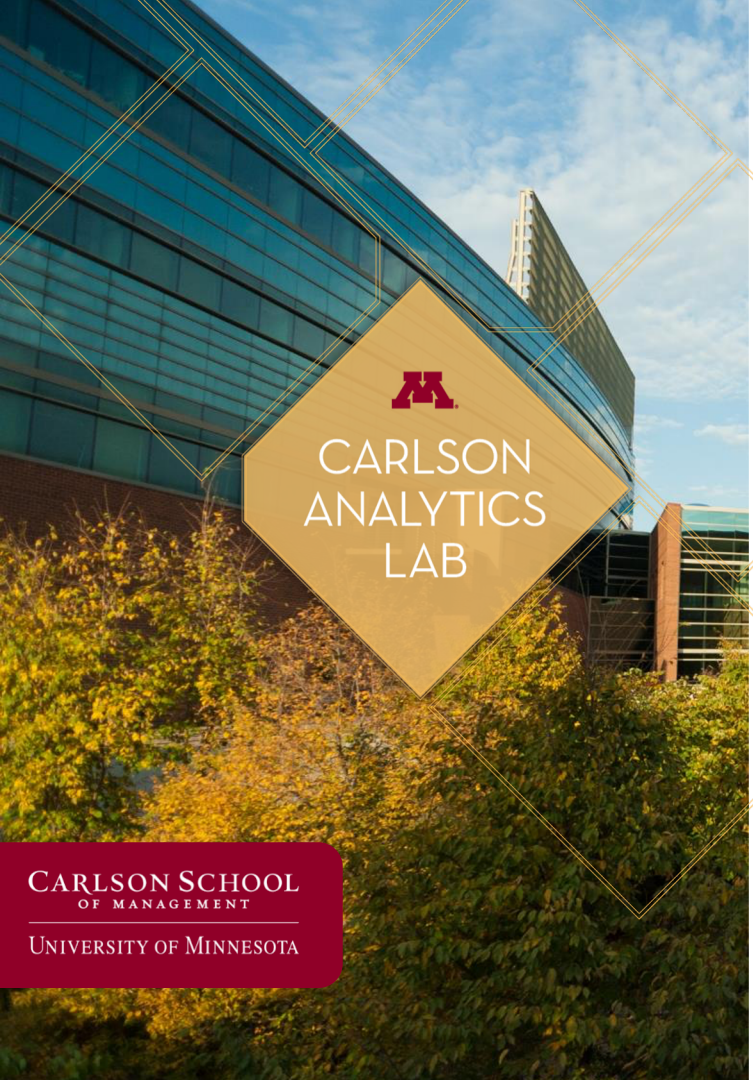


Hennepin County Graph Database Presentation

May 1st, 2020

Team O: Chia Hsuan Chava Chou, Brady Engelke, Yidan Gao, Minya Na, William Wu
Team P: Samira Arondekar, Maya Carnie, Mona Tai Hsuan Kan, Shobhit Mishra, Chuchen Leo Xiong

Hennepin County Aims to Mitigate Disparities for Citizens Across these 7 Domains

Hennepin County Wants to Move Fast to Address the 21 Disparity Initiatives Approved Across the 7 Domains

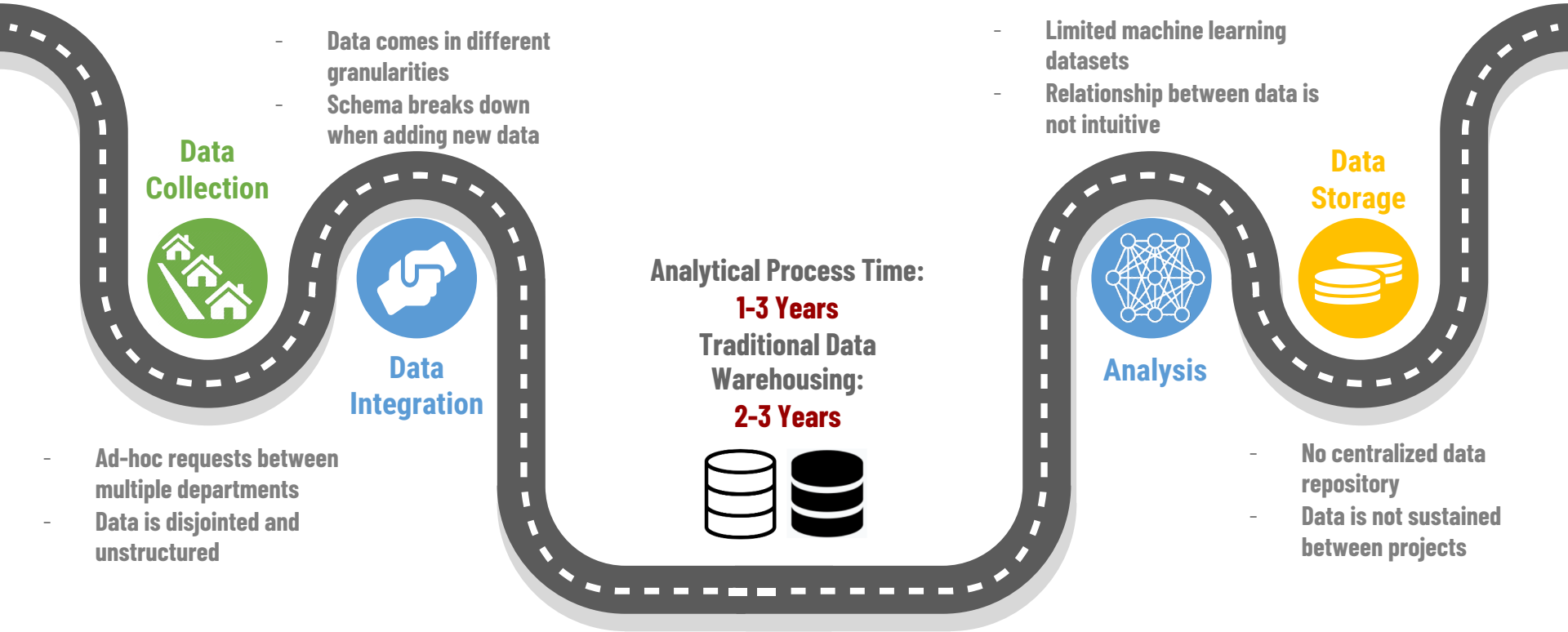






Education	Employment	Health	Housing	Income	Justice	Transportation
Improve school stability and consistency for youth involved in county systems	**Targeted recruitment to increase hiring opportunities at entry and leadership levels**	**Improve access to culturally specific mental health services and increase community wellbeing**	**Improve stable financing opportunities**	**Reform fees**	**Reform probation practices**	**Strengthen infrastructure through ongoing investment**
Increase early childhood education opportunities for youth of color	Improve retention rates at all levels of employment	Increase healthy births and create positive early starts	Increase housing stability through reducing evictions	Support pay equity	Reform countywide law enforcement charging and diversion practices	Enhance mobility options to ensure access to jobs
Increase post-secondary education and training opportunities	Increase advancement opportunities	Empower community to reduce chronic disease	Increase housing supports to reduce homelessness	Provide greater financial inclusion	Eliminate economic barriers in the justice system	Proactive management of activities that positively impact the environment and community

However, the Current Analytical Process is Lengthy & Inefficient



What type of database will minimize Hennepin County's analytical process lead time, and how can the database's capabilities help to reduce disparities in the County?

