전자신문 Lead the future

한국CIO포럼 CIO FORUM

CIO Summit 2015

CIO Summit 2015

디지털 기회선점을 위해 관점을 바꿔라!

TERADATA

차세대정보계 시스템의 비전

한국테라데이타 장동인 부사장

장동인

경력

- 현, 한국테라데이타 부사장
- 현, 빅데이터 전문가 협의회 의장
- 현, 경기도 빅데이터 자문위원
- 미래창조부 빅데이터 자문위원
- 미래읽기 컨설팅 대표
- Ernst & Young 컨설팅 본부장
- Deloitte consulting 전무(CRM부문 파트너)
- SAS Korea 부사장
- Siebel Korea 초대 지사장
- Oracle Korea 컨설팅 본부 이사
- Oracle HQ, Senior Principal Consultant
- Germany Amadeus, System Support Engineer
- American Airline Information Service, Consultant
- EDS, System Engineer
- VISA International, Programmer

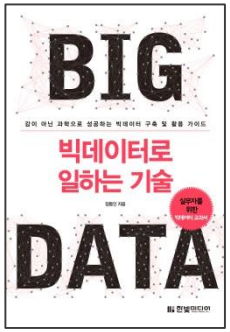
학력

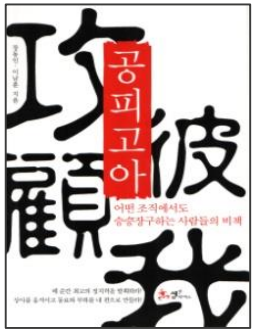
- 용산고등학교졸업
- 서울 공대 원자핵 공학과 졸업
- University of Southern California, 컴퓨터 공학 석사 졸업

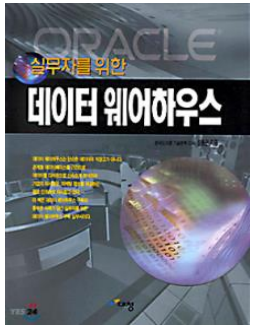
전문분야

- 빅데이터 전략 및 활용
- 클라우드 컴퓨팅
- 고객 및 마케팅 전략
- CRM 전략
- IT Architecture 및 전사 IT 전략
- 전사적 Data Warehouse 설계 자문
- Data Quality 자문
- IT Governance 자문

저서

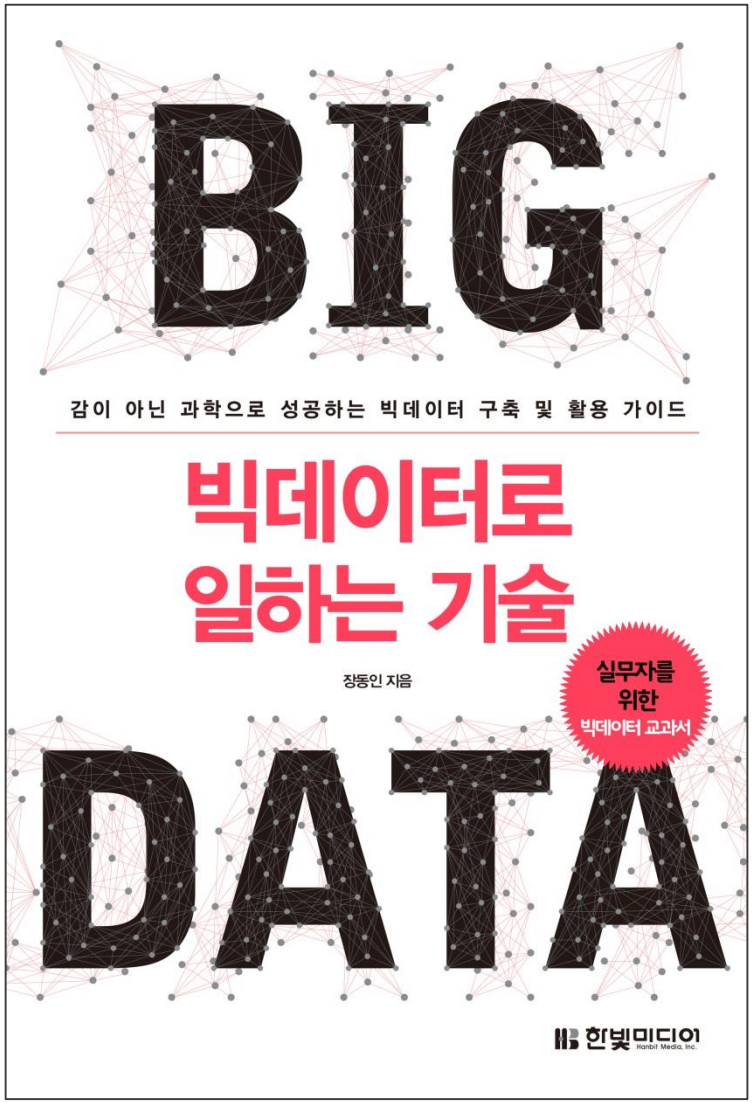






TERADATA

빅데이터로 일하는 기술



- 2014년12월15일 출판
- 한빛 미디어 (323p)
- 책을 쓴 동기;
  - ✓ 거의 모든 빅데이터 TF가 IT팀 위주
  - ✓ "우리 회사는 어떤 주제로 빅데이터를 해야 하는가?"
  - ✓ 현업은 분석을 안함
  - ✓ 기업 의사결정 문화는 "숫자" 채우기
  - ✓ 과거 CRM 무용론 (현업방관 IT위주)
- 책을 쓰는 동안 "소명의식"
  - ✓ 강한 위기의식
  - ✓ "빅데이터"라는 대중성
- 대상
  - ✓ 빅데이터를 도입하려고 하는 기업/공공기관의 빅데이터 TF팀, 임원, 현업
- 목적
  - ✓ 이제는 제대로 해보자
  - ✓ 빅데이터라는 냉철한 현실을 알리자

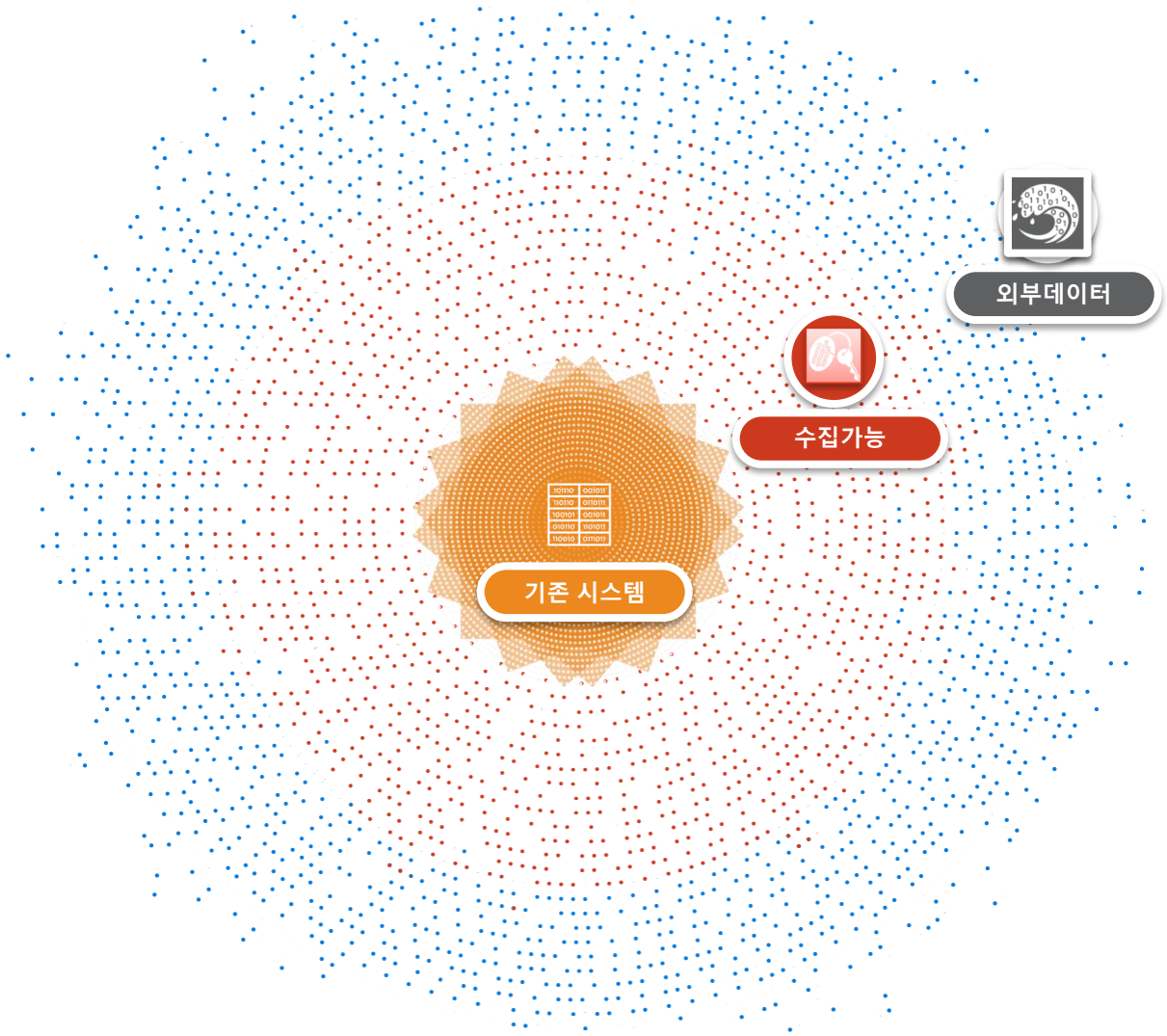
기존 정보계 시스템의 이슈

정보계는 차세대 시스템 구축에서 제외되었다...

- 현재 EDW의 이슈
  - ✓ Old, out-dated HW, SW (단종된 제품도 존재)
  - ✓ 현상태를 upgrade 할 것이냐 아니면 새로운 기술을 도입할 것인가?
  - ✓ 새롭게 등장하는 빅데이터들의 요구사항을 어떻게 수용할 것인가? 아니면 별도로 가야 하는가?
  - ✓ 현재 빅데이터 기술진보는 어디까지 왔는가? 검토단계
  - ✓ 기존의 시스템을 안전하게 migration할 수 있으며
  - ✓ 기존 시스템과 무리없이 integration 할 수있는가?
- 이러한 상황에서 빅데이터 기술은 어디까지 왔으며, 앞으로 어떤 방향으로 갈 것인가?

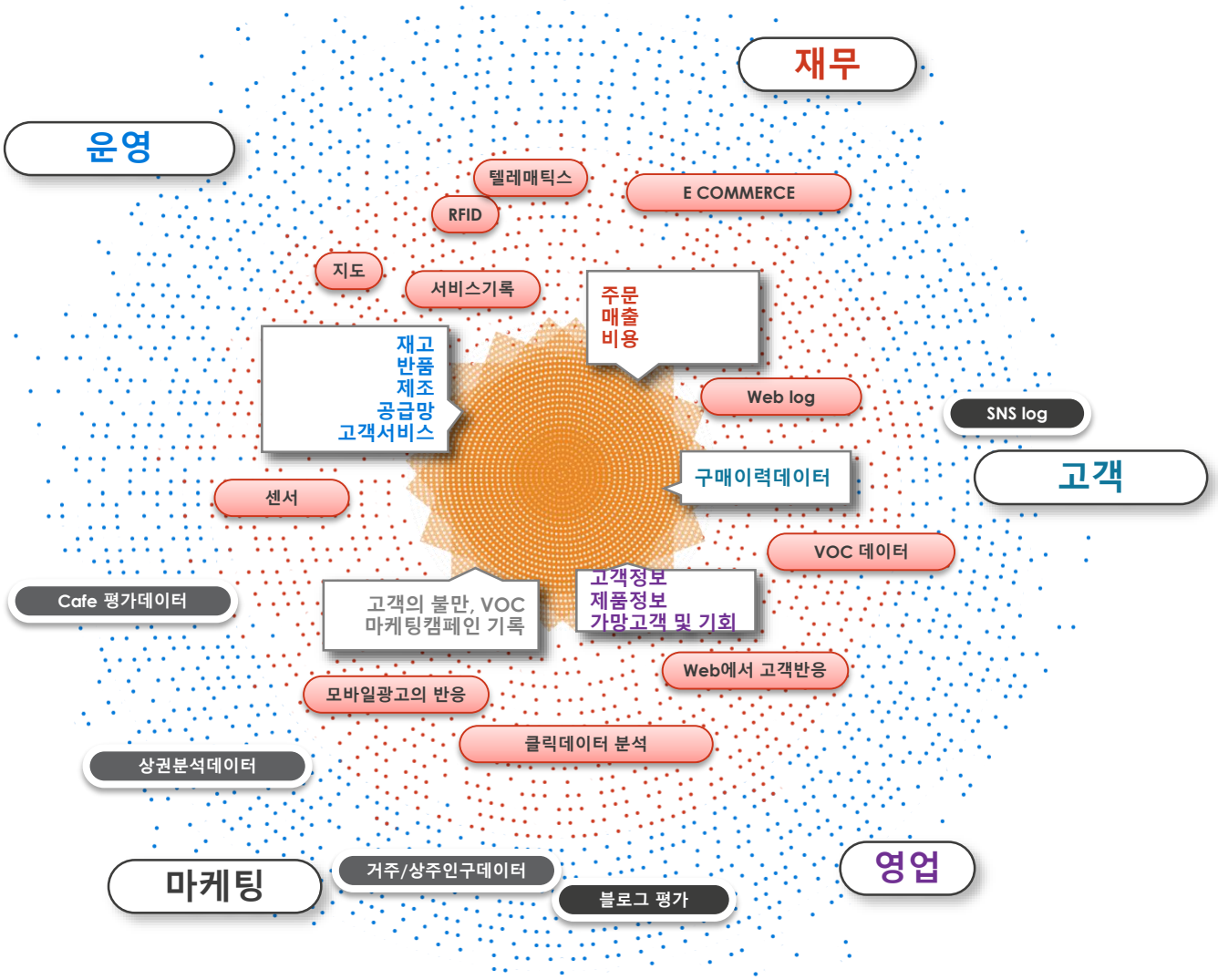
빅데이터가 무엇인가?

데이터는 우리 기업의 주변이 있다...



빅데이터가 무엇인가?

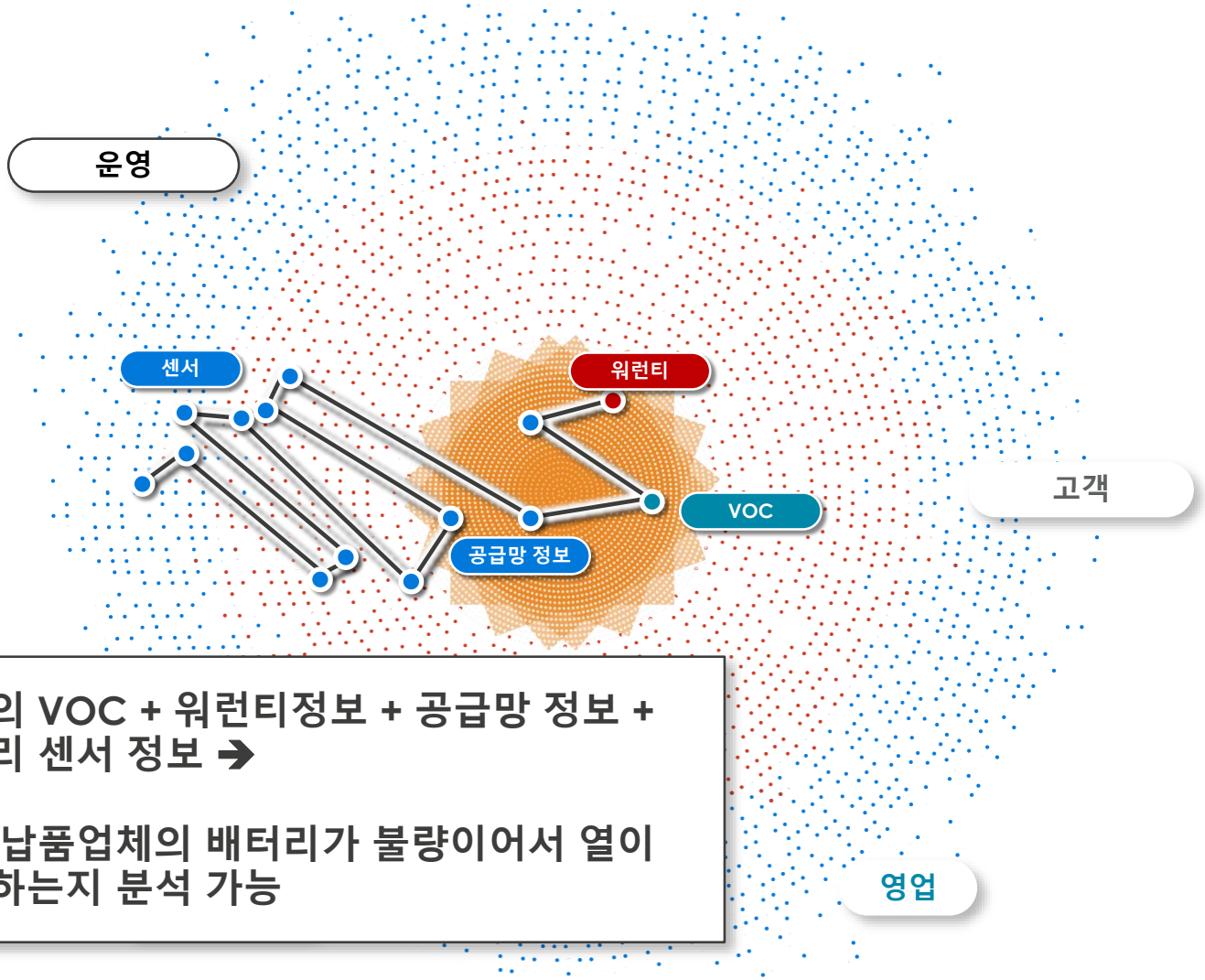
빅데이터는 우리 기업의 주변이 있다...



TERADATA

빅데이터를 가지고 어떻게 문제를 푸는가?

