

**나는 어쩌다 오픈소스 프로젝트
멤버가 되었나?**

권혁진 Databricks

누구세요?



Apache Spark Committer & PMC member



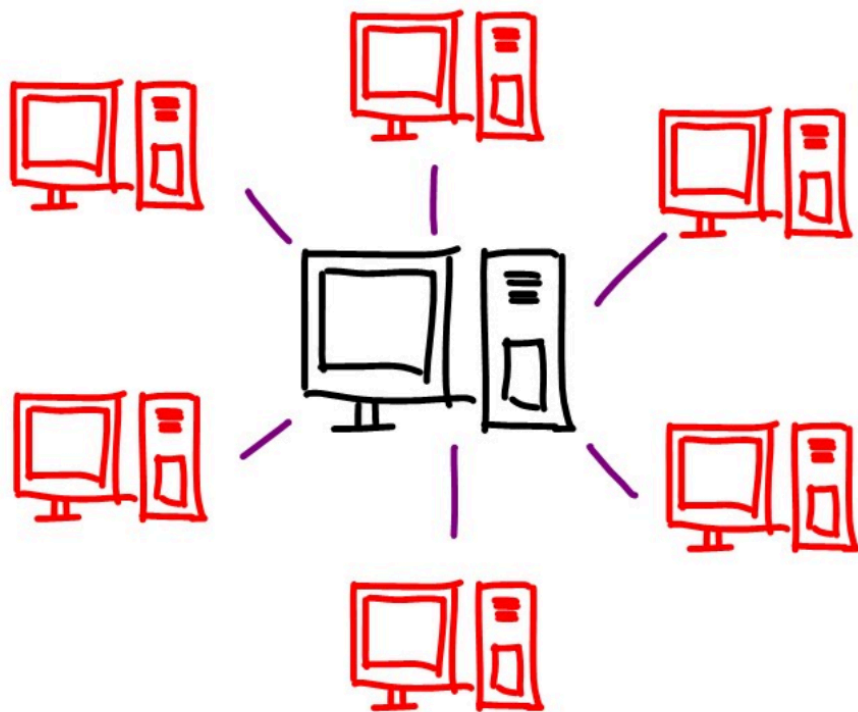
Databricks

전 Cloudera (Hortonworks)

전 Mobigen

Apache Spark?

Distributed
Computing



databricks®

Apache Spark?

