빅데이터 구축 전략:
IBM BigData Platform

IBM SWG

Introduction to IBM BigData Platform

1 폭발적인 Data 증가와 영향

2 현 DW의 한계와 전망

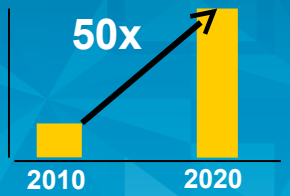
3 IBM이 제안하는 PureData System for Analytics

폭발적인 **Data** 증가와 영향

Mobile, Commerce, Social, Analytics, Big Data, Cloud 등 다양한 분야에서 발생하는 Data는 다양성, 대용량, 대량의 Transaction 등 Data System 에 대한 요건을 증가시키고 있습니다

Mobile *Commerce* *Social* *Analytics* *Big Data* *Cloud*

Increasing
Volume of data
requires growing capacity



35 ZB
by 2020

Increasing
Velocity of data
requires higher performance

Millions of transactions per second
Telco subscriber activity logging

Increasing
Variety of data
requires new techniques

Billions of devices & sensors
Smart Meters, RFIDs, GPS…

A smarter approach to meeting data challenges is required to:
Reduce complexity ▪ Accelerate time to value ▪ Improve IT economics

100개국 30개 이상의 산업에서 종사하는 임원, 매니져 3000명을 대상으로 조사한 내용입니다.

CIOs rank Analytics as the

#1 factor

Contributing to an organization's competitiveness.[1]

Organizations that embrace analytics are more than

2X

as likely to outperform their Peers.[2]

Financial outperformers are

64%

more likely to use analytics to evaluate talent supply and demand on an ongoing basis.[3]

Enterprises that apply advanced analytics have

33%

More revenue Growth and

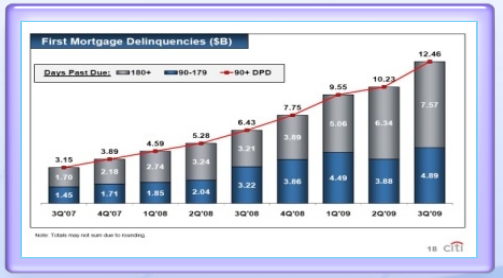
12X

more profit growth.[4]

[1] IBM CIO Study 2009
[2] IBM IBV/MIT Sloan Management Review Study 2011
[3] IBM CHRO Study 2010
[4] IBM CFO Study 2010

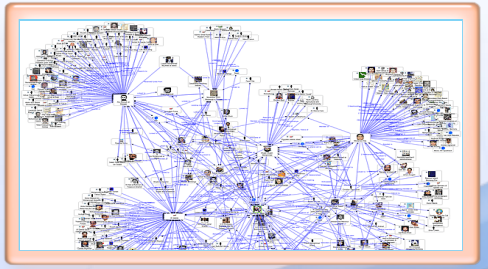
CIO는 예측분석을 통한 최적의 의사결정을 하기를 원합니다.

BI Reporting and Ad-Hoc Analysis



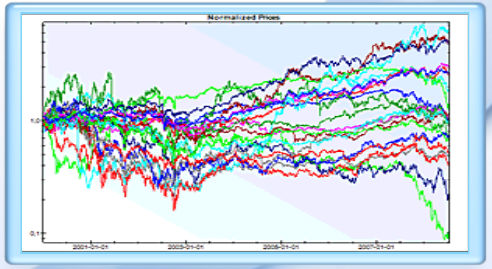
- What happened?
- When and where?
- How much?

Predictive Analytics



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

Optimization

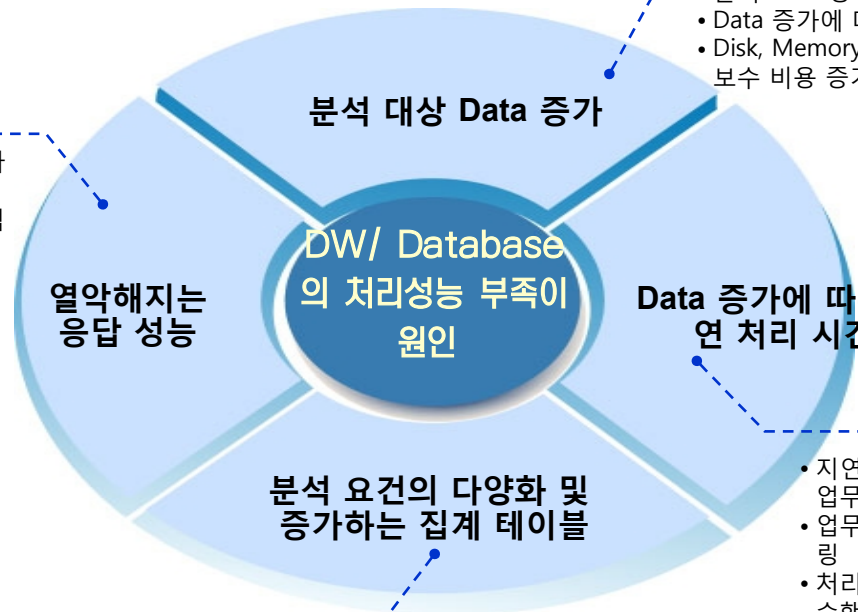


- What is the best choice?

현 DW의 한계와 전망

기존 DW 환경으로는 대용량, 다양성, 다량의 Transaction을 처리하는데에 한계가 있었습니다.

기존 DW 환경으로는 대용량, 다양성, 다량의 Transaction을 처리하는데에 한계가 있었습니다.



- 분석 Data 증가에 따른 관리 포인트 증가
- Data 증가에 따른 수행속도 이슈 발생
- Disk, Memory 등 시스템 자원의 소요 증가 및 유지 보수 비용 증가

- 분석 업무 데이터의 급속한 증가에 따른 응답속도의 저하
- 응답 성능의 저하로 인한 다양한 분석의 한계 (분석 요건의 다양화에 따른 비정형 쿼리 수행에 제한)
- 신규 업무에 대한 업무 분석의 한계

- 지연 처리 시간에 따른 분석 시간의 감소 및 분석 업무의 최소화 발생
- 업무 시간까지 영향을 미치는 Batch Job 스케줄링
- 처리 시간 향상을 위한 시스템 자원 소요 증가 및 수행속도 향상을 위한 튜닝필요

- 새로운 업무 분석 요건에 따른 계속적인 집계 테이블 생성 증가
- 집계 테이블 증가에 따른 ETL Job 및 프로세스 증가
- 분석 요건의 다양화로 비정형 쿼리 수행의 증가

Appliance는 목적에 맞게 최적화 한 전용장비 입니다.

