

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH



BÁO CÁO
THỰC TẬP NGOÀI TRƯỜNG
HỌC KỲ 2 NĂM HỌC 2023-2024

- **NGÀNH:** Khoa học Máy tính
- **CHƯƠNG TRÌNH ĐÀO TẠO:** Chất lượng cao - Tăng cường tiếng Nhật
- **ĐƠN VỊ/ DOANH NGHIỆP NHẬN THỰC TẬP:**
FPT Software HCM
- **CÁN BỘ HƯỚNG DẪN CHUYÊN MÔN TRỰC TIẾP CỦA ĐƠN VỊ/ DOANH NGHIỆP:**
Nguyễn Tiến Duy
- **CÁN BỘ HƯỚNG DẪN/ GIÁM SÁT/ CHẤM ĐIỂM BÁO CÁO CỦA KHOA (CBHD/CBGS/CĐBC):**
Nguyễn Quang Đức
- **SINH VIÊN THỰC HIỆN (SVTH):**
HỌ VÀ TÊN: Lê Hoàng Phúc **MSSV:** 2152239
HỌ VÀ TÊN: Lê Nguyễn Phước Lộc **MSSV:** 2153544

TP. Hồ Chí Minh, Tháng 08/2024

Mục lục

| | | |
|----------|--|-----------|
| 1 | LỜI CẢM ƠN | 2 |
| 2 | LỜI MỞ ĐẦU | 3 |
| 3 | TỔNG QUAN VỀ CÔNG TY THỰC TẬP | 4 |
| 3.1 | Giới thiệu công ty | 4 |
| 3.2 | Văn hóa làm việc ở công ty | 4 |
| 4 | NỘI DUNG CÔNG VIỆC THỰC TẬP TẠI CÔNG TY | 5 |
| 4.1 | Nội dung nhiệm vụ chính được giao trong quá trình thực tập | 5 |
| 4.2 | Chi tiết quá trình thực tập | 5 |
| 4.2.1 | Tuần 1: Làm quen với môi trường, văn hóa | 5 |
| 4.2.2 | Tuần 2: củng cố kiến thức về cơ sở dữ liệu, luyện tập SQL | 5 |
| 4.2.3 | Tuần 3: Làm quen với Snowflake | 6 |
| 4.2.4 | Tuần 4: Tìm hiểu thêm về API | 7 |
| 4.2.5 | Tuần 5: Bắt đầu dự án | 8 |
| 4.2.6 | Tuần 6: Sử dụng AWS | 9 |
| 4.2.7 | Tuần 7: Hoàn thiện dự án | 10 |
| 4.2.8 | Tuần 8: Báo cáo và trình bày kết quả đạt được | 11 |
| 4.3 | Cảm nhận | 12 |
| 5 | TỔNG KẾT | 13 |

1 LỜI CẢM ƠN

Để có kiến thức và kết quả thực tế ngày hôm nay, trước hết chúng em xin chân thành cảm ơn các Thầy Cô trong khoa Khoa học và Kỹ thuật máy tính trường Đại học Bách khoa TP.HCM đã giảng dạy và trang bị cho chúng em những kiến thức cơ bản trong các năm đại học. Bên cạnh đó, chúng em xin gửi lời cảm ơn chân thành đến các anh chị trong công ty đã giúp đỡ, chia sẻ kinh nghiệm và tạo mọi điều kiện thuận lợi giúp chúng em hoàn thành tốt quá trình thực tập của mình.

Với thời gian thực tập còn hạn chế và sự hiểu biết thực tế còn nhiều bỏ ngỡ nên bài nên bài báo cáo của chúng em sẽ không tránh khỏi những thiếu sót. Nên chúng em mong nhận được những ý kiến đóng góp để chúng em có thể đúc kết được nhiều bài học và kinh nghiệm cho bản thân, từ đó giúp ích được nhiều cho chúng em khi chính thức bước vào môi trường làm việc.

2 LỜI MỞ ĐẦU

Tại Việt Nam, "Chuyển đổi số" (Digital Transformation) là một khái niệm nổi bật trong thời đại Internet bùng nổ và đặc biệt trở nên cấp bách khi đại dịch Covid-19 diễn ra. Quá trình này đánh dấu sự chuyển mình sâu rộng từ các mô hình doanh nghiệp truyền thống sang mô hình doanh nghiệp số, nhờ vào việc áp dụng các công nghệ tiên tiến như dữ liệu lớn (Big Data) và điện toán đám mây (Cloud). Chuyển đổi số không chỉ thay đổi cách thức hoạt động của doanh nghiệp mà còn tạo ra cơ hội mới để tối ưu hóa quy trình, nâng cao hiệu quả và cải thiện trải nghiệm khách hàng.

Trong bối cảnh chuyển đổi số, Data Engineering giữ vai trò then chốt, là nền tảng vững chắc để khai thác giá trị từ dữ liệu. Các kỹ sư dữ liệu (data engineers) chuyên thiết kế và xây dựng các hệ thống lưu trữ và xử lý dữ liệu, đảm bảo rằng dữ liệu được thu thập, lưu trữ và xử lý một cách hiệu quả và chính xác. Họ phát triển các pipeline dữ liệu để tiếp nhận thông tin từ nhiều nguồn khác nhau, thực hiện các thao tác xử lý cần thiết và duy trì các hệ thống dữ liệu mạnh mẽ. Bằng việc áp dụng các công nghệ tiên tiến như cơ sở dữ liệu, hệ thống xử lý phân tán và các công cụ hỗ trợ hiện đại, kỹ sư dữ liệu giúp doanh nghiệp khai thác và phân tích dữ liệu lớn một cách tối ưu. Điều này không chỉ nâng cao khả năng ra quyết định thông minh mà còn tạo ra giá trị cạnh tranh bền vững, góp phần thúc đẩy sự phát triển và đổi mới trong doanh nghiệp.

