**CẨM NANG NGHỀ DATA**

Trong kỷ nguyên số, khi mà các công việc mang tính data trỗi dậy, người ta tuyển dụng nhiều các vị trí liên quan đến data nhưng lại không rạch ròi về đầu công việc, khiến các ứng viên nhầm tưởng. Do đó, một phần trong quyển sách này, tôi muốn giới thiệu về vai trò của họ trong một dự án tại doanh nghiệp.

Các công việc liên quan đến data được tuyển dụng với nhiều chức danh khác nhau như Data Analyst, Data Scientist, Business Intelligence Analyst, Data Engineer, Database Administrator. Mặc dù các vai trò này đều liên quan đến việc tiếp cận, xử lý và phân tích dữ liệu, nhưng mỗi vị trí lại có những trách nhiệm, kỹ năng và yêu cầu riêng.

Ví dụ, Data Analyst chịu trách nhiệm thu thập, làm sạch và phân tích dữ liệu để đưa ra những báo cáo, thông tin có giá trị phục vụ cho quá trình ra quyết định của doanh nghiệp. Họ cần có kỹ năng về Excel, SQL, các công cụ phân tích dữ liệu như Power BI, Tableau, Python,... Bên cạnh đó, họ cũng cần phải có những kỹ năng mềm như giao tiếp, trình bày, giải thích những phân tích của mình một cách dễ hiểu.

Trong khi đó, Data Scientist lại chịu trách nhiệm xây dựng các mô hình dự báo, học máy nhằm giúp doanh nghiệp đưa ra những quyết định chiến lược dựa trên dữ liệu. Họ cần có nền tảng toán học, thống kê vững chắc, cũng như kỹ năng lập trình, đặc biệt là Python và các thư viện phân tích dữ liệu.

Như vậy, mặc dù các vai trò này đều liên quan đến data, nhưng mỗi vị trí lại có những yêu cầu, trách nhiệm riêng. Trong chương này, chúng ta sẽ đi sâu vào từng vai trò cụ thể, hiểu rõ về nhiệm vụ, kỹ năng cần có, cũng như những cơ hội và thách thức của mỗi công việc liên quan đến data.

Ngoài ra, cuốn sách này sẽ cung cấp cho bạn một cái nhìn tổng quan về các ngôn ngữ, nền tảng và kỹ năng cần thiết để thành công trong nghề này. Nó cũng chia sẻ những câu chuyện thực tế về các thách thức thường gặp trong công việc, giúp bạn có thể lựa chọn con đường phù hợp nhất với mục tiêu và sở thích của bản thân.

MỤC LỤC

[1 DATA CARRER PATH 4](#_Toc168135659)

[1.1 Business Intelligence 8](#_Toc168135660)

[1.2 Data analyst 10](#_Toc168135661)

[1.3 Data engineer 10](#_Toc168135662)

[1.4 Data scientist 10](#_Toc168135663)

[**2** **Nền tảng và ngôn ngữ lưu trữ và phân tích dữ liệu** 12](#_Toc168135664)

[2.1 Câu lệnh SQL chuẩn 13](#_Toc168135665)

[2.2 Câu lệnh SQL mở rộng 15](#_Toc168135666)

[2.3 Tự động hóa với quan hệ SQL Excel 15](#_Toc168135667)

[**3** **Nền tảng và ngôn ngữ phân tích và tạo báo cáo** 15](#_Toc168135668)

[3.1 Câu lệnh Power BI 17](#_Toc168135669)

[3.2 Tự động hóa với quan hệ SQL Power BI 17](#_Toc168135670)

# DATA CARRER PATH

Trong kỷ nguyên số, khi mà các công việc mang tính data trỗi dậy, người ta tuyển dụng nhiều các vị trí liên quan đến data nhưng lại không rạch ròi. Bọn tay mơ như chúng tôi ngày ấy thậm chí còn không phân biệt được nó thuộc thể loại gì cho đến khi đọc chính xác trách nhiệm và kỷ năng yêu cầu trong mô tả công việc. Do đó, trong quyển sách này, tôi muốn nói một chút về nghề data và vai trò của họ trong một dự án

Để mọi người dễ hình dung, tôi có một ví dụ một vài tình huống thực tế trong doanh nghiệp như sau

Ví dụ có một website mạng xã hội tên Facepen. Phía dưới là giao diện cá nhân của tôi trên Facepen với dữ liệu tên, tuổi, email và ảnh đại diện của tôi:

|  |
| --- |
| Facepen website |

Data engineer sẽ là người xây dựng đường dẫn truyền dữ liệu (pipeline) để truyền dữ liệu từ website này lên Data Lake nơi có thể chứa tất cả các loại dữ liệu không có cấu trúc, bao gồm ảnh đại diện của tôi. Sau đó, từ Data lake anh ấy sẽ xử lý dữ liệu để thanh lọc chỉ còn dữ liệu được cấu trúc (Structured) đưa lên Data Ware House (DWH). Toàn bộ quá trình này người ta gọi đây là ETL. Trong quá trình ETL, tùy thuộc vào môi trường mà doanh nghiệp lựa chọn, anh ấy có thể sử dụng SQL, Python, Java,… (các ngôn ngữ này sẽ được nói chi tiết hơn ở các chương sau)

|  |
| --- |
| SQL server view |

Sau đó, Business Intelligence sử dụng các bảng dữ liệu này để vẽ dashboard trên Excel hoặc Power BI cho các cấp lãnh đạo hoặc khách hàng theo dõi số liệu một cách thường xuyên.

|  |
| --- |
| Power BI |

Mọi chuyện rất suông sẻ cho đến khi xuất hiện một requirement đến từ Product Owner. Chị ấy yêu cầu phải có thêm trường dữ liệu số điện thoại khách hàng để tăng tính bảo mật cho trang cá nhân của họ. Vậy là phát sinh công việc mới.

Data Analyst sẽ là người tham gia vào để xác định liệu giao diện màn hình nào, bảng dữ liệu nào sẽ bị ảnh hưởng trong dự án lần này. Đội Solfware Engineer xây dựng một trường dữ liệu nhận thông tin số điện thoại khách hàng trên website còn Data engineer lại được bỏ vào giỏ hàng cho việc đảm bảo số điện thoại được kéo từ Website về Server.

Một ví dụ khác trong ngân hàng

## Business Intelligence

Ở Việt Nam, bạn có thể tìm thấy Business Intelligence dưới các chức danh khác nhau tại các tên tuổi lớn trong và ngoài nước như Unilever, Nestlé, P&G, Vingroup, FPT, Viettel, Masan, Unilever, Nestlé, Procter & Gamble, Vinamilk,… Đối với nhóm ngành tài chính, những cái tên mới nỗi như Momo, Fe Credit, Mcredit,… thậm chí ở những ngân hàng già cỗi trong nước như BIDV, Techcombank, VPBank,… tôi còn thấy nhân sự đã quá quen với việc xuất dữ liệu trực tiếp từ hệ thống đặc biệt là những người làm phân tích.

Tùy vào cơ cấu và loại hình doanh nghiệp, các chức danh tương tự có thể là Business Intelligence còn có thể là Reporting analyst, Financial Analyst,… ở cấp độ cao hơn nó có thể là Head of reporting, Head of Business Intelligence,...

**Vai trò và trách nhiệm**

* Thiết kế và xây dựng các hệ thống báo cáo, bảng điều khiển (dashboard) để cung cấp các thông tin quan trọng cho các nhà quản lý và lãnh đạo.
* Phân tích dữ liệu từ nhiều nguồn khác nhau để tìm ra các mô hình, xu hướng và insights có giá trị cho việc ra quyết định kinh doanh.
* Xác định các chỉ số hiệu suất then chốt (KPI) và theo dõi chúng để đánh giá hiệu quả hoạt động của doanh nghiệp.
* Hỗ trợ các bộ phận khác như marketing, sales, tài chính trong việc ra quyết định dựa trên dữ liệu.
* Đề xuất các giải pháp công nghệ để cải thiện khả năng thu thập, quản lý và phân tích dữ liệu.

**Ngôn ngữ và môi trường dữ liệu**

* SQL để truy vấn và thao tác dữ liệu từ các hệ thống cơ sở dữ liệu.
* Python, R để thực hiện phân tích dữ liệu, xây dựng mô hình và tự động hóa các quy trình.
* Các công cụ BI như Power BI, Tableau, Qlik để xây dựng báo cáo và bảng điều khiển.
* Kho dữ liệu (data warehouse), hồ chứa dữ liệu (data lake) để lưu trữ và tổng hợp dữ liệu.

