THE VALUE OF PERFORMANCE.
NORTHROP GRUMMAN

Northrop Grumman Cybersecurity Research Consortium (NGCRC)

Spring 2014 Symposium

CYBERSECURITY RESEARCH Consortium

Crypsis: Secure Big Data Analysis in Untrusted Clouds

28 May 2014

Julian Stephen, Savvas Savvides, Russell Seidel and Patrick Eugster

Purdue University

Problem Statement

NORTHROP GRUMMAN

- Big data challenge
  - We are surrounded by massive amounts of data
  - Governments, corporations, academic institutions - all generate data
  - Need to process data and derive meaningful information
  - Need to process in parallel and provision for failures
- Potential of cloud
  - Computation infrastructure available on demand
  - Ability to adapt to changing requirements
  - Corporates opt for “in-house” clouds for data storage and analysis
- Challenges
  - How safe is it to trust a third party cloud provider?
  - How can banking, finance and insurance sectors leverage this potential?

“Big data absolutely has the potential to change the way governments, organizations, and academic institutions conduct business and make discoveries, and its likely to change how everyone lives their day-to-day lives” - Susan Hauser, corporate vice president, Microsoft’

Background

NORTHROP GRUMMAN

- Security in the cloud
  - ClusterBFT (Integrity, Availability, Isolation)
  - Crypsis (Confidentiality)

- Existing approaches
  - Communication-centric
    - Focus on messages exchanged between machines
    - Firewalls, anti-virus, etc.
  - Data-centric
    - Focus on data at rest
    - Encryption, access control, etc.
  - Computation-centric
    - Focus on computations generating correct output
    - Functional encryption
  - Solutions overlap, need to secure all three fronts

Homomorphic Encryption

NORTHROP GRUMMAN

- Holy grail: *Fully* homomorphic encryption (FHE)
  - Prohibitive overhead, getting more practical
  - Fine-print is in expressiveness
- *Partially* homomorphic encryption (PHE)
  - Allows for certain operations to be performed in encrypted form
  - E.g.,
    - Paillier [Paillier;EuroCrypt'99] ► AHE (+)
    - Unpadded RSA [Rivest et al.;CACM'78], ElGamal [ElGamal;IEEE ToIT'86] ► MHE (x)
    - DET (=), OPE (<)
- Conjecture
  - Partition programs according to attributes and use a different cryptosystem for each
  - We can use multiple PHE cryptosystems in parallel
  - Reencryption between PHE systems may be faster than FHE

