EXECUTIVE CERTIFICATION IN

CLOUD DATA ENGINEERING

IN COLLABORATION WITH IIT GUWAHATI

Ivy® Knowledge Services (P) Ltd.

5th Floor | 14 B, Camac Street | Kolkata – 700 017 | Tel-Fax : +91 33 400 11221 | Mob : +91 9748 44 1111

www.ivyproschool.com | info@ivyproschool.com

Learning Journey – PG in Cloud Engineering

Ivy® Knowledge Services (P) Ltd.

5th Floor | 14 B, Camac Street | Kolkata – 700 017 | Tel-Fax : +91 33 400 11221 | Mob : +91 9748 44 1111

www.ivyproschool.com | info@ivyproschool.com

A. SQL FOR DATA ENGINEERING

Module 1	Basic SQL Queries
	Writing SQL queries
	SELECT statements
	Filtering and sorting data
	Practical exercises
Module 2	Cleaning and Modifying Cloud Data using SQL
	Updating cloud data using SQL
	Data Transformations using SQL
	Adding and deleting records using SQL
	Data validation and error handling
	Practical examples with cloud data modifications
Module 3	Aggregating and Analyzing Data
	Using SQL functions
	Aggregating data using SQL
	Creating pivot tables and charts from SQL data
Module 4	Working with Multiple Data Tables using SQL
	Conditional logic with CASE statements
	JOIN operations in Excel
	Real-world examples and hands-on exercises
Module 5	Troubleshooting and Error Handling
	Identifying and resolving common Excel and SQL errors
	Debugging SQL queries in Excel
	Handling connectivity issues
	Hands-on troubleshooting exercises
Module 6	Advanced Filter and Organize Data using SQL
	Pattern matching using LIKE, Regular Expressions
	Searching using Wildcards, Combining using Wildcards
	Performing full-text search
	Rank Data with the RANK and DENSE_RANK Clauses Other Window functions like LEAD, LAG, SUM, PERCENTILE, nTILE

	Summarize Grouped Data using ROLLUP and CUBE Clauses Use PIVOT and UNPIVOT Operators
	Lab: Organize data using a credit card customer profiling data.
Module 7	Advance Methods to Work with Data from Multiple Tables
	Using various types of JOINS to work with more than two tables
	Combine the results from multiple tables using Union / Union ALL
	Compare the results of two queries with the INTERSECT and EXCEPT statements
	Lab: Summarize data from multiple sources and multiple database tables.
Module 8	Use DDL and DML Statements
	SQL Data Definition Language to create, alter, drop, truncate databases
	SQL Data Manipulation Language to perform inserts, updates, and deletes
	Lab: Work on creating a new database using ERD. Insert / manipulate data, take care of update errors, relationship constraints, etc.
Module 9	Use Subqueries & Common Table Expressions
	Write self-contained subqueries
	Write correlated subqueries
	Using the EXISTS / Not Exists / Any / All predicates with subqueries
	Using Common Table Expressions - Fundamentals
Module 10	Work with Advanced Table Expressions
	Building on Subqueries, Common Table Expressions, and Unions
	Using Derived Tables and Common Table Expressions
	Using a Common Table Expression to Solve a Complicated Join Problem
	Thinking About Performance
	Lab
Module 11	User Defined Functions and Stored Procedures
	Create user defined functions to automate SQL operations
	Create stored procedures
	Lab
Module 12	Hands On Case Studies - Capstone Projects
	-eCommerce, BFSI, Retail Real-life Datasets - Hands-on Projects on analyzing data using SQL Queries
Module 13	Final Project and Recap
	Apply knowledge to a real-world project (e.g., importing, analyzing, and reporting on cloud-based data in Excel)

	Recap of key concepts and skills learned
	Q&A session and resources for further learning

B. PYTHON ESSENTIALS FOR DATA ENGINEERING

Module 1	Foundations of Python Programming
	Understanding Variable and Data Type
	Python Terminal Walkthrough
	Understanding Objects and References
	Variables Rules Numbers
	Data Type and Math Operations
	Numbers – Exponentiation and Modulo
	Arithmetic Order of Precedence
	Boolean Data Type
	Working with String
Module 2	Advanced Data Structures and Basic Operations
	List and Accessing the Elements
	List Methods
	Working with Dictionary
	Nested Dictionary
	Dictionary Methods
	Working with Tuple
	Working with Comparators
	Understanding Boolean Operators
	Arithmetic Operators
	Assignment Operators
	Unary Minus Operator
	Relational Operators
	Logical Operators
	Hands-On Project: Essential Python data structures and operators.
Module 3	Program Control Flow and Loops
	Conditional Logic
	If-Else Conditions

