New Study On-Boarding

SOP

Table of Contents

Version Control				
Version	**Date**	**Description of Changes**	**Author**	**Approved By**
1.0	October 13, 2021	SOP for New Study On Boarding, first version	Risabh Jain, Parth P	Kamlesh Gupta

Objective:

The objective of this document is to explain the steps of new study on boarding in Data Lake, CPT and SDTM.

Scope:

1. Request for on boarding new study.
2. Arrowhead send clinical data files.
3. Circulants provide an analytic dashboard to client.

Reference:

Technical Design Document.

Responsibility:

Role	Responsibility
Arrowhead Data Manager/CRO/Vendor	• Send user's details to Circulants. • Send data files for New Study On-Boarding.
Circulants Team	• Create folder structure in sFTP Landing. • Create Storage Account and Containers. • Create Databricks configs. • Create Synapse Data Source, data views and table. • Create DQM mapping. • Files and target mapping files. • Test the process before commit to production. • Monitor & resolve all alerts in the process.

Pre-requisite:

Sr. No.	Necessary Details
1	Study Details (Name, Due date, Comments etc.)
2	sFTP access user’s details
3	sFTP Notification Recipients details
4	Vendor details (Name, File Type)
5	Users’ name, email id, access to data type
6	Email Distribution lists for success and failure notifications
7	Data files of the new study. (csv/sas/zip/other)

Table 1 Pre-Requisite Table

Study On-Boarding Process:

New Study On Boarding should be followed on board a new study in Data Lake, CPT and SDTM Data mart. Study on boarding comprises getting test data from Arrowhead, creating data structure based on provided data, on boarding the study in Test Environment, testing and finally migrating to production environment. Based on Access Request Form (ARF) access is given to required users thereafter.

Study on boarding steps are mentioned below:

1. New Study Pre-requisite Details:

   1.1. Make sure that all details of new study are received by development team, referring **Pre-Requisite Table**.

2. sFTP Directory Creation:

   2.1. sFTP stands for Secure File Transfer Protocol which helps us to receive the data in a secure manner with well-defined user access levels and business rules validation on received data.
   2.2. With the Help of details provided by Arrowhead folder structure is created in for user using format – User Name/Program Name/Study Name/Vendor Name/Data Type (Data/Unblinded). For example – *k.reis/apoc3/aroapoc32001/medpace-reference-lab/data.*
   2.3. Similarly, Folder structure is created in Landing Zone of SFTP server using format - /SFTPdata/arwruser/program name/study name/vendor name/data type. For example - */Sftpdata/arwruser/apoc3/aroapoc32001/medpace-reference-lab/da*ta will create the folder structure in the ARWR (sFTP landing).These folder structures are created by Circulants Infra team.
   2.4. To understand the file structure of sFTP please refer **Figure 3** & **Figure 4**.

3. Data Lake Storage Account Creation:
   3.1. Azure Data Lake is the file destination in which all the studies or programs are stored in a structured manner.
   3.2. A Storage Account is a parent storage location (example- AROHIF21001). In each storage account all the study specific vendor directories (containers) are created. As a result, there are several storage accounts, each carrying data from a different study or program.
   3.3. The DevOps team will be in charge of creating this Storage Account and assigning user's access.

4. Container Creation in Storage Account:
   4.1. Container is sub directory of Storage Account where the on boarded data is stored.
   4.2. Any given study may have numerous vendors, and any given program (cpt/sdtm) may have multiple studies. therefore, the container facilitates in the branching of the data correspondingly.
   4.3. For study: Containers will be generated for each vendor. The Storage Account is referred as the Study, and the naming convention will be Vendor_Name-Data_Type (ex. keystone-Unblinded).
   4.4. For program: Containers will be generated for each study when the Storage Account is referred to as a Program(cpt/sdtm). Create two sub directories named current and archive in the container.
   4.5. The DevOps team will be in charge of creating these containers and assigning users access.
   4.6. Please refer **Figure 5** and **Figure 6** for clear understanding.

5. Mount the container in Databricks:
   5.1. Databricks is a Platform as a Service (PaaS) that provides a unified data analysis system to organizations
   5.2. Databricks facilitates writing and executing programs in the cloud. It helps in processing of data to the Data Lake.
   5.3. Using the Databricks automated pipeline has been written in the notebooks to connect the Data Lake. Using the notebook user can write program to access the sFTP, Synapse, Cosmos DB, Power BI.

5.4. To access any storage account and underlying containers in Databricks, we need to mount(link) it in Databricks as a primary step. It needs to be mounted only once.
5.5. While study or vendor onboarding developer can check if the container is already mounted or not.
5.6. The development team is responsible to mount new containers in Databricks.

6. Target Mapping & DQM files Creation/Updation: (Only For Program(cpt/sdtm) based onboarding)
6.1. In the cpt/sdtm program there are multiple studies are on-boarded. All the studies are having similar file structure
6.2. To integrate the studies in a properly, the target Mapping file assists Databricks code in mapping all the study with the synapse target columns. this file needs to be placed in the publish folder on cpt/sdtm Storage Account.
6.3. To validate a study related file, DQM file contains the columns for the individual files in the cpt and sdtm. It consists of all the column names of individual files and the same is utilized by Databricks for the data quality checks. This DQM file is placed in the publish folder of cpt/sdtm Storage Account.
6.4. When New Program like cpt or sdtm gets On-Boarded, these two files (DQM & Source target mapping) needs to be created.
6.5. When New Study gets On-Boarded in any existing program, these two mapping files needs to be updated by development team manually.
6.6. The above steps are utilized to create the Common data model(cdm).
6.7. The Data Lake Development team is in-charge of this work.

7. Synapse Configuration:
7.1. Synapse is a serverless Azure Cloud service. It provides the execution engine of data warehousing and Big Data Analytics services.
7.2. In our implementation and study-based data or common data model (CDM)(cpt/sdtm) can be queries using synapse. It contains the Views and Tables used by POWER BI, SPOTFIRE or any Analytics tools.
7.3. To Configure Synapse First it requires to create Synapse Data Source which helps to connect and fetch the data from Data Lake (Source).
7.4. Secondly create different tables using source data as per requirements.
7.5. Finally, we may construct views using the tables based on business requirements.
7.6. For study: It is required to create above mentioned Data Source, tables, and views.
7.7. For program: If a new program is being on-boarded, then (Data source, Schema) needs to be created but if new study on-boarded then it just needs to be validated/updated.
7.8. The Data Lake development team is responsible for this task.

8. Mention Email Ids for Notification:
8.1. It is critical to be alerted for each automated pipeline, whether it is successful or not. If there is a failure, at what point did it fail?
8.2. To implement this in the project we have created well organized automated mail system at all the checkpoints of the pipeline.
8.3. In order to subscribe any user to this mail system for any certain study/program, add the mail ids to corresponding config files.
8.4. The Data Lake Developer team will initiate this task.

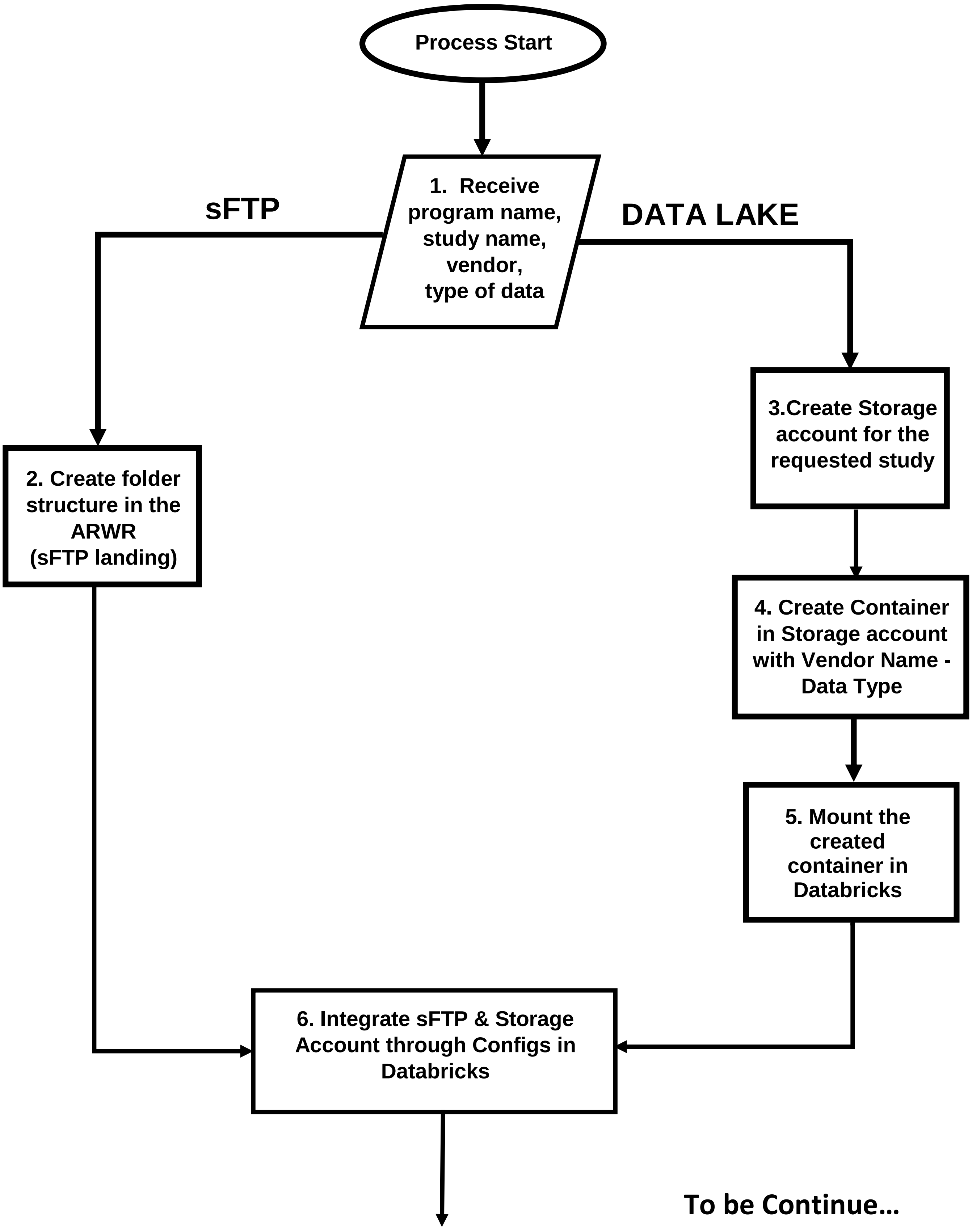
9. Power-BI Dashboard:
9.1. Power BI is a business analytics service by Microsoft. It aims to provide interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards.

