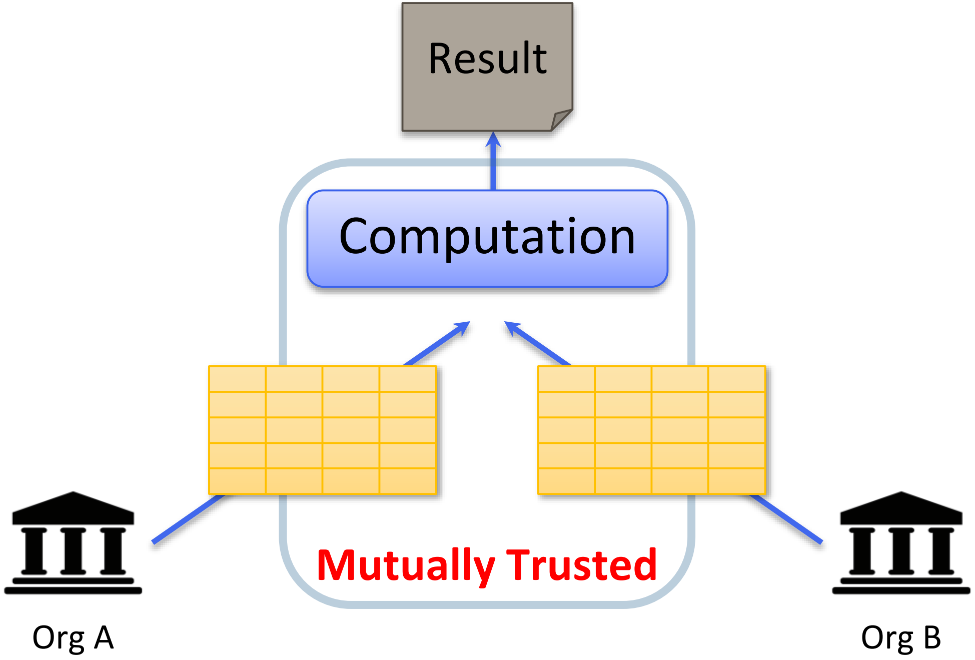
Confidential Computing

Analytics with data privacy and control

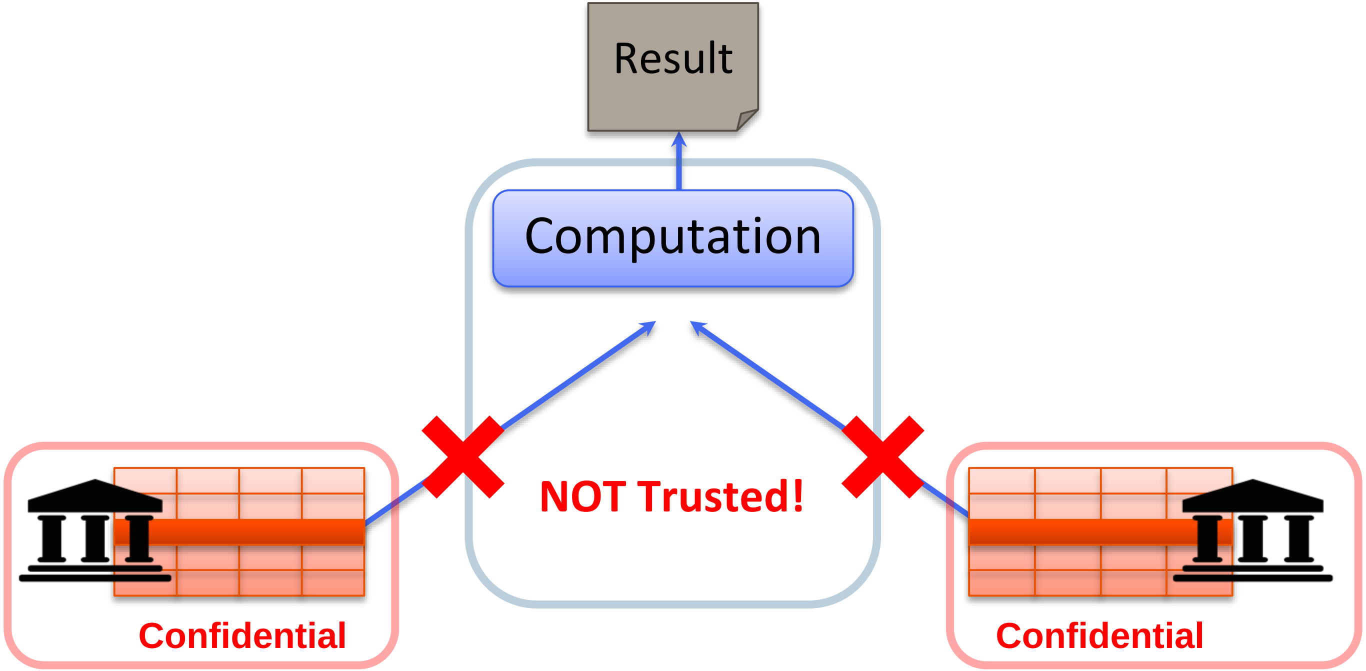
Brian Thorne

Privacy Preserving Linkage Motivation

Multi-Organisation Analytics Today

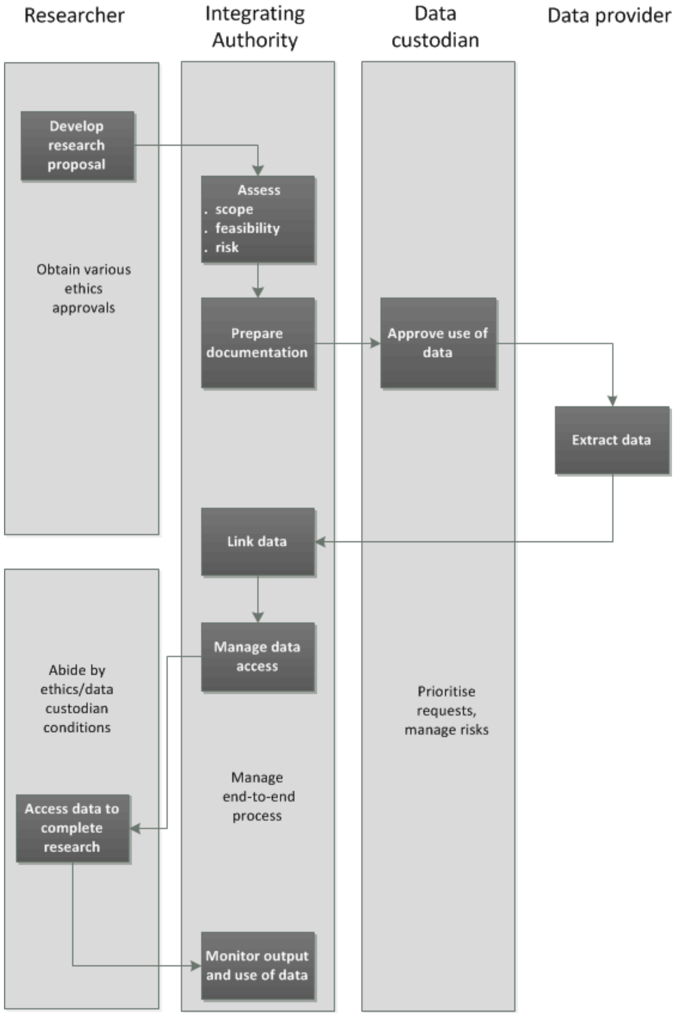


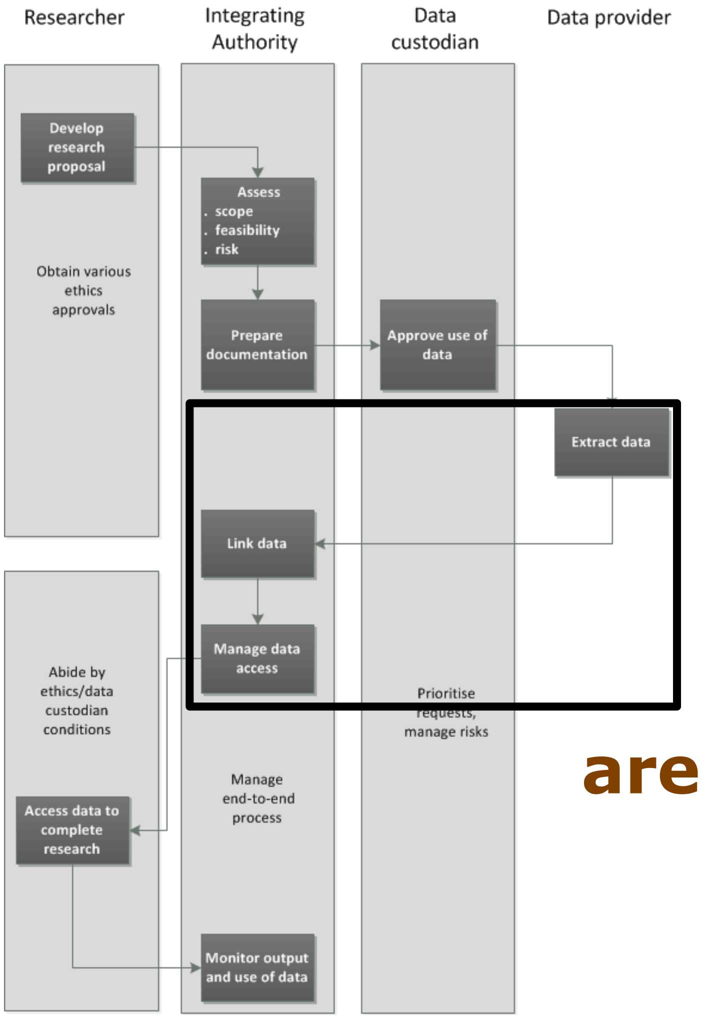
But many Opportunities are Blocked



Entity Matching

Overview of a typical data integration project within GOV

