

The background of the slide is a light gray gradient. It is decorated with numerous realistic water droplets of various sizes. Some droplets are large and prominent, while others are small and subtle. They are scattered across the slide, with a higher concentration in the top-left and bottom-right corners. Each droplet has a highlight and a shadow, giving it a three-dimensional appearance.

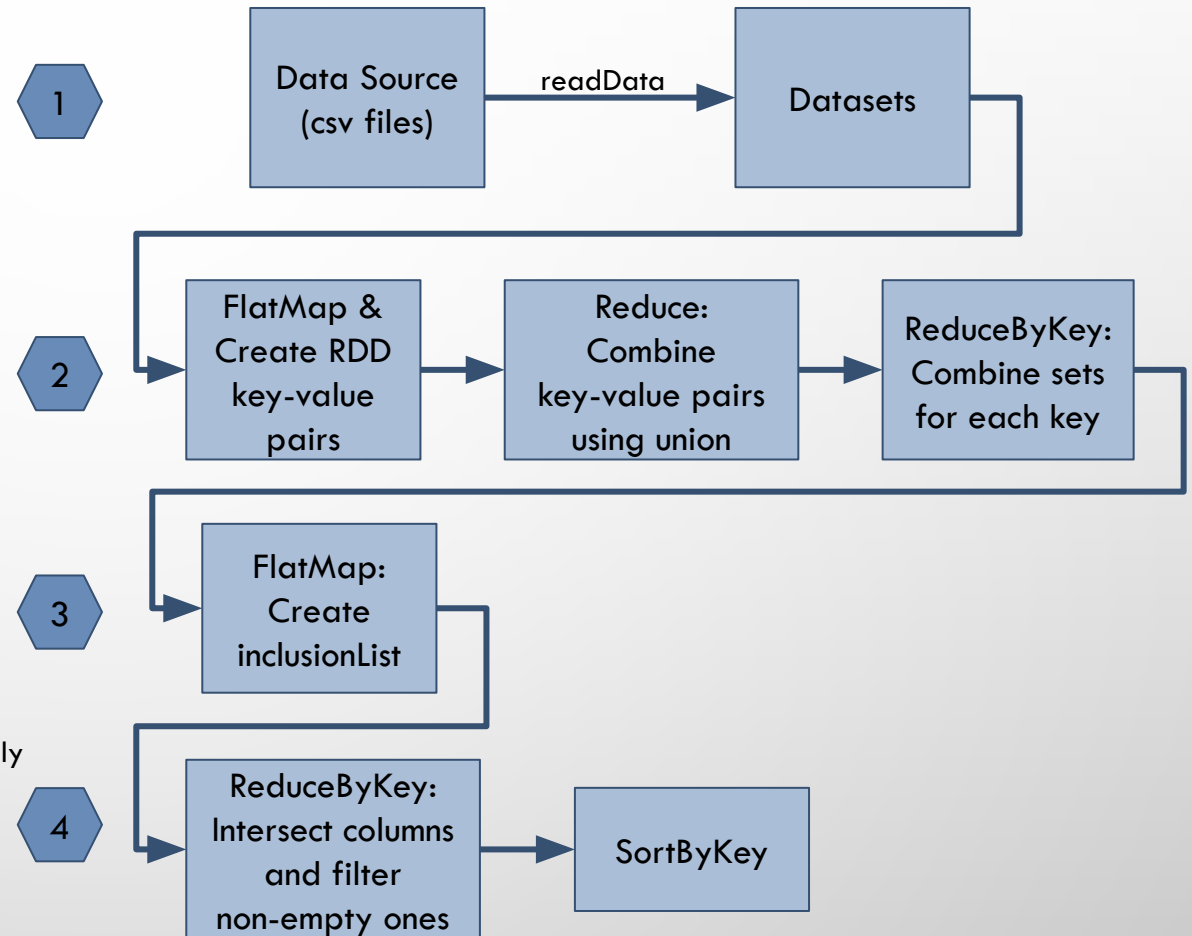
# BIG DATA SYSTEM

SPARK PROJECT

# CODE

```
def discoverINDs(inputs: List[String], spark: SparkSession): Unit = {  
  1 //Use readData to read input files as datasets  
  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  
  2 //-> Remove duplicate values  
  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

