진화하는 데이터웨어하우스(DW): DW 및 분석 어플라이언스의 미래

한국IBM 소프트웨어그룹, 정보관리사업부(Information Management)
김도윤 실장 (dyunkim@kr.ibm.com)

목차

- 왜 Big Data인가?
- Big Data와 Big Insight
- Big Insight를 위한 DW 및 분석 어플라이언스
  - Netezza Technologies
  - Netezza Advanced Analytics
  - Netezza Customers
- 진화하는 DW 및 분석 어플라이언스의 방향

목차

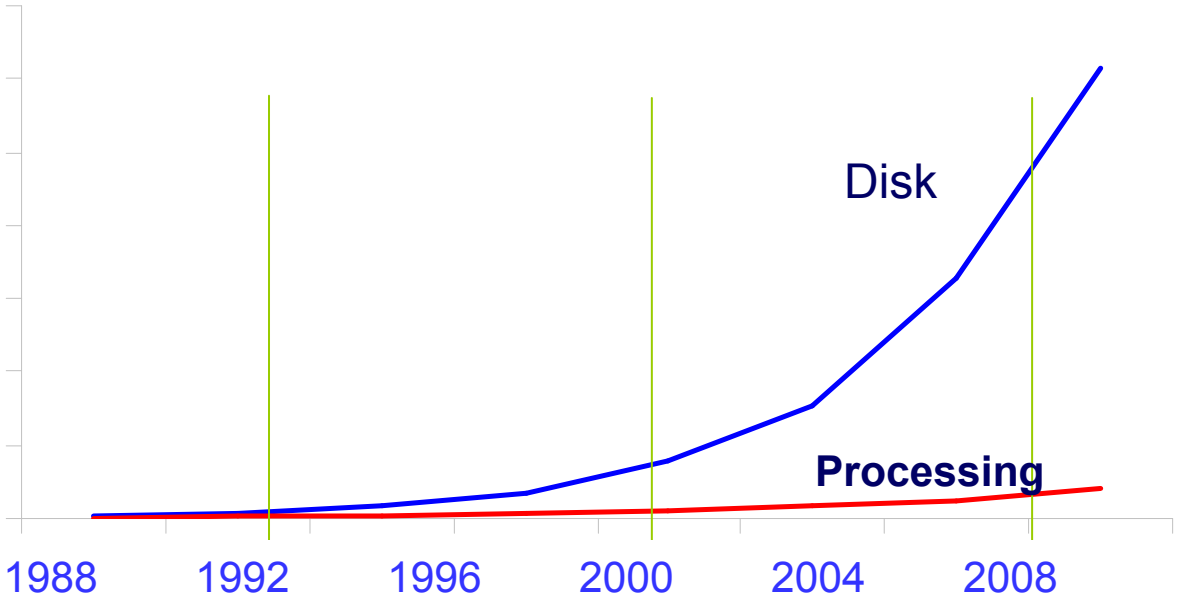
❖왜 Big Data인가?

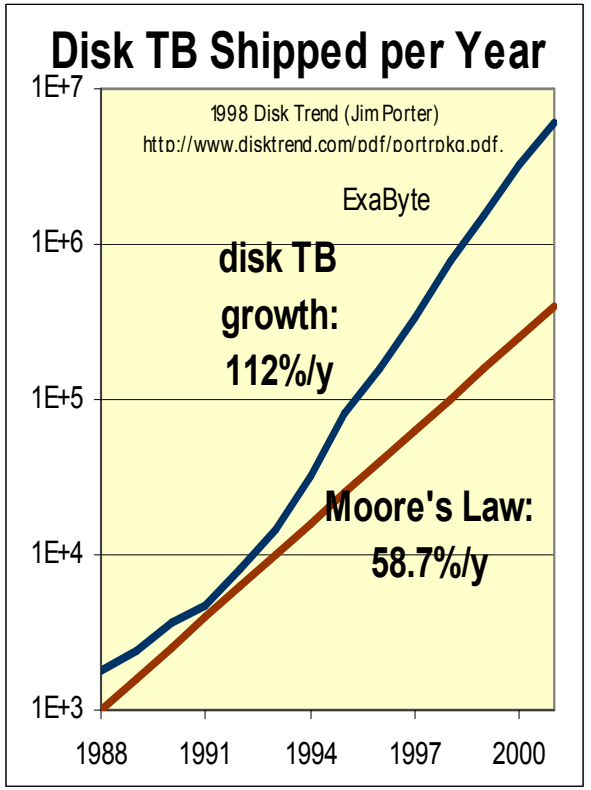
The BIG data gap…

Moore's law: processing "capacity" doubles every 18 months : CPU, cache, memory

It's more aggressive cousin:

- **Disk storage "capacity" doubles every 9 months**

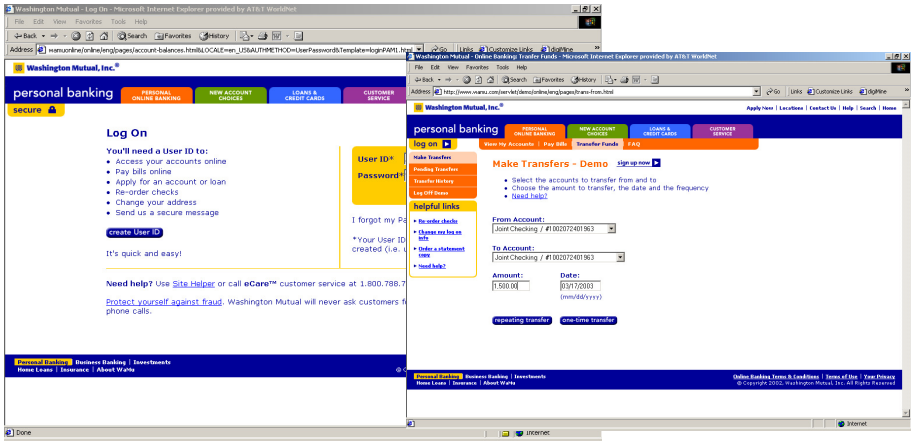




2000's Big Data from Online Transaction

Account Profile

Online Banking,Bill Pay,

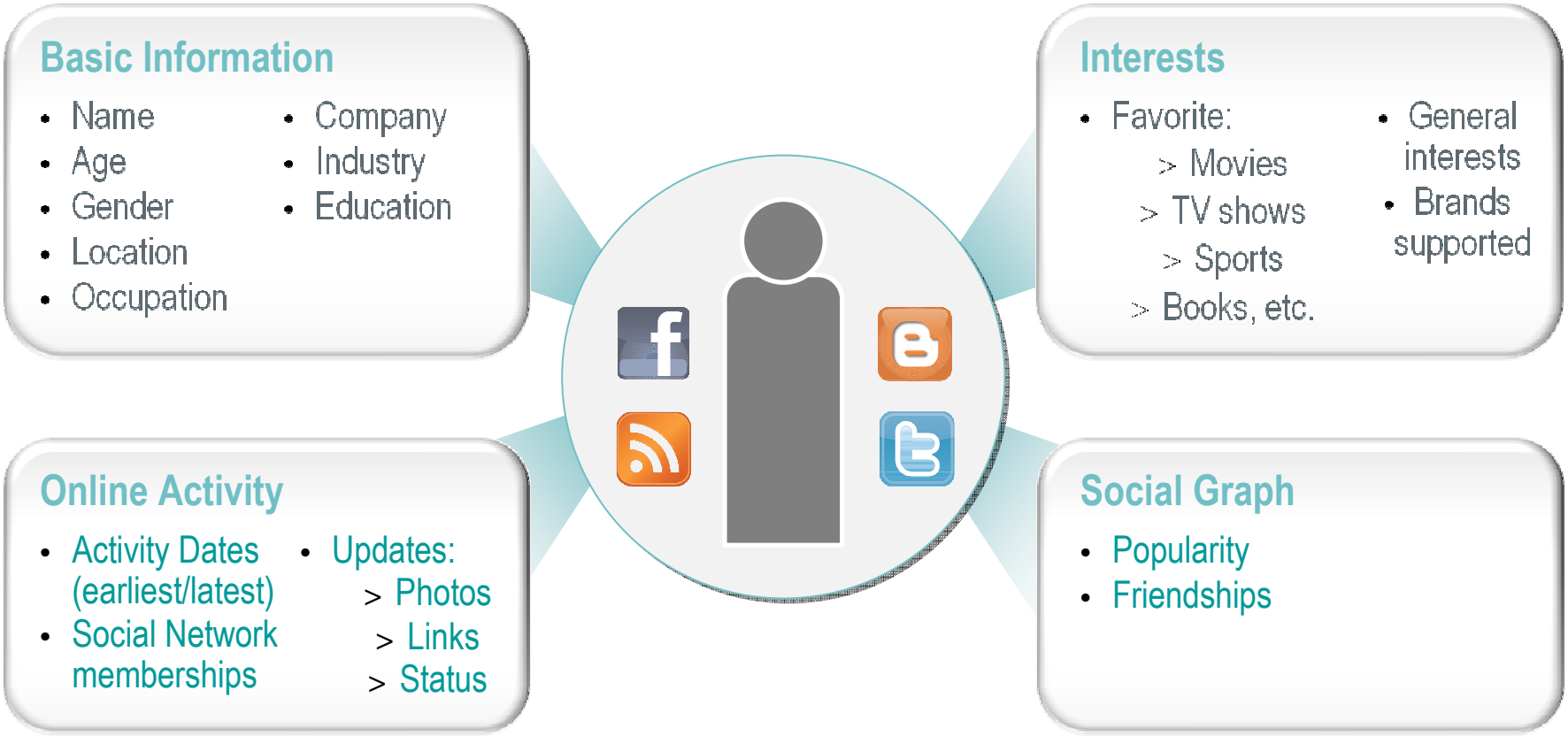


User Properties

Customer Interaction Base

Account Activity

2010's Big Data from Our Digital Footprint

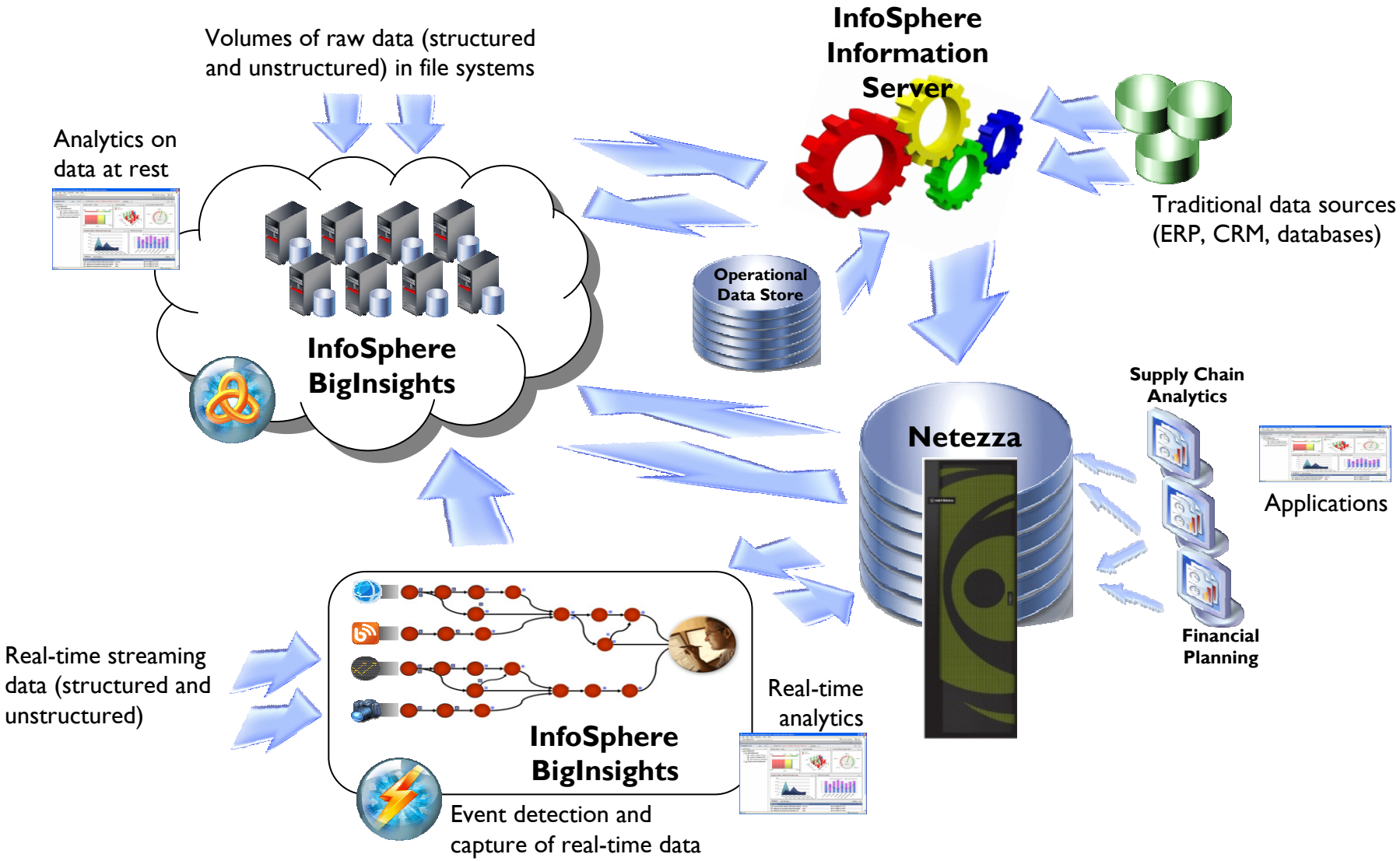


- 위치기반 데이터와 소셜 미디어에 의해 생산되는 사용자 활동 데이터가 새로운 차원의 정보로 양산되기 시작
- 새로운 정보를 활용해 실시간으로 서비스 사용자(고객)의 행동패턴, 선호도와 고객경험을 파악하는 상황인지(context awareness)가 일부 가능해졌으며 이를 바탕으로 기업들은 자사의 서비스나 제품 등 다양한 제안을 고객의 상황에 맞게 즉시 추천 가능해짐
- 클라우드 컴퓨팅, 스마트폰, 소셜 네트워크 서비스(SNS), RFID 등은 '빅 데이터' 시대를 가속화시킨 주요 요인

