Self-Service Data Preparation and Analysis by Business Users: New Needs, Skills, and Tools

Gregorio Convertino
Informatica LLC
Redwood City, CA 94063, USA
gconvertino@informatica.com

Andy Echenique
Informatica LLC
Redwood City, CA 94063, USA
aechenique@informatica.com

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA
© 2017 ACM. ISBN 978-1-4503-4656-6/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3027063.3053359

Abstract
This case study characterizes the new ecology of needs, skills, and tools for self-service analytics emerging in business organizations. A growing population of data analysts are being employed in departments such as Sales, Marketing, and Finance. These new users call for new tools for self-service analytics. The paper summarizes qualitative findings from three years of user research with business users performing data analysis tasks. The call for a broader theoretical framework is proposed based on these findings.

Author Keywords
Big Data; Data analysis; Data Preparation; Business Users; Data Analysts; Data Scientists; Sensemaking

ACM Classification Keywords
H.5.m. Information interfaces and presentation: Miscellaneous

Introduction and Motivation
Within the past 5 to 10 years, organizations have benefitted from unprecedented capabilities of continuous data gathering and cheap data storage. Combined with more readily available data analytics, it has triggered the start of the age of "big data" and

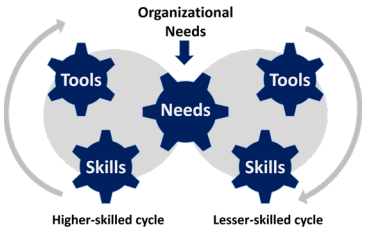

Figure 1: Framework for interlocking tools-skills-needs cycles.

"data-driven decisions" for businesses (e.g., Metcalf, 2016; Fisher 2014). This age gives organizations of all sizes new ways to understand and reengineer their business practices. At the same time, it has increased the demand for both data analytics roles and skills as well as tools that allow self-service analytics with big or medium-sized data sets. In response to this demand, "data science" programs have flourished in academia to train new professionals in this area.

Business organizations are thus increasingly reliant on data analytics to utilize data for a competitive advantage. Businesses commonly analyze their transaction and demographic information to improve efficiency, model customer engagement, and target sales and marketing efforts. This has resulted in a growing population of data analysts employed in departments such as Sales, Marketing, and Finance within organizations. These professionals have a variety of job titles, such as "Business [Intelligence] Analyst", "Data Analyst," Data Analyst in Operations" or "Data Scientist". However, they all share the need to prepare and use data to address specific business questions at hand (e.g., Kandel et al. 2012).

In addition to dedicated data professionals, there is also a growing population of casual, less specialized analyst calling for self-service analytics tools (e.g., Siegel et al. 2013). Casual data analysts fulfill the same responsibilities as their dedicated counterparts, yet do so while also performing other roles within the organization e.g., Product Managers, Line of Business or Team Managers, and Business Development Managers. For example, a Sales Manager may analyze data she has in her customer relationship management (CRM) system to assess the performance of her team in a specific geography over the past month.

Growing populations of dedicated and casual data analysts are adopting an increasing number of tools (see McKinsey 2011 and Gartner 2015). Traditionally, tools such as Microsoft Excel have been sufficient for data analysis. Yet the necessity to handle millions of rows and columns of data has pushed the adoption of new data analysis tools (Kandel et al. 2012). However, we observed that current efforts to develop new tools are targeted to a narrow set of user classes, yet little consideration for the broader ecology of the different tasks and roles. A broader framework that explains the new ecology of needs, skills, and tools is still missing.

As a first step towards creating broader framework of evolving roles and tools, this case study reflects on data collected during three years of user studies to understand how the changing skill sets have impacted data analysis tools. We present results on the variety of skills, needs and tools present for different categories of data analysts as well as data on how the change in tools have created changes within this community. Our results indicate a reduction in the barrier-to-entry for analysis tools which has accompanied a democratization of analysts utilizing large data sets. Our conclusions foreshadow the impact of these tools in creating new needs and designs for data analysis tools as well as the implications of an interlocked task-artifact cycle.

Prior Research and Framework

Prior work has highlighted the current incongruities between organizational needs, data analysts' skills, and current tools. Siegel and collaborators (2013)

conducted an in-depth qualitative research with 54 business professionals in eight organizations. They investigated numerous cases of casual data users and found a mismatch between the data-based decision-making practices that their organizations promote and the users' actual skills, motivations, and practices. Most casual users of data were not motivated to learn new skills beyond their own rudimentary quantitative practices. Thus, analysts could not take full advantage of the data available to them due to a gap between their quantitative analytic skills and the advanced skills necessary to use the business intelligence tools. Moreover, relevant qualitative information that could be used to interpret quantitative data tended to live in their heads rather than being shared, which represent another lost opportunity for the organization.

Previous research has also focused on describing the skills sets and roles underlying data analysis practices. Kandel and colleagues (2012) investigated 35 data analysts in various industries and analyzed their typical data analysis tasks. They identified three categories of data professionals: Application User, Scripter, and Hacker. The Application User is a data analyst with minimal to no programming skills that relies on the capabilities of advanced software to answer business questions. At the same time, they observed the scale and diversity of data sets in organization is increasing. They argue that future analytic tools which simplify analysis, such as visual analytics tools, can empower non-programmers to improve the quality and speed of their data analysis work.

Finally, previous research provides a proposes a feedback loop between users' behaviors or skills and the tools they use (Siegel et al. 2013). In Human-Computer Interaction, this loop is described as the task-artifact cycle by Carroll & Rosson (1992). This cycle proposes that tools (artifacts) and skills (operationally representable as tasks) co-evolve over time. As the skills of user class adapt to the tools available, then the tool features change, which in turn triggers new evolution of users' skills. The task-artifact cycle can explain the co-evolution occurred between data analyst skills and tools. In other words, as data analysts have continued to advance their own analytic skills and applications, tools have responded by leveraging these skills for more advanced analysis capabilities.

Research Questions

- **Analytics needs in business environments**. What are the new needs of business users performing data preparation and analysis?
- **Analytics skills**. What are the new skill sets of business users (data analytics literacy)?
- **Analytics tools and features**. What are the new tools emerging in responses?

We report qualitative findings in response to the individual questions above and propose them as an extension of the task-artifact cycle (Carroll and Rosson 1992) can help design future analytics tools for business users in organizations.

We build on the task-artifact cycle framework in two ways: First, we posit that the cycle should explain the co-evolution of three primitives: needs (e.g., the organization's needs for data analysis), skills, and tools (Figure 1). Second, we propose that for a more realistic description of the ecology of roles in organizations,

Sales Data Analyst
(task from an interview)

*"**Every quarter**, I pull together the deals won for my product [XXX]. I have a prebuilt **report created in Salesforce** that I download. Then I open a **worksheet that I have saved** that I paste the data into. Then, with a **combination of formulas and moving some data** (to get the spacing right), I have my sheet. I usually do this only once a quarter but would love to do it once or a month or so. I have a variety of **calculations**; I bucket the number of cores into ranges, calculate the price per core, determine product lead and identify ELA vs. licenses. I also **cleanup** the territory; [MyCorp] includes Healthcare as a region for example. I just want North America. Then I have a variety of pivot tables to calculate various summarizations (by year, by quarter, by region). This provides me all the information I need."*

multiple interlocked cycles (for different organizational roles) need to be considered when designing new tools.

Method

Over a period of three years, the authors (two user researchers) conducted a series of ten user studies. In each study, a researcher collected data from a small number of users (5-10) via remote interview sessions and online surveys. Each session included a semi-structured interview and a design evaluation as part of the work on two new software products for data analysis developed at Informatica LLC. The researchers collected video recording of the user's screen and answers to a post-study online survey data. The findings in this paper are based on the data from the semi-structured interview, or first part of the study session, and the online survey.

