Big Data and Advanced Analytics Survey 2014

Evans Data Corporation
340 Soquel Avenue
Santa Cruz, CA 95062
800-831-3080
www.evansdata.com

Table of Contents

EXECUTIVE SUMMARY 7

OVERVIEW 15

Objectives of the Survey 15
Survey Methodology 15
Research Design 15
Relative Rankings 15
The Sample – Software Developers 15
The EDC Panel 16
Other Evans Data Corp. Services 16
Multi-Client Survey Series 16
Tactical Survey Reports 18
Custom Surveys 18
Targeted Analytics 18

DRIVERS AND BARRIERS AND RISKS FOR BIG DATA 19

Top Drivers for Pursuing Big Data Projects 20
Top Drivers by Developer Segment 21
Top Drivers by Company Size 22
Top Barriers for using Big Data 23
Top Barriers for using Big Data by Company Size 25
Largest Risks from Using Big Data 26
Largest Risks from Using Big Data by Developer Segment 27
Largest Risks from Using Big Data by Company Size 28
Most Important Areas for Vendor Action to Help Secure Big Data 29
Traditional Security Mechanisms and Big Data 30

THE BIG DATA LANDSCAPE 31

Type of Information Feeding Data Stores 32
Type of Information Feeding Data Stores by Developer Segment 34
Type of Information Feeding Data Stores by Company Size 35
Departments Using Big Data Solutions 36
Departments Using Big Data Solutions by Company Size 37
Funding Source for Big Data Projects 38
Data Types for Big Data 39
Data Types by Company Size 40
Data Types by Departments Using Big Data Solutions 41
Homogeneity of Big Data Types 42
Homogeneity by Data Types 43
Target Industries for Big Data 44
Initial Reason for Transitioning to Big Data Solutions 46
Top Inhibitor for Transforming from Proof of Concept to Production 47

BIG DATA AND THE INTERNET OF THINGS 48

Client Devices Supported by Big Data 49
Connected Device Interaction with Big Data 50
Data Collected by Connected Devices 51
Greatest Difficulties in Big Data Implementation with Connected Devices 52
Connected Device Production 53
Concerns with the Internet of Things 54
Concerns with the Internet of Things for Developers in Small Companies.....55
Concerns with the Internet of Things for Developers in Mid-Size Companies....56
Concerns with the Internet of Things for Developers in Large Companies.....57

REAL TIME EVENT PROCESSING 58

Familiarity with Complex Real Time Products 59
Importance of Real Time Complex Event Processing 60
Importance of Real Time Complex Event Processing by Company Size 61
Relevance of Real Time Analytics for Organization's Needs 62
Relevance of Real Time Analytics for Organization's Needs by Company Size...63
Relevance of Real Time Analytics for Development Work 64
Expected Change in Real Time Event-based App Use 65
Most Important Types of Data to Analyze in Real Time 66

ADVANCED ANALYTICS TOOLS AND SERVICES 67

Data Feeds for Predictive Analytics 68
Network Provider Data Feeds for Predictive Analytics 69
Plans for Using Analysis Tools for Particular Tasks 70
Types of Tools Used for Analytics 71
Most Requested Improvement to Data and Analytics Tools 72
Most Requested Improvement to Data and Analytics Tools by Company Size...74
Primary Analytics Tools Being Used 75
Primary Analytics Tools Being Used by Company Size 76
Primary Analytics Tools Being Used by Most Requested Improvement 77
Big Data Analytics Provider Types Used 78
Big Data Analytics Provider Types Used by Company Size 79
Primary Reason for Choosing Predictive Analytics Tools and Vendors..... 80
Primary Reason for Choosing Predictive Analytics Tools and Vendors by Analytics Tools Used 81
Primary Reason for Choosing Predictive Analytics Tools and Vendors by Providers Used 82
Level of Interest in Analytics-as-a-Service Offerings 83
Level of Interest in Analytics-as-a-Service Offerings 84
Importance of Various Aspects of Data and Analytics Work 85
Acceptable Downtime for Big Data Solution Landscape 86
Acceptable Downtime for Big Data Solution Landscape by Company Size.....87
Necessary Features of an Analytics-as-a-Service Offering 88

DATA VISUALIZATION AND GRAPHICAL MODELING TOOL USE 89

Primary Data Visualization Tools Being Used 90

Primary Data Visualization Tool Use by Team Size 91
Primary Data Visualization Tool Use by Departments Using Big Data 92
Reason for Using Data Visualization Tool 93
Reason for Using Data Visualization Tool by Company Size 94
Importance of Graphical Modeling Tools for Managing Data Flow 95
Frequency of Using Graphical Modeling Tools for Managing Data Flow 96

