PhUSE 2014

Paper TT07 - Draft

Big and Bigger: A Big Data Project in the Pharmaceutical Industry, and Emerging Trends Beyond

Dimitri Kutsenko, Entimo AG, Berlin, Germany

Table of Contents

Introduction

More and more pharmaceutical companies are looking into Real World Evidence data as well as other data sources such as claim, financial, safety, and trial-management data, and trying to integrate this information into clinical analysis. Sponsors and regulatory authorities increasingly are investing in novel technologies and systems capable of processing large data volumes. The main goal of this paper is to provide a short overview of the current situation in the pharmaceutical industry regarding Big Data. After a brief summary of the major characteristics and driving forces behind the increasing interest in Big Data in the pharmaceutical industry, the author will guide the reader though a project in the area of Real World Evidence data that the author has been actively involved in. The key insights and developed best practices from this project as well as observations from other ongoing projects in this area will be touched upon. Encountered technological and human challenges related to large volume data, as well as their solutions and emerging trends inside and outside of the industry will be discussed at the end of the paper.

Big Data Characteristics

Before starting a tour through the Big Data project announced in the introduction, a fundamental question will be addressed: How big is "Big Data" and how can it be defined? A brief look across industries will help us to answer the first part of this challenging question.

My personal career has been inseparable from Big Data. I first became involved with it back in 2001 when working in the mobile telecommunications industry. At that time the mobile industry was preparing the roll-out of 3G mobile networks and collecting a vast amount of information from the networks under construction for testing and benchmarking. A few years later, I had a chance to work in the airline industry. By that time, the airline industry had collected over 30 years' of traffic and business data, which was managed in powerful data warehouses. The data aggregated online by high-performance queries over diverse data marts was massively used for analysis, reporting and steering purposes – and I was responsible for defining such queries to crunch data cubes.

Newer evidence from other industries is even more striking. Several years ago, Ebay was supposed to be managing and utilizing for searches, consumer recommendations and merchandising around 90 petabytes (PB) of data in data warehouses and Hadoop clusters [1]. Facebook claimed recently to have one of the largest data warehouses in the world, which stores more than 300 PB of data, used for a wide range of applications such as machine learning, graph and real-time interactive analytics [2]. Though genomics and other "-ics" in the pharmaceutical industry reliably generate data volumes that are competitive with other industries, the clinical area is rather lagging behind, and is just getting used to volumes of terabyte size.

In the quest to find a well-formed definition of Big Data, the fact stands out that even Wikipedia fails to provide a concise wording. The description is scattered over many pages, potentially indicating the ambiguity of the subject. Gartner coined with its definition the "3Vs" model for describing big data in three dimensions: **"***Big data is high-<u>volume</u>, high-<u>velocity</u> and high-<u>variety</u> information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."* [3]

Over time, several additional "V" dimensions have been suggested in the literature to describe Big Data, e.g. "Veracity" and "Variability", "Validity" and "Volatility" [4]. Recent voices suggest that the cost-reducing use case needs to be included in the description of the "Value" dimension [5].

All dimensions lead in the fact that Big Data management can become a very cumbersome process, especially when large volumes of data come from multiple, heterogeneous sources. These data need to be linked, connected and merged in order to deliver meaningful results, thus leading to a situation which can be termed as "complexity" in regard to Big Data.

An additional interesting aspect can be extracted from the Wikipedia text on Big Data: "Big data usually includes datasets with sizes beyond the ability of <u>commonly used</u> software tools…" [6] The phrase highlights the point that the big data size is relative, a constantly moving target, and is always industry-specific.

Joining together all the above-mentioned dimensions, we can express Big Data with the formula:

Big Data = f(Volume, Velocity, Variety, Variability, Veracity, Validity, Volatility, Value) -> Complexity

Big Data in the Pharmaceutical Industry

Why are we observing such a burning interest in Big Data in the pharmaceutical industry? What are the driving factors making such a previously rigid community explore new technological frontiers? Just a decade ago, companies in the pharmaceutical industry used to be protected castles surrounded by high stone walls and a complex network of heterogeneous, partly home-grown and partly purchased systems inside, without or with only limited interfaces in-between due to the lack of standardization.

