Big Data Platforms as a Service

Tiffany Fabianac
Indiana University
Bloomington, Indiana
tifabi@iu.edu

ABSTRACT

Big Data platform solutions allow data producers to use data to the fullest potential by combining processing engines with storage solutions and analytic technologies. Pharmaceutical clients are looking into platform solutions to safely store, analyze, and use clinical trial data, experimental data, drug development studies, drug production, regulation, and a number of other outlets. Just a few of the benefits of using a platform solution to manage these data outlets are possibly not having to change current work processes, that management and other research groups can access and use data without needing special access to systems, and scaleability of storage and analytic components is seamless. The problems faced to implementing big data platform solutions include the selection of a platform vendor, the design of appropriate data architecture, and establishing effective user interfaces.

KEYWORDS

i523, HID313, Big Data, Platform, Cloud Architecture

changed format

1 INTRODUCTION

Most pharmaceutical companies have adopted one or many Laboratory Information Management Systems (LIMS) and/or Electronic Laboratory Notebooks (ELN). These systems are often implemented as standalone systems within a single Research and Development (R&D) group or even within a single laboratory. A problem seen in large- or mid-sized pharmaceutical companies is that different research groups within the same organization often implement isolated LIMS or ELN. This severely restricts data sharing and reuse between groups which leads to many problems such as the same experiment being run multiple times between different groups, regulatory inefficiencies in tracking sample use and storage, and bottle necked development cycles due to missing data.

One of the emerging strategies to combat the problems arising from isolated systems is to combine systems using cloud computing. Platform as a Service (PaaS) provides an environment for the development and execution of applications and software tools. The platform is the heart of a cloud computing infrastructure that enables software on-top as well as data created from such software to be accessed and used by a multitude of users[8].

2 IMPORTANCE OF PLATFORMS

Many organizations struggle to share data and processing tools among researchers. PaaS provides a method of better resource utilization while reducing maintenance costs[7]. As pharmaceutical companies collect larger and larger masses of data through LIMS, ELN, and other systems the need for scaleable storage becomes inescapable. Cloud storage available with the implementation of a PaaS solves current and predictable future data storage needs as clinical trial data becomes truly digital and full genome analysis becomes more available. The surge of stored data requires access to tools with the capability of pulling insights from the data. These analytic tools are available in familiar formats that statisticians know and love such as SAS but new analytic tools have been built into platform environments as well as pushed the development of new market players like Tabeau and Spotfire[10].

A pharmaceutical company's R&D group is made up of several diverse units such as analytical chemistry, oncology, genomics, etc. Each of these groups has their own set of unique requirements and thus require multiple solutions to be implemented across the R&D organization. A problem arises now when an FDA regulator enters the lab space and requires an audit trail for a single sample. The sample was aliquoted and distributed across several groups and R&D management needs to be able to prove to the FDA regulator that the sample has been used only for its designated purpose and has been properly destroyed. The sample's use is recorded in several different LIMS and an ELN which the R&D manager does not have access to. With a properly implemented PaaS the manager can print the usage audit trail from each system without accessing them individually. The manager can pull destruction records and storage locations from current inventory and deliver these records to the FDA regulator without directly contacting any of the lab groups.

3 IMPLEMENTING PLATFORMS

Implementing a platform raises a number of concerns around security, selecting the right solution, designing the data architecture and associated relationships, and planning the user interface. All of the large platform providers have invested enormous amounts of resources into assuring the security of their data storage solutions. The right solution might be based on available applications, the storage solution's design, the cost, the learning curve for use, or a number of other client based requirements. Data architecture has the overarching purpose to design the data warehouse solution without limitations to growth, analysis tools, or query speed. User interface depends mostly on the user requirements, it could be driven by how much visibility is needed and how read and write privileges are designated.

The overarching concern with storing data outside of the organization is security. Numerous methods have been developed to assure cloud security such as integrated stacks used by Google and Microsoft Azure and Service Level Agreements (SLAs)[5]. Cloud companies are required to maintain high security at all levels. Google runs various vulnerability reward programs that pay developers, hackers, and security experts for finding security bugs. In addition

to the product bugs, Google also maintains high security at their data centers which includes laser beam intrusion detection, multi-factor access control, and biometrics to a limited population of less than 1% of Googlers[3].