The study involved 68 business users, each participated in a 90-minute remote interview. Table 1 summarizes the distribution of the participants by industry. Participants filled out a post-study online survey on typical questions, tools, data sources, and outputs for their analyses. The interview results were analyzed via open-coding to identify recurring themes.

Three User Classes

The results of open-ended coding surfaced three types of data analysts. Each data analyst differed in its analytic skill level and tools use. The descriptions below characterize each user class and their needs.

Data Analyst: Scenario

Data Analysts are primarily characterized by the time spent on analyzing data and their business knowledge. Their responsibilities originate with a business question to be answered using data available to the analyst. Business questions can include sales figures for a region or logistics data from a recent transaction. Data Analysts primarily use spreadsheet programs such as Microsoft Excel to clean and filter the data to answer this question, using formulas, pivot tables and sample charts to answer their business questions. Data Analysts spend the vast most of their time on data preparation (Figure 2, internal ring). Once the data is clean, it is visualized using pre-packaged charts in tools such as Microsoft Excel or Salesforce to represent the findings. The work flow can thus be represented as clean-and-report. The Data Analysts' skills allow simple data cleaning, analysis and data visualization operations.

Business Intelligence (BI) or Sr. Data Analyst Scenario

BI follow similar procedures as Data Analysts, yet with access to more varied data sources. Similar to data analysts, Sr. Data Analysts begin with a business question. Sr. Data Analysts spend most their time on data preparation (Figure 2, external ring). Yet rather than sourcing data from a specified location, Sr. Data Analysts are given wider areas of access within organizations and the ability to access basic forms of big data (e.g., relational databases). BI Analysts can use a wide variety of Business Intelligence (BI) tools such as Tableau or Microstrategy, which gives them more flexibility in representing the results. Sr. Data Analysts also follow the same clean-represent-report workflow as Data Analysts, yet with a greater emphasis on the levels of analysis provided.

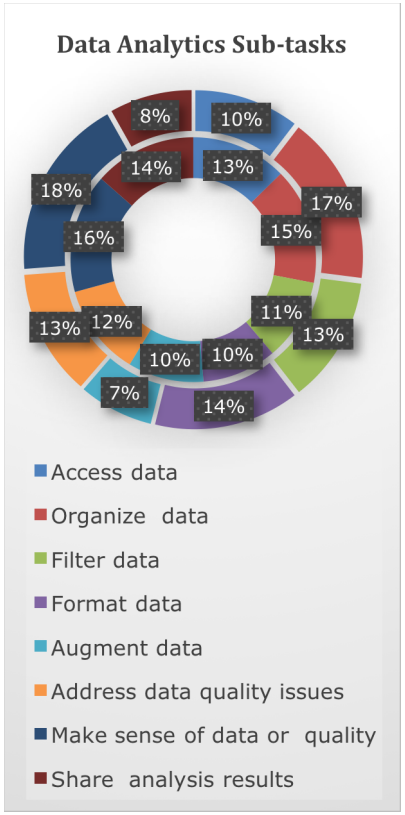

Figure 2: Percent of the total data analysis time spent weekly on each of these sub-tasks by Data Analysts (N=24), shown by the internal ring of the chart, and BI Analysts (N=10), external ring of the chart. For all sub-tasks, on average Data Analysts and BI Analysts spent of 24.6 and 20.0 weekly hours, respectively.

Data Scientist: Scenario

The final user class is the Data Scientists. Considered the most technical user within our study, the Data Scientists exhibit the most complex data preparation process. In addition to cleaning and filtering data, Data Scientists merge data sets and perform statistical analyses. Representing and reporting data are given less emphasis by data scientists, though they can use similar Business Intelligence tools. As opposed to the single clean-represent-report cycle, the Data Scientists' cycle includes several iterations of data preparation. Once the data is satisfactory, the result of the analysis is reported to the originator of the business question.

Needs, Skills, and Tools of User Classes

Our results define the needs, skills, and tools used by each class of data analysts. We analyzed survey data from 45 of the 68 study participants. Of the 45 participants, 24 classified as Data Analyst (i.e., regular users of Microsoft Excel and, in some cases, department-specific applications such as Salesforce), 10 as BI Data Analyst (i.e., user of Business Intelligence applications such as Tableau, QlikView, Salesforce Analytics, Business Objects, and local databases), and 11 as Data Scientists.

Needs

A significant result of this research is the identification of the needs of the data analyst community; needs without which they cannot fulfill their responsibilities. Data from the 68 participants in this study points to two main needs for further research: the need for handling large and diverse data sets and the need to keep track of combined data sets over time. We will describe these needs in depth and their implications.

The need to handle large and diverse data sets is becoming a pervasive responsibility for data analysts, regardless of skill and field of operation. Participants cited handling spreadsheets sets larger than 1 million rows on a regular basis. Handling data this large with traditional tools either creates performance issues or is simply impossible. It is also becoming increasingly common for analysts to integrate data of different formats and from different sources. Performing data integration requires specialized data preparation skills and a considerable amount of time. This necessitates participants to use more powerful analytic tools, such as SQL scripting or dedicated data preparation tools intended for large and varied data sets (see Informatica Intelligent Data Lake (IDL) and Trifacta Wrangler).

Managing "Big Data" also forces analysts to take a more holistic approach to data management. Storing data sets on private work computers was no longer possible. Specialized infrastructure would often be employed to handle large data sets, including servers appropriated specifically for data sets and multiple users. With the files now remote, importing or export large files also became a larger concern. Depending on bandwidth, retrieving files from a central server could take several minutes. Current data analyst tools need to account for this wait time into their design as well.

The second major need for data analysts is to track interconnected data sets. Analysts commonly combine data from multiple sources to receive the highest fidelity when answering complex business questions. Once the combined data set is created, analysts can

create more relevant cross sections of the data for more relevant results. An example within our data is a market researcher who combined sales data from multiple locations to analyze sales across all stores.

Once the data was aggregated, she singled out the specific item she was asked to investigate and analyzed that product's performance. Combining and deriving multiple data sets is a critical action for analysts to answer specific business questions (see sidebar). Current tools fail to manage the interlocking dependencies between data sets, especially when combined and derived data sets can be used an analyzed by several users.

Skills

The skills of an analyst are closest associated with the type of tools the analyst declared using. Data Analysts, who typically used small and medium data sets, performed tasks that were often limited to cleansing and reporting descriptive metrics. In contrast, Data Scientists, who used larger data sets, performed tasks that also included data modeling and prediction, such as predicting the best candidates for a new marketing campaign n. Thus, the variety of tools that an analyst can use acts as an indicator of her or his level of skill in analytics. This finding is consistent with the results reported by Kandel et al. (2012), where the tools are used to discriminate the Application Users from the more skilled Scripters and Hackers.

The differences in skills sets between Data Analyst and BI Data Analysts were modest since the two classes of users answered similar questions. The latter had additional skills such as performing basic operations on larger datasets (e.g., using SQL client) and use greater variety of tools. More evident was the difference in skills between these two classes, on the lower hand, and the Data Scientists on the higher end.

Another finding was that casual analysts are not commonly low-skilled analysts. This differs from prior characterizations of casual analysts (e.g., Siegel et al. 2013). The frequency of analyses was more commonly associated to the job role rather than the level of skills. This suggests that analytics duties are increasingly common for knowledge workers, even if their primary job is not data analysis (e.g., Product Managers).

Tools

Tools used by data analysts revealed the most apparent distinction between the roles of Data Analysts and BI Analyst vs. Data Scientists. Data Analysts and BI Data Analysts used spreadsheets software (e.g. Microsoft Excel) as their primary data analysis tool. Spreadsheets were followed by a variety of department-specific applications such as tools for storing data of Sales, Marketing, and Finance professionals, which often include prepackaged analytics capabilities (e.g., Salesforce) (Figures 3, 4).

The landscape of adopted tools is significantly different when considering the Data Scientists. For example, as shown in Figure 5, the most frequently used tools require scripting or programming (Python, Ruby, Java, etc.). Complexity and range of tool use is also tied with an analyst's skill level. The largest distinction between Data Scientists and Data Analysts is the number of tools they used and the skills necessary to operate them. The tool most used daily were scripting

languages such as Python and Ruby, with 82% of Data Scientists stating they used them daily.

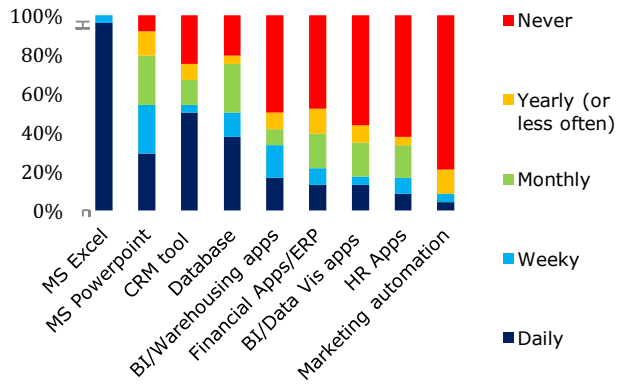

Figure 3: Tools used by Data Analysts (N=24). CRM: customer relationship management applications, BI: Business Intelligence tools w/ visualization capabilities.

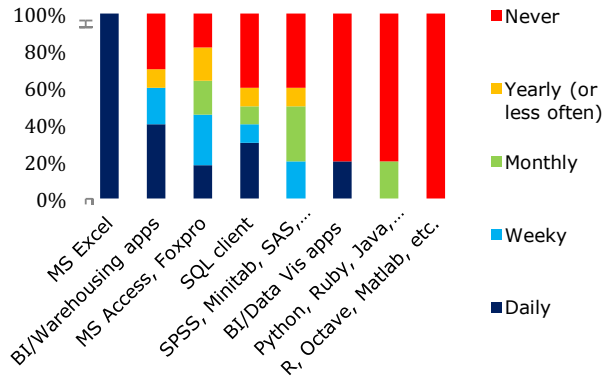

Figure 4: Tools used by BI Analysts (N=11).

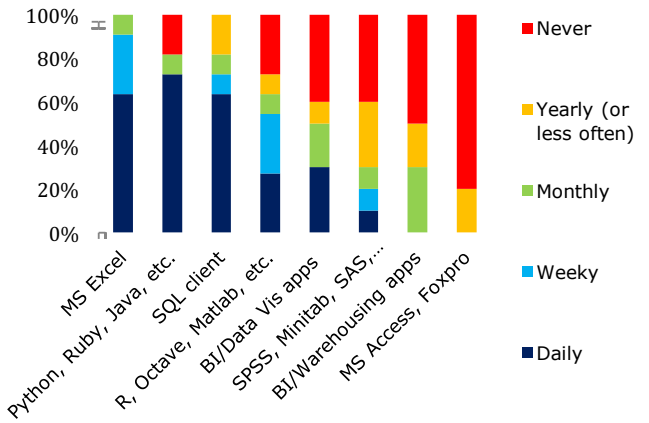

Figure 5: Tools used by Data Scientists (N=11).

By comparison, Data Analysts and BI analysts did not report using any scripting languages. The level of difficulty of tools extends to how Data Scientists find relevant data sources. Both BI Analysts and Data Analysts used predefined connections to local databases as a common activity (80% of BI analysts and 37% of data analysts perform this daily). The same task is performed by Data Scientists using SQL-based scripting to operate.

In addition to using more sophisticated tools, Data Scientists also differ in the breadth of tools used to answer business questions. Validating data, a task common to all data analysis, is undertaken with a larger number of tools by data scientists. Both BI and Data Analysts reported using Business Intelligence tools (such as Tableau) to understand the outputs of their data preparation. Data Scientists reported using statistical tools such as R for the same activity; tools

which require a greater level of skill with coding and statistical methods.

