Evolution of Databases in Big Data

Jubair Hassan

Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA

November 16, 2019

Why do you care?

Jubair said so

Why Do You Care?

"In the third century BC, the Library of Alexandria was believed to house the sum of human knowledge. Today, there is enough information in the world to give every person alive 320 times as much of it as historians think was stored in Alexandria's entire collection – an estimated 1,200 exabytes' worth. If all this information were placed on CDs and they were stacked up, the CDs would form five separate piles that would all reach to the moon."

— Kenneth Neil Cukier and Viktor Mayer-Schoenberger, Foreign Affairs

Summary

Background

Relational Databases

Relational Databases: Tables

- Data stored in tables
- Each row has an unique ID called KEY
- Columns are usually PRIMARY KEYS
- PRIMARY KEYS are an unique key for each record
- FOREIGN KEYS are columns in a table that refers to the PRIMARY KEY of another table

Relational Databases: Relational Model

Student ID	Name	Age
1	Jubair	25
2	Ahnaf	23
3	Dennis	23
4	Victor	23

Subject ID	Name	Teacher
1	Calc 1	Roberts, D
2	Calc 2	Roberts, D
3	Databases	McPhee, N
4	Software D.	KK, McPhee, N

Student ID	Subject ID	Marks
1	1	92.9
1	2	93.1
2	1	93.2
3	2	94.6

Figure: Relational Model

”Join” Statements

Relational Databases: "Join" Statements

```
SELECT
Orders.OrderID,
Customers.CustomerName,
Orders.OrderDate
FROM Orders
INNER JOIN Customers ON
Orders.CustomerID=Customers.CustomerID;
```

Figure: Sample Join Code

- Used to combine data from multiple tables

Code Snippet:

- OrderID from Orders
- CustomerName from Customers
- OrderDate from Orders
- CustomerID from Orders is joined with CustomerID in Customers
- Gives you NAMES of CUSTOMERS with the DATE of the ORDER and its ID

