# Apache MADlib (Incubating)

## User Survey Results

Oct 2016

Received ~40 responses from 27 different companies

# Summary (1)

- ~50% of respondents have 1 year or less of MADlib use
- Fraud detection is the most common use case
- Regression (various), clustering and random forest are the most commonly used MADlib algorithms
- Gradient boosting is the most commonly requested new algorithm

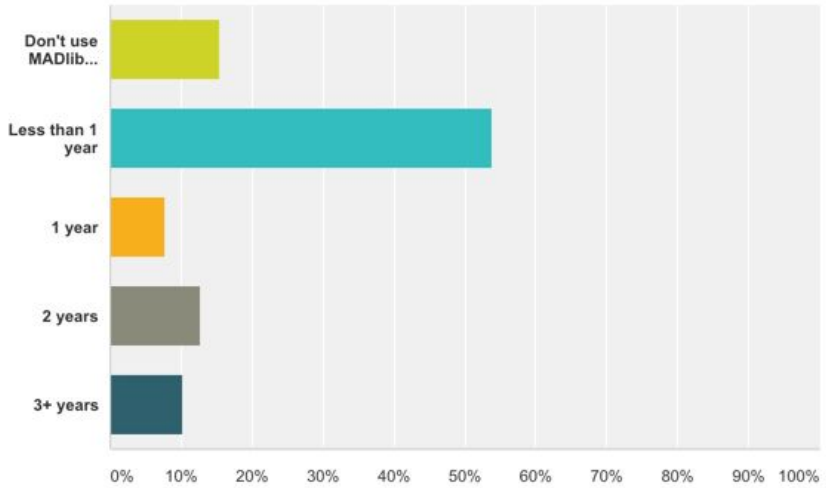
# Summary (2)

- Users prefer new algorithms more than improvements to existing algorithms by a 2:1 margin
- Improved documentation/examples and better performance are the biggest concerns
- The most common other tools used by respondents are R, Spark and Python (and associated libraries)

# Q1

## How long have you been using MADlib?

Answered: 39 Skipped: 0

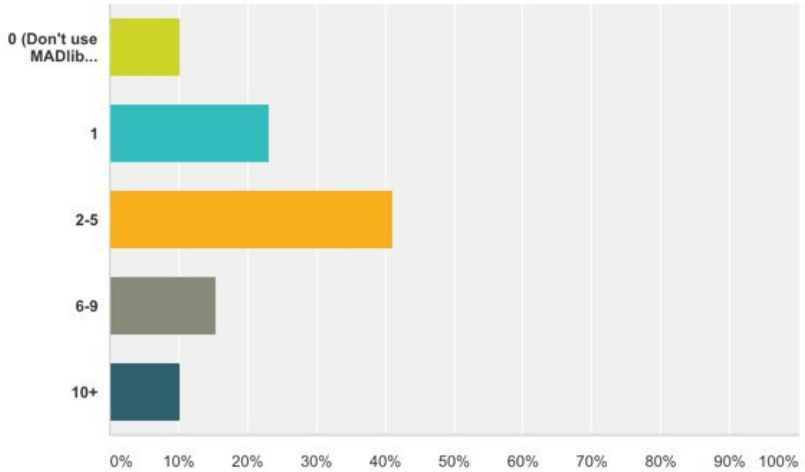



| Answer Choices | Responses | |
|---|---|---|
| Don't use MADlib currently | 15.38% | 6 |
| Less than 1 year | 53.85% | 21 |
| 1 year | 7.69% | 3 |
| 2 years | 12.82% | 5 |
| 3+ years | 10.26% | 4 |
| Total | | 39 |