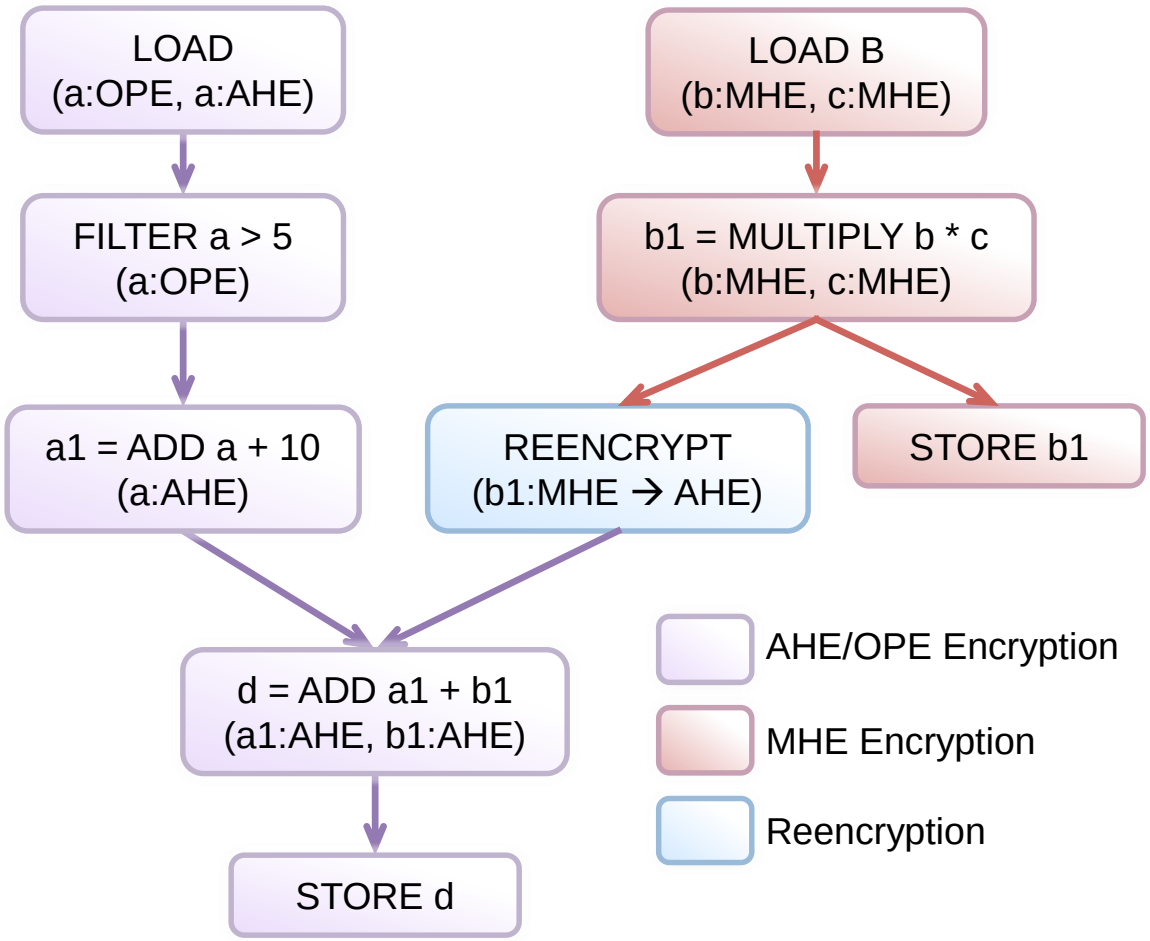
Related Work

NORTHROP GRUMMAN

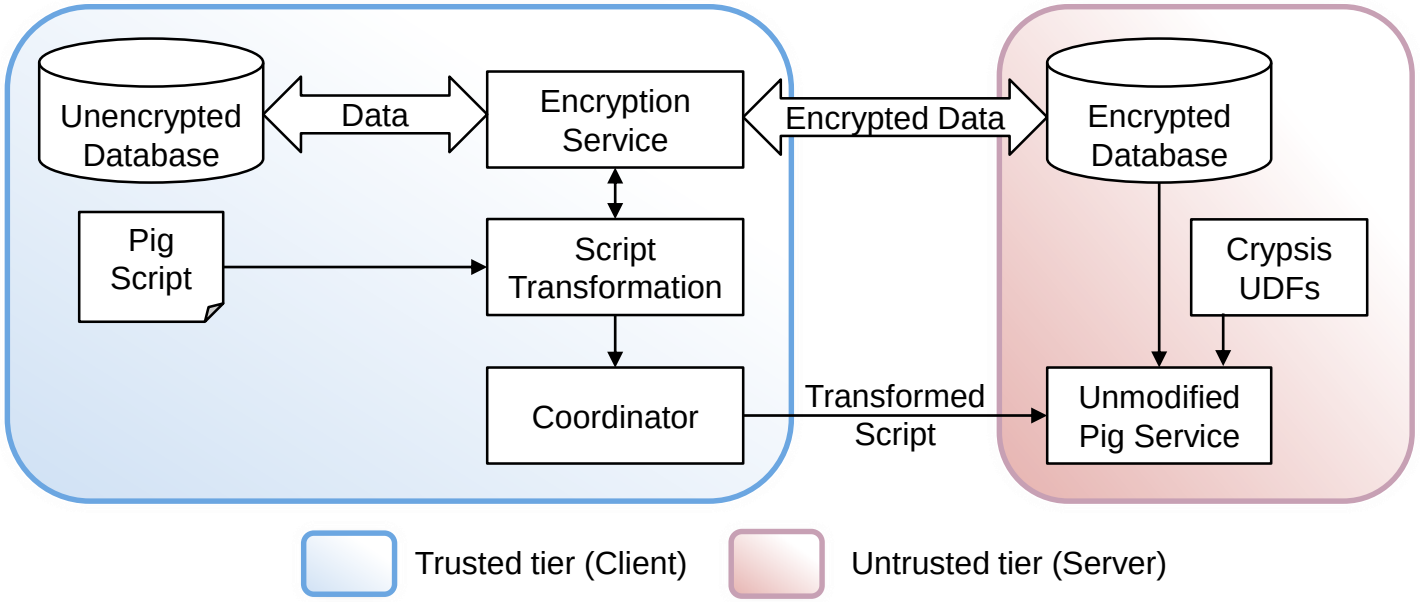
- DepSky [Bessani et al.;Eurosys'11]
  - Storage
  - Quorum-based replication with secret sharing for privacy and integrity
  - Only AHE
- CryptDB [Popa et al.;CACM'12]
  - Encrypted database for SQL (subset), No Parallelism
  - No reencryption; client-side query completion
- MrCrypt [Lesani et al.;OOPSLA'13]
  - Program analysis for individual MapReduce tasks
  - No reencryption
- Monomi
  - Uses techniques to improve performance of complex queries on encrypted data
  - Built on top of Postgres, Centralized Design

