DATA MANAGEMENT

SQL Segmento 1

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
Centro de Referencia Nacional en Comercio Electrónico y Marketing CRN Digital
GARANTÍA JUVENIL
UNIÓN EUROPEA
Barrabés
The Valley
"El FSE invierte en tu futuro"
Fondo Social Europeo

Índice

Antes de empezar...

Antes de empezar...

Métricas y dimensiones (variables numéricas y categóricas)

Una *métrica* es una variable donde el valor tiene un significado numérico; una *dimensión*, por el contrario, es aquella variable (normalmente en formato texto) que expresa una propiedad cualitativa, por tanto el rango de valores que puede adquirir es finito (limitado). Los términos métricas y dimensiones, comúnmente utilizados en analítica web, equivalen a los conceptos de variables numéricas y categóricas en estadística.

	Dimensiones		Métricas	
Row	Company	Sector	Employees	Revenue
1	Disney	Media	185000	52465
2	Twenty-First Century Fox	Media	20500	28987
3	Time Warner	Media	24800	28118
4	CBS	Media	19080	13886
5	Viacom	Media	9445	13268
6	Live Nation Entertainment	Media	12200	7246
7	Discovery Communications	Media	7000	6394
8	iHeartMedia	Media	18700	6242
9	Liberty Media	Media	3503	4795
10	TEGNA	Media	10020	3242

Antes de empezar...

Unidad de observación

Una *unidad de observación* es la persona o cosa a la cual corresponde la información provista en forma de métricas. Cada tabla de datos contiene una (solo una) y está compuesta por la combinación de todas las dimensiones. En el ejemplo debajo, las métricas reportan información para "empresas" y "sectores"; dado que las combinaciones entre empresas y sectores son únicas (cada empresa pertenece a un único sector), podríamos simplificar y decir que la unidad de observación de la tabla es "empresas".

Unidad de observación

Row	Company	Sector	Employees	Revenue
1	Disney	Media	185000	52465
2	Twenty-First Century Fox	Media	20500	28987
3	Time Warner	Media	24800	28118
4	CBS	Media	19080	13886
5	Viacom	Media	9445	13268
6	Live Nation Entertainment	Media	12200	7246
7	Discovery Communications	Media	7000	6394
8	iHeartMedia	Media	18700	6242
9	Liberty Media	Media	3503	4795
10	TEGNA	Media	10020	3242

Statements & Clauses Básicos

Statements & Clauses Básicos

Definiciones en esta sección

Existen dos *statements* (comandos) básicos que son obligatorios a ser incluidos en cada *query* (consulta de datos)

- **SELECT:** determina qué campos (columnas) serán extraídos de aquellos disponibles en una tabla de datos; los campos deben ser separados por coma. *Nota: las comas únicamente separan campos, por tanto el SELECT statement no debe comenzar con coma*
- **FROM:** especifica de qué tabla, de aquellas disponibles en una base de datos, estamos extrayendo información

Statements & Clauses Básicos

SELECT y FROM statements

SQL Query

```sql
SELECT
    Company
    ,Sector
    ,Employees
    ,Revenue
FROM `isdi-mda-256409.SQL_Basics.fortune` -- this is the name of the data table
```

Output

Row	Company	Sector	Employees	Revenue
1	Disney	Media	185000	52465
2	Twenty-First Century Fox	Media	20500	28987
3	Time Warner	Media	24800	28118
4	CBS	Media	19080	13886
5	Viacom	Media	9445	13268
6	Live Nation Entertainment	Media	12200	7246
7	Discovery Communications	Media	7000	6394
8	iHeartMedia	Media	18700	6242
9	Liberty Media	Media	3503	4795
10	TEGNA	Media	10020	3242

Statements & Clauses Básicos

Definiciones en esta sección

La **LIMIT** clause restringe el número de filas a ser incluidas en el resultado final. *Nota: este comando no cambia el orden de las filas, simplemente limita el resultado de las filas.*

De forma opcional, podemos renombrar los campos presentes en el SELECT *statement* utilizando el comando **AS**.

El **DISTINCT** *clause* en el SELECT *statement* fuerza valores únicos para el total de campos.

Nota: importante entender que el DISTINCT clause, al momento de evaluar si una fila es duplicada, tendrá en cuenta el total de las columnas en el output.