● Apache Hadoop
검색어

● Apache Spark
검색어

+ 비교 추가

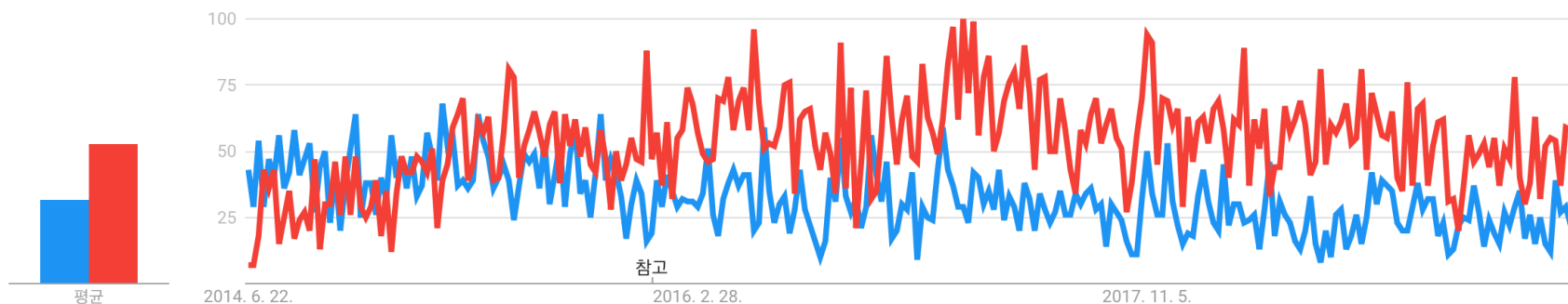
전 세계 ▼

지난 5년 ▼

컴퓨터과학 ▼

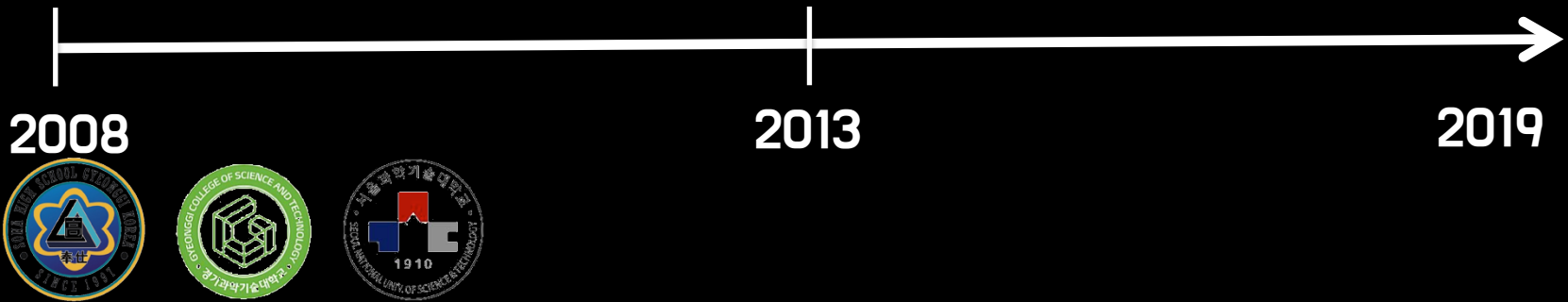
웹 검색 ▼

시간 흐름에 따른 관심도 변화 ⓘ

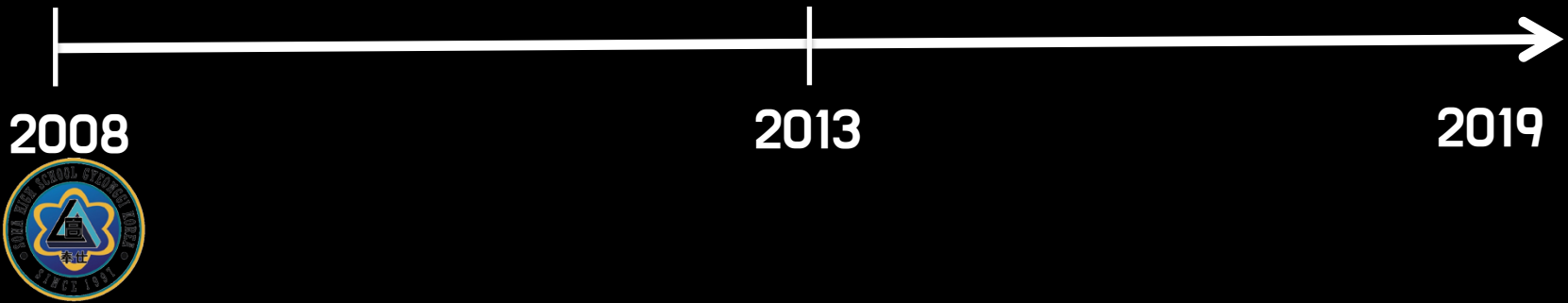




오픈 소스를 알기 전



오픈 소스를 알기 전



오픈 소스를 알기 전

2008



2013

2019

채연 눈물셀카

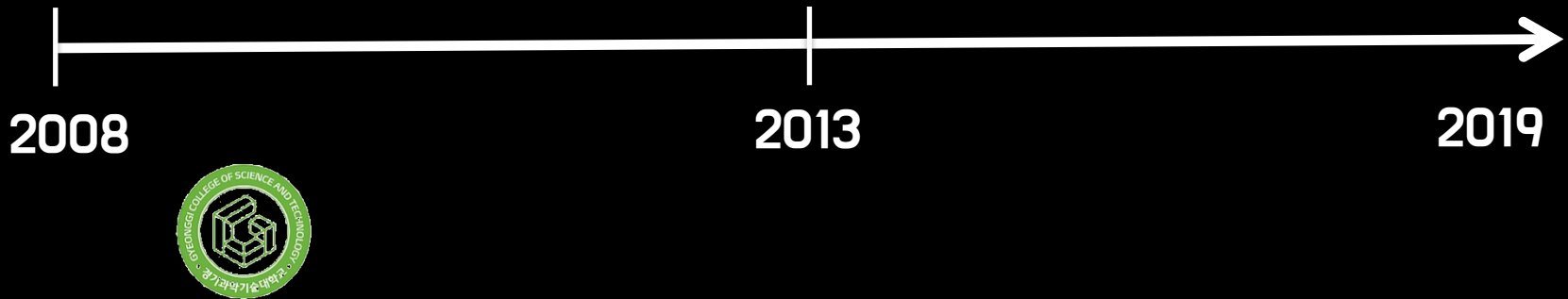


난...가...
눈물을 흘린...
가...
입이...
소...
...
난...
...
...
...
...

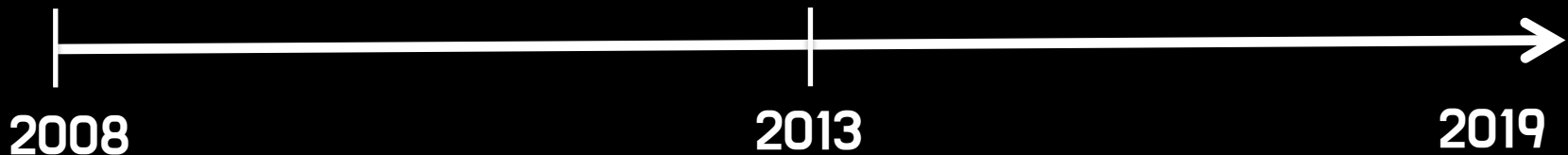


나도... 잔인한 놈이 될 거니까

오픈 소스를 알기 전



오픈 소스를 알기 전



$$I = \frac{dq_m}{dt}, V_m = \frac{d\Phi}{dt}, V_e = IR, q_s = CV_e, \Phi = LI$$

$$\begin{aligned} P = \frac{dW}{dt} &= I \cdot V_e + I \cdot V_m + q_s \cdot \frac{dV_e}{dt} \\ \Rightarrow &= I \cdot V_e + I \cdot \frac{d\Phi}{dt} + CV_e \cdot \frac{dV_e}{dt} \\ &= I^2 R + \frac{d}{dt} \left(\frac{1}{2} LI^2 \right) + \frac{d}{dt} \left(\frac{1}{2} CV_e^2 \right) \end{aligned}$$

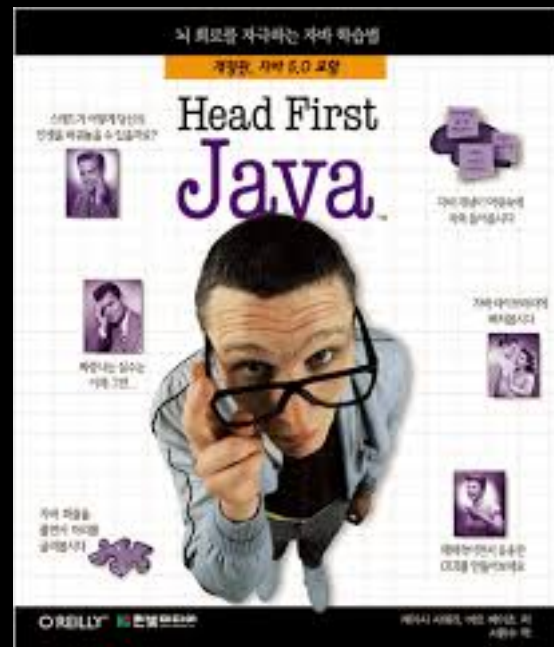
오픈 소스를 알기 전

2008

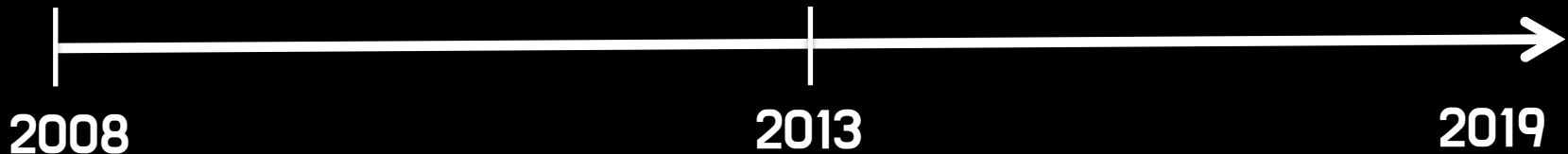


2013

2019

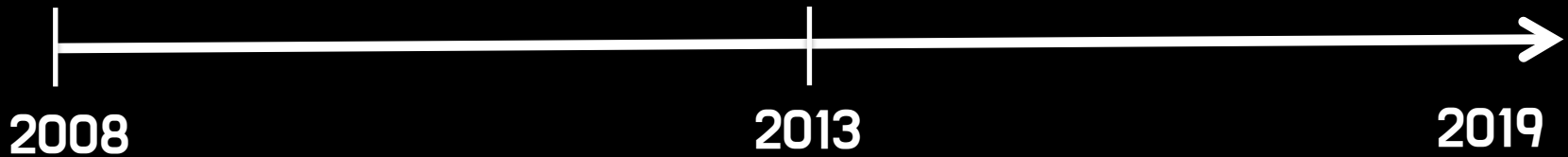


오픈 소스를 알기 전

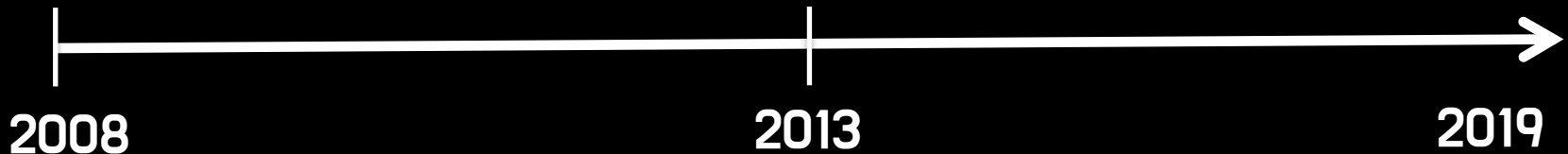


```
public class ThreadSafeInitialization {  
  
    private static ThreadSafeInitialization instance;  
    private ThreadSafeInitialization () {}  
  
    public static synchronized ThreadSafeInitialization getInstance () {  
        if (instance == null)  
            instance = new ThreadSafeInitialization();  
        return instance;  
    }  
  
    public void print () {  
        System.out.println("It's print() method in ThreadSafeInitialization instance.");  
        System.out.println("instance hashCode > " + instance.hashCode());  
    }  
  
}
```

