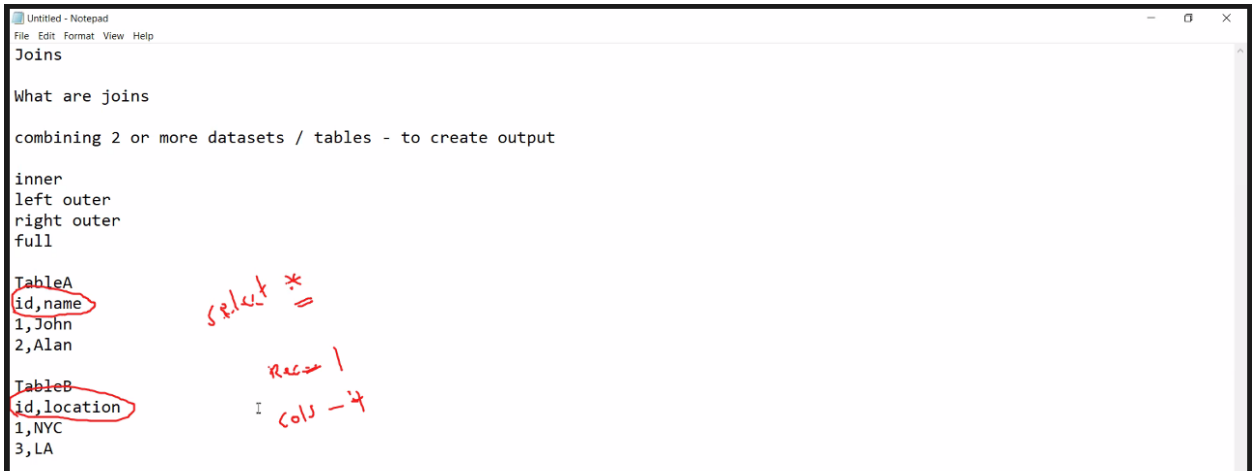
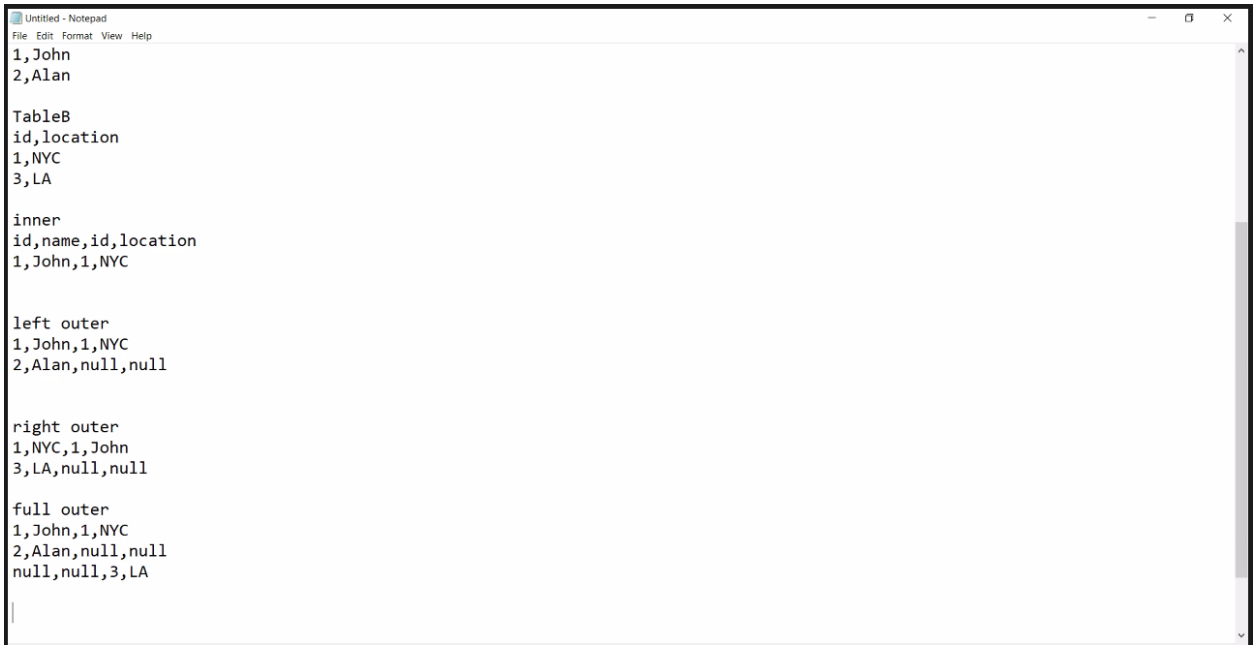
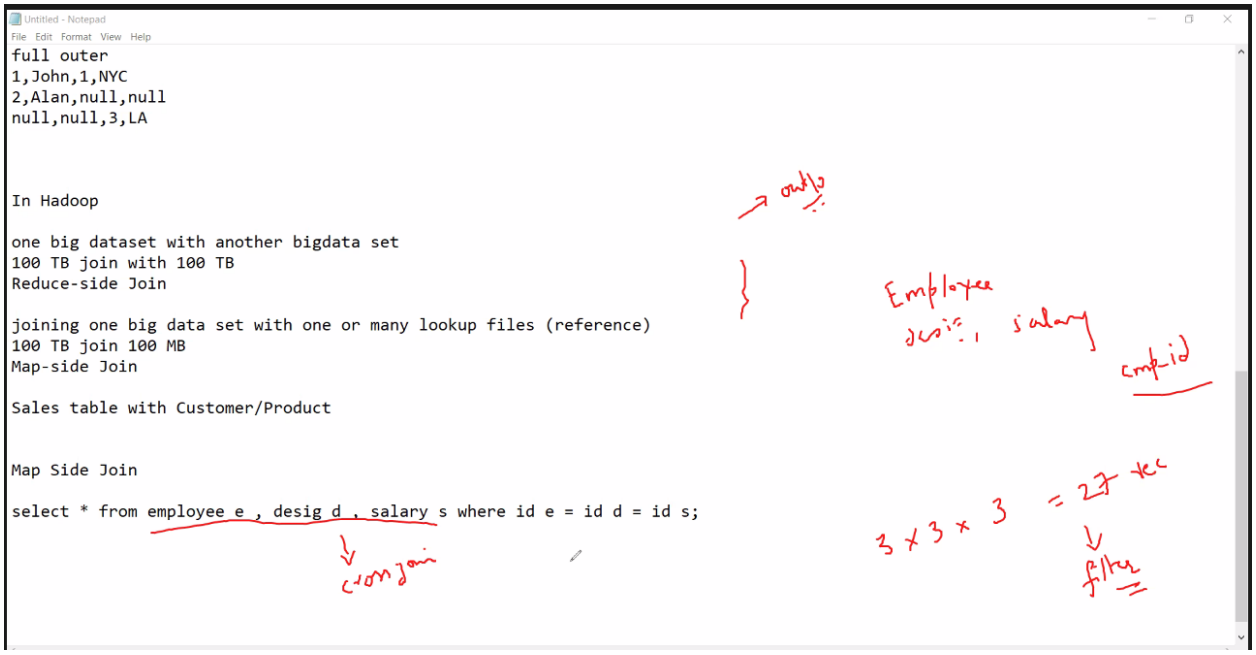
