ENHANCING RELATIONAL MODELS WITH GRAPH PROCESSING IN SQL SERVER 2017

TERRY MCCANN

@SQLSHARK

adatis

MVP Microsoft® Most Valuable Professional

We ***enable*** *our* ***clients***
make ***sense*** *of their* ***data***

DATA ARCHITECTURE

MANAGED SERVICES

DATA STRATEGY

DATA SCIENCE

DATA MANAGEMENT

DATA ANALYTICS

Shout it out!!

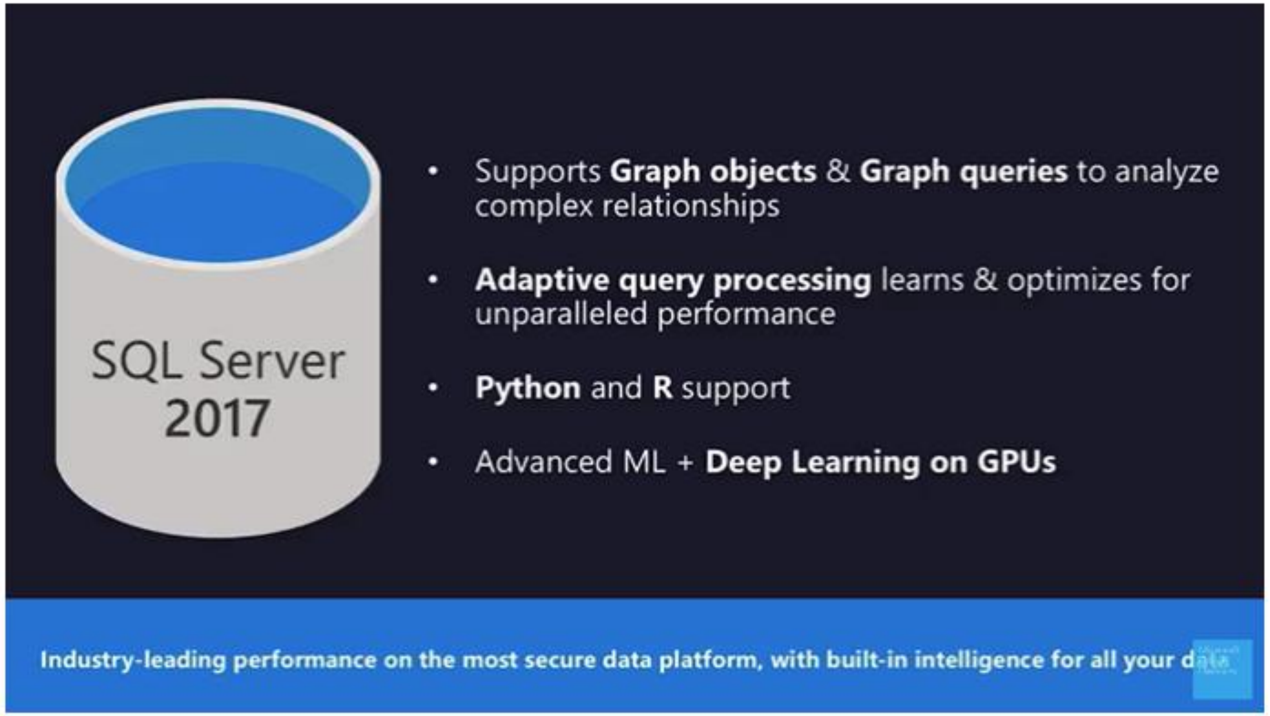
Where to find the ***slides*** and ***demos***

bit.ly/2uHQS74
Github/SQLShark

Terry McCann 2017

SQL Server 2017 Graph data processing. An introduction.

21. April 2017 terrymccann (2)



On the 19th of April 2017 at the Microsoft Data AMP, Microsoft announced SQL Server 2017 and a few new advanced analytics features (slide above). You can watch the full AMP here https://www.microsoft.com/en-us/sql-server/data-amp. One of the announcements related to the new support of Graph objects. There have been rumblings over the past few years of Microsoft working on a Graph engine. Project Trinity appeared to be just that https://www.microsoft.com/en-us/research/project/trinity/. Trinity never really made much of an impact, possibly due to other vendors having an easier to use product. Hopefully Microsoft are changing this by introducing graph querying directly in to the engine for SQL Server 2017. In this blog post I will look at the upcoming SQL Graph the graph processing engine in SQL Server 2017.

What is a Graph Database?

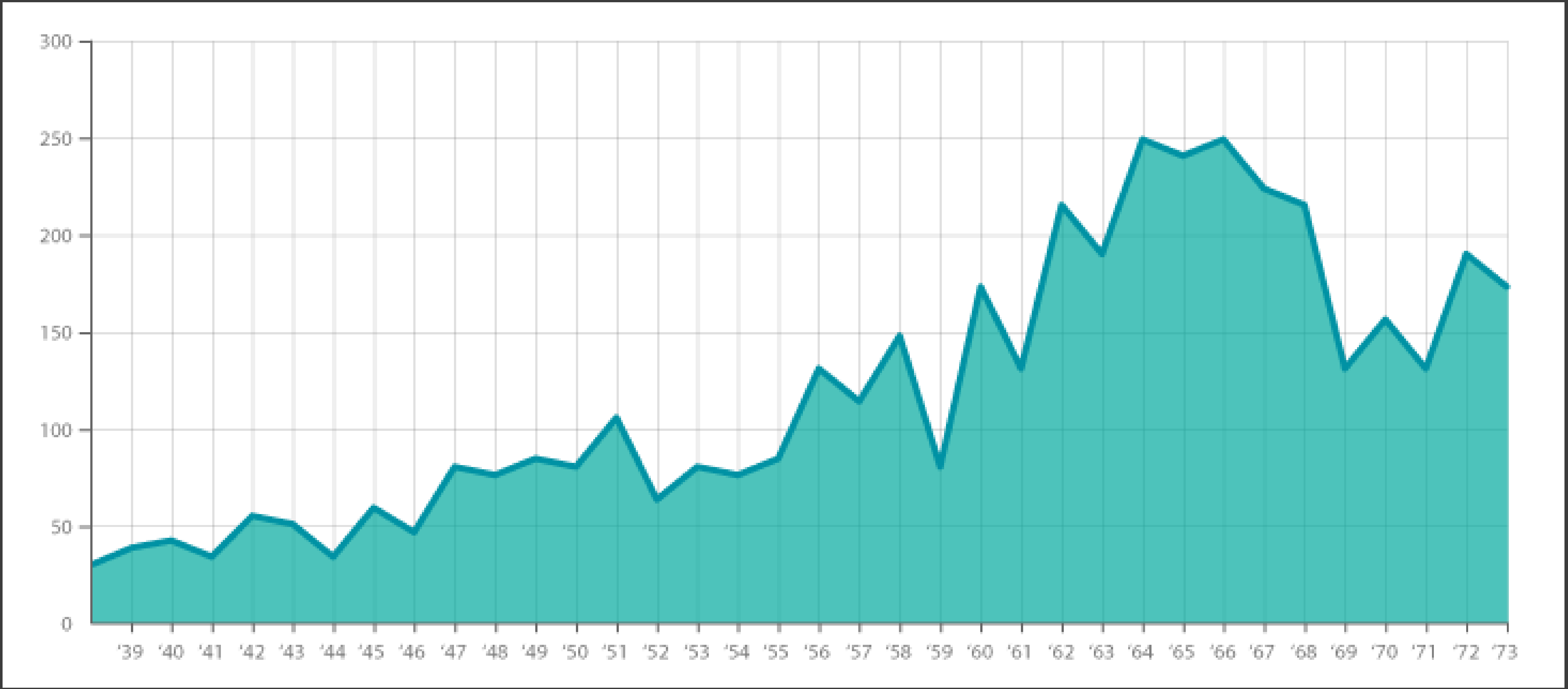
Relational databases are fantastic at answering many types of queries. There are developers who can write a

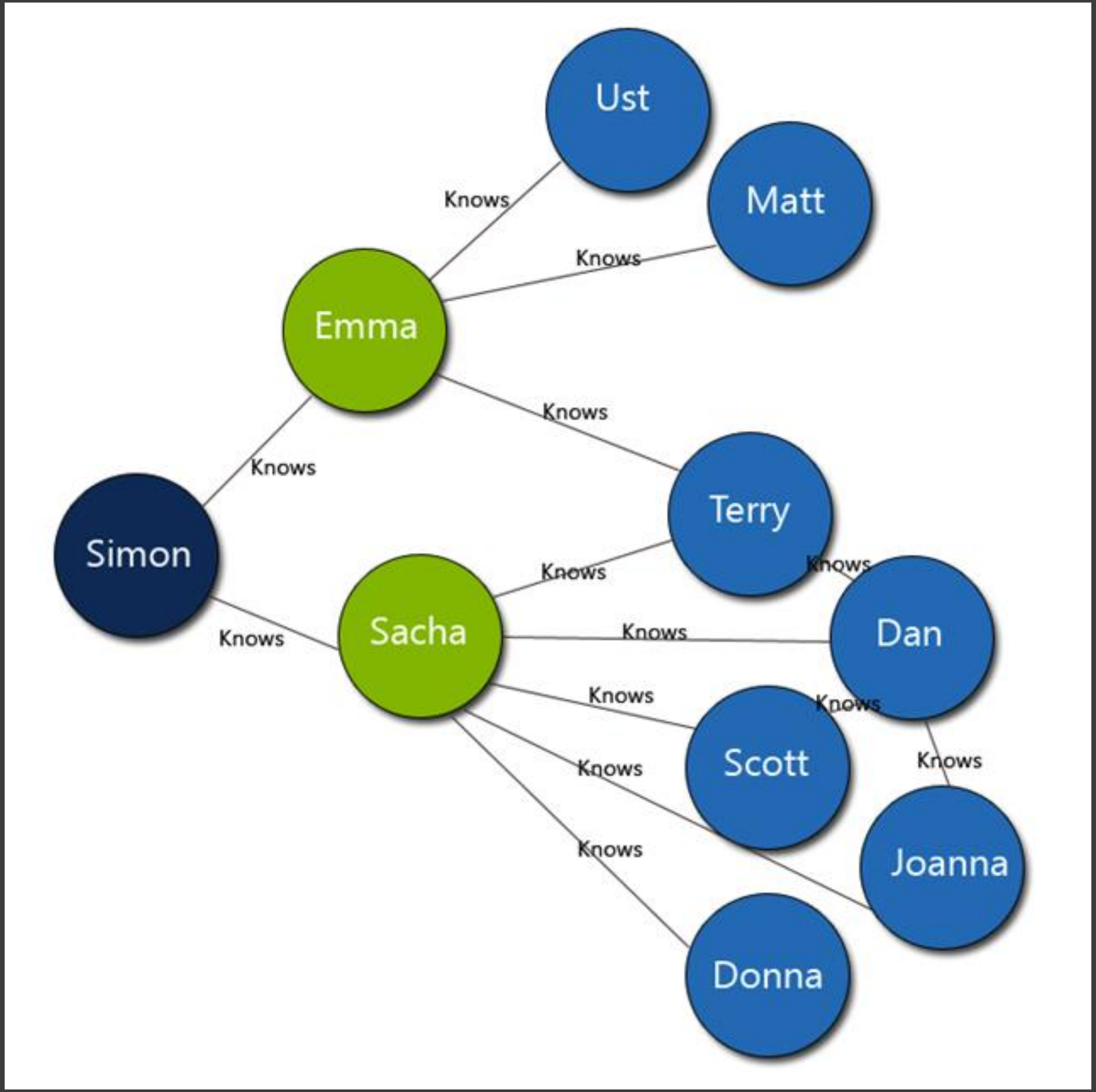
In this session we will look to understand what is ***Graph Processing***?

We will explore the history, the theory, the use cases, ***NODE***, ***EDGE*** & ***MATCH***.

We will understand why Graph is a ***big deal***, and why it will ***change*** how we work with highly connected data.

What is a Graph Database?


