오픈소스 Hive를 이용한 분석 Data Warehouse 구축 경험 공유

2013.12

윤종완

NCSOFT CORPORATION
12, Daewangpangyo-ro 644beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do, Korea

WWW.NCSOFT.COM

목 차

- 들어가는 말
  - Hive와 분석DW
  - 기업 데이터 분석 환경
  - NC 전면 전환 결과
- 전환 경험 이야기
  - 분석가 말 잘 듣기
  - 힘들었던 ETL 전환
  - 중요한 두 가지 전제
- 정리

Hive와 분석 DW

The Apache Hive data warehouse software facilitates querying and managing large datasets …

- hive.apache.org 발췌

Data warehouse is a database used for reporting and data analysis.

- en.wikipedia.org 발췌

"Hive만으로 기업 분석DW를 구축할 수 없다."

Hive 실체

Home

Tools

Added by Confluence Administrator, edited by Confluence Administrator on Jun 24, 2011

= What is Hive =

Hive is a data warehouse infrastructure built on top of [Hadoop]. It provides tools to enable easy data ETL, a mechanism to put structures on the data, and the capability to querying and analysis of large data sets stored in Hadoop files. Hive defines a simple SQL-like query language, called QL, that enables users familiar with SQL to query the data. At the same time, this language also allows programmers who are familiar with the MapReduce framework to be able to plug in their custom mappers and reducers to perform more sophisticated analysis that may not be supported by the built-in capabilities of the language.

Hive does not mandate read or written data be in the "Hive format"---there is no such thing. Hive works equally well on Thrift, control delimited, or your specialized data formats. Please see File Format and SerDe in the Developer Guide for details.

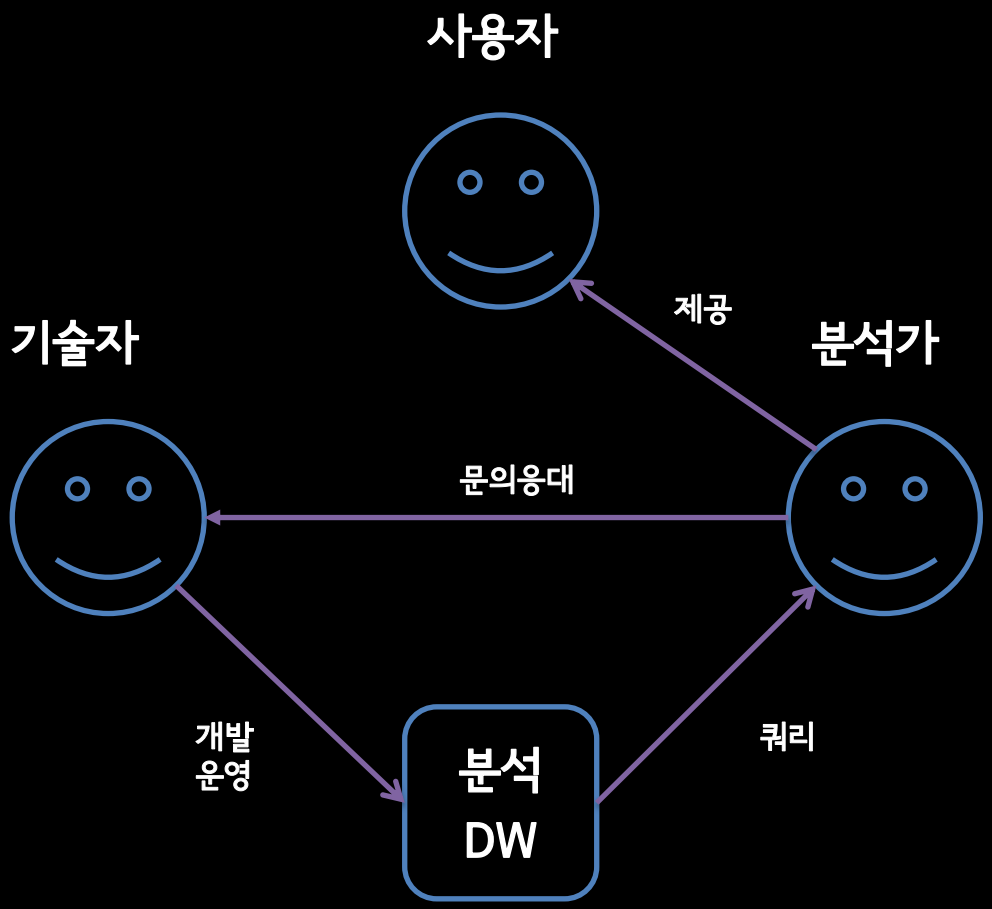
"Hive는 분석DW에서 DB와 같은 '데이터 그릇' 역할만 한다."

