

HADOOP 정글 헤쳐나가기

DATA TO VALUE

2015.4.3



WHY HADOOP?

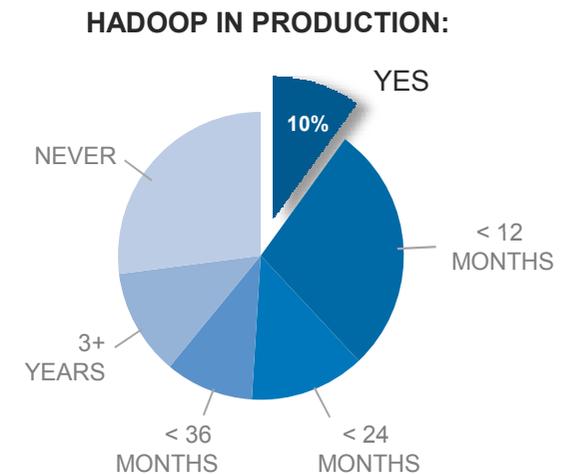


OPEN SOURCE
MASSIVE SCALE
FAST PROCESSING
COMMODITY COMPUTING
DATA REDUNDANCY
DISTRIBUTED

[Learn more >](#)

WHY HADOOP?

- Hadoop will soon become a *replacement complement* to:
 - ❑ Business Intelligence;
 - ❑ Data Warehousing;
 - ❑ Data Integration;
 - ❑ Analytics.
- 하둡 적용의 첫번째 목적 : Analytics (71%)
- 하둡 적용의 장벽 :
 - ❑ 내장된 분석 기능을 가지고 있지 않음
 - ❑ 비용 : 매우 많은 코딩으로 인한 비용 상승



SOURCE: [10 Myths About Hadoop - TDWI Best Practices Report](#)

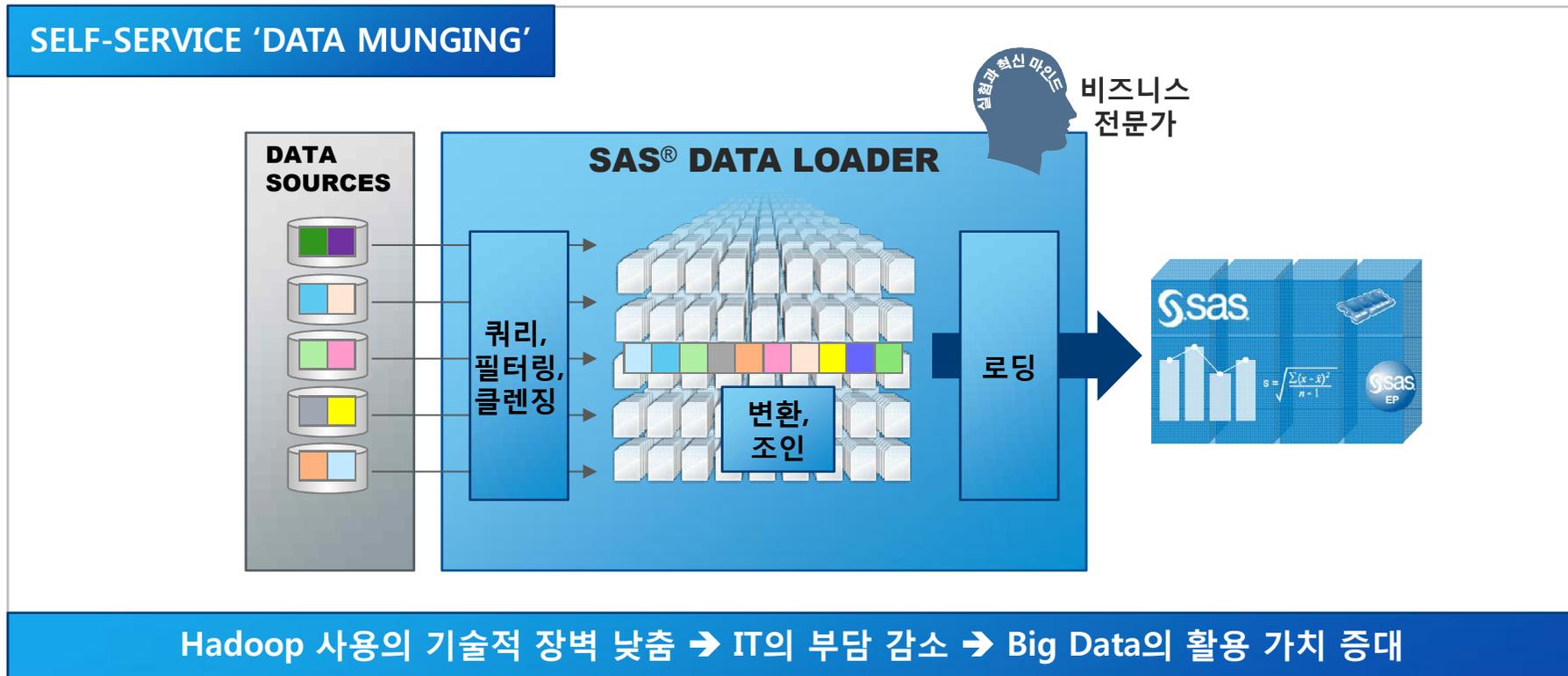
WHY SAS?

