2013 Data Science Salary Survey

Tools, Trends, What Pays (and What Doesn't) for Data Professionals

John King & Roger Magoulas

O'REILLY®

O'REILLY®
Strata
Making Data Work

Take the Strata Data Science Salary and Tools Survey

As data scientists and statisticians—as professionals who like nothing better than petabytes of rich data—we find ourselves in a strange spot: We know very little about ourselves.

But that's changing. This salary and tools survey is the second in an annual series. To keep the insights flowing, we need one thing: **People like you to take the survey.** Anonymous and secure, the survey will continue to provide insight into the demographics, work environments, tools, and compensation of practitioners in our field.

We hope you'll consider it a civic service. We hope you'll participate today.

Take the Survey ▶

2013 Data Science Salary Survey

Tools, Trends, What Pays (and What Doesn't) for Data Professionals

John King and Roger Magoulas

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'REILLY®

2013 Data Science Salary Survey

by John King and Roger Magoulas

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

January 2014: First Edition

Revision History for the First Edition:

2014-01-13: First release

While every precaution has been taken in the preparation of this book, the publisher and authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-1-491-94914-6

[LSI]

Building Data Science Teams
DJ Patil
What is Data Science?
Mike Loukides
Data Jujitsu
DJ Patil
Planning for Big Data
A CIO's handbook to the changing data landscape
O'Reilly Radar Team
O'REILLY

Table of Contents

2013 Data Science Salary Survey

Executive Summary

O'Reilly Media conducted an anonymous salary and tools survey in 2012 and 2013 with attendees of the Strata Conference: Making Data Work in Santa Clara, California and Strata + Hadoop World in New York. Respondents from 37 US states and 33 countries, representing a variety of industries in the public and private sector, completed the survey.

We ran the survey to better understand which tools data analysts and data scientists use and how those tools correlate with salary. Not all respondents describe their primary role as data scientist/data analyst, but almost all respondents are exposed to data analytics. Similarly, while just over half the respondents described themselves as technical leads, almost all reported that some part of their role included technical duties (i.e., 10–20% of their responsibilities included data analysis or software development).

We looked at which tools correlate with others (if respondents use one, are they more likely to use another?) and created a network graph of the positive correlations. Tools could then be compared with salary, either individually or collectively, based on where they clustered on the graph.

We found:

- By a significant margin, more respondents used SQL than any other tool (71% of respondents, compared to 43% for the next highest ranked tool, R).
- The open source tools R and Python, used by 43% and 40% of respondents, respectively, proved more widely used than Excel (used by 36% of respondents).
- Salaries positively correlated with the number of tools used by respondents. The average respondent selected 10 tools and had a median income of $100k; those using 15 or more tools had a median salary of $130k.
- Two clusters of correlating tool use: one consisting of open source tools (R, Python, Hadoop frameworks, and several scalable machine learning tools), the other consisting of commercial tools such as Excel, MSSQL, Tableau, Oracle RDB, and BusinessObjects.
- Respondents who use more tools from the commercial cluster tend to use them in isolation, without many other tools.
- Respondents selecting tools from the open source cluster had higher salaries than respondents selecting commercial tools. For example, respondents who selected 6 of the 19 open source tools had a median salary of $130k, while those using 5 of the 13 commercial cluster tools earned a median salary of $90k.

> We suspect that a scarcity of resources trained in the newer open source tools creates demand that bids up salaries compared to the more mature commercial cluster tools.

Salary Report

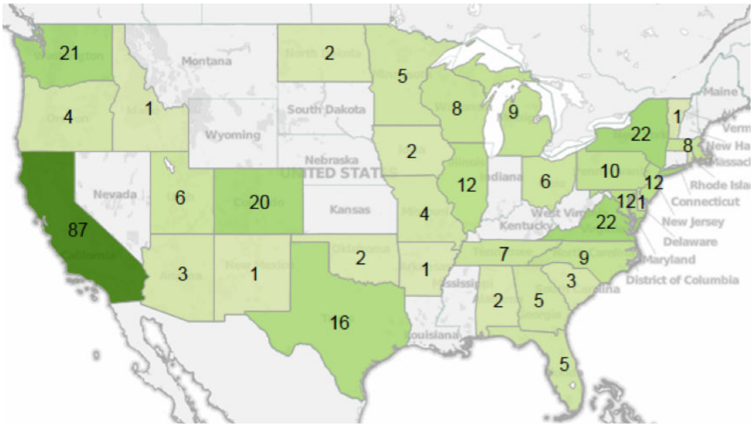
Big data can be described as both ordinary and arcane. The basic premise behind its genesis and utility are as simple as its name: efficient access to more—*much more*—data can transform how we understand and solve major problems for business and government. On the other hand, the field of big data has ushered in the arrival of new, complex tools that relatively few people understand or have even heard of. But is it worth learning them?

