1) What’s the difference between structured and unstructured data? Can you give examples that you’ve encountered for both types?

Structured data comes in predictable and reliable formats, so for instance a date of birth will always be given as dd/mm/yy.

Unstructured data – often text / string -does not have this format imposed, often because it has been entered as free text.

When working with data from Rockbuster – I had data that was always structured, like release dates of films – and that which wasn’t like names of Actors.

2) Given that much of big data is produced by machines and sensors, how trustworthy do you think that big data is? What characteristic of big data relates to the question of trustworthiness?

In many ways, that data should be more reliable, it is not liable to human error when inputting that data.

However, it does not have a human eye on it when being produced, so if things haven’t been set up correctly, or there is some kind of malfunction along the way, a systemic error could be introduced.

Data always needs assessing and potentially cleaning before use. If say, a lot of variables are missing – then a decision needs to be made on how to use that data.

3) Assume that you receive a table containing customer data. You notice that some values are missing or incomplete, and the formatting is inconsistent in some columns. Based on what you’ve learned so far, how would you go about cleaning this table? Think about what you would do first, second, third, etc.

If some data is missing entirely, then we need to go back to source and see if the data can be gathered, or not.

If some if missing, then we need to see how much. There may not be enough to make the data useful and we may need to delete that column. There may be nought that we can simply delete the missing values, or simply note they are missing. We may be able to intuit what the values should be.

Provided we are happy the data is accurate, it i’s inconsistent, we should work on the data to bring consistency.

4) Can you describe tools such as Hadoop and Apache Spark and their role in big data? What do they do and how do they work?

Because big data is so big, it is hard for it be stored and then drawn on and processed in a useful timeframe. The difficulty of getting the flow of data can actually produce a bottleneck and slow down your work.

Be storing the data distributed across a wider network, Hadoop and Apache Spark allow you to draw on and process the data faster, meaning you are not held up in your work.

5) How has the application of analytics to big data led to new discoveries and innovation? Can you give some examples?

The application of analytics to big data has unlocked patterns, correlations, and insights previously hidden in massive datasets, fueling discoveries and innovation across industries. In healthcare, big data analytics accelerates drug discovery and enables personalized treatments by analyzing genetic and clinical records. Retailers like Amazon use predictive analytics to optimize supply chains and personalize recommendations, boosting customer satisfaction. In transportation, Uber and Lyft analyze traffic data to improve routing and reduce wait times. Financial institutions employ big data to detect fraud in real time. Even climate science benefits, with analytics enabling more accurate weather modeling and sustainability strategies.

And this answer was written by AI, using data analytics to analyse Large Language Models.