오픈 소스를 알기 전

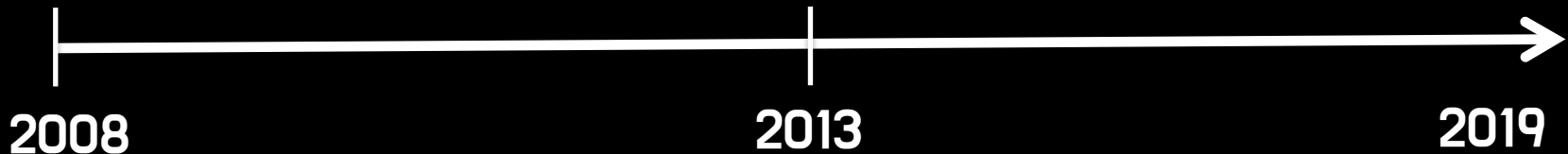


오픈 소스를 알기 전



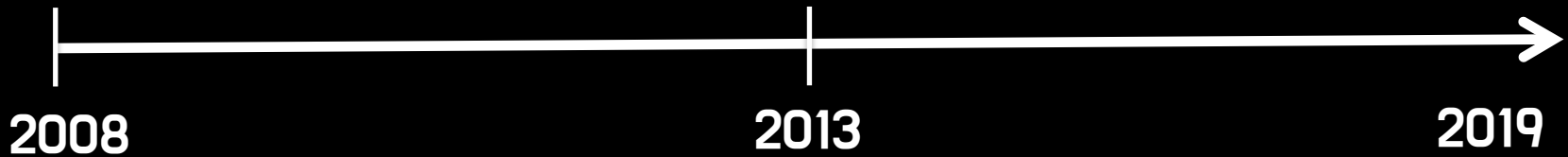
The screenshot shows a Naver search result for the query '리눅스가 뭔가요?'. The search bar at the top contains the text '리눅스가 뭔가요?'. Below the search bar, there are navigation tabs for '통합검색', '카페', '지식IN', '블로그', '웹사이트', '이미지', '동영상', '쇼핑', and '더보기'. The '카페' tab is selected. The search results are displayed under the heading '카페'. The first result is titled '리눅스가 뭔가요??' and dated 2019.01.14. The text of the post describes the author's experience with Linux servers and mentions a link to a Naver cafe. The second result is titled '리눅스가 뭔가요??' and dated 2018.07.07. The text explains that Linux is a computer operating system and mentions a link to a Naver cafe. The third result is titled '리눅스가 뭔가요?' and dated 2018.01.30. The text is partially visible and mentions a link to a Naver cafe.

오픈 소스를 알기 전

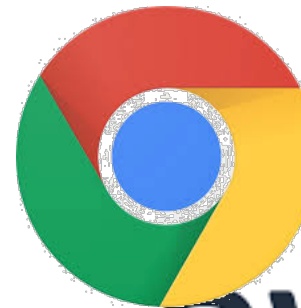
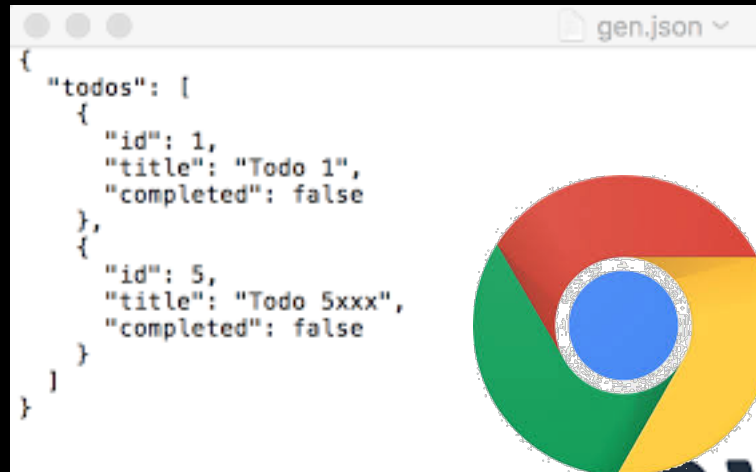
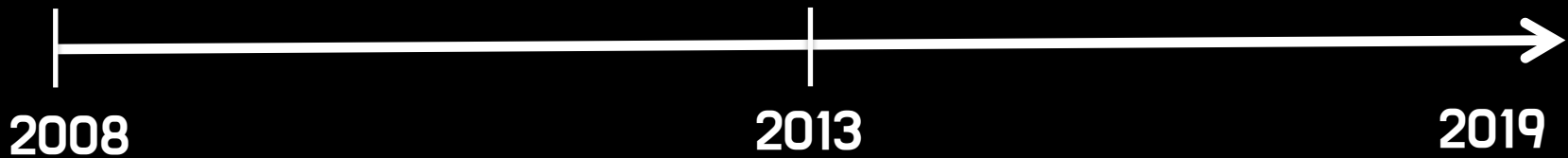


```
public class ThreadSafeInitalization {  
  
    private static ThreadSafeInitalization instance;  
    private ThreadSafeInitalization () {}  
  
    public static synchronized ThreadSafeInitalization getInstance () {  
        if (instance == null)  
            instance = new ThreadSafeInitalization();  
        return instance;  
    }  
  
    public void print () {  
        System.out.println("It's print() method in ThreadSafeInitalization instance.");  
        System.out.println("instance hashCode > " + instance.hashCode());  
    }  
  
}
```

