Appendix to Everything You Wanted to Know About Data Scientists in Software Teams

December 2, 2016
Technical Report
MSR-TR-2016-1127

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Data Culture at Microsoft

We are researchers in Microsoft Research investigating the emerging role of data scientists at Microsoft. With this survey, we want to learn about the role data science plays as part of your work. We would like to understand your background, the types of work you do, and the tools you use. Your answers will help us to improve data science at Microsoft.

We would be greatly appreciative if you would be willing to take the survey. The survey should take 20-30 minutes. Thank you!

This survey is anonymous. No personal information will be collected. Aggregate information may be shared with research collaborators outside of Microsoft and used in publications. We selected you as part of a sample of Microsoft employees worldwide based on your job role and/or subscriptions to distribution lists. Please contact tzimmer @ microsoft.com if you have any questions about this research project.

As a thank you for your time, you can enter your name into a raffle of four $125 Visa Gift Cards after completion of the survey (official rules of the sweepstakes).

Thanks,
Andrew Begel (ABEGEL), Rob DeLine (RDELINE), and Thomas Zimmermann (TZIMMER)

Please do not use the browser's Back button if you want to go back a page. Instead, use the survey's Back button, which you can find at the end of each page.

1) What group do you primarily work in? (required)*

( ) Advanced Technology (Eric Rudder)
( ) ASG - Applications and Services Engineering Group (Qi Lu)
( ) Business Development Group (Peggy Johnson)
( ) C&E - Cloud and Enterprise Engineering Group (Scott Guthrie)
( ) Dynamics (Kirill Tatarinov)
( ) Finance Group (Amy Hood)
( ) HR Group (Kathleen Hogan)
( ) LCA - Legal and Corporate Affairs Group (Brad Smith)
( ) Marketing Group (Chris Capossela)
( ) MDG - Microsoft Devices Group (Stephen Elop)
( ) OSG - Operating Systems Engineering Group (Terry Myerson)
( ) Operations (Kevin Turner)
( ) Strategy (Mark Penn)
( ) TnR - Technology and Research (Harry Shum)
( ) Other

Validation: Min = 0 Max = 99 Must be numeric Whole numbers only Positive numbers only

2) What is your age in years? (whole number) (optional)

_____

3) What is your gender?

( ) Female
( ) Male
( ) Other
( ) Prefer not to say

4) Do you have any direct reports?

( ) Yes

( ) No

5) Which of the following disciplines best describes your job role?

( ) Data & Applied Science.
Example job titles: Applied Scientist, Data Scientist, Data Scientist Lead, Applied Science Mgr, etc.

( ) Software Engineering.
Example job titles: Software Engineer, SDE, Software Engineering Mgr, etc.

( ) Program Management.
Example job titles: Program Manager, Program Manager Lead, PM Manager, etc.

( ) Other: _____

6) In what location do you work?

( ) North America: USA - WA (Puget Sound regions: Redmond, Bellevue, Seattle, Sammamish, etc.)

( ) North America: USA - Silicon Valley, CA

( ) North America: USA - Other

( ) North America: Canada, Mexico

( ) Central America and South America

( ) Europe

( ) Asia: China

( ) Asia: India

( ) Asia: Middle East

( ) Asia: Other

( ) Australia, New Zealand, Oceania

( ) Africa

( ) Other: _____

7) How long have you worked professionally, including your current job?

Validation: Must be numeric

Years: _____

Months: _____

8) How long have you worked at Microsoft?

Validation: Must be numeric

Years: _____

Months: _____

9) How long have you been analyzing data (for work or for fun)?

Validation: Must be numeric

Years: _____

Months: _____

10) What is the highest degree or level of school you have completed? If currently enrolled, please select the highest degree received.

( ) Secondary school degree (high school) or less

( ) Some university/college, no degree

( ) Associate's degree (2 year)

( ) Bachelor's degree (3 or 4 year university)

( ) Master's Degree (graduate degree)

( ) PhD (graduate degree)

( ) Other: _____

11) Please check all that apply. (required)*

[ ] I analyze product and customer data

[ ] I build predictive models from the data

[ ] I build data engineering platforms to collect and process a large quantity of data

[ ] I use big data cloud computing platforms (Cosmos, Dryad, MapReduce, AzureML) to analyze large data

[ ] I add logging code or other instrumentation to a system to collect the data required for my analysis

[ ] I communicate results and insights to business leaders

[ ] I manage one or more data scientists

[ ] None of the above

12) Please rank your skills. Drag the items from the left column to the right column. Place your best skills at the top. Select only the skills that you have.

