* **We are going to write pyspark code in notebook which will be written in python, But why we want to create a cluster databricks run time version which include scala and spark**

Basically, Apache Spark originally written in scala. On top of that, APIs like java , R python was written so we are selecting ADB run time version in scala only.

* **Use of SparkContext vs SparkSession?**

When we are coding in notebook, we need to say this is a spark program and we need to inhale all features which is available in spark , for that we need to create a object , that is nothing but spark context or spark session. Spark Context and Spark Session is an entry point to spark application. SparkSession introduced in spark 2.0 and above versions. Prior to that we have only spark context which we need to create before every program.

Using Spark Context , we can only be create RDD. If we need to use additional features of spark like data frames or streaming or hive related things , we need to create separate context like HiveContext, SQL Context , Streaming Context.

But in SparkSession, Hive Context, SQLContext, StreamingContext , SparkContext everything is encapsulated with in the Spark Session itself. Also, via spark session, major issue of multiple users accessing the same notebook environment also been addressed using the spark session. If we are using command line console, then spark object will be created by default.

**from pyspark.sql import SparkSession**

**def getsparkSession():**

**spark = SparkSession.builder.master(local[1])\**

**.appName(‘abc’)\**

**.getOrCreate()**

**return spark**

**Code Explanation :** In order to run this application, we are specifying master . master is a resource manager . so how much resource we are going to allocate to run this spark application. So, if this is stand-alone machine, let’s say if this is our personal laptop, we need to assign master as local and no of cores we are going to assign as local[1]. If we are using our real time production environment then master will be YARN. YARN is a resource manager which will allocate the resource to this spark application.

**appName usage in code:** In real time spark environment, there will be lot of spark applications will be running. We should know which is my application. If we want to know my application, we should give one appName as xxxxxx .

**getOrCreate() usage in code:** If spark session is already available, it will get it. If spark session is not exist then it will create it and we are returning spark object in the form of spark as shown in above code.

By using spark object, we can call a method called spark.read. etc

CSV files from source via ADF will be stored in ADLS Gen2 . This will be stored in ADLS Gen2 in zip format/zip folders . First, we have to extract those files and then we have to load it.

API will be hosted on cloud. So, we will fetch those records from the API.

Data from publish layer i.e; final layer data will be stored in 2 different places, ADLS Gen2 and Azure SQL/Azure DWH i.e; Azure Synapse. Because data lake is something not accessible by end users because they do not know how to write that query which means how to import the data. Azure SQL is something like everyone knows about it like writing simple statement.