NoSQL Databases: Neo4j

(and a bit on Graph Analytics)

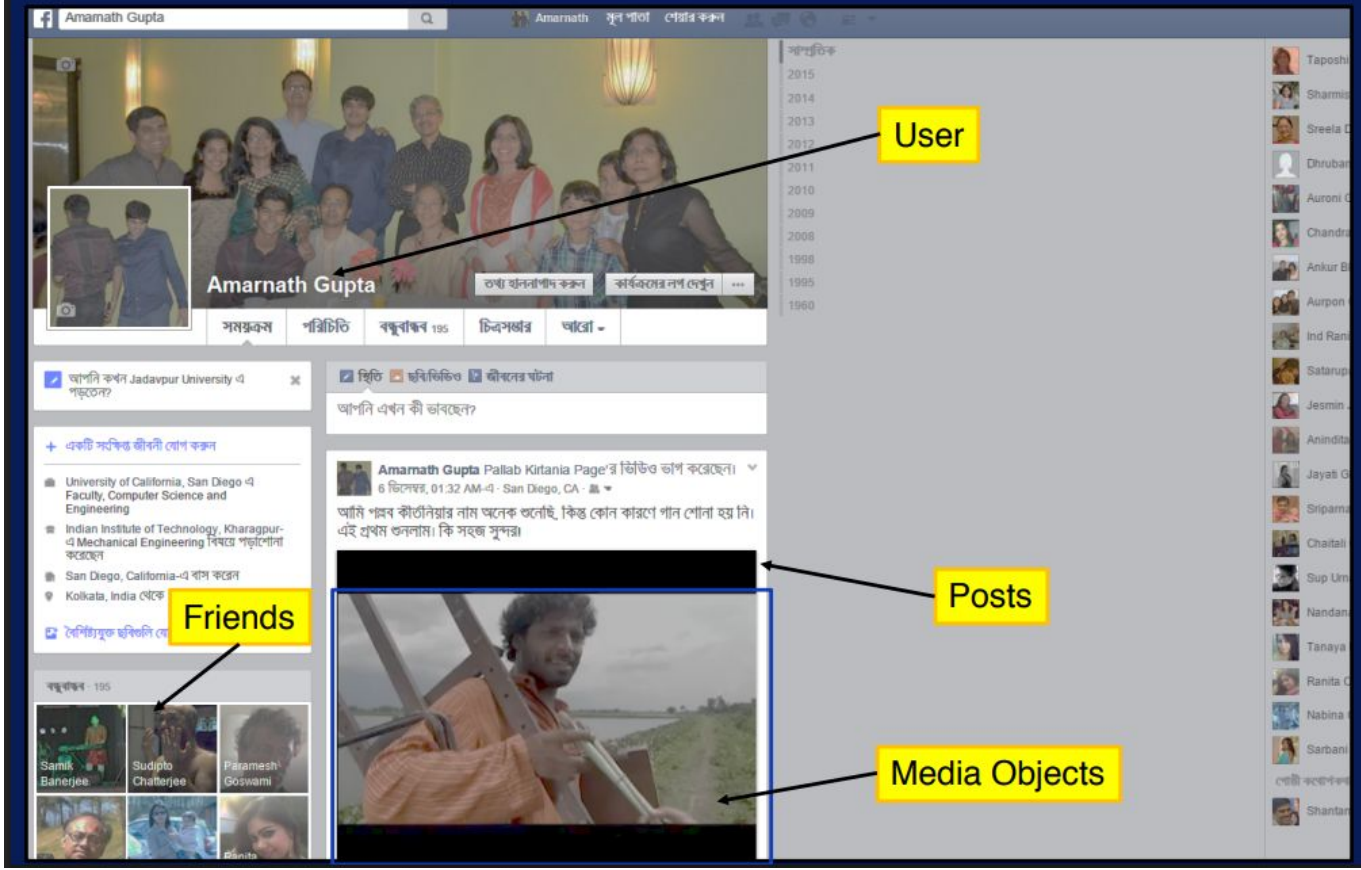
By Mahdis Rahmani

Overview

1. Motivation: Introduction to Graphs, Graph Analytics, Graph Analytics Techniques & Algorithms
2. Introducing NoSQL (and a brief history of DBMS)
3. Working with Distributed data (a flashback to distributed systems)
4. Graph Databases: Neo4j
5. Basics of Neo4j
6. Working with Neo4j: Querying and CRUD operations
7. Computing Platforms for (Big Data) Graph Analytics


Why Graphs?
Case Studies

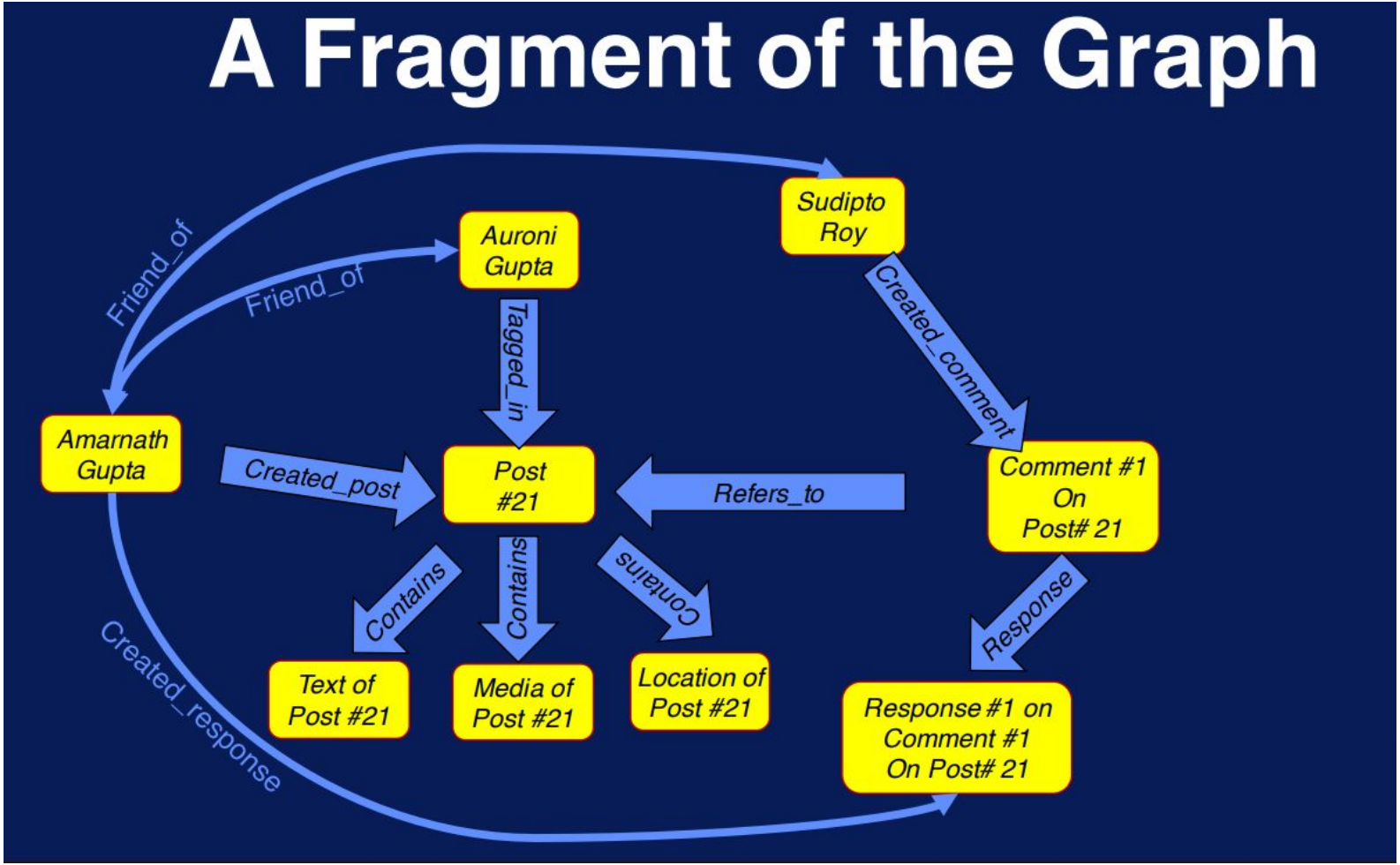
Social Media Produces Graphs



Can you see the graph?

