**Design Documentation**

**Query 1 :**

* For the join query two mappers and one reducer are used. Each mapper works with one input file (one table).
* The input format of the mappers is TextInputFormat, the key is the line number and the value is a single record.

Eg. (key : 10, value:”780,49,M,programmer,94560”)

* Mapper output : Key is the column value over which join takes place and the value is rest of the fields separated by a tab. Mapper also appends a string(“a.” or “b.”) in the beginning of value so that the reducer can identify whether a key value pair come from first input file or the second.

Eg. (key:94560, value:”a.780 49 M programmer”)

* The reducer iterates over all pairs of values for a particular key and if the two key value pairs belong to different table (have different appended string) it appends the two values.
* Lastly before giving writing to the output collector. The reducer checks for the where condition and writes only if the condition is satisfied.

**Query 2:**

* For the group query only one mapper and one reducer are used.
* The mappers input format is KeyValueTextInputFormat where in csv the first element is taken as key and all the rest are taken as value.

Eg. (key: 780, value: 49,M,programmer,94560)

* The mapper then concatenate all the field values over which group-by is taking place as key and value is the column’s value over which aggregation is taking place.

Eg ( key: “49 94560”, value:”49”) if grouping is over age and zip code and aggregation over age.

* The reducer just iterates over all values and takes the aggregation function according over the value and writes the expected output.