Have your cake, eat it, and hide the ingredients

Mayank, Larry, Merced, Azer, Orran, others?

January 23, 2017

Abstract

Data is important, sharing data is important, collaboration is important but sois protecting the data even from collaborators. Dataverse is an existing technology that provides curated datasets, tagged with both functional and legal access policies. The mass open cloud (MOC) is an existing cloud technology that permits multiple tenants to perform elastic computations on their data in an isolated, trust no-one fashion, through advanced federated login, hardware as a service facility with novel Hardware Isolation Layer and Bare Metal Imaging. Multi-party computation (MPC) is an existing technology that provides two or more parties to perform a computation on data in such a way that no party can learn something about the data that is not inferred from the output.

Combing these three technologies yields an unexpected consequence – the ability of "paranoid" parties to collaborate. One or more of the collaborating parties may have their own dataset whose contents they wish to remain confidential. They may not want to share the data even with their collaborators. However, they may want to perfroma a computation over one or more of the data sets, e.g. they want to share their cake. The collaborators may all see the output of the computation, e.g. "eat their cake". Different parts of the computation are performed by different parties in such a way, that no party can learn anything about the data, e.g. "hide the ingredients".

Dataverse provides the data set infrastructure, MPC provides the computational method, and MOC provdes the low latency communication between isolated hardware that is necessary for MPC to be practical.

1 Introduction

Why this is a good idea. – research and innovation is being held back by concerns over privacy, confidentiality, credit, isolation between goals

Why this is the right time. – explosion of data, benefit of data analysis, need for collaboration

Why we are the right folks. – build MOC, Dataverse, and practical MPC

The chances of success. – very high. Already have many users of DataVerse and MOC.

1.1 Motivation

case of average salaries without reveal each parties data and no one wants responsibility.

NIH requirement to make funded data pubic after time limit. Before time limit, collaboration is still possible

Better to be safe than sorry. Once data is "out of the box" cannot put it back in.

2 Preliminaries

Combining the three technologies gives something new, powerful, and unexpected.

2.1 MOC and MPC

Unlike most of the public cloud infrastructure, in the MOC, there is not trusted provider. Each tenant gets a physical server. The tenant may choose to install a virtual machine infrastructure on the server. The cloud provider, i.e. the MOC, makes it easy for the tenant to install a hypervisor and run VMs, but the MOC does not have control or access to the anything running on the server. The MOC also provides secure mechanism to measure and attest to the tenant that the firmware, operating system, and hypervisor have not been compromised. Once up and running, the environment can appear to the tenant just like a public (Closed) cloud, with one big difference: the tenant does not have to trust the provider and is isolated from all other tenants.

In addition, this isolation between tenants is like the isolation provided when tenants execute in their own private data centers. However, since the tenants are co-located within the same datacenter, the network latency between the tenants is very low. This is a crucial requirement for practical MPC. Under MPC, the parties frequently communication and the computation is usually are waiting for partial results from the other parties. The difference in latency speed between being co-located and geographically separate may be two or three orders of magnitude. An hour co-located computation may take 100 or 1,000 hours in separate data center computation, making MPC impractical. Note that because of the relatively large number of rounds, each of a small amount of data being communication, places a premium on latency. The general internet is optimized for bandwidth and not low latency.

The MOC, being an open market place, provides different types of network switches. For MPC the parties can make use of low latency switches (PODS reference) that are optimized for the MPC needs.

2.2 MPC and Cloud Dataverse/Datatags

-> Add computing to cloud dataverse

-> Allows secure creation of new data set

<- For classification storage based on synthesizing data

Input: where the data live

Output: where the access control decisions / attribution is done and provenance logging is performed.

2.3 MOC and Cloud Dataverse

Single shard location for lots of data (Analogy: interlibrary loan)

Policy: who can read data, why, how Orchestration

2.4 The synergistic payoff

People have been collaborating for years and there are many ways to collaborate with big data sets. There are many levels of security needs, levels of paranoa, potential legal implications, and the like. Often, collaboration requires some amount of "re-inventing the wheel" and access to experts.

The synergistic payoff of our proposed combining is two-fold. It eliminates teh need to re-invent the wheel. Second, it forsters "separation of concerns" so that each concern can be addressed fully and in depth.

Specifically, it provides the amortization of:

- Security programmers
- Lawyers
- IT Staff
- Classification expert and policy-agnostic programming

3 Details

4 Conclusion