PhUSE EU Connect 2018

Paper DH02

Data Engineering Project (Educating For The Future PhUSE WG)

Guy Garrett, Achieve Intelligence Ltd., Brighton, UK
Beverly Hayes, Janssen R&D, Spring House, PA, USA

ABSTRACT

With an expected 100% increase, over the next 3 years, of data from non-EDC sources (such as smartphones, wearables and custom apps) the traditional methods of managing data for clinical trials presents executives with a resourcing headache.

As such, many companies are looking for lower cost strategies to sure up this shortfall in resourcing. However, citing case studies from other industries, there are new methodologies/technologies in data engineering which could enable automation of much of the "heavy-lifting" currently practiced in clinical data management and statistical programming.

This paper discusses the Data Engineering Project within the PhUSE Computational Science (CS) Working Group, Educating For The Future, with a view to educate clinical data managers in data engineering principles so that they can be prepared, equipped and effective in dealing with the coming "data tsunami" heading to the shores of clinical research.

INTRODUCTION

Did you realise we are living in the age of the Fourth Industrial Revolution? Perhaps you have been busy downloading a myriad of "apps" designed to make your life easier or connecting on social media, uncovering relationships and associations you didn't even know you had. Perhaps you have been shopping a global marketplace, comparing prices, quality and availability, all at your fingertips and in a minutes' time. While this has been happening, the Fourth Industry Revolution has been evolving at exponential proportions [25]. Just ask Siri!

The term "Industrie 4.0", was originated in Germany, as a government-led initiative, to transform manufacturing through advanced digital capability. Thus creating the concept of a "smart factory", based on four key design principles [21]:

1. Interconnection of machines, devices, sensor and people
2. Vast amounts of useful information (data) to drive decision making
3. Technical assistance to aid humans, for example to visualise data or to perform tasks that may be of safety concern for a human.
4. The use of cyber-physical systems to make decisions on their own and to perform tasks as autonomously as possible.

Emerging from the premise of "Industrie 4.0" is the advent of the term "The Fourth Industrial Revolution" (also referred to as "4IR" or "I4.0"). This term originated in 2016 when described by Klaus Schwab (Founder and Executive Chairman of the World Economic Forum), as a "technological revolution that will fundamentally alter the way we live, work, and relate to one another". Klaus goes on to describe it as a digital revolution with innovative uses of a combination of technologies that build upon the premise of the third revolution (i.e. electronics and information technology to automate production). As a result, emerging technologies have brought forth advancements in fields such as artificial intelligence, robotics, the Internet of Things, autonomous vehicles, 3D printing, nanotechnology, biotechnology, materials science, energy storage, and quantum computing. This rapid evolution will undoubtedly affect industries world-wide, already disrupting many industries, such as travel agencies, video rentals and bookstores[32] .

The pharmaceutical industry is also experiencing the impacts of I4.0. Digital and mobile technologies has brought on significant advancements in data acquisition and accessibility as it relates to health care and patient data. As reported in the Tufts-Veeva eClinical Landscape study in 2017, data coming from sources such as, smartphones, custom applications, and mobile health are expected to double in the next 3 years [1]. Therefore requiring greater capabilities in handling large volumes of data, as well as data from coming in through various data streams and

formatting. As with other industries, data will become a critical asset to their business and the effective utilisation of this data can play a critical role in driving growth in the business and bringing novel therapies to the patients who need them.

In this paper, we will focus on the works of the Data Engineering Project within the Educating For The Future Working Group. With the formation of the Working Group in early 2018, the team had taken on the mission to explore how data engineering techniques, successfully deployed in other industries, could be utilised in the pharmaceutical industry, with a goal to *facilitate the education* of the pharmaceutical industry on these techniques.

We will share with you some introductory information about Data Engineering and Data Science and explore how embracing new data engineering techniques may affect the industry culture. You will learn about use cases of Data Engineering in other industries and how advances in digital capability have affected their business model. We will also share some of the many software packages and tools available to enable automation, commonly used in Data Engineering and Data Science.

Finally we will reflect on the benefits that data standardisation has brought to the pharmaceutical industry and share our vision for disseminating information to facilitate your learning going forward.

DATA ENGINEERING

To start this learning journey, exploring the term “Data Engineering” opens the door to the vast opportunities and roles available today centered around data. In doing a simple search on the internet, *“what is data engineering?*”, one will find many posts expressing their understanding of Data Engineering with some variation but also some similarity.

However, what is clear is that Data Engineering encompasses the many considerations that need to be taken into account to optimally curate, transform, secure and disseminate data suitable for analysis. As technology and tools have become more advanced, building such a platform and infrastructure requires engineers and architects of both general and specific expertise. The Data Engineer combines knowledge in areas such as software development, infrastructure, data architecture, data warehousing, cloud technology and data cleaning in order to design, build and test solutions that define the pipelines of data throughout the enterprise, making the data accessible to the organisation. [5] [27] [31]

Optimised Data Engineering appropriately balances the efficiency of an automated process against the cost of development and maintenance of that process, ensuring repetitive processes that require humans to write code, press keys, cut-and-paste and update documents are minimised or eliminated.

DATA SCIENCE

Often paired with the term “Data Engineering” is also the term “Data Science”.

According to Kelle O’Neal and Charles Roe:
“Data Science allows enterprises the ability to turn their data assets into a narrative. Data Science allows that narrative to be expanded across timelines, in different data spaces that trace from the past into the future, with much more involved questions and answers about an enterprise, different potential outcomes, and repercussions based on recommendations. Data Science employs a range of mathematical, business, and scientific techniques to solve complex problems about an organisation’s data assets.” [26]

In contrast, the focus of the Data Engineer is on the process from data curation to dissemination and the focus of the Data Scientist is on the analytics of the data, thus extracting knowledge from the data.

