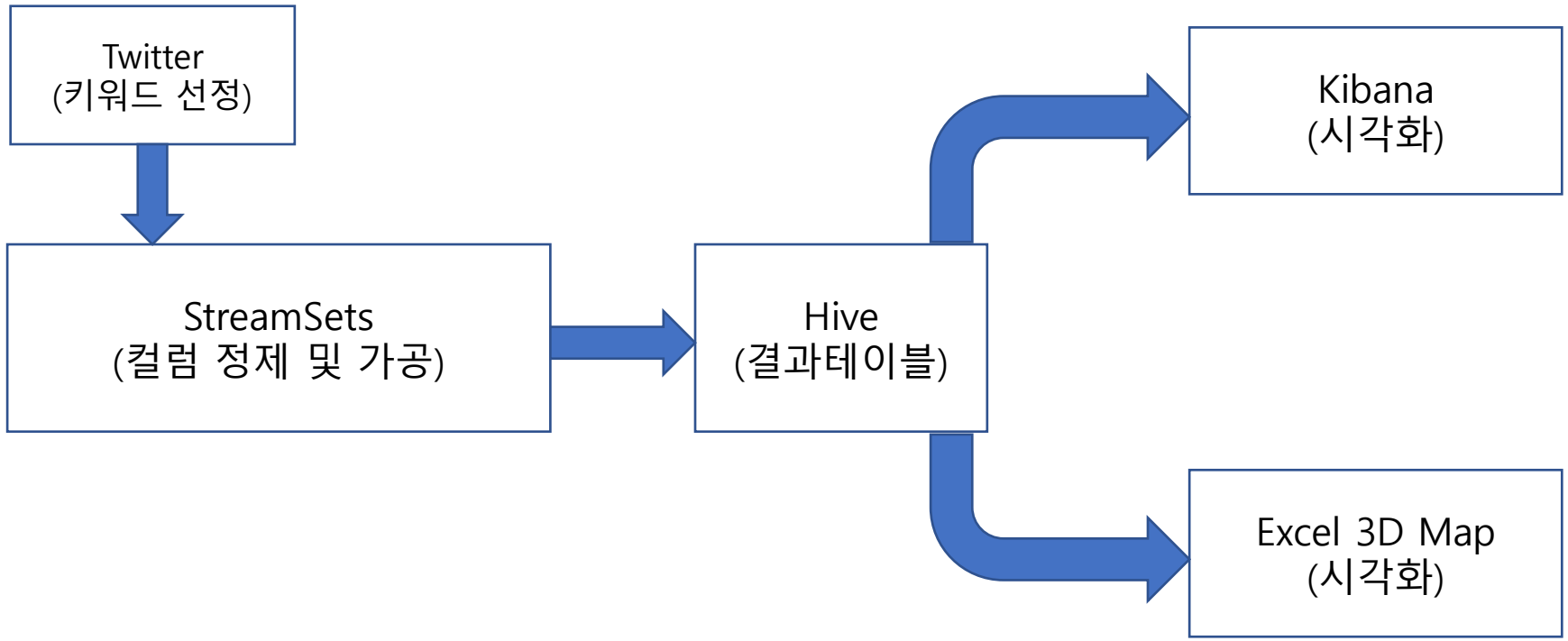
Twitter 감성분석 프로젝트

원본 : https://github.com/KimHyungkeun/Data_Engineer_Project/blob/main/Twitter_Sentiment_Analysis.pdf

(이삭엔지니어링 사내 프로젝트 / 2018년 11월 ~ 12월 진행)

Index

1. 프로젝트 구조

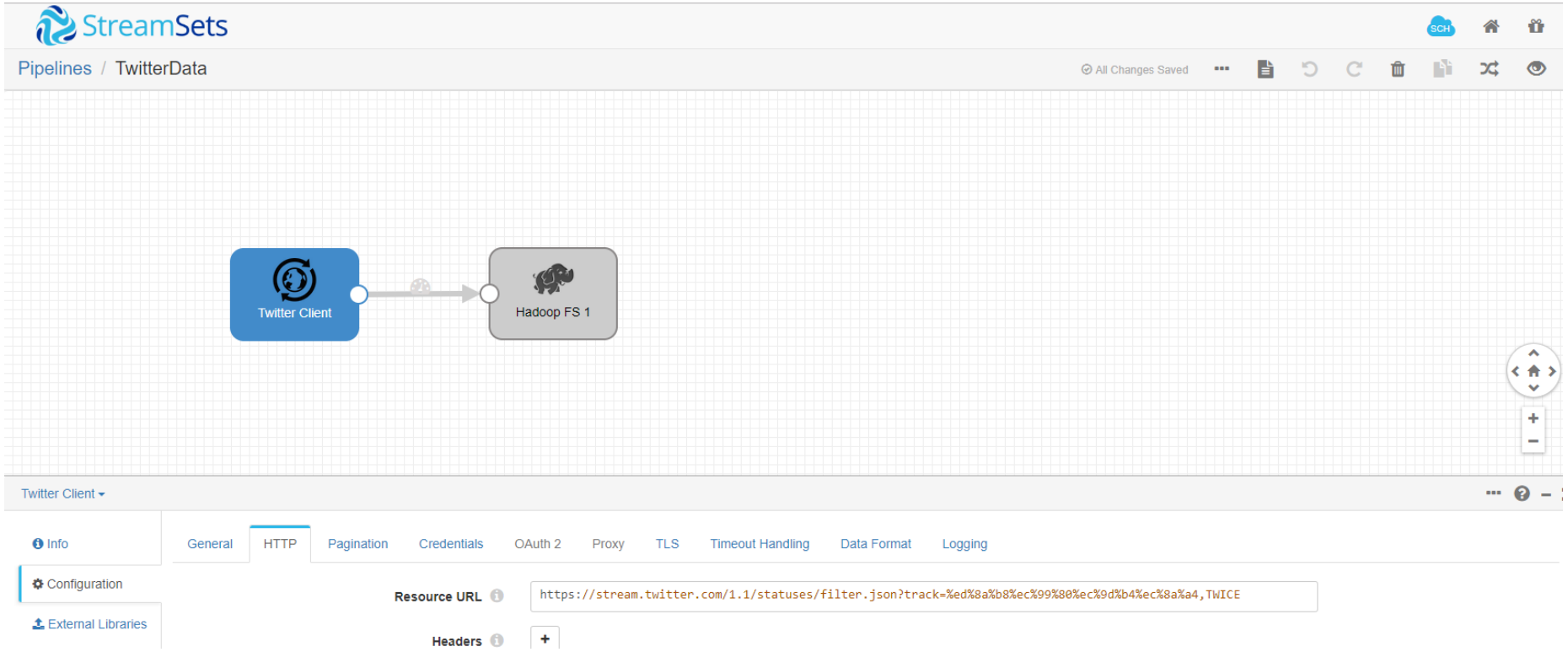

Twitter
(키워드 선정)
StreamSets
(컬럼 정제 및 가공)
Hive
(결과테이블)
Kibana
(시각화)
Excel 3D Map
(시각화)

2. 사용 언어 및 툴

- Cloudera Manager 5.x
- Hive 2.x
- Hue
- StreamSets
- Kibana 6.x

3. Streamsets로 트위터 데이터 받기

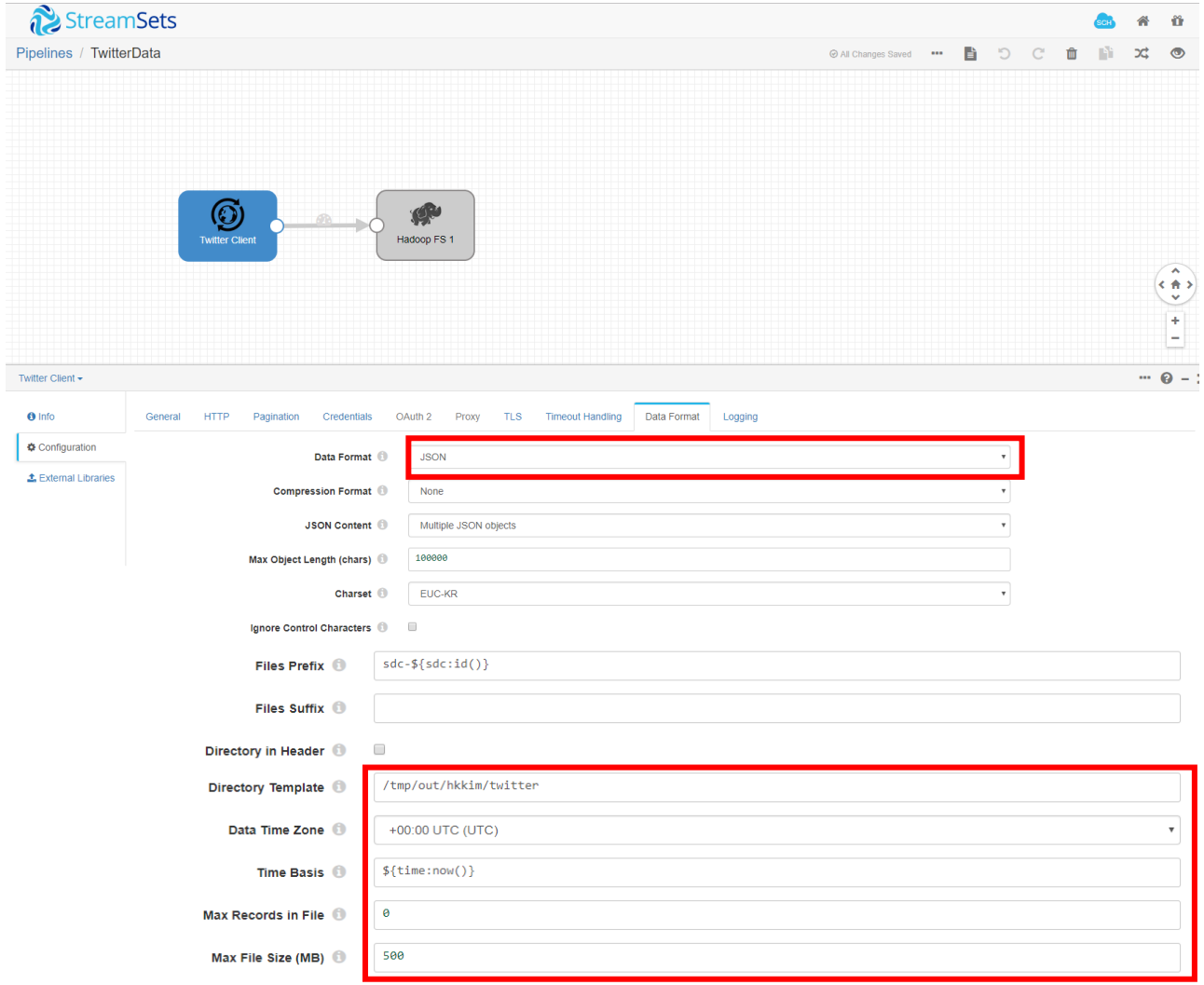
(1) 트위터 데이터를 받아오는 파이프라인



Resource : https://stream.twitter.com/1.1/statuses/filter.json?track=%ed%8a%b8%ec%99%80%ec%9d%b4%ec%8a%a4,TWICE

3. Streamsets로 트위터 데이터 받기

(1) 트위터 데이터를 받아오는 파이프라인



* Data Format : ***JSON***

* File당 사이즈 : ***500MB***

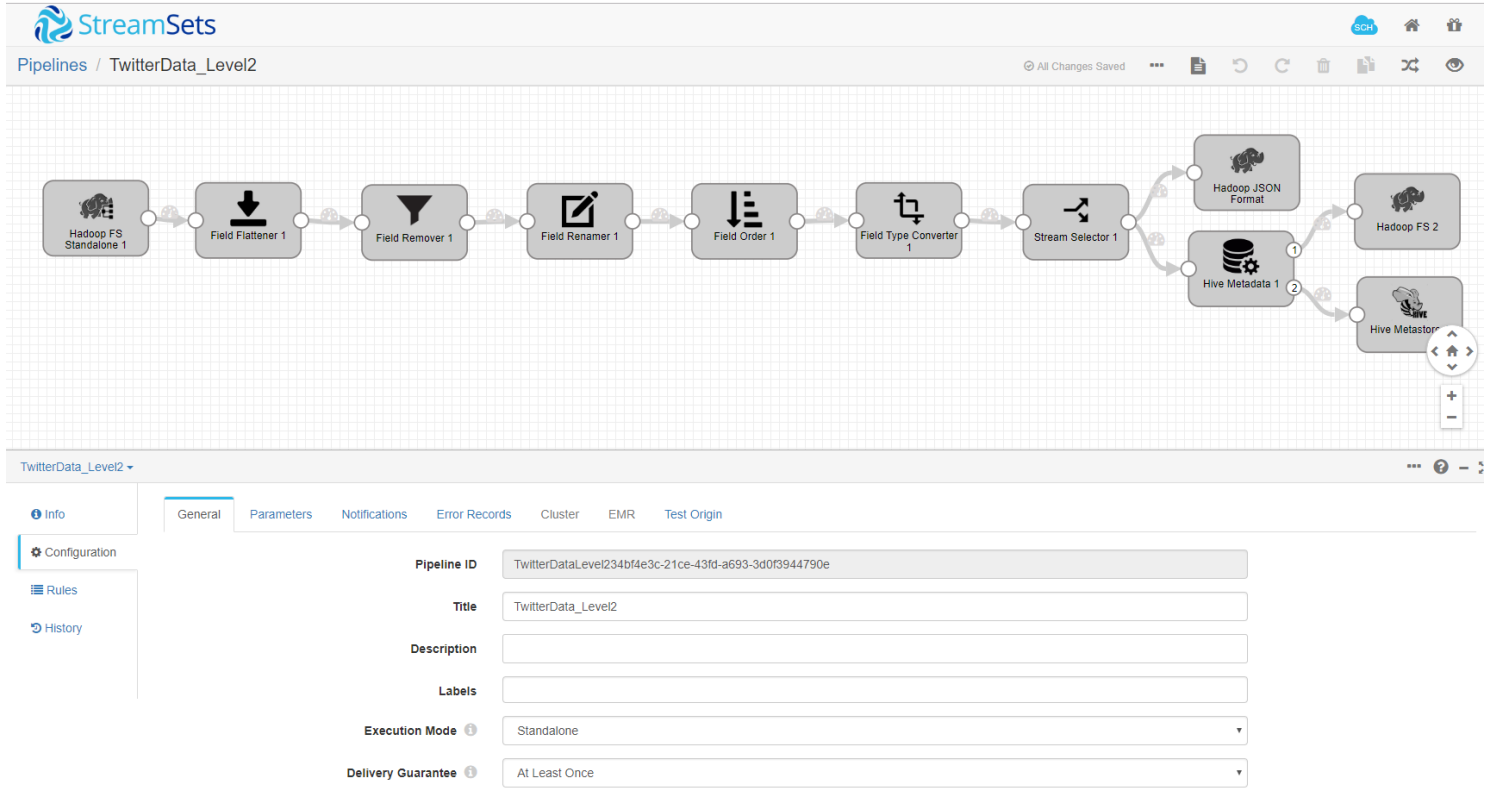
* Streaming한 데이터를 저장하는 위치 :

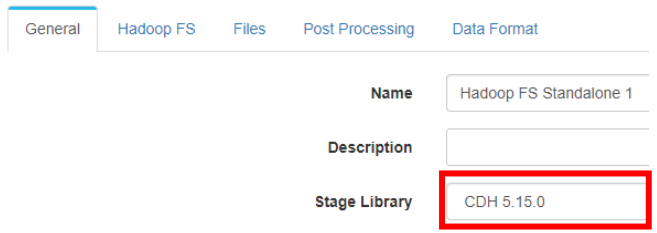
/tmp/out/hkkim/twitter

3. Streamsets로 트위터 데이터 받기

(2) 받아온 트위터들을 다시 재정리 하는 파이프라인

1) 받아온 Twitter Data들을 JSON파일로 정리하고, 다시 JSON의 Field들 중에서 몇몇 요소들만 골라낸다.



