

2014 데이터 그랜드 컨퍼런스

빅데이터 시대의 컴플라이언스 데이터 아카이브 전략

(개인 및 거래정보 데이터 아카이브 솔루션)

2014. 9. 30

박희선대표

(주)데이터웍스

주요내용

- 데이터 거버넌스 & 컴플라이언스
- RainStor 솔루션 소개
- RainStor 적용사례
- 요약/결론
- DEMO - 부스



데이터 거버넌스 & 컴플라이언스

- 데이터 거버넌스 & 컴플라이언스
- 국내 개인 및 거래정보 보관 관련 규정
- 전통적인 아카이브 접근 방식

데이터 거버넌스 & 컴플라이언스

➤ 데이터 거버넌스

데이터 거버넌스(Data governance)는 기업의 전사 데이터의 일관성 및 보안성 그리고 데이터 품질 유지를 위한 사람, 프로세스 그리고 기술을 포함하는 정책 또는 전략의 통칭

➤ 필요성

- 데이터 보안성 강화
- 데이터/정보의 품질 강화
- 중복 및 불필요한 데이터 최소화
- 데이터 관련 불필요한 재 작업 최소화
- 정보 생산성 향상을 위한 프로세스 정립

출처: www.ovum.com

데이터 거버넌스 & 컴플라이언스

➤ 컴플라이언스

컴플라이언스(Compliance)는 법규 및 규정에 대한 준수/준법감시/내부통제 등의 의미로 지역별(국가별), 산업별 그리고 형태별로 다양한 규정이 존재함



➤ 다양한 컴플라이언스 타입 존재

- 산업별(By Industry Sector)
- 지역별(By Geography)
- 데이터/정보 타입 별(By Data/Information Type)
- 용도별(By Usage)
- 시대별(Timeframe)

출처: www.ovum.com

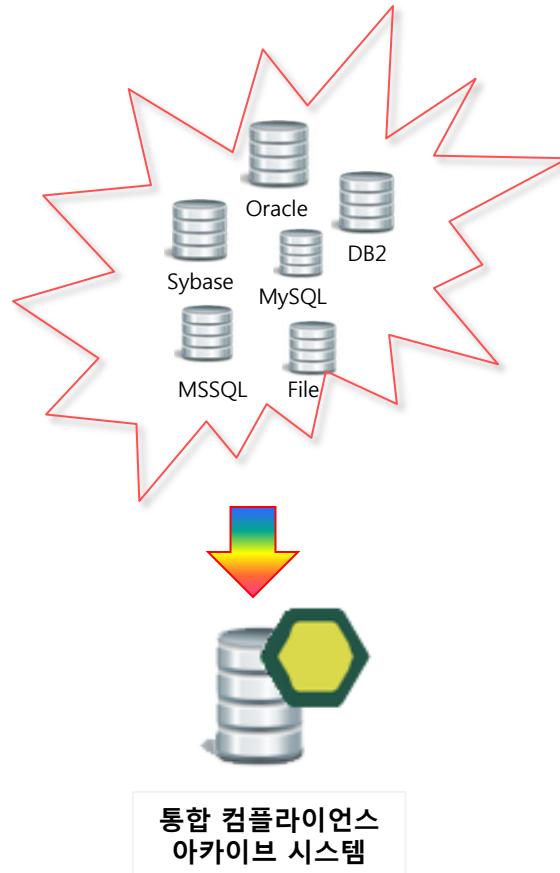
컴플라이언스 사일로 서버 존재

➤ 다양한 이해 당사자/외부 관계자들에 의한 복수의 컴플라이언스 사일로가 만들어짐

*사일로(silo): 격리된 단위 데이터 저장소를 의미함



- 최고 경영층
- 사업본부 경영층
- CIO
- IT부서
- 법무부서
- 재무부서
- 데이터 거버넌스
- 리스크 관리자



한국:

금융거래법 22조 및 금융거래 시행령
신용정보의 이용 및 보호에 관한 법률
개인정보보호법

...

미국:

Securities & Exchange Commission (SEC)
Commodity Futures Trading Commission (CFTC)[157]
Federal Reserve System ("Fed")
Federal Deposit Insurance Corporation (FDIC)
Financial Industry Regulation Authority (FINRA)

...

영국:

Financial Conduct Authority (FCA) – [154]
Prudential Regulation Authority (PRA) – [155]
Panel on Takeovers and Merges (PANNAL) – [156]

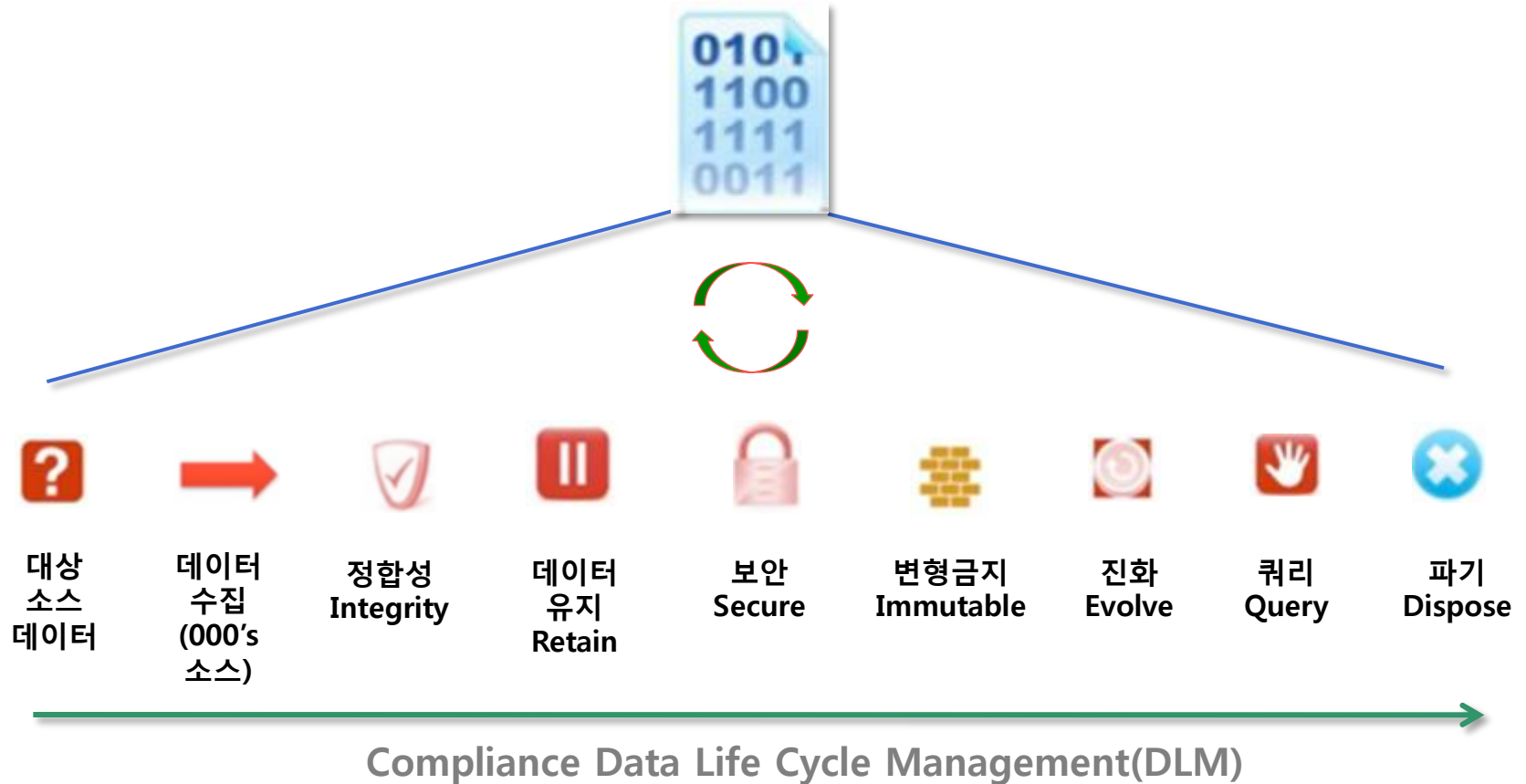
...

“대부분의 국가에서 컴플라이언스 규정준수 요구는 계속 엄격하게 강화되고 있음”

출처: www.ovum.com

컴플라이언스 데이터 라이프 사이클

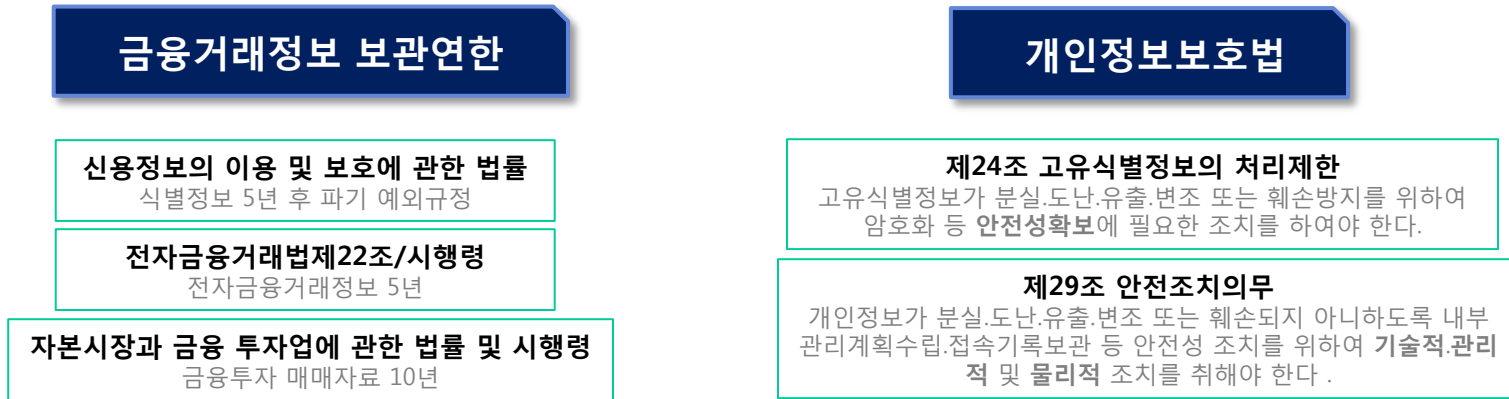
➤ **데이터**는 거버넌스와 컴플라이언스의 중심



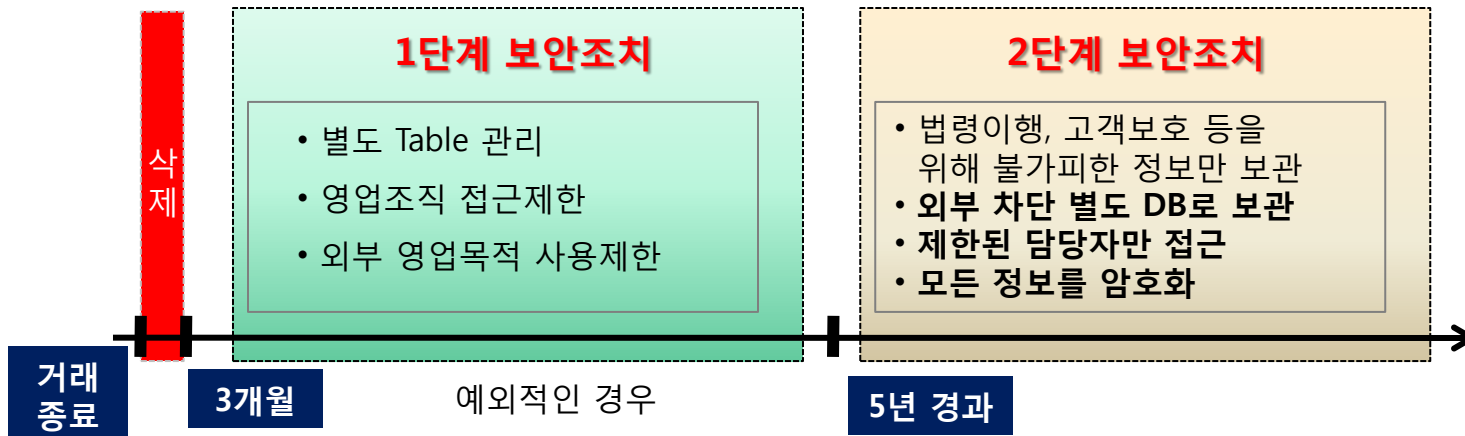
출처: www.ovum.com

국내 개인 및 거래정보 보관 관련 규정

➤ 개인정보 보유와 관련된 법령



➤ 개인정보 파기 및 보관 관련 금융권 공통기준(안)



출처: 금융위원회

전통적인 DB기반의 컴플라이언스 문제

- 전통적인 DB에 의한 접근방식에 의한 컴플라이언스 아카이브는 충분하지 않다.



~~Relational Database Management Systems~~

(e.g. Oracle)

Not Scalable. No Compliance
(Retention / Disposition)

~~Data Warehouse Analytical Systems~~

(e.g. Teradata, Netezza, SybaseIQ)

Costly to Scale. No Compliance.



~~Offline Tape Drivers~~

(e.g. In-house 또는 Outsourced)

Cannot Query. Very Slow.
Error-prone.

~~Legacy / MainFrame Systems~~

(Legacy file formats)

Costly. Difficult to integrate.
Non-standard



RainStor 솔루션 소개

- RainStor 소개
- RainStor 기능
- RainStor 특성
- RainStor와 Hadoop ecosystem
- RainStor 적용 사례

RainStor – Proven Solution

RainStor는 통합 온라인 **컴플라이언스 데이터 아카이브**의 업계 선두 기업입니다.

