PhUSE US Connect 2018

Paper AB13

Augmented Real-time Clinical DataMart

Phani S Srinivasan Ponnapalli, Syneos Health, New Jersey, USA
Subrahmanyam Rayaprolu, Syneos Health, New Jersey, USA

ABSTRACT

BIG DATA takes center stage in the current Digital World and is driving new opportunities across industries. In recent years, pharmaceutical companies started embracing digital information to optimize the time from "Lab to Life". The exciting challenge is to define a strategy and architecture to effectively handle and combine heterogeneous Big Data with traditional clinical data. Digital capabilities combined with rapid consumer adoption of wearable devices have endless possibilities to revolutionize clinical trial operations. Data from these new sources and social media combined with Traditional clinical data can augment analysis by discovering hidden patterns and produce actionable insights. "Augmented Realtime Clinical DataMart" (ARCDM) is a term we are coining for this approach. This paper explores the strategy to build a Data Lake (an architecture to collect vast pools of data from varied sources) and bridge a connection to traditional clinical data, driving new cognitive analytic framework.

INTRODUCTION (ARCDM)

Digital capabilities are disrupting multiple aspects of business and providing unparalleled opportunities across industries. In recent years, the Pharmaceutical industry has recognized this tremendous opportunity and started embracing digital information to optimize the time from "Lab to Life".

The rapid consumer adoption of wearable devices for the collection of health data is laying the foundation for the next revolution in clinical trial operations.

Clinical trial models based on the integration of digital capabilities demonstrate compelling benefits, including:

- Real-time patient monitoring and adherence
- Improved patient retention and compliance through convenient e-alerts and information sharing
- Cost reduction by reducing clinical visits
- Discovery of hidden patterns and actionable insights

Applying cognitive analytics on integrated Digital Data and Traditional Clinical Data will provide a new dimension of possibilities, including hidden patterns and actionable insights. "Augmented Realtime Clinical DataMart" (ARCDM) is a term we are coining for this approach.

This paper would explore the strategy for creating ARCDM and analytics framework in the following steps.

- Compare Traditional Clinical Data and Digital Data
- Create Modern Data warehousing architecture for Digital Data called "Data Lake"
- Bridge connection between Data Lake and Traditional Clinical Data by "Augmented Realtime Clinical DataMart"
- Apply Cognitive analytics on integrated information

TRADITIONAL CLINICAL DATA WAREHOUSE

Clinical Data Warehouse is a database that consolidates data from a variety of clinical sources to present a unified view for a single patient. Clinical data management and electronic data capture (EDC) systems often provide much of the source data, along with third parties such as laboratories, ECG, IVRS systems, PK/PD and many other source systems. In clinical research, one critical task is to ensure that the data collected or data entered into the database is valid, correct, and logically sound. Data entered should pass through some checks before it can be considered ready for analysis. Edit checks are a great mechanism to help improve data quality within an EDC system. For data from

other sources, the level of data checking may need to include metadata checking, data consistency checking and reconciliation checks.

Clinical Data Landscape has changed rapidly from past few years. Pharmaceutical and Biotechnology companies began to define companywide Data standards for clinical trials. As an Industry, it has realized the importance of common standards. Clinical Data Standards Consortium (CDISC) in collaboration with FDA and Pharmaceutical companies has been working very hard to create and adopt Industry standards to optimize the drug development. The main aim of these standards is traceability, ability to combine data across clinical studies and also in the reusability of tools. The challenge in implementing these standards lies in streamlining the different Data perspectives of data managers, clinical programmers, statistical analysts/statisticians and clinicians.

CDISC developed a series of standards, which include Operation Data Model (ODM), Study Data Tabulation Model (SDTM) and the Analysis Data Model (ADaM).

DIGITAL DATA

Clinical Data is experiencing another wave of change with the advent of digital data from health and medical devices. Mobile health apps are growing in popularity. These new data sources are very different from traditional data and are evolving at a much faster pace.

Over 318,000 health apps and 340 consumer wearable devices are now available worldwide. The value of these Digital Health tools to human health and the healthcare system is still evolving. Over 55% of the most downloaded health apps now use sensor data, with significant adoption of consumer wearables like Fitbit and Jawbone for wellness management being a key driver of the era. Digital sensors linked to health apps are bringing innovation in three key areas: the creation of smart devices, digital diagnostics, and new human-centered clinical trial designs.

