Hadoop과 오픈소스 소프트웨어를 이용한 비지니스 인텔리전스 플랫폼 구축

(Building Business Intelligence Platform Using Hadoop and OpenSource Tools)

PlatFromDay2009 | 2009. 6 . 12

김영우

warwithin@daumcorp.com

다음 커뮤니케이션

프리젠테이션 개요

- 비즈니스 인텔리전스 그리고 데이터 웨어하우스
  - 비즈니스 인텔리전스
  - 데이터 웨어하우스
- 대규모 데이터 분석과 데이터 웨어하우징
  - MapReduce vs. DBMS
  - 왜 Hadoop인가? 문제와 해결방법에 대한 고민
  - Hadoop을 이용한 MapReduce
- Hadoop 기반 데이터 웨어하우징 솔루션
  - 동기
  - Hive
  - CloudBase
  - 오픈소스를 활용한 비즈니스 인텔리전스 아키텍처
- Lessons Learned!

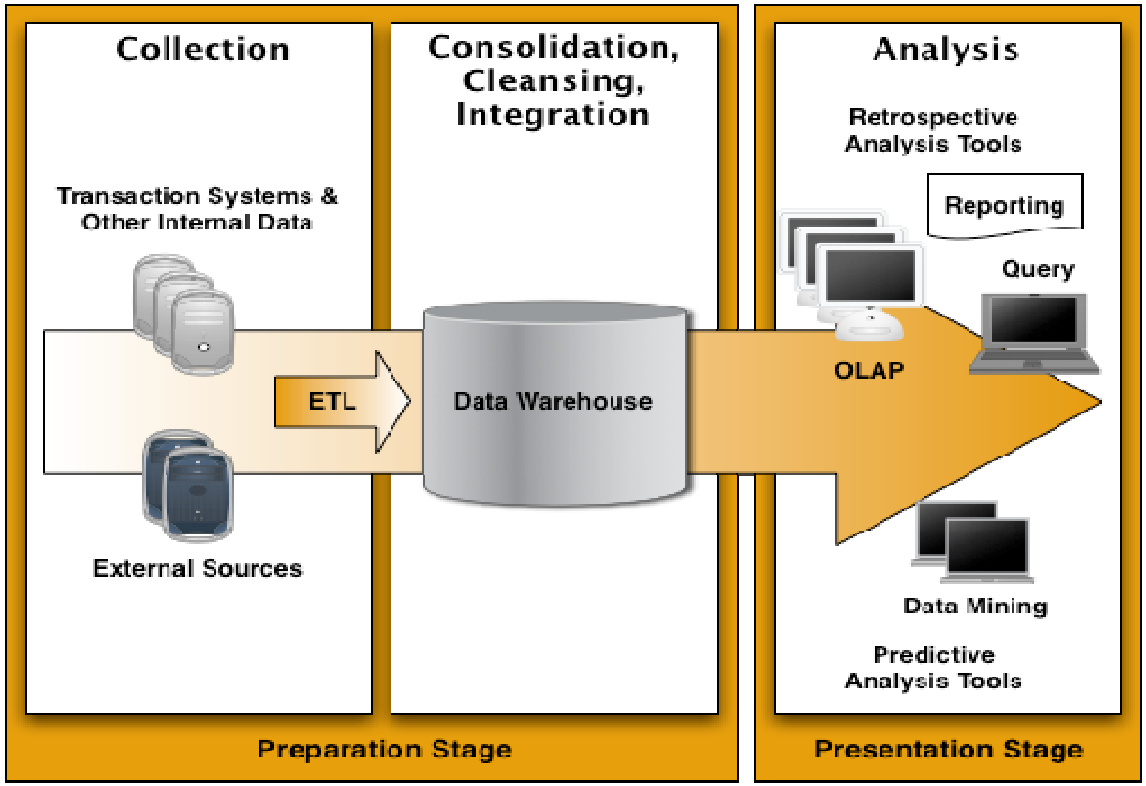
비즈니스 인텔리전스와 데이터 웨어하우스

- 비지니스 인텔리전스 (Business Intelligence)