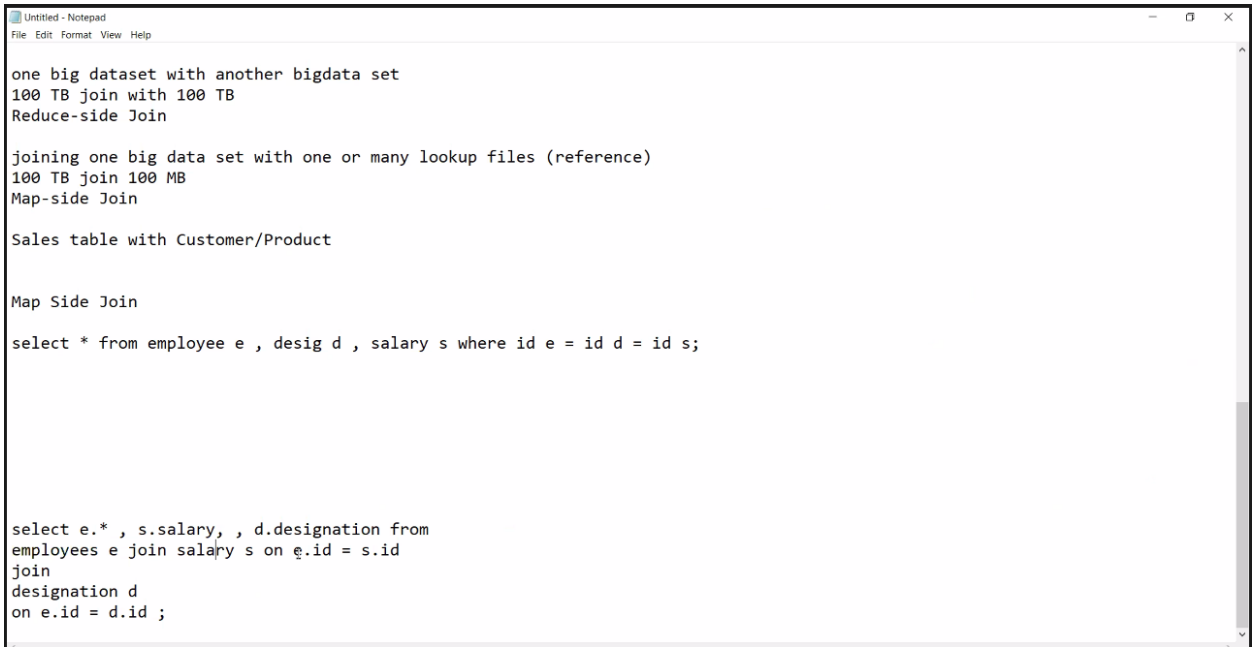
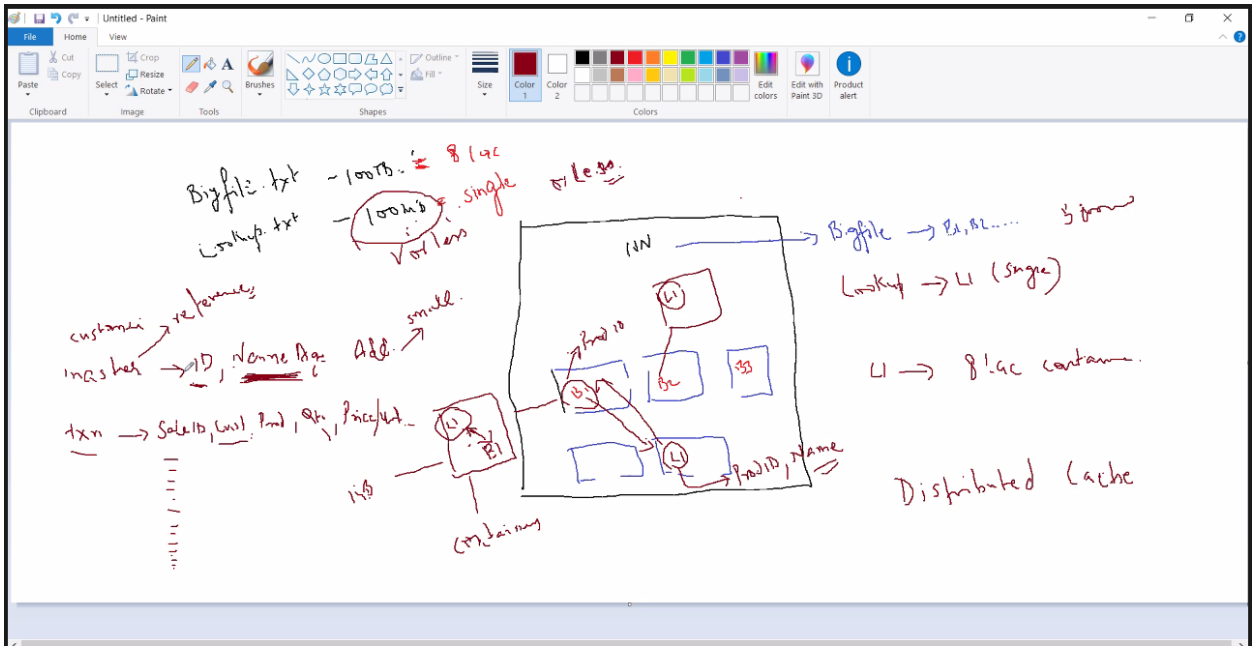
→ Joins

