Examen de Procesamiento

Certificación en Procesamiento con Scala

México, Marzo 2023

Instrucciones

De acuerdo a los temas vistos en el Curso: **Certificación en Procesamiento con Scala,** desarrolla la solución para el siguiente caso:

A un programador fanático de los Simpsons le interesa hacer un resumen de esta loca familia de Springfield y entregar premios a lo mejor de esta serie. Se interesa por premiar a los ganadores para las siguientes categorías:

- **Mejor temporada.**
- **Mejor Año.**
- **Mejor capítulo.**
- **Top 3 de los mejores capítulos por temporada.**

La **rama** a utilizar para este examen se llama: **examenMarzo2023**

Las tablas proporcionadas son:

- **src/test/resources/data/input/csv/t_fdev_simpsons.csv**: Contiene las estadísticas de las temporadas de los simpsons.

- **/src/test/resources/schema/t_fdev_simpsons.input.schema**: Contiene el schema del csv input **t_fdev_simpsons.csv**.

El schema de salida que debe cumplir tu dataframe es el siguiente:

- **src/test/resources/schema/t_fdev_winners.output.schema**: Contiene los campos de salida asociados.

El archivo para ingresar los parámetros iniciales son los siguientes **src/main/resources/config** y **src/test/resources/config**

Estas son las reglas que deberán implementarse sobre los datasets, recuerda organizar tu código en varios métodos, no dejes todo en único método:

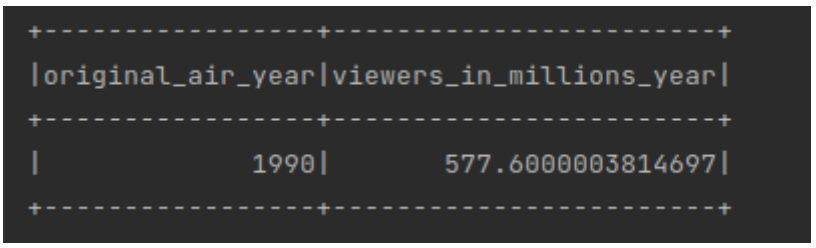
1. Limpiar el dataset, quitando los registros que tengan nulos en las columnas ***rating, votes y viewers_in_millions***.

2. Obtener la mejor temporada de los simpsons de acuerdo al mayor ***rating***.

```
+------+------------------+
|season|      total_rating|
+------+------------------+
|     7|208.10000038146973|
+------+------------------+
```

Resultado esperado, mejor temporada.

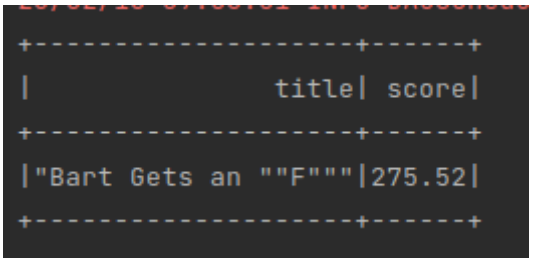
3. Obtener el mejor año de los simpsons de acuerdo al número de vistas de los capítulos de dicho año, de acuerdo a los siguientes criterios:
    a. Para saber a qué año pertenece un capítulo toma como referencia la columna ***original_air_year***.
    b. Considera que la columna ***viewers_in_millions*** indica el número de vistas que tuvo cada capítulo**.**



```
+-----------------+-----------------------+
|original_air_year|viewers_in_millions_year|
+-----------------+-----------------------+
|             1990|       577.6000003814697|
+-----------------+-----------------------+
```

Resultado esperado, mejor año.

4. Obtener el mejor capítulo: de acuerdo a las siguientes reglas.
    a. Agrega una columna ***score***, de acuerdo a la siguiente fórmula: ***rating*viewers_in_millions.***
    b. El mejor capítulo será el que tenga mayor ***score.***



```
+--------------------+------+
|               title| score|
+--------------------+------+
|"Bart Gets an ""F"""|275.52|
+--------------------+------+
```

Resultado esperado, mejor capítulo.

