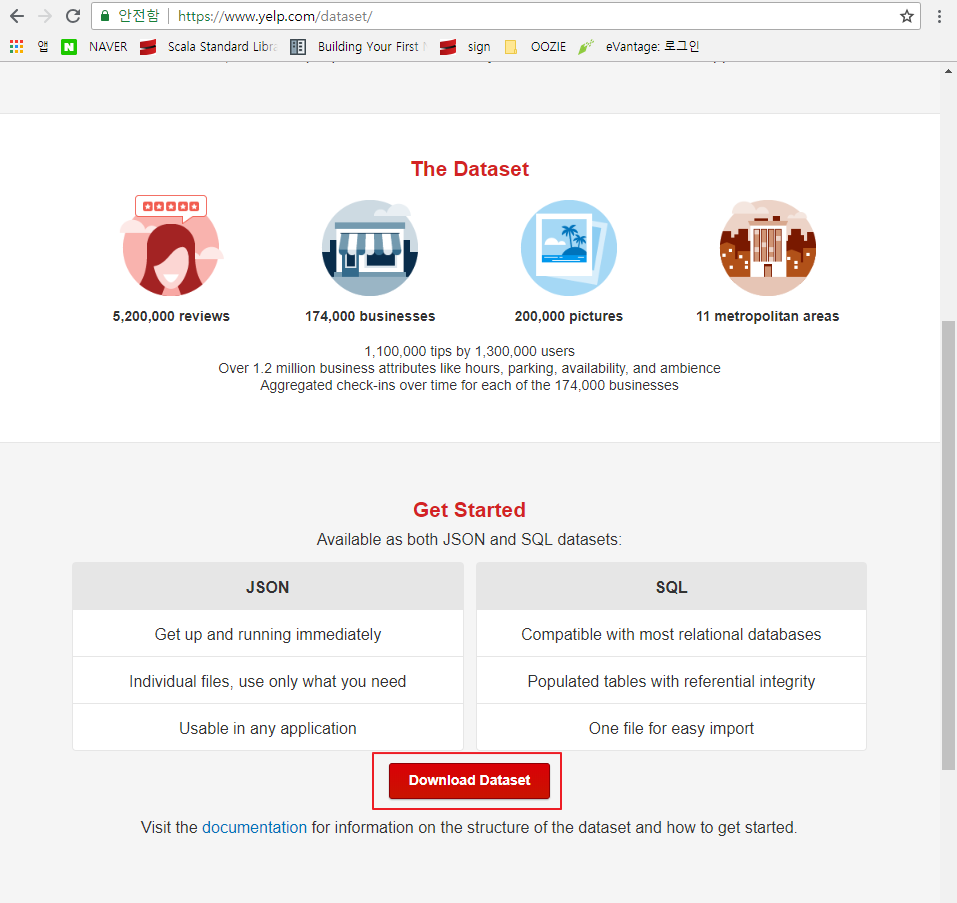
**Big Data Engineering Tutorial 1일차**

3조 신행철 백규진 이소연

1)     Download data and extract it.

