Assignment 7

Isaac Shoebottom (3429069)

Chapter 14 Review Questions

1.

Big data is defined by data that is so large in quantity or is being created so fast or in such variance that is requires systems other than relational databases to process/deal with this data being stored. This situation is known as polyglot persistence.

2.

Volume, velocity, and variety are known as the 3 V's of Big data.

Volume refers to the quantity of the data, referring to taking up a large size on disk. In terms of a relational database these might be known as a table with millions of entries

Velocity refers to the rate at which data is being defined. This could be data being entered into a database every millisecond, which might be a limit of disk IOPS or the CPU being able to process the entry of this into a relational database.

Variety is how data might come from an email, a web form, a physical letter of paper form or a desktop application.

4.

Scaling up is the process of adding faster hardware to your database, such as replacing your disks with SSD's or upgrading to the latest Xeon/EPYC CPU. Scaling out is the process of adding more servers and load balancing between these two computers, often connected over the internet, this might be more expensive than scaling up, but it reduces down-time as a new server can be added to a pool and become the responsibility of the load balancer immediately. Scaling up makes sense on something that is fine to go down when upgrades are taking place, or if purchasing a new server might be more of a burden than buying a new CPU or drives.

13.

The four common types of NoSQL databases are key-value databases which tend to simply store data with no indication of the kind of data it is storing, document databases which store data in a document format like XML or JSON which then become able to be queried on, column-oriented databases which store data in key-value pairs in columns which store that key-value as a key value to the column, and graph databases which represent the data as nodes edges and properties, which might be easier to visualize as something similar to a Venn diagram with each bubble being similar to a table, edges being similar to a foreign key and the properties the data fields of the bubble.

14.

A document vs key-value database differ in that documents are inherently grouped and become much easier to query that pure key-value pairs. In just a key-value pair system, there is no relation between two types of data. It primarily unstructured in that often two kinds of key value pairs cannot be linked together, and each key-value pair is independent from each other. It is most similar something akin to a

hash map in computer science, which offers no ability to express a relation to other hash maps other than to define you own data structure with does specify a relation in a custom format. A document however inherently implies a relation between to kinds of data due to the nature of it existing in the same file. This makes a document database much easier to query as well.