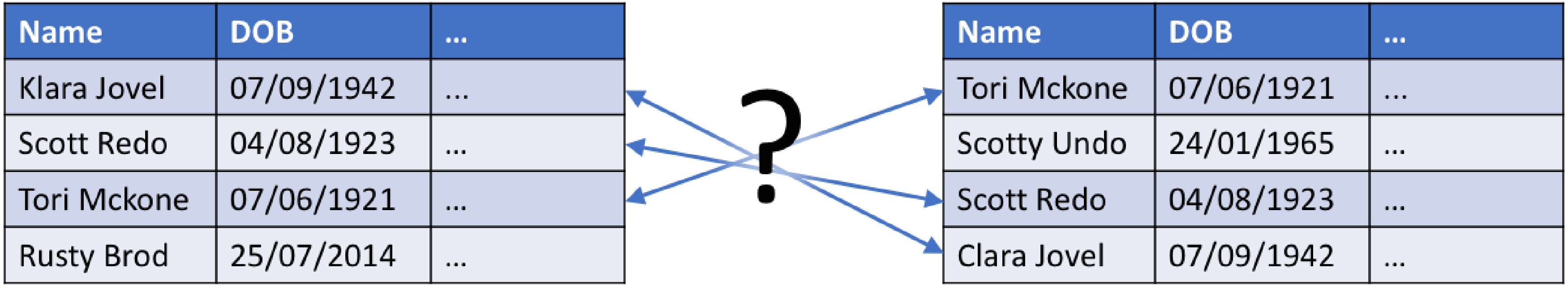

Researcher
Integrating Authority
Data custodian
Data provider
Develop research proposal
Assess
. scope
. feasibility
. risk
Obtain various ethics approvals
Prepare documentation
Approve use of data
Extract data
Link data
Manage data access
Abide by ethics/data custodian conditions
Prioritise requests, manage risks
Manage end-to-end process
Access data to complete research
Monitor output and use of data


Researcher
Integrating Authority
Data custodian
Data provider
Develop research proposal
Assess
. scope
. feasibility
. risk
Obtain various ethics approvals
Prepare documentation
Approve use of data
Extract data
Link data
Manage data access
Prioritise requests, manage risks
Abide by ethics/data custodian conditions
Manage end-to-end process
Access data to complete research
Monitor output and use of data