Relational Databases: "Join" Statements

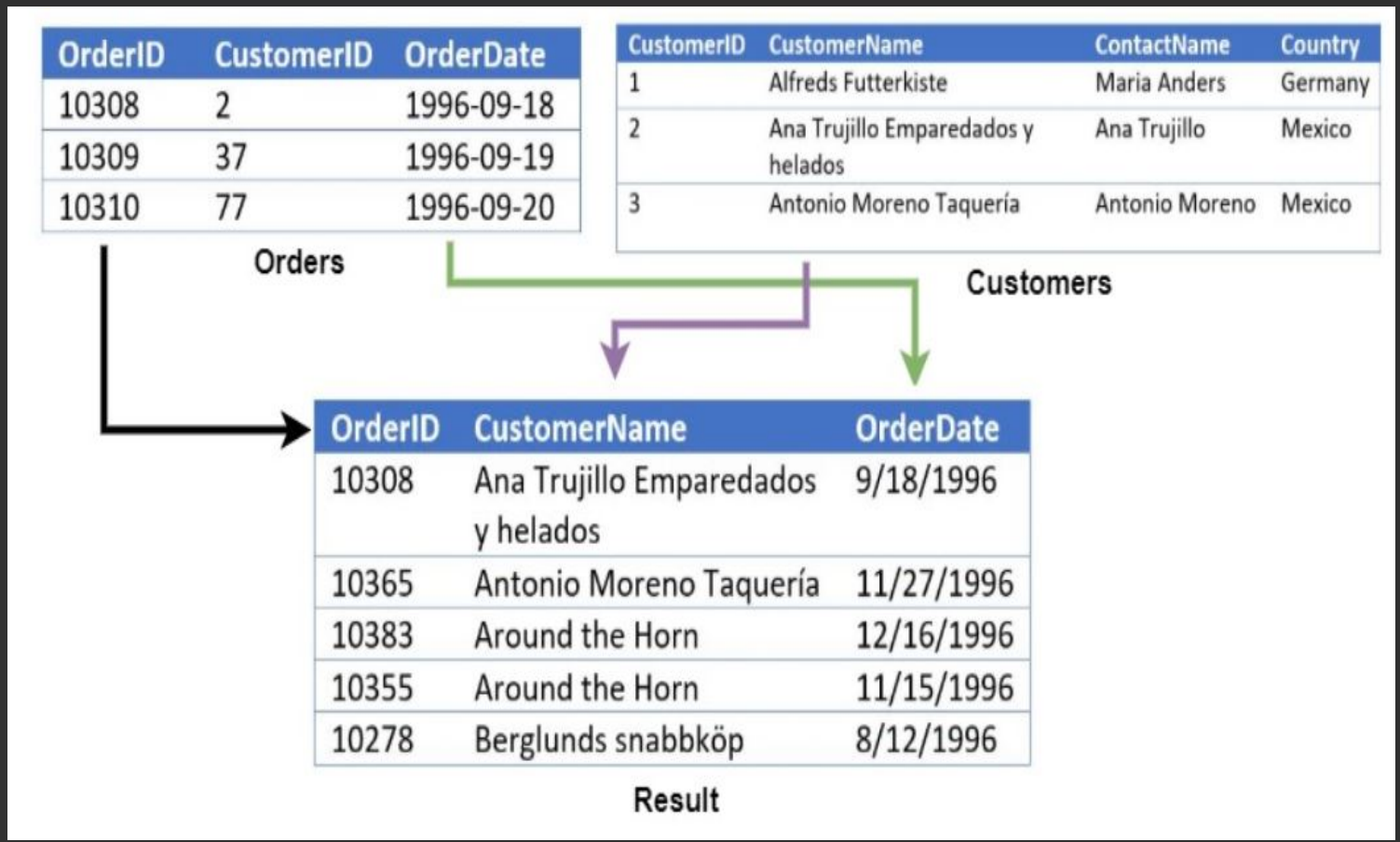

Figure: Source - W3Schools

Scalability

Relational Databases: Scalability - Horizontal Scaling

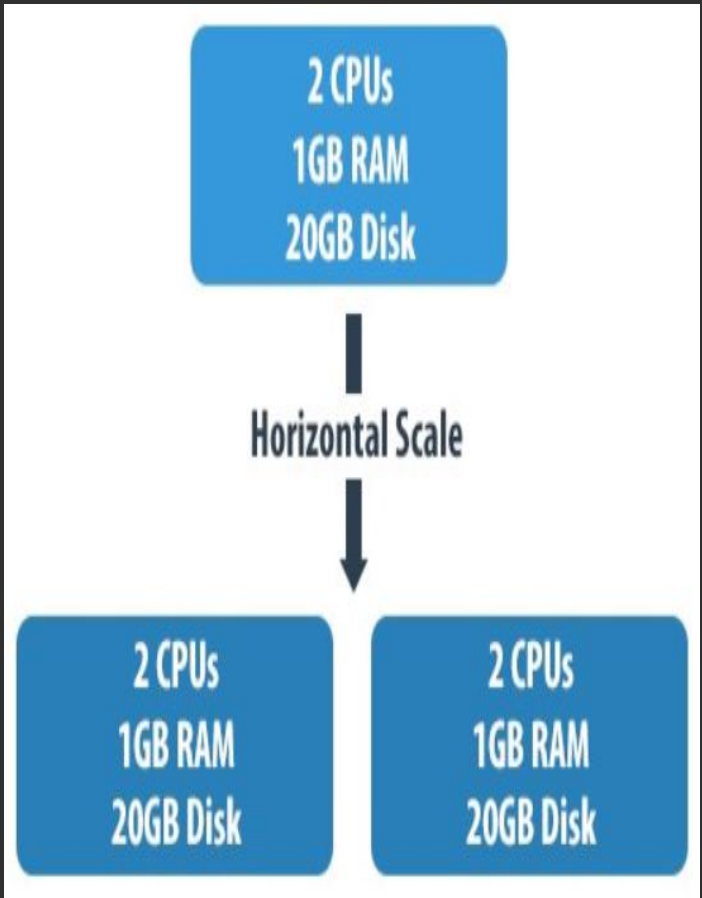

Figure: Source - Packt

- Add more machines
- Distribute data

Pros:

- Cheap
- Less load, better performance

Cons:

- Joins are harder

Relational Databases: Scalability - Vertical Scaling

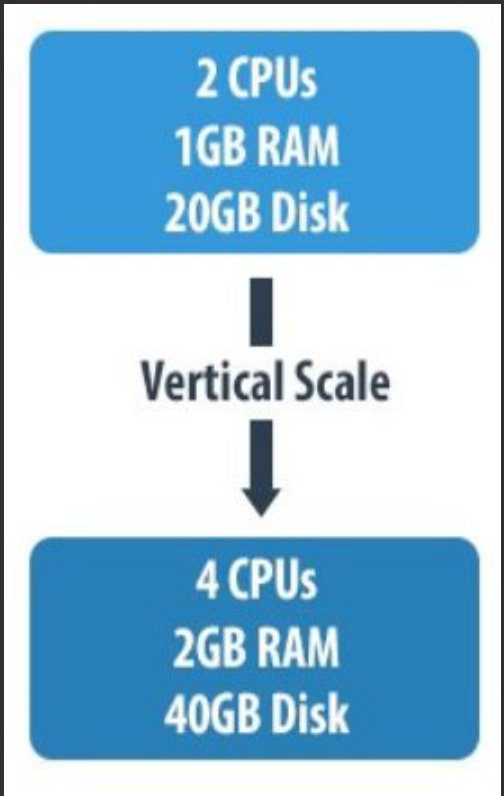

Figure: Source - Packt

- Upgrade to a more powerful machine

Pros:

- Simple
- Better performance

Cons:

- Multiple queries are harder to perform
- Expensive

Graph Databases

Graph Databases: Nodes and Relationships

- A NODE represents an entity (a person, place, thing, category or other piece of data)
- A RELATIONSHIP (vertices) represents how those two data are connected
- Relationships are prioritized
- The connections are always there
- No such thing as Foreign Keys

