MÁSTER EN BIG DATA Y DATA SCIENCE

02MBID Sistemas de almacenamiento y gestión Big Data

CURSO 2021-2022

ACTIVIDAD 1: Creación del esquema de una base de datos orientada a columnas

Hecho por el estudiante Carlos de la Morena Coco

1. Obtener los jefes provinciales que están asociados a una provincia buscando a través
del nombre de la provincia

Jefes provinciales por provincia	
COLUMNA	**CLAVE**
Provincia_Nombre	PK
Provincia_Jefe_provincial	CK

El nombre de la provincia es único para cada provincia, y ya que vamos a realizar las consultas por este parámetro pues es obvio que la deberíamos usar como Partition Key.
Los jefes provinciales son Clustering Key porque es lo que va a identificar de forma única esas filas.
Ya que un tributo de tipo conjunto no puede ser Clustering Key, usaremos de CK cada uno de los jefes provinciales que estén en los sets.

2. Obtener según la capacidad de cada subestación la longitud de las líneas a las que está
asociado.

Longitud de líneas según capacidad de subestación	
COLUMNA	**CLAVE**
Subestación_Capacidad	PK
Linea_CodLin	CK
Línea_Longitud	

Para buscar por la capacidad de subestación, esta debe estar de PK.
La clave primaria de cada línea nos servirá para tener un identificador único de cada línea.
La longitud simplemente la queremos mostrar, así que no estará en la clave primaria.

3. Consultar la información de la provincia a la que esté asignado un solo jefe provincial.

Información de provincia por jefe provincial	
COLUMNA	**CLAVE**
Provincia_ProCod	PK
Provincia_Jefe_provincial	CK
Provincia_Nombre	

Podemos usar los jefes provinciales de clustering key (por separado), de tal forma que agrupan en el mismo nodo una misma provincia.
Luego al hacer la consulta podemos comparar el número de entradas que hay en cada nodo, y mostrar aquellos nodos que solo tengan una entrada.

4. Consultar las estaciones y productores en las que estos últimos hayan provisto de una cantidad concreta en una fecha determinada.

Estaciones y productores por fecha y cantidad	
COLUMNA	**CLAVE**
Provee_Fecha	PK
Provee_Cantidad	PK
Estación_CodEst	CK
Productor_CodPro	CK
Estación_Nombre	
Productor_Nombre	

En esta ocasión usamos dos atributos como Primary Key, y es la fecha y la cantidad de la relación de Provee entre productores y estaciones.
Como es una relación varios a varios necesitamos un identificador único de cada uno, por eso añadimos sus claves primarias como Clustering Key.

5. Consultar según el CodDis de una distribución de red las subestaciones que esta suministra, incluyendo la longitud de la línea que suple a la subestación

Subestaciones por Distribución de red	
COLUMNA	**CLAVE**
Distribución_de_red_CodDis	PK

Línea_Codlin	CK
Subestación_CodSub	CK
Línea_Longitud	

Ya que vamos a realizar las consultas por el CodDis de la distribución de red, este debería ser partition key.
Dada la estructura que tenemos aquí (1:n:1:m), necesitamos identificadores únicos de subestaciones y línas. Estos atributos cumplirán los roles de Clustering Key

6. Obtener la capacidad sumada de todas las subestaciones que se encuentran en una zona determinada.

Capacidad sumada de subestaciones por zona determinada	
COLUMNA	**CLAVE**
Zona_ZonCod	PK
Subestación_CodSub	CK
SumCapacidad	+

Ya que vamos a buscar por zona determinada, la clave primaria de cada una de estas entidades servirá como Primary Key.
Necesitamos un identificador únicode las subestaciones, y es por eso que CodSub se usa aquí de Clustering Key. Por último, creamos un campo en el que añadamos todas las sumas. En este caso, lo llamaremos SumCantidad

7. Obtener los productores que estén asociados a través de la estación a una distribución de red buscando por la longitud máxima de esta distribución.

Productores por longitud máxima de distribución de red	
COLUMNA	**CLAVE**
Distribución_de_red_Longitud_máxima	PK
Distribución_de_red_CodDis	CK
Estación_CodEst	CK
Productor_CodPro	CK
Productor_Nombre	

La longitud máxima de cada distribución de red es una cifra que se va a repetir varias veces, y además es el parámetro por el que realizamos la búsqueda, así que sin duda alguna este va a ser nuestra Primary Key.

Por el tipo de relaciones que hay, necesitamos un identificador único de distribución de red (CodDis), otro de estación (CodEst) y otro de Productor (CodPro). Estos identificadores únicos cumplirán la función de Clustering Key.
También añaderemos el campo nombre, simplemente para que quede bonito al hacer la consulta.

8. Buscar productores según el origen de la energía que producen (eólica, nuclear, carbón, solar o gas).

Productor según origen de la energía	
COLUMNA	**CLAVE**
Productor_Origen_energía	PK
Productor_País	PK
Productor_CodPro	CK
Productor_Nombre	

En este caso, buscamos por el origen de la energía, que además es un valor que se repetirá bastante, de modo que este campo es una Partition Key sin discusión.
Debido al desequilibrio de resultados que hay y para no tener bajadas de rendimiento por tener los nodos desequilibrados, he decidido añadir otra Partition Key, que en este caso puede ser el país, ya que se repite para todos los casos, y no se indica que esté tan desproporcionado.
El identificador único de cada productor nos sirve de Clustering Key, y el nombre del productor lo añadimos, una vez más, para que haga bonito.