StarAgile

Phase wise Training Plan
HI-Labs

Table of Contents

Duration: 80 hours

TOC FOR PHASE 1 (FOUNDATION BUILDING) – 28 HOURS

Linux Essentials (5 Hours)

- **Introduction to Linux (1 Hour)**
  - What is Linux?
  - Linux distributions and their use cases
  - Why Linux is important for data professionals
- **Basic Linux Commands (1 Hours)**
  - File navigation (ls, cd, pwd)
  - File and directory management (mkdir, touch, rm, cp, mv)
  - Viewing and editing files (cat, less, nano, vi)
- **Linux File Permissions (1 Hour)**
  - Understanding users, groups, and others
  - File permission types (r, w, x)
  - Changing permissions with chmod and chown
- **Process Management (1 Hour)**
  - Viewing processes (ps, top, htop)
  - Managing processes (kill, jobs, bg, fg)
- **Networking in Linux (1 Hour)**
  - Checking connectivity (ping, curl, wget)
  - Managing network interfaces (ifconfig, ip, netstat)
- **Assessment for Linux**
  - Quiz: 15-20 questions on Linux commands, processes, and permissions.
  - Practical Task: Create directories, change permissions, and write a simple script for automating tasks.

SQL Basics (7 Hours)

- **Introduction to Databases (0.5 Hour)**
  - What is a database?
  - Difference between RDBMS and NoSQL
- **SQL Basics (1.5 Hours)**
  - SELECT statement
  - Filtering with WHERE
  - Sorting with ORDER BY
- **Working with Tables (1 Hours)**
  - Creating and dropping tables
  - INSERT, UPDATE, DELETE operations
  - Constraints (PRIMARY KEY, FOREIGN KEY)
- **Joins and Relationships (2 Hours)**
  - INNER JOIN, LEFT JOIN, RIGHT JOIN
  - Self Joins
  - UNION and INTERSECT
- **Functions in SQL (1 Hour)**
  - Aggregate functions (SUM, COUNT, AVG)
  - String functions (CONCAT, SUBSTRING)
  - Date functions
- **Basic Subqueries (1 Hour)**
  - Writing nested queries
  - Correlated subqueries
- **Assessment for SQL**
  - Quiz: 20-25 questions on SQL syntax and concepts.
  - Practical Task: Write queries to fetch data, perform joins, and manipulate tables.

Python Basics (11 Hours)

- **Introduction to Python (1 Hour)**
  - Python installation and setup
  - Python IDEs (Jupyter, PyCharm, VS Code)
- **Python Syntax and Basics (2 Hours)**
  - Variables, data types, and operators
  - Input/output in Python
  - Python comments and indentation
- **Control Flow (1 Hours)**
  - Conditional statements (if, elif, else)
  - Loops (for, while)
  - Loop control statements (break, continue, pass)
- **Functions and Modules (2 Hours)**
  - Defining and calling functions
  - Understanding return values
  - Importing and using modules
- **Working with Data Structures (2 Hours)**
  - Lists, tuples, and sets
  - Dictionaries and their operations
- **File Handling (2 Hours)**
  - Reading and writing text files
  - Handling exceptions during file operations
- **Debugging and Best Practices (1 Hours)**
  - Debugging with print and debugging tools
  - Writing clean and readable Python code
- **Assessment for Python**
  - Quiz: 20-25 questions on syntax and coding logic.
  - Practical Task: Solve coding problems on loops, functions, and file handling.

Case Study and Wrap-Up (2 Hours)

- **Comprehensive Case Study**
    - Scenario: End-to-end project involving Linux, Excel, SQL, and Python.
    - Retrieve a dataset from a Linux server. (This bullet point seems incomplete. Consider adding more detail about the case study's objectives and tasks.)
- **Final Assessment**
    - Quiz: 30-40 questions covering all topics from Phase 1.
    - Practical Task: Complete a multi-step task using all skills learned.

