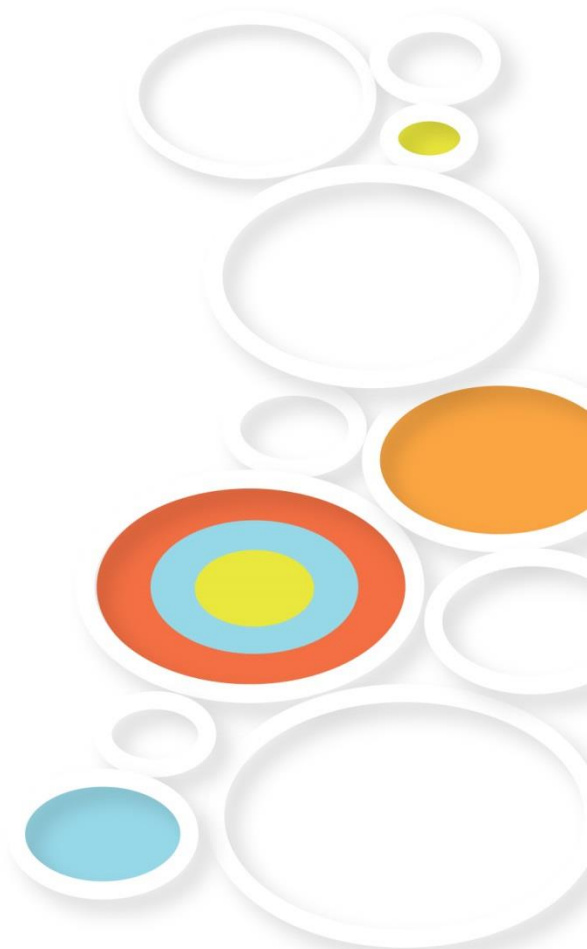


SAS High-Performance Analytics : Big Data Analytics를 위한 기술 혁신

SAS Korea 김근태



빨라진 의사 결정

빅데이터가 과거에는 불가능했던 새로운 기회를 제공합니다.

수 일 또는 수 주일이 소요되었던 분석 인사이트를 수 분 또는 수 초 내에 확보할 수 있습니다.

What if you could....



... Big Data를 **경쟁사보다 며칠 더 빠르게** 가망 고객의 구매 행위와 의사결정 기준을 예측할 수 있다면...

...**누구도 시도하지 못했던** 초 세분화된 마켓을 타겟으로 하는 상품 및 서비스를 출시함으로써 시장선점우위를 가져갈 수 있다면 ...

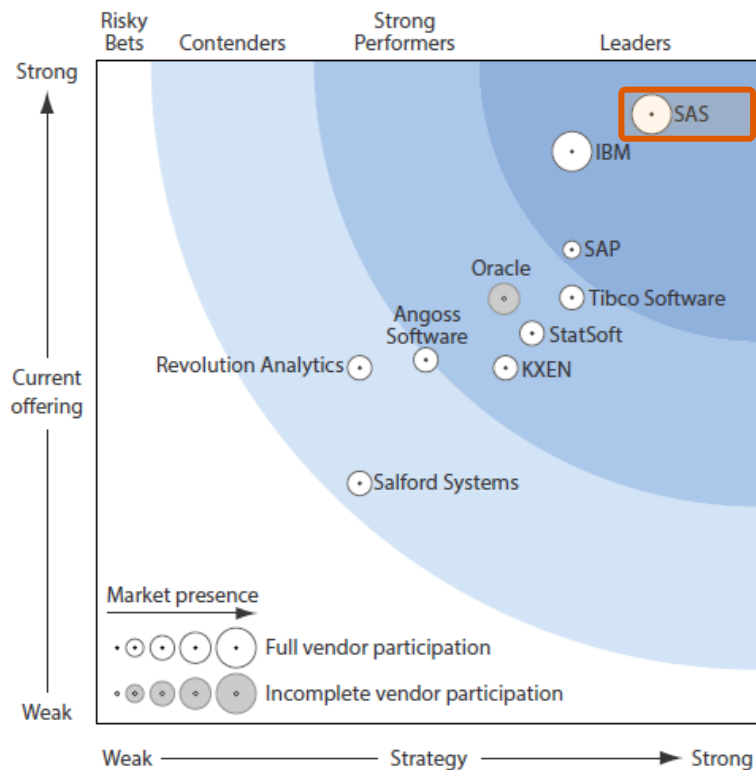
... 마케팅 캠페인 효과를 즉시 평가하여, **실시간에** **가깝게 전략 수정**을 할 수 있다면 ...

Big Data 시대의 분석 과제

분석 속성 Big Data 분석 요건

데이터 크기	Giga Bytes → 수백 Tera ~ Peta Bytes
데이터 종류	구조화된 데이터 → 비정형 Text Data로 확장
데이터 성격	Legacy Data → SNS, Log, CDR, FDC, Packet, Meter...
분석 시간	수 시간 이상 (Batch) → 수 초, 수 분 이내 (Near Real-time)
수행할 분석의 양	수십개의 모델 → 수천~수만개 이상의 세분화 된 모델
사용자 환경	전문가 환경 → Data Visualization, 분석 모델링 자동화

ANALYST REPORTS - FORRESTER WAVE™ (Q1 2013)



BIG DATA PREDICTIVE ANALYTICS GLOBAL TOP 벤더

The Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 2013, Forrester Research, Inc., January 3, 2013.

The Forrester Wave is copyrighted by Forrester Research, Inc. Forrester and Forrester Wave are trademarks of Forrester Research, Inc. The Forrester Wave is a graphical representation of Forrester's call on a market and is plotted using a detailed spreadsheet with exposed scores, weightings, and comments. Forrester does not endorse any vendor, product, or service depicted in the Forrester Wave. Information is based on best available resources. Opinions reflect judgment at the time and are subject to change.

Big Data Analytics



Big Data 분석을 위한 기술 혁신



Big Data 분석 1세대 – Big Data 저장 Platform과 고급 분석 기능 연계

MPP Architecture (Multi-Node)

DBMS Appliance / Hadoop



Teradata 700
appliance



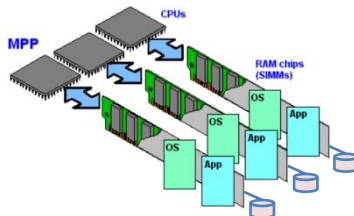
EMC
Greenplum
DCA



Oracle
Exadata



IBM ISAS



MPP
(Massively
Parallel
Processing)

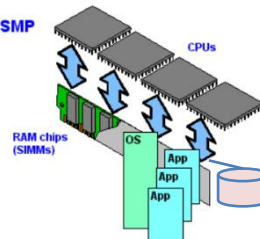
Access
Interface
/ ETL

SMP Analytic Architecture (Single-Node)

Analytics Functions

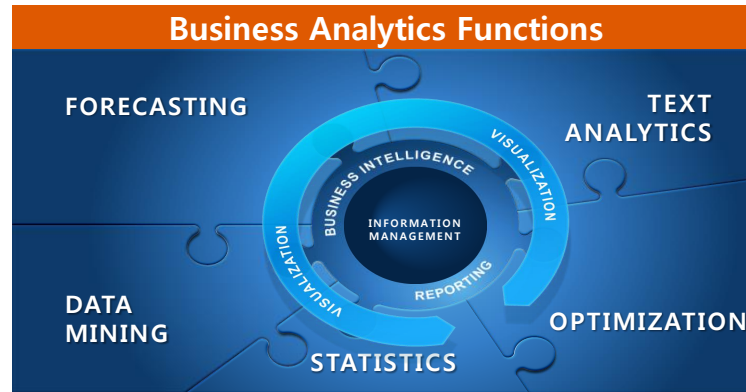


SMP



SMP
(Symmetric
Multi-Processing)

Big Data 분석 2세대 – Analytic Platform 혁신



MPP Hardware (Multi-Node)



Teradata 700
appliance



EMC²
Greenplum
DCA



Oracle
Exadata

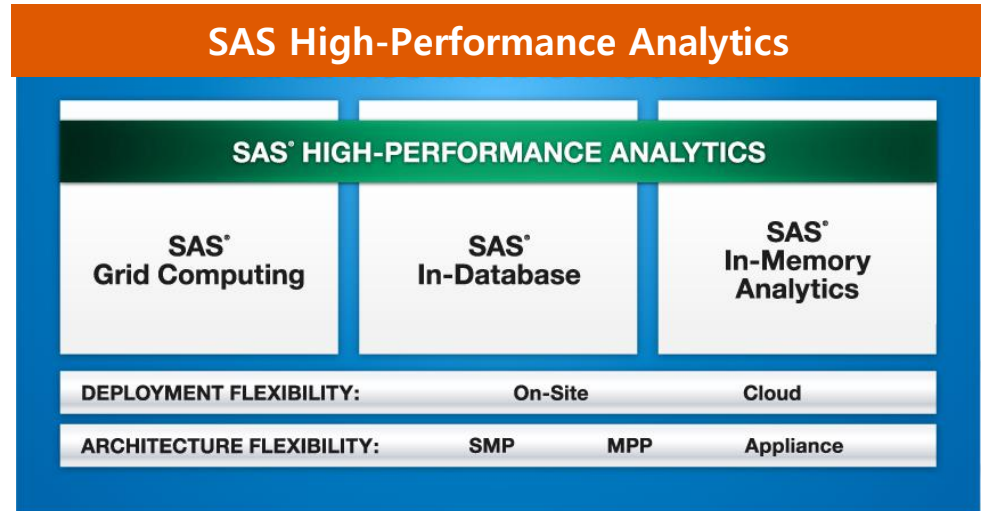


Analytic Platform 혁신 - SAS High-Performance Analytics

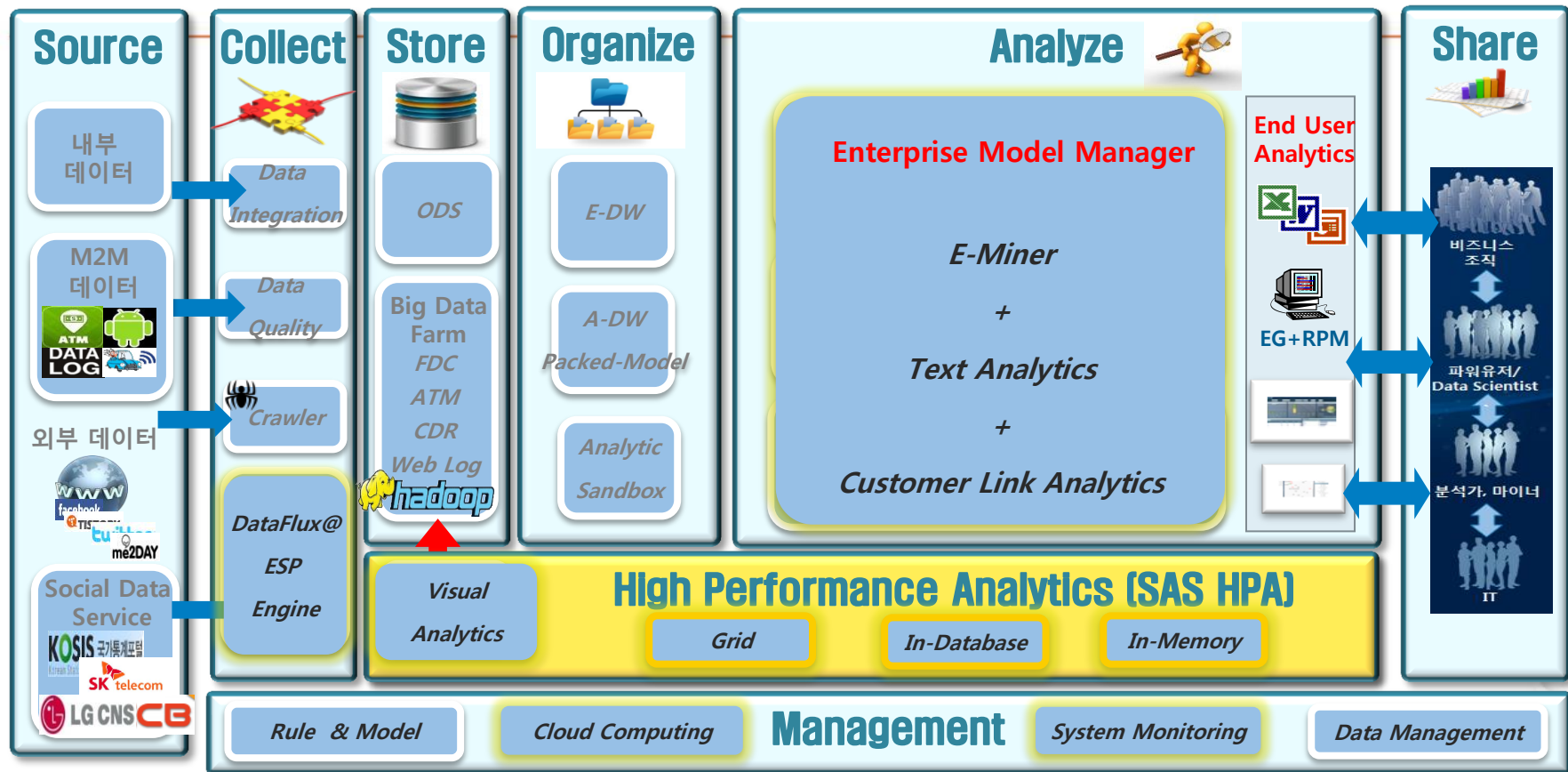
Business Analytics Functions



SAS High-Performance Analytics



SAS Big Data Analytics Platform



NOTE: SAS (r) Proprietary Software 9.3 (TS1M0)
Licensed to disney demo, Site 70068128.
NOTE: This session is executing on the X64_7PRO platform.

