<u>Pre-requisites and Important Notes:</u>

- Participants should have prior working experience on Airflow.
- Scope of this training is limited to the topics which are mentioned in this TOC. In case any other related or subtopic to be covered during the training, that needs to be explicitly mentioned in this TOC.
- All the participants need to pre-reed the study material and PPTs before coming to the session for next day.
- Every participant needs to invest at least 2 hours for offline study as well. However, 4 hours is recommended.
- This training will be 70% to 80% on Demonstration by the trainer.

Day 1 (8 Hours)

- Module 1: Airflow Fundamentals
    - Introduction to Airflow and its Architecture
    - Setting Up Airflow Environment and Basics
    - How does Airflow work?
    - How to setup Airflow on Google Cloud
    - Overview of Airflow UI
    - Overview of Airflow CLI
- Module 2: Advanced DAG Configuration Dynamic Task Generation
    - Dynamic task concepts
    - Mapping over the result of another operator
    - Mapping over multiple parameters
    - Repeated mapping
    - Task Groups and Dependencies
    - Mapping over task groups
    - Transform outputs with .map
    - Enable support for deferrable operators
    - Using Deferrable Operators

- Writing Deferrable Operators
- Triggering Deferral
- Writing Triggers
- View triggerer logs
- Monitor triggerer
- XCom
    - What is an Airflow XCom ?
    - How to use XCom in Airflow
    - How to Push a Value to Airflow XComs
    - How to Get the XCom Value through Airflow
    - XCom limitations

Day 2 (8 Hours)

- Module 3: Performance Optimization Managing Task Pools for Resources
    - Introduction
    - Use of Pools
    - Implementation Approach
    - Resource Management
    - Throttling
    - Prioritization
    - Preventing Resource Over-utilization
    - Load Balancing
    - Using multiple pool slots
    - Provision of Pools through DAG
    - Memory and Compute Intense Jobs
    - Scaling for 1000+ DAGs
        - Introduction
        - ARCHITECTURE
        - Airflow Settings
        - Instance Choice
        - Fine tuning the cluster configuration and parallalism
- Module 4: Advanced Configuration
    - Troubleshooting
        - Cleaning up Root Partition Space by Removing the Task Logs
        - Using macros with Airflow
        - When a DAG has X number of tasks but it has only Y number of running tasks
        - Tasks for a specific DAG get stuck
        - Which logs do I look up for Airflow cluster startup issues?

- Where can I find Airflow Services logs?
- What is $AIRFLOW_HOME?
- Where can I find Airflow Configuration files?
- Where can I find Airflow DAGs?
- Where can I find Airflow task logs?
- How do I restart Airflow Services?
- How do I delete a DAG?
- Tasks are Slow to Schedule or Aren’t Being Scheduled
- Task Logs are Missing or Fail to Display
- You receive an “unrecognized arguments” error
- Tasks Run Slow or Fail
- Tasks are Bottlenecked

Day 3 (8 Hours)

  - DAG Processor and Scheduler Service
  - Configuring DAG Parse Time
  - Custom XCom Backends
    - What is an Airflow XCom ?
    - How to use XCom in Airflow
    - Define a custom XCom class
    - Create and run your DAG to generate XComs
    - Create a custom serialization method to handle Pandas dataframes
    - Run a DAG passing Pandas dataframes via XCom
- Module 5: DAG Concurrency and Parallelism Managing DAG/Task Concurrency Levels
  - Parallelism in Workers
    - Installation level
    - Multiple schedulers
    - DAG level
    - Task level
  - Task Count Decisions
    - Pool
    - max_active_tis_per_dag
- Module: Mastering DAGs
  - Start_date and schedule_interval parameters
  - Manipulating the start_date with schedule_interval
  - Backfill and Catchup
  - Catching up non triggered DAGRuns
  - Dealing with timezones in Airflow
  - Making DAGs timezone aware

- How to make tasks dependent?
- Creating task dependencies between DagRuns
- How to structure a DAG folder?
- Organizing DAGs folder
- How does the Web Server work?
- How to deal with failures in DAGs?
- Retry and Alerting
- How to test DAGs?
- Unit testing DAGs

Day 4 (8 Hours)

- Module 6: Troubleshooting and Debugging Airflow and GKE Logs
    - Managed Composer (GCP) Mechanics
    - Ad Hoc Queries with the metadata database
    - Issue Identification
    - Identify the Failed Tasks
    - Analyze the Error Messages
    - Inspecting DAG Processor logs
    - Inspecting DAG parse times
    - Monitoring running and queued tasks
    - Troubleshooting issues at DAG parse time
    - Troubleshooting issues with running and queued tasks
    - Avoid task scheduling during maintenance windows
    - Make the Airflow scheduler ignore unnecessary files
    - Airflow scheduler processes paused DAGs
    - Usage of 'wait_for_downstream' in your DAGs
    - Symptoms of Airflow Database being under load pressure
    - Workarounds for issues encountered during backfilling DAGs
    - Restart Failed Tasks
    - Pools and priority_weights: Limiting parallelism - prioritizing tasks
- Module 7: Practical Case Studies and Best Practices
    - Real-world Use Cases
    - Hands-on Workshop with Complex DAGs: The Forex Data Pipeline
        - Introduction
        - Defining first DAG
        - Checking if the API is available - HttpSensor

- Checking if the currency file is available - FileSensor
- Downloading the forex rates from the API - PythonOperator
- Saving the forex rates in the HDFS - BashOperator
- Creating the Hive table forex_rates - HiveOperator
- Processing the forex rates with Spark - SparkSubmitOperator
- Sending an email notification - EmailOperator
- Sending a Slack notification - SlackAPIPostOperator
- Operator Relationships and Bitshift Composition
- Adding dependencies between tasks
- The Forex Data Pipeline in action!

Day 5 (8 Hours)

  - Scaling Strategies
    - Autoscaling environments
    - Scale and performance parameters
    - Multiple schedulers
  - Database disk space
- Additional Topics (Optional) Executors and Sensors
  - Monitoring and Alerts Integration
    - Introduction
    - How does the logging system work in Airflow?
    - Setting up custom logging
    - Storing logs in AWS
    - Elasticsearch Overview
    - Configuring Airflow with Elasticsearch
    - Monitoring your DAGs with Elasticsearch
    - Introduction to metrics
    - Monitoring Airflow with TIG stack
    - Triggering alerts for Airflow with Grafana
    - Airflow maintenance DAGs
  - Extending Airflow with Custom Operators
    - Hooks
    - User interface
    - Templating
    - Sensors