Master of Science Applied Data Science

Milestone Portfolio

Cherngywh Lee
334280765


Let's Go Orange!
Syracuse University
SYRACUSE

Instruction

M.S. in Applied Data Science

The Applied Data Science program at Syracuse university's School of Information Studies is a practitioner's degree – while the curriculum is founded upon firm theoretical underpinnings, the program is designed to be a professional program with a strong emphasis on the applications of data science to enterprise operations and processes, particularly in the areas of data capture, management, analysis and communication for decision.

Course list

IST 659 Data Admin Concepts & Database Management

IST 769 Advanced Database Management

IST 722 Data Warehouse

IST 687 Applied Data Science

IST 707 Applied Machine Learning

IST 664 Natural Language Processing

IST 772 Quantitative Reasoning in Data Science

IST 718 Big Data Analytics

IST 623 Introduction to Information Security

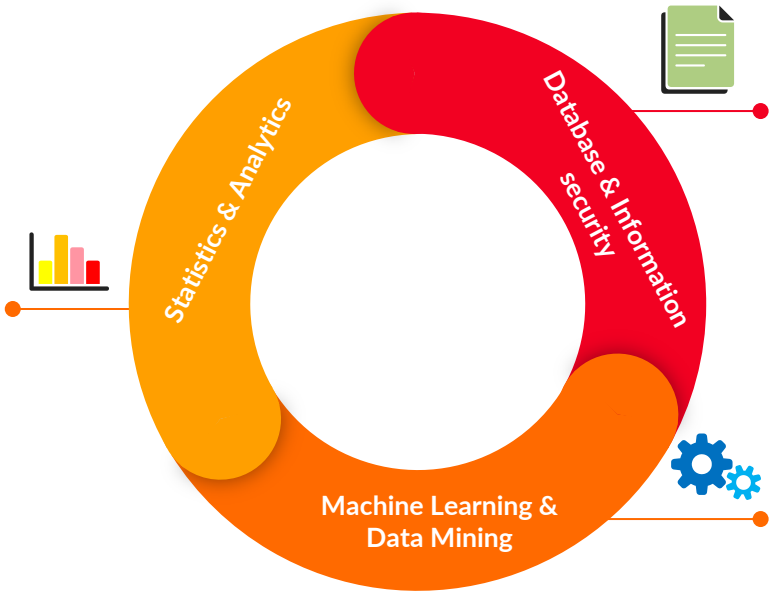
MBC 638 Data Analysis and Decision Making

SCM 651 Business Analytics

The elements of the program

Business Analytics

Analyze data with statistics theory like normal distribution, hypothesis test...etc. as well as tool like fishbone diagram to explore and make a data driven decision



Data Management

Well storage data with solid foundation of database building to various big data applications. Extract data to be data warehouse as the foundation of Business Analytics

Model & Prediction

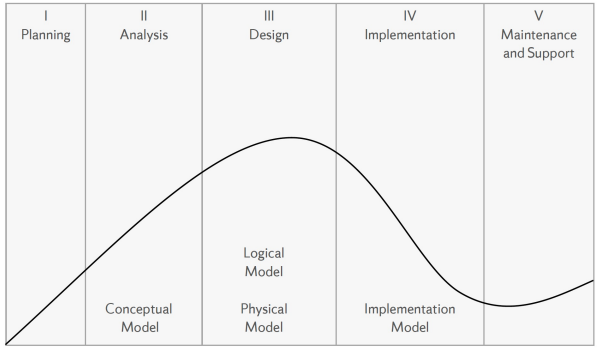
Find out useful variables, then applied various Machine Learning and Data Mining Model to interpret the data to predict the future

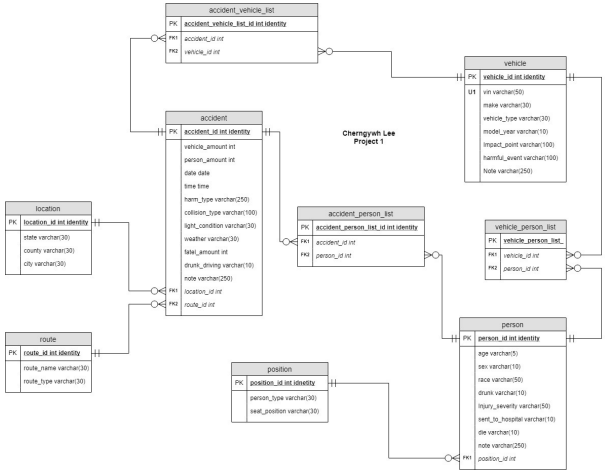
Database & Information Security

Database is the foundation of Data Science, like a Grocery store to a Chef or Ammunition to the Army