	While Loop Demo
	Break, Continue, and While/Else
	For Loop Demo
	Iterating Multiple Lists
	Using the Zip Function
	Using Range Function in For Loop
	Hands-On Project: Apply Python's conditional logic and loop structures for efficient data processing.
Module 4	Leveraging Functions for Data Preprocessing
	Understanding Methods
	Functions
	Working with Return Values
	Working with Positional / Optional Parameters
	Understanding Variable Scope
	More Built-In Functions
	Intro to Numpy
	Creating arrays
	Indexing Arrays
	Overview of OOP concepts: classes and objects.
	Understanding encapsulation, inheritance, and polymorphism in Python.
	Creating classes and objects in Python.
	Implementing encapsulation through class attributes and methods.
	Demonstrating inheritance and polymorphism for code reusability.
	Hands-On Project: Leverage Python's fundamental concepts, including methods, functions, and object-oriented programming (OOP) principles.
Module 5	Data Wrangling with DataFrames (Part 1)
	Establishing connections to various databases using Python.
	Utilizing libraries like pandas, SQLAlchemy, and pyodbc for database connectivity.
	Configuring database connection parameters and credentials.
	Executing SQL queries from Python using different methods.
	Utilizing Python libraries to seamlessly integrate SQL queries with Python code.
	Fetching and displaying query results in a Python environment.
	Accessing Rows and Columns in DataFrames
	Selecting Rows Based on Conditionals

	Replacing Values in DataFrames
	Handling Missing Values
	Dropping Duplicate Rows
	Looping Over a Column
	Applying a Function Over All Elements in a Column and Group
	Slicing DataFrames
	Hands-On Project: Tailored for an e-commerce platform, the goal is to establish connections to diverse databases using Python by leveraging key libraries such as pandas, SQLAlchemy and pyodbc for seamless database connectivity.
Module 6	Data Wrangling with DataFrames (Part 2)
	Merging and Concatenating DataFrames
	GroupBy Function
	Pivot Operations in DataFrames
	Dates Handling in DataFrames
	Feature Transformation: Standardizing and Normalizing
	Handling and Deleting Outliers
	Grouping Observations Using Clustering
	Advanced Filtering and Subsetting
	Efficient Memory Usage
	Parallel Processing with Dask
	Hands-On Project: The project covers merging and concatenating DataFrames.

C. BIG DATA PROCESSING

Module 1	Big Data Introduction
	Technical understanding of Distributed Computation & Storage
	Structured, Unstructured, Semi Structured Data
	File Formats : CSV, JSON, Parquet, AVRO, ORC
	Horizontal Vs Vertical Scaling
	File Compressions Techniques
	Understanding of theortical concepts mentioned in the topics
	Linux Commands and Introduction

	Quiz
Module 2	Hadoop
	Hadoop Complete Architecture
	Map-Reduce Functioning
	HDFS
	YARN (Yet Another Resource Negotiator)
	Blocks, Splits, Maps, Data Spilling, Heartbeats, Data Replication, FS Image, Checkpointing, High availibility
	Hadoop Daemons (Namenode, Datanode, Secondry Namenode, Standby Namenode)
	Setup hadoop in pseudo distributed mode in your machine , store large text file on HDFS and write Map-Reduce code to count frequency of each word
	Quiz
	Hands-on : Store large text file on HDFS and write Map-Reduce code to count frequency of each word
Module 3	Apache Hive
	Hive Installation
	Query Syntax
	Bulk Data Load
	Internal Vs External Tables
	Static & Dynamic Partitioning
	Map Side Join
	Hive SerDe
	UDF's in Hive
	Bucketing
	Query Optimization
	File Formats in Hive (ORC vs Parquet File Formats) & Compression Techniques
	Window Functions, Ranking, Sorting
	Perform a bulk load with dynamic partitioning
	Use Hive SerDe to create tables in hive for Json data
	Write and apply UDF in hive to flattern nested json file
	Quiz
	Hands-on : Create internal and external tables using data stored in HDFS
Module 4	Apache Spark - General Purpose Cluster Computing Framework
	Apache Spark Architecture