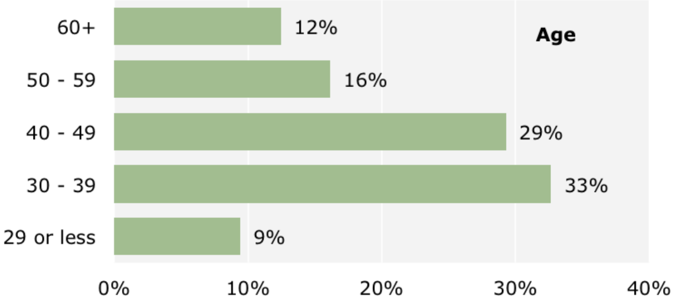
If you have any involvement in data analytics and want to develop your career, the answer is *yes*. At the last two Strata conferences (New York 2012 and Santa Clara 2013), we collected surveys from our attendees about, among other things, the tools they use and their salaries. Here's what we found:

- Several open source tools used in analytics such as R and Python are just as important, or even more so, than traditional data tools such as SAS or Excel.
- Some traditional tools such as Excel, SAS, and SQL are used in relative isolation.
- Using a wider variety of tools—programming languages, visualization tools, relational database/Hadoop platforms—correlates with higher salary.
- Using more tools tailored to working with big data, such as MapR, Cassandra, Hive, MongoDB, Apache Hadoop, and Cloudera, also correlates with higher salary.

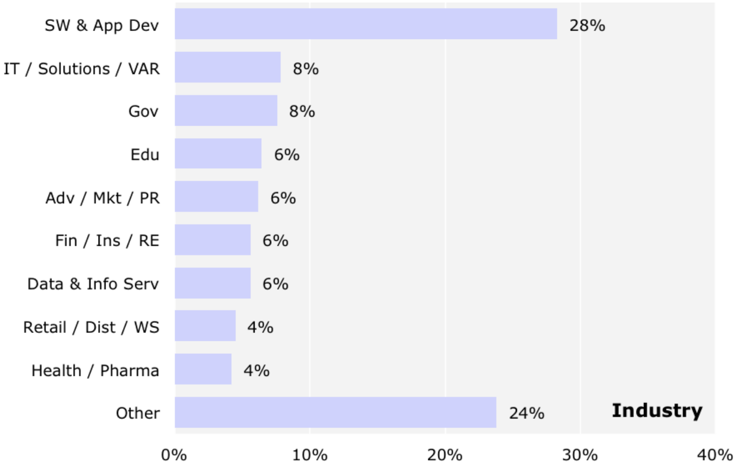
We should note that Strata attendees comprise a special group and do not form an unbiased sample of everyone who seriously works with data. These are people deeply involved with or interested in big data, seeking to network with others on the field's cutting edge and learn about the new technologies defining it—in short, they are ahead of the curve. If a trend observed in the sample is *not* consistent with what *would* be observed in the larger population (of analysts, data scientists, and so on), then this trend could represent the direction big data is headed. This is likely to be the case for tool usage.

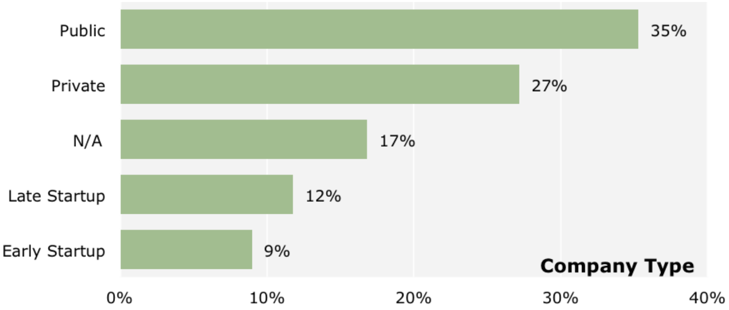
The majority of the survey's respondents were from the US, with most of the rest coming from Canada and Europe. Among those from the US, 68% were from states on either coast.





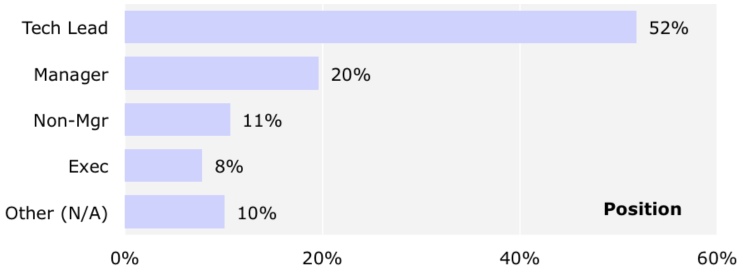
Our sample represented a wide range of ages, with most respondents in their thirties and forties. About 40% of respondents were based in the West, while the rest of the respondents were evenly distributed in the Northeast, Mid-Atlantic, South, and Midwest regions. California, Maryland, and Washington had the highest median salaries, while respondents in the South and Midwest reported the lowest median salaries.

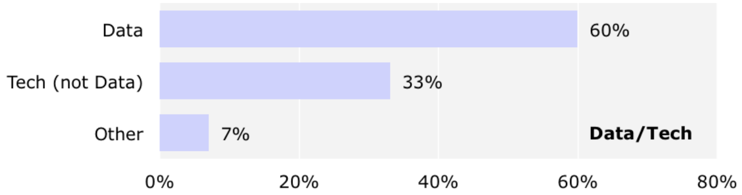




Twenty-three industries were represented (those with at least 10 respondents are shown above) and about one-fifth came from startups. A significant share of respondents, 42%, work in software-oriented segments: software and application development, IT/solutions/VARs, data and information services, and manufacturing/design (IT/OEM). Government and education represent 14% of respondents.[1] About 21% of those responding work for startups—with early startups, surprisingly, showing the highest median salary, $130k. Public companies had a median salary of $110k, private companies $100k and N/A (mostly government and education) at $80k.

[1] 60% of government and education respondents selected the "not applicable" category for company type.



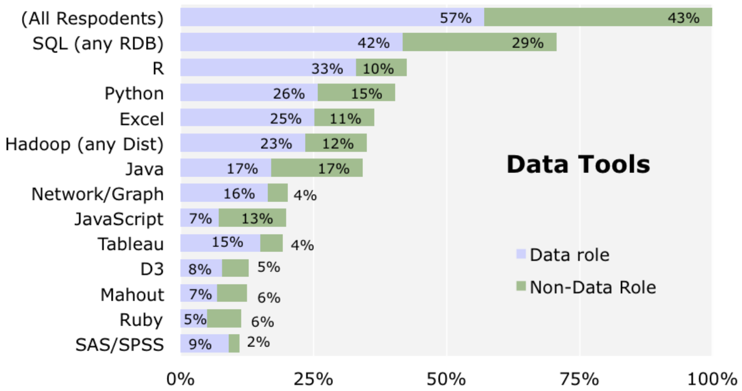


Most respondents (56%) describe themselves as data scientists/analysts. Choosing from four broad position categories—non-managerial, tech lead, manager, and executive—over half of the respondents reported their position as technical lead. The survey asked respondents to describe what share of their jobs was spent on various technical and analytic roles: 80% of respondents spend at least 40% of their time on roles like statistician, software developer, coding analyst, tech lead, and DBA. In other words, this was a very technical crowd—even those who were primarily managers and executives.

Tool Usage

The chart below shows the usage rate for the most commonly used tools. To show *who* these users are, for each tool, the share of respondents who use the tool and self-describe as primarily data analysts are shown in blue; those who use the tool and are not primarily data analysts are shown in green.[2]

[2] SQL/Relational Databases and Hadoop are categories of tools: respondents are included in their usage counts if they reported using at least one tool from the categories. The SQL/RDB list consists of 18 tools, the Hadoop list consists of 9.



That SQL/RDB is the top bar is no surprise: accessing data is the meat and potatoes of data analysis, and has not been displaced by other tools. The preponderance of R and Python usage is more surprising —operating systems aside, these were the two most commonly used individual tools, even above Excel, which for years has been the go-to option for spreadsheets and surface-level analysis. R and Python are likely popular because they are easily accessible and effective open source tools for analysis. More traditional statistical programs such as SAS and SPSS were far less common than R and Python.

