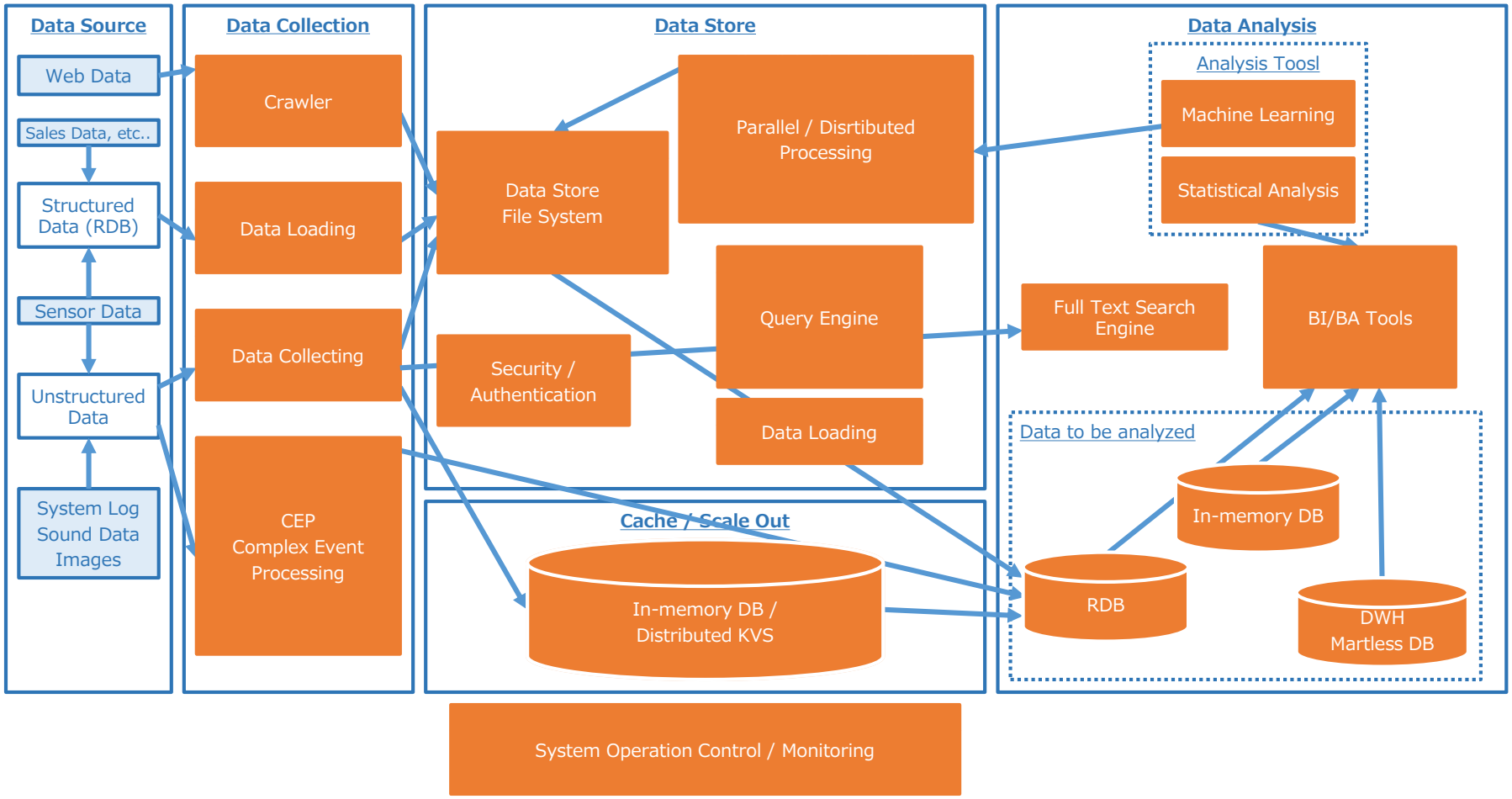
Survey of OSS in Big Data Platform

Contents

- Big Data Platform and Open Source Software
- Result of Data Analysis
  1. Developers' Activity
  2. Users' Activity
  3. Quality of Software

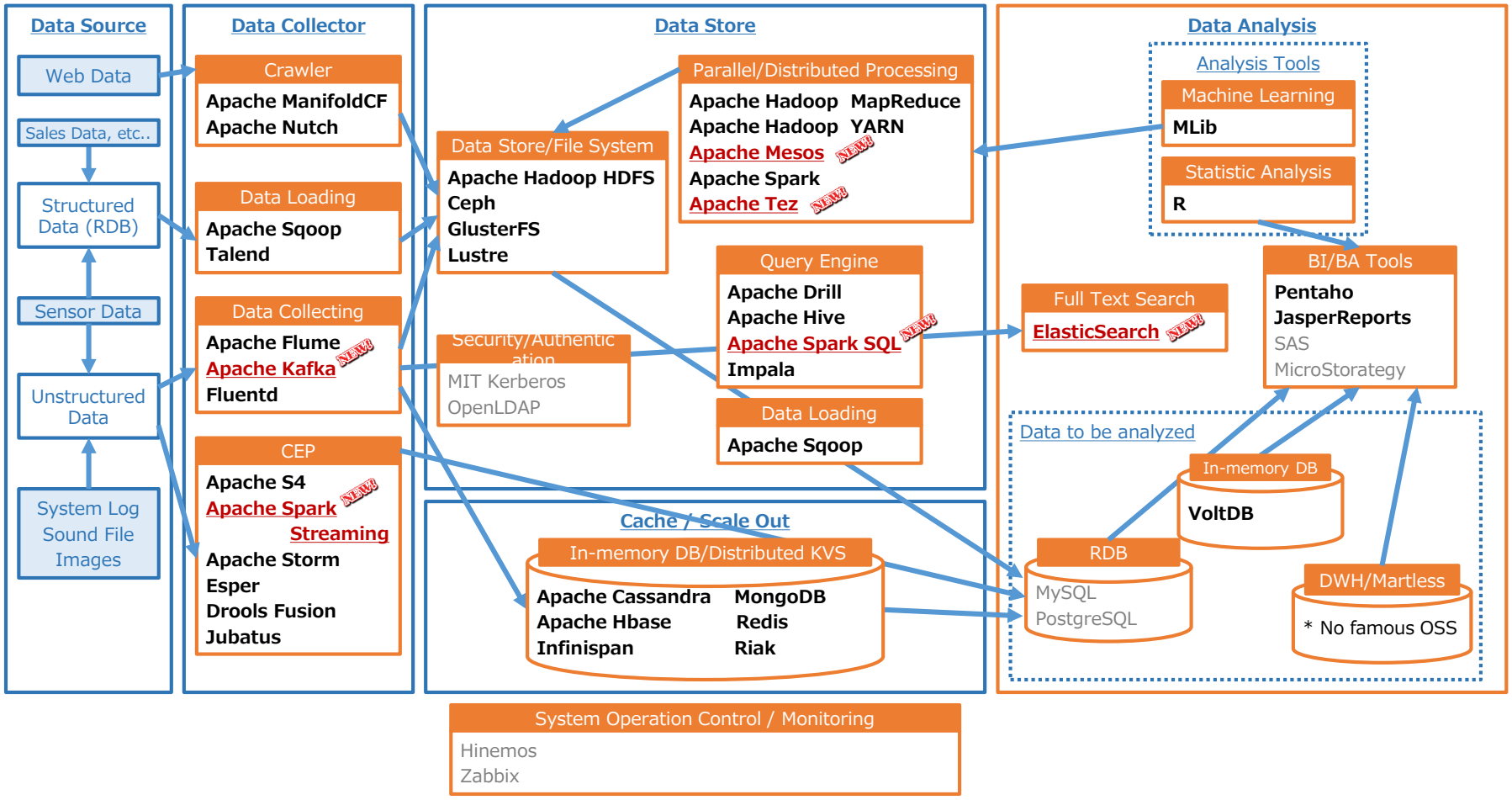
Introduction: Big Data Platform

- Most of the Big Data Platforms are composed of 3 functions : Data Collector, Data Store and Data Analysis



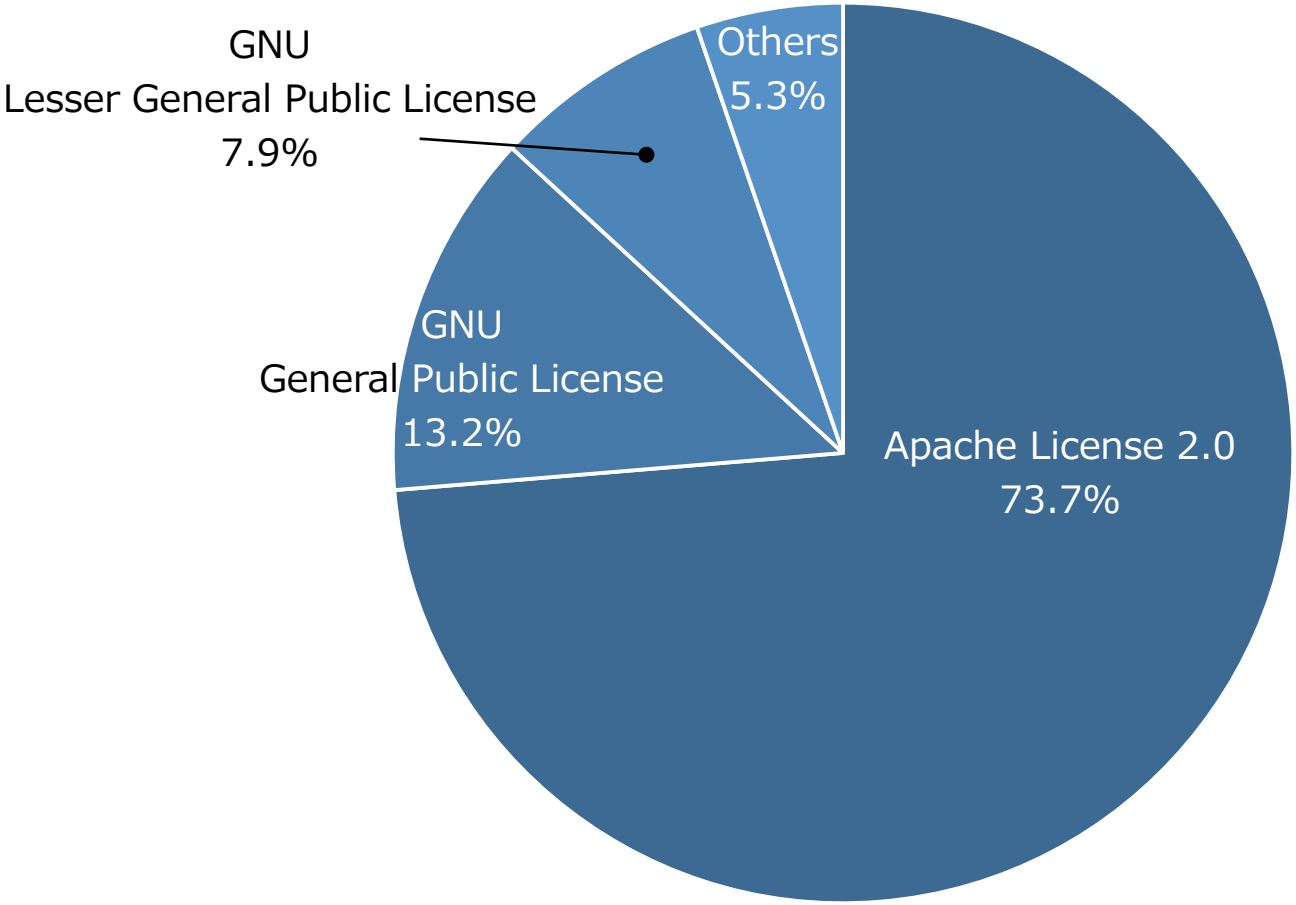
Introduction: OSS in Big Data Platform

- There are some Big Data Platform using only OSS
- Quality, Maturity and Functions of software are different, so **we need to decide** which software we should use



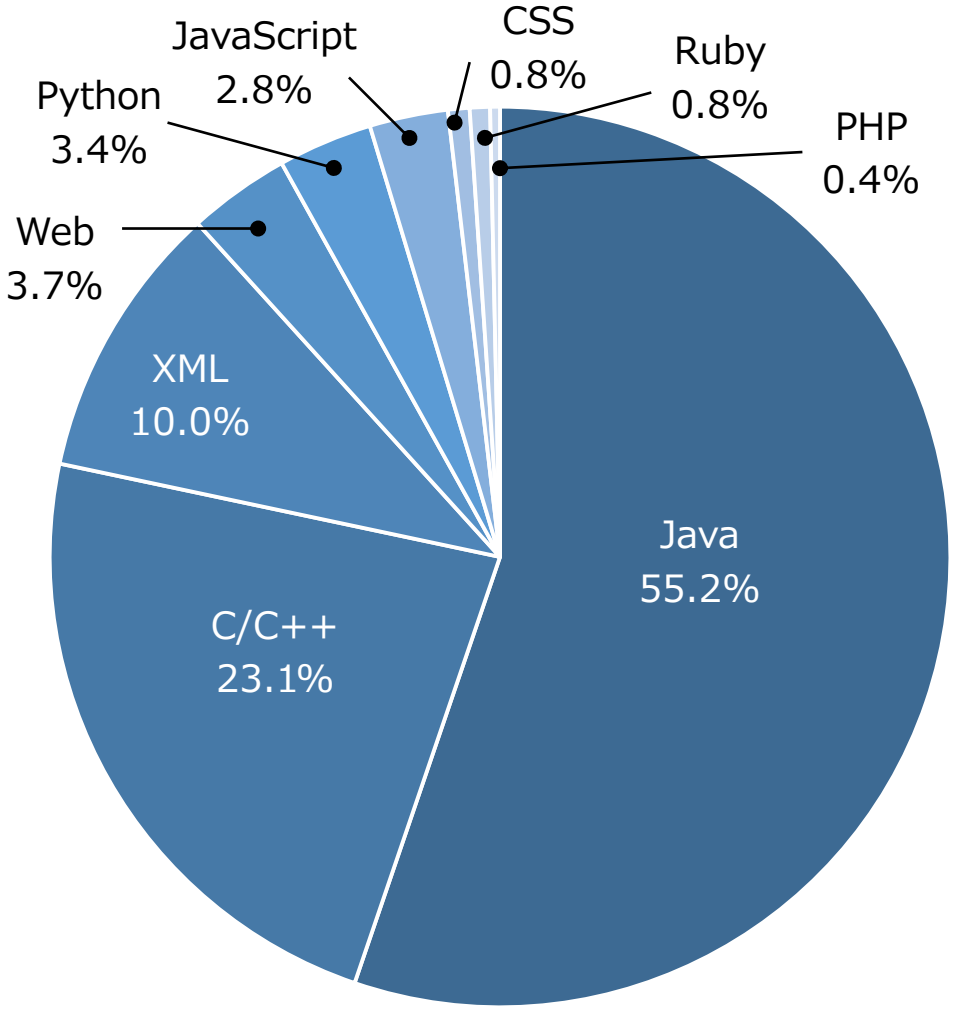
Licenses

- Over 70% of software uses Apache License 2.0
- One reason is that there are a large number of software under the Apache Software Foundation



Programming Language

- Java is the most popular language used in OSS (55.2%)
  - Second is C/C++ (23.1%)
  - The most popular *script language* is Python (3.4%)

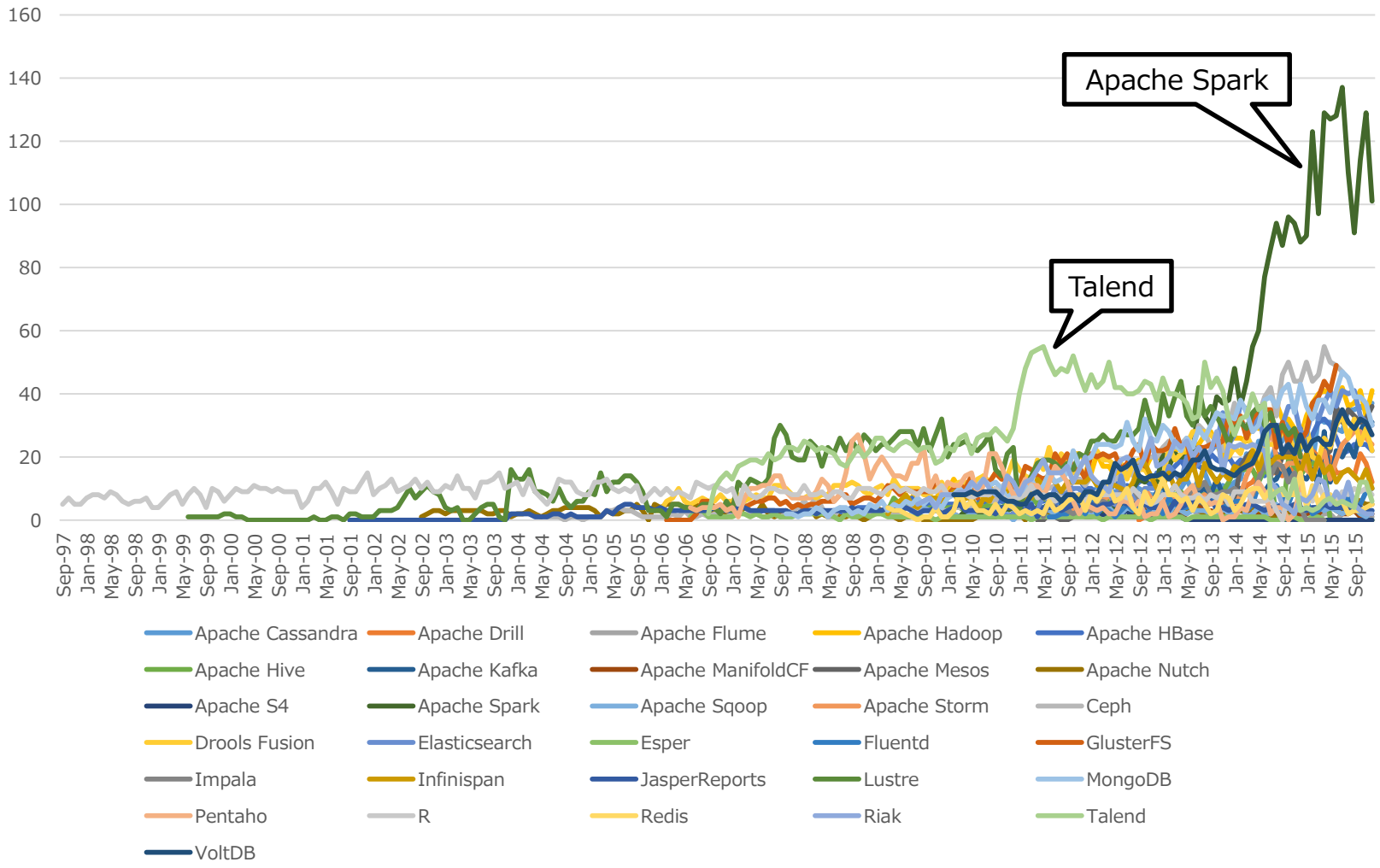


Developers' Activity

How active are developers?