서로 다른 Layer에 있는 데이터를 연결한다



고객의 VOC + 워런티정보 + 공급망 정보 + 배터리 센서 정보 ➔

어떤 납품업체의 배터리가 불량이어서 열이 발생하는지 분석 가능

TERADATA

차세대 정보계 시스템에 대한 요구사항

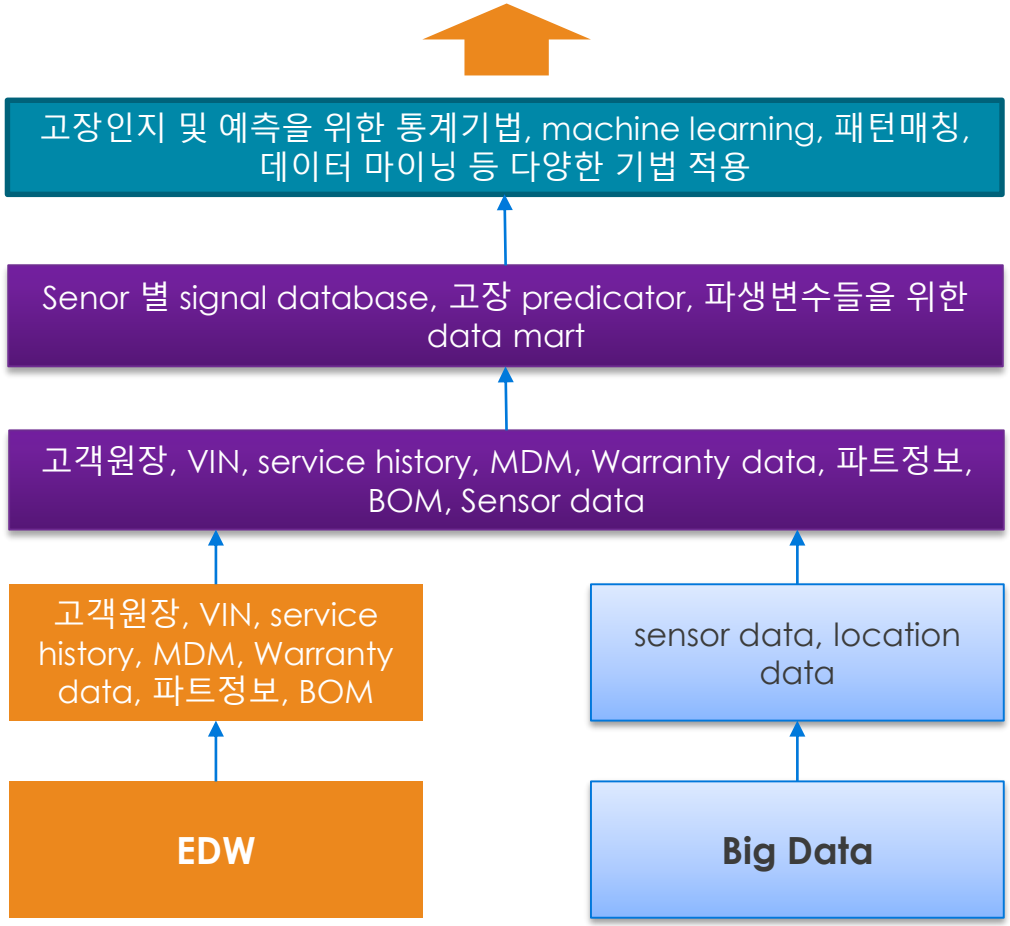
비즈니스 분야

- 현업은 기본 정보계 데이터이든 빅데이터이든 관계없이 비즈니스 분석을 원한다.
  - ✓ EDW: 고객원장, VIN(차대넘버), service history, MDM, Warranty data, 파트정보, BOM
  - ✓ 빅데이터: sensor data, location data
  - ✓ EDW+빅데이터➔ 원하는 분석
- 자연어 처리, 텍스트 마이닝, 그래프 분석, GIS 연결 분석, 머신러닝, 마이닝 등 전통적으로 사용하지 않았던 분석을 하게 된다.
  - ✓ 자동차 part의 문제를 진단하고 예측정비를 위한 패턴분석
- 전체 테이터를 분석하는 "데이터 탐색(data discovery)"을 하게 된다.
- 그러나, 이러한 요구는 한번에 나오지 않는다.
- 현업의 요구사항은 활용하면서 나온다.
- 현업은 지금까지 쓰던 것, 익숙하던 것을 고집한다. (reporting, OLAP tool)
- DW이든 빅데이터이든, 결국 분석하여 보여지는 것은 유사하다. (insight)
- 분석은 하면 할수록 업그레이된다. 변화관리가 생명이다.

차세대 정보계 시스템에 대한 요구사항

비즈니스 Question

차가 고장이 났다면, 센서데이터를 분석하여, 어떤 부속품이 고장났으며, 다른 부속품도 문제가 있는 것이 있으면, 발견하라



TERADATA.

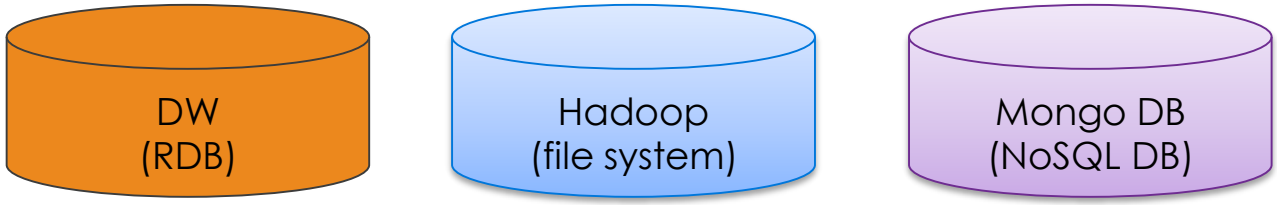
빅데이터 분석의 종류와 기술

분석의 종류	실제 내용	
단순분석(정형분석)	• 주로 Excel이나 SQL에 의해서 4칙 연산 및 그에 따른 그래프. 장표를 만듬 • 이미 그래프와 장표의 형식이 정해져 있음(예: 월별 매출액) • Hadoop에서는 이러한 단순 분석에 Map&Reduce(MR) 프로그래밍을 해야 함. • SQL 하나의 문장으로 같은 결과를 처리할 수 있음.	• MR • SQL on Hadoop • Aster
Ad-hoc분석 (임의분석)	• 미리 정해진 형식에 없는 분석. (ex. 갑자기 임원들이 만들어 오라는 장표) • Excel 같으면 데이터 가공을 다시 해야 하고, SQL같은 경우는 다시 SQL을 만들어야 함 • 최악의 경우에는 데이터 수집부터 다시 해야 함 • Hadoop도 MR 프로그램을 다시 짜야 함	• MR • SQL on Hadoop • Aster
OLAP분석	• OLAP은 On Line Analytical Processing 으로서 다차원 분석이라고 함. • 예를 들면 연도별, 회사별 매출액… 과 같이 ~별. 분석. • 차원과 팩트(fact)를 미리 정해 놓은 점에서는 정형분석이라고 할 수 있음 • 이것은 결국 RDB에서 SQL을 활용해서 나오는 결과임. (fact table과 dimension table의 join)	• RDB • Aster
실시간분석 (interactive query)	• 이것은 SQL on Hadoop 에서 SQL로 query를 던지면 바로 답이 나오는 경우. 데이터는 이미 Hadoop에 들어있어야 함. RDB경우는 이미 실시간 분석이라고 할 수 있음. • Hadoop의 MR 경우는 배치(batch job) 로 돌리므로 실시간 분석은 아님	• SQL on Hadoop • Aster
실시간분석 (CEP)	• 데이터가 계속적으로 들어올 때 (stream data) 이를 실시간으로 그래프를 그린다든가 간단한 분석을 하는 것. 분석된 데이터는 Hadoop이나 NoSQL DB로 들어가게 해서 나중에 배치 분석을 한다	
통계분석	• R, SAS, SPSS 등의 통계 패키지를 활용해서 분석 • - 통계알고리즘 및 데이터 마이닝 기법을 적용해서 forecasting, 시뮬레이션 등에 활용함 • - 빅데이터 분야에서도 여전히 활용 가능함	• R, SAS, SPSS • Aster
기계학습 예측	• 다양한 분석 및 예측모델을 만들어서 분석 • Machine learning	• R, SAS • Aster
감성분석	• 자연어처리, 감성분석은 taxonomy라고 하는 10만 단어이상의 사전을 만들어 일반 text를 계속 비교해나가는 computing-intensive한 job 임. • 과거에도 자연어 처리, 감성 분석은 있었으나, 그것을 처리하기 위해서는 매우 비싼 supercomputer 가 필요했음. • 그러나, in-memory 기술을 활용하고 hadoop, Nosql DB 등이 나와서 대용량데이터의 저장과 자연어처리를 할 수 있게 됨.	• SAS SMA • 한글분석 패키지 • Aster
기타분석	• 기타 pattern matching 기법을 활용한 자동차 표지판 같은 이미지 데이터 인식, CCTV 분석, 인공위성지도 분석 등도 있음. Social network Analysis,	• Aster

TERADATA

차세대 정보계 시스템에 주는 빅데이터의 영향 1

Multi-Database, Multi-platform



- Hadoop이나 NoSQL DB는 기존 IT환경(DW)에 친숙하지 않다
- Table, Access 방식, programs, interface 등 모든 것이 다르다
- 서로 다른 DB, platform 간의 호환성이 없다
- 데이터는 서로 다른 DB 에 존재한다.
- 데이터 관리/메타데이터의 필요
- 현업은 location(DB, platform)에 관계없이 분석을 원한다
- 정합성문제, 데이터 관리문제는 IT 팀에서.

차세대 정보계 시스템에 주는 빅데이터의 영향 2

Big Data Mart들이 양산된다

STT: Speech to Text

- 현업의 요구사항 충족시키기 위해서 그때그때 서로 다른 DB/tool 선정
- 이 Big Data Mart 들은 서로 연관성이 없고, 데이터가 중복
- 기존 DW와 통합이 어렵다 (그때 그때 기준 정보는 DW에서 ETL)
- 분석된 결과의 정합성 확인하기 어렵다
- 데이터 관리가 어렵다
- 용도별 DB 구매로 유지보수 skill이 많이 필요하다 (Multi-DB)
- 용도별 tool/DB 구매로 비용이 계속 들어간다

차세대 정보계 시스템에 주는 빅데이터의 영향 3

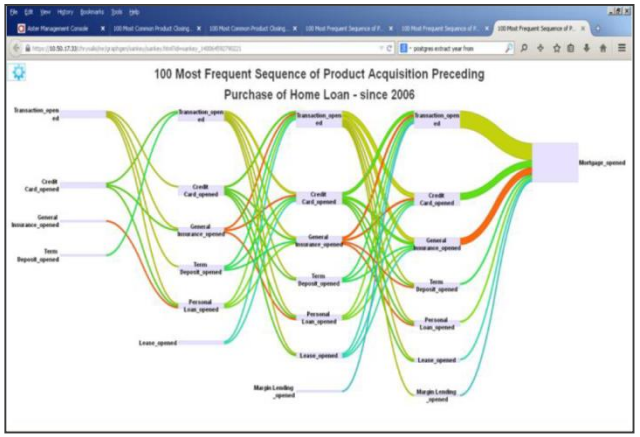
빅데이터 분석 기법이 도입된다

전통적인 통계분석 방식

Fraud detection
Customer seg.
이탈 score ….

- 데이터 마트 필요
- 정형 데이터 only
- 샘플링, Training set, test set, 적용
- Input 변수가 하나 더 생기면?
- Algorithm을 바꾸면?
- Blackbox !

빅데이터 분석 방식



- 전제 raw 필요
- 정형/비정형 both
- 전체데이터 분석하여 시각화
- Input 변수와 상관이 없음
- 다양한 Algorithm 적용
- Whitebox !

TERADATA.

차세대 정보계 시스템에 주는 빅데이터의 영향 4

데이터 레이크(Data Lake)의 등장

ODS(operational Data Store)	Data Lake (~ Hadoop)
Schema (정형only)	Schema Free (정형+비정형)
Schem가 다르면 먼저 DB에 반영	빠른 데이터 로드/축적
테이블/키를 가진 Relational 구조 only	복잡한 데이터구조도 Ok (hierachical data)
적재하면서/적재이후 cleansing	일단 데이터 적재후 cleansing
기존 RDB / DW	대용량 데이터 적재 비용절약

차세대 정보계 시스템에 주는 빅데이터의 영향 4

데이터 레이크(Data Lake)의 활용

#1 데이터 스테이징(data staging) → 일단 데이터를 적재후 처리

#2 ETL 작업이 매우 CPU/Disk를 필요로 하는 작업 → 싼 Hadoop 사용/비용

#3 비정형/복잡한 구조를 가진 데이터 적재

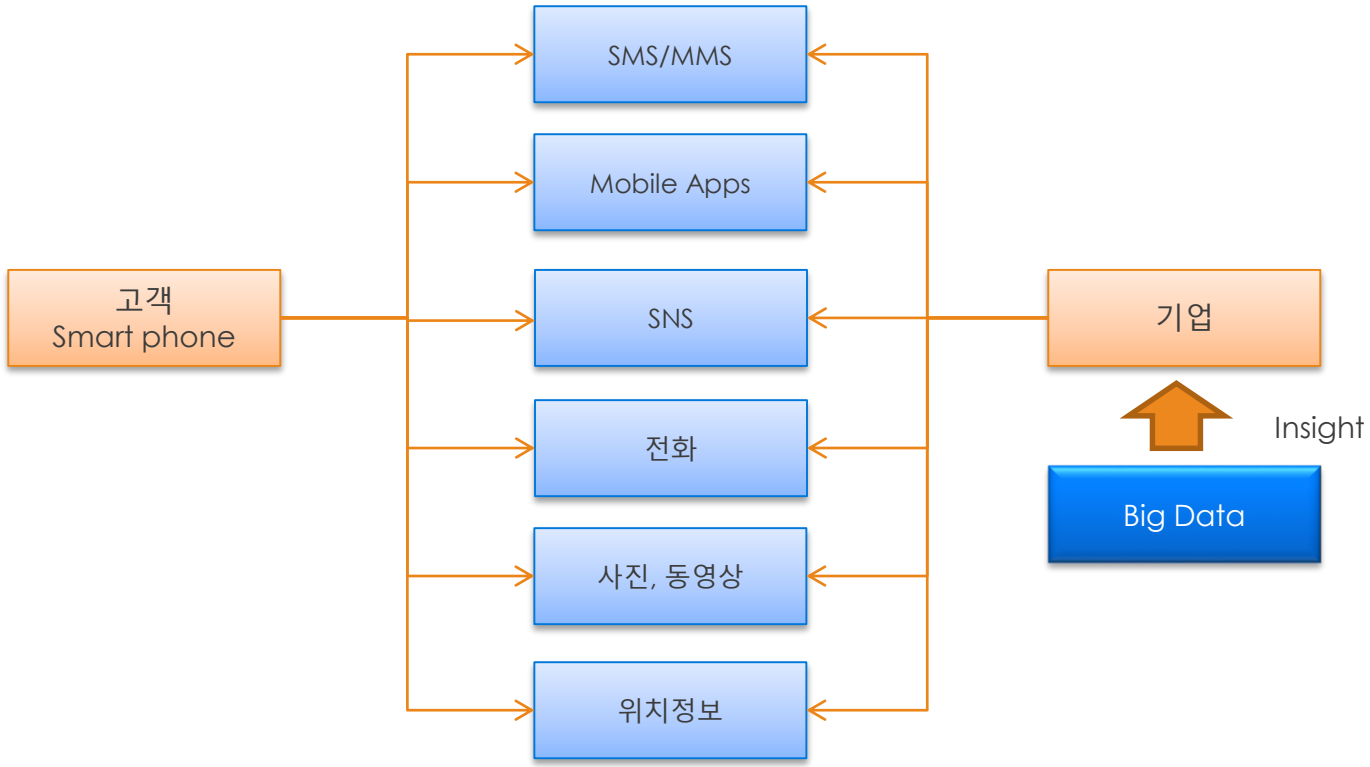
#4 장기간 많은 데이터 보관 (sensor 데이터)

#5 기존 아카이브용 Tape drive 대체

차세대 정보계 시스템에 주는 빅데이터의 영향 5

전사적 Data Driven Decision Making이 필요하다

- Digitizing business (Digital Transformation) 시대는 빠른 놈이 큰 놈을 이긴다
- 스마트폰을 가진 고객관리 시대와 스마트폰을 갖지 않은 고객관리 시대는 근본적으로 다르다
  ➔ 기존의 기업 내부 프로세스와 시스템은 스마트폰을 갖지 않은 고객관리 시대에서 만들어진 것들이다.
- 기업의 기본 업무 프로세스와 시스템에 대한 전면적인 재검토 필요
  (고객서비스/마케팅/영업/홍보..)



TERADATA.

현재까지 Big Data 기술의 발전 정도

빅데이터 기술은 많이 발전했으나 Hadoop의 약점...

- 현재 Hadoop의 발전 단계
  - ✓ 상용 Hadoop vendor는 SQL on Hadoop의 기능을 대폭 향상 시켰다
  - ✓ 그러나, 상용 Hadoop은 아직까지 full ANSI SQL을 지원하지 못한다
  - ✓ 상용 Hadoop 내에서 통계, 마이닝, 지리정보와 통합, 자연어 처리 등은 하지 못한다. 이것은 제3의 tool을 따로 사용해야 한다.
  - ✓ 상용 Hadoop에 있는 데이터를 관리하는 메타데이터 관리 tool은 자사 Hadoop에만 국한된다. (cross platform metadata 관리는 못함)
  - ✓ 상용 Hadoop에서 기존 RDB나 NoSQL DB와의 통합은 아직 미약하다
  - ✓ Hadoop, RDB, NoSQL DB를 아우르는 cross-platform query tool 은 hadoop 진영에서는 아직 없다
  - ✓ Virtual Mart 개념을 지원하지 못한다
- 이러한 상황에서 차세대 정보계 시스템을 구현하는 새로운 기술이 필요하다

Big Data 시대의 IT Architecture 입장에서 준비사항

지금부터 빅데이터 시대에 전사적인 Architecture를 그려야

- Hadoop에 대한 기술을 정확히 이해 (Hadoop의 장점/단점)
- 내부에서 collection이 가능한 빅데이터 정의/수집/적재 방안 탐구
- 외부에서 활용가능한 데이터 확인/구매 or 획득방안 연구
- Big data 분야의 솔루션맵
- Multi-database, query에 대한 기능/아키텍쳐
- Big data governance 컨설팅/내부 architecture 팀
  - ✓ IT적인 측면의 빅데이터 Architecture 설계
  - ✓ Multi-DB에 대한 데이터관리 및 메타데이터 구조 설계
  - ✓ 데이터 정합성 확보방안
  - ✓ Multi-DB 환경에서 query설계 및 개발 방안

빅데이터 솔루션 및 서비스 Map (외산솔루션)

