Sentiment Analysis for Dualcore Products

Beth Hilbert ~ Spring 2018

Business Question

What changes should be made to Dualcore's product line after considering customer product reviews?

Solution: Sentiment Analysis

- Turn words into numbers
- Match with word sentiment dictionary
- Calculate opinion measures

$$\text{Sentiment} = \frac{\#\text{positive} - \#\text{negative}}{\text{total count}}$$

range of -1 to 1

$$\text{Positivity} = \frac{\#\text{positive}}{\text{total count}} \text{ x } 100$$

range of 0 to 100%

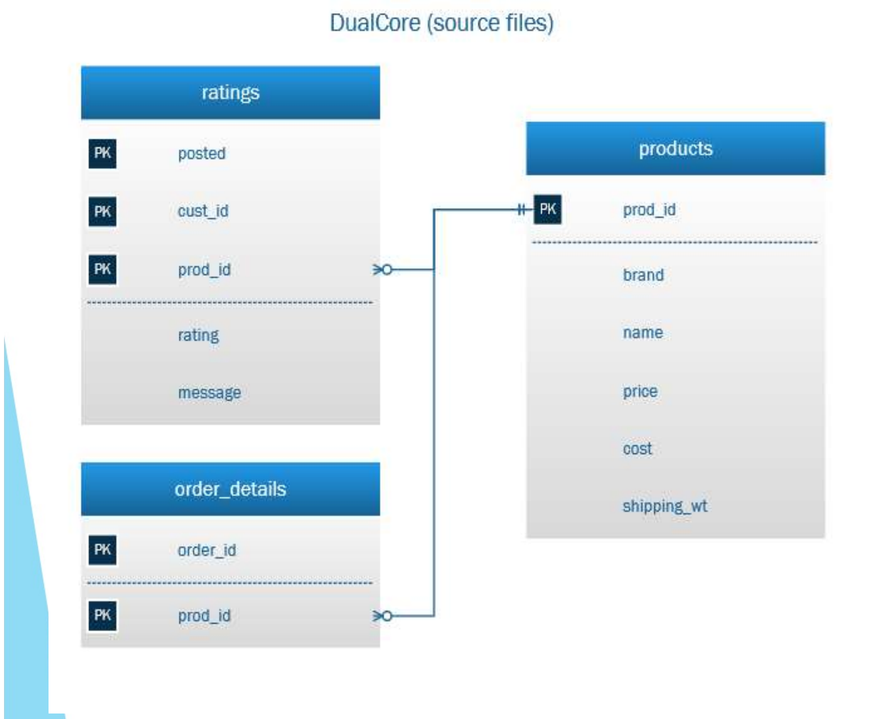
Overview: Tools

- Hadoop – store and process in parallel of key-value data
  - Scoop ingest
  - HDFS store chunked data
  - MapReduce, Hive, Impala process

- Microsoft Access – complex queries of dimensional data

Overview: Data Sources

Products, Order_Details and Ratings from Dualcore



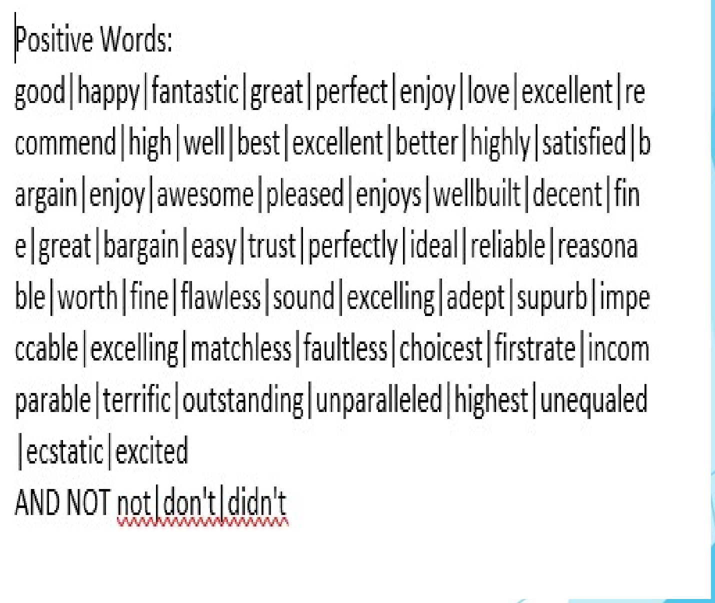
DualCore (source files)