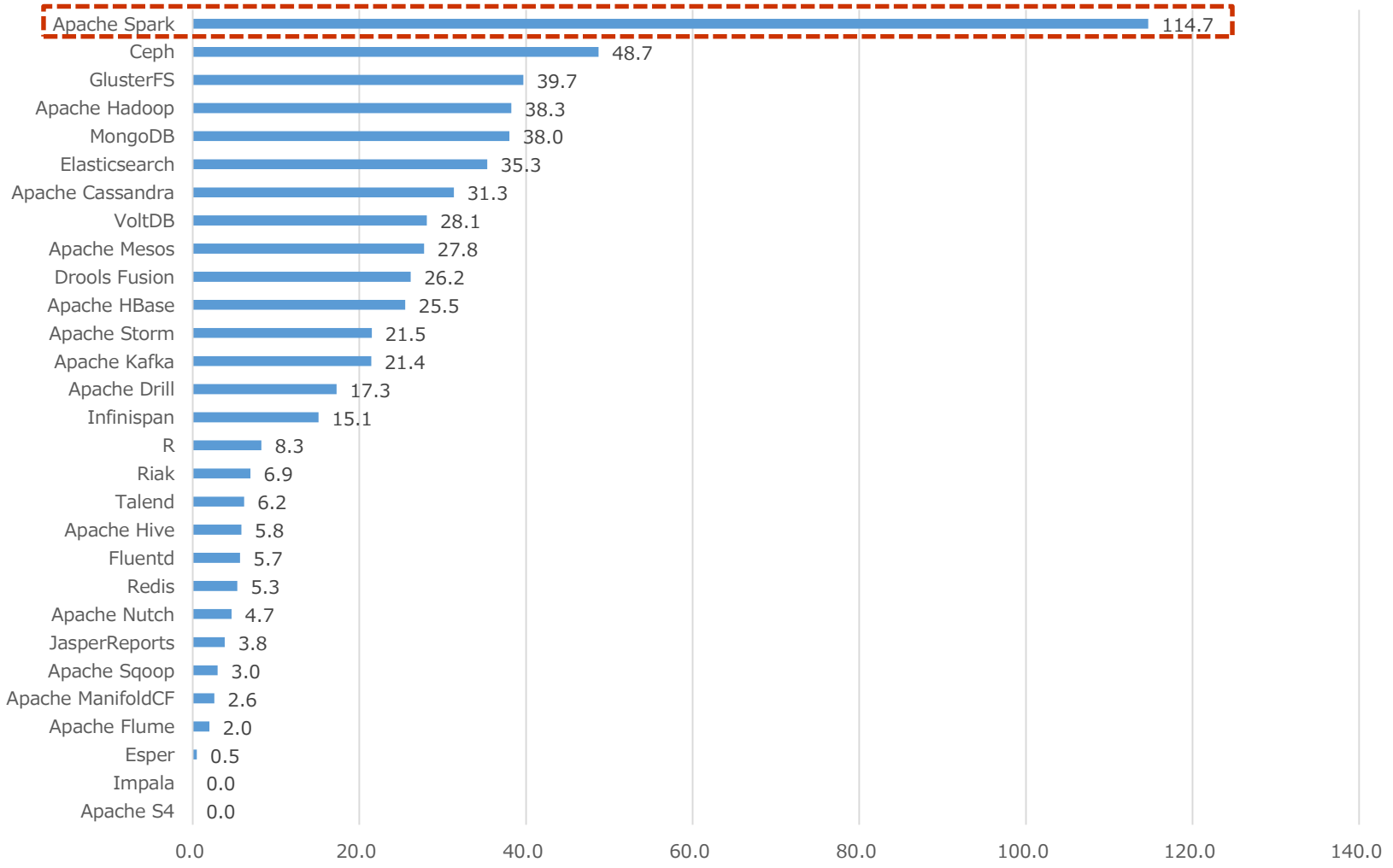
Number of committers (monthly)

- Apache Spark is rapidly increasing from 2014



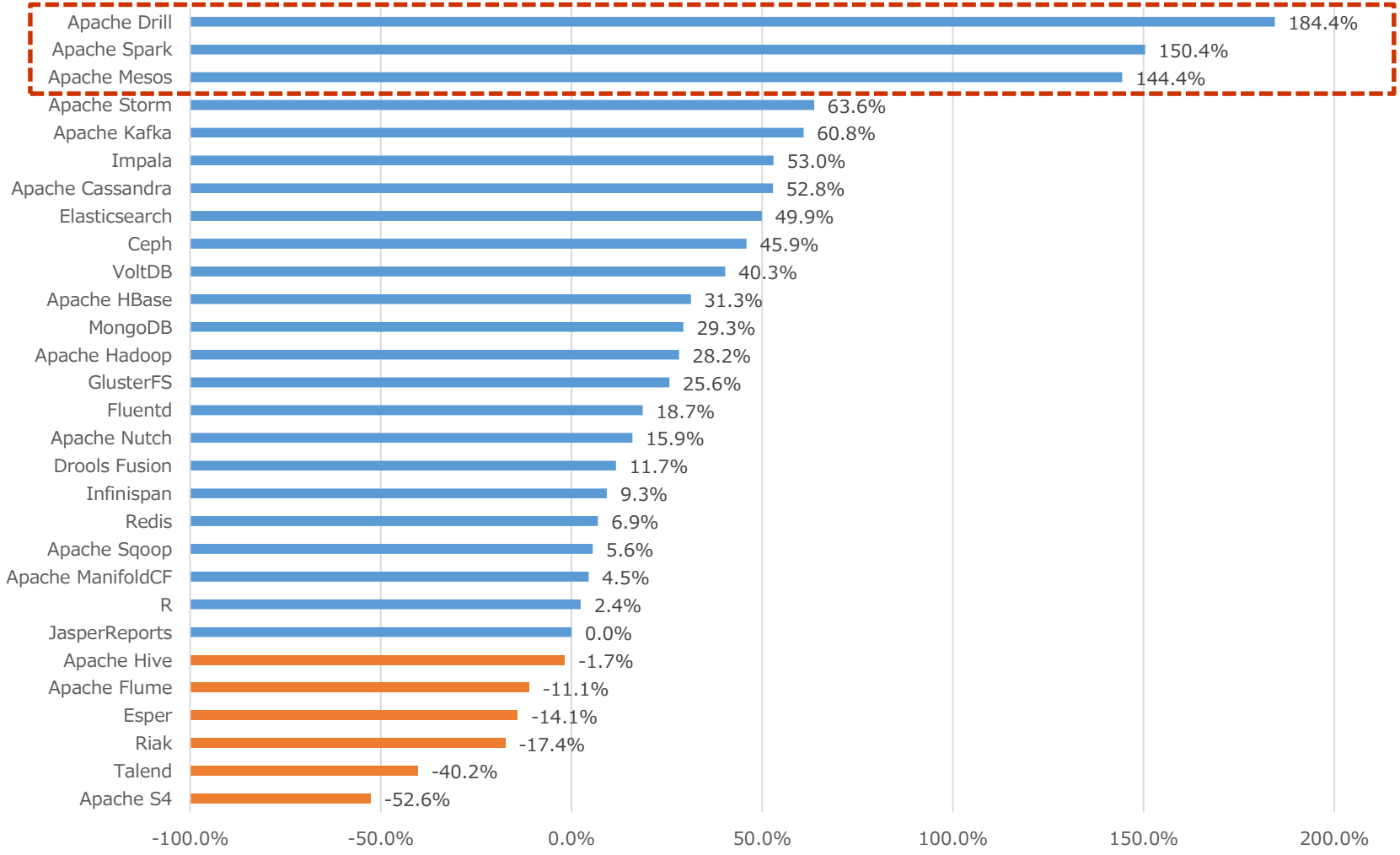
Average number of committers in month (2015)

- Over 100 committers make some commits to Apache Spark
  - 2013 : 27.8 → 2014 : 72.1 → 2015 : 114.7

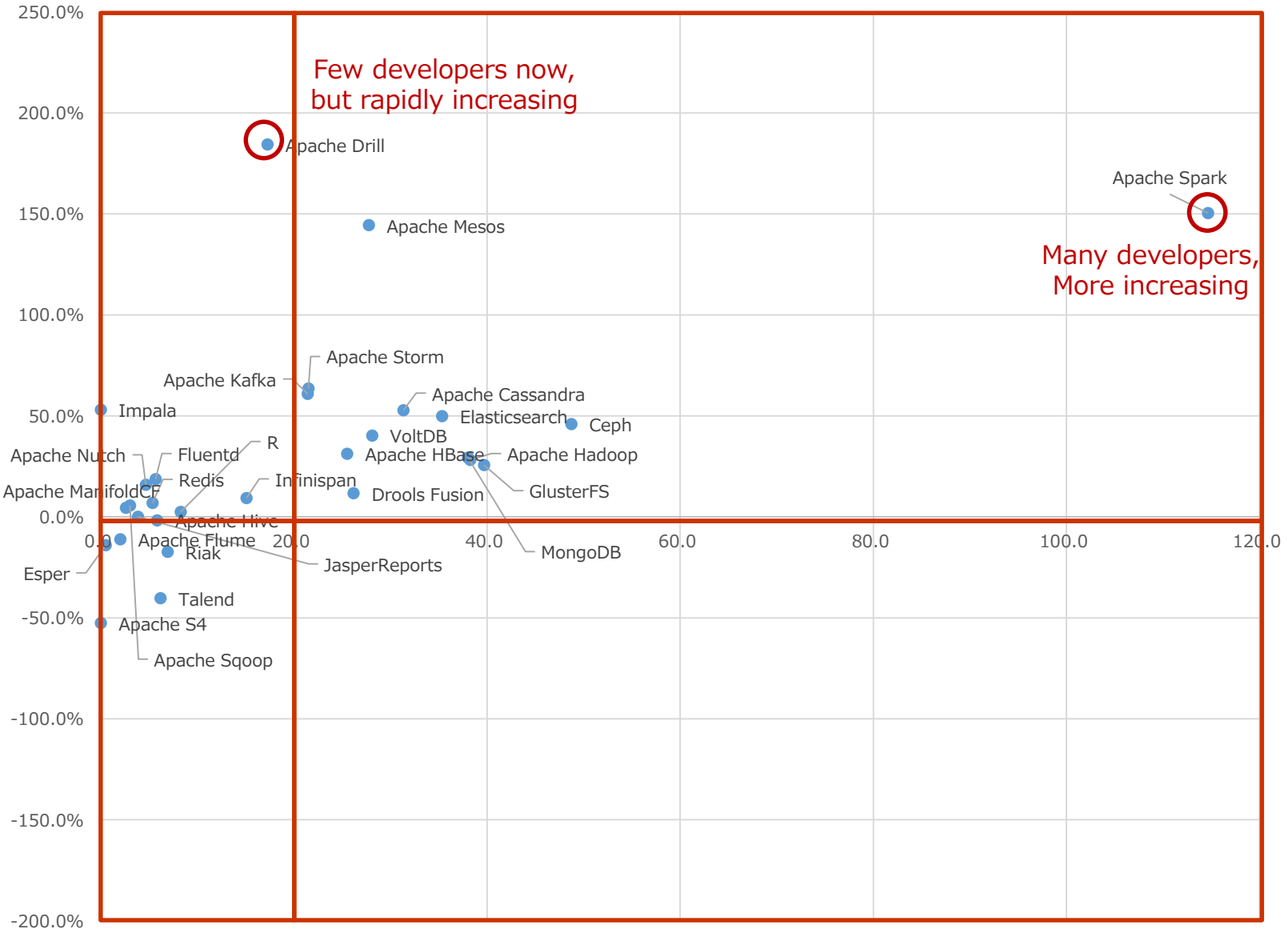


CAGR of committers (2011-2015)

- Apache Drill is the fastest-growing
  - Apache Spark and Apache Mesos are also growing

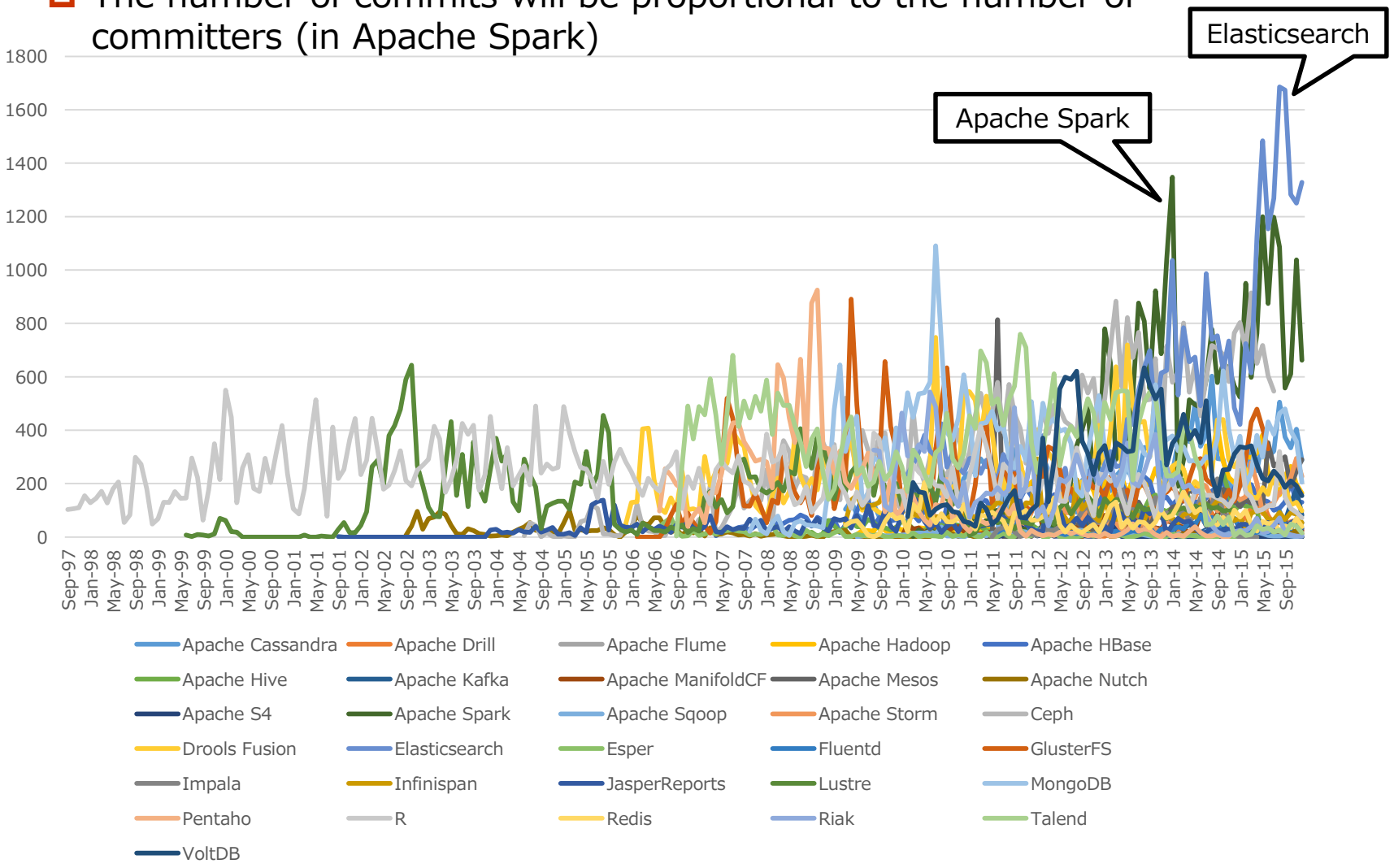


Average committers (x-axis) and CAGR (y-axis)



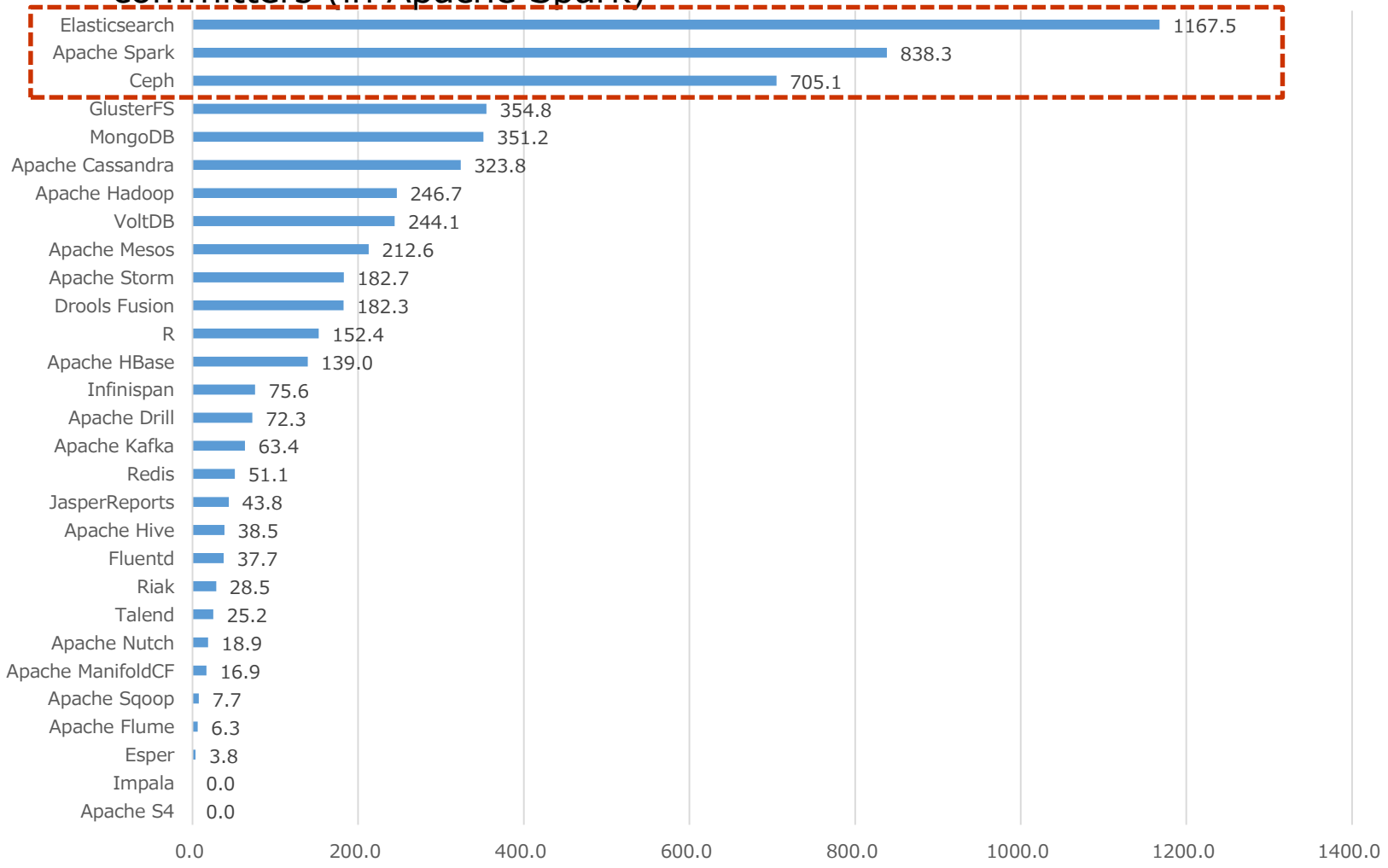
Number of Commits

- Recently, Elasticsearch gets the greatest number of commits
  - The number of commits will be proportional to the number of committers (in Apache Spark)

