성공적인 데이터 비즈니스를 위한 오픈소스 기술 활용 전략

2013. 02

심탁길

Agenda

I. 빅데이터 사업 동향

II. 주요 오픈 소스 기술 소개

III. 오픈 소스 기술 활용 전략

IT 산업 전망

국내 외 데이터 비즈니스 동향

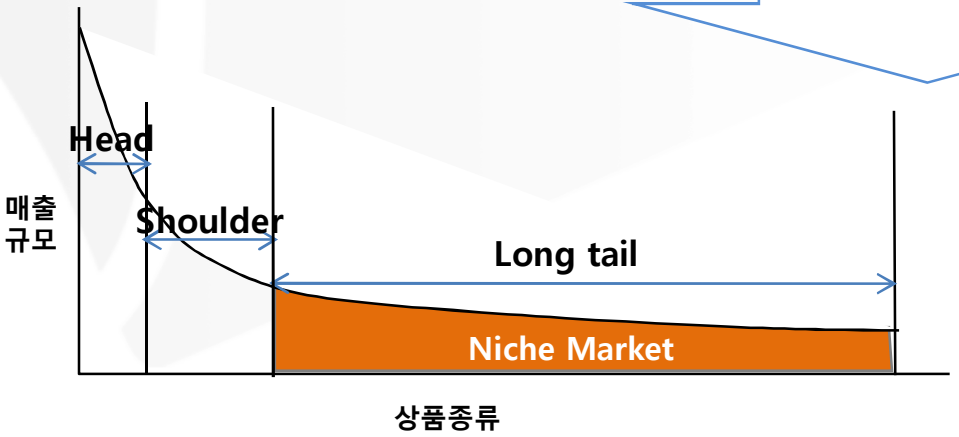
빅데이터 사업 유형

각 소비자 별 개인취향을 고려한 개인화 마케팅

[The Pareto principle]
상위 20% 상품이 전체 매출 중 80%를 차지하며,
베스트 셀러 상품위주로 판매전략 수립

온라인 시장의 발달로 소비자들은 획일화된 베스트 상품 구매보다 개인의 취향에 따라 상품을 검색/구매하는 개인화 구매 패턴을 보임

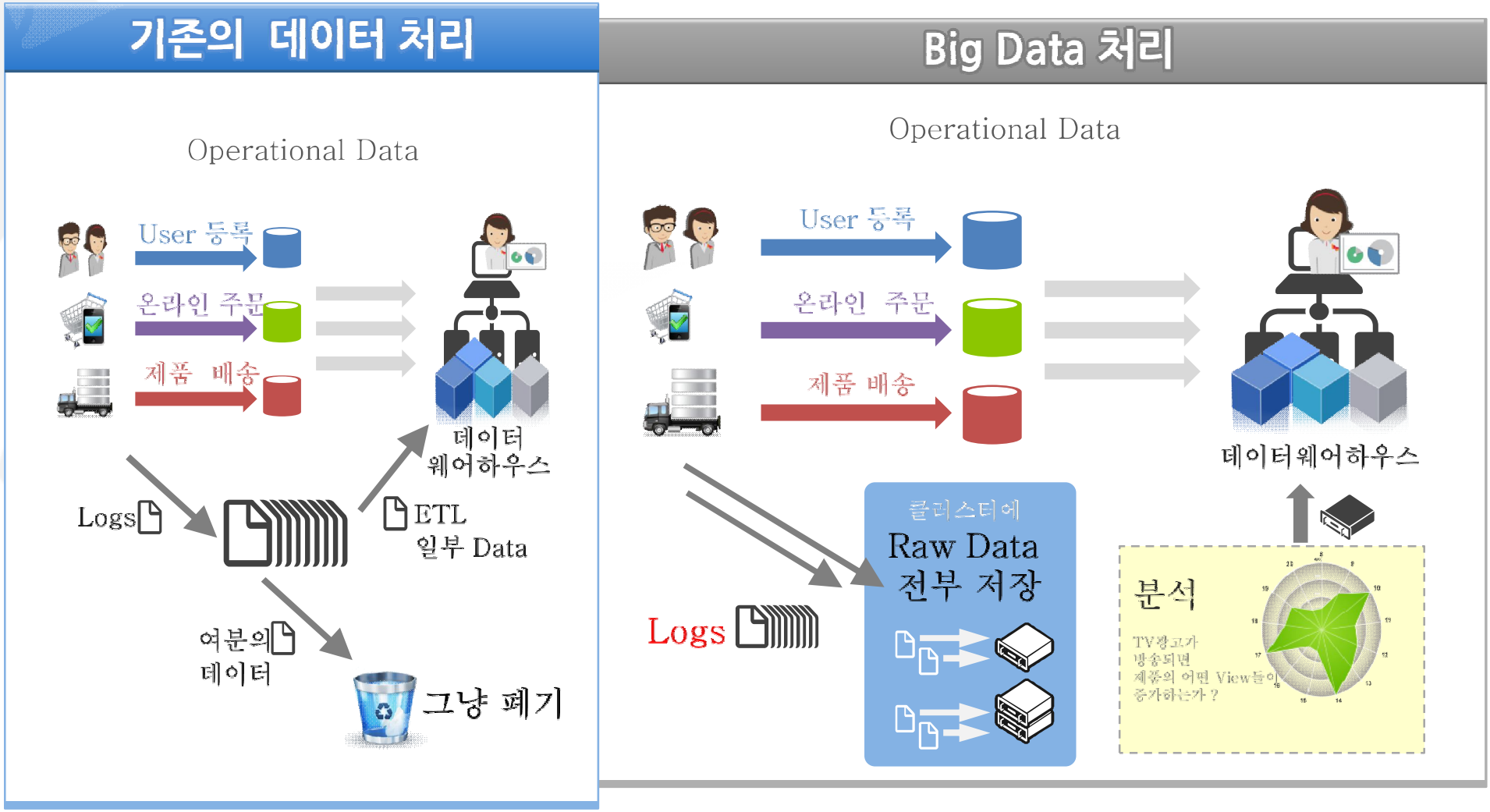
기업은 이러한 소비자 수요에 대응하기 위해 베스트 셀러 상품 개발 외 Long Tail 대상 Niche Market 마케팅에 집중



