빅 데이터를 위한 인프라 최적화 구현방법

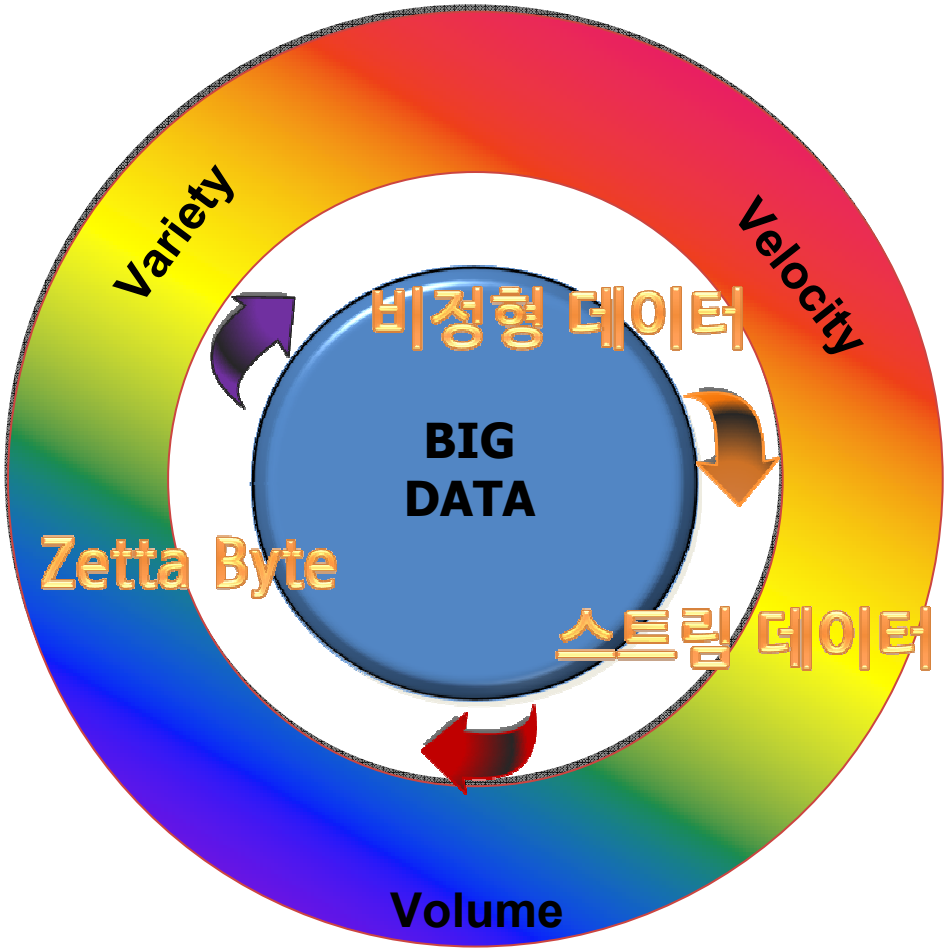
한국IBM 시스템테크놀로지그룹, 시스템 스토리지 사업부
김정림 차장 (jlkim@kr.ibm.com)

목차

1. 스토리지 관점에서의 3V 재해석
2. 분석기법 변화에 따른 플랫폼 변화
3. IBM BIG DATA 분석 플랫폼
4. BIG DATA의 데이터 관리 및 최적화
5. 결론