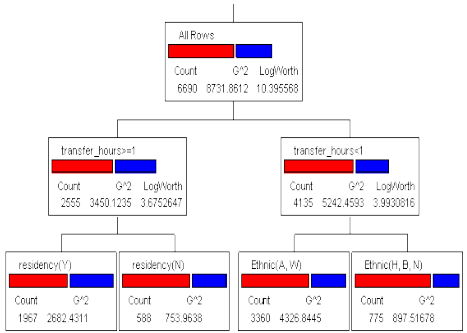
- 전용장비
- 목적에 맞는 최적화
- Complete solution
- 표준 인터페이스
- 간편한 설치
- 손쉬운 운영
- 관리 용이
- 저비용

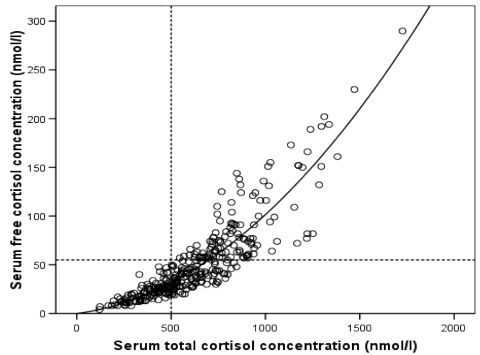
PureData System for Analytics 는 Database + Server + Storage를 하나로 통합한 새로운 개념의 데이터웨어하우스 어플라이언스 제품입니다.

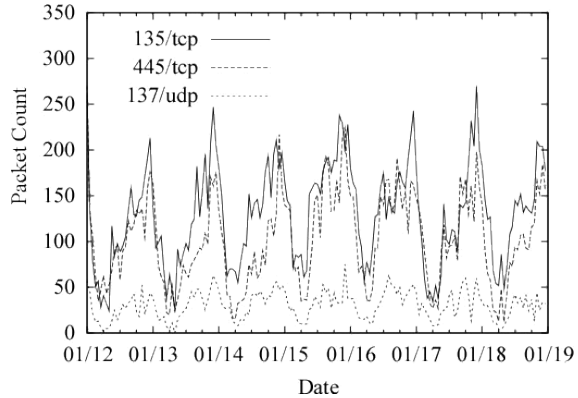
- PureData System for Analytics 제품은 Appliance 개념을 세계 최초로 DW 시스템에 적용하였으며, 가장 Appliance의 개념에 부합하는 DW 시스템을 제공합니다.
- Database + Server + Storage 를 하나로 통합하여 모든 구성을 최적화시킨 차세대 BI/DW Infrastructure 환경을 제공하는 DW 전용 어플라이언스 제품으로써, DW시스템으로 사용하는 데 있어 PureData System for Analytics 이외의 부가 장비를 전혀 필요로 하지 않습니다.

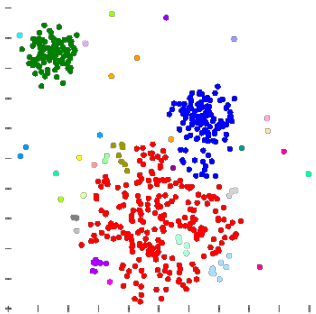
새로운 개념의 DataWarehouse Appliance

PureData System for Analytics는 기존 DW 요건과 함께 Data Mining과 통계 분석 요건을 추가하여 더욱더 빠른 통합 분석이 가능합니다.









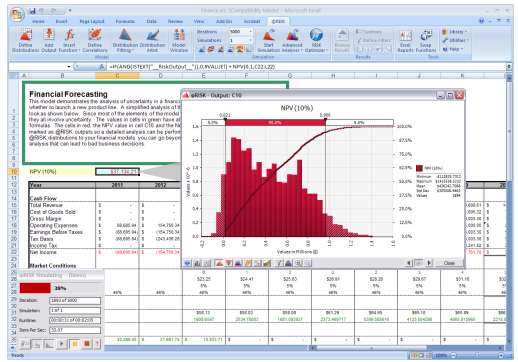
310

	301	302	303	304	305
TA	1st Item		2nd Item	3rd Item	4th Item
1	10		20	30	40
2	20		40		
3	10		30	50	
4	30		40	50	
5	20		40	70	80

Multi-Column Data

TA	Item
1	10
1	20
1	30
1	40
2	20
2	40
3	10
3	30
3	50
4	30
4	40
4	50
5	20
5	40
5	70
5	80

Transaction Data



IBM이 제안하는
PureData System for Analytics

PureSystems에는 PureFlex, PureApplication, PureData 로 구성됩니다

Infrastructure

PureFlex

Delivering Infrastructure Services

Application Platform

PureApplication

Delivering Platform Services

Data Platform

PureData

Delivering Data Services

8 Disk Enclosures
- 96 1TB SAS Drives (4 hot spares)
- RAID 1 Mirroring

2 Hosts (Active-Passive)
- 2 Quad-Core Intel 2.6 GHz CPUs
- 7x146 GB SAS Drives
- Red Hat Linux 5 64-bit

14 PureData for Analytics S-Blades™
- 2 Intel Quad-Core 2+ GHz CPUs
- 4 Dual-Engine 125 MHz FPGAs
- 24 GB DDR2 RAM
- Linux 64-bit Kernel

