12/17/2015

Jinsiwale, Nishant

U-20-99-4501

**Design & Implementation of a Data Warehouse**

**Final Project Report**

**Introduction –**

The reporting and sharing of information has been synonymous with databases as long as there have been systems to host them. Now more than ever, users expect the sharing of information in an immediate, efficient, and secure manner. However, due to the sheer number of databases within the enterprise, getting the data in an effective fashion requires a coordinated effort between the existing systems.

There is a very real need today to have a single location for the storage and sharing of data that users can easily utilize to make improved business decisions, rather than trying to traverse the multiple databases that exist today and can do so by using an enterprise data warehouse. The project on Amazon Reviews involves a description of data warehousing techniques, design, expectations, and challenges regarding data cleansing and transforming existing data, which was extracted from the transactional database and provided to us on HP Vertica Analytical Server

The project also includes a technical piece discussing database requirements and technologies used to create and refresh the dimensional attributes .

There are implementations on how to present the data in the enterprise data warehouse, such as through reporting and other business intelligence which here is Microsoft Excel’s Pivot Tables.

The Data Warehouse prototype developed will show how a pair of different databases undergoes the Extract, Transform and Load (ETL) process and loaded into an actual set of star schemas then makes the reporting easier.

**Purpose:**

The purpose of this project is to present central administrative data in an understandable format, allowing flexible presentation, and in a way that is easy for users to access

**Objectives:**

* Format the data such that it is easier for users to understand and to access.
* Develop An OLAP Cube For Powerful and easy to understand Analysis
* Deliver a set of reports based on the Dimensional Model and Measures
* Powerful Analytical Environment to modify, save, & run their own reports and the pre-written reports.

**Specific Goals:**

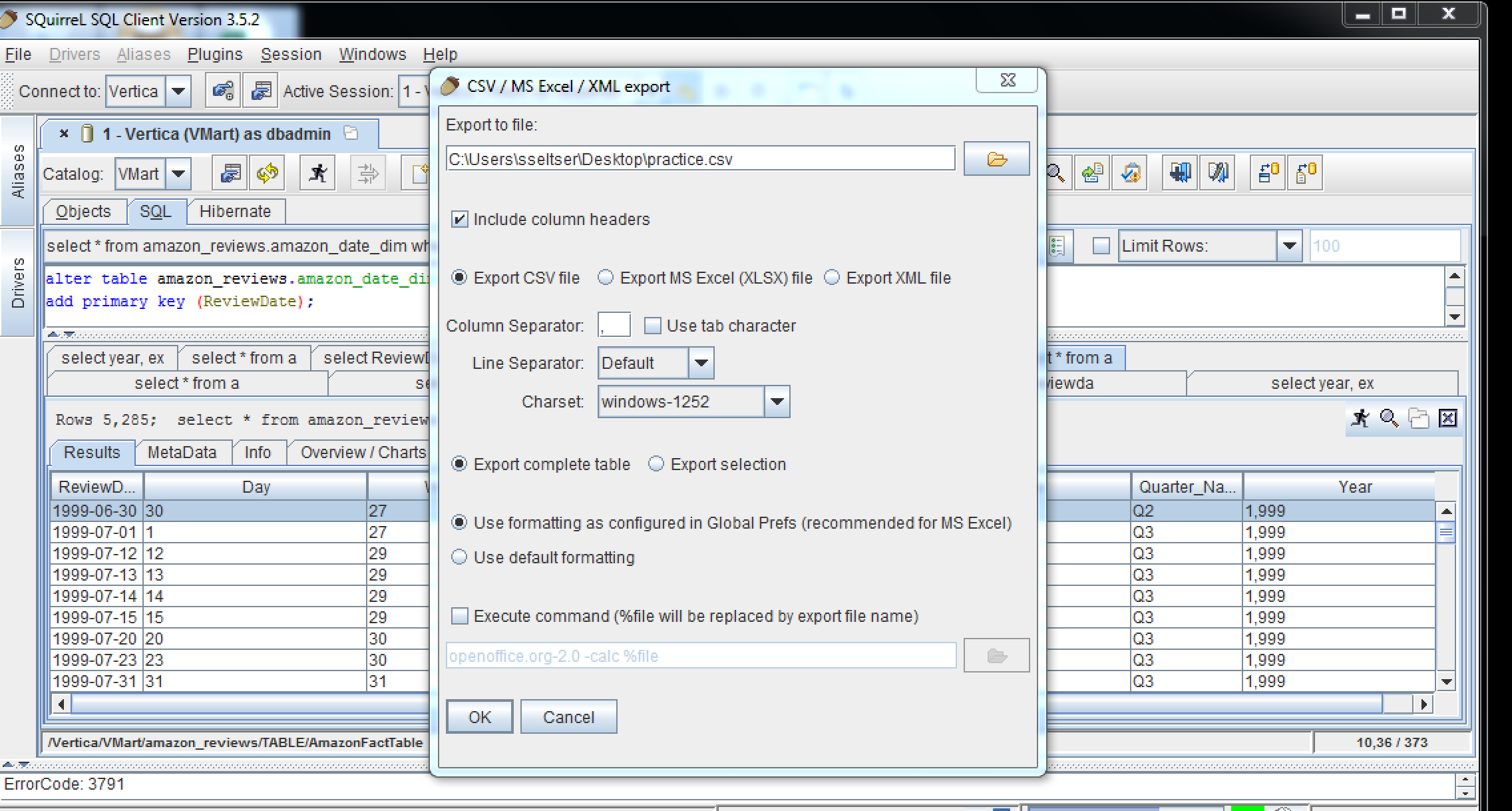
* Build a data warehouse for a large dataset on Amazon Product Reveiws from Year 1999 – 2014
* Reorganize the data tables into few tables (that we call dimensions)with an easier to understand format
* Rename tables and data fields to be more easily understood
* Pre-define table relationships so users do not need to understand or define them (Primary and Foreign Key Constraints)
* Develop an OLAP Cube based on Dimensions and Measures
* Develop a set of commonly used pre-written reports on Microsoft Excel Pivot Table
* Provide an easy method for users to develop their own reports and have easy access to both the common reports and their locally developed reports

**Status:**

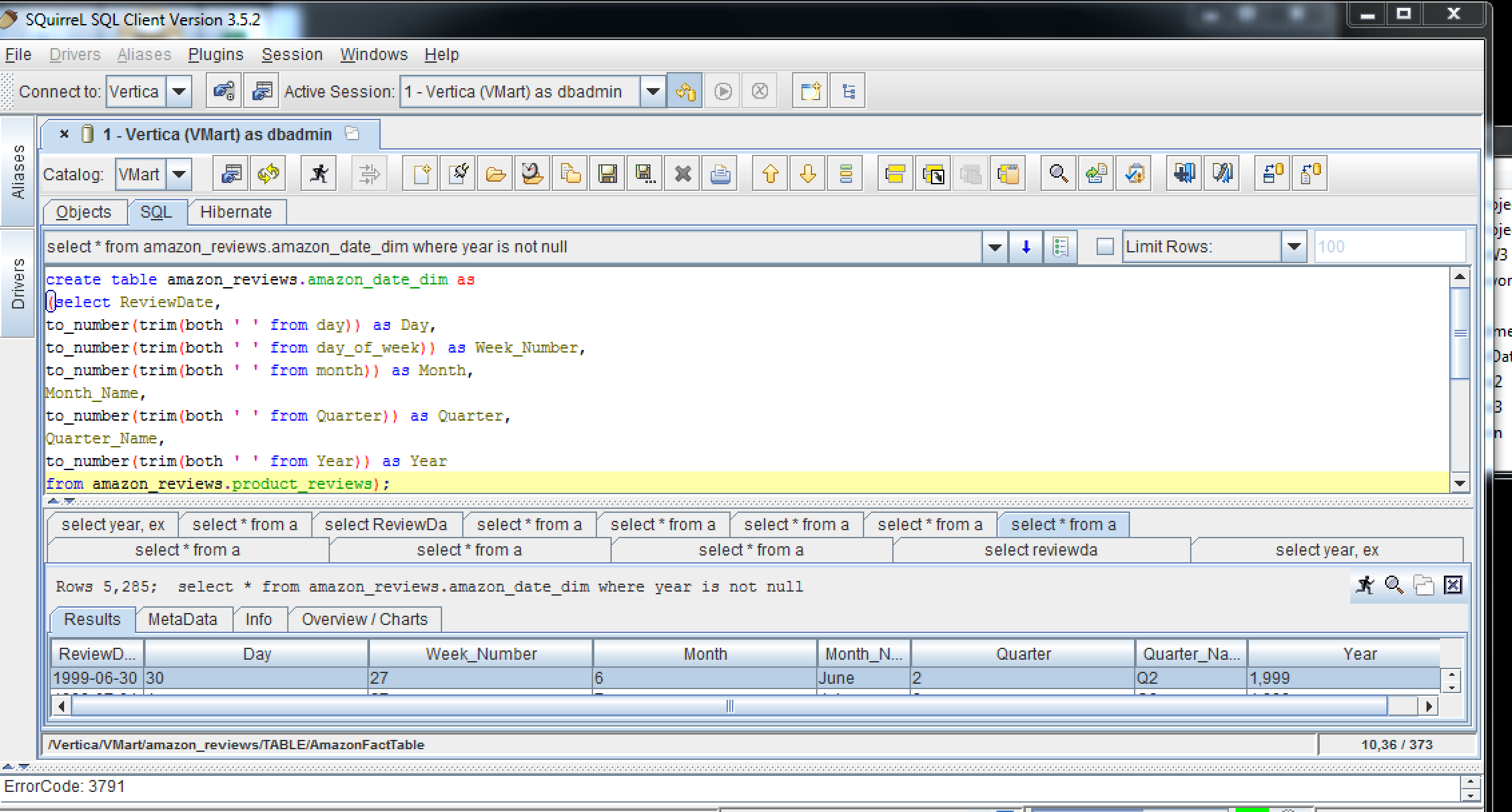
* Amazon Product Review Data Warehouse is in production with BI Model built. The model contains hundreds of pre-written Bi-Query reports for users to run.
* Human Resources Data Warehouse is in production with BI-Query Model built with user reports available.
* A number of Reports have been developed based on Product Categories and Date Dimensions

**Project Discription with Screenshots of Project Progress-**

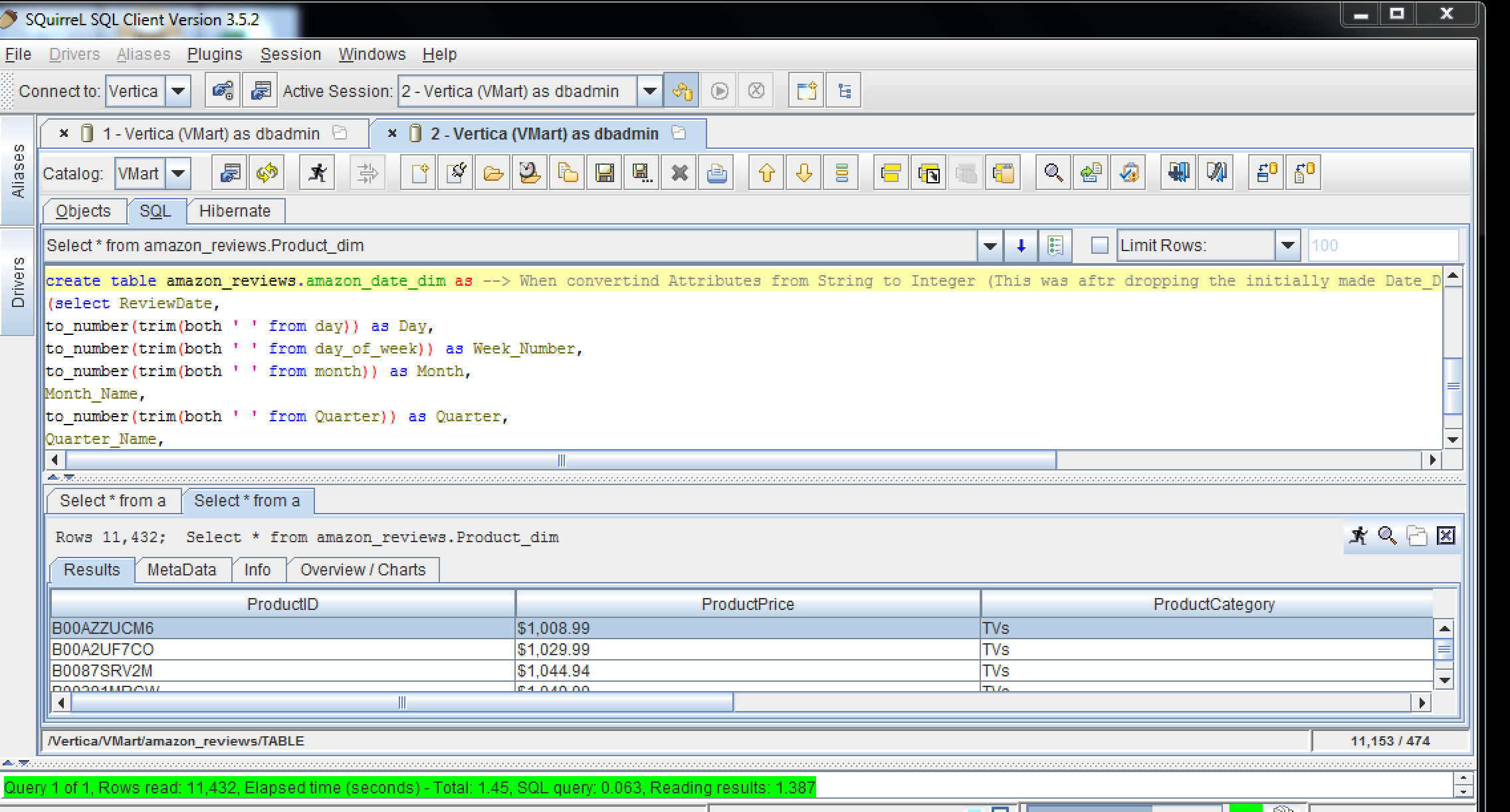
* **Exporting Amazon\_reviews.Product\_Reviews from SQLSquirrel to a .csv File –**



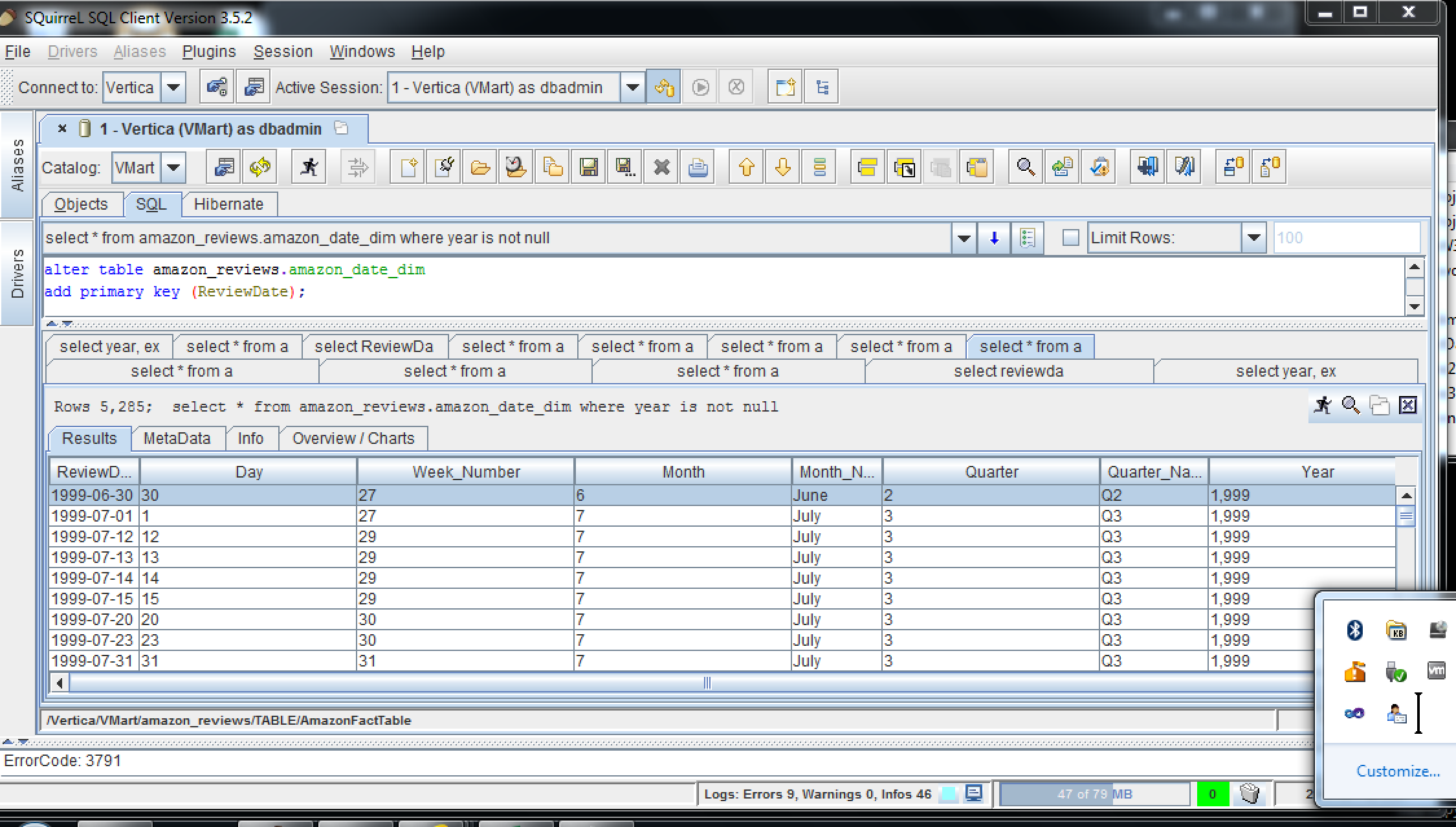
**Creation of Dimensions -**

* **Date\_Dimesion -**  
    
  

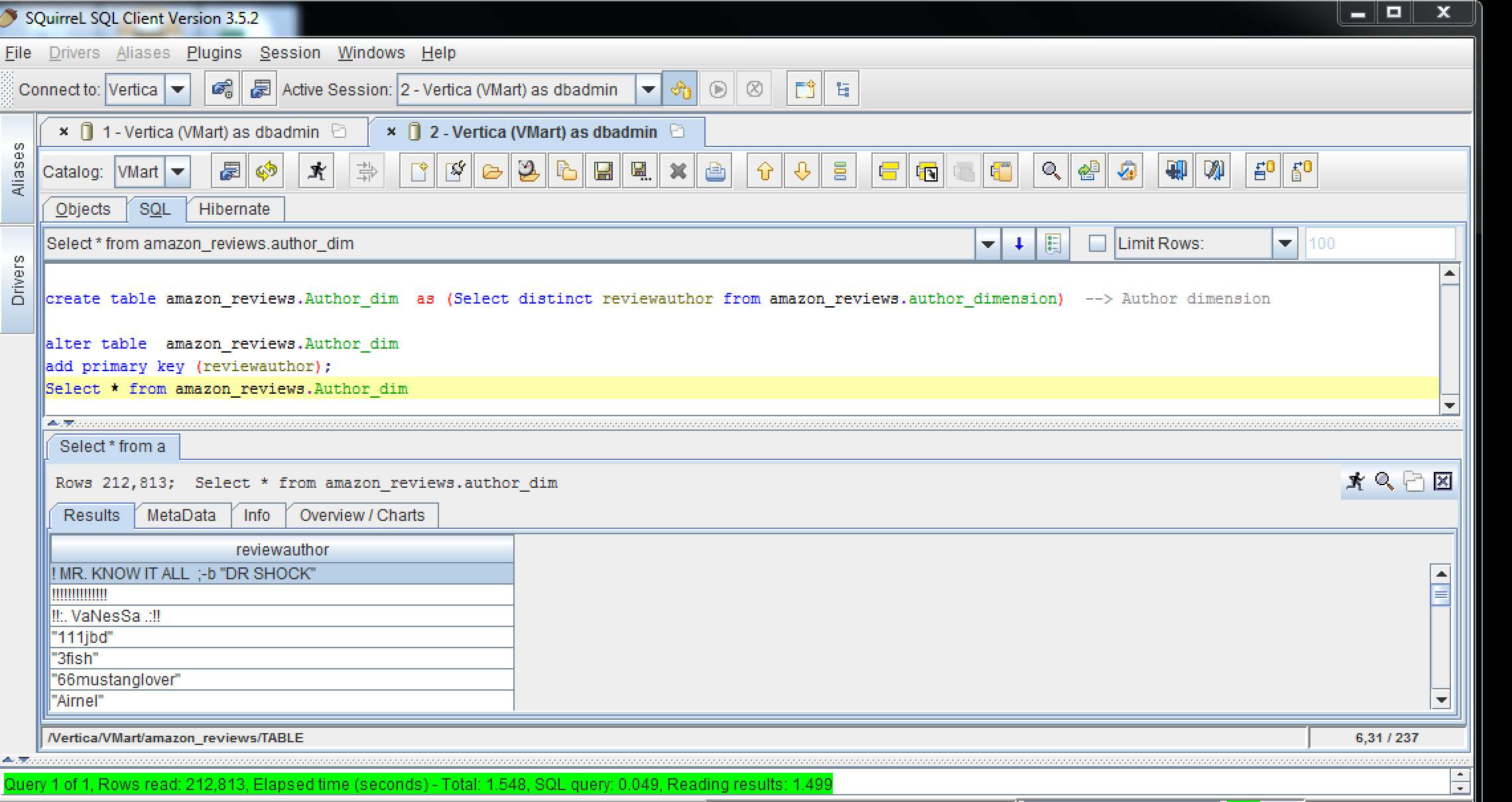
Creating another Date\_Dimension after converting String Datatypes to Integer(Double) –



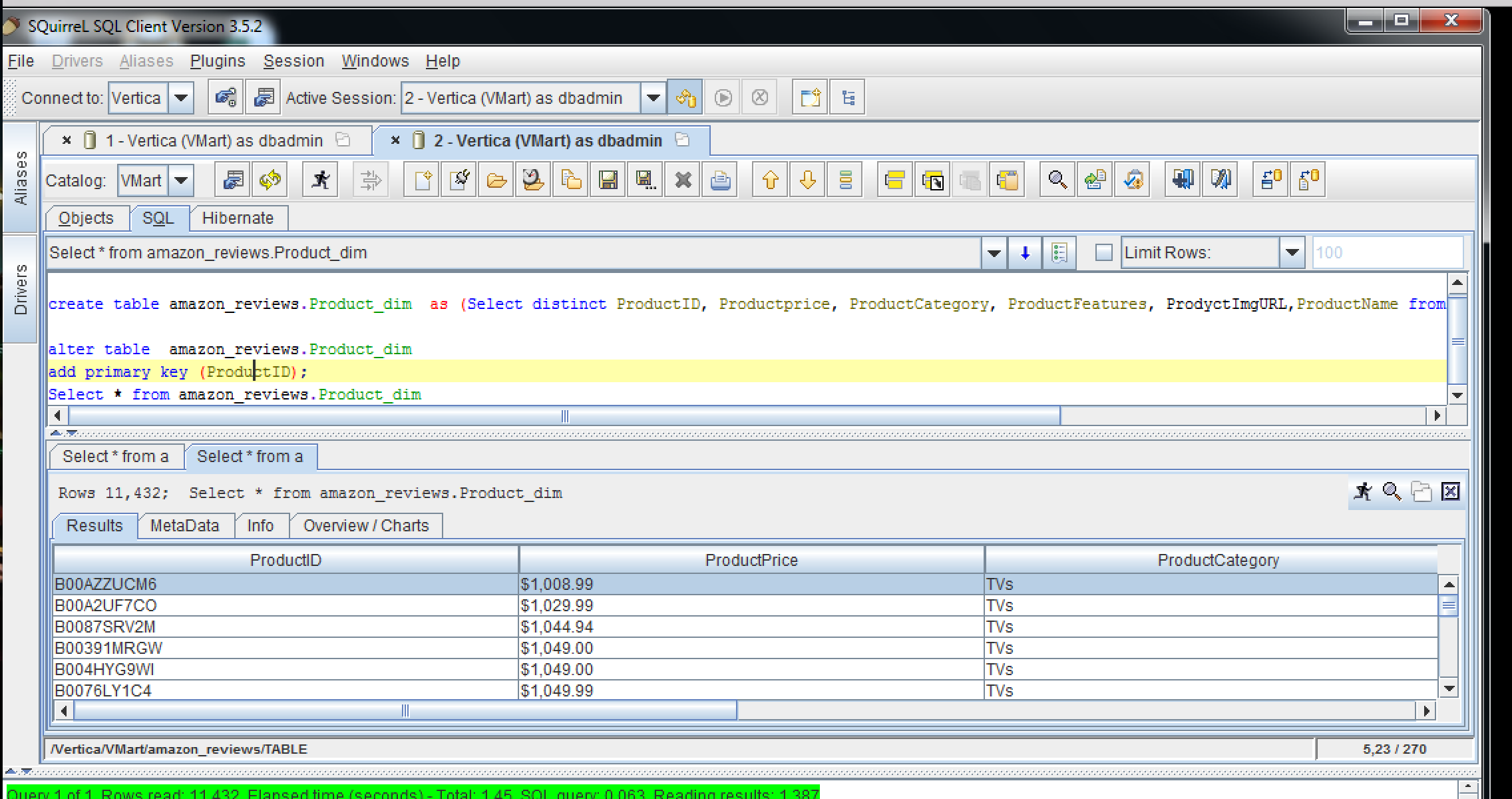
Making DateReview as PrimaryKey –



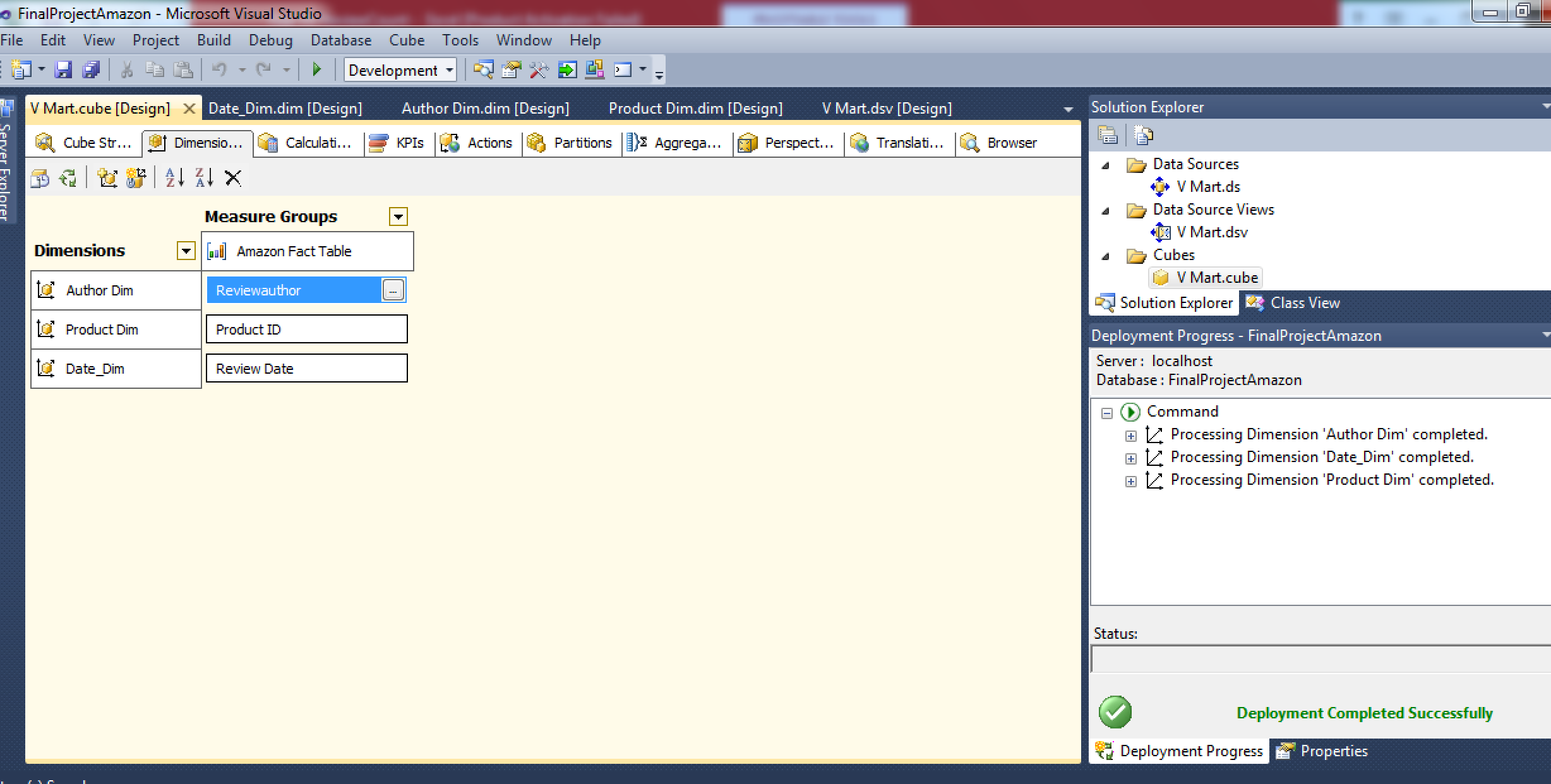
**Author Dimension and setting ReviewAuthor as the Primary key as well-**