1 스토리지 관점에서의 3V 재해석

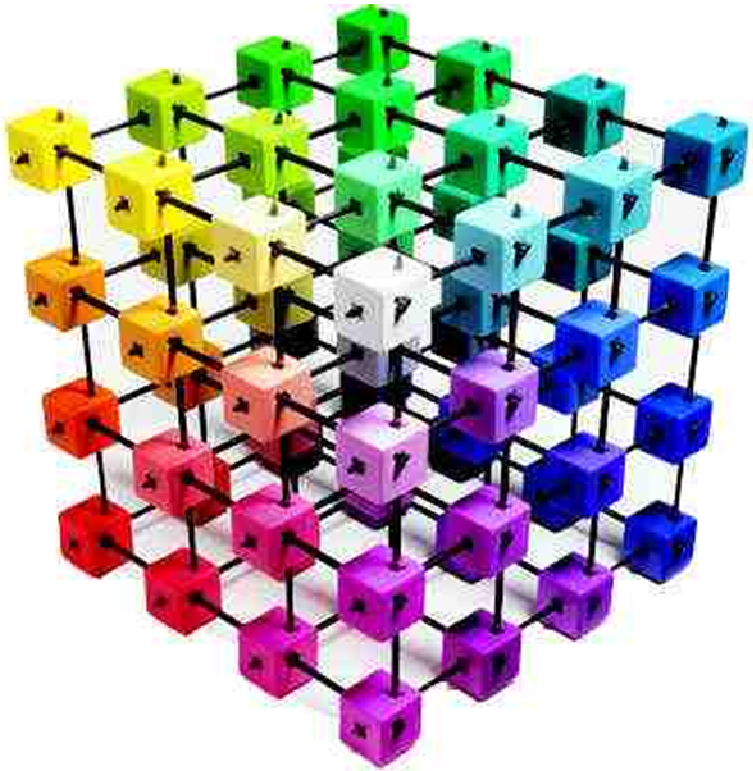
2012 IT TREND > BIG DATA



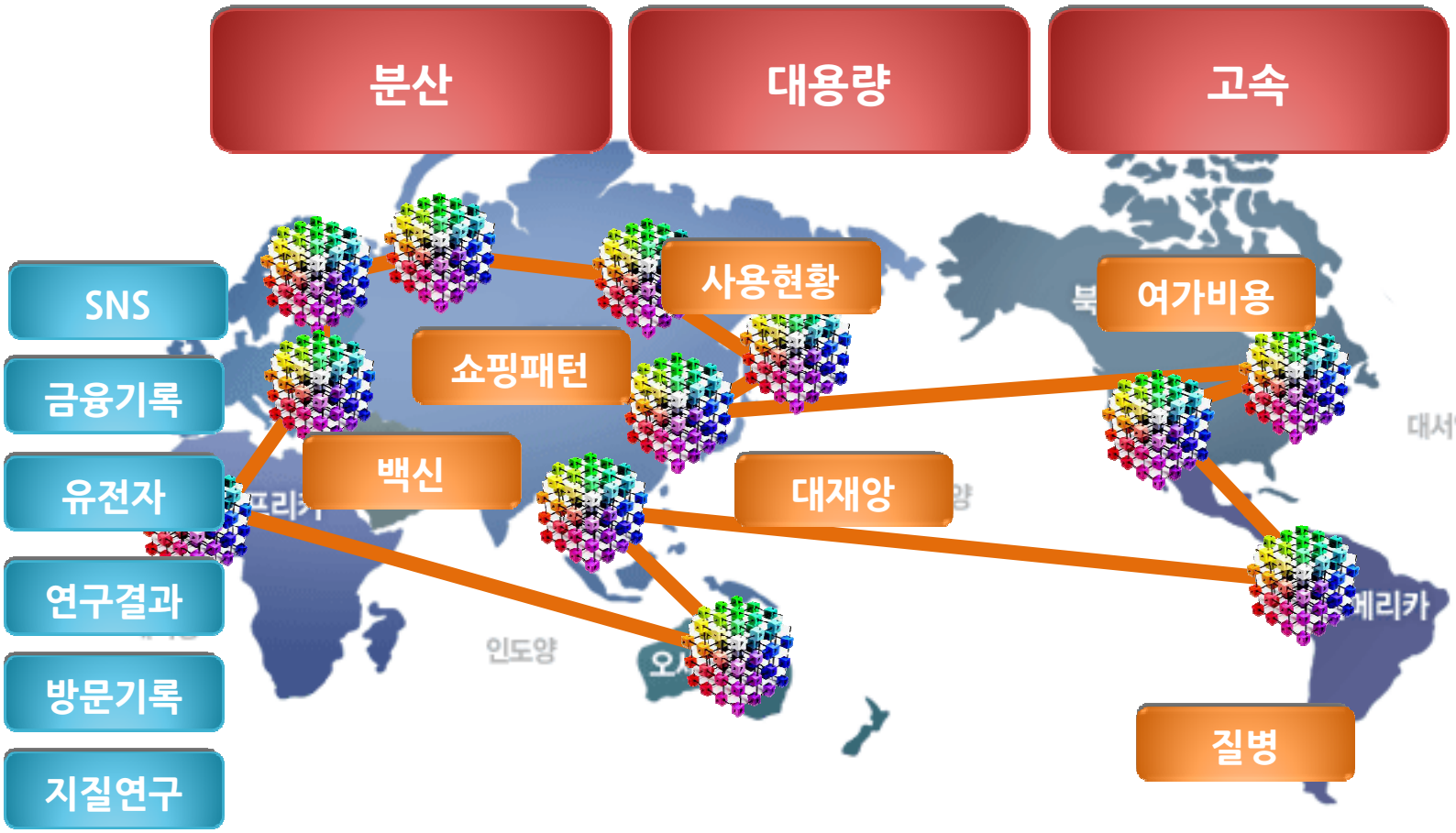
BIG DATA

기존 소프트웨어로 처리가
불가능한 데이터

데이터의 폭발 및 분석 요구

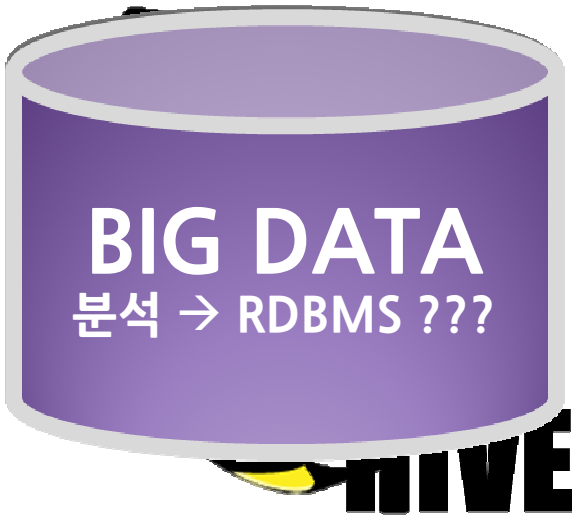


데이터의 폭발 및 분석 요구 > 분석 기법의 변화



분석기법 변화에 따른 플랫폼 변화

분석 플랫폼 요건 > 하둡



하둡(Hadoop)이란?

- 대용량 데이터 처리 분석을 위한 대규모 분산 컴퓨팅 지원 프레임워크
- 대용량 저장과 처리(계산)
- HDFS를 통해 분산 저장하고, 맵 리듀스를 통해 분산 처리함
- 하둡은 여러 개의 컴퓨터를 마치 하나인 것처럼 묶어주는 기술을 통해 저장 공간과 계산 능력을 늘림

분석 플랫폼 요건 > 하둡 > **IBM 병렬 분산 처리 시스템**

하둡(Hadoop)을 적용시켜 비즈니스를 ?

