AUSSIE BIKES CASE SCENARIO

Aussie Bikes (OZZB) specialises in manufacturing and selling three models of bicycles: (1) mountain bikes, (2) road bikes, and (3) touring bikes. Even though OZZB's primary business is focused on bicycle sales, it also sells accessories for bikers, such as bottles, bike racks, and brakes. In recent years, OZZB has extended to sports apparel, such as caps, gloves, and jerseys. Additionally, some portion of business includes sales of components like chains and derailleurs. While OZZB mainly manufactures bicycles, it purchases the apparel and the components from the other vendors. OZZB is not only in the business of manufacturing but also in the business of reselling.

OZZB does not own any traditional brick-and-mortar stores for retailing, but instead, it sells items in bulk to retail stores as a wholesaler. However, OZZB uses an internet platform for the retail sales to the individual customers. The OZZB business model divides customers into retail stores that sell bikes and individual customers. Overall, OZZB's customer base includes over 635 stores, over 18,484 personal customers, and a sales force of 17 salespeople who sell the products to customers. On the supply side, OZZB utilises services from over 100 vendor companies serving as suppliers of components, accessories, clothing, and raw materials. In recent years, OZZB has been a profitable and very successful business venture with a global customer base across the United States, Canada, Australia, the United Kingdom, France, and Germany. The company is eying an expansion of business operations but lacks a clear understanding of its market.

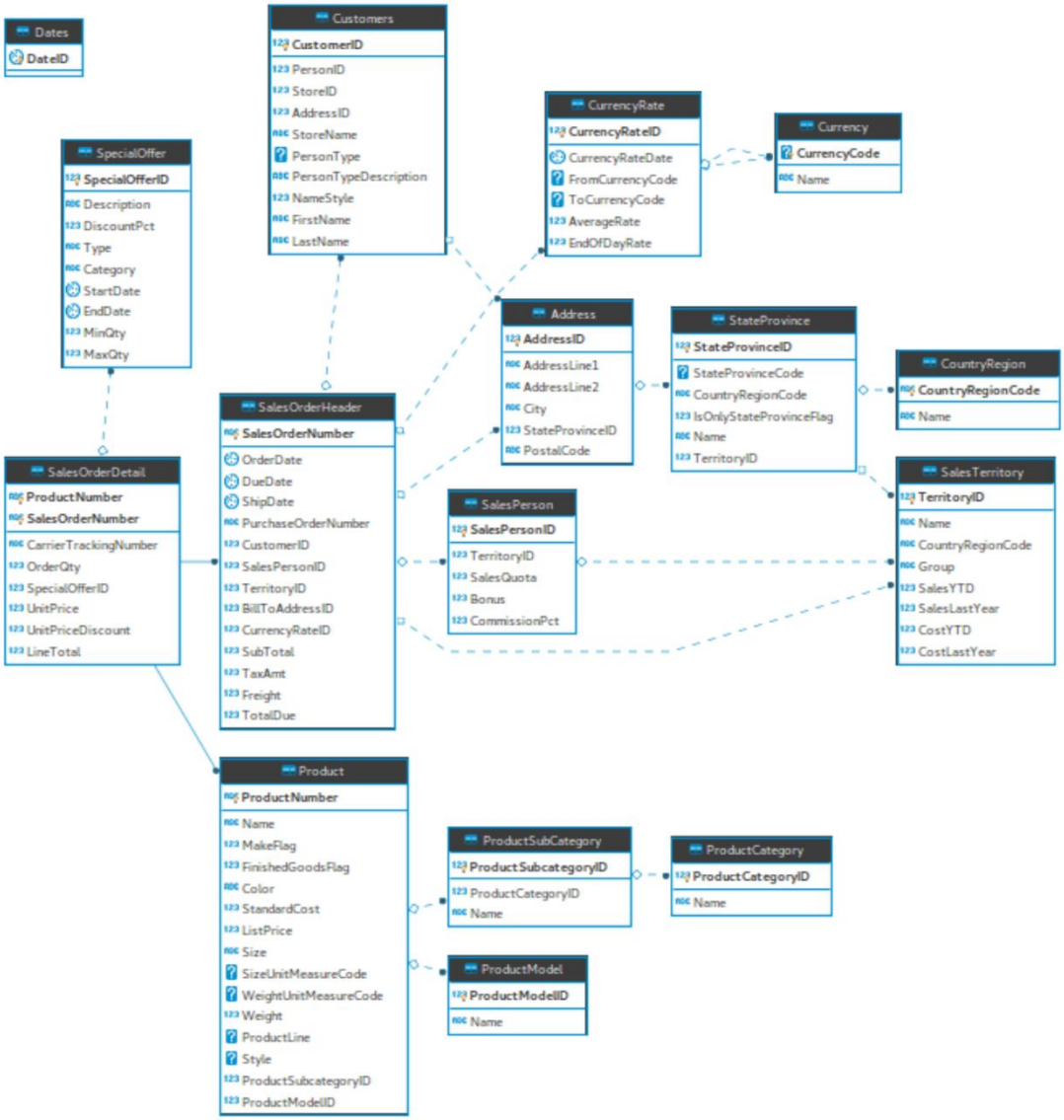
Amy is a newly hired manager and is tasked with the responsibility of building a better understanding of their current business before making the expansion decisions. In a recent business conference, Amy heard from vendors that business analytics can provide the business with the capability to make more informed decisions. She also discovered that the OZZB lacks the capabilities to make data-driven decisions as the board members rely on transactional databases to fetch the data. Amy identified the need for a data warehouse as a first step required for the expansion of business operations.

