cQube – Technical Specifications Document

July 2021

Current Version 3.0

Document released by:
Sreenivas Nimmagadda
(System Architect)

Document reviewed by:
Sateesh Pullela
(Technical head)

Document acceptance by:
Arvind Gopalakrishnan
(Delivery head)

Contents

1. Background and Context

EkStep and Tibil Solutions have embarked on a project named as 'cQube' to create an analytics and decision making tool for the education system. This product can be used for monitoring the education system in a state and on a broader scale for monitoring the schools across various levels of administration.

1.1 Use of this document

This document will cover the technical points of following areas:

1. Data emission process.
2. Role of Java and Python programming.
3. Nifi Processor information - Processor Groups, Data validations, Query configurations, Output file formats.
4. Prometheus and Grafana monitoring tool connectivity.
5. Database and Data modelling information.
6. Node connectivity with S3 bucket.
7. Angular JS, Chart JS and Leaflet roles in the visualization.
8. Keycloak integration with authentication process.

9. Logs.

1.2 State of this document

This document is an evolving document and is under change control. To request a change to the technical document please contact the system architect or the project manager.

1.3 Acronyms

The following is a list of acronyms which will be used throughout this document:

Table – 1: Acronyms

Acronym	Description
S3	Simple Storage Service
NIFI	Apache NIFI
PK	Primary Key
FK	Foreign key
RDAC	Restricted Database Access Control

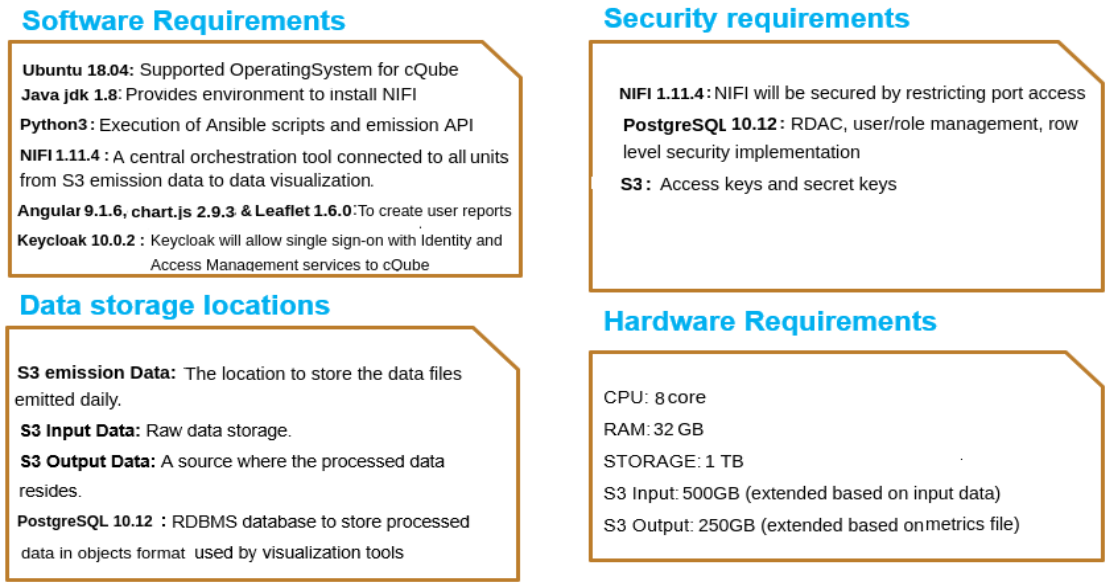
2. cQube Product setup

This section describes the prerequisites to install and configure the cQube product setup and the cQube product setup process. This section also describes the cQube network setup and the setup process

2.1 cQube product prerequisites

The cQube product installation process has a few system prerequisites that need to be followed. The system software, hardware, and security requirements have been derived and have to be adhered to before installation. The figure below gives an overview of the software requirements, security requirements, hardware requirements and the data storage locations:

Figure – 1: System hardware, software requirements, security requirements and data storage



Mentioned below are the prerequisites for the installation of the cQube product:

- The cQube product has to be installed in the AWS environment.
- Firstly an instance with Ubuntu 18.04 OS has to be created.
- All the hardware requirements mentioned in Section 4.1.4 have to be adhered to.
- Before starting installation, the network set-up has to be completed as mentioned in Section 5 of this document.

The below table describes the Infrastructure details to install the cQube

Activity	**Infrastructure used for cQube installation**	**Required Skills**
1-2 Weeks	2 Days	
Filling of the requirements in the required format and obtaining relevant approvals for procurement of the Infrastructure and securing funding for monthly/yearly spend.	AWS account 1. cQube Server (CPU - 8 Core RAM - 32 GB Storage - 500 GB) 2. S3 Buckets (S3 emission 100GB S3 input - 750 GB S3 Output - 250 GB)	AWS Basic Operation Skills - EC2 - S3 - Load Balancer - IAM - VPC - Route53 - Certificate Manager (SSL)
	OpenVPN Access Server (EC2 instance CPU - 1 Core RAM - 2 GB, Storage - 10 GB)	VPN admin Operations
	NGINX Reverse proxy(EC2 instance CPU - 2 Core RAM - 4 GB Storage - 30 GB)	NGINX configuration skills
	NAT gateway	AWS NAT gateway
	Ubuntu 18.04	Will be taken care while creating EC2 instance
	Java jdk 1.8	Will be installed through One-Step cQube Installation
	Python 3	Will be installed through One-Step cQube Installation
	NIFI 1.11.4	Will be installed through One-Step cQube Installation
	Angular 9.1.6, Chart.JS 2.9.3, Leaflet 1.6.0	Will be installed through One-Step cQube Installation
	PostgreSQL 10.12	Will be installed through One-Step cQube Installation

2.1.1 Software Requirements

Ubuntu 18.04: This is the operating system that supports the cQube product

- Java JDK1.8: This provides an environment for NIFI installation
- Python3: Python plays a role in the execution of Ansible scripts and data emission API using a virtual environment.
- NIFI 1.11.4: A central orchestration tool which is connected to all the units and all the different sections where the data flows, starting from the S3 emission data location to the data visualization stage
- PostgreSQL 10.12: An RDBMS database to keep all the processed data in relational data format. The data stored here is used by NIFI to prepare the JSON format files which can be used in the visualization charts.
- Angular 9.1.6 + ChartJS 2.9.3 + Leaflet 1.6.0: Angular, ChartJS and Leaflet are used to create dashboards/ user reports. The data stored in the S3 output bucket can directly be used to create the reports.

2.1.2 Security requirements

- cQube product security will be provided by implementing the private subnet with AWS load balancer as described in the Section 2 of this document.
- All the ports will be accessed by Nginx server only, So those ports will not be accessed directly from the internet.
- S3 buckets will be secured by the AWS default security.
- PostgreSQL will be secured by the following ways
  (The database security will be implemented in the future versions of cQube)
    - **RDAC:** Postgres provides mechanisms to allow users to limit the access to their data that is provided to other users. Database super-users (i.e., users who have `pg_user.usesuper` set) silently bypass all of the access controls described below with two exceptions: manual system catalog updates are not permitted if the user does not have `pg_user.usecatupd` set, and destruction of system catalogs (or modification of their schemas) is never allowed.

- **User and Role Management:** the user roles will be granted with one or more cQube database will have three different types or roles:
    1. role role (identified by prefix r_)
    2. group role (identified by prefix g_)
    3. user role (generally personal or application names)
- **Row Level Security:** In addition to the SQL-standard privilege system available through GRANT, tables can have row security policies that restrict, on a per-user basis, which rows can be returned by normal queries or inserted, updated, or deleted by data modification commands. This feature is also known as Row-Level Security. When row security is enabled on a table all normal access to the table for selecting rows or modifying rows must be allowed by a row security policy.

2.1.3 Data Storage Locations

- S3 emission data Location: The data emitted from the state education system has to be stored until the NIFI reads the data. S3 emission storage buckets are the storage locations where the emitted data files are stored.
- S3 Input Data: This is a location where the raw data resides for all future references.
- PostgreSQL: All the transformed and aggregated data will be stored in PostgreSQL tables.
- S3 Output Data: A location where the processed data resides in JSON format.