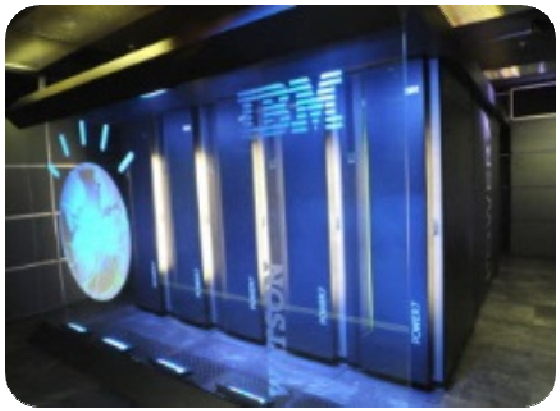
- CASE 1. 뉴욕 타임즈 - 130년 분량의 신문기사를 PDF화
  - 2008년 2월

→ 4TB → 24시간 소요

- CASE 2. 1분 SORT 대회 - 1,400노드 활용
  - 2009년 4월

→ 500GB → 59초 소요

IBM 병렬 분산 컴퓨팅도 비즈니스를 ?



- 슈퍼컴퓨터 '왓슨', 의료보험사 컨설턴트로 '취업'

슈퍼컴 왓슨의 경력 (2011. Feb)
•인간과 슈퍼컴의 격돌 (미국, 제퍼디 퀴즈쇼)
•3회 출현 7만7147달러의 상금 획득
•경쟁 대상 : 켄 제닝스 (74연승), 브래드 루터 (325만달러의 상금왕)을 격파 !!

분석 플랫폼 요건 > 하둡 > **IBM 병렬 분산 처리 시스템**

IBM 병렬 분산 컴퓨팅이 어떻게 비즈니스를 ?

똑똑한 왓슨(Watson) - 복잡하고 방대한 언어를 신속하게 분석하는 슈퍼컴퓨터
•특징 : '빅 데이터' 처리에 최적화된 IBM 파워7 서버에 의해 구동되며, 막대한 양의 작업과 데이터를 동시 처리
•특허 기술 1 - 데이터를 처리하면서 실시간으로 정보를 분석하는 기술
•특허 기술 2 - 사람만이 알아들을 수 있는 비유나 상징, 언어의 숨은 의도 파악

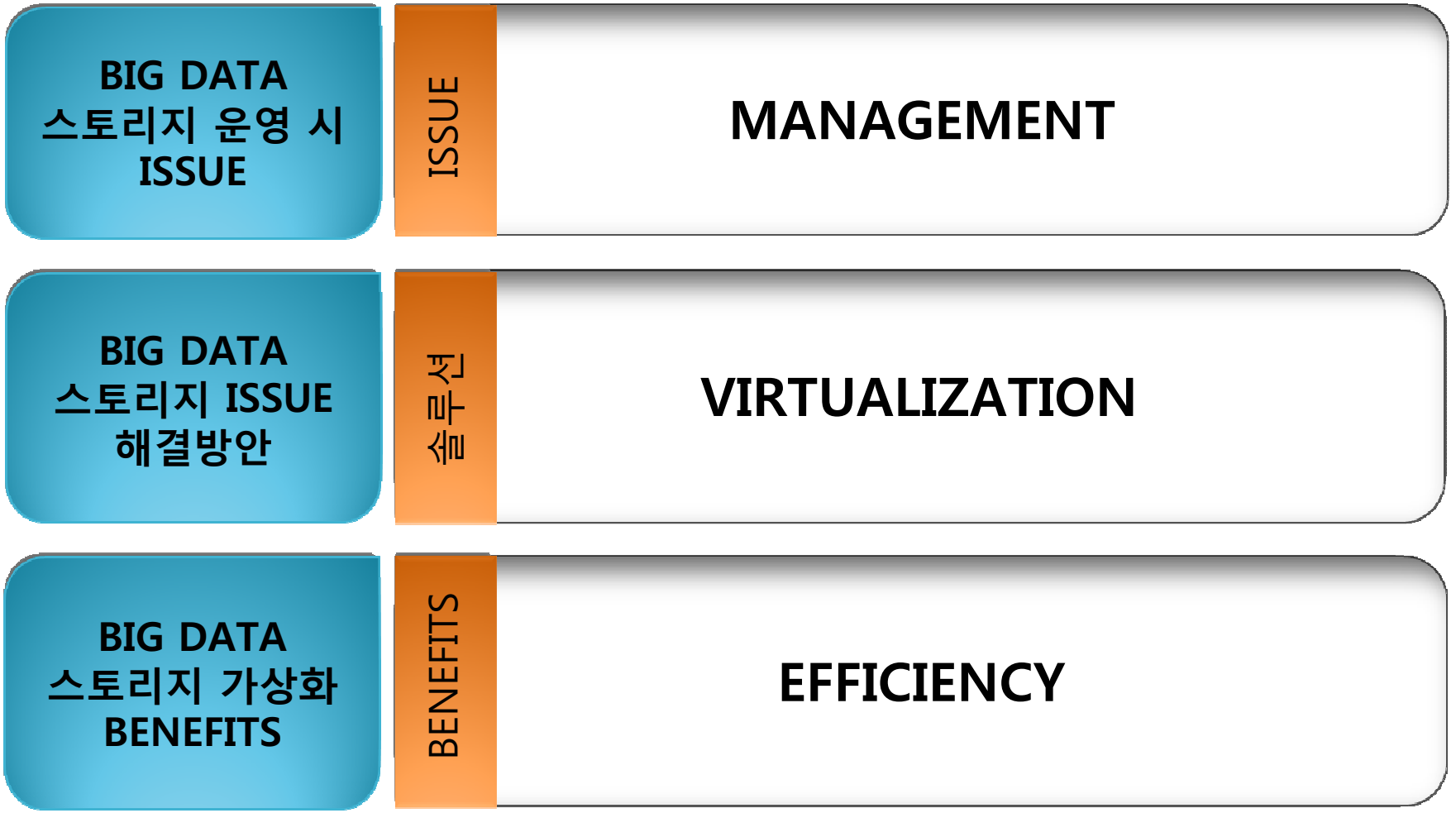
"왓슨의 빠른 데이터 처리속도를 바탕으로 건강보험 자료와 웰포인트 건강보험회사에 등록된 3420만명에 달하는 환자 정보를 통합하고, 이를 기초로 복잡한 의학적 치료법을 찾아내는 일을 담당하게 될 것"