3 TỔNG QUAN VỀ CÔNG TY THỰC TẬP

3.1 Giới thiệu công ty



FPT Software HCM (FSoft): FPT Software là công ty thành viên thuộc Tập đoàn FPT. Được thành lập từ năm 1999, FPT Software hiện là công ty chuyên cung cấp các dịch vụ và giải pháp phần mềm cho các khách hàng quốc tế, với hơn 30000 nhân viên, hiện diện tại 28 quốc gia và vùng lãnh thổ trên toàn cầu. Nhiều năm liền, FPT Software được bình chọn là Nhà Tuyển dụng được yêu thích nhất và nằm trong TOP các công ty có môi trường làm việc tốt nhất châu Á.

3.2 Văn hóa làm việc ở công ty

Quy định giờ giấc:

- Làm việc vào các ngày trong tuần (Từ thứ Hai đến thứ Sáu)
- Thời gian: 8:00 - 17:00 (Nghỉ trưa từ 12:00 - 13:00)

4 NỘI DUNG CÔNG VIỆC THỰC TẬP TẠI CÔNG TY

4.1 Nội dung nhiệm vụ chính được giao trong quá trình thực tập

| Tuần | Nội dung công việc |
|------|--|
| 1 | Học nội quy, văn hóa công ty & quy định về bảo mật thông tin |
| 2 | Củng cố kiến thức cần thiết về khoa học & kỹ thuật dữ liệu |
| 3 | Làm quen với Snowflake |
| 4 | Tìm hiểu về API và lập trình để sử dụng API |
| 5 | Bắt đầu Mock project, làm quen với Amazon Web Services (AWS) |
| 6 | Sử dụng các dịch vụ AWS cho Mock project |
| 7 | Hiện thực và sửa lỗi |
| 8 | Tổng hợp và báo cáo |

4.2 Chi tiết quá trình thực tập

4.2.1 Tuần 1: Làm quen với môi trường, văn hóa

Trong ngày đầu tiên, ngày bế giảng kỳ thực tập tại FSoft, chúng em được gặp gỡ các anh chị Mentor và Admin của các lớp thực tập. Sau khi giao lưu giới thiệu, các anh chị Admin đã chia sẻ về lịch sử phát triển của tập đoàn FPT nói chung và FSoft nói riêng, quy định & văn hóa làm việc tại công ty. Sau đó, các anh Mentor cũng có một chuyên mục hướng dẫn sơ bộ về công việc sắp tới của chúng em cùng với phương pháp làm việc sẽ thực hiện trong suốt quá trình thực tập tại công ty.

Trong buổi bế giảng, chuyên mục quan trọng nhất chính là **ISMS** (viết tắt của Information Security Management System, nghĩa là Hệ thống Quản lý An toàn Thông tin); đối với FSoft nói riêng và các công ty phần mềm nói chung, **ISMS** vô cùng quan trọng và ảnh hưởng rất lớn đối với mọi vấn đề về dữ liệu, doanh thu của cả hệ thống, một lỗi vi phạm về **ISMS** có thể dẫn tới thiệt hại không tưởng đối với doanh nghiệp hoặc cá nhân người vi phạm.

Những ngày đầu tiên đi đến văn phòng làm việc là khoảng thời gian chúng em được tìm hiểu thêm về văn hóa làm việc, được giao lưu làm quen với những anh chị Mentor, Admin và những người bạn cùng thực tập tại công ty. Bên cạnh đó, em thực hành một số bài tập lập trình Python và SQL để củng cố lại các kiến thức đã học cho những công việc sắp tới liên quan tới kỹ thuật dữ liệu.

4.2.2 Tuần 2: Củng cố kiến thức về cơ sở dữ liệu, luyện tập SQL

Trong tuần thứ hai, nhiệm vụ chúng em được giao tập trung vào việc củng cố các kỹ năng cơ bản về thiết kế cơ sở dữ liệu và thao tác với dữ liệu bằng **SQL**. Cụ thể, chúng em bắt đầu với việc ôn tập lại các nguyên tắc cơ bản trong thiết kế cơ sở dữ liệu, bao gồm cách xác định và xây dựng các thực thể, thuộc tính, mối quan hệ giữa các bảng, và cách tạo ra một cơ sở dữ liệu có cấu trúc rõ ràng và hiệu quả.

Quá trình ôn tập này giúp chúng em hiểu rõ hơn về cách tổ chức và tối ưu hóa dữ liệu để đảm bảo rằng dữ liệu có thể được truy xuất và quản lý một cách hiệu quả. Chúng em tập trung vào việc tạo các bảng dữ liệu với các khóa chính, khóa ngoại, và các ràng buộc (**constraints**) để đảm bảo tính toàn vẹn dữ liệu (**data integrity**).

Tiếp theo, chúng em chuyển sang phần thực hiện truy vấn dữ liệu bằng **SQL**. Đây là một phần quan trọng không chỉ trong việc thao tác với dữ liệu mà còn trong việc hiểu rõ cách dữ liệu được tổ chức và kết nối với nhau trong cơ sở dữ liệu. Chúng em được yêu cầu viết các câu truy vấn từ đơn giản đến phức tạp, bao gồm các truy vấn **SELECT** cơ bản, các phép nối (**JOIN**) giữa nhiều bảng, và các truy vấn con (**subqueries**). Ngoài ra, chúng em cũng thực hành việc sử dụng các hàm tổng hợp (aggregate functions), nhóm dữ liệu (**GROUP BY**), và lọc dữ liệu (**WHERE**, **HAVING**) để trích xuất thông tin chi tiết từ các tập dữ liệu lớn.

