Quality Thought

Transforming Dreams ! Redefining Future!

Celebrating 16 Years

A gateway to Your Bright Future

GCP Data Engineer

98491 78009

Plot No:850, Road, No: 45, 100 Feet Rd, Ayyappa Society, Madhapur, Hyd.-500081

qualitythought.in

Quality Thought
Transforming Dreams ! Redefining Future!

GCP Introduction

- The need for cloud computing in modern businesses.
- Key features and offerings of Google Cloud Platform (GCP).
- Overview of core GCP services and products.
- Benefits and advantages of using cloud infrastructure.
- Step-by-step guide to creating a free-tier account on GCP.

GCP Interfaces

Console

- Navigating the GCP Console
- Configuring the GCP Console for Efficiency
- Using the GCP Console for Service Management

Shell

- Introduction to GCP Shell
- Command-line Interface (CLI) Basics
- GCP Shell Commands for Service Deployment and Management

SDK

- Overview of GCP Software Development Kits (SDKs)
- Installing and Configuring SDKs
- Writing and Executing GCP SDK Commands

GCP Locations

Regions

- Understanding GCP Regions
- Selecting Regions for Service Deployment
- Impact of Region on Service Performance

Zones

- Understanding GCP Regions
- Selecting Regions for Service Deployment
- Impact of Region on Service Performance

Importance

- Exploring GCP Zones
- Distributing Resources Across Zones
- High Availability and Disaster Recovery Considerations

Quality Thought
Transforming Dreams ! Redefining Future!

GCP IAM & Admin

Identities

- Introduction to Identity and Access Management (IAM)
- Users, Groups, and Service Accounts
- Best Practices for Identity Management

Roles

- GCP IAM Roles Overview
- Defining Custom Roles
- Role-Based Access Control (RBAC) Implementation

Policy

- Resource-based Policies
- Understanding and Implementing Organization Policies
- Auditing and Monitoring Policies

Resource Hierarchy

- GCP Resource Hierarchy Structure
- Managing Resources in a Hierarchy
- Organizational Structure Best Practices

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

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
- Exception Handling
- File I/O operations

Modular Programming in Python

- Functions & Lambda Functions
- Classes

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
- 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
- Implementing best practices for optimizing Cloud Storage performance.
- Securing data in Cloud Storage with encryption and access controls.
- 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
    - Offline: Export and Import method
    - Online: 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
- SQL Queries and Performance Optimization

- Writing and optimizing SQL queries in BigQuery.
- Understanding query execution plans and best practices.
- Partitioning and clustering tables for performance.

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:

- Views
- Materialized Views
- Authorized Views

Integration with Other GCP Services

- Integrating BigQuery with Dataflow for ETL processes.
- Building data pipelines with BigQuery and Composer.

Case Study-1: Spotify
Case Study-2: Social Media

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.
- Use cases for DataProc in data processing and analytics.
- Cluster Creation and Configuration
    - Creating and managing DataProc clusters.
    - Configuring cluster properties for performance and scalability.
    - Preemptible instances and cost optimization.

Running Jobs on DataProc

- Submitting and monitoring Spark and Hadoop jobs on DataProc.
- Use of initialization actions and custom scripts.
- Job debugging and troubleshooting.

Integration with Storage and BigQuery

- Reading and writing data from/to Cloud Storage and BigQuery.
- Integrating DataProc with other storage solutions.
- Performance optimization for data access.
    - Reading and writing data from/to Cloud Storage and BigQuery.
    - Integrating DataProc with other storage solutions.
    - Performance optimization for data access.

- Automation and scheduling of recurring jobs.
- Case Study-1: Data Cleaning of Employee Travel Records
- End to End Batch Pyspark pipeline using Dataproc, BigQuery, GCS

Databricks on GCP

- What is Databricks lakehouse platform
- Databricks architecture and components
- Setting up and Administering a Databricks workspace
- Managing data with Delta Lake
- Databricks Unity Catalog
- Note books and clusters
- ELT with Spark SQL and Python
- optimize performance within Databricks.
- Incremental Data Processing
- Delta Live tables
- Case study: creating end to end workflows

DataFlow (Apache Beam development)

- Introduction to DataFlow
- Use cases for DataFlow in real-time analytics and ETL.
- Understanding the difference between Apache Spark and Apache Beam
- How Dataflow is different from Dataproc
- Building Data Pipelines with Apache Beam
  - Writing Apache Beam pipelines for batch and stream processing.
  - Custom Pipelines and Pre-defined pipelines
  - Transformations and windowing concepts.
- Integration with Other GCP Services
  - Integrating DataFlow with BigQuery, Pub/Sub, and other GCP services.
  - Real-time analytics and visualization using DataFlow and BigQuery.
  - Workflow orchestration with Composer.
- End to End Streaming Pipeline using Apache beam with Dataflow, Python app, PubSub, BigQuery, GCS
- Template method of creating pipelines

Cloud Pub/Sub

- Introduction to Pub/Sub
- Understanding the role of Pub/Sub in event-driven architectures.
- Key Pub/Sub concepts: topics, subscriptions, messages, and acknowledgments.
- Creating and Managing Topics and Subscriptions
  - Using the GCP Console to create Pub/Sub topics and subscriptions.
  - Configuring message retention policies and acknowledgment settings.
- Publishing and Consuming Messages
  - Writing and deploying code to publish messages to a topic.
  - Implementing subscribers to consume and process messages from subscriptions.
- Integration with Other GCP Services
  - Connecting Pub/Sub with Cloud Functions for serverless event-driven computing.
  - Integrating Pub/Sub with Dataflow for real-time stream processing.
- Streaming use-case using Dataflow

Quality Thought
Transforming Dreams ! Redefining Future!

