DATA ENGINEERING PROSPECTUS

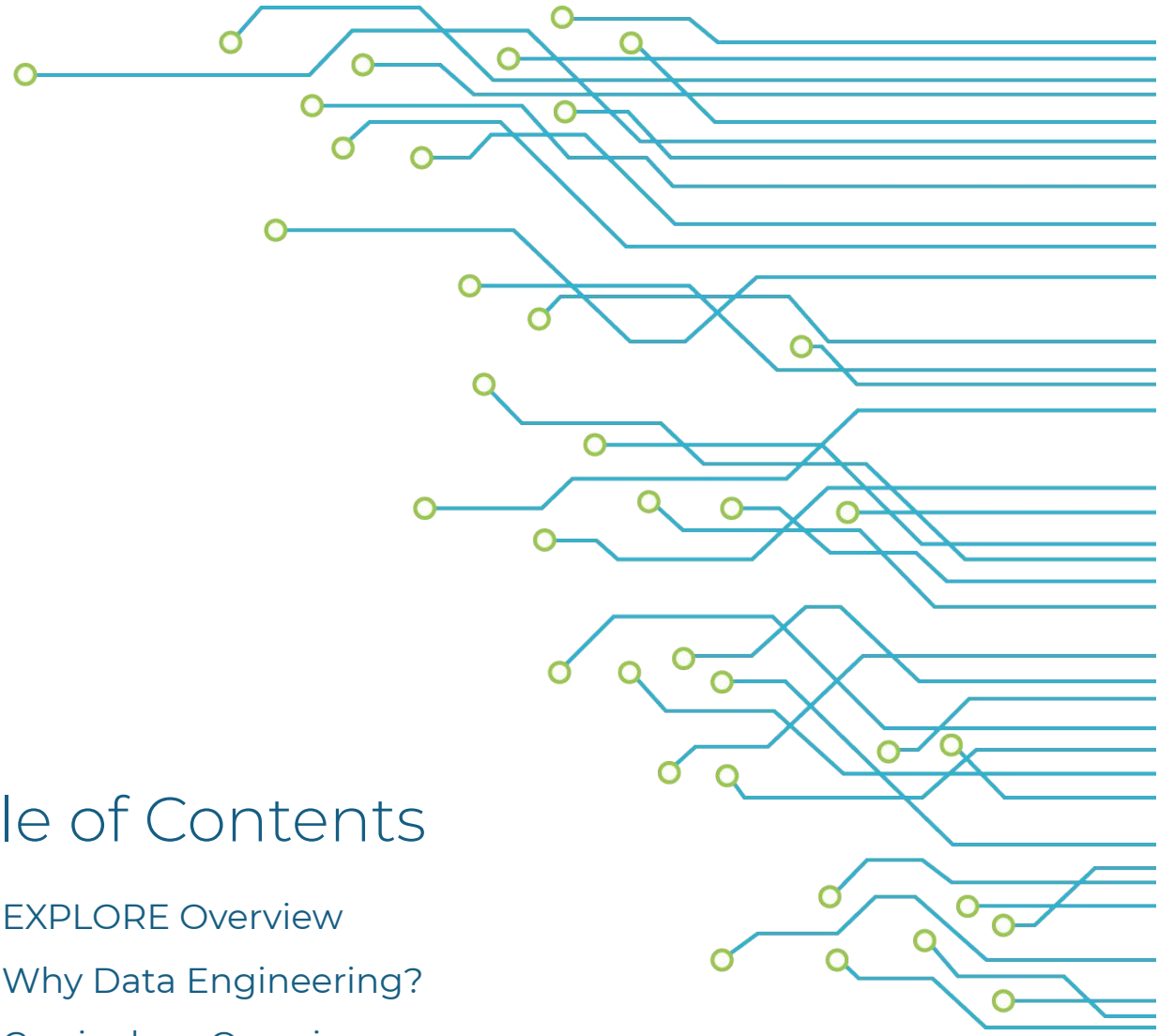


Table of Contents

EXPLORE Overview

EXPLORE is a next generation Learning Institution that teaches students the skills of the future. From Data Science to Data Engineering to Machine Learning to Deep Learning we deliver cutting edge courses to satisfy your hunger to learn. Our Programmes are built by an amazing Faculty - we employ some of the world's most talented Scientists who have experience solving difficult problems on a global stage.

Our philosophy is to teach our students how to solve problems in the real world. We emphasise team-work, collaboration and working within constraints, under pressure, with deadlines while understanding context, audience and implementation challenges. We are not a theoretical institution (although we cover the theory) - we are a 'practical, hands-on, roll-up-your-sleeves and get stuff done' kind of institution. As real-world Scientists who have delivered impact in the world of work we're well positioned to deliver these skills.

