Tecnologias de Processamento de Dados | **Fase 1**

Grupo 17 / 2019.20 / Faculdade de Ciências da Universidade de Lisboa

João Ferreira / 54121 / 6 horas de trabalho
Mariana Santos / 49525 / 6 horas de trabalho
Tiago Pereira / 49174 / 6 horas de trabalho
Vasco Leitão / 49455 / 6 horas de trabalho

"US-Accidents: A Countrywide Traffic Accident Dataset"

1. Describe your original data set. Identify the most important information. Identify missing or incomplete data, and identify possible strategies to use (or discard) them.

O dataset escolhido, extraído do Kaggle [1][2], apresenta cerca de 2,6 milhões de acidentes ocorridos entre fevereiro de 2016 e agosto de 2019 em 49 dos estados que compõem os Estados Unidos da América (EUA).

O conjunto de dados original é composto por 49 atributos que poderão ser classificados em 4 categorias: atributos temporais (data, hora), geoespaciais (cidade, rua, coordenadas GPS, etc), meteorológicos (temperatura, pluviosidade, etc) e relativos a sinalização rodoviária (sinalização vertical, existência de rotunda, etc), tendo estes sido recolhidos de diferentes fontes, tais como entidades estatais, forças de segurança, câmeras de vídeo-vigilância ou outros sensores posicionados em vias de comunicação, e extraídos através de diferentes APIs.

De um ponto de vista de substância, a escolha deste dataset está relacionada com a atualidade da temática e com a convicção de que um data warehouse construído sobre este conjunto de dados poderá contribuir decisivamente na melhoria da qualidade das análises produzidas, seja na perspectiva da identificação de locais sensíveis a sinistralidade rodoviária ou até no impacto de factores atmosféricos nessa mesma sinistralidade.

De um ponto de vista de forma, o desafio associado à construção de um data warehouse para acomodar um volume de dados relativamente extenso, e cujo potencial de crescimento é bastante generoso, contribuiu para a nossa escolha. Paralelamente, a presença no dataset de atributos do tipo hierárquico (temporais e geo espaciais), que poderão resultar em dimensões altamente pormenorizadas e diferenciadas, capazes de acomodar operações interessantes de roll-up, slice, dice e pivoting, foram também aspectos relevantes para esta escolha.

Em termos de processamento dos dados presentes no dataset original, serão feitas algumas alterações, seja de remoção ou adição de colunas, discretização de valores e tratamento de dados nulos/vazios.

Tratamento de dados nulos/incompletos

- Definição de valores para dados meteorológicos nulos será feita com base na média de valores em cada atributo, ao nível da cidade, na respetiva semana em que ocorreu o acidente;
- Os códigos TMC (*Traffic Message Channel*) que sejam nulos serão identificados com um número fora do domínio dos códigos existentes e não vão ser relevantes para possíveis interrogações. Estas linhas não serão removidas porque existem cerca de 700 mil registos, uma quantidade elevada que não pode ser descartada;
- As linhas que têm valores relativos ao *twilight* e período do dia vazios serão removidas (93 linhas);
- As linhas que não têm informação relativa à cidade na qual o acidente ocorre também serão removidas (83 linhas), dada a extensão do dataset e a dificuldade em obter a cidade com os dados existentes no mesmo;
- Relativamente ao fuso horário, cerca de 3000 registos não contêm informações sobre o mesmo. O preenchimento desses dados será feito através da correspondência entre a cidade onde o acidente ocorreu e o fuso horário no qual se encontra, analisando outros registos.

2. Describe what other data sets (if any) are going to be used for the construction of the data warehouse. How will they complement the existing information.

A recolha de dados externos tem como objetivo complementar alguns atributos específicos, e não funcionar como algo que contribua de forma global através da recolha de múltiplas colunas de uma vez. Sendo assim, encontramos dados úteis relativamente a:

- *Urban-Rural Classification (2013)* em cada município (*county*) dos EUA [3]. Estes dados dividem-se por seis classes dado o nível de urbanização do município em questão;
- O evento associado para cada *Traffic Message Channel* [4]. Dado que os TMC são enviados para aplicações de GPS ou dispositivos de navegação de modo a transmitirem informação relevante sobre o acidente, queremos, por exemplo, entender se há prevalência de determinados tipos de acidentes em determinados locais;
- Calendários das fases da lua entre 2016 e 2019 [5]. Depois de algumas leituras, queremos perceber se a fase da lua, aliada a outros fatores de tráfego, influencia o número de acidentes e a severidade dos mesmos.
- Limite máximo de velocidade por estado, nos EUA [6]. Dadas as várias diferenças de limites de velocidade dentro de cada estado, e estando a lidar com 49 estados no total, achamos que o processamento dos dados em cada estado seria dispendioso.
- Dados de consumo de álcool da população por estado (2017), nos EUA [7]. Após a discretização dos dados, queremos estabelecer uma visão geral dos dados dos acidentes em diferentes grupos de nível de consumo.
- Dados sobre a quantidade de automóveis registados em cada estado (2019), nos EUA [8]. Após a discretização dos dados, queremos estabelecer uma visão geral dos dados dos acidentes relativamente à quantidade de automóveis registrados no estado onde ocorrem.

3. Identify and characterize the facts table. Identify and characterize the grain of each element. If more than one fact table is present identify the grain for each of them.

Cada facto na nossa tabela representa a informação relativa à ocorrência de um determinado acidente. A tabela de factos é composta pelas respectivas chaves primárias de cada tabela de

dimensões mais duas medidas, severidade e duração do acidente. A granularidade temporal (horário local e padrão) é relativo ao minuto e a local é relativa à rua.

4. Identify all the possible dimensions with the data set and the accessory data found. Here we just need to identify the Dimensions, not define their structure.

Identificamos como *accessory data* qualquer dado que é criado através de fontes externas. Após a análise dos dados considerámos que o *data warehouse* se poderia repartir em seis dimensões, sendo elas:

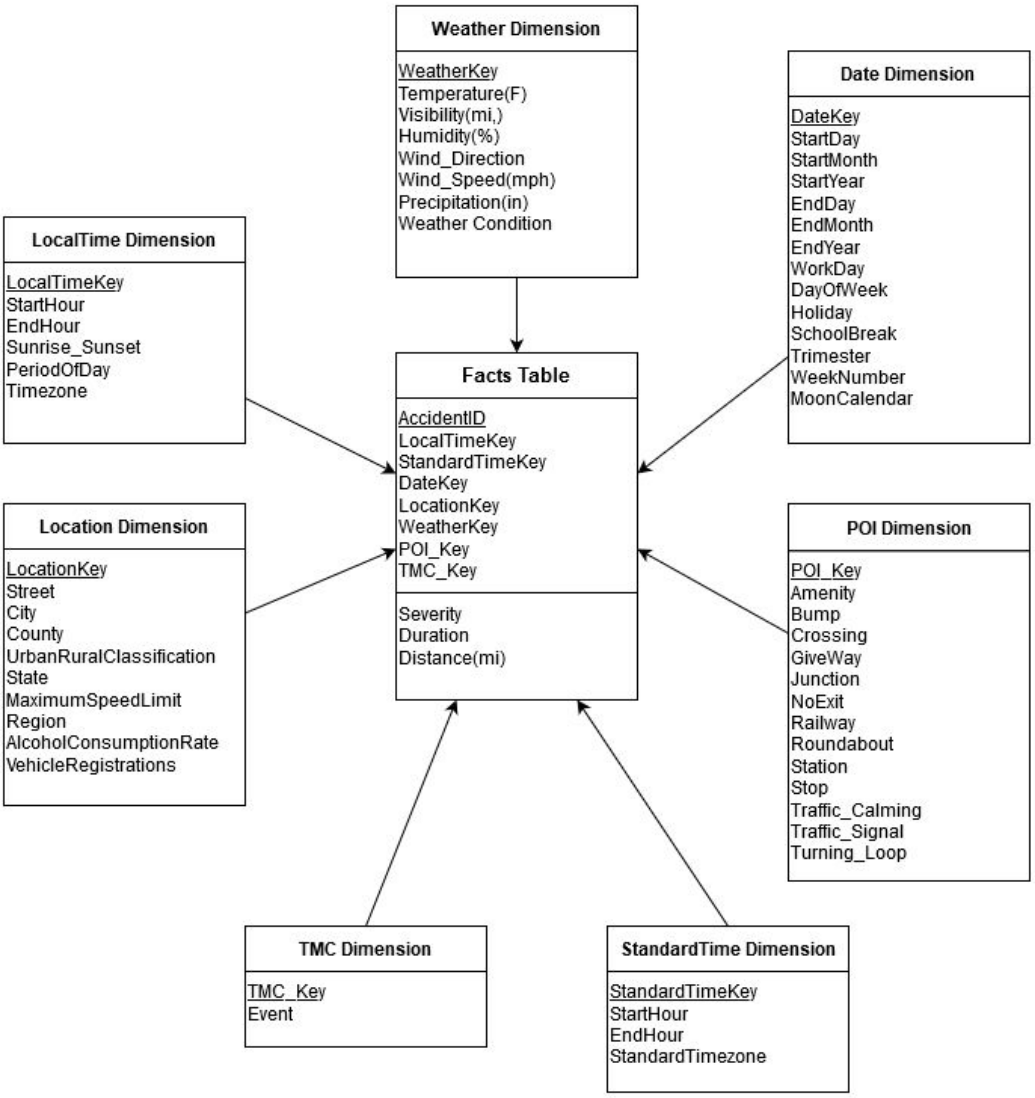
- **LocalTime:** dados temporais de caráter local.
- **StandardTime:** dados temporais relativos de caráter global, com base num determinado fuso horário padrão.
    - *Holiday (accessory data):* identificação de feriados especiais, como o Natal, *Thanksgiving*, ou o Dia da Independência, entre outros;
    - *School break (accessory data):* identificação de período escolar ou de pausa escolar;
    - *Moon Calendar (accessory data):* identificação da fase lunar.
- **Location:** dados relativos à localização do acidente, identificando outros atributos que sejam relevantes à análise do acidente e que dependem da localização.
    - *UrbanRuralClassification (accessory data)*: carácter urbano/rural da localização;
    - *MaximumSpeedLimit (accessory data)*: caracterização do limite de velocidade;
    - *AlcoholConsumptionRate (accessory data)*: caracterização do consumo de álcool *per capita* no estado em que ocorre o acidente;
    - *VehicleRegistrations (acessory data)*: caracterização do número de veículos registrados no estado em que ocorre o acidente.
- **Weather:** dados meteorológicos, devidamente discretizados.
- **POI:** dados relativos a sinalização rodoviária que sejam relevantes para a análise do acidente.
- **TMC:** códigos de Traffic Message Channel e respectivos eventos.

