UALR Master of Science in Information Quality
Capstone Project:
Requirements and Guidelines

Students pursing the Master of Science in Information Quality may choose to complete the 33-hour program by either a 6-hour master's thesis or 6-hour capstone project. This document outlines the requirements and guidelines for students selecting the **capstone project** option.

Table of Contents

The Capstone Project Process

The MSIQ Capstone Project process comprises the following steps:

1. Meet the minimum qualifications for starting a project
2. Identify a faculty advisor
3. Identify a sponsoring organization and supervisor
4. Identify an appropriate information quality project within the sponsoring organization
5. Submit a project proposal describing the project
6. Upon approval of the proposal, enroll in project hours
7. Complete the project
8. Upon completion of the project, submit a final project report to the MSIQ committee
9. Upon approval of the final report, make an oral defense of the project to the MSIQ committee and sponsor supervisor

Minimum Qualifications

Courses Completed

To be eligible to begin the capstone project you must have completed at least 9 hours of the MSIQ program. These 9 hours must include:

INFQ 7303	Principles of Information Quality
INFQ 7342	Information Quality Tools
Either INFQ 7367 or INFQ 7322	Information Quality Policy and Strategy Information Quality Theory

In Good Standing

You must also have regular admission status and otherwise be in good standing in the MSIQ program before starting a project.

Faculty Advisor

Every capstone project must have a faculty advisor. A project faculty advisor must be a member of the MSIQ faculty and must agree to serve in this role.

Sponsoring Organization and Supervisor Guidelines

Qualifying Organizations

The next thing that you must do is to find an organization willing to sponsor your project. The sponsoring organization must be external to the UALR Department of Information Science. The organization must agree to sponsor the project. The sponsoring organization may be another UALR academic or administrative department, a business, a government agency, or a non-profit organization.

Employers can be Sponsors

If you are employed either full-time or as an intern, your employer can qualify as a sponsor as long as the project proposal meets the required guidelines.

Sponsor Supervisor

In addition to identifying a sponsor, you must also identify a sponsor supervisor within the organization. A sponsor supervisor is an employee of the sponsoring organization who will agree to supervise your project and provide input to your advisor on your performance.

Qualifying Projects

Project Characteristics

Information Quality is a service science. Accordingly, by fulfilling your MSIQ requirement using the capstone project option you must address and improve some aspect of information quality. This includes producing a final report along with other deliverables appropriate to the project. Project sponsors are expected to help the student identify an information quality problem within the sponsor's organization and work with the student's faculty advisor to develop a plan for addressing the problem.

Must Involve Data and Information

The primary goal of any project should be for you to have an impact on information quality using the techniques that you learned in your coursework. You should be able to identify an information quality problem, then evaluate, measure, and improve the quality. While technology plays a primary role in information processing, the core focus should be on data and information. Technology improvement may be an important aspect. For example, a project to migrate information stored in a spreadsheet to a relational database in order to improve the dimensions of manipulability and accessibility of the underlying information. However, every project must at its core be related to data and information, and the data and information quality being improved must be clearly identified in the proposal.

Must Address Information Quality Dimensions

Your project must address and improve at least 3 dimensions of information quality. At least two of these dimensions must have quantitative measures or metrics defined as mathematical formulas. In addition, you should measure the level of quality in all of the dimensions you have identified at the start of the project and end of the project.

Projects are Unpaid

Unless you are an employee or intern of the sponsoring organization, the sponsor is not expected to pay for work related to the project.

Project Work Hours

A viable project should comprise 200-300 working hours. If you plan to complete the project in one 15-week semester should expect to spend approximately 15-20 hours per week actively working on the project. Accordingly, a two semester project would necessitate 7-8 weekly working hours to complete.

Two-Semester Projects versus One-Semester Projects

You are strongly encouraged to complete your projects over two consecutive semesters. It is very difficult to complete an entire project process of appropriate scale within the beginning and end of one semester. You are encouraged to begin your project in one semester and complete the project including the final report and oral presentation in the following semester. For administrative purposes, the final report must be submitted at least 2 to 3 weeks in advance of the oral presentation which must take place before the end of the semester. This leave only 10 to 11 weeks to identify and complete the actual project work and any unanticipated delays. If the project process is not complete by the last day of class in a semester, then you will be required to re-enroll in INFQ

7386 in each following semester until the project is complete. These additional hours are required for incomplete projects even if you have already completed 6 hours of qualifying project hours in previous semesters.

Project Proposal Guidelines

Your project proposal must be approved by the MSIQ Faculty Committee before you can enroll in either the INFQ 7386 or INFQ 7686 project course.

Formatting Guidelines

Please follow these formatting guidelines

- Letter page size in portrait orientation
- 1" top, bottom, and side margins
- Single column, with single or 1.25 line space
- 11 or 12 pt. font
- Use only black or dark blue font color
- Left justify paragraphs except for cover page
- Signal a new paragraph by either indenting the first line of a new paragraph or spacing between paragraphs
- Number all pages in the body of the proposal
- Figures and tables are encouraged, but all figures and tables should be numbered and captioned.
- Captions for figures and tables should fit on one line.
- All numbered and captioned tables and figures must be referenced by their number at least one time in the body of the proposal.
- All tables and figures must fit on one page. If an important table or figure is longer than one page should be included as an appendix to the proposal.
- Follow consistent indention of headings and following text throughout the proposal
- Do not split headings and following text between pages

Writing Style

- Use short sentences in direct, first-person language, e.g. "In this project I plan to design and …"
- Consider rewriting sentences longer than 2 lines.
- Avoid paragraphs of only one sentence
- Do not try to control page breaks by spacing with paragraphs. You should use a page break command
- Spell out "Figure", not "Fig"
- When citing a specific figure or table you should always capitalize e.g. instead of "figure 1 shows that …" write "Figure 1 shows that …"
- Avoid the use of the passive voice. Try to use active first person, active voice.
- Always give the complete spelling of an acronym the first time you use it, e.g. "Master data management (MDM) is an important component of …"
- Have someone else proofread your proposal for spelling, grammar, and punctuation. If you are not a native English speaker, you should have a native English speaker do the proofreading.

Proposal Sections

Cover Page

Your proposal must start with a cover page. The following items should appear on the cover page in centered paragraphs

- Start with the words "Project Proposal for the Master of Science in Information Quality Program"
- Project Title. The title should be descriptive of the project and must include the word "Quality" and the word "Information" or "Data"
- Your Name
- Proposal Submission Date
- Expected Project Completion Date
- Your Faculty Advisor's Name
- Name of your Sponsoring Organization
- Name, title and contact information for your Sponsor Supervisor

Project Description

Include a high-level (1- to 2-paragraph) description of the sponsoring organization followed by a

- **Problem Statement** that describes the information quality issues and the impact that they are having on the organization. Include any previous or current efforts by others in the organization to address these issues.
- **Objective Statement** that describes your goals for the project.
- **Proposed Solution** to the problem.

Impact on Information Quality

- Identify at least 3 dimensions of information quality to be improved in the project.
- Indicate how information quality will be improved, along with a metric for how the level of quality will be measured in each dimension (both at the start and end of the project). While some assessments can be anecdotal or subjective in nature, at least two of the measurements be a quantitative calculation based on a mathematic formula, .e.g. "For each file of N records, if M is the number of missing values for item X, then
  Completeness of Item X = 100 * (1 – M/N)".
- **WARNING**: In your metrics: do not claim to measure "accuracy" unless you are verifying the correctness of the data from the original source or another independent source know to be accurate. Don't confuse validation with verification. If you are cleaning the data to make all of the dates in the same format (xx/xx/xxxx), that is validating the dates. If these are dates-of-birth, then you are not measuring the accuracy unless you or someone else is going back to the original source (the person) to verify that is the correct date-of-birth or using some other independent source to verify that the date (such as birth certificate). Be careful with the use of the word "accuracy" in the proposal and in the final report.

Approach and Methodology

Describe the major activities you plan to carry out in solving the problem. If possible, you should show how these steps follow one of the IQ methodologies such as the MIT Total Data Quality Management (TDQM) methodology, Dr. Sebastian-Coleman's Data Quality Assessment Framework (DQAF), or McGilvray's Ten Step Method, or the Six-Sigma DMAIC.

Deliverables

Describe the major deliverables of the project including the artifacts you will leave with the sponsor, e.g. software, new procedure or process. Don't forget to include "Final Project Report" and "Slides for Project Defense".

Technology
To the extent you know at the time of the proposal, list the software, tools, systems, or other technology you expect to employ in the project.

Dependencies and Limitations
Describe any known issues that may delay or prevent a successful completion of the project. These might include delays in getting software, approval for system access, or dependencies other people in the sponsoring organization completing certain tasks critical to your project.

Ethical/Privacy Concerns
Describe any ethical/privacy concerns, challenges or limitations you may encounter in the project. Will you be working with sensitive information like personally identifiable information (PII) or health information? If so, how will it be secured and protected? Will you be required to sign any non-disclosure or security agreements before you can begin work?

Project Timeline
Provide a table project milestones where each row indicates a task and a date when you expect to finish the task. You do not need to include every task, only the major parts, but enough to follow your progress over the complete project. Project proposals should list about 6 to 10 milestones.

You may find it helpful to start at the end of the intended graduating semester and work backwards. All project work including the final defense must be completed by the last day of class in the last semester of the project. Include at least 2 weeks for compiling your findings and producing the Final Report and 2-3 weeks for revisions and approvals prior to your oral presentation.

A Proposal example is included as Appendix A of this document.

Enrolling in Project Hours

Once a student has an approved project proposal on file, he or she can enroll in project hours. An MSIQ capstone project requires 6 credit hours to complete. Students can successfully complete the 6-hour project requirements by choosing one of three different paths:

A. Two (2) 3-hour courses — **INFQ 7386**, Graduate Project, in two successive semesters.

B. One (1) 6-hour course — **INFQ 7686**, Graduate Project, in a single semester.

C. One (1) 3-hour course — **INFQ 7386**, Graduate Project, and
One (1) 3-hour course — **INFQ 7391**, Cooperative Education in Information Quality.

Note: In Option C ideally the work done during the Cooperative Education course is for your project sponsor and directly related to your project. However, this is not required. INFQ 7391 will satisfy 3 hours of the project requirement regardless of the employer and the duties performed or work completed during the internship.

It is very important for you to understand that the MSIQ Capstone Project is governed by a consecutive enrollment policy. This means that once you begin taking project hours, you must continue to enroll in project hours each semester (including the summer semester) until the project process is complete.

Project Execution

You should give your faculty advisor regular updates on the progress of the project. It is important for you to alert your advisor when problems or issues arise that could delay or prevent the completion of the project. Not all projects are expected to follow the project plan exactly as outlined in the project proposal. A change in direction or delay is not necessarily bad as long as there is a justifiable reason. On rare occasions projects are terminated before completion. In these cases, your faculty advisor and graduate coordinator will work with you to restart with a new project.

Final Project Report Guidelines

The final report is one of the most important components of the capstone project process. It should accurately and completely describe what the student accomplished in the project.

Formatting Guidelines

These are the same as for the project proposal

- The total number of pages in the final report including appendices must not exceed 50 pages. Longer is not necessarily better. Your goal should be a concise, yet complete report that tells the story of your project.
- Letter page size in portrait orientation
- 1" top, bottom, and side margins
- Single column, with single line space
- 11 or 12 pt. font
- Use only black or dark blue font color
- Left justify paragraphs except for cover page
- Signal a new paragraph by either indenting the first line of a new paragraph or spacing between paragraphs.
- Number all pages in the body of the report.
- All figures and tables are encouraged, but all figures and tables should be numbered and captioned.
- Captions for figures and tables should fit on one line.
- All numbered and captioned tables and figures must be referenced by their number at least one time in the body of the report.
- All tables and figures must fit and be placed on one page. If an important table or figure is longer than one page should be included as an appendix to the report.
- Follow consistent indention of headings and following text throughout the report.
- Do not split headings and following text between pages.

Writing Style

- Use first-person, past tense language. e.g. "In this project I designed and implemented …" Most of the final report should be written in the past tense describing activities you have already completed. Be careful about simply copying text from your proposal into the final report.
- Write in short, direct sentences. Consider rewriting sentences longer than 2 lines. Avoid paragraphs of only one sentence.
- Do not try to control page breaks by spacing with paragraphs. You should use a page break command.
- Spell out "Figure", not "Fig"
- When citing a specific figure or table you should always capitalize e.g. instead of "figure 1 shows that …" write "Figure 1 shows that …"
- Avoid the use of the passive voice. Try to use active first person voice.
- Always give the complete spelling of an acronym the first time you use it, e.g. "Master data management (MDM) is an important component of …"
- Have someone else proofread your report for spelling, grammar, and punctuation. If you are not a native English speaker, you should have a native English speaker do the proofreading.

Report Sections

Cover Page

The final report must start with a cover page. The following items should appear on the cover page in centered paragraphs:

- Project Title: The title should be descriptive of the project and must include the word “Quality” and the word “Information” or “Data”. As things change during the course of a project, the title on the final report maybe different than the title on the original proposal to reflect a change in direction, focus, or understanding.
- The subtitle should be “Submitted in Partial Fulfillment of Requirements for the Master of Science in Information Quality”
- Your Name
- Report Submission Date
- Your Faculty Advisor’s Name
- Name of the Sponsoring Organization
- Name and title of the Sponsor Supervisor

Executive Summary

The Executive Summary must be the first section of the final report. It is the most important part of the report. Don’t make it all background, it should clearly state the results of the project. The executive summary should cover three topics in the following order:

Background: First, start with a brief background. For example,
“XYZ is a *(company/agency/etc. - describe the sponsor and what the sponsor does, but be brief).* The information quality problem experienced by XYZ is that *(describe the problem or problems that were addressed by this project)*. The goal of this project was *(use the past tense in the summary because you should be describing things you have already done)* to *(design/build/implement/modify/etc. use a verb here) a (system/process/database/… – describe the main project deliverable)* for XYZ that (describe briefly how the deliverable addresses the sponsor’s IQ problem).”

What you did: Next, describe what you did in the project. For example,
“In this project, I *(designed/ programmed/ analyzed/ implemented/ tested/etc. – a verb that describes specifically what you did)* a *(dataset/GUI/SQL procedure/Run Book/… some deliverable) so that (describing the purpose of what you did)*.”

This sentence should be repeated for each major task. For example, if you profiled a data source, then developed and tested programs or scripts to cleanse and standardize the data.

The Benefits to the Sponsor: Finally, describe the benefits of your project. For example,
“As a result of my project *(describe a direct benefit to XYZ, e.g. the time required to receive a file and load it into the database has been reduced from an average of 2 days to 4 hours. In addition, the number of non-conforming data items has been reduced from 60% to 5%.)*”

This sentence should be repeated for each major benefit. The metrics or statistics stated here should also appear in the body and in the Conclusion Section of the report.

Other guidelines for the Executive Summary section include:

- The Executive Summary must not be longer than one page.
- It should be broken into different paragraphs.
- It should recap the most important quantitative and qualitative results of the project.
- It should describe what you did in the project including measurements and demonstrating improvements.

- It should be written retrospectively. It should be about what you have already done (past tense) and not what you propose to do.
- Don't save any surprises for the conclusion, disclose everything important in the executive summary.
- The last paragraph of the Executive Summary should be strong. It should be more quantitative and cite the most important improvement statistics from your report to back-up your claims of improvement.