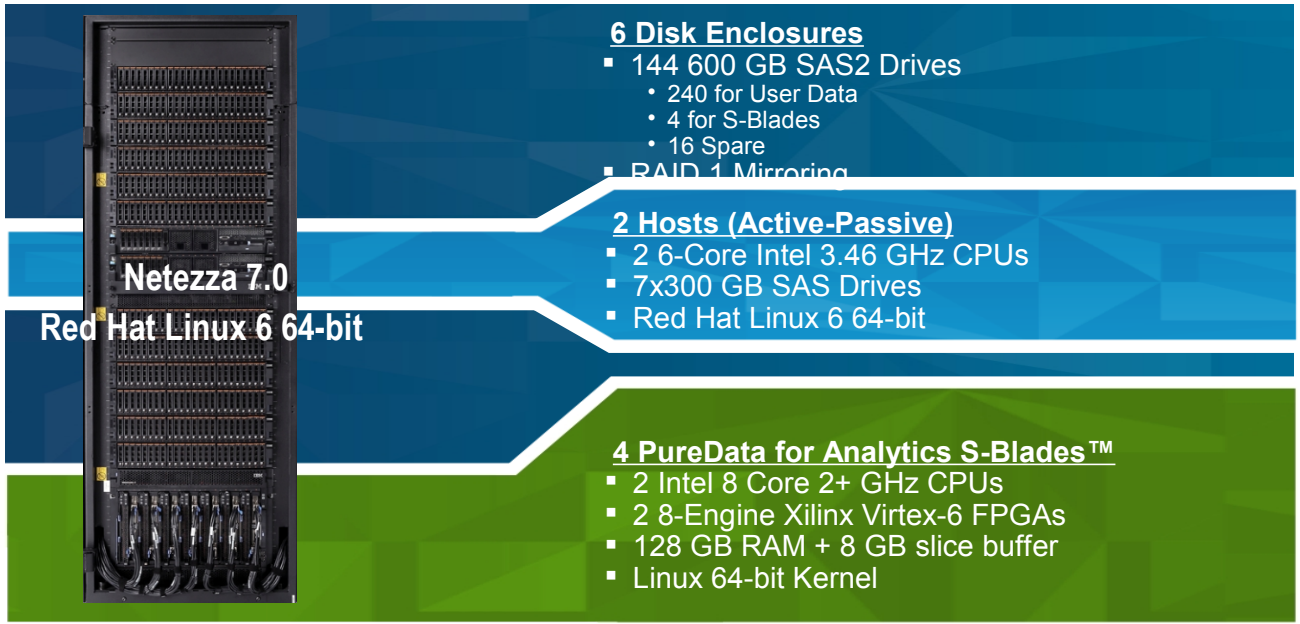
Scales from
¼ Rack to 10 Racks

32 TB to 1.2 PB of
User Data

- User Data Capacity: 128 TB**
- Data Scan Speed: 145 TB/hr**
- Load Speed (per system): 5+ TB/hr

- Power Requirements: 7.6 kW
- Cooling Requirements: 7.8 kW

**: 4X compression assumed



Scales from
½ Rack to 4 Racks

- User Data Capacity: 192 TB*
- Data Scan Speed: 478 TB/hr*
- Load Speed (per system): 5+ TB/hr

- Power Requirements: 7.5 kW
- Cooling Requirements: 27,000 BTU/hr

* Assuming 4X compression

IBM PureData System for Analytics

Optimized exclusively for analytic data workloads

속도

- 10-100x faster than traditional custom systems*
- Patented MPP hardware acceleration (Massively Parallel Processing)

간편성

- Data load ready in hours
- No database indexes
- No tuning
- No storage administration

확장성

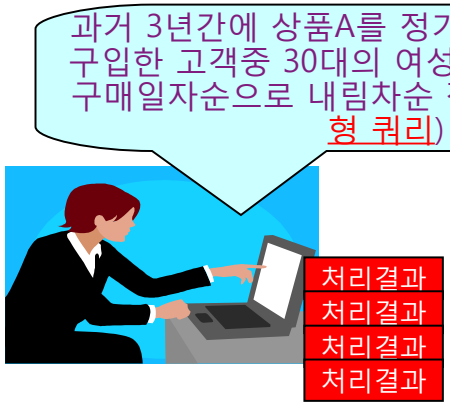
- Peta-scale data capacity

Smart

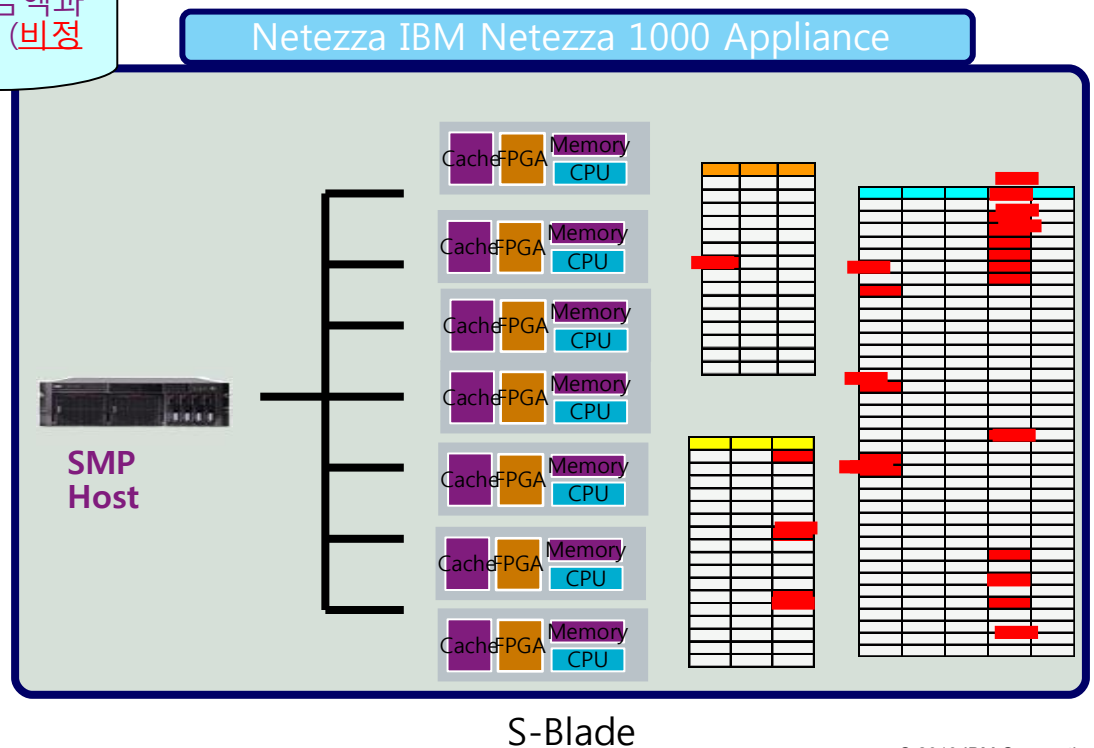
- Designed to runs complex analytics in minutes, not hours
- Richest set of in-database analytics

* Based on IBM customers' reported results. "Traditional custom systems" refers to systems that are not professionally pre-built, pre-tested and optimized. Individual results may vary.

PureData System for Analytic의 데이터 처리방식 입니다.



Disk를 스캔과 동시에 필요한 데이터를 각 SPU에서 판단하고, 필요한 데이터만을 전송. 네트워크나 호스트에서 취급하는 데이터 양은 기존 시스템의 1~2%.