20 of
World's Largest
Communications
Providers



10 of
World's Biggest
Banks & Financial
Institutions



15 Strategic
Solution &
Technology
Partners

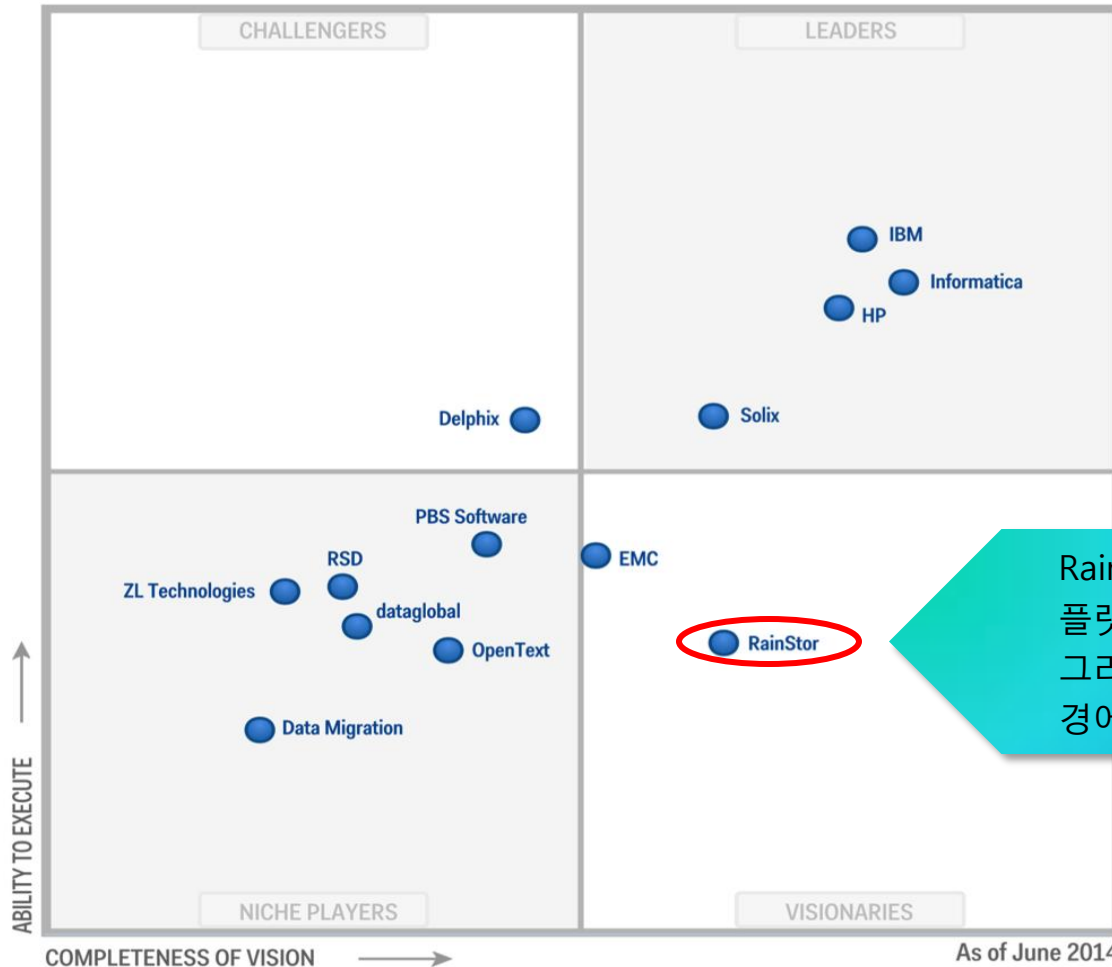


RainStor – 가트너 매직 쿼드런트

구조화된 데이터 아카이브 및 어플리케이션 리타이어먼트 부분

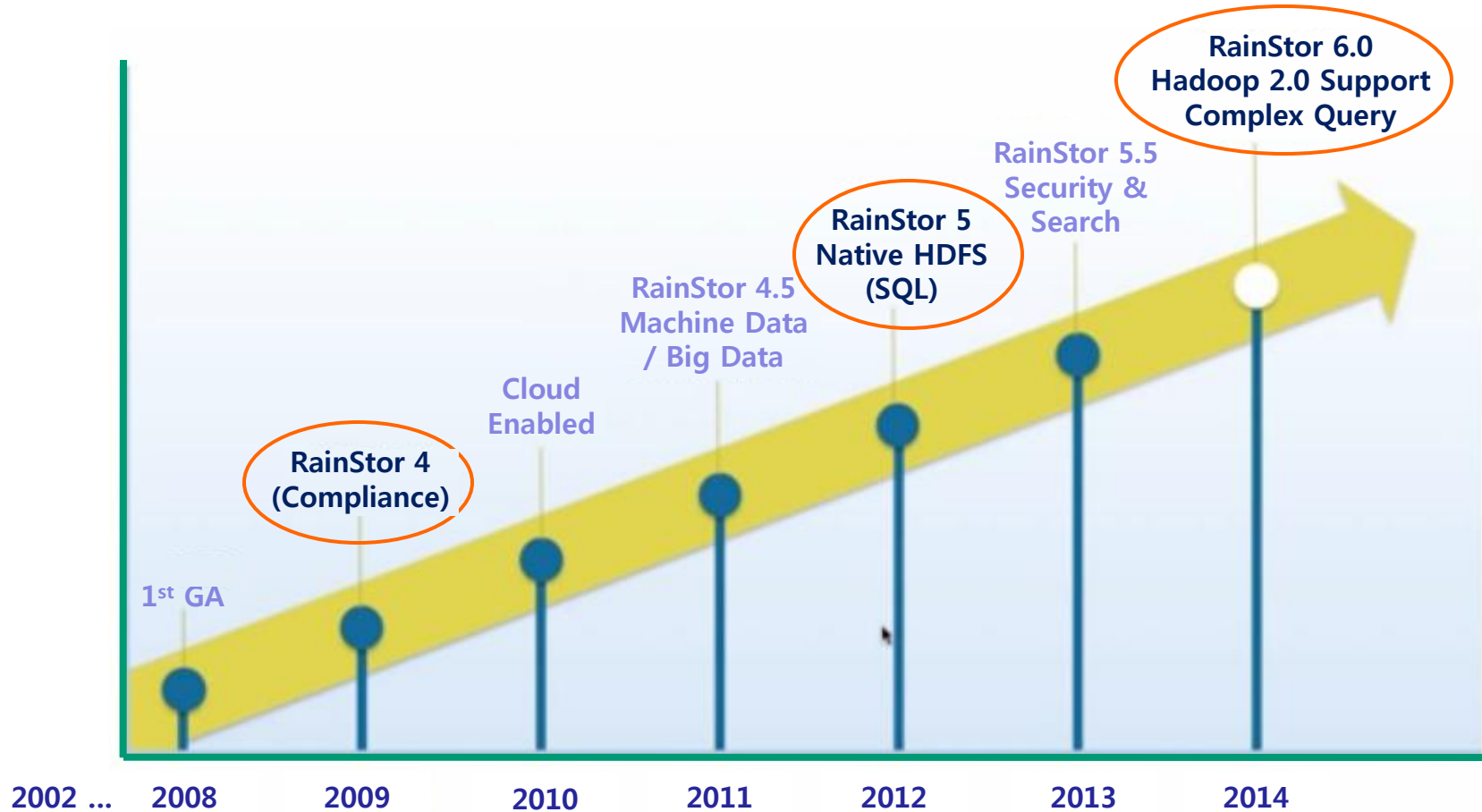
**Gartner
Magic Quadrant**

"데이터 통합은 기업 정보 인프라의 핵심이다. 원활한 데이터 공유를 추구하는 기업들은 빠른 ROI 요구, 통합 패턴, 비용 및 모델의 최적화, 데이터 관리 프로그램을 통한 시너지 효과 창출 등의 측면에서 유연한 기술력을 가진 툴을 선호하기 시작했다." **Gartner**



RainStor는 업계 최고의 데이터 압축 스토리지 플랫폼으로서 데이터웨어하우스 및 하둡의 통합 그리고 대용량 컴플라이언스 데이터 아카이브 환경에 특히 적합하다. **Gartner**

RainStor – Product Milestones



RainStor Solutions

Big Data와 Active Archive 구축 필요성은 동시에 제기되었다.



Network



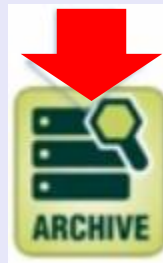
APPs



EDW



Tape



통합 데이터 아카이브

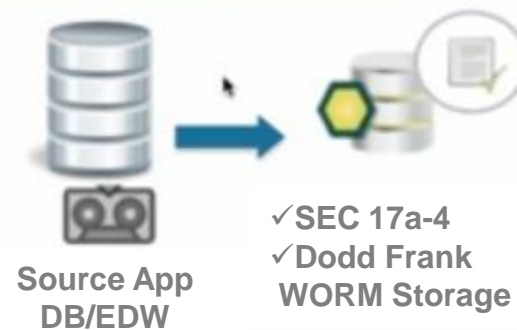
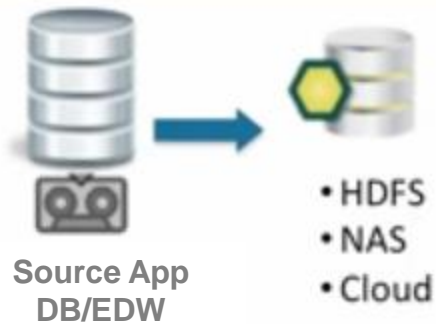
RainStor 통합 컴플라이언스 솔루션은 시계열성 또는 빅데이터에서 비즈니스 가치를 찾아내고 컴플라이언스 규정의 준수를 가능하게 한다.

RainStor Solutions

Solutions

Analytical Archive

Compliance Archive



Active Archive – 비즈니스 및 컴플라이언스 담당자들의 요구사항에 부합

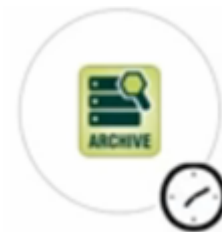
- 데이터 오프로드 : File/DB/EDW 또는 Tape
- 보안 : 빌트-인 기능 (LDAP / Kerberos / Encryption / Masking)
- 거버넌스 : 원본 데이터 유지 기능 / Immutable / Audit Trail
- 고성능 쿼리 : 상용 쿼리 툴, SQL, Hadoop 분석지원

거버넌스 & 컴플라이언스 데이터 보유 정책

거버넌스 :

- 위.변조 방지
- 접근기록 관리
- 레코드 단위 삭제
- 룰(rule) 기반의 데이터 파기 기능
- 컴플라이언스 법규 및 규정 지원
- 데이터 보호(Hadoop, DAS, NAS)

“지금부터 6개월간 해당 데이터를 보유하다가 자동으로 파기하시오.”



RainStor 는 모두 만족

컴플라이언스 :

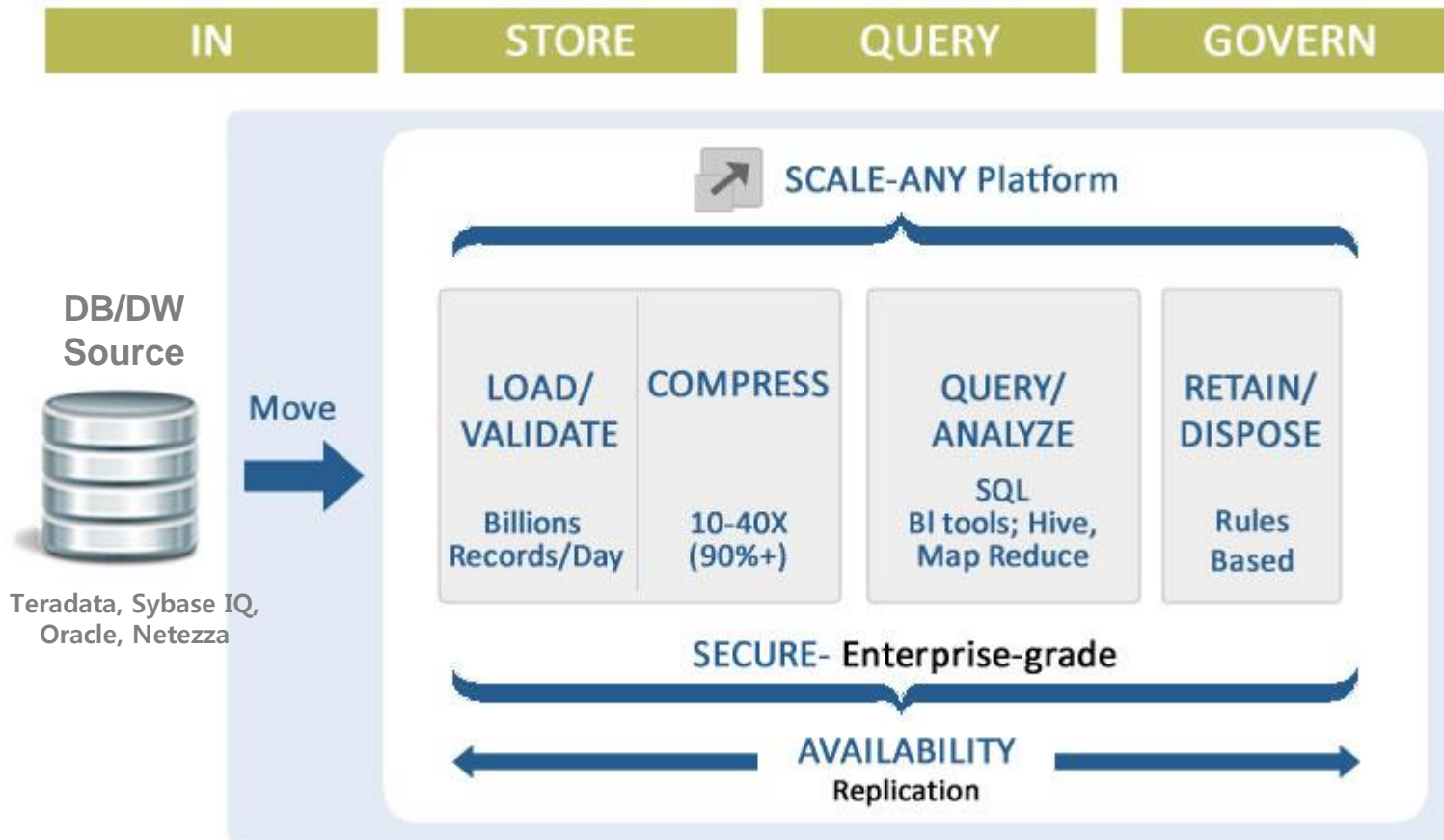
- EMC WORM 스토리지 / SEC 규정 준수
- 룰(rule) 기반의 데이터 보유정책 기능
- 데이터 값에 의한 데이터 보유기간 설정

```
SELECT      data(month, 6 transaction_ts)
            rs_retention_date
FROM        transactions;
```

“트랜잭션이 처리된 날짜를 기준으로 해당 레코드를 6개월간만 보유했다가 파기하시오.”

