上海科技大学
ShanghaiTech University

Sanitizing Clinical Data Using Big Data Frameworks

Presenter：姜虹竹

Date：2025/05/16

上海科技大学
ShanghaiTech University

Content

上海科技大学
ShanghaiTech University

01 Introduction

Sanitizing Clinical Data Using Big Data Framework

Problem:

- Clinical datasets contain sensitive information (e.g., age, gender, ZIP, diagnoses).
- Traditional anonymization methods like k-anonymity, l-diversity and t-closeness are hard to scale.

Motivation:

- Regulatory pressure (HIPAA, GDPR) demands both privacy and data utility.
- Need for scalable anonymization pipelines using tools like Apache Spark.

k-anonymity

K-anonymity demands that we group individual rows/persons of our dataset into group of at least ***k*** rows/persons.

l-diversity

L-diversity ensures that each k-anonymous group contains at least l different values of the sensitive attribute.

t-closeness

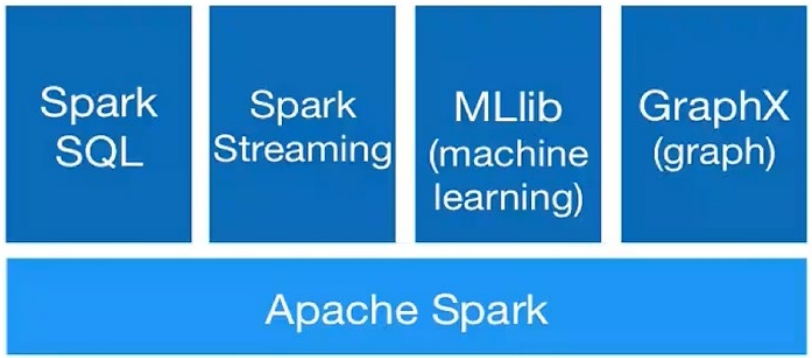
T-closeness demands that the statistical distribution of the sensitive attribute values in each k-anonymous group is "close" to the overall distribution of that attribute in the entire dataset.

This project aims at the application of big data frameworks, particularly Apache Spark, to implement these techniques efficiently on clinical datasets.

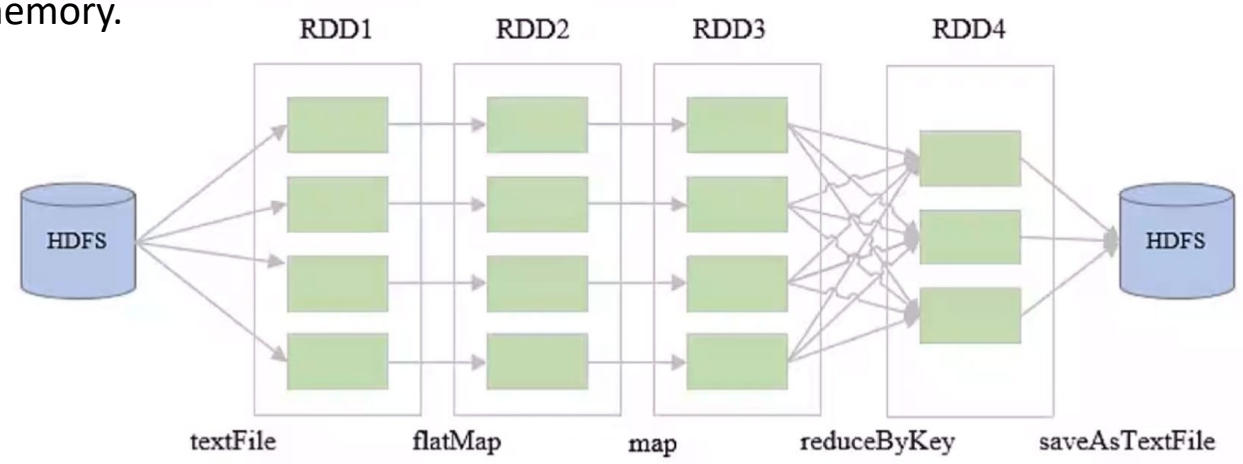
Why Apache Spark?

- The AMP Lab at the University of California, Berkeley, open-sourced Apache Spark.
- High throughput, low latency, general-purpose extensibility, and high fault tolerance.

- A general-purpose large-scale distributed computing engine.



- Spark performs computations based on RDD (Resilient Distributed Datasets).
- During data recovery, it only needs to load the RDDs from the previous stage in memory.



上海科技大学
ShanghaiTech University

02
Pilot Study

Sanitizing Clinical Data Using Big Data Framework

Dataset:

- "patients.csv" from SyntheticMass (116 synthetic EHR records)

