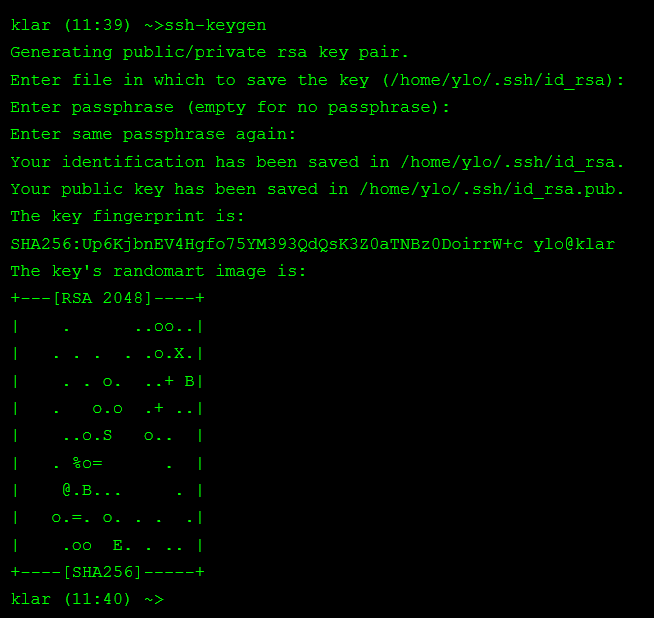
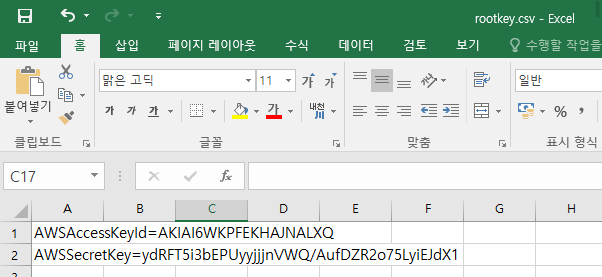
