SQL Written Component – Assignment Part 3

Question 1.

To compare the features and functionality of open-source and proprietary software for database design and management, I will compare the features of MySQL (Open-source) to Oracle (Proprietary).

Oracle is a proprietary software with the largest single market share than any other DBMS. They have a long list of compatible operating systems including Windows and many versions of Linux. First developed in 1977, Oracle has set the standard for a plethora of features seen in many other DBMS services and has developed a well established following of people over the years who are able to influence the design and implementation of new features – as users are the core of oracle and large sums of money are on the line. Users include: Bauerfeind AG, CAIRN India, Capcom Co., ChevronTexaco, Coca-Cola FEMSA, COOP Switzerland, ENEL, Heidelberger Druck, MTU Aero Engines, National Foods Australia, Spire Healthcare, Stadtwerke München, Swarovski, Tyson Foods, TVS Motor Company, Vilene.

MySQL is an open-source community driven software that includes many features similar to Oracle. It is one of the most popular open-source programs available due to a plethora of features similar to Oracle that comes in a well rounded package that is capable of running large companies like YouTube and EBay. Developed in 1995, MySQL has a following similar to Oracle many of which switched from Oracle due to the cost of the license and the community driven development of the system. Users include: GitHub, US Navy, NASA, Tesla, Netflix, WeChat, Facebook, Zendesk, Twitter, Zappos, YouTube, Spotify.

Cost is an important consideration when it comes to choosing what system you want to use as your main method of storing and accessing data, especially if your company relies on constant and reliable access to said data. MySQL clearly has an advantage here as it is free (for non-commercial use) compared to Oracle which has a very high charge for the license, for example – a person wanting a database for personal use would pay USD $17,500 for a license and USD $3,850 for an update and support license (for Oracle Standard Version 2), while MySQL is free for personal use. For business use, a company would pay USD $47,500 for the license and USD $10,450 for the update and support license (for Oracle Enterprise Edition pricing), while the same company would pay between USD $4,574.02 and $13,722.07 (for MySQL Enterprise Edition Subscription (1-4 socket server)). Note – All prices are correct as of 2nd January 2019. With the extra cost of Oracle comes extra services. As Oracle are directly responsible for the development of all features within the software, tech support is much easier and at a better quality due to a more intimate knowledge of the software they are supporting – while MySQL is more supported by the community. MySQL's website, like that of Oracle, has lots of documentation to support the software but not to the same extent. The Community page for MySQL version downloads have 10 different versions alone. This makes it harder to support as there are many minute differences between them which may drastically alter the support a user may require.

Integration with existing systems is another important thing to consider when choosing what software you want to use. E.G. if you only use Windows you may choose Microsoft SQLServer or Microsoft Access as these are designed specifically around the Windows Operating system. Of MySQL and Oracle, both are designed with full releases of Windows in mind. However, if you use a very niche version of Linux MySQL may be the better choice as someone in the community has likely used MySQL with the same version of Linux. Oracle may run on the Linux version however it may be unstable or may work just fine. This unpredictability may adversely affect how a company may be able to operate if they use Oracle on an unsupported system. If you want to move your database from one system to another, if you have the support license for Oracle then they can support you in moving the database, while with MySQL you likely need to ask the community or pay for an additional service like Oracle for a similar experience.

As said by https://blog.panoply.io/mysql-vs-oracle:

"*MySQL: MySQL is an open-source relational database management system (RDBMS). Just like other relational databases, MySQL uses tables, constraints, triggers, roles, stored procedures and views as the core components that you work with. A table consists of rows, and each row contains data for each column. MySQL uses primary keys to uniquely identify each row (a.k.a record) in a table, and foreign keys to assure the referential integrity between two related tables.*

Oracle: Oracle is a multi-model database with a single, integrated back-end. This means *that it can support multiple data models like document, graph, relational, and key-value within the database. "*

The main difference between these two programs is that Oracle can do all the things that MySQL can do, but MySQL can't do all the things that Oracle can do (MySQL is a subset of Oracle). This may be another reason why a user may choose Oracle over MySQL. Oracle provides more features than MySQL, yet what they have in common they both do well. If a user wants to keep records in a Relational Database then MySQL would be a good choice as it has the same functions in these areas. However, if the user needs a more flexible solution which can be used for more than just this then Oracle would be the better choice as it has more functions that can be used to create a more complex and tailored system to make data manipulation easier.

In conclusion – the RDBMS chosen will reflect what the company or user needs from the software. A small-scale database that stores the sales figures of a small business may use MySQL as it is cheaper and has all the features necessary for the task. A large mortgage company may prefer Oracle as they can store supporting documentation within the database and have better support from Oracle themselves if anything goes wrong.

Question 2.

Sources of data

Data stored within Big Data is likely the most important aspect of Big Data, as this is the whole reason for having the data. The main difference between Big Data and a Relational Database is that data within a relational database must be accurate and clearly defined. Big Data has poorly defined data which can be inaccurate or missing. This allows the data to be manipulated to show trends or predict missing or future data. E.G. Name is ambiguous for a data field for a relational database without prior knowledge about the table it is in. For Big Data this is a good field as any data that is a name can be entered – be it a person, building or any other noun. This concept is known as Bad Data. Bad Data is purely data that has inaccuracies within it – this can include but is not limited to: Duplicate data, Missing data, Incorrect SPAG, etc. If you ask someone on the street for their name and get a response, this data may be incorrect. If the data is known to potentially hold erroneous data then it can be more easily dealt with. If this data was entered into a relational database, the data would theoretically be assumed to be correct. This would be a problem for analysis as the relational database structure doesn't account for data inaccuracies. Big Data must have enough volume so that the amount of Bad Data, or missing data becomes statistically insignificant. This allows a user to analyse the data with as small a margin for error as possible while accounting for known and a likely percentage of unknown data inaccuracies.

Big Data can use less reliable sources than an equivalent relational database system as it is equipped to handle the data inaccuracies that may come with these unreliable data sources. If you decide to buy data from a company to process the data for your own gain but don't want to pay for a reputable source, Big Data is able to handle the reduced integrity well. This can reduce the expenses of a company and allow the company to make a larger profit. This is just one of the benefits of using Big Data within a commercial environment. Some of the top sources of Big Data include: Media, the Cloud, Internet of Things, Relational Databases, and the Internet. Companies are willing to share the figures about public exposure to media, including box office figures for films, Download figures for music, viewer count of a tv show, etc. This data can be processed to assess the likelihood of new media being a success or a failure. Cloud storage is not always secure. This can result in companies being able to see what data is being stored within the cloud to target adverts based on your use-case. The Internet of Things is now becoming popular as companies can directly access data about you – when you use the kettle, what's in your fridge, etc – and target ads accordingly. Relational databases pass pre-existing databases to companies for a price or in the event of a breach. The Internet stores lots of data about internet usage through cookies and history. This can be processed to assess criminal activity, search and spend habits, etc.

Tools required

Due to the nature of Big Data. Specific tools are required to be able to access, store, and manipulate Big Data in a useful and meaningful way. Apache Hadoop is an open-source tool that is one of the most popular for accessing Big Data. According to https://hadoop.apache.org/ , Apache Hadoop: "*allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.*" This describes many changes to the normal data accessing model that are required to access Big Data. It is necessary to spread the load over a network of machines as the data can be too large to be able to open on a single machine. E.G If a set of Big Data is 2 PB in size and a machine has 32 GB of RAM, a query may be able to run on the first 32 GB well (if the system is well optimised) but will struggle through the rest. If a network of thousands of low powered machines could be responsible for 8GB of the data, then this query would process much faster than a single machine. Sharing this load also means that the service used can dynamically allocate machines to tasks based on the processing power required. This means that the company that owns the machines being used can charge less as a client doesn't need to pay for the downtime of a machine and the company can own less machines in total. This scenario would be beneficial to all parties – like accessing the internet through a dedicated line versus packet switching across open channels. This method of operation is shared by many other services worldwide.

Another company with major links to Big Data is Amazon. Amazon could close their online store and still be a major worldwide company as their data servers and services like AWS are used just as much as their store, but with far more money being invested. Amazon allow it's AWS customers to create, manage and use Data Lakes – where a Data Lake is a large collection of data in it's raw format (Object blobs, files, etc). This allows Amazon to process clients data for them and make a profit in the process. As many businesses that rely on Big Data don't have the facilities themselves to process or even store the data, this is a major benefit to companies like Amazon or Apache Hadoop as they are a necessity in the market and the companies that rely on them are at their mercy. However this is a growing market, and competition is driving the prices down to get a competitive edge over other companies. This is making the processing of Big Data more accessible and more profitable.