PharmaSUG 2022 – Paper EP-156

Automation of Clinical Data Extracts Using Cloud Applications

Syam Chandrala, Allogene Therapeutics
Madhusudhan Nagaram, Allogene Therapeutics
Chaitanya Chowdagam, Ephicacy Consulting
Jegan Pillaiyar, Ephicacy Consulting
Kunal Chattopadhyay, ZS Associates

ABSTRACT

Data managers and external vendors post the data, in the form of zip files, into various storage locations such as BOX, SharePoint, sFTP, etc. Some of these data extracts are scheduled periodically, and some are scheduled bimonthly or on ad hoc basis. In addition to this, these data extracts could be in various formats and most of the time the files are password protected. It is the programmer's responsibility to make sure that current data and historic data are stored in the appropriate folders and in a consistent format such as SAS7BDAT as a source for creating SDTM, ADaM datasets. Programmers spend a lot of time and effort managing this process manually.

By leveraging the Cloud-based applications such as AWS secrets manager, AWS lambda functions in combination with python scripting, SAS program, and scheduling shell scripts, we have made this process automated and serverless. The latest data from sFTP will be identified, stored into AWS S3 location, unzipped using the correct password, the files then saved into appropriate folders, and an email will be sent to the team for success or failure of the data extracts along with the logs. This data will be posted on data lake to make it available for data analytics tools such as SAS, Spotfire, tableau, R etc., with minimal or no intervention of programmers.

This paper will demonstrate flow of data though multi-level cloud-based serverless architecture conjugated with python, SAS programs, and shell scripts.

INTRODUCTION

Statistical programmers map data from various sources such as EDC, PK, Biomarkers into SDTM and create ADaM datasets. This data is usually posted by external vendors as per the DTS (Data Transfer Specifications) generated by data managers. EDC data is obtained 2-3 times a week while other forms of data are transferred monthly, quarterly, or on ad hoc basis. It is crucial to organize this data in appropriate directories in order to generate analysis, datasets, and reports. It is always recommended to save or archive historic data in study folders in case any analysis is requested by clinical scientists based on this data. It takes a lot of effort, time and cost to maintain these study folders.

For non-EDC data, Data managers create DTS (Data Transfer Specifications) document, which includes all the information such as:

- Type of data
- Data transfer type
- File naming convention for zip file and the contents of zip file
- File names of all the files present in the zip file
- File format
- Transfer method and frequency
- Transfer structure consisting of data attributes, tests included, visit codes etc.

When external vendors transfer this data, statistical programmers compare the data against the data transfer specifications. It is an extra burden on programmers to make sure the data is consistent with the DTS document.

Data from all these sources are posted into various locations, including BOX, sharepoint, SFTP, etc. There is a need to standardize the process to minimize the effort, time and cost of managing this data. We have collaborated with our IT and Analytics colleagues in order to receive this data into Data Lake, which is present in AWS (Amazon workspace). Different types of AWS tools such as S3, Lambda functions, AWS Secrets manager, Amazon SES (Simple Email Service), etc., are used to store the EDC and non-EDC data and notify the users when the data is posted to Data Lake and SAS servers.

PROCESS

Below flow chart shows the transfer of data from various data sources, transformation of data, publishing the data using analytical tools to make the analysis results available for end users. Refer Figure 1.

