# PhUSE

# **Future Forum**

## *Other Industry Processes*

## Contents

## Overview: Purpose of this document

The Pharmaceutical industry has processed clinical data in a similar manner for many years. The PhUSE Future Forum People & Process Working Group has been tasked with determining whether there are lessons to be learned from other industries which may help streamline or optimize data processing in our industry.  The team has written a white paper to share information on established processes, tools and technologies used for the collection, storing and analysing of data. The team has also reviewed the profile and skill set of the people using these tools or technologies to perform data related tasks. From their research, the team have identified key themes from other industries data processes for the industry to consider applying to streamline clinical data driven processes. Mechanisms for addressing the identified challenges are also presented.

## Definitions:

| | |
|---|---|
| ADaM | Analysis Data Model (CDISC) |
| CDISC | Clinical Data Interchange Standards Consortium |
| CDM | Clinical Data Management |
| CRO | Clinical Research Organization |
| EDC | Electronic Data Capture |
| MDR | Metadata Repository |
| OC | Oracle Clinical |
| PhUSE | Pharmaceutical Users Software Exchange |
| QC | Quality Control |
| QMP | Quality Management Plan |
| ROI | Return on Investment |
| SCE | Statistical Computing Environment |
| SDTM | Study Data Tabulation Model (CDISC) |
| SME | Subject Matter Expert |
| SOP | Standard Operating Procedures |
| TFL | Tables, Figures, and Listings |

## Problem Statement:

The Pharmaceutical industry is not unique in its need to collect, store and analyse data, or in the requirement to comply with regulatory requirements. Many other industries also perform these activities. For many years, we have been attempting to improve our data processes by making small incremental changes with very limited success.  The team wanted to answer the question: What could we, as clinical data scientists, learn from other industries to improve our processes to make them fit for the future?

## Background:

The working group approached this problem statement from two different angles. Firstly, a survey was created (see Appendix A) by the team and sent to a broad range of individuals across various business sectors. Secondly, a literature review was performed to evaluate a number of case studies relating to data processing.

## Survey Results

The survey was sent to individuals across a broad level of roles at organizations, consequently the detail and perspective was often times unique to each respondent. Some had a higher level perspective and answered the questions from an organizational point of view, whereas other respondents answered from a department, a team or an individual contributor viewpoint. As well as the variety of perceptions the detail and quality of the response varied widely from the ten survey responses received.

**Question 1: Industry Demographics**

The purpose of this question was to ascertain the breadth of input across industries. The breakdown is as flows:

| Sector | Number of Respondents |
|---|---|
| Financial | 3 |
| e-commerce | 2 |
| High Energy Physics | 1 |
| Transportation Infrastructure | 1 |
| public safety | 1 |
| online travel agency | 1 |
| nutritional products industry | 1 |

While the number of respondents was low, we were able to gather data across a broad sector of industries. The team decided to interpret the data from the individual in the nutritional products industry with caution, as that industry is not far removed from pharmaceuticals.

**Question 2: Volume and Type of Data**
Almost one third of the respondents (3) explicitly mention that they work with unstructured data and/or NoSQL. The team made the decision to assume that all other respondents work with solely structured data. Almost half of the respondents (4) work with petabyte size data or larger.

**Question 3: Processes for the Capture, Management, Reporting and, Sharing of Data**
Two respondents referred their answer to this question to a process of manually entering data similar to the way that study personnel enter data in an eCRF. For the other respondents, data is captured via integrations with systems or machinery outside of the remit of the respondent (i.e. they are data consumers). A number of different central repositories were described by respondents. These are used to store data which is accessible to various users for analysis, reporting and sharing. In one case, the organization additionally stores analysed data in the repository for review and exploration in addition to “raw” data.

Half of the respondents (5) focus on data accessibility when asked about reporting and sharing. This indicates that this functionality is an important aspect of data flow for these organizations. A single respondent referred to static pdf and rtf reports and one respondent referred to reports in Excel, for ease of further filtering or drill down. The remainder (3) did not answer this part of the question.

**Question 4: Data Quality Issues**
A number of respondents (3) have issues with incomplete or missing data, while other respondents highlighted challenges with duplicate data, data bias, confidentiality and data quality.

There are a broad range of answers regarding the strategy to identify issues and clean the data. It goes from senior level review to the use of data integrity or mining model tools. To use data monitoring has been reported in two cases and controls at application-level once. One company is trying to improve the CRF design and to use only experienced and high quality investigational sites (development and production of nutrition products).

**Question 5: Technology and Tools**
A broad spectrum of tools and technologies were share by the respondents to the survey. Two respondents are heavily engaged in open source tools from the Apache suite, which has emerged in the Big Data era, and are using these tools throughout the entire data life cycle. A third respondent, also reported that they use Apache Kafka for data ingestion, making it one of the most frequently reported tools in the survey.