Crypsis Intuition

NORTHROP GRUMMAN



Architecture Overview



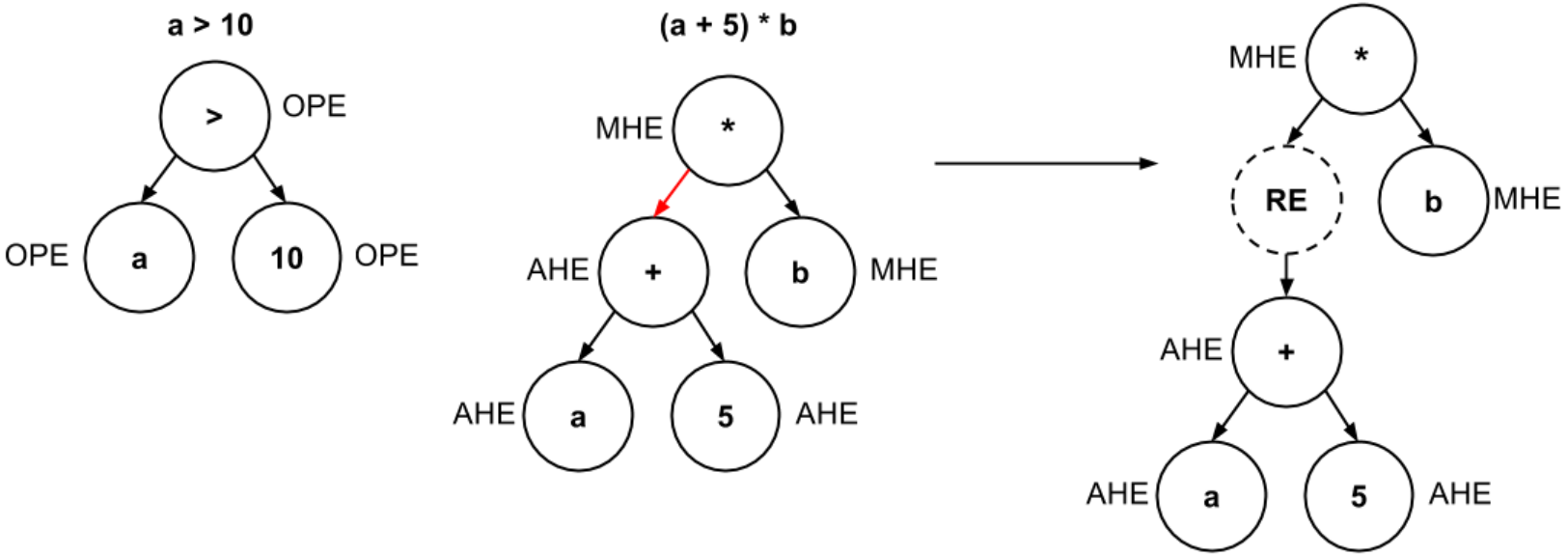
- Practical Confidentiality Preserving Big Data Analysis [J. Stephen, S. Savvides et al; hotcloud14]

Script Transformation

NORTHROP GRUMMAN

- Script analysis
  - Generate Data Flow Graph (DFG)
  - Nodes are relations (LOAD, FOREACH, etc…)
  - Edges are data flow between operators
- Generate Map of Expression Trees (MET)
  - Contains all expressions of the script
  - Keys are used to assign expressions to DFG
- Generate Set of Annotated Fields (SAF)
  - One entry for each <relation, field> of the script
  - <relation, field>, parent, available encryptions, required encryptions
  - Get available encryptions from lineage of field (parent)
  - Get required encryptions using MET

Identifying Reencryptions



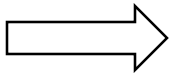
- Reencryption required when:
  - Required encryption not available
  - Incompatible operations e.g. addition followed by a multiplication
  - Reencryption is conceptual (can continue computation in client)

- 17 PigMix2 benchmarks (PigMix1 + 5)
  - Only script 8 requires reencryption (averaging)
  - 1 additional script requires same attribute available in 2 encryptions

