Crime Trends Analysis for Home Office



**SHANTA MOHASHIN**

**KUBRICK GROUP**

**SEPTEMBER 2016**

TABLE OF CONTENTS

[Introduction 3](#_Toc461195369)

[The approach 4](#_Toc461195370)

[Requirements Gathering and Analysis 5](#_Toc461195371)

[Initial Requirements List 5](#_Toc461195372)

[Extended Requirements List 5](#_Toc461195373)

[Design 6](#_Toc461195374)

[Data Sourcing 6](#_Toc461195375)

[Data Cleansing 6](#_Toc461195376)

[Data Loading 6](#_Toc461195377)

[Technical Design 7](#_Toc461195378)

[Entity Relationship Diagram 7](#_Toc461195379)

[Graphical User Interface Prototypes 7](#_Toc461195380)

[Implementation 8](#_Toc461195381)

[Storage Tier 8](#_Toc461195382)

[Examples of Storage Components Implementation 8](#_Toc461195383)

[Visualisation Tier 10](#_Toc461195384)

[Examples of Visualisation Components Implementation 10](#_Toc461195385)

[Testing 11](#_Toc461195386)

[Evaluation 12](#_Toc461195387)

[Issues and caveats 12](#_Toc461195388)

**TABLE OF FIGURES**

[Figure 1 - Tailored Solution Development Lifecycle 4](#_Toc461140950)

[Figure 2 - Entity Relationship Diagram 7](#_Toc461140951)

[Figure 3 - Initial UX Design Wireframe 7](#_Toc461140952)

# Introduction

Several studies had shown a high incline in violent crimes in the United Kingdom over last few years. According to [Telegraph](http://www.telegraph.co.uk/news/uknews/crime/12112024/Violent-crime-jumps-27-in-new-figures.html) violent crime in the UK had risen 27% between year 2014 and 2015.

Home Office (HO) have been increasingly concerned about the trend. A clear picture with specific data and evidence on the trend would clarify the existence of the problem and its worsening state. Which may potentially be used to develop a strategic plan on how to tackle the situation.

Kubrick consultant Shanta Mohashin undertook the task of data analysis and implementation of insights into the current and past crime and population trends which would directly support decision making for resource allocation and prove helpful in potential predictions of the future tendencies.

# The approach

Systematic and agile approach was selected. The method was based on both data lifecycle as well as proven software development lifecycle to ensure compliance with customer requirements and completeness of the technical solution.

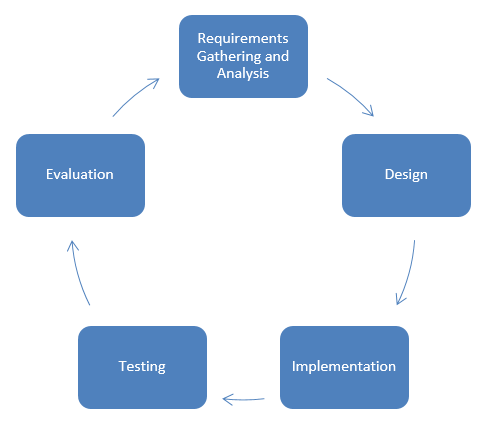


Figure - Tailored Solution Development Lifecycle

Requirements gathering is an initial phase of the project that focuses on customer needs for the technical solution. Requirements are then turned into a technical design of the solution that alongside of data sourcing, validation and cleansing forms the design phase. Using the design specification, the solution is implemented and appropriate data is being loaded which is followed by further implementation of the solution. Continuous testing forms integral part of the effort and is used to perform evaluation against initial customer requirements. The lifecycle and its phases can and will be reiterated until required outcome is reached.

# Requirements Gathering and Analysis

Due to concerning increase in violent crime, the Home Office has worked with Kubrick Group to source and utilise the existing data to provide valuable insights into the crime and its potential causes or provide any valuable information that may improve resource allocation for the police forces.

