

**Keywords Search on Relational Database**

Support multiple-keywords with relations Search



**KSE Report**

Deliverable Version 3.0

Report Version 3.0

Status: Final

**Researcher: Jiahui Zhao / Zhajy110**

**Supervisor: Dr. Jerry (Jixue) Liu**

**Project No. : #32**

**Contents**

[**1** **Introduction** - 6 -](#_Toc465100056)

[1.1 Background - 6 -](#_Toc465100057)

[1.2 Motivation - 6 -](#_Toc465100058)

[1.3 Requests - 7 -](#_Toc465100059)

[1.4 Solution - 7 -](#_Toc465100060)

[1.5 Scope - 8 -](#_Toc465100061)

[1.6 Plan and Implement Schedule - 8 -](#_Toc465100062)

[1.7 cost - 9 -](#_Toc465100063)

[1.8 Team member and roles - 9 -](#_Toc465100064)

[1.9 Meeting Agenda and Meeting Minutes - 9 -](#_Toc465100065)

[1.10 Impact statement - 9 -](#_Toc465100066)

[**2** **KSE search engine** - 10 -](#_Toc465100067)

[2.1 Design / approach - 10 -](#_Toc465100068)

[2.2 Functions - 12 -](#_Toc465100069)

[2.2.1 Menu function: - 13 -](#_Toc465100070)

[2.2.2 Search function: - 13 -](#_Toc465100071)

[2.2.3 Display function: - 13 -](#_Toc465100072)

[2.2.4 Export function: - 14 -](#_Toc465100073)

[2.3 Components and facilities - 14 -](#_Toc465100074)

[2.4 Result - 15 -](#_Toc465100075)

[2.5 Configure file - 15 -](#_Toc465100076)

[2.6 Data flow and procedure - 16 -](#_Toc465100077)

[2.6.1 Data flow - 16 -](#_Toc465100078)

[2.6.2 Configure procedure - 16 -](#_Toc465100079)

[2.6.3 Keywords search Procedure - 17 -](#_Toc465100080)

[2.7 Version of KSE - 18 -](#_Toc465100081)

[2.7.1 Version 1 - 18 -](#_Toc465100082)

[2.7.2 Version 2 - 19 -](#_Toc465100083)

[2.7.3 Version 3 - 21 -](#_Toc465100084)

[2.8 Testing - 22 -](#_Toc465100085)

[2.9 Use case story - 23 -](#_Toc465100086)

[**3** **Summary** - 24 -](#_Toc465100087)

[3.1 Management - 24 -](#_Toc465100088)

[3.2 Milestone - 25 -](#_Toc465100089)

[3.3 Deliverable - 26 -](#_Toc465100090)

[3.4 Risk - 27 -](#_Toc465100091)

[3.5 Discussion and future work - 27 -](#_Toc465100092)

1. **Introduction**
   1. Background

With the amount of available text data in relational databases growing rapidly, a huge amount of information is stored in relational databases, but information discovery on relational databases is not well supported. Even though the major RDBMSs have provided full-text search capabilities, the problem is that they still require users to have knowledge of the database schemas and use a structured query language to search information.

Keywords search is the most popular information discovery method now, such as people using keywords search on Google, Facebook to browsing internet. Because of using keywords search the user does not need to know either a query language or the underlying structure of the data. There are some keywords search method on relational database have been released in the current market, like “Oracle Text”, 什么 search

* 1. Motivation

The motivation of this project is that my client want to have is a simplified way to realize multiple keywords with relations search on relational database rather than writing a complex SQL query.



Figure 1: data set example

For example, If my client want to search 44 from full-text of this data set example (figure 1), as we know the schema of data set example, the SQL query should be like: *select \* from Sample-Table where age=”44” or workclass=”44” or edunum=”44” or occu=”44” or race=”44” or sex=”44” or country=”44” or salary=”44”.* If my client want to search 44 and female, the SQL query should be like: *select \* from Sample-Table where age=”44” or workclass=”44” or edunum=”44” or occu=”44” or race=”44” or sex=”44” or country=”44” or salary=”44” AND (age=”female” or workclass=” female” or edunum=” female” or occu=” female” or race=” female” or sex=” female” or country=” female” or salary=” female”)*

It needs to scan every row and every column of relational database to complete these search, how to write the SQL query what if we don’t know the schema of database? How complex SQL query if people want to search 10 keywords or more? How long does these need if the relational database has million tuples? What is the efficiency of these search?

This project aim to help my client to search information from relational database more quickly, as well as have more facility to search multiple keywords with “and” “or” relation on full-text of relational database.

* 1. Requests

The main request of my client is a simplified way to realize multiple keywords search on relational database rather than a complex SQL query.

|  |  |  |  |
| --- | --- | --- | --- |
| **Request Name** | **Request No.** | **Request Date** | **Current Status** |
| Multiple keywords search with ‘and’ relation | 2.1 | 06/09/2016 | Completed at every version |
| Multiple keywords search with ‘or’ relation | 2.2 | 06/09/2016 | Completed at every version |
| Result in table | 4.1 | 18/09/2016 | Completed at version 1 |
| Configuration file | 10.1 | 10/10/2016 | Completed at version 2 & 3 |
| Export result into file | 10.2 | 10/10/2016 | Completed at version 3 |
| ID column number auto catch | 11.1 | 17/10/2016 | Completed at version 3 |

Figure 3: request change log of project

* 1. Solution

We delivered this Keywords Search Engine (KSE) search engine at the end of the project to my client. The KSE search engine be used to support user to implement the keywords search on relational database, it will also resolve the search effective problem of relational database for the client, reduce the search time costs to improve the effective of search.

KSE is designed to ensure the proper development and implementation of multiple keywords search, as well as with “and” or “or” relations between the keywords. From and search perspective, when the time of searching is reduced, there is more efficiency and quickly, as they can rely on the KSE to search multiple keywords without any knowledge of the schema of database and complex SQL query. KSE is running on Hadoop + H-Base in Linux operating system, the H-Base provides Bigtable-like capabilities on top of Hadoop and HDFS. H-Base create key-value pairs with hash function to store keys and the related values.

* 1. Scope

The scope of this project is limited to textual relational database only, it cannot be used to search for any other type of databases. And the relational database needs to have trigger to recording any change of the table into a temp table if user want to use the auto update function.

The KSE search engine request the table of oracle has index column by itself. And the search engine needs to know which column is the ID column and specify the column name in configure file.

* 1. Plan and Implement Schedule

Based on the requests of client, this project is planned implement in 13 weeks, also 13 weeks is a whole study period of UniSA, and the plan of project has 4 steps:

The first step is 1-3 week, the main job of this step is collecting the information from client and all the requests, design and give client couple of solutions and discuss every possible solution with client to make sure which one is the best, and match all requests of client.

The second step is 3-6 week, build up the implement environment, debug the software will be used, analysis the type and structure of the data set which is given by the client, and design the structure with data flow of H-Base.

