

Due dates: March 19, April 9, and April 23, 2018

Project description

New Orleans was ranked #1 in population growth of traditional cities two years ago. On the other hand, the percentage of IT professionals within Louisiana's workforce is only 1.05%, which is 36% lower than the national average, and 93%, 121%, 195% lower than Alabama, Florida and Texas respectively. Louisiana deserves to claim more professional jobs by creating and enhancing programs that train the workforce for coding jobs, professional IT jobs. This is not only because coding jobs represent a large and growing part of the job market (there were nearly 7 million job openings in the U.S. last year for roles requiring coding skills), this is also because the jobs requiring coding skills are projected to grow 12% faster than the job market overall in the next 10 years; IT jobs are expected to grow 25% faster than the overall market¹. Jobs valuing coding skills have average salaries at or above the average for all jobs. The wages associated with each role type correspond with the intensity of programming skill requirements. Nationwide, the highest earning roles are in IT, with an average advertised salary of \$90,000².

While creating one's own business is a solution, finding a job is a more common way to start a college graduate's careers. To be qualified for a job, one often has to retool him/herself by learning and training. Realizing Louisiana had 45909 IT (see spreadsheet Year2016-LA-ITcluster), 4640 database (Year2016-LA-DatabaseSkill) and 1579 Java job postings (Year2016-LA-JavaSkill) in 2016, you should understand that the job market is on your side. (Sorry, I have only these old data.) Rather than jumping into any job position, you should be picky and strategic by having a career plan.

In this project, your task is to build an information system that can help optimize training programs, job assistance services, career planning, and company recruitment. The central efforts for this task is to design a database that supports these services. If your design is good and if Louisiana wants to keep the crown of workforce training in nation, the State should utilize your program. By feeding real-world data into your project, its outcome will also be your own job haunting tool.

Entities and Relationships

Workers are classified as staff (salary worker), wage-workers, and volunteers. Companies and organizations hold job positions. Workers acquire knowledge and skills through two types of learning activities, academic courses and commercial trainings. The elements of these two types of activities are courses. The mission of learning is to fill up knowledge/skill gaps between the workers' possession and the job position's requirements. In industry, practitioners refer this kind of learning activities as *training*, so do we. Generally, academic courses help students learn knowledge; training courses help trainees gain skills. Each course covers a set of knowledge/skills. For example, CSCI 4125 covers: data modeling (medium level), relational models (medium level), SQL (advanced level), JDBC programming (beginner level), RDBMS design (medium level), and database concurrency control (advanced level).

A person can work on one or more job positions. A job position is a slot that pays the worker. Every good-pay job position requires many skills and sometimes a certificate. Jobs can be classified into categories. Job categories form hierarchies. There are standards of job category descriptions such as those provided by the Department of Labor at the federal level or in many states.

Some important entities along with their important attributes are listed below. The attributes surrounded by a pair of braces ({ }) may have multiple values, the attribute that carries a plus (+) is a composite field which you should flatten in relational modeling. Note, the purpose of the following information is to describe the systems, **not** to give you an actual schema of the data model. For example, "company" in "Position(pos_code, type, pay_rate, pay_type, company)" means the identifier of a company, rather than the company object. A focus of your design is the database schema configuration. In your database, you should consider the nature of the data and the queries to process. You will need additional relationships such as **Has_Skill** and **Works** as well as entities defined by yourselves.

¹ Projections data are from the Bureau of Labor Statistics' Employment Projections program. <http://www.bls.gov/emp/>

² "BEYOND POINT AND CLICK", a report by Burning Glass Technologies Corp., June 2016, www.burning-glass.com

Person(per_id, name, address⁺, zip_code, email, gender, {phone})

where per_id is a the primary key.

Job_category(cate_code, title, description, pay_range_high, pay_range_low, {core_skill}, parent_cate)

where cate_code is a unique identifier, and core_skill refers to one or multiple skills represented by k_code. Parent_cate is a cate_code representing a higher level category that the current one belong to.

