PharmaSUG 2022 - AI-151

Leveraging AI to Process Unstructured Clinical Data in Real-time

Madhusudhan Nagaram, Allogene Therapeutics
Syam Chandrala, Allogene Therapeutics

ABSTRACT

Clinical trial data is typically posted as structured data in the form of SAS datasets or excel or csv formats. This structured data can be easily read using various analytics tools such as SAS, R, Tibco, Spotfire, etc., and can be presented as listings, summarized reports or visualizations or dashboards. Sometimes Translational data, including labs, Pharmacokinetics, Pharmacodynamics, and biomarker data, can be obtained from external vendors in various file formats. Often these external vendors lack the capability to send real-time data in structured file format. Instead, they send this data as PDF documents, scanned images of printer or handwritten documents, which are challenging to read and ingest into the tools for data listings and visualizations without human intervention.

Our goal is to minimize or eliminate human intervention, automate the dataset acquisition, pre-processing, validating and ingesting data into AWS data lake in real-time for functional users as an analytics platform. We have addressed this challenge by creating an automation process using various tools like Microsoft Power Automate, AWS S3, AWS Lambda functions, AWS Athena and MS AI models. So, the documents were processed in real-time and are available on data lake for analytics tools and dashboards for statistical analysis by utilizing serverless applications.

In this paper, we will demonstrate the step-by-step process of mapping the data from PDF files using Microsoft Power automate, storing it in AWS S3 bucket, querying the data using AWS Athena and reading it into Tibco Spotfire using the information link.

INTRODUCTION

Clinical trials are intended to investigate and find answers to a research question by generating data to prove or disprove the hypothesis. The quality of data generated plays an important role in the outcome of the study. For any clinical trial, sponsors must collect all protocol specified information as data on Case report form (CRF), which are then used to analyze and report clinical trial results. CRF data collected from patients are transferred to the sponsor in a structured format in the form of SAS datasets or csv or excel files.

With expanding knowledge in tumor biology and biomarkers, oncology therapies are increasingly moving towards biomarker-driven therapies tailored according to patient-specific characteristics, most commonly the tumor's molecular profile. Biomarker data plays a critical role in clinical research. It is evaluated as an indicator of biological, pathogenic, or pharmacologic responses to a therapeutic intervention.

Pharmaceutical companies collaborate with external vendors in obtaining different types of biomarker data generated from various assays. Below are a few examples of diagnostic assays using innovative technologies to identify biomarkers for microbial contamination, cell line identity, cytogenetics, and genetic stability.

a. Flow Cytometry
b. Fluorescence Immunohistochemistry
c. PCR assays
d. Gene expression assays
e. Next generation sequencing
f. Fluorescence in situ hybridization (FISH)

g. Chromosome microarray analysis
h. Mycoplasma detection by PCR
i. Short Tandem repeat analysis (STR)
j. Single nucleotide Polymorphism (SNP)
k. Spectral Karyotyping
l. G-banded Karyotype
m. Immunosequencing
n. Assays to track MRD data

Vendors usually post data in structured format once a month or quarter. They don't have the capability to transfer data very frequently in a structured format. Translational scientists like to see the biomarker data more often to identify the impact of biomarkers before any clinical effect is evident. As per the request from sponsor, vendors transfer the data in the form of scanned pdf documents, which can be either digital or handwritten.

Statistical programmers get requests from Translational and Clinical science departments to analyze this data and create reports from these documents. It is challenging for the programmers to read and generate analysis from these unstructured files. We have collaborated with IT and Analytics colleagues to process these files and generate visualizations using analytical tools like Spotfire and R.

This paper shows examples of unstructured clinical trial data from these above example assays and describes how this unstructured data is converted to a structured format, using a combination of various tools such as Microsoft Power Automate AI builder, AWS tools including AWS S3, Amazon Athena, Amazon SES and create visualizations using analytical tools like TIBCO Spotfire, Tableau and R.

TOOLS USED

POWER AUTOMATE

Power Automate is a service that helps to create automated workflows between apps and services to synchronize files, get notifications, collect data and more. Users can explore diverse set of templates and automate tasks. We have used AI builder in this Power Automate platform to read the pdf documents and handwritten images without coding or data science skills.

AWS S3

Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance.

AWS ATHENA

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon Simple Storage Service (Amazon S3) using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries. Athena scales automatically, running queries in parallel. So, the results are fast even with large datasets and complex queries.

AWS SES

Amazon Simple Email Service (SES) is a cost-effective, flexible, and scalable email service that enables developers to send mail from within any application securely, globally, and at scale. We can configure Amazon SES quickly to support several email use cases, including notifications, transactional, marketing, or mass email communications.

TIBCO SPOTFIRE

Tibco Spotfire is a data visualization and analytics software help to explore data and create interactive visual analytic dashboards.

We have chosen Microsoft Power Automate AI builder to create an AI model, trained the model with few documents and used the results from the model to automate the document processing and generate structured files in the form of csv format. CSV files generated from Power automate are stored in AWS S3 bucket. Using Amazon Athena, we read the data from AWS S3 and created tables and views which can be consumed by analytical tools such as Spotfire and R.

STEPS TO PROCESS UNSTRUCTURED DATA

STEP 1: TYPES OF UNSTRUCTURED CLINICAL TRIAL DATA

a. Scanned handwritten pdf files

b. Digital pdf files

c. Printed and scanned pdf file

Examples of Unstructured Data:

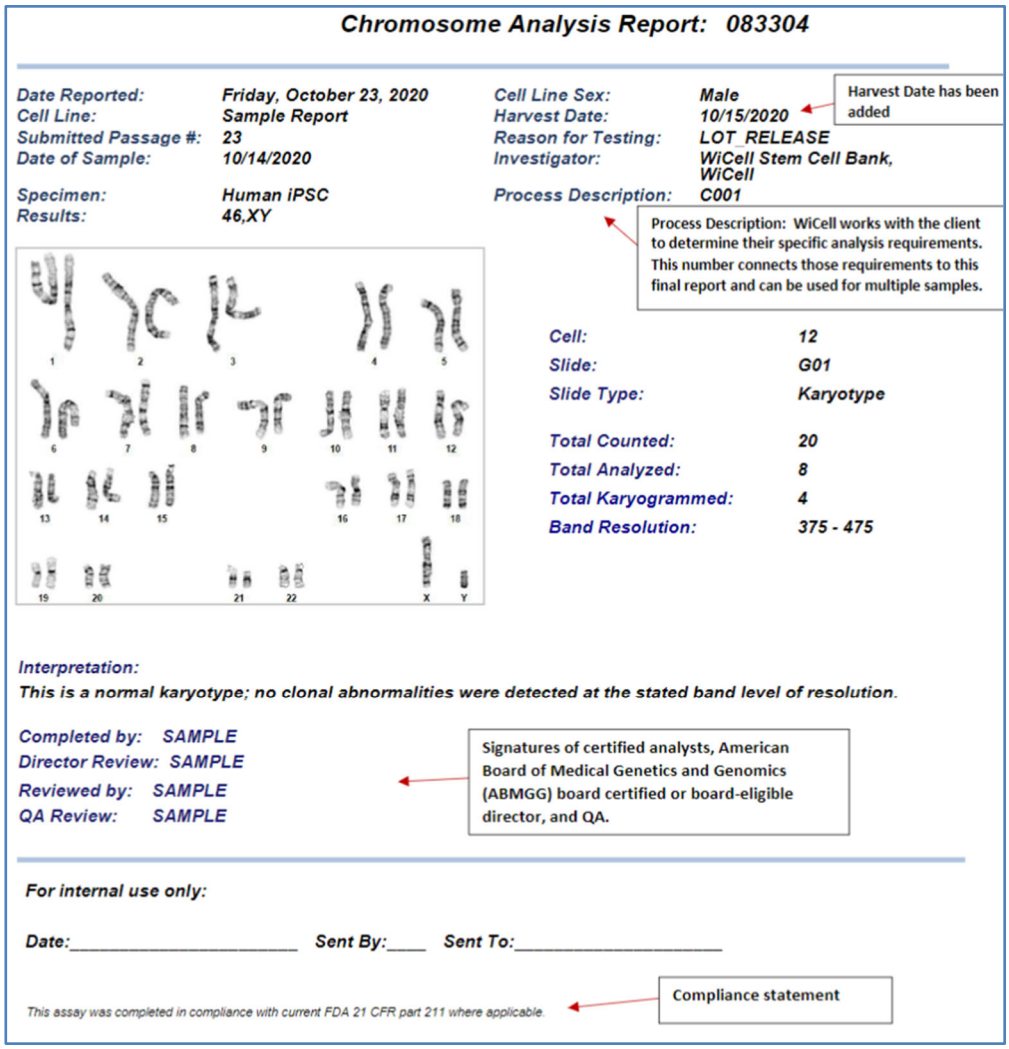
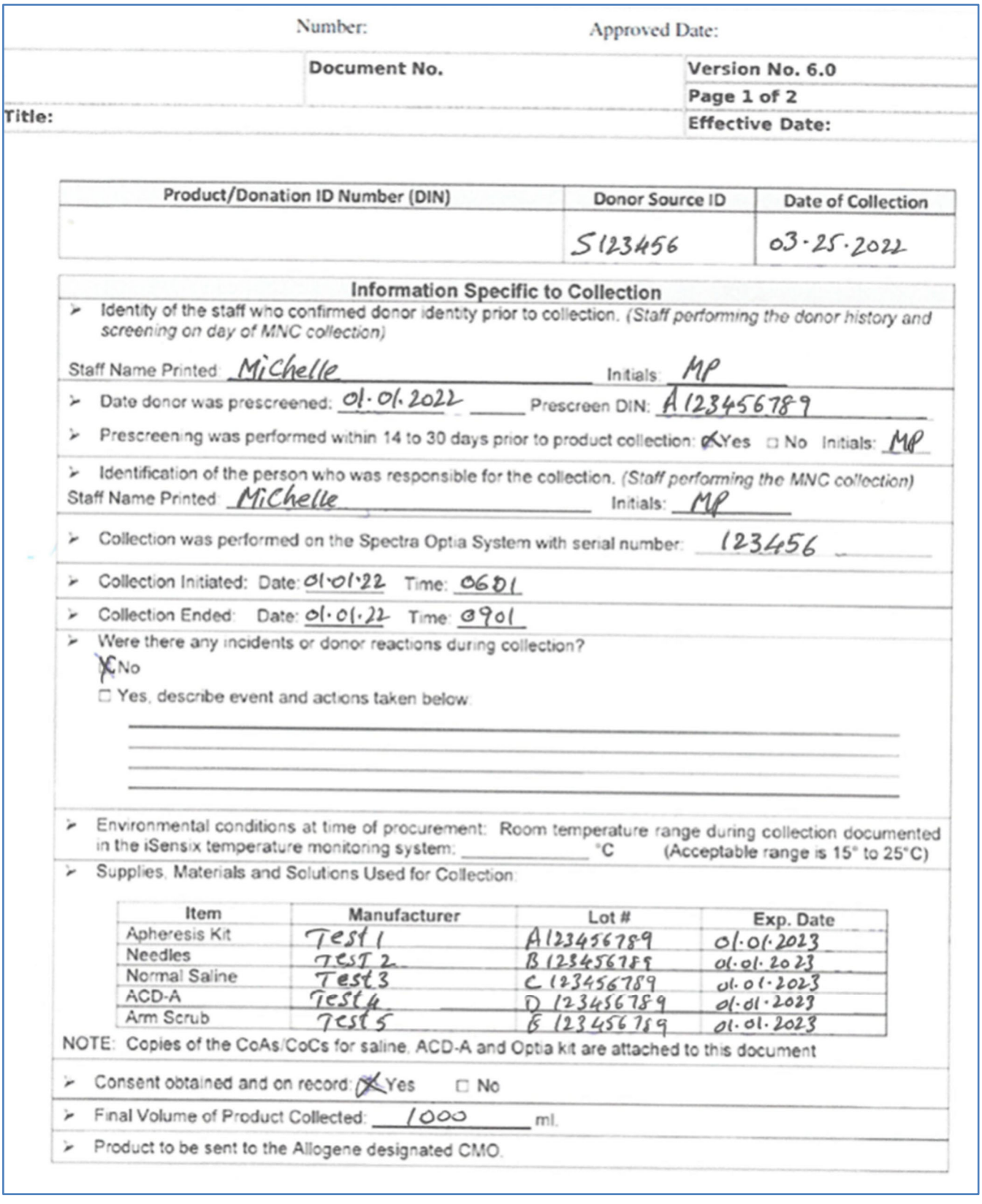


Figure 1. Scanned Digital pdf files



Number: Approved Date:

	Document No.	Version No. 6.0
		Page 1 of 2
Title:		Effective Date:

Product/Donation ID Number (DIN)	Donor Source ID	Date of Collection
	S123456	03·25·2022

Information Specific to Collection

- Identity of the staff who confirmed donor identity prior to collection. *(Staff performing the donor history and screening on day of MNC collection)*

Staff Name Printed: Michelle Initials: MP

- Date donor was prescreened: 01.01.2022 Prescreen DIN: A123456789
- Prescreening was performed within 14 to 30 days prior to product collection: ☒ Yes ☐ No Initials: MP
- Identification of the person who was responsible for the collection. *(Staff performing the MNC collection)*

Staff Name Printed: Michelle Initials: MP

- Collection was performed on the Spectra Optia System with serial number: 123456
- Collection Initiated: Date: 01.01.22 Time: 0601
- Collection Ended: Date: 01.01.22 Time: 0901
- Were there any incidents or donor reactions during collection?