area I’m covering today

Privacy-preserving entity resolution

- **Goal**: match *corresponding* rows in two distinct databases

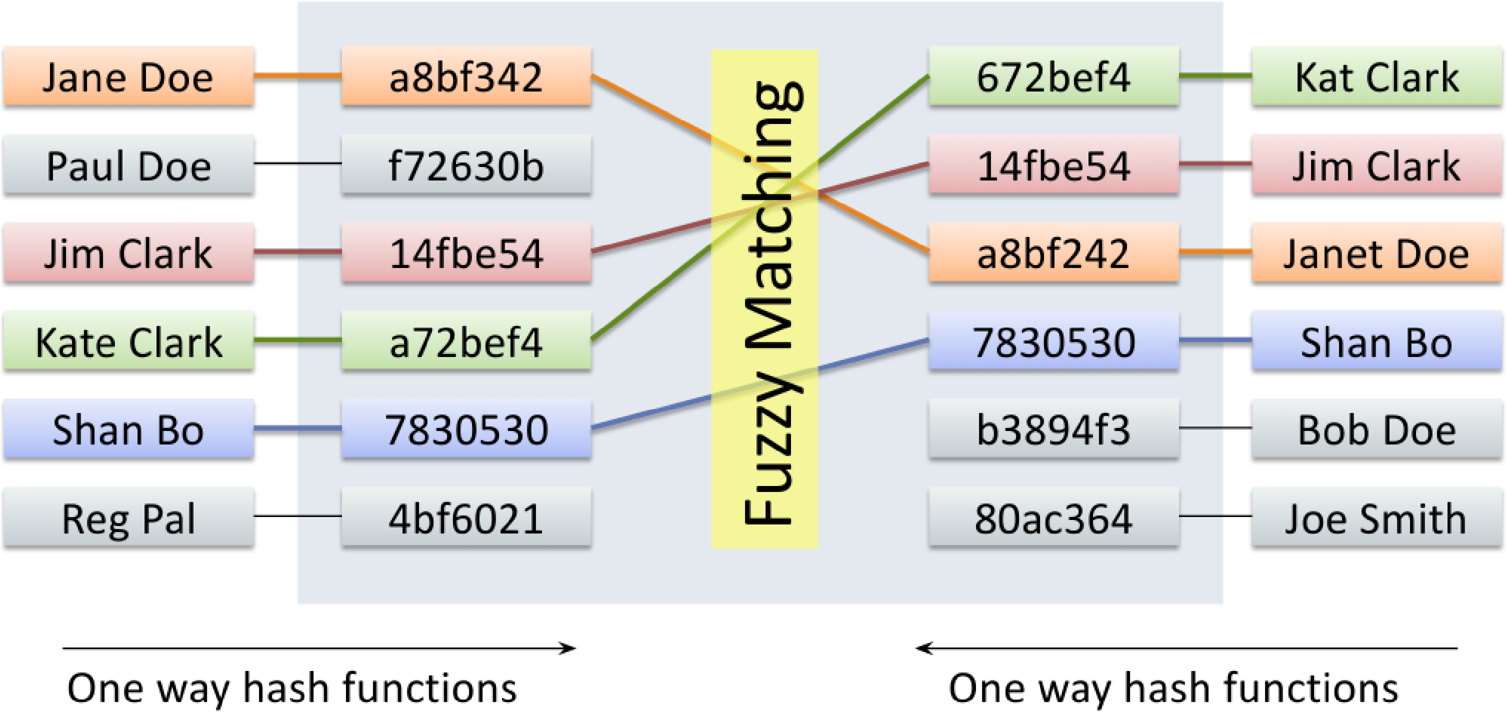


Name	DOB	...
Klara Jovel	07/09/1942	...
Scott Redo	04/08/1923	...
Tori Mckone	07/06/1921	...
Rusty Brod	25/07/2014	...

Name	DOB	...
Tori Mckone	07/06/1921	...
Scotty Undo	24/01/1965	...
Scott Redo	04/08/1923	...
Clara Jovel	07/09/1942	...

- **Constraint**: can’t share Personally Identifiable Information (PII)
- **Solution**: *fuzzy & private* matching

Privacy-preserving entity resolution



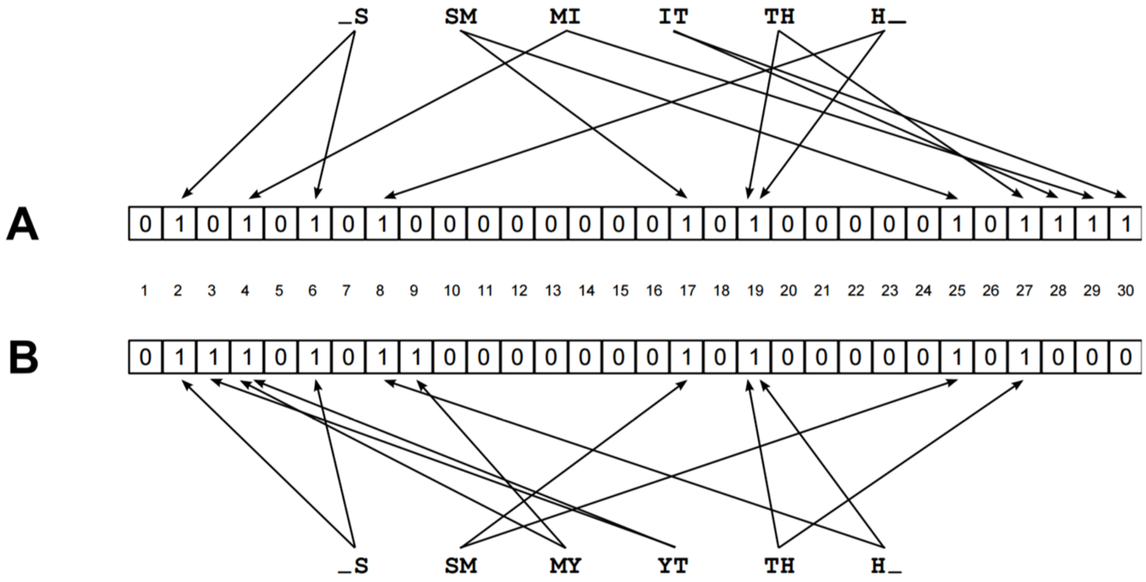
How?

For every record we process the PII into a **Cryptographic Longterm Key** or (CLK)

Briefly, we hash the bi-grams for each PII feature into a bloom filter.