Position(pos_code, emp_mode, {required_skill}, pay_rate, pay_type, cate_code, company, *more ...*)

where pos_code is a unique identifier, emp_mode can be “full-time”, “part-time”, and pay_type can be wage or salary, pay_rate is the hourly rate for wage or annual pay for salary. When the position belongs to multiple cate_codes in a hierarchy, you should choose the most specific one, the one that is closest to a leaf node.

Company(comp_id, address⁺, zip_code, primary_sector, {speciality}, website)

Where primary_sector refers to the business sector the company belongs to, such as tourism, oil and mining, entertainment, service, education, technology. Specialities can be construction, logistics service, computer hardware, software, electronics, basic materials, IT consulting, transportation.

Course (c_code, title, level, description, status, retail_price)

where c_code is a unique identifier, and status is either active or expired.

Secion(c_code, sec_no, complete_date, year, offered_by, format, price)

Where the triple (c_code, sec_no, year) form a unique identifier for every section, format can be (classroom, online-sync, online-selfpaced, or correspondence), offered_by refers to a university or a training company.

Knowledge_Skill(ks_code, title, description, level)

where ks_code is a unique identifier (you may use the skill code defined by the Department of Labor), and level can be “beginner”, “medium” or “advanced”.

In addition, many relationships are needed. A few of them are listed below.

Works(per_id, pos_code, start_date, end_date) where start_date ≤ end_date, and end_date < today indicates per_id no longer holds the job position.

Requires(pos_code, ks_code, prefer) where prefer indicates the necessity of the skill if prefer hold null.

Teaches(c_code, ks_code)

Prerequisite(c_code, required_code)

Has_Skill(per_id, ks_code)

Takes(per_id, c_code, sec_no, complete_date)

In a person’s career, working on multiple job positions is common. Sometimes, a person’s working history can make him/her invaluable at certain situation. Recruiters often seek applicants’ experience. For this reason, the database should not only record each person’s currently job position(s), but also track every worker’s job history.

The database will support at least three services, the *career planning service*, the *job hunting assistance*, and the *training service*. The *career planning service* needs information about workers’ education (transcripts), training and skills in order to help find workers’ suitable jobs. The *training service* can recommend the courses for a person who pursues a job position by recognizing the missing skills of the person for the job position. The features of *job hunting assistance* are up to your experience and imagination.

Data Preparation

In workforce management and career planning, strategically we need to know the category each job position belongs to and the structure of the job category hierarchy. This structure, the latest Standard Occupational Classification (SOC) system, is given by U.S. Bureau of Labor Statistics in 2009. See soc_structure_2010.xls and <https://www.bls.gov/SOC/>. You should focus on the 21 codes in the range of 15-1100 to 15-1199. Among the 1421 codes of major, minor, broad and detailed groups, let us start from the computer occupations.

We also need to know what kind of skills the IT job market is calling for. An initial set is in skill270.xlsx which is based on the software skills enumerated in the three spreadsheets Year-2016-LA-xxxx.xls.

Understanding the nature and roles of the skills is important in your learning, which can be found out in the IT skill standards developed by the [National Workforce Center for Emerging Technologies](#) (NWCET). The 2003 edition of the skills standards is in WCETSkillStandards03.pdf

(<http://www.tssb.org/sites/default/files/wwwpages/repos/pdf/NWCETSkillStandards03.pdf>).

This lengthy document categorizes skills into eight career clusters as shown in the list below.

Career Cluster code (cc_code)	Title
DDA	Database Development and Administration
DM	Digital Media
ESAI	Enterprise Systems Analysis and Integration
NDA	Network Design and Administration
PSE	Programming/Software Engineering
TS	Technical Support
TW	Technical Writing
WDA	Web Development and Administration

You should focus on the tables on DDA (p 18), DM (p42), ESAI (p 64), NDA (p 78), PSE (p 99) and WDA (p 154). In each table, you can limit your collection to the top row. For example, from the table on page 18, you can collect the six rows. The 7th and 8th rows are from the table on page 42. *[Once you complete the main part of your project, you may write a parser to collect complete data from these six tables as well as those “critical work function” tables for bonus points.]*

cc_code	Function_code	Work function
DAD	A	Analysis and design database
DAD	B	Develop and implement database
DAD	C	Perform administration and maintenance
DAD	D	Provide data assurance
DAD	E	Provide client and user services
DAD	F	Perform database test
DM	A	Perform analysis
DM	B	Produce visual and functional design
	...	

