**TỔNG HỢP KIẾN THỨC**

TỔNG QUAN VỀ KHOA HỌC DỮ LIỆU

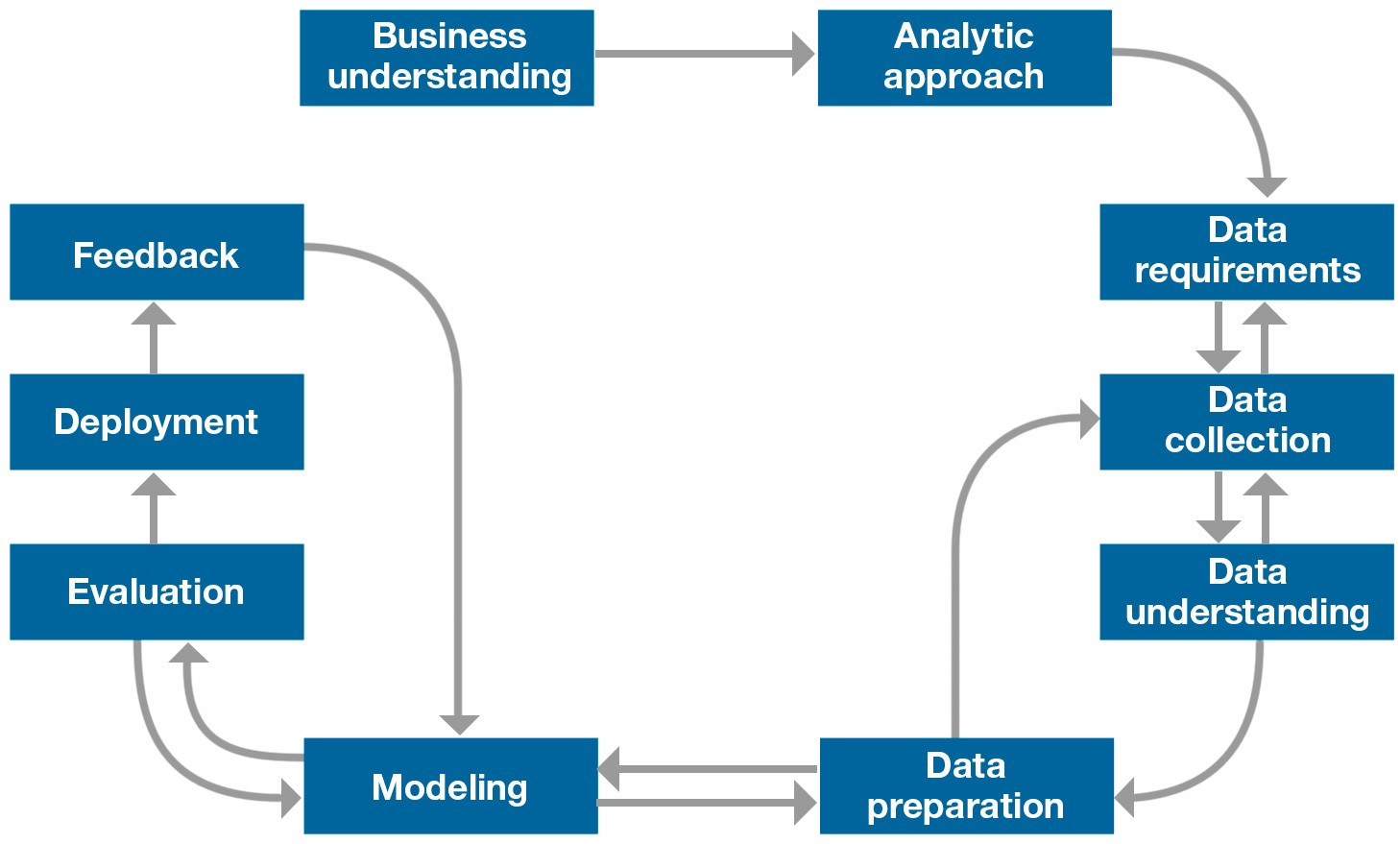
# **1. Định nghĩa về Khoa học Dữ liệu:**

* Sử dụng dữ liệu để hiểu về thế giới.
* Xác nhận mô hình/giả thuyết với dữ liệu.
* Khám phá và hiểu biết ẩn sau dữ liệu.
* Diễn giải dữ liệu thành câu chuyện.
* Trích xuất dữ liệu từ nhiều nguồn khác nhau.
* Nghiên cứu và trả lời câu hỏi từ dữ liệu.

