MapR Converged Data Platform

빅데이터 시장동향

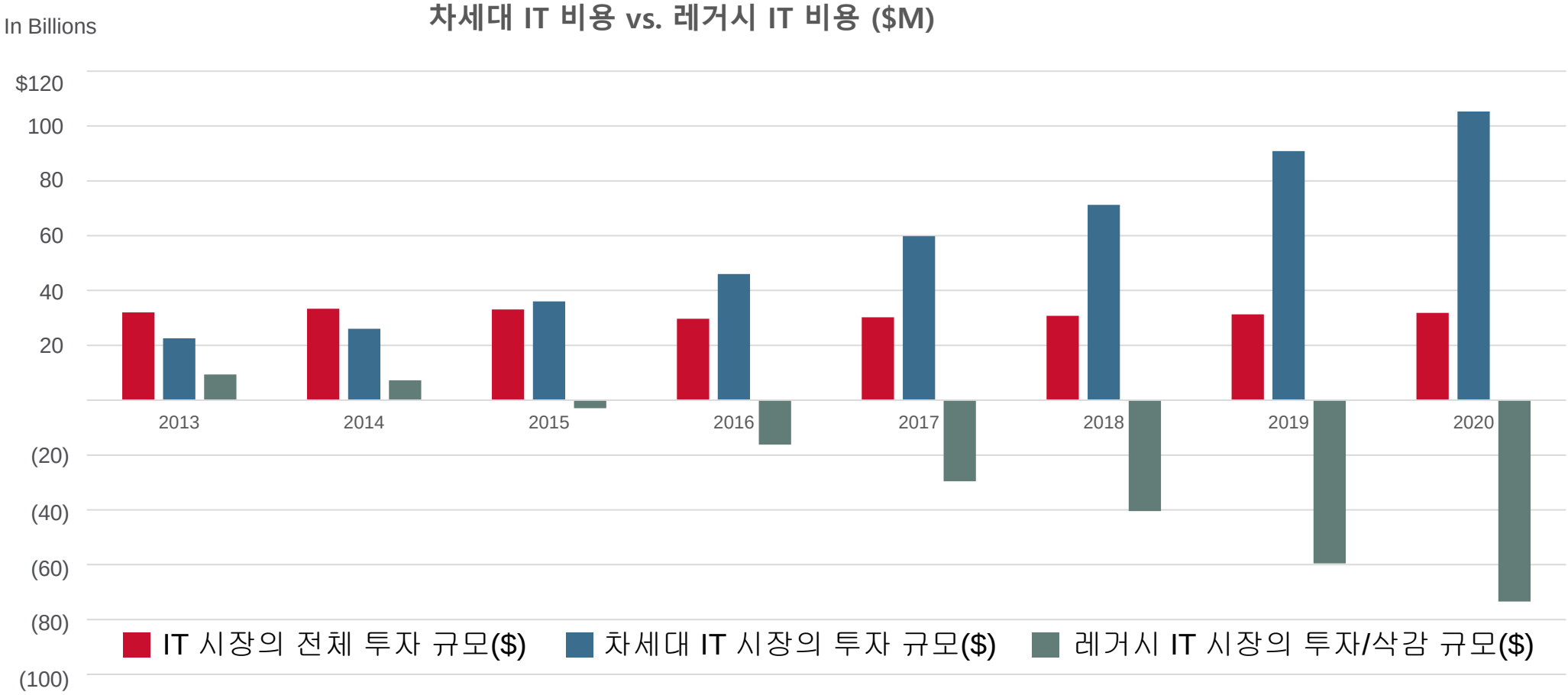
DIFFERENT
THAN
"THE OTHERS"

기업 IT 환경의 변화

"1980년대 모든 데이터를 플랫 파일로 관리하던 어려움을 극복하고자 데이터베이스 시스템이 시장에 출시 된 이후로 기업용 어플리케이션 등장, 인터넷의 등장, 디지털 변혁 접목 등 기업 혁신의 핵심에는 **항상 데이터가 중요한 역할**을 함"

기업 IT 환경의 변화와 데이터

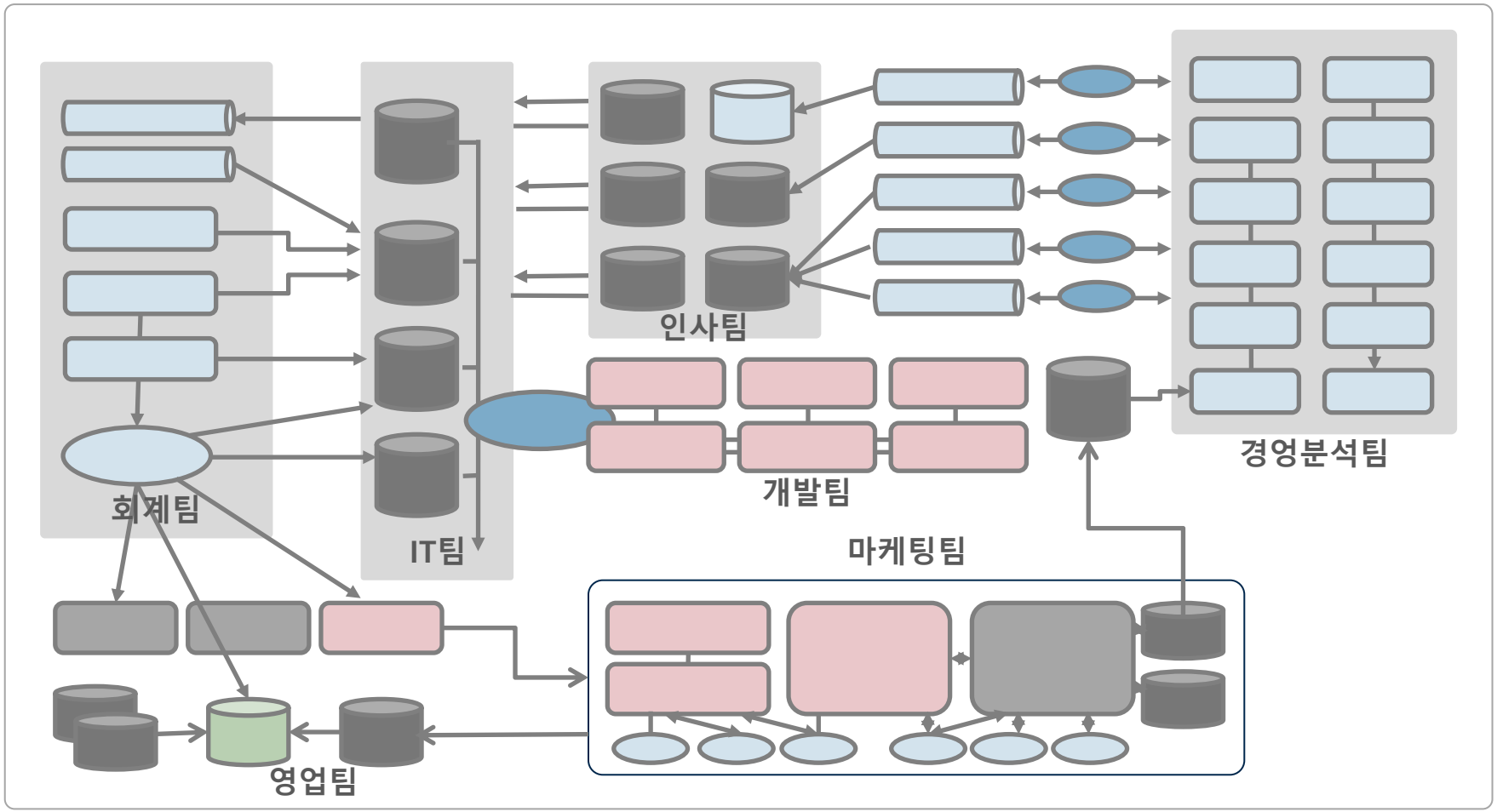
"가트너와 같은 시장 조사 기관들은 향후 4년 이내에 **기업내 데이터의 90%가 Hadoop과 같은 차세대 IT 환경에 저장**될 것으로 예측."



Source: IDC, Gartner; Analysis & Estimates: MapR
Next-gen consists of cloud, big data, software and hardware related expenses

기존 IT 환경의 특징

"기존 IT 환경은 전통적인 데이터 접근법에 의해 단위부서 별로 각기 다른 어플리케이션이 각기 다른 데이터 소스를 활용하는 구조"



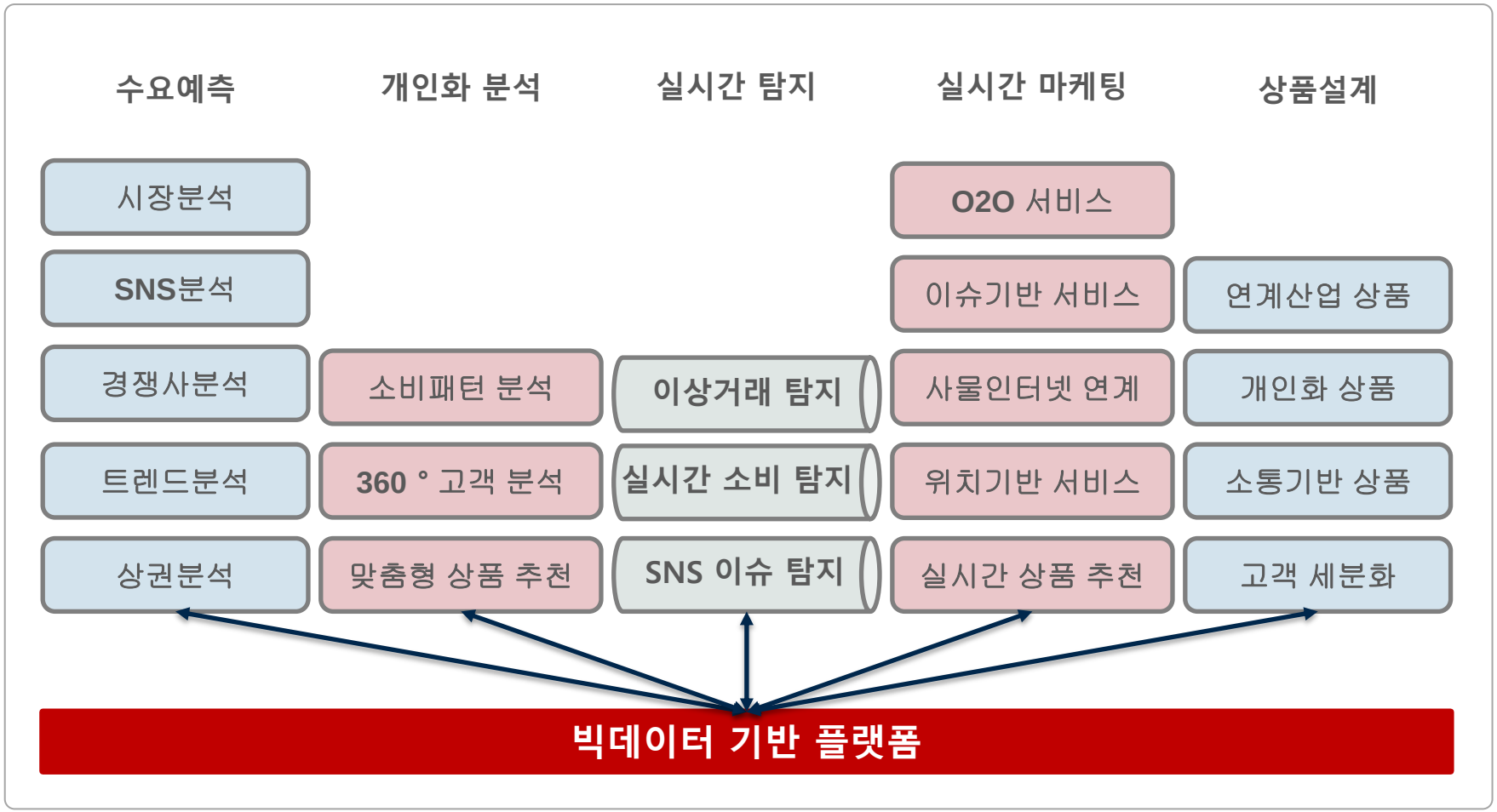
< 기존 IT 아키텍처 >

기존 환경의 특징

- 단위 사업별로 구축
- 사일로 형태의 구성
- 이기종 시스템들간의 복잡한 인터페이스 관계
- 정형 데이터 위주의 처리
- IT 에서 데이터를 발생 및 처리
- Oracle, SAP, Microsoft 등의 전통적 어플리케이션

차세대 IT 환경의 특징

"차세대 IT 환경은 고객니즈 또는 경영지표를 보다 심도있게 분석하기 위한 데이터 기반의 통합된 플랫폼 형태의 구조"



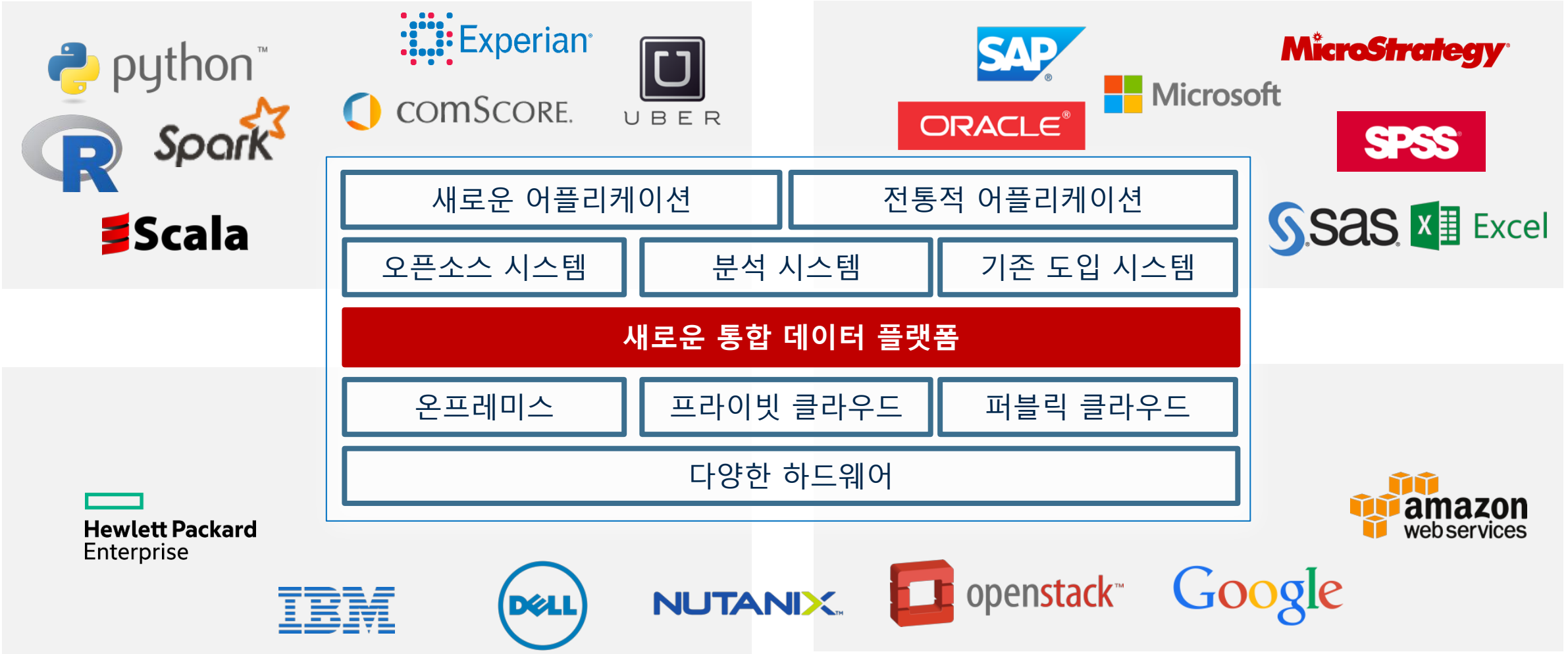
< 차세대 IT 아키텍처 >

차세대 환경의 특징

- 통합 사업으로 구축
- 이기종 시스템들을 데이터 기반으로 통합하는 형태
- 정형/반정형/비정형 데이터를 모두 통합
- IT/OT 모든 영역에서 데이터를 발생 및 처리
- 개인화 분석, 실시간 오퍼등의 새로운 패러다임

통합 데이터 플랫폼의 필요성

"새로운 차세대 IT 아키텍처는 전통적인 시스템과 새로운 시스템 모두와 호환이 가능 하면서 전사 통합이 가능한 아키텍처."



빅데이터 시대의 산업 방향

"McKinsey등의 시장조사 기관들은 금융업은 빅데이터 활용의 잠재 가치가 높은 산업으로, 기존 데이터 보유량이 많고 새로 유입되는 데이터 양이 많기 때문에 VoC 분석, 360° 고객 싱글뷰 구성 등 **활용가치가 매우 높은 산업으로 분류** 하고 있습니다."