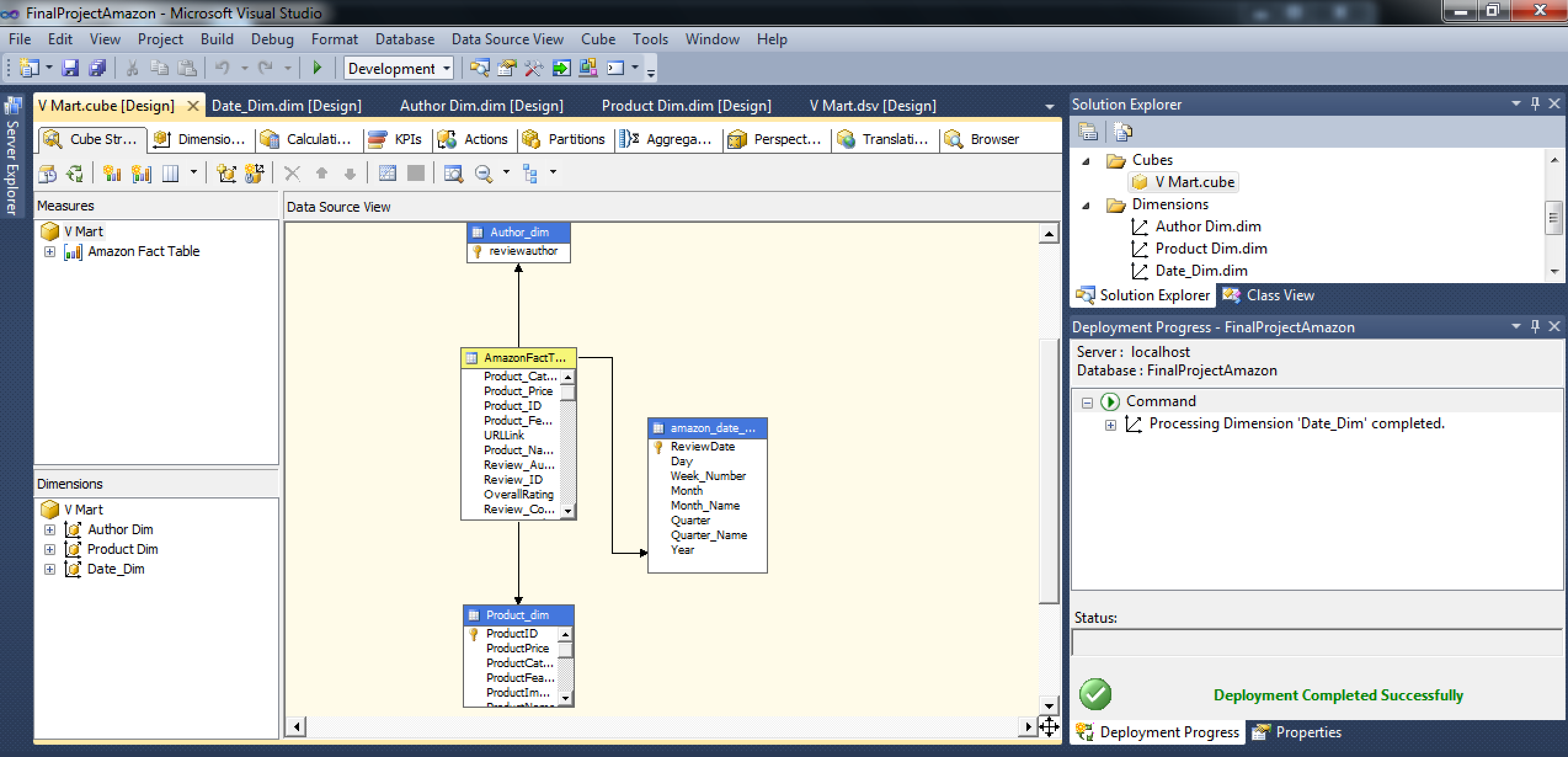


**Product Dimension and setting ProductID as the Primary key as well- –**



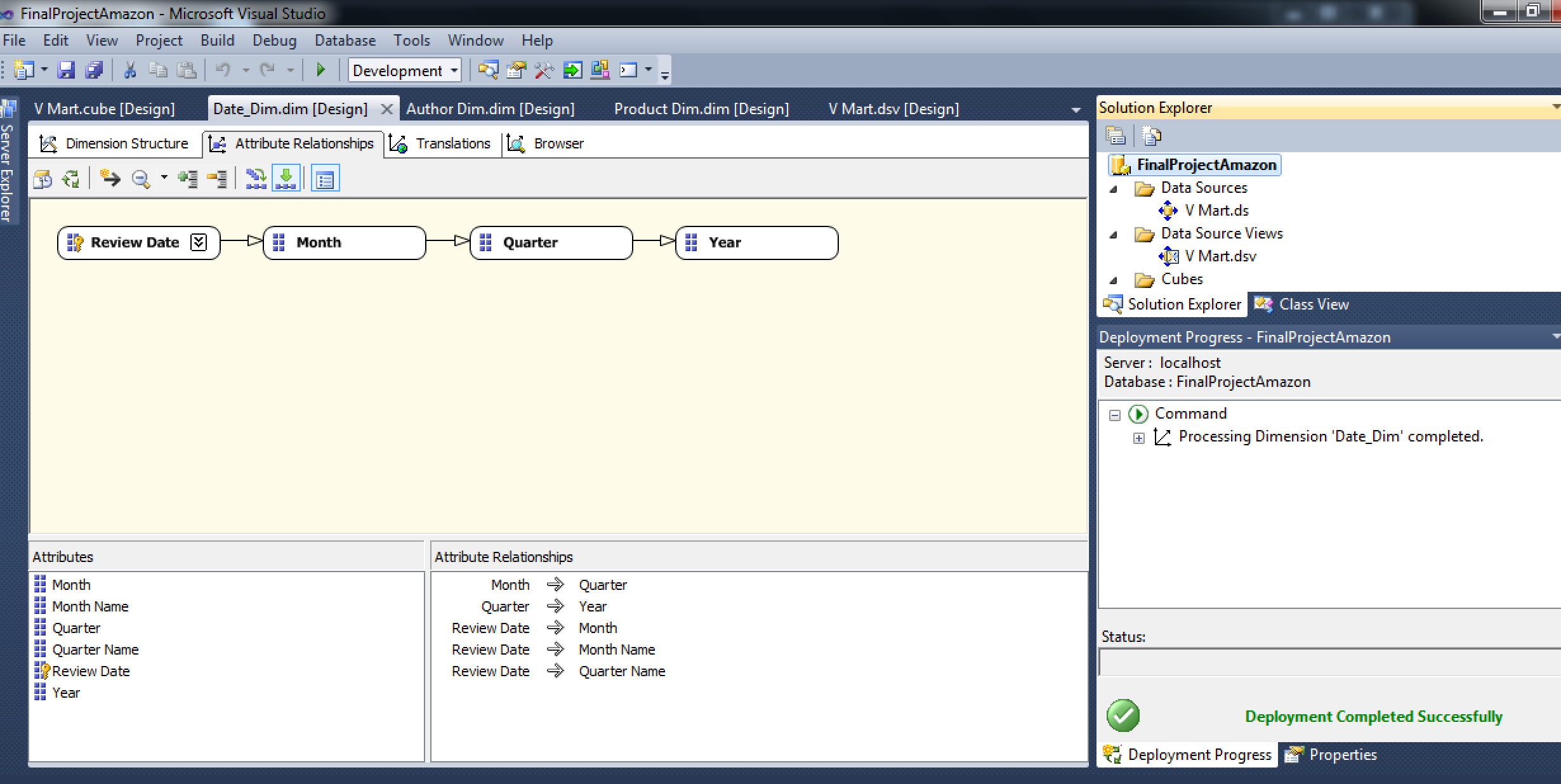
**VMart Cube Construction -**



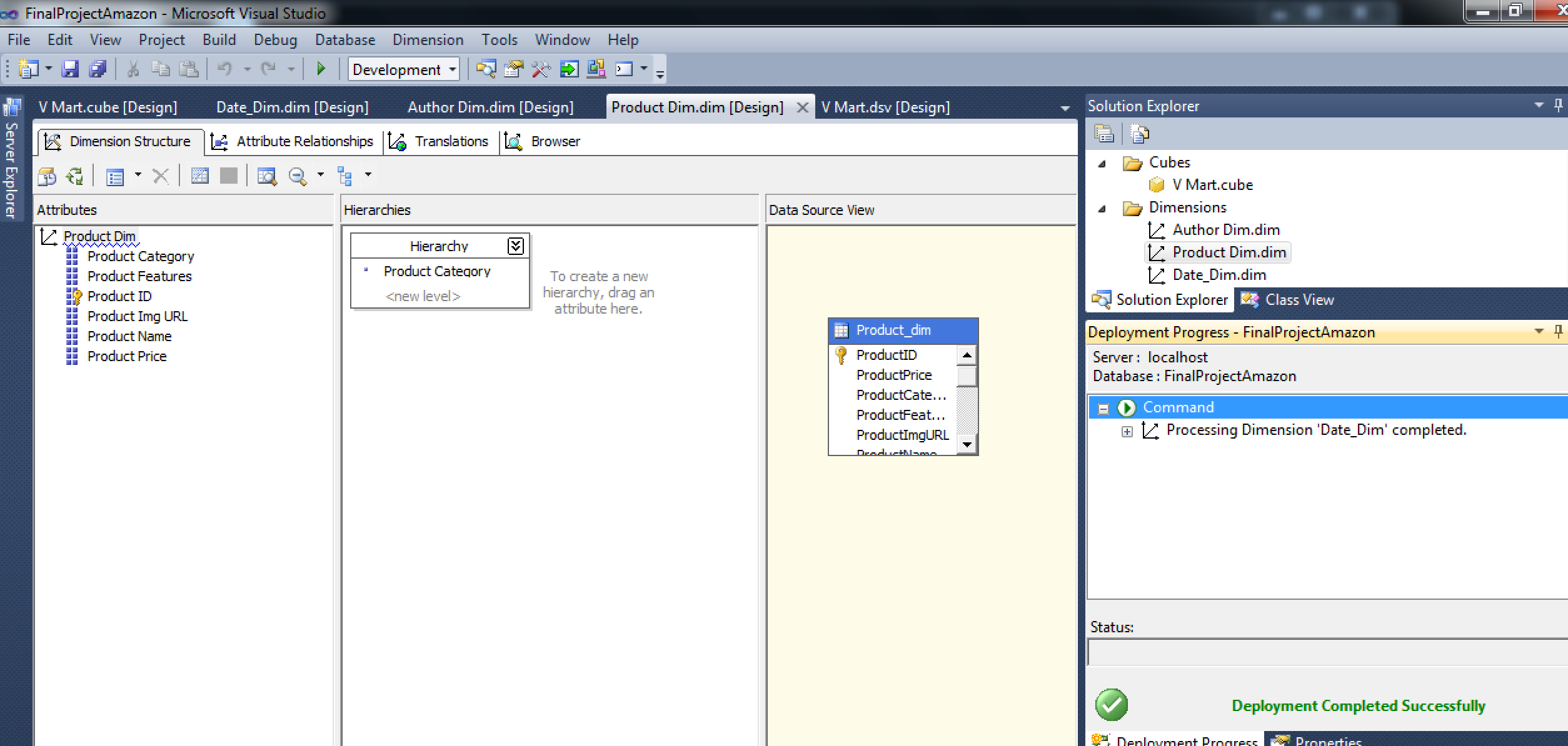


**OLAP Dimensions with Attribute Hierrachies –**

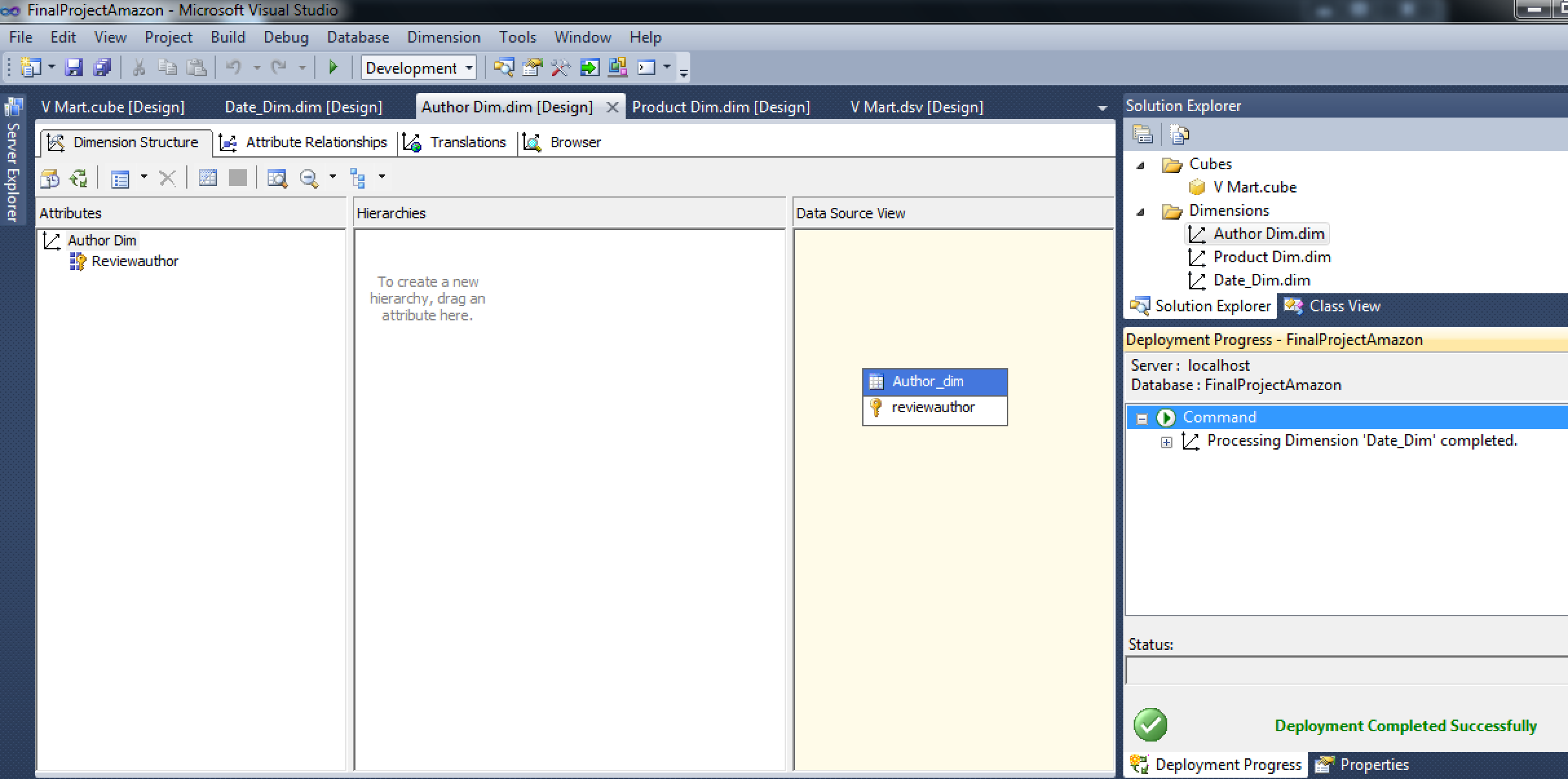
**Date\_Dimension –**

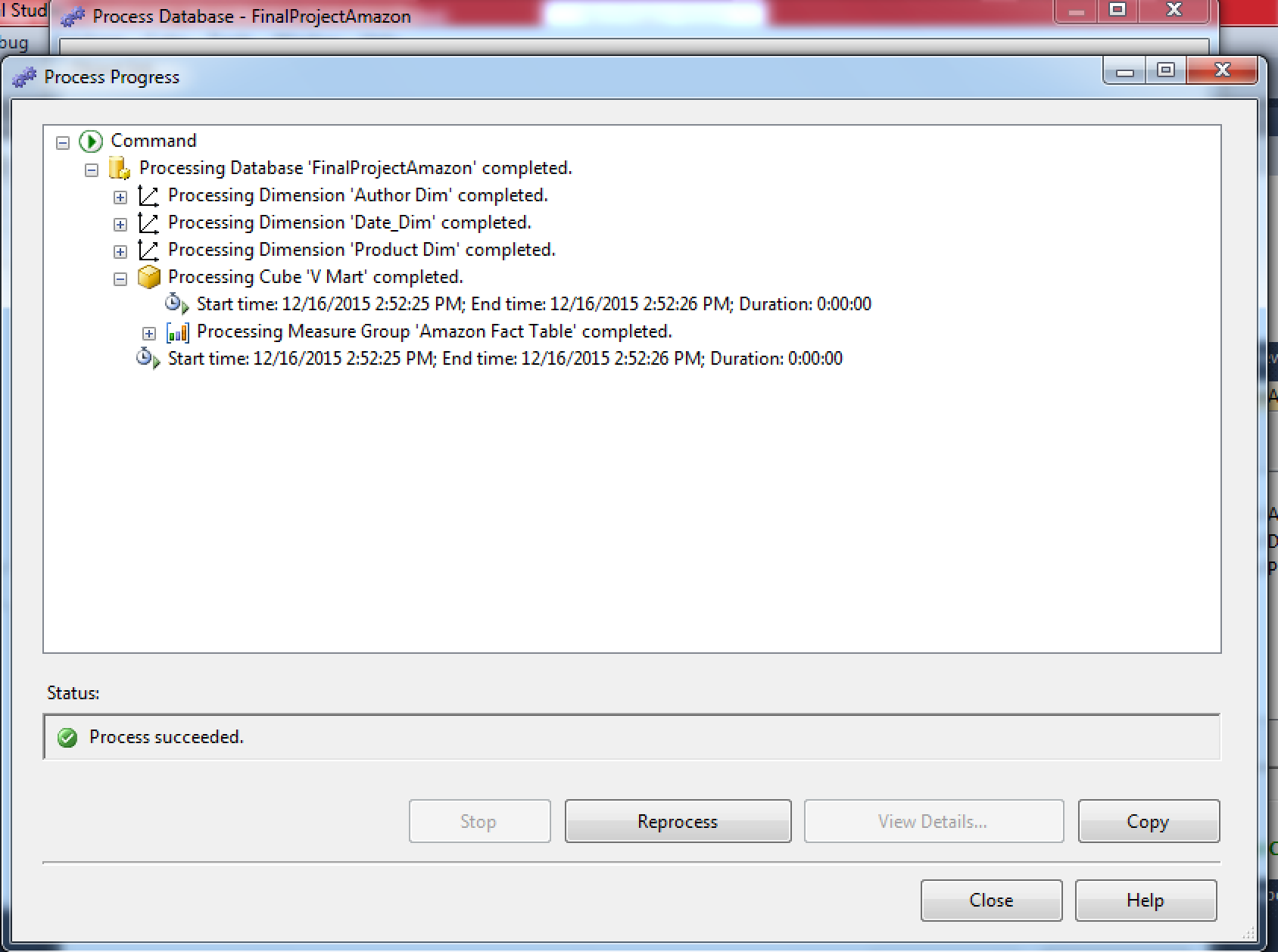


**Product\_Dimension –**



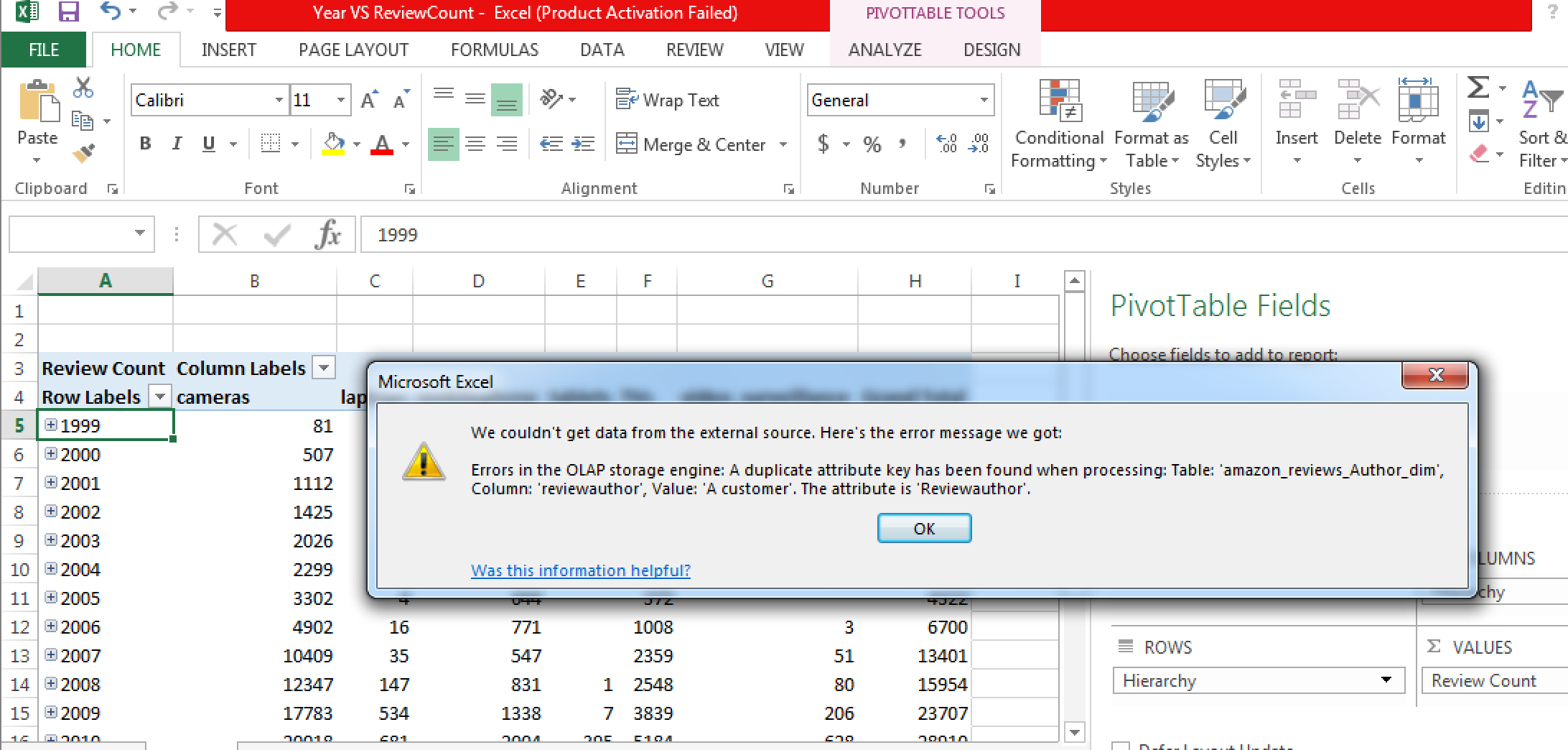