From the initial project brief, it was clear that Home Office wanted to specifically focus on the areas such as violent crime and its occurrence variables such trends for location and population of the selected locations.

In order to answer the questions posed by the customer a clear and easy to use technical solution would be devised that would be based on the initial requirements listed further in this section and also would aim to provide additional insights that would prove to be beneficial.

## Initial Requirements List

Below are the approved and reviewed requirements based on the initial project brief from the customer:

* Present an easy to understand information to support decision-making activities
* Provide evidence of worsening trends in violent crime
* Identify the data sources and validate reliability
* Source the appropriate datasets
* Cleanse the data and prepare for presentation
* Establish the link between the population and the crime
* Visualise relationship and trends between the time and location of the crime
* Design and implement the technical solution to enable the further data analysis as necessary
* Provide an ability to store large datasets using Microsoft SQL Server
* Present analysis using Microsoft Excel
* Evaluate the findings, solution as well as the project

## Extended Requirements List

Based on the iterative requirements and problem analysis throughout the lifecycle, Kubrick Group identified the following requirements to be applicable and valuable:

* Identify relationships between several location levels including country, region and the area
* Present the overall crime trend over the period of 4 years taking into account the area and the population of the area
* Extend analysis of the population and presentation of age groups at levels including country, region and area
* Provide insight into the gender ratio and the crime occurring at the country level
* Focus design on scalability to provide an ability to load additional data as required
* Reduce data processing times where possible
* Maximise maintainability and reusability of solution components

# Design

This section will describe the technical and data consideration and the necessary details to understand the technical solution provided by Kubrick Group as a response to customer requirements.

## Data Sourcing

Data sourcing was a vital part of the project ensuring that the appropriate decisions were being made based upon valid and accurate data from a trustworthy source. Throughout the design phase multiple data sources were identified and noted then basic checks of validation were performed to ensure data compliance.

The valid data sources that have been approved for use were:

* Police.gov database
* National Statistics database

The sources were then additionally cross-referenced to further improve the quality of validation.

The sourcing also required a focus on the initial requirements to ensure that all of the questions posed by the customer were appropriately addressed and the appropriate Data Governance is in place. Based on initial findings the limits of the dataset were established and have been placed to be 5 years for both crime and population data from June 2011 to June 2016 inclusive.

## Data Cleansing

Initially, the focus was on sourcing the relevant data, which later led to additional data cleansing activities to ensure only relevant data was used to support the implementation of the system according to the customer requirements.

Selected data cleansing activities included:

* Planning of cleansing based on initial data quality and structure
* Selection of cleansing tool for data partitions, MS Excel vs T-SQL
* Normalising the data into 1st, 2nd and 3rd normal form
* Identification of data relationships and dependencies
* Removing duplication as well as incomplete data

## Data Loading

A significant amount of effort was placed on data cleansing that ensured that only the relevant data was included in the implementation. In addition, a number of decisions were made to improve the effectiveness and efficiency of the solution, these included:

* Selection of appropriate approach and tool for data loading depending on dataset
* Automation of large volume dataset imports using appropriate tools such as Microsft SQL Server Integration Services (SSIS)
* Selection of the appropriate field types and size to maximise performance

## Technical Design

Industry standard Data Warehousing techniques are key for effective database structure and analysis. Inman’s Star and Snowflake Schema were compared side by side, and Snowflake approach was chosen as it focuses on factors such as intensive normalisation, in addition to well-defined logical structure and multi-level relationships.

### Entity Relationship Diagram

The following diagram was produced based on the initial identification of the data structure based upon the findings from the data sourcing and cleansing activities. The structure was then revised and normalised to ensure high performance and low levels of data duplication.

