SimpleGuru

Azure

Data Engineer

+91 9347664499 +91 7799244988 | info@simpleguru.in

SimpleGuru

Azure Data Engineer

Basics of Cloud Computing

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

Introduction to Big Data

- What is Data?
- What is Big-Data?
- Types of Data
  a. Structured
  b. Semi-Structured
  c. 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
  a. Subscription
  b. Resource Group
  c. Blob Storage, Data Lake Storage
  d. Azure SQL Server, Database
  e. Azure Data Factory
  f. Azure Databricks
  g. Azure Key Vaults
  h. Azure Logic Apps
  i. 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

Azure Data Factory(ADF)

- Introduction to Azure Data Factory
- Azure Data Factory UI Walkthrough
- Components of Azure Data Factory
- Integration Runtime (IR)
  a. Azure Auto Integration Runtime
  b. Selfhosted Integration Runtime
- Create Linked Service for
  a. BLOB, Azure Data Lake Storage
  b. Azure SQL
  c. On-premises Server
- Create Datasets from
  d. CSV, Parquet, Excel, Avro, Json etc.
  e. Azure and On-premises SQL Tables
- Pipelines
  f. Create a new pipeline
  g. Execute other Pipelines via REST API
  h. Debug pipeline
  i. Publish Pipeline
- Activities
  a. Copy Data
  b. Delete, Stored Procedures
  c. Get-Meta Data, Lookup
  d. For Each, IF Condition, Switch, Until
  e. Wait, Fail, Data Flow
  f. Set Variable, Append Variable
  g. Databricks Notebook
  h. Execute Pipelines
- Triggers
  a. Schedule Trigger
  b. Tumbling Window Trigger
  c. Storage Events
- Transformations
  a. Create Dataflow, Debug Dataflow
  b. Filter, Select, Sort, Aggregate, GroupBy
  c. Join, Lookup, Exists, Union, Alterrow
  d. Rank, Pivot, UnPivot
  e. Use Flowlet to avoid reduce steps.
- Parameters
  a. How to use parameters to dynamically manage multiple ADLS and SQL Servers, datasets, pipelines, Triggers
  b. Use parameters while pipeline execution
  c. Create Global Parameters
- Monitor Jobs
- Expression Language usage in ADF
- Send Failure notifications using Logic Apps
- Manage credentials using Azure KeyVault
- Repository, Change Management
  a. Create GitHub Repository
  b. Migrate ARM templates using Git
  c. ARM Templates - Export/Import Manually

+91 9347664499 +91 7799244988 | info@simpleguru.in

SimpleGuru

Azure Data Engineer

Azure Databricks

- Introduction to Spark
  a. Overview of Spark Architecture
  b. RDD Vs DataFlow Vs Dataset
  c. Transformations & Actions
- Introduction to DataBricks
  a. Create Databricks Workspace
  b. Create Clusters
  c. Create Databricks Notebooks
  d. Azure KeyVault Integration
- Databricks File System (DBFS)
  e. Create, Copy, Move files within DBFS
  f. Handle multiple files and folders
  g. Archive files in DBFS
- Databricks Utilities (dbUtils)
  a. File system, Secrets, Notebook, Widgets
- Integrate Databricks with External resources
  b. Create Mount point with ADLS, storage accounts, Azure SQL, Synapse etc..
  c. Read and write data from ADLS, SQL
- Delta Lakes
  a. DeltaLake Overview, Architecture
  b. Diff between DataLake & DeltaLake
  c. How to create DeltaLake Tables
  d. DML operations using Delta Tables
  e. How to manage SCD Type 1 and Type 2
  f. History Logs and Restore the Tables
- Optimization
  a. Cost optimization Techniques overview
  b. Catalyst optimizer, Cache, Persist

PySpark Programming

- 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
  a. Cast, Select, Filter, Sort, Aggregations
  b. Join, Union, Remove Duplicates
  c. Calculated Columns, Rename columns
  d. Window Functions
  e. String Functions
  f. Date Functions
  g. Conditional Statements
  h. Loops
  i. User Defined Functions
  j. Expression Language
- Run SQL queries in DataBricks using Spark SQL

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 Synapse

- Introduction to Azure Synapse
  a. Overview of Synapse Architecture
  b. Create Azure Synapse Account
  c. Configure access to Azure Synapse
- Overview of Pools in Synapse
  a. Serverless SQL Pool
  b. Dedicated SQL Pool
  c. Apache Spark Pool
  d. Data Explorer Pool
- Integration with DataLake Storage
  a. Load data to ADLS via Synapse UI
  b. Query data using SQL scripts
  c. Create a External tables from CSV and parquet file in ADLS
- Connect to External Resources
  a. Create External File Format
  b. Create External Data source
  c. Create External Table
  d. Create Views
- Use multiple languages in Synapse notebook using magic commands
- Overview of Mssparkutils
- Introduction to Spark in Synapse
  a. Create a notebook in Synapse
  b. Configure Cluster, Autoscaling
  c. Create Mount point to connect ADLS, storage accounts, Azure SQL and other external databases
  d. Transformations using Synapse
- Integration with Delta Lakes
  g. Create Tables from Delta Tables
  h. Create views from Delta Tables
- Monitor and Logging
  c. Monitor the pipelines
  d. 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.

