**Sushmita Arun Chaudhari**

**ID= INT-04**

**Subject: Salary Data Processing & Analysis Statement of Work (SoW)**

**## Documentation**

# **# OBJECTIVE**

# Here is the main objective was ,to gain valuable insights from given data of employee, department, and salary.

So as we start project handling the main task is review of data. So salary data was large enough containing 200000 rows. As number of tables ,rows are less in employee and department table as compare to salary, it was easy to review Data.

**# TOOLS:**

The required tools were, pyspark, power BI, SQL.

**Working:**

* After that according to given activity Medallion Architecture has to be followed. So task has begin with making containers blob storage for input and data lakegen storage for output with bronze, silver, gold layer
* As datasets were clean enough since duplicates were not also present, it took just checking Null values ,drop null values and then join the tables with their common tables.
* The main task was how to connect Databricks pyspqrk to azure SQL ,as silver layer and gold layer needed to be in SQL Server. But the JDBC code make it easy.
* Then the auto connection held between pyspark and sql server,all the tables from silver layer get connected to gold layer.
* And then according to desired ouput gold layer with salary\_gold has been formed for further transformation
* Aggregation was needed according to department wise ,salary wise
* So final output is a PowerBI dashboard

**## Final Project**

**Challenges:**

Firstly, looking at the data set with so much records, it was looking horrible to work on project, but as number of columns were les it went smoothly.

As well as the problem of dataset is missing value, but there were just few records with missing value, so the get easily.

**Gaining:**

* It was a great experience to make the different connections as per requirement.
* Power BI help to gain distribution of employee salary ,department wise maximum salary with the help bar chart ,pie chart,line chart.
* As I found there was a lack of some insightful columns in department, employee so not much more gainging possibled as it was needed.
* This projects defined ,it may possible to work on such big data like with having such 3 datasets . We can manage product wise cost, can manage region wise production but some more columns will be needed.

And here is final presentation of dashboard:

