FRESNO PACIFIC UNIVERSITY

Continuing Education
1717 S. Chestnut Ave.
Fresno, CA 93702-4709
(800) 372-5505
https://ce.fresno.edu

TEC-62: Data Engineering Program

Professional Education Course Syllabus

Course Description

The Big Data Engineering Program is designed to introduce the technically inclined student to the technologies and methodologies requested by hiring companies and used by real world data engineers. This program is fast paced and will cover a breadth of technologies, including Python programming, Hadoop and cloud-based services in Amazon Web Services (AWS). The student will also be introduced to such methods as data wrangling, munging, ingesting and modeling for analytics. By the end of the program the successful student will be prepared for an entry-level position as a data engineer, Python programmer, or business intelligence developer.

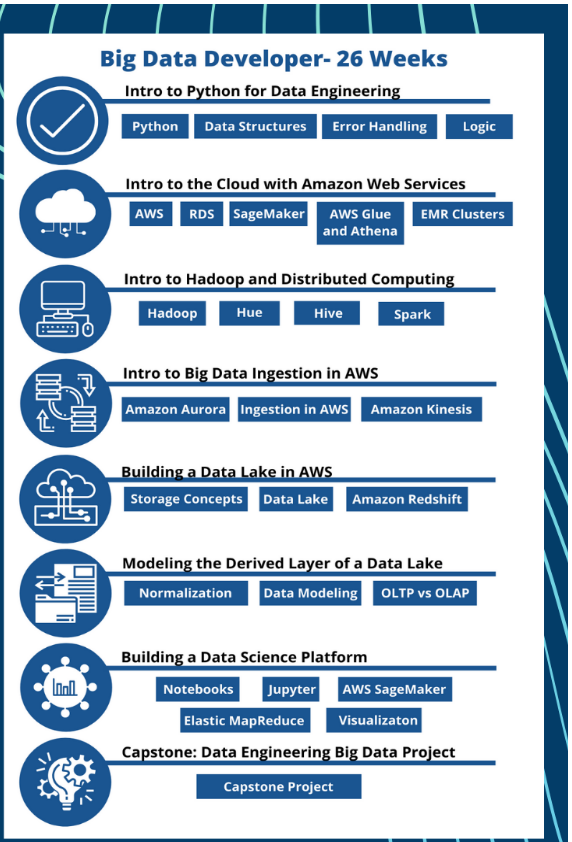
Pre Requisites for Data Engineering:

- Technical aptitude
- Logical thinker
- Comfortable with OS file system concepts
- Experienced downloading and installing software
- Experience working with data connections in Excel or other applications
- Experience creating macros and functions in Excel or other application
- Some exposure to programming is helpful but not required

The Big Data Engineering Program (26 Weeks)*

Promineo Tech's Big Data Engineering Program is designed to introduce the technically inclined student to the technologies and methodologies requested by hiring companies and used by real world data engineers. This program is fast paced and will cover a breadth of technologies, including Python programming, Hadoop and cloud-based services in Amazon Web Services (AWS). The student will also be introduced to such methods as data wrangling, munging, ingesting and modeling for analytics.

By the end of the program the successful student will be prepared for an entry-level position as a data engineer, Python programmer or business intelligence developer.



Big Data Developer- 26 Weeks

Intro to Python for Data Engineering
Python | Data Structures | Error Handling | Logic

Intro to the Cloud with Amazon Web Services
AWS | RDS | SageMaker | AWS Glue and Athena | EMR Clusters

Intro to Hadoop and Distributed Computing
Hadoop | Hue | Hive | Spark

Intro to Big Data Ingestion in AWS
Amazon Aurora | Ingestion in AWS | Amazon Kinesis

Building a Data Lake in AWS
Storage Concepts | Data Lake | Amazon Redshift

Modeling the Derived Layer of a Data Lake
Normalization | Data Modeling | OLTP vs OLAP

Building a Data Science Platform
Notebooks | Jupyter | AWS SageMaker | Elastic MapReduce | Visualizaton

Capstone: Data Engineering Big Data Project
Capstone Project

Career Outlook

According to an article on quanthub.com, "DICE's recent 2020 Tech Job Report reported Data Engineer as the fastest-growing job role in 2019, growing by 50% in 2019. The report also found it takes an average of 46 days to fill data engineering roles and predicted that the time to hire Data Engineers may increase in 2020 "as more companies compete to find talent they need to handle their sprawling data infrastructure.

