| **Buổi** | **Nội dung** | **Chi tiết** |
| --- | --- | --- |
| **1** | **Giới thiệu về Data Engineer**  **Python** | * Các khái niệm cơ bản về data, data engineer và các nhánh nghề nghiệp về data * Data Engineer roadmap * Các kiến thức cần có để phát triển thành data engineer * Các công cụ và ứng dụng được sử dụng phổ biến * Giới thiệu các công cụ sẽ có trong khóa học: AWS, VSCode, Anaconda, … * Ngôn ngữ lập trình và hệ điều hành * Lý thuyết căn bản về Python * Giới thiệu các tính năng của Python * Python áp dụng trong data science |
| **2** | **Database – SQL basic**  **Các loại data source thường gặp** | * Các loại cấu trúc dữ liệu * Giới thiệu về các khái niệm về SQL, NoSQL * Các loại cơ sở dữ liệu thường gặp * Lý thuyết về một số các câu truy vấn SQL * Các loại data source thường gặp trong thực tế * Nhận biết data source và một số pattern ingest data |
| **3** | **Data types and data modeling** | * Data types, data sources * Data warehouse vs Data Lake vs Data Mart * Data warehouse building concept and design. * SCD và CDC types * Demo một số mô hình DW và architect xây dựng data warehouse trong doanh nghiệp * Xử lý dữ liệu dạng batch – streaming * Data loader và concept ETL ELT * Ứng dụng và architect của ETL ELT trong cung cấp dữ liệu cho data warehouse |
| **4** | **Hands on Session 01** | * Thực hiện cài đặt CSDL AWS * Thực hành trên git (quản lý source code dự án) * Hướng dẫn kết nối CSDL * Thực hành truy vấn, thao tác CSDL bằng SQL * Demo các khái niệm căn bản * Hướng dẫn viết chương trình python * Định hướng nghiên cứu và học một loại ngôn ngữ lập trình hỗ trợ data engineer * Thực hành tạo lambda function * Demo các loại data ingestion (from sql, from api, from sftp, file storage…) |
| **5** | **Big data system** | * Giới thiệu về các loại hệ thống * Giới thiệu về hadoop * Giới thiệu về Hadoop ecosystem và ứng dụng * Advance concept xử lý big data và ứng dụng vào bài toán xử lý data * Giới thiệu một số tools hỗ trợ xử lý data |
| **6** | **Apache spark** | * Giới thiệu về concept xử lý dữ liệu * Spark introduction * Spark advance and components * Demo Spark simple jobs * Ứng dụng SparkSQL trong xử lý dữ liệu * Chuyên sâu về Spark architecture * Thực hanh viết job Spark đọc và ghi dữ liệu vào trong DW * Tối ưu Spark jobs, các lưu ý cần có khi phát triển một job Spark trong dự án thực tế * Spark Streaming |
| **7** | **Hands on session 2 - Apache Spark** | * Thực hành viết crawler dữ liệu bằng python * Thực hành xử lý dữ liệu thông qua Spark job * Giới thiệu và demo về spark streaming (optional) |
| **8** | **Data Engineering - Advance** | * Khái niệm và định hướng data engineering nâng cao * Các thành phần trong data pipelines * Data engineering flow advance * Khái niệm về điều hướng * Khái niệm về orchestration tools * Giới thiệu step function aws |
| **9** | **Hands on Session 3** | * Thực hành công cụ quản lý resource * Thực hành công cụ quản lý source code * Thực hành áp dụng step function trong orchestration tạo luồng pipelines xử lý data trên AWS |
| **10** | **Data open knowledge** | * Giới thiệu về data engineering pipelines trong dự án thực tế * Thiết kế chung hệ thống data warehouse * Hướng dẫn ý tưởng và triển khai một project data engineer * Đặt đề bài và mục tiêu khóa học * Đặt đề bài làm project bài tập lớn * Đưa yêu cầu về tài liệu và form team |
| **11** | **Hands on session 4** | * Standby hỗ trợ build project và fix bug |
| **12** | **Hands on session 5** | * Standby hỗ trợ build project và fix bug |
| **13-14** | **Project and Ending** | * Thực hiện chấm và nhận xét project * Các hướng đi tiếp theo sau khi kết thúc khóa học |