	Spark Core, Spark SQL
	Dataframes
	Datasets
	RDDs
	Spark Read/Write operations
	Lineage Graph, Lazy Evaluation
	Actions, Transformations, Optimized Joins, Brodcaster, Accumulator
	Understanding of Spark UI, Stages, Tasks
	Spark Submit Command Options
	Job optimization techniques
	Spark Catalyst Optimizer
	Static and Dynamic Resource allocation
	Understanding Memory Usage in Spark a) Cache & Persist b) Java Serializer vs Kryo Serializer
	Difference between nil, null, none & nothing
	Dealing with nulls in Scala
	What is yield
	Quiz
	Hands-on: Processing website user activity using Apache Spark.
Module 5	Apache Spark - Structured API
	Introduction to Spark Structured API
	Spark DataFrame
	Understanding SparkSession
	SparkSession vs SparkContext
	Dataframe with Various Transformations
	RDD vs DataFrame vs Datasets
	Challenges with DataFrame
	Spark Dataset API
	Difference Between DataFrame and Dataset
	Creating Data frame/Datasets from Various File Formats
	Read Modes & Schema
	Ways to Define the Schema
	Defining a Explicit Schema

	Registring UDF with Driver
	Transformations Hands on Examples
	Aggregate Transformations
	Simple Aggregations
	Grouping Aggregations
	Window Aggregations
	When to Use Simple Join When Use Broadcast Join
	Grouping Aggregation Real-time Example
	Infering Data in SparkSQL
	Quiz
	Hands-on : Enhance proficiency in Spark's Structured API and DataFrame functionalities using PySpark.
Module 6	Apache Spark - Optimization
	Level of Optimizations
	Resource level optimizations
	Application level optimizations
	Cluster level optimizations
	How to calculate no of Executors
	Thin Executor
	Fat Executor
	How to calculate no of Executors
	How to Calculate Memory allocation
	Static Resource allocation
	Dynamic Resource allocation
	Understanding Memory Usage in Spark
	Execution Memory
	Storage Memory
	Quiz
	Hands-on : Optimizing Apache Spark at various levels, including resource, application, and cluster optimizations.
Module 7	Apache Spark - Streaming
	Kind of Processing
	What is Real-tim Processing
	The Importance of Real-time Processing

	Batch processing vs Real-tim Stream Processing
	Spark Streaming Data
	Spark discretized stream or DStream
	Batch & Batch Interval
	Do Spark is a real-time streaming engine
	Stream Processing in Spark
	Transformed DStream
	Understanding Producer & Consumer Practical on Real-time Processing
	Stream Transformations
	Stateless Transformations
	Stateful Transformations Window Operations
	Batch Interval
	Window Size
	Sliding Interval
	Practical on Stateless Transformation
	Practical on Stateful Transformation
	reduceByKey vs updateStateByKey
	Working With Sliding Window
	reduceByKeyAndWindow Transformation reduceByWindow Transformation
	countByWindow Transformation
	Quiz
	Hands-on : Implement real-time data processing using Apache Spark Streaming.
Module 8	Real-Time Data Pipeline with Kafka
	Producer
	Consumer
	Kafka Cluster,Cluster Setup, Brokers
	Topics, Partitioning, Offset, Polling, Data Replication, Data Retention
	Consumer Group
	ZooKeeper
	Hands-on: Create realtime datapipeline using MySQL as source for incremental data stream, Apache Kafka for messaging Queue and Spark Streaming for data transformation. Store transformed realtime data in any NoSQL database for Analytical queries
Module 9	Real-Time Data Pipelines with MongoDB

	Understanding NoSQL databases and their relevance.
	Comparisons with traditional relational databases.
	Types of NoSQL databases
	Introduction to MongoDB and NoSQL Concepts
	Installation and setup of MongoDB.
	Basic MongoDB Operations and CRUD
	Basic commands and operations.
	Data Types and Indexing in MongoDB.
	Query operators: $gt, $lt, $in, etc.
	Data Modeling: Embedding and referencing documents.
	MongoDB Indexing and Performance Optimization
	Types of Indexes in MongoDB.
	Strategies for creating and using indexes effectively.
	Analyzing and optimizing query performance.
	Profiling queries: Understanding the database profiler.
	Monitoring tools and best practices for performance optimization.
	Real-Time Data Pipeline with MongoDB
	The role of MongoDB in real-time data pipelines.
	Using MongoDB as a source for incremental data stream.
	Integrating Apache Kafka as a messaging queue.
	Implementing Spark Streaming for data transformation.
	Storing transformed real-time data in MongoDB for analytical queries.
	Final project: Architecting and implementing a real-time data pipeline for an e-commerce company seeking to bolster its analytical capabilities.