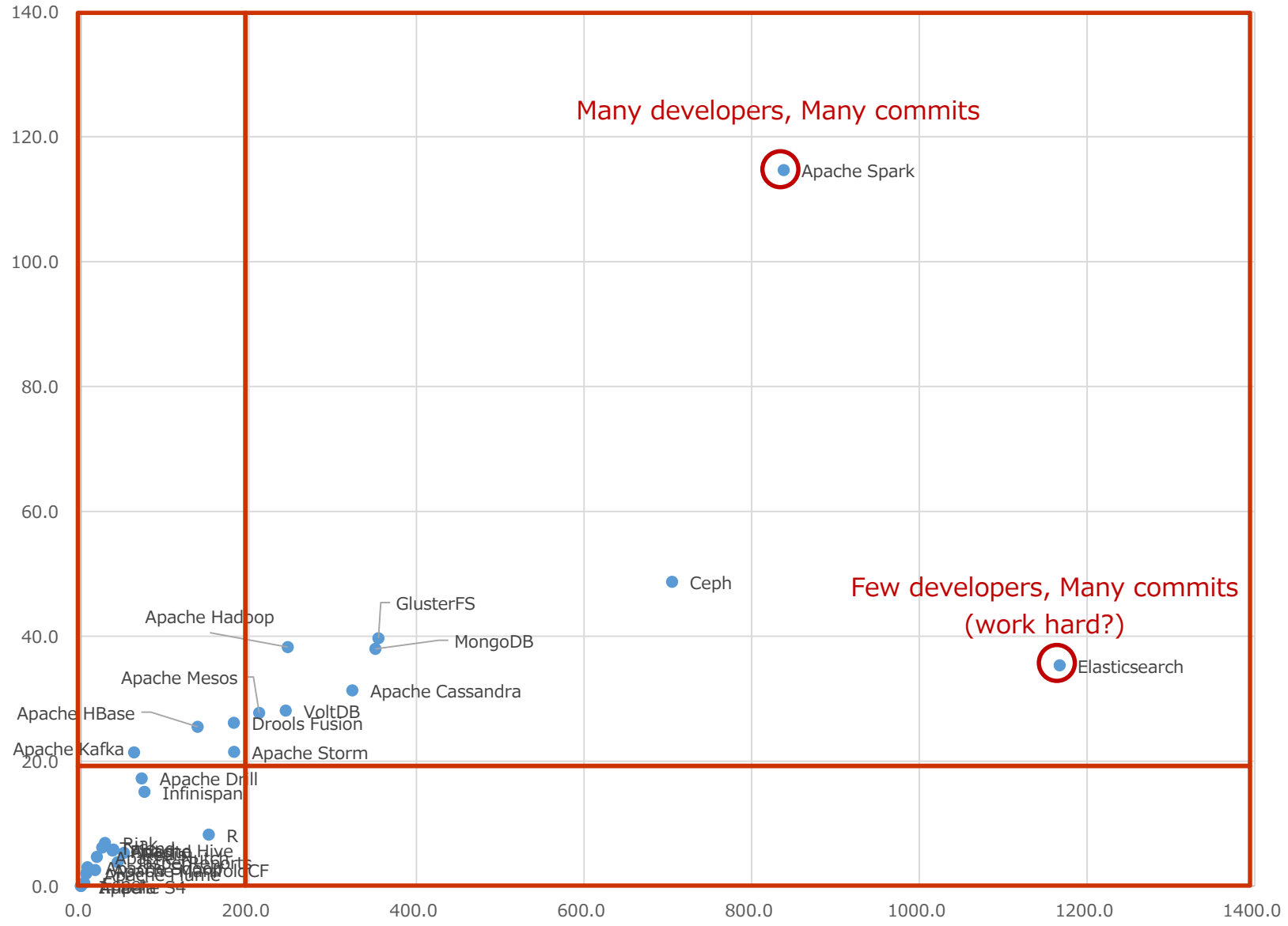


Average number of commits in month (2015)

- Elasticsearch got the largest number of commits
  - The number of commits will be proportional to the number of committers (in Apache Spark)

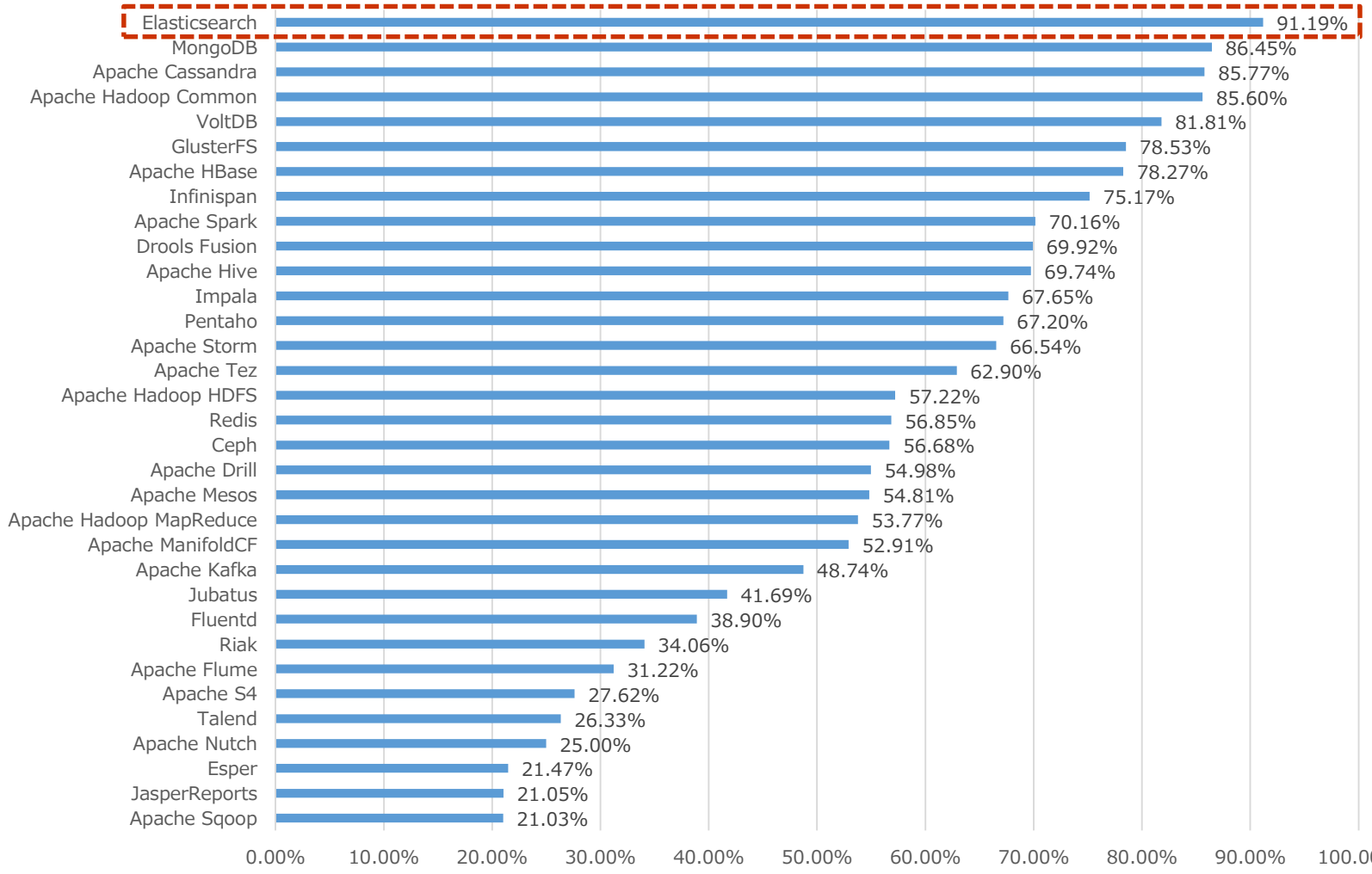


Commits (x-axis) and Committers (y-axis)

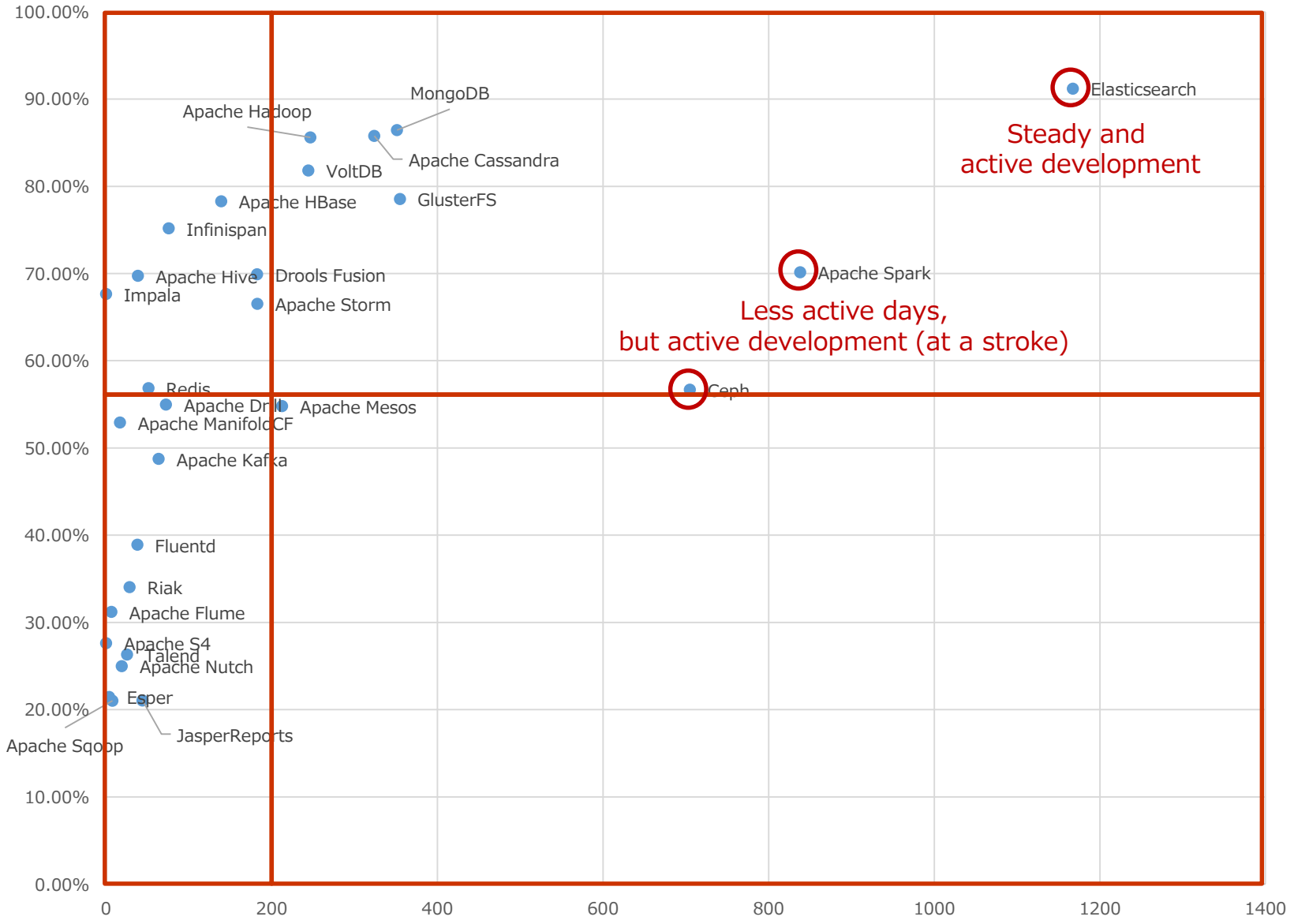


Ratio of active days

- Active days : days with commits in Git
  - Elasticsearch has the most active days in these software

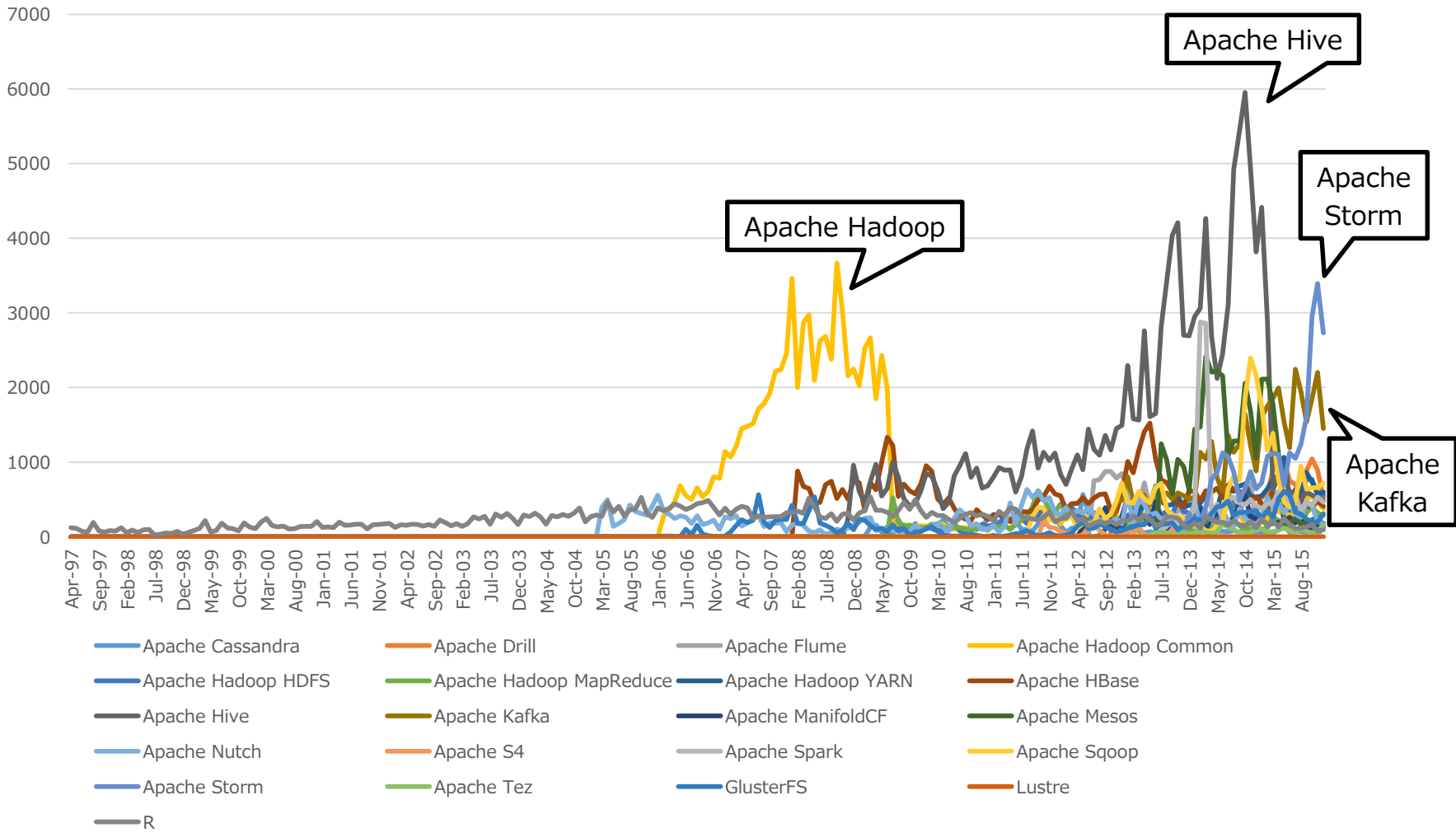


Commits (x-axis) ratio of active days (y-axis)



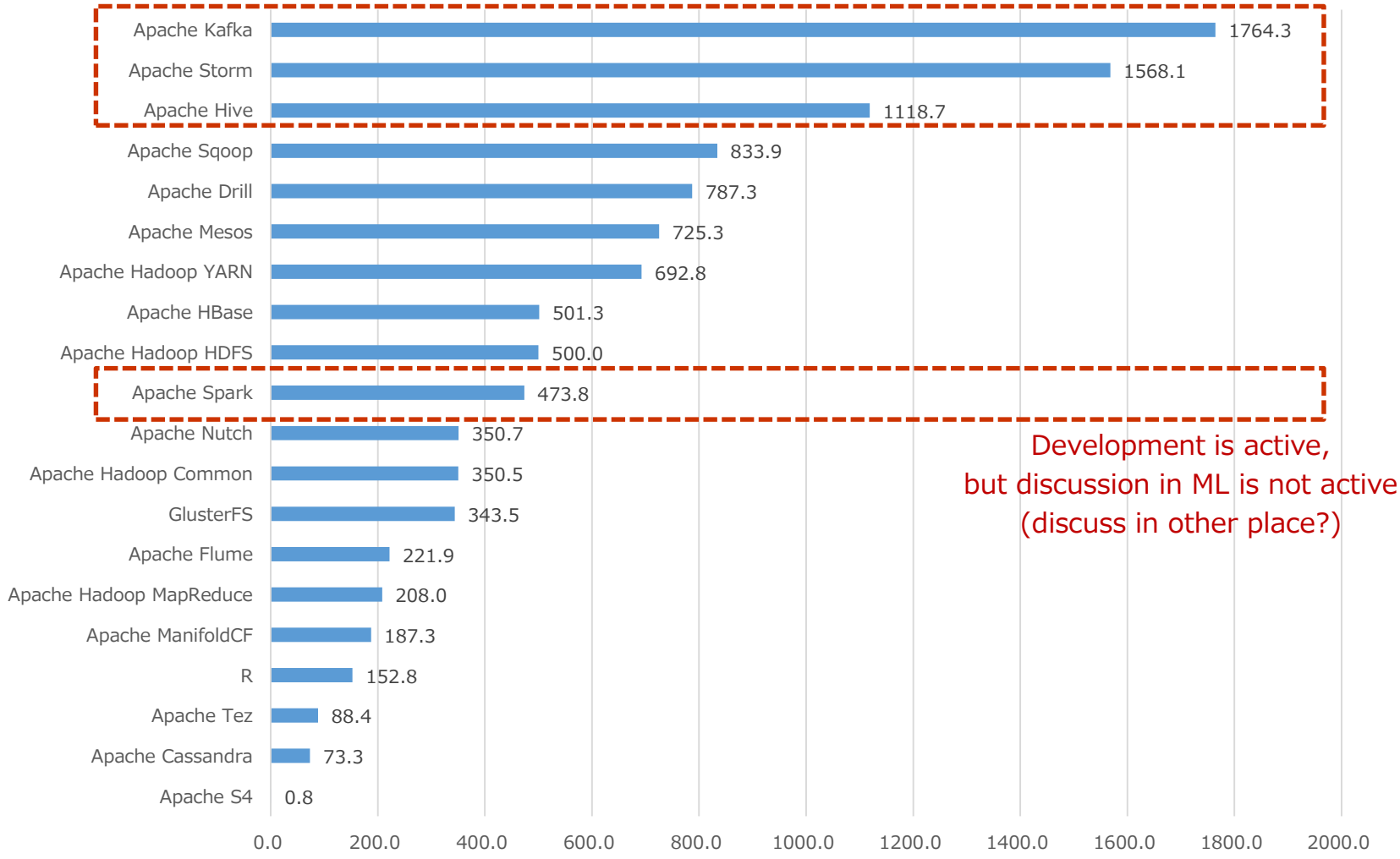
Number of mails in Mailing list for developers