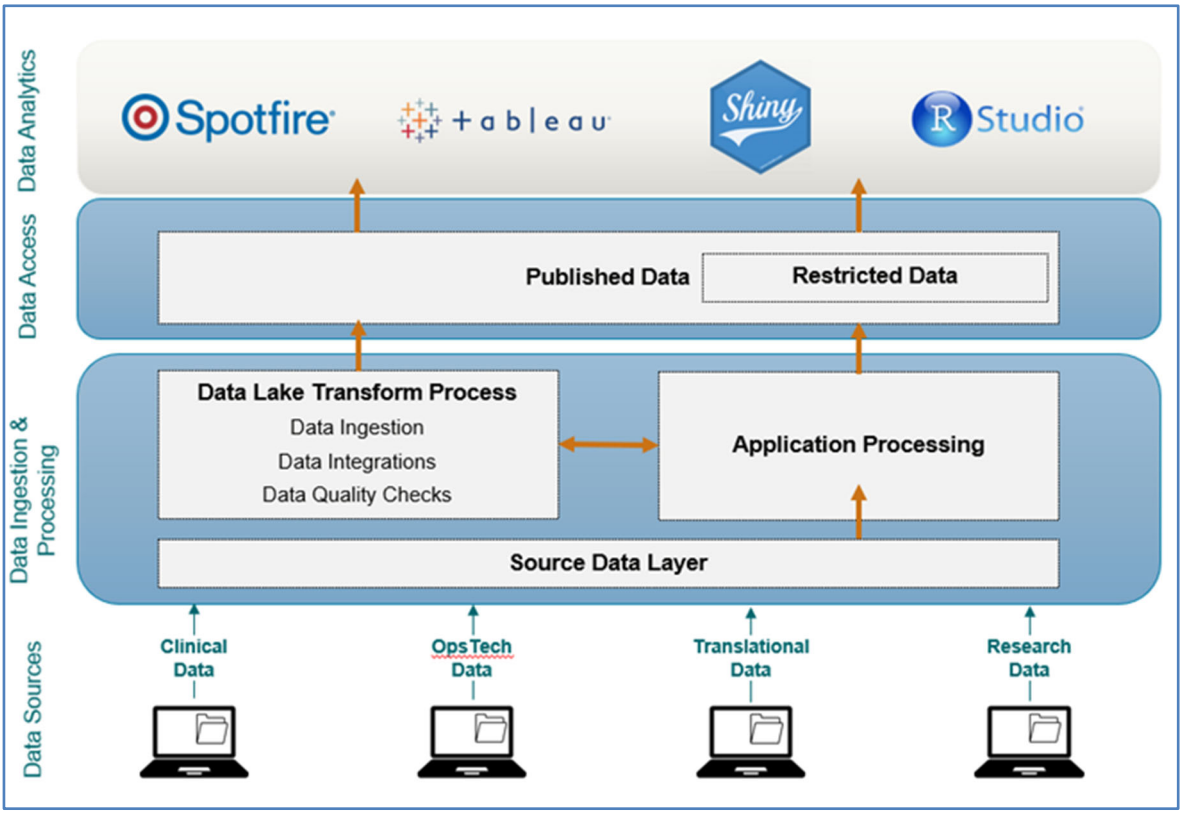


Figure 1: Overview of Data flow.

TOOLS USED FROM AMAZON WEB SERVICES

AWS S3

Amazon Simple Storage Service (S3) is the largest and most performant object storage service for structured and unstructured data. Amazon Simple Storage Service is the storage service of choice to build a data lake. With a data lake built on Amazon S3, AWS services can be used to run big data analytics, artificial intelligence (AI), machine learning (ML), high-performance computing (HPC), and media data processing applications to gain insights from unstructured data sets.

Figure 2: AMAZON S3 Storage architecture.

AWS LAMBDA

AWS Lambda is a serverless, event-driven compute service that lets you run code for virtually any type of application or backend service without provisioning or managing servers.

Figure 3: AWS lambda use case.

AWS SECRETS MANAGER

AWS Secrets Manager helps to protect secrets needed to access the applications, services, and IT resources. This service enables rotating, managing, and retrieving database credentials, API keys, and other secrets throughout their lifecycle. Users and applications retrieve secrets with a call to Secrets Manager APIs, eliminating the need to hardcode sensitive information in plain text.

AWS RDS

Amazon Relational Database Service (RDS) is a collection of managed services that makes it simple to set up, operate, and scale databases in the cloud.

Figure 4: AWS RDS use case

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 transactional, marketing, or mass email communications.

STEPS FOR THE AUTOMATION:

Below are the steps to automate the transfer of files to NFS Mount.