300
250
200
150
100
50
0
'39 '40 '41 '42 '43 '44 '45 '46 '47 '48 '49 '50 '51 '52 '53 '54 '55 '56 '57 '58 '59 '60 '61 '62 '63 '64 '65 '66 '67 '68 '69 '70 '71 '72 '73

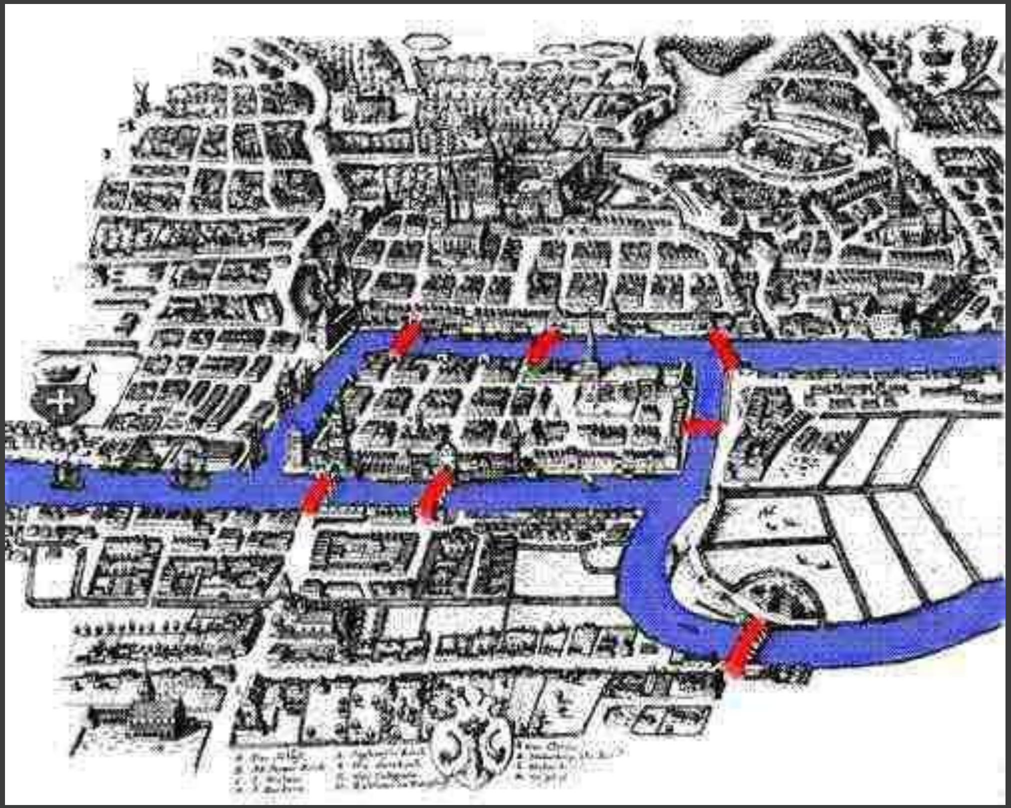

Ust
Matt
Knows
Knows
Emma
Knows
Knows
Terry
Simon
Knows
Knows
Sacha
Knows
Dan
Knows
Knows
Knows
Scott
Knows
Knows
Joanna
Knows
Donna

The History of Graph Databases

Königsberg Bridge Problem

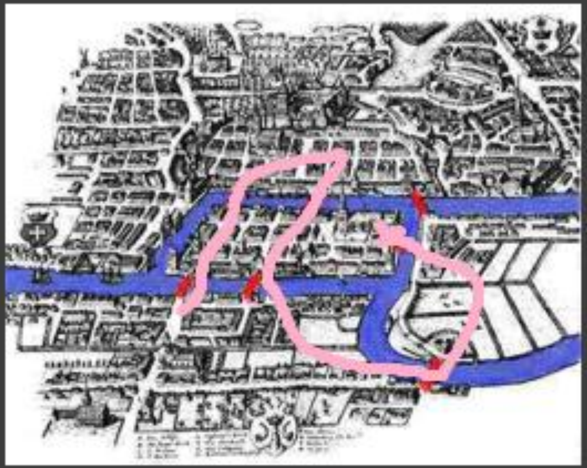
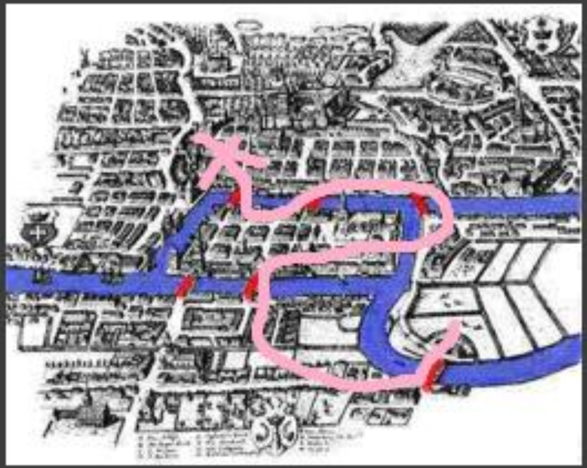
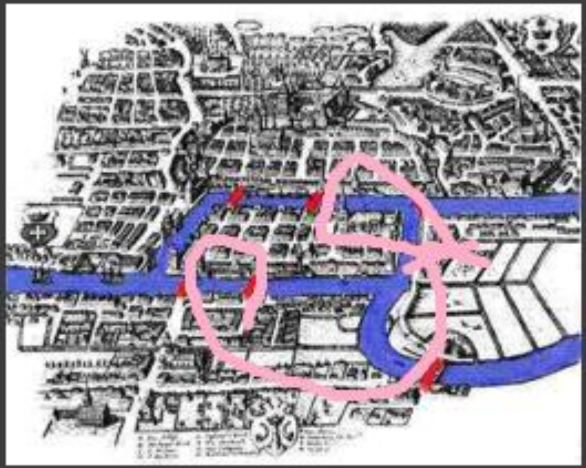
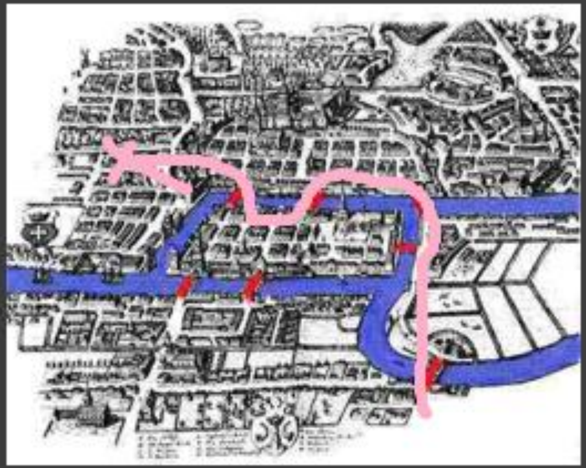
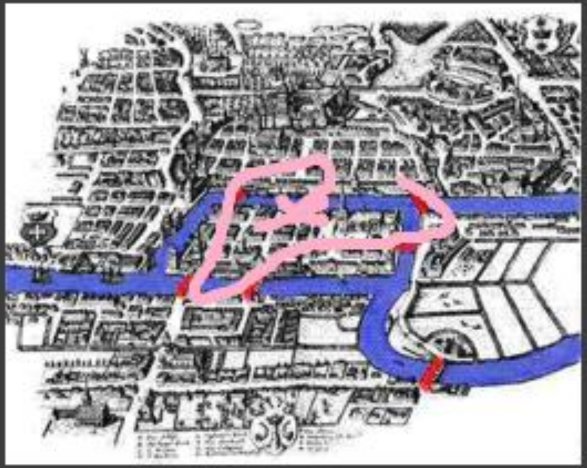
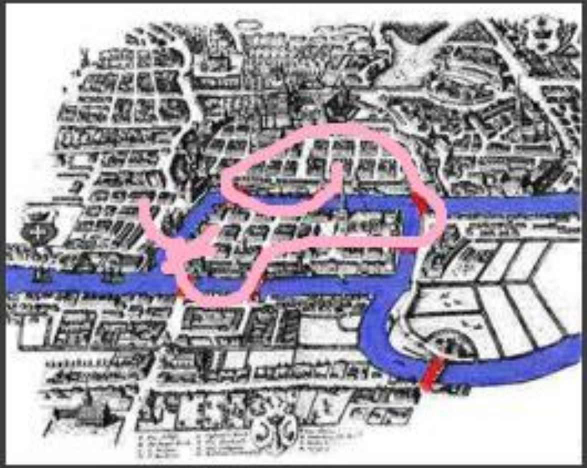
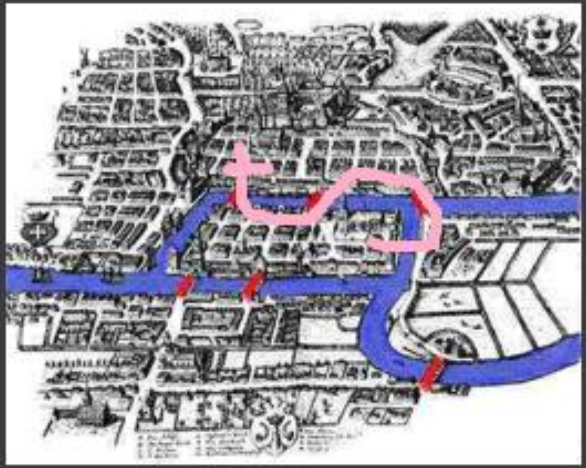
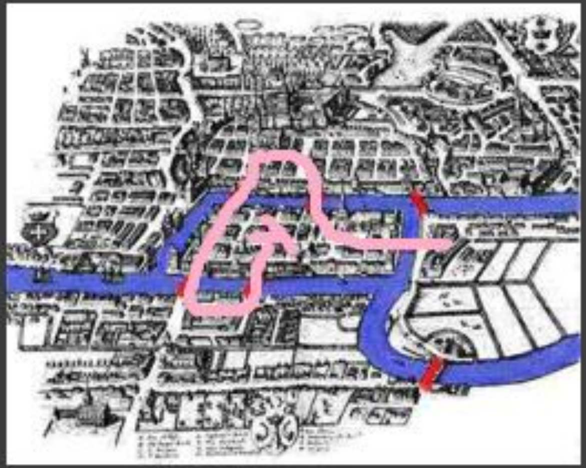
Königsberg Bridge Problem

“Which route would allow you to cross all 7 bridges without crossing any more than once”



Take a minute to work it out

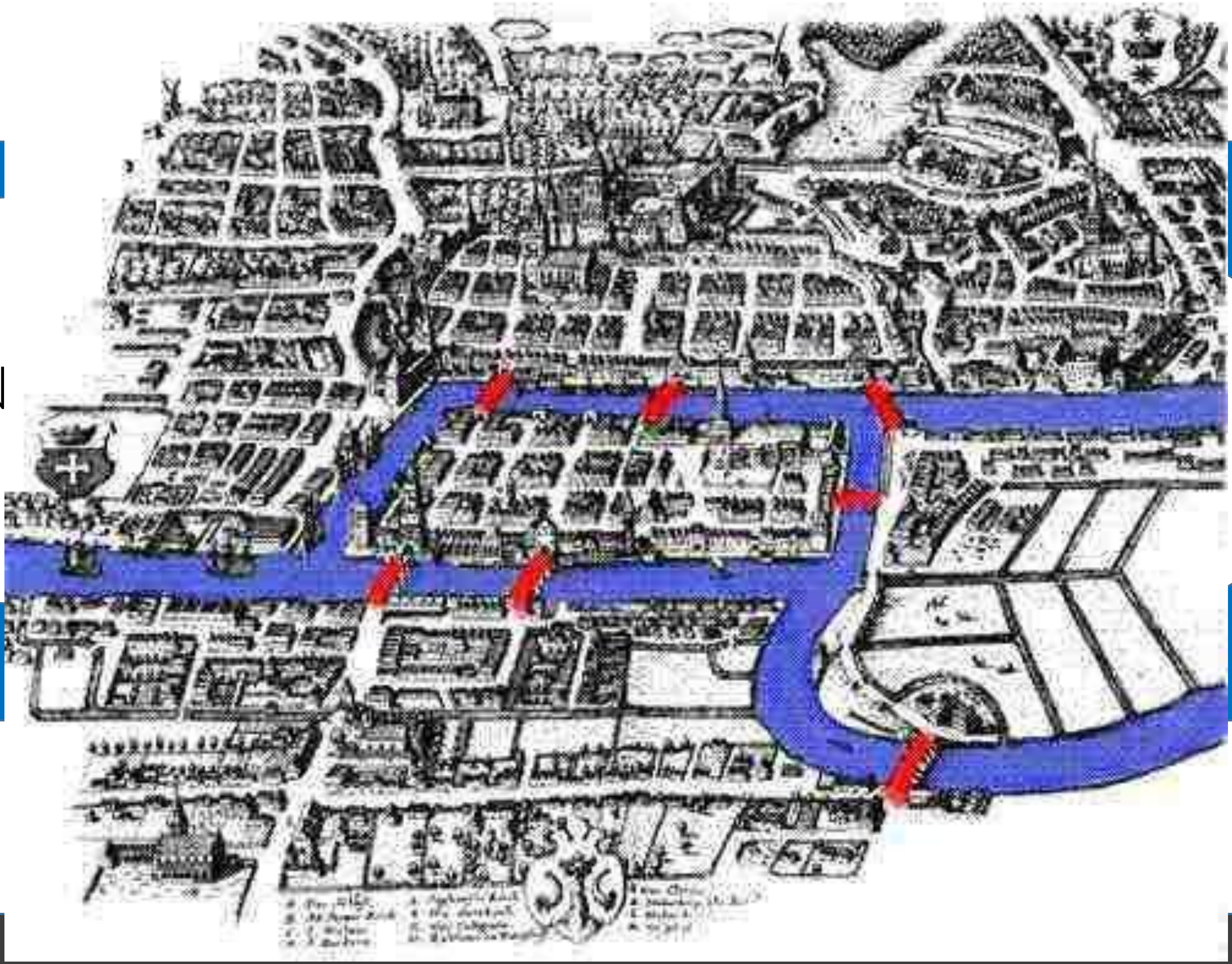
Solved it?

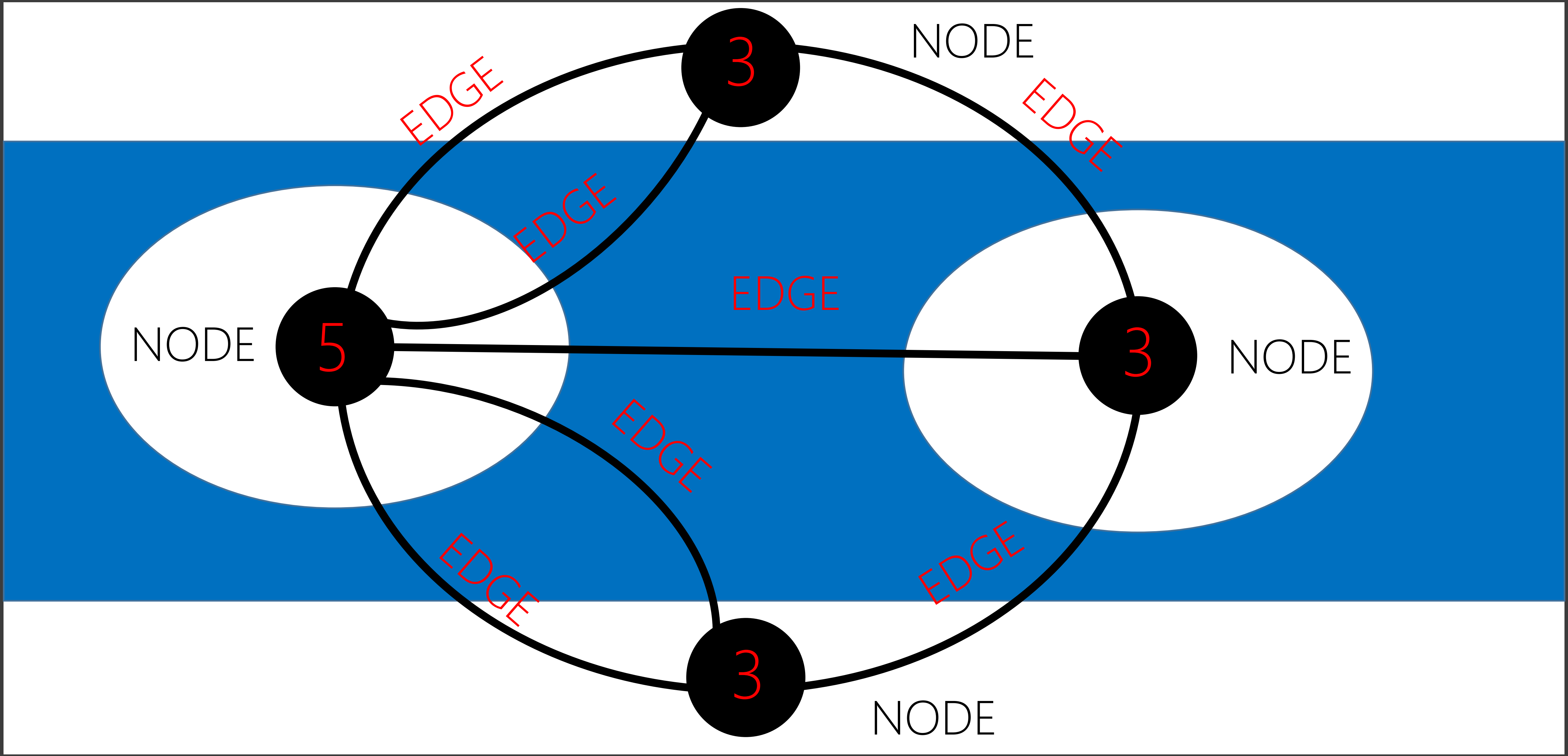
Leonhard Euler

https://www.youtube.com/watch?v=nZwSo4vfw6c

ISLAN



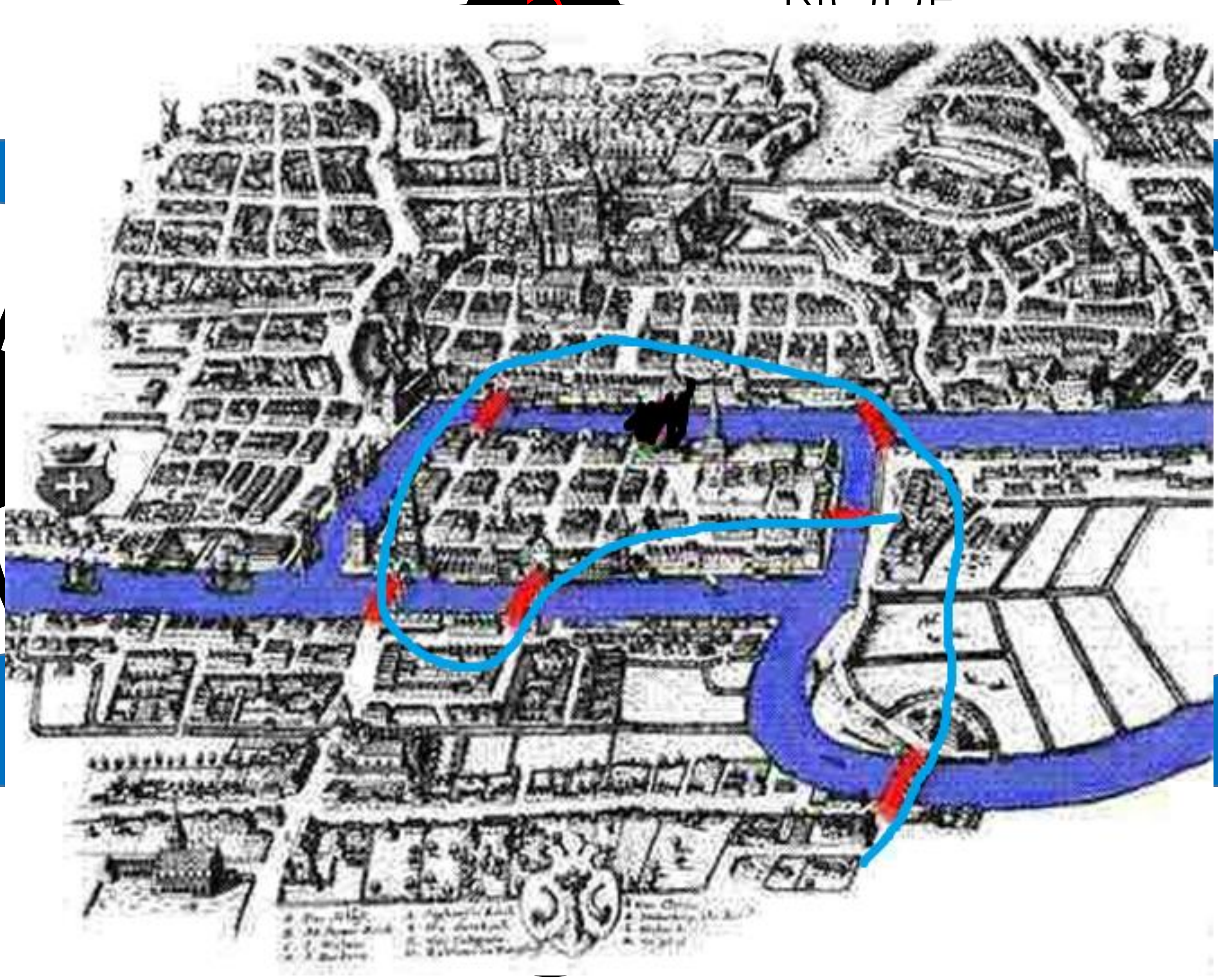
https://www.youtube.com/watch?v=nZwSo4vfw6c



https://www.youtube.com/watch?v=nZwSo4vfw6c

NODE

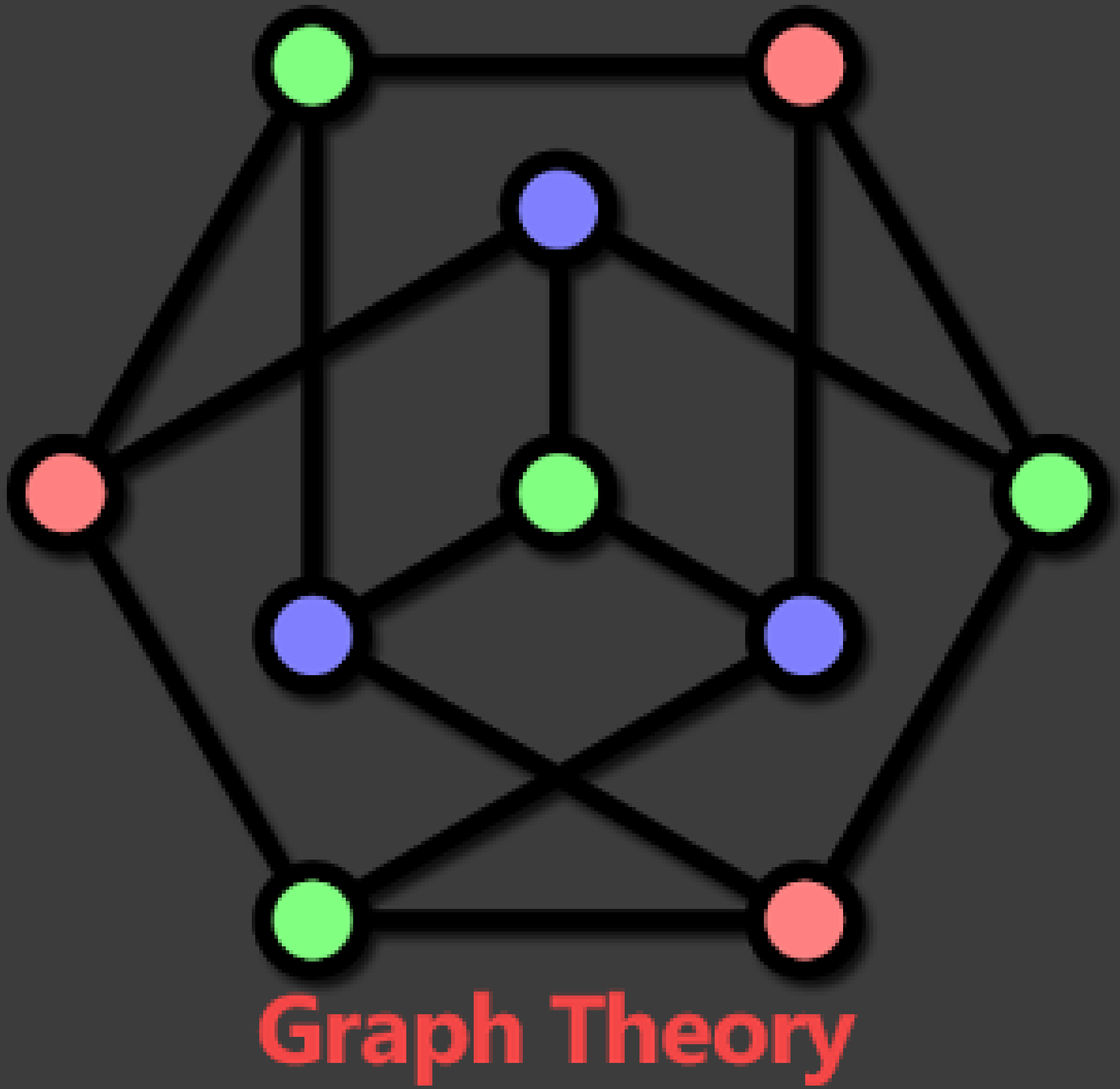
4



NODE

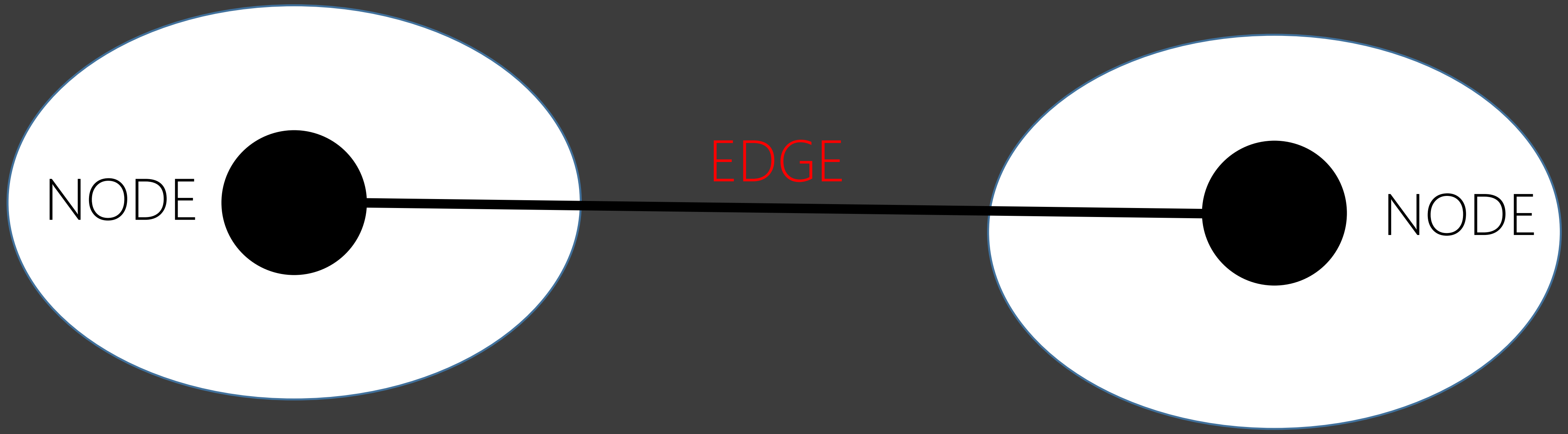
https://www.youtube.com/watch?v=nZwSo4vfw6c

Leonhard Euler



Terry McCann 2017

https://www.youtube.com/watch?v=nZwSo4vfw6c



NODE – An entity in a graph

EDGE – Relationship between NODEs

What is a Graph Database?

