DATA CENTER

Making Sense of your Data

The Intelligent Data Network: Proposal for Engineering the Next Generation of Distributed Data Modeling, Analysis and Prediction

David L. Brock

The Data Center, Massachusetts Institute of Technology, Building 35, Room 234, Cambridge, MA 02139-4307, USA

ABSTRACT

Computers today are faster, memory cheaper and bandwidth plentiful, yet the tasks we perform on our machines – email, documentation and data storage – are nearly the same as they were a decade ago. Things are about to change. Systems are now being created that allow computers to directly sense and interact with the physical world. What they lack is the ability to understand it. We propose a new infrastructure that enables computers to share models and automatically assemble these into simulations in order to better understand, manage, analyze, predict and plan the physical world.

ABOUT THE AUTHOR

David L. Brock is Principal Research Scientist at the Massachusetts Institute of Technology, and co-founder and a Director at the Auto-ID Center (now EPCGlobal, Inc. and Auto-ID Laboratories). The Center was an international research consortium formed as a partnership among more than 100 global companies and five leading research universities. David is also Assistant Research Professor of Surgery at Tufts University Medical School and Founder and Chief Technology Officer of endoVia Medical, Inc., a manufacturer of computer controlled medical devices. Dr. Brock holds bachelors' degrees in theoretical mathematics and mechanical engineering, as well as master and Ph.D. Degrees, from MIT. David can be reached at dlb@mit.edu

1. INTRODUCTION

The Data Project at the Massachusetts Institute of Technology is a new initiative aimed at creating the technologies and standards for widespread distributed modeling and simulation. The Project will research and develop the technology, standards and infrastructure that will enable globally shared models and distributed simulation components. The technologies and standards created by the Project will be open and freely distributed. This paper presents the vision and goals of the Data Project – and lays the foundation for a new Intelligent Data Network.

2. VISION

We are on the verge of a revolution that will transform industry, commerce and society. Computers today primarily store, manipulate and transmit data to people. Unless there is direct human interaction, computers essentially do nothing.

Yet computers have far greater capability. Computers can store and analyze vast quantities of information and share these results with other computers throughout the world. Computers can operate independently or collectively – with or without human interaction. And new initiatives, such as the Auto-ID Center, will enable computers to have instant access to real-world information (Brock 2000, 2001; Sarma 2000).

The failure to take full advantage of the computer's potential lies not in the hardware or communications technologies, but in lack of languages and standards that enable systems to share data and algorithms across multiple applications and domains.

Data analysis and modeling packages exist today, but these are typically vertically integrated, monolithic applications, with limited ability to extend or modify the base program. What we propose is a fundamentally new approach in which both fine and coarse grain modeling components can be created, published and integrated across the network automatically into on-going simulations.

We envision this new infrastructure will fundamentally transform the way we use computers and the network. These open standards will enable the automatic creation of both large and small-scale simulations for the purposes of data analysis and resource planning within a real physical environment.

3. BACKGROUND

Nearly every human endeavor uses models - implicit or explicit cognitive representations of physical systems and human behavior. From celestial motion to human emotion, we use mental representations to predict future behavior and evaluate past action.

Models are used universally in nearly every aspect of human life - politics, economics, business, commerce, manufacturing, maintenance, medicine, art and music. Psychologists and anthropologists use models to understand individual and collective human behavior. Engineers use models to predict, design and analyze physical structures and systems.

Scientists use models to represent the basic nature of the universe. Businesses use models to optimize their products and services for maximum return. In fact, the ability to model the world – that is to predict future action based on experience – could very well define the nature of intelligence.

As ubiquitous as models are, they are, for the most part, isolated from one another. In other words, a model from one domain, such as weather forecasting, does not interact with another, such as purchasing trends and behavior.

The reason for this is obvious. Until very recently humans were the only ones who built, used and shared models. Our limited cognitive ability naturally restricted the number and diversity of models we could accommodate.

Computers, on the other hand, have the ability to execute and communicate models with vast numbers of other computers. With their ever increasing processing power, data storage and networking bandwidth, the computer grid is poised to revolutionize our ability to understand and manage the physical world. And the Internet with its standards and languages provides the backbone for communication, but does not provide the mechanism for describing and integrating diverse analytic models.

