

VERSION IT

Multi-Cloud Data Engineer Program

(GCP + AZURE)

with Case Studies and Real-world Projects



#No -46, SAP street, Behind Mytrivanam,
Ameerpet, Hyderabad-500038



+91 9848015399, +91 9391237284



www.versionit.co.in

About Us Version IT

Version IT is not a mere software training institute, a team of IT professionals developed it as the best knowledge centre for hundreds of career-building conscious young people. Our training academy is the best training institute in Hyderabad offering various software courses with aptly placement orientation. We proudly announce that we achieved 100% placements in every batch we have taken up in the past two decades. Version IT Academy's strength is our academic excellence with which we have been placed in the top position among the software training institute in Hyderabad.

Corporate Training



Classroom Training



Online Training



Course duration: 4 months

Teaching Language: English

Mode: Online

Multi-Cloud Data Engineer (GCP + Azure) Curriculum

Cloud Introduction

- What is Cloud Computing?
- The need for cloud computing in modern businesses.
- Why organizations moved to Cloud
- Cloud Service Providers – GCP, Azure, AWS, ...

GCP Cloud Basics

GCP Cloud basics

- Introduction to Google Cloud Platform (GCP) and its core services
- GCP interfaces - Console, Cloud Shell and Cloud SDK
- GCP Locations - Regions, Zones & Global Infrastructure
- GCP Walkthrough - GCP Projects, Organizations, Folders & Billing Accounts
- GCP IAM & Admin - Principals, Roles, Service Accounts & Resource Hierarchy
- Setting up a Free Account in GCP

Linux Basics on Cloud Shell

- Getting started with Linux
- Linux Installation
- Basic Linux Commands
 - Cloud shell tips
 - File and Directory Operations (ls, cd, pwd, mkdir, rmdir, cp, mv, touch, rm, nano)
 - File Content Manipulation (cat, less, head, tail, grep)
 - Text Processing (awk, sed, cut, sort, uniq)
 - User and Permission related (whoami, id, su, sudo, chmod, chown)

Python for Data Engineer

- Introduction to Python and basic functions
- Data Types – Strings, Operators, Numbers (Int, Float), Booleans
- Data Structures – Lists, Tuples, Dictionaries, Sets
- Python Programming Constructs - if, elif, else statements, for loops, while loops
- File I/O operations - Read, write, append, rename, remove, copy, move
- Exception Handling and Logging
 - what is exception handling, why it is needed
 - try and except blocks, else, finally, raise blocks
 - Why Logging? And Logging Levels
 - Writing Logs to File

- Data and Time operations - Working date and time, formatting, Parsing, Time delta
- Modular Programming in Python
 - Functions & Lambda Functions
 - Classes
- REST API basics
 - What is API
 - HTTP methods (GET, POST, PUT, DELETE)
 - Status Codes, Request Headers, JSON Responses
 - Using the requests Library, Reading API Responses
- Modules and Packages
 - What is module and packages
 - Importing
 - Built-in modules
 - Install packages

GCP Data Engineering Tools

Google Cloud Storage

- Overview of Cloud Storage as a scalable and durable object storage service.
- Understanding buckets and objects in Cloud Storage.
- Use cases for Cloud Storage, such as data backup, multimedia storage, and website content
- Labs: using console & CLI to do below
 - Creating and managing Cloud Storage buckets.

- Uploading and downloading objects to and from Cloud Storage.
 - Setting access controls and permissions for buckets and objects.
 - Data Transfer and Lifecycle Management
 - Versioning and Object Versioning
- Integration with Other GCP Services
 - Monitoring and logging for Cloud Storage operations.