Relational and non-relational databases are equally popular, with explicit references to SAS data warehouse, Oracle DB, Fame DB, HDFS, HBASE, Redshift, DynamoDB, MongoDB and CouchDB.

For data analysis and reporting, Excel (3), SAS (3), Tableau (2), Python (2) and AWS Elasticsearch (2) are the most reportedly used tools, but the list of different languages and tools is long, including R, Java, Spark, Hive, Impala, SOLR, SAP (incl. BusinessObjects), ESRI GIS, Cognos, Grafana, TSDB, Adobe Analytics (Omniture) and dedicated self-build analysis software.

**Question 6: Data Processing Elapsed Time**
It was difficult to draw any conclusions to the responses to this question as there was significant diversity in answers provided.  These ranged from a minimum of seconds up to a maximum of one week.

**Question 7: Utilizing Artificial Intelligence**
Very few of the respondents reported the use of artificial intelligence at their companies. Only one is using deep learning for fraud and chatbot learning & decision making. While a pilot use of machine learning was reported by another respondent no decision had been made by the company as to the continued use of these tools. No other respondents were aware of such tools currently being utilized by their company at the present time.

**Question 8: Regulations**
From the responses received it is apparent that the financial area has the most external regulations to follow (e.g. General Data Protection Regulation/GDPR, Financial Services Authority/FSA, EU regulations). The nutritional products industry is required to follow Good Clinical Practice (GCP) and data privacy laws. Finally, for the public safety area, there are several regulations according to Human Rights and Data Protection that must be followed. The remaining respondents shared that there were no external regulations that they were required to follow.

**Question 9: Profiles or Skills Needed**
Between all of the respondents we received a long list of profiles and skills in response to this question. In total, we have received 31 different profile or skill names in response to this question, of which only Data Management was mentioned twice.

The full list of reported profile and skill names is listed in appendix B and C.

## Literature Review

Various methods are being used in data collection, storage and analysis of data based on requirements of end users in every industry. Internet, connected devices, sensors are taking data collection to new heights. Internet of Things (IoT) creates a huge opportunity for data collection in today's world. Latest advances in technology are providing solutions for this rapidly growing data. Big data analytics applications enable data scientists, predictive modelers, statisticians and other analytics professionals to analyse growing volumes of structured transaction data, plus other forms of data that are often left untapped by traditional business intelligence (BI) and analytics programs. To supplement the survey results the working group also conducted a literature review. Specifically, some of the practices being followed in the entertainment industry and Insurance companies are described in this section.

### Case Study: NetFlix[1]

Netflix collects and analyses upwards of 1.5 million events per second and has developed its' Suro framework to house and process this data. Suro consists of a producer client, a collector server, and a plugin framework that allows events to be dynamically filtered and dispatched to multiple consumers.

Netflix has deployed a large number of Amazon Elastic Compute Cloud to collect its event data. This event data can be log messages, user activity records, system operational data, or any arbitrary data that their systems need to collect for business, product and operational analysis. The data is largely unstructured.

When all these events go to Suro, it produces offline business reports (batch processing) using Hadoop Cluster run MapReduce jobs. For real-time operational reports Suro uses stream clusters. Stream consumers are typically employed to generate instant feedback, exploratory analysis, and operational insights.

### Case Study: Enterprise Data Analysis and Visualization

Stanford University, interviewed 35 data analysts across different industries including healthcare, retail, finance and social networking[2]. The interviewer asked open ended questions related to their typical

[1] The Netflix Tech Blog. 2013. *Announcing Suro: Backbone of Netflix's Data Pipeline*. [ONLINE] Available at: https://medium.com/netflix-techblog/announcing-suro-backbone-of-netflixs-data-pipeline-5c660ca917b6. [Accessed 4 September 2018].

[2] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, Jeffrey Heer. 2012. *Enterprise Data Analysis and Visualization: An Interview Study*. [ONLINE] Available at: http://vis.stanford.edu/files/2012-EnterpriseAnalysisInterviews-VAST.pdf. [Accessed 4 September 2018].

daily tasks, tools they used, and the challenges they encounter.

The authors categorized the analysts into 3 archetypes: Hackers, Scripters and Application users. Additionally, they provided further information regarding proficiencies, statistics and tools.

Hackers: Doing Manipulation of large data sets at early stage before modeling.

- Proficiency: Programming (R, MATLAB, Python, Perl, SQL, Pig)
- Statistics: Less sophisticated
- Tools used: R, MATLAB, Python, Perl, SQL, Pig, Statistical packages, Tableau, Excel, PowerPoint, D3 and Raphael

Scripting: Doing Advanced Statistical Modeling and little data manipulation.