접근이력 감사

비정형 데이터 통합

360° 고객 싱글뷰


융합 산업


개인화 상품 설계


DW 확장


VoC 분석

이상거래 탐지


사물인터넷 연계


AI 자동상담

기업들이 MapR 플랫폼을 선택하는 이유

"MapR 플랫폼은 단순한 빅데이터 플랫폼으로서의 기능이 아닌 엔터프라이즈 수준의 가용성과 안전성을 제공하는 기업용 플랫폼"

핵심 기술 요소

- 분산 파일 시스템 – MapR-FS
- NoSQL DB 시스템 – MapR-DB
- 글로벌 메세징 시스템 – MapR-Streams

고속처리를 위한 최적화

- 대용량 분산 병렬 처리 지원
- 정형 및 비정형등의 다양한 데이터를 통한 대규모 분석 및 기계 학습 구현

규모에 따른 확장

- **수 조 단위까지 확장되는 제한없는 파일 시스템**
- **대규모 분산 처리 환경에서 대용량 데이터 고속 처리**

엔터프라이즈 수준의 안정성

- **무중단 데이터 액세스를 지원하는 다양한 자동 복구 기능**
- 재해 복구와 스냅샷을 포함한 고가용성 제공
- **안정적인 고객 지원 체계**

MapR 플랫폼을 통한 비즈니스 목표

"디지털 변혁에 대한 혁신 성과와 비용 절감에 대한 경영 성과를 모두 달성할 수 있는 혁신 플랫폼으로서 MapR을 선택"

비즈니스 혁신

AMEX 카드사는 MapR 플랫폼 기반의 이상금융거래 탐지 시스템으로 매년 1조 달러의 거래에서 사기를 보호

TransUnion 사는 MapR 플랫폼에서 새로운 셀프서비스 분석 플랫폼을 통해 고객에게 더 나은 시장 통찰력을 제공하여 의사결정에 도움

AADHAAR 사는 인도의 12억5천만명의 생체인식 시스템을 개발하여 불법적으로 보험금을 수령하는 사기탐지 시스템을 개발해 1.3B$를 절감

비용 절감

Experian사는 기존 DB2 환경의 신용평가 스코어링 시스템을 MapR 플랫폼에 이식하여 20배의 비용 절감

UnitedHealth Group은 건강 보험료 지불과 클레임 관리 등 80여가지 유스케이스를 MapR 기반으로 도입하여 지불오류, 사기 등 월 2백만 달러의 비용 절감

IRi사는 기존 운영중이던 메인프레임에서 MapR 플랫폼과의 DW 오프로드를 통해 연간 2.5백만 달러의 비용을 절감

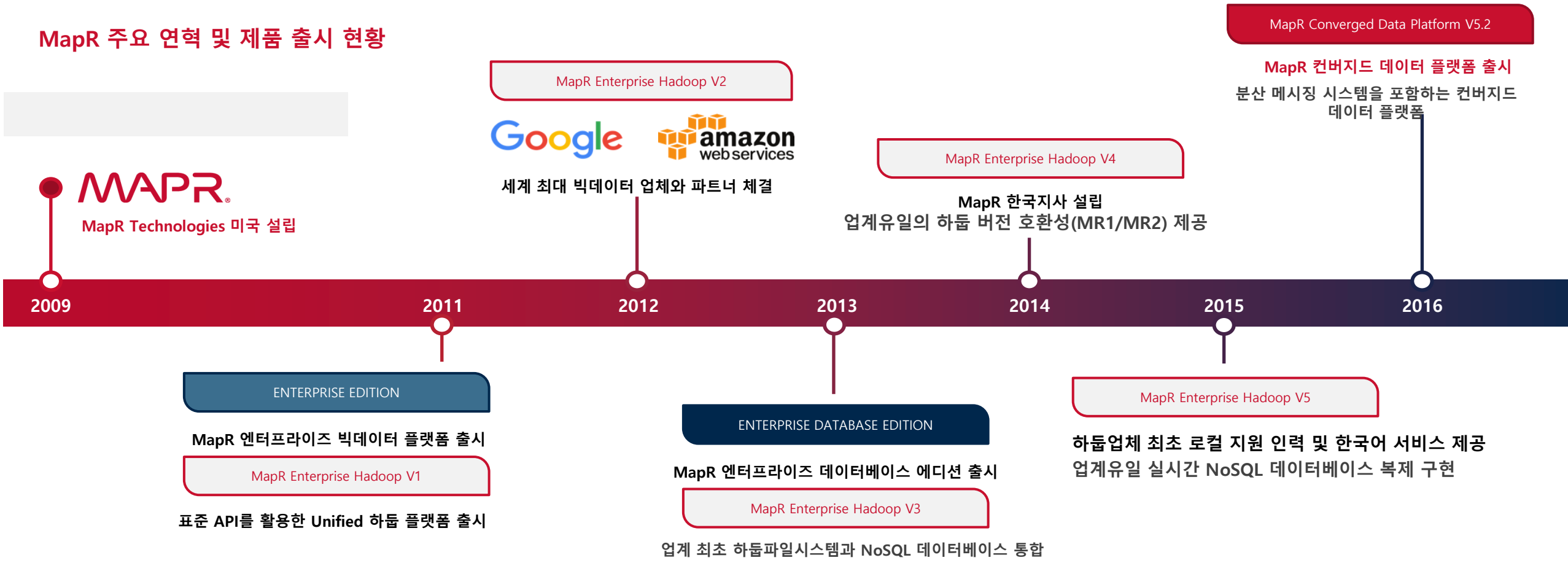
1. MapR Technologies 소개

DIFFERENT
THAN
"THE OTHERS"

MapR Technologies 연혁

"MapR Technologies는 2009년 설립되었으며, 엔터프라이즈 환경이 요구하는 신뢰성 높은 빅 데이터 플랫폼을 제공하는 회사입니다.
본사는 미국 산호세에 위치해 있으며, 빅 데이터 플랫폼 시장에서 **가장 많은 유료고객을 확보**하고 있습니다."

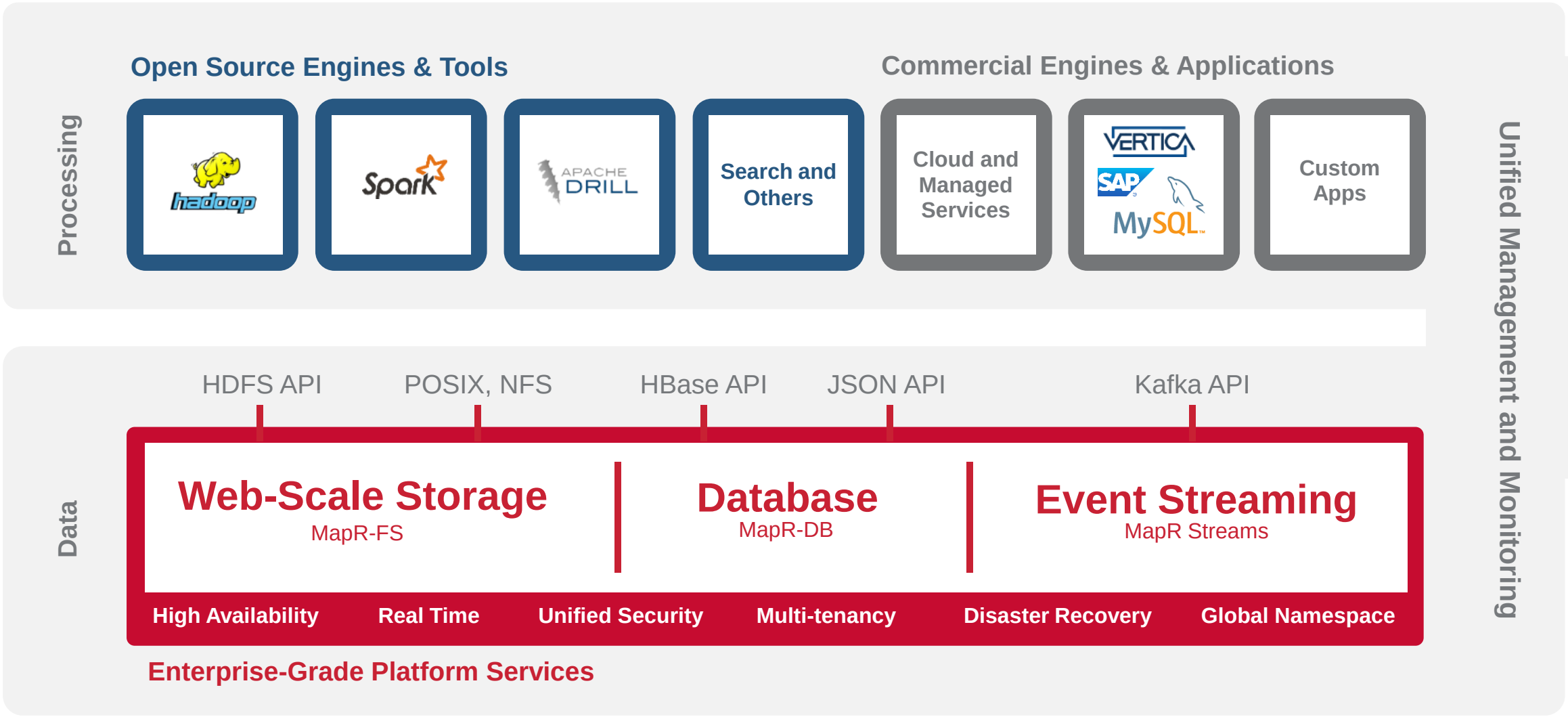
MapR 주요 연혁 및 제품 출시 현황



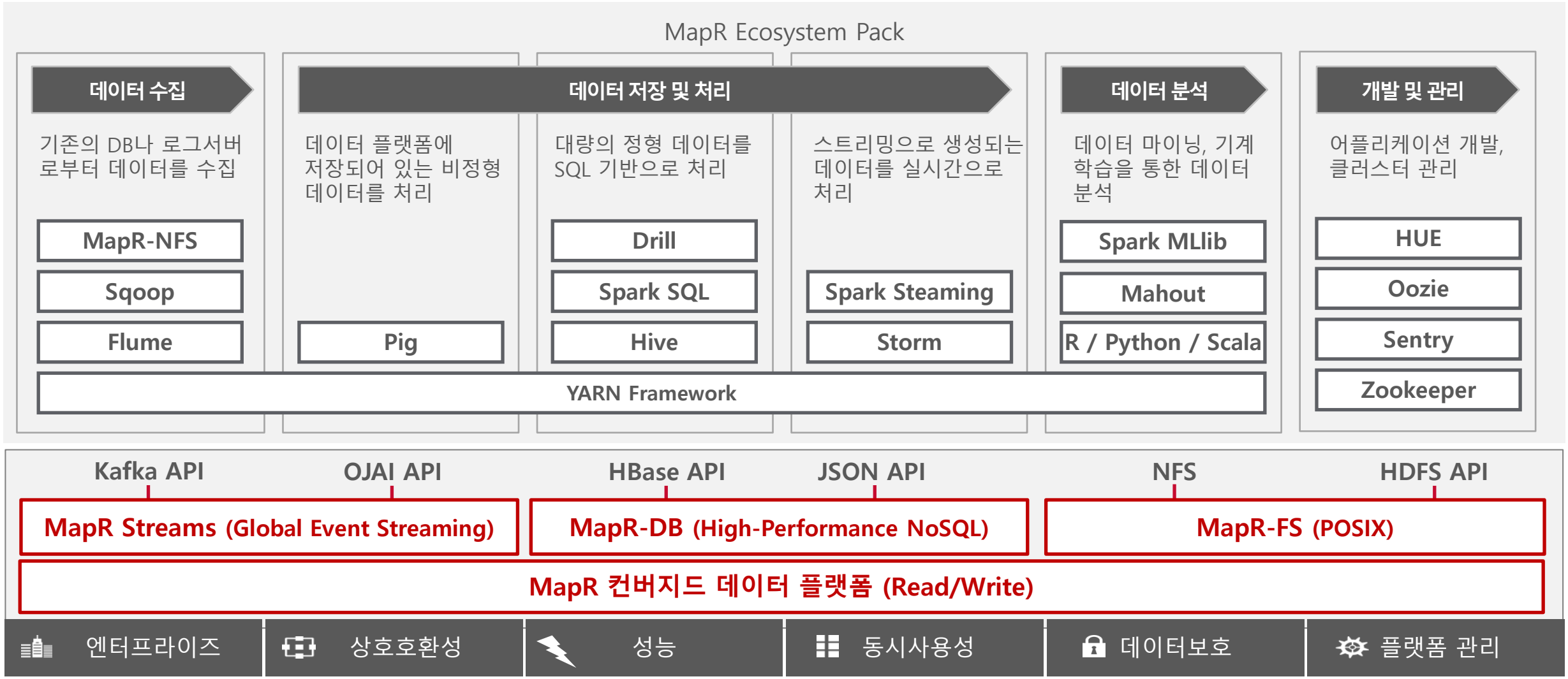
MapR 글로벌 파트너

MapR 산업별 글로벌 고객

MapR Converged Data Platform



MapR Core & MapR Ecosystem Pack

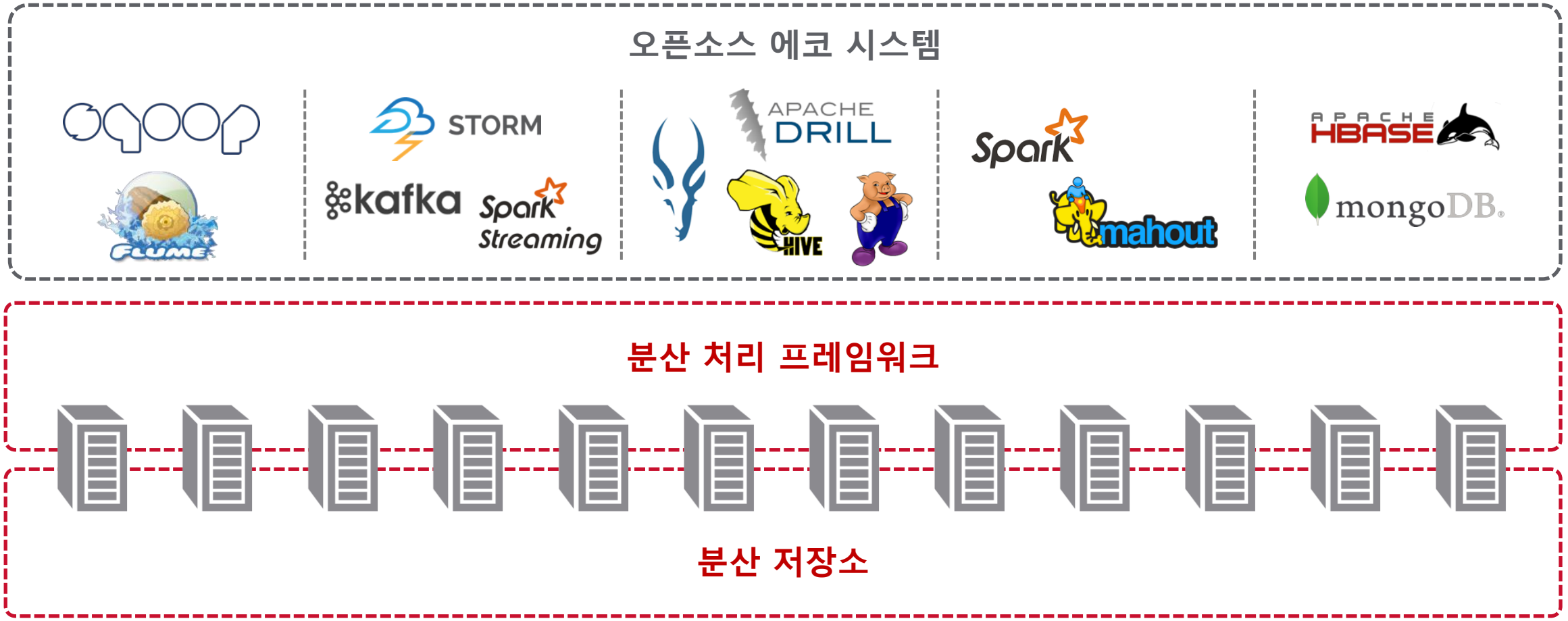


2. Hadoop 기반 플랫폼 구축과 한계

DIFFERENT
THAN
"THE OTHERS"

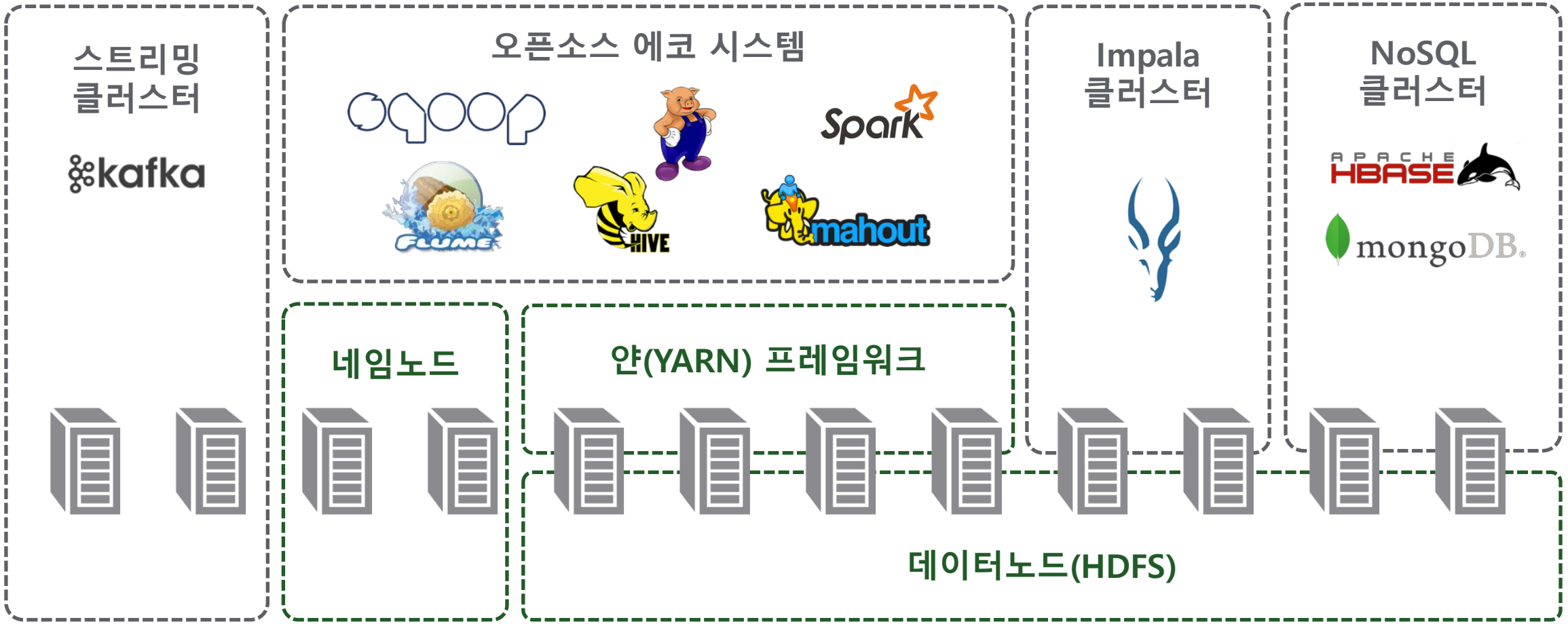
Hadoop 이란?

"Hadoop이란 대량의 데이터를 보관 및 처리할 수 있는 분산 저장소 및 분산처리 프레임워크 이며 Hadoop에 저장되어 있는 데이터를 수집, 저장, 처리, 분석 하기 위한 다양한 에코시스템이 클러스터 컴퓨팅 기반으로 구성되어 있는 환경"



하지만 실제 Hadoop 플랫폼을 구축해 보면...

