**CDE DATA ENGINEERING CURRICULUM**

* **DATA ENGINEERING FUNDAMENTALS**

Introduction to Data Engineering

* + Data Engineering Lifecycle:
    - Data Collection
    - Data Ingestion
    - Data Storage and Management
    - Data Transformation
    - Data Serving
    - Undercurrents
  + Data Pipelines
  + Data Modeling
  + Programming for Data Engineering
  + Cloud Data Engineering
  + Soft Skills and Best Practices
  + Case Studies
* **LINUX AND GIT** 
  + Basic linux commands
    - Introduction to the command line
    - Creating a directory and moving into directories
    - Listing files in directories
    - Creating files and viewing files
    - Copying files and moving files from one directory to the other
    - Renaming files
    - Absolute and relative paths
    - Seeing a process and killing process
  + Github
    - Creating repo
    - Cloning repo
    - Git add
    - Git commit
    - Git push
    - Git branch
    - Pull request
    - Git Conflict
    - Creating Git Readme and documenting projects
* **SQL** 
  + Databases and Data warehousing
  + Downloading the postgres server locally
  + Basic Queries
  + Introduction to SQL Data Cleaning
  + SQL Data Cleaning
  + Introduction to Window Functions
  + SQL Window Functions
  + Introduction to Advanced SQL (Subquery & CTE)
  + SQL DDL, DML commands
  + Creating Tables/Views (working with tables)
  + Stored Procedures
  + Entity Relationships Diagrams.
* **PYTHON** 
  + Python Basics
    - Control flow
      * Operators
        + Arithmetic operators
        + Assignment operators
        + Comparison operators
        + Logical operators
        + Identity operators
        + Membership operators
      * Logical statements
        + If and else statements
      * Loops
        + For loop
        + While loop
    - Functions
      * Normal functions
      * Generic function
        + Non Default arguments
        + Default arguments
      * \*Args and \*\*kwargs
    - Modules & Packages
      * In-builts modules
      * Custom modules
      * packages
    - Errors and Exceptions
  + Data structures
  + File handling
  + Data manipulation with pandas
  + Database interaction
  + API and web scraping
  + ETL process and Data Pipeline
  + Version control for project
  + Introduction to OOP
* **DATA MODELLING** 
  + Fundamental Concept
  + Basic Technique for Dimension tables
  + Basic Technique for Fact tables
  + Slowly Changing Dimensions
  + One or two case study
* **DBT** 
  + dbt fundamentals and installation
  + Understanding Jinja, Macros, and Testing in dbt
  + dbt packages
  + Introduction to dbt Cloud
* **DOCKER**

Overview of Docker and Internals of Docker

* + Docker File
  + Docker image
  + Docker containers
  + Understanding Docker volumes
  + Docker Networking
  + Introduction to YAML
  + Docker compose and Anchors in docker compose
* **DATA INTEGRATION**

**Airbyte**

* + - Airbyte concept
      * Source
      * Destination
      * Connection
      * Connector
      * Sync
    - Airbyte Architecture
      * Architecture Overview
      * WebApp
      * API server
      * Metadata Database
      * Temporal
      * Worker
    - Running airbyte in Docker
    - Understanding Source configuration
    - Understanding Destination configuration
    - Understanding connection
    - Configure a full synchronization between source and destination
      * S3 to Postgres Database
      * Postgres Database to Redshift
    - How sync works under the hood
* **ORCHESTRATION** 
  + Introduction to Apache Airflow
    - Airflow concept
      * Workflow
      * DAG
      * Task
      * Operators
      * Dependencies
    - Installation and Setup
      * Prerequisites
      * Installation
      * Configuration
    - Airflow Architecture
      * Architecture Overview
      * WebServer
      * Metadata Database
      * Scheduler
      * Worker
      * How Airflow Works
    - Creating Your First DAG
    - Understanding DAG configuration
    - Understanding Task configuration
    - Understanding airflow variables
    - Advanced DAG Concepts
    - Monitoring and Debugging
    - Airflow Configuration and Best Practices
    - Case Studies and Projects
* **CLOUD ( AWS )** 
  + Introduction to the cloud
  + AWS Global infrastructure
    - Regions and AZs
  + IAM
    - User
    - Groups
    - Policy
    - Roles
  + SSM
  + S3
    - Introduction to Data Lake and s3
      * Bucket creation
      * Bucket deletion
      * Object dump
      * Object deletion
      * Bucket versioning
    - Different storage class
    - Introduction to boto3
      * S3
        + List\_bucket
        + List\_object
        + Get\_object
        + Put\_object
    - Introduction to awswrangler
      * S3
        + Read\_csv
        + Read\_parquet
        + To\_csv
        + to\_parquet
  + Glue
    - Introduction to AWS Glue
    - Glue Data Catalog
      * Glue DB
      * Glue Table
  + Athena
    - Introduction to Athena
    - Query result location
    - Workgroup
    - Athena Usage Demonstration
  + RDS ( Relational Database Service )
    - Introduction to RDS
    - Difference between RDS instance and a Database
    - Security Group ( Security Firewall )
      * Inbound and Outbound Traffic
      * Security Group Rule
    - Provision an RDS instance ( Postgres Engine )
  + Redshift
    - What is Amazon Redshift
    - Cluster Overview
    - Redshift Architecture
    - Provision a Redshift Cluster
    - Users, Groups and access management
* **SPARK**
  + Introduction to Spark
  + Installation
  + Spark SQL and DataFrames
  + Spark Internals
  + GroupBy in Spark
  + Joins in Spark
* **TERRAFORM ( Infrastructure as Code )** 
  + What terraform is
  + How terraform works
  + Terraform state file
  + remote state file
  + Basic provisioning
    - Standalone resource provisioning
      * IAM
      * SSM
      * S3
      * Glue
      * RDS
      * Redshift
    - Resource referencing
    - Data source
    - Local usage
  + Modules
    - Module overview
    - Module structure
    - Create a simple resource module
* **CAPSTONE PROJECT**

TBD