Graph Databases: Graph Model

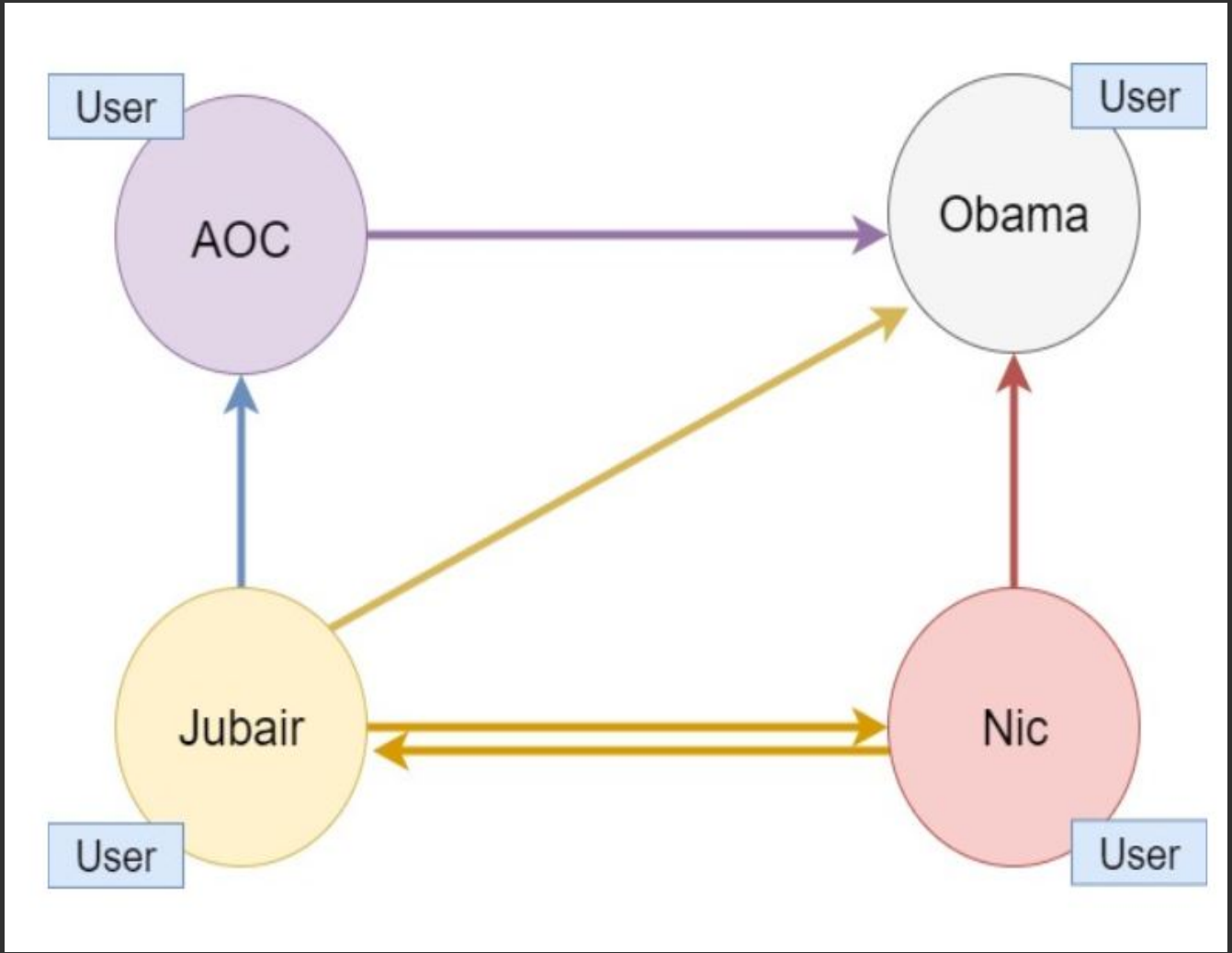

Figure: A Graph Model

Graph Databases: Important Properties

Graph Storage:

- Some graphs use native graph storage - a system designed to manage and store graphs
- Very fast and efficient

Graph Processing:

- The native graph processing is the most efficient way to process graphs
- The nodes physically point to each other in the database

Relational Databases in Big Data

Big Data

Big Data

- Large, diverse sets of information that is ALWAYS growing, EXPONENTIALLY
- Comes from multiple sources
- The three Vs:
  - The volume of information
  - The velocity or speed at which it is created and collected
  - The variety or scope of the data points being covered

Drawbacks of Using Relational Databases in Big Data

Drawbacks of Using Relational Databases in Big Data

- They do NOT scale well to very large sizes of data
- They do not handle unstructured data well (i.e. google type searching)
- It is harder to query some basic functions using Relational Databases (i.e shortest path between two points)
- Need to use more "join" statements and that decreases efficiency

USE GRAPH DATABASES

Graph Databases in Big Data

Comparative Analysis Study

Comparative Analysis

- Study by Unal and Oguztuzun [8]
- Study comparing MySQL and Neo4j
- Uses a Law Document System
- Eighteen data entity types and three level hierarchy for each data type