Source: If applicable, describe source origin

Index, Partition Table 등의 불필요로 생성 및 관리가 간편해 졌습니다.

Legacy DDL
Model Consi
Create Table

```
CREATE SET
    NO BEFC
    NO AFTE
    (
     Frt_Bill_I
     Load_Id
     Evnt_Typ
     Proc_Dt
     Adt_Dt
     Rcv_Dt
     Inv_Dt D
     Clt_Inv_
     Ship_Dt
     Rjct_Rsn
     ......
```

Indices
Tuning Modifications

```
PRIMARY INDEX X
INDEX ( Proc_Dt )
INDEX ( Ship_Dt )
INDEX ( Pro_Nb
INDEX ( R
INDEX
IND
IN
I
IN
IND
INDEX
INDEX ( F
INDEX ( BOL_N
INDEX ( Evnt_Typ_
INDEX ( Frt_Bill_Id
INDEX ( Cust_Id ,Pl
INDEX ( Frt_Bill_Id
INDEX ( Wgt )
INDEX ( Plnt_NCS
INDEX ( Acct_Nbr
```

Aggregation Tables
More Tuning Considerations

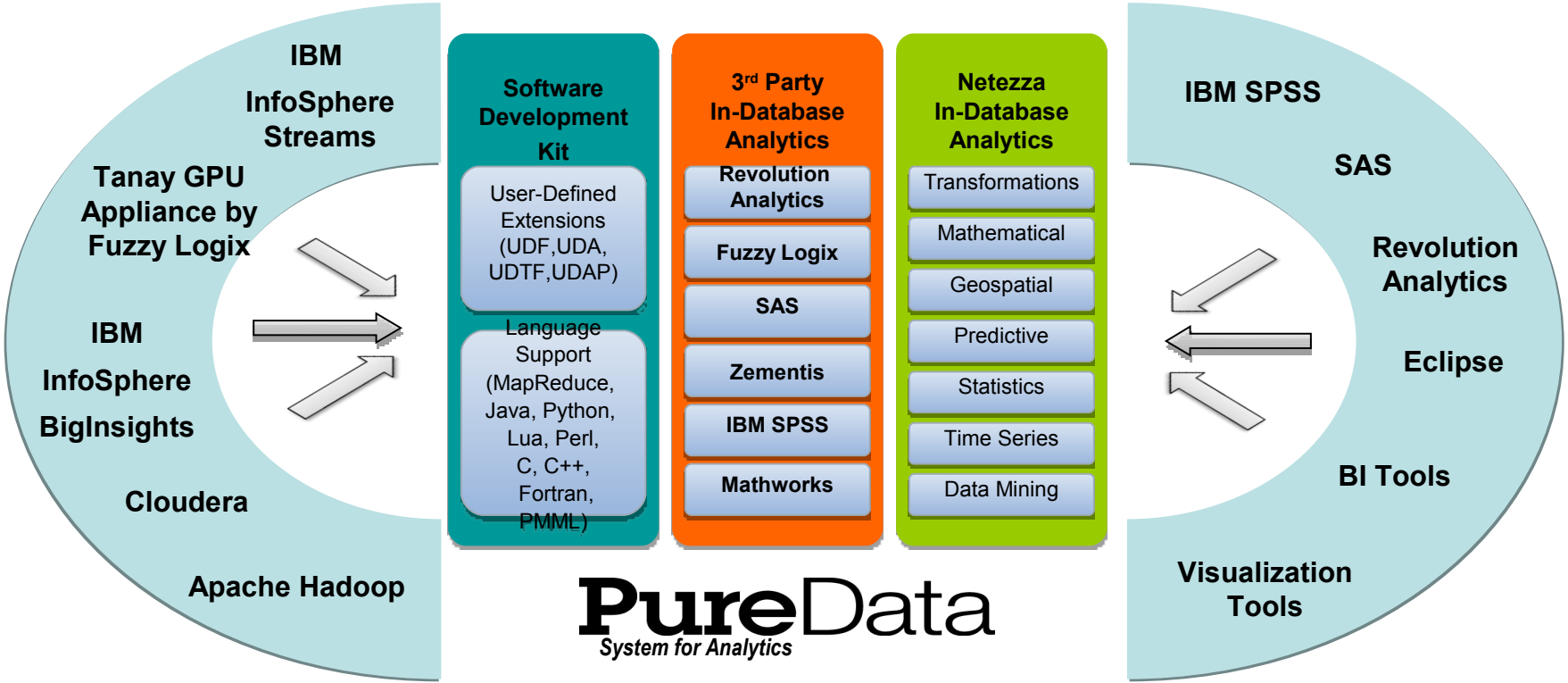
```
CREATE SET TABLE AGG_TISTBL.FRT_BI
    BEFORE JOURNAL,
    JOURNAL
    (
     Ct_Frt_Bi      INTEGER NOT NU
     Load_Id INT
     Evnt_Typ_Cd      (3) CHARACTER
     Proc_Wk DATE     MAT 'YYYY-MM
     Adt_Wk DATE      AT 'YYYY-MM
     Wk DATE          AT 'YYYY-MM
     k DATE           AT 'YYYY-MM-
Rjct_R      CHAR     HARACTER SET

CREATE SE       GG_TISTBL.FRT_BILL_EVNT_MONTH ,NO FALLBACK ,
    NO BE       URNAL,
                OURNAL

     Ct_Frt_Bill_Id_Mth INTEGER NOT NULL,
     Load_Id INTEGER,
     Evnt_Typ_Cd CHAR(3) CHARACTER SET LATIN NOT CASESPECIFIC NOT NULL,
     Proc_Mth DATE FORMAT 'YYYY-MM-DD' TITLE 'Freight Bill Process Month',
     Adt_Mth DATE FORMAT 'YYYY-MM-DD' TITLE 'Freight Bill Audit Month',
     Ship_Mth DATE FORMAT 'YYYY-MM-DD' TITLE 'Freight Bill Ship Month',
     Rjct_Rsn_Cd CHAR(2) CHARACTER SET LATIN NOT CASESPECIFIC,
```