In a recent meeting, the board approved hiring a business analyst consultant to provide analytic insights regarding the profitability of various products. You have been hired as a business analyst consultant to propose a business analytics solution to the management team. Your job is to present a prototype of a data warehouse and make a business case by identifying the key customers, profitable products, and sales territory in the last two years.

Amy has specifically requested, **a) Data Model for the Data Warehouse, b) Description of data integration process (ETL), c) Sample Analytical queries.**

QC1. Who are the key customers in Australia?
QC2. Which products were the most profitable in 2014?
QC3. Which sales territories are the most profitable during December 2013?

Sales System ER Model



REPORT

Executive Summary

Data is an asset that can allow us to get insights on business activities, consumer preferences, and make critical decisions. The transactional databases currently in place are normalised and contains many tables. They are designed for efficient data writing and simple data retrieval but are not ideal for answering analytical questions as it requires accessing and joining data from multiple tables before aggregating the data.

To optimise and organise data for answering analytical queries, performing numerical aggregations, and understand various dimensions, a data warehouse prototype has been created. The data warehouse is designed with a denormalised data model (the dimensional model) and contains a separate informational database to optimize data reading and aggregating capabilities, support complex querying, and meet analytical information needs.

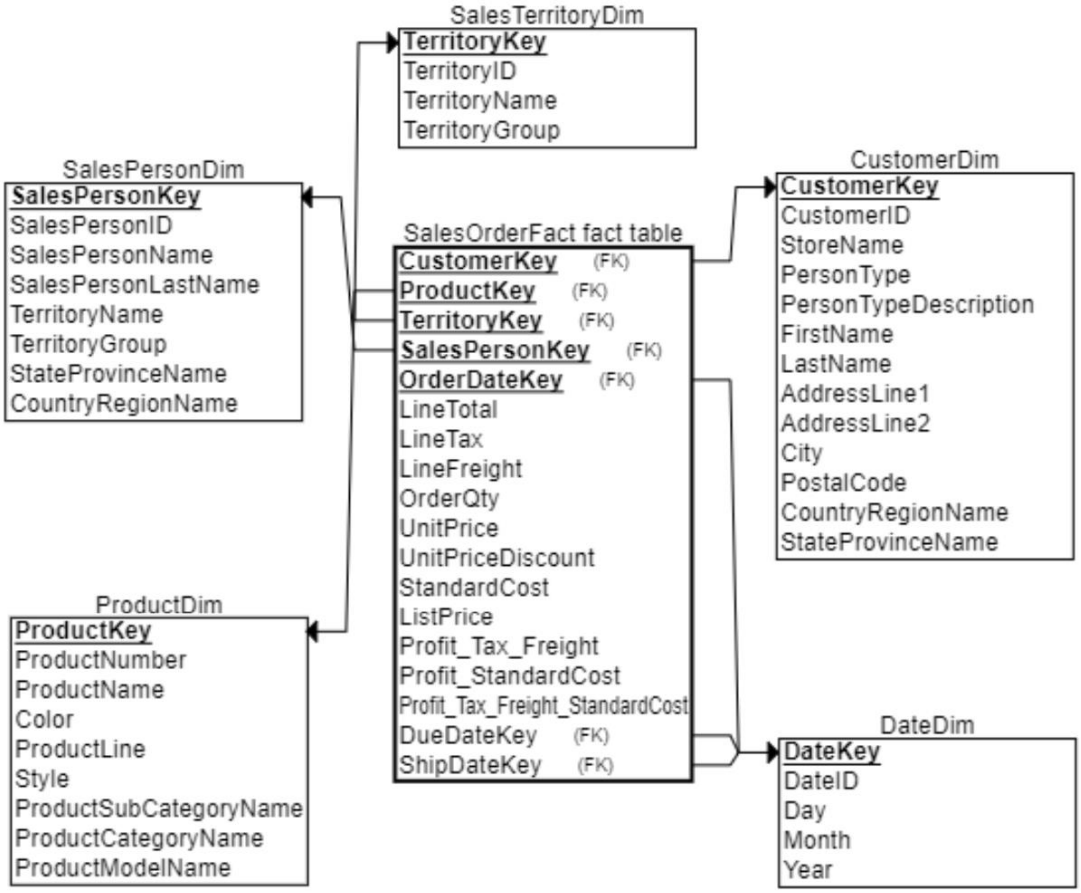
The dimensional model is based on the multi-dimensional model of data and is designed for retrieval-only databases. It is simple, intuitive, and allows us to focus on process measurement events such as profit and associated dimensions that provide descriptive context, i.e., fact tables and dimension tables. The dimensional model is designed before implementing the data warehouse.

