Quality Thought

Celebrating 16 Years

A gateway to Your Bright Future

TriCloud Data Engineering Architect Program

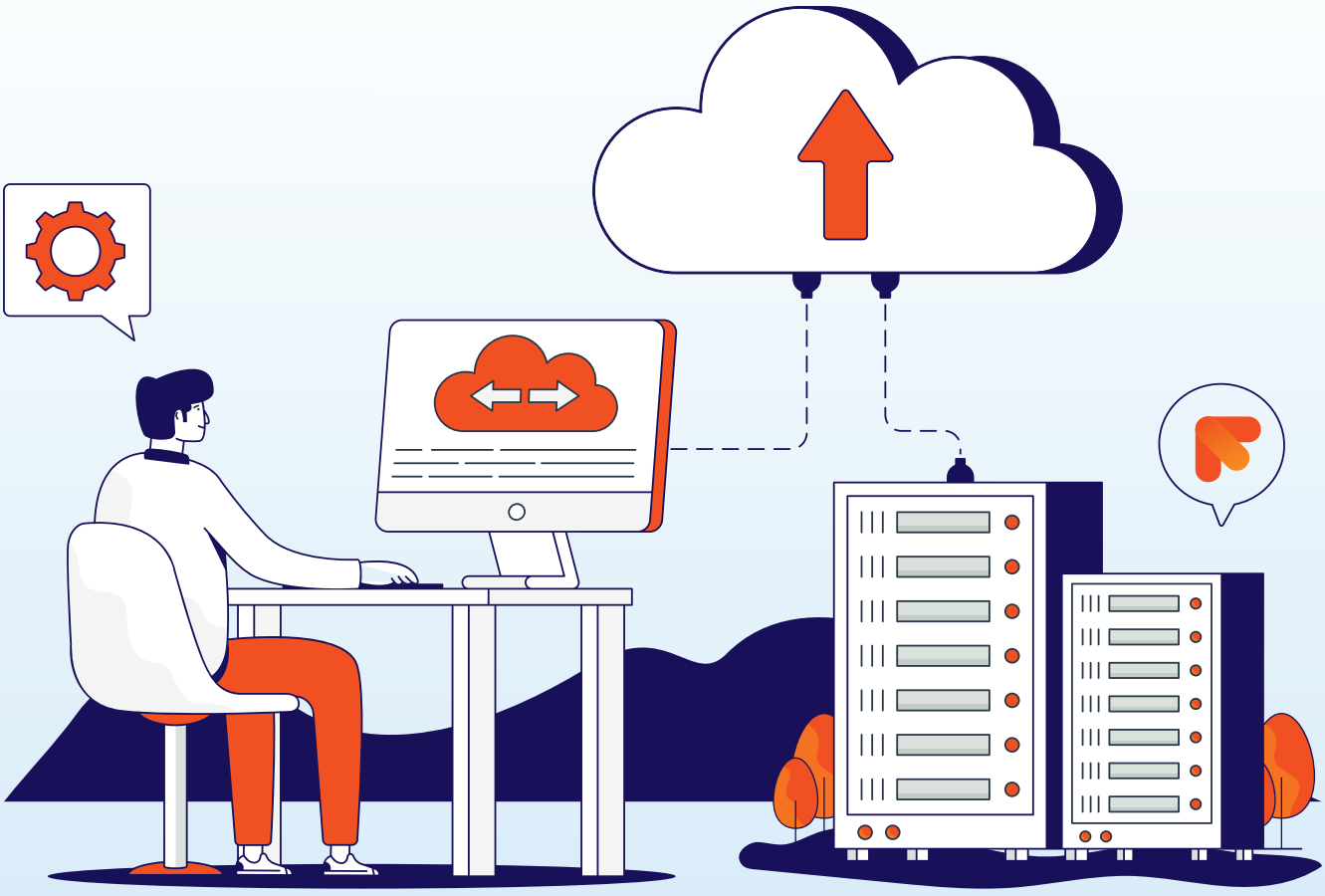
Topic

- Introduction to Python for Data Engineering, scripting fundamentals
- Variables, data types, memory model, type conversions
- String operations, slicing, formatting, cleaning raw text data
- Lists & tuples for batch processing, ETL use-cases
- Dictionaries & sets for fast lookups, config-driven ETL
- If-else, loops, nested loops for automation flows
- Functions, modular code design for ETL components
- Lambda, map, filter, reduce for scalable transformations
- Comprehensions for optimized data processing
- Virtual environments, structuring DE Python projects
- File handling: CSV, JSON, log parsing for ETL
- Exception handling & logging for production pipelines
- OOP for reusable pipeline frameworks
- Advanced OOP—inheritance, interfaces, ETL class design patterns
- Data Warehouse fundamentals—OLTP vs OLAP, batch vs real-time analytics
- Dimensional modeling—Star schema, Snowflake schema, SCD types
- Fact & Dimension tables, surrogate keys, business modeling
- ETL vs ELT, staging zones, data quality checks, DWH architecture
- SQL basics, DDL/DML, table design for analytical systems
- Filtering, sorting, grouping large datasets efficiently
- Join types with real-world DE use cases
- Subqueries, correlated subqueries, EXISTS patterns
- Window functions—ranking, moving averages, partitions
- Advanced aggregation—cube, rollup, grouping sets
- CTEs, recursive queries, query readability
- Indexes, partitioning, clustering strategies
- Stored procedures, reusable SQL logic
- Transactions & error handling in SQL pipelines
- SQL performance tuning—execution plans & cost optimization
- Cloud SQL differences—BigQuery slots, Redshift dist/sort keys, Synapse distributions
- Analytical SQL modeling for BI
- Retail SQL Case Study
- E-commerce SQL Case Study
- Finance SQL Case Study
- Data warehouse schema design in SQL
- Implementing fact/dimension tables
- BI SQL queries for dashboards
- SQL Review & Assessment
- Spark architecture—Driver, Executors, Cluster Manager
- SparkSession, reading large datasets, RDD intro
- RDD transformations & actions with use-cases
- DataFrame operations for large-scale ETL

- Spark SQL—views, optimizations
- Joins, aggregates, UDFs in distributed systems
- Partitioning, bucketing, caching—performance tuning