_____Algorithms (ex: computational complexity, CS theory)

_____Back-End Programming (ex: JAVA/Rails/.NET)

_____Bayesian/Monte-Carlo Statistics (ex: MCMC, BUGS)

_____Big and Distributed Data (ex: Hadoop, HBase, Cosmos, Map/Reduce)

_____Business (ex: management, business development, budgeting)

_____Classical Statistics (ex: general linear model, ANOVA)

_____Data Manipulation (ex: regexes, R, SAS, web scraping)

_____Front-End Programming (ex: JavaScript, HTML, CSS, ASP)

_____Graphical Models (ex: social networks, Bayes networks)

_____Machine Learning (ex: decision trees, neural nets, SVM, clustering)

_____Math (ex: linear algebra, real analysis, calculus)

_____Optimization (ex: linear, integer, convex, global)

_____Product Development (ex: design, project management)

_____Science (ex: experimental design, technical writing/publishing)

_____Simulation (ex: discrete, agent-based, continuous)

_____Spatial Statistics (ex: geographic covariates, GIS)

_____Structured Data (ex: SQL, JSON, XML)

_____Surveys and Marketing (ex: multinomial modeling)

_____Systems Administration (ex: *nix, DBA, cloud tech.)

_____Temporal Statistics (ex: forecasting, time-series analysis)

_____Unstructured Data (ex: noSQL, text mining)

_____Visualization (ex: statistical graphics, mapping, web-based viz)

13) Please tell us how you view yourself.

I think of myself as a/an ... (required)*

	Completely Agree	**Agree**	**Neutral**	**Disagree**	**Completely Disagree**
Scientist	( )	( )	( )	( )	( )
Engineer	( )	( )	( )	( )	( )
Business Person	( )	( )	( )	( )	( )
Artist	( )	( )	( )	( )	( )
Researcher	( )	( )	( )	( )	( )
Statistician	( )	( )	( )	( )	( )
Jack of All Trades	( )	( )	( )	( )	( )
Leader	( )	( )	( )	( )	( )
Entrepreneur	( )	( )	( )	( )	( )
Developer	( )	( )	( )	( )	( )
Data Scientist	( )	( )	( )	( )	( )

Working with Data

Please do not use the browser's Back button if you want to go back a page. Instead, use the survey's Back button, which you can find at the end of each page.

Validation: Must be numeric

14) We want to get a sense of your work week. In the table below, please enter roughly how many hours per week you typically spend on each of the activities.

If there are activities that you think should be included, please enter the time under "Other work activities" at the bottom of the table and list the activities in the next question.

	Hours
Query databases for existing data	
Build platforms to collect data and/or instrument code to gather data	
Prepare data: manipulating the data to fit your analytic needs, e.g., data merging, cleaning, data shaping	
Analyze data: statistics, machine learning, data mining, etc.	
Experiment: build capability and/or run experiments to test alternative designs	
Validate insight: check that analysis results are correct	
Disseminate insight: prepare presentations, write reports, create visualizations	
Engage with others about data and analysis: attend meetings, brown bags, etc.	
Operationalize insight: operationalize models, define automated actions, triggers based on models	

Act on insight: make decision based on data	
Other work activities *related* to data science	
Other work activities *not related* to data science	

15) If there is an important activity in your job role *related to data science* that we missed in the previous question, please fill it in here.

_____

_____

_____

_____

Validation: Min = 0 Must be numeric

16) How many hours per week do you typically spend in planned meetings? (decimals okay)

_____

17) What types of data do you analyze as part of your data science tasks? Check all that apply.

[ ] Business data (e.g., purchases, transactions)

[ ] Customer usage of the product (e.g., SQM data, feature usage, game play data)

[ ] Execution behavior of the product (e.g., crashes, performance data, load balancing)

[ ] Engineering data of the product (e.g., check-ins, work items, code reviews)

[ ] Survey data (e.g., from customer surveys)

[ ] Other (separate multiple data types with comma):

_____

18) What tools do you use for your data science tasks?

[ ] R
[ ] SPSS
[ ] JMP
[ ] Matlab
[ ] Minitab
[ ] Python
[ ] SQL
[ ] Excel
[ ] Office BI
[ ] Azure ML
[ ] Scope
[ ] TLC
[ ] C, C++, C#
[ ] Other (separate multiple tools with comma): _____

19) Please give an example of a problem related to data science that you worked on in the last six months. Describe what data you used for what purpose.

For example: "I worked on prioritizing which bugs to fix. I used telemetry data collected from customers."

_____
_____
_____
_____

20) What challenges do you frequently face when doing data science?

_____
_____
_____
_____

21) What new features, tools, processes, or best practices could improve how we do data science at Microsoft?

_____

_____

_____

_____

22) What advice related to data science would you give a friend who is looking to get started with data science?