To achieve quality data capture, near-real-time accessibility and meaningful analytics, one cannot function without the other, and effective teamwork optimises the value of each role. As such, an analytics team would be composed of distinct roles/capabilities [17]:

- Data Engineers (in areas such as database architecture, database development, machine learning architecture, ETL scripting , etc.)
- Data Scientists
- Business Analysts

Data Engineering brings together the broad expertise, of these roles, to ensure the data are curated and accessible to the Data Scientist, and in our environment today, this process is becoming more and more complex. Therefore,

expertise in curating big-data and data of varying formats (structured and unstructured) is a critical core competency to optimise the potential impact of these digital assets (i.e. the data).

The Data Scientist works deep in the data, utilizing various tools and techniques to discover patterns in the data that may drive decision making for the business. Optimising utilisation of the data to enable accurate conclusions can bear greater value to the organisation. As an example, per Tom Eunice's post, "a fraud-detection algorithm may be very accurate when based on many months of historical data. However, months of historical data may not always be available. Designing a fraud-detection model that is still accurate using historical data from only a few days would be of more use and more practical to implement." [17]

The Business Analyst helps the Data Scientist understand the meaning of the data and the relevance of any discovered relationships. Initially, uncovering relationships in the data and upon further investigation, identifies meaningful patterns that may reveal information that otherwise may not have been known. [17]

As you will see in the sections to follow, the full complement of the roles in an analytics team is what drives the business value. One discipline without the other (e.g. data engineering without data science) will result in missed opportunities. In the sections to follow, we often refer to Data Engineering, however, due to the close ties to Data Science, some examples elude to both Data Engineering and Data Science.

USE CASES FROM OTHER INDUSTRIES

In this section, we present three use cases from the transportation, retail, and agricultural industry. The use cases illustrate the importance and usage of Data Engineering. In each example the data collected, the consumer of the data, and the value of the organisation is reviewed. Similarities and potential applications to the pharmaceutical industry are discussed.

UBER

When it comes to moving people and making deliveries, few companies are more widespread and more widely-recognised than Uber. Uber is working to make transportation safer and more accessible, helping people order food quickly and affordably, reducing congestion in cities by getting more people into fewer cars, and creating opportunities for people to work on their own terms.[4]
But how do they do it? As it turns out, there's a great deal of data being collected, produced and visualised behind the scenes — all working to create a more efficient company and impact transportation as a whole. [22]

Data is Uber's biggest asset. They collect huge amounts of data to enable their business. They collect data from billions of GPS locations and their platforms are processing millions of events. Uber stores data about every trip for prediction about supply and demand. They also collect data on their drivers to understand their vehicle, location, speed, etc.

All of this data is then analysed and visualised to predict wait time, passenger demands, optimal driver locations, etc. They leverage data visualisation to understand safety, efficiency, and traffic. Visual analytics made their data actionable.[16]

Uber feels that if they don't use technology to analyse and interpret data, it is a missed opportunity to better understanding their business. They don't sit on their databases. They look for connections in every possible ounce of their data. [22]

This use case may make you wonder whether the pharmaceutical industry uses every ounce of data collected in a clinical trial or electronic health record? Is the pharma industry "sitting" on databases and missing opportunities? And can more data be collected and analysed in real time? If so, the pharmaceutical industry must identify opportunities to leverage both clinical and real world data and visual analytics to improve the efficiency and effectiveness of drug development. .

AMAZON

When Amazon first launched, it had a clear and ambitious mission to offer Earth's biggest selection and to be Earth's most customer-centric company. [13]

Their goal is still the same 20 years later as they continue to innovate new solutions to make things easier, faster, better, and more cost effective with a core focus being a customer-centric business.

According to founder and CEO, Jeff Bezos, technology is very important to supporting this focus on the customer. In their 2010 Annual Report (Amazon, 2011) he said:
"Look inside a current textbook on software architecture, and you'll find few patterns that we don't apply at Amazon. We use high-performance transactions systems, complex rendering and object caching, workflow and queuing systems, business intelligence and data analytics, machine learning and pattern recognition, neural networks and probabilistic decision making, and a wide variety of other techniques. And while many of our systems are based on the latest in computer science research, this often hasn't been sufficient: our architects and engineers have had to advance research in directions that no academic had yet taken. Many of the problems we face have no textbook solutions, and so we — happily — invent new approaches… All the effort we put into technology might not matter that much if we kept technology off to the side in some sort of R&D department, but we don't take that approach. Technology infuses all of our teams, all of our processes, our decision-making, and our approach to innovation in each of our businesses. It is deeply integrated into everything we do". [13]

Similar to Uber, Amazon collects multitudes of data including inventory levels, costs, sales, customer demographics, buying patterns, and competitor data. The data is both structured and unstructured from many diverse sources. The data is used in a variety of ways but most importantly it is used to understand customer behavior to drive customer growth, expansion, retention and personalisation. This is aligned with Amazon's customer-centric business model.