ratings
PK posted
PK cust_id
PK prod_id
rating
message

order_details
PK order_id
PK prod_id

products
PK prod_id
brand
name
price
cost
shipping_wt

Word Dictionaries from Stanford and Cloudera

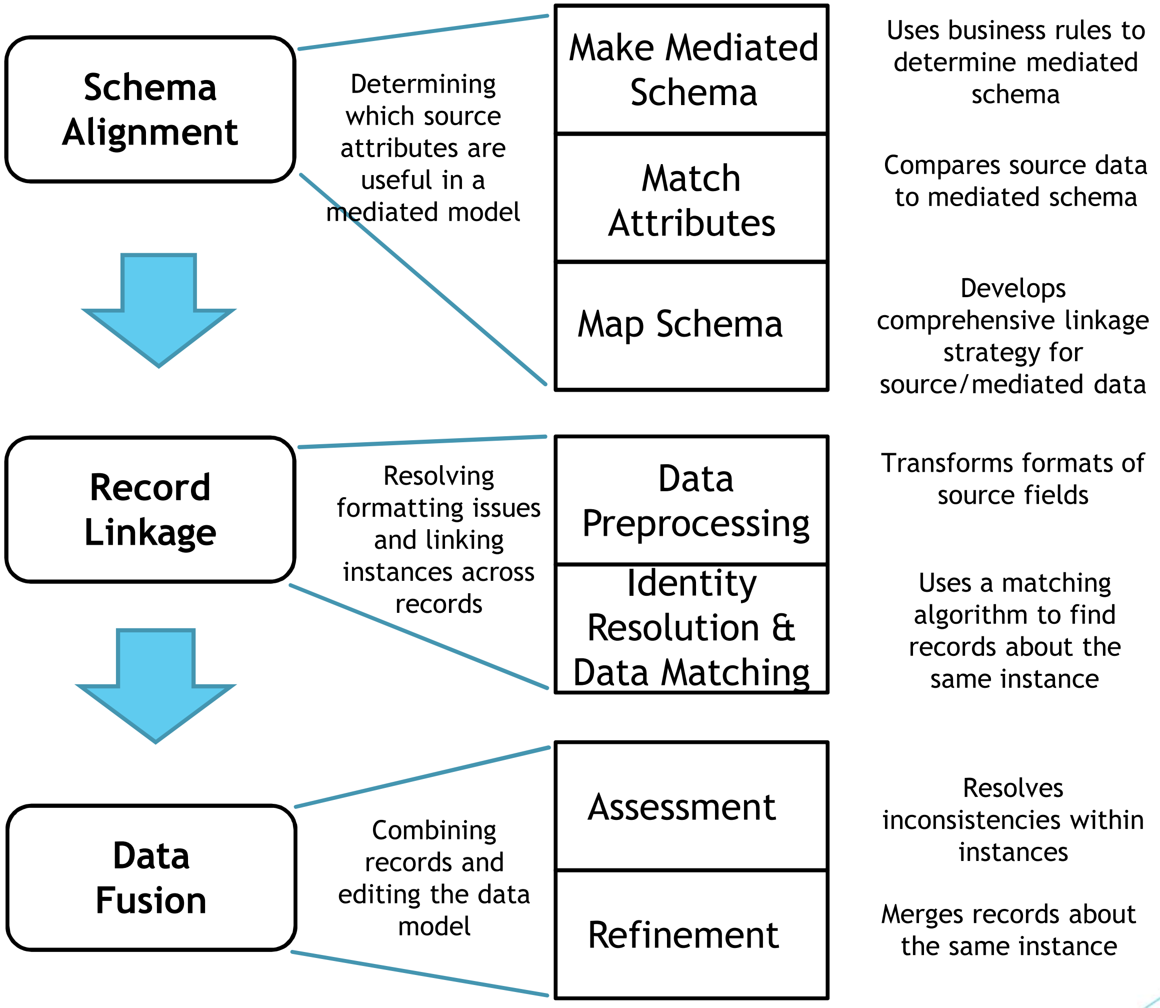


Positive Words:

good|happy|fantastic|great|perfect|enjoy|love|excellent|recommend|high|well|best|excellent|better|highly|satisfied|bargain|enjoy|awesome|pleased|enjoys|wellbuilt|decent|fine|great|bargain|easy|trust|perfectly|ideal|reliable|reasonable|worth|fine|flawless|sound|excelling|adept|supurb|impeccable|excelling|matchless|faultless|choicest|firstrate|incomparable|terrific|outstanding|unparalleled|highest|unequaled|ecstatic|excited