NOTE: SAS initialization used:
real time 0.66 seconds
cpu time 0.63 seconds

NOTE: AUTOEXEC processing beginning: file is C:\Users\olscha\autoexec.sas.

HPA_demo.sas

```
title1 "Logistic Regression With One Billion Records";
proc logistic data=GPlib.sgf_binary;
  class A B C;
  model y = a b c x1 x2 x3;
  performance details;
run;

title1 "Forward Selection, 120 Million Records, 2042 Parameters";
proc hpreg data=GPlib.reg_simtbl;
  class cls6-cls20;
  model y = cls6-cls20 x1-x100
           x1|x2|x3|x4|x5|x6|x7|cls7|cls8|cls9@2;
  selection method=forward hierarchy=single;
  performance details;
run;
```

Output - (Untitled)

Log - (Untitled)

HPA_demo.sas

Ln 1, Col 1

```
NOTE: SAS (r) Proprietary Software 9.3 (TSIM0)  
      Licensed to disney demo, Site 70068128.  
NOTE: This session is executing on the X64_7PRO platform.  
  
NOTE: SAS initialization used:  
      real time      0.66 seconds  
      cpu time       0.63 seconds  
  
NOTE: AUTOEXEC processing beginning; file is C:\Users\olscha\autoexec.sas.
```

In-Memory Analytics로 10억건 데이터에 대해 로지스틱 회귀분석 실행

결과 : proc logistic (기존 20시간) → proc hplogistic (50초)

32개 node의 DW Appliance (32 * 24 = 768 units of parallelism)



Big Data 분석을 위한 실체적인 End-to-End 기능 제공

SAS High-Performance Analytics - Procedures

데이터 가공

- HPDS2
- HPDMDDB
- HPSAMPLE

데이터 탐색

- HPSUMMARY
- HPDMDDB
- HPSAMPLE
- HPREDUCE
- HPIMPUTE*
- HPBIN*

변수 변환

- HPREDUCE
- HPIMPUTE
- HPBIN
- HPDS2

예측 모델링

- HPREG
- HPLOGISTIC
- HPNLIN
- HPNEURAL
- HPLMIXED
- HPFOREST
- HPSVM
- HPDECIDE

모델 선정, 테스트, 적용

- HPDS2
- HP4SCORE
- SAS Scoring Accelerator
- SAS Model Manager

시계열 분석

- HPNLIN, HPCOUNTREG, HPSEVERITY
- Mid-term: HPSIMILARITY
- Long Term: HPSIMULATE, HPCOPULA, HPQLIM, HPPANEL, HPMDC

텍스트 분석

- Integrate text into DM processes: Pattern Discovery, Predictive Modeling
- HPDM nodes for Text: Parse Node, Transform node
- HPDM procedures for Text: HPTXTPARSE, HPSVD

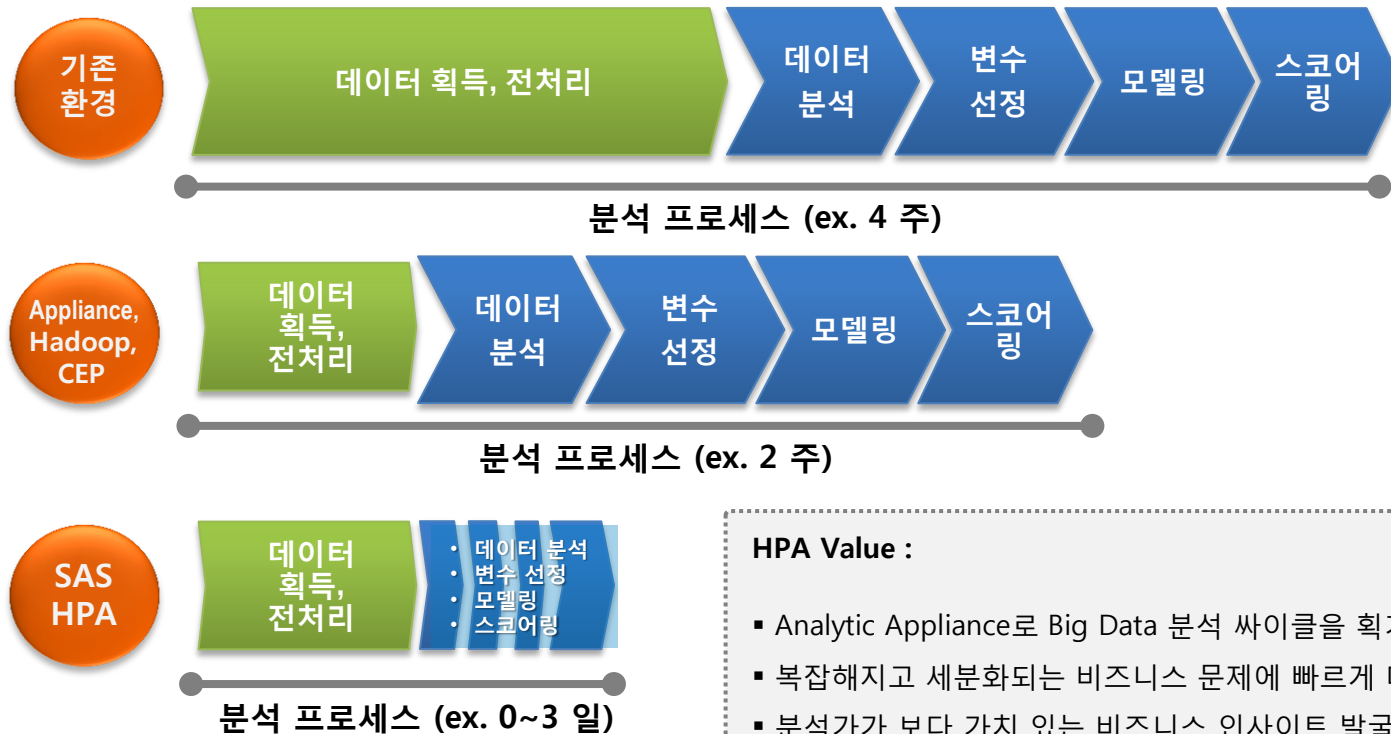
최적화

- HPMO (Marketing Optimization Solver)
- HPOPTGRAPH (Graph-algorithms & Network analysis)

HP Data Mining



High-Performance Analytics 기술이 필요한 이유 #1



High-Performance Analytics 기술이 필요한 이유 #2

기존 품질 관리, 캠페인, Risk,
Fraud, 최적화 모델의 한계

HPA는 어떻게 해결하는가 ?