Sau khi đã nắm vững các khái niệm và kỹ năng này, chúng em tiếp tục tìm hiểu về **Snowflake**, một nền tảng quản lý và phân tích dữ liệu trên đám mây. Việc này không chỉ là một phần của nhiệm vụ học tập mà còn là sự chuẩn bị quan trọng cho dự án mock sắp tới, nơi chúng em sẽ sử dụng **Snowflake** như một công cụ chính để xử lý và phân tích dữ liệu.

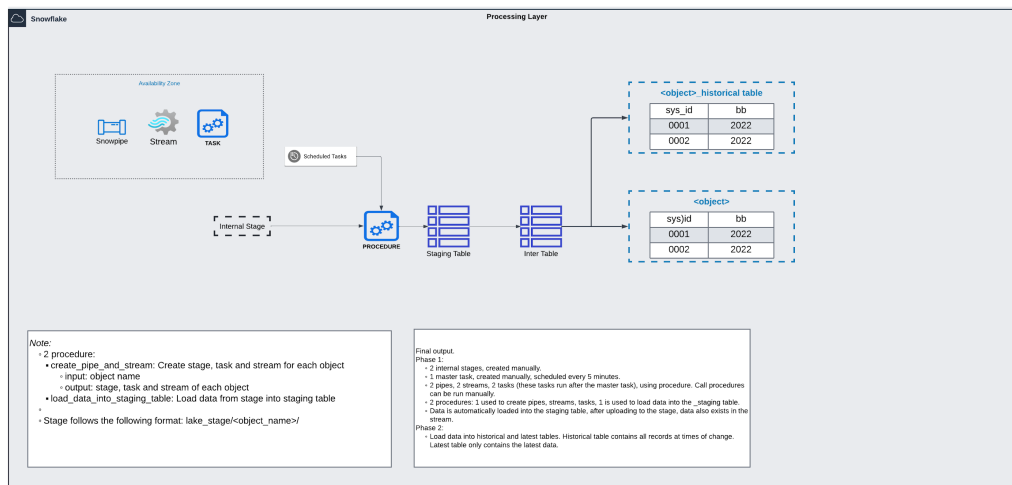
Việc tìm hiểu về **Snowflake** bao gồm việc làm quen với kiến trúc của nền tảng này, cách nó tích hợp các khái niệm cơ sở dữ liệu truyền thống với khả năng mở rộng của đám mây. Chúng em cũng học cách tạo và quản lý các

kho dữ liệu trên **Snowflake**, thực hiện các tác vụ liên quan đến lưu trữ và xử lý dữ liệu, và viết các câu lệnh SQL trên nền tảng này.

Hơn nữa, chúng em tìm hiểu về các tính năng nâng cao của **Snowflake** như khả năng chia sẻ dữ liệu an toàn, tự động tối ưu hóa truy vấn, và cách **Snowflake** hỗ trợ triển khai trên nhiều đám mây khác nhau như AWS, Azure và Google Cloud. Điều này giúp chúng em chuẩn bị hành trang sẵn sàng cho các yêu cầu thực tế trong dự án mock, nơi chúng em sẽ áp dụng những gì đã học để giải quyết các vấn đề thực tế, từ việc thiết kế cơ sở dữ liệu đến xử lý các tập dữ liệu lớn và phức tạp trên **Snowflake**.

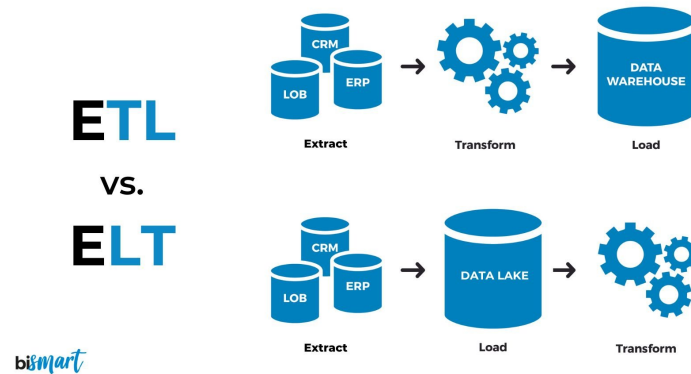
4.2.3 Tuần 3: Làm quen với Snowflake

Trong tuần 3, chúng em được giao một dự án nhỏ để thiết kế một data pipeline sử dụng **Snowflake**, chia làm 2 phase. Cùng với đó là làm quen về kiến trúc **ETL** (Extract, Transform, Load) và **ELT** (Extract, Load, Transform). Kiến trúc pipeline dự án nhỏ như sau:



Hình 1: Kiến trúc Pipeline đơn giản

- Phase 1: Thiết lập và quản lý luồng dữ liệu bên trong **Snowflake**. Nhiệm vụ chính của giai đoạn này là đảm bảo rằng dữ liệu từ các nguồn bên ngoài được tải lên và xử lý tự động vào bảng **staging** trong cơ sở dữ liệu (các bảng "staging" được dùng để chứa dữ liệu thô ban đầu, trước khi xử lý). Các bước thực hiện bao gồm:
 - Tạo **Internal Stage**: Đây là khu vực lưu trữ tạm thời, nơi chúng em tải dữ liệu thô từ các tệp như CSV, JSON, hoặc Parquet lên hệ thống của **Snowflake**.
 - Tạo **Pipe** để tự động tải dữ liệu vào bảng **staging**: **Pipe** tự động phát hiện dữ liệu mới trong **Stage** và nhập vào bảng **staging**, giúp bảng này luôn cập nhật với dữ liệu mới nhất từ **Stage**.
 - Tạo **Stream** để theo dõi thay đổi dữ liệu: **Stream** giám sát các thay đổi trong bảng **staging** (như thêm, sửa hay xóa), hỗ trợ quản lý và xử lý dữ liệu hiệu quả.
 - Tạo **Task** để Tự Động Hóa: **Task** được cấu hình chạy mỗi 5 phút, tự động kiểm tra **Stage** và tải dữ liệu vào bảng **staging**.
- Phase 2: Tiếp tục mở rộng và hoàn thiện quy trình quản lý dữ liệu bằng cách di chuyển dữ liệu từ bảng **staging** (được thiết lập trong Phase 1) sang hai bảng mới là **latest** và **historical**. Mục tiêu chính của giai đoạn này là đảm bảo rằng hệ thống có thể duy trì dữ liệu cập nhật và đồng thời lưu trữ lịch sử dữ liệu cho mục đích phân tích và tra cứu lâu dài. Các bước chính bao gồm:
 - Tạo **Task** để tải dữ liệu vào bảng Intermediate (bảng **inter**): Chuyển dữ liệu từ bảng **staging** sang bảng **inter**. Bảng **inter** dùng để lưu dữ liệu đã được xử lý và làm sạch dữ liệu trước khi chuyển tiếp đi tới các khâu tiếp theo.
 - Tạo **Task** để tải dữ liệu từ bảng **inter** vào bảng **latest** và **historical**: Di chuyển dữ liệu từ bảng **inter** sang bảng **latest** (nơi lưu trữ dữ liệu mới nhất và cập nhật) và bảng **historical** (lưu trữ toàn bộ lịch sử dữ liệu).



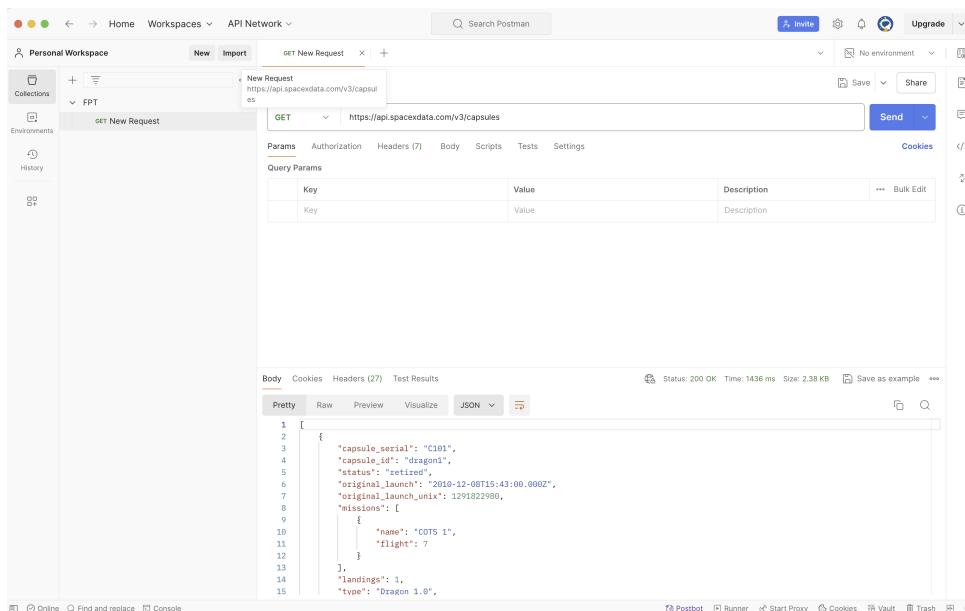
Hình 2: Kiến trúc ETL và ELT

Mục tiêu của dự án là thiết lập một quy trình pipeline hiệu quả, từ việc tải dữ liệu từ **Stage**, sau đó xử lý và làm sạch dữ liệu và cuối cùng là lưu trữ dữ liệu vào các bảng **latest** và **historical**, đảm bảo dữ liệu được cập nhật liên tục và lưu trữ đầy đủ.

4.2.4 Tuần 4: Tìm hiểu thêm về API

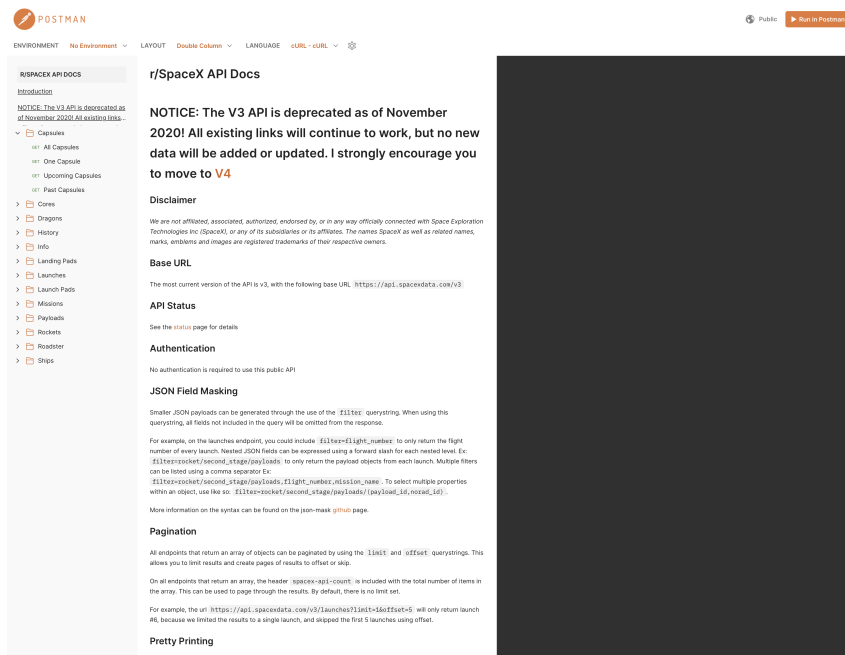
Trong tuần 4, chúng em tìm hiểu thêm về các **API** (Application Programming Interface) và cách lấy dữ liệu từ các **API**, cụ thể là:

- Khám phá các phương pháp để truy cập dữ liệu từ **API**, bao gồm việc gửi yêu cầu đến các điểm cuối (endpoints) của **API** và xử lý phản hồi nhận được.
- Sử dụng **Postman**, một công cụ mạnh mẽ để gửi các **yêu cầu API** (API request) và kiểm tra các phản hồi. Chúng em sẽ học cách tạo và gửi các **yêu cầu API**, cấu hình các tham số và tiêu đề yêu cầu, và phân tích dữ liệu phản hồi từ các **API**.



Hình 3: Giao diện Postman thực hiện lấy API bất kì

Sau khi tìm hiểu về các **API** và cách sử dụng **Postman**, chúng em chuyển sang bước tiếp theo là sử dụng **Python** để lấy dữ liệu từ các **API** và lưu trữ dữ liệu. Hình ảnh sơ qua về **API** free công khai của SpaceX, nơi cung cấp dữ liệu miễn phí cho dự án nhỏ này.



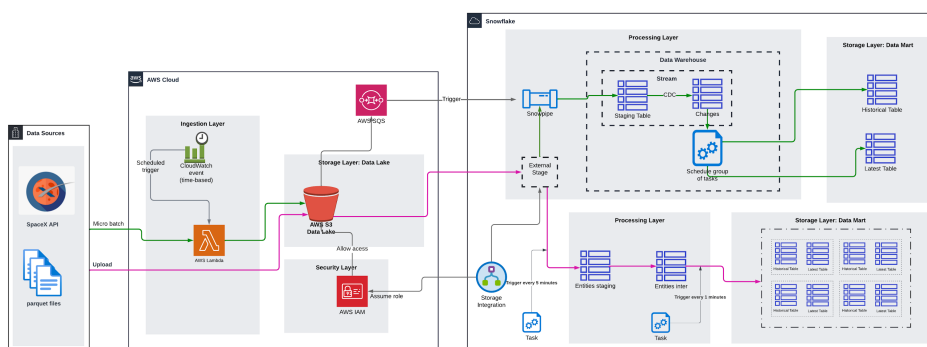
Hình 4: Hình ảnh trực quan về doc của SpaceX về API

- Sử dụng **Python** để gửi **yêu cầu API**, cụ thể là thư viện "request" để gửi yêu cầu đến các **endpoint** trong **API**.
- Dữ liệu sau khi nhận được từ **API** được lưu vào các file **JSON**.

Mục tiêu của các công việc này là hiểu cách thức lưu trữ dữ liệu từ **API** để có thể áp dụng vào **AWS Lambda**. Sử dụng **AWS Lambda** cho phép chúng em thực thi mã **Python** để lấy dữ liệu từ **API**, xử lý và lưu trữ vào các dịch vụ như **S3** hoặc **DynamoDB**, tối ưu hóa quy trình và nâng cao hiệu quả cũng như khả năng mở rộng.

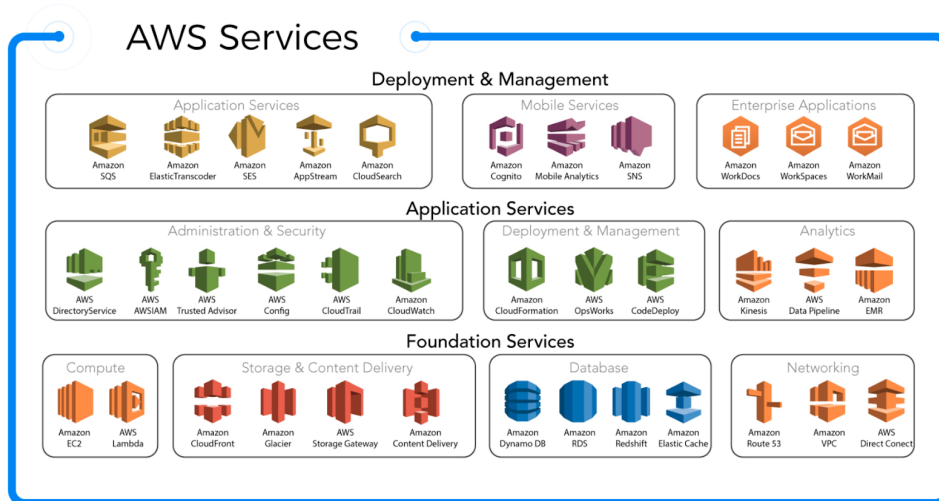
4.2.5 Tuần 5: Bắt đầu dự án

Tuần thứ 5 là tuần bắt đầu Final Project. Mục tiêu của dự án là áp dụng những kiến thức đã học ở trường và tại công ty để hiện thực hóa một data pipeline. Cụ thể, dự án sẽ tập trung vào việc thu thập dữ liệu từ dịch vụ **S3 Bucket** của **AWS**, tải dữ liệu vào **Snowflake**, và thực hiện các bước xử lý dữ liệu trong **Snowflake** để đảm bảo dữ liệu được tổ chức và phân tích một cách hiệu quả.

