이를 어떻게 정의하고 표준화하고 다양한 형태로 통합하고 공유하여 융합 비즈니스 모델을 구현하고 상업적 가치를 창출할 뿐 아니라 넓고 깊은 지식을 축적하여 인간의 개입 없이 활용할 것인가의 답은 바로 “빅데이터 플랫폼”이 될 것입니다.

DBMS 중심의 전통적인 데이터 분석 시대에는 데이터의 분석 용량의 한계가 존재했으므로 빅데이터 용량의 데이터를 전부 가공하고 정리하여 분석의 대상으로 삼기보다 일부 데이터를 샘플링해서 분석한 뒤, 모수를 추정하는 통계 기법을 활용하거나 데이터의 통합대상에서 빅데이터를 제외하고 비즈니스 시스템에 있는 정형데이터를 대상으로 대용량 처리용 DBMS를 이용하여 데이터의 분석을 하였습니다.

하지만 2009년 모바일 앱의 등장과 보급으로 개인별 실시간 발생하는 데이터가 폭증하게 되었고, IoT 기술 확산에 따라 기계에서 발생하는 센서 및 로그 데이터 등을 포함해서 초대용량 데이터 처리 기술이 발전하였습니다. 또한, 데이터 저장매체 가격의 하락 그리고 이러한 데이터를 분석해서 인사이트를 도출할 수 있는 AI 기술의 발달로 빅데이터 시장이 2010년부터 본격적으로 확대되기 시작했습니다.

빅데이터 플랫폼과 AI 학습 분석 기술의 발달로 현실 세계에서 발생하는 모든 유형의 데이터를 용량과 상관 없이 수집, 저장, 가공, 분석, 학습, 활용할 수 있게 되었습니다. 분석하고자 하는 전체 데이터를 대상으로 기존 방법으로는 알 수 없었던 새로운 정보와 통찰력을 도출하여 전혀 새로운 차원의 서비스를 만들어낼 수 있게 된 것입니다. 이러한 데이터 플랫폼 기술의 발달에 힘입어 디지털 전환의 가속화가 글로벌 하게 진행되고 있으며 병국가적으로도 디지털 뉴딜 정책을 시행하고 있고, 그 중 D.N.A (Data, Network, AI) 생태계 강화 사업을 위해 많은 예산을 집행하고 있어서 클라우드 서비스의 확산과 빅데이터 플랫폼의 활용이 폭발적으로 증가할 것으로 예측하고 있습니다.

디지털 고객과 밸류 체인의 스마트화로 변화되는 미래 비즈니스의 출발점은 “데이터” !!

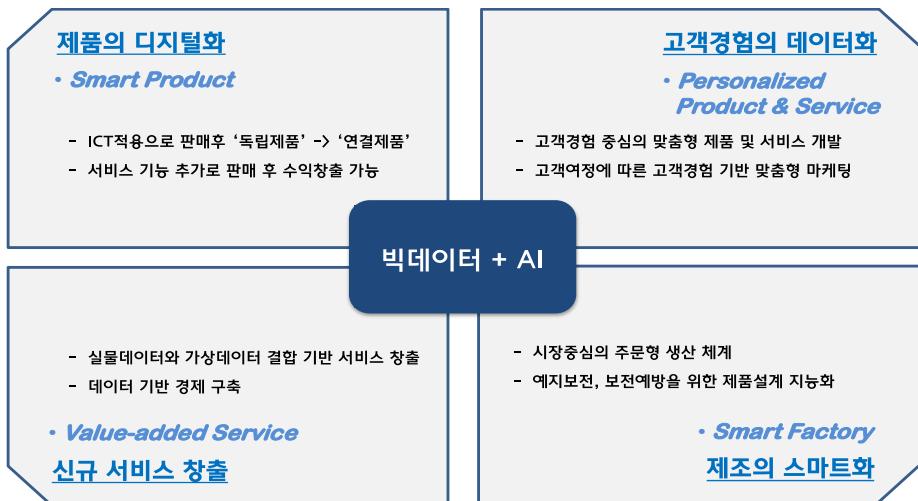


그림 37 디지털데이터 기반 미래 비즈니스 모델 예시

2. 제품의 개요

“빅데이터 플랫폼이라 함은 다양한 형태로 저장된 수많은 유형의 데이터를 쉽게 수집하고 효과적으로 통합하여 분석할 수 있는 형태로 저장하고 이후 학습까지 할 수 있도록 만들어진 소프트웨어 플랫폼을 이야기합니다.”

최신의 빅데이터 플랫폼 기술은 빅데이터 초창기 시대의 빅데이터 레이크(Lake)의 개념에서 진보하여 빅데이터 패브릭(Fabric)의 개념으로 확대 발전되고 있습니다. 즉, 기존의 데이터의 수집, 가공, 저장, 분석 기능은 물론이고 실시간 및 스트리밍 데이터의 효과적인 처리, 데이터 페더레이션(Federation) 기술을 이용한 데이터 가상화 기술 및 데이터 거버넌스 기능을 제공하여 디지털 전환시대에 옴니 채널(Omni-Channel) 데이터 서비스를 효과적으로 지원하기 위해 필요한 빅데이터 플랫폼으로 진화하고 있습니다.

당사의 주력 제품인 TeraONE™은 20년 기술 경험의 집약체로써 국내 1위 데이터 통합 제품인 TeraStream™과 역시 국내 1위 메타데이터 관리 제품인 MetaStream™을 바탕으로 하둡(Hadoop) 생태계에 포함된 공개 소프트웨어를 적극적으로 적용하고 창사이래 꾸준히 개발해온 실시간 데이터 통합 및 데이터 거버넌스 제품군을 통합하여 빅데이터 패브릭을 구현한 국내외적으로 가장 진보된 종합 빅데이터 플랫폼으로 개발되었습니다.

당사가 추구한 데이터 통합 제품의 개념은 대용량 데이터를 정렬하고 가공하기 위해서는 기존의 관계형 DBMS로는 불합리하다는 결론을 기초로 하여 병렬처리기술 기반의 파일시스템 중심으로 대용량 데이터를 처리하는 것이었습니다. 이는 90년대 말 2000년 초반에 대용량데이터에 대한 대응이 미숙했던 시대에 획기적인 시장의 호응을 불러일으켜 당사의 TeraStream™이 시장의 강자로 떠오르는데 결정적인 기여를 하였습니다. 빅데이터 시대를 촉발한 미국 실리콘밸리에서 개발된 하둡분산파일시스템(HDFS)도 결국 위와 같은 원리로 만들어진 것으로 이를 빅데이터의 가공 처리에 적용하면서 기존의 관계형 DBMS뿐만 아니라 수직컬럼형 DBMS 시장을 급격하게 잠식하여 빅데이터, AI 시대를 앞당기게 되었습니다.

아래 그림 [그림 2-38]은 DBMS 중심의 단일 데이터 처리 방식을 2가지의 복합방식으로 분리하여 처리 유형에 따라 최적화하는 대용량데이터 처리 아키텍처를 보여줍니다. 즉, 온라인(On-line)성 업무에는 관계형 DBMS를 적용하되 가공, 분석을 위한 대용량 배치(batch)성 업무는 병렬처리 기반의 파일시스템을 활용하는 복합 데이터 처리 방식입니다.

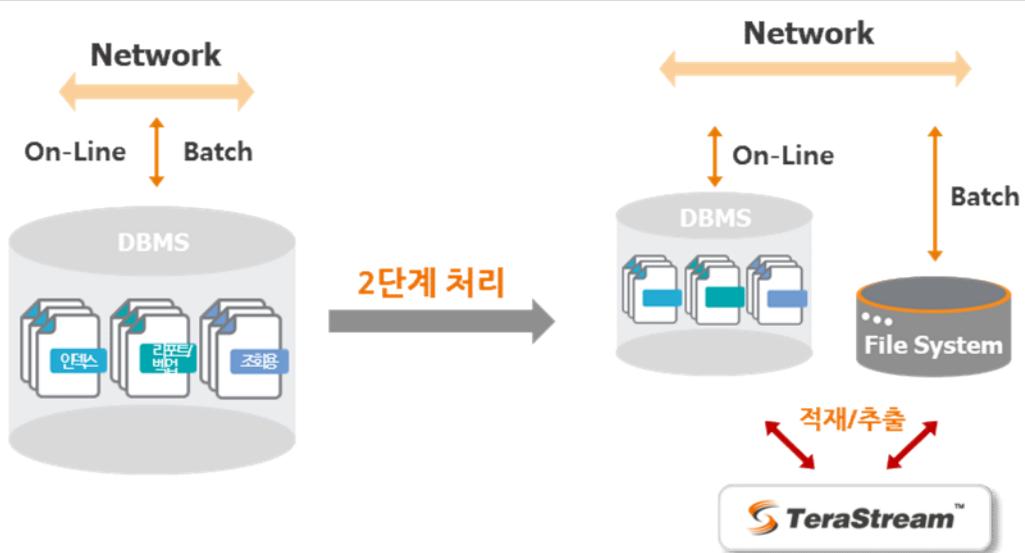


그림 38 대용량데이터 처리를 위한 2단계 데이터 처리 아키텍처

하둡분산파일시스템 기반의 현 빅데이터 플랫폼은 일반 병렬처리 기반의 파일시스템을 고속 네트워크 기반의 분산파일시스템으로 확장하고 2-Way 복합 데이터 처리 방식을 N-Way 복합 데이터 통합 방식으로 확대하여 빅데이터 레이크를 구축함으로써 데이터 처리 비용을 낮춤과 동시에 고성능을 확보하는 아키텍처를 구성하였습니다. 당사의 TeraONE™은 위 [그림2-38]의 TeraStream™ 기본 구조를 하둡 분산 파일시스템 기반으로 간단하게 확장함으로써 현 빅데이터 플랫폼의 기본 구조를 자연스럽게 구현하였습니다. 이는 과거 관계형 DBMS 중심의 단일 데이터 저장방식에서 수직 컬럼형 DBMS, 그래프DBMS, NoSQL DBMS, 하둡파일시스템, SPARK 등 다양한 데이터 저장체를 대상으로 데이터 통합 기술을 기반으로 분석 대상 데이터를 생성하여 이를 중심으로 서비스를 구현하는 빅데이터 플랫폼의 세계적인 추세에 N-Way 데이터 통합 아키텍처로 확장된 당사의 TeraONE™이 [그림2-38]에서 이 변화를 잘 반영하고 있음을 보여 줍니다.

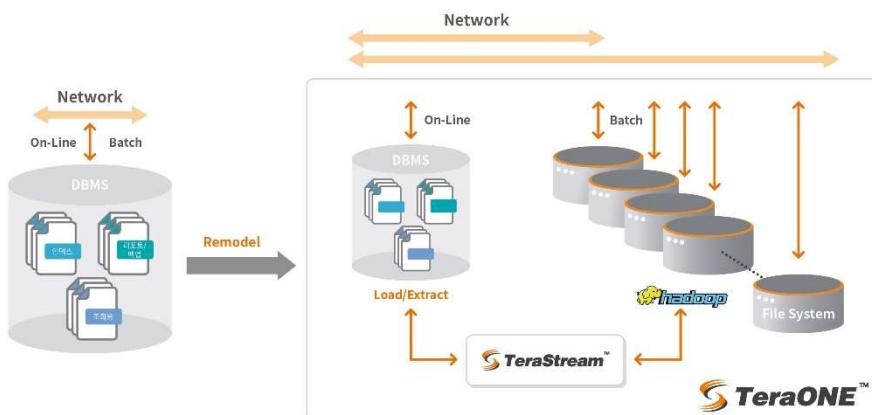


그림 39 TeraONE™ 의 N-Way 통합 기반의 빅데이터 아키텍처

2013년 TeraStream for Hadoop™이란 이름으로 첫 출시된 TeraONE™은 고객사의 모든 데이터를 대상으로 통합을 구현할 수 있도록 설계되었으며 궁극적으로 데이터 거버넌스 기술을 적용하여 전사 데이터가 표준화되고 효과적으로 관리 됨으로써 높은 품질의 데이터 성숙도를 가질 수 있도록 하고 있습니다. 이는 국내외 시장의 빅데이터 플랫폼이 단순히 DBMS기반의 데이터 카탈로그 구현을 통해 데이터의 검색 및 발견을 돋고 있는 것과는 대비되는 데이터관리체계 구축을 위한 프로페셔널한 개념으로 발전되었습니다.

위에서 설명한 개념을 기반으로 실시간 데이터 통합 개념을 추가하고 이를 스트리밍(Streaming) 데이터까지 확대하여 변경정보, 위치정보, 센서정보, 로그정보 등을 실시간으로 받아들이고 이를 비교 분석할 수 있도록 추가 제품으로 지원하고 있으며 실시간 메모리 기반으로 데이터를 통합하는 페더레이션(Federation) 기술을

적용하여 데이터거버넌스가 구현된 사내 환경에서 단 하나의 쿼리로 원하는 정보를 얻어 낼 수 있는 데이터 가상화를 구현하는 것이 TeraONE™을 통하여 가능하게 되었습니다.

3. 제품의 활용

(주)데이터스트림즈의 데이터 패브릭 기반의 차세대 빅데이터 플랫폼 TeraONE™은 아래와 같이 주요 제품들로 이루어져 있으며 각각의 활용 영역은 다음과 같습니다.

TeraONE™은 거의 모든 유형의 대용량 및 빅데이터를 다양한 원천으로부터 실시간 혹은 일괄 수집, 정제, 가공하여 분석에 활용할 데이터를 준비(Preparation)하고 궁극적으로 데이터를 분석 활용 대상 기준으로 통합(Integration)하고 통합된 데이터를 기반으로 학습, 분석, 예측을 가능하게 하는 미래 지향의 빅데이터 플랫폼입니다. TeraONE™의 버전별 제품 구성은 다음과 같습니다.

구분	Standard	Professional	Fabric
일괄 데이터 수집/준비	FACT™/TeraSort™	→	→
	TeraStream™	→	→
실시간 데이터 수집/준비		DeltaStream™	→
스트리밍 데이터 수집/준비		TeraStream BASS™	→
빅데이터 통합 저장소 및 관리시스템	하둡 및 에코시스템(오픈소스 패키징)	→	→
인공지능 분석 환경		TeraONE IDEA™	→
데이터 가상화			TeraONE SuperQuery™
데이터 거버넌스			IRUDA

가장 기본이 되는 스탠다드 버전은 대용량 데이터 일괄 수집, 정제, 가공을 통해 하둡 기반의 빅데이터 저장소에 적재함으로써 데이터 통합 및 저장 기능을 제공함으로써 기본적인 데이터 레이크를 지원합니다.