[The Long tail Principle]
과거 비즈니스 상에서 중요하게 평가되지 않았던
하위 80% 상품이 전체 매출의 50% 수준으로 확대

빅데이터 사업 유형

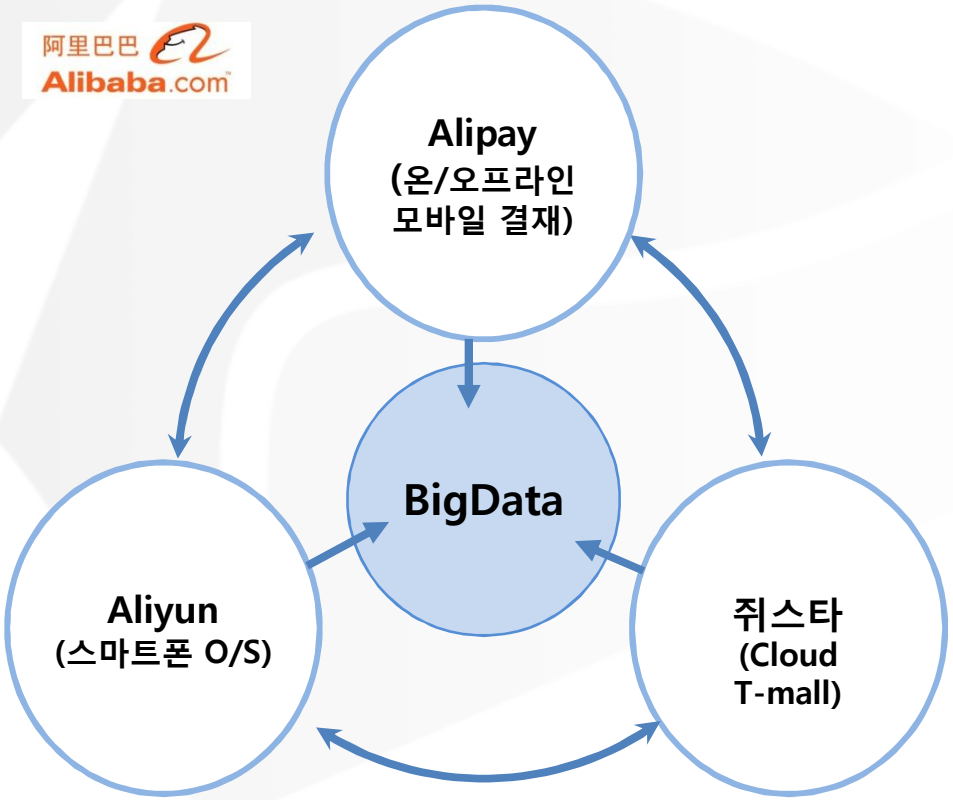
기업 내 원천 데이터의 전사적인 관리 및 다양한 비즈니스에 활용



해외 사례

알리바바는 '12년 매출 127조 규모 전세계 1위 e-Commerce 사업자이며, B2B/C2C/B2C 기반 온라인 상거래를 통해 발생되는 모든 데이터를 활용한 온라인 마케팅 플랫폼 확대 추진 중

