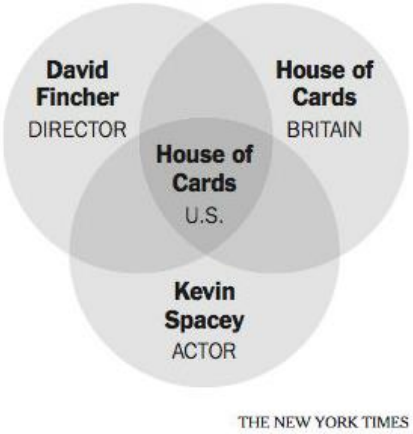
실사용자 중심의 빅데이터

2014.2.24

윤형기 hky@openwith.net

The Secret Sauce Behind Netflix's Hit, "House Of Cards": Big Data

BY ANALYZING ITS SUBSCRIBERS' PREFERENCES, NETFLIX CAN BE SURE ITS ORIGINAL CONTENT WILL FIND AN AUDIENCE. BUT IS THAT A GOOD THING?



목 차

- 도입
- 빅데이터 기술
  - 빅데이터 인프라 – Hadoop과 NoSQL
  - 분석기법 (Analytics) – 분석도구와 분석알고리즘
- 실사용자 중심의 빅데이터
  - 인프라 측면
  - 분석 측면
- 맺음말

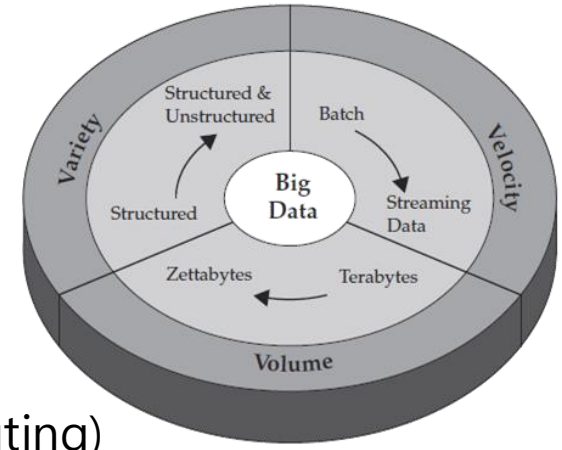
빅데이터 기술

- 배경
- 빅데이터 인프라
  - Hadoop
  - NoSQL
- 분석기법
  - 분석도구
  - 예측분석 알고리즘 (기계학습)

오늘날 IT 기술의 4인방

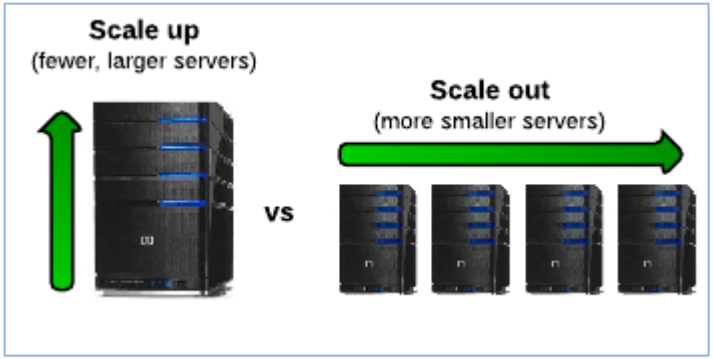
빅데이터의 배경

- Tidal Wave – 3VC



- Supercomputer
  - High-throughput computing
  - 2가지 방향:
    - 원격, 분산형 대규모 컴퓨팅 (grid computing)
    - 중앙집중형 (MPP)