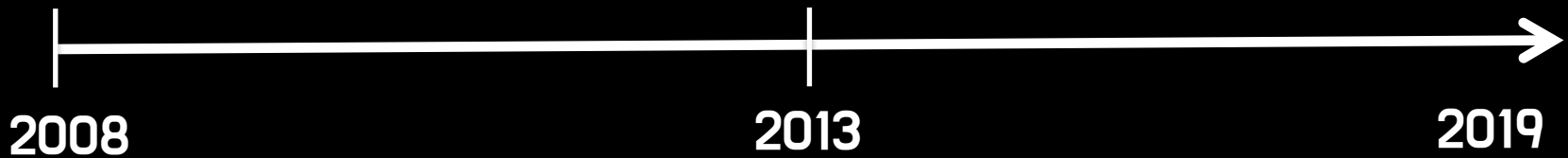
오픈 소스를 알기 전



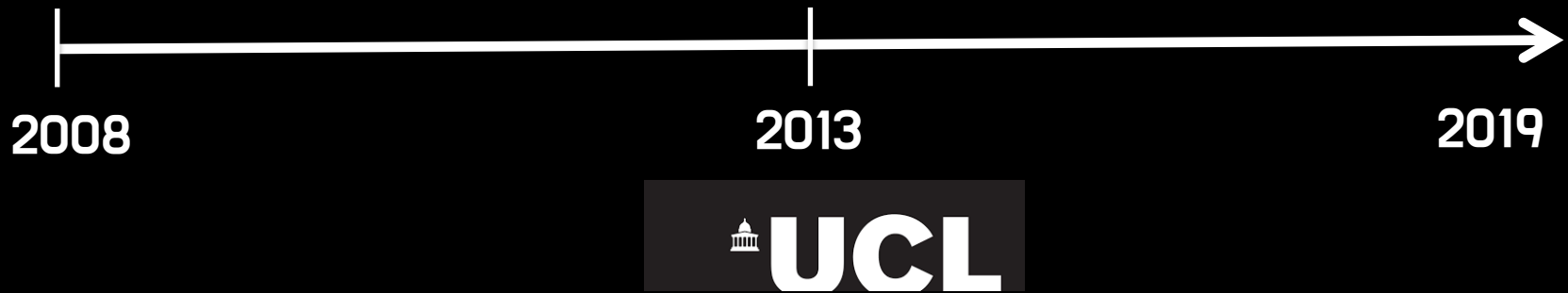
오픈 소스를 알기 전



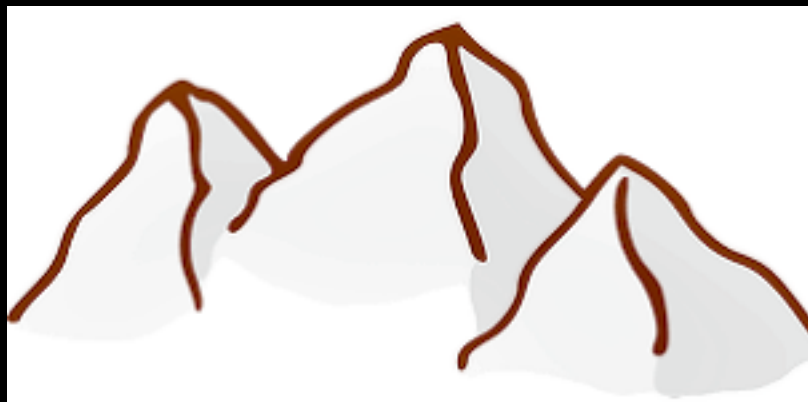
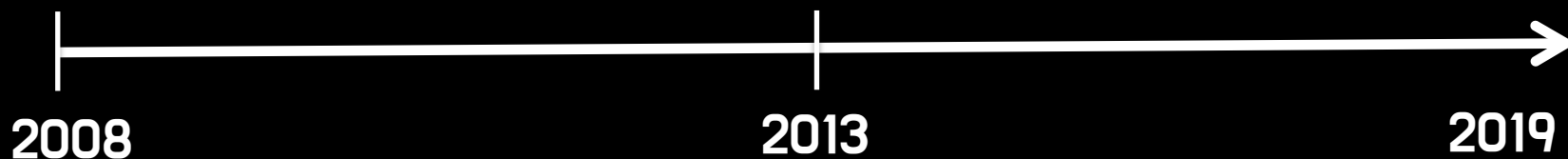
오픈 소스를 안 후



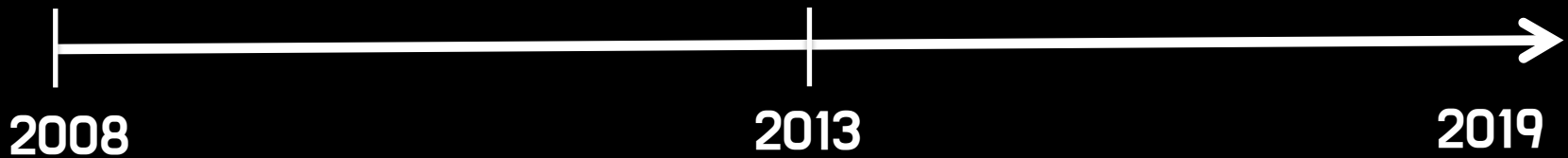
오픈 소스를 안 후



오픈 소스를 안 후

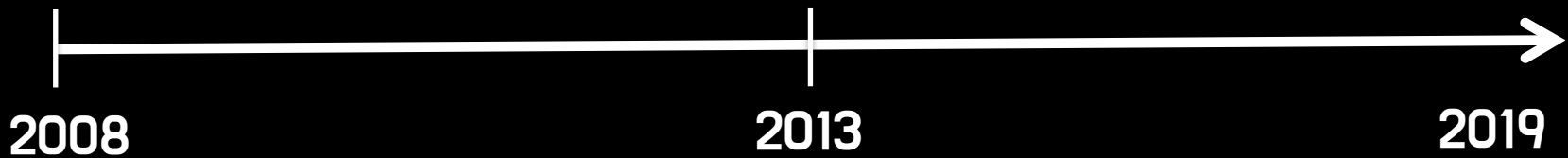


오픈 소스를 안 후



```
... print(
```

오픈 소스를 안 후



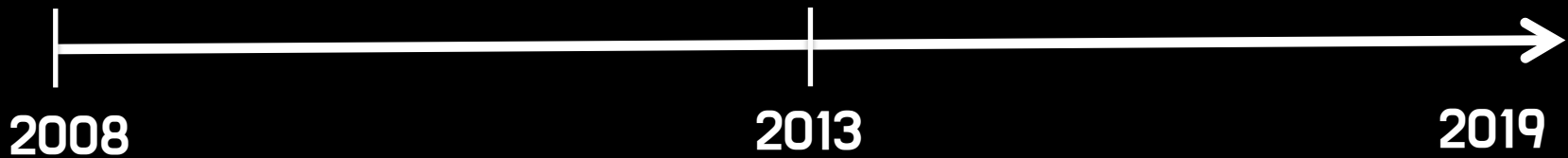
```
def __init__(self, data=None, index=None, columns=None, dtype=None,
             copy=False):
    if data is None:
        data = {}
    if dtype is not None:
        dtype = self._validate_dtype(dtype)

    if isinstance(data, DataFrame):
        data = data._data

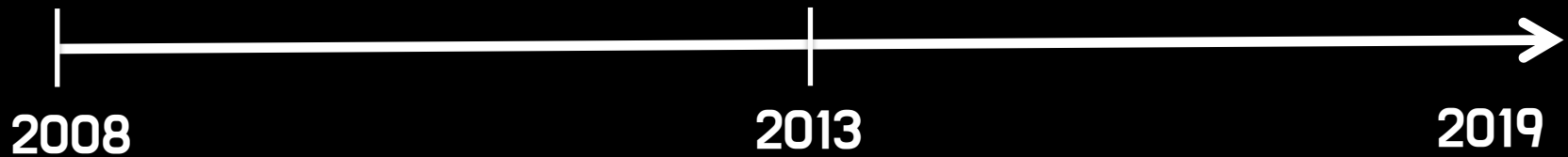
    if isinstance(data, BlockManager):
        mgr = self._init_mgr(data, axes=dict(index=index, columns=columns),
                             dtype=dtype, copy=copy)
    elif isinstance(data, dict):
        mgr = init_dict(data, index, columns, dtype=dtype)
    elif isinstance(data, ma.MaskedArray):
        import numpy.ma.mrecords as mrecords
        # masked recarray
        if isinstance(data, mrecords.MaskedRecords):
            mgr = masked_rec_array_to_mgr(data, index, columns, dtype,
                                           copy)

        # a masked array
    else:
        mask = ma.getmaskarray(data)
        if mask.any():
            data, fill_value = maybe_upcast(data, copy=True)
            data.soften_mask() # set hardmask False if it was True
```