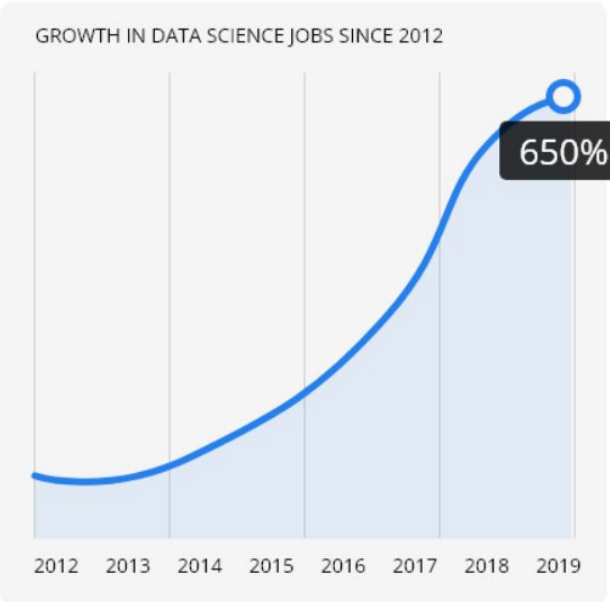
EXPLORE launched during 2013 and since then has taught 1,000's of students and solved many problems for businesses across multiple Industries across the world. We're reinventing education and invite you to join us to change things for the better.

Why Data Engineering?

Four megatrends are fundamentally changing the shape of our world:

1. Vast amounts of data are being generated every minute.
2. The processing speed of our machines is increasing exponentially.
3. We now have cloud providers who can store insane amounts of data for a few dollars.
4. Powerful open source algorithms that can read, write, translate and see are now available to everyone.

Data Engineering is the skillset used to harness the power of these tectonic shifts in our world. Over the last 4 years, the rise of job opportunities for data engineers has become almost level to those traditionally seen for software engineering positions.



EXPLORE

Curriculum Overview

This course will provide students with the knowledge, skills and experience to get a job as a data engineer - which requires a mix of programming, cloud computing, big data knowledge and the ability to apply these skills in new and challenging domains. This course will teach you how to create automated pipelines for a business.

Phase	*Module*	*Weeks*
Fundamentals	SQL Basics	3 weeks
	Python Programming	4 weeks
	Data Modelling	2 weeks
	Cloud Computing	3 weeks
Data Architecture	Big Data	4 weeks
	Data Warehouses and Lakes	4 weeks
	Data Automation	4 weeks

EXPLORE

Module 1

SQL Basics

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Project: Set-up and extract valuable information from the TMDB movie database

What is covered in Module 1:

Querying Data

SELECT statements

- ➔ Retrieving records from a database using SELECT
- ➔ Specifying conditions for your queries using FROM
- ➔ GROUP BY clause
- ➔ ORDER BY clause

JOIN datasets

- ➔ Python data types
- ➔ Lists, tuples, sets and dictionaries
- ➔ Using Numpy and Pandas libraries
- ➔ Store, retrieve and transform data

Changing Data

Create new data

- ➔ Create new databases and tables
- ➔ Create entries using SQL
- ➔ Create entries via CSV upload

Changing existing data

- ➔ Change existing data using SQL
- ➔ Change existing data using CSV upload

Module 2

Python Programming

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Project: Build a financial calculator

What is covered in Module 2:

Python Fundamentals

Python programming basics

- ➔ Installing Jupyter environment
- ➔ Pseudo code and debugging concepts
- ➔ Interactive vs scripting mode
- ➔ Working with primitive data types - variables, strings, integers, floating points, booleans

Logic and functions

- ➔ Conditional statements - IF and ELSE IF
- ➔ Working with lists
- ➔ For loops and while loops
- ➔ Break | Continue principles
- ➔ Creating and working with functions

Python Data Structures

Data types

- ➔ Working with Strings, Numbers, Booleans
- ➔ Lists and Tuples Semantics
- ➔ Working with Comparisons
- ➔ Working with Statements

Dataframes and using libraries

- ➔ Sets and Dictionaries Semantics
- ➔ Working with Comparisons
- ➔ Importing Data - using Numpy and Pandas libraries
- ➔ Working with Data Frames

Module 3

Data Modelling

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Project: Build a relational and NoSQL database

What is covered in Module 3:

Relational Data Modelling

Data Modelling Basics

- ➔ Purpose of data modelling
- ➔ Overview of data storage techniques and various databases
- ➔ Pros and cons of data storage techniques and databases

Relational databases

- ➔ When to use relational databases
- ➔ 1st, 2nd and 3rd normalization
- ➔ Creating normalized tables using primary and foreign keys
- ➔ Implement denormalized schemas

NoSQL Data Models

NoSQL databases

- ➔ Creating NoSQL databases
- ➔ Selecting appropriate primary keys

Relational vs NoSQL databases

- ➔ Differences of NoSQL vs relational databases
- ➔ When to use which

EXPLORE

Module 4

Cloud Computing

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Project: Create a pipeline to utilise data stored in S3 bucket

What is covered in Module 4:

Introduction to the Cloud

Cloud computing basics

- ➔ Intro to cloud computing
- ➔ Pros and cons of cloud computing
- ➔ Cloud providers