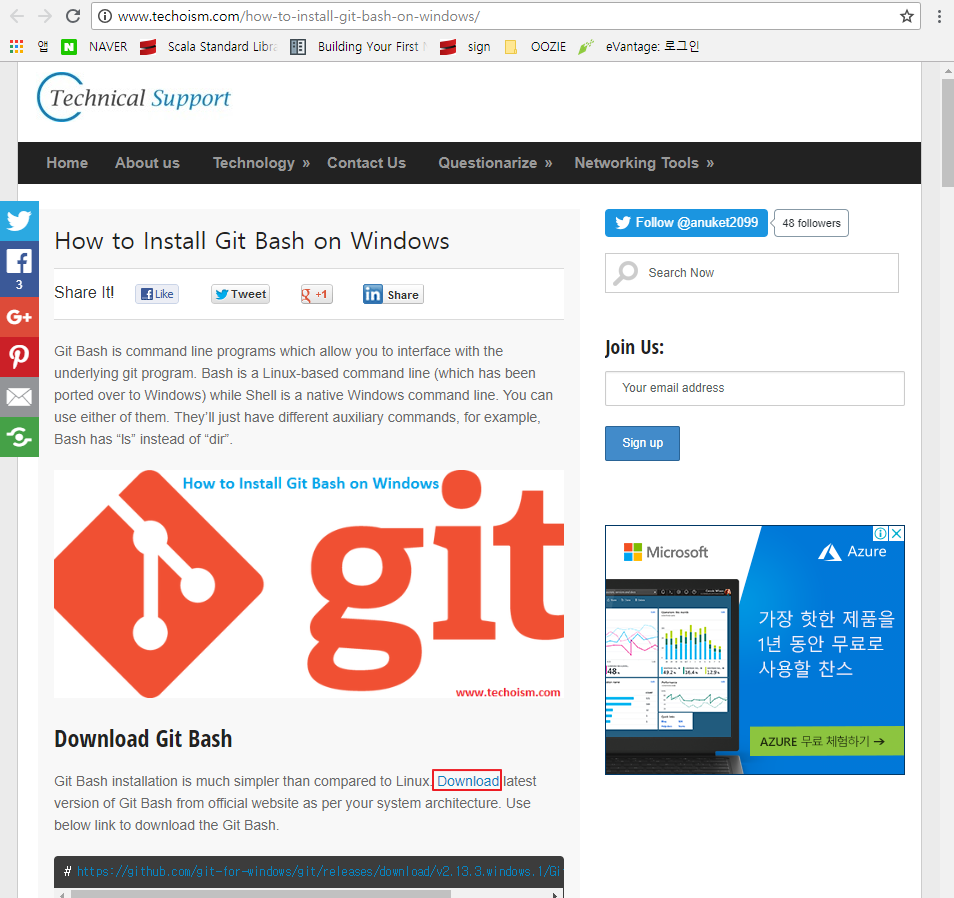
Data set URL and information about the data set:<https://www.yelp.com/dataset> ; Download **yelp\_dataset.tar.gz** file from



2) Putty, Git Bash, AWS CLI : 우리 조는 Git Bash로 진행

1. Use shell terminal or gitbash to untar yelp\_dataset.tar.gz file

• gitbash:  <http://www.techoism.com/how-to-install-git-bash-on-windows/>



Download 버튼을 누르고 설치 진행

• tar -zxvf yelp\_dataset.tar : usb로 진행

•   will generate 6 jsons and some pdf fiels: ./dataset/business.json, checkin.json, photos.json, review.json, tip.json, user.json

2. AWS CLI install

• Google AWS CLI download:<https://aws.amazon.com/cli/>

•   Windows:<https://s3.amazonaws.com/aws-cli/AWSCLI64.msi>

해당 링크 접속 시 바로 다운로드 진행됨

•   Git bash or shell terminal: which aws

• Need your key .pem to connect your S3 to upload json files

* <https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-key-pairs.html#having-ec2-create-your-key-pair>

•   Donwload pem from your AWS account: ~/.ssh/yours.pem

•   Aws Access Key ID, AWSSecretKey

|  |
| --- |
| 1. Open the Amazon EC2 console at <https://console.aws.amazon.com/ec2/>. 2. In the navigation pane, under **NETWORK & SECURITY**, choose **Key Pairs**.      1. Choose **Create Key Pair**.      1. Enter a name for the new key pair in the **Key pair name** field of the **Create Key Pair** dialog box, and then choose **Create**.      1. The private key file is automatically downloaded by your browser. The base file name is the name you specified as the name of your key pair, and the file name extension is .pem. Save the private key file in a safe place. |

• AWS Configure to connect to AWS:

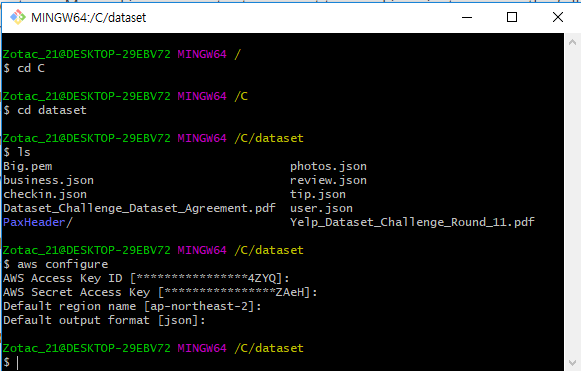
$ aws configure

AWS Access Key ID [None]: AKIAJ4DMPZFTQM3E4ZYQ

AWS Secret Access Key [None]: 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