DATABASE TECHNOLOGIES 97

Top Factors Affecting Database Technology Choice 98
Top Factors Affecting Database Technology Choice by Company Size 99
Best Database Architecture for Big Data 100
Types of Database Technologies In Use 101
Types of Database Technologies In Use by Company Size 102
Types of Database Technologies In Use by Primary Consideration When Choosing Database Technology 103
Percent of App Data Stored in Relational Data Structures 104
Percent of App Data Stored in Relational Data Structures by Company Size .. 105
Non-Relational Databases In Use 106
NoSQL Data Systems in Use 107
Plans for Specific Document Oriented Databases 108
Plans for In-Memory Databases, Caches, Appliances 109
Uses for Apps that Rely on In-Memory Databases 110
Uses for Apps that Rely on In-Memory Databases by Developer Segment 111
Fault Tolerance Processes Used 112
Fault Tolerance Processes Used by Company Size 113
Types of Database Instruction Architecture Used by Apps 114
Types of Database Instruction Architecture Used by Apps by Company Size .. 115
Experience with Sharding 116
Experience with Sharding by Company Size 117
Greatest Benefits of Sharding Data 118
Largest Hurdles for Sharding Data 119

DATA STORAGE AND THE CLOUD 120

Data Threshold for Switching to Big Data Solutions 121
Data Threshold by Primary Consideration When Choosing a Database 122
Integration with Enterprise Data Warehouse 123
Where are Big Data Applications Implemented? 124
Where are Big Data Applications Implemented? by Company Size 125
Servers Vendors Used for Private Cloud 126
Server Vendors Used for Private Cloud by Company Size 127
Anticipated Growth of Data Stores 128
Anticipated Growth of Data Stores by Company Size 129
Use of Compute Cluster for Storage and Computation 130
Use of Compute Cluster by Developer Segment 131
Number of Processors Used in Compute Cluster 132
Most Likely Processing Procedure 133
Preferred Interactivity Modalities: GUI-based vs. Script/Programming-based 134

Percent of Data Stored On Site vs. Off Site 135
Percent of Data Stored On Site vs. Off Site by Small Companies..... 136
Percent of Data Stored On Site vs. Off Site by Mid-Sized Companies..... 137
Percent of Data Stored On Site vs. Off Site by Large Companies 138
Expected Change in Data Storage Distribution 139
Expected Change in Data Storage Distribution by Company Size 140
Average Size of Database / Dataset 141
Average Size of Database/Dataset by Company Size 142

HADOOP 143

Plans for Using Hadoop Clusters..... 144
Number of Servers in Hadoop Cluster..... 145
Number of Servers in Hadoop Cluster by Company Size 146
Memory Required by Servers in Hadoop Cluster 147
Memory Required by Servers in Hadoop Cluster by Company Size 148
Types of Tools Used for Data Stored in Hadoop 149
Types of Tools Used for Data Stored in Hadoop by Company Size..... 150
Use of Commercial Hadoop Distribution 151
Hadoop Provider Used 152
Plans for Hadoop Data Lakes..... 153
Hadoop Analytics Integration..... 154
Likelihood of JavaScript Use with Hadoop 155
Analytics Software Used for Hadoop-based Predictive Models 156
Analytics Software Used for Hadoop-based Statistical Analysis 157

GENERAL TECHNOLOGY USE..... 158

Primary Host OS..... 159
Primary Host OS by Company Size..... 160
Primary Operating System for Deployment 161
Primary Linux Distribution Targeted..... 162
Language Use..... 163
Data Language Used for Big Data 164
Big Data Solution – Vertical solution or best of breed?..... 165
Technologies Used in Big Data and Advanced Analytics 166
Technologies Used in Big Data and Advanced Analytics by Company Size..... 167
Top Challenges to Building Scalable Big Data Apps 168
Top Challenges to Building Scalable Big Data Apps by Company Size..... 169
Top Open Source Projects Used for Big Data 170
Commercial vs. Open Source Application Use 171
Willingness to Use Open Source Solutions..... 172

TESTING AND SUPPORT 173

Big Data Dataset Creation for Testing 174
Dataset Creation for Testing by Company Size 175
Tools Used with Big Data Tests 176
Testing Tools Used by Company Size 177
Testing Tools by Dataset Creation 178

