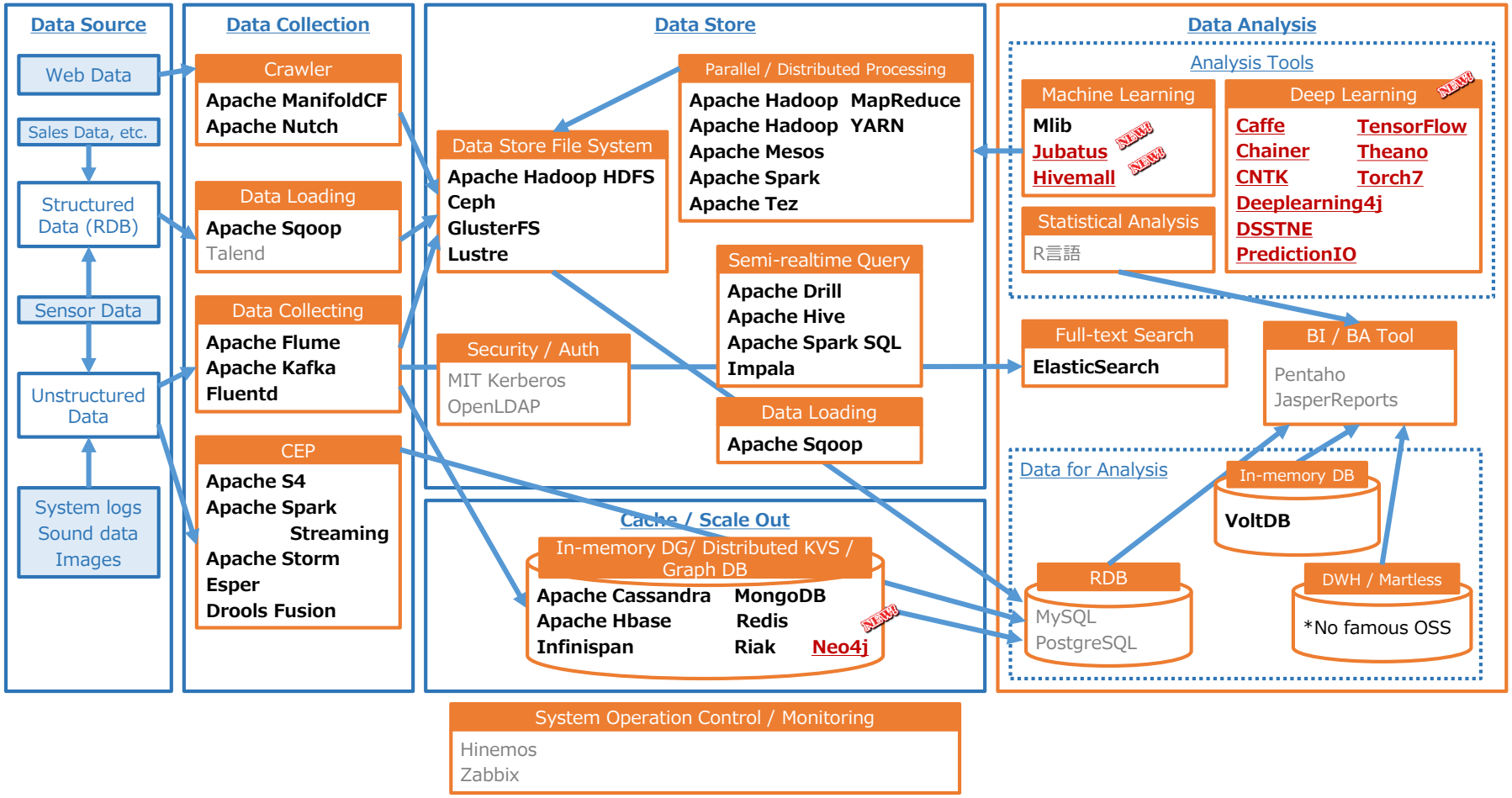
Survey of OSS in Big Data Platform (2016 update)

2017/6/1

Introduction: OSS in Big Data Platform

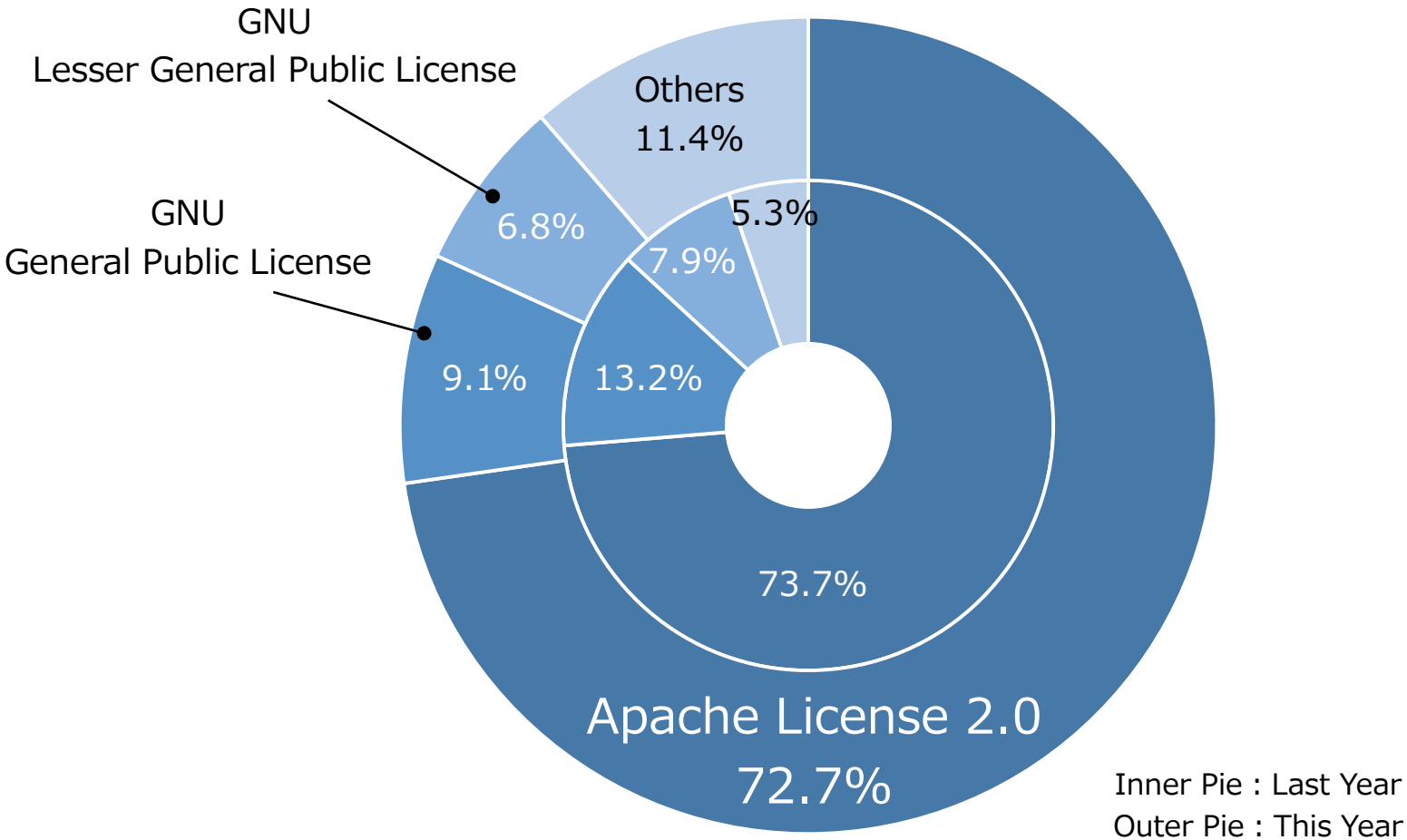
- Most of the Big Data Platforms are composed of 3 functions: Data Collection, Data Store, and Data Analysis
- This research targeted OSS written in black and red character



* OSS in gray characters are not targeted in this research

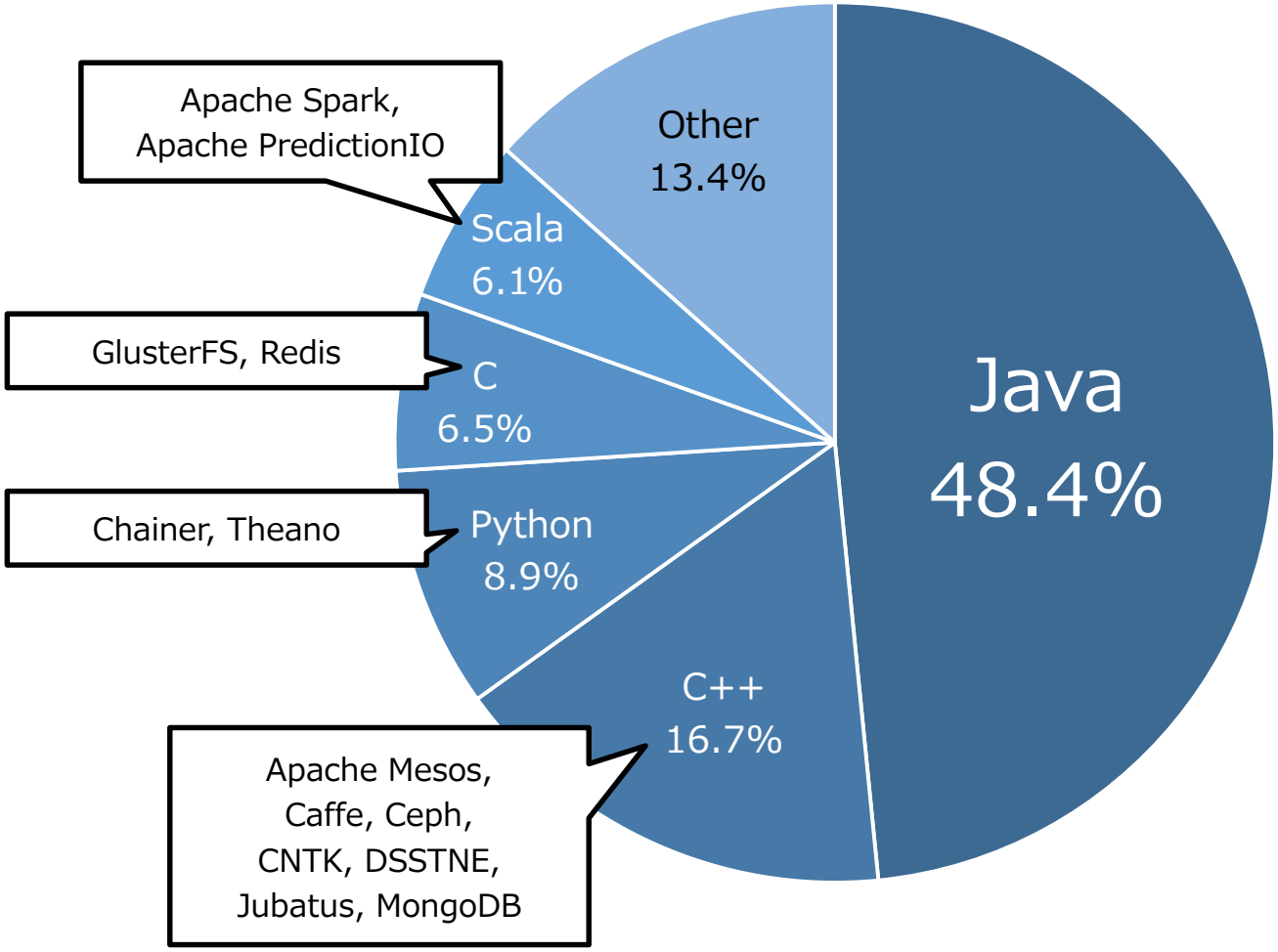
Licenses

- Over 70% of software uses Apache License
  - A large number of software are under the Apache Software Foundation
  - Some are Apache Incubator Projects (Impala, Hivemall, PredictionIO)



Programming Language

- Java is the most popular language used in OSS (48.4%)
  - C++ (16.7%), Python (8.9%), C (6.5%)



Viewpoints

- We used 4 viewpoints to rate the OSS

Developers' Activity
How active are developers of the OSS?

Use in Companies
How many companies do use the OSS ?

Engineers' Interest
How many developers are interested in the OSS?

Adequacy of technical information
How much information can we get about the OSS?

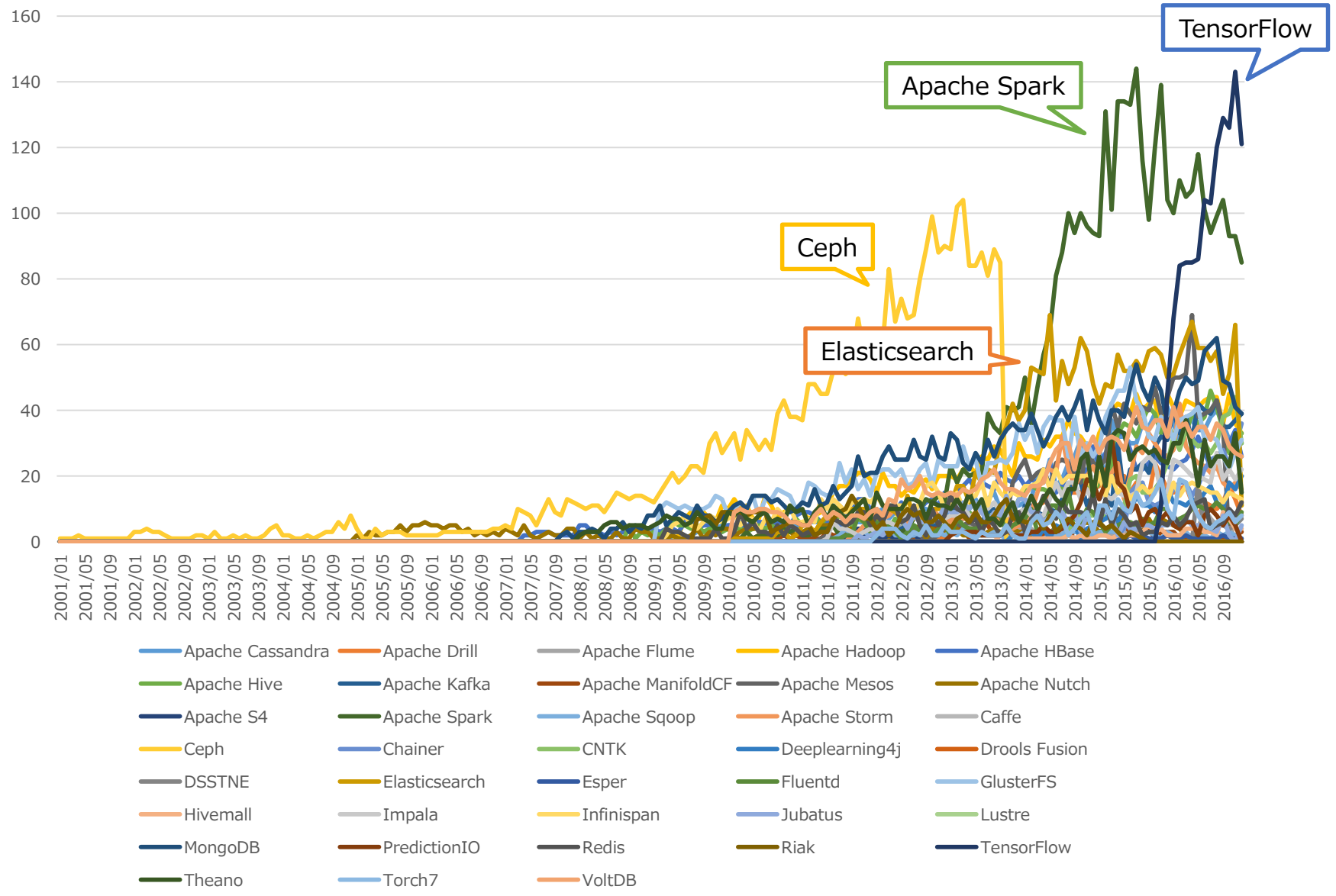
Developers' Activity

How active are developers of the OSS ?