Facebook

A Fragment of the Graph

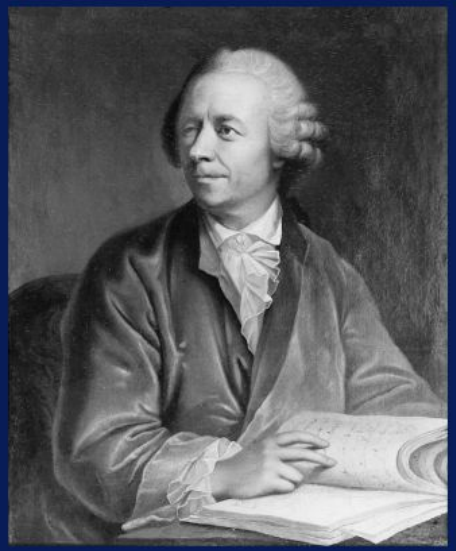


4 Use Cases

- **Example 1**
  - Social Media Analytics
- **Example 2**
  - Gene-Phenotype-Disease Networks
- **Example 3**
  - Human Information Network Analytics
- **Example 4**
  - Analysis and Planning for Smart Cities

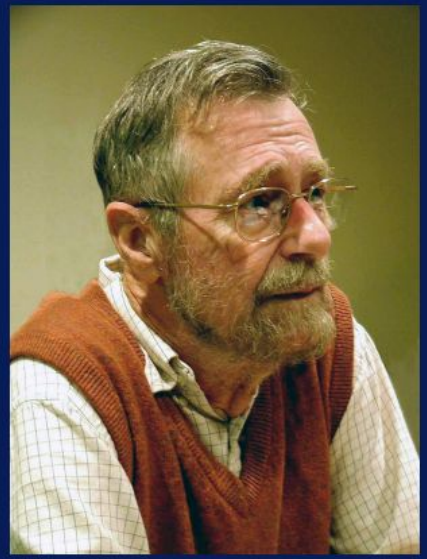
Graphs: Defined Differently by Different People

Mathematicians



Leonhard Euler

Computer Scientists

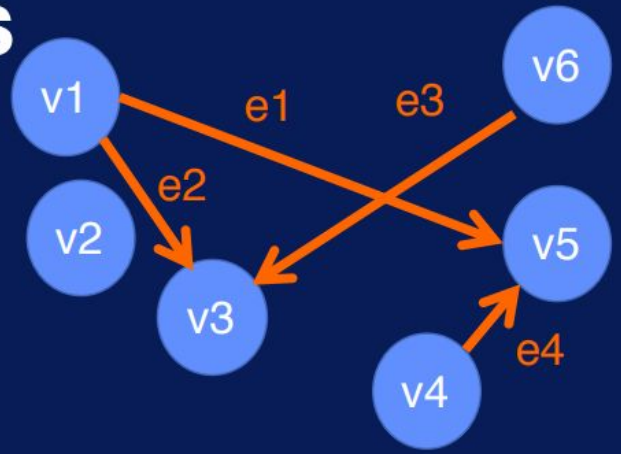


Edsger W. Dijkstra

Mathematical Definition:

- **V: a set of vertices**
- **E: a set of edges**

E = {e1, e2, e3, e4}



e: an edge designates a pair of vertices

E = {(v1,v5), (v1,v3), (v6,v3), (v4,v5)}

What about the Computer Science definition?

- **An abstract data type**

  1) Has a data structure to represent the mathematical graph

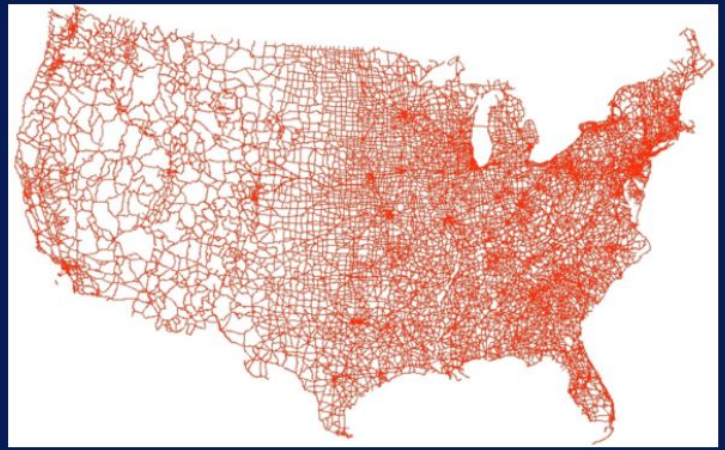
  2) Supports a number of operations (on that graph)
     - Add_edge
     - Add_vertex
     - Get_neighbor (and others)

Graphs and the V's of Big Data

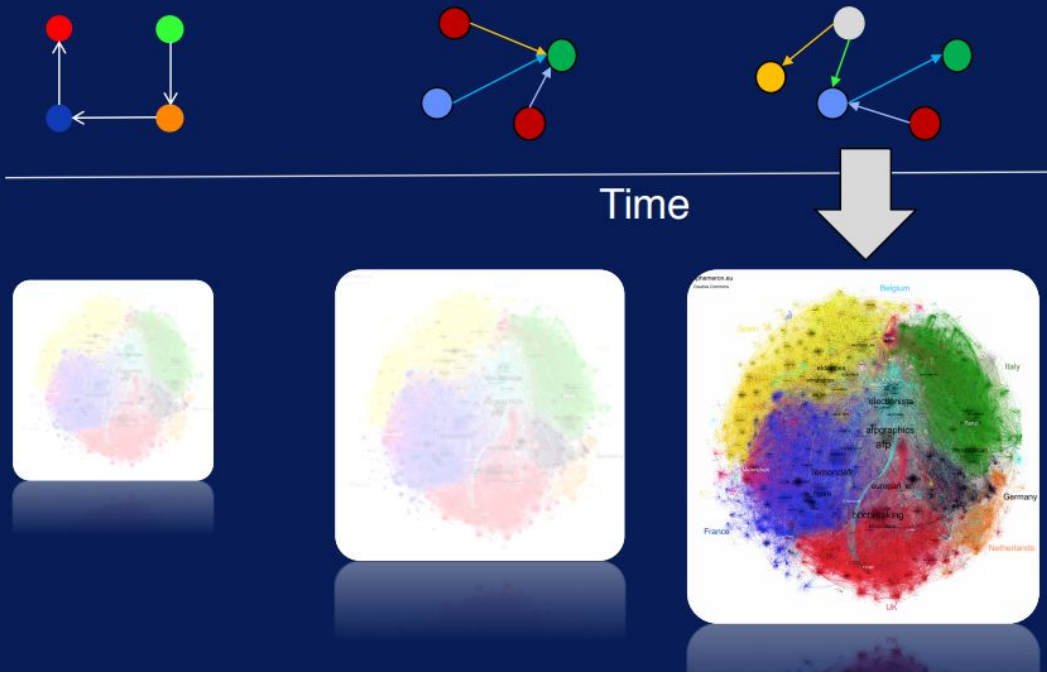
- Volume
- Velocity
- Variety
- Valence

Volume

- **Size**
  - # of nodes and edges

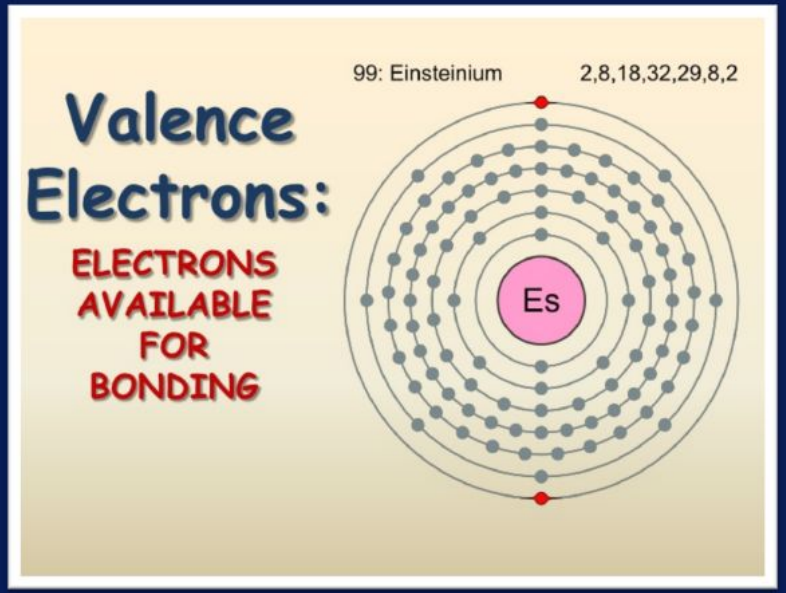


Velocity


Time
Belgium
Italy
Germany
France
Netherlands
UK
electronics
afpgraphics
afp
lemonde1
europarl
hbtbreaking