The data from the operational systems is then combined and transformed (ETL) into a multi-dimensional model to extract valuable data (integrated, non-volatile, subject oriented, and time-variant data) which we then deposit in the data warehouse (informational database as per the dimensional model). The data warehouse in turn allows us to access a unified version of truth that can support decision-making.

This dimensional model has helped us garner the following insights:

- Richard Carey is our top customer in Australia based on net profit contributions.
- In 2014, the top 10 profitable products are all bikes. 60% of them are mountain bikes, 30% are touring bikes, and 10% are road bikes.
- In December 2013, our most profitable sales territory was Australia.

The Dimensional Model



I used Ralph Kimball's four step process in developing the dimensional model:

Step 1: Select the business process.

We will be focussing on the Sales Process (*OrderQty*, *LineTotal*, *Profit*, etc.) as our objective is to analyse profitability across dimensions such as customers and products.

Step 2: Define the Data Grain.

Each data grain represents an individual customer, unique product (including its quantity per order), sales territory, salesperson involved in the transaction, and the day of transaction (day is at the atomic grain level despite *OrderDate* using datetime data type; time isn't recorded as per data realities of the system). Data is at the lowest possible details for provision of analytical flexibility.

Step 3: Identify Dimensions.

Here the dimensions are customer, product, territory, salesperson, and date. Surrogate keys were generated for each dimension to connect the dimensions and facts while preserving the natural key.

This allows us to save storage space and prevent potential disruptions when records of the operational sources change over time. A Date dimension has also been included as they are fundamental for tracking changes across time periods.

Hierarchies were collapsed across the dimensions. Information from the *Product, ProductSubCategory, ProductCategory, ProductModel* tables in the operational database were collapsed into the product dimension. The *Customers*, *Address*, *StateProvince*, and *CountryRegion* tables were collapsed into the customer dimension. The *SalesPerson SalesPersonDetails*, *SalesTerritory, StateProvince,* and *CountryRegion* tables were collapsed into the salesperson dimension. Adding additional attributes helps to manage the complexity while logically maintaining the hierarchical structure of the data in the same dimensional table. For example, *ProductSubCategoryName* and *ProductCategoryName* were additional attributes added to the product dimension to provide additional flexibility in the analysis.

Step 4: Identify the Fact Measure.

The fact table uses a composite key consisting of the keys of all the dimension tables (*CustomerKey, ProductKey, TerritoryKey, SalesPersonKey, OrderDateKey*).

The facts correspond to the measurements of sales events at a point in space and time. Here we include the facts *LineTotal, LineTax, LineFreight, OrderQty, UnitPrice, UnitPriceDiscount, StandardCost, ListPrice, Profit_Tax_Freight, Profit_StandardCost, Profit_Tax_Freigth_StandardCost*.

These are numerical and additive in nature and supports mathematical operations such as aggregation for analysis. The *LineTax* and *LineFreight* facts are the tax amounts and freight costs of the transaction distributed evenly across the line items.

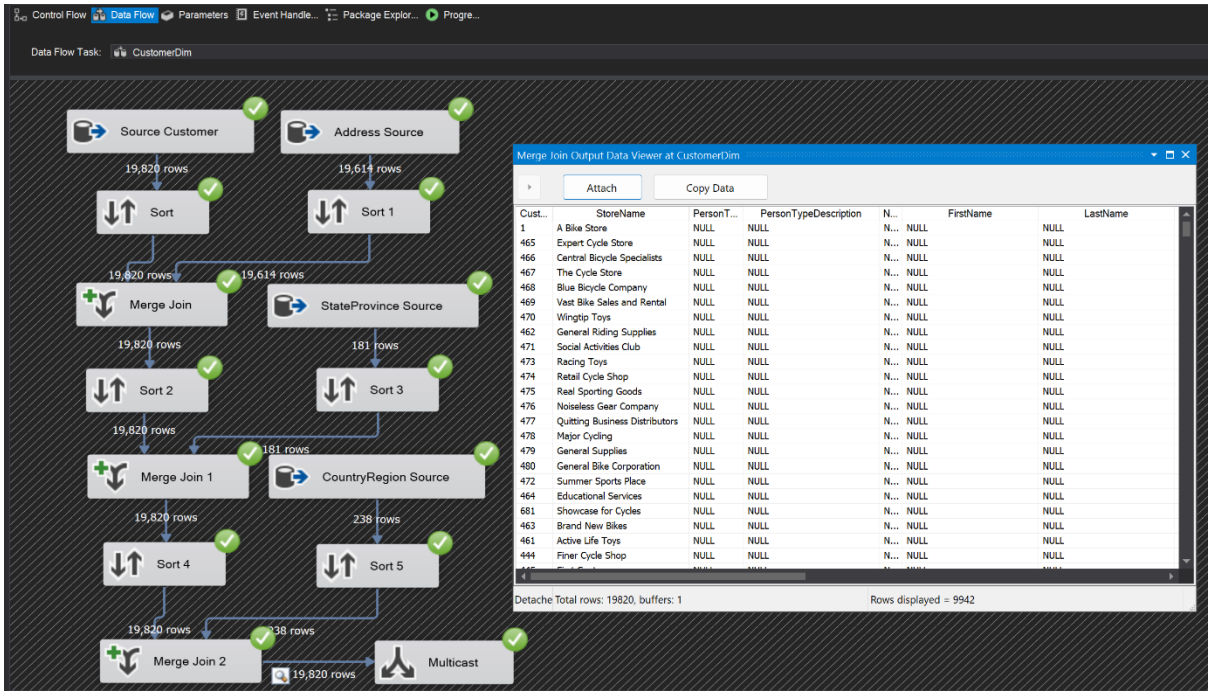
Profit_Tax_Freight, Profit_StandardCost, and Profit_Tax_Freigth_StandardCost represent profits calculated in different ways.

