Python (4 days)

Day 1

- Installing and setting up python
- Writing your very first program in python
  - Printing Hello World
- Operators and Expressions
- Slicing
  - Negative slicing
  - Using step in slicing
  - Slicing backwards
- Strings
  - String operators
  - String formatting
- Program Flow control in Python
  - if statement
  - elif
  - for loop
  - continue and break
  - while loop

Day 2

- List and Tuples
  - mutable vs Immutable objects
  - List
  - Sorting a list
  - Removing items from list
  - Replacing items in list
  - What are tuples
  - Performing basic functions to a tuple
- Dictionary and Sets
- Functions
  - Defining a function
  - Parameters and arguments
  - Returning values
  - Docstring
  - *args

Day 3

- Input and Output in python
  - Reading and writing to a text file
  - Appending to a file
  - Object persistence using shelve
- Exception handling in python
- Generators, Decorators and lambda expression

Day 4

• Introduction to external libraries in Python
• Deep dive into libraries:
  ○ NumPy, Pandas and Matplotlib

Assessment

SQL (4 Days)

Fundamentals of SQL

Day 1

• Introduction to SQL
  ○ Introduction
  ○ Work with Schemas
  ○ Explore the structure of SQL Statements DDL, DML, DCL
  ○ Examine the SELECT statements
  ○ Work with data types
  ○ Handle NULLs

Hands-on: Work with SELECT statements

Day 2

• Sort and filter results in SQL
  ○ Sort your results
  ○ Limit the sorted results
  ○ Page results
  ○ Remove duplicates
  ○ Filter data with predicates
• Combine multiple tables with JOINs in SQL
  ○ Understand joins concepts and syntax
  ○ Use Inner joins
  ○ Use Outer joins
  ○ Use Cross joins
  ○ Use Self joins
• Write Subqueries in SQL
  ○ Understand Subqueries
  ○ Use scalar or multi-valued subqueries
  ○ Use self-contained or correlated subqueries

Hands-on: Sort and filter query results Hands-on: Query multiple tables with joins
Hands-on: Use Subqueries

Day 3

• Use built-in functions and GROUP BY in SQL
• Categorize built-in functions
  ▪ Use aggregate functions - AVG SUM MIN MAX COUNT
  ▪ Use Mathematical functions - ABS, COS/SIN, ROUND RAND

- Use Ranking functions - RANK, DENDE-RANK
- Use Analytical function - LAG, LAST_VALUE, LEAD, PERCENTILE_CONT, PERCENTILE_DISC, PERCENT_RANK
- Use Logical functions - CHOOSE, GREATEST, LEAST
  - Summarize data with GROUP BY
  - Filter groups with HAVING
- Modify data with SQL
  - Insert data
  - Generate automatic values
  - Update data
  - Delete data
  - Merge data based on multiple tables

Hands-on: Use built-in functions
Hands-on: Modify data

Day 4

- Triggers
- Stored Procedure
  - Stored procedures
  - Create
  - Modify
  - Delete
  - Execute
  - Specify parameters
- Indexes
  - Heaps (Tables without Clustered Indexes)
  - Clustered & Non-Clustered Indexes

Hands-on: Stored procedure
Hands-on: Indexes
Assessment

PySpark (4 Days)

Pyspark

Day 1

- Fundamentals of PySpark
  - A Brief Primer on PySpark
  - Brief Introduction to Spark
  - Apache Spark Stack
  - Spark Execution Process
  - Newest Capabilities of PySpark
  - Cloning GitHub Repository
- Resilient Distributed Datasets
  - Resilient Distributed Datasets
  - Creating RDDs
  - Schema of an RDD

- Understanding Lazy Execution
- Introducing Transformations – .map(…)
- Introducing Transformations – .filter(…)
- Introducing Transformations – .flatMap(…)
- Introducing Transformations –. distinct(…)
- Introducing Transformations – .sample(…)
- Introducing Transformations – .join(…)
- Introducing Transformations – .repartition(…)
- **Project 1:** Count Data Project (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying map, filter, faltmap, distinct, join and repartition)
- **Project 2:** Weather Temperature Crunch (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying map, filter, faltmap, distinct, join and repartition on instream data)

Day 2

• Resilient Distributed Datasets and Actions
- Introducing Actions – .collect(…)
- Introducing Actions – .reduce(…) and .reduceByKey(…)
- Introducing Actions – .count()
- Introducing Actions – .foreach(…)
- Introducing Actions – .aggregate(…) and .aggregateByKey(…)
- Introducing Actions – .coalesce(…)
- Introducing Actions – .combineByKey(…)
- Introducing Actions – .histogram(…)
- Introducing Actions – .sortBy(…)
- Introducing Actions – Saving Data
- Introducing Actions – Descriptive Statistics
- **Project 3**: 10 Tasks in Students/Professor University Datasets (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying RDD actions.)
- **Project 4**: 8 Tasks in Customer Data Datasets (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying RDD actions through specified applicability)
- **Project 5**: Movie ratings

• DataFrames and Transformations
- Creating DataFrames
- Specifying Schema of a DataFrame
- Interacting with DataFrames
- The .agg(…) Transformation
- The .sql(…) Transformation
- Creating Temporary Tables
- Joining Two DataFrames
- Performing Statistical Transformations
- The .distinct(…) Transformation
- **Project 6:** CompanyMegaData (doing all the transformation logics,

columnal logic and aggregation and exploratory data analysis)
- **Project 7**: University Data (end to end pyspark execution of insight delivery on University Data)

Day 3

• Collaborative Filtering and Techniques
- Collaborative filtering
- Utility Matrix
- Explicit and Implict Rating
- Expected Results
- Dataset
- Joining Dataframe
- Train and Test Data
- ALS model
- Optimization Hyperparameter tuning and cross validation
- Best model and evaluate prediction
- **Project 8**: IMDB Rating project (Optimization logics focused on the project with extensive pyspark logic and clever techniques of manipulation)

• Spark Streaming
- Introduction to spark streaming
- Spark streaming with RDD
- Spark streaming Context
- Spark streaming Reading Data
- Spark streaming Cluster Restart
- Spark streaming RDD Transformation
- Spark streaming DF and Display
- Spark streaming DF Aggregation
- **Project 9**: Streaming Crunch Dataset(orchestration of a stream pipeline project of end to end execution of the ingestion of live data)

Day 4

• Spark ETL and Captone project
- Introduction to ETL
- ETL Pipeline
- Dataset
- Preprocessing, extraction, transformation
- Loading Data and cleaning
- RDS Networking
- Downloading PostGres
- Configuration and execution

Project 10: Completion of Captone Project (Full end to end project Streaming Crunch Dataset of entire pyspark concepts from data exploratory to applying techniques and finding out the logics to the requirement of the dataset along with applying multiple ways to solve a solution and figuring out the correct and most optimized way and efficient way)