Valence

- **In Chemistry**

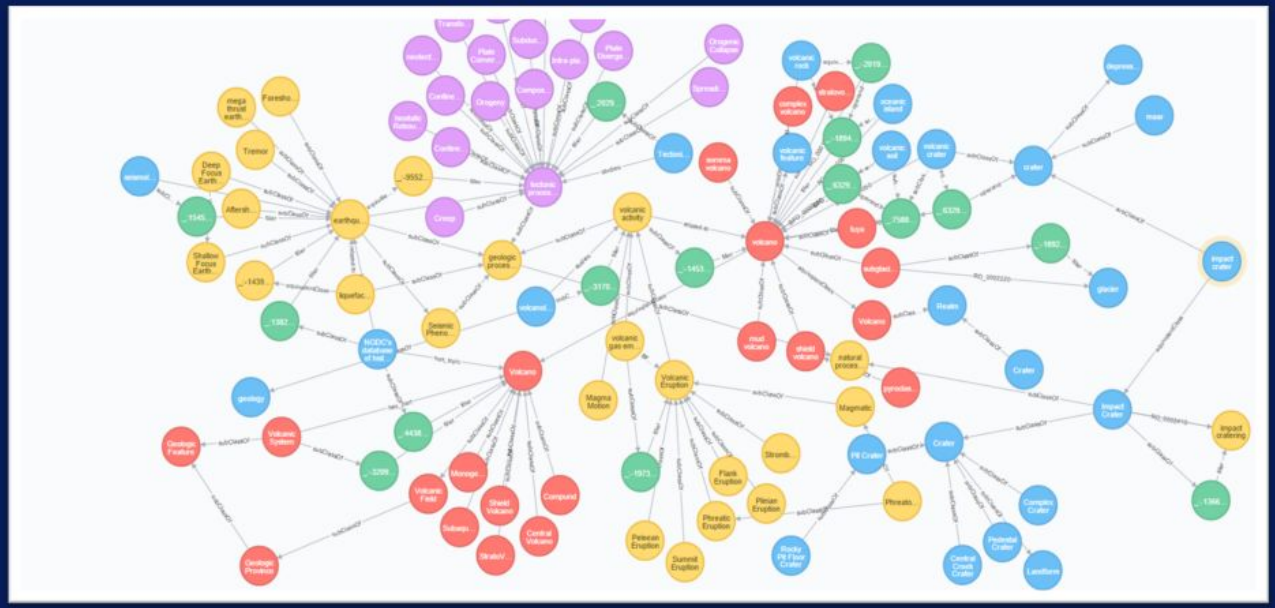


Valence

- **In Graphs**
  - A measure of **connectedness**

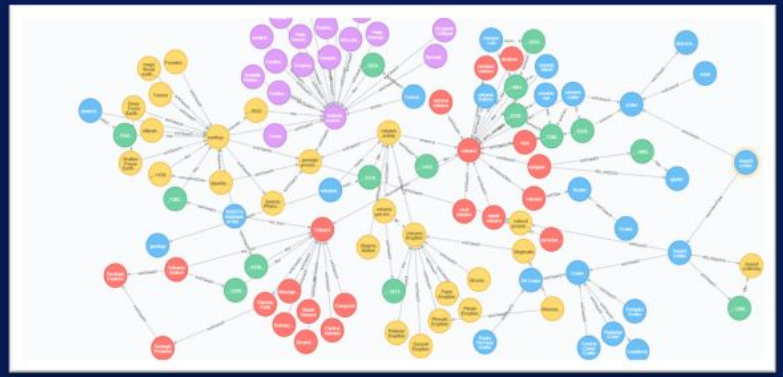
Variety…

…aka Heterogeneity



Variety

- Different kinds of graphs may have very different meanings



Social Networks

Citation Networks

Interaction Networks

Semantic Web/
Linked Data

Cities have Networks

- **Multiple interacting networks over the same spatial domain**
  - Transportation networks
    - Multiple modalities
  - Water and sewage network
  - Power transmission network
  - Broadband IP and M2M networks

What is Analytics?

- **Discovery and communication of meaningful patterns or interesting insights using**
  - Mathematical properties of data
  - Data computing for accessing and manipulating data
  - Domain knowledge to increase interpretability of data and results of analytics
  - Statistical modeling techniques for drawing inferences or making predictions on data

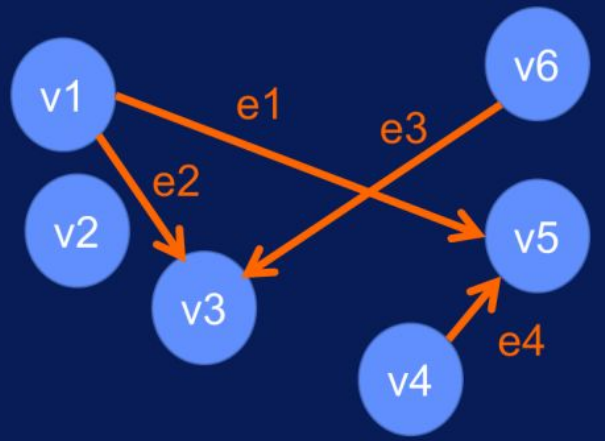
Some Broad Purposes of "Analytics"

- Uncover characteristics of data set based on its mathematical properties
- Answer specific questions from multiple data sets
- Develop a mathematical model for predicting the behavior of some variables
- Detect emergent phenomena and explain its contributing factors

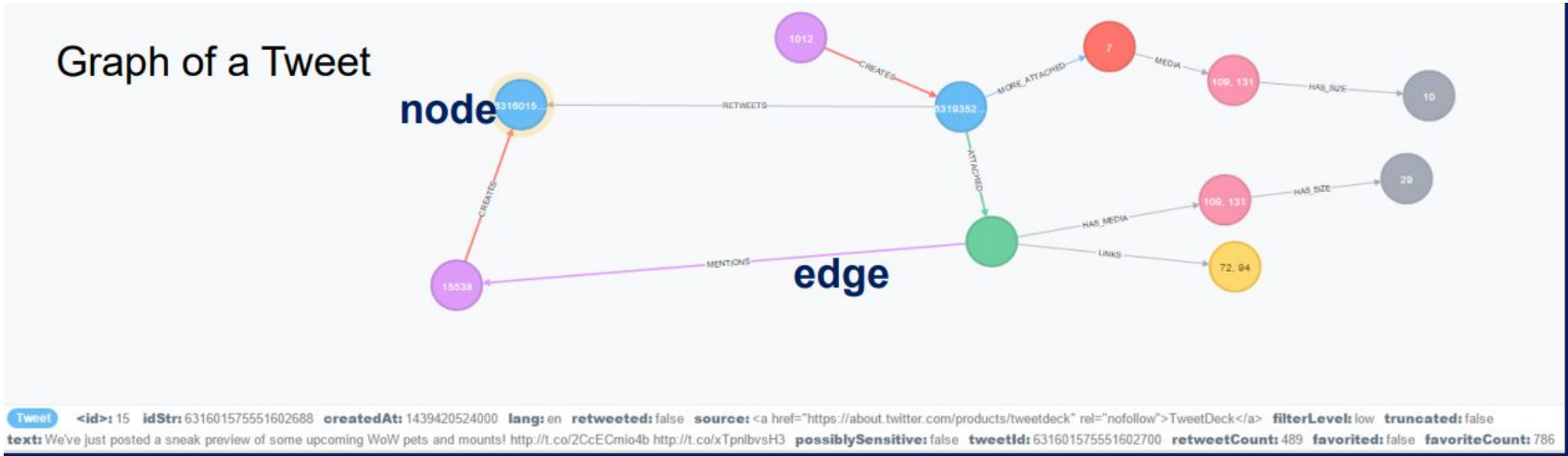
Graph Analytics

- Analytics where the underlying data is natively structured as or can be modeled as a set of graphs