Netezza DDL

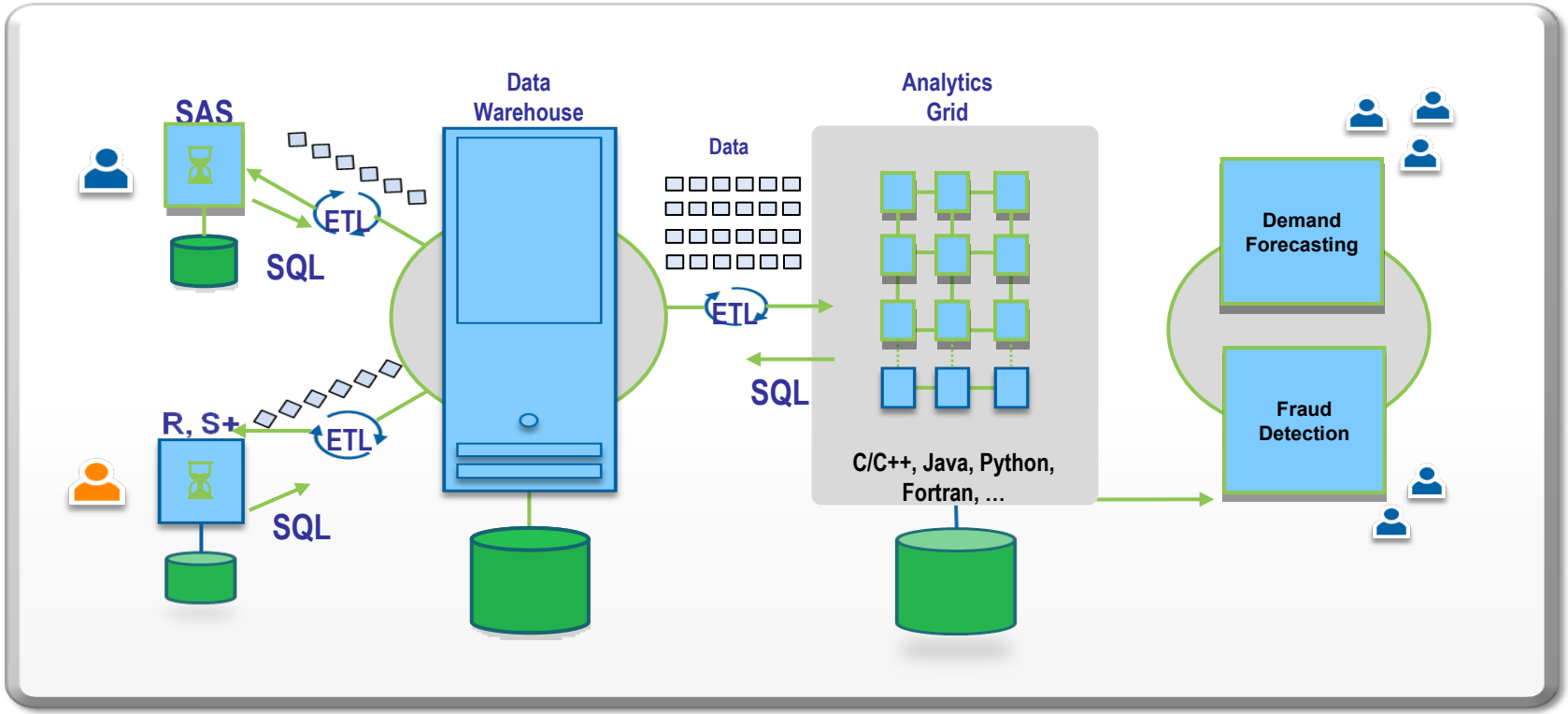
Create Table - Logical Model

```
CREATE TABLE frt_bill_evnt
(               Frt_Bill_Id INTEGER NOT NULL,
                Load_Id INTEGER,
                Evnt_Typ_Cd CHAR(3) NOT NULL,
                Proc_Dt DATE,
                Adt_Dt DATE,
                Rcv_Dt DATE,
                Inv_Dt DATE,
                Clt_Inv_Dt DATE,
                Ship_Dt DATE,
                Rjct_Rsn_Cd CHAR(2),
                .....
                Dlvy_Tm INTEGER,
                Actl_Cmdt_Ds CHAR(30),
                Carr_Rfrc_Nbr VARCHAR(25))
);
```

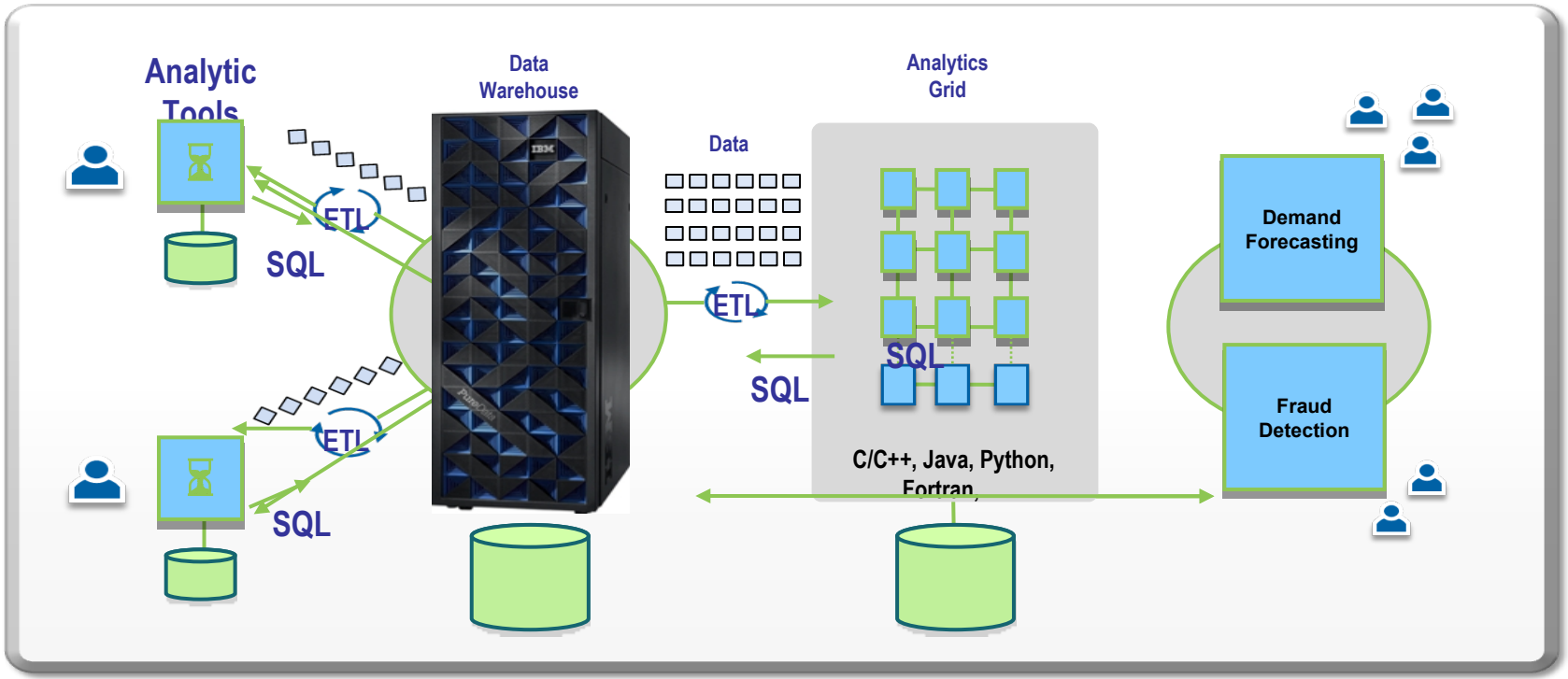
IBM Netezza Analytics를 통해 다양한 Big Data솔루션의 손쉬운 연동 및 통계 분석 솔루션들에 대한 PDA 내 고급 분석 수행 기능을 통해 효과적인 Big Data 통합을 제공합니다.