Data Admin Concepts & Database Management
Advanced Database Management
Introduction to Information Security
Data Warehouse

Build, Manage & Protect the Database





Examine data structures, file organizations, concepts, and principles of database management systems (DBMS) as well as data analysis, database design, data modeling, database management, and database implementation.

Follow the process below to build up a database from scratch:

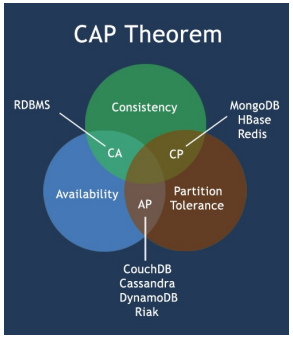
Conceptual → Logical → External → Internal → Physical

- Define the business rules and identification of the entities and relationships in the model. Set up attributes and keys and their types.
- Solve Problems by constructing database queries using SQL
- Recommend and justify strategies for managing data security, privacy, audit/control, fraud detection, backup and recovery with MS Access and SQL Server

Advanced concept and SQL skills with the relational data base model, such as transactions, concurrency control, performance and security. To build up a robust database can safely handle dynamic changes like temporal table.

- T-SQL Transaction Commands like Commit, Rollback to meet ACID (Atomic, Consistent, Isolated, Durable)
- Serializability and Locking Mechanisms to maintain concurrency control
- Define permissions and securable to ensure database security

Big Data Application

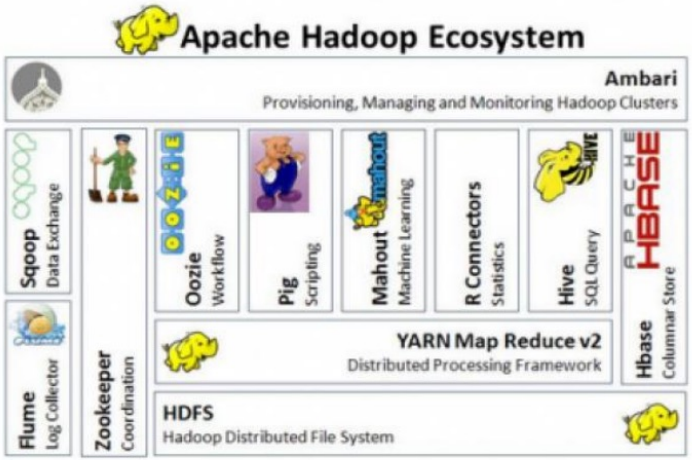


CAP theorem can be used on all kinds of database system

- CA – RDBMSs like Orcale, MySQL
- CP – Single-master systems like MongoDB, HBase and HDFS
- AP – Eventual consistency systems like CouchDB, Cassandra and Dynamo

No database can perfectly fulfill CAP theorem. However, we can freely conbine any of them based on our needs per Polyglot Persistence

- MongoDB – Product catalog and blog
- Redis – Shopping cart and web page cache
- Cassandra – Audit and activity logs
- MySQL – Order processing and payments
- Hadoop – Data analytics
- Neo4j – Social graph



Hadoop is the way from data to big data by the scale out design. Following HDFS

Data is stored as it is and chunked into blocks and distributed to data nodes across a cluster and managed by a name node.