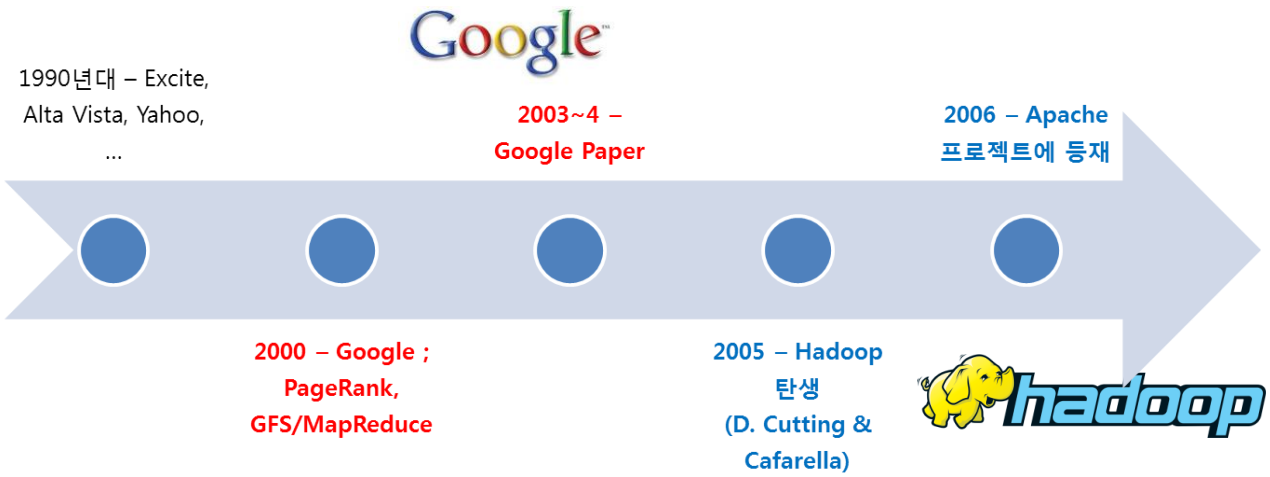
- Scale-Up vs. Scale-Out



- BI (Business Intelligence)
  - 특히 DW/OLAP/데이터 마이닝

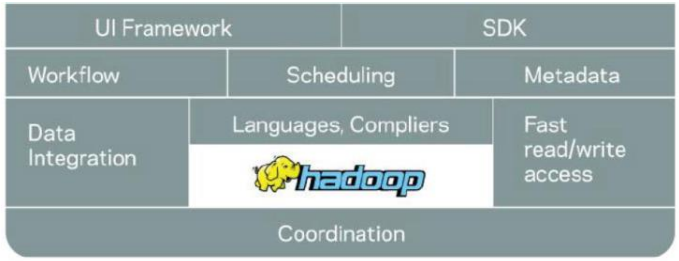
Hadoop

- Hadoop의 탄생?
  - 배경

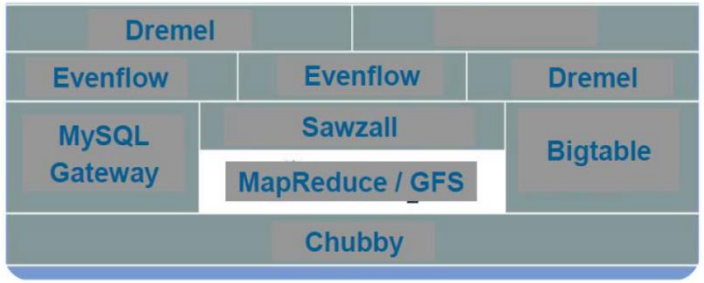


  - 특징
    - 대용량 데이터 분산처리 프레임워크
    - 프로그래밍 모델의 단순화로 선형 확장성 (Flat linearity)
      - "function-to-data model vs. data-to-function" (Locality)