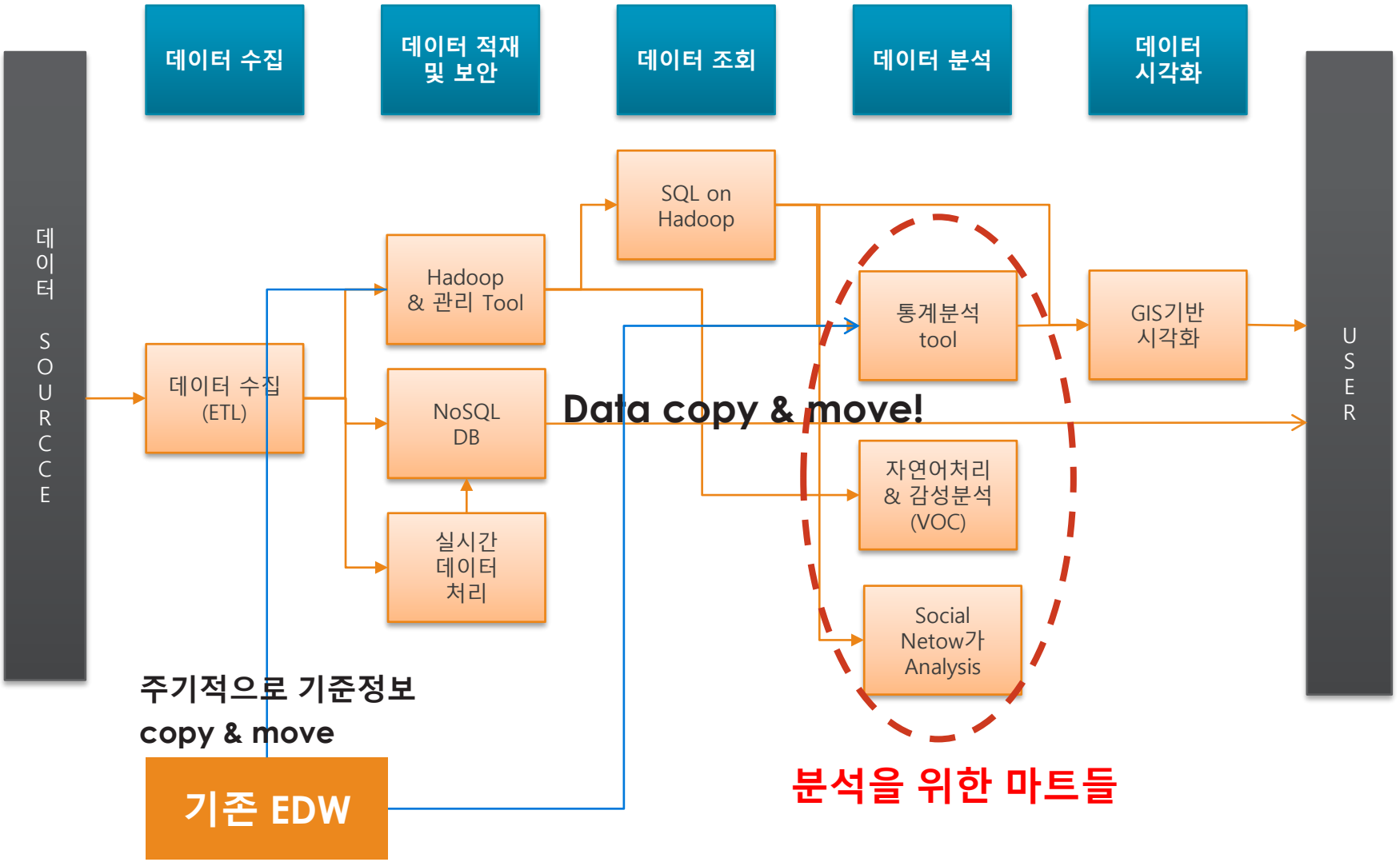
Open Source
유료 sw

빅데이터를 처리, 분석하기 위한 각 분야의 오픈 및 테라데이터 Aster위치

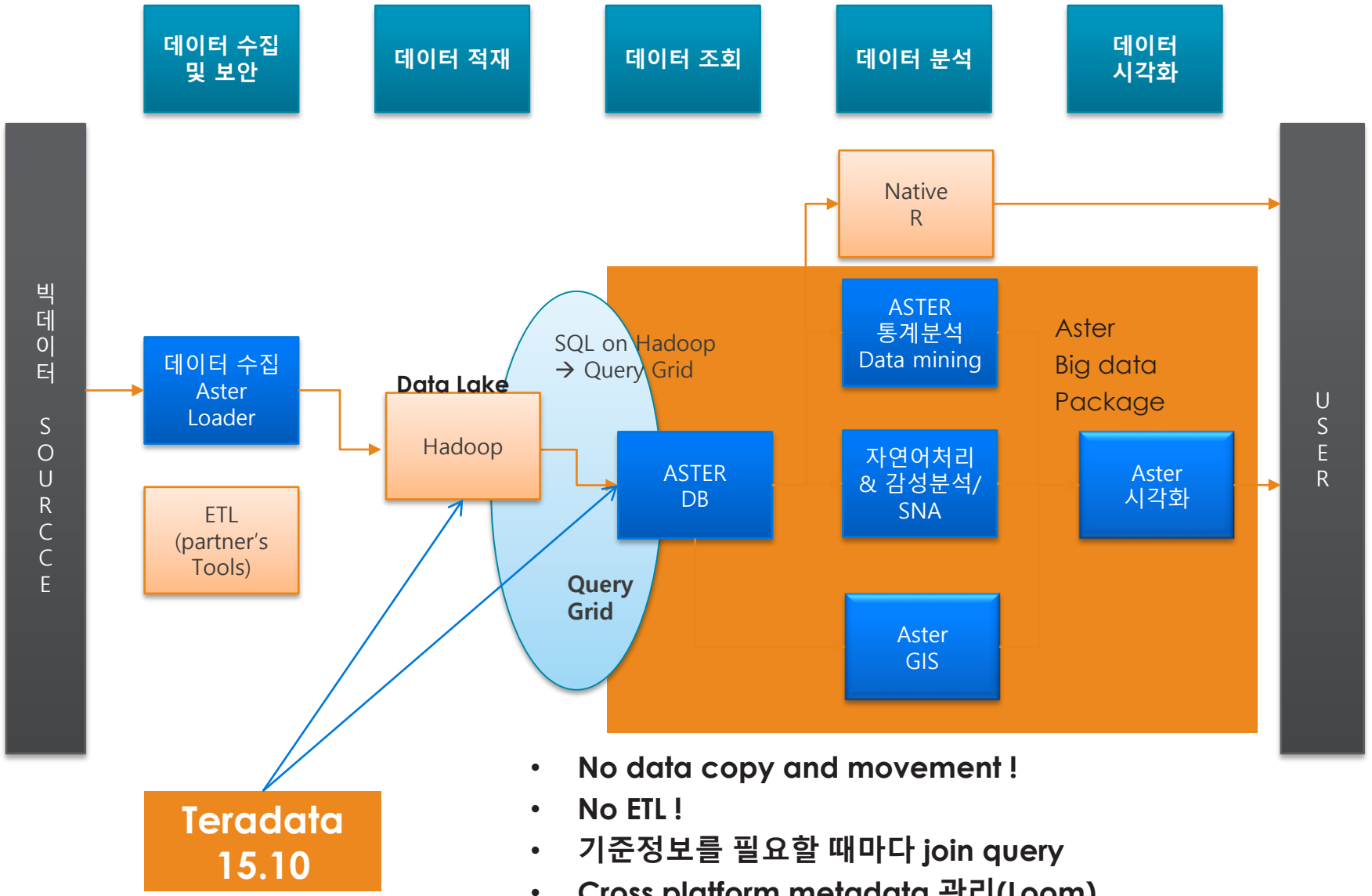
빅데이터 인프라									빅데이터 서비스	
데이터 수집	데이터 적재	데이터 조회 SQL	NoSQL	보안	실시간 데이터분석	통계분석	시각화	Hardware Appliance	Crawling/ 감성분석	Cloud for Big Data
Informatica	Apache Hadoop 2.0		Cassandra	Vormetric	SAP Hana	R	D3/ Visual.ly	Oracle Exadata	Salesforce.com Radian6	Amazon (IaaS+ Hadoop)
Talend (Open Studio)	Cloudera (CDH 4.0 Impala 2.0)		Mongo		CEP Esper	SAS	Qliktech	EMC Greenplum	SAS SMA	SoftLayer (IaaS+ Hadoop
IBM InfoSphere DataStage	HortonWorks (Data Plaform 2.0 Stinger)				Oracle	SPSS	Micro Strategy	Teradata Aster		Rackspace (IaaS+ Hadoop)
	MapR (M5 hadoop, M7 hbase)		Riak		Tibco	Tableau		IBM Netizza		Cloudant (DBaaS)
	Teradata Aster					Spotfire				Amazon Dynamo (DBaaS)
	Splunk (proprietary DB)					Aster	Aster		Aster	Teradata cloud

일반적인 Big Data System Architecture 문제점

다음과 같은 architecture로 가게 되면 필연적으로 data copy & move가 빈번



Teradata Unified Data Architecture



- **No data copy and movement !**
- **No ETL !**
- **기준정보를 필요할 때마다 join query**
- **Cross platform metadata 관리(Loom)**

TERADATA.

Teradata QueryGrid™

Business Users

IDW

TERADATA DATABASE

Discovery

TERADATA ASTER DATABASE

Analysts

HADOOP

Push-down to Hadoop System

TERADATA ASTER DATABASE

SQL, SQL-MR, SQL-GR

TERADATA DATABASE

Multiple Teradata Systems

RDBMS DATABASES

Push-down to Other Database

MONGODB DATABASE

Push-down to NoSQL Databases

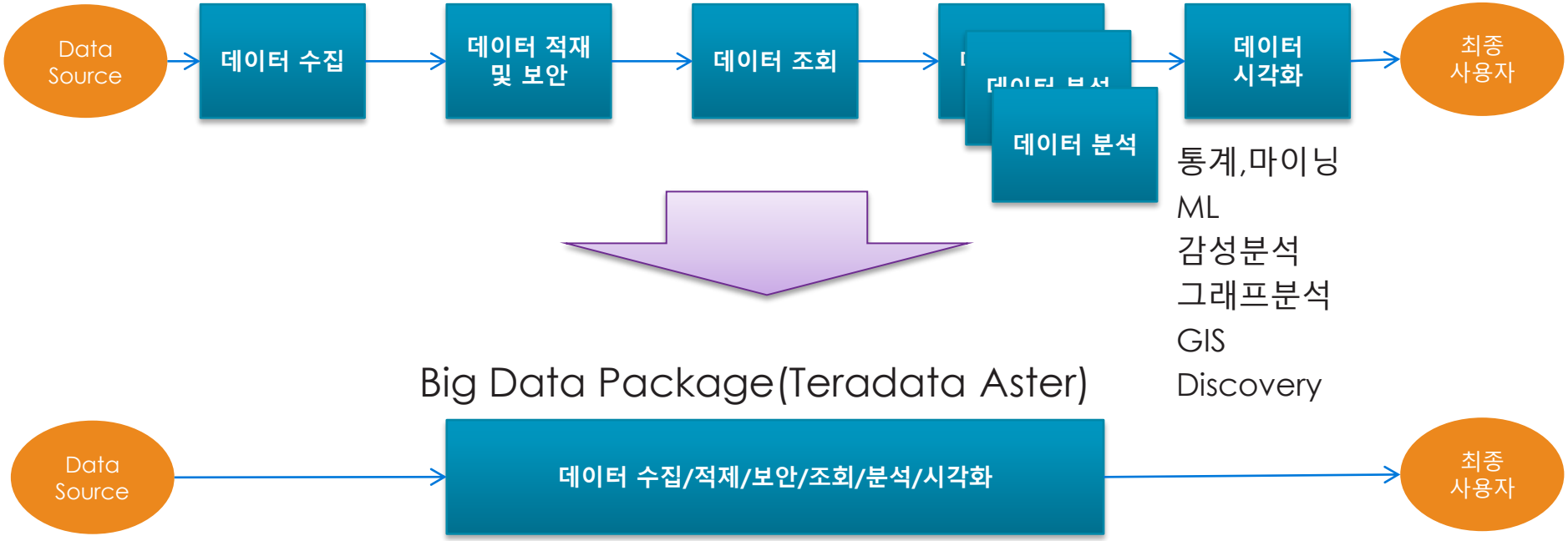
COMPUTE CLUSTER

Run SAS, Perl, Ruby, Python, R

Aster (통합 Big data 패키지)

Big data의 어려운 문제들을 해결하기 위한 통합 솔루션

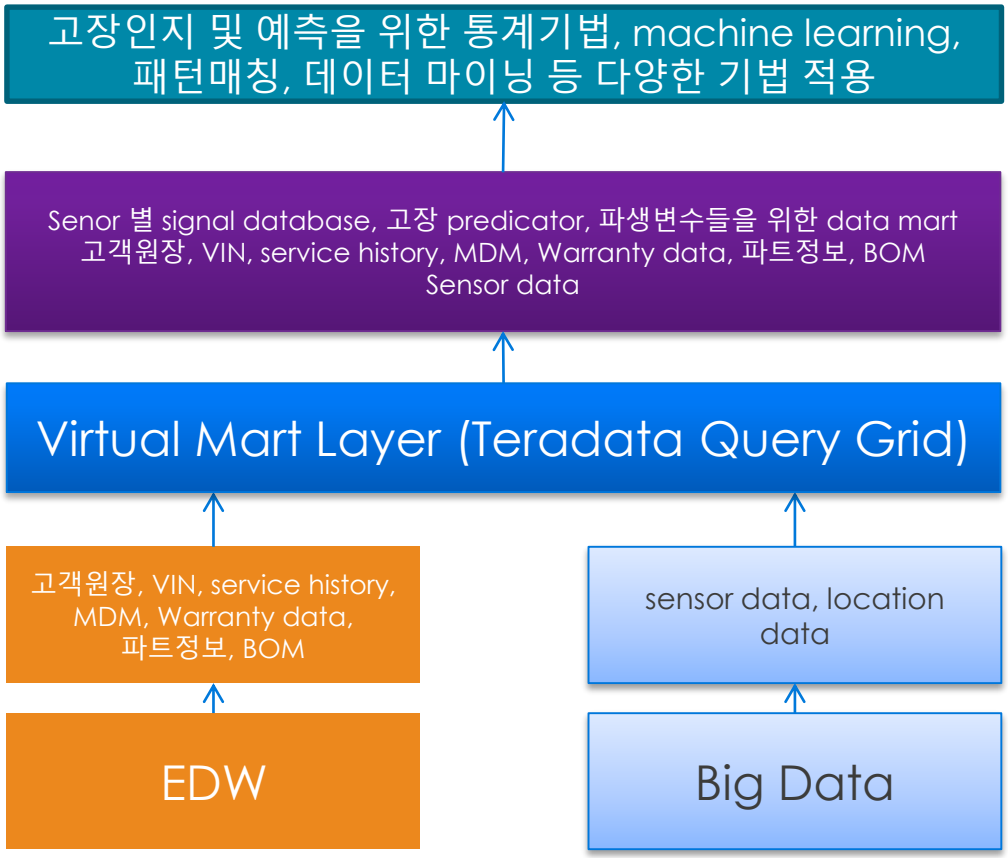
기존 big data 데이터 처리 단계



- Progress 기반 RDB (40PB-ebay)
- 종합분석 모듈 탑재(통계/마이닝/ML/감성분석/그래프분석/GIS/시각화
- Integration with R, 파이선
- Query Grid 통해서 기존 Hadoop/Oracle/Mongo DB 데이터 액세스

현업을 위한 가상 마트(Virtual Mart)

현업은 Aster 내에 있는 Virtual Mart를 통해 모든 데이터를 분석한다



- 현업은 필요한 정보를 mart 형태로 접근한다. (virtual mart)
- 실제 데이터는 기존 EDW와 big data 시스템에 있다.
- 이것을 중간 단계인 virtual mart layer 해주며, 이것은 Teradata의 Query Grid가 해준다.

ASTER

Big Data Apps and Teradata Aster AppCenter

Big Data Apps/AppCenter

Why is this important?

애스터의 앱센터는 이미 테라데이터에서 미리 개발된 분석로직과 코드를 마치 스마트폰의 앱스토어와 같은 개념으로 애스터의 분석엔진 위에 돌아가도록 만든 것입니다. 기존에 개발된 분석로직도 탑재가 가능합니다.

Bridge The Gap
With
Big Data Apps

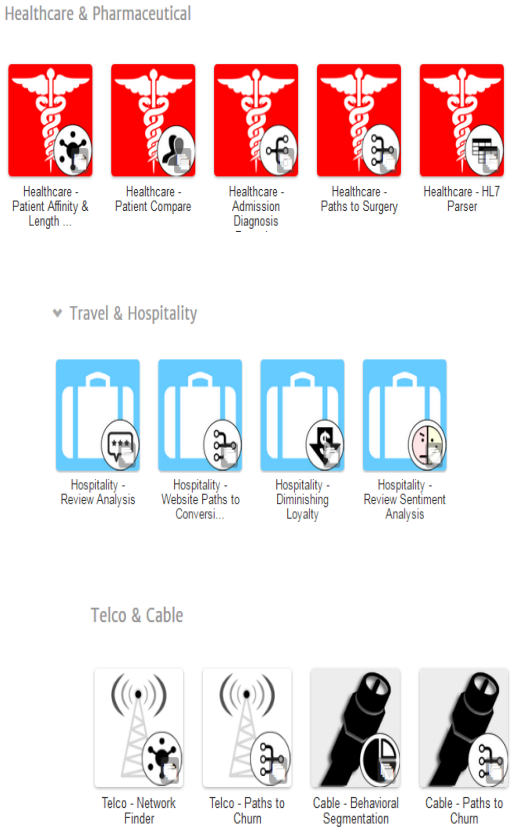
TERADATA.

Big Data Apps/AppCenter

- Industry focused to address specific business challenge
  - Path to Churn
  - Sentiment Analysis
  - Influencer Analysis
  - Marketing Attribution

- Delivered as ***pre-built templates*** that can be configured by Teradata Professional Services

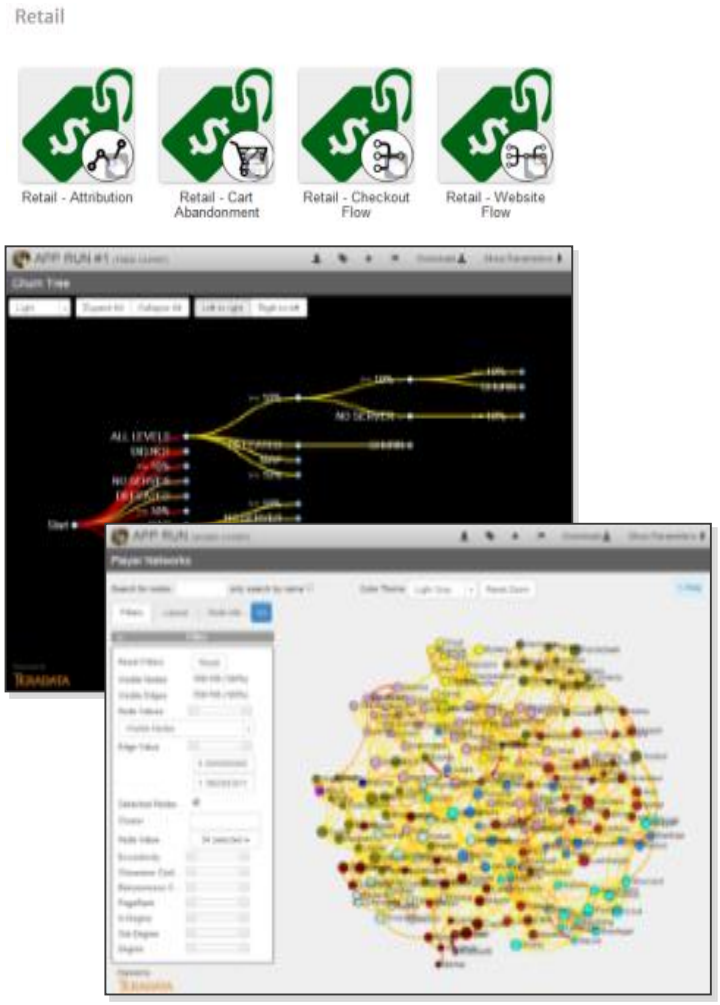
- Powered by Teradata Aster AppCenter ™ providing a common framework to build, deploy, shared and consume



TERADATA.