☒ No

☐ Yes, describe event and actions taken below:

- Environmental conditions at time of procurement: Room temperature range during collection documented in the iSensix temperature monitoring system: _____ °C (Acceptable range is 15° to 25°C)
- Supplies, Materials and Solutions Used for Collection:

Item	Manufacturer	Lot #	Exp. Date
Apheresis Kit	Test 1	A123456789	01.01.2023
Needles	Test 2	B123456789	01.01.2023
Normal Saline	Test 3	C123456789	01.01.2023
ACD-A	Test 4	D123456789	01.01.2023
Arm Scrub	Test 5	E123456789	01.01.2023

NOTE: Copies of the CoAs/CoCs for saline, ACD-A and Optia kit are attached to this document

- Consent obtained and on record: ☒ Yes ☐ No
- Final Volume of Product Collected: 1000 mL
- Product to be sent to the Allogene designated CMO.

Figure 2. Scanned Handwritten document

2000 Lab Rd.
Fremont, CA 92008
Phone: +001-XXX-XXX-XXXX

First last, M.D.
Laboratory Director
CLIA ID # 123456789
CAP # 123456

Subject ID:	100101001-001	**Accession #:**	AB111111-0001
Visit:	Month 1	**Protocol:**	XXX-111-001
Requisition ID:	6516594692	**Site #:**	100101001
Specimen:	PB EDTA	**Site Name:**	001
Collected:	02 Jan 2022 09:00	**Principal Investigator:**	
Received:	15 Jan 2022 11:16	**Reported:**	30 Jan 2022 12:57

Test qPCR assay

PB EDTA submitted for evaluation.

Results:

Reported Result

copies/ug DNA = 0.11 Average copies/ug

Average copies/cell = 0.12

Cor. Copies/uL= 0.13

Method:

The CAR qPCR assay is a test-based qPCR assay designed to detect copies of transgene in peripheral blood, bone marrow.

signature

Senior Hematopathologist — 25 Jan 2022

Electronic Signature — **Date**

Meaning: I have reviewed and approved these results.

This test was developed and its analytical performance characteristics were determined by XXXXXXXXXX Services, Inc. and validated for use per CLIA requirements.This test has not been cleared or approved. All tests are performed at Services, Inc., unless otherwise specified in the report.

Page 1 of 1

Figure 3. Printed and scanned digital document

STEP 2: TRAIN AI MODEL USING AI BUILDER IN POWER AUTOMATE

External vendors post the data to SFTP or Box or notify sponsors through email when data is posted in their web portal. The model must be trained using a few reports, as shown in the figure 4. User identifies the required data points from the document and value for each field and maps the value to the specific field. To train the model, this step should be repeated at least for five documents and publish the model in power automate for further use.

