Privacy-Preserving Analytics in and out of the Clouds

Jon Crowcroft,
http://www.cl.cam.ac.uk/~jac22

This work is funded in part by the EPSRC Databox (EP/N028260/1), NaaS (EP/K031724/2) and Contrive (EP/N028422/1) and Turing/Maru (EP/N510129/1) projects.
Compute-First Networking

• Should do to IP, what IP did to POTS
• Need to think 40-50 years ahead in s/w terms
• Scale up (10^23 nodes per person)
• Verification or Generative (correct & smaller by design) code
• Invariants (C (max latency), plank’s constant (min size)) matter more…
1 In Cloud. Trust Issues: Provider Perspective

Cloud provider does not trust users

Use virtual machines to isolate users from each other and the host

VMs only provide one way protection
Trust Issues: User Perspective

Users trust their applications

Users must implicitly trust cloud provider

Existing applications implicitly assume trusted operating system
Trusted Execution Support with Intel SGX

- Users create HW-enforced trusted environment (enclave)
- Supports unprivileged user code
- Protects against strong attacker model
- Remote attestation
- Available on commodity CPUs
Trusted execution environment (TEE) in process
- Own code & data
- Controlled entry points
- Provides confidentiality & integrity
- Supports multiple threads
- Full access to application memory
Intel Software Guard Extensions (SGX)

Extension of Instruction Set Architecture (ISA) in recent Intel CPUs

Protects confidentiality and integrity of code & data in untrusted environments
  – Platform owner considered malicious
  – Only CPU chip and isolated region trusted
SGX Enclaves

SGX introduces notion of **enclave**
- Isolated memory region for code & data
- New CPU instructions to manipulate enclaves and new enclave execution mode

Enclave memory **encrypted** and **integrity-protected** by hardware
- Memory encryption engine (MEE)
- No plaintext secrets in main memory

Enclave memory can be accessed only by enclave code
- Protection from privileged code (OS, hypervisor)

Application has ability to defend secrets
- Attack surface reduced to just enclaves and CPU
- Compromised software cannot steal application secrets
SGX SDK Code Sample

SGX application: untrusted code

```c
char request_buf[BUFFER_SIZE];
char response_buf[BUFFER_SIZE];

int main()
{
    ...
    while(1)
    {
        receive(request_buf);
        ret = EENTER(request_buf, response_buf);
        if (ret < 0)
            fprintf(stderr, "Corrupted message\n");
        else
            send(response_buf);
    }
    ...
}
```

Enclave: trusted code

```c
char input_buf[BUFFER_SIZE];
char output_buf[BUFFER_SIZE];

int process_request(char *in, char *out)
{
    copy_msg(in, input_buf);
    if(verify_MAC(input_buf))
    {
        decrypt_msg(input_buf);
        process_msg(input_buf, output_buf);
        encrypt_msg(output_buf);
        copy_msg(output_buf, out);
        EEXIT(0);
    }
    else
        EEXIT(-1);
}
```

Server:
- Receives encrypted requests
- Processes them in enclave
- Sends encrypted responses
SGX Enclave Construction

Enclave populated using special instruction (EADD)
- Contents initially in untrusted memory
- Copied into EPC in 4KB pages
- Both data & code copied before execution commences in enclave

```c
1 { char input_buf[BUFFER_SIZE];
2   char output_buf[BUFFER_SIZE];
3   int process_request(char *in, char *out)
    { copy_msg(in, input_buf);
      if(verify_MAC(input_buf))
      { decrypt_msg(input_buf);
        process_msg(input_buf, output_buf);
        encrypt_msg(output_buf);
        copy_msg(output_buf, out);
        EEXIT(0);
      } else
        EEXIT(-1);
    }
}
```
SGX Support for Spark

SGX-Spark
- Protect data processing from infrastructure provider
- Protect confidentiality & integrity of existing jobs
- No modifications for end users
- Acceptable performance overhead

Idea:
Execute only sensitive parts of Spark inside enclave
- Code that accesses/processes sensitive data

Code outside of enclave only accesses encrypted data
- Partition Spark
- Run two collaborating JVMs, inside enclave and outside of enclave
## Partitioning Spark

**Goal:** Move minimal amount of Spark code to enclave

<table>
<thead>
<tr>
<th>Outside</th>
<th>Enclave</th>
</tr>
</thead>
</table>
| **HadoopRDD**  
Provide iterator over input data partition (encrypted) |  |
| **MapPartitionsRDD**  
Execute user-provided function \( f \)  
(eg flatMap(line => { line.split(" ") }))  
(i) Serialise user-provided function \( f \)  
(ii) Send \( f \) and \( it \) to enclave JVM  
(iv) Receive result iterator \( it\_result \) |  
|  | (i) Decrypt input data  
(iv) Compute \( f(it) = it\_result \)  
(v) Encrypt result |
| **ExternalSorter**  
Execute user-provided reduce function \( g \)  
(eg reduceByKey\{case (x, y) => x + y\}) |  
|  | (iii) Decrypt input data  
(iv) Compute \( g(it2) = it2\_result \)  
(v) Encrypt result |
| **ResultTask**  
Output results |  |

