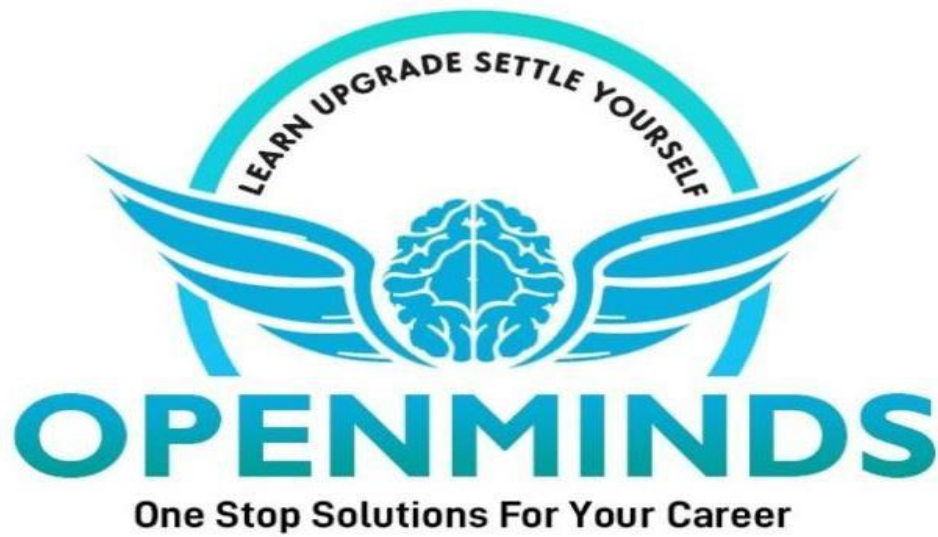
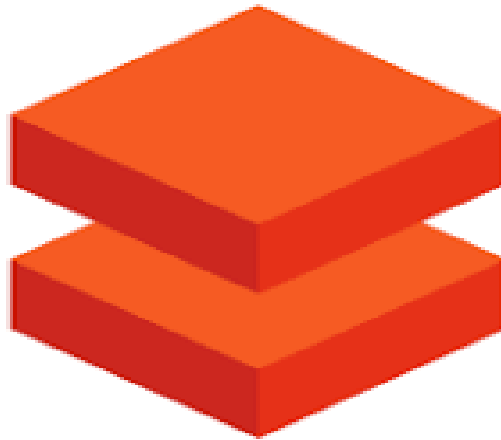




**OpenMinds Technologies**  
**Plot No:604/A, Aditya Enclave, Nilgiri Block , 6<sup>th</sup> Floor,Ameerpet,**  
**Besides Metro Station, Contact Mobile Number:+91-9542173327 /8522013676**  
**Website: [www.openmindstechnologies.com](http://www.openmindstechnologies.com)**



**Azure Databricks with Pyspark and Spark SQL**  
**Curriculum Learning Plan**

## 1 Introduction to Databricks

- What is the use of Databricks?
- Spark Architecture
- Workspace
- Types of Clusters and Runtimes
- Notebooks
- Jobs

## 2 DBFS(Databricks File System)

- Upload files into DBFS
- `dbutils.fs`
- `dbutils.data`
- `dbutils.data.summarize`
- `dbutils.fs.cp`
- `dbutils.fs.head`
- `dbutils.fs.mkdirs`

## 3 Notebooks utilities in Databricks

- `dbutils.notebook.run()`
- `dbutils.notebook.exit()`
- `dbutils.data`
- `dbutils.data.summarize`
- `dbutils.fs.cp`
- `dbutils.fs.head`
- `dbutils.fs.mkdirs`

## 4 Widget utilities in Databricks

- `Dbutils.widgets.combobox`
- `Dbutils.dropdown.dropdown`
- `dbutils.dropdown.multiselect`

## Widget utilities in Databricks

- `dbutils.dropdown.text`
- `dbutils.dropdown.get`
- `dbutils.dropdown.getArgument`
- `dbutils.widgets.remove`

## 5. Pass values to notebook parameters from another notebook using `run()` command

## 6. Create mount point in Azure Databricks using `dbutils.fs.mount()`

- Create mount point using Account Key
- Create mount point using SAS Token
- Connect ADLS Gen2 to 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

## 7. Secrets Management in Databricks

- Azure Key vault backed scopes
- Databricks-backed scopes

## 8. Read/Write Data from SQL Database using JDBC Connector

## 9. Access Notebook from ADF Pipeline

- Pass Parameter to Notebook from ADF Pipeline
- Send Exception or Error message from Pipeline to Notebook
- Send parameters from Notebook to ADF Pipelines

## 10 Delta Lake

- Time Travel/Versioning in Delta Lake
- RESTORE
- Get Delta Lake History
- Vacuum Command in Delta Lake
- Merge Command in Delta Lake
- Schema evaluation in Delta Lake
- Change Data feed in Delta Lake
- Z-Order index
- Table constraints
- Runbooks
- Run SQL Commands from Runbook
- Run Pipeline from Runbook

## 11 Optimization and Performance

- Optimize performance with caching on Azure Databricks
- Dynamic file pruning
- Low shuffle merge
- Adaptive query execution
- predictive IO
- Cost-based optimizer
- Auto optimize
- Higher-order functions
- Isolation levels
- column mapping

## 12 RDD

- Parallelize
- Collect()
- Repartition vs coalesce
- Broadcast variables
- Accumulator
- RDD Transformations
- Transformations Actions

## 13 Pyspark Dataframe

- Create an Empty DataFrame
- Create Empty DataFrame with Schema
- Convert Empty RDD to DataFrame

## 14 Pyspark Transformations

- show()/Display()
- StructType() and StructField()
- Column class
- Select()
- WithColumn()
- WithColumnRenamed()
- Where() and Filter()
- Drop() and DropDuplicates()
- orderBy() and sort()
- groupBy()
- Join()
- union() and unionAll()
- unionByName()
- map()
- flatmap()
- fillna() and fill()
- pivot()
- partitionBy()
- mapType()
- foreach()
- User Defined Functions

## 15 Pyspark SQL Functions

- Aggregate functions
- Window functions
- Date and Timestamp functions
- JSON functions

## 16 Pyspark DataSources

- Read and write CSV File
- Read and Write Parquet File
- Read and Write JSON File
- Read Hive Table
- Save to Hive Table
- Read JDBC in parallel
- Query Database Table
- Read and Write SQL Server
- Read JDBC Table

## 17 Pyspark Built-in functions

- when()
- expr()
- lit()
- split()
- concat\_ws()
- substring()
- translate()
- regexp\_replace()
- to\_timestamp()
- to\_date()
- date\_format()

## 19 Extra Features

- Autoloader → Ingest data efficiently into Delta tables.
- Unity Catalog → Govern and secure those Delta tables across workspaces/clouds.
- Mounting → Provide simple, persistent access to external storage locations.

## 20 Latest Announced Databricks CDC'S

- CDC (Change Data Capture) in Databricks
- CDC Approaches in Databricks
- Implementing CDC in Databricks with Delta Lake
- Medallion Architecture in Databricks
- What is it?
- Layers in Medallion Architecture
- Key Benefits

## 18 Pyspark Built-in functions

- struct()
- countDistinct()
- sum(), avg()
- row\_number()
- rank()
- dense\_rank()
- from\_json()
- to\_json()
- json\_tuple()
- get\_json\_object()
- schema\_of\_json()
- array()
- collect\_list()
- collect\_set()
- create\_map()
- map\_keys()
- map\_values()
- months\_between()
- explode()
- array\_constraints()
- datediff()