Paper CM07

Evolution of Data: From extraction to visualization

Ashish Kumar, Alkermes, Waltham, USA
James Dale, Alkermes, Waltham, USA

WHERE'S MY APPLE?

If you want to eat an apple pie you can either go to a farm to pick the apples right from the tree or go to a grocery store and buy the apples. These days the advent of ecommerce has ensured that you can even have the apples delivered to the door while you preheat the oven. You don't want to bother yourself with the process of cooking? Sure, just buy the entire apple pie then!

Are there reasons why we should or should not do the same with our clinical data?

INTRODUCTION

This paper will cover the data integration and visualization journey supporting clinical data review. All clinical trial data gets ingested, transformed, split, or merged and grouped with other similar data so that clinical or statistical decision making can be faster, efficient and more accurate. There are multiple tools that can and have been used for these activities by different sponsors and CROs.

We will focus on how Alkermes uses Knime, Python, SAS, MoveIT and Tableau to help our clinical study teams through this journey. This paper will highlight key aspects and features that we should look for in our ETL (Extract Transform Load) / ELT (Extract Load Transform) and visualization tool of choice. We will cover and showcase business cases across different clinical teams that we helped including Data Management,
Drug Safety & Pharmacovigilance, Clinical Operations and Medical Monitoring.

In this paper I will purposely shy away from any comparison of ETL or visualization tools. Instead will focus on what features of the chosen tool or technology Alkermes uses and how.

WHERE WE WERE

Data Managers are responsible to ensure all clinical data is collected, cleaned and made available to the clinical study teams so that on-going review and decisions on safety and efficacy of the drug can be made. Many times, the data managers end up managing the dataflow and issues all the way until it reaches the statistician or the medical monitor. This includes downloading the data directly from the source systems, copying them to the SAS server and troubleshooting back and forth with vendors when some files are not sent / received on time or if there are issues with the data.

Once all data is collected the data manger will notify the programmer and he/she will start working on converting the data into the required dashboards, listings or reports.

