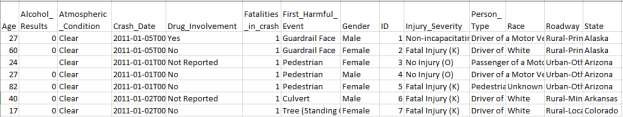
Data Warehouse design using sql server 2012

Data: A **.csv** file containing the information about the car accidents reported in 2011 in the United States, provided by The National Highway Traffic Safety Administration (NHTSA).

Data Mart Design Description:

The document will describe a data mart design that can be used to analyze car accidents reported in 2011 in the United States data set provided by The National Highway Traffic Safety Administration (NHTSA.) The design for the data mart was created and replicated in Microsoft SQL Server 2012. SQL Server was used to implement the analytical design in order to perform the analysis. The analytical design is based off of the flat data structure included in the Excel file provided by NHTSA.gov. It examines a variety of car accidents recorded by The Department of Transportation in the Unites States in 2011. The data mart will allow end users to extract data from Excel or CSV file, and load it to any reporting tool, such as Tableau.

This document will examine our analytical tables and relationships based on structure seen in our Dart Mart. A sample is as follows.

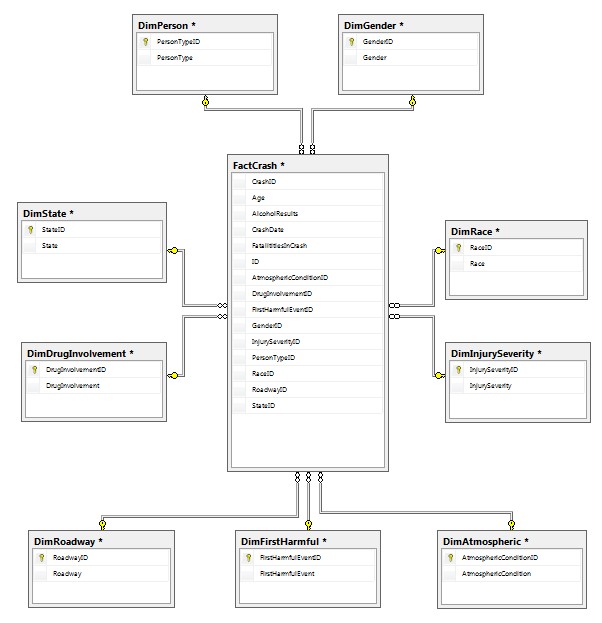
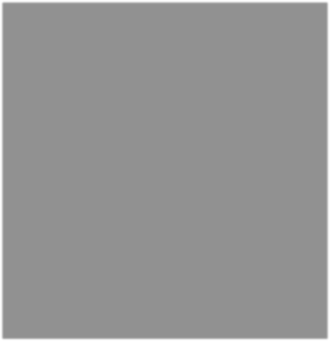


In our data set showing above taken from the source USA 2011 Car crash data Excel file, the analytical design will consider the measurable columns to be “ID”, “Age”, “AlcoholResults”, “CrashDate”,

“FatalitiesInCrash” are listed “FACTs”. The “AtmosphericCondition”, “DrugInvolvement”,

“FirstHarmfulEvent”, “Gender”, “InjurySeverity”, “PersonType”, “Race” “Roadway” , “State” are our listed columns that we used to created “Dimensions”. Which lead us to set of relational table objects that are related by a set of foreign to primary key relationships will be used together to form an analytical STAR schema. Our document will illustrate the STAR schema and describe the data mart design. The 2011 car accidents in the United States data set was extracted, transformed, and loaded (ETL.) The process is needed to populate the data mart with the transaction data. The ETL schema design uses a custom built SQL Server Integration Services (SSIS) packages develop with SQL Server Data Tools (SSDT) in Visual Studio.