5. Identify how many data marts are going to be used and what types of business processes might use them. If more than one data mart is present, define the Bus matrix and identify their position in a feasibility/value matrix.

Numa primeira abordagem, e no caso do nosso *dataset*, um *datamart* parece-nos suficiente para inquirir os dados de modo relevante e pertinente. O objectivo desta análise, da perspectiva do lado do negócio é fazer aferições que possam materializar-se em medidas que promovam a segurança rodoviária. Para isto pretendemos extrair dos dados padrões e fenómenos que explicam a causalidade dos acidentes.

Dado apenas termos um *datamart* não faz sentido a criação de um *bus matrix*.

6. Define the star schema of the proposed data warehouse.



a) For each dimension identify and describe its columns
Nota: A chave primária de cada tabela é assinalada pelo atributo sublinhado.

LocalTime Dimension

- **LocalTimeKey (inteiro)**: Identificador único do registo;
- **StartHour (time) :** hora de início do acidente (hh:mm:ss);
- **EndHour (time):** hora de fim do acidente (hh:mm:ss);
- **PeriodOfDay (string):** período do dia (hora de ponta, almoço, *dusk*, etc) em que ocorreu o acidente; O atributo *dusk* é obtido através dos atributos *Civil_Twilight, Nautical_Twilight, Astronomical_Twilight* do *dataset* original, e representa as fases de crepúsculo.
- **Timezone (string):** fuso horário a que corresponde a data e hora do acidente (US/Eastern, US/Pacific, etc).

StandardTime Dimension

- **StandardTimeKey (inteiro):** identificador único do registo;
- **StartHour (Time):** hora de início do acidente ajustado a um fuso horário comum (hh:mm:ss);
- **EndHour (Time):** hora de fim do acidente ajustado a um fuso horário comum (hh:mm:ss);

- **StandardTimeZone (string):** fuso horário de referência (p. Ex.: GMT, ou um dos referenciados no atributo *Timezone*).