+91 9347664499 +91 7799244988 | info@simpleguru.in

SimpleGuru

Azure Data Engineer

Unity Catalog

- Unity Catalog Overview & Architecture.
- Databricks Workspace Catalog vs. Unity Catalog
- Setting Up Unity Catalog in Azure Databricks
- Managing Data Governance with Unity Catalog
- Creating & Managing Catalogs, Schemas, Tables & Views.
- Access Control and Role-Based Permissions in Unity Catalog
- Managing External Tables and Storage Credentials
- Secure Data Sharing Across Workspaces (Delta Sharing with Unity Catalog)
- Best Practices for Data Security and Compliance

Realtime Scenarios(ADF/Synapse)

1.Create Azure SQL Server and Database

- Setting up Self-hosted Integration Runtime for accessing on-premises data
- Creating Linked Services for different data sources like SQL Server, Azure Blob, and ADLS
- Implementing dynamic Linked Services to avoid duplication across multiple environment
- Securely integrating ADF with Azure Key Vault for credential management

2. Dataset Management & Parameterization

- Creating a single dynamic dataset for multiple SQL servers using parameters
- Configuring a parameterized dataset to handle multiple Storage/ADLS accounts
- Using expression-based datasets to construct file paths dynamically

3. Copy Activity (ETL Operations)

- Copying structured data from Blob to SQL with auto schema mapping
- Copying data incrementally using watermark columns in SQL tables
- Copying data from multiple sources dynamically based on metadata tables
- Copying only specific file types (CSV, JSON) from source to destination
- Deleting source files after copy to avoid redundant processing
- Using wildcards to copy files from different date partitions
- Pass parameters dynamically to copy activity for multiple source-target pairs.
- Store pipeline parameters in a database table and fetch them dynamically.

4. Pipeline Execution & Error Handling

- Passing runtime parameters dynamically from an external source
- Executing a pipeline conditionally based on external triggers
- Retrying pipeline execution if it fails due to transient errors
- Sending failure notifications when a pipeline fails
- Skipping corrupted files and continuing the pipeline execution

5. Incremental & Delta Load Processing

- Incrementally copying new data using LastModifiedDate filter
- Merging new and changed data from SQL to Delta Lake
- Implementing SCD Type 1 & Type 2 using ADF Mapping Data Flows
- Maintaining historical data versioning in SQL for auditing
- Use ADF parameters to dynamically execute different stored procedures.

6. Data Flow Transformations

- Performing joins between two datasets dynamically
- Implementing pivot and unpivot operations for reporting
- Removing duplicate records before inserting into SQL
- Applying window functions for ranking or running totals
- Filtering only required rows before ingestion

7. Orchestration & Automation

- Triggering a pipeline when a file is uploaded to Blob Storage
- Re-executing failed pipelines after an hour automatically
- Stopping infinite loops caused by recursive pipeline triggers

8. Metadata Handling & File Management

- Extracting metadata of incoming files before processing
- Validating file schema before loading into SQL
- Archiving processed files to a separate folder
- Deleting empty files from Azure Blob

9. Monitoring & Logging

- Logging execution details (pipeline ID, timestamps) to a database
- Sending real-time alerts when pipeline execution exceeds threshold
- Monitoring pipeline failures and retrying automatically

10. DevOps & Security

- Using GIT integration to version-control ADF pipelines
- Deploying ADF pipelines across environments using ARM templates
- Managing access control using RBAC (Role-Based Access Control)

Realtime Scenarios (Databricks)

1. Data Ingestion & Processing

- Reading structured and unstructured data from ADLS, Blob, and SQL
- Performing schema evolution when ingesting JSON/Parquet files
- Auto-detecting file formats (CSV, JSON, Avro) before processing
- Handling corrupt records during ingestion
- Implementing Delta Live Tables for automated data pipelines

2. Delta Lake & Data Versioning

- Creating Delta tables from existing Parquet data
- Implementing Time Travel to retrieve old versions of data
- Upserting data into Delta tables using MERGE
- Implementing SCD Type 1 & Type 2 in Delta Tables

3. Data Transformation (ETL in Databricks)

- Using Window Functions for ranking and aggregations
- Handling NULL values and missing data efficiently
- Dynamically renaming and filtering columns in a DataFrame
- Implementing custom transformations using UDFs
- Build a Bronze layer to store raw data from multiple sources.
- Implement a Silver layer for data cleansing and standardization.
- Create a Gold layer for business reporting and aggregations.
- Use ADF to orchestrate Medallion Architecture workflows.

4. Orchestration & Integration

- Triggering Databricks jobs from ADF
- Integrating Databricks with Azure Event Hub for streaming
- Passing parameters dynamically to Databricks notebooks