By counting tool usage, we are only scratching the surface: who exactly uses these tools? In comparing usage of R/Python and Excel, we had hypothesized that it would be possible to categorize respondents as users of one or the other: those who use a wider variety of tools, largely open source, including R, Python, and some Hadoop, and those who use Excel but few tools beside it.

Python and R correlate with each other—a respondent who uses one is more likely to use the other—but neither correlates with Excel (negatively or positively): their usage (joint or separate) does not predict whether a respondent would also use Excel. However, if we look at *all* correlations between *all* pairs of tools, we can see a pattern that, to an extent, divides respondents. The significant positive correlations can be drawn as edges between tools as nodes, producing a graph with two main clusters.[3]

[3] Correlations were tested using a Pearson's chi square test with p=.05.

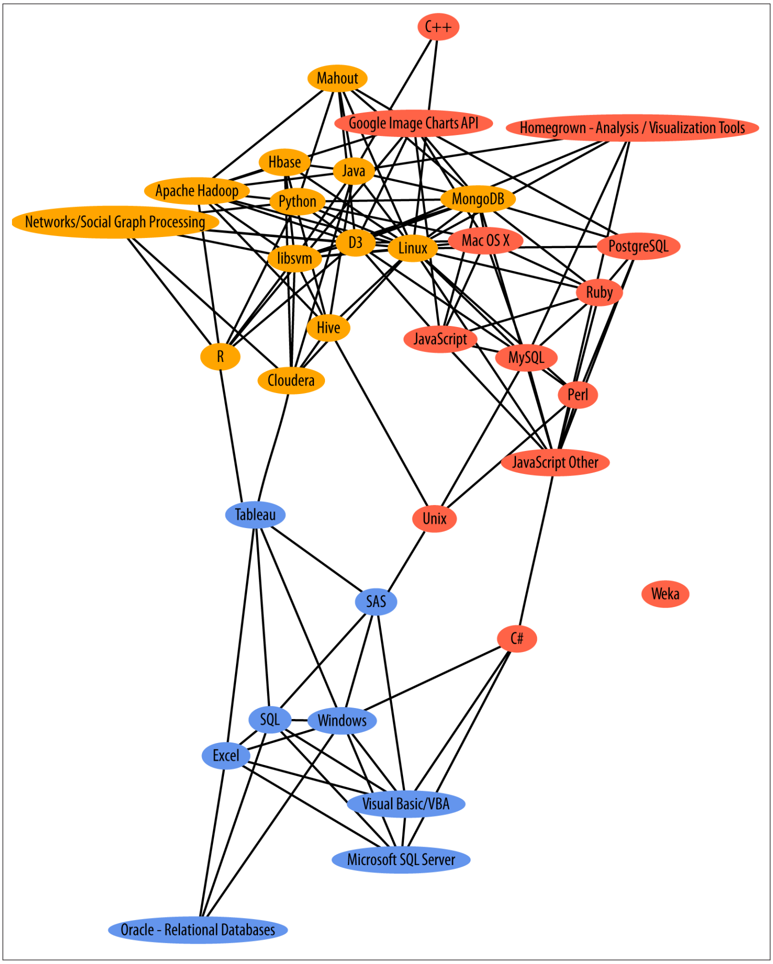


Figure 1. Tool correlations for tools with at least 40 users

One of the clusters, which we will refer to as the "Hadoop" group (colored orange in Figure 1), is dense and large: it contains R, Python, most of the Hadoop platforms, and an assortment of machine learning, data management, and visualization tools. The other—the "SQL/Excel" group, colored blue—is sparser and smaller than the Hadoop group, containing Excel, SAS, and several SQL/RDB tools. For the sake of comparison, we can define membership in these groups by the largest set of tools, each of which correlates with at least one-third of the others; this results in a Hadoop group of 19 tools and a SQL/Excel group of 13 tools.[4] Tools in red are in neither of the two major clusters, but most of these clearly form a periphery of the Hadoop cluster.

The two clusters have no tools in common and are quite distant in terms of correlation: only four positive correlations exist between the two sets (mostly through Tableau), while there are a whopping 51 negative correlations.[5] Interestingly, each cluster included a mix of data access, visualization, statistical, and machine learning–ready tools. The tools in each cluster are listed below.