PureData System for Analytic의 In-Database Analytics는 다양한 통계 분석 기법을 제공합니다



PureData System for Analytic의 In-Database Analytics는 다양한 통계 분석 기법을 제공합니다



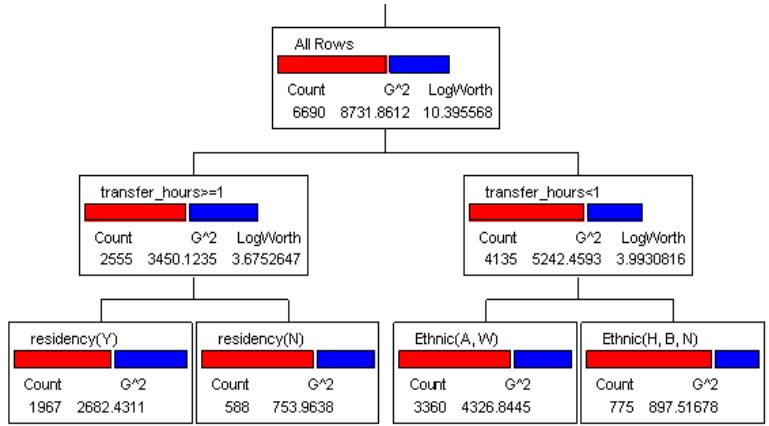
23

© 2013 IBM Corporation

PureData System for Analytic의 In-Database Analytics는 다양한 통계 분석 기법을 제공합니다

Predict discrete values

- Response to treatment
- Events
- Customer states
- Segment membership
- Credit Score



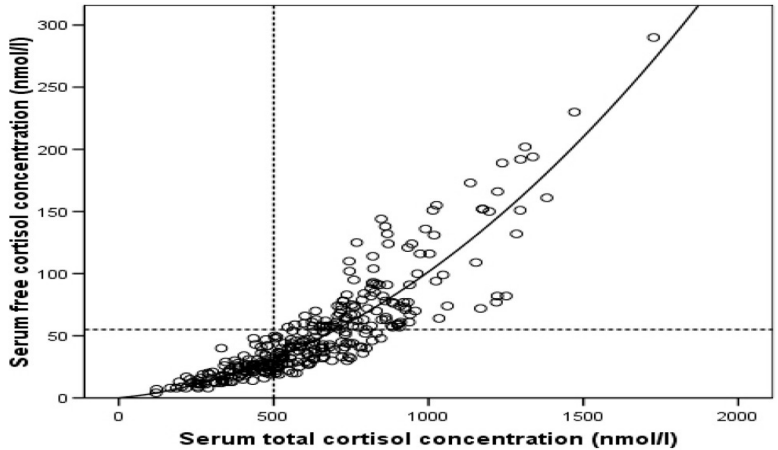
PureData for Analytics In-Database Functions

- ✓ Logistic Regression
- ✓ Probit Regression
- ✓ Decision Trees
- ✓ Naïve Bayes Classifier
- ✓ Nearest Neighbor

PureData System for Analytic의 In-Database Analytics는 다양한 통계 분석 기법을 제공합니다.

Predict point estimates

- Value of purchases
- Usage amount
- Population statistics



PureData System for Analytics In-Database Functions

- ✓ Linear Regression
- ✓ Mixed Linear Models
- ✓ Regression Trees
- ✓ Nearest Neighbor

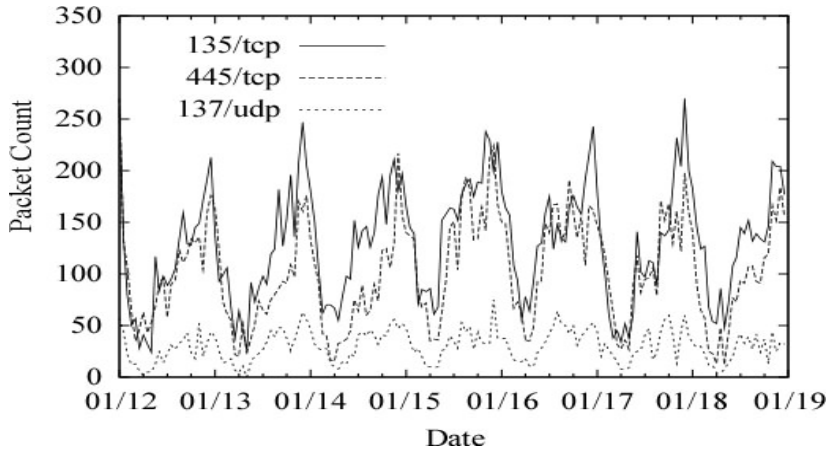
25

© 2013 IBM Corporation

PureData System for Analytic의 In-Database Analytics는 다양한 통계 분석 기법을 제공합니다.