Data Warehouse vs. Big Data

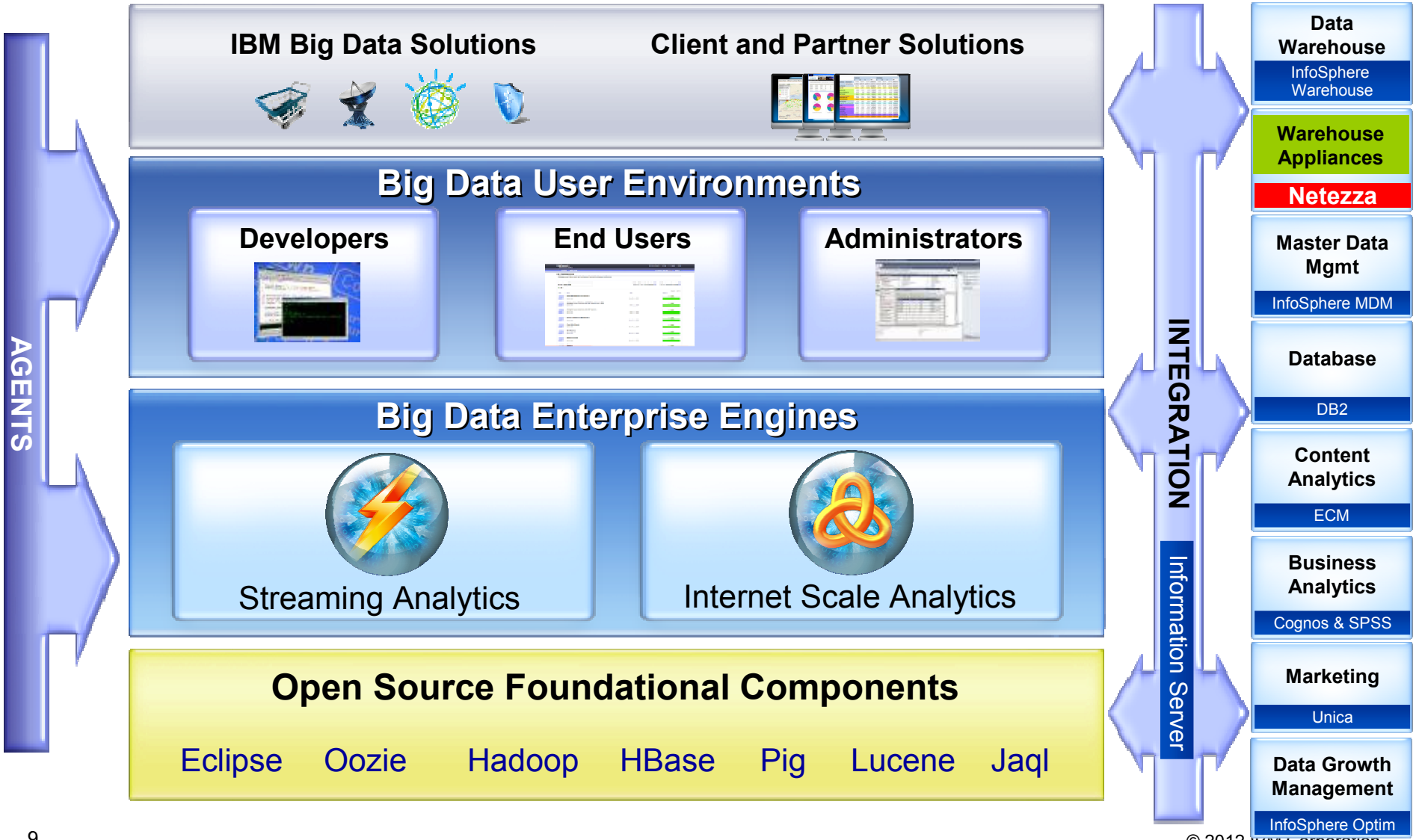
	Data Warehouse		Big Data
Variety	Structured data		Unstructured data Semi-structured data Structured data
Velocity	Data at rest		Data at rest Data in motion
Volume	Average size of new Netezza appliance: 10TB		Data at rest Data in motion

Result: Highly Flexible and Robust Warehouse



IBM's Big Data Platform Vision

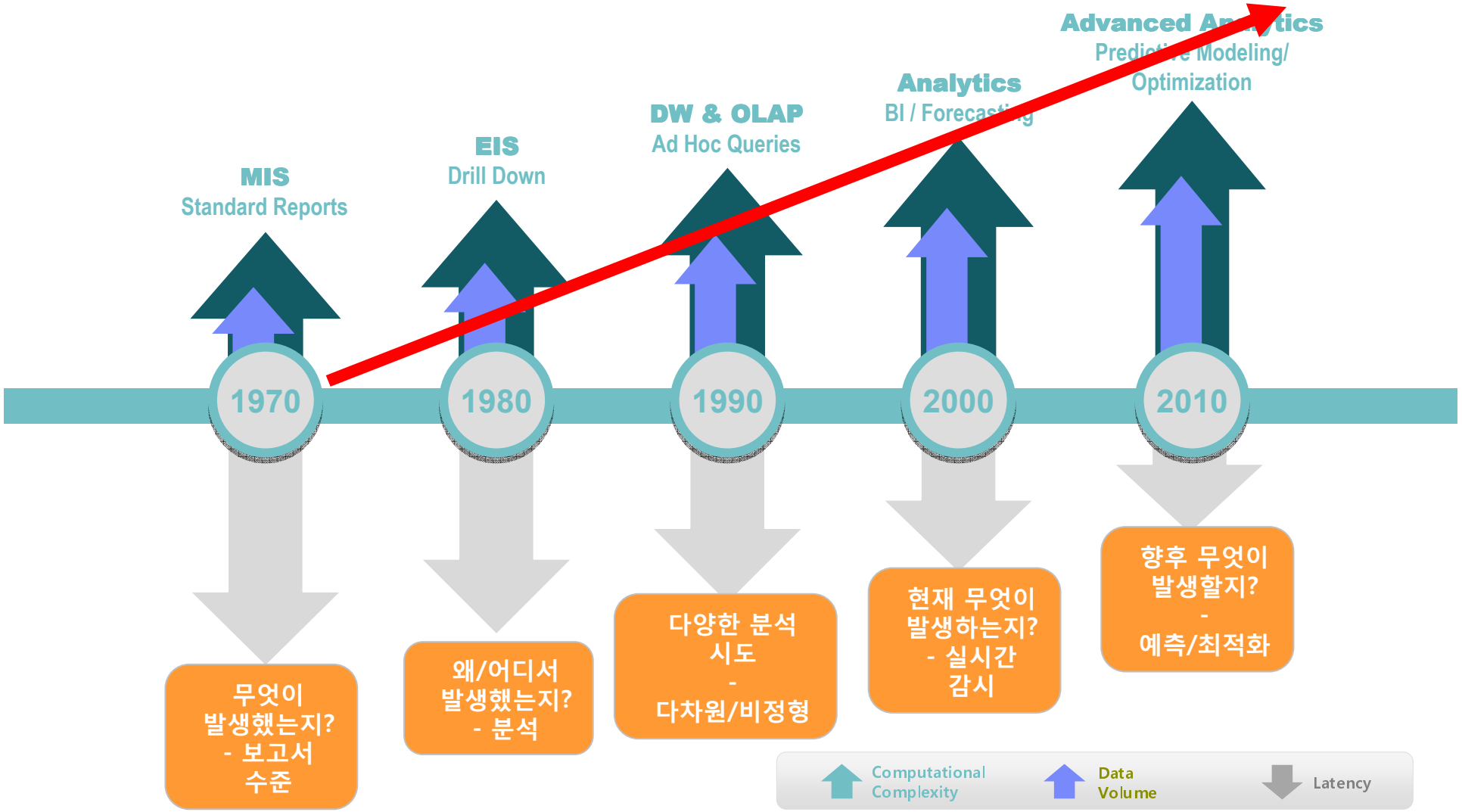
Bringing Big Data to the Enterprise



목차

- Big Data와 Big Insight

Evolution of Analytics – Reporting to Action

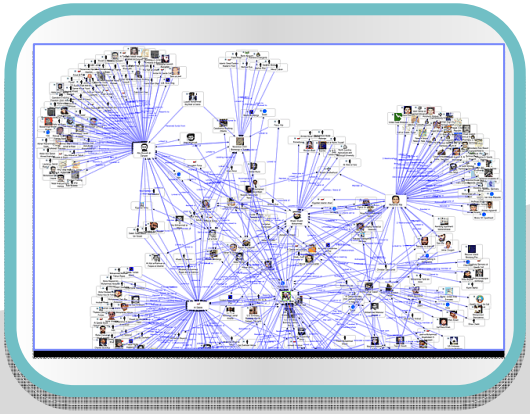


The Analytic Enterprise

BI Reporting and Ad-Hoc Analysis

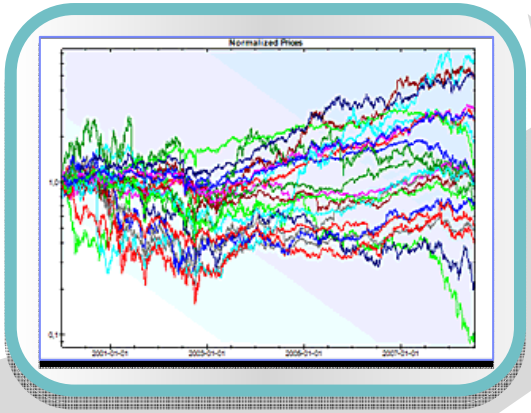
- What happened?
- When and where?
- How much?

Predictive Analytics



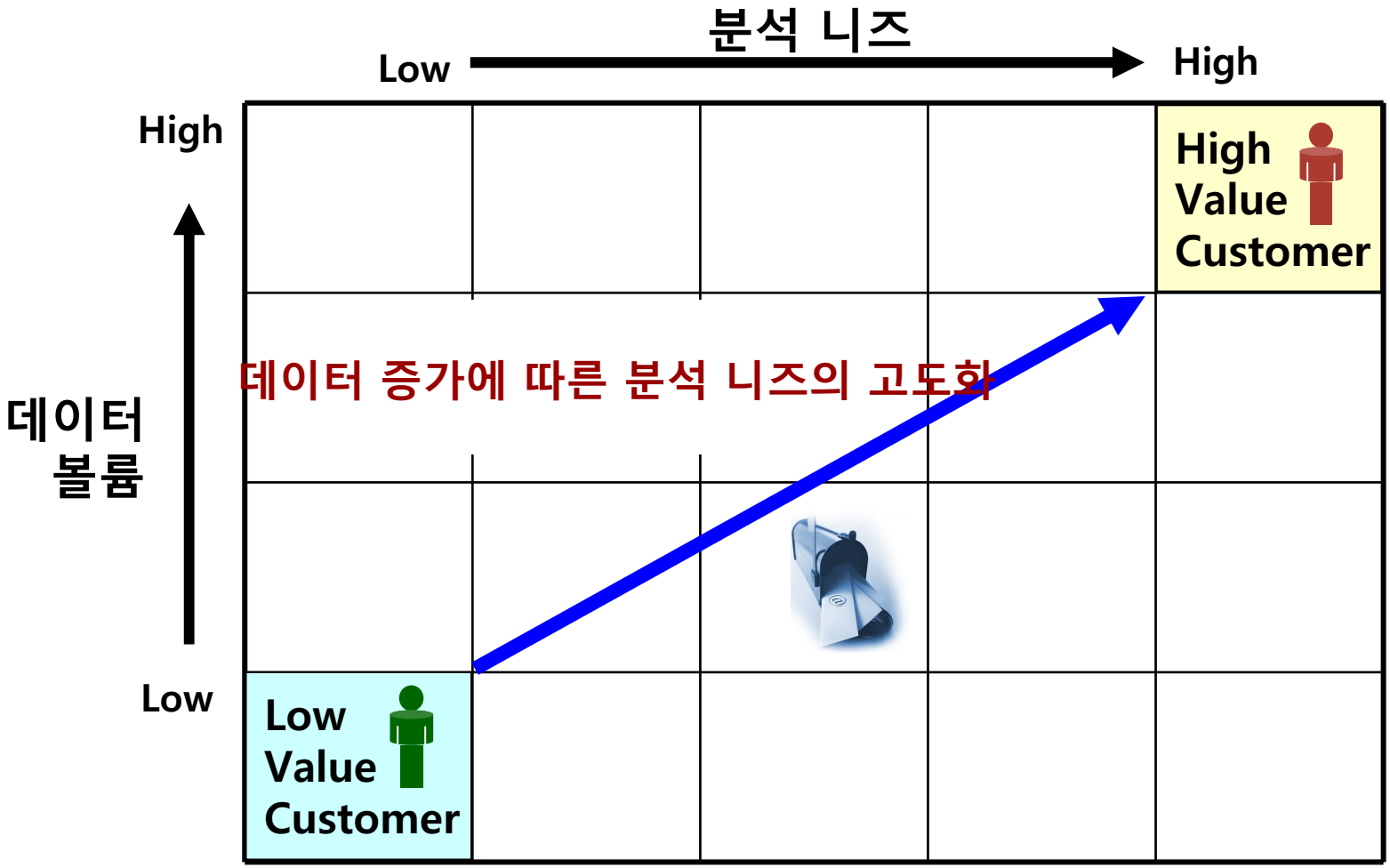
- What will happen?
- What will the impact be?