# Q2

## How many people in your organization use MADlib?

Answered: 39 Skipped: 0

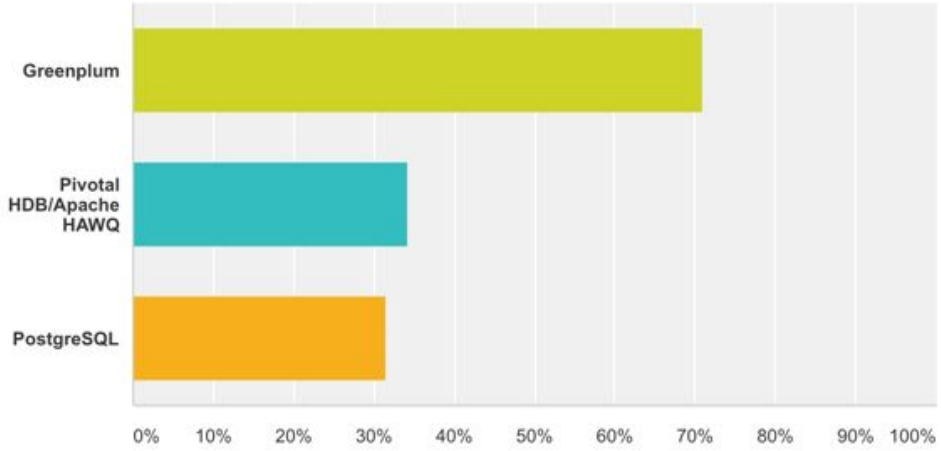



| Answer Choices | Responses | |
|---|---|---|
| 0 (Don't use MADlib currently) | 10.26% | 4 |
| 1 | 23.08% | 9 |
| 2-5 | 41.03% | 16 |
| 6-9 | 15.38% | 6 |
| 10+ | 10.26% | 4 |
| Total | | 39 |

# Q3

## Which platform(s) do you run MADlib on?

Answered: 38 Skipped: 1




| Answer Choices | Responses | |
|---|---|---|
| Greenplum | 71.05% | 27 |
| Pivotal HDB/Apache HAWQ | 34.21% | 13 |
| PostgreSQL | 31.58% | 12 |
| Total Respondents: 38 | | |