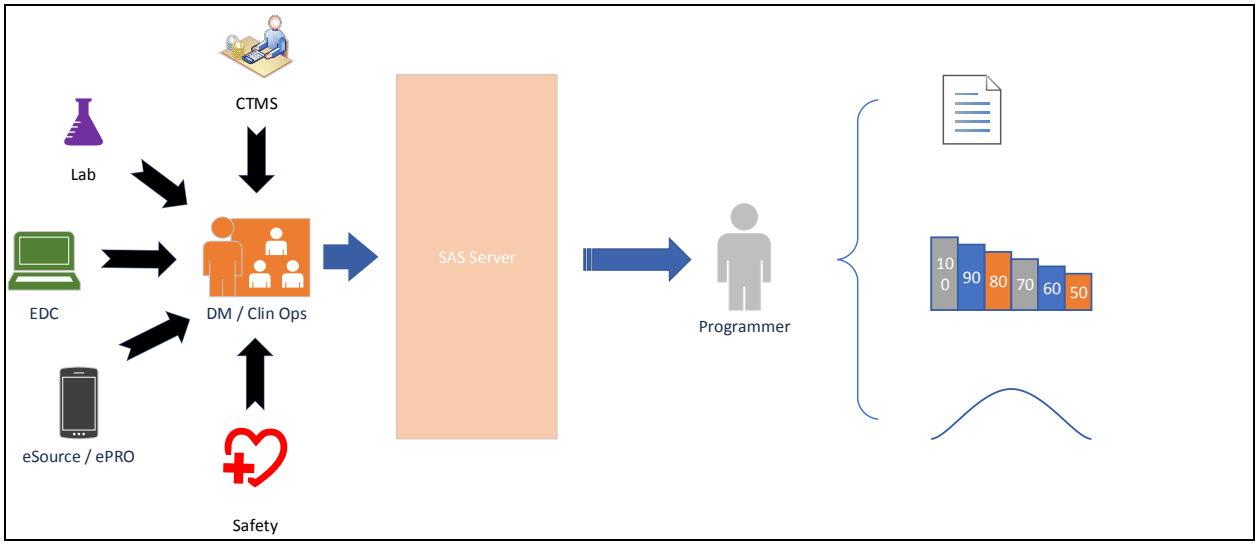

Figure 1: Historical Data Collection workflow

GETTING DATA OUT OF ESOURCE / EDC SYSTEM

Mostly all clinical trials use multiple systems and vendors to collect the clinical and operational data that a sponsor might use. Extracting the data automatically or on schedule from these systems depends on the maturity of the processes and technologies that the vendors use. It is not uncommon to be using more than one extraction medium. sFTP can be used where ease of setup is required but will demand a lot of attention. If the vendors and internal IT team have a well-established technology framework, then API and Webservices can be used to pull even the transactional data. If this is the preferred option,

then it is important to determine if the data is pushed by the vendor or pulled by you as a part of data ingestion requirements.

If the data being pulled is for data review only, the format of the files is not as important. But in instances where the data coming in can only be SAS files, the ETL process or tool being used must accommodate for conversion of SAS files.

The frequency in which the data should be pulled will always be dependent on the type of data and timing from database lock or other milestones in the study. For example, EDC and Lab data can be accepted weekly or monthly, while safety, AE data would be required daily or more frequently. Apart from scheduled data transfers the process for adhoc transfers must also be defined and tested, as these are typically done more frequently then planned.

Finally, what data should be pulled in? This again depends on the teams you are catering to, and the reports being generated. If the end users are Data Managers, Monitors and Biostatisticians, then having EDC, e-Source, Central lab and Safety data might suffice. For clinical trial managers and program managers it might be helpful to pull CTMS, enrollment and other vendor operational reports. If data science and machine learning is in scope, then bringing the transactional audit data across all systems should also be considered.

ETL OR ELT

With the advent of cloud technologies and lowering of data storage costs its frequently being asked whether the traditional Extract, Transform and then Load is still required? The advent of data sciences means that we need to have the data even in its raw format easily available and requiring the data to be loaded before any changes are made. But is this really a new line of thinking? Wasn't this being already done in a heavily regulated pharma industry?

Having all the raw data is good practice and helps fulfill regulatory requirements. But to make tools and processes reusable and efficient it's important to convert them into an intermediate standard. This will ensure that majority of the reports, dashboards and listings from an existing study will work across other studies as well. The intermediate standard can be defined based on every organization's needs. So, you decide if you want to term it as CDISC+, SDTM-, RAW+ or any other similar name.

During the data transfer and loading process we should verify the quality of the data as it flows from one state into another. The data should be tested at every

step for completeness, integrity and consistency using automated or manual validation steps.

WHY KNIME

KNIME, the Konstanz Information Miner, is a free and open-source data analytics, reporting and integration platform. It is widely used tool in the pharmaceutical industry albeit more on the drug discovery side. Knime caught our eye because it provides a visual platform for data transformation which makes it much easier for the Clinical data analytics team to document and maintain the entire data conversion process.

The workflows can be scheduled to run at any time of the day or can be event based. Knime can be used to filter and transform the data and make it ready for consumption by a dashboard. It can also be used to check for completeness of data and alert individuals or teams when an anomaly is detected. Finally, Knime can work with various visualization platforms – Tableau, Spotfire, Qlik and others to give the user a seamless guided analytics platform. Since it is open source, many useful tools like the SAS converter is easily available online.

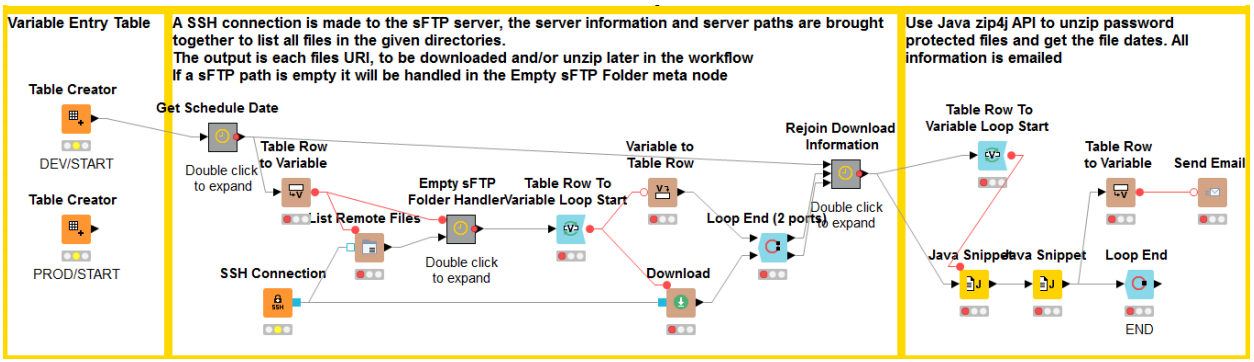

Figure 2: Data Pull from sFTP and Quality check workflow

VISUALIZATION

There are many visualization tools and technologies available and they should be able to accommodate your needs from a visualization standpoint. What is needed is not necessarily the best tool but the best-fit tool that fits your present and future needs.
If there is an ask within the organization for a large number of visualizations to be available in a short period of time, it might be best to look for an off-the-

shelf solution. These will come with some of the common visualization dashboards already built and ready to go. If you are a small to mid-size company where the team is still gauging through its data sciences and visualization requirements it might be best to build a dashboard internally using one of the visualization tools like Tableau, Spotfire, Qlik and the likes.

Tableau offers a contemporary visualization tool which provides the user ease to perform data analytics and exploration by themselves. If the reports are heavy on statistical analysis side it might be best to do the analysis outside and use Tableau for visualization only.

