HILDA 2016 Workshop: A Report

Arnab Nandi
The Ohio State University
arnab@cse.osu.edu

Alan Fekete
University of Sydney
alan.fekete@sydney.edu.au

Carsten Binnig
Brown University
carsten_binnig@brown.edu

Abstract

The first annual workshop "Human-in-the-Loop Data Analytics" (HILDA) was held on June 26, 2016 in association with ACM SIGMOD in San Francisco. The workshop sparked some excellent conversations between speakers and attendees from many disciplines spanning both industry and academia, and covered a variety of sub-themes related to human-in-the-loop data challenges.

1 Introduction

An often overlooked – but critical – part of the data management life cycle is the human-in-the-loop. The key focus of the HILDA workshop was to evaluate, understand, and improve the participation of humans in data management, with the eventual goal towards building optimized data management systems and techniques that treat humans as first-class citizens, alongside data. We aimed to bring together those in the database research community who are paying attention to the distinctive characteristics of people which impact the ways data management activities occur, and to reach out to other communities such as visualization, data mining, human-computer interaction as they grapple with data-related challenges.

HILDA 2016 accepted 16 papers [1] from 32 submissions. 8 of these were presented as longer talks, and others as short "taster" talks; all were also presented as posters. There were 53 registered participants and attendance reached over 90 at times during the workshop. There was a genuine buzz in the room with several great interactions, especially given the format where each session ended with a panel discussion among the presenters. We were very lucky to have two outstanding keynotes. Laura Haas (IBM Research) spoke about platforms for collaboration that have been built at the Accelerated Discovery Lab. These integrate information about people, data and tools, and support conversation as a metaphor for mixed-initiative interactions. Jeffrey Heer (University of Washington) spoke about declarative languages to describe (and thus allow automated generation and optimization of) data transformation and interactive presentation pipelines.

2 Research Themes

The presented papers spanned a multitude of facets of the HILDA theme, ranging from novel user interfaces, to infrastructural support for interactive analytics, to interactive data preparation.

User Interfaces: The usability of a data management system is hugely influenced by the nature of the interface

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

through which the users perform tasks such as express their queries, define the schema, tune for performance, and observe results. A significant part of the design towards supporting interactive analytics involves innovation in the visualization infrastructure. A tool for recommending visualizations was presented, as well as visual interfaces for particular tasks such as provenance, cluster adjustment, and identification of exceptional cases in data curation. There was also interest in conversational interfaces where a data platform and one or more users work together over multiple data manipulation steps through a (restricted) natural language.

Interactive Analytics: A second theme discussed during the workshop was centered around the question of how to design database systems to achieve the interactivity of analytical tasks – a crucial component towards accelerating human involvement. Papers tackled three important areas to achieve this goal: co-design of database system internals, interfaces, and visualizations; enabling ad-hoc analytics over new data sources by avoiding expensive data preparation and indexing costs; working with machine learning models and results (e.g., trends) in an interactive fashion.

One critical point raised during the workshop was that while enabling interactivity enables better and more widespread use of advanced analytical tools, it also significantly increases the risk of making spurious discoveries. Therefore, quantifying the risk is of utmost importance for the next-generation of interactive data exploration systems.

Data Cleaning, Extraction, and Labeling: Another key area where the involvement of humans was highlighted was that of bringing data into a structured form for downstream processing. In contrast to traditional data ingestion where data pipelines are considered a "one-time" effort, there is a growing understanding that this is typically an iterative and interactive process. Issues such as iterative data cleaning, supervised extraction, and guided data preparation are becoming quite common. Beyond traditional databases methods towards enabling humans to programmatically label data in machine learning pipelines was also discussed.

3 Bridging Communities

One major success of the workshop was the bringing together of people from the visualization, human-computer interaction, data mining, and data management communities. We were fortunate to see presentations and participation from students, researchers, and industry representatives from outside core data management areas. The trans-disciplinary interest in the area is bolstered by the presence of complementary workshops, such as the "Data Systems for Interactive Analysis" workshop at IEEE VIS, and the "Interactive Data Exploration and Analytics" workshop at ACM SIGKDD.

Industry Involvement: HILDA 2016 was supported with sponsorship from diverse companies interested in this new area of data analytics: IBM, Paxata, Tableau, and Trifacta. We are especially grateful to Paxata for providing scholarships for seven students to attend the workshop and the SIGMOD conference. There was also a strong industry presence in the program committee, in the paper submissions, and in discussions in the room.

Thoughts for the future: We are happy to announce that the 2nd annual HILDA 2017 workshop will be held at ACM SIGMOD 2017 in Raleigh, NC, USA on May 14, 2017. We encourage submissions to the workshop. More information is available at the HILDA website, `http://hilda.io`.

References

[1] C Binnig, A Fekete, A Nandi. *Proceedings of the First Annual Workshop on Human-in-the-Loop Data Analytics.* HILDA 2016.