Retail Big Data Apps/AppCenter

- Retail App Templates
  - Attribution (multi-channel)
  - Shopping Cart Abandonment
  - Checkout Flow Analysis
  - Website Flow Analysis
  - Customer Product Review Analysis
  - Market Basket & Product Recommendation
- Accelerates time to value
- Delivered as pre-built templates for PS (테라데이터 컨설팅) to configure and extend

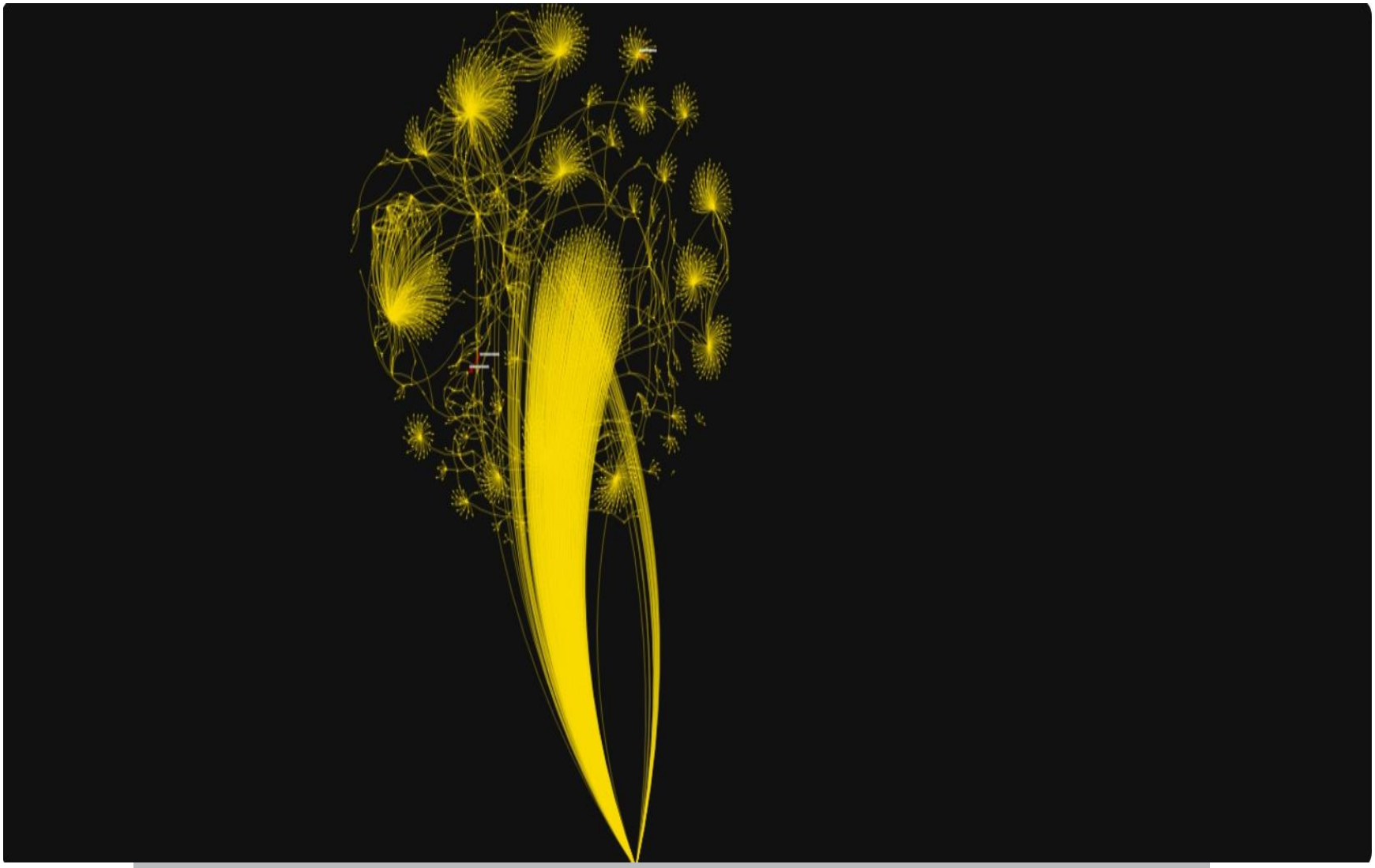


TERADATA.

Aster Visualization

Aster는 통합 big data 패키지로서 유려한 visualization tool을 갖고 있다

The Art Of Analytics Fund Chain View:
View of Fund Flow within Supply Chain (기업간 돈의 흐름)

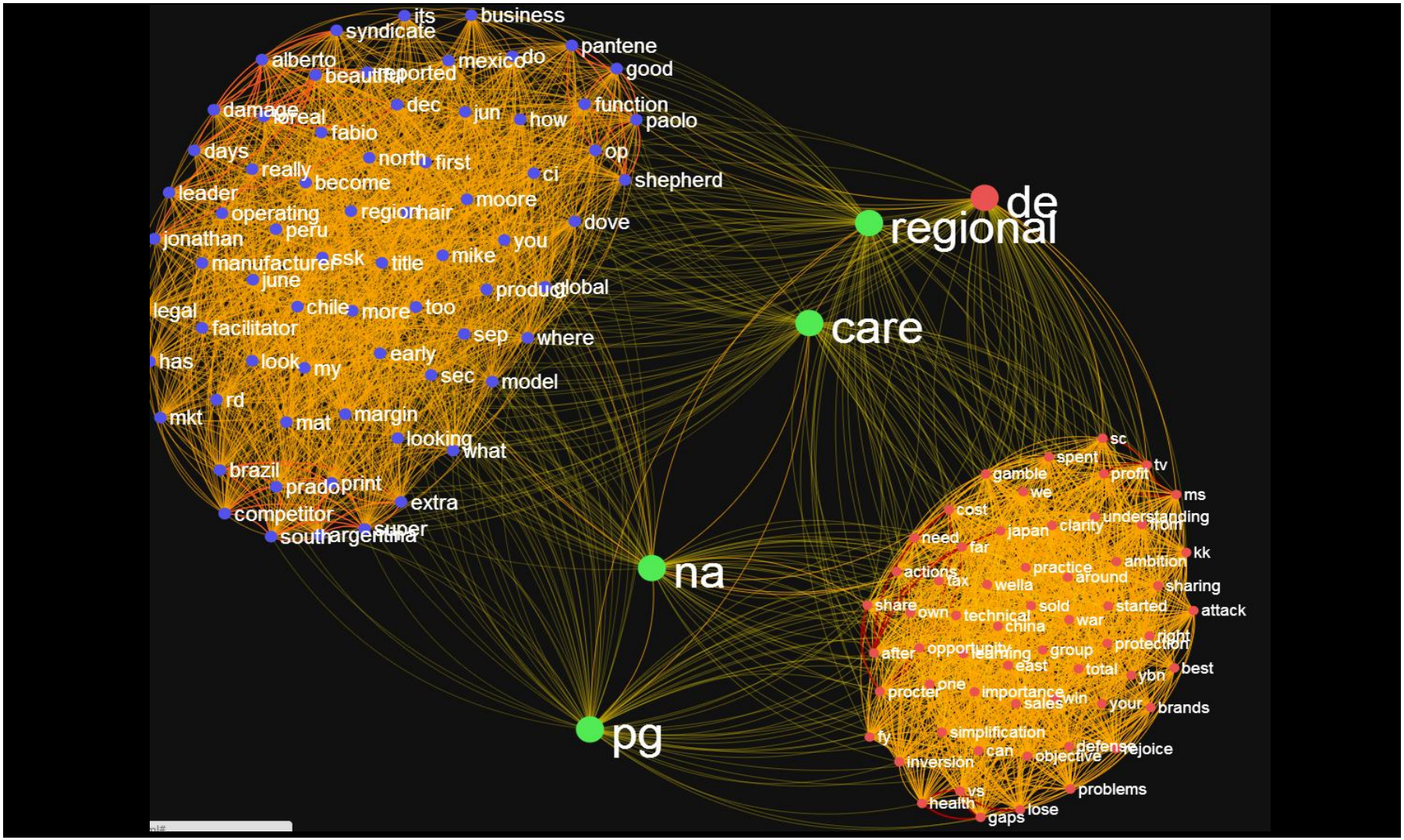


Analyst George Kong Beijing

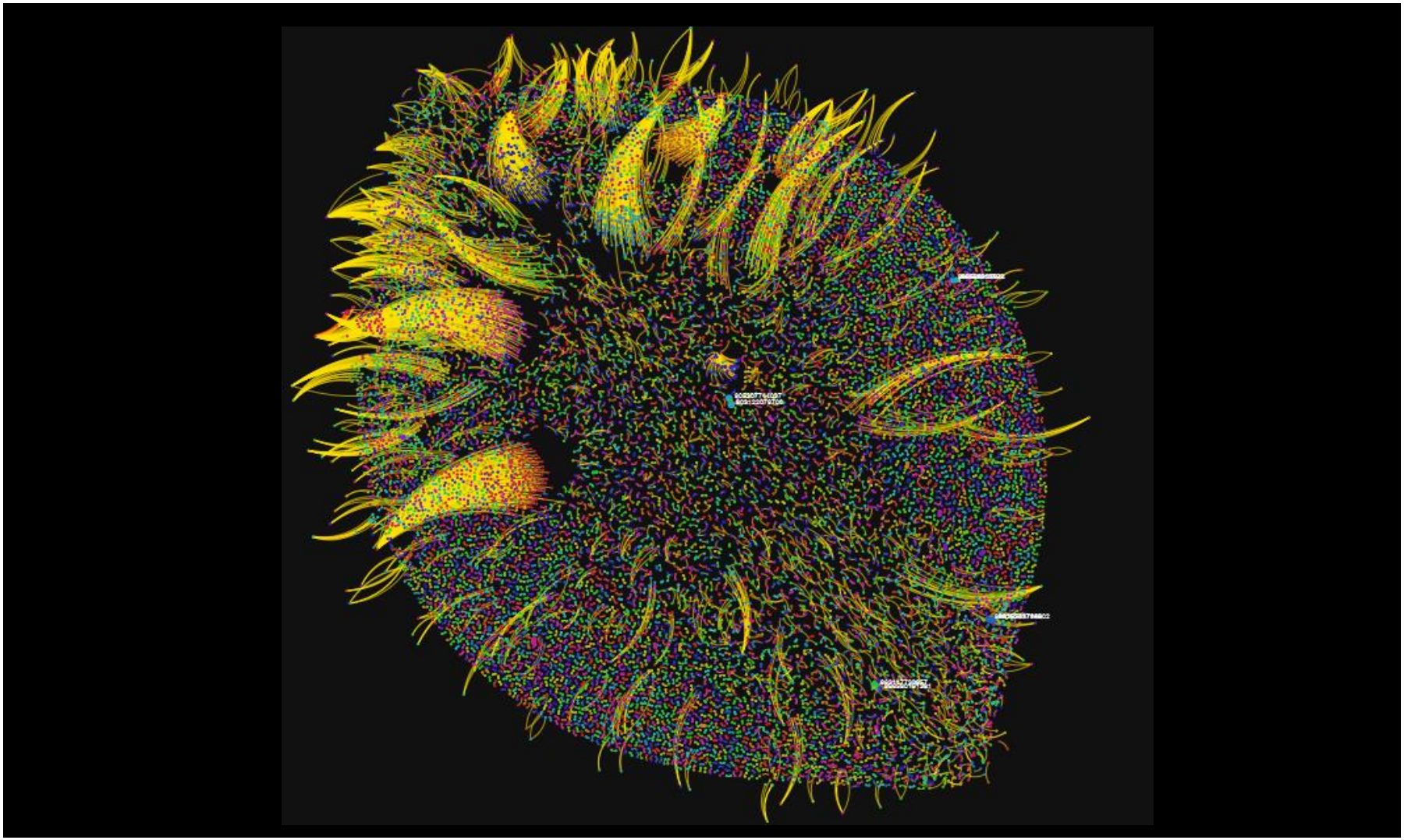
Working with the China Banks to better support the financing of Chinas automobile Industry produced this amazing image that shows the flow of funds through a supply chain, each dot a company

TERADATA.

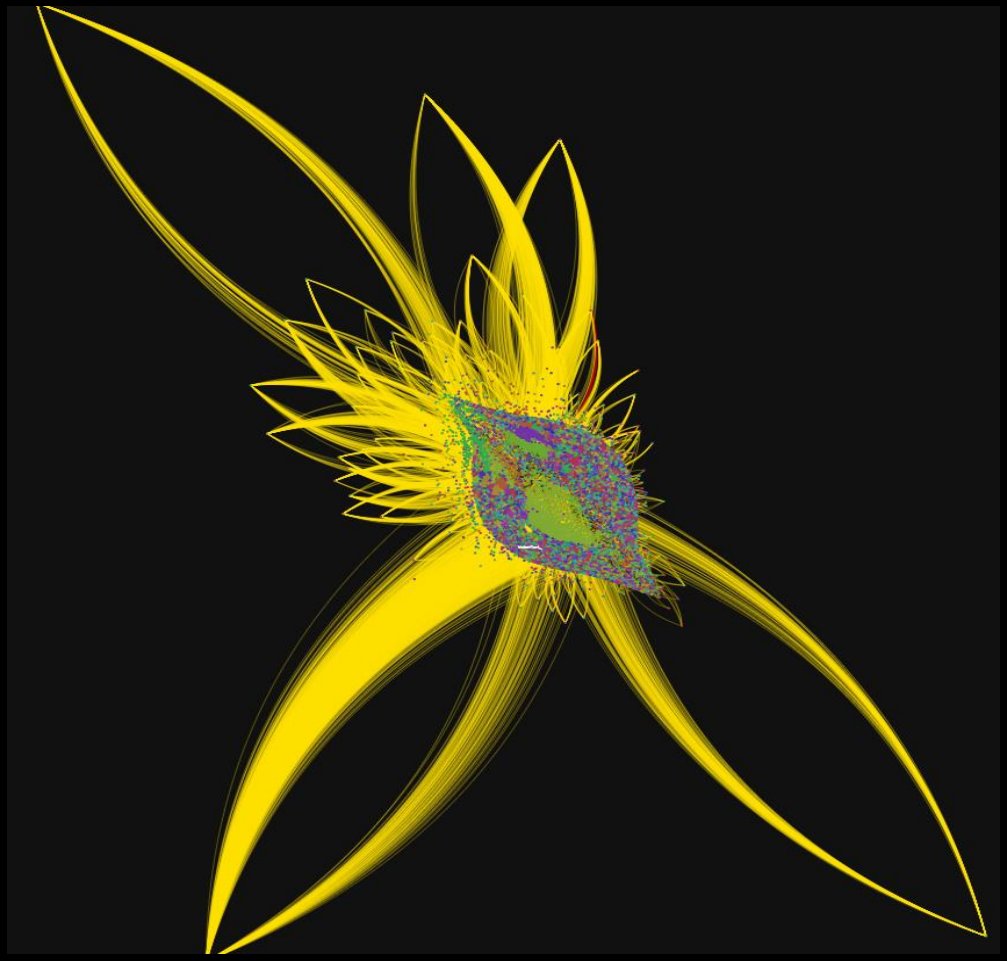
The Art of Analytics: Word co-occurence



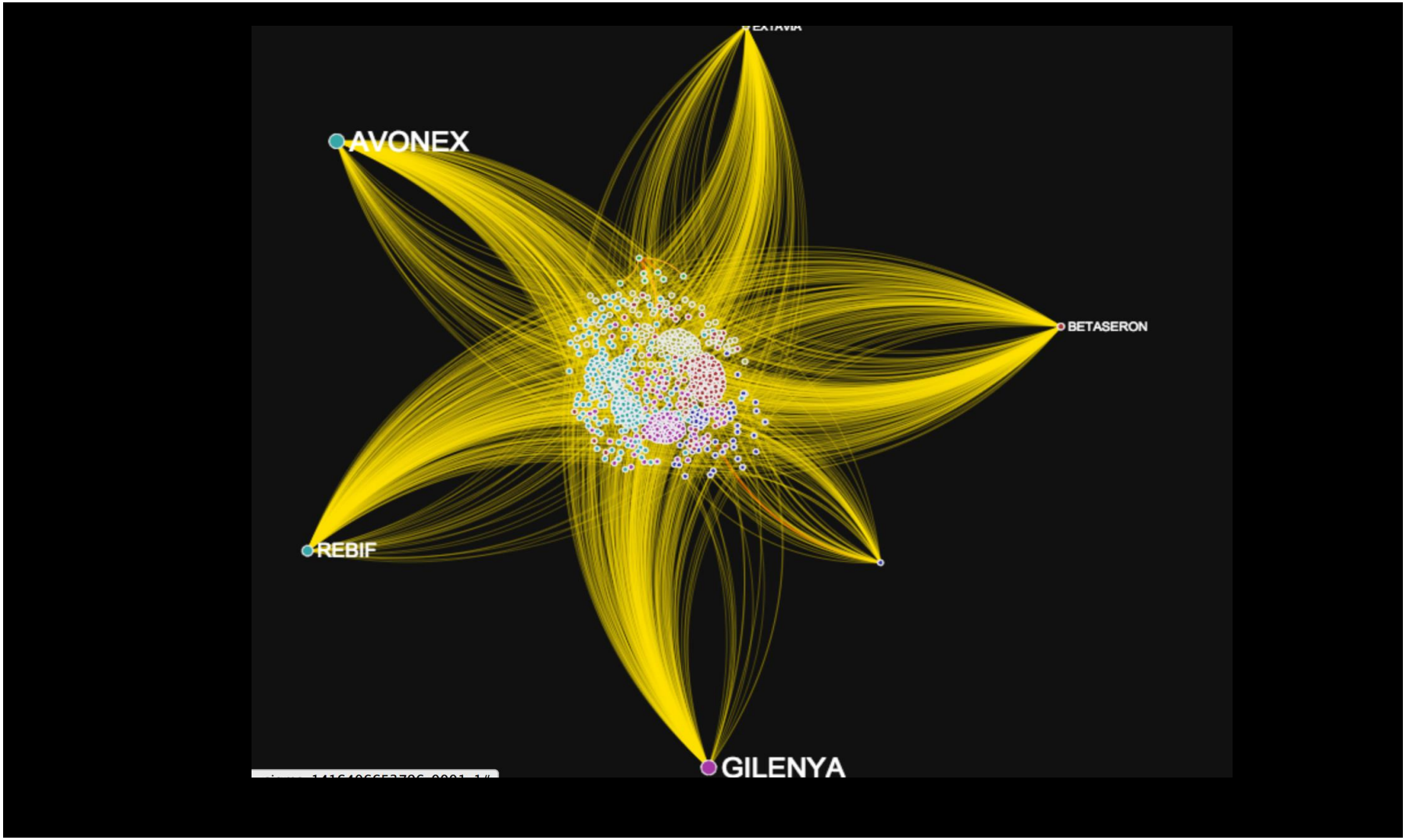
The Art of Analytics: Telco Calling circles



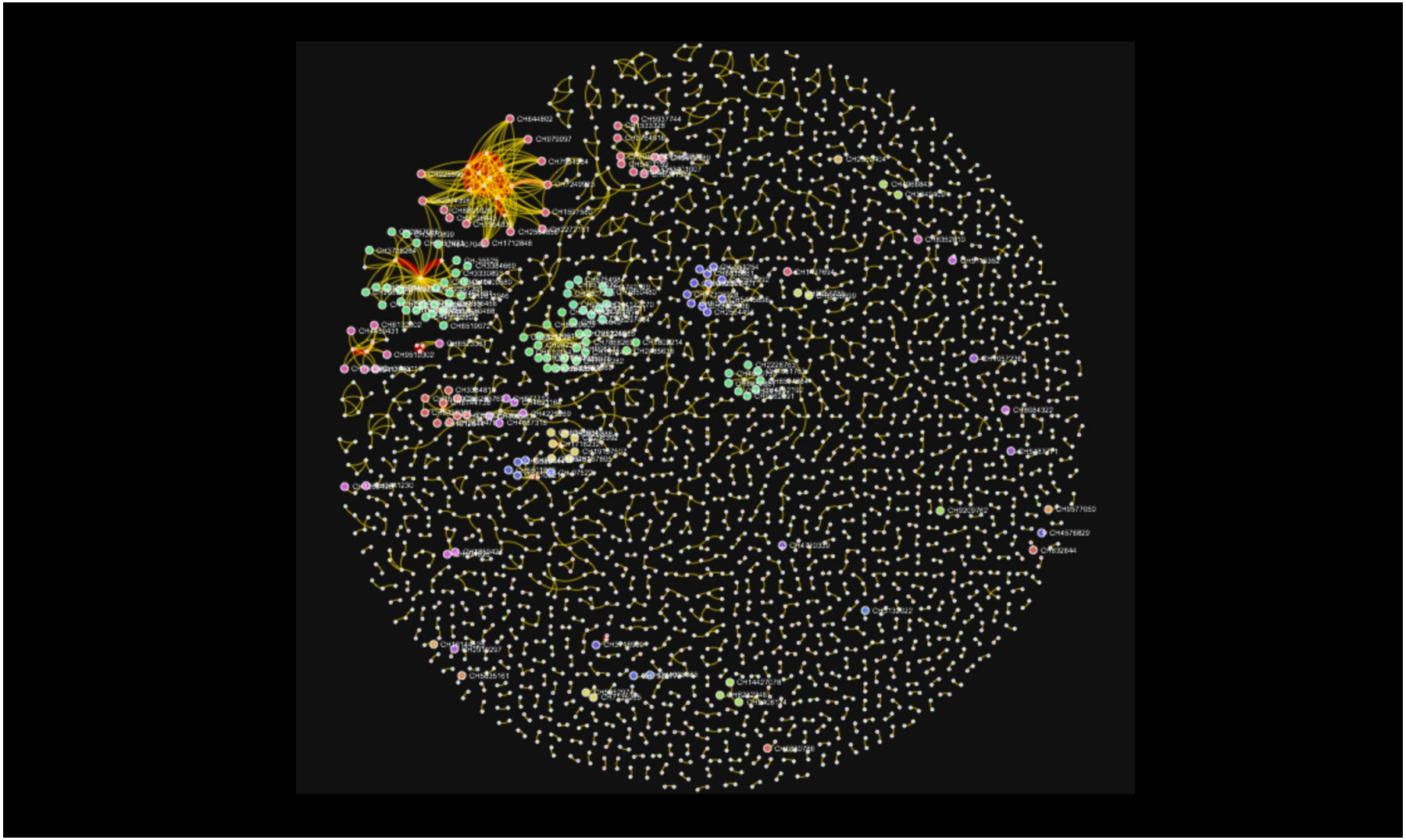
The Art of Analytics: Twitter Influencers



