



**04\_Employees Earning More Than Their Managers**

Write a Pyspark program to find Employees Earning More Than Their Managers

**Difficult Level :** EASY

**DataFrame:**

# Define the schema for the "employees"

employees\_schema **=** StructType**([**

StructField**(**"id"**,** IntegerType**(), True),**

StructField**(**"name"**,** StringType**(), True),**

StructField**(**"salary"**,** IntegerType**(), True),**

StructField**(**"managerId"**,** IntegerType**(), True)**

**])**

# Define data for the "employees"

employees\_data **= [**

**(1,** 'Joe'**, 70000, 3),**

**(2,** 'Henry'**, 80000, 4),**

**(3,** 'Sam'**, 60000, None),**

**(4,** 'Max'**, 90000, None)**

**]**



**INPUT**

| **INPUT** | | | |
| --- | --- | --- | --- |
| **ID** | **NAME** | **SALARY** | **MANAGERID** |
| **1** | **Joe** | **70,000** | **3** |
| **2** | **Henry** | **80,000** | **4** |
| **3** | **Sam** | **60,000** |  |
| **4** | **Max** | **90,000** |  |

**OUTPUT**

| **OUTPUT** |
| --- |
| **NAME** |
| **Joe** |



**# Creating Spark Session**

**from pyspark.sql import SparkSession**

**from pyspark.sql.types import StructType,StructField,IntegerType,StringType**

**from pyspark.sql.functions import when**

**from pyspark.sql import functions as F**

**from pyspark.sql.window import Window**

**#creating spark session**

**spark = SparkSession. \**

**builder. \**

**config('spark.shuffle.useOldFetchProtocol', 'true'). \**

**config('spark.ui.port','0'). \**

**config("spark.sql.warehouse.dir", "/user/itv008042/warehouse"). \**

**enableHiveSupport(). \**

**master('yarn'). \**

**getOrCreate()**

**# Define the schema for the "employees"**

**employees\_schema = StructType([**

**StructField("id", IntegerType(), True),**

**StructField("name", StringType(), True),**

**StructField("salary", IntegerType(), True),**

**StructField("managerId", IntegerType(), True)**

**])**

**# Define data for the "employees"**

**employees\_data = [**

**(1, 'Joe', 70000, 3),**

**(2, 'Henry', 80000, 4),**

**(3, 'Sam', 60000, None),**

**(4, 'Max', 90000, None)**

**]**

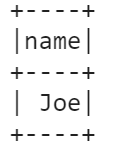
**# Create a PySpark DataFrame**

**emp\_df=spark.createDataFrame(employees\_data,employees\_schema)**

**emp\_df.show()**

| **emp\_df1 = emp\_df.alias("e1")**  **emp\_df2 = emp\_df.alias("e2")**    **self\_joined\_df = emp\_df1.join(emp\_df2, col("e1.id") == col("e2.managerId"), "inner") .select(col("e2.name"),col("e2.salary"),col("e1.salary").alias("msal"))** |
| --- |

**self\_joined\_df.filter(self\_joined\_df.salary>self\_joined\_df.msal).select("name").show()**

****

