Data Engineering Introduction:

- ✓ What is data engineering
- ✓ Introduction to Azure
- ✓ Comparison of Azure vs AWS Data Engineering services
- ✓ Bigdata introduction

Azure Basics:

- ✓ What is Azure and Cloud
- ✓ How to Create Resource Group in Azure
- ✓ Services Offered in Azure
- ✓ Azure Portal Walk Through
- ✓ SDKs or Tools for Azure Resources
- ✓ Create Free Azure Subscription

Python Overview

- ✓ History & features of Python
- ✓ Python vs Other Programming Languages
- ✓ First Python Program
- ✓ Python basic syntax
- ✓ Python Development Tools & Packages

Python Environment Setup

- ✓ Installing Python
- ✓ Verify & Setup Python environment

Python DataTypes

- ✓ Variables
- ✓ Data Types
- ✓ Strings
- ✓ Type Casting

Python Operators

- ✓ Arithmetic Operators
- ✓ Relational Operators
- ✓ Logical Operators
- ✓ Bitwise Operators
- ✓ Assignment Operator

Python Flow Control & Loops

- ✓ if, if else, if else if
- ✓ while, do while loops
- ✓ for loops

Python Functional Programming

- ✓ Function Declarations
- ✓ Calling Functions
- ✓ Functions Call-by-Name
- ✓ Functions with Named Arguments
- ✓ Functions with Variable Arguments
- ✓ Functions with Default Parameter Values
- ✓ Lambda Functions

Python Collections

- ✓ Lists
- ✓ Tuples
- ✓ Sets
- ✓ Dictionaries

Python Files I/O

- ✓ Reading input from console
- ✓ Reading data from Files
- ✓ Writing data to Files

Python Object Oriented Programming Python Classes

- ✓ Simple class
- ✓ Class objects
- ✓ Inheritance

Python Exception Handling

- ✓ Throwing Exceptions
- ✓ try, catch, finally
- ✓ Catching Exceptions
- ✓ The finally Clause

Python Miscellaneous

- ✓ Modules
- ✓ Dates
- ✓ RegEx

Spark:

- ✓ Spark Introduction
- ✓ Spark Overview
- ✓ Spark features
- ✓ Spark vs Hadoop MapReduce
- ✓ Programming Language choices in Spark
- ✓ Spark History
- ✓ Spark use cases

Spark Components or modules

- ✓ Spark Core
- ✓ Spark SQL
- ✓ Spark Streaming

Spark Architecture

- ✓ Spark Application flow
- ✓ Spark Driver, Executors
- ✓ Spark Context, Spark Session
- ✓ Spark dependency on Cluster Managers
- ✓ Spark execution modes
- ✓ Standalone cluster mode
- ✓ Spark on YARN mode

Spark Core

- ✓ Spark's main Data Abstraction
- ✓ About RDD
- ✓ RDD Features
- ✓ Creating RDDs
- ✓ Saving Files
- ✓ Data Manipulation using RDDs
- ✓ Transformations & Actions
- ✓ RDD Partitions & Coalesce
- ✓ Memory Management: cache & persist
- ✓ Data Loading and Saving through RDDs
- ✓ Aggregations, Joins through RDDs
- ✓ RDD Advanced concepts – Accumulators, Broadcast variables

Spark Structured APIs

- ✓ DataFrames
- ✓ Columns
- ✓ Rows
- ✓ Spark Types
- ✓ performance optimization with Spark Structured APIs
- ✓ Logical planning, Physical planning, and Execution

Spark DataFrames

- ✓ DataFrames basics
- ✓ Creating DataFrames
- ✓ Schemas
- ✓ DataFrame Operations
- ✓ Column wise operations
- ✓ Row wise operations

- ✓ Aggregations DataFrames
- ✓ Joins using DataFrames

Data Sources

- ✓ Reading & writing different files formats
- ✓ CSV Files
- ✓ JSON Files
- ✓ Parquet Files
- ✓ ORC Files
- ✓ SQL Databases
- ✓ TextFiles

Spark SQL

- ✓ Bigdata & SQL: Apache Hive vs Spark SQL
- ✓ Catalog, Tables, Views, Databases
- ✓ Data selection and manipulation using Spark SQL
- ✓ User Defined Functions
- ✓ Spark SQL integration with Hive

