上海科技大学
ShanghaiTech University

Sanitizing Clinical Data Using Big Data Frameworks

Presenter：姜虹竹

Date：2025/06/06

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, race, income).
- 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

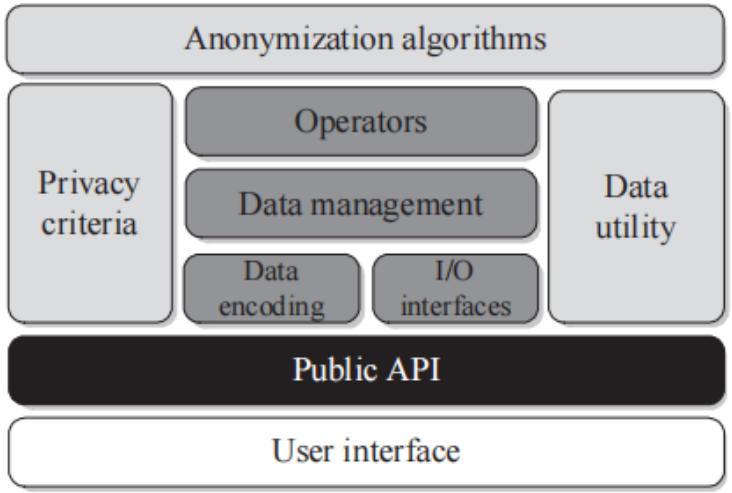
T-closeness demands that the statistical distribution of the sensitive attribute values in each k-anonymous group is "close" to the overall distribution of that attribute in the entire dataset.

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

Commonly used frameworks: ARX Data Anonymization Tool

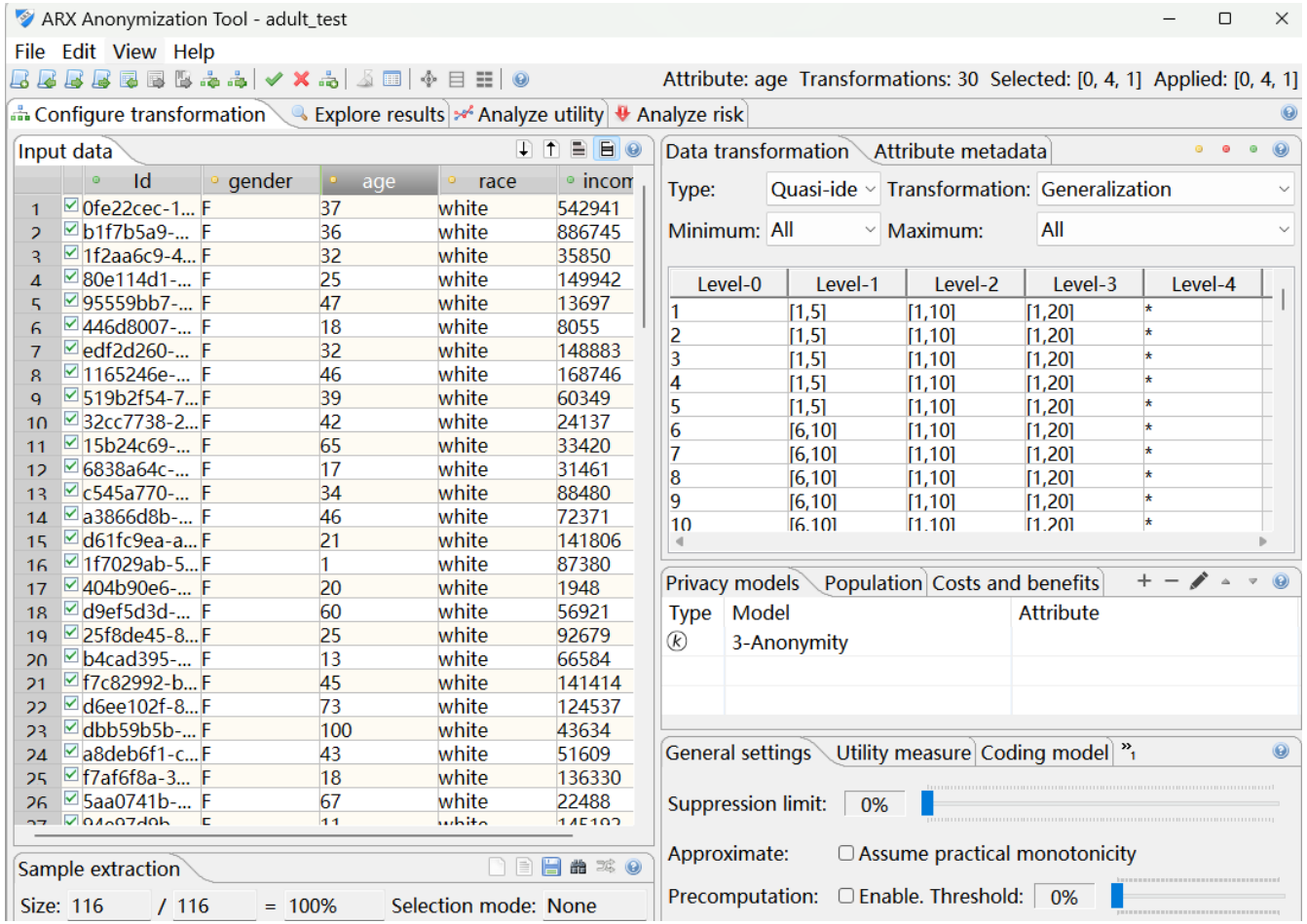
- ARX: a comprehensive open-source data anonymization framework.
  - A user-friendly, cross-platform graphical interface that simplifies the anonymization process
  - A rich library of privacy models and transformation methods
  - Active community support, frequent updates, and transparent development