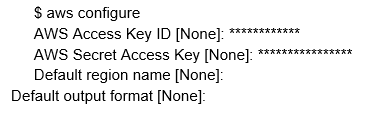
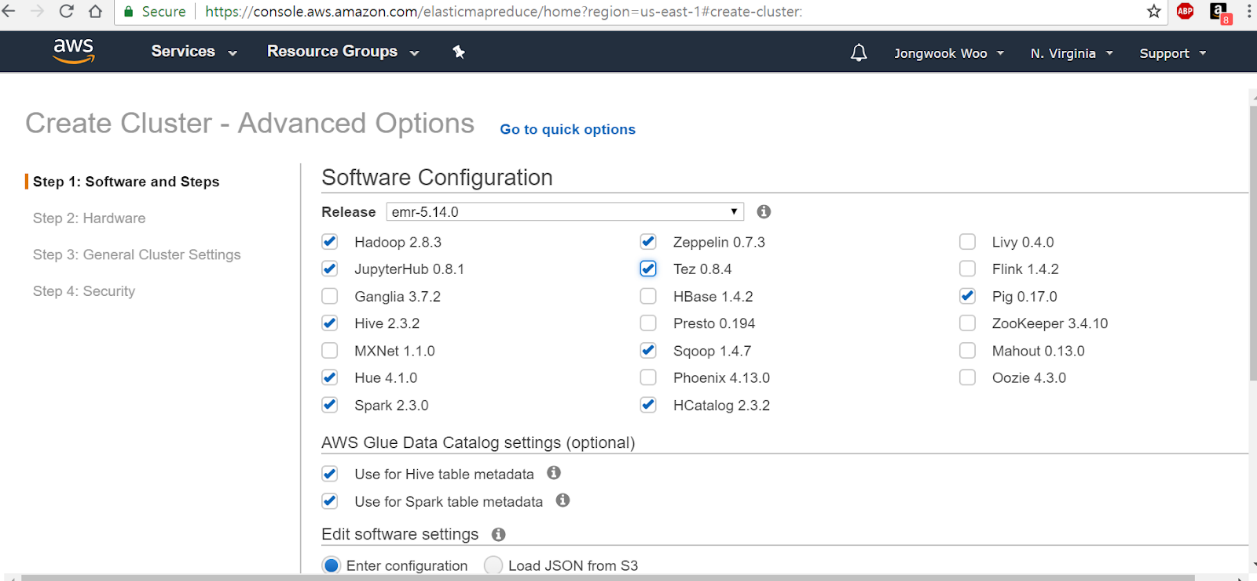
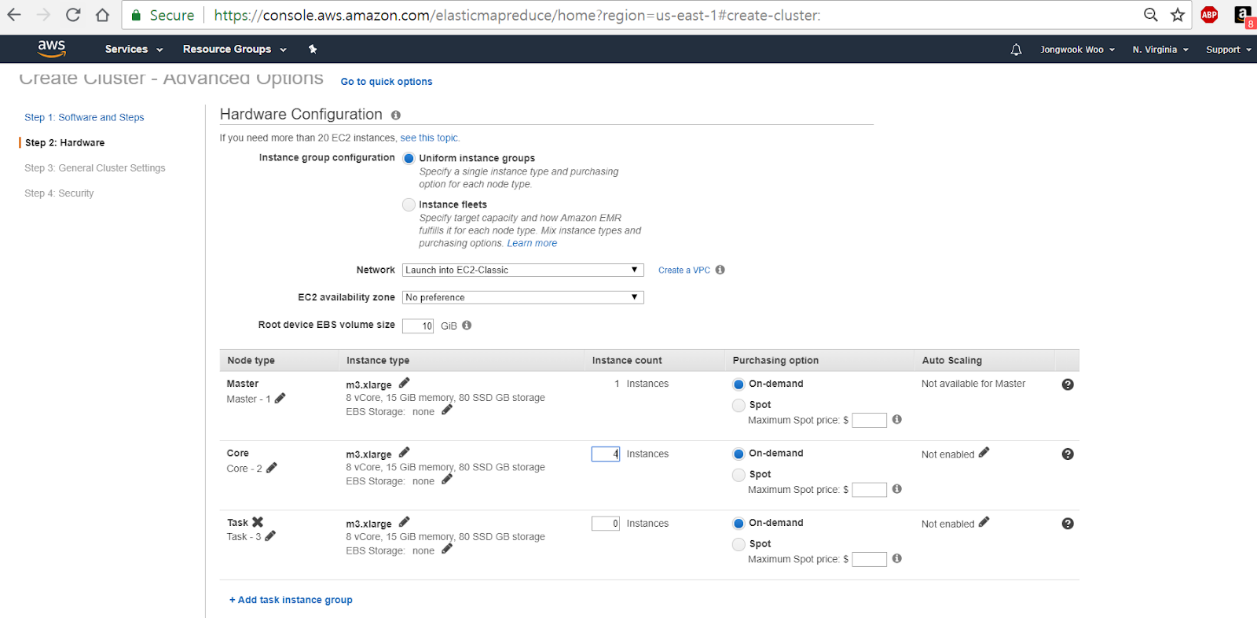