- Proficiency: R, MATLAB for data Modeling
- Statistics: Advanced
- Tools used: Applied models/functions, R, MATLAB, Tableau

Application User: Produce Visualization of smaller datasets especially charts.

- Proficiency: Business-side, Graphs, need data to be prepared
- Statistics: not much
- Tools used: SAS/JMP, SPSS, Excel, Crystal reports

The tasks within the Analysis process, based on the interviews, are:

- Discovery: Finding the data (i.e. data collection)
- Wrangle: Converting data in a desired format (e.g. integration of data from multiple sources)
- Profile: Verify data quality and its suitability
- Model: Data Summarization or prediction (e.g. summary statistics, running regression models, performing clustering or classification.)
- Reports: Reporting insights gained from Modeling to other analysts or business units.

The authors noted a number of challenges were repeated reported by the interviewees. These included, but were not limited to, the integration of data from distributed sources, data quality, visualizing data at scale and operationalizing workflows. From a pharmaceutical industry perspective, it is clear that the challenges that we are face are not unique and are also common in other industries.

As the scale and diversity of data increases, there is a need for visual analytic tools to improve the quality on analysis and speed at which it takes place. There is need for tools that simplify tasks across the analytic pipeline which can enhance collaboration among analysts and empower non-programmers to apply their statistical training and domain expertise to large and diverse dataset.

We see that the processes, tools and challenges faced across the survey respondents was similar to what we found and also what we understand to be the processes, tools and challenges within our industry.

## Case Study: A Reference Architecture for Self Service Analytics – Balancing Agility and Governance

The Eckerson report[3] focuses on the advent of self-service analytics, describing the processes involved as curate -> create -> consume in parallel to propose -> prototype -> promote. This is considered an iterative process or feedback loop creating a self-reinforcing data environment that accelerates time to insight as well as user productivity.

Roles described include Power Users (Data Scientists, Data Analysts and Statisticians) who collect and prepare the data, Casual Users (Data Explorers and Data Consumers) who consume the data and sometimes propose new ways of visualizing, analysing or reporting the data and Developers (System Analysts, Data Engineers, Business Developers and Application Developers).

They argue that the right tools can bring self-service analytics to the casual user. To avoid “all hell breaking loose” in this kind of environment, they suggest that the use of breadcrumbs (social media techniques such as tags, comments, ratings, follows) that can assist in navigation of the environment and helps users get up to speed quickly. A data governance gateway is also suggested to ensure data consistency and eliminate spreadmarts and data silos. This provides a seal of approval or watermark to show that the results have been vetted and are safe for use in decision making.

Eckerson’s reference architecture maps the users and technologies to an information supply chain that serves as the foundation for analytic processing. It supports iterative data workflows and the merge of top-down and bottom-up governance methods to ensure users get information they need quickly and with the context of governed vs ungoverned status. The report provides an interesting model together with an architecture and set of processes to support successful self-service analytics that may be applicable in the pharmaceutical industry.

[3] Eckerson Group. 2016. A Reference Architecture for Self-Service Analytics. [ONLINE] Available at: https://www.eckerson.com/register?content=a-reference-architecture-for-self-service-analytics. [Accessed 4 September 2018].

## Recommendation:

Effective access and use of data has the potential to be a game-changer in the quest to bring medicines to patients quicker, cheaper and with a higher certainty of success. Until now, the pharmaceutical industry has been attempting to improve our data processes by making small incremental changes with very limited success.

Within the Pharmaceutical industry significant investments of both time and finances have been made on the standardization of data, state of the art storage and management facilities, and high powered statistical computing environments. While we continue to use these advanced tools with out-dated processes, we fail to gain the promised benefits. These include the ability to respond rapidly to developments and events in clinical trials and in the market place, the ability to extract real-world value from the data we possess, and the ability to deliver safe and effective medicines to patients faster and in a more cost effective manner.

- Based on our survey results and the good (if not complete) alignment between the groups described in both the Standford and Eckerson reports indicate that having a dynamic relationship between those who work with and use the data is advantageous. We conclude that pharmaceutical companies would greatly benefit by incorporating such a relationship between the consumers of data/reports, the producers of data/reports and those supporting the develop of tools and systems. This would be reflected in the enabling of both a bottom-up and top-down approach to analysis and to their clinical data processes. A key component of this is to ensure teams are educated on how to use, challenge and work with data effectively.
- With the increase in visualisation techniques and availability of data static reports are no longer an efficient method of interrogating data to gain meaningful insights. By viewing data access as opposed to static reports as the end-product this would enable further visualization and deeper insights into the high quality datasets that are available.
- Data sharing and code sharing are of vital importance. It is critical that this also includes analysis efforts which have not born fruit. We should embrace automated processes for efficient data sharing and collaboration throughout the enterprise and supply chain.