"Business Intelligence is the process of gathering data, turning that data into information, and sharing that information such that it is useful for increasing top-line efficiency and bottom-line value"

데이터 웨어하우스 (Data Warehouse)

- Reporting
- Single Source of Truth
- Clean Source Data



MapReduce vs. DBMS

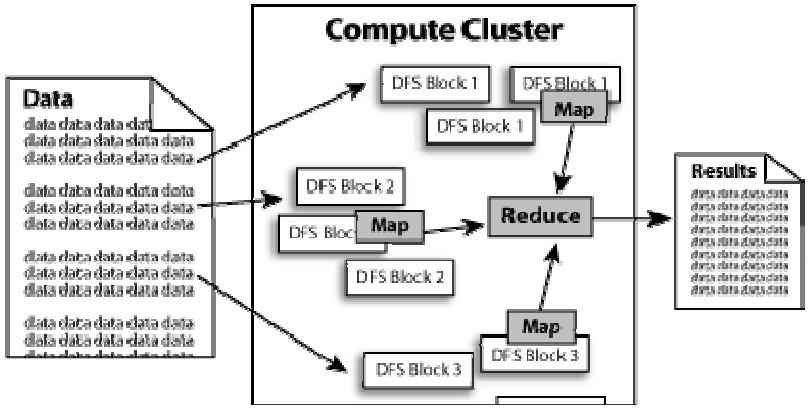
- Row-oriented DBMS vs. Column-oriented DBMS
- MapReduce vs. DBMS
  - "A Comparison of Approaches to Large-Scale Data Analysis: MapReduce vs. DBMS Benchmarks"
- MapReduce on MPP DBMS
  - Aster Database
  - Greenplum

MapReduce vs. DBMS

왜 Hadoop인가? 문제와 해결방법에 대한 고민

- 문제
  - 데이터, 데이터 그리고 데이터
  - Scale Up vs. Scale Out
  - 비용
  - 고가용성, 확장성 그리고 안정성

- 해결방법? Hadoop!
  - HDFS + MapReduce Framework
  - 확장성
  - 유연성
  - 저비용
  - 성능
  - 오픈소스



Hadoop을 이용한 MapReduce

- Hadoop MapReduce API
- Hadoop Streaming
- Pig
  - Yahoo!, 전체 Hadoop MapReduce 작업의 30%
- Cascading
- Java 이외 프로그래밍 언어를 위한 MapReduce 툴킷
- 데이터 웨어하우징 프레임워크
  - Hive
  - CloudBase

Hadoop MapReduce 활용사례

- http://wiki.apache.org/hadoop/PoweredBy

- 데이터 분석
- 검색 인덱싱
- 데이터 마이닝
- 광고 최적화
- 개인화
- 로그 분석
- 통계, 집계
- ……

"Global Information Platforms Evolving the Data Warehouse", Jeff Hammerbacher (Cloudera)

동기

- "Big Data: Viewpoints from the Facebook Data Team", Yahoo 2008 HackHouse, Jeff Hammerbacher (facebook)