Optimization



- What is the best choice?

데이터와 분석 니즈



분석의 한계

BI/DW 구축 목표 및 비전...

서비스향상

비용절감

고객이탈방지

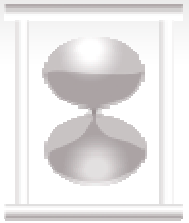
데이터신선도 유지

... 그러나 BI/DW의 현실은...

사용자 만족도 저하

느린응답시간

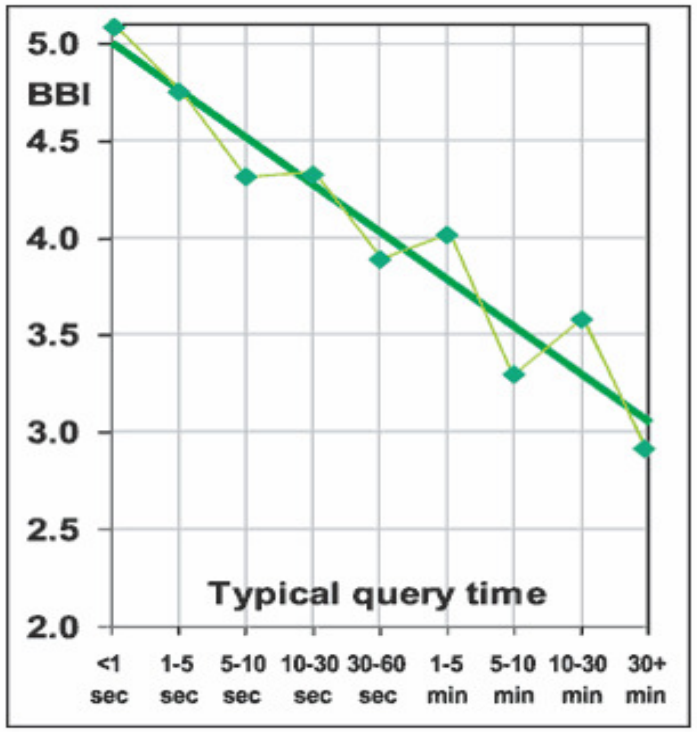
비용증대



데이터 신선도 저하

분석 성능의 중요성

Impact of Query Performance on Business Benefit



"기하급수적으로 데이터가 증가되고 있는 많은 BI/DW 시스템이 현재 성능 문제에 직면하고 있습니다."

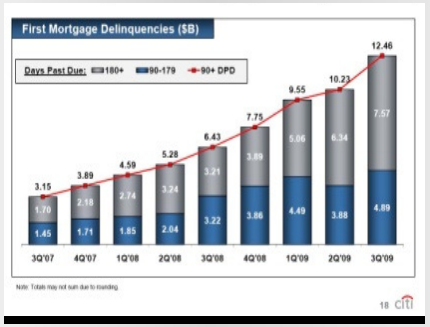
이제부터의 Big Data의 BI시대는

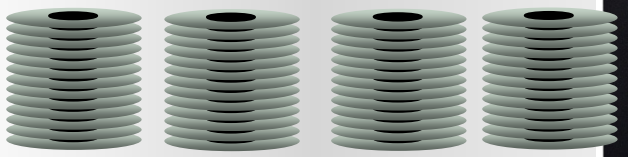
분석의『속도』가 기업의 경쟁력입니다.