2.1.4 Hardware Requirements

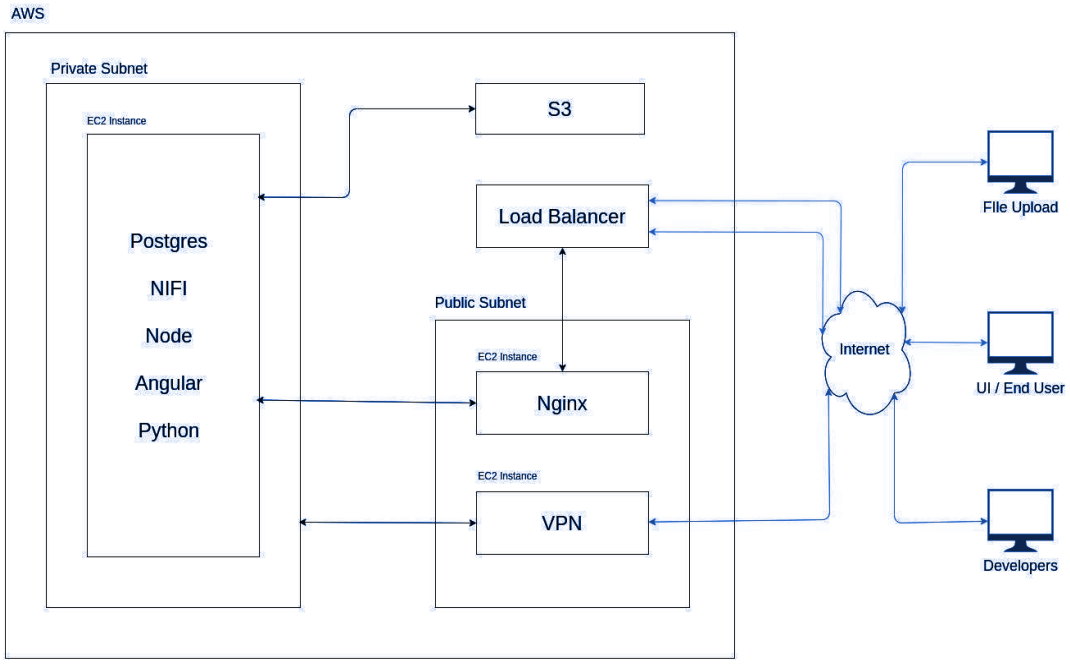
Listed below are the minimum hardware specifications/ requirements in order to install the cQube product

- 8 core CPU
- 32 GB RAM
- 500 GB - 1 TB harddisk
- Three S3 Buckets will be used - for cQube data emission, cQube data input and cQube data output

2.2 cQube network setup

The cQube network setup consists of the AWS, which encompasses the private subnet section which contain the EC2 instances Postgres, NIFI, node, angular and python, the public subnet section which comprises of the Nginx and VPN, S3 and the Load balancer. The cQube network setup process is described in the block diagram below:

Figure – 2: cQube network setup diagram



2.2.1 End User

- When a user accesses the cQube application through the browser, the request hits the load balancer of the AWS.
- Load balancer will forward the request to Nginx proxy server.
- Nginx will forward the request to angular application using private IP
- Angular sends the request to NodeJS server
- NodeJS will get the data from S3 bucket and respond to Angular
- Angular will process the data and display the results.

2.2.2 Emission User

- These users call the Rest API through python client-side script to upload the files
- After user authentication, Rest API (Written in python) provides the S3 one-time URL
- Client-side python upload the files using the S3 one-time URL

2.2.3 Developers

- Normally developers can deploy the changes through the Jenkins CI/CD pipeline.
- Developers use VPN to connect to the cQube production application if there are any direct changes, like technical changes, configuration or customizations are required.

2.3 cQube - for Installation

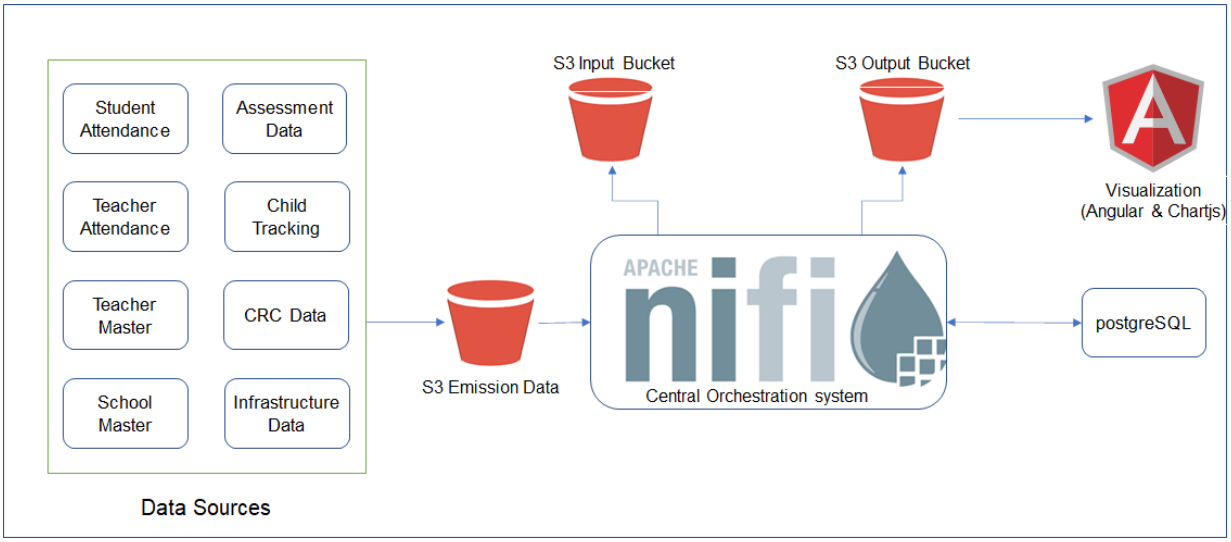
The cQube product can be installed as a one-step installation process. The one-step process installs the complete cQube stack, which includes Ansible automation scripts to install Java, Python, NIFI, Angular, ChartJS, Leaflet, S3 emission data, S3 input bucket, S3 output bucket and PostgreSQL installations.

Steps for Installation:

- As a first step, clone the files from GitHub using the following command:
  $ git clone https://github.com/project-sunbird/cQube.git.
- This downloads the cQube installation files. The cQube GitHub has all the installation files required for installing cQube.
- The README.md file has all the instructions that have to be followed for the cQube product installation.
- The install.sh script installs the complete cQube stack.
- The Install.sh file calls the Ansible playbooks in the background which will complete the cQube installation setup.
- The complete installation process takes approximately 30 minutes.
- Once the installation is complete, the message “cQube successfully installed” is displayed.

3. Software Architecture

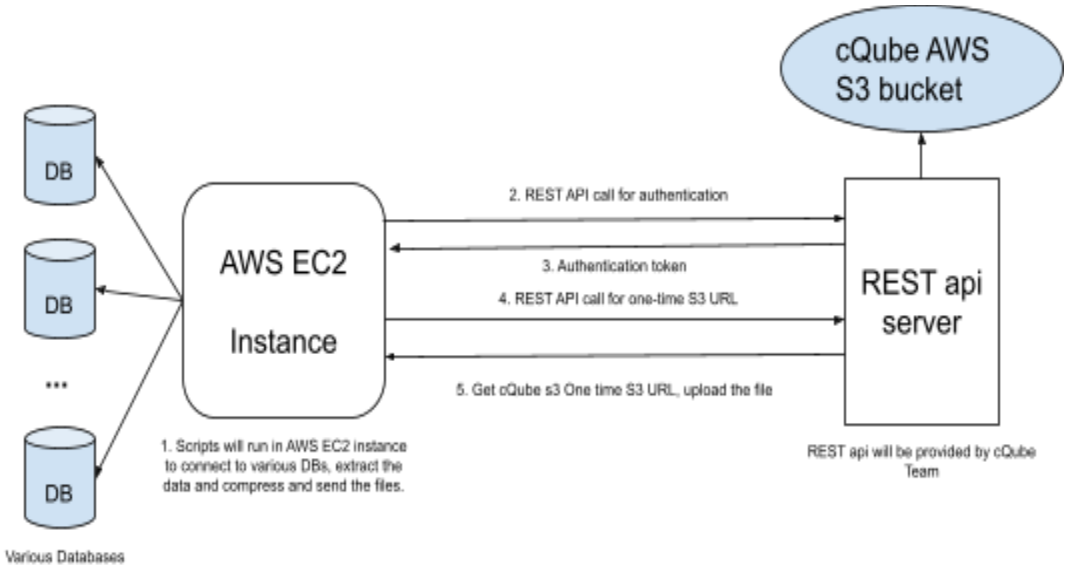
Figure – 3: cQube software architecture



3.1 Data emission process

cQube provides an API for the data ingestion. Authenticated API call provides onetime S3 URL and emits the data into the S3 emission bucket.