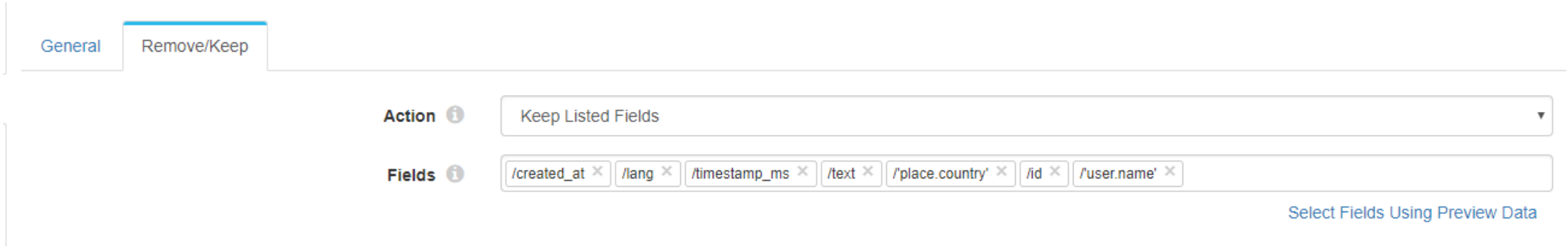


"General" 탭 -> "Stage Library" (CDH 5.15.0)

3. Streamsets로 트위터 데이터 받기

(2) 받아온 트위터들을 다시 재정리 하는 파이프라인

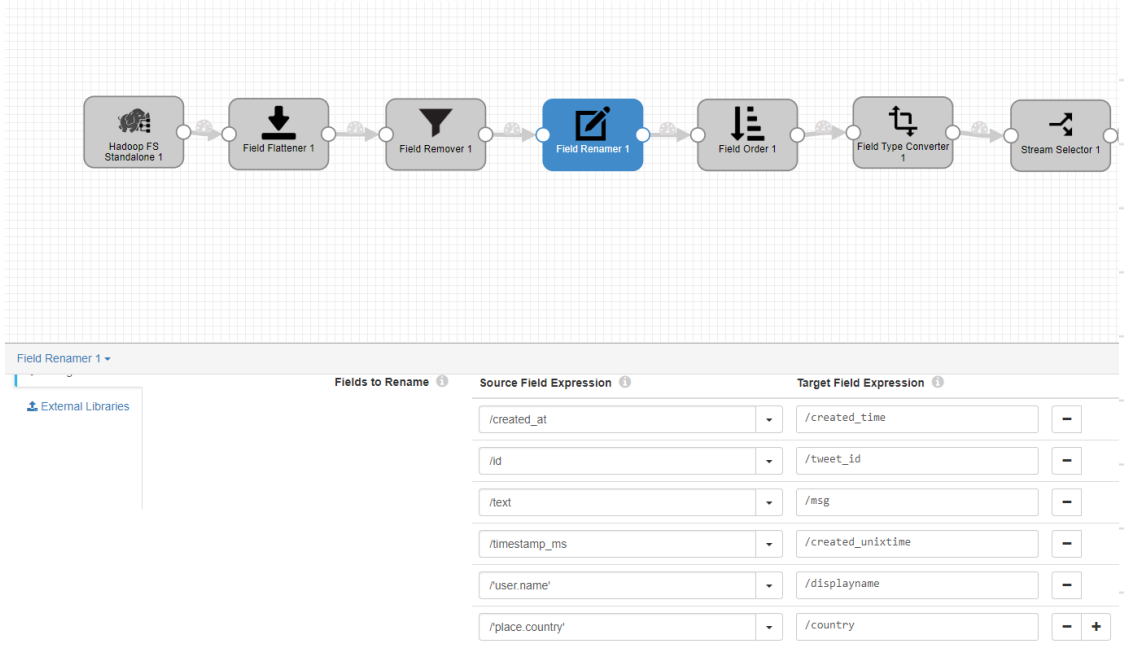
2) "Field Remover"에서 하단에 기입한 Field들만 남겨놓는다. 붉은색 field는 대부분이 NULL이지만, 가끔씩 값이 들어있는 경우가 있다.
(/id, /text, /created_at, /timestamp_ms,
/lang, /'user.location', /'user.screen_name',/'user.name', /'user.id', /'extended_tweet.full_text', /'retweeted_status.text',
/'retweet_status.extended_tweet.full_text', /'place.location', /'place.country')



3. Streamsets로 트위터 데이터 받기

(2) 받아온 트위터들을 다시 재정리 하는 파이프라인

3) Field Renamer에서 Field 이름을 변경한다.



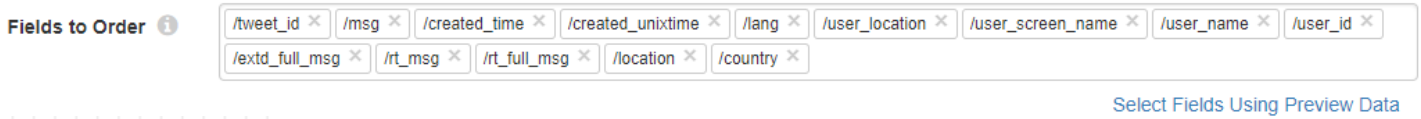
1	tweet_id
2	msg
3	created_time
4	created_unixtime
5	lang
6	user_location
7	user_screen_name
8	user_name
9	user_id
10	extd_full_msg
11	rt_msg
12	rt_full_msg
13	location
14	country

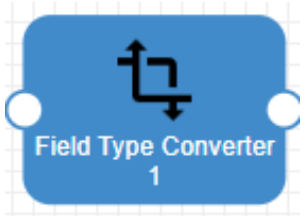
/Id -> tweet_id,
/text -> msg,
/'created_at' -> created_time,
/'timestamp_ms' -> created_unixtime
/'place.location' -> location,
/'place.country' -> country
/'extended_tweet.full_text' -> extd_full_msg,
/'retweeted_status.text' -> rt_msg
/'retweet_status.extended_tweet.full_text' -> rt_full_msg
나머지 Field에는 '.' 가 들어가 있다면 '_'로 대체한다.
(예 : **/user.location** -> /user_location)

3. Streamsets로 트위터 데이터 받기

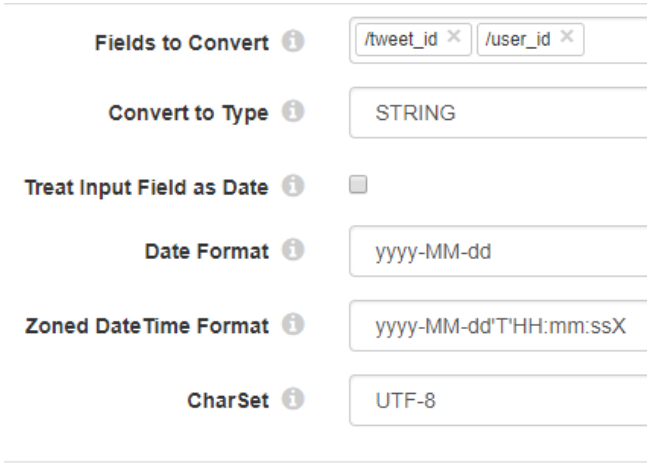
(2) 받아온 트위터들을 다시 재정리 하는 파이프라인

4) "Field Order" : 하단 순서대로 Field를 정렬한다.





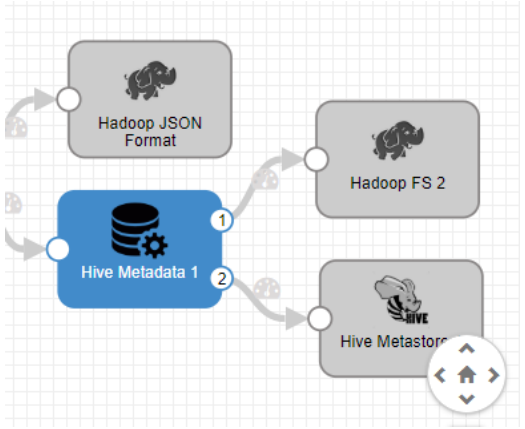
5) "Field Type Converter" : /tweet_id, /user_id를 STRING으로 바꾸어 준다.
(STRING으로 바꾸지 않으면 Preview시 Type 에러로 간주한다.)



3. Streamsets로 트위터 데이터 받기

(2) 받아온 트위터들을 다시 재정리 하는 파이프라인

6) 정리한 데이터들을 JSON파일로도 저장해두고, Hive Table도 생성해둔다.
(Table명 : tweets_text2)



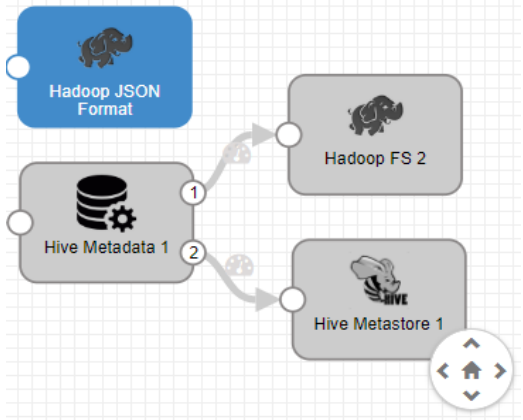
"General" 탭 -> "Stage Library" (CDH 5.15.0)

"Hive" 탭 -> "JDBC URL" **jdbc:hive2://10.100.3.210:10000/(Database)**

"Table" 탭 -> "Database Expression" (DB이름), "Table Name" (테이블 이름 짓기)

"Data Format" 탭 -> "Data Format" (Avro)

(Hive DB는 반드시 먼저 생성되어 있어야한다!!)



"General" 탭 -> "Stage Library" (CDH 5.15.0)

"Hadoop FS" 탭 -> "Hadoop FS Configuration Directory" (hadoop-conf)

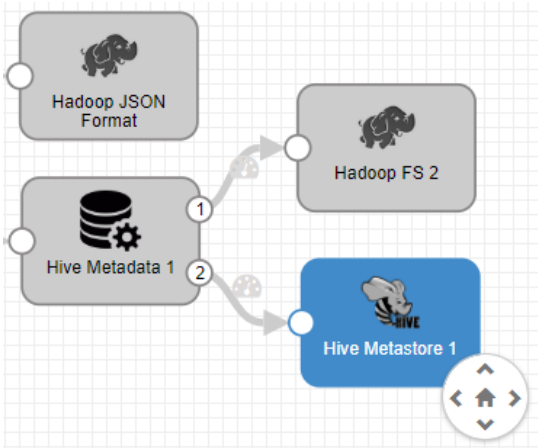
"Output Files" 탭 -> "Directory Template" (저장하고 싶은 곳의 경로를 넣는다.)

"Data Format"탭 -> "Data Format" (JSON)

3. Streamsets로 트위터 데이터 받기

(2) 받아온 트위터들을 다시 재정리 하는 파이프라인

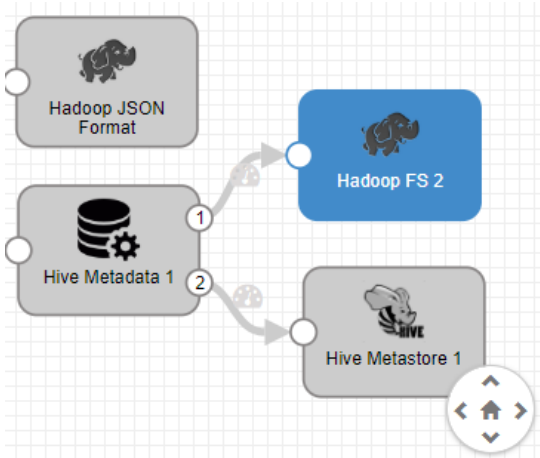
6) 정리한 데이터들을 JSON파일로도 저장해두고, Hive Table도 생성해둔다.
(Table명 : tweets_text2)



"General" 탭 -> "Stage Library" (CDH 5.15.0)

"Hive" 탭 -> "JDBC URL" **jdbc:hive2://10.100.3.210:10000/(Database)**

"Advanced" 탭 -> "Stored as Avro" (체크하기)



"General" 탭 -> "Stage Library" (CDH 5.15.0)