9.2. We are using this Power BI to create the visualization reports.
9.3. Based on Client Requirements the dashboard report been designed the reports are create in power and published with the access assigned to a designated user.
9.4. Power BI development team is responsible for this implementation.

10. Nightly Scheduler:
10.1. To execute the data ingestion pipeline in an automated manner Nightly Scheduler notebook is used.
10.2. It is scheduled to run between Monday-Friday (8:30 PM IST) for a period of 12 hours.
10.3. When new data will arrive in the program/study, nightly scheduler will execute. It internally calls the dependent notebooks like sdtm, cpt, study-onboarding and power BI trigger.
10.4. Data Lake development team is maintaining this Nightly Scheduler.

11. Testing & Deployment on Production:
11.1. After On-Boarding New Study/Program Testing of the Pipeline is must. Once it pass through all testing then only it can be move to production environment.
11.2. The Data Lake development team is responsible for this testing and Senior Developer of Data Lake team will be migrate the tested code to Production Environment.

Flow Chart 1: New Study On-Boarding in Data Lake


Process Start
1. Receive program name, study name, vendor, type of data
sFTP
DATA LAKE
2. Create folder structure in the ARWR (sFTP landing)
3.Create Storage account for the requested study
4. Create Container in Storage account with Vendor Name - Data Type
5. Mount the created container in Databricks
6. Integrate sFTP & Storage Account through Configs in Databricks

