

공개 SW 개발자 대회 Microsoft 지정과제 Hatchery Project

연구박재범

Contents

I 프로젝트 소개

II 시연

III 성능 분석

IV 기대효과 및 개선방안

프로젝트 소개



**big
data**



Shift

데이터 분석 과정 중에 겪는 문제점

1. 사용하고자 하는 데이터 분석 도구마다 데이터 파일에 대한 서로 다른 제약사항을 가지며 사용방법이 다르다.
2. 허용하는 데이터 파일 포맷이 한정적이다.
3. 서로 다른 Data Source로 부터 데이터를 필요로 하는 경우가 있다.
4. 데이터 파일 안에 값이 존재하지 않거나 consistency가 유지되지 않는 경우가 많다.
5. 분산되어 있는 데이터 파일을 합쳐야 하는 경우가 있다.
6. 프로그램 개발시에 데이터 파일마다 별도의 ETL 프로그램을 개발해야 한다.

데이터 분석 과정 중에 겪는 문제점

1. 사용하고자 하는 데이터 분석 도구마다 데이터 파일에 대한 서로 다른 제약사항을 가지며 사용방법이 다르다.
2. 허용하는 데이터 파일 포맷이 한정적이다.
3. 서로 다른 Data Source로 부터 데이터를 필요로 하는 경우가 있다.
4. 데이터 파일 안에 값이 존재하지 않거나 consistency가 유지되지 않는 경우가 많다.
5. 분산되어 있는 데이터 파일을 합쳐야 하는 경우가 있다.
6. 프로그램 개발시에 데이터 파일마다 별도의 ETL 프로그램을 개발해야 한다.

Apache Drill을 이용한 Simple ETL(Column Extraction)프로그램

개발 환경

OS

- Linux Ubuntu 14.04.5 LTS

주요 Library

- Apache Hadoop 2.7.2, Apache Hive 1.2.1, Apache Drill 1.8.0

User Interface

- Web Browser(Mozilla Firefox, Google Chrome)

개발 언어

- Java 8, HTML5, CSS, JavaScript

개발 도구

- Eclipse Neon, Sublime Text 3

구동 환경

- Apache Drill의 실행이 전제
- File System은 Local File System(Disk) 혹은 Distributed File System(HDFS)을 이용
- Apache Maven을 이용한 Build 및 실행

The logo for Apache Maven, featuring the word "Maven" in a bold, black, sans-serif font. The letter "a" is stylized with a colorful feather graphic integrated into it. A small "TM" trademark symbol is positioned to the upper right of the word.The logo for Apache Drill, consisting of a stylized grey feather icon on the left and the text "APACHE DRILL" on the right. "APACHE" is in a smaller, grey, sans-serif font, and "DRILL" is in a larger, blue, sans-serif font.

Hatchery

데이터 파일 포맷의 제약



다양한 파일 포맷 변환 기능 지원

데이터 파일 별 ETL 프로그램 개발



Hatchery

별도의 개발 없이 Simple ETL 가능

다른 데이터파일 간의 Join 기능 없음



다양한 데이터 파일의 Join 지원

세부 기능

1. 다양한 포맷의 Data file들로 부터 Metadata를 추출
2. 간단한 Web UI를 이용한 Column 추출 및 파일 포맷의 변환
3. Hive Metastore 접근성 제공
4. Multiple Data File에 대한 join 기능
5. JSON File에 대한 Flatten 옵션 제공



