The Sea of Stuff: a step towards the grand unified theory of data

Simone Ivan Conte – Adobe Systems (conte@adobe.com)
Alan Dearle – University of St Andrews (alan.dearle@st-andrews.ac.uk)
Graham Kirby – University of St Andrews (graham.kirby@st-andrews.ac.uk)

Status: This research is the outcome of my PhD.

Managing data is one of the main challenges in distributed systems and computer science in general. How we understand and interact with data has changed over the last few decades. Cloud storage, and web applications in general, have extended the types of interactions associated with files and folders from simple CRUD operations to more complex and diverse ones, such as sharing, synchronising, signing, or versioning. As a result, understanding data and how it should be managed effectively in a distributed heterogeneous system has become hard.

Cloud storage computing has revolutionised the way we build applications and services and think about data. Dropbox, AWS S3, or even Google Photos, to name a few, enable their users to operate on data as one would normally do on a local file system, but with the advantage that data is automatically versioned, encrypted, replicated and synchronised against a storage service or other user devices. Services and applications are modelled ad-hoc to provide such properties, but data is not. As a result, data is still tightly coupled to locations and services, unaware of how it has evolved over time, how it is being protected and what types of policies regulate its lifecycle.

Our hypothesis is that a data model based on immutable, self-describing, re-computable, and content-addressable entities can enable the construction of distributed data management systems where (I) data is accessible regardless of its location or the location of users; (II) metadata is accessible and computable independently of the data it describes; (III) access to data and metadata can be controlled over globally distributed users; (IV) arbitrary levels of resilience can be enforced; and (V) both data and metadata can be versioned.

This hypothesis is tested by the Sea of Stuff (SOS), a generic model describing data and computation over a distributed environment through immutable, discoverable manifests. The vision of the SOS is to provide a direction towards a grand unified theory of data. The SOS model captures the following fundamental aspects of data management: data itself, how data is composable, versioning, metadata, users and their role in owning and protecting data, and finally how data should be regulated across distributed nodes in an automatic manner.

Comparative and experimental evaluations show that the SOS model is a viable solution for managing data over a globally distributed system. Similarly, we show that computation over data (i.e. policies) is feasible over a distributed system of moderate size. The journey to a grand unified theory of data, however, is still long and many research questions remain unanswered. For example, adaptive data access/placement strategies based on the reputation of nodes or users are still expensive and therefore used only occasionally. Similarly, defining expressive and conflict-free policies over distributed and shared data is challenging.

- Dropbox – https://dropbox.com – Last accessed on 07/02/2019
- AWS S3 – https://aws.amazon.com – Last accessed on 07/02/2019
- Google Photos – https://photos.google.com – Last accessed on 07/02/2019