In recent years, many pharmaceutical companies started embracing new forms digital information to optimize the time for “Lab to Life”. Digital capabilities provide opportunities to improve clinical trial processes—from trial design, through trial start-up, to trial conduct and closeout. Wearable devices for the collection of health data are laying the foundation for the next revolution in clinical trial operations. Anticipating a digital future, major technology companies are also entering the clinical research space. Modern Clinical Trials and Research initiatives are incorporating these devices into ever more elaborate clinical trials focusing on Real World, Healthcare and Big Data sources.

To bring new therapies to patients over the next decade, Life Sciences organizations have to define a new strategy and process to integrate this digital data for more insights and to transform as new health solutions.

Life Sciences organizations are on a challenging path now to define answers for the following questions to handle these new sources of data.

A) HOW DIFFERENT IS DIGITAL DATA FROM TRADITIONAL DATA?

This digital Big Data is characterized by the 3 Vs:

- Volume – number of transactions / complexity
- Velocity - number of transactions / time
- Variety – disparate data types and formats

Example:
In traditional clinical data there will be one record per visit which is taken at site. Medical Devices are available as patch to a patient’s body that can monitor ECG every minute.

B) CAN WE HANDLE DIGITAL BIG DATA WITH TRADITIONAL DATA MODEL?

The Data Warehouse is typically used to store historical data (partially sourced from the Operations Data Store). Queries are usually subject-based and deep analytics are used to help gain business value out of integrated data sources and possibly external data sources as well.

Digital Data being mostly unstructured / semi-structured do not fit the enterprise relational data warehouse model. A new operational process has to be defined with the following focus areas to handle Big Data:

- Identify and use tools that easily manage unstructured data. Tools need to provide efficient text parsing and analytics, taxonomy and metadata management.
- Integrate unstructured data with existing information systems.
- Store and process data in volumes too large for a traditional database.

BIG DATA ARCHITECTURE

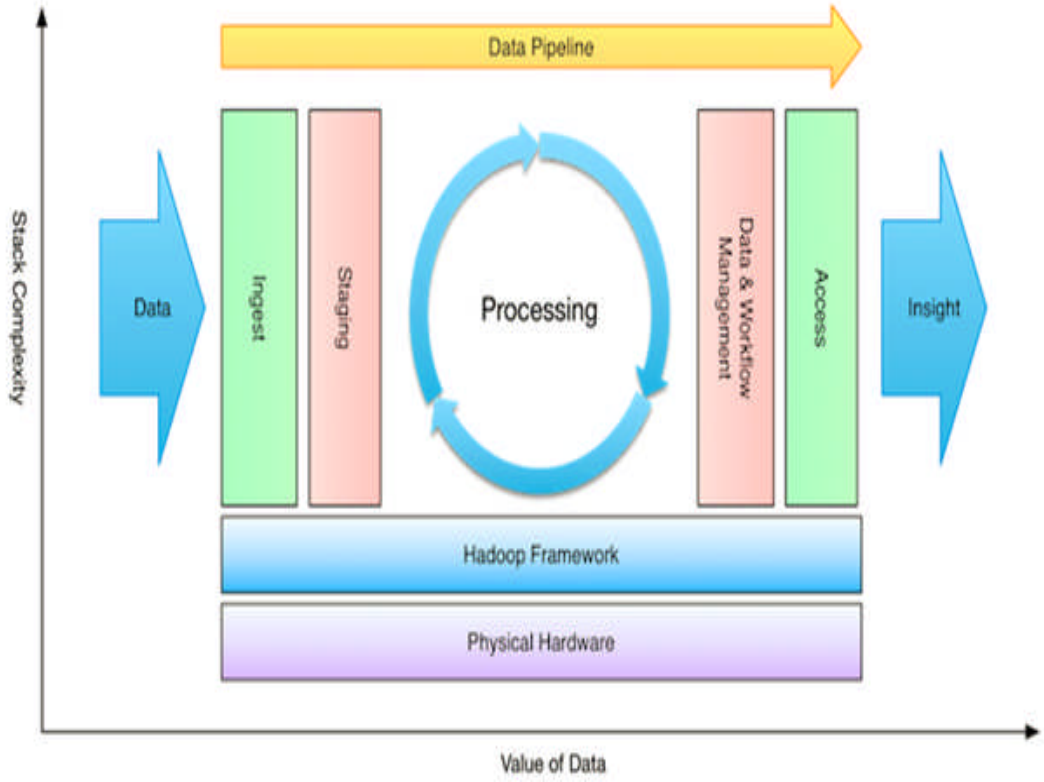
Big Data architecture is designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database systems. Big data solutions work on a fundamentally different principle to handle device data and streaming data. Big Data architecture has the following characteristics.