개요



주요 특징

알리바바 그룹의 티엔마오(天猫), 타오바오(淘宝) 등 e-커머스 플랫폼과 연결되어 하루 50만 건 이상 결제 및 각종 자료 수집

소비자 연령, 지역, 구매 패턴 등의 데이터를 수집 및 분석하여 상거래 기업이 원하는 정보를 정확하게 전달(인프라, SW, API 제공 포함)

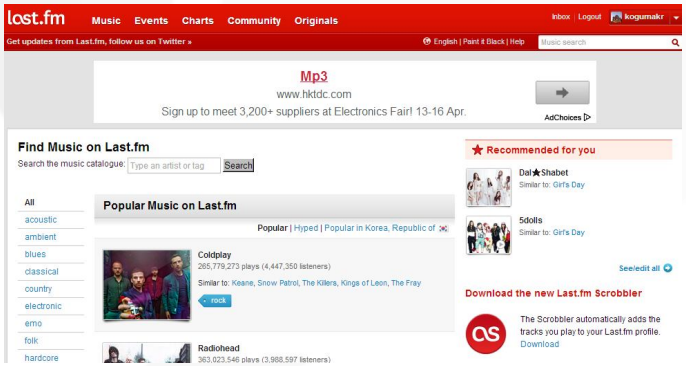
제공 받은 정보를 바탕으로 e-Commerce에서 필요한 제품 기획, 마케팅, 쿠폰/캠페인에 활용