From numerous articles on this topic, it seems that the following are among the most influential factors accelerating Big Data adoption in the pharmaceutical industry:

- Pressure for faster innovation with controlled risks
- The interconnected world with fast-spreading diseases
- Outcomes-based reimbursement
- Protection from drug theft
- Better monitoring of social media for brand protection

In order to address these factors, additional sources need to be taken into consideration. The following, certainly not exhaustive list presents major sources of Big Data relevant in the pharmaceutical industry [7]:

- Investigator and patient portals
- EHR systems
- Insurance and claim data
- Prescription data
- Safety data
- PK/PD data
- Lab data
- ePRO and EDC data
- Imaging data

Though social media are typically not listed in the current sources of Big Data in the pharmaceutical industry, they are certainly a good candidate for the future.

Many of them induce a significant load on the backend, which makes it impossible to process them with existing tools, architectures and approaches.

A Big Data Project

Preface

Entimo AG – my current employer - is known in the pharmaceutical industry as a highly specialized, product-focused software company and as a provider of software applications for clinical and pre-clinical R&D processes. Entimo has developed the Integrated Clinical Environment (entimICE) – a metadata-driven, configurable and regulatory compliant solution platform for life sciences. The product suite enables the customers to implement an end-to-end process from standards governance and study setup to statistical analytics and submission, and to do so in a validated and reproducible way. In order to address the challenge of Big Data and to enable its efficient processing, Entimo entered a strategic partnership with Teradata - the acknowledged industry leader in the area of Big Data processing and analytics. Teradata's Data Warehouse Appliance enables flexible, in-database analytics with extremely fast parallel processing, scalable to cope with massive data volumes. Entimo has integrated the Teradata database as a backend for its product platform. The ultimate goals of the combined solution are to handle extremely large data volumes without compromising governance, apply strict control of access rights, and provide the flexibility to use different analytical and reporting tools.

Project Initiation

At the beginning of 2014, a group from a major pharmaceutical company responsible for Real World Evidence research initiated a project with Entimo. The main project goal was to implement a unified system to manage the business workflows including program code and data for research and analytics. Entimo's main product, entimICE, delivered a framework for integration of Teradata, SAS and other tools in controlled and traceable workflows. It provided annotation capabilities, metadata management, search, indexing, auditing, security, and versioning. These features will facilitate full process compliance with the regulatory requirements of the pharmaceutical industry. entimICE drives and controls all workflows and serves as a broker between integrated analytical components, including SAS and the Teradata database. In addition, the requirements to "enable collaboration, knowledge sharing, and reuse of artifacts" and "handle data volumes of terabyte size with good performance (e.g. to identify and retrieve patients, process data, retrieve relevant data from sources)" were defined as project success factors.

An agile approach was chosen in the project to allow for feedback cycles. The project objectives were broken down into phases to install checkpoints throughout the project.

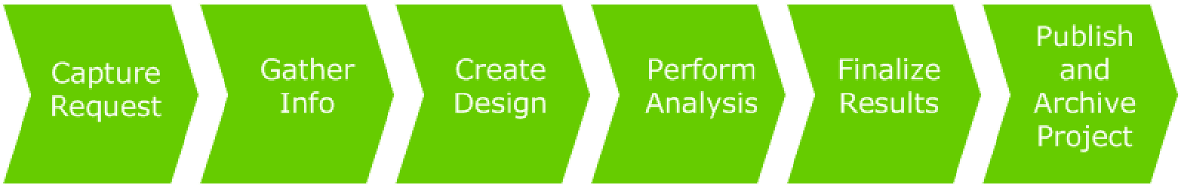
In addition, the project team was asked to obtain performance benchmarks in consideration of the governance overhead introduced and to work out best practices and effective coding methods within the integrated environment.

Process

Several important characteristics make projects in this RWE group unique. The variability of projects is extremely high; a project can last from one day to many months with project requests coming from various stakeholders such as clinical, medical, marketing and manufacturing among many others. Project outputs also vary, ranging from data to graphics and reports, depending on the requirements of the stakeholders.

The following illustration shows a top-level process which takes place within a project.

Note: The process will provide just a rough impression of the project flow and will not be described in detail in this paper.



Due to the process specifics, especially project variability, metadata (or project attributes) was found to play a key role in achieving project objectives. The entimICE product suite includes a flexible metadata layer which supports configuration of arbitrary metadata attributes at different levels including projects. The metadata attributes are searchable and allow fast access to historical projects in the RWE domain.

The following table gives selected examples of the metadata attributes which were identified in the project scope:

Level	Metadata Attributes
Request	Project Status, Project Description, Project Type, Rationale, Research Question, Business Purpose…
Design	Protocol, Study Objectives, Study Rationale, Study Design, Population, Data Collection Methods, Sample Size/Power, Limitations/Strengths…

Analysis	Drug Codes, Diagnoses, Treatment Patterns, Conditions, Trends, Code Comments, Methods Used…
Outcomes	Methods, Code…

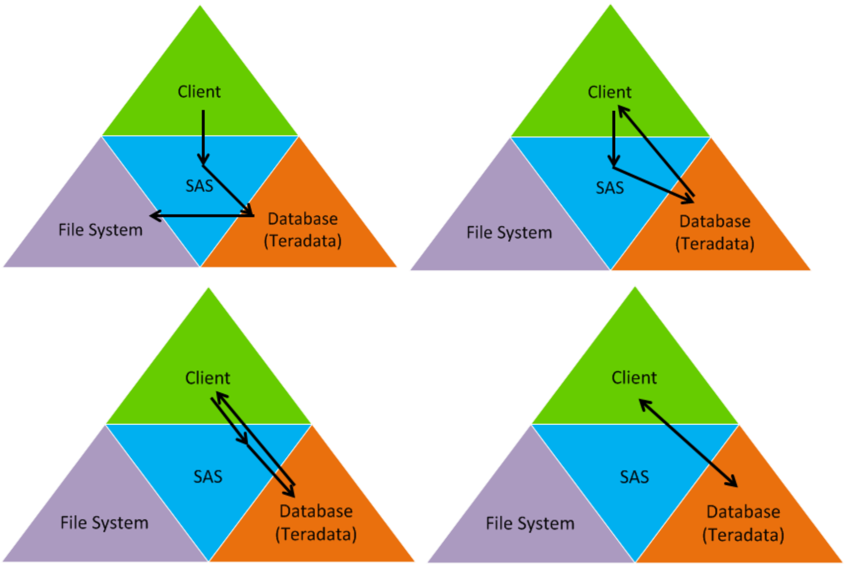
Analytical Approaches and Insights

In order to work out guidance on effective coding methods, Entimo suggested several test designs which were benchmarked in the project against the baseline – a pure SAS/Teradata infrastructure with the governance layer. The key test question was very simple and can be paraphrased as "benchmark the time needed to extract an increasing number of records from the Teradata repository for different access methods and write-back options".

The following test designs were compared for data based on the OMOP data model with up to 10 million records:

1. Run a SAS call using PROC SQL and store the result in a new SAS dataset in the file system.
2. Run a SAS call using PROC SQL and store the result as a new Teradata table.
3. Run a SAS call using PROC SQL with "SQL pass-through" and store the result as a new Teradata table.
4. Run SQL without SAS interaction directly via the Teradata interface.

The used test designs are illustrated in the following graphic:



Surprisingly or not, the assessment of the test results showed that:

1. The winner was #4.
2. As #3 supports almost all SQL flavors, the results were comparable with #4 with few exceptions.
3. The options #2 and #4 achieved comparable results if SAS "knew" specific SQL statements within the query. As #2 does not support all SQL flavors, in some cases the performance was degraded.