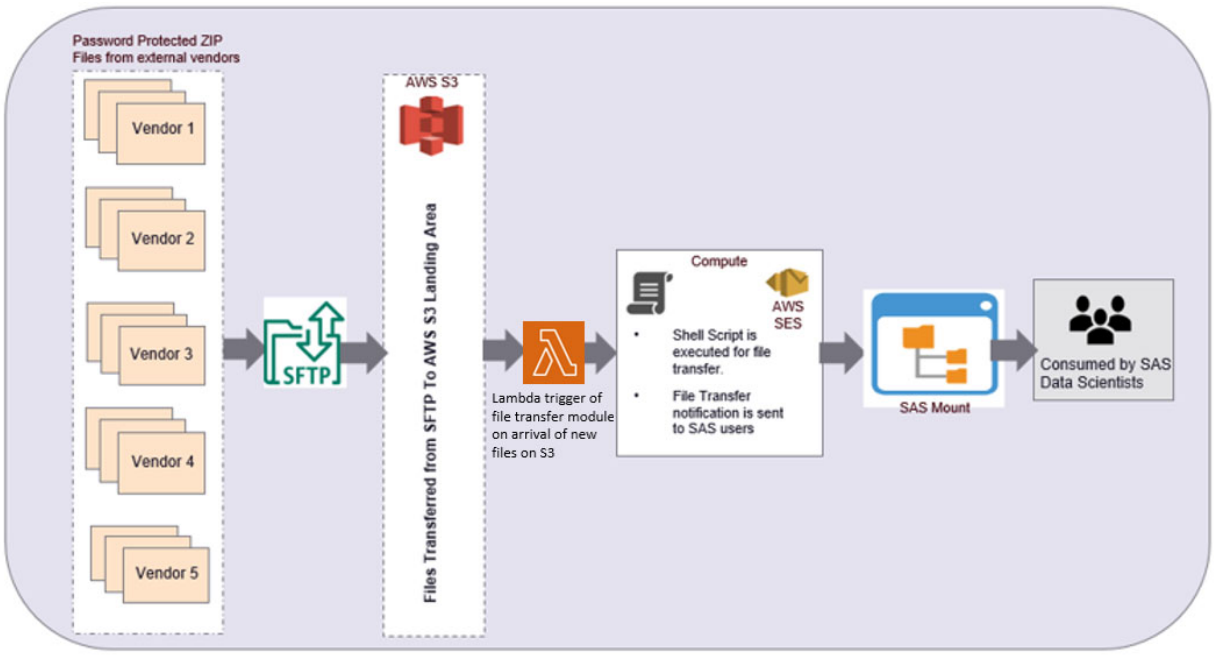


Figure 5: Sample High level flow diagram of the process.

STEP 1: Password protected zip files are posted to SFTP (S3 bucket) from EDC/Translational vendor.

STEP 2: Lambda events are configured on AWS S3 bucket, which are triggered on arrival of files from SFTP. Refer Figure 6.

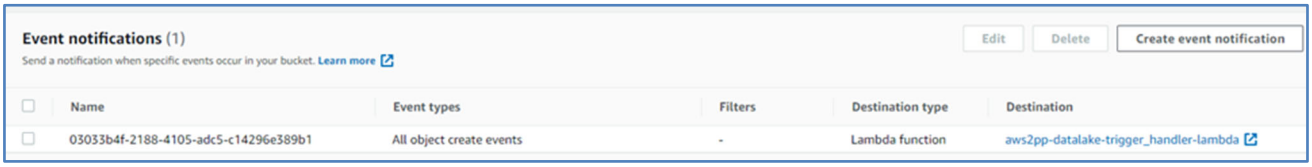


Figure 6: AWS Lambda configured in S3 bucket.

STEP 3: The Lambda events triggered, executes a .sh shell script. The utility of the shell script is to fetch the name of the secrets manager which stores the password for the external file received from vendors, the secret manager name is configured in AWS RDS resource mapped across specific vendors and study names. Refer Figure 7.

```
mysql> select * from ctl_zip_secret_storage ;
+-----+--------------+-----------------+----------------------+
| ID  | SUBJECT_AREA | VENDOR_NAME     | SECRETS_MANAGER_NAME |
+-----+--------------+-----------------+----------------------+
| 111 | study111     | study111_vendor | study111-secret      |
| 222 | study222     | study222_vendor | study222-secret      |
| 333 | study333     | study333_vendor | study333-secret      |
+-----+--------------+-----------------+----------------------+
3 rows in set (0.00 sec)
```

Figure 7: AWS RDS – MYSQL Instance to capture details of Secrets Manager.

STEP 4: Based on the secret name fetched from AWS RDS, we fetch the value of the password from the secrets manager configured in AWS to access the file

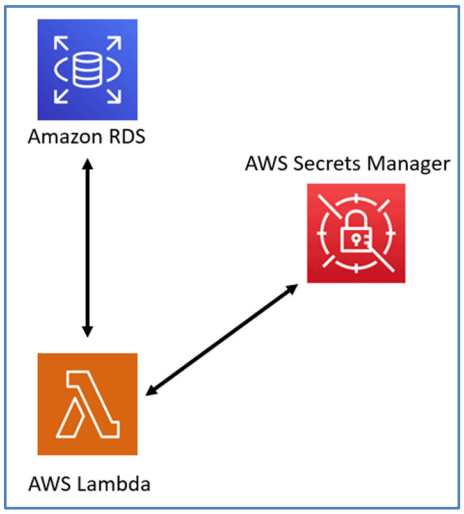


Figure 8: Interaction between Amazon RDS, AWS Secrets Manger, and AWS Lambda.