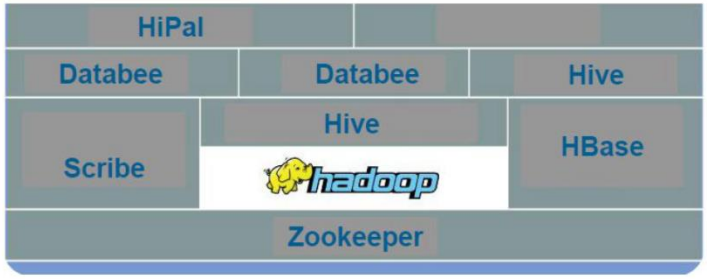
Frameworks



Google

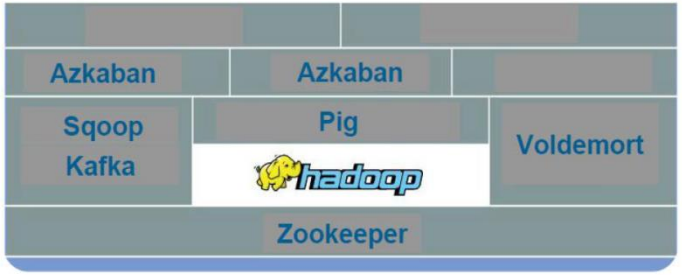


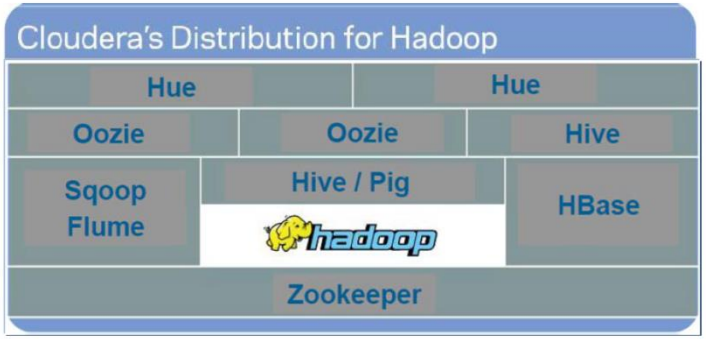
facebook



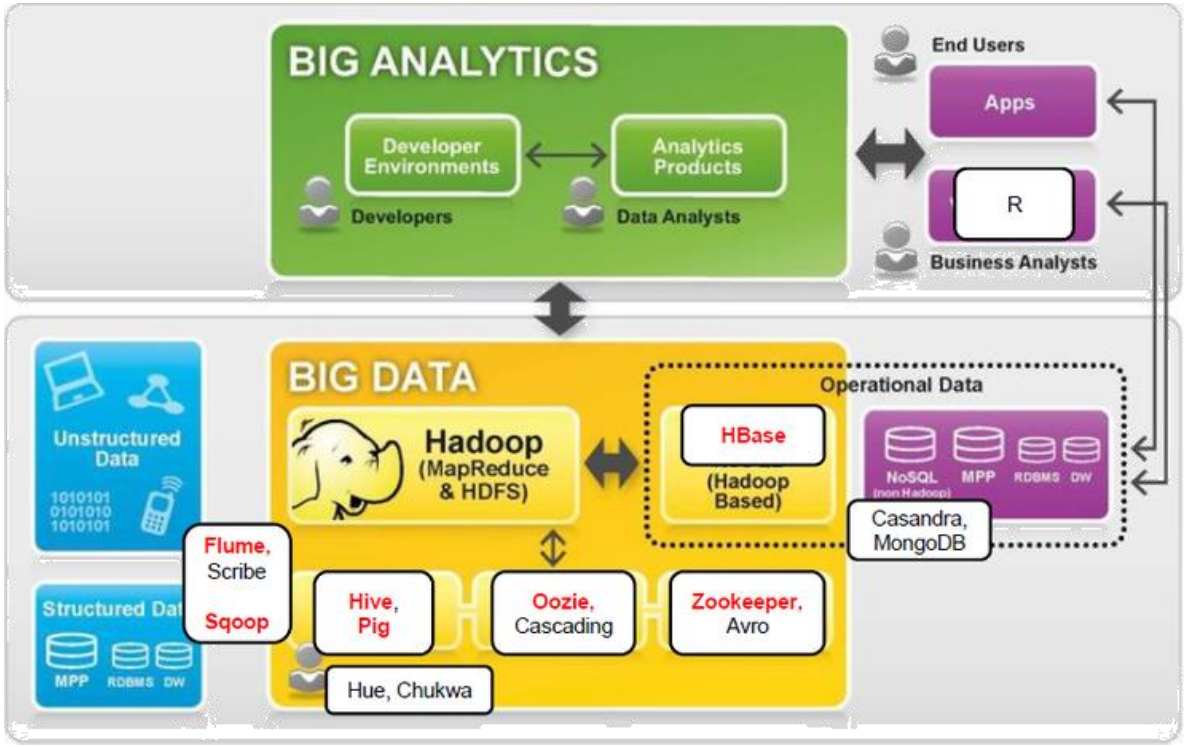
YAHOO!

