# Proyecto Big Data

# Mundiales de Fútbol FIFA


**Elaborado por:** Julia Dávila Salcedo

**Objetivos:**

1. Desarrollar un proyecto de big data aplicando diversas tecnologías como Hive, Spark y Hadoop.

2. Analizar los datos de los mundiales de fútbol de la FIFA, aplicar formatos y agrupaciones para obtener reportes.

**Tecnologías a usar:**

-pySpark: Lenguaje para manipular los dataframe de Spark en memoria.

-Hadoop: Permite el almacenamiento distribuido en un clúster

-Hive: motor de base de datos distribuido.

Se usará una máquina virtual en virtualbox con estas tecnologías instaladas para la parte de HDFS HIVE . Se usará Databricks para la parte de pySpark.

**Archivos de Datos a usar:**

-WorldCups.csv: lista de los mundiales de futbol, año en que se realizaron y países ganadores.

-WorldCupPlayers.csv: lista de los jugadores en cada partido.

-WorldCupMatches: lista de partidos, cantidad de goles anotados y asistentes.

A continuación se detalla las tablas y campos de cada archivo.

**Tabla WorldCups**

| |
|---|
| Year Year of the worldcup |
| Country Country of the worldcup |
| WinnerTeam who won the worldcup |
| Runners-UpTeam who was the second place |
| ThirdTeam who was the third place |
| FourthTeam who was the fourth place |
| GoalsScored Total goals scored in the worldcup |
| QualifiedTeams Total participating teams |
| MatchesPlayed Total matches played in the cup |
| Attendance Total attendance of the worldcup |

**Tabla WorldCupPlayers**

| |
|---|
| RoundIDUnique ID of the round |
| MatchIDUnique ID of the match |
| Team InitialsPlayer's team initials |
| Coach NameName and country of the team coach |
| Line-up S=Line-up, N=Substitute |
| Shirt NumberShirt number if available |
| Player NameName of the player |

PositionC=Captain, GK=Goalkeeper

Event G=Goal, OG=Own Goal, Y=Yellow Card, R=Red Card, SY = Red Card by second yellow, P=Penalty, MP=Missed Penalty, I = Substitution In, O=Substitute Out

**Tabla WorldCupMatches**

Home Team NameHome team country name

Home Team GoalsTotal goals scored by the home team by the end of the match

Away Team GoalsTotal goals scored by the away team by the end of the match

Away Team NameAway team country name

Win conditionsSpecial win condition (if any)

AttendanceTotal crowd present at the stadium

Half-time Home GoalsGoals scored by the home team until half time

Half-time Away GoalsGoals scored by the away team until half time

RefereeName of the first refree

Assistant 1Name of the first assistant referee (linesman)

Assistant 2Name of the second assistant referee (linesman)

RoundIDUnique ID of the Round

MatchIDUnique ID of the match

Home Team InitialsHome team country's three letter initials

Away Team InitialsAway team country's three letter initials

Archivos de configuracion, creación y carga de datos:

- **Script HDFS:** Contiene los comandos de creación de carpetas y permisos en HDFS
- **Script Hive:** Contiene las sentencias para la creación de tablas en Hive.

- **etl_proyectobigdata.py** : Contiene comandos en python para la creación de los spark Dataframe y su manipulación para generar los reportes . Se incluye también las versiones en notebook y html de este archivo con las salidas de los comandos.