To be Continue…

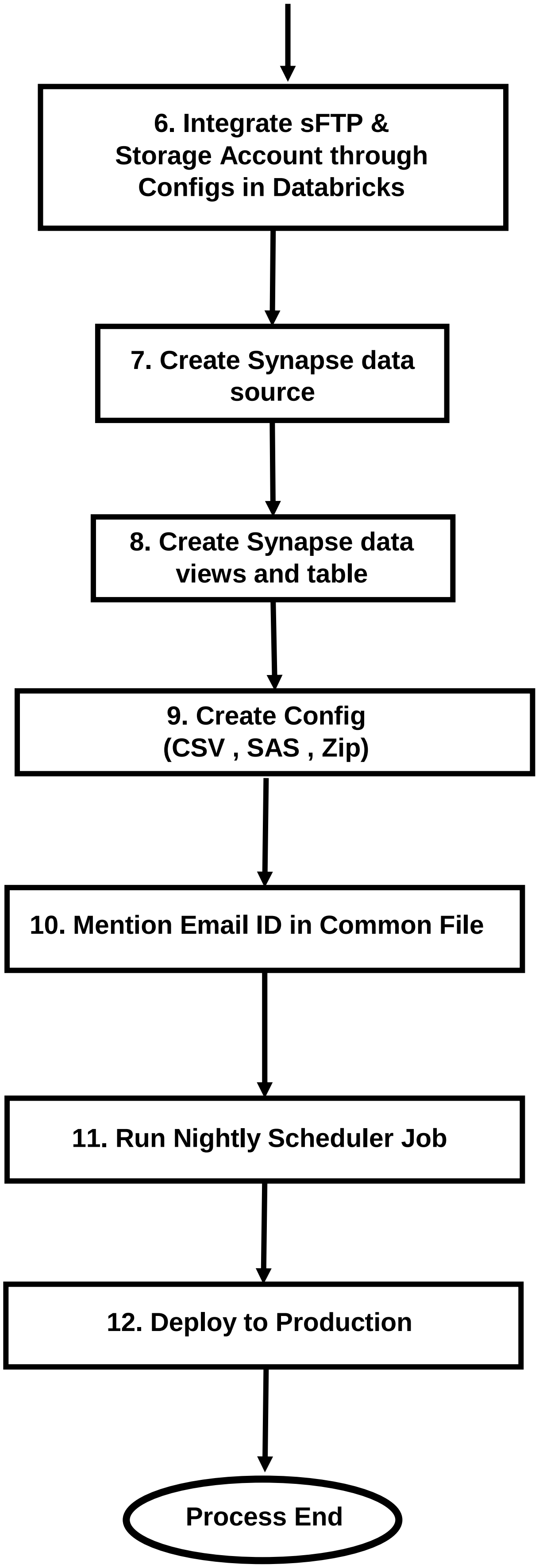
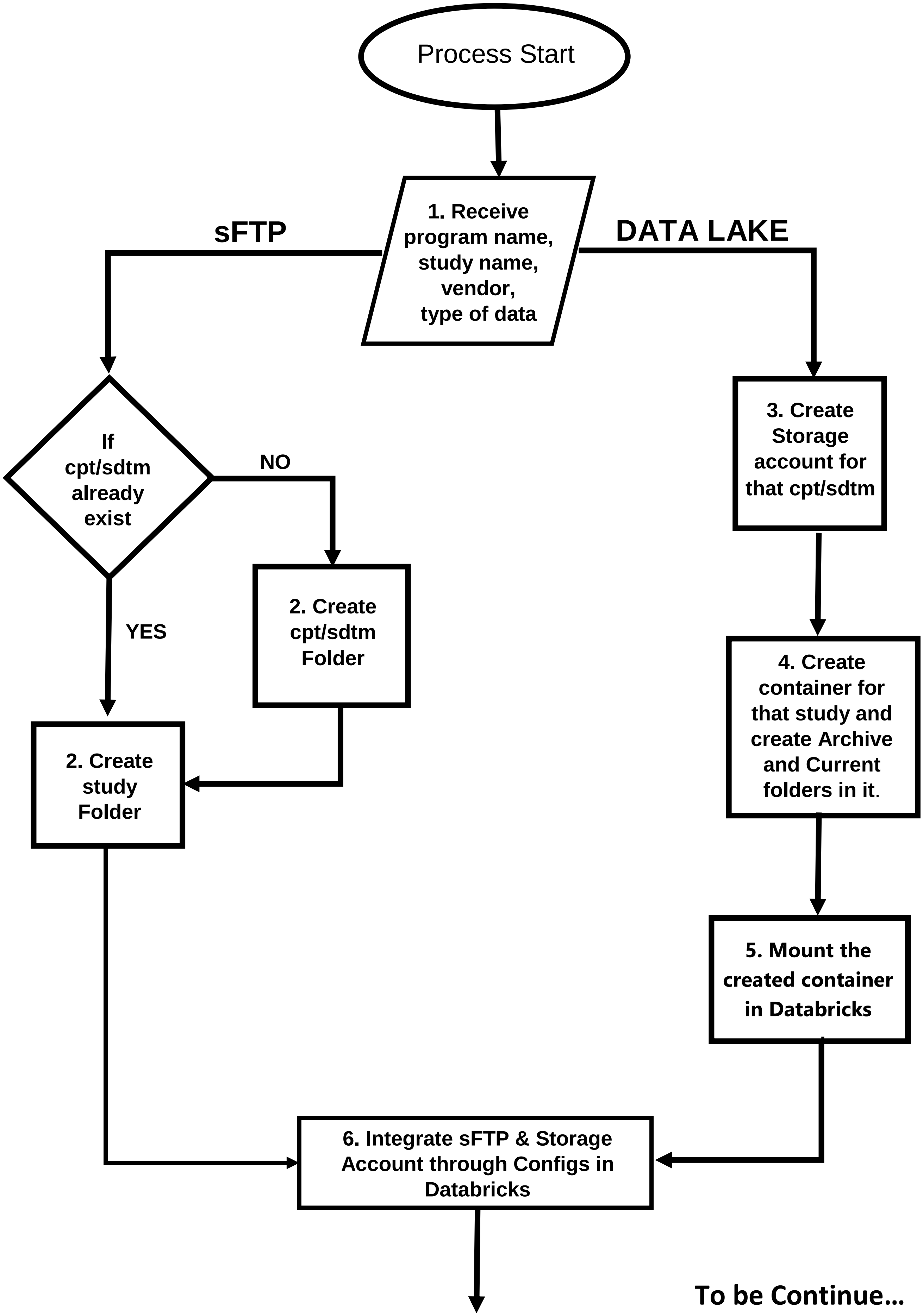

6. Integrate sFTP & Storage Account through Configs in Databricks
7. Create Synapse data source
8. Create Synapse data views and table
9. Create Config (CSV , SAS , Zip)
10. Mention Email ID in Common File
11. Run Nightly Scheduler Job
12. Deploy to Production
Process End

Figure 1 New Study On-Boarding for Data Lake

Flow Chart 2: New Study On Boarding for CPT/SDTM


Process Start
1. Receive program name, study name, vendor, type of data
sFTP
DATA LAKE
If cpt/sdtm already exist
NO
YES
2. Create cpt/sdtm Folder
2. Create study Folder
3. Create Storage account for that cpt/sdtm
4. Create container for that study and create Archive and Current folders in it.
5. Mount the created container in Databricks
6. Integrate sFTP & Storage Account through Configs in Databricks

To be Continue...