Source: OLAP Survey 5 analysis based on 2,100 participants

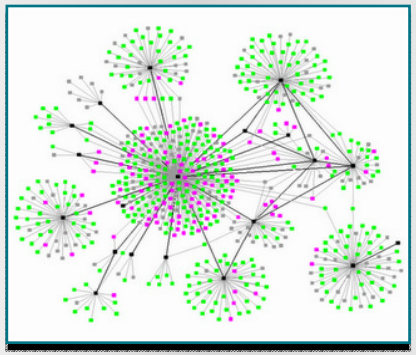
Big Data MEETS Big Math

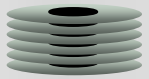
Big Data





Big Math



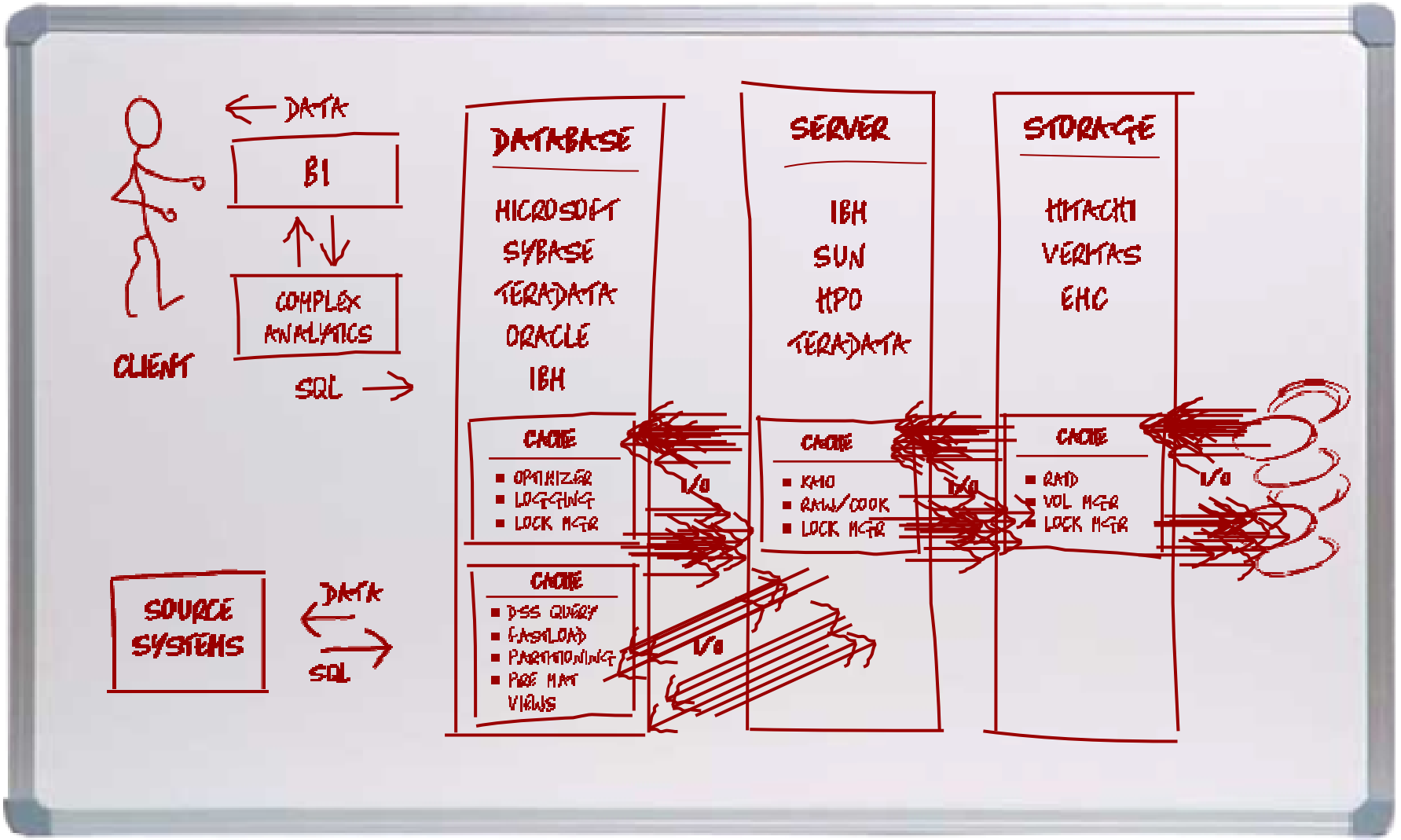


Analytics Without Constraints

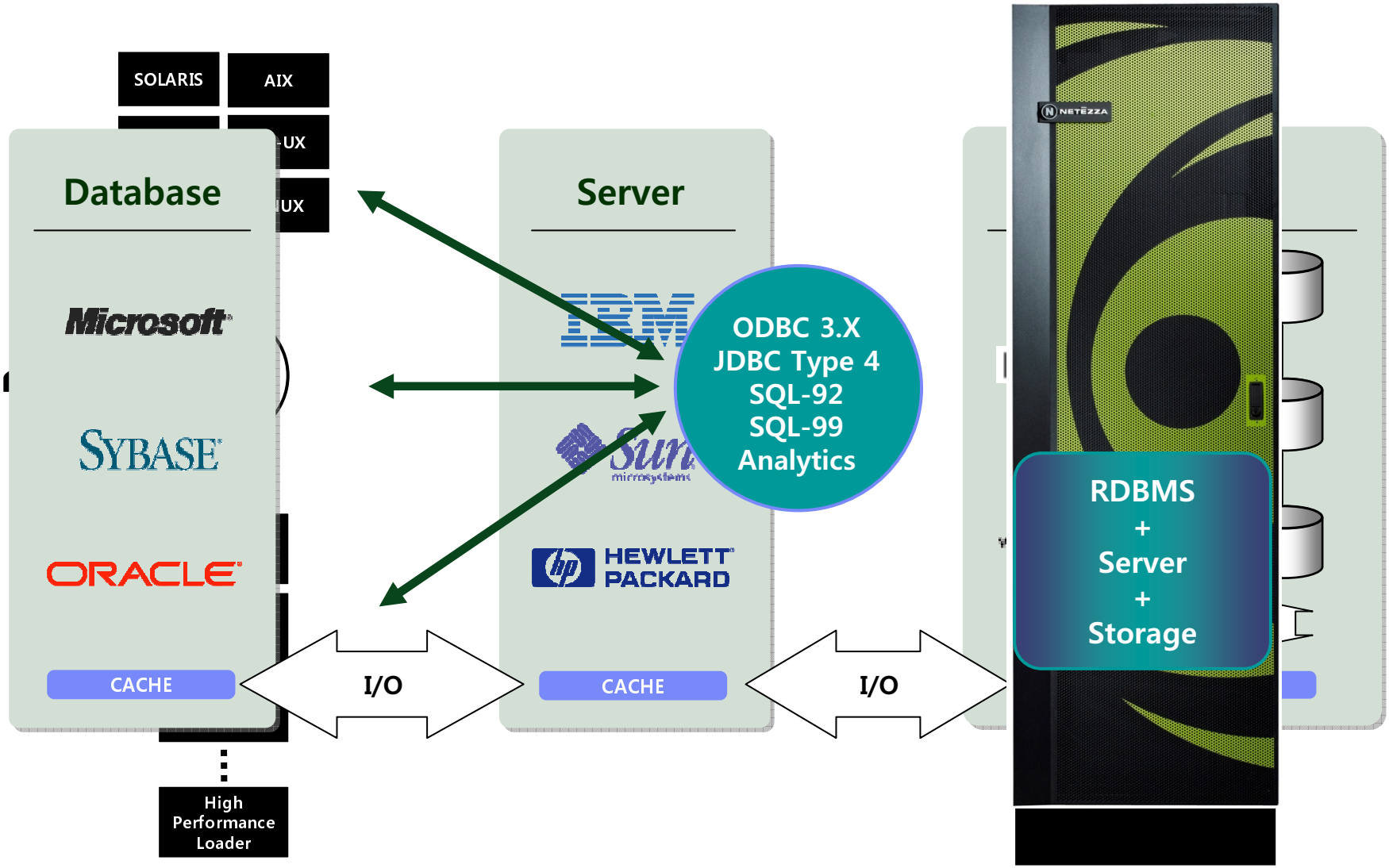
목차

- Big Insight를 위한 DW 및 분석 어플라이언스
  - Netezza Architecture

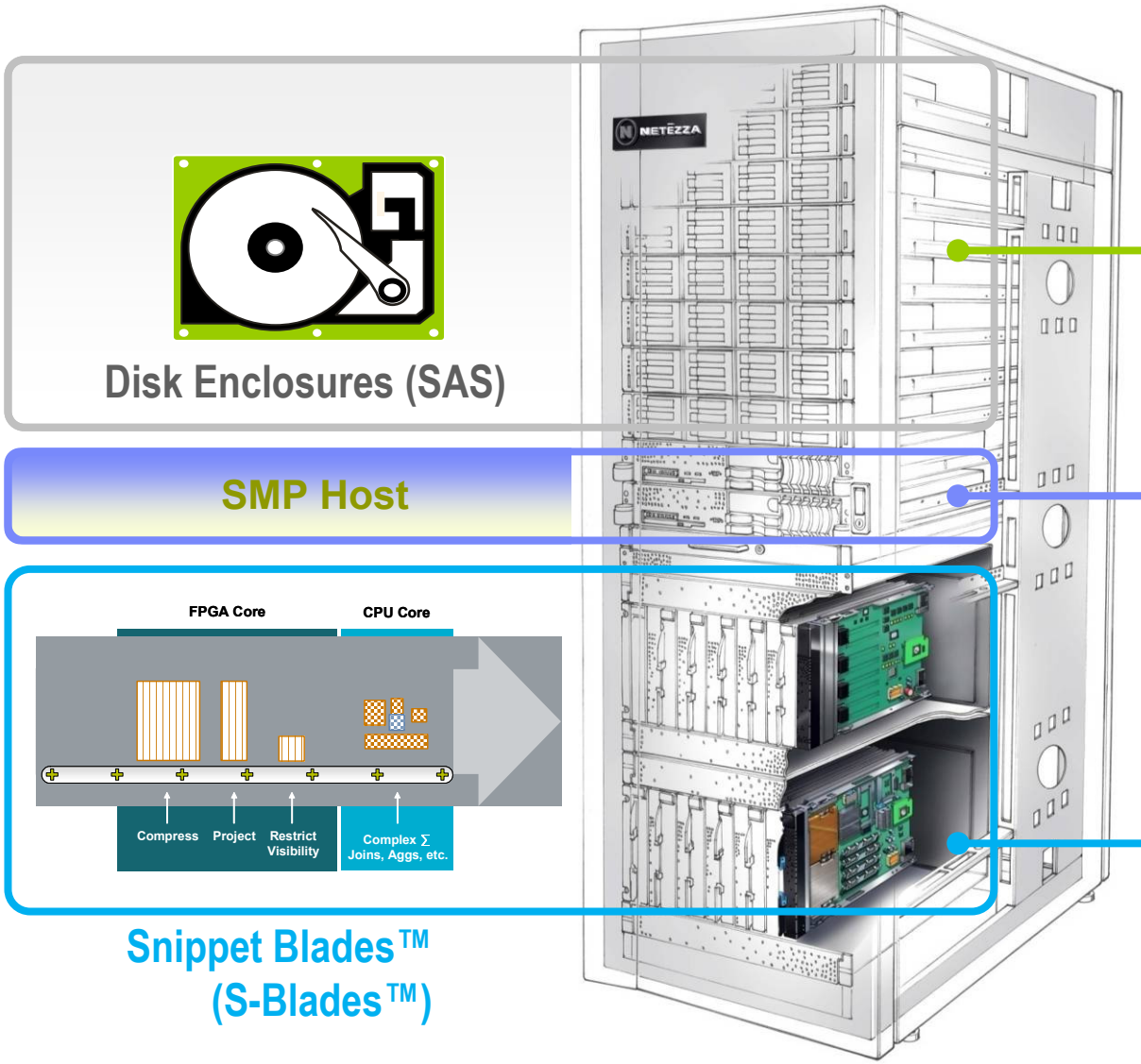
전형적인 DW 시스템의 데이터 처리 방식



Netezza Appliance Approach



The Netezza Appliance

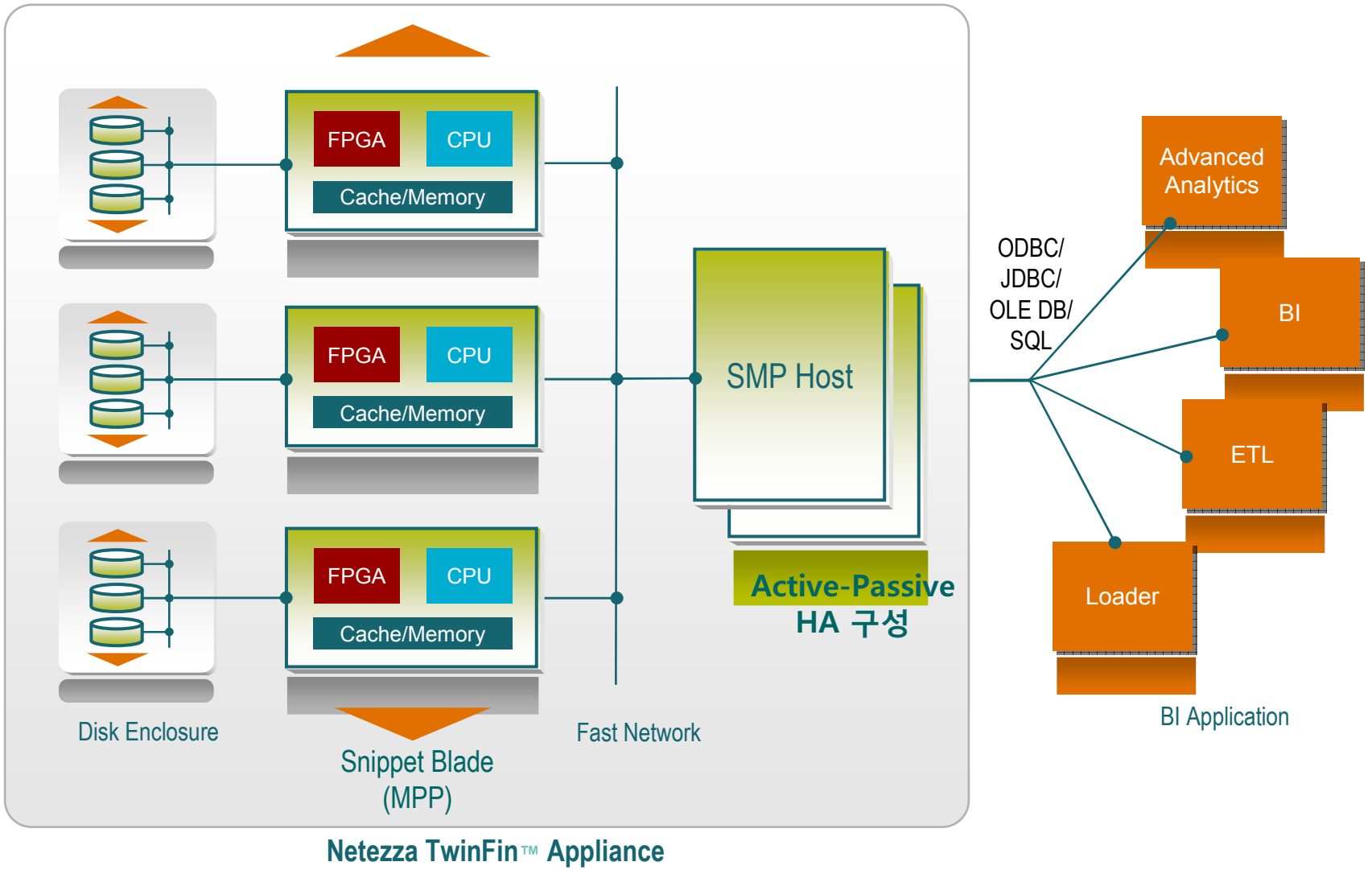


Slice of User Data
Swap and Mirror partitions
High speed data streaming

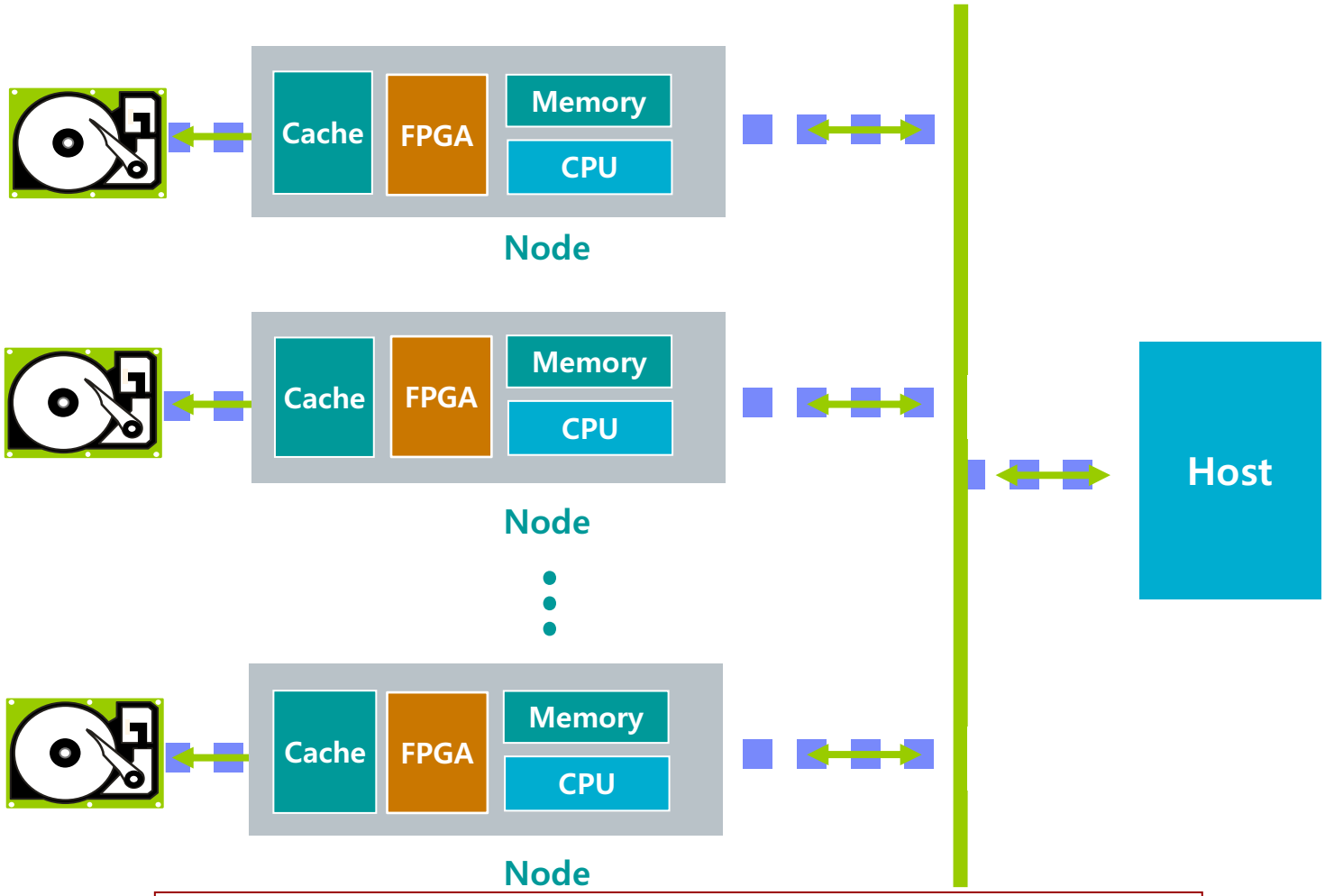
SQL Compiler
Query Plan
Optimize
Admin

Processor & streaming DB logic
High-performance database engine streaming joins, aggregations, sorts, etc.

Asymmetric Massively Parallel Processing™ (비대칭 초병렬 처리)