Note: I deliberately do not mention the exact benchmarks– in my experience, you will benefit from them only if you know and have exactly the same infrastructure. Otherwise, the numbers will only mislead you.

Project Achievements

As I am writing this paper, the first major project milestone was achieved: the system went live after about six months of intensive work. The project achievements can be grouped with regards to the project repository, analysis library and project tracking.

On one side, analysts in the RWE group with the delivered solution are now able to search AND FIND previous studies and projects to leverage the work done. The project and program ontology helps them constantly improve search results and thus reusability. The group has one centralized, controlled place to store, retrieve and later archive project documentation such as protocols, analysis plans, programs, analysis datasets, results and other project-related documents. All projects follow the same defined workflows; they demonstrate traceability and reproducibility, thus fulfilling regulatory requirements.

At the analysis level, the analysts maintain the central, reusable library of code lists and standard listings as well as templates of table shells. They design type-dependent project templates which make the setup of new projects easy and efficient.

Project managers are now able to track and document project decisions in real-time, making the whole process very transparent.

In general, the results obtained during the roll-out phase are very promising – time will be the best judge!

Constraints

It would be one-sided to talk about the achievements without mentioning bumps on the road. One could roughly classify constraints as related to technology and human (or social) factors.

The key technological factor determining the solution performance in the used architecture is certainly the network proximity of the solution components – the client farm, database cluster, SAS servers and the file server cluster (this fact is obvious to technical people, but should be mentioned in a business paper). In addition, the used storage concept has a significant effect on the overall solution performance: shared storage (SAN, NAS) has slower performance in comparison with direct-attached storage types (SSD, SATA). However, state-of-the-art data centers allow both concepts to be combined for frequently used vs. rarely used disk clusters, thus eliminating the latter disadvantage.

The tests described above confirmed that certain coding approaches are superior to others. The consequence of this finding is obvious. In most cases when SAS is used, the presented solution allows in-database capabilities to be used without changing the coding practice, which is certainly a big advantage. However, with changing research questions and data sources (e.g. usage of non-structured sources from the social media), a different skill mix might be required: SQL skills (e.g. PL/SQL or SQL/H) could become dominant in the future in comparison with SAS-centric approaches to underpin the implementation of RWE studies with the maximum performance. Consequently, educational strategies need to be developed to maintain the research capacity, enhanced by continued development of programming skills.

Recent Trends in Big Data Processing

In the unpredictable world of technological innovations, several major trends have recently been taking shape, mainly outside of the pharmaceutical industry. On one side and with improving interoperability, grid architectures of various kinds have become popular for resource-consuming tasks. Transactor-based architectures which abolish the transactional logic (e.g. Scala/Akka) and NoSQL Databases based on Key-Value, Graph and Document structures are blazing a trail in the technological landscape. The major players such as Facebook and Amazon, among many others, are successfully utilizing the MapReduce algorithms (e.g. based on Hadoop). Hadoop MapReduce and Hive are designed for large-scale, reliable computation, and are optimized for high data throughput volumes.

A newer development by Facebook is the quasi open source project "Presto" which is aimed at overcoming the limitations of technically sequential execution of MapReduce tasks, especially related to input/output (I/O) operations [2]. In the MapReduce algorithm, each task reads inputs from disk and writes intermediate output back to disk – what a waste! In the Presto architecture, all processing is done completely in-memory and is pipelined across the network. This avoids unnecessary I/O operations and associated latency. In addition, Presto was designed with a simple query abstraction which makes it possible to use with disparate, non-structured data sources.

New Trends in the Pharmaceutical Industry

The global trends in technologies processing Big Data have delivered very promising results. Why have they struggled to enter the pharmaceutical arena?