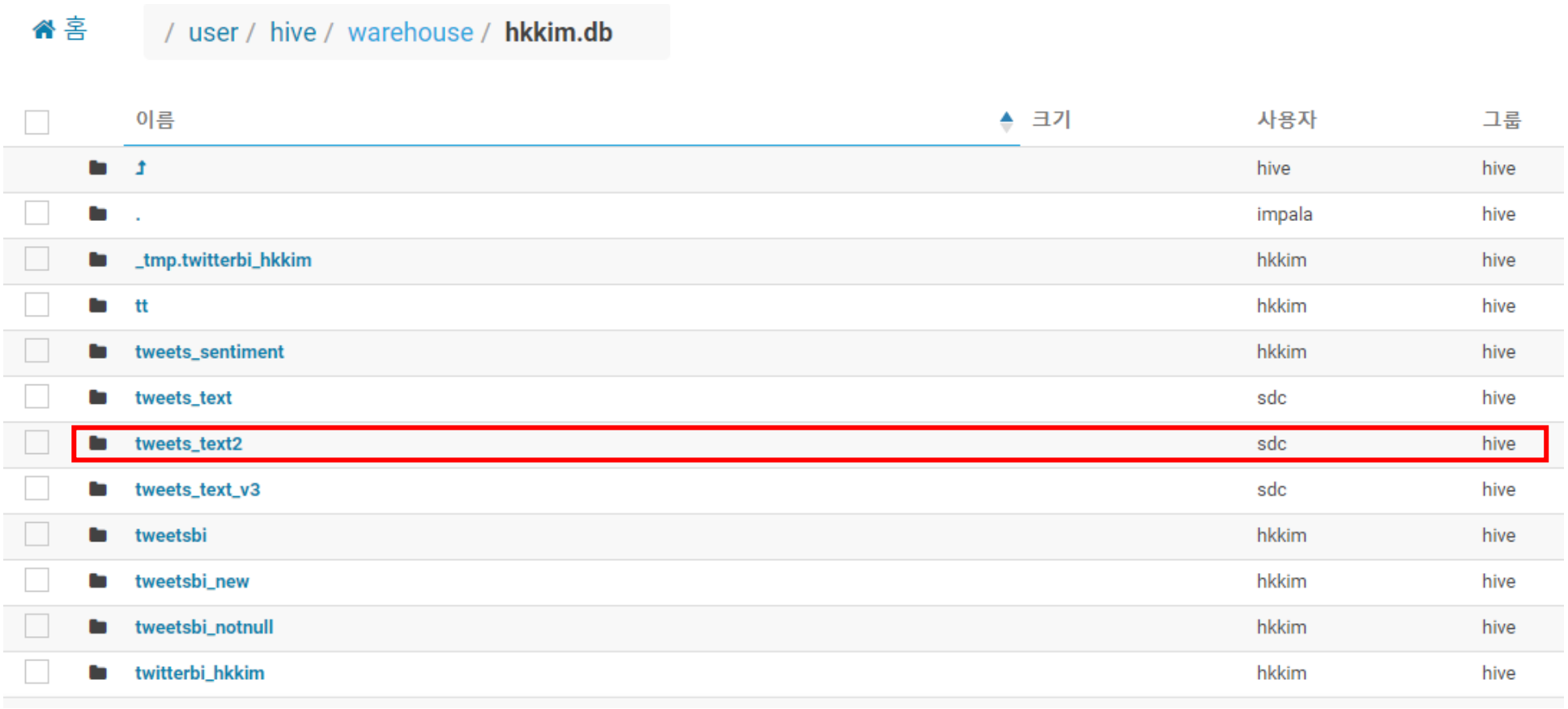
"Hadoop FS" 탭 -> "Hadoop FS Configuration Directory" (hadoop-conf)

"Output Files" 탭 -> "Directory in Header" (체크하기)

"Data Format"탭 -> "Data Format" (Avro)

3. Streamsets로 트위터 데이터 받기

(2) 받아온 트위터들을 다시 재정리 하는 파이프라인



4. Hue에서 Hive Table 만들기

(1) Hue에서 Hive Table 만들기

1) dictionary 테이블 만들기

(영어 단어에 대해 긍정, 중립, 부정 척도를 판별해주는 table)

```
CREATE EXTERNAL TABLE if not exists dictionary
( type string,
length int,
word string,
pos string,
stemmed string,
polarity string )
ROW FORMAT DELIMITED FIELDS TERMINATED BY 't'
STORED AS TEXTFILE LOCATION '/user/hkkim/twitter/data/tables/dictionary';
```

2) tweets_clean 뷰 만들기

(사용자가 만든 tweets_text2에서 tweet_id, timestamp, message, country만 뽑은 View를 만든다.)

```
CREATE VIEW IF NOT EXISTS tweets_clean AS
SELECT tweet_id,
cast( from_unixtime( cast ( substring ( created_unixtime,1,10) as bigint ) ) as timestamp) ts, msg, country
FROM tweets_text2;
```

4. Hue에서 Hive Table 만들기

(1) Hue에서 Hive Table 만들기

3) l1, l2, l3 뷰 만들기

(l1 : msg의 문장들을 단어 하나하나씩 쪼갠다.)

```
create view IF NOT EXISTS l1 as select tweet_id, words from tweets_text2 lateral view
explode(sentences(lower(msg))) dummy as words;
```

(l2 : l1에서 쪼갠 단어들 각각에 대해 나열한 view를 만든다.)

```
create view IF NOT EXISTS l2 as select tweet_id, word from l1 lateral view explode( words )
dummy as word;
```

(l3 : l2에 나열된 각 단어들에 대한 긍정, 중립, 부정의 척도를 1,0,-1로 나타낸다.)

```
create view IF NOT EXISTS l3 as select
tweet_id,
l2.word,
case d.polarity
when 'negative' then -1
when 'positive' then 1
else 0 end as polarity
from l2 left outer join dictionary d on l2.word = d.word;
```

4. Hue에서 Hive Table 만들기

(1) Hue에서 Hive Table 만들기

4) tweets_sentiment 테이블 만들기

(tweet_id. 즉, 사용자 별로 msg의 내용이 긍정, 중립, 부정 인지를 나타내는 표이다.)

```
create table IF NOT EXISTS tweets_sentiment  as select tweet_id,
case
when sum( polarity ) > 0 then 'positive'
when sum( polarity ) < 0 then 'negative'
else 'neutral' end as sentiment
from l3 group by tweet_id;
```

5) tweetsbi 테이블 만들기

(앞서 만들었던 tweets_clean view와 tweets_sentiment table을 서로 join 시킨다.)

```
CREATE TABLE IF NOT EXISTS tweetsbi AS SELECT
t.*,
case s.sentiment
when 'positive' then 2
when 'neutral' then 1
when 'negative' then 0
end as sentiment
FROM tweets_clean t LEFT OUTER JOIN tweets_sentiment s on t.tweet_id = s.tweet_id;
```

4. Hue에서 Hive Table 만들기

(1) Hue에서 Hive Table 만들기

6) tweetsbi_notnull 테이블 만들기

(sentiment 및 country에 NULL이 다수 포함되어 있어 NULL을 없앤 table을 다시 만듦.)

```
create table tweetsbi_notnull as

select * from tweetsbi where country != 'NULL' and sentiment is not null
```

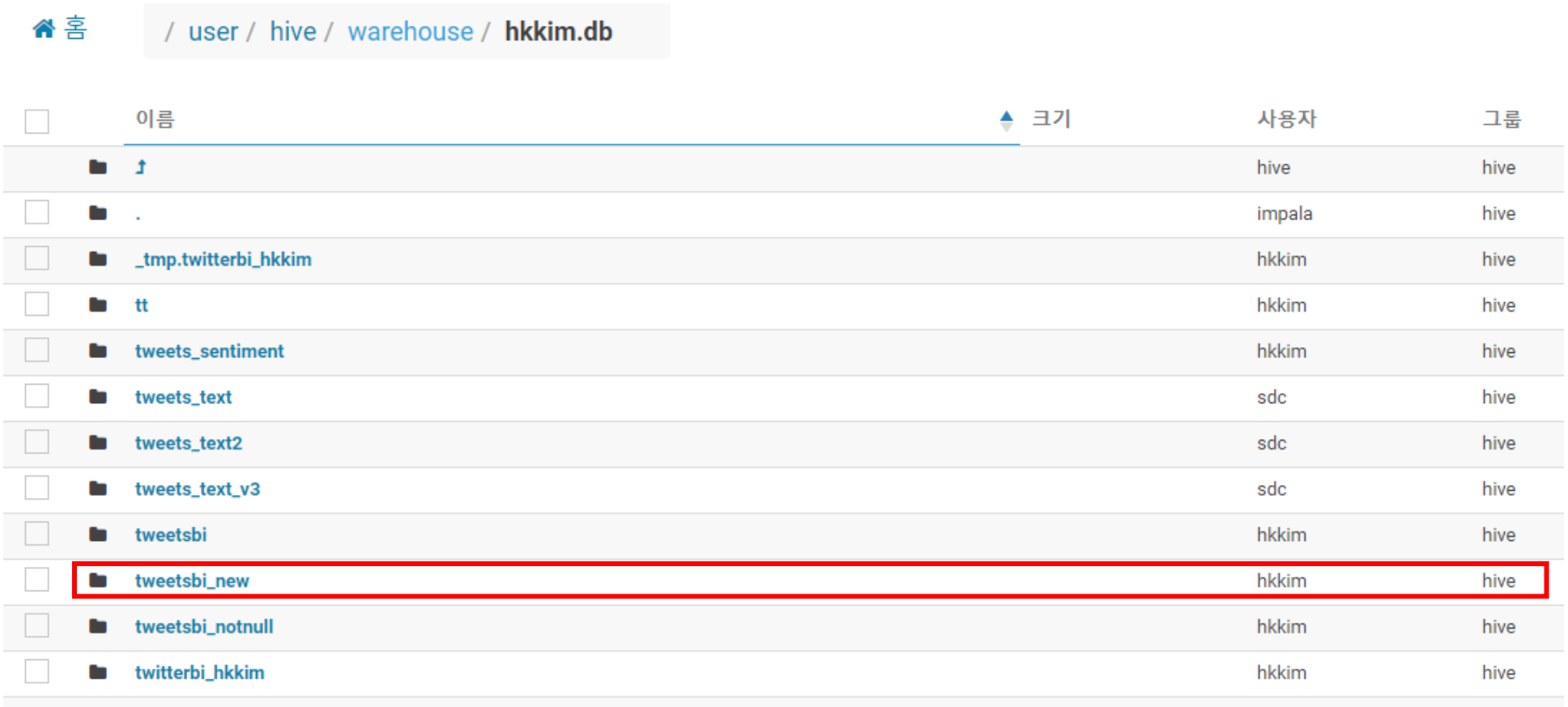
7) tweetsbi_new 테이블 만들기

(tweets_text2의 lang Column도 포함시킨 table을 만듦.)

```
create table tweetsbi_new as

select tweetsbi_notnull.*, tweets_text2.lang

from tweetsbi_notnull left outer join tweets_text2

on tweetsbi_notnull.tweet_id = tweets_text2.tweet_id
```

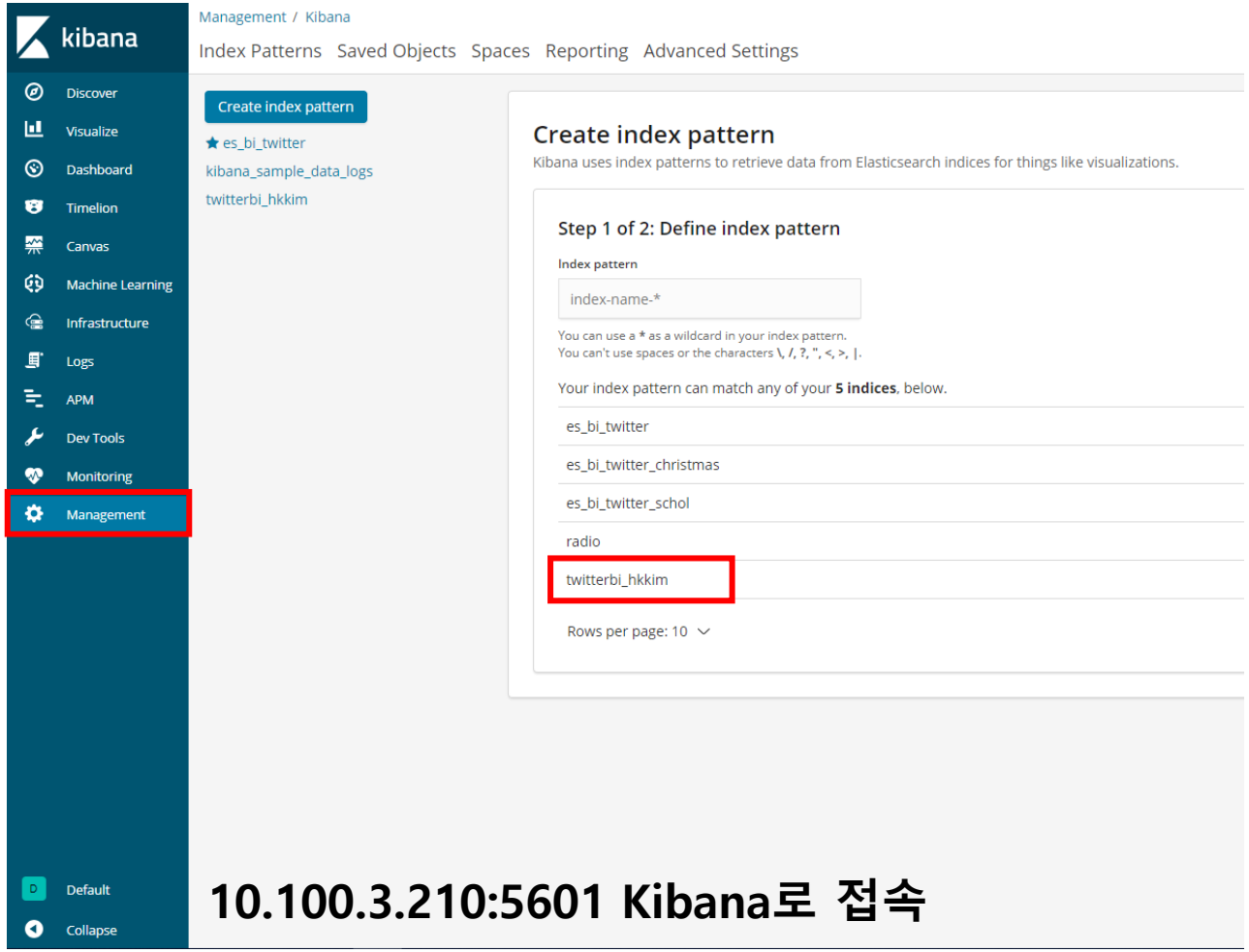
4. Hue에서 Hive Table 만들기

(1) Hue에서 Hive Table 만들기



5. Hive Table을 Kibana에서 분석

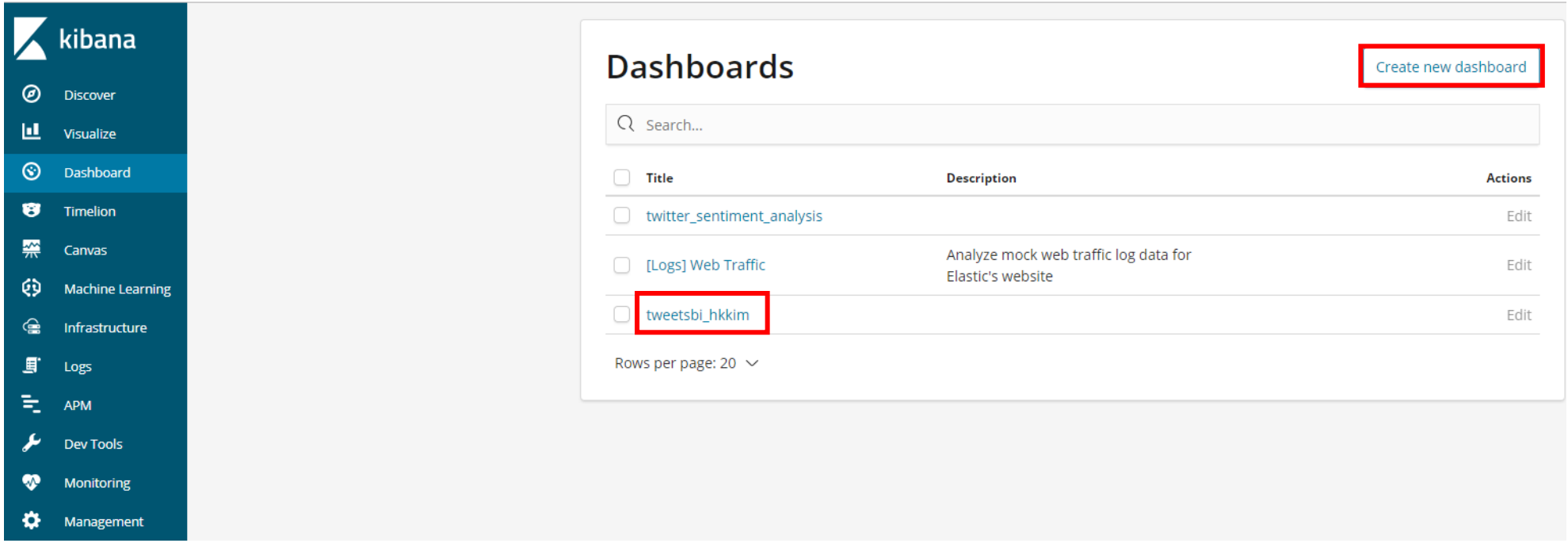
(1) Kibana에서 사용할 external 테이블을 생성




```
CREATE EXTERNAL TABLE twitterbi_hkkim (
tweet_id string,
ts string,
msg string,
country string,
sentiment int,
lang string
)
STORED BY 'org.elasticsearch.hadoop.hive.EsStorageHandler'
TBLPROPERTIES('es.resource' = 'twitterbi_hkkim/doc','es.nodes'='10.100.3.210');
INSERT OVERWRITE TABLE twitterbi_hkkim
SELECT * FROM tweetsbi_new;
```

