BASES DE DATOS AVANZADAS

DATA WAREHOUSE

Miguel Millán Alhambra
Desiré Saponaro
Antonio Manjavacas Lucas

Laboratory Book

Día 11 de febrero de 2020 (1 hora)

Después de un análisis previo, hoy hemos decidido el conjunto de datos a utilizar. Estos han sido estudiados, cargados en Google Collab y filtrados de acuerdo a las características con las que nos interesa trabajar:

- Carga de datos:

```
def upload_files ():
  uploaded = files.upload()
  for fn in uploaded.keys():
    print('User uploaded file "{name}" with length {length} bytes'.format(
        name = fn, length = len(uploaded[fn])))
    df = pd.read_csv(io.StringIO(uploaded[fn].decode('utf-8')))
    return df
```

```
data = upload_files()
data.head()
```

- Selección de características:

```
features = ['id', 'original_title','budget','genres',
            'original_language','popularity','production_companies',
            'release_date','revenue','vote_average']

df = data[features].copy()
df.head()
```

Las principales dificultades que hemos encontrado han sido: conocer las funciones de la librería pandas para el tratamiento de datasets y estudiar el formato de cada uno de los campos.

Día 13 de febrero de 2020 (1 hora)

En el día de hoy, hemos decidido la planificación del proyecto, teniendo en cuenta las fechas de entrega futuras y la disponibilidad de cada uno de los miembros del equipo.

11/2	**12/2**	**13/2**	**14/2**	**15/2**	**16/2**	**17/2**
Extracción de datos y atributos (ETL)		Planificación del proyecto		**Fecha límite planificación**		Limpieza de datos y carga en MongoDB Atlas
18/2	**19/2**	**20/2**	**21/2**	**22/2**	**23/2**	**24/2**
Fecha límite entrega parcial	Diseño de la base de datos y consultas	Diseño de la base de datos y consultas			Hacer presentación	**Fecha límite almacén de datos**

En próximas reuniones comenzaremos el diseño y elaboración del almacén de datos.

Día 18 de febrero de 2020 (2 horas)

Hemos dedicado el día de hoy al preprocesamiento de los datos. Concretamente, se han llevado a cabo las siguientes labores de preprocesamiento:

- Eliminación de registros con valores nulos.

```
df = df.dropna()
```

- Conversión de los registros de la columna "*genres*" y "*production_companies*" en listas.

```
for i, row in df.iterrows():
  genre = row['genres']
  genreList = json.loads(genre)
  listGenre = set()
  for g in genreList:
    listGenre.add(g["name"])
  df.at[i,'genres'] = listGenre
```

```
for i, row in df.iterrows():
  companie = row['production_companies']
  companiesList = json.loads(companie)
  listCompanie = set()
  for pc in companiesList:
    listCompanie.add(pc["name"])
  df.at[i,'production_companies'] = listCompanie
```

- Modificación del formato de las fechas de *AA/MM/DD* a *DD/MM/AA*.

```
for i, row in df.iterrows():
  date = row['release_date']
  x = str(date).split('-')
  new_date = x[2] + '-' + x[1] + '-' + x[0]
  df.at[i,'release_date'] = new_date
```

- Se han añadido las columnas "*month*" y "*year*" para cada película.

```
months = {
  1: 'jan',
  2: 'feb',
  3: 'mar',
  4: 'apr',
  5: 'may',
  6: 'jun',
  7: 'jul',
  8: 'aug',
  9: 'sep',
  10: 'oct',
  11: 'nov',
  12: 'dec'
}

for i,row in df.iterrows():
  date = row['release_date']
  x = str(date).split('-')
  df.at[i,'month'] = months[int(x[1])]
  df.at[i,'year'] = x[2]
```

Finalmente, se ha trabajado en la conexión con la base de datos y la carga de los mismos en MongoDB Atlas. Hemos encontrado problemas en la conexión mediante el cliente de MongoDB, algo que trataremos de solucionar en la próxima reunión.