- Distributed Data and parallel processing: Big Data solutions store huge data in a distributed manner in a file system. Process the data in parallel on a cluster of nodes. In simple words, Data is broken in to small pieces of information and stored in multiple blocks called HDFS. Tasks are executed in parallel by processing these blocks in parallel and results are merged back, is called map reduce.

- Fault Tolerance: Big Data solutions work on failure to tolerance by redundancy. The same information is stored in multiple places called racks. There are multiple machines available as cluster for processing. If any machine in the cluster goes down still system works due to data redundancy and multiple machines.

- Scalability: Big Data systems are very flexible in scaling storage space and computing power on fly whenever required.

- Cost effectiveness: Big Data systems like Hadoop are open source and uses commodity hardware. They do not require a very high-end server with large memory and processing power. This makes the system very cost effective.



PhUSE US Connect 2018

DATA LAKE

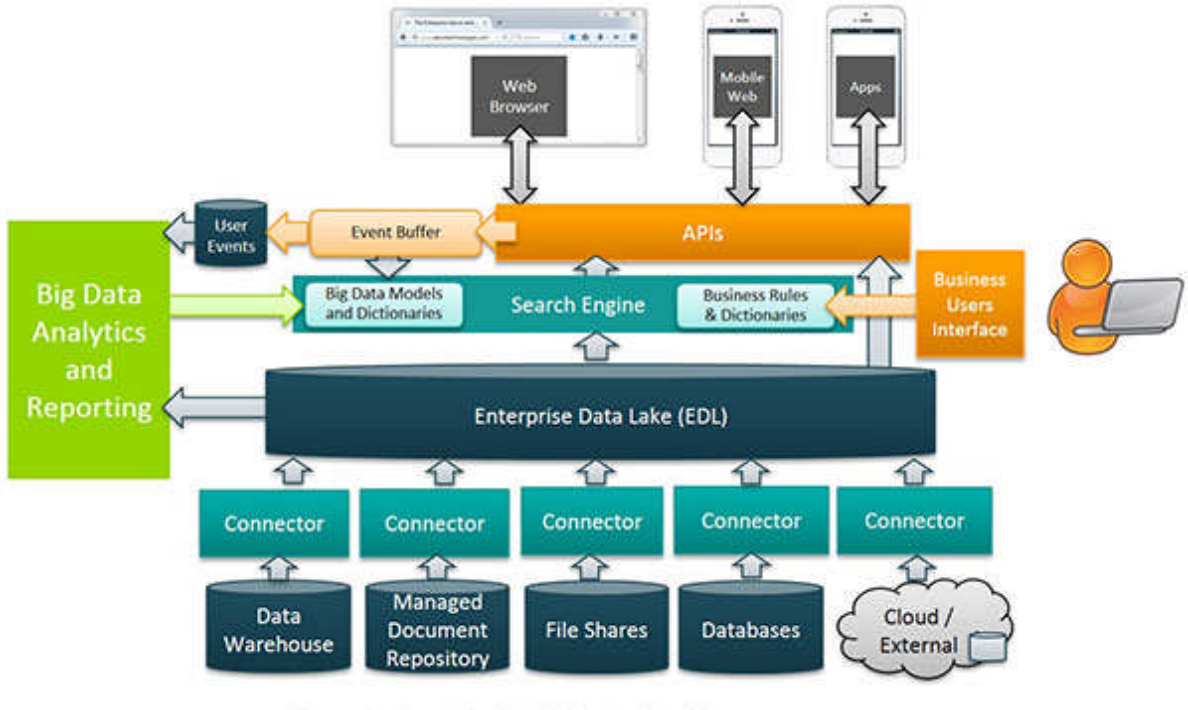
A data lake is a large storage repository that holds a vast amount of raw data in its native format until it is needed. An "enterprise data lake" (EDL) is simply a data lake for enterprise-wide information storage and sharing.

The advent of Big Data has also given rise to the concept of a "Data Lake" which is a storage repository that can hold vast amounts of data in raw data format including structured, semi-structured and unstructured data. A data lake accommodates data of varying structures, which are able to resolve at the application delivery level, i.e. when the data structuring requirements are known.

Data lakes are considered as a re-creation of data warehouses in the Big Data era. Data lakes are significantly different from warehouses, in terms of the ways they structure and manage data. The main difference is that data warehouses deal with structured data only, while data lakes store raw data and transform it to some structure when it is time to use the data (e.g. as part of an application). Also data lakes are closely affiliated to Big Data technologies as in most cases they leverage tools and techniques from the Hadoop/Big Data ecosystem. Finally, data lakes provide agility in terms of their processing. Data schemas can flexibly change, while data warehouses adhere to given schemas in order to benefit from well-structured data.