What if we could harness the power of multiple individual models into larger aggregates? What if we could make predictions based on not a few, but millions of diverse facts and functions? The result would be an unprecedented increase in productivity through the optimal use of resources. We could dramatically reduce the cost of goods and services through the elimination of production inefficiencies.

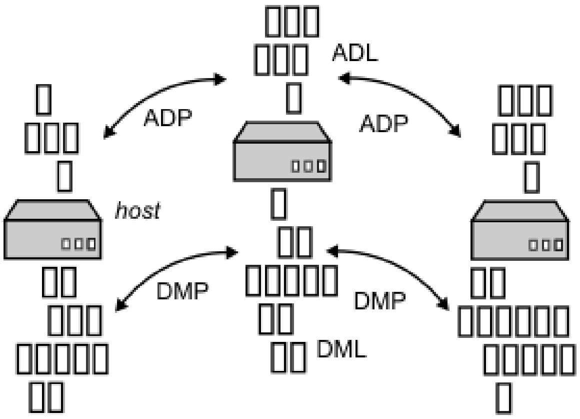
Not only in terms of economic efficiency, but also in terms of human lives. A global simulation network could, for example, analyze patterns of illness and predict epidemics long before humans recognized them. This is only the beginning. There are vast numbers other applications across the entire spectrum of human endeavor.

4. SYSTEM OVERVIEW

The fundamental idea is to create a family of standards that enable diverse models to be created separately and integrated automatically into an executing synthetic environment. In this way, developers can craft and refine their models within their particular areas of expertise and be assured that the resulting models will interoperate in a shared environment. We believe it is possible – with sufficient care in the definition – to create such a language that is both expressive in its description yet limited in its breadth to ensure compatibility.

We propose a family of standards that consist of four members – (1) Data Modeling Language (DML), a language for describing a model, (2) Data Modeling Protocol (DMP), a protocol for communicating the states of one simulation to another, (3) Automated Decision Language (ADL), a language for describing decision making systems (that is, algorithms that use the output of a simulation to provide feedback control to the physical world) and (4) Automated Decision Protocol (ADP), a protocol that allows these decision making systems to communicate with one another, as shown in Figure 1.

FIGURE 1. The proposed distributed simulation architecture is composed of four fundamental components: the Data Model Language (DML) an open standard for model components, Data Model Protocol (DMP) a communication standard between simulation hosts, Automated Decision Language (ADL) a language for decision systems and Automated Decision Protocol (ADP) a communication protocol between networks of decision systems.



The goal is to create *synthetic environments* that receive information from the physical world (for example through Auto-ID technology or *a-priori* databases) and produce inferences, interpretations and predictions about the physical states of an environment.

These interpolated or extrapolated state data are essential for any automated decision system. In other words, the estimated environmental states support networks of decision-making algorithms so that they may make informed decisions and efficient plans. The output of the simulation is essentially input to any automated control, monitoring, management and planning system.

The Data Modeling Language (DML) is intended to be a semantic for describing modular, interoperable simulation components. Models written in DML should automatically assemble into executable simulation environments. Consideration must be given to model type, description, breadth, resolution, ease of use, metrics, compensation, discover mechanisms and de/aggregation methods – just to name a few (Gershenfeld 1999).

We assume any simulation executes across multiple, heterogeneous platforms – that is a computational grid (Globus 2000). The semantic that describes the communication between simulation hosts is the Data Modeling Protocol (DMP). The DMP describes incremental results, as well as temporal and spatial subsets of an on-going simulation. The design of this protocol must consider variable update rates, bandwidth, latency, data aggregation and resolution.

A faithful reproduction of reality would be considered a successful simulation, but a central goal of this initiative is to provide a framework for intelligent decisions. Humans will make most of these decisions, but increasingly many of these will be augmented by synthetic decision systems. The Automated Decision Language (ADL) is a specification for describing these decision-making elements. We envision these elements will exist within networks of many – perhaps millions – of other decision-making elements.

If we succeed in creating such networks, it is likely the relation between decision-making elements will be complex; that is, hierarchical organizations of specialized components dynamically creating and readjusting network topology and subordinates to match the needs of a particular task. In any case, we will need some standardized protocol to enable disparate elements to communicate with one another in a common language.