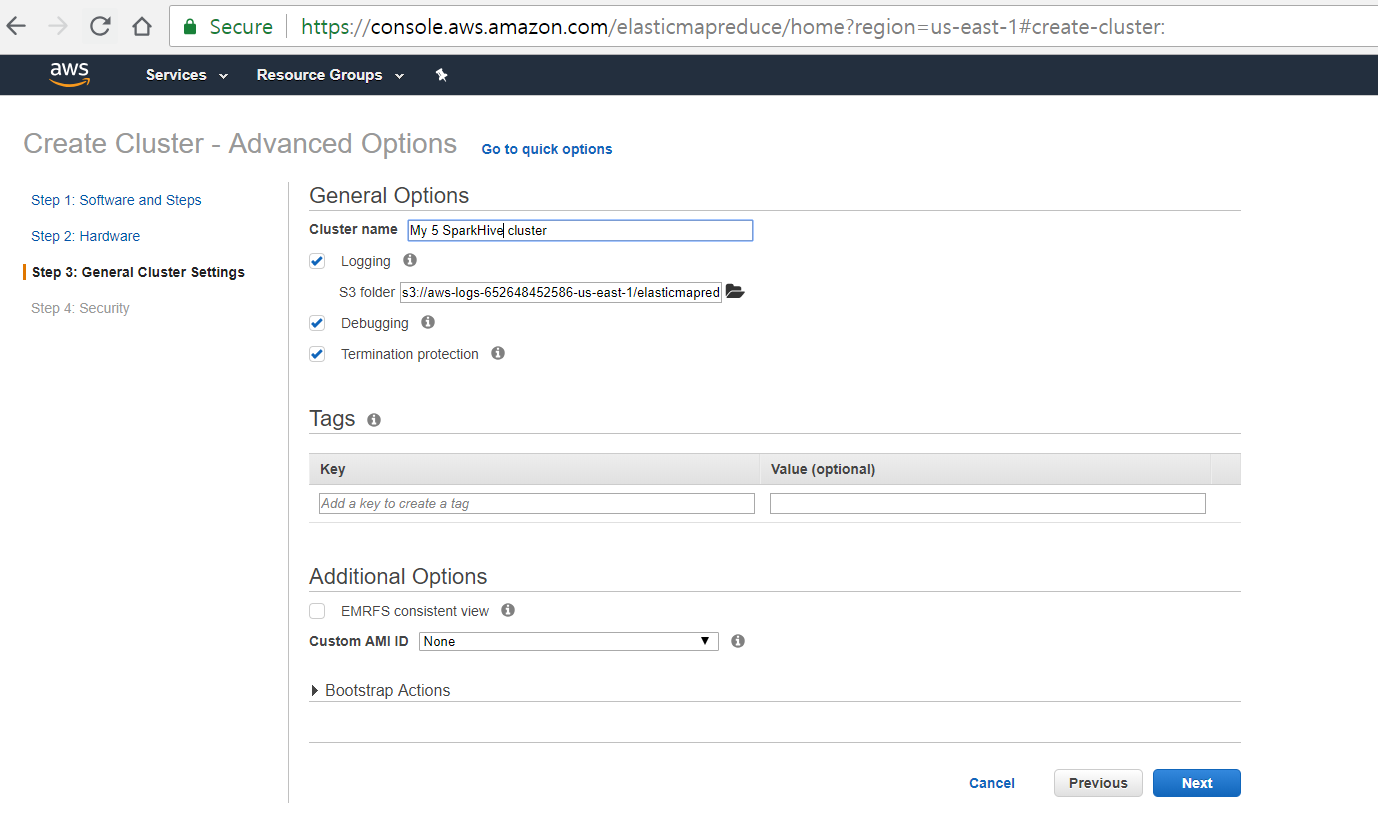
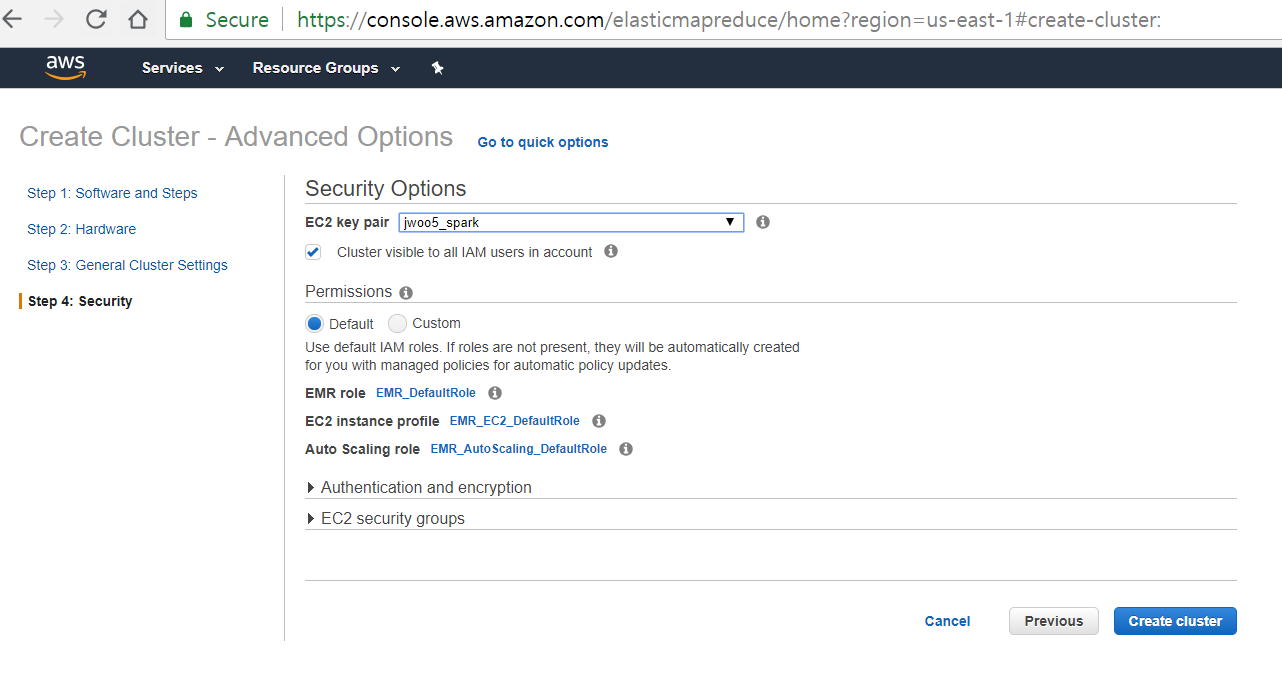