Agenda

1. Scope & Complexity of Data Requirements
2. Why Choose a Neo4j Graph Database
3. Modeling in Neo4j
4. Neo4j Live Demonstration
5. Implementation Considerations
6. Q&A -> Microsoft Teams Chat -> Prefix Questions with [Q&A]

Scope & Complexity of Data Requirements

The Data at Hand is Complicated and Difficult to Connect & Use

Housing Data
Census Tract Level

Tract ID	Number of Evictions
1	26
2	30

Average Income
Census Tract Level

Tract ID	Average Household Income (USD)
1	50000
2	30000

Join on Census Tract

Tract ID	Number of Evictions	Average Household Income (USD)
1	26	50000
2	30	30000

Different Data Granularities Create Integration Problems

Joined Housing and Income
Census Tract Level

Tract ID	Number of Evictions	Average Household Income (USD)
1	26	50000
2	30	30000

Homeless Shelters
Latitude/Longitude?

X	Y	Name
479487.4525	4980471	People Serving People
478151.2432	4977681	Simpson Emergency Shelter

Correlation with School Performance?
School District Level

District ID	School Name	Average ACT Score
273	Edina High School	28
833	East Ridge High School	24

County Level Data?
Cities?
Different Years?

Quickly Becomes Unmanageable When Data Scope Expands

Update Frequency	Employment	Education	Income	Health	Housing	Justice	Transportation	Misc.

City | Coordinates | Census Tract | County | School District | N/A

* Please refer to the Knowledge Transfer Document for a comprehensive data dictionary

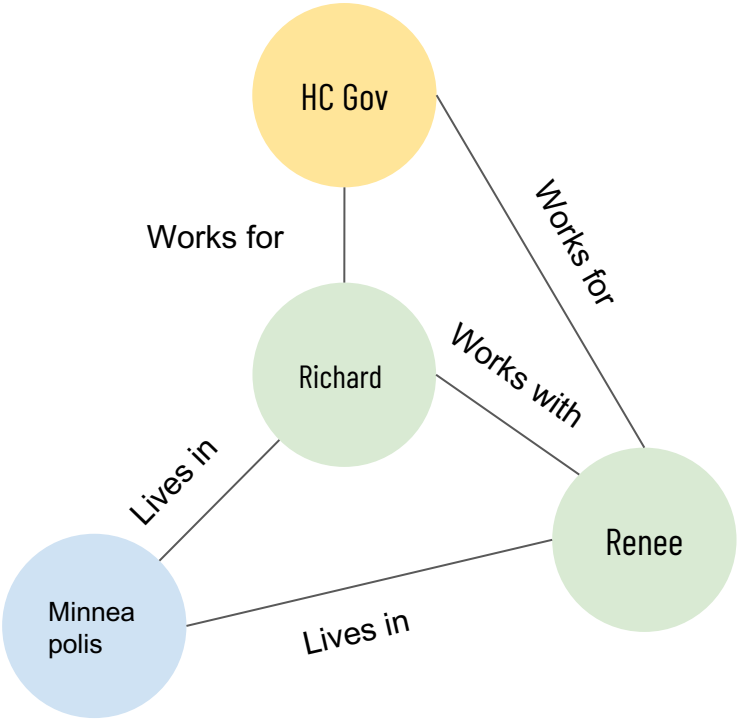
Why Choose a Neo4j Graph Database

Graph Databases are the Best Type of Database to Resolve Hennepin County's Pain Points

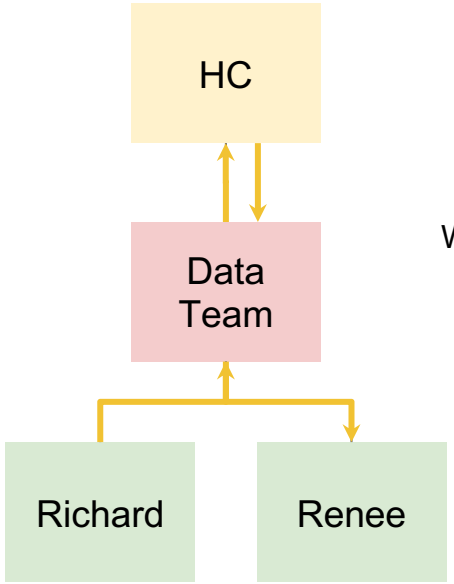
Pain Points	RDBMS	Aggregate Store	Graph DB
Ingest data with different granularities	Complex ETL process	Yes, but can produce duplicated data	Simply add new relationship & nodes
Integrate data without breaking existing schema	Lengthy schema migration process	Yes, but computation cost increases when querying past a single aggregation	Schema-free nature & flexible relationships
Intuitive visualization of data structure	User needs an understanding of complex ERDs	Built for high-velocity & live applications; data is stored for speed not connectivity	Yes, facilitates communication between tech and non-tech users

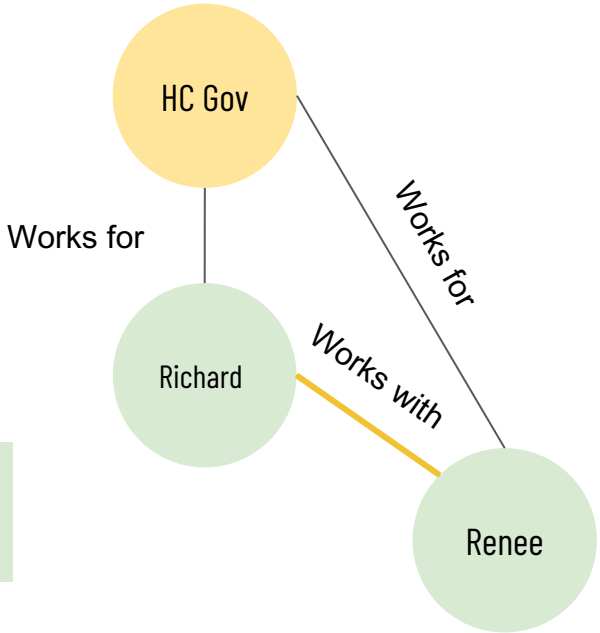
Brief Overview of a Typical Graph Database Schema

Graph DB stores both node & relationship



Graph DB accelerates query speed for highly interconnected data

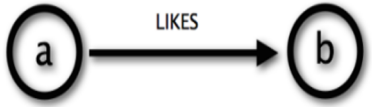




More Reasons Why Neo4j is Ideal

Ease of Programming

- Intuitive programming language
- Easy to learn
- Complex queries



Cypher

```
(a) -[:LIKES]-> (b)
```

Community Resources

- Open source
- Rich online resources and tutorials

GraphAcademy
Learn. Graph. Deploy.

