June 10, 2018

Course: CIS570 – Business Intelligence
Name: Robert Palumbo
Assignment: Reading Discussions – Week3-Session2
Due Date: Sunday, June 10 @ 11:59pm

Question 1. What is a *graph* database?

Simply put, a *graph* is a visual method that is used to show relationships between entities. In a traditional graph two entities or *nodes* are connected together with an *edge.* Entities can also be referred to as *points* or *vertices,* while the connections are also referred to as *lines* or *arcs.* A complex graph can have any number of interconnected nodes with complexity increasing as more nodes and edges are included. Further, nodes can have *properties* or *attributes* associated with them that provide context, while edges can have *direction* and *weights (importance).*

The key to a graph database is how it stores these relationships in an optimal way taking advantage of the very nature of the node and edge model. As we know, a traditional RDBMS maintains relationships between entities using tables and primary/foreign keys. At query time, these tables must be *joined* together in order to obtain the data which is applicable to the query. The join operations are quite expensive in terms of processing time and space. Further, the rigged structure of the schemas themselves make it difficult to represent the complex relationships and connectedness of our new world data.

Graph databases arose as an alternative method of capturing the essence of these complex data relationships and to do it in the most efficient manner as possible. In this type of database, the relationships themselves (connections) are of the same importance or value as that of the nodes. They are both treated as equal such that any operation that can be performed on a node can also be performed on a relationship. By doing this it makes it possible to explore and discover *hidden* relationships within the data leading to new realizations or findings.

This processing is achieved because in a graph database every node maintains a physical list of its connections to other nodes in the graph. When a node is retrieved by the database it has immediate access to the other connected nodes just by using the references in this list. The costly overhead of the indirect reference using a foreign key and join is removed from the process making an equivalent graph database query significantly faster as compared to an RDBMS.

Neo4j is among the most popular implementations of this type of database. Also, for reference, Gephi is a data visualization tool used to create visualizations of data which is based on node/edge pairing and is a great tool if you would need to create a visual representation of this type of data.

https://www.tibco.com/blog/2017/11/28/what-is-a-graph-database-and-why-is-it-important/
https://neo4j.com/developer/graph-db-vs-rdbms/
https://neo4j.com/
https://gephi.org/

Question 2. Identify an industry (other than agriculture), and discuss how big data analytics is transforming it.

Anadarko is an oil and gas company based out of Houston Texas with offices here in Denver and Platteville CO. The company has hundreds of oil wells across areas of Northern Colorado that are networked together which provide refined resources that we use daily. The well system has been developed over time and encompasses thousands of acres of land.

As previously mentioned in a related post, the associated costs with having to inspect and monitor this network of wells prior to the use of big-data was extreme. Teams were created whose sole purpose was to travel to each of the well sites and inspect the equipment for any signs of potential failure and to take preventative corrective action. Further, if a well happened to go into failure mode and stopped functioning, the latency between detecting and correcting the situation could amount to many hours or days of downtime leading to considerable loss of revenue for that well.

Anadarko and the oil and gas industry are now taking huge advantage of big data analytics to assist with these types of industry challenges. Today, modern well systems are making use of sensor driven diagnostics and IoT. Critical components of a well are now monitored using sensors and lasers which track the movement of parts (e.g. horsehead rotation), monitor inline pressures, temperature, resources flow, and many other important metrics. Many systems use cellular or other wireless technologies to allow these systems to be web enabled. Thus, the data can be live streamed to a central location employing OLAP providing near real-time analytics. As a result, better informed decisions can be made for directing applicable resources when anomalies are encountered during system processing. Through this process, a significant and tangible cost savings is now being realized.

Another aspect of how big data is used within this industry is with oil production itself. Seismic data is collected from the drilling and production processes which is routinely analyzed and used to make adjustments in the oil extraction process. This same data is also used to forecast oil production from the well allowing adjustments to be made to the process if expectations are not being achieved. And finally, similar seismic data is generated and used to locate new oil deposits which have not yet been tapped which leads of course to identification of potential new drilling locations.

https://mapr.com/solutions/industry/oil-and-gas-use-cases/
https://www.anadarko.com/