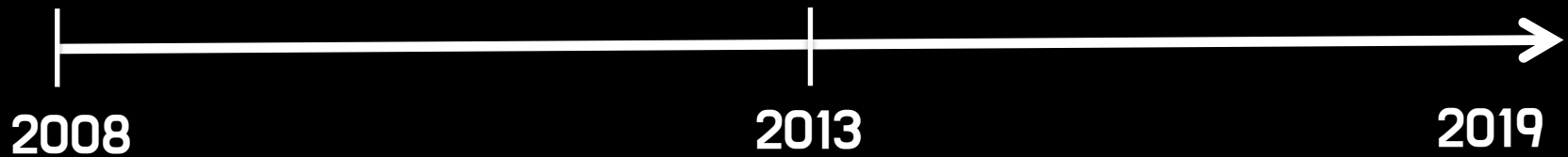
오픈 소스를 안 후



오픈 소스를 안 후

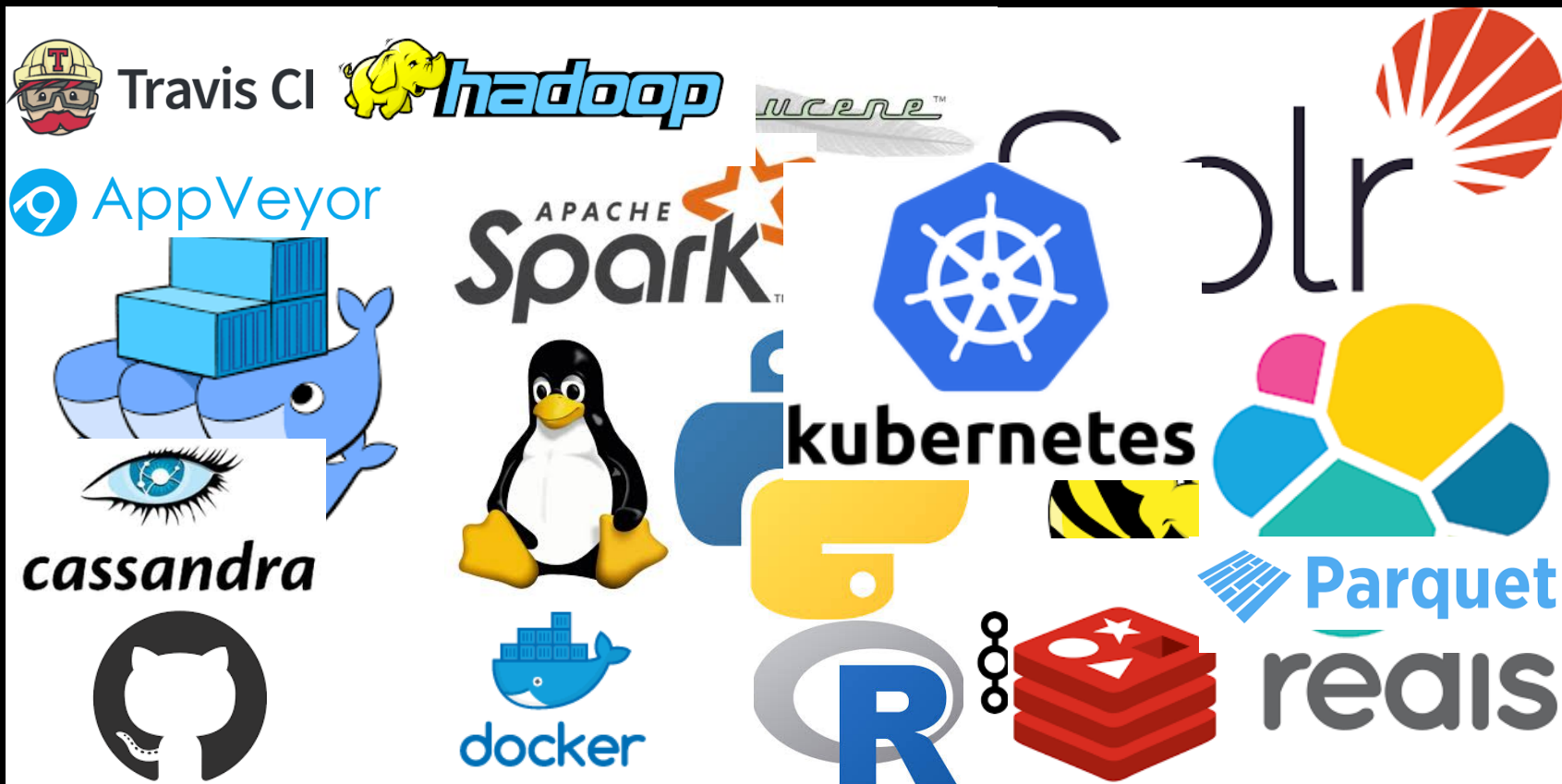
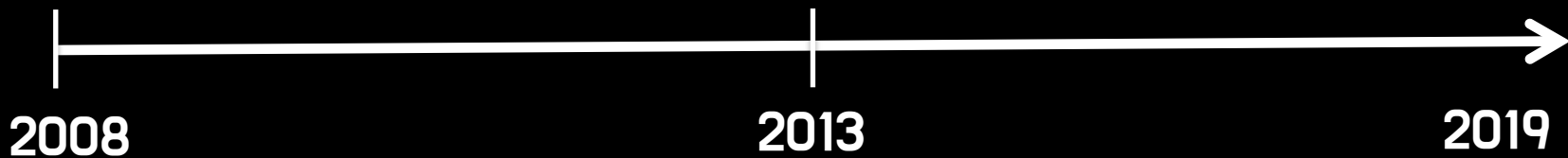


오픈 소스를 안 후

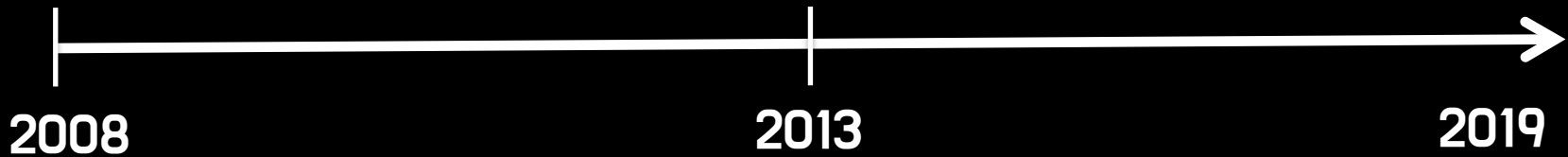


mobigen
Make The Best For Smart Business

오픈 소스를 안 하



오픈 소스를 안 후



Spark / SPARK-9814

EqualNullSafe not passing to data sources

Details

Type:	↑ Improvement	Status:	RESOLVED
Priority:	✓ Minor	Resolution:	Fixed
Affects Version/s:	None	Fix Version/s:	1.5.0
Component/s:	SQL		
Labels:	None		