LinkedIn



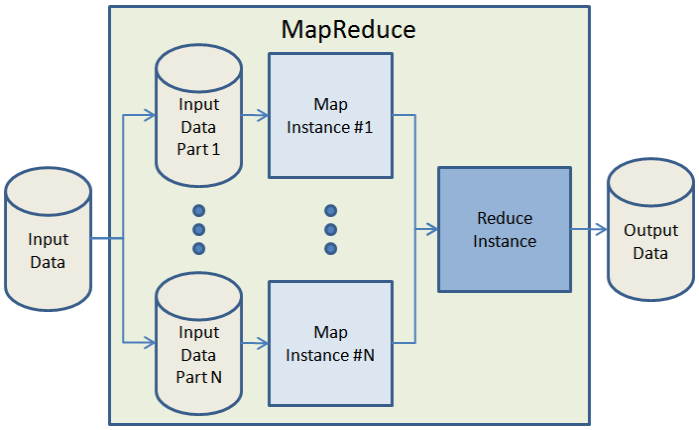


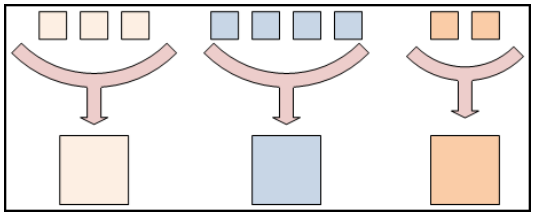
- Big Picture

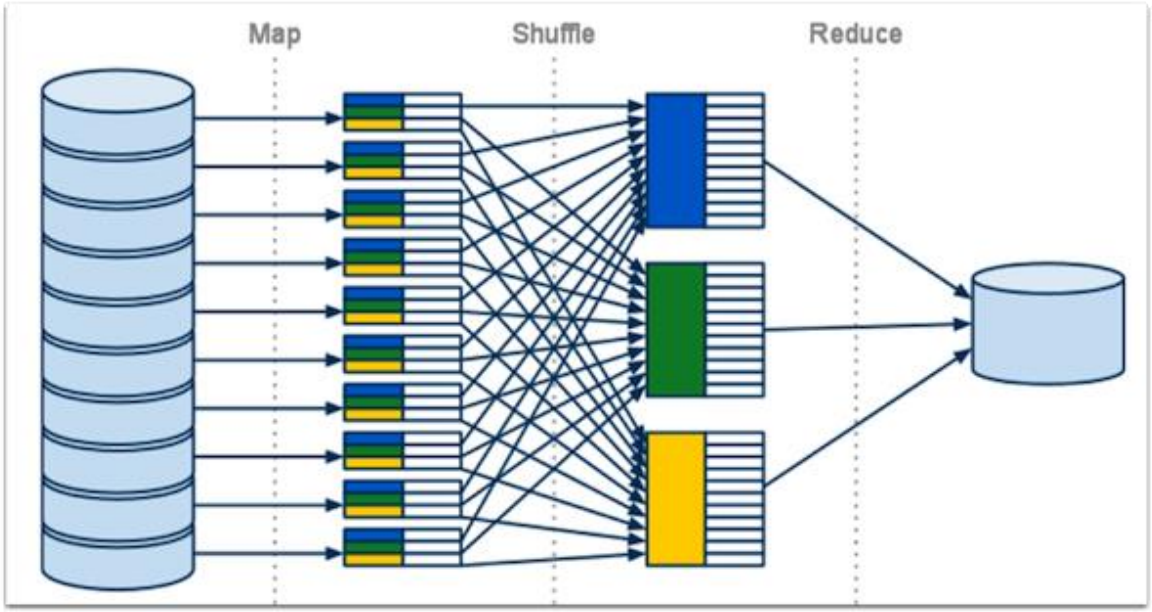


Store	Process
HDFS	Map/Reduce
Self-healing high-bandwidth clustered storage.	Fault-tolerant distributed processing.