Cloud Composer (DAG Creations)

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

Data Fusion

- Introduction to Data Fusion
  - Overview of Data Fusion as a fully managed data integration service.
  - Use cases for Data Fusion in ETL and data migration.
- Building Data Integration Pipelines
  - Creating ETL pipelines using the visual interface.
  - Configuring data sources, transformations, and sinks.
  - Using pre-built templates for common integration scenarios.
- Integration with GCP and External Services
  - Integrating Data Fusion with BigQuery, Cloud Storage, and other GCP services.
- End to End pipeline using Data fusion with Wrangler, GCS, BigQuery

Cloud Functions

- Cloud Functions Introduction
- Setting up Cloud Functions in GCP
- Event-driven architecture and use cases
- Writing and deploying Cloud Functions
- Triggering Cloud Functions:
  - HTTP triggers
  - Pub/Sub triggers
  - Cloud Storage triggers
- Monitoring and logging Cloud Functions
- Usecase-1: Loading the files from GCS to BigQuery as soon as it is uploaded

Terraform

- Terraform Introduction
- Installing and configuring Terraform.
- Infrastructure Provisioning
- Terraform basic commands
  - Init, plan, apply, destroy
- Create Resources in Google Cloud Platform
  - GCS buckets
  - Dataproc cluster
  - BigQuery Datasets and tables
  - And more resources as needed

By the End of the course What Students can Expect

Proficient in SQL Development

- Mastering SQL for querying and manipulating data within Google BigQuery and Cloud SQL.
- Writing complex queries and optimizing performance for large-scale datasets.
- Understanding schema design and best practices for efficient data storage.

Pyspark Development Skills

- Proficiency in using PySpark for large-scale data processing on Google Cloud.
- Developing and optimizing Spark jobs for distributed data processing.
- Understanding Spark's RDDs, Dataframes, and transformations for data manipulation.

Apache Beam Development Mastery

- Creating data processing pipelines using Apache Beam.
- Understanding the concepts of parallel processing and data parallelism.
- Implementing transformations and integrating with other GCP services.

DAG Creations with Cloud Composer

- Designing and implementing Directed Acyclic Graphs (DAGs) for orchestrating workflows.
- Using Cloud Composer for workflow automation and managing dependencies.
- Developing DAGs that integrate various GCP services for end-to-end data processing.

Notebooks, Workflows with Databricks

- Understand how to build and manage data pipelines using Databricks and Delta Lake.
- Efficiently query and analyze large datasets with Databricks SQL and Apache Spark.
- Implement scalable workflows and optimize performance within Databricks.

Architecture Planning

- Proficient in architecting end-to-end data solutions on GCP.
- Understanding the principles of designing scalable, reliable, and cost-effective data architectures.

Certification Readiness

- Prepare for the Google Cloud Professional Data Engineer (PDE) and
- Associate Cloud Engineer (ACE) certifications through a combination of theoretical knowledge & hands-on experience

The course will empower students with practical skills in SQL, PySpark, Apache Beam, DAG creations, and architecture planning, ensuring they are well-prepared to tackle real-world data engineering challenges and successfully obtain GCP certifications.

Our Students Are Placed In

tcs | vsoft | Cognizant | Microsoft | ENVISION | Sunseaz Technologies Pvt Ltd | wipro | AVSOFT | amazon
PrimeSoft | statista | Practera | PIERSOFT | KELTON TECH | Capgemini | KENSIUM | SOLIX | sumtotal | MIRABEL TECHNOLOGIES
VALUEMOMENTUM | PERVACIO | Apollo | LARSEN & TOUBRO | STIC Soft | Advi | NOOR | servicenow | ValueLabs | erwin
CyberTech | BLACK KNIGHT | Micron | highradius | HEXAWARE | NextSphere | GLOBAL | splashBI | archents | techforce.ai
Nendrasys | magnasoft | Innominds Powering the Digital Next | High Noon Consulting | JRD SYSTEMS | Medtronic | MAXO | PTW | Uber | techolution
SOLUGENIX | Collabera | iPrism Technologies | meslova | actlogica | S&P Global | FISSION | SOLIGENIX | TechnoFacts | SAHASYA

Quality Thought

98491 78009

Quality Thought Infosystems India (P) Ltd.

Plot No:850, Road, No: 45, 100 Feet Rd, Ayyappa Society, Madhapur, Hyderabad, Telangana 500081