OnMart Superstore

Requirements Specification
&
Architecture

1. The OnMart Superstore

The OnMart superstore is an online retailer that has a business model similar to Amazon, Walmart, and Target business models, and it is experiencing a significant growth in its online shopping business. The online superstore has several departments that are selling products in different categories: Electronics, Clothing, Grocery, Furniture, Sports, etc.

The OnMart superstore receives thousands of online orders every hour with package deliveries to different addresses in different cities and states. To cope with the significant growth in its online orders and deliveries, OnMart has decided to utilize a real-time stream processing and microservices platform in order to develop an application that integrates its different departments and functional units. The real-time data streaming platform will be utilized by OnMart's real-time application in order to perform a variety of data analytics tasks related to its business process and workflows including logistics, supply-chain, operational process, forecasting product demands, product recommendations, cross-promotion programs, package delivery network and delivery vehicles, customer satisfaction and product returns, fraud detection, risk mitigation and contingency planning.

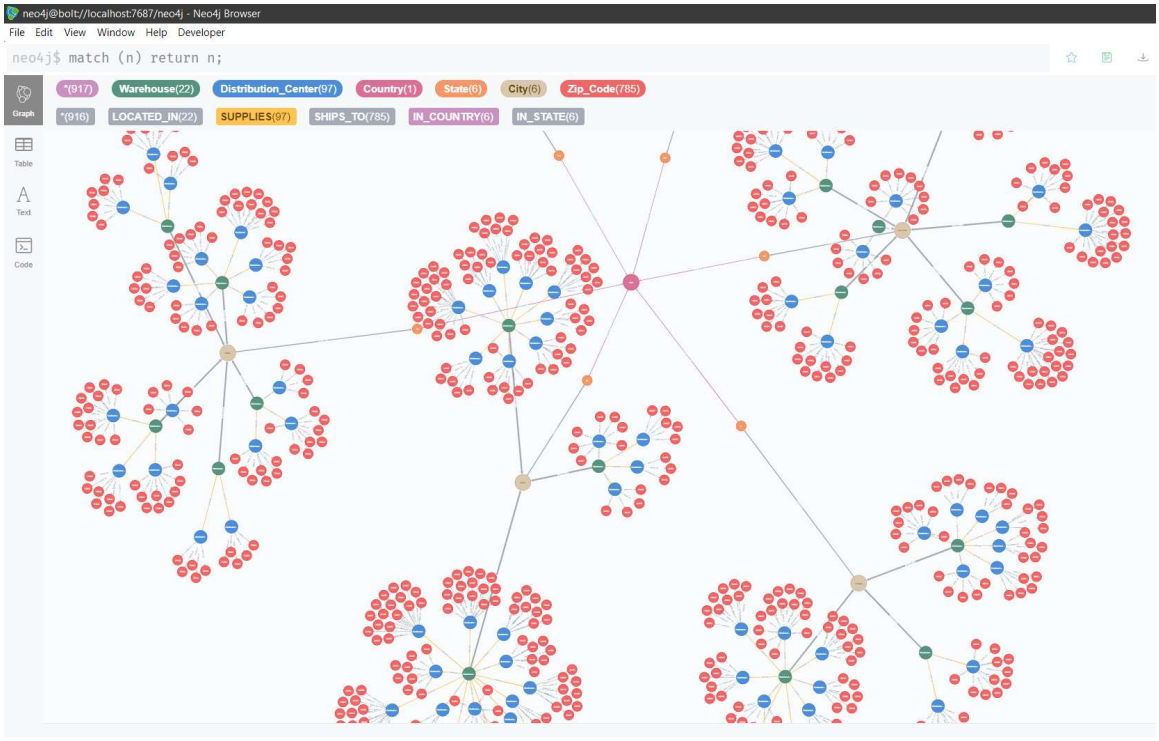
For its supply chain and its delivery network/vehicles, OnMart has several warehouses and distributions centers. Every warehouse supplies a number of distribution centers and every distribution center delivers packages in different zip codes.