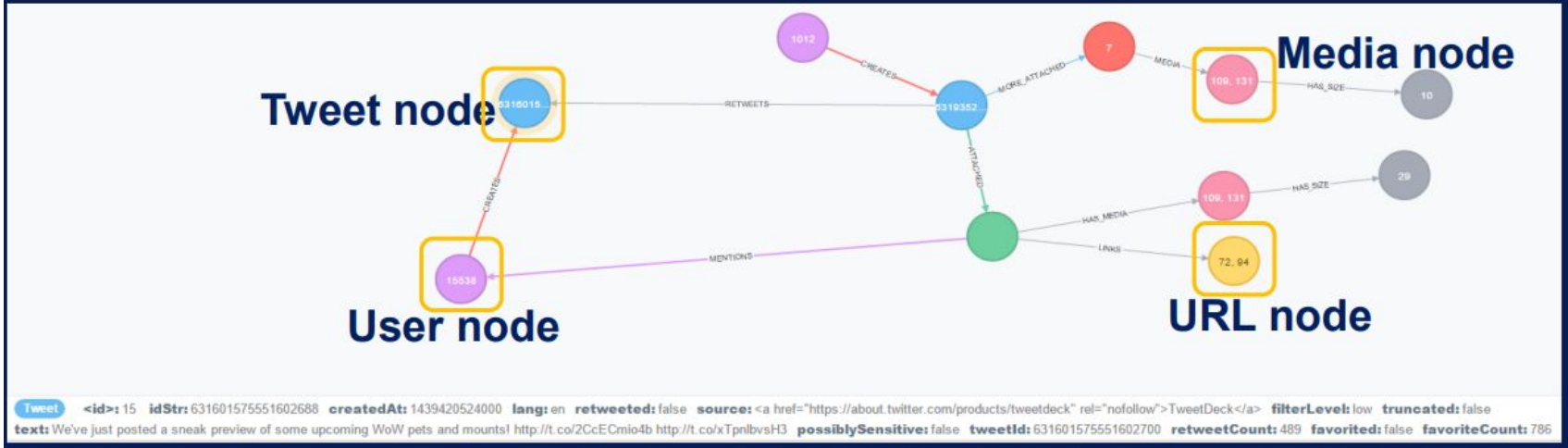
Our First Definition of Graphs



- V: a set of vertices
- E: a set of edges

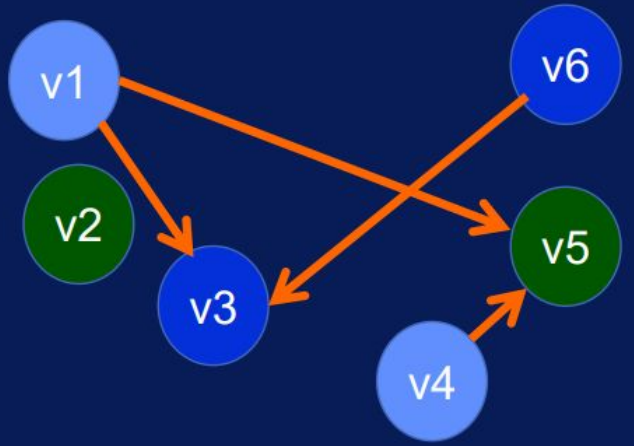


A Real Graph Has More Information Content

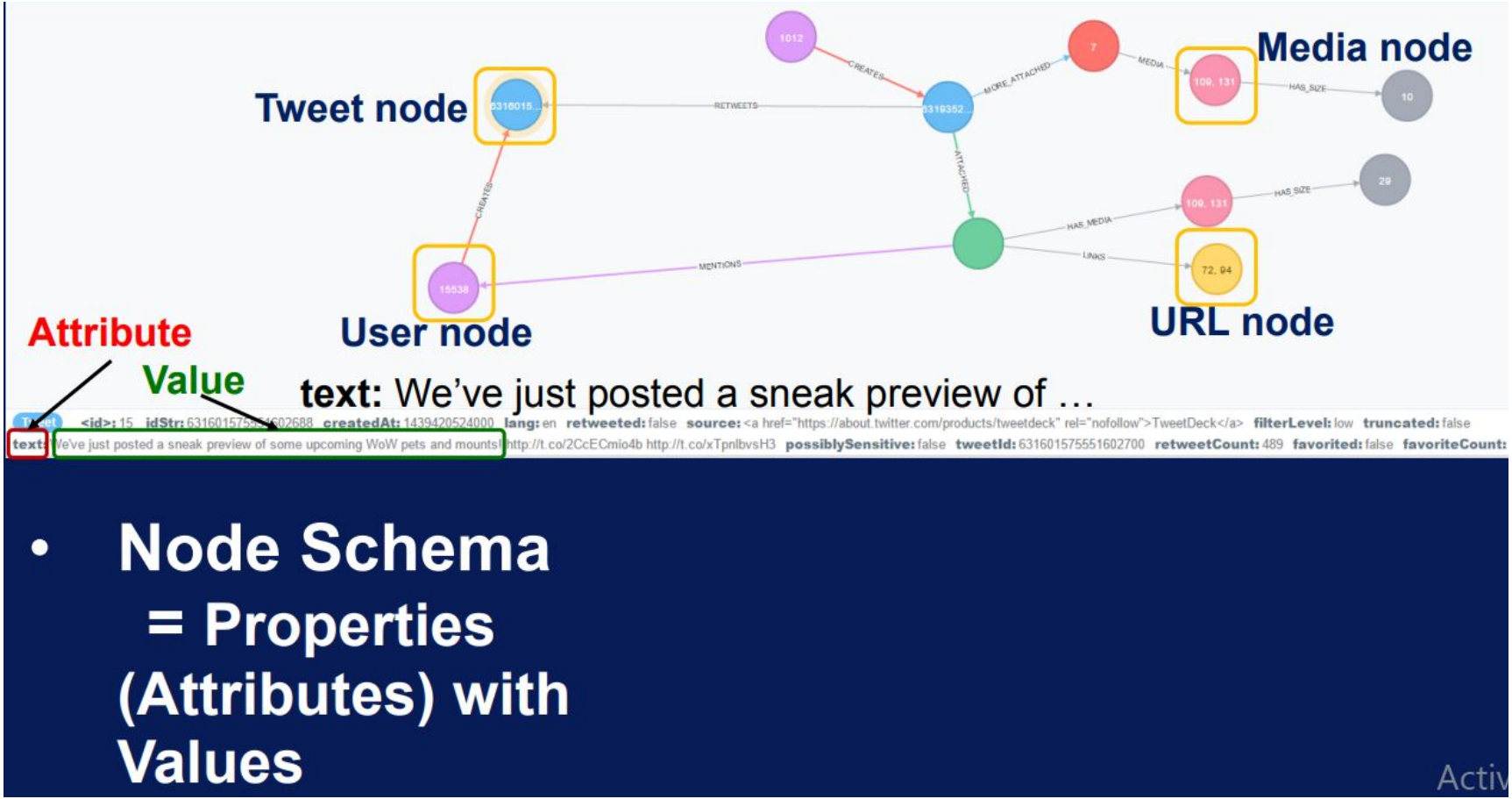


Node Types (aka Labels)