https://github.com/n1analytics/clkhash/

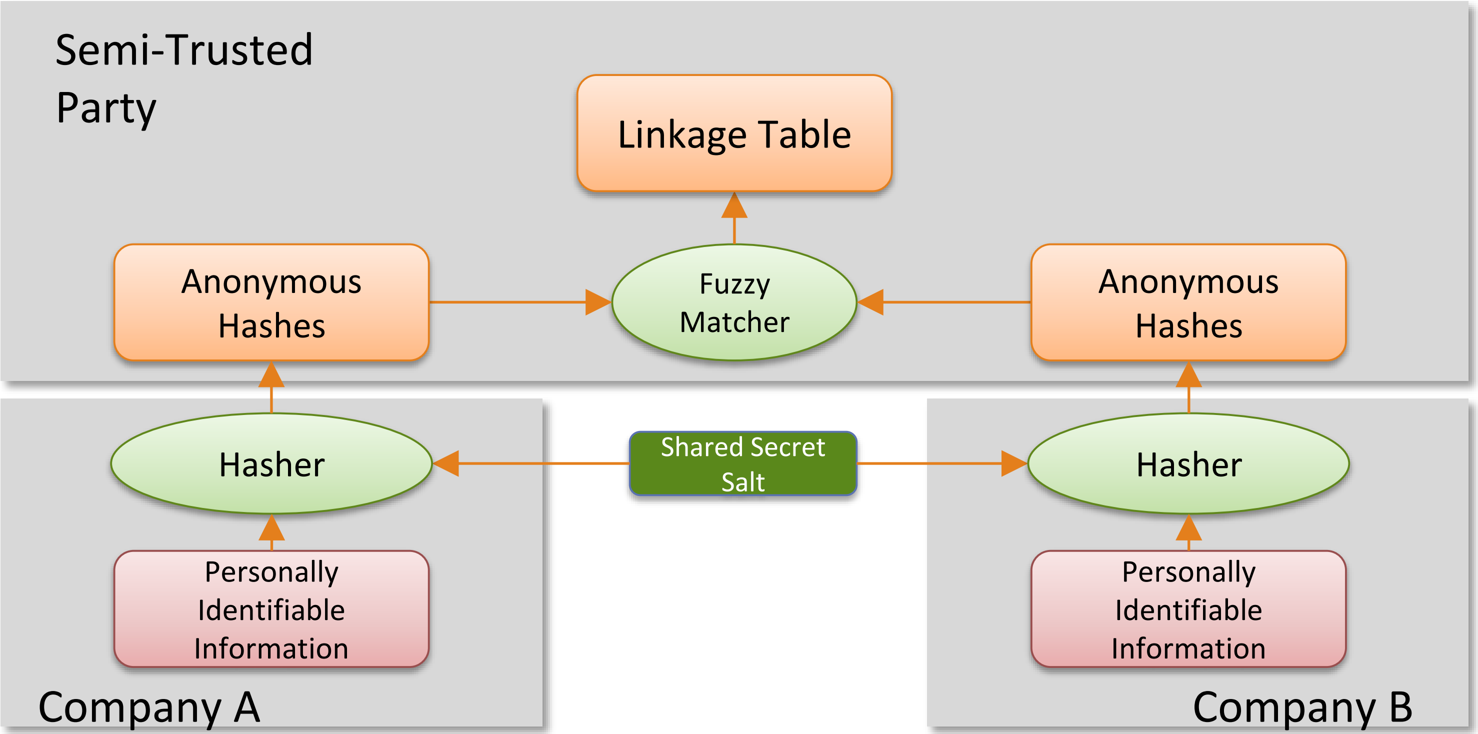
Cryptographic Longterm Key



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
A	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	1	1	1
B	0	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0

Figure 1: Example for the mapping of two names (SMITH, SMYTH) using bigrams and two hash functions to two Bloom-Filters (A, B) with 30 bits each.

Private Record Linkage



PII cannot be recovered from the hashes

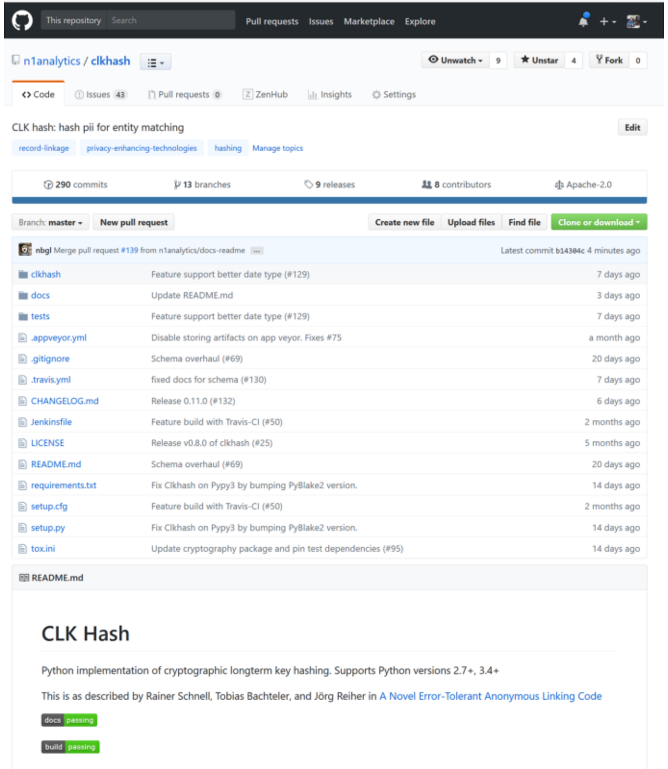
anonlink

Semi-trusted Third Party

- Only hashed data is uploaded to the entity resolution service
- Hash security relies on a shared secret between parties
- Implemented the service with a simple JSON + REST API
- All communication is secured with HTTPS
- Authentication tokens created for each job
- Result type and agreed schema is set at beginning

Client side: Command Line Utility

- Locally hashes PII data
- Creates new mapping jobs on the server
- Uploads hash data
- Retrieves results



CLK hash: hash pii for entity matching

nbgl Merge pull request #139 from n1analytics/docs-readme … Latest commit b14304c 4 minutes ago

