ZaliQL Demonstration

Babak Salimi, Corey Cole Dan R. K. Ports Dan Suciu
University of Washington
{bsalimi,coreyleoc,drkp,suciu}@cs.washington.edu

ABSTRACT

From CFP: Selection of demonstrations for presentation at SIGMOD is highly competitive. The evaluation criteria includes both the audience experience and the novelty of the system. The proposal should also describe in detail what SIGMOD attendees who view the demonstration will experience. What exactly will the audience see? Will they be able to interact with the system? Is there an interesting scenario or script that the demonstrators will use to motivate the demonstration? Authors may include a description of the novel intellectual content of the underlying system and of its architecture. The demonstration proposal must include citations of the most relevant papers about the proposed system and any previous publications of that same system. Priority will be given to systems that have never been demonstrated before. One of the demonstrations will be selected to receive an award of Best Demo.

ACM Reference format:
Babak Salimi, Corey Cole Dan R. K. Ports Dan Suciu . 2016. ZaliQL Demonstration. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 2 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Much of the success of Big data today comes from *predictive or descriptive analytics*: statistical models or data mining algorithms applied to data to predict new or future observations, e.g., we observe how users click on ads, then build a model and predict how future users will click. Predictive analysis/modeling is central to many scientific fields, such as bioinformatics and natural language processing, in other fields - such as social economics, psychology, education and environmental science - researchers are focused on testing and evaluating *causal hypotheses*. While the distinction between causal and predictive analysis has been recognized, the conflation between the two is common.

Causal inference has been studied extensively in statistics and computer science [**? ? ? ? ?**]. Many tools perform causal inference using statistical software such as SAS, SPSS, or R project. However, these toolkits do not scale to large datasets. Furthermore, in many of the most interesting Big Data settings, the data is highly relational (e.g, social networks, biological networks, sensor networks and more) and likely to pour into SQL systems. There is a rich ecosystem of tools and organizational requirements that encourage this. Transferring data from DBMS to statistical softwares or connecting these softwares to DBMS can be error prone, difficult, time consuming and inefficient. For these cases, it would be helpful to push statistical methods for causal inference into the DBMS.

Conference'17, Washington, DC, USA
2016. 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

Both predictive and causal analysis are needed to generate and test theories, policy and decision making and to evaluate hypotheses, yet each plays a different role in doing so. In fact, performing predictive analysis to address questions that are causal in nature could lead to a flood of false discovery claims. In many cases, researchers who want to discover causality from data analysis settle for predictive analysis either because they think it is causal or lack of available alternatives.

This work introduces ZaliQL,[1] a SQL-based framework for drawing causal inference that circumvents the scalability issue with the existing tools. ZaliQL supports state-of-the-art methods for causal inference and runs at scale within a database engine. We show how to express the existing advanced causal inference methods in SQL, and develop a series of optimization techniques allowing our system to scale to billions of records. We evaluate our system on a real dataset. Before describing the contributions of this paper, we illustrate causal inference on the following real example.

Example 1.1. FlightDelay. Flight delays pose a serious and widespread problem in the United States and significantly strain on the national air travel system, costing society many billions of dollars each year [**?**]. According to FAA statistics,[2] weather causes approximately 70% of the delays in the US National Airspace System (NAS). The upsetting impact of weather conditions on aviation is well known, however quantifying the causal impact of different weather types on flight delays at different airports is essential for evaluating approaches to reduce these delays. Even though predictive analysis, in this context, might help make certain policies, this problem is causal. We conduct this causal analysis as a running example through this paper. To this end, we acquired flight departure details for all commercial flights within the US from 2000 to 2015 (105M entries) and integrated it with the relevant historical weather data (35M entries) (see Section **??**). These are relatively large data sets for causal inference that can not be handseled by the existing tools. Table 1 presents the list of attributes from each data set that is relevant to our analysis.

When we make predictive analysis, whether we predict $\mathbb{E}[Y|X = x]$ or $\Pr(Y|X = x)$ or something more complicated, we essentially want to know the conditional distribution of Y given X. On the other hand, when we make a causal analysis, we want to understand the distribution of Y, if the usual mechanisms controlling X were intervened and set to x. In other words, in causal analysis we are interested in *interventional* conditional distribution, e.g., the distribution obtained by (hypothetically) enforcing $X = x$ uniformly over the population. In causal analysis, the difficulty arises

[1] The prefix Zali refers to al-Ghzali (1058-1111), a medieval Persian philosopher. It is known that David Hume (1711-1776), a Scottish philosopher, who gave the first explicit definition of causation in terms of counterfactuals, was heavily influenced by al-Ghzali's conception of causality [**?**].
[2] National Aviation Statistic http://www.faa.gov/

Attribute	Description
FlightDate	Flight date
UniqueCarrier	Unique carrier code
OriginAirportID	Origin airport ID
CRSDepTime	Scheduled departure time
DepTime	Actual departure time
DepDelayMinutes	difference in minutes between scheduled and actual departure time. Early departures set to 0
LateAircraftDelay	Late aircraft delay, in minutes
SecurityDelay	Ssecurity delay, in minutes
CarrierDelay	Carrier delay, in minutes
Cancelled	Binary indicator

(a) Flight dataset

Attribute	Description
Code	Airport ID
Date	Date of a repost
Time	Time of a report
Visim	Visibility in km
Tempm	Temperature in C°
Wspdm	Wind speed kph
Pressurem	Pressure in mBar
Precipm	Precipitation in mm
Tornado	Binary indictor
Thunder	Binary indictor
Hum	Humidity %
Dewpoint	De point in C°

(b) Weather dataset

Table 1: List of attributes from the flight(a) and weather(b) datasets that are relevant to our analysis.

from the fact that here the objective is to estimate (unobserved) *counterfactuals* from the (observed) *factual* premises.

Example 1.2. FlightDelay (Cont.). Suppose we want to explore the effect of low-pressure on flight departure delays. High pressure is generally associated with clear weather, while low-pressure is associated with unsettled weather, e.g., cloudy, rainy, or snowy weather. Therefore, conducting any sort of predictive analysis identifies low-pressure as a predictor for flight delays. However, low-pressure does not have any causal impact on departure delay (low-pressure only requires longer takeoff distance) [**?**]. That is, low-pressure is most highly a correlated attribute with flight delays, however ZaliQL found that other attributes such as thunder, low-visibility, high-wind-speed and snow have the largest causal effect on flight delays (see Sec. **??**); this is confirmed by the results reported by the FAA and [**?**].

2 DEMONSTRATION DETAILS

3 THE ZALIQL SYSTEM

4 CONCLUSIONS

ACKNOWLEDGMENTS

The authors would like to thank ..