

Prácticas BigData

HIVE

1. Prácticas adicional. Deslizamientos de tierra

- Vamos a realizar unas cuantas SELECT contra una DataSet de la NASA, que contiene información sobre deslizamientos de tierra ocurridos alrededor del mundo
- El fichero deslizamientos.csv tiene los datos
- Primero creamos la siguiente tabla en HIVE

```
create table deslizamientos
id
                    bigint,
fecha
                          string,
hora
                       string,
country
                        string
nearest_places
                        string
hazard_type
                        string
landslide_type
                        string
motivo
                       string
storm_name
                        string
  fatalities
                          bigint
  injuries
                          string
  source_name
                          string
  source_link
                          string
  location_description
                          string
  location_accuracy
                          string
  landslide_size
                          string
  photos_link
                           string
  cat_src
                          string
  cat_id
                          bigint
  countryname
                           string
  near
                          string
  distance
                          double
  adminname1
                           string
  adminname2
                           string
  population
                          bigint
  countrycode
                           string
```



```
continentcode
                          string
  key
                          string
  version
                          string
  tstamp
                          string
  changeset_id
                          string
  latitude
                          double
  longitude
                          double
                          string
  geolocation
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ';'
```

Luego la cargamos con los datos del fichero de los deslizamientos.
 Previamente, hemos copiado el fichero a /tmp

load data local inpath '/tmp/deslizamientos.csv' into table deslizamientos; No rows affected (7,21 seconds)

- Hagamos ahora algunos ejemplos de consultas
- Ver el nombre y fecha de las cinco primeras filas

```
select country, fecha from deslizamientos limit 5;
+-----+
| country | fecha |
+-----+
| United Kingdom | 01/02/2007 |
| Peru | 01/03/2007 |
| Brazil | 01/05/2007 |
| Brazil | 01/05/2007 |
```

 Averiguar el país, el tipo de deslizamiento y el motivo de aquellos sitios donde haya habido más de 100 víctimas.



	China	09/08/2008 Complex	Dam_Embankment_Collapse	277
	Brazil	11/24/2008 Landslide	Continuous_rain	109
	Taiwan	08/10/2009 Complex	Tropical_Cyclone	491
	Philippines	10/09/2009 Landslide	Tropical_Cyclone	104
	Uganda	03/01/2010 Complex	Downpour	388
	Brazil	04/07/2010 Mudslide	Downpour	196
	India	08/06/2010 Landslide	Downpour	234
	India	08/06/2010 Landslide	Downpour	182
	China	08/07/2010 Landslide	Downpour	1765
	Indonesia	10/04/2010 Landslide	Downpour	145
	Brazil	01/12/2011 Mudslide	Downpour	424
	Brazil	01/12/2011 Mudslide	Downpour	378
	Philippines	12/04/2012 Mudslide	Tropical_Cyclone	430
	India	06/16/2013 Debris_Flow	Downpour	5000
	Afghanistan	05/02/2014 Landslide	Continuous_rain	2100
	India	07/30/2014 Mudslide	Continuous_rain	151
	Nepal	08/02/2014 Landslide	Continuous_rain	174
		01/04/2006 Mudslide	Downpour	240
		12/12/2014 Landslide	Monsoon	108
		04/28/2015 Mudslide	Snowfall_snowmelt	250
		10/01/2015 Mudslide	Rain	280
		08/02/2015 Landslide	Downpour	253
		05/18/2016 Mudslide	Monsoon	101
		04/02/2016 Landslide	Unknown	104
+		++		

 Averiguar los deslizamientos ocurridos por tipos de deslizamiento (landslide_type)

select landslide_type, count(*) from deslizamientos group by landslide_type;



WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.

+	-+		-+
landslide_type		_c1	
+	-+		-+
1	1	18	
Complex		232	
Creep		5	
Debris_Flow		173	
Earthflow		3	
Lahar		7	
Landslide		6637	
Mudslide		1826	
Other		66	
Riverbank_Collapse		28	
Rockfall		484	
Rockslide		1	
Snow_Avalanche		7	
Translational_Slide		6	
Unknown		18	
landslide		4	
mudslide	1	7	
+	-+		-+;