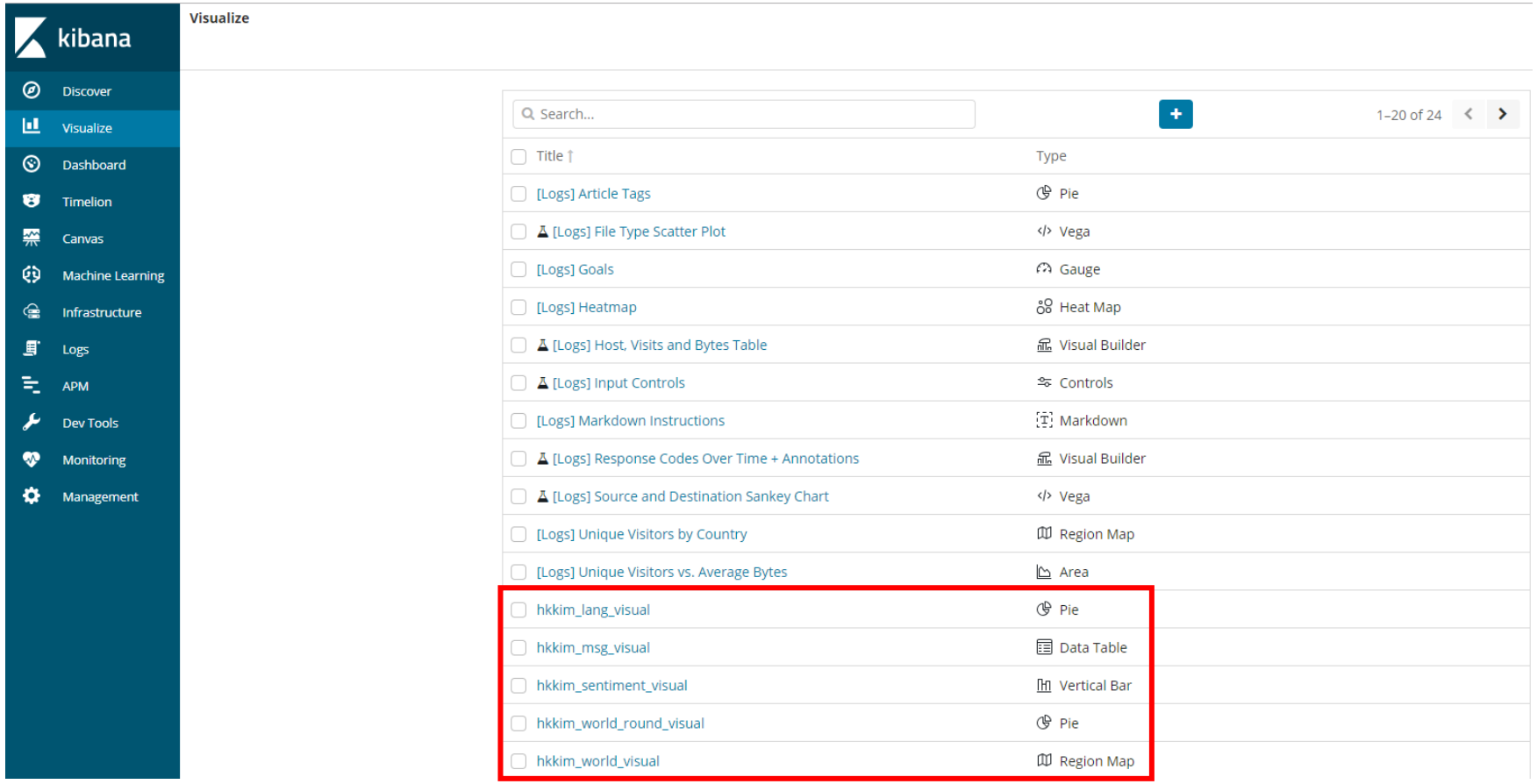
5. Hive Table을 Kibana에서 분석

(2) Dashboard 생성



5. Hive Table을 Kibana에서 분석

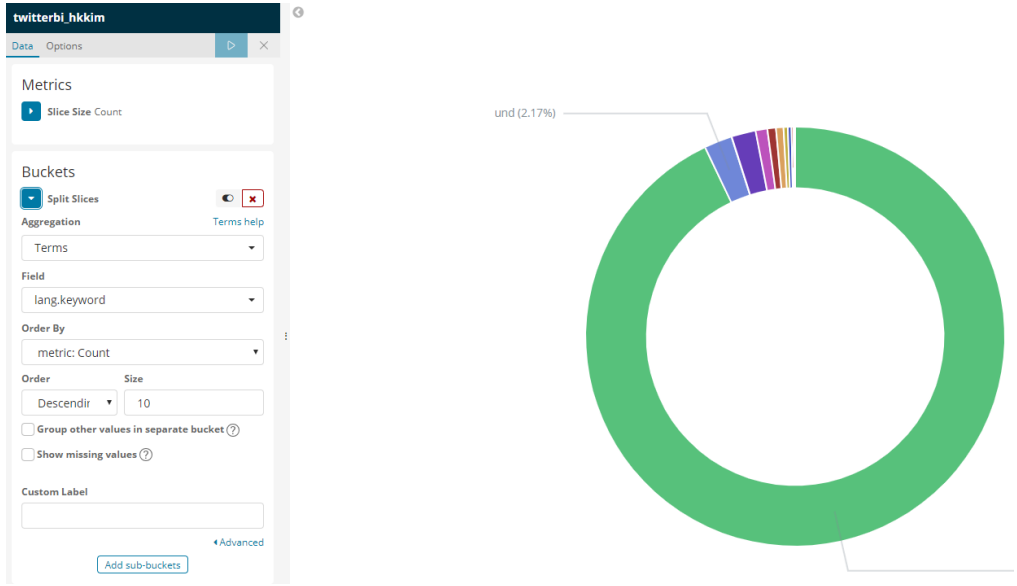
(3) Dashboard에 들어갈 Visualize(그래프) 생성



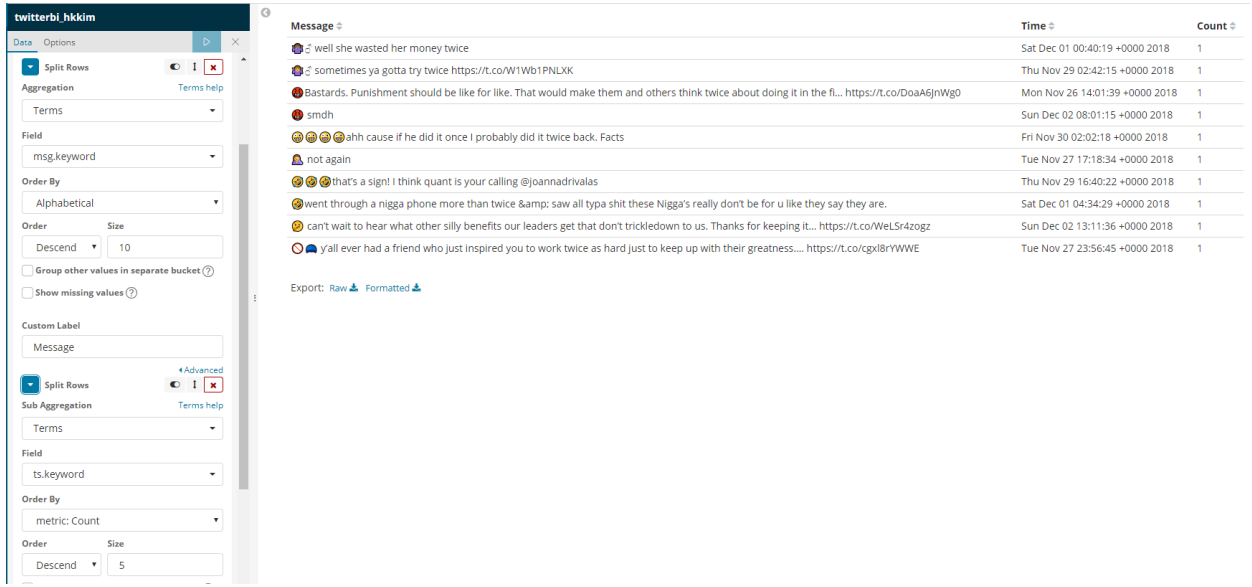
5. Hive Table을 Kibana에서 분석

(3) Dashboard에 들어갈 Visualize(그래프) 생성

hkkim_lang_visual : 언어별 사용자들에 대한 통계를 원그래프로 표현



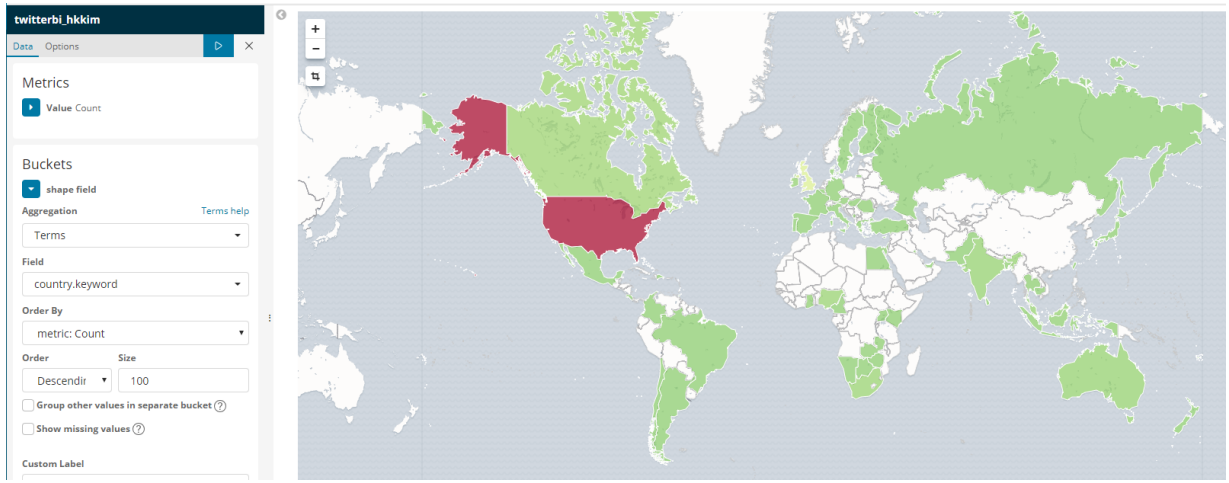
hkkim_msg_visual : 사용자들이 작성한 트위터 메시지들을 표시