In computing, a ***graph database*** is a database that uses graph structures for semantic queries with ***nodes***, ***edges*** and ***properties*** to represent and store data.

Graph databases, by design, allow ***simple*** and ***fast*** retrieval of ***complex*** hierarchical structures that are ***difficult*** to model in relational systems

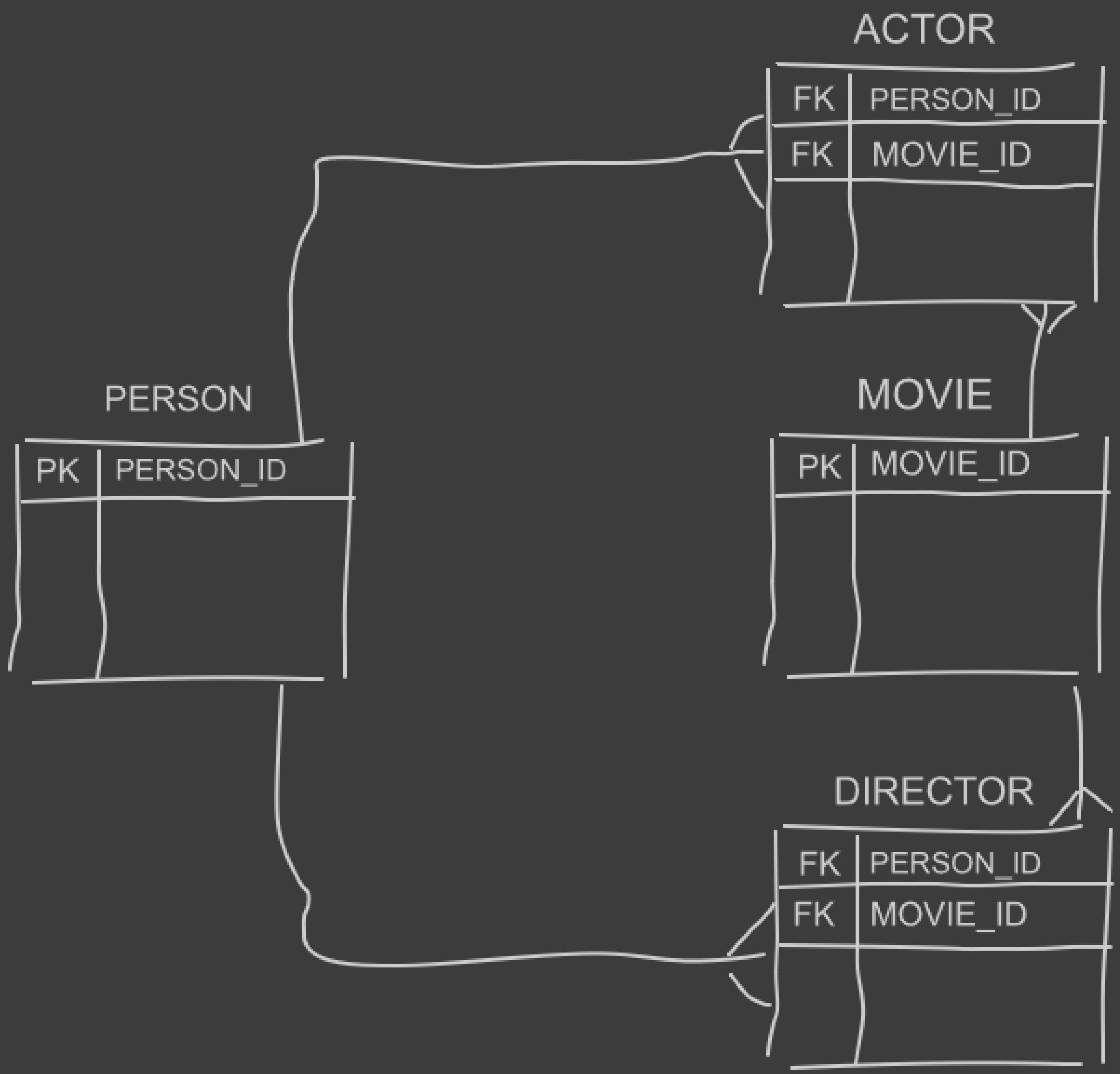
You cannot do that in a Relational Database

The ***problem*** with relational databases

Why ***relational*** database?

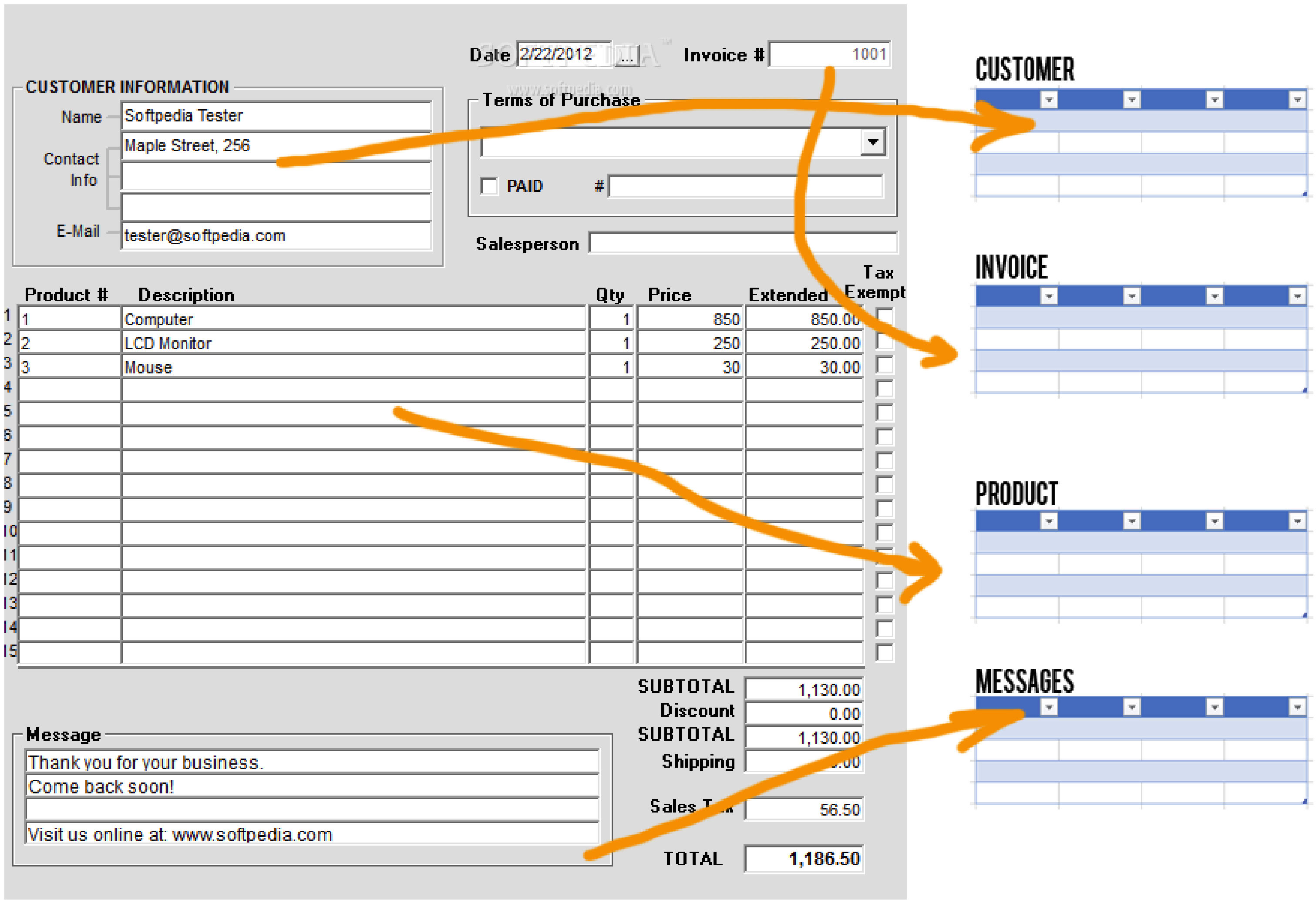
theory

A Simple Example



https://sonra.io/2017/06/12/benefits-graph-databases-data-warehousing/

Impedance Mismatch


Date 2/22/2012
Invoice # 1001
CUSTOMER INFORMATION
Name Softpedia Tester
Contact Info Maple Street, 256
E-Mail tester@softpedia.com
Terms of Purchase
PAID
#
Salesperson
Product # Description Qty Price Extended Tax Exempt
1 1 Computer 1 850 850.00
2 2 LCD Monitor 1 250 250.00
3 3 Mouse 1 30 30.00
4
5
6
7
8
9
10
11
12
13
14
15
SUBTOTAL 1,130.00
Discount 0.00
SUBTOTAL 1,130.00
Shipping
Sales Tax 56.50
TOTAL 1,186.50
Message
Thank you for your business.
Come back soon!
Visit us online at: www.softpedia.com
CUSTOMER
INVOICE
PRODUCT
MESSAGES

Complex to model and store relationships

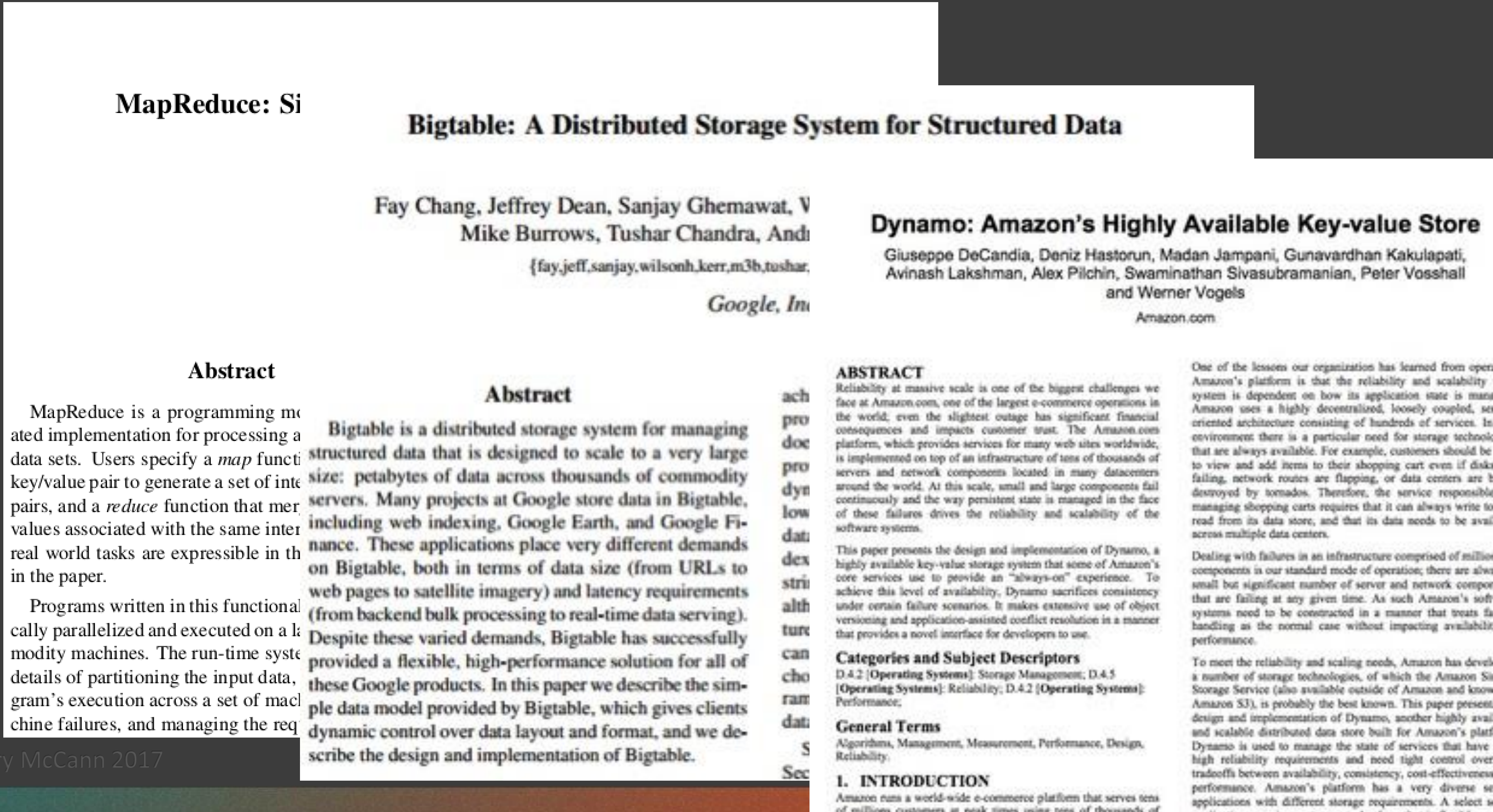
Performance ***degrades*** with more data

Queries can be ***long*** and ***complex***

What is complexity in a database?