## Disclaimer:

The opinions expressed in this document are those of the authors and should not be construed to represent the opinions of PhUSE members' respective companies or organizations or FDA's views or policies. The content in this document should not be interpreted as a data standard and/or information required by regulatory authorities.

# Appendices:

## A. Survey Questions

List of the initial questions provided to industry. All responses allowed for unrestricted text.

1. Please describe the high level business of your company including the industry that you belong to and whether this is regulated (controlled by government rules) or not.
2. What volumes and types of data do you typically deal with, and with which frequency is it updated (by the second, hour, day)?
3. Please describe your process in broad terms for the capture, management, analysis, reporting and sharing of data (please supply any non-confidential documentation in section XX that you think would be helpful.
4. Please tell us about the types of data quality issues you deal with including the strategies you have for identifying issues and cleaning the data.
5. Please tell us about the technology and tools you use for each process step
6. How long does it take for a data point on critical path to move through the data processing steps to be shared?
7. Do you use Artificial Intelligence (AI) for any part of your process? If yes, please provide details.
8. Which regulations, if any, do you need to adhere to when managing and processing data.?
9. What roles or job titles perform the tasks (please share job descriptions and skillsets if you can)?
10. Is there anyone else that you would recommend that we speak to (either in your organization or elsewhere)? Please provide contact details.

## B. Reported profiles and skills alphabetically ordered

| Profile / Skill | Industry | Count |
|---|---|---|
| Analyst | Infrastructure | 1 |
| Analytics | Online travel agency | 1 |
| Biostatistics | Nutritional products | 1 |
| Business Analyst | Financial | 1 |
| Clinical study management | Nutritional products | 1 |
| Clinical supplies | Nutritional products | 1 |
| Data Compliance | E-commerce | 1 |
| Data Engineer | Financial | 1 |
| Data Management | Nutritional products, onlince travel agency | 2 |
| Data Manager | Infrastructure | 1 |
| Data Steward | Infrastructure | 1 |
| Data Warehouse Administrator | Financial | 1 |
| Developer | Financial | 1 |
| Economist | Financial | 1 |
| Graduate Students | High energy physics | 1 |
| Head of Information Management | Public safety | 1 |
| Information Assurance Officer | Public safety | 1 |
| Information Officers | Public safety | 1 |
| Information Risk Owner | Public safety | 1 |
| IT Analyst | Financial | 1 |
| Machine Learning Engineer | Financial | 1 |
| Management Information Analyst | Financial | 1 |
| Medical Monitoring | Nutritional products | 1 |
| Medical writing | Nutritional products | 1 |
| PhD level Scientists | High energy physics | 1 |
| Quality management | Nutritional products | 1 |
| Records Manager | Public safety | 1 |
| Research Analyst | Financial | 1 |
| SAP super user | Infrastructure | 1 |
| Security Manager | E-commerce | 1 |
| Software Development Engineer | E-commerce | 1 |
| Statistical programming | Nutritional products | 1 |
| Statistician | Financial | 1 |

## C. Reported profiles and skills grouped by industry

| Industry | Profile / Skill |
|---|---|
| E-commerce | Data Compliance |
| E-commerce | Security Manager |
| E-commerce | Software Development Engineer |
| Financial | Business Analyst |
| Financial | Data Engineer |
| Financial | Data Warehouse Administrator |
| Financial | Developer |
| Financial | Economist |
| Financial | IT Analyst |
| Financial | Machine Learning Engineer |
| Financial | Management Information Analyst |
| Financial | Research Analyst |
| Financial | Statistician |
| High energy physics | Graduate Students |
| High energy physics | PhD level Scientists |
| Infrastructure | Analyst |
| Infrastructure | Data Manager |
| Infrastructure | Data Steward |
| Infrastructure | SAP super user |
| Nutritional products | Data Management |
| Nutritional products | Biostatistics |
| Nutritional products | Clinical study management |
| Nutritional products | Clinical supplies |
| Nutritional products | Medical Monitoring |
| Nutritional products | Medical writing |
| Nutritional products | Quality management |
| Nutritional products | Statistical programming |
| Onlince travel agency | Data Management |
| Online travel agency | Analytics |
| Public safety | Head of Information Management |
| Public safety | Information Assurance Officer |
| Public safety | Information Officers |
| Public safety | Information Risk Owner |
| Public safety | Records Manager |

## Acknowledgements

Sam Warden, Dirk Engfer, Sunil Gupta, Thomas Ellebaek, Brigitte Bernhoff, Amit Jain, Anders Vidstrup, Uday Dodla, Chris Price, Wendy Dobson

Apologies to contributors/reviewers that we may have missed.

## Project Leader Contact Information:

Sam Warden
d-Wise Europe
Geneva, Switzerland
+41 78 721 1970
sam.warden@d-wise.com