The third step is 6-11 week, this step is a main programming step. In this step, we regularly meet with client to notify them general information about project implement, we demonstrate the project deliverable to client every week, and collect any change or new request from the client.

The fourth step is 12-13 week, this is the final debug and reporting step, we debug and test the final deliverable and submit to client, finish the project report and presentation, make sure the client can fully understand the project and use the final deliverable effortless.

The following chart is the real implement schedule of project, the main frame is followed by the plan of project, but there are some adjustments in each step as client gave some new requests and changes.



Figure 4: implement schedule of project

* 1. cost

As this project is a UniSA internal project, the client also is the academic supervisor of this project is Dr. Jerry (Jixue) Liu, this project does not have any budget or cost.

* 1. Team member and roles

This is an individual project, only one team member and role is Jiahui Zhao, and Jiahui Zhao do all responsibilities for this project.

* 1. [Meeting Agenda](http://www.projectmanagementdocs.com/project-documents/meeting-agenda.html) and Meeting Minutes

We regular have a meeting per week with client to discuss current status and collect any change or request from client.

* Week 1 : 1pm-1.30pm 29th July 2016
* Week 2 : 2pm-4pm 1st August 2016
* Week 4 : 9am-10am 17th August 2016
* Week 6 : 1pm-1.30pm 31th August 2016
* Week 7 : 1pm-1.30pm 7th September 2016
* Week 10 :1pm-2pm 6th October 2016
* Week 12 : 17pm-18pm 17th October 2016

The [Meeting Agenda](http://www.projectmanagementdocs.com/project-documents/meeting-agenda.html) and Meeting Minutes and weekly project report were sent to client each week. Any review request please contact the client to get the copy.

* 1. Impact statement
     1. Improve search effective:

Quickly keywords search on relational database

* + 1. keywords search with relations:

Multiple keywords with “and” or “or” specific relation

* + 1. Reduce search cost:

Reduce both of search time cost and search engine cost.

1. **KSE search engine**
   1. Design / approach

KSE search engine to implement multiple keywords search with relations for relational database is to obtain a complete Key-Value pairs on H-table based on Hadoop plus H-Base. This approach is not making any change on relational database, it is convert the data of each cell of relational database to the keys of Key-Value pair, and store the ID of each tuple which is contained these keys as values of Key-Value pair.

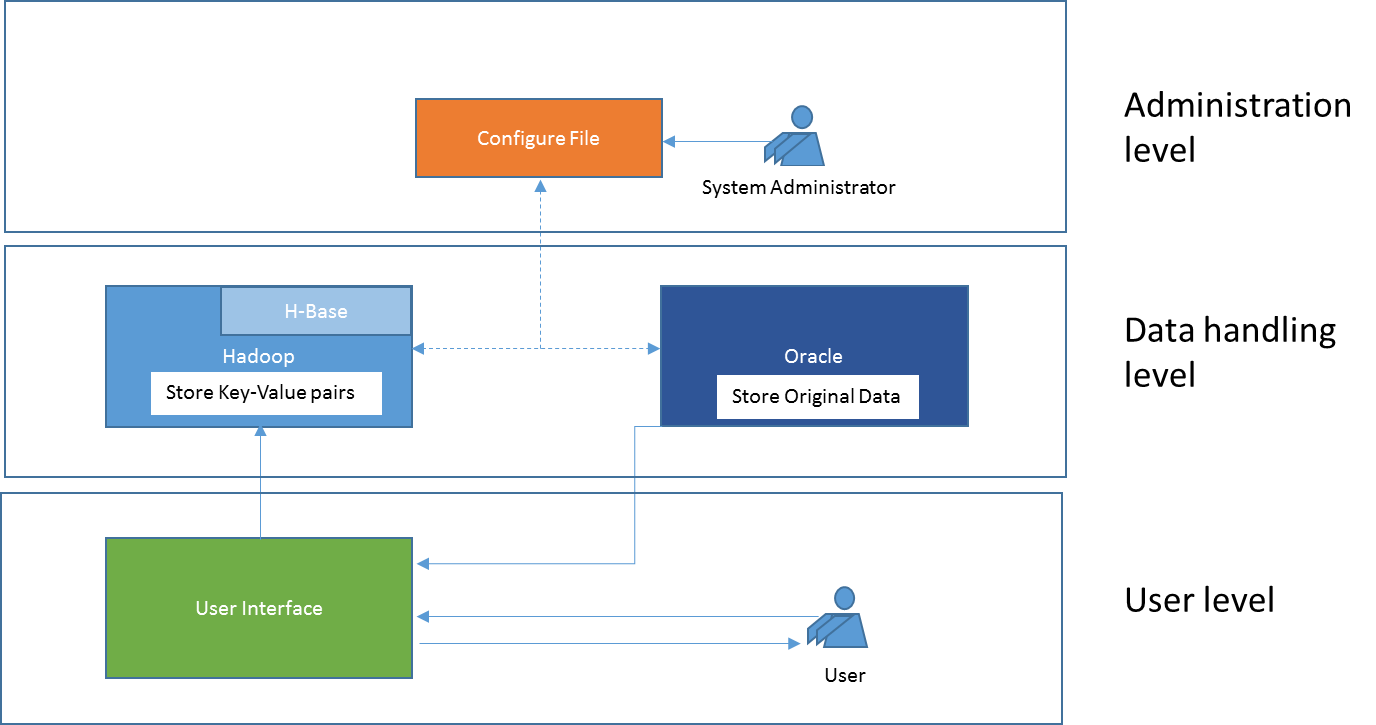
KSE typically consist of processing the keywords search on 3 levels (figure1), which includes the user level, data handling level, and administration level.

Figure 2: 3 levels of KSE

The user level is a visual level which has an interface to users to input the keywords, display result and other operation buttons such as menu button, navigation buttons and export button. The data handling level is background processing level which includes the inputted keywords convert and search from the Key-Values pairs on the H-Base and regular search with returned ID on relational database. The administration level is an assistance level, this level has an external configure file which has the basic configure information to support the KSE initial H-Base with the relational database as the client specified, and also it has some other information, for instance, the H-Base auto update period on seconds, when the relational data has some changes on data, this configure information will let the KSE frequently check the relational data if there has any changes, and make sure the Key-Value pairs of H-base is synchronous updated after each setting period.

To implement the key-value pair, we applied 4 possible solutions as following (figure4) and discuss which is most suitable to client.

* + 1. Multiple timestamp

All values stored in the same column family and column, different value has different timestamp, same value may have different timestamps to be distinguished. When the oracle table changed, the value would have a new version with new timestamp, however, the old one was still keeping in the pairs with old timestamp.

* + 1. Single timestamp

All values stored in the same column family and column, different value has different timestamp, same value can only have one version that means only one timestamp exit of the value. When the oracle table changed, the values have be deleted the old version and add new version.

* + 1. Column name + marked delete:

All values stored in the same column family but different columns with different column names and store empty value, the column name as the value should be stored. When the oracle changed, the value of column is change to “-” as marked delete, and add new column which store empty value.

