BDSD 25-26-1 Final Exam

Comment: Terrible, meaningless, and useless.

1. Single Choice (20 pts)

选择题记不太清楚了，大致涉及的知识如下：

1. Which of the following is **not** a characteristic of Spark?

2. What is KDD?（我还以为看错了，以为是 RDD，结果是 KDD。你的 PPT 里有 KDD 吗？）

3. （忘记了）

4. Which of the following is **not** unsupervised learning?

5. Which of the following is **not** a key-value database? (Options: Redis, Berkeley DB, etc.)

6. （忘记了）

7. Which is **not** a characteristic of Storm?

   - One of the options: *The process unit is "process"*.

8. （忘记了）

9. Which of the following statements is **not** a characteristic of the Deep Web? (All options came from the PPT description.)

10. What is heterogeneity in big data?

    - Options: Source, structure, etc.

> 很多题都涉及"特点/特征"，考点都很出乎意料。比如KDD 根本没有在PPT 中出现；又如考了 key-value 数据库的具体例子等。有的题老师上课时甚至说过"not important, skip"，但还是出现在了卷子上。

2. Fill-in-the-Blank (20 pts)

1. In Spark, RDD has two types of operations: ___ and ___.

   - The first operation has two types, which are ___ and ___.

2. Hadoop architecture has two main components: ___ and ___.

3. In HBase, the index consists of: __, ___, ___, and ___.

3. Short Answer (60 pts)

1. List the processing models and explain their characteristics. (10 pts)

2. Give the difference between **DML** and **federated learning**. (10 pts)

3. Apply 2-anonymization to the given table. (10 pts)

Original Table

DATE OF BIRTH	GENDER	ZIP CODE	SALARY
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

Table 1

2-Anonymized Table

DATE OF BIRTH	GENDER	ZIP CODE	SALARY
1/21/76	M	537**	50,000
4/13/86	F	537**	55,000
2/28/76	*	537**	60,000
1/21/76	M	537**	65,000
4/13/86	F	537**	70,000
2/28/76	*	537**	75,000

Table 2

(The original problem appeared in PPT-7.)

4. What are the key technologies in big data processing? Use a project as an example to explain. (30 pts)