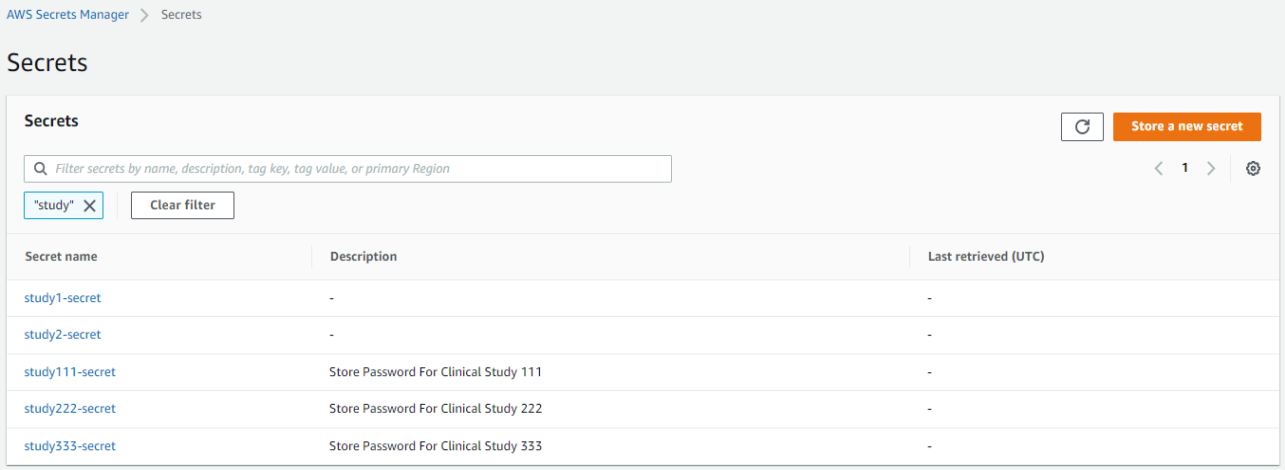


Figure 9: Screenshot of AWS Secrets Manager

STEP 5: Zip files are unzipped using the password fetched from AWS Secrets Manager and synced to NFS mount (SAS Server).

STEP 6: SAS users are notified through automated emails configured in AWS SES about the success or failure of any transfer.

STEP 7: Processing data in SAS server using below SAS programs and shell scripts. REFER Table 1

	Description
Metadata Setup	Create Metadata based on Data Transfer Specification (DTS) for each vendor in an excel file – This metadata will be a key driver in automating the process. We used it in SAS macros and programs. a. To check variable attributes b. to check program names that need to be executed.
Macro 1	Converts any file format to SAS dataset with all variables in Character.
Macro 2	Identifies the latest available data transfer file name & location.
Macro 3	Performs metadata check and compares previous versus the current extract and will notify any differences in log.
Macro 4	a. Dynamically checks for new files in given vendor folders and identifies its corresponding program as per metadata. b. Creates executable file .sh shell script, with a list of vendor programs. c. Creates a log summary of each executed script and notifies users through email.
Macro 5	a. Check the success or failure of the process by checking the datetime stamp of the SAS dataset. b. Create a summary table c. Notifies the users via email (shown in figure 9).
Schedule in CRON	Schedule executable file .sh shell script, created by macro4 in any terminal server.

Table 1: SAS macros and shell scripts

CRON

Cron command is a utility also known as a cron job. It is a job scheduler on linux operating systems, where users can schedule jobs to run periodically at fixed times, dates, or intervals.

CRONTAB SWITCHES

Crontab –l = lists user cron table

Crontab –e = creates a new cron table

Crontab –r removes a cron table and all scheduled jobs.