Loyalty & Target Marketing을 통한 소비자 만족도 및 매출 증대

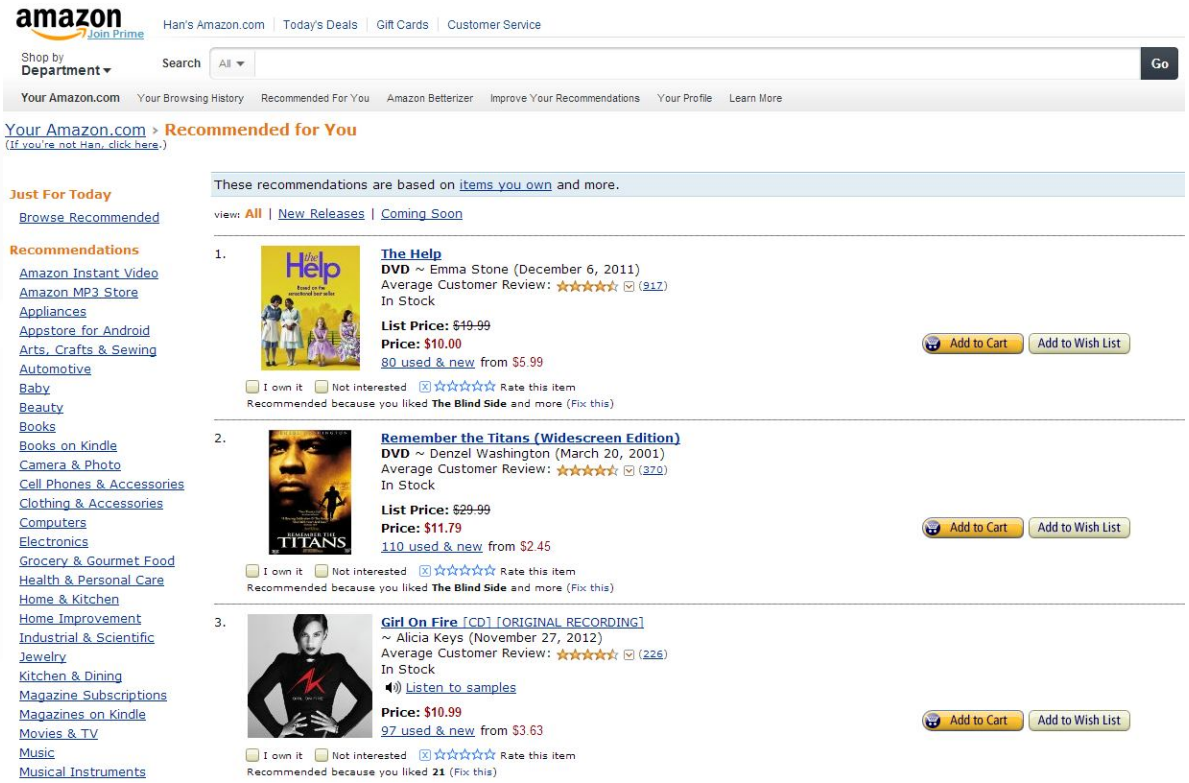
해외 사례

스트리밍 방식 주문형 비디오 추천

스트리밍 방식 음악추천



책, 음악, DVD, 의류 등 모든 상품에 적용



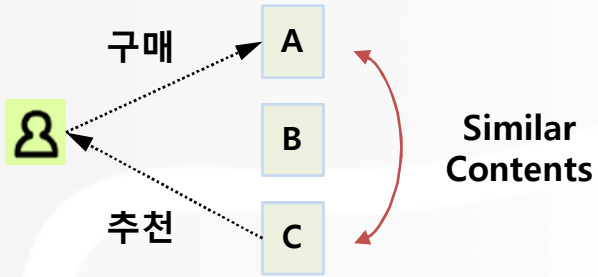
기존 방식

Content Based 분석

개요

- 회원의 과거구매내역과 유사한 다른 상품을 추천
- 아래 정보 기반으로 동작
  -도서 : 책의 제목, 저자, 분야, 도서평
  -영화 : 영화제목, 주연배우, 장르, 영화소개

개념도



적용 사례

- 박지성/이영표/축구를 좋아하는 사람에게
- 맨체스터/리버풀을 추천

특징

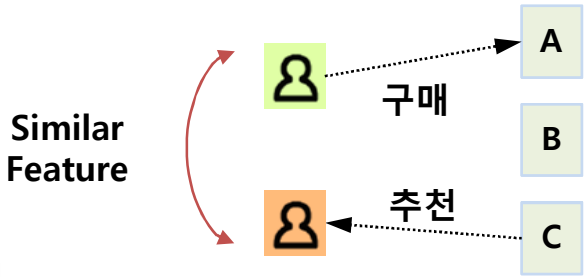
- 동일 Category 내에서만 추천 가능
  - 축구 콘텐츠 정보가 야구에서는 적용 불가
- 회원 속성이 Contents이용과 무관할 경우 존재
  - 박지성을 좋아하면 반드시 맨체스터를 좋아할까?

Demographic 분석

개요

- 회원 일반적인 정보(나이, 성별, 주소 등)를 인구통계학적으로 분석하여 상품 추천
  - IF age > 20 and age < 30 then Ballad
  - IF age < 20 then dance

개념도



적용 사례

- 댄스음악을 좋아하는 20대에게
- 댄스음악 베스트 앨범(소녀시대, 카라)를 추천

특징

- 회원정보가 불충분/부정확할 때 정확도가 떨어짐
- 회원 속성이 Contents이용과 무관할 경우 존재
  - 20대라면 반드시 댄스음악을 좋아할까 ?

개인화 알고리즘 방식

기존방식의 한계

- 최근 소비자들은 개인별 성향이 점차 강해짐에 따라 소비패턴분석이 성별/나이/거주지/직업 등으로 정보만으로 Categorize되기 어려움
- 따라서 기존 통계분석 방식으로는 광범위한 상품대상으로 소비자가 좋아할 만한 상품을 추천하기 어려움

- 강남 거주 20대 직장인 여성 싱글은 '네일아트'에 관심이 많을 것이다 ?

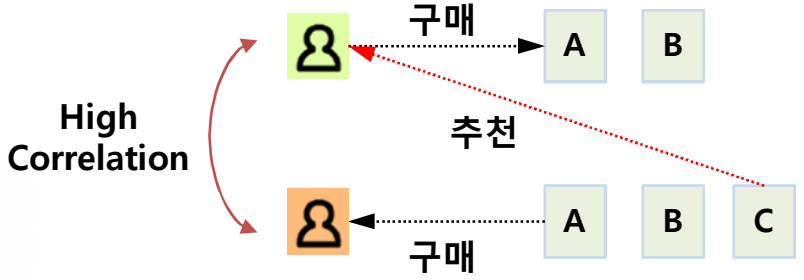
- 강북 거주 40대 주부는 유기농 초콜릿 비바니를 좋아하지 않을 것이다 ?

