Workshop No. 2

Data System Architecture and Information Retrieval

Kevin Santiago Avella Torres – 2021202096

Juan Esteban Oviedo Sandoval – 20192020064

Cristian Camilo Tuso Mozo – 20201020053

Databases II

Computer Engineering Program

Universidad Distrital Francisco José de Caldas

Introduction

The exponential growth of digital information has significantly increased the need for reliable and scalable cloud storage solutions. Users—ranging from students to professionals and small organizations require secure, accessible, and user-friendly platforms to store and share files efficiently. Traditional storage services can often present limitations such as restricted customization, complex integration with other systems, or high operational costs for small-scale users.

To address these challenges, this project proposes the design of a cloud-based file storage platform that enables users to upload, organize, and manage digital resources through an intuitive web interface. The system integrates authentication, access control, and metadata management to ensure information security and efficient retrieval. Moreover, it leverages cloud technologies such as object storage services and RESTful APIs to provide scalability and reliability.

This document focuses on the refinement and technical expansion of the design elements, together, these components establish the foundation for the database and system architecture design.

1 Business Model Canvas

1.1 Desing Business Model

The following model provides a structured framework to define the key aspects of the proposed file storage platform. Identifies the value proposition, customer segments, partners, resources, and revenue streams, providing a clear overview of how the system creates and delivers value to its users.