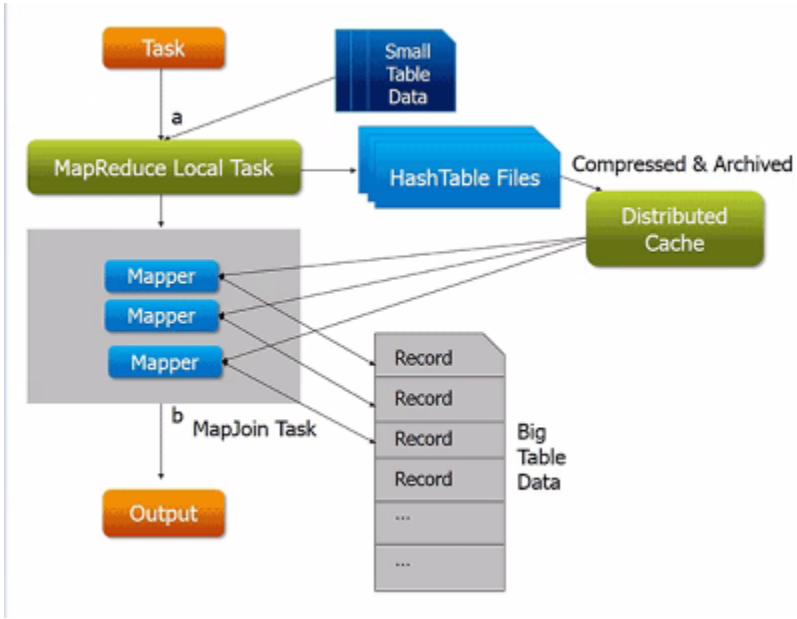

Untitled - Notepad
File Edit Format View Help
Joins
What are joins
combining 2 or more datasets / tables - to create output
inner
left outer
right outer
full
TableA
id,name
1,John
2,Alan
TableB
id,location
1,NYC
3,LA
select *
Rows 1
cols - 4

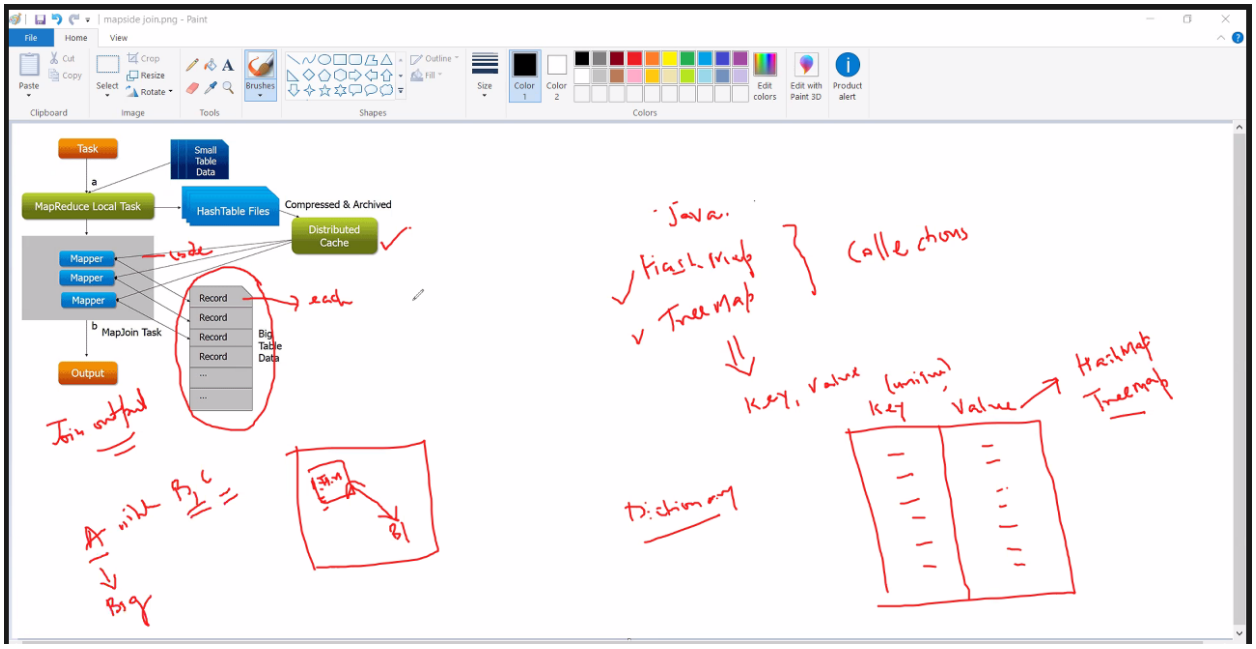

Untitled - Notepad
File Edit Format View Help
1,John
2,Alan
TableB
id,location
1,NYC
3,LA
inner
id,name,id,location
1,John,1,NYC
left outer
1,John,1,NYC
2,Alan,null,null
right outer
1,NYC,1,John
3,LA,null,null
full outer
1,John,1,NYC
2,Alan,null,null
null,null,3,LA