Comparative Analysis

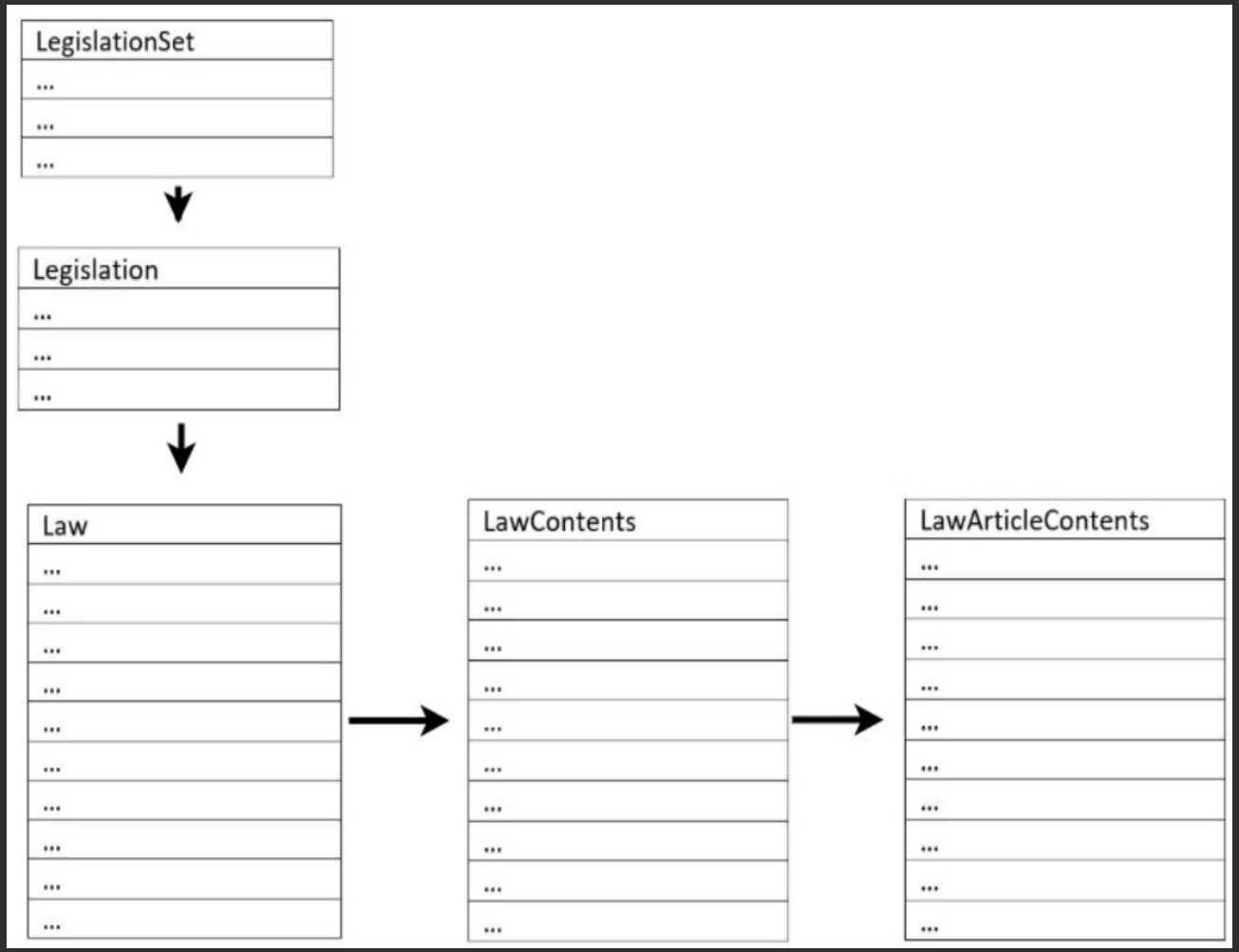

Figure: Relational Model of the Domain [8]

Comparative Analysis: Transformation Process

- Each entity table becomes a label on the nodes
- Each row in an entity table becomes a node
- Columns become node properties
- Join tables transformed to relationships [8]

