Secure JOINing Outsourced Databases by Secret Sharing

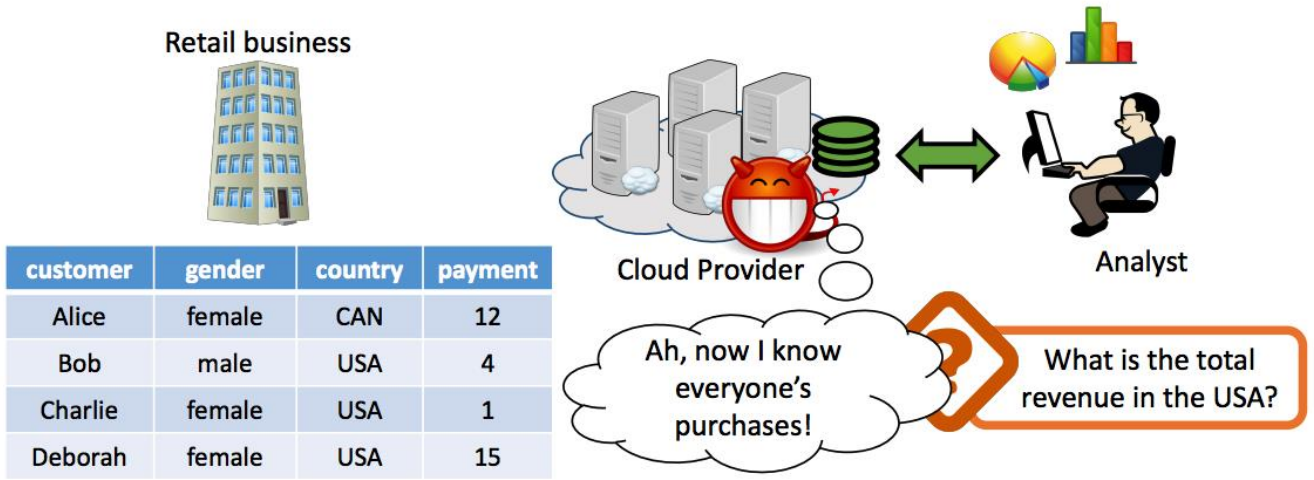
Shangyu Xie

CS595 Big Data Analytics

ILLINOIS TECH

Some Recap-Background

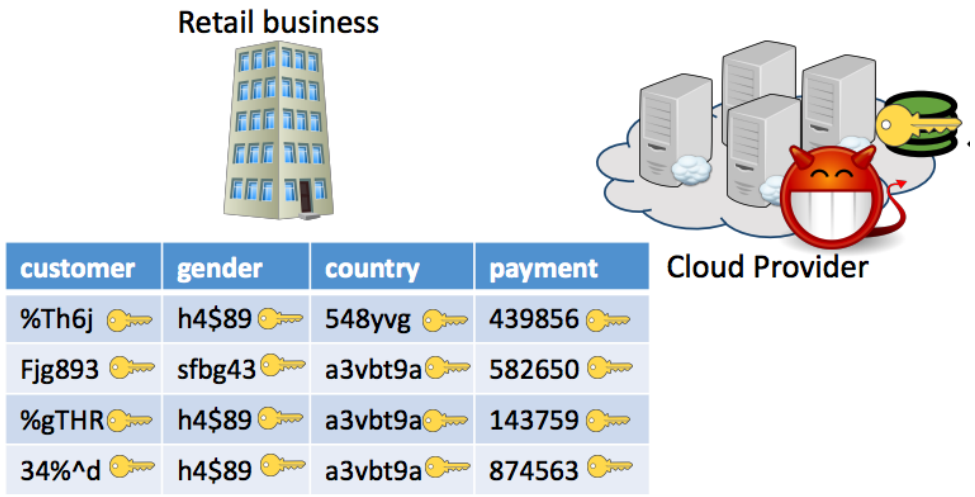
- Outsourcing big data analytics into cloud server
  - Outsourcing computation
  - Data privacy risks