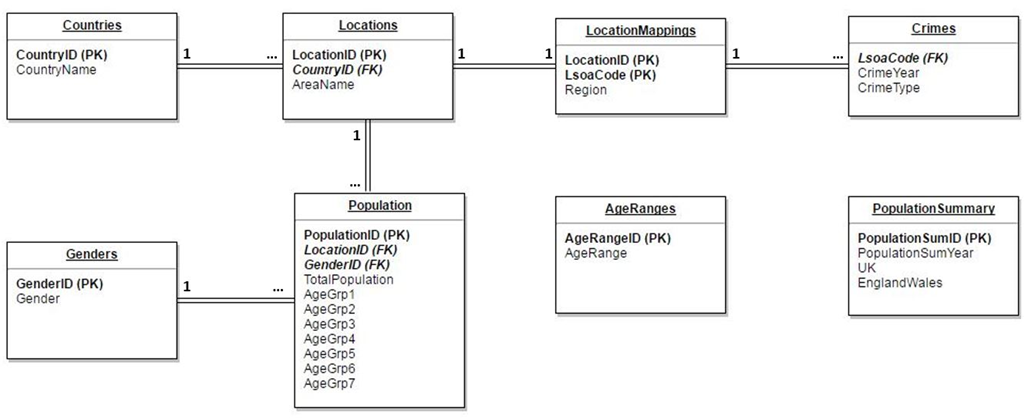


Figure - Entity Relationship Diagram

### Graphical User Interface Prototypes

In addition to the data storage design, some focus was placed on the design of the data representation and an initial design wireframe was presented to the customer and revised throughout.

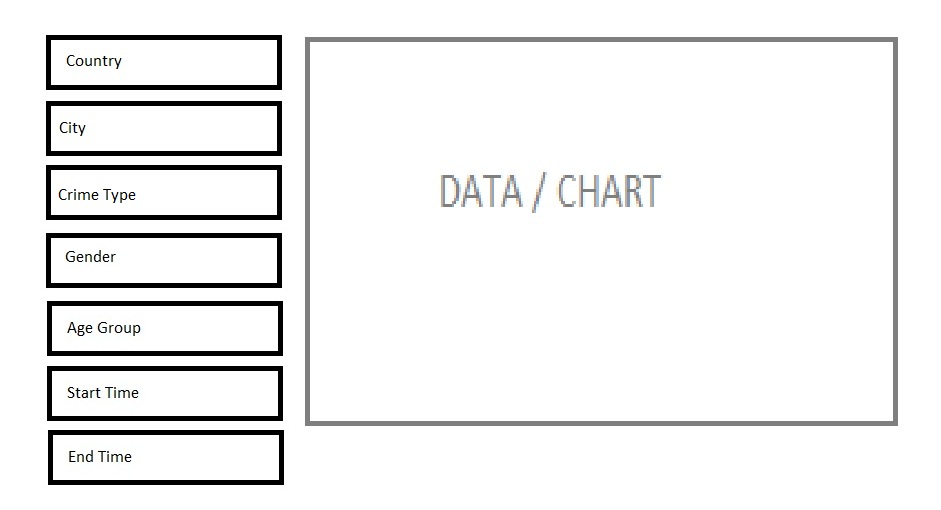


Figure - Initial UX Design Wireframe

# Implementation

Design of the solution resulted in inception of the implementation phase. Implementation was focused on actual development of the solution components that fulfilled the customer requirements.

The development was focused on two main components:

* Storage Tier
* Visualisation Tier

## Storage Tier

Development of the storage tier was the first logical step of the implementation as it is a backbone of the data visualisation and without the appropriate data and structure visualisation would not be possible.