"왓슨이 업데이트된 최신 정보를 과학적인 방법으로 제시하는 한편, 현재 환자에게 필요한 의학적인 요구사항을 포함해 모든 정보를 3초 안에 파악할 수 있다" - '샘 누스바움' 웰포인트 의학 최고책임자

IBM BIG DATA 분석 플랫폼

IBM BIG DATA 플랫폼 > Active Cloud Engine

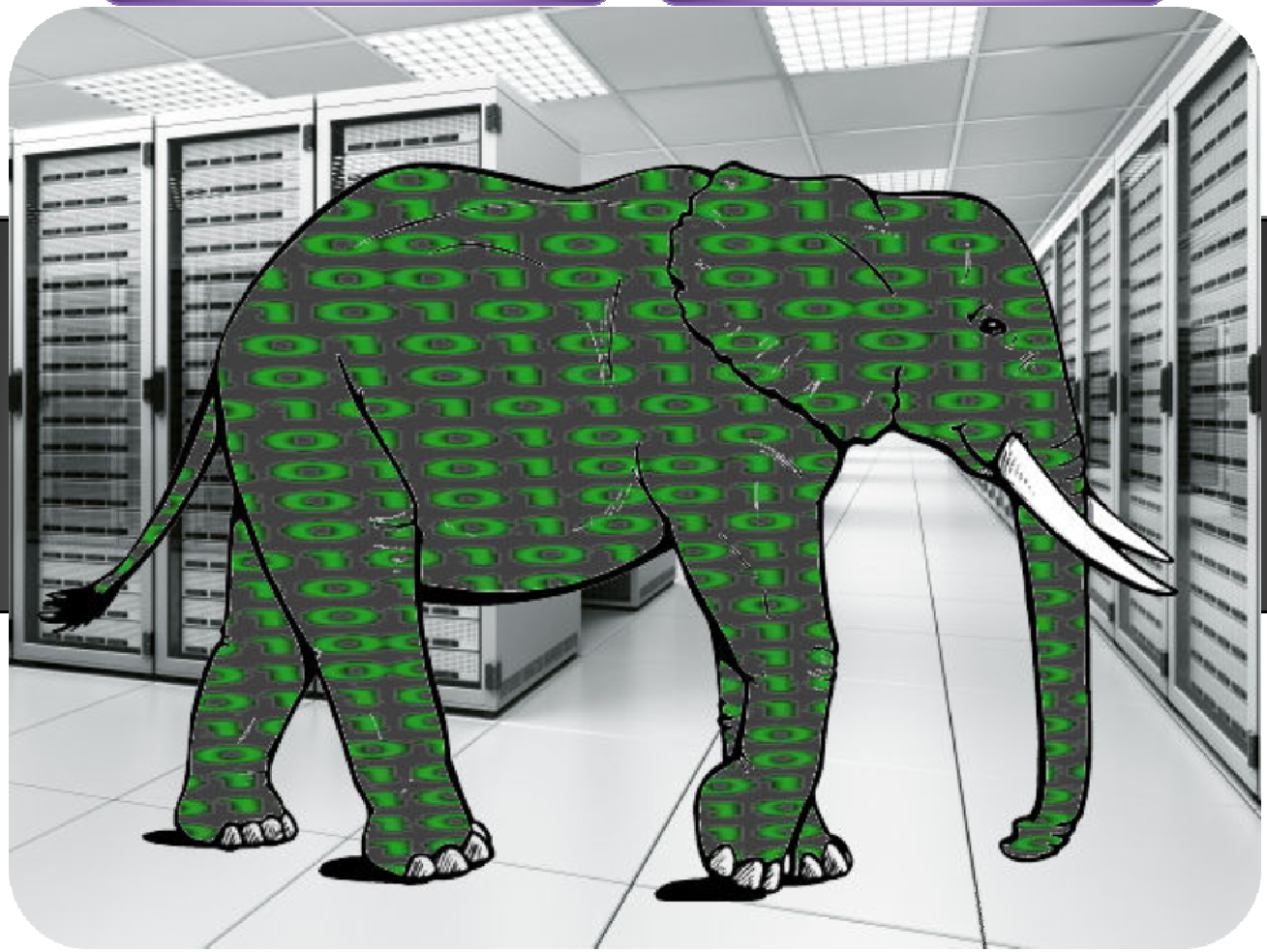
IBM BIG DATA 플랫폼 > **Active Cloud Engine** > SoNAS(Scale Out NAS)