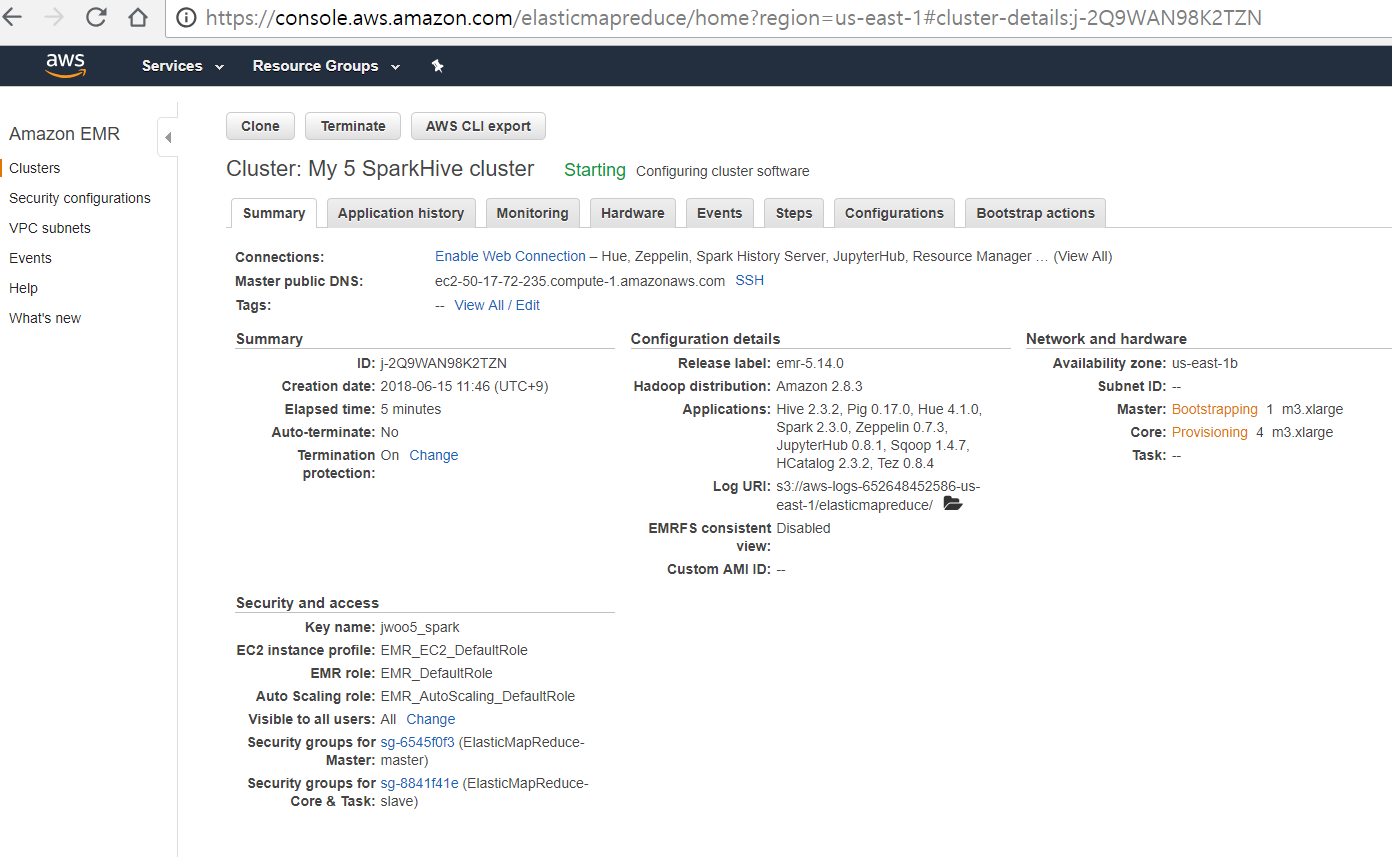
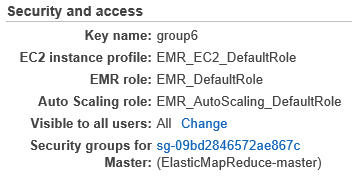