ANALYTICS
IN-MEMORY
HIGH-PERFORMANCE
DATA MANAGEMENT
BUSINESS INTELLIGENCE
DATA VISUALIZATION



Learn more >

(1) DATA PREPARATION

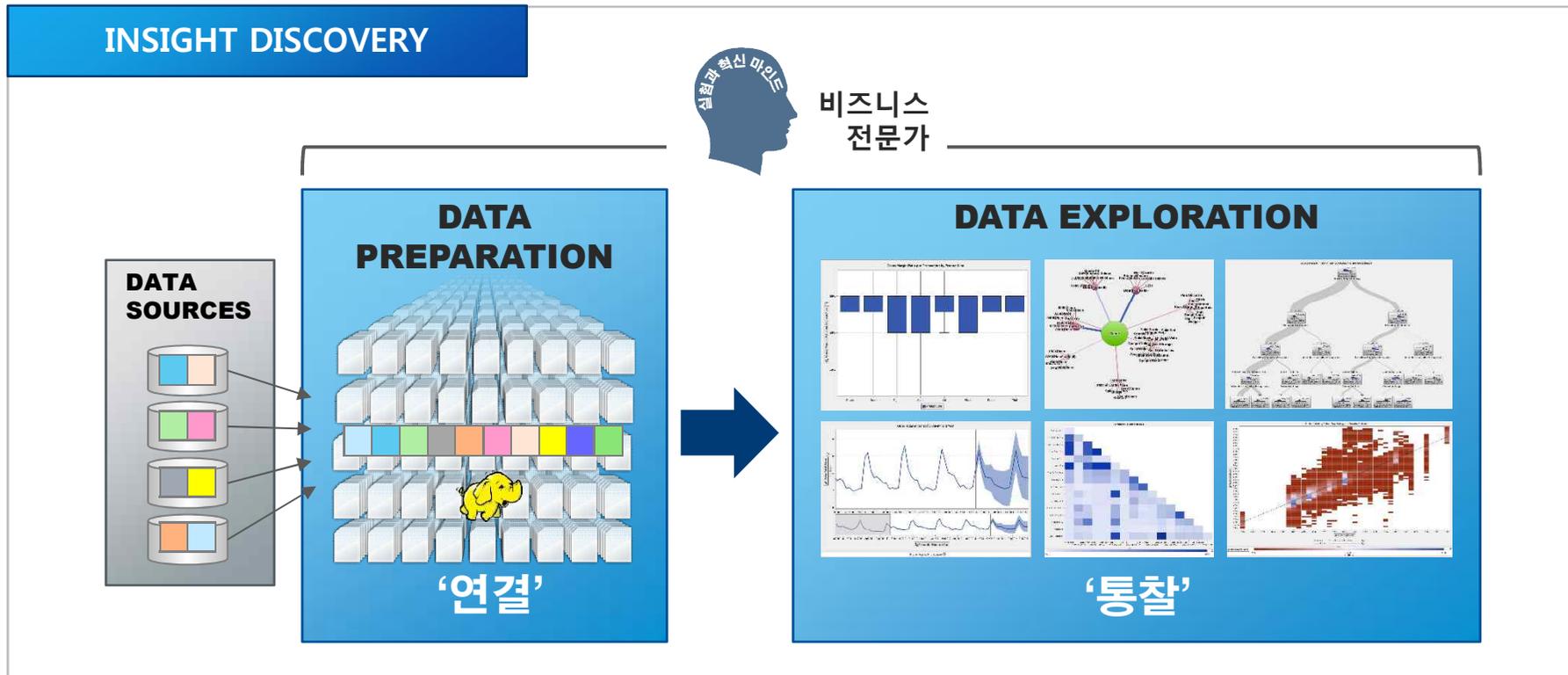


SAS DATA LOADER FOR HADOOP

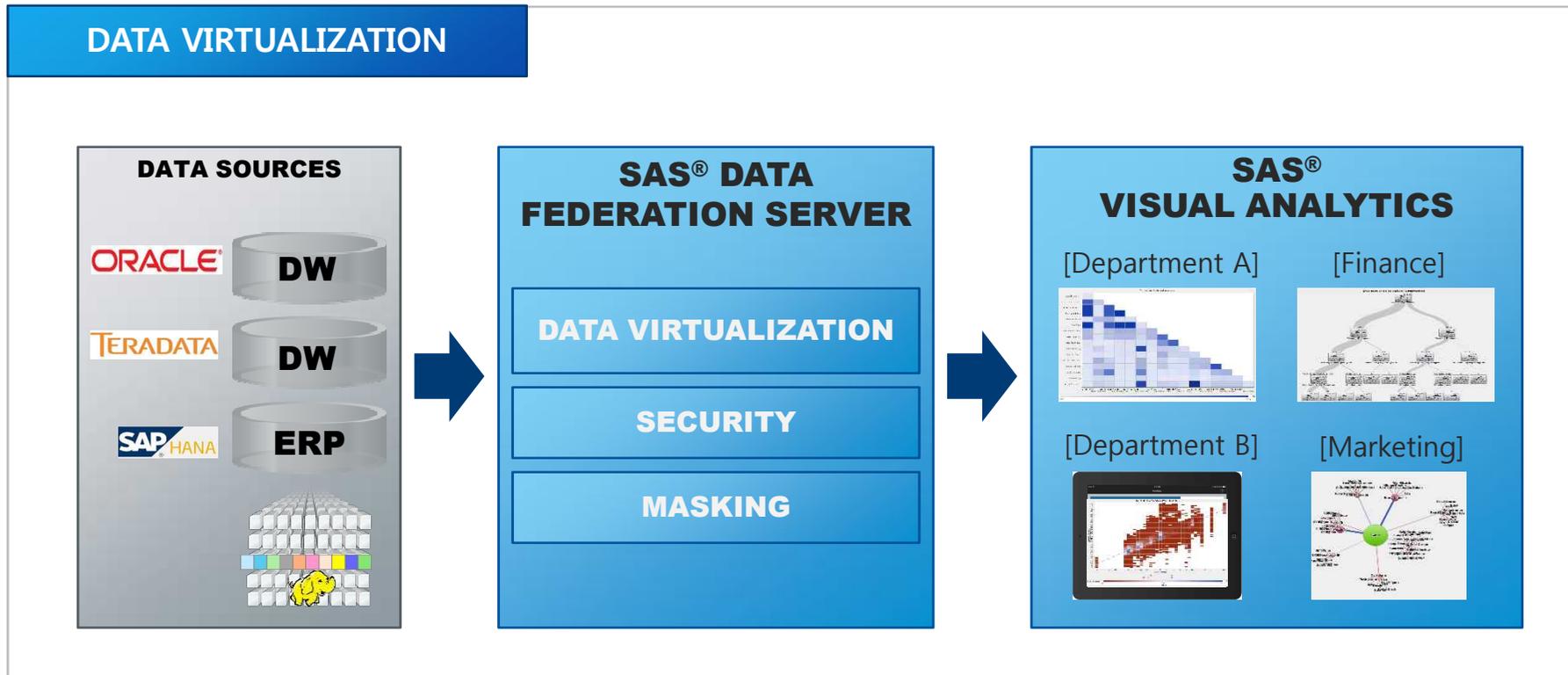
- **DIY HADOOP ;~)** FOR BUSINESS USERS



(2) DATA EXPLORATION

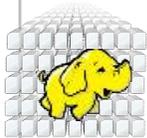


(2) DATA EXPLORATION



(3) MODEL DEVELOPMENT

MODELING ALGORITHMS ON HADOOP



SAS® ANALYTICS

Data Manipulation

- Aggregate
- Compute
- Update
- Append
- Set
- Schema
- DeleteRows
- DropTables
- PurgeTempTables

Evaluation, Deployment

- Assess Misclassification matrix Lift, ROC, Concordance
- Score
- Training / Validation

Data Exploration

- Boxplot
- Corr
- Crosstab
- Distinct
- Fetch
- Frequency
- Histogram
- KDE
- MDSummary
- Top K

Descriptive Modeling

- Association
- Path Analysis
- Clustering (k-means)
- Clustering (DBSCAN)

Predictive Modeling

- Decision Tree
- Forecast
- Gen Linear Model
- Linear Regression
- Logistic Regression
- Random Forests
- Neural Networks

Utilities

- Where
- GroupBy
- TableInfo, ColumnInfo, ServerInfo
- Partition, Balance
- Store, Replay, Free
- Table, Promote