HPA Value

샘플링 기반..

Big Data (col/row) 활용 예측

분석의 정확도 ↑

세분화되지 않은..

더 많은 모델 생성
→ 세부 단위 모델

분석의 깊이와 적용 범위 ↑

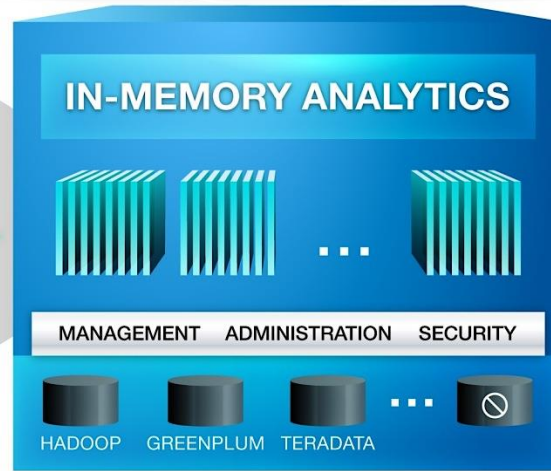
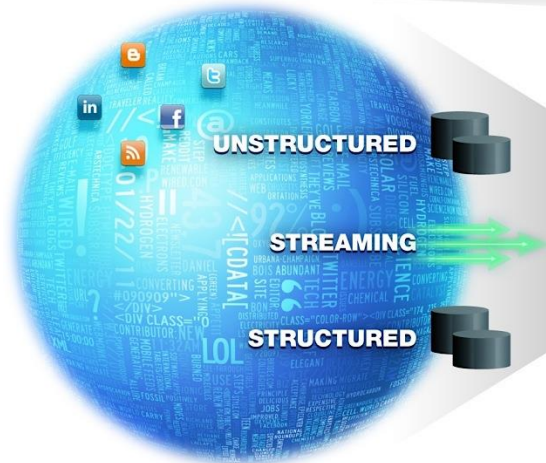
유통기한 없는..

Biz. 환경 변화에 빠르게 대응
(model life cycle 단축)

분석의 적시성 ↑

SAS High-Performance Analytics : In-Memory Analytics

TOOLS	ANALYTICS	ANALYTIC APPLICATIONS
Data Visualization Reporting	Descriptive Statistics Predictive Analytics Model Development Text Mining Forecasting Optimization ...	Retail Planning Revenue Optimization Marketing Optimization Stress Testing Liquidity Risk Management Fraud Detection ...



HPA 1.0 → HPA 2.0

High-Performance Analytic Server의 진화

Hadoop Cluster 기반 SAS Analytic Appliance

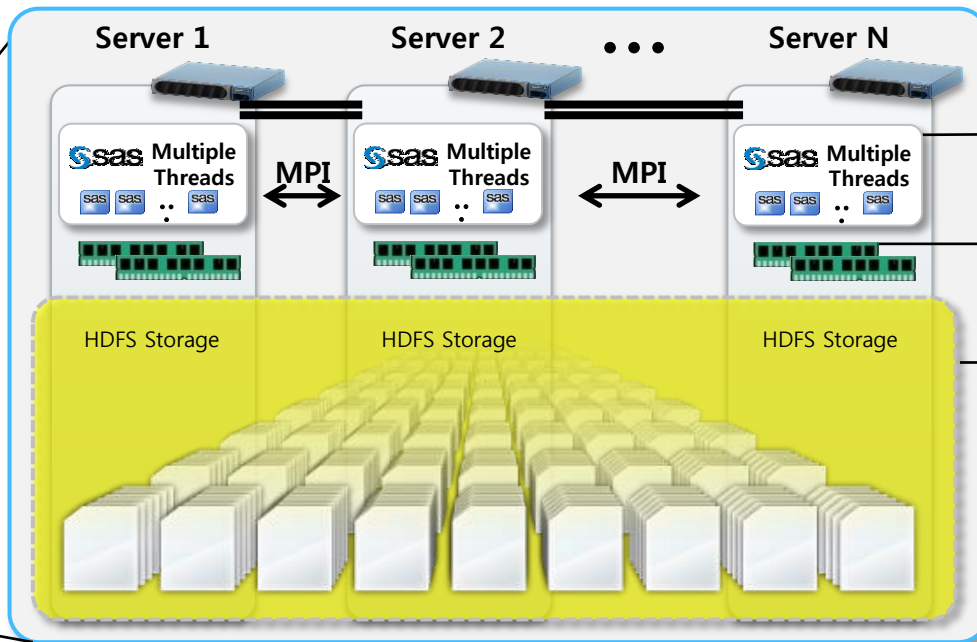
SAS STAT Client
SAS Enterprise Miner



HP C7000 Chassis
Holds BL460 Blades



Dell PowerEdge M1000e Chassis
M620 Blade Server



SAS High
Performance
Deployment

SAS In-Memory
Analytics



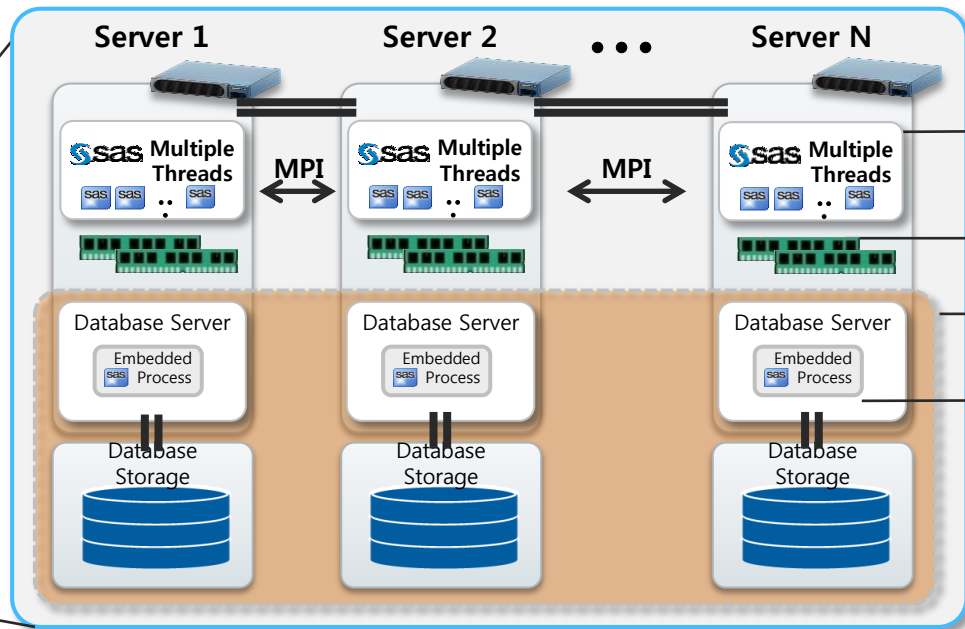
SAS High-
Performance
Analytics Platform

