AGM NoSQL Business Case Scenarios

DATASCI 205 – Section 6

Mohak Buch, Nicholas Lin, Soumik Mukherjee, Yoni Nackash

Berkeley
UNIVERSITY OF CALIFORNIA

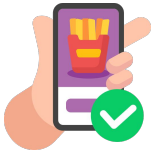
Enhancing our Customer Experience

- AGM has been growing, and leadership is interested in continuing to expand
- Tasked with coming up with several business cases to increase active users and overall sales
- Our NoSQL database use cases can be used individually, or in succession, with the overall objective of ***enhancing our customer experience lifecycle***
- These use cases span different components of the the customer lifecycle:

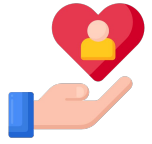
Identify several new high-density pickup locations

Calculate shortest pickup path on behalf of customer

Provide real-time order tracking

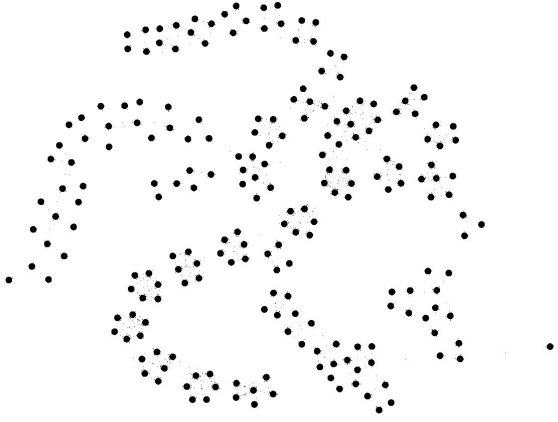


Tailor experience with personalized recommendations



Neo4j

- Neo4j is graph database designed to handle connected data and their relationships
- It offers native tools for visualizing, analyzing, and querying these relationships
- Uses an SQL-like language called Cypher to explore and manipulate graphs
- Has a suite of built-in functions for centrality and shortest path calculations
  - Includes PageRank, Louvain, Dijkstra's shortest path algorithms



Why Neo4j?

- Performance
  - Relational databases struggle with traversing relationships which require deep join operations
  - Leads to slow query performance especially when analyzing connectivity/influence of large networks
- Design
  - Data is modeled as nodes and edges, which directly represent entities and their relationships
  - Allows for more intuitive development

Neo4j: Identifying Most Influential Stations

Implementation

1. Run **PageRank** on graph database including all stations, connections, and weights (travel times)
    a. Measures the influence of each node within the graph, based on incoming relationships and the influence of its surrounding nodes
2. Average the influence of each Station (colors)
3. Identify the top Stations with most influence as a central pickup location

	Station	Average PageRank
1	Coliseum	0.689699
3	Daly City	0.674382
7	Pittsburg Center	0.657570
4	MacArthur	0.652321
0	Bay Fair	0.651678
6	Pittsburg	0.645705
9	West Oakland	0.644953
5	North Concord	0.641978
2	Concord	0.640689
8	Pleasant Hill	0.639865

Neo4j: Identifying Clusters and Top Stations as Pickup Locations

Implementation

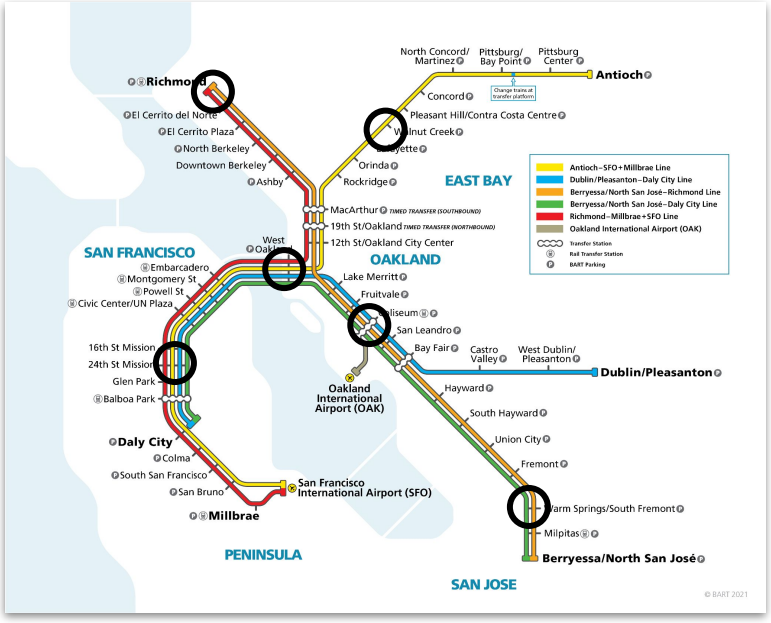
1. Run **Louvain Modularity** to group stations into clusters
    - Evaluates how dense nodes in a community are compared to how they would be in a random network
    - Accounts for weighted relationship between nodes
        - Travel Time between Stations
2. Use BART ridership data to find the average total exits over the past 3 months
3. Combine data with clustered stations
4. Identify most popular exits within each cluster
5. Setup pickup location for customers