프로페셔널 버전은 거의 모든 유형의 데이터를 실시간 및 일괄 수집 및 가공하여 하둡 기반의 빅데이터 통합 저장소에 저장할 뿐 아니라, IoT를 지원하여 실시간으로 스트리밍되는 데이터 흐름 속에 특정 패턴을 찾아내고 이와 연관된 데이터분석을 가능하게 하며 수집/통합된 데이터를 인공지능기반의 데이터 분석 플랫폼으로 연계하여 머신러닝 및 딥러닝 모델을 개발하고 시각화 할 수 있습니다. 이는 데이터 수집~정제~가공~통합~저장~분석~시각화까지를 데이터 사이언티스트들이 분석 도구로 활용할 수 있는 상위 버전입니다. 당사의 차세대 플랫폼 전략이 완성되는 빅데이터 패브릭 버전은 이러한 데이터 레이크 및 IoT지원 실시간 인공지능 분석 플랫폼에서 더 나아가 데이터 거버넌스와 데이터 가상화 제품을 포함합니다. 데이터 패브릭은 일관된 데이터 관리 체계를 통하여 데이터 거버넌스를 구현하고 이를 기반으로 양질의 데이터를 확보하고 이기종 데이터 플랫폼이나 데이터 소스에 상관없이 가상화 레이어에서 하나의 데이터베이스처럼 분석을 수행할 수 있는 실시간 옴니 채널 기반의 데이터 통합을 이루는 가장 진화된 데이터 전략을 구현할 수 있도록 지원합니다. TeraONE™은 데이터 거버넌스를 기반으로 데이터 레이크에서 데이터 가상화까지 확장함으로써 기존의 데이터 레이크를 넘어서 훨씬 활용 범위가 확장된 4차 산업 기반의 데이터 서비스를 지원하는 핵심 빅데이터 플랫폼으로 활용이 가능합니다.

빅데이터 패브릭 버전을 구성하는 주요 소프트웨어 제품인 IRUDA™는 당사의 데이터 거버넌스 제품군을 바탕으로 하나의 기능으로 동작하는 복합 제품화를 이루어 일관된 데이터 관리 수준을 보장하는 종합 데이터 거버넌스 플랫폼을 구성하고 있습니다. 특히, IRUDA Navigator™는 당사의 데이터 거버넌스 제품들로부터 메타데이터를 수집하여 효과적인 데이터 검색(Search)과 발견(Discovery)을 위한 저장체를 구성하고 검색을 위한 인덱싱 체계를 구성함으로써 데이터관리자 및 데이터 분석가들에게 필수적인 기능을 제공하고 있습니다.

또 다른 중요한 구성 제품으로 데이터 가상화를 가능하게 해주는 데이터 페더레이션 제품이 있으며 이는 서로 다른 이기종의 데이터 저장소 또는 다양한 클라우드 상에 존재하는 데이터를 물리적 이동없이 분석가가 필요한 시점에 실시간으로 통합하여 활용할 수 있도록 하는 TeraONE SuperQuery™를 통하여 구현하고 있습니다.

이 기능은 데이터 거버넌스가 잘 구축된 환경에서 데이터 수집 및 준비(Preparation)의 노력을 최소화하여

분석 효율을 극적으로 높여줄 뿐 아니라 데이터 페더레이션(Data Federation) 기능과 연계하여 한번의 쿼리로 필요한 정보를 얻어내는 궁극적인 데이터 가상화를 이룸으로써 4차 산업혁명의 핵심인 디지털 전환(Transform) 혁명을 이루기 위한 핵심 기능으로 활용될 수 있는 중요한 제품입니다.

TeraONE™과 같은 차세대 빅데이터 플랫폼을 기반으로 디지털 전환된 비즈니스 통합 플랫폼을 구축하면, 이를 기반으로 실시간 옴니 채널 기반의 산업간 혹은 기업간의 복합화된 페더레이션된(Federated) 서비스를 구현하여 히스토리에 기반한 깊은 통찰력이 있는 분석(Deep Insight), 차세대 CRM을 위한 옴니 채널 마케팅, 블록체인 기반의 대중에 의한 데이터 거래소, 거래를 지원하는 데이터 안심구역, 디지털 트윈 및 메타버스용 데이터 플랫폼, 스마트팩토리/팜/시티, CDP(Customer Data Platform)/CRM(Customer Relationship Management) 위한 고객 데이터 플랫폼, 마케팅 활용 분석 및 분석 서비스 등 다양한 데이터 서비스가 가능 합니다.

아래 [그림2-40]은 디지털 전환을 통한 4차 산업혁명 시대에 활용될 차세대 빅데이터 플랫폼과 이를 활용해 구현할 수 있는 서비스를 보여 주고 있습니다.

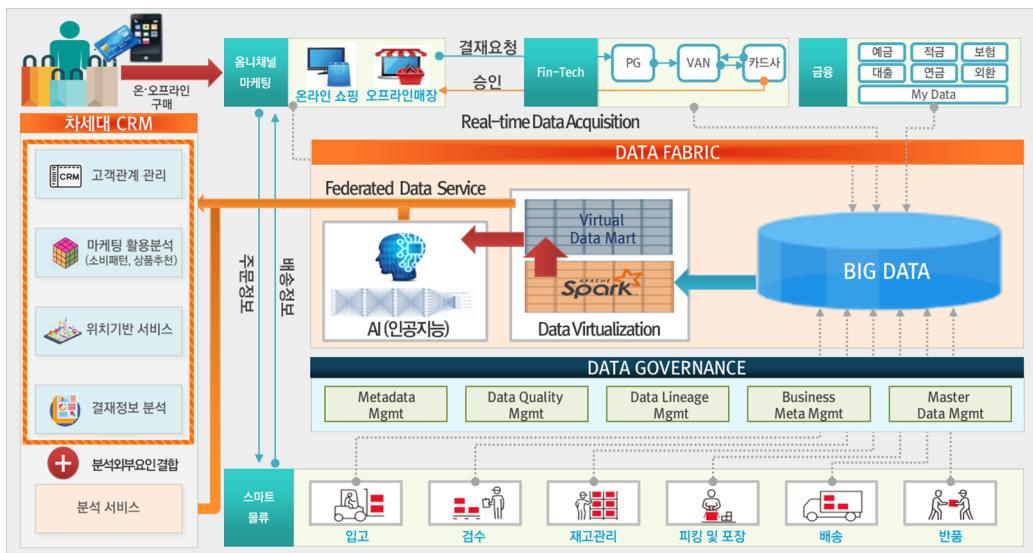


그림 40 차세대 빅데이터 플랫폼 TeraONE Fabric 활용 서비스 개념도

4. 활용 예시

구분	내용	예시 및 사례
데이터 레이크 구축	<ul style="list-style-type: none"> - 데이터 레이크는 기존의 데이터웨어하우스와는 다르게 특별히 정해진 목적없이 마신러닝 및 딥러닝 등의 인공지능 분석에 활용되는 데이터를 원시 형태로 저장하는 빅데이터 저장소 - 다양한 분야에서 수집된 정형/반정형 /비정형 등 모든 유형의 대용량 데이터를 수집/가공/적재 - 확장되는 데이터 용량에 맞춰 클라우드처럼 유연하게 확장 가능으로 데이터웨어하우스 대비 적은 비용으로 최적의 데이터 플랫폼 구성이 가능함 	<p>[당사 사례]</p> <p>포스코 클라우데라 원백 (2021) : 데이터 레이크 기반 스마트조업시스템 TeraONE™으로 원백 DB순회보험 클라우데라 원백 (2021) : 데이터 레이크 기반 IFRS TeraONE™으로 원백 서울시청 데이터거버넌스 기반 데이터 레이크 구축 (2020 ~ 2022)</p> <p>우체국금융 금융차세대시스템 구축 (2020 ~ 2022)</p> <p>농협생명 빅데이터 분석시스템 구축 (2020)</p> <p>국립서울헬스케어 빅데이터 플랫폼 구축 (2020) 외 다수</p>
옴니 채널 데이터 서비스	옴니 채널은 언제 어디서든 동일한 서비스를 받고 싶어하는 고객의 니즈를 충족시키기 위해 구매채널에 상관없이 모든 채널에서 고객이 동일한 상품/서비스를 제공받을 수 있도록 데이터를 통합 일관되게 제공하는 것	<ul style="list-style-type: none"> - 각 채널 간 동일한 고객 경험 제공을 통해 일관된 고객 응대 및 영업기회 발굴 - 온/오프라인 채널 간 정보 및 실시간 프로세스 연계를 통한 중단된 고객경험을 영업기회로 감지하여 상품판매 완성
데이터 거래소	데이터 거래소는 데이터 거버넌스 및 데이터 서비스를 구현하는 분석환경을 기반으로 데이터 공유 및 교환에서 수익을 창출하는 종합거래형 마켓플레이스를 완성하는 방향으로 발전하고 있음	<p>중국 상해/광저우 데이터 거래소</p> <p>3천 개의 데이터 속성을 기반으로 7억명 데이터 정보관리, 마케팅 부정사용탐지를 위한 고객 데이터 분석서비스를 제공하는 미국의 Axiom 데이터 거래소</p> <p>[당사 사례]</p> <p>광주 인공지능 플랫폼 (2021)</p> <p>NIA 디지털기술혁신 데이터 거래소 (2020)</p> <p>NIA 산림/교통 데이터 거래소 (2019)</p>

데이터 안심구역	<p>공공/민간이 보유한 미개방 데이터를 확보하여 보안이 철저한 안심구역을 구축,</p> <p>연구기관/학생/스타트업 등 대국민 대상으로 자유로운 분석이 가능하도록 환경 조성</p>	<p>[당사 사례] 전북도청 금융혁신 플랫폼 (2021) K-DATA 데이터 안심구역 (2019) : 데이터산업진흥원은 11 개 분야 22 개 민간 및 공공기관으로부터 미개방 데이터를 확보하여 보안환경이 구성된 안심구역을 구축하여 스타트업, 연구기관, 개인들이 자유롭게 분석하고 결과를 반출할 수 있는 여건을 마련함</p>
디지털 트윈 및 메타버스용 데이터 플랫폼	<p>현실에 적용하기 전에 디지털 트윈 기술을 활용하여 실제환경과 동일한 가상공간을 마련 다양한 실험을 시도하고 있고, 언택트 시대를 맞아하여 VR/AR을 활용한 기상환경인 메타버스 (Metaverse) 서비스가 활발하게 전개되고 있음</p>	<p>네이버의 메타버스 기반의 서비스 모델인 ZEPETO 인천광역시 디지털뉴딜 사업의 일환으로 추진하는 '현실세계 XR 메타버스 프로젝트'</p> <p>[당사 사례] 5G/디지털트윈을 활용한 자자체 및 국방 디중이용 건축물 시설안전대응 통합관리체계 구축 (2020)</p>
스마트팩토리/ 팜/시티	<p>다양한 IoT 센서에서 생성되는 데이터를 실시간으로 수집하여 제조공장, 농장물 재배현장, 도시 공공시설 등을 대상으로 인공지능 분석을 통한 스마트 서비스 구현</p>	<ul style="list-style-type: none"> - 제조공장의 공정 프로세스에 각종 센서를 통해 장애를 사전에 예측하고 제품 효율을 높이는데 활용 - 농작물 성장 과정에 IOT센서를 통해 발생된 데이터를 분석 온도, 습도 등 최적의 환경을 모색함 - 교통신호, CCTV 등 다양한 도시 공공시설에서 생성되는 데이터를 실시간 분석을 통해 교통의 원활한 흐름, 범죄예방 등 스마트 시티 구축에 활용
CDP/CRM 고객 데이터 플랫폼	<ul style="list-style-type: none"> - 고객데이터 플랫폼(CDP)은 다른 시스템에서도 접속할 수 있는 통합 고객 데이터를 지속적으로 수집/생성/관리 목적으로 구축 - 고객관계관리시스템(CRM)은 고객데이터를 중심으로 마케팅에 활용하는 목적으로 구축 - CDP와 CRM은 초기인화 마케팅을 지향하며, 온라인 채널과 실시간 데이터 처리가 가능한 기술을 활용 	<ul style="list-style-type: none"> - CDP는 다양한 고객 접점 채널에 대해 효과적인 최적의 마케팅 활동을 구성할 수 있는, 마케터가 직접 조작 가능한 통합된 고객 - 데이터베이스(가트너) CRM과 CDP를 통해 고객 데이터 기반으로 전방위로 이해해서 고객이 원하는 것을 알아내고 제품구매 상품체험 특정 광고 노출 등을 통해 여러 행동들로 유도함으로써 기업이 원하는 결과를 효율적으로 이끌어냄
마케팅 활용분석	<ul style="list-style-type: none"> - 구매나 열람 빈도 등의 통계를 근거로 하는 통계적 방식과 머신러닝, 딥러닝 기법을 활용한 AI 개인화 추천방식이 존재 - 최근에는 고객의 특성이나 취향을 반영하는 AI 개인화 상품 추천 방식이 대세 - 유저들의 유사성을 기반으로 나눠진 특정 세그먼트에 기준 고객의 방문이력 패턴의 분석 결과가 더해져, 첫 방문(신규고객) 유저에게 유리한 딥러닝 기반 상품 추천을 제공할 수 있음 	<p>연관 상품 추천 : 함께 본/함께 담은/함께 구매한 연관 상품 추천, 유사 고객 선호 상품 추천, 유사 상품 추천 세그먼트별 선호 상품 추천</p>
분석 서비스	<p>[이상거래 탐지] 데이터 객체의 패턴을 정형화한 뒤, 이를 기반으로 정상거래와 이상거래를 판별, 탐지 [그래프 알고리즘 기반 분석] <ul style="list-style-type: none"> - 객체 혹은 대상의 유사도를 기반으로 클러스터를 정의하고 그에따내 모든 노드를 연결하는 Edge의 기중치 합을 계산하여 관계성을 나타냄 - 그래프로 기사회하여 사용자가 직관적으로 이해할 수 있음 <p>[사이버 위협 Intelligence] <ul style="list-style-type: none"> - 행동특성 모델링 및 과거의 위협을 식별하는 예측분석을 통해 위협정보를 분류하여 전달하고 상황에 따른 빠른 판단을 통한 대응방안 제공 - 위협을 식별해 자동으로 위협을 탐지해 보안 경보 전송 및 발생한 보안 이벤트와 관련된 위험이^{무엇인지, 어떤 시스템이 어떤 취약점을 갖고 있는지} 등에 대한 정보 제공 가능 <p>[비정형 분석] <ul style="list-style-type: none"> - 비정형 데이터를 Opinion Mining (의견 미아님), </p> </p></p>	<p>부정수급 : 하위 신고로 세금 환급 및 지원금 등 공공 재정을 부정수급하는 범죄의 탐지</p> <p>e-커머스 : 고객 계정 도용, 상품정보 유출 및 부정 거래, 해킹 등 전자상거래 환경의 범죄 탐지</p> <p>관계 분석 : 특정 정책의 효과를 분석할 수 있는 다양한 데이터 (언론 기사, 지자체 보도 자료, 주택 가격 자료 등)를 그래프 데이터로 표현, 정책 효과에 대한 정량적 / 정성적 분석 수행으로 NLP 기반의 키워드 유형화 및 연관 관계의 예측 분석</p> <p>개인정보 탈취 관련 사고를 방지 및 파악 : 사용자의 로그인 패턴 키보드 이용 패턴 등을 분석해 디지털 신원을 파악하고 사이버 공격자의 패턴도 학습해 실제 주체가 누구인지 탐색</p> <p>감성 분석, 리뷰 분석 : 키워드와 연관된 감성 어휘의 빈도수를 분석해 중립, 긍정, 부정으로 분류하고 그 강도를 평가</p>