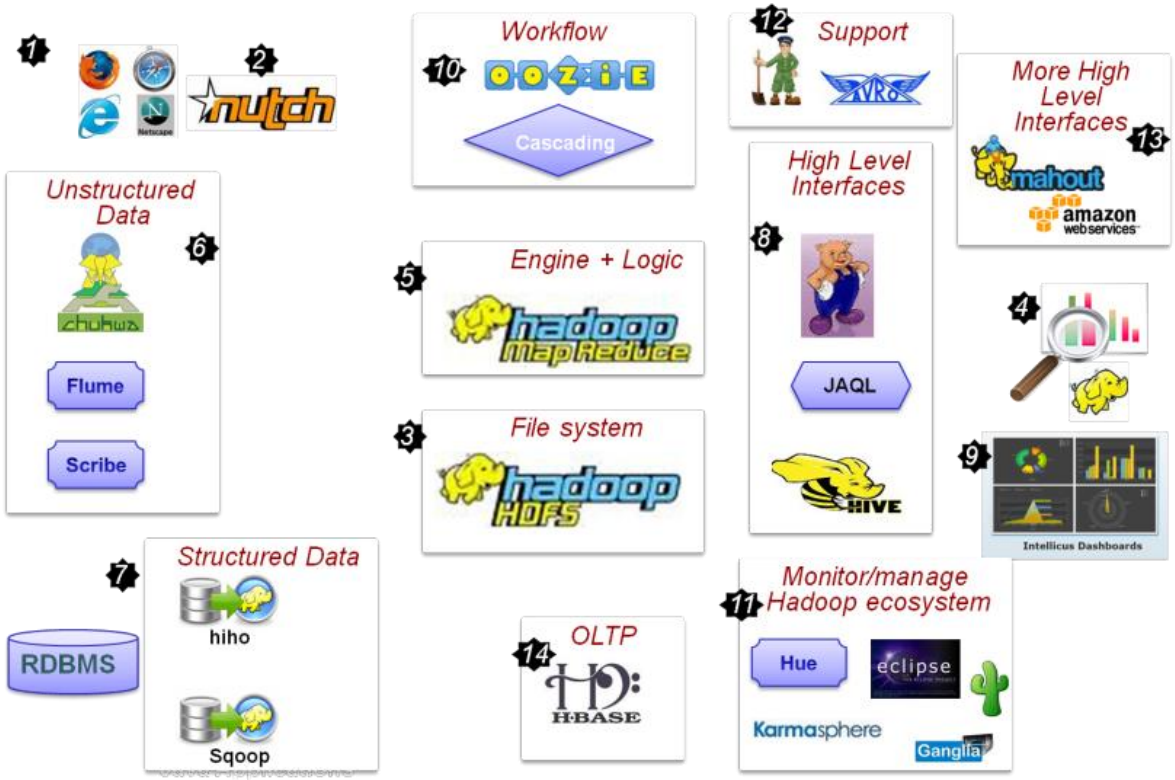
MapReduce


MapReduce
Input Data
Input Data Part 1
Input Data Part N
Map Instance #1
Map Instance #N
Reduce Instance
Output Data



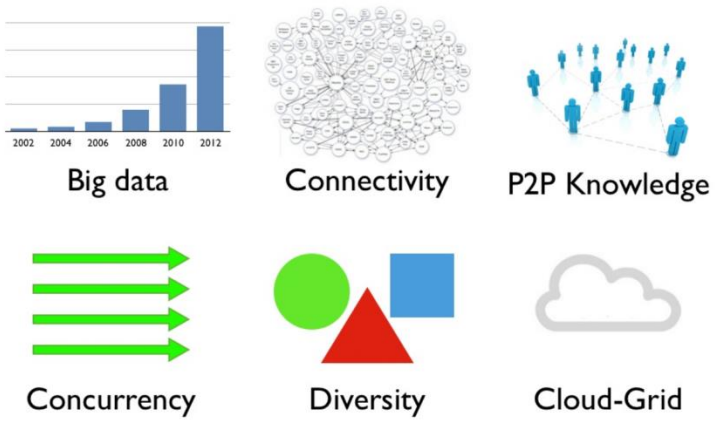

Map
Shuffle
Reduce

- Hadoop Ecosystem Map



NoSQL 데이터베이스

- 특징
  - 기존 RDB의 제한을 대폭 완화하여 단순화 → 성능향상, 유연화

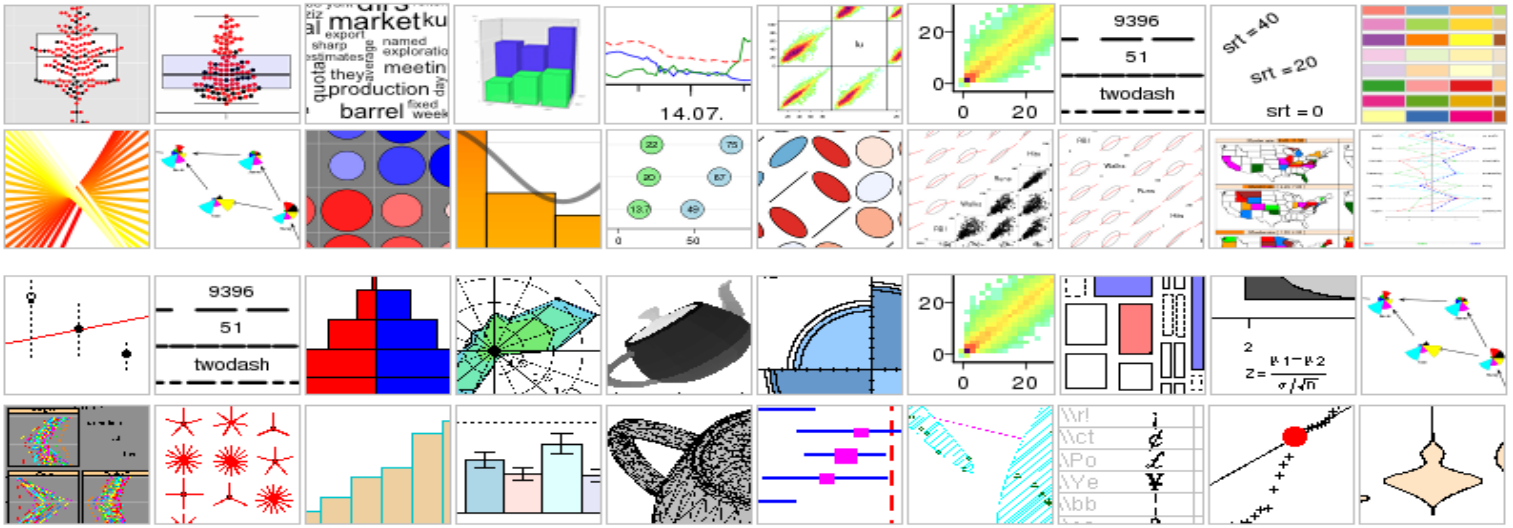


  - Key-Value Store, Document DB 등 다양한 종류
- 현황
  - 2014.1월 현재 150여 종

amazon web services™ HBASE Cassandra mongoDB Project Voldemort Google Bigtable CouchBase