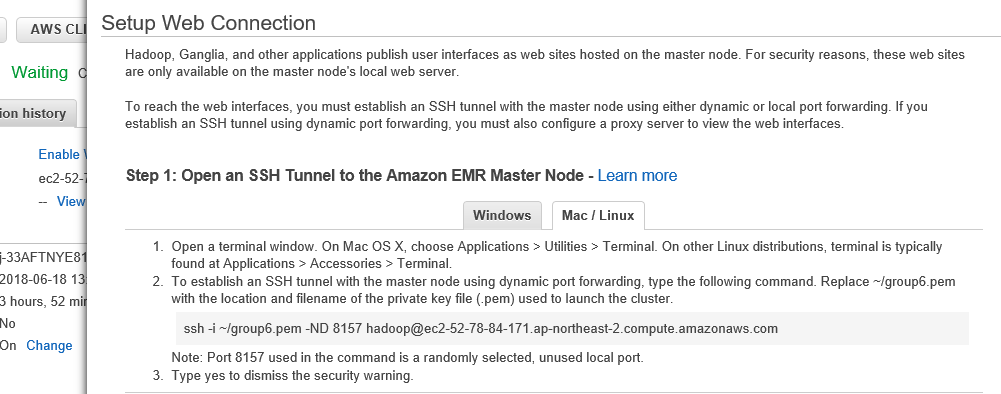
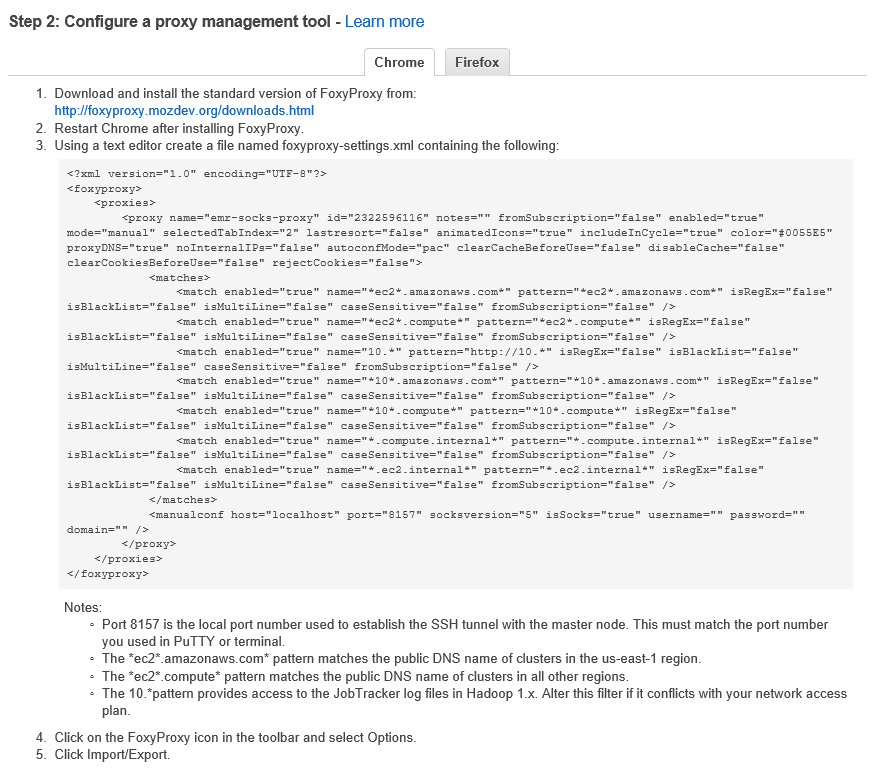
<https://goo.gl/UibNcT>