Collaborative Filtering (개인화 알고리즘)

개요

- 회원 프로파일 보다는 과거 구매정보만을 이용, 회원과 유사한 취향을 갖는 이웃 회원들을 선별하고 이들이 공통적으로 선호하는 상품을 추천

개념도



적용 사례

- 아마존 도서/음악/의류 추천에 CF 알고리즘 적용

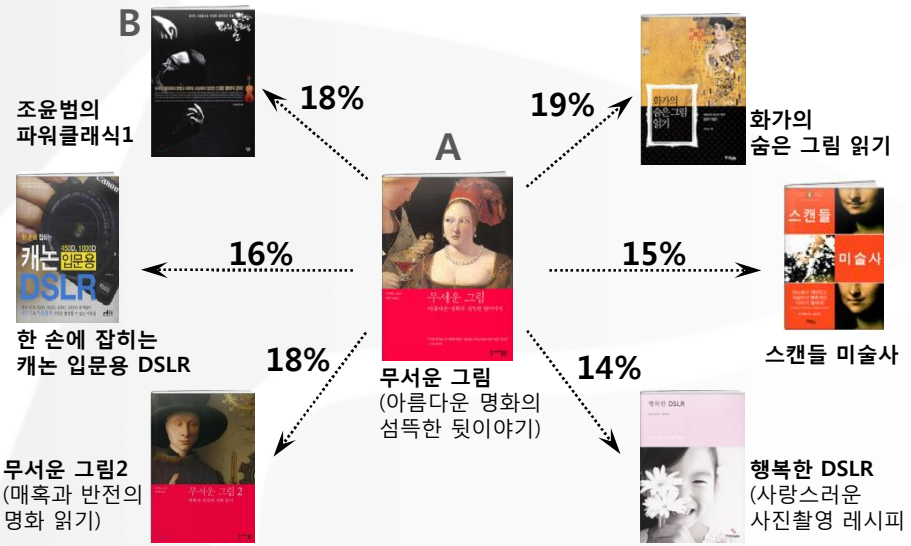
특징

- 회원들의 과거 구매를 가지고 정확도가 높은 추천
- 콘텐츠 내용, 사용자의 인구통계학적 정보 불필요
- 회원/상품수가 많아짐에 따라 Simulation 시간 증가
  - Big Data 기술을 통해 신속한 분석 제공
- 과거 구매이력이 적은 환경에서 적절한 추천이 어려움
  - 회원 수 구매이력이 많을수록 정확도는 높아짐

개인화 알고리즘 사례: 추천 시스템

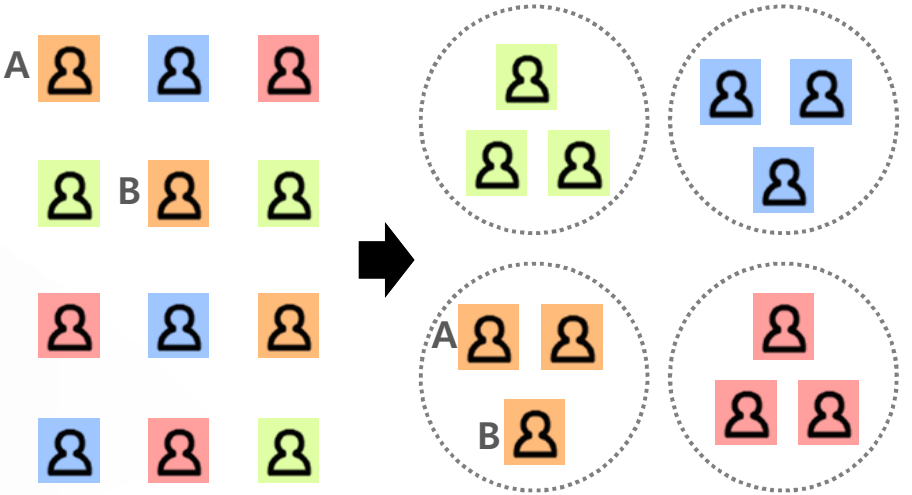
추천 시스템은 상품들간의 판매 상관관계를 파악하기 위한 상품 Network 분석과 구매성향이 비슷한 고객들을 그룹화하는 Clustering기법을 함께 사용하여 각 개인에게 가장 적합한 상품을 추천함

상품 Network 구성(도서 예시)



