티베로
인피니데이터 플랫폼을 통한
빅데이터 처리

티베로 컨설팅팀 수석컨설턴트 박근용

TIBERO
(주)티베로

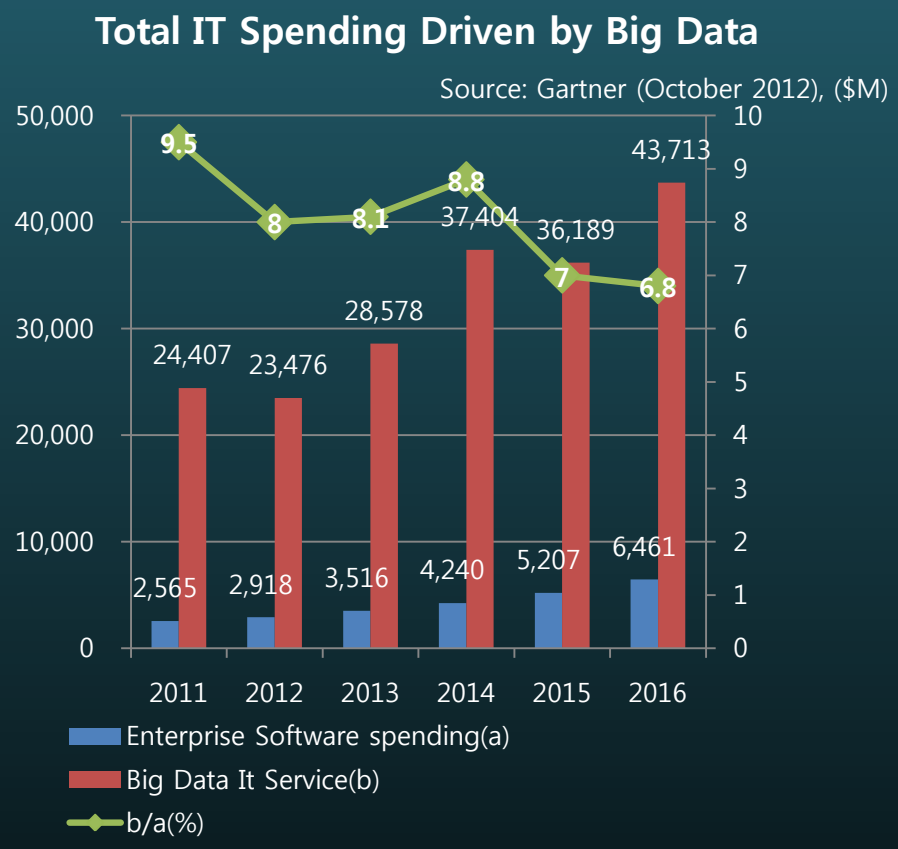
Contents

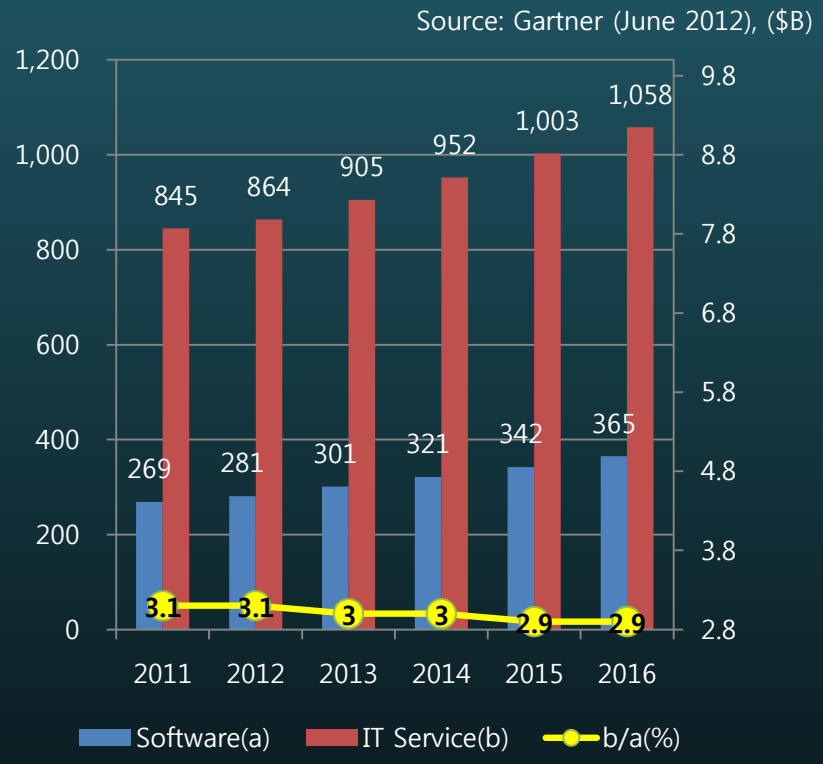
1. 빅데이터 기술 현황 : Hadoop 중심
2. 기존 빅데이터 기술 : Infinidata 2.0
3. 빅데이터 기술의 진화 : Infinidata 3.0

Contents

1. 빅데이터 기술 현황 : Hadoop 중심
2. 기존 빅데이터 기술 : Infinidata 2.0
3. 빅데이터 기술의 진화 : Infinidata 3.0

Big Data 시장 현황

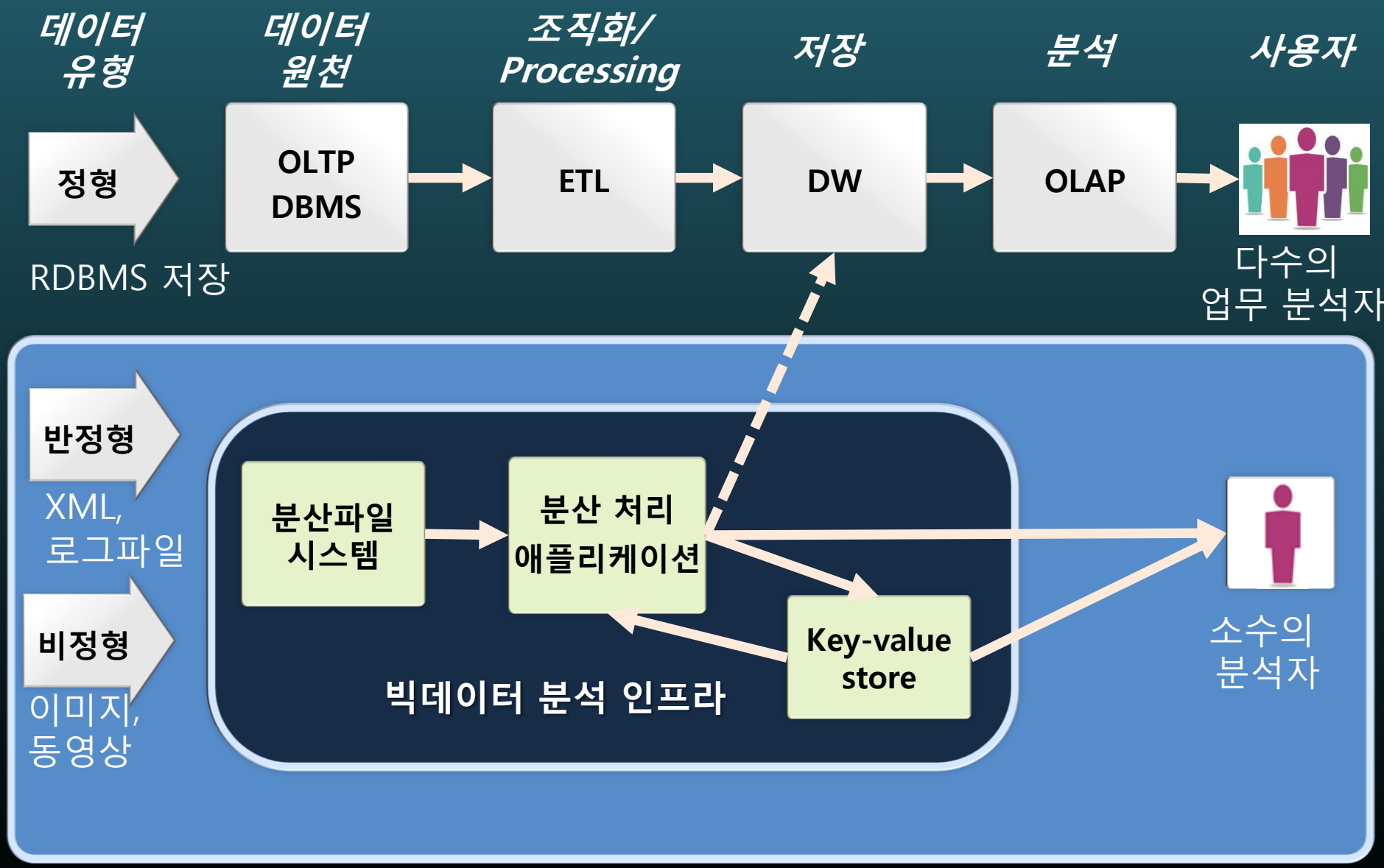




- IT Service 비용 : 240억달러
  (정규직/계약직 인건비)
- 상용 S/W 비용 : 27억달러
  ➔ 상용S/W 대비 인건비 : 약 9배

- IT Service 비용 : 8.5조달러
  (정규직/계약직 인건비)
- 상용 S/W 비용 : 2.7조달러
  ➔ 상용S/W 대비 인건비 : 약 3배

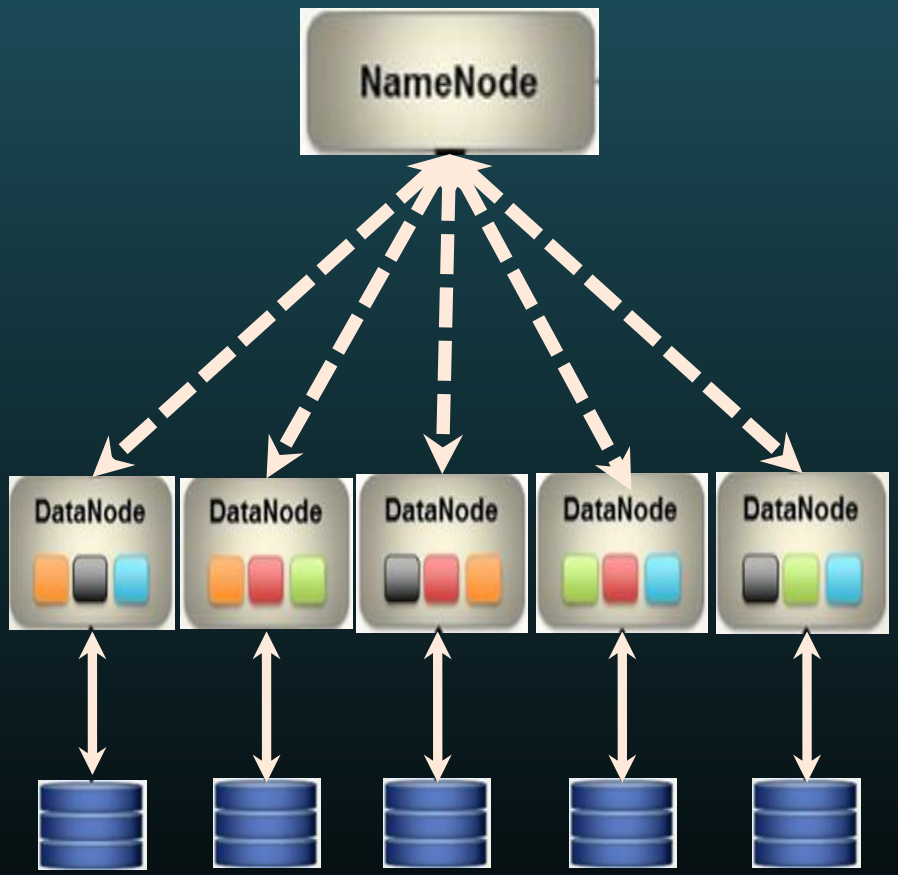
빅데이터 처리 개요


데이터 유형
데이터 원천
조직화/ Processing
저장
분석
사용자
정형
OLTP DBMS
ETL
DW
OLAP
RDBMS 저장
다수의 업무 분석자
반정형
XML, 로그파일
비정형
이미지, 동영상
분산파일 시스템
분산 처리 애플리케이션
Key-value store
빅데이터 분석 인프라
소수의 분석자

빅데이터 분석 인프라 : Hadoop

빅데이터 인프라 기술		주요 기능
분산 데이터베이스		▪ Key-value 데이터 저장 ▪ 임의 데이터 접근 지원
분산 처리 애플리케이션		▪ 비/반정형 데이터의 변환/연산 애플리케이션 개발 지원
분산 파일 시스템		▪ 데이터 파일의 읽기/쓰기 기능 제공

Hadoop HDFS



"Distributed File System"