* + 1. Column name

All values stored in the same column family but different columns with different column names and store empty value, the column name as the value should be stored. When the oracle changed, there is only one column exist with the same column name. Not exist column should be deleted to save the storage space.

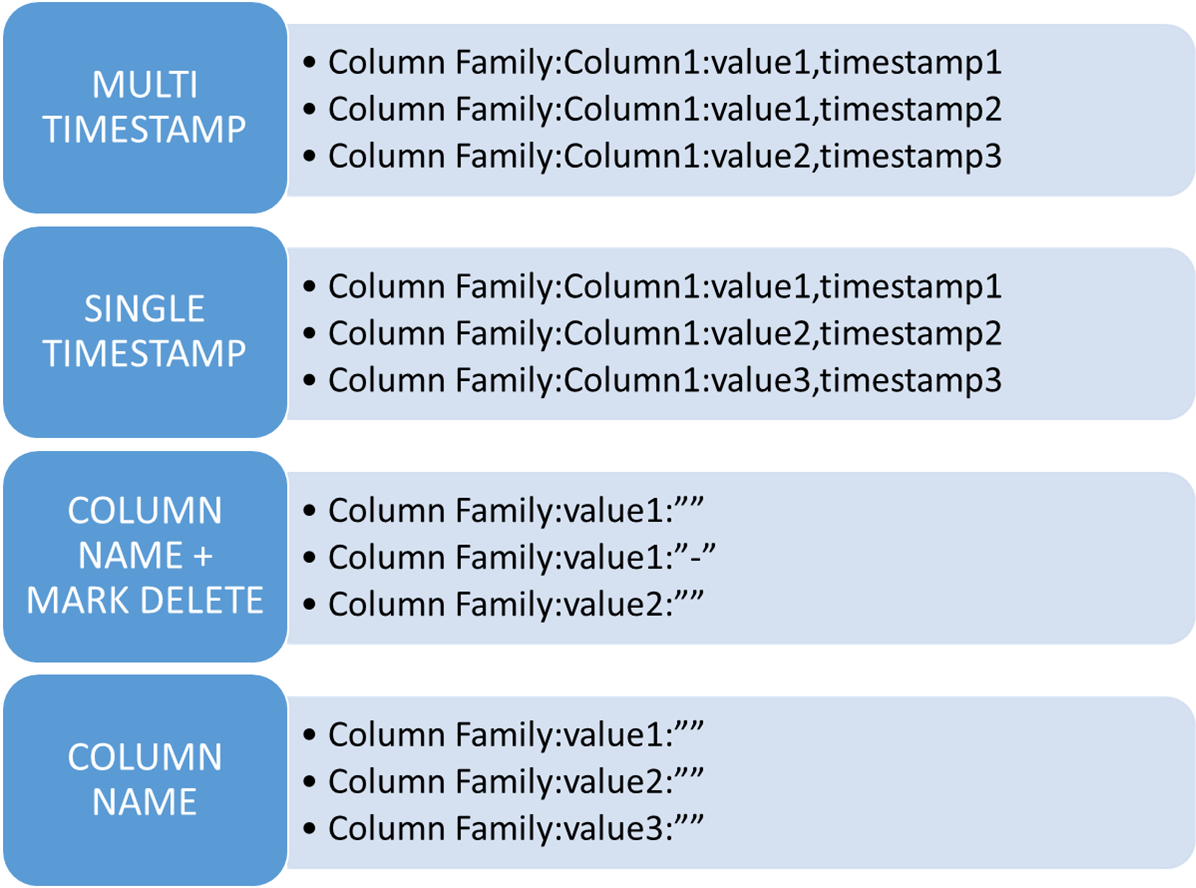
****

Figure 5: key-value pairs 4 implement ways compare

After we discussed with client the advantages and disadvantages of all 4 key-value pair implement ways, the client chosen the 2.1.4 column name way which would save more storage space and reduce the redundancy of the H-Base.

* 1. Functions

The functions of the KSE search engine are in 4 parts：

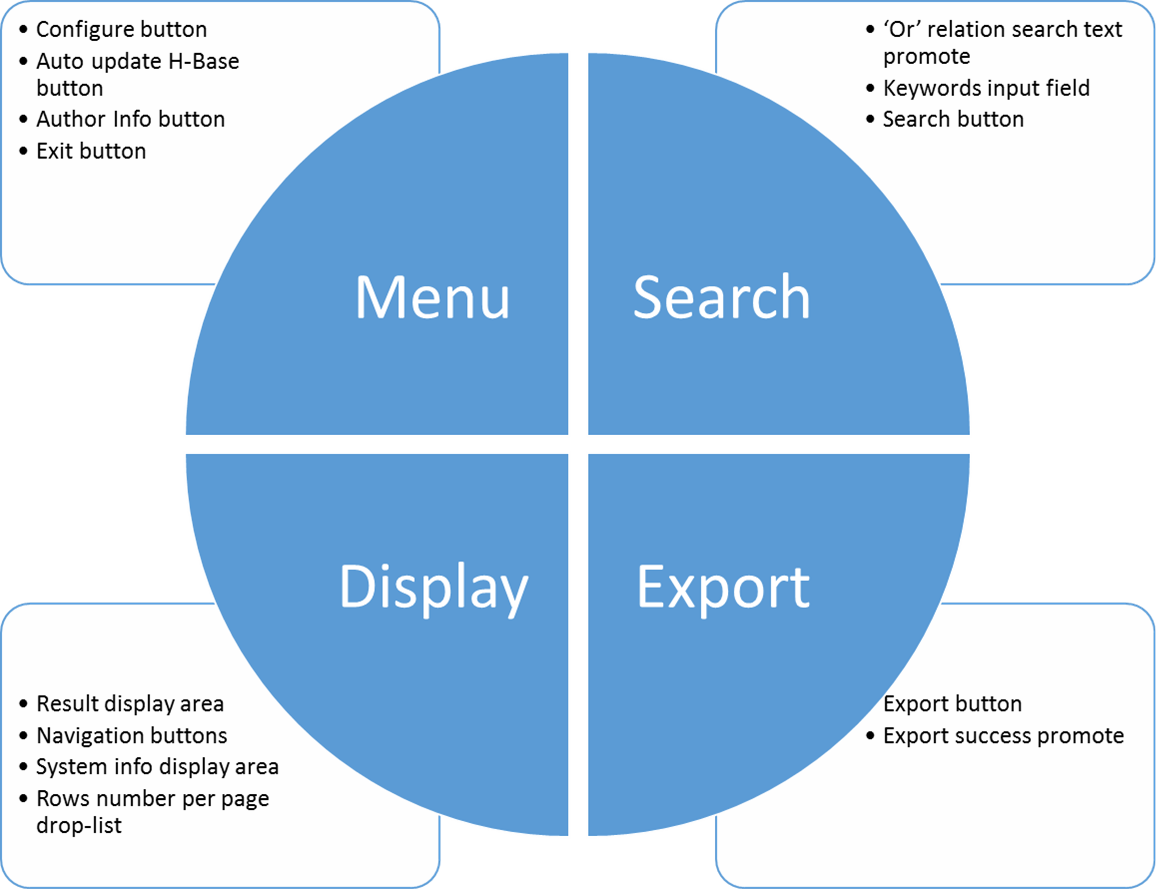
****

Figure 6: function chart of project

* + 1. Menu function:

Configure button: read the configure file, connect the oracle and H-Base, initial H-Base with oracle

Auto update H-Base switch: have two status, ON and OFF, when the status is ON, the user interface will auto update H-base with any change of oracle.