Text Analytics

- Parsing
- SVD
- Topic generation
- Document projection

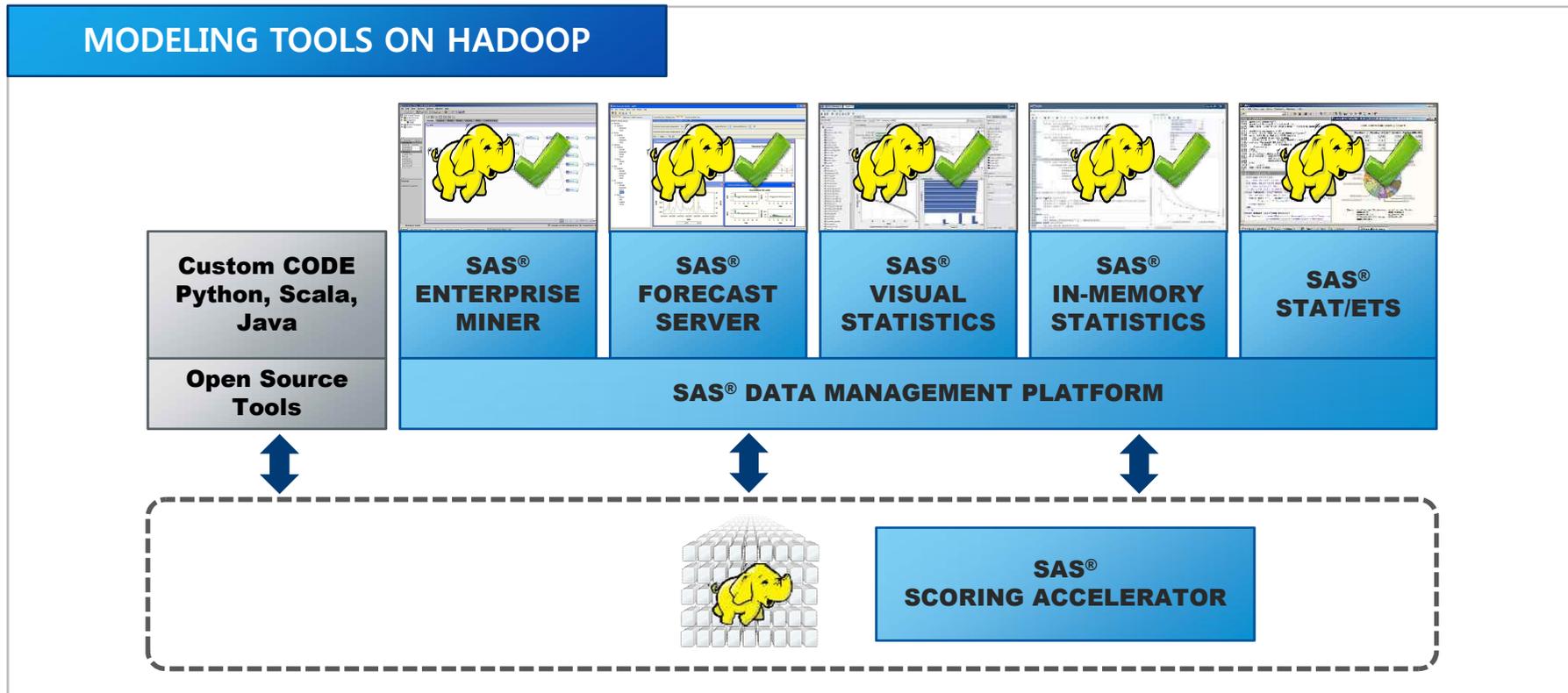
Recommendation Systems

- Association
- Clustering
- kNN
- SVD
- Ensemble

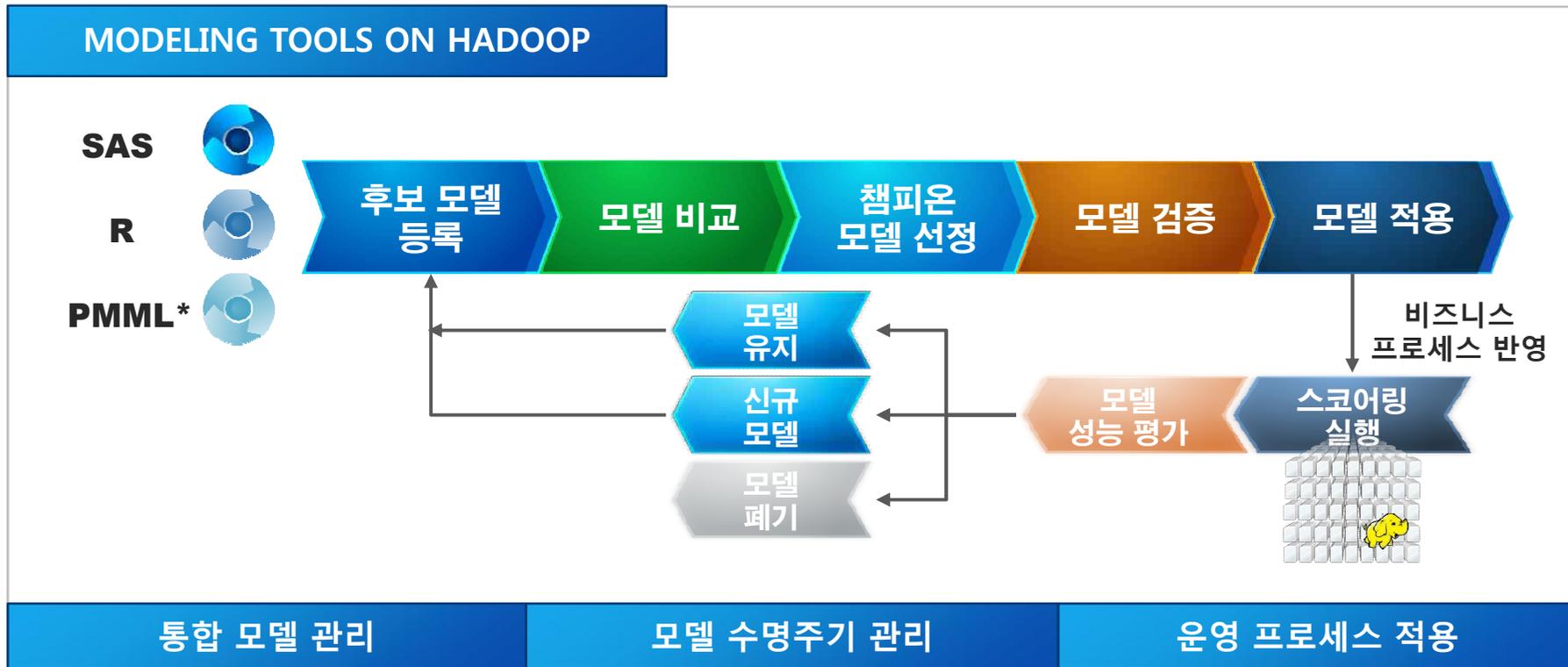
High Performance Modeling

- Statistics
- Data Mining / Text Mining
- Forecasting
- Optimization

(3) MODEL DEVELOPMENT



(4) MODEL MANAGEMENT



* PMML : Predictive Model Markup Language

(4) MODEL MANAGEMENT

[통합 의사결정 관리]

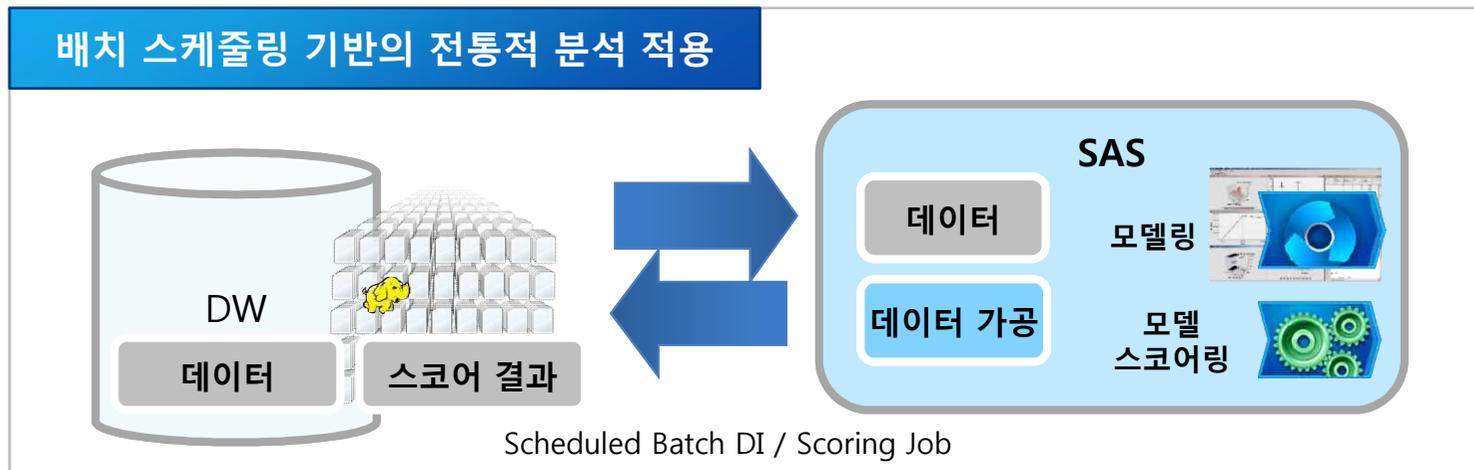
The screenshot displays the SAS Decision Management Workspaces interface. The main window is titled "Auto Auction Purchase Strategy" and contains a "Decision Flow Builder" diagram. The diagram shows a flow of decision nodes: "Request" (31,803 Records) leads to "Transmission Split" (31,803 Records, 0 Disposal), which branches into "Automatic Transmission" (29,804 Records) and "Unknown Transmission" (4 Records). "Automatic Transmission" leads to "Auto Qualification M..." (29,804 Records), which then leads to "Auto Qualification R..." (29,804 Records). "Auto Qualification R..." branches into "Auto Qualification S..." (29,804 Records, 0 Flipped) and "Other" (0 Records). "Auto Qualification S..." leads to "Good Car" (17,236 Records) and "Bad Car" (12,668 Records). "Good Car" leads to "Auto Price Model" (17,236 Records). The "Business Rule Manager" window shows a table with condition and action terms. The "Model Manager" window shows a "Lift Chart" for "Auto Purchase : Q1 2013 : Auto_Tree_Gini" comparing 2013Q1 and 2013Q2 performance.