Comparative Analysis

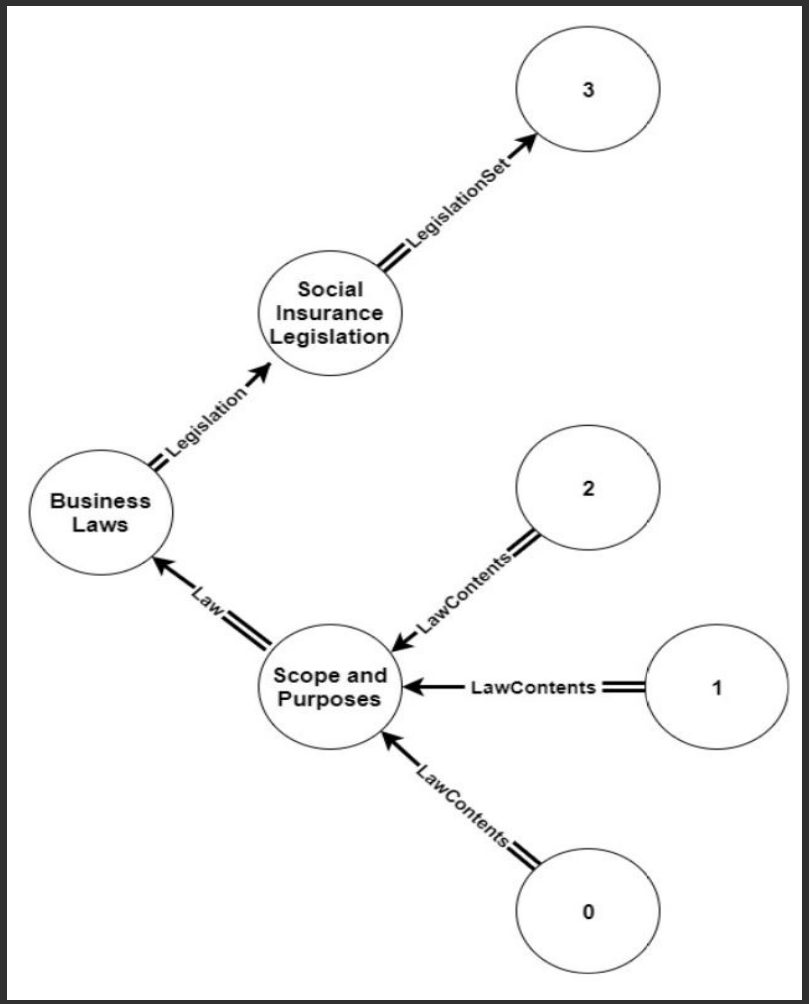

Figure: Graph Model of the Domain [8]