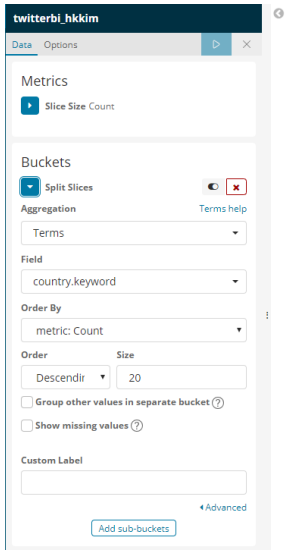
5. Hive Table을 Kibana에서 분석

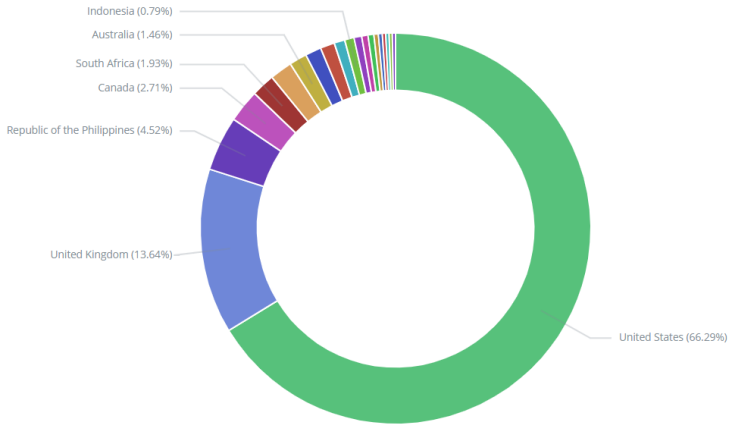
(3) Dashboard에 들어갈 Visualize(그래프) 생성

hkkim_world_visual : 나라별 사용자들에 대한 통계를 세계지도에서 표현



hkkim_world_round_visual : 나라별 사용자들에 대한 통계를 원그래프로 표현

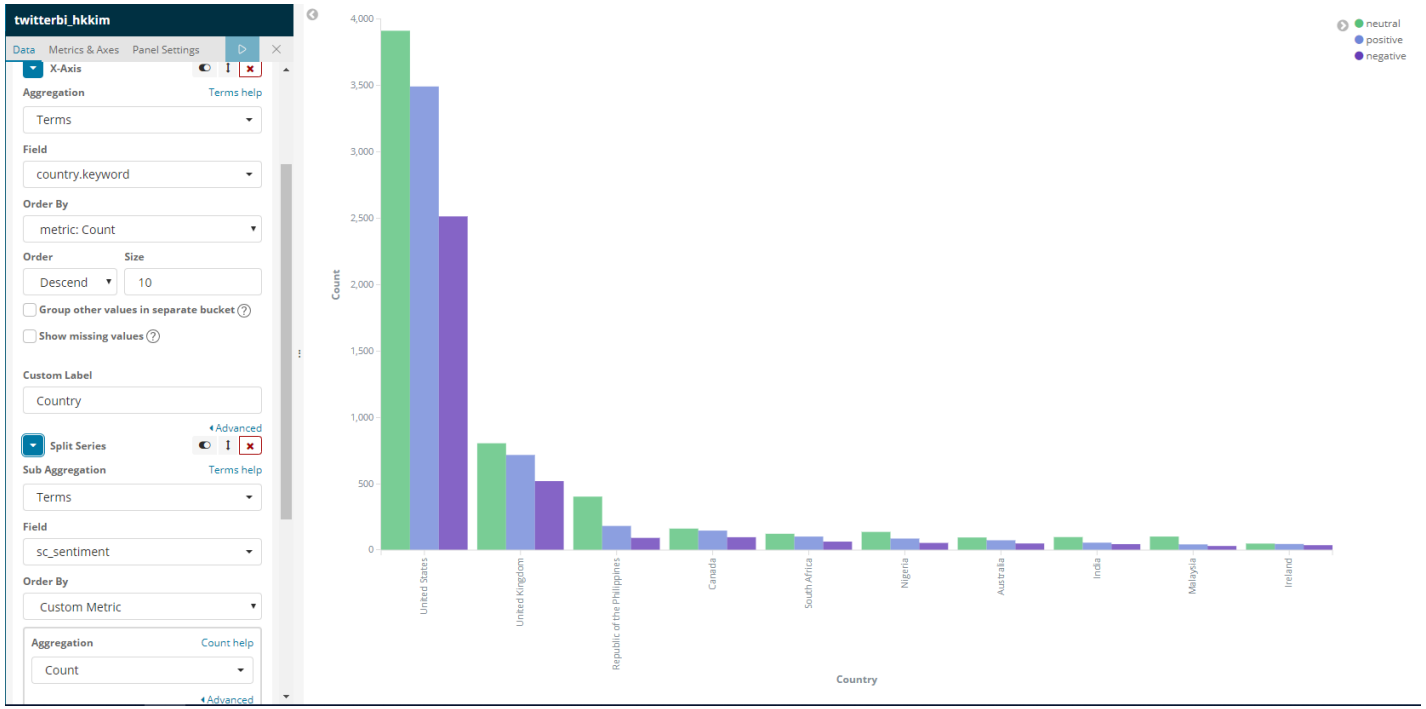




5. Hive Table을 Kibana에서 분석

(3) Dashboard에 들어갈 Visualize(그래프) 생성

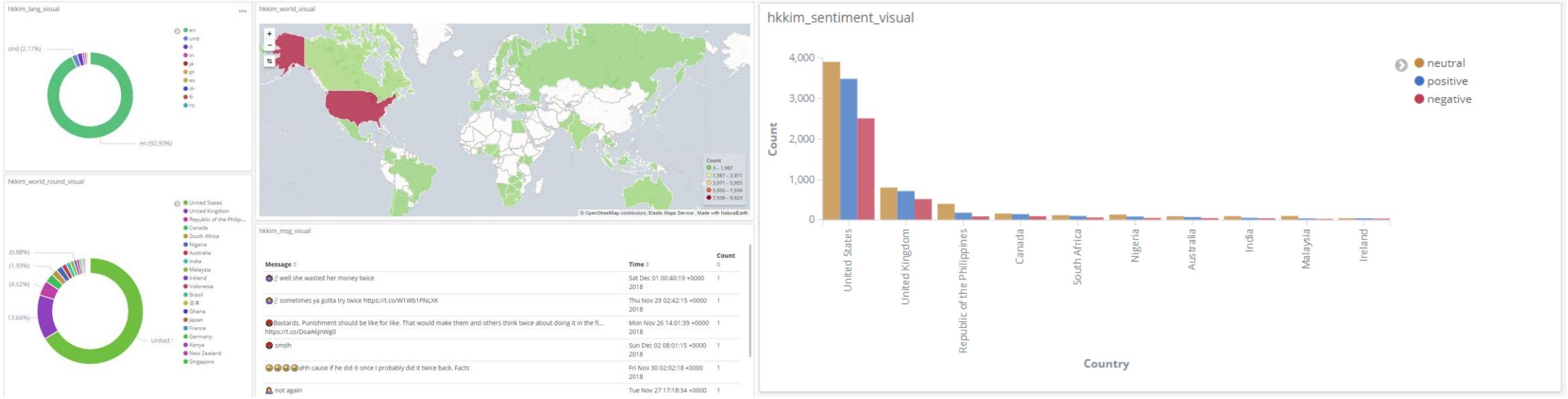
hkkim_sentiment_visual : 사용자들의 긍정, 중립, 부정의 정도를 표현



5. Hive Table을 Kibana에서 분석

(3) Dashboard에 들어갈 Visualize(그래프) 생성

(Dashboard에 적용된 모든 그래프들)

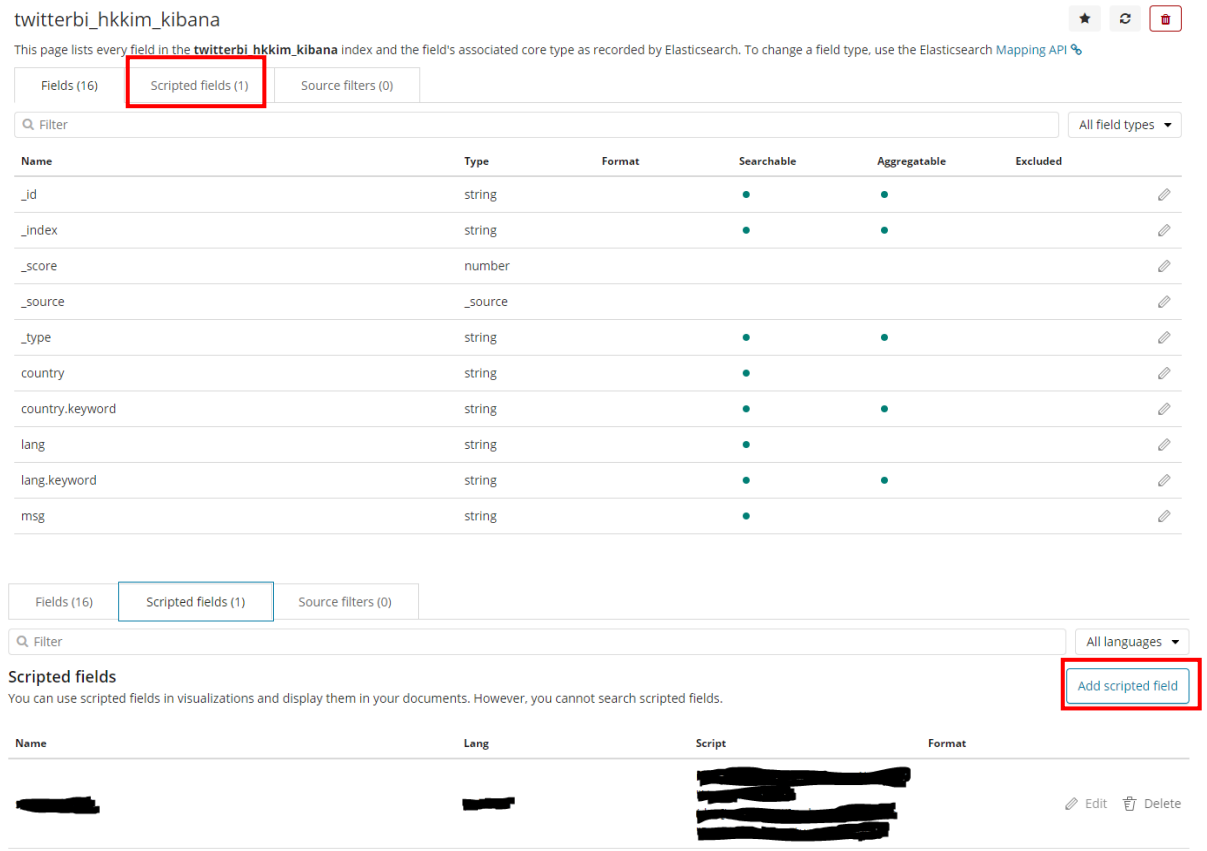


5. Hive Table을 Kibana에서 분석

(3) Dashboard에 들어갈 Visualize(그래프) 생성

(+ value명을 변경하는 법. 즉, 컬럼명을 변경한다)

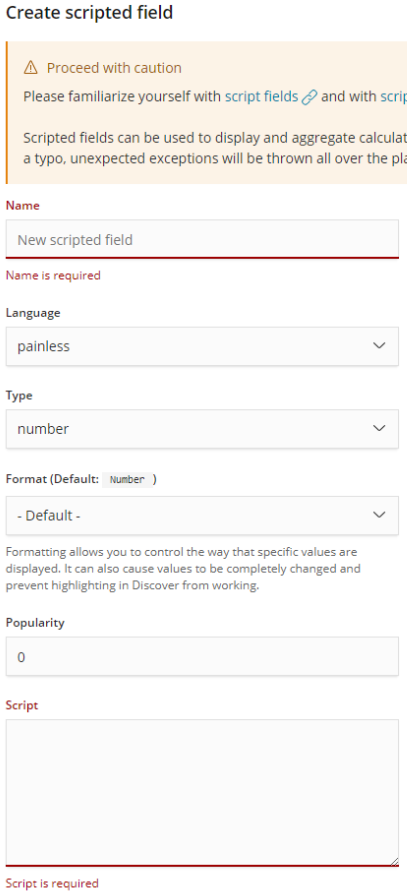
(1) Management -> twitterbi_hkkim_kibana -> Scripted fields -> Add scripted field



5. Hive Table을 Kibana에서 분석

(3) Dashboard에 들어갈 Visualize(그래프) 생성

(2) Script 작성



if (doc['sentiment'].value == 1) return "Neutral";

else if (doc['sentiment'].value == 2) return"Positive";

else return "Negative";

Name : scripted field 이름 짓기 *ex) str_sentiment*

Language : Painless, Expression 2 개 가 있 는 데 , default 는 Painless.
필자는 Painless로 설정되어 있는 것으로 바로 진행.

Type : 잘 모르겠음.

Format : 잘 모르겠음. (Type이 바뀌면 format의 default값 또한 변경된다.)

Popularity : 잘 모르겠음.

Script : Script내용 작성

(ex) sentiment field는 중립, 긍정, 부정을 '1', '2', '0'으로 표현.

그러나 그 표현방식을 "Neutral", "Positive", "Negative"로 바꾸어보는 script.)

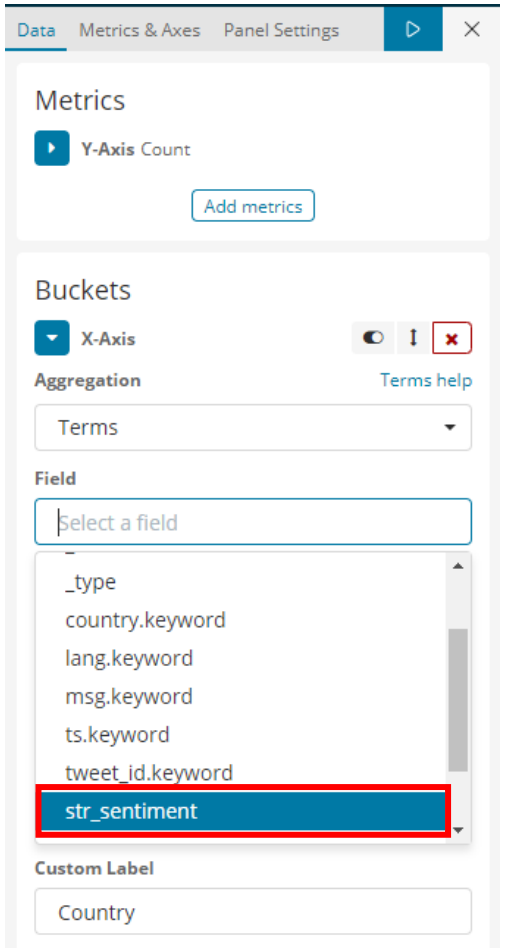
꼭 알아야할 점 :

*doc['field_name'].value*을 통해 해당 field에 있는 value들에게 access할 수 있다.

5. Hive Table을 Kibana에서 분석

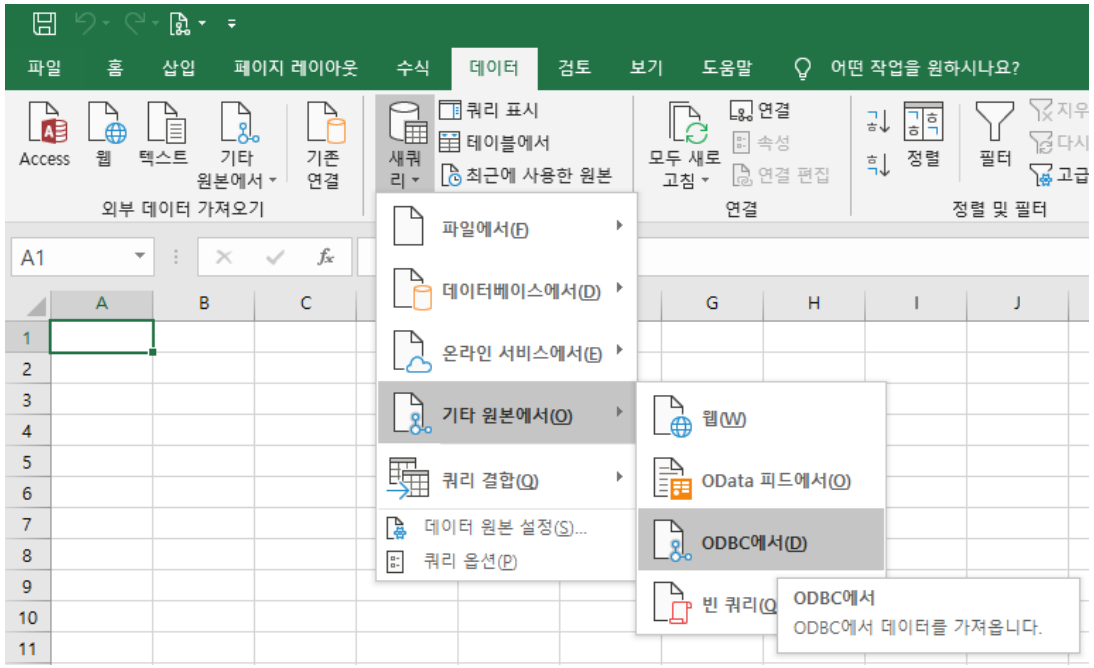
(3) Dashboard에 들어갈 Visualize(그래프) 생성

3) 위 과정을 거쳐서 "str_sentiment"란 이름을 가진 scripted field를 새로 생성하였다.