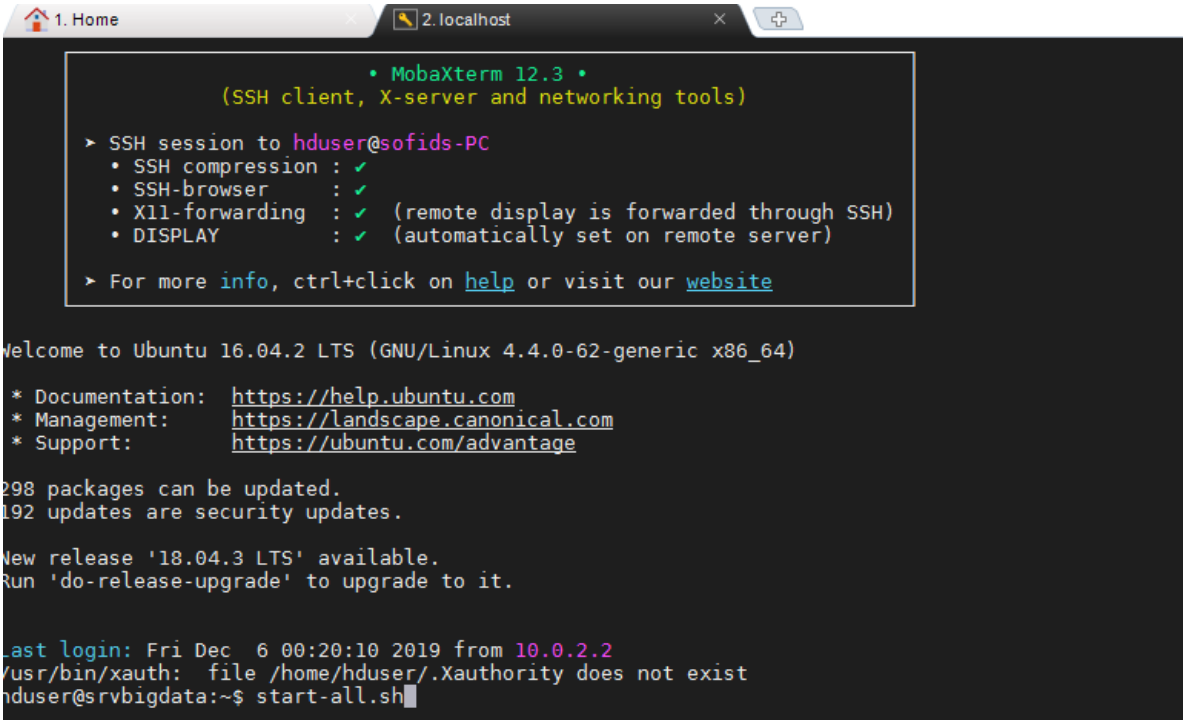
  **Pre requisitos**

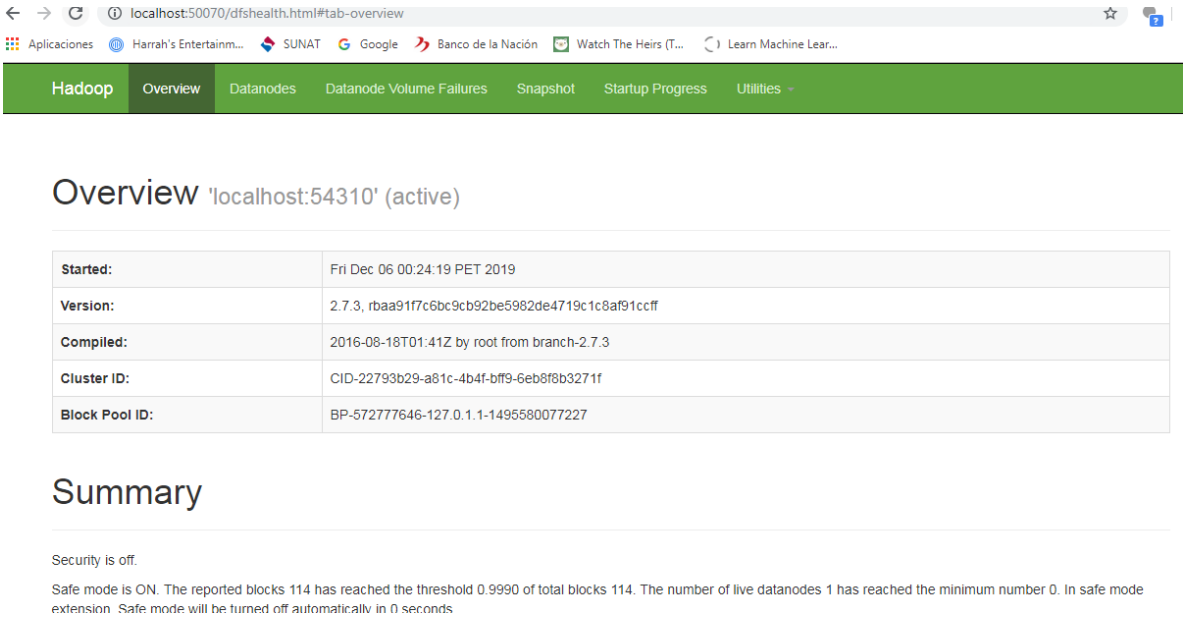
  1. Tener una cuenta en databricks

  2. Tener una maquina virtual con Hadoop y Hive instalado.

## Desarrollo del proyecto


1. Iniciar la máquina virtual y levantar el haddop






2. Crear la estructura de carpetas con el archivo **script_hdfs**

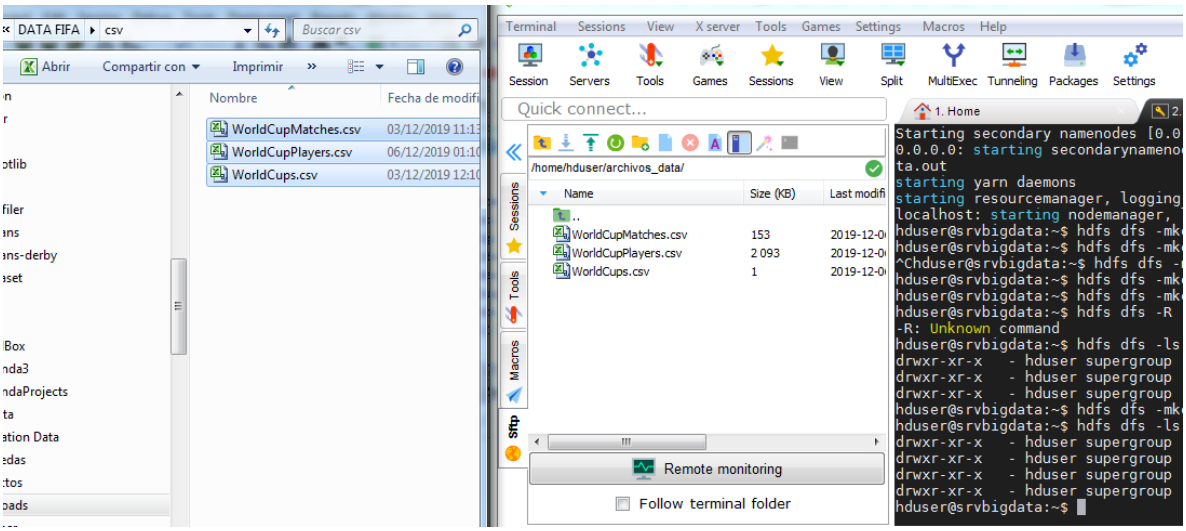
```
hduser@srvbigdata:~$ hdfs dfs -mkdir /ProyectoBigData
hduser@srvbigdata:~$ hdfs dfs -mkdir /ProyectoBigData/landing
^Chduser@srvbigdata:~$ hdfs dfs -mkdir /ProyectoBigData/smart
hduser@srvbigdata:~$ hdfs dfs -mkdir /ProyectoBigData/universo