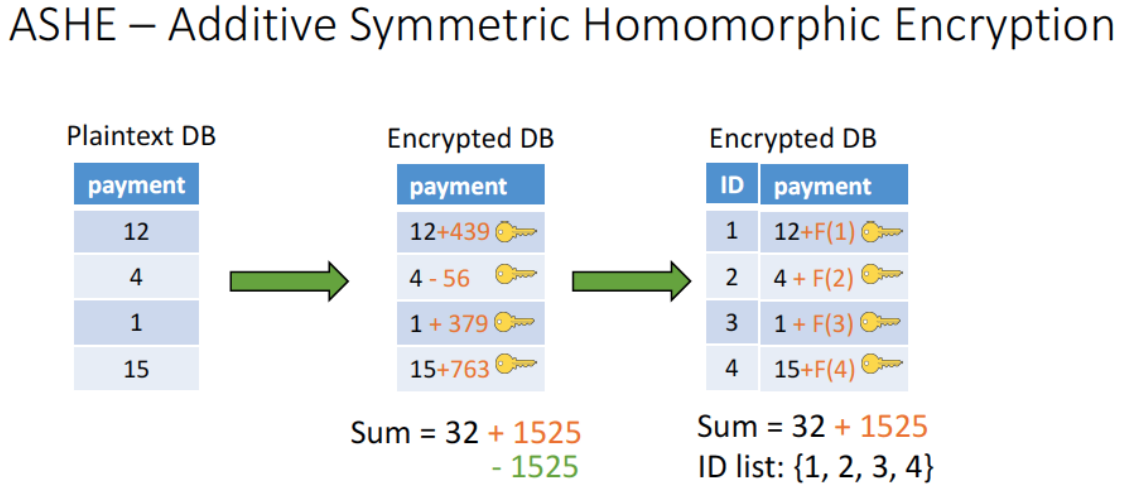


Papadimitriou, Antonis, et al. "Big data analytics over encrypted datasets with seabed." 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). 2016

Previous Solution-Encryption Scheme

- Encryption-based Scheme
  - High Computational overhead (large ciphertext & mathematical operation)
  - “Weak” privacy/security guarantee (Deterministic encryption)
  - Limited query/computation (e.g., aggregation, count)

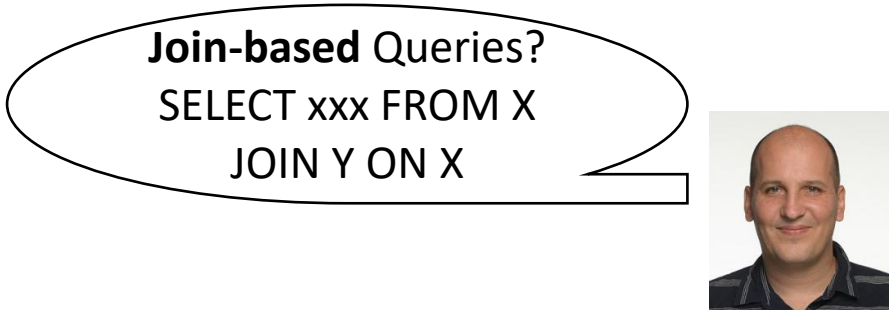




Papadimitriou, Antonis, et al. "Big data analytics over encrypted datasets with seabed." 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). 2016

Motivation for Alternative Solution

- JOIN operation?
  - Answer: Unluckily NOT
  - WHY?: Non-deterministic function.
  - Follow-up: Deterministic Encryption?