Transformation Example

NORTHROP GRUMMAN

```
A = LOAD ‘input1’ AS
      (a0, a1);
B = LOAD ‘input2’ AS (x0);
C = FILTER A BY a0 > 10;
D = GROUP C BY a1;
E = FOREACH D GENERATE group AS
       b0, SUM(C.a0) AS b1;
F = JOIN E BY b0, B BY x0;
STORE F INTO ‘out’;
```



```
A = LOAD ‘enc_input1’ AS
      (a0_ope, a0_ah, a1_det);
B = LOAD ‘enc_input2’ AS (x0_det);
C = FILTER A BY a0_ope > OPE(10);
D = GROUP C BY a1_det;
E = FOREACH D GENERATE group AS
       b0, ENCSUM(C.a0_ah) AS b1;
F = JOIN E BY b0, B BY x0_det;
STORE F INTO ‘out’;
```

Implementation

NORTHROP GRUMMAN

- MapReduce [Dean&Ghemawat; OSDI'04]
  - Parallel execution (map and reduce functions)
  - Hadoop version 1.2.1
- Pig and Pig Latin [Gates et.al; VLDB'09]
  - Pig Latin - High level data flow language for expressing data analysis programs
  - Pig - runtime environment, generates Map Reduce programs
  - Pig version 0.11.1
- Crypsis UDFs
  - Replace operations and aggregation functions with their encrypted version
  - Allows for an unmodified Pig service
  - +, -, ~, *, ^, ==, <, ≤, >, ≥
  - Aggregation functions: SUM, MAX, MIN, DISTINCT, ORDERBY, AVG, MEDIAN, ABS

Optimizations

NORTHROP GRUMMAN

- Minimize ciphertext overhead: Packing
  - Pack multiple numbers in a single plaintext before encrypting
  - Must handle overflows
  - Up to 90% reduction of ciphertext size
- Minimize number of reencryptions: Un-encrypted values
  - Often only parts of input data need to be encrypted
  - Some homomorphic cryptosystems have a secondary homomorphic property
  - E.g. $AHE(x)^y \rightarrow AHE(x * y)$ Note: y in not encrypted

