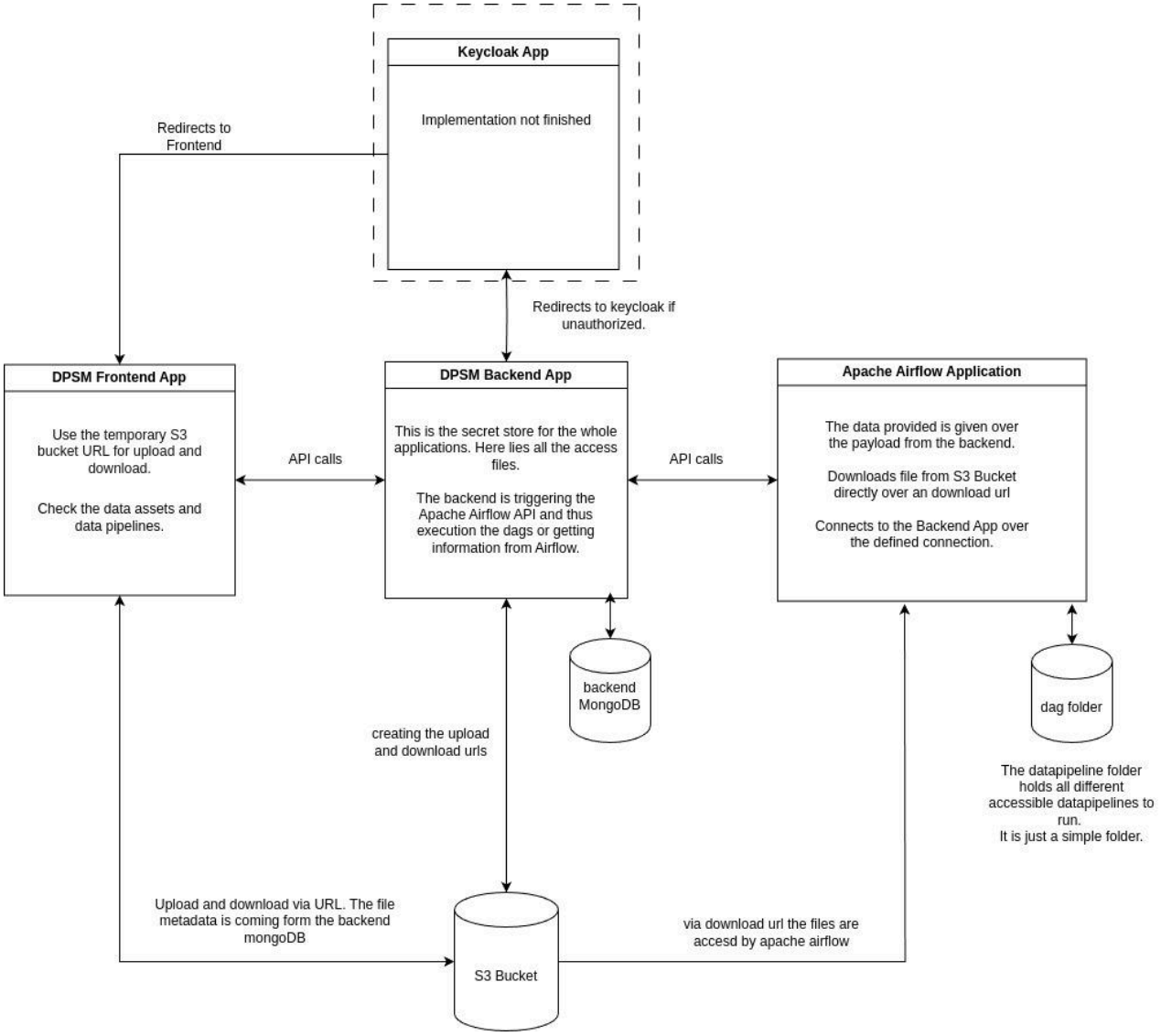
DESIGN DOCUMENTATION

1. Architecture



Following are setup as docker containers:

- DPSM Frontend App
- DPSM Backend App
- backend MongoDB
- Apache Airflow Application
- Keycloak App

The S3 Bucket is an external service and needs to be set up before.

Architecture Deployment Notes:

As Apache Airflow takes a lot of resources, it may make sense to split frontend, backend and the Airflow applications into their own EC2 instances.
But Keycloak **needs HTTPS** which means that you need a reverse proxy (nginx) for every instance to talk within and an SSL encryption. (We used certbot with nginx).
It also makes sense to deploy EC2 instances with elastic ips (which are static and don't change after restart of the instance)

2.Code Considerations

Frontend Considerations:

- While UI is not the primary focus, a filter search bar is implemented for efficient data retrieval and exploration.
- The search functionality extends to the frontend, allowing users to interact with and visualize the processed data.
- Users can upload the files to S3 bucket and download the files from S3 bucket directly.

Access Management (postponed):

- Authentication is handled through Keycloak, providing a more extensive login mechanism

- Differentiated access levels are not implemented; once a user is registered, they gain access to the entire database, ensuring a streamlined and uniform access control system.

Backend Functionality:

- **Database Interaction:**
  - The backend interacts with the database, facilitating data storage and Retrieval.
  - **Result Storage:**
    - Processed results are stored in the database in a key-value pair format, ensuring efficient and structured data storage.

- **S3 Bucket Functionality:**
  - **File Storage:**

    - The backend is responsible for storing files in the S3 bucket.
    - This includes uploading and managing various types of files within the S3 storage.
- **File Retrieval:**
    - Downloading files from the S3 bucket is also a backend-managed process.
    - Users can request files, and the backend facilitates the retrieval and delivery of the requested files.
- **AWS Credentials:**
    - Secure AWS credentials are utilized by the backend to access the S3 bucket.
- **Integration with Apache Airflow:**
    - Apache Airflow is integrated into the system to manage and execute data processing tasks.
    - It receives incoming data, performs operations like word count or key-value operations, and communicate the results back to the backend.

3.Technology Stack

- Frontend
    Typescript
    Angular
    Bootstrap
    Angular-material
- Backend
    Python
    Flask
- Database
    MongoDB
    S3 Bucket
- Data pipeline
    Apache Airflow
- Deployment
    EC2 instance