- DAGs, Lineage, Execution plans
- Broadcast joins & skew-handling strategies
- Window functions in PySpark
- Advanced PySpark functions—array, struct, explode
- Delta Lake—ACID, time travel, schema evolution
- Performance tuning—shuffle reduction, caching
- Error handling in distributed jobs
- Databricks overview—workspace, clusters
- DBFS, job scheduling, notebooks
- Autoloader—incremental ingestion patterns
- Medallion Architecture: Bronze/Silver/Gold
- Delta Live Tables automation
- Unity Catalog—Governance across clouds
- End-to-end ETL pipeline on Databricks
- Structured Streaming—design patterns
- Streaming ETL—microbatch vs continuous
- Real-time project: Kafka/Events → Delta
- Pipeline orchestration using Databricks Jobs
- Databricks production patterns review
- AWS Introduction—IAM, security, identity best practices
- S3 deep dive—versioning, lifecycle, storage classes
- EC2, VPC, networking fundamentals for Data Engineers
- AWS Glue Data Catalog & Crawlers
- Glue ETL Jobs with PySpark
- AWS Lambda for event-driven ETL
- Kinesis Streams + Firehose for real-time ingestion
- Amazon Redshift—dist/sort keys, compression, workloads
- Athena—serverless SQL pipelines
- End-to-end AWS pipeline: S3 → Glue → Redshift → Athena
- Azure introduction—IAM, RBAC, resources
- ADLS Gen2—folder structures, security, lifecycle
- Azure Data Factory—linked services, datasets
- ADF pipelines
- ADF triggers
- ADF Mapping Dataflow
- Synapse Analytics—dedicated pools, serverless SQL
- Azure Databricks—Spark ETL
- Event Hub + Stream Analytics—real-time ingestion
- Azure end-to-end pipeline: ADLS → ADF → Databricks → Synapse

Quality Thought
Transforming Dreams ! Redefining Future!

- GCP introduction—IAM, service accounts, projects
- GCS buckets—lifecycle rules, security, versioning
- BigQuery architecture—storage/compute separation
- BigQuery SQL—partitioning, clustering, optimizations
- Dataflow (Apache Beam) batch pipelines
- Beam transformations—ParDo, GroupByKey, Windowing
- Pub/Sub real-time streaming ingestion
- Dataproc—Spark on GCP, workflow execution
- Vertex AI—model integration in DE pipelines
- End-to-end GCP pipeline: GCS → Dataflow → BigQuery



Quality Thought
Transforming Dreams ! Redefining Future!

89771 69236

Quality Thought Infosystems India (P) Ltd.

#302, Nilgiri Block, Ameerpet, Hyderabad-500016 | www.qualitythought.in | info@qualitythought.in