PySpark Application Developer Tools

- ✓ pyspark interative shell
- ✓ PyCharm
- ✓ Spyder
- ✓ Jupyter Notebooks
- ✓ Zeppelin
- ✓ other developer tools

Executing PySpark Application

- ✓ Local mode
- ✓ Client mode
- ✓ Cluster mode
- ✓ Spark UI
- ✓ Monitoring Spark Applications
- ✓ Spark logs
- ✓ Spark Application Resource selection

Spark Streaming

- ✓ About Batch vs Streaming processing
- ✓ About Spark DStreams
- ✓ About Spark Structured streaming
- ✓ Transformation of Streams data
- ✓ Streaming sources & sinks
- ✓ Event Time
- ✓ Stateful processing

Azure Storage and Data Lake:

- ✓ Detail knowledge of Blob
- ✓ Create a container
- ✓ Upload a blob to Azure Storage
- ✓ Create a Blob
- ✓ List of the blobs in a container
- ✓ Delete a container
- ✓ Download the blob to your local computer
- ✓ Detail knowledge of Data Lake
- ✓ Create Azure Data Lake Gen 2 Account
- ✓ Create Folders
- ✓ Upload data
- ✓ Secure data
- ✓ Delete Azure Data Lake

Azure Key Vault:

- ✓ Introduction to Azure Key Vault
- ✓ Store Secrets in Azure Key Vault using Azure Portal

Azure Data Factory:

- ✓ Introduction to Azure Data Factory
- ✓ Top level Concepts in Azure Data Factory
- ✓ Create your First Azure Data Factory
- ✓ Different ways to work with Azure Data Factory
- ✓ Pipelines and Activities
- ✓ Linked Services and Datasets
- ✓ Triggers
- ✓ Schedule Trigger
- ✓ Tumbling Window Trigger
- ✓ Tumbling Window Trigger Dependency
- ✓ Event based Triggers
- ✓ Integration runtime
- ✓ Azure Integration runtime
- ✓ Self-Hosted Integration runtime
- ✓ Setting up Self Hosted Integration runtime
- ✓ Shared Self Hosted Integration runtime
- ✓ Parameterize Linked Services
- ✓ Parameterize Datasets
- ✓ Parameterize Pipelines
- ✓ System Variables

- ✓ Connectors Overview
- ✓ Supported File Formats
- ✓ Copy Data Activity
- ✓ Monitor Copy Data Activity
- ✓ Delete Activity
- ✓ Variables
- ✓ Set Variable Activity
- ✓ Append Variable Activity
- ✓ User Properties
- ✓ Execute Pipeline Activity
- ✓ Filter Activity
- ✓ ForEach Activity
- ✓ Get Metadata Activity
- ✓ If Condition Activity
- ✓ Wait Activity
- ✓ Until Activity
- ✓ Web Activity
- ✓ WebHook Activity
- ✓ Switch Activity
- ✓ Validation Activity
- ✓ Lookup Activity
- ✓ Transform Data Activities Overview
- ✓ Stored Procedure Activity
- ✓ Data flow
- ✓ Mapping Data Flow
- ✓ Data Flow Activity
- ✓ Mapping Data Flow Debug Mode
- ✓ Filter Transformation in Mapping Data Flow
- ✓ Aggregate Transformation in Mapping Data Flow
- ✓ JOIN Transformation in Mapping Data Flow
- ✓ Conditional Split Transformation in Mapping Data Flow
- ✓ Derived Column Transformation in Mapping Data Flow
- ✓ Exists Transformation in Mapping Data Flow
- ✓ Union Transformation in Mapping Data Flow
- ✓ Lookup Transformation in Mapping Data Flow
- ✓ Sort Transformation in Mapping Data Flow
- ✓ New Branch in Mapping Data Flow
- ✓ Select Transformation in Mapping Data Flow
- ✓ Pivot Transformation in Mapping Data Flow
- ✓ Unpivot Transformation in Mapping Data Flow
- ✓ Surrogate Key Transformation in Mapping Data Flow
- ✓ Window Transformation in Mapping Data Flow