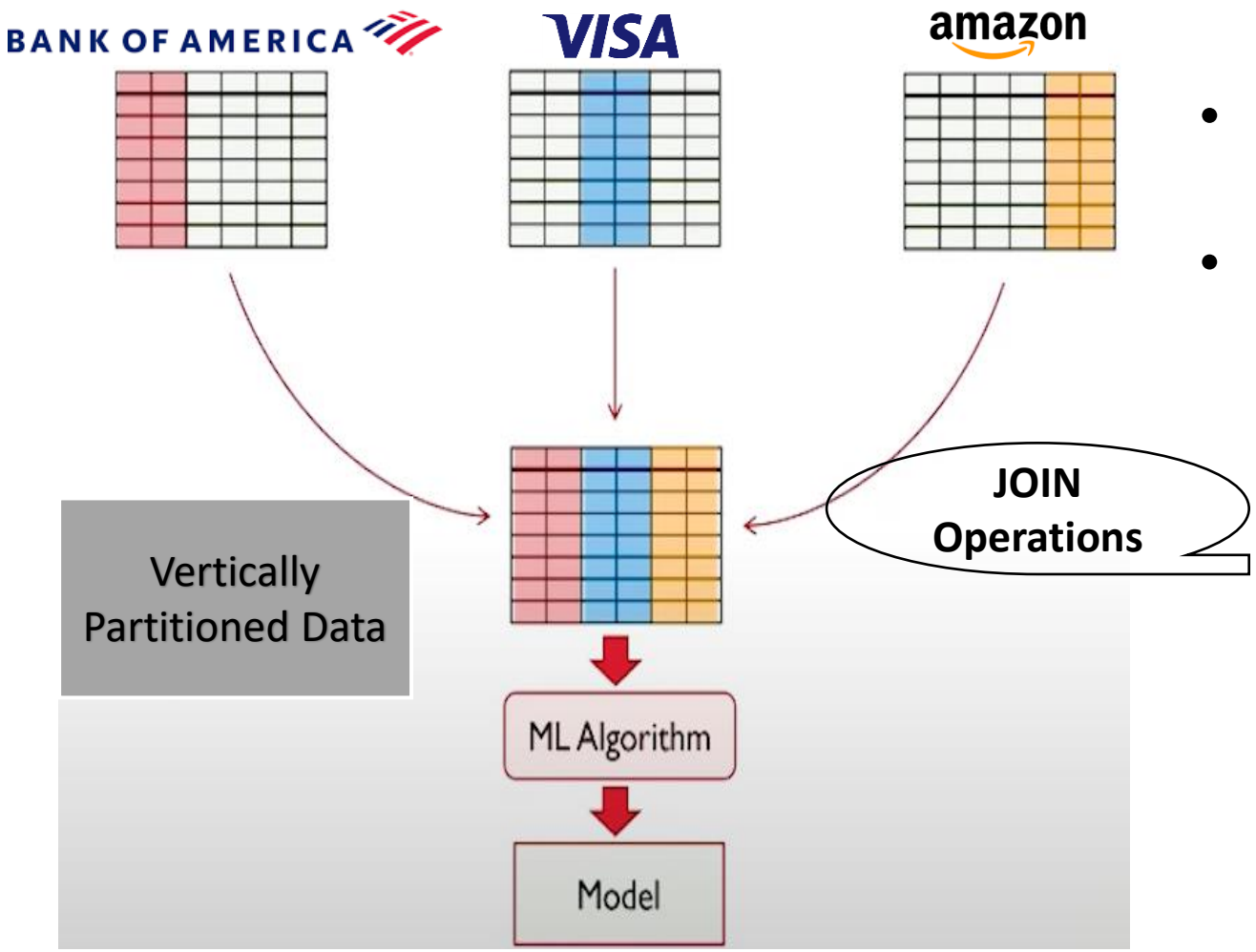


- There are other cryptographic-building blocks!
  - **Secret Sharing!**

Million Dollar Question!

Papadimitriou, Antonis, et al. "Big data analytics over encrypted datasets with seabed." 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). 2016
Grubbs P, Sekniqi K, Bindschaedler V, et al. Leakage-abuse attacks against order-revealing encryption[C]//2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017: 655-672.

The Essence of JOIN Table



- Fact: The data tables are assumed to be aligned but **NOT**!
- How: Secure JOIN the tables to align

Mohassel, Payman, et al. "ABY3: A mixed protocol framework for machine learning." Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018.

SSS on Outsourced Database!

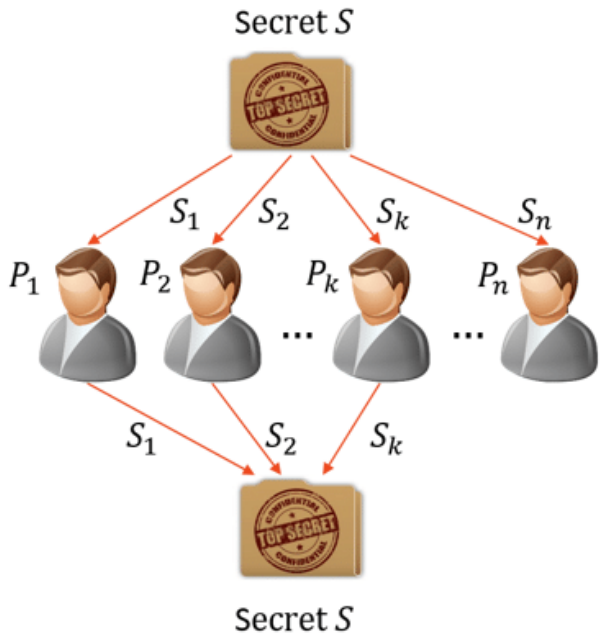
- Complex Function ✅
  - Secure Multiparty Computation (SMC) on SSS, e.g., ABY2.0 is a popular 2-PC secure machine learning framework.
  - Ready to design **JOIN function** for multiple tables or advanced queries.

