DSE203

Data Integration and Analytics for Demand Prediction

Query Capability & Learning

Homework 1 October 13, 2017

Salah Ahmad
Ehab Abdelmaguid
Disha Singla
Nolan Thomas
Sanjay Kenchareddy

Agenda

1. Overall Goal and objectives
2. System Architecture
3. Preliminary requirements from ML team
4. Preliminary requirements for Schema team
5. Analytical capabilities within database
6. Suggestions for ML team
7. Deliverables to ML team
8. Open items

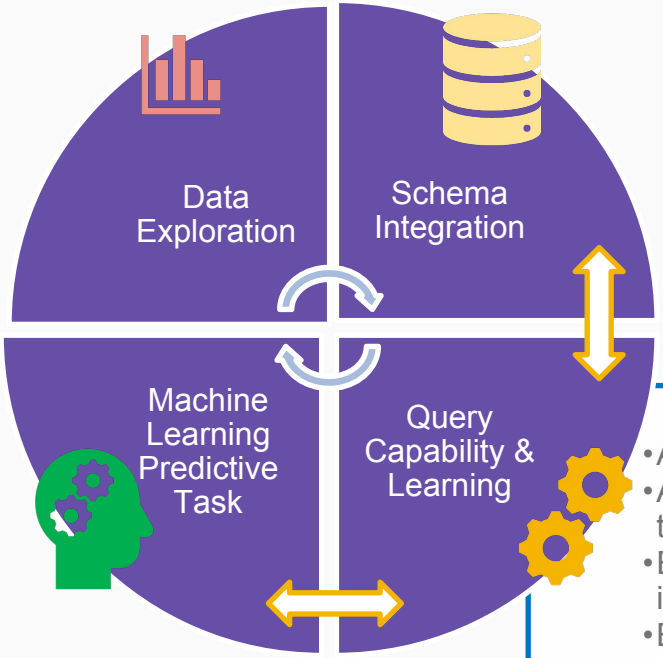
Goal and Objectives

Project Goal

Use **Customer purchasing behaviour** to build **integrated database system** and develop **demand analytics** to empower the client in making decisions for inventory management & demand forecast resulting in increased revenue and profit streams .

Q&L Team Objectives:

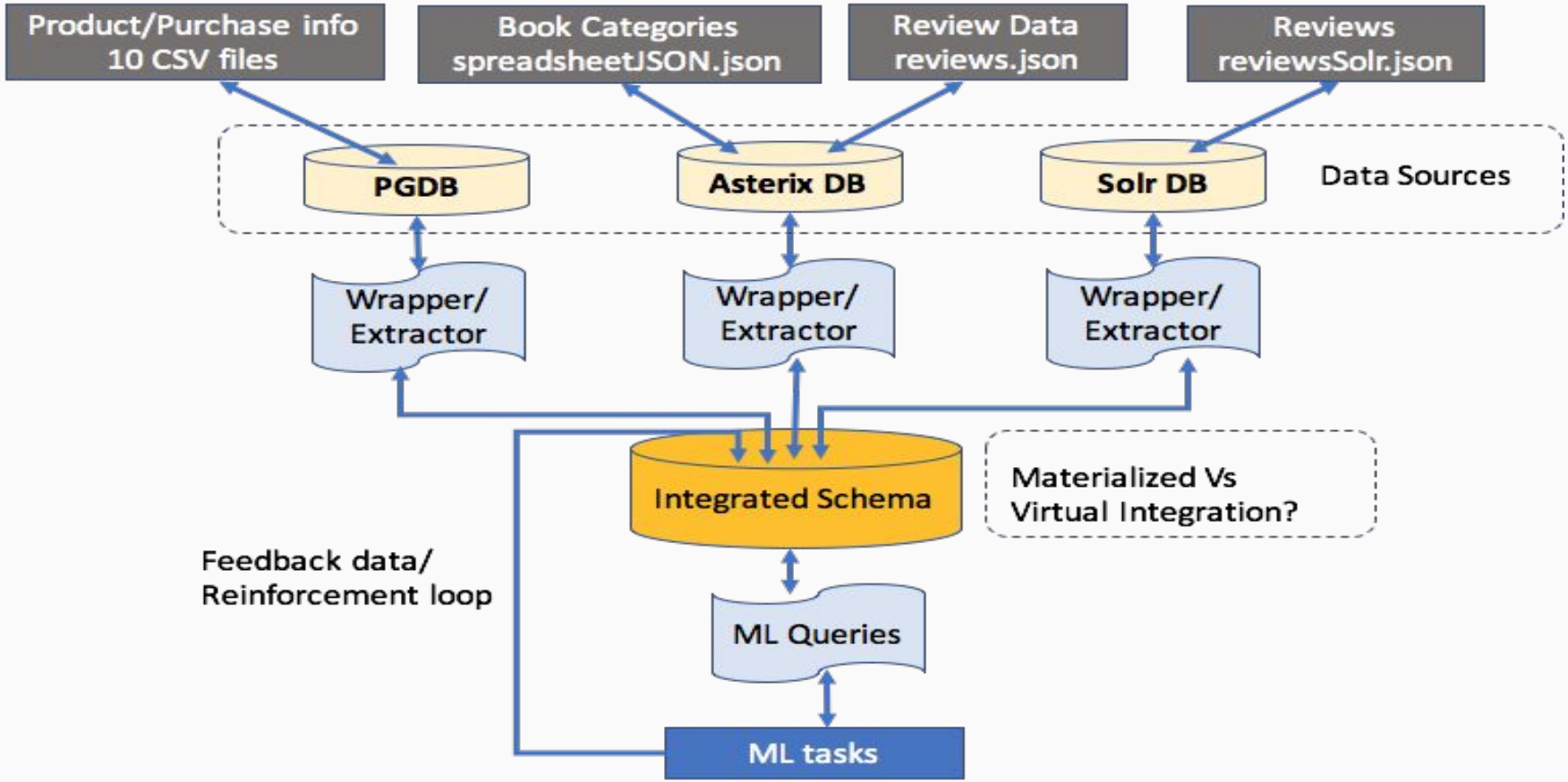
- Bridge the gap between schema and ML team
- Provide aggregated; non-aggregated and rolled up data
- Analytics within database