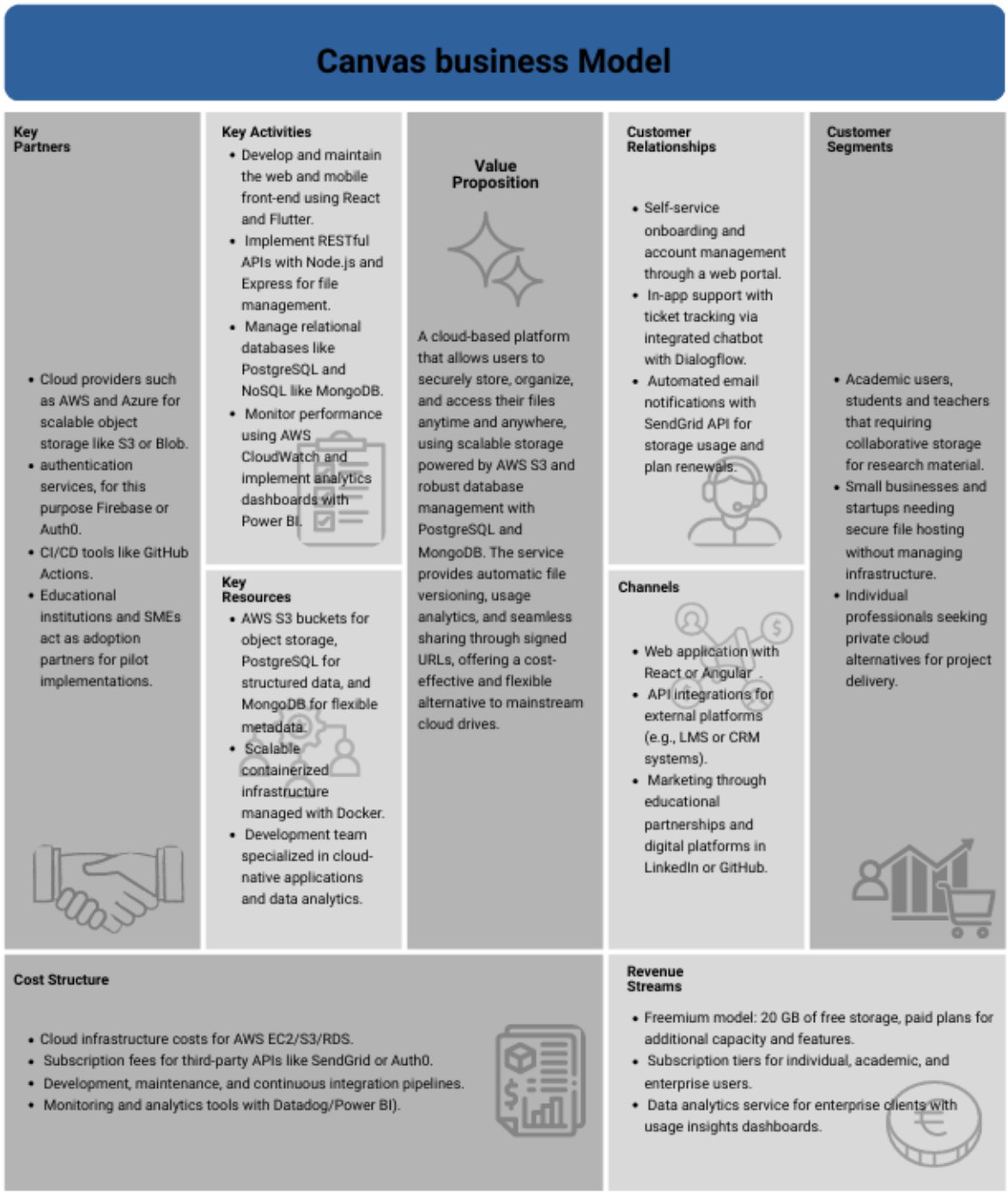

Figure 1: Business Model Canvas for the File Storage Platform.

2 Requirements Documentation

2.1 Functional Requirements

- FR1. Allow users to register with their details and relevant information (email address, personal details, and other details to be agreed upon).
- FR2. Allow already registered users to log in, taking into account their log-in credentials (email and password).
- FR3. Allow users to manage their storage space using folders, which must have specific options (view, organize, delete, move). Similarly, it must be possible to navigate between folders by hierarchical level according to the user's organization, listing the file elements of various permitted extensions, described specifically in 'RF4', which are classified in each of the folder layers.
- FR4. FR4. Users must be able to upload files to their workspace, organizing and separating items as they wish into folders. Possible events during file upload must be guaranteed, indicating whether the files have been uploaded correctly or whether errors have occurred during upload. Valid file types for user upload are (.docx, .pdf, .xls, .mp4, .mp3, .png, .jpg, .gif, .pptx) and the maximum upload size per file is 20 MB.
- FR5. Users must be able to view the current capacity of their storage space, detailing how much space is currently occupied and how much space remains available for use. It should be noted that this storage space depends directly on the user's plan.
- FR6. Allow the user to view important data about the files they have in their storage space, including file name, size, root or file type, and upload date. In addition, users should also be allowed to download files available in their storage space. However, Allow users to manage their files (delete, move, hide, and organize).
- FR7. Users shall have access only to files belonging to them and shall not have access to or be able to view items from other existing accounts.
- FR8. Users must have the ability to securely recover passwords (email with temporary token). This allows users to reset their passwords securely and reliably.
- FR9. The system must allow two plan options: The basic plan, which will be free for all accounts created and is acquired by default upon registration. This free plan will have limited trial storage space. The second plan will be the premium plan, which differs in that it increases storage capacity. Users can access the free plan by making a set payment to purchase premium membership, which will have a predetermined value and be valid for six months only. Users can only have one active premium membership. The system will also validate membership status and purchased storage capacity, ensuring that the limits set in either plan are not exceeded.

2.2 Non-Functional Requirements

Non-functional requirements establish the quality attributes and operational constraints of the file storage platform, each requirement has been refined to ensure technical feasibility and measurable implementation within the project scope.

- **NFR1. Security.** The system must ensure user authentication and data protection. Passwords must be stored securely using hash-based encryption, e.g., SHA-256, and all data transfers like login, upload, download must occur over HTTPS. Access control mechanisms will restrict unauthorized access to user content.
- **NFR2. Scalability.** The platform should be designed to support a moderate increase in users and files without significant performance loss. The implementation will rely on cloud storage services e.g., AWS S3 that allow horizontal scaling when necessary.
- **NFR3. Availability.** The prototype must maintain stable operation during normal usage and testing sessions. Full 24/7 availability is not required at this stage, but the system should be resilient to basic interruptions and restart gracefully.

