Processing UK Prescribing Data Using SparkSQL : CS784 Project Proposal

Derek Paulsen

April 1, 2019

1 Problem Statement

There is a large amount of freely available data from various governments which provide detailed information about important aspects of life (e.g. healthcare, crime, etc.). Due to the size and scope of this open data, producing useful insight from this data requires scaling both simple data exploration and anaylsis as well as advanced statistical procedures for extracting meaningful relationships between observational data. Effectively addressing the challenges has the potential to greatly inform a wide variety of decisions, from how best to provide healthcare to identifying possible risk factors for economic downturn.

2 Problems Encountered

Attempting address the above problem poses a few particular challenges,

2.1 Data Size

Many of the datasets of interest are large (e.g. greater than a GB in size). While dealing with a couple GB of data is possible on a single machine, depending on the relationship between data sets, 2GB of data can easily turn into 20GB after being joined. This problem is typically solved by using a distributed execution framework (e.g. MapReduce, TensorFlow, Spark), for our purposes interactive data querying is vital for efficiently exploring the data hence we opted to use SparkSQL [1].

2.2 Finding Linkable Datasets

While there is a massive amount of publicly available data from many sources, it is very difficult to find datasets that make sense to link. That is, finding two datasets which are from the same time period, concerned with the same population and may have some interesting interaction between them is far from trivial. This problem was addressed with a lot of googling and patience.

2.3 Joining Datasets

Even when datasets make sense to join it is frequently the case the there isn't a simple way to do so. In particular, we found that many of the datasets which include geographic location have different granularity. For example, census data is typically the most fine grained geographic data (by postcode), but statistics like mortality from diseases are much coarser grain and less

frequently updated. We applied extensive data transformations to get the data at the same granularity, making it possible to use standard SQL joins on the relations.

2.4 Finding Interactions

After finding datasets and figuring out how to link them, there still is the problem of finding meaningful/interesting interactions and insight. The number of possible directions for where to look for interactions in datasets is a massive search space when there are just 3 or 4 datasets. This problem is probably the most difficult to tackle since it is so data dependant. For our case we found it effective to first interactively query and get summary statistics of the data that we have to get a sense for possible directions and trends in the data. After doing this we ran many queries with cheap to compute statistical measures, in particular we looked at the pearson's correlation between two variables to as kind of a filter for directions to do finer analysis. Even when filtering with cheap statistical measures, the computation time is far to expensive to explore all possible interactions, hence we also had to use some notion of common sense as to what to look at. For example, higher crime might lead to greater stress which might lead to more heart medication being prescribed on average.

3 Progress

Currently we have been able to accomplish the following,

3.1 Data Acquisition

We have been able acquire the following dataset from the UK government's open data repositories, [5] [4]

- Census Data - Both official and intermediate estimates
- Indexes of Deprivation - measure of societal issues in a given area (e.g. crime)
- Mortality Rates - For given classes of diseases, including heart disease, cancer, stroke
- Drug Prescribing - From general practices, broken down by individual prescription written

The intersection of the data only applies to England and Wales (which covers roughly 85% of the total population of the UK), and accounts for 2011 to 2015.

3.2 Data Preprocessing

Each data set had a different granularity for the time and geographic size. To address this we did extensive preprocessing to normalize everything such that each data point was for the outer postcode and year. This allow for easy joining between the datasets using standard SQL queries at the cost of having to use coarser grain information.

3.3 Initial Profiling and Exploration

We instantiated a cluster of three machines on CloudLab and interactively queried the data using Spark [1] for basic statistics and checked that our normalization went as intended, that is, we were able to join the relations as we intended. After the verification and initial exploration/profiling, we ran basic queries looking for correlations between rates of drugs being prescribed and the

Índexes of Deprivation. For example we looked for a correlation between the rate of heart disease medication being prescribed and the crime rates in the area (which we found surprisingly high R^2 values).

4 Future Plans

Currently we are planning on doing finer grain analysis of the data for where we have found interesting correlations so far. In particular, we want to apply the methods from ZaliQL [3] such as CEM using SparkSQL to draw more robust conclusions from our data.

In addition to this we plan on creating various plots of the data including,

- Drug prescribing trends as a time series (i.e. basic line plots)
- Heat map of drug prescribing rates for various classes of drugs
- Heat maps for mortality rates to contrast with drug prescribing rates

5 Related Work

5.1 ZaliQL [3]

ZaliQL is a framework for doing causal inference on large observational datasets using PostgreSQL. To do this, they apply fuzzy matching between the control and treatment datasets using possible confounding variables and then do casual inference. The contribution of this paper is that this procedure (which is very common in statistical anaylsis) is implemented as a PostgreSQL package allowing it to scale to billions of observations.

5.2 SparkSQL [1]

Interactive querying of big data is very common task in nearly any data intensive field. In the past this kind of data analysis was done with a RDBMS (for example, PostgreSQL). While this provides interactive querying, SQL is not well suited for complex data analysis. SparkSQL provides a fully featured relational model for interactive data processing while allowing for the execution of arbitrary snippets of code, making it much more flexible for doing data exploration, especially when more advanced data processing is required.

5.3 Coarsened Exact Matching [2]

A common problem in statistical analysis is that it is often not possible or too expensive to obtain experimental data for a subject of interest. To address this issue, observational data is often used, due to its availability and price. Using observational data however doesn't allow for the control of confounding factors due to the lack of randomized assignment. Coarsened Exact Matching address this problem by performing a kind of fuzzy matching between the ćontroland ťreatmentģroups in the data and then doing analysis based on this matching.

References

[1] Michael Armbrust et al. "Spark SQL: Relational Data Processing in Spark". In: SIGMOD '15 (2015), pp. 1383–1394. DOI: 10.1145/2723372.2742797. URL: http://doi.acm.org/10.1145/2723372.2742797.

[2] Stefano M. Iacus, Gary King, and Giuseppe Porro. "Causal Inference Without Balance Checking: Coarsened Exact Matching". In: *Political Analysis* 20.1 (2012), pp. 1–24. URL: https://www.cambridge.org/core/journals/political-analysis/article/causal-inference-without-balance-checking-coarsened-exact-matching/5ABCF5B3FC3089A87FD59CECBB3465C0.

[3] Babak Salimi et al. "ZaliQL: Causal Inference from Observational Data at Scale". In: *Proc. VLDB Endow.* 10.12 (Aug. 2017), pp. 1957–1960. ISSN: 2150-8097. DOI: 10.14778/3137765.3137818. URL: https://doi.org/10.14778/3137765.3137818.

[4] *UK Office of National Statistics*. https://ons.gov.uk/.

[5] *UK Open Data*. https://data.gov.uk/.