Cloud SQL

- Introduction to Cloud SQL
- Creating and Managing Cloud SQL Instances
- Configuring database settings, users, and access controls.
- Connecting to Cloud SQL instances using Cloud SQL studio, Shell, Workbenches
- Importing and exporting data in Cloud SQL.
- Backups and High Availability
- Integration with Other GCP Services
- Managing database user roles and permissions.
- Introduction to DMS
- End to End Database migration Project
 - Manual: Export and Import method
 - Automation: Cloud SQL DMS method

BigQuery (SQL development)

- Introduction to BigQuery
- BigQuery Architecture
- Use cases for BigQuery in business intelligence and analytics.
- Various method of creating table in BigQuery
- BigQuery Data Sources and File Formats

- Native table and External Tables
- Working with complex data types
 - Working json data, nested, repeated and array data
- Data Integration and Export
 - Loading data into BigQuery from Cloud Storage, Cloud SQL, and other sources.
 - Exporting data from BigQuery to various formats.
 - Real-time data streaming into BigQuery.
- Configuring access controls and permissions in BigQuery.
- BigQuery Views:
 - Loading data into BigQuery from Cloud Storage, Cloud SQL, and other sources.
 - Exporting data from BigQuery to various formats.
 - Real-time data streaming into BigQuery.
- Optimization techniques in BigQuery
- BigQuery Slots – on demand, flat-rate, flex-slots
- Case Study-1: implement a real-world analytics data platform for Spotify
- Case Study-2: Enterprise Social Media analytics platform

DataProc (Pyspark Development)

- Introduction to Hadoop and Apache Spark
- Understanding the difference between Spark and MapReduce
- What is Spark and Pyspark.
- Understanding Spark framework and its functionalities

- Overview of DataProc as a fully managed Apache Spark and Hadoop service.
- Cluster Creation and Configuration
 - Creating and managing DataProc clusters.
 - Configuring cluster properties for performance and scalability.
 - Preemptible instances and cost optimization.
- Learning Pyspark:
 - How to read from multiple data sources – csv, text, json, parquet, database table, BigQuery tables
 - How to perform multiple transformations
 - How to write to multiple targets - csv, text, json, parquet, database table, BigQuery tables
- Running Jobs on DataProc
 - Submitting and monitoring Spark and Hadoop jobs on DataProc.
 - Use of initialization actions and custom scripts.
 - Job debugging and troubleshooting.
- Case study-1: Data Cleaning of Employee Travel Records
- Case study-2: Processing real-time patient health data
- Case study-3: Creating a pyspark job to support ML model creations
- Introduction to Databricks Lakehouse Platform
- Databricks Architecture and Core Components
 - Explore the architecture, including clusters, jobs, and the workspace environment.
 - Key components like Databricks Runtime, Spark, and Delta Lake.

- o Setting Up and Administering a Databricks Workspace
 - Step-by-step guide on workspace creation and management.
 - Admin tasks including user management, permissions, and billing.
- o Working with Notebooks and Clusters
 - Collaborative data science using Databricks notebooks.
 - Cluster setup, management, and autoscaling.
- o Databricks Unity Catalog
 - Unified governance solution for managing data and metadata across the lakehouse.
 - Organizing data assets and enforcing security policies.
- o Managing Data with Delta Lake
 - Understanding Delta Lake's ACID transactions and versioning capabilities.
 - Best practices for efficient data management.
 - Creating Delta Lake Tables
 - Querying and managing data across multiple file formats.
- o Building ELT Pipelines with Spark SQL and Python
- o Incremental Data Processing with Delta Lake
- o Case Study 1: Real-time data processing with Spark Structured streaming
- o Case Study 2: Databricks Autoloader for file-based ingestion utility
- o Performance Optimization in Databricks

GCP Data Engineering Tools

Google Cloud Storage

- o Overview of Cloud Storage as a scalable and durable object storage service.
- o Understanding buckets and objects in Cloud Storage.
- o Use cases for Cloud Storage, such as data backup, multimedia storage, and website content
- o Labs: using console & CLI to do below
 - Creating and managing Cloud Storage buckets.
 - Uploading and downloading objects to and from Cloud Storage.
 - Setting access controls and permissions for buckets and objects.
 - Data Transfer and Lifecycle Management
 - Versioning and Object Versioning
- o Integration with Other GCP Services
- o Monitoring and logging for Cloud Storage operations

Cloud SQL

- o Introduction to Cloud SQL
- o Creating and Managing Cloud SQL Instances
- o Configuring database settings, users, and access controls.
- o Connecting to Cloud SQL instances using Cloud SQL studio, Shell, Workbenches
- o Importing and exporting data in Cloud SQL.
- o Backups and High Availability
- o Integration with Other GCP Services
- o Managing database user roles and permissions.