DICE also noted that Amazon, Accenture and Capital One – all companies with deep pockets to pay high salaries – are hiring Data Engineers at high rates." (DuBois)

How much do data engineers make? "The average Data Engineer salary in the United States is $107,386 as of December 28, 2020, but the salary range typically falls between $89,861 and $125,293. Salary ranges can vary widely depending on many important factors, including education, certifications, additional skills, the number of years you have spent in your profession. With more online, real-time compensation data than any other website, Salary.com helps you determine your exact pay target." (Site built by: Salary.com)

Perquisites

- Technical aptitude
- Logical thinker
- Comfortable with OS file system concepts

- Experienced downloading and installing software
- Experience working with data connections in Excel or other applications
- Experience creating macros and functions in Excel or other application
- Some exposure to programming is helpful but not required

Course Outline

DE1 –AnIntroduction to Python for Data Engineering (4 weeks)

- A brief history of Python
- Python package management and repositories
- Installing the Anaconda distribution of Python
- Installing PyCharm
- Language structure and syntax
- Data Structures in Python
  - Variables
  - Strings
  - List
  - Sets
  - Tuples
  - Dictionaries
  - Other iterables
- Conditional logic
  - For Loop
  - If Else and If, Elif, Else statements
- Comprehensions
- Error handling in Python

DE2 – An Introduction to the Cloud with Amazon Web Services (5 weeks)

- Explain what the cloud is
- The major vendors in the cloud space
- Regions, Availability Zones (AZ) and VPCs
- How to create an AWS account
- How to monitor AWS account billing
- Understand various aspects of AWS technology
- IAM and security using roles and key pairs
- How AWS uses Infrastructure
- How to launch an EC2 instances
- Setup databases in RDS and in LightSail
- Store data and other file objects in an S3 bucket
- How to use AWS Glue and Athena
- How to use SageMaker
- How to spin up an EMR cluster
  - From console
  - From Boto3 using Python

DE3 – An Introduction to Hadoop and Distributed Computing (3 weeks)

- Introduction to Distributed Computing
- Apache Hadoop
  - HDFS
  - MapReduce
  - YARN
- Hadoop distributions
- Simple data ingesting

- Hue
- Hive
- Spark

DE4– An Introduction to Big Data Ingestion in AWS (3 weeks)

- Amazon Aurora in RDS
- Ways to use snapshots to move data
- Common ingestion methods in AWS
- Amazon Kinesis
  - Data Streams
  - Firehose
  - Analytics

DE5 – Building a Data Lake in AWS (2 week)

- A database vs. a data warehouse vs. a data lake
- Data storage concepts
  - Row oriented
  - Columnar
  - JSON
  - File formats
  - Parquet
  - Avro
  - Hive
  - CSV
  - ORC
- AWS S3 Buckets and File Compression
  - GZIP (.gz or .gzip)
  - SNAPPY (.snappy)
  - ZLIB (.zlib)
  - LZO (.lzo)
  - BZIP2 (.bzip or .bz2)

- Preparing an S3 Bucket for a Data Lake Storage
- AWS Lake Foundation
- Amazon Redshift

DE6 – Modeling the Derived Layer of a Data Lake(1 week)

- Why Model Data?
- Types of Data Models
- Entity Relationships
- An introduction to the Relational Data Model
- Normalization and Normal Forms
  - NF1 through NF3
- OLTP vs OLAP
  - The star schema for data warehousing

DE7 – Building a Data Science Platform (4 weeks)

- Notebooks
  - Popular notebooks
- Jupyter
  - Classic vs Lab

    - Data Access
    - Local vs. Server Clusters
    - JupyterHub
- Jupyter Notebooks with AWS SageMaker
- Jupyter Notebooks with AWS EMR – Elastic MapReduce
- Data Visualization
    - Tableau
    - Amazon QuickSight

DE8 – Capstone: Data Engineering Big Data Project (4 weeks)

- Student will use the technologies and methods they learned in the Big Data Engineering Program to build a complete data and analytics platform and will present their platforms to the class with slide deck presentation and live demos.

References:

DuBois, Jen. “Will Demand for Data Engineers Fuel a Talent Shortage in 2020?” QuantHub, 21 Jan. 2021, quanthub.com/data-engineer-demand/#:~:text=DICE's%20recent%202020%20Tech%20Job,growing%20by%2050%25%20in%202019.&text=Data%20Engineer%20job%20growth%20far%20outpaced%20other%20developer%20jobs%20in%202019.

Site built by: Salary.com. “Data Engineer Salary.” Salary.Com, www.salary.com/research/salary/listing/data-engineer-salary. Accessed 21 Jan. 2021.

*In an effort to provide up-to-date quality educational content, the topics, presentations, demos or course length in the outline is subject to change at any time without notice.