- **NFR4. Performance.** Core functions such as login, file upload, and download should respond within acceptable limits, thinkink at least in seconds for standard files. Database queries and file retrieval will be optimized for concurrent user requests.
- **NFR5. Usability.** The user interface must be simple, intuitive, and accessible on desktop and mobile browsers. Navigation and file management tasks should be easily performed without user training.
- **NFR6. Maintainability.** The system architecture must follow a modular design, allowing future updates and debugging with minimal effort. Code documentation and naming conventions will ensure ease of maintenance.
- **NFR7. Compatibility.** The web application must operate correctly across major browsers like Chrome, Firefox or Edge. Responsive design must ensure usability on mobile devices.
- **NFR8. Reliability.** The system should handle minor failures without losing user data. Uploads and downloads must include confirmation and error-handling mechanisms. Advanced redundancy is beyond the current scope but will be considered for future iterations.
- **NFR9. Backup and Recovery.** Basic backup mechanisms must be implemented using cloud storage versioning or scheduled export of metadata. Recovery from file deletion or corruption should be possible during testing.
- **NFR10. Logging and Auditability.** The platform must record relevant user operations, login, upload, download and delete in an event log to support monitoring and debugging. Logs will be stored securely and reviewed during system evaluation.
- **NFR11. Interoperability.** The platform must provide RESTful APIs that enable future integration with external applications, such as institutional systems or educational platforms. Initial endpoints will focus on authentication and file access.

3 User Stories

Title	Priority	Estimate
UH-01: User Registration	High	8 SP
As a new user, I want to register for an account by providing my email, password, and personal details, so that I can have my own private storage space on the platform.		
Given I am a new user on the registration page, when I enter a valid email, a secure password, and my required personal details and submit the form, then my account is created, and I receive an email confirmation.		

Title	Priority	Estimate
UH-02: User Login	High	3 SP
As a registered user, I want to log in using my email and password, so that I can securely access my files and storage space.		
Given I am a registered user on the login page, when I enter my correct email and password and click "Login", then I am authenticated and redirected to my personal dashboard.		

Title	Priority	Estimate
UH-03: Secure Password Recovery	Normal	5 SP
As a user who forgot my password, I want to request a password reset via email with a temporary token, so that I can regain access to my account without compromising security.		
Given I am on the login page and click "Forgot Password", when I enter my email address and submit the request, then I receive an email with a secure, temporary link to create a new password.		

Title	Priority	Estimate
UH-04: Upload a File	Must-Have	8 SP
As a logged-in user, I want to upload a file (e.g., .pdf, .jpg, .mp3) up to 100MB to a specific folder, so that I can store and organize my content in the cloud.		
Given I am in my workspace and have navigated to the desired folder, when I drag-and-drop or select a valid file for upload, then the system uploads the file and displays a success message; if there is an error, it clearly informs me.		

Title	Priority	Estimate
UH-05: Create and Navigate Folders	Normal	5 SP
As a user, I want to create new folders and navigate through my folder hierarchy, so that I can organize my files logically and find them easily.		
Given I am viewing the contents of my current folder, when I click "New Folder", provide a name, and confirm, then the new folder appears in my current view, and I can click on it to navigate inside.		

Title	Priority	Estimate
UH-06: View File List and Details	Normal	5 SP
As a user, I want to see a list of my files with their name, size, type, and upload date, so that I can get an overview of my stored content and its properties.		
Given I am logged in and viewing a folder, when the page loads, then I see a list of all items in that folder, displayed in a table or grid with the relevant columns.		

Title	Priority	Estimate
UH-07: Download a File	Normal	3 SP
As a user, I want to download a file from my storage space, so that I can access a local copy on my device.		
Given I am viewing the list of my files, when I click the "Download" button next to a file, then the file download begins immediately via a secure, temporary URL.		

Title	Priority	Estimate
UH-08: Delete a File or Folder	Normal	5 SP
As a user, I want to delete files or folders I no longer need, so that I can free up storage space and keep my workspace tidy.		
Given I have selected one or more files/folders in my workspace, when I click the "Delete" button and confirm the action, then the items are moved to a "Trash" (or permanently deleted with a warning), and the storage space is updated.		

Title	Priority	Estimate
UH-09: View Storage Capacity	High	8 SP
As a user, I want to see a visual indicator of my used and available storage space, so that I can manage my uploads and avoid running out of space.		
Given I am logged into my dashboard, when the page loads, then I see a clear display (e.g., a bar chart or text) showing "X MB used of Y MB available".		