- A high-level overview of the architecture of ARX



ARX Graphical User Interface

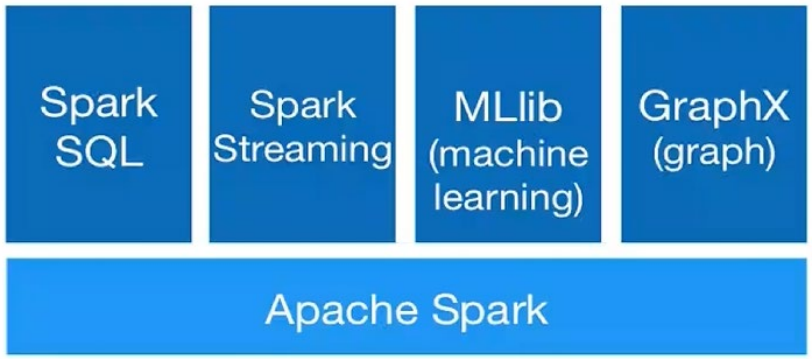
- Import dataset → Define Attribute Types → Import hierarchies → Configure Privacy →ModelExport Results



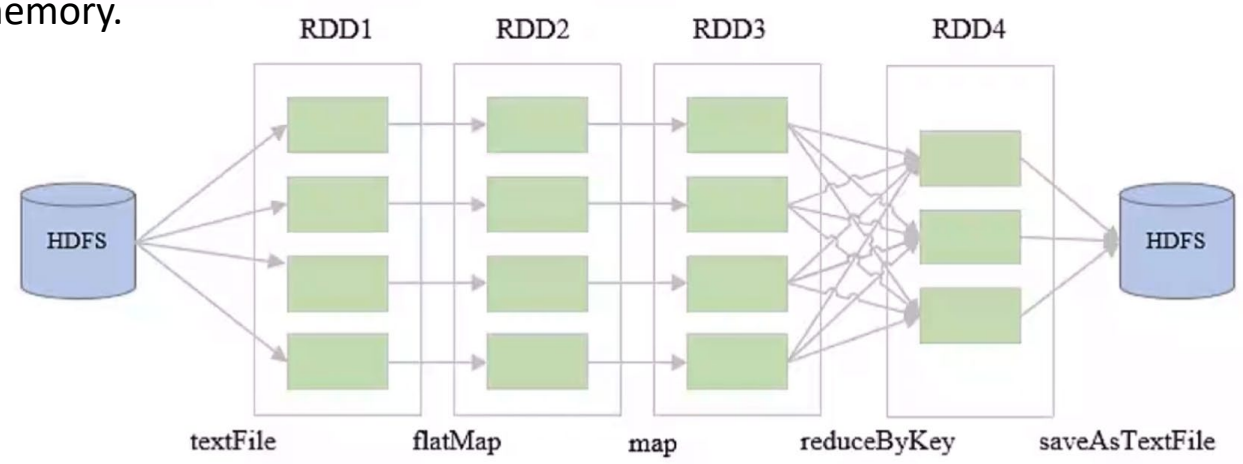
Why Apache Spark?