AND NOT not|don't|didn't

Data Integration Process



Schema Alignment

Determining which source attributes are useful in a mediated model

Make Mediated Schema

Uses business rules to determine mediated schema

Match Attributes

Compares source data to mediated schema

Map Schema

Develops comprehensive linkage strategy for source/mediated data

Record Linkage

Resolving formatting issues and linking instances across records

Data Preprocessing

Transforms formats of source fields

Identity Resolution & Data Matching

Uses a matching algorithm to find records about the same instance

Data Fusion

Combining records and editing the data model

Assessment

Resolves inconsistencies within instances

Refinement

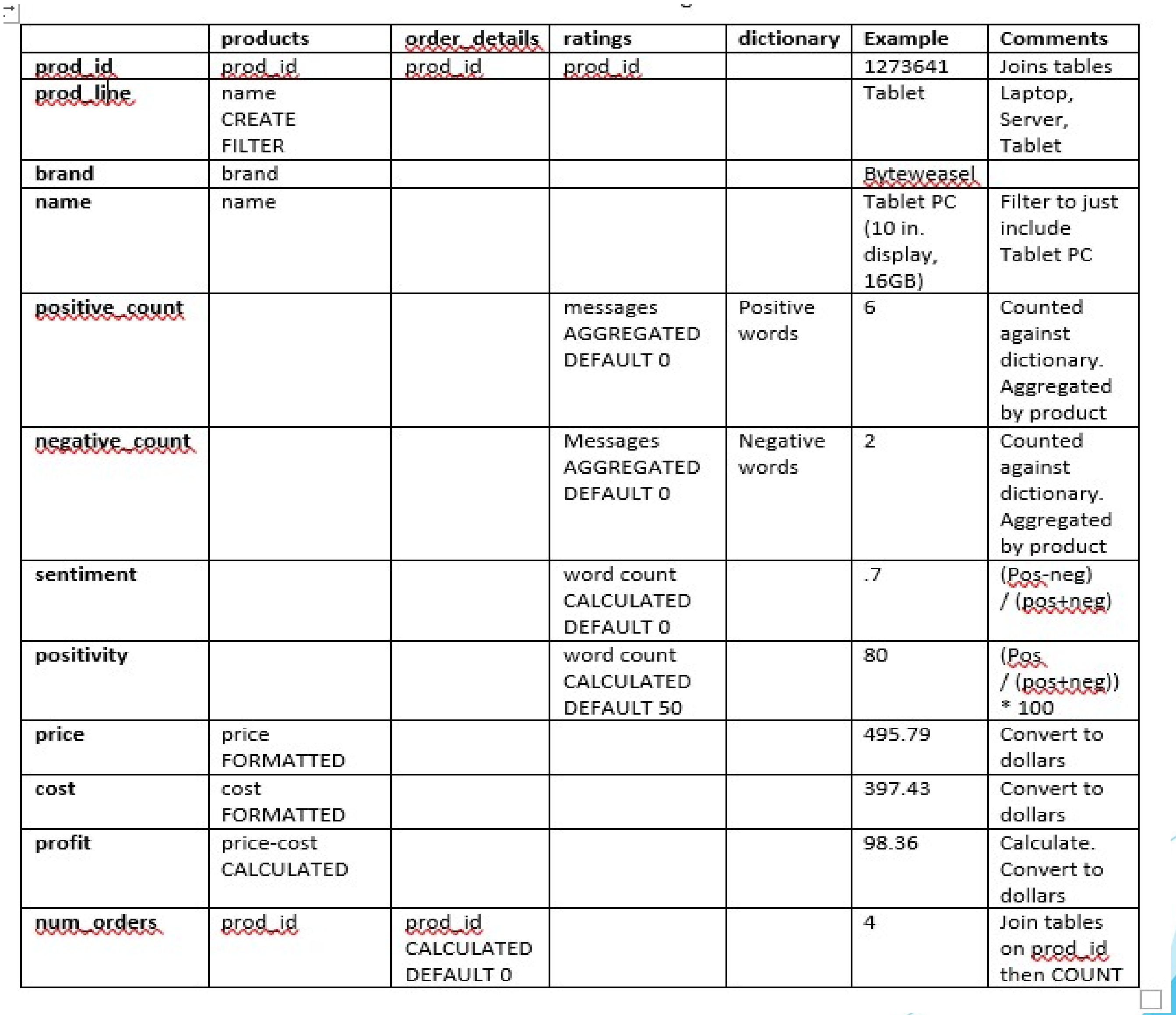
Merges records about the same instance

Data Integration Step 1: Schema Alignment

- Determine grain to answer business question (product)
- Determine attributes
- Locate data sources
- Map attributes from data sources

- Issues encountered:
  - Determine a grain which reflects magnitude of reviews
  - Dictionary (Stanford/movies, Cloudera/software version)