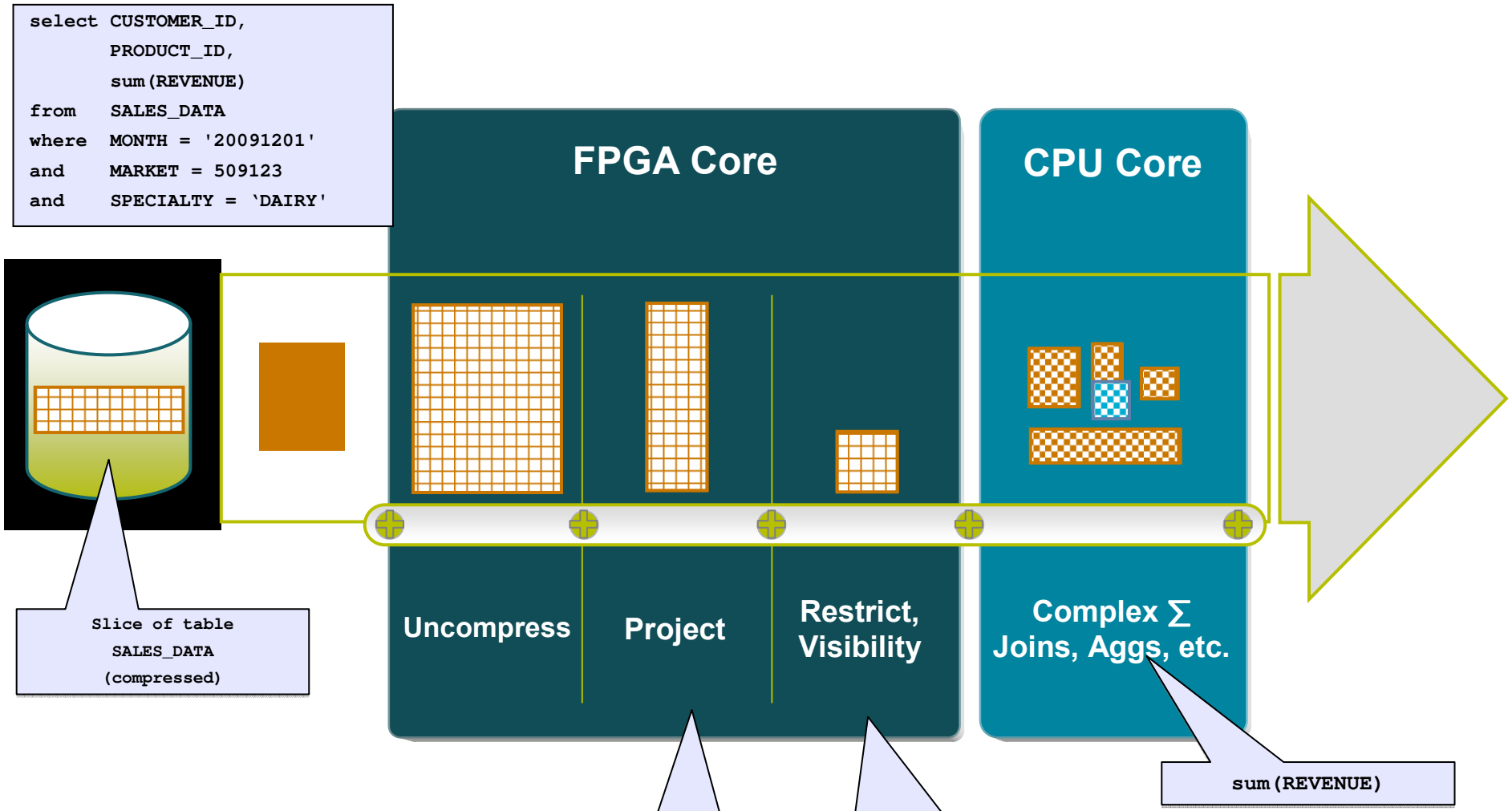


Netezza의 코어 기술인 「Data Streaming 처리」

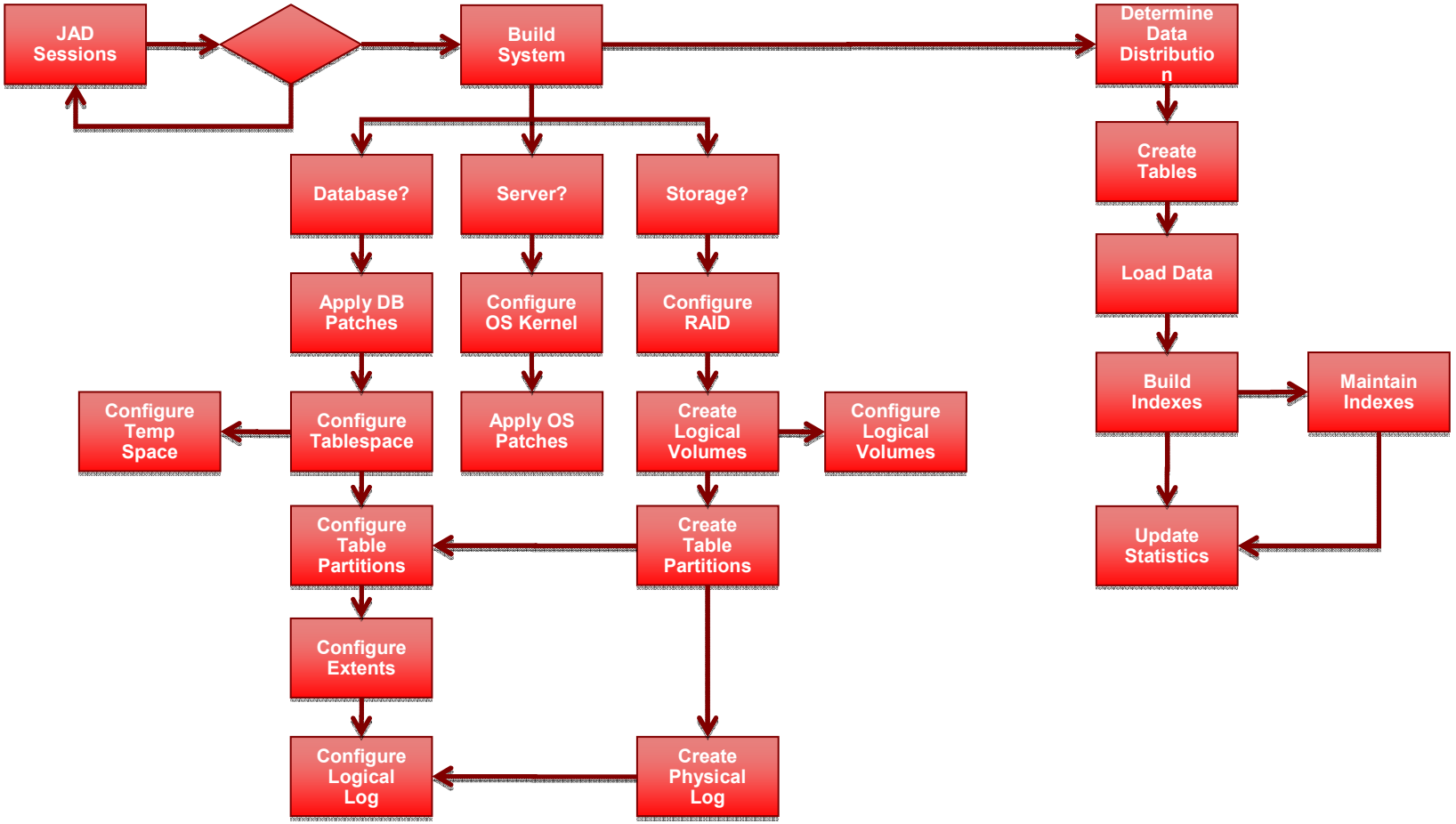


FPGA： Field Programmable Gate Array
각 칼럼과 레코드를 선택 추출하여 필요 최소한의 데이터를 필터링

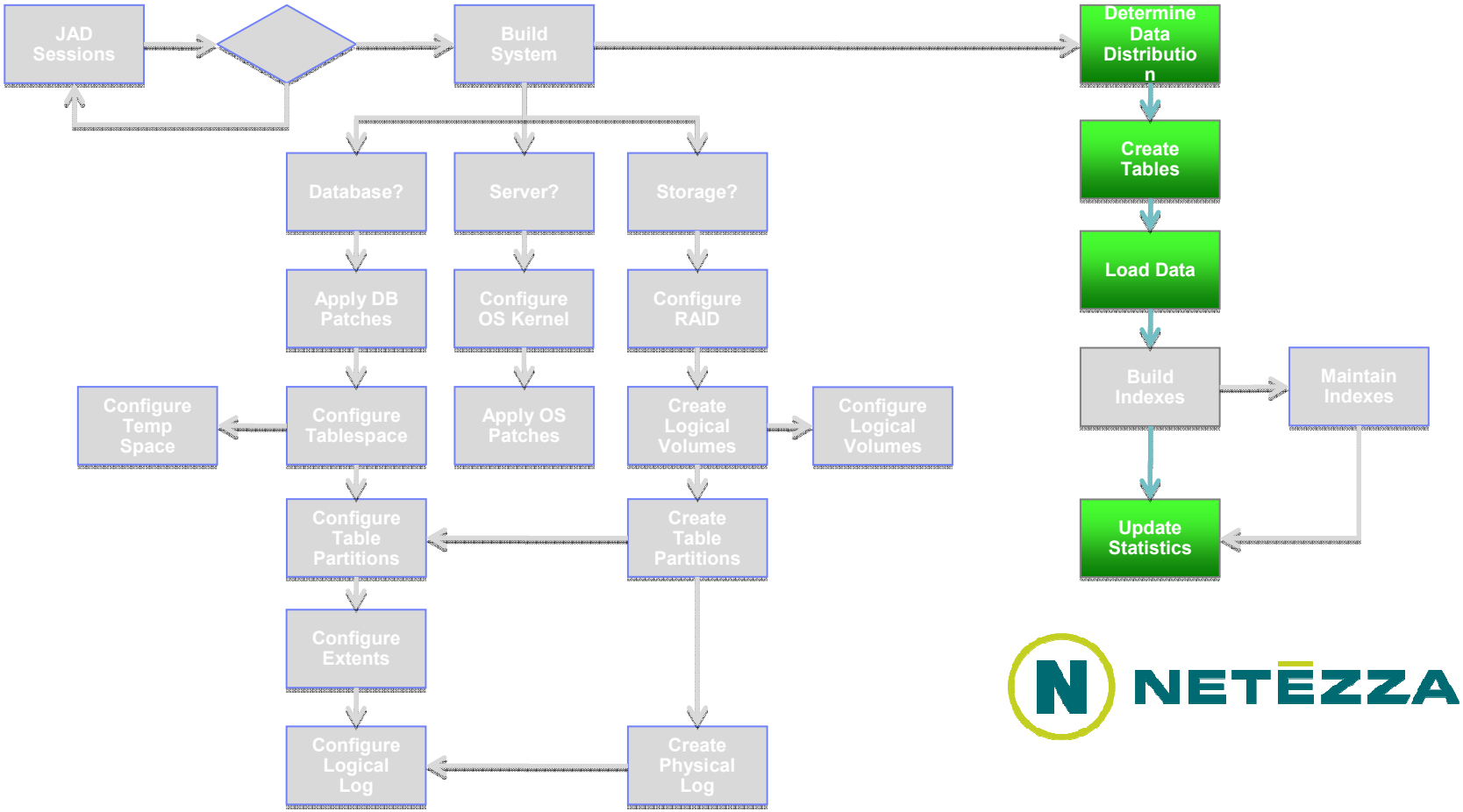
The Secret Sauce: FPGA



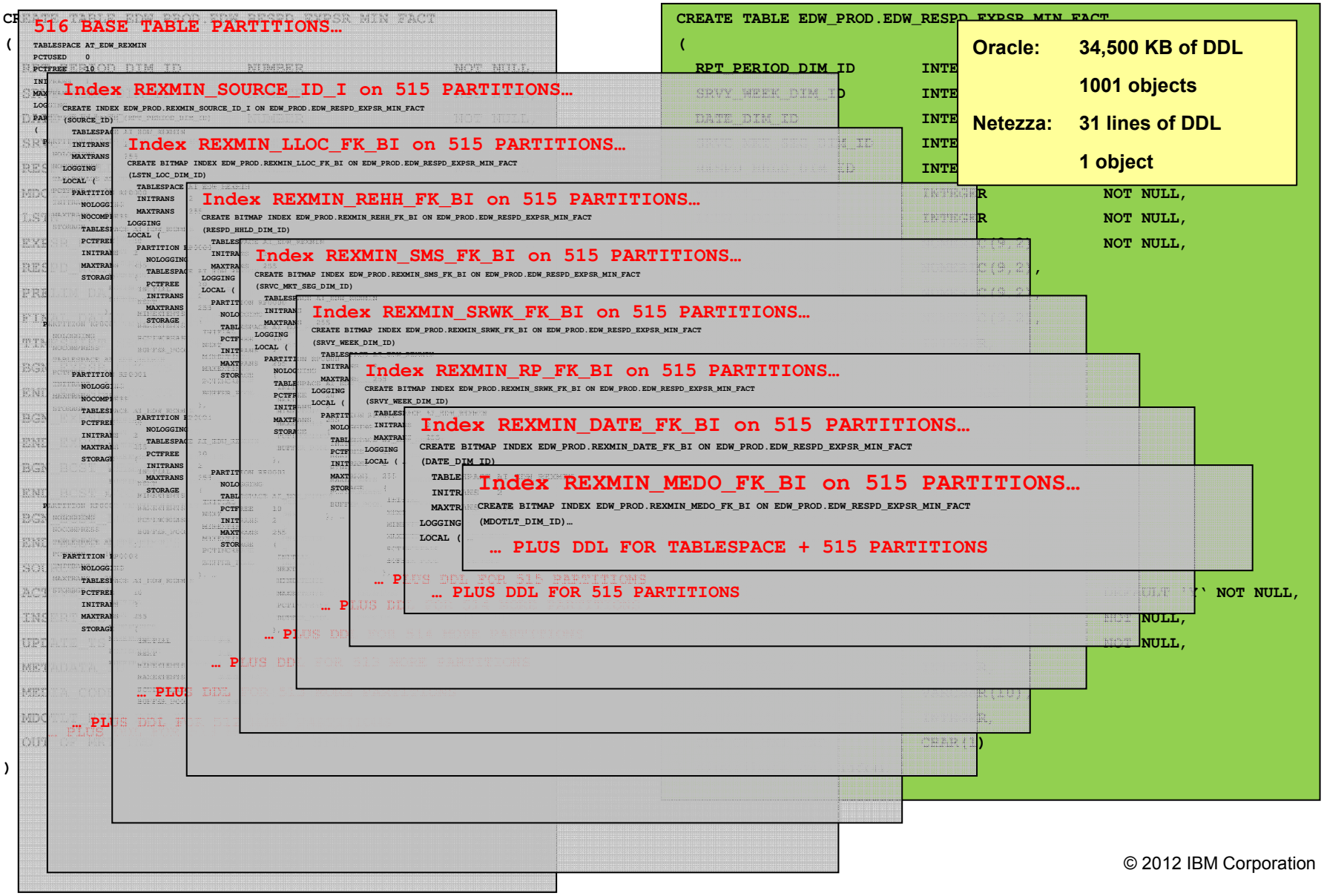
Typical Warehouse Implementation Process



Simplicity over Complexity



Traditional Complexity ... Netezza Simplicity (RDBMS 101)



Netezza Simplicity on TCO

Telecom Retailer and Service Provider

Telecom Call Detail Record FACT	Oracle Object Count *	Netezza Object Count
Tables	1	1
Indexes	12	
Table Partitions	47	
Index Partitions	564	
Table Partitions tablespaces	47	
Index Partitions tablespaces	47	
Table Data Files	170	
Index Data Files	122	
TOTAL	**1,010**	**1**

Netezza DDL 변환 작업

```
DROP TABLE DBAUSER.F_CMS_COMCSDAILYPUR CASCADE CONSTRAINTS;

CREATE TABLE DBAUSER.F_CMS_COMCSDAILYPUR (
  SALDATE            CHAR(8)          NOT NULL,
  COMCSNO            NUMBER(9)        NOT NULL,
  CHCD               VARCHAR2(10)     NOT NULL,
  CUSTCD             CHAR(6)          NOT NULL,
...
)
TABLESPACE TSD_CMS1
PCTFREE 10
PCTUSED 0
INITRANS 1
MAXTRANS 255
NOLOGGING
PARTITION BY RANGE (SALDATE)
(
    PARTITION F_CMS_COMCSDAILYPUR_P200610 VALUES LESS THAN ('20061101')
        NOLOGGING
        TABLESPACE TSD_CMS1
        PCTFREE 10
        PCTUSED -1
        INITRANS 1
        MAXTRANS 255
        STORAGE (
            INITIAL 102400 K
            MINEXTENTS 1
            MAXEXTENTS UNLIMITED
        ),
    PARTITION F_CMS_COMCSDAILYPUR_P200611 VALUES LESS THAN ('20061201')
        NOLOGGING
        TABLESPACE TSD_CMS2
        PCTFREE 10
        PCTUSED -1
        INITRANS 1
```