The logistics and supply chain network for OnMart has the following characteristics:

1. It delivers packages to customers distributed across 785 zip codes
2. It has 97 distribution centers that are located in 97 zip codes

3. It has 17 warehouses that are located in 17 zip codes
4. It serves customers in 8 cities located in different states in the US
5. Every serving facility (distribution center or warehouse) has a unique pair of latitude and longitude

Even though OnMart has many warehouses that supply distribution centers in the different cities, not every city has warehouses; Nashville and Atlanta do not have warehouses. Currently, Nashville is being supplied by products shipped from Chicago warehouses and Atlanta is being supplied by products shipped from Miami warehouses. The following figure illustrates the structure of the OnMart delivery network:



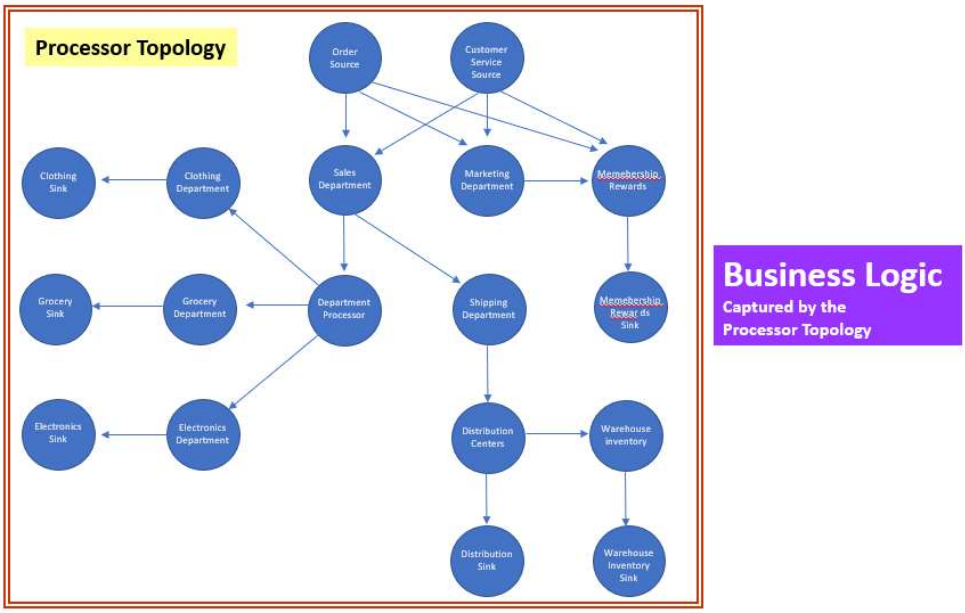
The following is a sample of the delivery zip codes, warehouses, distribution centers, cities, and state:

	A	B	C	D	E	F	G	H
1	**Zip**	**City**	**State**	**Latitude**	**Longitude**	**Facility_ID at given Location/Zip**	**Distribution Center ID Serving this Zip Code**	**Warehouse ID supplying this Distribution Center**
2	60651	Chicago	IL	41.901485	-87.74055	Warehouse_1	DistributionCenter_1	Warehouse_1
3	60697	Chicago	IL	41.811929	-87.68732		DistributionCenter_1	Warehouse_1
4	60667	Chicago	IL	41.811929	-87.68732		DistributionCenter_1	Warehouse_1
5	60694	Chicago	IL	41.811929	-87.68732		DistributionCenter_1	Warehouse_1
6	60684	Chicago	IL	41.811929	-87.68732		DistributionCenter_1	Warehouse_1
7	60644	Chicago	IL	41.881331	-87.75671	DistributionCenter_1	DistributionCenter_1	Warehouse_1
8	60668	Chicago	IL	41.811929	-87.68732		DistributionCenter_1	Warehouse_1
9	60646	Chicago	IL	41.995331	-87.7601		DistributionCenter_1	Warehouse_1
10	60654	Chicago	IL	41.888627	-87.63538		DistributionCenter_2	Warehouse_1
11	60683	Chicago	IL	41.811929	-87.68732		DistributionCenter_2	Warehouse_1
12	60616	Chicago	IL	41.8474	-87.63126		DistributionCenter_2	Warehouse_1
13	60647	Chicago	IL	41.921126	-87.70085	DistributionCenter_2	DistributionCenter_2	Warehouse_1
14	60606	Chicago	IL	41.882582	-87.6376		DistributionCenter_2	Warehouse_1
15	60623	Chicago	IL	41.850232	-87.718		DistributionCenter_2	Warehouse_1
16	60663	Chicago	IL	41.811929	-87.68732		DistributionCenter_2	Warehouse_1
17	60609	Chicago	IL	41.814283	-87.65282		DistributionCenter_2	Warehouse_1
18	60660	Chicago	IL	41.990631	-87.6667		DistributionCenter_2	Warehouse_1
19	60615	Chicago	IL	41.801525	-87.60215		DistributionCenter_2	Warehouse_1
20	60622	Chicago	IL	41.900332	-87.66927		DistributionCenter_2	Warehouse_1
21	60625	Chicago	IL	41.971614	-87.70256	Warehouse_2	DistributionCenter_3	Warehouse_2
22	60649	Chicago	IL	41.761734	-87.57072		DistributionCenter_3	Warehouse_2
23	60673	Chicago	IL	41.811929	-87.68732		DistributionCenter_3	Warehouse_2
24	60655	Chicago	IL	41.696283	-87.69912		DistributionCenter_3	Warehouse_2
25	60641	Chicago	IL	41.946431	-87.74576		DistributionCenter_3	Warehouse_2
26	60679	Chicago	IL	41.811929	-87.68732		DistributionCenter_3	Warehouse_2
27	60636	Chicago	IL	41.776633	-87.66854		DistributionCenter_3	Warehouse_2
28	60657	Chicago	IL	41.940832	-87.65852	DistributionCenter_3	DistributionCenter_3	Warehouse_2
29	60618	Chicago	IL	41.945681	-87.7048		DistributionCenter_3	Warehouse_2
30	60677	Chicago	IL	41.811929	-87.68732		DistributionCenter_3	Warehouse_2
31	60628	Chicago	IL	41.695434	-87.62255		DistributionCenter_4	Warehouse_2

2. Departments and Products

OnMart has several departments, functional units, warehouses and distribution centers located in different cities and states across the US. The following diagram illustrates the departments located in different cities that require real-time collaboration and data streaming platform:

The following diagram illustrates how the business logic for OnMart's business process and its workflows can be represented as a processor topology that can be programmed utilizing a real-time data streaming platform like Kafka:



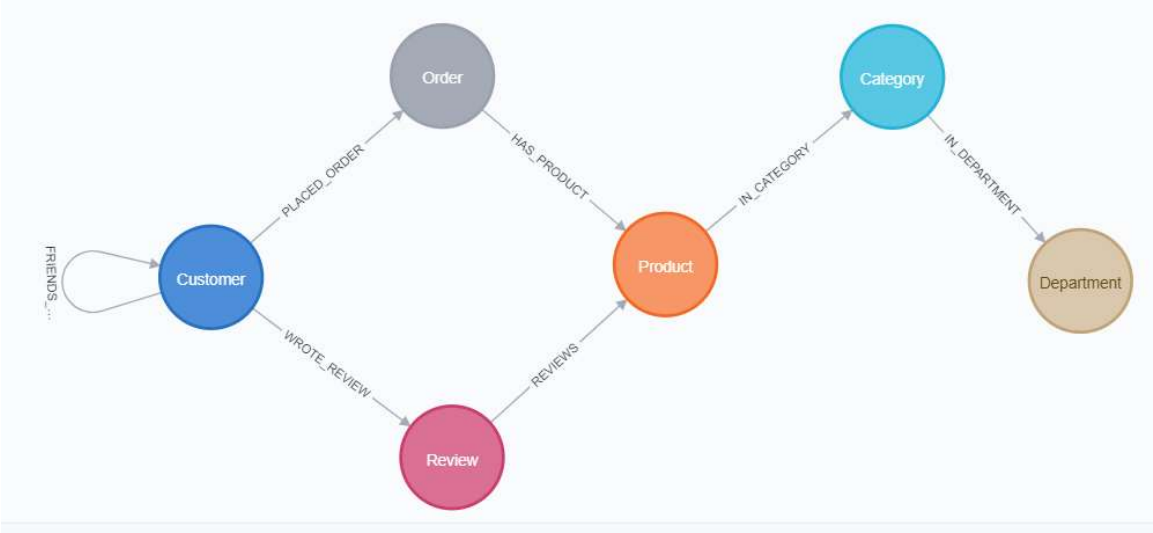
OnMart has several departments that are selling products in different categories. The following is a sample of these departments and the different product categories they sell:

Department	**Product Categories (Sample)**
Furniture	Bookcases, Chairs, and Tables
Grocery	Dairy, Beans, Pasta, Paper products, and cleaning supplies.
Office Supplies	Desk Supplies, Filing Supplies, Paper & Pads, Binding Supplies, and Stationery/mailing supplies
Clothing	Shoes, Socks, Dresses, Jackets, Shorts and Shirts
Electronics	Computers, TV, Phones, and Sound System

3. Customers and Purchases

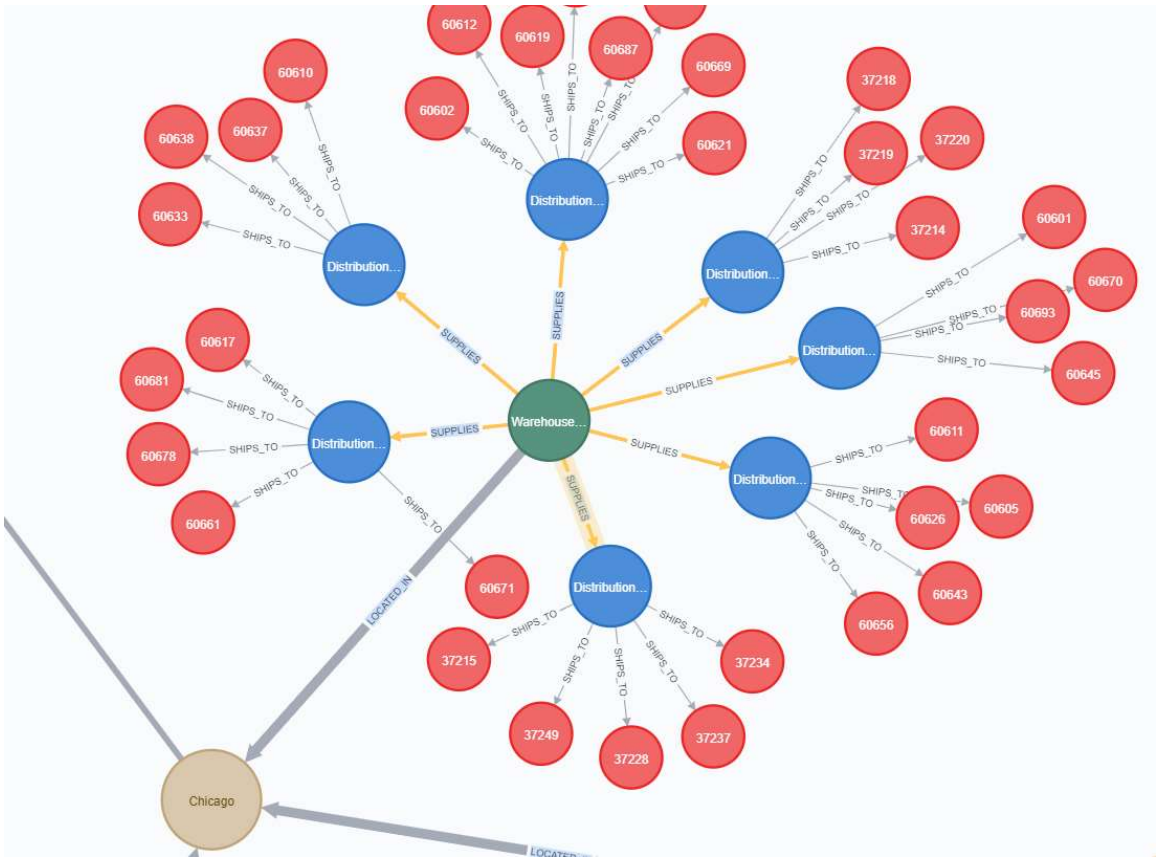
Customers **place** their orders online and the order might have products from the different categories in the different departments. For every order there will be a unique order number generated. When the customers place the order online, the customer will enter home address, delivery address, and shipping class.
The customer shall be provided with the following capabilities:

- **Return** purchased items.
- **Review** and **rate** purchased items.
- Some customers are connected with **friends** in a **Social Network** as illustrated in the **graph data** model listed below and can **share** product description of purchased items with **friends** in the **social network**.