Mediated Schema Map



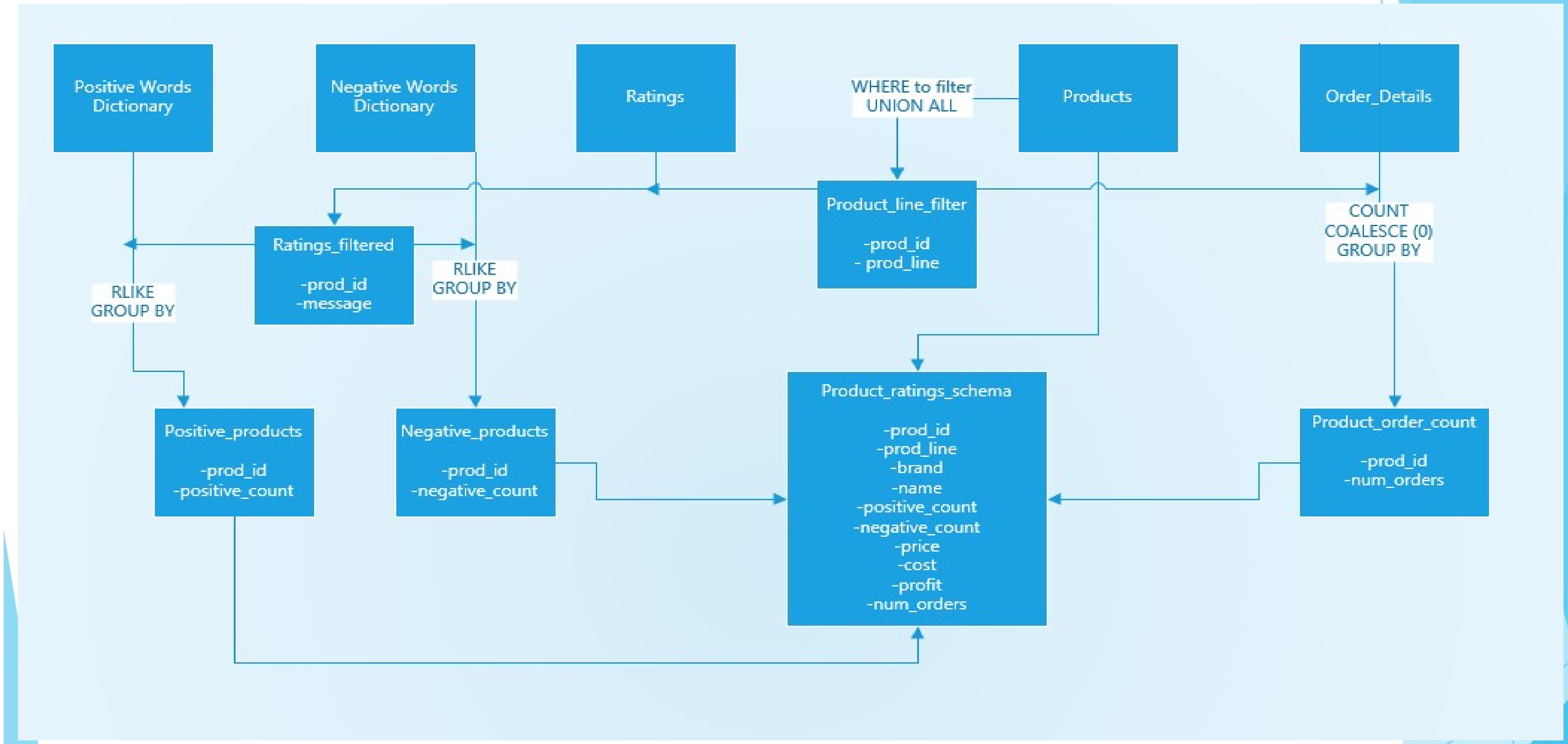
	products	order_details	ratings	dictionary	Example	Comments
prod_id	prod_id	prod_id	prod_id		1273641	Joins tables
prod_line	name CREATE FILTER				Tablet	Laptop, Server, Tablet
brand	brand				Byteweasel	
name	name				Tablet PC (10 in. display, 16GB)	Filter to just include Tablet PC
positive_count			messages AGGREGATED DEFAULT 0	Positive words	6	Counted against dictionary. Aggregated by product
negative_count			Messages AGGREGATED DEFAULT 0	Negative words	2	Counted against dictionary. Aggregated by product
sentiment			word count CALCULATED DEFAULT 0		.7	(Pos-neg) / (pos+neg)
positivity			word count CALCULATED DEFAULT 50		80	(Pos / (pos+neg)) * 100
price	price FORMATTED				495.79	Convert to dollars
cost	cost FORMATTED				397.43	Convert to dollars
profit	price-cost CALCULATED				98.36	Calculate. Convert to dollars
num_orders	prod_id	prod_id CALCULATED DEFAULT 0			4	Join tables on prod_id then COUNT

Data Integration Step 2: Record Linkage

- Formatting source attributes
  - prod_line, price, cost, profit, messages
- Transitioning rows to final grain
  - positive_count, negative_count, num_orders

- Issues encountered:
  - Scope/filter: tablets → tablet, server, laptop lines
  - Implement dictionary format
  - Attribute complexity with too much information
  - Aggregation, counts required intermediate tables

Intermediate Tables



Positive Words Dictionary

Negative Words Dictionary

Ratings

WHERE to filter
UNION ALL

Products

Order_Details

Product_line_filter

-prod_id
- prod_line

COUNT
COALESCE (0)
GROUP BY

Ratings_filtered

-prod_id
-message

RLIKE
GROUP BY

RLIKE
GROUP BY

Positive_products

-prod_id
-positive_count

Negative_products

-prod_id
-negative_count

Product_ratings_schema

-prod_id
-prod_line
-brand
-name
-positive_count
-negative_count
-price
-cost
-profit
-num_orders