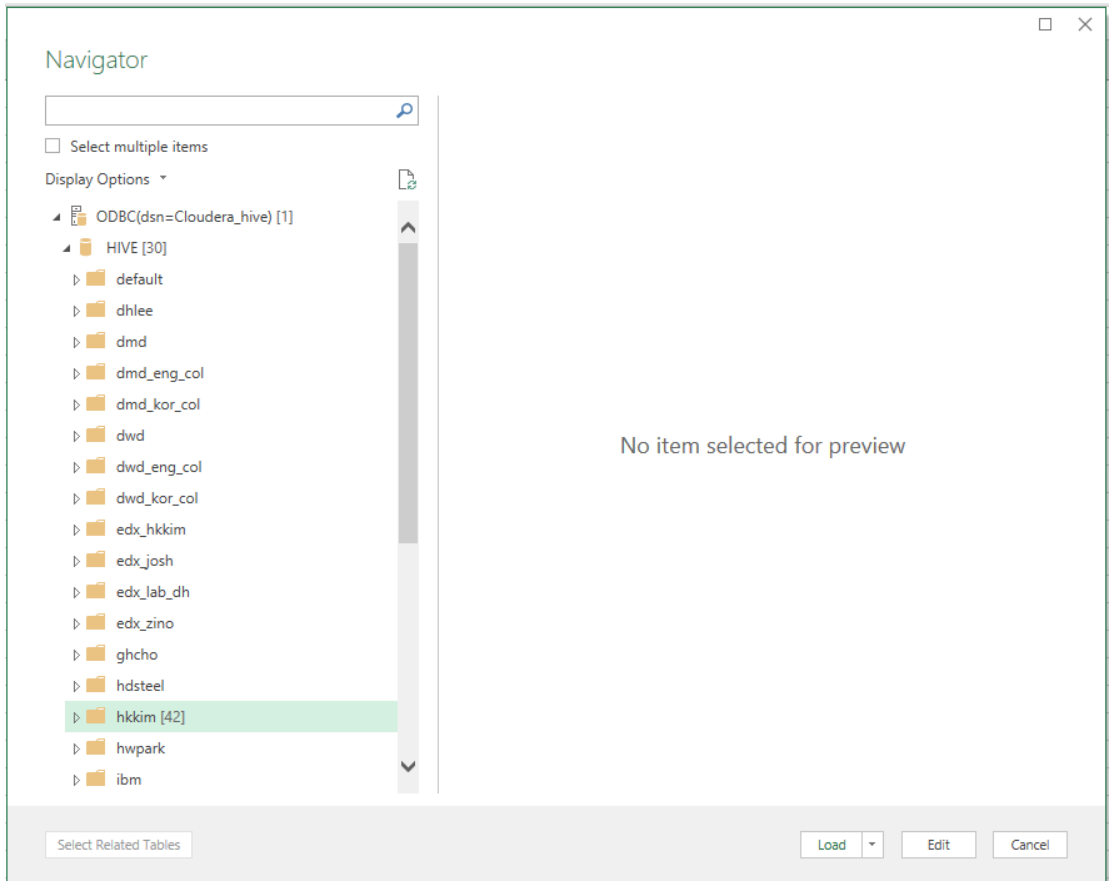
6. Excel 3D Map에서 분석 (부록 먼저 진행)

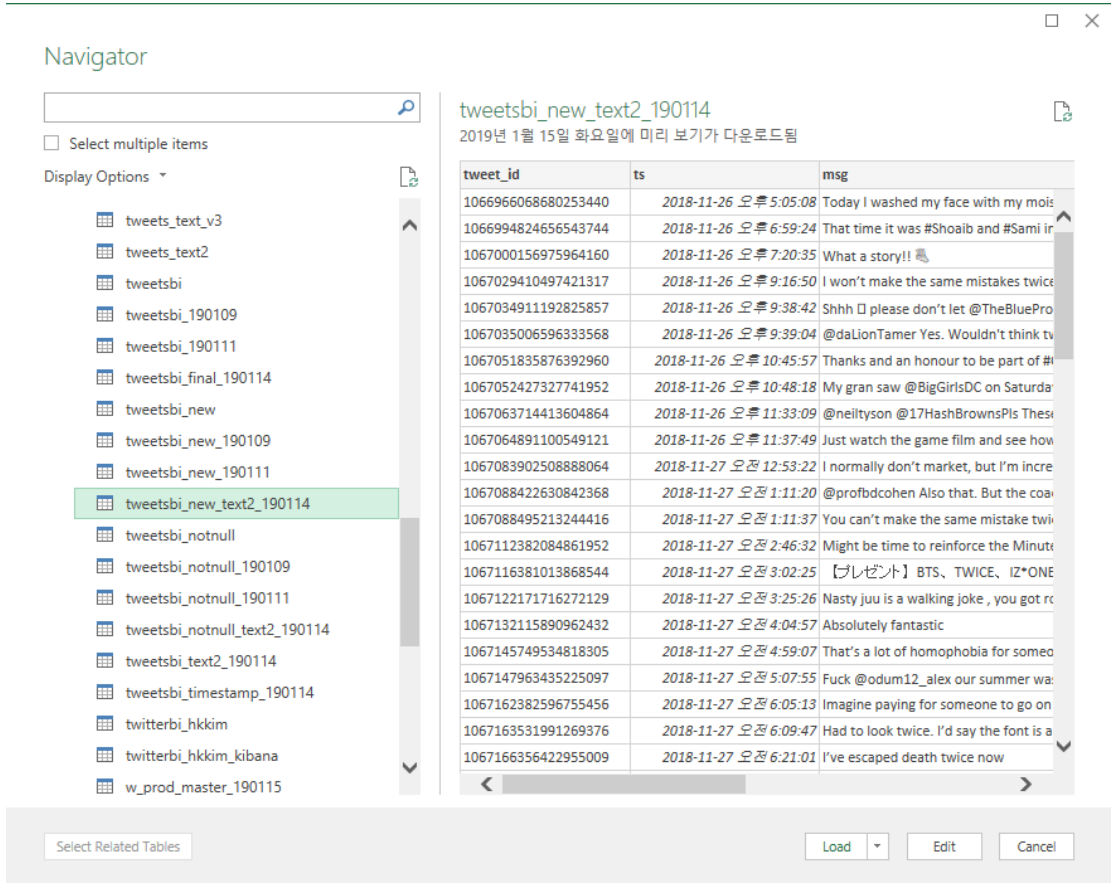
(1) Hive Table 가져오기 : "데이터 탭" -> "새 쿼리" -> "기타 원본에서" - > "ODBC에서"



6. Excel 3D Map에서 분석 (부록 먼저 진행)

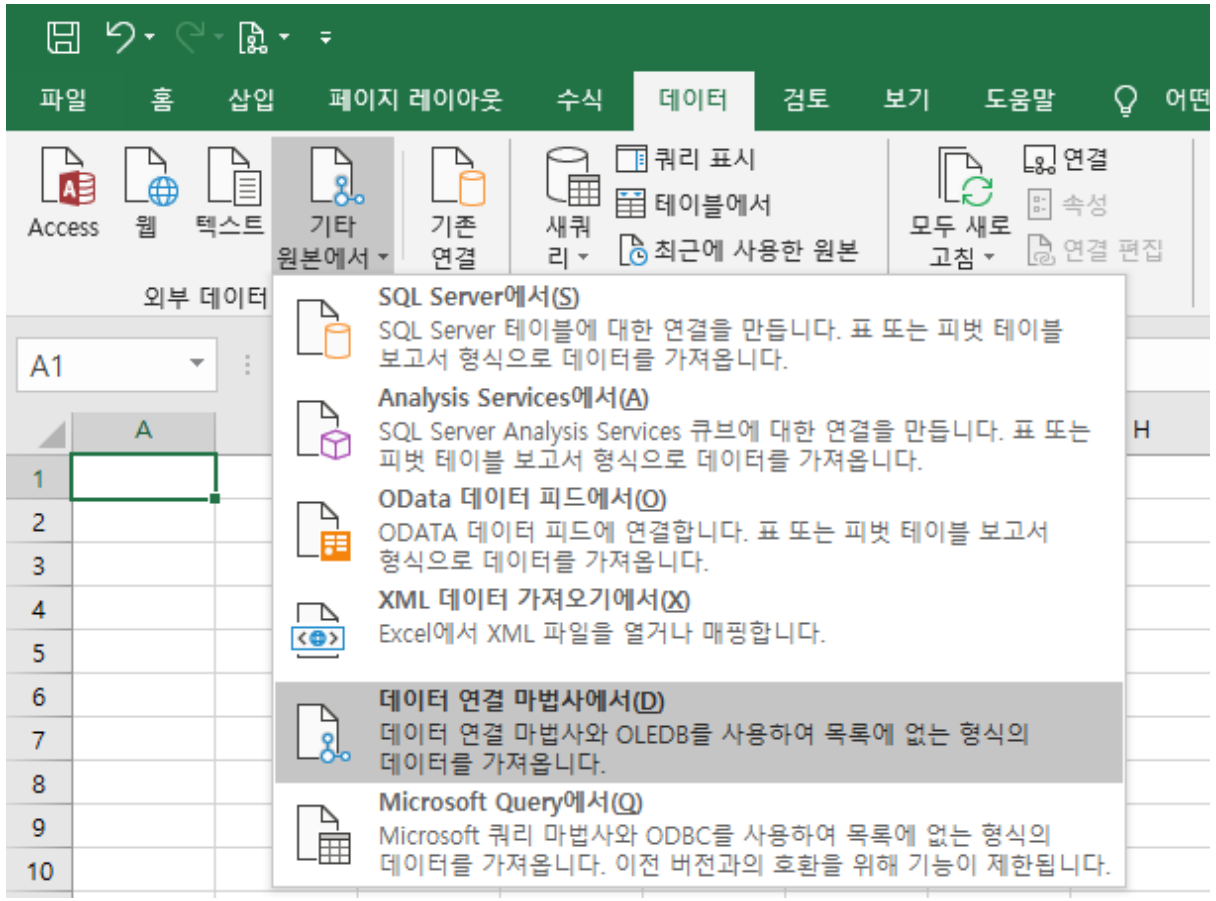
(1) Hive Table 가져오기 : "데이터 탭" -> "새 쿼리" -> "기타 원본에서" - > "ODBC에서"

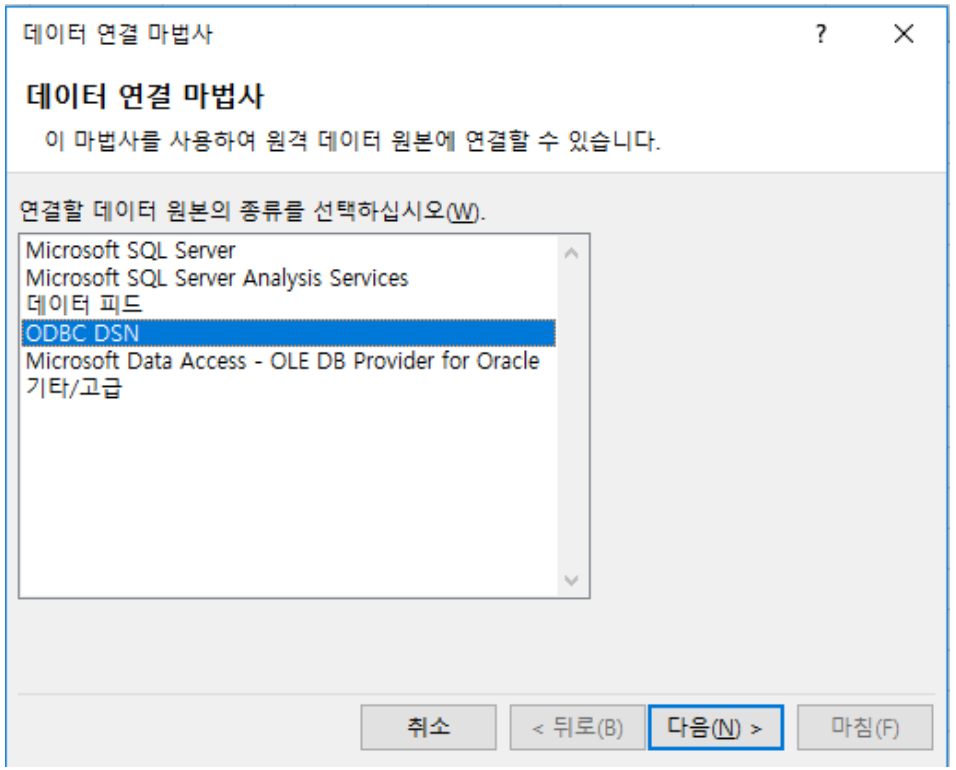




6. Excel 3D Map에서 분석 (부록 먼저 진행)

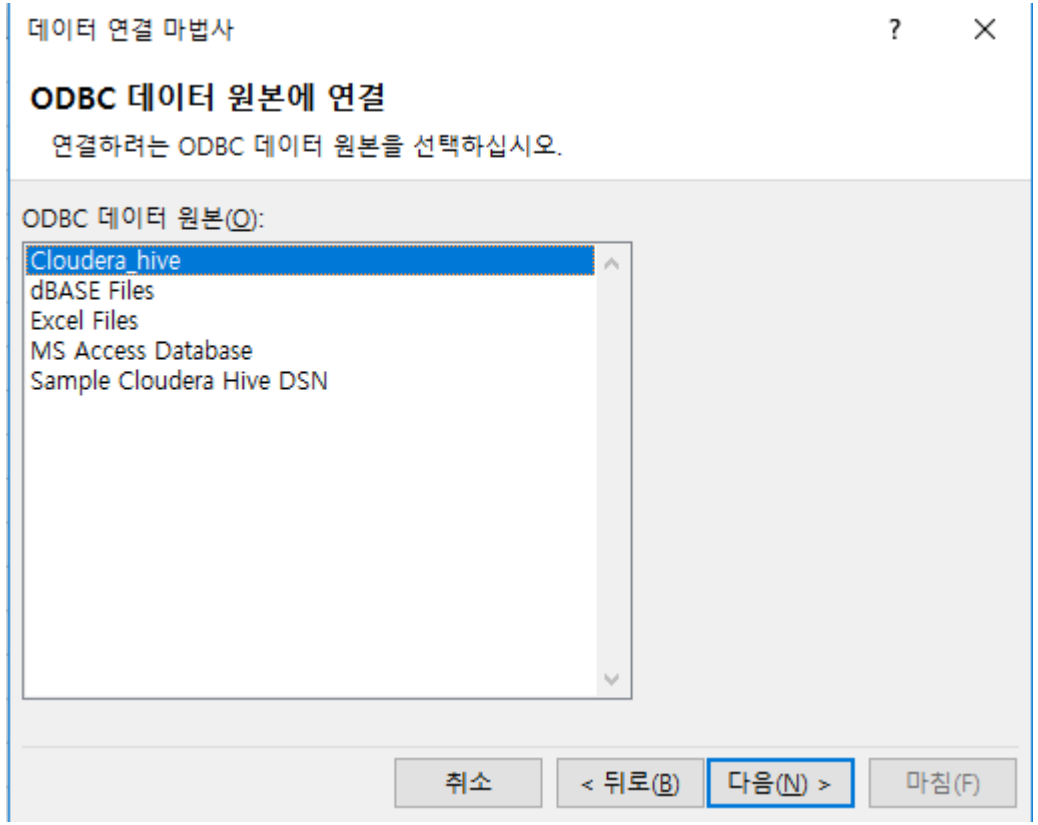
(2) Hive Table 가져오기 : "데이터 탭" -> "기타 원본에서" - > "데이터 연결 마법사에서"