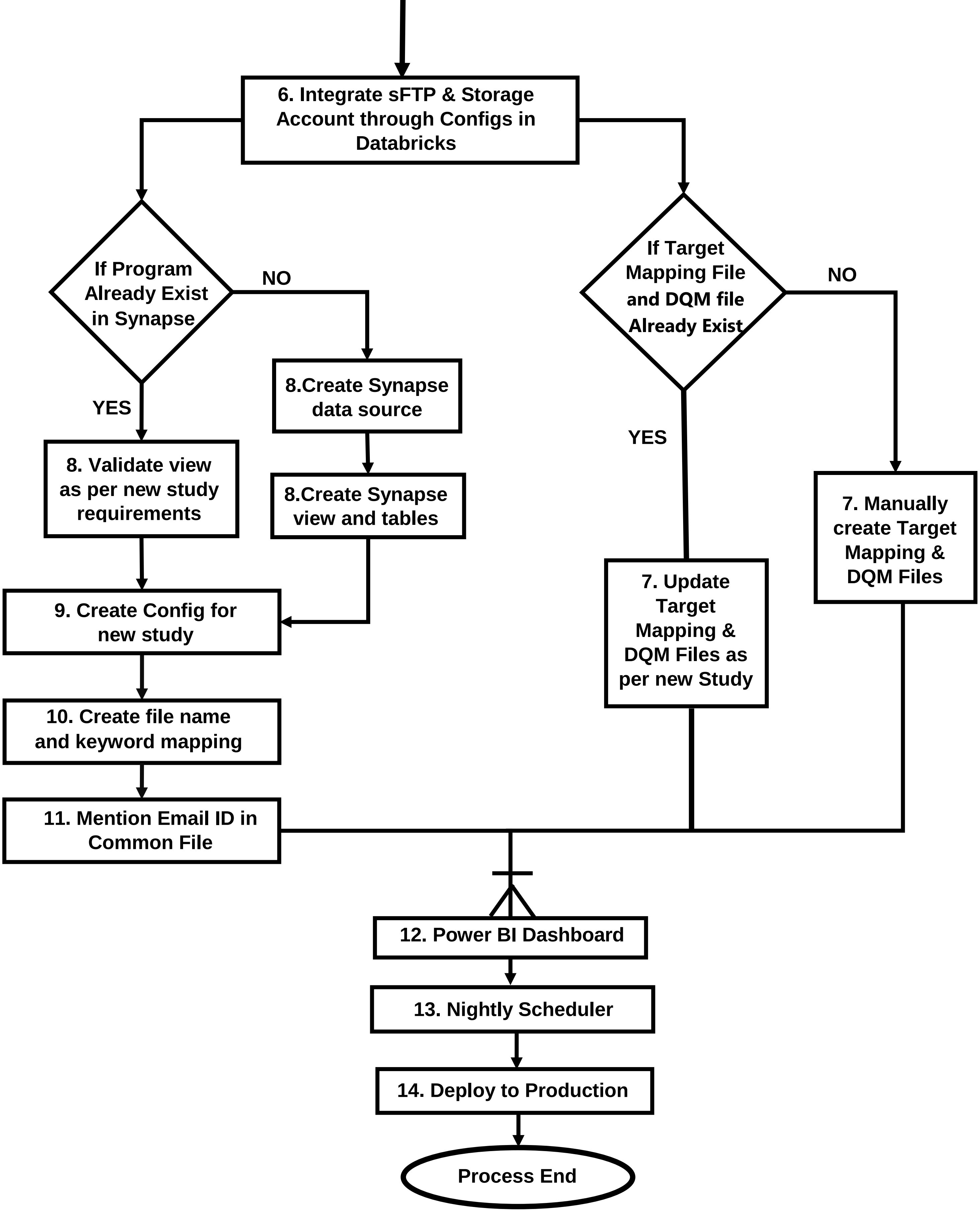

6. Integrate sFTP & Storage Account through Configs in Databricks
If Program Already Exist in Synapse
NO
YES
8.Create Synapse data source
8.Create Synapse view and tables
8. Validate view as per new study requirements
9. Create Config for new study
10. Create file name and keyword mapping
11. Mention Email ID in Common File
If Target Mapping File and DQM file Already Exist
NO
YES
7. Manually create Target Mapping & DQM Files
7. Update Target Mapping & DQM Files as per new Study
12. Power BI Dashboard
13. Nightly Scheduler
14. Deploy to Production
Process End