_____

_____

_____

_____

Interacting with Data

Please do not use the browser's Back button if you want to go back a page. Instead, use the survey's Back button, which you can find at the end of each page.

23) Do you have access to the following data?

	Yes	No	Don't know
Business data (e.g., purchases, transactions)	( )	( )	( )
Customer usage of the product (e.g., SQM data, feature usage, game play data)	( )	( )	( )
Execution behavior of the product (e.g., crashes, performance data, load balancing)	( )	( )	( )
Engineering data of the product (e.g., check-ins, work items, code reviews)	( )	( )	( )

24) Do you participate in the generation of the following data?

	No, I don't	I help choose what gets recorded	I add logging or other instrumentation to the code
Business data (e.g., purchases, transactions)	( )	( )	( )
Customer usage of the product (e.g., SQM data, feature usage, game play data)	( )	( )	( )
Execution behavior of the product (e.g., crashes, performance data, load balancing)	( )	( )	( )
Engineering data of the product (e.g., check-ins, work items, code reviews)	( )	( )	( )

25) How often do you view reports, dashboards, or other presentations based on the following data?

	Often (e.g. lobby kiosk, daily report)	**Regularly** (e.g. weekly meeting)	**Occasionally** (e.g. when someone has a question)	**Rarely**	**Never**
Business data (e.g., purchases, transactions)	( )	( )	( )	( )	( )
Customer usage of the product (e.g., SQM data, feature usage, game play data)	( )	( )	( )	( )	( )
Execution behavior of the product (e.g., crashes, performance data, load balancing)	( )	( )	( )	( )	( )
Engineering data of the product (e.g., check-ins, work items, code reviews)	( )	( )	( )	( )	( )

26) How often do you personally analyze the following data?

	Often (e.g. I produce a report)	**Regularly** (e.g. on a repeated basis)	**Occasionally** (e.g. when someone has a question)	**Rarely**	**Never**
Business data (e.g., purchases, transactions)	( )	( )	( )	( )	( )

Customer usage of the product (e.g., SQM data, feature usage, game play data)	( )	( )	( )	( )	( )
Execution behavior of the product (e.g., crashes, performance data, load balancing)	( )	( )	( )	( )	( )
Engineering data of the product (e.g., check-ins, work items, code reviews)	( )	( )	( )	( )	( )

27) Does your team use a standard framework for gathering data? Check all that apply.

[ ] Windows Events

[ ] Watson crashes

[ ] RAC

[ ] SQM

[ ] Asimov

[ ] Azure Event Hub

[ ] Azure MDS

[ ] Other: _____

[ ] I don't know

Validating Data and Insight

28) How do you ensure that the input data to your analysis is correct?

_____

_____

_____

_____

29) How do you ensure that you have high confidence about your analysis results?

_____

_____

_____

_____

30) The following statements are about debugging in the context of developing big data systems. Please indicate how you agree with following statements

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	I Don't Know	Not applicable
Distributed computation in Big Data Cloud Computing platforms (Cosmos, Dryad, Hadoop, Map Reduce, Spark) makes it difficult to see where and when a failure occurred.	( )	( )	( )	( )	( )	( )	( )
Bugs in big data systems are more difficult to find than bugs in traditional software systems.	( )	( )	( )	( )	( )	( )	( )
Errors in big data systems have higher impact than the bugs in traditional software systems due to cloud computing costs.	( )	( )	( )	( )	( )	( )	( )

It is difficult to trace the origin of failure-inducing data in cloud computing platforms.	( )	( )	( )	( )	( )	( )	( )
Lack of interactive debugging experiences in cloud computing makes the debugging process harder.	( )	( )	( )	( )	( )	( )	( )
It is difficult to conduct iterative trial-error type debugging while processing large data in cloud computing platform due to high computation cost and time.	( )	( )	( )	( )	( )	( )	( )
Many big data applications are hard to debug due to lack of predefined tests.	( )	( )	( )	( )	( )	( )	( )
Big data system debugging is challenging because we not only need to identify failure-inducing data, but also failure-inducing logic in code	( )	( )	( )	( )	( )	( )	( )

31) (This is the last question) If you have any comments regarding this survey in general, please enter them in the following text box.

_____

_____

_____

_____

Thank You!

Thank you for taking our survey. Your response is very important to us.

As another way of saying thanks, we're raffling off four $125 Visa Gift Cards (official rules of the sweepstakes).

Click here to enter the raffle by email

We're interested in following up with people on data science.

Click here to email us if you would be willing to talk with us

For Additional Information

Learn more about the Empirical Software Engineering group at Microsoft Research.

Recent papers:
The Emerging Role of Data Scientists on Software Development Teams

Analyze This! 145 Questions for Data Scientists in Software Engineering

Software Developers' Perceptions of Productivity