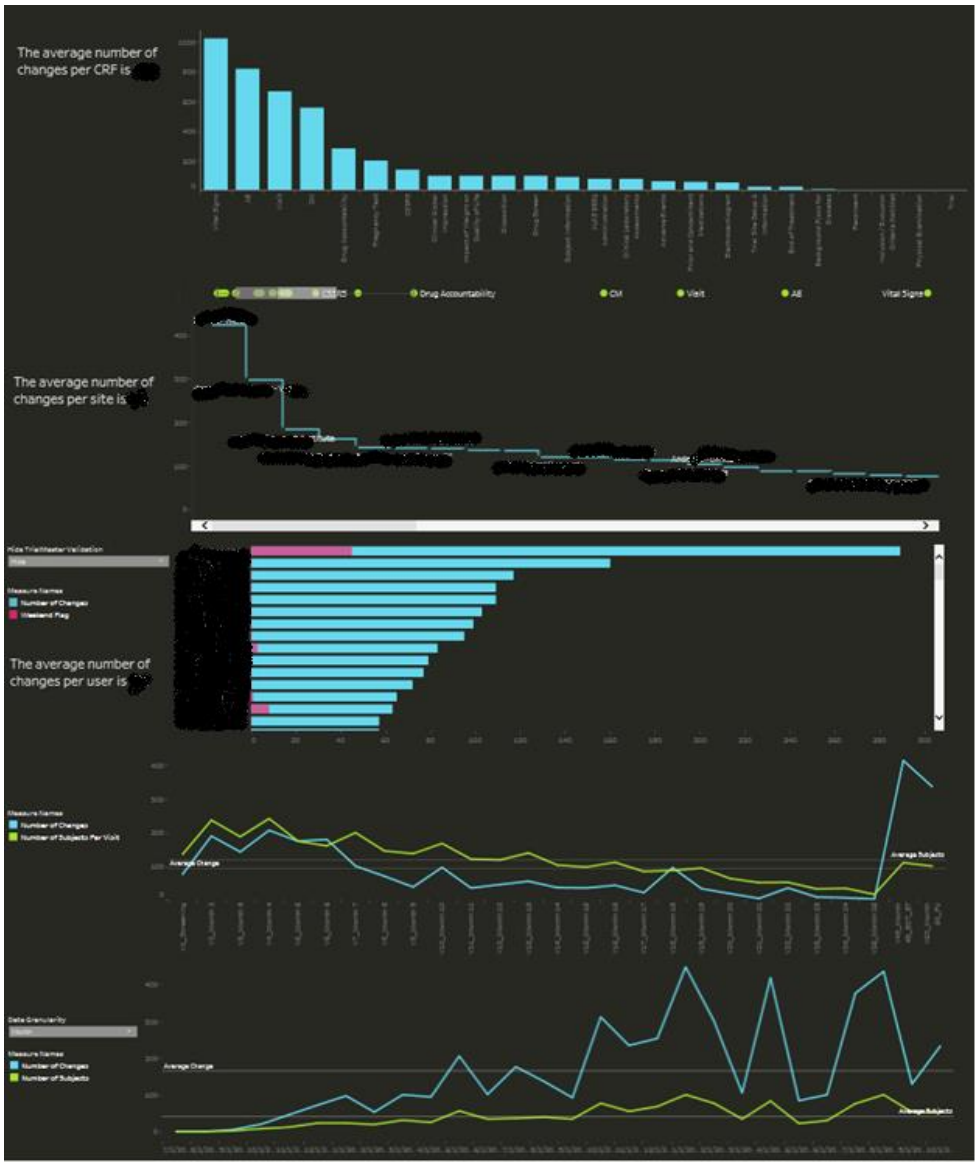

Figure 3: Audit Dashboard using Tableau visualization

Tableau's focus on being on the edge of technology is exemplary. They have a very wide base of user groups and strive to align themselves with the latest data science, analytics and visualization trends. It might be worthwhile to note that Tableau recently introduced their own ETL tool – Tableau Prep and chatbot - Ask Data.

WHERE WE ARE

Using tools like Knime and MoveIT we have been able to help Data Managers and Clinical Trial Managers to focus on execution of the trial instead of chasing the data. When there is a problem or missing data the team responsible including vendor, CRO, sponsor and other stakeholders are notified via emails generated directly from the system.

Having tools like Knime are extremely useful for transformation of data. But at the same time, we see a continued need for SAS or/and Python for some of the more complex algorithms. These can easily coexist, and you can even create Knime shells with Python code as a part of the entire workflow.