- The AMP Lab at the University of California, Berkeley, open-sourced Apache Spark.
- A relatively new distributed computing platform developed to overcome some shortcomings of Hadoop
- 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 Methods

Sanitizing Clinical Data Using Big Data Framework

Dataset:

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

Id	gender	age	race	income
34a210f9-5ce1-ad63-790:	M	31	white	71625
0fe22cec-1a19-99da-b67:	F	37	white	542941
b1f7b5a9-5cf5-6050-b23:	F	36	white	886745
1f2aa6c9-41bd-aa05-9da:	F	32	white	35850
80e114d1-013b-b546-7ab:	F	25	white	149942

  - Quasi_identifier: {gender, age, race}
  - Sensitive information: income

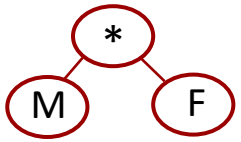
- "adults.csv" from UC Irvine Machine Learning Repository (32561 records)

age	workclass	education	marital_status	occupation	relationship	race	sex	income
39	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	<=50K
50	Self-emp-n	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	<=50K
38	Private	HS-grad	Divorced	Handlers-clear	Not-in-family	White	Male	<=50K
53	Private	11th	Married-civ-spouse	Handlers-clear	Husband	Black	Male	<=50K
28	Private	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female	<=50K

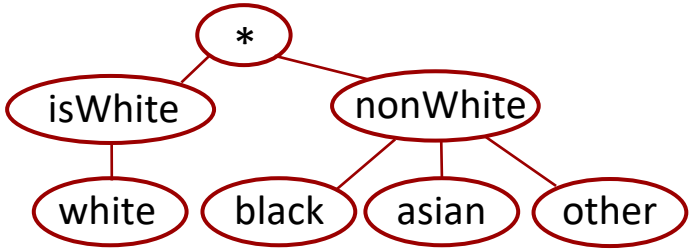
- Quasi_identifier: {education, marital_status, occupation, native_country, workclass, relationship, race, sex}
- Sensitive information: income

Examples of hierarchy generalization trees:

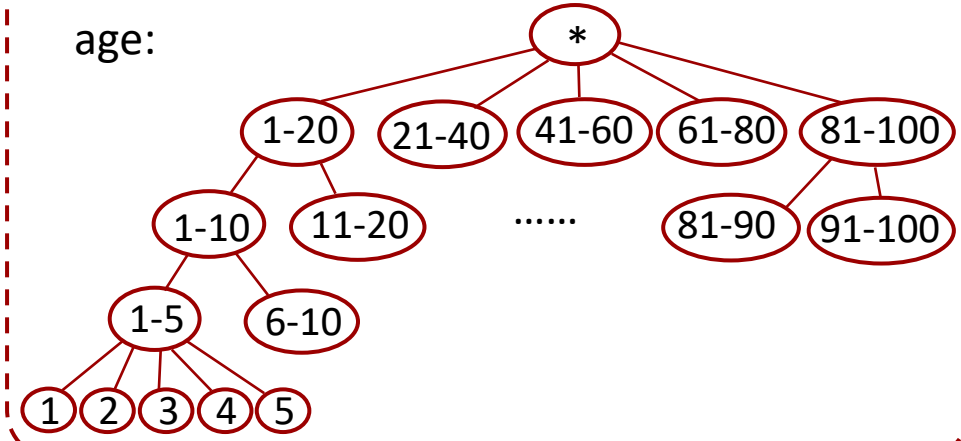
gender:



race:



age:



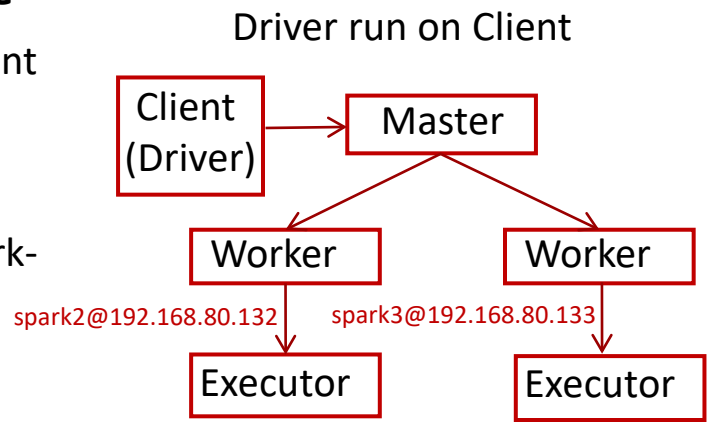
Spark Environment Configuration:

Cluster Setup

- Cluster mode: Standalone
- Master URL: spark://jhz:7077
- Number of worker nodes: 2
- Spark version: 4.0.0
- Scala version: 2.13.0
- Java version: 17.0.12

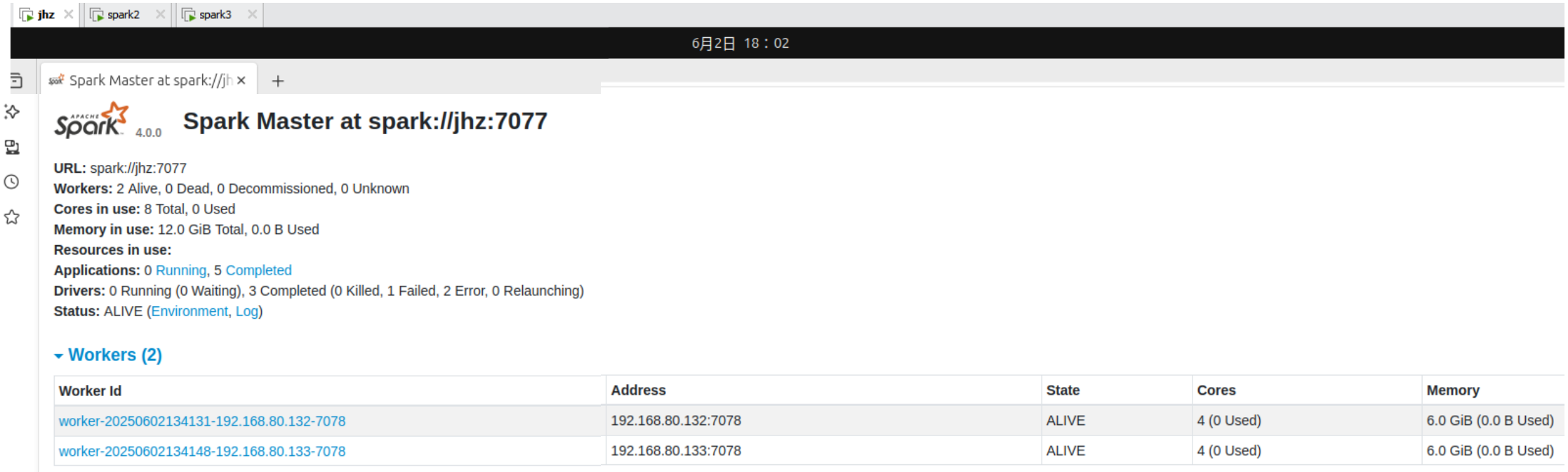
Submission Mode

- Deploy mode: client
- Driver location: Runs on the node that executes spark-submit（e.g. Master node）

