**Homework3 ---- SHOUKATH ALI (shshaik@iu.edu)**

**Part 1: Data Modeling**

**Data Description -**

For this project, I've chosen a dataset that provides demographic information about software developers, including their age, gender, country, development environment, languages and technologies they work with, and their preferred social media platforms. This data was gathered to analyze the trends and patterns within a developer community.

The dataset is in CSV format and includes the following columns: Id, Age, Gender, Country, DevEnvironment (Development Environment), LanguageWorkedWith, WebFrameWorkedWith, DatabaseWorkedWith, PlatformWorkedWith, and SocialMedia. Each row in the CSV file represents an individual developer's profile.

Kaggle - <https://www.kaggle.com/datasets/shariful07/demographics-technologies?resource=download>

(reduced the data manually to improve processing speed, attached csv file for reference)

The rationale behind choosing this dataset is to explore the relationships between different developers’ environments, skills, and preferences. The goal is to understand how these factors might be interrelated and could potentially inform decisions in technology stack choices, hiring, and community building efforts.

**Arrow Schema**

**A diagram of a software company

Description automatically generated**

The schema for the dataset was designed using the ARROWS tool to create a visual representation of the nodes, relationships, and properties. The following concepts are included in the graph model:

Developer Node: Represents an individual developer with properties extracted from the dataset such as age, gender, and country.

DevelopmentEnvironment Node: Represents the development environment used by the developer, with a relationship indicating what environment a developer works in.

SoftwareApplications Node: Represents the set of software applications, languages, and databases the developer is accustomed to, with relationships showing what the developer works with.

Platform Node: Represents the platform(s) the developer works on, with a relationship indicating the specific platform.

SocialMedia Node: Represents the social media platforms used by the developer, with a relationship to show which social media the developer uses.

Relationships are defined to show how developers are connected to their tools and platforms, thus creating a rich graph of the developer ecosystem.

Part 2

Loading data –

// Loading csv file

LOAD CSV WITH HEADERS FROM "file:///Users/shoukath/Downloads/demographic.csv" AS line

WITH line

// Create a Developer node for each row in the CSV

MERGE (dev:Developer {id: line.id})

ON CREATE SET

dev.age = toInteger(COALESCE(line.Age, '22')),

dev.gender = COALESCE(line.Gender, 'Unknown'),

dev.country = COALESCE(line.Country, 'Unknown')

// For each Developer, link to a DevelopmentEnvironment node

MERGE (env:DevelopmentEnvironment {name: COALESCE(line.DevEnviron, 'Unknown')})

MERGE (dev)-[:WORKS\_IN]->(env)

// For each Developer, link to the SoftwareApplications they work with

MERGE (soft:SoftwareApplication {language: COALESCE(line.LanguageWorkedWith, 'Unknown')})

MERGE (dev)-[:WORKS\_WITH]->(soft)

// For each Developer, create relationships to the Platform node

MERGE (plat:Platform {name: COALESCE(line.PlatformWorkedWith, 'Unknown')})

MERGE (dev)-[:WORKS\_ON]->(plat)

// developer uses social media, create and link a SocialMedia node

MERGE (social:SocialMedia {platform: COALESCE(line.SocialMedia, 'Unknown')})

MERGE (dev)-[:USES]->(social)

// Set additional properties for SoftwareApplication nodes from CSV

ON CREATE SET soft.webFramework = COALESCE(line.WebFrameWorkedWith, 'Unknown'),

soft.database = COALESCE(line.DatabaseWorkedWith, 'Unknown')

// Set properties for Platform nodes from CSV

ON CREATE SET plat.os = COALESCE(line.PlatformWorkedWith, 'Unknown')

// Index creation for faster lookup, should be run separately and only once

// CREATE INDEX ON :Developer(id)

// CREATE INDEX ON :DevelopmentEnvironment(name)

// CREATE INDEX ON :SoftwareApplication(language)

// CREATE INDEX ON :Platform(name)

// CREATE INDEX ON :SocialMedia(platform)

COALESCE is used to replace null or empty data with some identifiable value.