hduser@srvbigdata:~$ hdfs dfs -mkdir /ProyectoBigData/reportes
hduser@srvbigdata:~$ hdfs dfs -R -ls /ProyectoBigData
```

```
hduser@srvbigdata:~$ hdfs dfs -ls -R /ProyectoBigData
drwxr-xr-x   - hduser supergroup          0 2019-12-06 01:04 /ProyectoBigData/landing
drwxr-xr-x   - hduser supergroup          0 2019-12-06 00:54 /ProyectoBigData/reportes
drwxr-xr-x   - hduser supergroup          0 2019-12-06 00:52 /ProyectoBigData/smart
drwxr-xr-x   - hduser supergroup          0 2019-12-06 00:53 /ProyectoBigData/universo
```

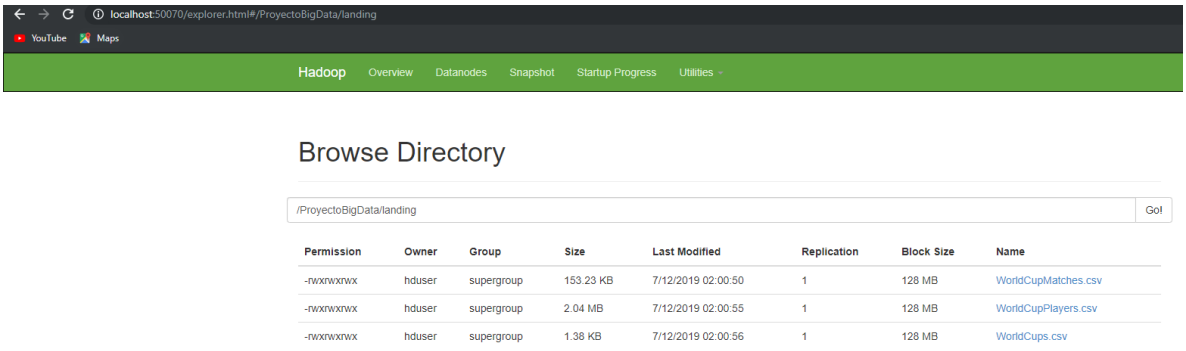
3. Modificando los permisos de la carpeta

