

Office of Technology Strategies (TS), Architecture, Strategy & Design (ASD)

A VA Executive's Guide to Hybrid Data Infrastructure

INTRODUCTION

This TS Note discusses Hybrid Data Infrastructure, which is a new type of data setup conceived to deal with data-intensive domains, such as the sciences and medicine. Hybrid Databases are systems that support both in-memory and on-disk storage. They are typically deployed to retain the high performance and small database footprint advantages of in-memory databases while leveraging the stability and possible cost savings of on-disk databases. This note will give an overview of Hybrid Database technologies, discuss why Hybrid Data Infrastructure is on the rise, and look into an example of this new and upcoming technology for highly scalable solutions.

OVERVIEW

Hybrid Data is a relatively new concept. It is an innovative approach based on the idea that several technologies, including private and public Cloud, can be combined to provide flexible access and usage of data and data management capabilities. In these domain spaces, datasets come in several forms, from big experiments to cross-lab, single lab, or even individual observations. The management and processing of such datasets is beyond the capacity of traditional technological approaches. Overall, the goal of Hybrid Data technology is to enable a data-management-capability delivery model in which computing, storage, data, and software are made available by the infrastructure-as-a-Service.

Established technological platforms are no longer able to address data and processing requirements of an emerging data-intensive model, and modern computer platforms are not capable of addressing global, changeable, and networked needs of the science communities that produce huge quantities and varieties of data.

Hybrid Data Infrastructure uses several technologies to create the management and usage capabilities required to implement the data-enabled scientific paradigm.

GRID AND CLOUD VS. HYBRID

Recent methods like Cloud Computing can only partially satisfy these needs, and Grid Computing was initially only created as a technological platform to overcome limited numbers of single labs by sharing and re-using computational and storage resources across labs. While this is a valid solution to some scientific domains, Grid Computing does not handle variety well and only supports a limited set of data types. In comparison, Cloud Computing provides an elastic usage of resources maintained by third-party providers. The technology is based on the notion that the management of hardware and middleware can be centralized, while applications remain in control of consumers. Though this reduces application maintenance and costs, it is not suitable to manage integration of resources deployed and maintained by diverse organizations.

Hybrid Data Infrastructure is the more effective solution for managing new types of scientific data. It understands that several technologies, including Grid and Cloud, can be integrated to provide flexible access and usage of data and data management capabilities.

THE GCUBE SYSTEM

gCube is one example of a technology using the Hybrid Database Infrastructure. It was initially created to manage distributed computing infrastructures, and evolved to operate large-scale Hybrid Database Infrastructure enabling a data-management-capability delivery model in which computing, storage, data, and software are made accessible.

Technology Strategies

Defining OI&T's
"To Be"
Technology
Vision



The TS office within OI&T's Architecture, Strategy & Design (ASD) interacts not only with the ASD pillar offices, but also with multiple stakeholders within OI&T and with strategic offices across the enterprise. TS works closely with IT and business owners to capture business rules and provide technical guidance as it relates to Data Sharing across the enterprise, specifically for interagency

gCube is more than just a software integration platform. It is also equipped with software frameworks for data management and workflow execution. This offers traditional data management facilities in an innovative way by taking advantage of the array accessible technologies.

These software frameworks can be used to implement different policies ranging from privacy through encryption and secure access control to the promotion of data sharing, while guaranteeing attribution. The infrastructure enabled by gCube is now being used to serve scientists operating in different domains (i.e. biologists generating model-based large-scale predictions of natural occurrences of species and statisticians managing and integrating statistical data).

THE RISE OF HYBRID DATA

Hybrid architectures address the various realities of big data environments and support the need to incorporate both established and new database approaches into one architecture. Concerning Hybrid architecture, each big data platform is fit for the purpose of the role which it is best suited. These roles can include data acquisition, collection, movement, transformation, cleansing,

A VA Executive's Guide to Hybrid Data Infrastructure

Continued from Page 1

staging, modeling, governance, access, delivery, archiving, and more.

Hybrid Data Infrastructure is the future of big data because users are realizing that no single type of platform is always best for all requirements. The inevitable trend is headed toward hybrid environments that address the following big data requirements:

- Extreme scalability and speed: hybrid will support scale-out parallel processing, optimized appliances, storage, dynamic query optimization, and mixed workload management
- Extreme agility and elasticity: hybrid will persist data in diverse physical and logical formats across virtualized cloud of unified memory and disk that can be flexibly scared up and out at a moment's notice
- Extreme affordability and manageability: hybrid will incorporate flexible packaging and pricing, including license software, modular appliances, and cloud approaches.