Product_order_count

-prod_id
-num_orders

- Further manipulation within Access:
  - Calculate sentiment and positivity measurements
  - Group by prod_line, brand, counts

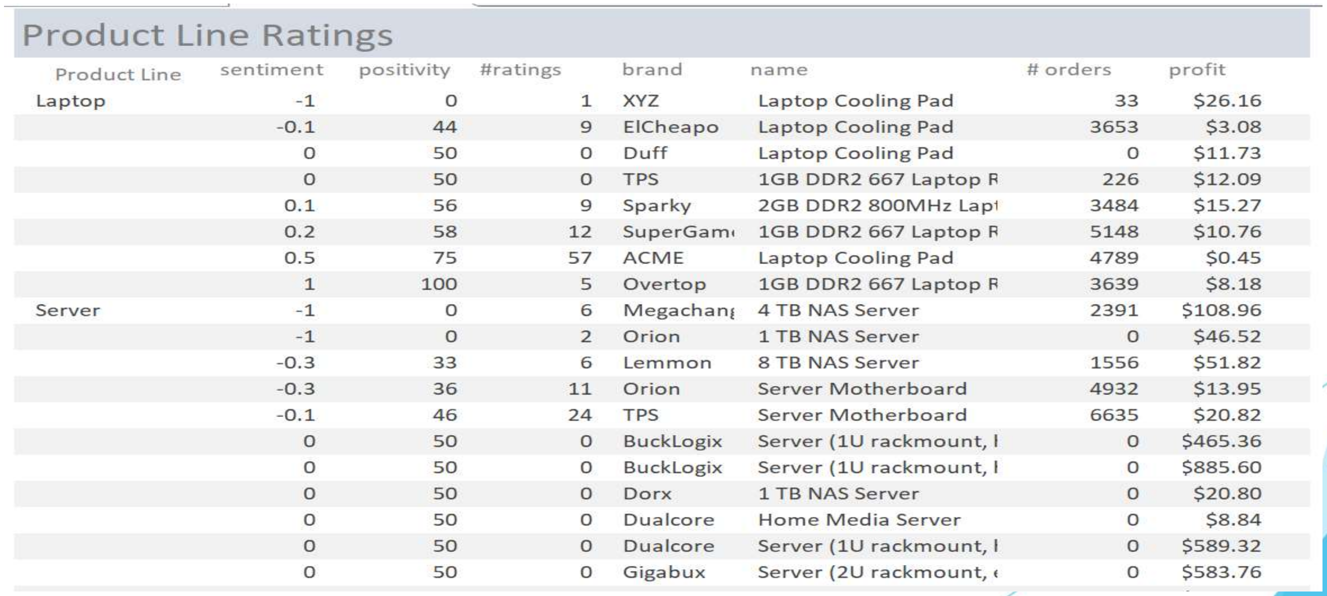
Data Integration Step 3: Data Fusion

- Consider incongruencies
- Identifying true values

- Issues encountered:
  - Handing Nulls: counts to 0, opinions to neutral (0 or 50).
  - Implementing Nulls: COALESCE and IF
  - Limitations of complex joins and aggregations (dimensional)

Business Insights

- Analyzed by Product Line
  - Laptop – only Acme cooling pads, highest laptop only neutral
  - Server -Megachamp Lemmon have high sales but low ratings. Olde-Grey highly rated. All motherboards poorly rated
  - Tablet – drop Orion and United Digital. Byteweasel sells the most and has highest positivity, but we are loosing $1 per sale.