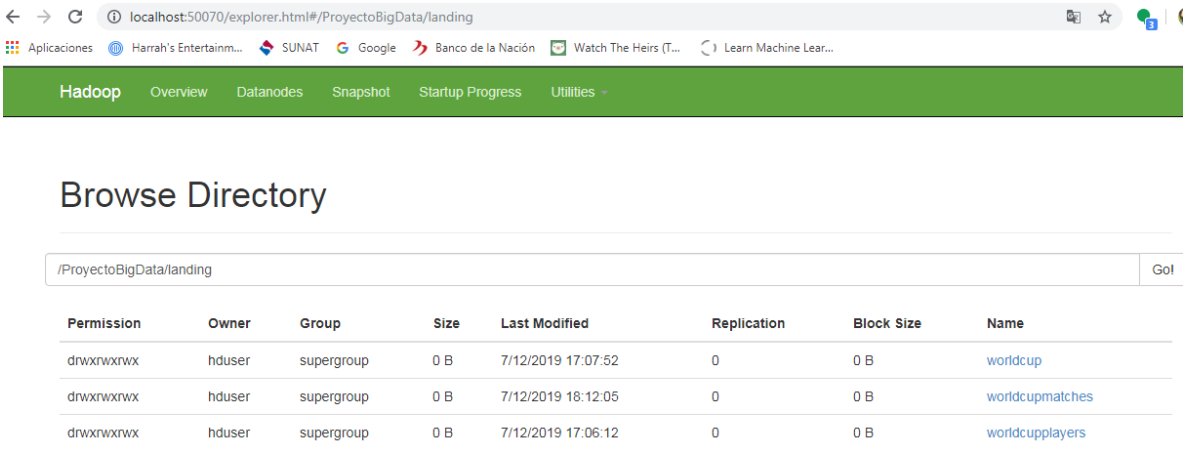
```
hduser@srvbigdata:~$ hdfs dfs -chmod -R 777 /ProyectoBigData
hduser@srvbigdata:~$
```

4. Pasando los archivos de la máquina local al Linux

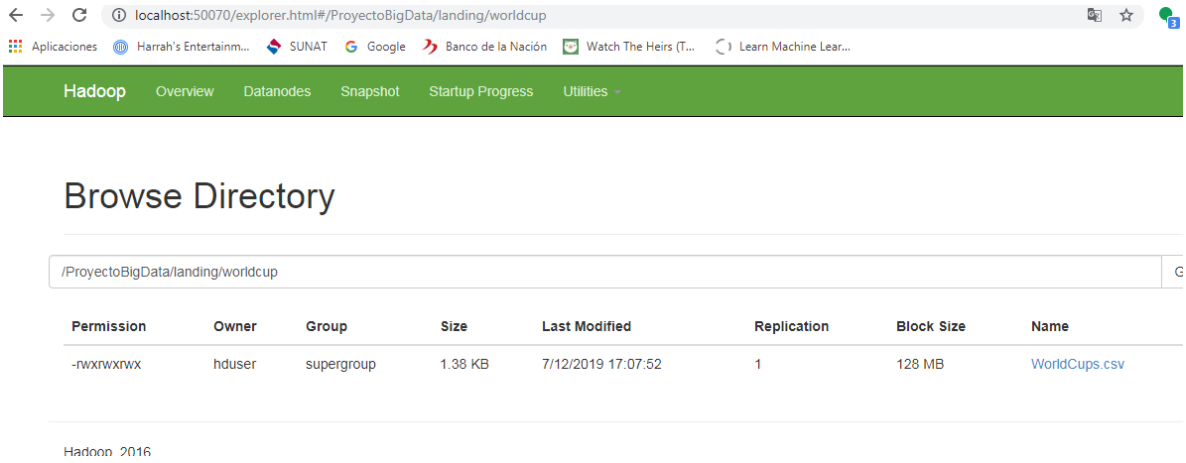



5. Pasando los archivos a hdfs






6. Viendo uno de los archivos en su tuta:




7. observando el árbol de carpetas creado:

```
hduser@srvbigdata:~$ hdfs dfs -ls -R /ProyectoBigData/
drwxrwxrwx   - hduser supergroup          0 2019-12-07 17:14 /ProyectoBigData/landing
drwxrwxrwx   - hduser supergroup          0 2019-12-07 17:07 /ProyectoBigData/landing/worldcup
-rwxrwxrwx   1 hduser supergroup       1412 2019-12-07 17:07 /ProyectoBigData/landing/worldcup/WorldCups.csv
drwxrwxrwx   - hduser supergroup          0 2019-12-07 18:12 /ProyectoBigData/landing/worldcupmatches
-rw-r--r--   1 hduser supergroup     156902 2019-12-07 18:12 /ProyectoBigData/landing/worldcupmatches/WorldCupMa
tches.csv
drwxrwxrwx   - hduser supergroup          0 2019-12-07 17:06 /ProyectoBigData/landing/worldcupplayers
-rwxrwxrwx   1 hduser supergroup    2144229 2019-12-07 17:06 /ProyectoBigData/landing/worldcupplayers/WorldCupPl
ayers.csv
drwxrwxrwx   - hduser supergroup          0 2019-12-07 01:54 /ProyectoBigData/reportes
drwxrwxrwx   - hduser supergroup          0 2019-12-07 01:53 /ProyectoBigData/smart
drwxrwxrwx   - hduser supergroup          0 2019-12-07 01:54 /ProyectoBigData/universo
hduser@srvbigdata:~$
```