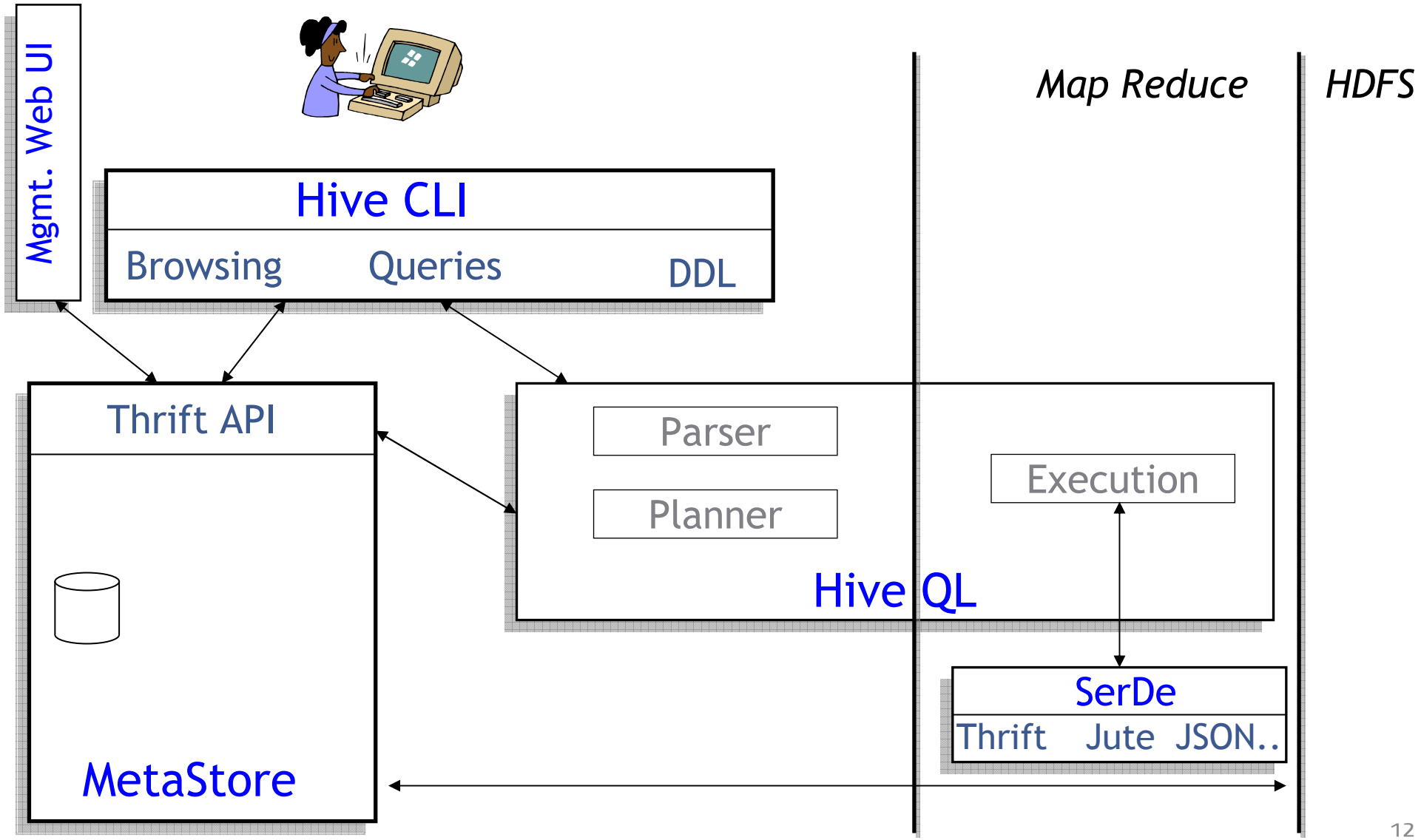
- 사용자는 Java나 다른 언어로 Map Reduce 작업을 직접 개발
- 비정형 질의 처리에 대한 유연성
- 사용자는 SQL에 익숙하다!
- 개발된 BI/리포팅 시스템과 통합 및 연동
- DBMS에서 제공하는 스키마
- 추상화된 프레임워크

Hive

Hive is a data warehouse infrastructure built on top of Hadoop that provides tools to enable easy data summarization, adhoc querying and analysis of large datasets data stored in Hadoop files.

- Facebook
- Hadoop Sub Project

Hive 컴포넌트


Mgmt. Web UI
Hive CLI
Browsing
Queries
DDL
Map Reduce
HDFS
Thrift API
Parser
Planner
Execution
Hive QL
SerDe
Thrift
Jute
JSON..
MetaStore

Hive 주요 기능

- 기본 SQL
  - SELECT … FROM … WHERE …
  - FROM 절의 서브쿼리
  - ANSI JOIN (현재 equi-join만 지원)
  - 다중 테이블 INSERT
  - 다중 GROUP BY
  - 샘플링
  - 파티셔닝
- Pluggable Map-reduce scripts using TRANSFORM
- MetaStore
  - 시스템 카탈로그
  - SQL 백엔드 (Derby, MySQL …)
- JDBC Driver
- Hive Web Interface

Hadoop 기반 데이터 웨어하우징 솔루션

Facebook의 Hive/Hadoop 활용

- 데이터 집계
  - 예: 일별/주별 노출/클릭수 집계
- 고객(사용자) 분석
- 비정형 분석
- 데이터 마이닝
- 스팸 판별
- 사용자 생산 컨텐츠에 대한 패턴 분석
- 애플리케이션 API 사용 패턴 분석
- 광고 최적화
- Lexicon

http://wiki.apache.org/hadoop/Hive/Presentations

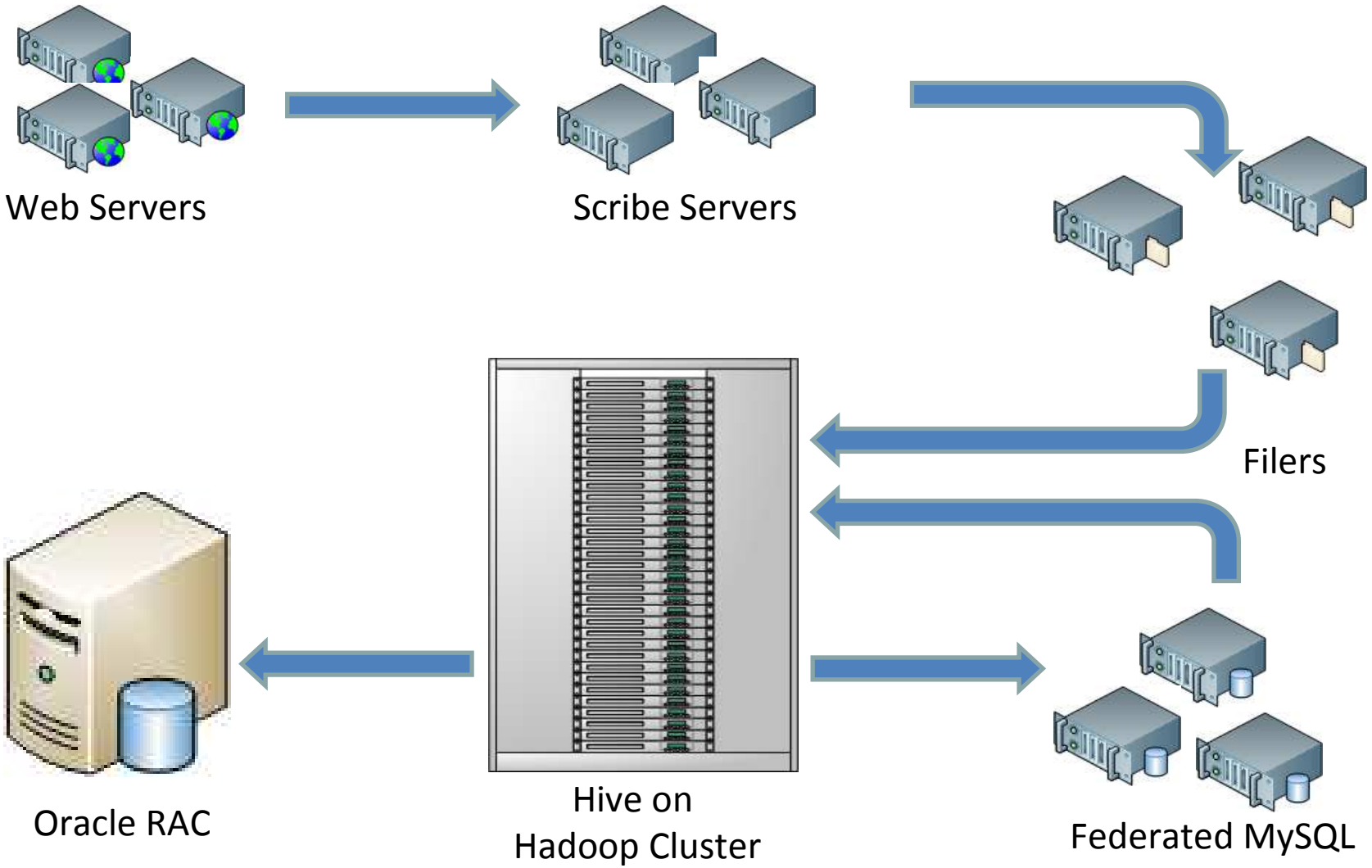