Statements & Clauses Básicos

LIMIT clause

SQL Query

```sql
SELECT
    Company
    ,Sector
    ,Employees
    ,Revenue
FROM `isdi-mda-256409.SQL_Basics.fortune` -- this is the name of the data table

LIMIT 3 -- we limit the number of rows in the output to only 3
```

Output

Row	Company	Sector	Employees	Revenue
1	Disney	Media	185000	52465
2	Twenty-First Century Fox	Media	20500	28987
3	Time Warner	Media	24800	28118

Statements & Clauses Básicos

Renombrando un campo (columna) AS

SQL Query

```sql
SELECT
    Company as comp -- we tag each field using the "as" command
    ,Sector as sect
    ,Employees as n_of_employees
    ,Revenue as revenue_in_2020
FROM `isdi-mda-256409.SQL_Basics.fortune` -- this is the name of the data table

LIMIT 3 -- we limit the number of rows in the output to only 3
```

Output

Row	comp	sect	n_of_employees	revenue_in_2020
1	Disney	Media	185000	52465
2	Twenty-First Century Fox	Media	20500	28987
3	Time Warner	Media	24800	28118

Statements & Clauses Básicos

DISTINCT clause

SQL Query

```
1  SELECT DISTINCT -- the DISTINCT clause forces unique values
2      Sector
3  FROM `isdi-mda-256409.SQL_Basics.fortune` -- this is the name of the data table
```

Output

Row	Sector
1	Media
2	Energy
3	Apparel
4	Chemicals
5	Materials
6	Retailing

Where Clause

Where Clause

Definiciones en esta sección

El **WHERE** *clause* establece condiciones de filtrado para las filas en nuestro resultado final. Las condiciones están separados por operadores **AND/OR**.

Nota: las condiciones en un WHERE clause no están separadas por coma.

Where Clause

Definiciones en esta sección

Un operador en computación es una expresión que permite realizar ejecuciones matemáticas o lógicas. Dentro de un WHERE *clause* se utilizan para definir criterios de filtrado en condiciones.

COMPARISON OPERATORS	DESCRIPTION
=	Equal to
>	Greater than
<	Less than
>=	Greater than or equal to
<=	Less than or equal to
<>	Not equal to

LOGICAL OPERATORS	DESCRIPTION
AND	TRUE if all conditions are met
OR	TRUE if at least one condition is met
BETWEEN *(X AND Y)*	TRUE if value within range
NOT	TRUE if condition is NOT met
IN *(X,Y,Z)*	TRUE if value is contained in list
LIKE *('pattern')*	TRUE if value matches pattern

Where Clause

Operadores: comparación & lógicos

SQL Query

```sql
SELECT
    Company
    ,Sector
    ,Employees
    ,Revenue
FROM `isdi-mda-256409.SQL_Basics.fortune` -- this is the name of the data table
WHERE
    Employees >100000                                -- condition 1: filtering a numeric variable
    AND Sector IN('Media','Retailing','Technology') -- condition 2: filtering a categorical variable
```

Output

Row	Company	Sector	Employees	Revenue
1	Disney	Media	185000	52465
2	Walmart	Retailing	2300000	482130
3	Target	Retailing	341000	73785
4	Macy's	Retailing	157500	27079
5	Sears Holdings	Retailing	178000	25146
6	Dollar General	Retailing	113400	20369

Order By Estatment

Order By Estatment

Definiciones en esta sección

El **ORDER BY** statement incluye condiciones que permiten establecer un criterio de orden en nuestro resultado final.

Por defecto, los criterios de orden son siempre en sentido ascendente, podemos revertirlo utilizando el comando DESC después del campo. Para campos de tipo texto, el orden será alfabético.

Nota: las condiciones en un ORDER BY statement van separadas por coma.

Order By Estatment

SQL Query

```sql
SELECT
    Company
    ,Sector
    ,Employees
    ,Revenue
FROM `isdi-mda-256409.SQL_Basics.fortune` -- this is the name of the data table
WHERE
    Employees >100000                                -- condition 1: filtering a numeric variable
    AND Sector IN('Media','Retailing','Technology') -- condition 2: filtering a categorical variable
ORDER BY
    Employees DESC  -- condition 1: sorting by numerical variable (descending)
    ,Company        -- condition 2: sorting by categorical variable (ascending, as default)
LIMIT 10
```

Output

Row	Company	Sector	Employees	Revenue
1	Walmart	Retailing	2300000	482130
2	IBM	Technology	411798	82461
3	Home Depot	Retailing	385000	88519
4	Target	Retailing	341000	73785
5	HP	Technology	287000	103355
6	Amazon.com	Technology	230800	107006
7	Lowe's	Retailing	225000	59074
8	Cognizant Technology Solutions	Technology	221700	12416
9	TJX	Retailing	216000	30945
10	Disney	Media	185000	52465

Case When Estatment

Case When Estatment

Definiciones en esta sección

El **CASE WHEN** statement permite crear campos con condiciones específicas, donde el valor que adquiera cada fila dependerá del criterio establecido. El statement se invoca utilizando la siguiente sintaxis,