D. AZURE CLOUD ENGINEERING

Module 1	Fundamentals of Azure
	Overview of Azure services
	Cloud service models: IaaS, PaaS, SaaS
	Azure global infrastructure
	Azure Active Directory (AAD) basics

	Azure Managed Identity
	Role-based access control (RBAC)
	Azure Security Fundamentals (NSG, Key Vault, Key Vault policies)
	Overview of Azure CLI
	Common CLI commands for Azure management
	Hands-on Practice: Delve into the fundamental aspects of Azure, gaining practical experience with key services and concepts.
Module 2	Virtual Machines, Storage, and Database Services
	Creating and managing VMs
	VM extensions and customization
	Virtual machine scalability and availability
	Azure Storage Services (Blob, Queue, Table)
	Azure SQL Database
	Azure Cosmos DB
	Azure Data Lake Storage Gen1 and Gen2
	Data Lake security and access control
	Hands-on Practice: Creating and managing virtual machines (VMs) on Azure.
Module 3	Advanced Data Engineering with Azure
	Overview of Azure Data Factory (ADF)
	Building end-to-end data pipelines
	Data movement and transformation activities
	Data integration and orchestration
	Triggers and scheduling in Azure Data Factory
	Looping and conditional constructs
	Parameterization and dynamic content
	Data Flow transformations and transformations using Mapping Data Flow
	Introduction to Azure Databricks
	Collaborative data science and engineering
	Notebook-based data processing
	Advanced transformations and analytics in Databricks
	Integration with Azure Data Factory
	Overview of Azure Event Hubs
	Integrating Event Hubs with Azure Data Factory
	Real-time data streaming and processing

	Overview and architecture
	Data warehousing concepts
	Query optimization and performance tuning
	Integration with Azure Data Factory
	Hands-On Project: Azure Data Engineering and Integration.
Module 4	Hybrid Cloud Data Engineering, Governance, and Compliance
	Connecting on-premises and multi-cloud environments
	Hybrid data scenarios and use cases
	Managing and orchestrating data workloads across diverse environments
	Implementing centralized control and governance
	Establishing secure access for on-premises and cloud-based data resources
	Identity federation and synchronization
	Enforcing data quality and compliance standards
	Policy-driven data management practices
	Understanding regulatory requirements for data engineering
	Aligning data practices with industry and regional standards
	Data security and privacy considerations
	Auditing and reporting for compliance verification
	Best practices for maintaining data integrity in a hybrid environment
	Hands-On Project: Hybrid Data Engineering Governance and Security.
Module 5	Azure DevOps, Containers, and Security
	Introduction to Azure DevOps for Data Projects
	Agile methodologies for data engineering
	Collaborative development practices
	Building CI/CD Pipelines for Data Workloads
	Automating data pipeline deployments
	Version control and continuous integration for data artifacts
	Source Code Management with Azure Repos
	Managing and versioning data engineering code
	Branching strategies for collaborative development
	Containerization Basics
	Introduction to containers in the context of data engineering
	Docker fundamentals for encapsulating data applications
	Deploying Data Workloads with Azure Kubernetes Service (AKS)

	Orchestrating and scaling data workloads using AKS
	Integration of AKS with data processing frameworks
	Azure Automation for Data Engineering Tasks
	Task scheduling and automation for data pipelines
	Scripting and configuration management best practices
	Application Insights for Data Engineering Monitoring
	Monitoring and optimizing data pipeline performance
	Analyzing telemetry data for continuous improvement
	Serverless Computing with Azure Functions in Data Workflows
	Event-driven data processing with Azure Functions
	Integration with data services for serverless architecture
	Azure Security Center for Data
	Implementing security policies for data workloads
	Continuous security monitoring and threat detection
	Azure Sentinel for Advanced Threat Hunting and Analysis
	Leveraging Azure Sentinel for proactive threat management
	Advanced analytics for identifying and responding to security incidents
	Hands-On Project: Azure DevOps for Secure and Efficient Data Engineering.