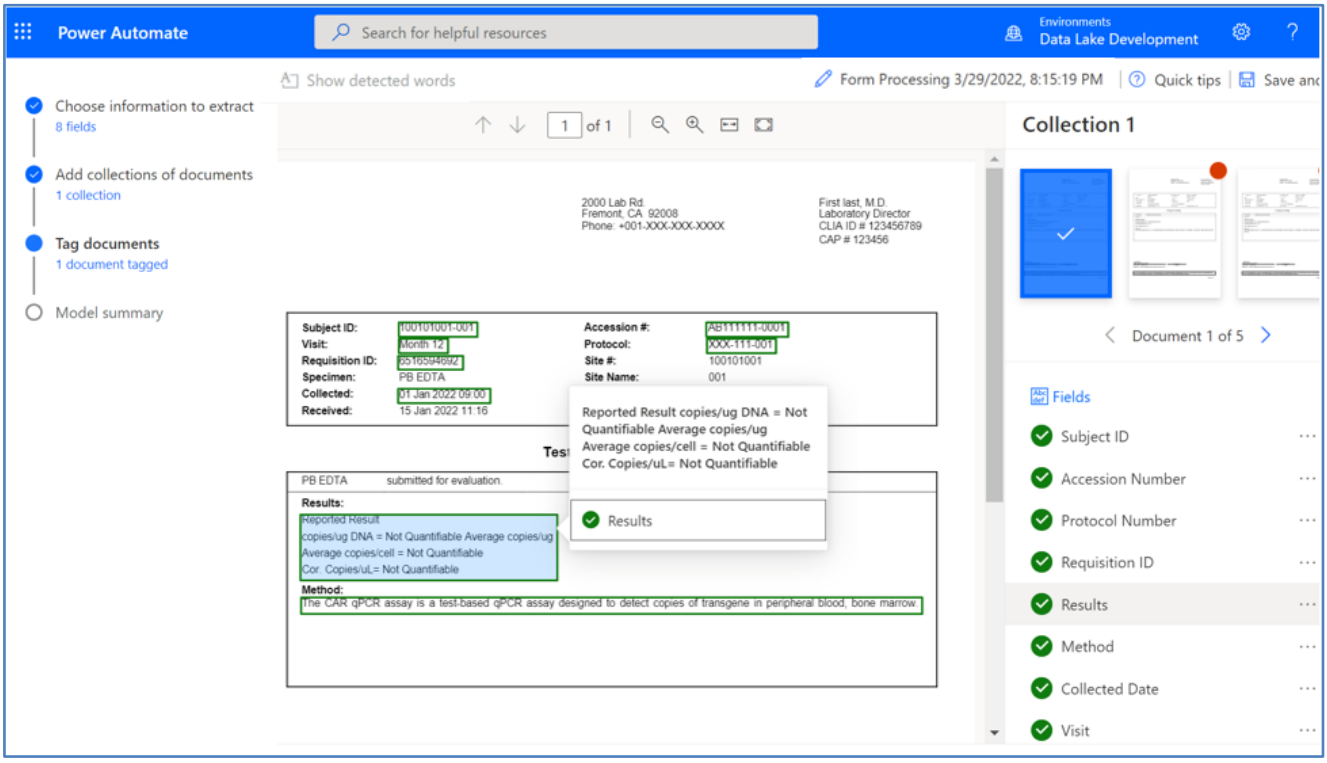


Figure 4. Training the AI model and mapping the digital pdf files

STEP 3: SETUP DATA PIPELINE

Data pipeline needs to be created by the user to automate the collection of the data from vendors.

This process includes following steps.

a. Trigger the flow, either scheduled or event driven (for example: when new email arrives from vendor)

b. Define required variables in the flow

c. Send the contents of pdf document to AI model

d. Identify the data using AI model and extract required fields.

e. Send the data record to the user for review and approval.

f. Create files in csv or json format once it is approved and copy it to AWS S3 bucket

The same process is repeated when similar files arrive from the vendor subsequently.