DB Appliance 기반 SAS Analytic Appliance

SAS STAT Client
SAS Enterprise Miner

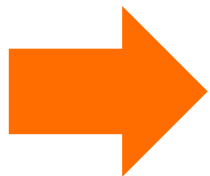


Analytic Appliance



HPA 1.0 → HPA 2.0

HPA 1.0
Fit
to
Appliance



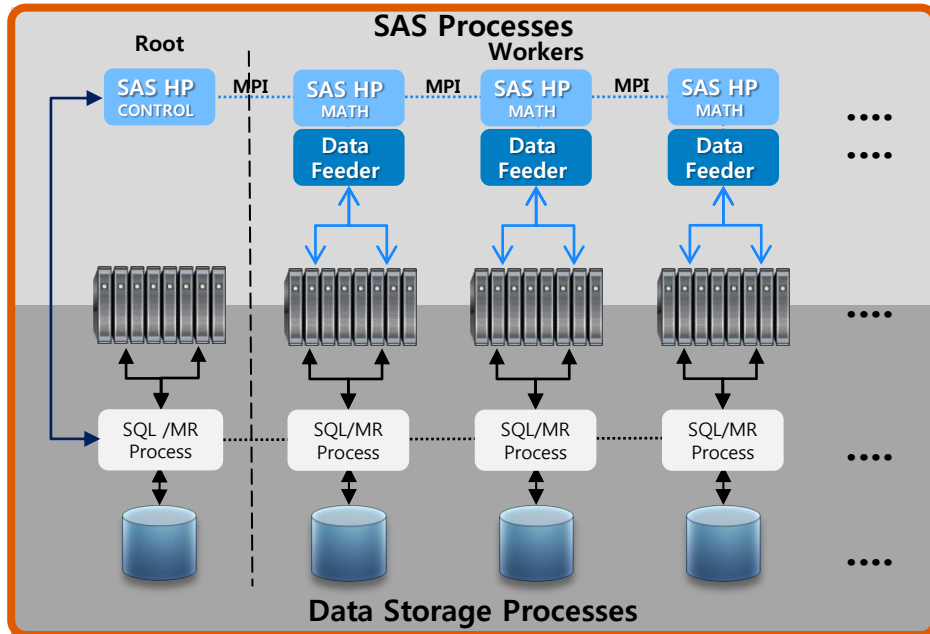
HPA 2.0 (SAS 9.4)
Now Fit
to
Data Center As Well

- ✓ 데이터 중복, 동기화 문제 해결
- ✓ 보다 유연한 아키텍처
- ✓ 보다 많은 H/W Platform 지원

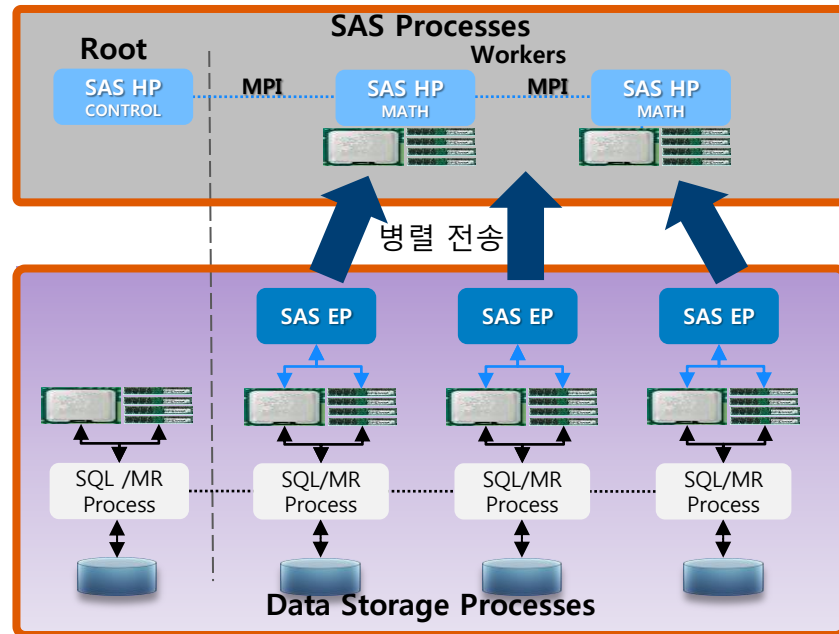


SAS Analytic Appliance – Symmetric / Asymmetric 구조 비교

HPA 1.0 (Symmetric Mode)

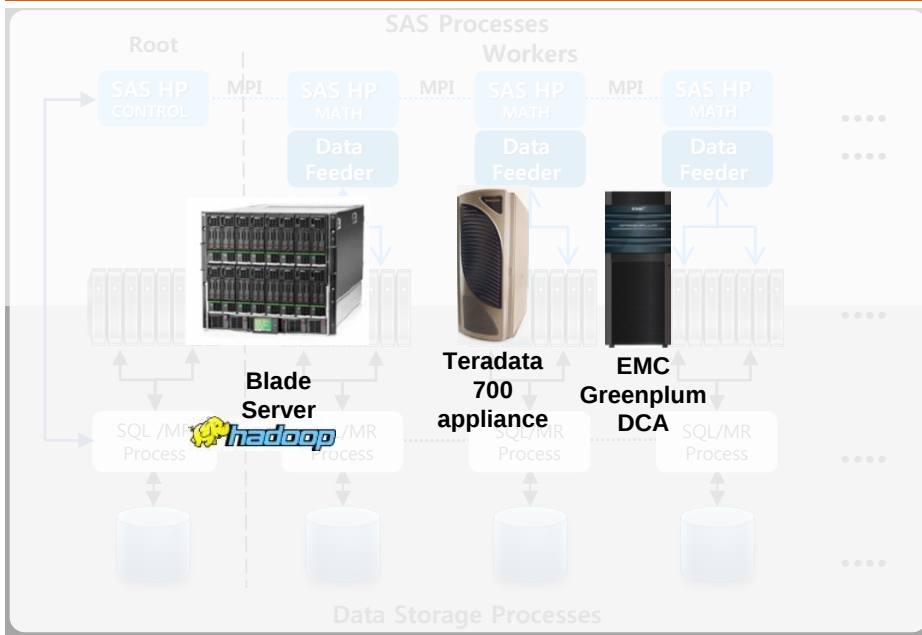


HPA 2.0 (Asymmetric Mode)