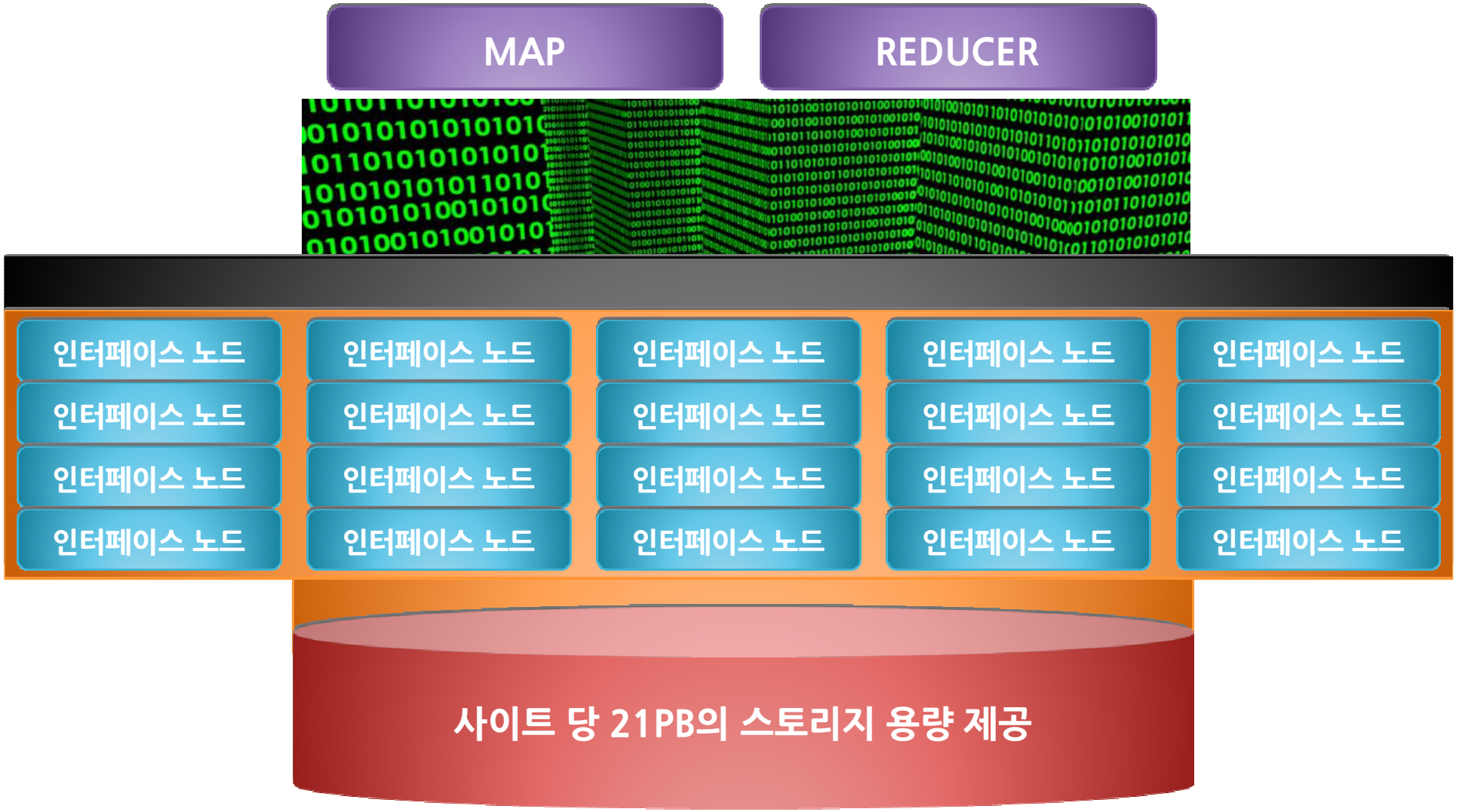
IBM BIG DATA 플랫폼 > **Active Cloud Engine** > SoNAS(Scale Out NAS)

MAP

REDUCER

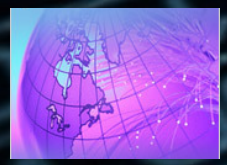


IBM BIG DATA 플랫폼 > **Active Cloud Engine** > SoNAS(Scale Out NAS)



IBM BIG DATA 플랫폼 > **Active Cloud Engine** > SoNAS(Scale Out NAS)

SoNAS to Scale out to Optimized Multiple Petabyte



싱글 네임 스페이스
단일 네임 스페이스를 제공하는 페타 스케일의 클러스터 NAS 시스템

ILM기반정보관리
사용자 데이터 기반의 통합된 데이터 관리와 TSM 연동을 통한 데이터 백업 및 복원

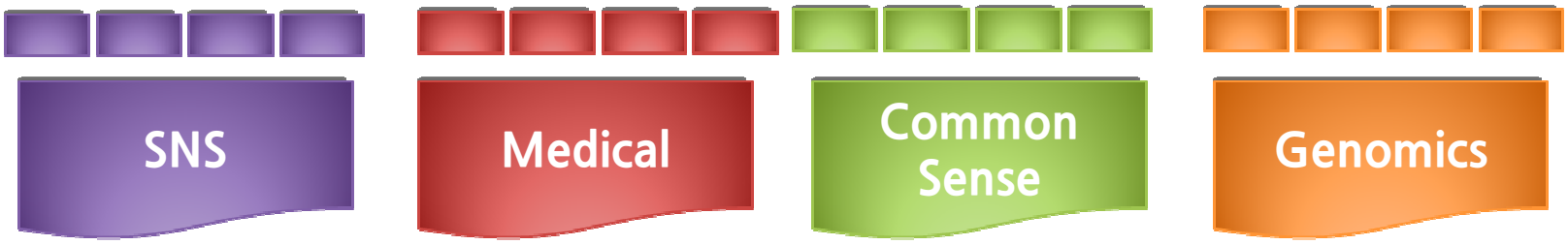
운영효율성
가상스토리지 리소스(Multiple-Tenancy) 제공 및 100% 디스크 스토리지 시스템 Utilization

