Big Data Analytics in Action

임상배
Principal Sales Consultant
DBU, Oracle Korea

10/08/2014

ORACLE

Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Program Agenda

1. 빅데이터 분석을 위한 아키텍쳐
2. 최근 기업들의 빅데이터 분석 아키텍쳐 트랜드
3. Oracle Advanced Analytics – Best Practices
4. Big Data 분석을 위한 Core Technology
5. Big Data 분석 활용 사례

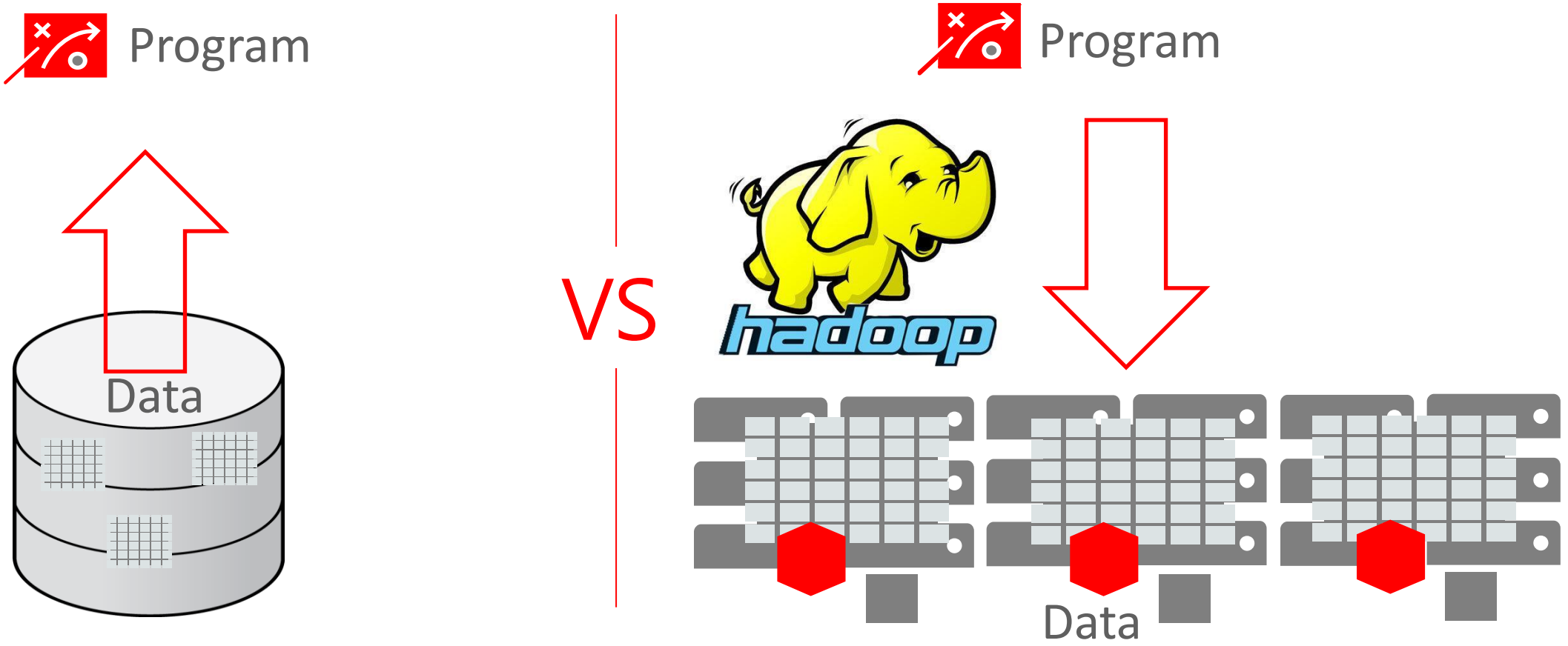
빅데이터 분석을 위한 아키텍쳐

분석의 기반이 되는 아키텍쳐

ORACLE

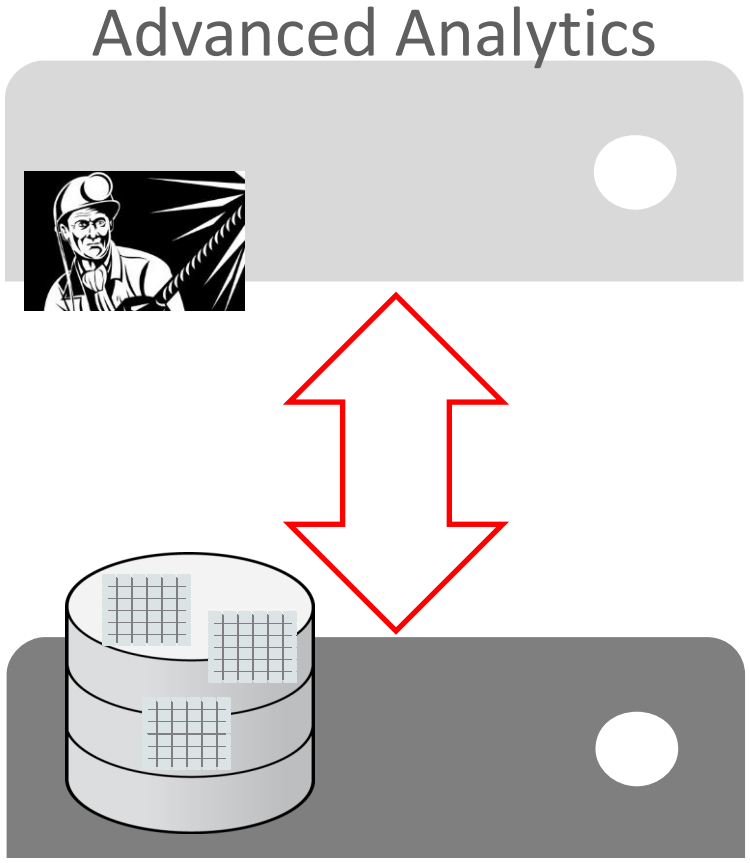
Big Data 환경, 가장 큰 기술적 변화(저장/처리)

Data 전송 구조에서 Program 전송 구조로 변경

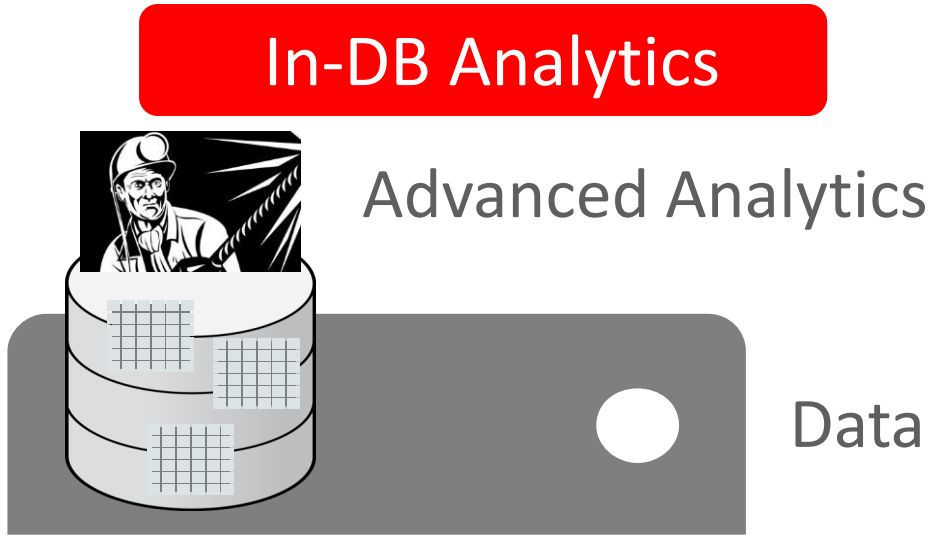


Big Data 환경, 가장 큰 기술적 변화(고급분석)

데이터 이동 없이, 데이터와 분석을 하나의 환경에서



VS



- 데이터 이동 없음
- 데이터 중복 제거
- 높은 보안성 및 경제성

ORACLE

Big Data 환경, 가장 큰 기술적 변화

생산과 활용의 차이를 줄이고

요약하면

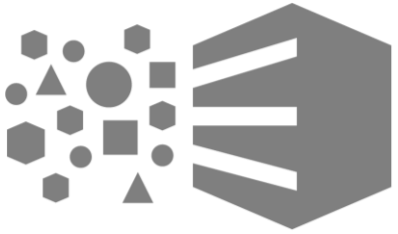
"데이터와 분석의 거리를 단축하는 방향으로 기술 혁신"

ORACLE

빅데이터 분석 시 Data Scientist역활

Data Scientist는 뭘 할 수 있는 사람?