- ✓ Alter Row Transformation in Mapping Data Flow
- ✓ Flatten Transformation in Mapping Data Flow
- ✓ Parameterize Mapping Data Flow
- ✓ Validate Schema in Mapping Data Flow
- ✓ Schema Drift in Mapping Data Flow
- ✓ Wrangling Data Flow Overview
- ✓ Merge Queries in Wrangling Data Flow
- ✓ Group By in Wrangling Data Flow
- ✓ Different Author Modes
- ✓ Setup GitHub Code Repository for Azure Data Factory
- ✓ Setup Azure DevOps Git Code Repository
- ✓ Use Azure Key Vault Secrets
- ✓ Continuous integration and deployment
- ✓ How to read JSON output of one Activity in to another Activity
- ✓ Annotations
- ✓ Templates Overview
- ✓ Global Parameters
- ✓ Rank Transformation in Mapping Data Flow
- ✓ Cache Sink and Cached lookup in Mapping Data Flow
- ✓ Session log in Copy Activity | Log Copied File names in Copy Activity
- ✓ Write Cache Sink to Activity Output
- ✓ Parse Transformation in Mapping Data Flow
- ✓ Fail Activity
- ✓ Inline Dataset
- ✓ Stringify transformation in Mapping Data Flow
- ✓ Assert Transformation in Mapping Data Flows
- ✓ Flowlets in Mapping data flow
- ✓ Script Activity in Azure Data Factory or Azure Synapse
- ✓ User defined Functions in Mapping data flows
- ✓ Fuzzy Joins Using mapping data flows
- ✓ Parameterize Linked Services using
- ✓ Cast Transformation in Mapping data flows
- ✓ Extract Data from table of website page
- ✓ Per Pipeline Billing View for Azure Data factory
- ✓ Time To Live(TTL) Setting in Azure IR to reduce cluster spin up time for dataflows
- ✓ Create Alert rules in Azure Data factory for Pipeline or activity Failures
- ✓ Pipeline return value in Set variable

Azure Synapse Analytics:

- ✓ Introduction to Azure Synapse Analytics
- ✓ Create Azure Synapse Analytics Workspace
- ✓ Basic Concepts in Azure Synapse Analytics
- ✓ Analyze data with a server less SQL pool and dedicated SQL Pool
- ✓ Analyze data with Server less Spark Pool
- ✓ Analyze data in Storage Account in Azure Synapse Analytics
- ✓ Integrate Pipelines in Azure Synapse Analytics
- ✓ Monitor your Azure Synapse Analytics Workspace
- ✓ Add an Administrator to your Azure Synapse Workspace
- ✓ Azure Synapse SQL Architecture
- ✓ Distributions(Hash, Round Robbin & Replicate)
- ✓ Server less SQL Pool Overview
- ✓ Create External Data source
- ✓ Create External File Format
- ✓ CETAS with Synapse SQL
- ✓ CTAS with Synapse SQL
- ✓ External Tables with Synapse SQL
- ✓ Create and query external tables from a file in ADLS
- ✓ Types of External Tables(Hadoop & Native) in Synapse SQL
- ✓ Administrative accounts in Synapse SQL
- ✓ Create Login and User for Server less SQL Pool
- ✓ Create Login and User for Dedicated SQL Pool
- ✓ Temporary Tables in Synapse SQL
- ✓ Using IDENTITY to create surrogate keys using dedicated SQL pool
- ✓ OPENROWSET() function in Synapse SQL
- ✓ Apache Spark in Azure Synapse Analytics
- ✓ Create a Spark Pool with Azure Portal
- ✓ Create a Notebook in Azure Synapse Analytics
- ✓ Pandas to read/write Azure Data Lake Storage Gen2 data in Apache Spark pool in Synapse Analytics
- ✓ Use FSSPEC to read/write ADLS Gen2 data in Apache Spark pool in Synapse Analytics
- ✓ Different ways to Create Notebooks in Azure Synapse Analytics
- ✓ Use multiple languages in Synapse notebook using magic commands

