# Day 1 - Python Basics

1. Python Overview
2. Features of Python
3. Python Environment Set-up and Installation
4. Jupyter Notebook Overview
5. Python Basics
   1. Basic Data Types
   2. Data structures
      1. List
      2. Tuple
      3. Dictionary
      4. Sets and sets operations
      5. List comprehension
   3. If, elif and else Statements
   4. For and While Loops
6. Python Built in functions
7. Standard libraries
8. Functions and Lambda Expressions
9. File handling
   1. Syntax
   2. How to Open File
   3. Read Lines
   4. Write to an Existing File
   5. Create a New File
   6. Delete a File
10. Using Pandas Package in Python
    1. Introduction
    2. Installation of Python
    3. Series
    4. DataFrames
    5. Missing Data
    6. Groupby
    7. Merging Joining and Concatenating
    8. Operations
    9. Data Input and Output
11. Logging in Python
    1. Why Use the logging Module
    2. Creating a Simple Logger
    3. Logging Exceptions
    4. Logging levels

# Day 2 - PySpark

1. What is Apache Spark
2. Apache Spark and Scala
3. Running Spark
   1. Downloading Spark Locally
   2. Launching Spark’s Interactive Consoles
4. An Introduction to Apache Spark
   1. Spark’s Basic Architecture
   2. Spark’s Language APIs
   3. Starting Spark
   4. The SparkSession
   5. DataFrames
      1. Partitions
   6. Transformations
      1. Lazy Evaluation
   7. Actions
   8. Spark UI
   9. An End-to-End Example
5. Deployment modes of spark
6. Structured API
   1. DataFrames and Datasets
   2. Schemas
   3. Overview of Structured Spark Types
      1. DataFrames Versus Datasets
      2. Columns
      3. Rows
      4. Spark Types
   4. Overview of Structured API Execution
      1. Logical Planning
      2. Physical Planning
      3. Execution
7. Persistence And Caching Mechanism in RDD
8. Understand Partitioning
9. Transformations and actions
   1. Lazy Evaluation
   2. Fault tolerance
10. Basic Structured Operations
    1. Schemas
    2. Columns and Expressions
    3. Records and Rows
    4. DataFrame Transformations

# Day 3 - PySpark

1. Introducing Apache Parquet file format
   1. What is Apache Parquet?
   2. Parquet Format vs. CSV
   3. Advantages of Parquet Columnar Storage
   4. Primitive data types in Parquet format
   5. Apache Parquet Spark Example
      1. Spark Write DataFrame to Parquet file format
      2. Spark Read Parquet file into DataFrame
      3. Append to existing Parquet file
      4. Using SQL queries on Parquet
      5. Spark parquet partition – Improving performance
      6. Spark Read a specific Parquet partition
2. Aggregations
   1. Aggregation Functions
      1. count
      2. min and max
      3. sum
      4. avg
   2. Grouping
   3. Window Functions

<https://towardsdatascience.com/spark-3-2-session-windowing-feature-for-streaming-data-e404d92e267>

* 1. User-Defined Aggregation Functions

1. Joins
   1. Join Expressions
   2. Join Types
   3. Inner Joins
   4. Outer Joins
   5. Left Outer Joins
   6. Right Outer Joins
2. Data Sources
   1. The Structure of the Data Sources API
   2. CSV Files
   3. JSON Files
   4. Advanced I/O Concepts
      1. Splittable File Types and Compression
      2. Reading Data in Parallel
      3. Writing Data in Parallel
      4. Managing File Size

# Day 4 - PySpark

1. Spark SQL
   1. What Is SQL?
   2. How to Run Spark SQL Queries?
   3. Catalog
   4. Tables
   5. Views
   6. Select Statements
   7. Databases
2. PySpark UDF Introduction
   1. What is UDF?
   2. Why do we need it?
3. Create PySpark UDF (User Defined Function)
   1. Create a DataFrame
   2. Create a Python function
   3. Convert python function to UDF
4. Using UDF with DataFrame
   1. Using UDF with DataFrame select()
   2. Using UDF with DataFrame withColumn()
   3. Registring UDF & Using it on SQL query
5. Resilient Distributed Datasets (RDDs)
   1. What Are the Low-Level APIs?
   2. About RDDs
      1. Types of RDDs
      2. When to Use RDDs?
   3. Creating RDDs
   4. Manipulating RDDs
6. How to read DAGs
7. Execution Plan
   1. Physical Execution Plan
   2. Logical Execution Plan
8. Catalyst Optimizer
9. How partitioning works
10. Driver vs Worker Nodes

# Day 5 - Advanced PySpark and Performance

1. Distributed Shared Variables
   1. Broadcast Variables
   2. Accumulators
      1. Basic Example
      2. Custom Accumulators
2. Developing Spark Applications
   1. Writing Spark Applications
   2. The Development Process
   3. Launching Applications
   4. Configuring Applications
3. Performance Tuning
   1. Indirect Performance Enhancements
   2. Direct Performance Enhancements