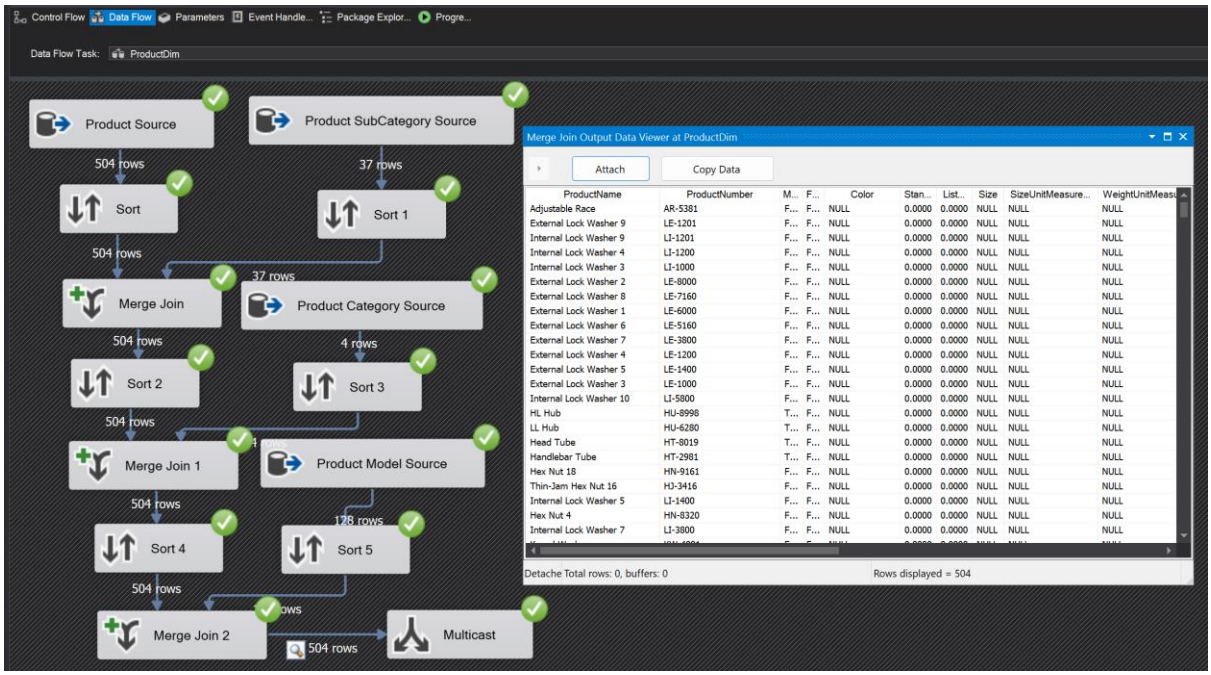
- Profit_Tax_Freight = LineTotal – LineTax – LineFreight
- Profit_StandardCost = LineTotal – (StandardCost * OrderQty)
- Profit_Tax_Freigth_StandardCost = LineTotal – LineTax – LineFreight – (StandardCost * OrderQty)

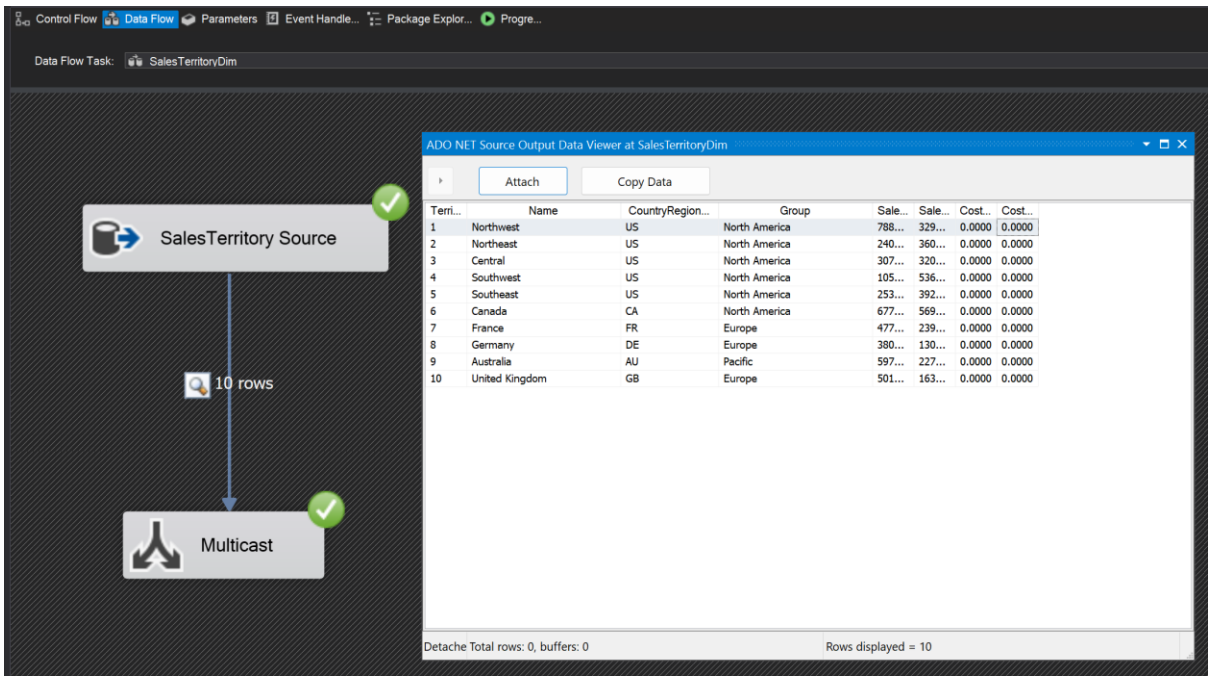
Note: *To navigate the appropriate equation for calculating profit we need to conduct an initial assessment as recommended by Kimball to understand the data realities, business context, and end-user requirements.* ***I will proceed with Profit_Tax_Freigth_StandardCost as profit for the purposes of this project and answer analytical queries****.*