Default region name [None]: ap-northeast-2 [서울]

Default output format [None]: json



# pem file should be in .ssh folder:  <https://www.ssh.com/ssh/keygen/>

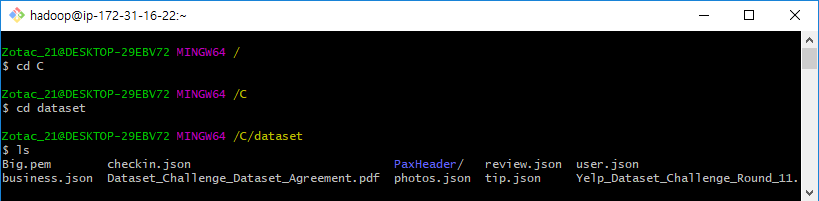
$ssh-keygen

# AWS CLI: <https://aws.amazon.com/cli/>

# AWS Region: <https://docs.aws.amazon.com/ko_kr/general/latest/gr/rande.html>

#

|  |  |
| --- | --- |
| 아시아 태평양(서울) | ap-northeast-2 |

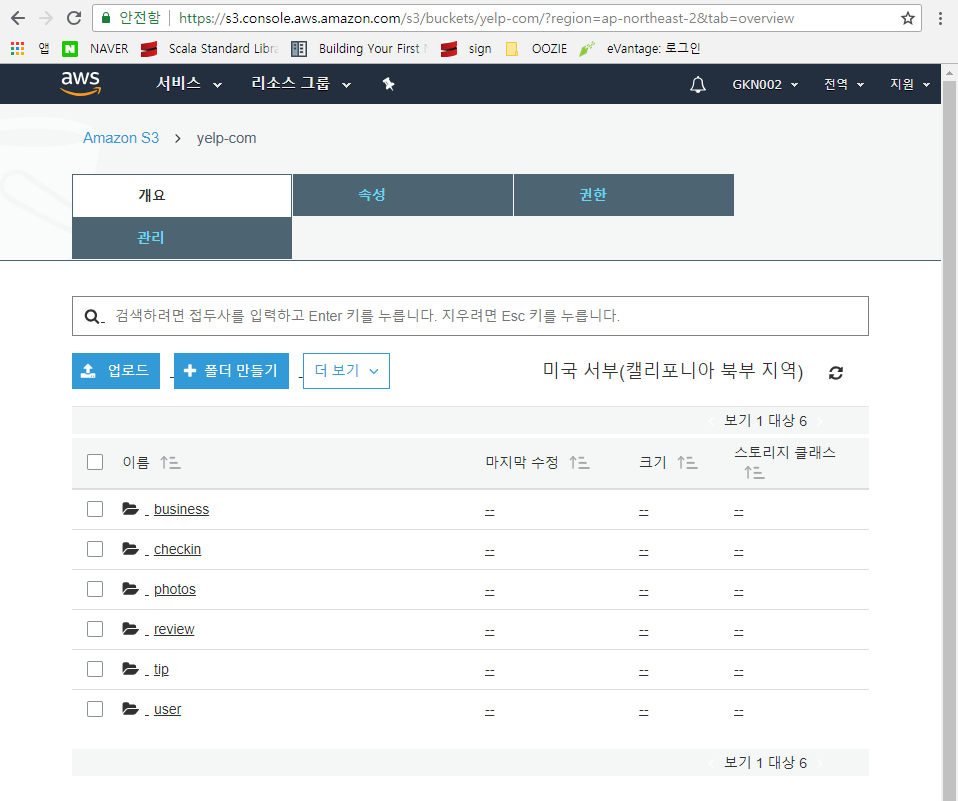


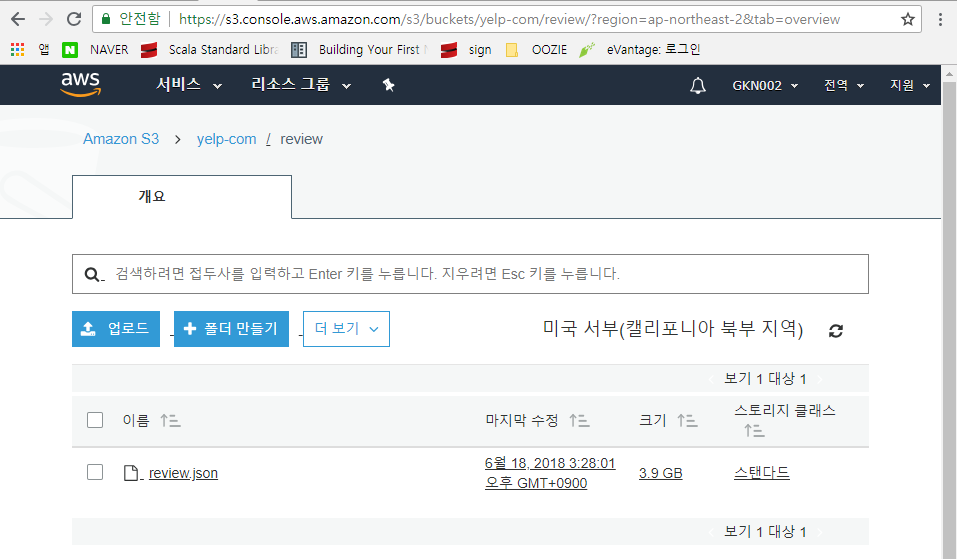
$ aws s3 cp photos.json s3://yelp/photos

upload: .\photos.json to s3://yelp/photos/photos.json

$ aws s3 cp review.json s3://yelp/review

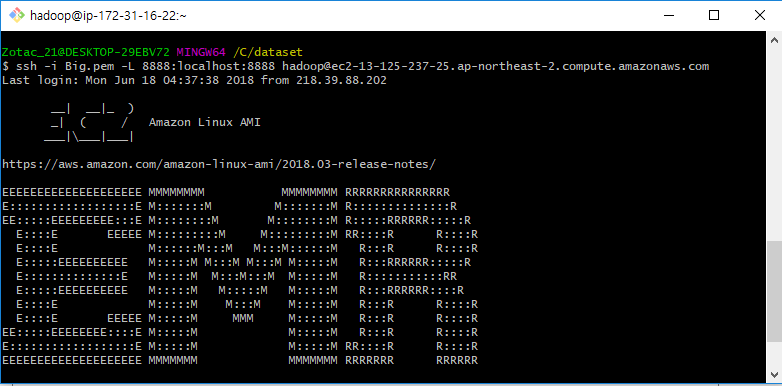
: 파일 copy 완료 후 아래 링크에서 파일을 확인할 수 있다.