Tools in the Hadoop Cluster

- Linux
- R
- Python
- Java
- D3
- Mahout
- Pig
- Hive
- MongoDB
- Hbase
- LIBSVM
- Cloudera
- Cassandra
- MapR
- Pentaho
- Amazon EMR
- Apache Hadoop
- Networks/Social Graph Processing
- IBM SystemML and Nimble

[4] This criteria for membership is somewhat arbitrary, especially for the Hadoop cluster —the level of internal connectedness increases gradually from the periphery to the core. For example, with a stricter (higher) proportion, we would define multiple, smaller, overlapping "Hadoop" clusters that span the previously defined cluster (proportion=.33), and include a number of other tools. The proportion of one third was chosen because the resulting sets are dense enough to be meaningful, they are unique (only one such set exists for each cluster, and these two sets are disjoint), and most tools with many users are included in at least one of them (e.g., 69% of tools with >50 users). Note that the graph shows only tools with at least 40 users, but we are considering all tools in the tool clusters. Most of the tools left out of the graph would be in red, but about a third of each cluster is not shown.

[5] A negative correlation between two tools X and Y means that if a respondent uses X, she is less likely to use Y as well. Of the 3,570 tools pairs, 141 have negative correlations —about 4%. Compare this to 51 negative correlations between the 247 pairs between the two clusters.

Tools in the SQL/Excel Cluster	
Windows	Microsoft SQL Server
Excel	Oracle RDB
SQL	Visual Basic/VBA
Tableau	BusinessObjects
SAS	Cognos
IBM DB2	Netezza (IBM)
Teradata	

The two clusters show a significant pattern of tool usage tendencies. No respondent reported using *all* tools in either cluster, but many gravitated toward one or the other—much more than expected if no correlation existed. In this way, we can usefully categorize respondents by counting how many tools from each cluster a respondent used, and then we can see how these measures interact with other variables.

One pattern that follows logically from the asymmetry of the two clusters involves the total number of tools a respondent uses.[6] Respondents who use more tools in the Hadoop cluster—the larger and denser of the two—are more likely to use more tools *in general* (shown in Figure 2).

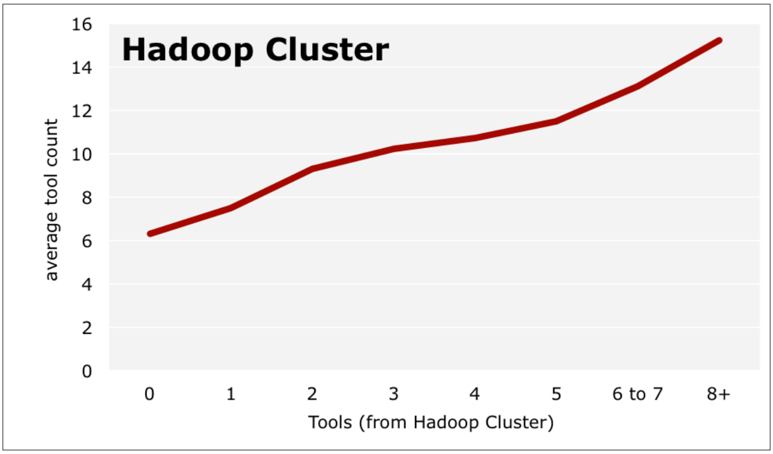


Figure 2. Tools (from Hadoop cluster)

[6] The total number of tools used by each respondent roughly followed a normal distribution, with a mean of 10.0 tools and a standard deviation of 3.7.

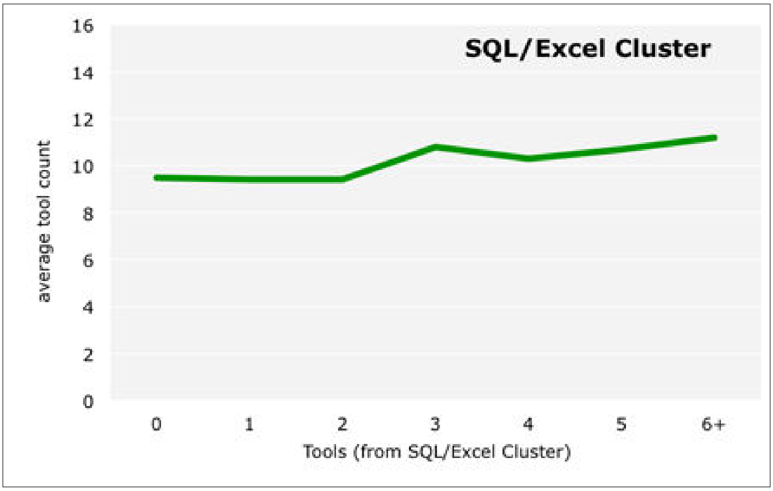


Figure 3. Tools (from SQL/Excel cluster)