Nodes and relationships -

A screenshot of a computer

Description automatically generated

Database information -

A screenshot of a computer

Description automatically generated

PART 3 –

Lets, start finding the required developer for the client who satisfies all the eligibility conditions, using the database below.

In Database visualization, we can say that we have 5 nodes and 4 relationships .

CALL db.schema.visualization()

A screenshot of a computer

Description automatically generated

Lets Explore -

Now, lets find how many values or components each node has,

MATCH (n)

RETURN labels(n) AS NodeType, count(\*) AS Total

A screenshot of a computer

Description automatically generated

It's used to count the total number of nodes for each label (node type) present in the Neo4j graph database. Here's a breakdown of the results displayed:

Developer Node: There are 241 nodes labeled as Developer. This indicates a sizable group of nodes representing individual developers or entities with developer-related attributes in the graph.

DevelopmentEnvironment Node: There are 17 nodes labeled as DevelopmentEnvironment. These nodes are likely used to represent various development environments that developers work in, such as IDEs, coding environments, or platforms.

SoftwareApplication Node: There are 18 nodes with the label SoftwareApplication. These could be representing different software applications, programming languages, or technologies that developers are using or are skilled in.

Platform Node: Also at 18 nodes, those labeled as Platform might represent the operating systems, server platforms, or computing platforms that the developers are working on.

SocialMedia Node: The smallest group, with 12 nodes labeled as SocialMedia, suggests that these nodes are used to represent various social media platforms that developers might use, perhaps for communication, marketing, or community engagement.

Count of total relationships –

MATCH ()-[r]-()

RETURN type(r) as RelationshipType, count(\*) AS Total

A screenshot of a computer

Description automatically generated

We got count of all the relationship same i.e 482, because each developer is connected with all the other nodes.

The table in the output consists of two columns: 'RelationshipType' displaying the type of the relationship, and 'Total' showing the count of relationships for each type. All relationship types have the same count, suggesting a uniform distribution across the types of relationships within this dataset.

Performing basic cyper queries –

Average age of developers data we have –

MATCH (d:Developer)

RETURN avg(d.age) as AverageAge

A screenshot of a computer

Description automatically generated

Average age of developers is – 30

Developers by Country and Sort by Age -

MATCH (d:Developer)

WHERE d.country = 'United States'

RETURN d.id, d.age

ORDER BY d.age

A screenshot of a computer

Description automatically generated

From the data, The min age of developer in united states is 21

Development Environments Used by Developers Above a Certain Age –

MATCH (d:Developer)-[:WORKS\_IN]->(env:DevelopmentEnvironment)

WHERE d.age > 30

RETURN env.name, count(d) as NumDevelopers

ORDER BY NumDevelopers DESC

A screenshot of a computer

Description automatically generated

As not of the data is unknown, its ranked first, let’s consider the second one – Visual studio code was the preferred development environment.

Most Popular Social Media among Developers –

MATCH (d:Developer)-[:USES]->(s:SocialMedia)