- Discussion about Apache Hadoop was active in 2006 to 2009
  - Recently Apache Hive, Apache Storm and Apache Kafka are also active

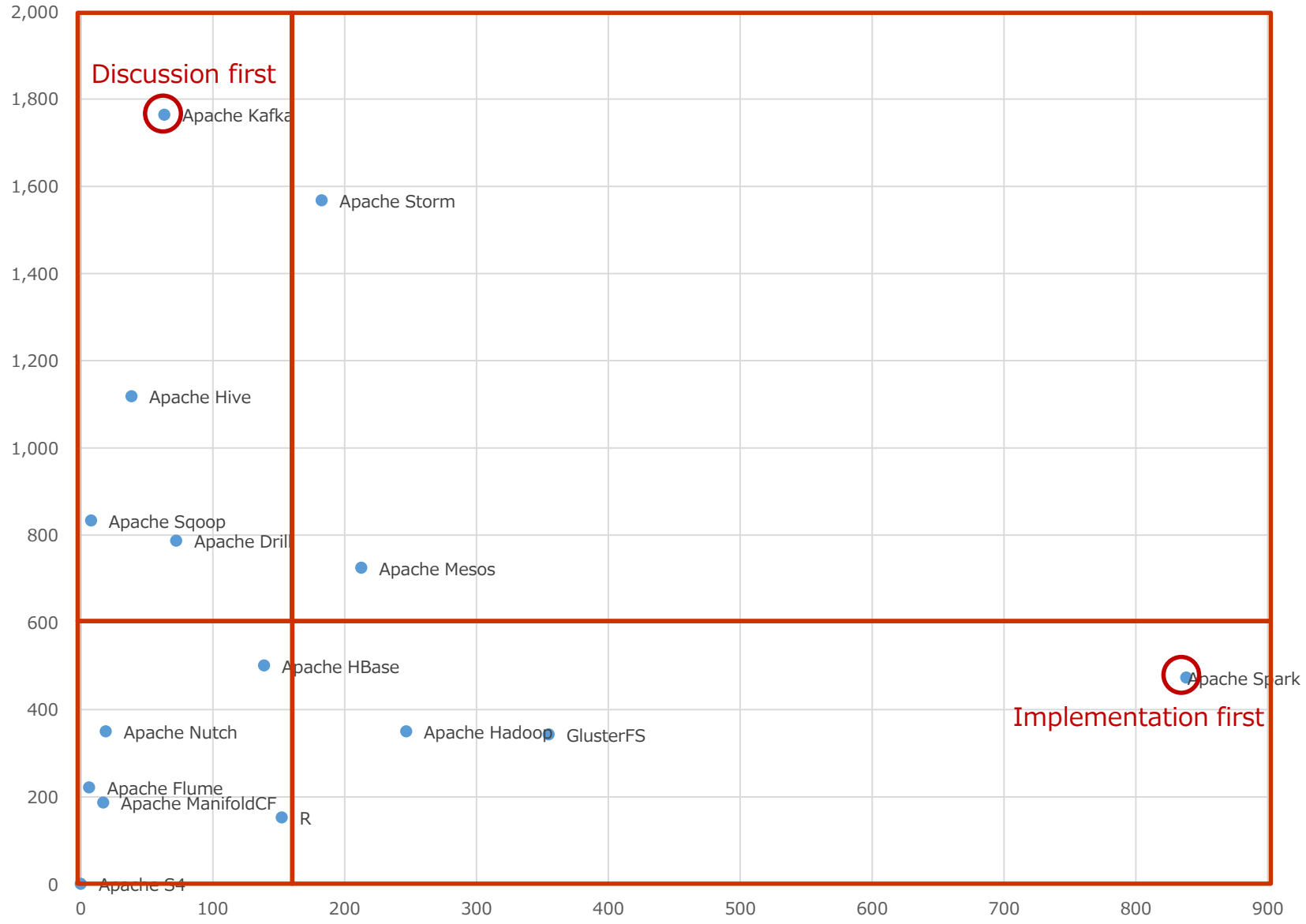


Average number of mails in ML for developers (2015)

- Developers actively discuss in mailing list of Apache Kafka, Apache Storm and Apache Hive

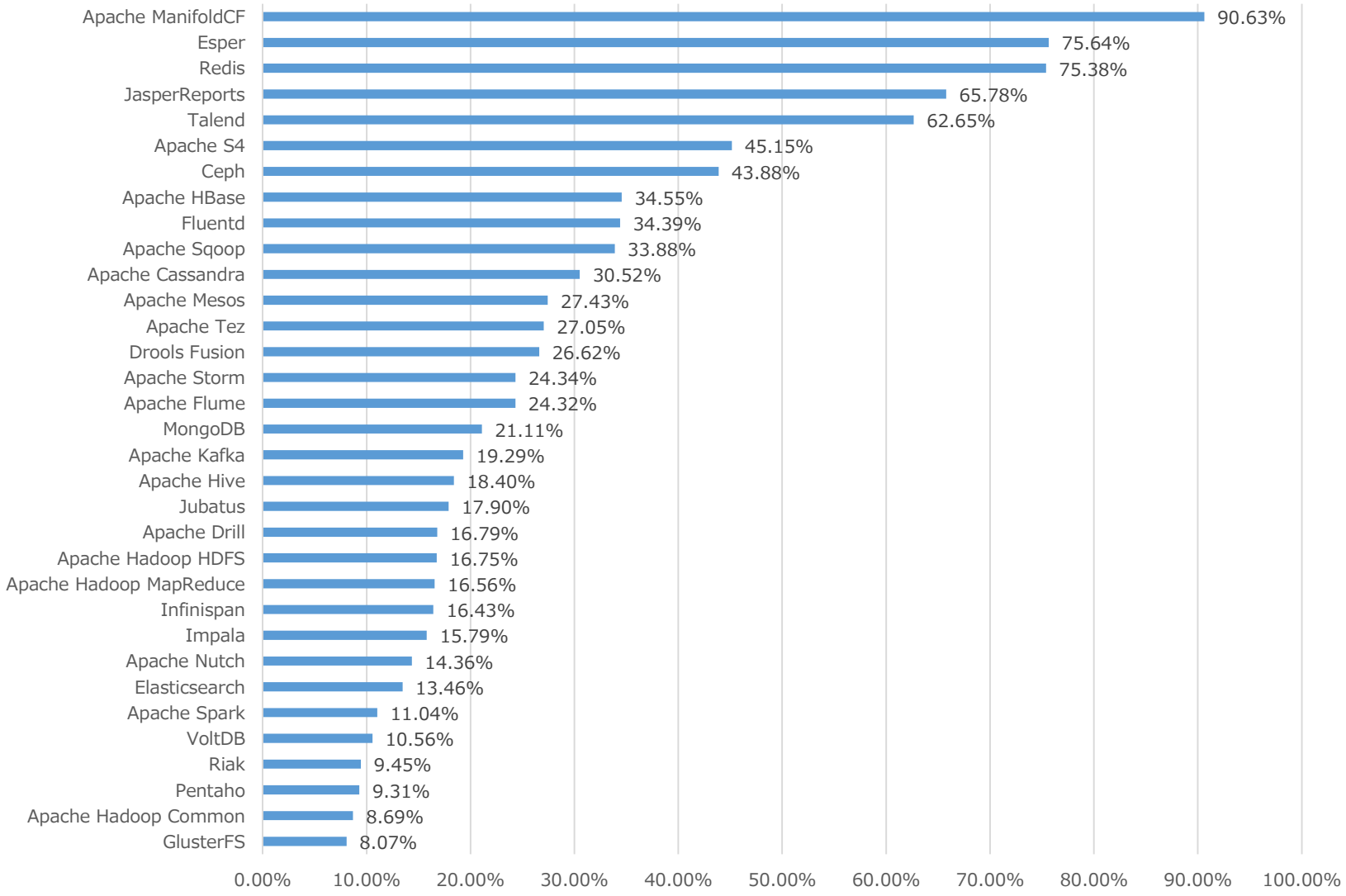


Correlation between commits and mails



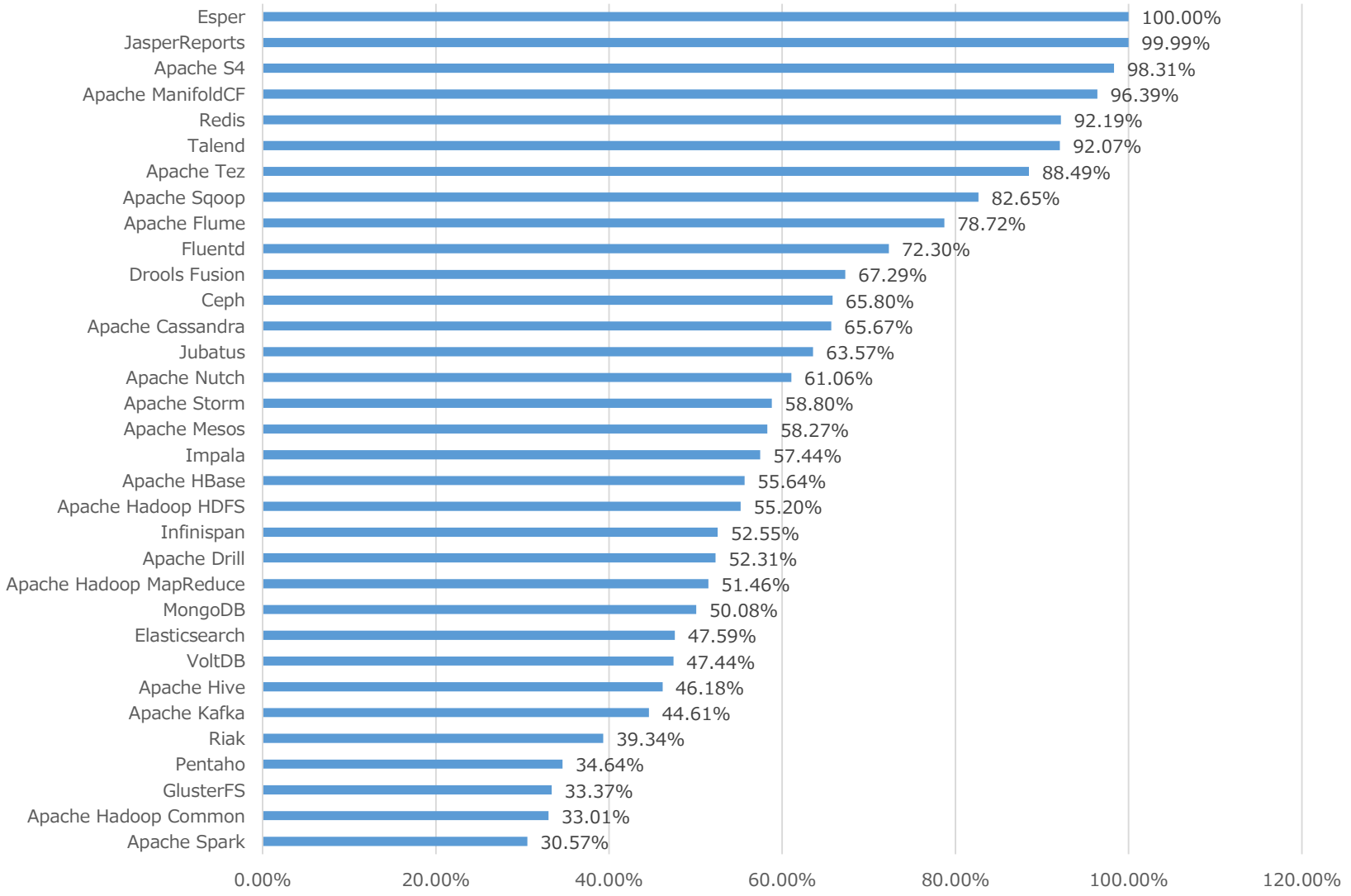
Ratio of commits by the top committer

- This ratio may show the power of top committer in the project (?)

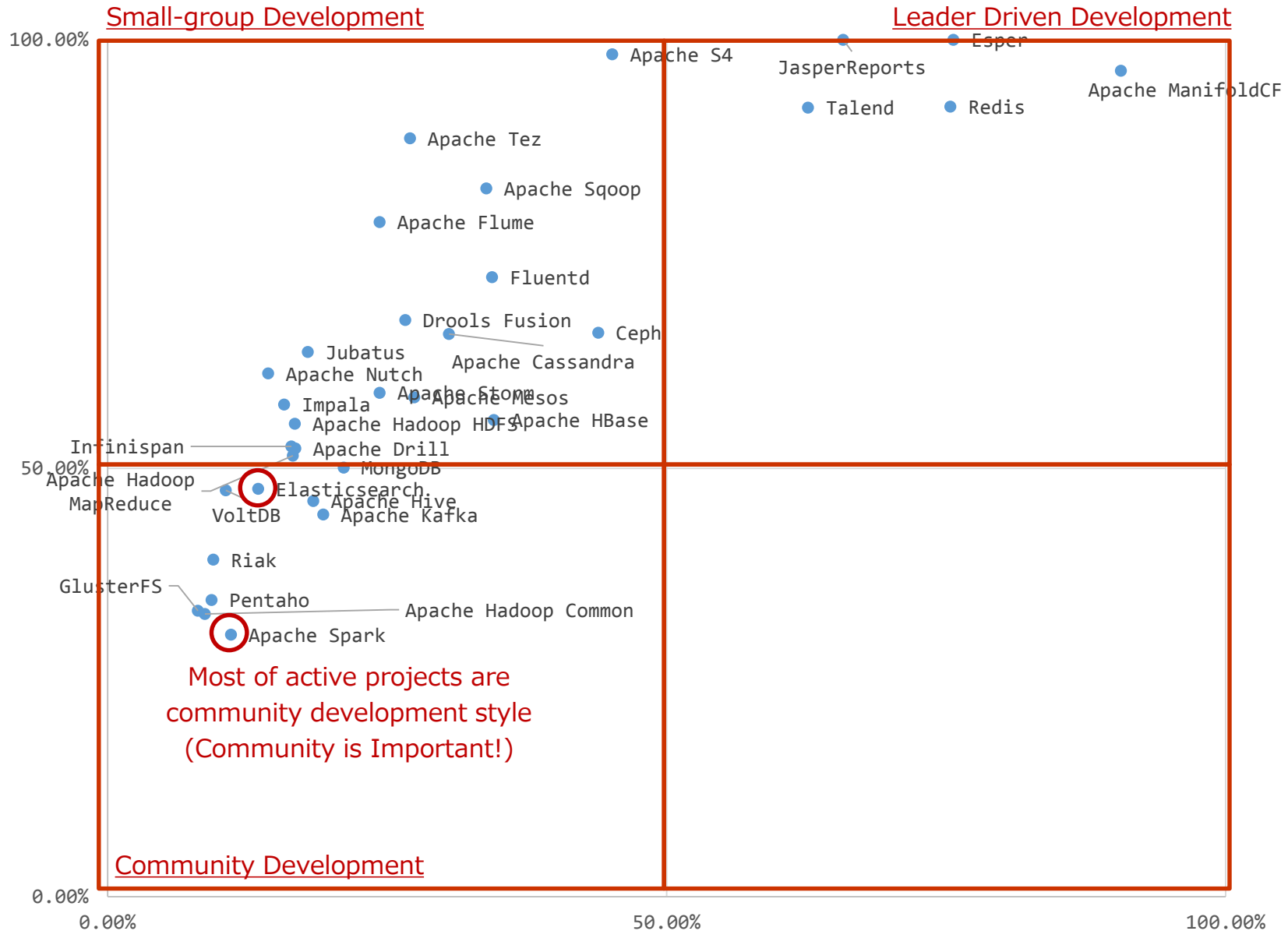


Ratio of commits by the top 1-5 committers

- Esper and JasperReports may be committed by only 5 committers



Top1 (x-axis) and Top1-5 (y-axis)

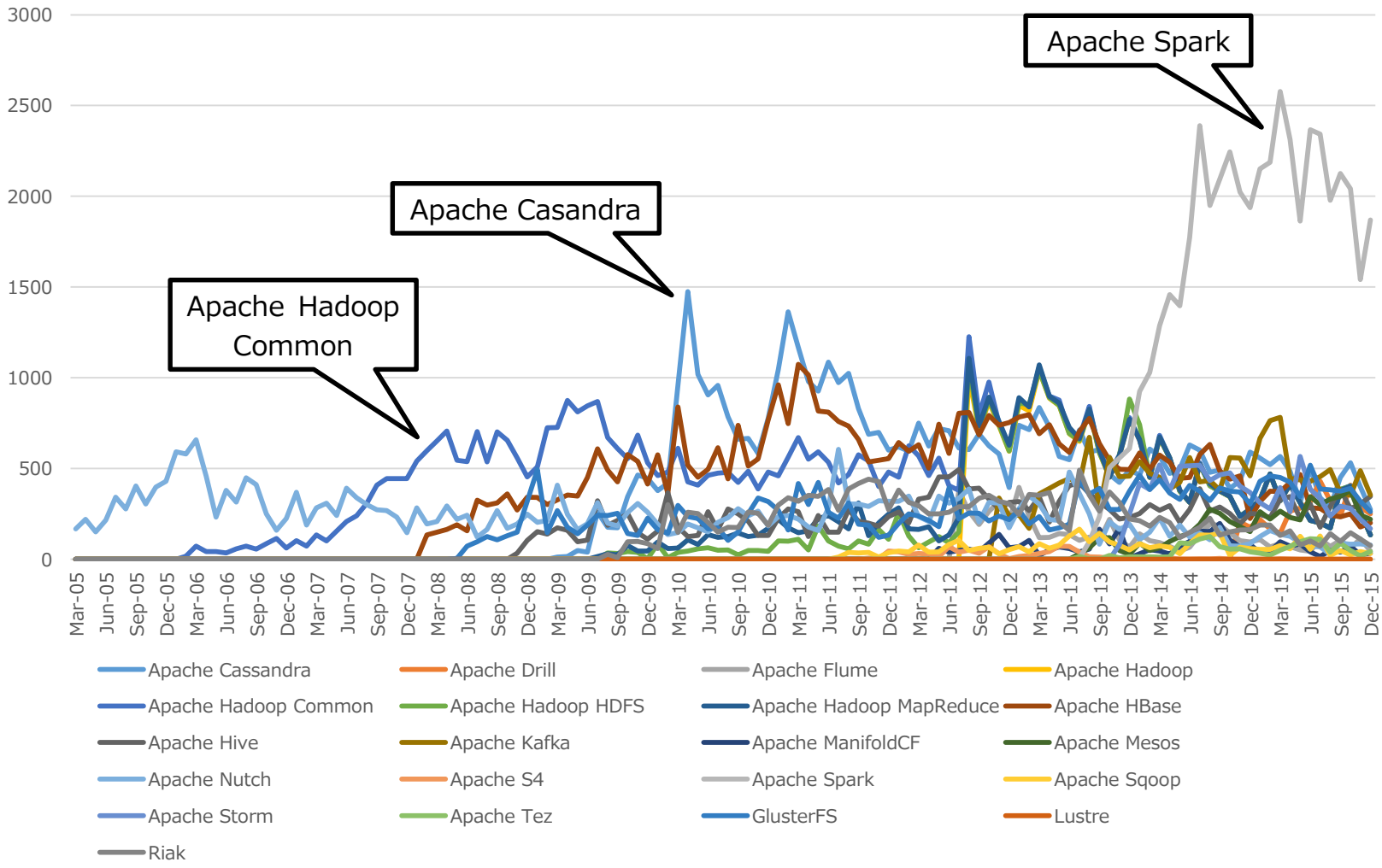


Users' Activity

How popular is the software?

