**Приветствие (1)**

Greetings ladies and gentlemen, today we'd like to talk about Privacy preserving distributed data mining based on secure multi-party computation.

**План (2)**

**Задачи (3)**

* To introduce what data mining is
* To define two problems to be solved
* Explain each task in detail
* Provide algorithms 1,2,3 as solutions

**Введение (4)**  
*In recent years, we have witnessed an unprecedented big data explosion than ever before, especially for private data in business environments. For most companies and organizations that own private big data, such as electronic commerce, online-banking and cross-field research data, the data infrastructure is built on a nationwide or worldwide distributed system.*

**Цель (5)**  
In this section, we will first introduce two real-world tasks of privacy preserving distributed mining. Then, we will illustrate the traditional way to solve these tasks by secret sharing, and point out its potential risk of privacy leakage. Finally, we will describe our solutions for these two tasks.

Two problems to be solved in this paper by introducing two real-world privacy preserving distributed mining tasks:  
  
Task 1: Calculating Statistics of Different Types of Data  
  
Task 2: Studying the Relationship between Variables by Linear Regression

**Задачи 1,2 (6-7)**

**Task 1**

* Here, we first define the problem of calculating statistics of different types of data under privacy preserving distributed data mining condition as:
* Given data in the form of 〈*type*, *value*〉 tuples, the problem is to calculate the statistical indicators of given data, such as mean, variance, maximum and minimum values, without privacy leak risk for both *type* and *value* in original data.

**Task 2**

* In general, we can use the linear regression to reveal the relationship between variables. The same concern of business privacy leakage exists if these companies are asked to disclose their trading records to perform linear regression analysis. Hence, we define the problem of performing linear regression under privacy preserving distributed data mining condition as:
* Suppose we have a set of tuples formed as

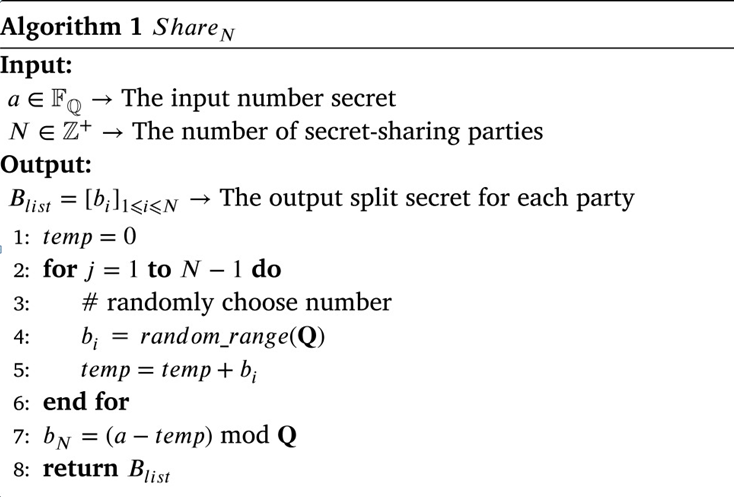
in which <dep\_var>  and  <indep\_var>  refer to dependent variable and independent variable, respectively. All independent variables in a row compose a matrix ***A***. All dependent variables compose a column vector ***b***. The problem is to find an optimal ***x*** for minimizing the error of  ***Ax= b***, without privacy leakage risk for both ***A*** and ***b*** in the original data.

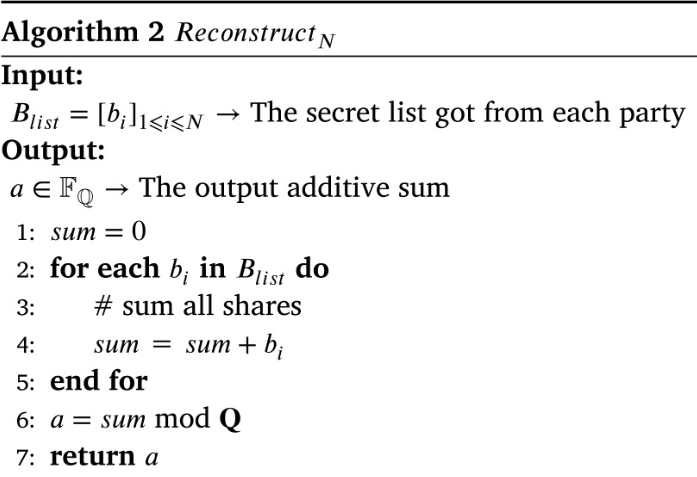
**Термины (8)**MPC - multi-party computation   
LU - lower–upper (LU) decomposition or factorization factors a matrix as the product of a lower triangular matrix and an upper triangular matrix (see matrix decomposition).

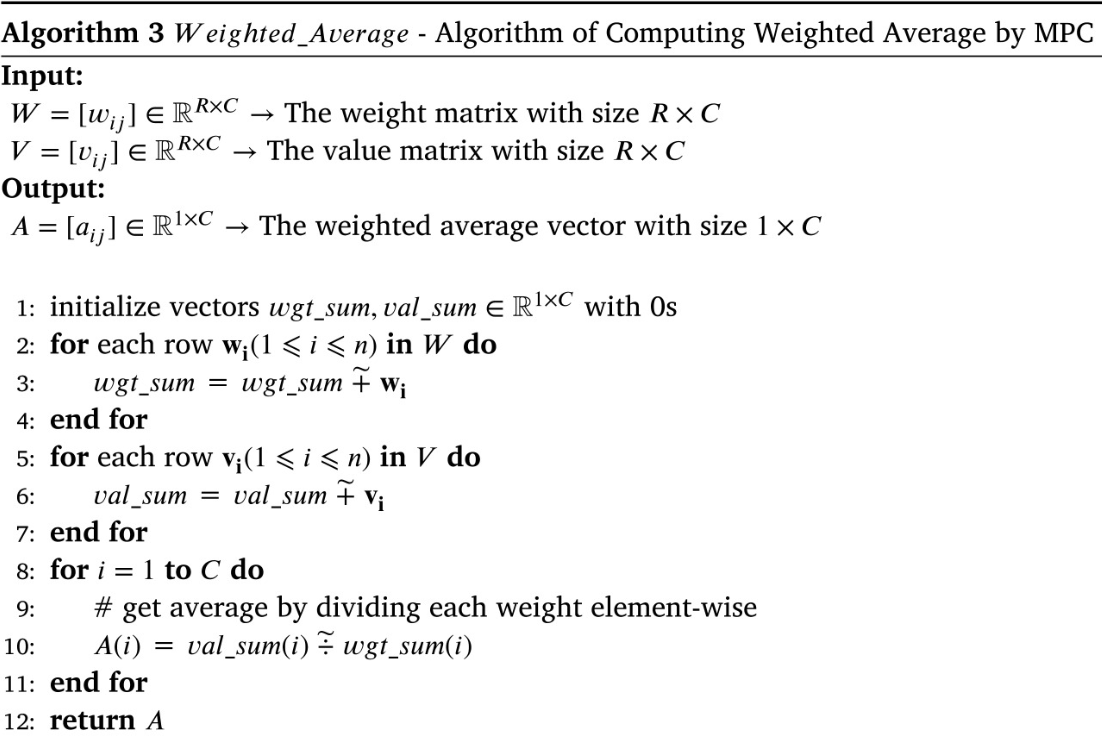
QR - a decomposition of a matrix A into a product A = QR of an orthonormal matrix Q and an upper triangular matrix R.