Complexity=

f(size , variable structure , connectedness)


MapReduce: Si

Abstract

MapReduce is a programming mo
ated implementation for processing a
data sets. Users specify a *map* functi
key/value pair to generate a set of inte
pairs, and a *reduce* function that mer
values associated with the same inter
real world tasks are expressible in th
in the paper.

Programs written in this functional
cally parallelized and executed on a la
modity machines. The run-time syste
details of partitioning the input data,
gram's execution across a set of mach
chine failures, and managing the req

Bigtable: A Distributed Storage System for Structured Data

Fay Chang, Jeffrey Dean, Sanjay Ghemawat, W
Mike Burrows, Tushar Chandra, Andr
{fay,jeff,sanjay,wilsonh,kerr,m3b,tushar,
Google, In

Abstract

Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers. Many projects at Google store data in Bigtable, including web indexing, Google Earth, and Google Finance. These applications place very different demands on Bigtable, both in terms of data size (from URLs to web pages to satellite imagery) and latency requirements (from backend bulk processing to real-time data serving). Despite these varied demands, Bigtable has successfully provided a flexible, high-performance solution for all of these Google products. In this paper we describe the simple data model provided by Bigtable, which gives clients dynamic control over data layout and format, and we describe the design and implementation of Bigtable.

Dynamo: Amazon's Highly Available Key-value Store

Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall and Werner Vogels
Amazon.com

ABSTRACT

Reliability at massive scale is one of the biggest challenges we face at Amazon.com, one of the largest e-commerce operations in the world; even the slightest outage has significant financial consequences and impacts customer trust. The Amazon.com platform, which provides services for many web sites worldwide, is implemented on top of an infrastructure of tens of thousands of servers and network components located in many datacenters around the world. At this scale, small and large components fail continuously and the way persistent state is managed in the face of these failures drives the reliability and scalability of the software systems.

This paper presents the design and implementation of Dynamo, a highly available key-value storage system that some of Amazon's core services use to provide an "always-on" experience. To achieve this level of availability, Dynamo sacrifices consistency under certain failure scenarios. It makes extensive use of object versioning and application-assisted conflict resolution in a manner that provides a novel interface for developers to use.

Categories and Subject Descriptors

D.4.2 [**Operating Systems**]: Storage Management; D.4.5 [**Operating Systems**]: Reliability; D.4.2 [**Operating Systems**]: Performance;

General Terms

Algorithms, Management, Measurement, Performance, Design, Reliability.

1. INTRODUCTION

Amazon runs a world-wide e-commerce platform that serves tens

Polyglot Persistence

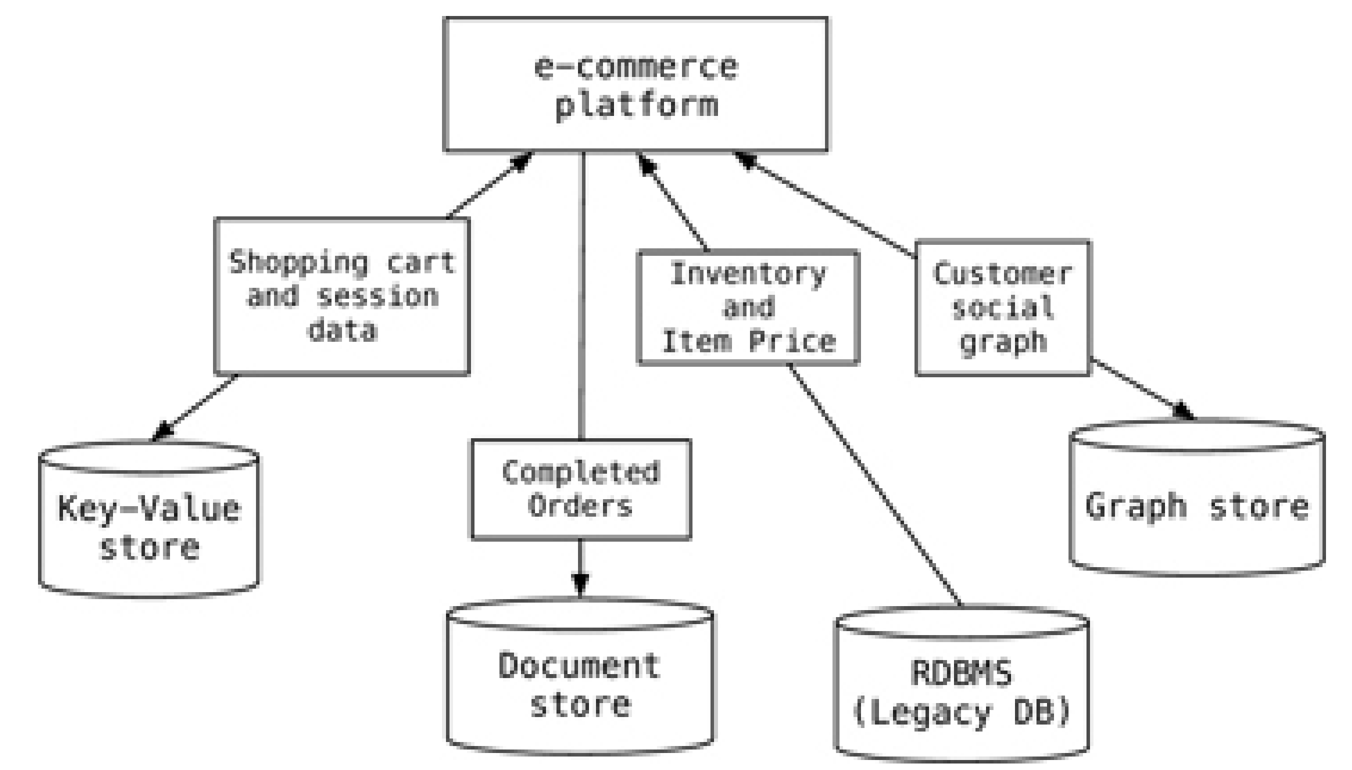
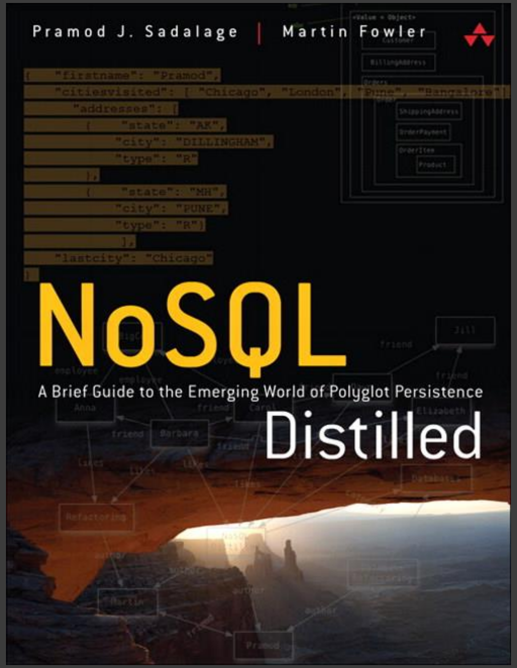


Figure 13.3. *Example implementation of polyglot persistence*



Martin Fowler

Relational is fantastic but ***Polyglot Persistence*** tells us to use the right data store for the challenge we are facing!

If we have ***highly connected data***, a RDBMS might not be the best too for the job!

A Property Graph


NODE
EDGE
NODE


Person
Acted In
Movie


Robin Williams
Acted In
Jumanji


Directed
Person
Movie


James
Cameron
Directed
Titanic

Entities become ***NODES***

FK/PK relationships become ***EDGES***

Many-To-Many relationships become ***EDGES***

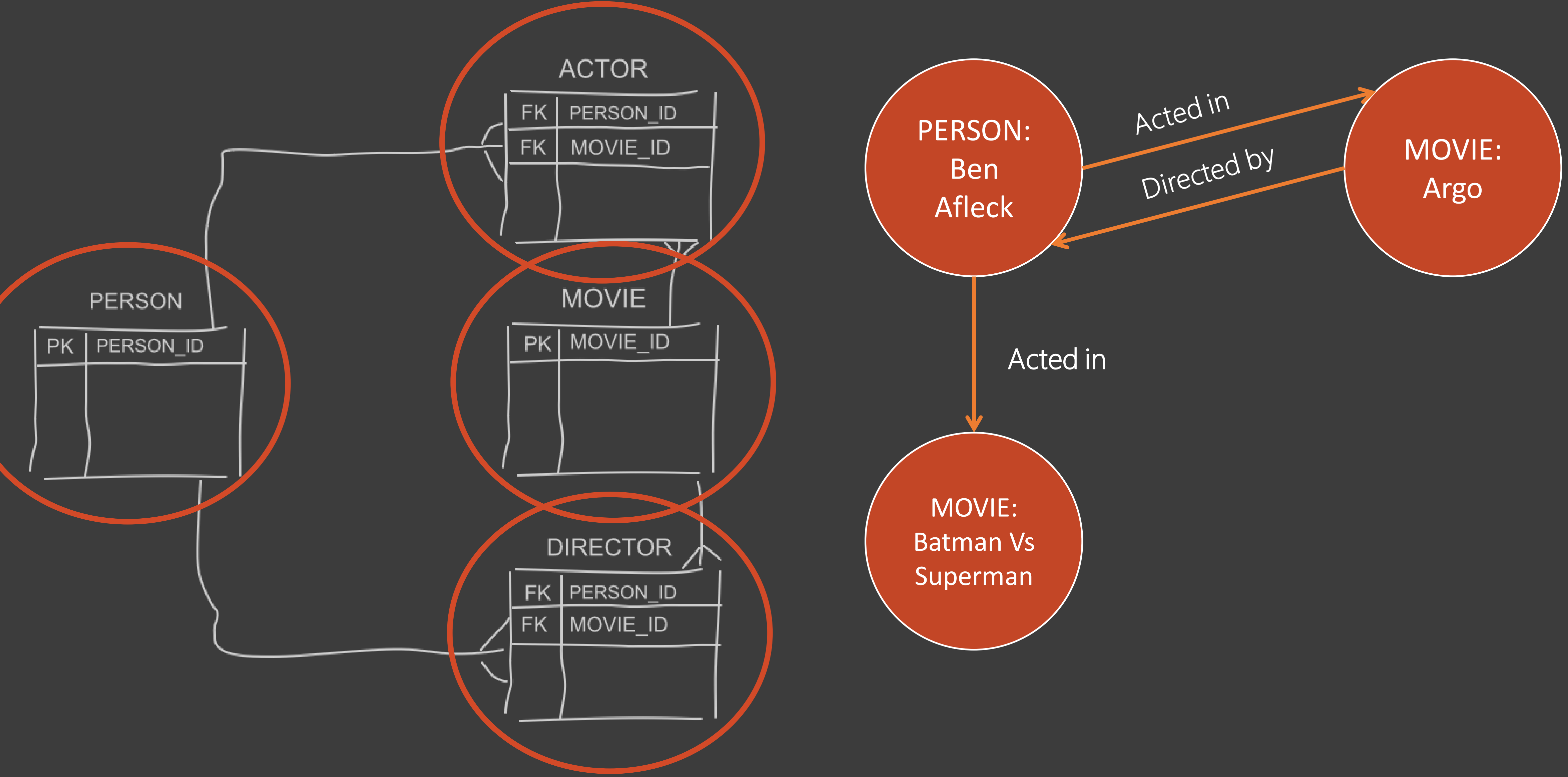
Attributed joins become ***EDGEs*** with properties