Predict temporal estimates:

- Monthly store sales
- Quarterly
- Daily impressions/clicks



PureData System for Analytics In-Database Functions

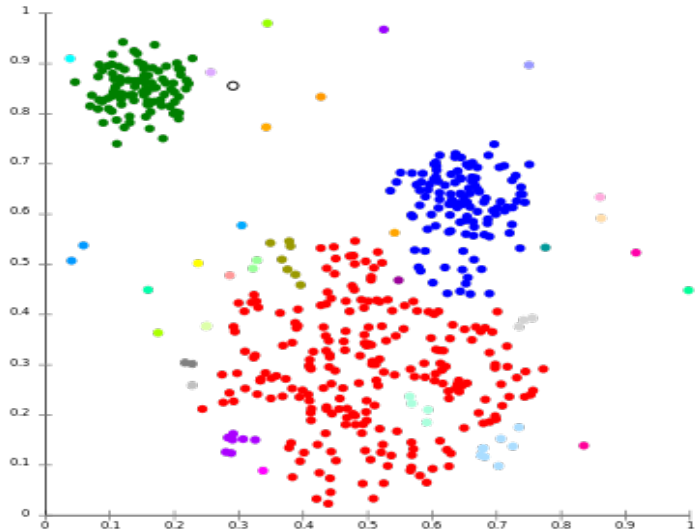
- ✓ Exponential Smoothing
- ✓ Holt-Winters
- ✓ ARIMA / ARMA
- ✓ Seasonal Trend Decomposition

© 2013 IBM Corporation

PureData System for Analytic의 In-Database Analytics는 다양한 통계 분석 기법을 제공합니다.

Group similar records:

- Customer segmentation
- Data exploration
- Text and image analysis



PureData System for Analytics In-Database Functions

✓ k-Means

✓ Divisive Clustering

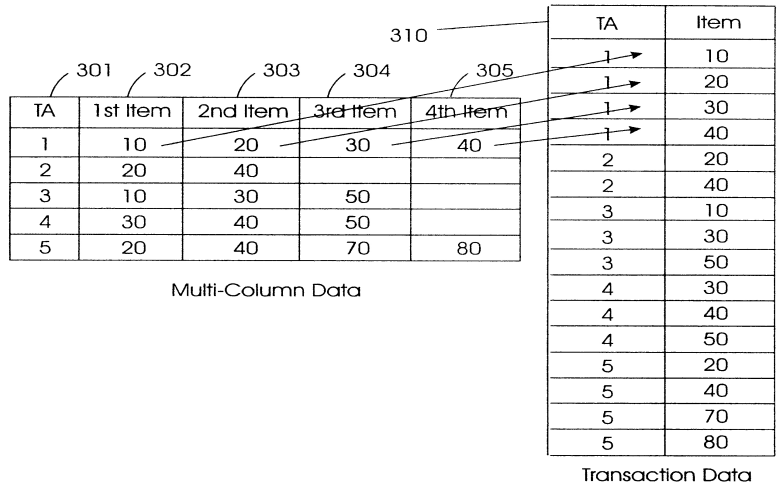
27

© 2013 IBM Corporation

PureData System for Analytic의 In-Database Analytics는 다양한 통계 분석 기법을 제공합니다.

Identify rules among variables:

- Data exploration and discovery
- Market Basket Analysis
- Cross-purchase propensity



PureData System for Analytics In-Database Functions

- ✓ FP-Growth
- ✓ A Priori

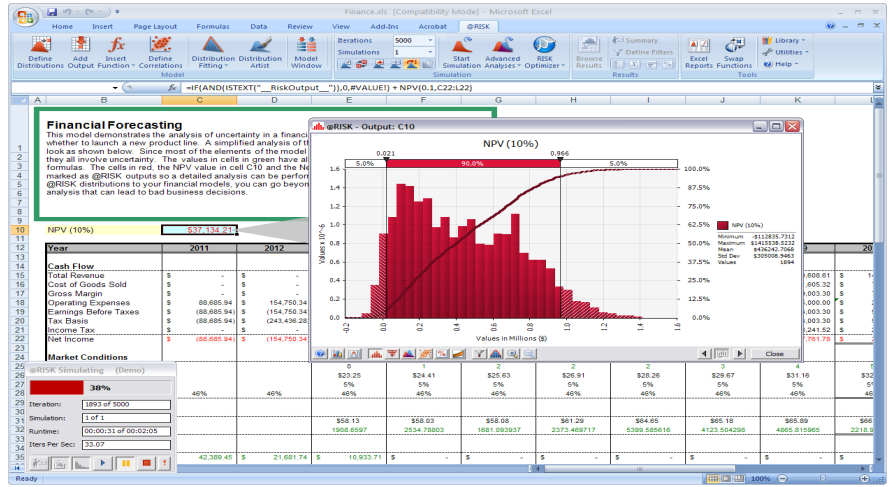
28

© 2013 IBM Corporation

PureData System for Analytic의 In-Database Analytics는 다양한 통계 분석 기법을 제공합니다.

Model complex systems:

- Climate and weather
- Portfolio risk
- Financial markets
- Consumer behavior



PureData System for Analytics Capabilities

- ✓ Matrix Engine
- ✓ Stochastic Functions
- ✓ Language Support
- ✓ Tools and Utilities

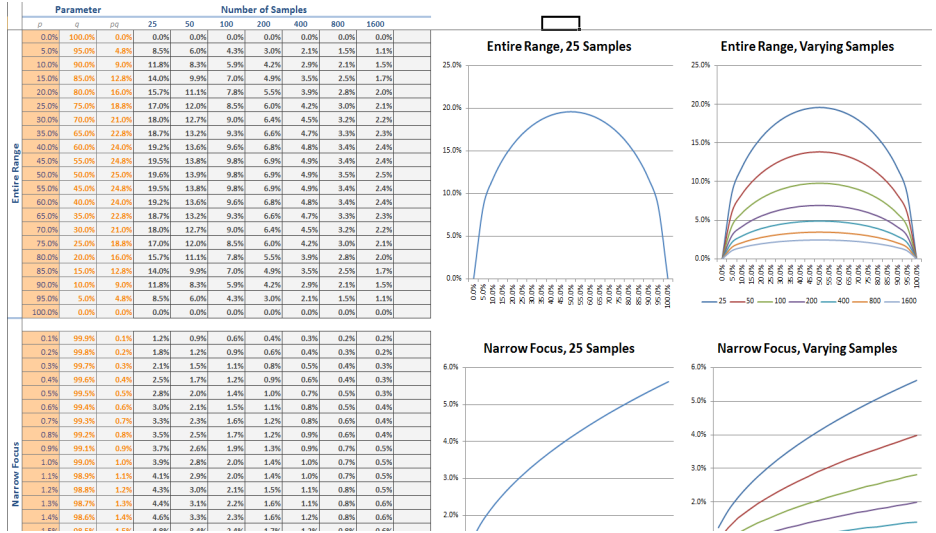
29

© 2013 IBM Corporation

PureData System for Analytic의 In-Database Analytics는 다양한 통계 분석 기법을 제공합니다.