In addition to customer focus analytics, Amazon uses data to optimise their supply chain, streamline fulfillment, anticipate shipping and recommend products. Since Amazon recognises the importance of technology and a focus on artificial intelligence and machine learning, they make the data widely available within their organisation to ensure data driven management styles. Amazon does not silo or restrict access to their database because they feel silo-ed data results in slow decision making. The databases are consumed across their organisation by both data scientists and machines.[13]

In this example, Amazon utilises data engineering and technology to unlock insights from their customer data and turn it into a competitive advantage. They make their data widely available within their organisation. Is this mindset and approach applicable in the pharmaceutical industry? What are the compliance concerns and risks of providing broad access to experimental clinical data? And how can the pharmaceutical industry utilise real world evidence data as well as historical data from clinical trials , similar to Amazon's customer database, as a competitive advantage?

AGRICULTURE

The field of agriculture (no pun intended!) has also moved into using formal analytical tools to aid producers (farmers) and oversight (government). A main area of focus is crop prediction. To that end, a variety of data is collected, for example: weather, soil, past crop results. Precise and accurate crop prediction can create an environment where yields are known and can be priced accordingly.

This analytical area is different from pharma in two major ways:

1. Predictor variables are highly environment-dependent (e.g., rain, air temperature) instead of highly controlled (e.g., fixed-dose regimens)
2. Regulatory requirements are not as stringent

Still, there is some potential to borrow tools and ideas from agriculture research to pharmaceutical research.

Here are four examples of software tools used in agriculture analytics:

HARVIST
The HARVIST (Heterogeneous Agricultural Research Via Interactive, Scalable Technology) project integrates multiple Earth Science data sources into a single graphical user interface that allows for the investigation of connections between different variables. In particular, the focus is on relationships between weather and crop yield.[2]

AgrometShell (AMS)
This software is designed to facilitate monitoring of the growing season, for governmental agency use. It is license-free, and bridges the gap between agromet, remote sensing and socio-economic datasets.[19]

Crop Yield Predictor
This software was designed as an interactive decision tool to predict crop yields and economic returns for deficit-irrigated crops. Users can designate potential irrigation schedules to optimise yields and net returns. These schedules can be tested with a range of annual precipitation to find yield and income risks from wet, average, and dry years. Alternative irrigation schedules could include pre-season irrigation, irrigation amounts and frequency, earlier or later commencement of irrigation, and earlier or later cessation of irrigation.[14]

Descartes Labs
This platform allows immediate, easy access to global data - historically and in near real-time. It uses Python APIs to access data about the earth within seconds.[15]

Users can write any Python function, and run it on thousands of images in parallel. They can run NDVI, BAI, NDSII, SAVI, or any other calculations over each pixel within your area of interest. Fuse satellite data, including high-resolution data from Airbus One Atlas, with weather data and Descartes Labs-generated data sets into a single global-scale analysis.

In all of the industry use cases above, data engineering and technology is utilised to unlock insights. Might there be applications of such concepts in pharmaceutical research?

AUTOMATION

In the age of I4.0, advancements in artificial intelligence and machine learning along with improvements in technologies for data storage, data access and computer processing are driving innovations in automation of data acquisition, integration and analytics. Automation software and tools continue to grow in availability and capability. However, the questions remain.

- Why automate?
- What should be automated?
- How might automation impact our business?

With these advancements in technology and automation capabilities, companies across all industries are looking at how they can use automation to improve their business. A recent post on DevOps.com stating "Today's systems are simply becoming too big and complex to run completely manually, and working without automation is largely unsustainable for many enterprises across all industries", [35] eludes to the criticality of automation to survive in today's competitive market.

However, determining what to automate requires careful consideration of how the automation will fit into your business workflows. Many organisations develop an automation strategy taking into consideration manual tasks that are often repeated or standardised and tasks that require human problem solving, particularly in un-predictable scenarios. As advised by Jim Higgins on Forbes.com, "The ideal level of automation is less about replacement and more about enablement. It helps users be better at their jobs, giving them analytics to make sound decisions for the business and freeing them from repetitive and monotonous tasks so they can be more strategic." In his post he emphasises the importance of the user feeling they have control of the technology and that the user feels the value of the automation, making them better at their job.[18] These are important considerations for successful integration of automations into your business processes.

As discussed in the previous use cases, data is needed in real-time to answer important business questions in every sector of our economy including pharmaceutics. The use of software and automation is essential in making this a reality.

SOFTWARE AND TOOLS TO ENABLE AUTOMATION

A few examples of software languages and tools to enable automation are included below.

Apache® Software Foundation: An open source software ecosystem with projects supporting big data, databases, graphics, and many more capabilities. A few of Apache projects are included below – visit https://www.apache.org/ to learn more.

Apache Spark™: a unified analytics engine for large-scale data processing for big data and machine learning. Apache Spark was originally developed at UC Berkeley in 2009.[36]

Apache Arrow™: Apache Arrow is a cross-language development platform for in-memory data. It specifies a standardised language-independent columnar memory format for flat and hierarchical data, organised for efficient analytic operations on modern hardware. It also provides computational libraries and zero-copy streaming messaging and interprocess communication. Languages currently supported include C, C++, Java, JavaScript, Python, and Ruby. [7]

Apache Airflow (undergoing incubation): Airflow is a platform to programmatically author, schedule and monitor workflows.[6]

Apache Ignite™ (a project managed by the Apache Ignite Committee)

Apache Ignite In-Memory Data Fabric is designed to deliver uncompromised performance for a wide set of in-memory computing use cases from high performance computing, to the industry most advanced data grid, in-memory SQL, in-memory file system, streaming, and more.[9]

Apache™ Hadoop® : The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.[8]