In order to comply with the customer requirements the Microsoft SQL Server was used as a tool to develop the storage tier and store all the necessary data. The initial development process included the following steps:

1. Create basic database structure (tables)
2. Validation of the basic database structure
3. Load all the necessary data
4. Validate that the data has been loaded successfully
5. Create data views to allow data visualisation

In addition to the standard development tasks additional measures have been taken to improve the efficiency and effectiveness of the solution and the development effort itself; these included:

1. Performance analysis and tuning including creation of keys and indexes, data compression
2. Query performance and redesign as well as and data integrity analysis with ATOM principle
3. Creation of a daily maintenance plan and a rolling data backup
4. Creation of a scheduled task to rebuild database indexes

### Examples of Storage Components Implementation

Below images represent selected technical aspects of solution component implementation:

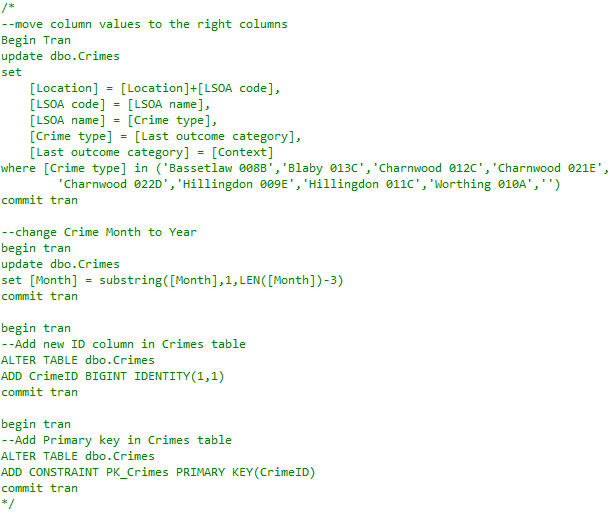


Figure - Data Cleasing using T-SQL

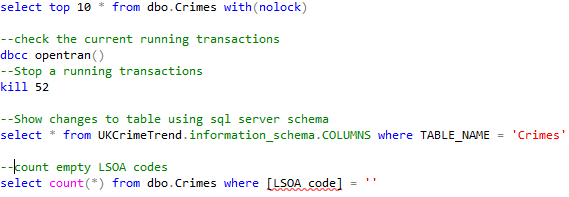


Figure - Clansing Validation

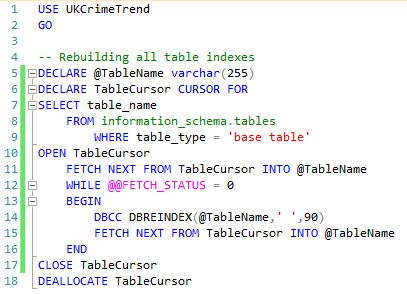


Figure - Rebuilding All Table Indexes

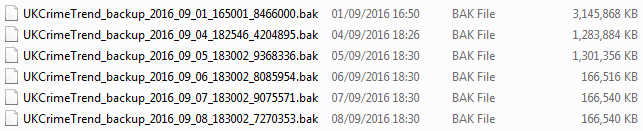


Figure - Automated Daily Backup and File Size Reduction

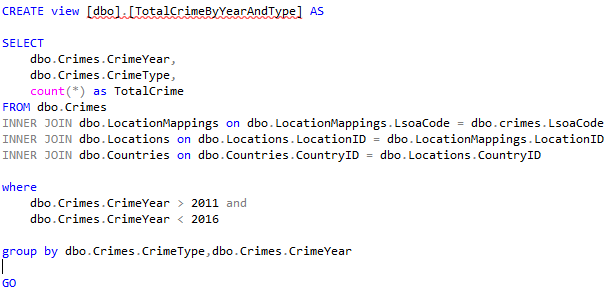


Figure - View: Total Crime by Year and Type

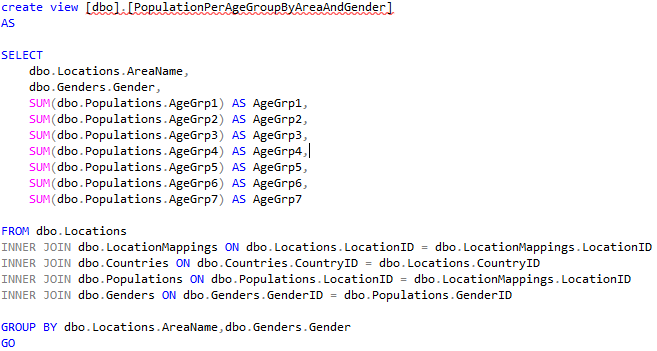


Figure - View: Population Per Age Group By Area And Gender

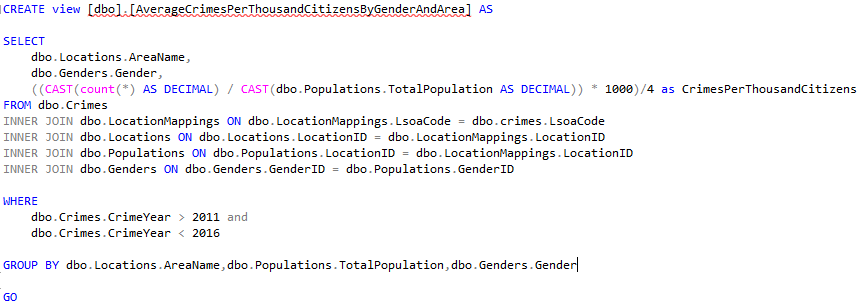


Figure - View: Average Crimes Per Thousand Citizens By Gender And Area

## Visualisation Tier

Visualisation Tier was developed in accoradance with best practices and based on the customer requirements gathered at the beginning of the project and used to provide the in-depth definition of the solution. This component of the solution was developed using Microsoft Excel as per customer specification.

Several tools and techniques were uses to fulfil the customer requirements and these included:

* Data connections to obtain data from the storage tier
* PivotTables and PivotCharts used for data representation and visualisation
* Visual Basic for Applications for the UX and Data automation
* Built-in UX components to enhance and improve usability

The above components were included in the iterative solution design that resulted in multiple versions being developed to ensure full compliance with customer needs.

### Example of Visualisation Components Implementation

[Crime Trends Analysis for Home Office Dashboard](Crime%20Trends%20Analysis%20for%20Home%20Office.xlsm) represents visualisation aspects of solution component implementation.