Graph Algorithm

- Community detection
- Graph similarity algorithms
- PageRank algorithm

Neo4j is the Most Popular Graph Database on the Market and is Used by Top Companies

> Neo4j is the most popular Graph DBMS among developers.
>
> - DB Engines

Ref: https://db-engines.com/en/ranking/graph+dbms

Expanding on Why Neo4j Will Resolve the County's Pain Points

Pain Point	Neo4j Solution
Data engineering must be conducted each time an ad-hoc analysis is needed	Accumulates the data engineering value from analysis to analysis
Connecting data that comes at a variety of time & spatial granularities	Richly connects disparate domains with intuitive relationships
Integrating data without breaking the existing normalized schema	Graph model is built to evolve as more data & domain knowledge are obtained
Normalized ERD does not map intuitively to reality	Graph model facilitates better communication between technical & non-technical stakeholders

Modeling in Neo4j

How to Model in Neo4j

Business Need	Add Nodes	Add Attributes	Define Relationship
Integrate data from **7 domains** with **different granularities** into a database to address disparities issue	Things can be labeled or grouped E.g. **county, tract, city, hospital**	Data in each nodes' level E.g. **mean income, total population, HDI index**	Connections between two nodes E.g. county **serves** tract, hospital **located in** tract, tract **adjacent to** tract

Basic Modeling

Add Nodes → Add Attributes → Define Relationship

Conceptual Modeling

County

Attributes
Name
Total Pop

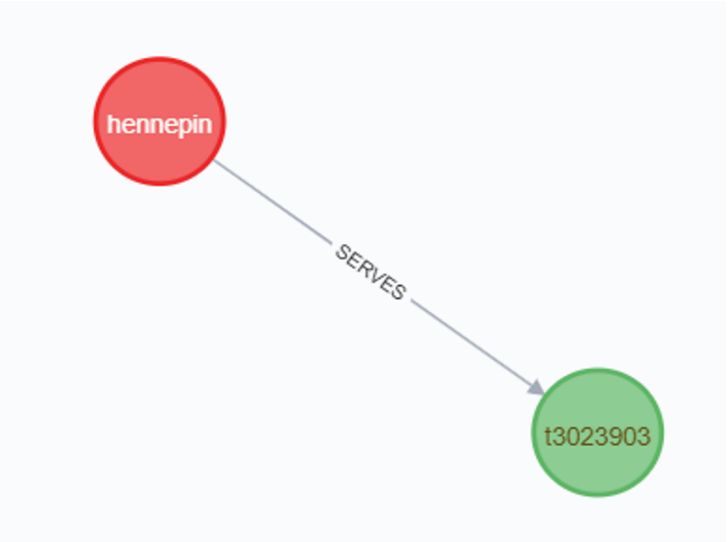
Serves

Tract

Census Tract 1004

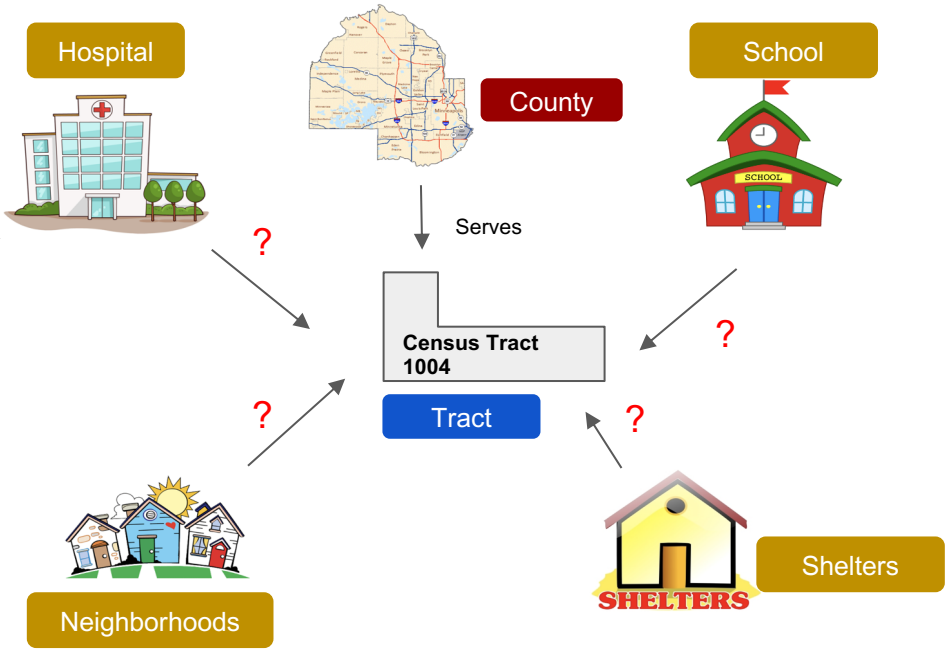
Attributes
Name
Median Income

Actual Modeling in Neo4j

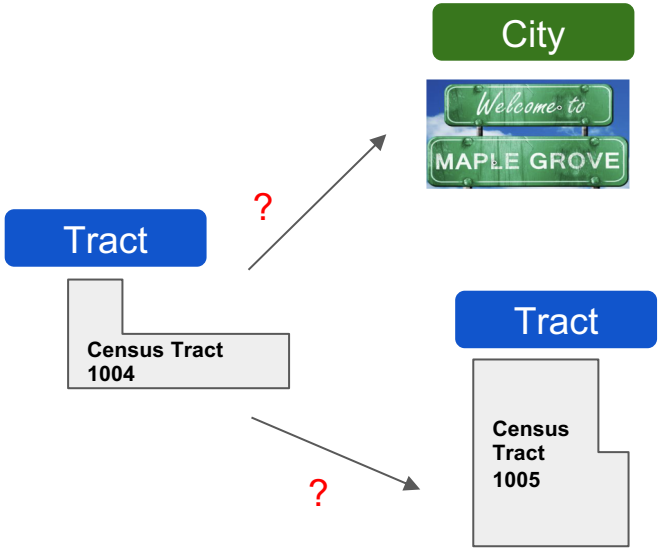


How to Accommodate Spatial Relationships with a Graph Database?