Día 20 de febrero de 2020 (2 horas)

Durante el día de hoy se ha implementado el código correspondiente a las Dimensiones 1 y 5 de nuestro almacén de datos. También se han preparado las secciones de código empleadas para la recuperación de las vistas:

```
query_dim1 = pd.DataFrame()

for document in db.earnings_movies.find({}, {"_id": 0}):
  query_dim1 = query_dim1.append(document, ignore_index=True)

query_dim1 = query_dim1[['year', 'film', 'earnings']]
query_dim1.sort_values(by='year')
```

	year	film	earnings
0	1916	Intolerance	8008844.0
1	1925	The Big Parade	21755000.0
2	1927	Metropolis	-91969578.0
3	1929	The Broadway Melody	3979000.0
4	1930	Hell's Angels	4050000.0
...	...	...	...

Día 25 de febrero de 2020 (2 horas)

Hoy se han finalizado el resto de dimensiones (2, 3 y 4). Aunque han supuesto una mayor complejidad, hemos logrado obtenerlas sin problemas. Los datos correspondientes a las mismas han sido introducidos en la base de datos.

Dimensión 4

`PRODUCTION_COMPANY | EARNINGS`

Ganancias totales por productora:

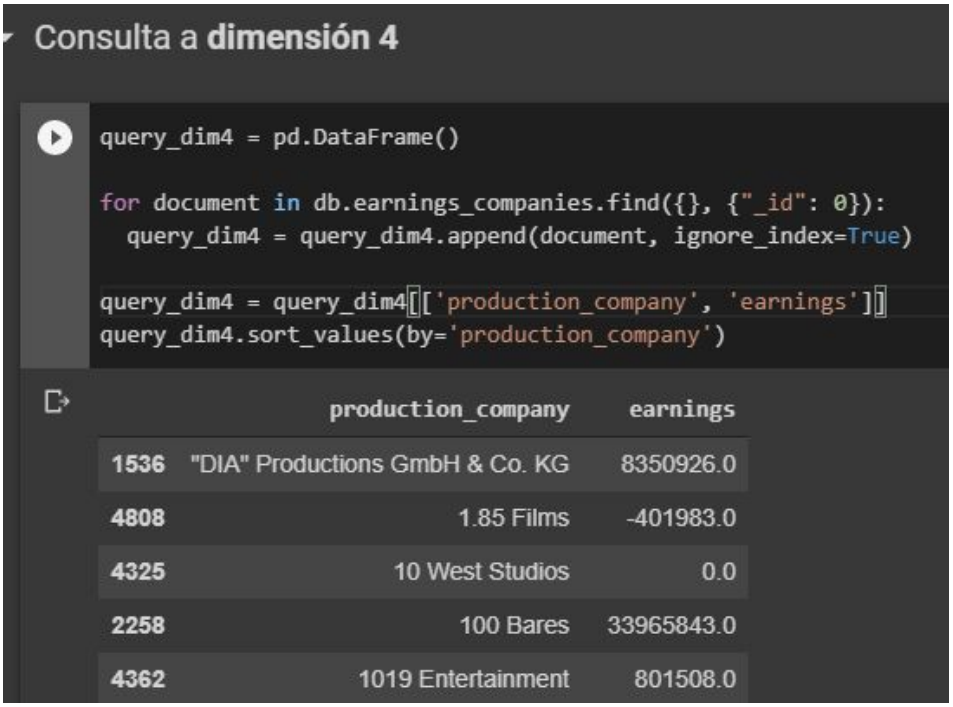
```
[72]  all_companies = []
      for list_companies in df['production_companies']:
        for companies in list_companies:
          if companies not in all_companies:
            all_companies.append(companies)
      all_companies
```

```
[75]  info_companies = {}
      for companie in all_companies:
        key_earns = companie + '_earns'
        info_companies[key_earns] = []
      info_companies
```

En próximas sesiones se completará el código para la obtención de las vistas mediante consultas a la base de datos y se prepararán las transparencias y la memoria.

Día 27 de febrero de 2020 (1’5 horas)

Hoy se han realizado las consultas a las distintas dimensiones para mostrar los datos previamente calculados.



Además, se ha comenzado el desarrollo de la memoria, indicando las fuentes de datos, el diseño del almacén de datos y el procedimiento para la limpieza y carga de datos.

Día 2 de marzo de 2020 (2 horas)

Hoy hemos realizado la memoria del trabajo así como preparado la presentación del mismo. También se ha creado el repositorio de GitHub donde se ha alojado el código y memoria del trabajo, dándolo así por concluido.

REPARTO DE PUNTOS

Antonio Manjavacas	Miguel Millán	Desiré Saponaro
33.33%	33.33%	33.33%