Author info button: display author’s information

Exit button: exit the search engine

* + 1. Search function:

Keywords input field: the keywords input field is the area let user to input the excepted keywords to do search.

Search button: to start the search, or click the “Enter” key of keyboard when the cursor focused on the input field.

Default relation promote: there some TEXT promote for the or relation between keywords search using “||or” keyword in the keywords input field, when the cursor not focused on the input field, the TEXT will display, when the cursor focused on the input field, the TEXT will not display.

* + 1. Display function:

Result display area: display the returned result from oracle in table and paged

Navigation button: navigate the pages of the result returned.

System info display area: display the original input text, display the searching time on

H-base and Oracle, display total pages and current page.

Rows number drop-list: the user can select the rows number of result at per page, after select the user interface will initial the result with the selected number.

* + 1. Export function:

Export button: export the result into excel file for each search. Auto create the file name with valid keywords and current system time.

Export info promote: promote the export success information, the file saved path and the file name.

* 1. Components and facilities

The components of KSE include the visual parts such as buttons, display area, and some non-visual parts such as configure file.

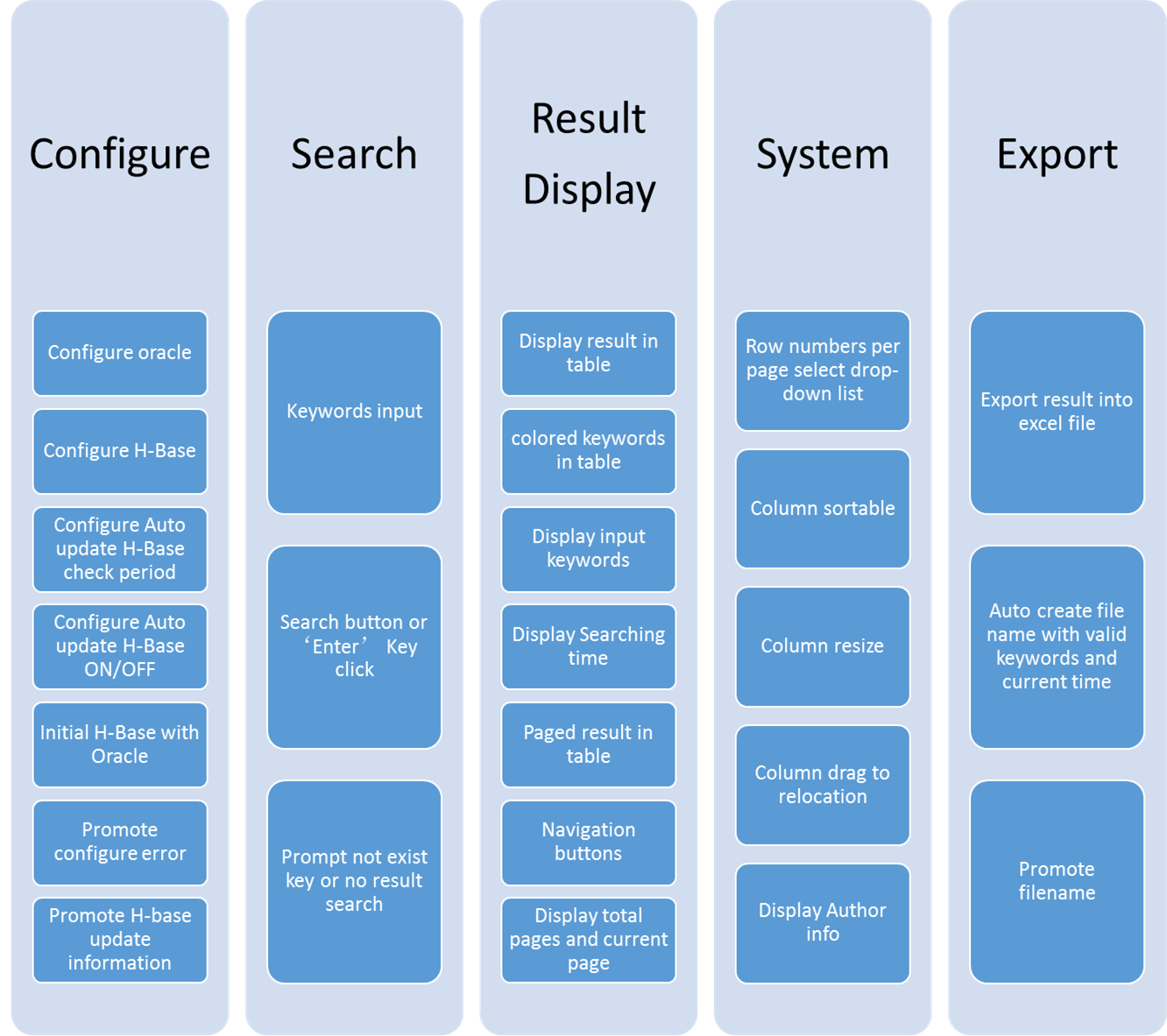
****

Figure 7: components chart of project

* 1. Result

The following results were produced for KSE keywords search overview on sample data set, the sample data set is a part of the relational database from client, it contains the first 1000 rows data only.

|  |  |  |  |
| --- | --- | --- | --- |
| **Search type** | **keywords** | **Return** | **Status** |
| Search single keyword | 44 | 28 tuples | Match SQL |
| Search multiple keywords  Default “and” relation | 44 female | 7 tuples | All match  SQL query |
| Search multiple keywords  Default “and” relation | 88 female | 0 tuples | Match SQL |
| Search multiple keywords  With “or” relation | 44 female ||or  33 male | 31 tuples | All match SQL query |
| Search multiple keywords  With “or” relation  And not exist keyword | 44 female ||or  33 male ||or  orange | 31 tuples | All match  SQL query |

Figure 8: Search results

* 1. Configure file

The KSE search engine need an external configure file to specify the oracle connection information and H-Base connection information, the configure file is wrote by the system administrator, it includes the oracle connection which have oracle URL, user login name, user login password, IP or Address, port number and SID of oracle, the H-Base connection information, which has the IP or address and the port number.(in this project we are using the local H-Base server and set the server login without user name and passwords, so we don’t need to specify the login information here.). The configure file have the table name of oracle, which is let the search engine know connecting the specify table of oracle when the search engine initial or auto updating H-Base. Specify ID column name is needed in this search engine, as every words will be the Key in Key-values pairs of H-Base, and ID as value in Key-value pairs. The location of ID column does not matter as the search engine will get the column number from metadata, even the ID column at the front or middle or last column. Last thing of configure file is auto update H-Base period, this is using to let the search engine auto check the temp table of oracle which content the changes of the table frequency by the specify period, when the temp table has some records, the search engine will read the records and update the H-Base with the changes (delete, add, update is separate to delete old record and add one record), and then search engine will clean the temp table to prepare the next change.