- o Introduction to DMS
- o End to End Database migration Project
 - Manual: Export and Import method
 - Automation: Cloud SQL DMS method

BigQuery (SQL development)

- o Introduction to DMS
- o End-to-End Database migration Project
- o Introduction to BigQuery
- o BigQuery Architecture
- o Use cases for BigQuery in business intelligence and analytics.
- o Various method of creating table in BigQuery
- o BigQuery Data Sources and File Formats
- o Native table and External Tables
- o Working with complex data types
 - Working json data, nested, repeated and array data
- o Data Integration and Export
 - Loading data into BigQuery from Cloud Storage, Cloud SQL, and other sources.
 - Exporting data from BigQuery to various formats.
 - Real-time data streaming into BigQuery.
- o Configuring access controls and permissions in BigQuery.
- o BigQuery Views:
 - Views
 - Materialized Views
 - Authorized Views
- o BigQuery Views:
- o Optimization techniques in BigQuery
- o BigQuery Slots – on demand, flat-rate, flex-slots

- o Case Study-1: implement a real-world analytics data platform for Spotify
- o Case Study-2: Enterprise Social Media analytics platform

DataFlow (Apache Beam development)

- o Introduction to DataFlow
- o Use cases for DataFlow in real-time analytics and ETL.
- o Understanding the difference between Apache Spark and Apache Beam
- o How Dataflow is different from Dataproc
- o Learning Apache Beam
 - How to read from multiple data sources – csv, text, json, parquet, database table, BigQuery tables
 - How to perform multiple transformations
 - How to write to multiple targets - csv, text, json, parquet, database table, BigQuery tables
- o Case study-1: Template method of creating pipelines
- o Case study-2: E-commerce Transaction Processing
- o Case study-3: End to End Streaming Pipeline using Apache beam with Dataflow, Python app, PubSub, BigQuery, GCS

Cloud Pub/Sub (Streaming)

- o Introduction to Pub/Sub
- o Understanding the role of Pub/Sub in event-driven architectures.
- o Key Pub/Sub concepts: topics, subscriptions, messages, and acknowledgments.
- o Creating and Managing Topics and Subscriptions
 - Using the GCP Console to create Pub/Sub topics and subscriptions.
 - Configuring message retention policies and acknowledgment settings.

- o Publishing and Consuming Messages
 - Writing and deploying code to publish messages to a topic.
 - Implementing subscribers to consume and process messages from subscriptions.
- o Case study-1: Streaming use-case using Dataflow

Cloud Composer (DAG Development)

- o Introduction to Composer/Airflow
- o Overview of Airflow Architecture
- o Use cases for Composer in managing and scheduling workflows.
- o Creating and Managing Workflows
 - Creating and configuring Composer environments.
 - Defining and scheduling workflows using Apache Airflow.
 - Monitoring and managing workflow executions.
- o Integration with Data Engineering Services
 - Orchestrating workflows involving BigQuery, DataFlow, and other services.
 - Coordinating ETL processes with Composer.
 - Integrating with external systems and APIs.
- o Error Handling and Troubleshooting
 - Handling errors and retries in Composer workflows.
 - Debugging and troubleshooting failed workflow executions.
 - Logging and monitoring for Composer workflows.
- o Level-1-DAG: Orchestrating the BigQuery pipelines
- o Level-2-DAG: Orchestrating the DataProc pipelines
- o Level-3-DAG: Orchestrating the Dataflow pipelines
- o Deploy DAGs: Implementing CI/CD in Composer Using Cloud Build and GitHub

Azure Data Engineering Tools

Azure Introduction

- o Introduction to Azure Cloud Platform and its core services
- o Azure interfaces - Console, Cloud Shell and Cloud SDK
- o Azure Locations - Regions, Zones & Global Infrastructure
- o Azure Walkthrough – Subscriptions, Resources, Resource group & Management group
- o Azure IAM & Admin - Identities, Roles, Service Accounts & Resource Hierarchy
- o Setting up a Free Account in Azure Cloud

Azure DataLake Storage (ADLS)

- o Introduction to Azure Data Lake Storage (ADLS)
- o Core Components of ADLS Gen2
 - Blobs, Containers, and Storage Accounts
 - Hierarchical Namespace
- o Why Use Azure Data Lake Storage?
 - Scalability, Cost-Effectiveness, Security
 - Integration with Analytics Services
- o Access and Management of ADLS
 - Data Management via Console, SDK, CLI, REST API
 - Role-Based Access Control (RBAC) and Access Control Lists (ACLs)
- o Pricing Overview
 - Pricing Tiers (Hot, Cool, Archive)
 - Cost Factors
- o Best Practices for ADLS
 - Data Partitioning
 - Tiering Strategy for Cost Optimization
 - Monitoring and Usage Tracking

- o Interactive Access Methods
 - Managing ADLS via Console
 - CLI/Shell Interaction
 - Programmatic Access via SDKs

Azure SQL DB

- o Introduction to Azure SQL Database
- o Key Features
- o Creating a SQL db for practice

Azure Data Factory (ADF)

- o What is ADF and Use cases
- o Core components of ADF and understanding & configuring
 - Integration runtime
 - Linked Services
 - Datasets
 - Pipelines
 - Activities
 - Triggers
- o Key Features of Azure Data Factory
- o Pipelines in ADF
 - Creating and Configuring Pipelines
 - Activities and Triggers
 - Monitoring and Managing Pipeline Runs
 - Orchestration
- o Multiple examples with ADF activities
- o Case study-1: End to End Incremental pipeline with ADF, ADLS, SQL DB.
- o Best Practices for ADF
 - Designing Efficient Pipelines
 - Data Partitioning and Parallelism
- o Monitoring and Optimization for Performance

Azure Synapse

- o Synapse Workspace, SQL Pools & Architecture
- o Dedicated & Serverless SQL concepts
- o Data ingestion using Synapse Pipelines
- o External Tables, Views & Data Lake integration
- o Query performance optimization, partitioning & workload management

Databricks on Azure/GCP

- o Introduction to Databricks Lakehouse Platform
- o Databricks Architecture and Core Components
 - Explore the architecture, including clusters, jobs, and the workspace environment.
 - Key components like Databricks Runtime, Spark, and Delta Lake.
- o Setting Up and Administering a Databricks Workspace
 - Step-by-step guide on workspace creation and management.
 - Admin tasks including user management, permissions, and billing.
- o Working with Notebooks and Clusters
 - Collaborative data science using Databricks notebooks.
 - Cluster setup, management, and autoscaling.
- o Databricks Unity Catalog
 - Unified governance solution for managing data and metadata across the lakehouse.
 - Organizing data assets and enforcing security policies.
- o Managing Data with Delta Lake
 - Understanding Delta Lake's ACID transactions and versioning capabilities.
 - Best practices for efficient data management.
 - Creating Delta Lake Tables
 - Querying and managing data across multiple file formats.

- o Building ELT Pipelines with Spark SQL and Python
- o Incremental Data Processing with Delta Lake
- o Case Study 1: Real-time data processing with Spark Structured streaming
- o Case Study 2: Databricks Autoloader for file-based ingestion utility
- o Performance Optimization in Databricks

By the End of the course What Students can Expect

1. Build end-to-end Data Engineering pipelines on GCP & Azure
2. Develop strong skills in SQL, Python, PySpark, Linux & Apache Beam
3. Work with industry standard data engineering services
4. Implement Batch & Real-Time Data Processing pipelines
5. Design Bronze → Silver → Gold / Medallion Architecture
6. Build production-ready projects and troubleshoot real-world data engineering scenarios
7. Understand Cloud, Data Lake, Data Warehouse & Lakehouse architectures
8. Prepare for Data Engineer interviews with practical, scenario-based knowledge