상품 간 관계 분석	▪특정 기간 동안 판매된 도서의 상관관계를 나타낸 상품 Network 구성
비고	▪책이 팔릴 때마다 상관관계가 변동되어 적절한 주기로 상품 Network를 업데이트 ▪가중치 적용(구매시기/판매수량/구매자수)

고객군 Clustering



고객 군 분석	▪총 상품 구매자들을 대상으로 취향이 비슷한 사람들을 묶어 클러스터 100개 구성
비고	▪정확도를 최대한 높일 수 있는 적절한 클러스터 개수를 정해야 함 ▪고객군을 나누는 적당한 경계를 찾아야 함

Agenda

I. 빅데이터 사업 동향

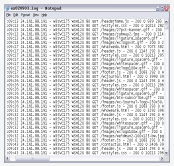
II. 주요 오픈 소스 기술 소개

III. 오픈 소스 기술 활용 전략

데이터 처리 흐름

데이터 소스 → 수집 → 저장 → 분석 → 표현

내부데이터



외부데이터

수집

- 로그 수집기
- 데이터 Integration
- 웹로봇
- RSS Feed
- Open API

저장

분산스토리지

배치처리

검색

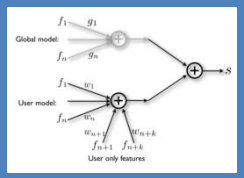
실시간&배치

NoSQL

관계형DB

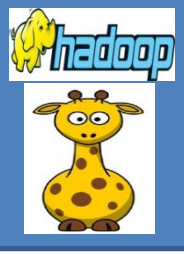
분석

분석알고리즘

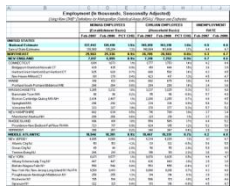


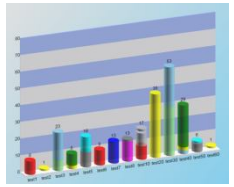
스크립트엔진

