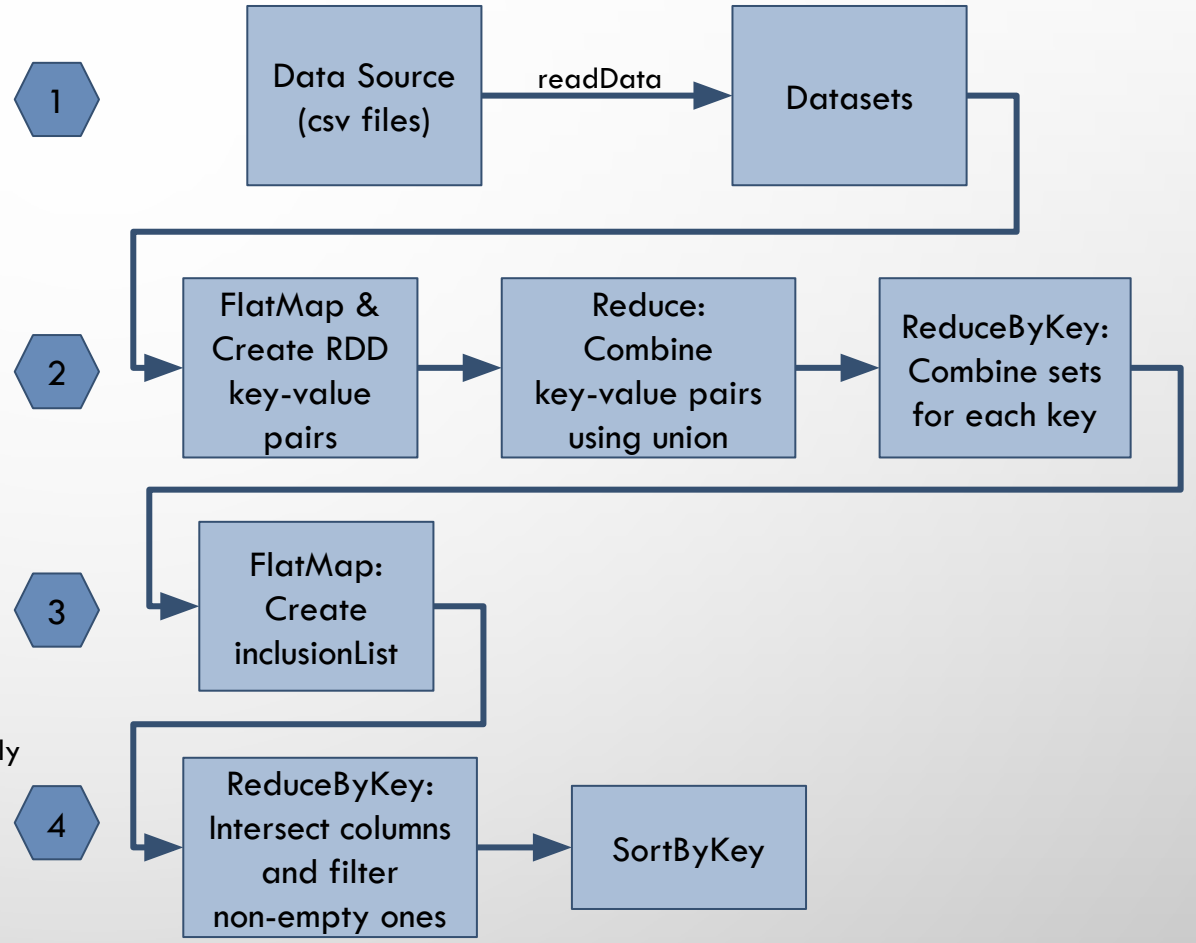
BIG DATA SYSTEM

SPARK PROJECT

CODE

```
def discoverINDs(inputs: List[String], spark: SparkSession): Unit = {
  //Use readData to read input files as datasets
1 val tables = inputs.map(input => readData(input, spark))

  //Retrieve columns values & names
  //-> Map them to (value, [ColumnName])
  //-> Group by value and add possible columns for value to
[columnName] set
  //-> Remove duplicate values
2 val valueSets = tables.flatMap(df => df.columns.map(col =>
df.select(col).rdd.map(r => (r(0).toString, Set[String](col)))))
    .reduce(_ union _)
    .reduceByKey(_ ++ _).values.distinct()

  //Create inclusionList (column, set/column)
3 val includeList = valueSets.flatMap(set => set.map(column =>
(column, set - column)))

  //Intersect includeList to find out common value, then filter to keep only
non-empty ones
4 val intersectSet = includeList.reduceByKey(_.intersect(_))
    .filter(_._2.nonEmpty).sortByKey(numPartitions = 1)

  // Print result
  intersectSet.foreach(columnName =>
    println(columnName._1 + " < " + columnName._2.mkString(", "))
  )
}
```

PIPELINE



<WL SQUAD>

NGOC THAO LY NGUYEN, NGUYENN6@STUDENTS.UNI-MARBURG.DE
CHIA-WEI CHOU, CHOUC@STUDENTS.UNI-MARBURG.DE