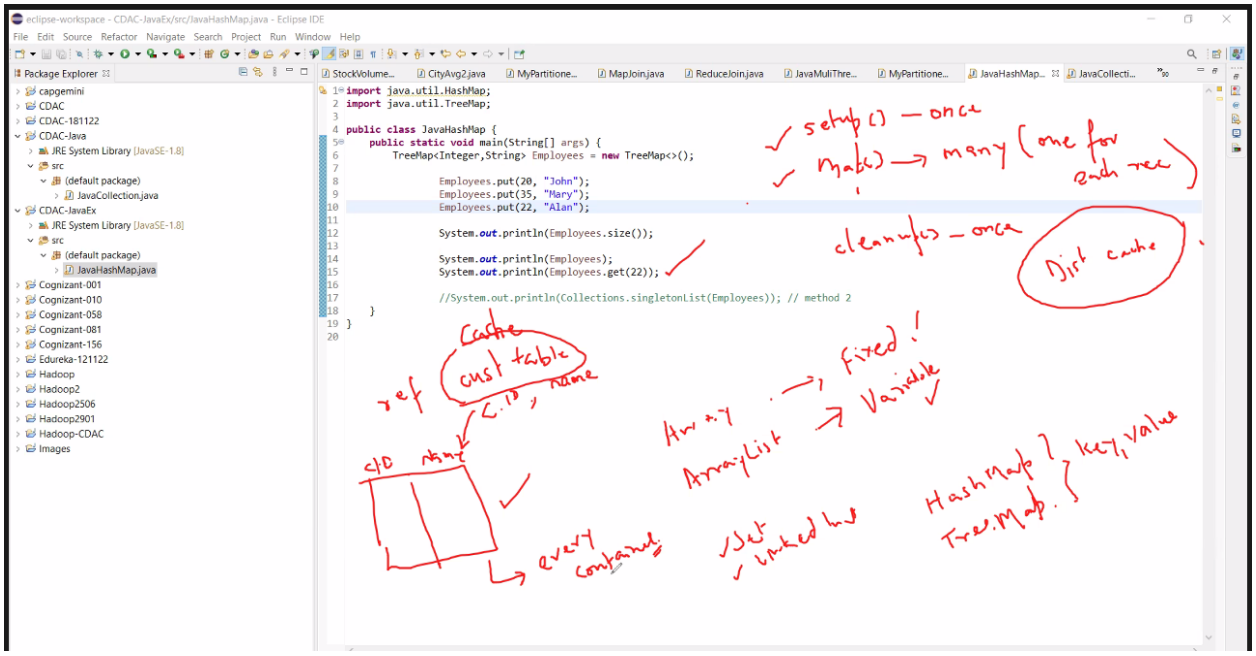

Untitled - Notepad
File Edit Format View Help
full outer
1,John,1,NYC
2,Alan,null,null
null,null,3,LA
In Hadoop
one big dataset with another bigdata set
100 TB join with 100 TB
Reduce-side Join
joining one big data set with one or many lookup files (reference)
100 TB join 100 MB
Map-side Join
Sales table with Customer/Product
Map Side Join
select * from employee e , desig d , salary s where id e = id d = id s;
outputs
Employee
desig
salary
emp_id
3 x 3 x 3 = 27 rec
filter
cross join

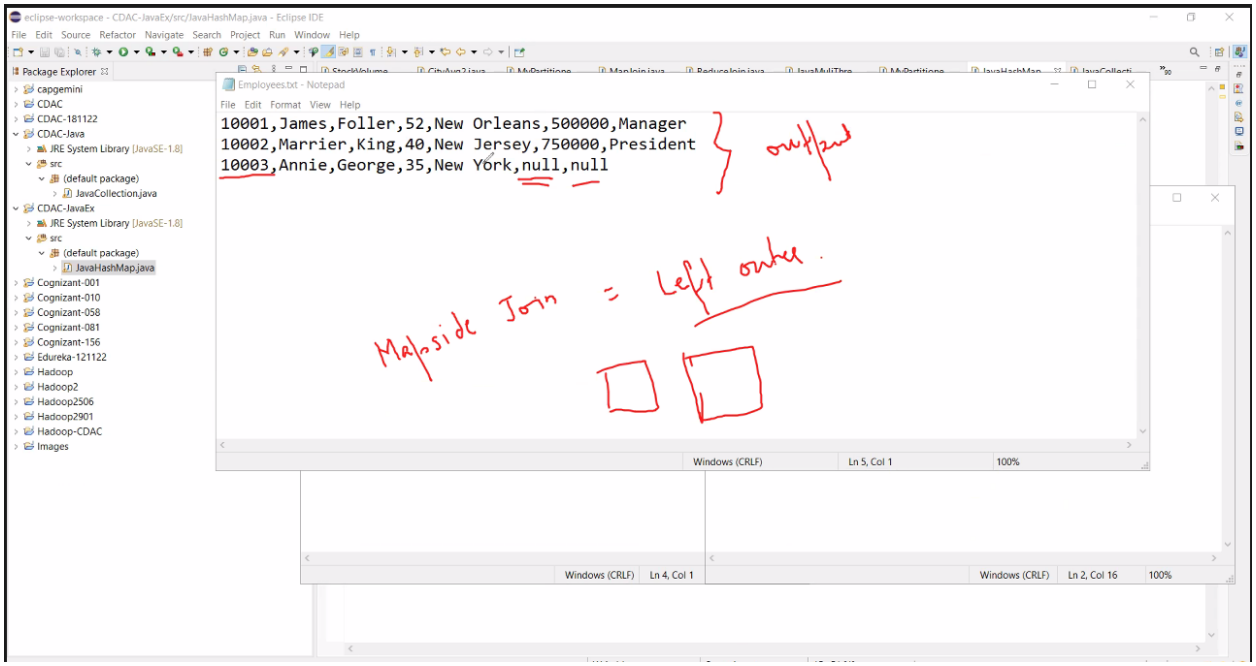

Untitled - Notepad
File Edit Format View Help
one big dataset with another bigdata set
100 TB join with 100 TB
Reduce-side Join
joining one big data set with one or many lookup files (reference)
100 TB join 100 MB
Map-side Join
Sales table with Customer/Product
Map Side Join
select * from employee e , desig d , salary s where id e = id d = id s;
select e.* , s.salary, , d.designation from
employees e join salary s on e.id = s.id
join
designation d
on e.id = d.id ;