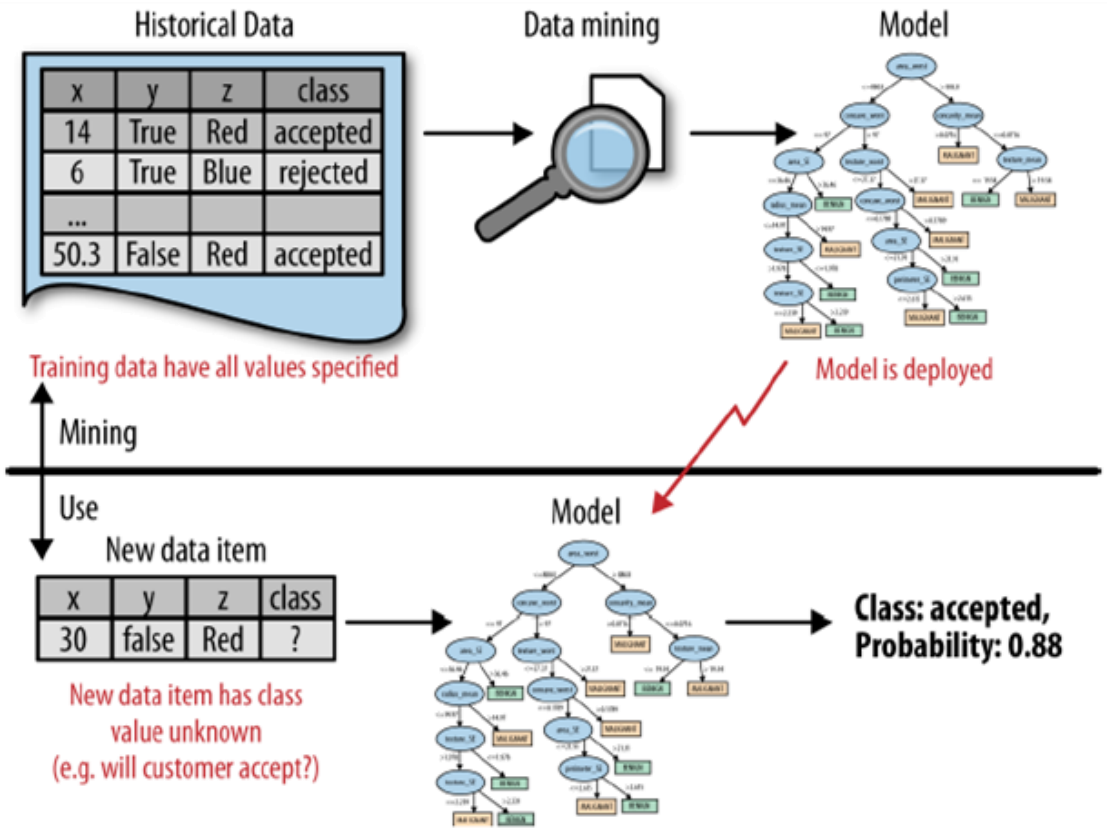
R

- open-source 수리/통계 분석도구 및 프로그래밍 언어
  - S 언어에서 기원하였으며 수 많은 package
    - CRAN: http://cran.r-project.org/
    - 현재 > 5,100 packages
  - 뛰어난 성능과 시각화 (visualization) 기능