Background

- Describe the information quality issues and their impact on the organization.
- Describe any previous and/or current efforts by others in the organization to address these issues.

Quality Assessment at Start and End of Project

It is critical that the report clearly shows the improvement of information quality in some quantitative way. Quantitative measures of information quality taken at the both the start and the end of the project should demonstrate this improvement. Although some assessment measures can be anecdotal or subjective in nature, at least two of the measurements should be quantitative.

For the quantitative measures:

- Identify the dimensions or characteristics of the information quality being measured.
- Describe the methodology for taking the measurements.
- For smaller sets of measurement data, include in the body of the report, otherwise include in an appendix.

Description of Project Activities

In this section, clearly explain which parts of the project were completed by you, and which parts were completed by other participants.

It is always helpful to include figures to illustrate your results and other work. Things like a schema for the new database, a new process flow, screen shots of a new data entry GUI, or an example business requirement document.

Be quantitative: The review committee expects to see tables and graphs of data quality measurements. The metrics that you defined in your proposal should be calculated and shown in the report with measurements at the beginning of the project, during the course of the project, and at the end of the project.

WARNING: If the data you are working with contains personally identifiable information (names, addresses, phone numbers), do not show these in your report. Be careful when including examples, tables, and screen shots. If you want to show a screenshot that has personal data you should edit out the identifying information or use dummy (not-real) information in your report.

Conclusion

The conclusion should:

- Clearly state the benefits of the project to sponsor.
- Recap and summarize your key improvement statistics for each dimension and focus on the benefits for the company that you achieved. You can add a table or a scorecard for this purpose.
- Focus on the results of the project and not the background or activities of the project.

Acknowledgement

It is always a good idea to acknowledge the help you received. If you worked with other people, and part of their work is included in the report, then be sure to acknowledge the specific work that they did.

Appendices

Appendices allow you to include detailed information that would be distracting in the main body of the report. Examples of items you might have in an appendix include things like reference tables, questionnaires, SQL scripts, and programming code.

Oral Defense

The final step in the capstone project process is to make an oral defense of your project to the MSIQ committee.

Scheduling

Scheduling is coordinated by the Information Quality Administrative Assistant well in advance of the end of the semester. All defenses must take place on or before the last day of class. However, your presentation cannot take place before your final report has been approved. Failure to get approval for your final report will result in the cancellation of your defense.

Time Limitation

Each defense presentation is limited to a total of 50 minutes including introduction, presentation, demonstrations (if any), and questions and answers by the committee.

You must leave the last 10 minutes open for questions and answers by the committee, therefore the total time for your part of the presentation must be no more than 40 minutes. Manage your time carefully.

Defense Attendees

Sponsor Supervisor: In addition to yourself and the members of the MSIQ Committee, you should invite your Sponsor Supervisor. Although his or her attendance is not mandatory, it is always beneficial to have the supervisor present to speak about the benefits of the project to the organization. If the supervisor is not present, then the faculty advisor will solicit feedback on your performance by email.

Guests: The oral project defense is a public event and you may invite guests to attend your defense.

Remote Presentation

Local students living within 50 miles of campus are expected to make their oral defense in person. Remote students should make arrangements at the time of scheduling for a live webcast presentation using Blackboard Collaborate or another mutually agreeable webcasting technology.

Presentation Content and Format

The oral presentation should be supported by PowerPoint slides that illustrate the major points of the project. You are expected to explain the content of the slides, not to read the slides. The slides should focus on the key elements of your project. The following is a suggested layout for your PowerPoint deck.

Suggested Layout

- Title Slide
  - Title of your project as it appears on your final project report
  - Your name
  - Name of the sponsoring organization

  - Name of the person in the organization supervising your project
  - Name of your faculty advisor
- Background Slide
  - One slide describing your sponsoring organization, Where is it? What does it do?
- Problem Statement
  - One or two slides that explain the problem you addressed in your project
  - Why was solving this problem was important the organization?
- Project Objectives
  - List all of the things that you originally planned to accomplish and deliver during the course of your project. List everything in the original plan even if some of them were not accomplished or changed during the course of the project.
- Project Plan or Approach
  - Explain your step-by-step strategy for solving the problem.
  - Use graphics where appropriate
- Activity Slides
  - One or more slides that explain your activities (what you did) in the project.
  - This is where you tell the story of your project.
  - Be sure to credit others who may have helped you
  - Use graphics where appropriate
  - Point out any problems you encountered and how you overcame these problems, or how these problems caused the project deliverables to change from the original design
- Project Accomplishments
  - List and explain what you accomplished.
  - Try to be quantitative (use numbers and statistics) as well as qualitative in your descriptions
  - If the project was to improve data quality, this is where you show the before and after measurements of the dimensions of DQ that you improved.
  - Use charts, graphs, and other graphics where appropriate
  - Explain any differences between your original plan of accomplishments and what you actually delivered, and what caused these differences.
- Conclusion
  - Explain how the project benefited the sponsor.
  - Explain what you learned from doing this project
- Acknowledgements
  - (Optional) Special thanks to anyone who was especially helpful to you
- Questions? Slide
  - A slide to leave on the projection screen while you answer questions from the committee

Overall Guidelines for PowerPoint Presentation

- Don’t prepare too many slides! You should be able to tell your story with about 15 slides, don’t go over 20 slides, otherwise you will exceed your 40 minute time limit for presenting
- Use graphics such as screen shots, charts, flow diagrams, and graphs as much as possible. Try not to make the entire deck one long set of bullet points.
- For slides where you need to use text, don’t make the text too small or too dense. Any text should be in a font size large enough to be easily read by the attendees.
- Check your slides to be sure they do not to reveal any confidential information from your sponsor. For example, check screen shots to be sure they don’t show any personally identifiable information.
- Only use dark color fonts on white or very light color backgrounds. Be careful about using color in your presentation. The colors rendered by the projector may look different than the way they

appear on your computer screen. When this happens, some combinations of colored fonts and backgrounds can become unreadable.

Live Demonstrations:
Live demonstrations of systems or websites can be an optional part of your presentation. Demonstrations can be very helpful, but can also be risky. Be sure to test your demonstration in advance using the same equipment you will use in your presentation, and if possible, in the same room where you defense will take place. If you plan to give a demonstration, make allowances for the extra time by reducing the number of presentation slides appropriately. In addition, it is also a good idea to prepare extra slides covering the same content as your demonstration in case your demonstration fails to operate properly.

Equipment

- You are expected to bring your own laptop for the presentation. If you don’t have one, please let the Information Quality Administrative Assistant know that at the time you schedule your presentation.
- Your oral defense will be scheduled in a room in the EIT building that has UALR WiFi Internet access and a digital projector.
- The projector will have a VGA input connector. Be advised! Many newer laptops do not have a VGA output connector. If your system does not have a VGA output connector, then you are responsible for bringing an appropriate adapter that will connect your laptop to the projector’s VGA input. For example, if your laptop only has HDMI output, you need to bring an HDMI-to-VGA adapter to your presentation.
- Bring a backup copy of your PowerPoint presentation on a USB drive in case your system fails or is incompatible with the projector. This will allow your presentation be made using a different computer.

APPENDIX A - EXAMPLE PROJECT PROPOSAL

Project Proposal for the
Master of Science in Information Quality Program

Improving the Quality of Alaska Law Enforcement Information through Inter-Agency Data Sharing

Submitted by: **Samantha Jones**

Faculty Advisor: **Dr. Emily Smith**

Sponsoring Organization: **Alaska State Police**

Sponsor Supervisor: **Sgt. Millard Fillmore**
Chief Technology Officer, Alaska State Police
mfillmore@aksp.org
907-456-3898

Date Submitted: **8/20/2014**

Estimated Completion Date: **4/30/2015**

Project Description

The purpose of this project is to help the Alaska State Police with the creation of a fusion center. A fusion center is simply a way of bringing together ("fusing") disparate information about the same case that is spread across different law enforcement agencies. A fusion center is an effective and efficient mechanism to exchange information and intelligence among agencies. A Fusion Center can also help to maximize resources, streamline operations, and improve the ability to fight crime and terrorism by merging data from a variety of sources (Taken from the "Fusion Center Guidelines: Executive Summary").

Problem: The U.S. Department of Justice's Global Justice Information Sharing Initiative sets out the guidelines of what a Fusion Center should be, but they did not offer any standards on how a Fusion Center should operate or how the data should be collected, used, or reconciled. The Alaska State Police has asked for help in evaluating the technologies that exist in this market space. The items of interest in the evaluation this includes Entity Resolution, Pattern Analysis, and Case Management.

Proposed Solution: My portion of the overall project is to help show the State Police how data from multiple sources can be consolidated to provide a more complete view. Possible sources of data include Police data from local city and county agencies, property taxes, DMV records, Marriage/Divorce records, civil court records, etc. Other possible sources of data include Department of Corrections (DOC), Alaska Crime and Information Center (ACIC), and the Department of Finance Administration (DFA).

Objective: The objective of my project is to demonstrate how the integration of data can be accomplished. In producing the report and other deliverables, I hope to provide the State Police with an understanding of how they can begin to assemble a Fusion Center from an IT perspective.

Impact on Information Quality

The primary information quality issues in this project are related to the dimensions of

- **Completeness—**Officers in the field and case managers often do not have access to all the information about a suspect or person of interest.
  Metric: I will provide a form to the case managers that they will use to rate completeness for the information on a monthly basis. The metric will be calculated as the average of the rating over all case managers.
- **Access—**Important case information resides in many different agencies and databases is not readily available in one place.
  Metric: The number of agencies participating and number of records contributed on a monthly basis.
- **Timeliness—**Currently it can take several days to request and receive information from another agency.
  Metric: All of the completed information request forms that were sent to other agencies will be analyzed each month. For each completed request I will calculate the difference in days between the "request time" and the "delivery time". The metric will be the average time difference calculated over all of the requests for that month.
- **Value-Added—**No single piece of information may be that important by itself, but when assembled into a complete picture, it could solve or prevent a crime.
  Metric: The percentage of cases closured each month as reported in the monthly statistics.

(NOTE: Other dimensions that may be relevant for a project include: Free-of-error (accuracy), believability, reputation, relevancy, amount-of-data, interpretability, representational consistency, conciseness of representation, manipulability, security)

Approach

Because this project will be about the reconciliation and recognition of disparate data sources and differing data collection methods, I have requested access to a sample dataset(s) from the participating parties. If unable to gain access, schemas for the data will be requested to create simulated datasets. The activities will be carried out over a **30-week** (two-semester) timeframe.

The source data will be stored in a MySQL database. A “query engine” will access the different tables to simulate data acquisition from multiple data sources. The results will be returned, cleansed and standardized, then presented to the user. Counts will be produced by running queries against the database.

Deliverables

1. **An Information Product (IP) Map** (or series of maps) of the proposed data sources. This will allow the Alaska State Police to see how the data can be integrated and where problems can occur in the integration process.
2. **A written plan** of how the disparate data can be integrated and reconciled. This plan will be based on what is learned from the IP Maps and the interviews with the participants.
3. If time permits, **a simple prototype** demonstrating the entity resolution and data integration process.
4. Final Project Report
5. Slides for Oral Project Defense

Technology Description

The software program language and specifications that I plan to use in this project are:

- SAS dfPower Studio—I will use this software as the entity resolution tool for this project.
- MySQL 5.0—a free Relational Database. I will use it to store the demo data.
- Global Justice XML (GJXML)—an XML specification for allowing interoperability between data sources containing law enforcement data.
- National Information Exchange Model (NIEM)—the new specification that will supersede GJXML as the exchange mechanism.
- I will use Java to develop any software tools needed for the project.

Problems and Limitations

The major challenges I see with the project are around the data, both the availability and the quantity of data. If the data are not available, then the schemas for the participating agencies will be requested, and I will have to create simulated data based on these schemas. That may make the project go longer than anticipated. However, by creating my own simulated data, I can control the size of the datasets and avoid overloading the test system, thereby solving the quantity problem.

Ethical and Privacy Issues:

There are several ethical issues that I will have to deal with in this project. The first is access to information. I will first have to sign a Confidentiality Agreement and undergo a standard State Police Background Check. Also, the data that will be used will be restricted to Freedom-Of-Information (FOI) data and other publicly available data.

Project Timeline

First Semester	
September 11, 2015	Secure necessary forms; Request data and/or schemas
October 16, 2015	Initial research on Fusion Centers, Entity Resolution and interviews with participants. Start IP Maps
November 6, 2015	Preliminary data analysis
December 4, 2015	Develop data validation rules and data cleansing routines
Second Semester	
February 5, 2016	Complete IP Map; Start Work on Prototype
March 7, 2016	Complete Prototypes
April 4, 2016	Submit First Draft Report for Approval; Schedule oral defense
April 15, 2016	Finish Final Project Report
April 29, 2016	Oral Defense

APPENDIX B: Example Project Reports

Improving Data Quality in Identifying Customers through Twitter and managing Customer Data

Final Project Report

Master of Science in Information Quality

University of Arkansas, Little Rock

By

Example Student

TNumber

Faculty Advisor

Name of Faculty Advisor

UALR IQ Graduate Program

Organizational Supervision

Name of Sponsor Supervisor

Position in Sponsor Organization

Contents

Executive Summary

Name of Sponsor is a partner owned company in Little Rock that offers advanced solutions for companies to manage their master data. An agency has approached the company to design a solution for them to identify new customers through twitter, manage their existing customer data and design a new process for gathering customer information. They wanted a web service to listen to twitter continuously for tweets from users enquiring about vacations and offer them deals by responding with the best offers. The agency had accumulated customer data over a period of time through different forms but these forms did not have any fields mandatory and this caused missing values in mandatory fields necessary for communication. These forms did not have any validations on the format of the data being entered which caused data to be missing and inconsistent. This inconsistency had caused difficulty in reaching the customers.

The main objective of this project was to improve the process of identifying prospective customer through twitter and improve data collection process through a new customer registration form. I have designed a web service that would listen to twitter continuously for tweets about a certain hashtags, search the tweets for phrases that describe services being offered by the agency, identify prospective customers and contact them by either posting a message on their wall or by communicating through their preferred mode of communication. In order to improve the process, solutions have been proposed and implemented to improve relevancy of the tweets identified. I have also designed a web page for customer registration to streamline the process of gathering customer information. During this process, data quality issues in existing data have been identified and solutions have been proposed in designing the new web page so they would not reoccur again.

As a result of this project, the web service for listening to twitter showed improvement in identifying relevant tweets and helped reach customers better and uniquely. This eliminated the problem of users being contacted more than once by maintaining the customer information uniquely in database. The relevancy of the tweets identified was improved by 88% from the initial stage of listening to tweets based on hashtags. The Web registration form designed helped resolve data quality problems existing in data by making mandatory fields required on the web form and by validating format of the data entered. The solutions proposed in data cleansing phase of customer data have improved the completeness of the mandatory fields to be 100% and representational consistency to be 100% in existing data. The new registration form eliminated possibility of null values and helped maintain completeness and consistency in data representation.