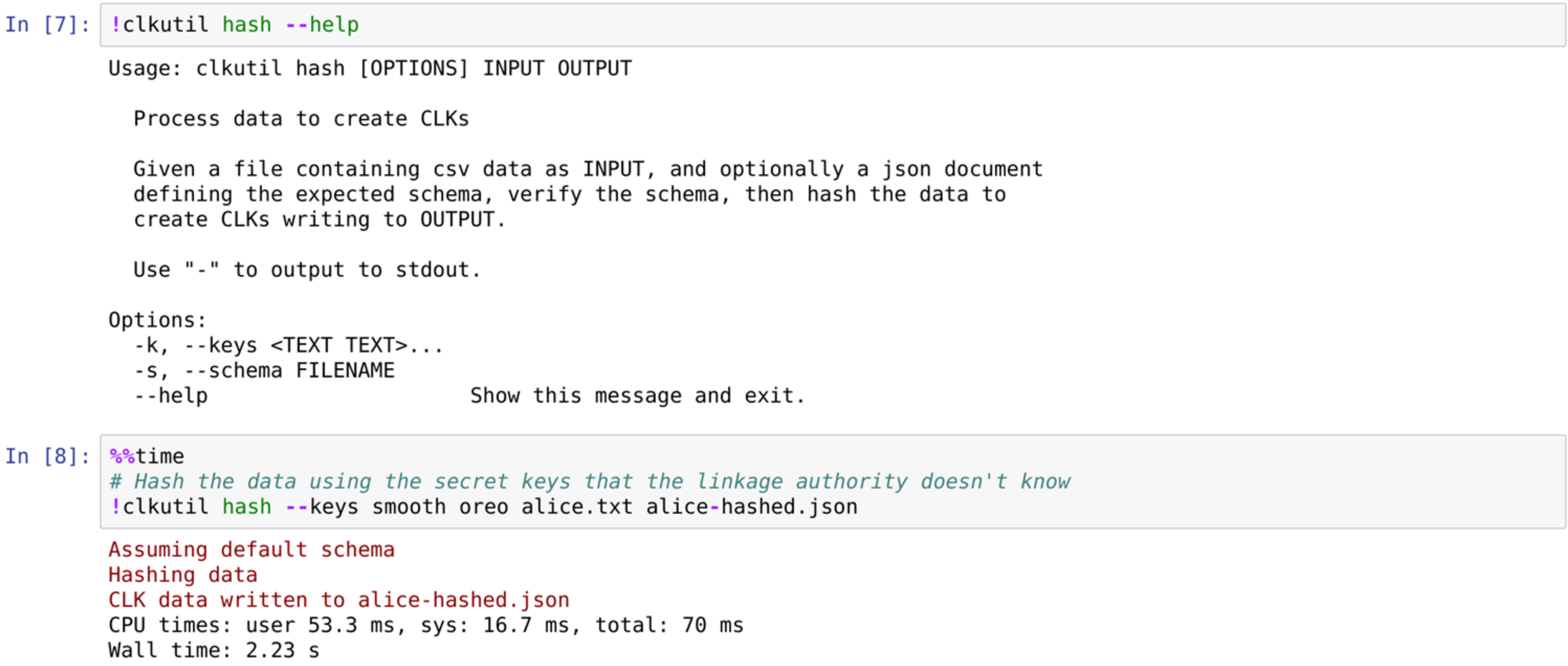
clkhash	Feature support better date type (#129)	7 days ago
docs	Update README.md	3 days ago
tests	Feature support better date type (#129)	7 days ago
.appveyor.yml	Disable storing artifacts on app veyor. Fixes #75	a month ago
.gitignore	Schema overhaul (#69)	20 days ago
.travis.yml	fixed docs for schema (#130)	7 days ago
CHANGELOG.md	Release 0.11.0 (#132)	6 days ago
Jenkinsfile	Feature build with Travis-CI (#50)	2 months ago
LICENSE	Release v0.8.0 of clkhash (#25)	5 months ago
README.md	Schema overhaul (#69)	20 days ago
requirements.txt	Fix Clkhash on Pypy3 by bumping PyBlake2 version.	14 days ago
setup.cfg	Feature build with Travis-CI (#50)	2 months ago
setup.py	Fix Clkhash on Pypy3 by bumping PyBlake2 version.	14 days ago
tox.ini	Update cryptography package and pin test dependencies (#95)	14 days ago

README.md

CLK Hash

Python implementation of cryptographic longterm key hashing. Supports Python versions 2.7+, 3.4+

This is as described by Rainer Schnell, Tobias Bachteler, and Jörg Reiher in A Novel Error-Tolerant Anonymous Linking Code



```
In [7]: !clkutil hash --help
```

```
Usage: clkutil hash [OPTIONS] INPUT OUTPUT

  Process data to create CLKs

  Given a file containing csv data as INPUT, and optionally a json document
  defining the expected schema, verify the schema, then hash the data to
  create CLKs writing to OUTPUT.

  Use "-" to output to stdout.

Options:
  -k, --keys <TEXT TEXT>...
  -s, --schema FILENAME
  --help                     Show this message and exit.
```

```
In [8]: %%time
        # Hash the data using the secret keys that the linkage authority doesn't know
        !clkutil hash --keys smooth oreo alice.txt alice-hashed.json
```