1. Conectarse a Hive
2. Usar el archivo **script_hive** para la creación de la base de datos y sus tablas.

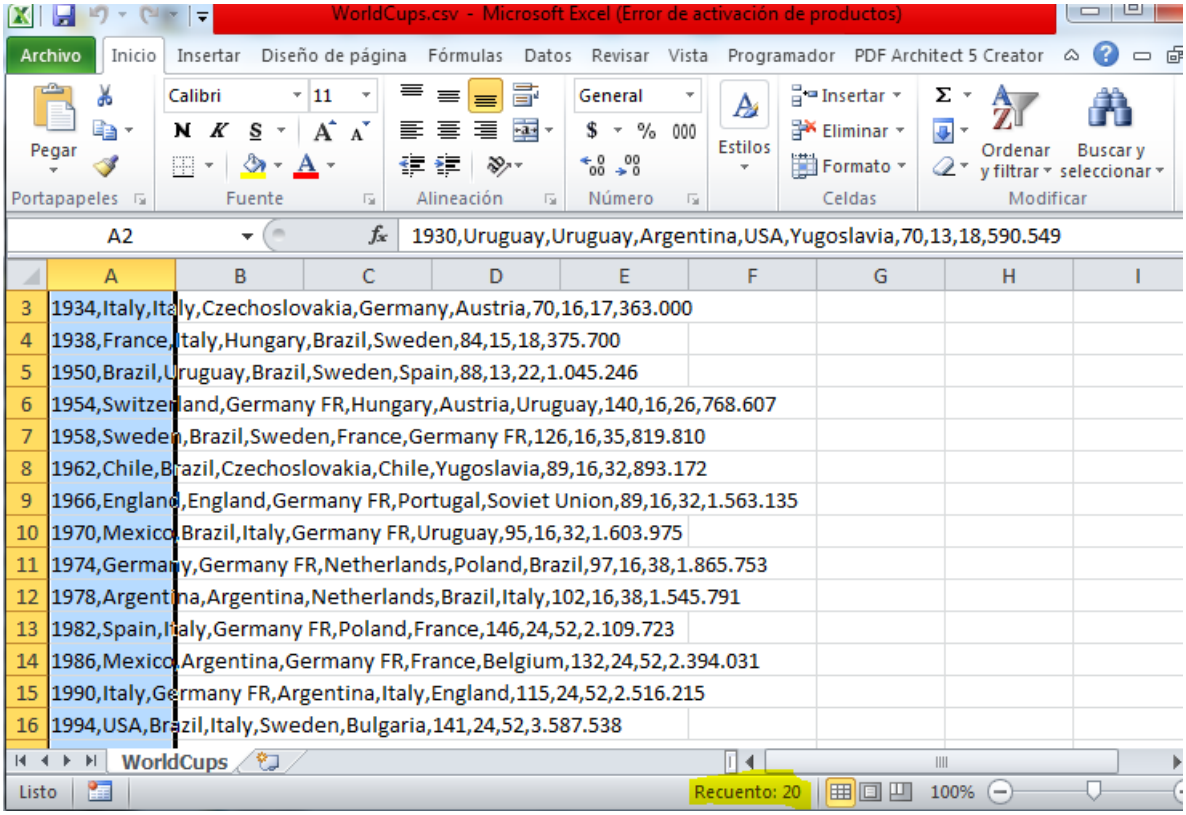
```
Logging initialized using configuration in jar:file:/home/hduser/hive/lib/hive-common-2.1.1.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X relea
hive> create database proyectobigdata;
OK
Time taken: 4.761 seconds
hive>
```

3. Creación de tablas

```
hive> create external table proyectobigdata.worldcup
    > (Year int,
    > Country string ,
    > Winner string ,
    > RunnersUp string ,
    > Third string ,
    > Fourth string ,
    > GoalsScored int ,
    > QualifiedTeams int,
    > MatchesPlayed int,
    > Attendance int
    > )ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > STORED AS TEXTFILE
    > LOCATION '/ProyectoBigData/landing/worldcup'
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.285 seconds
hive> create table proyectobigdata.worldcupplayers
    > ( RoundID int,
    > MatchID int,
    > TeamInitials string,
    > CoachName string,
    > LineUp string ,
    > ShirtNumber int,
    > PlayerName string,
    > Position string ,
    > Event string
    > )ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > STORED AS TEXTFILE
    > LOCATION '/ProyectoBigData/landing/worldcupplayers'
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.209 seconds
hive>
```