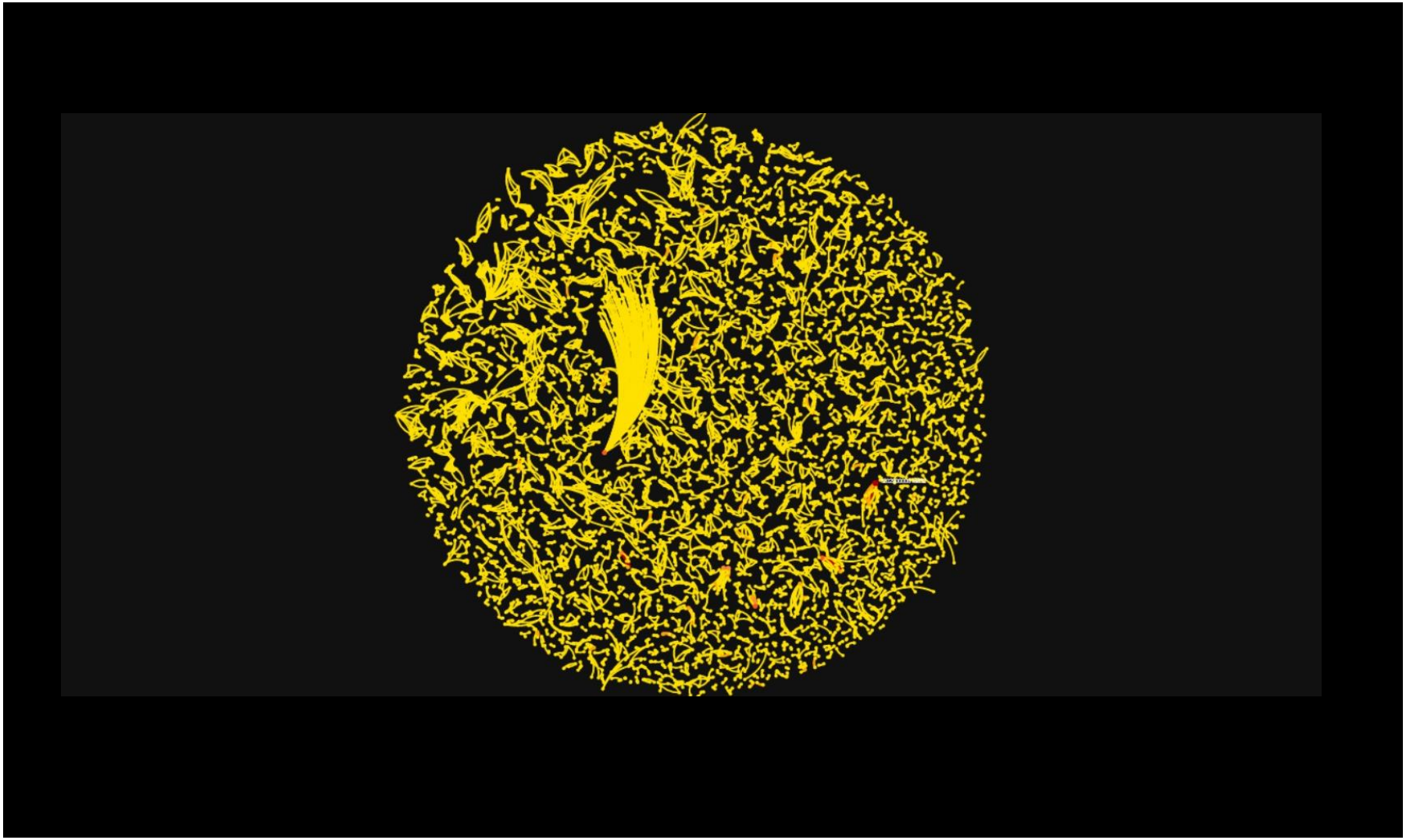
The Art of Analytics: Drugs v side-effects



The Art of Analytics: Fraud Networks

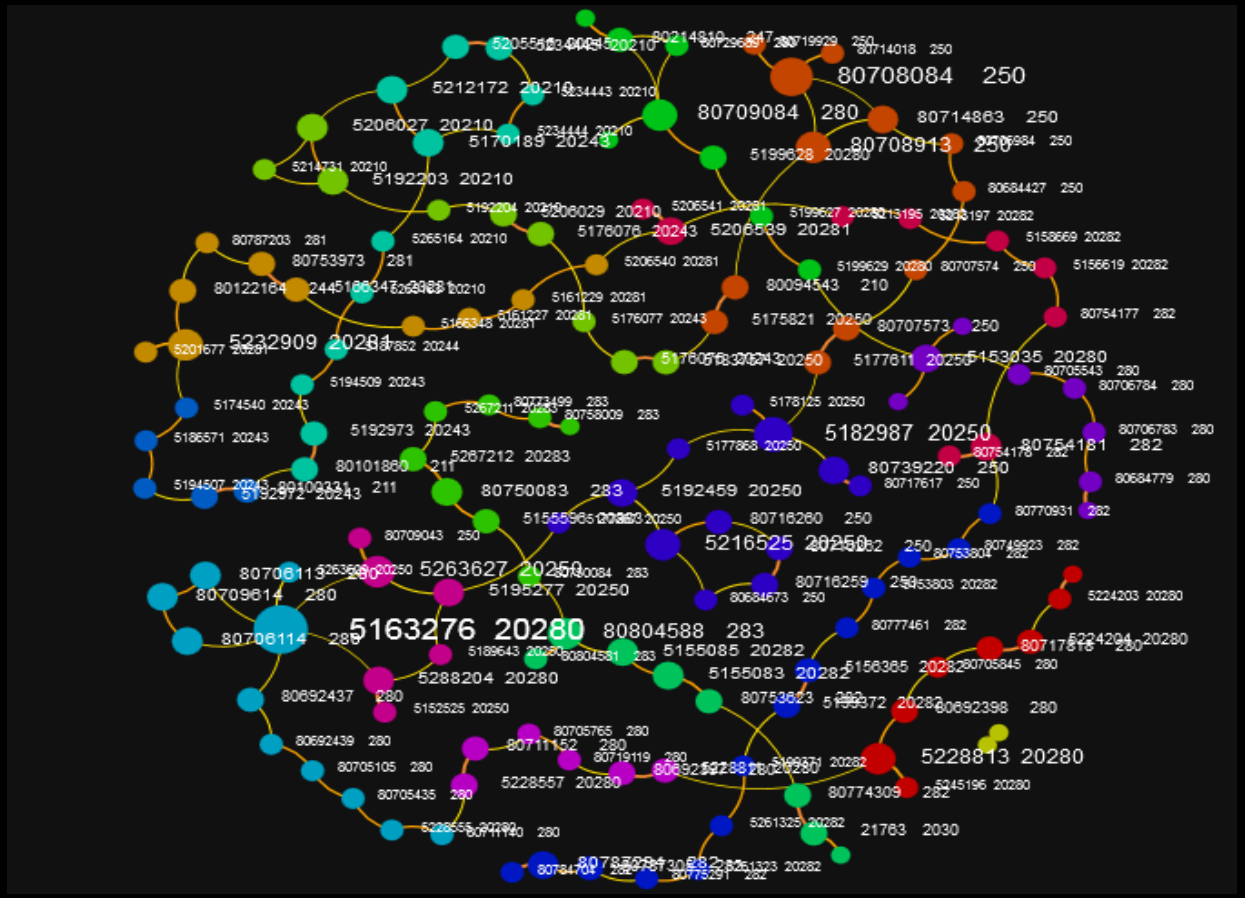


The Art of Analytics: Guarantee Solar Flare
An overview of guarantor & guaranteed enterprises (보증)



Art Of Analytics: Train Journey

A Train trip through Sydney (기차타고 갈 때 기지국이 통화 시스널을 연결하는 모습)



Analyst Sundara Raman Sydney

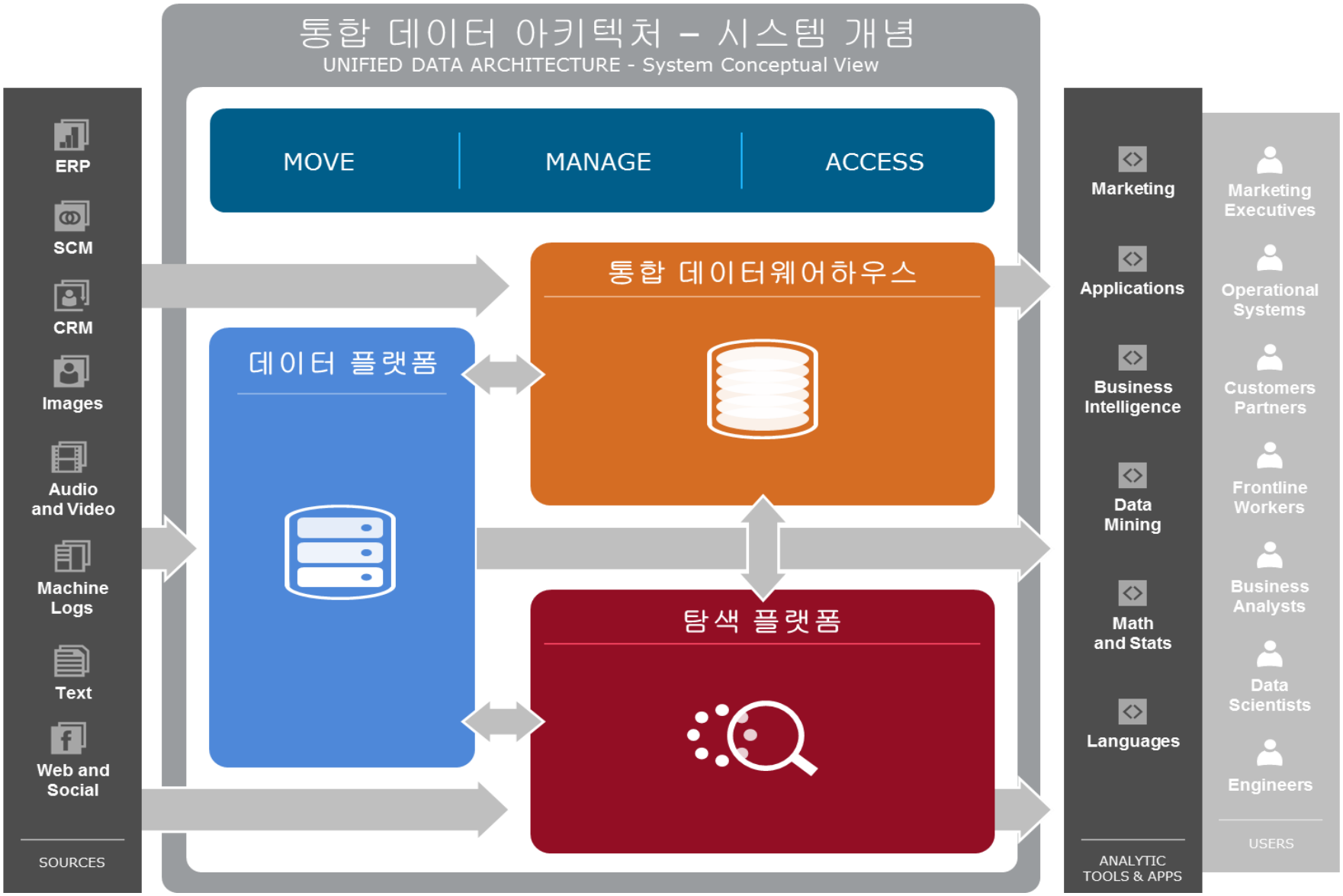
Taking a train ride in Sydney Australia armed with a Samsung Galaxy 3 Sundaras journey maps the cell towers that pick up and exchange signals as he goes.

TERADATA

The Art Of Analytics: Simbox Squid
Sim Box Fraud (simbox를 통한 국제전화 fraud)

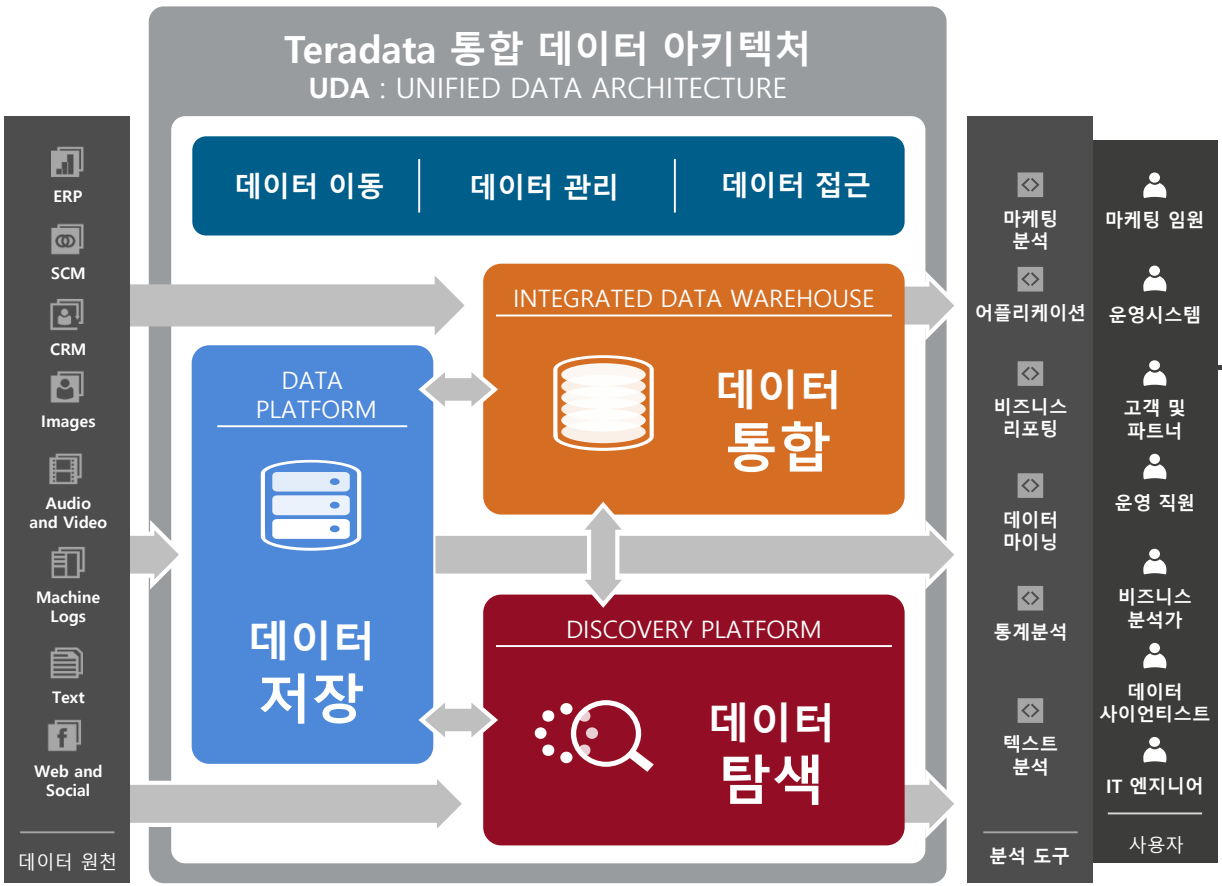
Analyst Ross Farelly Jakarta
Each dot a sim card number this awesome squid diagram appeared in simbox fraud analysis

UDA에 대한 소개



UDA에 대한 소개

선진 기업의 데이터 분석 경험을 기반으로 데이터 저장, 탐색, 통합을 위한 검증된 플랫폼과 제반 기술을 제공합니다.



데이터 저장

- 하둡 포트폴리오
- Teradata 1700

Teradata Hadoop

데이터 탐색

- 모든 데이터와 다양한 분석
- SQL-MapReduce

Teradata Aster

데이터 통합

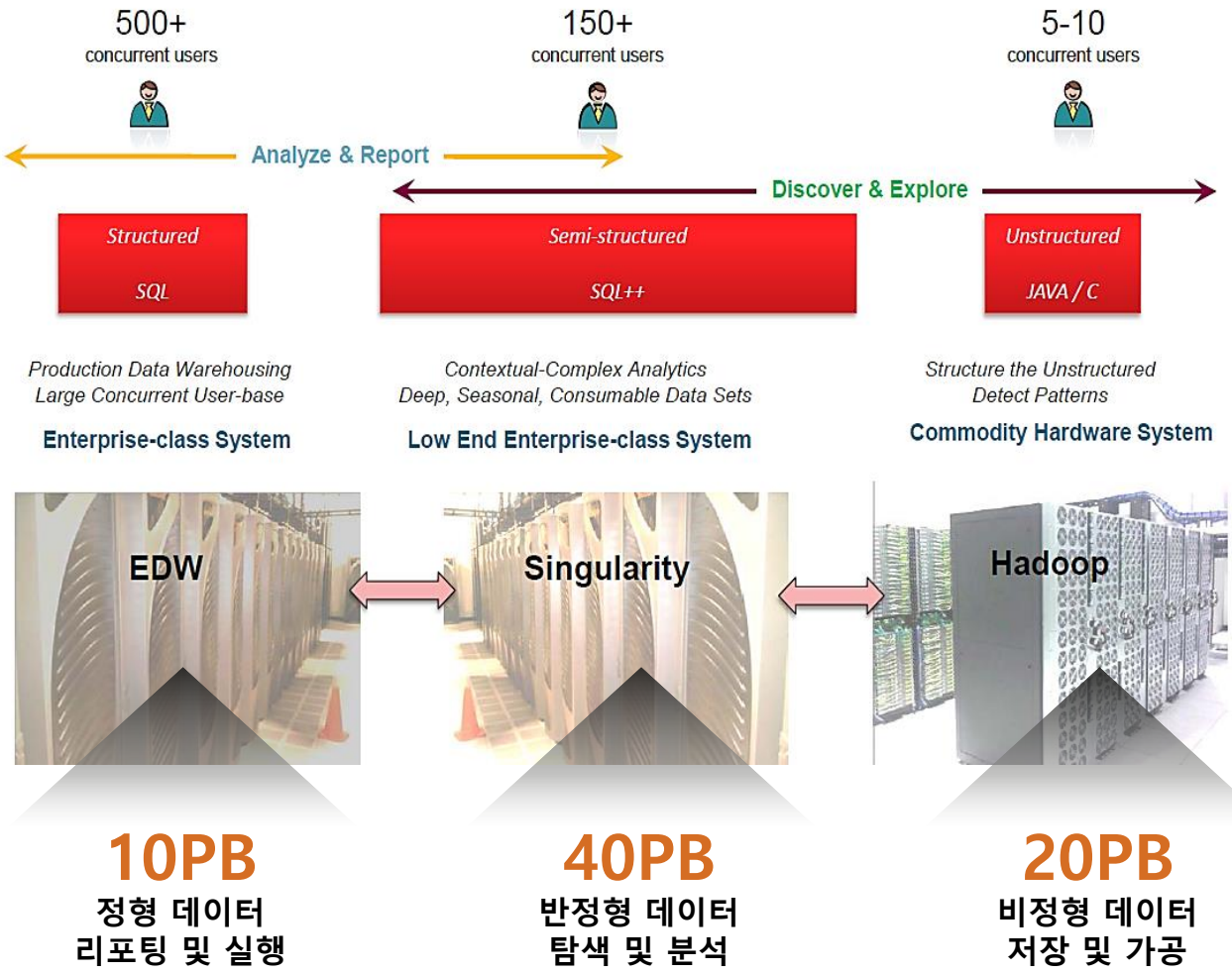
- 전사 사용자 데이터 기반 업무 활용 및 적용

Teradata IDW*

* IDW (Integrated Data Warehouse) : 통합 데이터 웨어하우스

TERADATA.