The companies hosting the jobs must matter the most in a worker’s career. The North American Industry Classification System (NAICS) is the coding used by U.S. Census Bureau. The NAICS code is a hierarchical structure in 2017_NAICS_index_File.xlsx. This is also available at

https://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart_code=54&search=2017%20NAICS%20Search .

The first two digits represent the top-level categories such as 11 refers to agriculture, 21 mining, and 54 professional, scientific and technical services. The NAICS codes related to computer and software are 511210, 518210, 541511-541519, 541715 and 611420.

Having the above three standards, you are ready to fit your project into the real world. First, you need to load soc_structure_2010.xlsx, 2017_NAICS_index_File.xlsx and skill270.xlsx into your database. Then describe some of the skills skill270.xlsx according to NWCETSkillStandards03.pdf.

So far, we have considered the stable data of your database. The actual job information is dynamic. You can start using the data in files indeed100.txt and WorkNOLA36.txt. In the initial testing of your SQL statements, use small data sets so that you can easily know what should be the correct result. 5 people, 10 job positions, 10 courses and 30 skills sound sufficient. Make sure you have job positions requiring multiple skills, courses teaching multiple different skills, persons possess multiple various skills. Some people are qualified for some job positions; other people do not.

As you can see the data given in the textual files are about a year old. A Java package is provided in the project section of our class Moodle site. You are strongly encouraged to use it to get daily/weekly feed of data in the local job market from Web site such as:

- NOLA.com <http://jobs.nola.com/Jobs/technology>. You can browse by category
- Monster <https://www.monster.com/jobs/q-it-jobs-l-new-orleans,-la.aspx>
- indeed <https://www.indeed.com/l-New-Orleans,-LA-jobs.html>
- and WorkNOLA.com <https://worknola.com/job-search>
<https://worknola.com/365-connect/web-developer-back-end-net>
<https://worknola.com/firstline-schools/2016-2017-systems-development-manager>

You can visit these Web sites and get an idea about the real job positions in the market. For example, on the left side of the Web page of WorkNOLA, you see panels of Job Function, Industry, Job Type, Employer, Required skills, and Preferred skills. indeed.com additionally shows salary estimate, location, experience level. The job ads in some site such as NOLA.com appear in a very loosely formatted manner. As a consequence, the same attribute may be referred in different terminologies. For instance, “Required skills” is sometimes called “Qualifications” or “Our Expectation”, “Qualifications & characteristics” means “Ideal Candidates”. You have your freedom to include or drop out the less essential attributes such as “Qualifications & characteristics”.

By referring the associated readme file, you most likely will be able to adapt the given code and fetch data from the above Web sites. Using live data can earn you points for your project.

Queries

The most important component of this project is the following queries that are given to illustrate the capabilities of your system. As we understand the job market Web sites better and agree on the features of your job hunting assistance.

1. List a specific company’s workers by names.
2. List a specific company’s staff by salary in descending order.
3. List companies’ labor cost (total salaries and wage rates by 1920 hours) in descending order.
4. Given a person’s identifier, find all the job positions this person is currently holding and worked in the past.
5. Given a person’s identifier, list this person’s knowledge/skills in a readable format.
6. Given a person’s identifier, list the skill gap between the requirements of this worker’ job position(s) and his/her skills.
7. List the required knowledge/skills of a pos_code and a job category code in a readable format. (Two queries)
8. Given a person’s identifier, list a person’s missing knowledge/skills for a specific pos_code in a readable format.
9. Given a person’s identifier and a pos_code, list the courses (course id and title) that each alone teaches all the missing knowledge/skills for this person to pursue the specific job position.
10. Suppose the skill gap of a worker and the requirement of a desired job position can be covered by one course. Find the “quickest” solution for this worker. Show the course, section information and the completion date.
11. Suppose the skill gap of a worker and the requirement of a desired job position can be covered by one course. Find the cheapest course to make up one’s skill gap by showing the course to take and the cost (of the section price).