Issue 1: Relationship Between Locations and Areas



Issue 2: Relationship Between Areas and Areas



Spatial Technical Solution - Adding Spatial Attributes into Database

Issue : How to add more spatial information

Solution: Add coordinates from WKT data

Shapefile from GIS Open Data

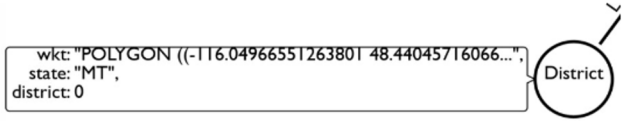
Data Engineering & Data Ingestion

neo4j

Nodes in Neo4j Database

Geometry Type	WKT	Object
Point	'POINT(-93.3 43.6)'	Hospital
Polygon	'POLYGON({-94.2 42.9, -93.6 42.4, -93.8 43.2})'	District

Nodes

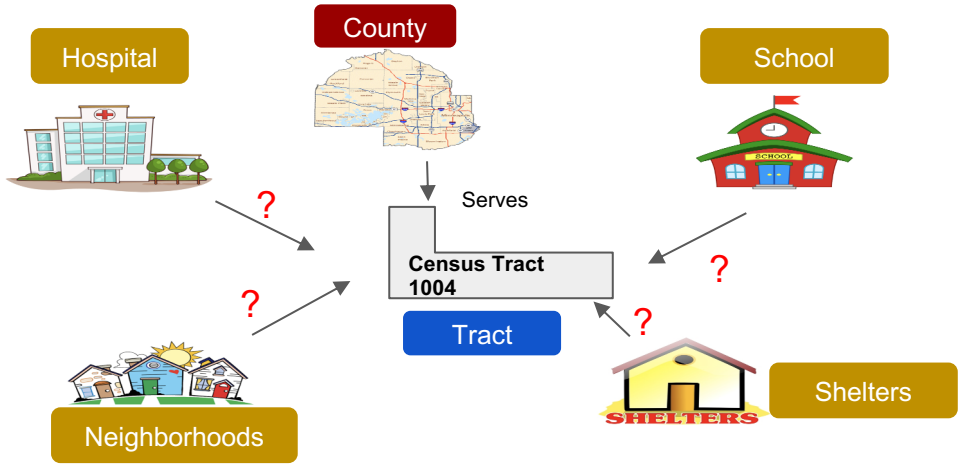


* Please refer to the KTD for the full script

Spatial Technical Solution - Relations Between Locations and Areas

Issue 1: Relationship Between Locations and Areas

Solution: Neo4j Spatial Plugin



Add Hospital Nodes

```
LOAD CSV WITH HEADERS FROM "file:///HC_hospitals.csv" AS row
CREATE (:Hospital {HospitalId: row.OBJECTID AS, Hospital_Name:row.NAME,
        Hopital_Address:row.ADDRESS, City:row.CITY, ZIP:row.ZIP,
        Hospital_Type:row.LIC_TYPE, HOSP_BEDS:row.HOSP_BEDS,
        Longitude:row.POINT_X, Latitude:row.POINT_Y}
```

Match Hospital with Tract Using Spatial Plugin WithinDistance Function

```
MATCH(h:Hospital)
CALL spatial.withinDistance('TractGeom',{latitude: toFloat(h.Latitude),
                                longitude:toFloat(h.Longitude)},0.01)
YIELD node AS t
```

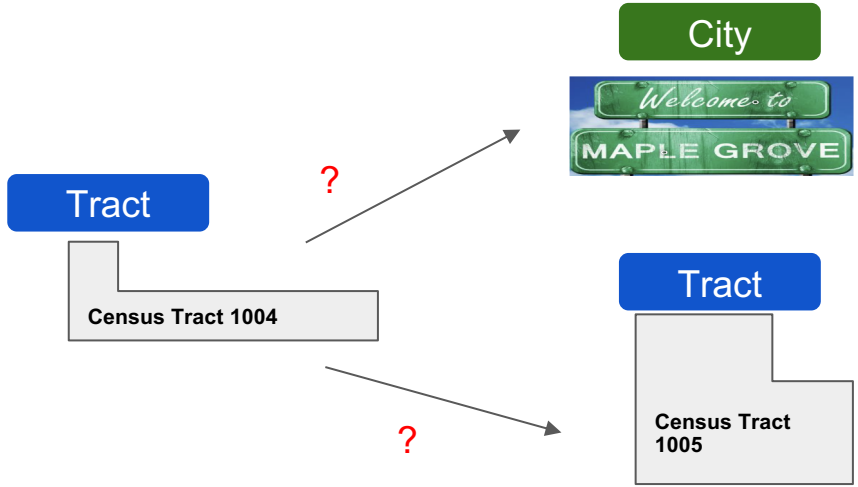
Define the relationships

```
MERGE (h)-[rel:locate_in]->(t)
return count(rel)
```

Spatial Technical Solution - Relationships Between Areas and Areas

Issue 2: Relationship Between Areas and Areas

Solution: Neo4j Spatial Plugin



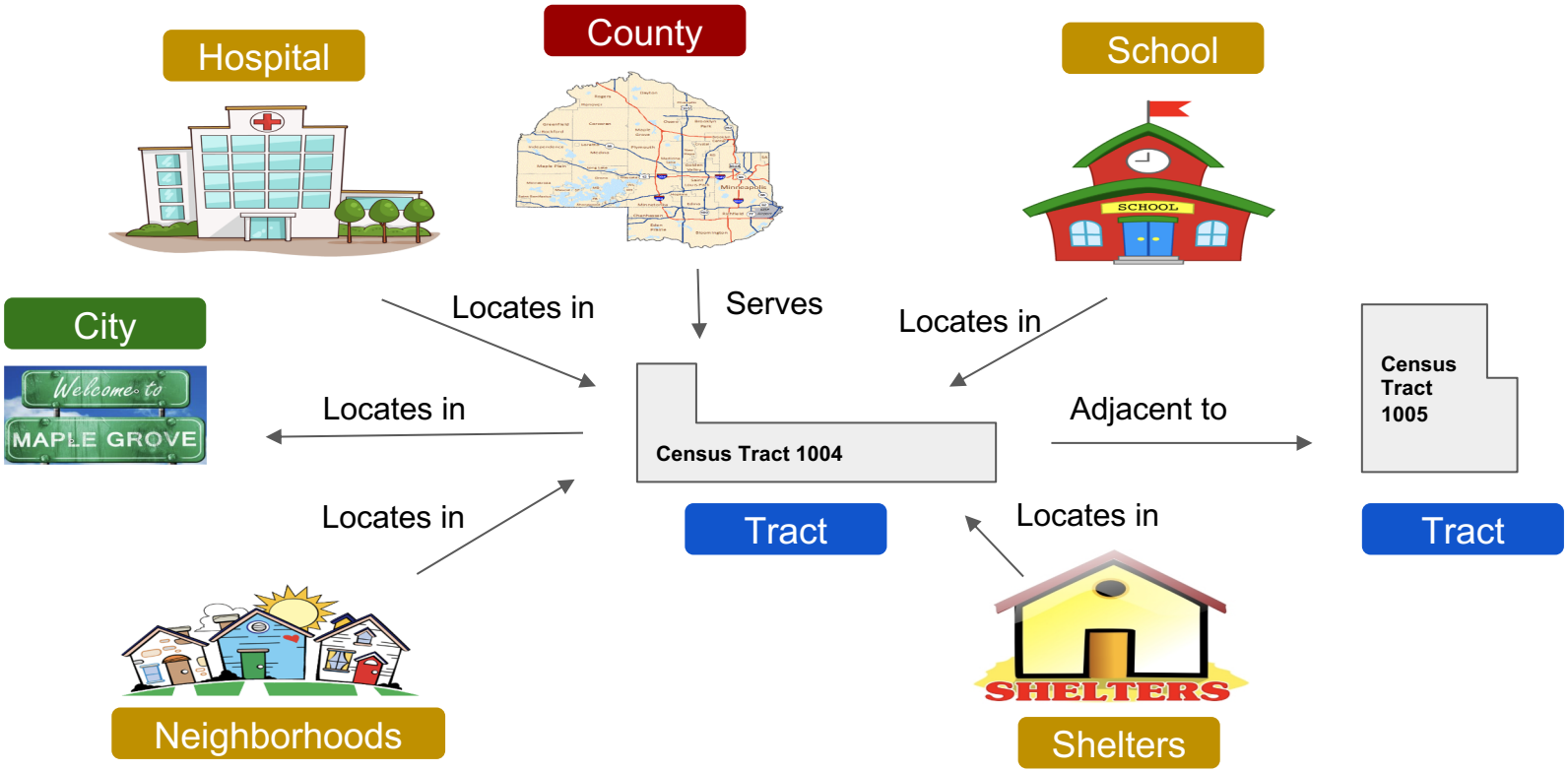
Match City with Tract Using Spatial Plugin WithinDistance Function

```
MATCH(t:Tract)
CALL spatial.withinDistance('CityGeom',{latitude: toFloat(t.Latitude),
                                        longitude:toFloat(t.Longitude)},0.01)
YIELD node AS n
MERGE (t)-[rel:locate_in]->(n)
return count(rel)
```