Peter Pietzuch - Imperial College London
Partitioning Spark

```
HadoopRDD A
  `<outfile`
  `<infile`

MapPartitions
  RDD C
  `i`, `j`, `k`

MapPartitions
  RDD B
  `i`

ResultTask
  `k`

Outside Enclave
Tasks
A,B,C,D
i,j,k
Iterators
Iterate via shm
```
Partitioning Spark

HadoopRDD A

MapPartitions RDD C

MapPartitions RDD B

HadoopRDD A

ItProvider i

ResultTask

ItConsumer k

k'

k

j

j'

j

k'

i

Communication

ItProvider k

SgxTask C

SgxTask B

ItConsumer i

i,j,k

A,B,C,D Tasks

i,j,k Iterators

Iterate via shm

Outside

Enclave

Peter Pietzuch - Imperial College London
2. Alternatives (as well as)

- **MPC**
  - Hard to comprehend
  - High latency

- **Homomorphic Cryptography**
  - Easier to understand
  - Lower throughput

- **Differential Privacy or K-Anon or Fuzzing**
  - Easier still to understand
  - Limited number of duty cycles
3. The HAT ecosystem – Not a Cloud

== Databox in the Wild

To evolve and emerge a digital service ecosystem of organisations and people where individuals are able to acquire, use, control and exchange their own data for their own good and the good of society

Ownership Model:

http://wrap.warwick.ac.uk/108357/
The HAT ecosystem

HAT users with their HAT data in their own microservice containers

Rumpel
Personal hyper data browser (to see & act on HAT data & content)
http://rumpel.hubofallthings.net

HAT service providers
Wholesale channel
Brand owners, IoT companies, SMEs, outsourcing personal data management & exchange
http://www.hatdex.org/hatdex-for-business/

Retail Channel: direct provisioning of HATs

HAT service providers

HAT Platform providers

HATDeX
http://hatdex.org
Certified

MarketSquare
Community space to browse data & content offers, interact & chat, ‘catalogue/store’, exchange data
http://marketsquare.hubofallthings.net

Other HAT Platform providers globally that host HATs on their own infrastructure

Dataplugs
Meta & aggregate level services
Data shoppers
Content shoppers
Social data sharing
Analytics, comparisons

Marketsquare add-ons

Dataplugs
Add new data into HATs
Social media, finance, Health, government data, supermarket purchases
Sign on services

Rumpel add-ons
Native applications (data does not leave the HAT/Rumpel space)
New views
Bundles/Contexts
Email, website, file service, messaging

Direct HAT applications
Applications that use HAT schema, logic and APIs directly

HAT-ready services contributing to and requesting HAT data & content

Brand owners, IoT companies, SMEs, outsourcing personal data management & exchange
Continuing research on HAT ecosystem
Managed by WMG, University of Warwick
http://hubofallthings.com/hat-rd/

HAT Global Festival
Hackathons
Funding for startups

Open source code repository

Community development
HAT Community Foundation

A members organisation (charity) startup for the Regulation, Innovation, Growth & Representation in the HAT personal data ecosystem
The HAT ecosystem

HAT users with their HAT data in their own microservice containers

Rumpel
Personal hyper data browser (to see & act on HAT data & content)
http://rumpel.hubofallthings.net

HAT service providers
Wholesale channel
Brand owners, IoT companies, SMEs, outsourcing personal data management & exchange
http://www.hatdex.org/hatdex-for-business/

HAT Platform providers

HATDEX
http://hatdex.org

Marketsquare
Community space to browse data & content offers, interact & chat, ‘catalogue/store’, exchange data
http://marketsquare.hubofallthings.net

Other HAT Platform providers globally that host HATs on their own infrastructure

HAT-ready services contributing to and requesting HAT data & content

Ready

Direct HAT applications
Applications that use HAT schema, logic and APIs directly

Rumpel add-ons
Native applications (data does not leave the HAT/Rumpel space)
New views
Bundles/Contexts
Email, website, file service, messaging

Marketsquare add-ons
Meta & aggregate level services
Data shoppers
Content shoppers
Social data sharing
Analytics, comparisons

Dataplugs
Add new data into HATs
Social media, finance, Health, government data, supermarket purchases
Sign on services