- ✓ Use temp tables to reference data across languages in Synapse notebooks in Azure Synapse
- ✓ %run command to reference another notebook with in current notebook
- ✓ Parameterize Synapse notebook in Azure Synapse Analytics
- ✓ Run Synapse notebook from pipeline | Pass values to Notebook parameters from pipeline in Synapse
- ✓ IPython Widgets in Synapse Notebook in Azure Synapse Analytics | ipywidgets in Synapse notebook
- ✓ Introduction to Microsoft Spark Utilities(MSSparkutils)
- ✓ Microsoft Spark File System(mssparkutils.fs) Utilities in Azure Synapse Analytics
- ✓ Microsoft Spark Utilities Notebook Utilities(mssparkutils.notebook)
- ✓ exit() function of notebook module in MSSparkUtils package
- ✓ run() function of notebook module in MSSparkUtils package
- ✓ Environment utilities (mssparkutils.env) in MSSparkUtils
- ✓ Configure access to Azure Data Lake Gen2(ADLS Gen2) for Synapse Notebook
- ✓ MSSparkUtils Runtime Utils in Synapse Notebook
- ✓ How to Mount ADLS Gen2 Storage using Linked Service
- ✓ How to Mount ADLS Gen2 Storage using Account Key or SAS Token in Synapse Notebook
- ✓ How to Unmount the Mount Point in Synapse Spark
- ✓ Creating a Spark Job Definition and Submitting it in Azure Synapse
- ✓ Manage Library Packages for Apache Spark
- ✓ Workspace packages for Apache Spark Pool
- ✓ requirements.txt File to Manage libraries for Apache Spark pool

Azure Data Bricks:

- ✓ Introduction to Azure Databricks
- ✓ Create an Azure Databricks Workspace using Azure Portal
- ✓ Create Databricks Community Edition Account
- ✓ Workspace in Azure Databricks
- ✓ Workspace assets in Azure Databricks
- ✓ Working with Workspace Objects in Azure Databricks
- ✓ Create and Run Spark Job in Databricks
- ✓ Azure Databricks architecture overview
- ✓ Databricks File System(DBFS) overview in Azure Databricks
- ✓ Databricks Utilities(dbutils) in Azure Databricks
- ✓

- ✓
- ✓ Data Utility(dbutils.data) in Azure Databricks in Databricks utilities
- ✓ File System utility(dbutils.fs) of Databricks Utilities in Azure Databricks
- ✓ exit() command of notebook utility(dbutils.notebook) in Azure Databricks
- ✓ run() command of notebook utility(dbutils.notebook) in Databricks Utilities in Azure Databricks
- ✓ Widgets utility(dbutils.widgets) of Databricks Utilities in Azure Databricks
- ✓ Pass values to notebook parameters from another notebook using run command in Azure Databricks
- ✓ Parameterize SQL notebook using widgets in Azure Databricks | Widgets in SQL in Azure Databricks
- ✓ Create Mount point using dbutils.fs.mount() in Azure Databricks
- ✓ Mount Azure Blob Storage to DBFS in Azure Databricks
- ✓ Delete or Unmount Mount Points in Azure Databricks
- ✓ mounts() & refreshMounts() commands of File system Utilities in Azure Databricks
- ✓ Update Mount Point(dbutils.fs.updateMount()) in Azure Databricks
- ✓ Secret Scopes Overview in Azure Databricks
- ✓ Install Databricks CLI and configure with your workspace | Azure Databricks
- ✓ Create an Azure Key Vault backed secret scope using the UI in Azure Databricks
- ✓ Create a Databricks backed secret scope in Azure Databricks
- ✓ Secrets Utility(dbutils.secrets) of Databricks Utilities in Azure Databricks
- ✓ Access ADLS Gen2 storage using Account Key in Azure Databricks
- ✓ Configure access to Azure storage with an Azure Active Directory service principal
- ✓ Access Data Lake Storage Gen2 or Blob Storage with an Azure service principal in Azure Databricks
- ✓ Access ADLS Gen2 or Blob Storage using a SAS token in Azure Databricks

Delta Lake usage in Databricks:

- ✓ Architecture
- ✓ Storage Understanding
- ✓ Table creation and API options
- ✓ DML Operations usage.
- ✓ Partitions
- ✓ Schema Enforcement
- ✓ Schema Evolution
- ✓ Versions
- ✓ Time Travel
- ✓ Vaccum
- ✓ Delta Lake Merge (SCD Type 1 and SCD Type2)

Azure Application Insights:

- ✓ Azure Application Insights Tutorial
- ✓ What is Azure Application Insights?
- ✓ Monitoring without code or Codeless Monitoring
- ✓ Monitoring with Code or Code based Monitoring
- ✓ Features of Azure Application Insights
- ✓ What exactly does Azure Application Insights monitor?
- ✓ How to create Azure Application Insights?
- ✓ Prerequisites
- ✓ How to enable Application Insights for Application?

Projects

Interview Questions

Recordings

Materials