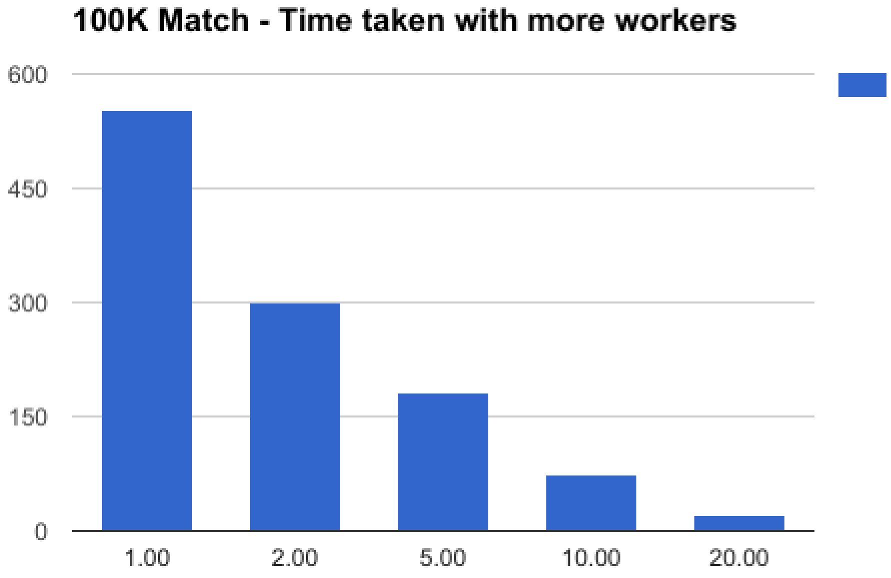
```
Assuming default schema
Hashing data
CLK data written to alice-hashed.json
CPU times: user 53.3 ms, sys: 16.7 ms, total: 70 ms
Wall time: 2.23 s
```

Performance & Case Study

Speed and Scale

- 1.3B hash comparisons/s
- Handle uploads of 35M hashes
- 1M x 1M match takes around 5 hours

Running on four r4.4xlarge instances on AWS



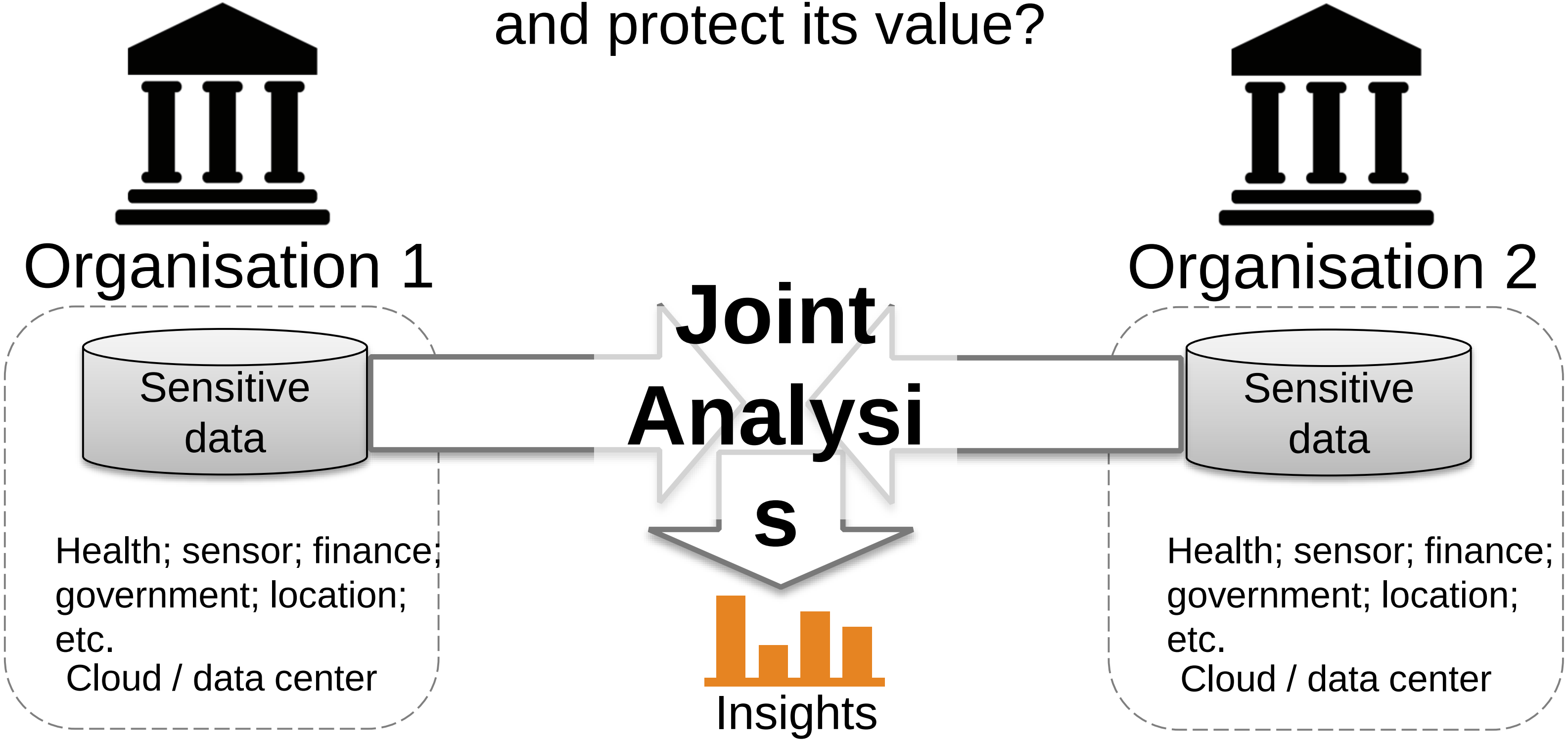
Computing similarity between CLKs is a very parallel problem. Our implementation utilizes multiple workers to carry out comparisons using a kubernetes cluster