Untitled - Paint
Distributed Cache


Task
Small Table Data
a
MapReduce Local Task
HashTable Files
Compressed & Archived
Distributed Cache
Mapper
Mapper
Mapper
Record
Record
Record
Record
...
...
Big Table Data
b
MapJoin Task
Output


mapside join.png - Paint
Task
Small Table Data
a
MapReduce Local Task
HashTable Files
Compressed & Archived
Distributed Cache
Mapper
Mapper
Mapper
Record
Record
Record
Record
Big Table Data
b
MapJoin Task
Output
Java.
Collections
HashMap
TreeMap
Key, value
Key
Value
Dictionary


eclipse-workspace - CDAC-JavaEx/src/JavaHashMap.java - Eclipse IDE
Package Explorer
capgemini
CDAC
CDAC-181122
CDAC-Java
JRE System Library [JavaSE-1.8]
src
(default package)
JavaCollection.java
CDAC-JavaEx
JRE System Library [JavaSE-1.8]
src
(default package)
JavaHashMap.java
Cognizant-001
Cognizant-010
Cognizant-058
Cognizant-081
Cognizant-156
Edureka-121122
Hadoop
Hadoop2
Hadoop2506
Hadoop2901
Hadoop-CDAC
Images
import java.util.HashMap;
import java.util.TreeMap;
public class JavaHashMap {
public static void main(String[] args) {
TreeMap<Integer,String> Employees = new TreeMap<>();
Employees.put(20, "John");
Employees.put(35, "Mary");
Employees.put(22, "Alan");
System.out.println(Employees.size());
System.out.println(Employees);
System.out.println(Employees.get(22));
//System.out.println(Collections.singletonList(Employees)); // method 2
}
}
setup() — once
make() → many (one for each rec
cleanup() — once
Dist cache
Cache
cust table
C.ID, Name
ref
C.ID
Name
every
Array → fixed!
ArrayList → Variable
Set
LinkedList
HashMap
TreeMap
key, value



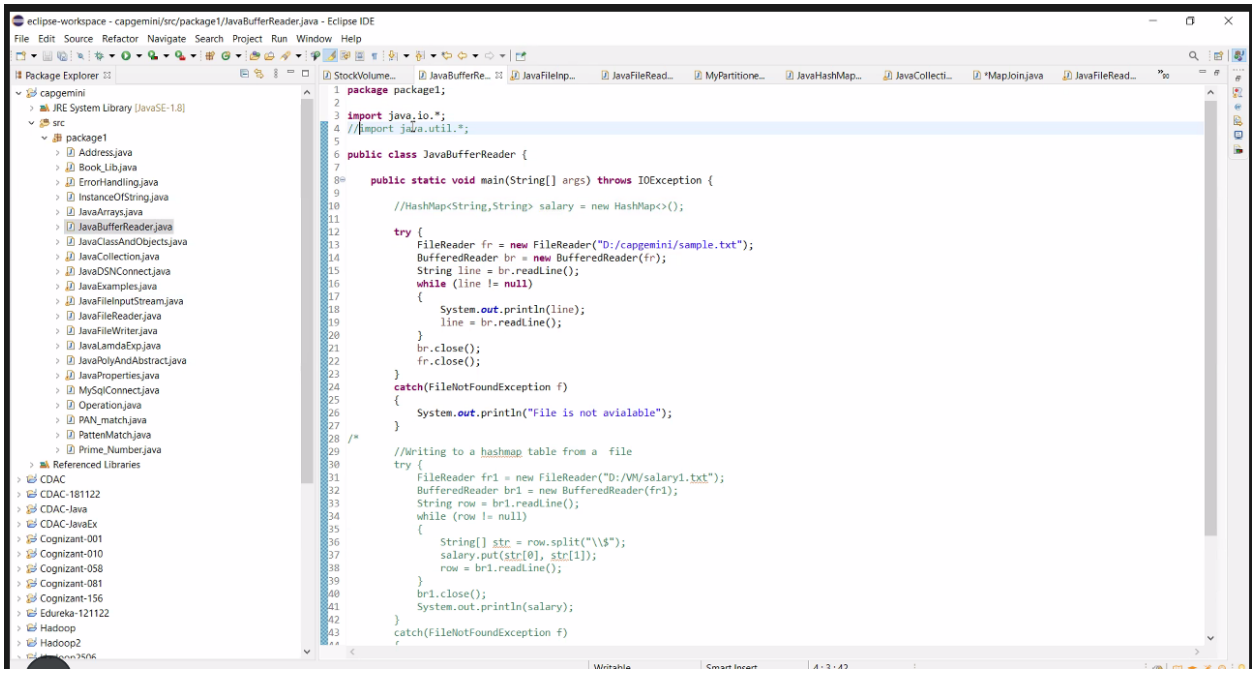

code:

```
try {
                FileReader fr = new
FileReader("D:/capgemini/sample.txt");
                BufferedReader br = new BufferedReader(fr);
                String line = br.readLine();
                while (line != null)
                {
                      System.out.println(line);
                      line = br.readLine();
                }
                br.close();
                fr.close();
          }
          catch(FileNotFoundException f)
          {
                System.out.println("File is not available");
          }
```

sample.txt :

```
Java is used as a native lang
Second line
Sixth line
Java is used as a native lang
Second line
Sixth line
Java is used as a native lang
Second line
```

Sixth line




## → For Map-side Join

a. produces Left Join
b. One big data set with one or many lookup / reference tables
c. 100TB with 100MB
d. Only mapper is needed in map-side join
e. One lookup-file data is in memory in map-side join

#

```
[bigdatalab456422@ip-10-1-1-204 ~]$ jar tvf  myjar.jar
    25 Mon May 22 12:17:26 UTC 2023 META-INF/MANIFEST.MF
   387 Thu May 18 15:53:20 UTC 2023 .project
  2459 Fri May 19 16:03:30 UTC 2023 AllTimeHigh$MapClass.class
  2392 Fri May 19 16:03:30 UTC 2023 AllTimeHigh$ReduceClass.class
  1722 Fri May 19 16:03:30 UTC 2023 AllTimeHigh.class
  2475 Fri May 19 16:53:46 UTC 2023 AvgClosingPrice$MapClass.class
  2454 Fri May 19 16:53:46 UTC 2023 AvgClosingPrice$ReduceClass.class
  1732 Fri May 19 16:53:46 UTC 2023 AvgClosingPrice.class
  2337 Fri May 19 17:41:44 UTC 2023 WordCount$IntSumReducer.class
  2461 Fri May 19 17:41:44 UTC 2023 WordCount$TokenizerMapper.class
  1790 Fri May 19 17:41:44 UTC 2023 WordCount.class
  2454 Fri May 19 15:53:50 UTC 2023 AllTimeLow$MapClass.class
  2388 Fri May 19 15:53:50 UTC 2023 AllTimeLow$ReduceClass.class
  1734 Fri May 19 15:53:50 UTC 2023 AllTimeLow.class
  1242 Sat May 20 17:37:44 UTC 2023
MyPartitioner$CaderPartitioner.class
  2365 Sat May 20 17:37:44 UTC 2023 MyPartitioner$MapClass.class
```

