The IBM Research Accelerated Discovery Lab: Objectives and Experience

(Extended Abstract of the Keynote Talk)

© David A. Pease
IBM Almaden Research Center, San Jose, California
pease@us.ibm.com

Abstract

The IBM Research Accelerated Discovery Lab is a unique, collaborative environment specifically designed to facilitate complex analytic projects by tackling the challenges of data-intensive scientific discovery. The environment provides access to diverse data sources, unique research capabilities for analytics such as domain models, text analytics and natural language processing capabilities derived from Watson, a powerful hardware and software infrastructure, and broad domain expertise including biology, medicine, finance, weather modeling, mathematics, computer science and information technology. This combination reduces time to insight resulting in scientific impact – ahead of the traditional pace of discovery. The lab collaborates with domain experts in areas such as healthcare analytics and genomics, and both the lab researchers and the domain experts get mutual benefits from the joint problem solving. Besides discussion of the basic ideas and experience of the design and formation of the Lab, platform and tools selection and evolving with time, we cover scenarios of interaction of the Lab with clients during problem definition and solving, how teams of the Lab are composed for specific problem solving, etc. Successful applications of the Lab infrastructure, from fields such as medical research, food safety, social media analytics and predictive equipment maintenance, are also included.

Extended Abstract

Today, businesses and scientists alike struggle to get to the value in their data. Their challenges include finding and gaining access to the data they need, "wrangling" the data into a form they can use, and setting up the systems and software to be used - all before even tackling the analysis. With no coordination, multiple groups may re-do the heavy lifting to ready the data for use, or struggle to figure out what data is already available. Further, the skills required to get from raw data to insight span a broad range from systems to data management, optimization, statistics, algorithms, story-telling and visualization. Rarely can you find such multi-disciplinary expertise in one team - it is typically scattered across multiple business units or departments.

The IBM Research Accelerated Discovery Lab (ADLab) is a unique, collaborative environment specifically designed to facilitate complex analytic projects by tackling the challenges of data-intensive scientific discovery. The environment provides access to diverse data sources, unique research capabilities for analytics such as domain models, text analytics and natural language processing capabilities derived from Watson, a powerful hardware and software infrastructure, and access to broad domain expertise including biology, medicine, finance, weather modeling, mathematics, computer science, and information technology. This combination reduces the time to insight, resulting in scientific impact ahead of the traditional pace of discovery. The lab collaborates with domain experts in areas as diverse as healthcare analytics, agriculture, sensors for the Internet of Things, and genomics, and both the lab researchers and the domain experts get mutual benefits from the joint problem solving.

In this talk we first discuss the ideas and goals behind the formation of the lab, including an overview of the physcial and software environments we set out to create in support of accelerating the pace of discovery. We describe the basic structure of the lab, including the physcial collaboration space, the Discovery Cloud data center environment, the custom-built collaboration and discovery software environment (LabBook), our Data Lake, and more.

As an example, we spend some time discussing our Discovery Cloud environment. While not the major goal nor interest of the ADLab, this environment is a necessary and important part of the discovery environment. For one thing, the type and volume of analytics targeted by the ADLab can require significant computing and storage resources, beyond that of many of our partners; additionally, because many of our projects include collaboration with organizations outside of IBM, it is necessary to have a collaborative data center environment that can be accessed by both IBM and non-IBM researchers while adhering to IBM's stringent security compliance and monitoring requirements. At one time the ADLab managed the largest externally-accessible data center (referred to internally as a "Yellow Zone") in all of IBM, with 480 compute servers (a total of 5,760 cores), about 70 TB of main memory, more than 11 PB of

Proceedings of the XVII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2015), Obninsk, Russia, October 13 - 16, 2015

local disk storage, and another 2.6 PB of shared SAN storage. Services provided by the ADLab Discovery Cloud include system configuration and provisioning, OS installation and configuration, user id management for both internal and external users, VPN administration, systems technical support, resource monitoring, storage configuration, provisioning, and monitoring, and security management, monitoring, and compliance. It is worth noting that a significant amount of the ADLab's resources have gone into building, management, and administering our Discovery Cloud environment in support of the Accelereated Discovery services.