Figure 2 New Study On-Boarding for CPT/SDTM

SFTP Folder Structure for Data Lake

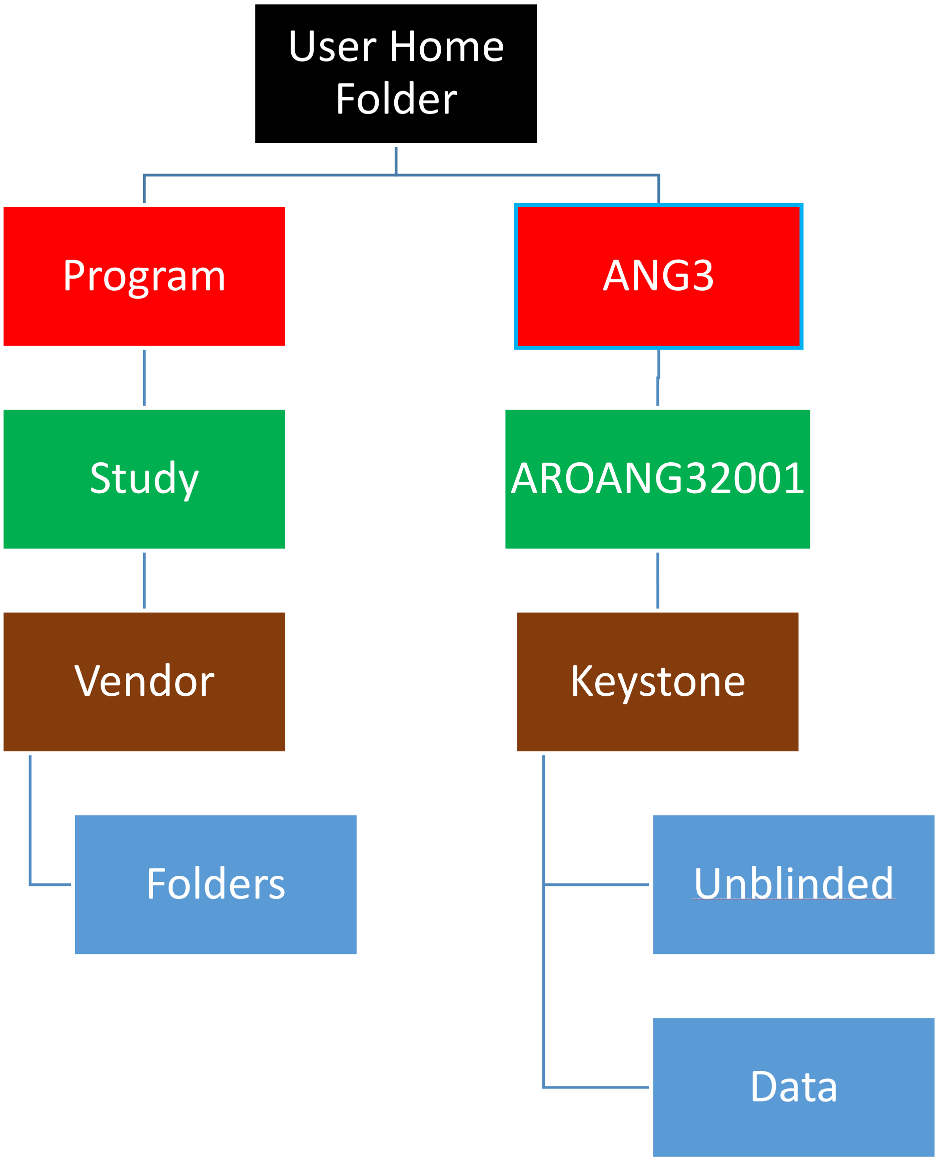


Figure 3 SFTP Structure – Data Lake