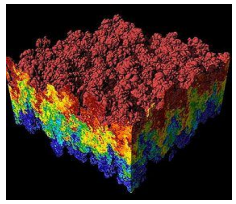
분산병렬처리



표현

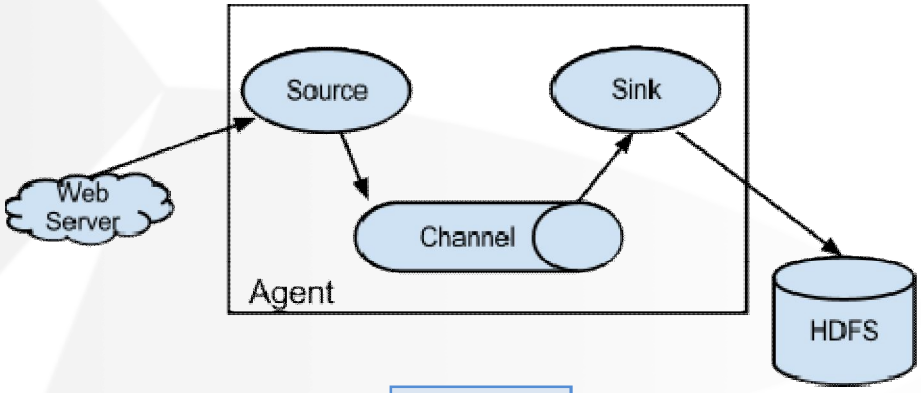




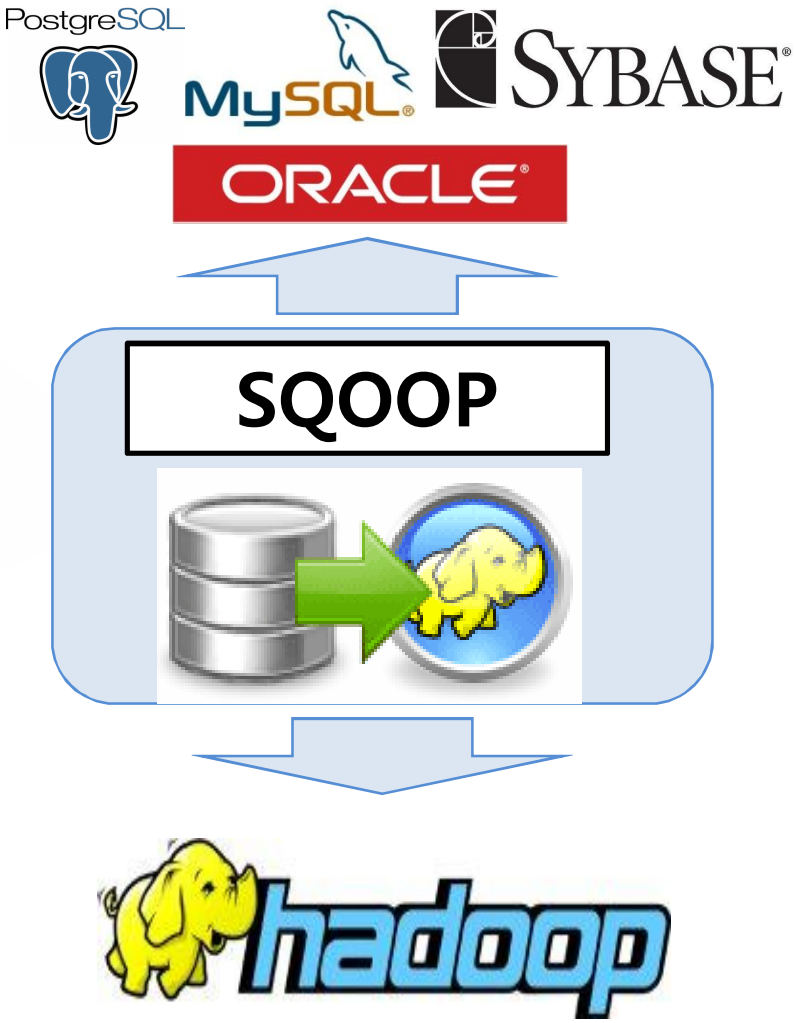


데이터 수집

실시간 이벤트, 로그, 스트림 데이터



관계형 데이터베이스 데이터



데이터 저장 관리

PB 이상의 대용량 비 정형 데이터 저장

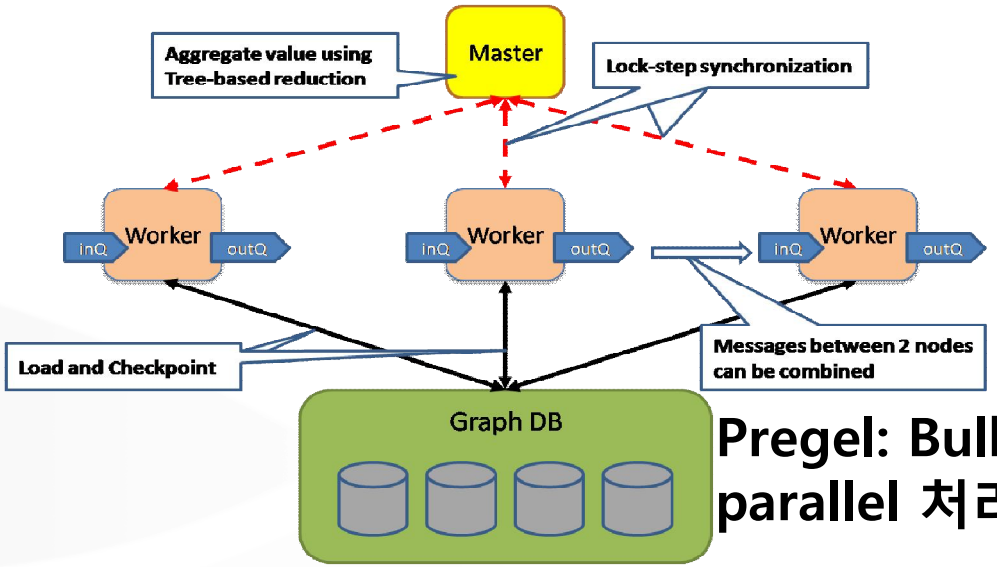
TB 이하의 실시간 조회용 데이터

TB 이상의 실시간 조회 및 분석 용 데이터

데이터 처리 및 연산

Giraph

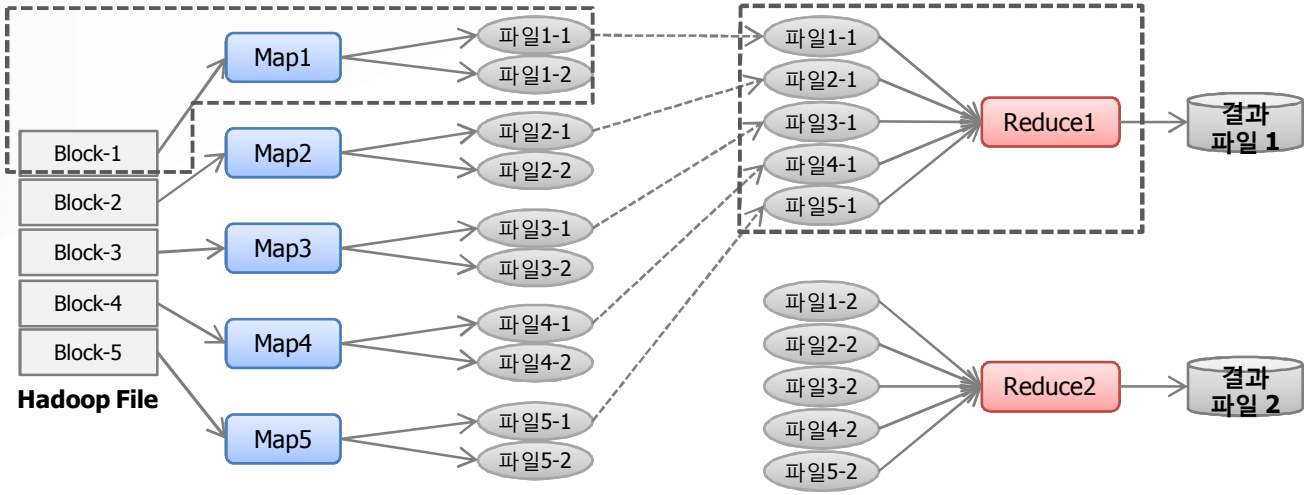
GoldenORB



Pregel: Bulk Synchronous parallel 처리 모델

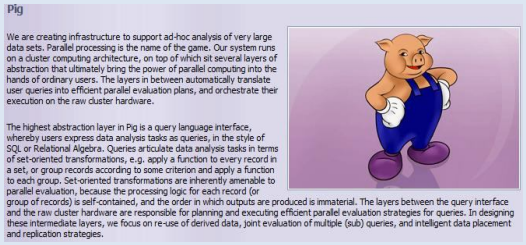
hadoop

MapReduce: 구글에 의해 고안된 분산 컴퓨팅 모델



데이터 모델링과 분석

Query 기반 대용량 데이터 배치 분석



Query 기반 대용량 데이터 준 실시간 분석

18 ~ 25세의 사용자가 가장 많이 방문하는 사이트 5개

사용자 정보

사용자	아이디	나이	성별
길동	kildong	20	남
철수	cheol	25	남
영희	young	15	여
영구	ygu	34	남

사이트 방문 데이터

