Kimball Lifecycle Proposal Template

CIS 9440 - Data Warehousing for Analytics
Final Project Milestone 1
Group Number - 2
Student(s) – David Freitag (working individually)

This Proposal is the beginning of your semester-long Final Project. The goal of the project is to develop a working Data Warehouse using a commercial database management system. Your project will use data from at least 2 sources, dimensionally model the data inside your Data Warehouse, and connect to a Business Intelligence application to produce valuable, actionable insights.

For motivation on project ideas, search for public datasets that interests you and your group. Then, think about how these datasets (maybe combined with other datasets) could address a problem or opportunity. Below are just a few (of many) public data sources:

1. Kaggle
2. NYC Open Data
3. Opendata.gov
4. Gapminder
5. Zillow
6. NOAA Climate Data
7. Google's public datasets

Data Warehouse Project Title:

- Environmental and Social Predictors of Chicago Crime

Motivation for Project idea:

- The ability to predict changes in the crime rate, as well as where those changes will occur, can lead to more effective and efficient deployment of police and security resources to discourage criminal activity.

Description of the issues or opportunities the project will address:

Issues presented by crime:

- The crime rate in Chicago results in significant financial cost to taxpayers as a result of the police department budget.
- Crime in Chicago has an economic cost on crime victims (i.e. theft of property).
- Crime in Chicago has a human cost on victims of crime, the families of those victims, and the families of the perpetrators.
- Crime in Chicago has a human and economic cost for perpetrators who find it challenging to join the job market after their interactions with the criminal justice system.

Opportunities that arise from this project:

- Reduce crime by more effectively deploying police resources to discourage criminal activity.
- Reduce costs of policing through a lower crime rate.
- Reduce the number of people interacting with the criminal justice system.
- Reduce recidivism of past offenders.
- Increased effectiveness of police activities.
- Increased efficiency of use of police resources.

Business Justification:

<u>High-level Business Initiative:</u>

- End product: intelligence dashboard to generate insights for data-driven decision making related to the factors that influence crime

<u>BI Sponsors and Stakeholders (who will own this project?)</u>

- BI Sponsor: CCCG (Chicago Crime Consulting Group) – a consulting group that is being contracted by the Chicago Police Department to provide intelligence and insights on crime and its predictors.
- End users: The Chicago Police Department, particularly employees involved in orchestrating the strategic deployment of officers in locations around the city.

<u>What's the Business Value?</u>

- This project will allow the Chicago Police Department to make informed decisions about where and how to deploy police resources, resulting in a significantly reduced crime rate and the costs associated with that crime.

How long will this take? How much will this cost?

Total cost to implement - $500,000
Ongoing costs - $300,000

- Storage
- DBMS
- Personnel

Technical Justification:

Which data sources do we already have for this project?

Dataset 1: Chicago Crime Data – 2001 to present
https://console.cloud.google.com/marketplace/product/city-of-chicago-public-data/chicago-crime

Dataset 2: US Inflation and Unemployment
https://console.cloud.google.com/marketplace/product/bls-public-data/cpi-unemployement

Dataset 3: Chicago Daily Summary Weather Data
https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/CITY:US170006/detail

What new data sources do we need (if any)?

Dataset 1: Education data for Chicago (i.e. graduation rates, attendance rates, dropout rates, etc.)

Dataset 2: Illinois subset of US Inflation and Unemployment data

Is the data we have conformed, consistent, and current? (data quality)

- Conformed: potentially – the uniting dimension for the data sources is time, and transformations in time will need to be done to the data to produce consistency across the source systems (i.e. datasets will need to be aligned and aggregated into daily, monthly, weekly, year timeframes).
- Consistent: yes – there is a single resource for the specific crime data that will act as the source of truth.
- Current: to a degree – crime data from 2001 to present is available, but “real time” data is not currently available.

What technical skills will we need to complete this project?

- Database Administration: setup and maintenance
- ETL skills to move data from data sources into the data warehouse
- Data Analysis skills to create the dashboard and interpret the data
- Domain knowledge to understand the meaning of information related to crime within the context of the city of Chicago

Will we need any new types of technologies?

- ETL pipelines will need to be coded
- A database will need to be set up and maintained
- A business intelligence dashboard will need to be created

Key Performance Indicators (KPI's) your Data Warehouse will display:

1. Total city crime rate for a given timeframe (day/week/month/year)
2. Total city arrests for a given timeframe (day/week/month/year)
3. Crime rate by location for a timeframe (district/ward/community area/beat)
4. Crime rate by IUCR code/FBI code for a timeframe (classification of type of crime)
5. Total arrests by location for a timeframe (district/ward/community area/beat)
6. Total arrests by code for a timeframe
7. Weather interaction (i.e. weather events like temperature/precipitation that could affect crime)
8. Unemployment interaction (i.e. employment information that could affect crime)
9. Education interaction (i.e. education trends like graduation rate that could affect crime)