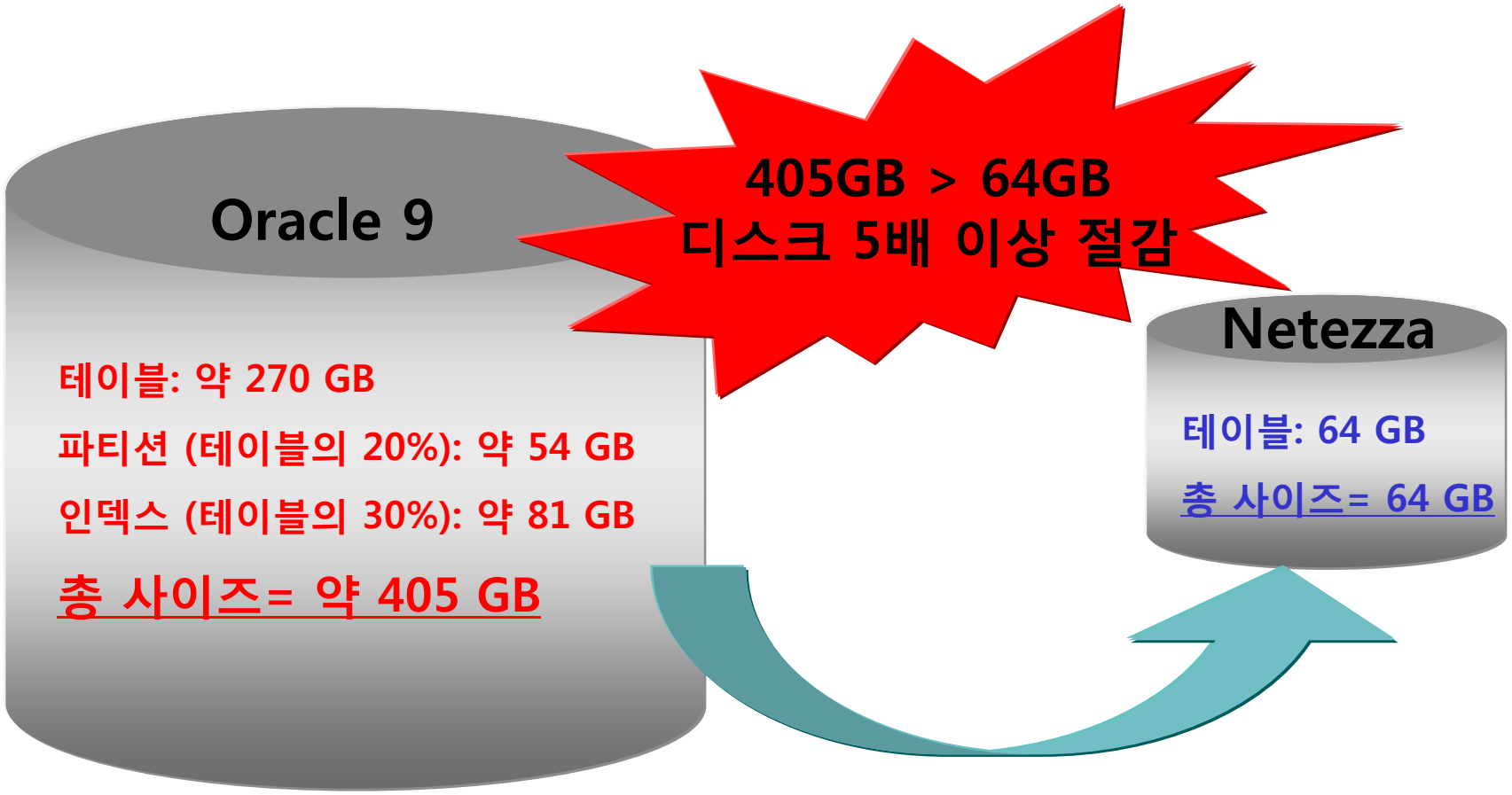
텍스트 841 라인
각종 옵션 정의

```
DROP TABLE DBAUSER.F_CMS_COMCSDAILYPUR ;

CREATE TABLE DBAUSER.F_CMS_COMCSDAILYPUR (
  SALDATE            CHAR(8)            NOT NULL,
  COMCSNO            INTEGER          NOT NULL,
  CHCD               VARCHAR(10)       NOT NULL,
  CUSTCD             CHAR(6)            NOT NULL,
  BRANDID            VARCHAR(10)       NOT NULL,
  LINEID             VARCHAR(10)       NOT NULL,
...
) ;
```

텍스트 46 라인
테이블 정의만
필요!

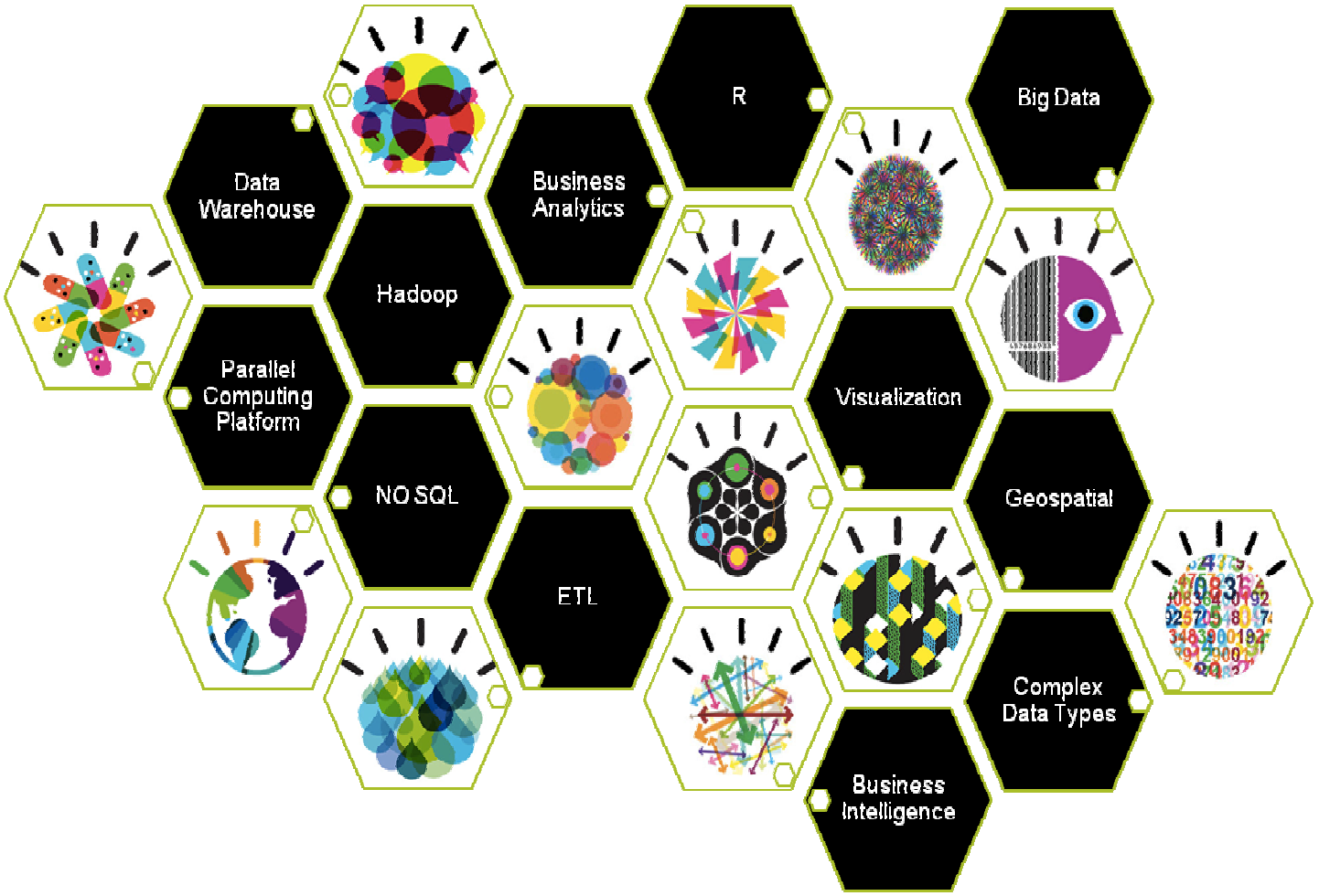
Low TCO (압축률)



목차

- Big Insight를 위한 DW 및 분석 어플라이언스
  - Netezza Advanced Analytics

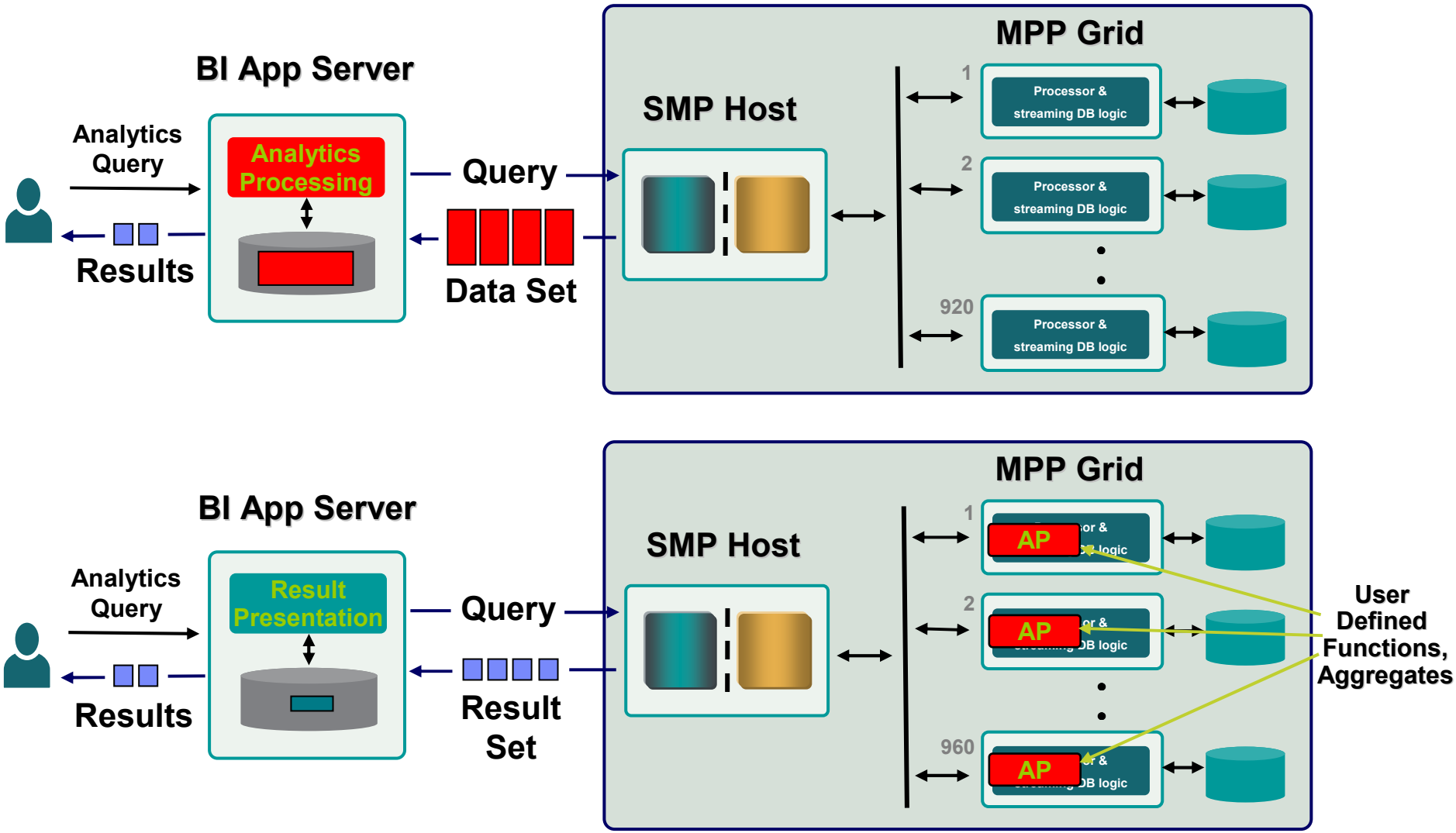
Powering the Age of Analytics



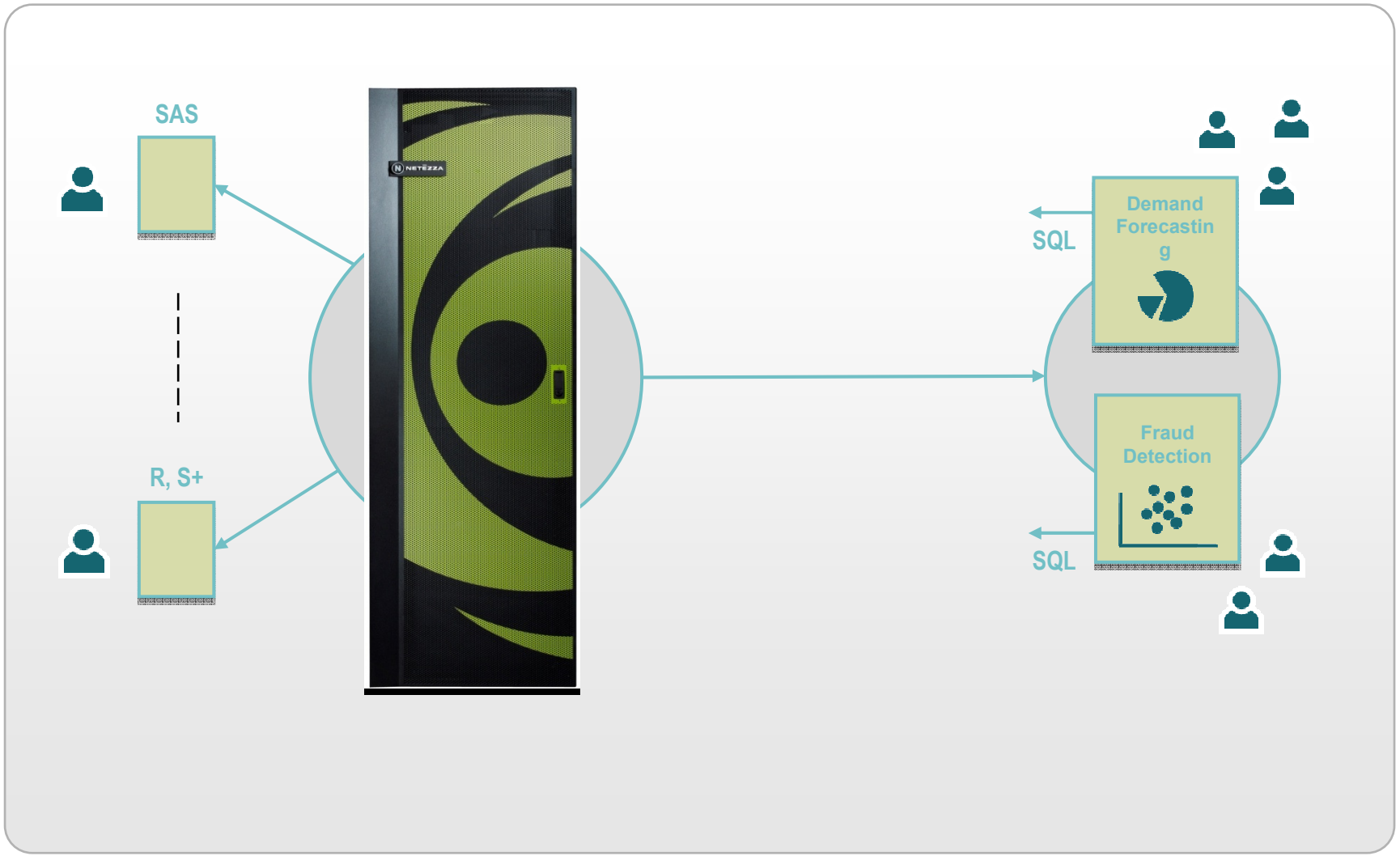
Powering the Age of Analytics

- Parallel Computing Platform
- Warehouse for Complex Data Types
- Hadoop Integration
- Business Analytics & R
- Business Intelligence
- ETL & ELT
- Geospatial