Introduction

Social media has reshaped the way marketing is done about a product. Increasingly, more and more users' everyday are using social media to voice their thoughts, know information about a product or compare the services of different companies. In this world of social media, if a business has not established its presence in social media and is not active then there is a little scope for that business to sustain in the market. Websites alone are no longer sufficient to retain customers. Presence on social media lets an organization know more closely about their competitors, know customers they didn’t know existed, interact with customers more closely and the sales would be highway. It is also very inexpensive.

With over 554,750,000 active users on Twitter and about 400 million tweets every day, Twitter has changed the way marketing is done. People follow pages for business and organizations and find out about the offers posted. This is one way businesses can market their products and get new customers. To take advantage of Twitter, an agency wants to partner with Name of Sponsor to market the offers to customers in a real-time and reach the customers early before other competitors.

The aim of this project was to build a web service that listened to Twitter continuously, analyze the text and respond to customer with different deals the agency offered. Each day twitter is flooded with numerous tweets. Identifying the relevant tweets was a major challenge. The agency has existing customer data they gathered through different means and forms. The data collected was incomplete and inconsistent and was not useful for the new system. This was because the existing forms did not mandate the fields that were necessary to reach customers and neither did it check the format of the data being entered. Different forms accepted data in different formats and left most of the important fields blank.

Two solutions have been designed to resolve the problems. A web service that would continuously listen to twitter about certain hashtags, analyze the text and respond to the customers based on whether the text is positive or negative or neutral. Then, a web form to capture the customer information that enforced rules on fields and format being entered. Data analysis and profiling has been performed on the existing customer data and data quality problems of completeness and representational consistency were identified and resolved so the data is consistent with the new system. Web service designed improved the relevancy of the tweets being captured there by identifying customers uniquely. Web registration form designed has eliminated the possibility of the data quality issues to reoccur and has made the process of gathering customer information much easier than before. Figure 12 under Appendix A shows the web page and the registration form designed for the agency.

Deliverables

Following were the list of deliverables that I delivered to Name of Sponsor as a part of my graduate project.

- Web registration form for the customers to register that ensures all the mandatory fields and formats checked.
- Web service designed in Provenir Studio 7.2 that listens for identifying customers through twitter.
- Database design to store Customer and Tweets information.

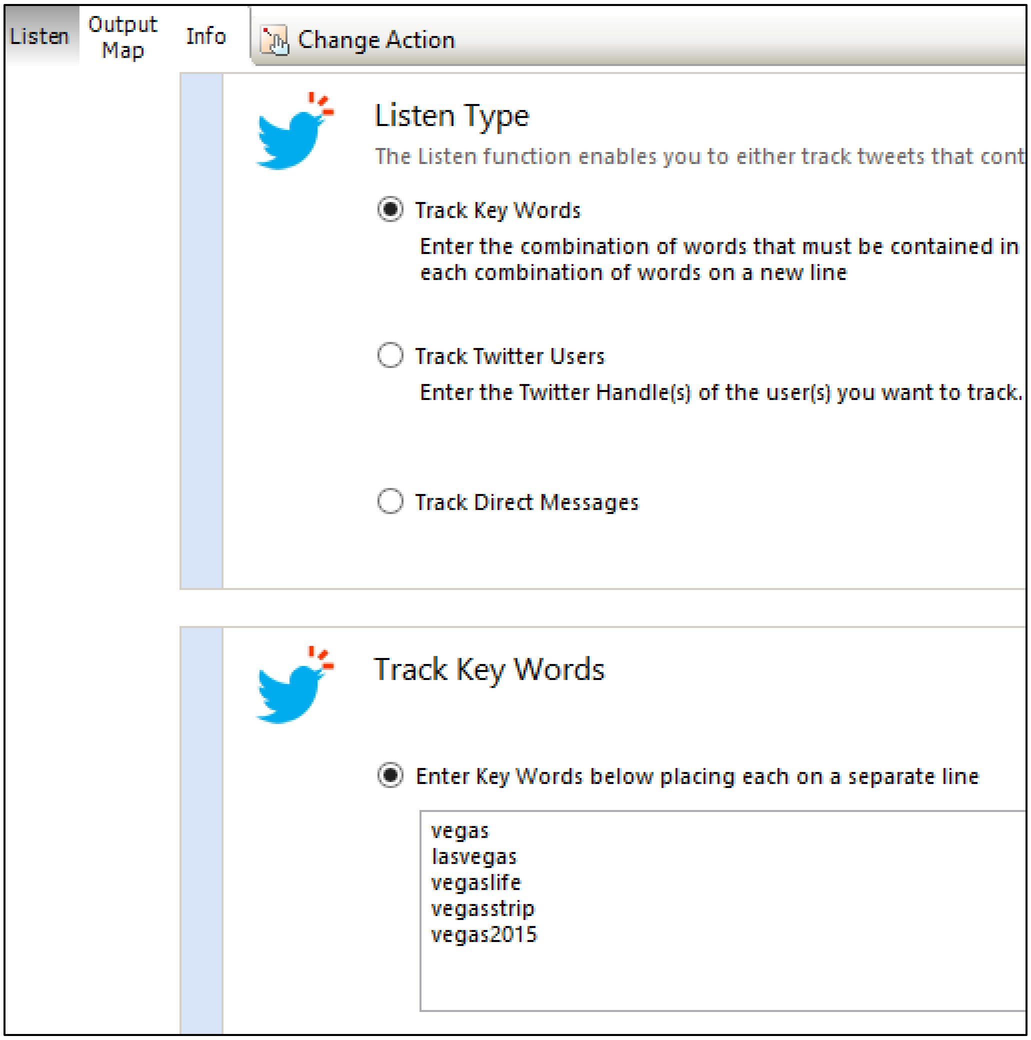
Data Collection and Extraction

Twitter Data

With millions of tweets flooding twitter every day, it is very important only to mine and extract those tweets that are relevant to the business analysis. Hashtags play a vital role in identifying tweets about an event or a place or about anything that is going on. The agency initially wants to market the offers they deal in Vegas.

Hashtags representing Vegas have been identified and I have extracted data from Twitter using Provenir Studio 7.2. It continuously listens to twitter and extracts all the information Twitter API provides about a tweet. Figure 1 shows the hashtags used and Twitter Listen process. Throughout this process, these hashtags have been listened to identify customers. Twitter API returns 57 attributes from each tweet but not all attributes are essential. I have only extracted the following attributes from Twitter API and mapped it to twitter Data in Provenir Studio. Figure 2 shows the process in Provenir Studio.

- **Name**: Name of the user as given by Twitter API
- **Screenname**: screen name of the user used to identify customer uniquely
- **Lang**: language of the tweet
- **Text**: content of the tweet
- **Tweet ID**: Twitter identification number of the tweet
- **Twitter ID**: Twitter identification of the user



Listen Type

The Listen function enables you to either track tweets that cont

● Track Key Words

Enter the combination of words that must be contained in each combination of words on a new line

○ Track Twitter Users

Enter the Twitter Handle(s) of the user(s) you want to track.

○ Track Direct Messages

Track Key Words

● Enter Key Words below placing each on a separate line

```
vegas
lasvegas
vegaslife
vegasstrip
vegas2015
```

Figure 1 Hashtags specified for Twitter Listen

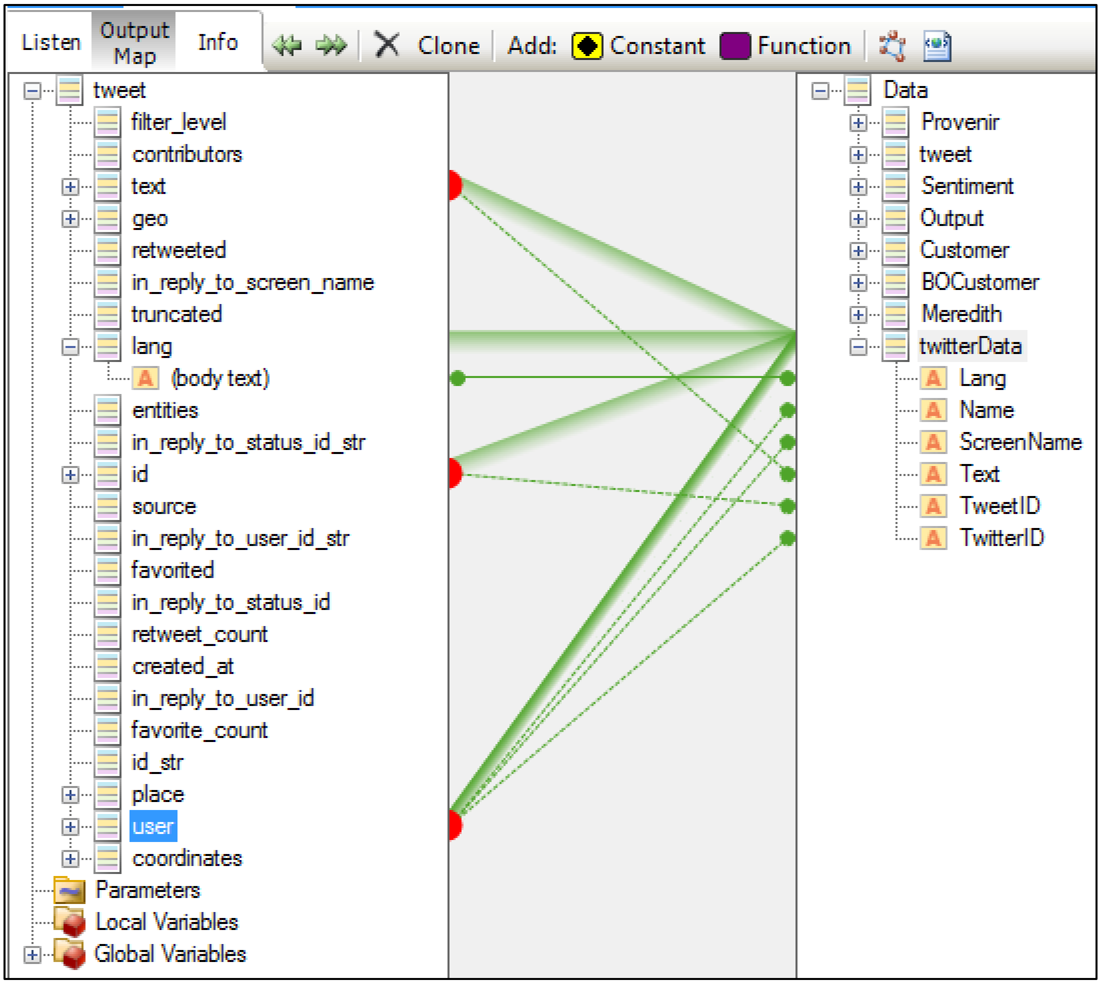


Figure 2 Data pulled from Twitter API

Customer Data

The agency has collected information about their customers through different forms and has extracted the data into an excel sheet. This process of gathering customer information was not properly implemented resulting in null values in mandatory fields and data being inconsistent. The new registration form for customer registration has resolved these issues. Information shared had the fields listed below.

- First Name
- Middle Name
- Last Name
- Email
- Mobile
- Twitter Screen Name
- Customer Type
- Level
- Contact Preference

Data Quality Assessment

Twitter Data

1. Data Relevancy

Twitter platform can be used to search for new clients and is considered second most successful way after content marketing [6]. This can be done by filtering out the tweets from twitter based on the hashtags that are most essential for business and then searching the tweets for relevant phrases or words that match the services offered [6]. After the tweets are found, customers can be contacted by leaving a comment on their wall.

Twitter is used to post a lot of content including the advertisements from competitors and user experiences. Eliminating irrelevant tweets by identifying phrases or words that correctly identify the services offered helps identify prospective clients better.

Data Relevancy Assessment

This metric is measured by the percentage of tweets reduced from initial stage until final stage where customer is contacted and helps to evaluate how well the process has performed.

Solution

Every tweet has a sentiment that could be positive, negative or neutral. Tweets could have URLs mostly posted by competitors for their offers or by users sharing their experiences. Retweets are posts forwarded or reposted by another user. Filter 1 filtered out all the tweets that have negative sentiments, tweets that had URLs and those that were retweets. After filtering out these irrelevant tweets, they are searched for key phrases that match the services offered by the agency. The agency offered services in for different categories listed below. Table 1(a) shows the usage most frequent words used in tweets for identifying these categories according to [5]. Figure 3 shows the relevance of these words in tweets according to [5]. Table 1(b) represents all the phrases used along with those in Table 1(a) in filter2 to identify the relevant tweets and grouped into four categories listed below.

- Travel
- Hotels
- Food
- Fun

Most frequent words
Attractions
hotel
park
accommodation
activities
shopping
events
tourism
restaurants
nightlife

Table 1(a) describing most frequent word used in Tweets according to [5]

Phrases looked for describing services offered by agency grouped into 4 categories			
Travel	**Hotels**	**Food**	**Fun**
break	hotel	restaurant	gamble
holiday	motel	bar	nightlife
weekend	inn	eatery	spa
vacation	lodge	food	event
tour	rest	breakfast	fun
trip	accommodation	bf	park
journey	stay	dinner	pastime
visit	overnight	cafe	nightlife
interval	quarter	grill	shopping
out		pizza	monument
fly		lunch	show
drive		eat	game
trek		bistro	activities
cruise		booth	strip
travel		buffet	casino
picnic		drink	amusement
season		booze	dance
attraction		alcohol	music
			pub
			party
			l

Table 1(b) representing phrases used in Filter 2 according to categories

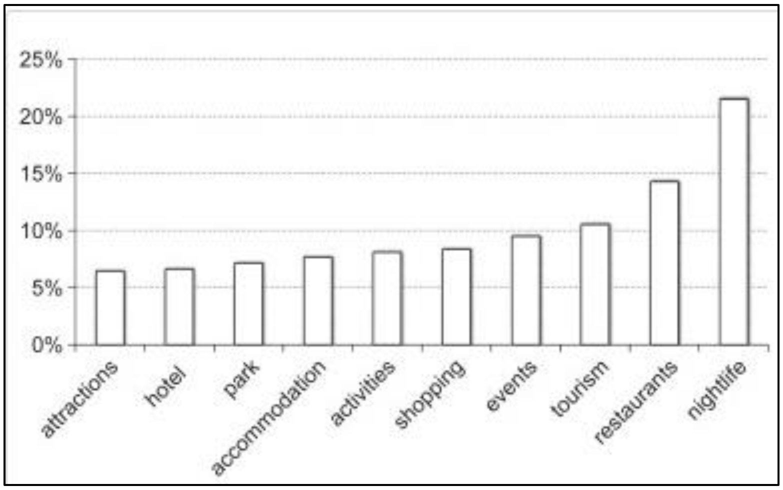


Figure 3 Bar chart explaining the relevance of these words in tweets [5]

Relevancy Metric Measurement

Data relevancy metric is used to measure how relevant is the data.

Relevancy Metric = 1 – number of tweets filtered / total number of tweets from initial stage

From the initial stage of listening to twitter for specific hashtags, irrelevant tweets are filtered out in two different filters. Filter 1 eliminated tweets having URLs, tweets that are not in English, tweets that have negative sentiment and retweets. Filter 2 searched tweets for phrases or words describing the services offered in four different categories. Table 1(c) describes the number of tweets reduced from initial stage to final stage along with relevancy metric for test runs for 10000 tweet. Figure 4 in Uniqueness section below shows the twitter run in progress.

no. of Tweets			Relevancy Metric		
Initial Tweets	**After Filter 1**	**After Filter 2**	**Initial Tweets**	**After Filter 1**	**After Filter 2**
10000	4760	1044	0	0.52	0.88

Table 1(c) describing number of relevant tweet and relevancy metric

This metric shows a huge improvement in identifying tweets from initial stage through different filters. Relevancy metric is increased from 0 in the initial stage to 0.52 after the Filter 1 and 0.88 after the Filter 2.

