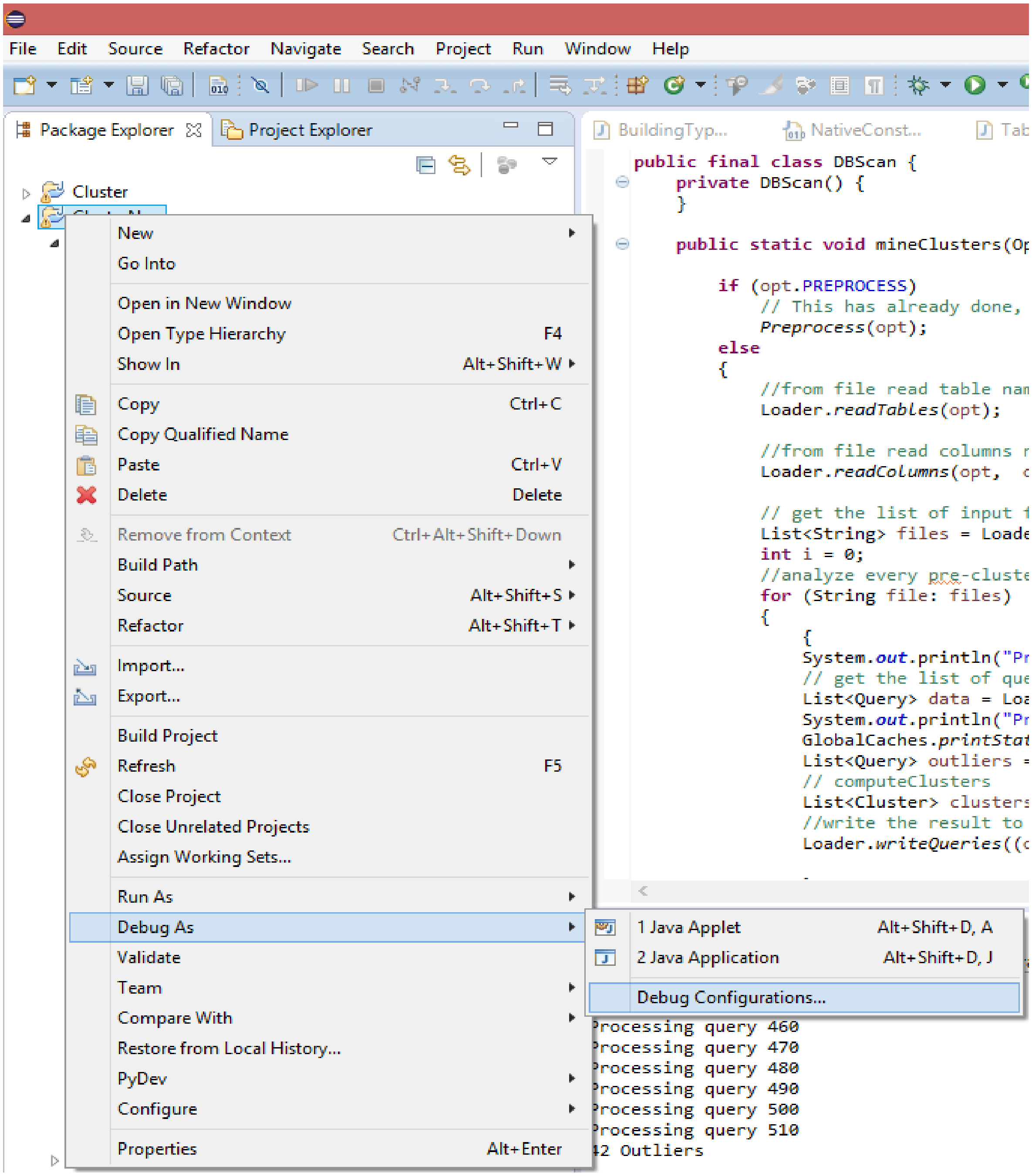
Task 0. First start

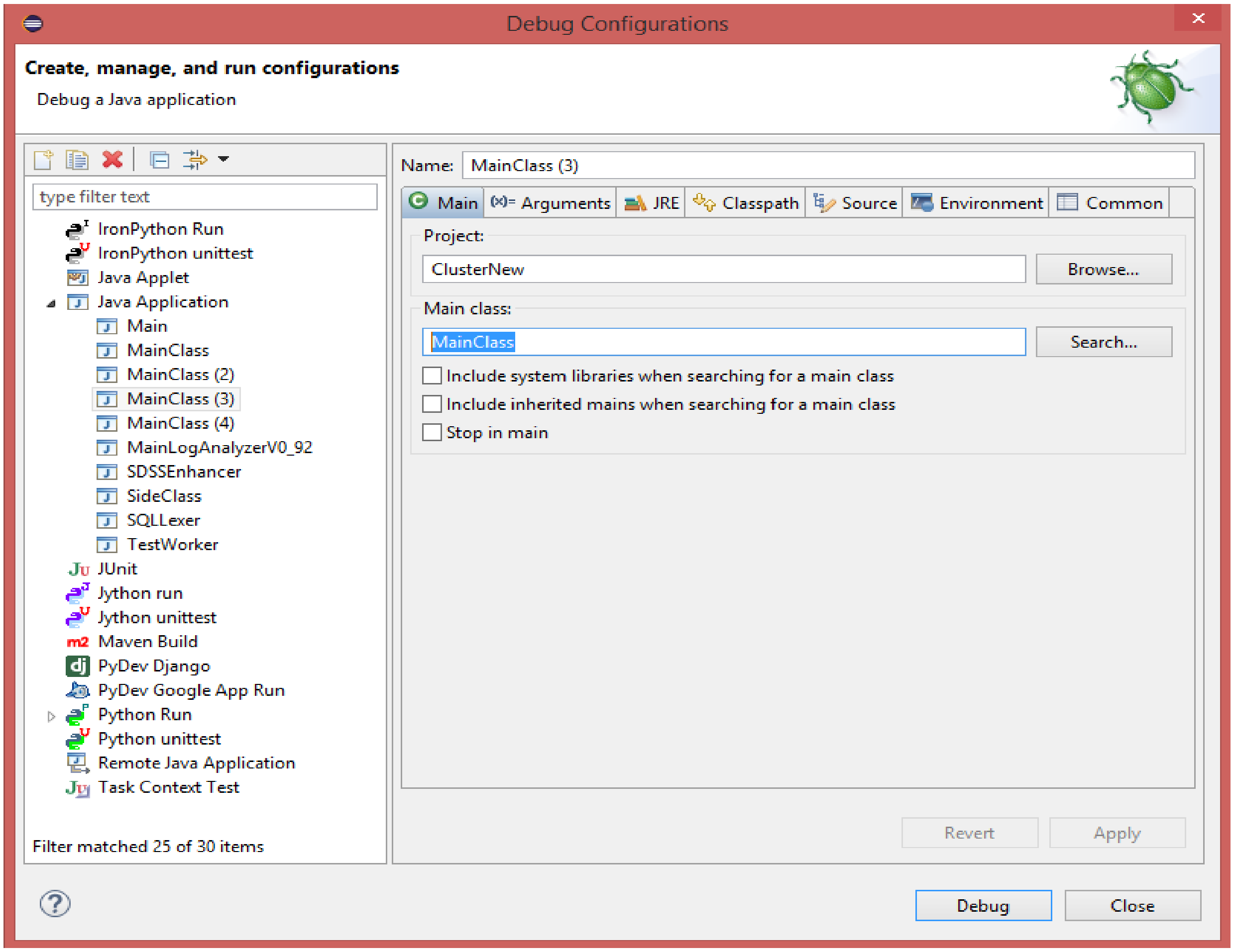
1. Download and install Eclipse: http://www.eclipse.org/downloads/packages/eclipse-ide-java-and-dsl-developers/mars2
    a. Import project ClusterNew (File -> Import-> Existing project into workspace)
    b. Set configuration



```
public final class DBScan {
    private DBScan() {
    }

    public static void mineClusters(Op

        if (opt.PREPROCESS)
            // This has already done,
            Preprocess(opt);
        else
        {
            //from file read table nam
            Loader.readTables(opt);

            //from file read columns r
            Loader.readColumns(opt,  c

            // get the list of input f
            List<String> files = Loade
            int i = 0;
            //analyze every pre-cluste
            for (String file: files)
            {
                {
                System.out.println("Pr
                // get the list of que
                List<Query> data = Loa
                System.out.println("Pr
                GlobalCaches.printStat
                List<Query> outliers =
                // computeClusters
                List<Cluster> clusters
                //write the result to
                Loader.writeQueries((c
```

```
Processing query 460
Processing query 470
Processing query 480
Processing query 490
Processing query 500
Processing query 510
42 Outliers
```

    c. Set Main Class as ‘MainClass’



Debug Configurations

Create, manage, and run configurations

Debug a Java application

Name: MainClass (3)

Project: ClusterNew

Main class: MainClass

□ Include system libraries when searching for a main class