Title	Priority	Estimate
UH-10: Move Files/Folders	Normal	8 SP
As a user, I want to move files and folders from one location to another within my storage, so that I can reorganize my content without having to re-upload it.		
Given I have selected a file or folder, when I select the "Move" action and choose a destination folder from my hierarchy, then the item is moved to the new location and removed from the original one.		

Title	Priority	Estimate
UH-11: Session Control and Security	Very High	5 SP
As a system security manager, I want all file accesses to be validated through session control, so that unauthorized users cannot access my files.		
Given a user is not logged in or their session has expired, when they try to access a direct file URL or an API endpoint, then the system denies the request and redirects them to the login page.		

Title	Priority	Estimate
UH-12: File Activity History	Normal	13 SP
As a user, I want to view a history of my recent actions (uploads, deletions, moves), so that I can track changes and recover from mistakes.		
Given I am on my dashboard, when I navigate to an "Activity" or "History" section, then I see a chronological list of my file operations with timestamps.		

Title	**Priority**	**Estimate**
UH-13: API for Third-Party Integration	Very High	13 SP
As a developer or power user, I want to use a well-documented API to interact with my files, so that I can integrate the storage service with other applications (e.g., an LMS).		
Given I have a valid API token, when I send a GET request to the ‘/api/files‘ endpoint, then I receive a JSON list of the files in my root directory.		

4 Initial Database Architecture

4.1 High-Level Architecture Proposal

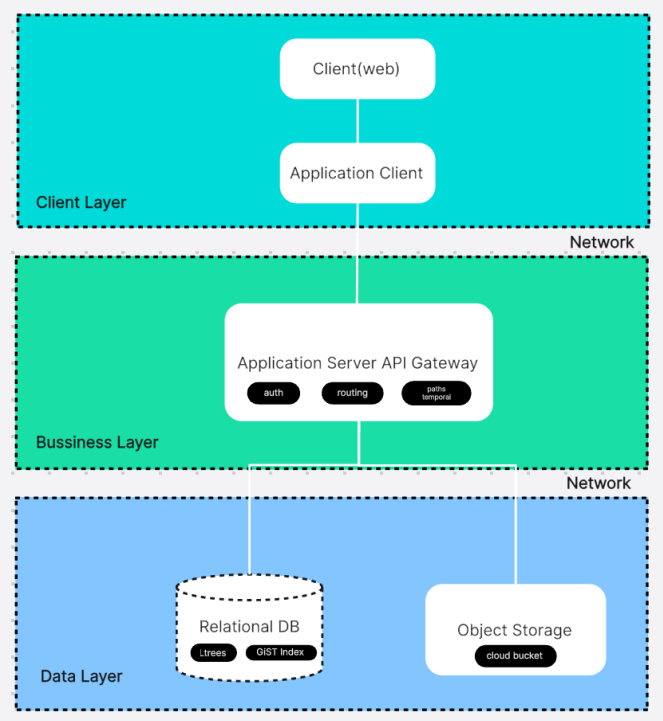

Figure 2: Architecture High Level Model for the File Storage Platform.