12. If query #9 returns nothing, then find the **course sets** that their combination covers all the missing knowledge/skills for a person to pursue a pos_code. The considered course sets will not include more than three courses. If multiple course sets are found, list the course sets (with their course IDs) in the order of the ascending order of the course sets' total costs.
13. Given a person's identifier, list all the **job categories** that a person is qualified for.
14. Given a person's identifier, find the job position with the highest pay rate for this person according to his/her skill possession.
15. Given a position code, list all the names along with the emails of the persons who are qualified for this position.
16. When a company cannot find any qualified person for a job position, a secondary solution is to find a person who is almost qualified to the job position. Make a "missing-one" list that lists people who miss only one skill for a specified pos_code.
17. List each of the skill code and the number of people who misses the skill and are in the missing-one list for a given position code in the ascending order of the people counts.
18. Suppose there is a new position that has nobody qualified. List the persons who miss the least number of skills that are required by this pos_code and report the "least number".
19. For a specified position code and a given small number k , make a "missing- k " list that lists the people's IDs and the number of missing skills for the people who miss only up to k skills in the ascending order of missing skills.
20. Given a position code and its corresponding missing- k list specified in Question 19. Find every skill that is needed by at least one person in the given missing- k list. List each skill code and the number of people who need it in the descending order of the people counts.
21. In a local or national crisis, we need to find all the people who once held a job position of the special job category identifier. List per_id, name, job position title and the years the person worked (starting year and ending year).
22. Find all the unemployed people who once held a job position of the given pos_code.
23. Find out the biggest employer in terms of number of employees and the total amount of salaries and wages paid to employees. (Two queries)
24. Find out the job distribution among business sectors; find out the biggest sector in terms of number of employees and the total amount of salaries and wages paid to employees. (Two queries)
25. Find out (1) the number of the people whose earnings increased, (2) the number of those whose earnings decreased, (3) the ratio of (# of earning increased : # of earning decreased), (4) the average earning changing rate of for the workers in a specific business sector (use attribute "primary sector" in table Company. [Hint: earning change = the sum of a person's current income – the sum of the person's earning when he/she was holding his/her the latest previous job position. For (4), only count the earning from the specified sector (companies' "primary sector")])
26. Find the leaf-node job categories that have the most openings **due to lack of qualified workers**. If there are many opening positions of a job category but at the same time there are many qualified jobless people. Then training cannot help fill up this type of job position. What we want to find is such a job category that has the largest difference between vacancies (the unfilled job positions of this category) and the number of jobless people who are qualified for the job positions of this category.
27. Find the courses that can help most jobless people find a job position by training them toward the jobs of this category that have the most openings due to lack of qualified workers.
28. List all the courses, directly or indirectly required, that a person has to take in order to be qualified for a job position of the given category, according to his/her skills possessed and courses taken. (required for graduate students only)

Requirements of SQL statements

- Every keyword must be in all-upper-case; table names and attributes had better not be in all-upper-case.
- The SELECT, FROM, WHERE, GROUP BY and ORDER BY clauses of each SQL statement must be lined up in separate lines. Nested SELECT statements must be indented properly. No indentation, no grade!

Development Tasks

- (1) Make an E-R diagram showing your data model. In the diagram, show the important attributes of each entity. For every relationship, show the cardinality and the participation status, as well as any important attributes associated with the relationship. Be sure to indicate the primary key of each strong entity. Note, this is a data-model E-R diagram,

not a detailed relation schema diagram for implementation. I will accept hand-draw diagrams. Feel free to discuss your E-R diagram in class or during my office hours.