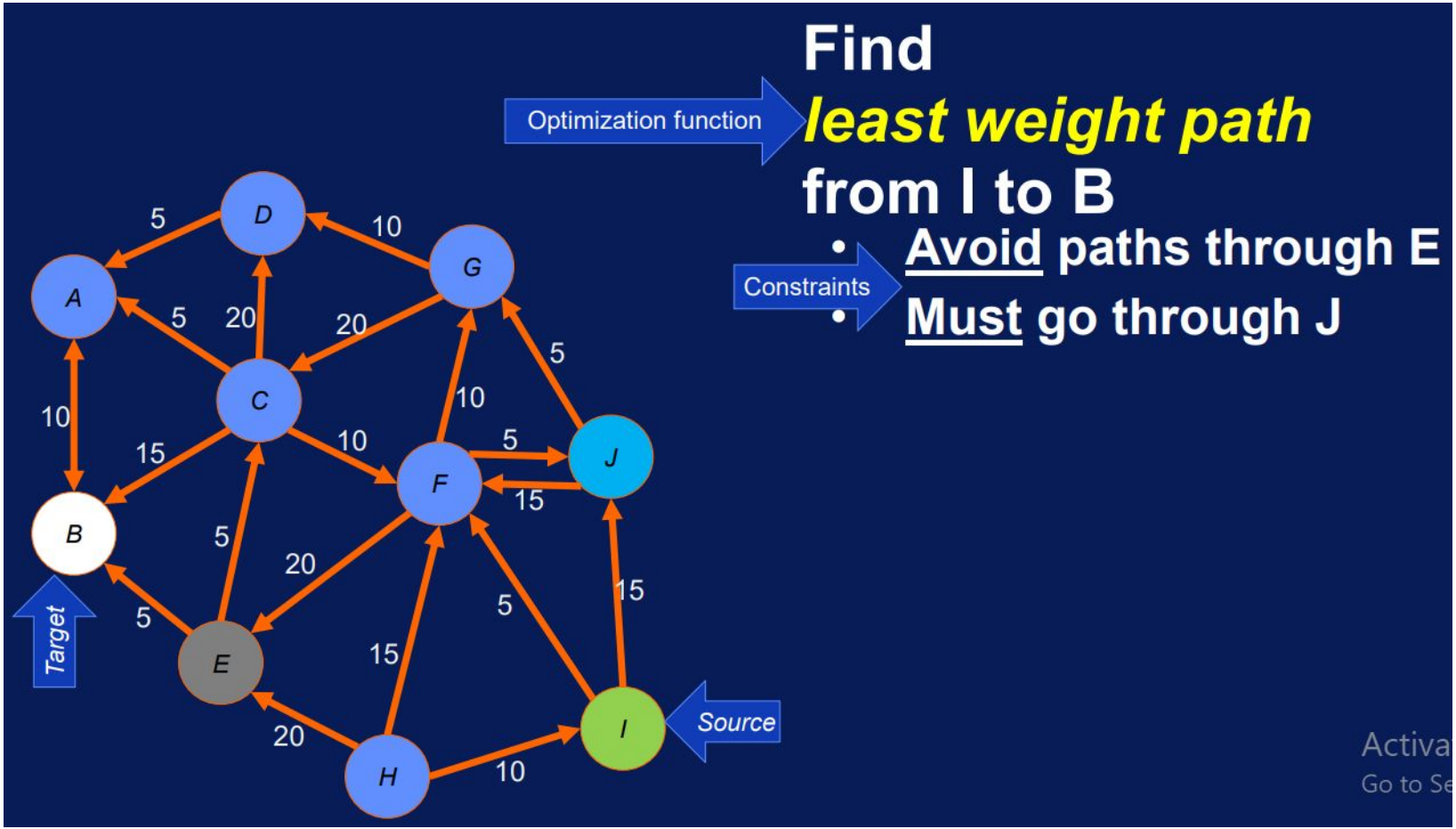
Graphs with Node Types



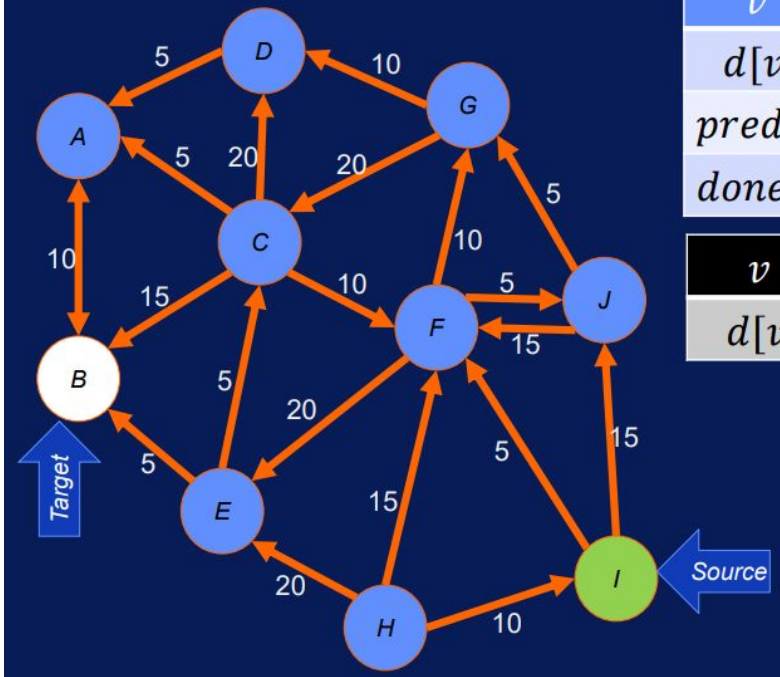
- V: a set of vertices
- E: a set of edges
- TN: a set of node types
- f (TN→V): type assignment to nodes



- **Node Schema = Properties (Attributes) with Values**


Find
least weight path
from I to B
Optimization function
Constraints
Avoid paths through E
Must go through J
A
B
C
D
E
F
G
H
I
J
Target
Source
5
10
15
20

Dijkstra's Algorithm



v	I	A	B	C	D	E	F	G	H	J
$d[v]$	0	∞	∞	∞	∞	∞	∞	∞	∞	∞
$pred[v]$										
$done[v]$	N	N	N	N	N	N	N	N	N	N

v	I	A	B	C	D	E	F	G	H	J
$d[v]$	0	∞	∞	∞	∞	∞	∞	∞	∞	∞

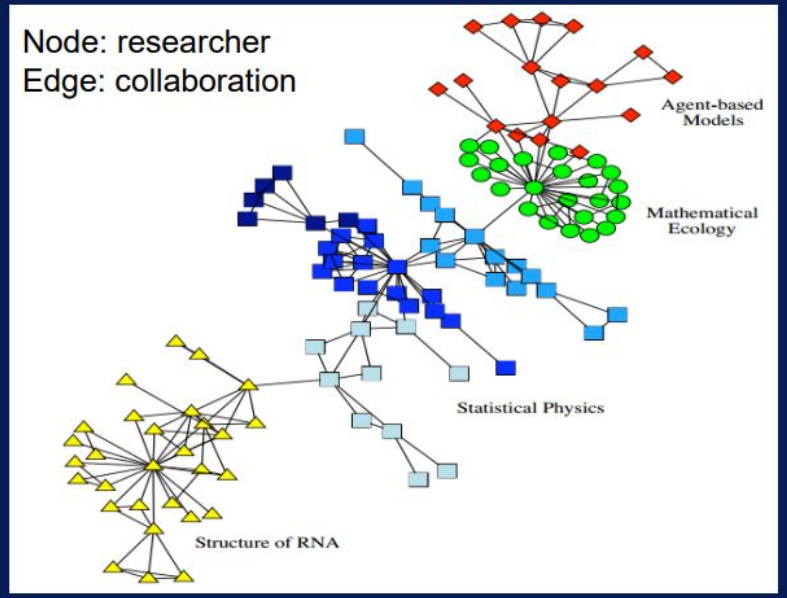
Dijkstra and Big Graphs

- The worst-case complexity of Dijkstra is proportional to the number of edges times log(number of nodes)
  - For1 Million nodes and 10 Million edges, the worst case complexity is proportional to ~14 Million!!

That's really high!!


Community Analytics

What is a Community?



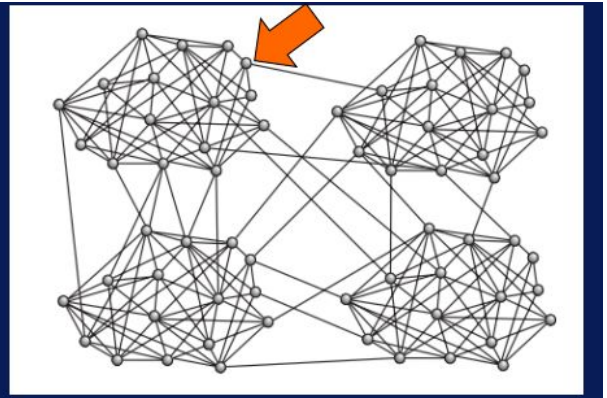
Which researchers collaborate?

- Entities often interact within groups
- Interactions form ***clusters***
- ***Community***
  - ***a dense subgraph (cluster) within a graph whose nodes are more connected within the cluster than to nodes outside the cluster***

Some Analytics Questions

- "Static" Analyses
  - What are the communities at time T?
  - Who belong to a community?
  - How closely knit is this community?
- Temporal/Evolution Analyses
  - How did this community form?
  - Which communities are stable?
  - Find strong *transient* communities – why did they form or dissolve?
- Predictive Analyses
  - Is this community likely to grow?
  - Will these nodes continue as a community in future?
  - Are dominant roles emerging in this community?

Detecting a Community

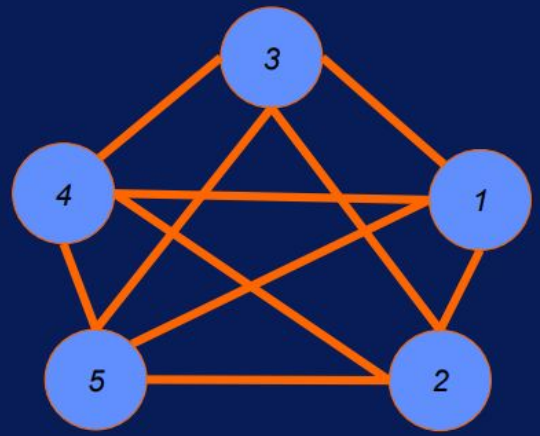


- C – **connected subgraph of graph** G
- **We can compute**
  - The internal and external degree of a vertex
    - Internal – within C
    - External – outside C
  - The internal and external degree of the cluster C
    - Sum of the internal/external degree of the vertices of C
    - Intra-cluster density -- $\delta_{int} = \frac{\#\ of\ internal\ edges\ in\ C}{n_c(nC-1)/2}$
    - Inter-cluster density -- $\delta_{ext} = \frac{\#\ of\ inter\ cluster\ edges\ of\ C}{n_c(n-nC)}$
  - For C to be a community
    - $\boldsymbol{\delta_{int}}$ ***should be high and*** $\boldsymbol{\delta_{ext}}$ ***should be low***