"기업환경에서 Apache Hadoop 기반의 빅데이터 플랫폼을 구축하여 보면, 다양한 기업니즈에 대응하기 위해 수많은 컴포넌트들이 활용되게 되고 이러한 컴포넌트의 불완전한 결합으로 인해 **빅데이터 통합 플랫폼의 복잡도가 오히려 증가하는 현상**이 발생"



하지만 실제 Hadoop 플랫폼을 구축해 보면...

"기업환경에서 Apache Hadoop 기반의 빅데이터 플랫폼을 구축하여 보면, 다양한 기업니즈에 대응하기 위해 수많은 컴포넌트들이 활용되게 되고 이러한 컴포넌트의 불완전한 결합으로 인해 **빅데이터 통합 플랫폼의 복잡도가 오히려 증가하는 현상**이 발생"

외부 데이터 처리 지연
서비스 가용성 축소

대면/비대면 결합 저해
비정형 데이터 처리 어려움

노드 수 증가에 따른 비용 증가

단일 장애점 발생

대고객 서비스 품질 저하
엔터프라이즈 수준 가용성 확보 불가

어느 빅데이터 프로젝트 매니저의 고민

"막상 Hadoop 프로젝트를 다 끝내놓고 보니..."

수많은 컴포넌트들...
수많은 클러스터들...
대체 **이 많은 관리포인트를**
어떻게 관리하라는거야...

뭐가 이렇게 어려운거야
익숙한 기술이라고는
하나도 없고...하아...
팀원 교육이 걱정이다...

그리고 Hadoop 이라는게
원래 이렇게 느린거였어?
이건 내 예상보다 훨씬
성능이 안나오잖아...

이거 정말 기**업에서**
사용하라고 만든거야?
재난대책은? 보안은?
걱정이다...

3. MapR Converged Data Platform 특장점

DIFFERENT
THAN
"THE OTHERS"

단일 장애점 제거로 서비스 레벨 품질 확보

단일 장애점 제거로 서비스 레벨 품질 확보

단일 장애점 제거로 서비스 레벨 품질 확보

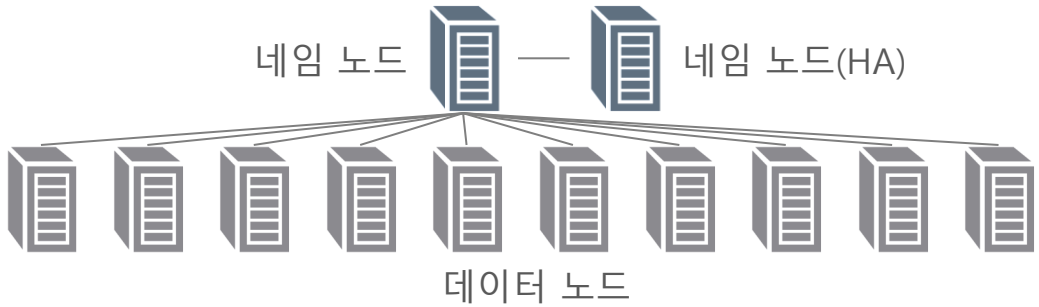
단일 장애점 제거로 서비스 레벨 품질 확보

단일 장애점 제거로 서비스 레벨 품질 확보

"Apache Hadoop의 필수적인 요소인 네임노드는 분산환경에서 데이터 블럭의 위치과 변경사항을 관장하는 핵심 요소이나 단일 구성으로 장애와 병목의 주요 요소. MapR은 **네임노드를 제거한 아키텍처**로 서비스 레벨의 품질 확보가 가능한 유일한 플랫폼"

기존 Hadoop File System

- 네임노드 장애시 전체 클러스가 중지되는 단일 장애점 발생
- 대기(Stand-by) 네임노드는 분산처리가 아닌 장애 대비
- 대량의 노드가 구성된 클러스터 환경에서 네임노드가 부하시 전체 클러스터의 성능이 저하되는 치명적인 구조

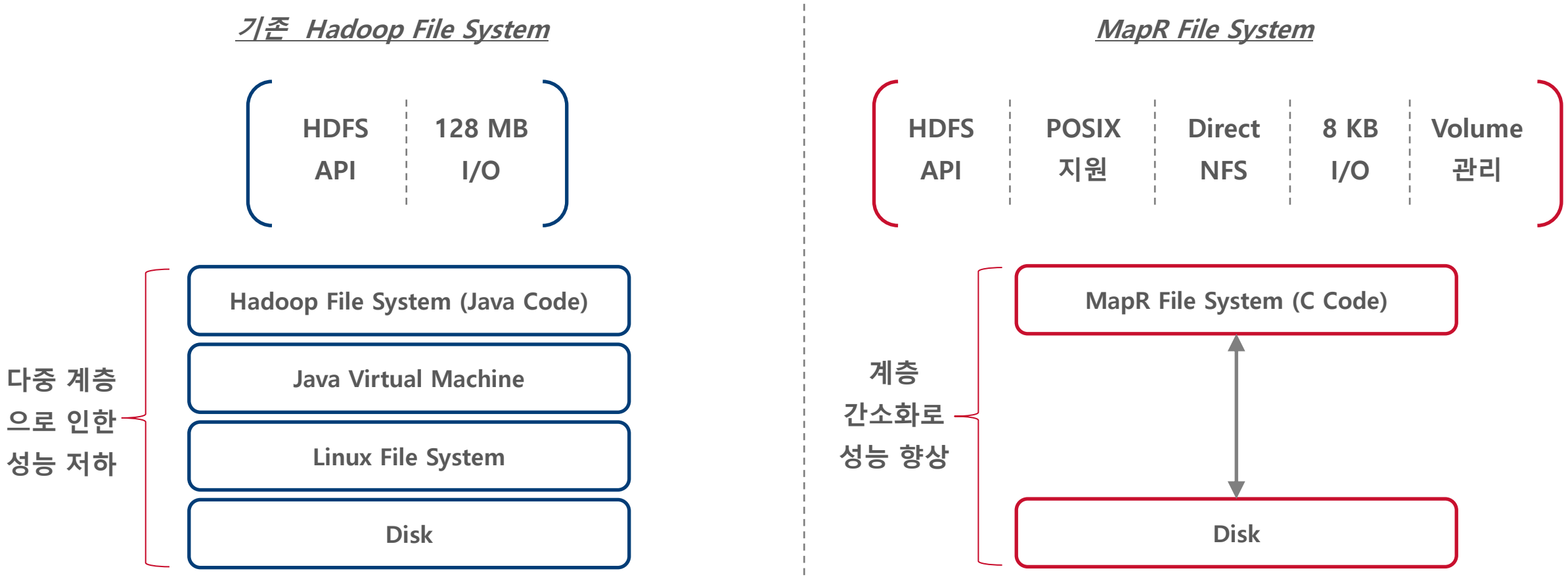


MapR File System

- 네임노드를 제거하여 **단일 장애점(SPOF) 제거**
- CLDB(Container Location Database) 기술기반 운영
- 대량의 노드가 구성된 클러스터 환경에서도 컨테이너 위치와 변경사항을 분산처리 하여 **성능에 영향을 미치지 않는 구조**

실시간 빅데이터와 서비스 성능 품질

"금융산업에서 빅데이터 플랫폼을 통한 선제적 의사결정 환경을 구현하기 위해서는 적시성을 확보하는 것이 핵심. MapR 플랫폼은 기존 Hadoop의 아키텍처를 완전히 혁신하여 **기업수준의 적시성 확보와 기존 개발 환경과 유사한 환경**을 제공"



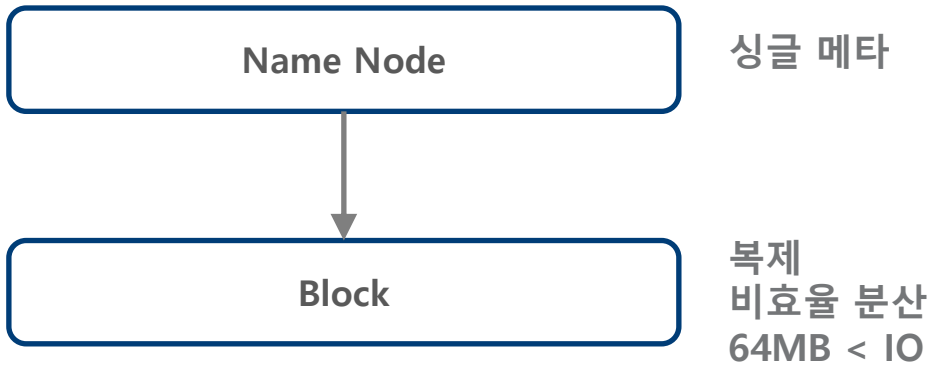
빅데이터 저장에 최적화된 MapR-FS의 시스템 구조

Volume	User-Defined
Container	10-30GB (Self-adjusting)
MapR-FS Files / MapR-DB Tables / MapR Streams	Sizes vary
Chunks / Tablets / Partitionlets	Sizes vary
Blocks	8KB (Fixed)
Storage Pools	3 disks (default)
Disks	Physical/Virtual
Data Node	Physical/Virtual

빅데이터 저장에 최적화된 MapR-FS의 시스템 구조

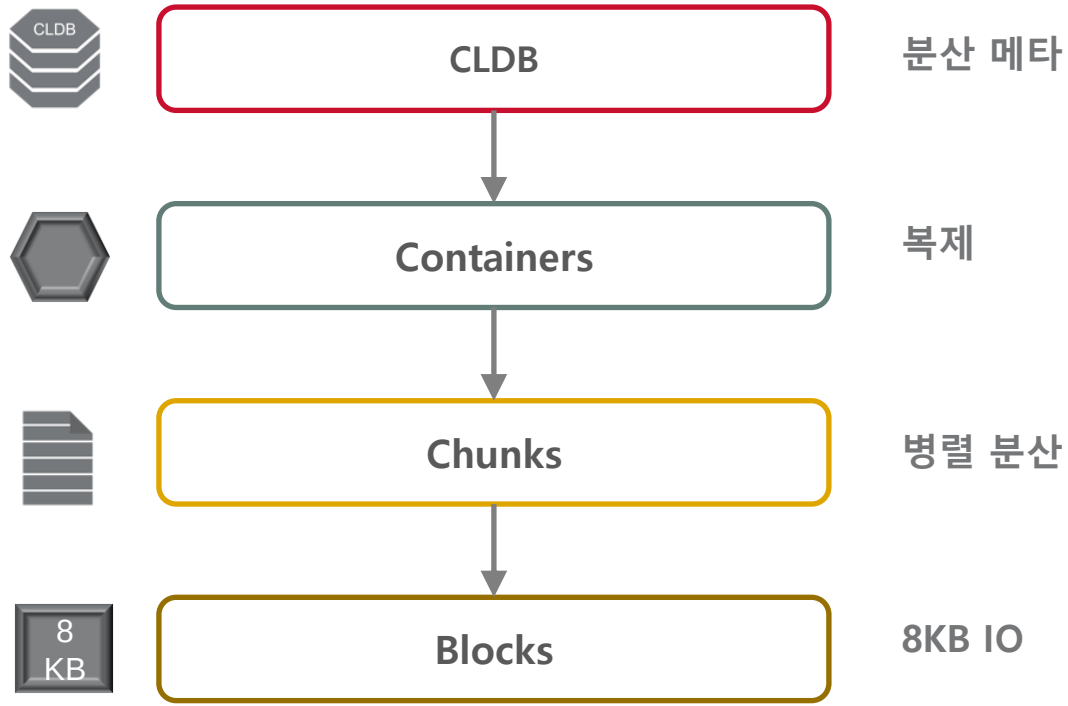
"MapR-FS는 Random Read/Write를 위한 8KB IO단위를 갖으면서도 데이터는 효과적으로 분배되어 병렬처리를 진행하며 GB 단위의 복제를 통해 메타관리를 보다 간소하게 관리하는 혁신적인 빅데이터 플랫폼 입니다."

기존 Hadoop File System



- ✓ *IO 처리 단위, 저장 단위, 복제 단위가 모두 동일하여 빅데이터를 효과적으로 처리가 불가능*
- ✓ *데이터 사이즈 증가로 대량의 Block이 생성되고 처리될때 Name Node의 부하로 인한 급격한 성능저하*
- ✓ *오로지 Read만을 위한 구조*

MapR File System



동일한 사용자 경험을 위한 MapR NFS

"MapR의 NFS 기능은 IT 관리자, 개발자, 데이터 사이언티스트이 특별한 Hadoop 커맨드를 익히지 않고도 **기존의 파일시스템과 동일한 방법**으로 MapR-FS를 사용할 수 있도록 지원하는 기능 입니다. 이 기능을 통하여 레거시 솔루션과도 연동이 가능 합니다"

드래그&드롭을 통한 파일의 활용 가능

일반적인 파일 브라우저를 사용하여 MapR-FS의 데이터를 손쉽게 활용

로그와 같은 파일을 MapR-FS 에 직접 Write

개별 서버 내부에 로그를 저장하고 Hadoop 으로 이관하는 방법이 아닌 MapR-FS에 바로 Write

커스터 마이징 없이 어플리케이션에서 활용

Read/Write 파일 시스템을 통해 Hadoop 전용 커넥터가 없는 어플리케이션도 활용

```
$ find . | grep log
$ cp /mapr/cluster
$ scp /mapr/cluster
$ vi results.csv
$ tail -f part-00000
```

표준 OS 유틸리티 사용

MapR-FS 는 POSIX 기반 파일 시스템으로 일반적인 리눅스 유틸리티를 그대로 활용 가능

데이터 압축

"MapR File System은 기본적으로 압축 기능이 적용. Hadoop 파일 시스템 압축 기능 사용시 과도한 성능저하가 발생하는 경쟁사 제품과 달리 MapR 파일 시스템은 IO 레이어의 간소화를 통해 압축 기능을 사용하여도 최소한의 오버헤드만 발생"

압축 타입	압축률	압축 속도	압축해제 속도
LZ4	2.084	330 MB/s	915 MB/s
LZF	2.076	197 MB/s	465 MB/s
ZLIB	3.095	14 MB/s	210 MB/s

엔터프라이즈 수준의 가용성 확보

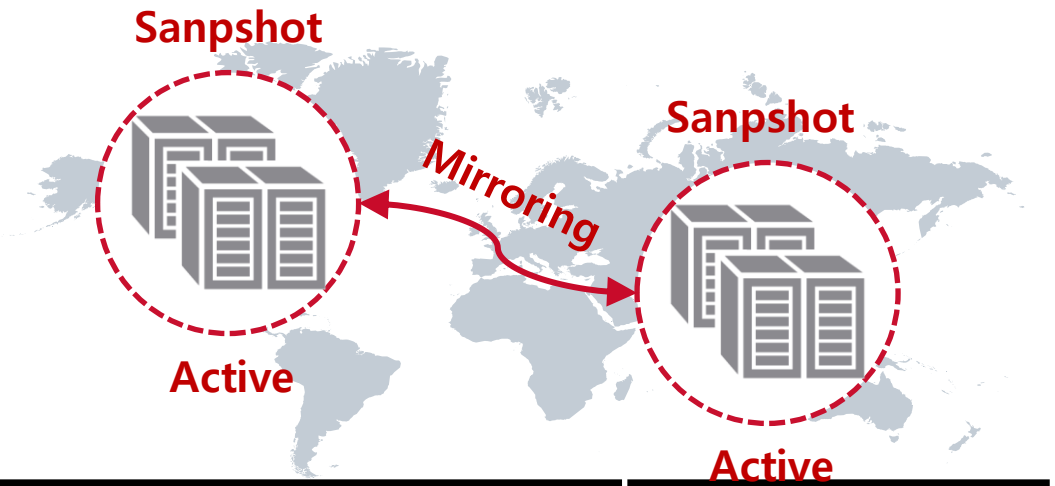
"금융업은 업의 특성상 DW 데이터를 포함하여 모든 데이터가 미션크리티컬 데이터로 분류되는 산업으로 빅데이터 역시 데이터의 안전성 확보가 매우 중요. MapR은 이러한 기업요건을 제품에 반영하여 **제품 엔진레벨에서 백업 및 HA/DR 환경을 구성**"

기존 Hadoop File System

- 태생적으로 HA/DR 구성이 고려되지 않은 아키텍처
- 단순 파일 복제 수준의 DR 센터 구성으로 데이터 유실 가능
- Active – Standby 수준의 DR 센터 구성

MapR File System

- 스냅샷 / 미러링 기능을 통한 내/외부로의 데이터 보호
- 압축 기능을 통한 저비용 고성능 DR 센터 구축 가능
- Active – Active HA/DR 센터 구성으로 클러스터 효율 극대화



엔터프라이즈 수준의 보안 기능 제공

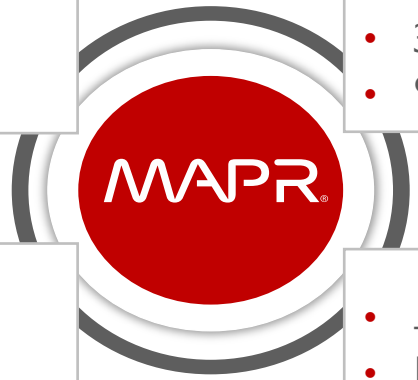
"기업의 보안 요건을 충족하기 위해 MapR은 권한, 인증, 암호화, 감사 보안 기능을 제공. 클러스터에서 파일 단위까지 미세한 보안설정으로 기업 내/외부의 플랫폼 사용자들에 대한 안전한 데이터 활용 환경을 제공"

Authorization

- Access Control Lists (ACLs) 관리
- 볼륨 단위 권한 제어
- MapR Table 접근 권한 제어
- 표준 UNIX 파일 시스템 퍼미션 지원

Encryption

- Wire-level 데이터 암호화
- NSA Suite B 암호화 (AES-256 및 SHA-256) 를 활용하여 표준 기반 기본 인증을 지원
- 3rd Party 솔루션연계를 통한 파일/컬럼 레벨
- 암호화