Intro to AWS

- ➔ Intro to AWS
- ➔ Creating an AWS account
- ➔ Overview of AWS services

AWS Services

Storage, Databases and Compute services

- ➔ Set up IAM policies
- ➔ Storing data in S3 buckets
- ➔ Spin up an EC2 instance
- ➔ RDS instances

Other services

- ➔ Security, Identity and compliance
- ➔ Networking and content delivery
- ➔ Using Lambdas

Module 5

Big Data

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Project: Create a cluster and optimise queries to extract results efficiently

What is covered in Module 5:

Big Data Concepts

Big data architecture

- The 4 V's of big data
- History of big data and its evolution
- Practical examples of using big data
- Concept of distributed data across servers

Hadoop vs Spark

- Overview of the Hadoop framework (HDFS, YARN, MapReduce, Hive)
- The workings of MapReduce
- Spark framework and differences to Hadoop
- The workings of Resilient Distributed Datasets (RDDs)

Big Data Software

HIVE and Spark SQL

- Overview of Hive and how it works
- Query data on Data Analytics Studio via Hive
- Alternatives to Hive

Optimize queries

- Partitioning and bucketing Hive tables
- Using appropriate file formats in Hive
- The use of indexing
- Avoiding / replacing certain clauses

EXPLORE

Module 6

Data Warehouses and Lakes

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Project: Build ELT pipelines for a data warehouse and a data lake

What is covered in Module 6:

Data Warehouses

Intro to data warehouses

- ➔ Data warehousing architecture overview
- ➔ Run an ETL process to denormalize a database
- ➔ Create an OLAP cube from facts and dimensions
- ➔ Compare columnar vs. row oriented approaches

Implementing data warehouses on AWS

- ➔ Extract data from S3 into Redshift using ELT process
- ➔ Set up AWS infrastructure using IaC
- ➔ Design an optimized table by selecting the appropriate distribution style and sorting key

Data Lakes

Intro to data lakes

- ➔ Purpose of data lakes
- ➔ Pros and cons of data lakes
- ➔ Data lakes vs data warehouses

Implementing data lakes on AWS

- ➔ Implement data lakes on Amazon S3, EMR, Athena, and
- ➔ Amazon Glue
- ➔ Use Spark to run ELT processes and analytics on data of
- ➔ diverse sources, structures, and vintages

EXPLORE

Module 7

Data Automation

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Project: Create a cluster and optimise queries to extract results efficiently

What is covered in Module 7:

Data Pipelines and Quality

Data Pipelines

- ➔ Creating data pipelines with Apache Airflow
- ➔ Setting up task dependencies
- ➔ Creating data connections using hooks

Data Quality

- ➔ Track data lineage
- ➔ Setting up data pipeline schedules
- ➔ Partition data to optimize pipelines
- ➔ Data quality checks

Production Data Pipelines

Build pipelines

- ➔ Build reusable and maintainable pipelines
- ➔ Build your own Apache Airflow plugins

Task boundaries and monitoring

- ➔ Implement subDAGs
- ➔ Setting up task boundaries
- ➔ Monitoring data pipelines

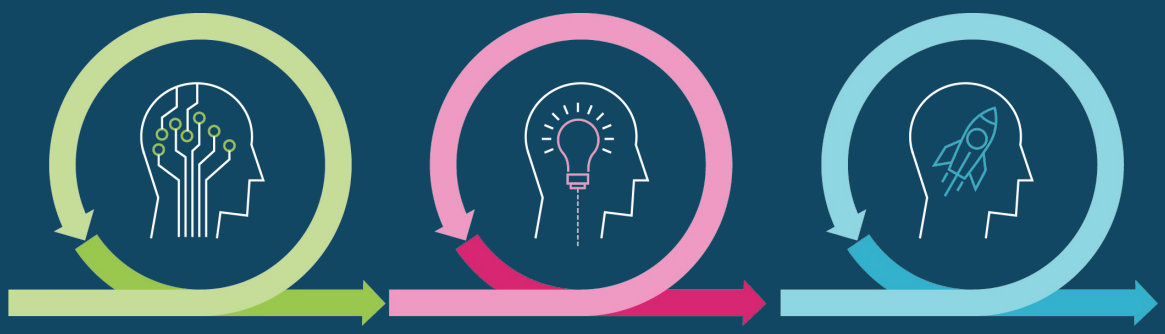
EXPLORE

EXPLORE Philosophy:
Solving problems in the real world

At EXPLORE we focus on building our student's ability to solve problems in the real world. Building things that work and make a difference is hard - that's what we teach.

We're not a traditional learning institution that spends weeks teaching matrix multiplication on a whiteboard (although understanding that is useful) - we're a practical, solution-orientated institution that teaches our students to work in teams, under pressure, with deadlines while understanding context, constraints and the audience.

Our courses are typically broken into Sprints where we teach a core set of concepts within the framework of solving a problem in a team with a tight deadline.



Students cycle from Sprint to Sprint solving different problems in different teams as they build this core muscle over the course.

EXPLORE

Contact Information

For any **admissions related enquiries** mail us on:
admission@explore-ai.net

For any **general enquiries** mail us on:
general@explore-ai.net

Visit: www.explore-datascience.net