Data Mart Schema Diagram:



Data Mart Meta Data:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **FACT TABLE: FactCrash** |  |
| **Column Name** | **Data Type** | **Description** | **Key** |
| CrashID | Int | Identity column ID | None |
| Age | Int | Person’s Age involved in the accident | None |
| AlcoholResults | Decimal(3, 2) | Person’s Alcohol level at the time of the accident | None |
| CrashDate | date | Date and time of the accident | None |
| FatalititiesInCrash | Int | Number of deceased at the accident | None |
| ID | Int | Unique Identification number for each accident | None |
| AtmosphericConditionID | Int | Unique identification number for weather condition | Foreign |
| DrugInvolvementID | Int | Unique Identification number for drug involvement | Foreign |
| FirstHarmfulEventID | Int | Unique Identification number for each harmful event | Foreign |
| GenderID | Int | Unique Identification number for person’s gender | Foreign |
| InjurySeverityID | Int | Unique Identification number for each injury | Foreign |
| PersonTypeID | Int | Unique Identification number for person’s type | Foreign |
| RaceID | Int | Unique Identification number for person’s race | Foreign |
| RoadwayID | Int | Unique Identification number for the accident location | Foreign |
| StateID | Int | Unique Identification number for each state | Foreign |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **DIMENSION TABLE: DimState** |  |
| **Column Name** | **Data Type** | **Description** | **Key** |
| StateID | Int | Unique Identification number for each state | Primary |
| State | Varchar(30) | Location of state where the incident took place | None |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DIMENSION TABLE: DimRoadway** | |  |
| **Column Name** | **Data Type** | **Description** | **Key** |
| RoadwayID | Int | Unique Identification number for the accident location | Primary |
| Roadway | Varchar(150) | Accident location | None |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DIMENSION TABLE: DimRace** | |  |
| **Column Name** | **Data Type** | **Description** | **Key** |
| RaceID | Int | Unique Identification number for person’s race | Primary |
| Race | Varchar(300) | Person’s race | None |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DIMENSION TABLE: DimPerson** | |  |
| **Column Name** | **Data Type** | **Description** | **Key** |
| PersonTypeID | Int | Unique Identification number for person’s type | Primary |
| PersonType | Varchar(100) | Person’s role in the accident | None |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DIMENSION TABLE: DimInjurySeverity** | |  |
| **Column Name** | **Data Type** | **Description** | **Key** |
| InjurySeverityID | Int | Unique Identification number for each injury | Primary |
| InjurySeverity | Varchar(50) | Type of injury | None |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DIMENSION TABLE: DimGender** | |  |
| **Column Name** | **Data Type** | **Description** | **Key** |
| GenderID | Int | Unique Identification number for person’s gender | Primary |
| Gender | Varchar(15) | Type of gender | None |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DIMENSION TABLE: DimFirstHarmful** | |  |
| **Column Name** | **Data Type** | **Description** | **Key** |
| FirstHarmfulEventID | Int | Unique Identification number for each harmful event | Primary |
| FirstHarmfulEvent | Varchar(200) | Type of harmful event | None |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DIMENSION TABLE: DimDrugInvolvement** | |  |
| **Column Name** | **Data Type** | **Description** | **Key** |
| DrugInvolvementID | Int | Unique Identification number for drug involvement | Primary |
| DrugInvolvement | Varchar(25) | Type of drug involvement (if applicable) | None |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DIMENSION TABLE: DimAtmospheric** | |  |
| **Column Name** | **Data Type** | **Description** | **Key** |
| AtmosphericConditionID | Int | Unique identification number for weather condition | Primary |
| AtmosphericCondition | Varchar(50) | Type of weather condition at the time of the accident | None |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **STAGING TABLE: Staging** |  |
| **Column Name** | **Data Type** | **Description** | **Key** |
| Age | Varchar(800) | Person’s Age involved in the accident | None |
| AlcoholResults | Varchar(800) | Person’s Alcohol level at the time of the accident | None |
| AtmosphericConditionID | Int | Unique identification number for weather condition | None |
| AtmpsphericCondition | Varchar(800) | Type of weather condition at the time of the accident | None |
| CrashDate | Varchar(800) | Date and time of the accident | None |
| DrugInvolvementID | Int | Unique Identification number for drug involvement | None |
| DrugInvolvement | Varchar(800) | Type of drug involvement (if applicable) | None |
| FatalitiesInCrash | Varchar(800) | Number of deceased at the accident | None |
| FirstHarmfulEventID | Int | Unique Identification number for each harmful event | None |
| FirstHarmfulEvent | Varchar(800) | Type of harmful event | None |
| GenderID | Int | Unique Identification number for person’s gender | None |
| Gender | Varchar(800) | Type of gender | None |
| ID | Varchar(800) | Unique Identification number for each accident | None |
| InjurySeverityID | Int | Unique Identification number for each injury | None |
| InjurySeverity | Varchar(800) | Type of injury | None |
| PersonTypeID | Int | Unique Identification number for person’s type | None |
| PersonType | Varchar(800) | Person’s role in the accident | None |
| RaceID | Int | Unique Identification number for person’s race | None |
| Race | Varchar(800) | Person’s race | None |
| RoadwayID | Int | Unique Identification number for the accident location | None |
| Roadway | Varchar(800) | Accident location | None |
| StateID | Int | Unique Identification number for each state | None |
| State | Varchar(800) | Location of state where the incident took place | None |