데이터 파일의
분할/분산 저장

+

데이터 파일의
정보관리

High
Scalability

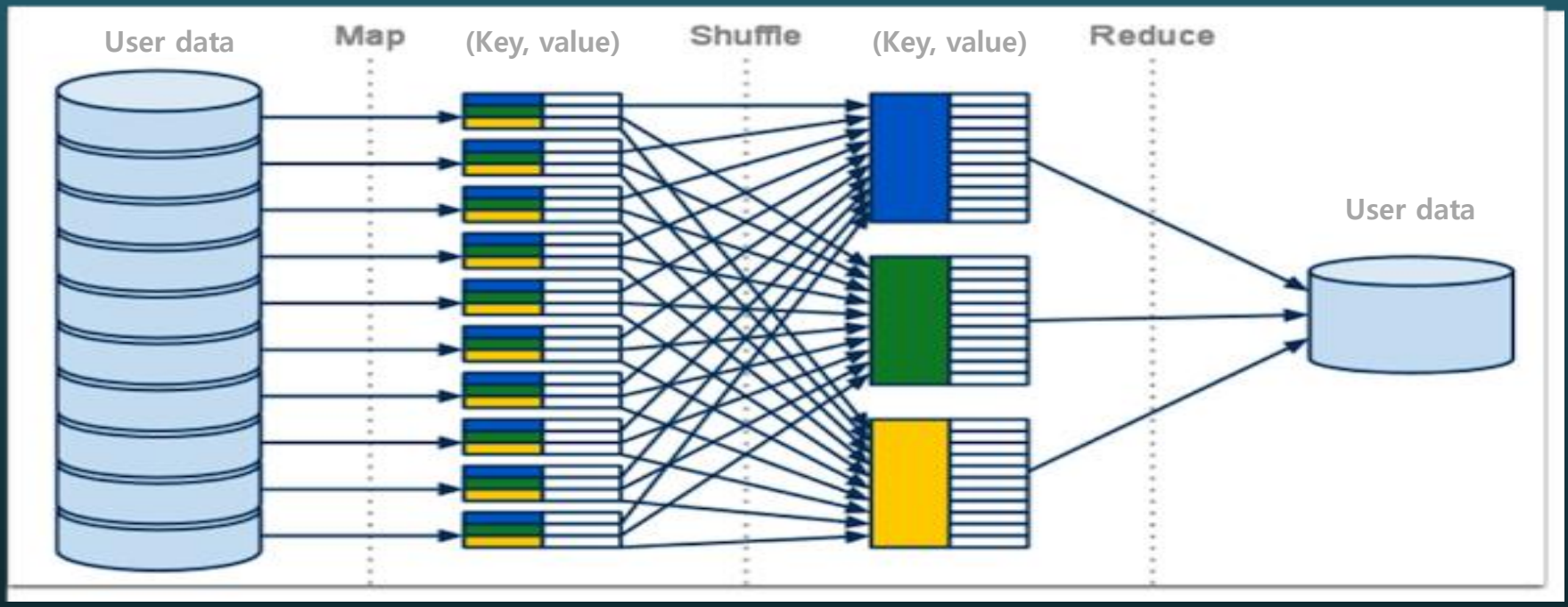
데이터 파일의 읽기/쓰기

+

데이터 파일의
복제

Availability

Hadoop MapReduce



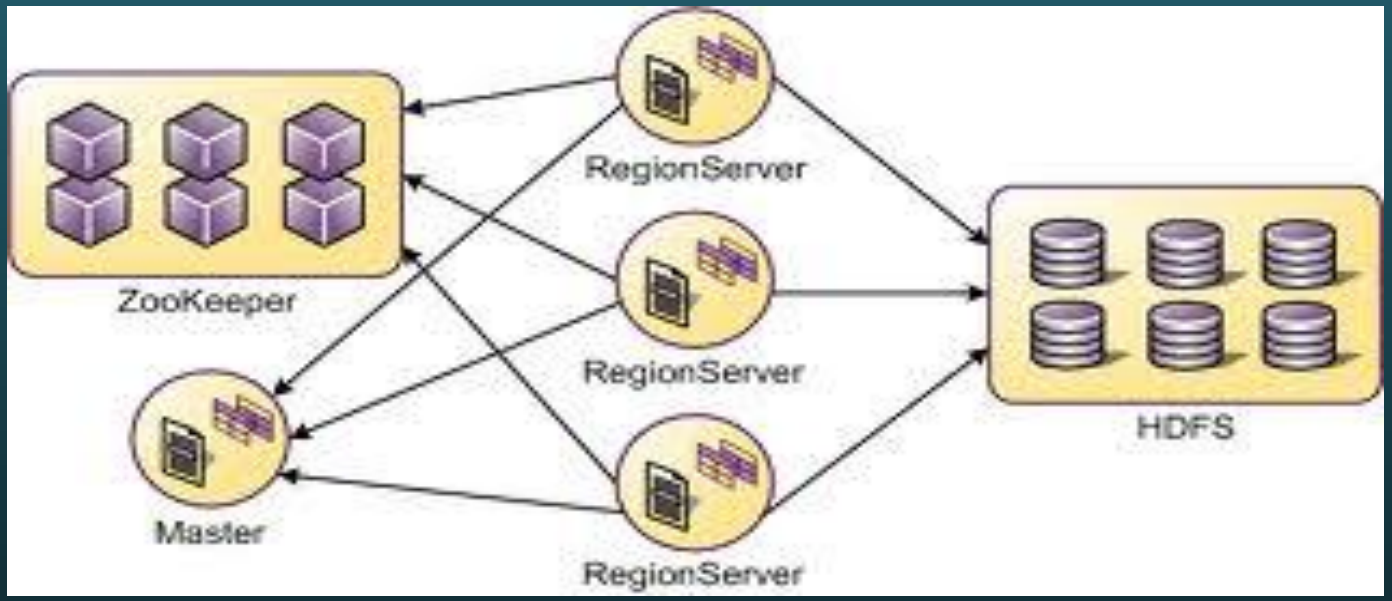
Map/Reduce함수를 이용한 애플리케이션 프레임워크

+ 비(반)정형 데이터의 변환/연산/적재 기능의 병렬처리 개발 지원

!

- 중간결과 데이터 파일 생성 ➔ 성능 저하
- 사용자의 고난이도 프로그래밍 필요

Hadoop Hbase



"Key-Value Store"

반정형 데이터의 부분검색 지원 ➔ 고성능 지원

➕ 실시간 읽기와 쓰기 기능 제공 ➔ 일부 RDBMS 기능 지원

! • SQL 미지원

Contents

1. 빅데이터 기술 현황 : Hadoop 중심
2. 기존 빅데이터 기술 : Infinidata 2.0
3. 빅데이터 기술의 진화 : Infinidata 3.0

Hadoop의 한계와 또 다른 접근방법

hadoop

- 기술인력 의존적 ➔ 시스템 품질 통제 어려움
- 일회적 배치성 분석 ➔ 대화식 분석(OLAP) 미지원
- 사용자 범위 제한적 : 소수의 분석가 활용

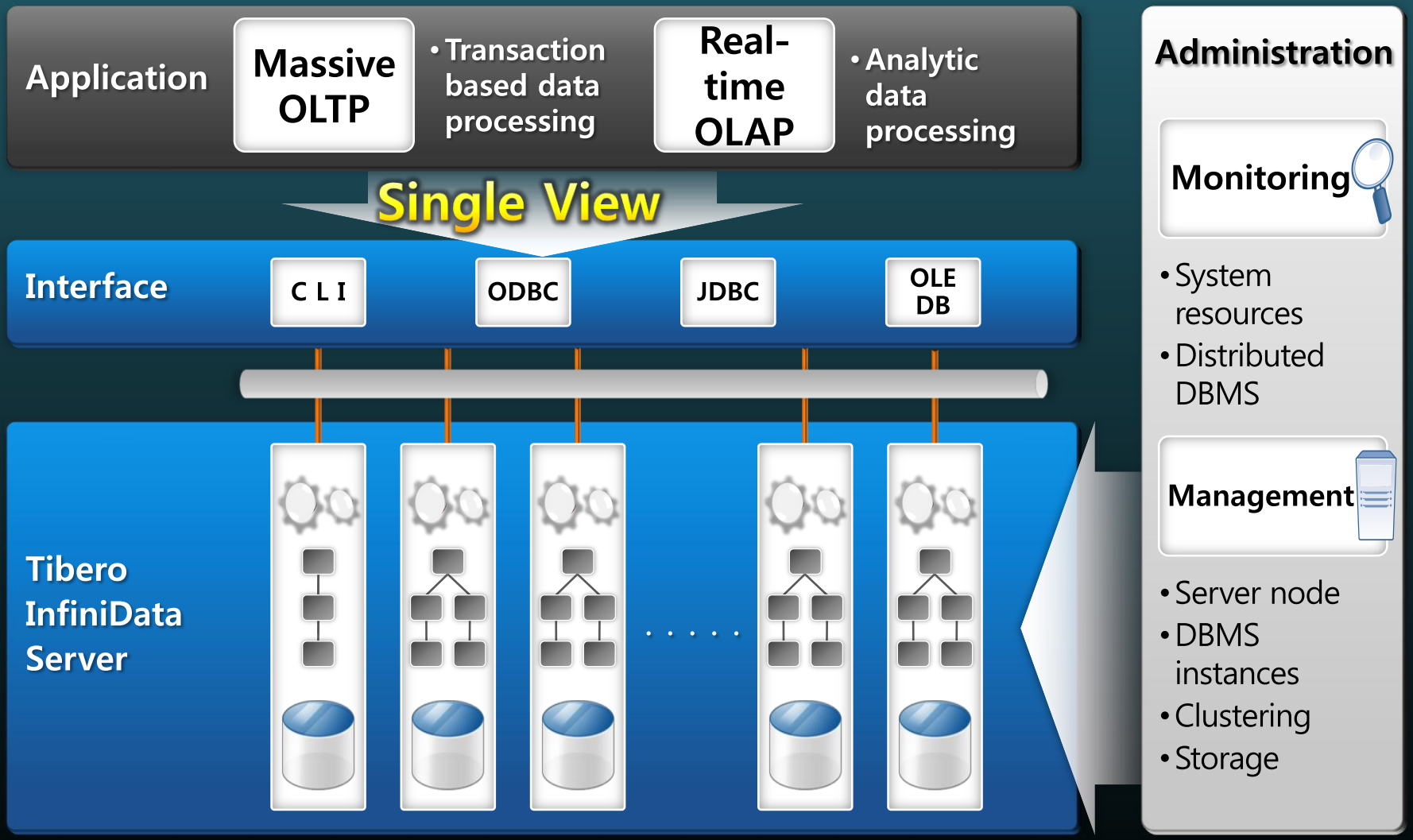
- 기관/기업의 Big Data 분포 : 정형>비정형
- 정형데이터를 반정형으로 전환/분석➔비효율
- RDBMS 추가 또는 DW로 데이터 통합 고려

➔ Gartner : “Logical DataWarehouse”

InfiniData

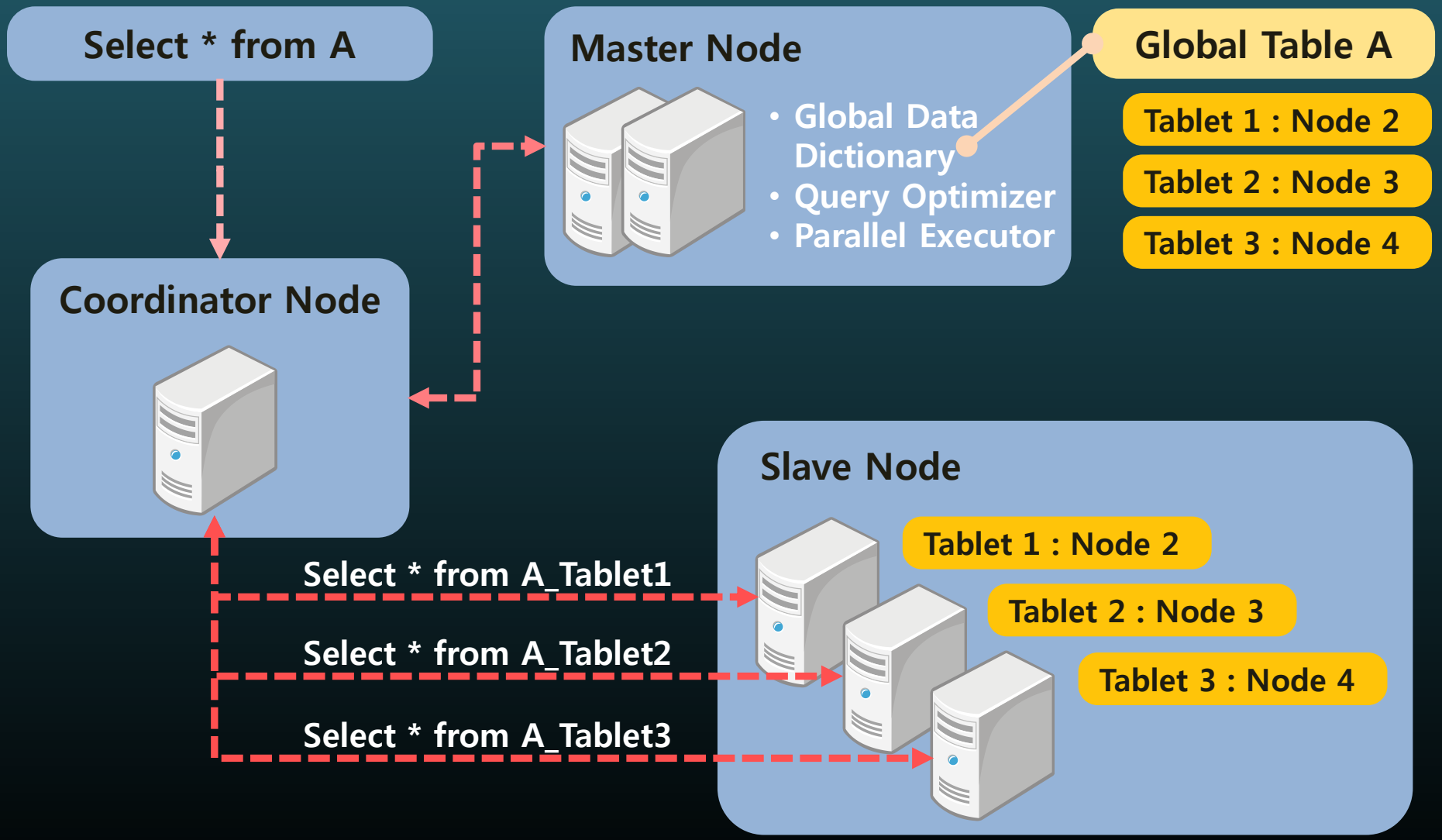
- 저가 서버 이용한 무한 확장성 제공 필요
- 빅데이터의 분산 처리 필요
- H/W 장애에 대한 고가용성 제공 필요