Averiguar los que han ocurrido agrupados por motivo

select motivo, count(*) fr	om desli	zamientos group by motivo;						
++								
motivo	_c1							
+	-+	-+						
	756							
Construction	52							
Continuous_Rain	36							
Continuous_rain	514							
Dam_Embankment_Collapse	9							
Downpour	4437							
Earthquake	76							
Flooding	49							
Freeze_thaw	26							
Mining_digging	74							



```
Monsoon
                         122
| No_Apparent_Trigger
                         | 2
| No Apparent trigger
                         | 18
| Other
                         | 15
Rain
                         1912
| Snowfall snowmelt
                        | 74
| Tropical_Cyclone
                        | 538
Unknown
                         748
| Volcano
                         | 1
monsoon
                         | 2
unknown
                         | 61
```

Indicar los 10 paises con más deslizamientos registrados

```
select country,count(*) as total from deslizamientos group by country
order by total desc limit 10;
             | total |
    country
+----+
               3387
| United States | 1439
| India
             884
| Philippines
             546
| China
             347
              324
Nepal
| Indonesia | 282
Brazil
               205
| United Kingdom | 147
Malaysia
               | 110
```

Crear la siguiente table de países.

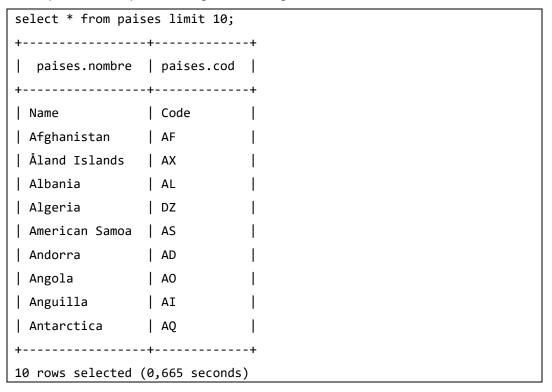
```
create table paises
  (
nombre string,
cod string)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
```

 Cargamos la tabla countries.csv que tenemos en los recursos de la práctica

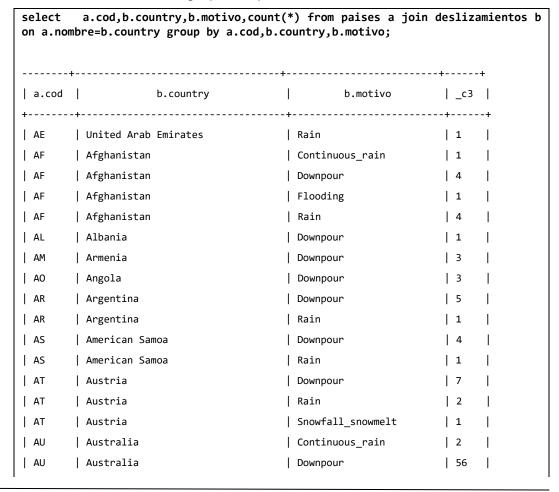
```
load data local inpath '/tmp/countries.csv' into table paises;
```



Comprobamos que ha cargado los registros



 Vamos ahora a visualizar el código del país, el nombre y el número de corrimientos de tierra agrupados por motivo





AU	Australia	Mining_digging	1	1
AU	Australia	Rain	15	1
AU	Australia	Tropical_Cyclone	1	1
AU	Australia	Unknown	4	1
AZ	Azerbaijan	Downpour	16	1
AZ	Azerbaijan	Snowfall_snowmelt	1	1
AZ	Azerbaijan	Unknown	2	1
BA	Bosnia and Herzegovina	Downpour	1	1
BA	Bosnia and Herzegovina	Rain	4	1
BB	Barbados	Downpour	1	1
BD	Bangladesh	Downpour	20	1
BD	Bangladesh	Monsoon	3	1
BD	Bangladesh	Rain	8	1
BD	Bangladesh	Unknown	2	

- Ahora, como práctica final, vamos a exportarlo a un fichero.
- Podemos hacerlo con este comando. Lo dejamos en un directorio denominado datos

insert overwrite local directory '/tmp/datos' row format delimited fields
terminated by ',' select a.cod,b.country,b.motivo,count(*) from paises a
join deslizamientos b on a.nombre=b.country group by a.cod,b.country,b.motivo;

- Dentro va a generar un fichero denominado 00000_0
- Ahora, por último vamos a importarlo en un Excel para ver el resultado y hacer un gráfico. Sería el punto y final de un trabajo con Big Data