# **2. Một nhà Khoa học Dữ liệu cần có:**

| **Phẩm chất** | **Kỹ năng** |
| --- | --- |
| * Tính tò mò. * Thích tranh luận và đưa ra ý kiến riêng. * Linh hoạt với các nền tảng phân tích. * Khả năng kể chuyện (storytelling). * Xác định lợi thế cạnh tranh. | * Lập trình / tư duy máy tính. * Đại số & Hình học. * Giải tích cơ bản. * Xác suất và thống kê cơ bản. * Cơ sở dữ liệu. |

# **3. Phương pháp luận Khoa học dữ liệu:**



**Vòng đời của dự án Khoa học dữ liệu**

| **Bước** | | **Mô tả** |
| --- | --- | --- |
| **Từ bài toán đến hướng tiếp cận giải quyết** | Business Understanding (Hiểu biết về doanh nghiệp) | Xác định mục tiêu, yêu cầu của doanh nghiệp |
| Analytical Approach (Phương pháp phân tích) | Xác định và áp dụng các phương pháp phân tích dữ liệu |
| **Làm việc với**  **dữ liệu** | Data Requirements  (Yêu cầu về dữ liệu) | Xác định loại dữ liệu cần thiết để phân tích và đáp ứng yêu cầu bài toán. |
| Data Collection  (Thu thập dữ liệu) | Thu thập dữ liệu từ nhiều nguồn khác nhau. |
| Data Understanding  (Hiểu dữ liệu) | Hiểu cấu trúc, tính chất và đặc điểm của dữ liệu. |
| Data Preparation  (Chuẩn bị dữ liệu) | Tiền xử lý dữ liệu, bao gồm xử lý dữ liệu thiếu, loại bỏ nhiễu, và chuyển đổi dữ liệu |
| **Tìm ra giải pháp** | Modeling (Mô hình) | Xây dựng mô hình dự đoán hoặc mô tả dữ liệu. |
| Evaluation (Đánh giá) | Đánh giá hiệu suất của mô hình. |
| Development (Phát triển) | Phát triển các giải pháp hoặc sản phẩm dựa trên kết quả đánh giá mô hình. |
| Feedback (Phản hồi) | Cung cấp phản hồi và điều chỉnh quy trình theo kết quả. |

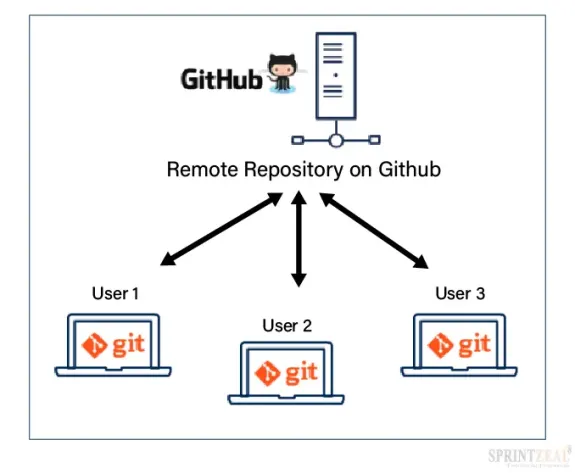
# **4. Cài đặt Anaconda, Jupyter Notebook, Markdown trên Jupyter Notebook:**

[Hướng dẫn cài đặt Anaconda](https://drive.google.com/drive/folders/145Qt_bQ26J53kfG1YpsFf70yXb1K8_XG)