InfiniData 2.0 : Scalable Cloud DBMS

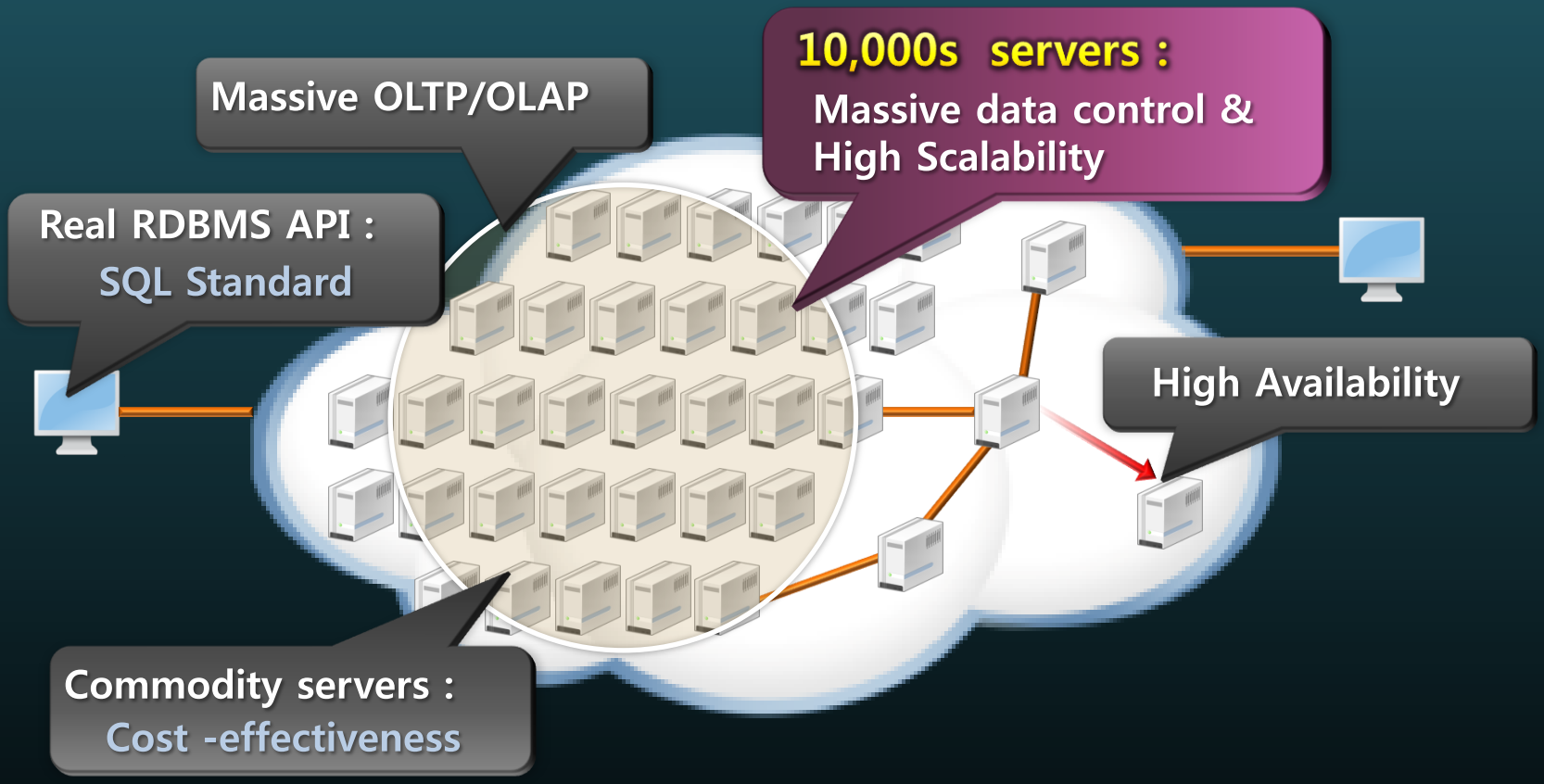


TIBERO

InfiniData 2.0 : Distribution & Single View Schema



Tibero InfiniData 2.0 : 주요 특징



InfiniData 2.0 : 분산 노드 관리 도구(Horde)

Monitoring

- Resource Monitoring
- Node Monitoring
- Session Monitoring

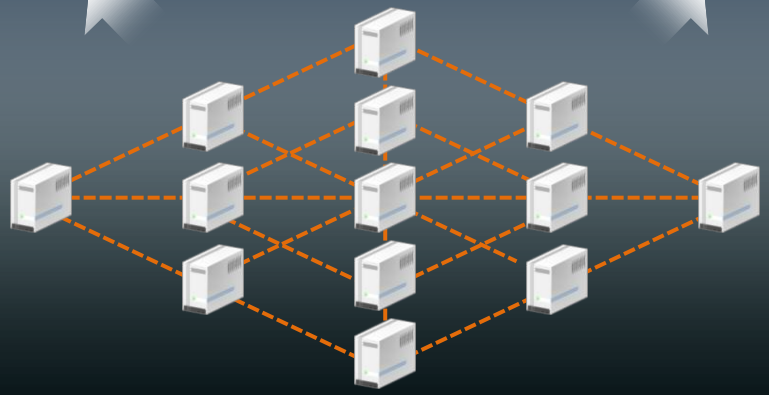
"Horde" Service Provider

Automatic Administration

Management

- Node Management
- Auto Mirroring
- Auto Recovery
- Load Balancing

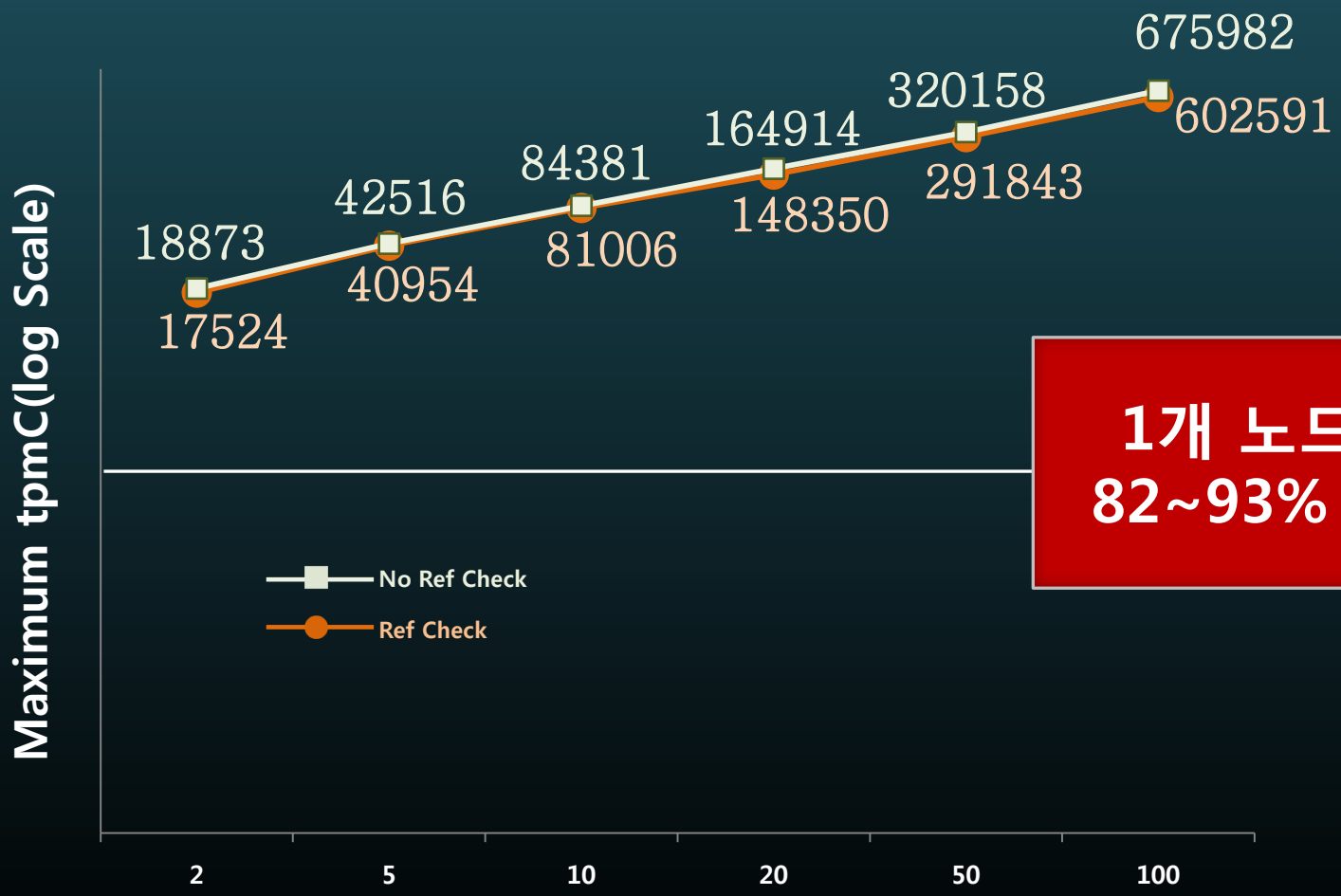
Node Membership & Resources



InfiniData 2.0 : 100개 노드 확장성 테스트(tpc-C)

High Scalability 제공

: InfiniData 100 Nodes에서 성능의 선형적 증가



1개 노드 확장 시
82~93% 성능 향상

TIBERO

Contents

1. 빅데이터 기술 현황 : Hadoop 중심
2. 기존 빅데이터 기술 : Infinidata 2.0
3. 빅데이터 기술의 진화 : Infinidata 3.0

InfiniData의 새로운 진화

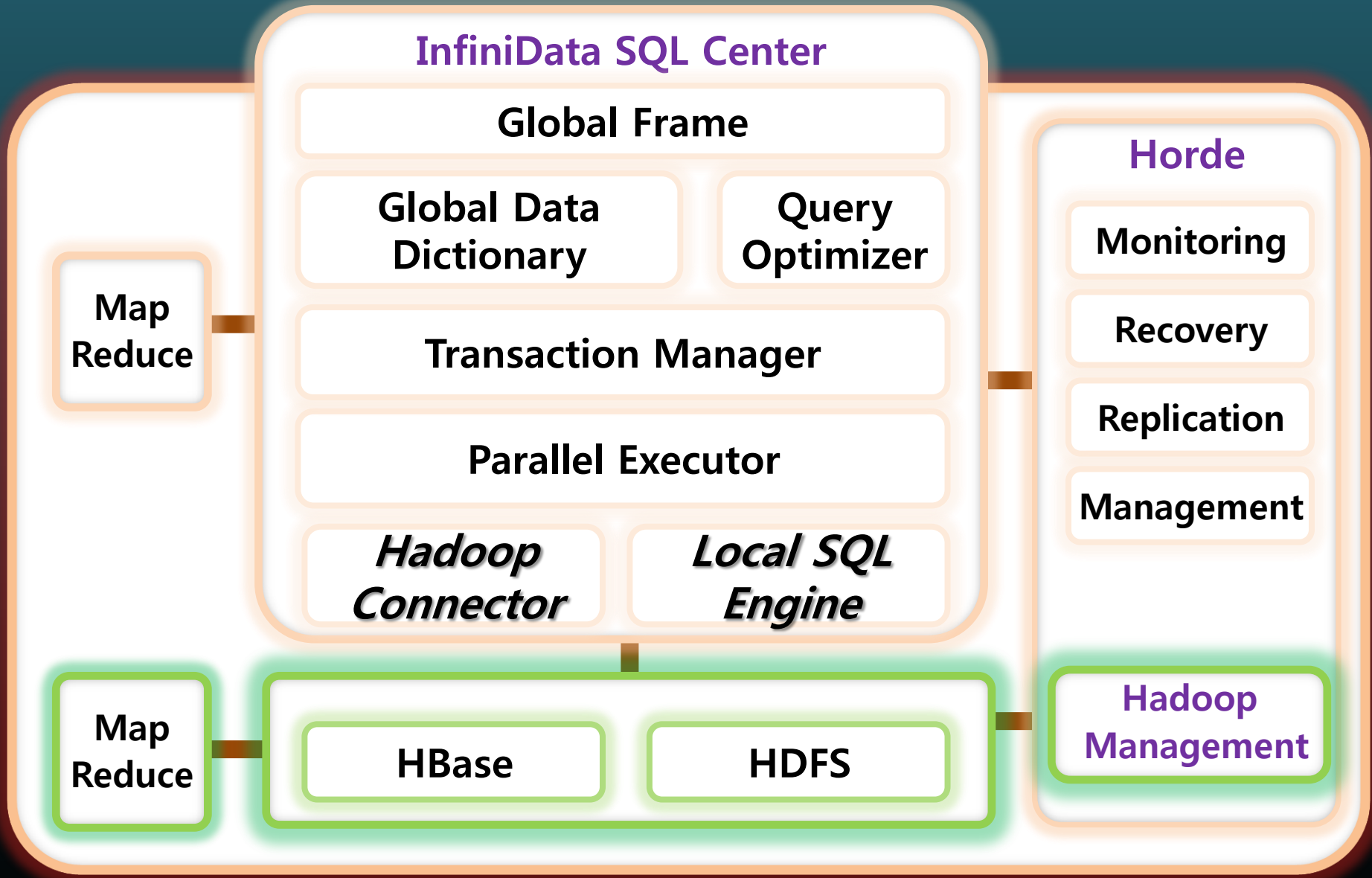
NoSQL	SQL
	InfiniData 2.0
반정형/비정형 데이터 분석	정형 데이터 분석
유연한 처리 (MapReduce)	편리한 분석(SQL)
다양하지만 일회적 배치 분석	범용적 대화식 분석(OLAP 연동)