4. Geospatial and Logistics

The logistics department is responsible for the logistics and supply chain network for OnMart Superstore. The department team members track the status of the warehouses, distributions centers, inventories, deliveries, delays, and returned products for the different zip codes that OnMart serves. The department tracks the connections between the cities, warehouses, distribution centers and the zip codes they serve within the OnMart supply chain network; the **data graph** below shows one of the distribution centers in Chicago along with the distribution centers associated with it and the zip codes that it serves.



The department is also responsible for creating contingency plans to cope with adverse weather conditions that might impact the daily operations for the warehouses, distribution centers and order deliveries.

To ensure quality of service for its customers and continuous improvement for customer satisfaction, OnMart has requested to conduct a study to produce an analytical research report utilizing **Facebook/Prophet** package and **Neo4j GDS** library to address the following:

1. Forecast the number of returned orders for every warehouse
2. Forecast the number of delayed deliveries for every zip code
3. Forecast the number of delayed deliveries for every distribution center
4. Identify top 10 distribution centers with highest number of order returns
5. Identify top 10 distribution centers with lowest number of order returns
6. Identify top 10 distribution centers with lowest number of deliveries
7. To offload busiest **warehouse** with highest number of **order returns**, identify the warehouse that has the **forecast** for the highest number of order returns, and find its nearest neighboring warehouse that has the lowest number of order returns (consider 50-200 mile radius)
8. To offload busiest **distribution center** with highest number of delayed deliveries, identify the warehouse that has the **forecast** for the highest number of delayed deliveries, and find its nearest neighboring distribution center that has the lowest number of delayed deliveries (consider 1-20 miles radius)

5. Marketing and Product Cross-Promotion:

5.1 Finding influential reviewers

OnMart customers write many reviews for the products that OnMart sells, and the challenge for the marketing team is to decide which reviews will be displayed first and what criteria is to be used for ordering the reviews for the potential new buyers (customers).

Since the customers who wrote product reviews are part of a social network and many of them are also connected to other customers in the social network, OnMart is requesting to leverage this information to find those customers who wrote at least three reviews for its products and are connected to other customers in the social network. In essence, OnMart is requesting to find the list of **influential reviewers** that will influence other customers, even though these customers are not direct friends in the social network. There are a few algorithms in the graph data science library (like Google **PageRank**, **ArticleRank**, etc.) that could be used to find **influential reviewers** and OnMart is requesting to utilize these algorithms to identify the **influential reviewers**.

5.2 Product Cross-Promotion

OnMart superstore is also interested in identifying products in different departments that it should target for cross-promotion campaigns.

The approach that OnMart is requesting is to find the well-connected customers who are also influential reviewers. These super-connectors can be used to attract customers from different types of communities/groups of customers by utilizing the **Betweenness Centrality** algorithm from the graph data science library. Effectively, these super-connectors will act as a connecting bridge between the different customers communities/groups. So the task that OnMart is requesting is to look at what other products these super-connectors like and use those for a cross-promotion campaign.

For example, OnMart is requesting to identify products from the **Furniture department** that it could use in a cross-promotion with the newly introduced Samsung TV in the **Electronics department.** This is an extremely valuable

feature for the retailer since an influential reviewer and super-connector customer who remodeled the living room and bought a new TV along with a new sofa, for example, can be used to attract new customers from different groups/communities of customers. In essence, we can use the **Betweenness Centrality** and **PageRank** algorithms to find which OnMart superstore reviewers are not only well connected across the whole **Social network** but might also act as a bridge between different customer groups.