□ Include inherited mains when searching for a main class

□ Stop in main

Filter matched 25 of 30 items

d. Set command line arguments

Debug Configuration-> Arguments

Command line arguments look like:

-FILE_C_OUTPUT "C:\Work\in_out\new\out.csv" -FILE_CLMN_OUTPUT "C:\Work\in_out\new\out_clmn.csv" -FILE_TBL_OUTPUT "C:\Work\in_out\new\out_tbl.csv" -FILE_INPUT "C:\Work\in_out\sample.csv" -FILE_TABLES "C:\Work\in_out\tables.csv"

Argument	Description
FILE_C_OUTPUT	Output file (file with clusters)
FILE_CLMN_OUTPUT	File with columns distributions
FILE_TBL_OUTPUT	File with tables names and counts of rows for each table
FILE_INPUT	Input file with SQL statements in intermediate format

The samples of the files you can find in 'Samples\Task0' folder

2. Build project
3. Become familiar with the code
4. Run, interpret and DESCRIBE the results. Clue: Use SkyServer data scheme description: http://skyserver.sdss.org/dr1/en/help/browser/browser.asp

5. Change threshold parameters (distance threshold and min count of queries to became a cluster), evaluate changes. What threshold parameters do you think are optimal? Why? How to set the threshold parameters?
6. What in the result doesn't have interpretation meaning? Why?
7. What queries can we exclude from the input datasets? Clue: this strongly depends on the answer to the previous question
8. Have a look at the query log in Oracle database.
   Connection's settings:

Server: marsara.ipd.kit.edu
Database (SID): student
Port: 1521
Username: bdcourse
Password: bdcourse

9. Become familiar with SkyServer database, we expect you to have some domain knowledge http://skyserver.sdss.org/dr12/en/home.aspx
10. Have a look how to collect logs of SkyServer http://skyserver.sdss.org/log/en/traffic/sql.asp?