Results

Result of the Comparative Analysis

Result of the Comparative Analysis

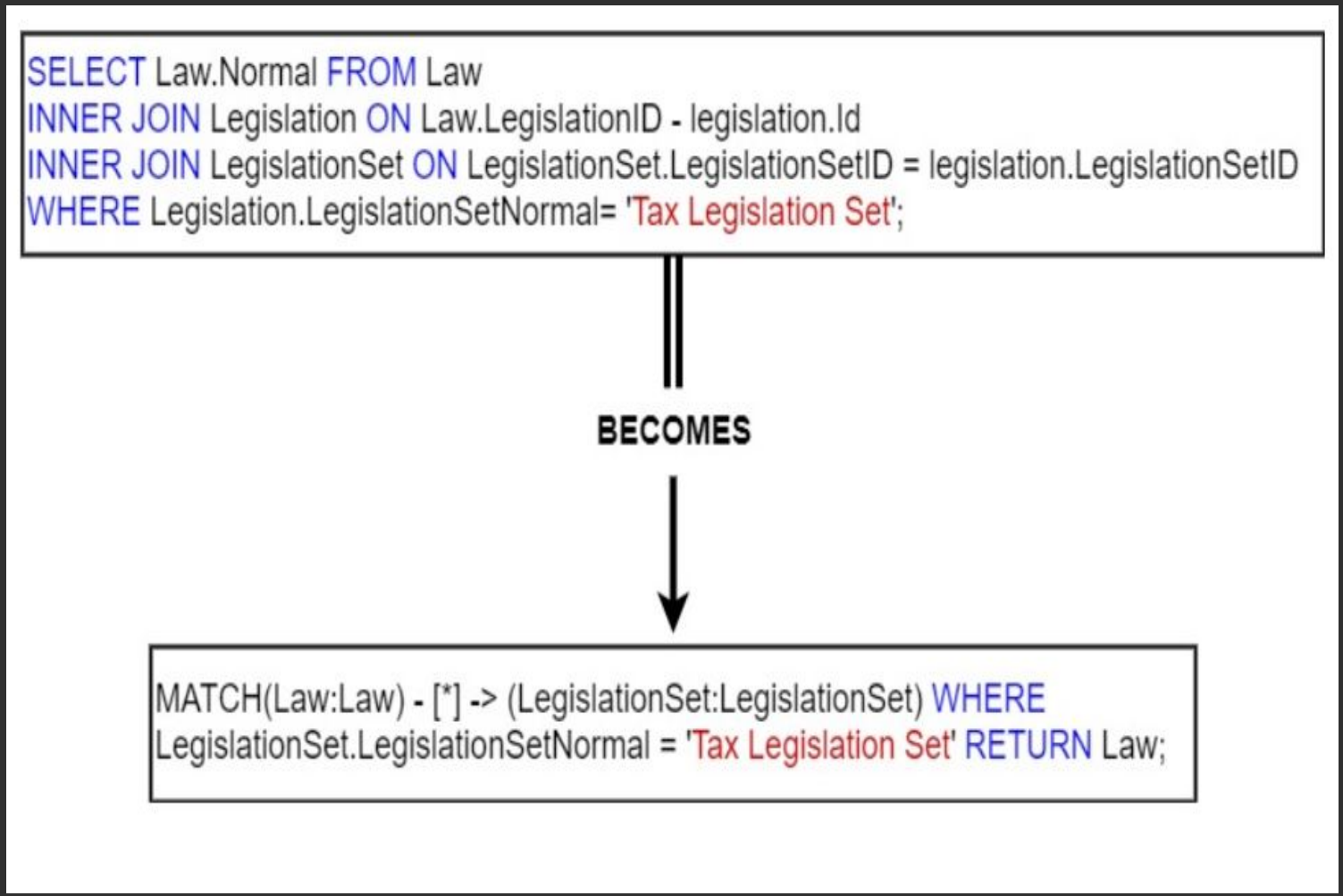

Figure: Two JOINS to NO JOINS! [8]

Result of the Comparative Analysis

- Same data was queried
- Data was retrieved TEN TIMES faster

Result of the Comparative Analysis

In the graph model, data was accessed:

- SIX TIMES faster when there were a thousand records
- THIRTY TIMES faster when there were ten thousand records

Graph Database worked better

Walmart Case Study

Walmart Case Study

What the customers wanted:

- Personalized suggestions while shopping online

The challenge:

- Need to go through the session browsing history in the website
- Need to go through connected products

The problem:

- These are complex queries for Relational Databases

What to do?

Use Graph Databases!

For Graph Databases:

- All the challenges mentioned are handled very easily

The Solution and Benefits:

- Started using Neo4j since 2013 after a one year trial period that yielded very successful results which meant it gaves very useful recommendations with low latency [10]

Graph Database worked better

Alternative Options

The Hybrid Approach

- Proposed by Vyawahare et. al.
- Use a combination of two databases: MySQL(relational) and Neo4j(graph)
- Would cater to ones who do not want to let go off Relational Databases
- Still in developing phases
- No concrete research done