Thank You

Our alumni work At



#No -46, SAP street, Behind Mytrivanam,
Ameerpet, Hyderabad-500038



+91 9848015399, +91 9391237284



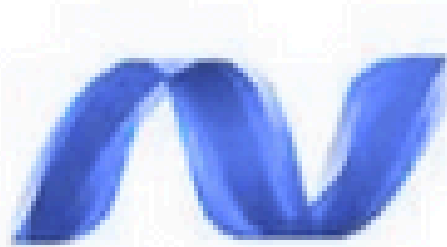
www.versionit.co.in

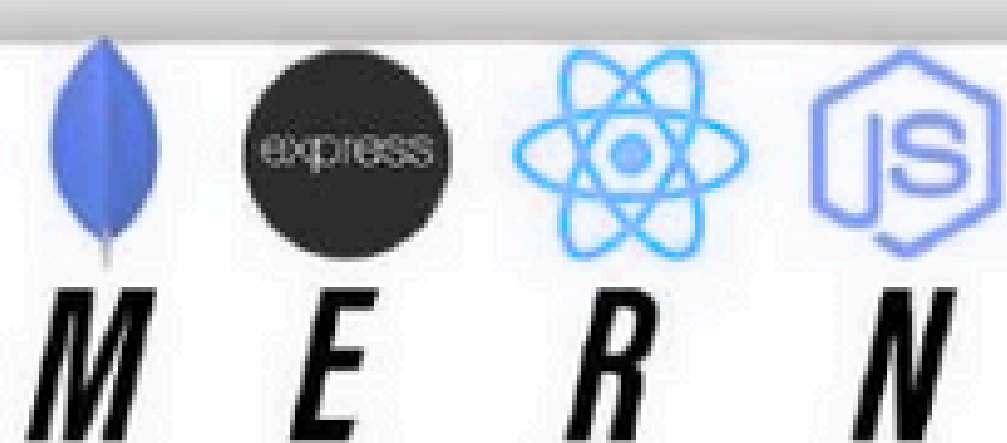
Our Other Courses

Development Technologies

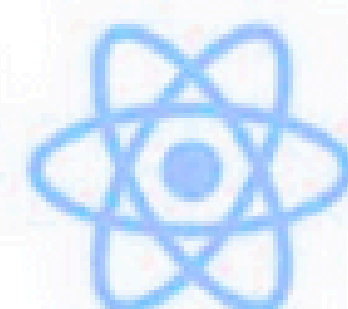
Java Full stack 

Python Full stack 

.Net Full stack 

M E R N 

M E A N 

React 