Hadoop



Relational

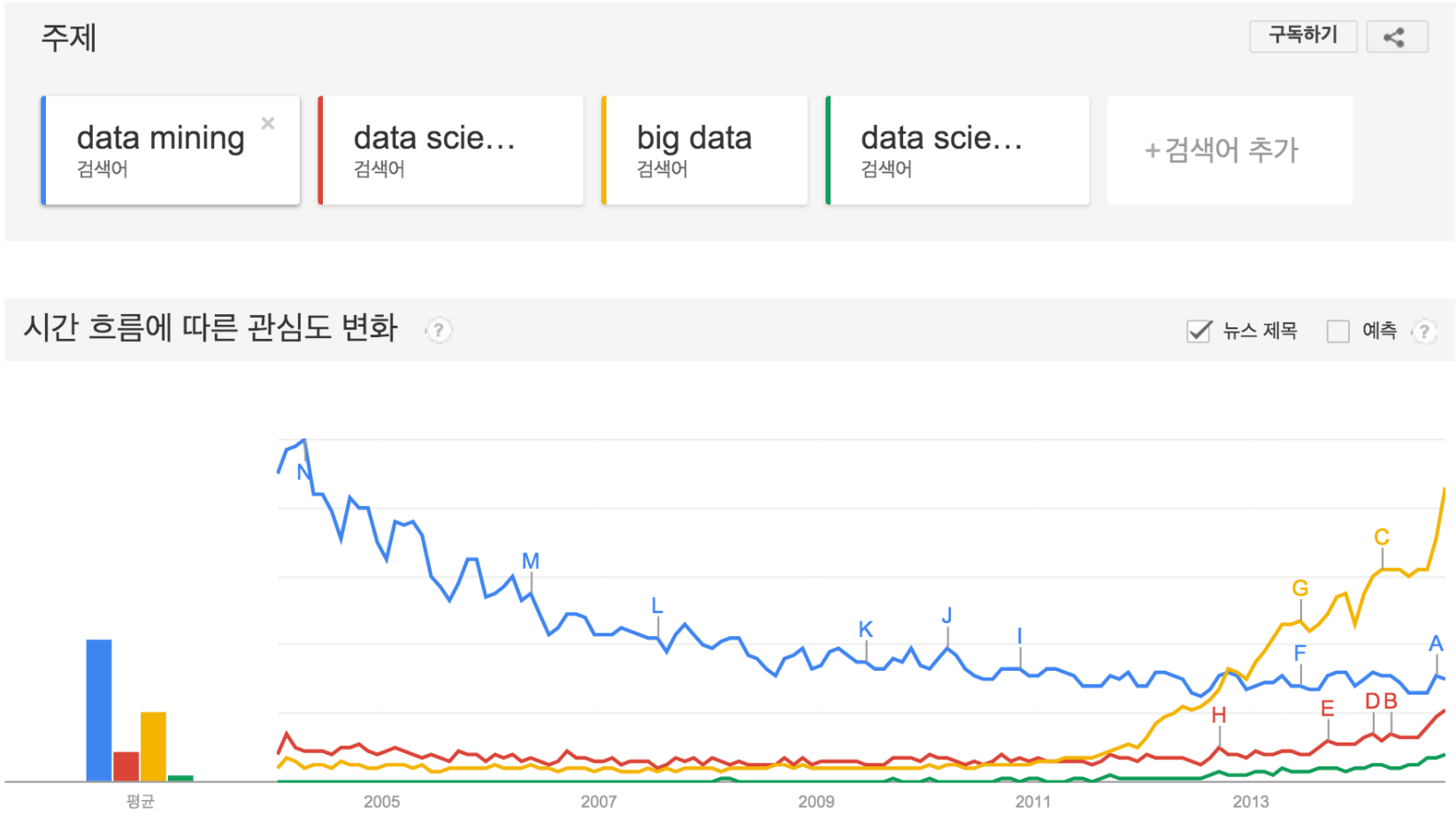
NoSQL

다양한 기술, 다양한 수준, 다양한 기대

ORACLE

Data Mining vs Data Scientist?

From google Trends



2014년 10월(데이터 일부)

- data mining: **30**
- data science: **21**
- big data: **86**
- data scientist: **8**

2015년 10월(데이터 예측)

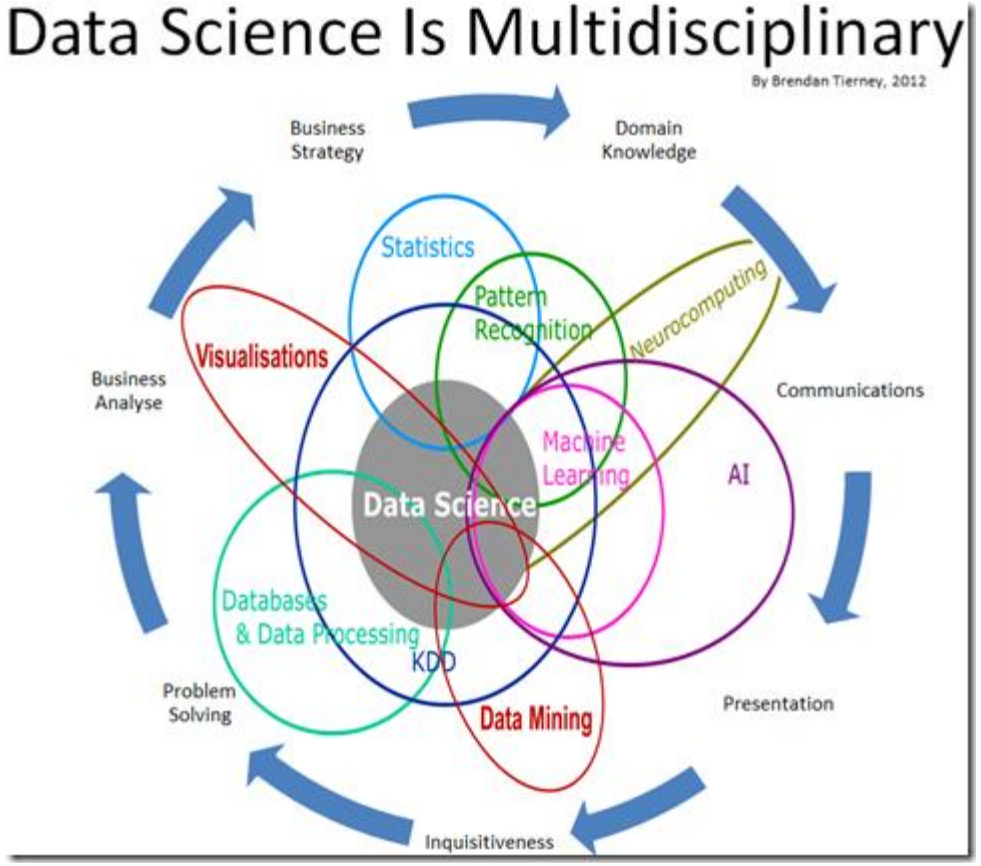
- data mining: **28**
- big data: **74**

출처 : https://www.google.co.kr/trends/explore#q=data%20mining%2C%20data%20science%2C%20big%20data%2C%20data%20scientist

Data Miner 요구 기술

- Database skills (SQL)
- Data Investigation
- Data ECTL (extraction, cleaning, transformation, loading) skills
- Hands-on with Data Mining software
- Some Knowledge of DM techniques
- Understanding of the DM outputs
- Industry knowledge
- Data visualization skills
- Interviewing and requirements gathering skills
- Presentation, writing, and communication skills
- Deployment / Implementation skills

Data Scientist?



출처 : http://www.oralytics.com/2012/06/data-science-is-multidisciplinary.html

그렇다면
국내에서는 어떻게
Big Data 분석을
진행할까요?

분석 시스템 운영 및 구축 시 기업이 바라는 것

확장성, 성능, 운영시스템 배치

- Scalability
- Performance
- Production Deployment

분석가가 원하는 것

기업의 요구를 충족하면서 기술적 접근성이 높고 이미 알고 있는 기술

- 기존 R의 환경, 문장, 문법을 가능한 그대로 사용하길 원함
- 기업에서는 분석에 사용된 자원(코드, 모델) 관리 및 보안에 대한 이슈를 해결하고자 함.
- 분석 수행 시 Laptop으로 내려온 데이터에 대한 보안 관리는 어떻게 해야 하는가?
- 일반 파일 시스템의 관리에 만족하지 않음