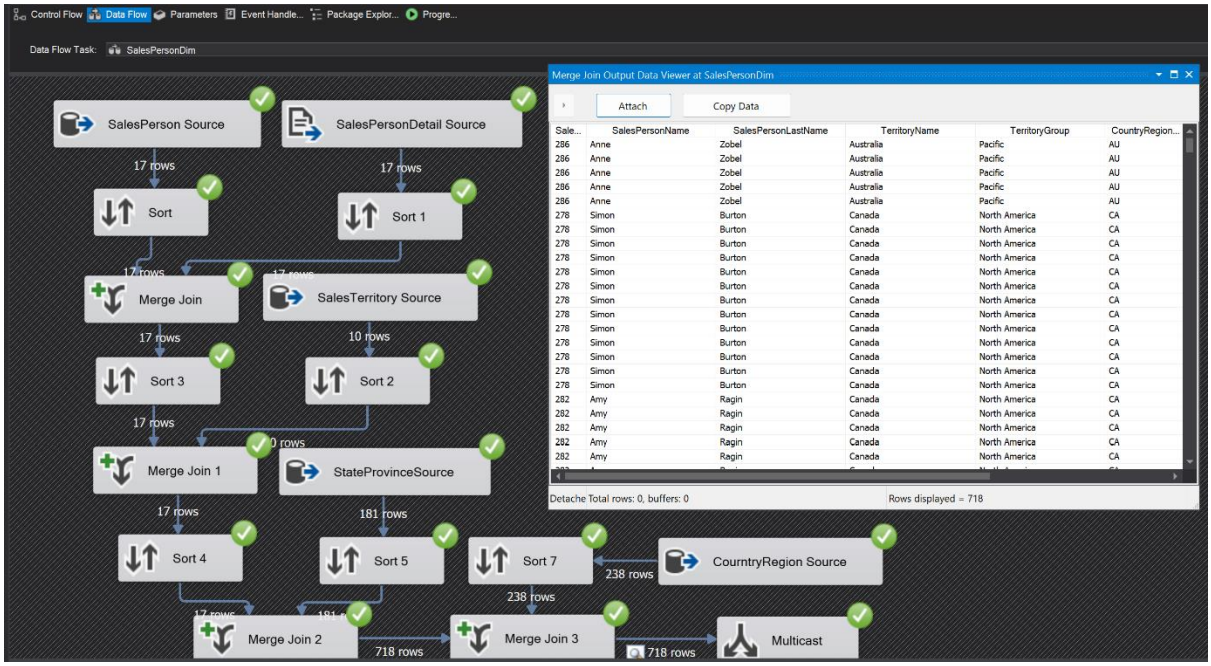
Implementation of the Dimensional Model (ETL with Microsoft SSIS)