Local Properties

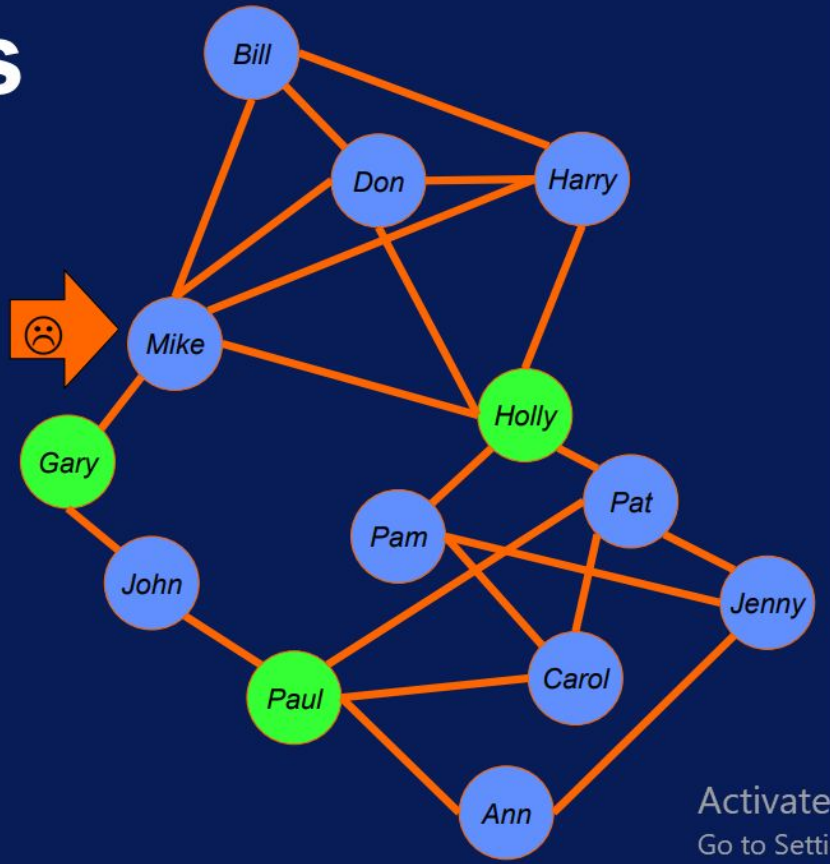
Properties of a subgraph and its neighborhood

- Clique
  - The perfect community
    - every two distinct vertices in the clique are adjacent
  - Finding the largest clique within a graph
    - Computationally hard problem
    - Simpler to find cliques of size k

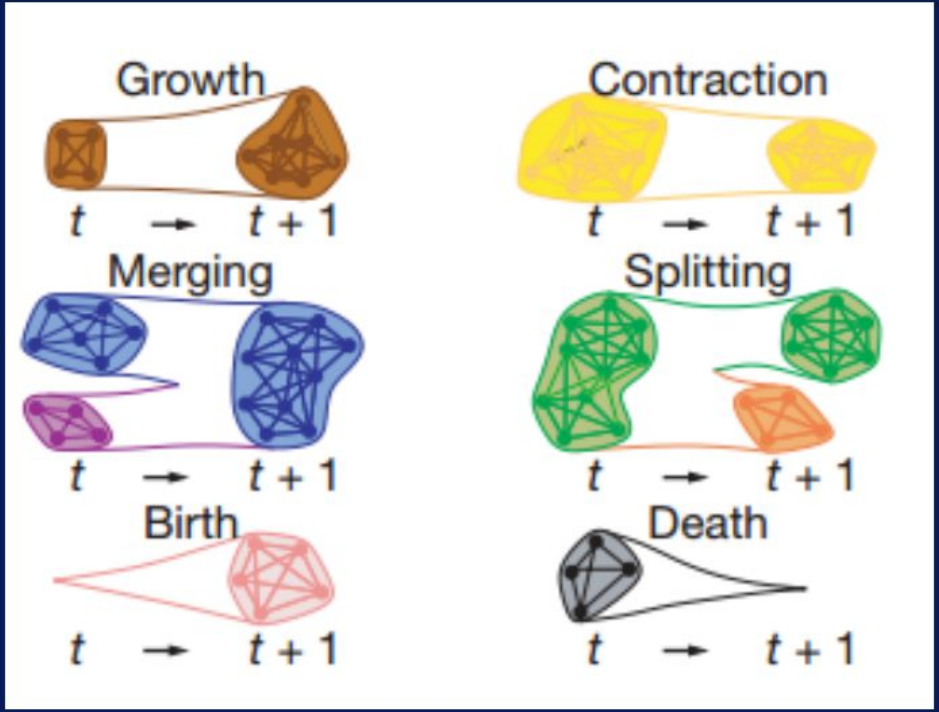


Near Cliques

- n-**clique**
  - Maximal subgraph such that the distance of each pair of its vertices is not larger than n
    - n = 1 for a clique



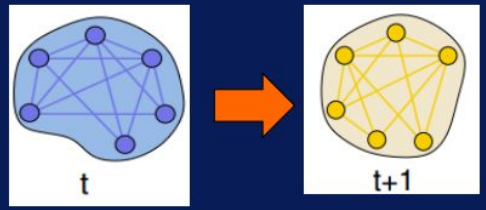
Evolving Communities



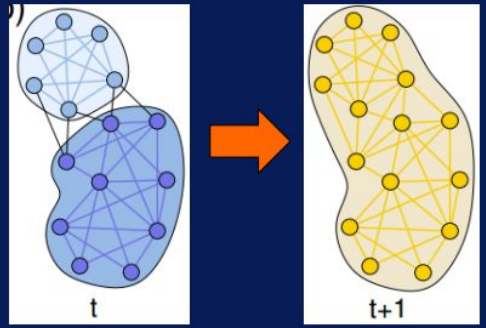
Measuring Evolution

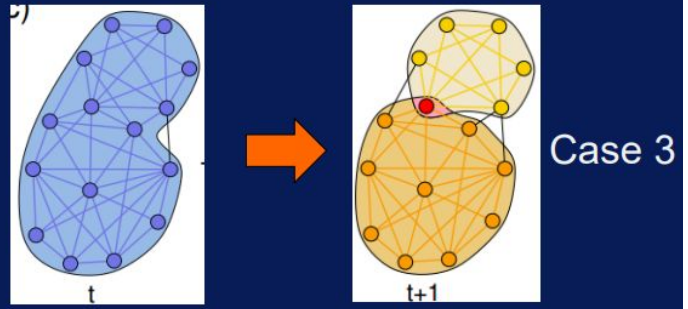
Find a graph at time t_0 and then at t_1

Case 1



Case 2



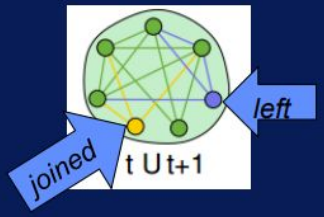


Case 3

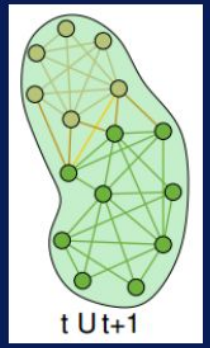
Measuring Evolution

Join the graphs

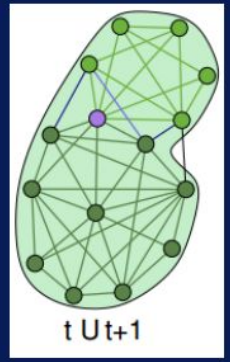
Case 1



Case 2



Case 3



2. Introducing NoSQL

What you will learn

Define the term NoSQL

Describe NoSQL technology

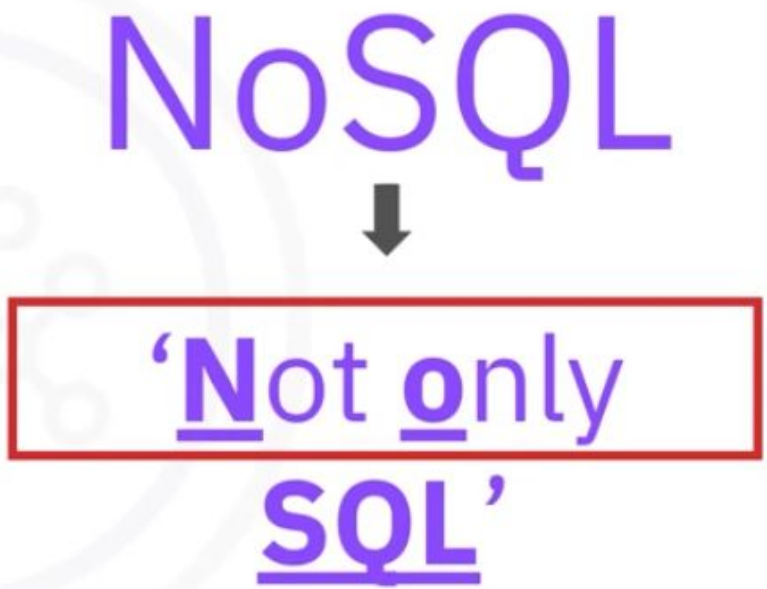
Describe the history of NoSQL

List four reasons for using NoSQL databases

What is NoSQL?

- The NoSQL name was introduced during an open-source event on distributed databases

- NoSQL doesn't mean 'No SQL'
  NoSQL means '**N**ot **o**nly SQL'



What is NoSQL?