**Kỹ năng cần có và cái khó của nghề**

Như ví dụ phía trên, bạn có thể thấy người

## Data analyst

## Data engineer

## Data scientist

**Các chức danh tương tự**

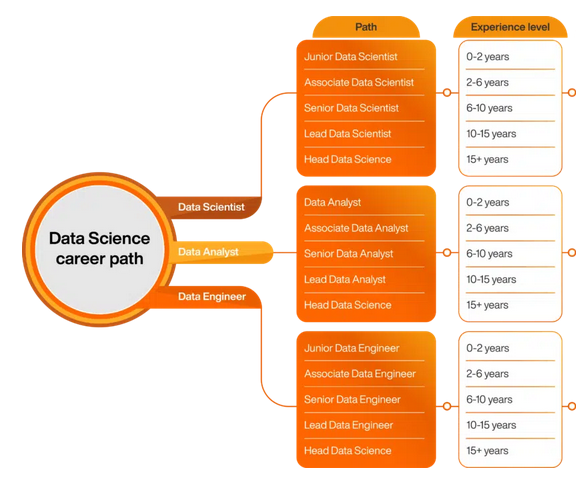
Các chức danh tương tự Business Intelligence còn có thể là Data Scientist, Data researcher,… ở cấp độ cao hơn nó có thể là Head of advance analytics, Sr/Mid Data Scientist,...

**Vai trò và trách nhiệm**

**Ngôn ngữ và môi trường dữ liệu**

**Kỹ năng mềm**

Sơ đồ Carrer path Data



# **Nền tảng và ngôn ngữ lưu trữ dữ liệu**

Ta đã nói nhiều về SQL, vậy SQL là gì?

SQL là viết tắt của Structured Query Language. Nói một cách dân gian, nó là ngôn ngữ giúp bạn xuất dữ liệu từ một hệ thống tập trung đã dược xử lý gọn gàn theo một trật tự được gọi là Server.

Việc sử dụng các ngôn ngữ SQL và nền tảng cơ sở dữ liệu (nơi bạn viết câu lệnh) khác nhau tùy thuộc vào nhu cầu và yêu cầu của mỗi doanh nghiệp.

* **Nền tảng MySQL,** thường được sử dụng trong các ứng dụng web, thương mại điện tử, blog, và các doanh nghiệp vừa và nhỏ. MySQL Sử dụng SQL chuẩn, với một số chức năng mở rộng đặc biệt của MySQL. Ưu điểm: Dễ sử dụng, miễn phí, tính năng phong phú, hiệu suất cao.
* **PostgreSQL**, phổ biến trong các ứng dụng yêu cầu tính toàn vẹn dữ liệu cao, được sử dụng nhiều ở các doanh nghiệp châu Âu, Mỹ. Ưu điểm: Hỗ trợ các tính năng nâng cao như lập trình, giao dịch, và phân tích dữ liệu. PostgreSQL cũng sử dụng SQL chuẩn, nhưng có thêm một số chức năng nâng cao như hỗ trợ lập trình PL/pgSQL. PL/pgSQL cho phép tạo các stored procedures, functions, triggers và các tính năng lập trình nâng cao khác trong PostgreSQL.
* **Oracle Database**: Được sử dụng rộng rãi trong các tập đoàn lớn, hệ thống doanh nghiệp, và các ứng dụng phức tạp. Oracle cũng sử dụng SQL chuẩn, nhưng có thêm một số chức năng riêng như package, cursor, và các tính năng quản trị nâng cao của Procedural Language/SQL (PL/SQL). PL/SQL cho phép tạo các stored procedures, functions, triggers và packages trong Oracle. Ưu điểm: Khả năng mở rộng, an ninh, và các tính năng quản trị dữ liệu mạnh mẽ.

Ví dụ trực quan:

* **Microsoft SQL Server**: Phổ biến trong các doanh nghiệp sử dụng các sản phẩm Microsoft khác như Windows, Office, và .NET. Ưu điểm: Tích hợp tốt với các công cụ Microsoft, thân thiện với người dùng. Microsoft cũng sử dụng SQL chuẩn, nhưng có thêm một số chức năng mở rộng của SQL Server như Transact-SQL (T-SQL). T-SQL cung cấp các tính năng nâng cao như biến, điều khiển luồng, xử lý lỗi, và nhiều chức năng khác
* **SQLite**: Được sử dụng rộng rãi trong các ứng dụng di động và nhúng, do có kích thước nhỏ và dễ cài đặt. Ưu điểm: Không yêu cầu cài đặt server, dễ sử dụng, mã nguồn mở.

## Câu lệnh SQL chuẩn

1. **SELECT:**

Dùng để truy vấn và lấy dữ liệu từ bảng trong cơ sở dữ liệu.

Ví dụ: SELECT \* FROM employees; hoặc SELECT first\_name, last\_name FROM employees;

1. **INSERT:**

Dùng để thêm dữ liệu mới vào bảng.

Ví dụ: INSERT INTO employees (first\_name, last\_name, email) VALUES ('John', 'Doe', 'john.doe@example.com');

1. **UPDATE:**

Dùng để cập nhật dữ liệu trong bảng.