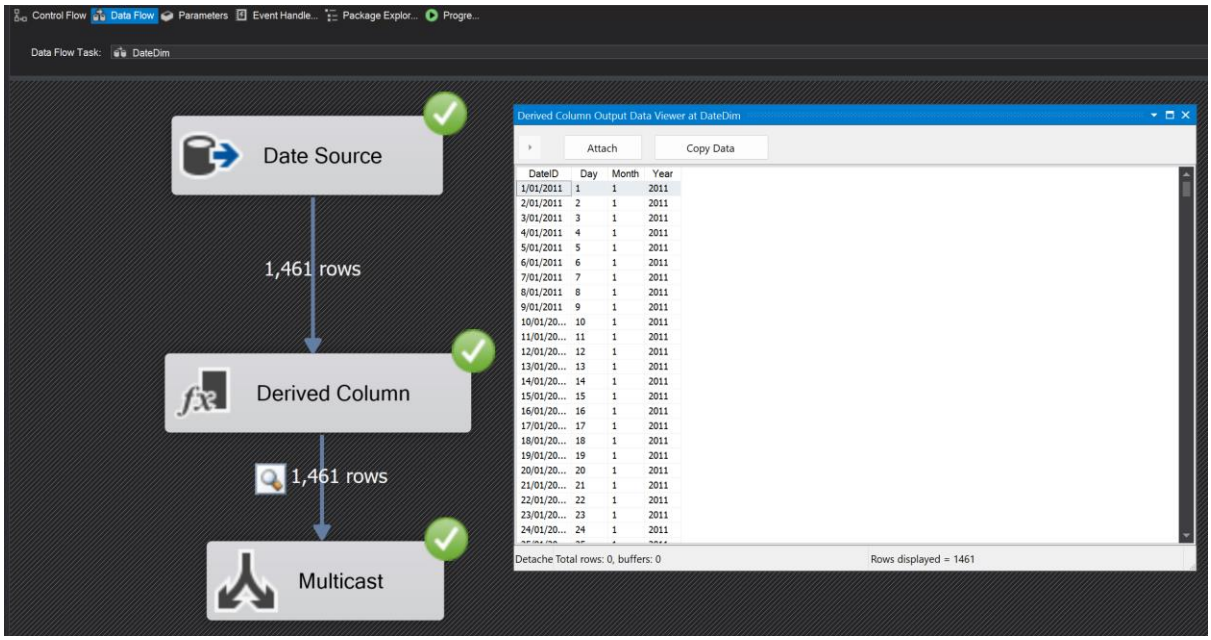
1. Dimensional Tables implementation



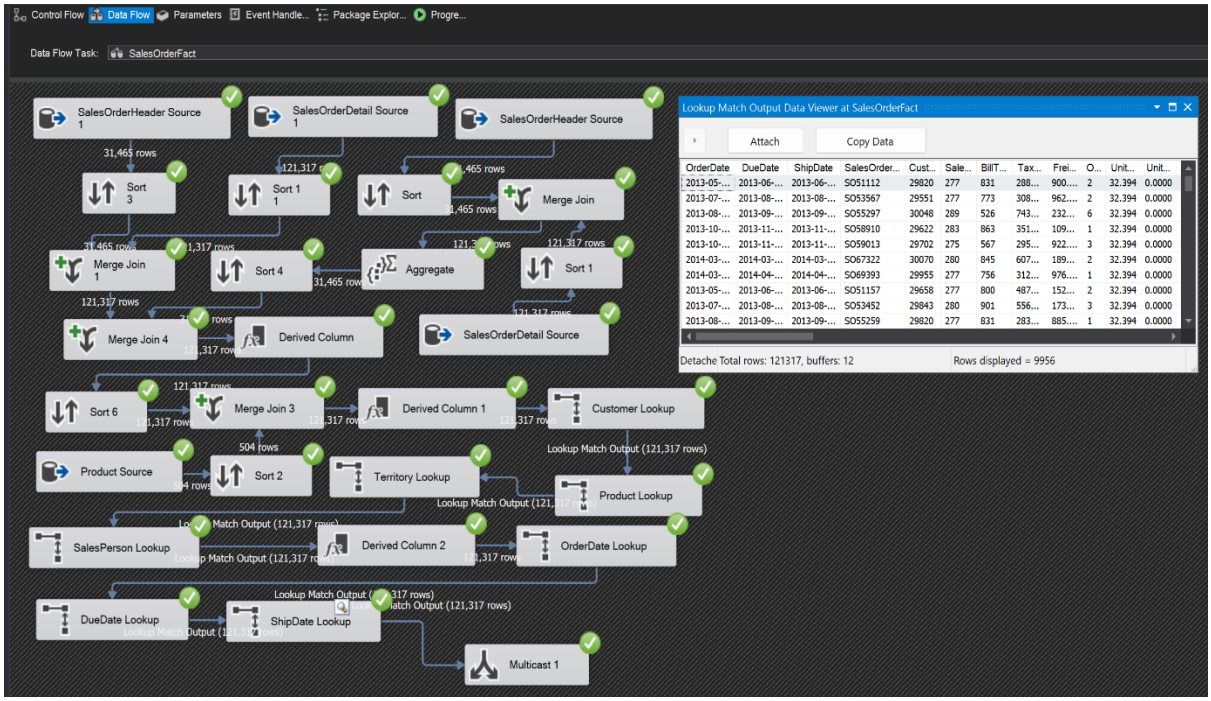



Control Flow
Data Flow
Parameters
Event Handle...
Package Explor...
Progre...
Data Flow Task: SalesTerritoryDim
SalesTerritory Source
10 rows
Multicast
ADO NET Source Output Data Viewer at SalesTerritoryDim
Attach
Copy Data
Terri... Name CountryRegion... Group Sale... Sale... Cost... Cost...
1 Northwest US North America 788... 329... 0.0000 0.0000
2 Northeast US North America 240... 360... 0.0000 0.0000
3 Central US North America 307... 320... 0.0000 0.0000
4 Southwest US North America 105... 536... 0.0000 0.0000
5 Southeast US North America 253... 392... 0.0000 0.0000
6 Canada CA North America 677... 569... 0.0000 0.0000
7 France FR Europe 477... 239... 0.0000 0.0000
8 Germany DE Europe 380... 130... 0.0000 0.0000
9 Australia AU Pacific 597... 227... 0.0000 0.0000
10 United Kingdom GB Europe 501... 163... 0.0000 0.0000
Detache Total rows: 0, buffers: 0
Rows displayed = 10