[Hướng dẫn cài đặt Jupyter Notebook](https://drive.google.com/drive/folders/145Qt_bQ26J53kfG1YpsFf70yXb1K8_XG)

[Hướng dẫn sử dụng Markdown](https://drive.google.com/file/d/14sGXVBRSmxhTHfkiOjVj-m0URpiOT4N8/view?usp=sharing)

# **5. Git/Github:**

Git là một phần mềm mã nguồn mở miễn phí, dùng để kiểm soát phiên bản phân tán, cho phép người dùng trên toàn thế giới có thể sao lưu dự án của bạn trên máy tính cá nhân của họ và đồng bộ hóa phiên bản với máy chủ từ xa để chia sẻ. 

| Git lưu trữ dữ liệu dưới dạng các snapshot (ảnh chụp) của mã nguồn tại mỗi điểm thời gian. Thay vì lưu trữ toàn bộ file mới, Git chỉ lưu trữ các phần thay đổi (diff) giữa các phiên bản. Điều này giúp tiết kiệm dung lượng lưu trữ và cho phép truy cập nhanh chóng đến các phiên bản trước đó. |  |
| --- | --- |

| **Khu vực làm việc** | * Working Tree: Nơi chứa các tệp mã nguồn đang được chỉnh sửa. * Staging Area (Index): Nơi lưu trữ các tệp đã được chọn để commit. * Local Repository: Nơi lưu trữ tất cả các snapshot của dự án. |  |
| --- | --- | --- |
| **Quá trình làm việc** | * Sửa đổi: Thay đổi mã nguồn trong Working Tree. * Thêm vào Staging Area: Chọn các tệp đã thay đổi và thêm vào Staging Area bằng lệnh git add. * Commit: Lưu trữ các thay đổi trong Staging Area vào Local Repository bằng lệnh git commit. * Push: Gửi các thay đổi đã commit lên kho lưu trữ remote (trên Github) bằng lệnh git push. |

| **Thuật ngữ** | **Giải thích** |  | **Lệnh** | **Tác dụng** |
| --- | --- | --- | --- | --- |
| Branch | Nhánh, phiên bản cụ thể trong kho lưu trữ tách ra từ project |  | git clone | Sao chép kho lưu trữ Git từ xa |
| Commit | Thời điểm cụ thể trong lịch sử thực hiện code |  | git status | Xem trạng thái các thay đổi trong khu vực làm việc |
| Check out | Chuyển đổi giữa các branch |  | git add | Thêm tệp tin vào khu vực Staging Area |
| Master/main | Nhánh chính trong kho lưu trữ |  | git commit | Lưu lại các thay đổi đã được thêm vào Staging Area |
| Merge | Bổ sung các thay đổi từ nhánh này sang nhánh khác |  | git push | Gửi các thay đổi đã commit lên kho lưu trữ từ xa |
| Pull | Yêu cầu thay đổi cho nhánh chính |  | git branch | Liệt kê các nhánh hiện có |
| Push | Cập nhật các branch từ xa |  | git checkout | Chuyển đổi sang nhánh khác |
| Repository | Kho lưu trữ Git |  | git merge | Hợp nhất các thay đổi từ một nhánh khác vào nhánh hiện tại |

**Xem thêm*:*** [Git/Github](https://docs.google.com/document/d/1c-TrYGojMfVnmSKpRFxEI8Sch-sF2JxiTvDFSyCbSBw/edit)

[Làm quen với Github](https://docs.google.com/document/d/1nQyovHIgTvO50bxp2mSnEXRLSaNjXWRV/edit?usp=drive_link&ouid=105307776553614830510&rtpof=true&sd=true)

[Tạo nhánh và gộp nhánh Github](https://docs.google.com/document/d/1dg0scFSHyhn4l1pijJKiLm5gopjKFhpy/edit?usp=drive_link&ouid=105307776553614830510&rtpof=true&sd=true)