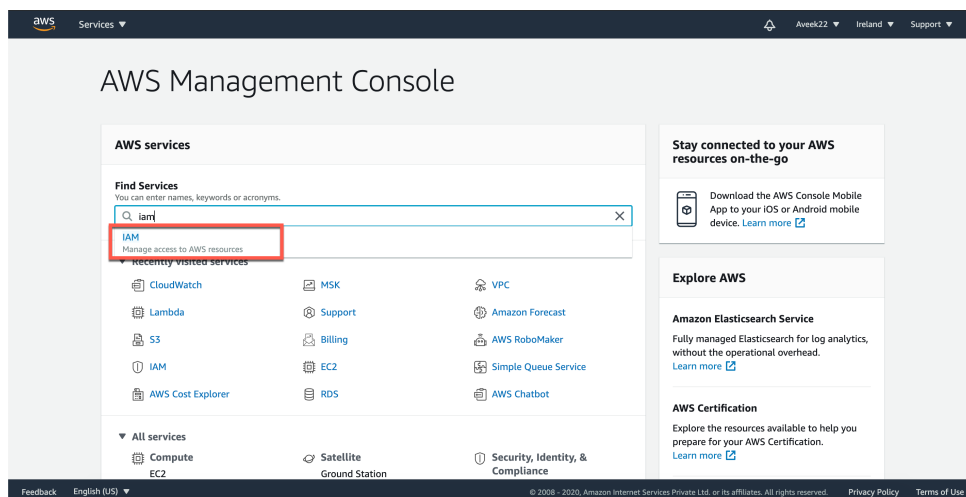


Hình 5: Kiến trúc Pipeline

- Làm quen với nền tảng điện toán đám mây **AWS** với các chức năng phục vụ cho công việc về kỹ thuật dữ liệu như **IAM**, **S3 Bucket**, **Lambda Function**,...



Hình 6: Một số dịch vụ mà AWS cung cấp

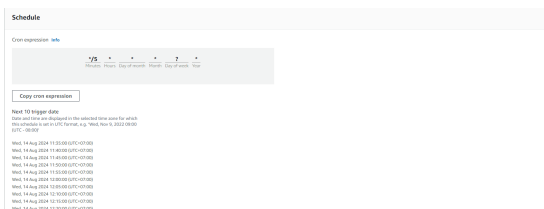


Hình 7: Giao Diện AWS IAM Service

4.2.6 Tuần 6: Sử dụng AWS

Trong tuần thứ 6, chúng em sử dụng các dịch vụ của AWS để tạo nguồn dữ liệu, đồng thời cũng tạo điều kiện thuận lợi cho việc tích hợp và xử lý dữ liệu trong data pipeline, từ đó tối ưu hóa quy trình và nâng cao hiệu quả phân tích.

- Cấu hình **CloudWatch Services** của AWS để tự động trigger **Lambda Function** mỗi 5 phút hoặc ít hơn.

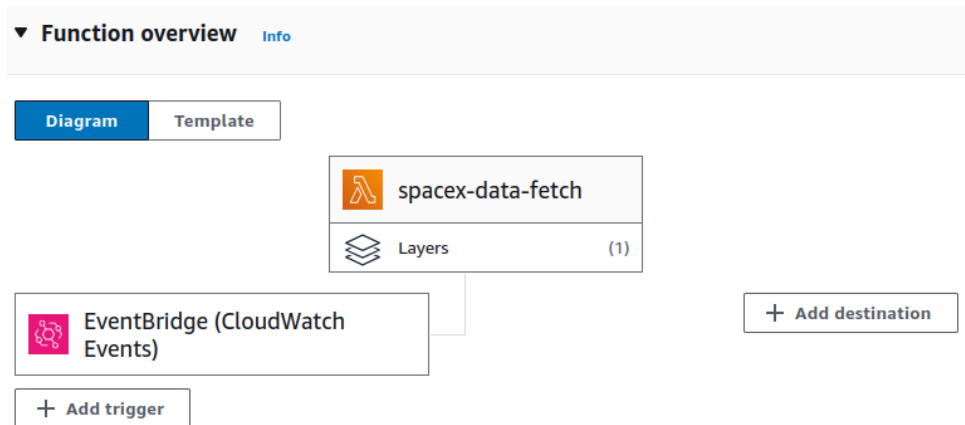


Hình 8: Cloudwatch để trigger Lambda Function

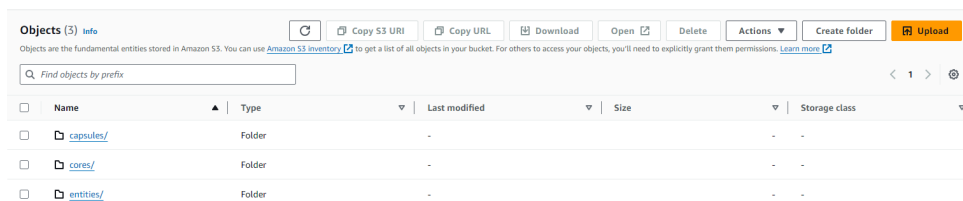
Hình 9: Chi tiết về lịch chạy của CloudWatch

- Thực hiện sử dụng **Lambda Function** để đẩy dữ liệu từ **SpaceX API** và đẩy vào **S3 Bucket**.
- Thực hiện cài đặt quyền để có thể đẩy dữ liệu từ **Lambda Function** vào **S3 Bucket**.