EDGE properties


User
Rated
Movie


Ash
Rated Score:9
Alien

Many-To-Many



Many-To-Many

Terry McCann 2017

https://sonra.io/2017/06/12/benefits-graph-databases-data-warehousing/

Entities become ***NODES***

FK/PK relationships become ***EDGES***

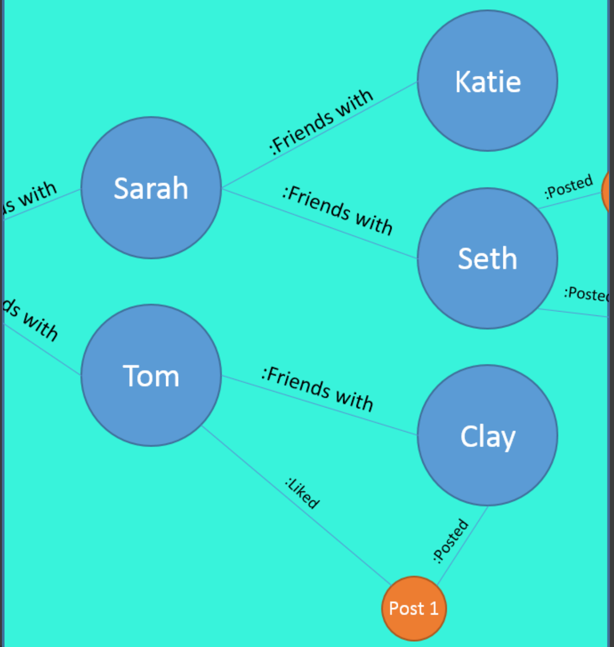
Many-To-Many relationships become ***EDGES***

Attributed joins become ***EDGEs*** with properties

1. Shortest path – What is the shortest path between 2 NODES
2. Transitive closure – Find a NODE *n* hops away – Find the NODE which is 10 hops away from where I am now
3. Polymorphism – Find any NODE connected to another NODE

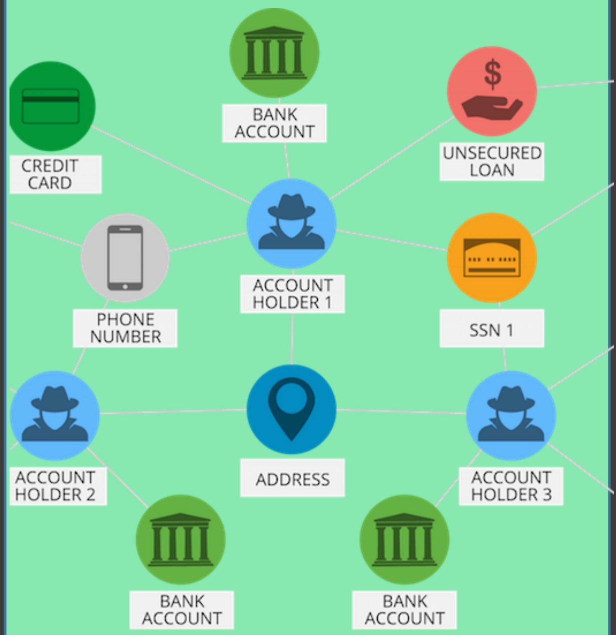
Social Graph

Who knows who?
Spot queen bees



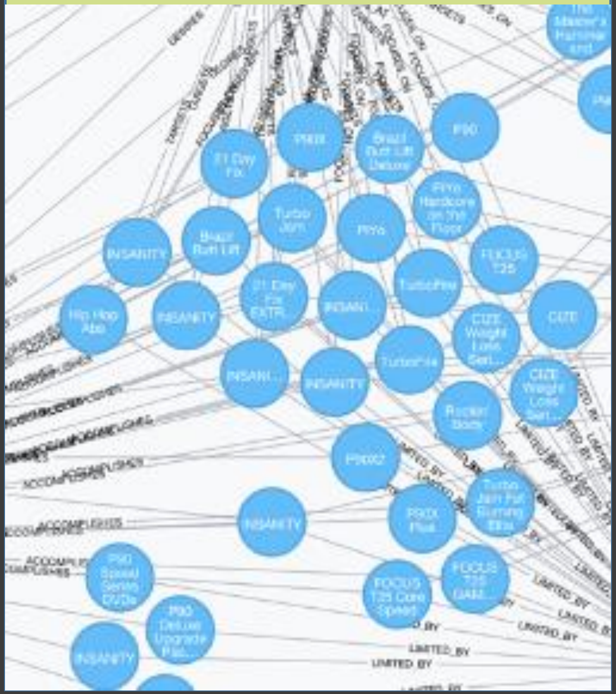
Fraud Detection

Spot rings of fraudulent activity



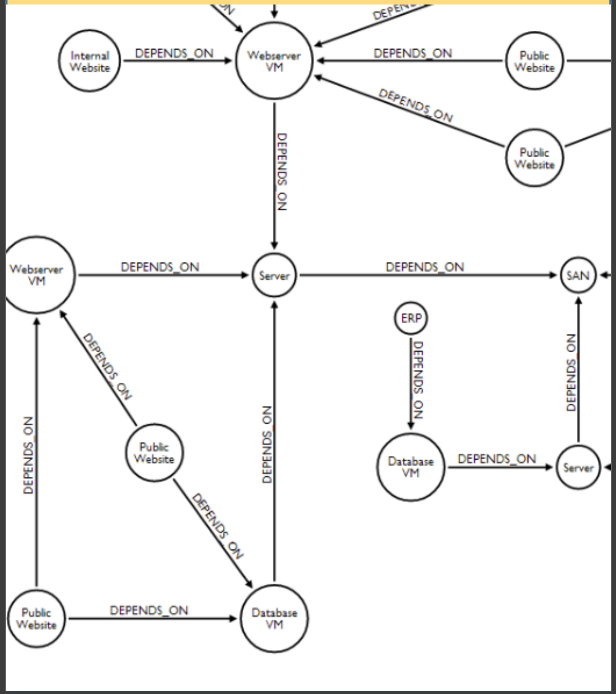
Recommendation

Bob bought X so did Jim, maybe Jim would also like…

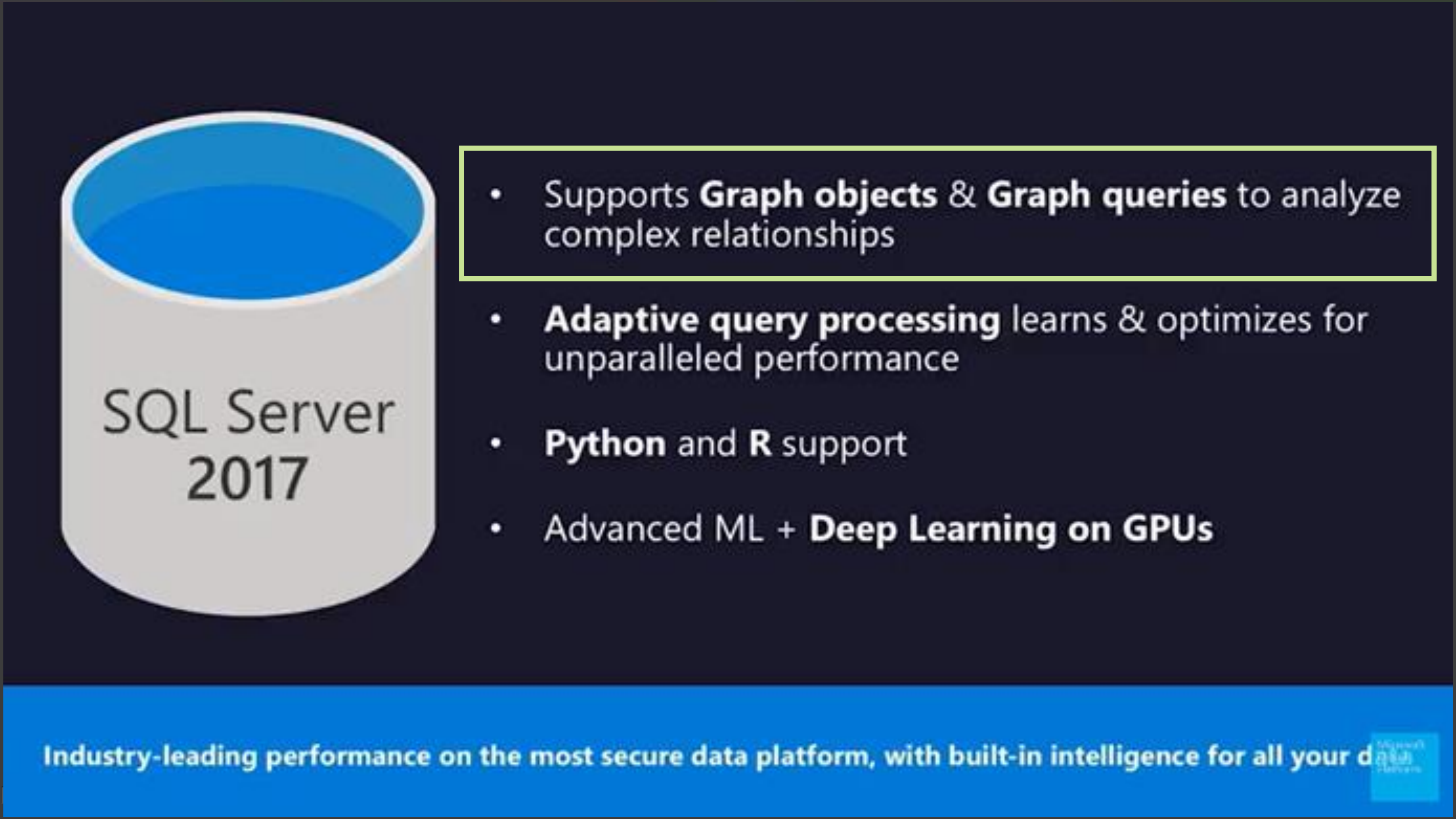


Network & IT

Understand the interconnectivity of your network



How does this relate to SQL Server?


SQL Server
2017
Supports Graph objects & Graph queries to analyze complex relationships
Adaptive query processing learns & optimizes for unparalleled performance
Python and R support
Advanced ML + Deep Learning on GPUs
Industry-leading performance on the most secure data platform, with built-in intelligence for all your

NODE aka vertexes

```sql
DROP TABLE IF EXISTS Genre;
CREATE TABLE Genre (GenreId INTEGER PRIMARY KEY, Genre VARCHAR(100)) AS NODE;
```

A node table represents an entity in a graph schema. Every time a node table is created, along with the user defined columns, an implicit $node_id column is created, which uniquely identifies a given node in the database.

$node_id - It is recommended that users create a unique constraint or index on the $node_id column at the time of creation of node table, but if one is not created, a default unique, non-clustered index is automatically created.

https://docs.microsoft.com/en-us/sql/relational-databases/graphs/sql-graph-architecture

EDGE aka relationship

```
DROP TABLE IF EXISTS ActedIn;
CREATE TABLE ActedIn AS EDGE;
```

An edge table represents a relationship in a graph. Edges are always directed and connect two nodes. An edge table enables users to model many-to-many relationships in the graph.

Each edge table has three pseudo-columns:
$edge_id: The id of the edge record
$from_id: One of the nodes in the edge record
$to_id: The other node in the edge record

MATCH

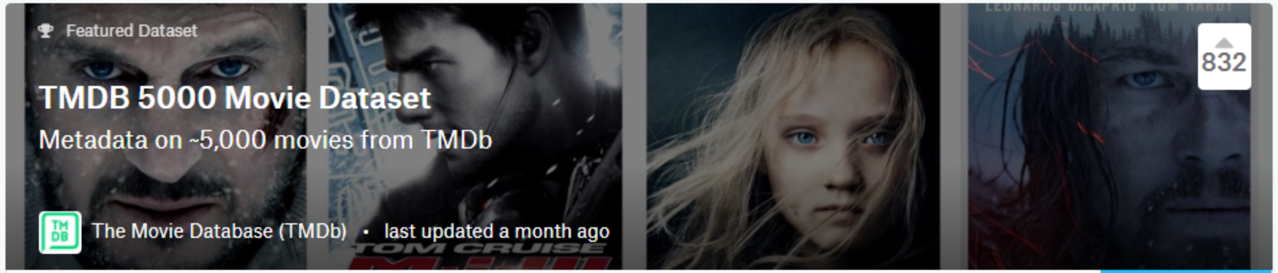
```
SELECT
        M.Movie
        , U1.FullName
        , U2.FullName
FROM dbo.[USER] U1, dbo.[USER] U2, dbo.Movie M, dbo.Reviewed R1,  dbo.Reviewed R2
WHERE MATCH (U1-(R1)->M<-(R2)-U2)
```

Specifies a search condition for a graph. MATCH can be used only with graph node and edge tables, in the SELECT statement as part of WHERE clause.

ASCII ART Syntax based on Cypher
Node-(edge)->Node OR Node<-(edge)-Node

Demo

kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs ••• Sign In



Overview Data Kernels Discussion Activity Download (9 MB) New Kernel

Tags film medium featured

Top Contributors

Anisotropic	1st
FabienDaniel	2nd
AdhokshajaPradeep	3rd

Kernels

Principal Component Analysi... run a month ago	184 votes
Film recommendation engine run 6 days ago	117 votes
EDA with Plotly run a year ago	43 votes

Discussion

Source for opening weeken... 2 days ago	0 replies
Updating datasets and deleti... 10 days ago	1 reply
Principal Component Analy... 10 days ago	52 replies

Description

Background

What can we say about the success of a movie before it is released? Are there certain companies (Pixar?) that have found a consistent formula? Given that major films costing over $100 million to produce can still flop, this question is more important than ever to the industry. Film aficionados might have different interests. Can we predict which films will be highly rated, whether or not they are a commercial success?