Apache Mesos™: Apache Mesos is a cluster manager that provides efficient resource isolation and sharing across distributed applications, or frameworks. It can run Hadoop, MPI, Hypertable, Spark, and other frameworks on a dynamically shared pool of nodes.[10]

Java™: A general purpose object oriented programming language released in 1995 by Sun Microsystems and later acquired by Oracle. Original goal of Java was to develop a language that could run on consumer appliances, like a refrigerator, and what we now call the internet of things. Their idea was to "write once, run anywhere" which in other words meant that one can write a piece of code and then it could be compiled to run on any device. Java however became more popular for its features of writing applets, small programs, which run inside a web browser. Java gained wide popularity and lots of success with this capability especially given the rise of the internet in the late 1990's. [37]

Jupyter Notebook: The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualisations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualisation, machine learning, and much more.[28]

Python™: Dating from 1991, the Python programming language was considered a gap-filler, a way to write scripts that "automate the boring stuff" (as one popular book on learning Python put it) or to rapidly prototype applications that will be implemented in other languages. However, over the past few years, Python has emerged as a first-class citizen in modern software development, infrastructure management, and data analysis. It is no longer a back-room utility language, but a major force in web application, creation and systems management, and a key driver of the explosion in big data analytics and machine intelligence.[38]

R and R Shiny: 2018 marks the 25 year anniversary of the creation of R. This open source statistical software used by millions of users around the world modernises the way we think of statistical computing. The Comprehensive R Archive Network (CRAN) provides open sharing of code libraries and the elimination of redeveloping code. R Shiny is an R package that provides end users web applications for visualising and analysing data.

R Markdown: R Markdown provides an authoring framework for data science. R Markdown documents are fully reproducible and support dozens of static and dynamic output formats.
You can use a single R Markdown file to both save and execute code and generate high quality reports that can be shared with an audience.

RDF data cube: The standard provides a means to publish multi-dimensional data, such as statistics, on the web in such a way that it can be linked to related data sets and concepts using the W3C RDF (Resource Description Framework) standard. The model underpinning the Data Cube vocabulary is compatible with the cube model that underlies SDMX (Statistical Data and Metadata eXchange), an ISO standard for exchanging and sharing statistical data and metadata among organisations. [34]

SQL: SQL is a domain-specific language used in programming and designed for managing data held in a relational database management system, or for stream processing in a relational data stream management system.

Tensorflow™: TensorFlow is an open source software library for high performance numerical computation. Its flexible architecture allows easy deployment of computation across a variety of platforms (CPUs, GPUs, TPUs), and from desktops to clusters of servers to mobile and edge devices. Originally developed by researchers and engineers from the Google Brain team within Google's AI organisation, it comes with strong support for machine learning and deep learning and the flexible numerical computation core is used across many other scientific domains. [3]

NoSQL database systems: NoSQL, is a form of database model other than that of a relational database. NoSQL databases have come to the fore with the ever expanding volumes of data and speed of data processing in the world

of e-commerce and social media. Benefits of a NoSQL data model may be a simpler design, distributed data stores and ultimately increased speed of processing.[24]

Although many of the software and techniques above play a role in enabling automation, current technologies will continue to evolve and mature. Many organisations are exploring automation and may test various solutions. However, automating for the sake of automation may not lead to the optimal solution. Developing a clear strategy, taking into consideration the problem, the people, and the desired outcome can lead to a greater effect of automation to improve efficiency and productivity.

CULTURE, BENEFITS & CHALLENGES

After now sharing some examples of data engineering techniques in industries such as travel, retail and agriculture and learning about Data Engineering and Data Science. How might this apply in pharmaceutical research? What might be the impact to the culture? How can the industry benefit from such techniques? And what may be the challenges or potential roadblocks?

CULTURE

Whenever a new methodology is introduced to an existing team, organisation or sector it is important to understand the prevailing culture of that incumbent organism. As the phrase often attributed to Peter Drucker states; “Culture eats Strategy for breakfast…”[12] If uptake depends on human decisioning then it doesn’t matter how innovative the technology, efficient or sensible the new methodology may be; if the people responsible for integrating, deploying and maintaining this new way of working don’t, for whatever reason, want it to happen.

For the pharmaceutical industry we must look inside at where we are, where we’ve come from and indeed where we’re heading to – to gain a holistic understanding of our industry culture – in order to make the case, for or against, with regards to adopting a data engineering approach to clinical data management.

Three main drivers seem to recur in data engineering discussions within the pharmaceutical industry:

Innovative Science

Traditionally the pharmaceutical sector would initiate drug discovery through two main avenues. Principally through internal development, discovering research projects that could be carried forward into human experimentation models, within an acceptable budget. Although this approach is highly rewarding for an organisation, to see a compound go from discovery to marketing approval, it is rarely enough to sustain a robust product development pipeline.

A second route would come from product in-licensing, co-development partnerships, and/or mergers and acquisitions to bring new products into the organisations’ development pipeline. Ideally this approach will lead to an eventual submission to the health authorities for marketing approval, although this is not always the case.. This approach often requires significant up-front investment, however, may also yield a high return.

As early as 2011, even before the Big Data boom, MIT’s Sloan School of management provided evidence that data-driven organisations performed 5-6% more effectively than their non-analytic cousins.[11] As such, there is now an increasing move towards a third way; that of the pharmaceutical organisations drawing more attention to Data Engineering and Data Science such that they may, leverage their vast amounts of historical clinical study data and augment it with external revolutionary data collections, such as wearables, genomics and electronic health records. This approach offers new opportunities to conduct research, leveraging data to its fullest potential. Combining this approach with the two described above creates multiple avenues for innovation for the organisation.