Condition Term	Action Term
1 <=3	=Yearly_Income-5500
2 >3 AND <=12	=Yearly_Income+1000
3 >12	=Yearly_Income+2500
4 -0	-Adjust
5 >0	-Adjust

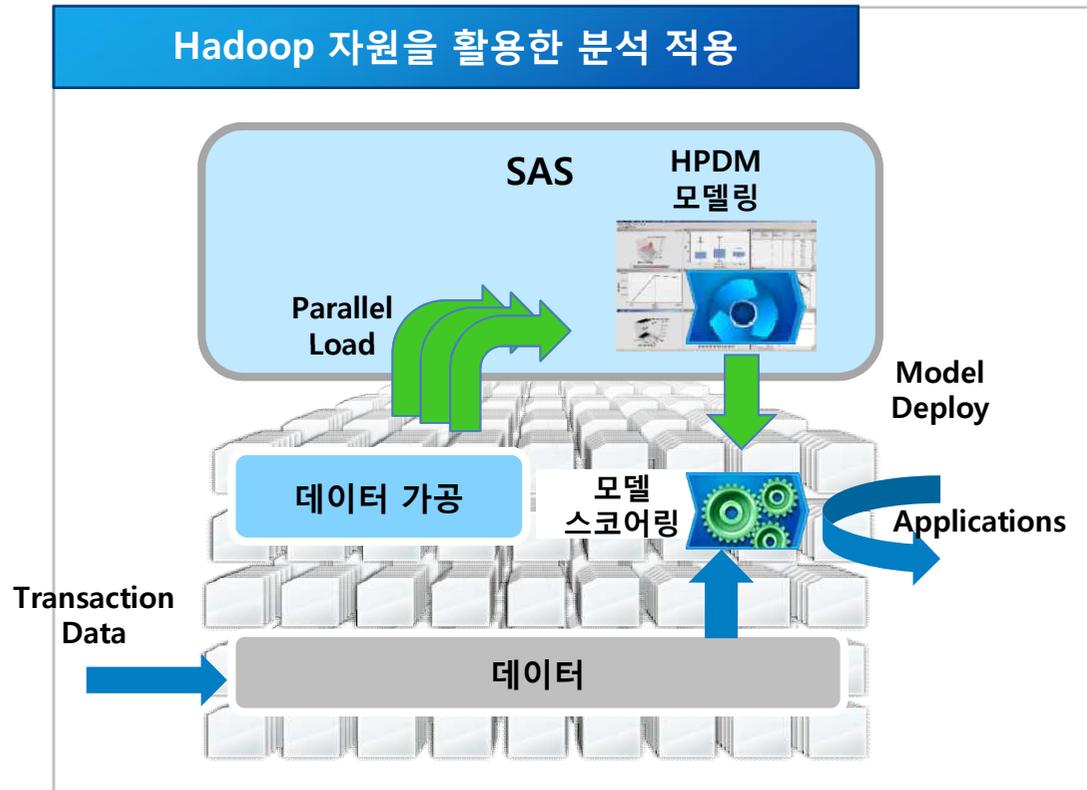
Variable	2013Q1	2013Q2
Stability	0.5	0.5
Lift	0.5	0.5
Gini	0.5	0.5
KS	0.5	0.5

(5) MODEL DEPLOYMENT

- ❖ 전통적인 분석 모델 적용 방법
- ❖ 사전 정의된 배치 스케줄링에 의해 SAS 서버 또는 응용프로그램에서 주기적으로 분석 결과 데이터 (스코어링) 산출
- ❖ 실시간 비즈니스, 대용량 데이터 환경 적용에 제약