2. Uniqueness

Uniqueness dimension of the data quality defines the extent to which expected attributes are unique in your data set. This metric defines that no event or metric will be recorded more than once and that there should be no duplicate records in the data.

Data Uniqueness Assessment

Screenname attribute in Twitter data table is unique identifier for that table and also the unique identifier of customer as provided by Twitter. In the process of listening to Twitter, it is possible that same user will tweet multiple times. Each time a user tweets, it is heard by Provenir Studio and is processed and saved in the database as a new record. During the initial stage, it was observed that there were multiple records for the same screen name violating the definition of uniqueness metric. With this problem, the users would be contacted more than once and if user tweeted more than once on same day, his wall would be filled with response message from agency as many times tweeted. The agency wanted the customers to be contacted only once.

Solution

(1) **Data Cleansing:** As part of data cleansing phase, all the records that were not unique or occurred more than once were identified by querying data base. All the records that were obtained were deleted from data base.

(2) **Data Validation:** The Screenname column has been made unique in the Twitter data table. Then, in the process of collecting Twitter data, new process has been added that first identifies if the screenname of the user already exists in Twitter data table before saving tweet to the database. If the screen name is found, the text field of that record is updated with the new tweet and further process is continued. If the screen name is not found, then the record is inserted into the database and a message would be posted to customer wall. Figure 4 below shows the IdentifyUniqueCustomer process in Provenir studio that identifies if screenname already exists in the data base and depending on the result, the new record is either inserted or updated.

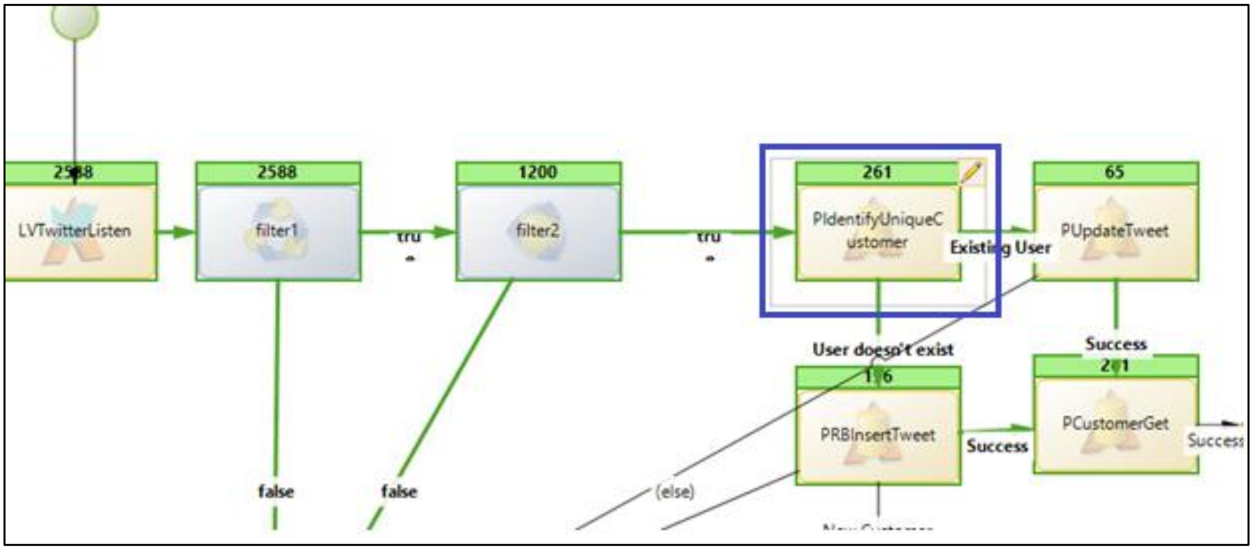


Figure 4 Process explaining identification of unique customer

Measurement

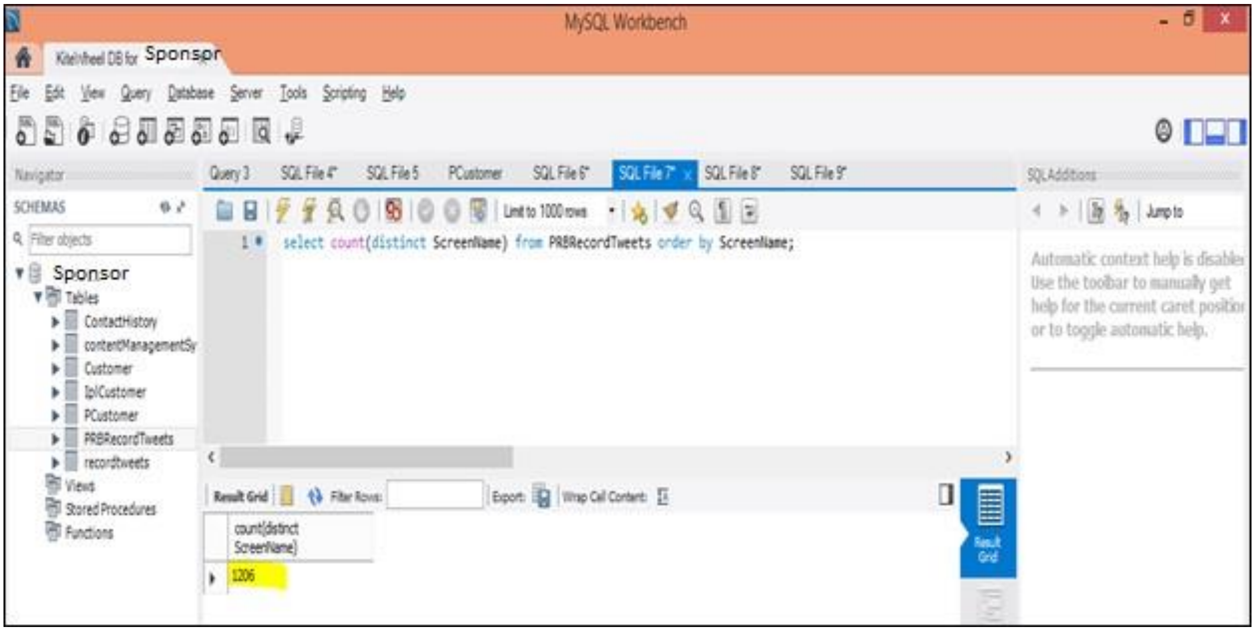
Data Uniqueness metric defines how unique is the data in primary fields of your data. It is calculated using the following formula.

Uniqueness = 1 – no. of duplicate records / total no. of records

All the duplicate records have been deleted in Twitter data table. Validation has been placed on database that Screenname field should be unique. New process has been added to the entire flow that when an existing user tweets, his record is updated in the table instead of creating a new record. Figure 5 below shows results of query run on database to identify unique customers. Table 2(a) shows the number of duplicate records and uniqueness metric.

	no of records	duplicate records	Uniqueness Metric
before	1300	94	0.92
after deleting duplicate records	1206	0	1

Table 2(a) describing duplicate records and uniqueness metric for Screenname field



```
select count(distinct ScreenName) from PRBRecordTweets order by ScreenName;
```

count(distinct ScreenName)
1206

Figure 5 Database query displaying unique screenname in Twitter data table

Customer Data

Data analyses and profiling of the customer data has been done using DQA Analyzer and the results of the analysis is listed below.

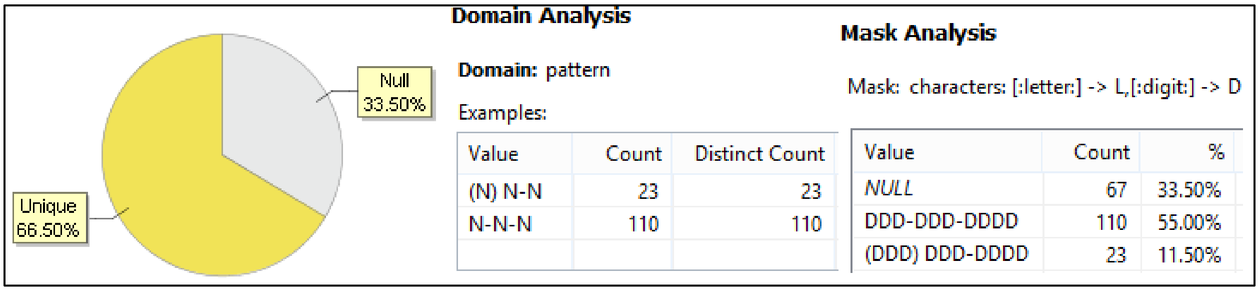
3. Completeness

Data completeness dimension of the data quality defines the extent to which the expected attributes in the data set are provided. It refers to whether or not all the necessary data required to meet the current and future business demand are available in the data source [3]. This metric requires that all mandatory fields in a data set must be complete fields and those fields that are not mandatory may not be complete.

Data Completeness Assessment

Field analysis for the mandatory fields' *first name, last name, email, mobile number, customer type, level* and *contact preference* was performed to evaluate completeness. Required fields *mobile number, customer type and level* that were essential for evaluating and contacting customer were found to have null values. It was identified that the earlier processes did not have any mechanism to mandate these fields. Figures [6, 7 and 8] display field analysis for mobile number, customer type and level fields respectively.

(a) Mobile number



Domain Analysis

Domain: pattern

Examples:

Value	Count	Distinct Count
(N) N-N	23	23
N-N-N	110	110

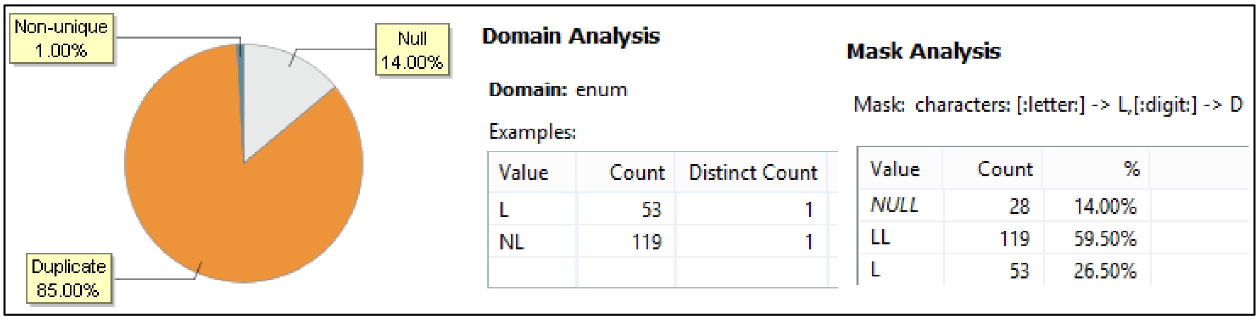
Mask Analysis

Mask: characters: [:letter:] -> L,[:digit:] -> D

Value	Count	%
NULL	67	33.50%
DDD-DDD-DDDD	110	55.00%
(DDD) DDD-DDDD	23	11.50%

Figure 6 Field analysis for Mobile Field

(b) Customer Type field



Domain Analysis

Domain: enum

Examples:

Value	Count	Distinct Count
L	53	1
NL	119	1

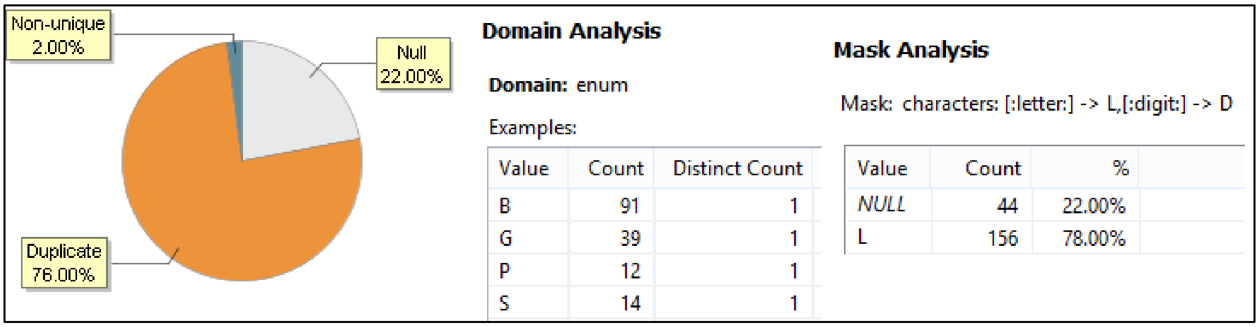
Mask Analysis

Mask: characters: [:letter:] -> L,[:digit:] -> D

Value	Count	%
NULL	28	14.00%
LL	119	59.50%
L	53	26.50%

Figure 7 Field analysis for Customer Field

(c) Level field



Domain Analysis

Domain: enum

Examples:

Value	Count	Distinct Count
B	91	1
G	39	1
P	12	1
S	14	1

Mask Analysis

Mask: characters: [:letter:] -> L,[:digit:] -> D

Value	Count	%
NULL	44	22.00%
L	156	78.00%

Figure 8 Field analysis for Level field

Solution

(1) **<u>Data Cleansing</u>**: The agency ranked customers as loyalty and non-loyalty members based on certain rules. Please refer to Appendix A for data dictionary for Customer Type and Level fields. Table 3(a) describes the rules as specified by the agency to fill out the missing values in these fields. As part of data preparation step, I have manually updated these

fields according to the rules in Table 3(a) in MySql table. I have updated the mobile number field to 'NA' for those values that were missing.

Customer Type	Default if Level null	Default if both fields null
L	G	
NL	B	NL - Customer Type, B -- Level

Table 3(a) Rules specified by agency to fill out missing values in customer type and level fields

(2) Data Validation: Mobile number field was on the new customer registration for designed to gather customer information. This field was made mandatory on the client side and evaluated before saving to data base. This will eliminate the null values in this field. Customer type and Level fields were internal for the agency and were not on the customer registration form. I have made these fields and mobile number field as required while creating data base for Customers. This prevents form to insert null values in this field thus eliminating null values. I have also placed the default values to be inserted in the data base for Customer Type and level fields. Figure 9 below shows the rules placed on database for these fields. Figure 11 in representational consistency section shows the validation rules placed on web registration form.