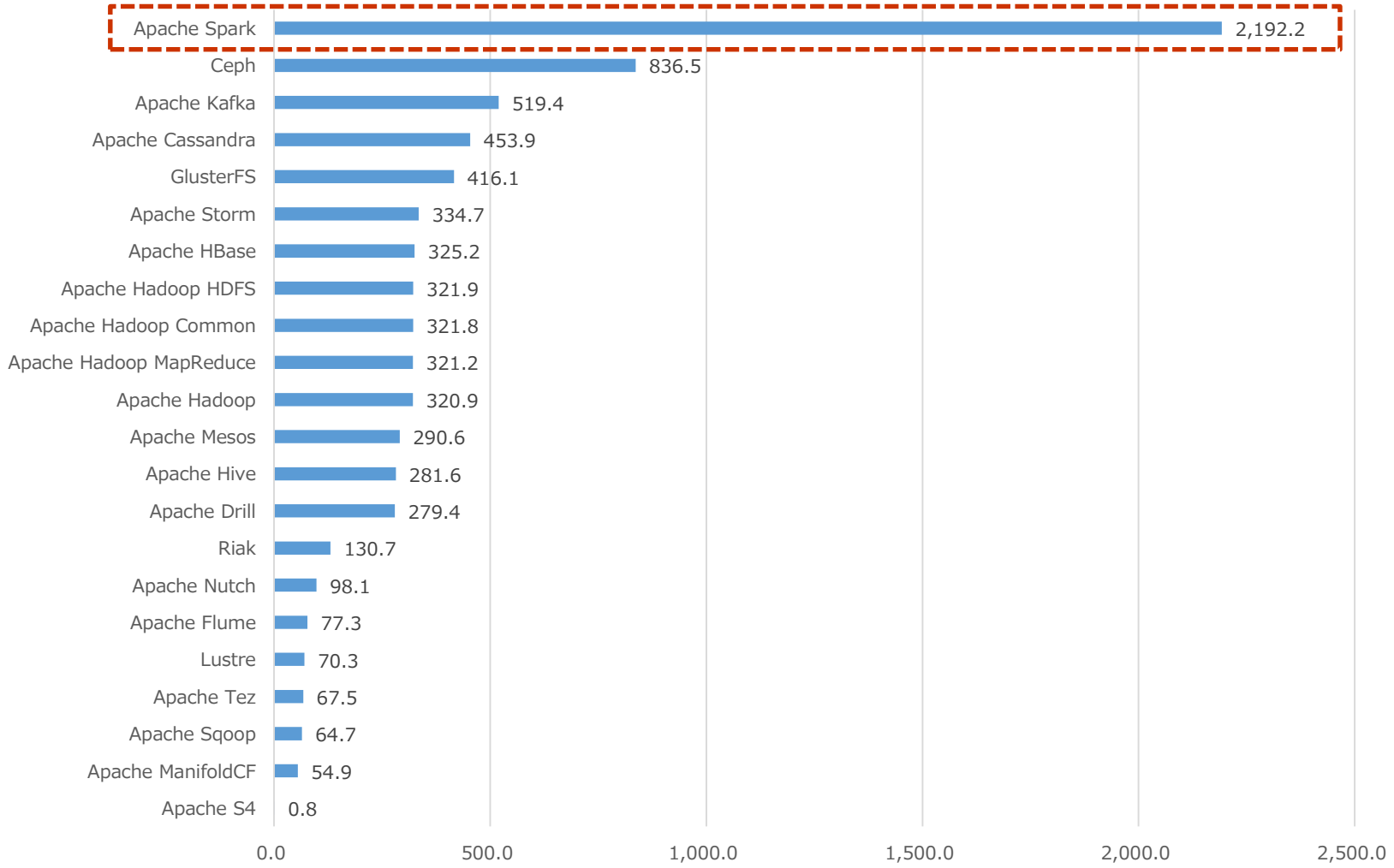
Mails in ML for users

- Apache Spark has very active user mailing list

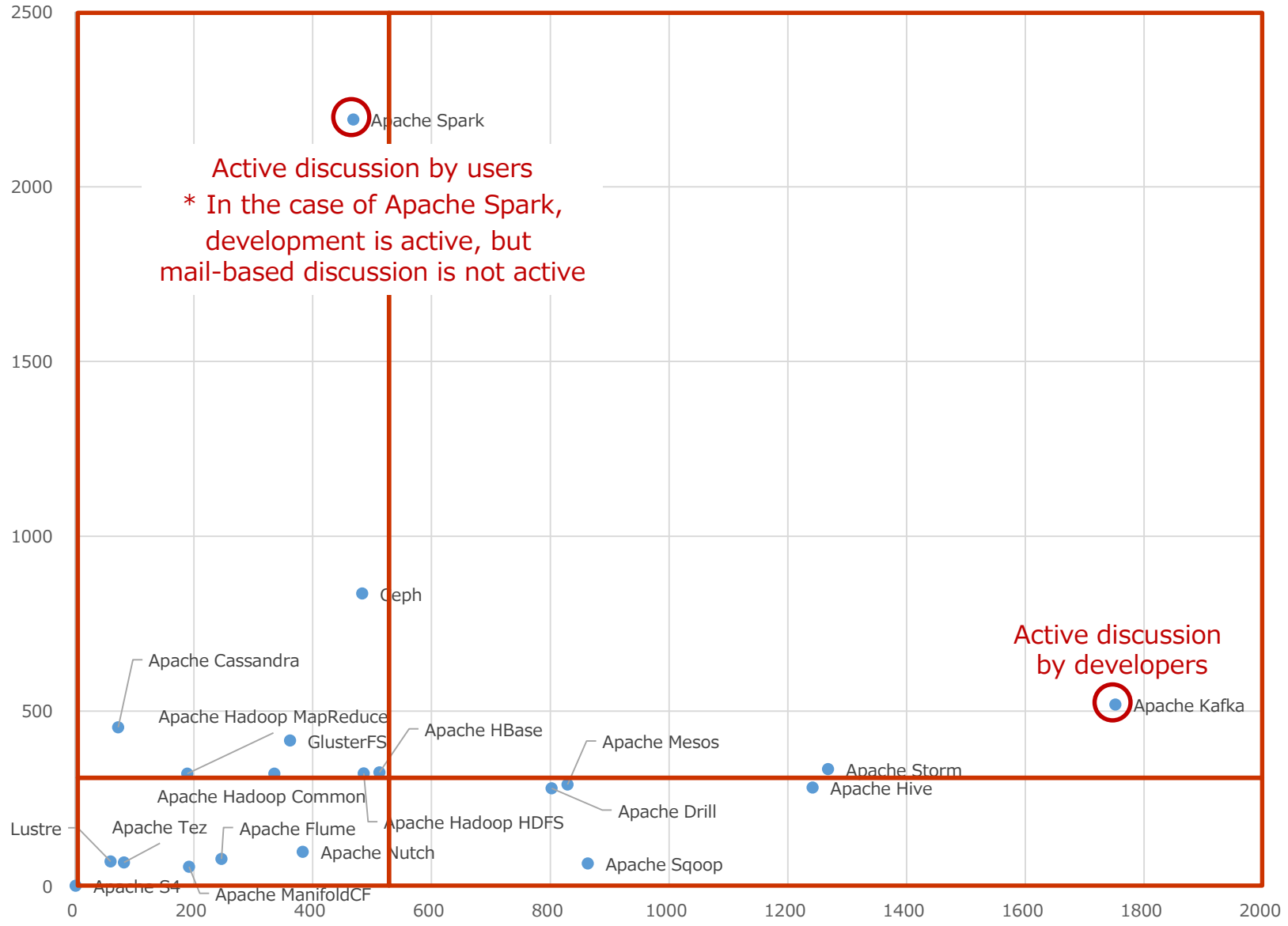


Average number of mails in ML for users (2015)

- Apache Spark has many active users and active developers

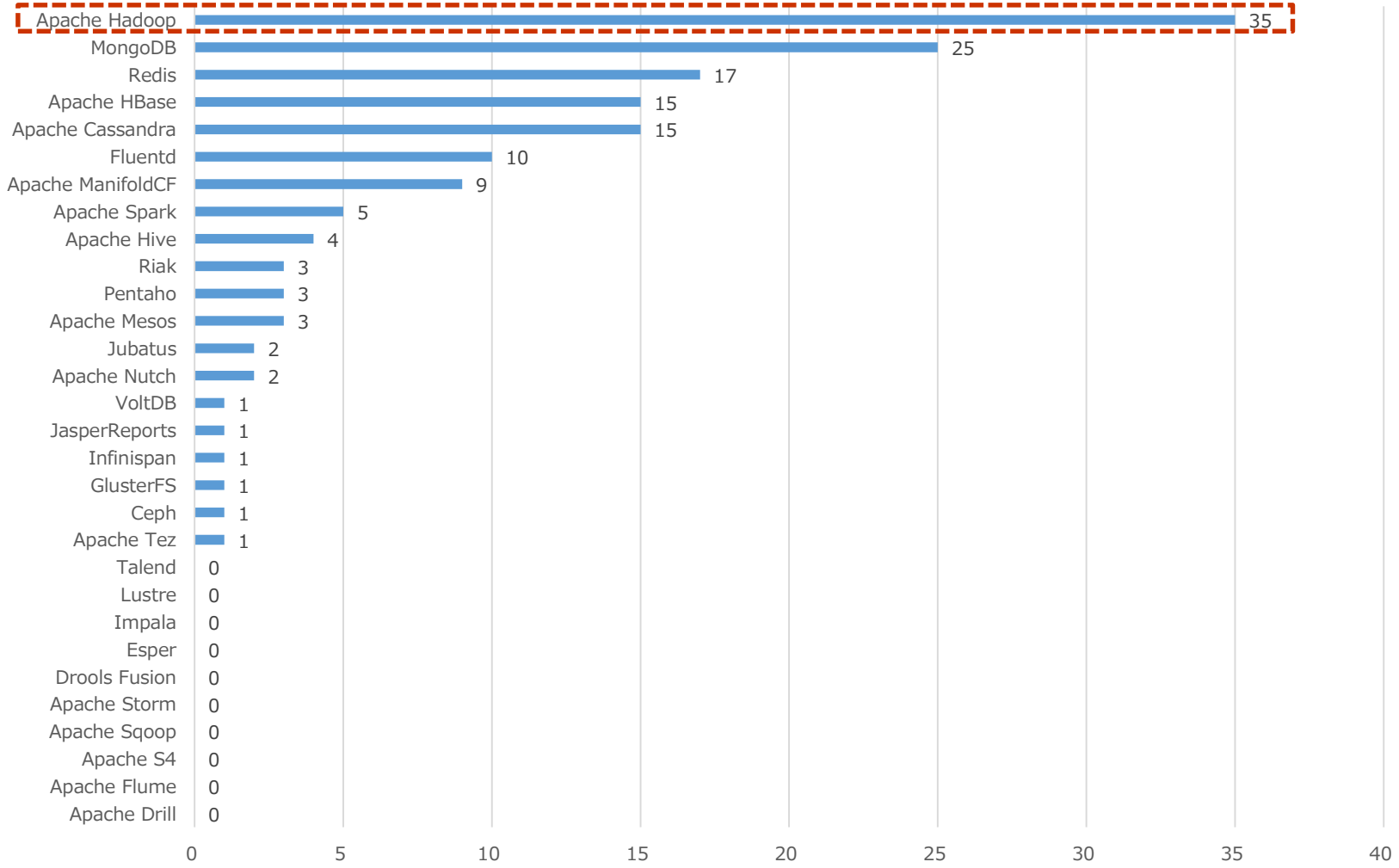


Mails by developers (x-axis) and users (y-axis)



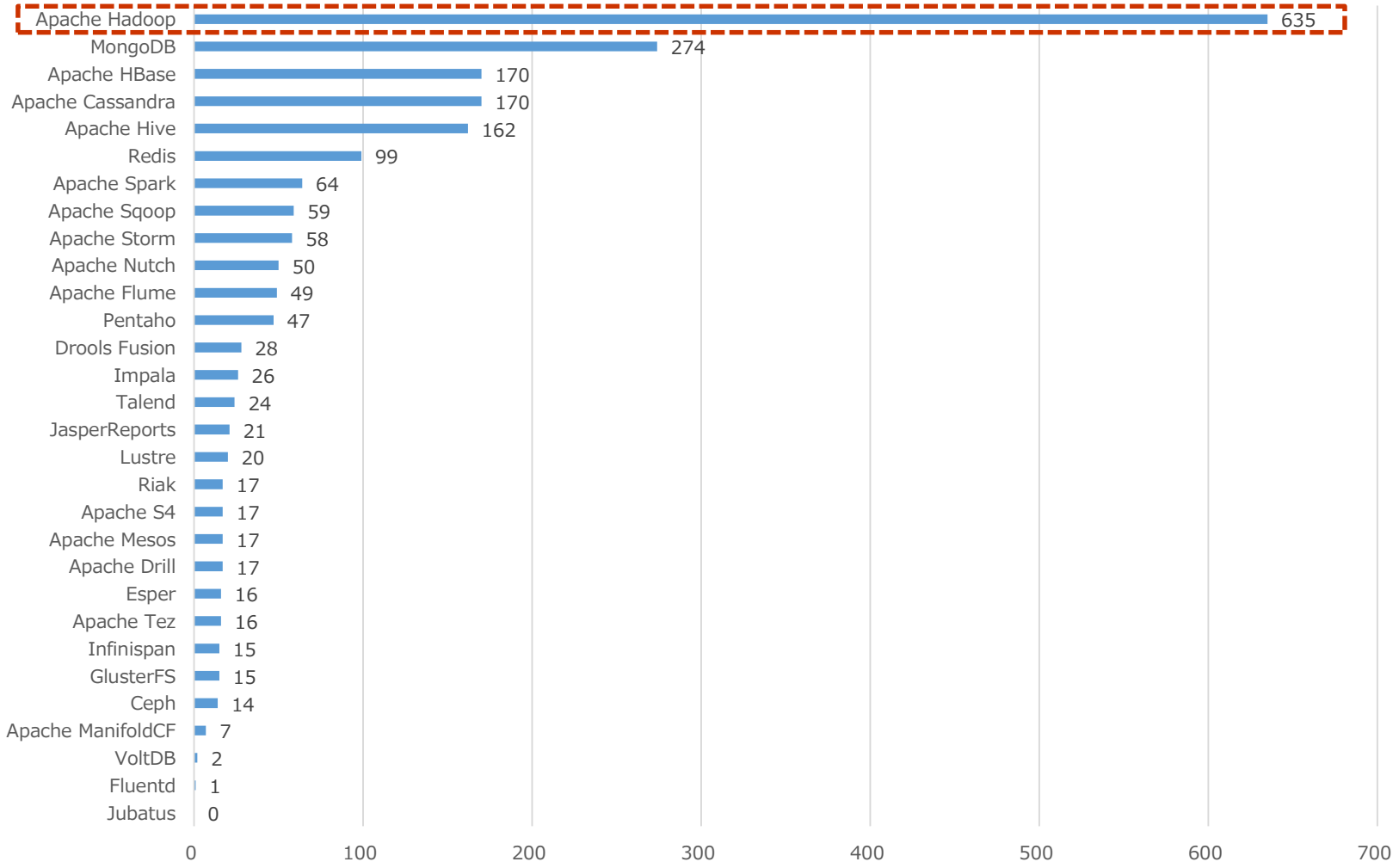
Books in Amazon (Japanese)

- There are many books about R (for statistic or programming, etc.) (450 Books, not in graph)
  - Then, major software like Apache Hadoop, MongoDB



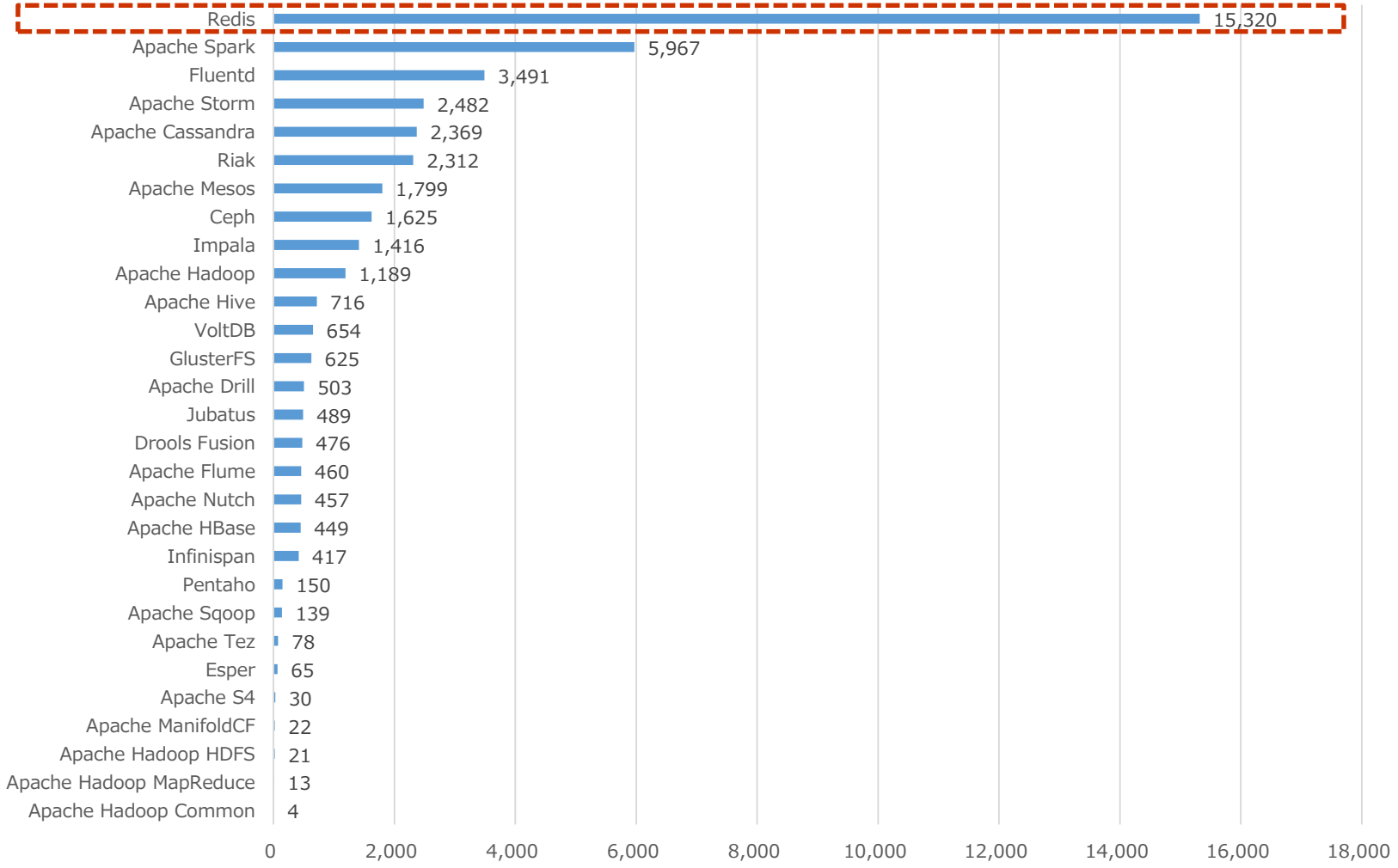
Books in Amazon (English)