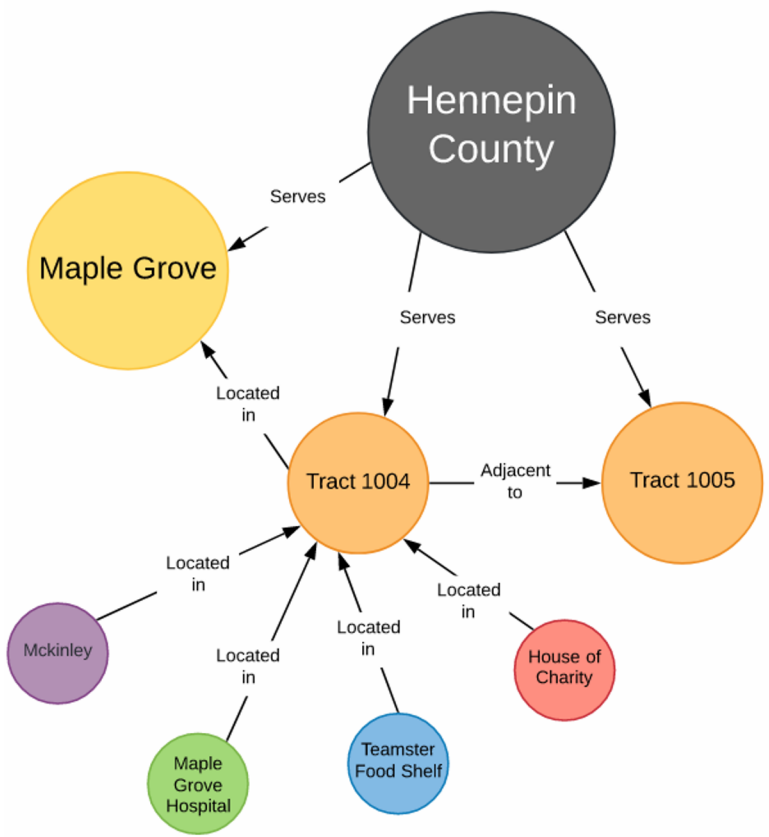
Match Tract with Tract Using Spatial Plugin Intersects function

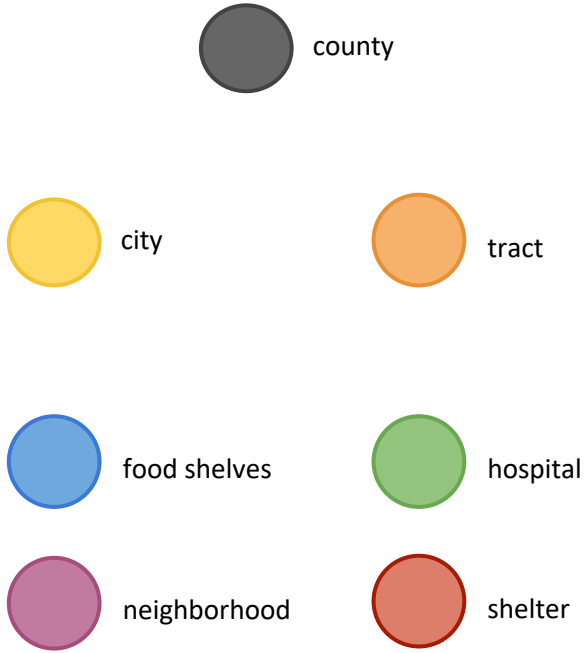
```
MATCH(t:Tract)
CALL spatial.intersects("TractGeom",t.TractWKT)
YIELD node AS n
WHERE n.TractName <> t.TractName
MERGE (t)<-[rel:adjacent_to]->(n)
```

Conceptual Database with Spatial Nodes and Relationships



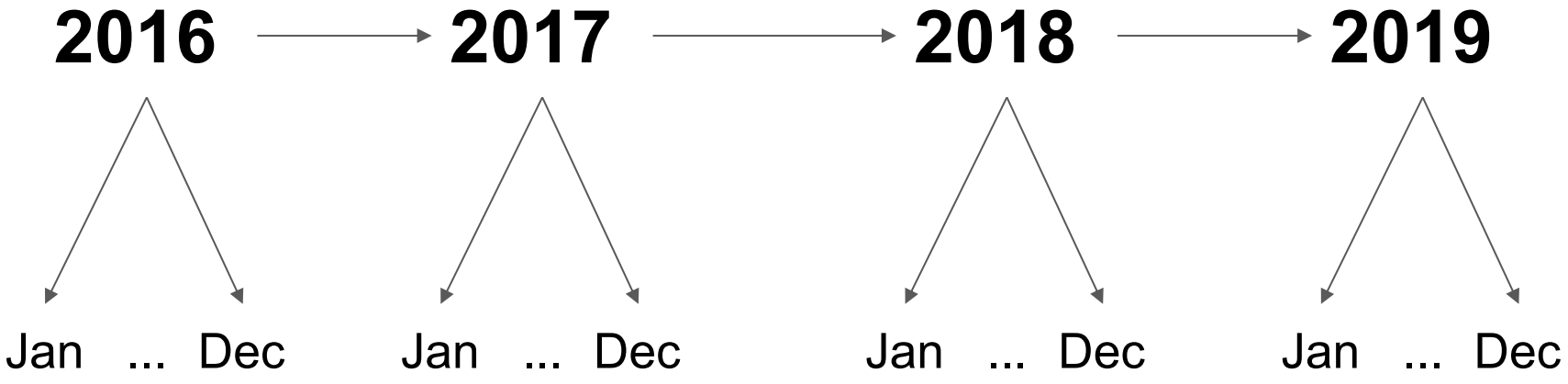
Evolving Database with Spatial Nodes and Relationships



