Programació per a la ciència de dades – PAC4

***Nom***: Albert Salvador Yuste

# Presentació

A partir de la solució oficial de la segona pràctica (PRA2), l’estudiant ha d’implementar els cubs multidimensionals necessaris per a l’explotació de la informació i l’anàlisi de dades posterior relativa a l’activitat volcànica mundial. D’aquesta manera, es facilitarà la presa de decisions als usuaris potencials.

L’objectiu d’aquesta activitat és implementar un model *online analytical processing* (OLAP) per a l’anàlisi multidimensional de la informació, amb la finalitat de respondre a les preguntes definides a l’anàlisi de requeriments.

Addicionalment, juntament amb aquest enunciat, es facilitarà el fitxer “**dw\_PRA2\_export.sql**”, que conté els scripts de càrrega de totes les taules plantejades a la solució oficial de la PRA2, perquè l’estudiantat pugui partir de la mateixa base.

# Descripció

Més concretament, aquesta tercera part del cas pràctic consisteix a dissenyar un model OLAP per a l’anàlisi multidimensional de la informació volcànica disponible en el magatzem de dades que permeti donar resposta a les qüestions següents:

1. Activitat volcànica per regió del món dels volcans submarins, ordenada per nombre d’erupcions (activitat volcànica).
2. Anàlisi evolutiva del nombre d’erupcions mundials iniciades des de 1991 fins a 2021.
3. Anàlisi del TOP5 de països amb volcans amb el nombre més elevat de dies en erupció el 2020.
4. Mitjana de dies per erupció dels volcans situats a les regions del Japó, Taiwan, illes Mariannes, Amèrica del Sud, oceà Atlàntic, Índies Occidentals i Indonèsia a la dècada dels noranta.
5. Mostrar el resultat de la consulta anterior arrodonit a dos decimals i ordenat per valor de la mitjana descendent.
6. Activitat volcànica per país entre els anys 2000 i 2010 (inici de l’erupció), ordenat per codi de país [Country Code 3] DESCENDENT.

Per a un correcte desenvolupament i disseny del model OLAP, el document ha d’incloure com a mínim la descripció detallada dels passos següents:

Disseny del model OLAP

1. Creació del projecte en *Visual Studio* i revisió de l’origen de les dades.
2. Creació de la vista de l’origen de les dades.
3. Creació i implementació del cub.
4. Configuració de jerarquies, dimensions i atributs.
5. Processament i resolució d’errors.

Explotació de dades

1. Dissenyar sis consultes per donar resposta a les qüestions plantejades en aquest mateix enunciat.

A més de lliurar el document amb la solució de la PRA3, es considerarà la implementació a la màquina virtual proporcionada en el curs.

En resum, el document de la solució de la PRA3 ha d’incloure:

* La descripció de totes les accions realitzades.
* Les captures que demostrin la realització correcta dels apartats enumerats anteriorment.
* Les captures que mostrin la definició correcta de les consultes, la visualització del resultat de les explotacions i la seva correcta interpretació.

Totes les respostes a les preguntes formulades a l’explotació de dades s’han de realitzar a partir del cub implementat a la PRA3, independentment de l’eina utilitzada per a la seva anàlisi i explotació, bé s’utilitzi el *browser* de *Visual Studio*, *Power BI* o *Excel*.

Índex de continguts

[1. Presentació 1](#_Toc104738528)

[2. Descripció 1](#_Toc104738529)

[3. Disseny del model OLAP 4](#_Toc104738530)

[A. Creació del projecte en Visual Studio 4](#_Toc104738531)

[B. Revisió de l’origen de les dades 6](#_Toc104738532)

[C. Creació de la vista de l’origen de les dades 7](#_Toc104738533)

[D. Creació i implementació del cub 11](#_Toc104738534)

[E. Configuració de jerarquies, dimensions i atributs 15](#_Toc104738535)

[i. DIM\_Date 15](#_Toc104738536)

[ii. DIM\_Subregion 18](#_Toc104738537)

[iii. DIM\_Rank 19](#_Toc104738538)

[iv. DIM\_Country 19](#_Toc104738539)

[v. DIM\_Volcano 20](#_Toc104738540)

[F. Processament i resolució d’errors 20](#_Toc104738541)

[4. Explotació de dades 21](#_Toc104738542)

[A. Anàlisi evolutiva del nombre d’erupcions mundials iniciades des de 1991 fins a 2021. 22](#_Toc104738543)

[B. Anàlisi del TOP5 de països amb volcans amb el nombre més elevat de dies en erupció el 2020. 23](#_Toc104738544)

[C. Mitjana de dies per erupció dels volcans situats a les regions del Japó, Taiwan, illes Mariannes, Amèrica del Sud, oceà Atlàntic, Índies Occidentals i Indonèsia a la dècada dels noranta. 24](#_Toc104738545)

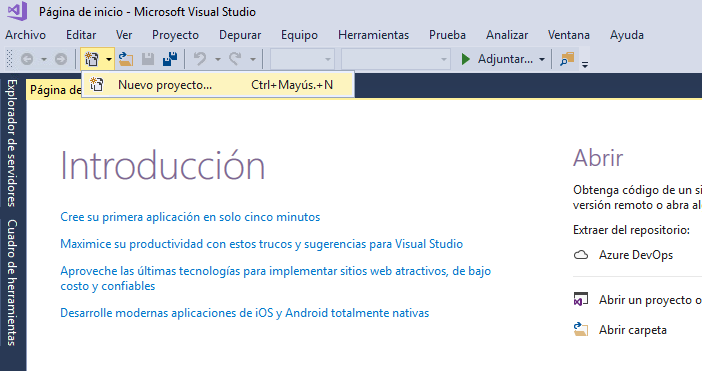
[D. Mostrar el resultat de la consulta anterior arrodonit a dos decimals i ordenat per valor de la mitjana descendent. 25](#_Toc104738546)

[E. Activitat volcànica per país entre els anys 2000 i 2010 (inici de l’erupció), ordenat per codi de país [Country Code 3] DESCENDENT. 25](#_Toc104738547)

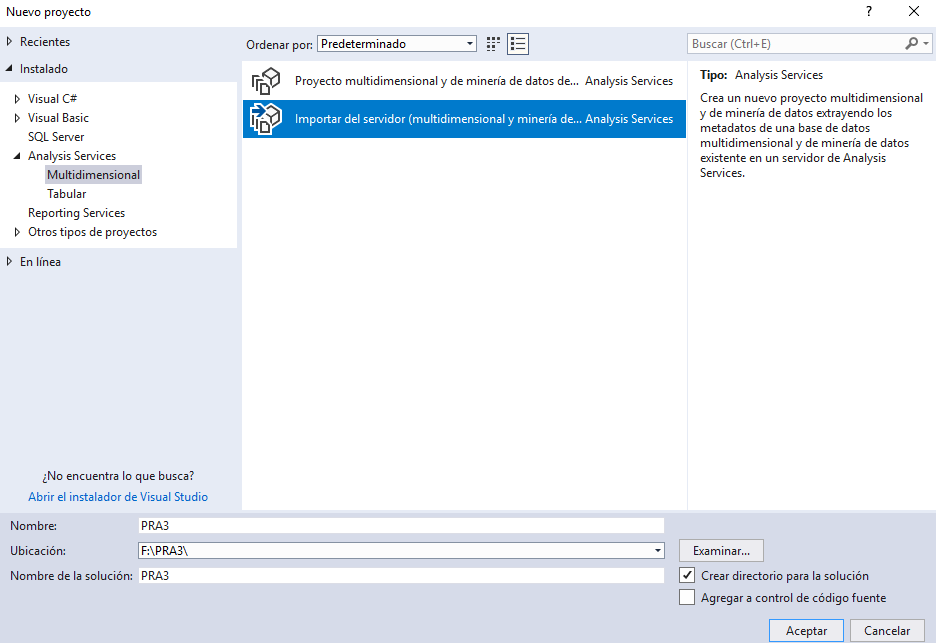
# Disseny del model OLAP

## Creació del projecte en *Visual Studio*

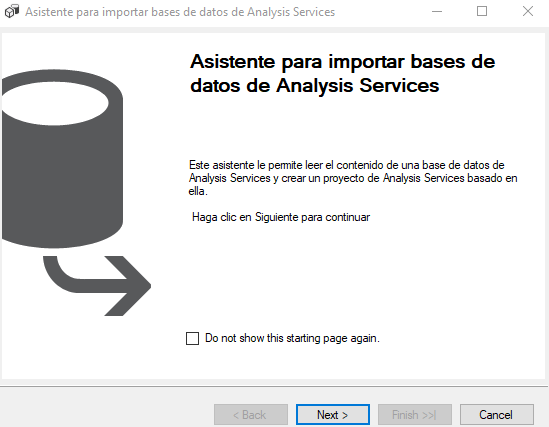
Iniciem *Visual Studio 2017* i seleccionem “Nuevo proyecto”:



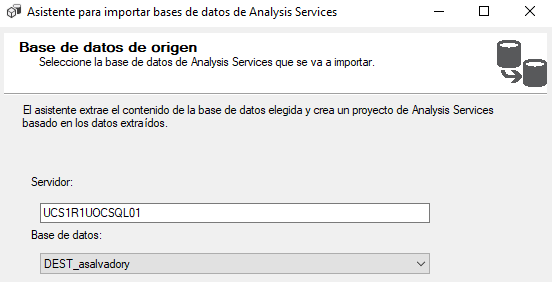
Tal i com es va realitzar anteriorment, es selecciona un *Analysis Services* → *Multidimensional*, el qual permetrà crear els cubs objectius d’aquesta pràctica. S’escull la opció “Importar del servidor (multidimensional y minería de datos)”, en modifiquem el nom (PRA3) i la seva ubicació (F:/PRA3)



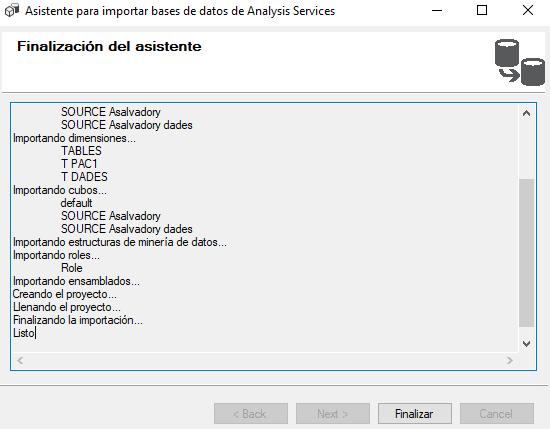
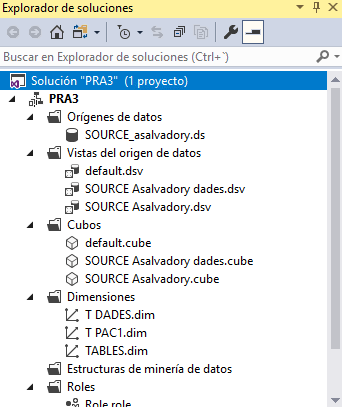
Després d’acceptar, procedim amb el procés d’importació:



En el següent pas indiquem sobre quin servidor volem importar la base de dades (UCS1R1UOCSQL01) i, un cop indicada, ens hauria d’aparèixer el nom de la base de dades adequada (DEST\_asalvadory)

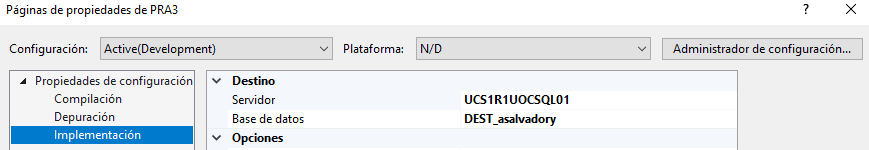


Un cop avançats en el procés ens apareix una finestra tal i com es pot observar seguidament, en la que se’ns confirma que l’origen de les dades s’ha importat de forma correcte. Alhora, un cop hem finalitzat, comprovem que podem tornar a observar totes aquelles vistes, cubs i dimensions que es van crear a la PAC1:

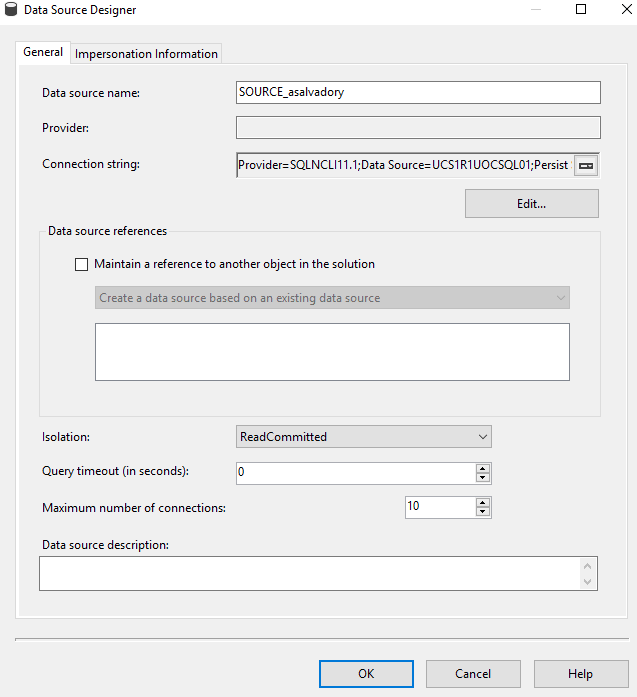
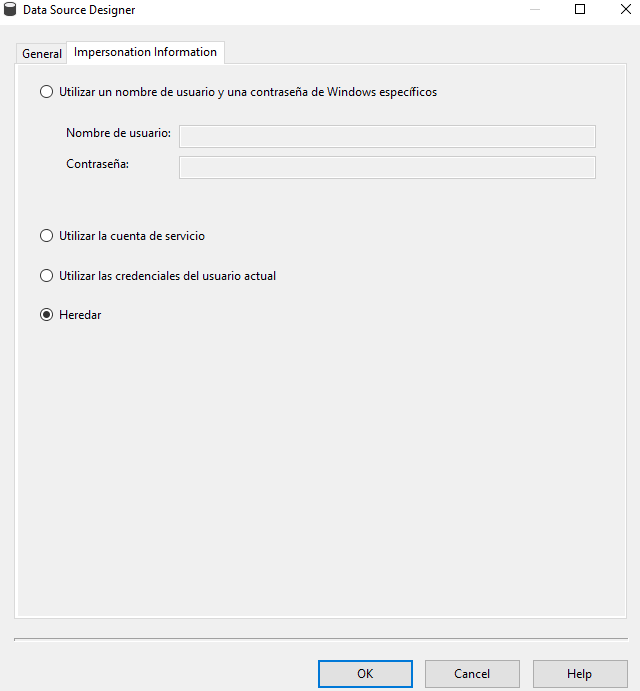
 

## Revisió de l’origen de les dades

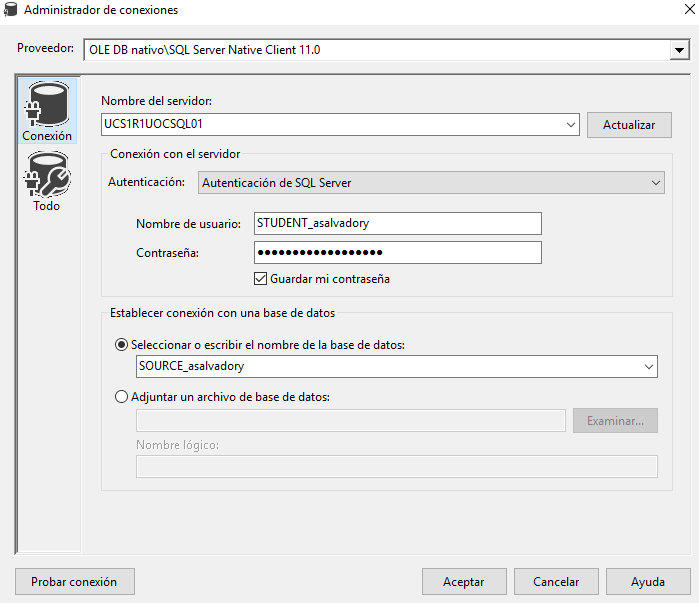
Primer de tot ens assegurem que la solució està ben configurada de la següent manera:

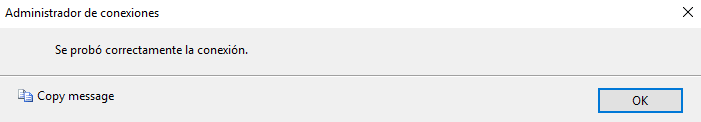


Per revisar l’origen de les dades seleccionarem amb un doble-click el “SOURCE\_asalvadory.ds”. En ell hi podem observar les següents pestanyes, una captura per cadascuna d’elles. De la segona sobretot ens fixem en què la opció “Heredar” està seleccionada:

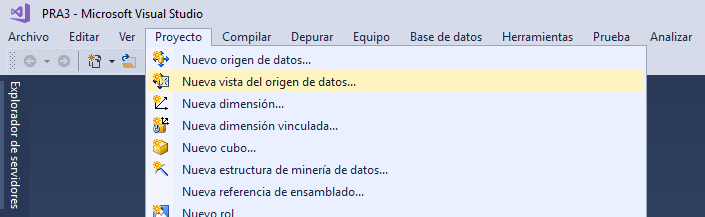
Sobre la primera de les pestanyes, seleccionem el botó “Edit...”. Ens n’apareixerà la següent finestra, de la qual ens interessa comprovar que estigui tot omplert de la manera adient, amb el nom del servidor, usuari i contrasenya, així com el nom de la base de dades adequat. Donant a “Probar conexión” hauríem de comprovar que la connexió en resulta satisfactòria:

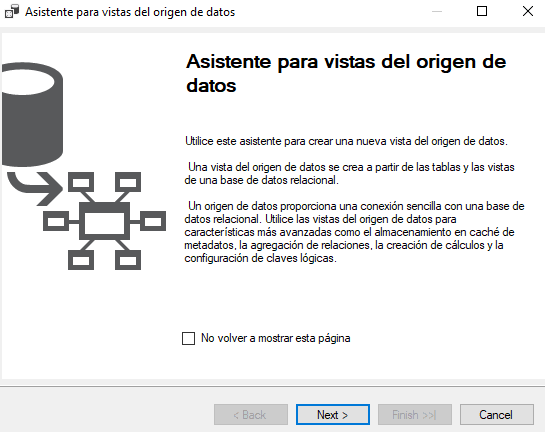




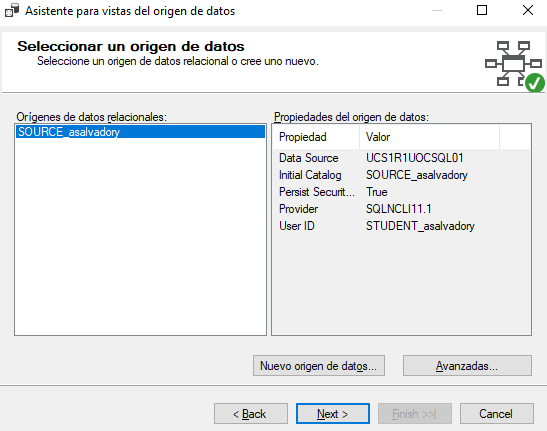
## Creació de la vista de l’origen de les dades

Per crear la vista de l’origen de les dades escollim la opció “Nueva vista del proyecto de datos...” i ens apareixerà un assistent:



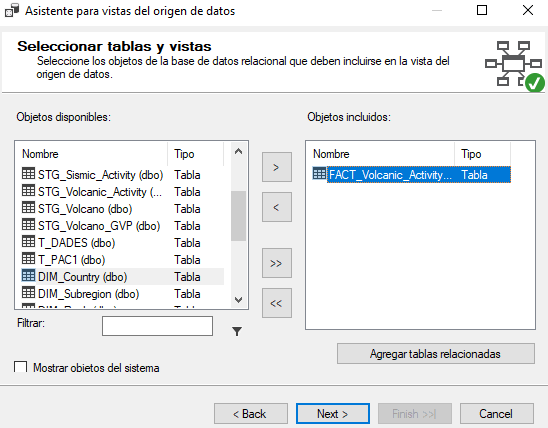


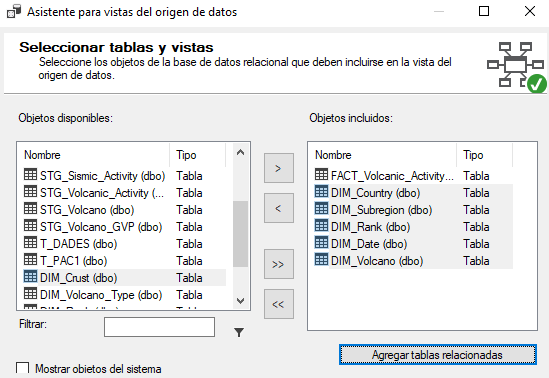
Seguint amb l’assistent, seleccionem l’origen de dades corresponent (SOURCE\_asalvadory):



A continuació crearem una vista per al cub en qüestió que es vol crear. La vista que definirem l’anomenen “**Vista\_Volcanic\_Activity**”, la qual ens donarà accés a les taules per dissenyar el cub d’anàlisi de l’activitat volcànica.

Per aquesta vista d’origen de dades seleccionem les taules necessàries per definir la vista. Per a tal fi, seleccionem la taula de fets i, seguidament, afegir totes aquelles taules relacionades amb la opció “Agregar tablas relacionadas”, de tal manera que s’utilitzen les claus foranes per dur les taules relacionades:

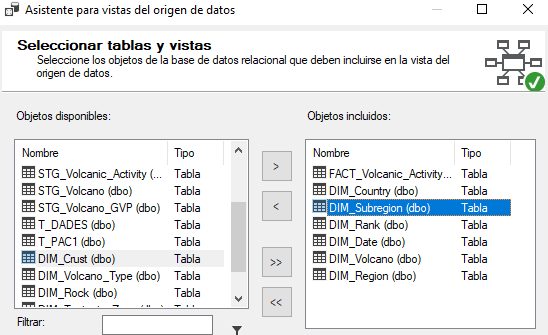
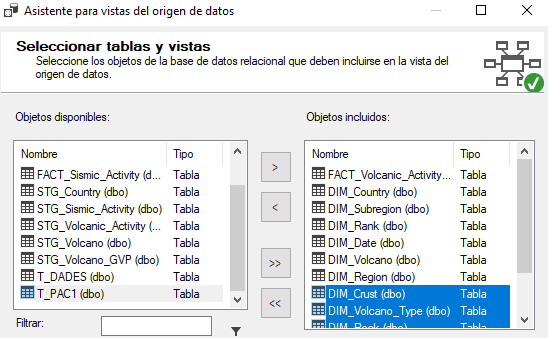




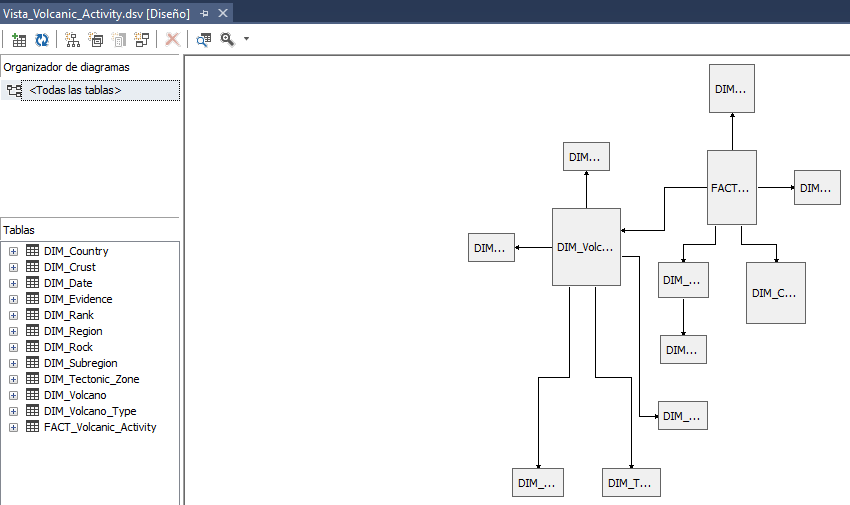
De fet, hem de tenir present en aquest punt que estem treballant amb una disseny en floc de neu, és a dir, tenim dimensions que depenen d’altres dimensions. Recordem el disseny conceptual de la solució oficial de la PRA1:



Observem com de “DIM\_VOLCANO” i de “DIM\_SUBREGION” en depenen més taules relacionades. Per afegir-les automàticament, les seleccionem amb l’assistent i tornem a clicar la opció de “Agregar tablas relacionadas” perquè finalment puguin estar totes les taules necessàries:

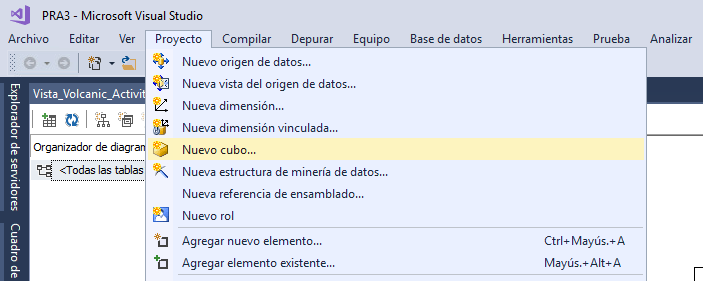
 

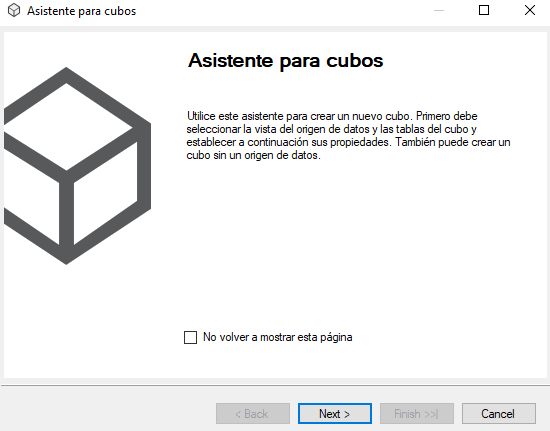
A l’últim pas finalment indiquem el nom que volem per aquesta vista (“**Vista\_Volcanic\_Activity**”) i finalitzem. Si fem doble clic a la vista generada n’obtindrem un disseny visual d’aquesta vista:



## Creació i implementació del cub

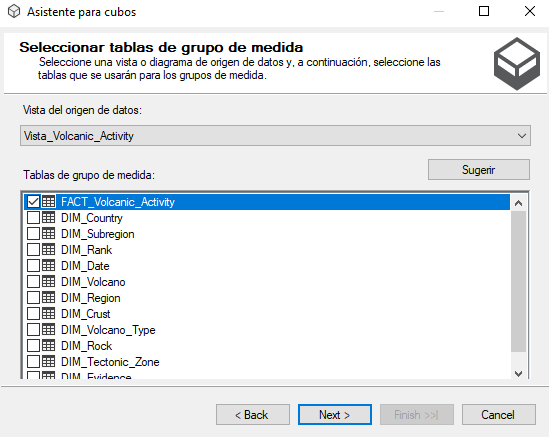
Un cop hem creat la vista per a l’origen de les dades, en aquest punt crearem el cub. L’objectiu d’aquest és realitzar els anàlisis OLAP per a l’explotació de les dades del *data warehouse*. Per crear-lo seleccionem la opció “Nuevo cubo...” i, novament ens n’apareixerà un assistent al respecte:



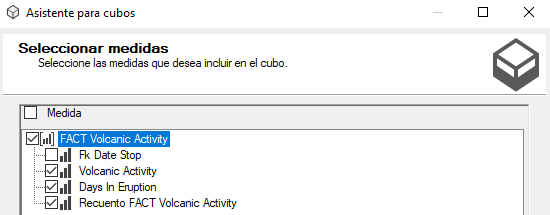


En el primer pas se’ns pregunta com desitgem crear el nostre cub. Dins de les opcions que se’ns mostren seleccionarem la opció “Usar tablas existentes” i seguim amb el assistent.

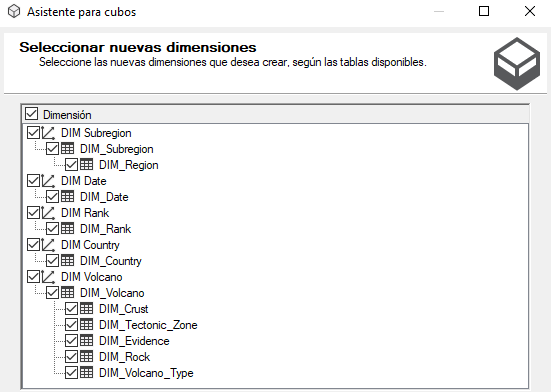
En el primer pas seleccionem la vista de l’origen de dades amb el que estem treballant (“**Vista\_Volcanic\_Activity**”) i en seleccionem tan sols la taula “FACT\_Volcanic\_Activity”:

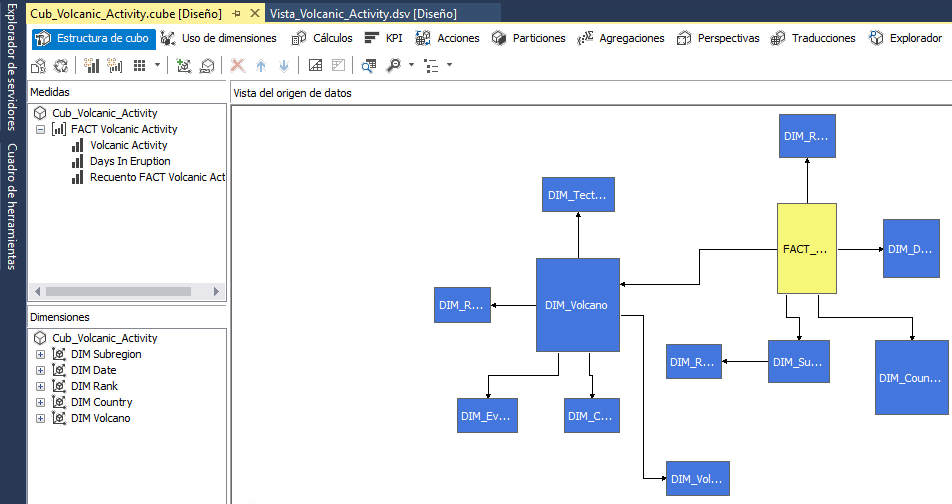


En el següent pas hem de seleccionar sobre quines mètriques volem treballar. En el cas de la taula “FACT\_Volcanic\_Activity” recuperem que les nostres mètriques d’interès són “**Volcanic Activity**” i “**Days In Eruption**”. Juntament amb la de “**Recuento FACT Volcanic Activity**” per quan necessitem fer recomptes, seran les que seleccionem:

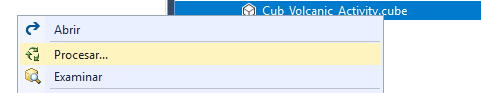


Seguim amb l’assistent i comprovem que automàticament ha seleccionat totes les taules dimensió que són necessàries, inclús les que en depenen d’altres degut a l’estructura de floc de neu. Ens assegurem que estan totes marcades i seguim amb l’assistent, el qual ens pregunta pel nom que li volem posar, el qual serà en aquest cas “**Cub\_Volcanic\_Activity**” i finalitzem.

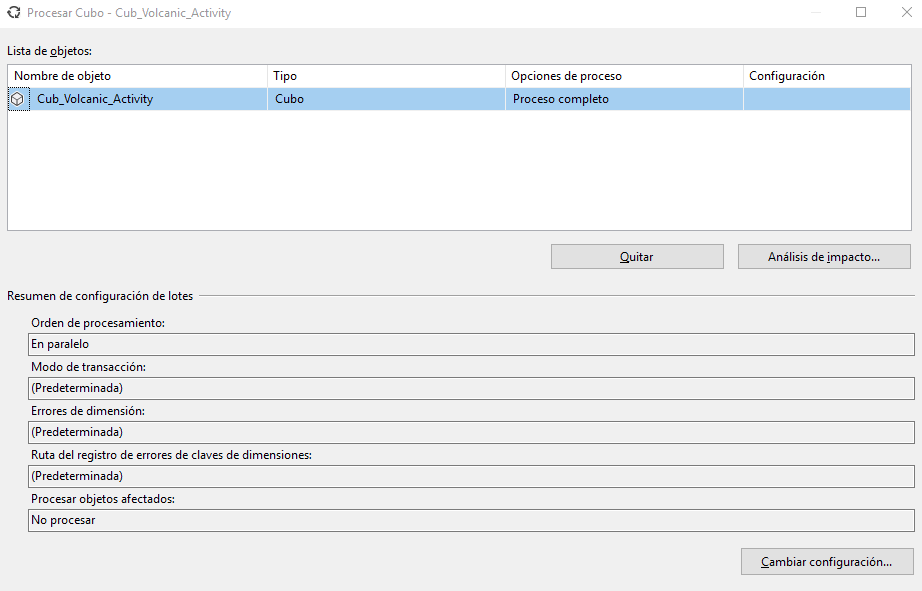


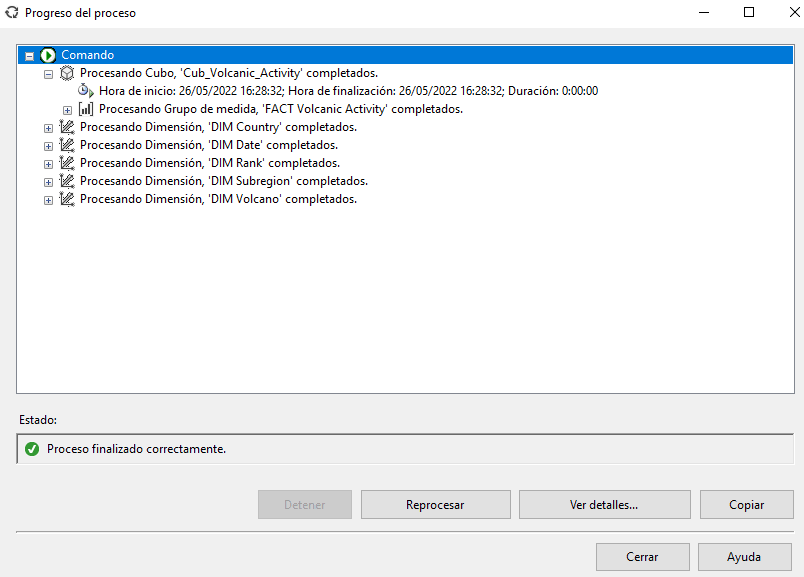


En el següent pas ens interessa poder visualitzar-ne les dades. Per poder-ho realitzar hem d’implementar el cub al servidor d’SSAS mitjançant l’opció “Procesar” al menú contextual del cub:

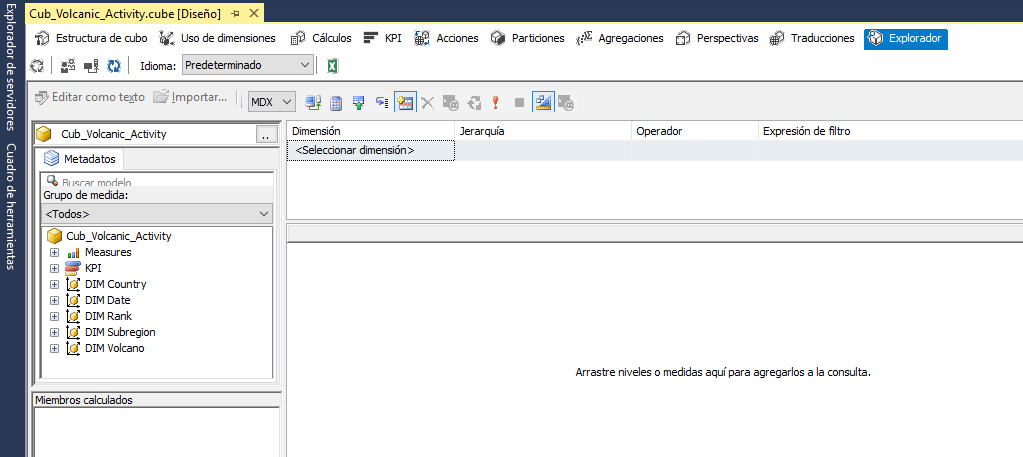


Després d’acceptar un seguit de passos mentre transcorre el procés d’implementació ens apareix per pantalla una finestra d’execució del cub com la següent. Si seleccionem “Ejecutar...” se’ns mostra novament per pantalla que el cub s’ha processat adequadament a la base de dades de destí:





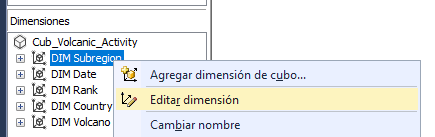
En aquest punt ja podrem navegar pel cub a través de la pestanya “Explorador”:



## Configuració de jerarquies, dimensions i atributs

Un cop hem definit totes les relacions del cub, es treballa amb les dimensions del projecte amb l’objectiu de definir els seus atributs i les seves jerarquies.

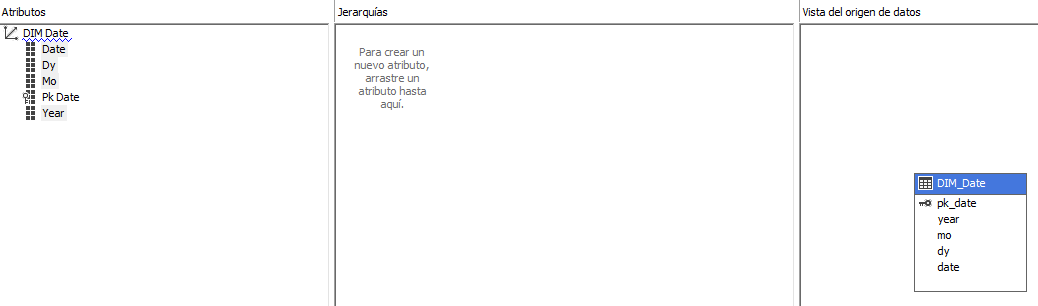
Per definir els atributs i jerarquies de cada dimensió triem la opció “Editar dimensión” des del menú contextual a la pestanya “Estructura” del cub:



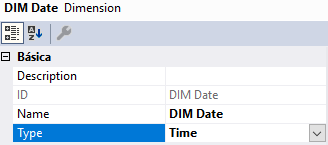
Només en modificarem aquelles dimensions amb els atributs necessaris per tal de realitzar posteriorment l’anàlisi de dades.

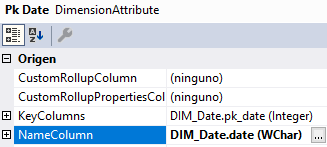
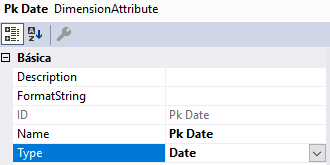
### DIM\_Date

Aquesta dimensió s’utilitzarà com a dimensió temporal. Inicialment només en tenim el valor de **Pk Date**, però ens interessa també obtenir-ne els valors de any, mes, dia i data. Per fer-ho afegim els atributs mencionats a través de la “Vista del origen de datos” de la dreta:

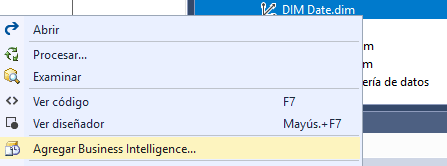


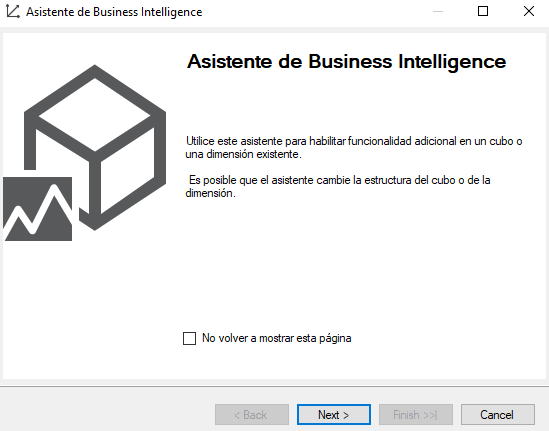
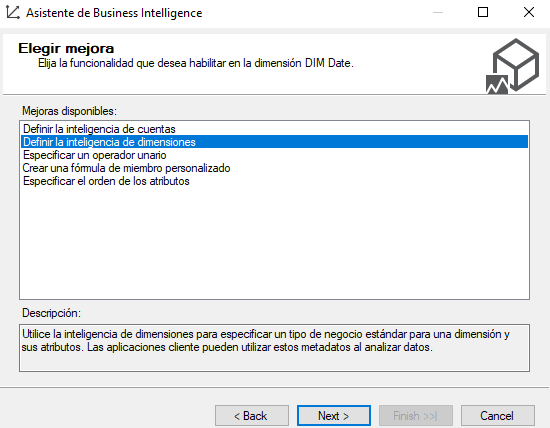
A més, a les propietats de la dimensió canviarem el tipus de la dimensió a *Time*. Alhora, indiquem que es consideri el **Pk Date** com a *Date* i la variable “NameColumn” que prengui els valors de l’atribut **date** de la DIM\_Date:



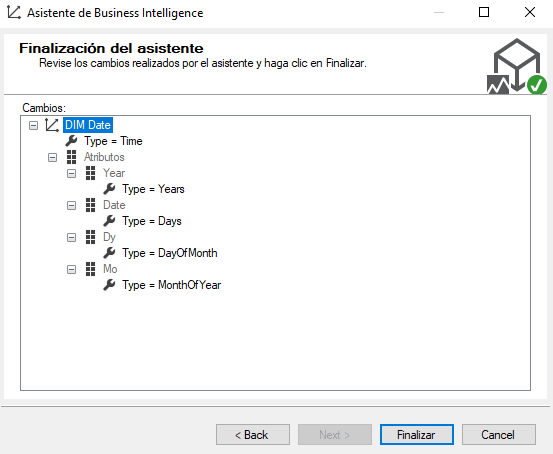


Un cop modificats aquests atributs, voldrem que es crei una jerarquia de dies, mesos i anys utilitzant la opció “Agregar Business Intelligence...”, de manera que ens apareix novament un nou assistent:

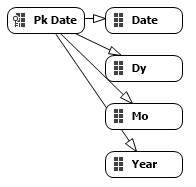
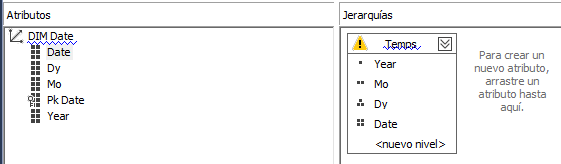


Dins de les opcions que ens apareixen seleccionem “Definir la inteligencia de dimensiones” per a què en el següent pas puguem atribuir a cada atribut el tipus que pertany per resoltar-ne el següent en l’últim pas:



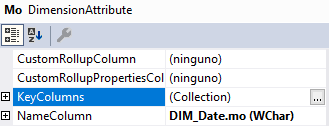
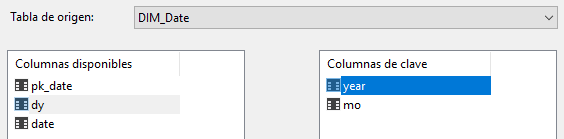
Un cop creada la en definirem la jerarquia temporal de la següent manera, amb les seva respectiva relació d’atribut:

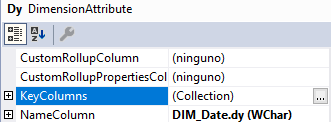
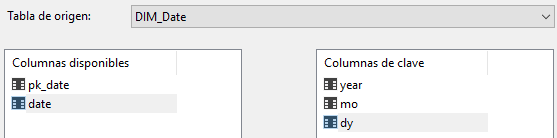


Ens interessa modificar-ne les relacions pertinents, així com definir relacions com a “Rígido” allà on pertoqui:

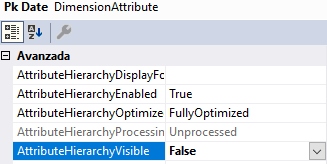


A l’hora d’anar a processar la dimensió DIM\_Date ens retorna errors amb les variables de mes (**Mo**) i dia (**Dy**). Per a corregir-ho modifiquem els “KeysColumns” i “NameColumn” adequats. Per a cada atribut, la captura de la dreta apareix al clicar els “...” de la propietat “KeysColumns”:

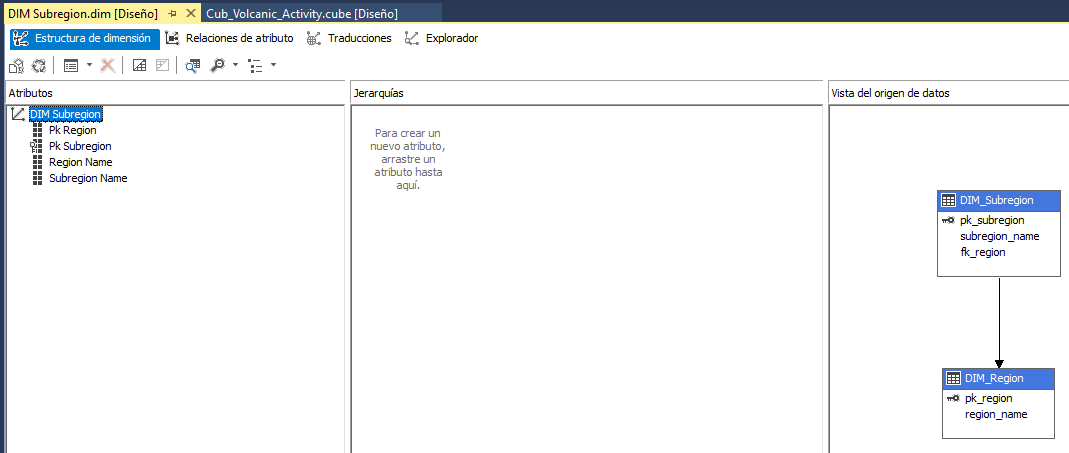
 

Finalment a l’hora de processar es retorna un avís de la necessitat d’ocultar els atributs de *primary key*. Per tal de realitzar-ho modifiquem la propietat “AttibuteHierarchyVisible” del **Pk Date** a *False*:

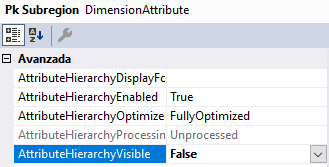


### DIM\_Subregion

Per a aquesta dimensió hem importat els noms de les subregions (**Subregion Name**) i els de les regions (**Region name**), obtinguts aquests últims de la taula relacionada DIM\_Region a través del **Pk Subregion**:

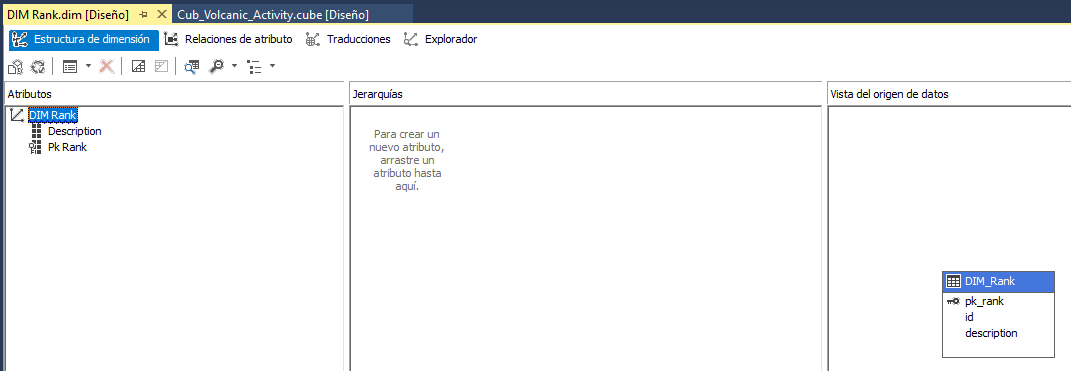


Alhora, hem modificat les propietats del Pk Subregion per tal de que no sigui visible modificant la propieta “*AttributeHierarchyVisible*”:

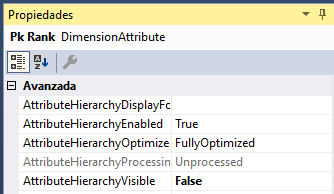


### DIM\_Rank

Per a aquesta dimensió hem importat les descripcions de cada rang (**Description**) obtinguts de la mateixa taula DIM\_Rank:

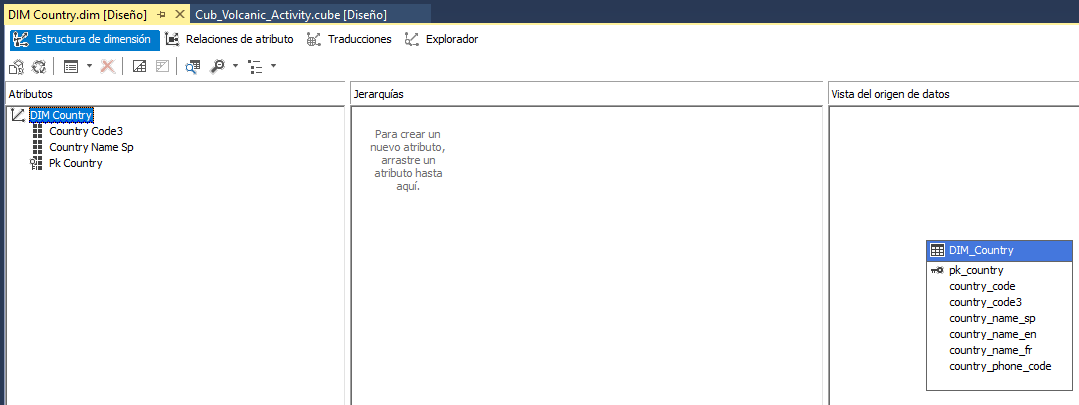


Igual que l’anterior cas, hem modificat les propietats del **Pk Rank** per tal de que no sigui visible modificant la propietat “*AttributeHierarchyVisible*”:

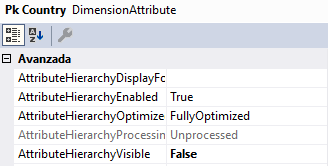


### DIM\_Country

Per a aquesta dimensió hem importat els noms de cada país en castellà (**Country Name Sp**) i el codi 3 de cada país (**Country Code3**) obtinguts de la mateixa taula DIM\_Country:

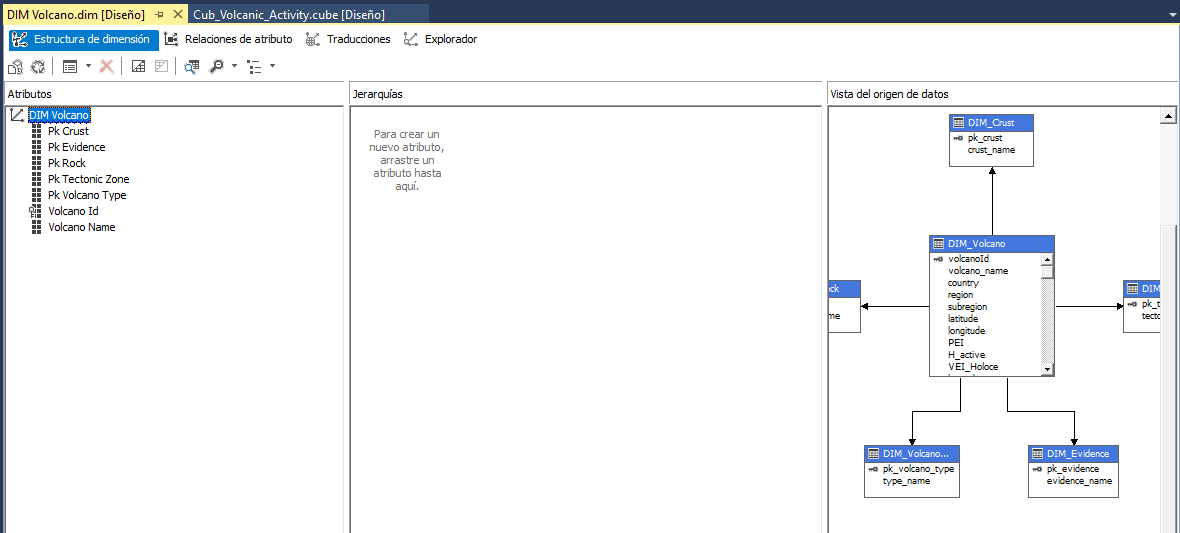


Novament per a aquesta dimensió, hem modificat les propietats del **Pk Country** per tal de que no sigui visible modificant la propietat “*AttributeHierarchyVisible*”:

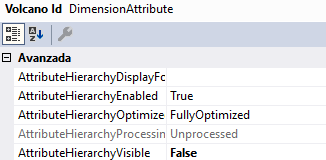


### DIM\_Volcano

Per a aquesta dimensió hem importat únicament els noms de cada volcà (**Volcano Name**) obtinguts de la mateixa taula DIM\_Country, doncs és la única variable que a priori ens pot resultar útil:

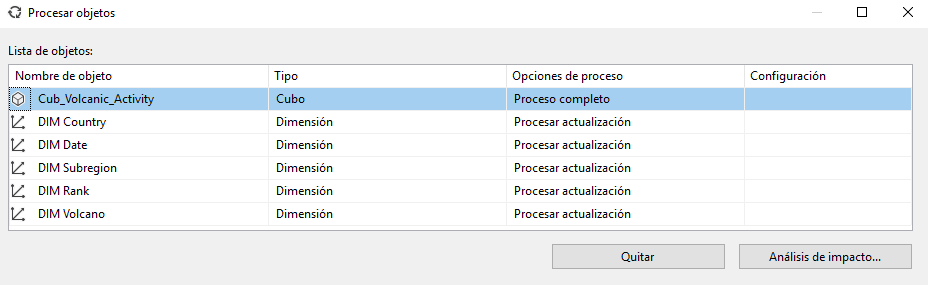


Novament per a aquesta última dimensió, hem modificat les propietats del **Volcano Id** per tal de que no sigui visible modificant la propietat “*AttributeHierarchyVisible*”:

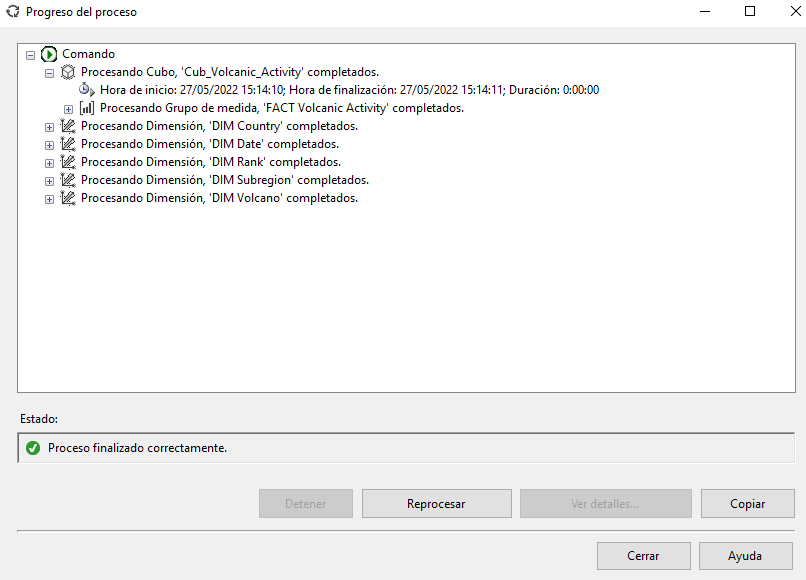


## Processament i resolució d’errors

Un cop realitzats tots els canvis necessaris a totes les dimensions, reprocessem tant el cub de nou com les dimensions modificades:

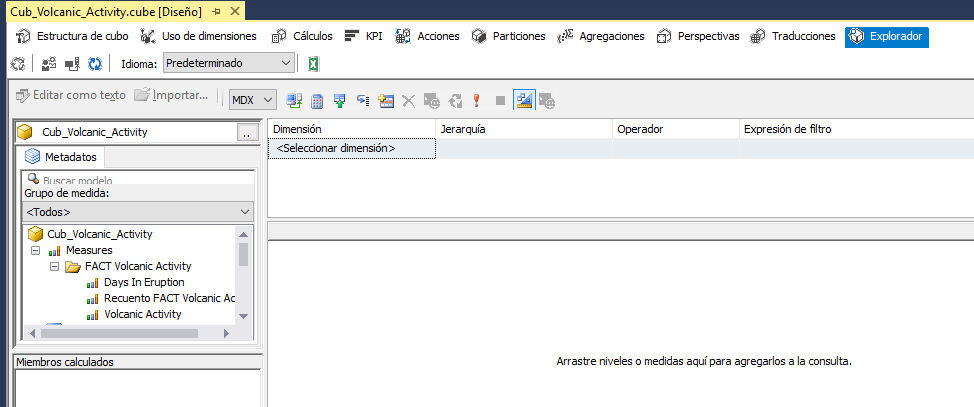


N’observem que no n’obtenim cap error a l’hora de processar-ho, per tant podem procedir al següent pas en què explotarem les dades sobre les que estem treballant:



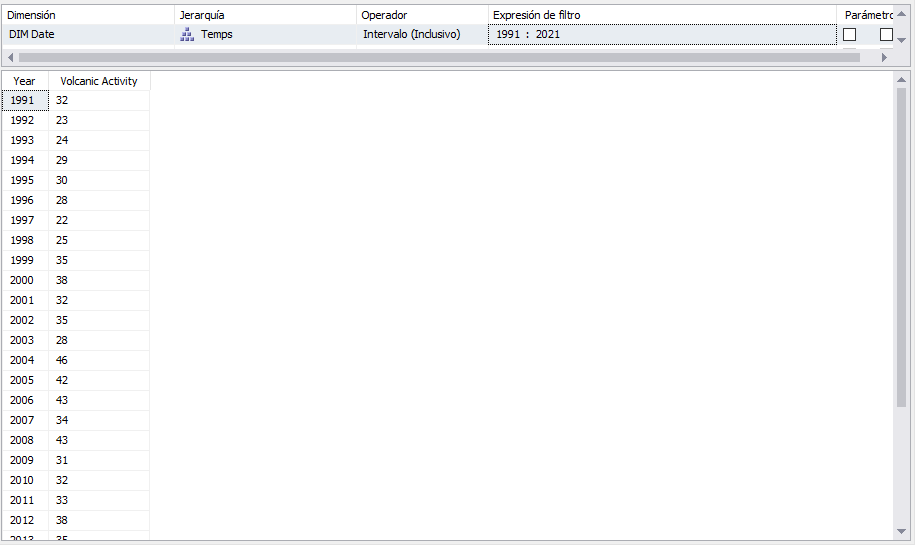
# Explotació de dades

Per realitzar les explotacions de dades per a cada cas utilitzarem la pestanya “Explorador” del cub que hem creat. En ella seleccionarem quines “Measures” volem calcular sobre quina/es dimensions en cada cas:



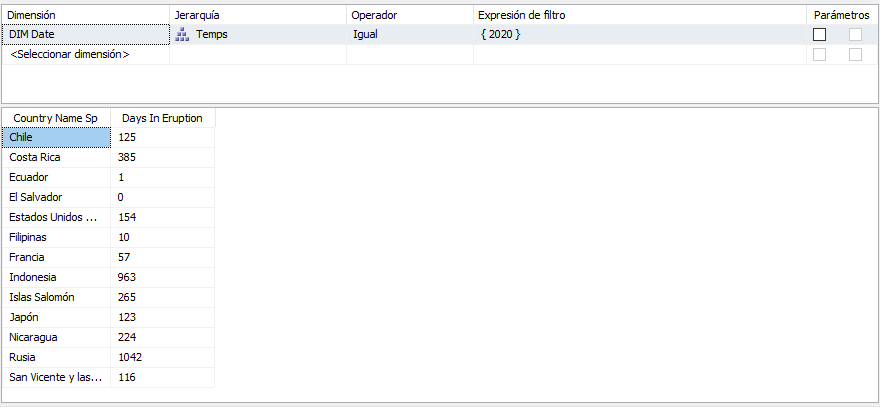
## Anàlisi evolutiva del nombre d’erupcions mundials iniciades des de 1991 fins a 2021.

