lifescale analytics analytics+insights for life science

client case study

Collaboration in the Cloud

How we helped researchers collaborate in the cloud using the tools they needed while complying with standards

CASE OVERVIEW

The bioinformatics team at a global donor research organization was tasked with the mission to find new and better ways to identity donors for leukemia patients. They needed to discover a way to collaborate with external researchers using a broad array of analytics tools while complying with platform and tool standards supported by corporate IT.

The focus was not only on data quality, but on new approaches to radically improve the selection and management of donor information. And like most informatics teams, the resources available for this task varied day-to-day and untested, prototypical software was used. Additional roadblocks included a lack of staffing, security concerns, software challenges, computing horsepower and lack of expertise in cloud computing.

This case study investigates the root of the problem, how we at Lifescale Analytics approached the issue and the results of employing a private cloud solution for this organization.

BRIEF OVERVIEW

The organization has a large platform of servers and storage used for day-to-day operational activities that are heavily secured and monitored. Management of their large volume of data must be HIPAA-compliant, and all software must be extensively tested, vetted, and written in a small subset of supported programming languages. There is little computing capacity to spare, and development processes have strict controls in place. Java is the approved development platform for all software development, and web-facing applications must be strictly code-reviewed for compliance purposes. Supported operating systems are Redhat Enterprise Linux, Solaris (legacy only), and MS Windows Server.

The Bioinformatics Research Team works on future-discovery projects, many with hypotheses that don't prove to be successful. They have a limited number of java programmers on the team. The Operations side of the corporation has several developers, however they are fully allocated on other work.

6200 Shingle Creek Parkway, Suite 580 Minneapolis, Minnesota 55430 | 763.585.5870 | info@lifescaleanalytics.com | lifescaleanalytics.com

THE CHALLENGES

Staffing

Because the Operations team has a full workload and is tasked with making improvements to their processes, they frequently borrow resources from the Bioinformatics Research Team since maintaining and improving day-to-day operations is priority number one. While it is important to dedicate resources to long-term projects, they are often reassigned in order to support operational needs.

Technologies

The organization uses Java for all internally developed solutions and Oracle databases for all data storage and management. Redhat Enterprise Linux is the officially supported operating system. Security policies and support concerns forbid the use of other technologies. As a result, all experimental code, no matter the language in which it was developed, is refactored into Java before it can even be evaluated for success.

Collaboration

In the bioinformatics, Oracle databases occupy little more than a niche,and Java is not the only programming language in use. While Redhat is used by some researchers, it does not occupy a leading share of the market. Researchers use a wide variety of programming languages and tools, ranging from commercial endeavors such as Matlab, to open source programming languages like R, Ruby, Python and Perl. The majority of development occurs using open-source technologies. As for databases, bioinformatics research spans the gamut of open-source databases, from traditional relational databases such as MySQL and PostgreSQL to a new wave of NoSQL databases such as Hadoop and Voldemort. Anyone wishing to collaborate with the academic research community is forced to interact with these tools. In terms of operating systems, Ubuntu Linux and Debian are widely used, as is CentOS to a lesser degree. Biology-focused Linux distributions have been developed, including CloudBioLinux, which is a tool of choice at several institutions.

The custom code developed by academic researchers is generally prototypical in nature, meaning that it is built for purpose and lacks flexibility. A lot of assumptions are made about what sort of environment is required to run it. Redhat Enterprise Linux rarely has the necessary development libraries easily available.

Computing Horsepower

DNA sequencing is becoming increasingly less expensive. In the near future, it will become economically feasible for prospective donors to submit their entire genome for analysis. A downside is that it requires approximately 3.2gb storage is required to persist a fully-sequenced human genome, and a significant amount of computing power is required to analyze it. Storage and processing requirements will increase as sequencing technology improves, as will capacity needed for marker and interpretive data. This organization maintains information on tens of thousands of donors, so as sequencing becomes more affordable and common, requirements for computing horsepower is going to grow rapidly.

Lack of Cloud Expertise
Like most research organizations, technology expertise was focused in the area of informatics and information delivery. Because dedicated resources are already scarce, having an expert on staff in the area of Cloud plan development and execution was not an option. Even though utilizing the Cloud is the best alternative for addressing many of these issues.

OUR APPROACH AND FINDINGS

In an effort to help the Bioinformatics Research Team accomplish their mission of discovering new and innovative ways to find more donors, Lifescale Analytics was tasked to take the following incremental steps:

- Standardize the existing Cloud computing environment
- Expand the Cloud for non-standard and experimental projects
- Investigate the development of an internal private Cloud

Here's what we found:

- Cloud computing provided a manageable platform that would support many of the non-standard software and solutions needs
- Cloud computing would also provide flexible and economic, on-demand computing capacity

IMMEDIATE ACTION AND THE RESULTS

Through our engagement with the organization, Lifescale Analytics accomplished the following:

- Through our engagement with the organization, Lifescale Analytics accomplished the following:
- Moved a stymied project into the Cloud, allowing for partnering researchers to examine and test the project, enabling publication of a scientific paper on a novel approach for identifying additional potential donors
- Freed up Java developers from Cloud administration
- Pushed the development of a private cloud through a series of approvals, so that implementation could begin
- Standardized the Cloud computing infrastructure around CloudBioLinux and puppet
- Implemented a virtual private cloud in Amazon, to allow for better system management
- Assisted researchers in moving anonymized genomic processing into the cloud
- Increased acceptance of open-source technologies internally

Lifescale Analytics continues to work with the organization to remove research roadblocks. The goal is not to push any given technology, but to ensure that issues are addressed quickly, allowing researchers to experiment more and discover new ideas worth pursuing. And because ideas can now be tested without refactoring into Java, only ideas which have panned out will be refactored. The Private Cloud will also be built around Cloudbiolinux and puppet, mirroring the Amazon setup as much as possible.

lifescale analytics analytics+insights for life science

CONCLUSIONS

No single technology is suitable for all problems. At Lifescale Analytics, we diligently pursue the selection of the best technology for any given business problem. When it comes to research in the Cloud, using Amazon EC2 has often been the best approach. For other projects, Private Cloud will be far more suitable (especially analytics that involves sensitive data or requires HIPAA compliance).

Without proper administration and management, any technology will rapidly become a chaotic mess. Throughout all projects, we advocate for centralized, automated and consistent systems resource management. In this particular case, 'puppet' has been the tool of choice.

WHAT CAN WE DO FOR YOU?

Our mission at Lifescale Analytics is to help life sciences companies gain insight from their data. By applying our expertise in advanced analytics and data management, we tackle the issues that keep them from getting the most from their data. Our job is to find ways to accelerate insight from bioresearch related data, leading to more accurate decisions and a greater impact of product and services in the market.

Our areas of expertise include:

- Data management
- Architecture
- Data science
- Business and Systems Analysis and
- Program execution

Whatever is holding you back from getting the most out of your research data, we can help. Call Ron Noden at (763) 585-5871 to get started.

6200 Shingle Creek Parkway, Suite 580 Minneapolis, Minnesota 55430 | 763.585.5870 | info@lifescaleanalytics.com | lifescaleanalytics.com