* 1. Data flow and procedure
     1. Data flow

The data flow of this project is crossing through all three levels, first the search engine needs to initial H-Base with oracle by configure file assistance, the data from oracle to H-base be Key-value pairs, the user give the keywords to user interface, then the user interface convert the keywords to key and search in the H-base, if these keys are available and could return the value of key-value pairs, the value to be ID to search into oracle with regular SQL query and return the original data tuples to user interface, the user interface organize the tuples into a table and display to user. The auto update H-base function also need the configure file assistance, any changes of oracle will be recorded into the temp table, the user interface frequency check the temp table, update H-base if there are any records in temp table, and then clean the temp table after update.

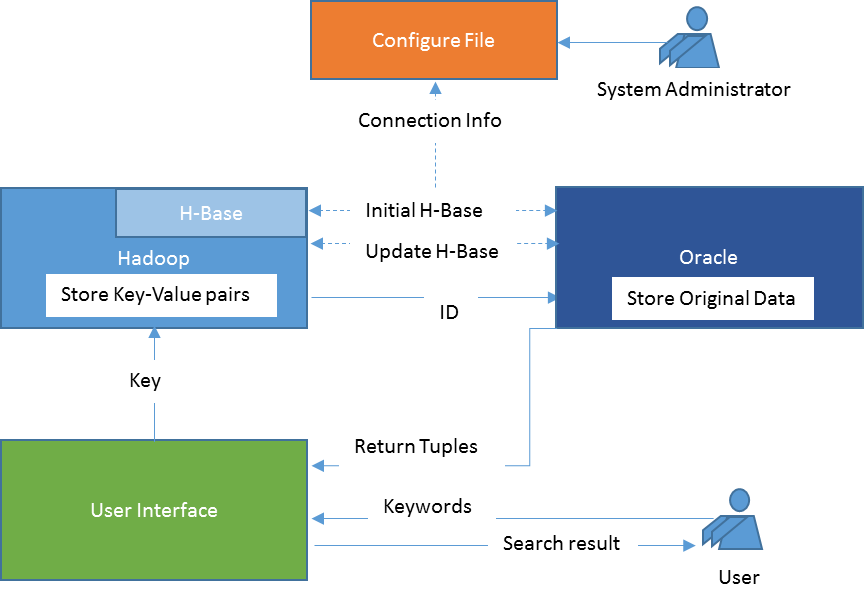
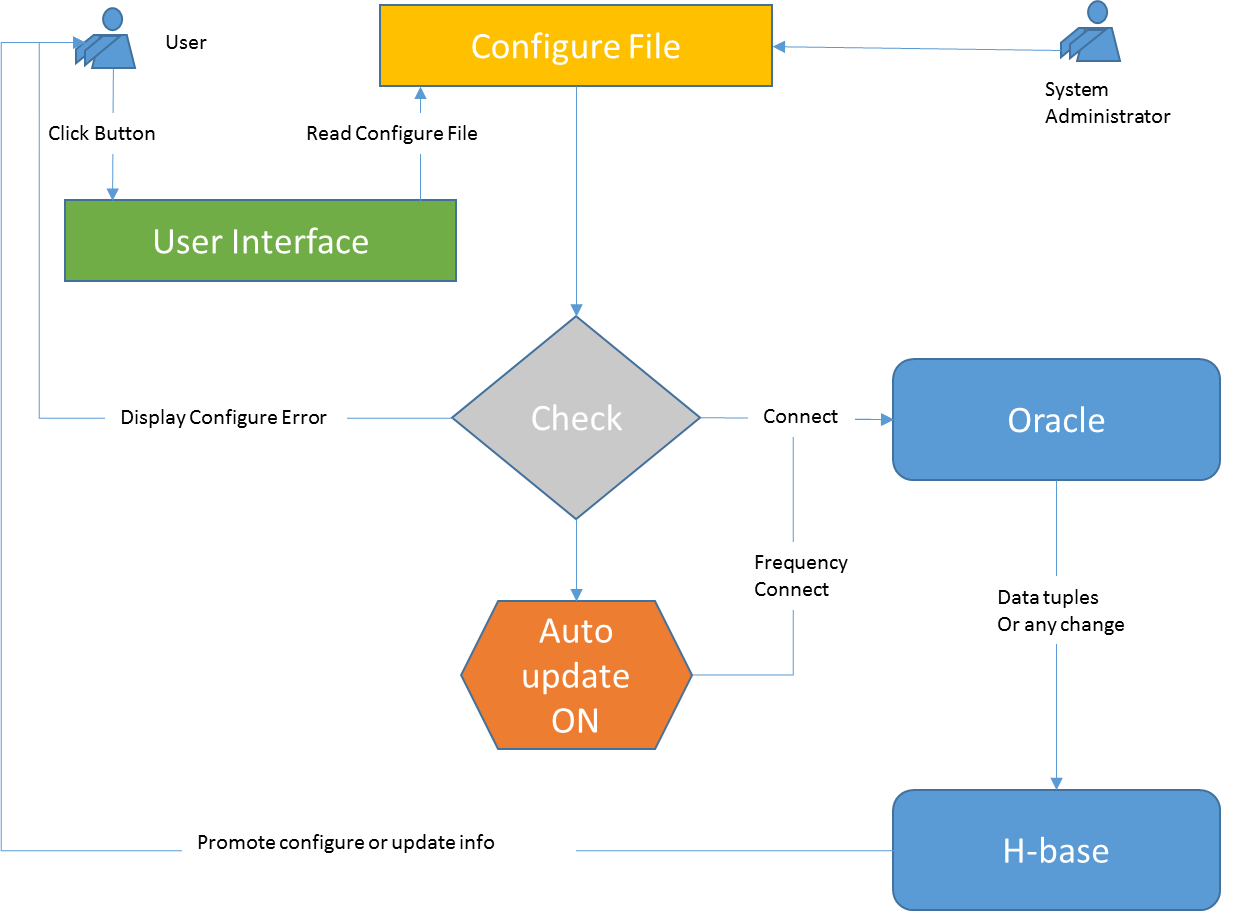
****

Figure 9: data flow chart of project

* + 1. Configure procedure

When user click the configure button, the user interface will read the configure file and set the oracle connection information and H-Base connection information, if all configure are correct, the user interface will do the initial H-Base with oracle data tuples, creates key-value pairs for oracle data in H-base, promote success information and keywords amount to user when the initial finished.