- 저가 서버(x86)의 무한 Scale-out
- 가용성
- In-Memory기술을 이용한 분산 병렬 처리

Hadoop과 InfiniData 2.0 통합

InfiniData 3.0

InfiniData 3.0 : Architecture



InfiniData 3.0 : 주요 특징

- Hadoop 내장
- Smart Filtering
- Distributed Aggregation
- Local Join

비(반)정형 데이터 분석

통합분석

- 단일 쿼리로 정형/비정형 데이터의 연계 분석
- 비정형데이터(처리 결과)를 정형데이터로 로딩 용이
- Connector방식과 Embedded방식 지원

- Column 압축 지원
- Zip 압축 해제 없이 로딩 지원
- Table data Encryption
- Privilege/Role등을 통한 권한 관리

데이터 압축 및 보안

In-Memory 기술

- 사용자 환경에 맞는 DBMS Mode 변경
- Single View for Disk and Memory
- Automatic Data Re-location

InfiniData 3.0 : 통합분석 프로세스

Hadoop

①데이터 유형 변환

비정형 데이터→ [MapReduce] → 반정형 데이터

InfiniData SQL Center

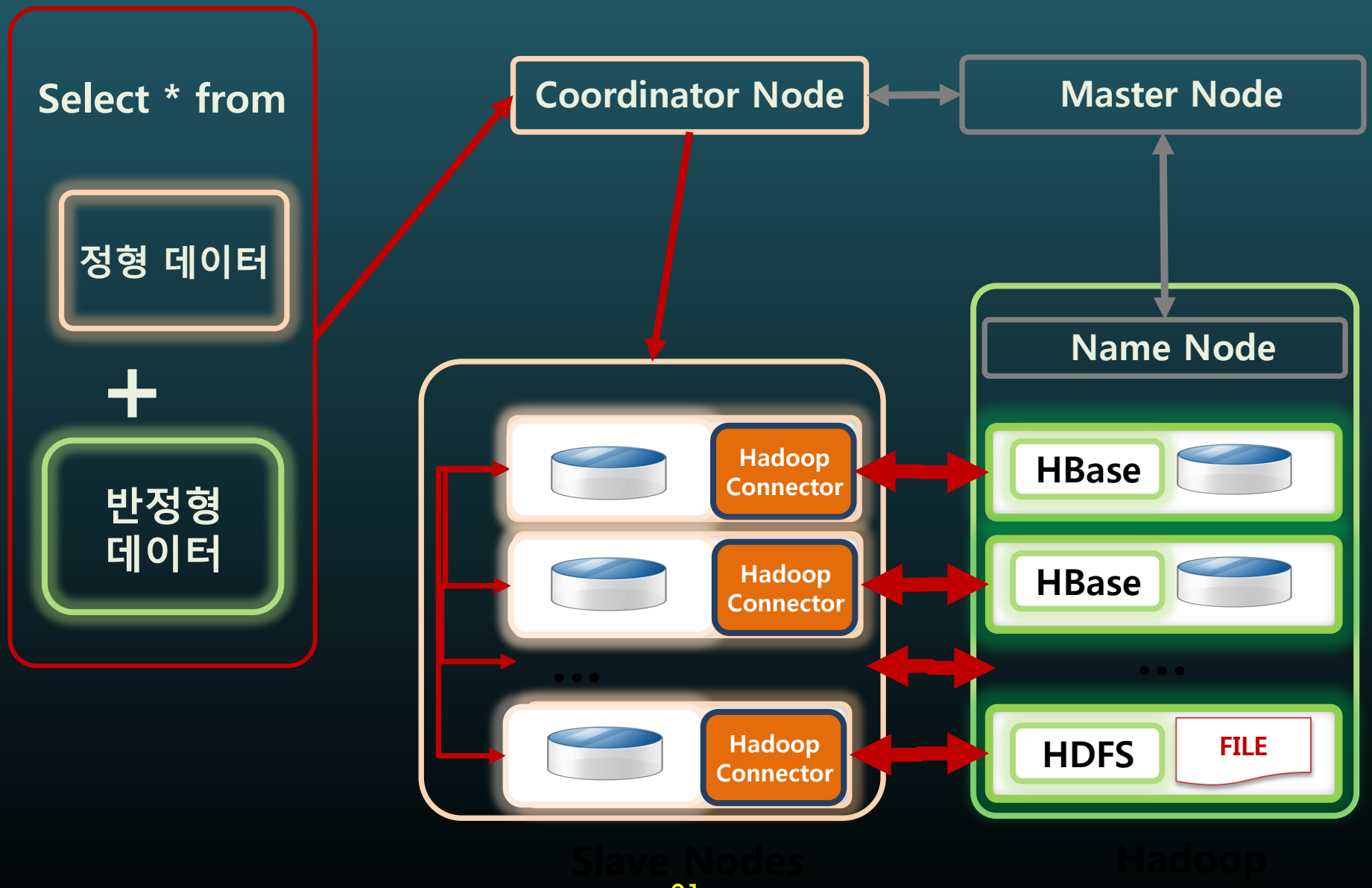
②Hadoop의 데이터 구조 정의

```
CREATE TABLE TO_CDR_201210_EXT_HDFS (
    a_switch_cd VARCHAR2(2), cmcc_branch_cd VARCHAR2(4), rollback_id CHAR(1),
    rollback_cnt NUMBER(5,0), usr_mob_nbr VARCHAR2(20), a_oper_cd VARCHAR2(15),
     ...,
    file_id VARCHAR2(10)
)
ORGANIZATION EXTERNAL
(
    a_switch_cd, cmcc_branch_cd, rollback_id, rollback_cnt, usr_mob_nbr, a_oper_cd,
    a_brnd_cd, a_usr_typ_cd, usr_pkg_typ, roam_typ_cd, visited_switch_cd, call_area_cd,
    call_dt, call_tim, hr_cd, debet_dt, tm_intrvl_cd, carry_typ_cd, cnct_nbr,
     ...,
    net_typ_cd, file_id
'hdfs://192.168.51.21:9010/user/tmc/data/CDR20121001.ZH.000003.pre.OUT',
'hdfs://192.168.51.21:9010/user/tmc/data/CDR20121001.ZH.000004.pre.OUT',
 ...,
'hdfs://192.168.51.21:9010/user/tmc/data/CDR20121001.ZH.000011.pre.OUT'
);
```

③통합 SQL 수행

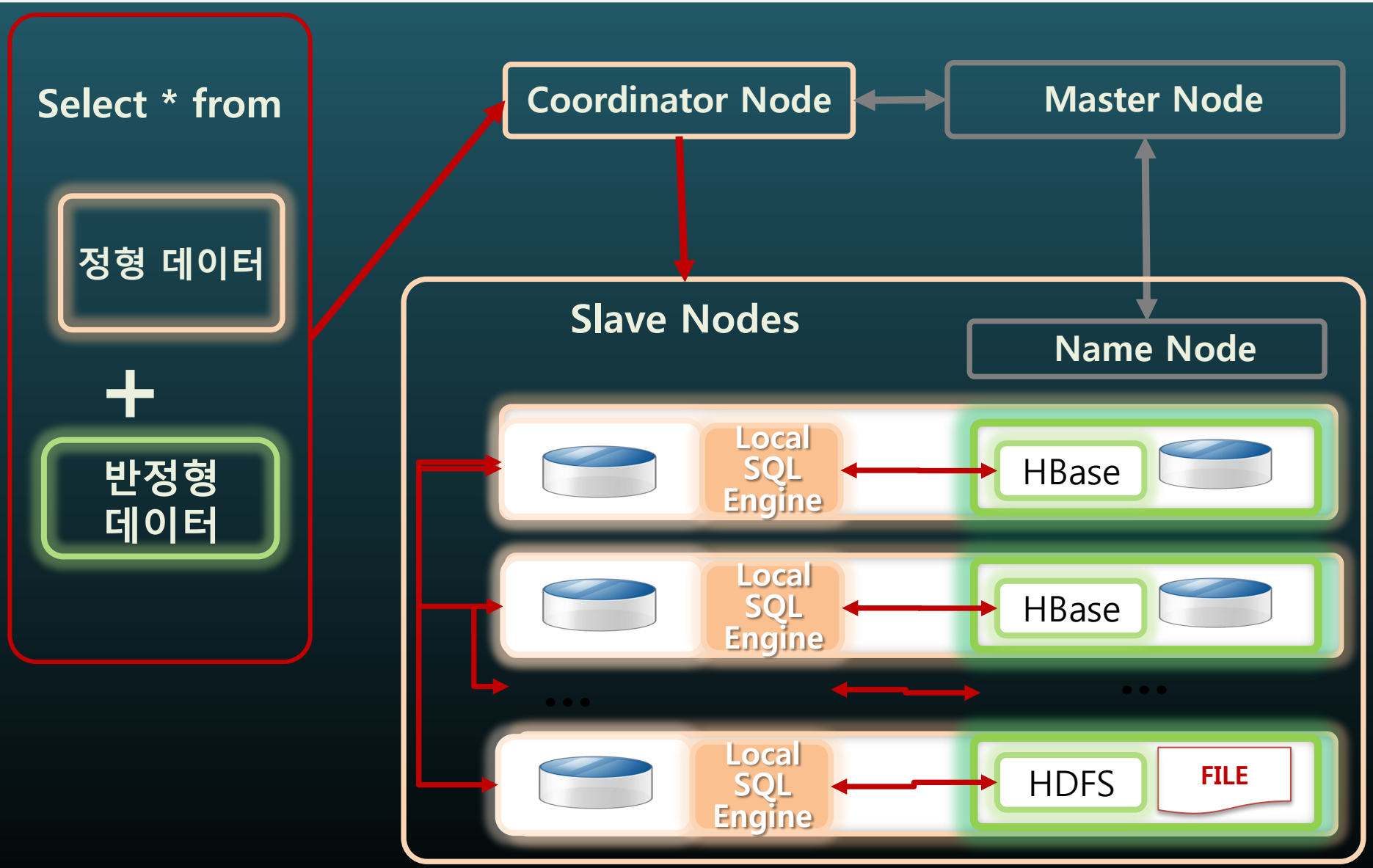
```
SELECT
        SUM(T1.CALL_CNT),
        SUM(T1.CALL_DUR),
        SUM(T1.CALL_FLUX),
        SUM(T1.CALL_FREE_DUR)
FROM    TO_CDR_201210_EXT_HDFS T1,
        TO_CDR_201210_TEST T2
WHERE   T1.USR_MOB_NBR =T2.USR_MOB_N
        T1.SRV_TYP_CD LIKE '0%'
        AND T2.DIR_TYP_CD NOT
```

