Subbu Kandhaswamy

ID – 743189972

[skandhas@syr.edu](mailto:skandhas@syr.edu)

MSADS

Portfolio Milestone

Syracuse University, March 2022

Skandhas->Google shared drive link

<https://drive.google.com/drive/folders/1hrhs_qrq-07ghDqtCAC2Pc0IAKQMo3rK?usp=sharing>

skandhas->Git repo with course projects covered in this portfolio

<https://github.com/subbukandhaswamy/IST782-MSADS>

# Table of Contents

1. [Introduction](#Intro)
2. [Learning Objective](#KLO)
3. [IST 659 – Database Administration](#ist659)
   1. [Project description](#Project659)
   2. [Learning Goals and Achievements](#LGA659)
4. [IST 722 – Data Warehousing and Business Solutions](#IST722)
   1. [Project description](#project722)
   2. [Learning Goals and Achievements](#LGA722)
5. [IST 718 – Big Data Analytics](#ist718)
   1. [Project Description](#project718)
   2. [Learning Goals and Achievements](#LGA718)
6. [Conclusions](#fconclusion)
7. [References](#ref)

# Introduction

Applied Data science program in Syracuse enables students to apply appropriate analytical and computational methods to handle real-world problems effectively. Understanding the data enhances clarity and communicates technical information to targeted audience through great visualization. Methods to collect, analyze and develop insights using data from multitude of domains using various tools and techniques. Wide range of topics focusing data analysis, warehousing, business intelligence, & deep learning etc., and each area of study further break into tasks such as data extraction, exploration, transformation, preprocessing, modeling and share business insights for making decisions.

All the projects and presentations covered in this curriculum exemplify the skills that we developed throughout the course, and I will be covering a subset of courses (below) as part of this portfolio milestone presentation

* + - IST 659 – DATABASE ADMIN CONCEPTS AND DB MANAGEMENT
    - IST 722 – DATA WAREHOUSING
    - IST 651 – SCRIPTING FOR DATA ANALYTICS
    - IST 718 – BIG DATA ANALYTICS

# Key Learning Objective

* Broad overview of the major practice areas in data science
* Collect and organize data.
* Identify patterns in data using visualization techniques, statistical analysis, and data mining process.
* Develop alternative strategies based on the data.
* Develop a plan of action to implement the business decisions derived from the analyses.
* Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization
* Synthesize the ethical dimensions of data science practice.

# IST 659 DATABASE ADMINISTRATION

Database admin concepts covered in this course helped with identifying the requirements to design and implement data-oriented solutions using relational database management system. How we are collecting and managing data, where to store and how stored data should be organized for a startup business- (Four seasons).

I designed and implemented database environments for the business to manage their sales, customers, and inventory.

### **Technology Used**

* + Microsoft SQL
    - TSQL
  + Microsoft Access
    - CRUD Operations
  + Visio
    - ER/ Modeling

## **IST659 – Project description [ Four seasons]**

***Problem statement* -**

Four Seasons International (FSI) Texas based quilt company specializes in handmade quilts, the company owns manufacturing and selling these quilts to businesses around the country. FSI participates in major trade shows nationwide to advertise and expand their growth displaying new and top selling quilts during these shows.

The Customers are categorized based on their purchase volume agreed with FSI. And over the years FSI grown rapidly and management decided to maintain Customer information, Sales & Product.

FSI primarily wants to maintain the following in a database.

Customer information (Name, Address, Phone, Email, YTD sales, Sales Target). Customers should be identified by unique value.

Category (Name, Sales Target, Duration, ). Category should be identified by unique category code and determined by sales volume and duration as customer with FSI.

***Stage 1 - Conceptual Model for the problem*** the conceptual model representation for upcoming Logical design and building actual database.

**A screenshot of a cell phone

Description generated with high confidence**

***Stage 2 - Normalized Logical Model*** for FSI reviewed and confirmed before creating the physical database.

## **A screenshot of a cell phone Description generated with high confidence**

***State 3 – Using TSQL Statements created physical structure*** (Tables / constraints / views etc.) for the databases.

Diagram

Description automatically generated

Used DDL, DML, AND TCL statements to build schema, update /move data into master tables, applied key constraints.

And enabled interface for end users of FSI to update master data (customer/product/types/category, etc.,) details and for doing transactions (Order/Sales)

*Below snapshot is the order transaction UI for FSI*

Table

Description automatically generated

And also include standard reports to help the business to review sales performance, stock availability and more.

Calendar

Description automatically generated with low confidence

## Graphical user interface, table Description automatically generated

## Learning Goal and **Achievement**

The course enabled me to understand the data, follow the database admin concepts, build entity relationship diagrams for the business requirements, convert the design into physical structures and finally onboard the business to use a relational database management system to improve their business operations.