- MapR Ticket 적용
- UserID & Password (PAM, LDAP)
- Kerberos 연동

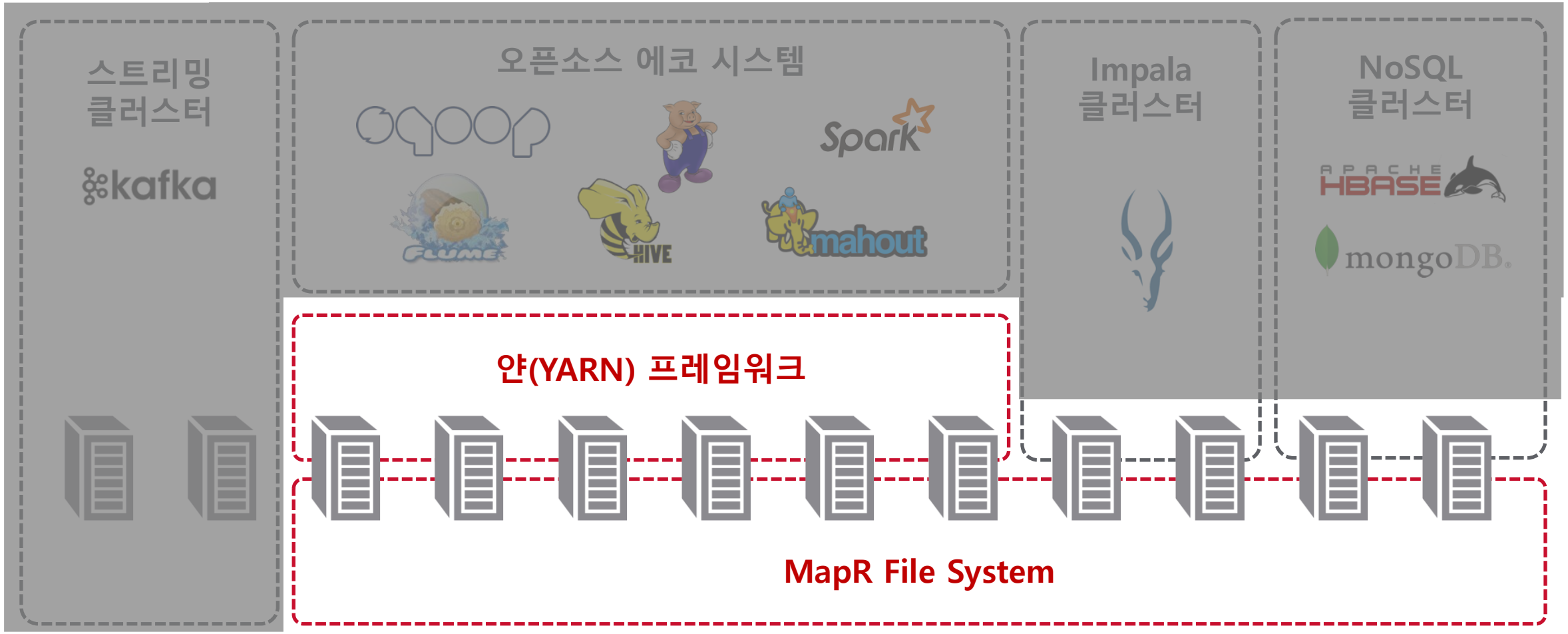
Authentication

- 모든 이벤트 JSON 로그파일로 저장
- 데이터 액세스 및 관리 포함
- SQL과 표준 BI툴을 통한 감시로그 질의 및 커스텀 리포트 생성

Auditing

MapR Converged Data Platform 통합

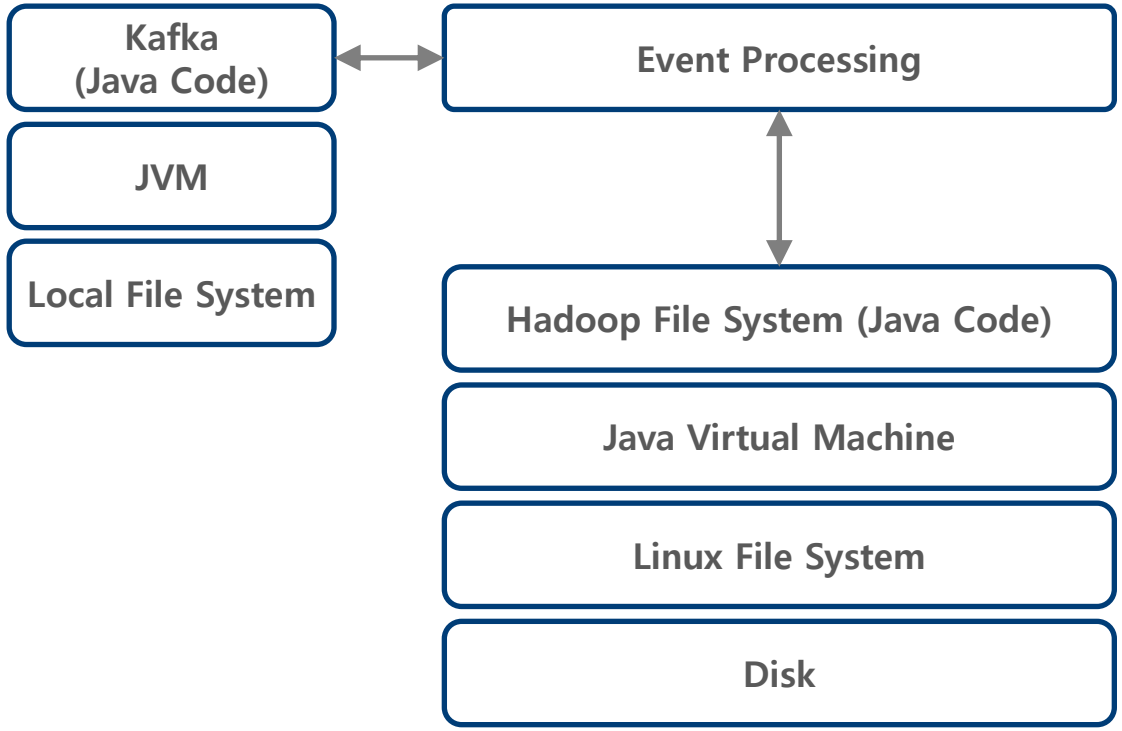
"MapR Converged Data Platform은 기업요건이 반영되지 않은 단순한 오픈소스 기반의 빅데이터 플랫폼이 아닌 엔터프라이즈 환경에서 필요로 되는 안전성, 보안, 성능을 강화한 업계 유일의 엔터프라이즈 대응이 가능한 빅데이터 플랫폼"



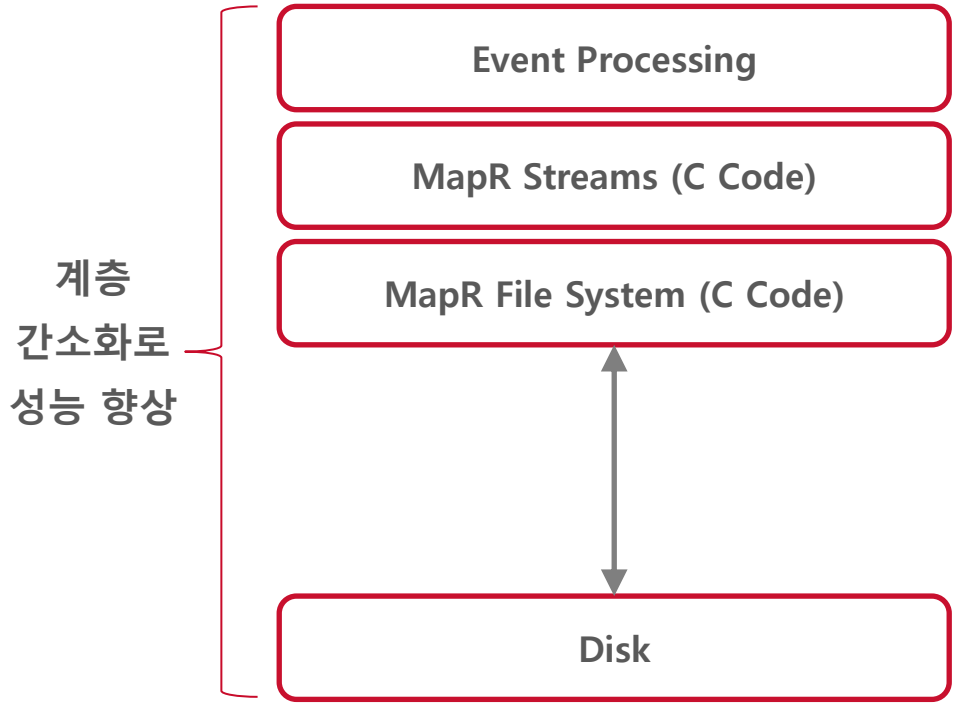
MapR Streams 개요

"스트림 데이터는 아직 저장되지는 않았으나 유입이 되고 있는 데이터로 빅데이터 환경에서 현재 발생되고 있는 상황이나 상태를 실시간으로 활용하고자 하는 요건이 증가. 하지만 오픈소스 Kafka 기반의 스트림 처리는 구조적 한계로 많은 이슈를 발생"

기존 Kafka 구동 방식



MapR Streams 구동 방식

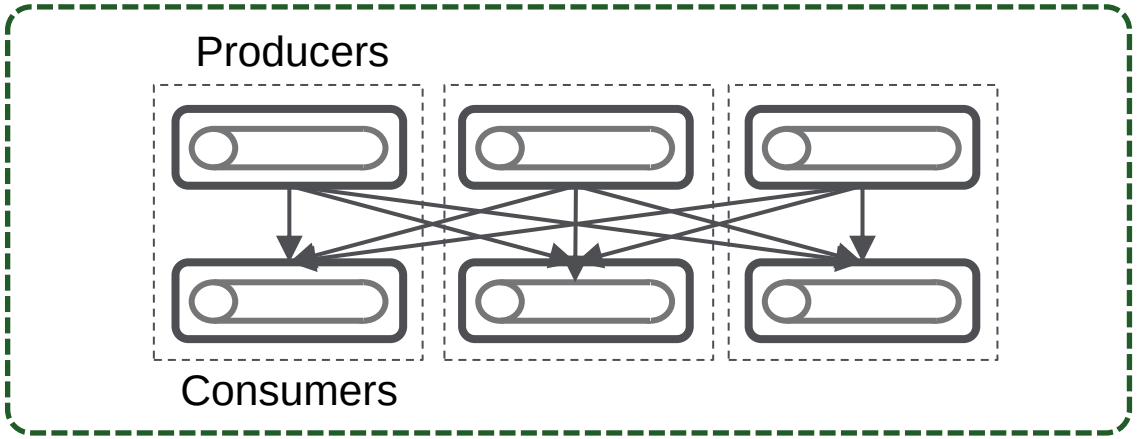


스트림 데이터의 적시성 확보

"금융업과같은 서비스 상품성 제품들은 오퍼링의 적시성이 매우 중요한 요소로 작용 합니다. 특히 시장 또는 개인의 상황을 실시간으로 파악할 수 있어야 합니다. SNS, IoT, Mobile 데이터에 대한 유실없는 안정적 스트림 처리와 고속 처리로 적시성 확보가 가능"

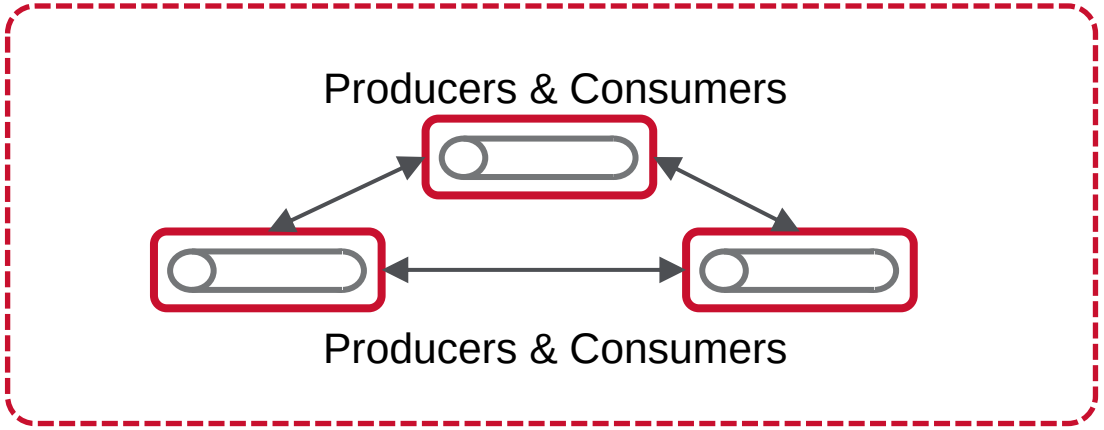
기존 Kafka 구동 방식

- 외부 모듈인 MirrorMaker 에 의한 복제 구조
- 메시지 오프셋들은 새 클러스터에 저장
- 메시지 루핑을 피하기 위한 복잡한 2-stage 클러스터



MapR Streams 구동 방식

- 코어 플랫폼에 의한 복제 구조
- 메타데이터/오프셋들과 함께 메시지 복제
- 메시지 루핑 방지 기술 내장

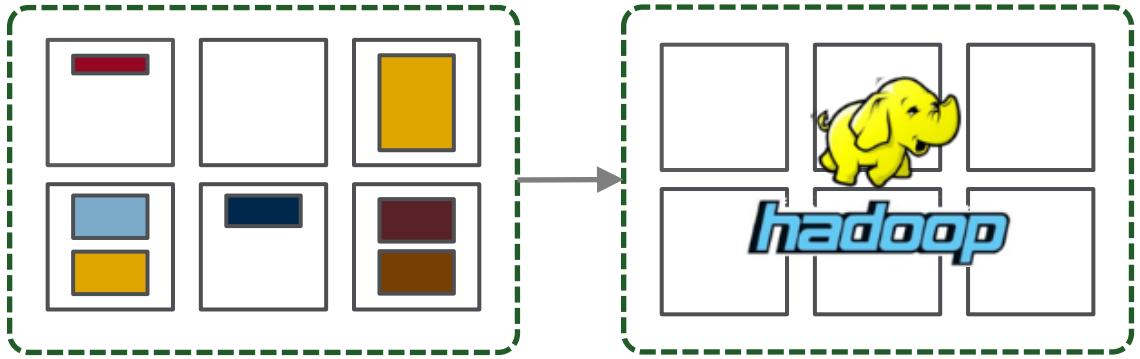


스트림 데이터의 분산 처리와 가용성

"기존 Kafka 스트림 클러스터는 Hadoop 외부에 위치하여 네트워크 통신에 의해 메시징 처리를 수행 합니다. 만약 스트림 처리를 위한 CEP와 같은 스트림 분석 솔루션의 장애가 발생시 Kafka 서버 내부 파일 버퍼의 사이징 한계로 **데이터의 유실 가능성이 발생**합니다."