	<p>Semantic (의미론), Syntactic (통사론)의 규칙에 따라 문서를 분류하며, 마신러닝 기반의 키포인트 주제어를 빌글하여 카테고리의 매팡을 통해 정확한 의견 분석 및 분류</p> <ul style="list-style-type: none"> - 텍스트 내의 의견 정보를 파악하기 위해 문장구조, 문장 간의 관계, 어휘 분석을 진행 	
5. 구성 제품의 개념 및 활용		
가. 데이터 통합(Data Integration) 기술		
1) 데이터 통합 및 빅데이터 준비 도구		
<p>(TeraONE Standard Version : FACT™/TeraSort™, TeraStream™ TeraONE Professional Version : TeraStream for BASS™, DeltaStream™)</p>		
가) 정의		
<p>다양한 서버 환경에서 원천 데이터를 빠르게 추출하여 빅데이터 저장소로 전송, 가공 처리 하는 데 데이터 통합 기능을 제공합니다.</p>		
나) 필요성		
<p>데이터 통합 기술은 빅데이터 플랫폼의 데이터 구축 및 운영을 위한 핵심 기술로, 빅데이터를 추출 및 가공 하는 기술 뿐만 아니라 기존 레거시 시스템으로부터 원천 데이터를 빅데이터 저장소로 전송하는 기술도 포함합니다. 따라서, 데이터 통합 기술은 빅데이터 분석을 위해 매우 중요한 전처리 기술이며, 수집된 데이터를 의미 있게 분석하기 위한 핵심적인 역할을 합니다. 이를 위해 다음과 같은 기능이 요구됩니다.</p> <ul style="list-style-type: none"> - 초고속 데이터 추출 및 소팅 (정렬) 엔진을 활용하여 데이터 추출 및 가공 성능을 향상, 시스템 자원을 효율적으로 이용하여 부하를 최소화해야 합니다. (FACT™/TeraSort™) - 대용량 데이터 배치 처리가 가능해야 하며, 데이터 분석에 필요한 데이터 전처리를 위해 다양한 데이터 변환 함수 및 사용자 정의 함수를 제공해야 합니다. (TeraStream™) - Oracle, Sybase, DB2 등 기존 Legacy RDBMS 외에 Hive, HDFS 등 빅데이터 저장소로 편리하게 연동할 수 있어야 합니다. (TeraStream™) - 실시간 데이터 변경 적재가 가능하여야 합니다. (DeltaStream™) - 빅데이터 시대가 되면서 데이터 형태가 일정하지 않은 반정형 및 비정형 데이터의 스트리밍 처리를 위해 Apache Flume, Apache Kafka와 같은 오픈소스를 패킹하여 제품의 기본 기능으로 제공해야 합니다. (아파치 라이선스 2.0에 의거 오픈소스 패킹 제공) - 그래픽 유저 인터페이스(GUI) 기반 분산 병렬 처리 프레임워크인 Map/Reduce 개발 지원이 용이해야 합니다. (TeraStream™) 		
다) 장점		
<ul style="list-style-type: none"> - 고성능 보장 : 대용량 데이터를 파일 처리 방식(초고속)으로 추출하여 대량의 데이터 배치 처리를 빠르게 처리할 수 있으며, 병렬 프로세싱을 통해 데이터 가공 작업의 속도를 향상시킬 수 있습니다. - 자원 이용시 고효율성 보장 : 데이터베이스와 분리된 시스템 자원을 활용함으로써 데이터베이스 자체 부하를 감소시킵니다. 온라인 데이터베이스 작업 시 자원에 부하를 주지 않고 배치 작업 수행이 가능하며, CPU, Memory, I/O의 효율적인 자원 할당이 가능합니다. - 높은 업무 생산성 : 빅데이터 통합 기술은 대부분 오픈소스로 이루어져 있어 개발 복잡성과 난이도를 줄여줍니다. 		

도가 높으며, 개발 공수 및 인건비 부담이 높은 반면 당사 제품은 꼭 필요한 기능은 오픈소스를 활용하지만 빅데이터 통합 주요 기능은 당사에서 개발한 제품을 활용하도록 구성하여, 통합 그래픽 유저 인터페이스 (GUI) 환경을 제공하여 프로그램 개발을 편리하게 할 수 있습니다. 오픈소스로만 데이터 통합할 때 대비 40% 이상의 개발 생산성이 제고됨을 50여개의 빅데이터 프로젝트를 수행하면서 검증된 제품입니다

- 효율적 기능 제공 : 데이터 배치처리 기술을 확장하여 온라인 배치 처리, 준 실시간 데이터 처리 업무에 적용 가능합니다.
 - 높은 호환성 : 서버 간의 분산 스케줄링 및 통합 스케줄링을 지원하며, 분산된 데이터 통합 솔루션 환경에서 통합 메타데이터 관리를 지원합니다.

나. 빅데이터 플랫폼(Big Data Platform) 기술

1) 빅데이터 통합 저장소

(아파치 라이선스 V2.0을 따르는 오픈 소스 하둡 에코시스템 패키징)

가) 정의

정형, 반정형, 비정형 등 모든 유형의 데이터를 실시간 수집하여 분석 및 서비스 목적으로 저장 및 가공할 수 있는 파일 분산 병렬 처리 시스템입니다.

나) 필요성

빅데이터 시대에 증대된 데이터의 다양성을 수용할 수 있는 효율적이고 생산적인 데이터 처리 기술이 필요하며, 이는 기존의 정형 데이터 처리 중심의 데이터베이스 기술의 제약을 해소할 수 있어야 합니다. 아래 [그림 2-41]는 저장소가 지원하는 다양한 데이터 형태와 사례를 표현 하였습니다.

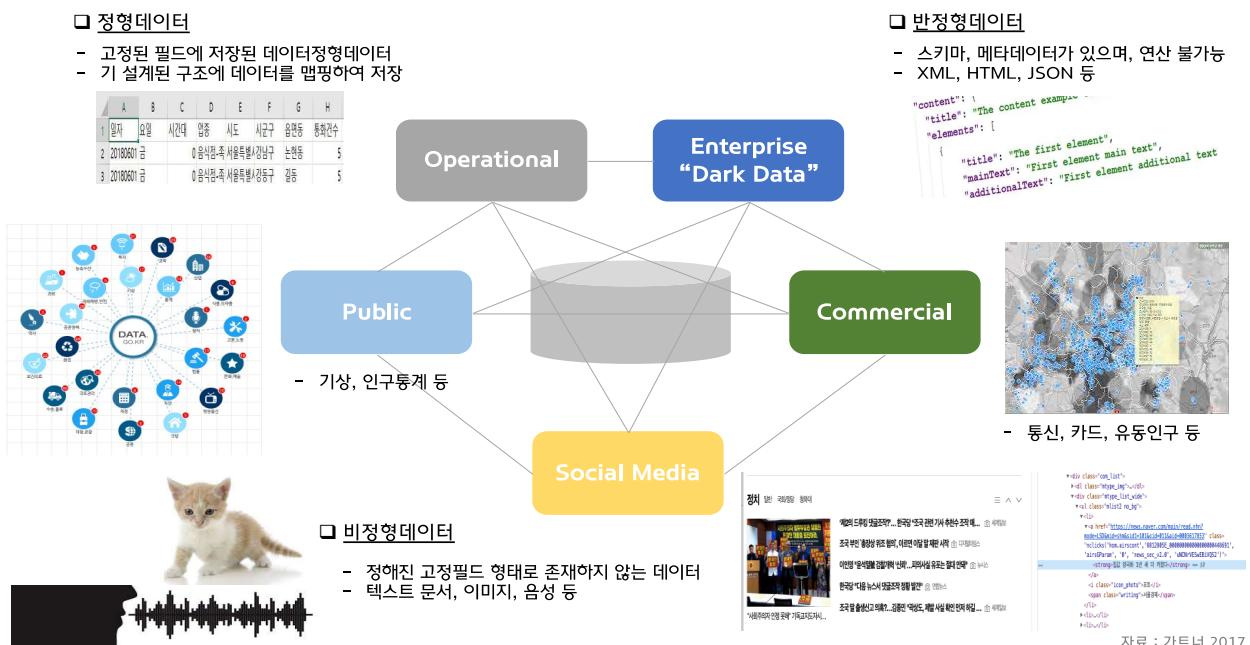


그림 41 빅데이터 시대, 확장된 데이터 유형

사람이 직접 컴퓨터에 입력하여 생성하는 데이터 외에도 모바일에서 실시간 발생하는 각종 로그 데이터, 인터넷 상에서 사용자가 메뉴를 클릭할 때 발생하는 이벤트 데이터, 각종 기계에서 발생하는 센서 데이터, 인터넷 미디어에서 발생하는 동영상, 음성, 이미지 등의 컨텐츠 데이터 등으로 데이터 처리 대상이 확대되고 있는데, 이러한 데이터는 기존의 기술인 데이터베이스로 처리할 수 없습니다. 특정한 모양이 없는 형태로 데이터가 무한 저장될 수 있어야 하며, 데이터 형태나 용량에 상관없이 저렴하게 저장 및 관리할 수 있는 저

장소가 필요합니다.

다) 장점

- 사용 용이성 : 빅데이터 관련 기술은 구축 및 운영이 어려운 오픈소스로 구성되는데, 당사는 20년간 시장에서 검증된 자체적인 데이터 통합 기술을 오픈소스와 접목하여 GUI 기반의 사용하기 쉬운 플랫폼으로 제공함으로써, 기존 데이터웨어하우스 담당자라면 새로운 기술에 거부감 없이 손쉽게 빅데이터 저장소로 접근하여 데이터를 처리할 수 있습니다.
- 확장성 : TeraONE™은 아파치 라이선스 2.0을 따르고 있으며, 빅데이터 저장소의 클러스터를 운영 및 관리하는 아파치 암바리를 기반으로 서비스 스택을 자체적으로 개발하여 패키징 한 상용 배포판으로, 당사의 핵심 원천 기술인 초고속 데이터 추출 및 적재 기술을 기반 기술로 하고 있습니다. 오픈소스 특성상 최신성을 유지하여야 하는데, 판교 기술연구소의 전담 개발조직이 제조사 레벨의 최신성 및 제품의 품질을 보장합니다. 필요한 기술은 자체 개발하여 확장하거나, 최신 오픈소스를 검증하여 신속하게 패키징하여 시장의 요구에 대응합니다.
- 비용 절감 : 경쟁하고 있는 외산 제품의 가격 정책은 구독료(Subscription) 모델만 제공하지만, 당사는 구독료 외에도 영구 라이선스로도 판매하고 있어 고객의 선택 폭에 제약을 두지 않습니다. 영구 라이선스와 구독료 방식을 비교했을 경우 5년 TCO (총 소유 비용, Total Cost of Ownership) 관점에서 영구 라이선스 방식이 2배 가량 저렴합니다.
- 유연성 : TeraONE™은 자체적으로 AI 플랫폼인 TeraONE IDEA™를 보유하고 있지만, 시장의 요구에 따라 타사의 제품과도 유기적으로 연계하여 제공할 수 있는 유연성을 자랑합니다. 삼성 SDS의 AI Brightics, SK AccuInsight와 같은 타사 AI 플랫폼과 TeraONE™이 연계하여 빅데이터 사업 생태계를 조성해 나가고 있습니다.

라) 활용 사례

고객사	주사업자	내용
우체국금융	SK C&C	• 금융 차세대 시스템 구축 (TeraONE)
서울특별시청	KT-DS	• 빅데이터 기반 데이터 저장소 구축 1차사업 • 전체 408개 시스템 대상으로 데이터 레이크 구축, 1차년도 수집대상 200여개
정보통신산업진흥원	컨소시엄	• 5G 디지털기반 국방 시설 안전 및 에너지 효율 체계 구축 사업하는 3개년 사업. Digital twin 기술을 활용한 민간주도 혁신성장 견인 및 시설을 안전관리 체계 마련을 통한 공공의 선도적 수요창출을 위해 실증 추진
인천국제공항공사	삼성SDS	• 빅데이터 플랫폼 구축 (TeraONE + 삼성 AI Brightics) • 인천국제공항에 특화된 AI/빅데이터/IoT 융합플랫폼 기반 구축
농협생명	삼성SDS	• 빅데이터 분석 시스템 구축 (TeraONE + SuperQuery + 삼성 AI Brightics)
근로복지공단	EDS	• 재정추계 시스템 및 산재 보험 요율 시스템 재 구축 (TeraONE + SuperQuery)
신용보증기금	데이터스트림즈	• 신용보증기금이 중소기업 마이데이터 활용지원 및 기업금융 환경의 변화를 기획 - 중소기업 마이데이터 전송, 기업성장MAP, 경기 예측

그림 42 2020년 주요 사례

고객사	내용
현대자동차	<ul style="list-style-type: none"> • 480개 노드 대상 (데이터 6PB) • 40개 멀티클러스터 관리기능 • 전사 빅데이터 Job 관리 워크플로우 기능 제품 반영 (기존 CDH 전환 검토 중)
국립암센터	<ul style="list-style-type: none"> • 하위 10개 빅데이터센터의 수집 데이터 관리를 위해 중앙메타 기반 하위메타 연계 기능 구현 (멀티 메타데이터 관리 기능 구현)
임업진흥원	<ul style="list-style-type: none"> • Hive를 활용한 빅데이터 저장소에서의 품질관리 기능 개발 • 데이터 상품검색 기능을 Elasticsearch 엔진에 태깅하여 Graph DB기반으로 데이터 맵 기능 구현 (품질, 데이터맵 기능 제품 반영) • 데이터 거버넌스 기반의 데이터 마켓플레이스 구현
데이터산업진흥원	<ul style="list-style-type: none"> • DATA LAKE 기반의 빅데이터 저장소 및 분석 플랫폼 구축 경험
한국철도공사	<ul style="list-style-type: none"> • 외산 C사 빅데이터 플랫폼 원백 사례

그림 43 2019년 주요 사례

다. 인공지능(AI) 및 어날리틱스(Analytics)를 위한 분석 플랫폼 기술

1) 인공지능 기반 빅데이터 분석 환경 (TeraONE IDEA™)

가) 정의

인공지능 서비스를 구현하기 위한 개발 및 배포 자원의 통합적인 관리(버전 및 형상 관리)와 인공지능 분석 환경 제공(머신러닝 및 딥러닝 개발 환경), 그리고 자원 모니터링 기능을 제공해야 합니다. 최근에는 클라우드 네이티브 기반의 개인 맞춤형 분석 환경 제공에 대한 수요가 증가되고 있으므로 클라우드 상에서 클릭 몇 번만으로 사용자 또는 해당 부서에 필요한 인공지능 분석 자원을 할당할 수 있어야 합니다.