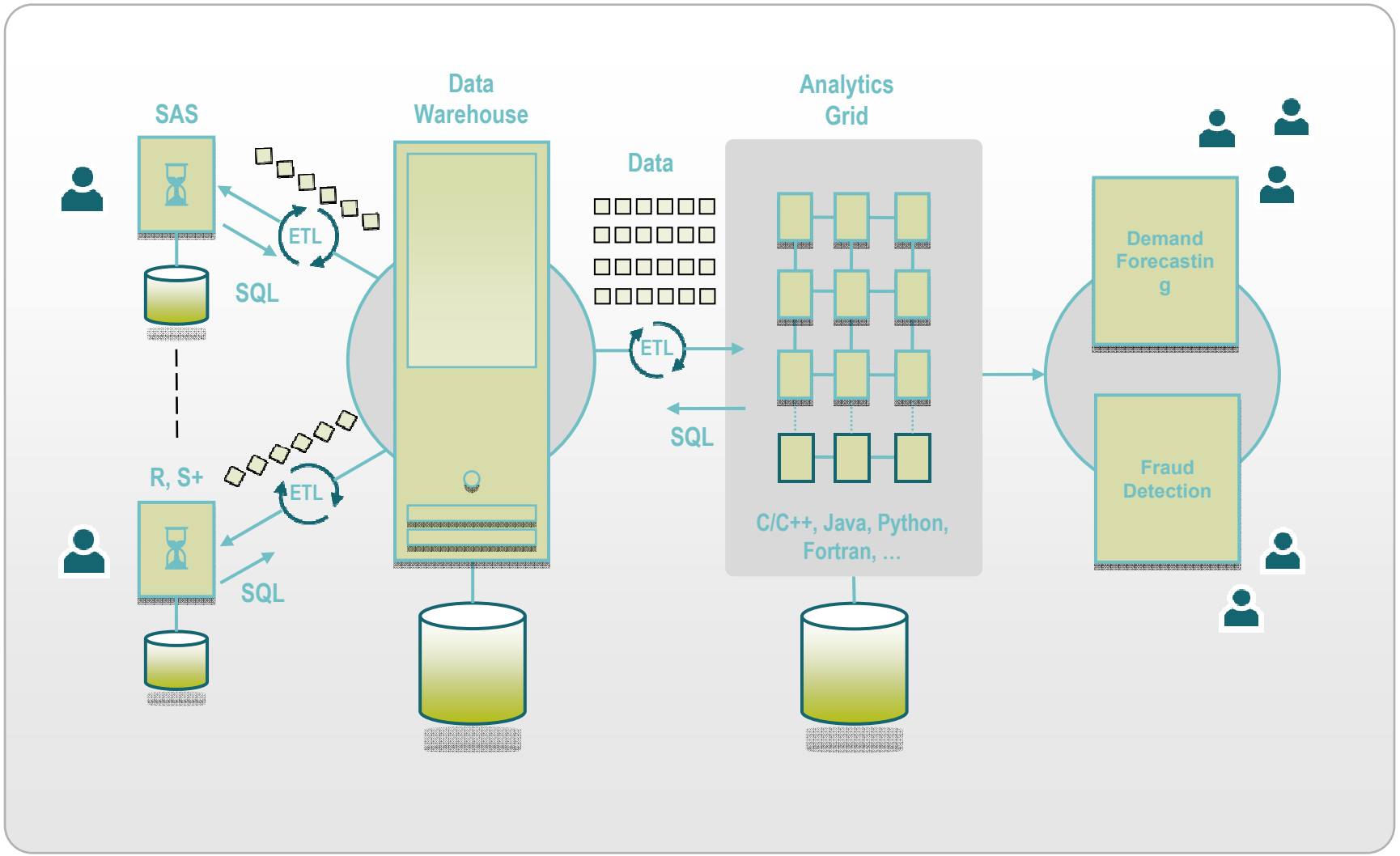
Netezza Embedded Analytics



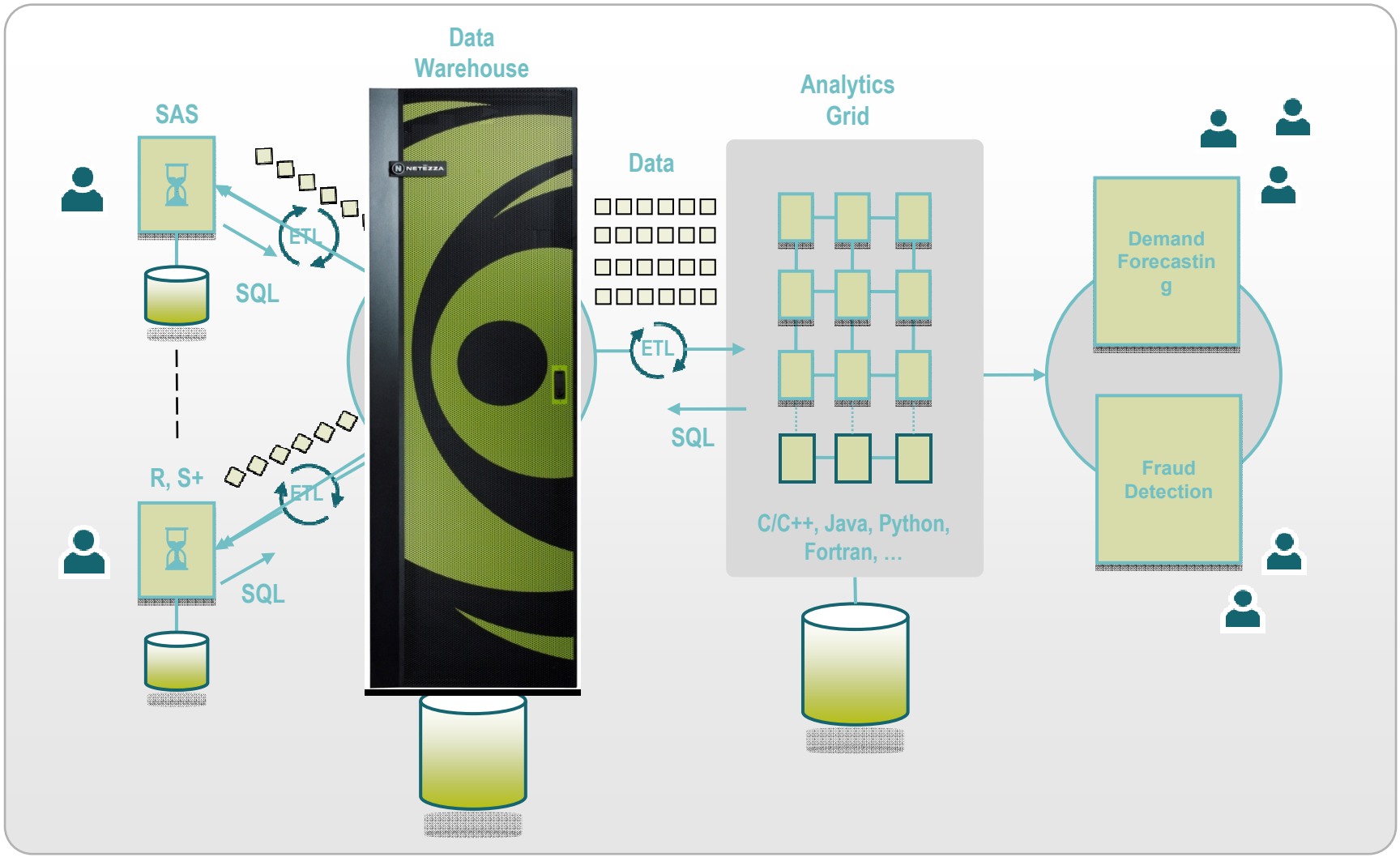
Advanced Analytics with Netezza



Advanced Analytics – the Traditional Way



Advanced Analytics with Netezza



Common use cases

unstructured data

→ hadoop → NETEZZA

Create context (classification, text mining)

Analyze

semi-structured data

→ hadoop → NETEZZA

Parse , aggregate

Analyze, report

structured data

→ NETEZZA → hadoop

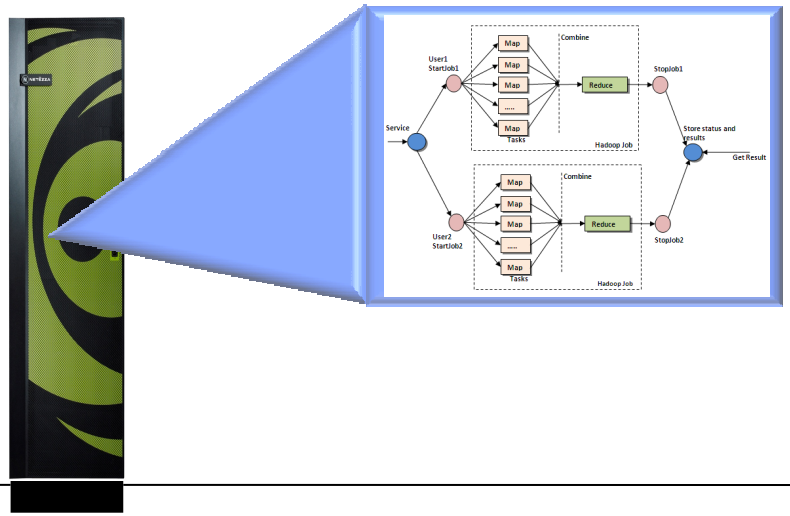
Analyze, report

Active archival
Long running queries

NING bluekai edmunds.com eHarmony tremor media

Strategy and Hadoop support

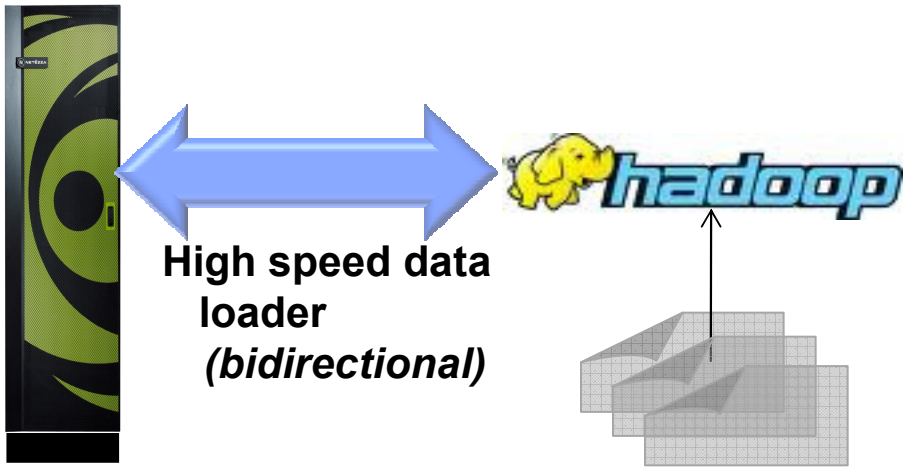
Scenario 1: Hadoop/Map-Reduce framework inside the appliance



- Invoke Hadoop jobs like UDFs
- Combine ubiquity of SQL with flexibility of Map Reduce
- Port existing jobs and functions as-is