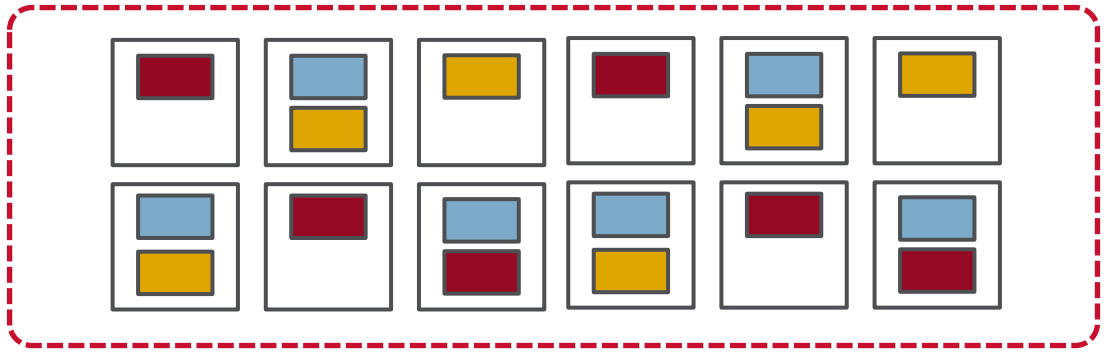
기존 Kafka 구동 방식

- 파티션을 단일 디스크에 할당
- 파티션의 유입률, 리텐션 정책등을 고려하지 않고 할당
- 수작업에 의한 파티션 이동



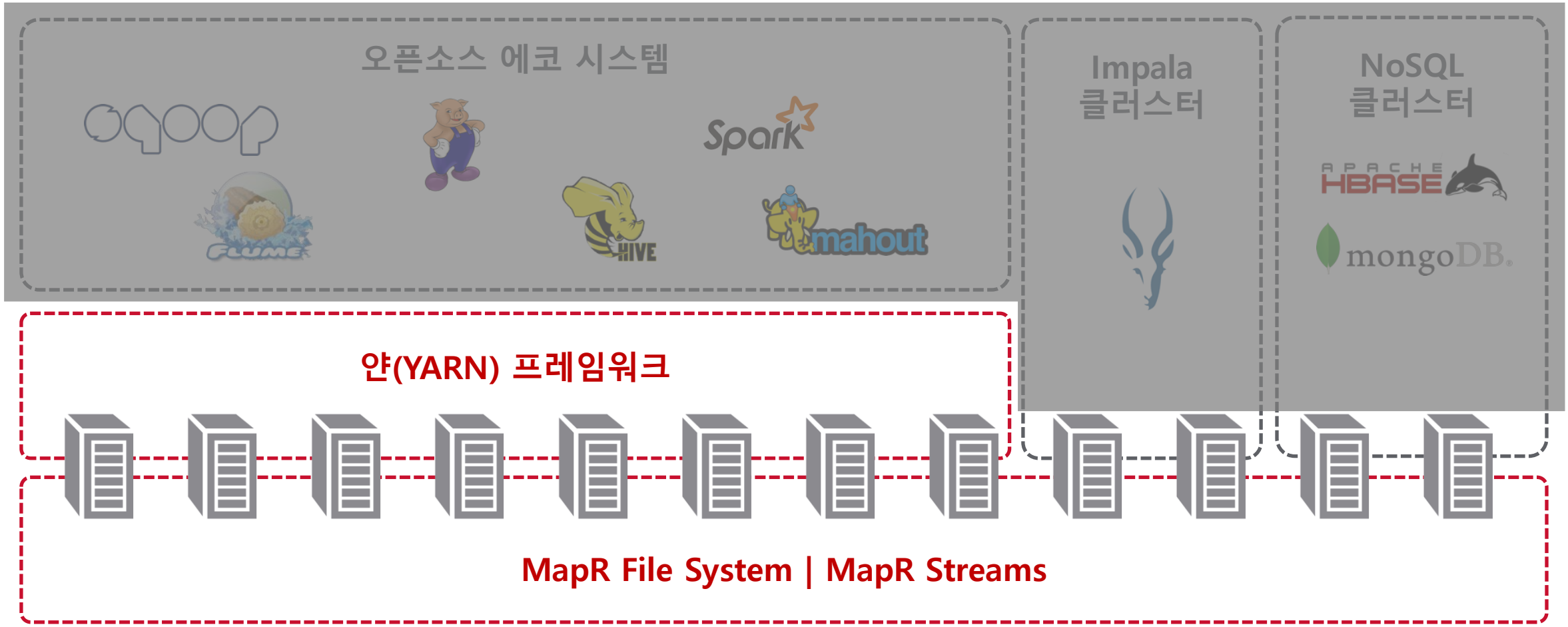
MapR Streams 구동 방식

- 파티션을 **MapR 파일시스템에 분할/분산**하여 할당
- 분할된 파티션을 활용율기반으로 할당
- 주기적으로 분할된 파티션의 리밸런싱 수행



MapR Converged Data Platform 통합

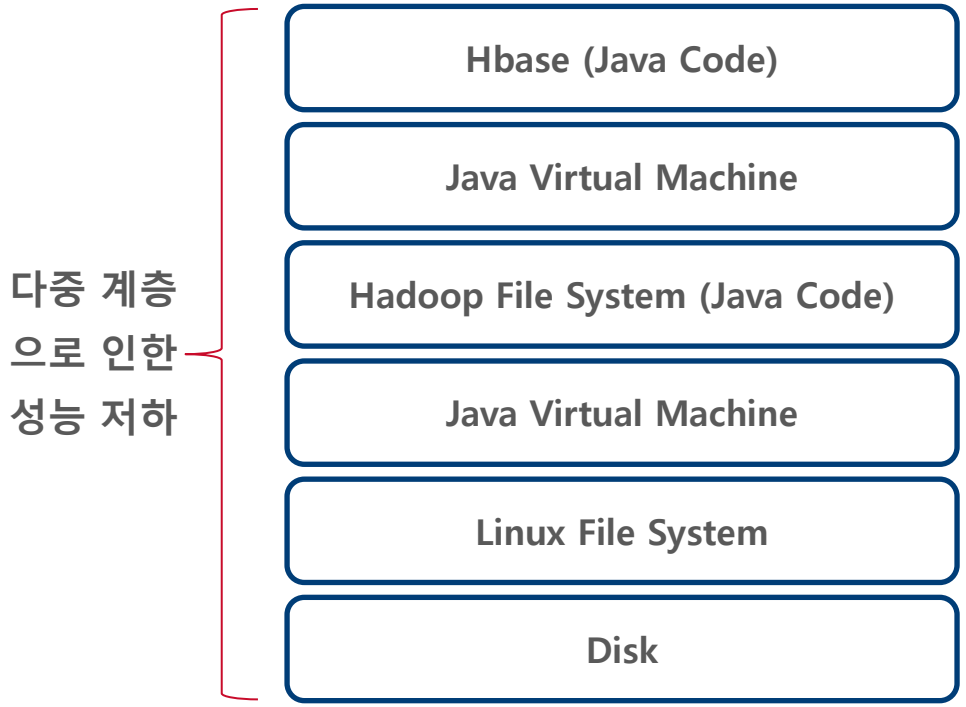
"MapR Converged Data Platform은 기업요건이 반영되지 않은 단순한 오픈소스 기반의 빅데이터 플랫폼이 아닌 엔터프라이즈 환경에서 필요로 되는 안전성, 보안, 성능을 강화한 업계 유일의 엔터프라이즈 대응이 가능한 빅데이터 플랫폼"



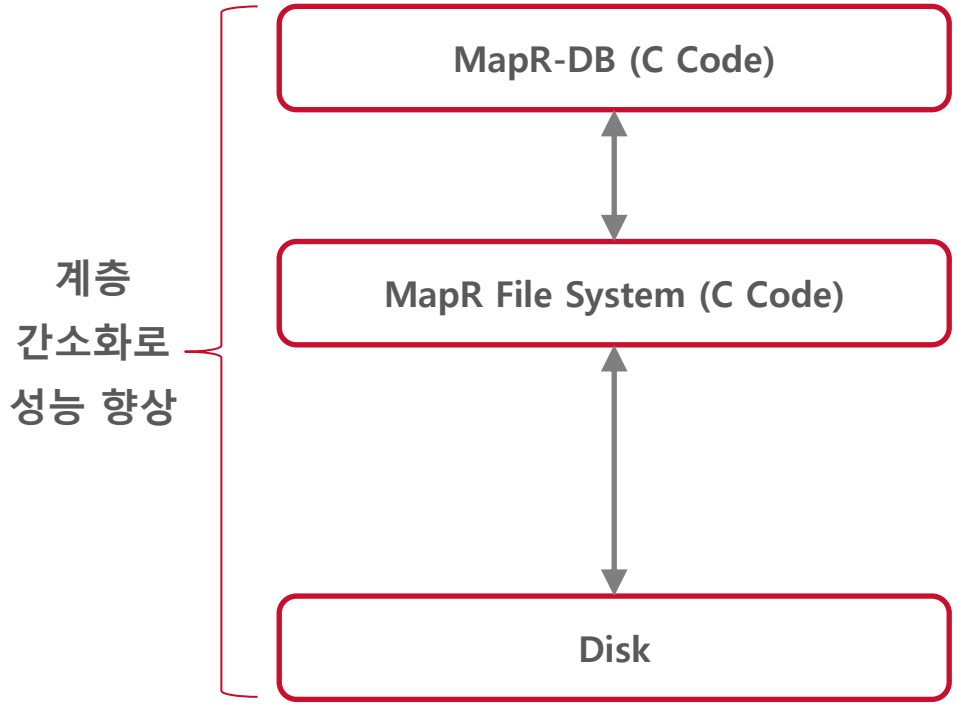
MapR-DB 개요

"Hbase는 Hadoop 환경에서 비정형 데이터를 저장하는 NoSQL 솔루션. 하지만 Hbase 구조적 한계로 인해 성능 및 안전성 확보가 매우 큰 이슈. MapR-DB는 이러한 제약 사항을 제거하여 MapR FS와 DB를 결합시켜 **고속으로 Schema less 데이터를 처리**"

기존 Hbase 구동 방식



MapR-DB 구동 방식

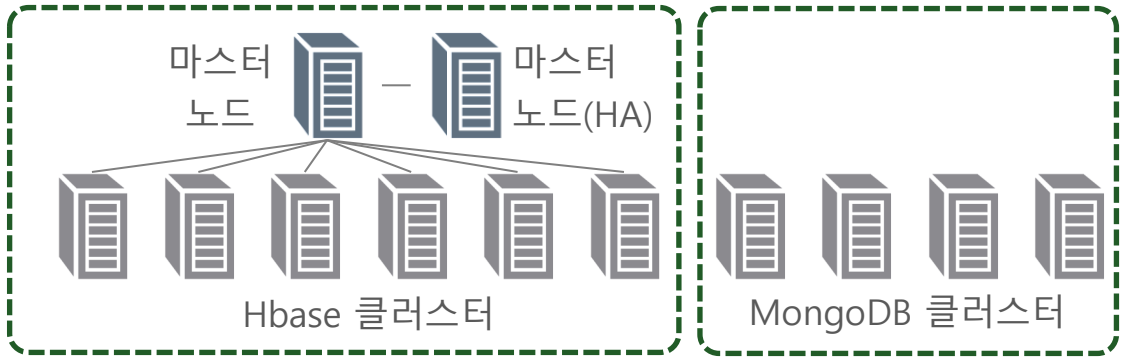


기업수준의 반정형 NoSQL 데이터를 처리하는 MapR-DB

"고객 대응 후 생성되는 데이터를 기존 정형 DBMS로 처리하는 것은 스키마 구성 및 관리에 대한 부담이 가중. 웹이나 모바일앱, ARS, VoC등으로 생성되는 데이터들에 대해 고객중심의 데이터를 구성하기 위해서는 **성능과 안전성이 확보된 NoSQL DB**가 필요."

기존 Hbase 구동 방식

- 마스터 노드가 단일장애점(SPOF)
- 연산영역이 독립된 영역에서만 사용
- 바이너리 테이블만 제공, 다큐먼트 NoSQL은 별도로 구성
- HA/DR 위한 기능 미제공



MapR-DB 구동 방식

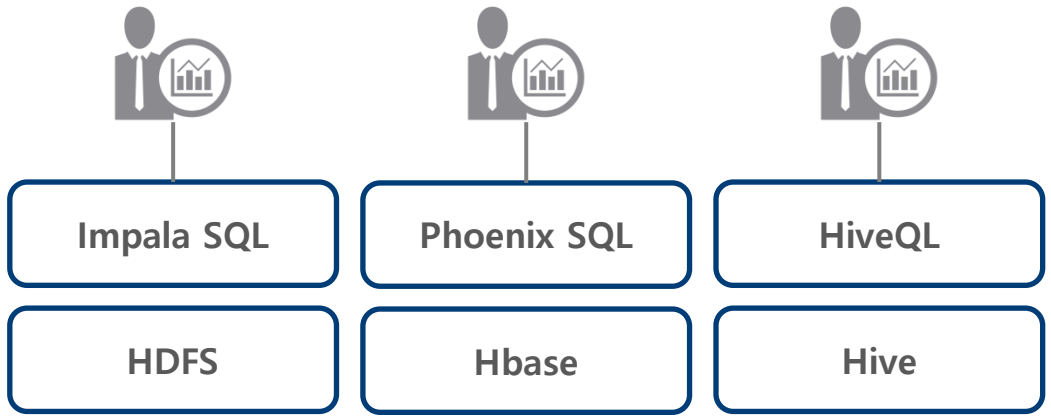
- 마스터 노드를 제거하여 단일 장애점(SPOF)을 제거
- MapR 클러스터 전체를 연산 영역으로 통합하여 사용
- 바이너리 테이블 뿐만 아닌 JSON 다큐먼트 테이블 제공
- HA/DR을 위한 Table 레벨 동기화 기능 제공

Apache Drill 소개

"Hadoop환경의 많은 SQL on Hadoop 기술들이 존재. Apache Drill은 MapR이 주도하는 프로젝트로 이기종 Hadoop 에코 시스템들에 대해 동일한 **ANSI SQL을 기준으로 개발 환경을 통합**하는 프로젝트로 가장 쉽게 사용할 수 있는 SQL on Hadoop기술"

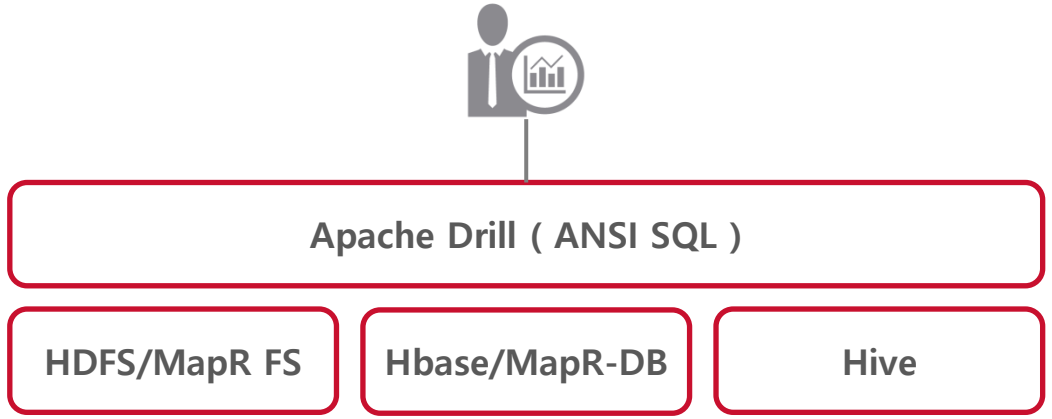
기존 SQL on Hadoop 기술

- 각각의 에코 시스템마다 독립적인 SQL on Hadoop 엔진
- ANSI SQL이 아닌 독립적인 SQL기반의 언어를 사용
- HDFS의 파일을 스키마 구성없이 SQL 직접접근이 불가



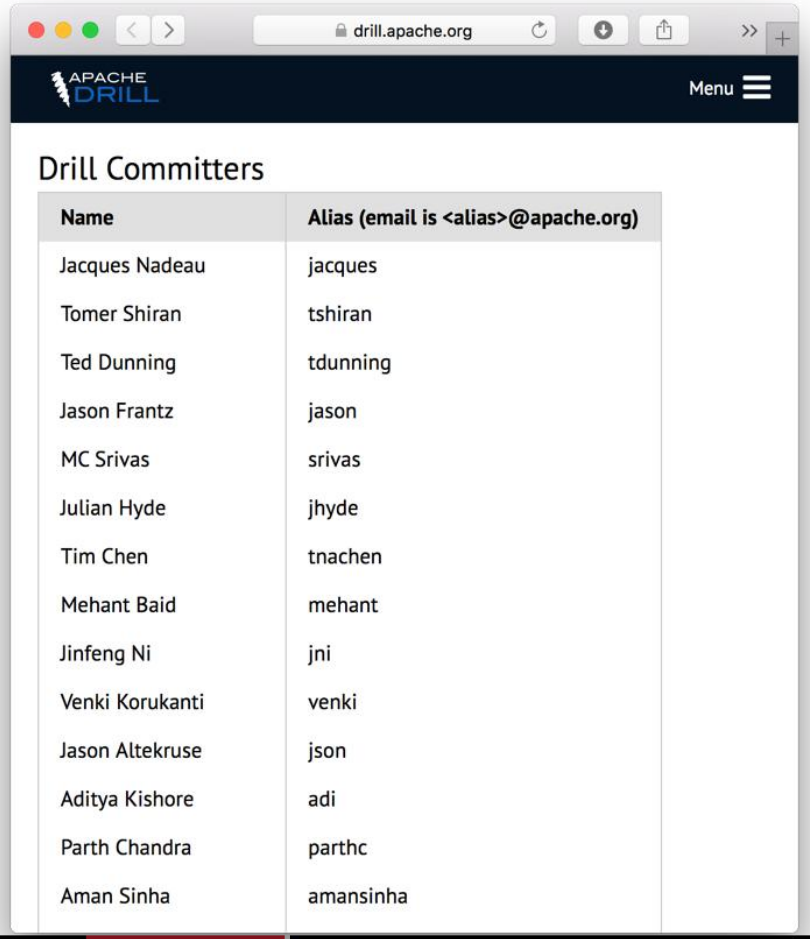
Apache Drill

- Drillbit 이라는 동일한 SQL on Hadoop 엔진 사용
- **ANSI 92 표준 SQL**을 모든 소스에 대해 동일하게 사용
- MapR FS 파일을 스키마 구성없이 SQL 직접접근 가능

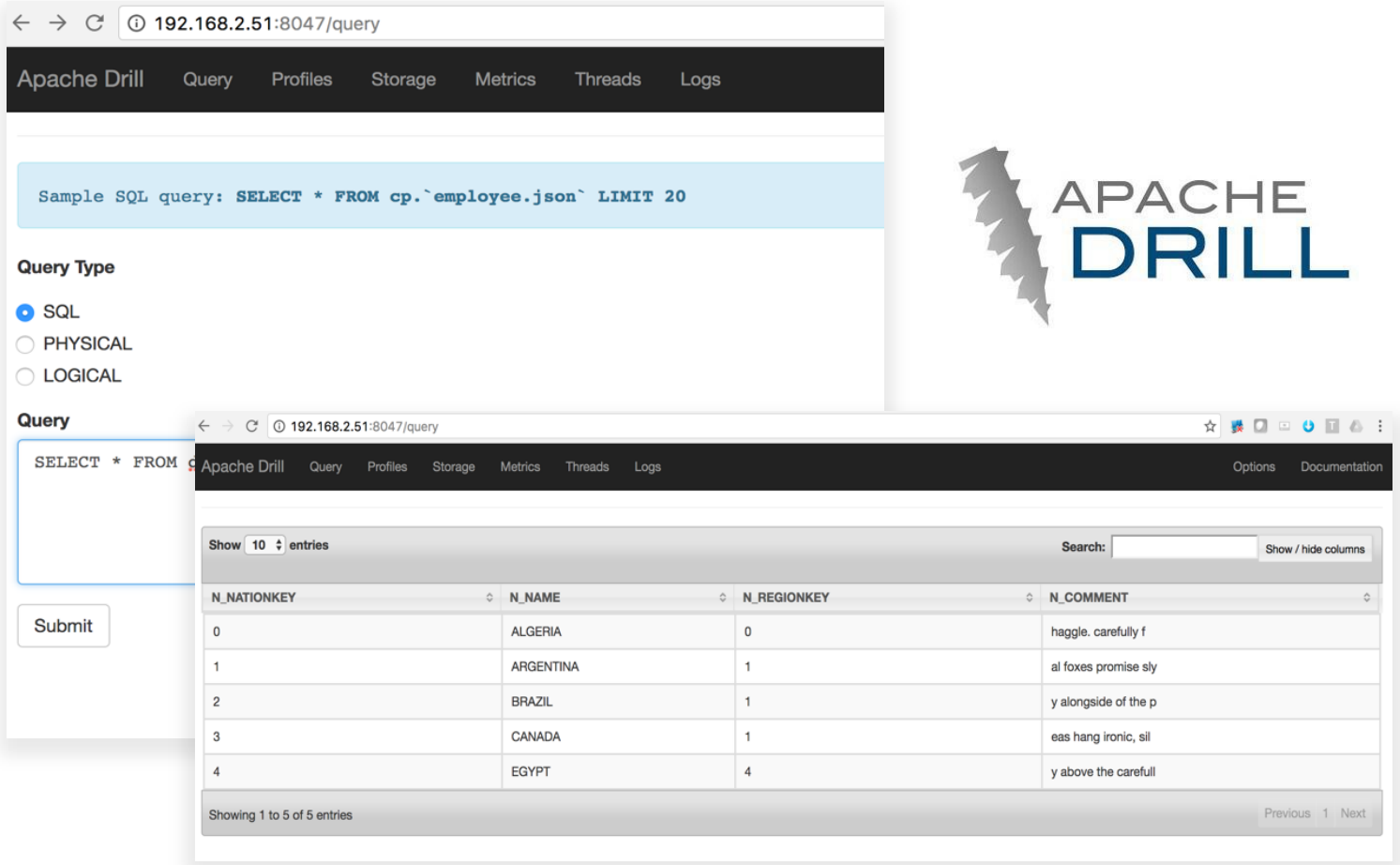


Apache Drill 소개

"Apache Drill은 상당 부분의 커미터들이 MapR 소속의 개발자들로 구성되어 코드를 기여하고 있으며 MapR-FS, HDFS, Hive, MapR-DB, Hbase 등 데이터 소스와 스키마에 자유로운 SQL on Hadoop 프로젝트. 자체 개발 Web UI 또는 Zeppelin 등과 통합하여 활용"



