Analytics-R'-Us

Schema Integration and Justification Team

Demand Prediction Analysis

DSE 203 Presentation #2

10/28/2017

Team:

Josh Wilson

Amisha Bhanage

Ken Kroel

Mai Huynh

Integrated Schema Provides

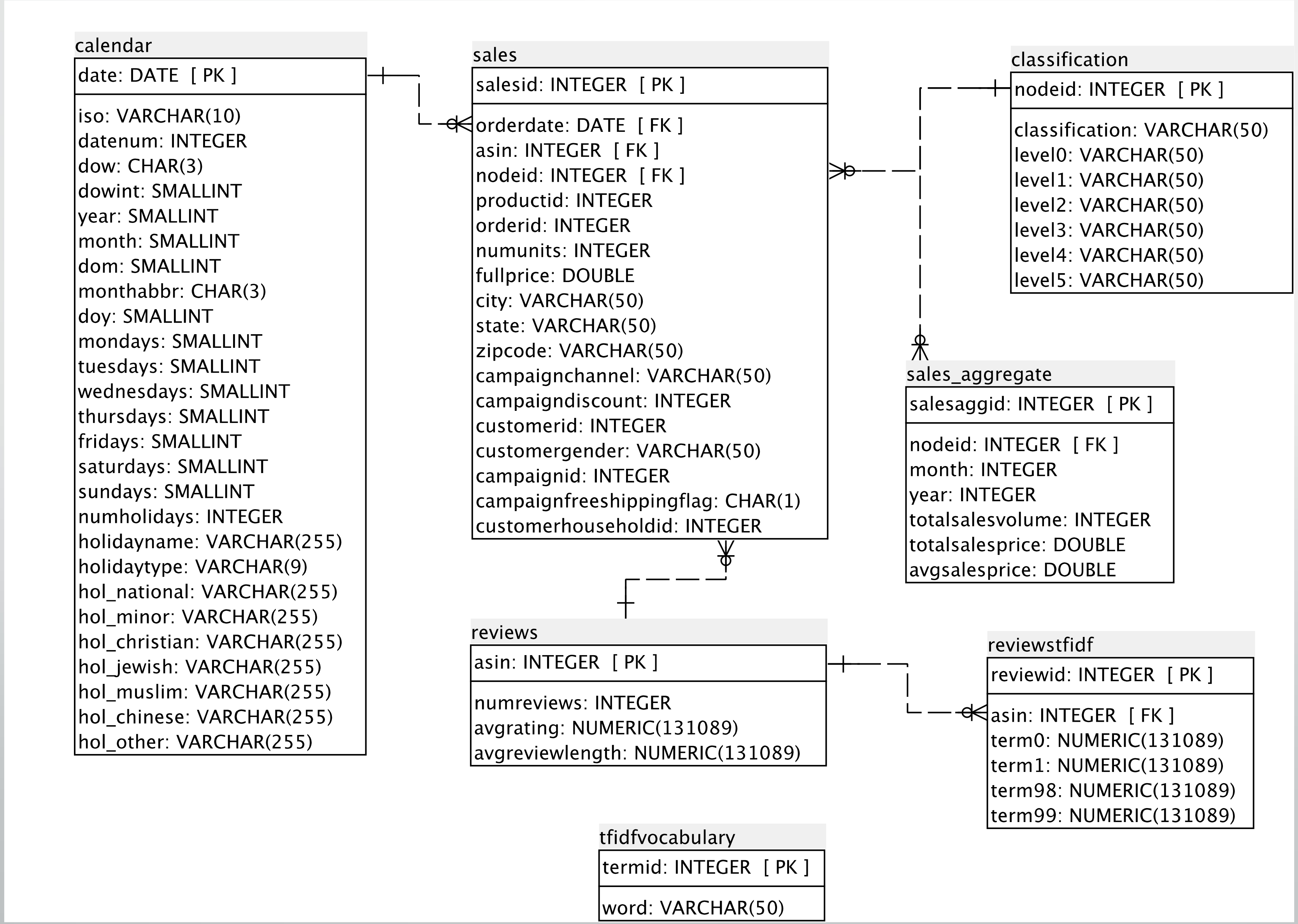
- Sales Information

Product ID (Books)	Unit Price
Product Classification (flattened)	Average Price
Order ID	Number of Units
Customer ID	Campaign ID
Customer Gender	Campaign Discounts
Period (Day, Week, Month, Year, Holidays)	Campaign Channels
Geography (State, City, Zip)	Free Shipping Flag