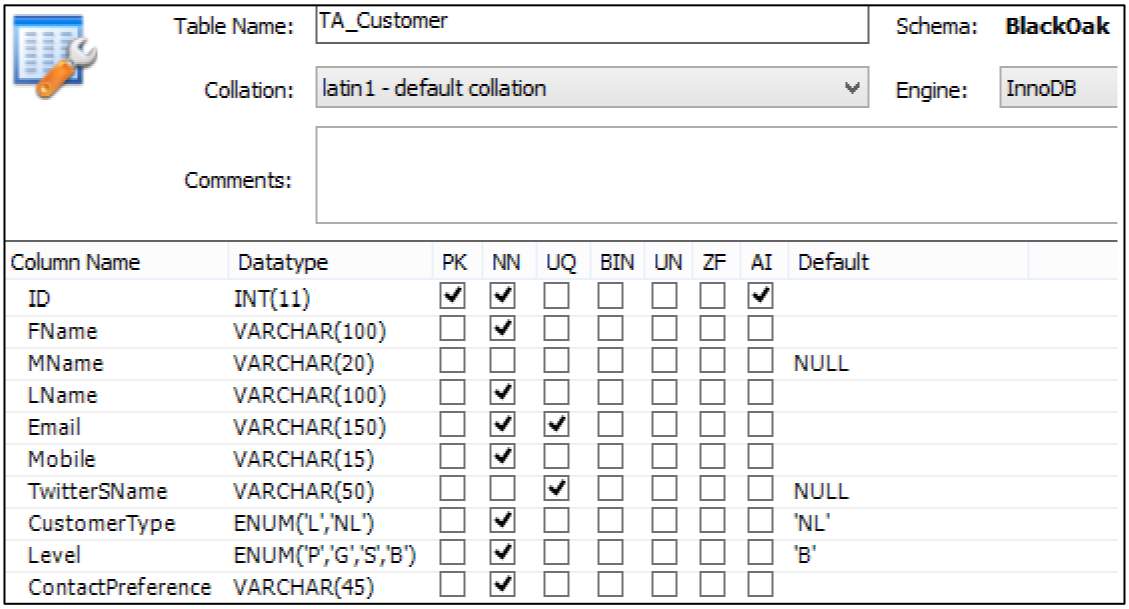


Table Name: TA_Customer
Schema: BlackOak
Collation: latin1 - default collation
Engine: InnoDB
Comments:

Column Name	Datatype	PK	NN	UQ	BIN	UN	ZF	AI	Default
ID	INT(11)	☑	☑	□	□	□	□	☑	
FName	VARCHAR(100)	□	☑	□	□	□	□	□	
MName	VARCHAR(20)	□	□	□	□	□	□	□	NULL
LName	VARCHAR(100)	□	☑	□	□	□	□	□	
Email	VARCHAR(150)	□	☑	☑	□	□	□	□	
Mobile	VARCHAR(15)	□	☑	□	□	□	□	□	
TwitterSName	VARCHAR(50)	□	□	☑	□	□	□	□	NULL
CustomerType	ENUM('L','NL')	□	☑	□	□	□	□	□	'NL'
Level	ENUM('P','G','S','B')	□	☑	□	□	□	□	□	'B'
ContactPreference	VARCHAR(45)	□	☑	□	□	□	□	□	

Figure 9 Database design showing mandatory fields and default values for Customer Type and Level Fields

Measurement

Data completeness metric is used to define how complete is your data in the mandatory fields.

Metric Definition

Completeness = 1 – number of incomplete records / total number of records

All the null values in mandatory fields have been updated as per agency rules. Validation rules have been placed on the registration form and database as well so null values cannot be entered again. Table 3(b) shows completeness metric before and after data cleansing.

Data Field	Incomplete values		Completeness metric	
	Before	After	Before	After
Level	44	0	0.78	1
Customer Type	28	0	0.86	1
Mobile	67	0	0.66	1

Table 3(b) representing incomplete values before and after data cleansing

Table 3(b) shows a huge improvement that all the values in mandatory fields are now complete and the metric is 1.

4. Representational Consistency

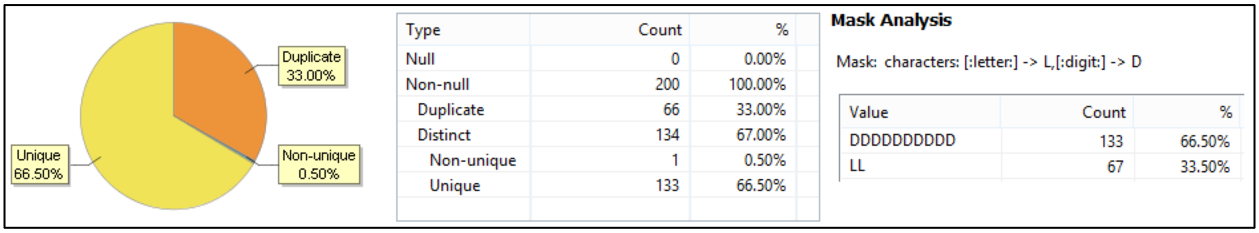
This metric defines that data representation should be consistent throughout the system. Same data can be represented in different formats causing difficulty in interpretation. Consumer of the data must be able to interpret the data and when data cannot be interpreted then data is no longer useful.

Representational Consistency Assessment

Field analysis of Mobile field showed that data in this field is represented in different formats. It was observed that the earlier forms did not follow a standard representation for the same data. Furthermore, in the new system developed, this mobile number field was used to contact customers with help of a Gateway that requires a totally different format to be used. With the existing data, I had to convert mobile number to a format accepted by Gateway programmatically to contact customers. Figure 6 in data completeness section explains field analysis for Mobile number revealing 2 different patterns.

Solution

(1) **<u>Data Cleansing:</u>** As part of the data cleansing process, to make all the data in mobile number field consistent, I have manually converted format of the data to the one accepted by Gateway so every time customer had to be contacted, conversion to acceptable format programmatically was eliminated. This resolved the problem of inconsistency in existing data. Figure 10 below explains the data field analysis after converting field format.



Type	Count	%
Null	0	0.00%
Non-null	200	100.00%
Duplicate	66	33.00%
Distinct	134	67.00%
Non-unique	1	0.50%
Unique	133	66.50%

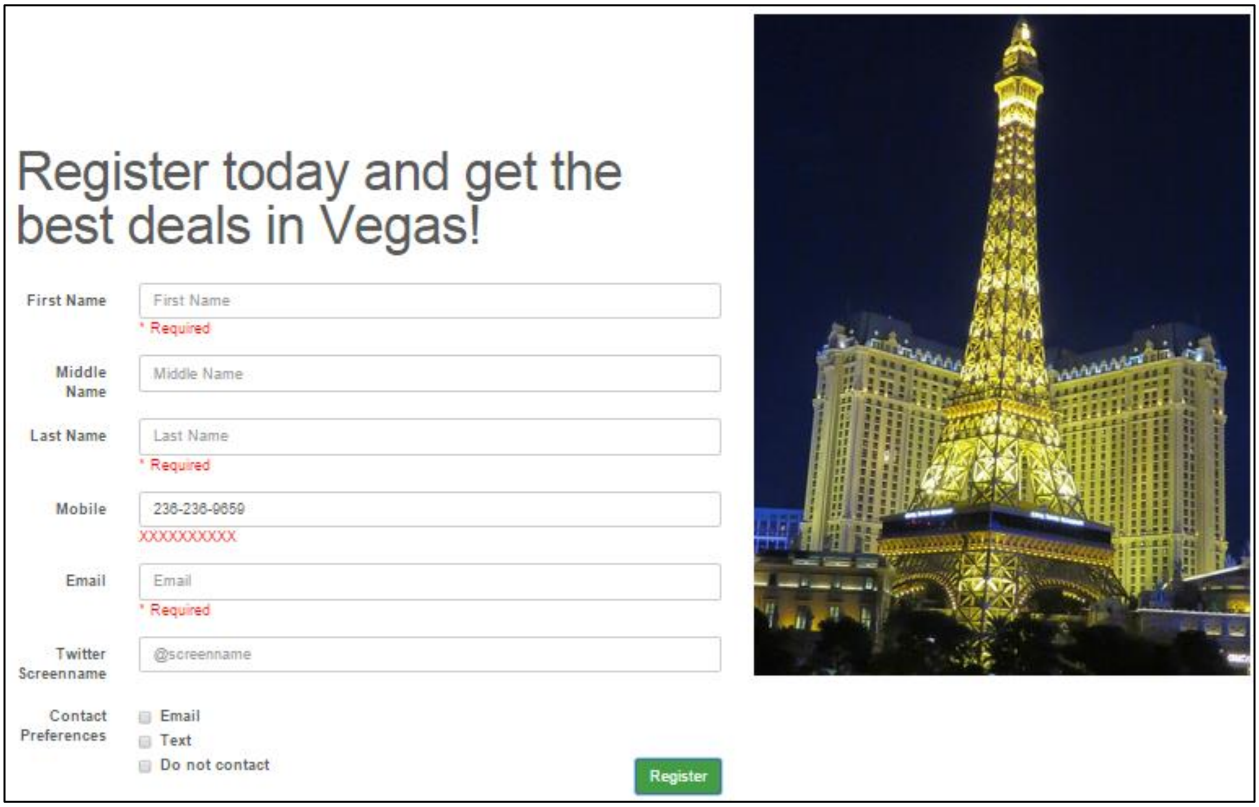
Mask Analysis

Mask: characters: [:letter:] -> L,[:digit:] -> D

Value	Count	%
DDDDDDDDDD	133	66.50%
LL	67	33.50%

Figure 10 Field Analysis of Mobile field after data cleansing

(2) **Data Validation:** To avoid the problem of inconsistency, I have placed a validation on type of data being entered in the new web form designed. This will eliminate the possibility of different formats being entered and force the user to enter only a specified format. Figure 11 shows the validation rules placed on the web registration form.



Register today and get the best deals in Vegas!

First Name	First Name * Required
Middle Name	Middle Name
Last Name	Last Name * Required
Mobile	236-236-9659 XXXXXXXXXX
Email	Email * Required
Twitter Screenname	@screenname
Contact Preferences	□ Email □ Text □ Do not contact

Figure 11 web registration form designed displaying validation rules

Measurement

Representational consistency is used to determine how consistent your data is.

Metric Definition

Consistency Metric = 1 – number of inconsistent fields / total number of fields

All the data in mobile field has been converted to a single format accepted by Text Gateway application. Format validation for the type of data to be entered in this field has been created on the new web registration form. Table 4(a) below describes the consistency metric before and after the data quality issue has been resolved.

Data Field	Inconsistent values		Consistency metric	
	Before	After	Before	After
Phone	133	0	0.335	1

Table 4(a) representing inconsistent fields and consistency metric before and after data cleansing

Metric shows a huge improvement as shown in Table 1(a) that the representational consistency metric has been improved from 0.33 to 1.This is a huge improvement the result of which is all the data in mobile field is now represented consistently.

Conclusion

The objective of this project was to design a web service to identify customers through twitter and reach them by posting on their wall about the deals offered by the agency and to design a web registration form that streamlined the process of gathering customer information. For any organization to succeed over their competitors, all it matters is how good the quality of their data is and how well it is maintained. During this entire project, I tried to maintain the data quality and offered solutions that resolved the existing problems in data and prevented them from reoccurring.

In designing the web service, the goal was to identify the relevant tweets and I have done to the best of my ability to identify relevant tweets. In the process of identifying relevant tweets, number of tweets listened on hashtags to the number of tweets responded were reduced by 88%. While storing the tweets in database, 100 % uniqueness of data has also been achieved. Data analysis and profiling was done as part of the data quality assessment for the agency's legacy data. Solutions were designed and implemented to eliminate these problems in existing data. In designing the new web registration form, all the issues identified during the data assessment phase were addressed and solution was provided so data quality problems would not reoccur in the future. 100% completeness of the mandatory fields has been achieved. This web registration form design streamlined the process of agency's customer data gathering and made it easy to maintain without any data quality problems. The new design of registration form made all mandatory fields to be required on client side and validated the format of the data being entered so the problems of completeness and inconsistency are avoided.

Acknowledgement

The success of this project is a reflection of the support each and every member of the Name of Sponsor team has shown to all whom I am very grateful. I take this opportunity to thank Name of Faculty Advisor who has given me an opportunity to work at Name of Sponsor. I would like to express my gratitude towards my mentor Name of Sponsor Supervisor without his support I would not have been able to complete this project. I would also like to thank Name of Another Person from Another Organization who has extended his tremendous support I needed for developing the web service. Above all this was the huge support morally I received from my parents and my husband without whom I would not have been able to complete the project.

References

[1]. "Data Quality Assessment: A Reviewer's Guide". *United Stated Environmental Protection Agency.* Print.

[2]. “The Six Primary Dimensions for Data Quality Assessment”. *Defining Data Quality Dimensions.* Print.

[3]. http://www.learn.geekinterview.com/data-warehouse/dw-basics/what-is-data-completeness.html

[4]. “Data Quality: Concepts, Methodologies and Techniques”. *Carlo Batini, Monica Scannapieco.*

[5]. Xiang, Z., & Gretzel, U. (n.d.). Role of social media in online travel information search. http://www.sciencedirect.com/science/article/pii/S0261517709000387

[6]. Putnam, J. (2011, June 14). 3 Simple Steps to Finding More Clients on Twitter http://www.copyblogger.com/prospecting-on-twitter/

Appendix A

A.1 Data Dictionary for Customer Type and Level Field

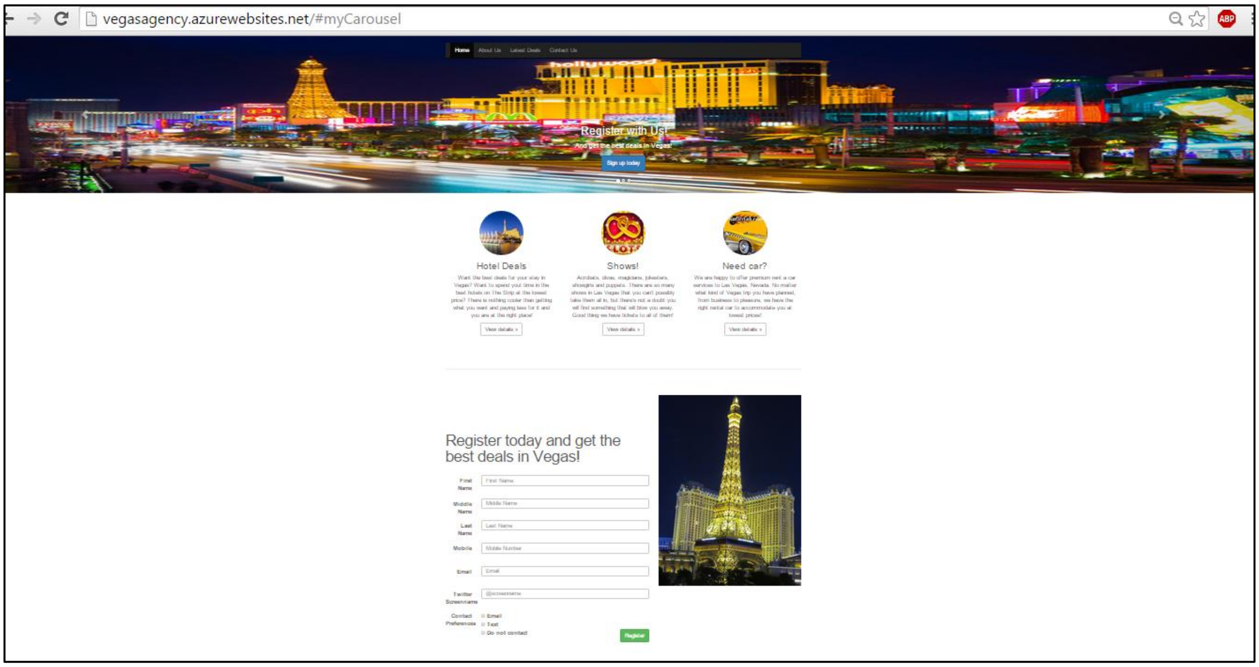
Customer Type	Meaning
L	Loyalty
NL	Non Loyalty

Customer Type field

Level	Meaning
P	Platinum
G	Gold
S	Silver
B	Bronze

Level field

A.2 Screenshot of the Web registration form



Register with Us!

