Improving Data Collection and Monitoring through Real-time Data Analysis

P. Lubell-Doughtie, P. Pokharel, M. Johnston, V. Modi
Columbia University
New York, New York
peter@helioid.com, {pp2427, mj2537, modi}@columbia.edu

ABSTRACT

Feedback based on real-time data is increasingly important for ICT-based interventions in the developing world. Applications such as facility inventories, summarization of patient data from community health workers, etc. need processes for analyzing and aggregating datasets that update over time. In order to facilitate such processes, we have created a modular web service for real-time data analysis: **bamboo**.

1. INTRODUCTION

To effectively monitor community health, allocate resources, and respond to crises, both countries and planners need real-time data. To address this, tools like EpiSurveyor, OpenDataKit, and formhub allow development planners and others to conduct data gathering exercises without the need for server infrastructure or in-house programmers. Nevertheless, data collection is only part of the picture.

In emergency response, the time lag in processing structured data causes significant delays; faster data analysis allows decision makers to address problems sooner and have a greater impact [4]. Health workers and managers can learn how to best modify their health programs and avert future deaths through access to audit trails based on timely data [5]. In mobile health systems, monitoring real-time reports can improve health programs and address key health risks [7].

Dynamic data analysis refers to viewing and analyzing information—such as ongoing survey data—in real-time as it is updated. Dynamic data analysis is a prerequisite for the aforementioned applications, yet it demands resources and skills that are often unavailable in the development context. Even simplistic systems that perform user-defined aggregations and calculations on dynamic datasets require high technical capacity. **bamboo** provides real-time aggregation, calculation, and summarization as a hosted web service.[1] Practitioners can interact with **bamboo**, and with their up-to-date datasets, using an easy to learn syntax.

[1] `http://bamboo.io/`

DEV '13, January 11-12, 2013 Bangalore India

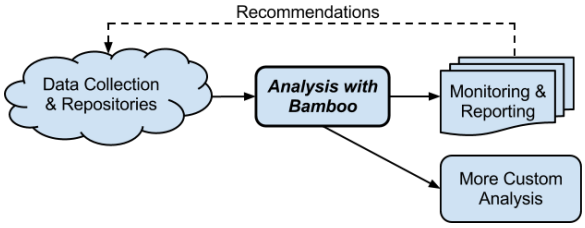

Figure 1: Systematizing data collection and reporting with bamboo.

2. RELATED WORK

We categorized existing solutions for dynamic data analysis as: *custom tools*, *hosted tools*, and *offline tools*. Custom tools are either built in-house or by a third-party, they are expensive and beyond the capacity of most organizations. Furthermore, custom tools often lead to a duplication of effort and are difficult to adapt to new tasks. Popular hosted tools, such as Google Fusion Tables and Google Docs, have functional limitations that made them unsuitable for our needs. Google Fusion Tables only allows a limited set of calculations, and the "one spreadsheet" model of Google Docs precludes aggregations [3]. Offline data analysis tools, e.g. Microsoft Excel, Python, R, SPSS, STATA, etc. have the flexibility required but have steep learning curves and require programmatic wrappers to allow for truly dynamic workflows, creating high barriers to entry.

In summary, custom tools are challenging to adapt and maintain, hosted tools are not flexible, and offline tools do not allow a centralized data store. We therefore built **bamboo** to provide the combination of updates, aggregations, and ease of use, which our and many other development research tasks require.

3. DESIGN

bamboo sits between data collection and reporting, as shown in Fig. 1. **bamboo**'s core functionality allows practitioners to:

1. store, update, and merge datasets,
2. build algebraic calculations and aggregations, and
3. generate summary statistics: means, counts, etc.

This exposes the split-apply-combine strategy of data analysis [8] through a web service. The dynamic statistical analysis within **bamboo** allow practitioners to easily build dash-

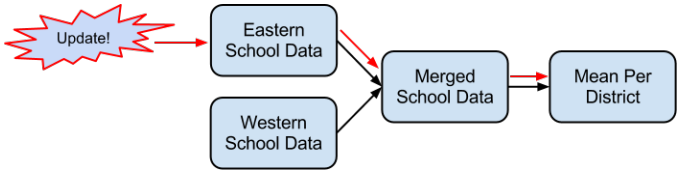

Figure 2: Red arrows show an update propagating to all downstream merged datasets and aggregations. Black arrows show structural dependencies created by the client.

boards, maps, and tables. These reporting tools will automatically update as new data arrives.

Generalizability is a fundamental principle of the design: **bamboo** accepts any CSV file and provides users with complete control over their calculations. Updates can be submitted to **bamboo**'s application programming interface (API) using JavaScript Object Notation (JSON) web requests, making it simple to bring together data from a variety of sources. These updates are propagated through the system ensuring that any aggregations or merged datasets are synchronized with the most recent data, as depicted in Fig. 2. To simplify updates, we have integrated **bamboo** with formhub, the mobile data collection platform, such that updates to formhub can be automatically passed onto **bamboo**. **bamboo** could be similarly integrated with ODK Aggregate and other data collection platforms.

To encourage community use and development we have made the code open source and structured it to be easily extendable. The **bamboo** web service uses Representational State Transfer (REST) conventions and we have written client libraries in Python and JavaScript that connect to it. For numerical computation **bamboo** uses pandas, a Python library for high-performance statistical data analysis [6].

4. CASE STUDY

During a large-scale data collection project in Nigeria, in which over 200,000 individual surveys were conducted, researchers used **bamboo** to monitor progress. Specifically, as a public water facilities survey was conducted across the country **bamboo** summarized the amount of data collected per state. This provided data collection monitors with the tools they needed to identify states in which data coverage was lower than expected.

To gain further insight, researchers used **bamboo** to conduct exploratory monitoring of this dataset through complex metrics. Fig. 3 shows the number of waterpoints surveyed per 1,000 people. **bamboo** analyzed the raw survey data and created the summary which powers this chart, ensuring that it was up-to-date as additional data was collected.

5. CONCLUSIONS

bamboo allows users without programming expertise to perform real-time data analysis. We envision immediate application in the development context, making performance monitoring [1] and real-time outlier detection within categorical data [2] much easier to implement. **bamboo** systematizes the process of creating indicators for domain specific datasets thereby reducing the time between collection and analysis, as well as the time between analysis and reporting.

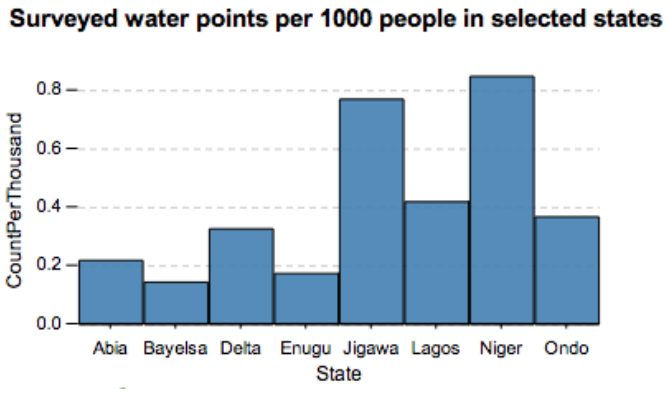

Figure 3: After merging datasets, the normalization formula required one API call. Thereafter, the same URL provides updated real-time data to create the graph above.

6. FUTURE WORK

In future work we will focus on providing additional analysis relevant to development work and methods to define and share this analysis. We will add analysis tools to encompass more real-world data processing tasks, such as Levenshtein distance, nearest neighbors calculations, time series analysis, and spatial analysis. From our field experience, we see the potential for a predefined library of common analysis techniques to reduce efforts duplicated across domains. Within certain domains we plan to offer "calculation libraries", which codify common metrics for that domain. For example, health researchers could create a set of maternal health indicators once and then apply these same indicators across all of their datasets.

References

[1] M. Berg. Childcount+ initial report. Technical report, Center for Global Health and Economic Development, Earth Institute, Columbia University, 2007.

[2] B. Birnbaum et al. Automated quality control for mobile data collection. In *Proceedings of the 2nd ACM Symposium on Computing for Development*, ACM DEV '12, New York, NY, USA, 2012. ACM.

[3] H. G. et al. Google fusion tables: Web-centered data management and collaboration. In *Proceedings of the ACM SIGMOD conference*, 2010.

[4] Internews. Dadaab, kenya - humanitarian communications and information needs assessment among refugees in the camps. Technical report, Internews, 08 2011.

[5] K. Krisberg. Worldwide maternal mortality on the decline, but much more work is needed. *The Nation's Health*, 40(9), 2010.

[6] W. McKinney. pandas: a foundational python library for data analysis and statistics. In *SC '11: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, New York, NY, USA, 2011. ACM.

[7] P. Mechael et al. Capitalizing on the characteristics of mhealth to evaluate its impact. *J Health Commun*, 17 Suppl 1, 2012.

[8] H. Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1), 4 2011.