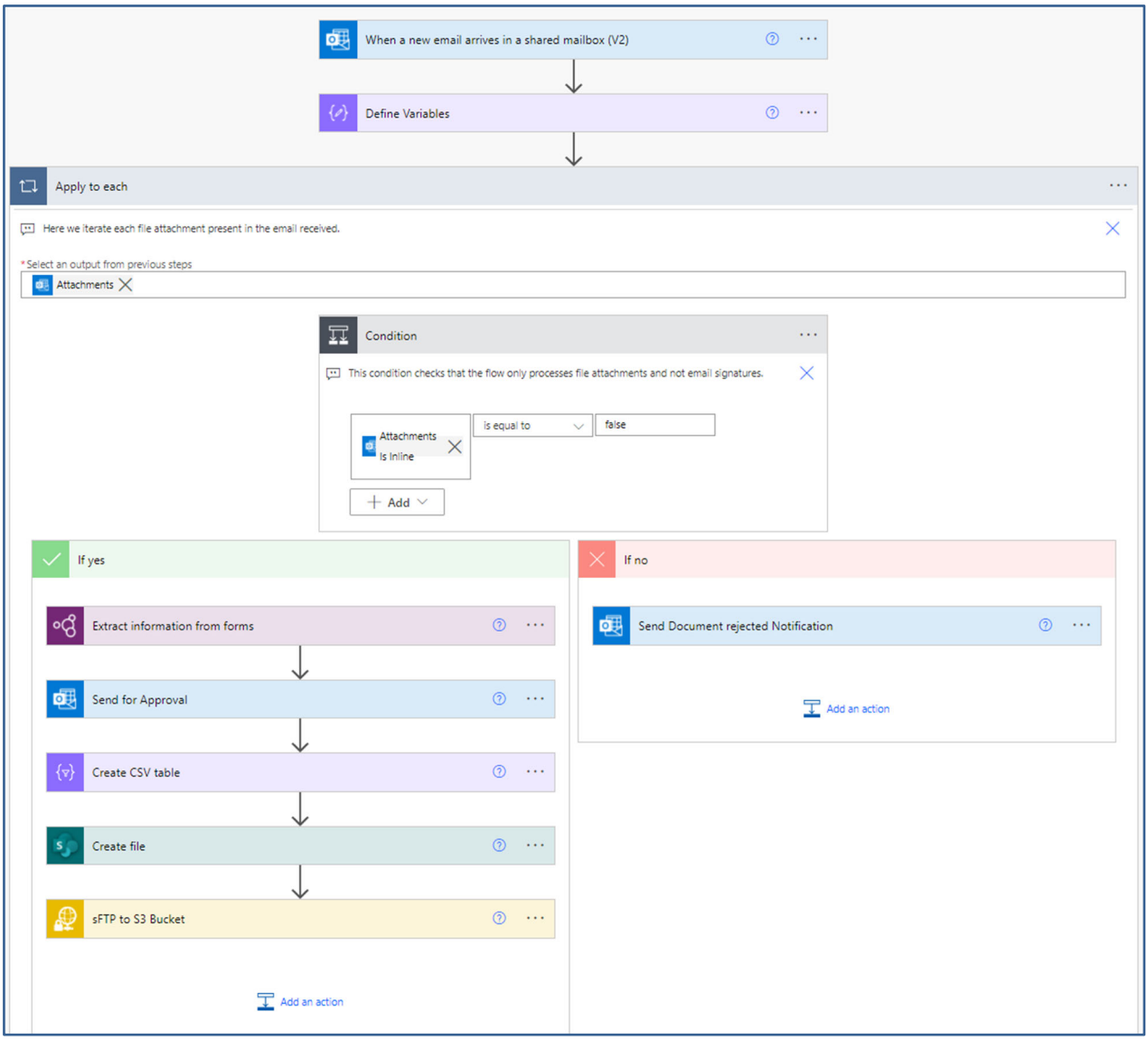


Figure 5. Example of data pipeline using Power Automate

STEP 4: CREATE ATHENA TABLES/ VIEWS

Once the data gets copied to AWS S3 bucket, Athena tables/ views are created using this structured data as shown in figure 6. The data can be queried using these tables.