Understand your data:

- Preliminary analysis
- Develop hypotheses
- Discovery



	Parameter			Number of Samples						
	p	q	pq	25	50	100	200	400	800	1600
Entire Range	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	5.0%	95.0%	4.8%	8.5%	6.0%	4.3%	3.0%	2.1%	1.5%	1.1%
	10.0%	90.0%	9.0%	11.8%	8.3%	5.9%	4.2%	2.9%	2.1%	1.5%
	15.0%	85.0%	12.8%	14.0%	9.9%	7.0%	4.9%	3.5%	2.5%	1.7%
	20.0%	80.0%	16.0%	15.7%	11.1%	7.8%	5.5%	3.9%	2.8%	2.0%
	25.0%	75.0%	18.8%	17.0%	12.0%	8.5%	6.0%	4.2%	3.0%	2.1%
	30.0%	70.0%	21.0%	18.0%	12.7%	9.0%	6.4%	4.5%	3.2%	2.2%
	35.0%	65.0%	22.8%	18.7%	13.2%	9.3%	6.6%	4.7%	3.3%	2.3%
	40.0%	60.0%	24.0%	19.2%	13.6%	9.6%	6.8%	4.8%	3.4%	2.4%
	45.0%	55.0%	24.8%	19.5%	13.8%	9.8%	6.9%	4.9%	3.4%	2.4%
	50.0%	50.0%	25.0%	19.6%	13.9%	9.8%	6.9%	4.9%	3.5%	2.5%
	55.0%	45.0%	24.8%	19.5%	13.8%	9.8%	6.9%	4.9%	3.4%	2.4%
	60.0%	40.0%	24.0%	19.2%	13.6%	9.6%	6.8%	4.8%	3.4%	2.4%
	65.0%	35.0%	22.8%	18.7%	13.2%	9.3%	6.6%	4.7%	3.3%	2.3%
	70.0%	30.0%	21.0%	18.0%	12.7%	9.0%	6.4%	4.5%	3.2%	2.2%
	75.0%	25.0%	18.8%	17.0%	12.0%	8.5%	6.0%	4.2%	3.0%	2.1%
	80.0%	20.0%	16.0%	15.7%	11.1%	7.8%	5.5%	3.9%	2.8%	2.0%
	85.0%	15.0%	12.8%	14.0%	9.9%	7.0%	4.9%	3.5%	2.5%	1.7%
	90.0%	10.0%	9.0%	11.8%	8.3%	5.9%	4.2%	2.9%	2.1%	1.5%
	95.0%	5.0%	4.8%	8.5%	6.0%	4.3%	3.0%	2.1%	1.5%	1.1%
	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Narrow Focus	0.1%	99.9%	0.1%	1.2%	0.9%	0.6%	0.4%	0.3%	0.2%	0.2%
	0.2%	99.8%	0.2%	1.8%	1.2%	0.9%	0.6%	0.4%	0.3%	0.2%
	0.3%	99.7%	0.3%	2.1%	1.5%	1.1%	0.8%	0.5%	0.4%	0.3%
	0.4%	99.6%	0.4%	2.5%	1.7%	1.2%	0.9%	0.6%	0.4%	0.3%
	0.5%	99.5%	0.5%	2.8%	2.0%	1.4%	1.0%	0.7%	0.5%	0.3%
	0.6%	99.4%	0.6%	3.0%	2.1%	1.5%	1.1%	0.8%	0.5%	0.4%
	0.7%	99.3%	0.7%	3.3%	2.3%	1.6%	1.2%	0.8%	0.6%	0.4%
	0.8%	99.2%	0.8%	3.5%	2.5%	1.7%	1.2%	0.9%	0.6%	0.4%
	0.9%	99.1%	0.9%	3.7%	2.6%	1.9%	1.3%	0.9%	0.7%	0.5%
	1.0%	99.0%	1.0%	3.9%	2.8%	2.0%	1.4%	1.0%	0.7%	0.5%
	1.1%	98.9%	1.1%	4.1%	2.9%	2.0%	1.4%	1.0%	0.7%	0.5%
	1.2%	98.8%	1.2%	4.3%	3.0%	2.1%	1.5%	1.1%	0.8%	0.5%
	1.3%	98.7%	1.3%	4.4%	3.1%	2.2%	1.6%	1.1%	0.8%	0.6%
	1.4%	98.6%	1.4%	4.6%	3.3%	2.3%	1.6%	1.2%	0.8%	0.6%

PureData System for Analytics Capabilities

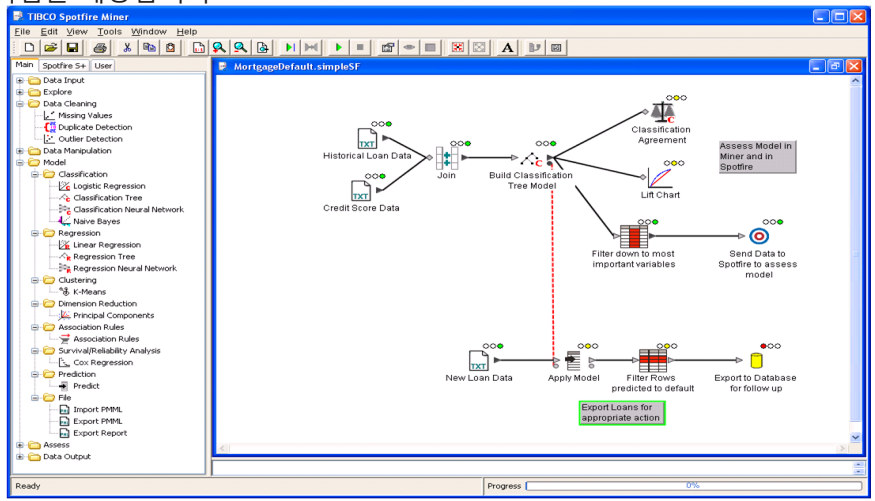
- ✓ Bayesian Network
- ✓ Multivariate Statistics
- ✓ Univariate Statistics
- ✓ Statistical Tests

(*) dbLytix

PureData System for Analytic의 In-Database Analytics는 다양한 통계 분석 기법을 제공합니다.

Prepare data for analysis:

- Select cases for analysis
- Transform variables
- Treat data quality issues



PureData System for Analytics Capabilities

- ✓ Random Sampling
- ✓ Standardization
- ✓ Normalization
- ✓ Missing Value Treatment
- ✓ Binning
- ✓ Utilities