The Automated Decision Protocol (ADP) is intended for just such a purpose. Using the ADP, decision-making elements can locate, communicate and coordinate with one another, even though individuals may exist on different hosts and in different organizations.

The combination of these elements – DML, DMP, ADL and ADP – represents the foundation for building general-purpose synthetic environments. The idea is that synthetic environments can be constructed automatically and then modified in real-time in response to the needs of a given task.

5. APPLICATIONS

The applications for such an architecture are manifold – impacting nearly every aspect of industry and commerce.

Businesses can optimize their production and sales based on predicted use of their products. Logistics, transportation and routing could be modified in real-time to avoid anticipated weather and traffic problems (Simchi-Levi 2002).

Hospitals could monitor and predict patient health based on real-time biometrics. Assisted living facilities and home care services could adjust medication in response to expected activity and individual metabolism.

Automobiles could dynamically adjust power, transmission, suspension and braking given driving and road conditions. Trans-metropolitan traffic signal optimization could drastically reduce delays and improve network efficiency.

Aerospace design and flight control systems are almost entirely composed of interdependent simulation models. Air transportation, routing and loading, while efficient, are often constrained along predefined pathways – allowing little optimization and individual autonomy. Path optimal "free flight" would result in a more efficient and reliable transport, increasing air capacity and minimizing delays.

Agriculture and live stock management could use a wide range of diverse data to regulate day-to-day operations such as feeding and harvest. Pesticides, fertilizer and feed could be dispensed in complex patterns – optimized for individual efficiency.

The entertainment industry, particularly electronic games and motion picture visual effects, relay, to an ever-greater degree, on complex physical models and engaging character behavior. These industries could not only benefit from an open modeling environment, but could also contribute to the technologies and modeling components across a broad range of applications. Furthermore, their ability to produce compelling visuals will help communicate abstract data sets and predicted physical environments for many diverse applications.

Environmental impact studies and public policy are dictated to a large degree by physical models and sensory data. A shared, open standard for simulation components could allow these environmental projections to be validated using multiple independent models. Furthermore, the propagation of hazardous material, the dispersion of chemical agents and the flow of recycled material could be anticipated and controlled to an even greater level with accurate analytic models

The financial services industry including securities, insurance, banking and housing are regulated almost entirely through analytic models and data projections. An open modeling infrastructure would allow economic models to be exchanged and enhanced in real-time to allow far greater precision in financial projection and economic efficiency.

Legal services, from corporate law to criminal defense use models to form their language and plead their case. These models are created implicitly or on an ad-hoc basis according to the needs of a particular argument. An interoperable modeling environment, however, could allow the legal profession to share the physical and human behavioral models that were created by other industries.

Engineering and the sciences classically use models in every aspect of their work. Clearly, the ability to create and share models in an open environment would have tremendous benefit in advancing these fields.

6. CONCLUSION

The prospect of sharing – through standard languages and protocols – the collective efforts of data modelers and system planners throughout the world is very enticing. It has the potential to revolutionize nearly every aspect of human endeavor, as well as provide unprecedented benefit and savings across the industry and commerce. Yet the challenges and difficulties are extraordinary – from theoretic achievability to practical implementation. Still the potential rewards make the effort well worth pursuing, and may ultimately result in a true *Intelligent Data Network*.

REFERENCES

Brock, D. (2000) "Intelligent Infrastructure – A Method for Networking Physical Objects," *MIT Smart World Conference*, April 2000.

Sarma, S., Brock, D. and Ashton, K. (2000) "The Networked Physical World – Proposal for Engineering the Next Generation of Computing, Commerce and Automatic Identification," *The MIT Auto-ID Center*, Cambridge, MA, MIT-AUTOID-WH-001, October 2000.

Brock, D. (2001) "The Electronic Product Code (EPC) – A Naming Scheme for Physical Objects," *The MIT Auto-ID Center*, Cambridge, MA, MIT-AUTOID-WH-002, January 2001.

Gershenfeld, N., (1999) *The Nature of Mathematical Modeling*, Cambridge University Press, Cambridge, United Kingdom.

The Globus Project (2000) (*http://www.globus.org*).

Simchi-Levi, David, Philip Kaminsky and Edith Simchi-Levi (2002), *Designing and Managing the Supply Chain*, McGraw-Hill/Irwin.

NOTES