Description

[SPARK-9814][SQL] EqualNotNull not passing to data sources

New issue

#8096

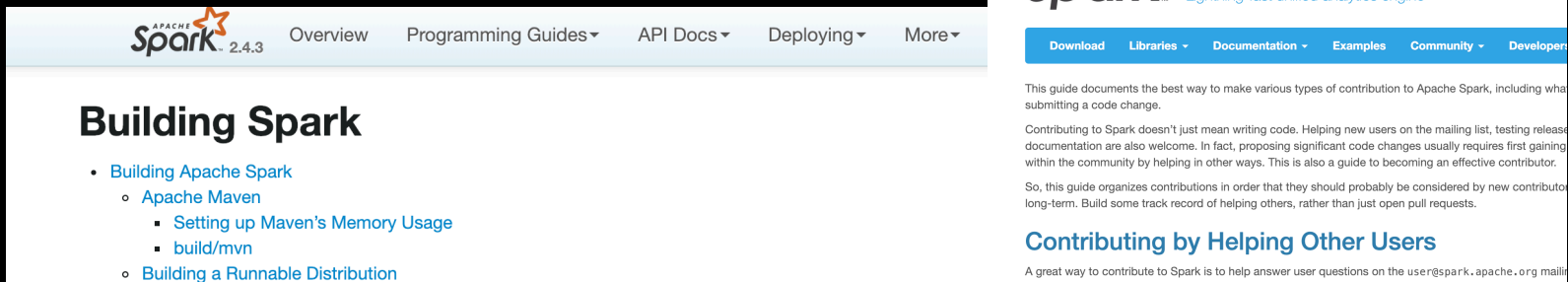
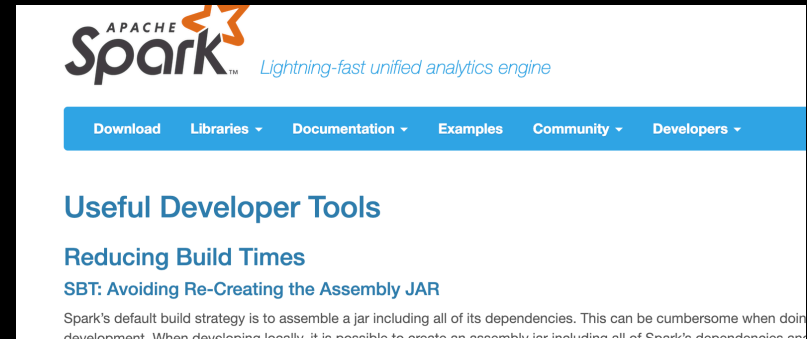
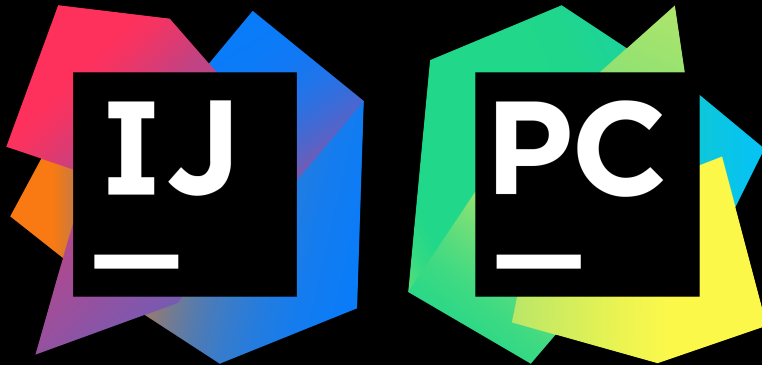
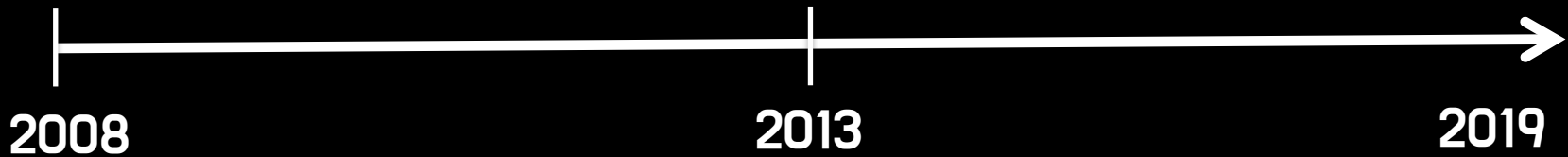
Closed HyukjinKwon wants to merge 3 commits into [apache:master](#) from [unknown repository](#)

Conversation 9 Commits 3 Checks 0 Files changed 3 +15 -0

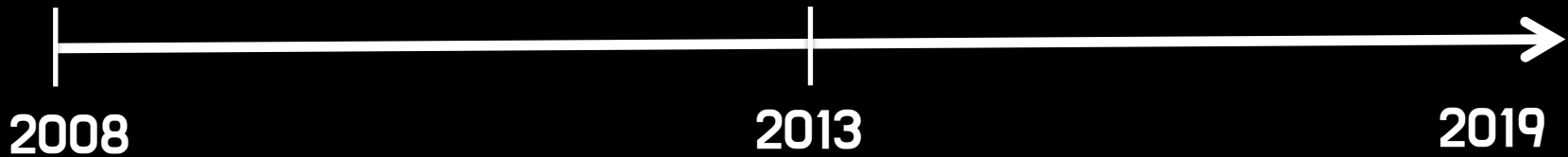
Changes from all commits File filter... Jump to...

```
5  sql/core/src/main/scala/org/apache/spark/sql/execution/datasources/DataSourceStrategy.sca...  
@@ -349,6 +349,11 @@ private[sql] object DataSourceStrategy extends Strategy with Logging {  
349 349      case expressions.EqualTo(Literal(v, _), a: Attribute) =>  
350 350          Some(sources.EqualTo(a.name, v))  
351 351  
352 +      case expressions.EqualNullSafe(a: Attribute, Literal(v, _)) =>  
353 +          Some(sources.EqualNullSafe(a.name, v))  
354 +      case expressions.EqualNullSafe(Literal(v, _), a: Attribute) =>  
355 +          Some(sources.EqualNullSafe(a.name, v))  
356 +
```

오픈 소스를 안 후



오픈 소스를 안 후



FILTERS

<<

New filter

Find filters

My open issues

Reported by me

All issues

Open issues

Done issues

Search

Save as

Spark ▾

Type: All ▾

Status: All ▾

Assignee: All ▾

Contains text

More ▾

🔍

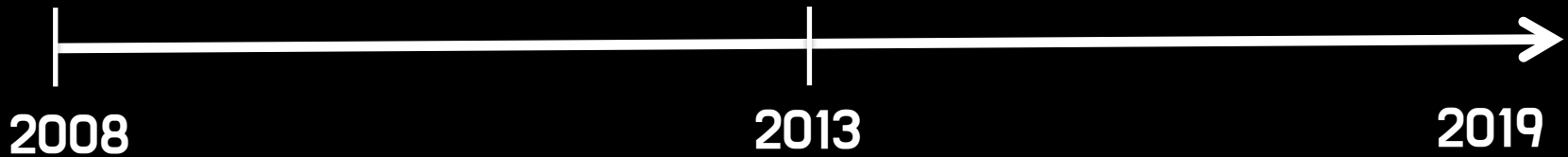
Advanced

Resolution: Unresolved ▾ ×

1–50 of 1872 🔗

T	Patch Info	Key	Summary	Components	Assignee	P ↓
🔴		SPARK-27815	do not leak SaveMode to file source v2	SQL	Wenchang Fan	🚫

오픈 소스를 안 후



519 Open ✓ 24,464 Closed		Author ▾	Labels ▾	Projects ▾	Milestones ▾	Reviews ▾	Assignee ▾	Sort ▾
	[SPARK-28205][SQL] useV1SourceList configuration should be for all data sources							2
	#25004 opened 24 minutes ago by gengliangwang							
	[SPARK-28204][SQL][TESTS] Make separate two test cases for column pruning in binary files							1
	#25003 opened 1 hour ago by HyukjinKwon							
	[SPARK-28203][Core][Python] PythonRDD should respect SparkContext's hadoop configuration							1
	#25002 opened 3 hours ago by advancedxy							
	[SPARK-28083][SQL] Enhance ANSI SQL: LIKE predicate: ESCAPE clause.							3
	#25001 opened 4 hours ago by believer							

apache / spark

Watch 2,138 Star 22,416 Fork 19,348

Code Pull requests 519 Projects 0 Security Insights

Branch: master ▾

Commits on Jun 28, 2019

[SPARK-28077][SQL] Support ANSI SQL OVERLAY function. ...

believer authored and ueshin committed 3 hours ago

832ff87 <>

[SPARK-28185][PYTHON][SQL] Closes the generator when Python UDFs stop... ...