5.3 Product Recommender

Due to the evolving and continuous changes in the behavior of its online customers and in an effort to invest more in focused advertisement and targeted marketing, OnMart requested a study to evaluate the different recommendation models and provide experimental results to help it in the final decision making process for the selection of a recommender system for its business. There are a variety of models and categories for recommender systems:

- **Collaborative-Filtering Models**: These models are **behavioral** and based on the interaction patterns between the users and items through purchases, ratings, and reviews in order to detect similarities between the users or the items.
- **Content-Based Models**: These models are hybrid and they utilize the **structural** properties and **behavioral** user-item interactions to detect similarities between items or users; for example, two items are considered similar if they are in the same category, or have similar image, textual descriptions, or reviews. These models are utilized in the **cold-start** scenarios where we have limited to no interaction data for the newly introduced item or a new user.
- **Specialized Models**: These models include trending and popularity-based models, periodicity-based models for cycles of repeating user behavior in different seasons and holidays.

OnMart has demonstrated its bias toward using the **collaborative-filtering** model because it has access to the data of its large customer network of product reviews and purchases, and they are very much interested to utilize that in their recommender system.

We have been asked by OnMart to utilize the **collaborative-filtering** model to find customers who have given similar ratings to the same products, in essence, finding similar customers to suggest products that have high ratings. In this approach we are looking to find links and associations between users rather than the rated products. This approach will discover

the groups of customers who rated certain products in a similar manner and then make recommendations for customers in every group for products they didn't purchase or rate yet since it is highly likely these same customers in every group will share their ratings and preferences in future ratings and purchases.

OnMart wants to utilize the recommender system to identify products that it can recommend through email to its customers who made recent purchases from its different departments. There are many recommendation algorithms that fall into different categories (collaborative-filtering, content-based, matrix factorization, hybrid-model, etc.) and there are many freeware python packages that have implemented popular recommendation algorithms; **SurpRise** and **TensorFlow Recommenders** are examples of these packages.

OnMart is requesting that different experiments be conducted to evaluate and develop any **two** recommender systems that we could choose from the list of the recommender systems published on GitHub (https://github.com/grahamjenson/list_of_recommender_systems) and provide a formal evaluation for the performance and product recommendation accuracy of these algorithms when running these algorithms on CPUs and GPUs; note that not all packages (algorithm implementation) are developed to run on GPU.

As part of the research study, you need to identify the values for the threshold to use in order to filter products and customers in the same or different geographical locations; consider the following in your experiments:

1. Filter the products with more than 100 ratings.
2. Filter the customers that have rated more than 10 products.
3. Filter the products with more than 100 ratings in a certain city.
4. Filter the customers that have rated more than 10 products in a certain zip-code.
5. Exclude products that are in the top 5 highest product returns from the dataset used to build the recommender system.

6. Forecasting Sales Revenue and Product Demand Analysis:

There are many algorithms to forecast sales revenue and product demands. OnMart needs a formal evaluation of the popular forecasting algorithms to predict its product demand and sales per zip code, distribution centers, and warehouses. This feature requires comparative analysis and research of common forecasting models and the application of descriptive analytics methods part of the evaluation study.

For this feature, OnMart's requirement is to perform a number of experiments to forecast sales revenue and product demands using time-series analysis utilizing the three packages **Facebook/Prophet**, **TensorFlow Time Series (TFTS)**, and **StatsModel**.

Part of the exploratory/descriptive data analysis and predictive work that will be done for this feature:

1. The day of the week that has the maximum number of purchases made
2. The month of the year that has the highest sales revenue
3. The weekly sales revenue forecast per zip-code
4. The daily number of purchases forecast per distribution center
5. The weekly number of product returns forecast per warehouse
6. Whether the number of delayed deliveries per zip-code positively correlated to the number of returned purchases/products.
7. The busiest zip code
8. The most selling product in every zip-code
9. Analyze whether the sales of the Electronics department and the Clothing department have similar seasonal patterns
10. Product purchases seasonality through the year; for example, Back-to-School, Christmas, Superbowl, Weather, etc.
11. Build time series analysis model for product X in zip-code Y to forecast daily sales revenue

7. Fraud Detection and Contingency Planning

7.1 BOT-Generated and Fake Ratings Detection

Customers of OnMart utilize their credit cards for their online purchases and they can write reviews and rate the products they purchased online.