InfiniData 3.0 : 통합분석 - Connector 방식

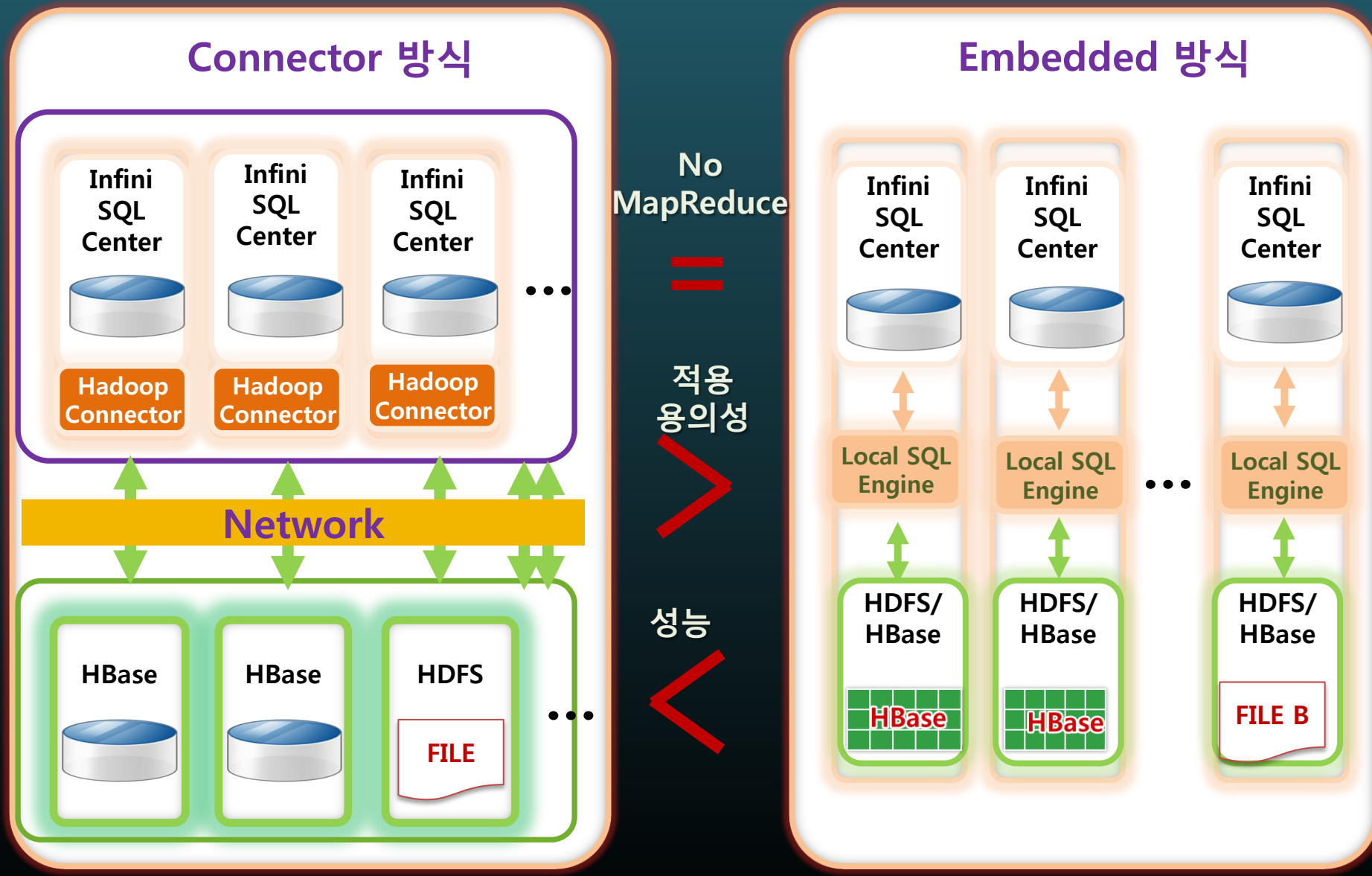


TIBERO

InfiniData 3.0 : 통합분석 - Embedded 방식



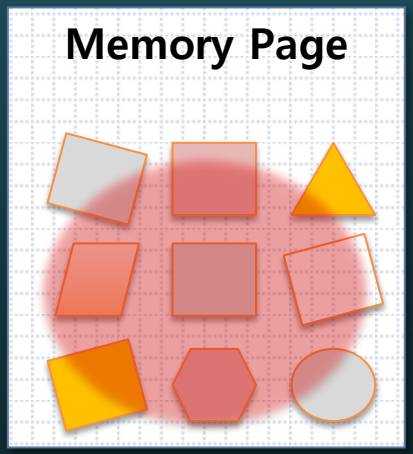
InfiniData 3.0 : Hadoop 연동방식 특징



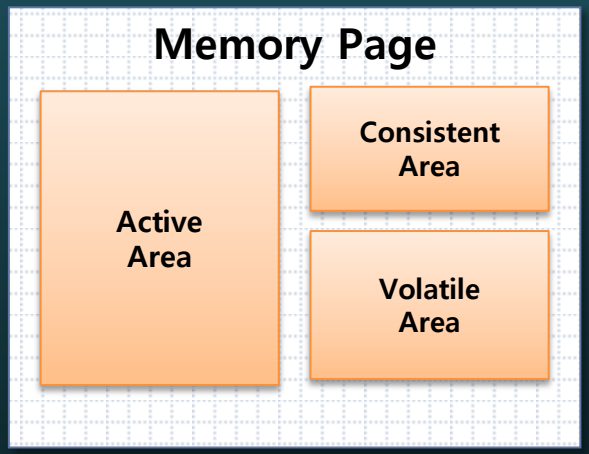
InfiniData 3.0 : In-Memory Technology

- 비효율적인 메모리 사용
- Garbage Collection 필요
- Disk IO 많음

일반적인 In-Memory DBMS



Tibero 6 In-Memory DBMS

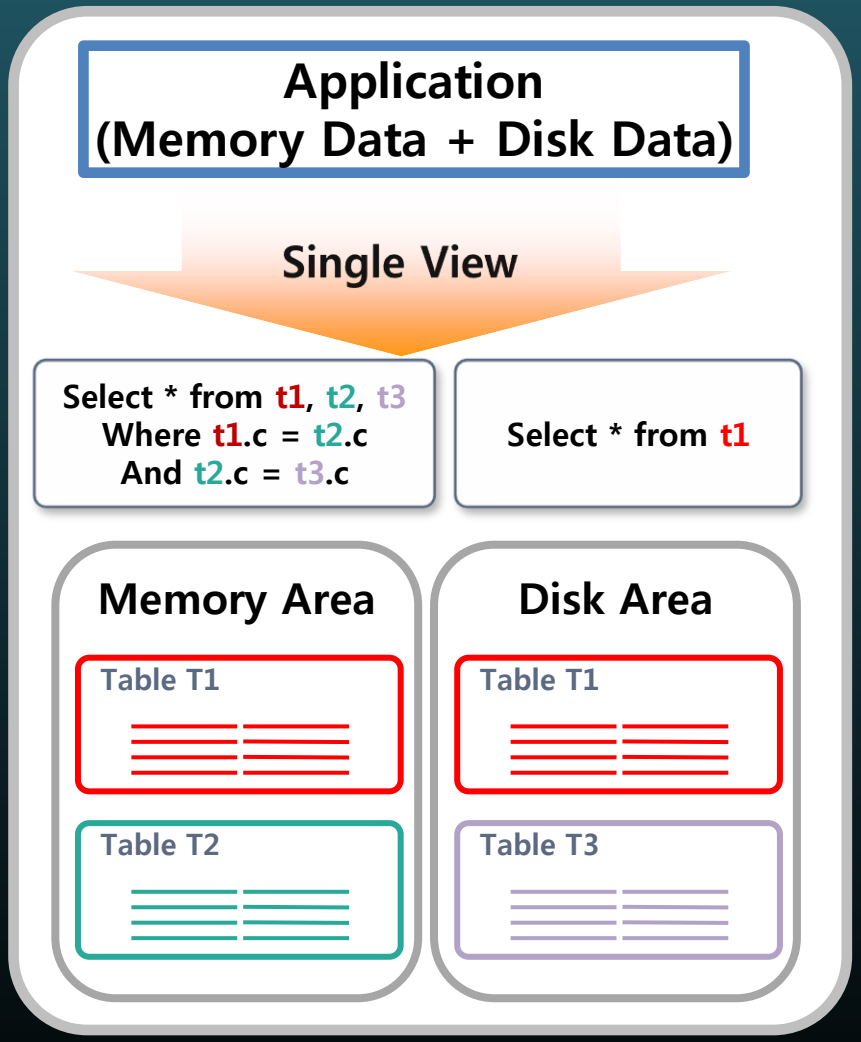


- Policy에 따른 Consistent Area 관리
- Garbage Collection 불필요
- Disk IO 감소

사용자 환경에 맞는 DBMS Mode 변경

1. Only In-Memory DBMS 옵션
2. Only Disk DBMS 옵션
3. 자동관리 옵션
4. 테이블별 설정 옵션

InfiniData 3.0 : Single View for Disk and Memory

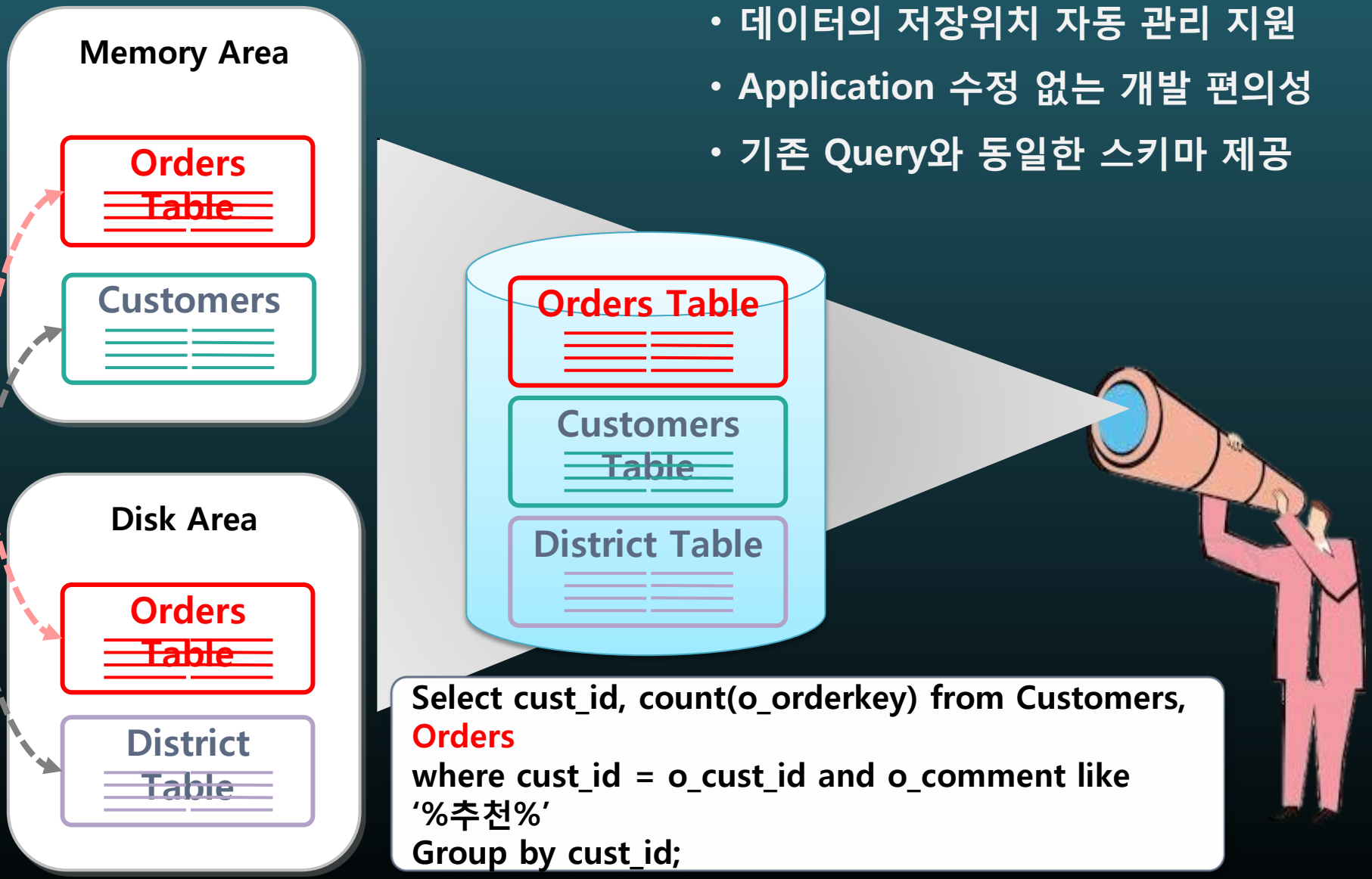


Single View 핵심 기술

- Location Partition
  - ➢ 데이터 위치를 기준으로 한 internal partition
- Location Aware Query Processing
  - ➢ 저장 장소 별로 최적화된 data access plan

InfiniData 3.0 : Location Aware Query Optimization

- 데이터의 저장위치 자동 관리 지원
- Application 수정 없는 개발 편의성
- 기존 Query와 동일한 스키마 제공




```
Select cust_id, count(o_orderkey) from Customers,
Orders
where cust_id = o_cust_id and o_comment like
'%추천%'
Group by cust_id;
```


TIBERO

InfiniData 3.0 : Automatic Data Re-location

사용 빈도 추적

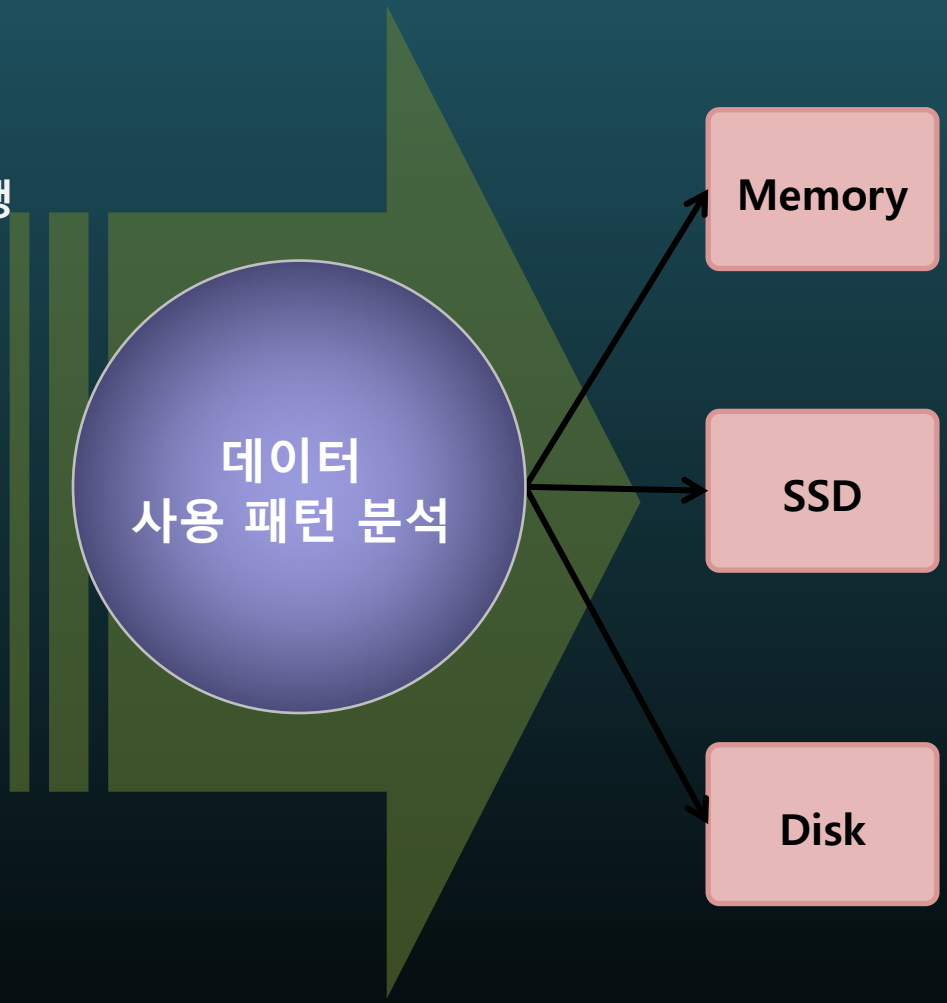
- 조회/수정이 빈번함
- 조회/수정이 가끔 발생
- 조회/수정 거의 없음