Date Dimension

- **DateKey (inteiro):** identificador único do registo;
- **StartDay (inteiro):** dia de início do acidente;
- **StartMonth (string):** mês de início do acidente;
- **StartYear (inteiro):** ano de início do acidente;
- **EndDay (inteiro):** dia correspondente ao fim do acidente;
- **EndMonth (string):** mês correspondente ao fim do acidente;
- **EndYear (inteiro):** ano correspondente ao fim do acidente;
- **WorkDay (string):** identificação se a data de início corresponde a um dia de trabalho, fim-de-semana ou feriado nacional;
- **DayOfWeek (string):** dia da semana a que corresponde a data de início do sinistro;
- **Holiday (string):** data de início do sinistro corresponde a feriado e a que feriado;
- **SchoolBreak (string):** data de início do sinistro corresponde a pausa escolar;
- **Trimester (inteiro):** trimestre a que corresponde a data de início do sinistro;
- **WeekNumber (inteiro):** Identificador do número da semana do ano a que corresponde a data de início do sinistro;
- **MoonCalendar (string):** Identificador da fase lunar.

Weather Dimension

- **WeatherKey (inteiro):** identificador único do registo;
- **Temperature (float):** temperatura do ar, em graus Fahrenheit (F), com referência a uma hora e estação de medição próximas ao acidente;
- **Visibility (float):** intensidade do vento, em milhas por hora (mph), com referência a uma hora e estação de medição próximas ao acidente;
- **Humidity (float):** humidade do ar, em percentagem (%), com referência a uma hora e estação de medição próximas ao acidente;
- **Wind_Direction (string):** direção do vento com referência a uma hora e estação de medição próximas do acidente;
- **Wind_Speed (float):** intensidade do vento, em milhas por hora (mph), com referência a uma hora e estação de medição próximas ao acidente;
- **Precipitation (float):** quantidade de precipitação, em polegadas (in), com referência a uma hora e estação de medição próximas ao acidente;
- **WeatherCondition (string):** descrição das condições meteorológicas na altura do acidente (neve, chuva, etc).

Location Dimension

- **LocationKey (inteiro):** identificador único do registo;
- **Street (string):** nome da via/rua/estrada em que ocorreu o acidente;
- **County (string):** munícipio de ocorrência do acidente;
- **UrbanRuralClassification (string):** classificação do munícipio quanto ao seu carácter urbano ou rural;
- **City (string):** cidade de ocorrência do acidente;
- **State (string):** estado onde ocorreu o acidente;
- **MaximumSpeedLimit (inteiro):** velocidade máxima de circulação na via, no estado, em milhas por hora (mph);
- **Region (string):** região dos EUA onde ocorreu o acidente, definida através do *U.S. Census Bureau*;

- **AlcoholConsumptionRate (float):** taxa de consumo de álcool per capita, por estado;
- **VehicleRegistrations (inteiro):** número de veículos registrados no estado de ocorrência do sinistro.

POI Dimension

- **POI_key (inteiro)**: identificador único do registo;
- **Amenity (booleano):** existência de estabelecimentos de lazer na proximidade;
- **Bump (booleano):** existência de lomba na proximidade;
- **Crossing (booleano)**: existência de cruzamento na proximidade;
- **GiveWay (booleano):** existência de sinalização de cedência de prioridade na proximidade;
- **Junction (booleano):** existência de junção de vias de trânsito na proximidade;
- **NoExit (booleano):** existência de sinalização de via sem saída na proximidade;
- **RailWay (booleano):** existência de via de caminho-de-ferro na proximidade;
- **Roundabout (booleano):** existência de rotunda na proximidade;
- **Station (booleano):** existência de estação de transportes públicos na proximidade;
- **Stop (booleano):** existência de sinalização de stop na proximidade;
- **Traffic_calming (booleano):** existência de mecanismos de redução de tráfego/velocidade na proximidade;
- **Traffic_Signal (booleano):** existência de semáforos na proximidade;
- **Turning_Loop (booleano):** existência de possibilidade de inversão de marcha na proximidade.

TMC Dimension

- **TMC_Key (inteiro)**: identificador único do código de Traffic Message Channel (TMC);
- **Event (string):** evento correspondente ao código TMC respectivo.

