Count Everything: Secure Count Query Framework Across Big Data Centers

Presenters: Ida Sim (UCSF) & Xiaoqian Jiang (UCSD)
November 30, 2016
Team Organization
The Big Problem

- Big data vision includes computing across multiple types of data for the same person, e.g.,
  - electronic health record (EHR) data
  - mobile health (mHealth) data
  - genome sequence data

- But EHR, mHealth, and genome data for the same person are spread across different hosts, each exposing data via different application program interfaces (APIs)
  - faced by BD2K, PMI, Cancer Moonshot, etc. researchers
Problems
1. Identity resolution is difficult
2. Person has not given permissions
3. Institutional privacy policies and tech are a barrier
4. Different APIs (syntax and semantics differ)
Count Everything: counts first, then more advanced queries

Problems:
1. Identity resolution is difficult
2. Person has not given permissions
3. Institutional privacy policies and tech are a barrier
4. Different APIs (syntax and semantics differ)
Select `count(*)` from MD2K A, i2b2 B, GA4GH C where 
A.blood Glucose>110 and B.smoker=true and C.r123140=C

**Problems**
1. Identity resolution is difficult
2. Person has not given permissions
3. Institutional privacy policies and tech are a barrier
4. Different APIs (syntax and semantics differ)

**Solution**
1. Identity resolution is difficult
2. Person has not given permissions
3. **Count Everything:**
   - queries are done locally
   - patient-level data not shared
   - only counts are shared
4. Different APIs (syntax and semantics differ)
Problems
1. Identity resolution is difficult
2. Person has not given permissions
3. Institutional privacy policies and tech are a barrier
4. Different APIs (syntax and semantics differ)

Solution
1. Identity resolution is difficult
2. Person has not given permissions
3. Count Everything:
   - queries are done locally
   - patient-level data not shared
   - only counts are shared
4. Standard APIs (aligned syntax and semantics)
Problems
1. Identity resolution is difficult
2. Person has not given permissions
3. Institutional privacy policies and tech are a barrier
4. Different APIs (syntax and semantics differ)

Scalable Solution
1. Identity resolution is difficult
2. Person has not given permissions
3. Count Everything:
   - queries are done locally
   - patient-level data not shared
   - only counts are shared
4. Count Everything containers with standard API endpoints
Project Aims and Accomplishments

Completed

1. Interoperable APIs
2. Secure querying
3. Working demo

Unfunded

1. Federated security (current)
2. Docker containers (future, if funded)
Project Aim 1: Interoperable APIs

Query APIs created by each of the data providers:

PIC-SURE

n = 41,000
1,300 variables

http://bx2k-picure.hms.harvard.edu/

EHR Data

MD2K

n = 2504
synthetic data

http://ce.gaqomhio.8083/v1.0.M2/

mHealth Data

Global Alliance for Genomics & Health

n = 2504

http://coega4gh.westus.doudapp.azure.com/bc/

Genomic Data

Count Everything
Secure Query

BD2K Researcher
Data Elements

PIC-SURE API
- Mean Systolic/Diastolic Blood Pressure (mmHg)
- Gender
- Age
- Race
- Height (cm)
- Weight (kg)
- BMI
- Glucose, Serum (mg/mL)
- Uric acid (mg/dL)
- Total Cholesterol (mg/dL)

n = 41,000
1,300 variables

Open mHealth API
- Systolic/Diastolic Blood Pressure (mmHg)
- Sleep duration (hours)
- Height (cm)
- Weight (kg)
- Temperature (C)
- Body Fat Percentage (%)
- Blood Glucose (mg/dL)

n = 2504
synthetic data

GA4GH API
- Reference name
- Start position
- Alternate bases
- Assembly ID

n = 2504
Synthetic Cohort

- Unique global ID assigned to 2504 “patients” from each data set
Project Aim 2: Secure Querying

EHR Data
- n = 41,000
- 1,300 variables
- http://bd2k-picsure.hms.harvard.edu/

mHealth Data
- n = 2504
- synthetic data
- http://ce.qa.ohio.edu/v1.0.M2/

Genomic Data
- n = 2504
- http://ce.qa4gh-westus.cloudapp.azure.com/ce/

BD2K Researcher

Count Everything Processor

Count Everything Crypto Service Provider
Project Aim 3: Demonstration

- **n = 41,000**
  - 1,300 variables
  - EHR Data

- **n = 2,504**
  - synthetic data
  - mHealth Data

- **n = 2,504**
  - Genomic Data

**1000 Genomes**
- A Deep Catalog of Human Genetic Variation

---

Current Work: Auth0 Federated Security

1. Synthetic cohort
2. Auth0 security framework
3. Count Everything:
   - queries are done locally
   - patient-level data not shared
   - only counts are shared
4. Count Everything standard API endpoints → Docker containers

- EHR Data
  - http://chdi-persur.hms.harvard.edu/
- mHealth Data
  - http://o.eqi.ohio.io/8000/v1.0.M2/
- Genomic Data
  - http://cegs4gh.westus.cloudapp.azure.com/ei
Homomorphic encryption (HME)

Three roles:

**Data Owner:** Providing data for counting;

**Processor:** Collecting data from data owners in ciphertext form and process the data for CSP evaluation;

**Cryptographic Service Provider (CSP):** Providing the cryptography service and evaluate the ciphertext of final results.
Homomorphic Encryption
Homomorphic encryption (Palliar)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>E(A)</th>
<th>E(B)</th>
<th>E(C)</th>
<th>D(E(A)*E(B)*E(C))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>rand</td>
<td>0</td>
<td>21...572</td>
<td>45...865</td>
<td>35...545</td>
<td>rand</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>rand</td>
<td>72...839</td>
<td>34...214</td>
<td>23...424</td>
<td>rand</td>
</tr>
<tr>
<td>rand</td>
<td>0</td>
<td>rand</td>
<td>77...717</td>
<td>57...324</td>
<td>34...325</td>
<td>rand</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>53...943</td>
<td>34...541</td>
<td>45...234</td>
<td>0</td>
</tr>
<tr>
<td>rand</td>
<td>rand</td>
<td>rand</td>
<td>13...758</td>
<td>65...434</td>
<td>23...423</td>
<td>rand</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>09...121</td>
<td>23...885</td>
<td>28...424</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ c = g^m \cdot r^n (mod \ n^2) \]

- \( c \) is the ciphertext, \( m \) is the plaintext
- \( r \) is a random number, \( g, n \) are the public keys

Counts of zeros = 2
Summary

Created a synthetic cohort by assigning unique global IDs to convenience sets of mobile, 1000 Genome project, and clinical CDC NHANES data

Exposed the data through Open mHealth, GA4GH, and PIC-SURE APIs

Developed a semi-centralized system using iDASH as the HME processor and bioCADDIE as a cryptographic service provider (CSP) to support computation on encrypted data

Consortium Contributions

- Used data APIs and tooling from 3 BD2K Centers
- Leveraged bioCADDIE/iDASH work on secure querying architectures
- BD2K Consortium funding enabled quick integration and scaling across centers
- Demonstrates key parts of a scalable solution to big data computing that takes into account reality of multiple data hosts