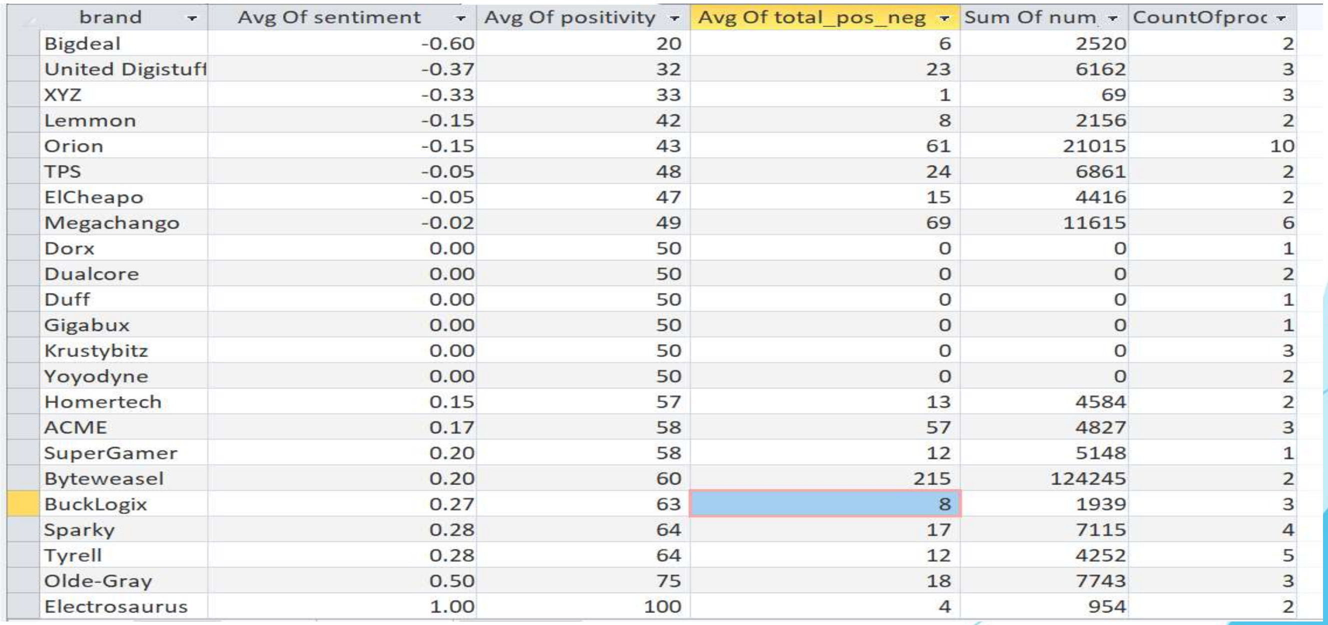


Product Line Ratings

Product Line	sentiment	positivity	#ratings	brand	name	# orders	profit
Laptop	-1	0	1	XYZ	Laptop Cooling Pad	33	$26.16
	-0.1	44	9	ElCheapo	Laptop Cooling Pad	3653	$3.08
	0	50	0	Duff	Laptop Cooling Pad	0	$11.73
	0	50	0	TPS	1GB DDR2 667 Laptop R	226	$12.09
	0.1	56	9	Sparky	2GB DDR2 800MHz Lapt	3484	$15.27
	0.2	58	12	SuperGam	1GB DDR2 667 Laptop R	5148	$10.76
	0.5	75	57	ACME	Laptop Cooling Pad	4789	$0.45
	1	100	5	Overtop	1GB DDR2 667 Laptop R	3639	$8.18
Server	-1	0	6	Megachan	4 TB NAS Server	2391	$108.96
	-1	0	2	Orion	1 TB NAS Server	0	$46.52
	-0.3	33	6	Lemmon	8 TB NAS Server	1556	$51.82
	-0.3	36	11	Orion	Server Motherboard	4932	$13.95
	-0.1	46	24	TPS	Server Motherboard	6635	$20.82
	0	50	0	BuckLogix	Server (1U rackmount, I	0	$465.36
	0	50	0	BuckLogix	Server (1U rackmount, I	0	$885.60
	0	50	0	Dorx	1 TB NAS Server	0	$20.80
	0	50	0	Dualcore	Home Media Server	0	$8.84
	0	50	0	Dualcore	Server (1U rackmount, I	0	$589.32
	0	50	0	Gigabux	Server (2U rackmount,	0	$583.76

Business Insights

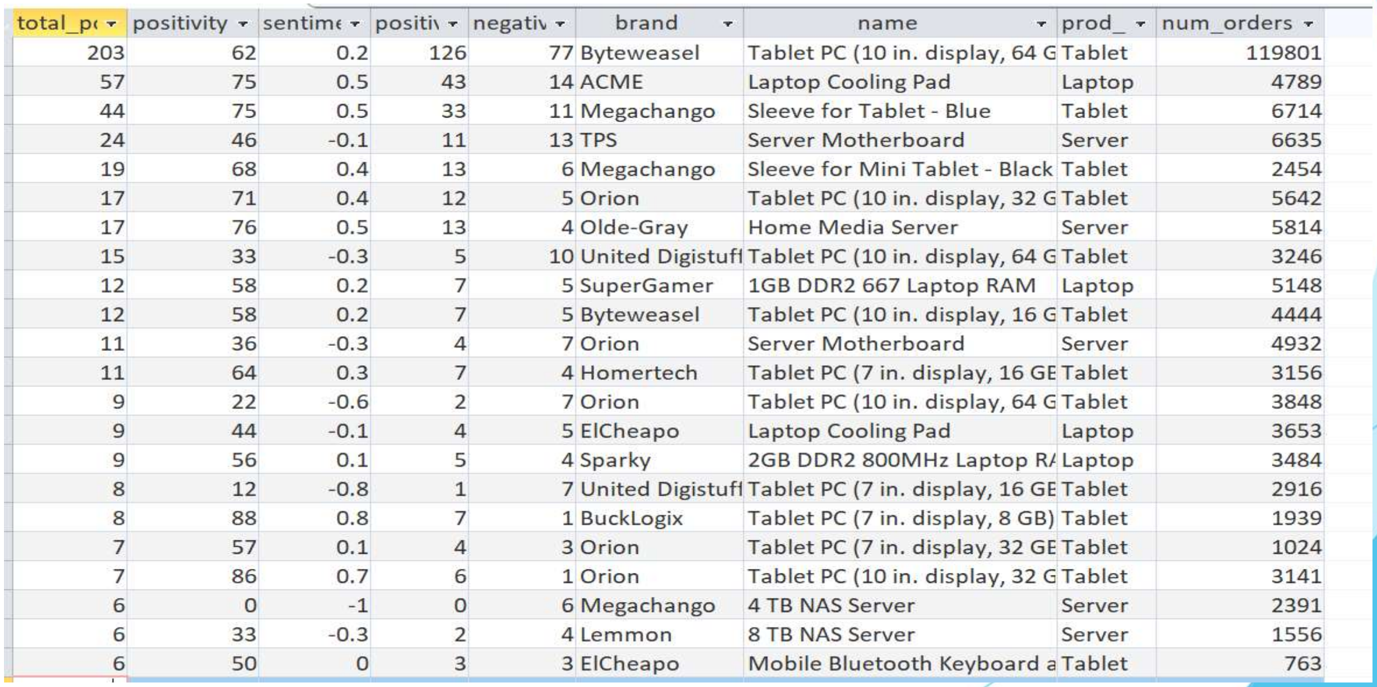
- Analyzed by Brand
  - 5 brands elicit strong negative opinions
  - Bigdeal – 2 products total, brand sentiment of -0.6, high sales
  - United Digistuff – 3 products, brand sentiment of -0.37, high sales