Per realitzar la primera de les consultes, com hem dit hem utilitzat la pestanya “Explorador” del cub. En aquest cas volem conèixer el nombre d’erupcions per data, així doncs, la primera dimensió que utilitzem és la de DIM\_Date, indicant que volem que s’incloguin aquells anys entre el 1991 i 2021. Si llavors incorporem com a columnes de resultats els anys (**Year**) i l’activitat volcànica (**Volcanic Activity**), observem com es calcula per a cada any dins del periode indicat:

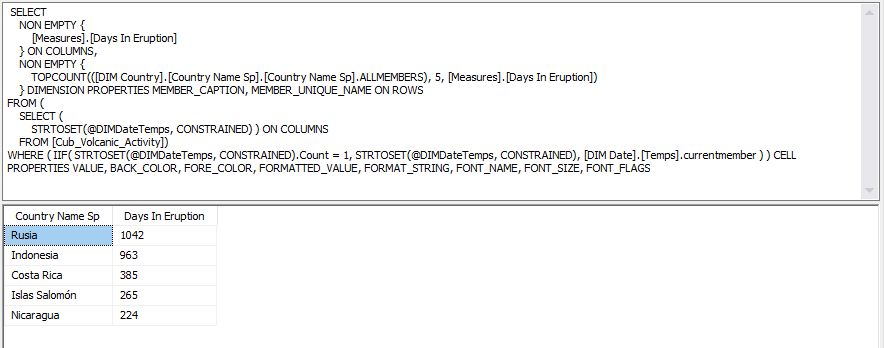


## Anàlisi del TOP5 de països amb volcans amb el nombre més elevat de dies en erupció el 2020.

Inicialment comencem com en el cas anterior, és a dir, indicant sobre quin anys volem realitzar la consulta, en aquest cas, any 2020. Com que volem saber els noms del països i els dies en erupció dels volcans que tenen, seran aquestes dues variable, **Country Name Sp** i **Days In Eruption** les que introduïm com a columnes:

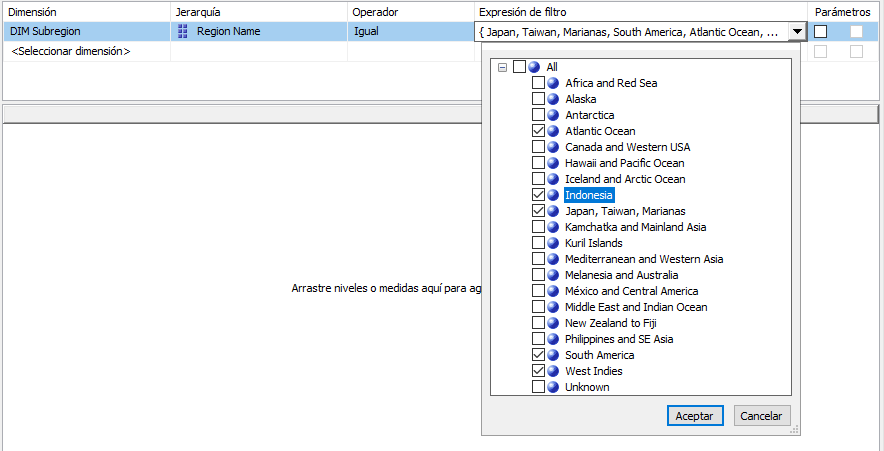


En aquest cas però no és suficient amb això. Per acabar de completar la consulta hem realitzat una modificació al MDX per fer servir una funció **TOPCOUNT**. Aquesta la podem observar introduida a la sisena línia del codi, on indiquem que en volem obtenir els primers 5 valors segons [Measures].[Days In Eruption]. S’observa doncs com finalment només obtenim els 5 primers països que tenen volcans amb més dies en erupció, ordenats de forma descendent per aquest últim:

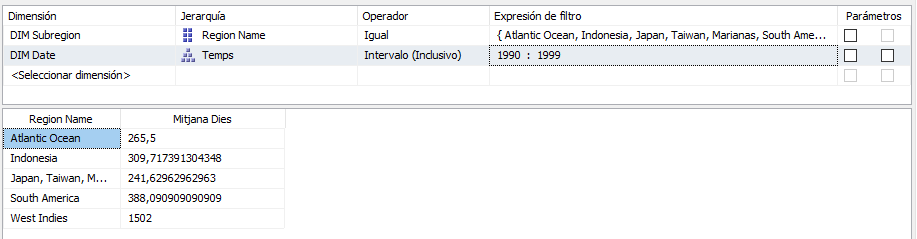


## Mitjana de dies per erupció dels volcans situats a les regions del Japó, Taiwan, illes Mariannes, Amèrica del Sud, oceà Atlàntic, Índies Occidentals i Indonèsia a la dècada dels noranta.

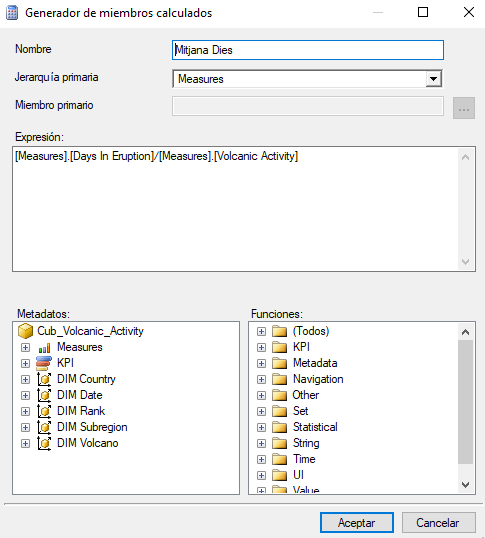
El primer cas consisteix en filtrar sobre quines regions volem fer la consulta. Aquestes es poden consultar a l’explorador del cub tal i com s’indica a la següent imatge:



Tot seguit indiquem sobre quines dates es vol treballar. En aquest cas s’indica que es volen selecciones aquelles erupcions compreses a la dècada dels ’90, és a dir, 1990-1999, ambdós anys inclosos. Per això en el resultat volem que se’ns mostri els noms de les regions (**Region Name**) i un comptador de mitjana de dies per erupció, tal i com s’observa a continuació:



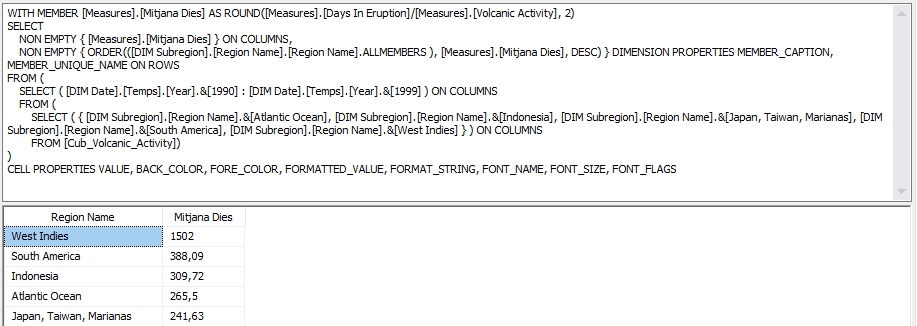
On “Mitjana Dies” l’hem creat amb un “Agregar miembro calculado” amb la següent expressió:



## Mostrar el resultat de la consulta anterior arrodonit a dos decimals i ordenat per valor de la mitjana descendent.

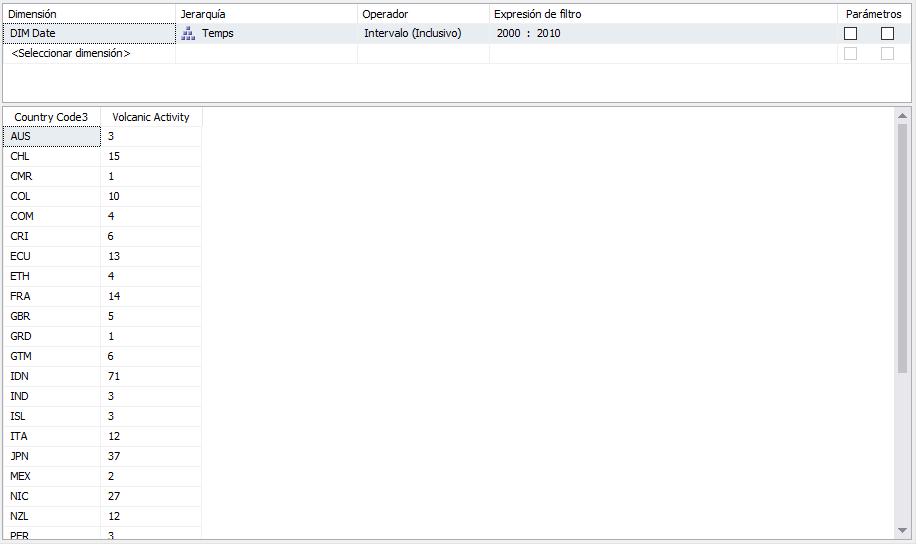
Per realitzar-ho s’ha modificat sobre la consulta anterior al MDX fent servir una funció **ROUND** (primera línia), on especifiquem fins a quants decimals volem obtenir el resultat del càlcul.

Alhora, a la quarta línia hem afegit una funció **ORDER**, tot indicant que volem els valors de les regions a les files, però ordenades per el valor descendent del valor mitjà de dies d’erupció:



## Activitat volcànica per país entre els anys 2000 i 2010 (inici de l’erupció), ordenat per codi de país [Country Code 3] DESCENDENT.

Inicialment fem una consulta normal amb l’explorador del cub, limitant les dates que volem incloure en la consulta (entre el 2000 i 2010), indicant que volem extreure els valors de **Country Code 3** i de **Volcanic Activity**:



Veiem que s’ordenen per ordre alfabètic ascendent del codi del país. Per modificar-ho entrem al MDX. En ell novament utilitzem una funció **ORDER**, indicant que volem que s’ordenin segons el mateix codi de país, especificat en aquest cas per la variable “*CurrentMember*”, tal i com es pot observar a la tercera línia del codi. Tot seguit es pot observar el resultat final d’aquesta consulta:

