Letter from the Special Issue Editor

People are always trying to do more with their data than current technology allows. This leads to a lot of pain, as they fight a battle to get the software to do what they want, without unmanageable amounts of money, time, complexity, or general frustration. When enough people are fighting the same war against their data, new data models and architectures might emerge which promise to bring peace to the data wars. Thus, the rise of relational systems, object oriented systems, XML systems, and so on, each dealing with one or more specific pain points. Unfortunately, each new generation of systems does not end the battle, but merely transform it, as practitioners try to use their new weapons to do even more interesting things with their data.

The latest battle is being fought against "Big Data" - data that is large, complex or dynamic enough to be too painful to handle with existing systems. In the last decade or so, the ease of collecting a mind-numbing amount of data, often but not always from web applications, has led to a desire to do something with all that data. Although a variety of systems have been developed to store, query, analyze and transform big data, many practitioners still feel that they are fighting a war against their data - with newer, more powerful and sophisticated (and some may say, deadly) software tools, but a war nonetheless.

In this issue, we collect "Big Data War Stories" - examples of practitioners trying to use big data tools, finding that it was still painful, and having to adopt new strategies and tools to overcome their difficulties.

- From Facebook, we hear about experiences trying to get a new and promising but relatively untested technology (HBase) to manage a huge number of user messages.
- From Microsoft, we hear about efforts to take noisy maps and shopping data and clean it well enough to make it useful to users.
- From Yahoo, UC Santa Cruz and UC Irvine, we hear about a key limitation of big data storage systems: poor support for machine learning, and new strategies to deal with this limitation.
- From LinkedIn, we hear about how logs of activity data can be very useful in driving web applications, datacenter management and so on, but also how new techniques are needed just to move the right data to the right place to be analyzed.
- From Google, we hear about how just collecting and analyzing big data is not enough, if it cannot also be easily shared and visualized.

In some of these cases, the right tactic was to modify and extend an existing platform or technique. In other cases, a whole new tool was needed to overcome the limitations of big data software.

I hope these stories, and the lessons contained within them, are useful to other practitioners fighting similar battles against their data. I would like to thank all of the authors who agreed to share their experiences, as well as Dave Lomet who, as always, provided excellent advice during the process of putting together this issue.

Brian F. Cooper
Google
Mountain View, CA USA