- Almost same tendency as Japanese (R is not in graph)
  - Software produced by Japanese developers (e.g. fluentd, jubatus) get low rank in English books ranking



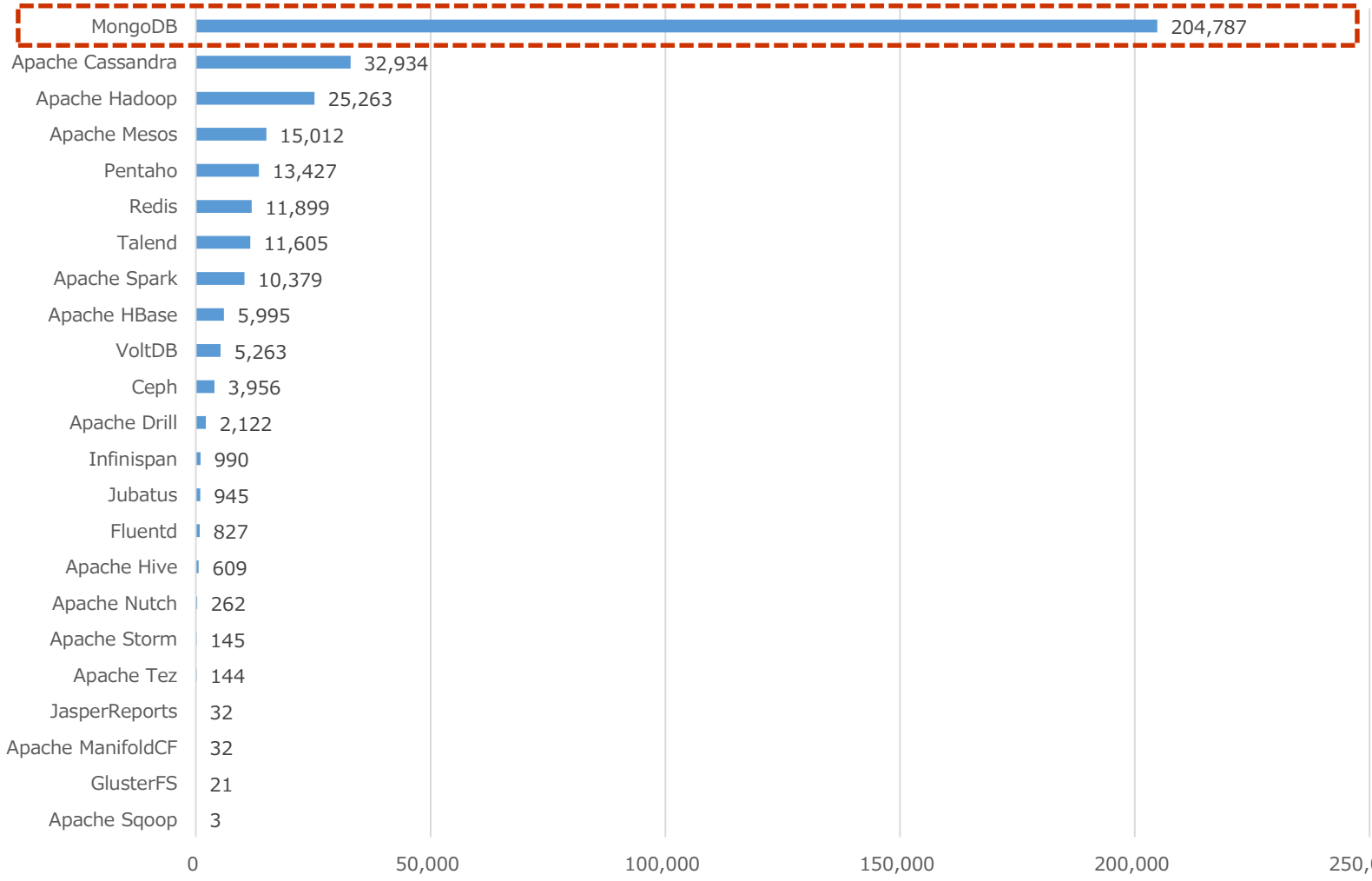
Stars in GitHub

- Redis gets a large amount of stars
  - The number of stars can be related to the number of enterprise users(?)



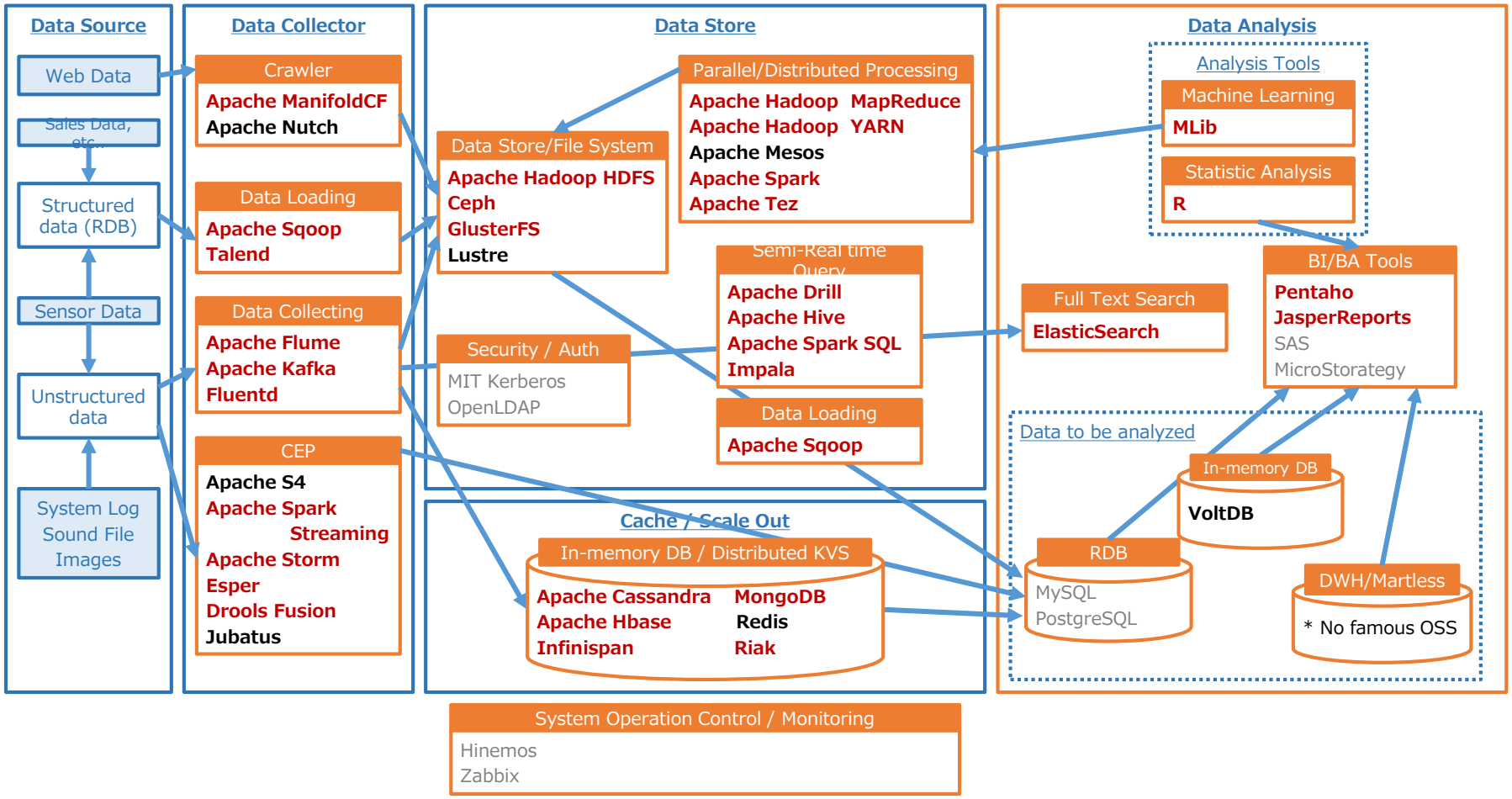
Followers on Twitter

- MongoDB gets the greatest followers
  - It might depend on how early the account start



Enterprise support

- We can receive Enterprise Support for software in red letters in Japan
  - We can build the Big Data Platform only with Enterprise supported OSS
  - Some software are provided as a service in Cloud (e.g. Jubatus)



Enterprise track record

- From The Linux Foundation SI Forum 2015 in Japan

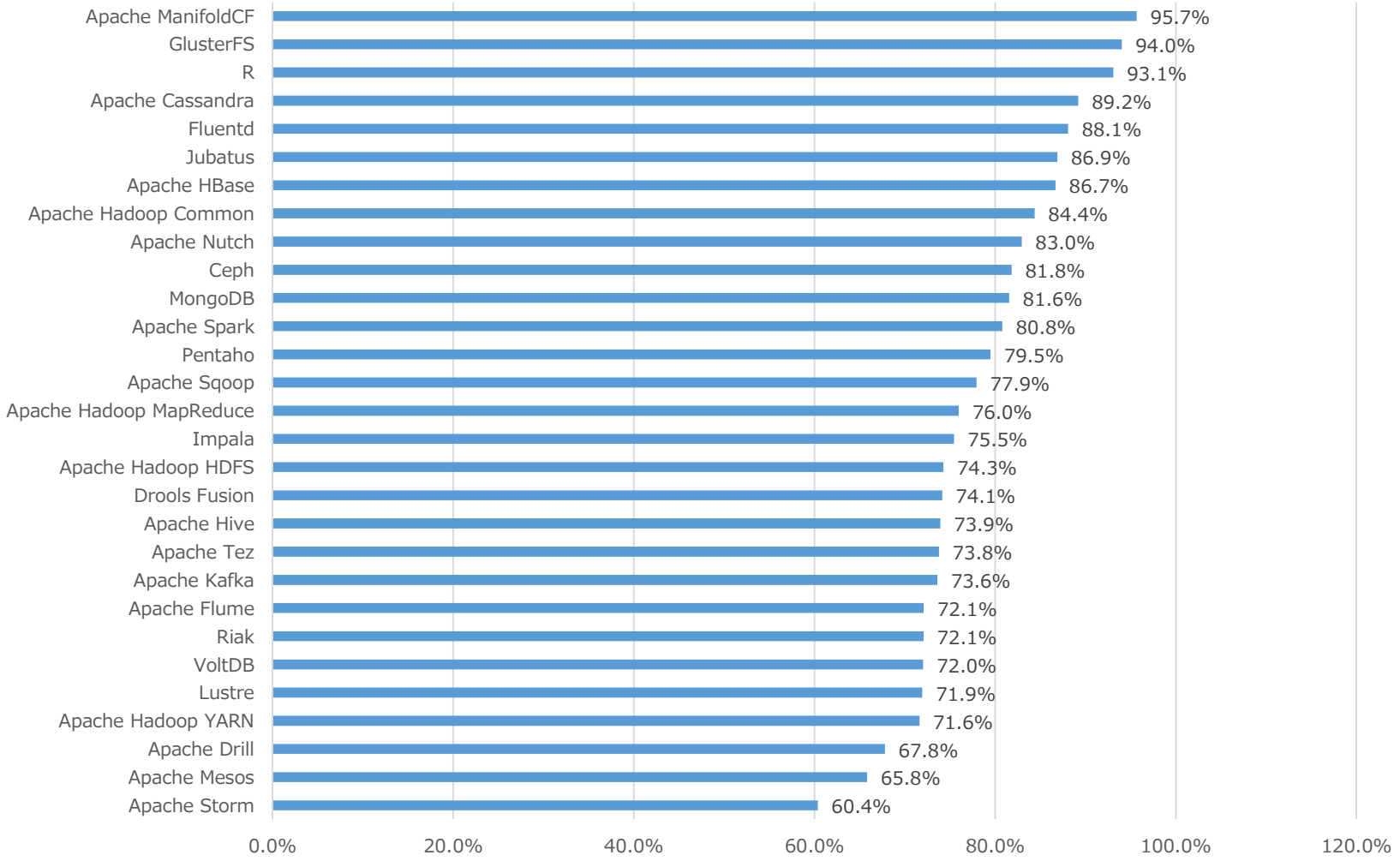
Many Introduction Track records	Apache Hadoop, GlusterFS, MongoDB, JasperReports
Some Introduction Track records	Talend, Fluentd, Jubatus, Apache Spark, Ceph, R, Lustre, Apache Cassandra, Apache Hbase, Redis, Elasticsearch, Pentaho
Some Verification Track records	VoltDB
No track records	The Others (that is a little disappointing result)

Quality of Software

Can we use the software without defects?

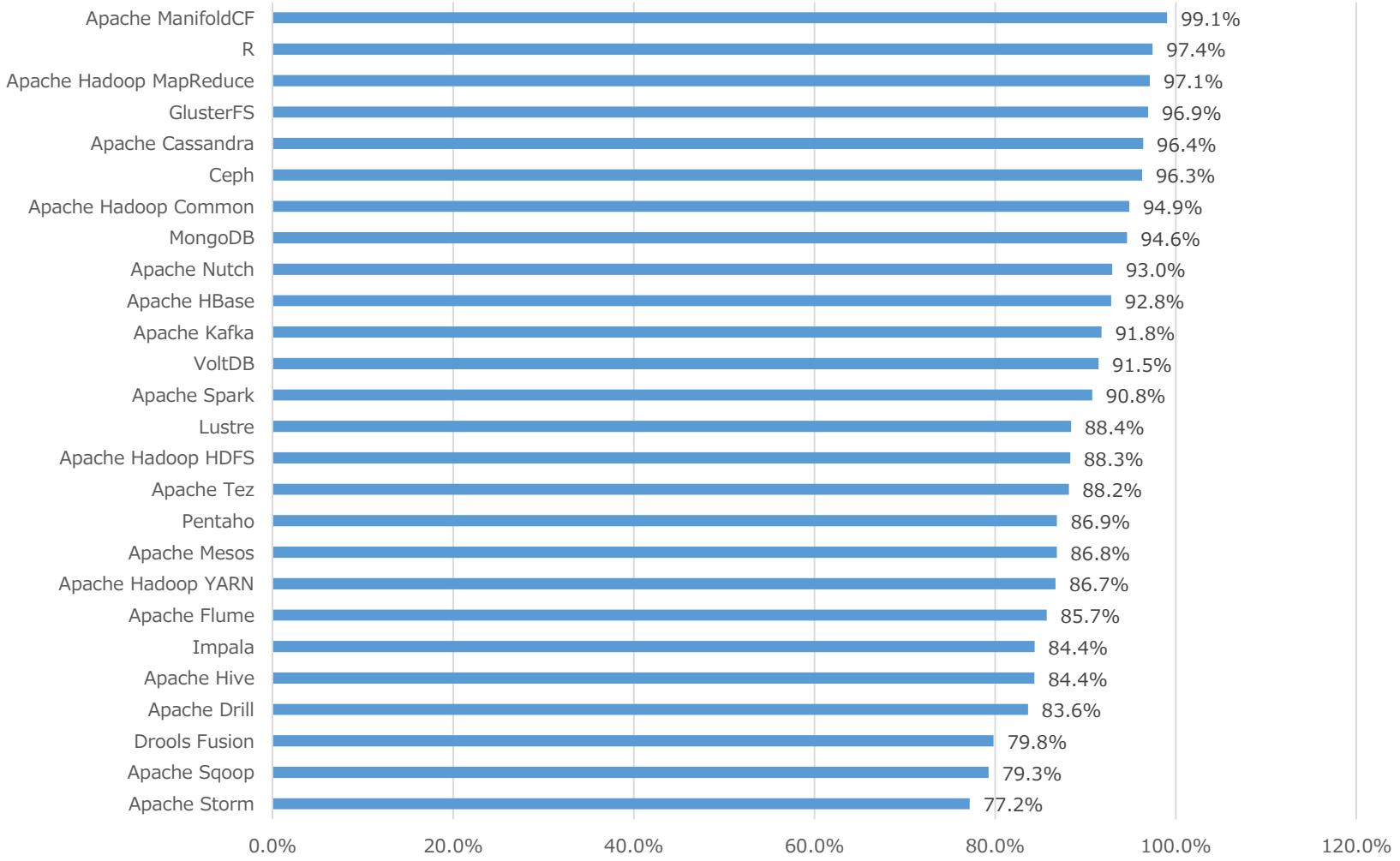
Bug resolution rate

- Apache ManifoldCF, GlusterFS and R perform high bug fix rate
  - Apache Storm performs the lowest bug fix rate, but the rate is 60%

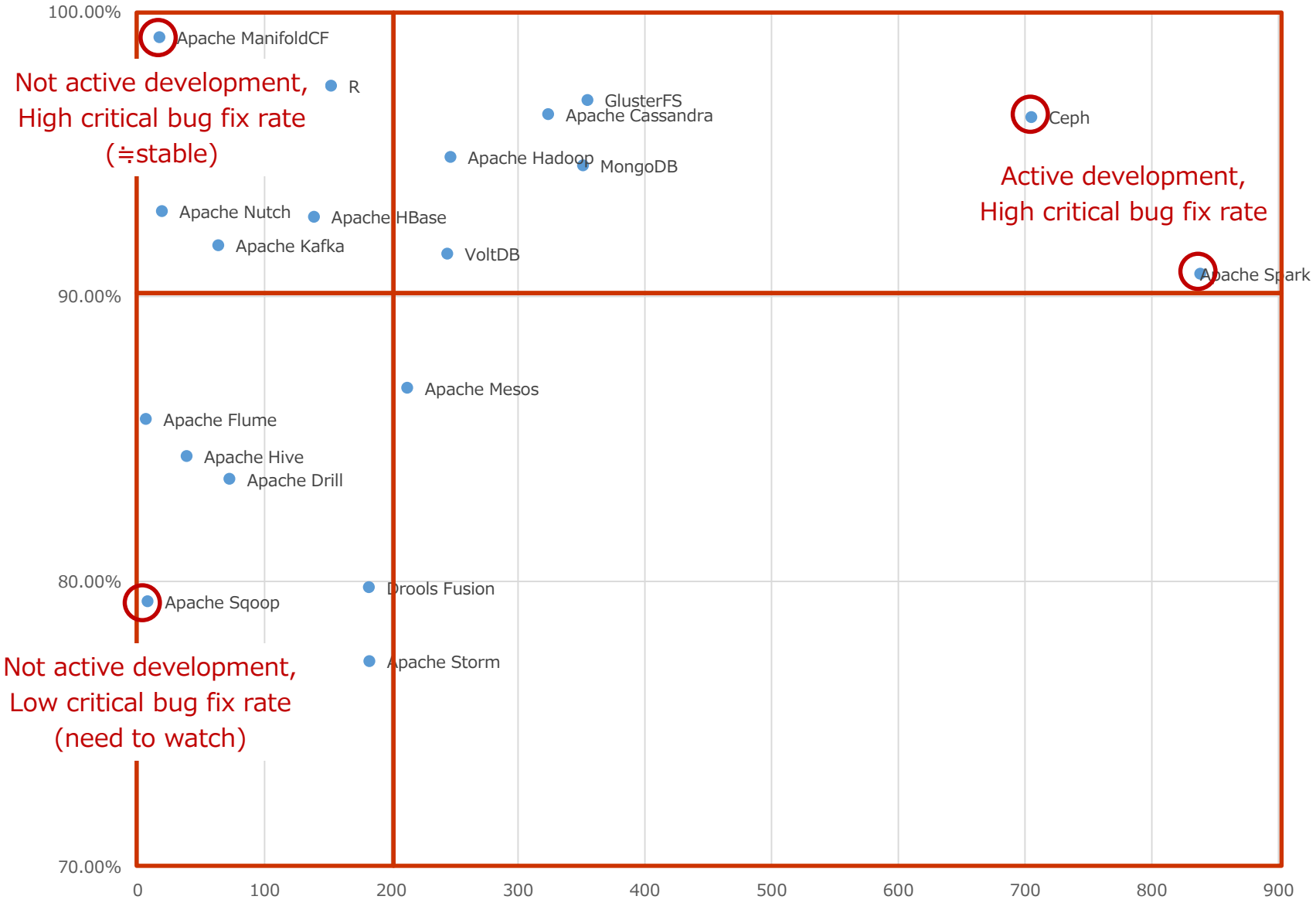


Blocker/Critical bug resolution rate

- About critical bugs, over 80% of them are resolved
  - Apache ManifoldCF also performs high fix rate

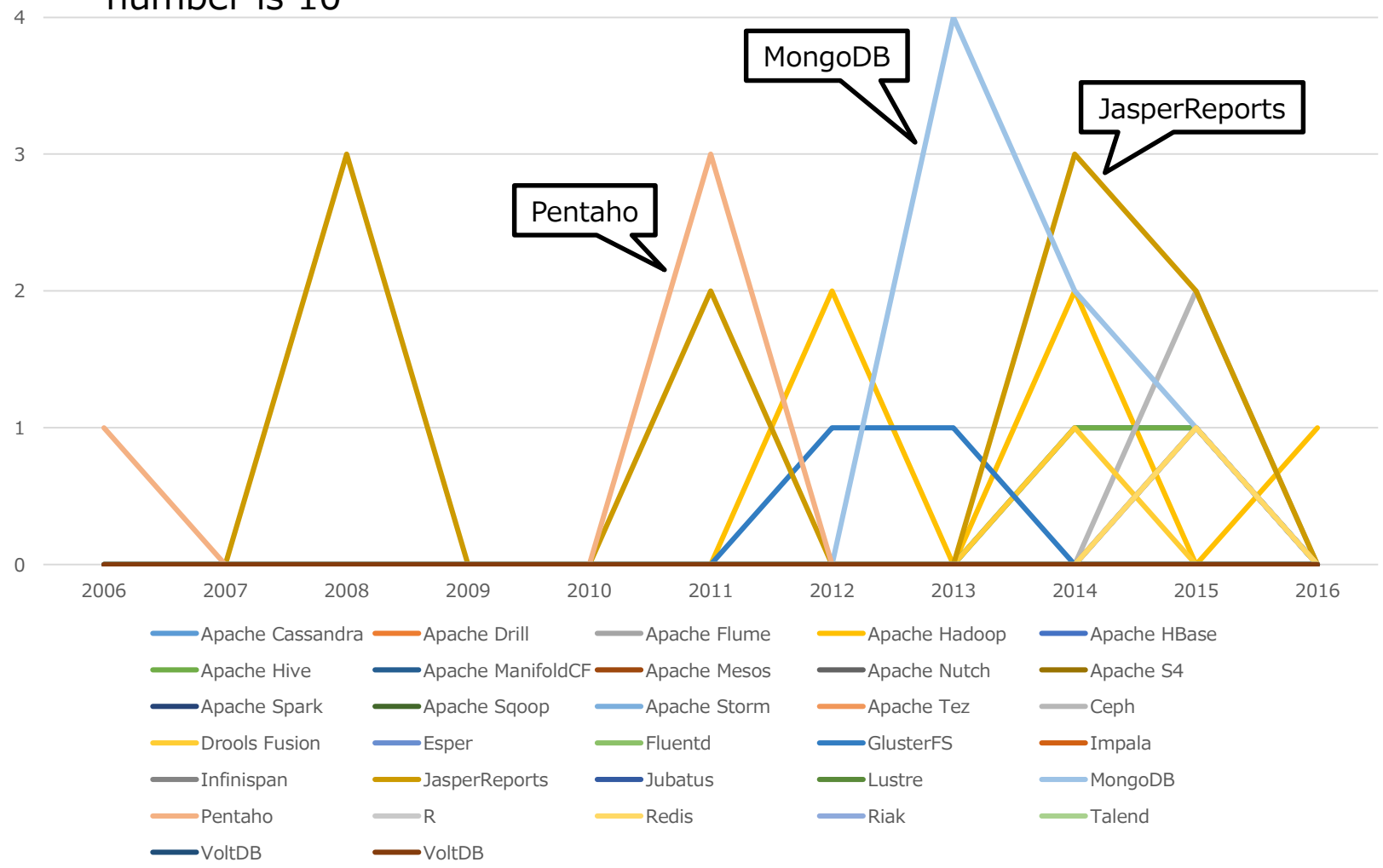


Commits (x-axis) and Critical bug fix rate (y-axis)



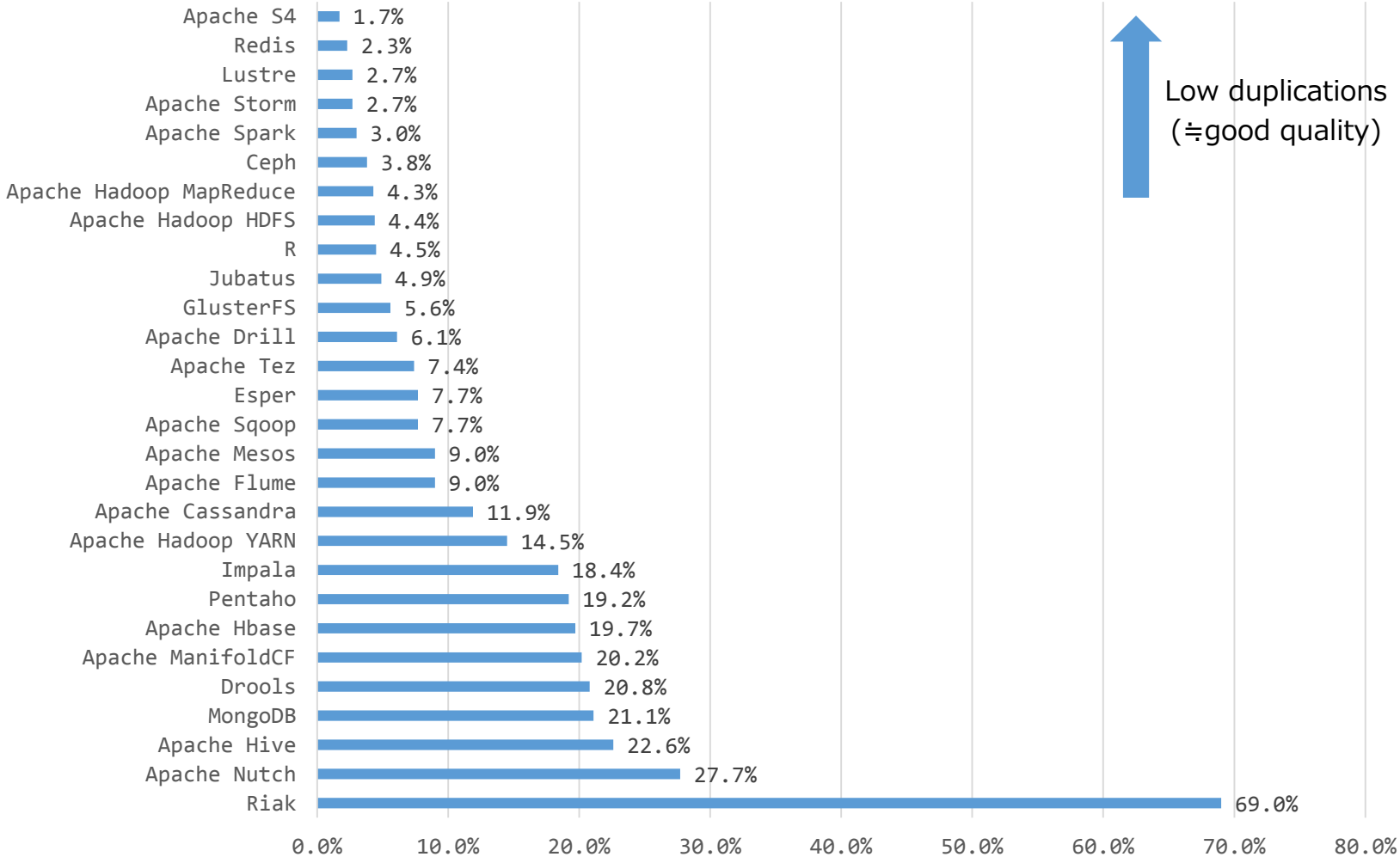
Vulnerabilities (2006 – 2016)

- Totally there are small number of vulnerabilities
  - JasperReports had the largest number of vulnerabilities, but the number is 10



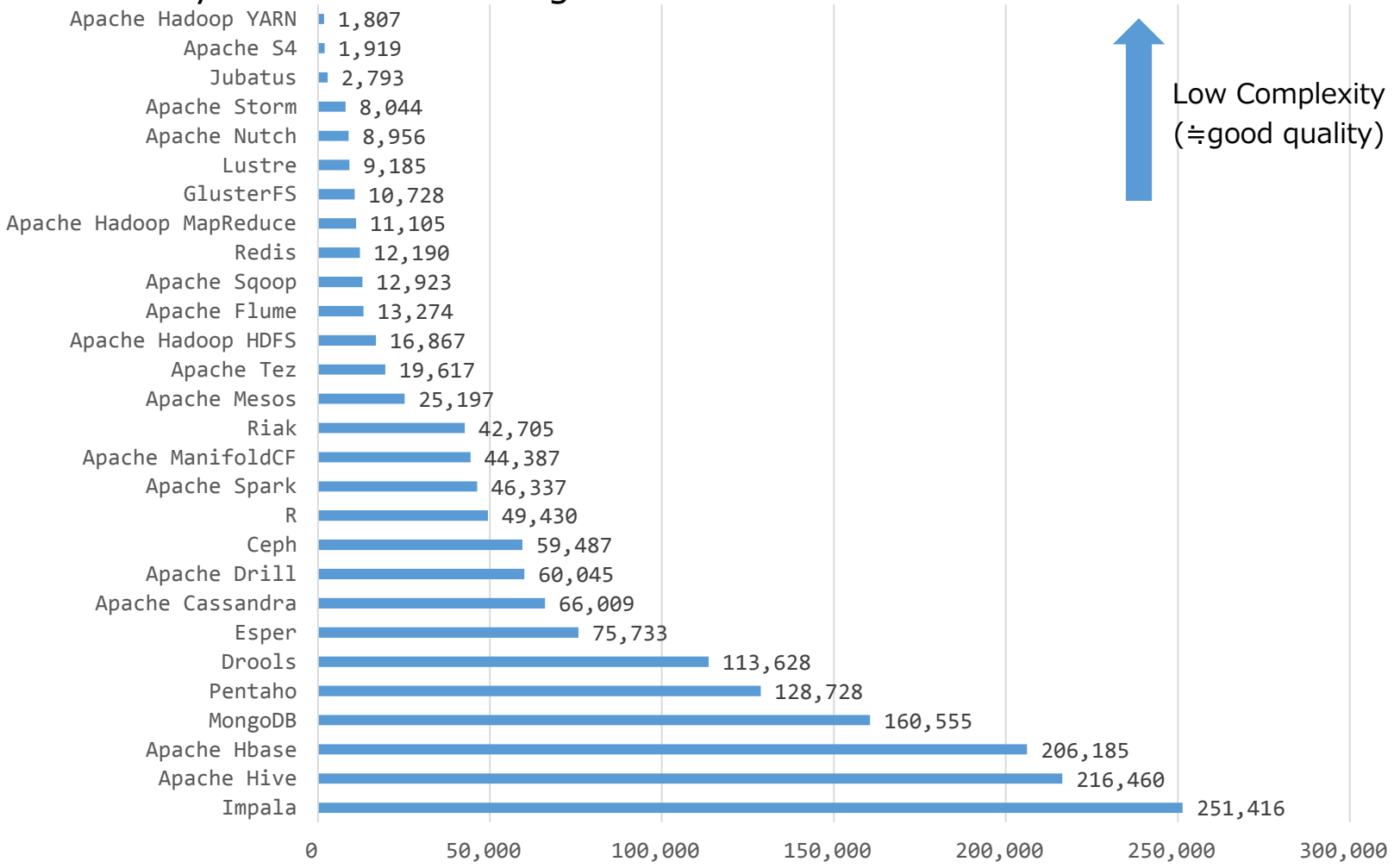
Duplications in source code

- Result of static source code analysis by SonarQube
  - Apache S4 and Redis have less duplicated source code (maybe good)



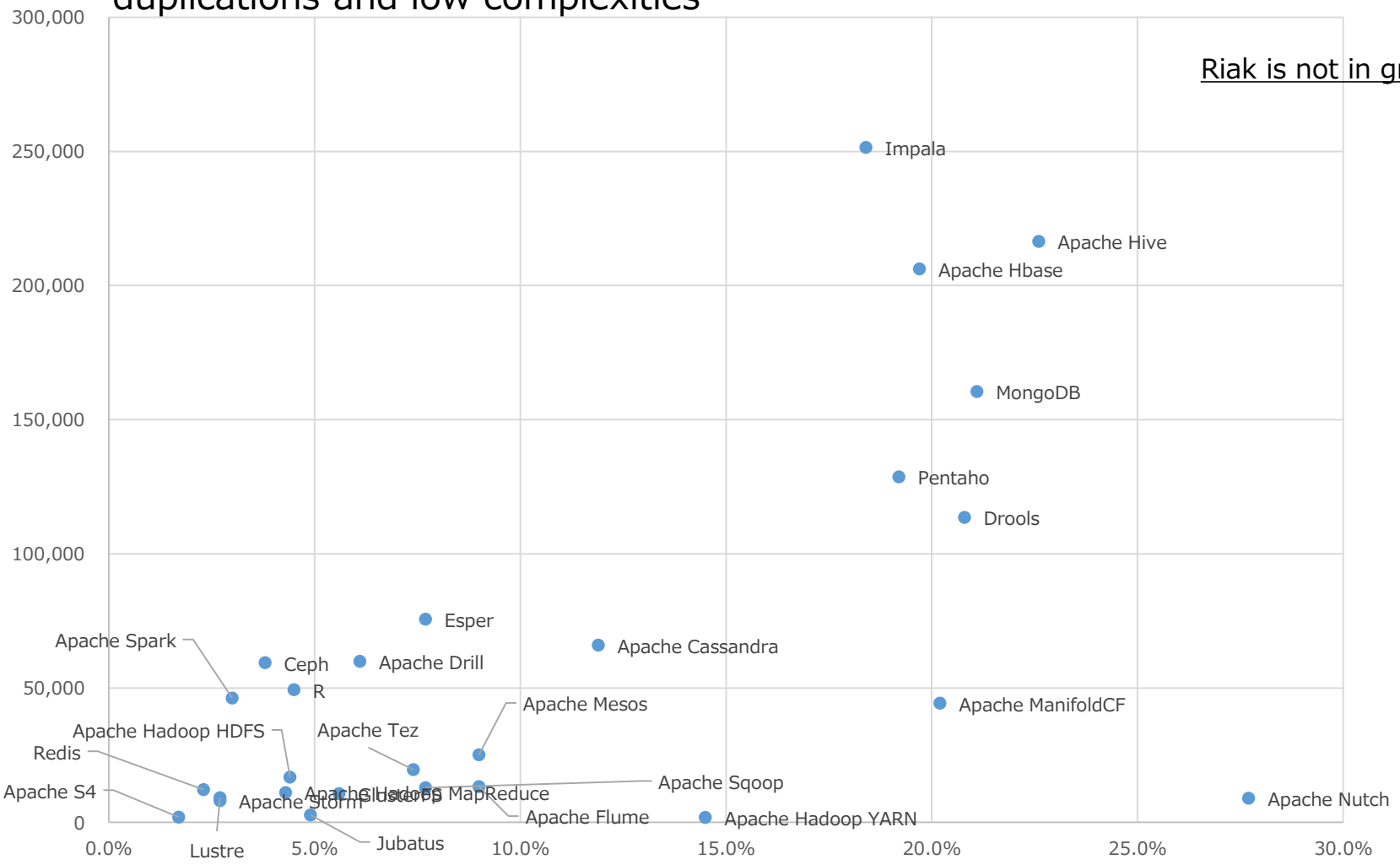
Complexities of source code

- Result of static source code analysis by SonarQube
  - Apache Hadoop YARN, Apache S4 and Jubatus may have relatively easily understandable logic



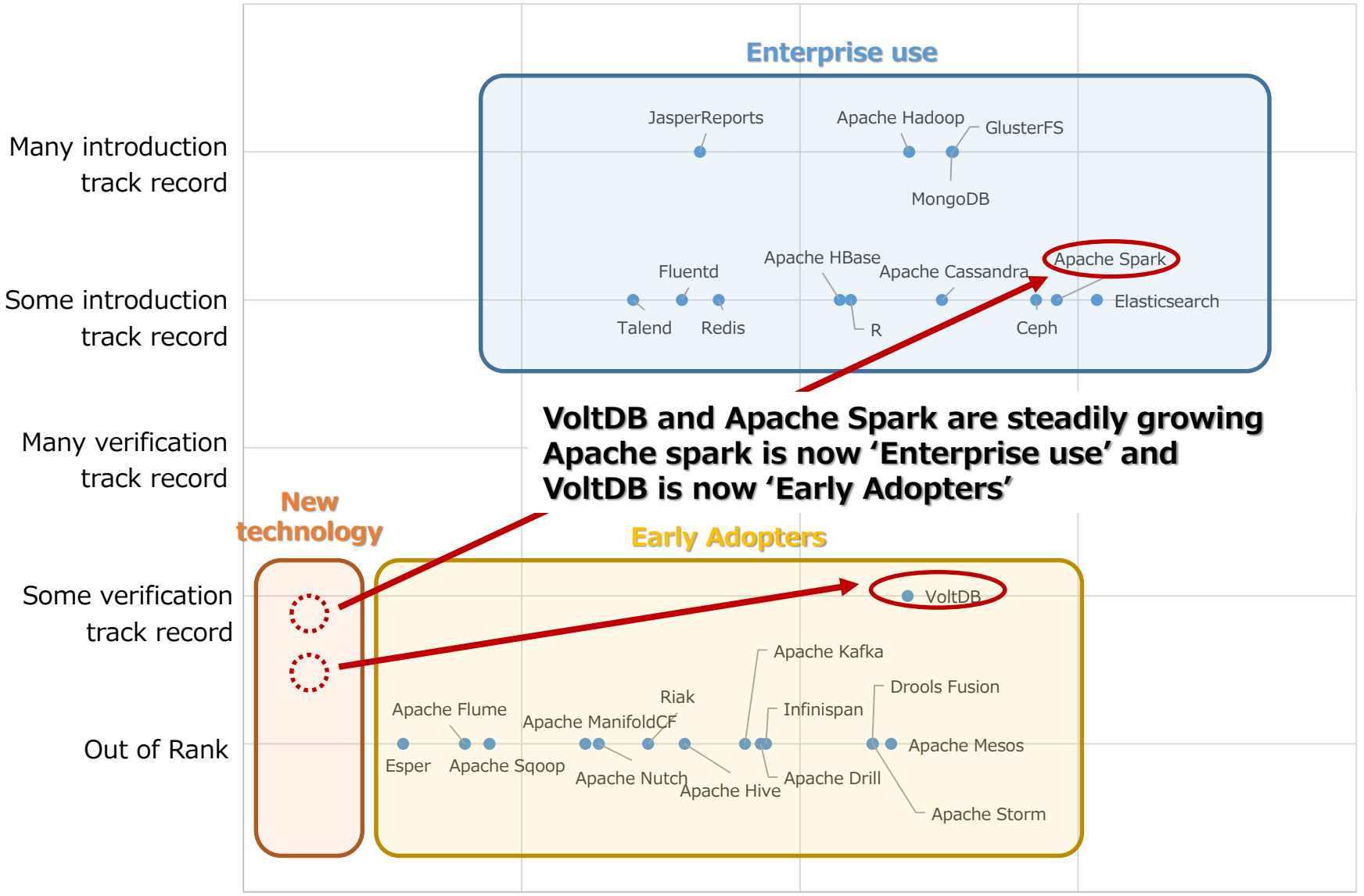
Duplications (x-axis) and Complexity (y-axis)