The project helped FSI business to move from their legacy applications/interfaces that involved both manually and semi-automated process into a complete system driven business to achieve more sales.

# IST 722 – DATA WAREHOUSING AND BUSINESS SOLUTIONS

This is the next level of data management for business, especially for large scale operations with more than one type of operational data sources involved that need to be organized, refined, and restructured for business solutions (built by BI applications) that is used for decision making purposes.

We learned various database constructs (Operational Data Source, Warehouse, Data Mart), understanding the difference between Kimball and Inman’s approach of data warehousing, integration of staged data using ETL, Designing and implementing data warehouse and BI components, understanding modeling techniques (High level and detailed dim modeling) and data governance concepts.

## Project Description (FudgeMart)

Fudgemart, Inc.’s subsidiary businesses have been collecting data for many years in the fudgemart\_v3 and fudgeflix\_v3 databases. However, the data for each of these subsidiaries were difficult to access, and there were currently no consequential Business Intelligence or Decision Support Systems to help Fudgemart make more data-driven decisions.

Our goal of this project was to create a data warehouse and a business intelligence (BI) platform for two different business processes of Fudgemart, Inc. Our BI solution allowed business users to easily access operational, financial, and supply chain data to make better data-driven decisions. The data warehouse is designed with an enterprise bus architecture, and contains staging, ETL, data marts, and a Power BI/Excel user interface.

### Business Requirements

* Access analytical data from Fudgemart / Fudgeflix users from centralized location
* Data to perform analysis to increase revenue, reduce cost, manage product, and cross functional opportunities between two entities.
* DW should feed data to generate reports to track sales, product movement (most selling to least selling), order fulfillment (shipping time, order accuracy)
* Our DW will integrate data from FudegeMart and FudgeFlix to provide Top sellers, Order frequency, Time to deliver (Shipdate – Orderdate)

### Functional Requirements

* Best-selling products.
* Most popular Title.
* Required stock to hold.
* Titles queued.
* How long does it take to deliver an order?
* Order frequency of product
* Cross customers with accounts in both businesses.
* Order frequency of products

### Business processes (related to the above questions)

* Order Fulfilment
* Product Inventory trend
* Sales
* Product review
* Sales coverage

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Charter** | | | |
| **Project Name** | IST 722 Group Project Assignment – Fudgeflix & Fudgemart | | |
| **Project Description** | Development of data warehouse and business intelligence solution for Fudgemart, Inc. | | |
| **Project Manager** |  | **Date Approved** | **5/9/21** |
| **Project Sponsor(s)** | **Humayun Khan** | **Signature** | **On file** |
| **Business Case** | | **Expected Goals/Deliverables** | |
| **A conformed data warehouse will allow Fudgemart to move towards an organization that can utilize all of its cross-functional data for efficient and effective business intelligence.** | | Goals:   * Identify functional requirements * 5 business processes to model * 1 integrated process in data warehouse * Meaningful business intelligence * Demo satisfies functional requirements. | |
|  | | Deliverables | |
| **Team Members** | | * **Project Document** * **High-level dimensional modeling worksheet** * **Detail level dimensional modeling worksheet** * **Initial ETL done in SSIS.** * **Data Warehousing** * **Business Intelligence** * **Presentation and Demo** | |
| **Name** | **Role** |
| **Keeley Ables** | **Dimensional Modeling** |
| **Subbu Kandhaswamy** |  |
| **Zachary Agrue** | **ETL Dev** |
| **Steven LaMarre** | **BI developer** |
| **Risks and Constraints** | | **Milestones** | |
| **Time** | **The team has a short time frame and limited availability for completion of the project.** | **5/9/21** | **Project document** |
| **Experience** | **The student team is learning the subject matter at the same time applying the concepts for the project.** | **5/16/21** | **High-level dimensional modeling worksheet** |
|  |  | **6/6/21** | **Detail-level dimensional modeling worksheet, SQL implementation, Initial ETL** |
|  |  | **6/20/21** | **Completed Project** |

### Business Process

|  |  |
| --- | --- |
| Facts | Dimensions |
| Ship date | Product (Flix & Mart) |
| Order Date | Customer (Flix & Mart) |
| Stock Units | Orders |
| Returned Date | Suppliers |
| Order Units | Shippers |

### Business Requirement

* Products with highest sales.
* Sales trend by category/by region/by period
* Product Reviews
* Conformed Customer dimension across businesses

### Schema Building

Diagram

Description automatically generated

### ETL - Extract Transform Load Operation

Graphical user interface

Description automatically generated

### **Technology Used**

* Microsoft SQL Server (SSAS/SSDT/CRUD)
  + Database Warehouse DB Management
  + ETL
* MS Excel
  + High Level & Detail Dim Modeling
* Microsoft Power BI
  + Business Solutions (Reporting)

## Learning Goals and Achievements

* Primary learning goal of this course is to understand the OLTP Sources (technically, the transaction data), bringing the data together into a common pool (Staging) and use Data transformation tools (SSDT) to refine and build data warehouse environments.
* Use refined DW data for OLAP applications to build reporting (Online Analytical Processing
* Build OLAP reports using MS Power-BI/Tableau) for effective decision making.