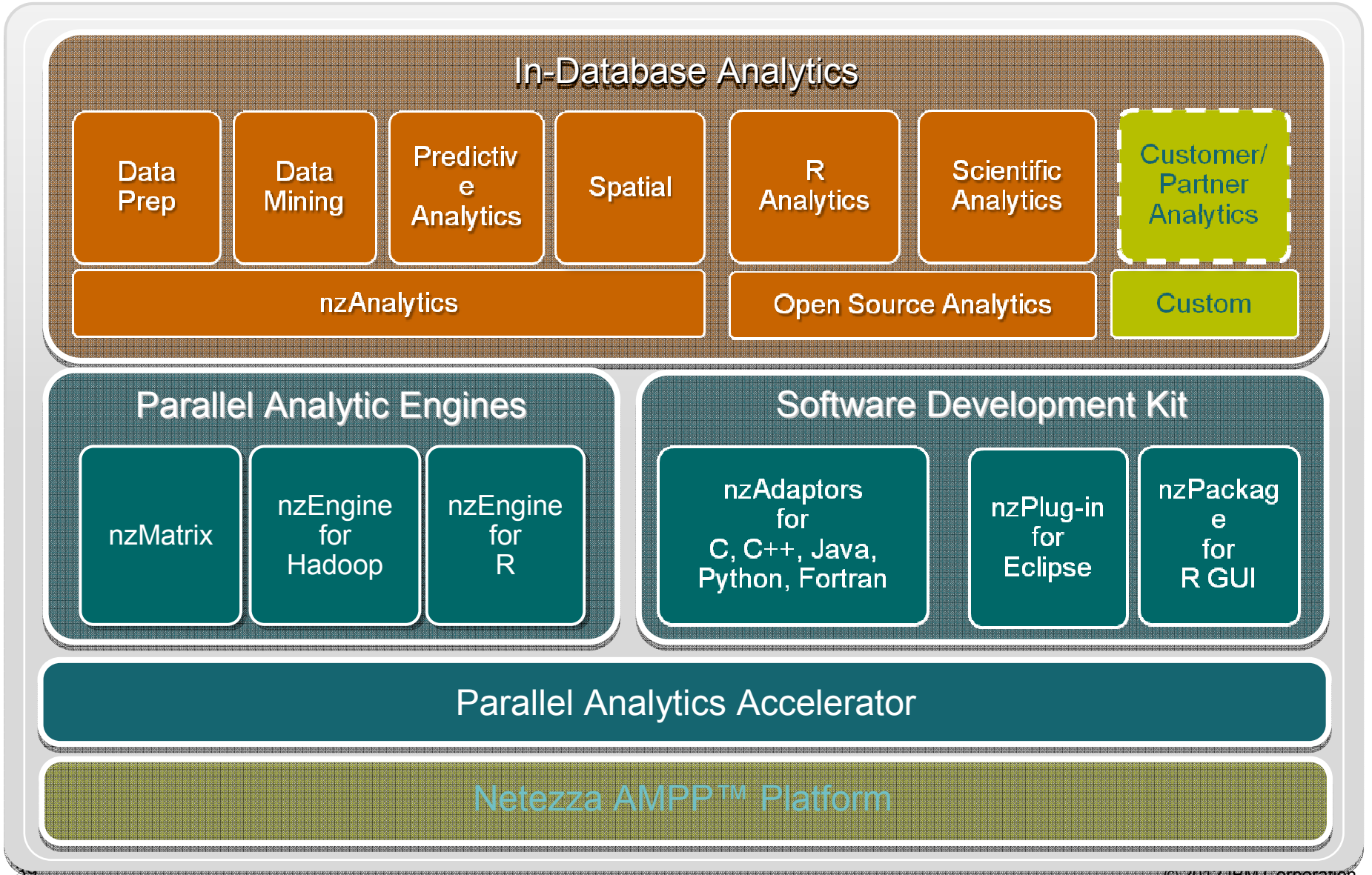
Scenario 2: Hadoop integration (Cloudera Connector)

cloudera



- Move data back and forth between Netezza and Hadoop cluster
- Use Hadoop for ingesting/parsing web logs, offline analytics
- Port existing jobs and functions as-is

Advanced Analytics Framework



목차

- Big Insight를 위한 DW 및 분석 어플라이언스
  - Netezza Customers

Performance

15,000명의 유저가 하루 800,000개 이상의 쿼리를 네티자 도입 이전과 비교해서 50배의 빠른 속도로 분석

*"…when something took 24 hours I could only do so much with it, but when something takes 10 seconds, I may be able to **completely rethink the business** …"*

- SVP Application Development, Nielsen

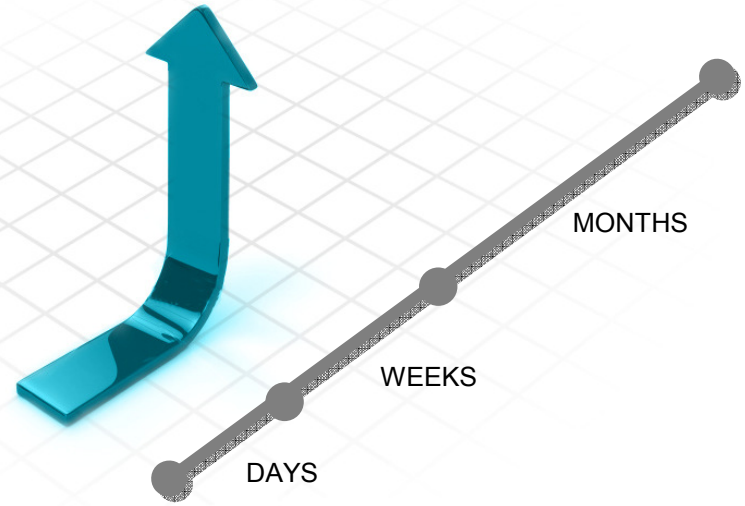
nielsen

Source: http://www.youtube.com/watch?v=yOwnX14nLrE&feature=player_embedded

Simplicity

도입에서 운영까지 6개월 만에 구축 완료, 1-Day 교육만으로 운영부터 구축까지

3개월 만에 ROI 달성, 기존 Oracle 대비하여 200배 이상의 빠른 분석 성능



"Allowing the business users access to the Netezza box was what sold it."

XO Communications

Steve Taff,
Executive Dir. of IT Services

Big Data

Netezza 도입으로 7년간 거래 이력 데이터가 매년 2-3배로 빠르게 증가하여 1PB로 확장

*"NYSE … has replaced an Oracle IO relational database with a data warehousing appliance from Netezza, allowing it to **conduct rapid searches of 650 terabytes** of data."*

ComputerWeekly.com

NYSE Euronext

Source: http://www.computerweekly.com/Articles/2008/04/14/230265/NYSE-improves-data-management-with-datawarehousing.htm

Smart, Advanced Analytics

고객이 다시 매장을 방문할 경우 무엇을 구매할지 예측

고객에게 발송된 Coupon 이용률이 최대 25% 증가

*"Because of (Netezza's) in-database technology, we believe we'll be able to do 600 predictive models per year (**10X** as many as before) **with the same staff**."*

CATALINA MARKETING

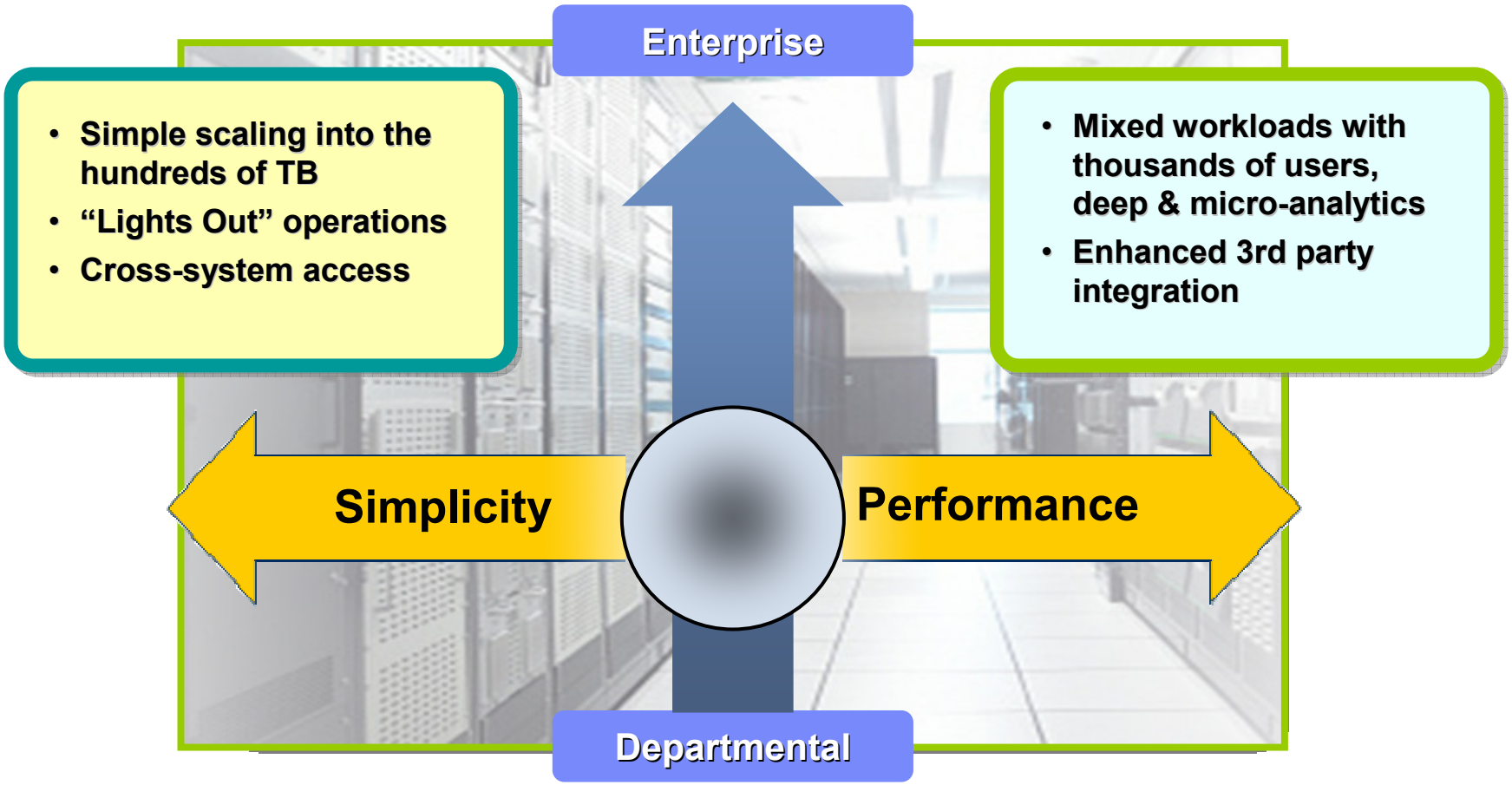
Eric Williams,

CIO and executive VP

목차

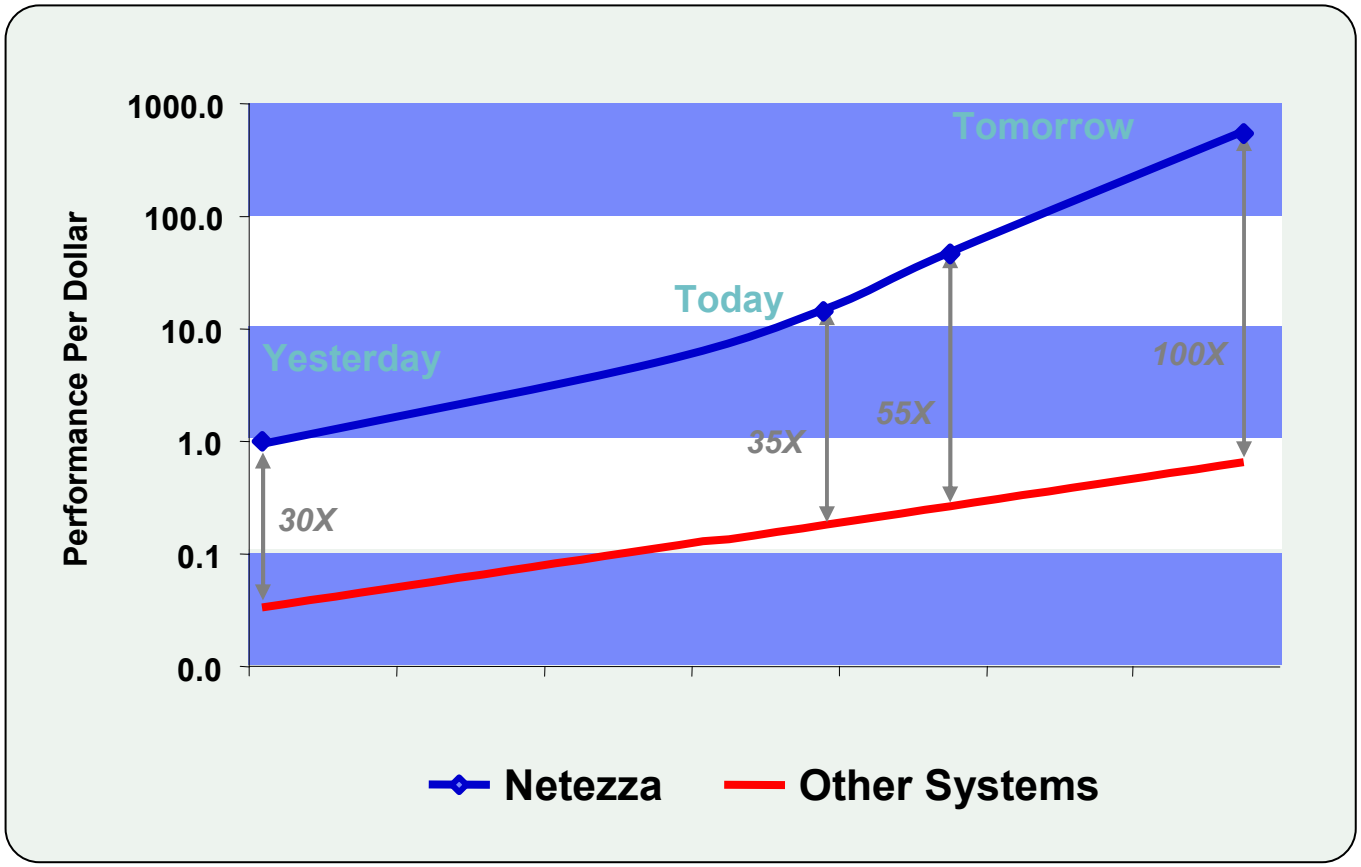
- 진화하는 DW 및 분석 어플라이언스의 방향

The Mental Shift: Big Data Storage to Big Data Strategy



Netezza Roadmap: Appliance and Architecture

Changing the Rules of BI



Global Data Warehouse Deployment