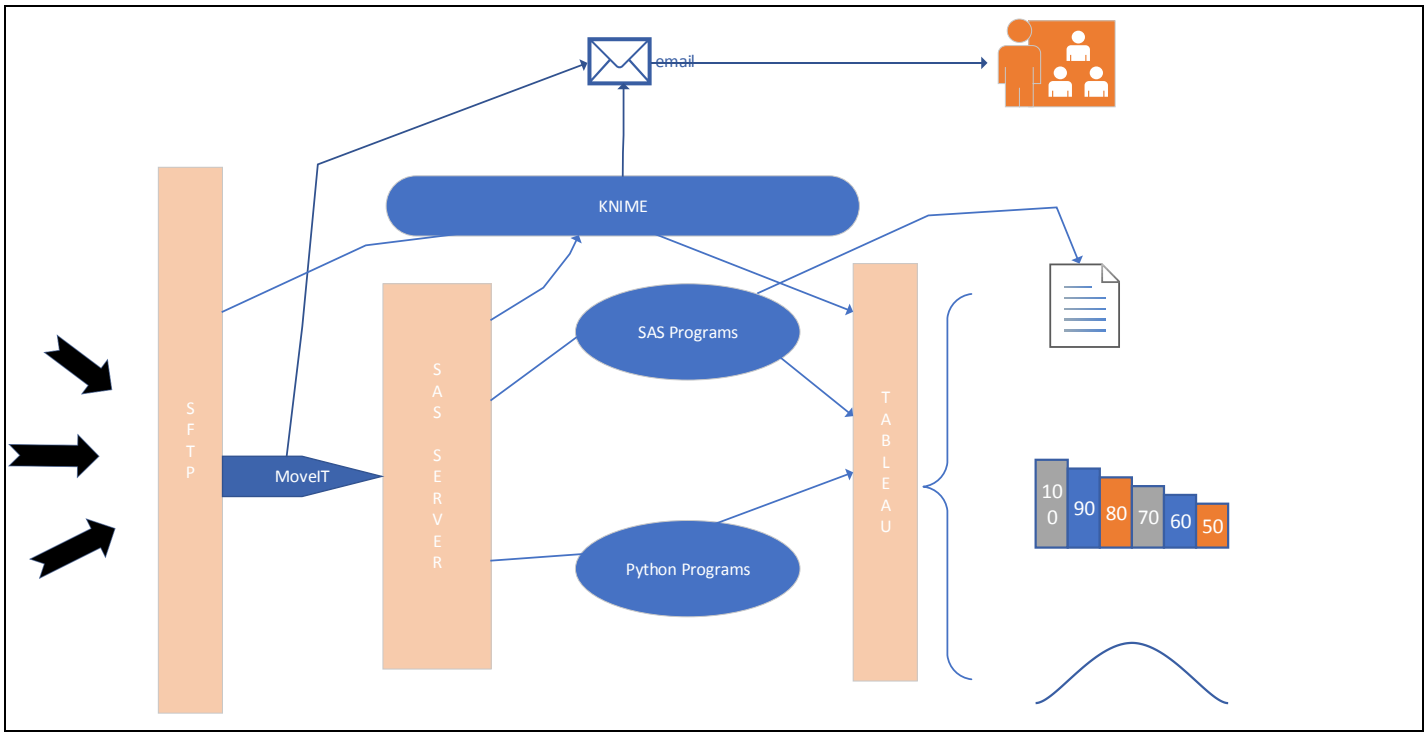

Figure 4: Alkermes data operations as of now

VALIDATION

This brings us to the next logical question – What is the validation strategy that we should adopt? We answered that question by defining our users and how the data is being used. If the dashboard, listing, or visual is for ongoing data review, the validation efforts can be kept lighter. But in case your deliverable is being looked at by the Biostatistician and might be used in any kind of submission to the FDA or any other external review a full cycle validation should be done.

Currently at Alkermes, all our deliverables are for data review only. So, we are in the process of validating software platforms like Knime, Tableau and even MoveIT. But apart from validating the vanilla software packages we have also documented our development process, defined workflows to be followed when transforming data and/or building dashboards and provided standard templates / forms where required. The workflow dictates requirements documentation, peer review of code and testing against a test script. This ensures that the process is nimble and doesn't have a huge impact on our development timelines.

DATA STRATEGY GOING FORWARD

As we evolved through this cycle, few of our deliverables and decisions were reactive. But now that our team and processes have matured its imperative for us to extend our vision and look beyond. Currently we go digging for our data for every requirement or dashboard built. Therefore, one of the first problems that we are trying to solve is to have a central repository of data.

We have started discussions around the needs to have a data lake which can be used for all our data analytics and data science work. In a recent brainstorming session, we defined the key stakeholders and framework which should be used to build this data lake. This has given us a good blueprint of what needs to be implemented which will be worked on in 2020.

