

# CITS3401 Data Exploration & Mining Project 1

## Healthy Burgers Fast Food Chain

Aleck Greenham 20362627

Ash Tyndall 20915779

26th April 2013

## Abstract

This document details the design of a data cube suitable for the *Healthy Burgers* franchise, a fast-food restaurant chain with stores in three countries, to aid management in its plans for improving sales and global expansion. Where the system requirements [1] were ambiguous or incomplete, reasonable assumptions were made and documented.

## Introduction

The client, Healthy Burgers, wishes to mine historical sales data from their nine stores spread across three countries in Australasia to aid management in making decisions regarding global expansion of the franchise and improving sales at existing restaurants. Each restaurant maintains a Online Transaction Processing (OLTP) database and there are aggregate databases at the state, country and worldwide levels. The company maintains data on sales, products, suppliers and restaurants. The purpose of this document is to detail the design of a data cube to extract data from these databases that will support the decisions faced by Healthy Burgers management.

## Limitations

It is outside the scope of this assessment to address the importing of data from the heterogeneous OLTP restaurant databases, including the cleaning of data. It shall be assumed that the data is complete, easily available and in the desired formats. It is acknowledged, however, that this would not be the case in real industry applications.

A detailed interpretation of the results of the OLAP operations is also outside the scope of this assessment. It falls to Healthy Burgers management to fully interpret and make recommendations or decisions based on the data stored within the data warehouse.

## Requirements

Table 1 is a summary of authors' interpretation of the initial requirements:

Object	Properties	Restrictions
Country		Australia, New Zealand, Singapore
Store	Interior Design Facility Type	3 restaurants in each country 3 different interior designs Facilities are 'dine in', 'drive through' and 'both'
Combo Meal	Price Category Calorie Value Products	6 different combo meals Not all stores have combo meals Divided into 3 different price categories
Supplier		There are 2 suppliers Suppliers do 3 combo meals each
Sale	Date Combo	Date Between 2008 and 2012
Sales Period		Has 2 combo meals Can be 'breakfast', 'lunch' or 'dinner'
Promotional Period		6 promotional periods per year

Table 1: System Requirements Summary

## Assumptions

The following assumptions were made:

- To make the initial requirements [1] sufficiently complete
  - To make the design scenario as realistic as possible
  - For the sake of simplicity or clarity (where realism was impossible, impractical or outside the scope of assessment)
1. The type of ingredients are not important, only their supplier - i.e. the supplier is a proxy for how desirable the ingredients they supply are.
  2. There are six combo meals globally.
  3. There are 3 sales periods, each with 2 combo meals, and only 6 combo meals to choose from. If it's true that not all stores sell all the combo meals, then some combo meals must be shared between sales periods.
  4. All non-combo products sold are not of interest (because they have no relation with any of the other data fields).
  5. The same suppliers provide the ingredients for same combo meals, globally.
  6. A promotional period has its own 6 combo meals (thus the combo meals change every 2 months).
  7. The price of combo meals is set at the store level.

## Schema Design

A Star Schema was chosen as the design document [1] states Healthy Burgers is only interested in sales, which removes the need to have multiple fact tables (i.e. a Constellation Schema). Table 2 shows the dimension tables used in the solution implementation.

Dimension	Fields
Store	store_key country
Design	design_key
Facilities	facility_key
SalesPeriod	sales_period_key
PromotionalPeriod	promotional_period_key
Years	years_key
Combos	combo_key calories
Suppliers	supplier_key
Price Range	price_range_key
Calorie Range	calorie_range_key

Table 2: Dimension Tables

Field Type	Field
Dimension Keys	store_key
	design_key
	facility_key
	sales_period_key
	promotional_period_key
	years_key
	combo_key
	supplier_key
Measures	price_range_key
	calorie_range_key
	total_sales_in_dollars

Table 3: Fact Table

The time values have been split into a number of different dimensions to facilitate the comparison of different time periods against each other. An example query is comparing the total sales during each sales period (Breakfast, Lunch or Dinner) across each promotional period (1st promotion, 2nd promotion, etc.) to find out whether a promotion was particularly successful for a given sales period.

The remaining data has been split up into dimensions to accommodate queries that would be reasonably expected to be of interest to Healthy Burgers management.