Cloud Technologies

aws 

Azure 

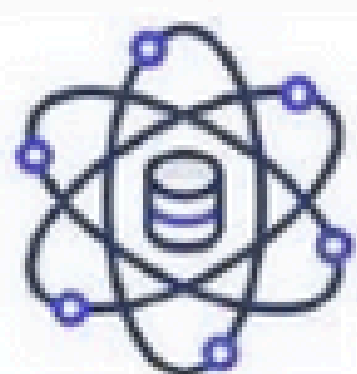
GCP 

Dev Ops 

Salesforce 

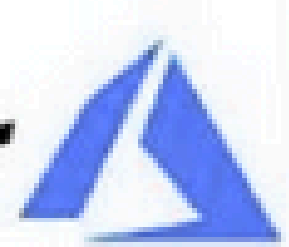
servicenow 

Triending Technologies

Data Science 

Data Analytics 

Cyber Security 

Azure Data Engineer 

Aws Data Engineer 

GCP Data Engineer 

#No -46, SAP street, Behind Mytrivanam,
Ameerpet, Hyderabad-500038

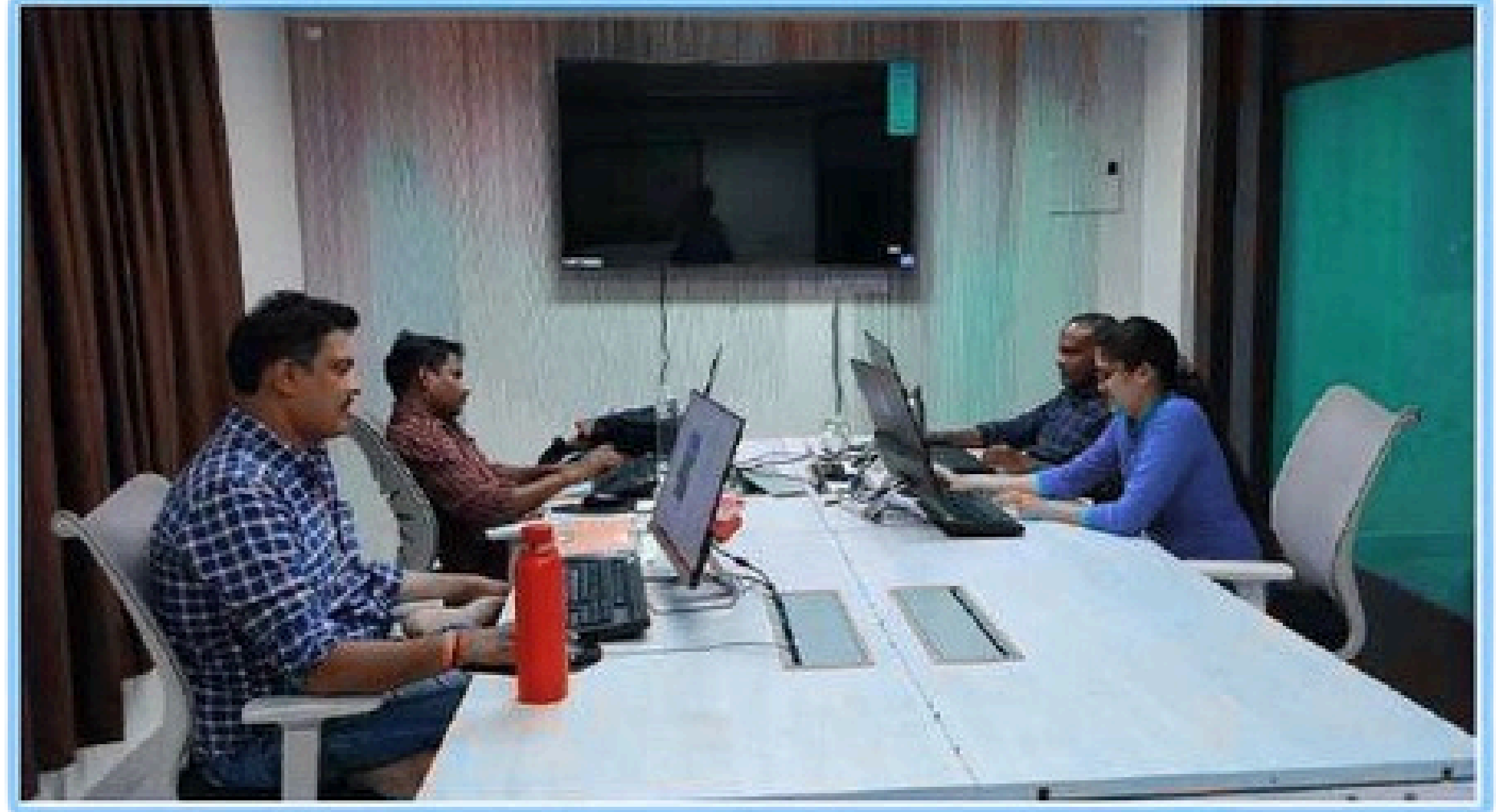