Control Flow
Data Flow
Parameters
Event Handle...
Package Explor...
Progre...
Data Flow Task: SalesPersonDim
SalesPerson Source
SalesPersonDetail Source
17 rows
Sort
Sort 1
Merge Join
SalesTerritory Source
10 rows
Sort 3
Sort 2
Merge Join 1
StateProvinceSource
181 rows
Sort 4
Sort 5
Sort 7
238 rows
CourntryRegion Source
Merge Join 2
718 rows
Merge Join 3
Multicast
Merge Join Output Data Viewer at SalesPersonDim
Attach
Copy Data
Sale... SalesPersonName SalesPersonLastName TerritoryName TerritoryGroup CountryRegion...
286 Anne Zobel Australia Pacific AU
278 Simon Burton Canada North America CA
282 Amy Ragin Canada North America CA
Detache Total rows: 0, buffers: 0
Rows displayed = 718



2. Fact Table implementation



3. The Implementation Process

Relevant components of the data warehouse architecture:

Source Data Systems:

Before we can integrate data into a dimensional model, we need to first identify the source data systems. Our main data source is the internal sales system modelled as a transactional database. We also have an internally sourced flat file that contains the names of the Aussie Bikes salespeople.

Data Staging Area:

After identifying the data sources, we then conduct Extract, Transform, Load (ETL) with SSIS. ETL facilitates creation of the dimensionally modelled data through data integration, transformation, cleaning and is vital in ensuring the quality of data and smooth functioning of the data warehouse.

We extract the data in the sales system from SQL Server Management Studio using an ADO Net Source in SSIS. To extract the flat file, which is stored as a .xlsx file, the file had to be converted to a CSV file before it could be read as a flat file source in SSIS.

In the transformation phase, we fix the data quality issues and prepare the data to load into the dimensional model.

- In the flat file, the *SalesPersonID* attribute is converted into a 4 byte signed integer from a general string data type to match the data type of *SalesPersonID* in the *SalesPerson* table. The data types need to match to perform an SSIS merge join task.
- For the dimension and fact tables, we merge data from various tables in the sales system using a merge join task.
- Artificially generated surrogate keys are added through SQL table creation queries in the ADO Net Destination before loading data into the dimensional tables in the data warehouse.
- In the *DateDim* table, we store additional information regarding Day, Month, and Year, based on the *DateID* attribute from the Dates table in the operational database via a derived column task.
- To calculate tax and freight amounts for each line item, we get the aggregate sum of the *OrderQty* attribute per order (since *LineTax*=*TaxAmt**(*OrderQty*/Sum(*OrderQty*))) via an aggregate task. With the calculated sum of *OrderQty* per order, we can derive the *LineTax* and *LineFreight* attributes via a derived column task. Once we have *LineTax* and *LineFreight* attributes, *Profit_Tax_Freight*, *Profit_StandardCost*, and *Profit_Tax_Freigth_StandardCost* (as defined earlier) can be calculated via an additional derived column task.
- We convert *OrderDate*, *DueDate*, and *ShipDate* from datetime data type to date data type via type casting and the derived column SSIS task. We do this for looking up matching date keys in the *DateDim* dimensional table, since *DateID* in *DateDim* has date data type.
- For the fact table, we can obtain the surrogate keys from the constructed dimensional tables via a lookup task by matching the natural key in the fact table with the corresponding key in the dimensional table.

Once the data is transformed, we can load the processed data into the data warehouse.

Data and Meta-Data Storage Area:

The final raw data is loaded onto the data warehouse, '*OZZB_SalesDataWarehouse',* a database in SQL Server Management Studio via SSIS ADO Net Destination.

Key Customers in Australia

	CustomerKey	CustomerID	CustomerProfitAustralia	FullName	City	State
1	3026	29628	14246.8	Richard Carey	Lavender Bay	New South Wales
2	1033	11767	3226.5	Meagan Madan	Hawthorne	Queensland
3	1618	11112	3214.2	Crystal Wang	Hervey Bay	Queensland
4	2214	12338	3199.7	Monica Vance	South Melbourne	Victoria
5	2130	11101	3186.5	Abby Sai	Cranbourne	Victoria
6	3399	11120	3185.6	Beth Jiménez	Sydney	New South Wales
7	2386	11109	3183.6	Ruben Kapoor	South Melbourne	Victoria
8	3344	11451	3182.7	Ruben Muñoz	Newcastle	New South Wales
9	1866	11995	3178.1	Kelvin Carson	Sunbury	Victoria
10	709	11901	3177.5	Stacy Alvarez	Darlinghurst	New South Wales

Richard Carey is our top customer in Australia based on net profit contributions, contributing ~14247$ in profits. His profit contribution is ~4.6 times that of our 2nd top customer in Australia, Meagan Madan. Furthermore, 80% of our key customers from Australia seem to be from either New South Wales or Victoria.