The Hybrid Approach

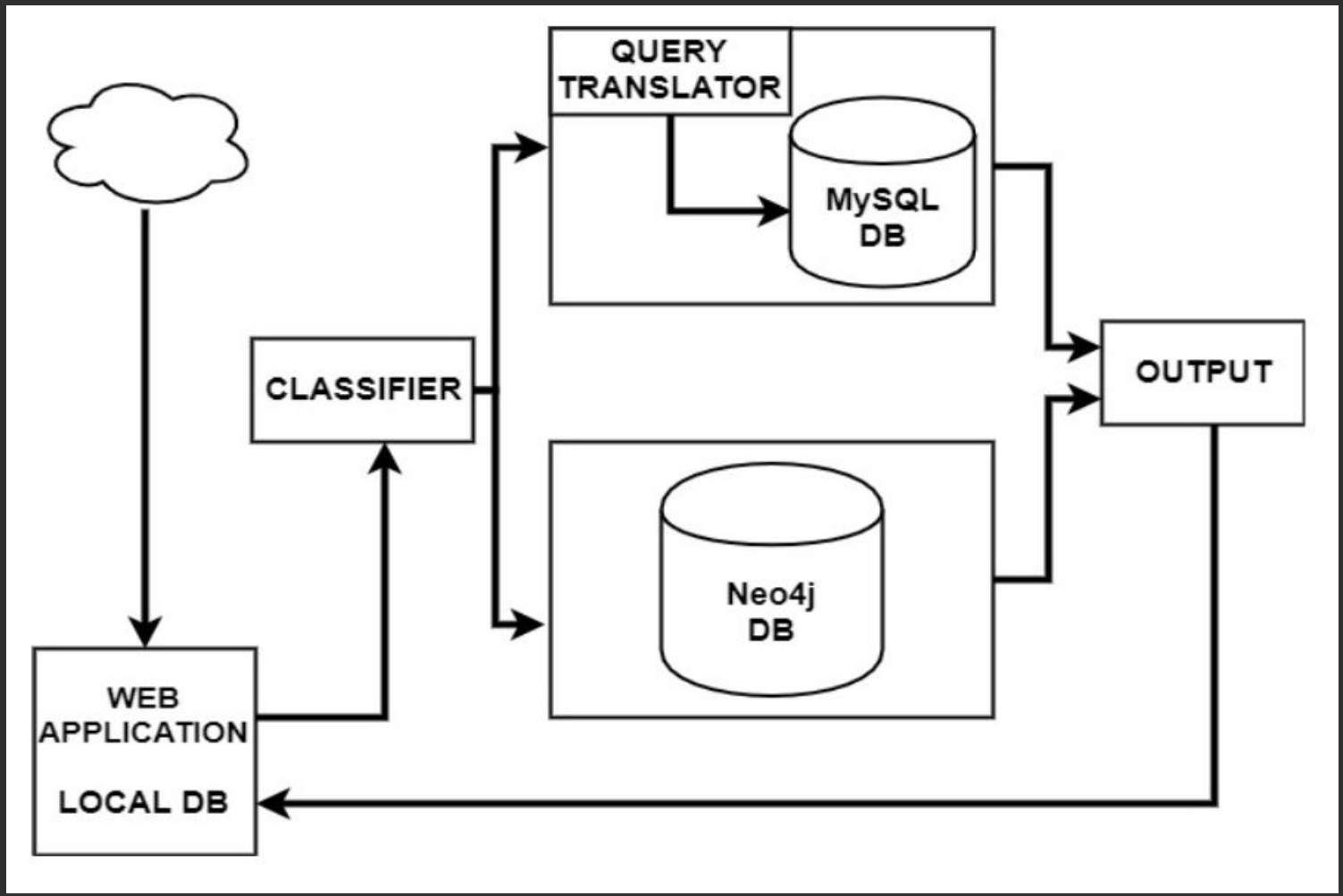

Figure: The Hybrid Model [8]

Conclusion

Conclusion

- Relational Databases has its uses
- Graph Databases are better for handling Big Data
- They scale well and are very fast at querying large amount of data

Graph Databases are better at handling Big Data!!

Acknowledgments

Prof. Hussam Ghunaim, Prof. Kristin Lamberty and Prof. Nic McPhee
Friends and Family

References I

- *Big Data.* http://https://www.investopedia.com/terms/b/big-data.asp. Accessed: 2019-10-27.
- *Big Data.* https://infocus.dellemc.com/april_reeve/big-data-and-nosql-the-problem-with-relational-databases/. Accessed = 2019-11-12.
- *Database Scaling.* https://hackernoon.com/database-scaling-horizontal-and-vertical-scaling-85edd2fd9944. Accessed = 2019-11-12.
- *Graph Databases for Beginner: Why Graph Technology is the Future.* https://neo4j.com/blog/why-graph-databases-are-the-future/. Accessed: 2019-10-22.
- *Neo4j.* https://neo4j.com. Accessed = 2019-11-12.

References II

Relational Databases Are Not Designed For Scale. https://www.marklogic.com/blog/relational-databases-scale/. Accessed: 2019-10-27.

Scaling Horizontally and Vertically for Databases. https://medium.com/@abhinavkorpal/scaling-horizontally-and-vertically-for-databases-a2aef778610c. Accessed: 2019-10-27.

Unal, Y. & Oguztuzun, H. *Migration of Data from Relational Database to Graph Database.* in *Proceedings of the 8th International Conference on Information Systems and Technologies* (ACM, Istanbul, Turkey, 2018), 6:1–6:5. ISBN: 978-1-4503-6404-1. http://doi.acm.org.ezproxy.morris.umn.edu/10.1145/3200842.3200852.

References III

Vyawahare, H. R., Karde, P. P. & Thakare, V. M. *A Hybrid Database Approach Using Graph and Relational Database.* in *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)* (Aug. 2018), 1–4.

Walmart Case Study. https://neo4j.com/case-studies/walmart/?ref=blog. Accessed = 2019-10-22.

Thank You!

Questions?