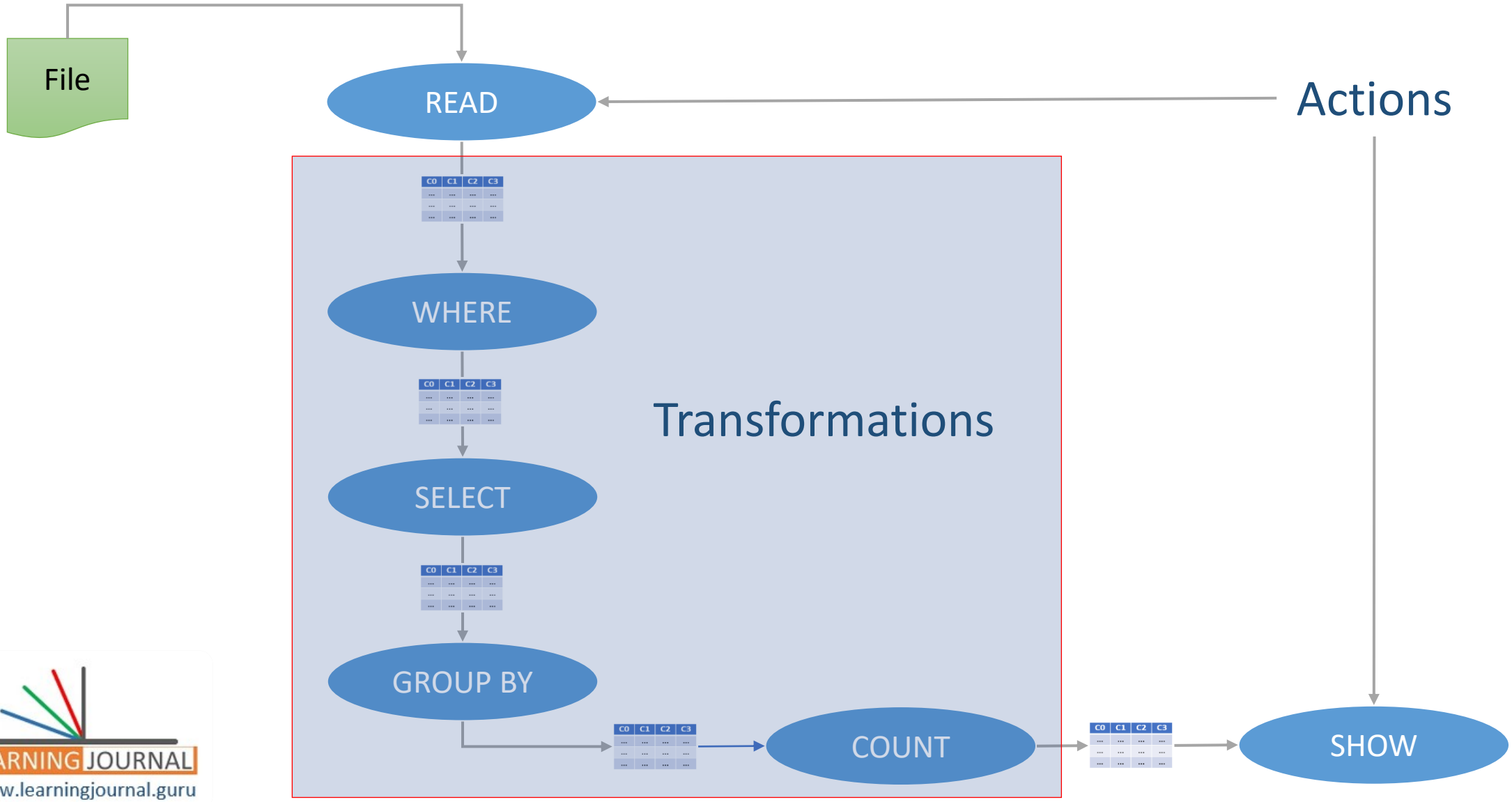
Apache Spark

Spark Jobs, Stages, and Tasks

Spark Operations



Spark Execution Plan

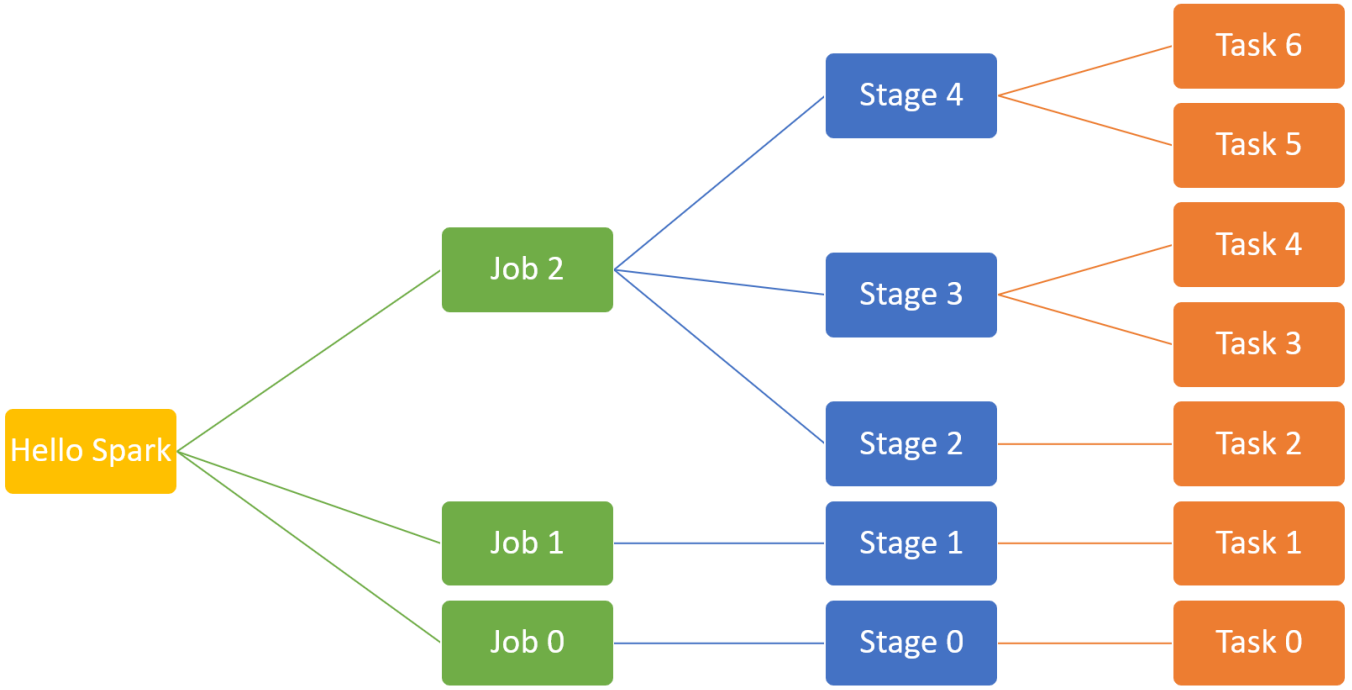
```scala
def main(args: Array[String]): Unit = {
  if (args.length == 0) {
    logger.error("Usage: HelloSpark filename")
    System.exit( status = 1)
  }
  logger.info("Starting Hello Spark")
  val spark = SparkSession.builder()
    .config(getSparkAppConf)
    .getOrCreate()
  //logger.info("spark.conf=" + spark.conf.getAll.toString())

  val surveyRawDF = loadSurveyDF(spark, args(0))
  val partitionedSurveyDF = surveyRawDF.repartition( numPartitions = 2)
  val countDF = countByCountry(partitionedSurveyDF)

  logger.info(countDF.collect().mkString("->"))

  logger.info("Finished Hello Spark")
  scala.io.StdIn.readLine()
  spark.stop()
}

def countByCountry(surveyDF: DataFrame): DataFrame = {
  surveyDF.where( conditionExpr = "Age < 40")                    : Dataset[Row]
    .select( col = "Age",  cols = "Gender", "Country", "state")  : DataFrame
    .groupBy( col1 = "Country")                                  : RelationalGroupedDataset
    .count()                                                     : DataFrame
}
```




LEARNING JOURNAL
www.learningjournal.guru