🧐

스쿱 (Sqoop)

HDFS 외부의 스토리지 저장소 접근 API

- 회사의 주요데이터는 관계형 RDBMS

아파치 스쿱은 구조화된 데이터 저장소에서 데이터를 추출해서 하둡으로 보내 처리할 수 있게 해준다.

HBase로 옮기는데도 사용

물론!.

- 이러한 처리는 하이브,MapReduce프로그램으로도 할 수 있다.

스쿱 설치

[Sqoop] Sqoop 설치 및 MySQL과 Import, Export하기 (tistory.com)

스쿱 커넥터

스쿱이 임포트와 익스포트하게 해주는 모듈식 컴포넌트

MySQL,PostgreSQL,오라클 SQL 서버, DB2 다양한 커넥터 제공

또한 자바의 JDBC 프로토콜을 지원하는 어떤 DB에도 연결가능 제네릭 JDBC 커넥터 제공

임포트 예제

mysql 설치후

```
mysql -u root -p
create database hadoopguide
GRANT all Privileges on hadoopguide .* TO ''@'localhost' //특권 모드 모든 연산 , 로컬호스트 접근
```

mysql hadoopguide

create table 생략

삽입 과정 생략

위과정을 통해 생성된 테이블을 HDFS 로 임포트할 수 있다.

sqoop import —connect jdbc:mysql://localhost/hadoopguide \

—table widgets -m 1

m옵션 : 병렬 연산 갯수

스쿱의 import 도구는 맵 리듀스 잡을 실행, MYSQL DB 접속을 하여 , 테이블을 읽는다.

읽어온 데이터는 hadoop fs -cat widgets/part-m-00000을 통해 확인가능

기본적으로 스쿱은 임포트한 데이터를 콤마로 구분된 텍스트파일 생성

컬럼 구분자 파일 포맷 압축 임포트 세부제어 ⇒ http://sqoop.apache.org

기본적으로 텍스트파일로 읽는데 , 바이너리 필드를 못가지며, null값과 “null”값을 구분못한다.

import 하면 , 자바소스 파일이 생성됨 (위에서는 widgets.java) ,특정 테이블 데이터를 가져오기전 미리 생성된 자바코드를 생성하고 사용

import 수행하지 않고, 자바 소스파일만 생성할 수도 있다

-m 5 옵션으로 줄 경우, if db record id 10000

5개의 sql문 where 를 통해 5개로 나뉘어 실행

—query로 컬럼 변환과 같은 세밀한 제어 수행 가능

HDFS에 있는 데이터와 DB 데이터 동기화를 유지하기 위해서는 주기적인 임포트 작업이 필요.

하둡 스트리밍 스크립트 : 명령줄로 mapreduce 코드 작성없이 사용하는 것

스쿱실행과 동시에 maprduce 컴파일된 jar을 지정하여 작업가능

ex) HADDOP_CLASSPATH=$SQOOP_HOME/sqoop-versionjar hadoop jar \

sqoo-examples.jar MaxWidgetId -libjars $SQOOP_HOME/sqoop-version.jar (이해 안 됨)

익스포트 수행

db에 hdfs 를 목적지로 결과를 쓸 수 있따.

타깃 테이블을 미리 준비

```
sqoop export --connect jdbc:mysql://localhost/<타깃db> -m <병렬처리갯수> \
--table <타깃테이블> --export-dir /user/hive/warehouse/zip_profits \
input-fields-terminated-by '\001'
```

—input-fields-terminated-by 구분자 이스케이프 시퀀스 지원

임포트와는 다르게, 요구가 많네..