Barriers for Big Data Testing 179
Barriers for Big Data Testing for Small Companies 180
Barriers for Big Data Testing for Mid-sized Companies 181
Barriers for Big Data Testing for Large Companies 182
Best Provider of Support for Big Data and Analytics 183
Best Provider of Support for Big Data and Analytics by Company Size 184

DEMOGRAPHICS 185

Developer Segment 186
Industry 187
Job Role 188
Company Size 189
Number of Developers in Team 190
Average Number of People Involved in Particular Development Activities 191
Skill Sets in Data Science Expertise 192
Annual Revenue 193
Time Spent with Databases and Analytics 194
Involvement in Big Data or Advanced Analytics Projects 195
Experience with Analytics 196

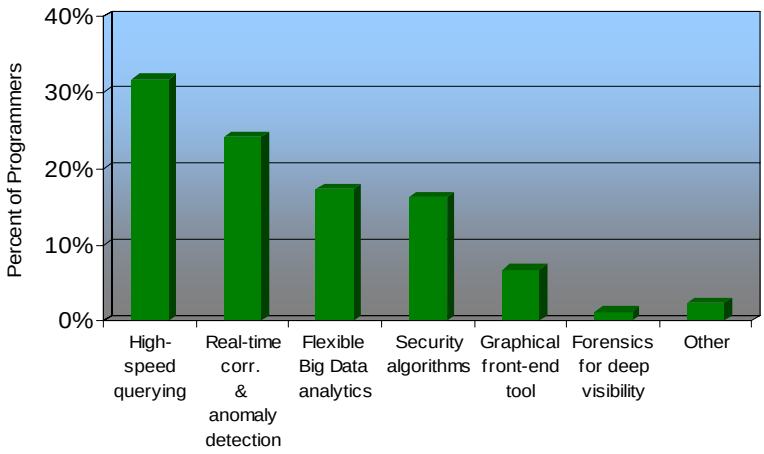
Evans Data Corp. (EDC) has made every effort to produce the highest quality research product in this effort. The customer understands that EDC uses those statistical and data gathering techniques which, in its opinion, are the most accurate possible. However, inherent in any statistical study is a possibility of error, which must be taken into account in evaluating the results. Evaluations and interpretations of statistical research findings and decisions based on them are solely the responsibility of the customer and not EDC. The conclusions, summaries and interpretations provided by EDC are based strictly on the analysis of the data gathered, and are not to be construed as recommendations; therefore EDC neither warrants their viability or accuracy nor assumes responsibility for the success or failure of any customer actions subsequently taken.

This report is the exclusive and proprietary property of Evans Data Corp. and is subject to limited distribution and restricted disclosure only. Any unauthorized use, reproduction, or transfer of this material is strictly prohibited. Clients of this study may freely distribute this report to full-time employees within their corporate entity, providing that the above requirements are adhered to by all recipients. Clients may not distribute any of this information outside of their organizations without written permission from EDC.

Most Important Areas for Vendor Action to Help Secure Big Data

Information security and maintaining the privacy of sensitive data are two of the top problems or risks that developers see in Big Data adoption, and -- as later analysis reveals -- more than 58% of Big Data developers have run into a situation in which traditional security mechanisms for data don't work with Big Data or unstructured data. Adopting the right security practices throughout their organizations and baking security into their Big Data projects is a critical component of securing Big Data, but developers may also hope for tech assistance from their Big Data vendors. In order to meet their security requirements, developers look for tools that query security intelligence data and provide flexibility in algorithms, analytics, and correlation across various types of data.

Which of the following would be MOST helpful things a vendor could supply to you to help secure big data?



Big Data & Advanced Analytics Survey © 2014 Evans Data Corp.

High-speed querying of security intelligence data is the top area for vendor action; 32% of Big Data developers report that high-speed querying is the most helpful thing a vendor could supply for them to help secure Big Data, and this consideration reprises its role atop the list. High-speed querying can help match the rate at which Big Data projects receive data, providing greater security to both structured and unstructured data.

An additional 24% of developers would find real-time correlation and anomaly detection of diverse security data to be the most helpful thing a vendor could supply for them to help secure Big Data. As diverse security streams come in, this feature would process the data and correlate the information. This area has increased nominally over the previous period.

Which of the following would be MOST helpful things a vendor could supply to you to help secure big data?

	Valid Percent
High-speed querying of security intelligence data	31.8
Real-time correlation and anomaly detection of diverse security data	24.1
Flexible big data analytics across structured and unstructured data	17.3
Security algorithms for various scenarios	16.3
Graphical front-end tool for visualizing and exploring big data	6.8
Forensics for deep visibility	1.2
Other	2.5
Total	100.0

Big Data & Advanced Analytics Survey © 2014 Evans Data Corp.

Network Provider Data Feeds for Predictive Analytics

About 7% of Big Data developers report that their company's data analytics primarily receive data from network providers. When they do, developers are slightly more likely to report that their company's predictive analytics receive data feeds from mobile apps and mobile internet usage history than they are to report that their organizations receive information regarding user's call and SMS history.

Of the 7% of Big Data developers whose company's predictive analytics primarily receive data from network providers, 76% report that their company's predictive analytics receive information about mobile apps and mobile internet usage history. Thus, these developers' companies are looking for how often a person uses an Internet radio app when in the car, or how often a person visits their favorite department store's website after a trip to the mall. Another 69% report that their companies' predictive analytics receive data about the calls and SMS sent out and received.

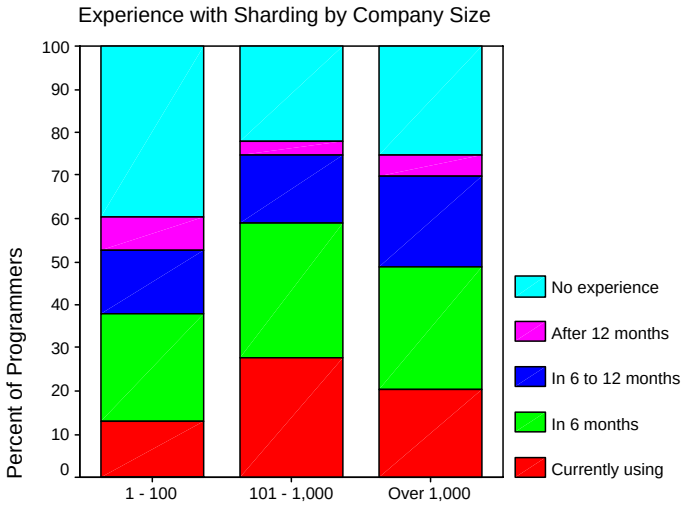
Note: This question was asked only of those developers who indicated that they perform predictive/advanced analytics based on data that is obtained from network providers.

What type of data from network providers feeds your company's predictive analytics?	Count	Percent of Responses	Percent of Cases
Mobile apps and mobile internet usage history (e.g., site/app, date, hour, duration)	32	39.5	76.2
Calls and SMS history (e.g., destination, date, hour, duration)	29	35.8	69.0
Geolocation data	18	22.2	42.9
Other	2	2.5	4.8
	-----	-----	-----
Total Responses	81	100	192.9
Big Data & Advanced Analytics Survey © 2014 Evans Data Corp.			

Note that this multiple response question allowed the developers to select as many responses as they wished, and thus the total number of cases will not come to 100%. The response column shows the percent of total responses, while the case column shows the percent of actual developers (cases) who responded.

EXPERIENCE WITH SHARDING BY COMPANY SIZE

Sharding is valuable for distributed computing applications, so it is no surprise that the medium-sized and large companies, most heavily involved with Big Data projects, are more likely than small companies to either be currently using database sharding or planning to do so.



Big Data & Advanced Analytics Survey © 2014 Evans Data Corp.

Specifically, only 13% of Big Data developers in small companies are currently using database shards, as contrasted with 28% in medium-sized companies and 20% in large companies. Forty percent of Big Data developers in small companies have no experience with sharding and no plans for sharding; 22% and 25% of those in medium-sized and large companies, respectively, report no experience with or plans for sharding.

About how many people are in your company at all locations?
*** What is your experience with data shards?**

			What is your experience with data shards?					
			Currently using them	Plan to use in 6 months	Plan to use in 6 to 12 months	Plan to use after 12 months	No experience or plans	Total
About how many people are in your company at all locations?	1 - 100	% within row	13.0%	25.0%	14.8%	7.4%	39.8%	100.0%
		% within column	14.3%	20.6%	20.3%	36.4%	34.4%	23.7%
	101 - 1,000	% within row	27.6%	31.5%	15.5%	3.3%	22.1%	100.0%
		% within column	51.0%	43.5%	35.4%	27.3%	32.0%	39.8%
	Over 1,000	% within row	20.5%	28.3%	21.1%	4.8%	25.3%	100.0%
		% within column	34.7%	35.9%	44.3%	36.4%	33.6%	36.5%
Total		% within row	21.5%	28.8%	17.4%	4.8%	27.5%	100.0%
		% within column	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Big Data & Advanced Analytics Survey © 2014 Evans Data Corp.