TOC FOR PHASE 2 (INTERMEDIATE LEVEL)

Advanced SQL (8 Hours)

- **Advanced Joins and Set Operations (2 Hours)**
    - CROSS JOIN and its applications
    - FULL OUTER JOIN
    - UNION, UNION ALL, INTERSECT, and EXCEPT
- **Window Functions (2 Hours)**
    - Understanding OVER and PARTITION BY
    - ROW_NUMBER, RANK, and DENSE_RANK
    - Aggregate functions with windowing
- **Advanced Subqueries (1 Hour)**
    - Correlated subqueries in-depth
    - Subqueries in FROM and SELECT clauses
- **CTEs and Recursive Queries (1 Hour)**
    - Writing Common Table Expressions (CTEs)
    - Recursive CTEs for hierarchical data
- **SQL Optimization and Performance Tuning (2 Hours)**
    - Understanding execution plans

- Indexing strategies
- Avoiding common performance bottlenecks

- **Assessment for SQL**
  - Quiz: 20-25 questions on advanced SQL concepts.
  - Practical Task: Write complex queries using window functions, CTEs, and optimization techniques.

Python Advanced Concepts (6 Hours)

- **Object-Oriented Programming (OOP) in Python (2 Hours)**
  - Classes and objects
  - Inheritance, polymorphism, and encapsulation
  - Abstract classes and interfaces
- **Python for File Operations (1.5 Hours)**
  - Working with large files efficiently
  - Reading and writing binary files
  - File compression and decompression
- **Working with Python Packages (2.5 Hours)** *(Re-numbered for logical flow)*
  - **Data Manipulation with NumPy and Pandas:**
    - NumPy arrays, slicing, and broadcasting
    - Pandas DataFrames, groupby, and aggregation
  - **Data Visualization with Matplotlib and Seaborn**
  - **Working with JSON and XML files**
- **Assessment for Python**
  - Quiz: 25-30 questions on OOP, file handling, and packages.
  - Practical Task: Automate an end-to-end workflow involving EC2, file transfer, and data processing with Pandas.

AWS and Pipeline Basics (7 Hours)

- **Introduction to AWS Services (1 Hour)**
  - Overview of AWS services (EC2, S3, IAM)
  - Understanding AWS CLI and Management Console
- **Working with S3 (2 Hours)**
  - Uploading, downloading, and deleting objects
  - Managing permissions and buckets
- **Pipeline Triggers and Automation (3 Hours)** *(Combined with 3.5 for better flow)*
  - Understanding CI/CD concepts
  - Using Apache Airflow for pipeline orchestration
  - Writing Python scripts to trigger pipelines
  - Logging and monitoring pipeline executions
  - Integrating Python and AWS for Pipelines:
    - Triggering AWS Lambda functions with Python
    - Using Python for S3 and EC2-based workflows
- **Assessment for AWS and Pipelines (1 Hour)**
  - Quiz: 15-20 questions on AWS services and pipelines.
  - Practical Task: Build and trigger a pipeline using Airflow, incorporating Python scripts and AWS S3.

Case Studies and Wrap-Up (4 Hours)

- **Intermediate Case Study 1 (2 Hours)**
  - Scenario: Automate the deployment of a file processing pipeline.
  - Launch EC2 instances using Python.
  - Transfer files via SFTP to EC2.
  - Process the data using Pandas and save it to S3.
- **Intermediate Case Study 2 (2 Hours)**
  - Scenario: Develop a data ingestion pipeline.

- Fetch data from an external source using SFTP.
- Load the data into a database using Python.
- Trigger a notification or downstream job using Airflow.

- **Wrap-Up and Discussion**
  - Recap of key learnings.
  - Addressing participant questions and concerns.