- Good environment ✅
  - Cloud-based outsourcing scenerio.
  - Industrial development, Facebook (Meta) with CrypTen. Cape with TFEncrypted.

Mohassel, Payman, et al. "ABY3: A mixed protocol framework for machine learning." Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018.
Patra, Arpita, et al. "ABY2. 0: Improved mixed-protocol secure two-party computation." 30th {USENIX} Security Symposium ({USENIX} Security 21). 2021.
https://ai.facebook.com/blog/crypten-a-new-research-tool-for-secure-machine-learning-with-pytorch/
https://github.com/tf-encrypted/tf-encrypted

Secret Sharing

- Known as Shamir's Secret Sharing (SSS) in cryptography
- Split a secret s into **n** parts, where any **k out of n** parts can reconstruct the secret but any **less than k** parts cannot. **(k,n)** threshold.
- Construct Principle
  - Based on **Lagrange interpolation theorem**: k points is enough to uniquely determine a polynomial of degree less than or equal to k-1.
  - 2 points determine a line, 3 points for parabola.



Shamir Adi. How to share a secret [J]. Communications of the ACM, 1979, 22(11): 612-613.
https://en.wikipedia.org/wiki/Shamir%27s_Secret_Sharing

Toy Example (Integer arithmetic)

- Given a secret 1234, which is to divided into **5** shares and any **3** shares can be used to reconstruct the secret. **(3, 5)** threshold.
- Share Preparation
  1. construct a 2-degree polynomial : $f(x) = \mathbf{1234} + 166x + 94x^2$
  2. Compute at $x = 1, 2, 3, 4, 5$, as 5 shares $(1, 1494), (2, 1942), (3, 2578), (4, 3402), (5, 4414)$.
- Reconstruction from 3 shares

$$\ell_0(x) = \frac{x - x_1}{x_0 - x_1} \cdot \frac{x - x_2}{x_0 - x_2} = \frac{x - 4}{2 - 4} \cdot \frac{x - 5}{2 - 5} = \frac{1}{6}x^2 - \frac{3}{2}x + \frac{10}{3}$$

$$\ell_1(x) = \frac{x - x_0}{x_1 - x_0} \cdot \frac{x - x_2}{x_1 - x_2} = \frac{x - 2}{4 - 2} \cdot \frac{x - 5}{4 - 5} = -\frac{1}{2}x^2 + \frac{7}{2}x - 5$$

$$\ell_2(x) = \frac{x - x_0}{x_2 - x_0} \cdot \frac{x - x_1}{x_2 - x_1} = \frac{x - 2}{5 - 2} \cdot \frac{x - 4}{5 - 4} = \frac{1}{3}x^2 - 2x + \frac{8}{3}$$

$$\begin{aligned} f(x) &= \sum_{j=0}^{2} y_j \cdot \ell_j(x) \\ &= y_0\ell_0(x) + y_1\ell_1(x) + y_2\ell_2(x) \\ &= 1942\left(\frac{1}{6}x^2 - \frac{3}{2}x + \frac{10}{3}\right) + 3402\left(-\frac{1}{2}x^2 + \frac{7}{2}x - 5\right) + 4414\left(\frac{1}{3}x^2 - 2x + \frac{8}{3}\right) \\ &= 1234 + 166x + 94x^2 \end{aligned}$$

Some Notations about SSS

- Perfect secrecy on the **Finite filed**.
  - Security Setting: **non-colluded servers.**
- Property
  - Simple yet effective to compute (vs. homomorphic encryption)
  - Additive homomorphic (friendly useful for aggregation)
- Variant SS
  - Additive Secret Sharing (light computation)
  - Split into several shares, where reconstruction require all the shares
  - Example, S=4, create 2 shares, s1=1 and s2=3