Ví dụ: UPDATE employees SET email = 'jane.doe@example.com' WHERE id = 2;

1. **DELETE:**

Dùng để xóa dữ liệu từ bảng.

Ví dụ: DELETE FROM employees WHERE id = 3;

1. **CREATE TABLE:**

Dùng để tạo mới một bảng trong cơ sở dữ liệu.

Ví dụ: CREATE TABLE employees (id INT PRIMARY KEY, first\_name VARCHAR(50), last\_name VARCHAR(50), email VARCHAR(100));

1. **ALTER TABLE:**

Dùng để thay đổi cấu trúc của bảng, như thêm/xóa/sửa đổi cột.

Ví dụ: ALTER TABLE employees ADD COLUMN phone VARCHAR(20);

1. **DROP TABLE:**

Dùng để xóa hoàn toàn một bảng khỏi cơ sở dữ liệu.

Ví dụ: DROP TABLE employees;

1. **JOIN:**

Dùng để kết hợp dữ liệu từ nhiều bảng dựa trên một điều kiện liên kết.

Ví dụ: SELECT employees.first\_name, employees.last\_name, departments.department\_name FROM employees INNER JOIN departments ON employees.department\_id = departments.id;

1. **WHERE:**

Dùng để lọc dữ liệu dựa trên một điều kiện cụ thể.

Ví dụ: SELECT \* FROM employees WHERE email LIKE '%@example.com';

1. **ORDER BY:**

Dùng để sắp xếp dữ liệu kết quả theo thứ tự tăng dần hoặc giảm dần.

Ví dụ: SELECT \* FROM employees ORDER BY first\_name ASC, last\_name DESC;

**Tình huống**

Để nói chi tiết hơn về SQL, tôi dành hẵn một quyển sách “Advanced SQL” nói về mọi thứ liên quan SQL bao gồm các câu lệnh phổ biến, giải quyết các tình huống phức tạp, data modeling, xây dựng datamart, các chứng chỉ uy tín và việc ôn luyện cho các buổi phỏng vấn

## Câu lệnh SQL mở rộng

## Tự động hóa với quan hệ SQL Excel

# **Nền tảng và ngôn ngữ phân tích và tạo báo cáo**

Hãy bắt đầu với một cụm từ quá quen thuộc với người làm nghề data ở Việt Nam, Power BI. Power BI là một nền tảng hoặc công cụ BI (Business Intelligence) được phát triển bởi Microsoft giúp tạo báo cáo trực quan.

Ngoài Power BI của Microsoft, còn có một số nền tảng phân tích dữ liệu và tạo báo cáo tương tự khác, bao gồm:

1. **Tableau:**

* Là một trong những công cụ BI phổ biến và mạnh mẽ nhất.
* Cung cấp khả năng kết nối với nhiều nguồn dữ liệu, biến đổi dữ liệu và tạo các biểu đồ, báo cáo trực quan.
* Có các tính năng như dashboard, story, data blending và các tính năng nâng cao khác.

1. **Google Data Studio:**

* Là một công cụ BI miễn phí của Google.
* Cho phép kết nối với nhiều nguồn dữ liệu của Google như Analytics, Ads, Sheets.
* Cung cấp các mẫu sẵn và khả năng tạo báo cáo và dashboard tương tác.

1. **Qlik Sense:**

* Là nền tảng BI của Qlik, cung cấp các tính năng phân tích dữ liệu mạnh mẽ.
* Sử dụng kỹ thuật "associative" để phân tích dữ liệu và tạo các visualization.
* Hỗ trợ nhiều nguồn dữ liệu và có các tính năng như dashboard, storytelling.

1. **Looker:**

* Là một nền tảng BI được mua lại bởi Google.
* Cung cấp các tính năng phân tích dữ liệu, tạo báo cáo và dashboard
* Sử dụng LookML để mô hình hóa dữ liệu và tạo các metric, dimension.

1. **Microsoft Excel:**

* Mặc dù không phải một nền tảng BI chuyên dụng, Excel vẫn là một công cụ phổ biến và mạnh mẽ để phân tích dữ liệu.
* Cung cấp các tính năng như bảng tính, biểu đồ, pivot table, Power Query, Power Pivot.

## Câu lệnh Power BI

## Tự động hóa với quan hệ SQL Power BI

# **Ngôn ngữ lập trình**

## **Python**

Các thư viện phổ biến

Xử lý dữ liệu: Pandas,

Visualize: Plot, Stre

Crawl dữ liệu