Hadoop 기반 데이터 웨어하우징 솔루션

Hadoop Statistics @ facebook

- 610 노드 (1000노드로 확장 예정)
- 2.5 PB (압축 후, 400 TB)
- 매일 15 TB 데이터 유입
- 매일 4000 개의 작업, 55 TB 데이터 엑세스
- 매일 15 TB 중간 데이터 생성

Hadoop 기반 데이터 웨어하우징 솔루션

Facebook의 데이터 웨어하우스 구성



Hive 로드맵

- BI 플랫폼 통합
  - JDBC/ODBC
- Columnar 스토리지 (HIVE-352)
- 통계 수집 및 비용기반 쿼리 옵티마이저
- JOIN 알고리즘 개선
- 인덱스 (HIVE-417)
- 압축
- SQL 지원 추가
- 고급 기능: Cube, Frequent Item Sets …
- Sqoop ("SQL-to-Hadoop")
  - http://www.cloudera.com/blog/2009/06/01/introducing-sqoop/

http://wiki.apache.org/hadoop/Hive/Roadmap

CloudBase

High-performance Data Warehouse System for Terabyte and Petabyte scale analytics

- Hadoop의 Map Reduce 아키텍처 기반으로 개발 (Java)
- Business.com
- 오픈소스, GPL v2
- http://cloudbase.sourceforge.net