Secure JOIN on Tables

- Problem definition
  - Multiple k data owners hold data tables D_i (have common columns), we want make JOIN on one specific attribute A_{join} with these tables without disclosing the tables' information
- Solution overview
  - Given the domain cardinality of A_{join} is C. Every data table can map every row value of A_{join} to a vector of size C, $[v_1, v_2, v_3, \cdots, v_C]$.
  - If the mapping value exists, the v is set to 1, otherwise to 0. We can outsource such vectors in secret share to the servers
  - The servers will aggregate the share of vectors (do computation) and send back to data owners.

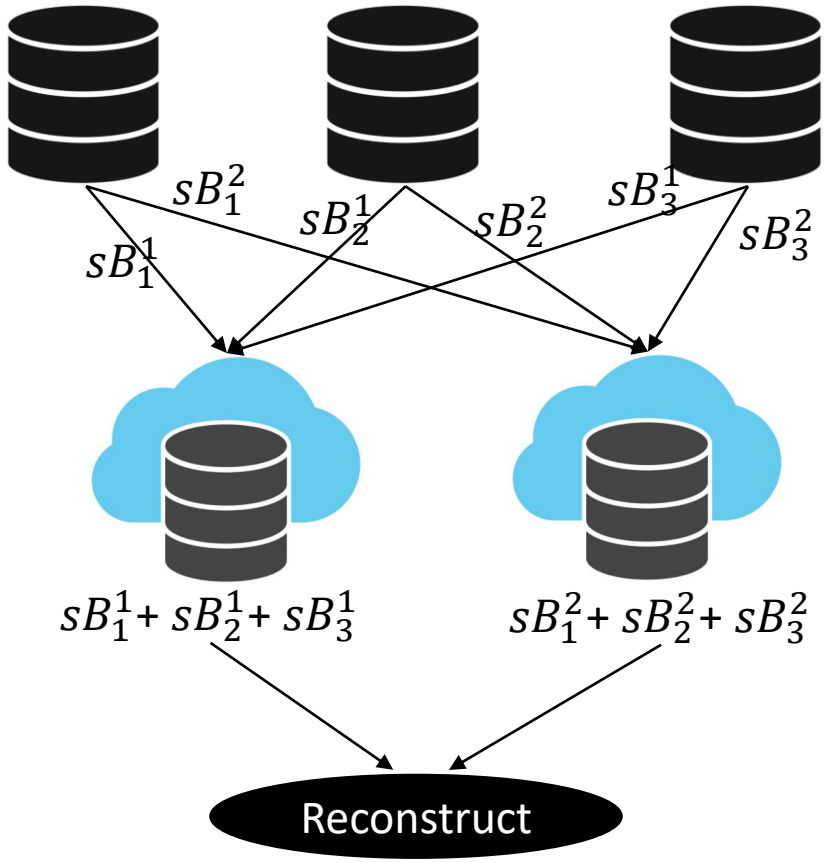
Example Demonstration (integer addition)

- The correctness of secret sharing
  - We have three data owners' tables

  B1= [0, 1, 0]

  B2= [1, 1, 0]