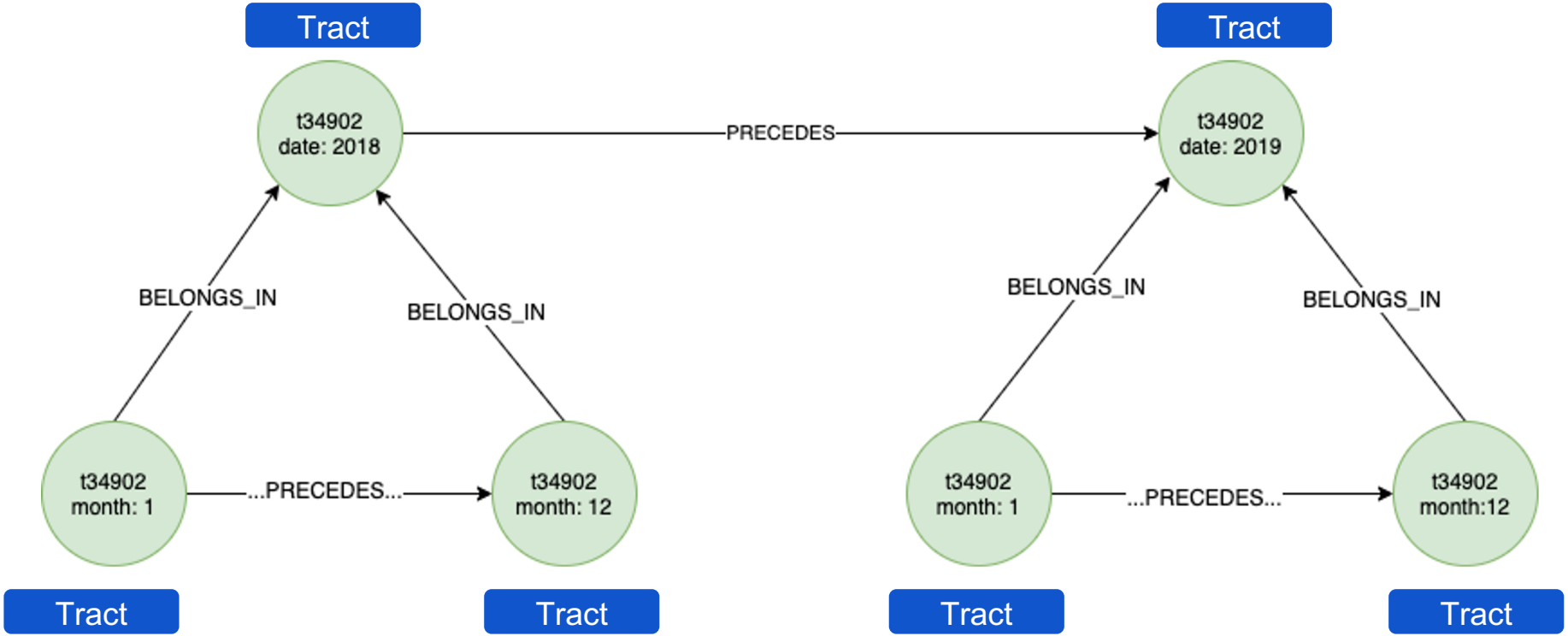


* Please refer to the KTD for the full script

How to Accommodate Historical Data with a Graph Database?



Temporal Technical Solution



Neo4j Live Demonstration

What Would it be like for an Analyst to Inform a Disparity Initiative with this Graph Database?

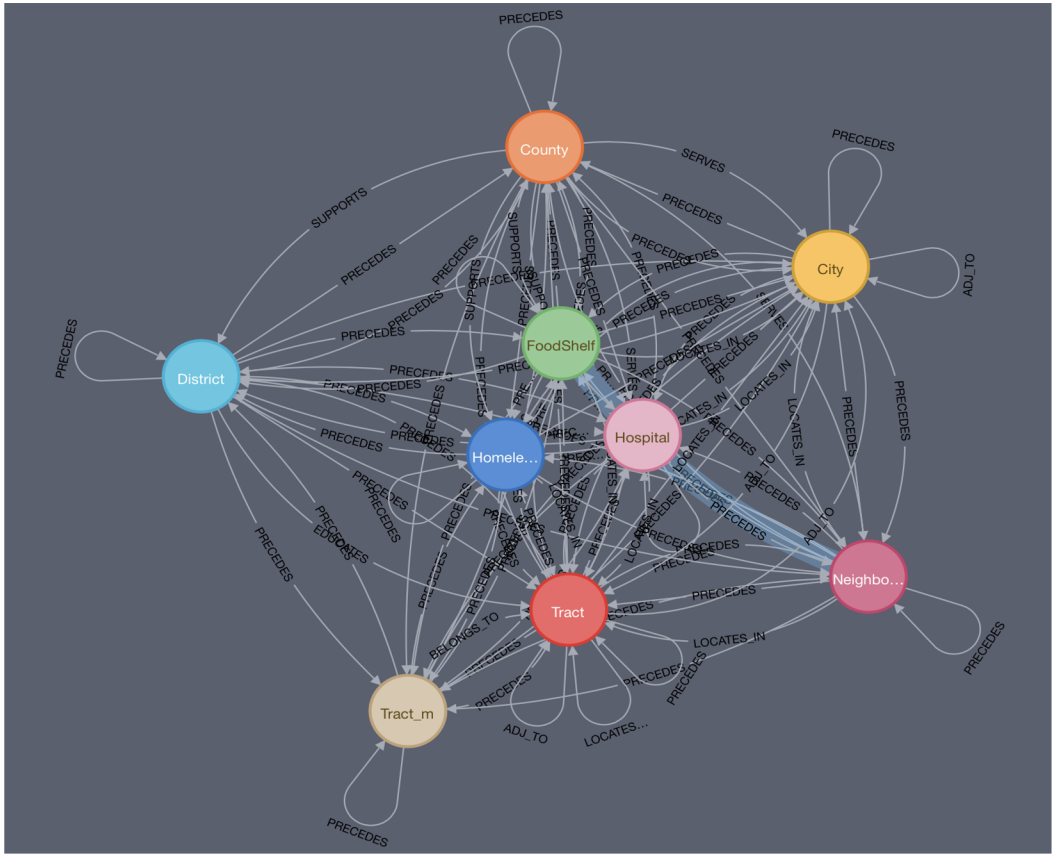
Domain: Housing

Initiative: Increase housing supports to reduce homelessness

Profiling the Graph Database

Schema Summary Stats

- 16,344 Nodes
- 9 Different Types of Nodes
- 36,878 Relationships
- 150+ Unique Node Properties