4 IDENTIFYING THE RIGHT PAAS

Every organization has a unique set of user requirements and every organization shares a certain number of user requirements. Something as simple as requiring a username and password to access content is a requirement shared across the great majority of systems while the need to create complex animal breeding plans that produce offspring with genetic content for 20 specific alleles may be a requirement for one unique client. A market analysis weighing a platform's capabilities against the organization's requirements will often help to narrow down this expanding market. Some of the largest PaaS providers are Microsoft, Amazon, and Google.

Microsoft big data solutions have taken advantage of open source technologies by setting Hadoop as the center of their big data platform. Hadoop is implemented through Hortonworks Data Platform (HDP) which has been developed as a open source solution with Apache and other open source components. Microsoft allows cloud and on-premise implementation, but generally local environments are only used as proof of concept testing. Microsoft platform solutions allow for data to be manipulated and used in Microsoft tools such as Sharepoint and Excel while big data analysis, visualization, and mining can be performed using SQL Server Analysis Services or HDInsight. The Hadoop-based platform has no limitations with structured or unstructured data, a number of additional tools are available for data storage, and efficient queries provide a potential boost to discovery. Microsoft Azure storage runs $40 a month per 1TB and employs a pay for use plan to resource use within the platform's toolbox[4].

Amazon Web Services (AWS) offers data storage solutions in NoSQL and Relational Database models. Interactions with these data engines can be done using Hadoop, Interactive Query Service, or Elasticsearch. Amazon has designed their storage sources in such a way that clients can use any preferred open source application, but Amazon has also developed a toolbox of analytic tools. Amazon offers data warehousing through Amazon Redshift which allows for management, query, and analysis at the petabyte-scale. Amazon storage runs around $80 a month per 1TB. AWS offers Business Intelligence, Artificial Intelligence, Machine Learning, Internet of Things, Serverless Computing, and a number of data interface tools available in a pay-as-you-use billing form[1].

Google Cloud Platform (GCP) offers a complete end-to-end data storage solution which allows the use of GCP developed systems and open source tools. BigQuery is Google's data warehouse tool which is serverless and requires no infrastructure management with the assist of Google Cloud Dataflow. Dataflow eliminates the need for resource management and performance optimization. GCP storage runs $10 a month per 1TB. GCP has a number of applications for data manipulation. Dataproc allows dataset management through Hadoop and Spark, data visualization can be generated through Datalab, Data Studio, and Dataprep which are all Google developed applications[2].

5 DESIGNING THE DATA ARCHITECTURE

All data storage solutions from relational databases to noSQL data stores to cloud data warehouses have to start with a defined architecture. The data architecture model will illustrate how data components will be organized and connected. The mindset of a data architect should be focused on reducing complexity of the data model while maintaining the highest level on utilization. This can be a fine line to walk as a designer. Complexity can be reduced by breaking user requirements down to the most basic and generalized principles to define the simplest data modules. An example of this might be a system that requires a number of different requests and instead of designing a component for vendor requests, user requests, and management requests the component is designed for request and request type. This generality allows for easy future scaling or additional system requirements not yet defined. Cloud systems maintain high utilization by manipulating data using strategic layering. One layer for storage, one layer for defining storage keys, another for combining query tools, another for consolidating query results and so on. With the more established cloud offerings a lot of these layers have already been supplied, but the transitions and interconnections still have to be outlined by a designer[9].

6 DESIGNING THE USER INTERFACE

A system's user interface (UI) must be laid out in a simple and intuitive manner that allows users to perform the tasks required while exploring new insights provided by generated data. There are a number of influences leading to the development of user interfaces such as familiarity; users are familiar and comfortable performing a search in Google or Amazon interfaces and maintain the same high expectation with their working environment. If a user requires sample tracking or auditing they may relate the need to how a package is tracked with FedEx or UPS and expect the same level of access and insight to sample tracking within their working environment. Users may even have an information management system that they use and are comfortable with so switching to a new UI can be daunting as it requires additional training and most likely new work processes.