TeraONE IDEA™는 머신러닝 및 딥러닝 응용프로그램을 구축, 학습, 배포하는 인공지능 분석 플랫폼으로 기업의 분석 생산성을 촉진하는 데이터 사이언스 도구로 활용할 수 있습니다. TeraONE™ 프로페셔널 버전에서 자동 배포할 수 있도록 구성하였으며, 도커와 컨테이너 기반으로 개인 맞춤형 분석 환경을 제공함으로써 분석을 하는 데이터 사이언스팀과 데이터를 준비하는 IT팀 간의 협업이 원활할 수 있도록 지원합니다. 데이터 사이언티스트가 인공지능 분석을 위해 사용하는 R, Python 개발 환경 뿐 아니라 편리한 셀프 분석 및 시각화 기능도 제공합니다. 분석 프로젝트 공유 및 재활용, 분석 모델 및 형상 관리, 데이터 분석 자원 및 환경을 요청하고 제공받을 수 있는 편의성을 제공합니다. 또한 TeraONE™와 함께 사용할 경우, 빅데이터 플랫폼과 연계하는 분석 파이프라인을 제공하여 인공지능 플랫폼에서 활용되는 다양한 데이터 셋에 쉽고 편리하게 접근할 수 있습니다.

나) 필요성

대량으로 발생하는 빅데이터를 수집하는 이유는 기준에 관리하지 못하던 데이터를 분석하여 부가가치를 창출하는 서비스를 구현하기 위함입니다. 인공지능 서비스 구축에 필수적인 기계학습 과정은 “데이터 준비 -> 학습 데이터 생성 -> 인공지능 모델 생성 및 학습 -> 검증”的 단계를 거치게 되는데 이러한 각 프로세스를 효과적으로 실행할 수 있는 인터페이스의 제공은 인공지능에 대한 숙련도가 낮은 개발자들의 서비스 구축에 대한 진입 장벽을 낮추는 것이 필요합니다. 오픈소스 기반의 딥러닝 라이브러리 추천 및 최신 버전 관리 등이 이러한 프로세스의 품질을 높여주는 중요한 요소들입니다.

다) 장점

- 사용 용이성 : 빅데이터 저장소와 인공지능 분석 환경의 기능을 포털 환경으로 제공하여, 데이터

사이언티스트들은 복잡한 빅데이터 인프라를 이해하지 못하더라도 직관적인 포털을 통해 분석 자원 할당 요청, 분석 툴 선정, 분석 모델과 데이터 공유 및 활용, 형상관리, 커뮤니티 참여, 개인 맞춤형 분석 환경 구성 등의 기능을 쉽게 활용할 수 있습니다.

- 제품간 호환성 : TeraONE IDEA™는 TeraONE™을 통해 배포하여 쉽게 설치할 수 있습니다. 빅데이터 저장소와 유기적으로 연계되어 분석의 생산성을 제고합니다. 인공지능 알고리즘을 코드 레벨에서 작성할 수 있는 파이썬, 스칼라, R등의 언어를 지원하며 내장된 시각화 툴과 연계하여 사용 가능합니다. GUI 기반으로 모델링할 수 있는 기능도 추가적으로 제공합니다.
- 비용 효율성 : 오픈소스 기반 코드 레벨 분석 환경과 GUI 기반 분석 환경을 옵션으로 나누어 제공함으로써, 초기 단계 대규모 투자를 하지 않더라도 사용자의 분석 수준에 맞춰 확장해 나갈 수 있습니다.

라) 활용 사례

고객사	사업명	내용
정보통신산업진흥원	5G 기반 디지털 트윈 시설물 안전 실증	시설물의 안전 이상 징후 및 노후화 예측 시뮬레이션 인천광역시, 전라남도, 안양시, 여수시 4 개 시설물에 안전계측센서와 환경센서, CCTV, 전력검침 등 5G를 활용한 데이터 계측
우정사업본부	금융 차세대 시스템 구축	빅데이터 기반 정보활용 고도화
한국지역정보개발원	차세대 지방세입정보시스템 구축	지방세 데이터 서비스(체납분석, 세수예측)

상기 사례 중 5G 기반 디지털 트윈 시설물 안전 실증은 KT GiGAsafe 플랫폼을 통하여 디지털트윈, 빅데이터, AI 기술을 활용하여 실시간 시설안전계측, 시설위험진파, 시설 노후화 상태 파악, 이상 징후 및 위험 시점 예측분석, 재난대피 경로지원, 감염자 이동경로 분석, 환경요소변동 시뮬레이션 등 시설관리/재난대응/체계관리 관련 제반 서비스를 제공하는 사업입니다.



그림 44 디지털트윈 기반 시설물별 공통서비스 시나리오 예시

세부 사업 내용으로 인천광역시 상업시설, 전라남도 의료시설, 안양시 체육시설, 여수시 문화시설 총 4개 랜드마크 시설물을 대상으로 하여 당사는 IoT센서에서 수집되는 데이터와 각종 레거시 데이터를 TeraONE 플랫폼에 적재한 후 AI분석 기법을 통해 디지털트윈 기반 시설안전 서비스를 실증하였습니다.

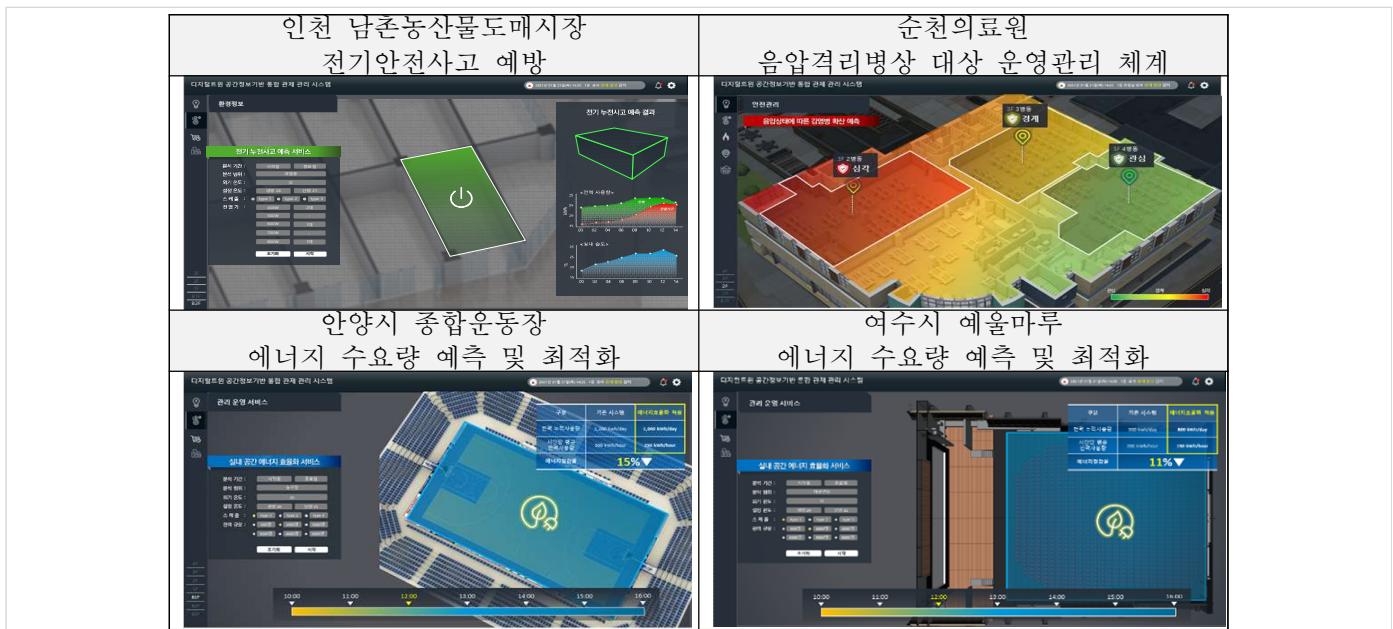


그림 44 디지털트윈 기반 시설물별 시뮬레이션 서비스 예시

라. 데이터 패브릭(Data Fabric) 기술

1) 데이터 가상화 (TeraONE SuperQuery™)

가) 정의

데이터 플랫폼의 기술이 다변화 되면서 데이터마트 및 데이터웨어하우스 저장소, 클라우드 플랫폼, 빅데이터 플랫폼 등 이기종의 다양한 데이터 소스를 물리적으로 하나의 거대한 빅데이터 저장소로 통합하지 않고도, 데이터 가상화 기술을 이용해 하나의 데이터처럼 분석할 수 있습니다.

(기존 데이터베이스 : IBM, Oracle, SybaseIQ, Teradata, Netizza, Vertica 등)

오픈소스 기반 데이터베이스 : PostgreSQL, MariaDB 등

하둡 기반 빅데이터 플랫폼 : HIVE, SPARK, NOSQL 등

클라우드 플랫폼 : 아마존, 구글, 마이크로소프트, G-Cloud 등)

나) 필요성

빅데이터 플랫폼은 기존의 데이터베이스와는 다르게 대용량의 데이터를 얼마든지 확장하여 물리적으로 저장이 가능하다는 장점이 있습니다. 하지만 빅데이터로 수집할 수 있는 데이터가 무한히 확대되면서 언제까지 모든 데이터를 하나의 저장소에 저장해야 하는지에 대한 이슈가 대두되었습니다. 또한, 클라우드 기반 인프라가 기본이 되면서 다양한 클라우드에 존재하는 데이터를 통합하는 것도 현실적으로 불가능한 일입니다. 서로 다른 기술 기반의 데이터 소스를 통합하려면 표준화 작업에 엄청난 노력을 들여야 합니다. 당사의 데이터 가상화 기술은 당사의 제품 전략의 핵심인 메타데이터를 기반으로 합니다. 데이터 모델링 없이 (물리적 또는 논리적 데이터 모델이 없어도) 사용자가 수행하고자 하는 쿼리를 즉시 해석하여 해당 데이터 소스가 존재하는 물리적인 플랫폼 정보를 메타데이터를 기반으로 파악하여 데이터를 가상화 레이어로 가져올 수 있습니다. 데이터 가상화 레이어에 일단 탑재되면 분산 메모리 기반으로 분석이 가능하므로 매우 빠른 속도로 원하는 결과를 얻을 수 있습니다. 저렴한 리눅스 기반 하둡 아키텍처를 활용하므로, 고가의 메모리 데이터베이스와는 차원이 다른 기술입니다.

다) 장점

- 사용 용이성 : 메타데이터 기술을 기반으로 데이터 모델 사전 정의 없이, 사용자가 수행한 쿼리를 해석하여 필요한 데이터가 저장되어 있는 위치에서 데이터를 즉시 가져올 수 있습니다.

- 유연성 : 사전에 정해진 쿼리를 수행하는 것이 아니므로 제약 사항이 없으며, 사용자가 활용하고 싶은 데이터에 대한 확장이 가능합니다.
- Self-BI 지원 : 인공지능 분석 환경과 연계하여 IT담당자의 지원없이 필요한 데이터를 즉시 제공 받아 활용 가능합니다.
- 분석 성능 향상 : 분산 메모리 기반으로 분석 및 쿼리를 수행할 수 있어, 기존 방식에 비해 10배 이상의 분석 성능을 보장합니다.

라) 활용 사례

고객사	사업명	내용
근로복지공단	재정추계 요율 시뮬레이션	RDBMS인 Oracle과 빅데이터 Hive 데이터를 물리적 통합 없이 데이터 가상화 레이어에서 분석
NH농협생명	빅 데이터 분석 시스템	삼성SDS AI Brightics에서 분석하고자 하는 이기종의 데이터를 가상화에서 통합 지원 (Oracle, Hive 등)
지역정보개발원	차세대 지방세입 시스템	전국 시군구 체납 데이터를 가상화로 통합하여 분석 결과 제공

마. 데이터 거버넌스(Data Governance) 기술

- 정의

기업 내·외부에서 산재해 있는 다양한 데이터 소스들의 메타데이터 정보, 즉, 용어, 표준, 속성, 도메인, 오너십, 개인정보, 보안정보, 수집/저장되는 위치 및 경로 등 참조, 분석, 활용을 위한 정보들을 정의하고, 분석가나 데이터 관리자 등 데이터 사용자들이 정해진 보안 정책 아래에서 누구나 쉽게 찾고 공유 할 수 있게 하여, “관리자들만의 시스템이 아닌 데이터 소비자들이 공유하면서 정확한 의미로 이해할 수 있으며 데이터 내부까지 정밀하게 분석하고 추적할 수 있는 데이터 관리체계”가 요구되고 있습니다.

- 필요성

데이터 거버넌스 사업은 빅데이터 활용 분석 시장이 확대되면서 동반 확장세를 보이고 있습니다. 그동안 IT시스템을 운영하면서 보유하고 있는 데이터의 종류와 양이 엄청나게 늘어났지만 어디에 어떤 데이터가 존재하는지, 내가 필요한 데이터가 확보 가능한지, 수집된 데이터가 활용이 가능한지, 누가 데이터를 만들어서 어디로 이동되는지 등을 파악하여 그 구조를 적시에 한눈에 볼 수 있도록 통합된 뷰를 제공함으로써 데이터 기반의 업무를 지원하는 전사 관점의 데이터 거버넌스 체계가 필요합니다.