Among the many reasons is certainly the technological reliability required in clinical trials, along with such requirements as reproducibility and traceability. The pharma world would be technologically more advanced, but simultaneously ethically more questionable, if we could adopt any new technology without validation. NoSQL databases can be very fast, but lack the robustness of relational databases. Learning algorithms can deliver improved results, but can hardly be validated due to the evolutionary component. It will take time to prove the reliability of such technologies which are involved in decisions related to human lives.

However, the situation is not hopeless - there are positive signs on the horizon.

Roche talks a lot about "social analytics" and "data exploration" - the analysis and detection of meaningful structures within unstructured texts [8]. Astra Zeneca (AZ) strives to be "the industry leader in a payer-driven world" with the increased usage of Real World Evidence data [10]. Numerous alliances have been forged between industry leaders: AZ and Roche, GSK and Pfizer, AZ and BMS, TransCelerate to mention a few [9]. More are certain to come.

However, the most prominent evidence of an opening up toward Big Data technologies is the Nextgov initiative by the FDA. Recently, a tender was conducted for a project to "to crawl 20 million biomedical journal abstracts and citations housed on a National Library of Medicine database to uncover drugs that are disproportionately associated with "adverse events."" Another initiative by the FDA is "… a program to monitor the Web for illegal sales of drugs, medical devices, cosmetics and veterinary products as well as counterfeit food and adulterated or misbranded vaccines." [11] Following the example of the American regulatory authority, the UK government has published its "Vision for Real World Data – Harnessing the Opportunities in the UK" [7].

Conclusions

Big Data practitioners have faced much criticism over the last few decades. They have been blasted among many other things for lacking robust scientific methodology, and for insight deficits due to inadequate skills in understanding Big Data. It is said that Big Data analysis can tell us about "the world as it was in the past, or, at best, as it currently is". Its ability to predict the future is limited because of changing systems' dynamics.

Despite the criticisms, Big Data is successfully continuing its march in the pharmaceutical industry, promising better informed decision-making based on new data sources and performance dimensions. What does the technological future of the pharmaceutical industry look like? In my opinion, this is the point of singularity nobody can currently predict. But certainly, Big Data will always be there, and its data sources will increase in their volume and variety.

References

1. http://www.itnews.com.au/News/342615,inside-ebay8217s-90pb-data-warehouse.aspx (accessed on 2014/10/10)
2. https://www.facebook.com/notes/facebook-engineering/presto-interacting-with-petabytes-of-data-at-facebook/10151786197628920 (accessed on 2014/10/10)
3. http://www.gartner.com/it-glossary/big-data/ (accessed on 2014/10/10)
4. http://inside-bigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/ (accessed on 2014/10/10)
5. http://blogs.sap.com/innovation/big-data/2-more-big-data-vs-value-and-veracity-01242817 (accessed on 2014/10/10)
6. http://en.wikipedia.org/wiki/Big_data (accessed on 2014/10/10)
7. The Vision for Real World Data – Harnessing the Opportunities in the UK, White Paper, abpi, September 2011
8. http://www.roche.com/de/media/roche_stories/roche-stories-2014-07-21.htm (accessed on 2014/10/10)
9. http://online.wsj.com/news/articles/SB10001424127887323998604578567682985587790 (accessed on 2014/10/10)
10. Keohane P. (2011, November 7). The reality of 'Real World Evidence'. Presentation for ISPOR.
11. http://www.fedtechmagazine.com/article/2013/04/fda-launches-three-big-data-initiatives (accessed on 2014/10/10)

Abbreviations

BD – Big Data
entimICE – Entimo Integrated Clinical Environment
NAS – Network-Attached Storage
SAN – Storage Area Network
SATA – Serial Advanced Technology Attachment
SSD – Solid State Drive
RWE – Real World Evidence

Contact Information

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

Dimitri Kutsenko
Director, Business Development
Entimo AG | Stralauer Platz 33-34 | Berlin 10243 | Germany
Email: dku [at] entimo.de
Work Phone: +49 30 520 024 100
Web: www.entimo.com