Question 3:

As Data Engineer you are requested to write a Python application in order to connect to Adventure Works database on SQL Server and do data manipulation. https://docs.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-serverver15&tabs=ssms

Steps include:

• Restore AdventureWorks2019.bak database on SQL Server. It is important to use the 2019 OLTP version.

• Connect to the database using pyodbc or any other library you prefer

• Using tables Sales.SalesOrderHeader and Sales.SalesOrderDetail create and populate a Fact table called Fact.Sales to the best of your abilities

• Using Sales.Customer and tables in Person schema create and populate a Dimension called Dim.Customer. You do not need to use all the tables in Person schema to create this dimension.

• You can create Fact.Sales and Dim.Customer using either DDL or Python Notes: • Refer to AdventureWorks ERD.jpg which can help you in the above assignment

• It is highly recommended to write structured and well documented code

Solution:

After loading data to sql server, let’s create and populate fact and dimension table.

**Fact:**

As we’ve given two table SalesOrderHeader and SalesOrderDetail. Prior one gives us information of on order and the later gives product level details.

So, we’ll choose our **GRAIN** to be lowest possible so it is easier to summarize afterwards.

Our fact table will be transactional which means it will contain information about every transaction till the level of what product was sold.

Now we need to decide what columns we’ll be keeping from both of the table and what will be the **MEASURES**.

* Column is selected for fact table
* Column is not Selected for Fact table
* **Column in measure**

SalesOrderHeader

* SalesOrderID
* ShipMethodID
* RevisionNumber
* OrderDate
* DueDate
* ShipDate
* Status
* OnlineOrderFlag
* SalesOrderNumber
* PurchaseOrderNumber
* AccountNumber
* TerritoryID
* CustomerID
* BillToAddressID
* ShipToAddressID
* CreditCardID
* CreditCardApprovalCode
* CurrencyRateID
* **SubTotal**
* **TaxAmt**
* **Freight**
* **TotalDue**
* Comment
* rowguid
* ModifiedDate

SalesOrderDetail

* SalesOrderID
* SalesOrderDetailID
* CarrierTrackingNumber
* OrderQty
* ProductID
* SpecialOfferID
* UnitPrice
* UnitPriceDiscount
* LineTotal
* rowguid
* ModifiedDate

**DDL:**

This will be DDL of our fact table:

DDL can be find in factDDL.sql

**Dimension:**

As we need to only use person schema so we won’t join with columns such as **TerritoryID** or **StoreID.** We’ll only join it with Person.Person table using PersonId. For simplicity we’ll add data of only Person.Email and Person.Phone as these are most relevant.

Here is the DDL Statement of our dimension table.

CREATE TABLE [Dim].[Customer](

    [CustomerKey] [int] NOT NULL,

    [PersonKey] [int] NULL,

    [StoreID] [int] NULL,

    [TerritoryID] [int] NULL,

    [AccountNumber] [varchar](10) NOT NULL,

    [Customerrowguid] [uniqueidentifier] NOT NULL,

    [CustomerModifiedDate] [datetime] NOT NULL,

    [PersonType] [nchar](2) NULL,

    [NameStyle] [dbo].[NameStyle] NULL,

    [Title] [nvarchar](8) NULL,

    [FirstName] [dbo].[Name] NULL,

    [MiddleName] [dbo].[Name] NULL,

    [LastName] [dbo].[Name] NULL,

    [Suffix] [nvarchar](10) NULL,

    [EmailPromotion] [int] NULL,

    [AdditionalContactInfo] [xml](

        CONTENT [Person].[AdditionalContactInfoSchemaCollection]

    ) NULL,

    [Demographics] [xml](

        CONTENT [Person].[IndividualSurveySchemaCollection]

    ) NULL,

    [Personrowguid] [uniqueidentifier] NULL,

    [PersonModifiedDate] [datetime] NULL,

    [EmailAddress] [nvarchar](50) NULL,

    [EmailAddressID] [int] NULL,

    [Emailrowguid] [uniqueidentifier] NULL,

    [EmailModifiedDate] [datetime] NULL,

    [PhoneNumber] [dbo].[Phone] NULL,

    [PhoneNumberTypeID] [int] NULL,

    [PhoneModifiedDate] [datetime] NULL

) ON [PRIMARY] TEXTIMAGE\_ON [PRIMARY]

GO