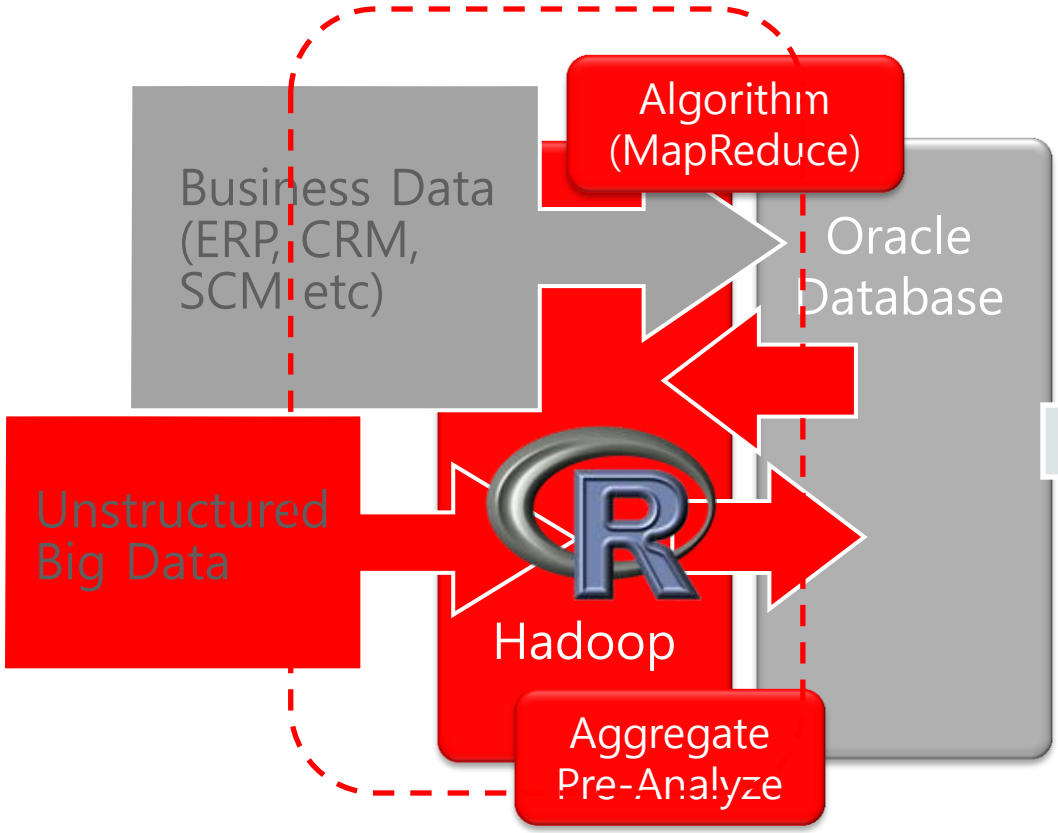
최근 기업들의 빅데이터 분석 아키텍쳐 트랜드

새로운 기술의 도입 및 현실적 대안

ORACLE

Hadoop 기반 R 활용 고급 분석

Hadoop의 비용효율적 분산병렬처리 기술 활용



개발역량이 높은 기업에서는 Java-Mapreduce 프로그램으로 알고리즘을 구현

Oracle BI Enterprise Edition

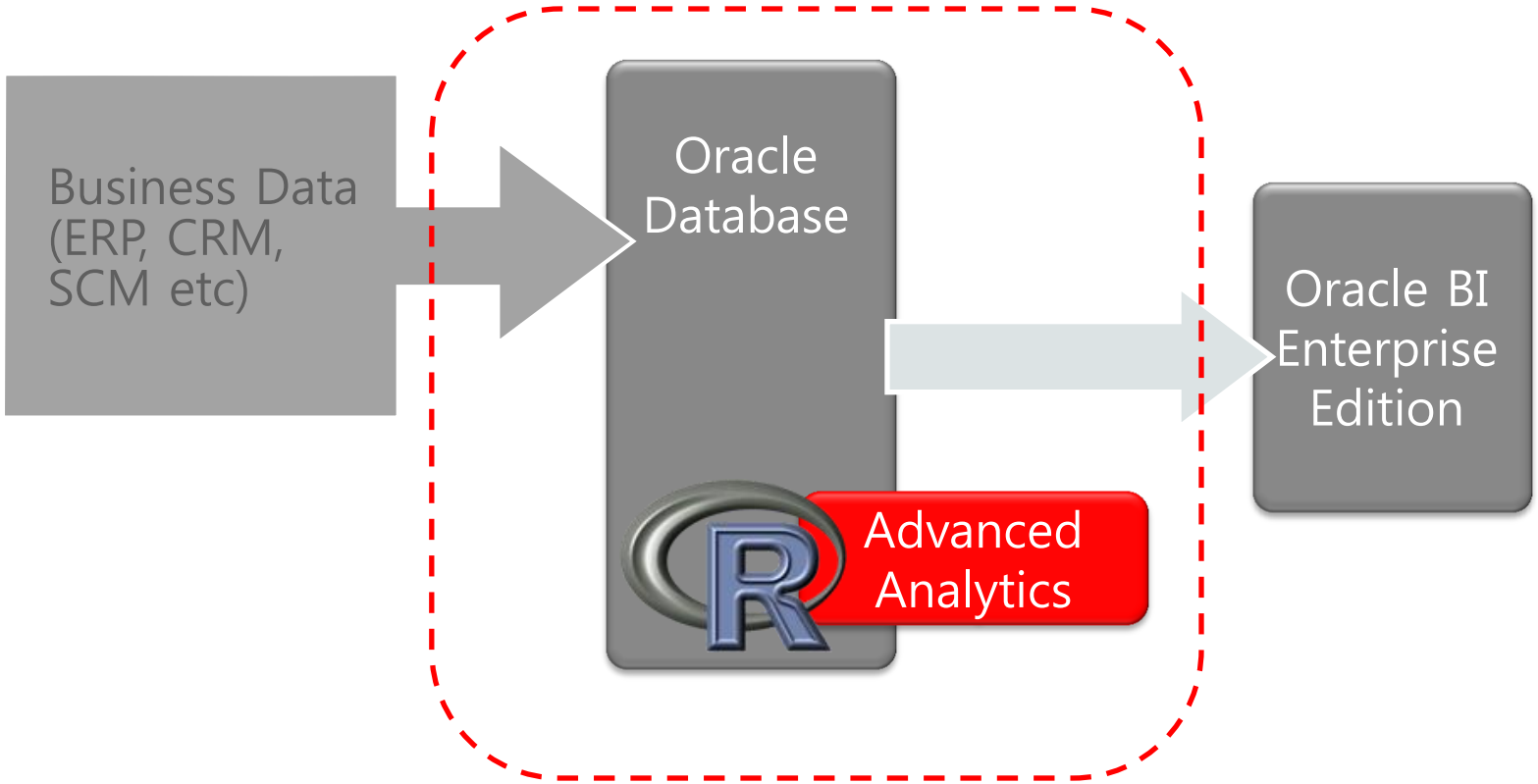
Hadoop&R을 연계하여 일부 알고리즘을 Hadoop 기반으로 재작성하여 활용

Acquire → Organize → Analyze → Decide

ORACLE

DBMS 기반 R 활용 고급 분석

데이터의 이동 없이 데이터가 있는 곳에서 고급 분석 수행



최근 Hadoop 기반으로 빅데이터 분석을 했던 국내 기업들은 In-DB 분석으로 확장 하고 있음

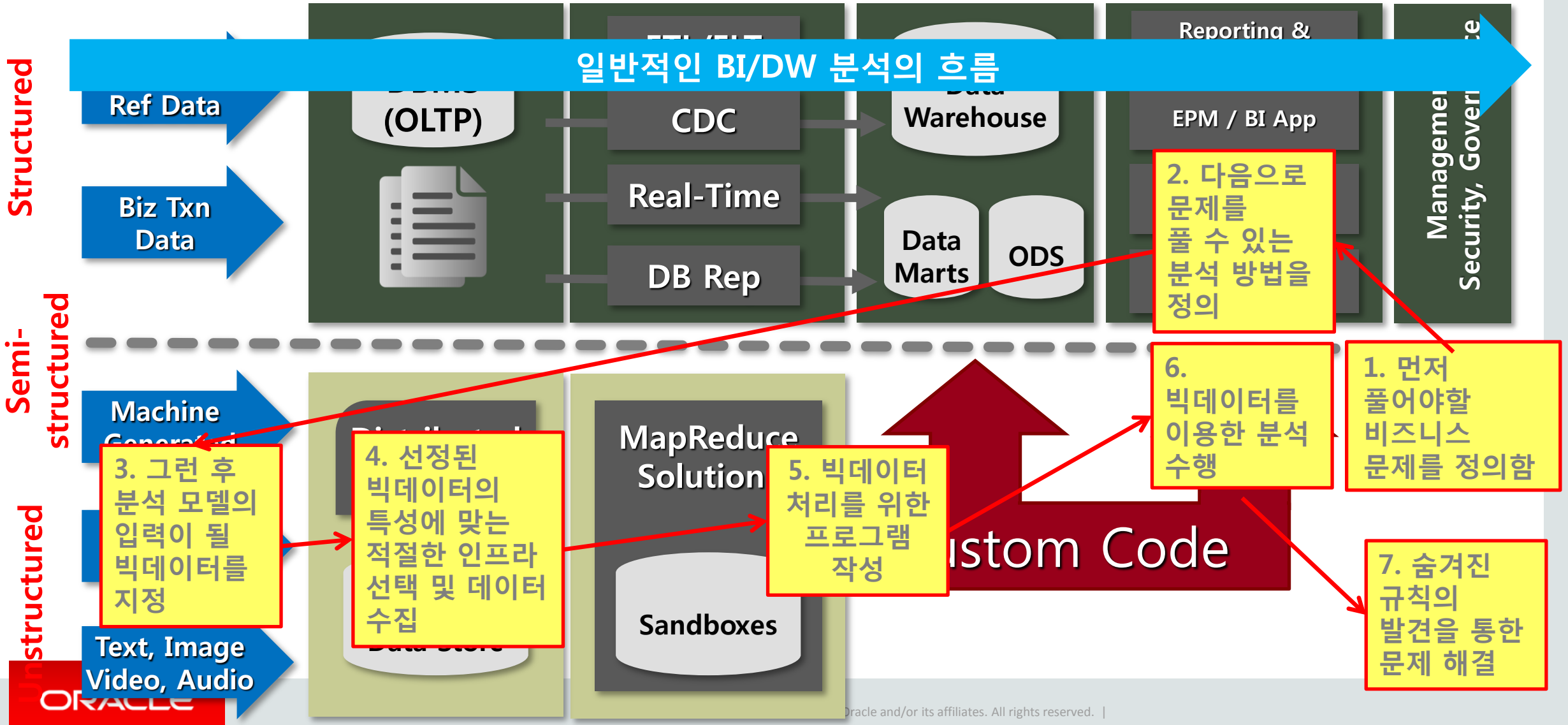
- 분석 대상 데이터는 대부분 DB에 존재
- 이미 개발된 알고리즘/기능으로 신속한 분석을 원하고 있음

Acquire → Organize → Analyze → Decide →

Oracle Advanced Analytics – Best Practices

Nothing is Different; Everything is Different

빅데이터 분석의 접근 단계



Best Together : RDBMS, NoSQL, Hadoop

RDBMS

최적 사용:

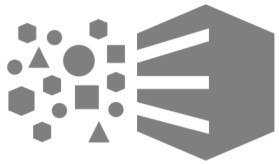
- 비즈니스 데이타 (계좌, 고객 등) High density data
- 엄격한 트랜잭션 처리(ACID)
- 다수의 사용자에 대해 정합성과 안정성 보장
- 100% SQL Compliance
- 고비용

NoSQL

최적 사용:

- SNS, 저밀도 초대용량 데이터
- Put/Get 연산 위주
- Partial Consistency→ Delay 허용
- 유연성과 효율성
- 특화된 용도에 맞게 사용
- RDBMS와는 보완 관계

  → 선택의 폭이 넓어짐



Hadoop

최적 사용:

- 웹/센서 로그 등의 low density data
- 기존 데이터의 Archival
- Parallel Batch Processing

  ➔ 트랜잭션 지원 안함
- 데이터 전 처리 및 집계에 적합
- 저비용

데이터의 특성에 맞추어 적절한 아키텍쳐에 저장하는 것이 TCO 절감의 출발점

ORACLE

Technology 혁신을 통한 비즈니스 가치 창출

업무 특성에 적합한 Tool을 사용, 함께 사용하여 시너지 효과 창출

Hadoop

비즈니스를 변화

- Disrupt competitors
- Disintermediate supply chains
- Leverage new paradigms
- Exploit new analyses

NoSQL

비즈니스 확장

- Serve data faster
- Meet mobile challenges
- Scale-out economically

Relational

비즈니스 운영

- Integrate existing systems
- Support mission-critical tasks
- Protect existing expenditures
- Ensure skills relevance

Oracle Advanced Analytics—*Best Practices*

Nothing is Different; Everything is Different

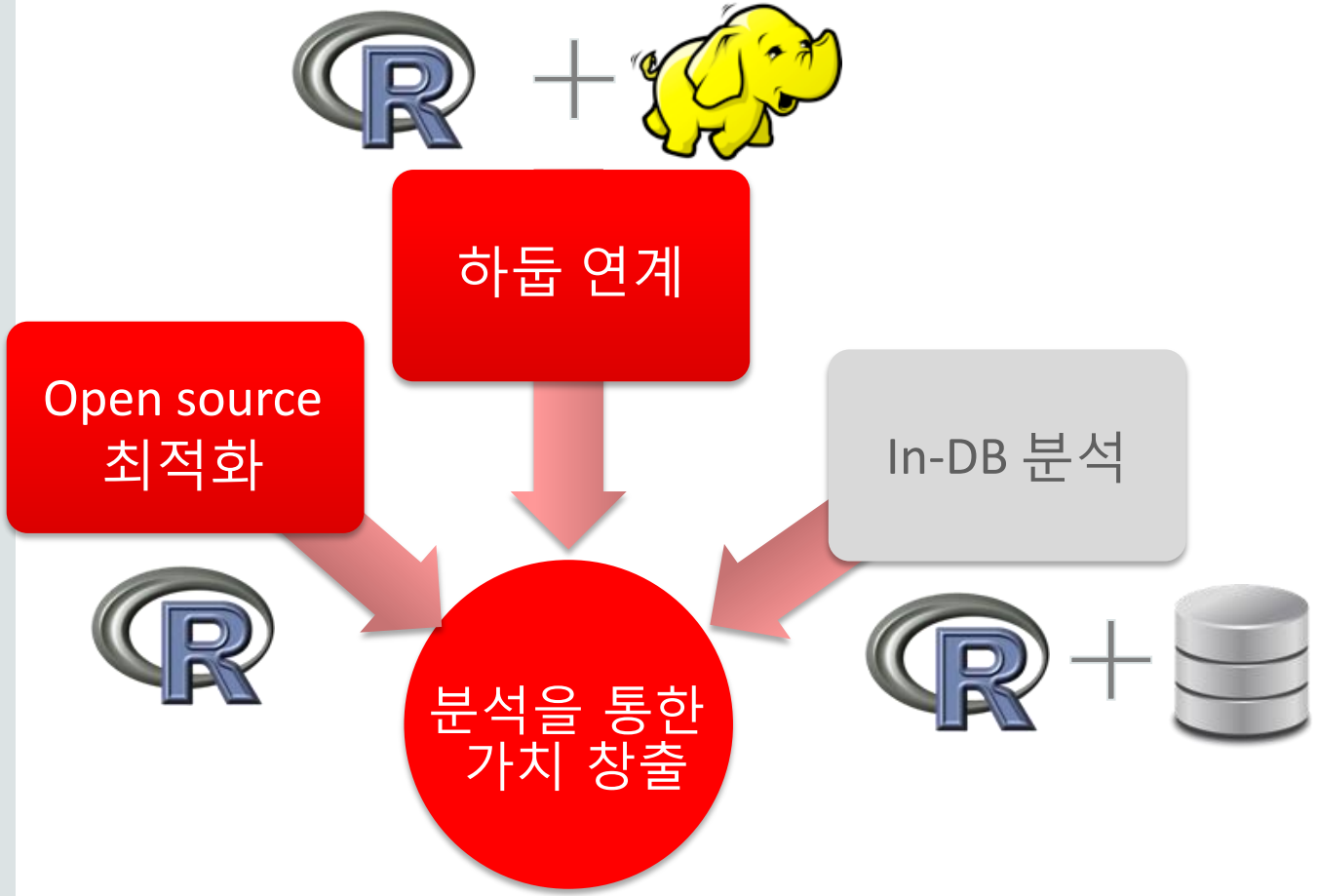
1. Start with a Business Problem Statement
2. Don't Move the Data
3. Assemble the "Right Data" for the Problem
4. Create New Derived Variables
5. Be Creative in Analytical Methodologies
6. Quickly Transform "Data" to "Actionable Insights"
7. Automate and Deploy Enterprise-wide