- Refers to family of databases that vary widely in style and technology
- However, they share common traits:
  - Non-relational
  - Not standard row and column type RDBMS
- Could be referred to as ‘Non-relational databases’

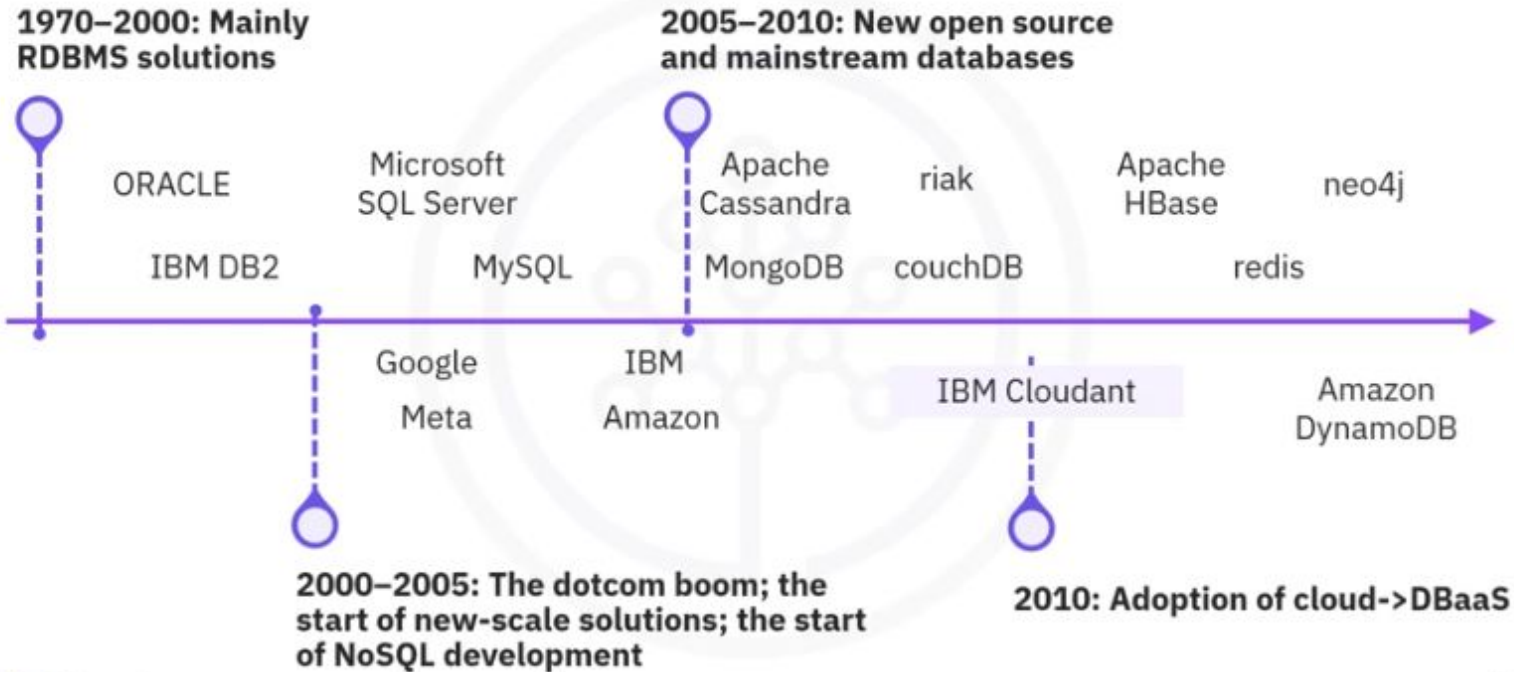
What is NoSQL?

NoSQL databases:

- Provide new ways of storing and querying data
- Address several issues for modern apps
- Provide a flexible schema
- Scale to meet demand
- Are distributed systems

1. Fault tolerance 2. Availability

History of NoSQL



Why NoSQL?

Flexible data models

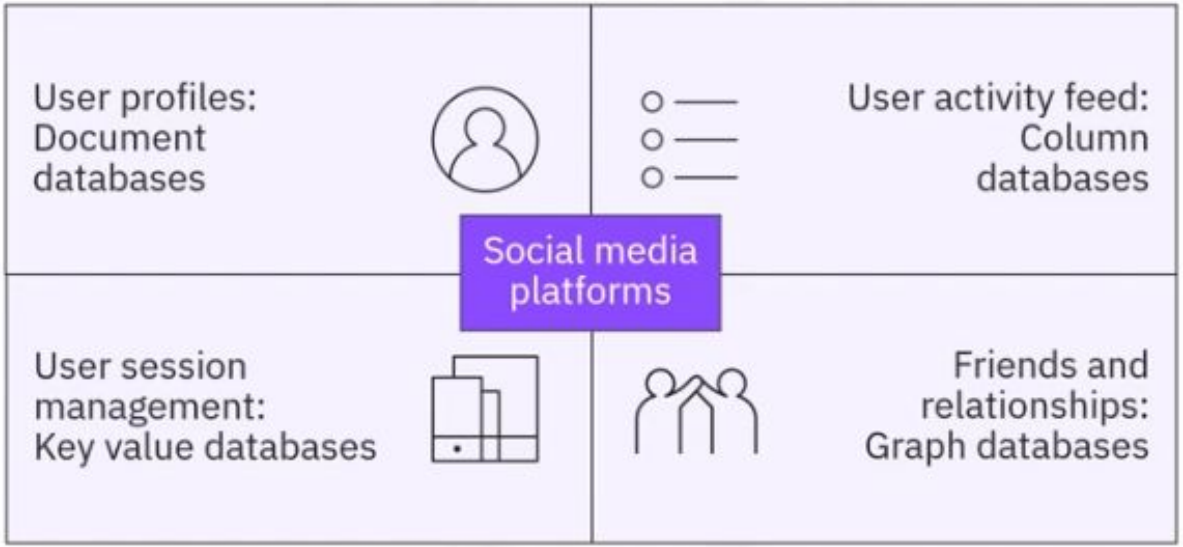
Built-in scalability

Developer productivity

Distributed databases

NoSQL in use: Social media example

Businesses can use a mix of NoSQL databases



Recap

- The name ‘NoSQL’ stands for **N**ot **o**nly **SQL**
- NoSQL refers to a class of databases that are non-relational in architecture
- Implementations of NoSQL databases differ technically but share common traits
- NoSQL databases are distributed systems that easily horizontally scale, and provide native fault tolerance and high availability

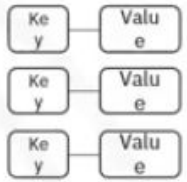
NoSQL database categories

- The most common trait among NoSQL databases is that they are non-relational in architecture
- What types of NoSQL databases are available?
- What is common to them?

NoSQL database categories

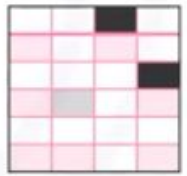
General consensus is...

...NoSQL databases fit into four categories

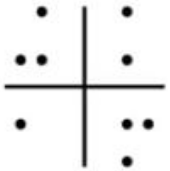


Key-Value

Document



Column



Graph

NoSQL database characteristics

But what ties NoSQL databases together?

- Majority have their roots in the open source community
- Many have been used and leveraged in an open source manner
- Open source community support is fundamental to their industry growth

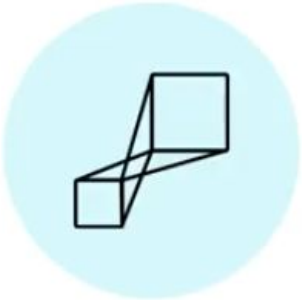
Benefits of NoSQL databases

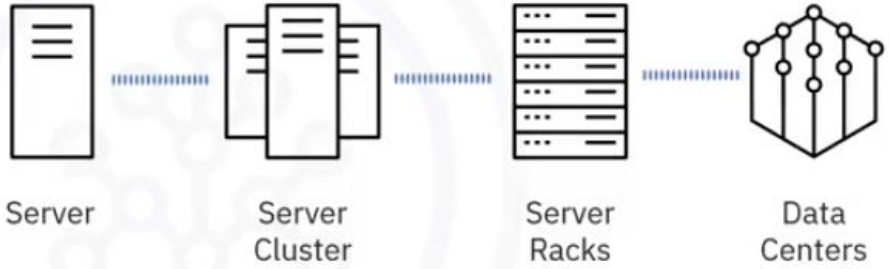
Why use a NoSQL database?

Why is their popularity growing so rapidly?