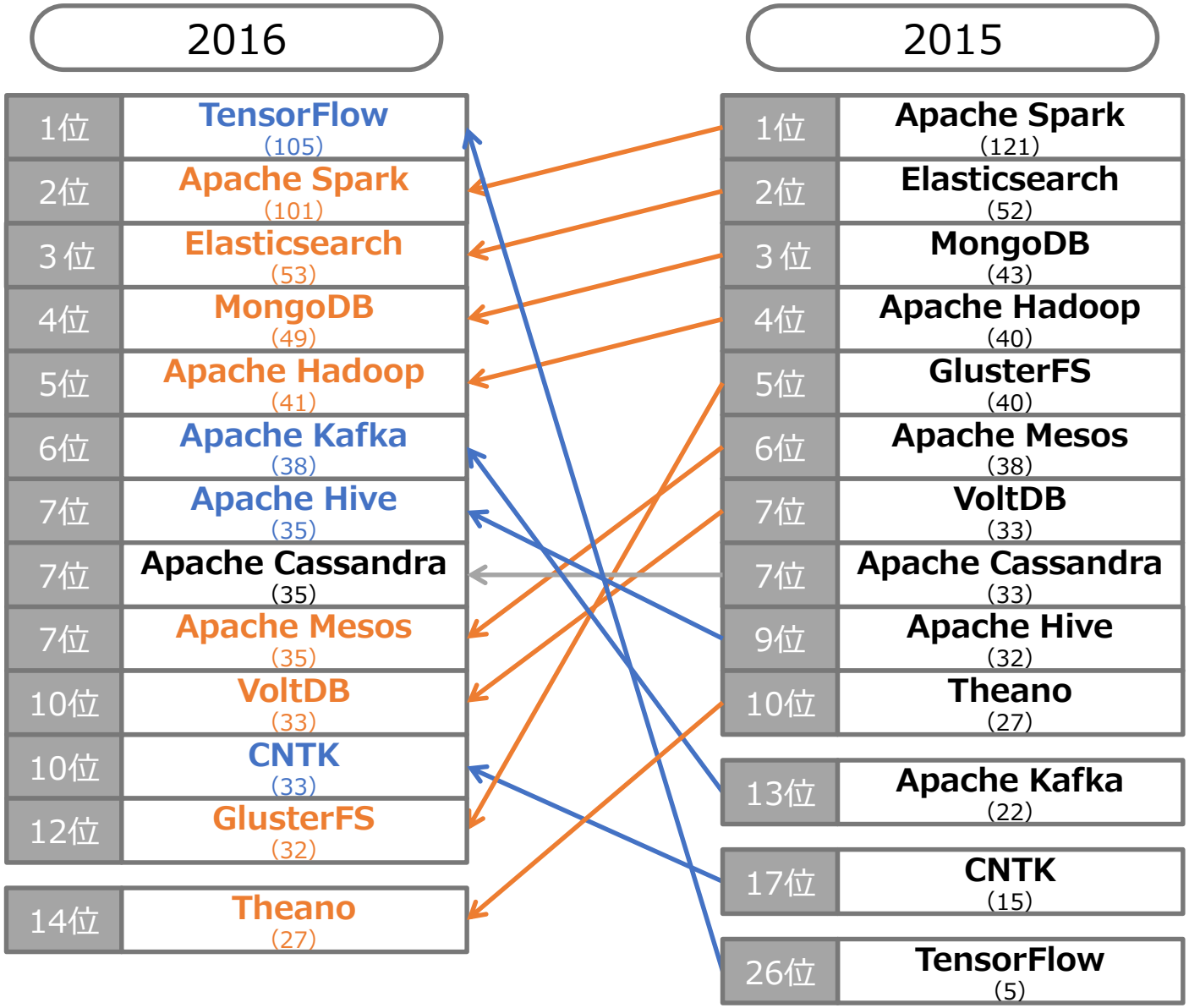
Viewpoints of developers' activity

- Committers
  - Number of developers who committed the repository (ex. GitHub) in 2016
  - And growth rate from 2015
- Commits
  - Number of commits made into the repository (ex. GitHub) in 2016
  - And growth rate from 2015
- Active days
  - Ratio of days with commits in Git since the initial commit
- Mails in mailing list for developers
  - Number of mails posted to the mailing list for developers in 2016
- Issue resolution rate
  - Ratio of the closed issue to all issues
    - In this research, we did not consider the priority of issues

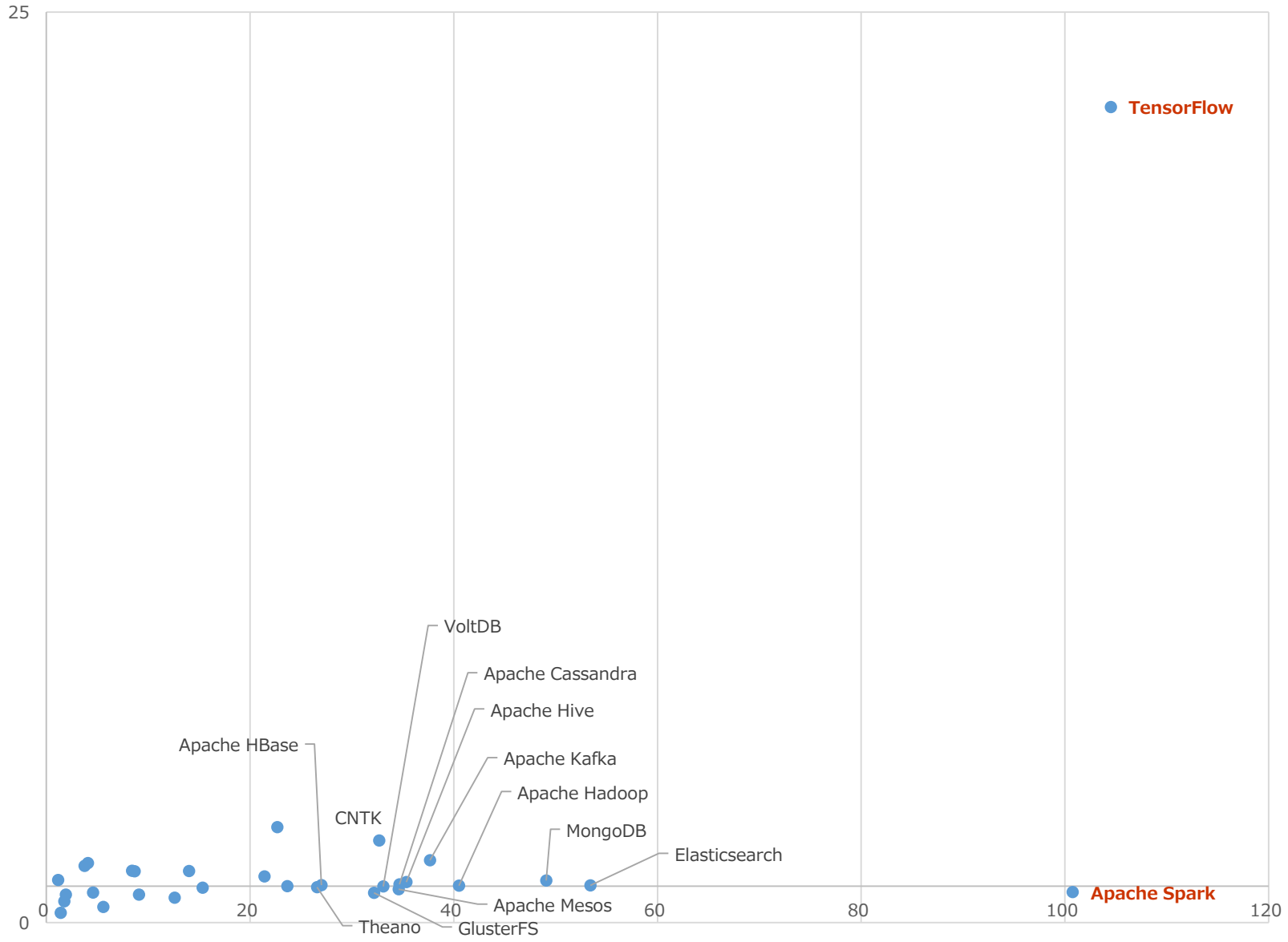
Committers (monthly)



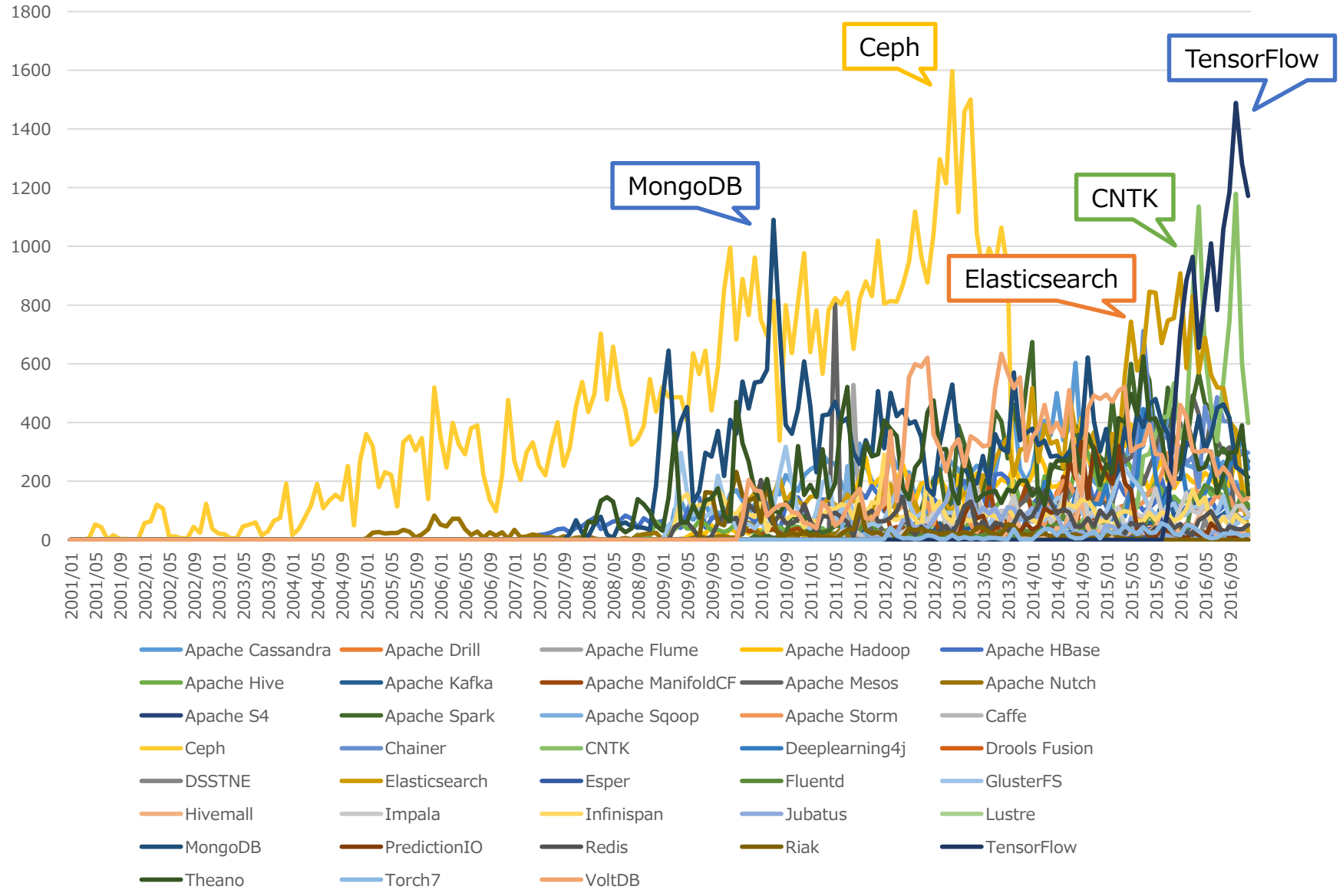
Monthly average number of committers



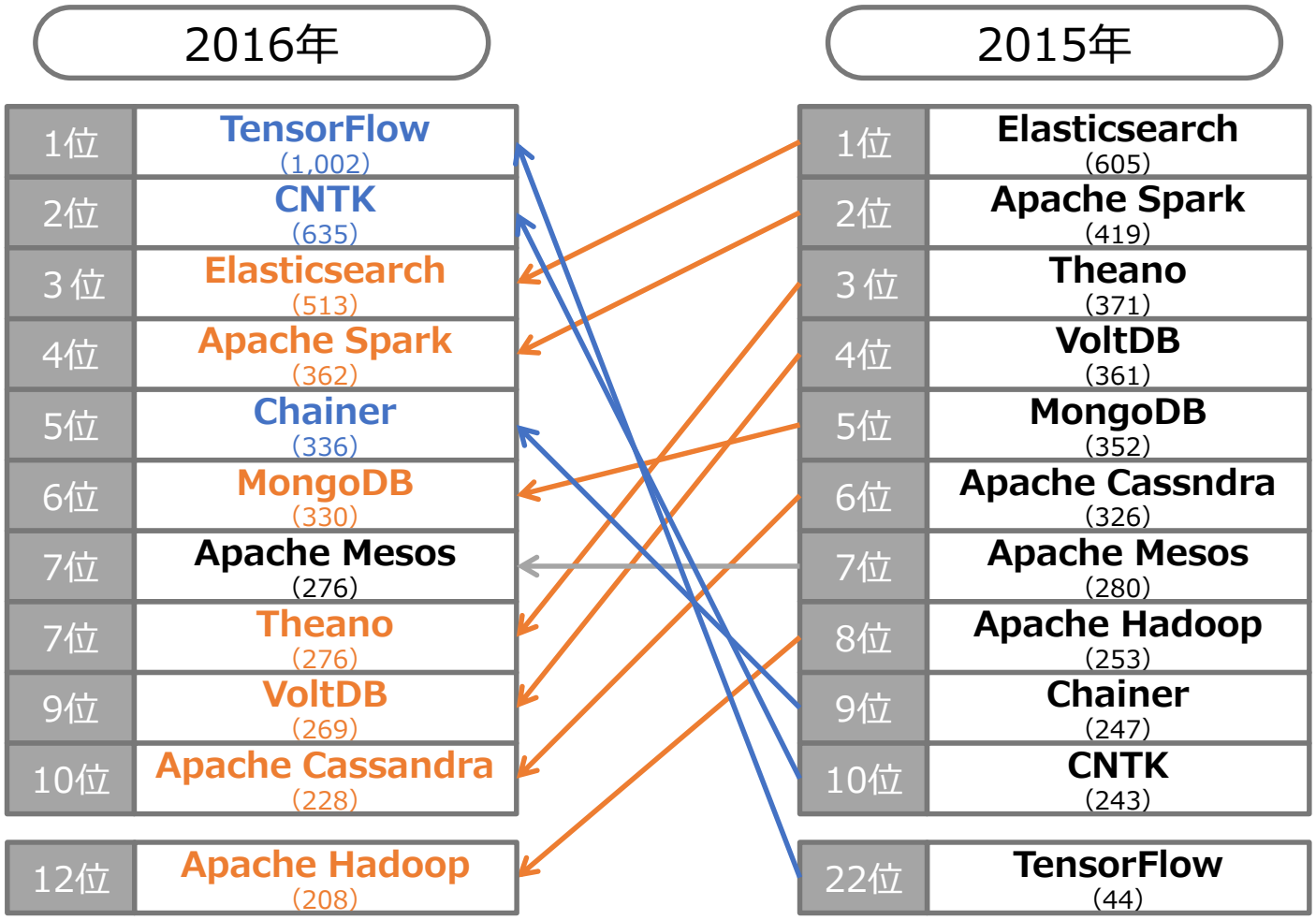
Committers [x-axis] and growth rate [y-axis]