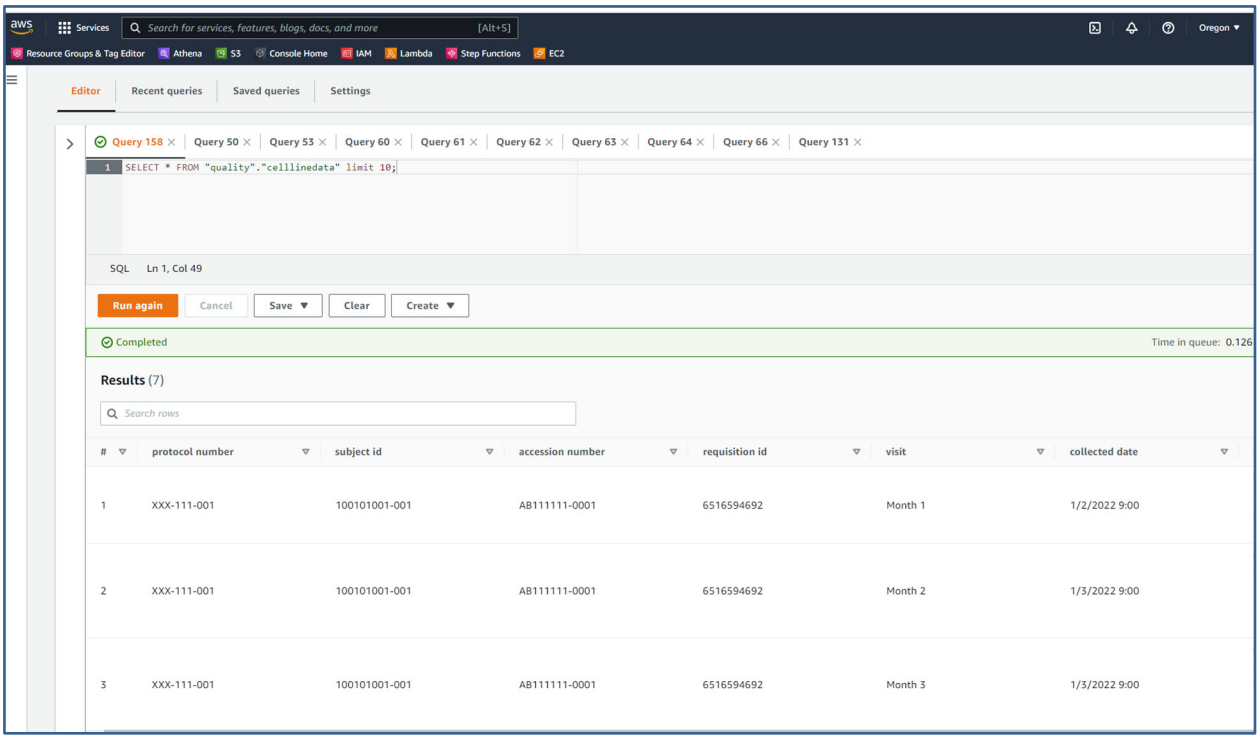


Figure 6. AWS Athena table from csv file in AWS S3 bucket

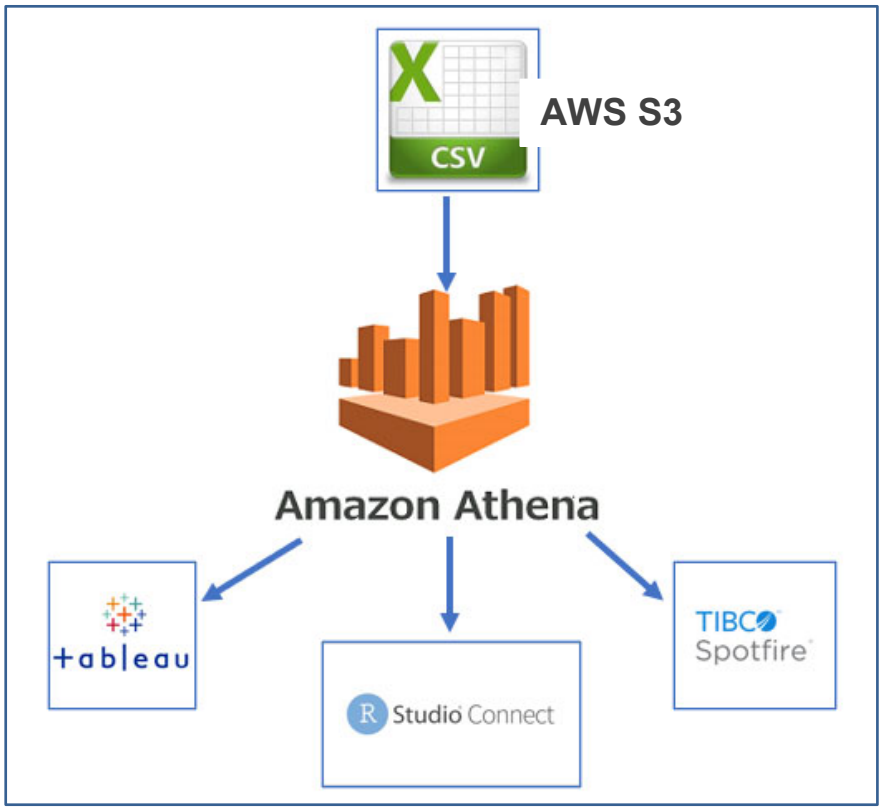


Figure 7. Data flow from AWS S3 to analytical tools via Amazon Athena

STEP 5: CREATE DASHBOARDS/ REPORTS USING ANALYTICAL TOOLS

AWS Athena is linked to TIBCO Spotfire using the Information link, and the data is available for creating the dashboards and reports, as shown in figure 8. This data from AWS Athena can be pulled into Tableau or R for performing data analysis and creating dashboards or reports. This data from AWS S3 bucket can be compared with monthly SAS data extracts that are posted to SAS Server by vendors.