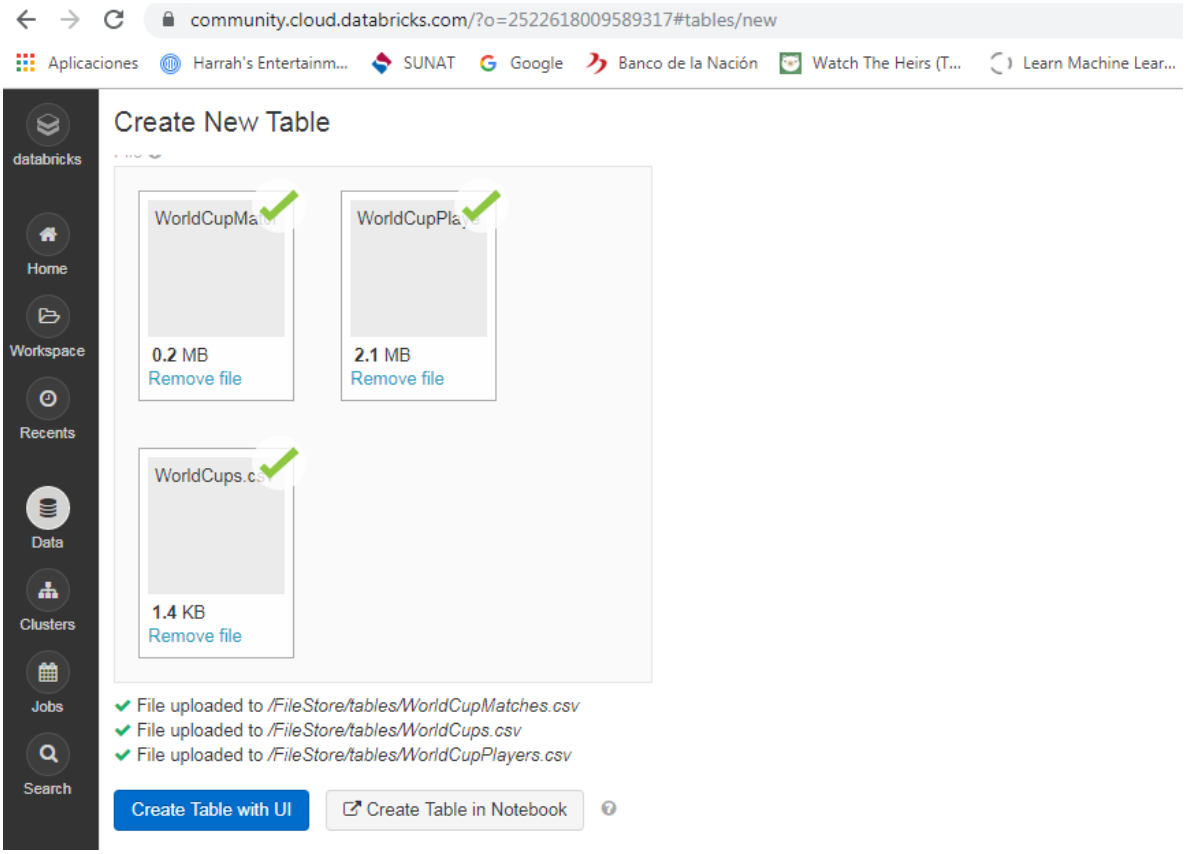
```
hive> create external table proyectobigdata.worldcupmatches
    > (Year  int  ,
    > Datetime  string  ,
    > Stage  string  ,
    > Stadium  string  ,
    > City  string  ,
    > HomeTeamName  string  ,
    > HomeTeamGoals  int  ,
    > AwayTeamGoals  int  ,
    > AwayTeamName  string  ,
    > Winconditions  string  ,
    > Attendance  int  ,
    > HalfTimeHomeGoals  int  ,
    > HalfTimeAwayGoals  int  ,
    > Referee  string  ,
    > Assistant1  string  ,
    > Assistant2  string  ,
    > RoundID  int  ,
    > MatchID  int  ,
    > HomeTeamInitials string ,
    > AwayTeamInitials string )
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY '\;' LINES TERMINATED BY '\n'
    > STORED AS TEXTFILE
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.284 seconds
hive> load data inpath '/ProyectoBigData/landing/worldcupmatches/WorldCupMatches.csv' into table proyectobigdata.worldcupmatches;
Loading data to table proyectobigdata.worldcupmatches
OK
Time taken: 0.582 seconds
```