Figure – 5: Data emission process



The steps involved in the data Emission Process

- The data emitters will work from the data source location.
- Data emission will be performed periodically as per the specified time interval from different data sources with the help of an automated extraction process.
- Once the emission process extracts the data fields which are used by cQube, it is converted into csv formatted, pipe delimited files.
- The CSV data files will then be placed into the state data center by the automated process.
- Emission automated processes will invoke cQube data-ingestion APIs to emit the data.
- The API will have different end points as mentioned below:
    - Data emission users will request the cQube admin for the emission API token.
    - Data emission users incorporate the API token into the emission process code.
    - The Emission process makes an API call to generate AWS S3, one time presigned URL.
    - API calls to emit the data files using the https protocol into cQube.
    - API takes the data file as a parameter.
    - The API sends an acknowledgement on successful emission.

3.1.2 Certificate based authentication for emission & download API

The certificate based authentication is an additional security layer to authenticate the client based on the client certificate & client key along with the user credentials. Without the certificate & key user would not be able to emit the files to cQube or download the data from cQube. Below are the steps to add the certificate based authentication.

- Create a new subdomain for api and assign that api domain name to nginx server.
- Create a server config in nginx on nginx server.
- Using letsencrypt (or anything is fine), create an ssl certificate for that api domain and configure the ssl certificates to api domain.
- Create a self-signed client certificate and configure it.
- Copy the client certificate and key to the client machine (where emission happens).
- Restart the nginx on nginx server.
- Unassign the api domain name to nginx server.
- Comment out or delete the api selection in main nginx server configuration.

- Create a new load balancer with TCP listener.
- Assign api domain name to load balancer's A record.
- Add the load balancer's sg to nginx sg on port 443.

- The example spec for the emission API will be as like below

```
headers = {'Authorization': 'Bearer access_token', 'Content-Type': 'application/json'}

GET https://cqube.tibilprojects.com/data/list_s3_buckets
Body: {"input": "cqube-qa10-input", "output": "cqube-qa10-output", "emission":
"cqube-qa10-emission"}

POST https://cqube.tibilprojects.com/data/list_s3_files
Body: { "bucket": "cqube-qa10-emission"}

POST https://cqube.tibilprojects.com/data/download_uri
Body:
{"filename":"school_master/2020/2020-06/2020-06-04_school_master/04-06-2020_13:12:59.574_5e1
60862-c5b3-4121-9a96-ecefa34fc264_school_mst.zip","bucket":"cqube-gj-input"}
```

3.2 NIFI Processor Groups

- NIFI will have processor groups to combine similar NIFI processors, i.e the processes that are related to the same functionality into batches.
- Separate processor groups enable only the required data source transformation and this supports further scaling.
- The processor groups are loosely coupled, which enables them to make specific changes required to any particular processor group without affecting all the other processor groups.

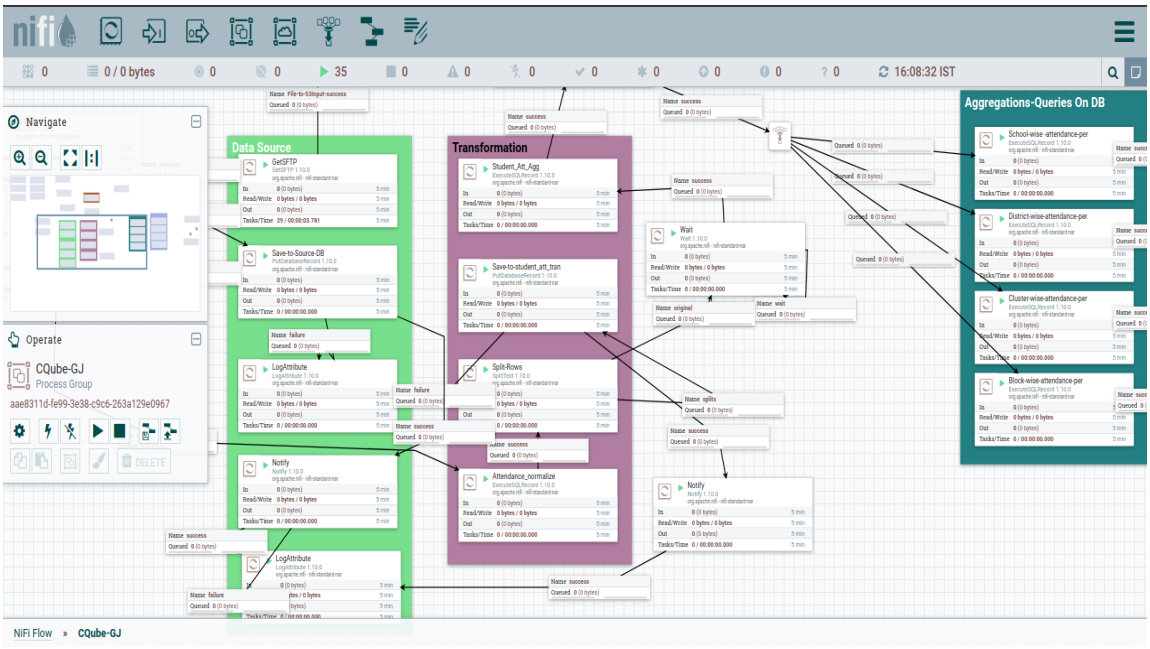
3.2.1 NIFI Data Process

- NIFI acts as the central orchestration system and all the aggregations are performed in the NIFI using PostgreSQL databases.
- NIFI fetches the aggregated data from the PostgreSQL database and sends them to the S3 output bucket.

3.2.2 NIFI data pipeline

cQube uses Apache NIFI to automate the data flow

Figure – 6: NIFI flow screenshot



- Apache NIFI Data Source Processor fetches the data from S3 emission data bucket.
- Stores raw data into S3 Input bucket.
- NIFI processes perform the data validations.
- Transformation Processor: Data is transformed using the queries in PostgreSQL and the transformed data is sent to the PostgreSQL.
- Aggregation Processors: Data aggregations are performed on the transformed data.
- NIFI stores the aggregated data and static data into S3 output buckets in JSON format.

3.2.3 NIFI Data Validations

- Data validations will take place at the following different levels:
    - The emitted data undergoes their first set of data validations before being copied into the S3 Input bucket
    - Second set of data validation takes place after the files are copied into the S3 Input bucket and before the Nifi process starts processing the data for cQube report creation.

NIFI fetches the data from the S3 emission data bucket and sends it to the S3 input bucket, after performing the following validations on the data:

- **Emitted data file size check:** NIFI gets the file size of the emitted data from the Manifest file and performs a check to see if the emitted file size matches with the original file size.
  NIFI does not allow the file into the S3 Input bucket if the file size does not match. A notification is sent through an email to the data emitter to check the file size and re-emit the data file.
- **Record count at the emitted file check:** NIFI gets the count of the number of records in the emitted data from the Manifest file. Nifi checks if the emitted file records count matches with the original file records count.
  If the file records do not match, the NIFI will not process the file into the S3 Input bucket. A notification email is sent to the data emitter to re-emit the data file.

NIFI fetches the data from the S3 emission data bucket and sends it to the S3 input bucket, after performing the following validations:

- **Column level validations**: Column datatype mismatch, Number of columns, Data exceeding the column size.
- **Improper Data handling**: Missing/ null data values for mandatory fields, Empty data files, Special characters, Blank lines in data files.
- **Duplicate records validation:** NIFI validates the duplicate records by grouping the same kind of records together.

For the record which is having duplicate values for all fields (mirror image record) NIFI will consider the first record and the rest of the records will be eliminated.

For the rest of the duplicate records where the records are having the same ID (student ID/ assessment ID/ infra ID/ CRC visit ID) and different values will not be inserted into the database tables as ID is the primary key.

For the Duplicate records with different lat long details, NIFI eliminates the records which have the same id and different lat long details and the records which have different ids and the same lat long details.

For semester report, The records which are having the same values for fields Student ID, School ID, semester, studying class and different values for the subjects then NIFI will eliminate those records.

- **Overlapping data validation**: Overlapping data validation takes place based on the data source.