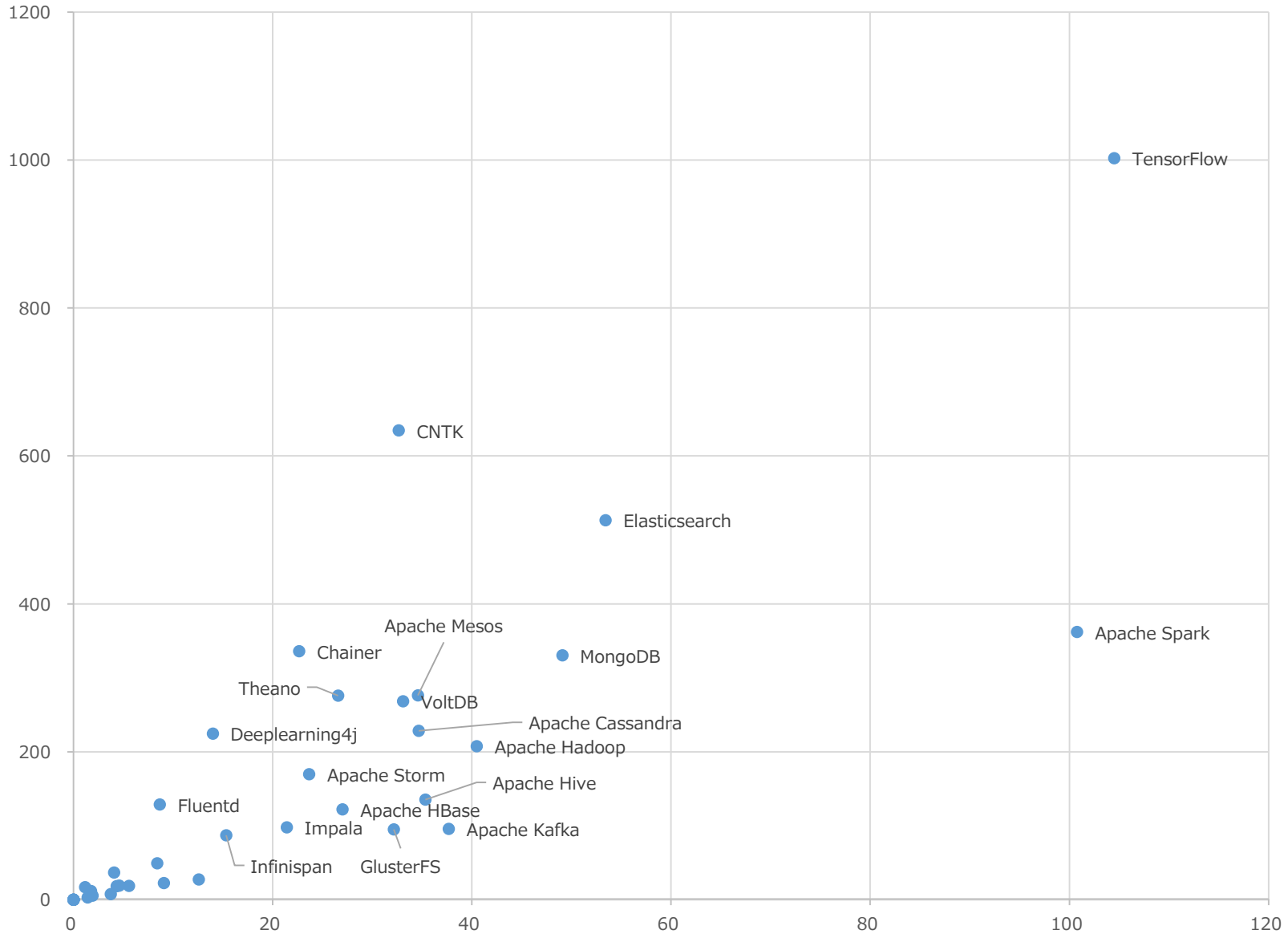
Commits (monthly)



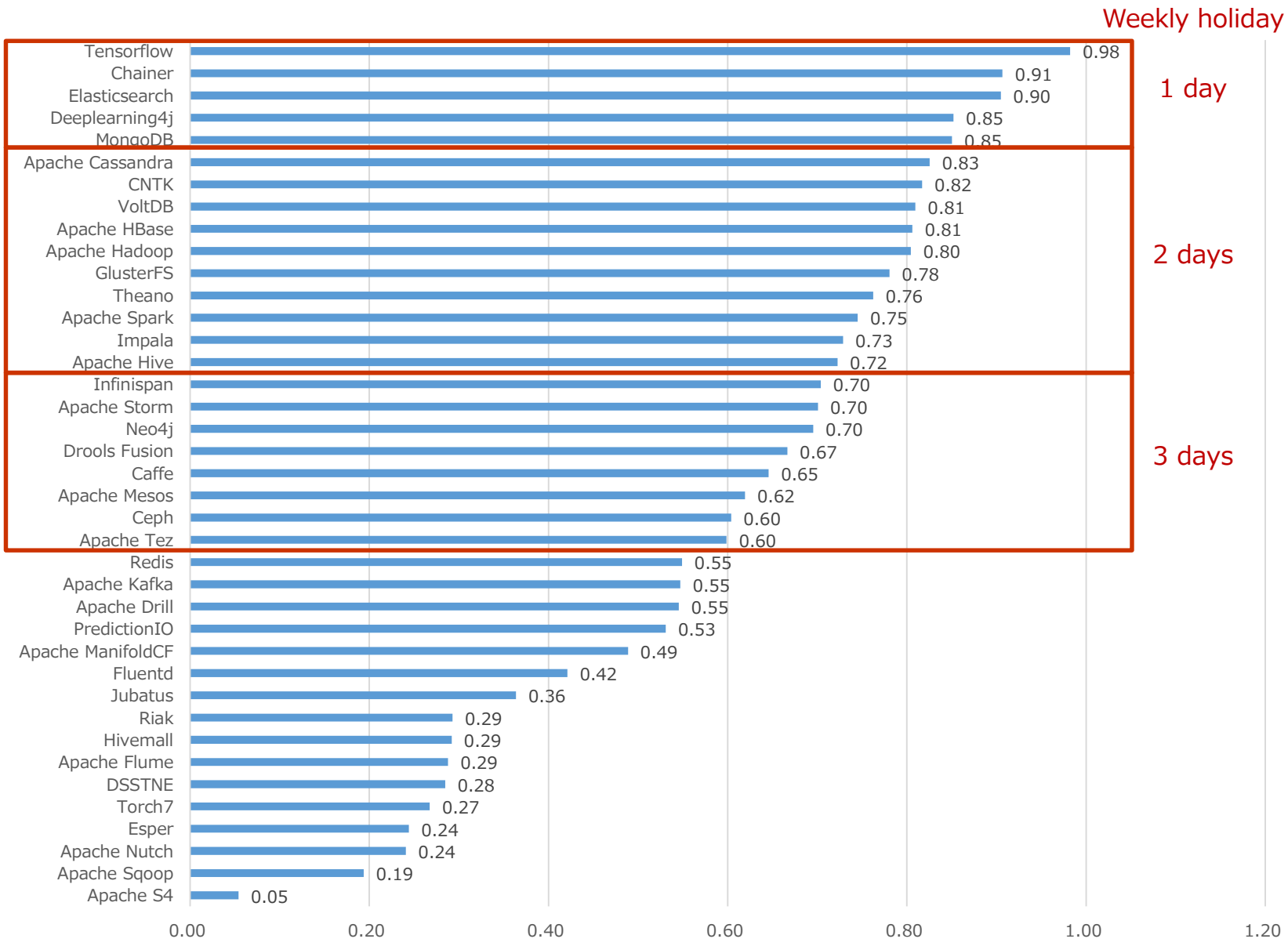
Monthly average number of commits



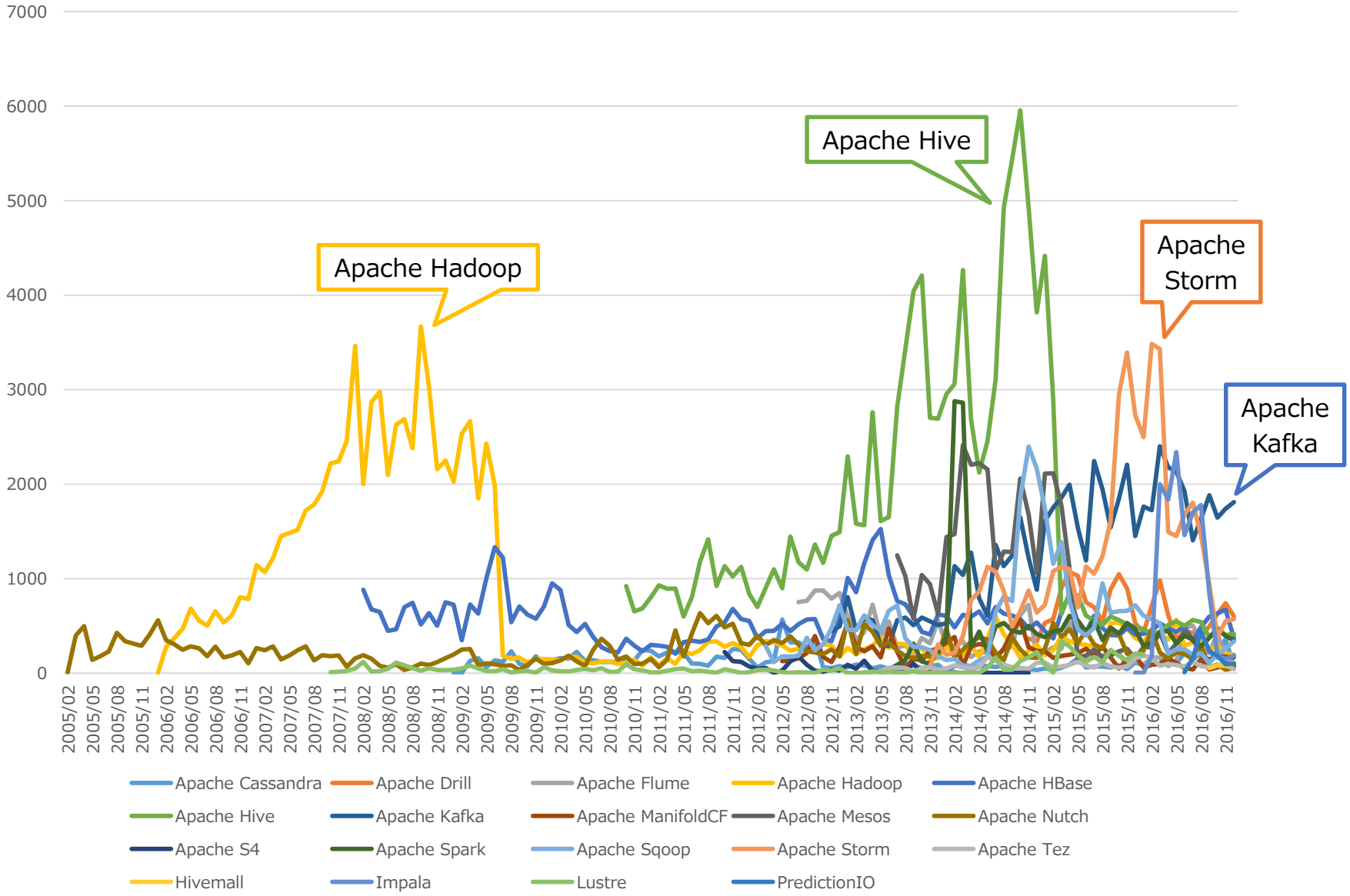
Committers [x-axis] and Commits [y-axis]



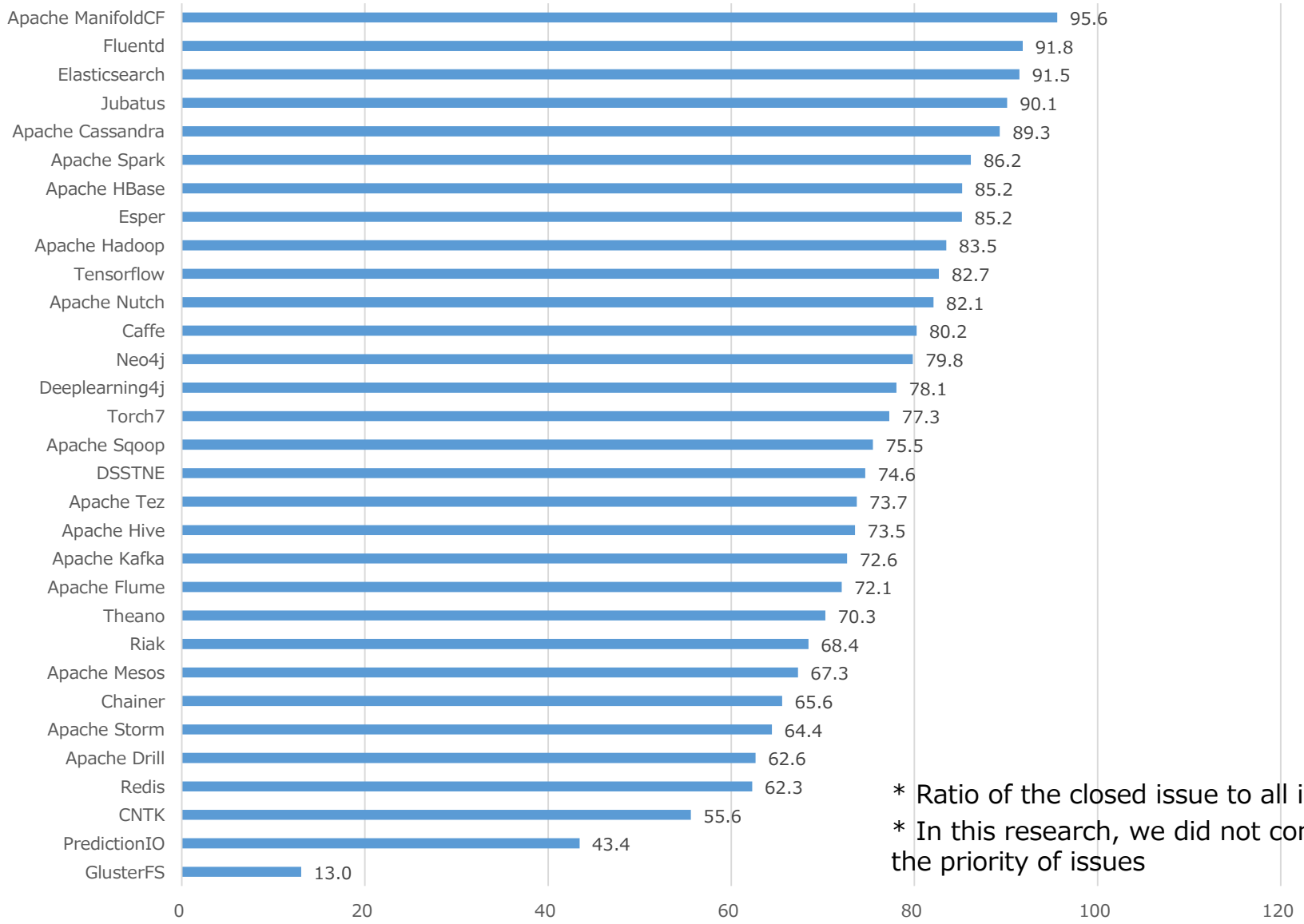
Active days



Mails in mailing list for developers

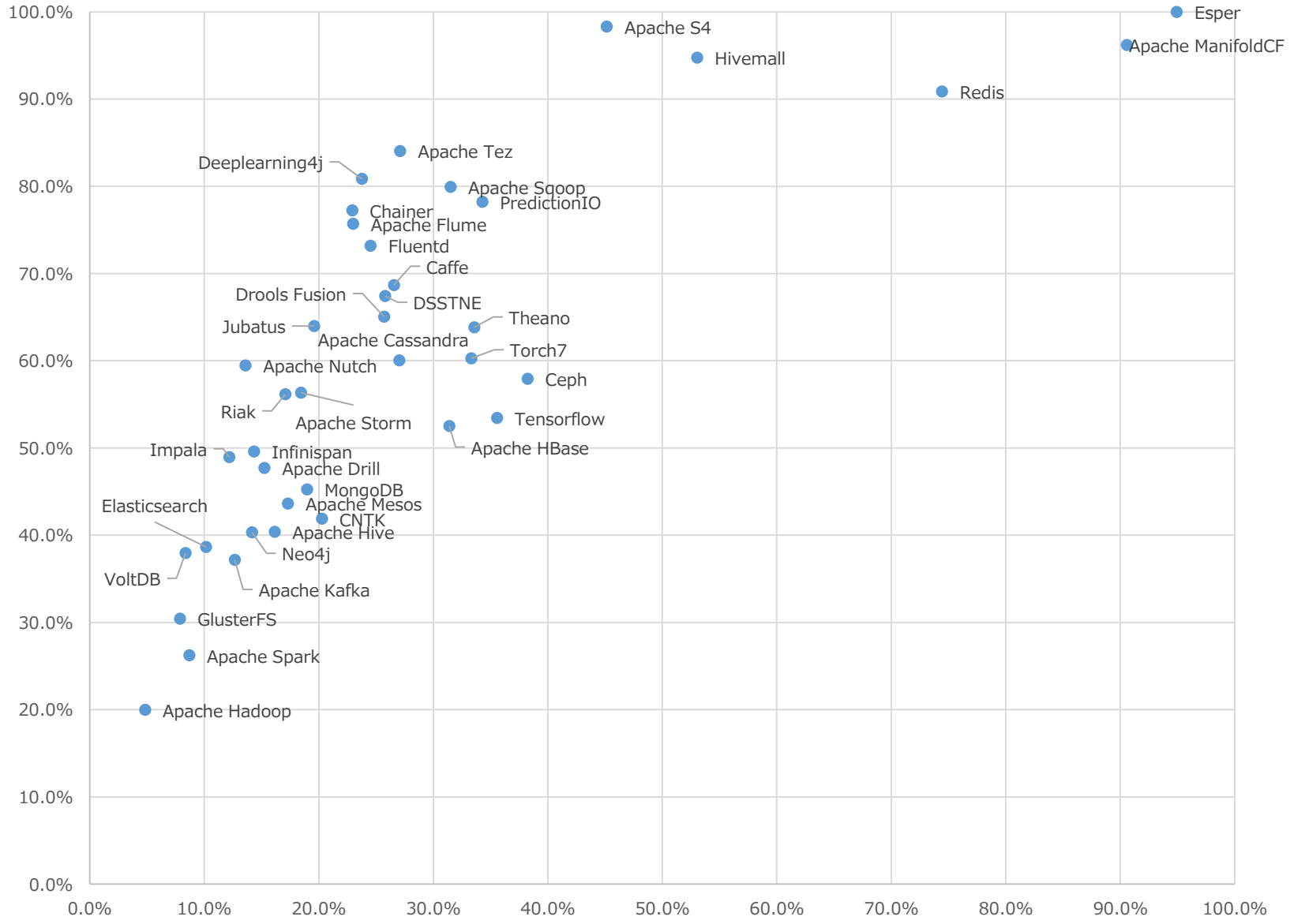


Issue resolution rate



* Ratio of the closed issue to all issues
* In this research, we did not consider the priority of issues

Top1 [x-axis] and Top1-5 [y-axis] developers



Ranking of “Developers’ Activity”

1位	**TensorFlow**
2位	**Elasticsearch**
3位	**Apache Spark**
4位	**MongoDB**
5位	**CNTK**

* compare with deviation value of each viewpoints

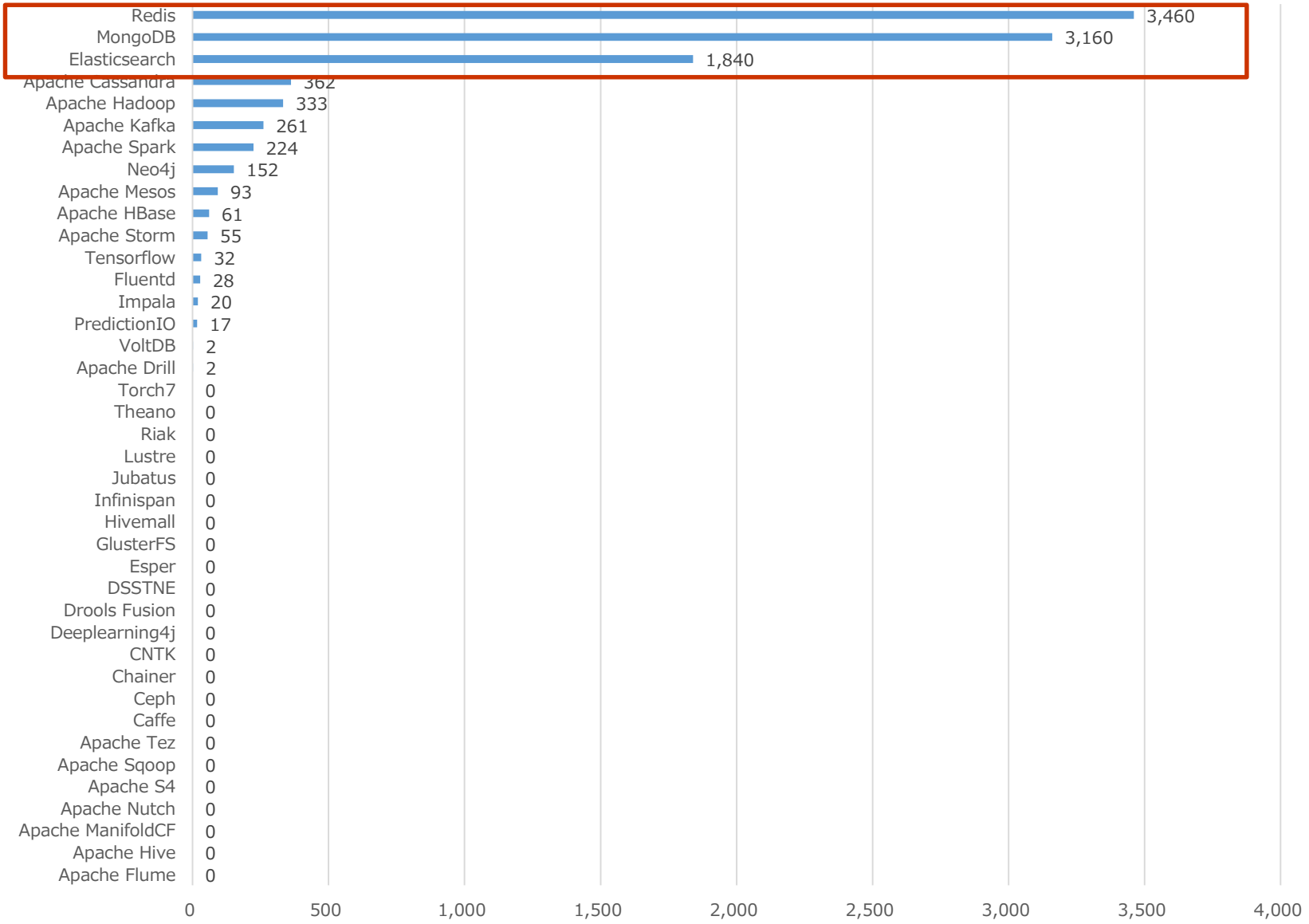
Use in Companies

How many companies do use the OSS?

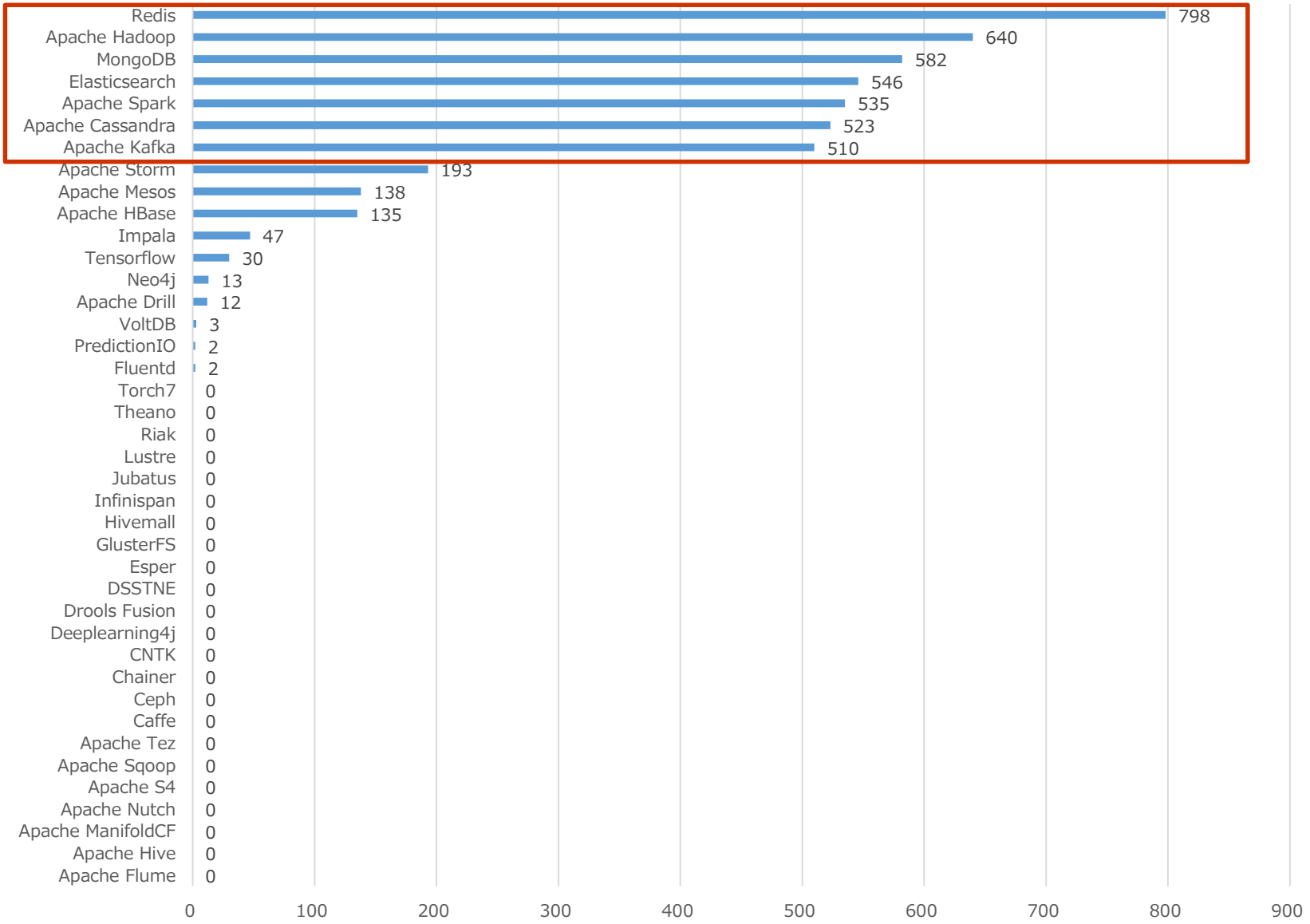
Viewpoints of use in companies

- Number of companies using the OSS
  - Stacks registered in StackShare (https://stackshare.io/)
- Number of jobs
  - Jobs registered in StackShare (https://stackshare.io/)
  - Jobs registered in StackOverflow (http://stackoverflow.com/)

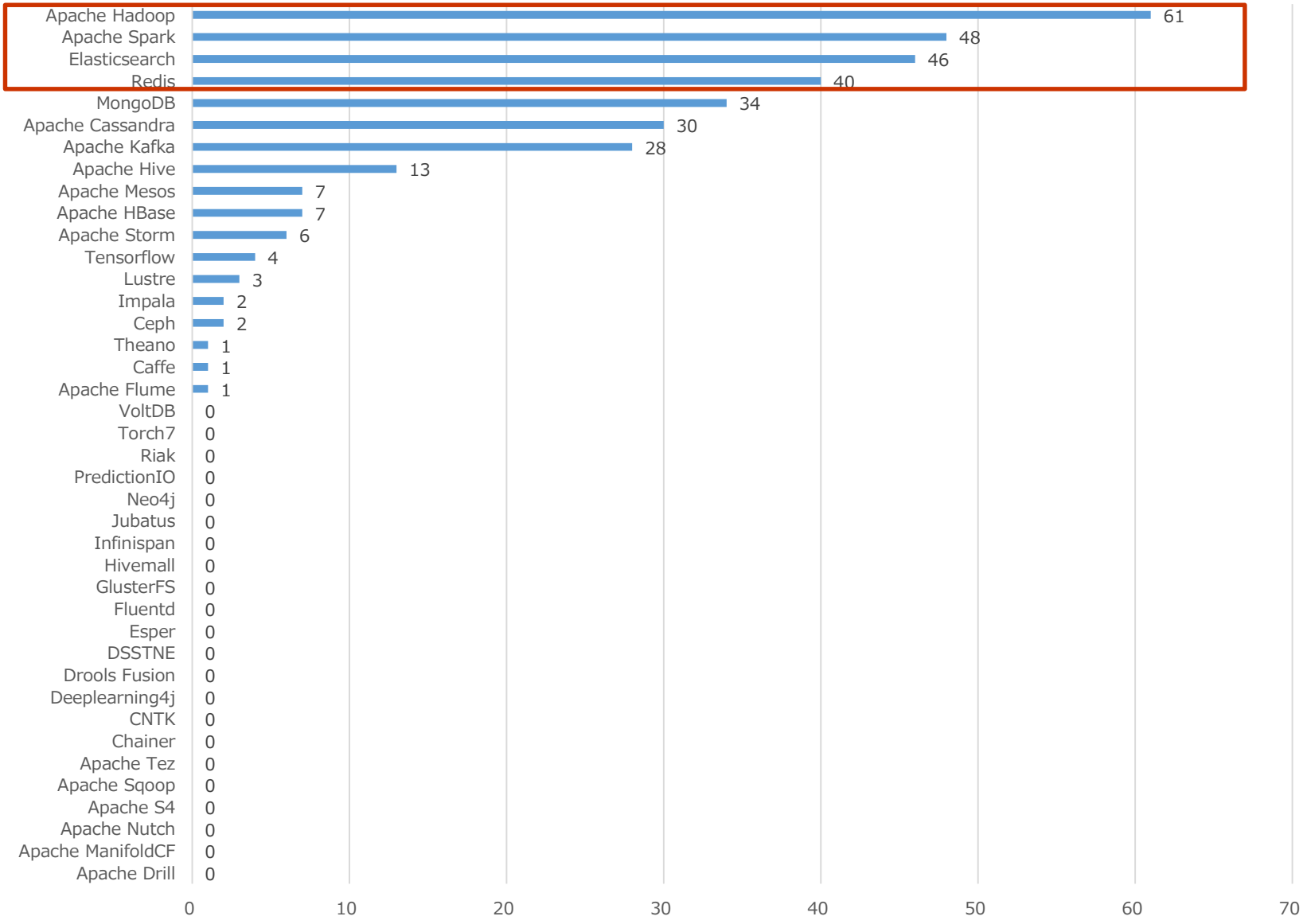
Number of companies using the OSS (StackShare)



Number of jobs (StackShare)



Number of jobs (StackOverflow)



Ranking of “Use in Business”

1位	**Redis**
2位	**MongoDB**
3位	**Elasticsearch**
4位	**Apache Hadoop**
5位	**Apache Spark**