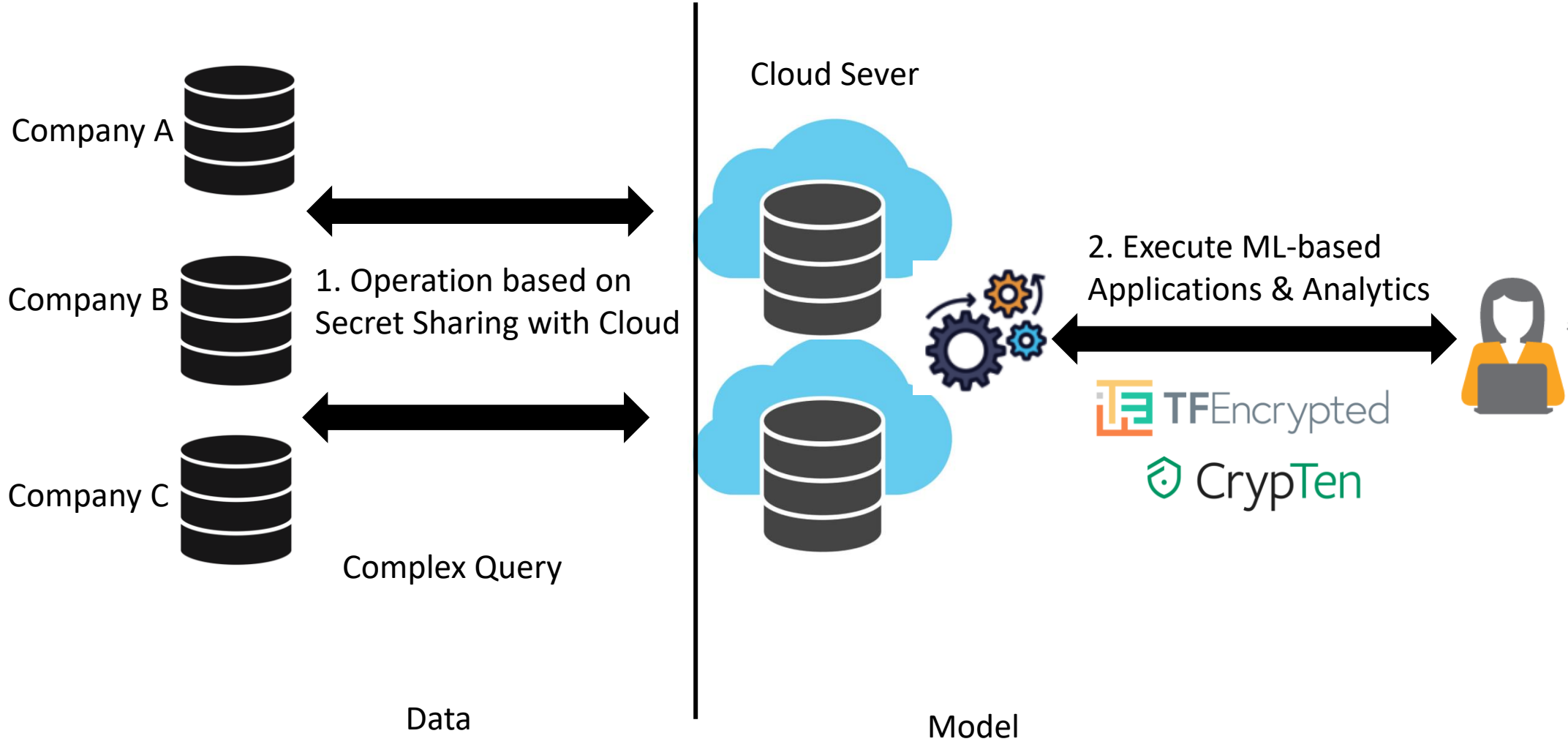
  B3= [1, 1, 1]
  - Aggregate the 3 tables, we can get a vector [2, **3**, 1].
- Security guarantee
  - Servers does not know the values
  - Utilize cyclic group under modulo multiplication to protect intermediate results
  - G={1,3,4,5,9} mod 11. 5



