Analytics-R'-Us

Schema Integration and Justification Team

Demand Prediction Analysis

DSE 203 Presentation #3

11/8/2017

Team:

Josh Wilson

Amisha Bhanage

Ken Kroel

Mai Huynh

Specific Stakeholder Queries Addressed

1. What are the top 3 categories of books that are most read around Christmas?
2. What time of the year are the sales of "Education" books the highest?
3. Given month m and category c, predict the amount of sales for the category.
4. Which book categories show a downward trend in demand in Winter and Spring?
5. Is there a category that we should discontinue stocking?

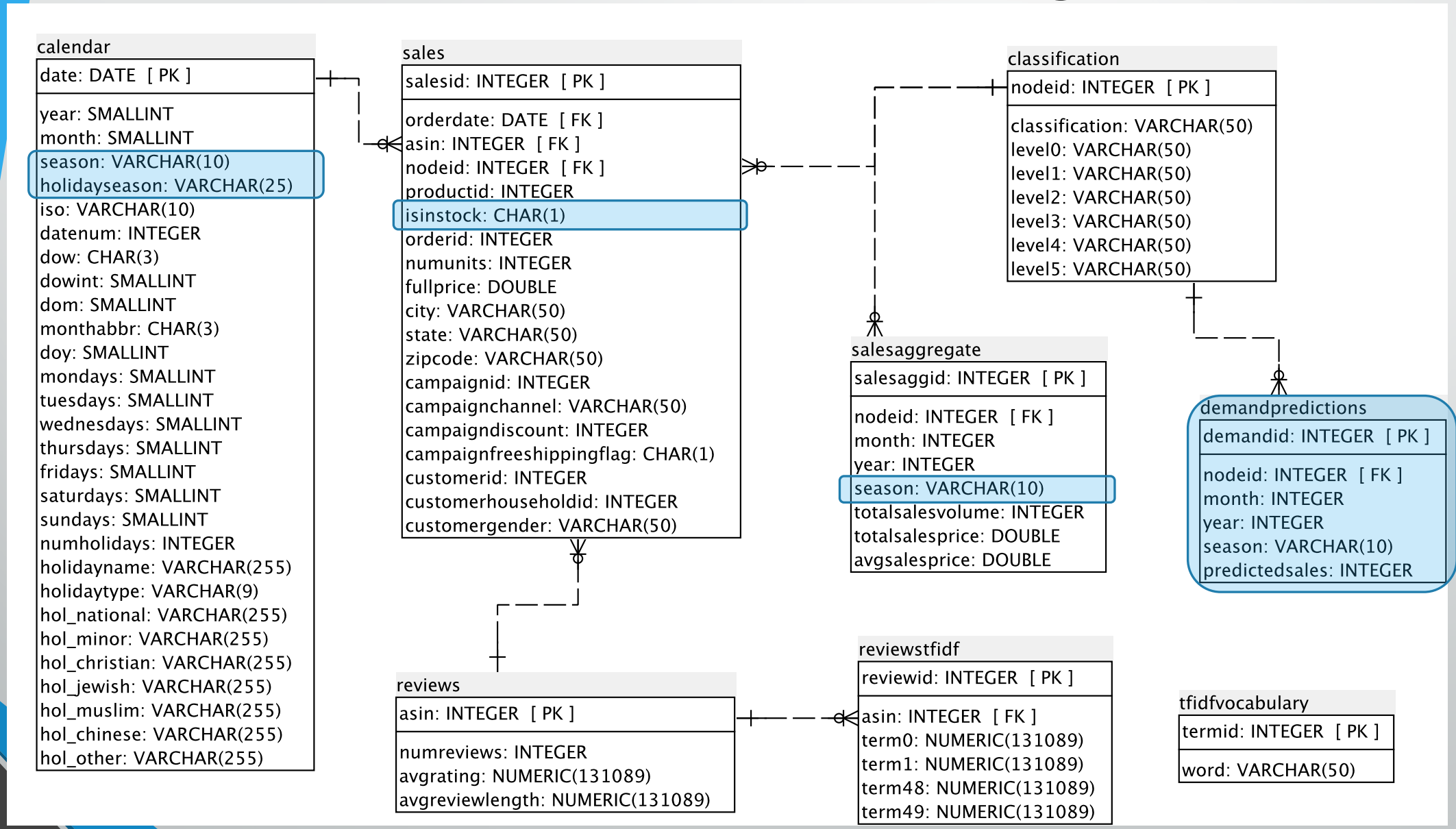
High Level Assumptions

1. Queries on mediated schema return the data required for the query execution team to provide data in the proper format to answer specific end user queries
2. Seasons are defined as follows:
   - Spring = March, April, May
   - Summer = June, July, August
   - Fall = September, October, November
   - Winter = December, January, February
3. "Around Christmas" means December
   - Could also create an additional view that aggregates sales by holiday season
4. Per discussions with ML team, category is defined by "nodeId" / "classification" rather than nested categories
   - Queries can be revised if necessary
5. Collaboration with other Demand Prediction teams via GitHub and Trello:
   - https://github.com/kshannon/dse203-demand-pred/blob/master/machine-learning-code/datalog-queries-presentation-3.txt
   - https://trello.com/b/5fDcmYGJ/dse-203-project-board

Summary of Schema Changes to Support Additional Stakeholder Queries

- Added ”Season” (meteorological seasons) and “HolidaySeason” attributes to Calendar relation
- Added “Season” attribute to ”SalesAggregate” relation
- Added “DemandPredictions” relation to store predictions from ML team
- Added “isinstock” from customer Products relation to mediated schema Sales relation
- Script to Update the Integrated Schema is available in Github

Updated Schema ER Diagram



Query 1: What are the top 3 categories of books that are most read around Christmas?