Patient Protection

Patient safety is, rightfully so, a major concern that needs to be addressed when introducing the topic of Data Engineering within the pharmaceutical industry. Risk needs to be managed effectively to ensure the safety of patients, hence the structure incorporated in all sponsor companies for identifying, managing and reporting serious adverse events. Regulations have developed over time and are continually being monitored and reviewed to ensure data around this important and sensitive area are handled appropriately [for instance 21 CFR Part 11].

However, we must not be so averse to risk that we miss the potential improvement opportunities offered by data augmentation, automation, enrichment and standardisation (key components of Data Engineering). Hiring programmers to re-program tasks which are standard use within the organisation could be deemed an inefficient use of a resource which could be better utilised if invested in discovering further treatments. Data Engineering, with the

appropriate level of checks and balances, could offer a “safer” data environment by adopting “write-once, use-often” programs building in automation where value added. Such a change may increase overall productivity while also mitigating the risk of human error.

Value Generation

As with other industries, value generation is an important factor to ensure cost-efficient drug development in the pharmaceutical industry. The cost of investing in research, change management and new technology, often with no guarantee of return, has long been a closely monitored metric. Investing in Data Engineering techniques that enable organisations to reduce costs and shorten timelines while also utilising Data Science across legacy datastores delivers a double impact of value generation for executive leadership teams.

Therefore, utilising Data Engineering to preserve or advance innovative science, patient protection and value generation within the organisation may enable quicker adoption and motivation for the change.

With these cultural precepts in mind we can begin to look into the benefits and challenges of applying emerging techniques in Data Engineering within the pharmaceutical industry.

BENEFITS

With much of study data being similar, automating the collection, validation, reporting and analysis could vastly reduce timescales for set-up, database lock and time to submission. If eCRF screens are created using standard variables for patient numbers, investigator details, substance lists etc., this can free up the data architect/engineer to determine how best to deal with non-conformed data (such as exploratory biomarker data or real world evidence).

Data validation, which can be a main source of contention in a study timeline, if automated, can reduce human error, whilst automatically auditing how errors are dealt with. Regulations are clear on how modification to data items need to be handled appropriately. A locked down standardisation routine can manage a majority of “common” data errors (e.g. missing or incorrect zip code for an address) with a formal method of query management auditing the automatic responses from investigators.

If centralised data is also adopted, this lays a good foundation for Data Science to come to the fore. Being able to analyse across studies to determine answers to “never before thought of” questions is a huge benefit to companies. It can also pinpoint problems such as patient scarcity, by identifying patients from previous studies who may benefit from new trials. [33]

Risk-based monitoring and adaptive trials already look to centralised data for their source, however due to the siloed study culture in most organisations these departments often rely on pooling data after the event, rather than a fully automated, near-real-time, centrally populated conformed database.

These are just a few examples of how data engineering techniques can improve quality and efficiency when processing the information captured in clinical trials. As we further explore Data Engineering and learn of the many techniques and tools available, many more applications will arise, bringing forth greater value and benefit to the industry.

CHALLENGES

The disadvantages of adopting data engineering generally fall into the following categories:

- Initial Investment
- Ongoing Maintenance

In order to create a data engineering culture within an organisation time and money is required, yet this is often at odds with the financial model, as this investment is not directly associated to a compound our clinical study. Additionally, a certain amount of centralisation is required to enable digital connectivity between systems/tools, bringing along additional challenges as it relates to security, data protection and maintenance of the study blind.Therefore, strategic investment is required, with a robust security model, which needs executive buy-in from senior leaders who can envision the longer term benefits of data automation.

The type of upfront organisational design required is in the areas of data capture, data validation, data storage and reporting/submission. Data engineering works most optimally within a set of standards, which the pharmaceutical industry has no shortage of. For example, CDASH for collection, SDTM/Adam for submission can be utilised effectively within an engineered data flow – but each study needs to have architects, right from the outset, involved in study set-up to ensure that those studies are adhered to, or to capture innovative ideas and bake them into ongoing standards, or alternatively, determine methods to cater for these non-conformed data elements.

From an ongoing perspective, organisations need different skills to that of the traditional clinical Data Manager. The ability to think big picture, yet have a precise focus on detail to ensure individual studies don't lose their innovative science, whilst conforming to centralisation, is essential for the Data Engineer who will create and maintain these automated data pipelines. An automated test harness needs to be employed to ensure the ongoing maintenance does not break what is already being captured. Lastly, a pragmatic mindset to ensure overengineering doesn't occur is essential for an optimized data engineering model. [29]

INDUSTRY STANDARDS

Since the formation of CDISC, significant progress has been made in standardising the format of data collected, analysed and submitted to the health authorities. This has allowed organisations to explore methods to build automation in their data flows and reporting tools.

Various pharmaceutical companies are/have explored automation in creation of a standard CRF, conversion of data into SDTM , checking of data for submission compliance, production of standard tables, listings & figures, creating components of the e-submission data packages, etc. However, in many cases the tools require rather rigid adherence to a standard, making it difficult to truly realise the benefits of the automation and requiring additional manual upkeep and maintenance.