- Thực hiện cài đặt quyền để Snowflake có thể lấy dữ liệu từ **S3 Bucket**



Hình 10: Hình ảnh từ hệ thống AWS về EventBridge và Lambda Function

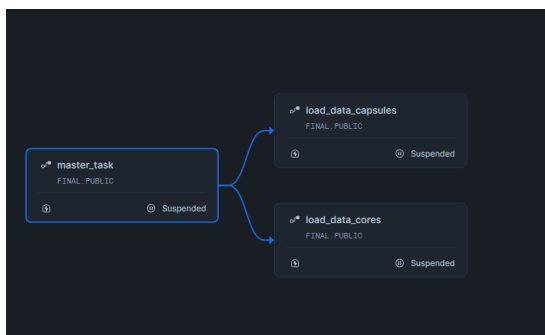


Hình 11: Sơ lược về S3 Bucket sau khi đã được lambda function trigger

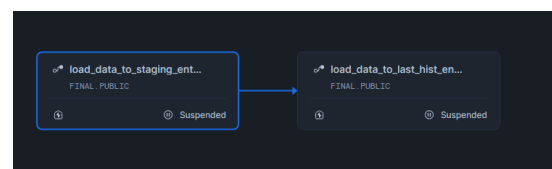
4.2.7 Tuần 7: Hoàn thiện dự án

Sau khi đã hoàn thành việc cài đặt việc lưu trữ dữ liệu trong data lake **S3 Bucket**, chúng em đã hiện thực trong **Snowflake** các bước sau:

- Thực hiện các thao tác trên các bảng được load từ **S3 Bucket** để tách các bảng đó ra thành các bảng con nếu có và các bảng **latest** và **historical** để có thể quản lý dữ liệu mới nhất và thực hiện các thao tác trên dữ liệu đó bằng các **Tasks** nối liền nhau.
- Viết Query để thực hiện toàn bộ quá trình một cách tự động hoàn toàn bằng **SnowSQL**.
- Thực hiện viết các **Master Task** và các **Task** con của nó để thực hiện tự động theo thời gian định trước.



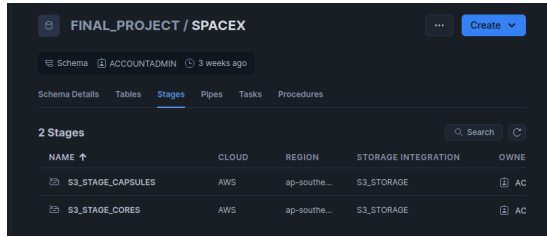
Hình 12: Master Task hiện thực quá trình load dữ liệu từ spacex datalake vào bảng latest và historical



Hình 13: Master Task hiện thực quá trình load dữ liệu từ entities datalake vào bảng latest và historical

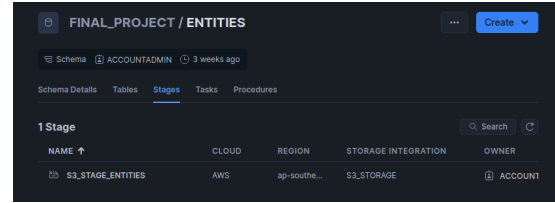
4.2.8 Tuần 8: Báo cáo và trình bày kết quả đạt được

Sau khi đã hoàn thành xong dự án, chúng em báo cáo với anh Mentor kết quả quá trình cũng như các khó khăn gặp phải trong quá trình làm việc cùng với một số kết quả đạt được sau quá trình làm việc.



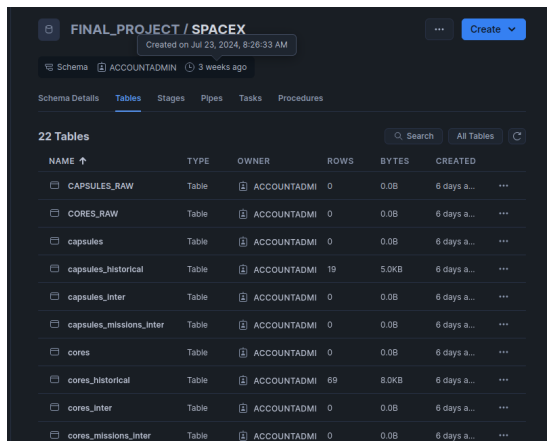
| NAME | CLOUD | REGION | STORAGE INTEGRATION | OWNER |
|-------------------|-------|--------------|---------------------|-------|
| S3_STAGE_CAPSULES | AWS | ap-southe... | S3_STORAGE | AC |
| S3_STAGE_CORES | AWS | ap-southe... | S3_STORAGE | AC |

Hình 14: SpaceX Stage để lưu trữ các file được load vào AWS S3 Bucket



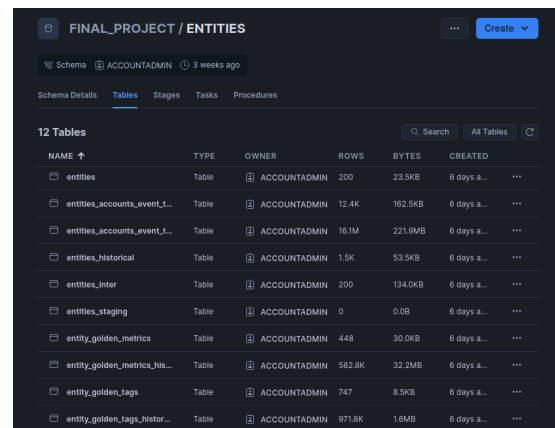
| NAME | CLOUD | REGION | STORAGE INTEGRATION | OWNER |
|-------------------|-------|--------------|---------------------|----------|
| S3_STAGE_ENTITIES | AWS | ap-southe... | S3_STORAGE | ACCOUNT1 |