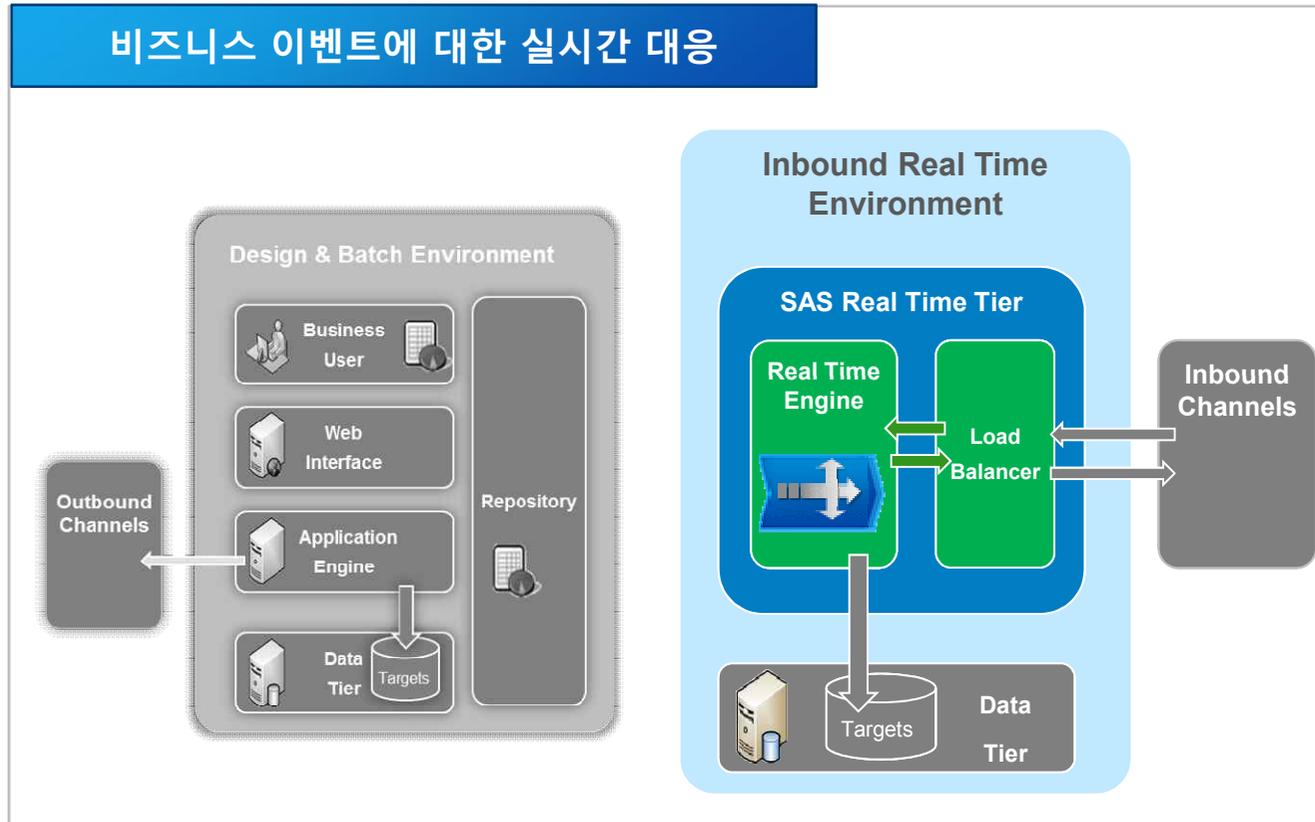
(5) MODEL DEPLOYMENT



- ❖ Hadoop 내부에서 분석을 위한 데이터 가공
- ❖ Hadoop 데이터를 활용한 빠른 모델링
- ❖ 생성된 모델을 자동화된 방법으로 Hadoop에 적용
- ❖ Hadoop 내부에서 이루어지는 모델 적용 결과 값 산출 (스코어링)