However, recent advances in data engineering techniques, such as machine learning and big data processing, has allowed companies to more easily curate data that are in a structured, as well as, unstructured format. These concepts have been applied in the healthcare industry as shared at the PhUSE single day event in Ridgefield, CT. At the event, Dr. Wade Schulz, of the Yale School of Medicine, shared the various technologies and tools used to build a data lake that integrates with their clinical information systems to provide historic and real-time data for research studies and clinical decision support. Tools in his "Data Science Toolkit" include, but is not limited to, kafka, Storm, Apache Spark, Hadoop, Apache HBASE and python to ingest, process, store and analyse clinical and healthcare data. He also notes the anticipation of a significant amount of future healthcare data in unstructured format, posing greater challenge to ingest and process data. [31]

Reflecting on the extensive standards and structured data formats that exist in the pharmaceutical industry today and the advancements brought forward through I4.0, the pharmaceutical industry is well positioned to take advantage of emerging technologies in automation, particularly on standardised data. Thus allowing for focus on novel data types and unstructured data.

EDUCATION METHODS

Now that we have shared information on data engineering, uses in other industries and tying it back to the pharmaceutical industry, we will reflect on the ultimate goal of our working group, which is to *facilitate the education* of the pharmaceutical industry on these concepts and techniques.

As a first step for our working group, we discussed various possible educational methods. Taking into consideration that educational methods have changed significantly over the past few decades, we set our goal to cater to different learning preferences by using a wide variety of resources available today. [23]

Therefore we plan to employ much, if not all, of the following educational methods and resources to build our repository and develop learning pathways:

- Videos
- Textbooks or articles on selected topics
- Blogs
- Podcasts
- Social Media Channels
- Interactive Components
- Professional trainings
- Pages of/Links to other websites

ACQUISITION OF INFORMATION

Now that we defined the various educational methods we want to cover in our "Educating For The Future" repository, an important question is how we plan to gather the information.

A simple approach to get information about data engineering is to "Google it". That is, conduct digital searches on terms such as "Data engineer", "data science", "database modelling", "NoSQL database", "RDF triplestore", "Spark",

etc… However, a quick Google search for “Data Engineer” alone delivers 416.000.000 results. While we will certainly include this approach and seek out trustworthy links, we have also identified other means to gather information.

One method is to utilise machine learning capabilities from platforms like Pinterest. Pinterest allows users to create “pinboards”, where they can pin digital material of interest. The user then receives automated recommendations of similar pins, or pinboards, from other users on Pinterest. This method instantly brings together users of a similar interest to share information and knowledge. As a first step, we have created a pinboard (https://pin.it/nup5bju2qhbso7) and pinned material related to the education methods noted above.

Another method is to acquire information by following discussion forums where subject matter experts share their knowledge. Here we might explicitly use Reddit. Reddit (www.reddit.com) represents a massive collection of forums, where people can share news and content or comment on other people’s posts. It also contains a great deal of information on the data engineering topic: (https://www.reddit.com/r/dataengineering/)

In addition to the resources noted above, we see this as a great opportunity to tap into the vast knowledge of the PhUSE community. With over 8,500 PhUSE members worldwide, many of them are experts in their field who want to share their expertise in order to advance the industry. Through the use of videos, blogs and social media, we plan to create content drawing from the collective knowledge of PhUSE members around the globe.

These methods are a few examples of how we will acquire information on our data engineering topic. As we continue in our learning journey, our approach and methodologies will also evolve as we learn of new technologies and innovations in this area.

OUR SQUARESPACE

In the Educating For The Future Working Group our goal is to present information in an organised and meaningful way to facilitate learning for the user. More than just a file-share, we plan to build learning pathways, allowing the user to consume the relevant material to facilitate their learning process. The intent is not to “train” the user, but instead to provide information that allows the user to learn more about the topic (or technique) and how it is utilised in other industries, and perhaps within the pharmaceutical industry as well.

The working group has selected Squarespace as a tool to develop our learning portal. The portal will be constructed to highlight materials gathered by each of the sub-teams: Design Thinking, Machine Learning & Artificial Intelligence and Data Engineering. In the Data Engineering sub-team portal, learning pathways will be organised by category, such as industry use cases, data model types and automation software.

Link to our Squarespace portal: http://bit.ly/PhUSE_Education

CONCLUSION

We have shared with you the mission and goal of the Data Engineering Project within the Educating For The Future Working Group. You have gained an awareness of what is meant by Data Engineering and how it contrasts with Data Science, and learned of the importance of teamwork between Data Engineering, Data Science and Business Analytics to create value through the optimisation of data utilisation to gain a competitive advantage.

We have also learned there are many techniques, tools and software employed in the Data Engineering space, only some of which have been presented in this paper. Although, evident from the information shared, there are many new techniques for handling very large data bases, maximising data processing speed, reducing time writing programming code and eliminating manual steps.

In the use cases, it is shown that data can play a critical role to drive decision making, regardless of the industry, whether it be transportation, retail or agriculture. It is also clear that effective Data Engineering can optimise automation, enable scalability and improve the speed and quality of business operations.

As we contemplate these new concepts and learn how other industries are employing these methodologies in their day-to-day operations we reflect on the pharmaceutical industry and what the potential impacts and considerations may be.

In the pharmaceutical industry, our mission is to bring novel therapies to the patients who need them in order to improve, prolong and/or save lives. Critical to our mission is the safety, efficacy and availability of the drug/product.

As such, data plays a key role in our quest. It is the mode by which we evaluate the safety and efficacy of the drug before making it available for use in hospitals and doctors' offices, world-wide.

Over the past twenty-years, the industry has partnered with health authorities to establish standardisations in data capture, analysis and submissions. These initial efforts in establishing standards has positioned the industry well to improve their ability to quickly and accurately capture and summarise data. However, we have learned that in the age of I4.0, with the advent of new technologies allowing data to be captured directly from the source, such as wearables and smartphones, we will be faced with new challenges to curate data from vast sources, of significant volume, as quickly and efficiently as possible, in order to achieve our mission.