RainStor 기능 소개

RainStor는 Data 수집부터 거버넌스까지 전체 라이프 사이클을 관리한다.

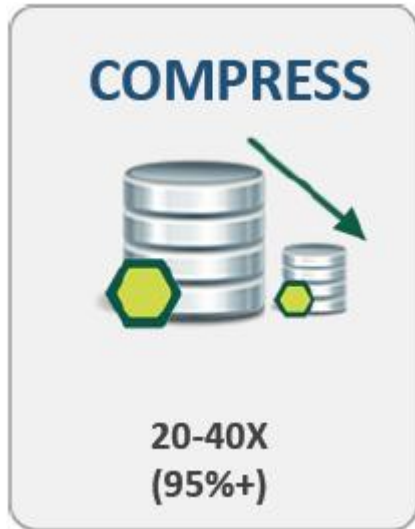


RainStor 특징 – Data Load



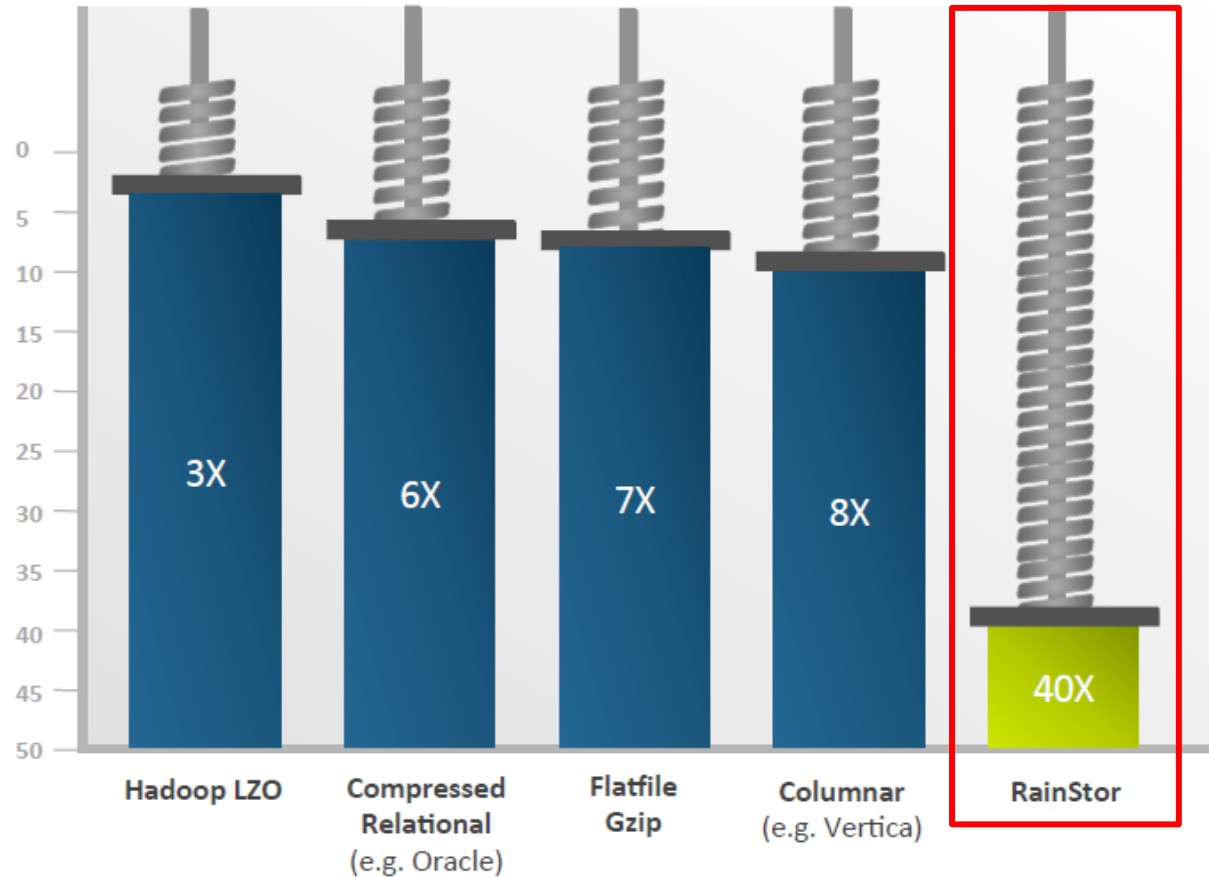
- 병렬로드(core, node)가 기본
- 파일타입: CSV, BCP, XML
- 기존 툴 적용: ESB, ETL, Utilities
- 데이터타입: Machine data/log & DW
- 로드방식: Bulk, continuous, landing (auto data) area
- 국내 SQL Canvas ILM과 연동 – 자동화 가능

RainStor 특징 – Data Compress



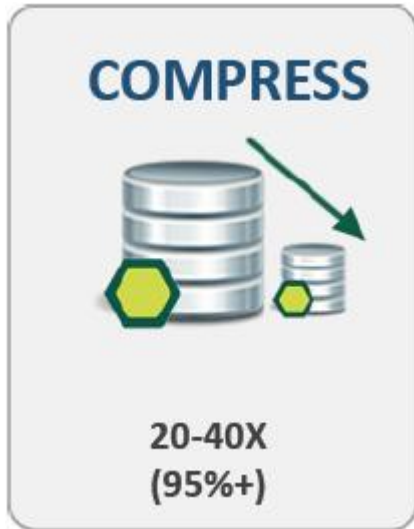
압축레벨

1. Field(Value)
 2. Pattern
 3. Algorithmic
 4. Byte
- ↓



Source: Ratio vs. Raw –RainStor Benchmarks using customer data(2012)