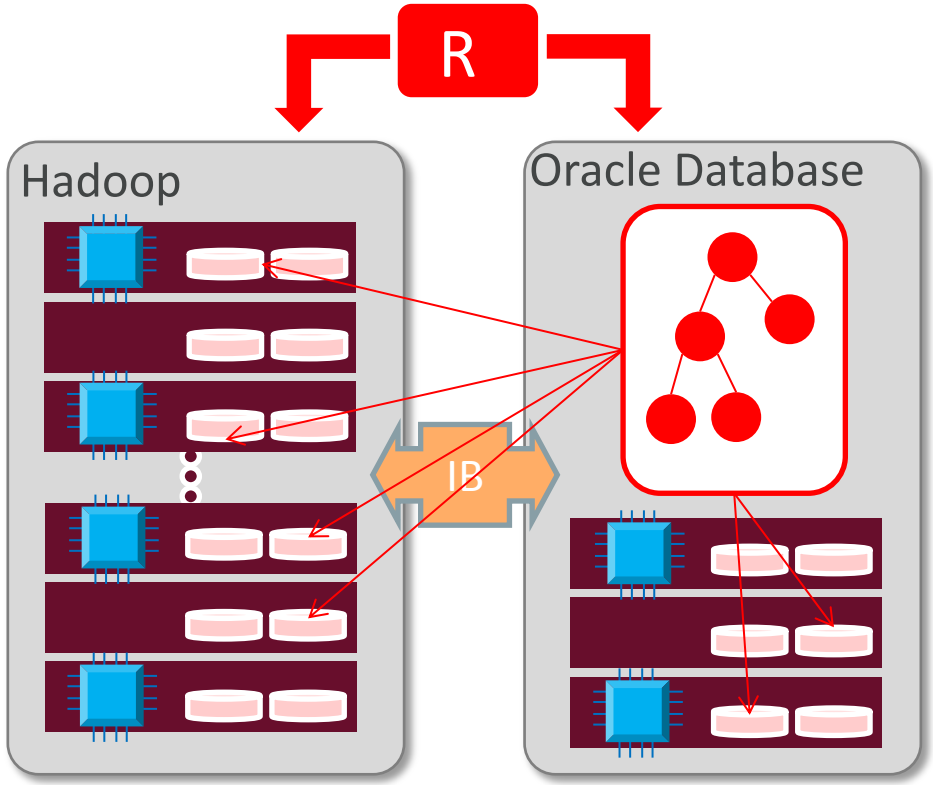
Big Data 분석을 위한 Core Technology

분석가에게 필요한 지식

Big Data 분석 플랫폼 전략

R을 활용한 Any Data, Any Where 분석 지원





Oracle R Distribution : 오픈 소스 R 성능 극대화

오라클이 무상으로 배포

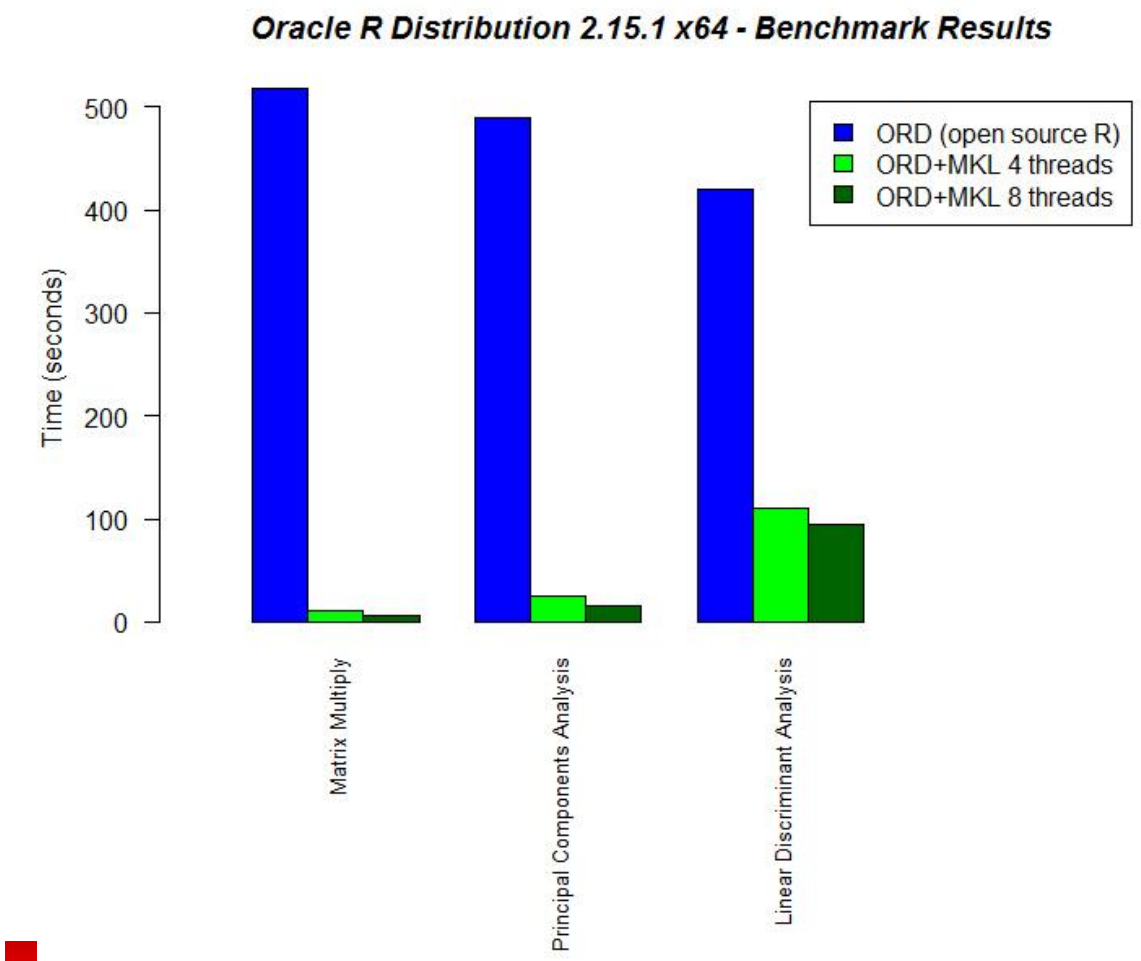
+

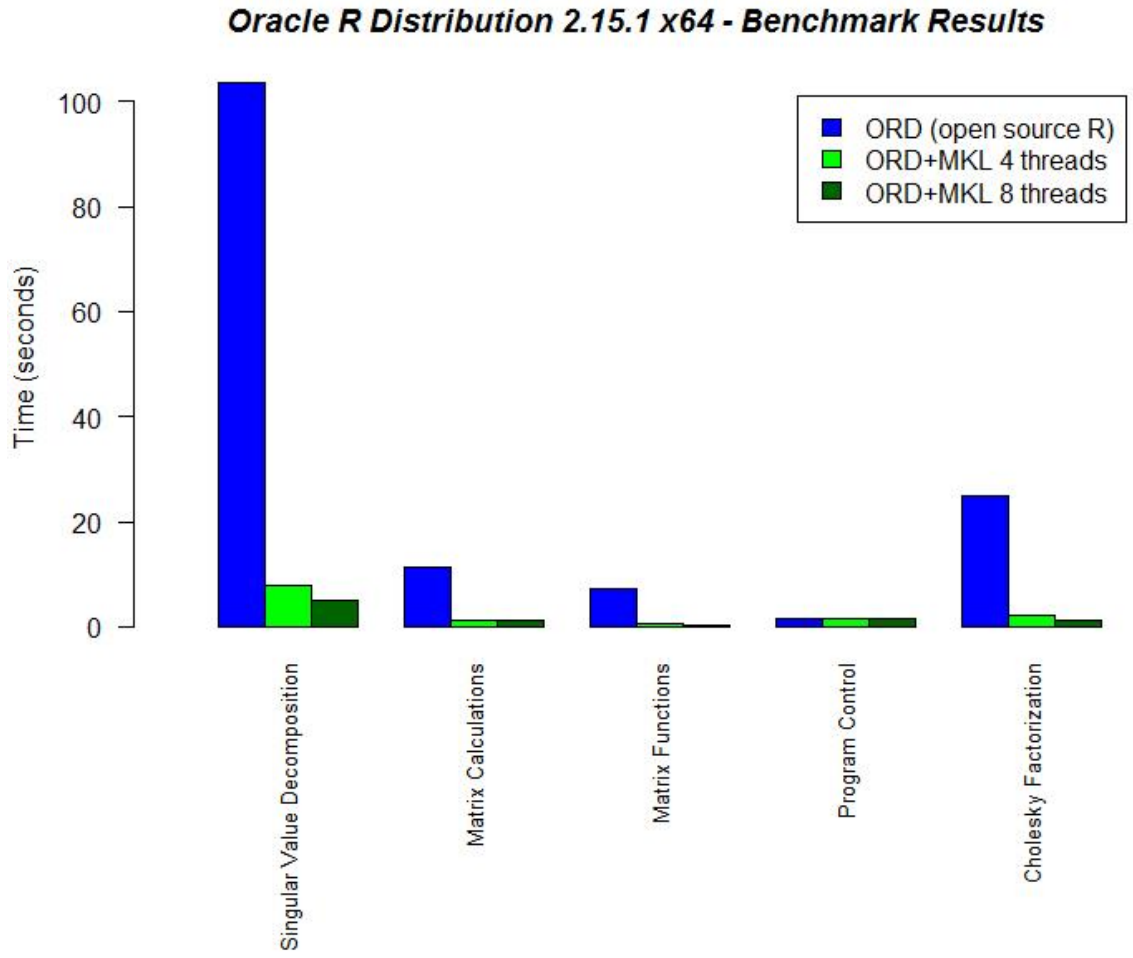
Ability to dynamically load
Intel Math Kernel Library (MKL)
AMD Core Math Library (ACML)
Solaris Sun Performance Library

- 행렬연산 수행 시 성능 극대화 필수
- Compiler 선택 및 MKL 도입한다면 큰 성능 향상을 이룰 수 있음
- 최적화된 플랫폼 구축은 항상 대가를 원함(시행착오)

Oracle R Distribution Performance benchmark

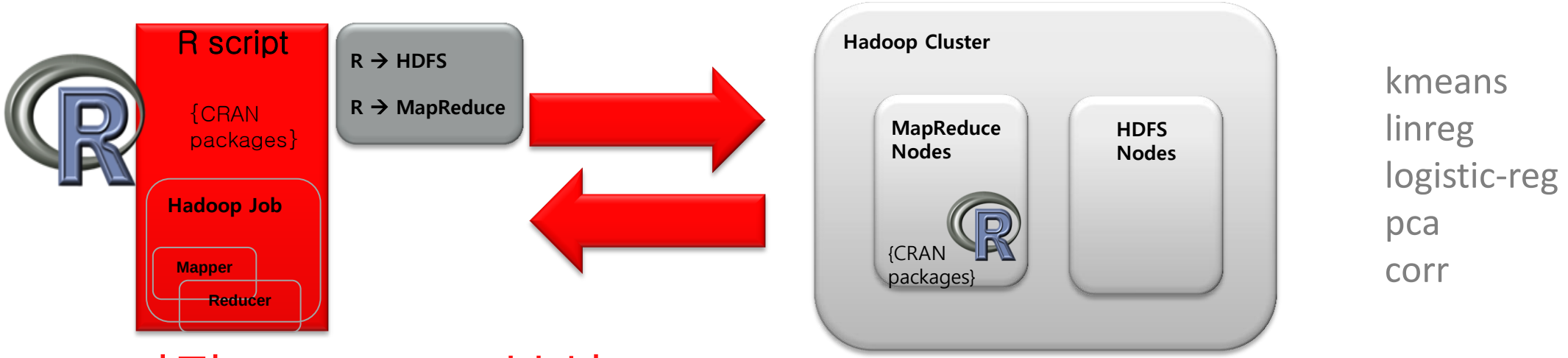
Benchmark Results





Hadoop과 R을 연계

기존 알고리즘을 하둡 기반 방식으로 전환



kmeans
linreg
logistic-reg
pca
corr

- 마음껏 저장하고 마음껏 분석하자
- 기존 분석 알고리즘을 하둡의 MapReduce 방식으로 재구성하여 방대한 데이터를 대상으로 빠른 속도의 분석 수행
- 다양한 최신 기법의 분석 알고리즘을 비용 효율적으로 개발 및 활용

Hadoop 연계 시 필요한 사항

Using MapReduce

Function	Description
hadoop.exec	Starts the Hadoop engine and sends the mapper, reducer, and combiner R functions for execution. You must load the data into HDFS first.
hadoop.jobs	Lists the running jobs, so that you can evaluate the current load on the Hadoop cluster.
hadoop.run	Starts the Hadoop engine and sends the mapper, reducer, and combiner R functions for execution. If the data is not already stored in HDFS, then hadoop.run first copies the data there.
orch.dryrun	Switches the execution platform between the local host and the Hadoop cluster. No changes in the R code are required for a dry run.
orch.export	Makes R objects from a user's local R session available in the Hadoop execution environment, so that they can be referenced in MapReduce jobs.
orch.keyval	Outputs key-value pairs in a MapReduce job.
orch.keyvals	Outputs a set of key-value pairs in a MapReduce job.
orch.pack	Compresses one or more in-memory R objects that the mappers or reducers must write as the values in key-value pairs.
orch.temp.path	Sets the path where temporary data is stored.
orch.unpack	Restores the R objects that were compressed with a previous call to orch.pack.

Functions for Using MapReduce

Hadoop & R 연계 코드

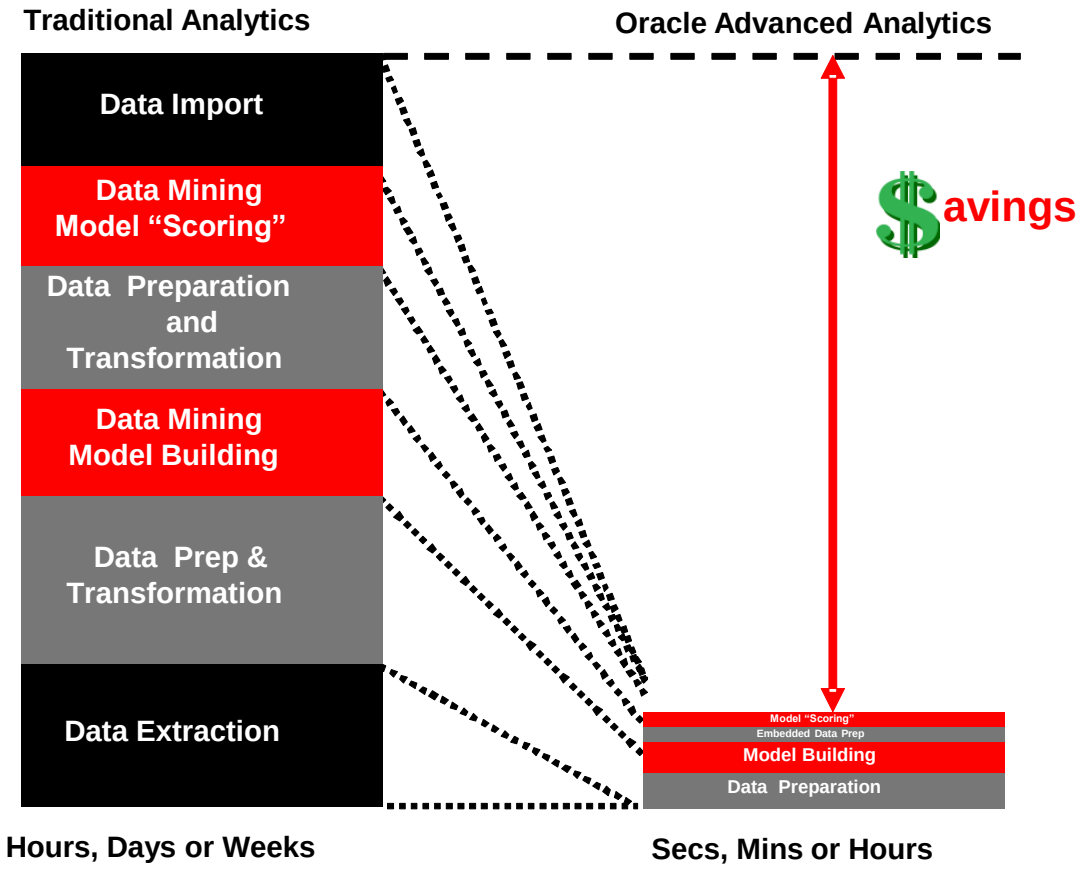
- Average the arrival delay for all flights with destination of SFO
- Map function filters records that have destination of SFO
- Reduce function computes the mean of arrival delay

```
dfs <- hdfs.attach('ontime_R')
res <- NULL
res <- hadoop.run(
   dfs,
   mapper = function(key, ontime) {
      if (key == 'SFO' & !is.na(ontime$ARRDELAY)) {
         keyval(key, ontime) }
   },
   reducer = function(key, vals) {
      sumAD <- 0
      count <- 0
      for (x in vals) {
        sumAD <- sumAD + x$ARRDELAY
        count <- count + 1
      }
      res <- sumAD / count
      keyval(key, res)
   }
)
```

```
> hdfs.get(res)
  key    val1
1 SFO  17.44828
```