Syntax: 1 2 3 4 5 COMMAND

1= MINUTE (0-59)

2= HOUR (24 HOUR FORMAT, 0-23)

3= DAY OF MONTH (1-31)

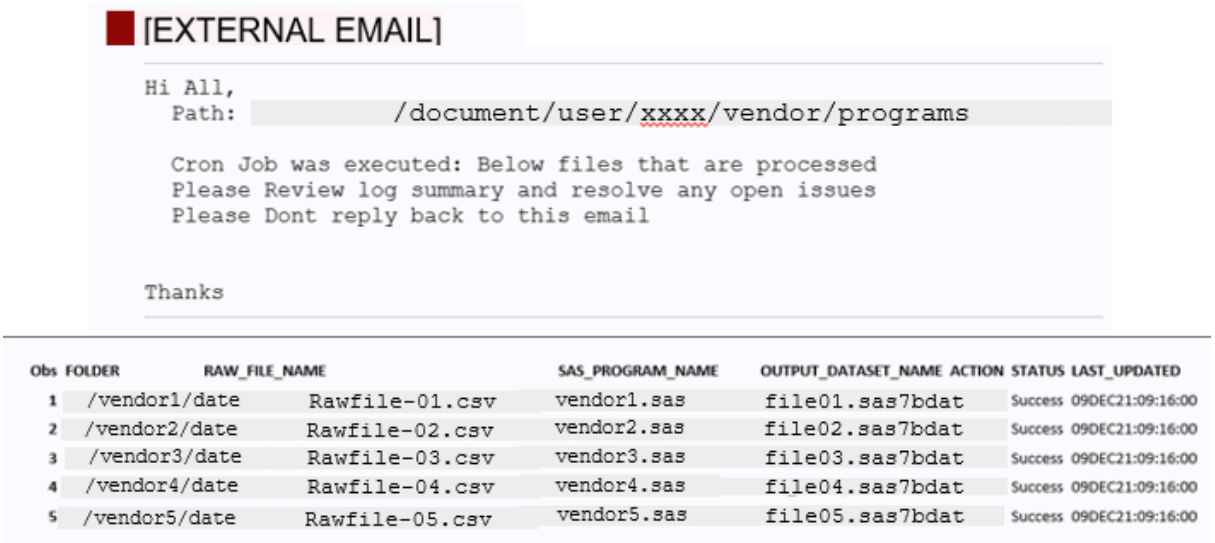
4= MONTH OF YEAR (1-12)

5= DAY OF WEEK (0-6, 1-6, 0,7=SUNDAY).

EXAMPLE TO SCHEDULE SCRIPT FILE AT 1AM EVERY DAY

01 00 * * * /study/level_1/level_2/scripts /shell_script.sh

shell_script.sh contains sas program/macros (refer Table 1) that are designed for processing data in sas environment. Macro 5 in Step7 will send out an email notification as shown in figure 10 below.



Obs	FOLDER	RAW_FILE_NAME	SAS_PROGRAM_NAME	OUTPUT_DATASET_NAME	ACTION	STATUS	LAST_UPDATED
1	/vendor1/date	Rawfile-01.csv	vendor1.sas	file01.sas7bdat		Success	09DEC21:09:16:00
2	/vendor2/date	Rawfile-02.csv	vendor2.sas	file02.sas7bdat		Success	09DEC21:09:16:00
3	/vendor3/date	Rawfile-03.csv	vendor3.sas	file03.sas7bdat		Success	09DEC21:09:16:00
4	/vendor4/date	Rawfile-04.csv	vendor4.sas	file04.sas7bdat		Success	09DEC21:09:16:00
5	/vendor5/date	Rawfile-05.csv	vendor5.sas	file05.sas7bdat		Success	09DEC21:09:16:00

Figure 10: Status email notification to users.