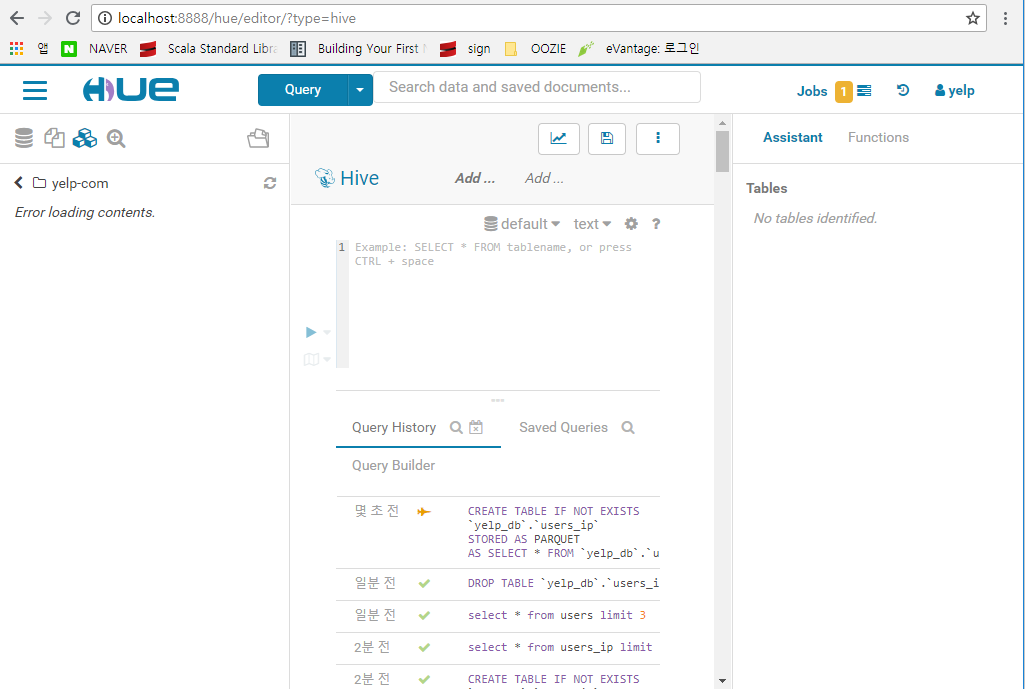


3) 하이브 external table, Json SerDe, S3 데이터 로드

- Hue: 8888

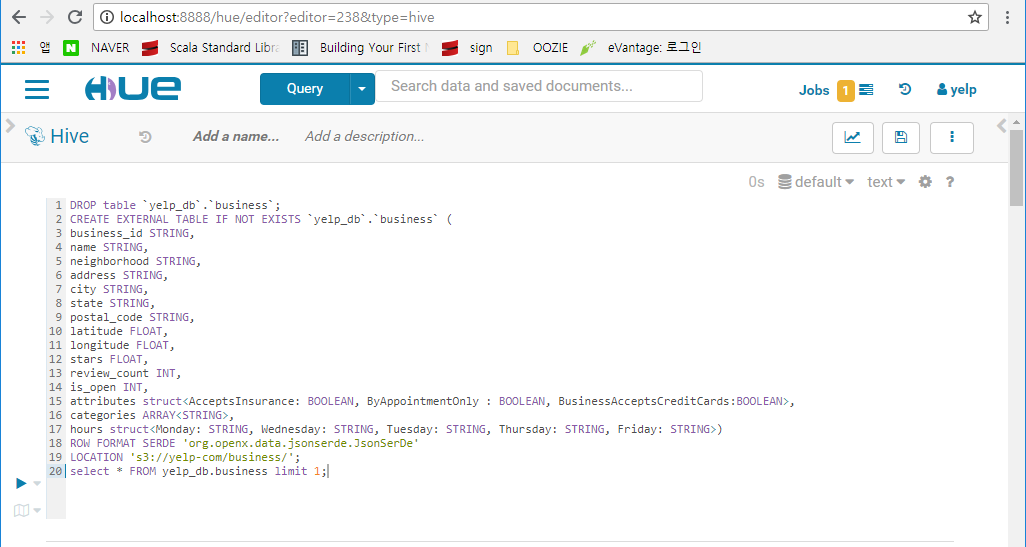


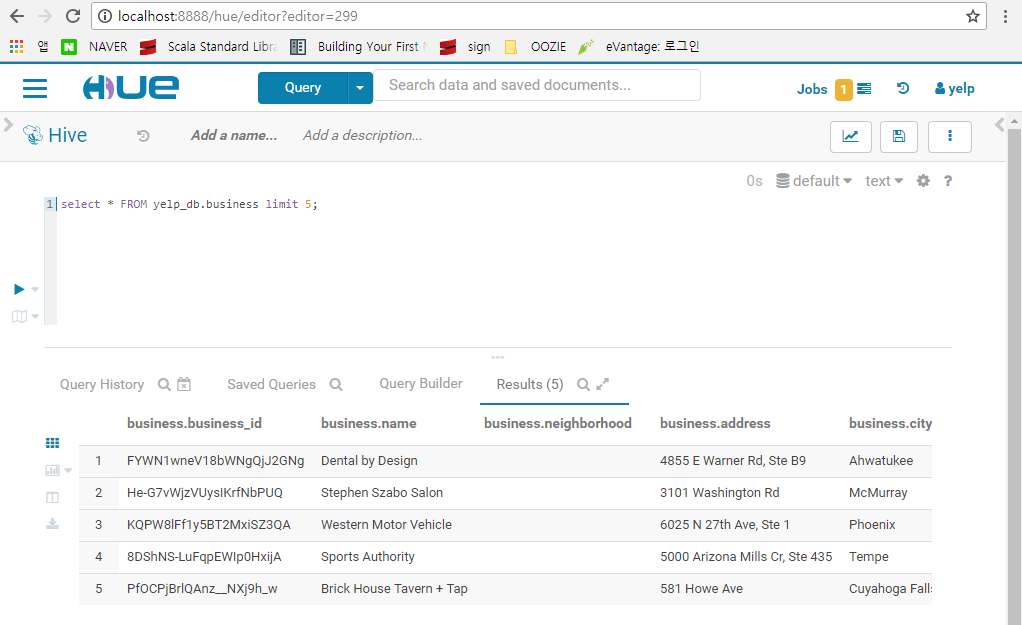
* Hue



* Review table : 하이브 external table, Json SerDe, S3 데이터 로드

[business table]





* Review, user, tip도 같은 방식으로 진행
* Query

|  |
| --- |
| DROP table `yelp\_db`.`business`;  CREATE EXTERNAL TABLE IF NOT EXISTS `yelp\_db`.`business` (  business\_id STRING,  name STRING,  neighborhood STRING,  address STRING,  city STRING,  state STRING,  postal\_code STRING,  latitude FLOAT,  longitude FLOAT,  stars FLOAT,  review\_count INT,  is\_open INT,  attributes struct<AcceptInsurance: BOOLEAN, ByAppointmentOnly : BOOLEAN, BusinessAcceptsCreditCards:BOOLEAN>,  categories ARRAY<STRING>,  hours struct<Monday: STRING, Wednesday: STRING, Tuesday: STRING, Thursday: STRING, Friday: STRING>)  ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'  LOCATION 's3://yelp-com/business/'; |

|  |
| --- |
| DROP table `yelp\_db`.`users`;  CREATE EXTERNAL TABLE IF NOT EXISTS `yelp\_db`.`users` (  `user\_id` STRING,  `name` STRING,  `review\_count` INT,  `yelping\_since` STRING,  `friends` ARRAY<STRING>,  `useful` INT,  `funny` INT,  `cool` INT,  `fans` INT,  `elite` ARRAY<String>,  `average\_stars` FLOAT,  `compliment\_hot` INT,  `compliment\_more` INT,  `compliment\_profile` INT,  `compliment\_cute` INT,  `compliment\_list` INT,  `compliment\_note` INT,  `compliment\_plain` INT,  `compliment\_cool` INT,  `compliment\_funny` INT,  `compliment\_writer` INT,  `compliment\_photos` INT)  ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'  LOCATION 's3://yelp-com/user/'; |

|  |
| --- |
| DROP table `yelp\_db`.`tip` ;  CREATE external TABLE IF NOT EXISTS `yelp\_db`.`tip` (  `user\_id` STRING,  `business\_id` STRING,  `text` STRING,  `date` STRING,  `likes` INT)  ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'  LOCATION 's3://yelp-com/tip'; |

|  |
| --- |
| DROP table `yelp\_db`.`review`;  CREATE EXTERNAL TABLE IF NOT EXISTS `yelp\_db`.`review` (  `review\_id` STRING ,  `business\_id` STRING,  `user\_id` STRING,  `stars` INT ,  `date` STRING ,  `text` STRING ,  `useful` INT ,  `funny` INT ,  `cool` INT )  ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'  LOCATION 's3://yelp-com/review/'; |

* 인터널, PARQUET 타입으로 저장하기

|  |
| --- |
| create table if not exists business\_p  stored as parquet  as select \* from business; |

|  |
| --- |
| create table if not exists users\_p  stored as parquet  as select \* from users; |

|  |
| --- |
| create table if not exists tip\_p  stored as parquet  as select \* from tip; |

|  |
| --- |
| create table if not exists review\_p  stored as parquet  as select \* from review; |