And as achievements

We combined more than one data source for fudgemart , staged, processed and built a DW environment that holds refreshed set of transaction and sales data, and built business intelligence solution for fudgemart and fudgeflix combined to track Subscription types, highest selling products, states with top subscriptions and highest profiting subscription types and products and other data points that helped the organizations to invest in critical operations for increased productivity.

Graphical user interface

Description automatically generated

These are the steps we followed as a team to achieve the key results for Fudgemart.

Graphical user interface

Description automatically generated with low confidence

# IST 718 Big Data Analytics

Want to cover this course as it’s one of the critical topics of data science explaining how to translate a business challenge into an analytics challenge. The world is moving into handling data in various formats when it comes to data analysis, most of the businesses in the cloud and on-premises interact with and handle unstructured/free form data in addition to relational data structures. And from an analytics point of view, building statistical models, and identifying insights are essential to lead into actionable results.

And we used Python and Spark for building big data analytics pipelines and for understanding the machine learning techniques (algorithms). And learned to explain how we can use advanced analytics to create competitive advantage.

## Project Description (Tour de France)

In this project, we want to predict what the future Tour de France will look like – specifically the pace and time of the next winner. Additionally, we hope to explore the data and describe the ideal cyclist. This is important because it allows us to understand which cyclist, we should invest in.

***About Tour de France*** *–*

*Every summer, the world’s top cyclists gather in France for a grueling race across the country. It’s a cycling competition held in France every year (sometimes neighboring countries) and takes place in 21 states with a course covering approximately 3,500 kilometers*

For this, we used Tour de France dataset directly from the official tour de France [website](https://www.letour.fr/en/history) that contains riders’ information for every year’s race from 1903 thru current.

Python libraries required for our analysis to pull, clean, and prepare data required for our analysis, then merged additional dataset to build performance information (winders dataset), and analyzed the clean & merged datasets using Matplotlib, seaborn, plotly libraries for visualization.

Chart

Description automatically generated

Text

Description automatically generated with low confidence

Graphical user interface, website

Description automatically generated

And we used prediction modeling (prophet model) using average pace and total seconds, to determine the winners for upcoming events.

Chart

Description automatically generated

## Learning Goals and Achievements

And as a team we were able to use historic datasets of the world’s largest cycling event to predict the winners effectively. Tour de France is a highly competitive race with many interested in understanding how to estimate a winner and thru our analysis we determined that

* Winners of Tour de France will win by continuing to beat the previous years' winner's paces and times
* Winners of Tour de France will likely be around 28 years old, 1.78 meters tall, and around 153 pounds

# Conclusions

The courses covered in this portfolio milestone deck, helped me to understand various forms of data, gave clarity to apply the methods and analyze data to present the insights and predict future outcomes that I can share to make business decisions and handle real-world problems.

With other courses of my Applied Data Science curriculum that includes

* Data Science introduction
* Data Analysis & Decision making
* Data & Business Analytics
* Management Service Principles
* Cloud Management &
* Information policy.

helped me gain the critical knowledge around cloud solutions, security & privacy, machine learning, mining, and business/management science techniques.

And I want to thank everyone were part of my course journey (respected Professors and fellow classmates) for sharing their knowledge, experience to help me learn various processes, tools, and techniques, and influenced me to the next level in my career.

# References

* **James, Gareth, Daniela Witten, Trevor Hastie,** and Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013. Available at <https://wwwbcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf>.
* **Goodfellow**, Ian, Yoshua Bengio, and Aaron Courville, Deep Learning (DL), MIT Press, 2016. Available at <http://www.deeplearningbook.org/>.
* **Michael Fudge** - School of Information Studies at Syracuse University - <https://github.com/mafudge/learn-databases/>
* **IST 722 - Data warehousing Techniques** – by KeyData1 <https://www.1keydata.com/datawarehousing/datawarehouse.html>
* **Ralph Kimball** <https://www.kimballgroup.com/category/articles-design-tips/>
* **Business Intelligence Best Practices** <http://www.bi-bestpractices.com/view-articles/4737>
* **Dimensional Modeling** <https://tombreur.wordpress.com/2017/04/30/the-past-and-future-of-dimensionalmodeling>
* **Architecture** <https://s3-us-west-2.amazonaws.com/syr-mac/prod/2017-0929+SYR-MSIM_0023+Data+Warehouse/Documents/WhichDWArchitectureIsMostSuccessful.pdf>