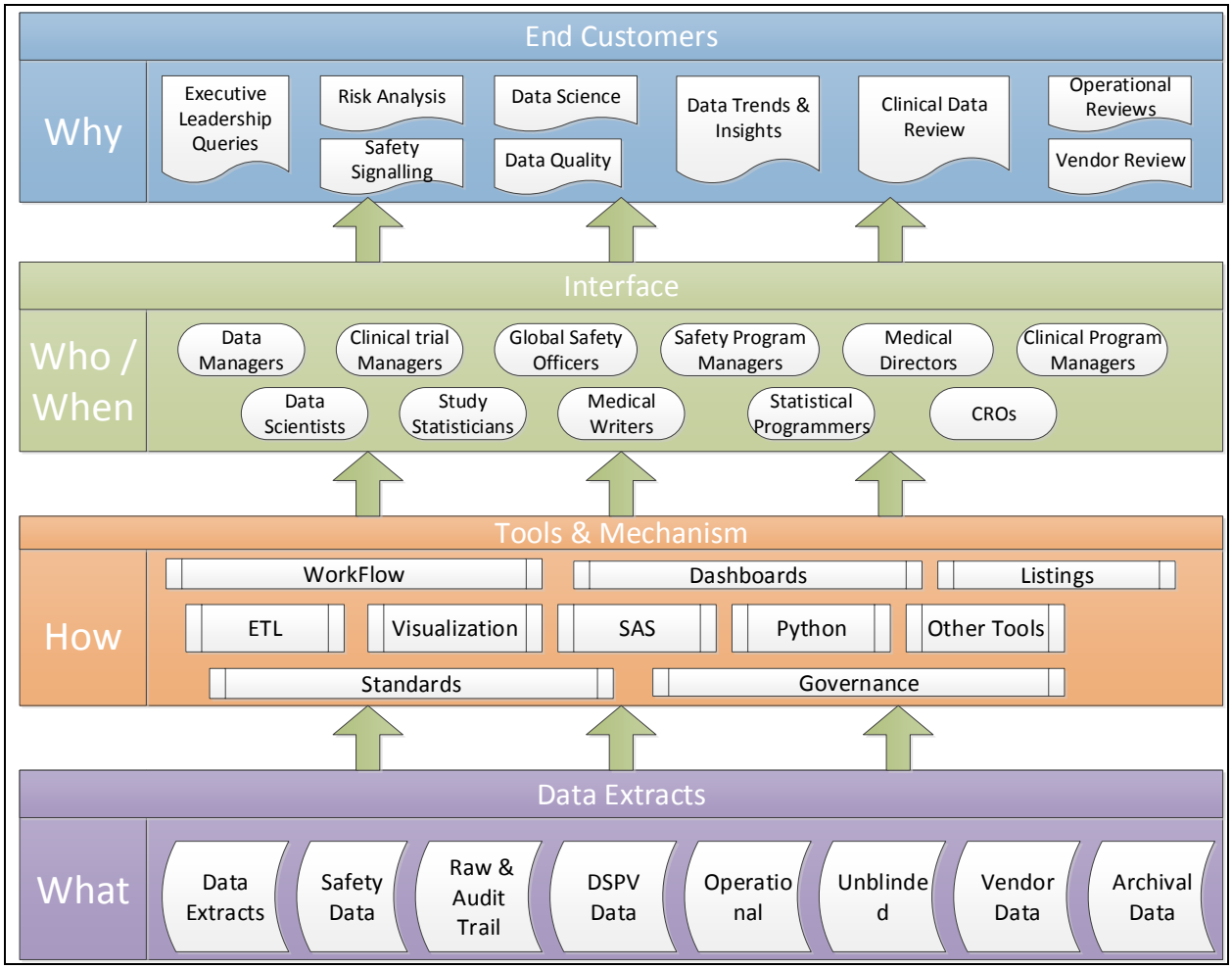

Figure 5 : Data Strategy Framework

CASE STUDY 1

Alkermes uses patient profiles to review completed or early terminated patients. Historically this review was done using SAS programs or JMP Clinical producing listing reports. The domains and fields included in the profiles vary from study to study or may be different even for a single study based on the intent of the review.

We were able to produce Patient Profiles in Tableau with dynamic filtering and data analysis capabilities. This will reduce the usage of data listing reports, making the data review faster and more efficient.

Knime was used to convert raw data into the format as per requirements for the data review being done. Additional calculations like the PCS criteria was incorporated within the Knime workflow.

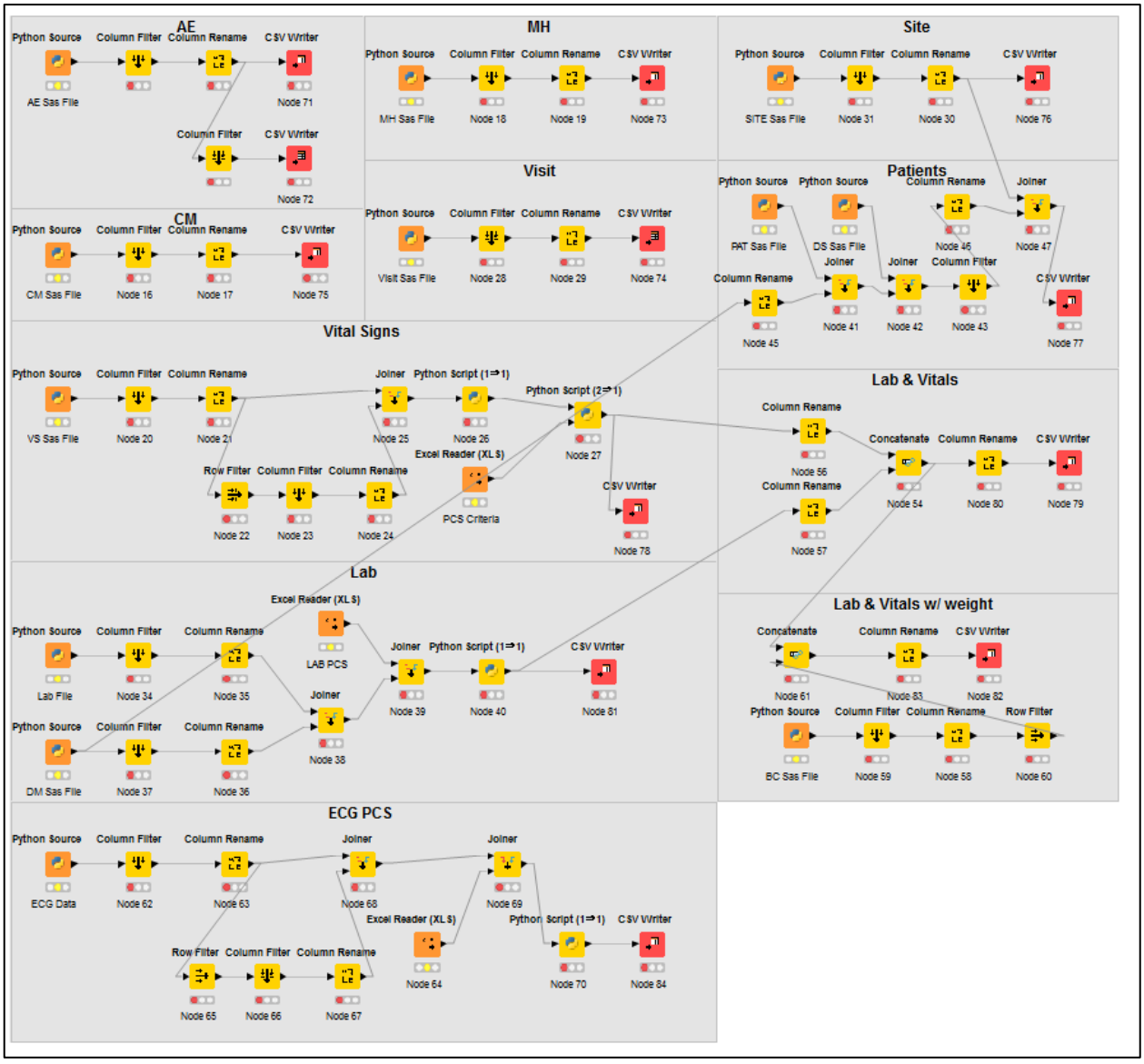

Figure 6 : Data Transformation using Knime

Tableau was used for visualization of the subject data by creating a dashboard for each domain being reviewed. Hovering over a data point provides additional details about that field as a tooltip. Instead of providing a huge listing with all labs we provided a lab dropdown which the monitor can use to review desired lab.