업무 성격 지정

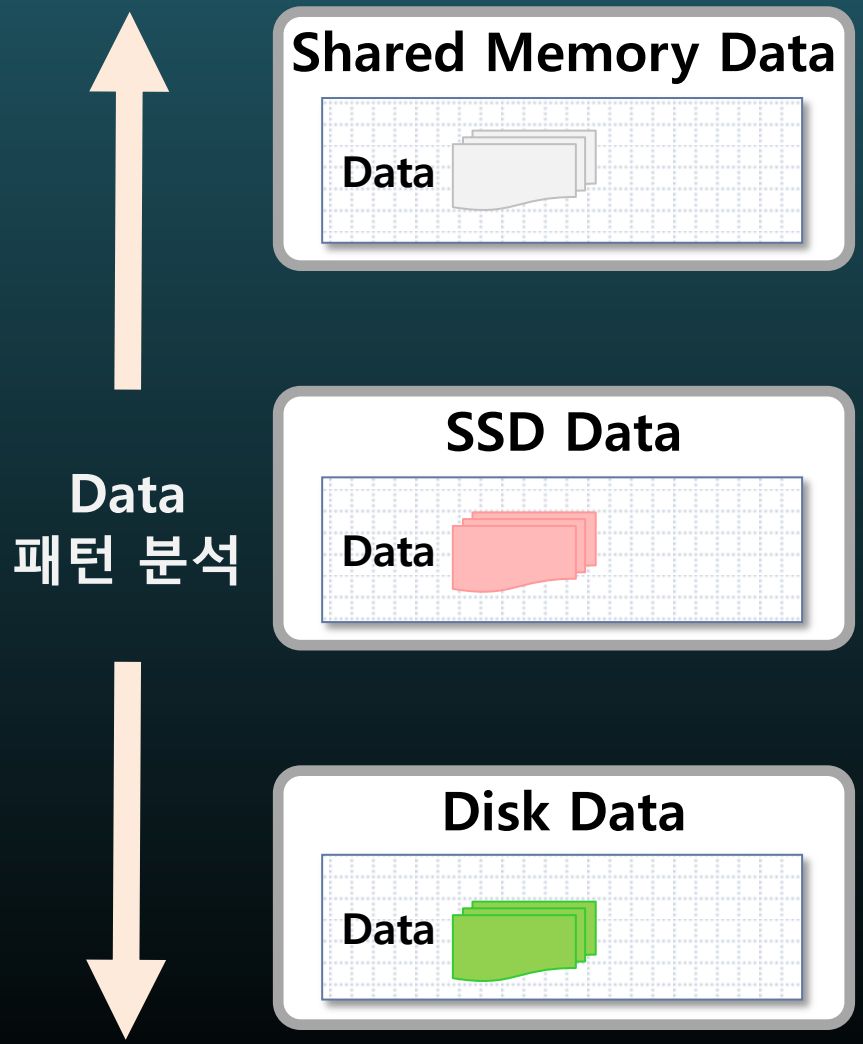
- 실시간 OLTP
- 대규모 분석 업무
- 로그 성격의 데이터

데이터 Age 추적

- 최근에 생성됨
- 생성 후 1분기 경과
- 생성 후 2분기 경과



InfiniData 3.0 : In-Memory 데이터 패턴 분석

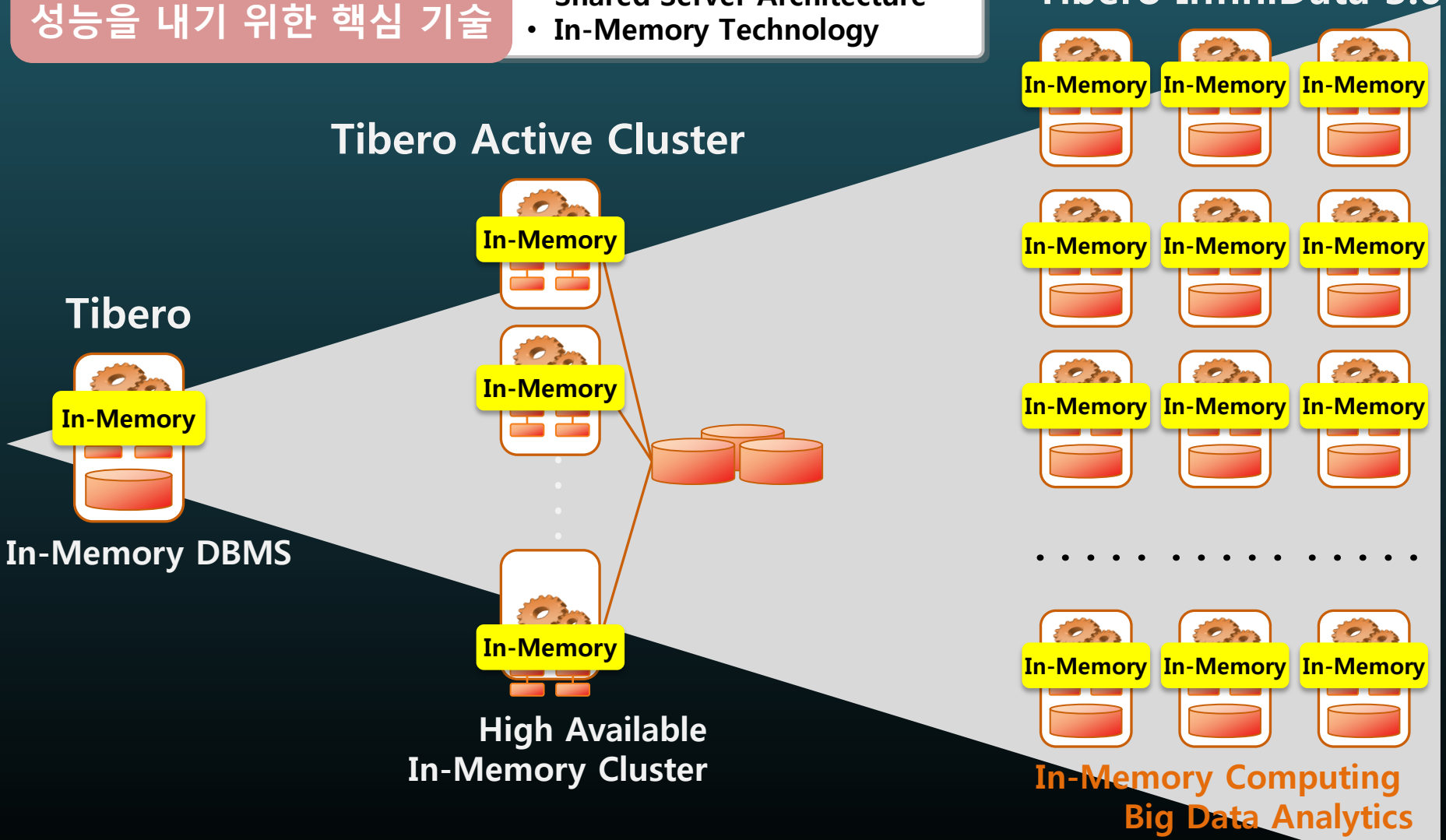


- Data Usage Tracking
  - ➢데이터 사용패턴 추적
  - ➢빈번하게 접근되는 데이터 사용기록 유지
- Data Age Tracking
  - ➢데이터 생성, 변경 이력 추적
- Automatic Data Transfer
  - ➢데이터 사용패턴 및 보관 정책에 따라 memory 또는 disk로 데이터 자동 이관
- 데이터 패턴 분석
  - ➢사용빈도, 업무성격, Age를 종합적으로 분석하여 최적 데이터 위치 자동 결정