- (2) Reduce the data-model E-R diagram to the database relations schema. Draw a relation-schema diagram for your implementation.
- (3) Enumerate the concerned functional dependencies (to be learned in Chapter 8) in your system.
- (4) Revise the database schema design by producing a lossless-join 3NF schema that preserves functional dependencies.
- (5) Populate the tables with adequate data such that all the questions can be demonstrated. Be sure to write your INSERT statements in SQL scripts. You are also required to write a SQL script that cleans up everything in the database; this will be a good exercise to realize the dependencies enforced by foreign key references. In your development process, you must re-populate your database after cleaning up it for multiple times.
- (6) Write the SQL statements that carry out the 27 queries listed above, or 28 for the graduate students.
- (7) Design and implement *some* of the Java classes such as Course, Job Category, Job, and Person with the JDBC technique. These Java classes should support creation (such as creating a course, a job category, or a job position) and deletion (such as removing a job position or setting a course *inactive*.) *Note, do not spend too much time on too many "classes".*

Write a user interface in Java to drive each of these queries with JDBC. A graphical interface (GUI) will be appreciated but not necessary.

- (8) Design and implement database application that manages *business processes*. Write at least three programs to support business *processes* listed as the following:

a) A company accepts a new employee;

You can assume the person is already in the system. So the process starts with a per_id and a pos_code. Then the application should automate the steps shown below by interacting with the user.

- i. Upload the person's transcripts and input the course taking information into table Takes;
- ii. Populate table Has_Skill with more rows derived from the courses this person have taken;
- iii. Verify if this person has every skill required by the given pos_code;
- iv. If a skill gap is identified, propose a training plan for this person.

b) A person's job hunting; This could be just a user interface for Queries #13 and #14.

c) A company finds the best choice of people for a job position with training; This could be just a user interface for Queries #15 and #18.

Team organization

You are required to form a team made of two people. Each team member should work on one or more different services. Every team member should perform every aspect in the development process of database applications such as E-R modeling, SQL, and Java. Do not divide your tasks horizontally such as one writes SQL only, the other writes Java only. Such kind of task division will give the Java guy serious disadvantages in earning a fair grade.

Every group must sign up for an hour-long project inspection with me in the week of **November 27**. At least two test cases should be prepared and documented for queries 8 through 28, as well as for services (a) through (c) specified in Development Task (8). Each test case must include the prepared input data/setting and a specification of the expected output/results. Early submission and demonstration of project earn bonus points.

Due Dates

Due date of Phase 1 (15%) – March 19: The deliverables of task (1), hand in a hardcopy of your E-R diagram in the class on Tuesday, March 20.

Due date of Phase 2 (35%) – April 9: The deliverables of tasks (2), (5) and (6). You are required to hand in a hardcopy of SQL statements for the 27 (or 28) queries in class on Tuesday, April 10. #1 through #24 will be graded. SQL statements for #25 and beyond can earn bonus.

Due date of Phase 3 (50%) – April 23: The project report including everything such as the revised E-R diagram, schema diagram of tasks (1) and (2), as well as revised SQL statements of task (6).

Graduate Students (CSCI 5125) Are Required to Complete Additional Work [Undergraduate students can earn extra points by working out the following requirements.]

(1) Some job position require one or more certificates. One can obtain a certificate by passing a set of exams provided by some companies. Most certificates can expire. Many certificates are bound to a tool. A set of certificates are potential for CSCI 4125 students including Oracle Database Developer I and Oracle Database Developer II. These two certificates bound to tool Oracle 11g. To obtain a certificate, one has to pass a number of courses. The courses are among the courses we considered previously. They could be offered either by colleges or training companies. A conceptual entity of certificate is given below.

Certificate(cer_code, title, description, expire_date, issued_by, t_code)

where cer_code is a unique identifier, t_code shows the bounded tools of this certificate if any, and issued_by refers to a company.

Graduate students' implementation of the project should manage certificates. The queries regarding jobs and skills should handle considerations of certifications.

(2) Implementation of Query 28 is mandatory for graduate students, optional for undergraduate students for bonus points.

(3) Design and implement experiments on concurrency control using this database system. Each experiment can be demonstrated in SQLPLUS or in JDBC implementations. Every experiment must be documented with problem specification, applied solution, and presented with three test cases (with input data/setting and the expected output/results). Every graduate student must have at least four different experiments prepared and documented.