$AHE(a) * AHE(b) = AHE(a + b)$

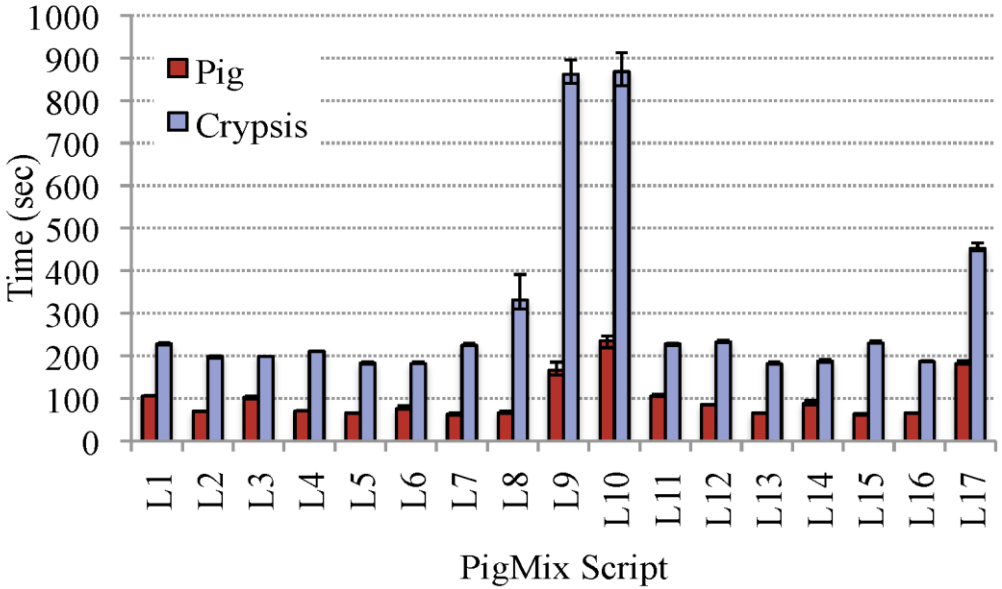
REENCRYPT$(AHE(a + b)) = MHE(a + b)$

$MHE(a + b) * MHE(c) = MHE((a + b) * c)$

$AHE(a) * AHE(b) = AHE(a + b)$

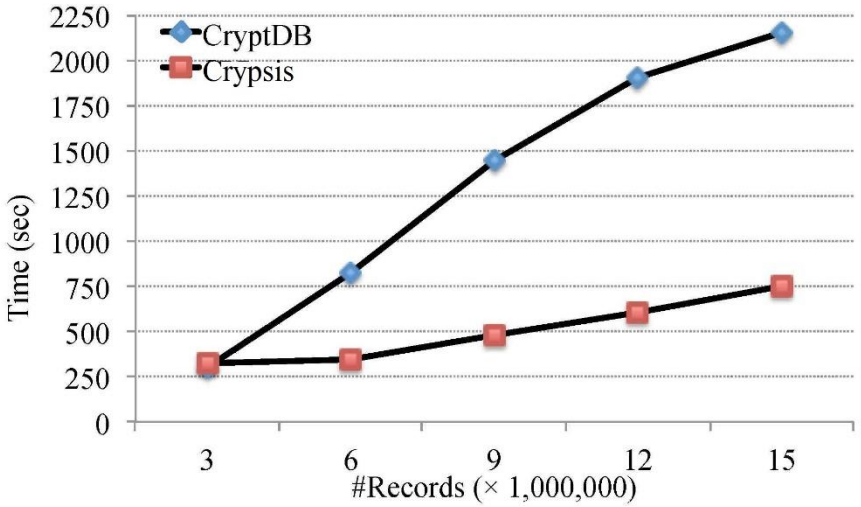
$AHE(a)^c = AHE((a + b) * c)$

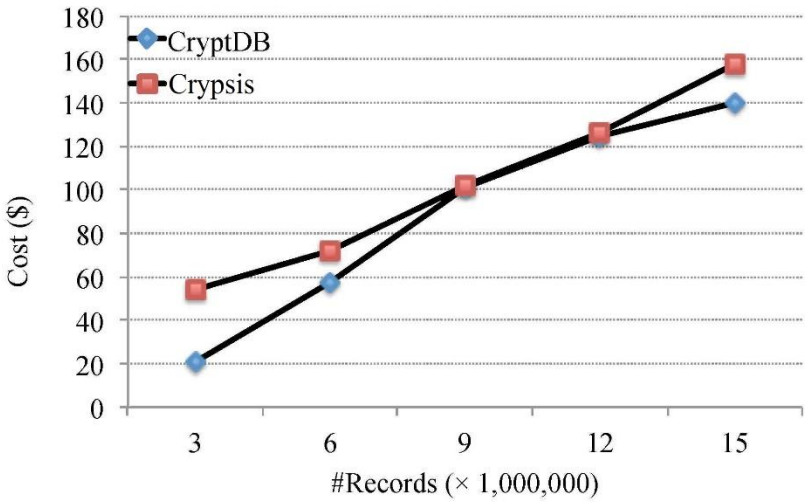
Evaluation (PigMix)



- 11 ec2 c3.large instances (2 vCPUs, 3.75GB ram)
- 5GB of data (over 3 million rows)
- An average of 3x overhead in terms of latency
- FHE can exhibit several 100 times overhead

Evaluation (CryptDB cimparison)





- 3 m3.medium instances
- ~ 3x faster for 15 million records
- Similar overall cost

Limitations

NORTHROP GRUMMAN

- Iterations – Recursion
  - Pig Latin is a query based language
  - No support for iterations or recursion
  - Iterations over encrypted data can be very expensive (reencryptions)
  - Use of UDFs for more complex operations
- User Defined Functions (UDFs)
  - Many big data analysis languages propose UDFs vs. built-in/pre-defined operators and functions
  - Black-boxes from perspective of program analysis (Cannot support transformation)
  - Can analyze byte-code on Java UDFs

Opportunities for Reducing Reencryption

NORTHROP GRUMMAN

- Statement re-ordering, e.g.,

```
a1 < 0  (OPE)
...     (CS1 != OPE)
a1 > 20 (OPE)
```

```
a1 < 0  (OPE)
a1 > 20 (OPE)
...     (CS1 != OPE)
```

- Computation re-writing, e.g.,

```
t1 = t2 * (t3 + t4)
...
t5 = t1 + t6
```

```
t1 = t2 * t3
t7 = t2 * t4
...
t5 = t1 + t7 + t6
```

- Condition re-writing, e.g.,

```
(t1 + t2 > 0)
...
```

```
(t1 > 0 AND t2 > 0)
...
```

Demonstration

Conclusions and Future Work

NORTHROP GRUMMAN

- More fine-grained encryption system
  - Augment computational capabilities not involving reencryptions.
  - Reduce the cost of reencryption operations
- Modified Pig Service
  - cost of performing the computation is visible to the programmer at compilation time
- Integration with ClusterBFT
  - Secure Cloud-based Data Analysis with ClusterBFT [J. Stephen et al; Middleware'13]
  - In collaboration with Prof. Cristina Nita-Rotaru
  - Can serve as a lower level infrastructure on top of which Crypsis is built.
- Collaboration with Prof. Bharat Bhargava
  - Confidential Reputation management module (Secure trust calculations)

THE VALUE OF PERFORMANCE.

NORTHROP GRUMMAN