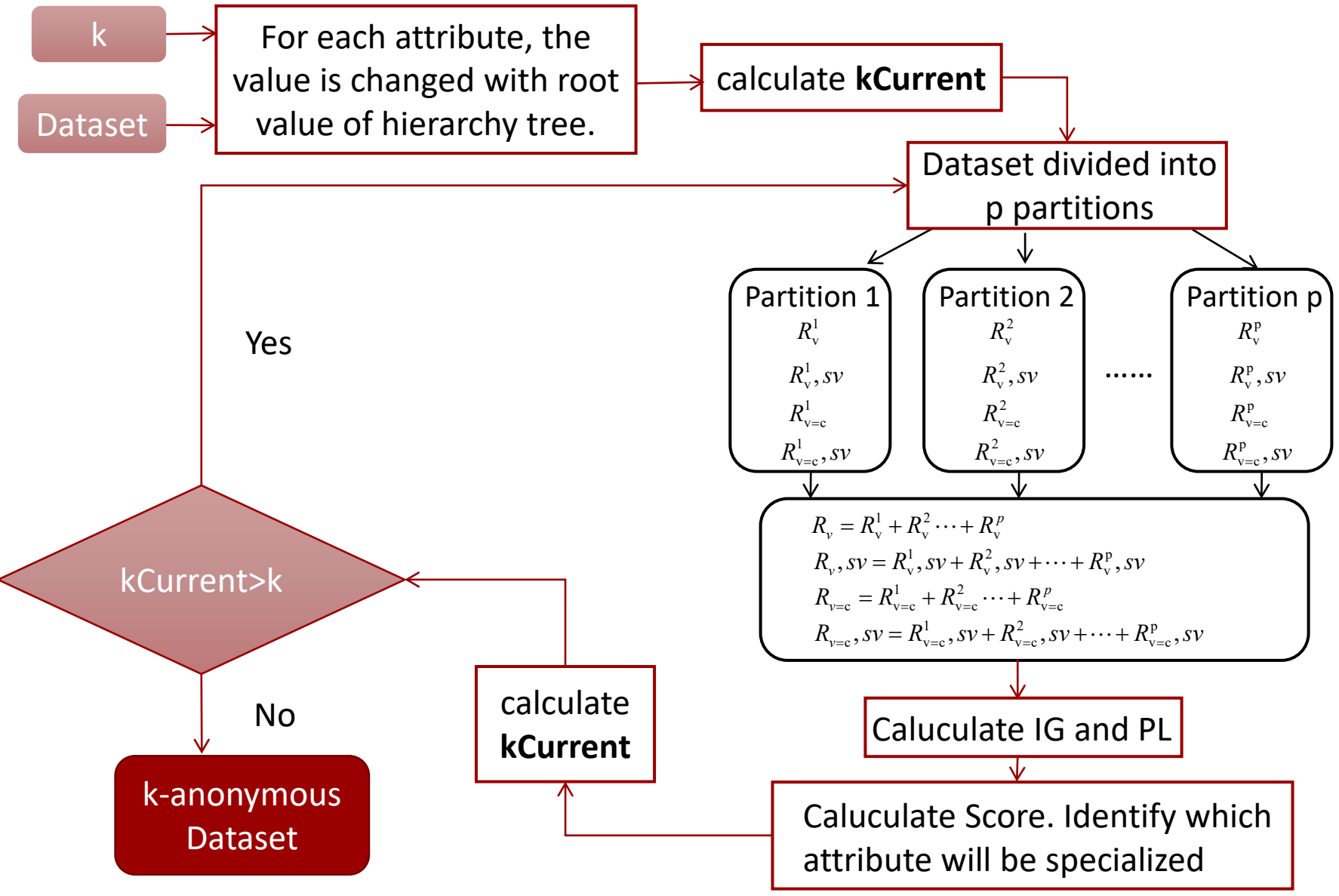


Resource Allocation

Parameter
--executor-memory 2G
--executor-cores 2
--total-executor-cores 4
--driver-memory 2G



Top-down Specialization(TDS) Algorithm



Notation used in the algorithm:

Notation	Description
$kCurrent$	the calculated k value at each iteration of anonymization
p	number of partition
Rv	set of records containing attribute values which can be generalized to v
Rv, sv	set of data records which contains sensitive values sv in Rv
IG	Information Gain
PL	Privacy/anonymity Loss
$IGPL$	Information Gain Privacy Loss

Two metrics:

(1) $AnonyLoss(v) = A(v) - A'(v)$

(2) $InfoGain(v) = I(R_v) - \sum_{c \in children(v)} \frac{|R_{v=c}|}{|R_v|} I(R_{v=c})$

$$Score(v) = \begin{cases} \frac{InfoGain(v)}{AnonyLoss(v)}, & \text{if } AnonyLoss(v) \neq 0 \\ InfoGain(v) & \text{otherwise} \end{cases}$$