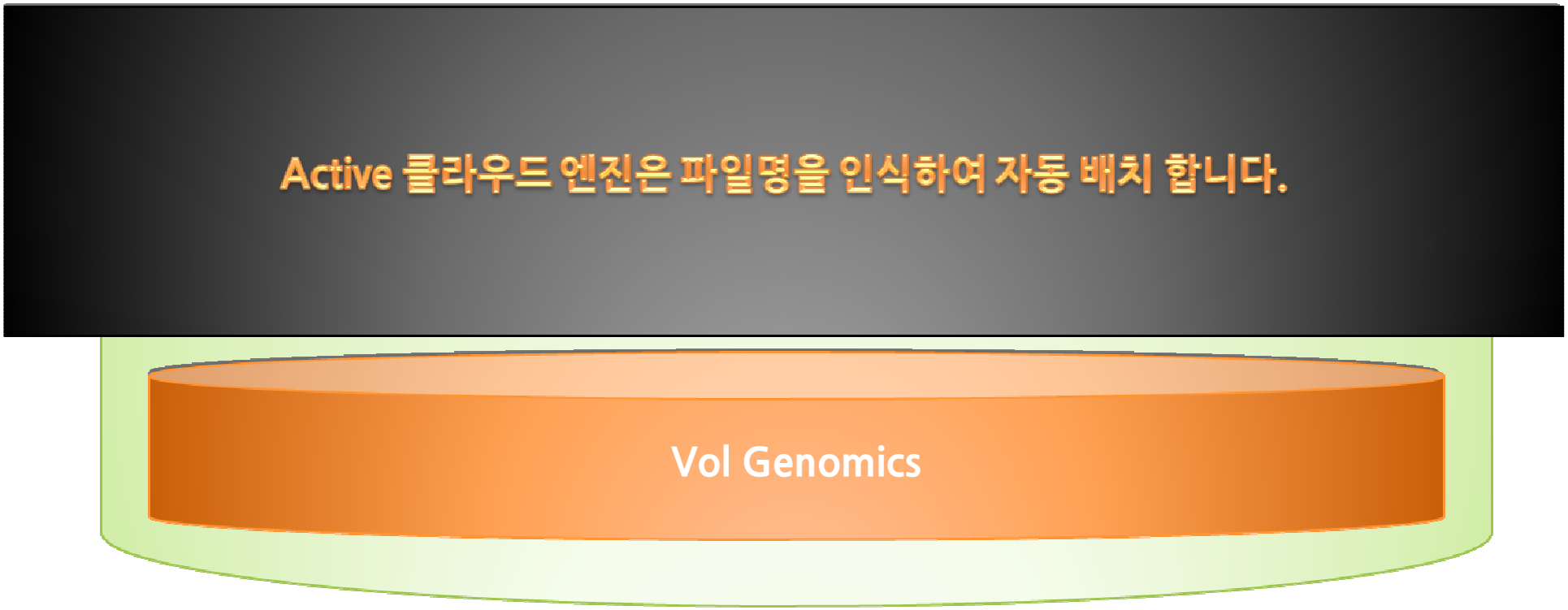
TCO 격감 및 ROI 최대화의 경제적 솔루션
비즈니스 환경을 위한 서비스 중심의 Platform 제공 및 서비스 및 어플리케이션 가속화의 경제적인 솔루션