The requirements state that Healthy Burgers is only interested in a single measurement: sales in dollars; this is reflected in the fact table which consists of keys to all the of available dimensions, and the total sales in dollars.

## Implementation

### Data Generation

A Python script (`genvals.py`) authored by Tyndall generated random data that fit within the restrictions imposed by the initial requirements and what was considered realistic. Product names were generated from random dictionary words; product attributes such as price and calorie values were normally variate around a realistic mean and standard deviation. The rest of the data was similarly generated.

In an endeavour to make the data as realistic as feasibly possible, care was taken to influence it in a variety of sensible ways. Sales were positively correlated with increasing caloric content, and similarly correlated with decreasing price. Different suppliers and store designs affect sales and countries enjoy different sales periods to different extents.

### Data Import

An Excel workbook (`data-import.xlsx`) was developed with two worksheets to handle the generation of the attribute structure required by Palo; data cleaning and categorisation; and fact insertion. You can see examples of their function in `create-structure.png` and `create-data.png`.

1	The Motivate Australia	No roof	Both	Breakfast	1	2008	Organized Husi Tyndall Wholefoo 1018	14548	345	Original Data
2	=SAS1	=SCS1	=SDS1	=SES1	=VLOOKUP(SF\$1:SD\$21:SE\$26,2)	=SG\$1	=SH\$1	=SIS1	=VLOOKUP(SIS1:SD\$29:SE\$34,2,TRUE)	=SK\$1
3	StoreName	Not Used	StoreDesign	StoreFacilities	SalesPeriod	PromotionalPeriod	SalesYear	ComboName	ComboSupplier	ComboCalories
4										
5	Config									
6	Database	localhost/BurgerStore								
7										
8			Dimensions (in order)				Warning: Magic			
9							=PALO.SETDATA(SK\$2,FALSE,SB\$7,SB\$10,SA\$2,SC\$2,SD\$2,SE\$2,SF\$2,SG\$2,SH\$2,SIS\$2,SK\$2,SL\$2)			
10	Cube Name	Sales					Add number of sales dimension			
11			Stores							
12			Designs							
13			Facilities							
14			SalesPeriods							
15			PromotionalPeriods							
16			Years							
17			Combos							
18			Suppliers							
19			PriceRange							
20			CalorieRange							
21			Month/Number							
22			1	Jan/Feb						
23			2	Mar/Apr						
24			3	May/Jun						
25			4	Jul/Aug						
26			5	Sep/Oct						
27			6	Nov/Dec						
28			PriceRange creation							
29			0	\$0 - \$1.99						
30			200	\$2 - \$3.99						
31			400	\$4 - \$5.99						
32			600	\$6 - \$7.99						
33			800	\$8 - \$9.99						
34			1000	\$10+						
35			CalorieRange creation							
36			0	0 - 249						
37			250	250 - 499						
38			500	500 - 749						
39			750	750 - 999						
40			1000	1000 - 1249						
41			1250	1250 - 1499						
42			1500	1500+						
43										

Palo's Import Wizard takes the input CSV file (`fulldata.csv`) generated by the Python script, and places it line-by-line in the first worksheet ("1. Create Structure"). This worksheet's formulas are recalculated for each row as Palo iterates through them, and thus detects new categories as they appear in the data, creating the appropriate attributes and nesting structure in the OLAP database.

The Import Wizard is then used again on the second worksheet ("2. Create Data"). This worksheet then adds the Sales facts to the Sales data cube.

This 2-step process is necessary, as Palo strongly recommends against combining the `PALO.EADD` and `PALO.SETDATA` formulas in the same worksheet when using the Import Wizard.