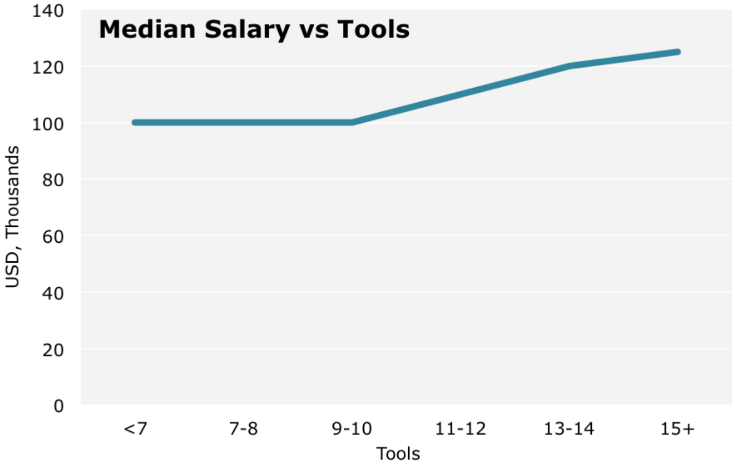
Figure 2 and Figure 3 can be read as follows: in each graph, all respondents are grouped by the number of tools they use from the corresponding cluster; the bars show the average number of tools used (counting any tool) by the respondents in each group.[7] While the bars rise in both graphs, it should be remembered that a positive correlation would be expected between these variables.[8] In fact, the real deviation is in the SQL/Excel graph, which is much flatter than we would expect. This pattern confirms what we could guess from the correlation graph: respondents using more tools from the SQL/Excel cluster use few tools from outside it.

Whether or not this matters is another question: it may be possible for some analysts, for example, to rely on tools taken only from the SQL/Excel cluster to perform their tasks. However, our data shows that using more tools generally correlates with a higher salary. The following graph shows the median base salary of respondents using a certain

[7] These bins were chosen to have a sufficient number of respondents in each.

[8] Both variables are counting tools: each total tool count value contributing to the average (for the y-value) cannot be less than the in-cluster count (the x-value). A similar graph using a random set of tools would almost always produce a rising pattern, albeit not as steep as the one shown by the Hadoop cluster.

number of tools. Median base salary is constant at $100k for those using up to 10 tools, but increases with new tools after that.[9]



Given the two patterns we have just examined—the relationships between cluster tools and respondents' overall tool counts, and between tool counts and salary—it should not be surprising that there is a significant difference in how each cluster correlates with salary. Using more tools from the Hadoop cluster correlates positively with salary, while using more tools from the SQL/Excel cluster correlates (slightly) negatively with salary.

[9] Salary figures are for US respondents only.

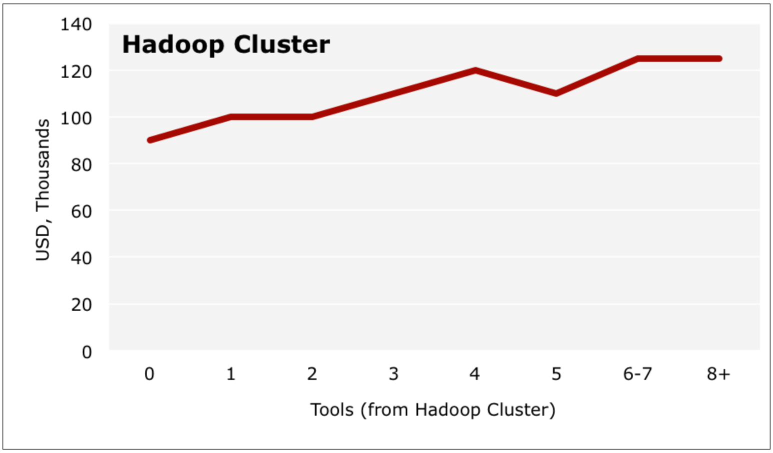


Figure 4. Tools (from Hadoop cluster)