CASE

WHEN condition_1 THEN value_if_true

WHEN condition_2 THEN value_if_true

ELSE value_if_none_above_is_true

END as field_name

Case When Estatment

Ejemplo utilizando un CASE WHEN statement

SQL Query

```sql
SELECT
Company
, Revenue
, CASE
    WHEN Revenue > 200000 THEN 'top_revenue_company'
    WHEN Revenue > 100000 THEN 'mid_revenue_company'
    ELSE 'not_in_the_top'
  END AS revenue_segment

FROM `isdi-mda-256409.SQL_Basics.fortune`

ORDER BY Revenue DESC
```

Output

Row	Company	Revenue	revenue_segment
1	Walmart	482130	top_revenue_company
2	Exxon Mobil	246204	top_revenue_company
3	Apple	233715	top_revenue_company
4	Berkshire Hathaway	210821	top_revenue_company
5	McKesson	181241	mid_revenue_company
6	UnitedHealth Group	157107	mid_revenue_company
7	CVS Health	153290	mid_revenue_company
8	General Motors	152356	mid_revenue_company
9	Ford Motor	149558	mid_revenue_company
10	AT&T	146801	mid_revenue_company

Práctica

Práctica

Ex.1: Descripción

Extraer campos *name*, *species* y *homeworld* de la tabla de datos de Star Wars

- *Tabla*: star_wars_characters
- *Descripción de tabla*: esta tabla incluye datos sobre personajes de la saga Star Wars

Resultado esperado

Row	name	species	homeworld
1	Bossk	Trandoshan	Trandosha
2	IG-88	Droid	*null*
3	R5-D4	Droid	Tatooine
4	R2-D2	Droid	Naboo
5	Nute Gunray	Neimodian	Cato Neimoidia
6	Mas Amedda	Chagrian	Champala
7	Adi Gallia	Tholothian	Coruscant
8	Mon Mothma	Human	Chandrila
9	Luke Skywalker	Human	Tatooine
10	Jek Tono Porkins	Human	Bestine IV

Práctica

Ex.2: Descripción

¿Cuáles son los planetas (*homeworlds*) incluídos en la tabla de Star Wars?

- *Tabla*: star_wars_characters
- *Descripción de tabla*: esta tabla incluye datos sobre personajes de la saga Star Wars

Resultado esperado

Row	homeworld
1	Trandosha
2	*null*
3	Tatooine
4	Naboo
5	Cato Neimoidia
6	Champala
7	Coruscant
8	Chandrila
9	Bestine IV
10	Eriadu

Práctica

Ex.3: Descripción

Extraer campos *film*, *director*, *year* y *actor* de la tabla de James Bond; filtrar por películas publicadas hasta el año 2000, cuyo director sea *Lewis Gilbert* o *Martin Campbell.* Excluir aquellas películas protagonizadas por Roger Moore.

- *Tabla*: James Bond
- *Descripción de tabla*: esta tabla incluye datos sobre personajes de la saga Star Wars

Resultado esperado

Row	Film	Director	Year	Actor
1	You Only Live Twice	Lewis Gilbert	1967	Sean Connery
2	GoldenEye	Martin Campbell	1995	Pierce Brosnan

Práctica

Ex.4: Descripción

Extraer países que cumplan con alguna de las siguientes condiciones: (i) sean países africanos con un índice de alfabetismo entre el 25% y 75% o (ii) países europeos con un ratio de población viviendo en áreas urbanas menor al 50%.

- *Tabla*: world_health_org
- *Descripción de tabla*: contiene información de países provista por la Organización Mundial de la Salud

Resultado esperado

Row	Country	Continent	Adult_literacy_rate	Population_in_urban_areas
1	Uganda	Africa	68.1	13
2	Bosnia and Herzegovi	Europe	96.7	46
3	Turkmenistan	Europe	98.8	47
4	Sierra Leone	Africa	34.8	41
5	Niger	Africa	28.7	17
6	Sudan	Africa	60.9	42
7	Central African Republic	Africa	48.6	38
8	Liberia	Africa	60.0	59
9	Angola	Africa	67.4	54
10	Chad	Africa	25.7	26

Práctica

Ex.5: Descripción

¿Cuáles son los 5 países africanos con mayor PIB (GIPC) per cápita?

- *Tabla*: world_health_org
- *Descripción de tabla*: contiene información de países provista por la Organización Mundial de la Salud

Resultado esperado

Row	Country	Continent	Gross_income_per_capita
1	Equatorial Guinea	Africa	16620
2	Seychelles	Africa	14360
3	Botswa	Africa	11730
4	Gabon	Africa	11180
5	Mauritius	Africa	10640

Práctica

Ex.6: Descripción

Extraer las 10 películas con mayor *IMDB score*, filtrar por películas publicadas a partir de la década del 80, excluir aquellas producidas en los EEUU.

