# ADE (AZURE DATA ENGINEERING) WITH SPARK (PYSPARK)

**Mr. GOPAL KRISHNA, Sr. Data & Cloud Architect,**
**18+ Years Of Real Time IT Exp, 11+ Years On BIGDATA Implementation Projects Exp**
**CLOUDERA CCA 175 & Microsoft Certified Consultant**

## ADE (Azure Data Engineering) with SPARK (PYSPARK) Course Content

### Introduction to Cloud Computing

- Need for Cloud Computing?
- Essentials of Cloud Computing?
- Cloud Deployment Models
  - Public cloud
  - Private cloud
  - Hybrid cloud
- Varieties of Cloud Services
  - PaaS [ Platform as a Service ]
  - IaaS [Infrastructure as a Service].
  - SaaS [ Software as a Service ]
  - AaaS [ Anything as a Service ]
- Real world examples of Cloud Computing
- On-premises Vs Cloud Integrated project implementation.

### Data Warehousing Basics

- Overview of Data Warehouse
- Overview of ETL pipeline in real world
- ETL(Extract, Transform & Load) VS ELT( Extract, Load & Transform)
- Data Warehouse Architecture

### Azure Cloud – at a Glance

- Overview of Azure Portal
- Different Services in Azure Portal
- Create an Azure Account.
- Azure Subscription
- Azure Subscription Types
  - Free tier
  - Pay-As-You-Go(PAYG)
  - Enterprise Agreement(EA)
- What is Resource Group (RG)? Why is it essential?
- What are Azure Regions & Region Pairs
- Azure Availability Zones

# ADE (AZURE DATA ENGINEERING) WITH SPARK (PYSPARK)

**Mr. GOPAL KRISHNA, Sr. Data & Cloud Architect,**
**18+ Years Of Real Time IT Exp, 11+ Years On BIGDATA Implementation Projects Exp**
**CLOUDERA CCA 175 & Microsoft Certified Consultant**

### Azure Storage

- Azure Storage Account Creation
- Types of Azure Storage Accounts
  - Blob Storage
  - File Storage
  - Queue Storage
  - Table Storage
  - ADLS Gen2 Storage
- What are the advantages & disadvantages of Azure storage accounts?
- What is ADLS [Azure Data Lake Storage] Gen2 ?
- Need for Azure Data Lake Storage(ADLS) Gen2
- How to create an ADLS Account?
- How to Load the data to ADLS?
- Read & Write the data to ADLS

### Azure SQL

- Introduction to Azure SQL DB
- Create Azure SQL Server and Database
- What is Elastic Pool in Azure SQL?
- Configuring Elastic Pool
- Different Service Tiers
- DTU (Database Transaction Units) Vs vCores.
- High Availability and Disaster Recovery in Azure SQL DB

# PYSPARK on DATABRICKS

### PYTHON BASICs

- Introduction to Python
- Why Python is a "Dynamically Typed Language"?
- Variable Declaration in Python
- Conditional Statements in Python
  - If…else
  - If …else..if
- Python Loops
  - While loop
  - For loop
- How to pass Command line arguments in Python

# ADE (AZURE DATA ENGINEERING) WITH SPARK (PYSPARK)

**Mr. GOPAL KRISHNA, Sr. Data & Cloud Architect,**
**18+ Years Of Real Time IT Exp, 11+ Years On BIGDATA Implementation Projects Exp**
**CLOUDERA CCA 175 & Microsoft Certified Consultant**

- Collections in Python
  - List
  - Set
  - Dictionaries
- Python Lambda
- Python Functions
- Python Try…Except
- How to read JSON Data in Python
- Overview of Python OOP

### DATABRICKS

- Introduction to Datariicks
- Different Components of Databricks
  - Workspace
  - Clusters
  - Notebooks
  - Databricks Runtime(DBR)
- How to create Databricks Cluster with different parameters
- How to create Notebooks with different language options
- Introduction to Databricks File System(DBFS)
- Databricks Integrations
  - Azure Data Lake Storage(ADLS)
  - Azure SQL
  - Different File Formats
- Data Lake VS Delta Lake
- Delta Lake Table creation
- Database Operations on Delta Lake tables

## PYSPARK

- Motivation for Spark
- Advantages of IN_MEMORY Processing over DISK Based
- Where to use Spark
- ROI Comparison of Spark Processing Vs Other BigData Frameworks.
- Why Spark Processing is Faster than other BigData Frameworks.
- **Spark Architecture**
  - Spark Master
  - Spark Driver
  - Spark Executor

# ADE (AZURE DATA ENGINEERING) WITH SPARK (PYSPARK)

**Mr. GOPAL KRISHNA, Sr. Data & Cloud Architect,**
**18+ Years Of Real Time IT Exp, 11+ Years On BIGDATA Implementation Projects Exp**
**CLOUDERA CCA 175 & Microsoft Certified Consultant**

  - Spark Worker Node
  - Spark Cluster Manager
- **Different Types of Cluster Manager in Spark**
  - Standalone
  - YARN
  - Apache Mesos
- **Different Modules of Spark**
  - Spark Core
  - Spark SQL
  - Spark Streaming
  - Spark Graph-X
  - Spark MLLib