SFTP Folder Structure for SDTM and CPT

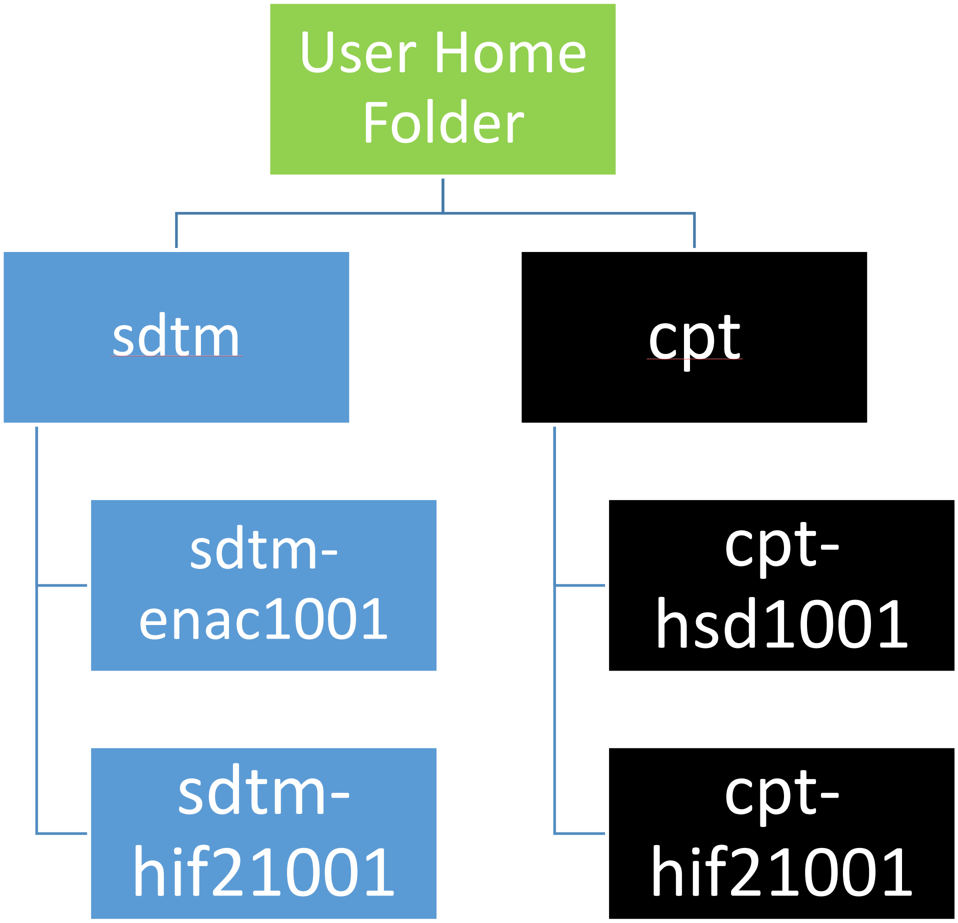


Figure 4 SFTP Structure – SDTM and CPT

SDTM Storage A/C Architecture

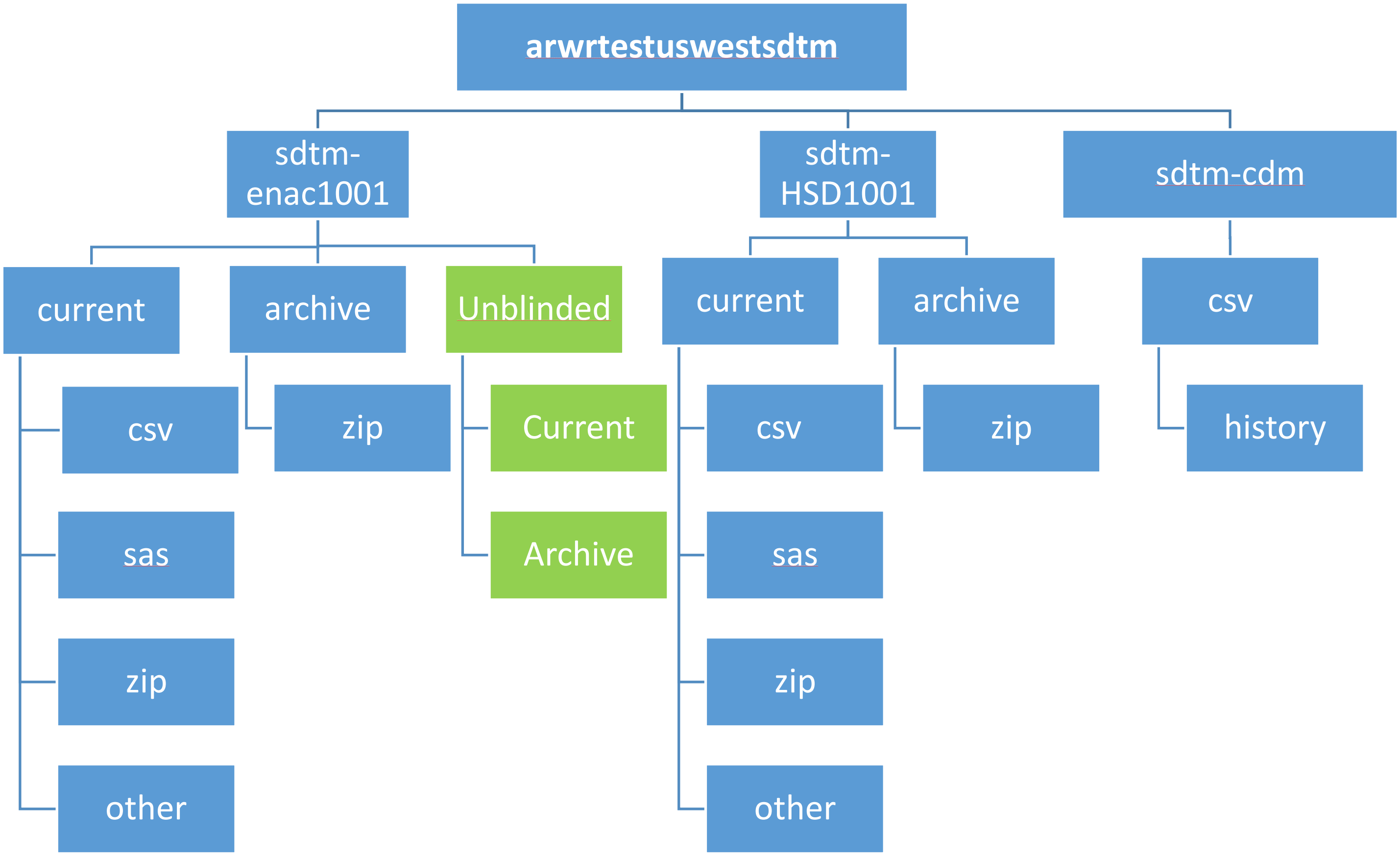