**** Figure 10: configure procedure chart of project

When user click the auto update H-Base button and switch the function status is ON, the user interface is going to read the configure file and set the wait period to the specific seconds, and then the user interface check the temp table of oracle if it has any records in there, if the temp table have some records, the user interface will update H-Base with these records, delete the records from temp table when update finish, and then going to waiting status till the next check,

* + 1. Keywords search Procedure

When the user input some keywords, the user interface will check if it is valid, like only contain blank, if there are some other keywords other than blank, the user interface will check if the input does contain or relation is specific grammar “||or”, if there has or relation between keywords, The user interface will separate the keywords before the first “||or” and check the left if there is more “||or” left using induction, if there is no more “||or” left, the user interface send the separated input to H-base as key to search, the H-base return the values related the keys to the user interface, and then using or relation which contained in the user input to joint all returns of separated input, the joint result will be the ID list and using general SQL query to search from oracle, the oracle return the original data tuples to user interface, the user interface fill the data into a table and paged as user selected row number in the drop list, automatically colored the keywords with model of table and display to the user. If there is any error such as, no result related the input of search, or not exist key in H-base, the user interface will promote the information to user when the final result displayed.

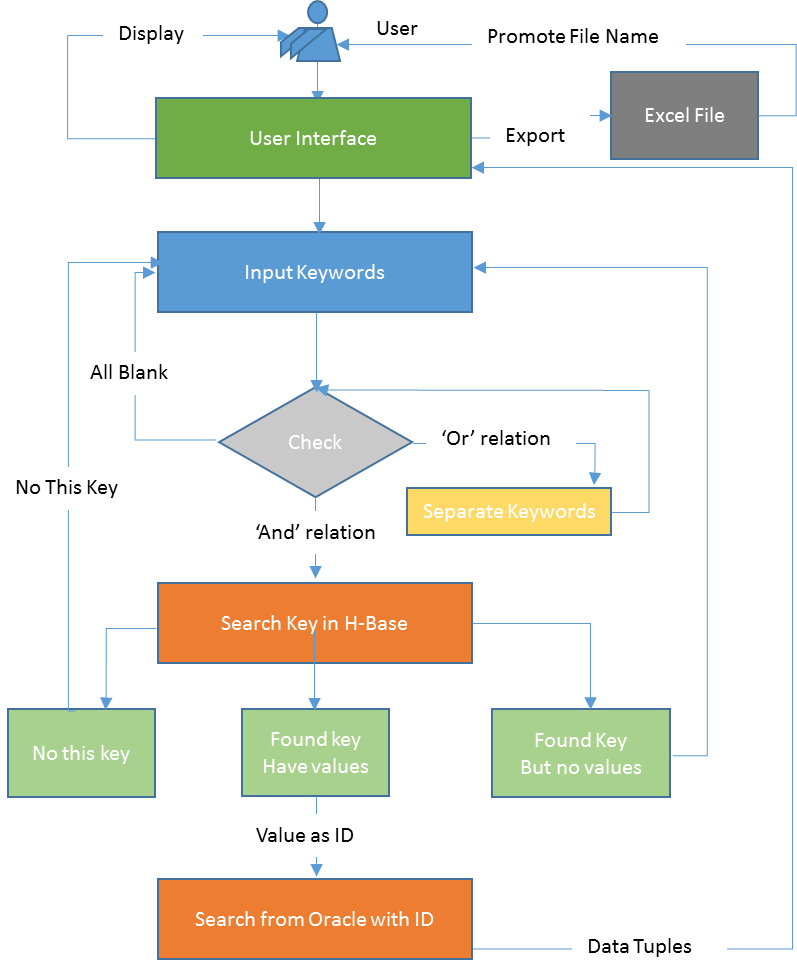
****

Figure 11: keywords search procedure chart of project

* 1. Version of KSE

This project we implemented three versions of KSE as the client requested some new feature of the searching.

* + 1. Version 1

The version 1 is only basic search function, input complex keywords with relation, the user interface display the result in table.

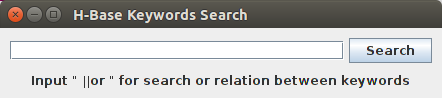


Figure 12: version 1 main frame

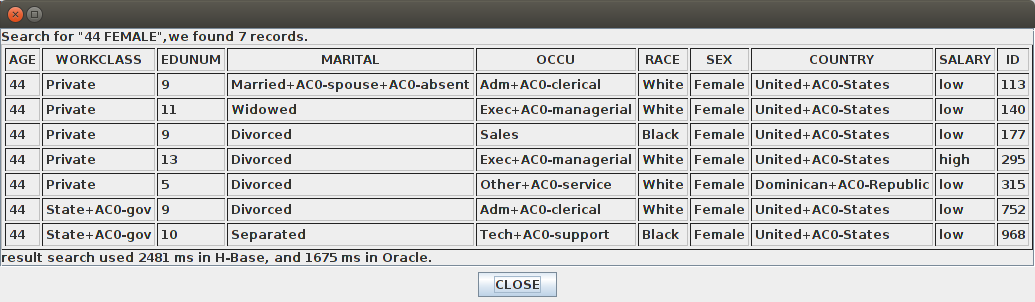


Figure 13: version 1 deliverable result frame

* + 1. Version 2

After the client requests some new feature, we implement the version2, the result is separately to several pages as some search should return huge amount tuples from oracle, it is hard to read and find useful information on one page, highlight the keywords in the result display will help client easy read and find the different between tuples.