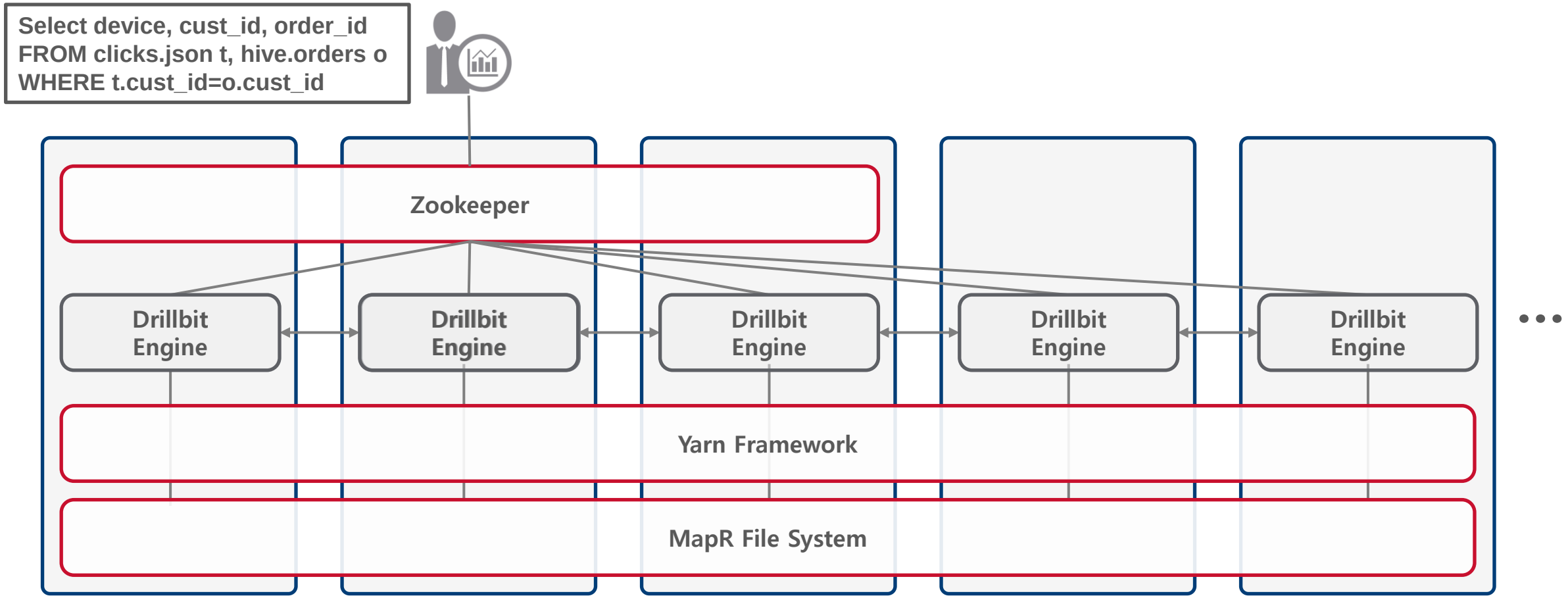
Name	Alias (email is <alias>@apache.org)
Jacques Nadeau	jacques
Tomer Shiran	tshiran
Ted Dunning	tdunning
Jason Frantz	jason
MC Srivas	srivas
Julian Hyde	jhyde
Tim Chen	tnachen
Mehant Baid	mehant
Jinfeng Ni	jni
Venki Korukanti	venki
Jason Altekruse	json
Aditya Kishore	adi
Parth Chandra	parthc
Aman Sinha	amansinha



APACHE DRILL

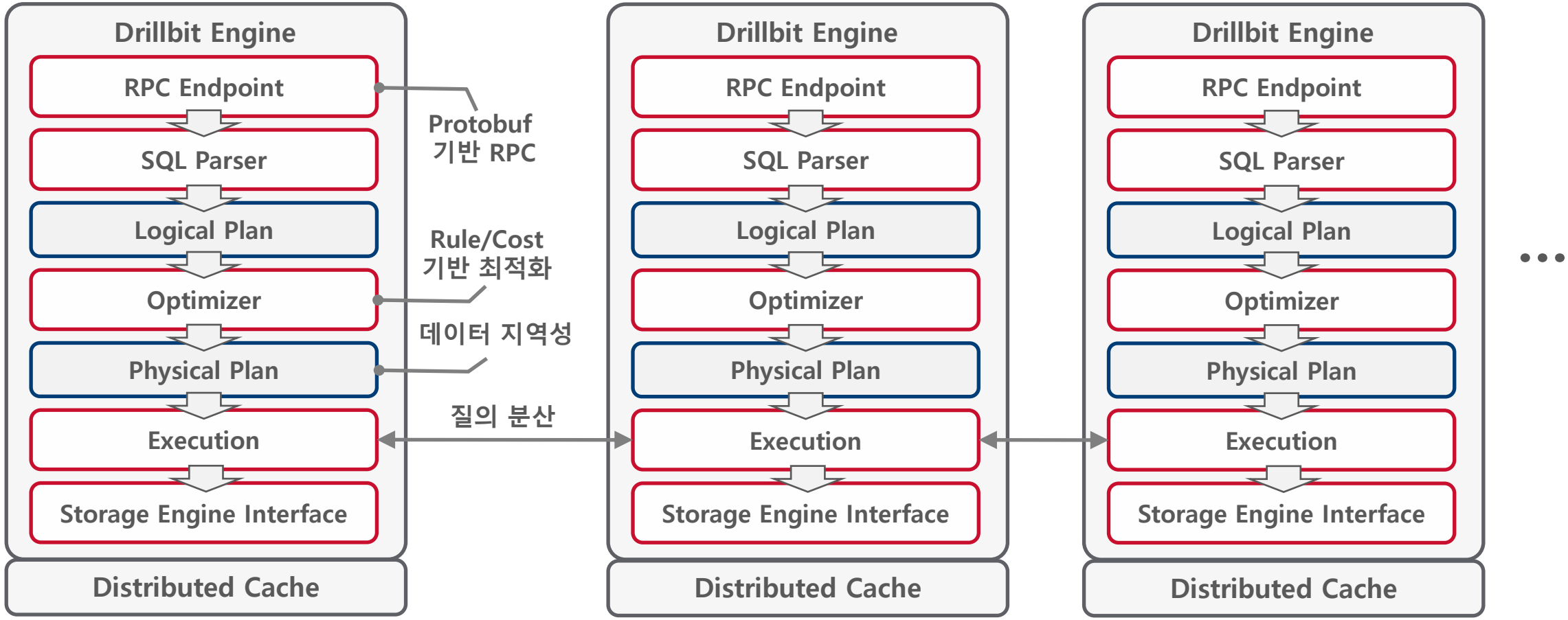
Apache Drill 소개

"Apache Drill은 MapReduce 엔진이 아닌 Drillbit이라는 전용 데이터 처리 엔진을 보유. 사용자는 Zookeeper로부터 질의를 전달할 Drillbit엔진을 할당받고 질의를 수행 하며 각 엔진은 캐시된 메타 데이터를 조회하여 데이터 지역성을 기반으로 데이터를 탐색"



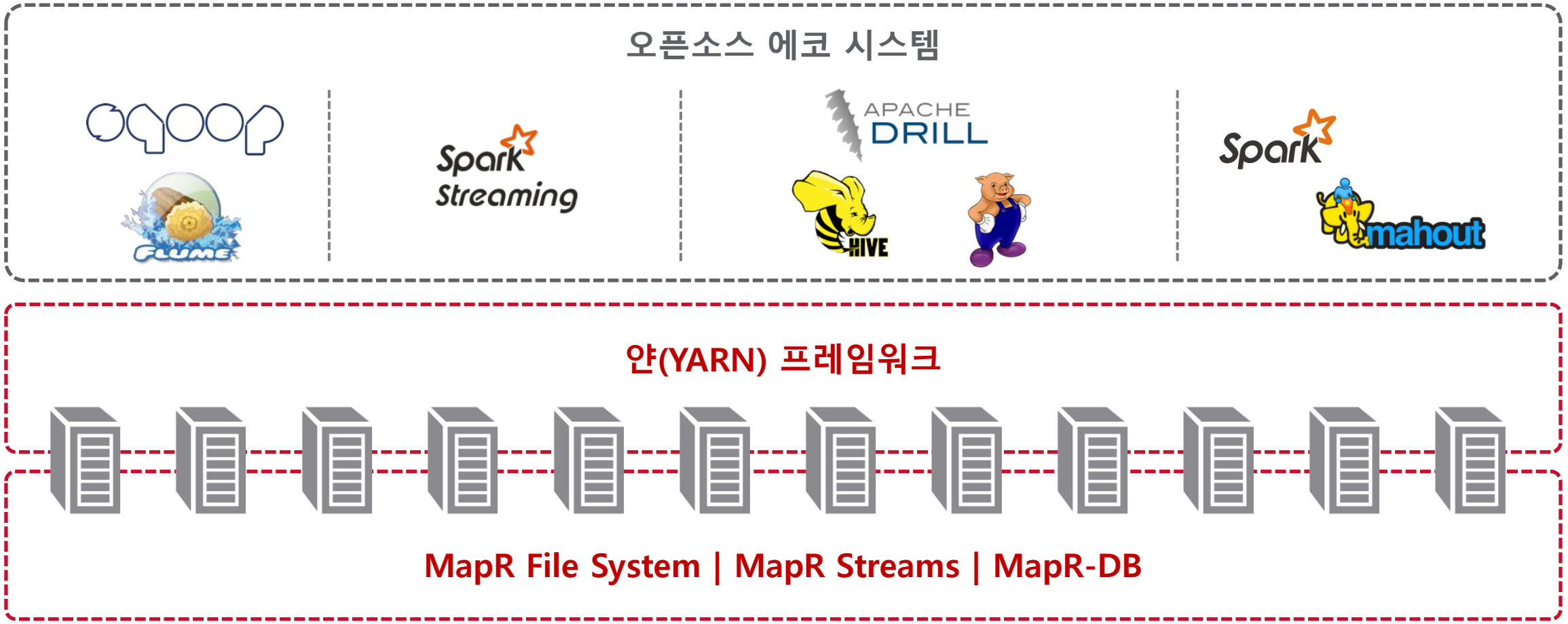
Apache Drill 소개

"질의를 파싱하기 위한 Drillbit 엔진이 선정되면 쿼리를 Logical Plan과 Physical Plan으로 각기 수행계획을 생성. 이때 데이터의 지역성을 최대한 활용하되기 위한 정보를 수집하고 각 계획을 트리구조로 조각내어 각 분산 노드의 Drill 엔진에 전달"



MapR Converged Data Platform 통합

"MapR Converged Data Platform은 기업요건이 반영되지 않은 단순한 오픈소스 기반의 빅데이터 플랫폼이 아닌 엔터프라이즈 환경에서 필요로 되는 안전성, 보안, 성능을 강화한 업계 유일의 엔터프라이즈 대응이 가능한 빅데이터 플랫폼"



MapR Converged Data Platform의 특장점 요약

"MapR Converged Data Platform은 기업이 직면하고 있는 빅데이터 문제를 해소한 업계 유일의 **기업용 빅데이터 플랫폼** 입니다"

오픈소스 기반 Hadoop Platform

- 네임노드 및 마스터노드의 단일 장애점 문제
- JVM및 OS File System의 복잡한 단계로 인한 성능저하
- Hadoop 전문가만이 시스템 운영 및 개발이 가능
- HA/DR 및 보안기능의 취약으로 기업수준 대응 불가
- 다량의 관리노드 구성이 필요로 되어 TCO 증가

MapR Converged Data Platform

- 네임노드 및 마스터노드를 제거하여 단일 장애점 제거
- JVM 및 OS File System 영역을 제거하여 성능 향상
- POSIX, NFS 기능 등으로 기존의 인력으로 운영 및 개발 가능
- 코어레벨로 HA/DR 및 보안기능 제공으로 기업수준 대응
- 간소화된 클러스터 구성으로 노드를 감소하여 TCO 절감

4. 오픈소스 에코시스템

DIFFERENT
THAN
"THE OTHERS"

하둡 에코 시스템 – 데이터 수집

"다양한 원천 시스템에 있는 로우 데이터를 데이터 저장 영역인 MapR-FS(HDFS)와 연계하기 위해 아래와 같은 기술을 사용합니다. "

컴포넌트	설 명
MapR-NFS	• 읽기/쓰기가 가능한 NFS를 제공하여 소스 데이터 시스템에 마운트하여 신속히 데이터를 로드 할 수 있음 • 하둡 파일시스템 접근에 대한 POSIX 사용 지원
Apache Sqoop	• 간단한 CLI(Command Line Interface)로 Oracle, MySQL 등의 RDBMS의 특정 테이블 또는 특정 조건에 맞는 데이터를 HDFS로 쉽게 옮길 수 있으며, Hive, Pig, HBase 등으로 바로 옮겨 확인 가능 • HDFS에 저장되어 있는 데이터를 RDBMS로 적재 또는 RDBMS 데이터를 HDFS로 적재 기능 • 하둡용 ETL 도구
Apache Flume	• 대용량의 로그 데이터를 분산,안정성,가용성을 바탕으로 효율적으로 수집,집계,이동이 가능한 로그수집 기능 • 장애가 나더라도 로그,이벤트를 유실 없이 전송함을 보장 • 수평확장(Scale-Out)이 가능하여 분산수집이 가능한 구조로 설계됨 • 다양한 소스로부터 데이터를 수집하여 다양한 방식으로 데이터를 전송이 가능하지만, 아키텍처가 단순하고 유연하며 확장 가능한 데이터 모델을 제공하여, 실시간 분석 애플리케이션을 쉽게 개발 • 각종 Source, Sink등 제공으로 쉽게 확장 가능

하둡 에코 시스템 – 데이터 처리

"MapR은 수집된 데이터에 대한 조회 및 처리를 위해 SQL on Hadoop과 같은 다양한 오픈 소스 기술들을 제공 합니다. "

컴포넌트	설 명
Apache Drill	• 다양한 정형 또는 비정형 데이터(Hbase, Hive, MongoDB, Json., CSV등)에서 즉각적인 데이터 탐색을 제공 • ANSI SQL 사용으로 손쉽게 쿼리를 작성
Apache Spark	• 하둡 기반의 고급 실시간 분석 엔진 • 고속 쿼리 수행 툴, 머신 학습 라이브러리, 그래프 프로세싱 엔진, 스트리밍 분석 엔진 제공
Apache Hive	• 전문적인 분석코딩 능력이 필요 없이 간편하게, 접근하기 쉽도록 RDB의 SQL문과 유사한 HQL(Hive Query Language)을 제공 • 테이블과 파티션과 관련된 메타정보를 하이브 메타스토어에 저장
Apache Pig	• 대용량의 데이터 집합을 분석하기 위한 플랫폼 • 특수목적에 맞게 사용자 함수를 만들어 확장성을 제공

하둡 에코 시스템 – 데이터 보관 및 워크플로우 처리

"MapR은 보안과 성능이 강화된 비용효율적인 스케일 아웃 방식의 분산 저장 시스템인 MapR-FS, MapR-DB를 제공합니다. "

컴포넌트	설 명
MapR-FS (HDFS)	• 분산 데이터 처리를 위해서 데이터를 저장하기 위한 파일 시스템 • 비 NameNode 아키텍처 • 미러링을 통해 복제본을 원격 사이트에 생성하여 재해 복구(DR)를 지원 • 대용량 데이터 (terabyte 또는 petabyte)를 저장하도록 고안. 이를 위해서 데이터를 여러 대의 컴퓨터에 나누어 저장. 즉, NFS보다 훨씬 큰 파일도 지원 • 데이터 액세스에 대한 성능개선도 용이하며 확장 시 클러스터에 컴퓨터 node만 추가하면 됨
MapR-DB (NoSQL)	• 로그 데이터, 센서 데이터, 메타데이터, 클릭 동향, 사용자 프로파일, 세션 상태 및 링크/의미/관계 데이터를 포함한 광범위한 운영 데이터 형식을 관리하기 위해 열 기반(column-oriented) NoSQL 데이터 모델을 지원 • 높은 처리량 및 지속적인 짧은 대기시간을 제공
Apache Oozie	• 하둡 잡을 관리하기 위한 워크플로우 스케줄러 시스템 • 우지 워크 플로우 잡은 하둡의 다양한 잡을 실행할 수 있는 액션(action)으로 구성되며, 방향성이 있는 비순환 그래프 (DAG: Directed Acyclic Graph)를 구성 • 하둡 스택의 다양한 하둡 잡과 동작하도록 통합되었으며, 여기이는 자바 맵 리듀스, 스트리밍 맵 리듀스, 피그, 하이브, 스쿱, Distcp 등이 포함되며 자바 프로그램이나 쉘 스크립티와 같은 특정 시스템에 특화된 잡도 실행 가능

하둡 에코 시스템 – 외부 솔루션 연계

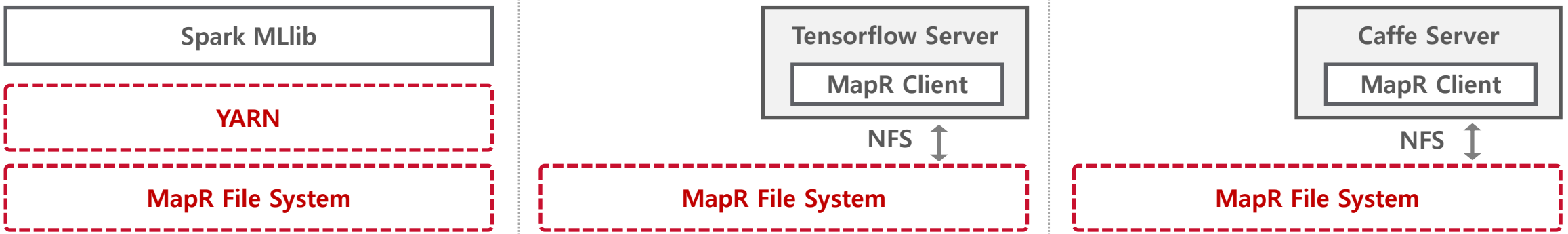
"MapR은 가공된 데이터를 활용하기 위해 다양한 프로토콜 및 API 를 제공 합니다. 다양한 SW들과 유연한 연결성을 제공합니다"