Data61 Privacy Projects

Protari, SENDA, Risk Identification, N1, Private Linkage

Confidential Computing

How can we learn insights from data from multiple sources and protect its value?



Organisation 1

Sensitive data

Health; sensor; finance; government; location; etc.

Cloud / data center

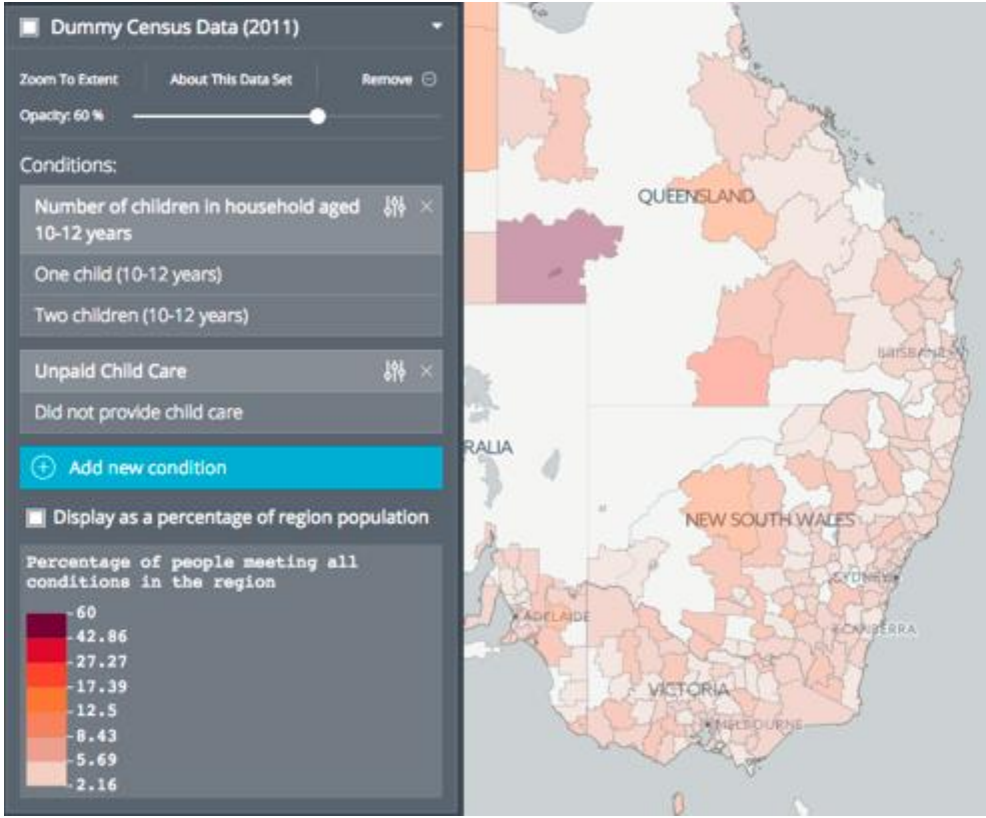
Organisation 2

Sensitive data

Health; sensor; finance; government; location; etc.

Cloud / data center

Protari



■ Dummy Census Data (2011)

Zoom To Extent | About This Data Set | Remove ⊖

Opacity: 60 %

Conditions:

Number of children in household aged 10-12 years

One child (10-12 years)

Two children (10-12 years)

Unpaid Child Care

Did not provide child care

⊕ Add new condition

□ Display as a percentage of region population

```
Percentage of people meeting all
conditions in the region
```

- 60
- 42.86
- 27.27
- 17.39
- 12.5
- 8.43
- 5.69
- 2.16

N1 Analytics

Goals

:

Release your data without losing control

Access data that is currently too sensitive

Technologies

: Fully, Somewhat, Partially Homomorphic encryption

Secure Multiparty Compute

Learning from Aggregates

Capabilities

:

Learn and deploy models

Secure aggregation of data

Clustering/Anomaly Detection

Confidential Computing

Analytics with data privacy and control

www.n1analytics.com