- **Resilient Distributed Dataset (RDD)**
  - What is RDD and why it is important in Spark
  - RDD Key Features
    - Immutable
    - Lazily Evaluated
    - Partitioned
    - Cacheable
  - How to create a RDD
  - Different types of RDDs.
  - RDD Operations
    - Transformations
    - Actions
  - Different Transformations in RDD
  - Different Actions in RDD
  - Loading Data through RDD
  - Saving Data
  - Key-Value pair RDD
  - Loading and Saving Data – through different File Formats
    - Text,csv,tsv,Object files
    - As a Hadoop file
  - Key-Value Pair RDD operations
  - Spark Storage Persistence Levels
  - Running Spark in a Clustered Mode
  - Deploying Application with spark-submit
  - Cluster Management
  - **Accumulators**
    - Introduction to Accumulators
    - Practical applicability of accumulators
    - Real time examples on Accumulators

# ADE (AZURE DATA ENGINEERING) WITH SPARK (PYSPARK)

**Mr. GOPAL KRISHNA, Sr. Data & Cloud Architect,**
**18+ Years Of Real Time IT Exp, 11+ Years On BIGDATA Implementation Projects Exp**
**CLOUDERA CCA 175 & Microsoft Certified Consultant**

- **Broadcast variables**
  - Introduction to Broadcast variables
  - Practical applicability of Broadcast variables
  - Real time examples on Broadcast variables

## SPARK SQL

- Introduction to Spark SQL
- The SQL Context
- Hive Vs Spark SQL
- Spark SQL support for Text Files, Parquet and JSON files
- Data Frames
- How to process the Data Frame object data
  - SQL Option
  - DSL Commands Option
- Joins in Data Frames
- Handling null values in Spark SQL
- Examples of Spark SQL Real Time
- Different File Formats Supported in Spark SQL
  - Text
  - JSON
  - CSV
  - ORC
  - TSV
  - Parquet
- **Different Integrations with Spark SQL**
  - Spark SQL integration with Hive
  - integration with RDBMS Spark SQL

## SPARK STREAMING

- Introduction to Spark Streaming
- Architecture of Spark Streaming
- RDD vs Discretized Streams(DStreams)
- DStream Operations
- Introduction to SparkStreamingContext(SSC)
- Transformations on DStreams
  - Window Operations
  - Transform Operations
- Spark Streaming Vs Flume
- Introduction to Structure Streaming

# ADE (AZURE DATA ENGINEERING) WITH SPARK (PYSPARK)

**Mr. GOPAL KRISHNA, Sr. Data & Cloud Architect,**
**18+ Years Of Real Time IT Exp, 11+ Years On BIGDATA Implementation Projects Exp**
**CLOUDERA CCA 175 & Microsoft Certified Consultant**

## Azure Data Factory (ADF)

- Introduction to Azure Data Factory
- Components of Azure Data Factory
  - Pipelines
  - Activities
  - Datasets
  - Linked Services
  - Data Flows
  - Integration Runtimes(IR)
  - Triggers
  - Parameters & Variables
  - Monitoring
  - Debugging Tools
  - Management Hub
- What is Integration Runtime(IR)
- What is Azure Integration Runtime
- What is self-hosted Integration Runtime
- What is Azure-SSIS(SQL Server Integration Service) Integration Runtime
- Linked Service Creation for
  - BLOB Storage
  - Azure SQL
  - Azure Data Lake Storage
  - On-Premises Server
- Dataset Creation from
  - CSV File
  - JSON File
  - AVRO File
  - Parquet File
  - Excel File
  - Azure SQL Tables & On-premises SQL DB Tables
- Pipelines
  - Create a new pipeline in ADF
  - How to Publish a pipeline
  - How to Debug a pipeline
- Activities
  - To copy the data
  - To Delete the data
  - Lookup
  - For each

# ADE (AZURE DATA ENGINEERING) WITH SPARK (PYSPARK)

**Mr. GOPAL KRISHNA, Sr. Data & Cloud Architect,**
**18+ Years Of Real Time IT Exp, 11+ Years On BIGDATA Implementation Projects Exp**
**CLOUDERA CCA 175 & Microsoft Certified Consultant**

  - If condition
  - Switch Case
  - Until
  - Wait
  - Fail
  - Databricks Notebook
- Triggers
  - Schedule Trigger
  - Tumbling Window Trigger
  - Storage Events Trigger

## Azure Synapse Analytics

- Introduction to Azure Synapse
- Walk through of Azure Synapse Studio
- How to create a Synapse Account
- Introduction to MPP(Massively Parallel Processing)
- Different Pools in Synapse
  - Serverless SQL Pool
  - Dedicated SQL Pool
  - Apache Spark Pool
  - Azure Data Explorer(ADX) Pool

## Snowflake

- Introduction to Snowflake
- Data Ware House(DWH) Vs Snowflake
- Snowflake Architecture
- Real world examples on Snowflake

## PRE-REQUISITES FOR THE COURSE

  - SQL Commands Basic Knowledge [ reference to Basic SQL Commands will be provided]
  - Python Basics are mandatory [ will be provided as part of course itself ]

# ADE (AZURE DATA ENGINEERING) WITH SPARK (PYSPARK)

**Mr. GOPAL KRISHNA, Sr. Data & Cloud Architect,**
**18+ Years Of Real Time IT Exp, 11+ Years On BIGDATA Implementation Projects Exp**
**CLOUDERA CCA 175 & Microsoft Certified Consultant**

**What we are offering as part of the Course?**
---

- 2 REAL TIME Hadoop Projects End-to-End Explanation with architecture.
- Detailed Assistance in RESUME Preparation on a one-to-one basis with Real Time Projects based on your technical back ground.
- All the Real time interview questions and answers will be provided.
- Discussing the new happenings in Azure Cloud
- Discussing the Interview Questions on a daily basis
- Discussing Certification (Azure Data Engineering Certification – Azure DP-203) Related topics on a daily basis.
- Academic Projects will be provided for pursuing students