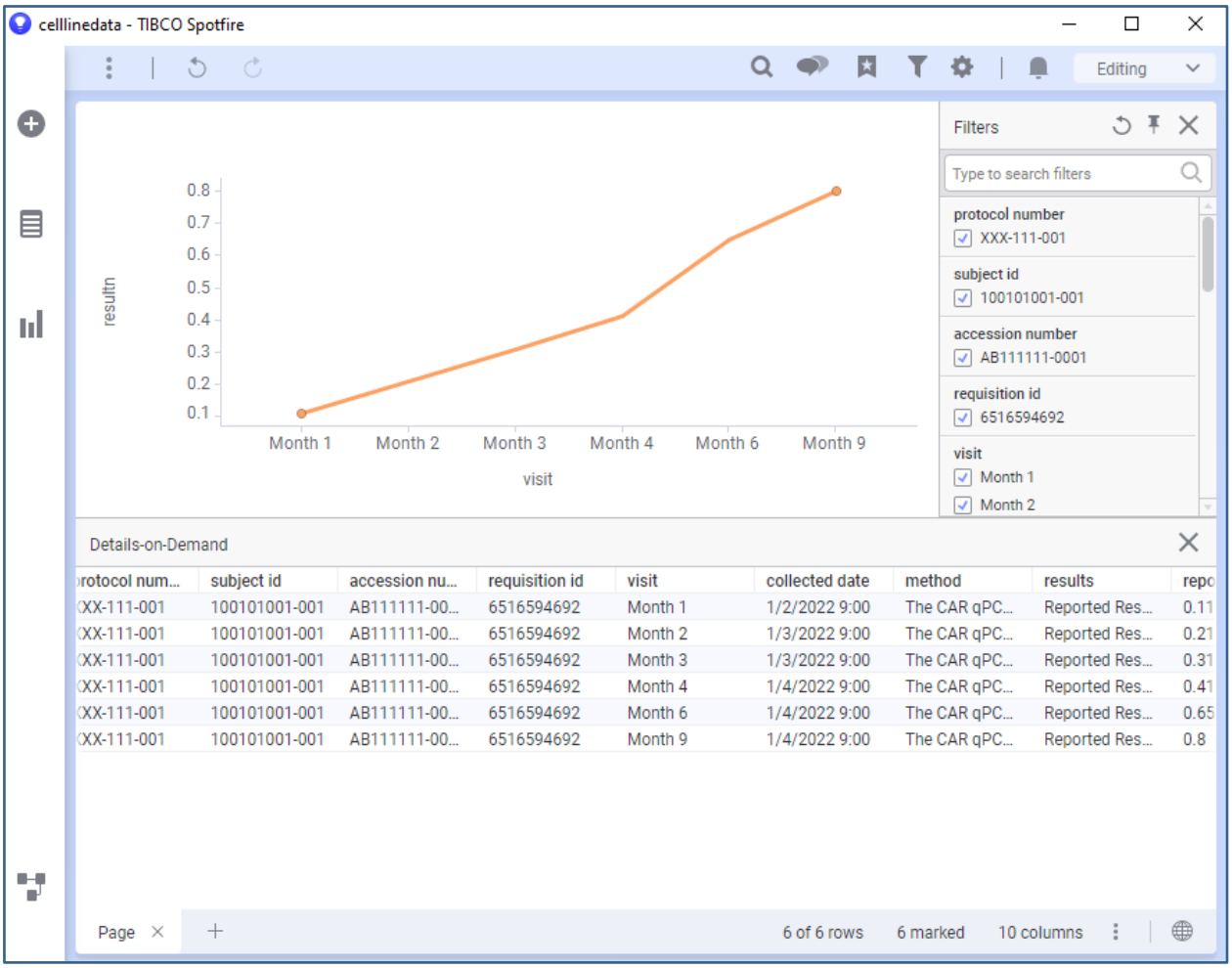


Figure 8. Tibco Spotfire

CONCLUSION

In this paper, we have demonstrated the automation of converting unstructured data to structured format. This process was designed by combining cloud-based AWS tools, AI builder from power automate, and analytics tools for making the data available to the users without any human intervention. The files are ingested into Power Automate and the process was repeated a few times for each vendor data, which helps in training the AI model to identify the data accurately. This process helps Clinical and Translational scientists in interpreting the data immediately and efficiently without any errors. This process helps not only statistical programmers to obtain translational data but also helpful for other cross functional departments with similar challenge.

REFERENCES

1. Data management in clinical research: An overview
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3326906/2

2. Biomarker-Driven Oncology Clinical Trials: Key Design Elements, Types, Features, and Practical Considerations
   https://ascopubs.org/doi/10.1200/PO.19.00086

ACKNOWLEDGMENTS

We would like to thank Saurav Mahanti, Senior Director, Data Management Analytics and Integration and Debi Prasad Roy, Executive Director, Head of R&D Tech, Data Science, Chaitanya Chowdagam, Principal Programmer for their support and suggestion/comments.

RECOMMENDED READING

- *Microsoft Power Automate Documentation*
- *TIBCO Spotfire Documentation*
- *Amazon Athena user guide*

CONTACT INFORMATION

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

Name: Madhusudhan Nagaram
Enterprise: Allogene Therapeutics
Address: 210 E Grand Ave., South San Francisco, CA 94080
E-mail: mnagaram@gmail.com

Name: Syam Chandrala
Enterprise: Allogene Therapeutics
Address: 210 E Grand Ave., South San Francisco, CA 94080
E-mail: Syam.chandrala@gmail.com