	Cluster	Station	Avg Total Exits Over 3 Months
0	0	MacArthur	88943.0
1	1	24th Street Mission	144064.0
2	2	Glen Park	111682.0
3	3	San Leandro	76013.0
4	4	Castro Valley	27956.0
5	5	Richmond	52704.0
6	6	West Oakland	380779.0
7	7	Warm Springs	51099.0
8	8	Lake Merritt	92132.0
9	9	Powell Street	324129.0
10	10	South San Francisco	43878.0
11	11	Walnut Creek	64991.0

Combining PageRank & Louvain Modularity to Determine Pickup Locations

Average PageRank

	Station	Average PageRank
0	Coliseum	0.689699
1	Daly City	0.674382
2	Pittsburg Center	0.657570
3	MacArthur	0.652321
4	Bay Fair	0.651678
5	Pittsburg	0.645705
6	West Oakland	0.644953
7	North Concord	0.641978
8	Concord	0.640689
9	Pleasant Hill	0.639865



Louvain Modularity Clusters

	Cluster	Station	Avg Total Exits Over 3 Months
0	0	MacArthur	88943.0
1	1	24th Street Mission	144064.0
2	2	Glen Park	111682.0
3	3	San Leandro	76013.0
4	4	Castro Valley	27956.0
5	5	Richmond	52704.0
6	6	West Oakland	380779.0
7	7	Warm Springs	51099.0
8	8	Lake Merritt	92132.0
9	9	Powell Street	324129.0
10	10	South San Francisco	43878.0
11	11	Walnut Creek	64991.0

Berkeley
UNIVERSITY OF CALIFORNIA

Neo4j: Identifying Shortest Path for Pickup

Implementation

1. Identify customer's closest station to their home by geodesic distance
2. Leverage **Dijkstra's algorithm** to calculate the shortest path between the customer's home station and our new key pickup locations
    - The cost of a path is calculated via the weighted relationship between stations, which is travel time in seconds
    - Guaranteed to find shortest path via exhaustive search
3. Compare the shortest path cost of traveling to each key pickup location, and determine which station is most ideal for the customer
4. Provide customer with path instructions


```
This customers ideal pickup station is:  West Oakland

---------------------------------
   Total Cost:  600
   Minutes:  10.0
---------------------------------
depart MacArthur, 0, 0
yellow MacArthur, 0, 0
yellow 19th Street, 180, 180
yellow 12th Street, 120, 300
yellow West Oakland, 300, 600
arrive West Oakland, 0, 600
```

Berkeley
UNIVERSITY OF CALIFORNIA

Neo4j: Identifying Shortest Path for Pickup

Shortest Path Algorithm Considerations

A* Algorithm	• Performance improvement over Dijkstra's through cost estimation, but no guarantee of accuracy • Willing to compromise slightly on performance in favor of guaranteed accuracy → increase customer satisfaction
Single Source Algorithm	• Could be leveraged to find shortest path between customer location and all key nodes, however path is not returned, only cost • Including the path is critical for clear customer pickup instructions

Redis

In Memory Database	• Stores data in memory, used as a cache or quick access database
Business Use	• Track progress of any given order in real time
Advantages	• Provides rapid data access by storing frequently accessed data in memory • Redis Geospatial Indexing functions may also be easily used to provide more information to customers based on their location.

MongoDB

Document Database	• A collection of documents, JSON/JSON like objects
Business Use	• Storing order, user information
Advantages	• Well suited for performing aggregations and analytical queries on the data • Distribute data across different servers

Conclusion & Next Steps

- As AGM has been growing choosing BART stops for food pick up locations will be critical for our next phase of growth
- Neo4j can be leveraged to identify these pick up locations and the shortest path for customer pickup
    - Coliseum, 24th Street, West Oakland, Warm Springs, Richmond, Walnut Creek
- Redis: Real Time Order Tracking
- Mongo: Storing user, order information
- **Next Steps**: proof of concept with one or two key pick up locations to gauge value and collect insight

Thank You

Berkeley
UNIVERSITY OF CALIFORNIA