4.2 Entity Relationship Diagram - First Version

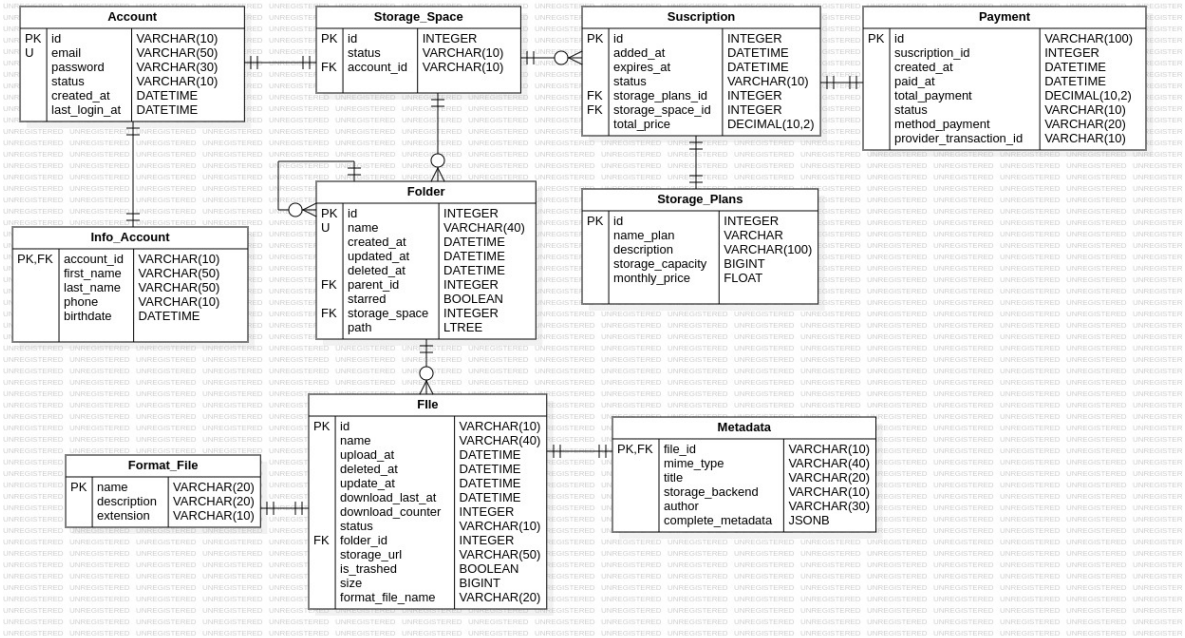

Figure 3: First version of the entity relationship diagram for the file storage platform.

4.2.1 Description of entities

- The **Account** entity represents the user account within the system. It contains main fields such as *id*, *email*, *password*, *status*, *created_at*, and *last_access*. Its role is to be the starting point for all user information, since both personal settings and storage space are derived from it.
- **Info_Account** complements the account with personal data: *first_name*, *last_name*, *phone*, and *birthdate*. It is directly associated with Account and serves to store user identification information.
- The **Storage_Space** entity manages the storage space allocated to each account. It includes *id*, *status*, and a foreign key *account_id*. Its role is to serve as a link between the user and their storage subscriptions.
- **Subscription** stores the data for an active subscription: *id*, *added_at*, *expires_at*, *status*, *total_price*, along with references to the contracted plan and storage space. Its function is to record the conditions of use of the service for a given period.
- The **Payment** entity represents the payments made for a subscription. It contains *id*, *created_at*, *paid_at*, *total_payment*, and *method_payment*. It allows you to keep track of the charges associated with each service contract.
- The **Folder** entity organizes files into hierarchies. Its most important fields are *id*, *name*, *created_at*, *updated_at*, *deleted_at*, *parent_id*, and *starred*. It represents directories that can contain files and subfolders.
- **File** stores user file information. It includes *id*, *name*, *created_at*, *deleted_at*, *download_counter*, *size*, and *folder_id*. It is responsible for representing the digital content within the folders.
- The **Format_File** entity describes the file format. It contains *name*, *description*, and *extension*. Its function is to identify the file type and its basic characteristics.
- Finally, **Metadata** complements the file with specific information such as *mime_type*, *title*, *author*, *storage_backend*, and *complete_metadata*. It serves to expand the technical and descriptive details of each file.

4.2.2 Description of Relationships between Entities

- Account is related to Info_Account, as each account has associated personal information.
- Account is linked to Storage_Space, indicating the storage space allocated to each user.
- Storage_Space is connected to Subscription, which defines the terms of the contracted service.
- Subscription is associated with Payment, which allows the payments for each subscription to be recorded.
- Folder is organized hierarchically by its parent_id, and in turn contains multiple Files.
- File is related to Format_File, to define the file type, and to Metadata, which expands its descriptive information.

4.3 Data Flow and Storage Solutions

Here describes the overall data flow of the platform and the adopted storage solutions. It explains how information moves through the system and how files and metadata are managed across different storage layers.