5. Obtener el top 3 de capítulos por cada temporada de acuerdo a su ***score***, guardar el resultado en formato parquet, de acuerdo a los siguientes criterios:
    a. La última temporada no tiene suficiente información para sacar un top 3, entonces tome en cuenta que si una temporada no tiene al menos 3 capítulos no se incluye en el resultado.
    b. Almacenar estos tops en un particionado por ***season***, tomar los datos de acuerdo al schema de salida de la ruta **src/test/resources/schema/t_fdev_winners.output.schema** y almacene el archivo en formato parquet en la ruta **src/test/resources/data/output.**

season	title	number_in_season	original_air_date	original_air_year	production_code	rating	votes	viewers_in_millions	score	top
1	Krusty Gets Busted	12	1990-04-29	1990	7612	8.3	1716	30.4	252.32	1
1	Life on the Fast ...	9	1990-03-18	1990	7611	7.5	1578	33.5	251.25	2
1	The Crepes of Wrath	11	1990-04-15	1990	7613	7.8	1539	31.2	243.36002	3
2	"Bart Gets an ""F"""	1	1990-10-11	1990	7F03	8.2	1638	33.6	275.52	1
2	Simpson and Delilah	2	1990-10-18	1990	7F02	8.3	1588	29.9	248.17	2
2	Treehouse of Horror	3	1990-10-25	1990	7F04	8.2	1786	27.4	224.68	3
3	Homer at the Bat	17	1992-02-20	1992	8F13	8.6	1637	24.6	211.56001	1
3	Flaming Moe's	10	1991-11-21	1991	8F08	8.8	1618	23.9	210.32	2
3	Radio Bart	13	1992-01-09	1992	8F11	8.5	1365	24.2	205.70001	3
4	Lisa's First Word	10	1992-12-03	1992	9F08	8.5	1350	28.6	243.1	1
4	Duffless	16	1993-02-18	1993	9F14	8.3	1209	25.7	213.31001	2
4	Mr. Plow	9	1992-11-19	1992	9F07	8.8	1595	24.0	211.20001	3
5	Treehouse of Horr...	5	1993-10-28	1993	1F04	8.7	1437	24.0	208.79999	1
5	Homer and Apu	13	1994-02-10	1994	1F10	8.3	1171	21.8	180.94	2
5	Cape Feare	2	1993-10-07	1993	9F22	9.0	2010	20.0	180.0	3
6	Treehouse of Horr...	6	1994-10-30	1994	2F03	9.0	1690	22.2	199.8	1
6	Homer the Great	12	1995-01-08	1995	2F09	8.9	1457	20.1	178.89	2
6	Bart's Comet	14	1995-02-05	1995	2F11	8.6	1221	18.7	160.82	3
7	Treehouse of Horr...	6	1995-10-29	1995	3F04	8.5	1304	19.7	167.45001	1
7	King-Size Homer	7	1995-11-05	1995	3F05	9.0	1633	17.0	153.0	2
7	Two Bad Neighbors	13	1996-01-14	1996	3F09	8.7	1264	16.5	143.55	3
8	The Springfield F...	10	1997-01-12	1997	3601	9.0	1793	20.9	188.09999	1
8	Treehouse of Horr...	1	1996-10-27	1996	4F02	8.3	1186	18.3	151.89	2
8	Simpsoncalifragil...	13	1997-02-07	1997	3603	7.7	1088	17.7	136.29001	3
9	The Principal and...	2	1997-09-28	1997	4F23	7.4	1158	14.9	110.26	1
9	Lisa's Sax	3	1997-10-19	1997	3602	8.1	980	12.9	104.490005	2
9	The City of New Y...	1	1997-09-21	1997	4F22	9.1	1918	10.5	95.55	3