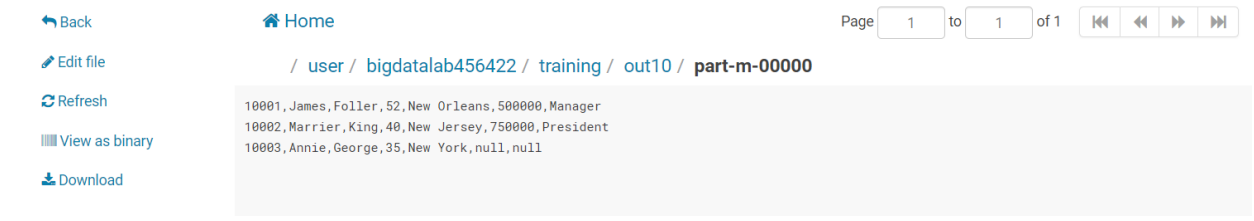
```
 2905 Sat May 20 17:37:44 UTC 2023 MyPartitioner$ReduceClass.class
 2632 Sat May 20 17:37:44 UTC 2023 MyPartitioner.class
 2408 Thu May 18 17:48:56 UTC 2023 StockVolume$MapClass.class
 2349 Thu May 18 17:48:56 UTC 2023 StockVolume$ReduceClass.class
 1697 Thu May 18 17:48:56 UTC 2023 StockVolume.class
 2648 Sat May 20 15:42:48 UTC 2023 CityAvg2$CityCombineClass.class
 2269 Sat May 20 15:42:48 UTC 2023 CityAvg2$CityMapClass.class
 2639 Sat May 20 15:42:48 UTC 2023 CityAvg2$CityReduceClass.class
 2034 Sat May 20 15:42:48 UTC 2023 CityAvg2.class
 4760 Mon May 22 12:05:50 UTC 2023 MapJoin$MyMapper.class
 1817 Mon May 22 12:05:50 UTC 2023 MapJoin.class
 2456 Sat May 20 15:04:24 UTC 2023
StockVolumeWithCombiner$MapClass.class
 2397 Sat May 20 15:04:24 UTC 2023
StockVolumeWithCombiner$ReduceClass.class
 1813 Sat May 20 15:04:24 UTC 2023 StockVolumeWithCombiner.class
  640 Thu May 18 17:00:00 UTC 2023 .classpath
```

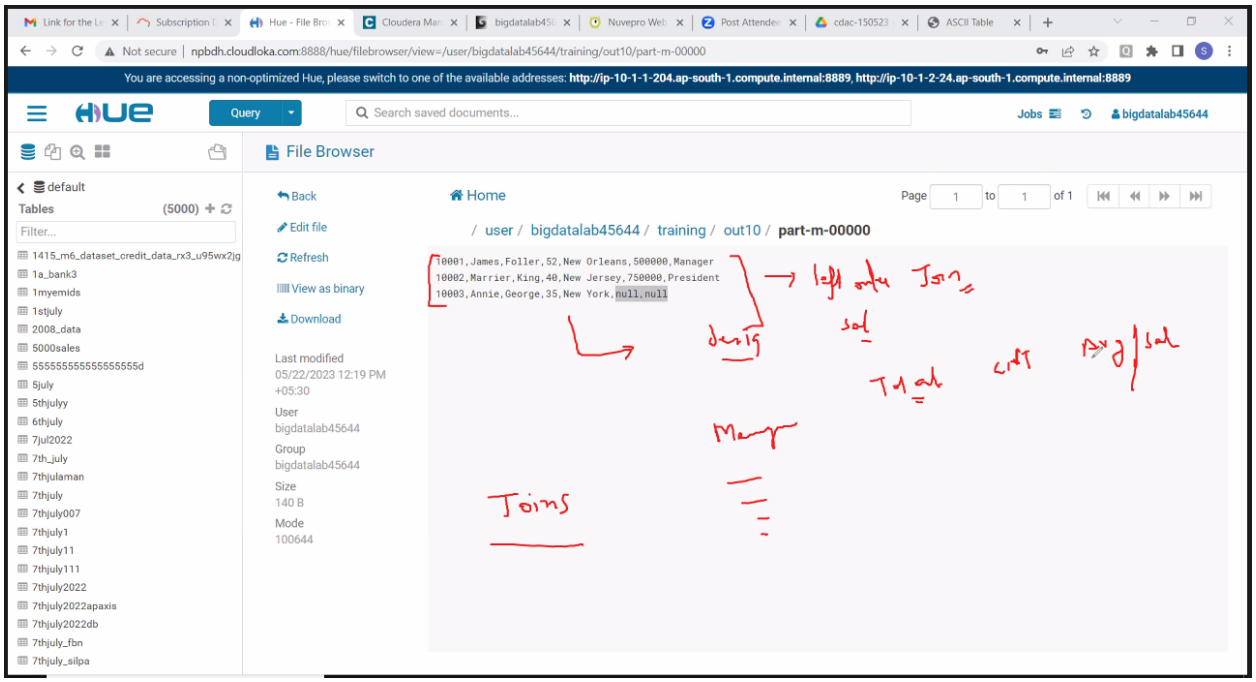
**#**

```
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop jar myjar.jar MapJoin
training/Employees.txt training/salary.txt training/desig.txt
training/out10
WARNING: Use "yarn jar" to launch YARN applications.
23/05/22 06:52:19 INFO client.RMProxy: Connecting to ResourceManager
at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/22 06:52:20 WARN mapreduce.JobResourceUploader: Hadoop
command-line option parsing not performed. Implement the Tool
interface and execute your application with T
oolRunner to remedy this.
23/05/22 06:52:20 INFO mapreduce.JobResourceUploader: Disabling
Erasure Coding for path:
/user/bigdatalab456422/.staging/job_1684298513961_1023
23/05/22 06:52:20 INFO input.FileInputFormat: Total input files to
process : 1
23/05/22 06:52:20 INFO mapreduce.JobSubmitter: number of splits:1
23/05/22 06:52:20 INFO Configuration.deprecation:
yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
Instead, use yarn.system-metrics-publisher.enable
d
23/05/22 06:52:20 INFO mapreduce.JobSubmitter: Submitting tokens for
job: job_1684298513961_1023
23/05/22 06:52:20 INFO mapreduce.JobSubmitter: Executing with tokens:
[]
```

```
23/05/22 06:52:20 INFO conf.Configuration: resource-types.xml not
found
23/05/22 06:52:20 INFO resource.ResourceUtils: Unable to find
'resource-types.xml'.
23/05/22 06:52:20 INFO impl.YarnClientImpl: Submitted application
application_1684298513961_1023
23/05/22 06:52:20 INFO mapreduce.Job: The url to track the job:
http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/applicati
on_1684298513961_1023/
23/05/22 06:52:20 INFO mapreduce.Job: Running job:
job_1684298513961_1023
23/05/22 06:52:28 INFO mapreduce.Job: Job job_1684298513961_1023
running in uber mode : false
23/05/22 06:52:28 INFO mapreduce.Job:  map 0% reduce 0%
23/05/22 06:52:33 INFO mapreduce.Job:  map 100% reduce 0%
23/05/22 06:52:33 INFO mapreduce.Job: Job job_1684298513961_1023
completed successfully
23/05/22 06:52:34 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=223551
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=326
                HDFS: Number of bytes written=140
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots
(ms)=2749
                Total time spent by all reduces in occupied slots
(ms)=0
                Total time spent by all map tasks (ms)=2749
                Total vcore-milliseconds taken by all map tasks=2749
                Total megabyte-milliseconds taken by all map
tasks=2814976
        Map-Reduce Framework
                Map input records=3
                Map output records=3
                Input split bytes=129
```

```
        Spilled Records=0
        Failed Shuffles=0
        Merged Map outputs=0
        GC time elapsed (ms)=57
        CPU time spent (ms)=540
        Physical memory (bytes) snapshot=262836224
        Virtual memory (bytes) snapshot=2590892032
        Total committed heap usage (bytes)=362807296
        Peak Map Physical memory (bytes)=262836224
        Peak Map Virtual memory (bytes)=2590892032
File Input Format Counters
        Bytes Read=101
File Output Format Counters
        Bytes Written=140
```