Benefits of NoSQL databases

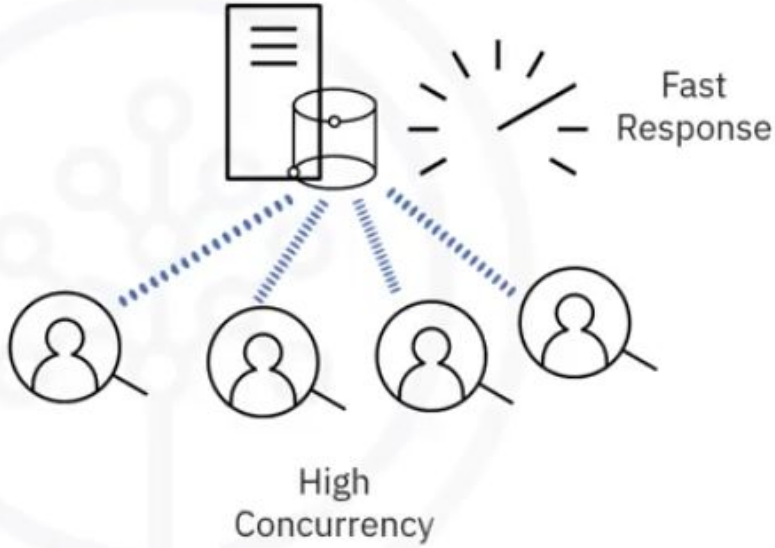
Scalability





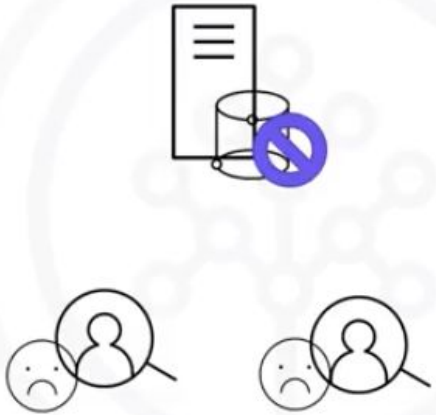
Benefits of NoSQL databases

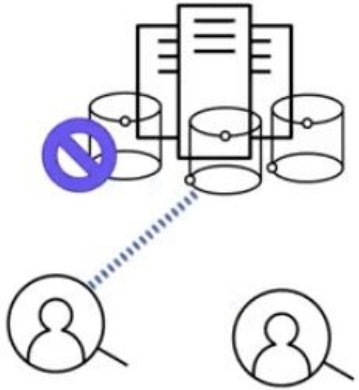
Performance



Benefits of NoSQL databases

Availability



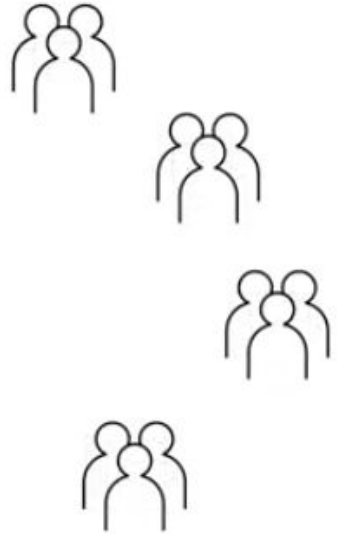


Benefits of NoSQL databases

Cloud Architecture

Benefits of NoSQL databases

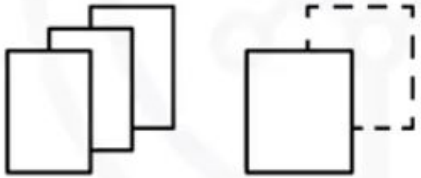
Flexible Schema



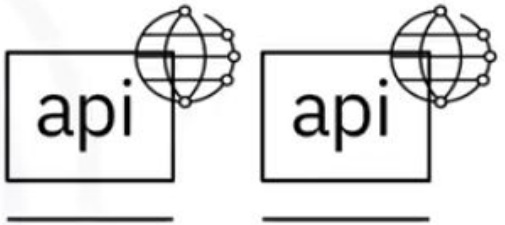
Benefits of NoSQL Databases

Specialized capabilities

Indexing and Querying

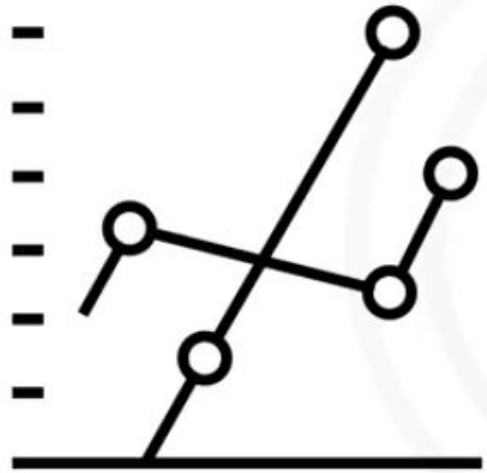


Data Replication Robustness



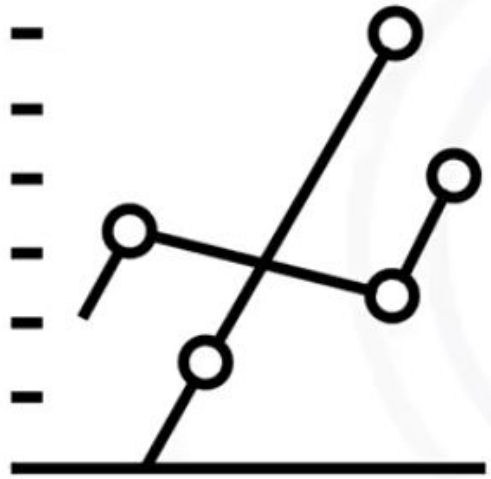
Modern HTTP APIs

Graph NoSQL database architecture



- Graph databases store information in entities (or nodes) and relationships (or edges)
- Graph databases are impressive when your data set resembles a graph-like data structure

Graph NoSQL database architecture



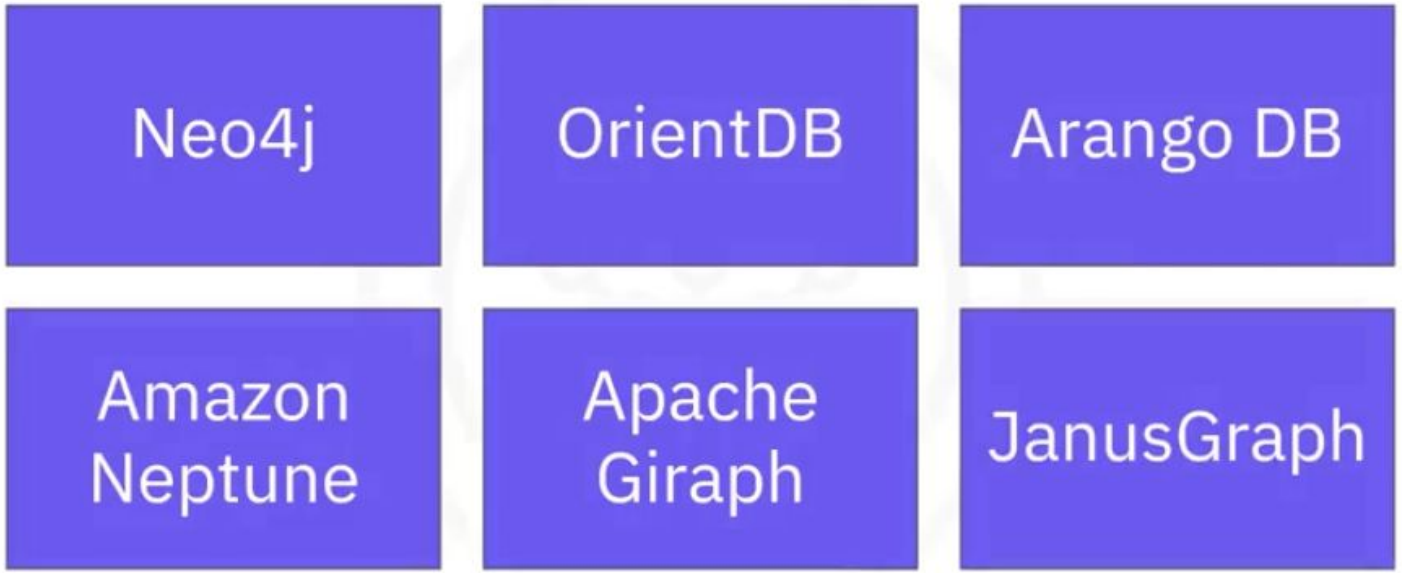
- Graph databases do not shard well
  - Traversing a graph with nodes split across multiple servers can become difficult and hurt performance
- Graph databases are ACID transaction compliant
  - Unlike other NoSQL databases discussed

Graph NoSQL database typical use cases

Cases for a Graph NoSQL database:

- For highly connected and related data
- Social networking
- Routing, spatial, and map apps
- Recommendation engines

Graph NoSQL database example vendors



Hands-On: Getting Started With Neo4j

Five Nodes
N1 = Tom
N2 = Harry
N3 = Julian
N4 = Michele
N5 = Josephine

Five Edges
e1 = Harry 'is known by' Tom
e2 = Julian 'is co-worker of' Harry
e3 = Michele 'is wife of' Harry
e4 = Josephine 'is wife of' Tom
e5 = Josephine 'is friend of' Michele