- Assumptions:
  - Sales volume is an appropriate proxy for “most read”
- Datalog query heads from ML Team:
  - Query1(nodeID, count_orderID):-
- Datalog query (against mediated schema) to obtain needed information:
  - Q1(classification, totalsalesvolume, year) :-
    classificationinfo(nodeid, classification, _ , _ , _ , _ , _ , _),
    salesaggregate (_ , nodeid, month, year , _ , totalsalesvolume, _ , _),
    month = 'December’

Query 2: What time of the year are the sales of "Education" books the highest?

- Assumptions:
  - Query is asking for months or seasons with highest education sales
- Datalog query heads from ML Team:
  - Query2 (month):-
- Datalog query (against mediated schema) to obtain needed information:
  - Q2 (month, year, season, totalsalesvolume, totalsalesprice) :-
    classificationinfo (nodeid, classification, _ , _ , _ , _ , _ , _),
    salesaggregate (_ , nodeid, month, year , season , totalsalesvolume, totalsalesprice , _),
    classification = 'Education'
  - Allows identification of top months or seasons by either total sales volume or total sales price

Query 3: Given month m and category c, predict the amount of sales for the category.

- Assumptions:
  - Query cannot be answered directly without input from ML team, unless "DemandPredictions" relation has been populated with ML demand prediction info
  - First version of Q_3 datalog provides information needed to execute ML demand prediction models to populate "DemandPredictions" relation
  - Second version of Q_3 assumes "DemandPredictions" relation is populated with current demand predictions
- Datalog query heads from ML Team:
  - Query3 (*) :-
  - Query3 (date_agg_month, inventory_sold_ratio, dollar_sold_ratio, volume_moved, product_rating_average, product_rating_delta, total_sales, contains_sold_out_product , large_inventory_drop , is_pos_sentiment , is_neg_sentiment, is_neutral_sentiment, count_of_nodeIDs, is_in_campaign) :-
- Datalog query (against mediated schema) to obtain needed information:
  - Q3 (classification, orderdate, isinstock, numunits, fullprice , campaignid, avgrating) :-
    classificationinfo (nodeid, classification, _ , _ , _ , _ , _ , _),
    sales (orderdate, asin, nodeid, _ , isinstock, _ , numunits, fullprice, _ , _ , _ , campaignid, _ , _ , _ , _ , _ , _),
    reviews (asin, _ , avgrating, _),
    reviewstfidf (_ , asin, termo, term1, … , term48, term49),
    tfidfvocabulary (termid, word)
  - Q3 (predictedsales) :-
    classificationinfo (nodeid, classification, _ , _ , _ , _ , _ , _),
    demandpredictions (_ , nodeid, month, _ , _ , predictedsales),
    month = $M,
    classification = $C

Query 4: Which book categories show a downward trend in demand in Winter and Spring?

- Assumptions:
  - Downward trend is within Winter and Spring season rather than from Fall to Winter and Spring to Summer
- Datalog query heads from ML Team:
  - Query4_spring (nodeID, spring_sale_trend) :-
  - Query4_winter (nodeID, winter_sale_trend) :-
- Datalog query (against mediated schema) to obtain needed information:
  - Q4 (classification, month, year, season, totalsalesvolume, totalsalesprice) :-
    classificationinfo (nodeid, classification, _ , _ , _ , _ , _ , _),
    salesaggregate (_ , nodeid, month, year , season , totalsalesvolume, _ , _),
    season = 'Winter' | season = 'Spring'

Query 5: Is there a category that we should discontinue stocking?

- Assumptions:
  - Business decision will be made to discontinue stocking categories when total sales volume and/or price in previous M months has not exceeded user-defined threshold
  - Alternative is to use pre-populated demand predictions to identify categories with low predicted sales
- Datalog query heads from ML Team:
  - Query5 (nodeID) :-
- Datalog queries (against mediated schema) to obtain needed information:
  - Q5 (classification, month, year, totalsalesvolume, totalsalesprice) :-
    classificationinfo (nodeid, classification, _ , _ , _ , _ , _ , _),
    salesaggregate (_ , nodeid, month, year, _ , totalsalesvolume, totalsalesprice , _)
  - Q5 (classification, month, year, predictedsales) :-
    classificationinfo (nodeid, classification, _ , _ , _ , _ , _ , _),
    demandprediction (_ , nodeid, month, year , _ , predictedsales)