**Author\_Dimension –**



**Cube after Successful Processing -** 

The reports regarding the Project are in the Project Document Folder named – FinalProject-DataWarehouse-Nishant\_Jinsiwale.zip

**Errors Encountered in the Project –**



**Softwares tools involved –**

**Data Transformation** – Python   
 SQLSquirrelClient

**DataWarehouse** – HP Vertica  
  
**OLAP Engine** – Microsoft SSAS  
  
**Reporting Tool** – Microsoft Excel(Pivot Tables )

**Conclusion -**

This project introduced the idea of the design and implementation of an Data Warehouse within the context of a higher education environment, in order to better integrate systems for simpler and improved reporting. In order to implement a data warehouse, it is necessary to have a strong understanding of the concepts and software tolls involved .

When it comes to reporting, there are many tools available and I have used Microsoft Excel Pivot Table which integrates very well with the Microsoft OLAP(SSAS). The project concludes by studying some actual studies, involving the design and implementation of a Data Warehouse example as well as studying an actual case of improving queries.

With this practical exercise, we looked into running the same query against the same data but in very different evnironments.. Such practical examples as well as the study of data warehouse architecture and reporting capabilities, show the advantages of implementing a Data Warehouse for an improved experience for users who are querying and reporting on data for better-informed decisions.