上海科技大学
ShanghaiTech University

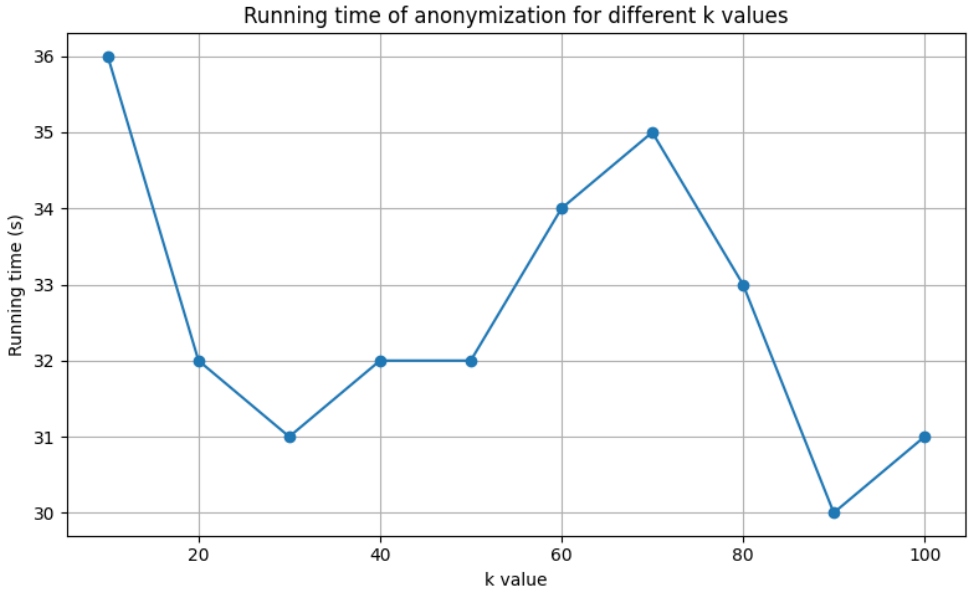
03
Results

Sanitizing Clinical Data Using Big Data Framework

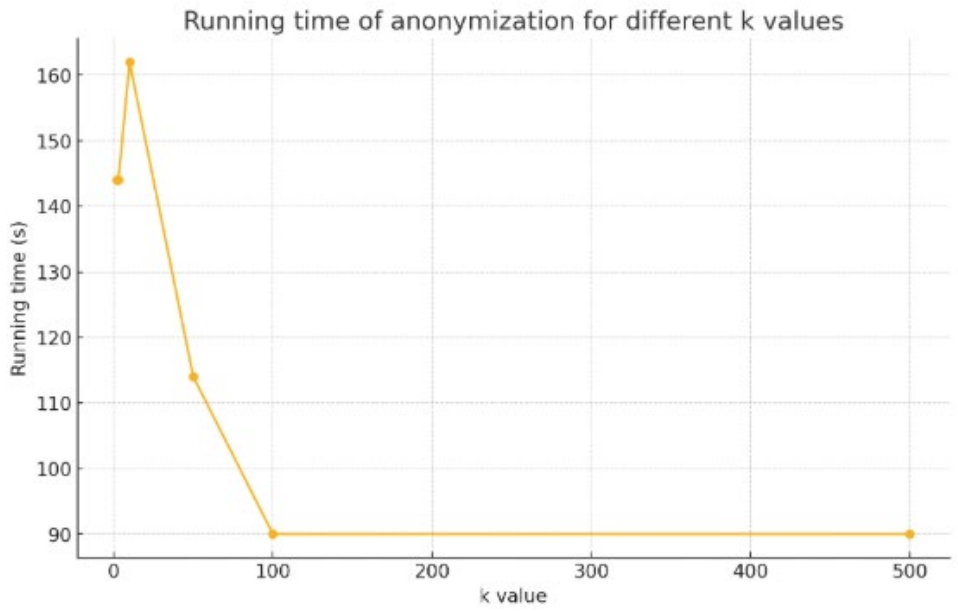
Efficiency results

- Two datasets: one with 116 rows, the other with 32,561 rows.
- Smaller k values lead to longer running times for both datasets.
- Larger datasets (32,561 rows) show more pronounced variations in running time as k。
- Larger k values generally result in reduced processing time, especially in larger datasets.

Dataset 1 (116 records)



Dataset 2 (32561 records)



Data Utility Comparative Analysis: Spark-based K-Anonymity vs. ARX Tool

TDS results

- Generalization level: {gender, race, age} = {1,2,3}

gender_general	race_generalized	age_generalized
*	*	61-80
*	*	21-40
*	*	21-40
*	*	21-40
*	*	61-80

- The k-anonymous dataset (k=3) generated by the TDS algorithm meets the 3-diversity requirement.

Attribute	Generalization intensity
gender	100%
age	80%
race	100%

ARX results

- Generalization level: {gender, race, age} = {0,4,1}

gender	age	race
F	*	nonWhite
F	*	isWhite
F	*	nonWhite
F	*	isWhite
F	*	isWhite

- also meets the 3-diversity requirement

Additional quality metrics:

Output data | Classification performance | Quality models

Attribute-level quality

Attribute	Data type	Missings	Gen. intensity	Granularity	N.-U. entropy
gender	String	0%	100%	100%	100%
age	String	100%	0%	0%	0%
race	String	0%	50%	87.93103%	75.52098%

上海科技大学
ShanghaiTech University

Thanks for listening

Presenter：姜虹竹