# Worksheet 6 Big data Answers

**Task 1**

1. Explain what is meant by these three characteristics of Big Data:

 (a) Volume

 Generated and stored data that is too voluminous to be handled by one server.

 (b) Velocity

 Continuous streams of data being collected in real time, that may also require a response within milliseconds.

 (c) Variety

 Data that could be in the form of text, numbers, images or sounds. It can also create ‘missing pieces’.

2. Why can “Big Data” not be held in a traditional database structure?

The links between data items are too complex for traditional database relationships to represent.

3. Research examples of Big Data and Smart Cities and make notes about what you can find out about

Summarise your findings in a one-minute presentation.

 Smart Cities can use data from sensors to drive decisions about energy use, traffic management, health, crime, public safety and many other issues

 Each building might provide data about energy consumed and generated

 Each junction might report traffic flows for vehicles, cycles and pedestrians with details such as time, direction, speed and size.

 Some decisions may be immediate, eg controlling traffic lights to allow emergency vehicles a green light; others may be long-term such as designing a land use policy.

 This topic might also lead to discussion of the consequences of uses of computing.

 See <https://www.theguardian.com/sustainable-business/2014/nov/21/smart-city-sensors-big-data-internet>

 <http://www.computerworld.com/article/3039218/internet-of-things/smart-city-tech-shows-complex-connections-between-data-and-sensors.html>

City of Barcelona Deploys Big Data BI Solution to Improve Lives and Create a Smart-City Template: <https://youtu.be/E18vfMV0Yrc>

<http://eprints.soton.ac.uk/394830/1/IJIM_Big_Data_in_Smart_City_Final.pdf>

**Task 2**



4. How many degrees of separation are there between Roger and Angie?

Three: Angie’s son Paul is a friend of Lucy, who knows (even if the friendship is one sided) Roger. Note that there are multiple longer connections, including two via Gemma

5. Identify some of the nodes and edges of the above graph.

 Nodes are the Users, who have *names.* Edges are relationships which have both *type* and *direction.*

6. Discussion: can ‘colleague\_of’ work in only one direction? Can ‘employer\_of’ work in both directions? Are human relationships simple to define – or immutable?

‘Colleague\_of’ is a reciprocal or bi-directional relationship like ‘married\_to’; ‘employer\_of’ is one-way.

There are issues both with relationships - which can mutate, like ‘married\_to’ – and the semantics of description. “Petite amie” in French (en. *Girlfriend*) is not the same as its direct translation “little friend” in English. Politicans may use the term “my right honourable friend” while insulting a bitter enemy. In terms of big data analytics, fuzzy logic, artificial intelligence and machine learning combine with the functional programming paradigm to deal with diverse datasets of variable quality to deliver useful results.

 <http://www.businesscloudnews.com/2016/06/06/what-is-the-promise-of-big-data-computer-will-be-better-than-humans/>

 <https://en.wikipedia.org/wiki/Apache_Spark>

**Task 3**

7. Watch the video about Hadoop distributed programming compared to SQL, at <https://youtu.be/MfF750YVDxM> (6:13)

 You are a Big Data consultant. Prepare a one minute presentation to brief a client on which approach they should use, and why, for:

* managing the payroll for a group of UK academy schools.
* researching student attitudes to school uniform policies as expressed through social media.
* managing a traffic light control system in a Smart City, with the aim of both reducing congestion and giving priority to emergency service vehicles responding to incidents.

 The payroll application involves highly structured and low-velocity data where accuracy and confidentiality are important, so an SQL approach would be best.

 In the other two cases, although the volume might vary, the data is both unstructured and frequently changing so that a scalable, distributed approach makes sense, especially in traffic control where speed of processing could be critical.

 Bear in mind that Big Data is often characterised by lack of structure at least as much as by volume.

8. Research the PageRank algorithm at the heart of the Google story. Consider how the scale of web search activity has grown in your lifetime.

 Answer: In 1998, Brin and Page predicted that “it is likely that top search engines will handle hundreds of millions of queries per day by the year 2000”. By 2012, Google was reporting daily searches of 3.5 billion.

 The significance of scalable solutions is clear.

 <http://www.internetlivestats.com/one-second/#google-band>

 Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, *56*(18), 3825-3833. [http://ilpubs.stanford.edu:8090/361/1/1998-8.pdf](%20http%3A//ilpubs.stanford.edu%3A8090/361/1/1998-8.pdf)

 Further reading:

<http://pr.efactory.de/>

<http://www.cs.princeton.edu/~chazelle/courses/BIB/pagerank.htm>

<https://youtu.be/v7n7wZhHJj8>