- Data stored as they are, schema applied when data are read
- NameNode is the master node to determine and maintain how the blocks of data
- Data split into blocks and are distributed over physical DataNodes
- Default replication for failover

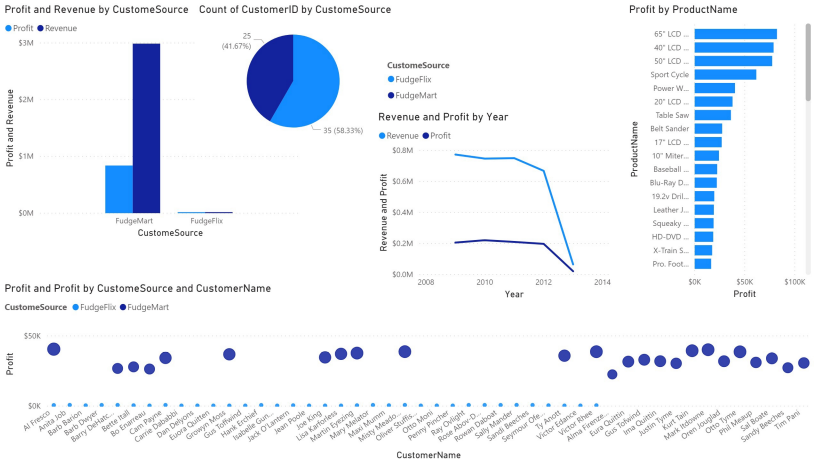

ANALYSIS OF THE SOURSE FILE
0,0587230-9
0,05
0,04
0,03
0,02
0,01
OBJECT: B/31/99
STATUS:
mode:
32/673
FILE MANAGER
FILE DIRECTORY
026/57
SYSTEM FILE
PROFILE
79933-2B
15%
IDENTIFICATION
PROFILES
SETTINGS
HISTORY
LOG
DOCUMENTS
MATCH
MATCH

Data Warehouse Project – Process

Data Warehouse is the foundation of BI (business intelligence).

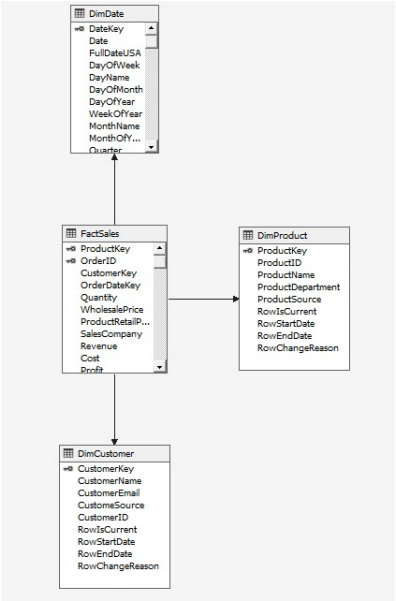
In this project, we performed ETL (Extract, Transform, Load) form OLTP database to BI.

- 2 OLTP data sets – Fudgemart & Fudgeflix. Fudgemart is and online retailer and Fudgefilx is a online DVD seller by mail/video.
- We have a few questions and improvement targets for these 2 data sets:
  - Who are the most frequent customers for Fudgemart?
  - What are the most profitable products for Fudgemart?
  - How do sales compare between Fudgeflix and Fudgemart? Profits and Revenues?
- Identify Business Process to Model, these represent the data mart in the data warehouse:
  - Business Process and Gain – one row per product order to figure out who is the most frequent customer
  - Dimensions – Sales, Date, Order, Customer
  - Facts – Order amount, product price
- Create a Star Schema, Data mart is implemented as a star schema in a RDBMS. It's also called ROLAP
- Define the attributes in each dimension from high-level to detail-level.
- Perform ETL to extract data from OLTA to Data Warehouse based on the ROLAP
- Conduct Data Analytics by BI tools with data from Data Warehouse