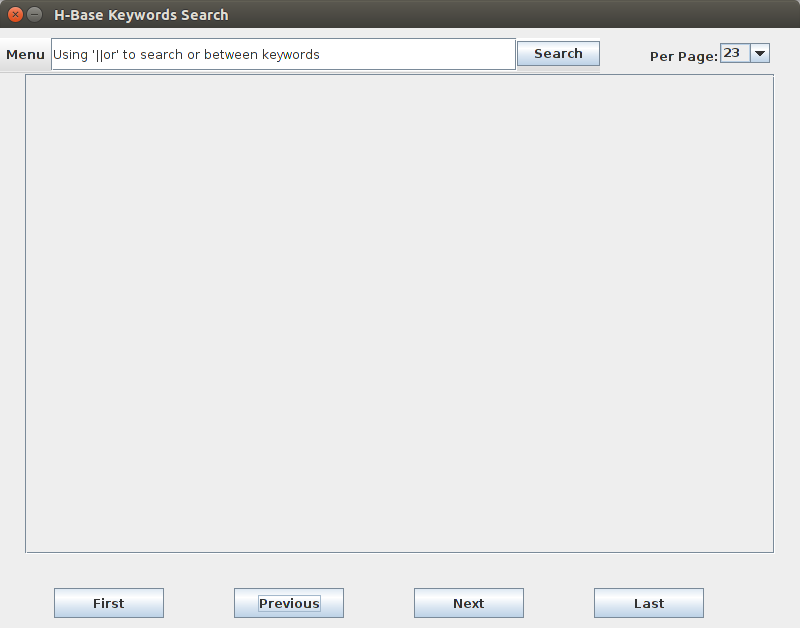


Figure 14: version 2 of deliverable main frame

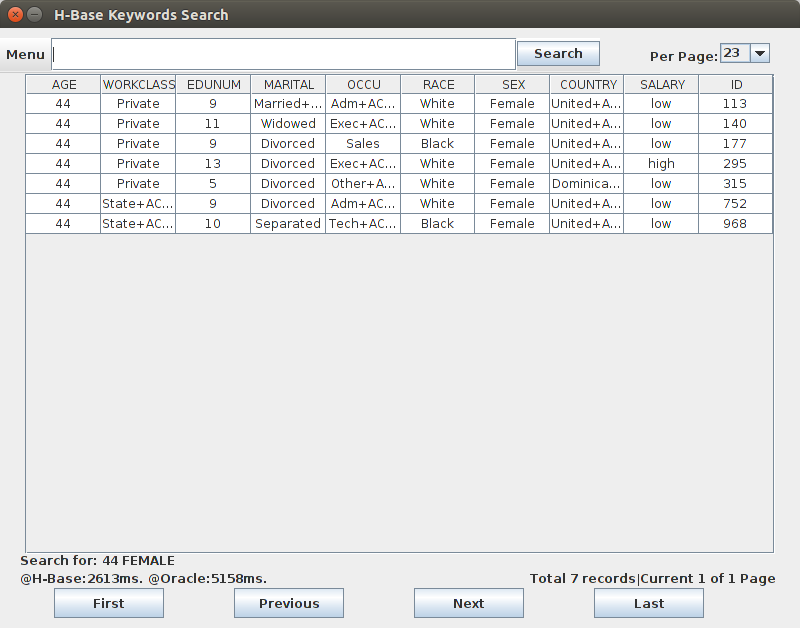


Figure 15: version 2 result frame

* + 1. Version 3

After demonstration the version 2 to client, the client wants to save some search result to a file, that means the client does not need to do search every time. Therefore, in the version 3, we add an export button in order to give the client a choice to export this result or not, the result will export into an Excel file, and the user interface would auto create the file name with only valid search keywords plus a current system time to unique each search result.

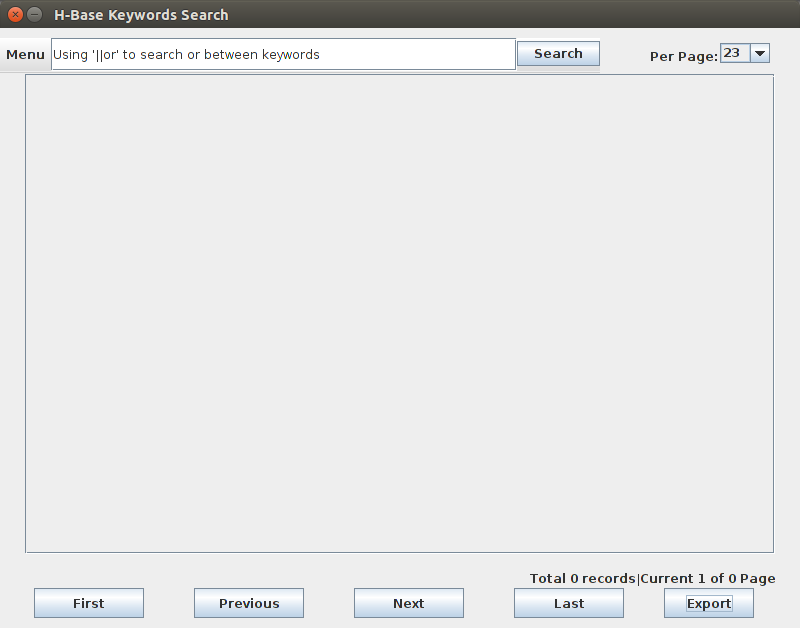
****

Figure 16: version 3 main frame

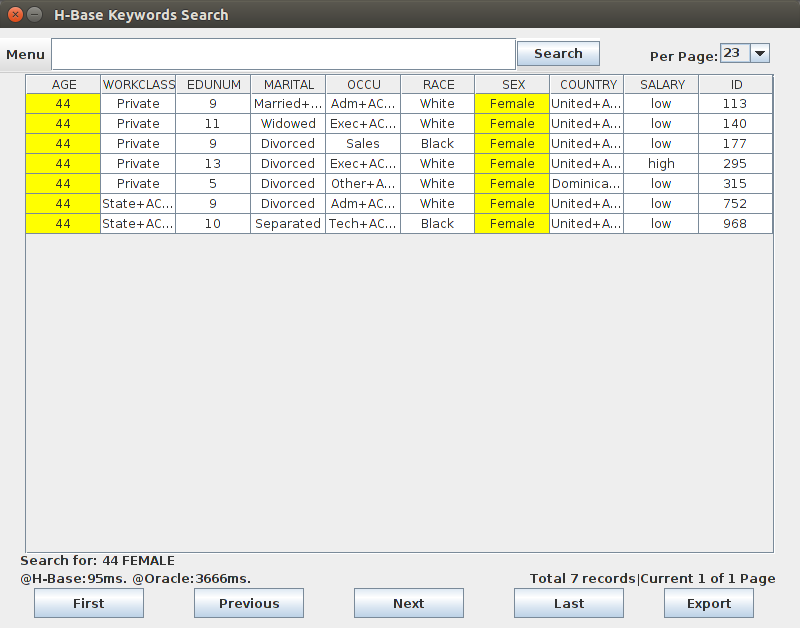
****

Figure 17: version 3 result frame

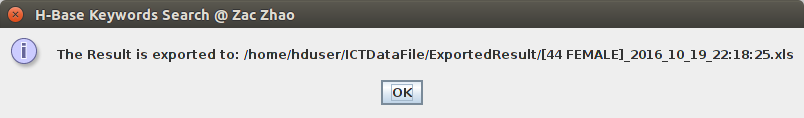


Figure 18: version 4 export promote frame

* 1. Testing

To make sure the quality of deliverable, we check and test several cases depends on the requests of client.

|  |  |  |
| --- | --- | --- |
| **Check or test name** | **No.** | **Status** |
| Check the H-Base working condition | 3.1 | pass |
| Check the H-Base add and delete key-value pair | 4.1 | pass |
| Check the input valid | 5.1 | pass |
| Check the single key search | 5.2 | Pass |
| Check multiple key with and relation search | 6.1 | pass |
| Check multiple key with or relation search | 6.2 | pass |
| Check multiple key with complex relation search | 8.1 | pass |
| Check retune result in table | 9.1 | pass |
| Check highlight keywords in result | 10.1 | pass |
| Check result in table and paged | 10.2 | pass |
| Check export result into file | 11.1 | pass |

Figure 19: check and test list of project

* 1. Use case story

For the deliverable we assume a user

|  |
| --- |
| Multiple keywords with mix relation search on relational database |
| Use case ID: 100300 |
| Actor: the oracle user |
| Starting condition: the user need to finish the configure file and initial H-Base before search |
| Flow of events:   * User fill the oracle and H-Base information into the configure file if have not do that before * User click the KSE executing file to start the KSE search engine * User click configure button of search engine to initial H-Base if have not do that before * User input keywords * Click search button or click “Enter” key of keyboard * Click ok button if there is any promote for key or result * View the result with navigation buttons * Click export button for export result into file if needed |
| End condition: the user Input more keywords to start another search or click exit button to exit the search engine. |

Figure 20: use case story of KSE

1. **Summary**

The KSE search engine is 100% satisfy all requests of my client.

