Big Data

Question 1 : The Term Big Data refers to?

Big Data refers to both structured and unstructured data. Structured data refers to information with a high degree of organization, such that inclusion in a relational database is seamless and readily searchable by simple, straightforward search engine algorithms or other search operations [1 BrightPlanet]. Whereas unstructured data is data with a lack of structure and therefore very hard to analyze and search. Unstructured data consists of ‘human data’ like emails and social media information [2 CU].

Typically 10% of big data is structured and the other 90% is unstructured [2. CU]

Question 2: What is considered the primary goal of looking at big data/large data sets?

Big data analytics is the process of examining large amounts of data of a variety of types big data to uncover hidden patterns, unknown correlations and other useful information. Such information can provide competitive advantages over rival organizations and result in business benefits, such as more effective marketing and increased revenue. [3]

Big data helps businesses make better decisions by examining the data they have collected and find repeatable business patterns.

Question 3: Big data as a service (BDaaS) takes advantage of the predictive analytics of an outside provider.

True, BDaaS is when one organisation hires an external company to analyse their data so they may gain a competitive advantage in their field.

BDaaS is intended to free up organisational resources by using the predictive analytics of an outside provider to manage and assess large data sets, rather than using in-house staff [4].

The Service may be software to help analyse the data or a team of data scientists.

Question 4 Which of the following bid data frameworks was developed by google? What is it?

MapReduce: Simplified Data Processing on Large Clusters.

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model.[5 Google]

MapReduce is the programming paradigm that allows for massive scalability across hundreds of servers.

MapReduce first takes a set of data and converts into another set of data where individual elements are broken down into tuples(Key/value pairs). It then take the output and then combines those data tuples into a smaller set of tuples. [6 IBM]

Question 5: Is it possible to gather information based on people's activities while maintaining privacy?

//\\//\\

Question 6: Collecting and storing big data alone isn't enough to produce real business value. Analytics is necessary to:

Extract valuable insights from the data.

Question 7: What is Data Warehousing? How does fit in with Big Data ?

Data warehousing is the storage of a large amount of information by a business. A business might use data warehousing for data mining and analysis so that they might find repeatable business patterns to gain a competitive advantage in their field.

This data must be stored in a manner so it can be easily managed and analysed, Ralph Kimball stated in his book that:

“A data warehouse is a copy of transaction data specifically structured for querying and reporting”

Data warehousing is similar to archiving, except the main purpose of the data is so that it can be analysed.

Question 8: What is a Data Set? What is an open Data Set? How are they beneficial ?

A Data set is a collection of data. Most commonly a data set corresponds to the contents of a single database table [7].

“A data set is a set of data that is collected for a specific purpose. There are many ways in which data can be collected—for example, as part of service delivery, one-off surveys, interviews, observations, and so on. In order to ensure that the meaning of data in the data set is clearly understood and data can be consistently collected and used, data are defined using metadata…” [8. A guide to data development (2007)]

## 

## 

## 

## Refrences

[1. <http://www.brightplanet.com/2012/06/structured-vs-unstructured-data/> ]

[2. <http://www.prezi.com/-k-qer5fnlnn/big-data-intro-talk/>] - coventry university material.

[3. <http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics>]

[4. <http://searchcio.techtarget.com/definition/big-data-as-a-service-bdaas>]

[5. <http://research.google.com/archive/mapreduce.html>]

[6. <http://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/>]

[7. <http://en.wikipedia.org/wiki/Data_set>]

[8. <http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=6442458038>]