InfiniData 3.0 : In-Memory를 활용한 Big Data 처리

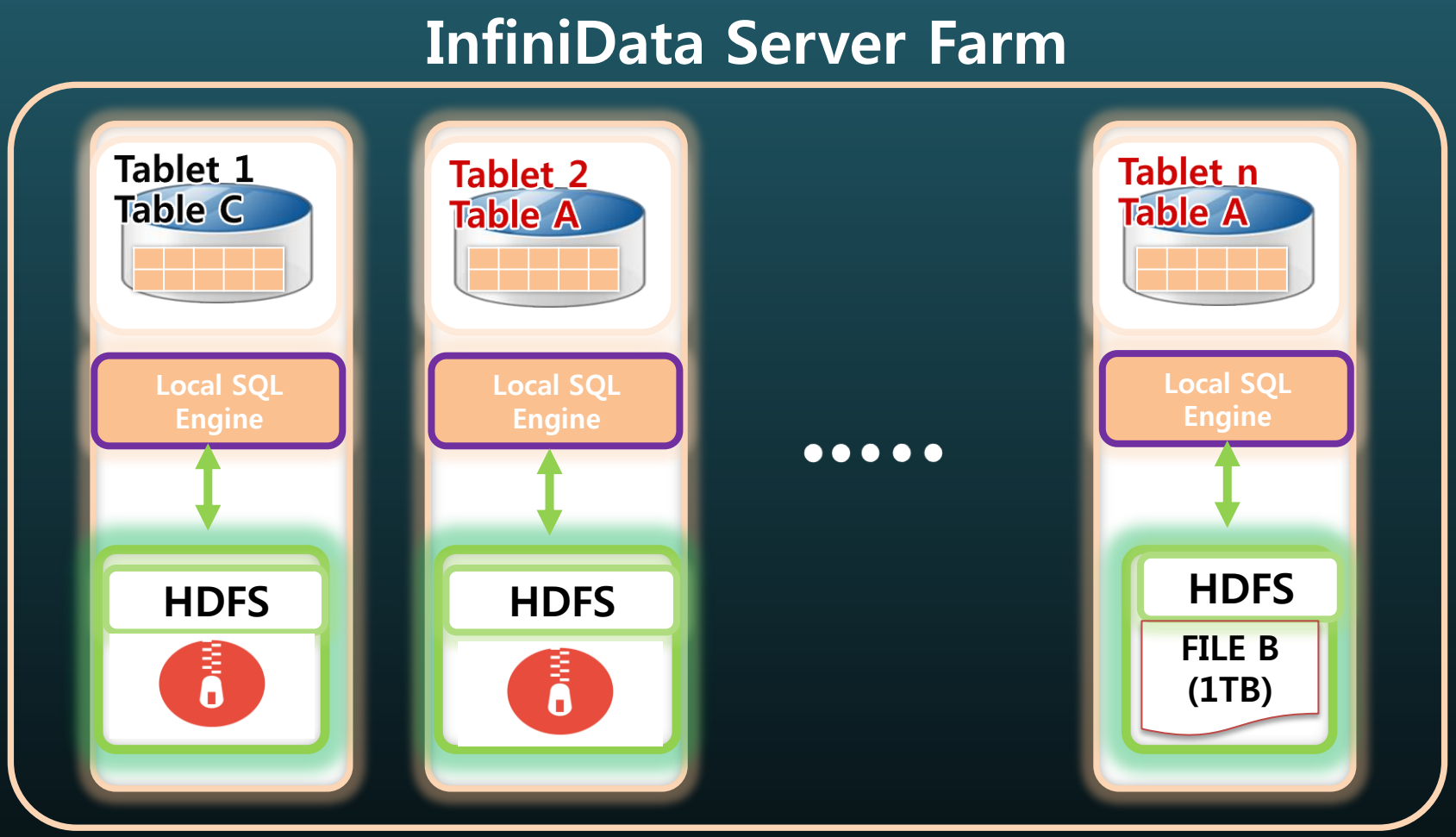
대규모 트랜잭션에서 탁월한 성능을 내기 위한 핵심 기술

- Hyper Thread Architecture
- Shared Server Architecture
- In-Memory Technology



TIBERO

InfiniData 3.0 : 압축 지원

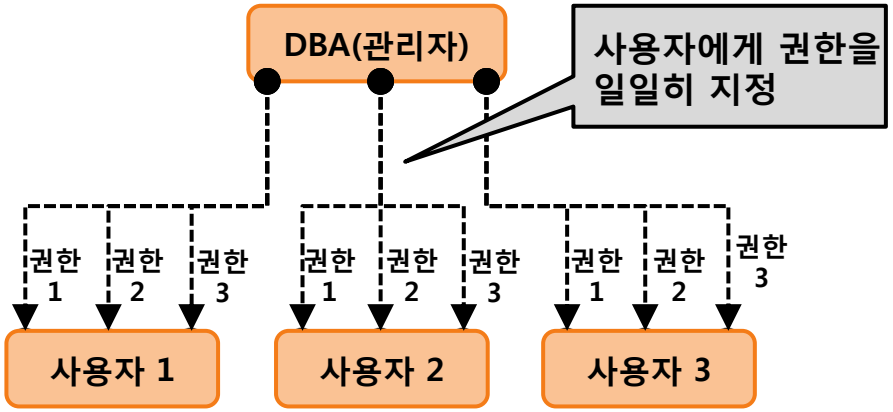


- Zip 압축 파일 해제 없이 Direct Loading 사용
- Column 압축을 통한 고밀도 압축 기능 지원

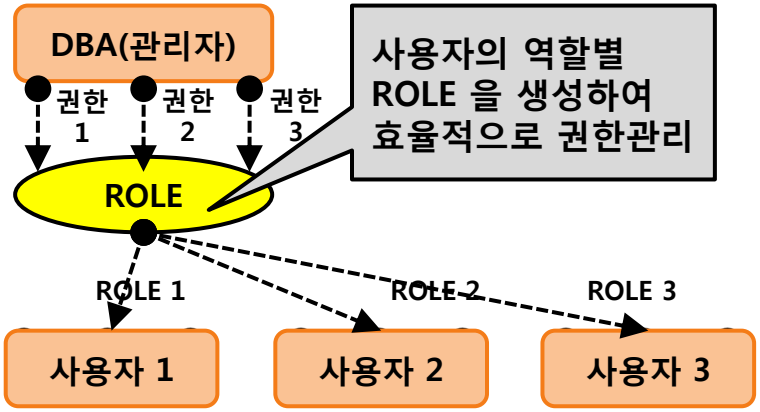
InfiniData 3.0 : 보안 기능 제공

접근권한 제어

ROLE 사용하지 않은 권한부여

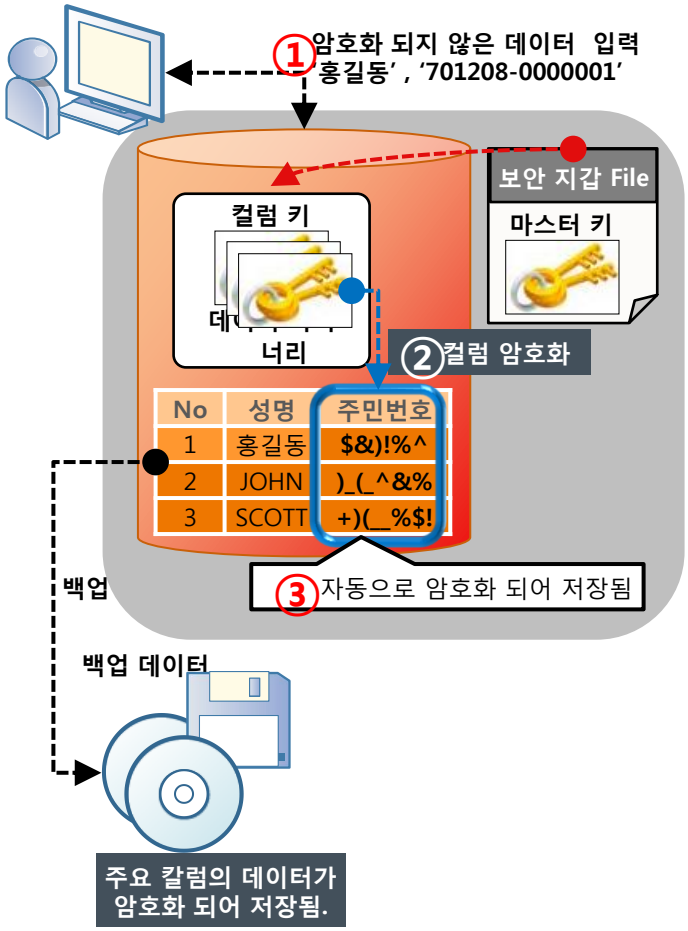


ROLE 사용한 권한 부여



암호화 지원

Transparent Data Encryption

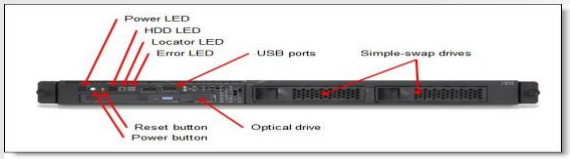


TIBERO

InfiniData 3.0 : 개발/테스트 환경

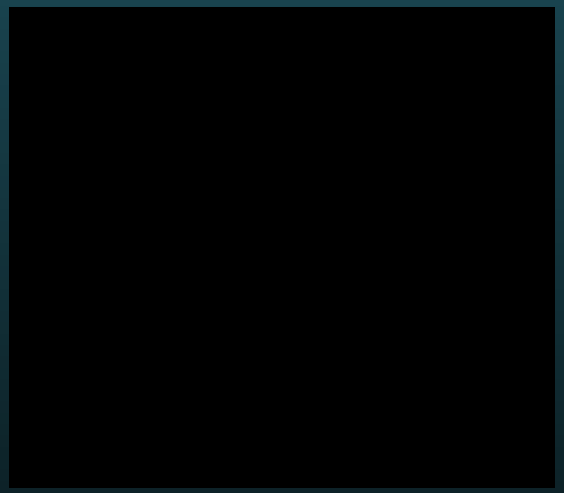
서버 사양

- Model : IBM X3250M4
- OS : Ubuntu 11.10 x86_64
- CPU : 3.10GHz , 4 Core
- Memory : 8G
- Disk : 1T * 2EA



Server Farm 구성

InfiniData 3.0 : 100대 Cluster 관리기능 시연



1. 분산노드관리도구 Horde 접속 : *hrdcmd*

2. cluster 구성 물리적 서버 조회 : *show machine*
→ 서버ID, CPU/메모리/읽기/쓰기 상태 표시

3. DBMS cluster의 논리적 서버(instance) 조회 : *show cluster이름*

→Active Instance의 상태정보 : 전체 Instance 서버수(100대), Instance id, 서버 id, 가동상태(OFF), 메모리상태
→Standby Instance의 상태정보

4. cluster의 모든 instance 부팅 : *hrdt -a boot*
→ 100개 Active Instance와 100개 standby Instance 부팅

5. DBMS cluster의 논리적 서버(instance) 조회 : *show cluster이름*
→Active Instance 100대, Standby Instance 100 대 상태 정보 : 가동상태(ON)

InfiniData 3.0 : Connector 방식 시연 - 시나리오

구성 환경

- Hadoop 10개 노드 + Infinidata 10개 노드
- Hadoop 버전 : 1.2.1

시나리오개요

Hadoop(HDFS)에 분산 저장된 반정형 데이터 파일을
InfiniData Connector을 이용해
InfiniData SQL Center에서 “Count All Records” 수행

선행 작업

- Hadoop(HDFS) 데이터 파일 의 메타정보 생성

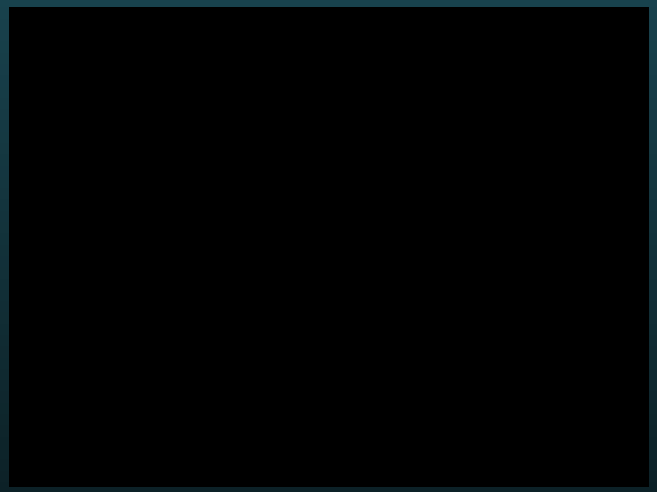
→ Hadoop(HDFS)의 데이터 파일을 InfiniData SQL Center에서 테이블 형식으로 정의

```
CREATE TABLE TO_CDR_201210_EXT_HDFS (
    a_switch_cd VARCHAR2(2), cmcc_branch_cd VARCHAR2(4), rollback_id CHAR(1),
    rollback_cnt NUMBER(5,0), usr_mob_nbr VARCHAR2(20), a_oper_cd VARCHAR2(15),
    a_brnd_cd NUMBER(5,0), a_usr_typ_cd VARCHAR2(2), usr_pkg_typ CHAR(1),
    roam_typ_cd NUMBER(5,0), visited_switch_cd VARCHAR2(2), call_area_cd VARCHAR2(8),
    call_dt NUMBER(10,0), call_tim VARCHAR2(6), hr_cd NUMBER(5,0),
    debet_dt NUMBER(10,0), tm_intrvl_cd NUMBER(10,0), carry_typ_cd VARCHAR2(2),
    cnct_nbr VARCHAR2(8), toll_typ_cd NUMBER(5,0), dir_typ_cd VARCHAR2(2),
    srv_typ_cd VARCHAR2(2), subsrv_typ_cd VARCHAR2(2), rate_typ_cd CHAR(1),
    b_nbr VARCHAR2(20), b_switch_cd VARCHAR2(2), b_oper_cd VARCHAR2(15),
    b_brnd_cd NUMBER(5,0), b_usr_typ_cd VARCHAR2(2), b_area_cd VARCHAR2(8),
    call_dur NUMBER(10,0), toll_dur NUMBER(10,0), call_flux NUMBER(10,0),
    call_cnt NUMBER(10,0), mob_fee NUMBER(7,2), toll_fee NUMBER(7,2),
    inf_fee NUMBER(7,2), imsi VARCHAR2(15), local_cd VARCHAR2(4),
    cell_cd VARCHAR2(4), out_route_cd VARCHAR2(7), in_route_cd VARCHAR2(7),
    route_typ CHAR(1), switch_nbr VARCHAR2(4), roam_nbr VARCHAR2(20),
    imei VARCHAR2(15), bill_typ_cd VARCHAR2(3), disc_cd VARCHAR2(20),
    vpmn_ind CHAR(1), disc_mob_fee NUMBER(7,2), disc_toll_fee NUMBER(7,2),
    disc_inf_fee NUMBER(7,2), disc_fee NUMBER(7,2), call_free_dur NUMBER(10,0),
    acct_bal NUMBER(10,2), icp_stat_cd VARCHAR2(2), load_id NUMBER(10,0),
    file_typ VARCHAR2(2), cx_file_ind CHAR(1), net_typ_cd VARCHAR2(2),
    file_id VARCHAR2(10)
```

```
ORGANIZATION EXTERNAL
(
        DEFAULT DIRECTORY _extld_tmp_dir
        ACCESS PARAMETERS
        (
                LOAD DATA CHARACTERSET EUCKR
                INTO TABLE TO_CDR_201210_EXT_HDFS
                FIELDS TERMINATED BY "|"
                (
    a_switch_cd, cmcc_branch_cd, rollback_id, rollback_cnt, usr_mob_nbr, a_oper_cd,
    a_brnd_cd, a_usr_typ_cd, usr_pkg_typ, roam_typ_cd, visited_switch_cd, call_area_cd,
    call_dt, call_tim, hr_cd, debet_dt, tm_intrvl_cd, carry_typ_cd, cnct_nbr,
    toll_typ_cd, dir_typ_cd, srv_typ_cd, subsrv_typ_cd, rate_typ_cd, b_nbr,
    b_switch_cd, b_oper_cd, b_brnd_cd, b_usr_typ_cd, b_area_cd, call_dur,
    toll_dur, call_flux, call_cnt, mob_fee, toll_fee, inf_fee, imsi, local_cd,
    cell_cd, out_route_cd, in_route_cd, route_typ, switch_nbr, roam_nbr, imei,
    bill_typ_cd, disc_cd, vpmn_ind, disc_mob_fee, disc_toll_fee, disc_inf_fee,
    disc_fee, call_free_dur, acct_bal, icp_stat_cd, load_id, file_typ, cx_file_ind,
    net_typ_cd, file_id
                )
        )
        LOCATION
        (
'hdfs://192.168.51.21:9010/user/tmc/data/CDR20121003.ZH_000001.pre.OUT',
'hdfs://192.168.51.21:9010/user/tmc/data/CDR20121003.ZH_000002.pre.OUT',
'hdfs://192.168.51.21:9010/user/tmc/data/CDR20121003.ZH_000003.pre.OUT',
'hdfs://192.168.51.21:9010/user/tmc/data/CDR20121003.ZH_000004.pre.OUT',
'hdfs://192.168.51.21:9010/user/tmc/data/CDR20121003.ZH_000005.pre.OUT',
'hdfs://192.168.51.21:9010/user/tmc/data/CDR20121003.ZH_000006.pre.OUT',
'hdfs://192.168.51.21:9010/user/tmc/data/CDR20121003.ZH_000007.pre.OUT',
```