Profiling the Graph Database

County

City

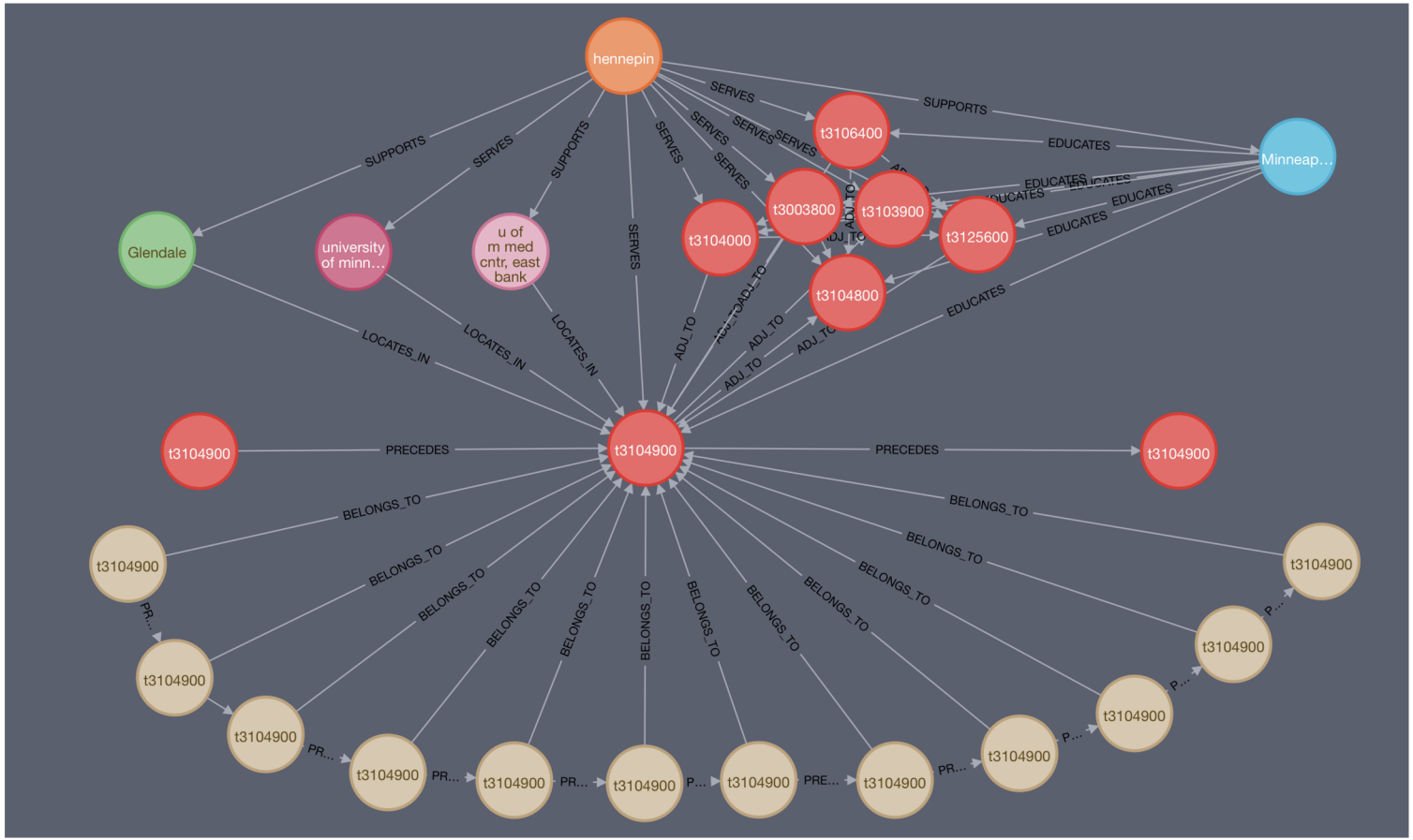
Tract

Tract- Monthly

Neighborhood

Food Shelf

Hospital

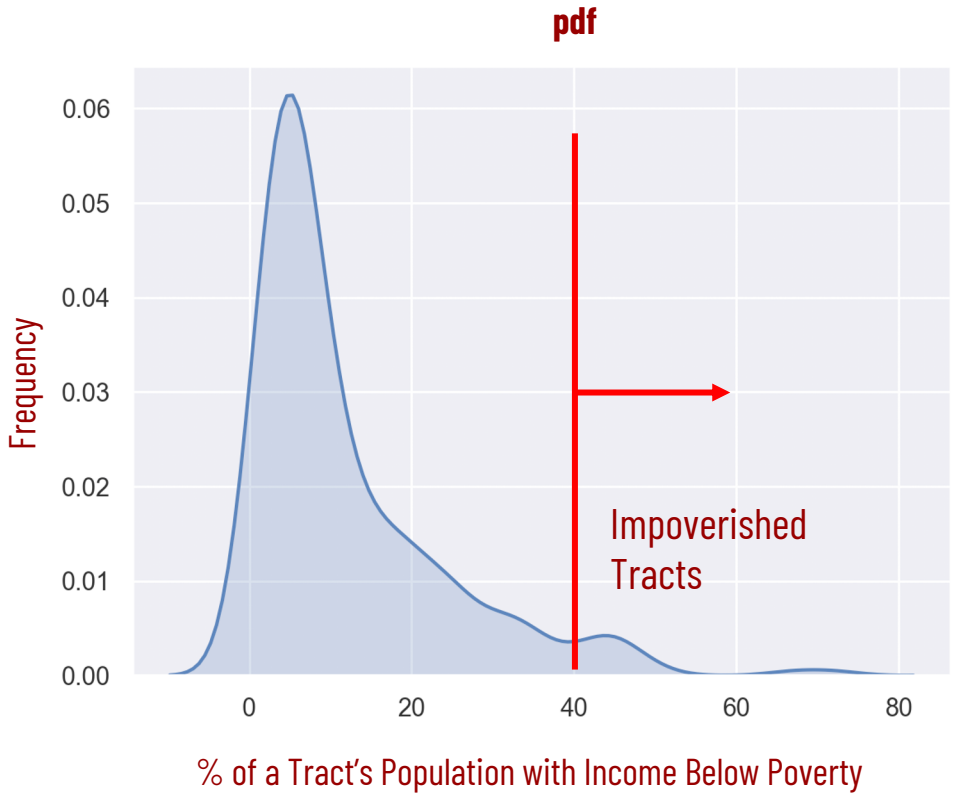


Logical Chain of Analysis for Demo

- Which tracts in Hennepin County are most at risk of homelessness?
  1. Is there a pocket of these tracts in Hennepin County adjacent to one another?
  2. What housing support resources are at the disposal of Hennepin County?
  3. Are the necessary housing supports in place for this large pocket of adjacent at risk tracts?