기업용 분석 DW 역할

- 분석의 기본은 '비교'의 반복
  - SAS
  - R
  - Excel Pivot
  - OLAP Tools
- 비교 데이터는 수집, 정제, 그룹 과정으로 생산
  - ETL Tools
  - Hive
  - Parallel DBMS
  - …

"수집, 정제, 그룹 추출을 잘 하는 추가 도구 구축이 필요하다."

기업용 분석DW 이해 당사자



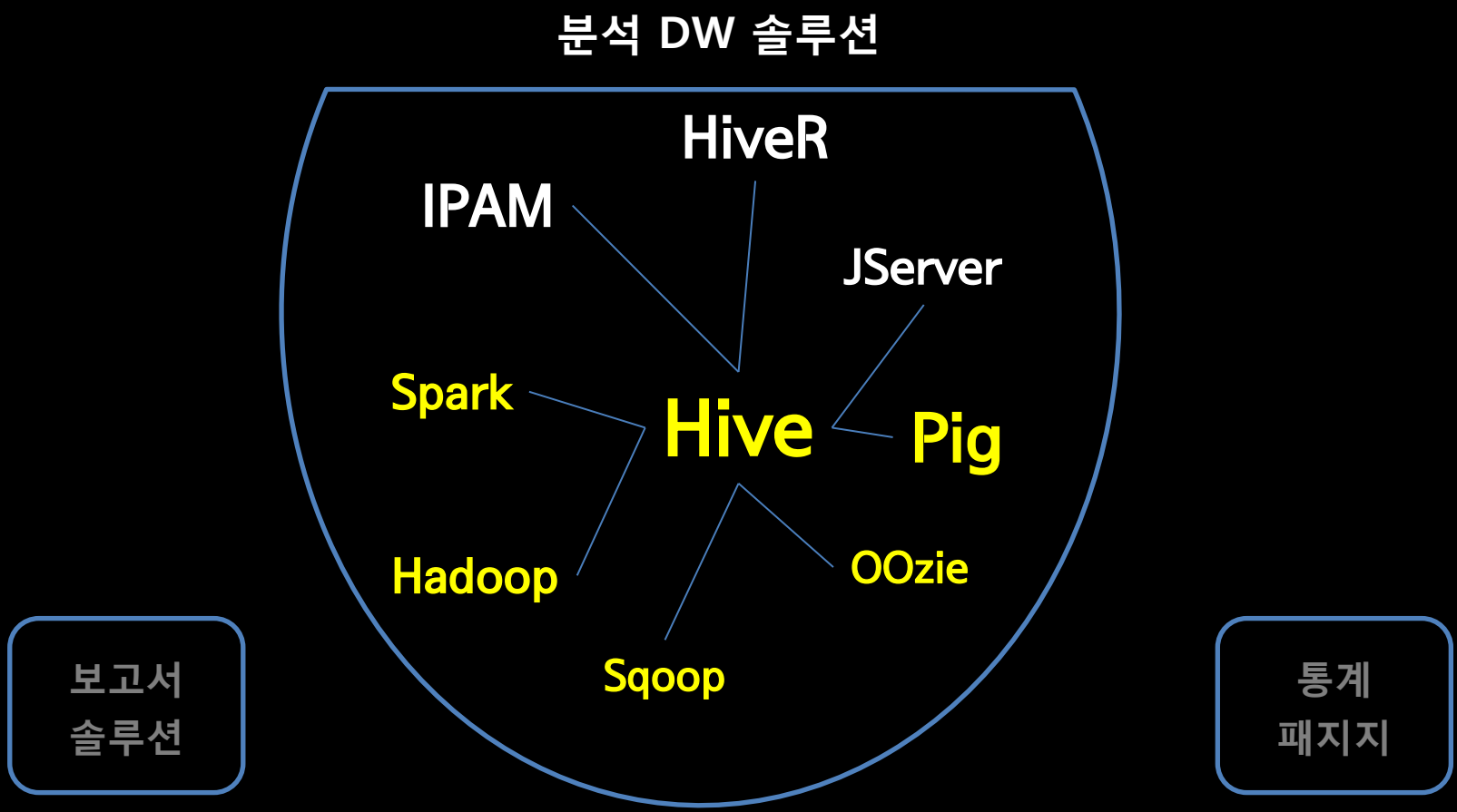
"Hive 분석 DW 구축에는 이해 당사자의 동참이 절실하다."