1. Download Data  
   [https://www.yelp.com/dataset 에서 yelp\_dataset.tar.gz](https://www.yelp.com/dataset%20에서%20yelp_dataset.tar.gz) file 다운로드
2. Dwonload Git Bash  
   <http://www.techoism.com/how-to-install-git-bash-on-windows/>  
   에서 다운받고  
   tar –zxvf yelp\_dataset.tar 하면 압축 풀린다.(business, checkin, photos, review, tip, user json file이 있어야 함.)
3. Download Google AWS CLI  
   <https://s3.amazonaws.com/aws-cli/AWSCLI64.msi>  
   이걸 설치하면 git bash에서 aws 명령이 먹힌다.?
4. JSON File을 S3에 올리기 위해서 Pem file과 key가 필요함.  
   <https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-key-pairs.html#having-ec2-create-your-key-pair>  
   EC2 – Network&Security – Key Pairs 선택 후 Create Key Pair 하면 Pem file을 다운로드 받을 수 있게 된다.(pem file은 ~/.ssh/yours.pem 이런식으로 .ssh 폴더에 있어야 한다.)  
   .ssh 폴더 생성 방법   
    $ssh-keygen 명령어 실행하면 생김.( <https://www.ssh.com/ssh/keygen/>) 계속 enter 치면 됨  
   