Data Mart ETL Description:

The Extraction, Transformation, and Load (ETL) process is an important data integration method designated to fulfill strategic decision support and overall analytical design to improve user respond time. The ETL process allows reading data from a database, convert and put that data into a data warehouse.

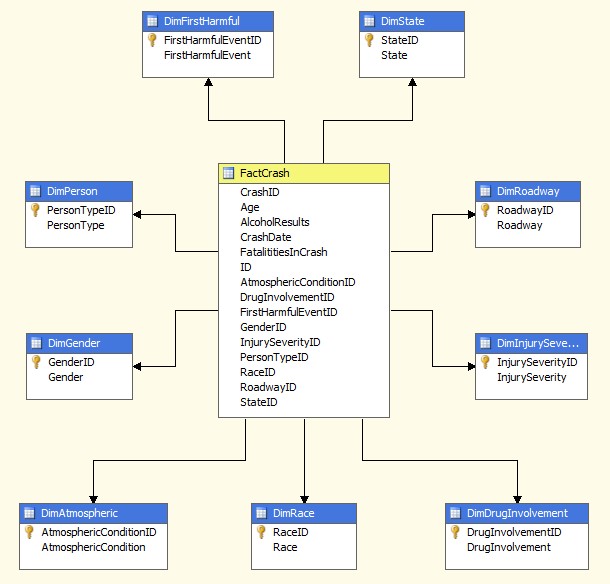
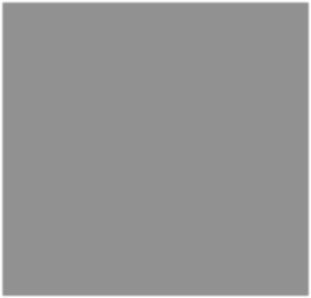
The process has been designed with SQL Server Data Tools (SSDT) using the SQL Server Integration Services (SSIS) Template in Visual Studio to provide an advantage of rapid redesign and implementation. First, design changes can be implemented to the table structure. Then, the SSIS package would be modified appropriately so that the data can be easily truncated and reloaded into all Schemas.   
SSIS Package Details:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CONTROL FLOW NAME** | **TYPE OF TASK** | **DATA FLOW NAME** | | **DESCRIPTION** |
| Get Crash Data from  Source CSV and Load to  Staging | Data Flow Task | 1.  2. | Flat File Source  OLE DB Destination | Obtains and connects the .csv crash data file and extracts the data into  staging table |
| Load State Dimension | Data Flow Task | 1.  2.  3. | OLE DB Source  Data Conversion  OLE DB Destination | Get States from staging table, converts the datatype’s length and load  States into the State  Dimension table |
| Update StateID into Staging Table | Execute SQL Task | N/A  (Execute SQL Server Task  Editor) | | Executes SQL Statement  UPDATE query to assign  StateID into Staging Table |
| Load Roadway Dimension | Data Flow Task | 1. OLE DB Source 2. Data Conversion 3. OLE DB Destination | | Get Roadway from staging table, converts the datatype’s length and load  Roadway into the Roadway  Dimension table |
| Update RoadwayID into Staging Table | Execute SQL Task | N/A  (Execute SQL Server Task  Editor) | | Executes SQL Statement  UPDATE query to assign  RoadwayID into Staging  Table |
| Load Race Dimension | Data Flow Task | 1. OLE DB Source 2. Data Conversion 3. OLE DB Destination | | Get Race from staging table, converts the datatype’s length and load  Race into the Race  Dimension table |
| Update RaceID into Staging Table | Execute SQL Task | N/A  (Execute SQL Server Task  Editor) | | Executes SQL Statement  UPDATE query to assign  RaceID into Staging Table |