The Hybrid Data environment will continue with centralized and "hub-and-spoke" topologies toward the new cloud-oriented and federated architectures. This platform is evolving away from the single master diagram and more toward database virtualization behind a semantic concept layer. The biggest hope is that it will provide big data professionals and users with logically unified access, modeling, deployment, optimization, and management of this diverse resource.

HOW HAS VA UTILIZED HYBRID DATA INFRASTRUCTURE

While Hybrid Data usage is still being developed and learned, VA does have one important tool that uses the new technology. This tool is an app called My HealtheVet. The VA announced the feature back in 2011 as an online resource that gives active duty and retired Servicemembers access to their electronic health records (EHRs). The application utilizes Hybrid Data by pulling from both the VA's Health Information Systems and Technology Architecture (Vista) and a NoSQL data store. The primary purpose of its NoSQL data store is that My HealtheVet stores patient-generated data (PGD) created by users. Veterans and beneficiaries using the app can only create and edit PGD in that NoSQL data store. It is arguably the most advanced EHR system that exists and won the 2009 TEPR (Towards the Electronic Patient Record) Award for Personal Health Records from the Medical Records Institute. It has more subscribers than any other commercial EHR deployment and serves as a model by many vendors for their own products.

HYBRID DATA ACCESS IN THE PROPOSED VA ENTERPRISE ARCHITECTURE

Hybrid data access capabilities are at the core of the proposed data layer for the VA Enterprise Architecture (VA EA) described in the Hybrid Data Access (HDA) design pattern. The proposed data layer contains a variety of relational, NoSQL, and other data stores which are encapsulated by and accessed through an enterprise Create, Read, Update, Delete (eCRUD) service. eCRUD itself is accessed through VA authoritative information services, as shown in the diagram below.

If you have any questions about hybrid data infrastructure, don't hesitate to ask TS (askTS@va.gov) for assistance or more information.

Check out earlier TS Note editions [here](http://www.techstrategies.oit.va.gov/docs_ctsnotes.asp). (http://www.techstrategies.oit.va.gov/docs_ctsnotes.asp).

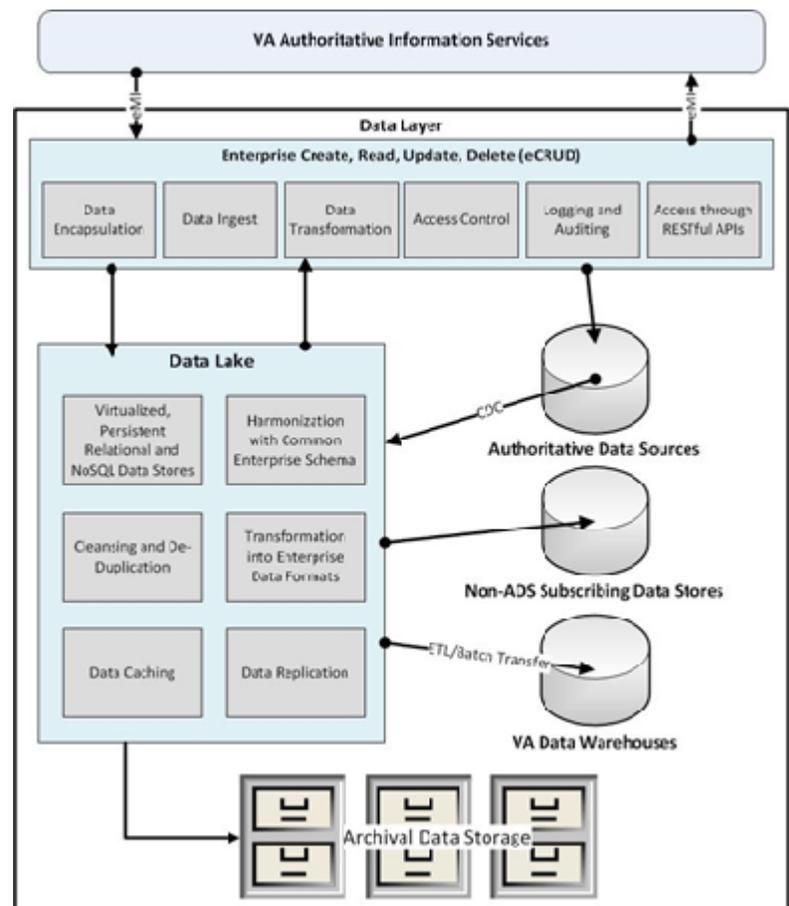


Figure 1: Hybrid Data in VA's SOA Architecture