1. For each of the top three categories:
   1. Aggregate sales volume per month/year. I.e, query total sales volume for computer programming category in July/2013.

**SELECT** nodeid,

MONTH,

YEAR,

total\_sales\_volume

**FROM** is\_sales\_by\_month sm

**INNER** **JOIN** categories cats

**ON** sm.nodeid = cats.nodid

**WHERE** MONTH = 7

**AND** YEAR = 2013

**AND** classification = “for computer programming”

* 1. Average monthly sales volume

**SELECT** nodeid,

MONTH,

YEAR,

avg\_sales\_volume

**FROM** is\_sales\_by\_month sm

* 1. Average monthly sales price

**SELECT** nodeid,

MONTH,

YEAR,

avg\_sales\_price

**FROM** is\_sales\_by\_month sm

1. For each of the top three categories (from EDA team):
   1. We need to know the distributions of every product in each of the categories. For example, a product name Python for beginners made up of certain % of sale volume of computer programming category.
   2. We need to know the mean and variance of monthly sales volume.
   3. We need to know the mean and the variance of average sales price.
2. Sales Buckets:
   1. Find the category with the greatest sales but the lowest variance in sales per month. that is bucket 1 (flat sales)
   2. Find the category with the most sales but where all months have small variance except for one month. (seasonal sales).
3. Aggregate sales volume per month/year for each of the top 100 categories
   1. Calculate the mean/variance of each category
   2. Based on some criteria (i.e. exceed 30% deviation of the mean for 1 month), we will categorize the category as time-invariant, season, and etc. Bucket the 100 categories to its label group
4. What percentage of the time do we sell more than 90% of the books in inventory?
5. How many times did we nearly sell out our stock? And can we predict when we will sell out of a given title based on time of year? / current sales trend? For example, does the number of times the book/category is mentioned in the reviews lead to greater sales?
6. What fraction of book categories have low variance (e.g. less than 20% of the mean) in terms of books sold per month?
7. Determine the number of times a category was mentioned in a given month to get a 'hot category' score?
8. What are the mean/median price of top products/product categories that were sold for the past 3 years?
9. What are the top sale products/product categories for each season?

SELECT productid, count(numunits) AS NumberOfSales,

CASE

WHEN orderdate BETWEEN CAST (&#39;2017-01- 01 00:00:00.000&#39; AS DATETIME) AND CAST

(&#39;2017-03- 31 00:00:00.000&#39; AS DATETIME) THEN 1

WHEN orderdate BETWEEN CAST (&#39;2017-04- 01 00:00:00.000&#39; AS DATETIME) AND CAST

(&#39;2011-15- 31 00:00:00.000&#39; AS DATETIME) THEN 2

WHEN orderdate BETWEEN CAST (&#39;2017-07- 01 00:00:00.000&#39; AS DATETIME) AND CAST

(&#39;2012-09- 30 00:00:00.000&#39; AS DATETIME) THEN 3 ELSE 4

END AS orderdate

FROM Sale s

GROUP BY CASE

WHEN orderdate BETWEEN CAST (&#39;2017-01- 01 00:00:00.000&#39; AS DATETIME) AND CAST

(&#39;2017-03- 31 00:00:00.000&#39; AS DATETIME) THEN 1

WHEN orderdate BETWEEN CAST (&#39;2017-04- 01 00:00:00.000&#39; AS DATETIME) AND CAST

(&#39;2011-15- 31 00:00:00.000&#39; AS DATETIME) THEN 2

WHEN orderdate BETWEEN CAST (&#39;2017-07- 01 00:00:00.000&#39; AS DATETIME) AND CAST

(&#39;2012-09- 30 00:00:00.000&#39; AS DATETIME) THEN 3 ELSE 4

END

1. For each category find how many NodeId it belongs too? (How many nodeIDs per bucket?)

SELECT count(nodeid) as count

FROM sales s

GROUP BY nodeid

1. For each category find value for feature from (invariant, variant, indeterministic demand based on Holiday season or months - One hot encoding)?
2. Number of books in each category sold per month per year?

select nodeif, count(nodeid), date\_trunc('month', orderdate)

from sales

where 1=1

group by nodeid, date\_trunc('month', orderdate)

order by nodeid, date\_trunc('month', orderdate)

1. Counts of orders per month, per year

select count(\*), date\_trunc('month', orderdate)

from sales

where 1=1

group by nodeid, date\_trunc('month', orderdate)

order by nodeid, date\_trunc('month', orderdate)

1. What percentage of the time do we sell more than 90% of the books in inventory?
2. How many times did we nearly sell out our stock? And can we predict when we will sell out of a given title based on time of year? / current sales trend?. For example, does the number of times the book/category is mentioned in the reviews lead to greater sales?
3. Average “Overall” (rating from 1 to 5) of a Category per month over time (Asterix)?

select avg(overall), date\_trunc('month', Unixreviewtime)

from sales

where 1=1

group by nodeid, date\_trunc('month', Unixreviewtime)

order by nodeid, date\_trunc('month', Unixreviewtime)

1. We would like to use the solr data to determine for the top 3 categories what the overall sentiment of each category is for a given month.

sentiment analysis can be done by using OpenNLP or

machine learning (Mahout or Weka). It is not a simple query in solr

As per the request, we are updating the queries into prediction variable ‘Y’ and feature attributes ‘X’we have thought of so far. We will update more after further analysis.

For a given category...

1. [y = num\_books\_sold] count number of books sold. (type = int) => Prediction variable
2. [x1 = inventory\_sold\_ratio] the ratio of inventory sold compared to all other categories (type = float)
3. [x2 = dollar\_sold\_ratio] the ratio of dollar amount sold compared to all other categories (type = float)
4. [x3 = volume\_moved] find the total number of books sold (type = int)
5. [x4 = product\_rating\_avg] find its avg product rating (type = float)
6. [x5 = product\_rating\_delta] find its product rating delta change from the previous month (type = float)
7. [x6 = total\_sales] calculate the total dollar sales (type = float)
8. [x7 = contains\_sold\_out\_product ] return T/F if any product within the category is sold out (type = boolean, 1/0)
9. [x8 = large\_inventory\_drop ] has any product in the category dropped below 90% of the inventory count from the month before (type = boolean 1/0)
10. [x9 = is\_pos\_sentiment ] are the product review’s avg sentiment positive (type = bool, 0/1)
11. [x10 = is\_neg\_sentiment] are the product review’s avg sentiment negative (type = bool, 0/1)
12. [x11 = is\_neutral\_sentiment] are the product review’s avg sentiment neutral (type = bool, 0/1)
13. [x12 = count\_of\_nodeIDs] count the number of nodeIDs that currently belong to the category (type = int)

Note: - Detailed information on Model and respective thoughts on features will open up more queries and will be updated during the presentation in class - Reviews information from solr doesn’t have temporal information but we can join that with reviews in relational DB based on asin and nodeid