This is a great place to start digging in to those questions, with data on the plot, cast, crew, budget, and revenues of several thousand

https://www.kaggle.com/tmdb/tmdb-movie-metadata

Demo

Depth	RDBMS execution time(s)	Neo4j execution time(s)	Records returned
2	0.016	0.01	~2500

MATCH caveats

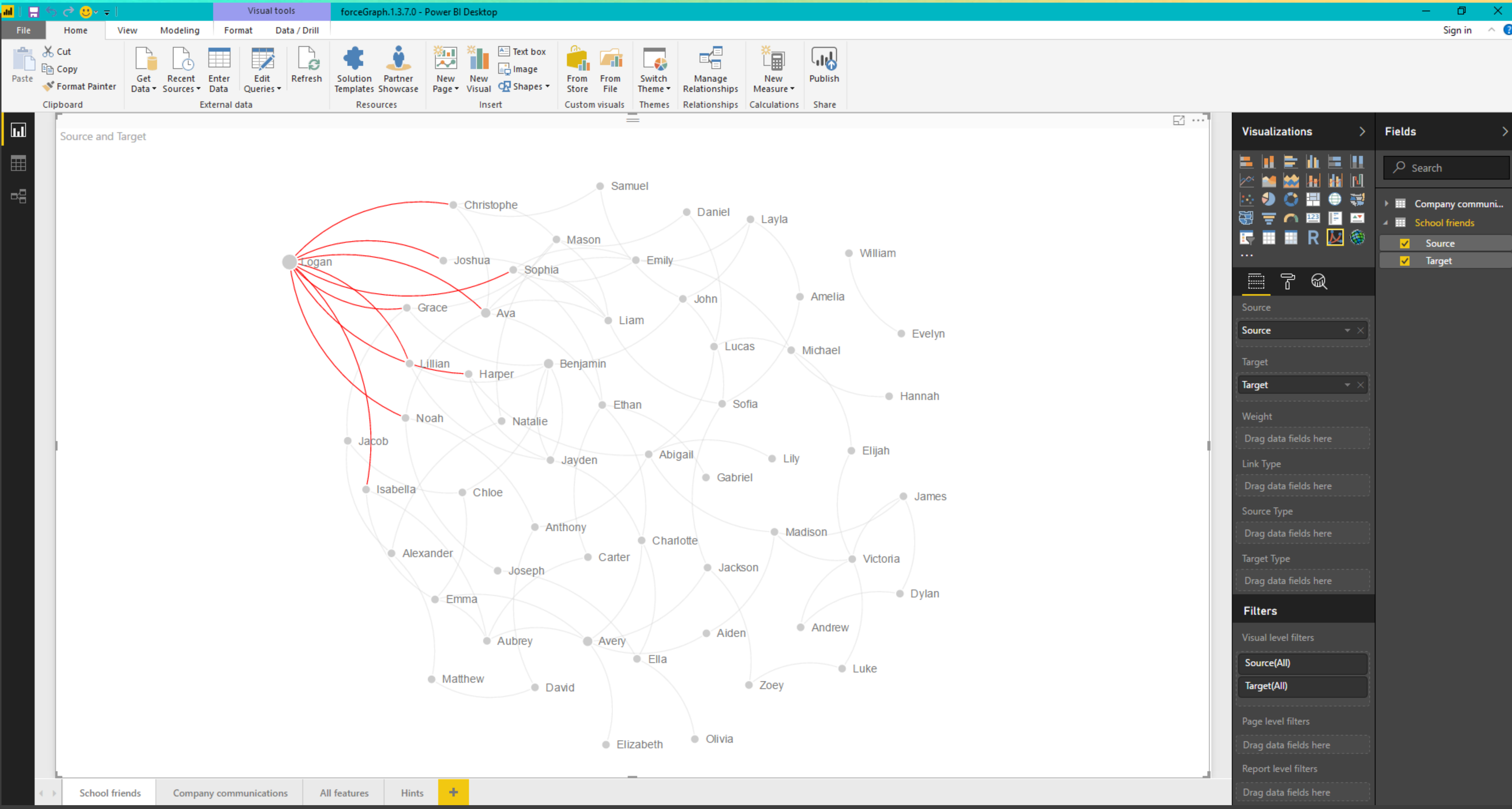
- The node names inside MATCH can be repeated. In other words, a node can be traversed an arbitrary number of times in the same query.
- An edge name cannot be repeated inside MATCH.
- An edge can point in either direction, but it must have an explicit direction.
- OR and NOT operators are not supported in the MATCH pattern. MATCH can be combined with other expressions using AND in the WHERE clause. However, combining it with other expressions using OR or NOT is not supported

Limitations

1. **Shortest path** – What is the shortest path between 2 NODES
2. **Transitive closure** – Find a NODE *n* hops away – Find the NODE which is 10 hops away from where I am now
3. **Polymorphism** – Find any NODE connected to another NODE

Limitations continued

- Local or global temporary tables cannot be node or edge tables.
- Table types and table variables cannot be declared as a node or edge table
- Node and edge tables cannot be created as system-versioned temporal tables.
- Node and edge tables cannot be memory optimized tables.
- Users cannot update the $from_id and $to_id columns of an edge using UPDATE statement. To update the nodes that an edge connects, users will have to insert the new edge pointing to new nodes and delete the previous one.
- Cross database queries on graph objects are not supported.


Visual tools
forceGraph.1.3.7.0 - Power BI Desktop
File
Home
View
Modeling
Format
Data / Drill
Sign in
Paste
Cut
Copy
Format Painter
Clipboard
Get Data
Recent Sources
Enter Data
Edit Queries
Refresh
External data
Solution Templates
Partner Showcase
Resources
New Page
New Visual
Text box
Image
Shapes
Insert
From Store
From File
Custom visuals
Switch Theme
Themes
Manage Relationships
Relationships
New Measure
Calculations
Publish
Share
Source and Target
Samuel
Christophe
Daniel
Layla
Mason
Logan
Joshua
Emily
Sophia
William
John
Amelia
Grace
Ava
Liam
Evelyn
Lucas
Michael
Lillian
Benjamin
Harper
Hannah
Ethan
Sofia
Noah
Natalie
Jacob
Elijah
Abigail
Lily
Jayden
Gabriel
Isabella
Chloe
James
Anthony
Madison
Charlotte
Alexander
Carter
Victoria
Joseph
Jackson
Dylan
Emma
Andrew
Aiden
Aubrey
Avery
Ella
Luke
Matthew
David
Zoey
Elizabeth
Olivia
School friends
Company communications
All features
Hints
Visualizations
Fields
Search
Company communi...
School friends
Source
Target
Source
Source
Target
Target
Weight
Drag data fields here
Link Type
Drag data fields here
Source Type
Drag data fields here
Target Type
Drag data fields here
Filters
Visual level filters
Source(All)
Target(All)
Page level filters
Drag data fields here
Report level filters
Drag data fields here

<Appendix> - Research

- https://www.mssqltips.com/sqlservertip/4883/sql-server-2017-graph-database-example/
- http://sqlblog.com/blogs/john_paul_cook/archive/2017/06/19/modeling-many-to-many-relationships-in-sql-server-2017-graph-database.aspx
- https://stephanefrechette.com/sql-graph-sql-server-2017/#.WVbCP2grKUI
- https://blogs.technet.microsoft.com/dataplatforminsider/2017/04/20/graph-data-processing-with-sql-server-2017/
- https://stackoverflow.com/questions/20979831/recursive-query-used-for-transitive-closure
- https://blogs.msdn.microsoft.com/sqlcat/2017/04/21/build-a-recommendation-system-with-the-support-for-graph-data-in-sql-server-2017-and-azure-sql-db/
- https://github.com/arvindshmicrosoft/MillionSongDatasetinSQLServer
- https://github.com/Microsoft/sql-server-samples/blob/master/samples/applications/iot-connected-car/README.md
- https://www.youtube.com/watch?v=3vleFxDGoEs
- https://www.simple-talk.com/sql/t-sql-programming/sql-graph-objects-sql-server-2017-good-bad/
- https://sonra.io/2017/06/12/benefits-graph-databases-data-warehousing/
- https://social.technet.microsoft.com/wiki/contents/articles/37993.t-sql-graph-based-tables-in-sql-2017.aspx

Graph processing is amazing!

However, SQL Server is not quite there yet! When Graph Processing in SQL Server can rival Neo4j, we will have a very powerful system.

