BIG DATA MANAGEMENT

CZ/CE4123

Tutorial – Big data 5V's

QUESTION1

Which of the following are related to big data applications, why?

(1) A computer scientist at MIT is trying to prove NP!=P (note: a famous computer science conjecture).

(2) A manager in Walmart wants to understand the purchasing behavior of customers through the purchase records in 2021.

(3) A programmer spends 10 hours in debugging his code.

(4) Artificial Intelligence can play games.

(1) A computer scientist at MIT is trying to prove NP!=P (note: a famous computer science conjecture).

(1) A computer scientist at MIT is trying to prove NP!=P (note: a famous computer science conjecture).

Ans: No. This is not that related to big data applications. It requires mathematical skills.

(2) A manager in Walmart wants to understand the purchasing behavior of customers through the purchase records in 2021.

(2) A manager in Walmart wants to understand the purchasing behavior of customers through the purchase records in 2021.

Ans: Yes. The purchase record of 2021 forms a large **volume** of data, from which we can discover a lot of **values** (e.g., co-purchase patterns) of users. The records are generated fast (**velocity**) due to Walmart's high popularity, and they are highly trustworthy (**veracity**) because they are from the Walmart system. The records may come from a **variety** of sources.

(3) A programmer spends 10 hours in debugging his code.

(3) A programmer spends 10 hours in debugging his code.

Ans: No. This is about a programmer using his own knowledge in debugging. Not related to (big) data.

(4) Artificial Intelligence can play games.

(4) Artificial Intelligence can play games.

Ans: Yes. Many of the AI players train themselves by machine learning (e.g., using neural networks) to become a top player. Typical machine learning models generate/require a lot of data (**volume**) for training. The training data may be frequently generated by other components (**velocity**). It will finally generate an AI player (**value**). The data generated by correct AI program should be highly trustworthy (**veracity**). The training data can come from different sources (**variety**).

E.g. https://www.youtube.com/watch?v=qv6UVOQ0F44

(4) Artificial Intelligence can play games.

Note: It is also acceptable to say that some AI is not related to big data. For example, some earlier AI uses rule-based algorithms.

QUESTION2

[Open Question]:
Is it possible that an application is not considered a big data application now, but can possibly become a big data application in the future?

Please also give some concrete examples during the discussion.

More and more applications/scenarios that were not considered to be data related previously become data-driven applications now. Examples include AI for game-playing. Early AI players were designed based on certain human-designed rules. Now, many AI players are based on big data to improve their playing strategies.

Therefore, for some of the above questions we gave an answer of “No”, they might become “Yes” in the future. For example, is it possible to make use of big data to help programmers debug their code? Let’s see how it goes in the future.

QUESTION3

Question 3:

For each of the following descriptions, which of the 5V’s is the most related?

QUESTION3

(1) There are more than 60000 searches per second in the Google search engine.

QUESTION3

(1) There are more than **60000 searches per second** in the Google search engine.

Ans: **60000 searches per second → velocity**

QUESTION3

(2) Suppose there is a database storing all kinds of enterprise related data, including the ratios of male/female in different companies. Ben wrote the following C program to collect the ratio of male/female for a company.

```
int cal_ratio(){
  int num_Male=getMales();
  int num_Female=getFemales();
  return num_Male/num_Female;
}
```

QUESTION3

(2) Suppose there is a database storing all kinds of enterprise related data, including the ratios of male/female in different companies. Ben wrote the following C program to collect the ratio of male/female for a company.

```
int cal_ratio(){
  int num_Male=getMales();
  int num_Female=getFemales();
  return num_Male/num_Female;
}
```

Suppose num_Male=1500, num_Female=1000, then the program returns “1” instead of “1.5”.

Ans: The code in red indicates that the ratio is not a fractional number. It is a bug in the algorithm, which outputs biased data. → Veracity

QUESTION3

(3) ImageNet is a 150GB dataset that holds 1,281,167 images for training and 50,000 images for validation, organized in 1,000 categories.

QUESTION3

(3) ImageNet is a 150GB dataset that holds 1,281,167 images for training and 50,000 images for validation, organized in 1,000 categories.

Ans: 150GB dataset that holds 1,281,167 images → Volume

(4) A website allows users to upload different forms of documents such as Excel, JPG, PDF, video.

(4) A website allows users to upload different forms of documents such as Excel, JPG, PDF, video.

Ans: Structured data (Excel) and Unstructured data(video, PDF, JPG)→Variety

(5) A researcher performs data mining algorithms on Amazon’s purchase records, and he successfully predicts the best seller in the next month.

(5) A researcher performs data mining algorithms on Amazon's purchase records, and he successfully predicts the best seller in the next month.

Ans: Data mining → Value