Hình 15: SpaceX Stage để lưu trữ các file được load vào AWS S3 Bucket



| NAME | TYPE | OWNER | ROWS | BYTES | CREATED |
|-------------------------|-------|--------------|------|-------|-------------|
| CAPSULES_RAW | Table | ACCOUNTADMIN | 0 | 0.0B | 6 days a... |
| CORES_RAW | Table | ACCOUNTADMIN | 0 | 0.0B | 6 days a... |
| capsules | Table | ACCOUNTADMIN | 0 | 0.0B | 6 days a... |
| capsules_historical | Table | ACCOUNTADMIN | 19 | 5.0KB | 6 days a... |
| capsules_inter | Table | ACCOUNTADMIN | 0 | 0.0B | 6 days a... |
| capsules_missions_inter | Table | ACCOUNTADMIN | 0 | 0.0B | 6 days a... |
| cores | Table | ACCOUNTADMIN | 0 | 0.0B | 6 days a... |
| cores_historical | Table | ACCOUNTADMIN | 69 | 8.0KB | 6 days a... |
| cores_inter | Table | ACCOUNTADMIN | 0 | 0.0B | 6 days a... |
| cores_missions_inter | Table | ACCOUNTADMIN | 0 | 0.0B | 6 days a... |

Hình 16: Các bảng thuộc API SpaceX gồm Latest và Historical cùng với các bảng con



| NAME | TYPE | OWNER | ROWS | BYTES | CREATED |
|------------------------------|-------|--------------|--------|---------|-------------|
| entities | Table | ACCOUNTADMIN | 200 | 23.9KB | 6 days a... |
| entities_accounts_event_t... | Table | ACCOUNTADMIN | 12.4K | 162.5KB | 6 days a... |
| entities_accounts_event_t... | Table | ACCOUNTADMIN | 16.1M | 221.9MB | 6 days a... |
| entities_historical | Table | ACCOUNTADMIN | 1.5K | 53.9KB | 6 days a... |
| entities_inter | Table | ACCOUNTADMIN | 200 | 134.0KB | 6 days a... |
| entities_staging | Table | ACCOUNTADMIN | 0 | 0.0B | 6 days a... |
| entity_golden_metrics | Table | ACCOUNTADMIN | 448 | 30.0KB | 6 days a... |
| entity_golden_metrics_his... | Table | ACCOUNTADMIN | 582.8K | 32.2MB | 6 days a... |
| entity_golden_tags | Table | ACCOUNTADMIN | 747 | 8.5KB | 6 days a... |
| entity_golden_tags_histor... | Table | ACCOUNTADMIN | 971.8K | 1.6MB | 6 days a... |

Hình 17: Các bảng thuộc Entities gồm Latest và Historical cùng với các bảng con

4.3 Cảm nhận

Sau khi thực tập ở vị trí Data Engineer, chúng em đã có những trải nghiệm quý giá và học hỏi được rất nhiều điều. Trước hết, chúng em đã hiểu sâu hơn về cách xây dựng và quản lý các pipeline dữ liệu, từ việc thu thập, xử lý đến lưu trữ và phân tích dữ liệu. Công việc của một Data Engineer đòi hỏi không chỉ kiến thức kỹ thuật mà còn cả khả năng giải quyết vấn đề, tư duy logic và sự tỉ mỉ trong từng chi tiết.

Qua quá trình thực tập, chúng em cũng nhận thấy tầm quan trọng của việc làm sạch và chuẩn hóa dữ liệu, bởi dữ liệu chất lượng cao là yếu tố then chốt để đưa ra những phân tích chính xác và quyết định đúng đắn. Đặc biệt, việc tiếp cận với các công nghệ mới và công cụ hiện đại như AWS, Snowflake, và các hệ thống xử lý phân tán đã mở rộng tầm nhìn của chúng em về lĩnh vực này, giúp chúng em nắm bắt được cách các doanh nghiệp tận dụng dữ liệu để tạo ra giá trị thực tế.

Ngoài ra, môi trường làm việc chuyên nghiệp cùng với sự hỗ trợ từ anh Mentor cùng các đồng nghiệp đã giúp chúng em nhanh chóng hòa nhập và phát triển. Những thách thức mà chúng em đã gặp phải trong suốt quá trình thực tập không chỉ giúp chúng em rèn luyện kỹ năng mà còn củng cố sự tự tin trong việc đối mặt với các vấn đề phức tạp trong lĩnh vực dữ liệu.

Trải nghiệm thực tập này không chỉ mang lại cho chúng em kiến thức và kỹ năng cần thiết mà còn giúp chúng em khẳng định niềm đam mê với lĩnh vực Data Engineering, đồng thời định hướng rõ ràng hơn cho con đường sự nghiệp trong tương lai.

5 TỔNG KẾT

Trong suốt quá trình thực tập, em đã được hưởng nhiều quyền lợi và cơ hội học hỏi quý báu, giúp em phát triển kỹ năng và kiến thức trong lĩnh vực phát triển phần mềm nói chung và kiểm thử phần mềm nói riêng. Hơn nữa, em đã được tiếp xúc với các vấn đề và thách thức mà các doanh nghiệp đang đối diện liên quan đến chuyển đổi số.

Môi trường làm việc chuyên nghiệp, luôn hòa đồng, vui vẻ, giúp đỡ nhau trong công việc. Các thành viên trong nhóm luôn chia sẻ kiến thức, kinh nghiệm để mọi người đều có thể hoàn thành công việc một cách tốt nhất cũng như phát triển bản thân một cách hoàn thiện nhất.

Các thành viên Ban Quản lý luôn hòa đồng với nhân viên, thăm hỏi và giúp đỡ nhân viên trong quá trình làm việc cũng như sinh viên thực tập tại công ty. Môi trường làm việc tại công ty luôn mang đến cho em sự thoải mái nhưng không thiếu sự nghiêm túc trong công việc. Thực tập tại công ty, em được học hỏi nhiều điều mới, môi trường làm việc thực tế giúp em mạnh dạn hơn trong công việc và giao tiếp. Thời gian thực tập tại công ty sẽ là một hành trang vững chắc cho em trên con đường làm việc sau này.

TP. Hồ Chí Minh, tháng 8 năm 2024

Sinh viên

Lê Hoàng Phúc

Lê Nguyễn Phước Lộc