- There are weak correlation between duplications and complexities
  - However, there are software like Apache Nutch, which has many duplications and low complexities



Summary

Summary



Conclusion

- We can build the Big Data Platform only with OSS
  - Enterprise supports are getting better
  - However, it is necessary to check the functions and quality of softwares
- Apache Spark and the ecosystem are hot
- Developers for Elasticsearch may be hardworkers
- MongoDB and Ceph are going to be stable

- We should continue to watch the situation of OSS

Appendix (Data source and Enterprise Service in Japan)

- Data sources
- Enterprise Service in Japan

Official Sites

Function	Software	Official Site
Crawler	Apache ManifoldCF	http://manifoldcf.apache.org/
	Apache Nutch	http://nutch.apache.org/
Data Loading	Apache Sqoop	http://sqoop.apache.org/
	Talend	https://www.talend.com/
Data Collecting	Apache Flume	https://flume.apache.org/
	Apache Kafka	http://kafka.apache.org/
	Fluentd	http://www.fluentd.org/
CEP	Apache Storm	http://storm.apache.org/
	Apache S4	http://incubator.apache.org/s4/
	Jubatus	http://jubat.us/
	Esper	http://www.espertech.com/products/esper.php
	Drools Fusion	http://www.drools.org/
	Apache Spark Streaming	http://spark.apache.org/streaming/
Parallel / Distributed Processing	Apache Hadoop Commons	http://hadoop.apache.org/
	Apache Hadoop MapReduce	http://hadoop.apache.org/
	Apache Hadoop YARN	http://hadoop.apache.org/
	Apache Mesos	http://mesos.apache.org/
	Apache Spark	http://spark.apache.org/
	Apache Tez	https://tez.apache.org/
Data Store / File System	Apache Hadoop HDFS	http://hadoop.apache.org/
	Ceph	http://ceph.com/
	GlusterFS	http://www.gluster.org/
	Lustre	http://lustre.org/
Query Engine	Apache Drill	https://drill.apache.org/
	Apache Hive	https://hive.apache.org/
	Apache Spark SQL	http://spark.apache.org/sql/
	Impala	http://impala.io/
In-memory DB / Distributed KVS	Apache Cassandra	http://cassandra.apache.org/
	Apache HBase	http://hbase.apache.org/
	Infinispan	http://infinispan.org/
	MongoDB	https://www.mongodb.org/
	Redis	http://redis.io/
	Riak	http://docs.basho.com/
Full Text Search	Elasticsearch	https://www.elastic.co/products/elasticsearch
Machine Learning	Apache Spark MLib	http://spark.apache.org/mllib/
Statistical Analysis	R	https://www.r-project.org/
BI/BA Tools	JasperReports	http://community.jaspersoft.com/
	Pentaho	http://community.pentaho.com/
In-memory DB	VoltDB	https://voltdb.com/

Licenses

function	software	License
Crawler	Apache ManifoldCF	Apache License 2.0
	Apache Nutch	Apache License 2.0
Data Loading	Apache Sqoop	Apache License 2.0
	Talend	Apache License 2.0
Data Collecting	Apache Flume	Apache License 2.0
	Apache Kafka	Apache License 2.0
	Fluentd	Apache License 2.0
CEP	Apache Storm	Apache License 2.0
	Apache S4	Apache License 2.0
	Jubatus	GNU Lesser General Public License v2.1
	Esper	GNU General Public License v2
	Drools Fusion	Apache License 2.0
	Apache Spark Streaming	Apache License 2.0
Parallel / Distributed Processing	Apache Hadoop Common	Apache License 2.0
	Apache Hadoop MapReduce	Apache License 2.0
	Apache Hadoop YARN	Apache License 2.0
	Apache Mesos	Apache License 2.0
	Apache Spark	Apache License 2.0
	Apache Tez	Apache License 2.0
Data Store / File System	Apache Hadoop HDFS	Apache License 2.0
	Ceph	GNU Lesser General Public License v2.1
	GlusterFS	GNU General Public License v3
	Lustre	GNU General Public License v2
Query Engine	Apache Drill	Apache License 2.0
	Apache Hive	Apache License 2.0
	Apache Spark SQL	Apache License 2.0
	Impala	Apache License 2.0
In-memory DB / Distributed KVS	Apache Cassandra	Apache License 2.0
	Apache HBase	Apache License 2.0
	Infinispan	Apache License 2.0
	MongoDB	GNU Affero General Public License v3
	Redis	BSD License
	Riak	Apache License 2.0
Full Text Search	Elasticsearch	Apache License 2.0
Machine Learning	Apache Spark MLib	Apache License 2.0
Statistical Analysis	R	GNU General Public License
BI/BA Tools	JasperReports	GNU Lesser General Public License
	Pentaho	Apache License 2.0
In-memory DB	VoltDB	GNU General Public License v3

Mailing List for Developers

functions	software	Mailing list for developers
Crawler	Apache ManifoldCF	dev@manifoldcf.apache.org
	Apache Nutch	dev@nutch.apache.org
Data Loading	Apache Sqoop	dev@sqoop.apache.org
	Talend	-
Data collecting	Apache Flume	dev@flume.apache.org
	Apache Kafka	dev@kafka.apache.org
	Fluentd	*Google groups
CEP	Apache Storm	dev@storm.apache.org
	Apache S4	s4-dev@incubator.apache.org
	Jubatus	*Google groups
	Esper	dev@esper.codehaus.org
	Drools Fusion	*Google groups
	Apache Spark Streaming	*Same as the ML of Apache Spark
Parallel / Distributed Processing	Apache Hadoop Common	common-dev@hadoop.apache.org
	Apache Hadoop MapReduce	mapreduce-dev@hadoop.apache.org
	Apache Hadoop YARN	yarn-dev@hadoop.apache.org
	Apache Mesos	dev@mesos.apache.org
	Apache Spark	dev@spark.apache.org
	Apache Tez	dev@tez.apache.org
Data Store / File System	Apache Hadoop HDFS	hdfs-dev@hadoop.apache.org
	Ceph	ceph-devel@vger.kernel.org
	GlusterFS	gluster-devel@gluster.org
	Lustre	lustre-devel@lists.lustre.org
Query Engine	Apache Drill	dev@drill.apache.org
	Apache Hive	dev@hive.apache.org
	Apache Spark SQL	*Same as the ML of Apache Spark
	Impala	*Google groups
In-memory DB / Distributed KVS	Apache Cassandra	dev@cassandra.apache.org
	Apache HBase	dev@hbase.apache.org
	Infinispan	infinispan-dev@lists.jboss.org
	MongoDB	*Google groups
	Redis	*Google groups
	Riak	*For users only
Full Text Search	Elasticsearch	*Google groups
Machine Learning	Apache Spark MLib	*Same as the ML of Apache Spark
Statistical Analysis	R	r-devel@r-project.org
BI/BA Tools	JasperReports	-
	Pentaho	-
In-memory DB	VoltDB	-

Mailing List for Users

function	software	Mailing list for users
Crawler	Apache ManifoldCF	user@manifoldcf.apache.org
	Apache Nutch	user@nutch.apache.org
Data Loading	Apache Sqoop	user@sqoop.apache.org
	Talend	-
Data Collecting	Apache Flume	user@flume.apache.org
	Apache Kafka	users@kafka.apache.org
	Fluentd	*Google groups
CEP	Apache Storm	user@storm.apache.org
	Apache S4	s4-user@incubator.apache.org
	Jubatus	*Google groups
	Esper	user@esper.codehaus.org
	Drools Fusion	*Google groups
	Apache Spark Streaming	*Same as the ML of Apache Spark
Parallel / Distributed Processing	Apache Hadoop Common	user@hadoop.apache.org
	Apache Hadoop MapReduce	*Same as the ML of Apache Hadoop Common
	Apache Hadoop YARN	*Same as the ML of Apache Hadoop Common
	Apache Mesos	user@mesos.apache.org
	Apache Spark	user@spark.apache.org
	Apache Tez	user@tez.apache.org
Data Store / File System	Apache Hadoop HDFS	*Same as the ML of Apache Hadoop Common
	Ceph	ceph-user@lists.ceph.com
	GlusterFS	gluster-users@gluster.org
	Lustre	lustre-discuss@lists.lustre.org
Query Engine	Apache Drill	user@drill.apache.org
	Apache Hive	user@hive.apache.org
	Apache Spark SQL	*Same as the ML of Apache Spark
	Impala	*Google groups
In-memory DB / Distributed KVS	Apache Cassandra	user@cassandra.apache.org
	Apache HBase	user@hbase.apache.org
	Infinispan	*For developers only
	MongoDB	*Google groups
	Redis	*Google groups
	Riak	riak-users@lists.basho.com
Full Text Search	Elasticsearch	*Google groups
Machine Learning	Apache Spark MLib	*Same as the ML of Apache Spark
Statistical Analysis	R	(r-help@r-project.org?)
BI/BA Tools	JasperReports	-
	Pentaho	*Google groups
In-memory DB	VoltDB	-

Source Code Repository Service

function	software	Source code repository service
Crawler	Apache ManifoldCF	GitHub(apache/manifoldcf)
	Apache Nutch	GitHub(apache/nutch)
Data Loading	Apache Sqoop	GitHub(apache/sqoop)
	Apache Kafka	GitHub(apache/kafka)
	Talend	-
Data Collecting	Apache Flume	GitHub(apache/flume)
	Fluentd	GitHub(fluent/fluentd)
CEP	Apache Storm	GitHub(apache/storm)
	Apache S4	GitHub(apache/incubator-s4)
	Jubatus	GitHub(jubatus/jubatus)
	Esper	GitHub(espertechnic/esper)
	Drools Fusion	GitHub(droolsjbpm/drools)
	Apache Spark Streaming	*Same as the repository of Apache Spark
Parallel / Distributed Processing	Apache Hadoop Common	GitHub(apache/hadoop-common)
	Apache Hadoop MapReduce	GitHub(apache/hadoop-mapreduce)
	Apache Hadoop YARN	-
	Apache Mesos	GitHub(apache/mesos)
	Apache Spark	GitHub(apache/spark)
	Apache Tez	GitHub(apache/tez)
Data Store / File System	Apache Hadoop HDFS	GitHub(apache/hadoop-hdfs)
	Ceph	GitHub(ceph/ceph)
	GlusterFS	GitHub(gluster/glusterfs)
	Lustre	-
Query Engine	Apache Drill	GitHub(apache/drill)
	Apache Hive	GitHub(apache/hive)
	Apache Spark SQL	*Same as the repository of Apache Spark
	Impala	GitHub(cloudera/impala)
In-memory DB / Distributed KVS	Apache Cassandra	GitHub(apache/cassandra)
	Apache HBase	GitHub(apache/hbase)
	Infinispan	GitHub(infinispan/infinispan)
	MongoDB	-
	Redis	GitHub(antirez/redis)
	Riak	GitHub(basho/riak)
Full Text Search	Elasticsearch	GitHub(elastic/elasticsearch)
Machine Learning	Apache Spark MLib	*Same as the repository of Apache Spark
Statistical Analysis	R	-
BI/BA Tools	JasperReports	-
	Pentaho	GitHub(pentaho/pentaho-platform)
In-memory DB	VoltDB	GitHub(VoltDB/voltdb)

Twitter

funciton	software	Twitter
Crawler	Apache ManifoldCF	@ApacheManifold
	Apache Nutch	@ApacheNutch
Data Loading	Apache Sqoop	@sqoopit
	Talend	@Talend
Data Collecting	Apache Flume	-
	Apache Kafka	@apachekafka
	Fluentd	@fluentd
CEP	Apache Storm	@ApacheStorm
	Apache S4	-
	Jubatus	@JubatusOfficial
	Esper	-
	Drools Fusion	-
	Apache Spark Streaming	*Same as the account of Apache Spark
Parallel / Distributed Processing	Apache Hadoop Common	@hadoop
	Apache Hadoop MapReduce	*Same as the account of Apache Hadoop
	Apache Hadoop YARN	*Same as the account of Apache Hadoop
	Apache Mesos	@Apache Mesos
	Apache Spark	@ApacheSpark
	Apache Tez	@ApacheTez
Data Store File System	Apache Hadoop HDFS	*Same as the account of Apache Hadoop
	Ceph	@Ceph
	GlusterFS	@glusterfs
	Lustre	-
Query Engine	Apache Drill	@ApacheDrill
	Apache Hive	@ApacheHive
	Apache Spark SQL	*Same as the account of Apache Spark
	Impala	-
In-memory DB / Distributed KVS	Apache Cassandra	@Cassandra
	Apache HBase	@Hbase
	Infinispan	@infinispan
	MongoDB	@MongoDB
	Redis	@redisfeed
	Riak	-
Full Text Search	Elasticsearch	@Elasticsearch
Machine Learning	Apache Spark MLib	*Same as the account of Apache Spark
Statistical Analysis	R	-
BI / BA Tools	JasperReports	@jasperreports
	Pentaho	@Pentaho
In-memory DB	VoltDB	@VoltDB

Issue Tracker

function	software	Issue Tracker
Crawler	Apache ManifoldCF	JIRA(https://issues.apache.org/jira/browse/CONNECTORS)
	Apache Nutch	JIRA(https://issues.apache.org/jira/browse/NUTCH)
Data Loading	Apache Sqoop	JIRA(https://issues.apache.org/jira/browse/SQOOP)
	Talend	-
Data Collecting	Apache Flume	JIRA(https://issues.apache.org/jira/browse/FLUME)
	Apache Kafka	JIRA(https://issues.apache.org/jira/browse/KAFKA)
	Fluentd	GitHub(https://github.com/fluent/fluentd/issues)
CEP	Apache Storm	JIRA(https://issues.apache.org/jira/browse/STORM)
	Apache S4	-
	Jubatus	GitHub(https://github.com/jubatus/jubatus/issues)
	Esper	-
	Drools Fusion	JIRA(https://issues.jboss.org/projects/DROOLS)
	Apache Spark Streaming	*Same as the account of Apache Spark
Parallel / Distributed Processing	Apache Hadoop Common	JIRA(https://issues.apache.org/jira/browse/HADOOP)
	Apache Hadoop MapReduce	JIRA(https://issues.apache.org/jira/browse/MAPREDUCE)
	Apache Hadoop YARN	JIRA(https://issues.apache.org/jira/browse/YARN)
	Apache Mesos	JIRA(https://issues.apache.org/jira/browse/MESOS)
	Apache Spark	JIRA(https://issues.apache.org/jira/browse/spark)
	Apache Tez	JIRA(https://issues.apache.org/jira/browse/TEZ)
Data Store File System	Apache Hadoop HDFS	JIRA(https://issues.apache.org/jira/browse/HDFS)
	Ceph	Redmine(http://tracker.ceph.com/projects/ceph)
	GlusterFS	Bugzilla(https://bugzilla.redhat.com/)
	Lustre	JIRA(https://jira.hpdd.intel.com/secure/Dashboard.jspa)
Query Engine	Apache Drill	JIRA(https://issues.apache.org/jira/browse/DRILL)
	Apache Hive	JIRA(https://issues.apache.org/jira/browse/HIVE)
	Apache Spark SQL	*Same as the account of Apache Spark
	Impala	JIRA(https://issues.cloudera.org/secure/Dashboard.jspa)
In-memory DB / Distributed KVS	Apache Cassandra	JIRA(https://issues.apache.org/jira/browse/CASSANDRA)
	Apache HBase	JIRA(https://issues.apache.org/jira/browse/HBASE)
	Infinispan	JIRA(https://issues.jboss.org/secure/Dashboard.jspa)
	MongoDB	JIRA(https://jira.mongodb.org/secure/Dashboard.jspa)
	Redis	Google Project Hosting(https://code.google.com/p/redis/issues/list)
	Riak	GitHub(https://github.com/basho/riak/issues)
Full Text Search	Elasticsearch	-
Machine Learning	Apache Spark MLib	*Same as the account of Apache Spark
Statistical Analysis	R	Bugzilla(https://bugs.r-project.org/bugzilla3/)
BI / BA Tools	JasperReports	-
	Pentaho	JIRA(http://jira.pentaho.com/secure/Dashboard.jspa)
In-memory DB	VoltDB	JIRA(https://issues.voltdb.com/browse/VDM)

Enterprise Services in Japan

function	software	Enterprise License	Maintenance Support Service	Training Service	Service providing on Cloud
Crawler	Apache ManifoldCF	–	○	○	–
	Apache Nutch	–	–	–	–
Data Loading	Apache Sqoop	–	○	△	△
	Talend	○	–	○	○
Data Collecting	Apache Flume	–	○	△	△
	Apache Kafka	–	○	–	○
	Fluentd	–	○	–	–
CEP	Apache Storm	–	○	–	△
	Apache S4	–	–	–	–
	Jubatus	–	–	–	○
	Esper	○	–	–	–
	Drools Fusion	○	–	–	–
	Apache Spark Streaming	–	○	△	△
Parallel / Distributed Processing	Apache Hadoop Common	○	○	○	○
	Apache Hadoop MapReduce	○	○	△	○
	Apache Hadoop YARN	○	○	△	△
	Apache Mesos	–	–	–	○
	Apache Spark	–	○	○	○
	Apache Tez	–	○	–	–
Data Store File System	Apache Hadoop HDFS	○	○	△	△
	Ceph	○	○	△	–
	GlusterFS	○	○	○	–
	Lustre	–	–	–	○
Query Engine	Apache Drill	○	○	–	△
	Apache Hive	–	○	○	△
	Apache Spark SQL	–	○	△	△
	Impala	–	○	○	△
In-memory DB / Distributed KVS	Apache Cassandra	○	○	○	○
	Apache HBase	–	○	○	○
	Infinispan	○	–	–	–
	MongoDB	○	○	○	○
	Redis	–	–	○	○
	Riak	○	–	–	○
Full Text Search	Elasticsearch	–	○	○	○
Machine Learning	Apache Spark MLib	–	○	–	△
Statistical Analysis	R	○	○	○	○
BI / BA Tools	JasperReports	○	○	○	○
	Pentaho	○	○	○	○
In-memory DB	VoltDB	–	–	–	–