And get the best deals in Vegas!

Sign up today

Hotel Deals

Want the best deals for your stay in Vegas? Want to spend your time in the best hotels on The Strip at the lowest price? There is nothing cooler than getting what you want and paying less for it and you are at the right place!

View details »

Shows!

Acrobats, divas, magicians, jokesters, showgirls and puppets. There are so many shows in Las Vegas that you can't possibly take them all in, but there's not a doubt you will find something that will blow you away. Good thing we have tickets to all of them!

View details »

Need car?

We are happy to offer premium rent a car services to Las Vegas, Nevada. No matter what kind of Vegas trip you have planned, from business to pleasure, we have the right rental car to accommodate you at lowest prices!

View details »

Register today and get the best deals in Vegas!

First Name: First Name

Middle Name: Middle Name

Last Name: Last Name

Mobile: Mobile Number

Email: Email

Twitter Screenname: @screenname

Contact Preferences: □ Email □ Text □ Do not contact

Register

Figure 12 Web site and the new registration form designed

"The goal is to turn data into information, and information into insight."

-- Carly Fiorina

Improving the Quality of a Compliance Training Information System

Project Report for

Master of Science in Information Quality Program
University of Arkansas at Little Rock

By

Student Name

Supervised by:

Faculty Advisor Name

Organizational Supervision by:

Supervisor Name

Director of Data and Research

(SPONSOR)

Contents

Executive Summary:

SPONSOR Data and Research was established to meet the growing data reporting requirements of the Individuals with Disabilities Education Act of 2004 (SPONSOR) and to undertake related research. They give trainings to the local education agencies to achieve accurate, valid and timely data to meet all state and federal reporting. The staff of the education agencies and other related organizations registers for these trainings and attend them to understand proper requirements of data at different cycles of an educational year. Certificate will be given for everyone who attends the training. So, the training staff maintains the records of who register and attend the trainings. They used Survey Monkey for the registration system, Excel to maintain the records of trainees and Mail Merge, PDF – Splitter to prepare and mail certificates.

The data setup of this system was of low quality with recurring problems and therefore not appropriate to use. Tracking registrations list was difficult using Survey Monkey, preparing certificates manually took lot of time failing to send certificates immediately and after all these hard work of manual process, the mails used to bounce back as many of the email addresses were wrong failing to deliver the certificates. This created a redundant, inconsistent, missing data and hard to use interface. The goal of my project is to design a solution that is easy to use, create workflows and automate the business process. Also, to have a reliable and efficient solution that is consistent and streamlined with the actual processes of the task.

With my project I replaced Excel based reporting system to relational database system. I used stringent conditions and constraints on the data with careful observations that helped to guide to a strict format to maintain consistency. With SQL Server, I have designed a new database structure and implemented normalization. Also I have applied conditions using triggers in the database to avoid adding/updating missing and irrelevant data. With this, I have applied conditions to the entry/update of the data from front end and back end. The quality of the data is measured in several Information Quality Dimension sections. Now, the staff can directly use the website (where I automated the workflows) to get all the data in the required view. Also, as when the trainee attendance is registered on the website, using that data, the Certificates are auto generated, and for this I used Visual Studio .Net and Business Objects Crystal Reports. I wrote code in C# which helped in normalization process, enhanced the measures of the quality and streamlined the processes. Also, 'bcrypt' password hashing function based on Blowfish symmetric-key block cipher is used to encrypt the passwords for proper security [1].

My project helped the organization to overcome the manual approach of using trainee data in excel. It reduced manual work, saved lot of time. This process has remarkably improved the poor data quality issues present within the data. This has also eliminated errors and enforced standards, indicating that the measures helped reduce data quality issues for information quality dimensions within the organization. The automation of workflows improved the accessibility and easy of data usage by the staff. The value added in terms of correctness of data showed an improvement from 48% to 100%. The other dimension to which value is added is completeness that showed an improvement from 36% to 100% and data timeliness shows an improvement from 7-10 business days to one business day after the end of training.

Introduction:

There is always a chance that the data might not be in appropriate form, inconsistent and could be wrongly interpreted while filling data which does not have any constraint. This leads to bad quality of data and this data cannot be used as expected. The main issue in the organization is with data given by trainee. Data is entered in survey monkey where it is not validated during entry. Also, it was required to overcome the issues of saving trainee data in Excel as it could be lost from the system if it was not properly handled. Also, it became difficult for the organization to deal with the emails which were not properly filled resulting in failure of certificate delivery. Tracking on these email issues was tedious and cumbersome process as sometimes it was difficult to find the trainee to send the certificate.

Other problem is with certificates preparation. Preparing certificates for 50 to 100 trainees for each workshop manually takes lot of time delaying in certificates delivery. When the data is improper, it is hard to find trainees attendance and send them certificates. One major issue with the trainee data that happened in the organization was some of the email addresses that were present Excel were incorrect and the mails were received by different persons. At times, the important notifications and confidential data were delivered to unintended individuals or failure in delivery. It was also observed that other duplicate issues in information and missing data was found in Excel which had to be corrected to errors and maintenance issues.

These both problems were solved using technologies of Microsoft .NET and Microsoft SQL Database systems, Crystal Reports and following an approach when data is dealt.

For the first problem, I had designed a few User Interface pages using Microsoft .NET. When a trainee registers in the registration page, account confirmation email will be sent to their email with a unique activation code. If they click on the link sent to their email, only then their account gets activated and then they can register for the workshops. This helps in the verification of the email addresses. All the data is being validated from both User Interface and from back end database as well using triggers when the trainee enters the data. This addressed the problems of data inconsistency, duplicates, missing values, incorrect data and other issues with the data. Also, the concept of normalization was applied on the data when worked in creating the tables in new database as this would further increase the better working with the data.

For the second one, I had created an automated workflow in Microsoft .NET and this helped the organization in overcoming the manual approach of preparing and delivering certificates for trainees who attended workshops without failure as the emails are verified. With this process, the trainees are not required to wait for long time for the certificates, as they will get it instantly with the available data. The implementation with screenshots displaying the validations and metadata used is discussed in the report. After clear understanding of the existing architecture and rigorous analysis on the data, issues related to redundancy and operability in the training system was resolved. During this process, the metrics, qualitative and quantitative analysis of the data was observed and is explained in the report.

Project Approach and Methodologies:

The approach that I used for this project is Cross Industry Standard Process for Data Mining (CRISP-DM) as shown in Figure 1, to deal with the trainee data issues [2].

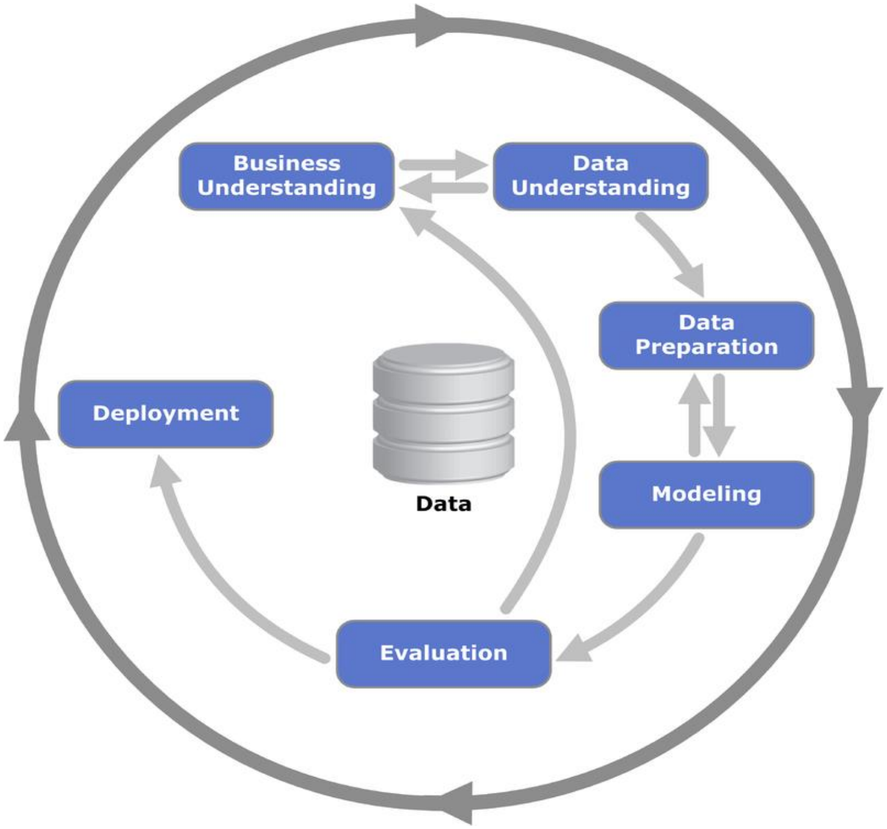


Figure 1: CRISP-DM Architecture

I first understood the requirements clearly with required constraints on the fields and also how the data has to be displayed and then went to data understanding phase. In this phase, I observed the possible implications and did data analysis to prepare proper database schema and understood required validations for the data.

After this, in the third phase it was important to prepare the data and improve the quality of the data considering the dimensions of Information quality. It is required to make sure appropriate data has been prepared so that it can be further applied. Now, once the data has been prepared and understood properly, it was important to design a model so that the data can be entered with proper validations and with proper relationship between the tables, the data can be queried easily.

Data is entered into the new system with different formats and with improper test data. This is evaluated and tested for working based upon the business requirements. Finally, the new system after improving the quality of data, the database can be queried for the accurate results and the model has been deployed for actual use.

Business Understanding:

In this phase I listed down all the important business and revenue aspects of the data. In this stage, the organization does not want to spend lot of time on manually preparing certificates and also verifying the email addresses. So, it wants to set objectives looking at the proper trainee data and automating the workflow process. The focus is on how much it has impact on the organization. For instance how satisfied is the staff with the appropriate data and processes followed. Also, this stage tries to answer questions related to describing the criteria for being a successful outcome. For example, knowing how better is the organization with the new process and system. So, answering these questions and the impact that it has on the organizations with respect to data and its fitness for use is critical. This stage helped me understand the different aspects of the training system data and get to know what exactly is required for the organization.

Data Understanding:

The data used here is very crucial for the organization to carry out training activities. The amount of quality information put to use out of the data present is very vital. As we already know that the design of the existing system and structure is poor with redundant data and incorrect data. As the data is considered as organization asset and that too if the organization could generate revenue from the data, proper care needs to be taken. It is essential to see that data is entered without any errors into the database and should be maintained for consistency with time.

From data understanding phase, I came to know that the main attribute is email address. Based on this attribute the trainee is marked attended and certificates will be delivered with help of the email addresses. Other important attribute are Trainee Name, Trainee Organization, and attendance. I came to a conclusion that the training system should make sure trainee enters mandatory fields like First Name, Last Name, Email, User ID, Password, Security Question, Security Answer and Date of Birth. All these fields are to be validated and Email address should be verified.

Data Preparation:

This phase helps in preparation of the data and overcoming all the issues with the data quality. In this phase, based on the data understanding, I built the database schema and the tables in this phase. The entity relationship diagram of the new database system is shown in Figure 2 displaying the relations between tables in the database. These relationships between tables help in connecting the data among tables and avoid creating orphan records. The user id from the user table is taken as foreign key for user id in registration table and user is marked attended for the workshop in registration table based on the user id eliminating any orphans creation.

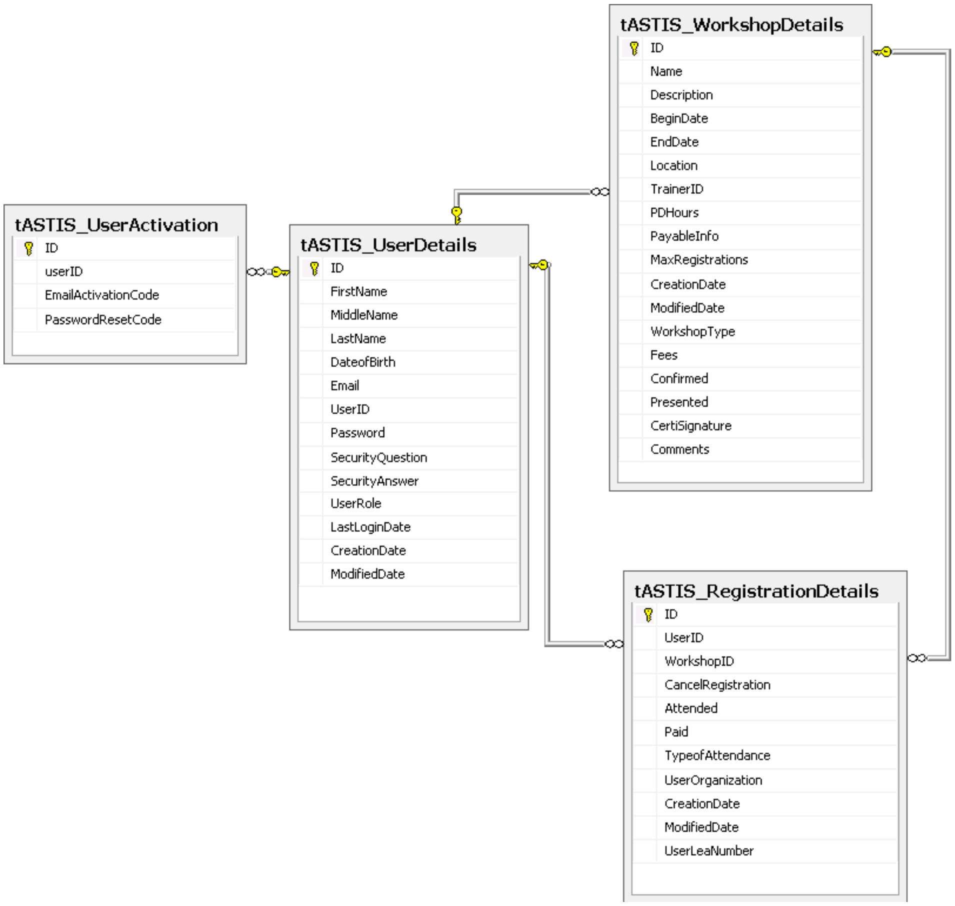
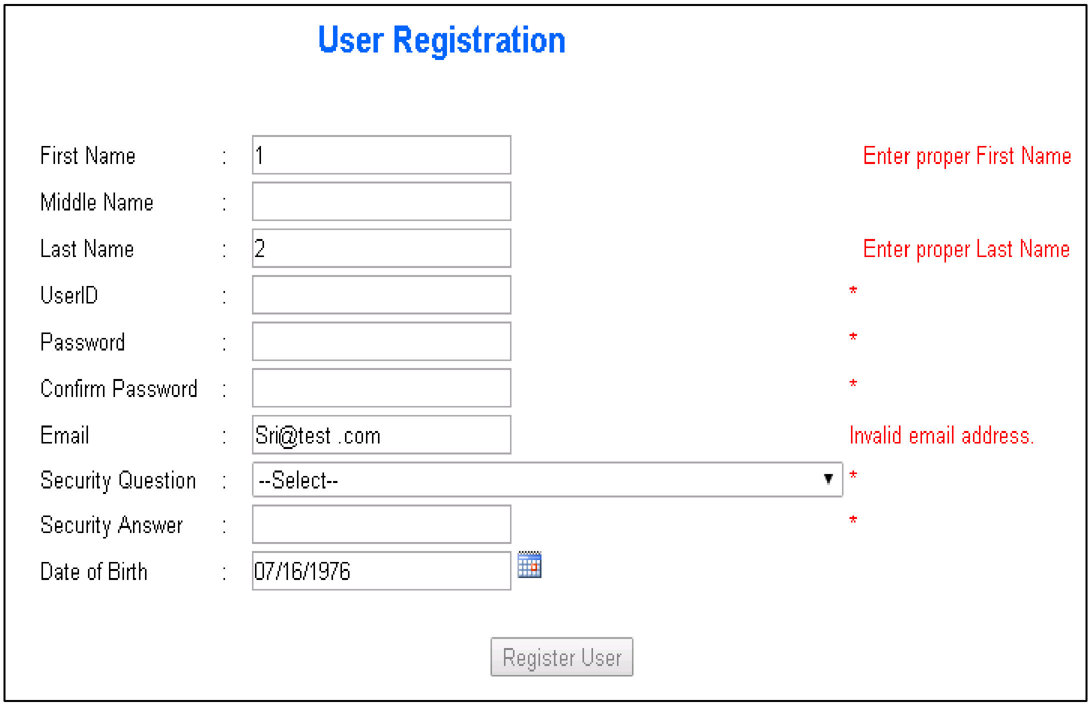


Figure 2: Entity Relationship Diagram

To register in the training system, user is required to enter mandatory fields and those will be validated based on the field conditions as show in Figure 3. If invalid data is entered by the user, the system does not accept it and so is not entered into the database. Also, few attribute values will be validated in the database with the help of triggers during inserting or updating of the records [Appendix I]. User cannot be less than 18 years or more than 100 years. This validation is taken care by the Ajax calendar in the Microsoft Visual Studio. All the dates for which the user is less than 18 years or greater than 100 years are disabled, avoiding user to enter their date of birth which is a mandatory field.

