CELEBRATING 14 YEARS

QualityThought®

Transforming Dreams! Redefining Future!

Azure Data Engineer

CELEBRATING 14 YEARS

QualityThought®

Azure Data Engineer

Azure Data Factory

Basics of Cloud Computing

- What is Cloud Computing?
- Types of Cloud deployment models
  1. Private Cloud
  2. Public Cloud
  3. Hybrid Cloud
- Types of Cloud services
  1. IaaS – Infrastructure as a Service
  2. PaaS – Platform as a Service
  3. SaaS – Software as a Service

Introduction to Big Data

- What is Data?
- What is Big-Data?
- Types of Data.
  1. Structured
  2. Semi-Structured
  3. Unstructured

Introduction to Big Data

- What is Data?
- What is Big-Data?
- Types of Data.
  1. Structured
  2. Semi-Structured
  3. Unstructured
- What is Datawarehouse?
- Overview of various
- Datawarehouse architecture
- Data sources of Big-Data
- Characteristics of Big-Data
- Variety, Velocity, Volume, Veracity, Value

Introduction to Azure

- Create an Azure account
- Overview of Azure portal.
- "Overview and practical implementation of below services"

1. Subscription
2. Resource Group
3. Blob Storage, Data Lake Storage
4. Azure SQL Server, Database
5. Azure Data Factory
6. Azure Databricks
7. Azure Key Vaults
8. Azure Logic Apps
9. GitHub Repository

Data Lake Storage(ADLS)

- Create a Storage account
- Types of Storage accounts
- Create a ADLS account
- Configure Access to ADLS
- Load data to ADLS
- Read and write Data to ADLS
- Configure Backup and Disaster Recovary

Introduction to Azure SQL

- Create Azure SQL Server and Database
- Configure Elastic pools
- Configure Compute resources
- Configure Access and Security
- Configure Azure SQL Connection to Data Factory and Databricks

Azure Data Factory

- Introduction to Azure Data Factory
- Azure Data Factory UI Walkthrough
- Components of Azure Data Factory
- Integration Runtime (IR)

Azure Data Factory

- Azure Auto Integration Runtime
- Selfhosted Integration Runtime
- Azure SSIS Integration Runtime
- Create Linked Service for
  1. BLOB Storage
  2. Azure Data Lake Storage
  3. Azure SQL
  4. On-premises Server
- Create Datasets from
  1. CSV, Parquet, Excel, Avro, Json etc.
  2. Azure SQL Tables
  3. On-premises SQL Tables
- Pipelines
  1. Create a new pipeline
  2. Execute other Pipelines via REST API
  3. Debug pipeline
  4. Publish Pipeline

CELEBRATING 14 YEARS

QualityThought®

Azure Data Engineer

- Activities
  1. Copy Data
  2. Delete
  3. Stored Procedures
  4. Get-Meta Data
  5. Lookup
  6. For Each
  7. IF Condition, Switch
  8. Until
  9. Wait
  10. Fail
  11. Data Flow
  12. Set Variable, Append Variable
  13. Databricks Notebook
  14. Execute Pipelines
- Triggers
  1. Schedule Trigger
  2. Tumbling Window Trigger
  3. Storage Events
- Transformations
  1. Create Dataflow, Debug Dataflow
  2. Filter
  3. Select, Sort
  4. Aggregate, GroupBy
  5. Joins6. Lookup, Exists
  7. Union
  8. Alterrow
  9. Rank
  10. Pivot, UnPivot
  11. Use Flowlet to avoid reduce steps
- Parameters
  1. How to use parameters to dynamically manage multiple ADLS and SQL Servers, datasets, pipelines, Triggers
  2. Use parameters while pipeline execution
  3. Create Global Parameters
- Monitor Jobs
- Expression Language usage in ADF
- Send Failure notifications using Logic Apps
- Manage credentials using Azure KeyVault
- Repository, Change Management
  1. Create GitHub Repository
  2. Migrate ARM templates using Git
  3. ARM Templates – Export/Import Manually

- Monitor Jobs
- Expression Language usage in ADF
- Send Failure notifications using Logic Apps
- Manage credentials using Azure KeyVault
- Repository, Change Management
  1. Create GitHub Repository
  2. Migrate ARM templates using Git
  3. ARM Templates – Export/Import Manually

Databricks + Pyspark

Introduction to Azure Databricks

- Introduction to Spark
  1. Overview of Spark Architecture
  2. RDD Vs DataFlow Vs Dataset
  3. Transformations & Actions
- Introduction to DataBricks
  1. Create Databricks Workspace
  2. Create Clusters
  3. Create Databricks Notebooks
  4. Azure KeyVault Integration
- Databricks File System (DBFS)
  1. Create, copy, move files within DBFS
  2. Handle multiple files and folders
  3. Archive files in DBFS
- Databricks Utilities (dbUtils)
  1. File system
  2. Secrets
  3. Notebook
  4. Widgets
- Integrate Databricks with External resources
  1. Create Mount point with ADLS, Storage accounts, Azure SQL, Synapse etc..
  2. Read and write data from ADLS, SQL
- Delta Lakes1. DeltaLake Overview, Architecture
  2. Diff between DataLake & DeltaLake
  3. How to create DeltaLake Tables
  4. DML operations using Delta Tables
  5. How to manage SCD Type 1 and Type 2
  6. History Logs and Restore the Tables
- Optimization