- 장점

IRUDA™의 장점은 고객사가 데이터 관리체계에 대한 개념이 성숙되지 못해서 데이터 거버넌스 제품이 종합적으로 구성되어 구축 요건이 없다고 하더라도 데이터 거버넌스의 기반이 되는 메타데이터 관리 솔루션인 MetaStream™ 도입을 통하여 빅데이터 활용 기반의 데이터 거버넌스 체계를 완성해갈 수 있다는 점입니다. 당사의 데이터 거버넌스 제품군인 MetaStream™, MetaStream for BizData™, QualityStream™, Q-Track™은 데이터 표준 사전을 생성하여 데이터베이스에서 쓰이는 기술용어를 표준에 따라 관리하고 이를 기준으로 하여 데이터 품질을 검증하고 데이터 흐름을 관리하는 등 데이터베이스 차원의 데이터 활용의 정확도를 보장 해주는 역할을 수행합니다. 특히, 최근에는 기술데이터 뿐 아니라 업무 용어까지 표준화하고 관련 메타데이터를 초상세화하여 전사 혹은 시스템 별로 데이터의 공유를 원하는 시장이 증가 추세이며 당사의 제품군은 이를 지원하여 초상세화 된 업무용어와 관련 시스템의 기술용어를 연결하고 공유하는 기능을 제공하고 있습니다. 주기적인 데이터베이스 카탈로그(DBC) 추출 정보와 모델 기반의 메타승인 정보와의 정합성 검증 모니터링, 승인된 테이블의 데이터 값이 데이터 규칙 또는 업무 규칙에 맞는지에 대한 정확성 검사, 오류가 발생했을 때 그 원인이 어디에 있는지 오류를 추적하는 기능을 확인할 수 있습니다.

또한, 데이터를 만들 때부터 폐기할 때까지 생애주기(Life Cycle)에 따라 데이터 가치를 높이기 위한 제반 관리체계를 지원합니다.

- 활용 사례

고객사	사업명	내용
KB국민은행	차세대 거버넌스 체계 수립 및 구축	메타데이터, 품질관리, 비즈니스 메타데이터 및 데이터 흐름을 포함한 전사 데이터 거버넌스 포털 구축
금융결제원	데이터 통합 체계 구축	비즈니스 메타데이터 기반 데이터 카탈로그 구축
신한오렌지 생명	IT통합 차세대 구축	메타데이터 기반 신한생명과 오렌지 생명 통합, 탐색과 발견 기능 제공
서울시청	빅데이터 통합 저장소 기반 데이터 거버넌스 구축	메타데이터 기반 빅데이터 통합 저장소 구축, 탐색과 발견 기능 제공

상기 활용 사례 중 특히 KB국민은행 사례는 금융권의 벤치마킹 대상이 될 정도로 빅데이터 시대, 꼭 필요한 데이터 거버넌스 모델로 성공적으로 수행되었습니다. 데이터 거버넌스 기반의 데이터 플랫폼 구축을 통해 필요한 데이터를 적시에 정보를 검색하고 사용자 중심의 데이터 서비스 기반 구축을 통해 활용성과 접근성을 높여 사용자 만족도가 증가하고 있습니다. 메타기반 통합 검색 제공으로 활용성을 제고하고 데이터 위치, 설명, 의미 및 연관정보를 제공하고 계정계-정보계-BI보고서 간 전사 데이터 흐름을 한눈에 파악하고 데이터 오류 등을 손쉽게 추적할 수 있습니다. 데이터 사용, 미사용 현황 파악 및 미사용 데이터 연관 관계 분석을 포함하여, 데이터 이동경로와 업무프로세스 간 연관분석을 통해 데이터 활용도 측정을 지원합니다. 전사 각 업무별 비즈메타를 활용하여 은행 전체 또는 각 그룹/부서별 데이터 현황을 한 눈에 파악하고 사용자 중심의 Self BI 서비스 및 사용 편의성 제고로 IT 부서가 아닌 현업부서 중심의 분석 데이터 활용으로 신속·정확한 의사결정에 기여합니다.

1) AI 기반 데이터 카탈로그 서비스 (IRUDA Navigator™)

가) 정의

데이터 카탈로그는 기업이 보유한 데이터 자산 목록을 생성하고 관리하는 도구로, 사용자들이 기업 내외에 분산된 데이터를 쉽게 찾아 활용할 수 있도록 지원합니다. 데이터 거버넌스 기술이 총망라되어야 구현이 가능합니다. 데이터 표준화를 기반으로 한 용어사전, 기술 메타데이터, 비즈니스 메타데이터, 활성 메타데이터 등이 관리되어야 합니다. 또한, 데이터의 품질수준 확보를 위해 데이터 관리 정책에 의거한 데이터 품질 관리가 포함됩니다. 사용자들이 빅데이터 저장소에 저장되어 있는 데이터 셋을 검색하거나 분석 환경에서 활용한 데이터 셋에 대한 로그를 활성 메타데이터로 정의합니다. 이러한 사용자 경험 데이터를 학습시켜 넷플릭스나 유튜브처럼 필요 데이터 셋 추천 서비스, 데이터 셋 간 유사도 및 연관 분석 서비스 등을 AI 기반으로 제공합니다.

나) 필요성

기업에 존재하는 데이터가 어디에, 어떤 형태로 존재하는지 알 수 있습니다. 즉, IRUDA Navigator™는 전사 메타데이터를 기반으로 관련 솔루션과 연동, 연계되어 서비스 되고 있습니다. 데이터가 있는 곳에 찾아 갈 수 있고, 데이터 항목을 열람하고, 필요 시 데이터를 볼 수 있습니다. 전사 거버넌스 관점으로 데이터를 찾고, 현상을 이해하고, 통합 키를 생성하여, 데이터를 서로 연결하여 통합적인 인사이트를 제공할 수 있는 기반을 제공합니다.

기존의 데이터 거버넌스는 IT부서에서 데이터 관련 표준을 시스템 운영에 적용하여 운영 효율성을

제고하기 위한 목적이 컸습니다. 빅데이터 시대에 데이터에 대한 관심은 IT부서를 떠나 데이터를 전문적으로 다루는 조직 (CDO, 데이터/디지털 총괄 부문)으로 이관되면서 빅데이터 활용성 강화를 위한 데이터 거버넌스 필요성이 대두되고 있습니다.

- 기업의 데이터 자산에 대한 용이한 파악이 가능해야 합니다.
 - 어떠한 데이터를 보유하고 있는가?
 - 보유하고 있는 데이터의 내용이 무엇인가?
 - 데이터는 신뢰할 수준인가?
 - 개인정보 등 컴플라이언스 관련 데이터는 어떻게 관리할 것인가?
 - 이를 활용해 어떠한 비즈니스 가치를 도출할 수 있는가?
- 데이터 관리 관점의 변화가 요구되고 있습니다.
 - IT 중심의 시스템 운영 관점에서 협업 중심의 데이터 활용관점으로 전환
 - 데이터 저장 및 통제 중심에서 데이터 접근 및 사용 중심으로 변화
- AI 기반의 데이터 활용성 강화 서비스 제공이 요구되고 있습니다.
 - 활성 메타데이터 (검색이력, 사용이력 등) 기반 메타데이터 특징 학습
 - 유사도 판별 및 연관 분석을 통해 데이터 셋 추천 서비스 제공
(용어/도메인/데이터 셋 추천)
 - 데이터 자산의 시각화 (그래프 모델링 기반의 데이터 맵 제공)

다) 장점

- 메타데이터 중심의 기술 확장성 : 당사의 제품 전략은 m-DOSA(metaData Oriented Service Architecture)로써 하나의 단일한 메타데이터를 기반으로 전 제품이 유기적으로 연계되어 데이터에 대한 토클 솔루션을 제공하는 것입니다. 이는 Gartner와 같은 글로벌 리서치 회사에서 최근 트렌드로 주장하고 있는 Unified Metadata의 개념과 일맥 상통하지만, 당사는 이미 2005년부터 메타데이터 기반으로 데이터 관리 체계에 대한 기술 연구 및 제품 공급으로 시장에서 충분히 검증을 해 오고 있습니다. 경쟁사 제품들이 기술메타데이터 제공에 머무르고 있지만, 당사는 비즈니스 메타데이터, 컨텐츠 메타데이터, 텍스트 메타데이터, 통계 메타데이터, 도서 메타데이터, 활성 메타데이터로 지속적으로 확장하며 시장을 선도하는 기술 역량을 자랑합니다.
- 빅데이터 활용성 제고 : 데이터 카탈로그를 통해 분석하고자 하는 데이터 셋을 쉽게 식별하고 탐색 할 수 있도록 지원할 수 있습니다. 또한 데이터 사이언티스트들이 검색하거나 분석에 활용한 데이터 셋에 대한 정보를 인공지능 기반으로 학습시켜 사용자별 데이터셋 별 다양한 추천 서비스를 제공하여 사용자 관점의 활용성을 강화할 수 있습니다.
- 활용 가능 분야

구분	내용
데이터 카탈로그 서비스	빅데이터 플랫폼 도입한 기업이나 기관에서 빅데이터 활용성 강화를 위해 데이터 카탈로그 구축(SK이노베이션, LG에너지솔루션 등)
계열사 간 데이터 공유 체계	지주 회사 중심으로 계열사 간 데이터 통합이 어려운 경우 데이터 카탈로그 기반 데이터 공유 체계 구현(SK지주, 하나금융지주, 우리금융지주 등)
데이터 거래소	데이터 거래소 내의 데이터 상품 카탈로그 구현(산림/교통/디지털자산혁신 빅데이터 거래소 등)
데이터 안심구역 서비스	폐쇄된 빅데이터 저장소 내에서 데이터를 분석할 수 있는 서비스로 불특정 다수의 분석가들이 활용하는 플랫폼으로 데이터 카탈로그 서비스가 필수적임(K-DATA, 전북도청 금융혁신 플랫폼, 광주 인공지능 플랫폼 등)

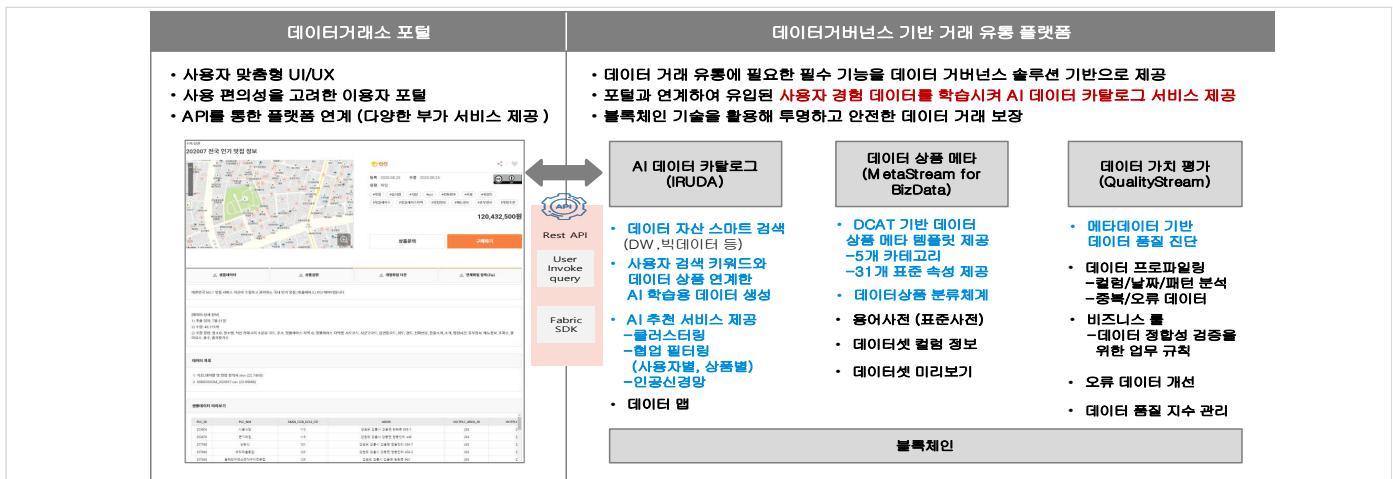


그림 45 데이터 거버넌스 기술 기반 AI 데이터 카탈로그 서비스 활용 가능 분야 및 주요 기능

라) 활용 사례

고객사	사업영	내용
임업진흥원	산림 데이터 거래소	데이터 거래를 위한 데이터 상품용 카탈로그 서비스 제공
교통연구원	교통 데이터 거래소	
한국산업기술평가원	디지털기술혁신 데이터 거래소	Apriori 알고리즘 기반 데이터 맵
한국데이터산업진흥원	데이터 안심구역	포털에서 데이터 탐색을 위한 데이터 카탈로그 서비스 제공
한국기업데이터	데이터 현황관리 시스템 구축	메타데이터 기반 데이터 탐색 및 데이터 미리보기 기능 제공
LG에너지솔루션	데이터 카탈로그 구축	

2) 데이터 표준화 및 메타데이터 관리 도구 (MetaStream™)

가) 정의

메타데이터는 데이터에 대한 설명 정보로 기업이나 기관 내의 모든 사람들이 데이터의 정의와 내용에 대한 공통 이해를 공유하고, 데이터가 어디에 어떻게 저장되어 있는지, 데이터를 어떻게 사용할 것인지 이해할 수 있도록 도와줍니다. 데이터 표준화 과정에서 결정된 지침과 원칙을 제공하며 사용하고 있는 단어, 용어, 도메인 등에 대한 표준 사전을 관리합니다. 표준에 대한 사전 신청 승인 및 결과 조회, 표준항목 조건 검색 및 관련 테이블 컬럼 등에 대한 연관 정보를 조회할 수 있습니다.

나) 필요성

각종 문서, 보고서, 사용자 화면에 표시되는 업무 용어 등을 표준화하여 전사 공유함으로써, 비즈니스 관점에서 데이터의 의미에 대한 이해를 도와줍니다. IT관점에서는 시스템의 정보, 데이터베이스 테이블 정의, 구조, 사이즈, 테이블 간의 관계, 속성 등 기술적 구성요소 간의 상호 참조 및 연결구조의 이해를 도와줍니다. 메타데이터 관리시스템의 가장 중요한 효과는 기업이 자신의 데이터에 대한 “who, what, why, when, where, how”에 대한 해답을 제시하는 것입니다.