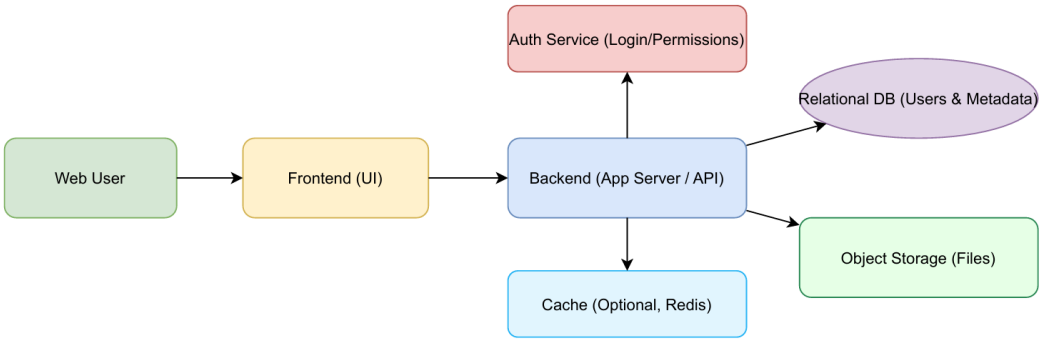

Figure 4: Data Flow and Storage Solutions for the file storage platform.

For the dataflow, it proposed system manages two primary data types: user information and files. The data flow begins when a user interacts with the platform through the web or mobile interface. Requests are sent to the backend application server, which validates user credentials and permissions through the authentication service. Once authenticated, the backend processes the request according to its type:

- For file uploads, the binary object is stored in the cloud object storage, while metadata, e.g., file name, size, type, upload date, and owner is recorded in the relational database.
- For file downloads, the backend verifies user access rights and generates a temporary signed URL, allowing secure file retrieval from the object storage.
- For queries such as listing files or browsing folders, the backend retrieves metadata from the database and may leverage a cache layer to accelerate frequent lookups.

This flow ensures that every operation is validated, traceable, and optimized for performance while maintaining strict access control.

Regarding storage solutions, its design adopts a hybrid model:

- Relational Database: like PostgreSQL or MySQL is used for structured information such as user accounts, roles, permissions, and file metadata. This allows enforcing relationships and maintaining consistency.
- Object Storage: e.g., AWS S3, Azure Blob is used for binary files, ensuring scalability, durability, and redundancy. This approach allows efficient handling of large files with virtually unlimited capacity.
- Cache Layer: for supports faster access to frequently requested metadata, reducing database load.
- Backup and Replication strategies: ensure high availability and disaster recovery, minimizing the risk of data loss.

This architecture separates file storage from metadata management, providing scalability and flexibility. It also ensures that the platform can handle a growing number of users and files without compromising security, reliability, or performance.

5 Information Requirements

This section identifies and describes the main types of information for the file storage platform that must retrieve and manage to support its business processes and user interactions.

5.1 Main Information Types

- **User Account Information.** Includes essential user details such as email, password, registration date, and status. This information supports authentication and user management processes. It relates to user stories *UH-01 User Registration*, *UH-02 User Login*, and *UH-03 Secure Password Recovery*.
- **Personal Profile Data.** Contains user-identifying information. Name, last name, birthdate, contact data. This data improves personalization and communication features, aligning with the *Customer Relationship* block in the Business Model.
- **File Metadata.** Describes the properties of each stored file: name, size, format, upload date, and folder location. It enables browsing, searching, and organizing functionalities. Directly related to user stories *UH-04 Upload a File*, *UH-05 Create and Navigate Folders*, and *UH-06 View File List and Details*.
- **Storage Usage Data.** Includes information about total and available space per user and plan type. This supports the system's scalability control and provides feedback to users about their consumption levels, as seen in *UH-09 View Storage Capacity*.
- **Access Control and Permissions.** Defines which user has access to each file or folder for view, edit or share. Ensures security and privacy according to user roles and authentication tokens. This type of data supports *UH-07 Download a File*, *UH-08 Delete a File or Folder*, and *UH-11 Session Control and Security*.
- **Subscription and Payment Records.** Stores plan details, payment methods, and subscription validity periods. It enables the implementation of the freemium model defined in the *Revenue Streams* block and supports future billing automation. It relates to user story *UH-09 View Storage Capacity* and the business requirement of premium upgrades.
- **Activity Logs.** Logs all relevant system events, including uploads, downloads, deletions, and failed access attempts. This information supports auditing, reliability, and fault diagnosis, as defined in non-functional requirements *NFR10 Logging and Auditability* and user story *UH-12 File Activity History*.
- **API Integration Data.** Contains the tokens, permissions, and endpoints associated with third-party system access. It supports the integration described in the *Interoperability* requirement and relates to user story *UH-13 API for Third-Party Integration*.