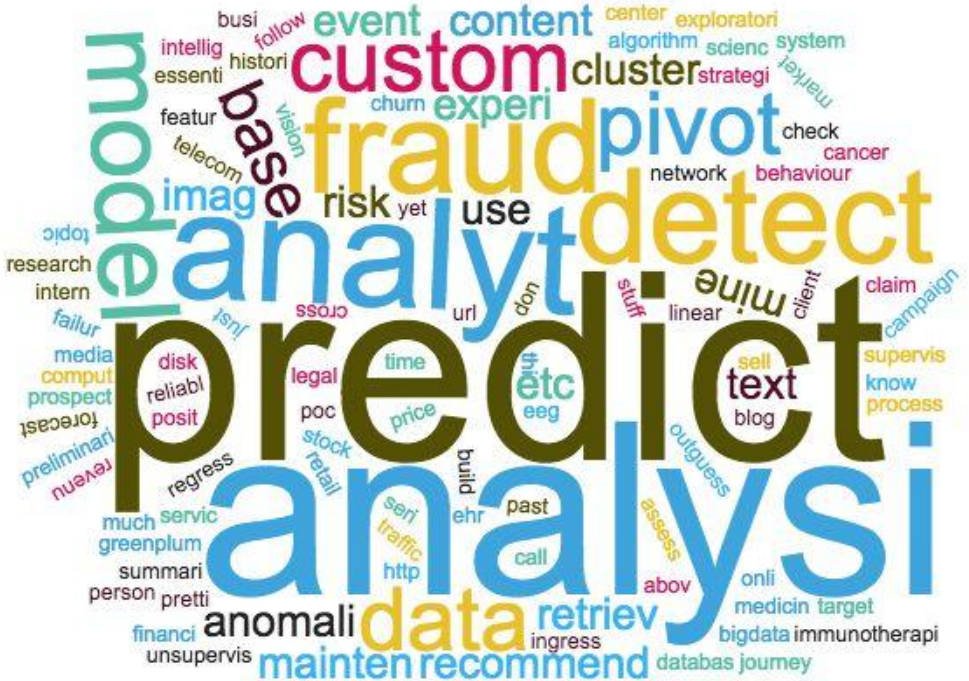
# Q4 - Top Use Cases

| Use case |
| --- |
| Fraud detection |
| Predictive modeling (general) |
| Clustering |
| Financial risk analysis |
| Client prospecting |
| Customer experience analytics |
| Marketing |
| Predictive maintenance |
| Recommendation |
| Text mining |

# Q4 - Other Use Cases

| Use case |
| --- |
| Ad targeting |
| Anolmaly detection |
| Business intelligence |
| Call center analytics |
| Cancer research |
| Claims prediction |
| Computer vision processing |
| Content based image retrieval |
| Cross sell |
| Disk failure |
| Electroencephalogram (EEG) |
| Electronic health records |
| HR analytics |
| Immuno therapy |
| Legal |
| Network traffic time series |
| Personalized medicine |
| Reliability analysis |
| Retail |
| Revenue forecasting |
| Service strategy assessments |
| Stock price analysis |
| Telecom customer churn |
| Topic modeling |

# Q4 - Use Cases




Stemmed, stop words removed

# Q5 - Frequently Used Algorithms

| MADlib algos used |
| --- |
| Regression (various) |
| K-means |
| Random forest |
| LDA |
| Elastic net |
| Summary |
| ARIMA |
| Association rules |
| Correlation |
| Decision tree |
| PCA |
| Low rank matrix factorization |
| Classification (various) |
| PivotalR |
| SVM |

# Q6 - Top Requested Features

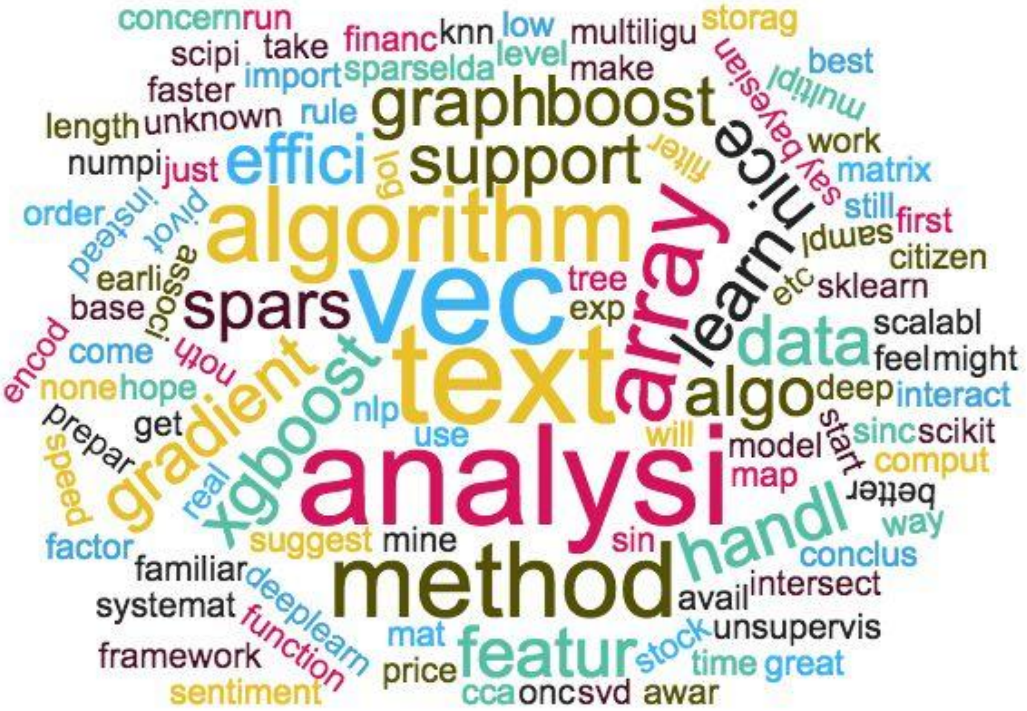
| Desired new features |
| --- |
| Gradient boosted trees |
| Data preparation |
| Graph algorithms |
| Better interaction with PL/Python and libraries |
| R interface* |
| Deep learning |
| Improve speed of association rules |


*Note that there is an R interface called PivotalR
https://cran.r-project.org/web/packages/PivotalR/

# Q6 - Other Requested Features

| Desired new features |
| --- |
| Sentiment analysis |
| More sparse vector support |
| More scalable SVD |
| More scalable matrix factorization |
| More low level algorithms (vec-vec, vect-mat) |
| More efficient map functions on vectors (e.g., log, exp, sin) |
| Other optimization algorithms besides SGD |
| NLP |
| Hot encode >1600 features |
| More text mining |
| More array support |
| Sampling methods |
| More unsupervised learning |
| More Bayesian methods |
| More finance algorithms |
| kNN |
| Canonical-correlation analysis (CCA) |
| Important features from scikit-learn |
| Sparse LDA |

# Q6 - Requested Features

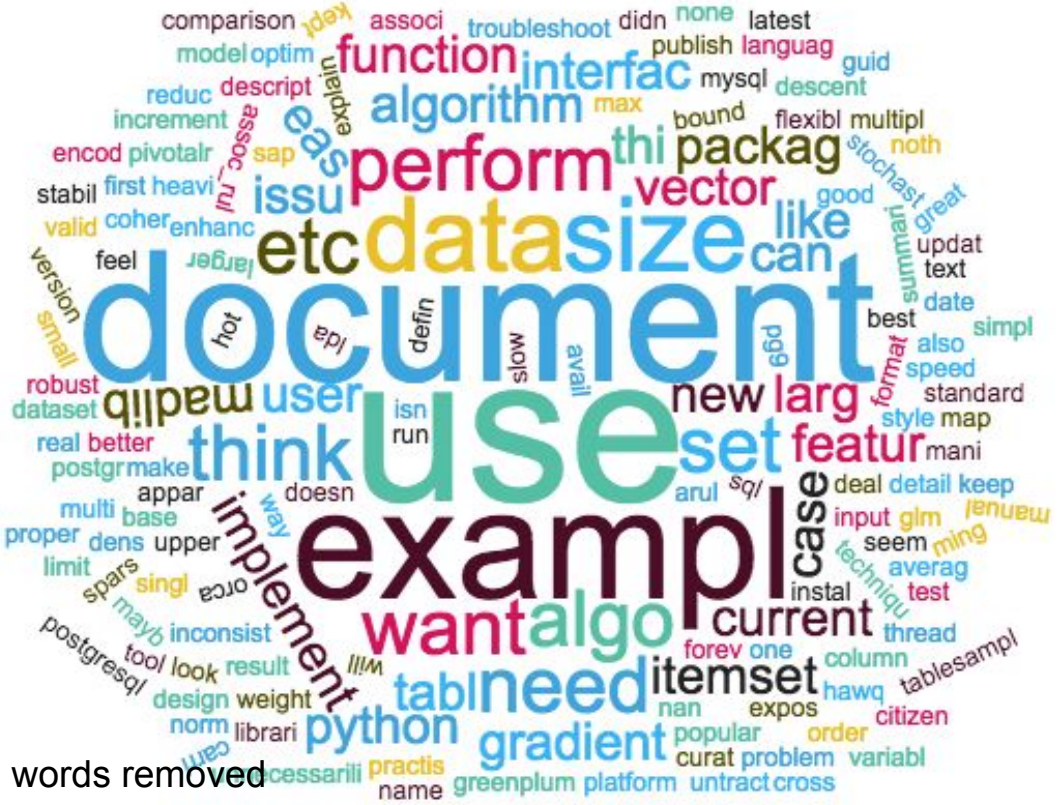



All responses, stemmed, stop words removed

# Q7 - Main Concerns

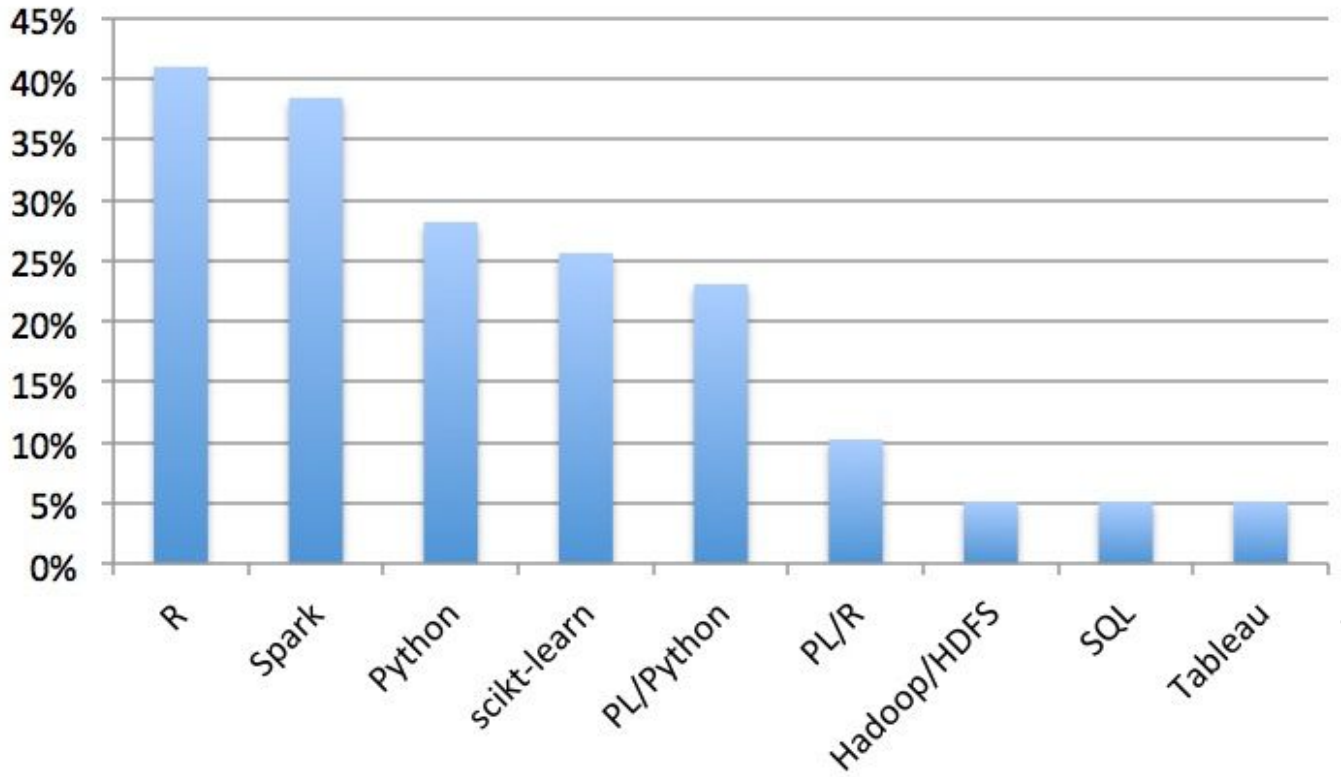
| Biggest issue |
| --- |
| Better docs and examples |
| Performance |
| Robustness, stability |
| Inconsistent input formats |
| Ease of use |

# Q7 - Main Concerns




All responses, stemmed, stop words removed

# Q8 - Other Tools Used




+Several others...