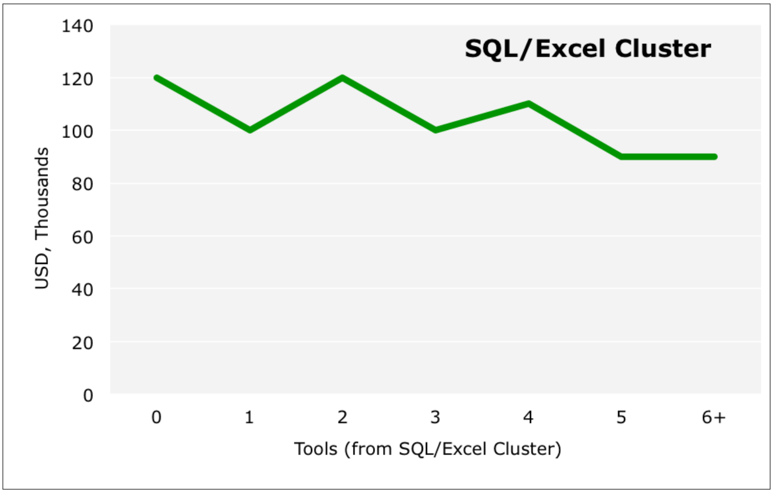
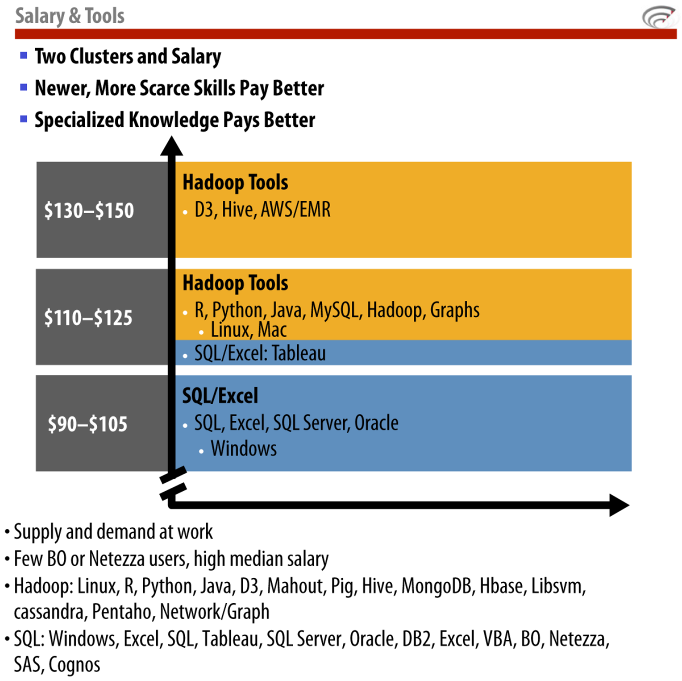


Figure 5. Tools (from SQL/Excel cluster)

Median base salary generally rises with the number of tools used from the Hadoop cluster, from $85k for those who do not use any such tools to $125k for those who use at least six. The graph for the SQL/Excel cluster is less conclusive. The variation in median salary in the lower range of tool usage seems to vary randomly, although there is a definite drop for those using five or more SQL/Excel cluster tools.

The same pattern can be seen in a different way by looking at tool usage versus salary on a tool-by-tool basis. The median base salary of all US-based respondents was $110,000, against which we can compare the median salaries of those respondents who use a given tool.[10]



Salary & Tools

- **Two Clusters and Salary**
- **Newer, More Scarce Skills Pay Better**
- **Specialized Knowledge Pays Better**

$130–$150

Hadoop Tools
- D3, Hive, AWS/EMR

$110–$125

Hadoop Tools
- R, Python, Java, MySQL, Hadoop, Graphs
  - Linux, Mac
- SQL/Excel: Tableau

$90–$105

SQL/Excel
- SQL, Excel, SQL Server, Oracle
  - Windows

- Supply and demand at work
- Few BO or Netezza users, high median salary
- Hadoop: Linux, R, Python, Java, D3, Mahout, Pig, Hive, MongoDB, Hbase, Libsvm, cassandra, Pentaho, Network/Graph
- SQL: Windows, Excel, SQL, Tableau, SQL Server, Oracle, DB2, Excel, VBA, BO, Netezza, SAS, Cognos

Tools in the blue boxes are from the SQL/Excel cluster, tools in orange boxes are from the Hadoop cluster. Of the 26 tools with at least 10 users that "have" a median salary above $110k—that is, the median salary of the users is above $110k—12 are from the Hadoop cluster, but only 3 are from the SQL/Excel cluster (Tableau and the lightly used BusinessObjects and Netezza). Conversely, out of 12 tools with median salaries below $110k, 7 are from the SQL/Excel cluster, while *none* are from the Hadoop cluster.