Detailed Protocol

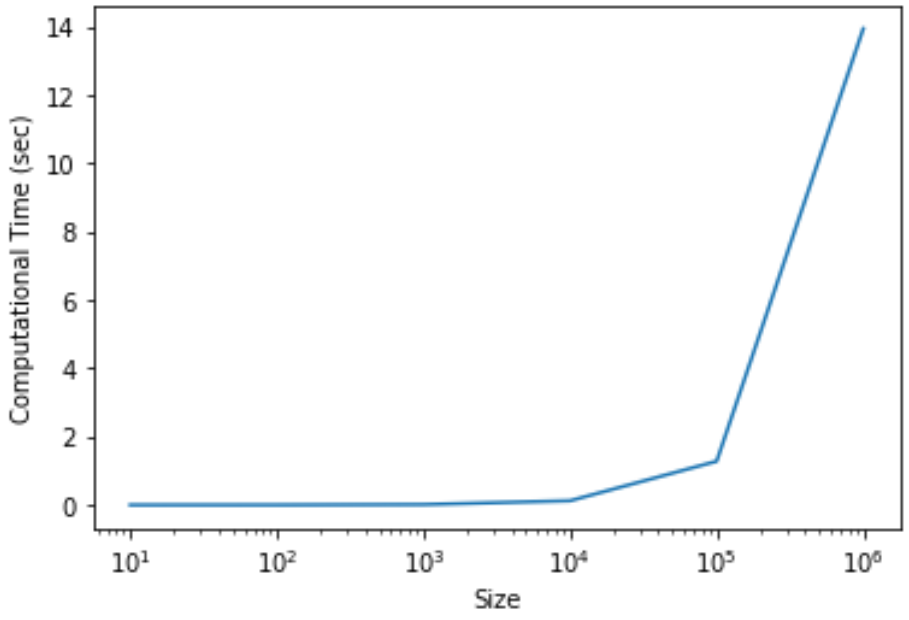
1. Each data owner i=1,2, ..k hashes the domain value of join columns
   a) Compute $\mathrm{B^i} = \left[v_1^i, v_2^i, \cdots, v_C^i\right], v = 1/0.$
   b) Construct m shares (m is the number of server) $SSS\left(B^i\right) = B_1^i, B_2^i, \cdots, B_m^i$
   c) Send the shares to the server 1, 2, … m
2. Each server j aggregates such shares received from the data owner
   a) Compute $\Sigma^{\mathrm{k}}\mathrm{B}_{\mathrm{j}}^{\mathrm{i}}$, construct as the normal secret share
   b) Send such aggregated shares back to the data owner
3. Each data owner will locally aggregate the values from the server
   a) If the value corresponding to one cell is equal to the number of data tables. This indicates that this is the intersection value.

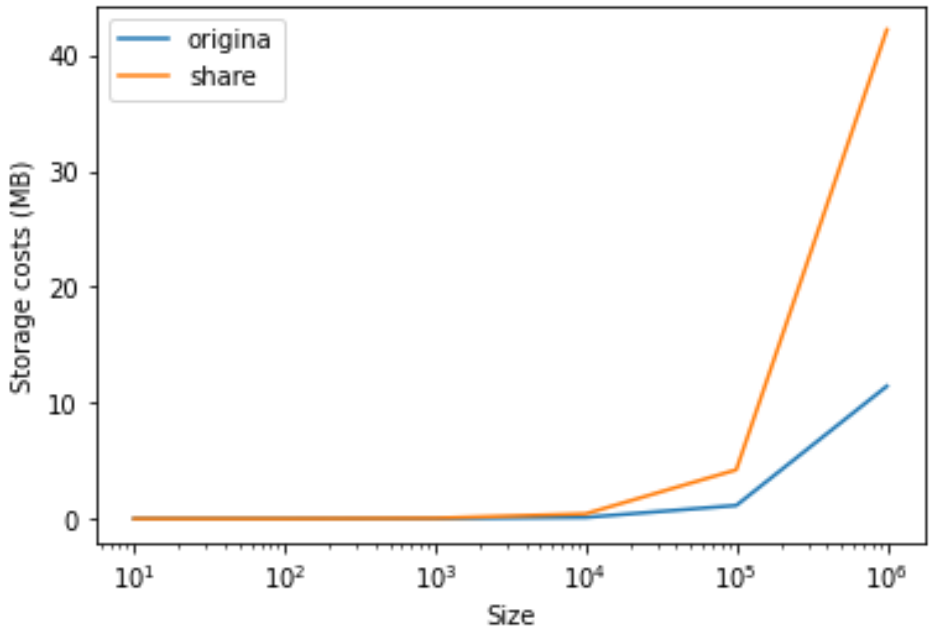
SSS-based Big Data Analytics Scenario



Experimental Evaluation

- Simulated data table of cardinality [10, 100, …., 1M]
- Evaluate the running time and storage costs





https://github.com/xhhhaha/cs595db

Conclusion & Future Work

- Secure JOIN could be essential for outsourced computation, e.g., ML
- Secret sharing can server as a strong cryptographic building block

- Mitigate/alleviate the model setting (under malicious setting/verifiability)
- Design SMC protocols to support direct/flexible queries

- Integrate with the Spark
- More system optimization & design, e.g., reduce the interaction and speed up/ evaluate on the real-world dataset
- Build top-level ML applications with secure-shared based function