brand	Avg Of sentiment	Avg Of positivity	Avg Of total_pos_neg	Sum Of num	CountOfproc
Bigdeal	-0.60	20	6	2520	2
United Digistuff	-0.37	32	23	6162	3
XYZ	-0.33	33	1	69	3
Lemmon	-0.15	42	8	2156	2
Orion	-0.15	43	61	21015	10
TPS	-0.05	48	24	6861	2
ElCheapo	-0.05	47	15	4416	2
Megachango	-0.02	49	69	11615	6
Dorx	0.00	50	0	0	1
Dualcore	0.00	50	0	0	2
Duff	0.00	50	0	0	1
Gigabux	0.00	50	0	0	1
Krustybitz	0.00	50	0	0	3
Yoyodyne	0.00	50	0	0	2
Homertech	0.15	57	13	4584	2
ACME	0.17	58	57	4827	3
SuperGamer	0.20	58	12	5148	1
Byteweasel	0.20	60	215	124245	2
BuckLogix	0.27	63	8	1939	3
Sparky	0.28	64	17	7115	4
Tyrell	0.28	64	12	4252	5
Olde-Gray	0.50	75	18	7743	3
Electrosaurus	1.00	100	4	954	2

Business Insights

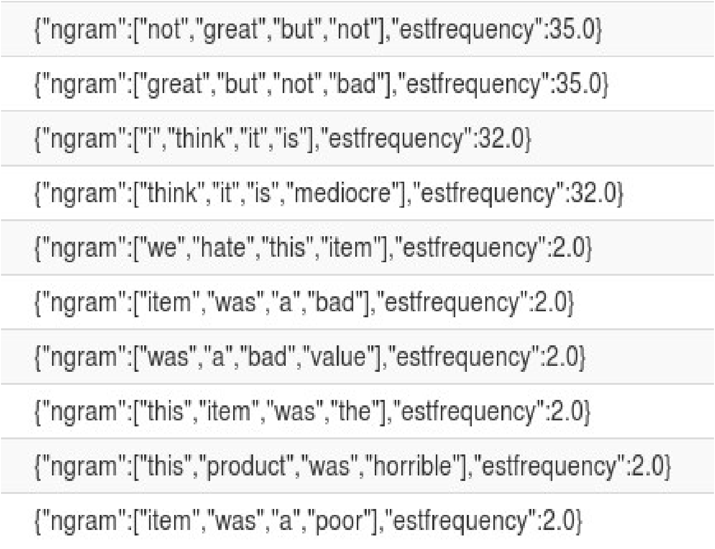
- Analyzed by Counts
  - Top 7 counts, all but one had better than average reviews
  - Motherboards have high number of reviews, and they are negative