1	The Motivated Trotskyism Australia	No roof	Both	Breakfast	1	2008	Organized Hush Puppies Combo	Tyndall Wholefoods 1018	14548	345	Original Data
2	=SAS1	=SCS1	=SDS1	=SES1	=VLOOKUP(SF\$1:SD\$21:SE\$26,2)	=SG\$1	=SH\$1	=SIS1	=VLOOKUP(SIS1:SD\$29:SE\$34,2,TRUE)	=SK\$1	Conversions
3	StoreName	StoreCountry	StoreDesign	StoreFacilities	SalesPeriod	PromotionalPeriod	SalesYear	ComboName	ComboSupplier	Not Used	Not Used
4											
5	Config										
6	Database	localhost/BurgerStore									
7											
8			Dimensions (in order)				Warning: Magic				
9			Stores				=PALO.EADD(SB\$7:SD\$9,"c","All Stores")	< Make "All Stores"			
10			Designs				=PALO.EADD(SB\$7:SD\$10,"c","All Designs")	< Make "All Designs"			
11			Facilities				=PALO.EADD(SB\$7:SD\$11,"c","Any Facilities")	< Make "Any Facilities"			
12			SalesPeriods				=PALO.EADD(SB\$7:SD\$13,"c","All Periods")	< Make "All Periods"			
13			PromotionalPeriods				=PALO.EADD(SB\$7:SD\$14,"c","All Years")	< Make "All Years"			
14			Years				=PALO.EADD(SB\$7:SD\$12,"c","All Periods")	< Make "All Periods"			
15			Combos				=PALO.EADD(SB\$7:SD\$15,"c","All Combos")	< Make "All Combos"			
16			Suppliers				=PALO.EADD(SB\$7:SD\$16,"c","All Suppliers")	< Make "All Suppliers"			
17			PriceRange								
18			CalorieRange				=PALO.EADD(SB\$7:SD\$9,"n",SB\$2,"All Stores",1)	< Add country to "All Stores"			
19											
20			Month/Number				=PALO.EADD(SB\$7:SD\$9,"n",SA\$2,SB\$2,1)	< Add store to elements			
21			1	Jan/Feb			=PALO.EADD(SB\$7:SD\$10,"n",SC\$2,"All Designs",1)	< Add design to elements			
22			2	Mar/Apr			=PALO.EADD(SB\$7:SD\$11,"n",SD\$2,"Any Facilities",1)	< Add facility to elements			
23			3	May/Jun			=PALO.EADD(SB\$7:SD\$13,"n",SF\$2,"All Periods",1)	< Add month to elements			
24			4	Jul/Aug			=PALO.EADD(SB\$7:SD\$14,"n",SF\$2,"All Years",1)	< Add year to elements			
25			5	Sep/Oct			=PALO.EADD(SB\$7:SD\$12,"n",SE\$2,"All Periods",1)	< Add period to elements			
26			6	Nov/Dec			=PALO.EADD(SB\$7:SD\$15,"n",SH\$2,"All Combos",1)	< Add combo to elements			
27							=PALO.EADD(SB\$7:SD\$16,"n",SG\$2,"All Suppliers",1)	< Add supplier to elements			
28			PriceRange creation								
29			0	\$0 - \$1.99			=PALO.EADD(SB\$7:SD\$17,"c","All Prices")	< Make "All Prices"			
30			200	\$2 - \$3.99			=PALO.EADD(SB\$7:SD\$17,"n",SE\$29,"All Prices",1)	\$0 - \$1.99			
31			400	\$4 - \$5.99			=PALO.EADD(SB\$7:SD\$17,"n",SE\$30,"All Prices",1)	\$2 - \$3.99			
32			600	\$6 - \$7.99			=PALO.EADD(SB\$7:SD\$17,"n",SE\$31,"All Prices",1)	\$4 - \$5.99			
33			800	\$8 - \$9.99			=PALO.EADD(SB\$7:SD\$17,"n",SE\$32,"All Prices",1)	\$6 - \$7.99			
34			1000	\$10+			=PALO.EADD(SB\$7:SD\$17,"n",SE\$33,"All Prices",1)	\$8 - \$9.99			
35							=PALO.EADD(SB\$7:SD\$17,"n",SE\$34,"All Prices",1)	\$10+			
36			CalorieRange creation								
37			0	0 - 249			=PALO.EADD(SB\$7:SD\$18,"c","Any Calories")	< Make "Any Calories"			
38			250	250 - 499			=PALO.EADD(SB\$7:SD\$18,"n",SE\$37,"Any Calories",1)	0 - 249			
39			500	500 - 749			=PALO.EADD(SB\$7:SD\$18,"n",SE\$38,"Any Calories",1)	250 - 499			
40			750	750 - 999			=PALO.EADD(SB\$7:SD\$18,"n",SE\$39,"Any Calories",1)	500 - 749			
41			1000	1000 - 1249			=PALO.EADD(SB\$7:SD\$18,"n",SE\$40,"Any Calories",1)	750 - 999			
42			1250	1250 - 1499			=PALO.EADD(SB\$7:SD\$18,"n",SE\$41,"Any Calories",1)	1000 - 1249			
43			1500	1500+			=PALO.EADD(SB\$7:SD\$18,"n",SE\$42,"Any Calories",1)	1250 - 1499			
44							=PALO.EADD(SB\$7:SD\$18,"n",SE\$43,"Any Calories",1)	1500+			

