



4.11 Big Data

1) One characteristic of a data set that might result in it being classified as Big Data is that it contains a variety of different forms of information.

a) Describe **two other** characteristics that might result in a data set being classified as Big Data.

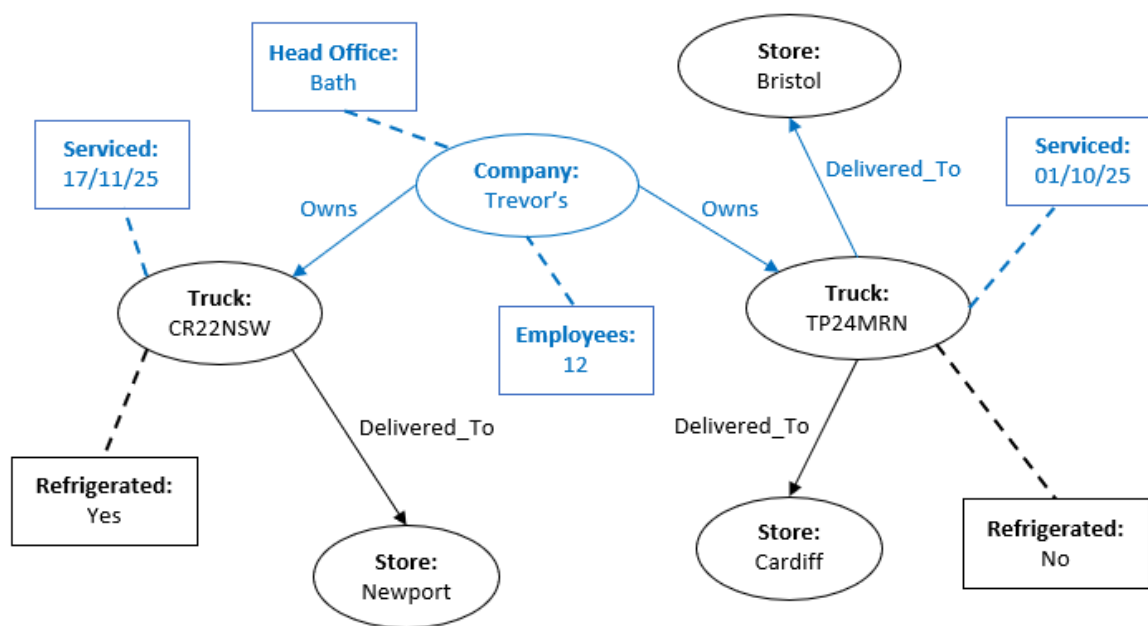
[2 marks]

- There is a high volume of data // data cannot be fit on one server
- The data is received / must be processed at high velocity

b) In a fact-based model, data is represented as atomic facts, which are immutable (i.e. will never change). Fact-based models can be represented visually using a graph schema.

Figure 1 shows part of a graph schema for a data set about deliveries made to stores by trucks.

Figure 1





Complete the graph schema in **Figure 1** to represent the following additional facts.

- Truck TP24MRN has made a delivery to the Bristol store.
- Truck CR22NSW was last serviced on 17 November 2025 and truck TP24MRN was last service on 01 October 2025.
- Both trucks are owned by a haulage company called Trevor's which has 12 employees and has a head office in Bath.

[3 marks]

Answer shown above on the graph schema



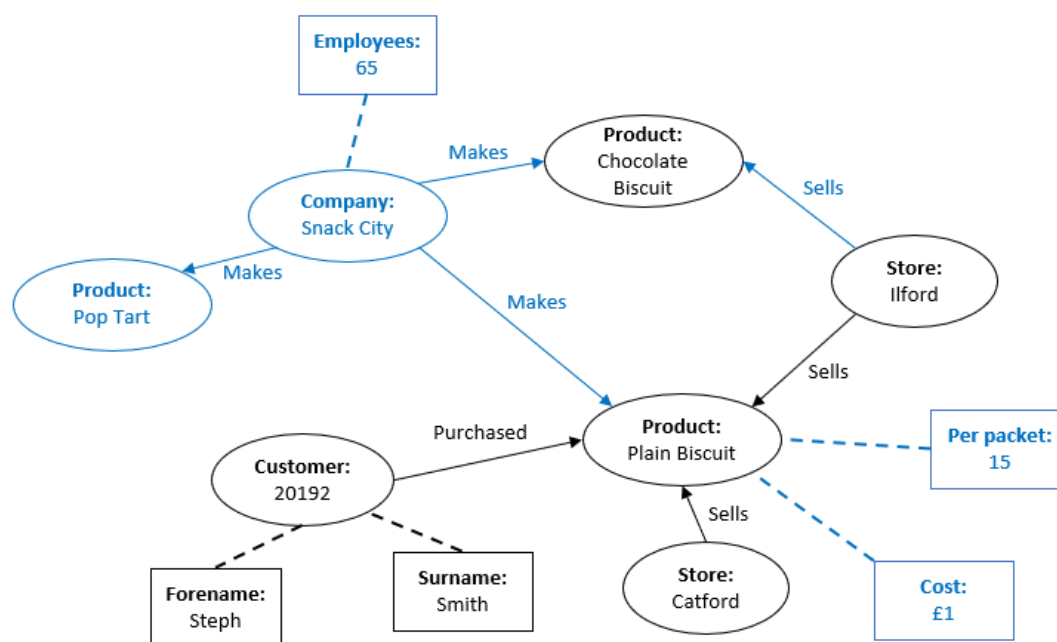


- 2) A supermarket chain uses a system which stores details of all of the products that it sells and the sales that it makes.

The data that the supermarket stores is Big Data.

- a) The graph schema in **Figure 2** represents part of a fact-based model of the dataset that the supermarket has built.

Figure 2



Modify the graph schema in **Figure 2** to represent the following additional facts.

- The Ilford store sells chocolate biscuits.
- There are 15 individual biscuits in a packet of plain biscuits and each packet costs £1.
- Both chocolate biscuits and plain biscuits are made by the company Snack City. The company has 65 employees and also makes Pop tarts.

[3 marks]

Answer shown above on the graph schema





- b) One approach to dealing with Big Data is to write code that can be distributed to run across more than one server.

State **two** features of functional programming languages that make it easier to write code that can be distributed to run across more than one server.

[2 marks]

- **Immutable data structures // the state of a data structure cannot be changed after creation**
 - **Statelessness // functions do not have side-effects // all functions are pure**
 - **Functions can be distributed to server and executed on data sets then the results can be combined // map-reduce**
 - **Higher-order functions can compose the results of processing on multiple processors/cores // functions are first-class objects**
 - **The order of execution can be determined at run-time // the order of execution can be determined by the translator // the order of execution is not defined by the program code // programs are not a sequence of instructions that must be followed in a specific order**
-





- 3) Big Data is increasingly used by organisations to store and analyse large amounts of information.

Discuss the characteristics of Big Data, the challenges involved in storing and processing it, and the techniques that can be used to overcome these challenges.

You should also consider ethical and legal issues associated with storing and processing personal data.

[12 marks]

What Big Data is

- Variety of different forms of information // data may lack structure
- E.g. Email messages, videos, images, website contexts, facial recognition
- There is a high volume of data // more data than will fit on one server
- E.g. Large medical datasets for diagnosis, gene sequencing, predicting disease outbreaks, results of large scale scientific experiments
- The data is generated/processed at high velocity
- E.g. Thousands of items to process per second , card payment fraud detection, recommendation systems

Challenges

- Data cannot be stored on one server
- Not possible to process data quickly enough with one computer
- Data cannot be represented in a table // by relational database
- Some forms of data are difficult to analyse

How overcome

- Distributed database systems across multiple servers
- Use of functional programming
- Massively parallelising the execution of programs
- Functional makes it easier to write distributable code // determine which parts of code can run independently
- Functional programming makes it easier to write correct code
- Use of many thousands of commodity servers
- Use of servers with multiple CPUs
- Machine learning can identify patterns // use of predictive models
- Use of languages such as XML or JSON to describe semi-structured data
- Using a fact-based model can manage bigger data sets better than a relational model.





Ethical and legal issues

- How can data be kept securely?
 - Who should have access to what data?
 - Will people know what data is being stored about them?
 - Where should all the data be stored // concerns about data stored in other countries
 - What rights do people have in relation to data stored about them
 - Example laws: Computer misuse act, GDPR, data protection act, Regulation of Investigatory powers act.
 - Who owns data about individuals?
-