Data Warehouse Project - ETL

1. Star Schema



2. High Level Dimension

Group 3 *4/20/2020*

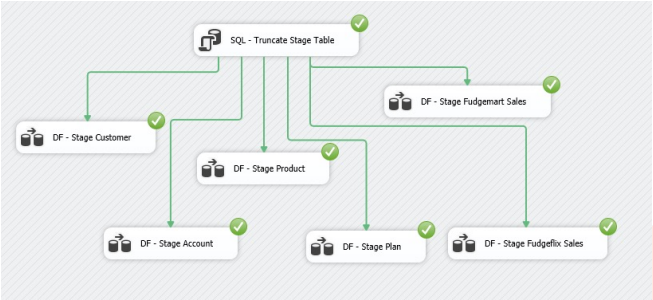
Business Process Name	Fact Table	Fact Grain Type	Granularity	Facts
most frequent customers	Frequent_customers	accumulating snapshot?	one row per orders	number of orders,
most ordered product per customer	top_product_per_customer	accumulating snapshot?	one row per order detail	number of products
rating of each title	title_rating	transaction	one row per title	Title, Rating, director
accounts per zipcode	accounts_zipcode	Periodic Snapshot	one row per zipcode	Name, Address, Zipcode, Account Plan
fudgeflix vs fudgemart spending	fudgeflix_fudgemart_spending	Accumulating snapshot	row for subscription or order	Customer, Accounts, total sales,
fudgeflix customer subscription tier, spending at fudgemart	fudgeflix_fudgemart_tier_spending	Accumulating snapshot	row for subscription amd order	customer, account, order
Fudgeflix customer plan change	fudgeflix_plan_change	slow changing dimesion	one row per plan	customer, account, order

3. Detailed Level Dimension

Table Name	DimProduct	
Table Type	Dimension	Home Page
Display Name	Product	
Database Schema	foodtraders	
Table Description	Products on an order (with supplier and category info)	
Comment	rollup into suppliers and categories	
Biz Filter Logic		
Size		
Generate Script?	Y	

Column Name	Display Name	Description
ProductKey	ProductKey	Surrogate primary key
ProductID	ProductID	Business key from source system (aka natural key)
ProductName	ProductName	Name of product
ProductDepartment	ProductDepartment	department of product
WholesalePrice	WholesalePrice	Cost of the product
VendorName	VendorName	Name of vendor
RowIsCurrent	Row Is Current	Is this the current row for this member (Y/N)?
RowStartDate	Row Start Date	When did this row become valid for this member?
RowEndDate	Row End Date	When did this row become invalid? (12/31/9999 if current row)
RowChangeReason	Row Change Reason	Why did the row change last?

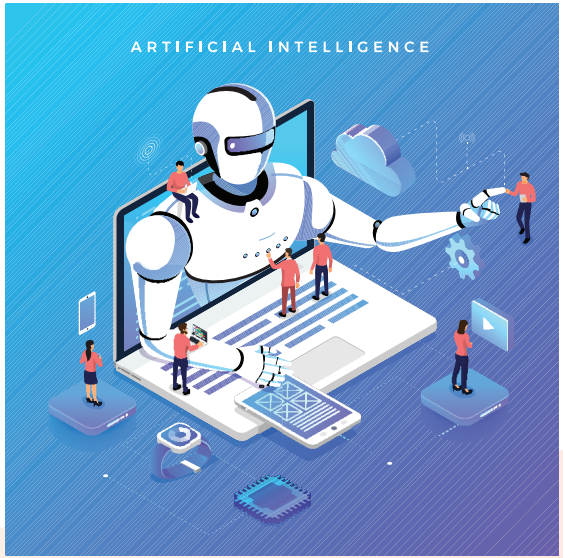
4. Load Data by SSIS



Machine Learning & Prediction

Predictive modeling solutions are a form of Data Mining by analyzing historical and generating model to predict the future

Applied Data Science
Applied Machine Learning
Natural Language Processing
Big Data Analytics



Model & Prediction

IST 687 Applied Data Science

The course introduces students to applied examples of data collection, processing, transformation, management.

IST 718 Big Data Analytics

A broad introduction to analytical processing tools and techniques for information professionals.

IST 707 Applied Machine Learning

Introduction to data mining techniques, familiarity with particular real-world applications, challenges involved in these applications, and future directions of the field.

IST 664 Natural Language Processing

This course is designed to develop and understanding of how natural language processing can process written text and produce a linguistic analysis that can be used in other application.

Statistics & Analytics

Explore data with statistical methods and technologies for analyzing historical data in order to gain new insight

Quantitative Reasoning in Data Science
Data Analysis and Decision Making
Business Analytics

Statistics & Analytics

IST 772 Quantitative Reasoning in Data Science

Statistical inference is the process by which we make sense of uncertainty in data, and this course focuses on establishin a thoughtful and

MBC 638 Data Analysis and Decision Making

This course will familiarize students with the assumptions underlying various statistical techniques and assist in identifying their appropriateness in

SCM 651 Business Analytics

This course is intended for the graduate student who is interested in developing a portfolio of skills in business analytics.