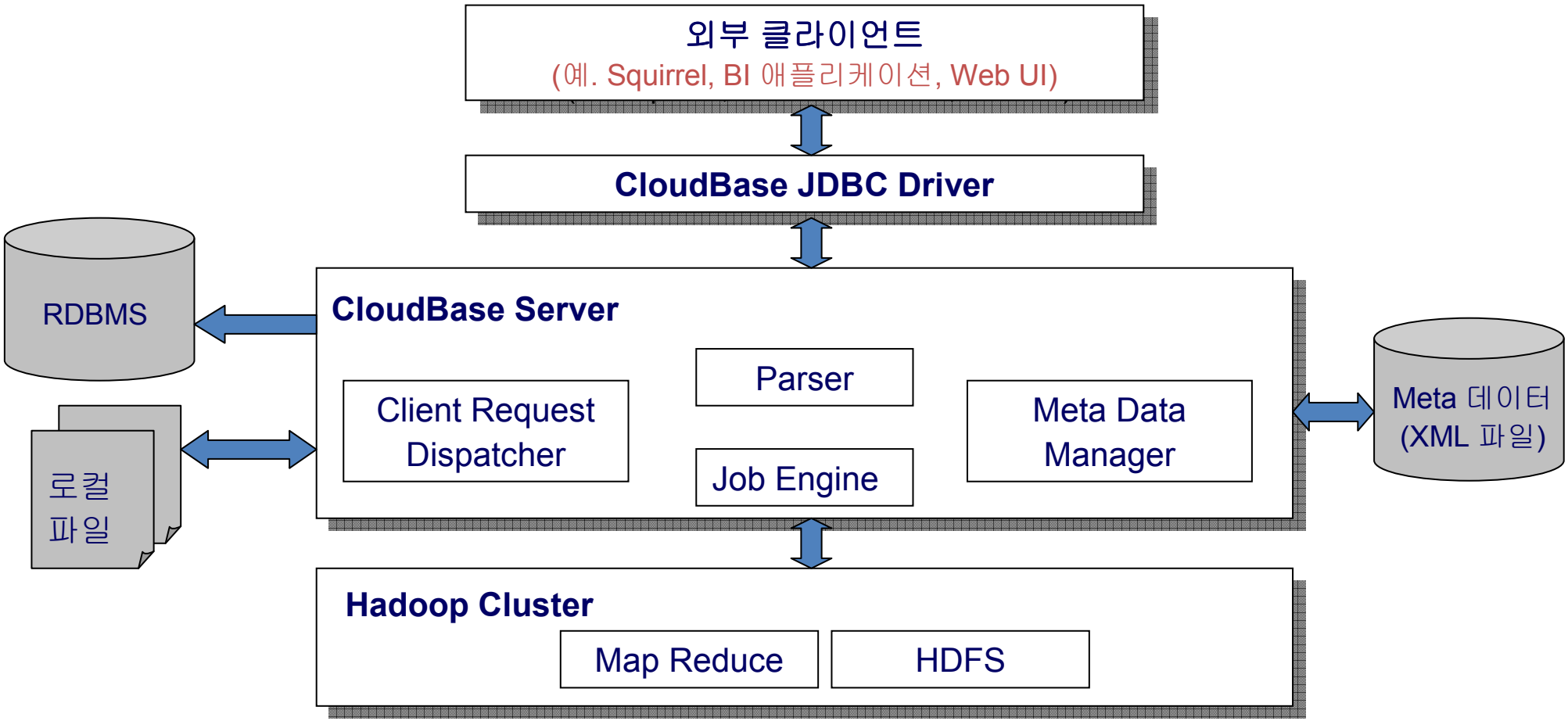
BUSINESS.COM
Business Begins Here™

CloudBase

CloudBase 주요기능

- 질의언어로 ANSI SQL 지원
- 데이터 타입과 SQL NULL 지원
- JDBC 드라이버
- DBLINK를 통한 RDBMS로 데이터 연동
- 문자 함수, 날짜/시간 함수, NULL 처리 함수 지원
- 집계 함수 지원
  - SUM, COUNT, MAX, MIN, AVG, BCAT
- LIKE 구문에서 정규표현식 지원
- 서브 쿼리, 뷰(VIEWS) 지원
- TOP n, ORDER BY, GROUP BY, HAVING 구문 지원
- 테이블 인덱싱 지원
- 사용자 정의 함수(UDF), 사용자 정의 타입(UDT) 지원

CloudBase 아키텍처



CloudBase 최적화

- 테이블 인덱싱
  - 해쉬 인덱스
- Join 알고리즘 구현
  - Semi Join과 비슷함
  - Inner Join, Outer (left, right, full) Join
- ORDER BY 구현
- 질의를 최적의 Map Reduce 작업으로 변환
  - SQL을 MapReduce로 변환

CloudBase on EC2

- CloudBase on Amazon's Elastic Compute Cloud (EC2) cluster.
- Public AMI
- http://cloudbase.sourceforge.net/index.html#ec2

Business.com의 CloudBase 활용

- 원본 로그 파싱(ETL) 및 분석 리포팅 시스템 연동
- 비정형 질의 시스템 제공
- 주간, 월간 리포트 생성

CloudBase Scalability Benchmark @ Business.com

- UserSessions 테이블
  - 40GB 데이터
  - 1억 7천 7백만 로우
- Business.com 프로덕션 데이터베이스에서 질의 실행
- 위와 동일한 질의를 CloudBase(Amazon EC2 cluster)에서 실행