* compare with deviation value of each viewpoints

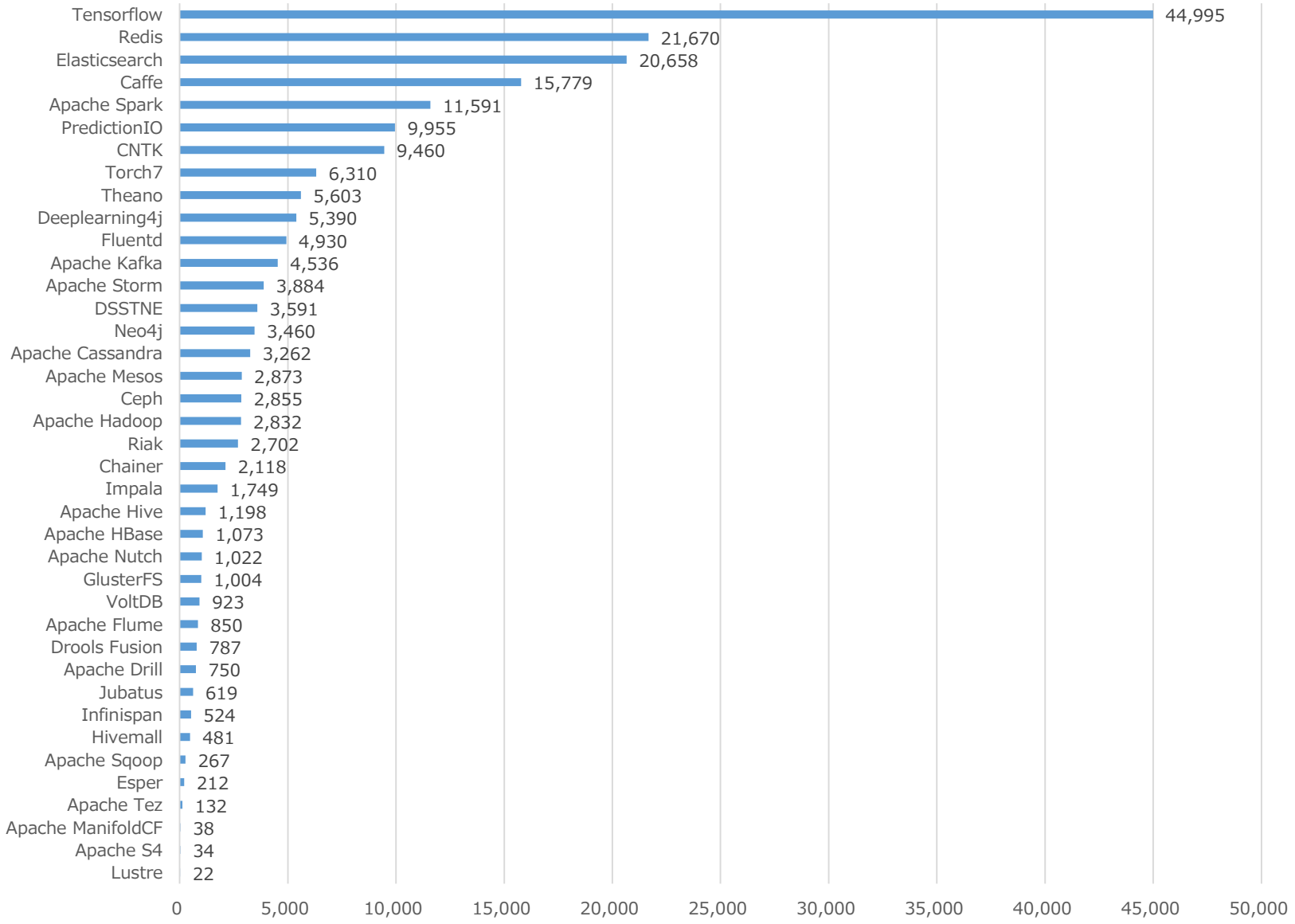
Engineers' Interest

How many engineers are interested in?

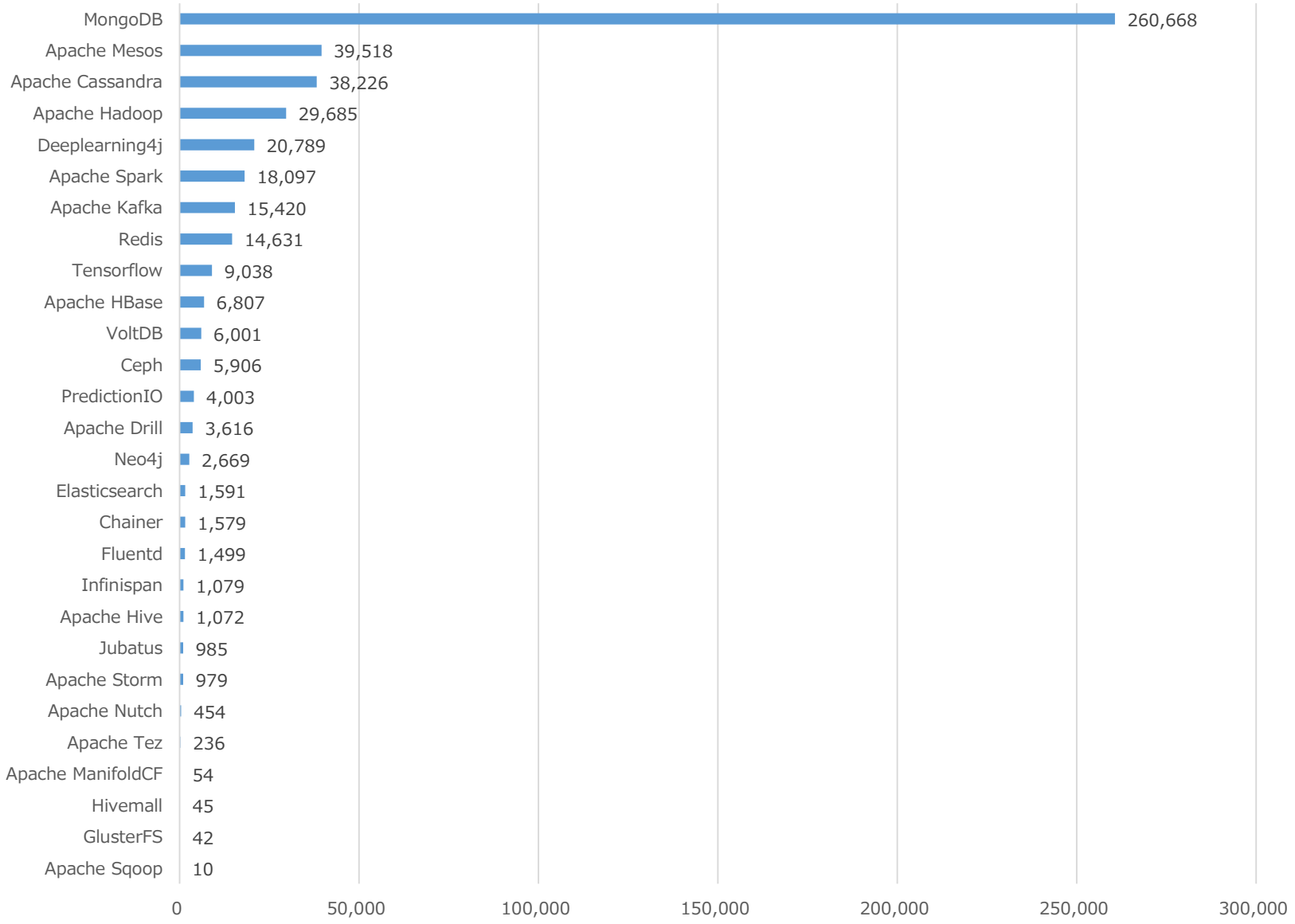
Viewpoints of engineers' interest

- Number of fans
  - Stars on GitHub (https://github.com/)
  - Followers on Twitter (https://twitter.com/)
  - Followers on Qiita (http://qiita.com/) *technical blog service in Japan
  - Fans on StackShare (https://stackshare.io/)
- Google Search
  - Points of Google Trends (https://www.google.co.jp/trends/)
    - Worldwide / Japan
    - Relative values to the value of Node.js