DATA LAKE ARCHITECTURE

All content will be ingested into the data lake or staging repository and then searched using a search engine. Where necessary, content will be analyzed and results will be fed back to users via search to a multitude of UIs across various platforms.



PhUSE US Connect 2018

BRIDGE CONNECTION BETWEEN DATALAKE AND TRADITIONAL CLINICAL DATA

The exciting challenge is to define a strategy and architecture to effectively handle and combine heterogeneous Big Data with Traditional Clinical Data.

COGNITIVE ANALYTICS

Cognitive analytics offer a way to bridge the gap between Big Data and the reality of practical decision making.

Cognitive analytics applies human like intelligence to certain tasks, such as understanding not only the words in a text, but the full context of what is being written or spoken, or recognizing objects in an image within large amounts of information. It is a data forward approach that starts and ends with what's contained in information. This unique way of approaching the entirety of information (all types and at any scale) reveals connections, patterns and collocations that enable unprecedented, even unexpected insight.

Applying this Cognitive Analytics framework, a hybrid information layer is created from the data lake that could be integrated with a traditional data warehouse for discovering hidden patterns and produce actionable insights, thus the "Augmented Real-time Clinical DataMart" (ARCDM).

We suggest a three step process for building an ARCDM that is outlined below:

1. **IDENTIFY THE QUESTIONS TO BE ANSWERED**

Analytics is about asking—and answering—smarter questions.

What questions should you are asking to drive more value in your organization? What information is needed to answer these questions? Which tools can efficiently deliver the answers? Clinical Data is complex and diverse, Pharma Organizations have to identify and prioritize the questions trying to be answered.

For example:

- In a typical clinical trial, each patient responds differently to the same drug. Pharma company's challenge is to find a pattern associated with individual characteristics.
- Patient enrolment for a rare disease.
- Plan for monitoring trials in real time to identify signals requiring action to avoid significant and potentially costly issues.
- Understand how individuals are thinking about the marketed drug.

2. **FISHING THE DATA LAKE**

Data Lake and Big Data Ecosystems contain tons of data in raw format. We need to extract the relevant information with which we will be able to answer the questions which are not available in traditional data warehouses. Fishing in a lake and a data lake are much the same. Data scientists must not only go where the fish are for Big Data insights, but also find a way to quickly build the data pipeline that turns raw data into usable business results.

In this step, Data Lake is queried with Big Data tools like Cloudera to extract relevant information that help you find correlations and patterns you may not have seen before to assist with decision-making processes.

3. **AUGMENTED REALTIME CLINICAL DATAMART**

"Augmented Realtime Clinical DataMart" (ARCDM) is a term we are coining to define a data abstraction layer which combines the traditional clinical data and the data fish we extracted. Selectively move data that are deemed to have the highest business value in a complementary warehouse.

In this step, Information extracted from the data lake is curated and massaged. Later it would be augmented with traditional clinical Data Warehouse to create new analytic layer of information.