RETURN s.platform as SocialMediaPlatform, count(d) as NumUsers

ORDER BY NumUsers DESC

A screenshot of a computer

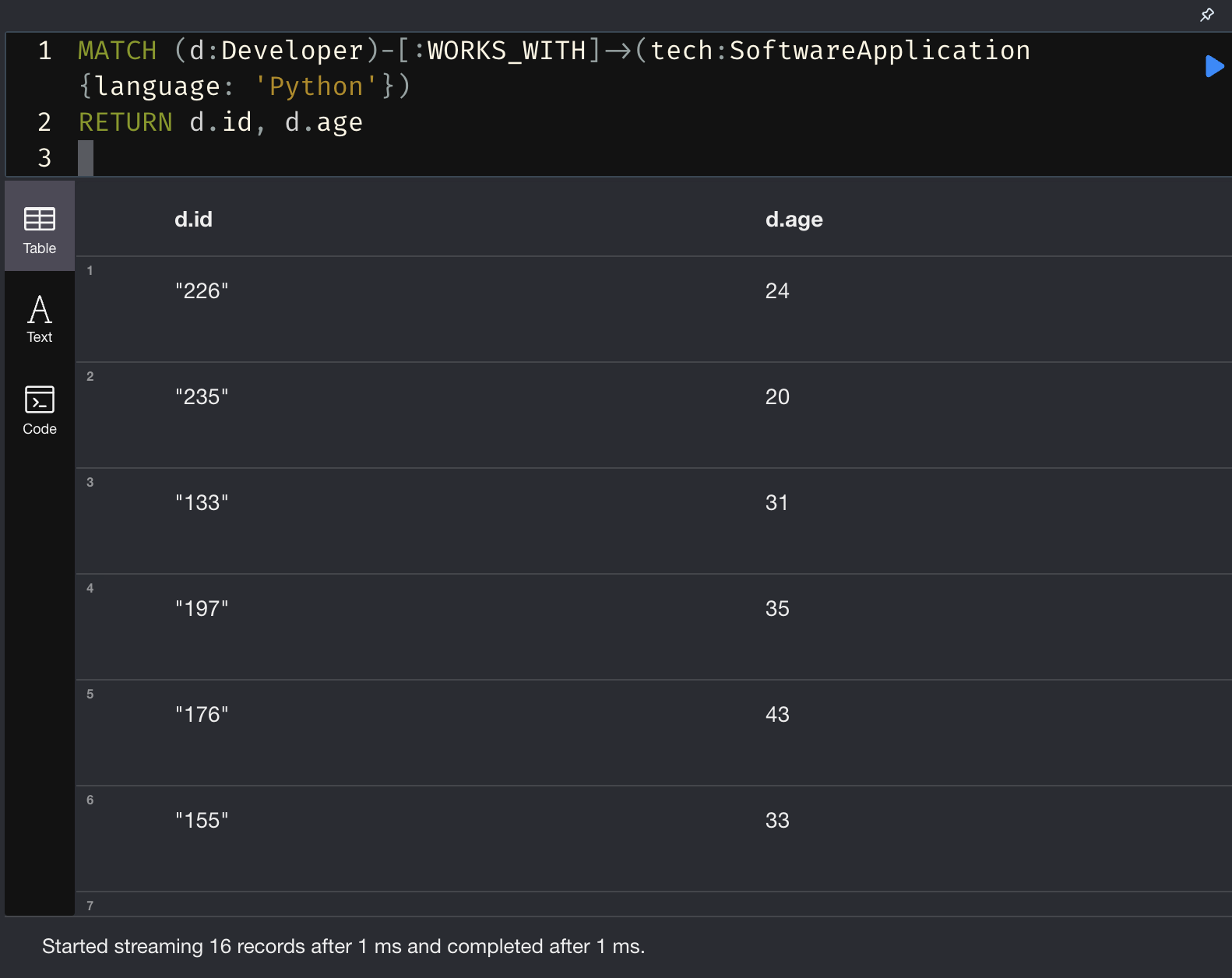
Description automatically generated

Reddit was the most used social media application by developers.

Developers Who Work with a Specific Technology –

MATCH (d:Developer)-[:WORKS\_WITH]->(tech:SoftwareApplication {language: 'Python'})

RETURN d.id, d.age



Looks like 16 people use Python, from the data.

Search for developers who work with both HTML/CSS and SQL database –

// Match developers who have a WORKS\_WITH relationship to any SoftwareApplication

MATCH (d:Developer)-[:WORKS\_WITH]->(tech:SoftwareApplication)

// Collect all the programming languages and databases associated with each developer into lists

WITH d, collect(tech.language) AS Languages, collect(tech.database) AS dbs

// Filter to include only those developers who work with both Python and MySQL

WHERE 'HTML/CSS' IN Languages AND 'MySQL' IN dbs

// Return the developer's name, the list of languages, and the list of databases they work with

RETURN d.id AS DeveloperName, Languages, dbs



Looks like there are 29 developers who work with the combination of the language – HTML/CSS and Databases -MySQL.

---- XXX ---