유형(Type)	비즈니스 메타데이터 효과 지표 (ROI지표)	기술 메타데이터 효과 지표 (ROI지표)
정량적 효과	<ul style="list-style-type: none"> ▪ 비즈니스 정보 검색시간 단축 ▪ 데이터 품질의 개선 ▪ 직원 교육/훈련 비용의 감소 	<ul style="list-style-type: none"> ▪ IT정보 검색 시간 단축 ▪ 정보시스템 개발 생산성 증대 ▪ 비생산적인 작업 감소 ▪ 재작업 감소 ▪ 데이터 중복 감소 ▪ 중복 프로세스 감소
정성적 효과	<ul style="list-style-type: none"> ▪ 데이터에 대한 일관된 정의 및 이해 촉진 ▪ 데이터에 대한 협업의 신뢰성 증가 ▪ 비즈니스 사용자의 IT전문가들 사이의 간접적인 소통 증가 ▪ 데이터 분석가들이 가치있는 의사결정을 내리도록 지원 ▪ 부정확한 결정을 감소시킴 ▪ 정보 접근의 편리성 증대 ▪ 데이터의 일관성 증가 ▪ IT시스템에 대한 협업의 신뢰성 증대 	<ul style="list-style-type: none"> ▪ 정보시스템 전체 관리기능의 개선 ▪ 정보시스템의 대응속도 향상 ▪ 표준 준수도의 향상 ▪ IT직원교체/직무 순환시 지식전달 용이 ▪ 작업결과에 대한 타 팀에 전달효과 증대 ▪ 정보시스템의 활용현황 파악용이 ▪ 프로젝트 실패 가능성 줄임 ▪ IT포트폴리오 관리 가능 ▪ 데이터 관리 개선으로 데이터 보안 위협감소 ▪ 정보시스템 감사기능 개선

그림 46 메타데이터 관리시스템 효과 지표

다) 장점

- 데이터 설계 도구인 모델링 툴과의 연계 기능을 통해 모델 승인 결재선 및 결재 진행 관리를 통해 주기적으로 모델 정보를 추출하여 표준사전과 모델과의 정합성과 모델과 테이블간 정합성 등을 검증할 수 있도록 관리할 수 있습니다.
- 운영중인 전체 데이터베이스 시스템에 대한 카탈로그 정보 관리 기능과 관리 대상 데이터베이스 추가 시 유연하게 적용할 수 있으며 역할별(모델러, DA, DBA등) 결재 승인 프로세스 관리 기능을 제공합니다.
- 표준모델 승인 기반으로 자동 DDL 생성 및 다운로드 기능 제공함으로써 개발자, 모델러가 쉽고 빠르게 메타데이터를 관리 할 수 있도록 기능을 제공합니다.

라) 활용 사례

고객사	사업명	내용
KRX	해외 거래용 메타관리 시스템	국내 메타데이터와 분리하여 해외 거래용 추가 구축
기획재정부	차세대 디지털예산회계시스템	메타데이터 관리시스템 업그레이드
MG새마을금고	데이터 거버넌스 기반	지티원 원백
관세청	빅데이터 통합 분석을 위한 메타데이터 관리시스템 구축	빅데이터용 메타 관리 체계 구축
인천공항	빅데이터 기반 데이터 거버넌스	Hive DDL을 통한 빅데이터 메타 관리 기능 제공
서울시청	빅데이터 통합 저장소 기반 데이터 거버넌스 구축	267 개 시스템 통합 메타 기반 탐색과 발견 기능 제공
한국철도공사	시스템 증가에 따른 메타스트림 업그레이드	승객, 철도차량, 예약발매 통합 메타 업그레이드
SK그룹	클라우드 환경 메타 기반 지주사 데이터 통합 허브	아마존 웹 서비스 (AWS) 상에 지주사 데이터 통합 메타 관리 포털
삼성카드	실시간 온라인 마케팅시스템 통합 구축	전자 메타 연계한 통합 허브 플랫폼 메타체계 구축

3) 비즈메타관리 도구 (MetaStream For BizData™)

가) 정의

비즈니스 사용자 및 IT 사용자 관점에서 데이터의 접근성, 편의성, 활용성, 보안성 고려하여 최적화된 분류 및 정체를 통해 데이터를 찾기 쉽고, 접근하기 용이하며 내가 원하는 형태로 다양한 분석정보 제공하는 것이 요구됩니다. 비즈메타관리 도구를 통하여 표준계수와 경영지표를 식별 및 정의하고 관리항목과 보고기준 등을 정의하고 표준계수 및 경영지표를 통한 회사 내 일관된 소통이 가능하도록 하고, 어느 보고서를 통해서도 동일한 결과를 도출할 수 있도록 지원합니다.

나) 필요성

보고서는 많은데 어느 조직, 담당자가 어떤 데이터를 보고 있고, 업무에 필요한 데이터가 있는지 확인이 필요하며, 누구나 같은 의미로 이해하고 같은 결과를 도출하기 위해 업무용어에 대한 정의 및 설명을 공유하고 협업이 IT에 편리하게 접근하여 다양한 분석에 활용을 할 수 있도록 비즈니스 메타와 IT 메타를 결합한 통합 메타데이터 서비스 포털 제공이 필요합니다.

다) 장점

데이터 거버넌스 솔루션인 MetaStream For Bizdata™는 “사용자 관점의 통합. 데이터 서비스” 솔루션으로써, 협업 사용자, 관리자 및 데이터 분석가 입장에서의 활용하는 데이터에 대한 설명을 제공합니다. 각종 보고서, 화면, 장표에서 표현되는 업무적 용어, 용어별 산출식, 산출근거, 보고서 작성자 등 사용자 관점에서 필요한 항목을 관리하고 다양한 형태로 시각화할 수 있습니다. 업무 용어 및 분류체계 표준계수, KPI 산출식, 산출근거 및 항목별 SQL북 보고서 분류 검색과 데이터 링크 기능을 제공하여 세부내역을 사용자가 쉽게 확인할 수 있습니다.

라) 활용 사례

고객사	사업명	내용
신한금융투자	클라우드 기반 데이터 레이크 거버넌스 구축	비즈니스 메타데이터 구축 및 전사 포털 연계 서비스
한국투자증권	전자정보 관리체계 혁신	비즈니스 메타데이터 기반 데이터 포털 및 분석 포털 연계 서비스
통계청	マイ크로데이터 거버넌스 구축	표준 메타데이터 기반 통계메타 구축
한국산업기술시험원	데이터 거래소 구축	데이터 거래 메타데이터를 비즈니스 메타데이터 엔진을 활용해서 구현

4) 메타데이터 초(超)상세화 (메타데이터 확장 서비스)

가) 정의

디지털 전환 시대의 가장 큰 장애물은 “메타데이터화 하기 힘든 데이터를 어떻게 다룰 것인가?” 하는 것입니다. 즉, 이미지나 오디오 정보는 디지털화 되어 있으므로 필요한 정보를 변형하거나 연관성을 찾아내어 학습시킬 수는 있지만 메타데이터가 없으면 이를 효과적으로 저장하고 검색하기가 힘들어 집니다. 비단 멀티미디어 데이터 뿐만 아니라 기업의 업무 데이터도 “메타데이터 초(超)상세화”를 통하여 디지털 전환 시대에 대응을 할 수 있습니다. 즉, 메타데이터 초(超)상세화 구현을 통하여 기업에서 생성, 변형, 유통되는 다양한 데이터를 메타데이터 기반으로 검색, 통계, 분석, 예측, 학습 등 다양한 데이터 활용 요구사항을 지원합니다.

나) 필요성

세계 최고의 ERP기업인 SAP의 전통적인 ECC/R3나 마스터데이터관리솔루션인 MDG의 경우 자재 및 협력사 등 업무데이터에 대한 메타데이터 상세화가 기본적으로 제공됩니다. 이를 통하여 강력한 데이터 거버넌스 정책을 적용하여 표준화된 양질의 디지털 데이터를 확보함으로써 업무의 정확성과 효율성을 대폭 끌어

올림으로써 세계 최고의 제품으로 성장하였습니다. 빅데이터 시대에 접어 든 오늘날 전통적인 시장에서 성공적인 성과를 보인 데이터 솔루션 벤더 중에 메타데이터 상세화를 주장하는 경쟁자들은 국내에는 전무하고 글로벌하게도 SAP를 제외하고는 별로 눈에 띄이지 않습니다. 하지만 디지털 전환 트랜드와 마스터 데이터 관리를 계기로 전사적 데이터에 대한 메타데이터 상세화가 중요한 이슈로 제시되고 있습니다.

당사의 메타데이터 초(超)상세화 제품은 앞서 기술한 기술(Technical) 메타 데이터 이외에 사용자가 활용하는 보고서, 화면, 자료 등을 표현하는 비즈니스 메타 데이터, 이미지 동영상을 포함한 음악, 문화재, 영화 등 멀티미디어 자료를 표현하는 컨텐츠 메타 데이터, 각종조사, 설문에 근거하여 작성되는 통계자료의 통계 메타 데이터, 신문 등 텍스트 형태소를 분석 하여 정의하는 텍스트 메타 데이터, 데이터 셋을 기반으로 데이터 거래 상품을 구성하는 상품 메타 데이터는 물론 업무 절차에 직접적으로 활용되는 마스터 데이터(기준정보)를 중심으로 기업이 보유하고 있는 모든 유형의 데이터를 메타 데이터로 초(超) 상세화하여 디지털화 한 다음 DBMS 및 빅데이터 저장소에 저장하고 가상화된 하나의 질의어(Query)로 추출, 가공, 분석할 수 있는 데이터 가상화 환경 기반을 제공합니다.

5) 데이터 품질관리 도구 (QualityStream™)

가) 정의

정확한 데이터 분석을 위해서는 데이터의 표준화와 품질 확보가 중요합니다. 데이터의 많은 기업들이 품질향상을 목적으로 전사 데이터관리 체계 기반 하에 데이터 표준화 관리체계, 메타데이터 통합 저장소, 메타데이터 관리 시스템 등의 데이터 거버넌스 기반을 구축하고 있습니다. 신규 시스템 구축을 수행할 때에는 구축 수행사가 데이터관리의 정합성 유지를 위해 조직과 프로세스를 유지관리 하지만 프로젝트 종료 시점에서 인수인계를 기능적인 측면에서만 진행하다 보니 프로젝트기간 중에 도출된 비즈니스적 요건과 핵심관리 요소가 전반적으로 인수인계 되지 않는 문제점 발생합니다.

나) 필요성

데이터조직의 권한과 책임이 유연해 지면서 강력한 데이터의 신규/변경에 따른 의사결정의 역할을 수행하지 못하게 되고 힘있는 부서나 담당자에게 밀려 표준에서 벗어난 것을 승인하게 되어 잘못된 정보를 제공하게 되는 원인이 발생됨에 따라 데이터 확보의 효과성이나 활용 결과의 신뢰성을 잊게 됩니다. 당사의 강력한 데이터 표준관리 솔루션인 MetaStream™은 데이터 관리 정책, 프로세스, 표준화, 조직 구성을 아우르는 DA Framework으로써 데이터 품질 확보의 기반 솔루션으로 활용됩니다. 하지만 전사 차원의 데이터 관리 정책은 궁극적으로 데이터의 일관성 유지와 신뢰성을 확보하는 것을 비전과 목표로 함으로 MetaStream™이 추구하는 표준 원칙과 데이터 관리 정책 및 데이터 거버넌스의 활동을 효과적으로 달성하기 위해서는 표준을 기준으로 정합성 관리 중심의 데이터 품질 관리 프로세스를 체계화 하여야 합니다.

다) 장점

메타데이터 수집 및 매핑 관리 및 품질 검증대상 관리기능 제공으로 운영중인 데이터에 대하여 대상데이터에 직접 접근 또는 일정 저장소에 추출한 후 품질 규칙에 기반 후 주기적 측정을 실행합니다. 규칙에 맞지 않는 오류데이터가 발견되면 그 리스트를 확인한 후, 원인을 분석하여 다시 재발하지 않도록 원천 시스템의 개선을 진행할 수 있도록 지원합니다. 원천 데이터 중 반복적인 품질관리가 필요한 대상 데이터를 직접 또는 특정 영역으로 해당 데이터를 추출한 후 정해진 품질 측정식과 측정 스케줄에 의해 주기적인 검사 프로세스를 수행하고 그 분석 결과를 제공합니다. 대용량의 데이터는 자사의 강력한 데이터 처리 도구인 TeraStream™을 활용하여 빠른 품질 측정을 지원합니다. 자사 솔루션의 장점 중 하나인 파일 기반 검증체계를 추가적으로 제공함으로써 정형데이터 뿐만 아니라 반정형, 비정형 데이터에 대한 메타데이터 정보 연계 및 메타데이터 기반의 데이터 품질관리 정보의 통합 레파지토리 기반으로 측정정보, 오류정보, 기준정보에 관한 다양한 통계를 제공하며 관련 정보를 보고서 파일로 출력 기능을

제공합니다.

라) 활용 사례

고객사	사업명	내용
한국기업데이터	데이터 현황관리 시스템	데이터 품질 진단 및 개선 관리
한국산업기술시험원	NIA 빅데이터 플랫폼 구축	수집 데이터 품질 검증 및 데이터 셋 진단체계 구축
경복대학교	학생성공 데이터 플랫폼 구축	데이터 품질 검증 및 정제
대통령기록관	차세대 대통령 기록관리 시스템	데이터 품질 검증 및 정제
신한오렌지 생명	IT통합 차세대 구축	신한생명과 오렌지생명 통합 품질 검증 체계 구축
서울시청	빅데이터 통합 저장소 기반 데이터 거버넌스	정형데이터 및 파일 검증 및 진단 개선 관리

6) 데이터 흐름관리 도구 (Q-Track™)

가) 정의

데이터 소스 및 이동 경로를 포함하는 데이터 흐름 정보를 시각적으로 제공하기 때문에 데이터의 생성부터 폐기에 이르는 생명주기를 추적하고 관리할 수 있습니다. 데이터 흐름을 분석하고 시각화하여 기업 내 데이터 흐름을 가시적으로 관리할 수 있습니다.

나) 필요성

업무 분석가들은 30% 이상의 시간을 데이터 작업에 사용하고 있습니다. 데이터의 확인, 통합, 수정 등에 많은 노력과 비용이 소모되고 있으며 오류 데이터는 직간접적으로 비용 이슈와 직결됩니다. 데이터는 운영에서 분석 데이터까지 여러 단계 (ODS → 데이터웨어하우스 → 데이터마트)로 이어지는 데이터 흐름을 가지고 있으나, 데이터 추출, 가공, 로드 기술이 다양하게 존재하여 이 흐름을 가시적으로 관리하기 어려운 것이 현실입니다.