Tasks

- Analyze schema/data fields
- Analyze requirements from ML team
- Evaluate Query Capabilities, identify/fix gaps
- Build queries to extract data
- Optimize queries for performance

Ideal System Architecture



Preliminary Requirements from ML team

Build queries to extract following data to develop regression models for demand prediction

1. Book category
2. Sales from Ad campaign
3. Seasonality
4. Book's popularity
5. Sentiment analysis about book

Requirements for Schema Team

Shema Relationships	• Product_Id <-> asin <-> node_id(s) • Customers to Reviewers to Product • Customers to Subscribers
Schema Gaps	• Campaign duration (start and end date) • Are specific products targeted in a campaign rather than entire order • Subscription benefits: discounts, specific campaigns, free shipping • Subscription history [only know current state] • Product can be under multiple classifications (nodes) [GRE books]
Data Gaps	• i) Lack Inventory Data, ii) Shipping Fees (Impact of Free Shipping), iii) Taxation • Customer Zips: location information for location analysis • No order line data prior to 2009 (Is there historical data)
Data Types/Codes	• Reviews.helpful is an array [3,3] • Definition of subscribers.stoptype (M,V,I) codes
Data Anomalies	• Order line balancing (Full Amount less Campaign Discount) != Unit Price • Order lines with Unit Price $0 (Possible Campaign at Product Level) • Products without names • Household with 746 customers?

High level expectations from Schema Team

- ERD and Data Dictionary
- Hosted integrated solution:
  - Account for scalability, security
- Implementation/Technology Choices: optimal performant solution
  - Classification (support node hierarchy)
  - Book Reviews (textual sentiment analysis)
  - Book Store Data

Analytics capabilities within database(1/2)

Aggregation and Analytics at different levels

- Customer (for customer segmentation)
  - Total spending
  - Time based weekly, monthly, etc.
- Product (for campaign success, seasonality based predictions, product popularity)
  - Total units sold
  - Sales on weekly, monthly, yearly level
  - Sales based on campaigns

Analytics capabilities within database(2/2)

- Geography (for Geography based market segmentation)
  - sales at zip level of customers and rolling up to county, state, zone levels
- Category
  - Top seller in a category
  - Performance in a category
- Subscription
  - Sales contribution by subscribers
  - If part of any special subscription benefits like pick-N
  - subscription based discounts
- Text (for sentiment analysis)
    - SOLR search results; counts, length normalization, etc.
    - Potentially Tf-Idf vectors, etc.

Suggestions to ML Team

1. Seasonal campaign based demand predictions
2. Customer segmentation based on purchase history (gold, platinum, etc.)
3. Least popular product analysis
4. Customer segmentation-demographics (buying history based; time of the day; subscriptions; promotions, etc.)
5. Geography area based segmentation
6. Subscription based predictions

Deliverables to ML team

Deliverables

- Queries to extract data from integrated database
- Query outputs
- Query capabilities and limitations
- Analytical capabilities within DB
- Data Exploration capabilities within DB

Ways for provisioning data

- Flat files (data dumps)
- Self service through integrated transformed data warehouse
- API access
- Other suggestions ?

Questions

- All
  - Should we care about scalability of the data ?
  - How/where sampling of data performed ?

- Stakeholders
  - Are there any e-books /hardcover ?
  - Are there Personally identifiable information (PII) attributes in the customer table ?
  - Data governance & Security consideration

Q&A

Thank You