- *Tabla*: imdb_movies
- *Descripción de la tabla*: esta tabla incluye data de películas publicada por IMDB (imdb.com)

Resultado esperado

Row	movie_title	director_me	imdb_score
1	The Lord of the Rings: The Fellowship of the Ring	Peter Jackson	8.8
2	Queen of the Mountains	Sadyk Sher-Niyaz	8.7
3	City of God	Ferndo Meirelles	8.7
4	Spirited Away	Hayao Miyazaki	8.6
5	The Pianist	Roman Polanski	8.5
6	Children of Heaven	Majid Majidi	8.5
7	The Lives of Others	Florian Henckel von Donnersmarck	8.5
8	Airlift	Raja Menon	8.5
9	Amélie	Jean-Pierre Jeunet	8.4
10	Baahubali: The Beginning	S.S. Rajamouli	8.4

Práctica (individual)

Ex.7.1: Descripción

Cambia los nombres de las columnas (campos) que están en inglés por su traducción en español.

- *Tabla*: loan-data
- *Descripción de la tabla*: esta tabla incluye data para la entrenamiento de un algoritmo de clasificación para la concesión de créditos

Práctica (individual)

Ex.7.2: Descripción

¿Qué solicitudes de crédito tienen un plazo de devolución entre 12 y 24 meses?

- *Tabla*: loan-data
- *Descripción de la tabla*: esta tabla incluye data para la entrenamiento de un algoritmo de clasificación para la concesión de créditos

Práctica (individual)

Ex.7.3: Descripción

¿Qué solicitudes de crédito corresponden a hombres solteros?

- *Tabla*: loan-data
- *Descripción de la tabla*: esta tabla incluye data para la entrenamiento de un algoritmo de clasificación para la concesión de créditos

Práctica (individual)

Ex.7.4: Descripción

¿Qué solicitudes de crédito corresponden a personas que en algún momento han solicitado otros créditos y los han pagado?

- *Tabla*: loan-data
- *Descripción de la tabla*: esta tabla incluye data para la entrenamiento de un algoritmo de clasificación para la concesión de créditos

Práctica (individual)

Ex.7.5: Descripción

¿Qué solicitudes corresponden a personas que tienen 4 créditos o más en curso?

- *Tabla*: loan-data
- *Descripción de la tabla*: esta tabla incluye data para la entrenamiento de un algoritmo de clasificación para la concesión de créditos

Práctica (individual)

Ex.7.6: Descripción

¿Qué solicitudes de crédito corresponden a un crédito de negocio?

- *Tabla*: loan-data
- *Descripción de la tabla*: esta tabla incluye data para la entrenamiento de un algoritmo de clasificación para la concesión de créditos

Práctica (individual)

Ex.7.7: Descripción

¿Qué solicitudes de crédito corresponden a un crédito de reparaciones?

- *Tabla*: loan-data
- *Descripción de la tabla*: esta tabla incluye data para la entrenamiento de un algoritmo de clasificación para la concesión de créditos

Práctica (individual)

Ex.7.8: Descripción

¿Qué solicitudes de crédito corresponden a personas que viven en su vivienda de propiedad?

- *Tabla*: loan-data
- *Descripción de la tabla*: esta tabla incluye data para la entrenamiento de un algoritmo de clasificación para la concesión de créditos

Práctica (individual)

Ex.7.9: Descripción

¿Qué solicitudes de crédito corresponden a personas con más de 60 años de edad?

- *Tabla*: loan-data
- *Descripción de la tabla*: esta tabla incluye data para la entrenamiento de un algoritmo de clasificación para la concesión de créditos

Práctica (individual)

Ex.7.10: Descripción

¿Qué solicitudes de crédito corresponden a personas entre 35 y 50 años de edad?

- *Tabla*: loan-data
- *Descripción de la tabla*: esta tabla incluye data para la entrenamiento de un algoritmo de clasificación para la concesión de créditos

Práctica (individual)

Ex.7.11: Descripción

¿Qué solicitudes de crédito se han aprobado?

- *Tabla*: loan-data
- *Descripción de la tabla*: esta tabla incluye data para la entrenamiento de un algoritmo de clasificación para la concesión de créditos

Práctica (individual)

Ex.7.12: Descripción

¿Qué solicitudes de crédito se han rechazado?

- *Tabla*: loan-data
- *Descripción de la tabla*: esta tabla incluye data para la entrenamiento de un algoritmo de clasificación para la concesión de créditos

"El FSE invierte en tu futuro"

Fondo Social Europeo