시연

성능 분석

 **big
data**

 Shift

분석 환경

OS

- Linux Ubuntu 14.04.5 LTS on Virtual Machine

Memory

- 4 GB

Apache Hadoop

- Pseudo Distribute Mode

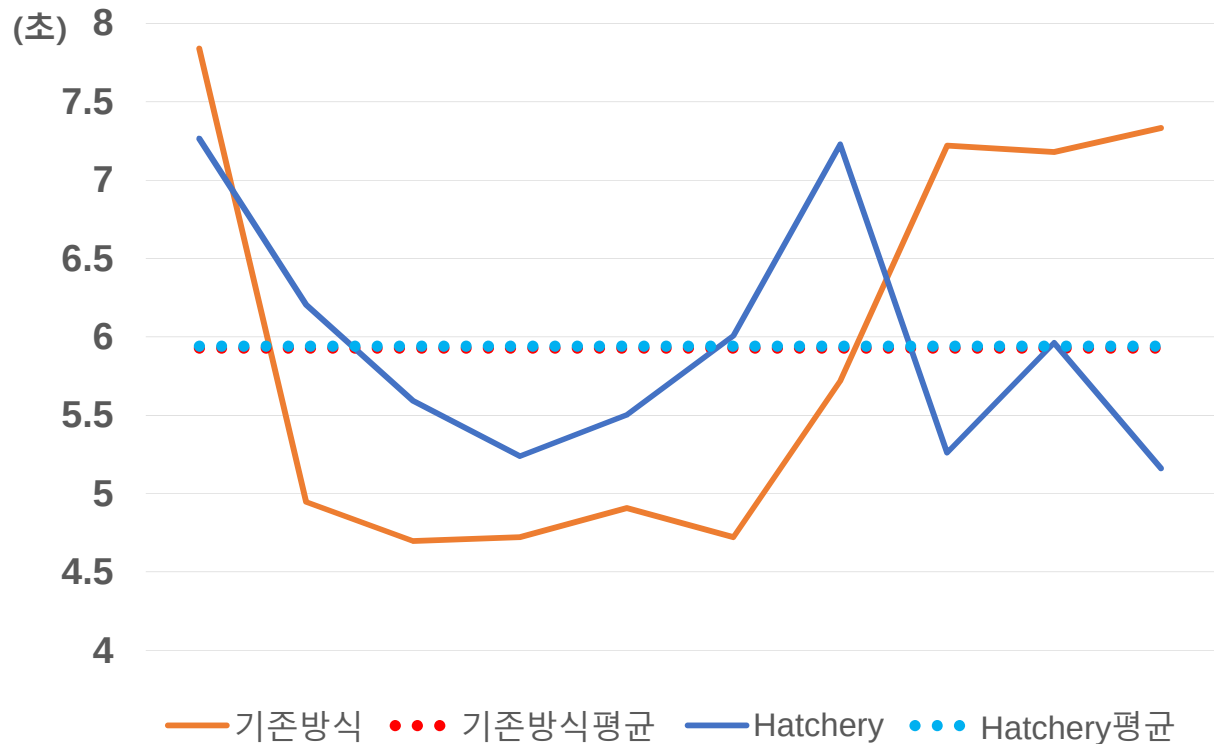
Apache Drill

- Embedded Mode

성능 분석 시나리오

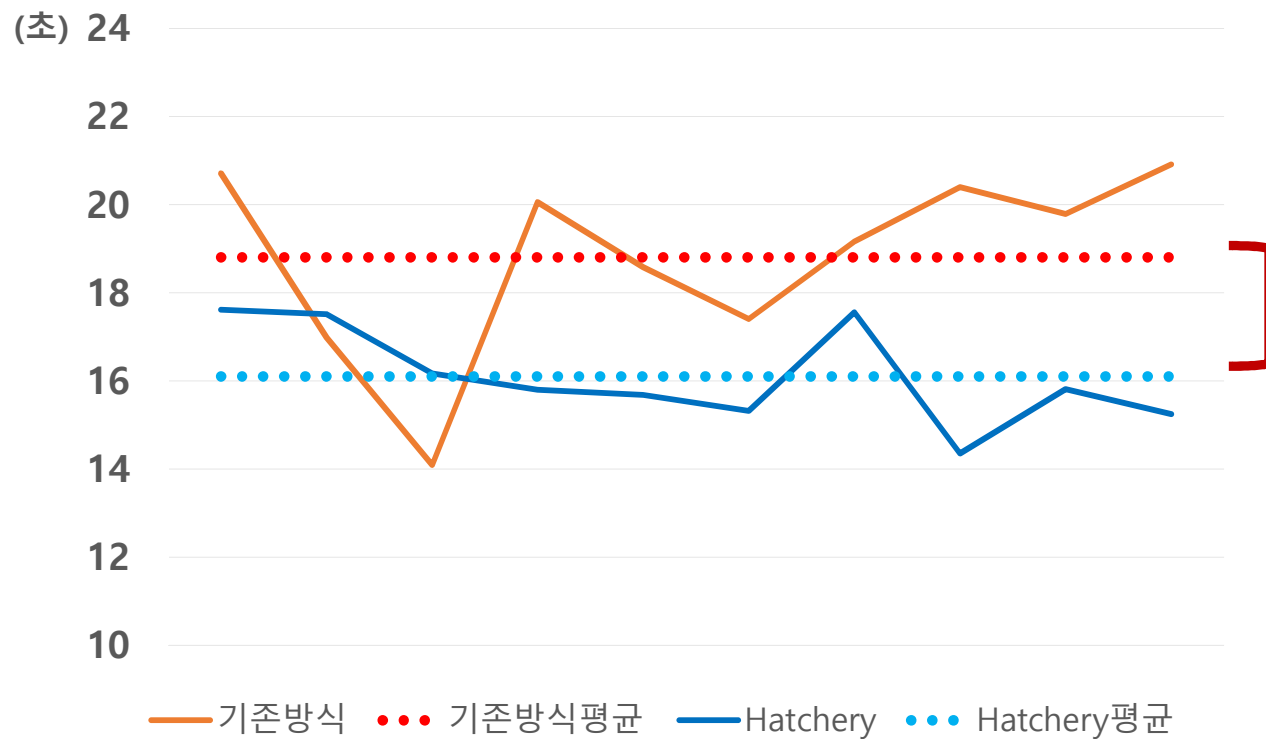
Scenario #1		Scenario #2	
단일 파일 [csv, 689.4MB]		디렉토리 [csv, 672,1MB, 702.9MB, 689.4MB]	
대조군	실험군	대조군	실험군
HDFS 내부의 CSV 파일로 부터 컬럼 추출		HDFS 내부의 여러개의 CSV 파일로 부터 컬럼 추출	
Java native	Hatchery	Java native	Hatchery

Scenario #1



기존 방식과 평균 속도의
차이가 없음

Scenario #2



평균 18.81초 -> 16.11초
14.35% 속도 향상

기대효과 및 개선방향



기대효과

1. 간단한 웹 UI로 데이터 파일에 대한 접근성 증가
2. 자유로운 파일 포맷 변환으로써 데이터 파일 자체의 활용성 증가
3. 데이터 분석에 있어서 개발자가 데이터의 전처리 코드를 직접 개발하는 불편함을
대폭 감소
4. 데이터 파일의 최소화를 통한 분석 Query의 최적화 도모

개선방향

1. 다양한 데이터 분석 도구에 대한 사전 규약 처리
2. 서로 다른 Data Source간 통합 전처리 제공
3. 데이터 파일의 Inconsistency 개선 기능 제공
4. Azure Cloud Service를 활용한 작업 성능 개선

감 사 합 니 다