4. Verificando el número de registros comparando con el csv:



```
hive> select  count(*) from proyectobigdata.worldcup ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a differen
ne (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hduser_20191214034553_5acf12f3-c806-4cd8-9108-800a6f884009
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1576310675085_0001, Tracking URL = http://srvbigdata.edutronic.com:8088/proxy/application_157631067
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1576310675085_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-12-14 03:48:12,999 Stage-1 map = 0%,  reduce = 0%
2019-12-14 03:49:13,875 Stage-1 map = 0%,  reduce = 0%
2019-12-14 03:49:30,008 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.31 sec
2019-12-14 03:50:18,299 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 12.31 sec
2019-12-14 03:50:30,919 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 20.03 sec
MapReduce Total cumulative CPU time: 20 seconds 30 msec
Ended Job = job_1576310675085_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 20.03 sec   HDFS Read: 10151 HDFS Write: 102 SUCCESS
Total MapReduce CPU Time Spent: 20 seconds 30 msec
OK
20
Time taken: 287.391 seconds, Fetched: 1 row(s)
hive>
```


1. Se subirá los archivos al Filestore de Databricks :




Ruta de archivos en databricks:

File uploaded to */FileStore/tables/WorldCupMatches.csv*
File uploaded to */FileStore/tables/WorldCups.csv*
File uploaded to */FileStore/tables/WorldCupPlayers.csv*

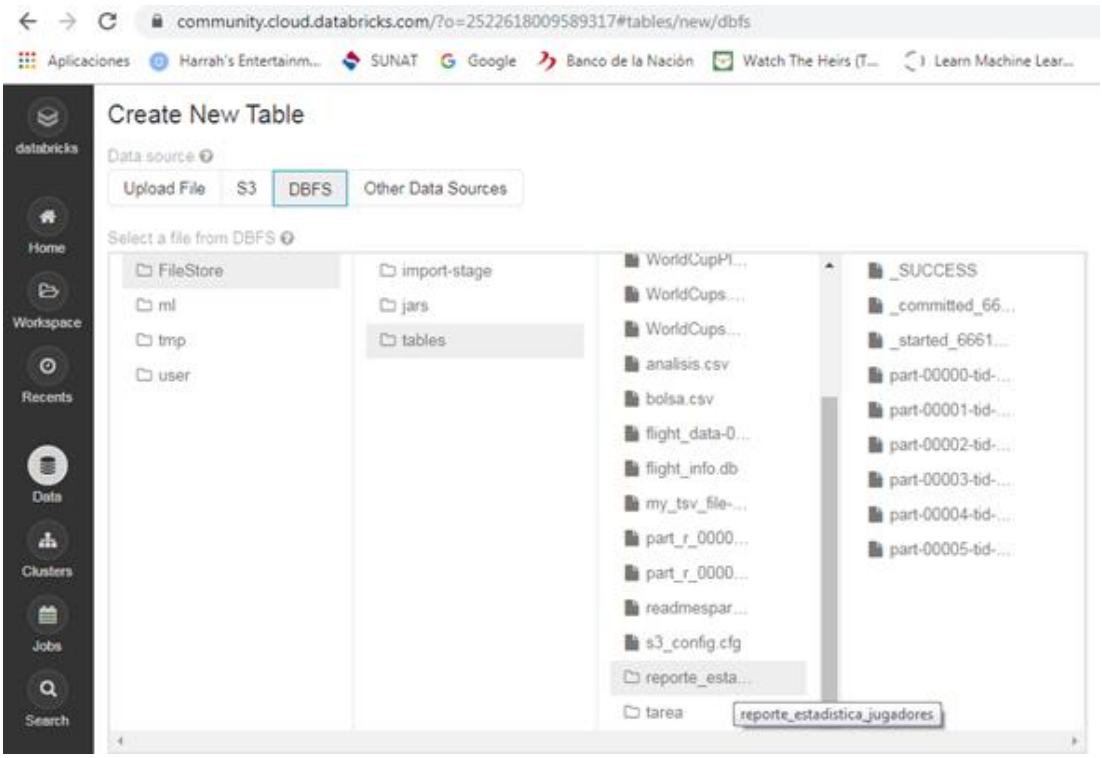
2. Usar el archivo etl_proyectobigdata.py para realizar el etl de los csv en spark Dataframes para la posterior generación de los reportes. Este archivo se creó como un notebook en Databricks.

En el archivo etl_proyectobigdata.pynb es el mismo archivo en formato notebook y el archivo

etl_proyectobigdata.html es en formato web.