5. Key 만들기  
   IAM에서 Create New Access Key 하면 rootkey.csv file을 확인 할 수 있다.   
   
6. Aws configure에 key를 입력한다.  
   $aws configure 명령어를 실행하면 아래와 같이 나옴.  
   access key id랑 secretkey를 위 rootkey.csv file을 보면서 입력하고  
   default region name은 ap-northeast-2로 한다. 이건 aws region 을 검색해보면 해당 지역에 대한 code를 찾을 수 있음.  
   
7. Aws s3 명령어 실행  
   git bash에서 aws s3 ls 를 날리면 s3 에 있는 현재 폴더를 확인 할 수 있다.
8. S3 bucket 만들기  
   amazon s3에 서비스에 들어가서 create bucket 클릭하면 bucket을 만들 수 있음. 이 bucket에 json file을 옮겨야 함. 아래와 같은 명령어로 copy 가능.  
   aws s3 cp review.json s3://yelpgroup6/review/.
9. EMR 클러스터 구동하기  
   아래와 같이 구동하면 됨.  
     
     
     
     
   
10. Tunneling 후 hue를 통해 s3에 있는 json file을 읽어 hive에 테이블로 저장하기.  
    tunneling은 EMR에 master group에 들어가서 master의 inboud에 ssh 선택 후 22번 포트를 오픈해주면 됨.  
      
    hue를 통해 s3에 접속하려면 git bash에서 아래와 같이 web connection을 하면 됨. Pem file 경로를 수정해주고 명령어 실행하면 행이 걸리는 것을 확인 가능.  
    
11. Fireproxy 설치(in Chrome)  
      
    여기까지 하면 hue에서 s3 file을 읽어 쿼리 할 수 있음.
12. S3를 읽어 hive에 table 생성하기.  
    create external table photos2 (photo\_id STRING, business\_id STRING, caption STRING, label STRING)ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'  
    LOCATION 's3://yelpgroup6/photos';
13. More work)s3에서 file을 읽어서 테이블 생성 하던 것을 hdfs로 file을 옮겨서 parquet 테이블을 생성한다.

<http://ec2-52-78-84-171.ap-northeast-2.compute.amazonaws.com:8888/>

2day

Hue에 있는 hive 데이터를 이용하여 데이터 분석하기

1. Exploded table 생성  
   business table에 있는 categories array 컬럼을 flatten 하여 exploded table을 만들도록 한다.  
   query) create table exploded row format serde ‘org.openx.data.jsonserde.JsonSerDe’ stored as textfile location ‘/user/dalgual/yelpgroup6/exploded’ as  
   select \* from business lateral view explode(categories) c as cat\_exploded;  
   1) exploded table에서 10 record 추출하기.  
   query) select business\_id, categories, cat\_exploded from exploded limit 10;  
   2) dataset에서 business의 total number 찾기  
   query) select count(business\_id) from business;
2. Restaurants table 생성  
   exploded table에서 cat\_exploded = ‘Restaurants’ 인 것만 골라서 table을 만들도록 한다.  
   query) create table if not exists restaurants row format serde ‘org.openx.data.jsonserde. JsonSerDe’ stored as textfile as select \* from exploded where cat\_exploded=’Restaurants’;  
   1) restaurants table에서 5개 record만 뽑으면서 cat\_exploded 컬럼을 category로 변경해서 추출하시오.  
   query) select name, review\_count, stars, cat\_exploded category from restaurants limit 5;  
   2) struct형태인 attributes컬럼에 ambience 중 romantic 을 추출하라.  
   query) select name, attributes.ambience.romantic from restaurants limit 5;  
   3) restaurants 테이블의 row num 찾기  
   query) select count(business\_id) from restaurants;
3. Review\_filtered table 생성  
   review table과 restaurants table을 같은 business\_id로 join하여 만들어라.  
   query)  
   1) review\_filtered table에서 restaurants의 review 수를 구하라.
4. User table에서 distinct한 user 수 구하기
5. Elite\_users table 생성  
   1) user table에서 4개 record 추출….  
   2) elite column을 explode해서 elite\_year라는 column을 가진 elite\_users 테이블을 만들어라.