NCSoft 분석 DW 전환 이야기

- ‘2012년 가을 착수 ~ 2013년 여름 완료
- 오픈소스 + 인하우스 제품으로 구성
- RDMBS를 사용 않는 100% 하둡 솔루션
- 주 2만 쿼리, 90명 분석가 사용 중
- 매출, 이탈, 인 게임 등 모든 분석에 활용

“1년간의 프로젝트로 하둡 기반 분석 DW 전환을 완료했다.”

2013년 가을 NCSoft 분석 DW 모습



"Hive를 중심으로 8개 솔루션을 결합해서 분석DW를 만들었다."

분석 모습 예

"엔씨 분석가는 HiveR로 600T 공간내 분석 데이터를 스스로 추출한다."

프로젝트에 대한 고민

1. 하둡 기반 분석 기반으로 '통째로' 이전 사례가 없는데?
2. 과연 분석가들이 빠른 RDB를 버릴까?
3. 1만개가 넘는 ETL 전환이 가능할까?
4. 매출 Fact도 하둡 ETL로 만들 수 있을까?

지당하지만 까다로웠던 분석가의 요구

1. 쉽고 편한 쿼리 도구
2. 상업용 DB같은 빠른 쿼리 처리 속도
3. 데이터에 대한 접근 및 권한 제어
4. 다중 사용자 지원

“분석가들은 DB + 하둡 장점만을 가진 분석 DW를 원했다.”

발전하는 오픈소스

- SQL on Hadoop
  - ✓ Spark
  - ✓ Impala
- Toad for Cloud Database
- Hiveserver2

"분석가 요구들은 오픈 소스 진영에 의해 빠르게 개선되고 있다"

교훈:짧은 쿼리 속도 튜닝

- Hiveserver를 개선하여 속도를 튜닝함
- 짧은 쿼리 긴 쿼리를 다른 인프라로 처리함
- 사용자가 인프라의 다름을 인지하지 못함
- 15초 ~ 150초 대 쿼리의 속도 개선을 목표
- 약 40% 쿼리를 Shark로 처리, 220% 개선함

"많이 쓰는 짧은 쿼리에 대한 대책이 꼭 필요하다"

확신은 없고 부담만 컸던 기술자

1. 6백여 개의 적재 대상 Hive 테이블
2. 1만여 변환 세션을 가진 ETL
3. Pig 학습에 대한 부담
4. 하둡 ETL 성능에 대한 불안

“분석 DW 사용자 요구만이 전부는 아녔다.”

교훈: 처리 성능

레코드 수	INFA	PIG
18백만	3mins, 32sec	6mins, 3sec
21만		
17백만		
2백만		1mins, 39sec
총 소요시간	3mins, 34sec	7mins, 42sec

INFA 대 Pig ETL 성능 비교표

"Pig가 항상 성능이 좋진 않다."