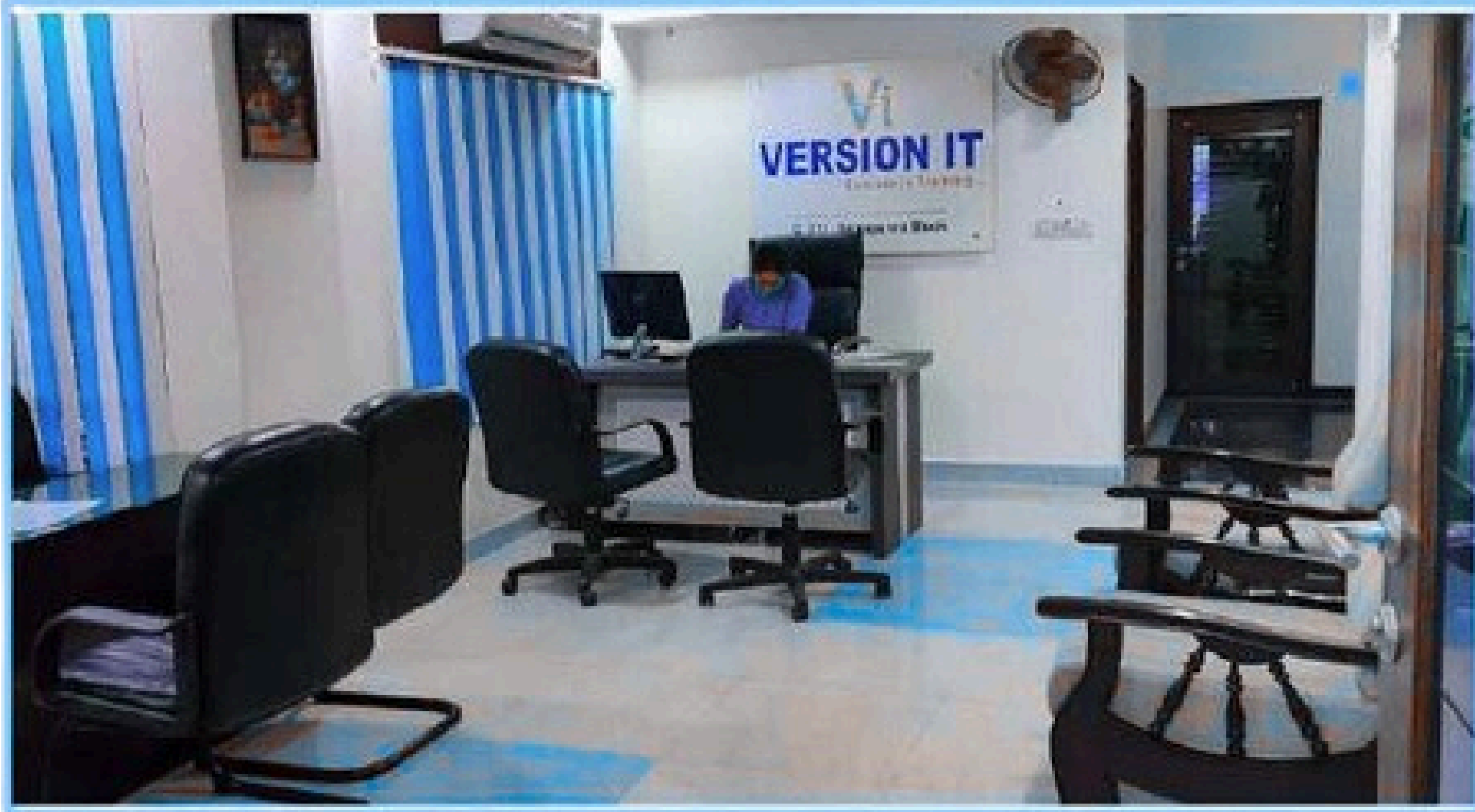


+91 9848015399, +91 9391237284



www.versionit.co.in

Our infrastructure



#No -46, SAP street, Behind Mytrivanam,
Ameerpet, Hyderabad-500038



+91 9848015399, +91 9391237284



www.versionit.co.in

Multi-Cloud

ADDRESS:

**Version IT: No: 46, SAP street, Behind Mytrivanam,
Ameerpet, Hyderabad-500038**



Call / Whatsapp : +91 9848015399 | +91 9391237284
www.versionit.co.in