Figure 7 : Visualization of Patient Profile data

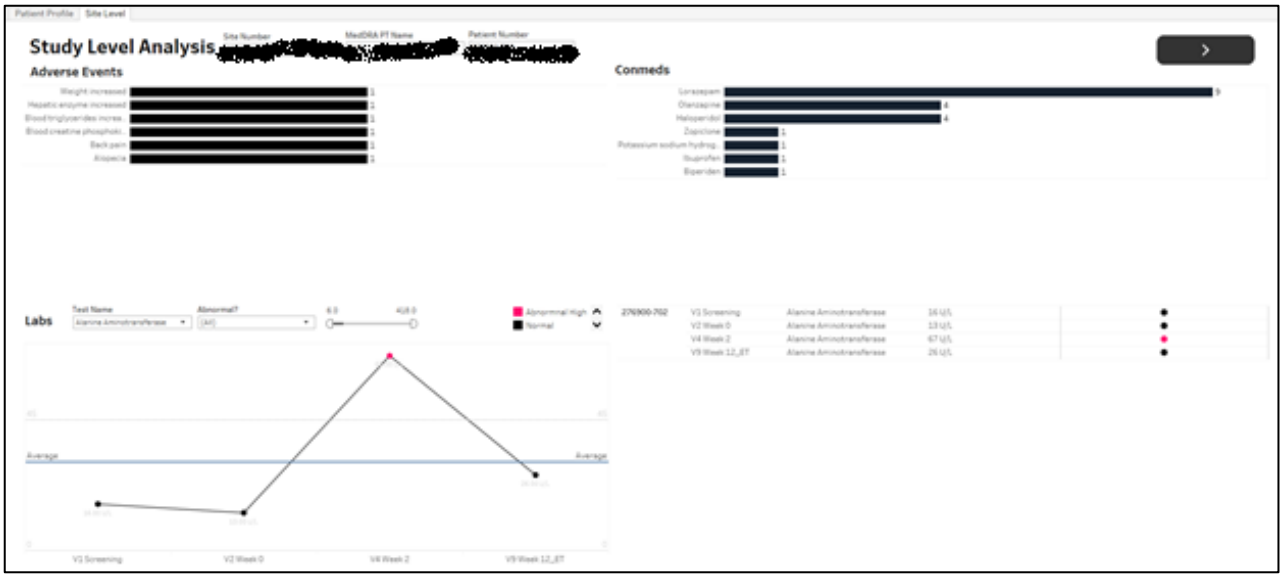

Figure 8 : Patient Lab data

CONCLUSION

The data journey has started for us and we have made good decisions regarding tools and have partnered with some great teams comprising of both internal departments and external vendors. We are actively working on our

long-term data strategy and are optimistic that it will help us to make our processes and deliverables better.

Irrespective of where the apple pie came from we want to make sure that it is made to order for our teams who will be consuming it.

CONTACT INFORMATION

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

Ashish Kumar
Alkermes Inc.
852 Winter St,
Waltham. MA 02451-1420
Work Phone: +1 781 609 6620
Email: ashish.kumar@alkermes.com
Web: www.alkermes.com

James Dale
Alkermes Inc.
852 Winter St,
Waltham. MA 02451-1420
Work Phone: +1 781 609 6792
Email: james.dale@alkermes.com
Web: www.alkermes.com