# Testing

Iterative nature of the development process implied necessity of modular approach to solution testing. Based on the standard Agile software development method, a short sprint was established to deliver a piece of functionality catering a single customer requirement that would then be tested against that requirement. The chosen approach ensured that the released functionality met the requirement and complied with the standards.

Four levels of testing were performed for the solution:

1. Solution component testing (low-level component module)
2. Solution component integration
3. Solution tier testing
4. Full system testing

Test plans were created based on the initial requirements and then functional requirements of the solution and were used to ensure quality and compliance.

# Evaluation

The testing results had shown that the functionality of the technical solution provided meets the customer requirements. The solution is also fully functional and ready for deployment.

In addition to basic customer requirements an iterative approach had proven to be reliable and in alignment with the industry standards the following was achieved:

* Extension of the basic customer requirements
* Use of the best practices for naming conventions and focus on code quality within areas such as readability, maintenance and re-usability.
* Focus on performance and storage optimisation to reduce the overheads

## Issues and caveats

Below table lists all the issues and resolutions faced throughout the development of the project that may have impacted the data and the outputs as well as the overall solution design and implementation.

|  |  |
| --- | --- |
| **Issue** | **Resolution** |
| Lack of data for violent crime for the whole period selected. | Included a broader picture for the timeline where the data was not available. |
| No crime data was avaialable for Scotland and Northern Ireland. | Scotland and Northern Ireland datasets were removed from the scope of the solution. |
| Crime location data reported using geolocation. | Instead of geolocation area name was used to provide an in-depth insight into the location in addition to higher level region and country. |
| Performance for crime data queries. | Due to very large volumes of the data associated with the crime statistics the timeline of reports was reduced to 4 years. |
| Extensive effort for crime data cleansing | Due to very large volume of the data associated with the crime statistics data loading automation was implemented using SSIS and cleansing was performed using SQL Server rather then Microsoft Excel. |