Figure 5 SDTM Storage A/C Architecture

CPT Storage A/C Architecture

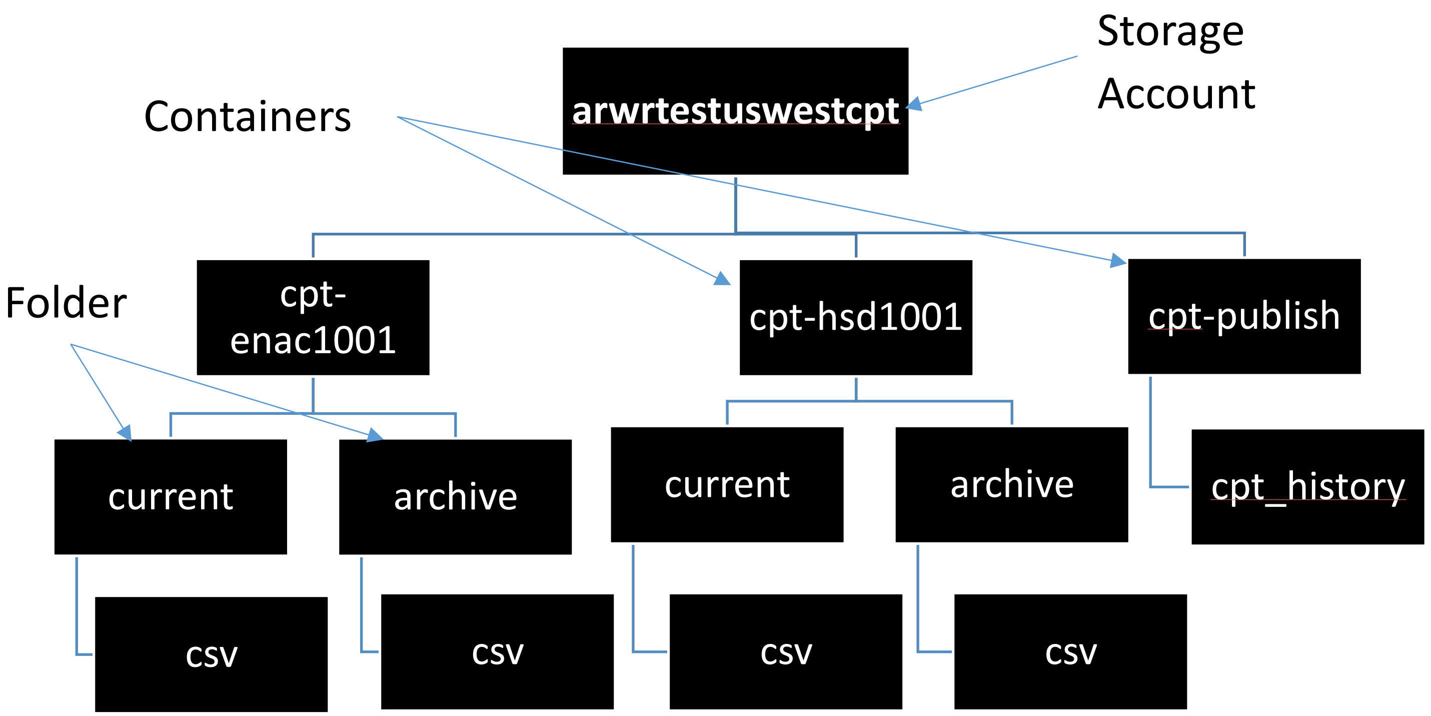


Figure 6 CPT Storage A/C Architecture