RDD and DataFrame Operations

**Dataframe Operations:**

**1. Creating New column, Renaming New column, Changing DataType of column, selecting subset of column**

**1.1) Creating New column**

emp.withColumn("Ecopy", $"EMPNO").show()

**1.2) Renaming new column**

emp.withColumnRenamed("EMPNO", "Ecopy2").show()

**1.3) Changing DataType of the column**

emp.withColumn("MGR", $"MGR".cast("double")).printSchema()

**Functional way:**

import org.apache.spark.sql.Column

import org.apache.spark.sql.functions.udf

def toDouble(df: Column): Column = {

df.cast("double")

}

emp.withColumn("dd", toDouble(emp("MGR"))).show()

**1.4)** **selecting subset of column**

**Use of when:**

val emp1 = emp.withColumn("COMM", when($"COMM" === "NULL", 0).otherwise($"COMM"))

emp1.groupBy("DEPTNO").agg(sum(when($"COMM".isNull, $"SAL").otherwise($"SAL"+$"COMM"))).show()

**3. Counting null values, handling and filling null values**

**3.1 Counting null values:**

**pushing null values where value is NULL:**

val emp2 = emp.withColumn("COMM", when($"COMM" === "NULL", null).otherwise($"COMM"))

emp2.filter("COMM is null").show()

**3.2 Filling null values:**

emp2.na.fill("0", Seq("COMM")).show()

**3.3 Droping null values:**