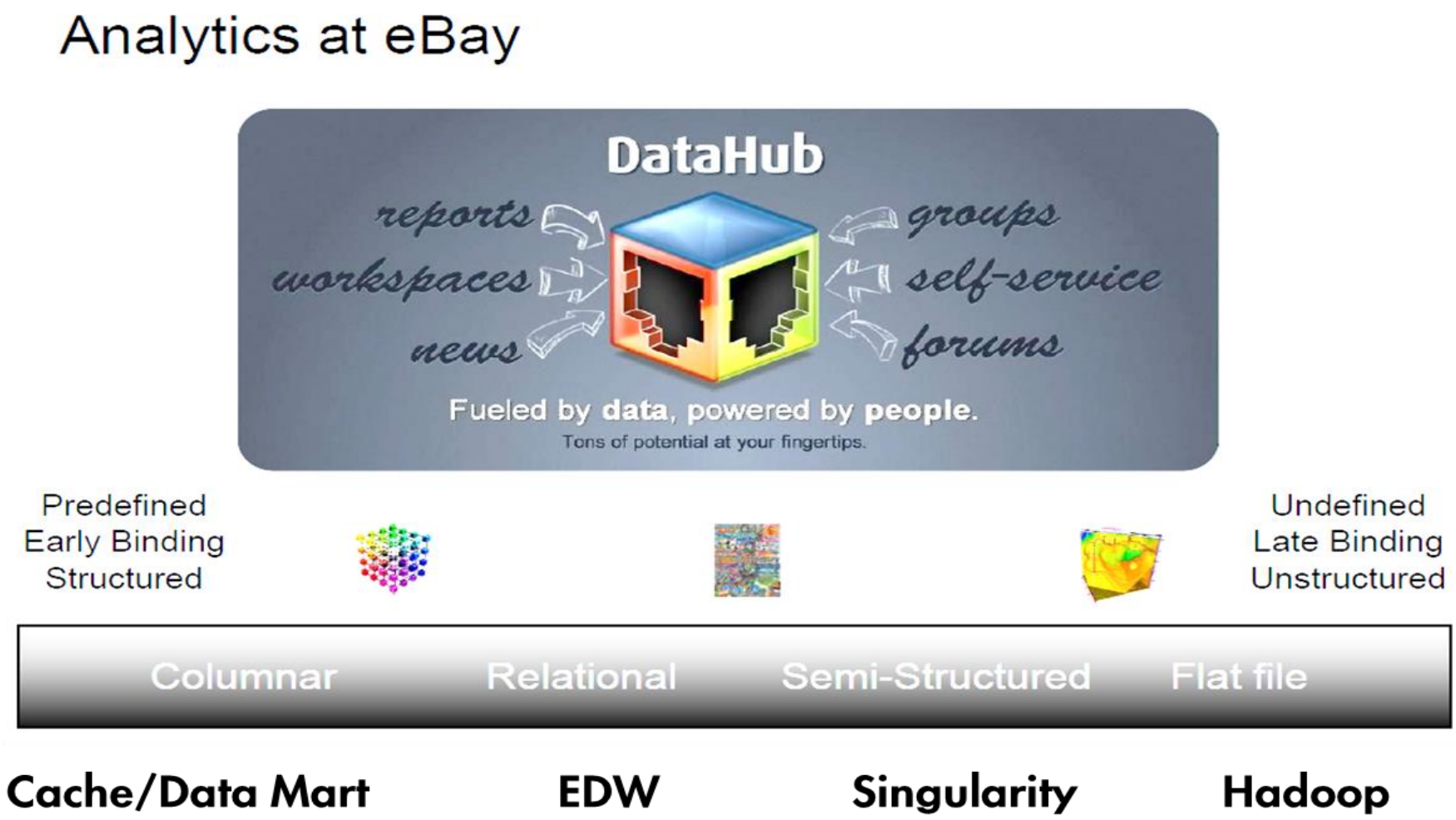
■ eBay의 UDA 사례



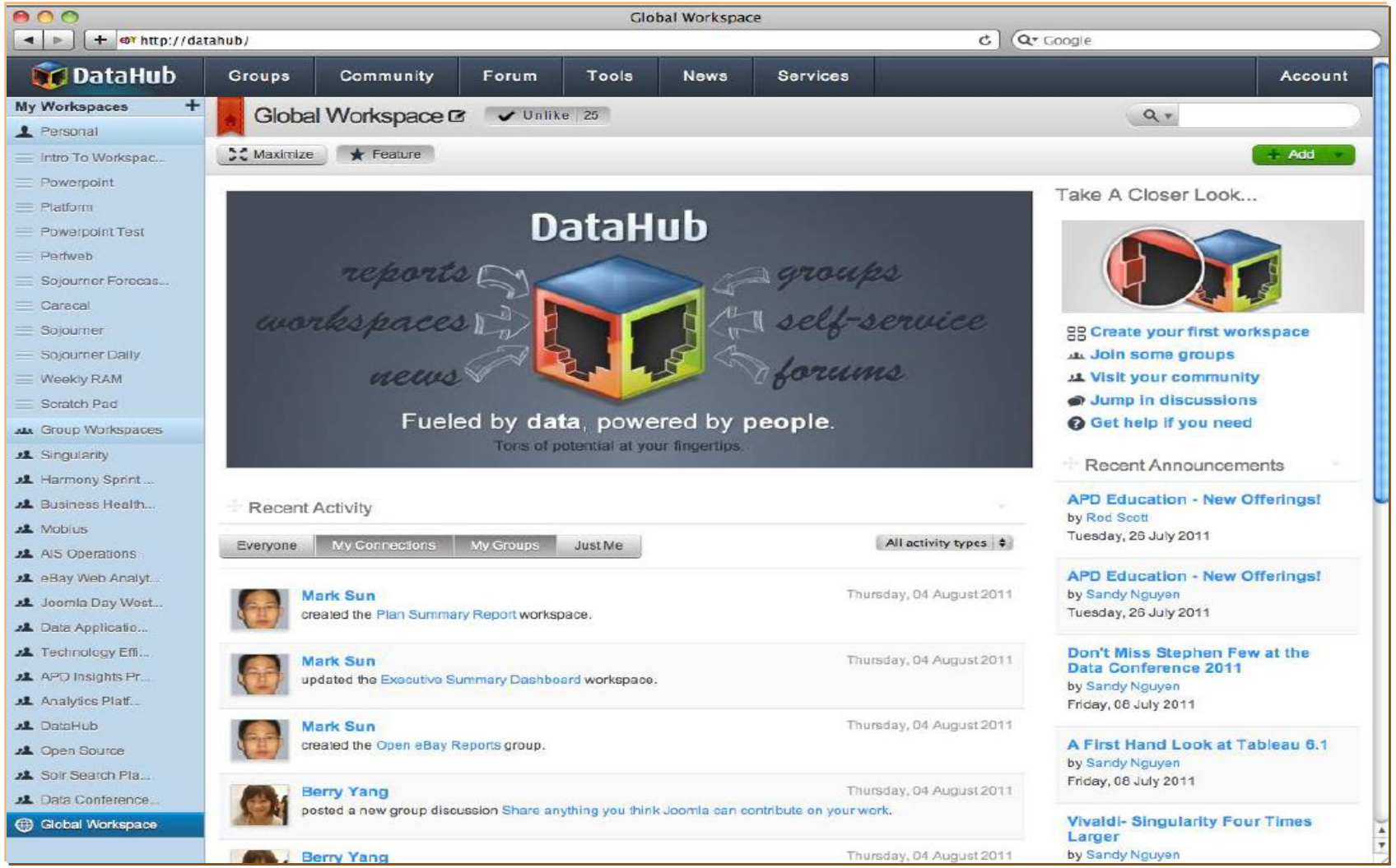
비즈니스 활용

- 고객 행동 추적 분석
- 상품 속성 분류
- 1:1 개인화 추천
- 마케팅 인사이트 도출
- 원가 및 재무 분석
- 시스템 보안 및 운영
- 신사업 개발

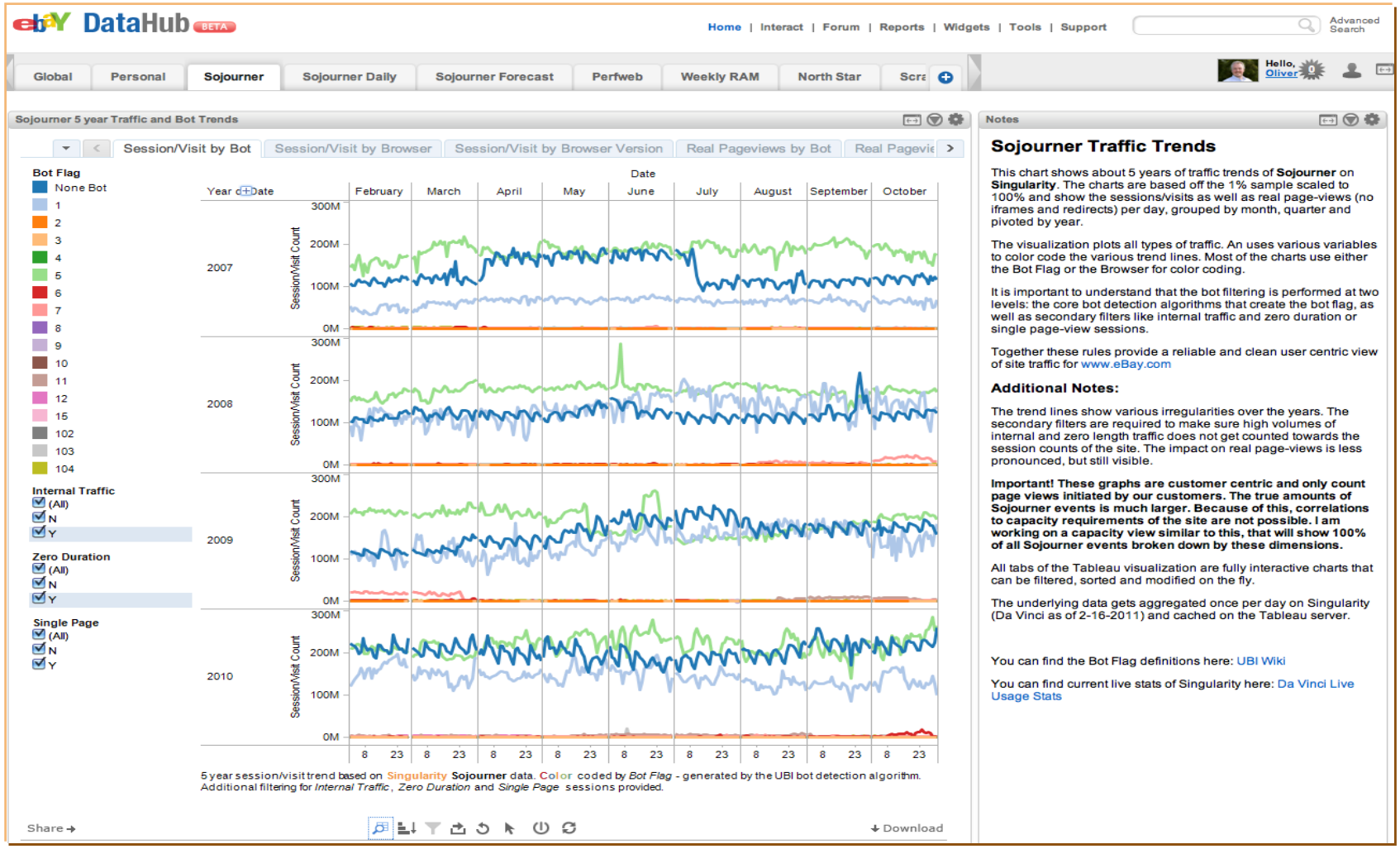
■ eBay의 UDA 사례



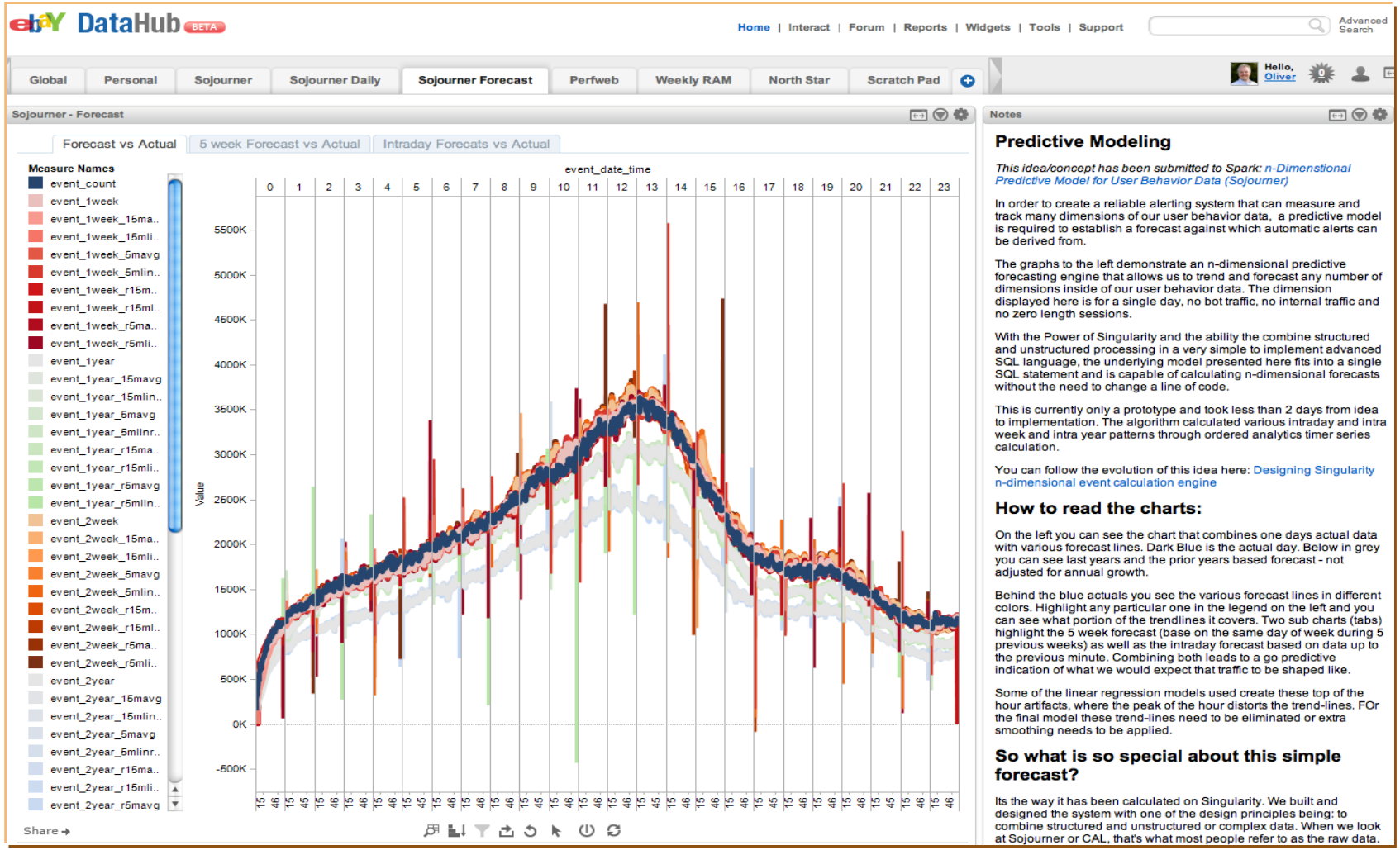
■ eBay의 UDA 사례


Global Workspace
http://datahub/
Google
DataHub
Groups
Community
Forum
Tools
News
Services
Account
My Workspaces
Personal
Intro To Workspac...
Powerpoint
Platform
Powerpoint Test
Perfweb
Sojourner Forecas...
Caracal
Sojourner
Sojourner Daily
Weekly RAM
Scratch Pad
Group Workspaces
Singularity
Harmony Sprint ...
Business Health...
Mobius
AIS Operations
eBay Web Analyt...
Joomla Day West...
Data Applicatio...
Technology Effi...
APD Insights Pr...
Analytics Platf...
DataHub
Open Source
Solr Search Pla...
Data Conference...
Global Workspace
Global Workspace
Unlike 25
Maximize
Feature
Add
DataHub
reports
workspaces
news
groups
self-service
forums
Fueled by data, powered by people.
Tons of potential at your fingertips.
Take A Closer Look...
Create your first workspace
Join some groups
Visit your community
Jump in discussions
Get help if you need
Recent Activity
Everyone
My Connections
My Groups
Just Me
All activity types
Mark Sun
created the Plan Summary Report workspace.
Thursday, 04 August 2011
Mark Sun
updated the Executive Summary Dashboard workspace.
Thursday, 04 August 2011
Mark Sun
created the Open eBay Reports group.
Thursday, 04 August 2011
Berry Yang
posted a new group discussion Share anything you think Joomla can contribute on your work.
Thursday, 04 August 2011
Berry Yang
Thursday, 04 August 2011
Recent Announcements
APD Education - New Offerings!
by Rod Scott
Tuesday, 26 July 2011
APD Education - New Offerings!
by Sandy Nguyen
Tuesday, 26 July 2011
Don't Miss Stephen Few at the Data Conference 2011
by Sandy Nguyen
Friday, 08 July 2011
A First Hand Look at Tableau 6.1
by Sandy Nguyen
Friday, 08 July 2011
Vivaldi- Singularity Four Times Larger
by Sandy Nguyen

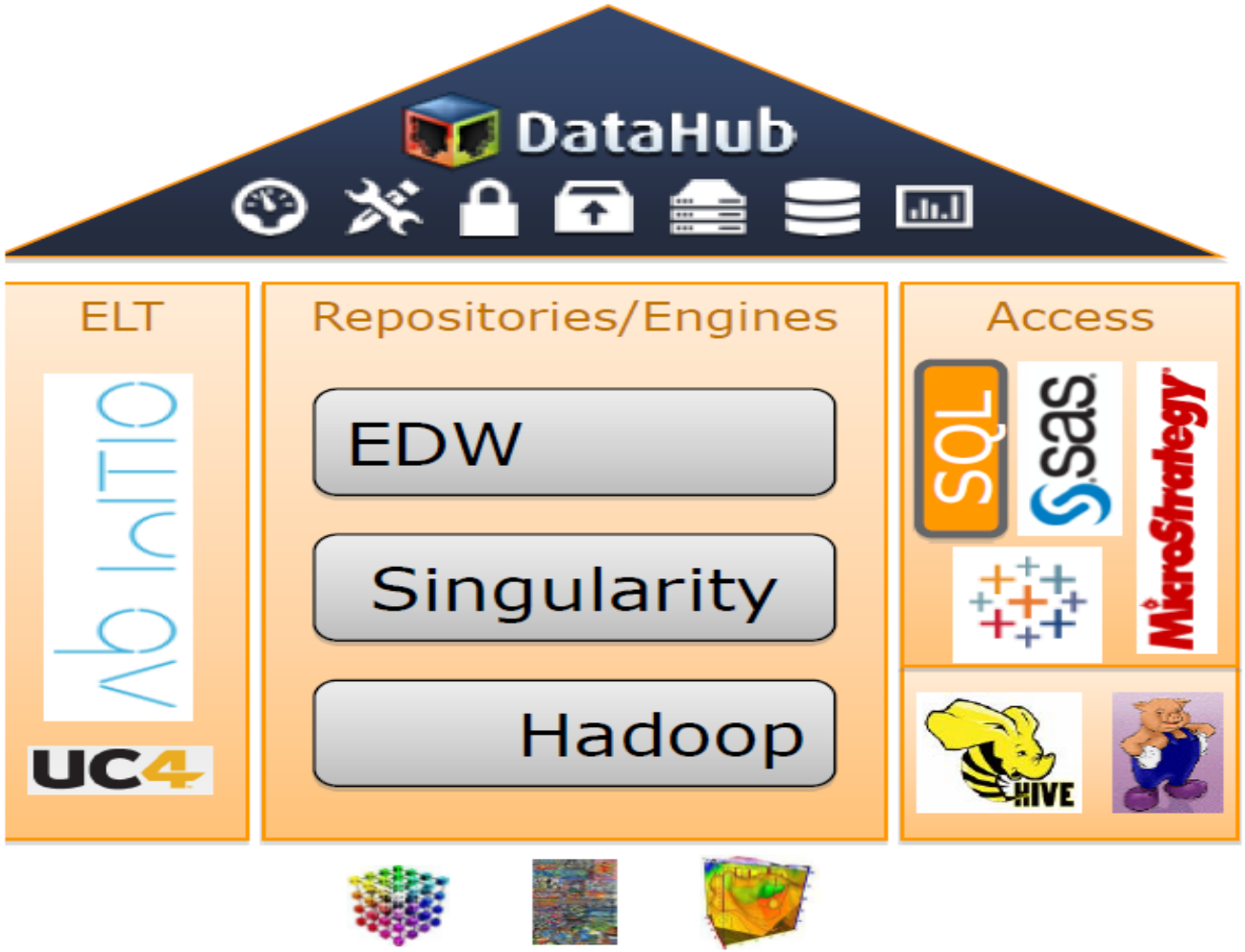
■ eBay의 UDA 사례



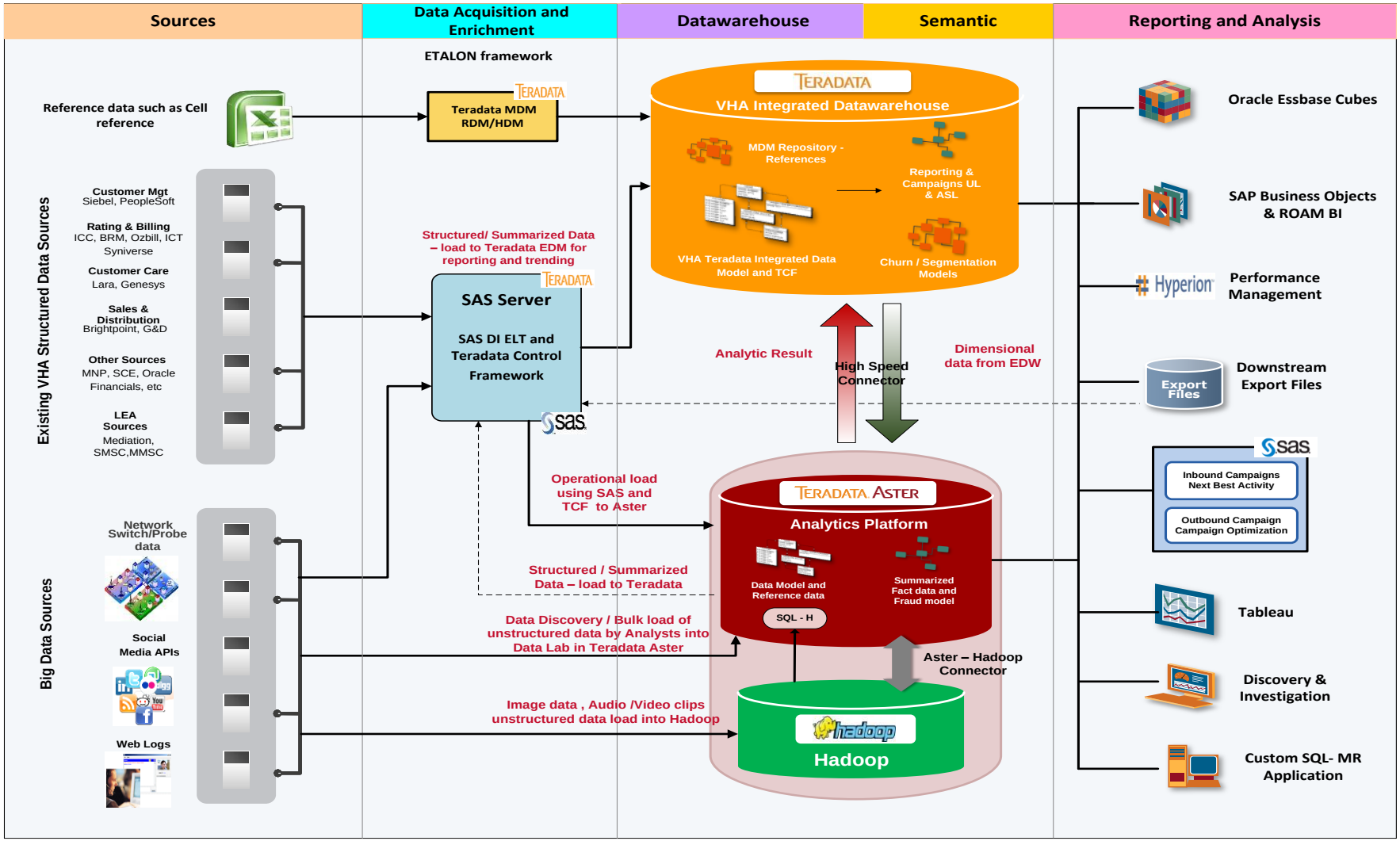
■ eBay의 UDA 사례


DataHub BETA
Home | Interact | Forum | Reports | Widgets | Tools | Support
Advanced Search
Global
Personal
Sojourner
Sojourner Daily
Sojourner Forecast
Perfweb
Weekly RAM
North Star
Scratch Pad
Hello, Oliver
Sojourner - Forecast
Forecast vs Actual
5 week Forecast vs Actual
Intraday Forecats vs Actual
Measure Names
event_count
event_1week
event_1week_15ma..
event_1week_15mli..
event_1week_5mavg
event_1week_5mlin..
event_1week_r15m..
event_1week_r15ml..
event_1week_r5ma..
event_1week_r5mli..
event_1year
event_1year_15mavg
event_1year_15mlin..
event_1year_5mavg
event_1year_5mlinr..
event_1year_r15ma..
event_1year_r15mli..
event_1year_r5mavg
event_1year_r5mlin..
event_2week
event_2week_15ma..
event_2week_15mli..
event_2week_5mavg
event_2week_5mlin..
event_2week_r15m..
event_2week_r15ml..
event_2week_r5ma..
event_2week_r5mli..
event_2year
event_2year_15mavg
event_2year_15mlin..
event_2year_5mavg
event_2year_5mlinr..
event_2year_r15ma..
event_2year_r15mli..
event_2year_r5mavg
event_date_time
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Value
5500K
5000K
4500K
4000K
3500K
3000K
2500K
2000K
1500K
1000K
500K
0K
-500K
Share
Notes
Predictive Modeling
This idea/concept has been submitted to Spark: n-Dimenstional Predictive Model for User Behavior Data (Sojourner)
In order to create a reliable alerting system that can measure and track many dimensions of our user behavior data, a predictive model is required to establish a forecast against which automatic alerts can be derived from.
The graphs to the left demonstrate an n-dimensional predictive forecasting engine that allows us to trend and forecast any number of dimensions inside of our user behavior data. The dimension displayed here is for a single day, no bot traffic, no internal traffic and no zero length sessions.
With the Power of Singularity and the ability the combine structured and unstructured processing in a very simple to implement advanced SQL language, the underlying model presented here fits into a single SQL statement and is capable of calculating n-dimensional forecasts without the need to change a line of code.
This is currently only a prototype and took less than 2 days from idea to implementation. The algorithm calculated various intraday and intra week and intra year patterns through ordered analytics timer series calculation.
You can follow the evolution of this idea here: Designing Singularity n-dimensional event calculation engine
How to read the charts:
On the left you can see the chart that combines one days actual data with various forecast lines. Dark Blue is the actual day. Below in grey you can see last years and the prior years based forecast - not adjusted for annual growth.
Behind the blue actuals you see the various forecast lines in different colors. Highlight any particular one in the legend on the left and you can see what portion of the trendlines it covers. Two sub charts (tabs) highlight the 5 week forecast (base on the same day of week during 5 previous weeks) as well as the intraday forecast based on data up to the previous minute. Combining both leads to a go predictive indication of what we would expect that traffic to be shaped like.
Some of the linear regression models used create these top of the hour artifacts, where the peak of the hour distorts the trend-lines. FOr the final model these trend-lines need to be eliminated or extra smoothing needs to be applied.
So what is so special about this simple forecast?
Its the way it has been calculated on Singularity. We built and designed the system with one of the design principles being: to combine structured and unstructured or complex data. When we look at Sojourner or CAL, that's what most people refer to as the raw data.

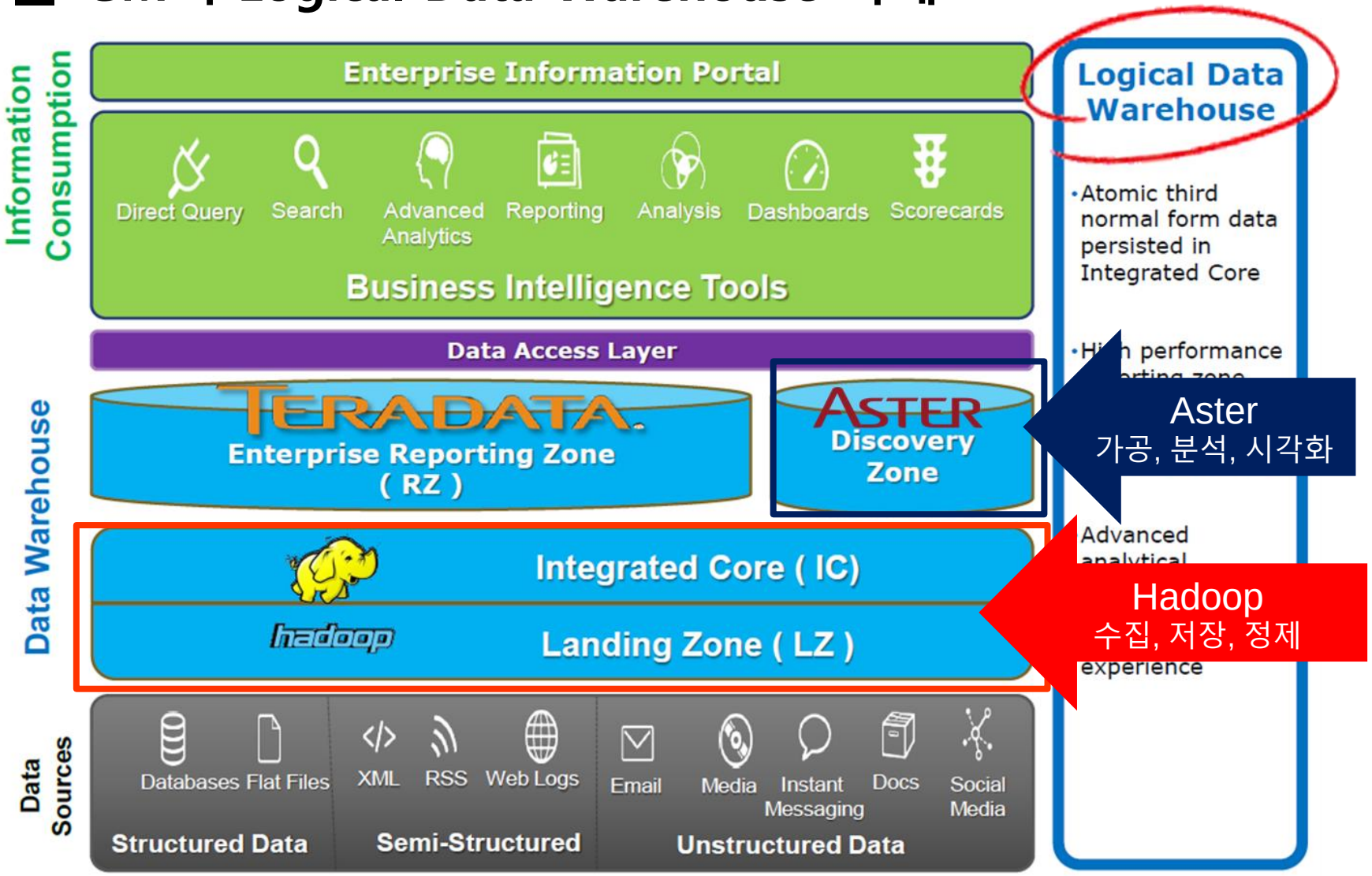
■ eBay의 UDA 사례


DataHub
ELT
Ab Initio
UC4
Repositories/Engines
EDW
Singularity
Hadoop
Access
SQL
sas
MicroStrategy
HIVE