RainStor 특징 – Data Compress

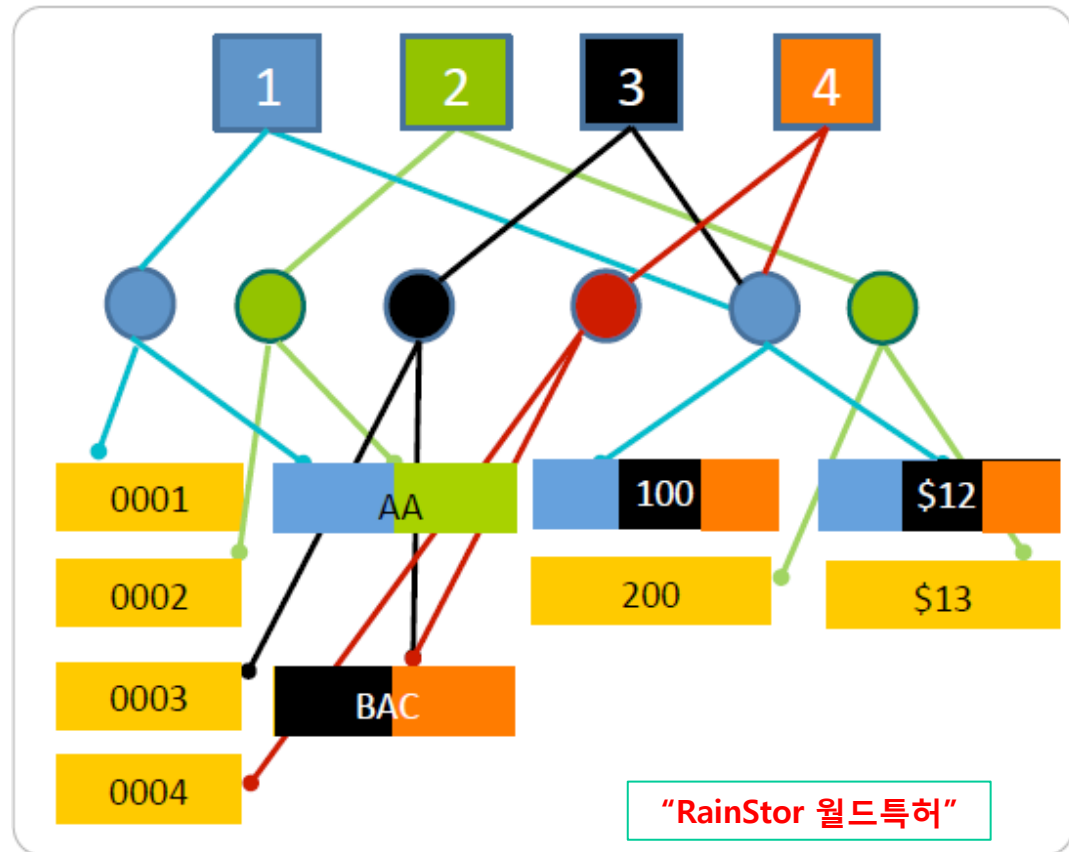


압축레벨

1. Field(Value)
2. Pattern
3. Algorithmic
4. Byte



Value & Pattern De-Duplication 기술



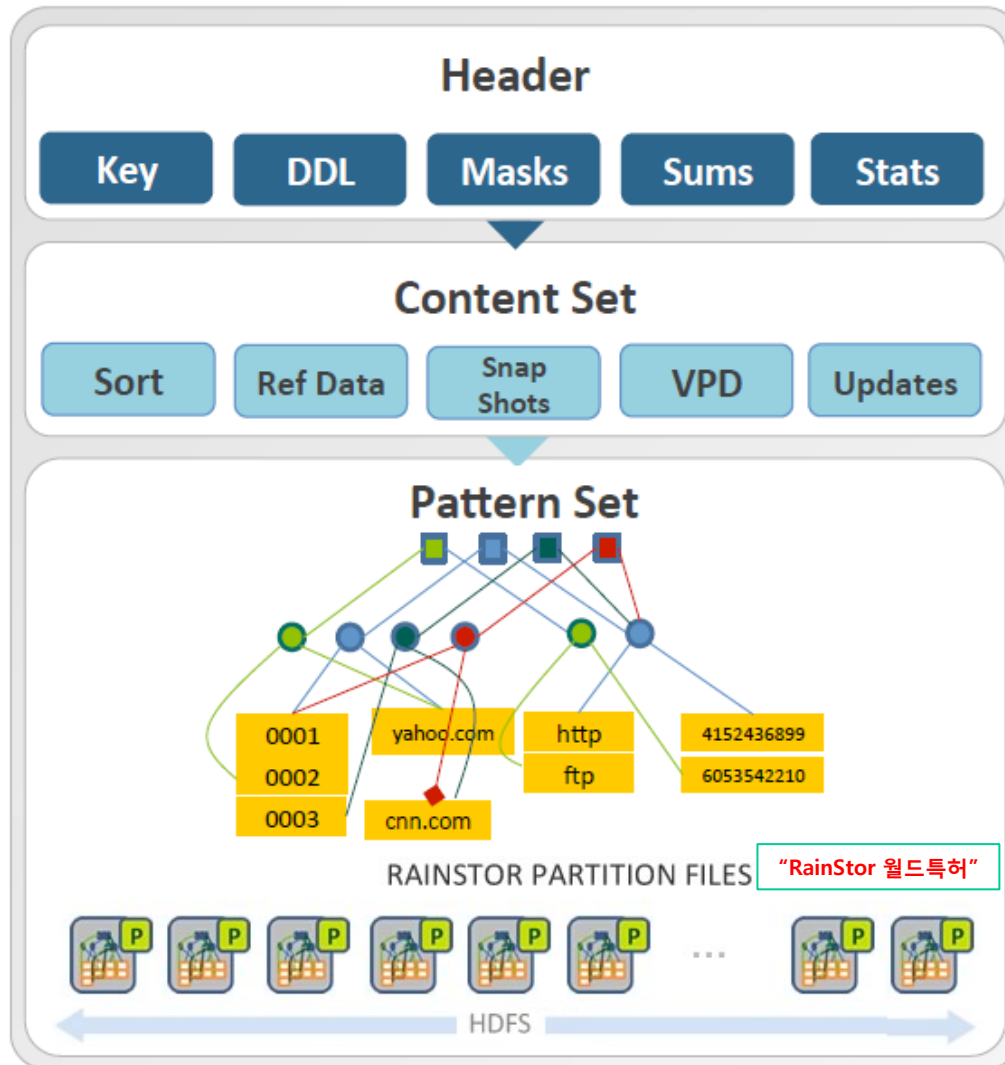
Example of Stock Trades

RainStor 특징 – Data Store

STORE

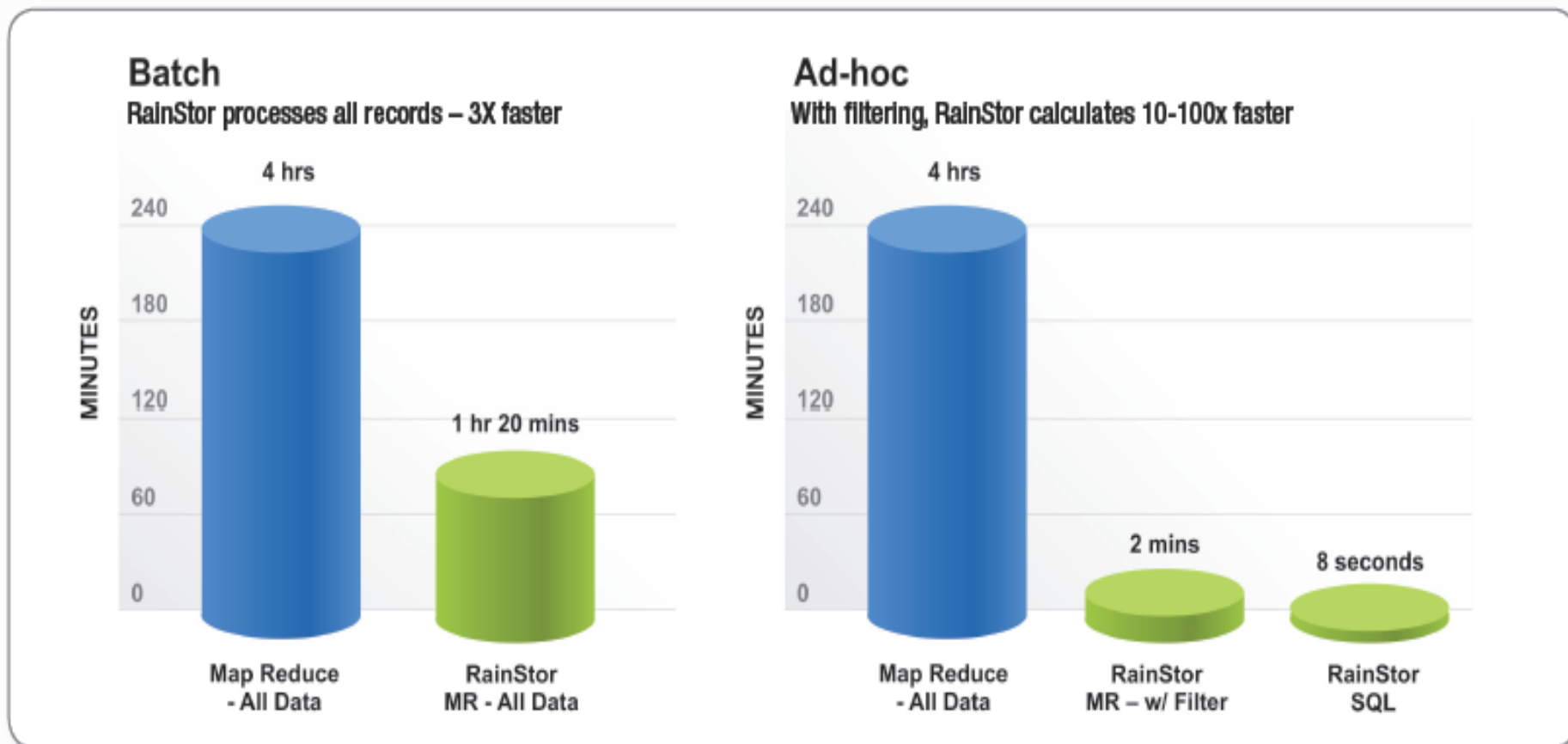


20-40X
(95%+)