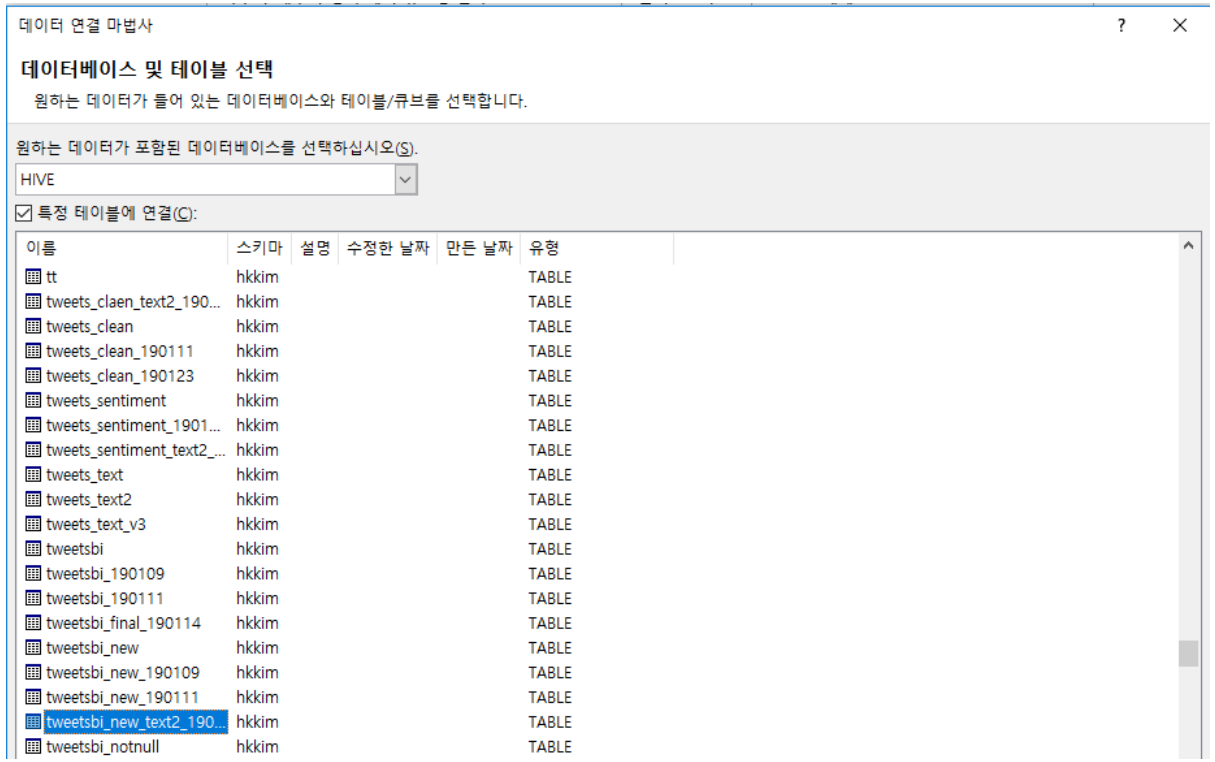




6. Excel 3D Map에서 분석 (부록 먼저 진행)

(2) Hive Table 가져오기 : "데이터 탭" -> "기타 원본에서" - > "데이터 연결 마법사에서"





6. Excel 3D Map에서 분석 (부록 먼저 진행)

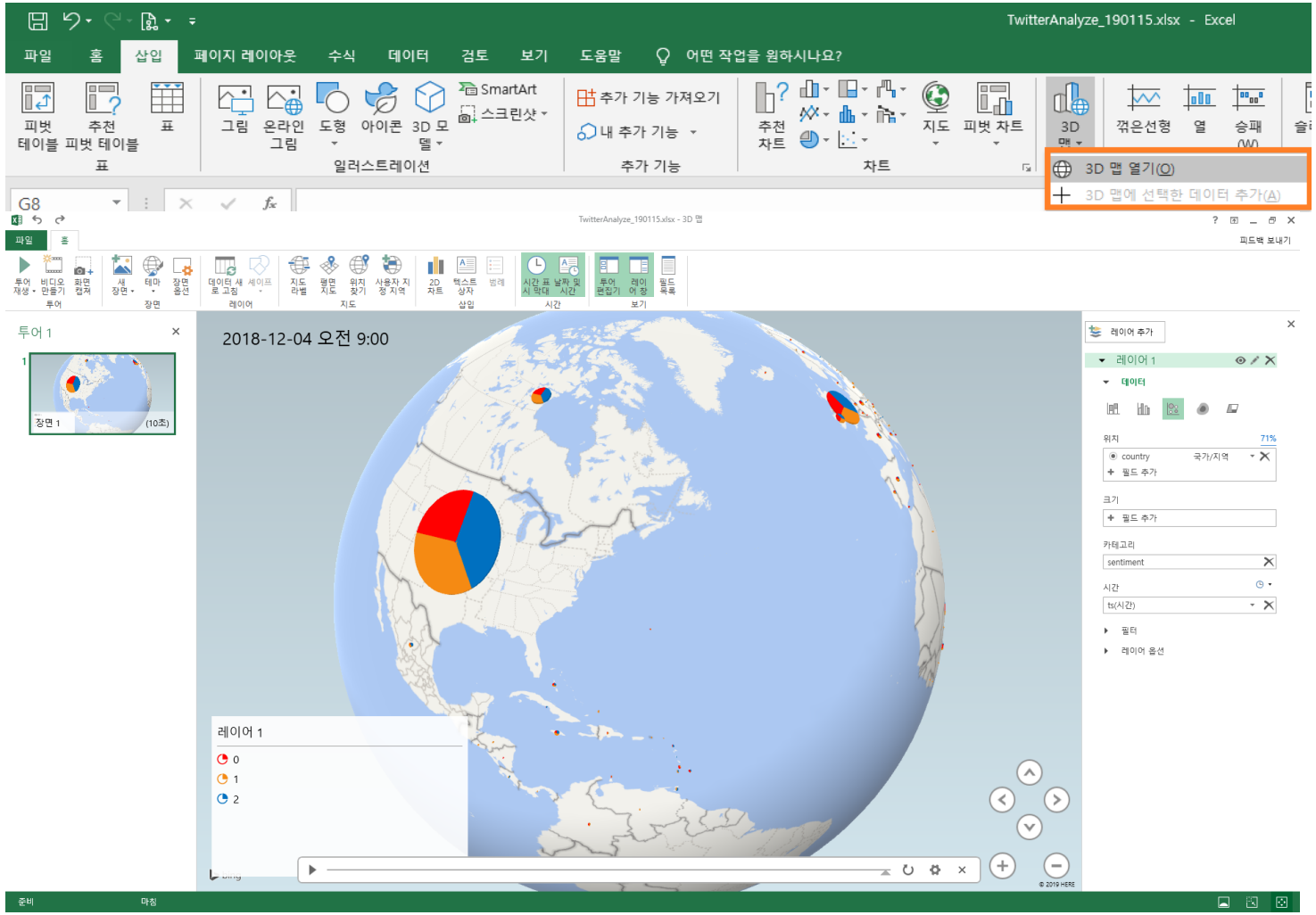
(2) Hive Table 가져오기 : "데이터 탭" -> "기타 원본에서" - > "데이터 연결 마법사에서"

결과 : 데이터를 Excel에 Import 했을 시의 모습

tweet_id	ts	msg	country	sentiment	lang
1066966068680253440	2018-11-26 17:05	Today I washed my face with my moisturizer. Twice.	United States	1	en
1066994824656543744	2018-11-26 18:59	That time it was #Shoaib and #Sami in this time it is Yasir alone.		1	en
1067000156975964160	2018-11-26 19:20	What a story!! 🔥	United Kingdom	1	en
1067029410497421317	2018-11-26 21:16	I won't make the same mistakes twice anymore	United States	0	en
1067034911192825857	2018-11-26 21:38	Shhh 😘 please don't let @TheBlueProphet see this ... lol	Ghana	2	en
1067035006596333568	2018-11-26 21:39	@daLionTamer Yes. Wouldn't think twice.	Hong Kong	2	en
1067051835876392960	2018-11-26 22:45	Thanks and an honour to be part of #C4PV making the Conservative case for #PeoplesVote	United Kingdom	1	en
1067052427327741952	2018-11-26 22:48	My gran saw @BigGirlsDC on Saturday night and is still raving about it. Proudly shown me he	United Kingdom	0	en
1067063714413604864	2018-11-26 23:33	@neiltyson @17HashBrownsPls These days he might have thought twice. Can you say The Tor	United States	2	en
1067064891100549121	2018-11-26 23:37	Just watch the game film and see how bad FSU's D played. Peep the busted coverages, how th	United States	2	en
1067083902508888064	2018-11-27 0:53	I normally don't market, but I'm incredibly happy with the quality f stickers made by @StickerY	United States	2	en
1067088422630842368	2018-11-27 1:11	@profbdcohen Also that. But the coaching situation had to be changed. Embarrassing losses t	United States	0	en
1067088495213244416	2018-11-27 1:11	You can't make the same mistake twice. The second time you make it, it's no longer a mistake	Germany	0	en
1067112382084861952	2018-11-27 2:46	Might be time to reinforce the Minuteman movement! Organized, United, Strong, Ethical Volu	United States	2	en
1067116381013868544	2018-11-27 3:02	【プレゼント】BTS、TWICE、IZ*ONE、Wanna One出演！12/12開催「2018 MAMA」日本公演に5組	日本	1	ja
1067122171716272129	2018-11-27 3:25	Nasty juu is a walking joke , you got rocked on ur own side lines N started crying in front of e	United States	2	en
1067132115890962432	2018-11-27 4:04	Absolutely fantastic	United Kingdom	2	en
1067145749534818305	2018-11-27 4:59	That's a lot of homophobia for someone who would get their shit ran by dwight howard. Thin	United States	1	en
1067147963435225097	2018-11-27 5:07	Fuck @odum12_alex our summer wasn't good enough. 😂🤣	United States	2	en
1067162382596755456	2018-11-27 6:05	Imagine paying for someone to go on holiday but they can't even hold off a gig :/ of someon	United Kingdom	2	en
1067163531991269376	2018-11-27 6:09	Had to look twice. I'd say the font is a little too close together. Would make an excellent stock	United States	0	en
1067166356422955009	2018-11-27 6:21	I've escaped death twice now	The Netherlands	0	en
1067171716894474240	2018-11-27 6:42	and the eagles won both games. isn't that the part that matters?	United States	1	en
1067177692716589058	2018-11-27 7:06	Genuinely thought this was just me for ages. @TonyMartiallah	United Kingdom	2	en
1067202785249296384	2018-11-27 8:45	@Rogg_2016 I told kole I just wanted my hair done BUT I also want this really pretty ring that	United States	2	en
1067204108057305088	2018-11-27 8:51	i drop my phone and got it replaced twice (2x!!!!!) this weekend should i just get a pink motor	United States	2	en
1067209020866363393	2018-11-27 9:10	@elonmusk @margrethmpossi Sounds like a permanent, manic state. The light that burns twic	United States	2	en
1067209198419800064	2018-11-27 9:11	@Bungie fix this please I got error code guitar twice tonight 😢 https://t.co/DbSAhTwqwG	Ireland	0	en
1067210619475148800	2018-11-27 9:16	don't gotta tell me twice 😏	United States	1	en
1067221204182937600	2018-11-27 9:58	Eh blue pound	United Kingdom	1	en
1067222671451201536	2018-11-27 10:04	@Blondiegrrl @nytimes Hard to imagine this to happen. Once a candidate lost he or she has	United States	0	en
1067230750255267841	2018-11-27 10:36	@AbbeyRoseJ I've been at myrtle where the j/m intersect for 6 months now but I've only taker	United States	1	en
1067237365985488896	2018-11-27 11:03	@BravenakBlog oh dear God. I mean, I love me some beto. Think he's super. But now I have	United States	2	en

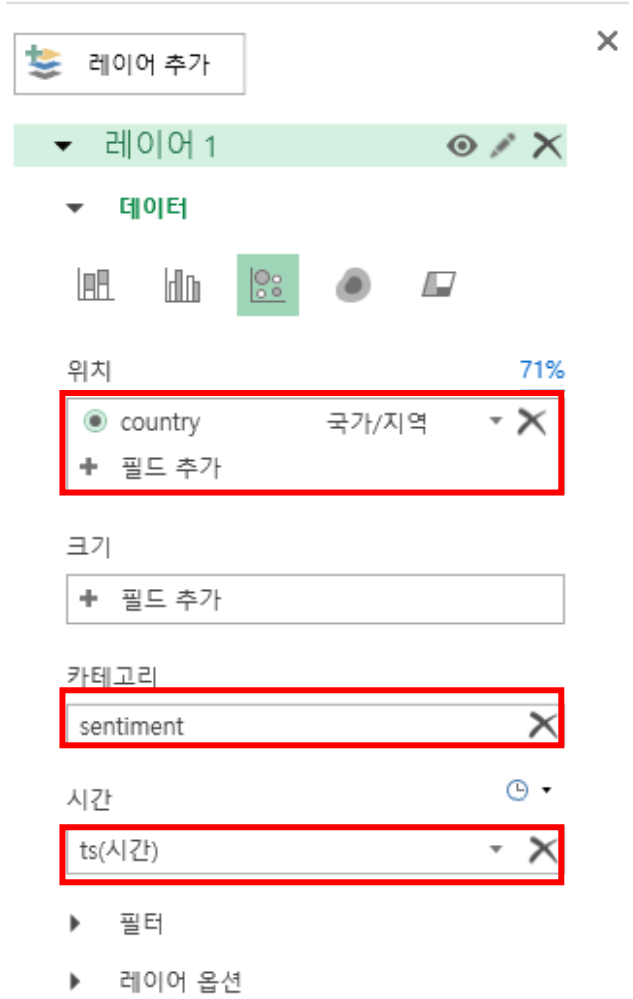
6. Excel 3D Map에서 분석 (부록 먼저 진행)