교훈: Pig ETL 생산성

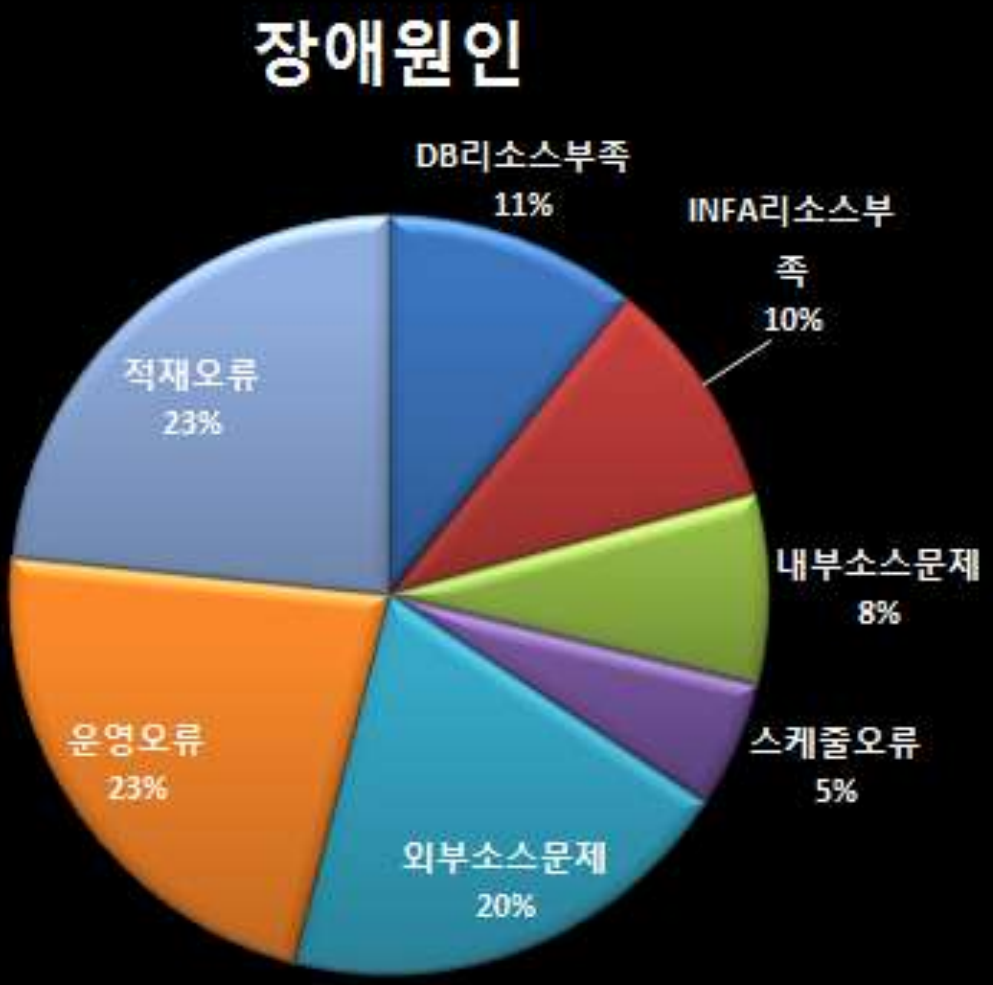
- SQL 기반 ETL은 집합 개념에 대한 이해가 핵심
- Pig는 형태만 다를 뿐 SQL과 같은 집합 처리 기술임
- Pig syntax는 무척 간단함.

```
IMPORT ‘pigs/ACC/CC/LOAD_GC_ACCOUNT.pig’;
IMPORT ‘pigs/ACC/CC/LOAD_FGC_LAST_PS_GOODS.pig’;

DEFINE GET_ACTIVE_ACCOUNT(dir, execute, yyyymmdd) return rst {
  …
  GAME_PLAY_AGREEMENT = LOAD_GAMEPLAYAGREEMENT (…);
  GAME_PLAY_AGREEMENT = FOREACH AN_ACCOUNT_ETC GENERATE
    GAME_ACCOUNT_NO,
    LAST_LOGIN_DATE,
    LAST_LOGOUT_DATE
  ;
  GAME_PLAY_AGREEMENT = FILTER GAME_PLAY_AGREEMENT BY SERVICE_CODE==27;
  …
}
```