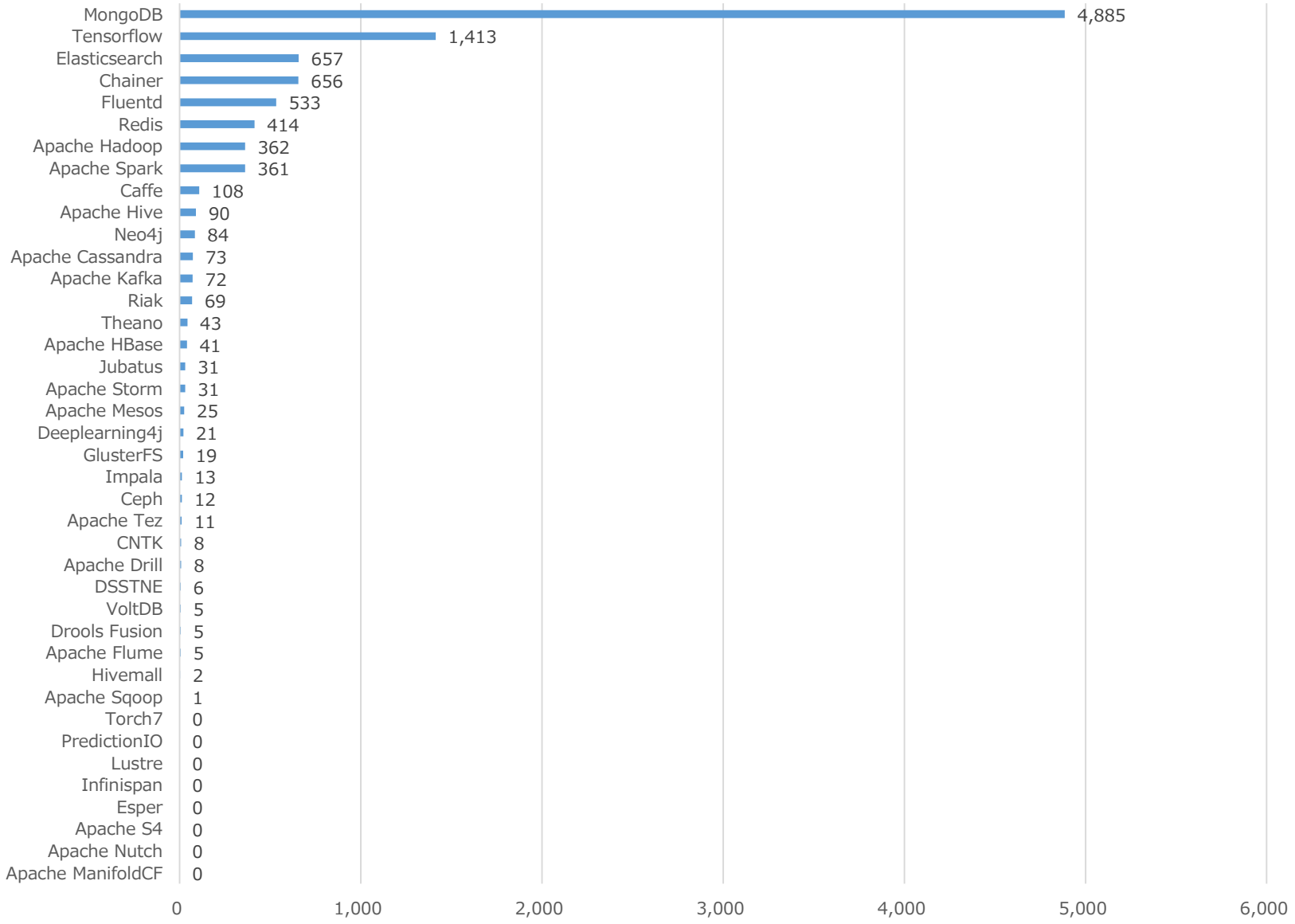
Stars on GitHub



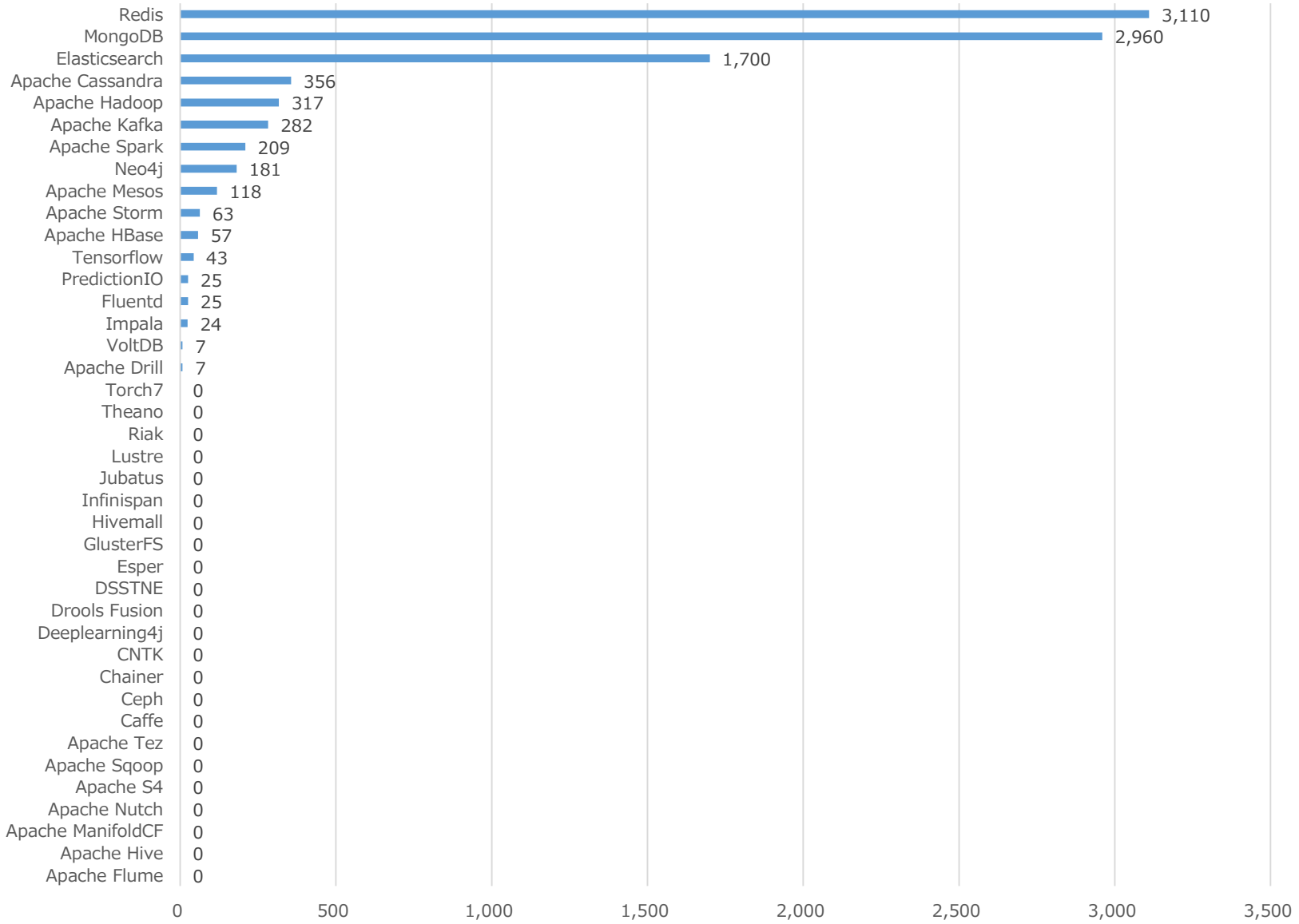
Followers on Twitter



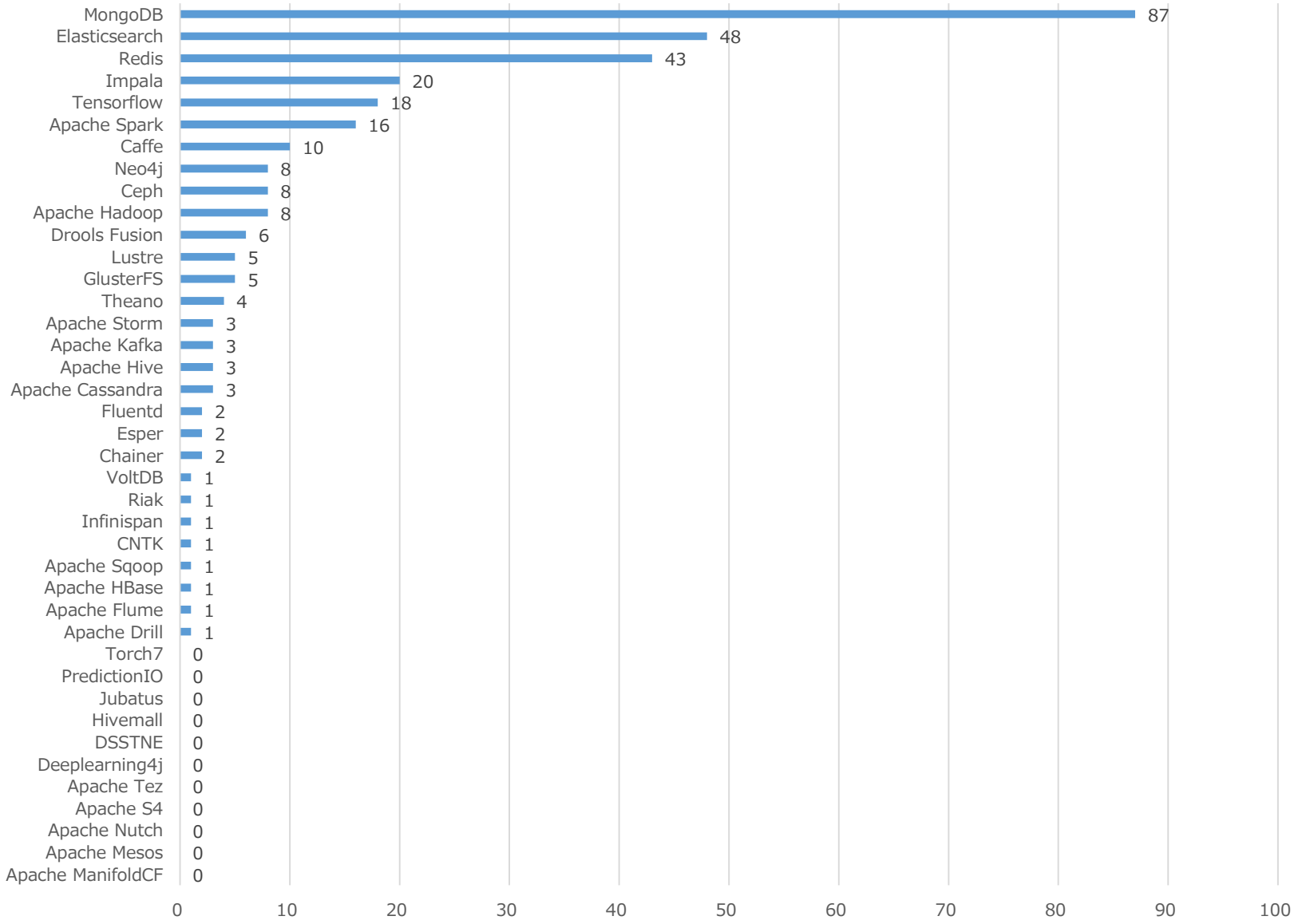
Followers on Qiita



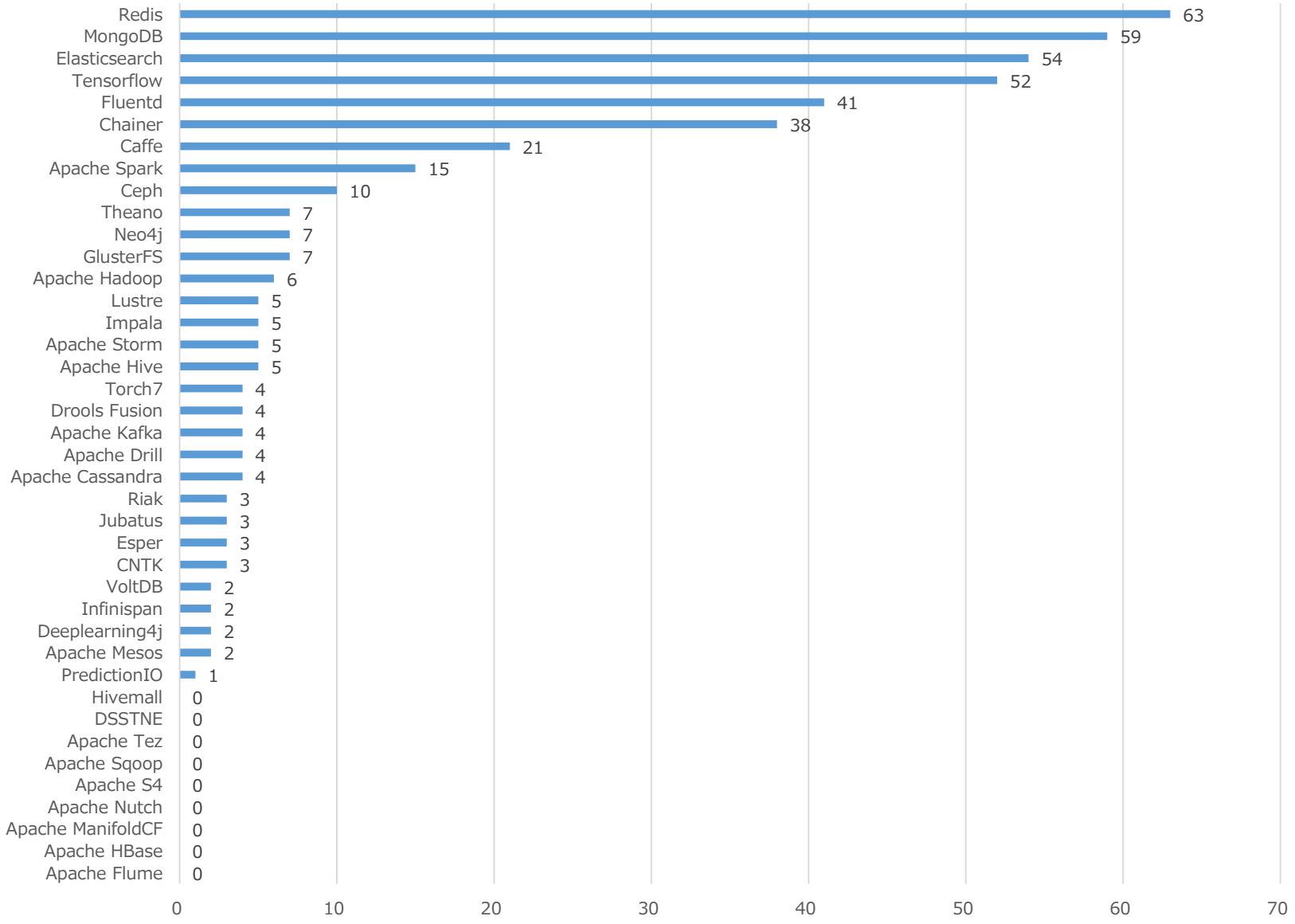
Fans on StackShare



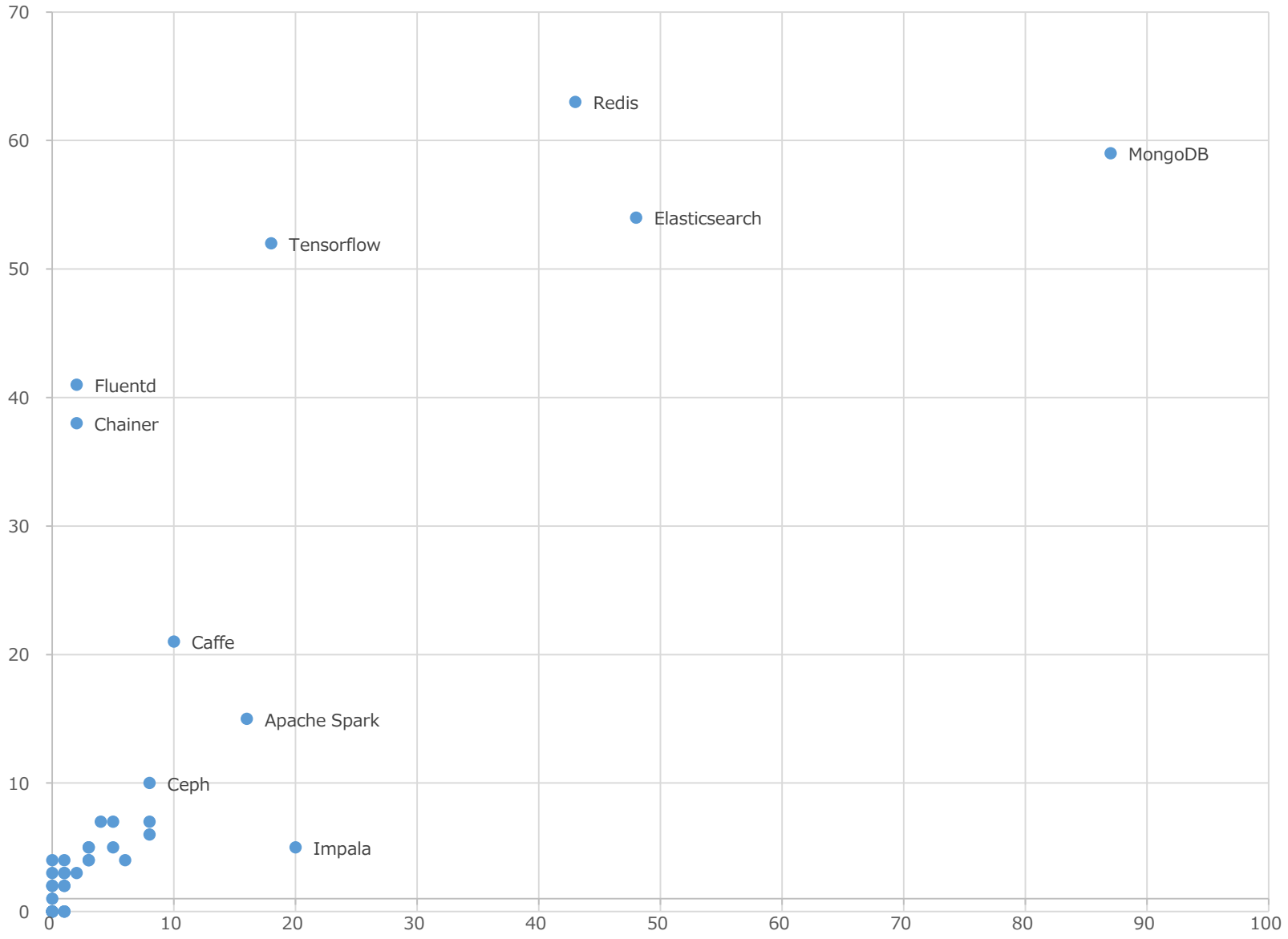
Google Trends (Worldwide)



Google Trends (Japan)



Google Trends Worldwide [x-axis] and Japan [y-axis]



Ranking of “Engineers’ Interest”

1位	**MongoDB**
2位	**Redis**
3位	**Elasticsearch**
4位	**Tensorflow**
5位	**Caffe**

* compare with deviation value of each viewpoints

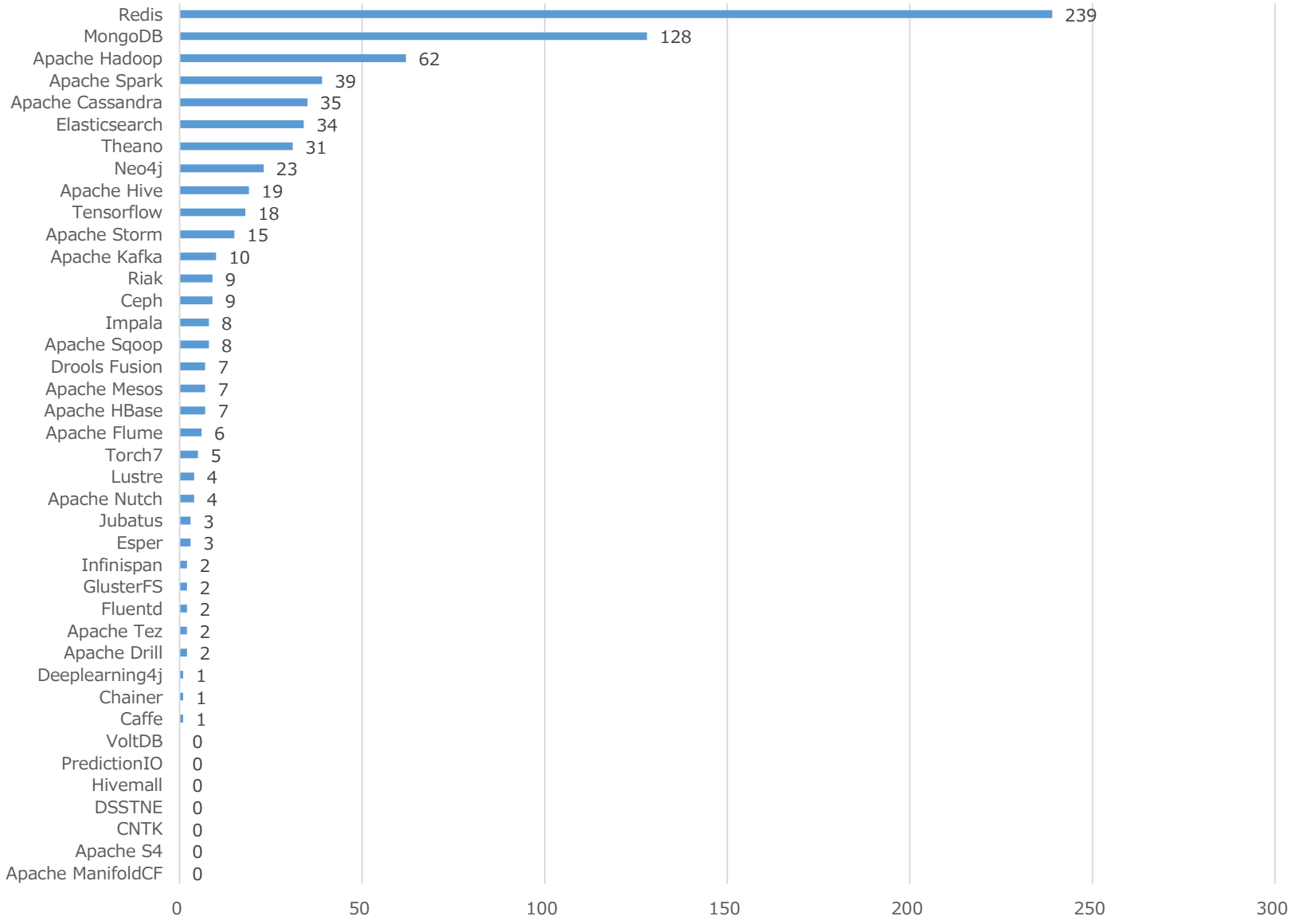
Adequacy of Technical Information

How much information can we get about the OSS?