|  |  |  |  |
| --- | --- | --- | --- |
| Load PersonType Dimension | Data Flow Task | 1. OLE DB Source 2. Data Conversion 3. OLE DB Destination | Get Person Types from staging table, converts the datatype’s length and load PersonType into the PersonType Dimension table |
| Update PersonTypeID into Staging Table | Execute SQL Task | N/A  (Execute SQL Server Task  Editor) | Executes SQL Statement  UPDATE query to assign  PersonTypeID into Staging  Table |
| Load InjurySeverity Dimension | Data Flow Task | 1. OLE DB Source 2. Data Conversion 3. OLE DB Destination | Get Injury Severity from staging table, converts the datatype’s length and load Injury Severity into the Injury Severity Dimension table |
| Update InjurySeverityID into Staging Table | Execute SQL Task | N/A  (Execute SQL Server Task  Editor) | Executes SQL Statement  UPDATE query to assign  InjurySeverityID into  Staging Table |
| Load Gender Dimension | Data Flow Task | 1. OLE DB Source 2. Data Conversion 3. OLE DB Destination | Get Gender from staging table, converts the datatype’s length and load Gender into the Gender’s  Dimension table |
| Update GenderID into Staging Table | Execute SQL Task | N/A  (Execute SQL Server Task  Editor) | Executes SQL Statement  UPDATE query to assign  GenderID into Staging  Table |
| Load FirstHarmfulEvent Dimension | Data Flow Task | 1. OLE DB Source 2. Data Conversion 3. OLE DB Destination | Get First Harmful Event from staging table, converts the datatype’s length and load FirstHarmfulEvent into  the First Harmful Dimension table |
| Update FirstHarmfulID into Staging Table | Execute SQL Task | N/A  (Execute SQL Server Task  Editor) | Executes SQL Statement  UPDATE query to assign  FirstHarmfulID into Staging  Table |
| Load DrugInvolvement Dimension | Data Flow Task | 1. OLE DB Source 2. Data Conversion 3. OLE DB Destination | Get Drug Involvement Event from staging table, converts the datatype’s length and load  DrugInvolvementID into the Drug Involvement  Dimension table |
| Update  DrugInvolvementID into  Staging Table | Execute SQL Task | N/A  (Execute SQL Server Task  Editor) | Executes SQL Statement  UPDATE query to assign  DrugInvolvementID into  Staging Table |
| Load Atmospheric Dimension | Data Flow Task | 1. OLE DB Source 2. Data Conversion 3. OLE DB Destination | Get Atmospheric Condition from staging table, converts the datatype’s length and load AtmosphericConditionID into the Atmospheric  Dimension table |
| Update AtmosphericID into Staging Table | Execute SQL Task | N/A  (Execute SQL Server Task  Editor) | Executes SQL Statement  UPDATE query to assign  AtmosphericID into Staging  Table |
| Update AlcoholResults | Execute SQL Task | N/A  (Execute SQL Server Task  Editor) | Executes SQL Statement  UPDATE query to assign  AlcoholResults into Staging  Table |
| Load Fact Table | Data Flow Task | 1. OLE DB Source 2. OLE DB Destination | Get Staging Data and load  Fact table (FactCrash) Data |

