**9.1: Explain the impact that Big data, data warehousing and data mining have had on an organisation. (P1 – Task 3)**

Explain in your own words the meanings of the following phrases;

|  |
| --- |
| **Big Data - large sets of online data that can be analysed to see patterns, characteristics and trends about humans, to be used for predictions, detections of certain things like criminals and epidemics and improving traffic flow**  **Data Warehousing - storing data in large amounts from multiple analytical sources.**  **Data Mining - ways of gathering data from a large storage of it such as a warehouse.** |

|  |  |
| --- | --- |
| Big Data Characteristic | **Description of the characteristic.** |
| **Volume** | How much data a server can hold. |
|  |
| **Velocity** | How fast data is stored or processed |
|  |
| **Variety** | How diverse data is, in types of media and different forms of structure. |
|  |
| **Storage** | How much data is and can be stored. |
|  |
| **Processing** | How much data can be processed. |
|  |

|  |  |
| --- | --- |
| Key features of a data warehouse: | **Description of the feature.** |
| **Subject Orientation** | Data warehouses focus on past subjects, like for example, sales, revenue, and not on ongoing and current organization data. This enables it to be used for data analysis which is a key element of decision-making |
|  |
| **Data denormalization** | A way of improving the read speed of a database and making queries simpler by optimising it. The way of doing this is by adding redundant data to one or more tables. |
|  |
| **Non-Volatility** | ensures that your previous data is not lost as new data is updated |
|  |
| **Large amounts of historical data usage** | Data can be identified via a time stamp. This allows you to store a lot of historical data which can be used for predictions of what will happen in the future so you or your company can make the best decisions for the future based on these predictions. |
|  |
| **Use of Queries** | Queries are used to request for information in a data warehouse and can also be used to process data alongside tables, charts or graphs. |
|  |
| **Controlling of data load** | Timestamps on data can be used to identify when data was uploaded, using load and unload utilities such as adding rows to a table while retaining all current data and read it from a file. |
|  |

|  |  |
| --- | --- |
| Aspects of Data Mining: | **Description of this aspect of Data Mining.** |
| **Methods used to automatically search large stores of data to discover patterns and trends beyond simple analysis** | Traditional statistical methods require a great deal of user interaction in order to validate the correctness of a model. As a result, statistical methods can be difficult to automate. Moreover, statistical methods typically do not scale well to very large data sets. Statistical methods rely on testing hypotheses or finding correlations based on smaller, representative samples of a larger population.  Data mining methods are suitable for large data sets and can be more readily automated. In fact, data mining algorithms often require large data sets for the creation of quality models.  Different types of data mining techniques are recognising patterns in data, classifying different attributes of data into different categories, associating different variables with each other, detecting outliers, clustering similar groups of data together, regression to identify the likelihood of a certain variable, prediction by using previously collected data to predict likely future outcomes. |
|  |
| **Use of sophisticated mathematical algorithms to segment the data and predict future events** | Data mining algorithms are often sensitive to specific characteristics of the data: outliers, irrelevant columns, columns that vary together (such as age and date of birth), data coding, and data that you choose to include or exclude. |

Describe how Bury College could make use of developments in Big data, data warehousing and data mining:

|  |
| --- |
| Bury college can use these developments to make it easier for students and teachers alike to get the information they need from the internet and from the college servers. For example, if someone is looking for Science material, they could easily find it in the college shared drives. |