→ For MapInnerJoin

#

```
[bigdatalab456422@ip-10-1-1-204 ~]$ jar tvf myjar.jar
   25 Mon May 22 12:31:44 UTC 2023 META-INF/MANIFEST.MF
  387 Thu May 18 15:53:20 UTC 2023 .project
 2459 Fri May 19 16:03:30 UTC 2023 AllTimeHigh$MapClass.class
 2392 Fri May 19 16:03:30 UTC 2023 AllTimeHigh$ReduceClass.class
 1722 Fri May 19 16:03:30 UTC 2023 AllTimeHigh.class
 2475 Fri May 19 16:53:46 UTC 2023 AvgClosingPrice$MapClass.class
 2454 Fri May 19 16:53:46 UTC 2023 AvgClosingPrice$ReduceClass.class
 1732 Fri May 19 16:53:46 UTC 2023 AvgClosingPrice.class
 2337 Fri May 19 17:41:44 UTC 2023 WordCount$IntSumReducer.class
 2461 Fri May 19 17:41:44 UTC 2023 WordCount$TokenizerMapper.class
 1790 Fri May 19 17:41:44 UTC 2023 WordCount.class
 2454 Fri May 19 15:53:50 UTC 2023 AllTimeLow$MapClass.class
 2388 Fri May 19 15:53:50 UTC 2023 AllTimeLow$ReduceClass.class
 1734 Fri May 19 15:53:50 UTC 2023 AllTimeLow.class
 1242 Sat May 20 17:37:44 UTC 2023
MyPartitioner$CaderPartitioner.class
 2365 Sat May 20 17:37:44 UTC 2023 MyPartitioner$MapClass.class
 2905 Sat May 20 17:37:44 UTC 2023 MyPartitioner$ReduceClass.class
 2632 Sat May 20 17:37:44 UTC 2023 MyPartitioner.class
 2408 Thu May 18 17:48:56 UTC 2023 StockVolume$MapClass.class
 2349 Thu May 18 17:48:56 UTC 2023 StockVolume$ReduceClass.class
 1697 Thu May 18 17:48:56 UTC 2023 StockVolume.class
 2648 Sat May 20 15:42:48 UTC 2023 CityAvg2$CityCombineClass.class
 2269 Sat May 20 15:42:48 UTC 2023 CityAvg2$CityMapClass.class
 2639 Sat May 20 15:42:48 UTC 2023 CityAvg2$CityReduceClass.class
 2034 Sat May 20 15:42:48 UTC 2023 CityAvg2.class
 4839 Mon May 22 12:31:06 UTC 2023 MapInnerJoin$MyMapper.class
 1850 Mon May 22 12:31:06 UTC 2023 MapInnerJoin.class
 4760 Mon May 22 12:05:50 UTC 2023 MapJoin$MyMapper.class
 1817 Mon May 22 12:05:50 UTC 2023 MapJoin.class
 2456 Sat May 20 15:04:24 UTC 2023
StockVolumeWithCombiner$MapClass.class
 2397 Sat May 20 15:04:24 UTC 2023
StockVolumeWithCombiner$ReduceClass.class
 1813 Sat May 20 15:04:24 UTC 2023 StockVolumeWithCombiner.class
  640 Thu May 18 17:00:00 UTC 2023 .classpath
```