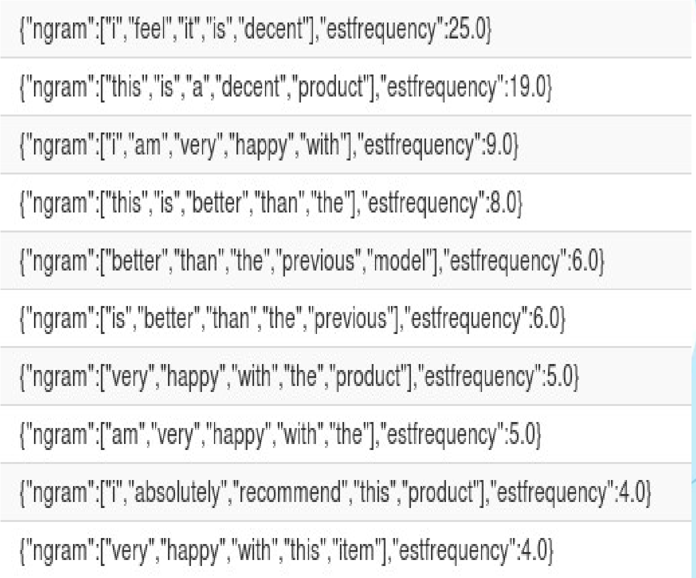
total_p	positivity	sentime	positi	negativ	brand	name	prod_	num_orders
203	62	0.2	126	77	Byteweasel	Tablet PC (10 in. display, 64 G	Tablet	119801
57	75	0.5	43	14	ACME	Laptop Cooling Pad	Laptop	4789
44	75	0.5	33	11	Megachango	Sleeve for Tablet - Blue	Tablet	6714
24	46	-0.1	11	13	TPS	Server Motherboard	Server	6635
19	68	0.4	13	6	Megachango	Sleeve for Mini Tablet - Black	Tablet	2454
17	71	0.4	12	5	Orion	Tablet PC (10 in. display, 32 G	Tablet	5642
17	76	0.5	13	4	Olde-Gray	Home Media Server	Server	5814
15	33	-0.3	5	10	United Digistuff	Tablet PC (10 in. display, 64 G	Tablet	3246
12	58	0.2	7	5	SuperGamer	1GB DDR2 667 Laptop RAM	Laptop	5148
12	58	0.2	7	5	Byteweasel	Tablet PC (10 in. display, 16 G	Tablet	4444
11	36	-0.3	4	7	Orion	Server Motherboard	Server	4932
11	64	0.3	7	4	Homertech	Tablet PC (7 in. display, 16 GE	Tablet	3156
9	22	-0.6	2	7	Orion	Tablet PC (10 in. display, 64 G	Tablet	3848
9	44	-0.1	4	5	ElCheapo	Laptop Cooling Pad	Laptop	3653
9	56	0.1	5	4	Sparky	2GB DDR2 800MHz Laptop R/	Laptop	3484
8	12	-0.8	1	7	United Digistuff	Tablet PC (7 in. display, 16 GE	Tablet	2916
8	88	0.8	7	1	BuckLogix	Tablet PC (7 in. display, 8 GB)	Tablet	1939
7	57	0.1	4	3	Orion	Tablet PC (7 in. display, 32 GE	Tablet	1024
7	86	0.7	6	1	Orion	Tablet PC (10 in. display, 32 G	Tablet	3141
6	0	-1	0	6	Megachango	4 TB NAS Server	Server	2391
6	33	-0.3	2	4	Lemmon	8 TB NAS Server	Server	1556
6	50	0	3	3	ElCheapo	Mobile Bluetooth Keyboard a	Tablet	763

Business Insights

- Negative NGRAMS
  - mediocre, not great, 30+
  - bad, horrible, poor, 2x
- Positive NGRAMS
  - decent, 25x
  - happy, better than



```
{"ngram":["not","great","but","not"],"estfrequency":35.0}
{"ngram":["great","but","not","bad"],"estfrequency":35.0}
{"ngram":["i","think","it","is"],"estfrequency":32.0}
{"ngram":["think","it","is","mediocre"],"estfrequency":32.0}
{"ngram":["we","hate","this","item"],"estfrequency":2.0}
{"ngram":["item","was","a","bad"],"estfrequency":2.0}
{"ngram":["was","a","bad","value"],"estfrequency":2.0}
{"ngram":["this","item","was","the"],"estfrequency":2.0}
{"ngram":["this","product","was","horrible"],"estfrequency":2.0}
{"ngram":["item","was","a","poor"],"estfrequency":2.0}
```



```
{"ngram":["i","feel","it","is","decent"],"estfrequency":25.0}
{"ngram":["this","is","a","decent","product"],"estfrequency":19.0}
{"ngram":["i","am","very","happy","with"],"estfrequency":9.0}
{"ngram":["this","is","better","than","the"],"estfrequency":8.0}
{"ngram":["better","than","the","previous","model"],"estfrequency":6.0}
{"ngram":["is","better","than","the","previous"],"estfrequency":6.0}
{"ngram":["very","happy","with","the","product"],"estfrequency":5.0}
{"ngram":["am","very","happy","with","the"],"estfrequency":5.0}
{"ngram":["i","absolutely","recommend","this","product"],"estfrequency":4.0}
{"ngram":["very","happy","with","this","item"],"estfrequency":4.0}
```

Future Steps …

- Sentiment of phrases
- Scale of positive/negative
- Further train dictionary

Questions?