Data mining - process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

**Алгоритмы 1,2,3 (9-10)**







**Algorithm 3** – A simple yet useful algorithm to calculate the weighted average value of data in different types

**Решения (11)**

**Task 1**

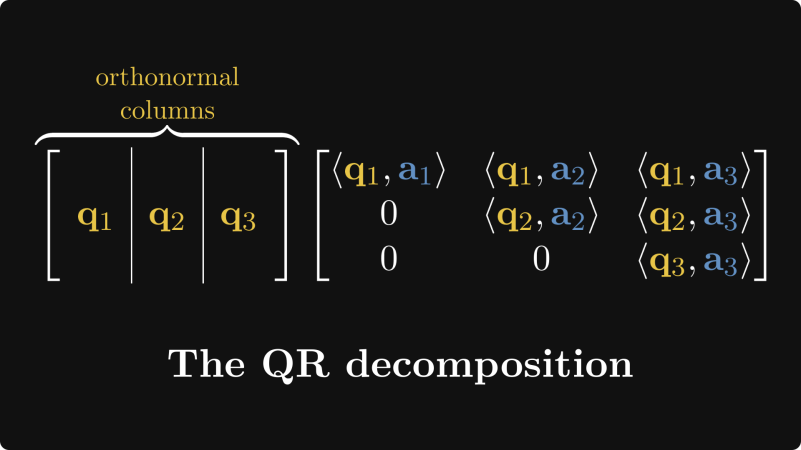
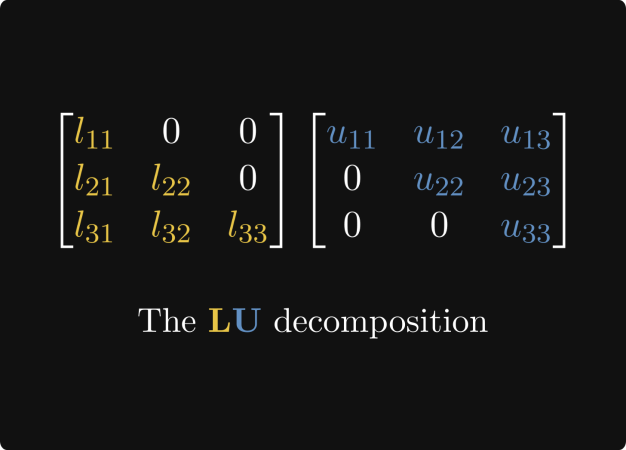
We consider to mix up different types of data to avoid revealing the amount of cotton of a certain types. In this way, we can compute statistical indicators by selecting data of each type rather than transfering data of a certain type between involved parties. The key point to achieve this goal is the capability of distinguishing different types in a mixed dataset.

**Task 2**

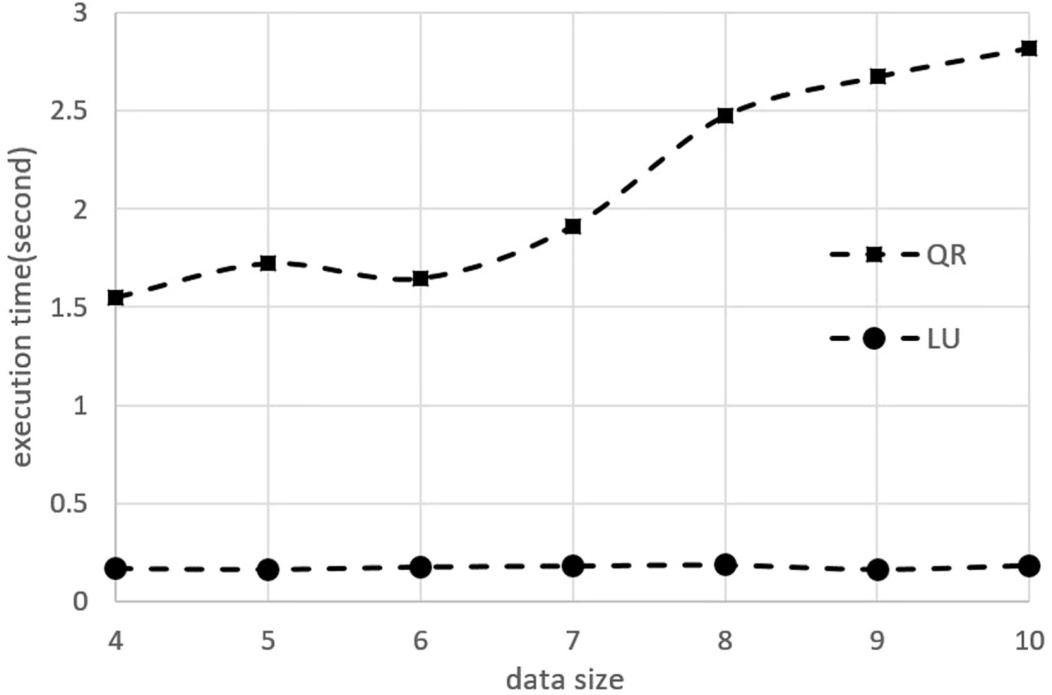
To accomplish task 2 we use least square method to solve linear regression.   
We choose QR and LU decomposition methods as candidates for matrix decomposition.

QR decomposition - Gram–Schmidt

LU - the product of a lower triangular matrix and upper triangular matrix

**Результаты (12)**



* To compare the performance of our implemented QR and LU decomposition for MPC, we generate seven datasets. In each dataset, every participant has a matrix with 4 columns and N rows. N varies from 4 to 10 for these seven datasets.

**Вывод (13)**

Privacy preserving distributed data mining is an important task for big data.. In this paper, we have proposed to utilize the MPC and SPDZ protocol to perform this task.

To accomplish these data mining tasks, we have proposed two solutions based on matrix computation with one-hot encoding and LU decomposition.

**Прощание (14)**

Thank you for your attention