컴포넌트	설 명
HBase-API	• MapR-DB에 저장된 결과 데이터를 HBase-API를 이용하여 다양한 외부 도구과 연계
Kafka-API	• MapR Streams의 데이터를 Kafka-API를 통하여 활용 가능
Hadoop Connector	• Hadoop Connector를 이용하여 다양한 외부 도구과 연계
Apache Drill ODBC/JDBC Driver	• Apache Drill의 ODBC 및 JDBC 드라이버를 이용하여 다양한 UI 툴과 연계
JSON Interface	• MapR-DB, MapR Streams, MapR FS의 데이터를 JSON 기반으로 인터페이스

머신러닝 도구 활용

"MapR 플랫폼은 Spark Mllib, Mahout 머신러닝 라이브러리를 제공하며 고객사의 요구에 따라 Tensorflow, Caffe, H2O와 같은 머신러닝 혹은 딥러닝과 같은 AI 분석 도구들을 활용함에 있어 MapR NFS 기능을 통해 손쉽게 분석 도구들과 연계"

Spark MLlib	Tensorflow	caffe
Spark의 기본 머신러닝 라이브러리로 인메모리 기반으로 머신러닝 알고리즘을 분산처리	구글에서 개발한 딥러닝, 머신러닝 라이브러리로 CPU/GPU 모드로 수행	버클리대에서 개발한 딥러닝 라이브러리로 CPU/GPU 모드로 수행
Classification, Regression, Decision Tree, Recommendation, Clustering, Topic Modeling, Association Rule 등의 알고리즘 제공	Classification, Regression, Clustering, Markov, Neural Network 등의 알고리즘 제공	Artificial Neural Network, Convolutional Neural Network 등 Neural Network 중심의 딥러닝 알고리즘 제공



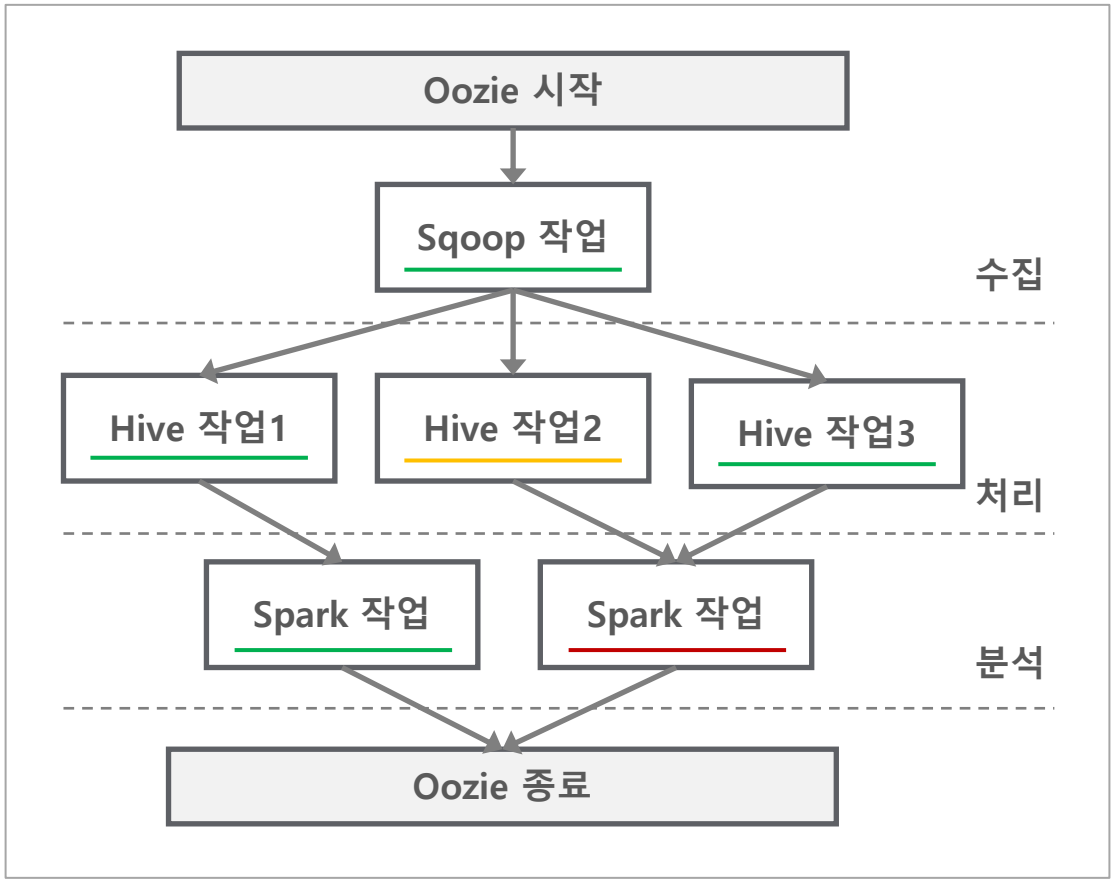
Spark가 지원하는 기계학습 알고리즘

"Apache Spark는 인메모리 기반의 분석을 위해 Mllib 라이브러리를 통해 다양한 기계 학습 알고리즘을 제공"

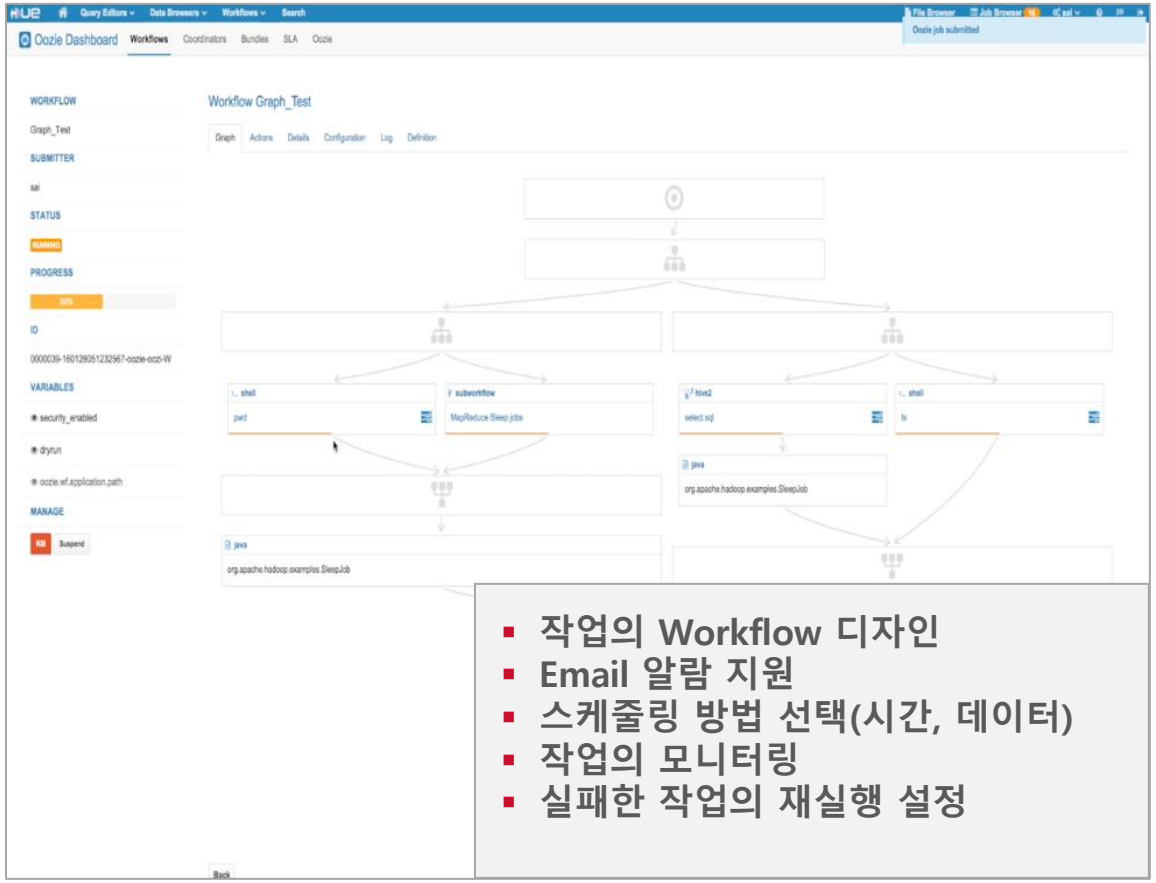
Basic statistics	Classification and regression	Collaborative filtering	Clustering	Dimensionality reduction
▪ Summary statistics	▪ SVMs, logistic regression, linear regression	▪ alternating least squares (ALS)	▪ k-means	▪ Singular value decomposition (SVD)
▪ Correlations	▪ Naive Bayes		▪ Gaussian mixture	▪ Principal component analysis (PCA)
▪ Stratified sampling	▪ Decision trees		▪ Power iteration clustering (PIC)	
▪ Hypothesis testing	▪ Random Forests		▪ Latent Dirichlet allocation (LDA)	
▪ Random data generation	▪ Gradient-Boosted Trees		▪ Streaming k-means	

ETL 및 Workflow 처리

"수집, 처리, 분석 등의 작업을 워크플로우 단위로 스케줄링 하기위해 Apache Oozie를 지원. Apache Oozie를 HUE와 함께 사용하면 GUI 기반의 드래그&드롭 기반으로 설계 할 수 있으며 작업의 시작, 중지, 실패 등의 모니터링도 GUI 환경에서 손쉽게 확인"



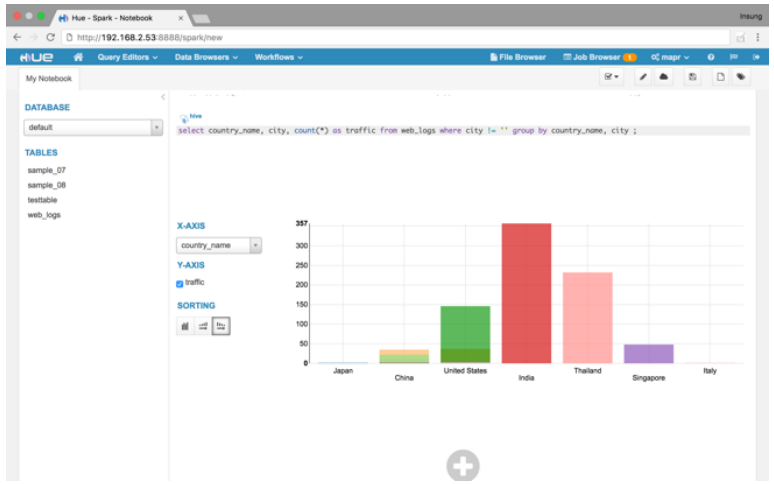
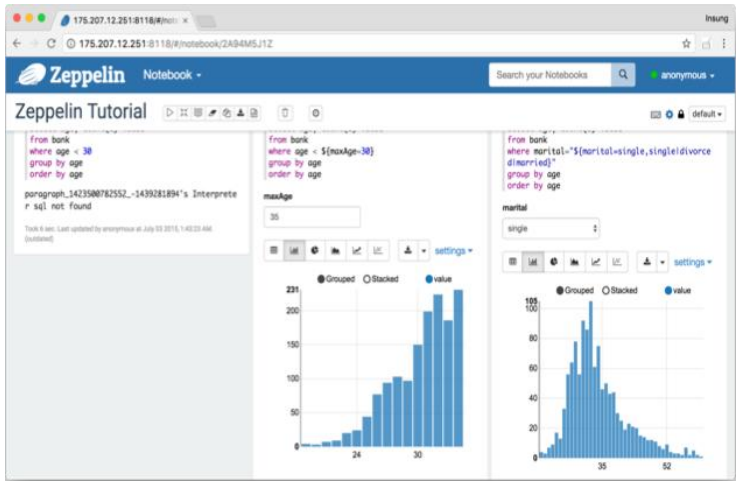
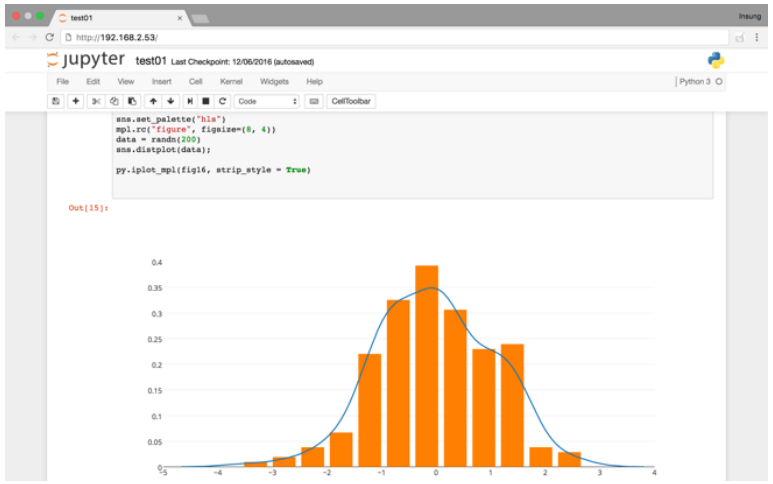
< Oozie 작업의 논리 구조 >



< HUE에서 Oozie 수행 >

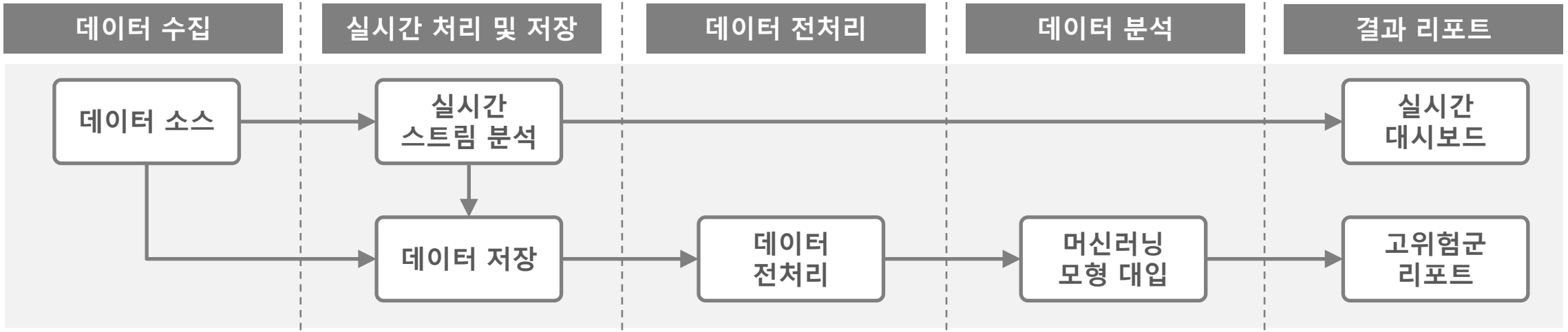
분석 도구 활용

"MapR 플랫폼은 다양한 분석도구 활용을 지원. 분석 환경의 목적과 특성에 따라 적절한 분석 도구를 선정하여 적용. 인터프리터 종류나 시각화 방법에 따라 분석 도구의 선택이 달라질 수 있으나 공통적으로 모든 도구가 Spark를 적극적으로 지원"

HUE	Zeppelin	Jupyter
Hadoop 기반의 가장 대중적인 분석 도구로 쿼리도구 및 노트북 분석 제공	가장 많은 기본 인터프리터를 제공하며 노트북 기반의 분석	Python 노트북 분석에 특화 되어 있으며 다양한 무료/유로 라이브러리 지원
MapR 파일시스템 관리, 파일 권한 관리, Hive, Impala, Spark(Scala, Python, R)등 분석 지원	Hive, Drill, Impala, Spark(Scala, Python, R) 등 분석 지원하며 타 인터프리터 설치 지원	Python 인터프리터만 기본 제공하고 타 인터프리터는 별도 플러그인을 설치하여 지원
		

이상징후 예측 시스템 적용 예

"이상징후 포착 시점을 1차 실시간, 2차 비실시간으로 구분하여 1차 분석에서 발견된 이상징후를 실시간으로 리포트하고 머신러닝 기법이 적용된 2차 심층분석을 통하여 발견된 징후를 리포트하여 **전체 현상을 수리적으로 분석한 결과를 이상징후에 대한 평가 기준**으로 활용."