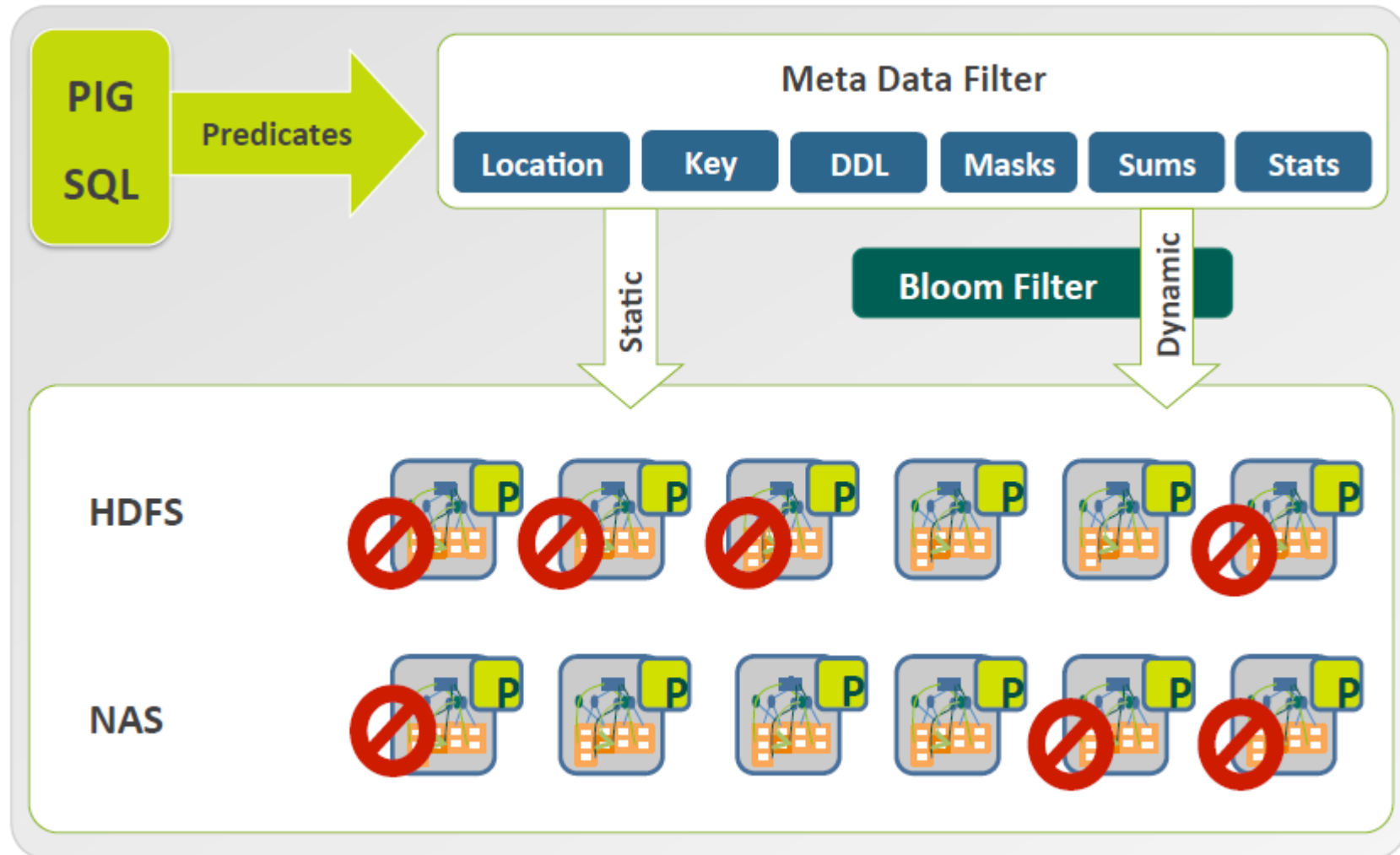
- 메타 데이터 기반의 액세스
- Encapsulated
- 고성능 압축 오브젝트
- 스키마와 독립적
- 다중 구조화된 데이터 지원
- 파티션 형태로 저장

RainStor 특징 – Query Speed



Query Test: 하루 평균 15억 건이 거래되는 symbol 데이터로 테스트 진행 (미국의 주요 금융 서비스 데이터, 2011)

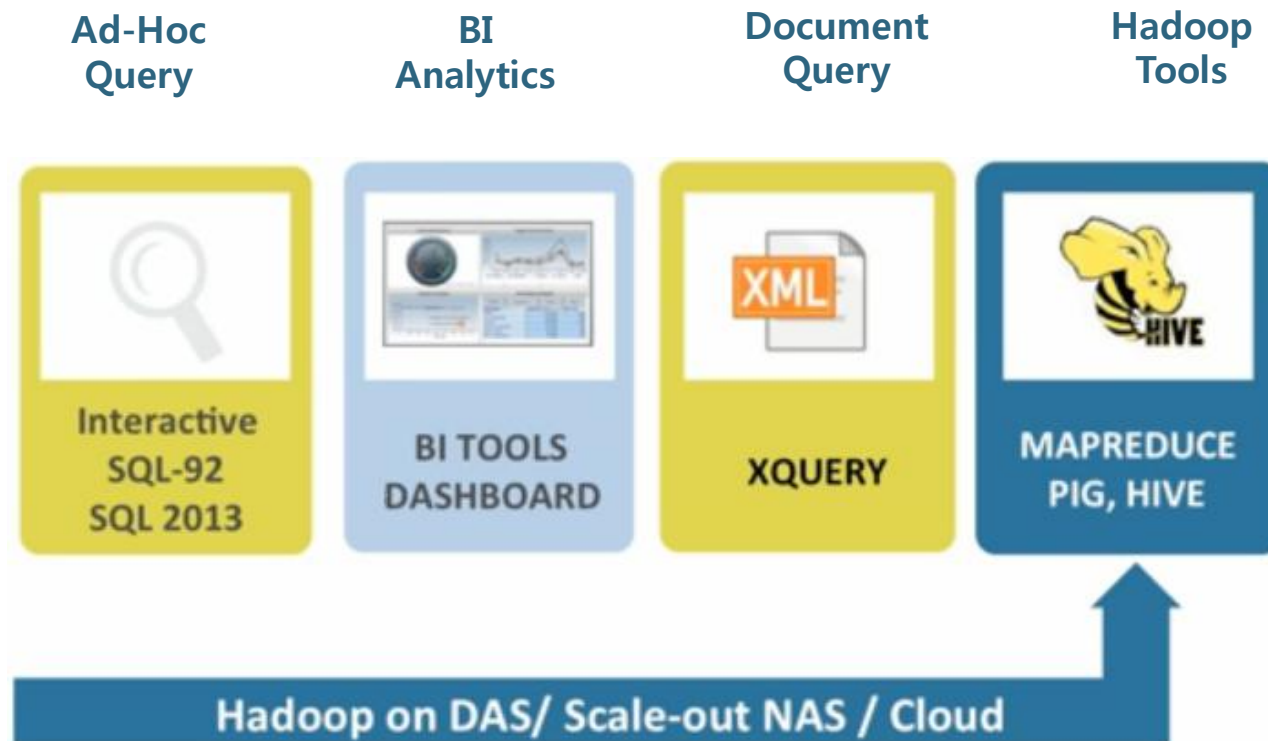
RainStor 특징 – Query Speed



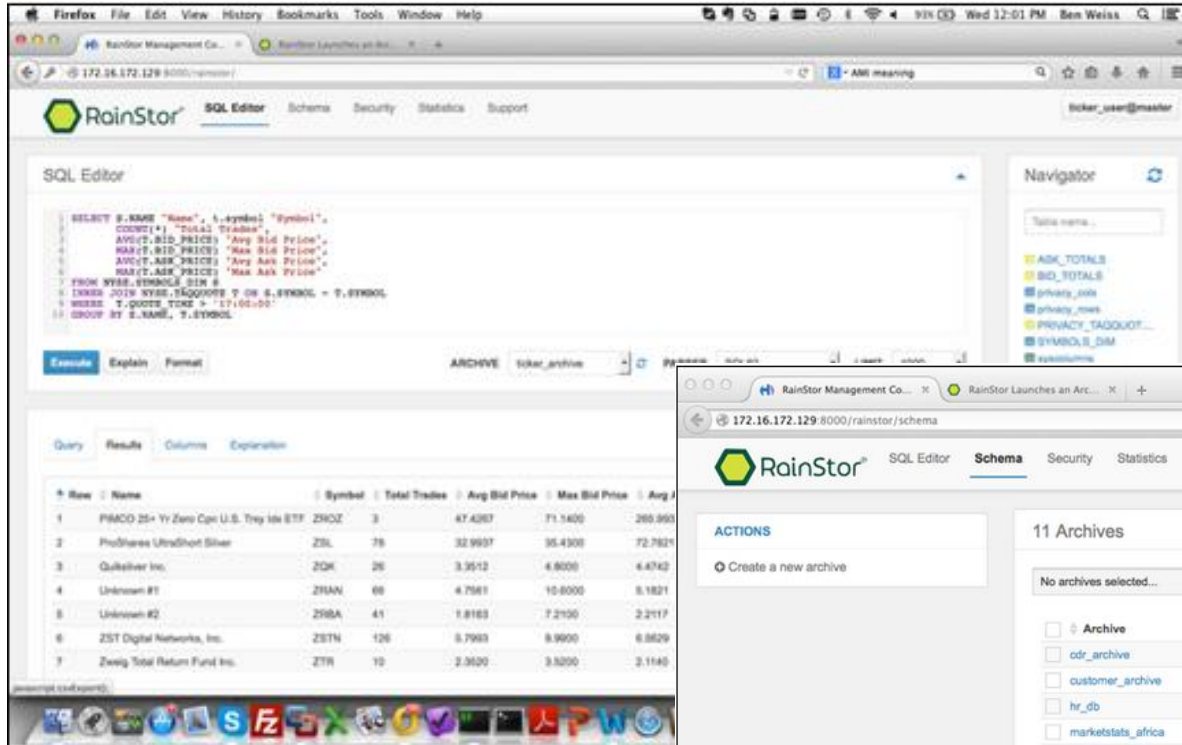
RainStor 특징 – *Data Visualization*

ANALYZE

- 고성능 Text 검색 지원
- ANSI SQL-92, 2003
- 표준 ODBC, JDBC 지원
- 상용 BI Tool 연동 지원
- 하둡 PIG 및 HIVE 쿼리 지원
- Cross-Schema Query
- Point-In-time hist. view