OnMart recently noticed an increase in the number of customer complaints related to fake reviews and fraudulent transactions. OnMart requested a formal analytics study to identify the root cause of the problem in order to rectiy the issues underneath.

To detect fake or bot-generated reviews, there are few guidelines (https://consumerfed.org/review-bots-false-advertising-or-opinion-spam/) that OnMart suggested for the development team to consider in order to identity fake reviews/ratings:

- High volume of reviews in a short time span
- High percentage of five-star ratings
- Reviews with similar/exact language from different users for the same product
- Multiple ratings of the same product from the same user in 2 or more zip codes in a short period of time

7.2 Credit Card Fraud Detection

To detect fraudulent transactions for credit card purchases, OnMart requested a set of rules and a decision tree model (classifier) to help detect fraudulent credit card transactions.

To classify the real-time credit card transactions, a classifier will use a pool of classification rules to classify the transaction in Real-Time (RT) as **Fraudulent** or **Approved**. Majority of the credit card detection systems send alerts when a fraud detected to a fraud detection department for further follow up by an investigator and another alert for the credit card holder.

The fraud detection process for OnMart will have a pool of rules that it can apply to check the current transaction with the previous transactions. For example, some of the features and descriptive metrics that could be used are average spending, average number of daily/monthly transactions,

product category, profile of the card holder, etc. The goal is to create a decision tree model (classifier) that predicts whether the transaction is fraudulent or not fraudulent (target variable) by learning the decision rules inferred from the transactions log dataset features.

OnMart has requested that the following rules be considered and analyzed for labeling/classifying the transactions when building the predictive model::

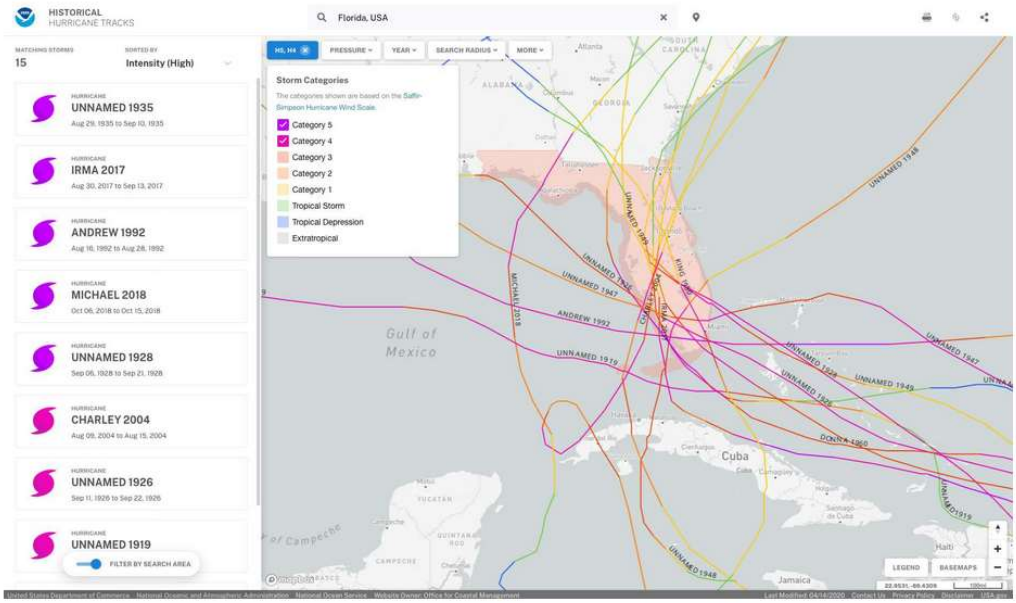
1. If a credit card is used in a short period of time for multiple deliveries in different zip codes, then fraudulent transaction
2. If transaction amount > $1,000 and credit card is used for first time in the delivery zip-code, then fraudulent transaction
3. If transaction amount <= $1,000 and credit card is used for first time in the delivery zip-code, then not fraudulent transaction
4. If transaction amount > $1,000 and credit card is not used for first time in the delivery zip-code, then not fraudulent transaction
5. If transaction amount <= $1,000 and credit card is not used for first time in the zip-code delivery, then not fraudulent transaction