In-DB 분석(별도 분석 플랫폼 없이 즉시 분석)

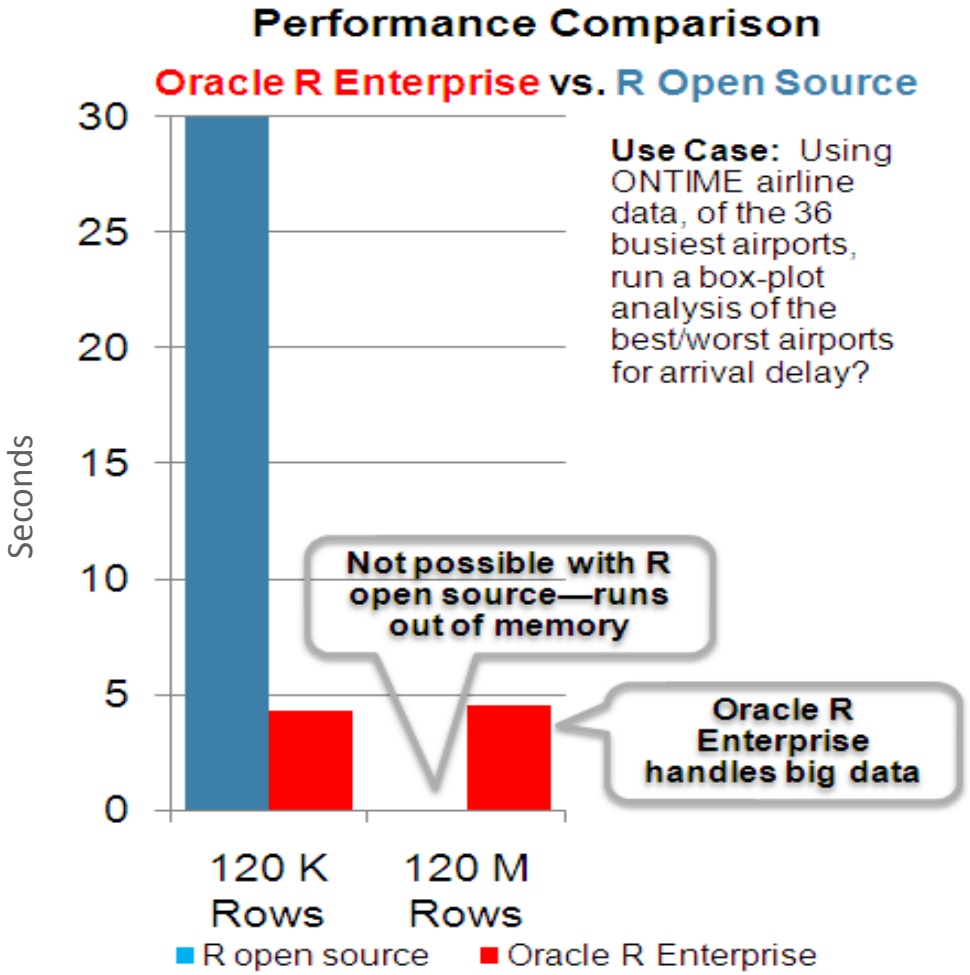
플랫폼 간 데이터 이동 없이 높은 보안환경에서 즉시 분석 지원



- **데이터는 데이터베이스에 존재**
  - SQL 커널에서 확장성있고 병렬처리 가능한 데이터 마이닝 알고리즘 구현
  - 데이터 준비가 자동화됨
- **데이터를 이용해 인사이트를 얻을 수 있는 지름길**
  - 예측 어플리케이션을 위한 생산성이 높은 개발환경 제공
  - 데이터베이스 스코어링 엔진이 SQL 마이닝 함수를 Exadata 스토리지 티어로 옮김
- **적은 총 소유 비용(TCO)**
  - 데이터 중복 제거
  - 별도의 분석용 서버들을 제거
  - 확장성, 관리성, 보안 지원

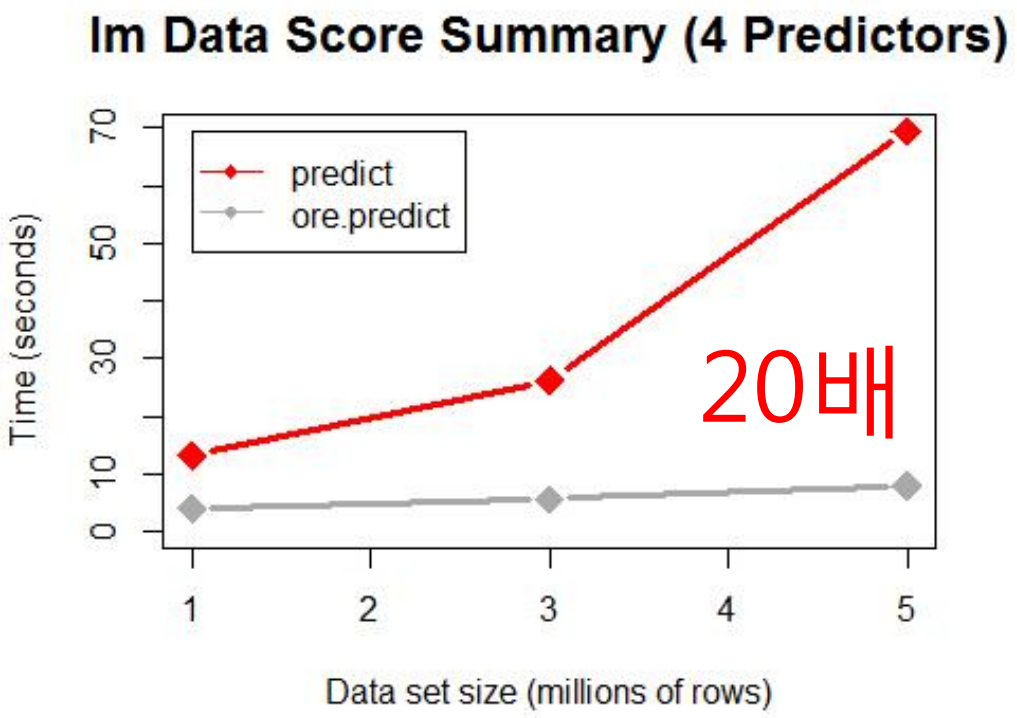
In-DB 분석의 효과

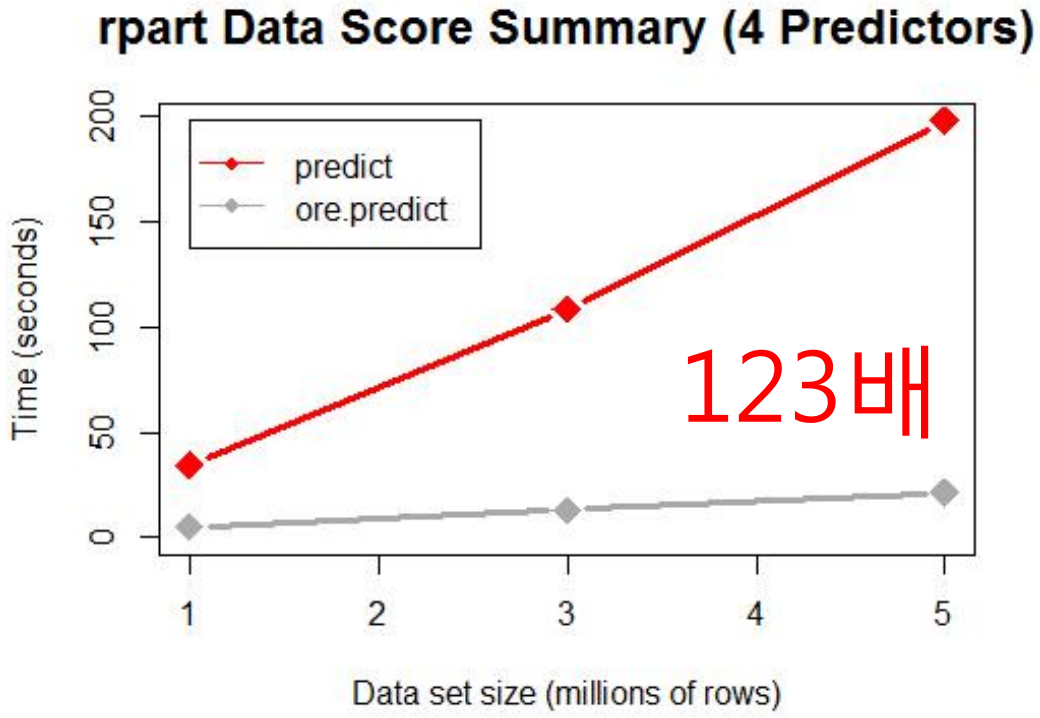
- Transparently function-ships R constructs to database via R → SQL translation
  - –Data structures
  - –Functions
    - Data manipulation functions (select, project, join)
    - Basic statistical functions (avg, sum, summary)
    - Advanced statistical functions(gamma, beta)
- Performs data-heavy computations in database
  - –R for summary analysis and graphics
- Transparent implementation enables using wide range of R “packages” from open source community