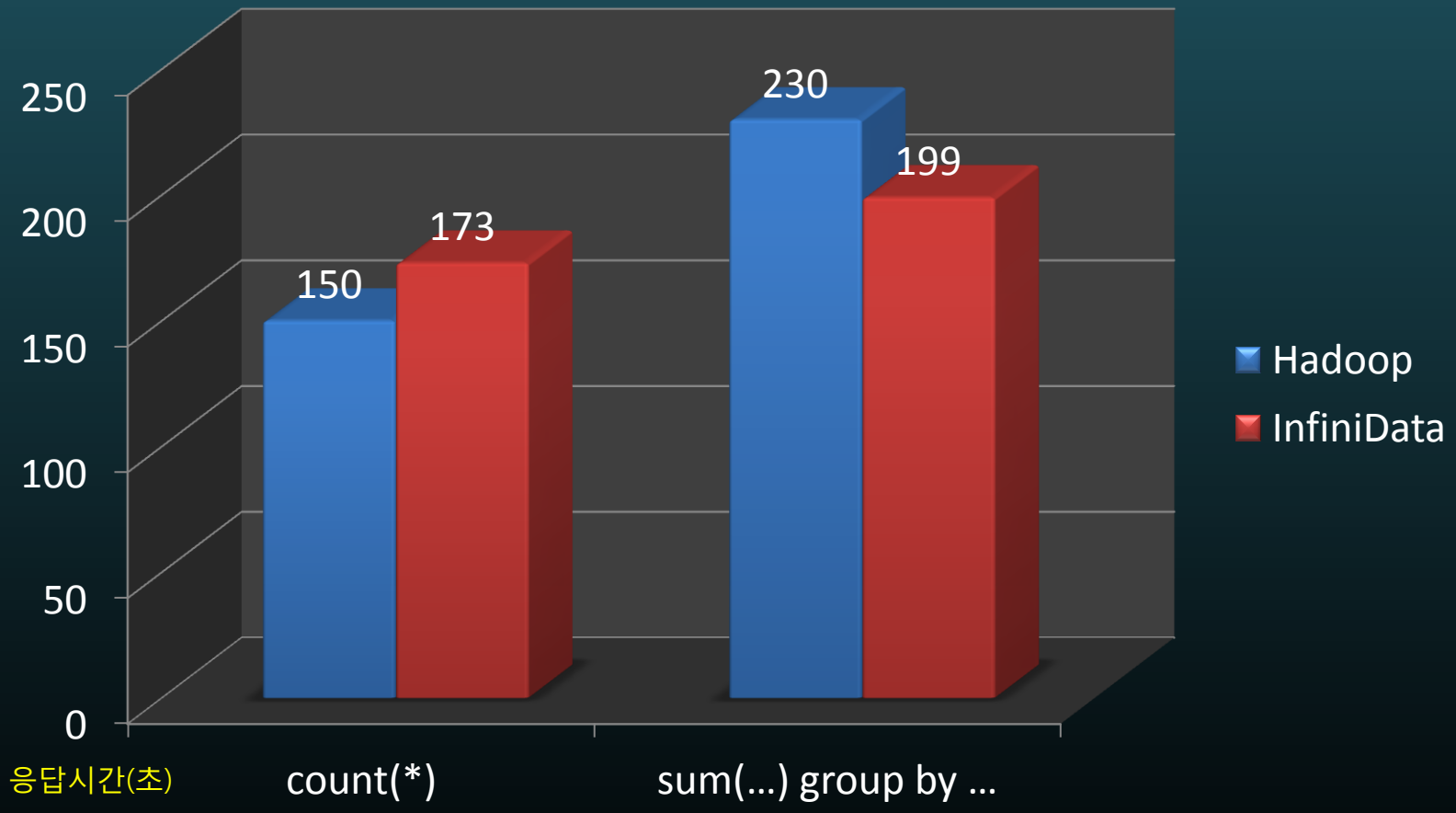
HIBERO

InfiniData 3.0 : Connector 방식 시연



InfiniData 3.0 : InfiniData와 Hadoop 성능 비교

총 2.5억건 데이터를 대상으로 2종류 Query의 응답시간(초)을 제품별로 비교



TIBERO

InfiniData 3.0 : Hadoop-InfiniData의 통합분석 시연-시나리오

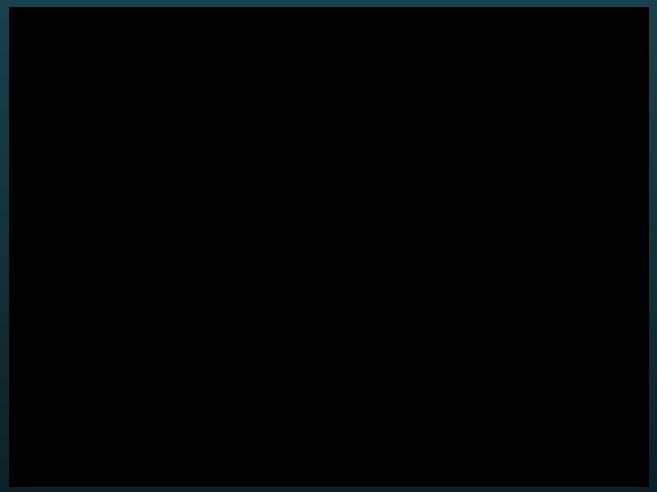
Hadoop의 분산 파일
'TO_CDR_201210_EXT_HDFS'

JOIN & SUM

InfiniData의 분산 테이블
'BASS_USER_MO_201210'

```
SELECT
        SUM(T1.CALL_CNT),
        SUM(T1.CALL_DUR),
        SUM(T1.CALL_FLUX),
        SUM(T1.CALL_FREE_DUR)
FROM
        TO_CDR_201210_EXT_HDFS T1,
        BASS_USR_MO_201210 T2
WHERE
        T1.USR_MOB_NBR = T2.USR_NBR AND
        T2.BRND_CD NOT IN('5','6','7','8')
```

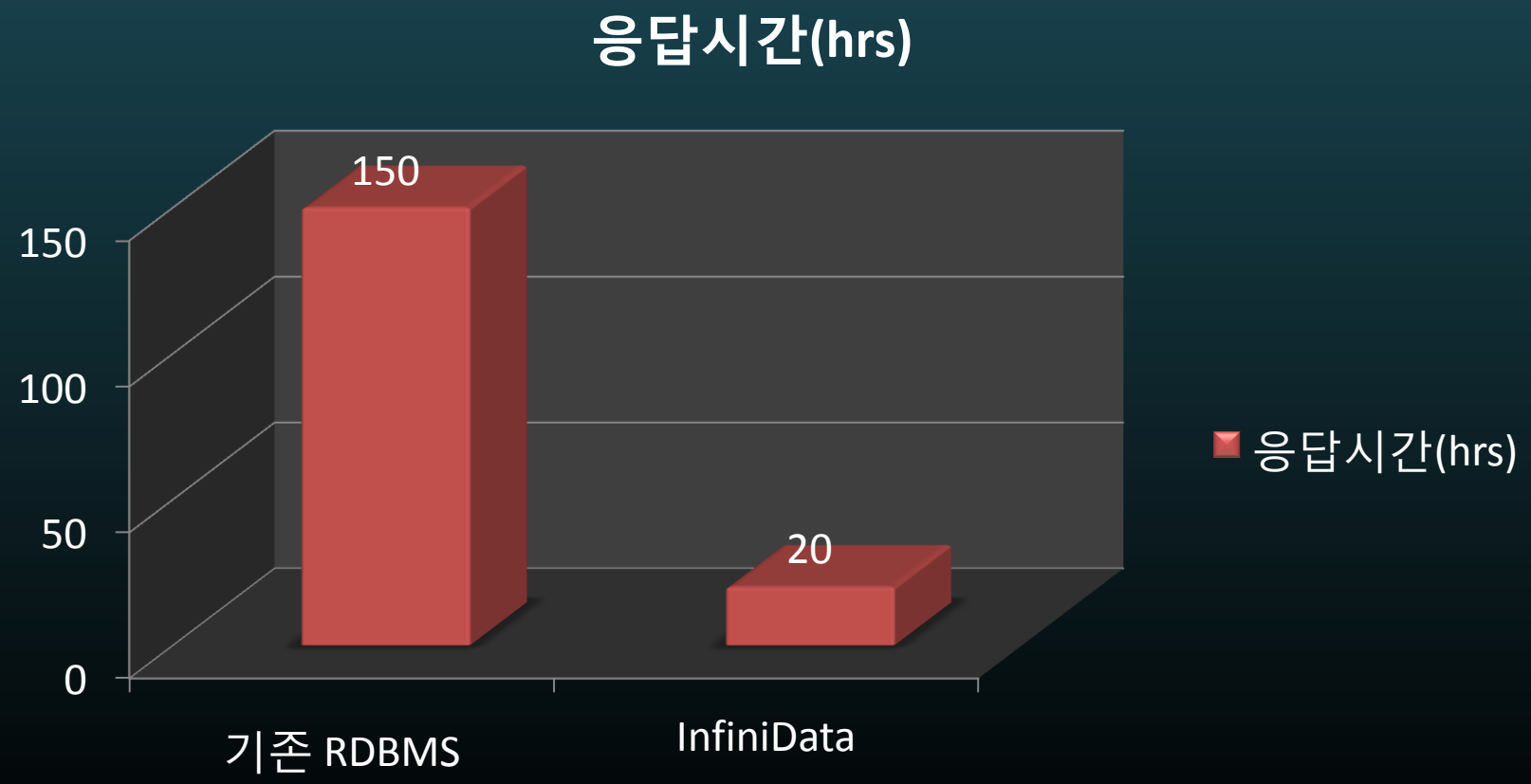
InfiniData 3.0 : Hadoop-InfiniData의 통합분석 시연



W 온라인 게임회사 : 고객행동패턴 분석 사례(1/2)

- 월별 잔존율 분석

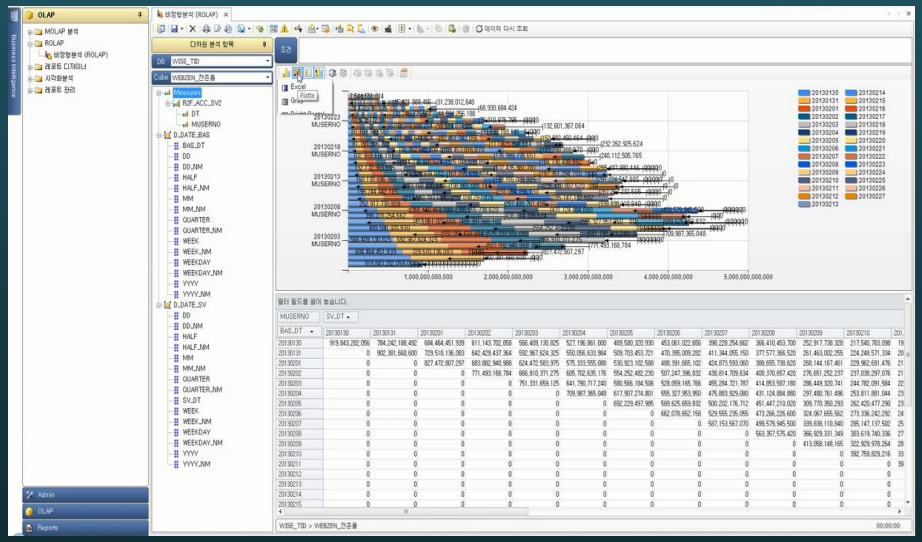
매월 1일 10만명의 접속자중, 다음날에도 접속한 사용자를 찾아내어 게임에
하루도 쉬지 않고 계속 접속하는 충성도 높은 사용자 그룹을 찾는 것(마케팅활동에 적용)

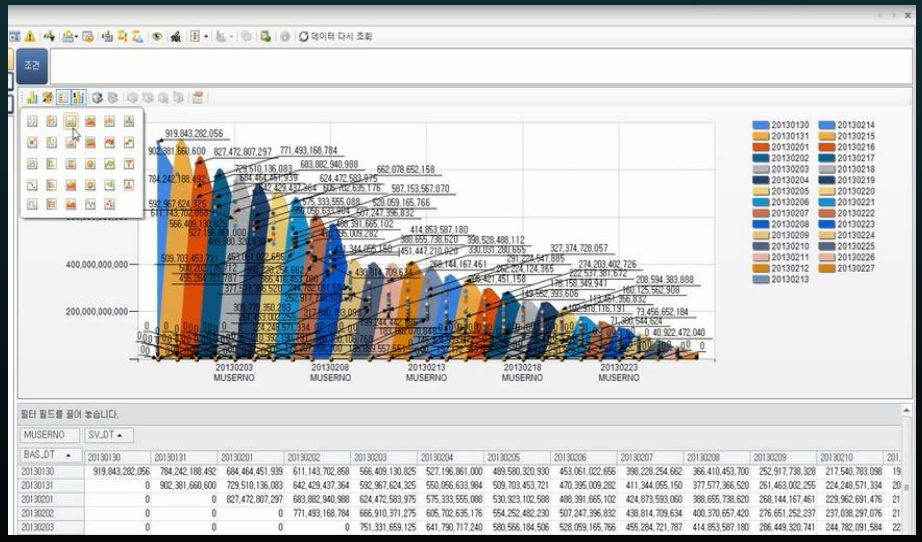


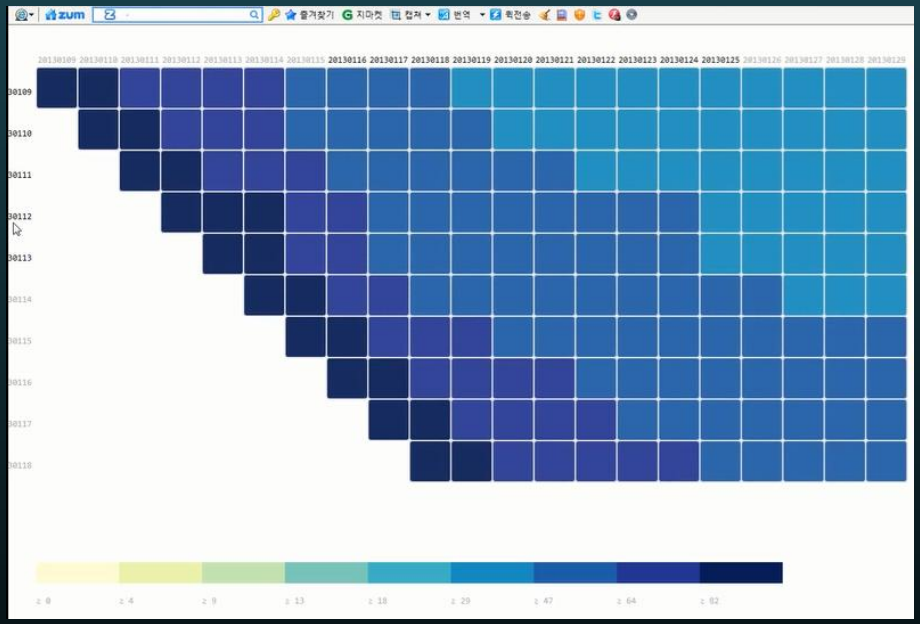
TIBERO

W 온라인 게임회사 : 고객행동패턴 분석 사례(2/2)

- 분석결과의 OLAP(WISE OLAP) 활용







TIBERO

Thank you!