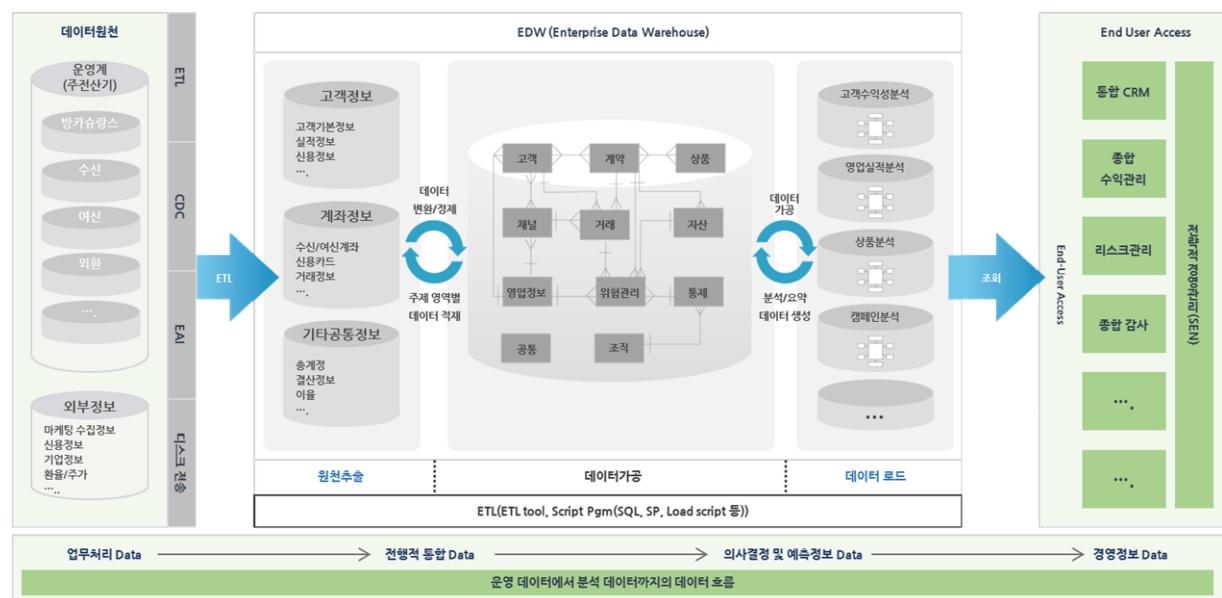


그림 47 데이터 관리 환경의 복잡성

데이터 흐름관리 도구는 데이터를 효과적으로 관리할 수 있도록 도와줍니다. 데이터 라이프사이클 관리와 전사데이터에 대한 현황파악 및 최적화된 데이터 관리를 지원합니다. 전사 데이터에 대한 현황 파악 및 최적화 된 데이터 관리와 정확성, 안정성을 보장하며, 데이터 신뢰도는 지속적으로 향상됩니다.

다) 장점

- 데이터 투명성 확보 : 데이터의 생성부터 폐기까지 전체 생명주기 및 이와 관련한 모든 프로세스를 추적하여 데이터의 투명성 및 이해도를 높여줍니다.
- 영향도 분석 : 테이블/컬럼 레벨의 데이터 흐름을 시각화해 프로그램 또는 데이터의 오류 사항이 어떤 하위 시스템 및 분석 내용에 영향을 줄 수 있는지 쉽게 파악할 수 있도록 지원합니다. 데이터의 변경에 따른 연관 업무 및 담당자 정보제공 데이터의 생성 및 가공 과정에 대한 정확한 정보를 통하여 담당자 간의 명확한 의사소통 기반을 제공합니다. 드러나지 않은 데이터의 문제점 도출이 가능하며, 데이터의 흐름 분석을 통한 비합리적 연결구조 파악, 중복된 데이터 생성 프로세스 파악, 특히 데이터 가공 장애 시 영향받는 작업 및 테이블 정보 파악에 유용합니다.
- 규정 준수 : 데이터의 전체 흐름을 기록하여 데이터 무결성 및 투명성을 보장합니다. 이를 통해 사용자는 데이터의 흐름을 정확하게 확인하고 사업 내용이 법규 및 표준 범위를 벗어나지 않도록 보장할 수 있습니다.
- 데이터 오류 교정 : 회사 전반에 걸쳐 프로그램이 데이터에 어떻게 접근 및 변환, 공유되는지 파악할 수 있으므로 오류 소스로 인하여 어떤 보고 내용이 영향을 받았는지 확인하여 교정할 수 있습니다.

라) 활용 사례

고객사	사업명	내용
KB국민카드	전자 데이터거버넌스 구축	단말 계정계 정보계 백데이터 리포트까지 이어지는 데이터 흐름을 분석하여 기시화 서비스 제공
KB국민은행	차세대 거버넌스 체계 수립 및 구축	마이데이터 품질관리 비즈니스 마이데이터 및 데이터 흐름을 포함한 전자 데이터 거버넌스 포털 구축
대구은행	대구은행 정보계 고도화	계정계에서 정보계로 이어지는 전자 데이터 흐름 분석

7) 마스터 데이터 관리 도구 (MasterStream™)

가) 정의

마스터 데이터(기준정보)는 전사 업무부문 간 공통의 의미를 가져야 할 정보로, 업무 수행 시 참조되는 데이터를 의미합니다. 전사 관점의 데이터의 기준을 통합적으로 관리, 데이터 처리 과정에서 발생할 수 있는 오류를 바로잡고 데이터의 품질을 확보 전사 관점의 단일버전으로 유지해야 하는 대상이며 모든 데이터의 뿌리가 되는 데이터의 근간입니다.

나) 필요성

마스터 데이터가 중복이나 누락 등의 이유로 부정확할 경우, 마스터 데이터를 참조해 생성되는 트랜잭션 데이터부터 이를 가공해 생성되는 분석 데이터까지 신뢰성과 품질에 치명적인 문제가 발생될 수 있습니다. 이는 기업의 운영에도 심각한 문제를 초래할 수 있습니다. 기업 경영의 근간이 되고 있는 마스터 데이터의 품질수준 향상과 데이터 오류를 원천적으로 제거하기 위해서는 표준화된 기준정보 통합 모델을 기반으로 일관된 프로세스를 정의하고, 실시간 데이터 분석 및 업무간 연계를 실현하고, 투명한 정보의 흐름 체계를 완성하여 신뢰할 만한 리포팅으로 실시간 경영정보 기반을 확보하여야 합니다. MasterStream™은 마스터 데이터 관리 제품으로 서로 다른 기준과 표준을 가진 시스템으로부터 통합된 표준을 만들어내고 (Data Consolidation) 이를 중심으로 일관된 값의 데이터를 관련 시스템에 공급함으로써 정확한 데이터 공유가 이루어 질 수 있습니다.

이를 위해서는 전사적으로 공유되고 기업의 중요한 정보를 담고 있는 기준 데이터에 대해서는 데이터 생애주기 관리에 기반한 검증·동시배포의 체계가 정비되어야 하며, 최소한 레거시 시스템에서 생성된 데이터

터를 통합하여 각 응용시스템에서 참조되기 전에 업무규칙에 의한 검증과정을 거쳐 동일한 시점에 동일한 활용이 가능하도록 자동화된 통제 및 모니터링 솔루션을 필요로 합니다. 또한 마스터데이터는 데이터 분석의 가치가 재조명 받는 빅데이터' 시대에 마스터데이터 관리 중요성이 다시금 대두되고 있습니다.

다) 장점

마스터데이터관리 솔루션인 MasterStream™은 전사 데이터 거버넌스 기반으로 구성되어 있습니다. 즉, 메타 표준에 의한 모델링, 마스터데이터 분류체계와 분류별 속성과 오너십 정의가 연결되어 담당자별 권한에 의하여 데이터 생성, 변경 및 승인절차가 이루어질 수 있도록 워크플로우 기능이 내재되어 있습니다. 멀티 도메인을 지원하여 마스터데이터가 추가 되더라도 적은 비용으로 손쉽게 추가할 수 있습니다. 또한 오류가 발견되면 언제든지 즉시 해당 단계에서 데이터 흐름을 통제할 수 있고, 또 다른 오류 발생을 제어할 수 있는 체계를 갖추고 있습니다. ETL 솔루션과 데이터 품질관리 도구 그리고 검증 규칙관리 시스템과 연동하여 품질검증 프로세스를 적용하는 검증 체계를 지원하고, 검증규칙에 대한 관리 및 검증결과에 대한 모니터링을 수행하며, 데이터 추출시점 데이터 표준체계에 정의된 검증규칙을 활용, 필요한 테이블, 컬럼에 적용하여 사전 검증 작업이 실행할 수 있습니다.

라) 활용 사례

고객사	사업명	내용
행정안전부	재난관리 자원통합 구축	자원, 조직, 협력사 및 인적자원 마스터데이터 관리 시스템 구축
고려아연	설비마스터 통합 MDM 구축	설비마스터 표준화 및 마스터데이터 관리 시스템 구축
한솔제지	전사 데이터 거버넌스 기반 MDM 구축	고객/벤더/자재/ 등 운영마스터 데이터 외 40 종 관리
한국도로공사	마스터데이터 공동활용체계 구축	마스터데이터 관리 체계 수립 및 노선 마스터데이터 관리 시스템 구축
현대글로비스	기준정보관리체계 구축	고객/벤더/조직/전사코드 등 9 종 마스터데이터 구축

◎ 국내외 기술동향

1. 전통적인 Data Warehouse기술에서 빅데이터 분석 환경으로 변화

가. 분석 대상 데이터의 확대

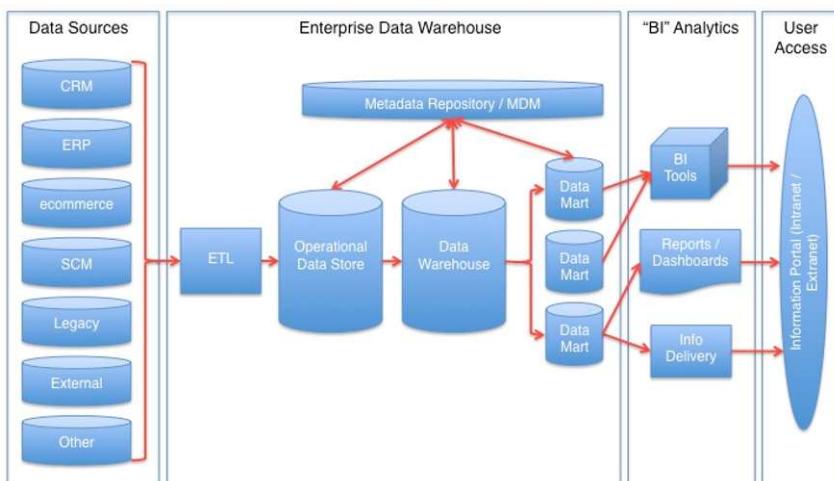


그림 48 EDW 아키텍처 개념도

전통적인 Data Warehouse는 RDBMS로 운영되고 있는 운영 시스템으로부터 데이터 분석에 적합한 형태의 분석 시스템을 구축하고 이를 기반으로 데이터에 대한 분석과 인사이트를 제공하는 시스템입니다. 분석 대상 데이터는 기존의 운영 시스템이 생성하는 정형 데이터 중심이고 이런 기업의 운영 과정에서 생성되는 데이터의 분석을 통한 비즈니스 분석을 중심으로 운영되었습니다.

하지만 2010년을 전후한 전 세계적 스마트폰의 보급은 개인 장비에서 생성하는 이미지, 텍스트, 비디오 데이터가 폭발적으로 증가시켰고 고객과 잠재적 대상 고객인 각 개인의 성향을 기반으로 한 비즈니스 의사결정이 중요성이 대두되면서 개인이 생성하는 데이터에 대한 분석 요구가 급격히 증가하게 되었습니다.

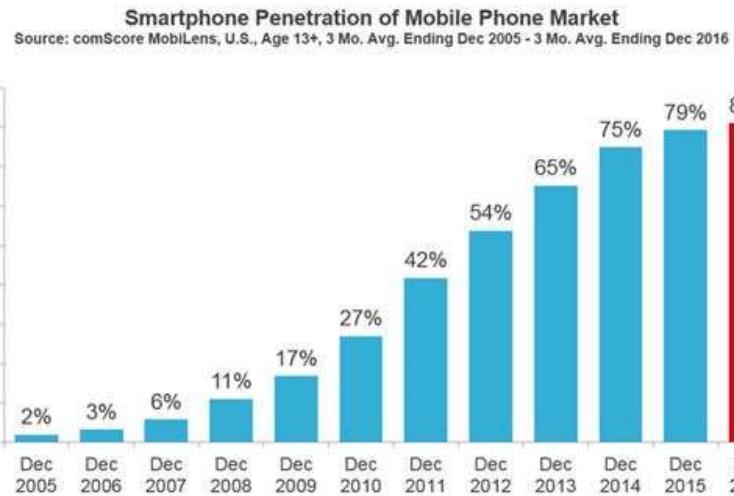


그림 49 미국 스마트폰 보급률

그리고 2010년을 전후한 IoT 시장의 급격한 성장은 모든 사물이 데이터를 생성할 수 있는 기반을 만들게 되었습니다. 스마트 팩토리, 스마트 빌딩, 스마트 카 등 IoT 기반의 새로운 비즈니스를 만들어내게 되면서 다양한 센서가 지속적으로 대량의 데이터를 만들어 내는 기반이 되었습니다.

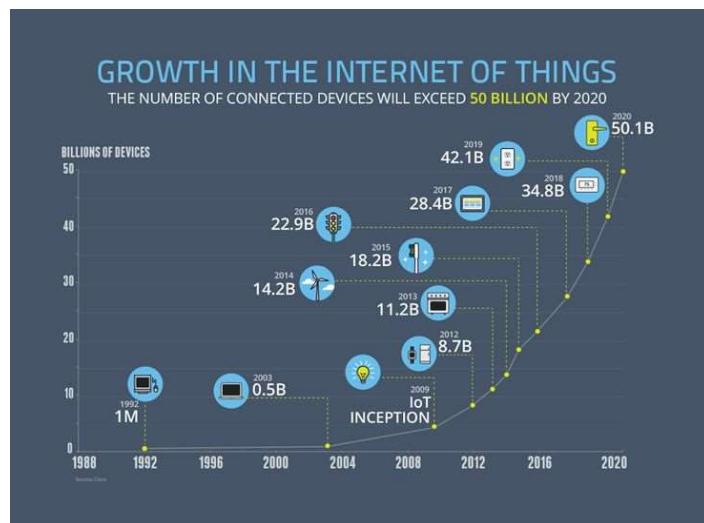


그림 50 글로벌 IoT 시장의 성장 (출처 : KOTRA, 2017)

나. 데이터 분산처리 기술의 등장과 성장

2006년 분산파일시스템인 HDFS, 분산처리프레임워크인 MapReduce를 핵심으로 공개소프트웨어로 등장한 Hadoop은 그 효율성을 인정받으며 급격한 성장을 하였습니다. 이 Hadoop을 기반으로 한 다양한 공개 소프트웨어가 폭발적으로 등장하면서 짧은 시간에 Hadoop 생태계를 형성하였습니다.



그림 51 빅데이터 소프트웨어 생태계

이러한 흐름에서 급격히 증가한 데이터로 고민하던 많은 기업들의 참여로 하나의 기반 기술로 빅데이터 기술 생태계를 형성하였습니다.