WeichenXu123 authored and HyukjinKwon committed 6 hours ago

31e7c37 <>

[SPARK-28179][SQL] Avoid hard-coded config: spark.sql.globalTempDatabase ...

wangyum authored and HyukjinKwon committed 12 hours ago

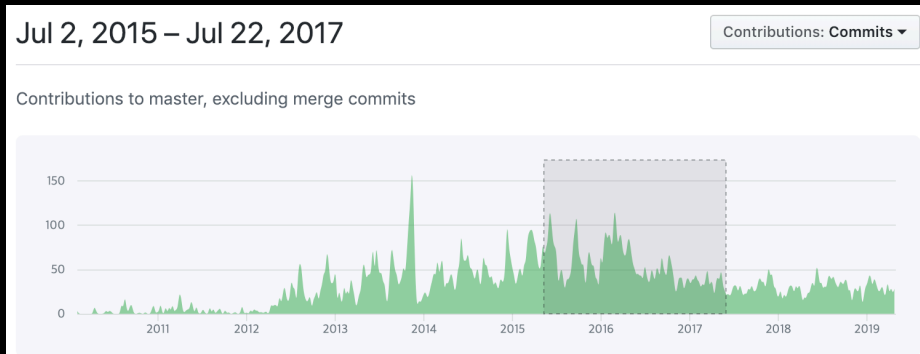
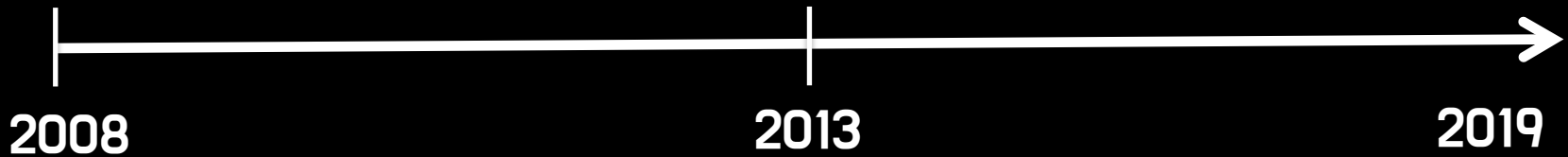
410a898 <>

[SPARK-28187][BUILD] Add support for hadoop-cloud to the PR builder. ...

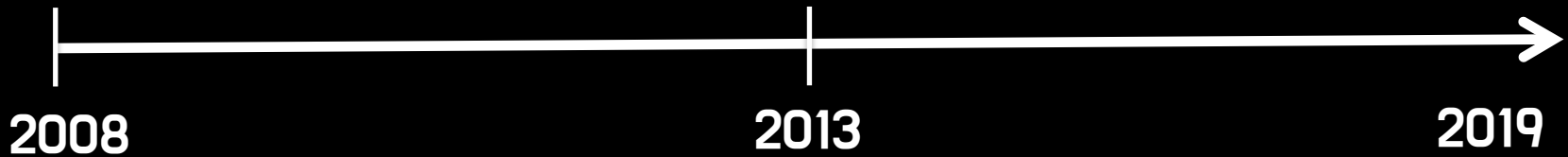
vanzin committed 15 hours ago

11e21cc <>

오픈 소스를 안 후



오픈 소스를 안 후



Current Committers

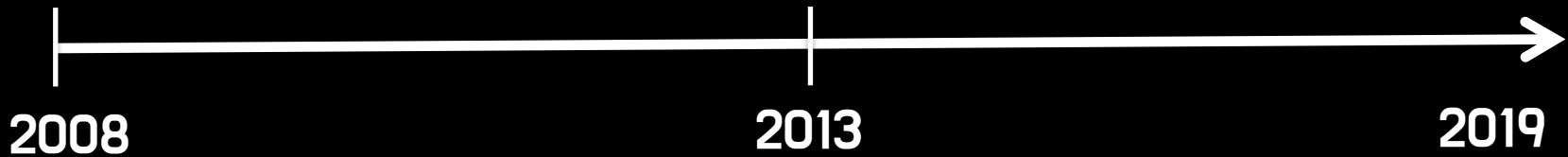
Name

Organization

Hyukjin Kwon

Mobigen

오픈 소스를 안 후



A screenshot of a GitHub comment thread. The top comment is from user 'yuwillrun', posted 5 days ago, with 'Author' and 'Contributor' badges. The comment asks for code review and advice on solving a 'commit' issue, specifically mentioning 'git pull from databricks/master'. The bottom comment is from user 'ueshin', also posted 5 days ago, with a 'Collaborator' badge. The comment responds that they will review later and provides instructions on how to create a feature branch from the HEAD of the updated master branch. The instructions are shown in a code block with a light blue background. The comment concludes with a request to look into the details of each command.

yuwillrun commented 5 days ago Author Contributor ...

@ueshin Need code review. There are too many commit above.
And can you give me some advice how to solve these `commit` ? How to clean branch and commit after
`git pull from databricks/master` ?

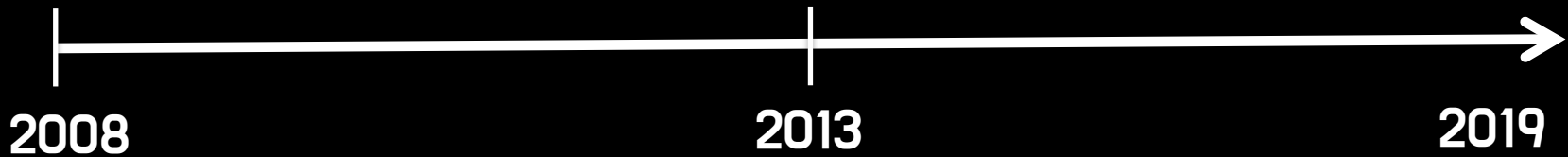
ueshin commented 5 days ago Collaborator ...

I will review later, but as for the commits, you can just cut each feature branch from HEAD of the updated master branch, something like:

```
$ git status
On branch master
...
$ git pull
...
$ git checkout -b feature-branch
... work for the feature
```

Please look into the details of each command.