사이트	방문자	시간
chosum.com	kildong	08:00
ddanji.com	tiffany	12:00
flickr.com	yuna	11:00
espn.com	ygu	21:34


```
select t2.url, count(1) as visits
from userinfo t1 join webdata t2 on
(t1.id=t2.id)
where t1.age > 17 and t1.age < 26
group by t2.url
sort by visits DESC
limit 5;
```

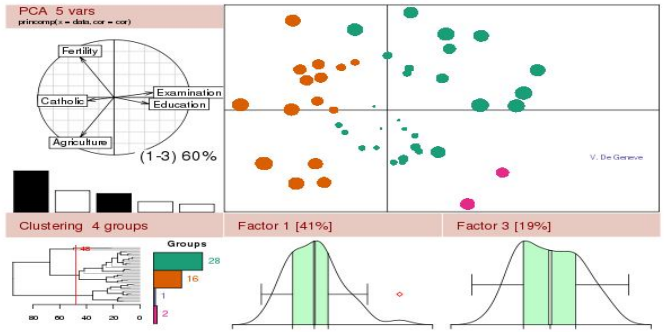
알고리즘과 분석 도구

R 프로젝트

오픈소스 통계 패키지(http://www.r-project.org/)

시각화 기능

기계 학습 기능



아파치 Mahout 프로젝트

아파치 Mahout 프로젝트 (http://mahout.apache.org)

하둡 분산 파일시스템과 NoSQL 데이터베이스 상의 데이터를 대상으로 MapReduce 작업 실행

다양한 정보 분석 알고리즘 제공

- 행렬곱, 벡터 연산 작업
- 클러스터링: Canopy, K-Means등
- 협업 필터링(Collaborative Filtering)

Agenda

I. 빅데이터 사업 동향

II. 주요 오픈 소스 기술 소개

III. 오픈 소스 기술 활용 전략

사업 유형 별 오픈 소스 기술 활용 방법

Questions & Answers