1. Cost optimization Techniques overview
2. Catalyst optimizer, Cache, Persist

CELEBRATING 14 YEARS

QualityThought®

Azure Data Engineer

Pyspark

- Introduction to Python
- Variables, Datatypes, Operators
- Introduction to PySpark
- Read and write data from CSV, Json, Parquet, Azure SQL, DBFS, ADLS etc.
- Transformations using PySpark
  1. Cast
  2. Select
  3. Filter
  4. Sort
  5. Aggregations
  6. Join
  7. Union
  8. Remove Duplicates
  9. Calculated Columns, Rename columns
  10. Window Functions
  11. String Functions
  12. Date Functions
  13. Conditional Statements
  14. Loops
  15. User Defined Functions
  16. Expression Language
- Run SQL queries in DataBricks using Spark SQL

Synapse

Introduction to Azure Synapse

- Overview of Synapse Architecture
- Create Azure Synapse Account
- Configure access to Azure Synapse

Overview of Pools in Synapse

- Serverless SQL Pool
- Dedicated SQL Pool
- Apache Spark Pool4. Data Explorer Pool

Integration with DataLake Storage

- Load data to ADLS via Synapse UI
- Query data using SQL scripts
- Create a External tables from CSV and parquet file in ADLS

Connect to External Resources

- Load data to ADLS via Synapse UI
- Query data using SQL scripts
- Create a External tables from CSV and parquet file in ADLS

Connect to External Resources

- Create External File Format
- Create External Data source
- Create External Table
- Create Views

Use multiple languages in Synapse notebook using magic commandsOverview of Mssparkutils

Introduction to Spark in Synapse

- Create a notebook in Synapse
- Configure Cluster, Autoscaling
- Create Mount point to connect ADLS, storage accounts, Azure SQL and other external databases
- Transformations using Synapse

Integration with Delta Lakes

- Create Tables from Delta Tables
- Create views from Delta Tables

Monitor and Logging

- Monitor the pipelines
- Notify the failure message using Logic Apps

Monitor and Logging

- Monitor the pipelines
- Notify the failure message using Logic Apps

Integrate credentials using Azure Key Vault
Use parameters to integrate multiple Pipelines datasets, triggers, Linked service, notebooks etc.

Connecting Azure Synapse Views with BI tools(Power BI)

Realtime Scenarios: ADF

- Create Azure, Selfhosted Integration runtime
- Creating linked services
- Creating Dynamic Linked services using the Parameters
- Create Datasets
- Parameters - dynamically use a single dataset for multiple SQL servers
- Parameters - dynamically use a single dataset for multiple storage/ADLS accounts
- Copy activity – BLOB to BLOB
- Copy activity – BLOB to azure SQL
- Copy activity – copy based on wildcard search
- Copy activity – copy the filtered file formats
- Copy activity – copy multiple files from blob to another blob
- Copy activity – Delete source files after copy activity
- How to pass parameters to the pipeline
- Convert one file format to another file format
- Delete the files from blob with does not have the data
- How to use getmetdata activity
- Integrate keyvault in ADF
- How to use databricks activity activity and pass paraemeters to it
- Validate file schema before processing the pipeline
- Execute copy activity if source file is available, Else send a failure notification
- Re-run the pipeline automatically after 1 hour incase of first time failure
- Dynamically copy the data based on daily file list
- Copy data from onpremises to Bronze layer
- Incrementally copy new and changed data in destination table
- Log successful execution of activity to SQL table
- Dynamically pass parameters from ADF to SQL Stored procedures
- Log pipeline failure in SQL table
- How to stop a pipeline to going into infinity loops
- Dataflows – select the rows
- Dataflows – Filter the rows
- Dataflows – join Transformations
- Dataflows – union Transformations
- Dataflows – look up Transformations
- Dataflows – window functions transformations
- Dataflows – pivot, unpivot transformations
- Dataflows – Alter rows transformations
- Dataflows – Removing Duplicates transformations
- Use Execute pipeline to run other pipelines from same DataFactory
- Run other datafactory pipeline using the REST API
- Send email alert when a pipeline fails
- How to send mail notifications using logic apps
- How to create alerts and rules

- How to set global parameters
- Log the pipeline and runid details in log file
- How to import and export ARM templates
- How to integrate ADF with Devops
- How to use GIT repository
- Monitor the pipelines
- Debug the pipelines
- How to create a scheduling trigger
- Create tumbling window trigger to execute the historical dates data
- Execute a pipeline based on Storage account activity
- Create variables using set variable activity
- How to use if condition using if condition activity
- Iterating files using for loop activity

Real Time Scenarios: Databricks

- Create a mount point connection to ADLS
- Create a mount point connection to Azure SQL
- Create a Dataframe from a CSV file
- Create a Dataframe from a parquet file
- Create a Dataframe from Json File
- Change the datatype of a column(String to Date, String to Integer)

CELEBRATING 14 YEARS

Quality Thought®

Our Students Got Placement at

tcs | Capgemini | LTIMindtree | accenture | 66degrees | AFFINE | High Noon Consulting

AGILISIUM | ATIMETRIK | brillio | CGI | CISCO | Cloud4C | amazon

CLOUD9 | CloudCover | Coforge | Cognizant | Convergytics | CSS CORP | Medtronic

databricks | dataflex | DXC TECHNOLOGY | Deloitte. | EMIDS | Foray | STIC Soft

genpact | HCL | HSBC | i2o | IBM | CAPRUS IT

Infosys | Legato | MINDGRAPH | Mindtree | motivitylabs | photon | mesleva

pwc | quantiphi | RENAULT | SIGMASOFT Virtual Molding | SOFTILITY | syren cloud | splashBI

TIGER ANALYTICS | wipro | ENVISION | virtusa | ValueLabs | zensor | SOLIX

JRD SYSTEMS | ValueLabs | archents | mapsoft | Innominds Powering the Digital Next | techforce.ai | techolution

SOLUGENIX | Nendrasys | S&P Global | Aadvi | iPrism | ENVISION | Sunseaz Technologies Pvt Ltd

QualityThought

90595 14148

Quality Thought Infosystems India (P) Ltd.

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