TOC FOR PHASE 3 (ADVANCED LEVEL)

Big Data Technologies (9 Hours)

- **Introduction to Big Data (1 Hour)**
  - What is big data?
  - Overview of big data ecosystems (Hadoop, Spark, Snowflake, Databricks)
- **Working with Hadoop (2 Hours)**
  - Hadoop architecture and HDFS overview
  - Writing and running MapReduce jobs
  - Basic Hadoop CLI operations
- **Introduction to Snowflake (3 Hours)**
  - Snowflake architecture and data warehousing
  - Creating tables and loading data into Snowflake
  - Querying Snowflake using SQL
  - Integrating Snowflake with Python
- **Comparative Overview of Big Data Tools (1 Hour)**
  - Use cases for Hadoop, Snowflake, and Databricks
  - Choosing the right tool for specific scenarios
- **Assessment for Big Data (1 Hour)**
  - Quiz: 15-20 questions on Hadoop, Snowflake, and Databricks basics.
  - Practical Task: Load data into Snowflake and query it. Perform basic transformations using Databricks or Hadoop CLI.

Advanced Pipeline Development (8 Hours)

- **Orchestrating Pipelines with Apache Airflow (3 Hours)**
    - Setting up Airflow and understanding DAGs
    - Creating and scheduling workflows
    - Managing dependencies between tasks
- **Python for Pipeline Automation (4 Hours) (Renamed for clarity)**
    - Automating file ingestion pipelines using SFTP and Python
    - Writing scripts to trigger Airflow jobs programmatically
    - Generating reports and logging pipeline executions
- **Assessment for Pipeline Development (1 Hour)**
    - Quiz: 20-25 questions on Airflow and Python for pipelines. (Simplified)
    - Practical Task: Design an automated pipeline using Python and Airflow, triggered by file uploads.

Advanced Python Workflows (9 Hours)

- **Python with AWS (3 Hours)**
    - Advanced boto3 operations for EC2 and S3
    - Automating AWS Lambda function triggers
    - Using Python for resource monitoring and logging
- **Python for Data Engineering (3 Hours)**
    - Handling large datasets with Pandas
    - Data cleaning and transformation pipelines
    - Using Dask for parallel processing
- **Error Handling and Logging in Python (1 Hour)**
    - Best practices for error handling
    - Implementing robust logging in Python scripts
- **Advanced File Operations (2 Hours) (Re-numbered for flow)**

- Handling multiple file formats (CSV, JSON, XML, Parquet)
- Efficient file I/O for large-scale applications

- **Assessment for Python**
  - Quiz: 20-25 questions on Python with AWS, data engineering, and error handling.
  - Practical Task: Build a Python script to automate data processing and save outputs to S3.

Comprehensive Capstone Project (4 Hours)

- **4.1 Project Overview (1 Hour)**
  - Define goals, deliverables, and milestones.
- **4.2 Case Study 1: Big Data Pipeline (3 Hours)**
  - Scenario: Extract data from multiple sources. Transform data using Python and Pandas. Load data into Snowflake and perform queries. Automate the process with Airflow and log results.

Assessments and Wrap-Up (Combined)

- Final Quiz (1 Hour): 30-40 questions covering all topics across the three phases.
- Practical Task (2 Hours): An integrated task combining Python, pipelines, and big data tools. (This likely refers to a final practical task beyond the individual assessments within each section.)

About StarAgile

StarAgile has expanded from its humble beginnings to compete with industry heavyweights over the years. By upholding the quality and consistency, StarAgile has exponentially grown since 2016 training 100k+ professionals to access 100+ countries. Moreover, our strong commitment and growthoriented mindset have supported countless individuals and teams in upskilling and gaining a competitive edge in their respective fields, fostering community trust and earning global recognition.

What We Do

To stay abreast with the everchanging technologies in this industry, there is a need for professionals to expand their knowledge restrictions, solve complex problems and successfully deliver outcomes.

StarAgile acts as a facilitator for professionals to develop newer skills, certifications, and bridge their knowledge gaps. We constantly expand our course portfolio to meet the global requirements and consistently address the evolving learning needs.

Global Accreditation

StarAgile is proudly partnered with globally recognized accreditation bodies, ensuring the highest standards of education and delivering exceptional value to our students. Our partnerships guarantee that our training programs meet industry benchmarks and provide students with the skills and credentials that truly matter.