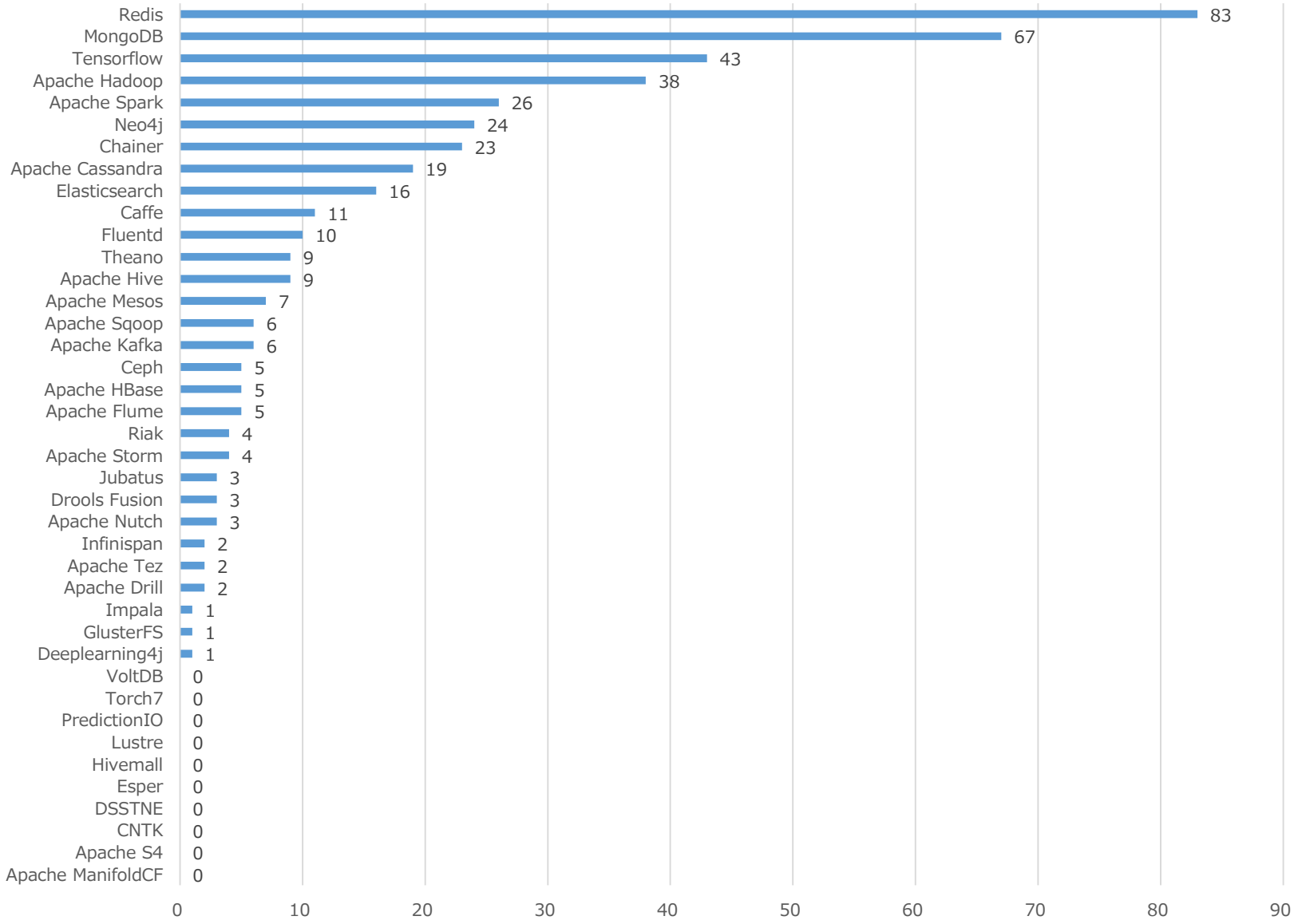
Viewpoints of Adequacy of technical information

- Books
  - Number of books available in Amazon (https://www.amazon.com/)
    - English books / Japanese books
- Seminars / workshops
  - Number of seminars / workshops on Connpass (https://connpass.com/)
- Slides
  - Number of presentations written in Japanese on SlideShare (http://www.slideshare.net/)
- Questions
  - Number of questions on StackOverflow (http://stackoverflow.com/)
  - Numebr of questions on following websites
    - QA@IT (http://qa.atmarkit.co.jp/)
    - Teratail (https://teratail.com/)
    - スタック・オーバフロー (http://ja.stackoverflow.com/) *StackOverflow for Japanese
- Technical posts
  - Number of posts in Hacker News (https://news.ycombinator.com/)
  - Number of posts in Qiita (http://qiita.com/)
- Mails in mailing list for users
  - Number of mails posted to mailing list for users in 2016

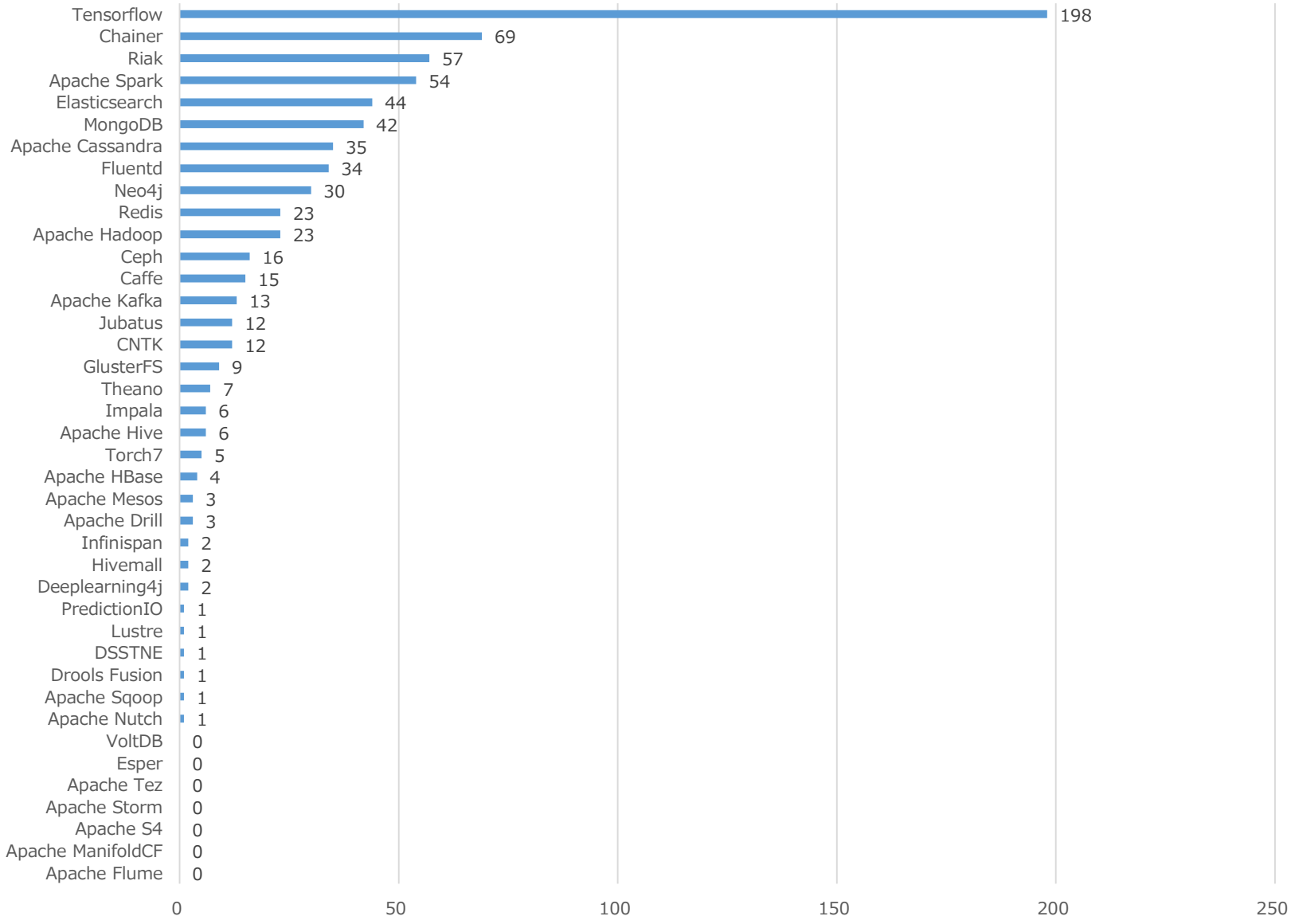
English books (Amazon)



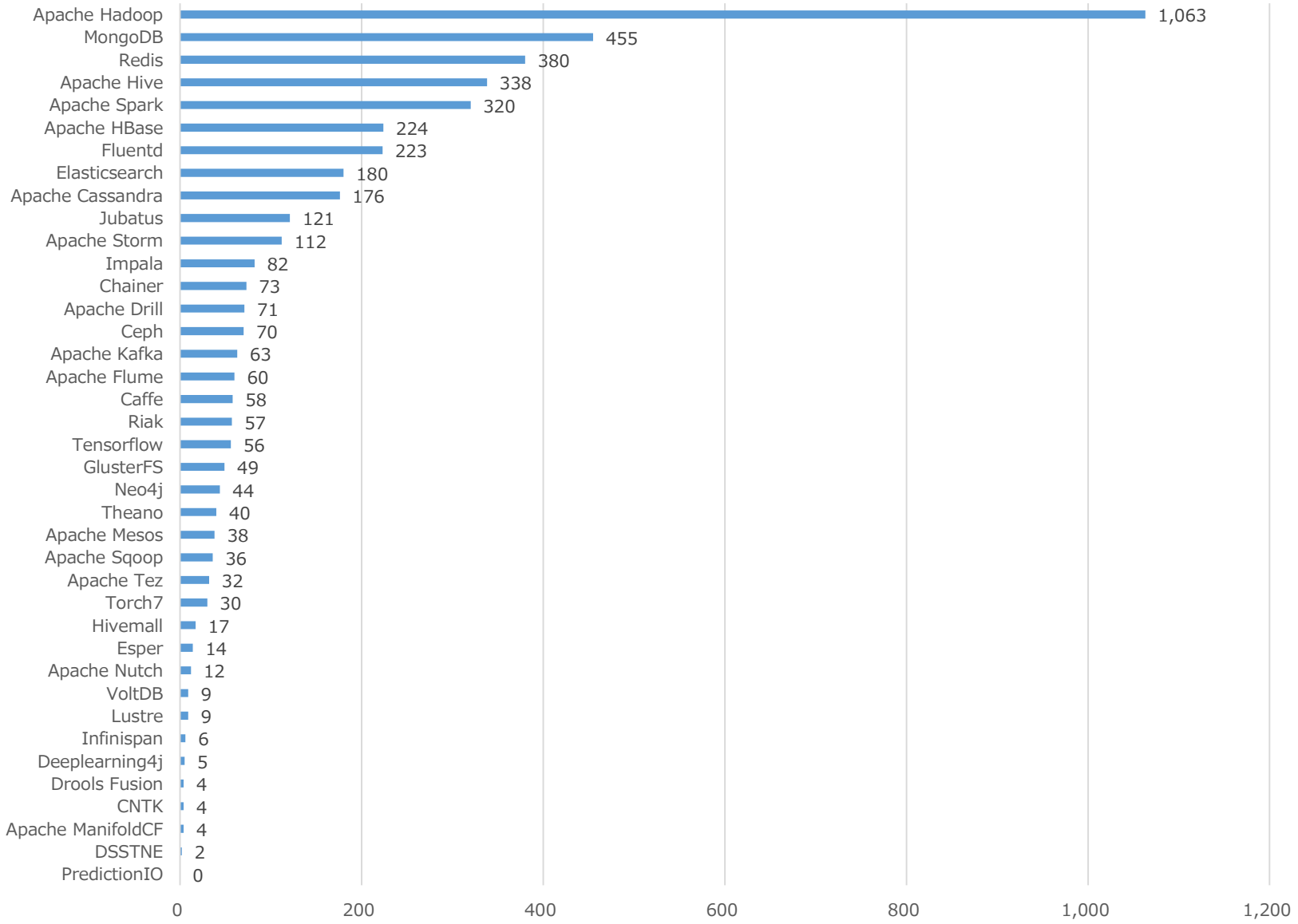
Japanese Books (Amazon)



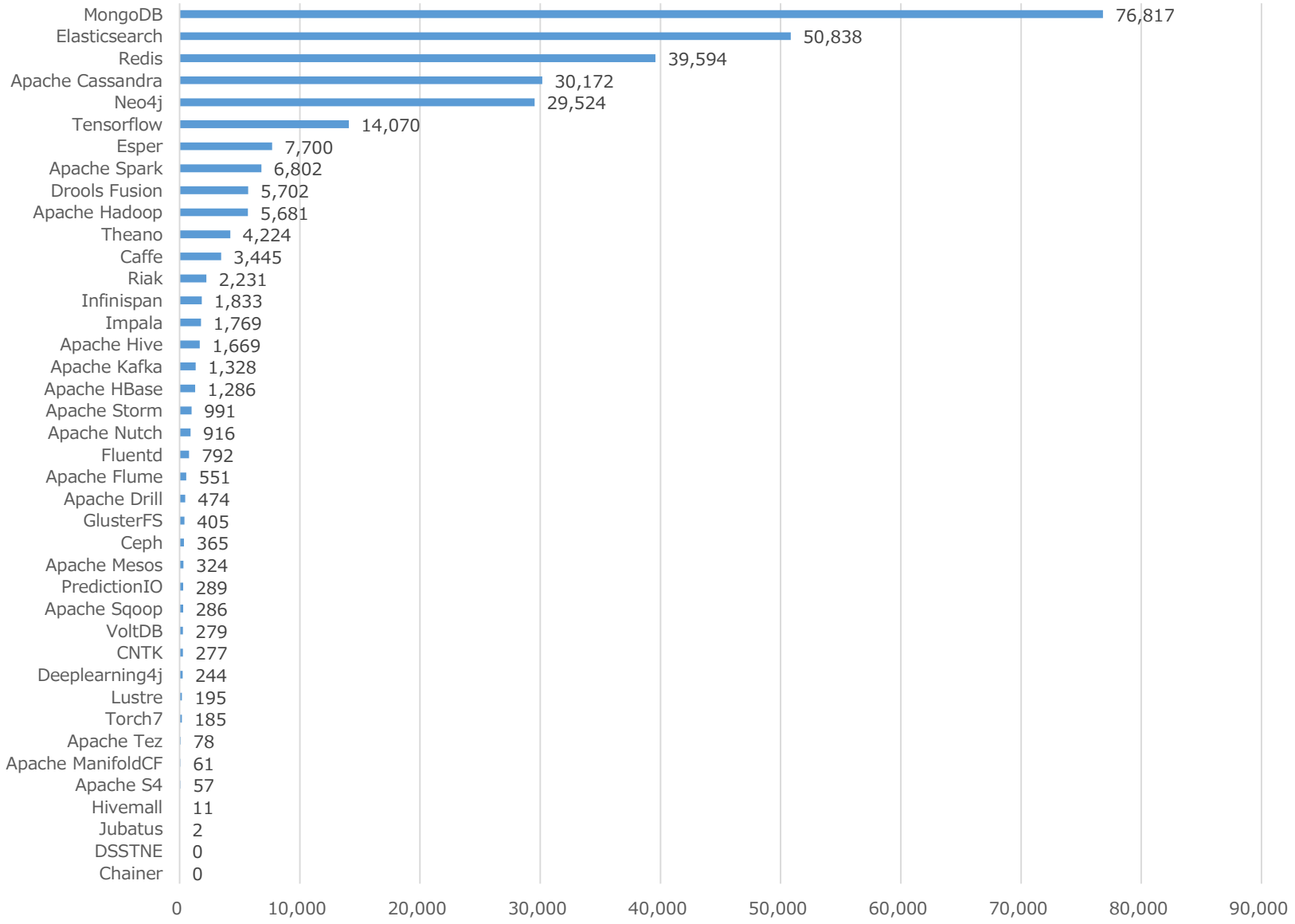
Seminars / workshops (Connpass)



