Big Data Lab01
Group members: Liang Haoxuan, Liu Sihan, Wang Changpeng

Basic Tasks

a/ Big Data means volume informations, it include a huge amount of informations. Big Data has 4 characteristics: Volume, Variety, Velocity and Veracity. For example, when different countries' users access Tiktok, people can receive a push of different types of contents. In a same country, different people can also receive a push of different content, which they like. So, Tiktok use big data, as volume information, to 'guess' what are the customers' favourite things, what they want to see on the Tiktok platform.

b/ With the big data analyse, people can get something useful. Such as the market purchasing power, Competitiveness of peers and the buying tendency of customers. The information produced by the big data analyse can help company to improve the products and change sals strategy etc, which can help business to increase their revenue and reduce risk.

c/ The simplest different between structured and unstructured data is that structured data can be represent by numbers or diagram but un structured data can't. And we can get information from structured data more easily and quickly. But the information of them are less than unstructured data.

d/ Smart Watch

Types	Blood	Sleep	Sport
Subdivision	pressure	time	number of steps
	oxygen content	quality	time
			heat

Sleep	time	quality
8.29	7h34m	GOOD
8.28	7h02m	GOOD
8.27	7h12m	GOOD
8.26	7h21m	GOOD

e/ Data-intensive system is the system which is to deal with a huge amount of data. Because there are billions of data have been created everyday, so without those system, our life will be terrible. Those system usually used to analyze the big data, machine learning or data processing. The Google Big Query, Amazon Redshift, and Hadoop are commen Data-intensivee systems.

f/ Data storage: distributed file system,such as Hadoop Distributed File System, can provide hide throughput data access, is suitable for big data storage and processing.

Data visulisation and analyse: e.g Tableau, allow customers to bilud a interactive charts and panel to visualize and explore big data set.

Compute and distribute: massively parallel processing framework allows parallel processing and big data set analysis, ist suitable for all kinds of computing tasks.

Data warehouse: e.g NoSQL, NoSQL databases offer flexible data models and are suitable for handling unstructured and semi-structured data, providing high scalability and performance in data warehousing scenarios, enable organizations to manage and analyze vast amounts of data effectively for various applications and insights.

Medium Tasks

a/ Data has been coined "The oil of the 21st century", WHY? We can talk something about the value of oil to the world. With oil, we can get a variety of chemical products, which have great economic value and promote the development of the world. A classic example, engine, the engine of a car or a plane can't work without oil. Without engine, our technology will go back a lot. So these are the importance of oil to the world. Same, data has a equal significance to 21st century as oil. 21st century is a data era. Data is produced, transmitted and processed in large quantities in society. No matter is any industry, data will participate in it, the data carries the information in the industry, transmits the information. Without data, the development of the world will slow down a lot or even reverse. Some of the data also has high commercial value, such as the manufacturing data of various chips. In the 21st century, data acts as the carrier of information. The development of society, technology and science can not be separated from data.

b/ 1/ Because big data is volume, so the data of big data are acquired by people is impossible, and need to use some program or machine to acquire data. There is a problem, something are not comform to the requirements, but the program also put these into the big data. They can not like people to distinguish data. So it will bring inconsstency， incompleteness and so on to big data.

2/ The veracity of data also stand for accuracy, believability, reputation, obiectivity, factuality, consistency, correctness, and unambiguousness. The data should be accurate, otherwise it is dummy data, and than the people can believe in that data. And the data must be objective, there are no emotion in the data at all.

Advanced Tasks

a/ APISCRAPY (2023). AI & ML Training Data | Artificial Intelligence (AI) | Machine Learning (ML) Datasets | Deep Learning Datasets | Easy to Integrate | Free Sample [Dataset]. https://datarade.ai/data-products/ai-ml-training-data-ai-learning-dataset-ml-learning-dataset-apiscrapy

https://datasetsearch.research.google.com/search?src=3&query=machine%20learning&docid=L2cvMTF2ejhtMHgzeQ%3D%3D

b/ This data set is about ai learning and has more than 50M varieties of records from 61 countries. The data from varieties of area, include healthcare and banking, as well as e-commerce and natural language processing. This data set has a large amount of data and ist complex enough.