  The NIFI process for student attendance reports will check the last updated day's record from the transactional table and will process the records from the day after the last updated date. The records from all of the previous days will not be considered for NIFI processing.

  For the other data sources, duplicate records where the records are having the same ID (student ID/ assessment ID/ infra ID/ CRC visit ID) and different values will not be inserted into the database tables as ID is the primary key.

- **Other data issues**: Data handling in cases like job failures, missing data for certain days and late receipt of the data (receiving data after a few days), updating the wrong data, upon request (when issue identified at the report)

3.3 Prometheus and Grafana monitoring tool connectivity

Prometheus:

- Configure a job to scrape for the server's RAM, CPU and Storage status on port 9100

- Configure a job to scrape for the Nifi's process and Memory status on port 9092

- Configure a job to scrape prometheus itself health check on port 9090

Node exporter:

- Configure node_exported to send the metrics about the node (cQube server)

Grafana:

- Configure prometheus as new Datasource

- Import the dashboard given in cQube git repository located on cQube/ development/ grafana/ cQube_Monitoring_Dashboard.json

3.4 cQube Data Model

This section describes the student attendance, student assessment, infrastructure data and the teacher attendance data structure.

Tables classification

All the entities are classified into four types

- Static tables
- Metadata tables
- Hierarchical tables
- Dynamic tables
- Aggregated tables

Static tables: The tables which contain the static information of the entities like geo master, school master, student master, subject master

Metadata tables: The tables which are used to store the processing information of the emission process like incoming files and process status. To store the information of the days when the student attendance data is processed.

Hierarchical tables: These tables contain the rarely updated values or dimensions that change rarely such as student class, academic year.

Dynamic tables: These tables contain the daily or frequently changing values such as the student attendance, Teacher attendance, CRC data, Semester assessment.. These tables are updated more frequently.

Aggregated tables: These tables will hold the school-wise aggregated values which will be ready to be converted into JSON files. Multiple aggregation tables will be created based on the visualization report requirements.

Initialization stage for Infrastructure dataset: After the configuration stage is completed, the infrastructure master tables are initialized with the infrastructure columns of that state, based on the columns present in the infrastructure transaction table columns. Once initialization is completed the infrastructure weights can be configured through the emission API.

Click on the link below for the data dictionary:

https://github.com/project-sunbird/cQube/blob/release-1.9/documents/cQube_data_dict.pdf

Click on the link below for the latest ERD:

https://github.com/project-sunbird/cQube/tree/release-1.9/documents/ERD

3.5 Node connectivity with S3 bucket

1. Connectivity of aws s3 to nodejs is being done by an package called "aws-sdk"

2. Stored all the aws access related keys in the .env file

3. Maintained the config file to store all the s3 connect details (Ex: accessKeyId, secretAccessKey, bucketName)

4. Common function to read the files from s3 by passing the filename as a parameter to the s3 file reading code

Ex: s3Key parameter variable -> stores the filename of the every api request and pass to the function

```
const readS3File = (s3Key) => {
    return new Promise((resolve, reject) => {
        try {
            const_data['getParams']['Key'] = s3Key;
            const_data['s3'].getObject(const_data['getParams'], function (err, data) {
                if (err) {
                    logger.error(err);
                    reject({ errMsg: "Something went wrong" });
                } else if (!data) {
                    logger.error("No data found in s3 file");
                    reject({ errMsg: "No such data found" });
                } else {
                    var jsonData = JSON.parse(data.Body.toString());
                    resolve(jsonData)
                }
```

```
            });
        } catch (e) {
            reject(e)
        }
    })
}
```

4. Once the response received from the file, the required logics are done on the data set received from the s3 file

5. Logging the application information and error using the library called winston

6. All the data sent back as a json format to the Angular through the different apis.

3.6 AngularJS, chart js and leaflet roles in the visualization

1. Angular - is used as a web application of the project where all types of visualization reside on.

2. All the headers and token for backend api are passed using HTTP Interceptors

3. The API calling functions are written in the different service files according to the reports.

4. Integrated the keycloak for authentication with the application

5. All the environment variables (backend api url, keycloak url, keycloak clientId, realm name) are stored respectively in environment.prod.ts file

6. Leaflet - The open source library has been used to show the visualization of maps related reports.

7. Chartjs - Has been used to show the visualization of Line charts, bar charts, scatter plots for the reports developed

8. Some of the reports are showcased using the bootstrap datatables in a tabular format.

3.7 Logs

Following list log files are linked in <base_dir>/cqube/logs directory

Nifi logs:

- <base_dir>/cqube/nifi/nifi/logs/nifi-app.log as nifi-app.log

- <base_dir>cqube/nifi/nifi/logs/nifi-bootstrap.log as nifi-bootstrap.log

Postgres logs:

- /var/log/postgresql/postgresql-10-main.log as postgresql-10-main.log

Emission app logs:

- <base_dir>/cqube/emission_app/python/access.log as emission_app-access.log

- <base_dir>/cqube/emission_app/python/error.log as emission_app-error.log

UI & Admin UI logs:

- /home/<system_user_name>/.pm2/logs/client-side-error.log as client_side-error.log

- /home/<system_user_name>/.pm2/logs/client-side-out.log as client_side-out.log

- /home/<system_user_name>/.pm2/logs/server-side-error.log as **server_side-error.log**

- /home/<system_user_name>/.pm2/logs/server-side-out.log as server_side-out.log

- /home/<system_user_name>/.pm2/logs/admin-client-side-error.log as admin_client_side-error.log

- /home/<system_user_name>/.pm2/logs/admin-client-side-out.log as admin_client_side-out.log

- /home/<system_user_name>/.pm2/logs/admin-server-side-error.log as admin_server_side-error.log

- /home/<system_user_name>/.pm2/logs/admin-server-side-out.log as admin_server_side-out.log

System logs:

- /var/log/syslog as syslog

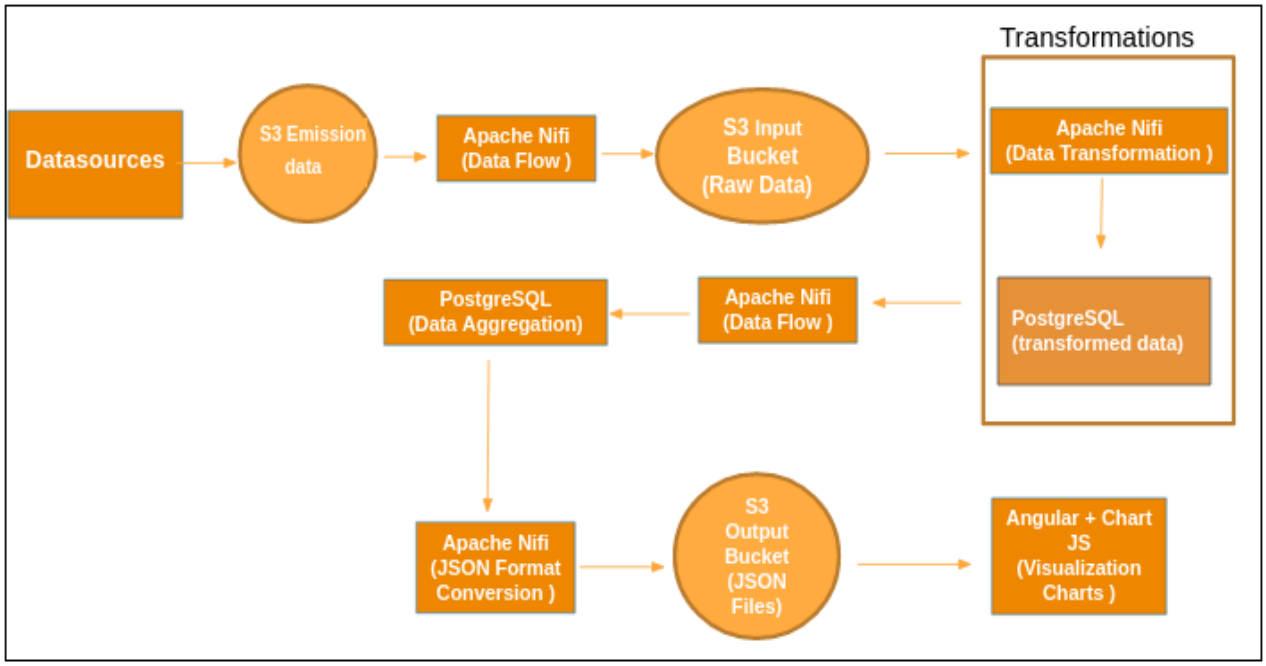
Keycloak logs:

- <base_dir>/cqube/keycloak/standalone/log/server.log as server.log

- <base_dir>/cqube/keycloak/standalone/log/audit.log as audit.log

4. cQube data flow

Figure – 4: cQube data process flow



Description of the data flow:

Table – 2: cQube data flow tables

S3 emission data	Is the location where the emitted data resides. This folder is created in AWS S3 Raw data files are emitted to S3 emission data buckets periodically.
S3 emission data - NIFI	NIFI collects the raw data files from the S3 emission data location and after that removes them from the S3 emission data folders.
NIFI	NIFI sends the raw data to the S3 (input) bucket. NIFI does the validation of data fields and values. NIFI inserts the transformed data to PostgreSQL. NIFI specifies the scripts to PostgreSQL to perform the necessary aggregations for visualization reports. NIFI picks the aggregated data from PostgreSQL and places them into the S3 bucket (output) to create JSON files which can be used for the visualization reports
PostgreSQL	PostgreSQL base tables are populated by NIFI PostgreSQL transformed tables are populated by NIFI PostgreSQL aggregate tables are populated by NIFI
S3 Output Bucket	S3 Output bucket contains the files in JSON format Angular reads the data from S3 output bucket by using Node API
Angular, ChartJS/ Leaflet	Angular fetches the data using Node JS from S3 to create reports using ChartJS/ Leaflet

5. Security Implementations

- **EC2**: A pair of public and private keys are generated, and the public key is stored in the EC2 server. The client with the private key gets authenticated with the server during login only if the keys match.
- **S3 emission data bucket**: cQube provides a data-ingestion API to emit the data. The API will have different secured endpoints like:
    1. User authentication
    2. API call to emit the data files using the https protocol into cQube.
    3. API takes the data file as a parameter.
    4. The API will provide acknowledgement on successful emission.
- **S3 Input & Output buckets**: A pair of access keys and secret keys are generated to secure the S3 location.
- **NIFI**: Only authorized users can gain access to the NIFI Dashboard. The confidential keys such as username and password will be encrypted.

- **PostgreSQL**: PostgreSQL will be secured as follows:
    1. User Access Control (RDAC)
    2. Server configuration
    3. User and role management
    4. Logging
    5. PostgreSQL audit extension (pgAudit)
    6. Security patches
- **Angular**: Role based authentication will be provided to prevent access to unauthorized users. A reverse proxy server has been used that usually stays behind the firewall of a private network. Reverse proxies are also used as a means of caching common content and compressing inbound and outbound data, resulting in a faster and smoother flow of traffic between the clients and servers.
- Role based authentication: Will be provided to admin users, and based on the user roles, relevant access will be provided to users at district levels, block levels, cluster levels and school levels.
- Normal/ regular users can access public reports without any authentication.

6 S3 bucket partitioning

S3 buckets will contain partitions for the data files to store. The partitions are created at the S3 input bucket & the S3 Output bucket.

6.1 S3 emission bucket partitions

- All the files in the S3 emission bucket will be in the CSV format.
- S3 emission bucket follows the folder hierarchy based on the data sources.
  S3 -> Bucket name -> Data source -> emitted zip files with timestamp
- The emitted zip file contains the CSV data files with timestamp and a manifest file with a timestamp.
- The folders and the files will be removed from the S3 emission bucket once NIFI copies the data.
- Unprocessed data files will remain in the S3 emission bucket for one week and then they will be deleted automatically at the end of the week.

6.2 S3 input bucket partitions

- All the files in the S3 input bucket will be in the CSV format.
- S3 input bucket follows a hierarchical partitioning based on the Data source, Year, Month, date and timestamp
  S3 -> Bucket name -> Data source -> Year -> Year - Month -> date_Source name

  Example for the S3 input bucket:

  ```
  S3/cqube-gj-input/student_attendance/2020/2020-05/2020-05-29_student_attendance
  ```

6.3 S3 output bucket partitions

- All the files in the S3 Output bucket will be in the JSON format.
- S3 output bucket follows the hierarchical partitioning based on the data source, Year, Month, date and timestamp, similar to the partitioning that the S3 input bucket follows.
- Metadata files will have information of the latest updated output files which helps cQube to consider the latest output file during the visualization stage.

7. cQube users - Technical activities

The cQube product can be used by a variety of users and each user will have different functions that he/she can perform and different user privileges The cQUbe users can be divided into the following categories:

1. Admin
2. Ad-hoc Analyst

The various roles and the functions that each role can perform have been described in the table below.

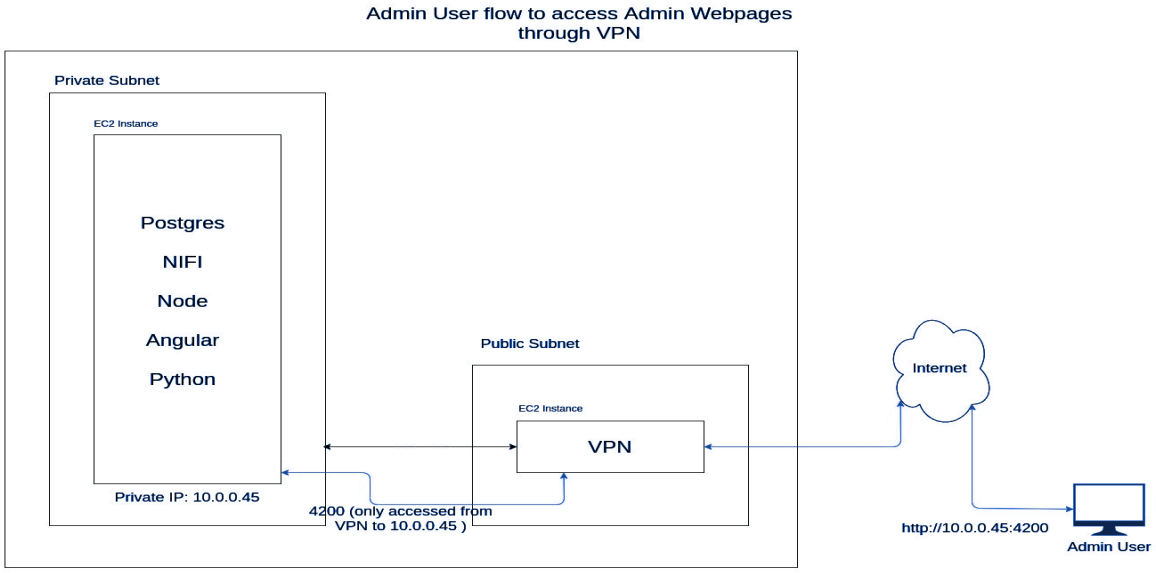
7.1 Admin

cQube has two different interfaces, an interface for the cQube administrator and another interface/ dashboard for the cQube reports:

- Administrator pages: Administrator activities will run separately in the VPN and admin must login through the 2 factor authentication process to perform the admin tasks.Change password functionality will be included in this login.
- cQube dashboard for report: This is a normal login which will have the cQube insights of the metrics.

7.1.1. Admin login process:

Figure - 7: Admin user flow through VPN



cQube

Admin must follow the 2 factor authentication process to perform the admin tasks. The Admin user should connect to the VPN to avail the 2 factor authentication by following the steps mentioned below:

- Admin has to request the devops team for OpenVPN access details.
- Admin has to download the google authenticator app to his phone and register the app with the QR code showing in the OPENVPN Access server page.
- By providing the credentials & Google authenticator code to download the user-locked profile(client.ovpn) file.
- Admin has to install the client openvpn-connect application
- Admin has to create the OpenVPN profile from the client.ovpn file
- Admin has to validate the user authentication and Google authenticator code to login to the cQube VPN.
- With the successful authentication admin can access the cQube admin features by opening the local ip in the browser.