Once the user registration is successful, an email will be sent to the user to the email address entered during the registration with a unique activation code as in Figure 4. User will be able to login to the training system only if their account is activated. So they have to click on the activation link which is sent to their email address. If the email address is the user actual email address, then the user can activate their account and register for workshops. If it is wrong email address, then they cannot login to the training system and register for workshops.

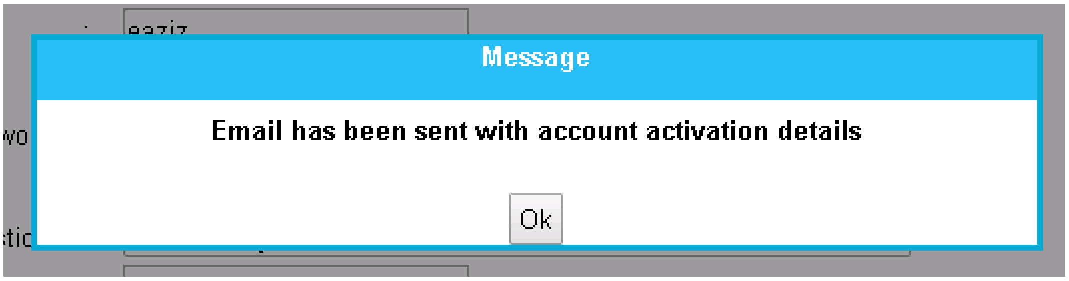


User Registration

First Name	:	1	Enter proper First Name
Middle Name	:		
Last Name	:	2	Enter proper Last Name
UserID	:		*
Password	:		*
Confirm Password	:		*
Email	:	Sri@test .com	Invalid email address.
Security Question	:	--Select-- ▾	*
Security Answer	:		*
Date of Birth	:	07/16/1976	

Figure 3: User Registration System

The date of birth text box is disabled and the user can enter date of birth only by using the calendar as shown in Figure 3. The calendar gives the date of birth only in one format which is MM/DD/YYYY. This avoids the formatting issues. The email address text box validates if the email is in the required format or not. If the email address has any space or any other irrelevant characters, the system does not accept it. The registration button is enabled only after all mandatory fields are entered and validated. Even if one of the entered fields is invalid, the registration button is not enabled and user cannot register in the training system. While registering for any workshop, the user is required to select the existing school district name or has to enter valid organization name. This helps the trainer understand from which organizations the trainees are coming.

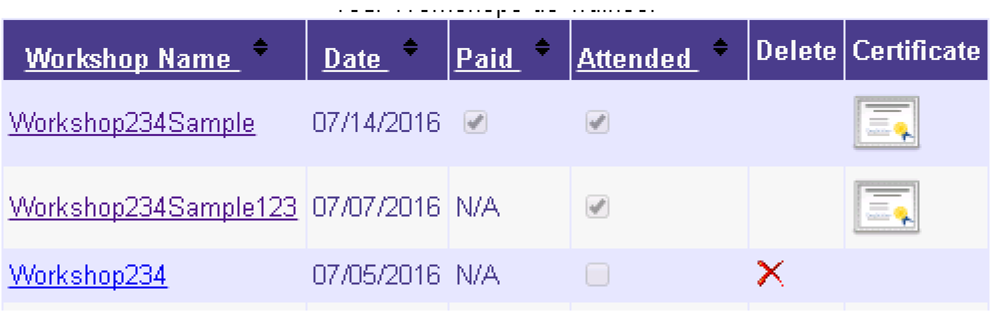


Message

Email has been sent with account activation details

Figure 4: Email Verification

After the user registers for the workshop, the trainer will mark the attendance of the user if they had attended the workshop. Based on the user details, the certificate will be generated automatically once the trainer marks the user as attended as shown in Figure 5. The name on the certificate will be same as the name provided by the user during registration.



Workshop Name	Date	Paid	Attended	Delete	Certificate
Workshop234Sample	07/14/2016	☑	☑		📜
Workshop234Sample123	07/07/2016	N/A	☑		📜
Workshop234	07/05/2016	N/A	☐	✗	

Figure 5: Automated Certificate Preparation

Modelling:

This is the important phase in the process of automation. During this phase, the implementation of the new system replacing the old one takes place. The data that was present in the old database that had no relationships and distinctly present in isolation from each other will be removed and new database with relationships will be added into the system. I used Microsoft Visual Studio and C# to validate the data before it enters into the database. Few of the attributes are validated in the database with the help of triggers when values for that attributes are being inserted or updated. Clean and appropriate data is loaded into the new database with required relationships between the tables and new rules were imposed as per the requirement. During this, it was ensured that the data was applied normalization of 1NF, 2NF, 3NF. Business Objects Crystal Reports are used to generate certificates automatically based on the trainee attendance to that particular workshop.

In the new training information system, the user needs to register in the system first. An activation code will be sent to the user's email address. On click of the activation link, the user's account gets activated and will be able to register for workshops. While registering for the workshop, user is required to enter their organization name. Once the user registers for the workshop, all the details of the workshop will be mailed to the user. After the workshop, the trainer marks all the attended users as attended in the training system. As soon as the trainer marks attendance, the certificate will be auto generated and will be available for download. User can login into the training system and download their certificates. This process is clearly explained in the sequence diagram Figure 6.

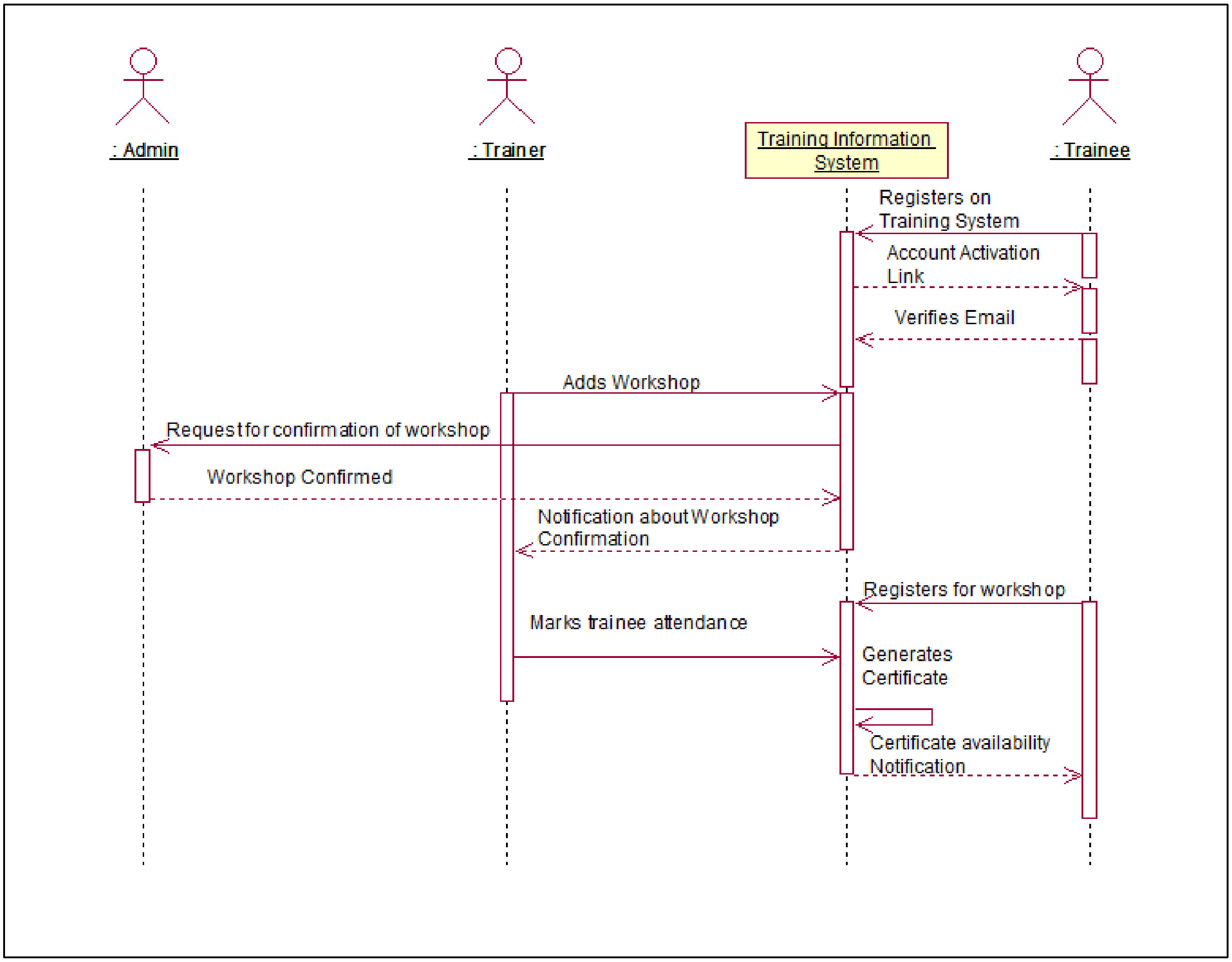

: Admin
: Trainer
Training Information System
: Trainee
Registers on Training System
Account Activation Link
Verifies Email
Adds Workshop
Request for confirmation of workshop
Workshop Confirmed
Notification about Workshop Confirmation
Registers for workshop
Marks trainee attendance
Generates Certificate
Certificate availability Notification

Figure 6: Sequence Diagram

Evaluation:

Once the modelling is done, the new system will be tested and evaluated to measure Information Quality dimensions for instance maintainability and overall integrity of the system for working conditions. The evaluation is done manually considering random set of data with both valid and invalid values. This helps us check if data is properly created and updated in the new database. There were few places where validation was not proper like Paid attribute which should accept only 'N/A' for workshops having no fees. I added these validations using C# and tested them. All the errors were fixed as and when appeared. The automation workflow also was tested for use and few steps failed when invalid data is given but later all of them were handled soon and finally overall system was implemented correctly.

Deployment:

The deployment phase of the project was carried out to see if the automated workflow was supported by the system for maintainability and it was successfully implement on local server. The system will be deployed in to production server by May 29th, 2016. The SPONSOR is now planning to use this system not only for internal trainings but also for data summits coming up in the near future. The code for validations and data normalization which was developed using C# can be found in appendix section of code sample [Appendix III].

IQ Dimensions:

- **Completeness:**

  Completeness of the data is referred to as the presence of the necessary information in the database. It is calculated as the ratio of the total number of incomplete records to the total number of records in the data base.

  Completeness Rating = 1 – number of incomplete items/total number of items

  The staff of the organization sends certificates to the trainees through the emails given during registration. But not all the trainees had complete email addresses resulting in the failure of certificate delivery. Same is the case with Organization account numbers also. The certificates should have both first name and last name of the trainee. As the trainees did not provide both the names while registering, the certificates did not have their full name. The organization had a list of all these requests and the incomplete email addresses. Based on these, completeness rating can be calculated.

  Measure of completeness for organization account number is as follows

  Measure of completeness = 1 – number of incomplete items/total number of items
  = 1 - 364/571
  = 1 – 0.64= 0.36

  This indicates only 36% of the data is complete for phone number and remaining records have issues in terms of format and should be validated for completeness.

  Similarly other attributes are also calculated for completeness and the below was observed.

  The Table 1 indicates those only 36.25%, 49.91% and 68.12% of taken attributes in the old system show some percentage of completeness and this data needs to complete and needed to be validated and remaining records have issues in terms of format and were validated for completeness in C#.

	Old System			New System		
Attribute Name	Incomplete Records	Total Records	Completeness Rating%	Incomplete Records	Total Records	Completeness Rating%
Email	182	571	68.12	0	362	100
Last Name	286	571	49.91	0	362	100
Organization Account Number	364	571	36.25	0	362	100

Table 1: Completeness Rating of User Table Comparison in Old and New System

With the validation done in C# and SQL for correctness, it was ensured to see that all confirm to a standard and unique value. It was observed that the completeness after the enhancement was 100% in the new system.

- **Accuracy:**

It is crucial for any data that it is accurate so that it can be used across the organization. From the bounce back emails list, it is easy to find the total number of inaccurate emails and can be measured against the total number of emails to find the free of error rating.

The accuracy measure for emails is as follows.

Measure of accuracy = 1- Number of items in errors/total number of items

= 1 - 297/571

= 0.4798

This indicates only 47.98% of the data is complete and remaining records of the old system are not accurate and should be validated for accuracy. These needs to be accurate and the emails are now verified by creating activation code using C#.

In the new training system, the measure of accuracy is 100% as all the emails are verified first before the trainee attends any workshop. The accuracy is achieved completely. By having the verified email addresses, the organization was very happy as it was easy to manage the operations related to trainee data and send them notifications and certificates.

	Old System			New System		
Attribute Name	Incomplete Records	Total Records	Accuracy Rating%	Incomplete Records	Total Records	Completeness Rating%
Email	297	571	47.98	0	362	100

Table 2: Accuracy Rating of Email Attribute Comparison in Old and New System

- **Timeliness:**
  Timeliness refers to the degree to which data represents reality from the required point of time. It is calculated as the average turnaround time since the information is received and the trainee is contacted. Whenever there is a workshop, validating the trainee data and sending them certificates required minimum of 7-10 days. This caused issues in terms of operability within the organization.