3. Revisión de los reportes guardados en el DBFS de databricks

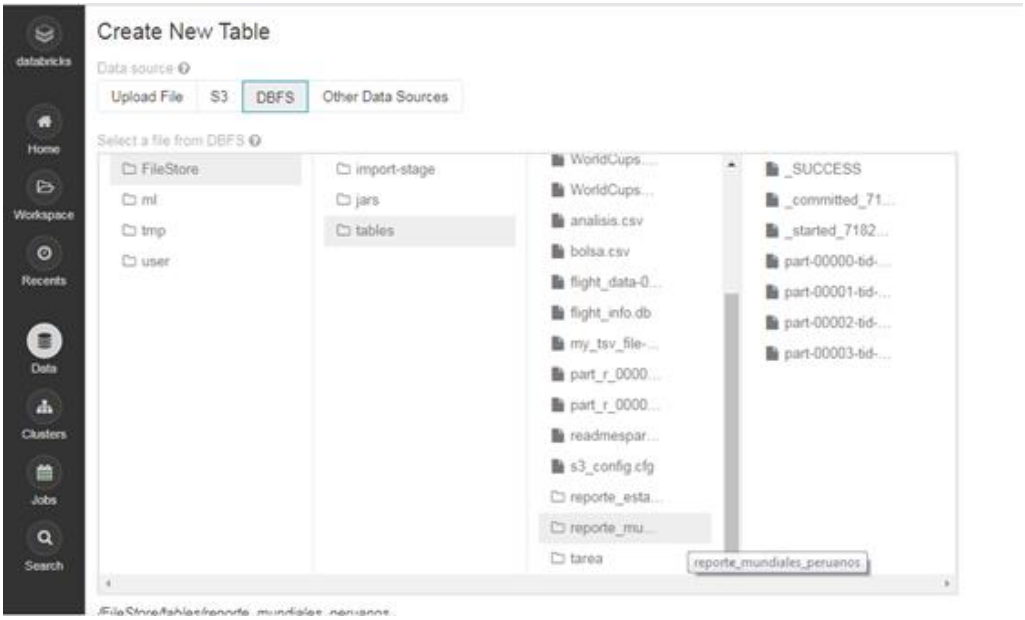
**Reporte de estadística de jugadores:** Este reporte contiene el número de goles, penales, penales fallidos y tarjetas rojas por jugador de los mundiales. Sirve como estadística para otorgar los premios respectivos a los jugadores con mejor desempeño.




| | NOMBRE_JUGADOR | POSICION_JUGADOR | INICIALES_PAIS | NUMERO_GOLES | NUMERO_PENALES | NUMERO_PENALES_FALLADOS | NUMERO_TARJETAS_ROJAS |
|---|---|---|---|---|---|---|---|
| 8376 | Vava | Other | BRA | 9 | 0 | 0 | 0 |
| 8377 | Paolo Rossi | Other | ITA | 9 | 0 | 0 | 0 |
| 8378 | Jairzinho | Other | BRA | 9 | 0 | 0 | 0 |
| 8379 | Ademir | Other | BRA | 8 | 0 | 0 | 0 |
| 8380 | Guillermo Stabile | Other | ARG | 8 | 0 | 0 | 0 |

**Reporte de mundiales en los que Perú participó :**

Con la nueva partipación de Perú en el último mundial la Federación Peruana de Fútbol solicitó a la FIFA un reporte de los últimos mundiales en los que Perú participó para recordar a todos sus jugadores.




Vista del reporte en un dataframe Pandas

| | MUNDIAL | FECHA_PARTIDO | EQUIPO DE CASA | EQUIPO EXTERNO | NOMBRE_JUGADOR | POSICION_JUGADOR | INICIALES_PAIS | NOMBRE_REFEREE |
|---|---|---|---|---|---|---|---|---|
| 0 | Argentina-1978 | 03 Jun 1978 | Peru | Scotland | Ramon Quiroga | Goalkeeper | PER | Eriksson Ulf (swe) |
| 1 | Argentina-1978 | 03 Jun 1978 | Peru | Scotland | Jaime Duarte | Other | PER | Eriksson Ulf (swe) |
| 2 | Argentina-1978 | 03 Jun 1978 | Peru | Scotland | Rodolfo Manzo | Other | PER | Eriksson Ulf (swe) |
| 3 | Argentina-1978 | 03 Jun 1978 | Peru | Scotland | Hector Chumpitaz | Captain | PER | Eriksson Ulf (swe) |
| 4 | Argentina-1978 | 03 Jun 1978 | Peru | Scotland | Ruben Diaz | Other | PER | Eriksson Ulf (swe) |