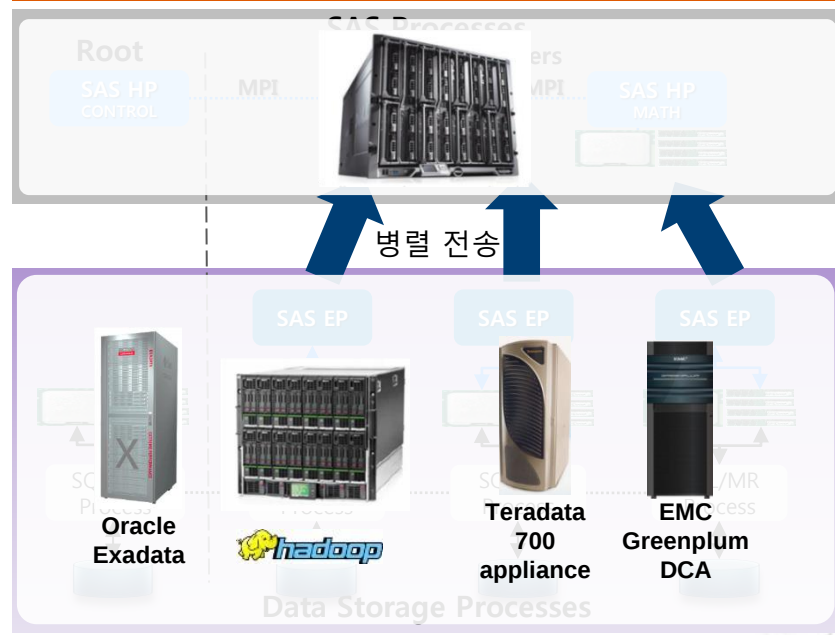
SAS Analytic Appliance – Symmetric / Asymmetric 구조

HPA 1.0 (Symmetric Mode)



Hadoop HDFS, Greenplum, Teradata

HPA 2.0 (Asymmetric Mode)



+ Oracle Exadata, Common Hadoop

High-Performance Analytic Server

- 적용 사례

SAS HPA : 산업별 Big Data Analytics 적용 사례



금융	공공	통신	유통	제조
				
리스크 관리	탈세 적발	캠페인 최적화	재고 관리	Warranty Analysis
• 365배 빠른 리스크 값 산출 • 신속한 시장 변화 대응	• 정교한 탈세 적발 • 부정 환급 적발	• 15% 향상된 캠페인 응답률	• 매장, 상품 속성별 최적화 30시간 → 2시간	• 워런티, 콜센터 텍스트 데이터 분석 • 문제점 사전 해결을 통한 품질확보 및 고객만족

Business Problem	Analysis	Before HPA Server	After HPA Server
대출 연체 가능성 예측	• 10억건 데이터에 대한 회귀분석	11 ~ 20 시간	54 초 소요
리스크 익스포저 산출	• 회귀분석	167시간 시간 (1주)	84 초 소요
신규 오퍼링을 위한 고객 타겟팅	• 뉴럴 네트워크	• 5 시간 – 한 Model 수행 • 하루에 1 모델 처리 • 단일 알고리즘 • 뉴럴네트워크 트레이닝 7회 • 모델 리프트 :1.6%	• 3 minutes – 한 Model 수행 • 1 모델에 30분 소요 • 다양한 알고리즘 : Random Forest, SVM, Logistic Regression, NN • 뉴럴네트워크 트레이닝 5000회 • 모델 리프트 : 2.5%

United Healthcare Group

BUSINESS ISSUE

- 전자 의료 기록(Electronic medical records :EMRs)이 데이터 폭증을 가져옴.
- 모든 비정형 텍스트 데이터 활용 (레코드, 임상 노트, 이메일, 처방 내용등)
- 어떻게 의료 서비스의 질을 높이고 비용을 줄일 것인가?
 - 처방전을 쓰기 전에 그 내용이 환자에게 줄 부작용 파악
 - 고객을 잃기 전에 그 고객의 불만을 이해
 - 비용청구 금액 지불 전에 허위 청구여부 적발

SOLUTION

- SAS® High-Performance Analytics Server with Greenplum (HP Text Mining 포함)

RESULTS

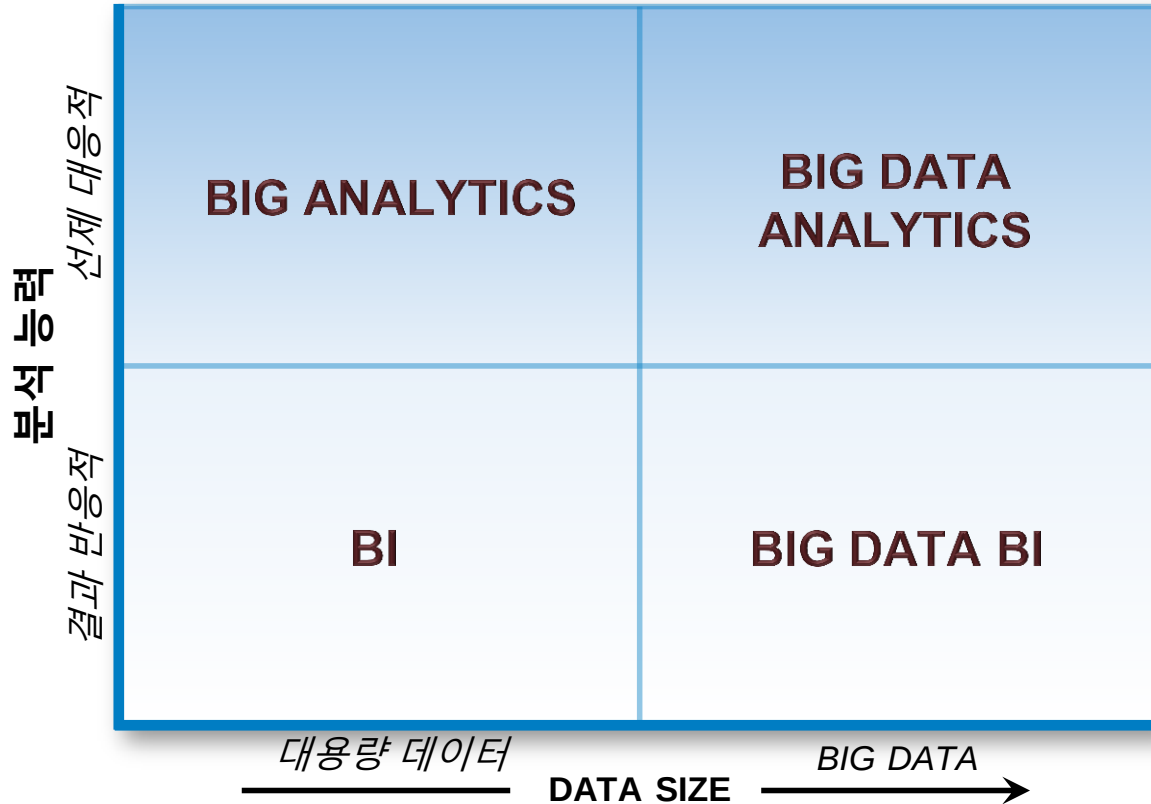
- 모델 프로세싱 시간이 4시간에서 10초로 단축
- 오분류 비율이 30% 에서 10%로 줄었으며, 분석 정확도 향상 (more than 10% lift)
- 발전 방향 : HPA로 최적화된 분석 사이클로 전사에 다양한 분석 서비스 제공
(**DAAS : Data Analytics as a Service**)