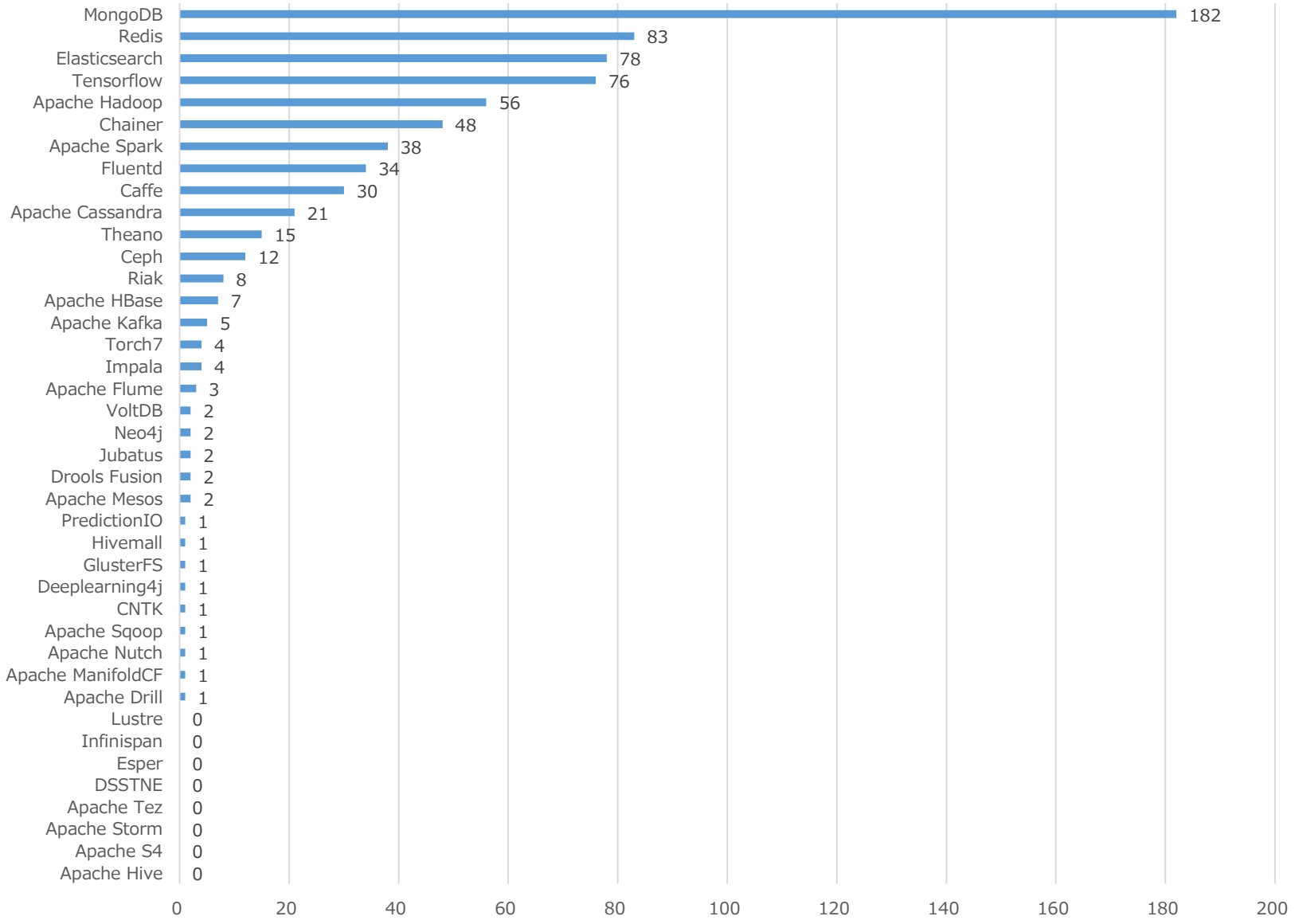
Slides (SlideShare)



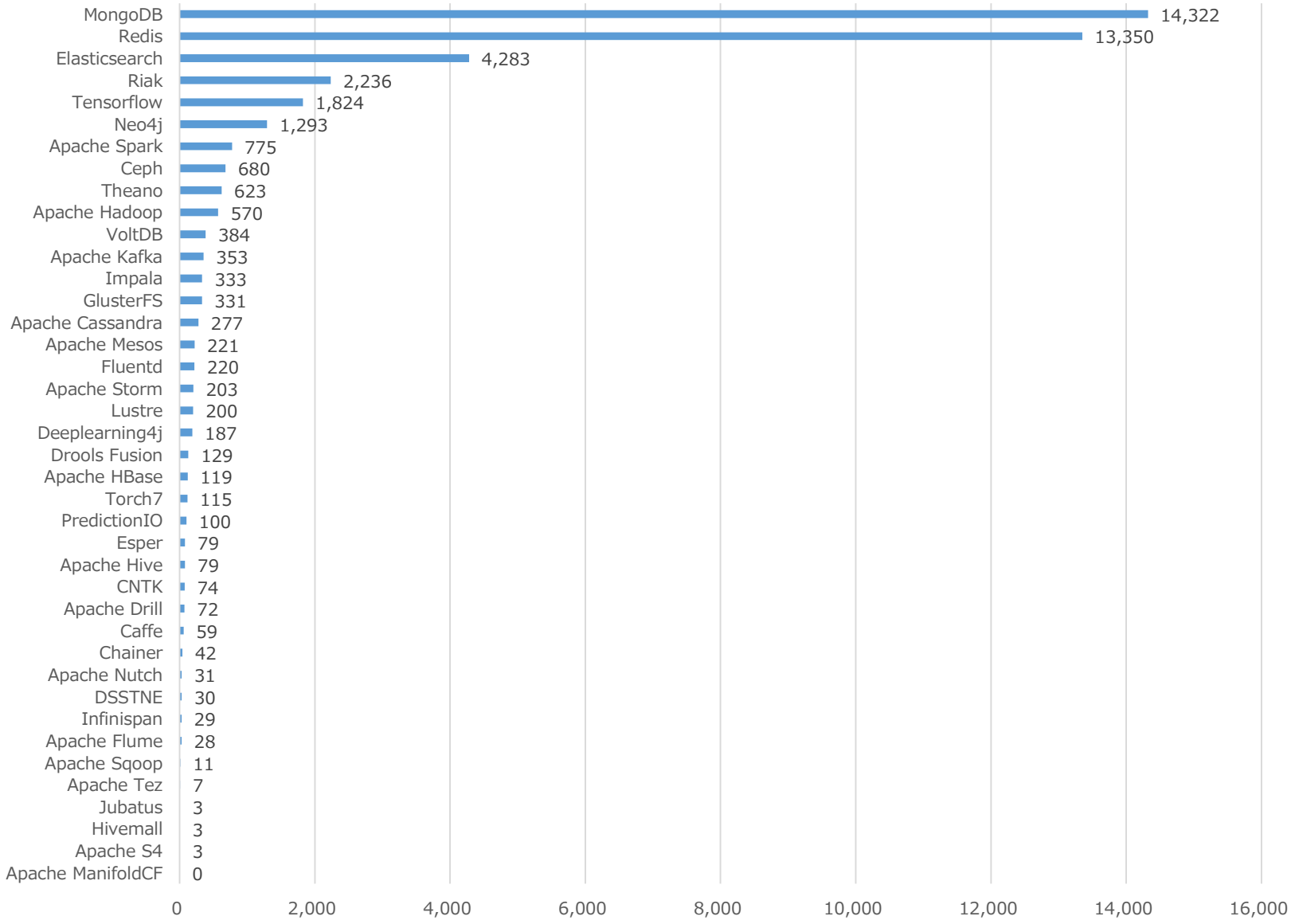
Questions (StackOverflow)



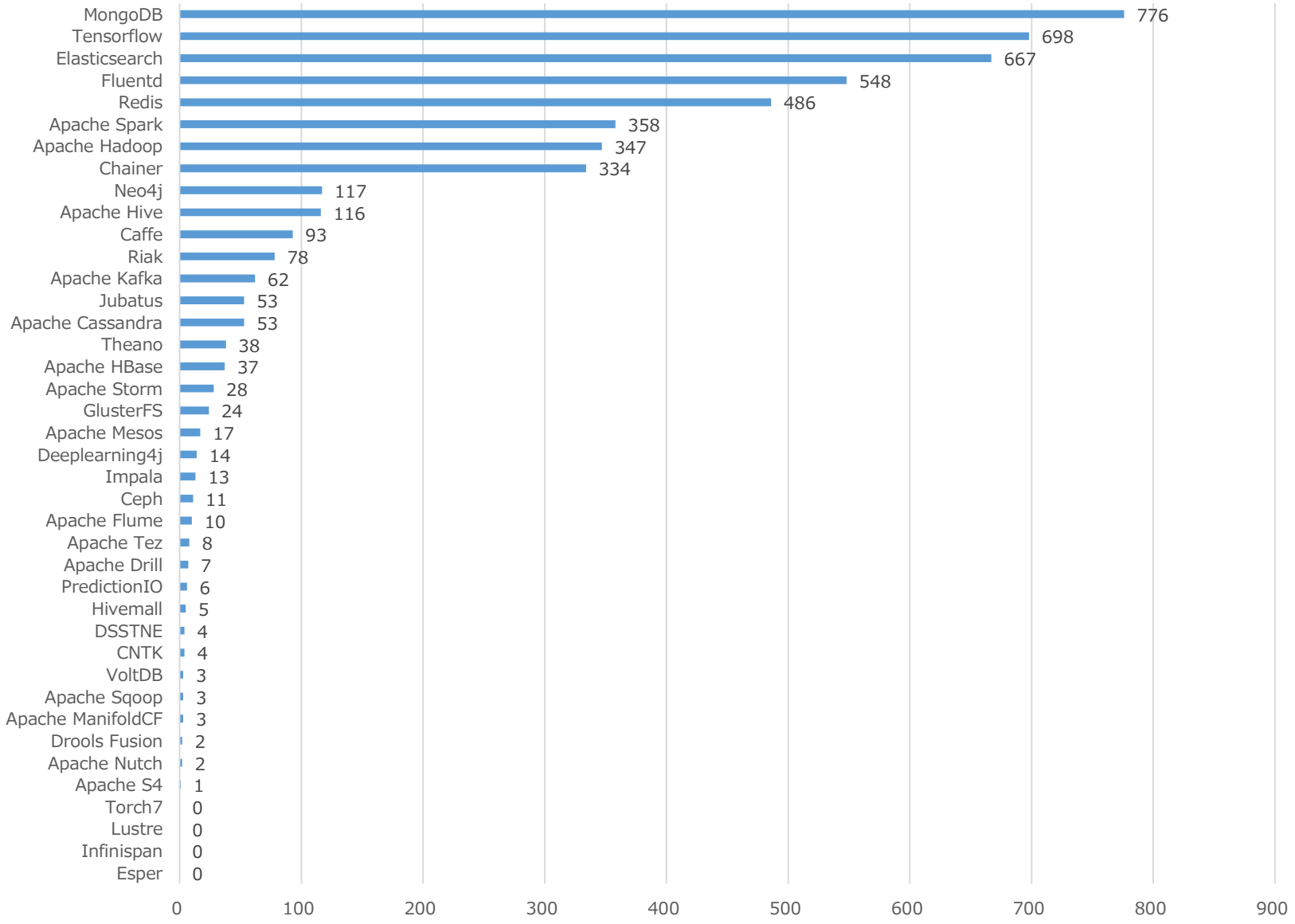
Questions (Other web services)



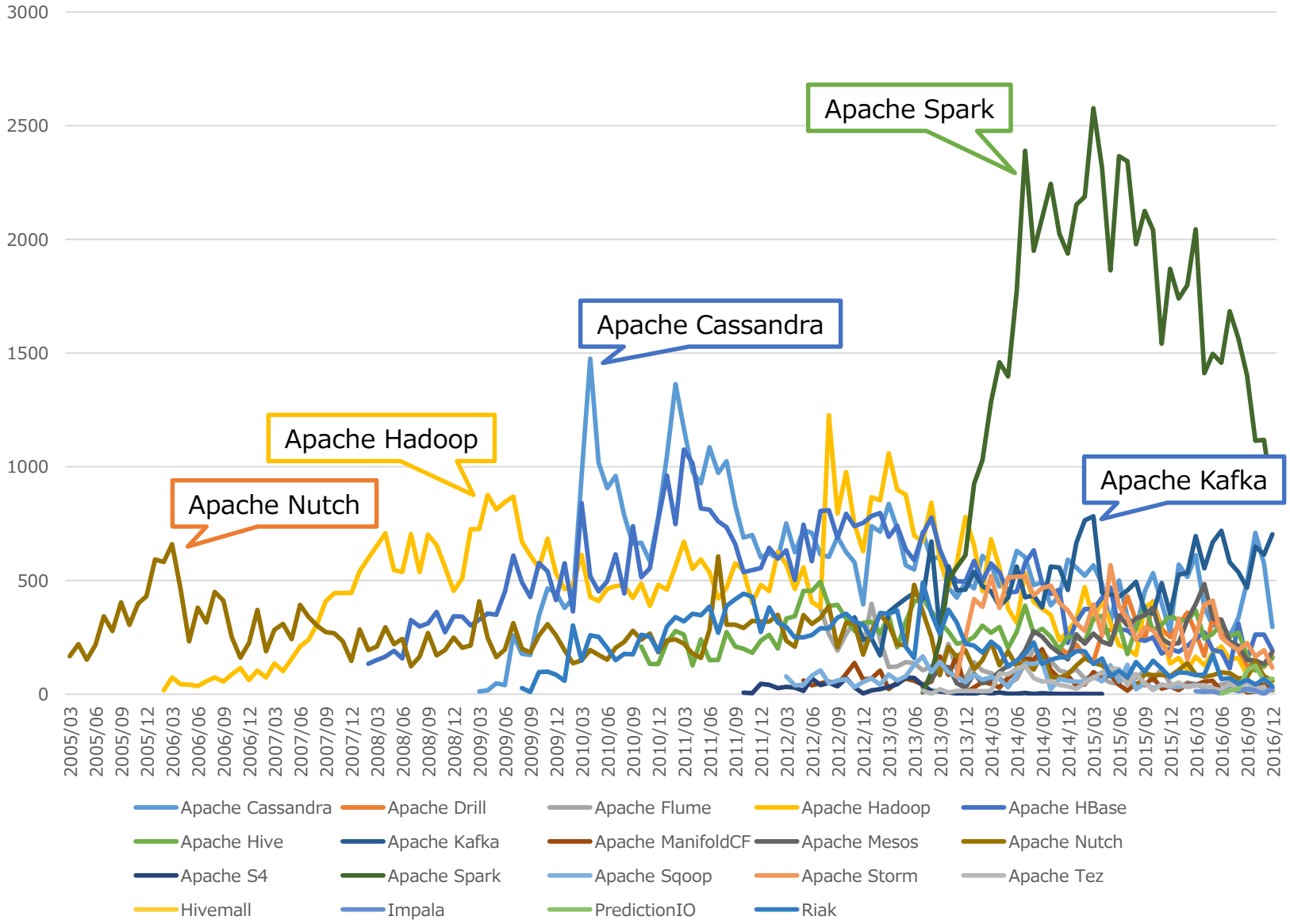
Technical posts (Hacker News)



Technical posts (Qiita)



Mails in mailing list for users



Ranking of “Adequacy of Technical Information”

1位	**MongoDB**
2位	**Redis**
3位	**TensorFlow**
4位	**Elasticsearch**
5位	**Apache Hadoop**

* compare with deviation value of each viewpoints

Summary

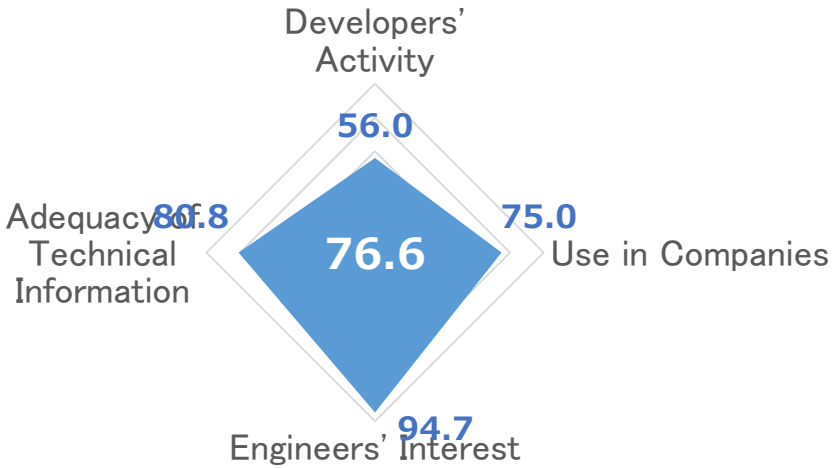
Overall ranking

1位	**MongoDB**
2位	**Redis**
3位	**Tensorflow**
4位	**Elasticsearch**
5位	**Apache Spark**

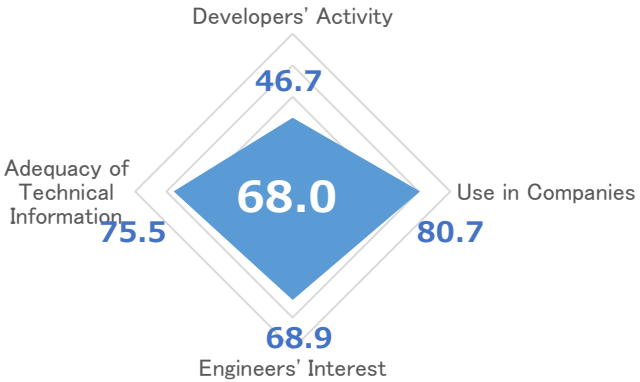
* compare with deviation value of each viewpoints

Overview of Top3

MongoDB



Redis



Tensorflow