UI developers have the challenging job of creating the face of an application. A poorly designed face may not attract as many customers as something with a higher graphical output. Even a strong performing system can be downgraded or completely ignored by users if its front end is poorly laid out. Considerations for a UI design include font-size, space between elements, interactive space, and line-width which can all differ across devices such as between a tablet, desktop, or smartphone[6].

7 CONCLUSION

As more and more companies realize the value of their data, platforms and associated tools become more and more vital to organizational success. The pharmaceutical industry knows that data is king, but is experiencing major bottlenecks in deploying platform solutions for the reasons discussed: the cost and complexity of implementation, the concern over security, the frustration of changing or creating new work processes. Current information management systems help scientists and researchers work exponentially faster than they ever could on paper, but current systems are not designed

to facilitate sharing of ideas. This is where platforms come in. A regulatory supervisor should not need training on every information management system to effectively regulate the use and disposal of clinical samples. A laboratory technician should not need to wait for specific system privileges to access a study that the organization did in a different lab space, whether it's in the same building or on the other side of the globe. Platform services are allowing scientists and managers to share ideas more efficiently then they ever have before and the pharmaceutical industry has the potential to exploit this new technology to improve life expectancy, make drugs safer, and research smarter.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor Von Laszewski and Teaching Assistants Hyungro Lee, Juliette Zerick, Saber Sheybani Moghadam, and Miao Jiang.

REFERENCES

[1] 2017. Big Data on AWS. Website. (Oct. 2017). https://aws.amazon.com/big-data/
[2] 2017. Big Data Solutions. Website. (Oct. 2017). https://cloud.google.com/products/big-data/
[3] 2017. Google Security Whitepaper. Website. (Oct. 2017). https://cloud.google.com/security/whitepaper#state-of-the-art_data_centers
[4] 2017. Understanding Microsoft big data solutions. Website. (Oct. 2017). https://msdn.microsoft.com/en-us/library/dn749804.aspx
[5] Valentina Casola, Alessandra De Benedictis, Massimiliano Rak, and Villano Umberto. 2014. Preliminary design of a platform-as-a-service to provide security in cloud. *ResearchGate* (01 2014), 752–757. https://www.researchgate.net/publication/289573602
[6] Miroslav Macik, Tomas Cerny, and Pavel Slavik. 2014. Context-sensitive, cross-platform user interface generation. *Journal on Multimodal User Interfaces* 8, 2 (01 Jun 2014), 217–229. https://doi.org/10.1007/s12193-013-0141-0
[7] Sungyoung Oh, Jieun Cha, Myungkyu Ji, Hyekyung Kang, Seok Kim, Eunyoung Heo, Jong Soo Han, Hyunggoo Kang, Hoseok Chae, Hee Hwang, and Sooyoung Yoo. 2015. Architecture Design of Healthcare Software-as-a-Service Platform for Cloud-Based Clinical Decision Support Service. *Healthcare Informatics Research* 21, 2 (April 2015), 102–110. https://doi.org/10.4258/hir.2015.21.2.102
[8] Arto Ojala and Nina Helander. 2014. Value creation and evolution of a value network: A longitudinal case study on a Platform-as-a-Service provider. In *47th Hawaii International Conference on System Science*, Vol. 47. 975–984.
[9] Jerome H. Saltzer and M. Frans Kaashoek. 2009. *Principles of Computer System Design: An Introduction.* Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-374957-4.00010-4
[10] Domenico Talia. 2013. Clouds for scalable big data analytics. *Computer* 46, 5 (2013), 98–101.