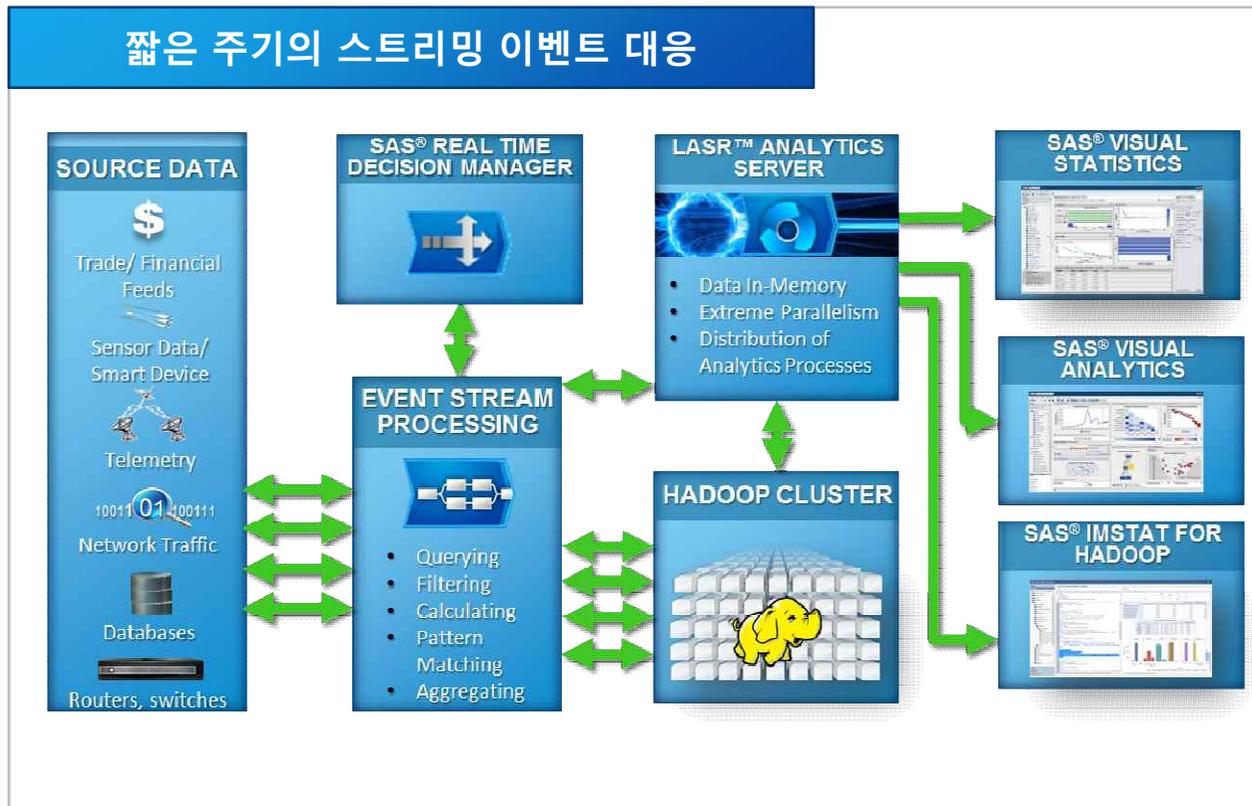
(5) MODEL DEPLOYMENT

비즈니스 이벤트에 대한 실시간 대응



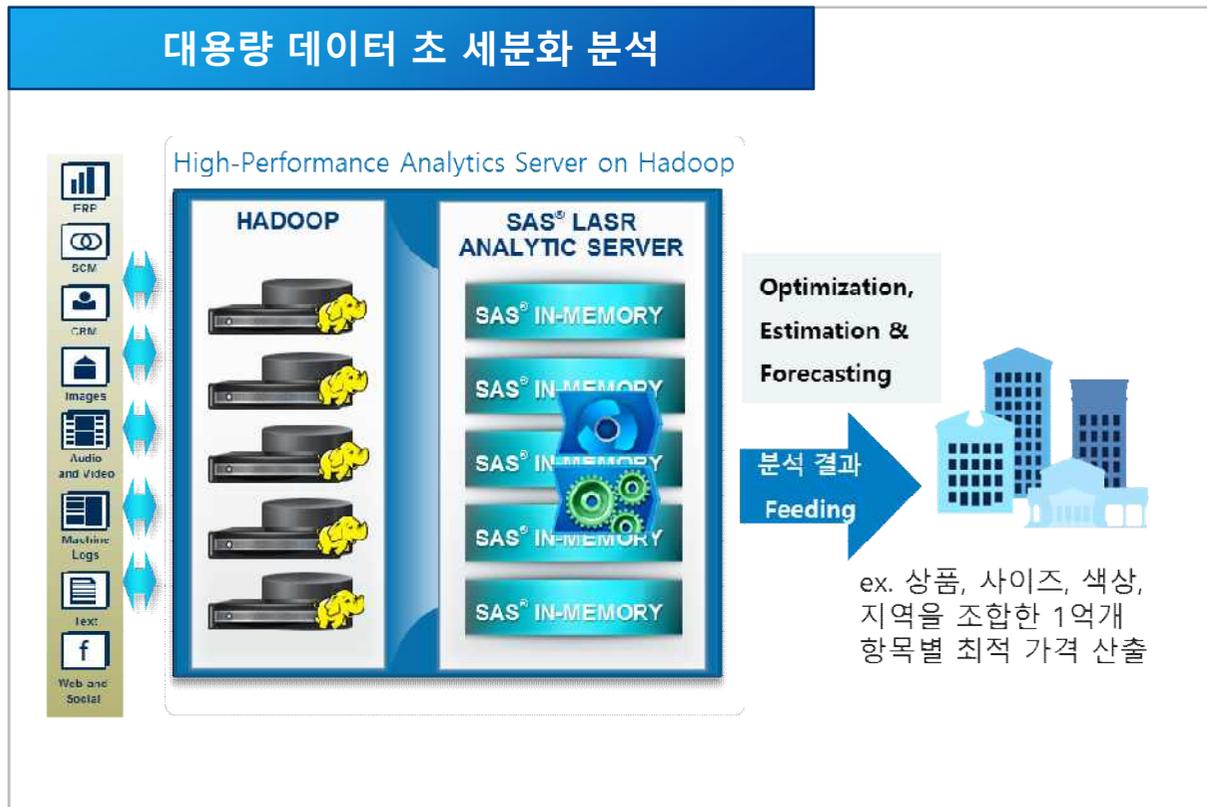
- ❖ 외부 발생 이벤트에 대한 실시간 대응
- ❖ ex) 인바운드 콜 대응 실시간 마케팅
- ❖ 분석 모델 기반의 실시간 의사결정 프로세스 지원

(5) MODEL DEPLOYMENT



- ❖ 매우 짧은 주기의 실시간 스트리밍 이벤트 처리
- ❖ 이벤트 처리 결과를 하둡 및 인메모리 분석과 연계
- ❖ 실시간 의사결정 프로세스 및 분석 연계

(5) MODEL DEPLOYMENT



- ❖ 대용량 데이터에 대해 초세분화된 분석을 적시에 빠르게 수행
- ❖ MPP 분산서버 환경에서 고급통계분석 병렬 수행
 - SAS High-Performance Statistics
 - SAS High-Performance Data Mining
 - SAS High-Performance Forecasting
 - SAS High-Performance Optimization
 - SAS High-Performance Econometrics
 - SAS High-Performance Text Mining

하둡 비전문가를 위한 셀프서비스 HADOOP 분석



Business Analyst
/ IT



SAS Data Loader for Hadoop



SAS Visual Analytics / Statistics



Copyright © 2015, SAS Institute Inc. All rights reserved.





THE
POWER
TO KNOW.