7.3 Contingency Planning and Disaster Recovery

OnMart has a number of warehouses and distribution centers across the US in different states and cities to ensure on time deliveries for the online purchases. However, the hurricane season in Florida that lasts from June until the end of November presents a number of challenges and uncertainties for OnMart logistics, supply chain, and business opportunities due to the uncertain weather conditions. The following map shows category 4 and 5 hurricane tracks that crossed over the state of Florida between 1910 and 2018 (Source: https://oceanservice.noaa.gov/news/historical-hurricanes/).

NOAA Historical Hurricane Tracks

Explore more than 150 years of **historical hurricane landfalls**.



Due to the uncertainties of the weather conditions related to the hurricane season and their catastrophic impacts on its logistics and supply chain network in state of Florida, OnMart has requested to find a backup warehouse for every serving warehouse in the state of Florida; the backup warehouse must be the **farthest** warehouse that has the **lowest number of shipments** within **150-200 miles radius** for every serving warehouse.

8. Mock Data Producer for Purchases

To simulate the real-time transactional data stream, we will use kafka as the streaming platform and we will create a mock data producer to generate the real-time transactions. The simulator (the mock data producer) must be able to produce 10, 100, and 200 transactions/second. Every transaction will have the information related to the customer, delivery address, product name, product category and department, the distribution center responsible for purchase/order delivery and its supplying warehouse, etc.

The Mock Data Producer will produce real-time transactions that have the following data fields. Also provided below are examples of the data values for some of these fields:

CustomerID
- 596-64-3428
- 047-57-4034
- 265-57-4688

FirstName

LastName

CreditCardNumber
- xxxx-xxxx-xxxx-6796
- xxxx-xxxx-xxxx-0475
- xxxx-xxxx-xxxx-3492

OrderID
- 830540e1-8559-4ae9-b262-15cd5108d3fe
- 58735979-f7a4-453b-b90a-462308d4b835
- 7a799356-d831-4bfc-82dc-68bdaf38e1f5

PurchaseDate
- Mon Dec 28 23:22:32 CST 2020
- Mon Dec 28 23:36:12 CST 2020
- Mon Dec 28 23:29:59 CST 2020

DeliveryDate
- Sat Jan 02 23:28:22 CST 2021
- Sat Jan 02 23:29:31 CST 2021
- Sat Jan 02 23:29:58 CST 2021

ShippingClass

Standard

Two Days

Next Day

ProductID
- CLO-SH-0374-496
- FUR-BO-9493-010
- GRO-BE-0741-135

Department
- Furniture
- Grocery
- Electronics

Category

- Bookcases
- Beans
- Accessories

ItemPurchased
- Lava Classic Bookcase - PVC
- RANCH STYLE Black Label Black Beans
- Blu-ray Single Disc - 10/Pack

Quantity
Price
ShippingCost
Discount
Sales
Profit
DistributionCenterID
- DistributionCenter_53
- DistributionCenter_86
- DistributionCenter_21

DeliveryZipCode
HomeZipCode
TransactionStatus
- Approved
- Declined

City
Region
- East
- South
- West
- South
- North

State
Country
Segment
- Corporate
- Consumer
- Home Office

OrderPriority
OrderReturned
- Yes
- No

OrderDeliveryDelayed
- Yes
- No

Rating
- 1 to 5

ReviewID
Friends
SharedWith
ZipCodeInSupplyChainNet
CityInSupplyChainNet
StateInSupplyChainNet
LatitudeInSupplyChainNet
LongitudeInSupplyChainNet

FacilityIDInSupplyChainNet
    Warehouse_7
    DistributionCenter_31
DistributionCenterIDInSupplyChainNet
    DistributionCenter_51
    DistributionCenter_89
    DistributionCenter_6
WarehouseIDInSupplyChainNet
    Warehouse_15
    Warehouse_1
    Warehouse_8