7.2 Ad-hoc analyst

- They are the dynamic report creators of cQube.
- The ad hoc analyst can make use of third-party visualization tools like Metabase, Tableau or any other visualization tool to create dynamic dashboards and develop dynamic reports by directly accessing the database.
- Ad-hoc users can download the JSON files from the S3 output bucket through the API.
- The example spec for the download API will be as like below

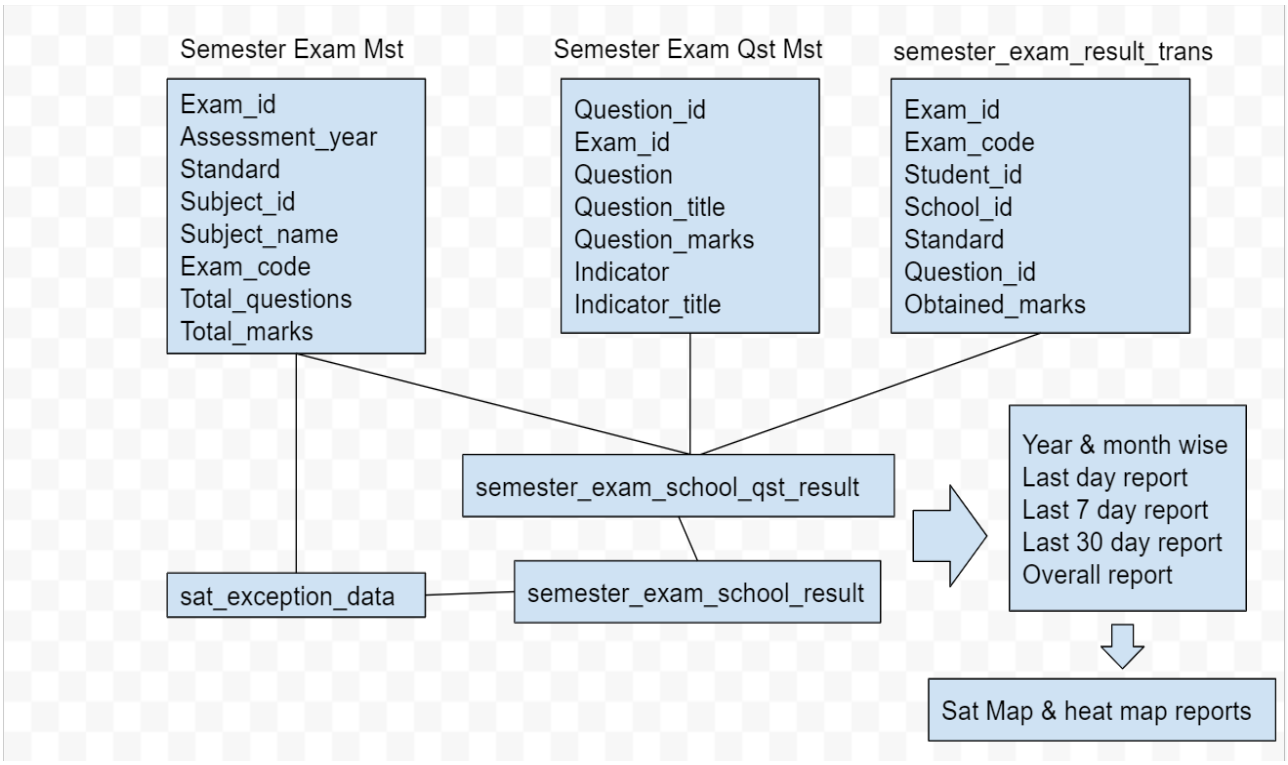
```
headers = {'Authorization': 'Bearer access_token', 'Content-Type': 'application/json'}

GET https://cqube.tibilprojects.com/data/list_s3_buckets
{"input": "cqube-qa10-input", "output": "cqube-qa10-output", "emission": "cqube-qa10-emission"}

POST https://cqube.tibilprojects.com/data/list_s3_files
Body: { "bucket": "cqube-qa10-emission"}

POST https://cqube.tibilprojects.com/data/download_uri
Body:
{"filename":"school_master/2020/2020-06/2020-06-04_school_master/04-06-2020_13:12:59.574_5e1
60862-c5b3-4121-9a96-ecefa34fc264_school_mst.zip","bucket":"cqube-gj-input"}
```

8. Semester Assessment Test Configuration



- Data is aggregated based on the above tables at question level for each grade & subject in a school for every year and the results are stored into semester_exam_school_qst_result.
- Aggregation of data at subject & grade using the semester_exam_school_qst_result is performed and stored into semester_exam_school_result.
- From the aggregated data the last day, 7 days, 30 days, overall reports are queried and stored to the files.
- The exception report is generated based on the latest data available in aggregation tables and stored to the s3 files.

9. cQube data replay process

The data replay process takes place based on the data source. Mentioned below are the steps that are involved in the data replay process:

- Admin will be provided with a screen to select the options to clear the data for each of the data sources. The admin screen will contain the following selection options:

1. For student attendance, Teacher attendance the admin will be able to select the ‘year and month’ using the year & month drop down.

2. For CRC and Diksha summary rollup the admin will have the calendar selection. The data will be deleted for the selected dates.

3. For the Semester reports, admin will be able to select the required semester from the available semesters. The selected semester data will be deleted.

4. For Periodic Assessment Test, admin selects the exam code option from the multiple select box which is having all the available exam codes. The complete data which is related to the selected exam code will be deleted.

5. For Diksha TPD, Admin selects the Batch ID option from the select box which is having all the available Batch IDs. The complete data which is related to the selected Batch ID will be deleted.

6. For UDISE & Infrastructure data sources, admin can delete overall data with the selection of ‘Yes or No’ option from the select box. Full refresh will happen with the new data.

7. For the static data sources, admin can delete overall data with the selection of ‘Yes or No’ option from the select box. Full refresh will happen with the new data.

- A submit and Reset all buttons will be given in the admin screen to Submit the request and reset the options.
- When admin clicks on submit button, All the data sources will be created as JSON file as shown below


```
{"student_attendance":  {
    "year":"2020",
    "months":["01", "03"]
 },
 "teacher_attendance":  {
     "year":"2020",
    "months":["01", "03"]
 },
 "crc":  {
     "year":"2020",
    "months":["01", "03"]
 },
```

```
 "diksha_summary_rollup":  {
  "from_date":"",
  "to_date":""
 },
  "semester":  {
   "semester":[1,2]
 },
 "periodic_assessment_test":  {
   "exam_code":["PAT010101012021", "PAT010201012021"]
 },
 "diksha_tpd":  {
   "batch_id":["03052315462389", "046789546783"]
 },
  "udise":  {
   "selection":"yes/no"
 },
  "Infrastructure":  {
   "selection":"yes/no"
 },
  "static":  {
   "selection":"yes/no"
 }
}
```


- The JSON file containing the values selected by the admin will be placed in the S3 emission bucket.
- A scheduler will be created for the data replay process for all reports. And the scheduler will run based on the schedule defined by the admin.
- The scheduler will initiate the NIFI to get the file from S3 input bucket. NIFI performs the data deletion operation based on the inputs given by the admin (for all the data sources).

9.1 Data deletion process

Once the file is emitted to the S3 bucket, NIFI function will be invoked at the scheduled time and get the input parameters from the JSON file. The queries will be executed and delete the data from transactional tables. The output files will be updated according to the deleted data.

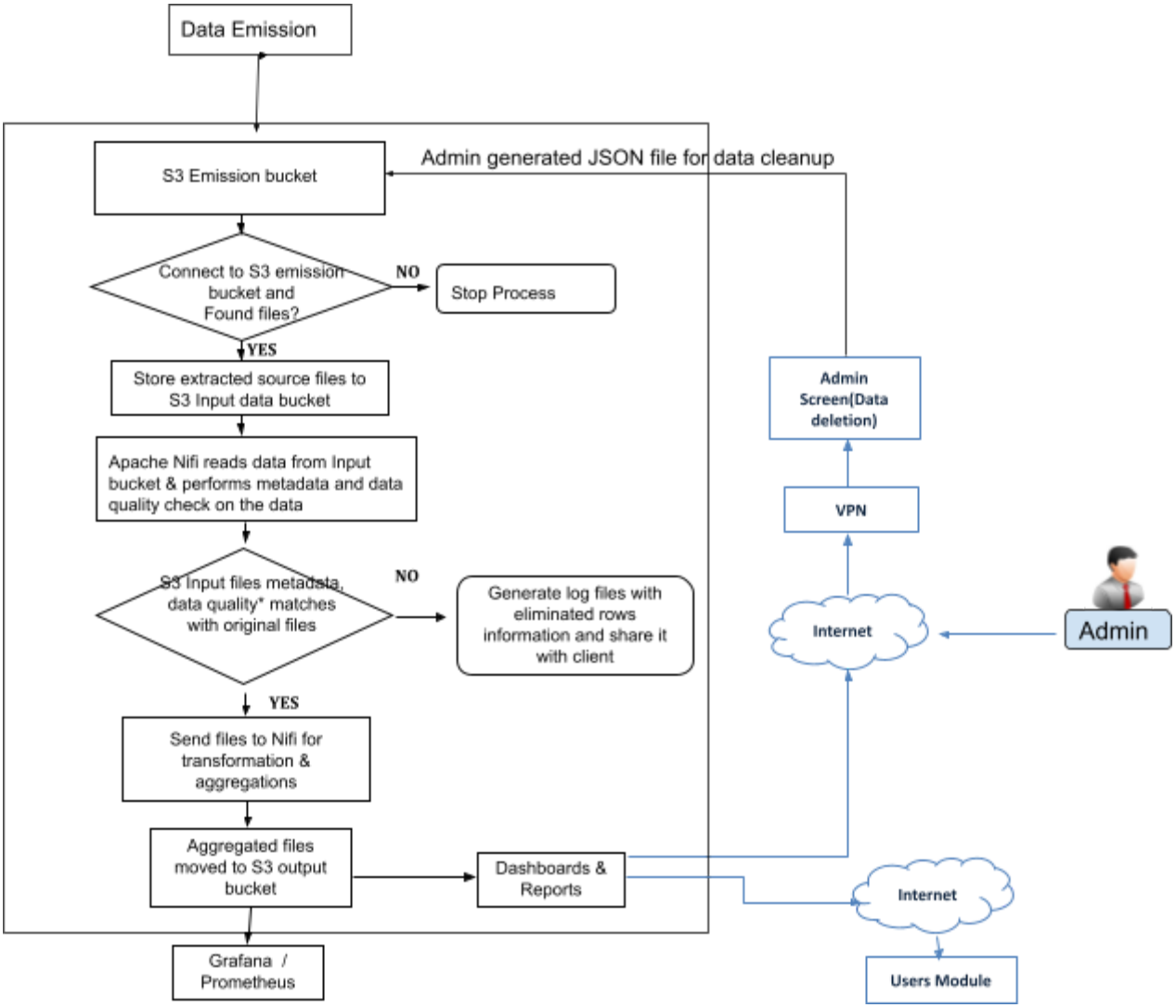
9.2 Data reprocessing (for previously deleted data) flow

Data reprocessing will take place in the normal cQube emission process.

- The latest data file will be emitted to S3 emission bucket

- The file will be processed as the regular data process from NIFI
- All the validations will be performed by NIFI and the validated data will be inserted into the transactional tables.
- All the metrics will be re-calculated and updates of the output files.
- The new metrics will be affected in the reports.

The complete workflow process will be like below.



List of tables cleared for the data source

datasource	parameter	list of tables	function call
student_attendance	month,year	student_attendance_meta,student_attendance_staging_1,student_attendance_staging_1,student	select del_data(p_data_source=>'student_attendance',p_year=>2022,VARIADIC

		nt_attendance_temp,student_attendance_trans,school_student_total_attendance	p_month=>array[1,2]);
teacher_attendance	month,year	teacher_attendance_meta,teacher_attendance_staging_1,teacher_attendance_staging_1,teacher_attendance_temp,teacher_attendance_trans,school_teacher_total_attendance	select del_data(p_data_source=>'teacher_attendance',p_year=>2022,VARIADIC p_month=>array[1,2]);
crc	month,year	crc_location_trans,crc_inspection_trans,crc_visits_frequency	select del_data(p_data_source=>'crc',p_year=>2022,VARIADIC p_month=>array[1,2]);
semester_assessment_test	exam_code/semester	semester_exam_mst,semester_exam_result_staging_2,semester_exam_school_qst_result,semester_exam_result_temp,semester_exam_school_result,semester_exam_qst_mst,semester_exam_result_staging_1,semester_exam_result_trans	
periodic_assessment_test	exam_code	periodic_exam_mst,periodic_exam_result_staging_2,periodic_exam_school_qst_result,periodic_exam_result_temp,periodic_exam_school_result,periodic_exam_qst_mst,periodic_exam_result_staging_1,periodic_exam_result_trans	select pat_del_data(p_data_source=>'periodic_assessment_test',VARIADIC p_exam_code =>array['PAT0302290720201','PAT0302290720202']);
diksha_tpd	batch_id	diksha_tpd_agg,diksha_tpd_trans,diksha_tpd_content_temp,diksha_tpd_staging	select diksha_tpd_del_data(p_data_source=>'diksha_tpd',VARIADIC p_batch_id => array['0302290720201','0302290720202']);
diksha_summary_rollup	from_date,to_date	diksha_content_staging,diksha_content_temp,diksha_content_trans,diksha_total_content	select diksha_summary_rollup_del_data('diksha_summary_rollup','2022-12-27','2022-12-31');
infrastructure	all	infrastructure_temp,infrastructure_trans	select all_del_data('infrastructure');
static	all	block_tmp,block_mst,district_tmp,district_mst,cluster_tmp,cluster_mst,school_master,school_tmp,school_hierarchy_details,scho	select all_del_data('static');

		ol_geo_master	
udise	all	udise_sch_incen_cwsn,udise_nsqf_plcmnt_c12,udise_sch_enr_reptr,udise_nsqf_basic_info,udise_sch_incentives,udise_nsqf_trng_prov, udise_sch_exmmarks_c10,udise_nsqf_class_cond,udise_school_metrics_trans,udise_sch_exmmarks_c12, udise_sch_pgi_details,udise_nsqf_enr_caste, udise_sch_enr_age,udise_sch_exmres_c10, udise_sch_profile,udise_nsqf_enr_sub_sec, udise_sch_enr_by_stream,udise_sch_exmres_c12, udise_sch_recp_exp,udise_nsqf_exmres_c10,udise_sch_enr_cwsn,udise_sch_exmres_c5, udise_sch_safety,udise_nsqf_exmres_c12,udise_sch_enr_fresh, udise_sch_exmres_c8, udise_sch_staff_posn,udise_nsqf_faculty, udise_sch_enr_medinstr,udise_sch_facility, udise_tch_profile,udise_nsqf_plcmnt_c10,udise_sch_enr_newadm	select all_del_data('udise');

10. New use-case creation

10.1 New Use-Case creation

New Dataset Setup: To create the new data sets please follow the below steps

1. Select the required columns in the emission dataset based on the metrics required to visualize in the dashboard.
2. Design the data model for the new dataset.
3. The data model should have relation with the static tables, and also it should have transaction, aggregation tables.

4. Create the table definitions & store them in the SQL file. Example for reference. https://github.com/project-sunbird/cQube_Workflow/tree/cQube-release-qa/development/postgres/table_definitions
5. Write the queries to perform the below data validations

- Null Validation
- Mirror Validation
- Same id Validation

6. Create the count tables if required to store the transaction counts, so that the old data in the transaction tables will have no dependencies for calculating metrics and they can be cleared to keep the db light.
7. Write the upsert queries with the metric calculation logic to process the incremental data load to transaction/count & aggregation tables.
8. Create the materialized views to write the s3 output files based on requirement for last 1 day/last 7 days/last 30 days/overall/year & month data from aggregation tables, which will be visualized in the dashboard. Store the functions/views/Materialized view queries configuration files. Example for reference https://github.com/project-sunbird/cQube_Workflow/tree/cQube-release-qa/development/postgres/datasource_config
9. Write the queries for s3 output meta files which store data such as the year & month details.

NIFI Processor Creation: To create the new NIFI processor group please follow the below steps.