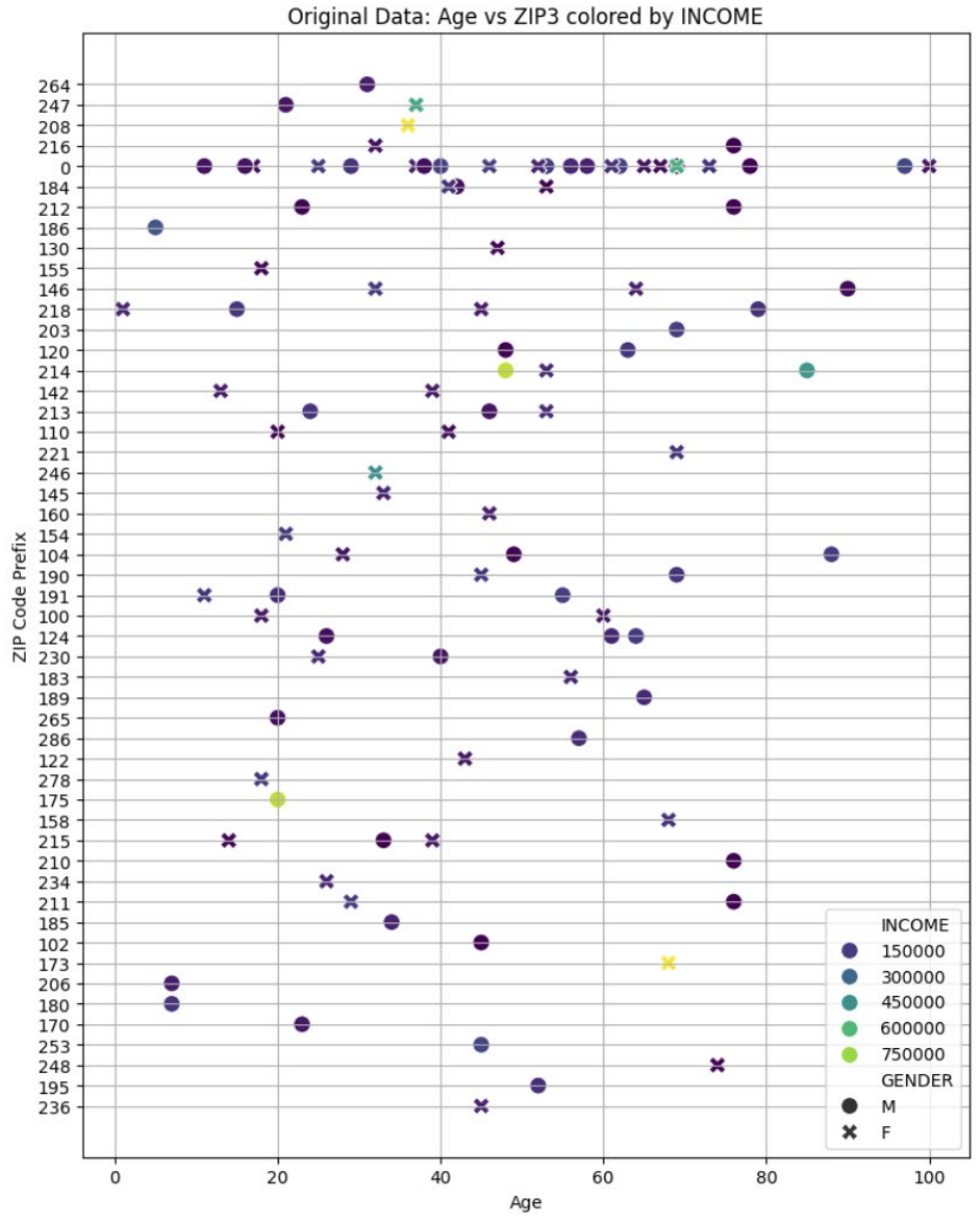
ID	BIRTHDATE	RACE	GENDER	COUNTY	ZIP	INCOME
34a210f9-	1993/11/11	white	M	Barnstab	02644	71625
0fe22cec-	1987/8/2	white	F	Middlese	02472	542941
b1f7b5a9-	1989/3/5	white	F	Norfolk (	02081	886745
1f2aa6c9-	1993/2/2	white	F	Norfolk (	02169	35850
80e114d1-	1999/10/27	white	F	Hampshir	00000	149942

Outcome:

- Successful proof-of-concept on a small dataset

Summary of groups (k-anonymity and l-diversity):

	AGE	GENDER	ZIP3	total_count	l_diversity
0	13-17	F,M	0,142,215,218	6	6
1	18-25	F,M	0,154,155,191,212,213,247	7	7
2	18-25	F,M	100,110,170,175,230,265,278	8	7
3	1-11	F,M	0,180,186,191,206,218	6	6
4	26-34	F,M	104,124,185,211,215,234	6	6
5	29-36	F,M	0,145,146,208,216,246,264	7	7
6	37-45	F,M	0,110,142,184,190,247	9	9
7	39-45	F,M	102,122,215,218,230,236,253	7	7
8	46-53	F,M	0,120,130,213,214	9	8
9	46-55	F,M	104,160,184,191,195	5	5
10	56-61	F,M	0,100,124,183,286	8	7
11	62-67	F,M	0,120,124,146,189	8	8
12	68-73	F,M	0,158,190,203,221	8	7
13	68-76	F,M	173,210,211,212,216,248	7	3
14	78-90	M	0,104,146,214,218	7	5
15	97-100	F,M	0	8	2



上海科技大学
ShanghaiTech University

03
Extended Project & Timeline

Sanitizing Clinical Data Using Big Data Framework

The roadmap for experiments:

- Distributed anonymization (k-anonymity, l-diversity and t-clossness; dataset: MIMIC-III)
- Analyzing re-identification risk quantification using game-theoretic framwork
- Evaluation and performance optimization

Potential Risks and Mitigation Strategies:

Risk	Description	Mitigation
NP-hard problem	Find the optimal partition into k-anonymous groups is an NP-hard problem.	Try "Mondrian" algorithm or ARX anonymization tools.
Data access	Access to MIMIC-III may be delayed due to approval processes or restrictions.	Use synthetic datasets (e.g., Synthea, MIMIC-demo) for development and testing.

Timeline:

- 2025/4/25-2025/5/9: Topic selection
- 2025/5/10-2025/5/16: Design and complete the pilot study
- 2025/5/17-2025/5/31: Run experiments and optimize Spark
- 2025/6/1-2025/6/13: Final report and presentation

上海科技大学
ShanghaiTech University

Thank You

Presenter：姜虹竹