Growing

Community Foundation

HAT Dex
http://www.hatdex.org

Continuing Research on HAT ecosystem
Managed by WMG, University of Warwick
http://hubofallthings.com/hat-rd/

Wellbeing * blockchains * content * others

Other HAT Platform providers

Dataplugs
Meta & aggregate level services

Malfunction

Retail Channel: direct provisioning of HATs

HAT service providers

Marketsquare add-ons
Community space to browse data & content offers, interact & chat, ‘catalogue/store’, exchange data
http://marketsquare.hubofallthings.net

Other HAT Platform providers globally that host HATs on their own infrastructure

Rumpel add-ons
Native applications (data does not leave the HAT/Rumpel space)
New views
Bundles/Contexts
Email, website, file service, messaging

Marketsquare add-ons
Meta & aggregate level services
Data shoppers
Content shoppers
Social data sharing
Analytics, comparisons

Dataplugs
Add new data into HATs
Social media, finance, Health, government data, supermarket purchases
Sign on services

Growing

Community Foundation

HAT Dex
http://www.hatdex.org

Continuing Research on HAT ecosystem
Managed by WMG, University of Warwick
http://hubofallthings.com/hat-rd/

Wellbeing * blockchains * content * others

Other HAT Platform providers

Dataplugs
Meta & aggregate level services

Malfunction

Retail Channel: direct provisioning of HATs

HAT service providers

Marketsquare add-ons
Community space to browse data & content offers, interact & chat, ‘catalogue/store’, exchange data
http://marketsquare.hubofallthings.net

Other HAT Platform providers globally that host HATs on their own infrastructure

Rumpel add-ons
Native applications (data does not leave the HAT/Rumpel space)
New views
Bundles/Contexts
Email, website, file service, messaging

Marketsquare add-ons
Meta & aggregate level services
Data shoppers
Content shoppers
Social data sharing
Analytics, comparisons

Dataplugs
Add new data into HATs
Social media, finance, Health, government data, supermarket purchases
Sign on services

Growing

Community Foundation

HAT Dex
http://www.hatdex.org

Continuing Research on HAT ecosystem
Managed by WMG, University of Warwick
http://hubofallthings.com/hat-rd/

Wellbeing * blockchains * content * others

Other HAT Platform providers

Dataplugs
Meta & aggregate level services
HAT Research

Increasing knowledge in the HAT personal data ecosystem
The HAT ecosystem

HAT users with their HAT data in their own microservice containers

- **Rumpel**
  - Personal hyper data browser (to see & act on HAT data & content)
  - Personal HAT address: yourname.hubofallthings.net

- **HAT service providers**
  - Wholesale channel
    - **HATDeX**
      - http://hatdex.org
      - Community space to browse data & content offers, interact & chat, ‘catalogue/store’, exchange data
    - **MarketSquare**
      - http://marketsquare.hubofallthings.net
      - Community space to browse data & content offers, interact & chat, ‘catalogue/store’, exchange data

- **HAT Platform providers**
  - Other HAT Platform providers globally that host HATs on their own infrastructure

- **HAT service providers**
  - Marketsquare add-ons
    - Meta & aggregate level services
    - Data shoppers
    - Content shoppers
    - Social data sharing
    - Analytics, comparisons
  - Dataplugs
    - Add new data into HATs
    - Social media, finance, Health, government data, supermarket purchases
    - Sign on services

- **HAT-ready services contributing to & and requesting HAT data & content**
  - **Direct HAT applications**
    - Native applications (data does not leave the HAT/Rumpel space)
    - New views
    - Bundles/Contexts
    - Email, website, file service, messaging
  - **Rumpel add-ons**
    - Social data sharing
    - Email, website, file service, messaging
  - **Marketsquare add-ons**
    - Meta & aggregate level services
    - Data shoppers
    - Content shoppers
    - Social data sharing
    - Analytics, comparisons
  - **Dataplugs**
    - Add new data into HATs
    - Social media, finance, Health, government data, supermarket purchases
    - Sign on services

- **Retail Channel: direct provisioning of HATs**
  - HAT users with their HAT data in their own microservice containers

- **Community development**
  - HAT Global Festival
  - Hackathons
  - Funding for startups

- **Open source code repository**
  - Applicaions that use HAT schema, logic and APIs directly

- **Brand owners. IoT companies, SMEs, outsourcing personal data management & exchange**
  - http://www.hatdex.org/hatdex-for-business/

- **Developed Rumpel**
  - EPSRC Funded £385k
  - Open sourced for community
HAT Data Exchange

A commercial startup managing and operating the HAT personal data ecosystem
Who Am I? & lets not speculate further 😊

Thanks to EPSRC/databox
• &Liang Wang, et al, Cambridge

Thanks to Turing/Maru
• &Peter Pietzuch, Imperial