(3) Excel 3D Map 사용



6. Excel 3D Map에서 분석 (부록 먼저 진행)

(3) Excel 3D Map 사용

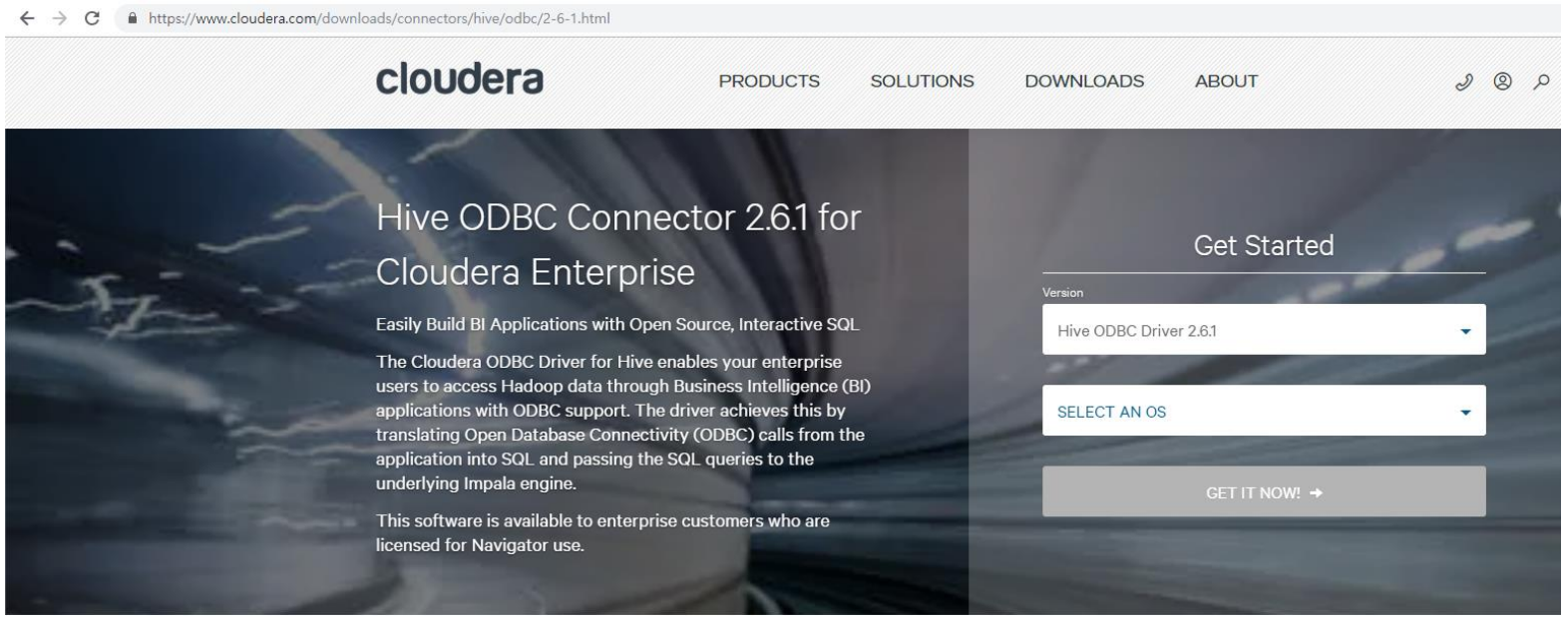


*** 위치 : 각 나라별 위치를 뜻한다**

*** sentiment : 각 나라별로 응답의 긍정,중립,부정의 정도를 나타낸다.**

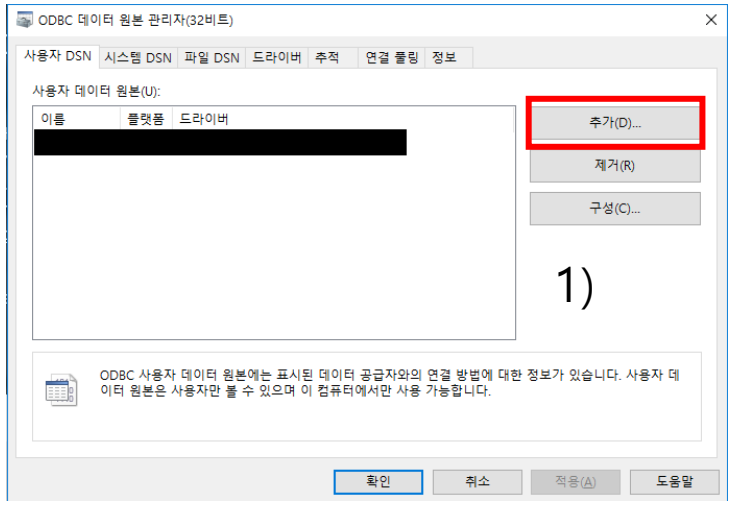
*** ts : 시간의 흐름에 따라 각 나라별 sentiment의 정도의 변화를 볼 수 있다.**

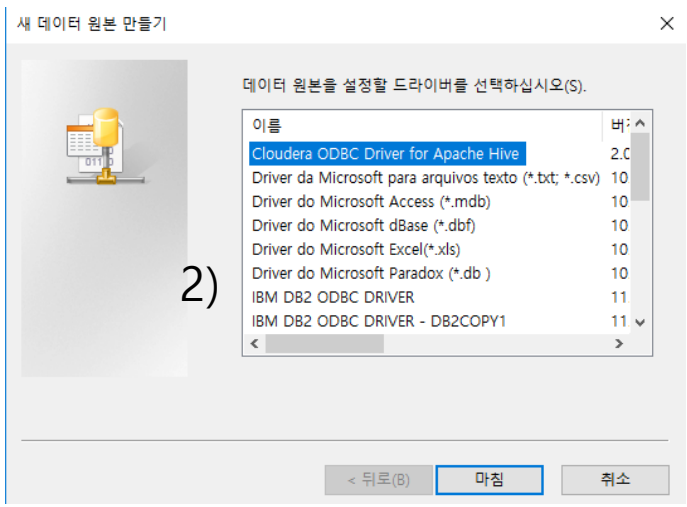
부록 1 : Cloudera ODBC Connector 설치하는 방법

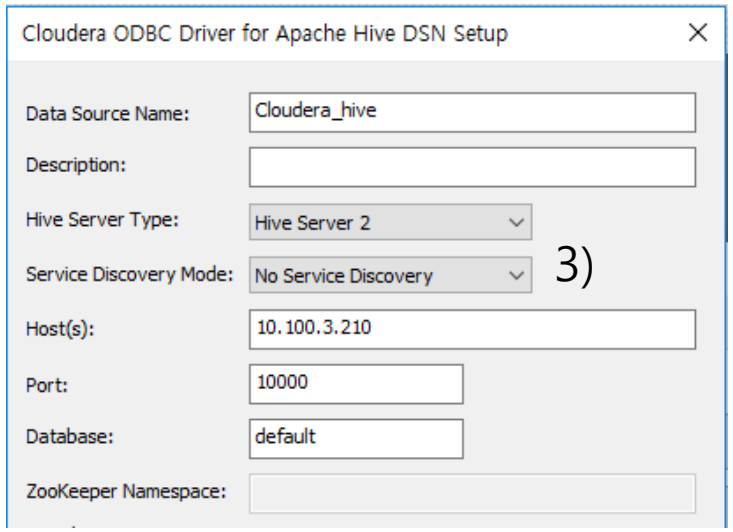


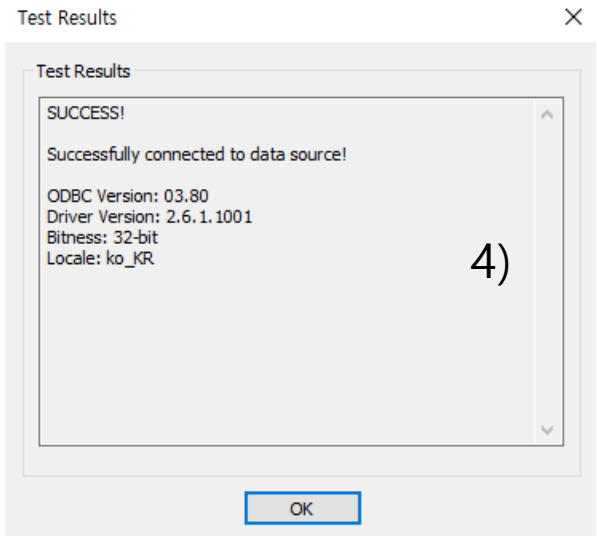
다운로드 경로 : https://www.cloudera.com/downloads/connectors/hive/odbc/2-6-1.html
(32bit, 64bit 2가지가 있는데 본인에게 알맞은 bit로 선택한다.
작성자의 경우에는 32bit 버전을 다운받았다.)

부록 1 : Cloudera ODBC Connector 설치하는 방법







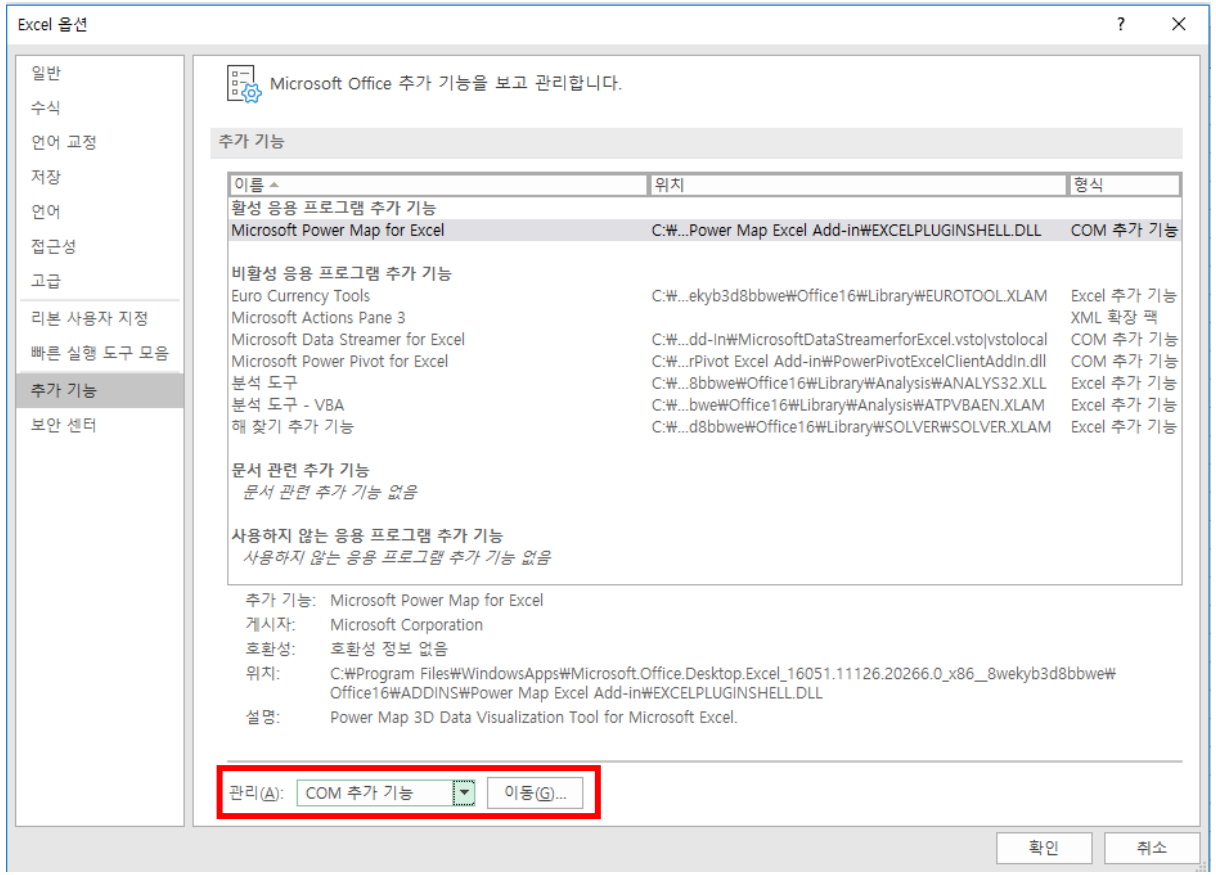


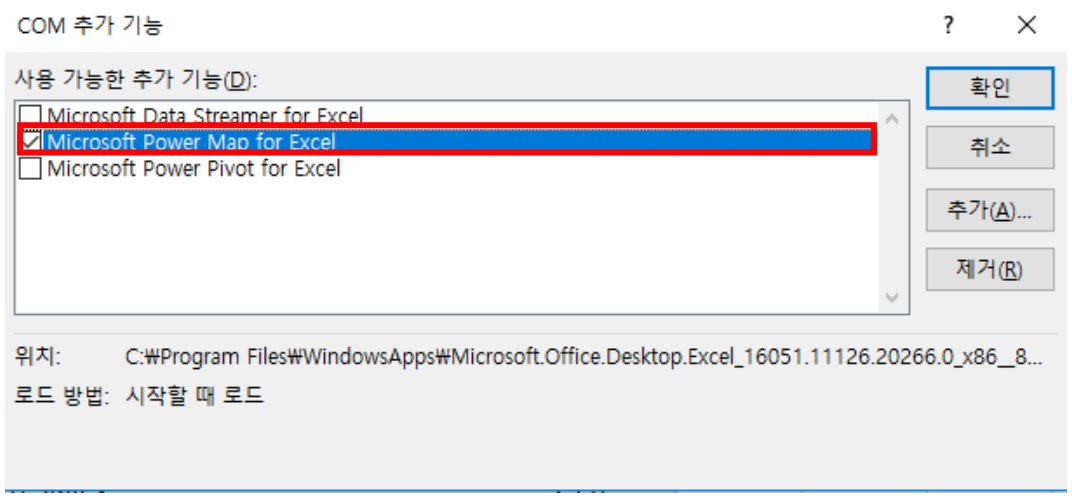
(1) 다운을 받아 ClouderaHiveODBC32(2.6).msi 파일을 열어 설치 진행

(2) "시작" -> "ODBC 데이터 원본 (32비트)"

(3) 아래와 같이 진행

부록 2 : Excel 3D Map 활성화





참고자료

Reference ：

https://github.com/nberah/tutorials/blob/hdp/tutorials/hortonworks/analyzing-social-media-sentiment-data/tutorial.md