다. EDW와 빅데이터 분석

빅데이터 기술의 등장으로 시장의 변화를 예상한 동사는 기존의 EDW (Enterprise Data Warehouse)와 빅데이터기술이 공존할 것으로 예상하여 EDW와 빅데이터가 융합된 Hybrid 분석 환경을 준비하였고 이를 기반으로 SuperDW Architecture라 명명하고 각 제품을 개발하였습니다.

빅데이터 기술 기반 환경에서도 데이터 통합은 매우 큰 비중을 차지하고 있기 때문에 기존의 ETL을 Hadoop 기반으로 전환하면서도 기존의 EDW를 지원할 수 있는 형태로 진화 발전 시켜 TeraStream for Hadoop 제품을 출시하였습니다.

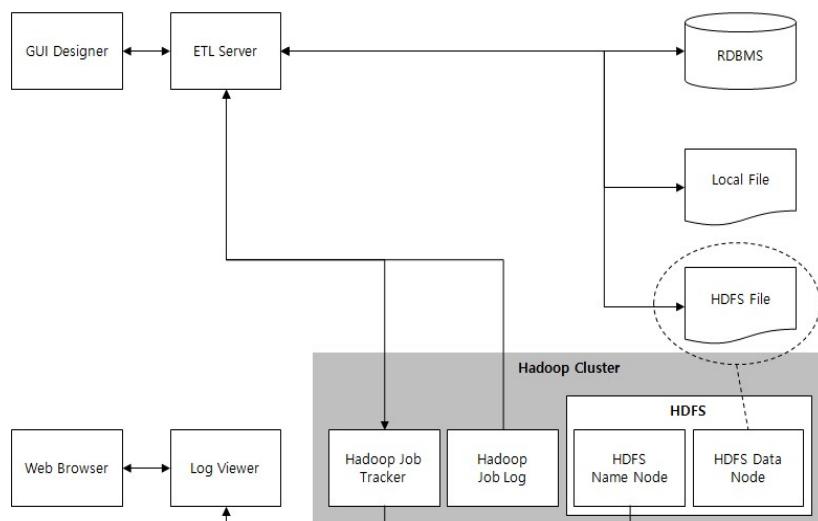


그림 52 RDBMS, HDFS를 동시 지원하는ETL,TeraStream for Hadoop™

이를 기반으로 SuperDW Architecture 기반 빅데이터 플랫폼으로 발전시켜 다양한 사업을 성공적으로 수행하였습니다.

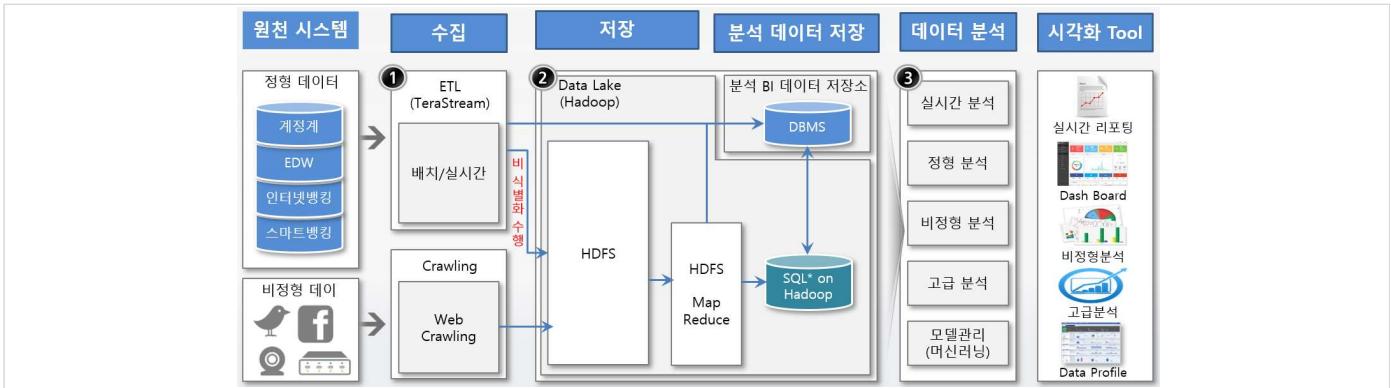


그림 53 TeraStream for Hadoop 기반의 SuperDW Architecture 예시

라. 빅데이터 플랫폼으로 확장

당사는 데이터 통합 중심의 빅데이터 플랫폼에서 빅데이터 생태계 전체에 대한 활용성을 확보하기 위해 초고속 실시간 데이터 처리 기술(TeraStream BASS™), 데이터거버넌스 기술을 통합하고 빅데이터 통합 관리 기술을 융합해 빅데이터 플랫폼으로 발전시켜 TeraONE™을 출시하였습니다.

2. 클라우드 환경으로 변화

가. 빅데이터 분석에 대한 인식 변화

빅데이터 전문 인력을 중심으로 운영되던 빅데이터 분석의 흐름이 도구의 사용에 대한 지식만으로도 가능하도록 기술의 뒷받침이 필요하다는 "Citizen Data Scientist" 개념의 등장과 공공 데이터 허브의 급격한 성장으로 복잡한 빅데이터 분석 인프라에 대한 지식 없이 단순한 도구로 활용할 수 있도록 자동화된 빅데이터 플랫폼이 요구되고 있습니다.

이러한 인식 변화는 누구나 쉽게 데이터 수집, 저장, 분석 환경을 구성하고 일상의 업무에서 빅데이터 기술을 활용한 데이터 분석을 하고자 하는 요구가 증가하고 있습니다.



그림 54 빅데이터 분석에 대한 인식 변화로 예상되는 미래

나. 빅데이터 인프라 기술의 변화

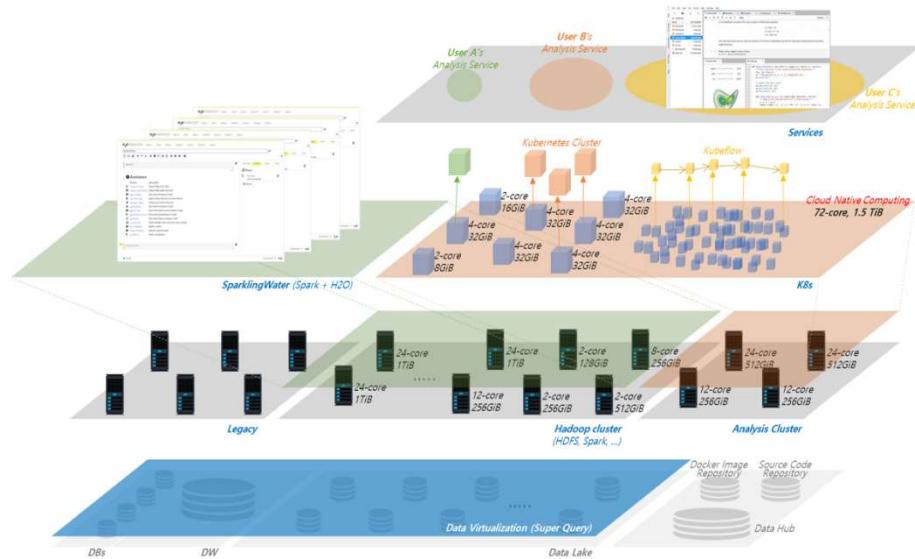


그림 55 Cloud Native Computing 기술에 의한 빅데이터 인프라 기술 변화

리눅스의 컨테이너 가상화 기술을 활용한 “Cloud Native Computing” 기술의 등장으로 수백개 노드의 복잡한 빅데이터 인프라에 대한 급격한 변화가 진행되고 있습니다. 물리적 하드웨어는 단지 컴퓨팅 파워를 제공하는 요소이고 이를 활용하는 것은 Cloud Native Computing을 통해 가상화하여 쉽고 빠르게 변화에 대응하는 방향으로 변화하고 있습니다.

“Cloud Native Computing” 기술에 의한 빅데이터 인프라 기술 변화는 물리적인 인프라 구축과 별개로 사용자의 필요에 따라 쉽게 빅데이터 분석 환경을 생성하고 사용자의 목적에 따라 다양한 형태로 자동 구성되는 기술의 개발을 촉진하게 될 것입니다.

다. 클라우드 기반의 빅데이터 플랫폼

위의 기술변화에 대응하기 위해서는 Private Cloud, Public Cloud, Hybrid Cloud(Private, Public Cloud 환경의 공존) 환경에서 Hardware 환경에 구속되지 않는 필요와 목적에 따라 동적으로 구성되는 빅데이터 플랫폼의 개발이 필연적으로 이어질 것으로 예상되고 있습니다.

클라우드 기반의 빅데이터 플랫폼의 핵심 기술은 다음과 같을 것으로 분석하고 있습니다.

- 1) 구성에 대한 요구사항을 입력하고 수분 내에 자동적으로 구성하는 기술
- 2) 데이터 탐색을 지원하는 데이터 허브 탐색, 데이터 큐레이션을 통한 인사이트 도출 지원 기술
- 3) 접근 가능한 데이터에 대한 데이터 맵 생성 기술
- 4) 데이터의 특성을 인지하여 수집 방법, 저장 방법, 분석 방법에 대한 추천 기술
- 5) 데이터 맵에 기반한 분석 자동화 기술

3. Data Fabric 환경으로 변화

2010년부터 전 세계적으로 하둡 기반의 빅데이터 플랫폼인 데이터레이크 구축 사업에 많은 투자가 수반되었습니다. 데이터웨어하우스의 데이터까지도 데이터레이크로 물리적 통합하는 시도가 이루어졌습니다. 하지만 데이터를 무한정 하나의 저장소로 통합하는 것도 한계가 있었으며, 이러한 데이터레이크에 축적한 데이터는 용처가 불확실한 데이터 자산일 가능성이 높아졌습니다(센서, 웹로그 데이터 등 기존에 관리하지 않던 데이터 관리 기준 부재가 문제). 즉 빅데이터의 잠재적 가치와 데이터 영구 보관에 따른 낭비 사이의 균형 점 도출이 필요하게 되어 데이터 거버넌스의 중요성이 재조명되었습니다.

글로벌 리서치 회사인 포레스터와 가트너에서는 2018년부터 데이터를 물리적으로 데이터레이크에 통합하지 않고도 다양한

이기종 플랫폼 (데이터웨어하우스, 빅데이터, 클라우드 등)의 다양한 데이터소스 (DB2, Oracle, Tibero, SybaseIQ, Hive, PostgreSQL 등)를 메타데이터를 참조하여 데이터 가상화 레이어에서 하나의 쿼리로 데이터를 분석할 수 있는 개념인 데이터 패브릭을 발표하였습니다.

“..lake of clarity of data in data lakes lead to data chaos. It becomes more difficult to ensure data quality and governance.. ..bad data costs companies on average one third of their revenue.

FORRESTER®

“..data lakes lacks semantic consistency and data governance.
This makes data analysis highly individualized..

Gartner

데이터 패브릭의 3가지 핵심 구성요소는 데이터 레이크, 데이터 거버넌스, 데이터 가상화입니다. 동사는 이 핵심 구성요소를 이루고 있는 모든 기술들을 자체적으로 개발하여 확보하고 있었고, 20여년간 데이터 통합 및 관리 소프트웨어를 개발하여 사업을 진행하면서 이러한 미래 트렌드를 예측하고 준비를 해왔습니다. 그 결과 시장의 동향 분석과 고객의 요구 사항을 기반으로 2019년 데이터 패브릭 전략을 완성하는 데이터 가상화 솔루션을 출시하여 근로복지공단을 시작으로 농협생명, 지역정보개발원에 납품하였으며 2021년 하반기에도 많은 대기 수요가 있는 상황입니다. 이는 데이터 관련 보유하고 있는 모든 요소 기술을 융합하여 신속하게 시장의 요구를 수용할 수 있는 동사의 역량에 기인합니다. 2021년 가트너에서 발표한 확장된 데이터 패브릭 전략에는 다음과 같은 비즈니스 메타 데이터 기반의 데이터 카탈로그, 데이터 가상화 등이 포함되고 있으며, 이는 동사의 IRUDA™, TeraONESuperQuery™의 기능에서 제공됩니다.



그림 55 전통적인 데이터 플랫폼에서 데이터 패브릭

또한, 빅데이터 플랫폼 또는 데이터 거래소 등에서 사용자들이 자주 활용하는 데이터 셋에 대한 정보를 학습시켜 넷플릭스나 유튜브와 유사하게 AI를 활용한 추천서비스를 제공하고 있습니다. 2019년 산림 분야 빅데이터 플랫폼 및 거래소에 시범 적용 이후 학습 데이터 모델링 개선 및 알고리즘 고도화를 통해 2021년 4월 특허 등록 결정을 받았습니다.

(2021.04 특허등록 완료)

“인공지능 추천 모델을 사용하여 추천 정보를 제공하는 데이터 카탈로그 제공 방법 및 시스템”

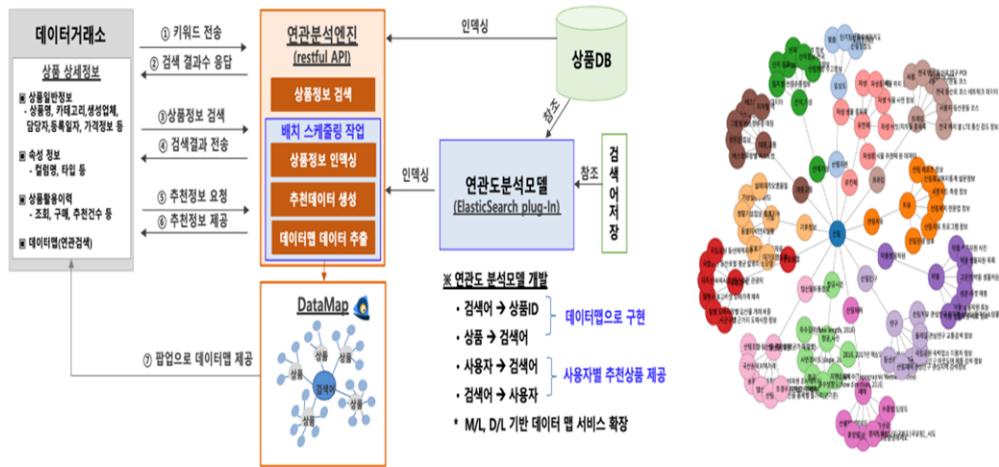


그림 56 AI 추천 정보를 사용하는 데이터 카탈로그 제공 방법 및 그 시스템

2020년부터 데이터 사이언티스트들이 필요한 정보를 쉽게 찾아 분석에 활용할 수 있도록 지원하는 데이터 카탈로그 사업이 확대되고 있습니다. 동사는 산림/교통/헬스케어/디지털기술혁신 분야에 구축한 경험을 기반으로 제품을 고도화하여 특히 등록을 결정 받게 되었으며, 관련 시장을 선도하고 있습니다.