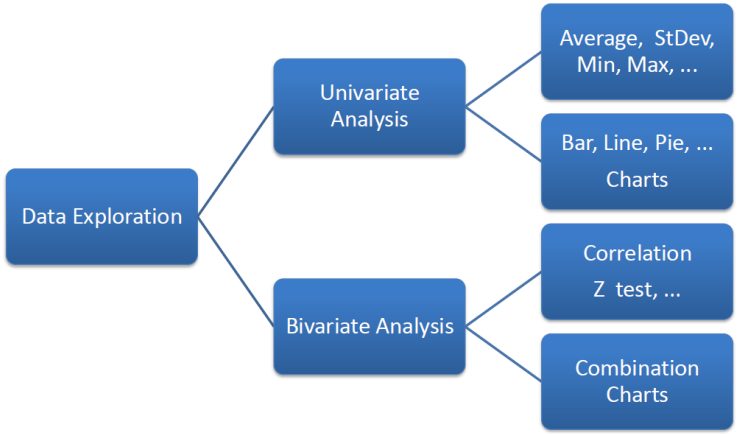


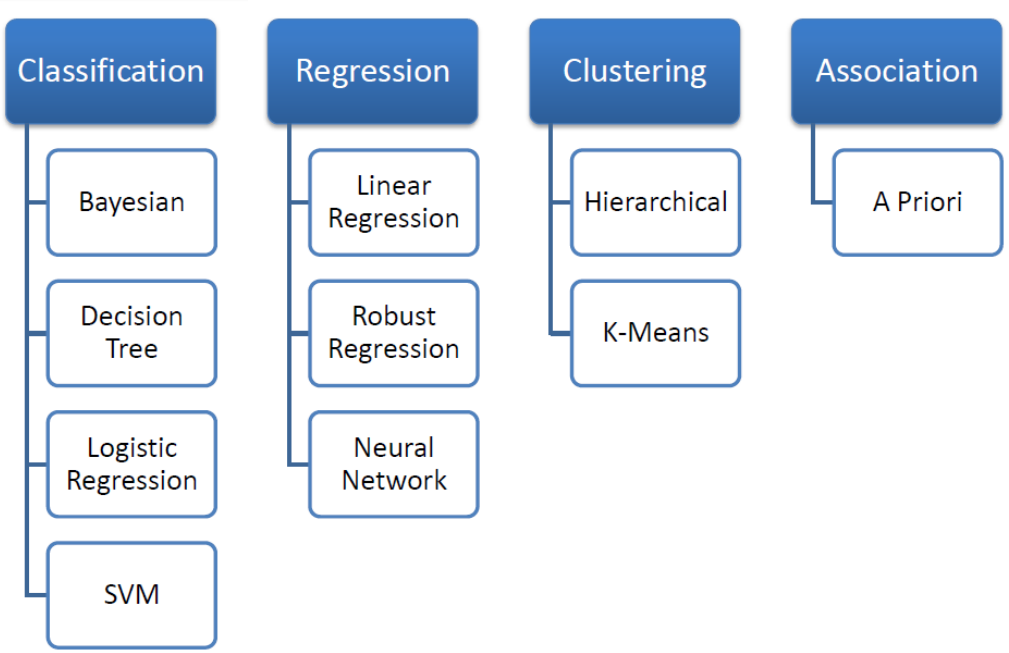
분석 기법

- 일반적인 기계학습 절차



- 분석 알고리즘
  - 탐색
  - 모델링





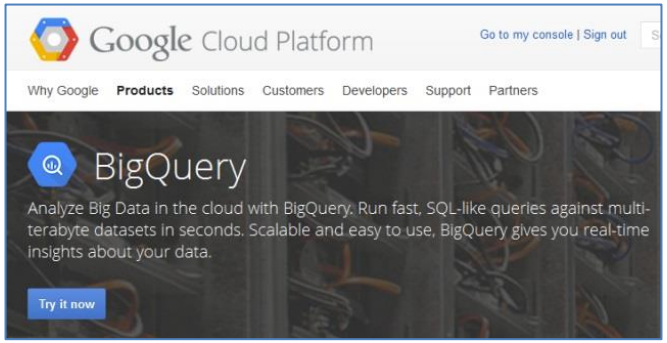
실사용자 중심의 빅데이터

- 인프라 측면
- 분석 측면
- 기타

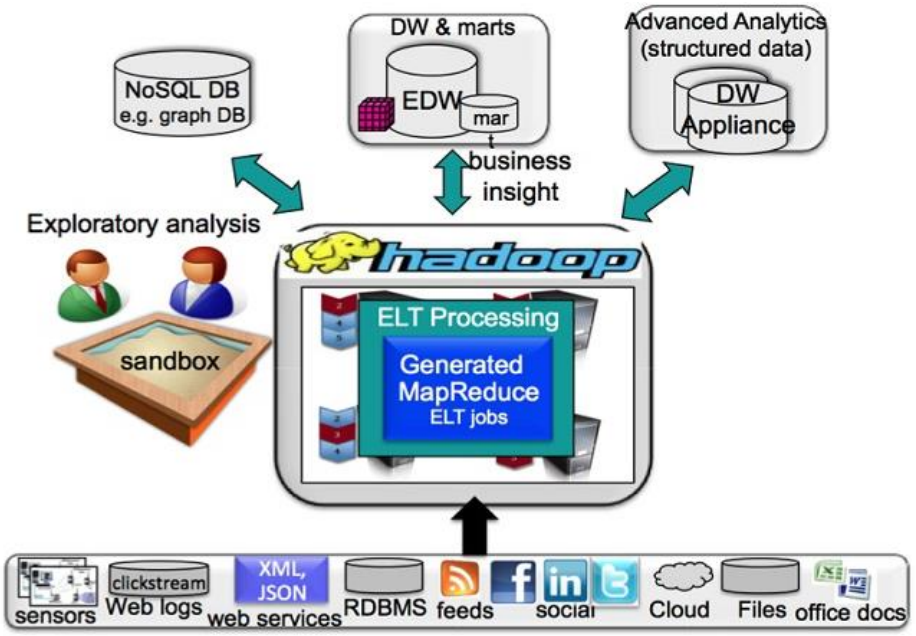
인프라 측면

- 빅데이터 어플라이언스 (Appliance)
  - Oracle Big Data Appliance
    - Sun 서버 + Cloudera 솔루션 (CDH) + NoSQL + ...
  - MarkLogic의 빅데이터 어플라이언스
    - SGI DataRaptor (HPC Appliance) + MarkLogic NoSQL DBMS
  - EMC의 Pivotal HD
    - = Hadoop 2.0 + Greenplum DBMS (SQL질의)
  - ASTER BIG Analytics Appliance
    - Aster DBMS + SQL-H (80여 SQL-MapReduce 함수) + Hortonworks HDP + 서버 + 관리tools + ...
  - ...

- 클라우드 컴퓨팅과 빅데이터
  - Cloud 서비스
    - Amazon EMR
    - 국내 클라우드 서비스
  - Big data as a service (BDaaS)
    - = 외부의 서비스 업체가 분석 도구 또는 분석결과를 제공
    - 일종의 managed services (SaaS와 흡사)
      - Cloud 스토리지와 연계되는 경향
  - Google의 BigQuery
    - Dremel, REST API
  - …



- Enterprise Data hub
  - Cloudera



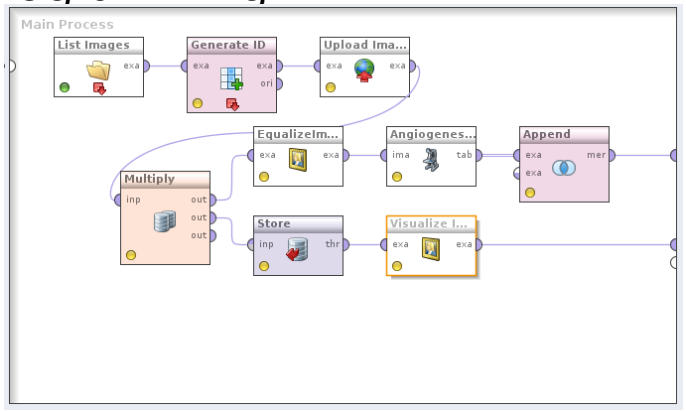
    - 그림출처: http://inside-bigdata.com/2013/12/02/enterprise-data-hubs/
  - 아직은… 그러나 2~3년 내에 …
    - Spark Streaming

분석측면

- 기능특화


Analytics Tools Landscape