PhUSE US Connect 2018

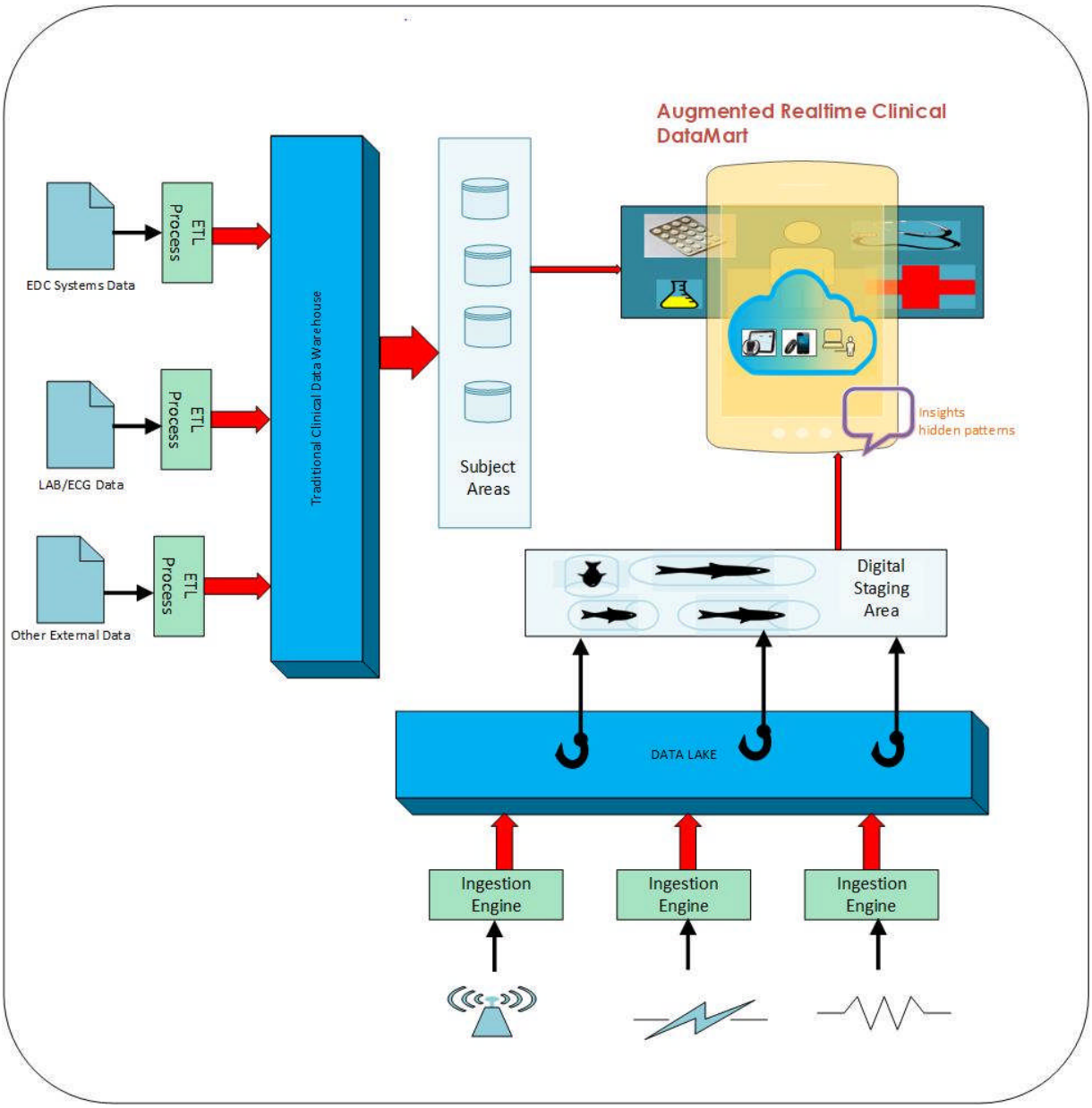
FEATURES OF ARCDM

- Centralized data repositories streamline translational clinical research
- Provide relevant data to researchers for research
- Facilitate storage and maintenance of datasets
- Promote collaboration

BENEFITS OF ARCDM

- Discovering hidden patterns
- Producing actionable insights
- Augmenting traditional data with real time digital data
- Answering the open questions

ARCDM ARCHITECTURE



CONCLUSION

The vast volume of digital data generated by health devices and sensors will be transformative across the entire health care spectrum, from wellness and prevention to treatment and research. During the next five years, Big Data and AI technologies will mature to enable advanced research models, including cloud-based health databases of continuously uploaded patient data and Internet-based trials conducted remotely. ARCDM is one of the possible strategies to combine traditional clinical data and this new form of enormous data. Cognitive analytics framework on digital data augmenting the traditional data will open doors for a new era of clinical research.

REFERENCES

1. http://blog.cloudera.com/blog/2014/09/getting-started-with-big-data-architecture/
2. https://www.searchtechnologies.com/blog/search-data-lake-with-big-data
3. https://blogs.informatica.com/2016/03/30/fishing-for-big-data-insights-with-an-intelligent-data-lake-2/#fbid=4g3xvtj52zd
4. http://www.itexchangeweb.com/blog/fishing-in-the-big-data-lake/
5. https://itknowledgeexchange.techtarget.com/iot-agenda/gone-fishing-in-iot-data-lakes/
6. http://support.sas.com/resources/papers/proceedings10/183-2010.pdf
7. http://www.oracle.com/us/industries/life-sciences/guide-clinical-data-warehouse-ar-1563726.pdf
8. https://www.cognizant.com/perspectives/information-architecture-trolling-in-the-data-lake-get-yourself-a-fish-finder
9. https://www.analyticsvidhya.com/blog/2013/07/big-data/

ACKNOWLEDGMENTS

The authors would like to thank Carol Frazee, Associate Director, Syneos Health, for her valuable comments and feedback.

CONTACT INFORMATION

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

Name: Phani Ponnapalli
Enterprise: Syneos Health
Work Phone: +1 732.666.8572
E-mail: phani.ponnapalli@syneoshealth.com

Name: Subrahmanyam Rayaprolu
Enterprise: Syneos Health
Work Phone: +1 972.370.3552
E-mail: subrahmanyam.rayaprolu@syneoshealth.com