- Book Review Information

Review Text TF-IDF of 100 word	Number of Reviews
Average Rating	Average Review Length

Virtualized Integrated Schema



calendar

date: DATE [PK]

iso: VARCHAR(10)
datenum: INTEGER
dow: CHAR(3)
dowint: SMALLINT
year: SMALLINT
month: SMALLINT
dom: SMALLINT
monthabbr: CHAR(3)
doy: SMALLINT
mondays: SMALLINT
tuesdays: SMALLINT
wednesdays: SMALLINT
thursdays: SMALLINT
fridays: SMALLINT
saturdays: SMALLINT
sundays: SMALLINT
numholidays: INTEGER
holidayname: VARCHAR(255)
holidaytype: VARCHAR(9)
hol_national: VARCHAR(255)
hol_minor: VARCHAR(255)
hol_christian: VARCHAR(255)
hol_jewish: VARCHAR(255)
hol_muslim: VARCHAR(255)
hol_chinese: VARCHAR(255)
hol_other: VARCHAR(255)

sales

salesid: INTEGER [PK]

orderdate: DATE [FK]
asin: INTEGER [FK]
nodeid: INTEGER [FK]
productid: INTEGER
orderid: INTEGER
numunits: INTEGER
fullprice: DOUBLE
city: VARCHAR(50)
state: VARCHAR(50)
zipcode: VARCHAR(50)
campaignchannel: VARCHAR(50)
campaigndiscount: INTEGER
customerid: INTEGER
customergender: VARCHAR(50)
campaignid: INTEGER
campaignfreeshippingflag: CHAR(1)
customerhouseholdid: INTEGER

classification

nodeid: INTEGER [PK]

classification: VARCHAR(50)
level0: VARCHAR(50)
level1: VARCHAR(50)
level2: VARCHAR(50)
level3: VARCHAR(50)
level4: VARCHAR(50)
level5: VARCHAR(50)

sales_aggregate

salesaggid: INTEGER [PK]

nodeid: INTEGER [FK]
month: INTEGER
year: INTEGER
totalsalesvolume: INTEGER
totalsalesprice: DOUBLE
avgsalesprice: DOUBLE

reviews

asin: INTEGER [PK]

numreviews: INTEGER
avgrating: NUMERIC(131089)
avgreviewlength: NUMERIC(131089)

reviewstfidf

reviewid: INTEGER [PK]

asin: INTEGER [FK]
term0: NUMERIC(131089)
term1: NUMERIC(131089)
term98: NUMERIC(131089)
term99: NUMERIC(131089)

tfidfvocabulary

termid: INTEGER [PK]

word: VARCHAR(50)

Virtualized Integrated Schema Mapping

- Sales → PostgreSQL
- Calendar → PostgreSQL
- Classification → flattened info from JSON / AsterixDB
- Reviews → ASIN/product level summary from JSON / AsterixDB
- ReviewsTfIdf → TF-IDF info for top 100 vocabulary terms for each review
- TfIdfVocabulary → maps each vocabulary word to its numeric term
- Sales_aggregate → summary of sales volume and price by classification, month, and year

- Full mappings between integrated schema and source data available on Github

Example of Query Decomposition

- Query:
  - Total sales volume for “Computer Programming” in in July-2013
- Source data:
  - Sales (orderdate, nodeid, numunits, orderdate) ⟵ PostgreSQL
  - Classification (nodeid, classification) ⟵ JSON from AsterixDB
- Query breakdown

```
SELECT agg.nodeid, agg.MONTH, YEAR, agg.total_sales_volume
FROM sales_by_month agg INNER JOIN classification cls
ON      agg.nodeid      = cls.nodid
WHERE  MONTH       = 7 AND YEAR          = 2013
AND classification = “for computer programming”
```

Additional Query Ideas for Stakeholders

1. What is the average rating for each book?
2. What is the average rating of books within each category?
3. What is the variance in ratings within each category?
4. What is the geographic distribution of orders for each category?
5. What is the variance in number of orders by book in each category? (i.e. do some books within a category sell more than others?)
6. What is the variance in number of orders in each category by month/quarter? (i.e. is demand seasonal?)
7. What percentage of sales in each category are associated with a campaign? With each specific type of campaign?
8. How correlated is demand with number of reviews, average review rating, discount percentage, days until holiday season?
9. Are specific customers more influenced by certain types of campaigns?
10. Are products from the same category more likely to be bought in the same order?

Q & A

Thank you!