8 BIBTEX ISSUES

Warning–no key, author in www-gcp-security
Warning–no author, editor, organization, or key in www-gcp-security
Warning–to sort, need author or key in www-gcp-security
Warning–no key, author in www-msdn
Warning–no author, editor, organization, or key in www-msdn
Warning–to sort, need author or key in www-msdn
Warning–no key, author in www-aws
Warning–no author, editor, organization, or key in www-aws
Warning–to sort, need author or key in www-aws
Warning–no key, author in www-gcp
Warning–no author, editor, organization, or key in www-gcp
Warning–to sort, need author or key in www-gcp
Warning–no key, author in www-aws
Warning–no key, author in www-aws
Warning–no key, author in www-gcp
Warning–no key, author in www-gcp-security
Warning–no key, author in www-msdn
Warning–no key, author in www-aws
Warning–no author, editor, organization, or key in www-aws
Warning–empty author in www-aws
Warning–no key, author in www-gcp
Warning–no author, editor, organization, or key in www-gcp
Warning–empty author in www-gcp
Warning–no key, author in www-gcp-security
Warning–no author, editor, organization, or key in www-gcp-security
Warning–empty author in www-gcp-security
Warning–no key, author in www-msdn
Warning–no author, editor, organization, or key in www-msdn
Warning–empty author in www-msdn
Warning–no number and no volume in Casola
Warning–empty publisher in Ojala
Warning–empty address in Ojala
Warning–empty address in Saltzer
(There were 33 warnings)

9 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

9.1 Assignment Submission Issues

Do not make changes to your paper during grading, when your repository should be frozen.

9.2 Uncaught Bibliography Errors

Missing bibliography ffile generated by JabRef
Bibtex labels cannot have any spaces, _ or & in it

Citations in text showing as [?]: this means either your report.bib is not up-to-date or there is a spelling error in the label of the item you want to cite, either in report.bib or in report.tex

9.3 Formatting

Incorrect number of keywords or HID and i523 not included in the keywords

Other formatting issues

9.4 Writing Errors

Errors in title, e.g. capitalization

Spelling errors

Are you using *a* and *the* properly?

Do not use phrases such as *shown in the Figure below.* Instead, use *as shown in Figure 3*, when referring to the 3rd ffigure

Do not use the word *I* instead use *we* even if you are the sole author

Do not use the phrase *In this paper/report we show* instead use *We show.* It is not important if this is a paper or a report and does not need to be mentioned

If you want to say *and* do not use *&* but use the word *and*

Use a space after . , :

When using a section command, the section title is not written in all-caps as format does this for you

`\section{Introduction}` and NOT `\section{INTRODUCTION}`

9.5 Citation Issues and Plagiarism

It is your responsibility to make sure no plagiarism occurs. The instructions and resources were given in the class

Claims made without citations provided

Need to paraphrase long quotations (whole sentences or longer)

Need to quote directly cited material

9.6 Latex Errors

Erroneous use of quotation marks, i.e. use “quotes” , instead of ” ”

To emphasize a word, use *emphasize* and not “quote”

When using the characters & # % _ put a backslash before them so that they show up correctly

Pasting and copying from the Web often results in non-ASCII characters to be used in your text, please remove them and replace accordingly. This is the case for quotes, dashes and all the other special characters.

9.7 Structural Issues

Acknowledgement section missing

Incorrect README file

In case of a class and if you do a multi-author paper, you need to add an appendix describing who did what in the paper

The paper has less than 2 pages of text, i.e. excluding images, tables and figures

The paper has more than 6 pages of text, i.e. excluding images, tables and figures

Do not artifficially inffate your paper if you are below the page limit

9.8 Details about the Figures and Tables

Capitalization errors in referring to captions, e.g. Figure 1, Table 2

Do use *label* and *ref* to automatically create figure numbers

Wrong placement of figure caption. They should be on the bottom of the figure

Wrong placement of table caption. They should be on the top of the table

Images submitted incorrectly. They should be in native format, e.g. .graffle, .pptx, .png, .jpg

Do not submit eps images. Instead, convert them to PDF

The image files must be in a single directory named ”images”

In case there is a powerpoint in the submission, the image must be exported as PDF

Make the ffigures large enough so we can read the details. If needed make the ffigure over two columns

Do not worry about the ffigure placement if they are at a different location than you think. Figures are allowed to ffoat. For this class, you should place all ffigures at the end of the report.

In case you copied a ffigure from another paper you need to ask for copyright permission. In case of a class paper, you must include a reference to the original in the caption

Remove any ffigure that is not referred to explicitly in the text (As shown in Figure ..)

Do not use textwidth as a parameter for includegraphics

Figures should be reasonably sized and often you just need to add columnwidth

e.g.

`/includegraphics[width=\columnwidth]{images/myimage.pdf}`