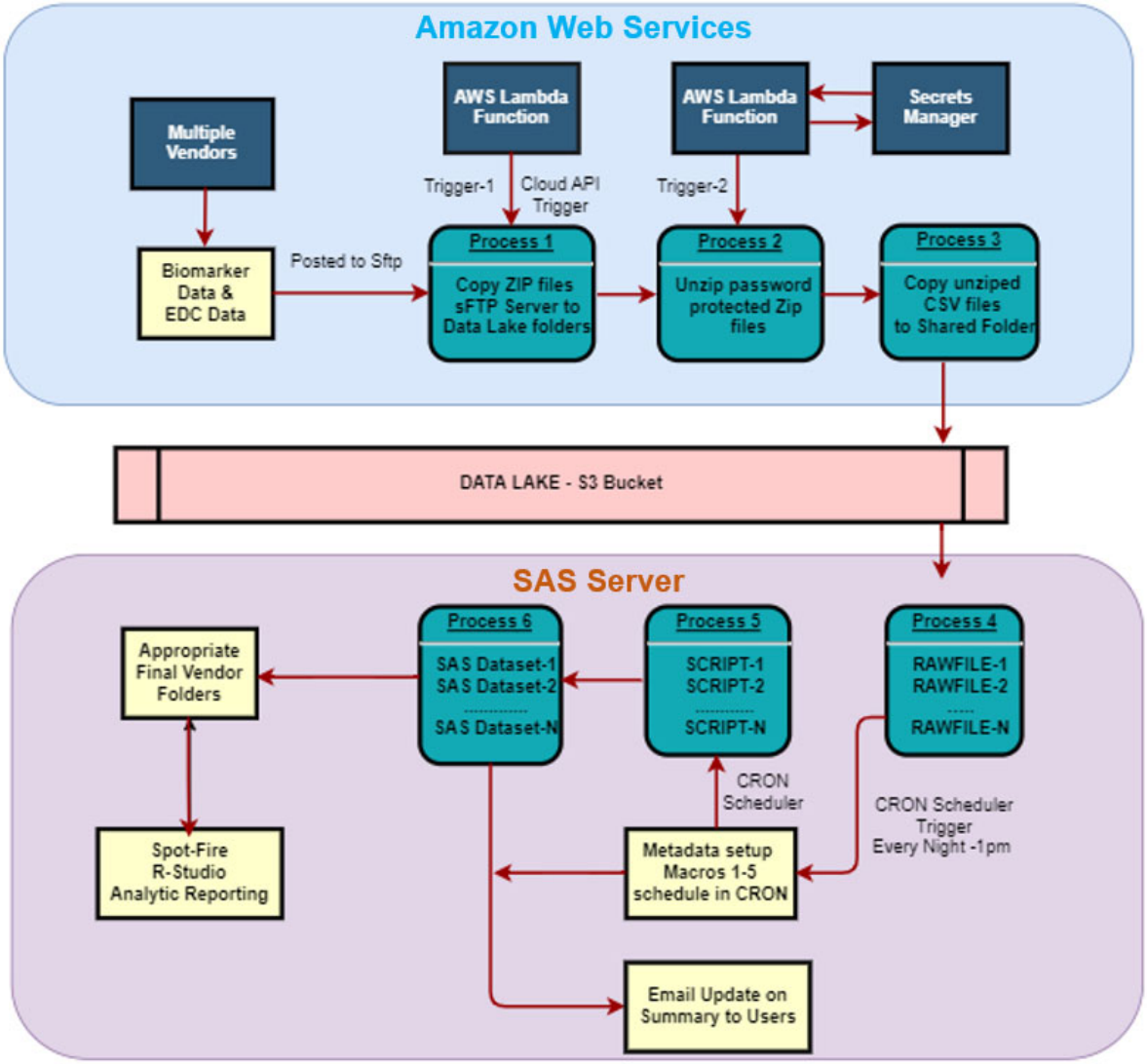


Figure 11: Overall Flow diagram.

CONCLUSION

By harnessing the power of cloud-based tools integrating with SAS, we can create highly efficient and cost-effective process of transferring, storing, transforming, and analyzing clinical data. Large amounts of genomic and biomarker data can be handled with high durability, availability, and scalability. With the help of in-built analytical tools and automation tools, clinical data can be made available securely to the end users such as biostatisticians, clinical and translational scientists as soon as the data gets posted into the data lake environment without any programmer's intervention.

REFERENCES

1. AWS lambda

   https://aws.amazon.com/lambda/

2. Amazon Relational database service documentation

   https://docs.aws.amazon.com/rds/?id=docs_gateway

3. Amazon Simple Storage Service Documentation.

   https://docs.aws.amazon.com/s3/?id=docs_gateway

4. AWS Secrets manager
   https://aws.amazon.com/secrets-manager×

ACKNOWLEDGMENTS

We would like to thank Babu Kish, Senior Manager, IT Infrastructure and Cloud, Saurav Mahanti, Senior Director, Data Management Analytics and Integration, and Debi Prasad Roy, Executive Director, Head of R&D Tech, Data Science for their support in this automation process and suggestion/comments.

RECOMMENDED READING

- *Amazon S3 user guide*
- *Amazon lambda user guide*

CONTACT INFORMATION

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

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

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

Name: Chaitanya Chowdagam
Enterprise: Ephicacy Consulting Group, Inc.
E-mail: chowdagam@gmail.com

Name: Jegan Pillaiyar
Enterprise: Ephicacy Consulting Group, Inc.
E-mail: jeganpillayar@gmail.com

Name: Kunal Chattopadhyay
Enterprise: ZS Associates
E-mail: kunal.chattopadhyay@zs.com