These challenges come with many complexities due to the nature of the industry. Considerations related to health authority regulations, data protection and patient privacy, to name a few, must be carefully assessed to ensure overall compliance and protection of the patient's health and well being are maintained. This is also a well-established industry, dating back several decades of data collection and analysis. Therefore, adopting new practices, tools and technologies, requires significant investment, in both time and money, to effect change.

Although adopting new methodologies in data processing brings forth many challenges, when considering our overall mission, optimised Data Engineering methodologies will become critical in order to bring new therapies to patients as quickly and efficiently as possible.

To do this we ask ourselves these questions:
In what areas might current Data Engineering methodologies make the greatest impact in clinical data management?
In what areas might automation bring the greatest value?
What are the biggest challenges in applying new methodologies? How might the industry have to change?

The challenges facing us with the imminent "data tsunami" also bring great opportunity to employ many of the new techniques and technologies available today. Undoubtedly, to make large scale changes in such a deep-rooted and regulated industry, like the pharmaceutical industry, will take careful consideration with managed risks. However, as alluded to earlier in the paper, with the rapid advancements in technology, operational challenges will become greater and greater as data sources and the volume of data increase over time.

Taking into consideration the new challenges ahead, one is often reminded of the famous quote by Winston Churchill, "To improve is to change; to be perfect is to change often"

HOW YOU CAN GET INVOLVED

The Data Engineering Project within the Educating For The Future Working Group is striving to create a learning portal to *facilitate the education* of the pharmaceutical industry on Data Engineering techniques and methodologies used in other industries. Our Squarespace portal provides us with a medium to share knowledge and explore various capabilities and their application, across various industries. This provides a framework to curate and share knowledge as we continue on our quest to leverage data, as quickly and efficiently, to bring novel therapies to the patients who need them.

Bringing the collective knowledge within the pharmaceutical industry together with information gathered external to the industry may uncover new and innovative solutions that employ Data Engineering in pharma. The PhUSE Working Group is designed to bring together these diverse perspectives. Our team is comprised of members internal and external to the pharmaceutical industry, including academia. If you have an interest to contribute and join the Working Group, please contact us so we can join forces along this learning journey.

REFERENCES

[1] *Tufts – Veeva 2017 EClinical Landscape Study*. Tufts University, 2018, pp. 11–13, *Tufts – Veeva 2017 EClinical Landscape Study*.

[2] “About HARVIST.” *NASA*, Jet Propulsion Laboratory, California Institute of Technology, harvist.jpl.nasa.gov/.

[3] “About TensorFlow.” *TensorFlow*, www.tensorflow.org/.

[4] “About Uber - Our Story - Vision for Our Future | Uber.” *Driver Requirements | How To Drive With Uber | Uber*, 2018, www.uber.com/about/.

[5] Aghabozorgi, Saeed, and Polong Li. “Data Scientist vs Data Engineer, What’s the Difference?” *Cognitive Class Blog*, 2016, cognitiveclass.ai/blog/data-scientist-vs-data-engineer/.

[6] “Apache Airflow (Incubating) Documentation¶.” *Apache Airflow (Incubating) Documentation - Airflow Documentation*, Apache Incubator, airflow.apache.org/.

[7] “Apache Arrow.” *Apache Arrow Homepage*, The Apache Software Foundation, arrow.apache.org/.

[8] “Apache Hadoop.” *Apache Hadoop*, The Apache Software Foundation, hadoop.apache.org/.

[9] “Apache Ignite.” *Apache Ignite*, The Apache Software Foundation, ignite.apache.org/.

[10] “Apache Mesos.” *Apache Mesos*, The Apache Software Foundation, mesos.apache.org/.

[11] Brynjolfsson, Erik, et al. “Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?” *SSRN*, 22 Apr. 2011, ssrn.com/abstract=1819486.

[12] Campbell, David J., et al. *Business Strategy: an Introduction*. Palgrave Macmillan, 2011.

[13] Chaffey, Dave. “Amazon.com Case Study - 2018 Update.” *Smart Insights*, Smart Insights, 14 Aug. 2018, www.smartinsights.com/digital-marketing-strategy/online-business-revenue-models/amazon-case-study/.

[14] “Crop Yield Predictor.” *Crop Water Allocator | K-State Mobile Irrigation Lab*, K-State Research & Extension Mobile Irrigation Lab, www.bae.ksu.edu/mobileirrigationlab/crop-yield-predictor.

[15] “Descartes Labs: Home.” *Descartes Labs*, Descartes Labs, www.descarteslabs.com/.

[16] “Engineering Intelligence Through Data Visualization at Uber.” *Uber Engineering Blog*, 2016, eng.uber.com/data-viz-intel/.

[17] Eunice, Tom. “Do Data Scientists Need Data Management.” *IBM Big Data & Analytics Hub*, IBM, 2015, www.ibmbigdatahub.com/blog/do-data-scientists-need-data-management.

[18] Higgins, Jim. “How Much Automation Is Too Much?” *Forbes*, Forbes Magazine, 6 Apr. 2018, www.forbes.com/sites/forbestechcouncil/2018/04/06/how-much-automation-is-too-much/#96106e2f9696.

[19] Hoefsloot, Peter. “AgrometShell.” *Hoefsloot Spacial Solutions*, Peter Hoefsloot, hoefsloot.com/new/?software=agrometshell.