We continue with a deeper look at the other important components of our environment, starting with LabBook; LabBook is designed to be the principal collaboration tool used by both the ADLab researchers and its project partners. It incorporates information about projects, people, applications, various types of data sets, and experimental results in a simple notebook metaphor. It captures metadata about everything from data set provenance and use to the social interactions of its users, and uses this information to assist the researchers in tasks from data selection to automatic conversion and visualization.

LabBook also provides the primary interface to the Lab's Data Lake, which contains data sets from such varied sources as ESRI (geo-spatial information), various government bureaus, Twitter, worldwide patent databases, and many others. Data may be added to the Data Lake through various means: we began with what we considered to be a useful set of both public and private (for-cost) data sources, and continue to grow that set; we have a formal process for requesting additions to the current set, which of course includes questions of cost, ownership, appropriateness, etc.; finally, internal and external project requirements can drive the inclusion of new data. Not all data is available to all ADLab users; some may be obtained with specific project use, legal, or other restrictions attached. (Confidential research partner data is not considered part of the Data Lake, and is managed separately.) We discuss some of the challenges in collecting and curating such diverse data sets.

We examine some of the scenarios of interaction of the Lab with clients during problem definition and solving, as well as how teams of the Lab are composed for specific problem solving. This leads us to a discussion of some successful applications of the Lab infrastructure, from diverse fields such as medical research, food safety, smart agriculture, social media analytics, and predictive equipment maintenance. An example of such an application that we present is our work with Baylor College of Medicine to accelerate the rate of discovery of information on a critical tumor protein (called p53). Because of prior projects related to chemical structures and medicine, the ADLab had expertise in these fields and immediate availability of both medical literature and patents; in addition, we could leverage our infrastructure for external collaboration and our expertise in appropriate systems and analytics. All of this allowed the ADLab and Baylor to very quickly begin to predict potential p53 kinases at a rate many times that of traditional research methods. The system developed from this work is being released as Watson Discovery Analytics.

Finally, we conclude with a recap of some of the lessons we've learned so far, and a look at plans for the future. In addition, we discuss our ongoing internal research to try to measure and quantify the rate of discovery, in order to evaluate and tune the impact of the ADLab itself.

Related Literature

[1] L. Haas, et al., "The IBM Research Accelerated Discovery Lab", SIGMOD Record, June 2014

[2] E. Kandogan, C. Christodoulakis, M. Roth, P. Schwarz, J. Hui, I. Terrizzano, H., Lee, R. Miller, “LabBook: Metadata-driven Social Collaborative Data Analysis”, submitted to IEEE Conference on Visual Analytics Science and Technology (IEEE VAST 2015)

[3] M. Roth, E. Kandogan, J. Hui, P. Schwarz, H. Kache, K. Shank, “From Information Management to Information Orchestration”, 7th Biennial Conference on Innovative Data Systems Research (CIDR 2015)

[4] E. Kandogan, et al., “Data for All: A Systems Approach to Accelerate the Path from Data to Insight”, Big Data Congress, 2013 IEEE International Congress, 427-428

[5] C. Kieliszewski, L. Anderson, S. Stucky, “A case study: Designing the service experience for big data discovery” In Proc. 5 th Int’l Conference on Applied Human Factors and Ergonomics (2014)

[6] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, E. Mueller, “Watson: Beyond Jeopardy!”, Artificial Intelligence, 199-200: 93-105, 2013

[7] J. Fan, A. Kalyanpur, D. Gondek, D. Ferrucci, “Automatic knowledge extraction from documents” IBM Journal of Research and Development, 56(3.4), 5:1-5:10, 2012

[8] A. Ghoting, et al, “SystemML: Declarative machine learning on MapReduce” ICDE 2011 (April 2011), 231-242

[9] E. Kandogan, A. Balakhrisnan, E. Haber, J. Pierce, “From data to insight: work practices of analysts in the enterprise”, IEEE Computer Graphics and Applications, Special Issue on Business Intelligence Analytics, 34 (5): 42-50, 2014

[10] Sequencing the Food Supply Chain, IBM Research, 2015, http://www.research.ibm.com/client-programs/foodsafety/index.shtml