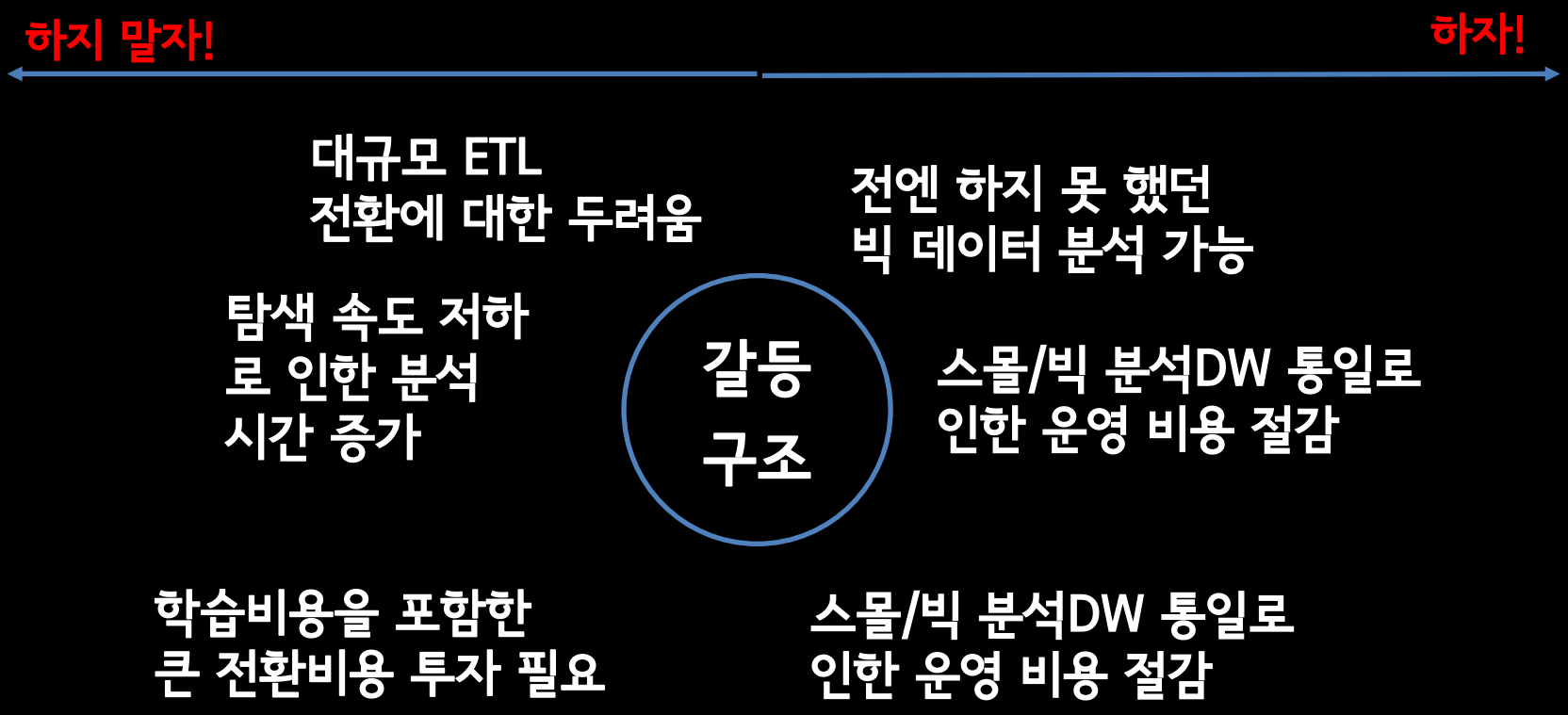
“생각보다 나쁘지 않은 Pig ETL 생산성”

교훈: ETL은 ETL



"Pig를 도입해도 ETL 강건성은 22% 개선 효과만 기대할 수 있다."

성공적인 구축을 위한 +alpha



"Hive 분석DB로 100%전환을 위해서는 꼭 필요한 것이 있다."

Alpha: 임원의 지원

- 결국, Hive DW 도입은 단기, 중장기 이득의 싸움
- 분석가는 자주하는 단기 분석의 효율화에 주안점
- 임원은 회사 발전에 필요한 중장기 장점에 주안점
  - ✓ 적재 후 분석에 큰 관심 (데이터 자산화)
  - ✓ 사용자, 분석 결과의 폭에 더 관심 (인뎁스 분석)
  - ✓ 통일된 DW 인프라 (운영 비용)

"중장기적인 시각을 가진 임원의 지원이 중요"

Alpha: 프로그래머 채용

- 오픈 소스 기업도 역시 기업이다. 분석DW 구축을 도와주지 않는다. 스스로 알아서 구축하거나, 상업용 BI솔루션을 구입해야 한다.
  - ✓ Pentaho for Cloudera
  - ✓ TERADATA ASTER와 Horthonworks

- 오픈 소스 도입은 그럭저럭한 상용 솔루션을 목표로 삼으면 안 된다. 동종 기업에서 못하는 딱 맞춤에 초점을 둬야 한다.

"상용용 솔루션 구입 비용으로 프로그래머를 채용하라."

요약

- 기업용 Hive 분석 DW 구축이 가능하다.
  - ✓ 프로그래머를 채용하라.
  - ✓ 임원의 지지를 얻어라.
  - ✓ 분석가 요건을 잘 들어라.
  - ✓ 기업 ETL을 Pig로 전환하라.
- 문의는 jongwanyun@gmail.com 로 연락