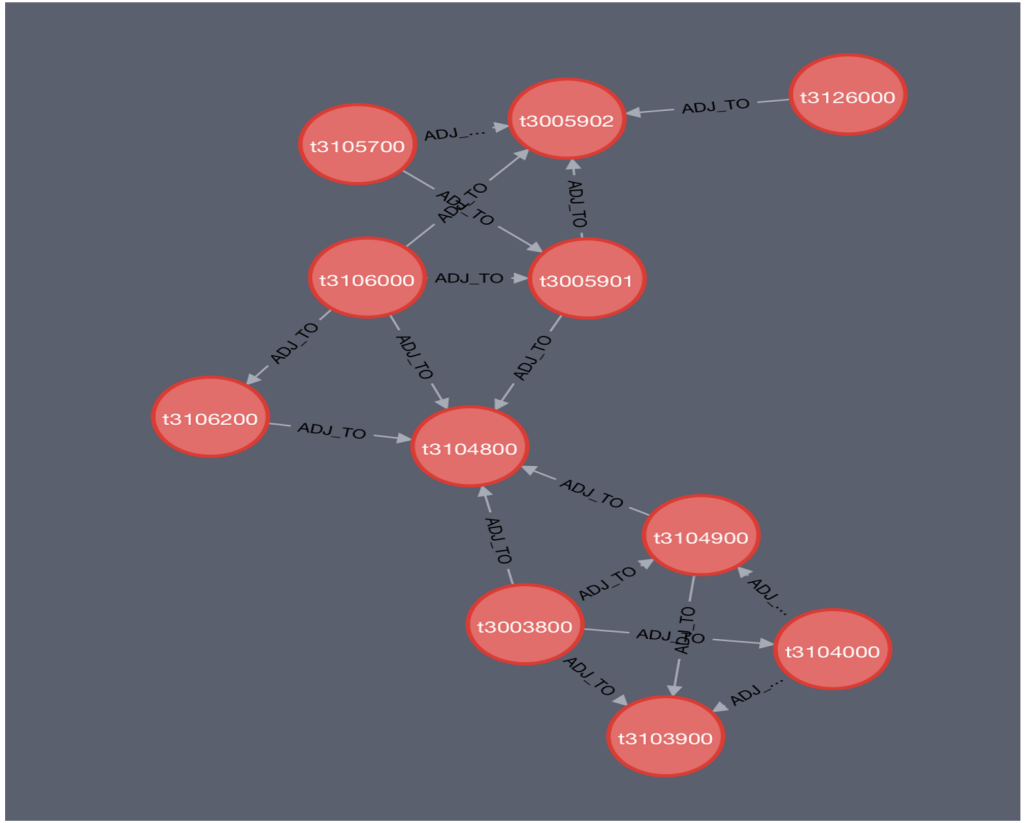
Q: Which Tracts in Hennepin County are Most at Risk of Homelessness?

A: There are 14 impoverished tracts within Hennepin County



Q: Is there a Pocket of these Tracts in Hennepin County Adjacent to One Another?

A: 11 out of the 14 impoverished tracts are adjacent to each other



Q: What Housing Support Resources are at the Disposal of Hennepin County?

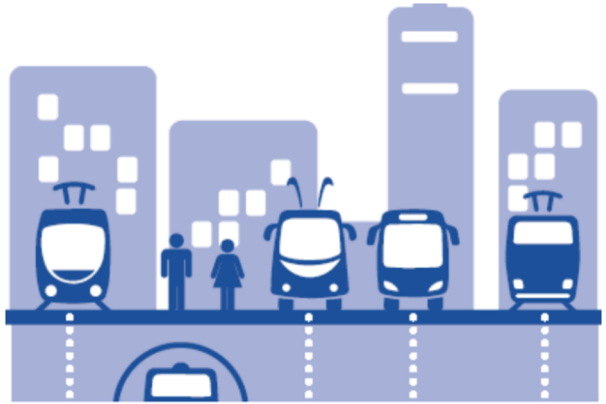
Housing Support Resource	Total Available
Average % of housing units subsidized within a tract	5.8%
# of Homeless Shelters	19
# of Food Shelves	48

Q: Are the Necessary Housing Supports in Place for this Large Pocket of at Risk Tracts?

Housing Support Resource	Amount Allocated to Pocket of 11 Impoverished Tracts
Average % of housing units that are subsidized within an impoverished tract	25.6%
# of Homeless Shelters	1
# of Food Shelves	6

A: Only **6** out of **48** food shelves & **1** out of **19** homeless shelters are allocated to this pocket of 11 adjacent impoverished tracts

4 Things to Consider When Implementing the Database



Add more transportation data

Community Detection Algorithms

Centrality Algorithms

After adding individual-level datasets, explore graph algorithm applications

```
CREATE CONSTRAINT ON (t:Tract) ASSERT t.name is UNIQUE
CREATE CONSTRAINT ON (t:Tract) ASSERT exists(t.hdi_index)
CREATE CONSTRAINT ON (t:Tract) ASSERT (t.name, t.date) IS NODE KEY
```

To ensure data integrity & to minimize data redundancy, enforce constraints on the graph database

Microsoft Azure

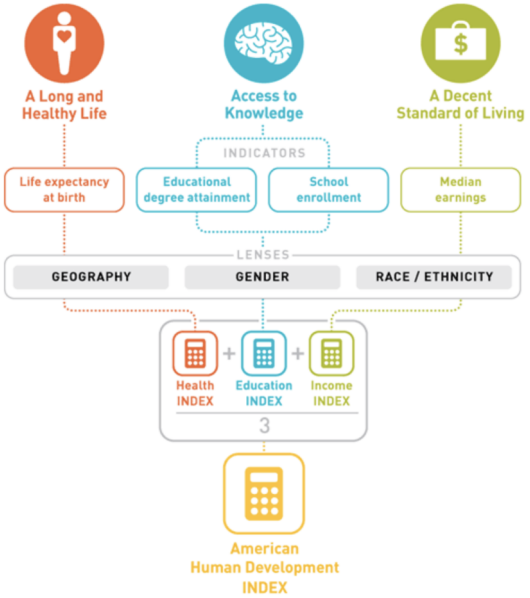
Refer to KTD on how to use Neo4j in Azure

Recap

- Why Choose a Neo4j Graph Database
- Modeling in Neo4j
- Capabilities of the Neo4j Graph Database
- Implementation Considerations

Stick Around for the Following Use Case Presentations!

SM SNAP Supplemental Nutrition Assistance Program



- **SNAP Market Penetration**: Graph DB enables Hennepin County to create a variety of views that would be difficult to create in an RDBMS
- **HDI Predictive Modeling**: Graph DB enables Hennepin County to accumulate datasets well-suited for ML

Thank You, Hennepin County!

Q&A