High-Performance Analytic Server

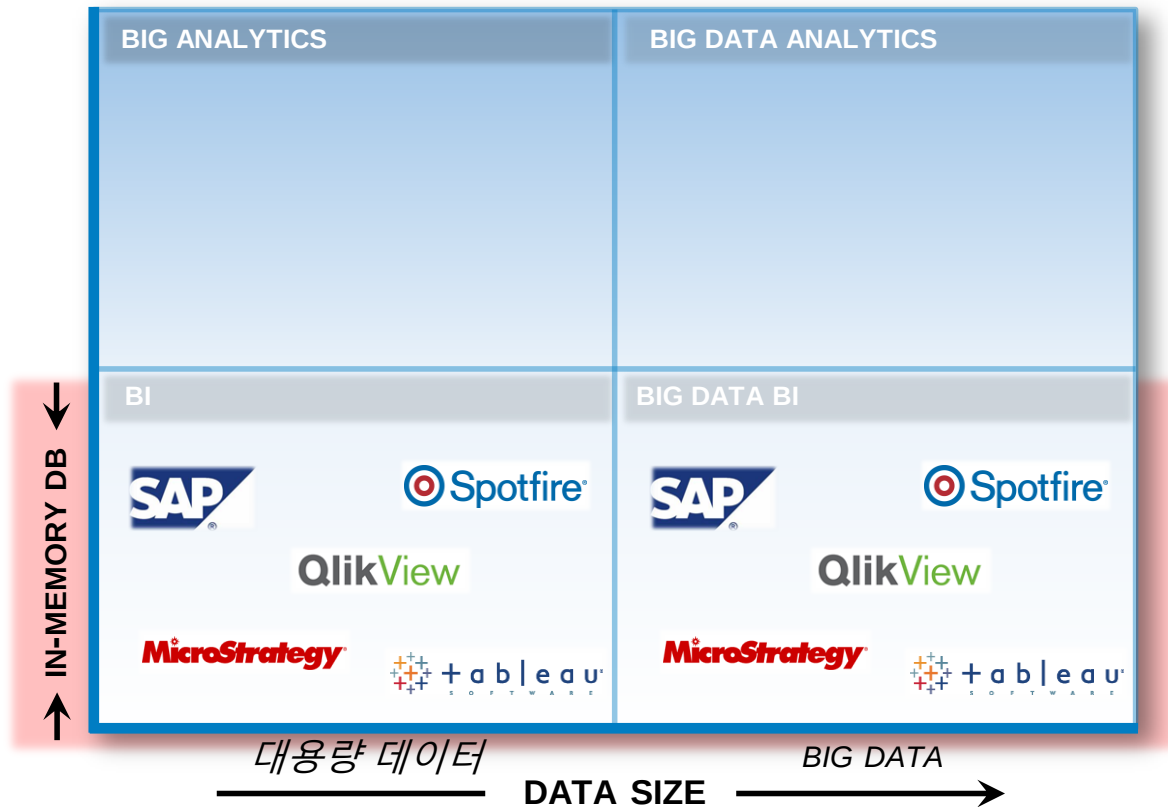
- Wrap Up



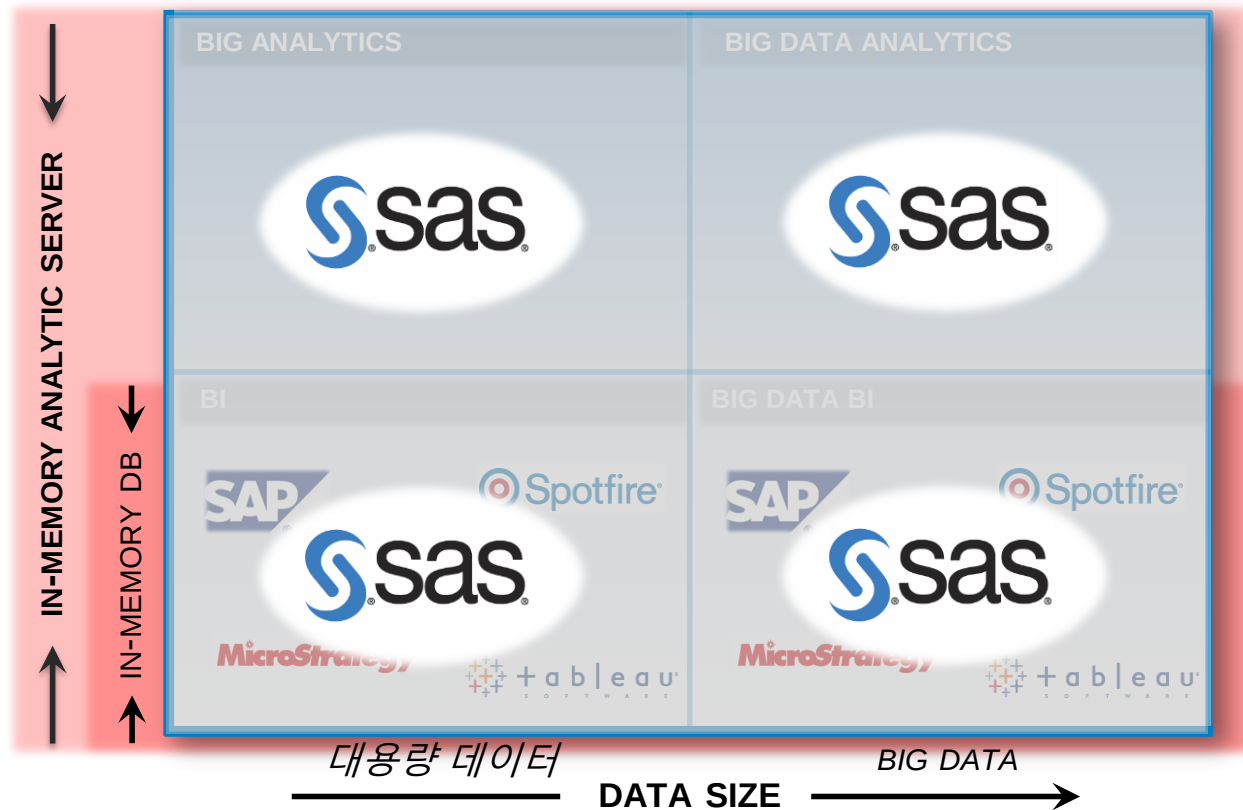
Big Analytics를 설명하는 새로운 관점들..