성능 Benchmark – Predict 수행

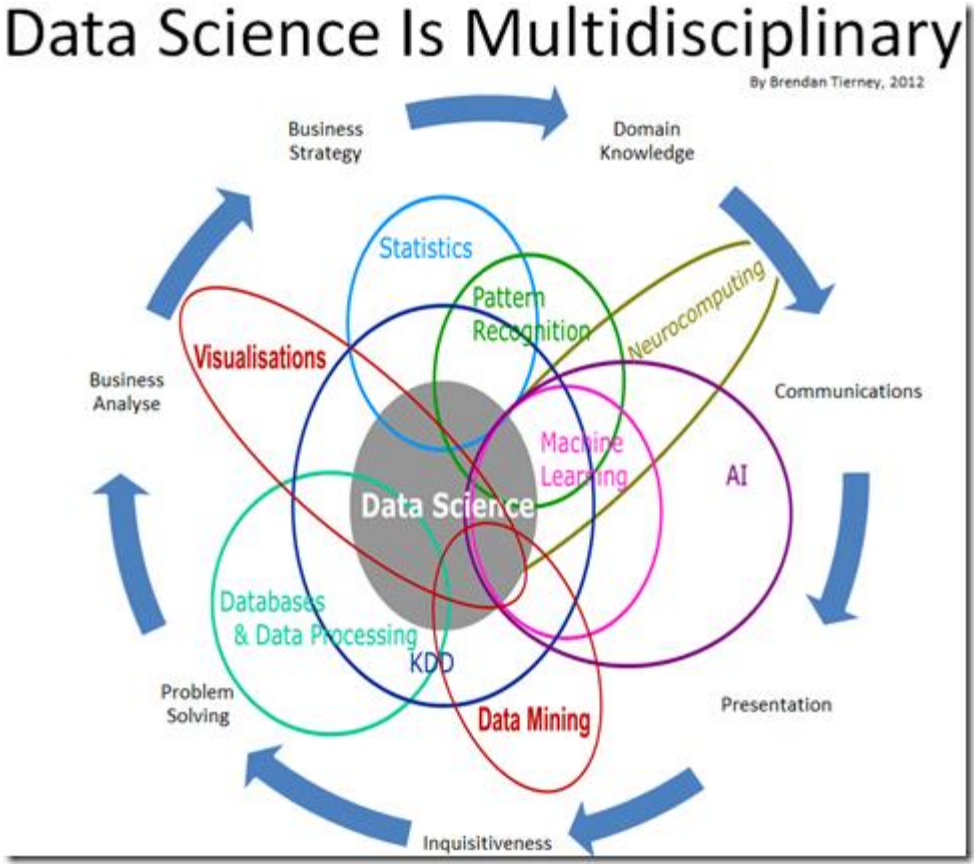
lm, rpart Predict 수행 비교(20배, 123배 빠른 성능)



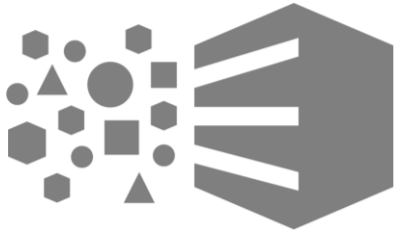


R predict vs. ORE ore.predict

만약 데이터가 여러 플랫폼에 있다면?



Hadoop



NoSQL

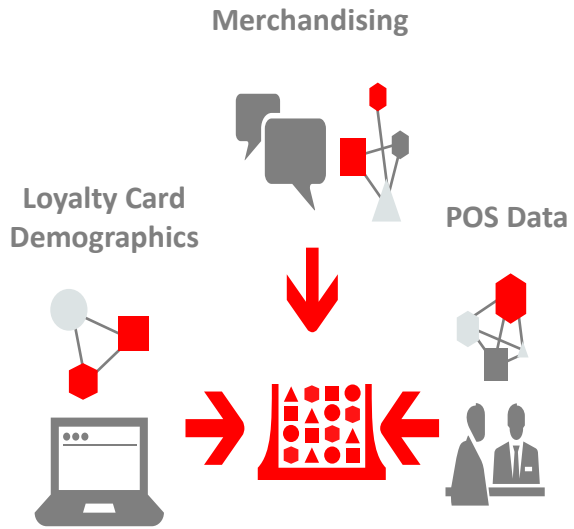
RDBMS

빅데이터 도입에 대한 장벽
Lessons Learned

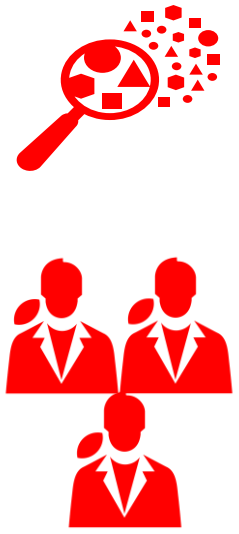
- **기술**
  - 빅데이터를 활용하기 위한 도구나 교육의 부족
- **통합**
  - 현 아키텍처에 복잡하다고 느껴지는 빅데이터 아키텍처 추가
- **보안**
  - 명확한 데이터 통제 방안 부재

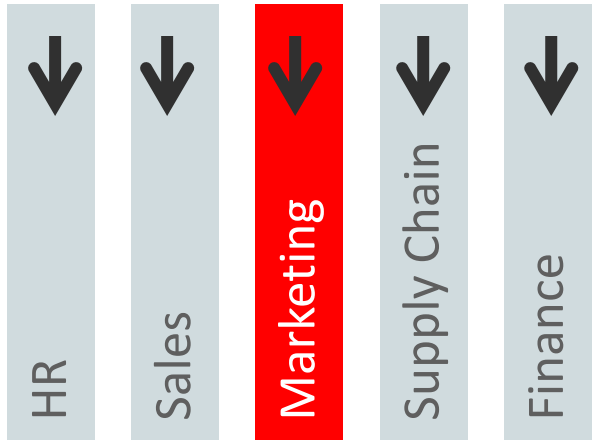
Enterprise Analytics and the Data Warehouse

Gather, Wait , Analyze, Repeat



N days
N Weeks





Warehouse Platform
다수의 업무 사용자

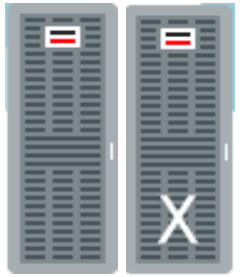
Analytics Platform
소수의 분석가

Repeat in every LOB

오라클의 해결 방안

통합

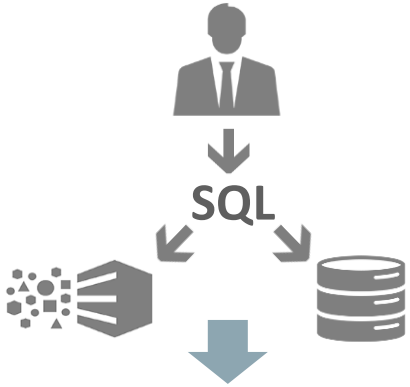
기존 아키텍쳐에 Big Data를 추가하는 것은 매우 복잡한 일



엔지니어드 시스템

스킬

Big Data 활용에 필요한 도구와 교육의 부족



모든 데이터를 SQL로 분석

보안

신기술에 대한 관리 체계 수립 및 실행을 위한 명확한 방식의 부재

모든 데이터에 데이터베이스 보안 적용

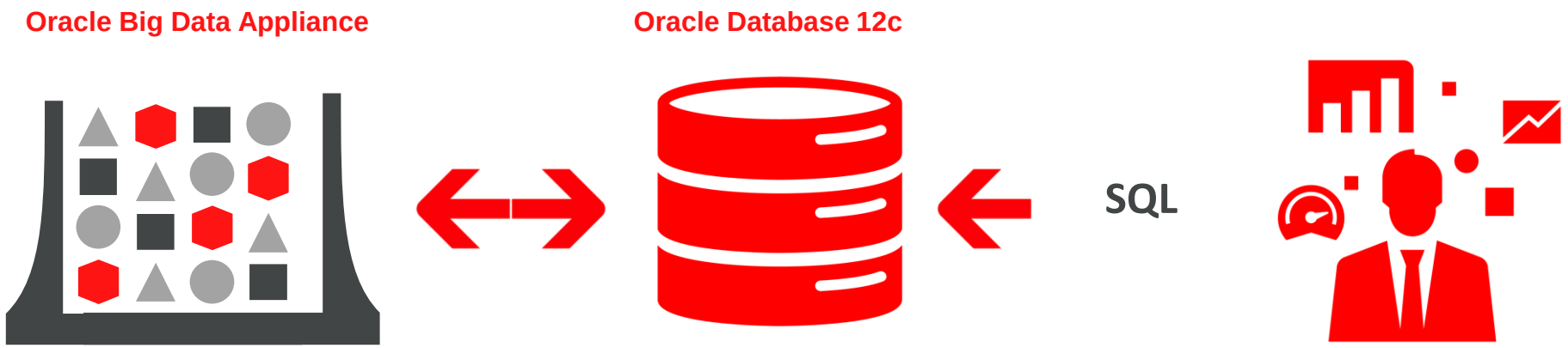
오라클 Big Data SQL 목적

One fast SQL query, on all your data.