In this tool context, a general finding from our survey suggests that Data Analysts and BI Data Analysts generally spend more time preparing and organizing the data - i.e., about 60% of their data analysis time (on average, 58% for 24 Data Analyst and 61% for 10 BI Data Analyst) - than the activities of accessing, making sense, and sharing results combined - which account for 40% of their data analysis time (on average, 42% for 24 Data Analyst and 39% for 10 BI Data Analyst). This suggest a key bottleneck in efficiency of the data analysts is still in the limited support for data preparation; i.e., tool support for combining, organizing, cleansing, and filtering the data for the question at hand.

Conclusions

This study condenses findings on the co-evolving tool and skills of Big Data analysts. Qualitative research findings indicate a broader framework to understand the ecology of needs, skills, and tools within data analytic. This is motivated by a visible gap: organizations are internally pushing for more data-based decision-making to derive a competitive edge. Coupled with the challenge of increasingly large and diverse data sets, this raises the bar for the required skill level for data preparation and analysis. However, the number of highly-skilled individuals, such as Data Scientists, is limited. Simultaneously, current tools do not empower most Data Analysts, which are medium-skilled, to perform the necessary analysis. The result gap is the organization's lost opportunity to utilize available data and gain a competitive edge.

We argue that this gap is the result of a myopic approach to the design of data analysis software. Tools gravitate either towards the low number of highly-skilled Data Scientists or the larger, under-supported population of Data Analysts with very little middle ground and interplay (Siegel, 2013). In conclusion, we argue for the development of platforms of an integrated platform: a muti-tool, multi-role environment. The design of the envisioned environment needs to account for interlocking tools-skills-needs cycles specific to different job roles in the same organization (Figure 1).

Implementing the proposed multi-tool, multi-role platform requires affording both synergy and co-evolution. Affording synergy means supporting complementary job roles with related tasks. E.g., as a Data Scientist creates projected sales by quarter, Data Analysts can synchronously utilize up-to-date baseline values for comparison. Affording co-evolution of skills and tools requires that the different sets of roles-specific tools will evolve in concert. Thus the degree of flexibility of the platform to adapt to new skill configurations will determine its efficacy with helping the organization derive value from data. Future research is needed to further refine and specialize the classes of data analysts outlined in this case study and, on this basis, define a more detailed and implementable versions of the proposed framework.

Acknowledgements

The authors wish to thank their colleagues in User Experience and Product Management as well as the respective team managers for their support. We also thank the REV and IDL teams for their valuable feedback over the years conducting the user studies which made this research possible.

References

1. Carroll J.M. and Rosson M.B. Getting around the task-artifact cycle: how to make claims and design by scenario. ACM Trans. Inf. Syst. 10, 2, 181-212. 1992.
2. Fisher D, Drucker S, Czerwinski M. IEEE Comput Graph Appl. 2014 Sep-Oct;34(5):22-4. Business intelligence analytics.
3. Gartner report by Randall L., Sallam R.L., Hostmann B., Zaidi E., Market Guide for Self-Service Data Preparation for Analytics, March 5, 2015.
4. Kandel S., Paepcke A., Hellerstein J. M., and Heer J. Enterprise data analysis and visualization: An interview study. Visualiz. & Computer Graphics, IEEE Trans. 2012.
5. McKinsey report by Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A., Big data:The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011.
6. Metcalf, J., Keller E. F., boyd D. Perspectives on Big Data, Ethics, and Society. Council for Big Data, Ethics, and Society. 2016.
7. Siegel D., Sorin A., Thompson M., and Dray S. Fine-tuning user research to drive innovation. interactions, 20(5), 2013.
8. Microsoft Excel: www.microsoftstore.com/office (2017/02/20).
9. Informatica IDL: www.informatica.com/products/big-data/intelligent-data-lake.html (2017/02/20).
10. Trifacta Wrangler: www.trifacta.com/products/ (2017/02/20).
11. Salesforce: http://www.salesforce.com/ (2017/02/20).
12. Tableau Software: http://www.tableau.com/ (2017/02/20).