[20] Hyde, Brigham. “The Spotify/ITunes Model For AI In Health Care.” *Forbes*, Forbes Magazine, 5 July 2018, www.forbes.com/sites/forbestechcouncil/2018/07/05/the-spotifyitunes-model-for-ai-in-health-care/#501791c67e4a.

[21] “Industry 4.0.” *Wikipedia*, Wikimedia Foundation, 2018, en.wikipedia.org/wiki/Industry_4.0#cite_note-Definition-I4.0-1.

[22] Jacob, Sherice. “How Uber Uses Data to Improve Their Service And Create The New Wave of Mobility.” *NEILPATEL*, neilpatel.com/blog/how-uber-uses-data/.

[23] Kamel Boulos, Maged N, et al. “Wikis, Blogs and Podcasts: a New Generation of Web-Based Tools for Virtual Collaborative Clinical Practice and Education.” *BMC Medical Education*, BioMed Central Ltd, 15 Aug. 2006, bmcmededuc.biomedcentral.com/articles/10.1186/1472-6920-6-41.

[24] Moniruzzaman, A B M, and Syed Akhter Hossain. “NoSQL Database: New Era of Databases for Big Data Analytics - Classification, Characteristics and Comparison.” *Academia.edu - Share Research*, International Journal of Database Theory and Application, www.academia.edu/5352898/NoSQL_Database_New_Era_of_Databases_for_Big_Data_Analytics_-_Classification_Characteristics_and_Comparison.

[25] Montresor, Fulvia. “The 7 Technologies Changing Your World.” *World Economic Forum*, 2016, www.weforum.org/agenda/2016/01/a-brief-guide-to-the-technologies-changing-world.

[26] O'Neal, Kelle, and Charles Roe. “Business Intelligence versus Data Science: A DATAVERSITY 2015 Report.” DATAVERSITY, DATAVERSITY, 2015, http://whitepapers.dataversity.net/content54237/.

[27] Paruchuri, Vik. “Data Engineering Series.” *Dataquest*, Dataquest, 15 Dec. 2017, www.dataquest.io/blog/what-is-a-data-engineer/.

[28] “Project Jupyter.” *Project Jupyter*, 2018, jupyter.org/.

[29] Savva, Nicos, and Gabriel Straub. “Making Big Data Deliver.” *London Business School*, London Business School, 2018, www.london.edu/faculty-and-research/lbsr/making-big-data-deliver.

[30] Scavicchio, Julia, and Mike West. “What Is the Difference between Data Engineer, Data Architect, Data Infrastructure and Machine Learning Engineer?” *Quora*, 2016, www.quora.com/What-is-the-difference-between-data-engineer-data-architect-data-infrastructure-and-machine-learning-engineer.

[31] Schulz, Wade. “Baikal – Implementing and Deploying Clinical Models with a Real-Time Data Lake.” PhUSE SDE. Focus on the Patients - Bridging Data to Solutions, 26 July 2018, Ridgefield, Boehringer Ingelheim.

[32] Schwab, Klaus. “The Fourth Industrial Revolution: What It Means and How to Respond.” *World Economic Forum*, 2015, www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

[33] Swanger, David. “Abundance of Data + Scarcity of Patients = Clinical Complexity.” *Geeks Talk Clinical*, Medidata, 2018, blog.mdsol.com/abundance-data-scarcity-patients-clinical-complexity.

[34] “The RDF Data Cube Vocabulary.” Edited by Richard Cyganiak and Dave Reynolds, *W3C - World Wide Web Consortium*, Government Linked Data Working Group, 2014, www.w3.org/TR/vocab-data-cube/.

[35] Wells, Marshall, and Miha Kralj. “3 Reasons Why Automation Is Critical.” *DevOps.com*, DevOps.com, 21 Mar. 2016, devops.com/3-reasons-automation-critical/.

[36] “What Is Apache Spark?” *Databricks*, Databricks, databricks.com/spark/about.

[37] Wintrich, David. “Java: What Beginners Need to Know Now.” *Course Report*, Course Report, 2017, www.coursereport.com/blog/what-is-java-programming-used-for.

[38] Yegulalp, Serdar. “What Is the Python Programming Language? Everything You Need to Know.” *InfoWorld*, IDG Communications, Inc., 1 June 2018, www.infoworld.com/article/3204016/python/what-is-python.html.

ACKNOWLEDGMENTS

We would like to take this opportunity to recognize the great efforts made by the Data Engineering team within the Educating For The Future Working Group. Their contributions, conducting the research, preparing materials and designing the Squarespace portal, has made this a success. Thank you to the team!

Amy Gillespie, Beate Hientzsch, Jagdev Bhogal, Jatin Patel, Karnika Dalal, Mark Bynens, Mike Carniello, Paul Slagle, Sascha Ahrweiler, Shaaz Ansari, Vijay Pasapula, Vince Marinelli

Also a special thanks to the leaders of the Educating For The Future Working Group, Ian Fleming and James McDermott, for their leadership and to Wendy Dobson for her continued support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Beverly Hayes
Janssen R&D, LLC
1400 McKean Road, PO Box 776
Spring House, PA 19477
Phone 215-793-7358
Fax 215-986-1020
Email bhayes2@its.jnj.com

Guy Garrett
Achieve Intelligence Ltd.,
90-92 High Street,
Evesham, Worcs WR11 4EU
United Kingdom.
Phone +44-(0)-7887-954-496
Fax n/a
Email guy.garrett@achieveintelligence.com