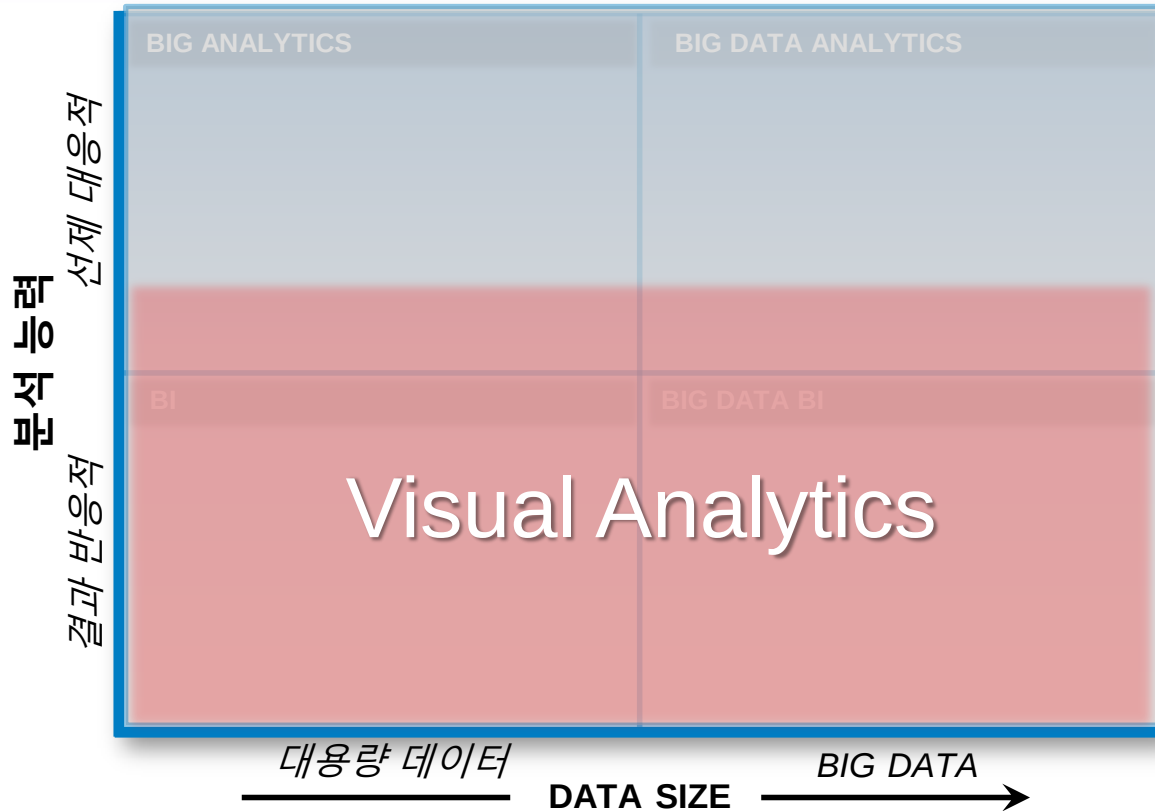
Big Analytics를 설명하는 새로운 관점들..



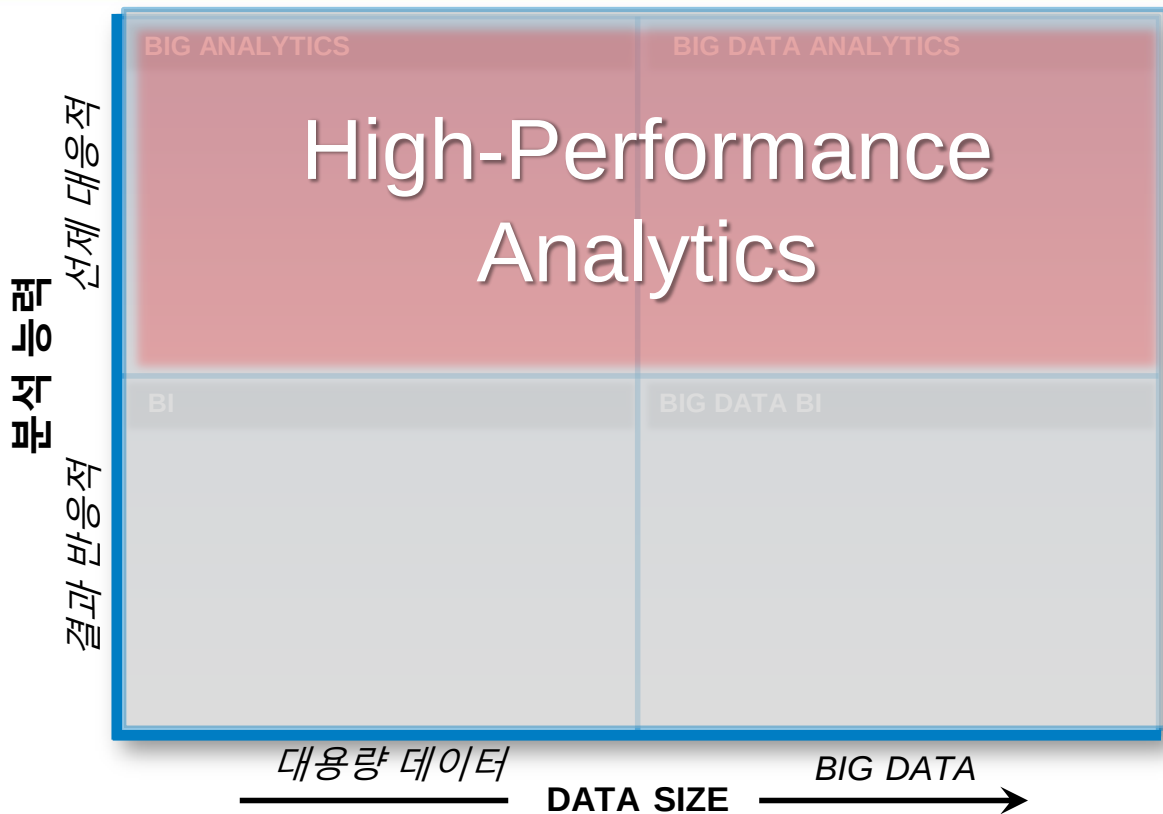
Big Analytics를 설명하는 새로운 관점들..



Big Analytics를 설명하는 새로운 관점들..



Big Analytics를 설명하는 새로운 관점들..



Big Data Analytics를 위한 기술 혁신, SAS® High-Performance Analytics



기존에 처리할 수 없었던 Big Data를 유용한 비즈니스 가치로 전환합니다.



세계 최고의 분석을 통해 정확한 통찰력을 제공합니다.



훨씬 빨라진 대응을 통해 비즈니스 운영 방식을 혁신시켜 줍니다.



미래 지향적이고 확장 가능한 분석 인프라로 진화시킵니다.

See Your Big Results with
SAS® Visual Analytics