IV

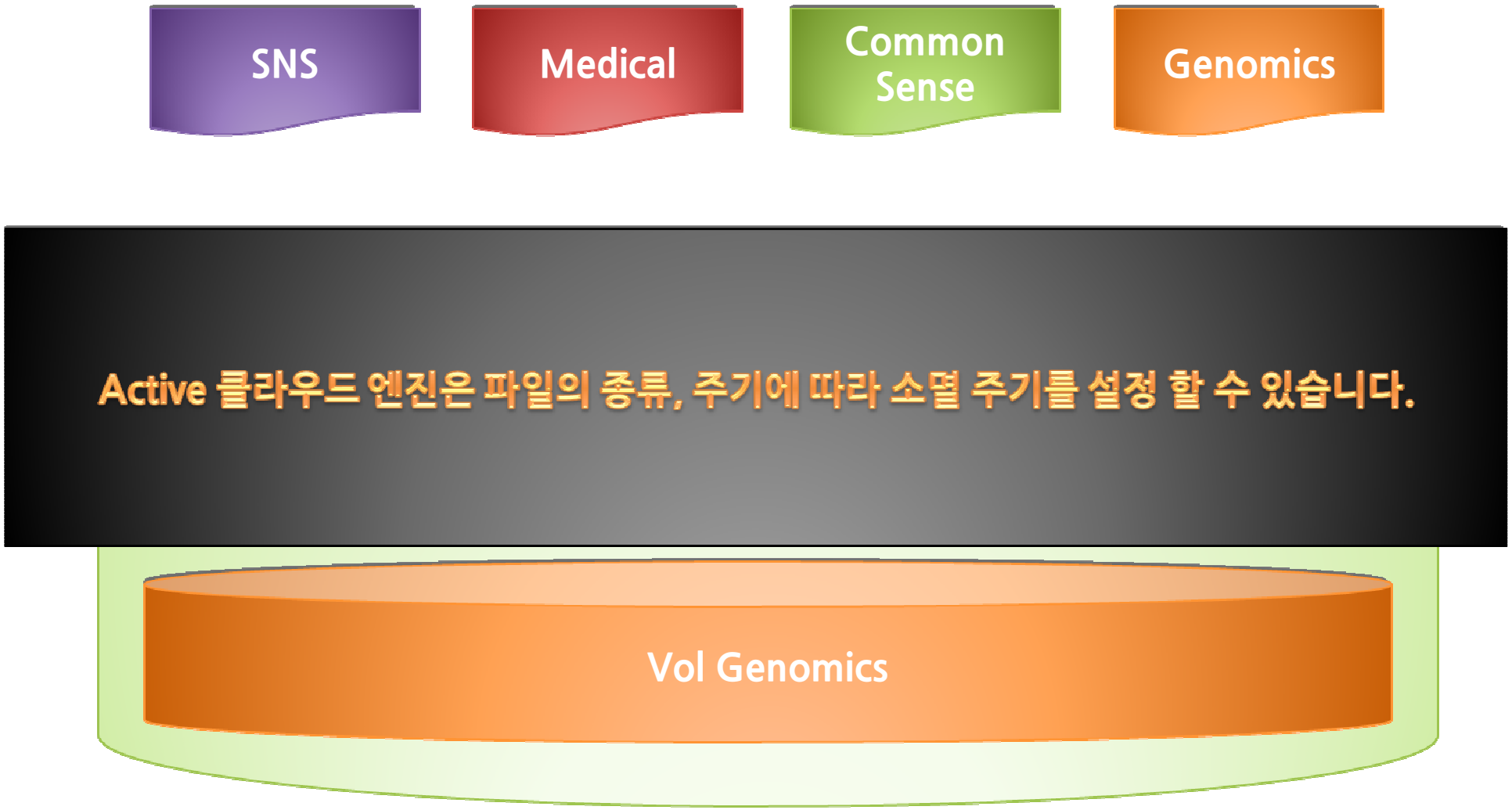
BIG DATA의 데이터 관리 및 최적화

IBM BIG DATA 플랫폼 > Active Cloud Engine > 데이터 관리 1

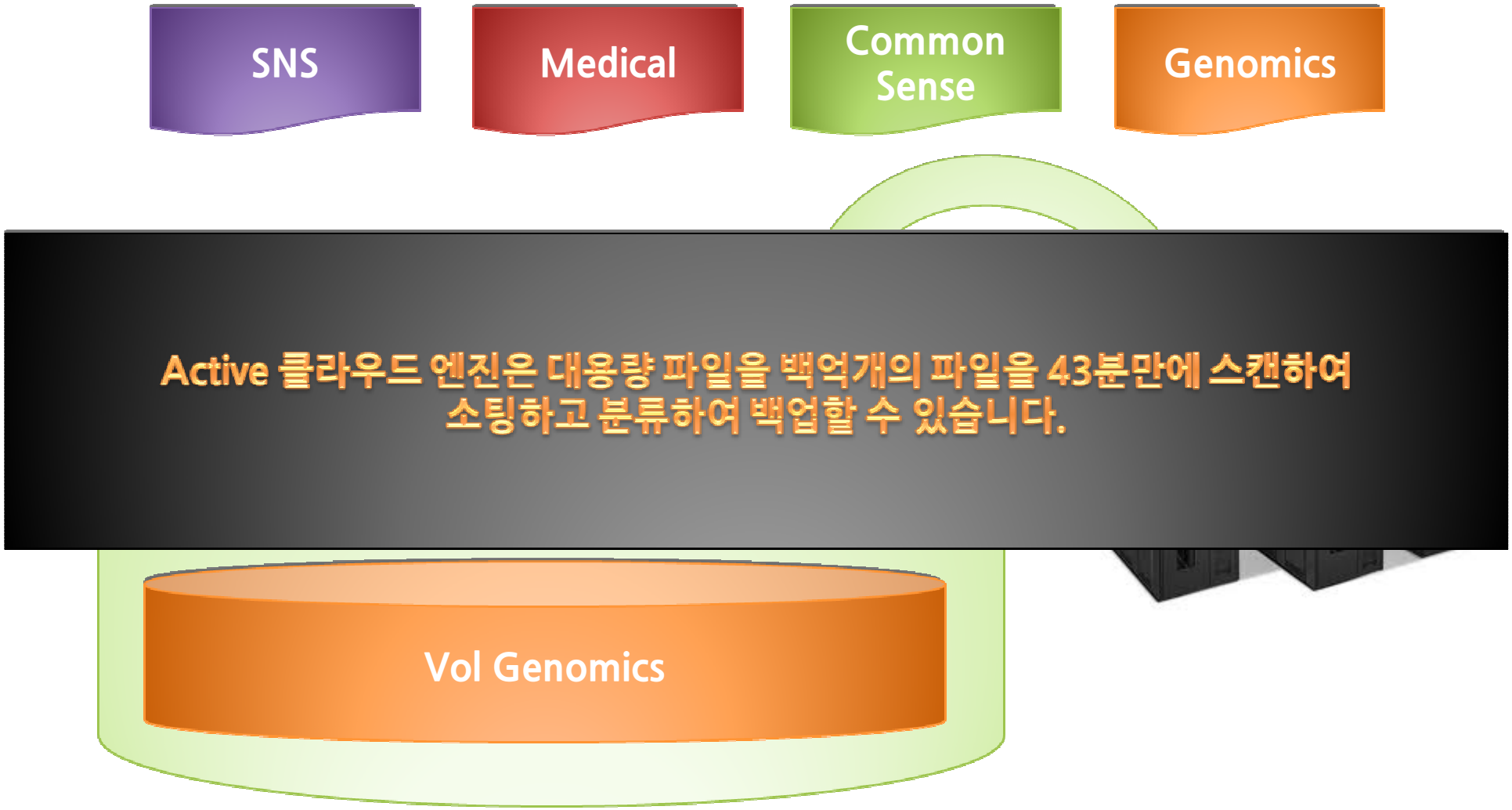




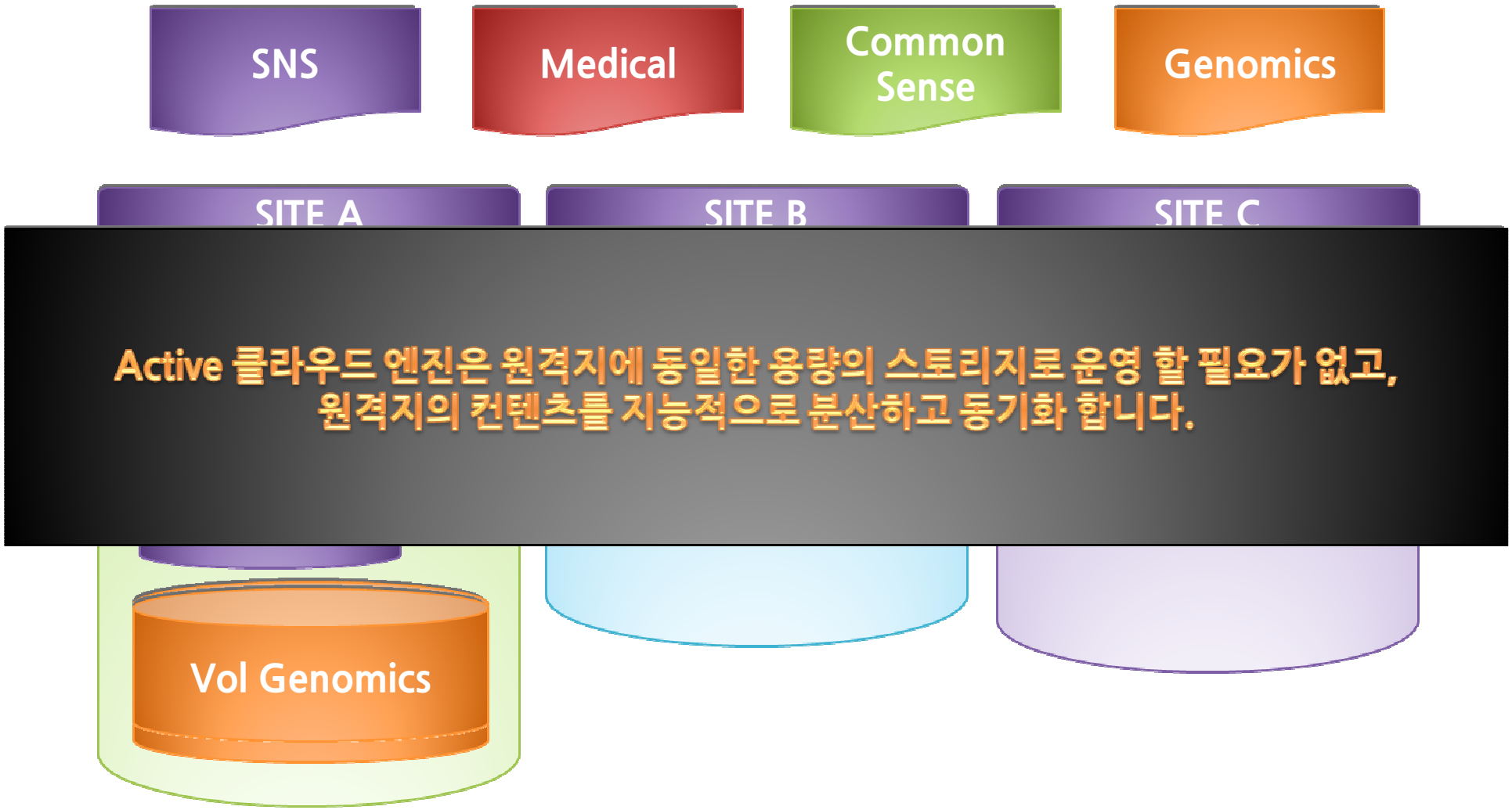
IBM BIG DATA 플랫폼 > Active Cloud Engine > 데이터 관리 2



IBM BIG DATA 플랫폼 > Active Cloud Engine > 데이터 관리 3



IBM BIG DATA 플랫폼 > Active Cloud Engine > 스토리지 관리



IBM BIG DATA 플랫폼 > Active Cloud Engine > 스토리지 관리

IBM Active Cloud Engine

IBM SoNAS Platform
w/ DCS6620, XIV, V7K

IBM V7K Unified

Active 클라우드 엔진을 탑재하여 다양한 BIG DATA를 고속으로 처리하며,
자동으로 주기 관리를 합니다.

V

결 론

IBM BIG DATA 플랫폼 > 결론

결론1 누구도 피해 갈 수 없는 BIG DATA 는 IT의 필수 요소입니다.

결론2 BIG DATA의 활용은 BIG DATA의 체계적인 저장에서 시작되며, 분석을 통하여 비즈니스에 적용 할 수 있습니다.

결론3 가상화된 BIG DATA의 플랫폼에 Active Cloud Engine을 탑재한 IBM 의 SoNAS, V7K 플랫폼은 BIG DATA 분석을 고속화 하고, 데이터 주기를 자동관리, 스토리지 관리를 최적화 합니다.

결론4 IBM의 BIG DATA 플랫폼을 통하여 BIG DATA의 고민으로부터 자유로워질 수 있습니다.

참석하신 여러분께 진심으로

감사 드립니다.

* 문의처 : 김정림 차장 / 010-4995-8576 / jlkim@kr.ibm.com