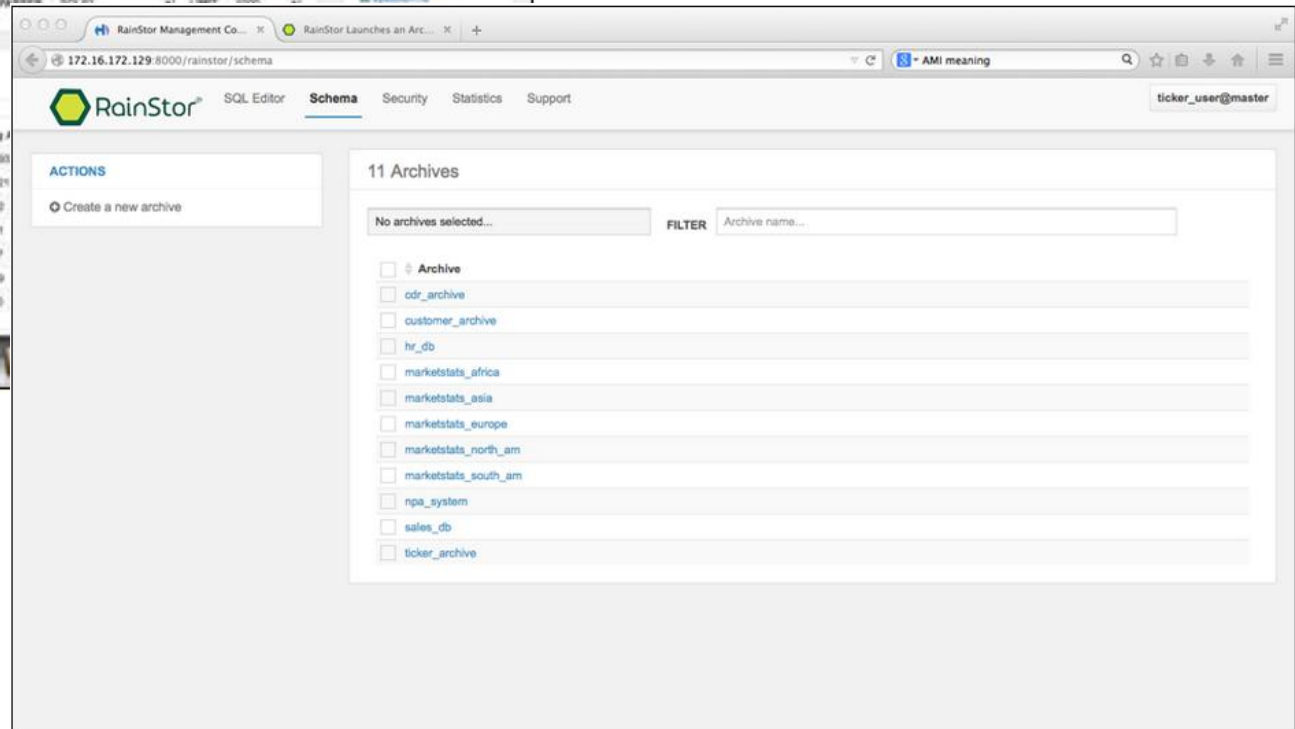


RainStor 특징 - 쿼리 및 관리 툴



RainStor Archive App HUE 기반의 쿼리 및 관리/모니터링 툴

*HUE(Hadoop User Experience) : Hadoop Data 분석을 위한 Web 인터페이스



RainStor 특징 – *Security*

Perfect

- 불변성 데이터 보관 (WORM)
- Immutable 스토리지에 사용이 가능한 유일한 DB
- 데이터 변경이력관리
- 접근내역 관리

Adapt

- 크로스 스키마 지원
- 데이터 임포트 시점별 쿼리 지원
- 스키마 독립적

“RainStor 월드특허”

Retain

- 정책기반의 데이터 유지 및 파기(파티션 또는 레코드 단위)
- 컴플라이언스 룰(rule) 기반의 데이터 유지 및 분석
- 레코드 별 Legal-hold 부여
- Encryption 및 Column Masking 지원
- User 및 Role 관리



RainStor 특징 – *Security*



PRIVACY

- Data Encryption
- Data Masking
- Views

INTEGRITY

- Tamper-proofing
- Audit Trail
- Record-level Delete
- Data Disposition

TRUST

- Kerberos Authentication
- Authorization
- LDAP/Active Directory
- Linux PAM Support
- WORM Storage Support

RainStor 특징 - 운영모델

Single Node

All Storages



Single Node



All Platforms

- Online Archiving & ILM/DLM 데이터 처리
- DISK 구성 : SAN, NFS, CAS, DAS 기반

RDR

Multi Node

SAN, NFS, NAS



Cluster Multi Node



Linux

Hadoop Multi Node

HDFS



Hadoop Multi Node

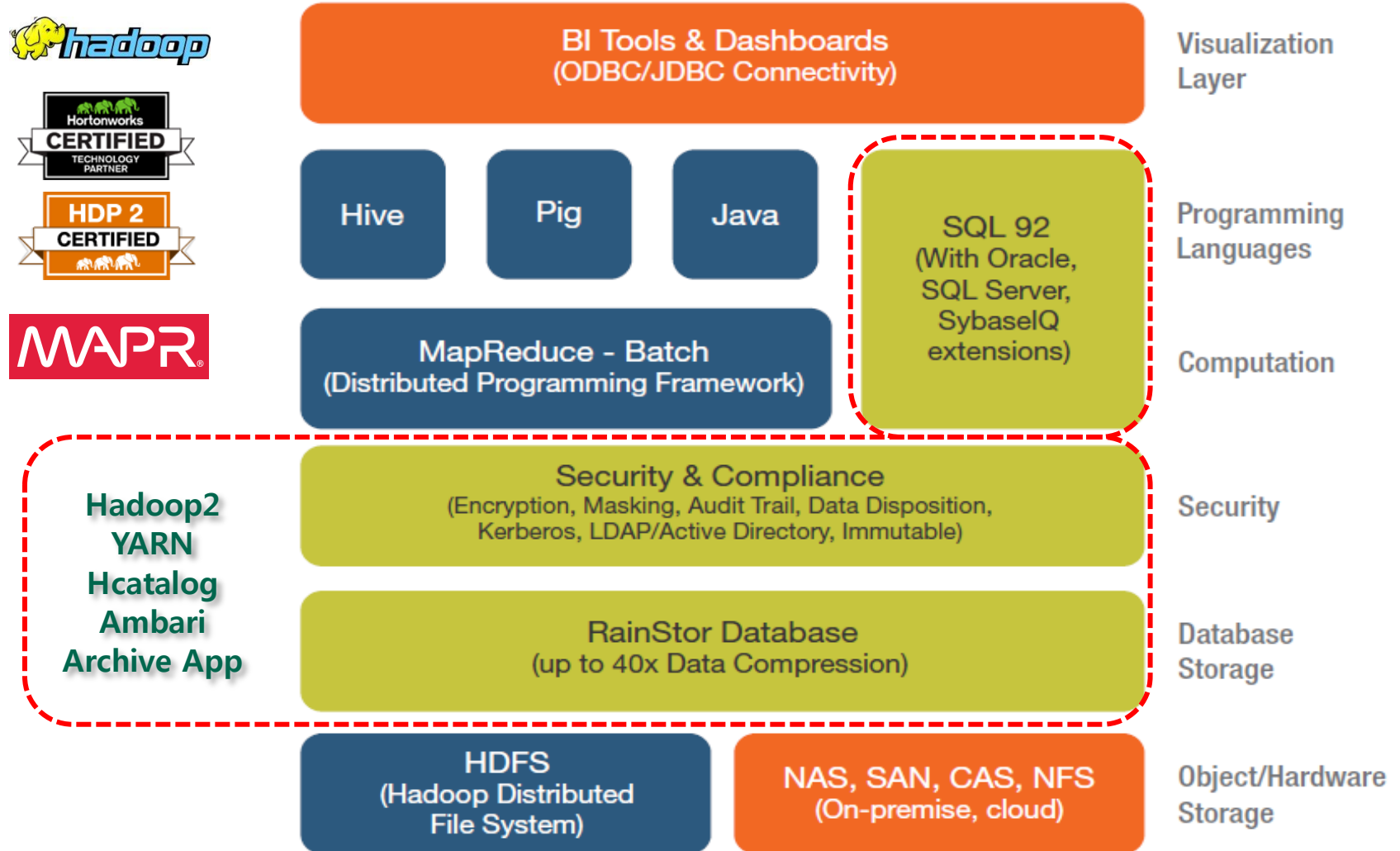


Linux

- 고성능 병렬 빅데이터 플랫폼
- DISK 구성 : Hadoop HDFS 기반

RDA

RainStor + Hadoop ecosystem



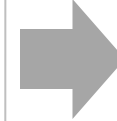
RainStor 기능(요약)