* 1. Management

A successful project must have a good project plan in the first place that could help us to prevent bad surprises so that could help clients to save time costs. Consultation skill is the key in a project. Fully understand every details of the client requests and communicate them effectively that could improve the efficiency of a project. Last but not the least, keep simple and acceptable to our client when we prepare the deliverable as we assume clients do not have any technical / professional knowledge. At the end of this project, our deliverable is 100% match the requests of client, the client is very happy with the final version of deliverable.

Follow this project I have gained real project management experience at several fields,

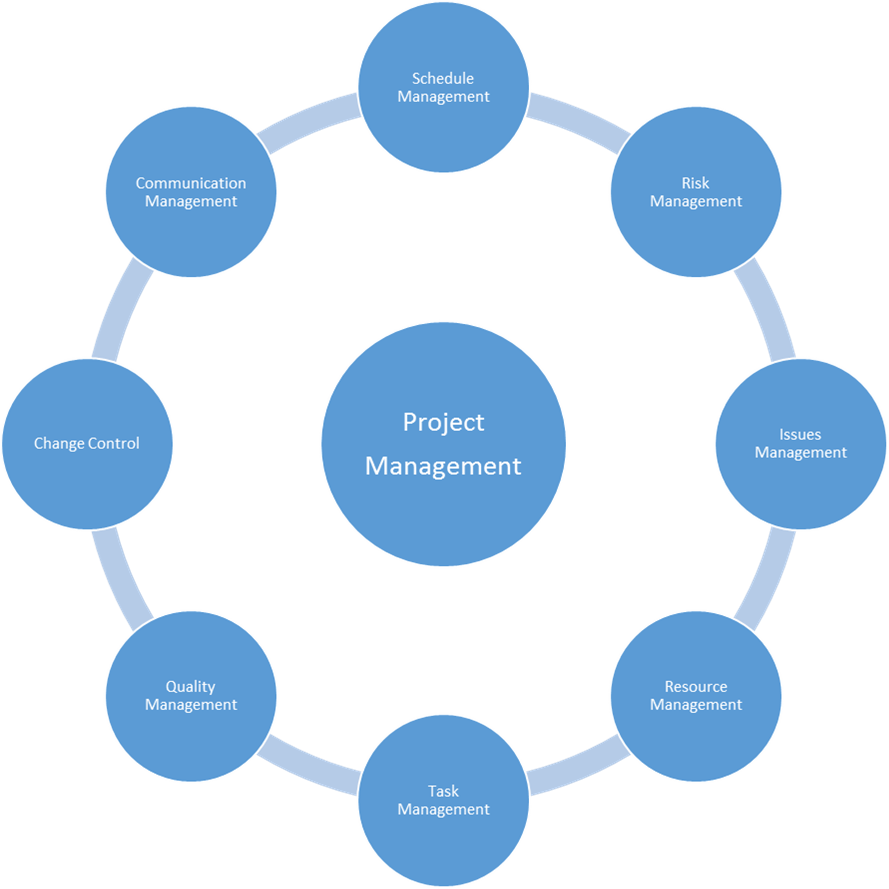
****

Figure 21: management of project

* 1. Milestone

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Milestone** | **WBS** | **Planned** | **Forecasted** | **Actual** | **Status** |
| Research the topic | 2.1 | Week 1-4 | Week 4 | Week 4 | completed |
| Computer environment building | 5.1 | week 3 | week 3 | week 3 | completed |
| Install and test software | 5.2 | Week 3 | Week 3 | Week 3 | Completed |
| Oracle configure | 6.1 | Week 6 | Week 6 | Week 6 | Completed |
| H-Table create and store | 6.2 | Week 6 | Week 6 | Week 6 | Completed |
| Get row number for keyword | 7.1 | Week 7 | Week 7 | Week 7 | Completed |
| Get original data from Oracle | 7.2 | Week 7 | Week 7 | Week 7 | Completed |
| Colored key | 10.1 | Week 10 | Week 10 | Week 10 | Completed |

Figure 22: milestone list of project

* 1. Deliverable

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Deliverable** | **WBS** | **Planned** | **Forecasted** | **Actual** | **Status** |
| Weekly meeting minutes and Report | 1.1 | Week 2 | Week 2 | Week 2 | Completed |
| Charter | 4.1 | Week 3 | Week 3 | Week 3 | Completed |
| Project presentation | 5.3 | Week 4 | Week 4 | Week 4 | Completed |
| Search single word in H-base and return original data from Oracle | 7.3 | Week 7 | Week 7 | Week 7 | Completed |
| Search multiple words | 8.1 | Week 8 | Week 8 | Week 8 | Completed |
| Oracle update trigger | 8.2 | Week 8 | Week 8 | Week 8 | Completed |
| Search or relation(version 1) | 9.1 | Week 9 | Week 9 | Week 9 | completed |
| Paged table(version 2) | 9.2 | Week 9 | Week 10 | Week 10 | Completed |
| Configuration file | 10.2 | Week 10 | Week 10 | Week 10 | Completed |
| Export result into excel(version 3) | 11.1 | Week 11 | Week 11 | Week 11 | Completed |

Figure 23: deliverable list of project

* 1. Risk

There are some risks when using the KSE search engine, such as

|  |  |  |
| --- | --- | --- |
| **Risk name** | **Risk No.** | **Current Status** |
| May have huge among rows key-value pair | 4.1 | Depends on the data set of client |
| Search efficient | 4.1 | Depends on the size of data set of client, and the keywords of the search input |

Figure 24: Risks of KSE

* 1. Discussion and future work

Perhaps a feature that could be implemented in the future is to have “input error correction” function on KSE. Such as user has some input error in the wrong sequence, for example “girl” spell to “gril”, or input error in incorrect/missing character of keyword, like the word “Canberra” spell to “Canbarra” or “Canbera”. The “input error correction” function can detecting the error and KSE will suggest the most similar keyword and result for the error keyword.

Thanks for my client and my project supervisor Dr. Jerry Liu for gave me such good project and a lot of good advices, since some requests of this project are hard to implement and that were some big challenges, but that also was a very good chance to improve my consultation skills programming skills and project management experiences. I am really enjoy this project at this moment.