#

```
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop jar myjar.jar MapInnerJoin
training/Employees.txt training/salary.txt training/desig.txt
training/out11
```

```
WARNING: Use "yarn jar" to launch YARN applications.
23/05/22 07:05:19 INFO client.RMProxy: Connecting to ResourceManager
at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/22 07:05:20 WARN mapreduce.JobResourceUploader: Hadoop
command-line option parsing not performed. Implement the Tool
interface and execute your application with T
oolRunner to remedy this.
23/05/22 07:05:20 INFO mapreduce.JobResourceUploader: Disabling
Erasure Coding for path:
/user/bigdatalab456422/.staging/job_1684298513961_1086
23/05/22 07:05:20 INFO input.FileInputFormat: Total input files to
process : 1
23/05/22 07:05:20 INFO mapreduce.JobSubmitter: number of splits:1
23/05/22 07:05:20 INFO Configuration.deprecation:
yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
Instead, use yarn.system-metrics-publisher.enable
d
23/05/22 07:05:20 INFO mapreduce.JobSubmitter: Submitting tokens for
job: job_1684298513961_1086
23/05/22 07:05:20 INFO mapreduce.JobSubmitter: Executing with tokens:
[]
23/05/22 07:05:21 INFO conf.Configuration: resource-types.xml not
found
23/05/22 07:05:21 INFO resource.ResourceUtils: Unable to find
'resource-types.xml'.
23/05/22 07:05:21 INFO impl.YarnClientImpl: Submitted application
application_1684298513961_1086
23/05/22 07:05:21 INFO mapreduce.Job: The url to track the job:
http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/applicati
on_1684298513961_1086/
23/05/22 07:05:21 INFO mapreduce.Job: Running job:
job_1684298513961_1086
23/05/22 07:05:28 INFO mapreduce.Job: Job job_1684298513961_1086
running in uber mode : false
23/05/22 07:05:28 INFO mapreduce.Job:  map 0% reduce 0%
23/05/22 07:05:33 INFO mapreduce.Job:  map 100% reduce 0%
23/05/22 07:05:34 INFO mapreduce.Job: Job job_1684298513961_1086
completed successfully
23/05/22 07:05:34 INFO mapreduce.Job: Counters: 33
       File System Counters
               FILE: Number of bytes read=0
               FILE: Number of bytes written=223556
               FILE: Number of read operations=0
               FILE: Number of large read operations=0
               FILE: Number of write operations=0
```

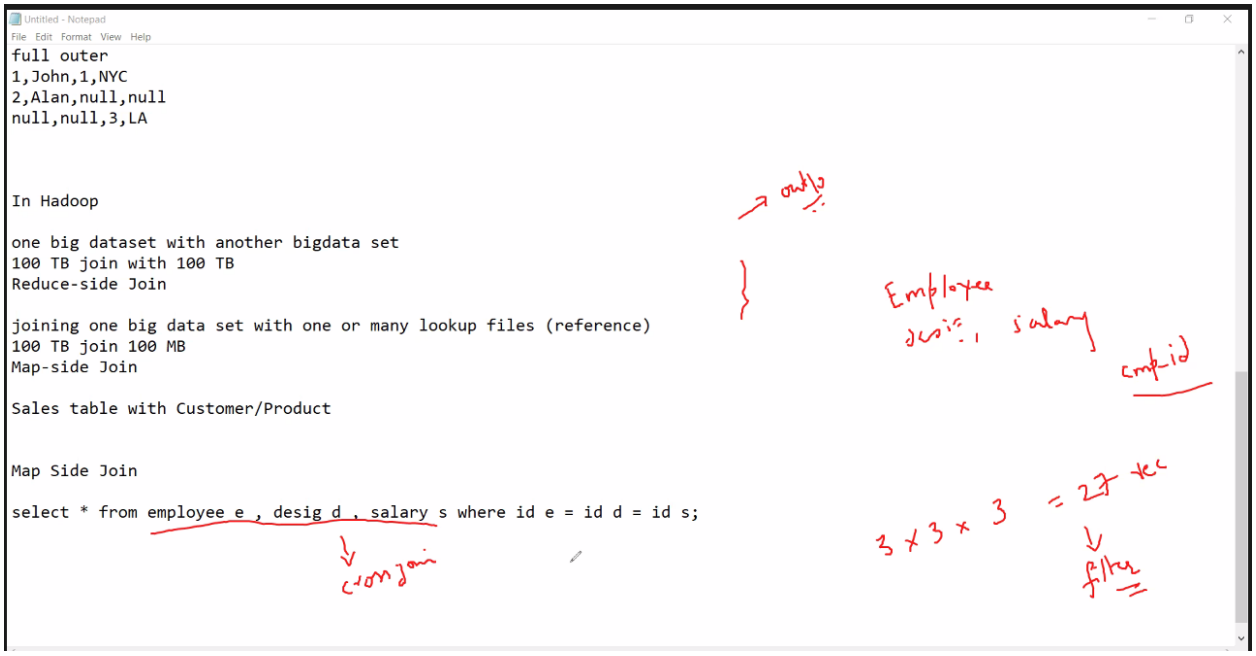
```
                HDFS: Number of bytes read=326
                HDFS: Number of bytes written=99
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots
(ms)=3267
                Total time spent by all reduces in occupied slots
(ms)=0
                Total time spent by all map tasks (ms)=3267
                Total vcore-milliseconds taken by all map tasks=3267
                Total megabyte-milliseconds taken by all map
tasks=3345408
        Map-Reduce Framework
                Map input records=3
                Map output records=2
                Input split bytes=129
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=62
                CPU time spent (ms)=660
                Physical memory (bytes) snapshot=231632896
                Virtual memory (bytes) snapshot=2574462976
                Total committed heap usage (bytes)=239599616
                Peak Map Physical memory (bytes)=231632896
                Peak Map Virtual memory (bytes)=2574462976
        File Input Format Counters
                Bytes Read=101
        File Output Format Counters
                Bytes Written=99
```

→

Untitled - Notepad

File Edit Format View Help

```
full outer
1,John,1,NYC
2,Alan,null,null
null,null,3,LA



In Hadoop

one big dataset with another bigdata set
100 TB join with 100 TB
Reduce-side Join

joining one big data set with one or many lookup files (reference)
100 TB join 100 MB
Map-side Join

Sales table with Customer/Product


Map Side Join

select * from employee e , desig d , salary s where id e = id d = id s;
```




Untitled - Notepad

File Edit Format View Help

```
joining one big data set with one or many lookup files (reference)
100 TB join 100 MB
Map-side Join

Sales table with Customer/Product


Map Side Join

select * from employee e , desig d , salary s where id e = id d = id s;



select e.* , s.salary, , d.designation from
employees e left outer join salary s on e.id = s.id

left outer join
designation d
on e.id = d.id

join
table4 on .........
```