  With the new automated workflows and database system available in the system now, the data received is validated and standardized. It was made sure the completeness, consistency and other dimensions of quality are ensured and then the data is updated into the tables. Also, this made sure quick update to the database without any erroneous or duplicate data when update is done in the system. Now the trainee certificate is accessible with in a business day after the end of training.

- **Ease of understanding:**
  In the old system, trainer used to get confused with the trainee data as few may have same first name and no last names or vice versa with invalid email addresses. It was difficult to understand which trainee exactly attended the workshop to send the certificates. Data in excel sheet would be very confusing at times with duplicate columns and inconsistent data.

  In the new system, trainees must enter both their first name and last name. It also has verified email address of the trainee and valid organization account number. This makes the trainer easy to understand which trainee had actually attended the workshop and then mark them attended accordingly as the new database system do not allow duplicates and has consistent data.

- **Data Specification:**
  In the old system, all data is entered into excel sheet with no standards or documentation. Now in the new system, all tables and columns are created following standards and are documented which would help in managing metadata and reference data in the future. (If any Data Governance Project is implemented in organization)

- **Security:**
  In the old system, the data is not secure. Data is exposed to everyone who had access to the shared drive where excel is saved. In the new system, the windows security feature of the Visual Studio and ‘bcrypt’ password hashing function based on Blowfish symmetric-key block cipher are used to secure the data and encrypt the passwords for proper security [Appendix II].

Schema Quality:

This was an important phase in the project. It is more critical to see that a schema of appropriate and good design is created [3]. A great deal of importance was given to data quality and the main goal is to implement this well-designed and operable schema for the data. During this analysis and implementation for schema quality, some important aspects and concepts were applied to improve the schema quality. Eventually the overall working of the system was in a better and efficient way. Below are some of the implementations that were observed, implemented and applied to the system for improving metrics and dimensions of data quality.

Schema Standardization:

The existing data is being saved in Excel. There were two tables with no relation between them and at sometimes there used to be only one table with all data in it without attribute minimization. By working on this system rigorously, I found that there were many duplicates that were commonly spread across excel. The data was redundant as same columns were repeating.

- **Correctness based on model:**
  It is very important that all the tables in database confirm to a standard model and needed to be in conformance with it. Once this has been formed as basis, then all the attributes should be in agreement with the standard format. This was implemented to resolve the discrepancy with name attribute in the user table. The name field was divided into 2 attributes – [firstname], [lastname] but in most of the tables the name was only 1 attribute - [name]. These irregularities in schema were needed to be looked at first.

- **Consistency based on attributes:**
  Along with completeness, consistency is also an important criterion for the creation of schema. Consistency refers to the requirement that any given transaction must change affected data only in allowed ways. Any data written to the database must be valid according to all defined rules, including constraints, cascades, triggers, and any combination thereof. For example, The Paid attribute can have only values ‘Yes’, ‘No’ or ‘N/A’. Any other value is not allowed to be entered for this attribute. In new system, if a user attempts to enter something else, then consistency rule kicks in and disallows the entry of such a value.

- **Completeness based on requirements:**
  It is very important to have completeness in representation of attributes in accordance with the requirements. As it was observed that, there were no relational dependencies among the tables and had many duplicates, resulted in having redundancy and inconsistency in attributes. Also, there was no typical process followed to get the valid data. In one table, the column name is Trainee Name which includes both first name and last name which might have only first name or only last name or both. In other excel

sheet the name is in two columns named as first name and last name. The completeness with respect to requirements was missing in the previous system. I made sure it is applied properly in the new system.

Normalization:

Database Normalization is a technique of organizing the data in the database. Normalization is a systematic approach of decomposing tables to eliminate data redundancy and undesirable characteristics like Insertion, Update and Deletion Anomalies. It is a multi-step process that puts data into tabular form by removing duplicated data from the relation tables. [4].
It was observed that the existing database schema had no relationships and all the tables were isolated. In order to query something or update/insert the information in database was a tedious process as this also led to redundancy in data and undesirable anomalies were formed. It was the need of the hour to implement a relational database overcoming redundancy and minimization of the attributes in the tables to perform data manipulation operation on schema. It was made to ensure that all the tables were analyzed and designed in such a way so that they are in some form of normalized form with no redundancy of the data. In Figure 7, the tables and the normalized forms are shown.

Table	Normalized Form
Users	2 NF
Workshops	2 NF
Registration	3 NF

Figure 7: Tables and Normalized Forms

Hence, by implementing normalization it was observed that the redundant data was eliminated and ensured that the data dependencies are appropriate i.e., data is logically stored. Thus, by doing this to bringing to a highest level possible for normalization not worsening the situation with tables gave the best quality of the database.

Conclusion:

The main focus of the project was to improve the quality of the data with respect to trainee information and automation of training system workflow process in Microsoft Visual Studio. These processes helped in acquiring the highest quality of data within the organization. I used several methodologies and approaches to implement the automation and data quality goals. The approaches ensured that the data quality and automation processes enabled better understanding and maintainability. The existing data had data quality issues like duplication, inconsistency, redundancy in the data and incompleteness and invalid data is found.

To overcome the outlined data quality issues, automated workflows processes was implemented using C# in Microsoft Visual Studio and Business Objects Crystal Reports. These measures resulted in enhancing the speed of operability of the processes. They also lead to effective maintenance of the processes and made easily accessible. The Information Quality dimensions were measured to correctly weigh the qualitative and quantitative analysis of the data that were statistically significant. These approaches helped to analyze the data issues and validate the improvement in the schema quality. Finally, these improvements helped the organization to transform the trainee related data for efficient working and operation thereby decreasing the time spent by the trainers in finding the valid data and generating certificates.

The improvements in statistical analysis of the metrics indicate that the measures overcame data quality issues for information quality dimensions. The use of written code in normalization and improvement demonstrated that the quality enhancing measures streamlined the processes. The automation of workflows and improvement in metrics of data quality dimensions had considerable impact on trainee data usage with respect to the time and accessibility of the data. The value added in terms of correctness of data showed an improvement from 48% to 100% of the data post quality measures. Similarly, the value added to completeness was from 36% to 100% and finally for timeliness of the data was from 7-10 business days to one business day after the end of training.

These improvements added huge value to the trainers in the organization in terms of usability of the data, ease of handling of workflow processes and increased efficiency of trainee data management. The automation processes reduced manual workload of the trainer resulting in improved time management. These processes had significant effect on data governance within the organization leading to the mitigation of data related issues and burden on the corporate executives. In summary, all of the above improvements resulted in no data quality issues thereby immensely adding benefit to the organization.

Acknowledgement:

I take this opportunity to thank and express my gratitude to my Project Supervisor, Supervisor Name, for her excellent guidance with the project, keen observation and constant support with the project. I would also like to thank SPONSOR Data and Research for sponsoring and giving me the chance to work with the resources required for the project, otherwise this project could not have been possible to meet the timelines. I would also thank my Faculty Advisor for his priceless support and invaluable help and guidance and support for me with the project without which this could not have been possible. This will definitely be very good experience for me on projects and I shall take this forward. I would also thank Dr. Elizabeth Pierce as her classes IQ Policy and Strategy and IQ Theory. Her teaching style helped me a lot with my project and documentation.

References:

1. "NIST.gov - Computer Security Division - Computer Security Resource Center", Csrc.nist.gov, 2016. [Online]. Available: http://csrc.nist.gov/groups/STM/cmvp/.
2. "Cross Industry Standard Process for Data Mining", Wikipedia, 2016. [Online]. Available: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining.
3. O. Herden, Measuring Quality of Database Schemas by Reviewing – Concept, Criteria and Tool, 1st ed. 2016.
4. Support.microsoft.com, 2016. [Online]. Available: https://support.microsoft.com/en-us/kb/283878.
5. Information Quality Masters Classroom material.

Appendix I: Validation from Database

Trigger used for validation: Sample Code

```
create trigger check_paid on tASTIS_RegistrationDetails for insert, update
as
begin
        declare @fees decimal(6,2), @paid varchar(3)
        set @fees = (select fees from tASTIS_WorkshopDetails W join inserted on
        W.ID=inserted.WorkshopID)
        select @paid = paid from inserted
        if(@fees>0)
                begin
                        if(@paid!='Yes' or @paid!='No')
                                begin
                                        raiserror('Paid Value is wrong', 1, 1)
                                        rollback transaction
                                end
                end
        else
        begin
            if(@paid!='N/A')
                begin
                    raiserror('Paid Value is wrong', 1, 1)
                    rollback transaction
                end
        end
end
GO
```

Appendix II: bcrypt Password Hashing

```
// Pass a logRounds parameter to GenerateSalt to explicitly specify the
// amount of resources required to check the password. The work factor
// increases exponentially, so each increment is twice as much work. If
// omitted, a default of 10 is used.
string hashed = BCrypt.HashPassword(password, BCrypt.GenerateSalt(12));

// Check the password.
bool checkPassword = BCrypt.CheckPassword(password, hashed);
```

Appendix III: Code Sample

```
using System;
using System.Collections.Generic;
using System.Web;
using System.Web.UI;
using System.Web.UI.WebControls;
using System.Data;
using System.Configuration;
using System.Data.SqlClient;
using System.Web.Security;

namespace ASTIS
{
    public partial class Registration : System.Web.UI.Page
    {

        protected void Page_Load(object sender, EventArgs e)
        {
            //Making sure the register button is disabled till user enters all mandatory fields and
data is validated
            Master.BodyTag.Attributes.Add("onload", "Validate()");
                            //The trainer age can only be between 18 and 100
            calDOB.StartDate = DateTime.Now.AddYears(-100);
            calDOB.EndDate = DateTime.Now.AddYears(-18);
                            //Date of Birth textbox is disabled and date is entered from calendar
alone to maintain single date format and consistency
            txtDOB.Attributes.Add("ReadOnly", "ReadOnly");
            if (!IsPostBack)
            {
                calDOB.SelectedDate = DateTime.Now.AddYears(-40);
                if (this.Page.User.Identity.IsAuthenticated)
                {
                    FormsAuthentication.SignOut();
                }

            }
        }


        protected void RegisterUser(object sender, EventArgs e)
        {
            if (Page.IsValid)
            {
                //
                string userId;
                SqlConnection sqlConn = new
SqlConnection(ConfigurationManager.ConnectionStrings["dASTISConnStr"].ToString());
                {
                    using (SqlCommand cmd = new SqlCommand("sASTIS_UpdateUser"))
                    {
                        cmd.CommandType = CommandType.StoredProcedure;
                                                    //Adding data into User Table
                        cmd.Parameters.Add("@FirstName", SqlDbType.VarChar).Value =
txtFirstName.Text.Trim();
                        cmd.Parameters.Add("@MiddleName", SqlDbType.VarChar).Value =
txtMiddleName.Text.Trim();
                        cmd.Parameters.Add("@LastName", SqlDbType.VarChar).Value =
txtLastName.Text.Trim();
                        cmd.Parameters.Add("@DateofBirth", SqlDbType.Date).Value =
txtDOB.Text.Trim();
                        cmd.Parameters.Add("@Email", SqlDbType.VarChar).Value = txtEmail.Text.Trim();
                        cmd.Parameters.Add("@UserID", SqlDbType.VarChar).Value =
txtUserID.Text.Trim();
                        cmd.Parameters.Add("@Password", SqlDbType.VarChar).Value =
txtPassword.Text.Trim();
                        cmd.Parameters.Add("@SecurityQuestion", SqlDbType.VarChar).Value =
ddlSecurityQ.SelectedItem.Value;
                        cmd.Parameters.Add("@SecurityAnswer", SqlDbType.VarChar).Value =
txtSecurityA.Text.Trim();
                        cmd.Parameters.Add("@UserRole", SqlDbType.VarChar).Value = "Trainee";
                        cmd.Parameters.Add("@UID", SqlDbType.Int).Value = "-1";


                        cmd.Connection = sqlConn;
                        try
                        {
                          sqlConn.Open();
                          userId = Convert.ToString(cmd.ExecuteScalar());
                        }
                        catch (Exception ex)
                        {
                          HttpContext.Current.Response.Cookies["errorMessage"].Value = ex.Message;
                          Server.Transfer("Error.aspx");
                          throw ex;
                        }
                        finally
                        {
                          sqlConn.Close();
                          sqlConn.Dispose();
                        }
                    }


                    switch (userId)
                    {
                                                    //Making sure that no duplicates are created
                        case "-1":
                          lblMessage.Text = "Username already exists.<br /> Please choose a different
username.";
                          break;
                        case "-2":
                          lblMessage.Text = "Supplied email address has already been used.";
                          break;
                        case "-3":
                          lblMessage.Text = "User already exists";
```

```
                            break;
                        case "-5":
                            lblMessage.Text = "Please try again";
                            break;
                        case "-6":
                            lblMsg.Text = "Registration successful. <br />Mail has been sent to activate the
account";
                            mpeMsg.TargetControlID = "btnMsg";
                            mpeMsg.Show();
                            sendUserActivationMail();
                            break;
                        default:
                            lblMessage.Text = "Please try again";
                            break;
                    }


                }
            }
        }

        protected void sendUserActivationMail()
        {
            string emailStatus = "-1";
            string activationCode = Guid.NewGuid().ToString();
            SqlConnection sqlConn = new
SqlConnection(ConfigurationManager.ConnectionStrings["dASTISConnStr"].ToString());
            {
                using (SqlCommand cmd = new SqlCommand("sASTIS_UserActivation"))
                {
                    cmd.CommandType = CommandType.StoredProcedure;
                    cmd.Parameters.AddWithValue("@UserID", txtEmail.Text.Trim());
                    cmd.Parameters.AddWithValue("@ActivationCode", activationCode);
                    cmd.Parameters.AddWithValue("@action", "UserActivationEmail");
                    cmd.Connection = sqlConn;
                    try
                    {
                        sqlConn.Open();
                                                                                //Sending activation link to user to verify the email
address
                        emailStatus = Convert.ToString(cmd.ExecuteScalar());
                    }

                    catch (Exception ex)
                    {
                        HttpContext.Current.Response.Cookies["errorMessage"].Value = ex.Message;
                        Server.Transfer("Error.aspx");
                        throw ex;
                    }
                    finally
                    {
                        sqlConn.Close();
                        sqlConn.Dispose();
                    }
                }
                if (emailStatus.Contains("@"))
                {
                    string Subject = "ASTIS Account Activation Details";
                    string body = "Dear ASTIS User";
                    body += "<br /><br />Please click the following link to activate your account";
                    body += "<br /><a href = '" + 
Request.Url.AbsoluteUri.Replace("UserRegistration.aspx", "Login.aspx?ActivationCode=" + 
activationCode) + "'>Click here to activate your account.</a>";
                    body += body += "<br /><br />Thanks and Regards<br />ASTIS Team";
                    Common.sendEmail(emailStatus, null, Subject, body);

                    lblMsg.Text = "Email has been sent with account activation details";
                    mpeMsg.TargetControlID = "btnMsg";
                    mpeMsg.Show();

                }
                else
                {
                    lblMessage.Text = "Please type proper Email address given during registration";

                }
            }
        }

        protected void btnMsg_Click(object sender, EventArgs e)
        {
            Response.Redirect("Default.aspx",false);
        }


    }
}
```