Colunas originais removidas que não serão convertidas em outras e sem qualquer uso para o warehouse:

- Source
- Weather_Timestamp
- Airport_Code
- Side
- Wind_Chill
- Pressure
- ID

b) Identify and characterize the measures of the facts table

Identificam-se duas medidas na tabela de factos: severidade (*severity*), duração (*duration*) e distância (*distance*).
A severidade corresponde a um valor entre 1 e 4 que pretende qualificar, numa escala crescente de gravidade, a ocorrência quanto ao impacto que teve no normal fluir do tráfego.
Por sua vez a duração corresponde à duração efetiva da ocorrência resultando da diferença entre data e hora de início e data e hora de término.
Finalmente, a distância reflete a extensão de via pública afetada pelo acidente, em milhas (*mi*).

c) Estimate its size and growth over time (in number of lines in tables)

Considerando que em média ocorrem anualmente cerca de 6 milhões de acidentes rodoviários nos Estados Unidos é de esperar que a tabela de factos cresça em média cerca de 500 mil linhas por

mês, embora este valor tenha seguramente alguma variação sazonal e variação associada a ciclos económicos. O crescimento da tabela de factos vai também, consequentemente, levar ao aumento em todas as dimensões exceto a dimensão referente às TMC (esta que funciona de forma estática).

7. Identify the main usage of the data system and propose different analysis queries for the system. It's just necessary to write them down in plain language.

As interrogações procuram fazer a análise relativamente às medidas associadas à tabela de factos, como a frequência de acidentes, a sua severidade e duração, e perceber como estas se relacionam com as dimensões do *data warehouse*.

- Há algum período do dia onde ocorrem mais acidentes? Esses períodos correspondem a horas de maior movimento, como *rush hours*?
- É possível verificar um aumento significativo de acidentes nas pausas escolares e épocas festivas?
- Em acidentes que ocorrem durante a noite, existe correlação entre as fases da lua e o número de sinistros, e a severidade dos mesmos?
- Assumindo que os diferentes níveis de urbanismo nos municípios afetam a quantidade de tráfego nos mesmos, esse aumento também vai ocorrer no número de acidentes?
- A presença de sinalização/*points of interest* tem que impacto nas medidas registadas na tabela de factos? Quais são os *points of interest* que têm maior impacto na redução do número de acidentes?
- Qual a influência que os diferentes limites de velocidade têm na severidade dos acidentes? Serão acidentes mais severos associados a estradas com limites maiores?
- O encadeamento por luz solar é um factor de causa de acidentes. Qual o impacto do crepúsculo no possível encadeamento do condutores?
- Algumas condições meteorológicas estão associadas a 'piso escorregadio'. Essas condições aumentam de forma significativa a severidade e a distância dos acidentes?
- A sinistralidade aparenta um carácter sazonal. Quais são os períodos do ano (trimestres e meses) onde se verifica o maior número de sinistros?

Referências

1. Moosavi, S. (2020), *A Countrywide Traffic Accident Dataset (2016 - 2019),* Kaggle, https://www.kaggle.com/sobhanmoosavi/us-accidents;
2. Moosavi, S. (2019), *A Countrywide Traffic Accident Dataset (2016 - 2019),* https://smoosavi.org/datasets/us_accidents;
3. CDC/National Center for Health Statistics (2017), *NCHS Urban-Rural Classification Scheme for Counties*, https://www.cdc.gov/nchs/data_access/urban_rural.htm;
4. OpenStreetMap (2020), *TMC - Event Code List,* https://wiki.openstreetmap.org/wiki/TMC/Event_Code_List;
5. Redelmeier et al. (2017), *The full moon and motorcycle related mortality: population based double control study,* https://doi.org/10.1136/bmj.j5367*;*
6. Wikipedia (2020), *Speed limits in the United States,* https://en.wikipedia.org/wiki/Speed_limits_in_the_United_States#Overview;
7. Elflein, J. (2019), *Per capita alcohol consumption of all beverages in the U.S. by state - 2017,* https://www.statista.com/statistics/442848/per-capita-alcohol-consumption-of-all-beverages-in-the-us-by-state/;
8. U.S. Office of Highway Policy Information (2018), *State Motor-Vehicle Registrations - 2018, https://www.fhwa.dot.gov/policyinformation/statistics/2018/mv1.cfm*.