- Create the list of processors based on the list of requirements. The NIFI processor creation would be the drag and drop of NIFI UI.
- Create a processor group by including all the newly created processors.
- Create the link processors in the cQube_data_storage processor group for the newly created processor group.
- Link the newly created processor group to cQube_data_storage processor group

Note: The new NIFI processor group should be created based on the data sources mentioned in the Ansible code

Visualization creation:

- New dash page has to be created for the new use case
- Icons should be added if required in the below folder https://github.com/project-sunbird/cQube_Workflow/tree/cQube-test/development/angular/client-side/src/assets
- New components can be added for the side bars at the below location https://github.com/project-sunbird/cQube_Workflow/tree/cQube-test/development/angular/client-side/src/app/containers

The command to create the New component is as
> Open the terminal
> cd “Location of containers folder path”
> ng generate component “Component_Name”

- Write a switch statement for the new use-case in the below file https://github.com/project-sunbird/cQube_Workflow/blob/cQube-test/development/angular/client-side/src/app/app-routing.module.ts

10.2 New use-case integration with cQube

Data set Configurations:

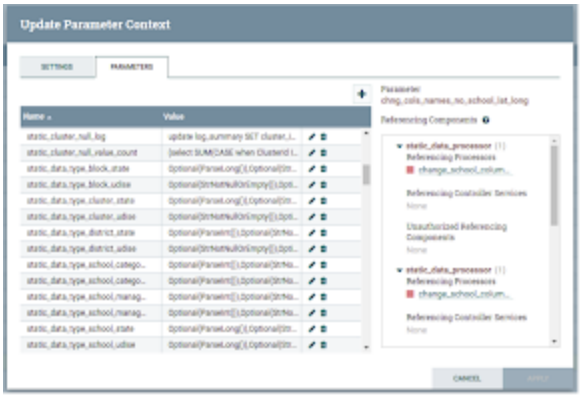
- Integrate the data sources with the data reply to clear the data based on the parameters.
- Integrate the data sources with the data retention to clean up the old data from the transaction table.

NIFI configuration changes:

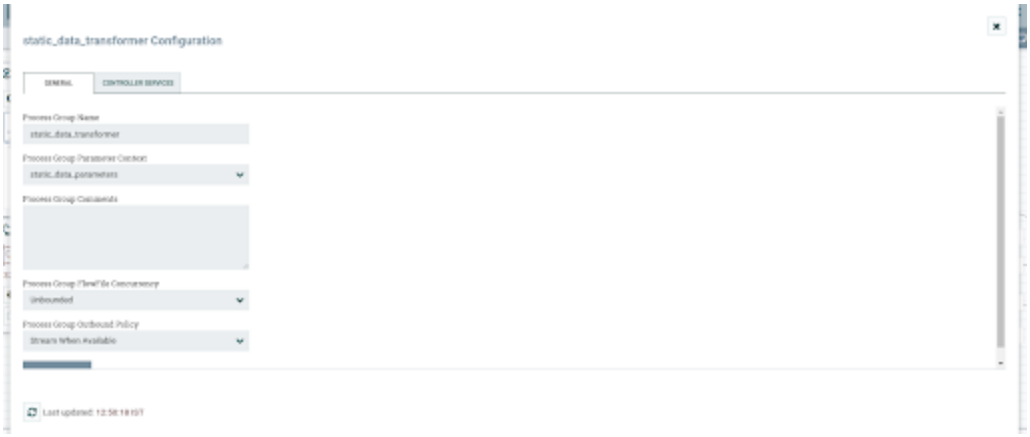
- Create the set of parameters based on the requirement
- Link the parameters with the processor groups
- For S3/Local connectivity no changes required in the new processor groups as they already taken care in the cQube_data_storage processor group

For reference please see the below Images

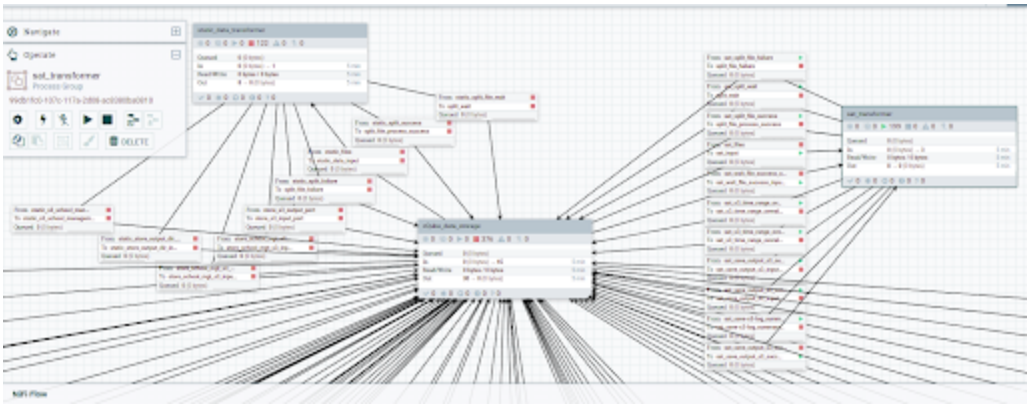
For the set of parameters



To Link the parameters with the processor groups



The below image represent the link with cQube_data_storage processor group



Visualization code changes:

-

11. NIFI configurations for S3/In-house

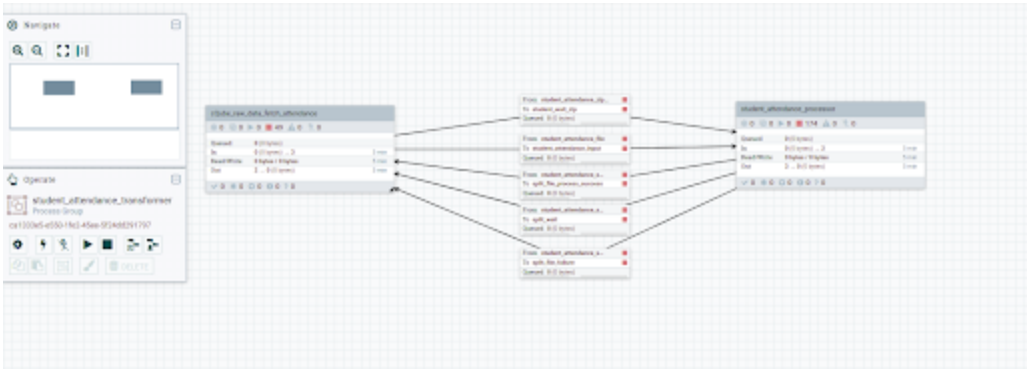
For the connectivity of the S3 and Inhouse, cQube is having a new processor group which was created in cQube release 3.0 that is cQube_data_storage processor group.

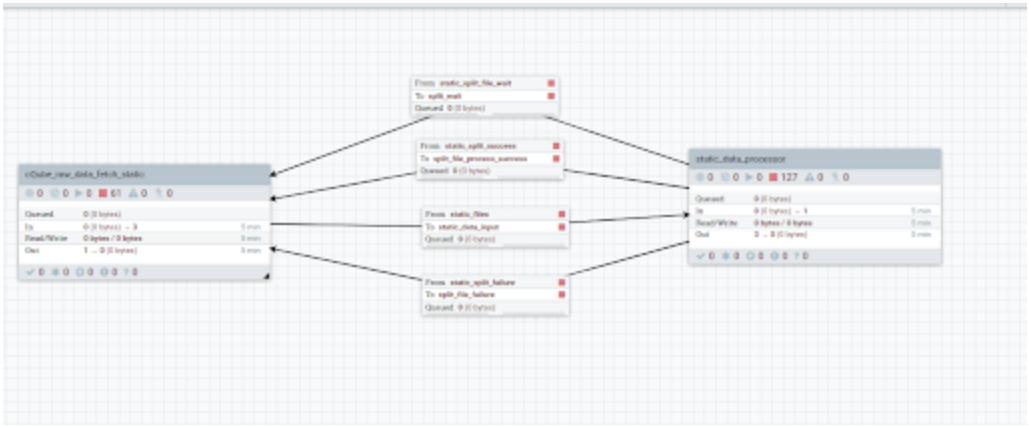
The below changes have been made in the existing processor groups as mentioned in the below points.

- Create a new processor group for which cQube_data_storage processor group which will handle both S3 and Local data storage.
- For S3 and In-house data center, A separated new processors were added Inside the cQube_data_storage processor group
- Transformations, validations were grouped as a separate processor groups which are together in one previously
- Removed the S3 connections from the all processor groups and now linked the connectivity with the single point of processor that is cQube_data_storage processor group

For example, Please refer the below image for more information

The Old Student attendance & Static processor groups were like the below





After the changes made the above processor groups are like below.

(i) cQube common Issues and Resolutions

Please refer to the below document for understanding the most common issues and its resolutions were added in the below documents.

https://github.com/project-sunbird/cQube_Workflow/blob/release-3.0/documents/Solution%20Doc%20-%20Error%20Log.pdf

(ii) List of NIFI processor groups & UI Components

Please refer to the document below for understanding the NIFI processor groups and the UI Reports and its descriptions.

https://github.com/project-sunbird/cQube_Workflow/blob/release-3.0/documents/Nifi%20Processor%20Group%20%26%20UI%20Repost%20List.pdf