We must be careful in jumping to conclusions: correlations between salary and tool *usage* do not necessary equate to salary trends before and after *learning* a tool. For example, we can expect that learning tools from the SQL/Excel cluster does not *decrease* salary.

[10] Only tools used by at least 10 US-based respondents are considered here; tools with lower usage counts may not produce reliable medians.

Other variables could affect both tool usage and salary. For example, more respondents from startups had salaries above $110k (53%) than other company types (41%), and they tended to use more tools from the Hadoop cluster and fewer from the SQL/Excel cluster. However, having 21% of respondents working for startups mutes their effect on the overall survey. No other variables in the survey were found to influence these patterns.

Even considering the issues above, it seems very likely that knowing how to use tools such as R, Python, Hadoop frameworks, D3, and scalable machine learning tools qualifies an analyst for more highly paid positions—more so than knowing SQL, Excel, and RDB platforms. We can also deduce that the more tools an analyst knows, the better: if you are thinking of learning a tool from the Hadoop cluster, it's better to learn several.

The tools in the Hadoop cluster share a common feature: they all allow access to large data sets and/or support analysis of large data sets. The demand for analysts who know how to work with large data sets is growing, in particular for those who can perform more advanced machine learning, graph and real-time tasks on large data sets. Until the supply of such analysts catches up, their salaries will naturally be bid up.

Our data illustrates a landscape of data workers that tend toward one of two patterns of tool usage: knowing a large number of newer, more code-heavy, scalable tools—which often means higher salary—or knowing smaller numbers of more traditional, query-based tools.

The survey results help address whether data analysts need to code—coding skills are not necessary but provide access to cutting-edge tools that *can* lead to higher salaries. While the survey shows that tools in the SQL/Excel group are widely used, those who can code and know tools that handle larger data sets tend to earn higher salaries.

As exceptions to the broader pattern, three tools in the SQL/Excel cluster—Tableau, Business Objects, and Netezza—did correlate with higher salaries (Business Objects and Netezza had few users). Tableau is an outlier in the correlation graph, somewhat bridging the two clusters, as Tableau correlated with R, Cloudera, and Cassandra usage. We placed Tableau in the SQL/Excel cluster based on the cluster definitions, but we could also have excluded Tableau from both groups; this would have created an even stronger correlation between the clusters and salary (i.e., raising the Hadoop cluster salary, reducing the SQL/

Excel salary), as Tableau is one of the few SQL/Excel tools that correlates positively with salary.

Open source tools such as R and Python are not popular just because they are free—they are powerful and flexible and can make a big difference in what an analyst can do. Furthermore, their usage has expanded enough that employers are likely to begin assuming their knowledge when considering job candidates. As for Hadoop, it is not a fad: new technologies that handle Big Data are transformative, and those who know how to operate them should be among the most in-demand workers of our increasingly data-driven society.

Conclusion

While the results of this survey clearly indicate certain patterns of tool usage and salary, we should remember some of the limitations of this data. Sampled from attendees at two conferences, these results capture a particular category of professionals: those who are heavily involved in big data or highly motivated to become so, often using the most advanced tools that the industry has to offer. This study shows one perspective of modern data science, but there are others.

We would like to continue this study in several ways. Comparing these results with data from job postings, or more in-depth investigations of individuals' exact tool usage within their workflow, could expand our findings in interesting ways. More fundamentally, we will continue to ask our Strata attendees about their tool usage at subsequent conferences. Some new tools with only a handful of users among the respondents at last year's event would be expected to have dozens this time around. The required tasks of big data change rapidly, requiring ongoing attention to how these changes are reflected in the data tool landscape.

About the Authors

John King is a data analyst at O'Reilly Media. Having previously worked on survey-based sociolinguistic research in the Republic of Georgia, he now runs surveys at O'Reilly, using the results not just for internal use but also to share his findings with the public.

Roger Magoulas is research director at O'Reilly Media and co-chair of the Strata conferences. Roger and his team build the analytic infrastructure and provide analysis services, including technology trend analysis, to business decision makers at O'Reilly and beyond.