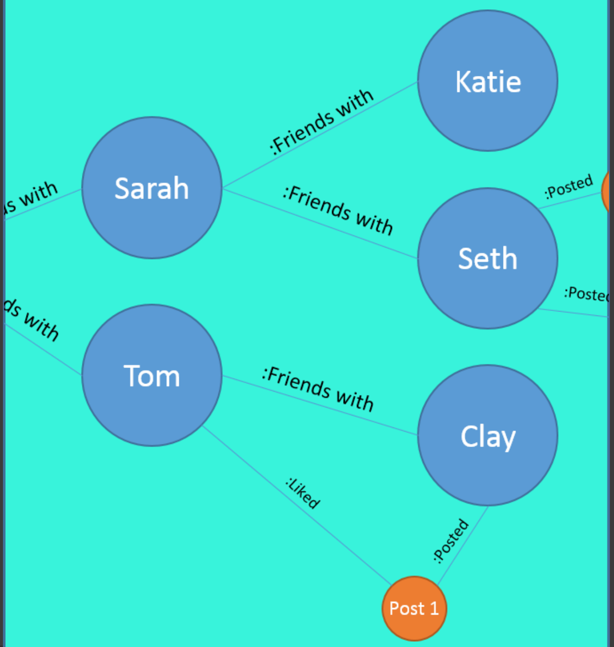
Thank you

@SQLShark
tpm@adatis.co.uk
www.adatis.co.uk
@AdatisBI

So we have extra time…

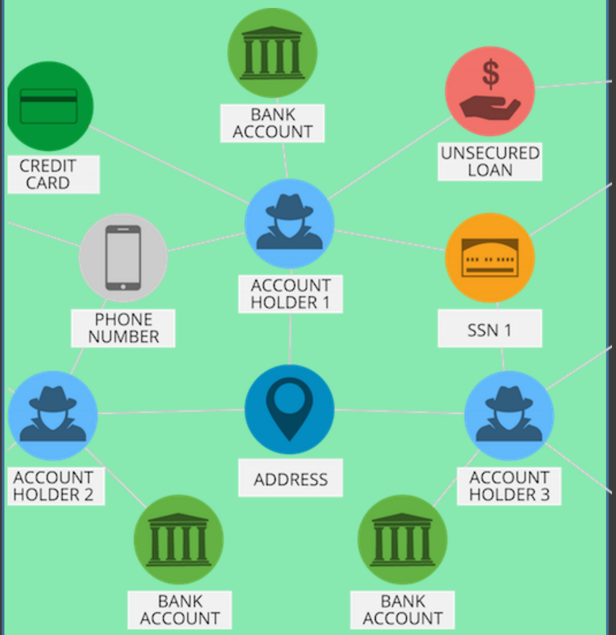
Social Graph

Who knows who?
Spot queen bees



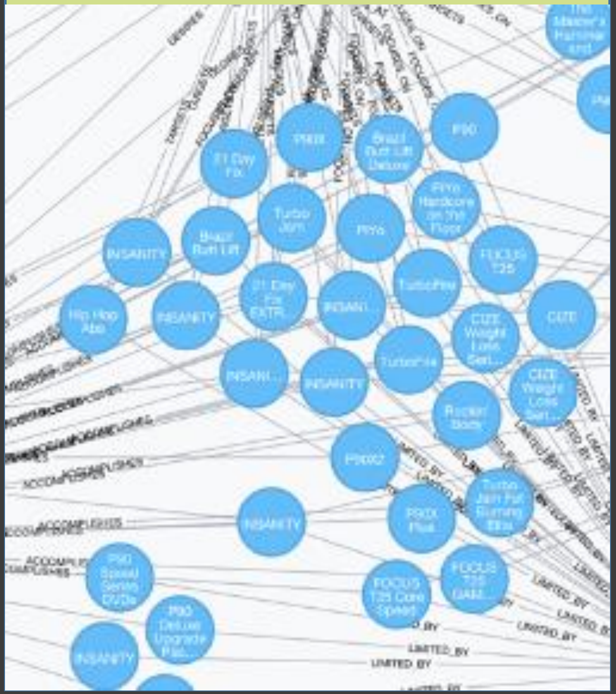
Fraud Detection

Spot rings of fraudulent activity



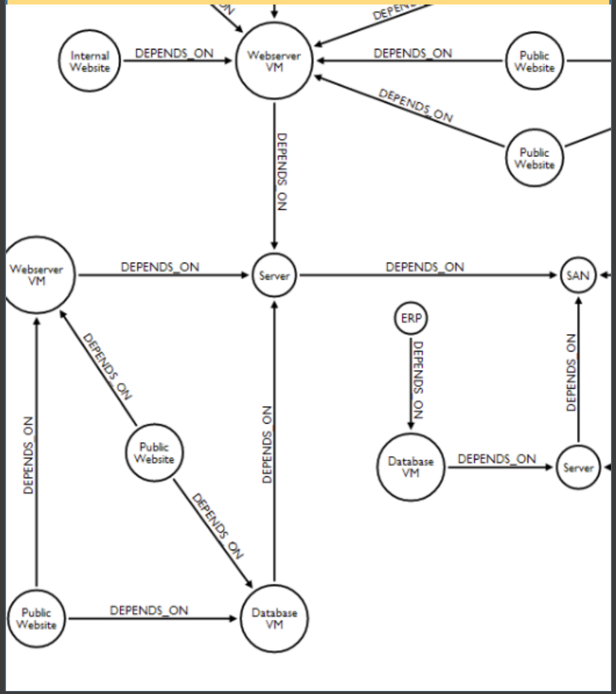
Recommendation

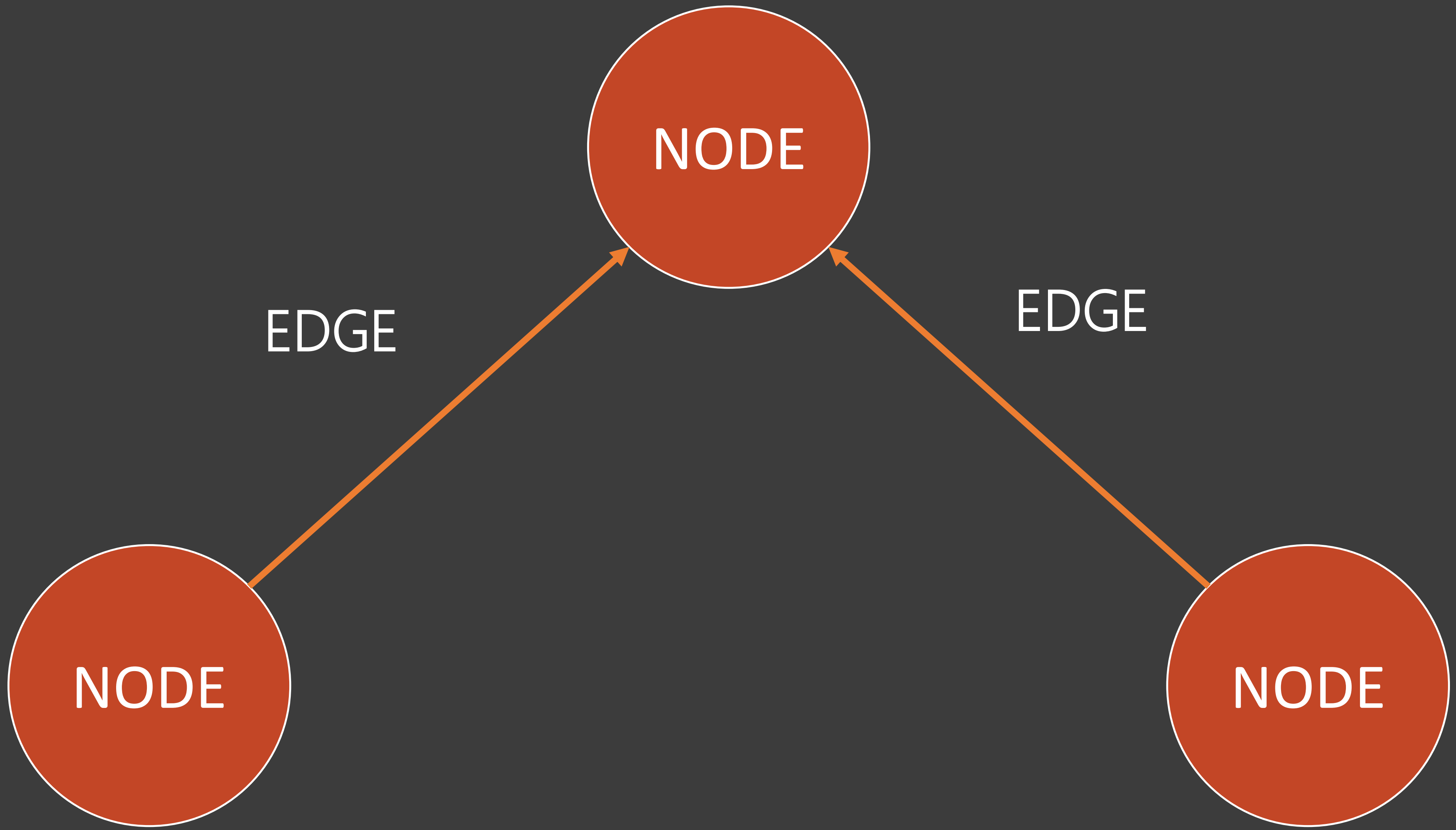
Bob bought X so did Jim, maybe Jim would also like...

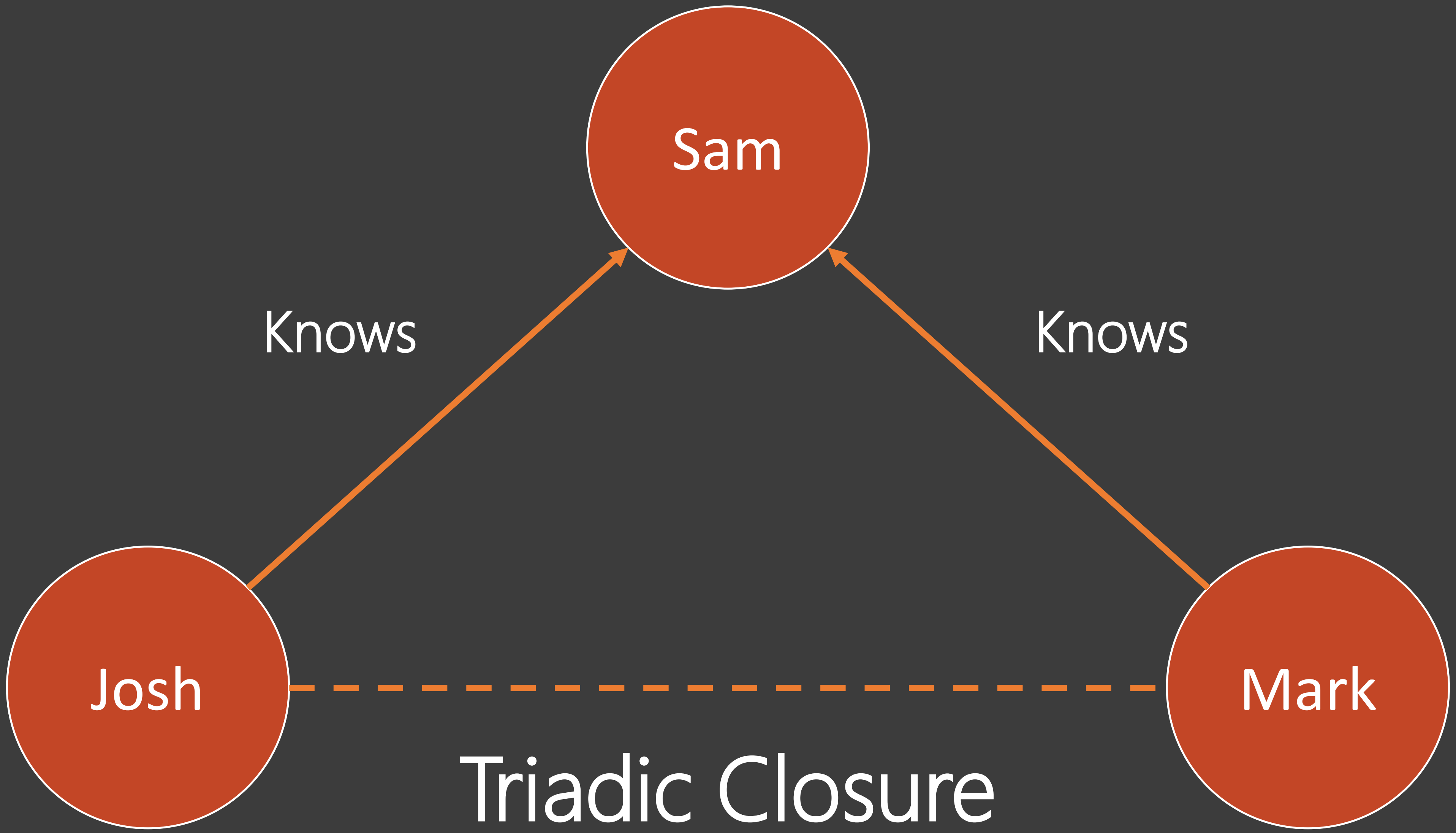


Network & IT

Understand the interconnectivity of your network

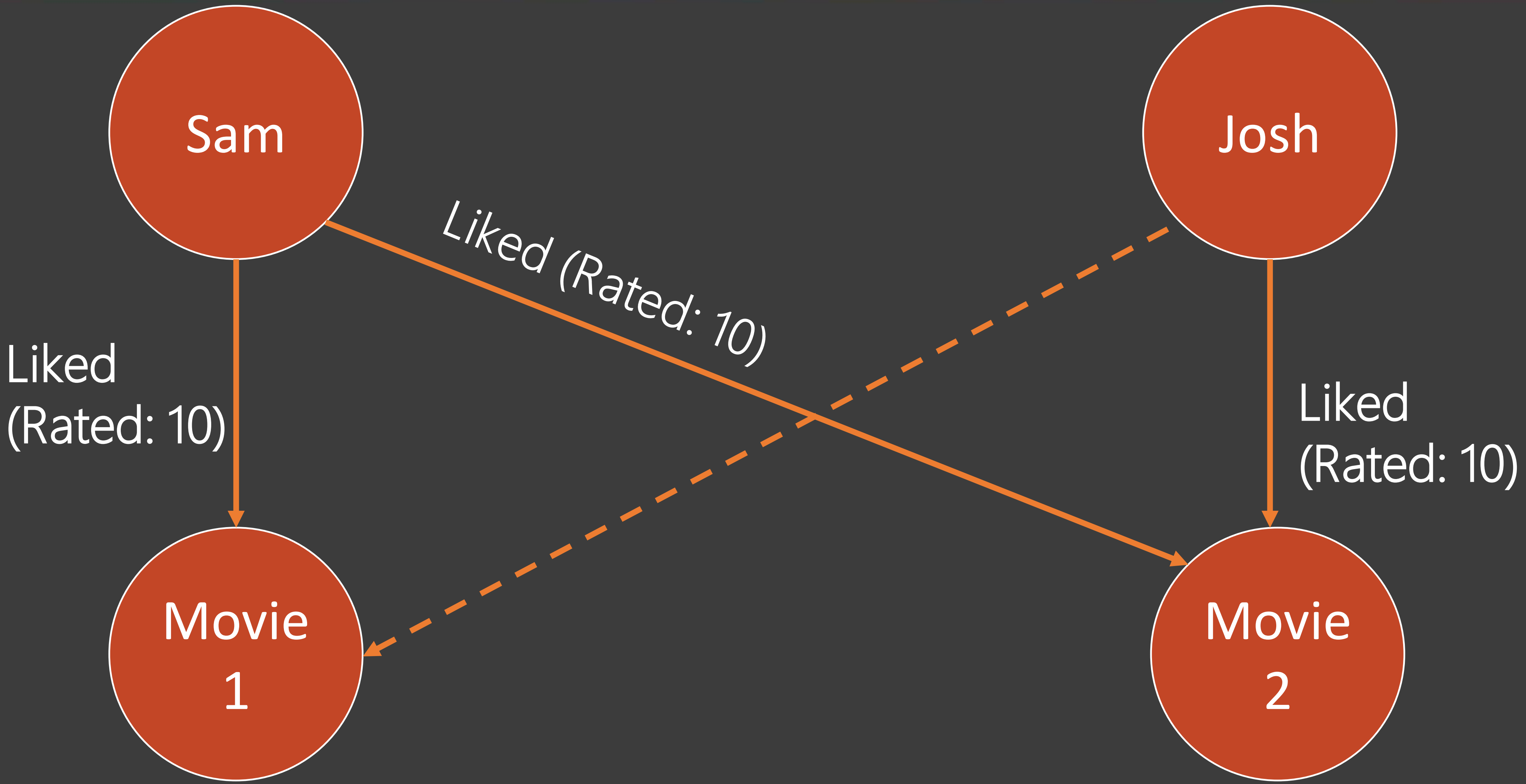



NODE
EDGE
EDGE
NODE
NODE



Triadic Closure

Recommendation Demo



Sam
Josh
Liked (Rated: 10)
Liked
(Rated: 10)
Liked
(Rated: 10)
Movie
1
Movie
2