Each type of information is strategically connected to the business model and user experience:

- **Customer Segments and Relationships:** User and profile information support personalized and secure access for different user categories, for an initial case: students, professionals, and small businesses.
- **Value Proposition:** File metadata, ACLs, and activity logs guarantee reliable and organized file management, reinforcing the core service value.
- **Revenue Streams:** Subscription and payment data ensure the sustainability of the freemium and premium plans.
- **Key Activities:** Monitoring user behavior and storage usage enables continuous system optimization and scalability planning.

Whith these information requirements that define the data backbone of the system, guiding the subsequent database query design and ensuring consistency between storage, performance, and business objectives.

6 Query Proposal

6.1 User Account Information

This query retrieves the user's stored credentials and account status to verify their identity and check if the account is active.

```
SELECT id, email, password, status
FROM account
WHERE email = 'user@example.com';
```

- Purpose: Authenticate a user during login.
- Insight: Verifies user credentials and retrieves basic account status.

6.2 Personal Profile Data

This query joins the account and info account tables to get a complete view of the user's profile.

```
SELECT a.email, ia.first_name, ia.last_name, ia.phone, ia.birthdate
FROM account a
JOIN info_account ia ON a.id = ia.account_id
WHERE a.id = 'USER123';
```

- Purpose: To display a user's profile information on their account settings page.

6.3 File Metadata

This query retrieves non-trashed files from a specific folder, including their format details, sorted by upload date.

```
SELECT f.id, f.name, f.upload_at, f.size, ff.extension
FROM file f
JOIN format_file ff ON f.format_file_name = ff.name
WHERE f.folder_id = 555 AND f.is_trashed = FALSE
ORDER BY f.upload_at DESC;
```

- Purpose: To list all files within a specific folder for navigation..

6.4 Storage Usage Data

This analytical query aggregates (SUM) the size of all files a user owns and compares it to their plan's capacity.

```
SELECT sp.storage_capacity AS plan_limit,
    COALESCE(SUM(f.size), 0) AS used_space
FROM storage_space ss
JOIN storage_plans sp ON ss.id = sp.id
JOIN file f ON f.storage_space_id = ss.id
WHERE ss.account_id = 'USER123'
GROUP BY sp.storage_capacity;
```

- Purpose: To check a user's total used storage against their plan's limit.

6.5 Access Control and Permissions

This query checks complex ownership and sharing rules. The relational model is strong for enforcing these multi-conditional access controls.

```
SELECT f.id
FROM file f
LEFT JOIN folder fo ON f.folder_id = fo.id
WHERE f.id = 'FILE789'
AND (f.owner_id = 'USER123' OR fo.shared_with @> ARRAY['USER123']);
```

- Purpose: To verify if a user has permission to download a specific file.

6.6 Subscription and Payment Records

This query involve multiple joins across transactional tables (Payment, Subscription, Account) to generate a business report.

```sql
SELECT a.email, s.id AS subscription_id, p.created_at, p.total_payment
FROM payment p
JOIN subscription s ON p.subscription_id = s.id
JOIN account a ON s.account_id = a.id
WHERE p.status = 'FAILED'
AND p.created_at > CURRENT_DATE - INTERVAL '7 days';
```

- Purpose: To find all users with a failed payment for a follow-up process.

6.7 API Integration Data

Similar to the SQL approach, but the flexible document model allows for easily adding new types of permission scopes or metadata to the token without schema changes.

```
db.apiCredentials.findOne(
    { api_key: "tok_xyz123", is_active: true },
    { projection: { user_id: 1, permissions: 1 } }
)
```

- Purpose: To validate an API token and retrieve its associated permissions.