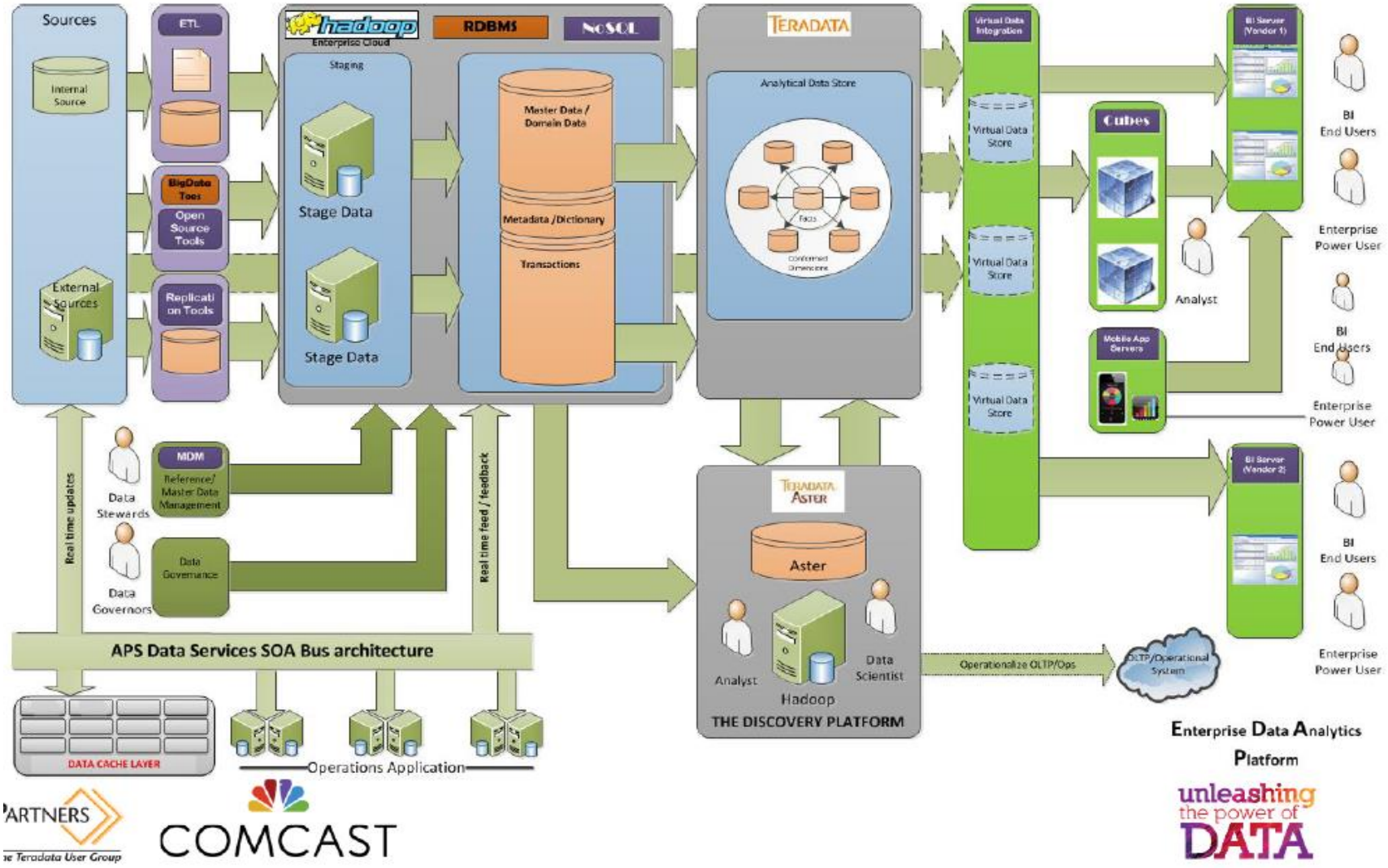
■ Vodafone의 UDA 사례



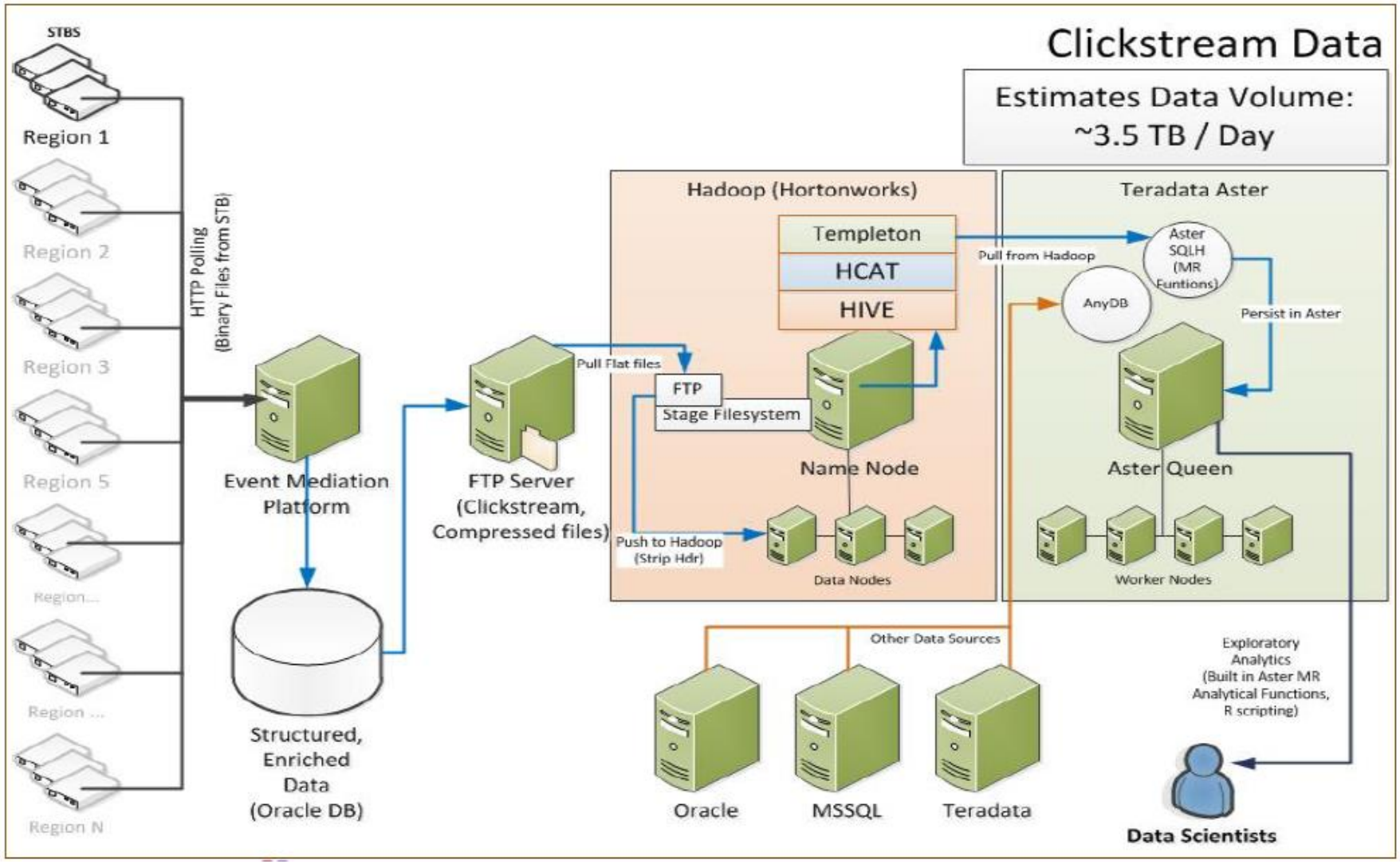
■ GM의 Logical Data Warehouse 사례



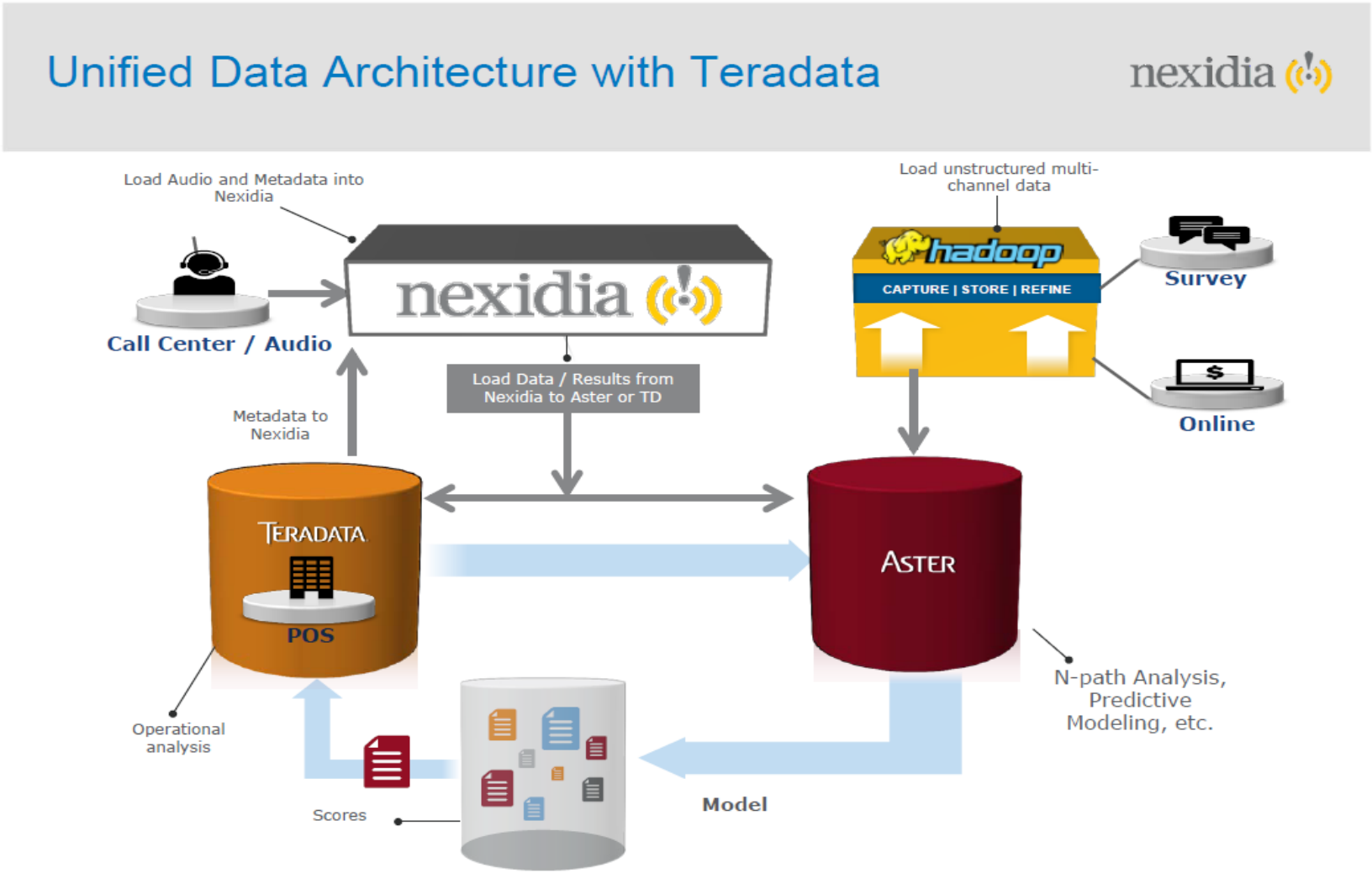
■ Comcast 사례


Sources
Internal Source
External Sources
ETL
BigData Tools
Open Source Tools
Replication Tools
hadoop
Enterprise Cloud
RDBMS
NoSQL
Staging
Stage Data
Stage Data
Master Data / Domain Data
Metadata /Dictionary
Transactions
TERADATA
Analytical Data Store
Facts
Conformed Dimensions
Virtual Data Integration
Virtual Data Store
Virtual Data Store
Virtual Data Store
Cubes
Mobile App Servers
Analyst
BI Server (Vendor 1)
BI Server (Vendor 2)
BI End Users
Enterprise Power User
BI End Users
Enterprise Power User
BI End Users
Enterprise Power User
Real time updates
Data Stewards
MDM
Reference/ Master Data Management
Data Governors
Data Governance
Real time feed / feedback
APS Data Services SOA Bus architecture
DATA CACHE LAYER
Operations Application
TERADATA ASTER
Aster
Analyst
Hadoop
Data Scientist
THE DISCOVERY PLATFORM
Operationalize OLTP/Ops
OLTP/Operational System
Enterprise Data Analytics Platform
unleashing the power of DATA
TERADATA
PARTNERS
the Teradata User Group
COMCAST

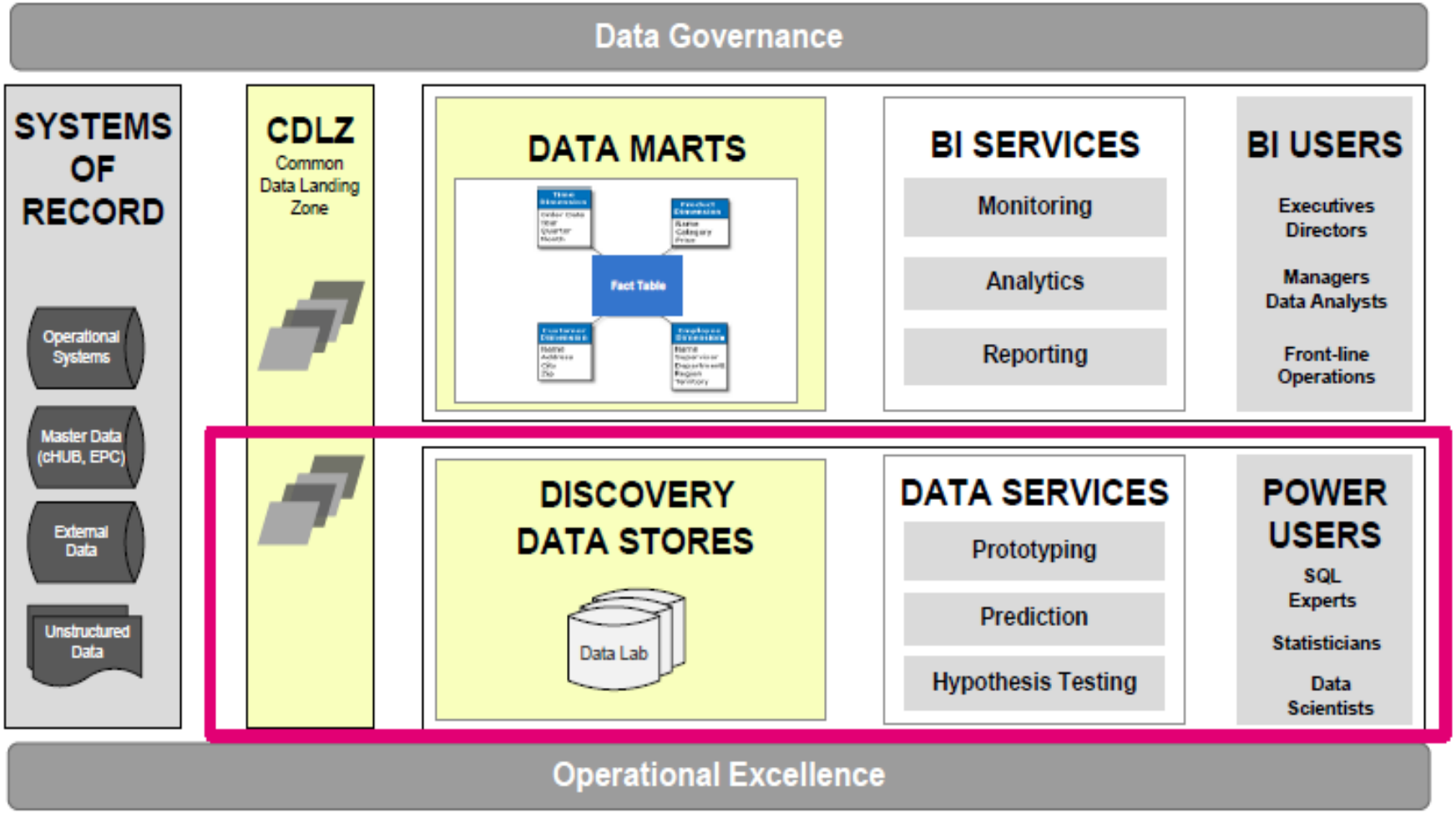
■ Comcast 사례



Comcast 사례

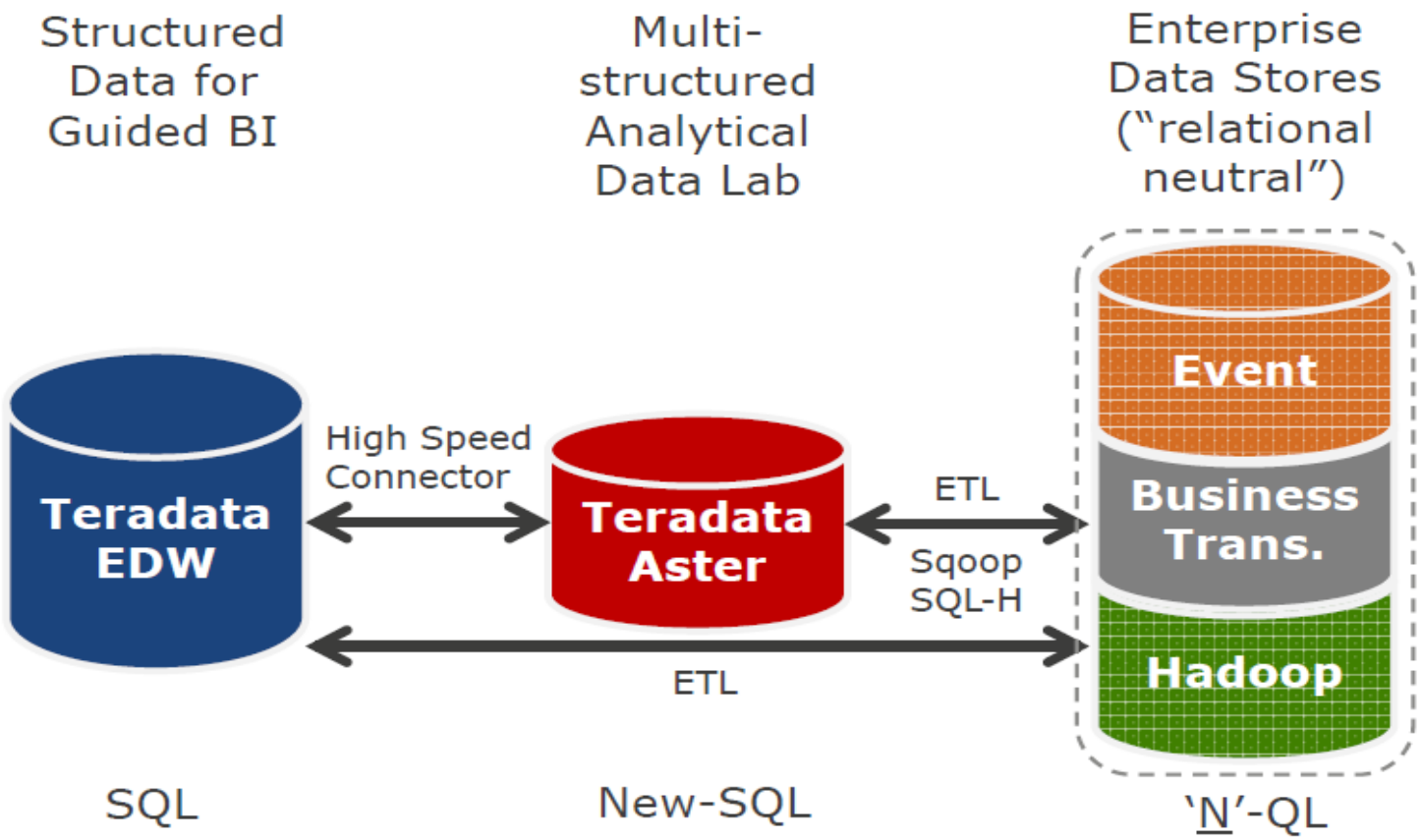

Unified Data Architecture with Teradata
nexidia
Load Audio and Metadata into Nexidia
Call Center / Audio
nexidia
Load unstructured multi-channel data
hadoop
CAPTURE | STORE | REFINE
Survey
Online
Load Data / Results from Nexidia to Aster or TD
Metadata to Nexidia
TERADATA
POS
ASTER
Operational analysis
Scores
Model
N-path Analysis, Predictive Modeling, etc.

■ T-mobile 사례

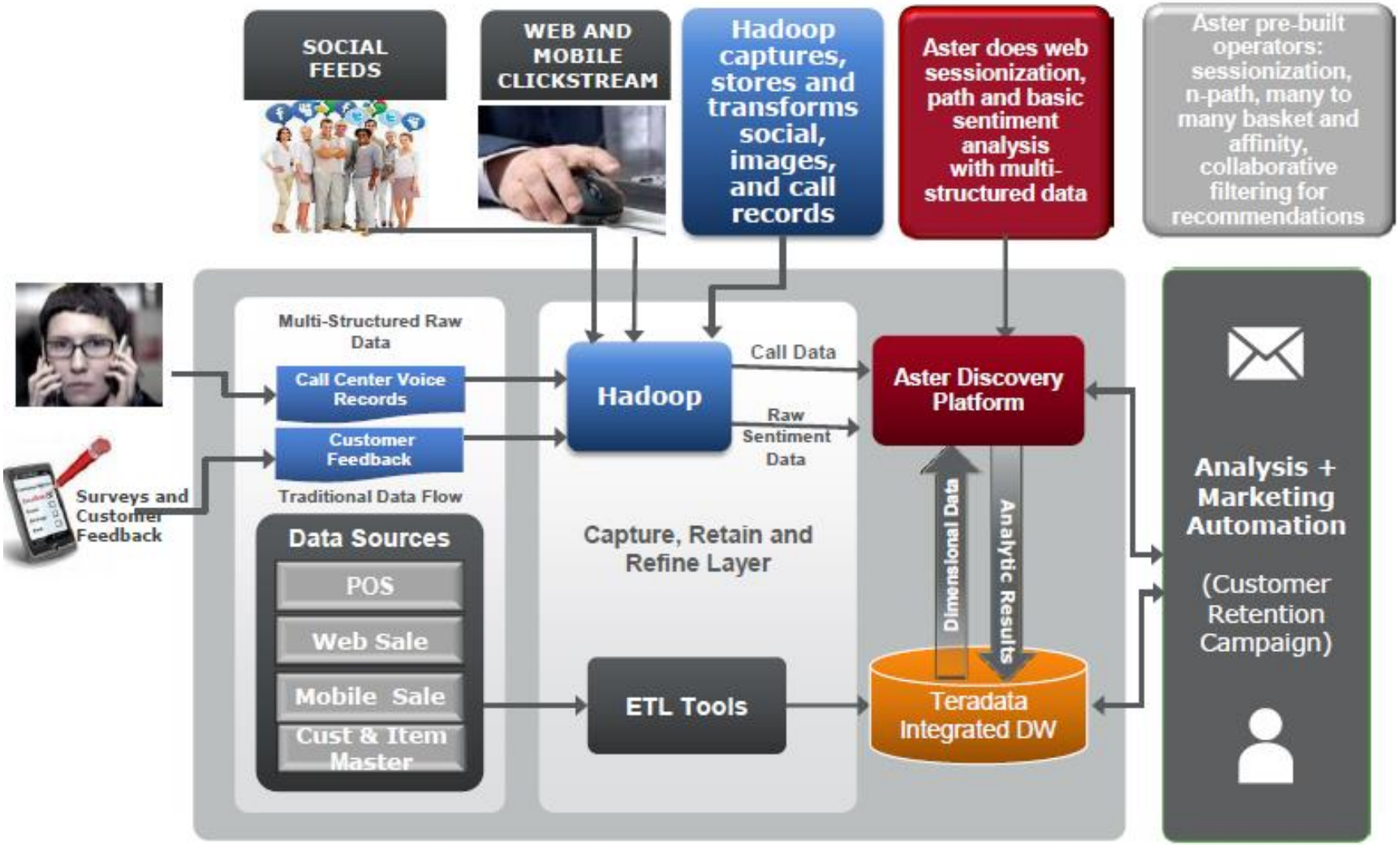


■ Discover Financial Service 사례

Next Generation of Discover DW/BI Data Environment



■ Wells Fargo 사례



TERADATA.

TERADATA.

장동인

010-5259-9509

Don.chang@teradata.com
donchang@hanmail.net

Facebook:
http://www.facebook.com/jang.cloud