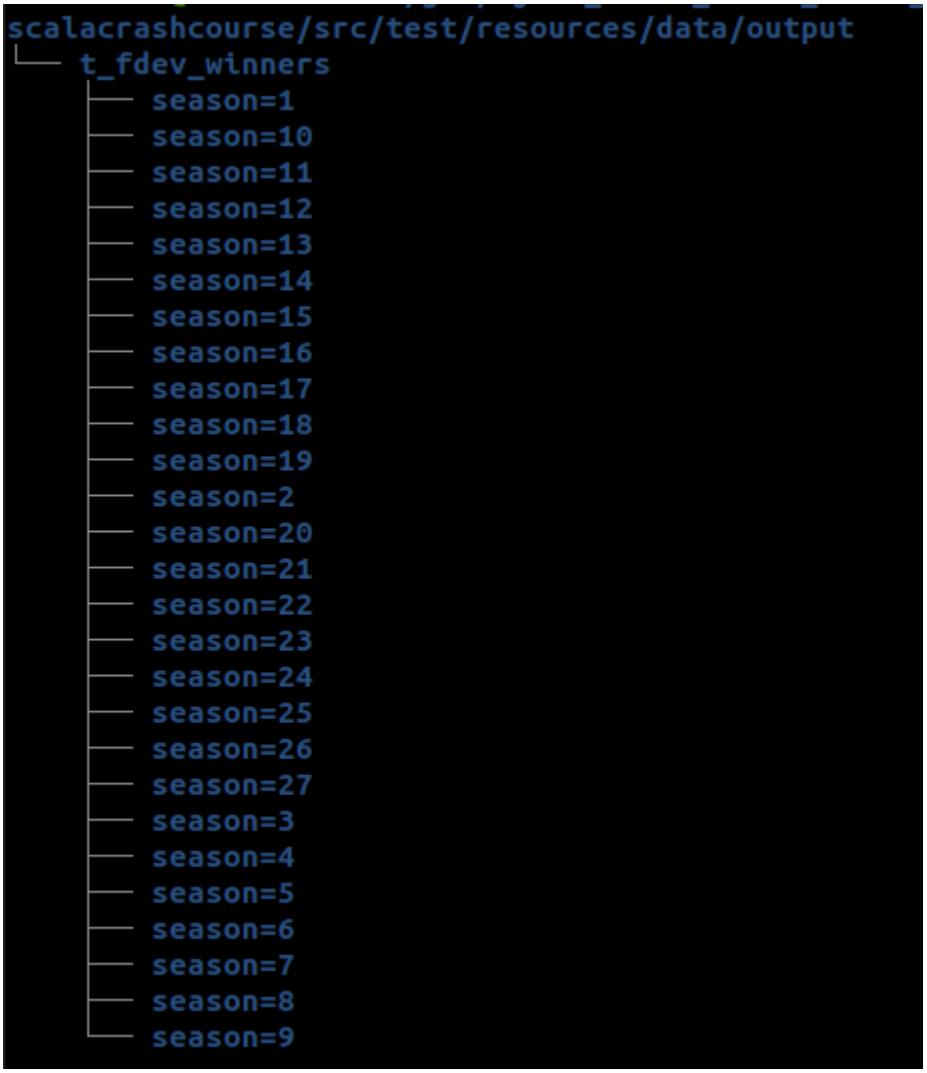
Resultado esperado, Top 3 por temporada.

PRUEBAS UNITARIAS

Toma un método y ejecuta una prueba unitaria para el caso success y otra para un caso failed.

PRUEBAS DE ACEPTACIÓN

Realiza las comprobaciones que creas pertinentes para que el negocio apruebe tu desarrollo, aquí una muestra del resultado esperado.



Particionado en parquet esperado.

EVIDENCIAS:

Crea la **PR** con tu solución hacia la rama del examen (**examenMarzo2023**) incluyendo tus parquets resultantes, posteriormente dentro de la descripción del classroom dónde

bajaste este examen, también encontrarás un link que te llevará a la carpeta de "**evidencias"** y en esta deberás crear una carpeta nueva con tu **nombre completo** donde subirás lo siguiente:

- Sube una imagen con el resultado (show) del paso 2 al 4.
- Agrega también la evidencia de tus ejecuciones de pruebas unitarias y pruebas de aceptación.

NOTA IMPORTANTE:

Sin evidencias NO se tomará en cuenta el trabajo realizado. Recuerda que NO se aceptarán documentos fuera de horario.

¡Mucho éxito!