오라클 SQL on Hadoop, 그 이상의 기능을 제공

- **Smart Scan** 서비스 제공
- **풍부하고 다양한 SQL 연산** 제공
- 오라클 데이터베이스의 **보안성** 제공

모든 데이터에 대한 관리 체계 수립



하둡에 JSON 데이터를 저장

중요 비즈니스 데이터 오라클 데이터베이스에 저장

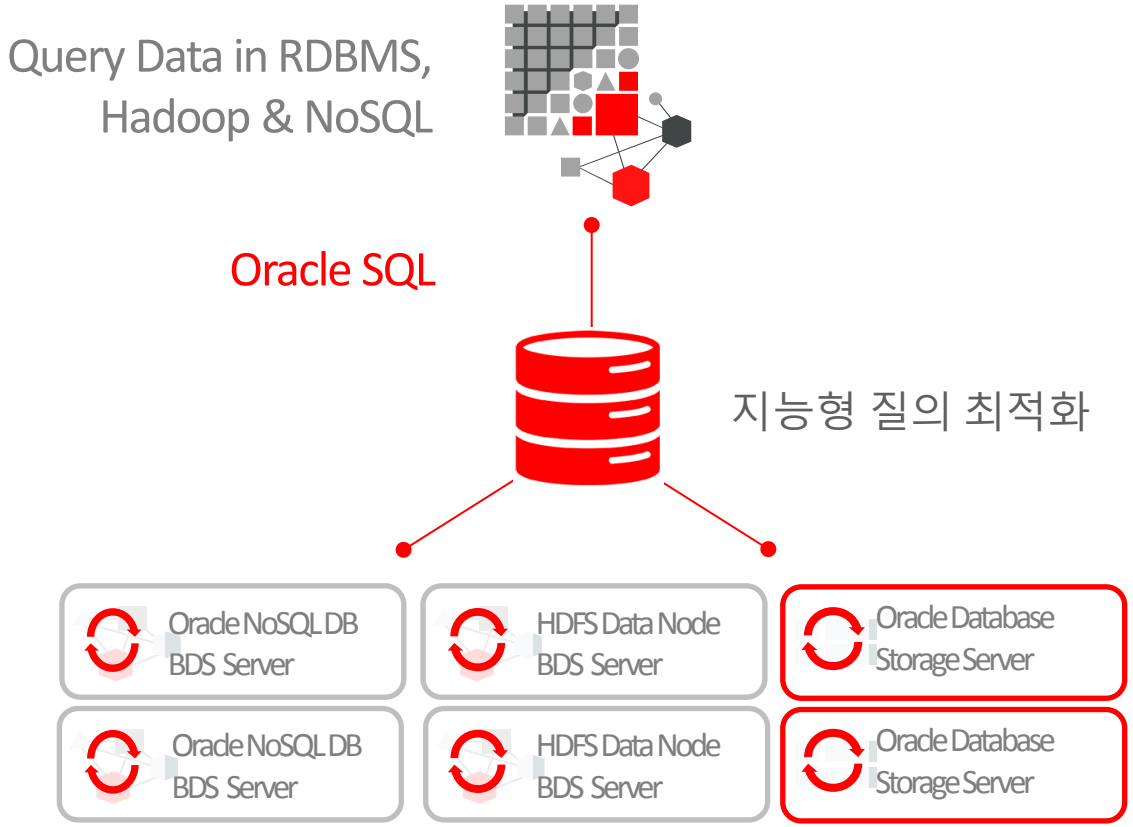
SQL을 통해 데이터 분석

- 하둡에 저장된 데이터에 고급 보안 적용
  - Masking/Redaction
  - Virtual Private Database
  - 세밀한 수준의 접근 제어

```
DBMS_REDACT.ADD_POLICY(
  object_schema => 'hr',
  object_name => 'employee',
  column_name => 'social_sec_num',
  policy_name => 'redact_ssn',
  function_type => DBMS_REDACT.FULL,
  expression => '1=1'
  );
```

Oracle Big Data SQL 기능 및 성능 제공 방식

성능 최적화 : 오라클 Big Data Appliance 기반의 스마트 스캔



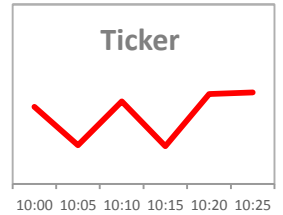
Fast

병렬처리 극대화

데이터 필터링 로컬에서 수행

데이터 이동 최소화

Oracle SQL을 통한 패턴 매칭 수행
오라클 SQL 분석 Functions



주식 시장 데이터 패턴 매칭- Double Bottom (W)

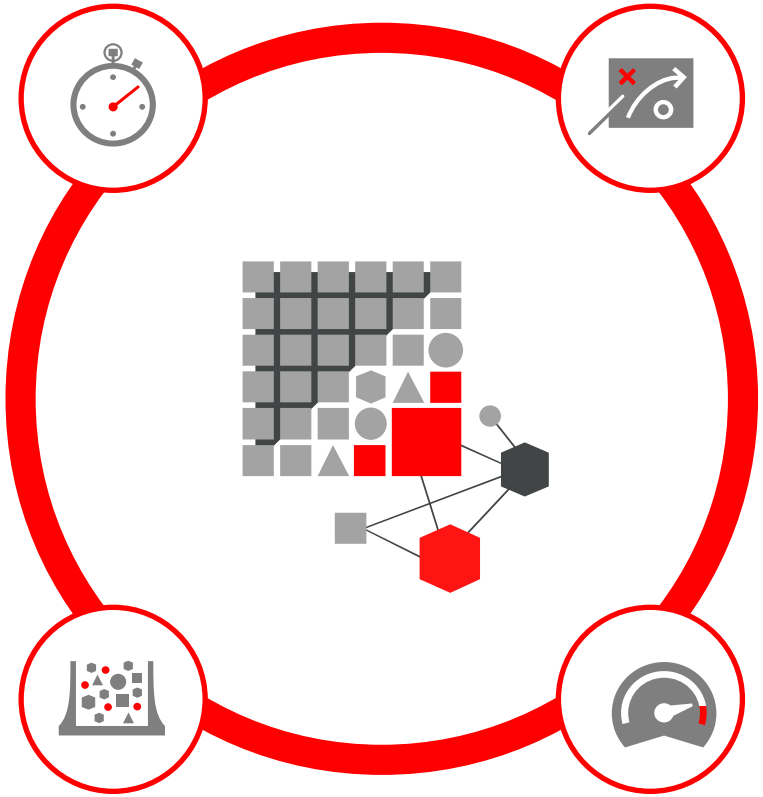
```
package pigstuff;

import java.io.IOException;
import java.util.ArrayList;
import java.util.Iterator;
import org.apache.pig.EvalFunc;
import org.apache.pig.PigException;
import org.apache.pig.backend.executionengine.ExecException;
import org.apache.pig.data.BagFactory;
import org.apache.pig.data.DataBag;
import org.apache.pig.data.DataType;
import org.apache.pig.data.Tuple;
import org.apache.pig.data.TupleFactory;
import org.apache.pig.impl.logicalLayer.FrontendException;
import org.apache.pig.impl.logicalLayer.schema.Schema;

/**
 *
 * @author nbayliss
 */
public class W_FINDER extends EvalFunc<Tuple> {

    private class V0Line {

        String state = null;
        String[] attributes;
        String prev = "";
        String next = "";

        public V0Line(String[] atts) {
            attributes = atts;
        }

        public String[] getAttributes() {
            return attributes;
        }
```

```
SELECT first_x, last_z
FROM ticker MATCH_RECOGNIZE (
        PARTITION BY name ORDER BY time
        MEASURES FIRST(x.time) AS first_x,
                 LAST(z.time)  AS last_z
        ONE ROW PER MATCH
        PATTERN (X+ Y+ W+ Z+)
        DEFINE X AS (price < PREV(price)),
               Y AS (price > PREV(price)),
               W AS (price < PREV(price)),
               Z AS (price > PREV(price)  AND
                     z.time - FIRST(x.time) <= 7 ))
```

250+ Lines of Java UDF

12 Lines of SQL

20x 적은 코딩



ORACLE®

Big Data Management System

고급 질의 및 분석
SQL의 모든 기능과 고급 분석 활용

모든 데이터를 효과적으로 활용
관계형, 하둡, NoSQL

안전성 높은 데이터 관리
모든 데이터에 대한 단일화된 관리 체계 수립

가장 빠른 성능
전 플랫폼을 SQL 처리 시 활용

응용프로그램에 영향을 주지 않음
응용프로그램 코드 변경 필요 없음

ORACLE

Big Data 분석 활용 사례

고객 LTV 스코어링 및 수요 예측

EDW 환경을 분석인프라로 활용, In-DB Analytics 도입

Objectives

- Exadata DW 기반으로 In-DB로 수행되는 R로 작성된 모델을 개발
- 분석가가 스코어링을 위한 R 스크립트를 사용
- 수요예측 및 고객 LTV 분야에 활용

Solution

- 데이터 분석가는 EDW위에 데이터 이동 없이 바로 모델을 구축
- 분석 모델 각 버전을 DB에 저장하여 협업 및 반복 수행 및 감사가능성 제공
- R 스크립트는 운영계에 활용되며 수초 내에 수행
- 향후 마케팅 및 세일즈 계획이 가능

"오라클의 Exadata 기반으로 Oracle Advanced Analytics 기능을 활용하여 데이터 분석가가 오픈 소스 R 기술을 이용하여 데이터베이스에 저장된 데이터를 바로 분석할 수 있게 되었습니다. 이를 통해 우리는 기존의 DW 인프라와 프로세스를 재활용하여 신속한 플랫폼을 만들 수 있었습니다. 그리고 향후 확장의 가능성을 보았습니다."

– A사 수석 아키텍트

ORACLE

리스크 모델 개발-금융 서비스 제공사

기존 정보계 데이터 추출 없이 빠른 모델 개발

Objectives

- 신용 스코어링 모델의 성능 개선
- 대용량 데이터 지원 및 확장성 제공

Solution

- 오라클의 Advanced Analytics 옵션을 활용하였음
- Commercial, risk, collections 분야에 예측 분석 모델을 개발

"오라클의 Advanced Analytics 기능을 통해 우리에게 분석을 위해 기존의 데이터웨어하우스에서 데이터를 추출하는 작업을 하지 않게되었습니다. 그리고 기존의 데이터 추출 작업을 하지 않아도 되어 빠르게 commercial, collections 부분의 리스크 모델을 개발할 수 있었으며 강력한 In-DB 분석의 기능의 효과를 경험하였습니다. "

- B사 리스크 개발 매니저

ORACLE

요약

- Big Data 시대, **새로운 관점의 기술**에 대한 이해 필요
- 기존 **관계형** 분석환경의 **강화**(In-DB 분석) + 새로운 **비관계형** 인프라(하둡) **도입**, **하나의 분석 환경** 구축
- 빅데이터 **분석 시장**은 Hadoop 기반 분석에서 **In-DB 기반** 분석으로 **진화**하고 있음(생산성, 유지보수 편의성)
- 데이터 플랫폼의 종류에 상관 없이 분석가에게 단일한 분석환경 제공이 필요.

Hardware and Software
Engineered to Work Together