SQL:

```
WITH cte1 AS (SELECT C1.CustomerKey,
CAST(ROUND(SUM(F.Profit_Tax_Freigth_StandardCost),1) AS FLOAT) AS
CustomerProfitAustralia
      FROM SalesOrderFact F
      JOIN CustomerDim C1
      ON F.CustomerKey=C1.CustomerKey
      WHERE C1.CountryRegionName='Australia'
      GROUP BY C1.CustomerKey
)
SELECT TOP 10 cte1.CustomerKey, C2.CustomerID,
cte1.CustomerProfitAustralia, C2.FirstName + ' ' + C2.LastName AS
FullName, C2.City, C2.StateProvinceName AS State
FROM cte1
JOIN CustomerDim C2
ON cte1.CustomerKey=C2.CustomerKey
ORDER BY cte1.CustomerProfitAustralia DESC;
```

Most Profitable Products in 2014

	ProductKey	ProductProfit2014	ProductName	Category	SubCategory	Model
1	322	263799.4	Mountain-200 Black, 38	Bikes	Mountain Bikes	Mountain-200
2	320	263643.1	Mountain-200 Silver, 38	Bikes	Mountain Bikes	Mountain-200
3	321	260559.7	Mountain-200 Black, 42	Bikes	Mountain Bikes	Mountain-200
4	324	240766.9	Mountain-200 Silver, 46	Bikes	Mountain Bikes	Mountain-200
5	323	234601.9	Mountain-200 Silver, 42	Bikes	Mountain Bikes	Mountain-200
6	325	227075.6	Mountain-200 Black, 46	Bikes	Mountain Bikes	Mountain-200
7	398	77168.5	Touring-1000 Blue, 54	Bikes	Touring Bikes	Touring-1000
8	401	75009.5	Touring-1000 Blue, 46	Bikes	Touring Bikes	Touring-1000
9	399	74805.4	Touring-1000 Yellow, 54	Bikes	Touring Bikes	Touring-1000
10	359	73512	Road-350-W Yellow, 44	Bikes	Road Bikes	Road-350-W

In 2014, our top 10 profitable products are all bikes! Among the top 10 bike products, 60% are mountain bikes, 30% are touring bikes, and 10% are road bikes. Also, the most profitable mountain bike makes ~3.4 that of the most profitable touring bike. Most of these top bikes are either black or silver in colour.

SQL:

```
WITH cte1 AS (SELECT P.ProductKey,
CAST(ROUND(SUM(F.Profit_Tax_Freigth_StandardCost),1) AS FLOAT) AS
ProductProfit2014
      FROM SalesOrderFact F
      JOIN ProductDim P
      ON F.ProductKey=P.ProductKey
      JOIN DateDim D
      ON F.OrderDateKey=D.DateKey
      WHERE D.Year=2014
      GROUP BY P.ProductKey
)
SELECT TOP 10 cte1.ProductKey, cte1.ProductProfit2014, P.ProductName,
P.ProductCategoryName AS Category, P.ProductSubCategoryName AS
SubCategory, P.ProductModelName AS Model
FROM cte1
JOIN ProductDim P
ON cte1.ProductKey=P.ProductKey
ORDER BY cte1.ProductProfit2014 DESC;
```

Most Profitable Sales Territories in December 2013

	TerritoryKey	TerritoryProfitDec2013	UnitSales	DollarSales	TerritoryName	TerritoryGroup
1	9	148812.5	1091	468650.3	Australia	Pacific
2	1	95444.5	1725	637728.4	Northwest	North America
3	4	94715.8	1911	749160.2	Southwest	North America
4	8	63245.3	973	308301.2	Germany	Europe
5	10	59685.1	790	306343.3	United Kingdom	Europe

In December 2013, our most profitable sales territory is Australia in the pacific group with ~148812.5$ in profits for the month. Although Australia is the most profitable sales territory, our 2nd and 3rd most profitable territories, Northwest and Southwest respectively, contributed to more unit sales and dollar sales for the month. This could perhaps be because of lower costs incurred within Australia.

SQL:

```
WITH cte1 AS (SELECT T1.TerritoryKey,
CAST(ROUND(SUM(Profit_Tax_Freigth_StandardCost),1) AS FLOAT) AS
TerritoryProfitDec2013, SUM(F.OrderQty) AS UnitSales,
CAST(ROUND(SUM(F.LineTotal),1) AS FLOAT) AS DollarSales
      FROM SalesOrderFact F
      JOIN SalesTerritoryDim T1
      ON F.TerritoryKey=T1.TerritoryKey
      JOIN DateDim D
      ON F.OrderDateKey=D.DateKey
      WHERE D.Year=2013 AND D.Month=12
      GROUP BY T1.TerritoryKey
)
SELECT TOP 5 cte1.*, T2.TerritoryName, T2.TerritoryGroup
FROM cte1
JOIN SalesTerritoryDim T2
ON cte1.TerritoryKey=T2.TerritoryKey
ORDER BY cte1.TerritoryProfitDec2013 DESC;
```

By denormalizing the data, the dimensional model has made it simple to answer analytical queries (as observed from the simpler SQL queries), perform numerical aggregations, and understand various dimensions.