CloudBase Scalability

Query	RDBMS	4 Nodes	8 Nodes	16 Nodes	32 Nodes	64 Nodes
select distinct(dateid) from table	**7 sec**	**17 min**	**10 min**	**5 min 23 sec**	**2 min 17 sec**	**1 min 28 sec**
select sum(bdcrev) as rev from table where lcase(referingengineword) like '%business%card%` and geographyid = 332	**13 min 38 sec**	**26 min**	**14 min**	**7 min 53 sec**	**3 min 19 sec**	**2 min 3 sec**
select * from table where lcase(referingengineword) like '%xybtq%'	**13 min 13 sec**	**20 min**	**11 min**	**6 min 34 sec**	**2 min 27 sec**	**1 min 26 sec**
select dateid, sum(bdcrev) as bdcrev from table group by dateid	**14 min 8 sec**	**27 min**	**14 min**	**8 min 10 sec**	**3 min 25 sec**	**2 min 9 sec**
select userid, count(userid) as cnt_userid from table group by userid order by cnt_userid	**37 min**	**52 min**	**33 min**	**22 min 57 sec**	**17 min 11 sec**	**14 min 36 sec**

CloudBase 로드맵

- 성능
- 개발자, 사용자를 위한 커뮤니티 구축
- 라이선스

오픈소스를 활용한 비즈니스 인텔리전스 아키텍처

- Hadoop
- RDBMS
  - MySQL, PostgreSQL
  - MonetDB, LucidDB
  - CUBRID
- ETL/EAI
  - Pentaho Data Integration (Kettle)
  - Talend Open Studio (TOS)
- BI/OLAP/리포팅
  - In-House Apps
  - BIRT, JasperSoft, Pentaho, Palo
- 데이터 마이닝

Hadoop 기반 데이터 웨어하우징 솔루션

- 로그 분석
- 쇼핑검색 리포팅 및 의사결정지원
- 광고 최적화
- 쇼핑 개인화

- Hadoop
- CloudBase
- Pentaho Data Integration (Kettle)
- Oracle
- Java/Flex

- Pig

Lessons Learned!

- Keep It Simple!
  - 로그
- Size Matters!
  - The Unreasonable Effectiveness of Data, Google
- PoC, 테스트 그리고 성능평가
- 비용
  - 문제를 해결하기 위한 최선의 선택
- Data Integration (EAI, ETL)
- 리포팅, 시각화, 비즈니스 성능 관리, 데이터 분석
- 데이터 품질
  - Garbage in, garbage out!
- 버그, 패치 그리고 워크어라운드
- 커뮤니티, 블로그, 포럼, JIRA, 메일링 리스트

Lessons Learned!

- 협업
  - 데이터 아키텍트, 개발자, 비지니스 사용자, 데이터 분석가, ...
- 오프소스!
- 시스템 관리
  - 시스템 설정 관리
  - 성능 모니터링/관리

- '기술'과 '전략'의 문제
- 씹을 수 있는 만큼만 물어라!
- 역시, 어려운 문제!

참고문헌

- Hadoop, http://hadoop.apache.org/
- Hive, http://hadoop.apache.org/hive/
- CloudBase, http://cloudbase.sourceforge.net
- Pig, http://hadoop.apache.org/pig/
- Clarise Z. Doval Santos and Joseph A. di Paolantonio, "The Economics of BI: How to Drive Cost Effective Strategies", Campus Technology 2007
- Presentations About Hive, http://wiki.apache.org/hadoop/Hive/Presentations
- Trandeep Singh, "CloudBase", Business.com
- Jeff Hammerbacher, "Global Information Platforms Evolving the Data Warehouse", Cloudera
- Kun Tong, "Web scale data mining using PIG", 2nd Hadoop in China Salon
- DBMS2, http://www.dbms2.com/
- "BI의 비즈니스 가치 증대를 위한 데이터 통합 플랫폼", 한국인포매티카

피드백?

- 질문?
- 아이디어?

감사합니다