Data Mart Cube:

The following diagram illustrates the Cube Structure:

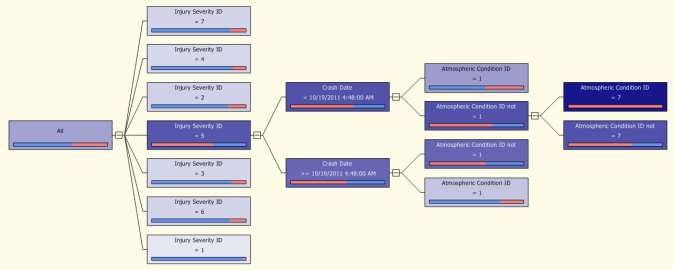
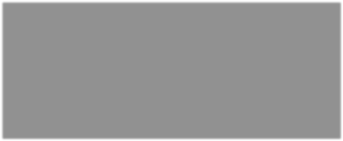
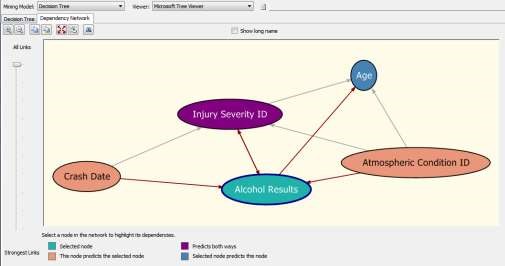
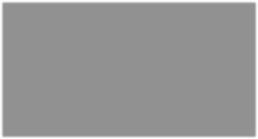
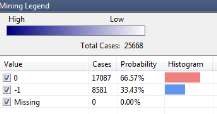


The “AlcoholResultPerformance” KPI is defined with the following criteria:

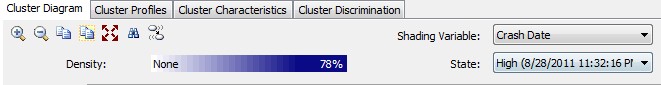
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Associated**  **Measure**  **Group** | **Value Expression** | **Goal Expression** | **Status Indictor** | **Status Expressions** | **Trend**  **Indicator** |
| FactCrash | [Measures].[Alcohol  Results] | "More than 0.08  alcohol level is illegal" | Gauge | CASE  WHEN  [Measures].[Alcohol  Results] > 0.08  THEN 1  ELSE 0 END | Standard Arrow |

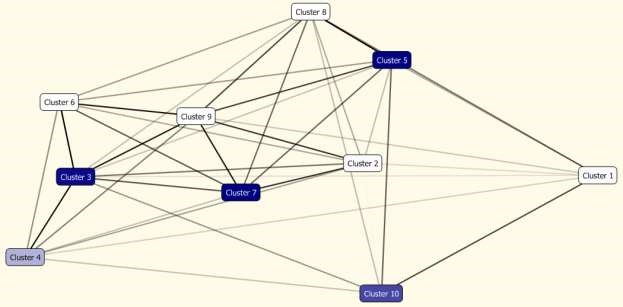
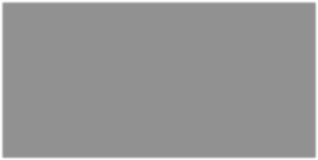
Data Mining Results:

The ***Decision Tree Algorithm*** was used to evaluate the age, alcohol results, and injury severity in car accidents. The results showed the most predictable pattern in Alcohol Results tree selection when Injury severity ID is equal to 5 (fatal injury) along with a crash date on or before October 19, 2011 at 4:48 AM with variant Atmospheric conditions were a total of 17,087 or a 66% cases reported.



The ***Clustering Algorithm*** was used to evaluate car accidents in specific dates. The results of this algorithm show that the most predictable pattern can be seen in Cluster 3, 5, and 7 where Crash Date is high between August and December 2011 around 10:00 - 11:00 PM with a total of 78%





Data Mart Reports:

The below report example describes the total amount of fatalities occurred in a car accident shown per State during 2011.

# Fatalities in Crash per State 2011