구분	RainStor 기능
<p>✓ Load</p>	<ul style="list-style-type: none"> ▪ 고성능 병렬 bcp, xml 데이터 적재(load) 및 추출(export) ▪ 자동적인 데이터 적재를 위한 랜딩(Landing area) 기능 ▪ 상용 ETL 도구 또는 유사한 솔루션 적용 가능 ▪ TERADATA와 native 연동 지원(RainStor FastConnect, FastForward) ▪ 국산 데이터 이관 솔루션인 SQL Canvas(Oracle, Sybase, DB2)와 연동지원
<p>✓ Query</p>	<ul style="list-style-type: none"> ▪ ODBC/JDBC 와 같은 표준 인터페이스 방식 지원 ▪ 상용 클라이언트 BI/쿼리 툴 및 RainStor 자체 쿼리 도구 사용 ▪ CPU/core 간, node 간 고성능 병렬 쿼리 지원 ▪ ANSI-SQL 92, Oracle, MSSQL 쿼리 및 SQL온하둑: Pig, Hive 분석 제공 ▪ Schema Evolution(Cross Schema) 및 Point-in-time 쿼리 지원
<p>✓ Scale</p>	<ul style="list-style-type: none"> ▪ 단일 node에서 클러스터 또는 하둑기반의 대규모 node까지 유연한 확장 지원 ▪ 대용량 데이터 증가/보유 시에도 동일한 쿼리 및 로드 성능 유지 ▪ 상용 서버 또는 저가형 일반 모든 서버 및 OS에서 운용가능 ▪ 다양한 하둑 빅데이터 에코 솔루션(Apache, Cloudera 등)들과 유연한 연동 지원
<p>✓ Store</p>	<ul style="list-style-type: none"> ▪ 최고의 압축저장 알고리즘 탑재(데이터 유형 별 40배까지 압축) ▪ 저가형 디스크 및 일반 범용 서버에서 운영 지원 ▪ 다양한 저장 방식 지원(예, NFS, SAN, DAS, CAS, NAS, Hadoop HDFS) ▪ 하둑 HDFS 적용시 데이터 리플리카 특성에 의한 데이터 폭발 방지
<p>✓ Security</p>	<ul style="list-style-type: none"> ▪ 상용 DBMS 정도 수준의 User 및 Role 관리 기능 ▪ Data Encryption(암호화) 및 민감한 칼럼에 대한 Masking 기능 ▪ 모든 액세스 및 접근 기록 유지 기능 ▪ 변경이 불가능한 CAS(Content Addressable Storage)을 지원하는 유일한 DBMS ▪ 데이터 수정이 원천적으로 불가능(예, Update 미 지원)
<p>✓ Admin</p>	<ul style="list-style-type: none"> ▪ 간단한 환경설정 및 자동화된 운용환경(1주일 이내 업무적용가능) ▪ 데이터 레코드 별 보유기간 사전 정의 및 자동 폐기 등 관리비용 절약 ▪ 대용량 데이터에 대한 full/인크리멘탈 방식의 백업 지원 ▪ 클러스터 및 하둑 HDFS 인프라 적용으로 자동 Fail Over 구성 ▪ 데이터 압축 및 고성능 쿼리지원으로 하둑 구성시 H/W node 수 감축 가능

Case Study – 월 스트리트 투자 은행



EMC
WORM Storage



온라인 트랜잭션 시계
열성 데이터
데이터 폭발

Hadoop 사용
~60TB 시계열성 데
이터 로드

40x 압축
고성능 쿼리
Hadoop 인프라 운용

- 10,000 사용자
- 10여 년 운용 경험

- Hadoop 클러스터 (Cloudera)/Dell 서버 사용
- 키 아카이브 기능에 대한 비즈니스 부서에서 검증
- (Sales, Legal, Finance, Marketing, Customer Server)

- 온라인으로 데이터 보유
- 쿼리 가능
- Hadoop을 실 비즈니스에 사용

Good ROI

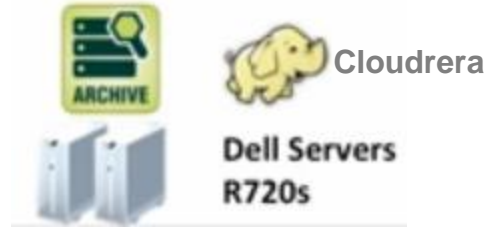
정책기반의 데이터
아카이브

고성능, 예측 가능
한 확장성, 저비용

Case Study - 저비용 EDW 운용



✓FastConnect
✓FastForward



고객 애플리케이션
Oracle DB
IBM DW's

40+개 APPs에서
~2TB/1일 데이터를
Archive로 로드

최적의 압축
고성능 쿼리
WORM

- 40+ 다양한 어플리케이션 시스템
- 보안이 필요한 민감한 데이터의 보유

- 고성능 병렬 로드
- 효율적인 Archive 시스템 관리를 위하여 GUI 기반의 모니터링 및 관리 툴 제작

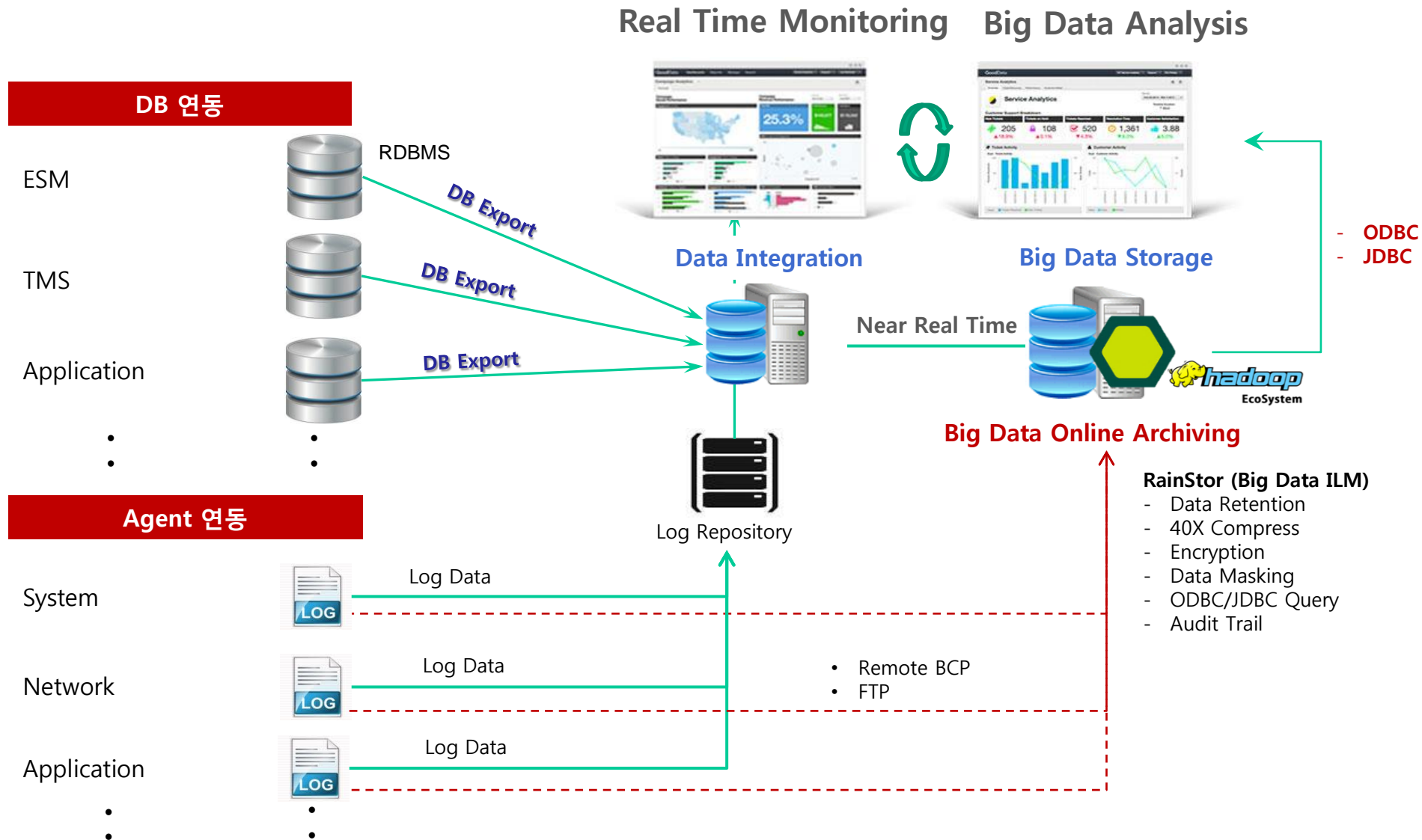
- SEC 컴플라이언스
- 벌금 및 과징금 제외
- EMC WORM 스토리지 사용

SQL 쿼리 지원
자동 Data Load
Connector 지원

고성능 응답시간
(DW에 필적한 만큼)

고성능 압축 및
저비용

Case Study - 통합 로그 관리



컴플라이언스 아카이브 구축 전략(결론)

- **통합된 컴플라이언스 아카이브 DB 구축**
 - 전사적으로 유사한 시스템을 통합하여 관리 효율 및 보안성 강화
 - 장기적으로 통합된 컴플라이언스 기반 인프라 준비 필요
- **컴플라이언스 아카이브에 특화 된 솔루션 도입**
 - 보안성 강화 및 비용(H/W, S/W, 인건비 등) 절감을 위한 전략 필요
 - **룰(rule)** 기반의 데이터 **보유** 및 **파기** 기능 등 - 관리의 유연성 제공
 - 태생적 접속기록(Audit Trail) 유지 필요
 - WORM 스토리지 적용 가능 등, 불변의 데이터 보유 강화
- **감독기관에서 요구하는 규정 준수 노력**
 - 컴플라이언스 규정을 준수하려는 노력(투자)이 모두 보존됨
 - 최근 정보유출 판례에서 물리적 실 투자에 관한 노력이 인정됨
- **신기술 기반의 솔루션 도입 검토**
 - 하둡 등, 빅데이터 기반의 인프라 활용 필요 - 비용절감
 - 저비용의 일반 서버, OS 그리고 디스크 시스템 적용가능
 - 전통적인 DBMS는 요구사항이 다른 다양한 업무를 지원하기 어려움

Q/A

&

Demo(전시부스)

감사합니다.