Online web analytics
Social Media Analytics
Operation Research
Reporting & Visualization
Data Analytics
Support (OLAP)
Data management & Analytics
ETL Tools
Digital Ads Analytics
Email analytics
Analytics-Tools.com
OMNITURE
crazyegg
crimson hexagon
Brandwatch
sysomos
Coremetrics
CLICKY
lithium
responsys
radian6
Google Analytics
ILOG
LINDO
Yesmail
Evolver
ExactTarget
RISK Optimizer
GAMS
epsilon
FICO
dialog
X'CELSIUS
ValueClick
doubleclick
tableau
QlikView
atlas
Spotfire
eyeWONDER
mediamind
Elixir
sas
INFORMATICA
SPSS
CloverETL
Minitab
Centerprise
KNIME
Greenplum
SAP BusinessObjects
kxen
aster data
Hyperion
NETEZZA
splunk>
VERTICA
COGNOS
MicroStrategy

- 기타 – 새로운 모색
  - Ayasdi: Model-free, Query-free Insight Discovery
  - RapidMiner: Analytics with no programming



  - Dendrite3 –Lab41의 graph 분석 솔루션.
    - Titan (분산 그래프 DBMS) + GraphLab (graph analytics) + front-end 도구 (AngularJS)

맺음말

- 전략
  - 빅데이터 마인드
    - 비즈니스 모델적 사고
  - "자산으로서의 데이터"
    - Metadata관리의 중요성 – 데이터 품질
    - 조직의 데이터 공유/협업을 막는 문화와 관행
  - 기술의 내재화와 분석적 사고
    - 오픈소스 S/W 전략
    - 분석적 기업문화
- 기타
  - 복잡계 이론과 System Dynamics
  - 데이터 잔해 (Data Exhaust)
- "Big Data is All Data"