## Interesting Scenarios

Scenarios were chosen for their perceived relevance and interest to Healthy Burgers management for their goals of expanding the Healthy Burgers franchise into other countries, and to replicate the success of existing stores and products where possible, to increase sales and popularity.

Accompanying screenshots for each scenario may be found in the Appendices. The reader

should note that although these images are technically not full screen, they do include enough details to verify they are indeed legitimate outputs of the Palo server/Excel plugin. Electronic copies should also be available with this submission.

### **Total sales in dollars per year by facility type**

It is of interest to Healthy Burgers to know the facility types positively correlated with large amounts of sales to aid in the decision of what facilities to offer in new restaurants. It is also useful for deciding whether to expand existing restaurants to offer new facilities, or whether to close existing facilities because they do not warrant the additional cost of running them.

### **Total sales in dollars per year by interior design**

Similar to facility types, the correlation between interior designs and total sales can be used for deciding the layouts of new restaurants or whether or not to renovate existing branches to be consistent with the more profitable operations - if this is indeed deemed an important factor in their success.

### **Total sales in dollars per category of calories by price bracket**

If high calorie meals are indeed more popular - and the data should be consulted to confirm this suspicion - then it is of interest to Healthy Burgers to ascertain how much people are willing to pay per calorie when it comes in the form of the different meals offered in the combos. This can be used to anticipate the success and pricing of future meals offered at both existing and future restaurants, to maximise profits.

### **Total sales in dollars per products supplied by suppliers by year**

Essentially, this offers an indication of how much profit Healthy Burgers is making from the products whose ingredients are provided by each supplier. This will reveal any correlations or trends regarding consumer preference for the ingredients of one supplier over another. This may also be used to help shape the future pricing of products, identifying and meals or combos that are currently over or under priced.

### **Total sales in dollars per sales period by country**

Total sales per country is of obvious utility for expanding the franchises stores. Countries may be identified as having significantly large sales and prove a sound choice for future restaurants. Alternatively, it may be decided that these countries are saturated in terms of similar restaurants and future expansion should take place elsewhere.

## **References**

- [1] CITS3401 Data Exploration and Mining - Project 1, <http://undergraduate.csse.uwa.edu.au/units/CITS3401/labs/proj1-2013.html>

## A Total sales in dollars per year by facility type

The screenshot shows a Microsoft Excel spreadsheet titled 'snapshot.xlsx'. The ribbon is set to 'Home', and the 'Font' and 'Alignment' groups are visible. The active cell is D19. The spreadsheet contains a pivot table with the following data:

	A	B	C	D	E	F	G	H
1	localhost/BurgerStore							
2	Sales							
3	All Stores							
4	All Designs							
5	All Periods							
6	All Periods							
7	All Combos							
8	All Suppliers							
9	All Prices							
10	Any Calories							
11								
12		Dine-in	Drive-through	Both				
13	2008	3,278,289	3,260,388	2,097,224				
14	2009	3,379,143	3,319,011	2,194,353				
15	2010	3,362,891	3,278,076	2,213,186				
16	2011	3,331,692	3,329,445	2,204,859				
17	2012	563,618	559,616	371,054				
18								
19								
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								

The bottom of the spreadsheet shows several worksheet tabs: 'Facilities vs Years', 'Designs vs Years', 'Price vs Calories', 'Suppliers vs Years', and 'Country vs Sales Periods'.

## B Total sales in dollars per year by interior design

snapshot.xlsx - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View Team

Cut Copy Paste Format Painter Clipboard

Arial 10 Bold Italic Underline Font Color Background Color

Wrap Text Merge & Center Alignment Number

G10 fx

	A	B	C	D	E	F	G	H
1	localhost/BurgerStore							
2	Sales							
3	All Stores							
4	Any Facilities							
5	All Periods							
6	All Periods							
7	All Combos							
8	All Suppliers							
9	All Prices							
10	Any Calories							
11								
12		Avant-garde	No roof	Upside-down				
13	2008	3,875,214	3,639,131	1,121,556				
14	2009	3,959,781	3,794,831	1,137,895				
15	2010	3,906,224	3,814,124	1,133,805				
16	2011	3,884,370	3,858,049	1,123,577				
17	2012	656,542	645,577	192,169				
18								
19								
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								

Facilities vs Years Designs vs Years Price vs Calories Suppliers vs Years Country vs Sales Periods

## C Total sales in dollars per category of calories by price bracket

The screenshot shows an Excel spreadsheet with a pivot table. The pivot table is located in the range A1:H19. The data is organized as follows:

	A	B	C	D	E	F	G	H
1	localhost/BurgerStore							
2	Sales							
3	All Stores							
4	All Designs							
5	Any Facilities							
6	All Periods							
7	All Periods							
8	All Years							
9	All Combos							
10	All Suppliers							
11								
12		\$0 - \$3.32	\$3.33 - \$6.65	\$6.66+				
13	0 - 249	39,659	502,404	655,141				
14	250 - 499	471,914	2,225,416	3,194,643				
15	500 - 749	693,547	3,133,056	3,216,969				
16	750 - 999	1,085,811	4,122,365	3,577,612				
17	1000 - 1249	1,321,554	3,293,541	2,999,576				
18	1250 - 1499	792,149	2,224,614	3,192,874				
19	1500+	0	0	0				
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								

The pivot table is titled "localhost/BurgerStore" and is located in the range A1:H19. The data is organized by "Sales" (rows 2-10) and "All Suppliers" (rows 11-19). The columns represent price brackets: "\$0 - \$3.32", "\$3.33 - \$6.65", and "\$6.66+". The values represent total sales in dollars.



## D Total sales in dollars per products supplied by suppliers by year

snapshot.xlsx - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View Team

Cut Copy Paste Format Painter Clipboard

Arial 10 Bold Italic Underline Font Color Background Color

Wrap Text Merge & Center Alignment

General Number \$ % , .00

C34

	A	B	C	D
1	localhost/BurgerStore			
2	Sales			
3	All Stores			
4	All Designs			
5	Any Facilities			
6	All Periods			
7	All Periods			
8	All Combos			
9	All Prices			
10	Any Calories			
11				
12		Greenham Suppliers Incorporated	Tyndall Wholefoods and Logistics Pty Ltd	
13	2008	4,733,528	3,902,373	
14	2009	3,901,927	4,990,580	
15	2010	4,295,970	4,558,183	
16	2011	3,784,274	5,081,722	
17	2012	803,812	690,476	
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				

Facilities vs Years Designs vs Years Price vs Calories Suppliers vs Years Country vs Sales Periods

## E Total sales in dollars per sales period by country

The screenshot shows a Microsoft Excel spreadsheet titled 'snapshot.xlsx'. The 'Home' tab is active, displaying the ribbon with options for File, Home, Insert, Page Layout, Formulas, Data, Review, View, and Team. The ribbon includes sections for Clipboard, Font, Alignment, and Number. The spreadsheet data is as follows:

	A	B	C	D	E	F	G	H
1	localhost/BurgerStore							
2	Sales							
3	All Designs							
4	Any Facilities							
5	All Periods							
6	All Years							
7	All Combos							
8	All Suppliers							
9	All Prices							
10	Any Calories							
11								
12		Australia	New Zealand	Singapore				
13	Breakfast	2,072,646	4,912,840	3,695,538				
14	Lunch	6,497,830	5,978,464	1,675,544				
15	Dinner	5,020,201	2,027,055	4,862,727				
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								
30								

The bottom of the spreadsheet shows a navigation bar with tabs: Facilities vs Years, Designs vs Years, Price vs Calories, Suppliers vs Years, and Country vs Sales Periods. The 'Country vs Sales Periods' tab is currently selected.