- 다양한 소스의 데이터 수집
- 실시간 분석 데이터 분류
- 배치 분석 대상 데이터 분류
- Apache Sqoop, Apache Flume, MapR-NFS 기술활용

- MapR-Streams 와 Apache Spark 연계를 통한 룰 기반의 실시간 데이터 처리 및 분석
- MapR-DB, MapR-FS에 원본 및 처리된 데이터 저장

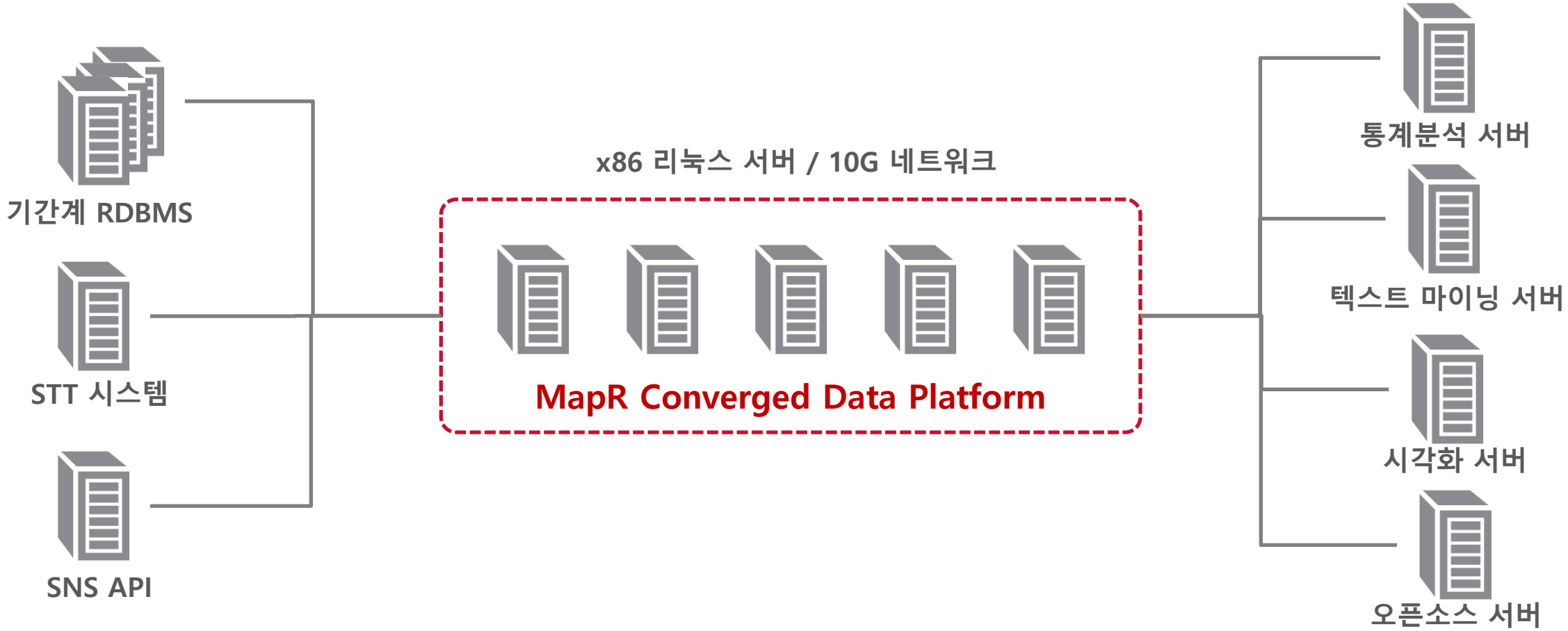
- 머신러닝 분석을 위한 데이터 정제 및 가공등의 전처리 수행
- 데이터 수치화, 정형화 수행
- Apache Hive, Apache Spark 등의 처리 기술을 활용

- 모형 훈련 및 평가가 완료된 분류모델(Classification)에 새로 유입된 데이터 적용
- Apache Spark 또는 Apache Mahout의 머신러닝 기술활용

- 분석 완료된 결과 데이터를 시각화 및 리포트로 생성
- Apache Drill 기술을 활용하여 ANSI SQL 기반으로 결과 데이터 조회

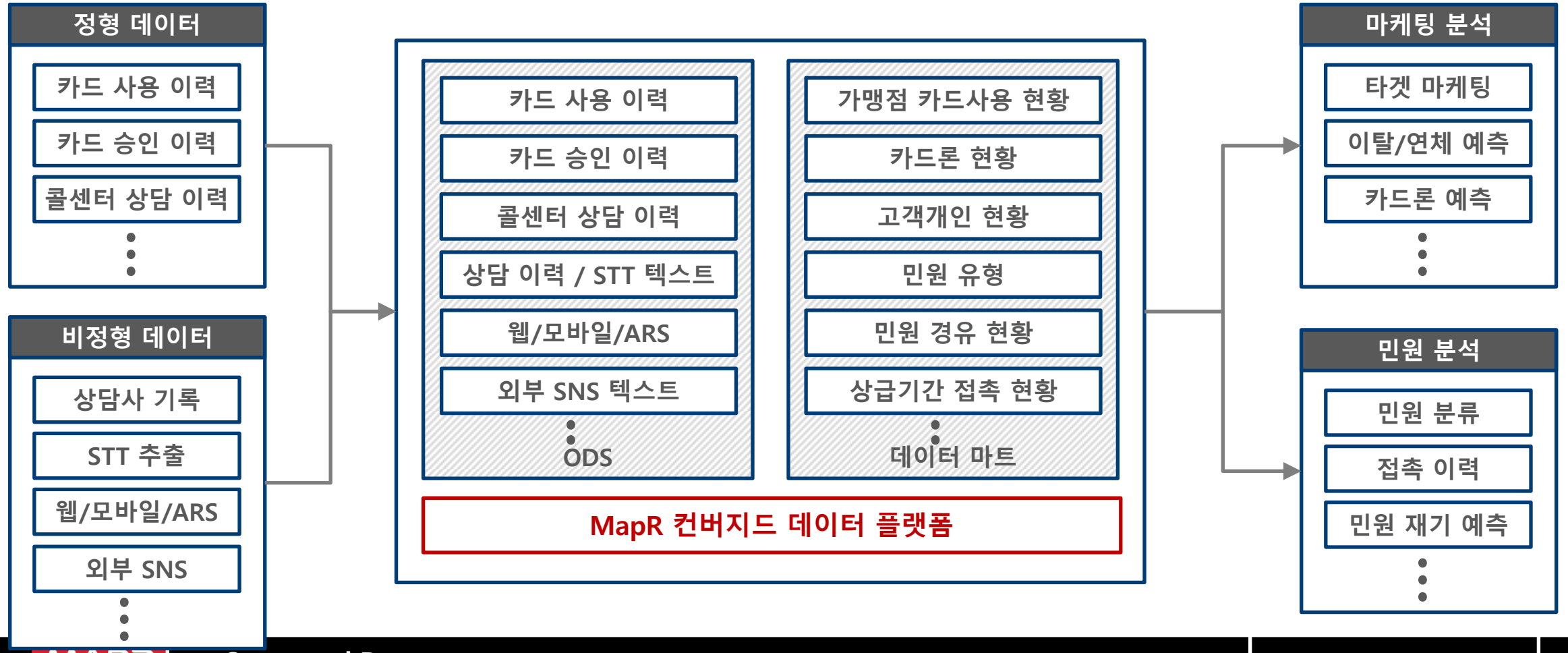
개인화 마케팅 시스템 적용 예

"다양한 소스로 부터 데이터를 수집하여 클러스터로 구성되어 있는 MapR 컨버지드 데이터 플랫폼에 데이터를 수집, 저장, 처리 후 분석 서버가 저장된 데이터를 분석. 분석 서버는 분석 목적에 따라 각기 다른 서버로 구성되어 있으면 신기술을 위한 별도 서버 구성"



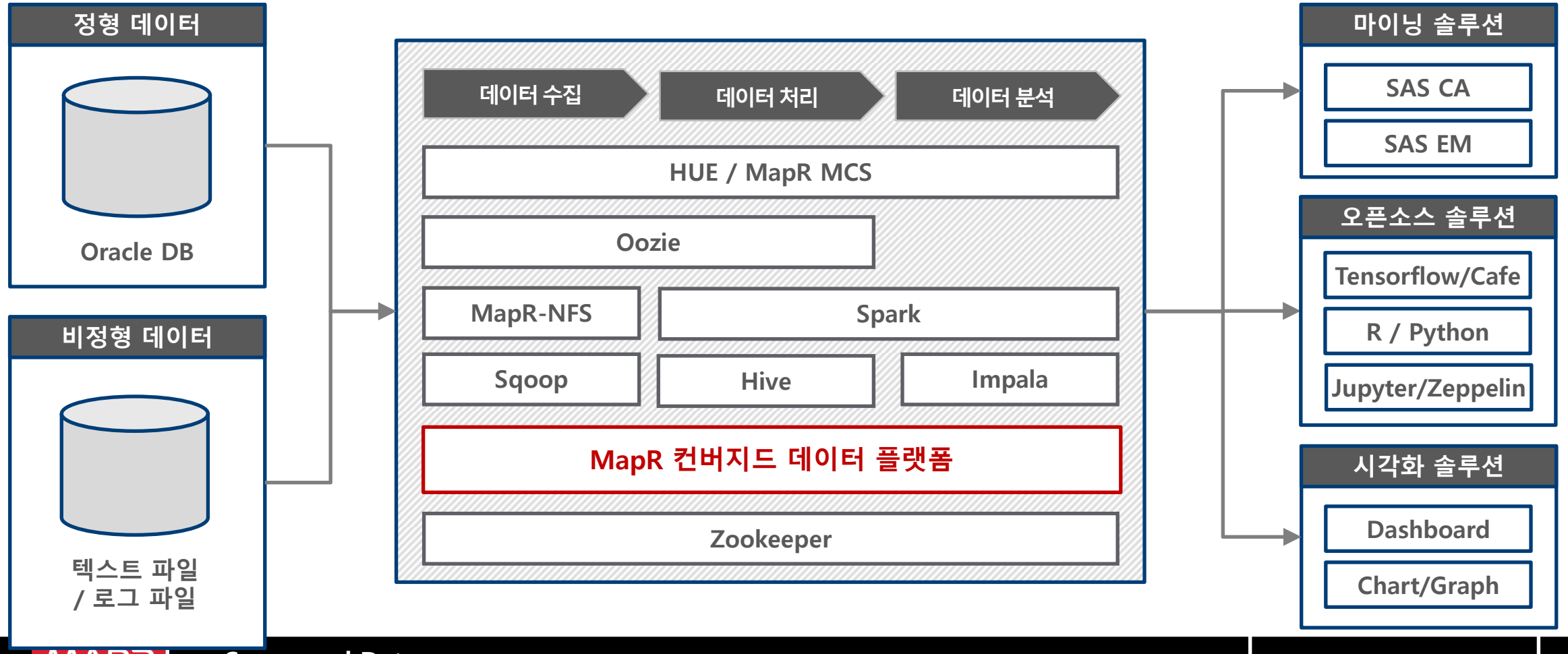
개인화 마케팅 시스템 적용 예

"정형 데이터와 비정형 데이터를 수집후 ODS 영역에 저장후 데이터 처리를 통해 데이터 마트 구성. 구성 된 마트는 기존 분석방법으로 정형 데이터를 분석하고 새로운 분석방법으로 비정형 데이터를 분석 후 모델을 결합하여 보다 정확한 예측 모델을 생성"



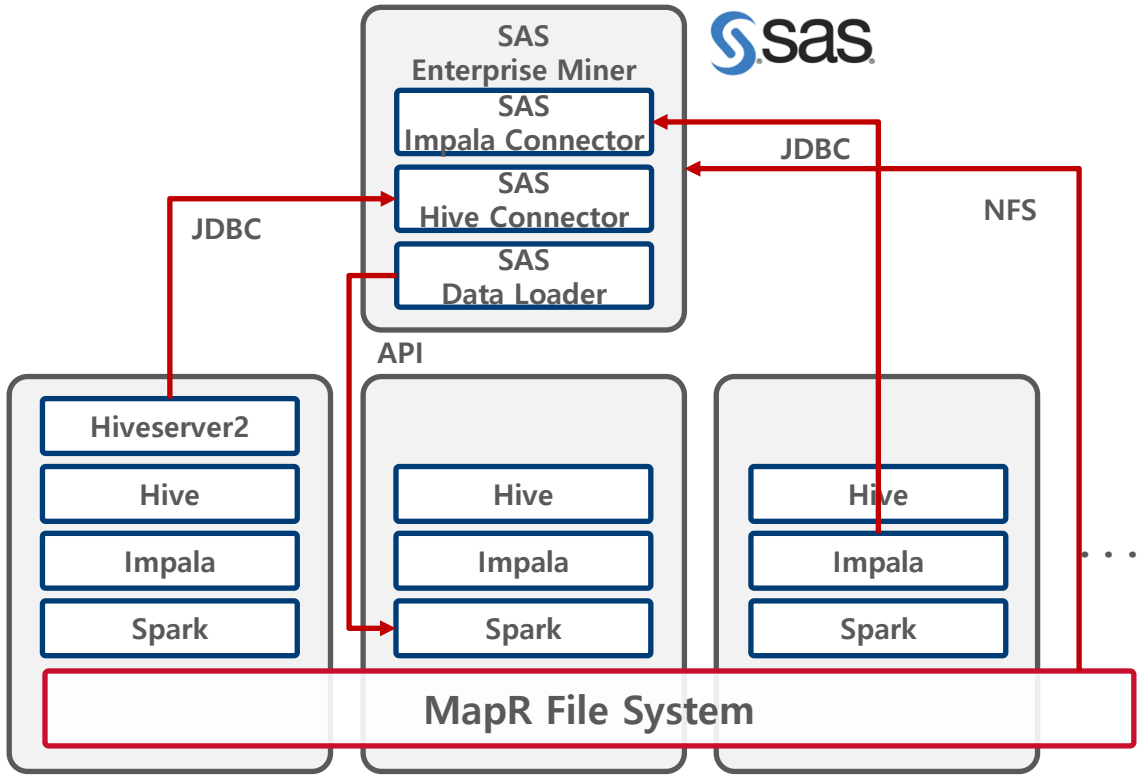
개인화 마케팅 시스템 적용 예

"MapR의 다양한 데이터 처리 에코시스템과 SAS의 분석 솔루션을 연동하여 데이터를 분석. MapR의 고성능을 활용하여 기존 반년 주기의 배치 및 분석 작업을 월주기로, 월 주기의 분석 작업을 일주기로 단축 하였고 영향 변수의 수 도 확장하여 분석"



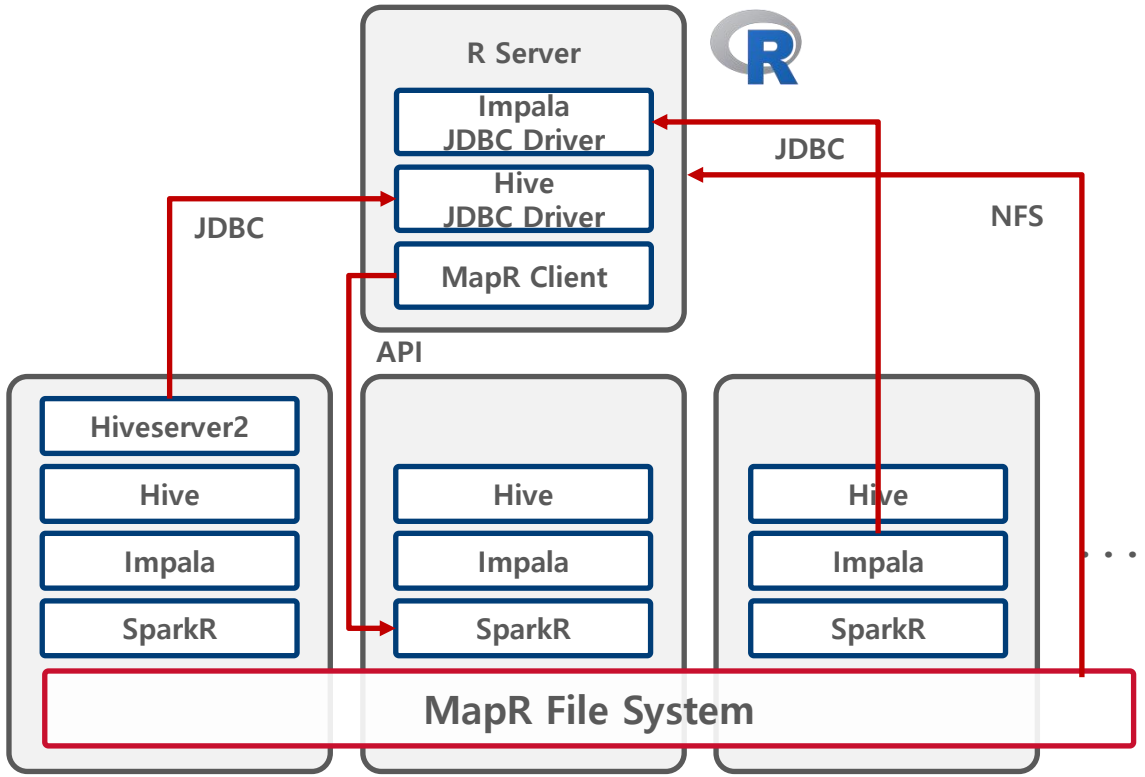
개인화 마케팅 시스템 적용 예

- 전용 커넥터를 통해 Impala 와 Hive를 연계하여 SAS의 데이터 소스로 활용
- SAS Data Loader를 통해 Spark와 연계하여 SAS의 데이터 정재작업을 위임



< MapR & SAS 연계 아키텍처 >

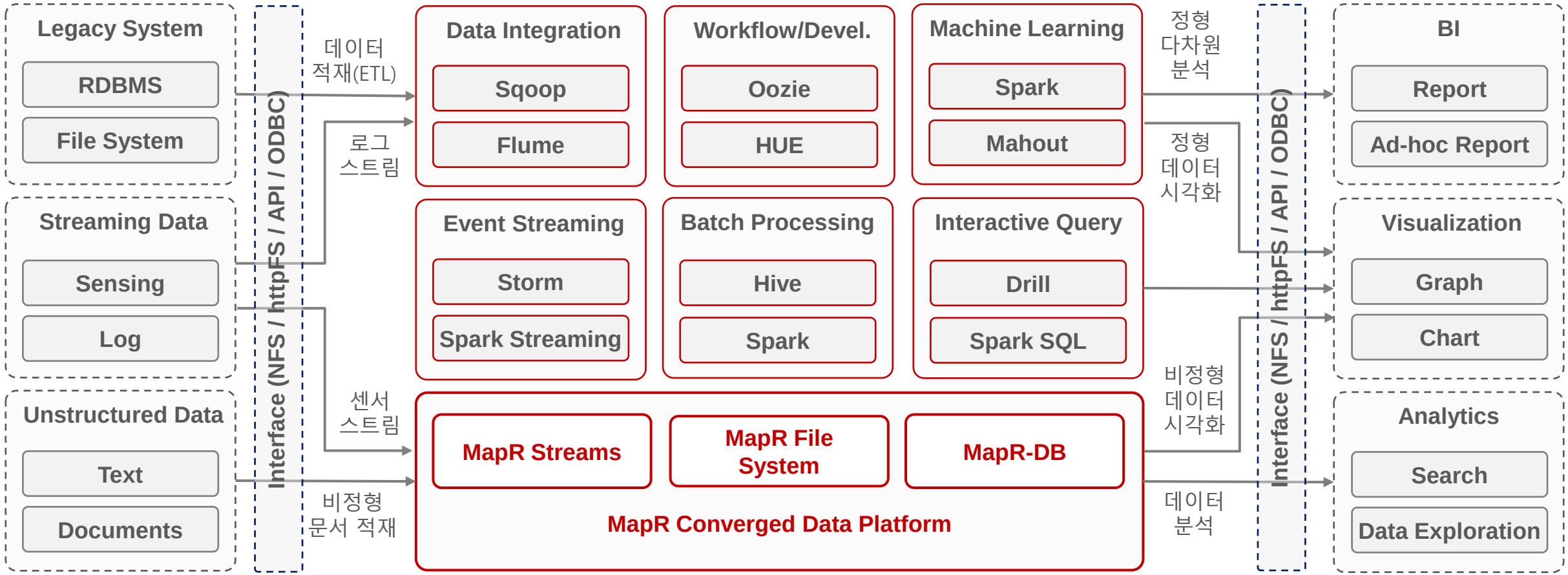
- JDBC 드라이버를 통해 Impala 와 Hive를 연계하여 R의 데이터 소스로 활용하여 R Server에서 분석을 수행
- MapR Client를 통해 SparkR의 API를 호출하여 Spark 클러스터에서 분산처리



< MapR & R 연계 아키텍처 >

구성 아키텍처 예시

하둡기반 플랫폼 구축



구성 아키텍처 예시

하둡을 사용한 하이브리드 DW