오픈 소스를 안 후



Diversity #23588

Closed jflittner wants to merge 2 commits into apache:master from jflittner:patch-1

Conversation 15 Commits 2 Checks 0 Files changed 1

Changes from all commits File filter... Jump to... ⚙

3 examples/src/main/resources/people.json

```
... @@ -1,3 +1,6 @@
1 1 {"name":"Michael"}
2 2 + {"name":"Rachel"}
3 3 {"name":"Andy", "age":30}
4 4 {"name":"Justin", "age":19}
5 5 + {"name":"Xiao", "age":22}
6 6 + {"name":"Vinitha", "age":34}
```



HyukjinKwon commented on 19 Jan

Member



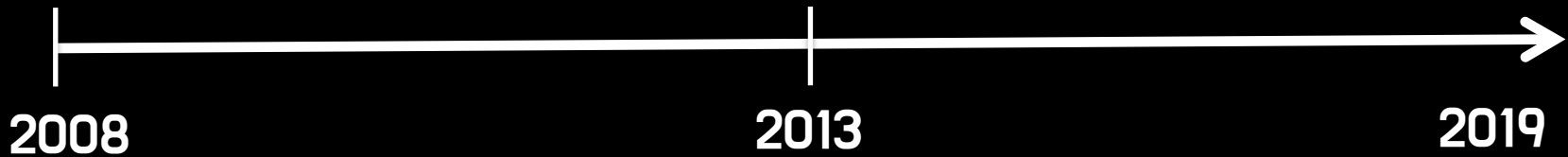
Please review <http://spark.apache.org/contributing.html> before opening a pull request.


Also, I don't think it's worth to fix.



HyukjinKwon closed this on 19 Jan


오픈 소스를 안 후



 kmarekspartz commented on 20 Jan ...


To clarify, which portion of those guidelines is this not compliant with?

There's no JIRA, but the guidelines are clear that no JIRA is required for contributions like these.

 HyukjinKwon commented on 20 Jan Member ...


This

The PR title should be of the form [SPARK-xxxx][COMPONENT] Title, where SPARK-xxxx is the relevant JIRA number, COMPONENT is one of the PR categories shown at spark-prs.appspot.com and Title may be the JIRA's title or a more specific title describing the PR itself.


 kmarekspartz commented on 20 Jan ...

A bug may be reported by creating a JIRA but without creating a pull request

However, trivial changes, where the what should change is virtually the same as the how it should change do not require a JIRA. Example: Fix typos in Foo scaladoc


 HyukjinKwon commented on 20 Jan Member ...

Yea so do you think "Diversity" implies what the PR proposes? Take a look at other PRs

 HyukjinKwon commented on 20 Jan Member ...


See all the minor PRs <https://github.com/apache/spark/pulls?q=is%3Apr+minor+is%3Aclosed>

Why do you think the guide doesn't work for this PR specifically? What makes you think this PR complies the doc? Shall we fix the doc then?

 kmarekspartz commented on 21 Jan ...

The guidelines say nothing about requirements for minor PRs. If you are going to close minor PRs for being non-compliant, please upgrade the guidelines so that people know the undocumented hoops they have to jump through to make a contribution.

Note, too: the term "guidelines" implies optional. You may want to rebrand them as "contribution requirements" when you add the currently undocumented requirements.

 HyukjinKwon commented on 21 Jan • edited Member ...

No, read the comment and document above. I closed this PR because the change doesn't look worth to fix. Title should be specific: Title may be the JIRA's title or a more specific title describing the PR itself.

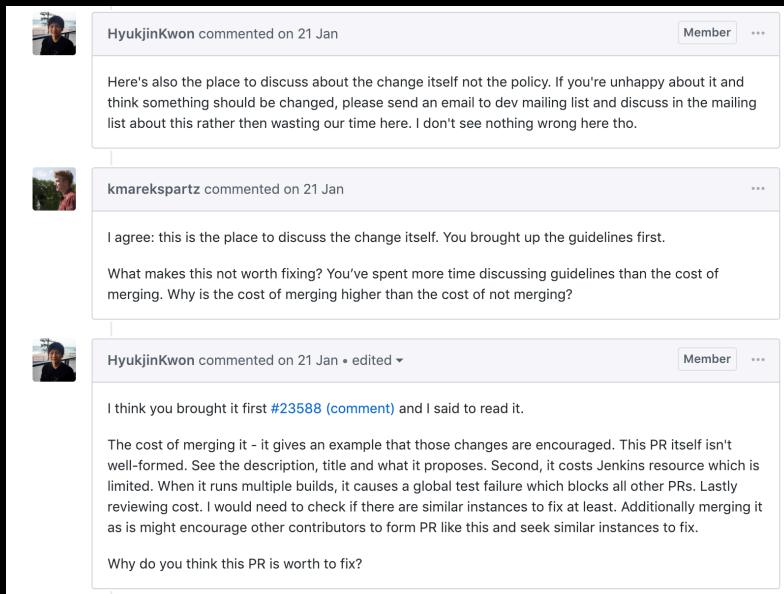
Note that, I said "Please review <http://spark.apache.org/contributing.html> before opening a pull request.". There's nothing required here.

오픈 소스를 안 후

2008

2013

2019



HyukjinKwon commented on 21 Jan

Here's also the place to discuss about the change itself not the policy. If you're unhappy about it and think something should be changed, please send an email to dev mailing list and discuss in the mailing list about this rather than wasting our time here. I don't see nothing wrong here tho.

kmarekspartz commented on 21 Jan

I agree: this is the place to discuss the change itself. You brought up the guidelines first.

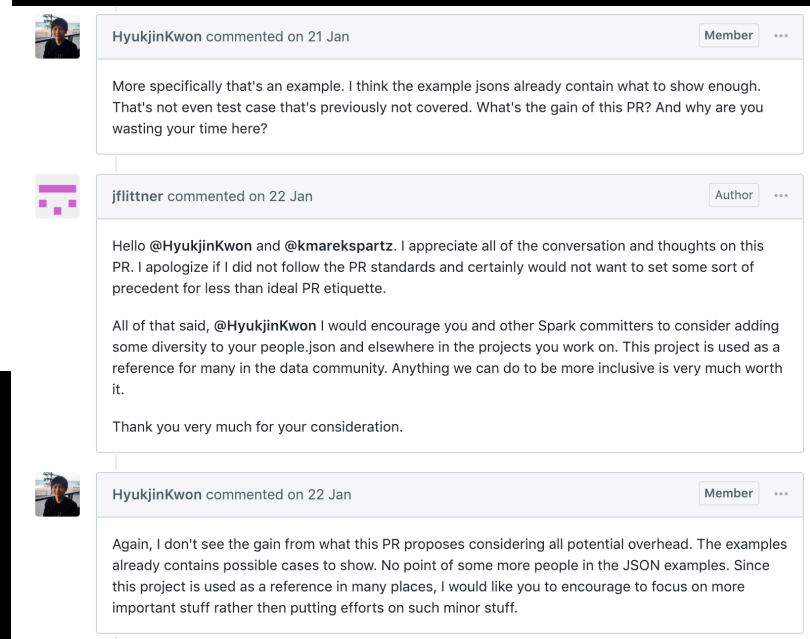
What makes this not worth fixing? You've spent more time discussing guidelines than the cost of merging. Why is the cost of merging higher than the cost of not merging?

HyukjinKwon commented on 21 Jan • edited

I think you brought it first [#23588 \(comment\)](#) and I said to read it.

The cost of merging it - it gives an example that those changes are encouraged. This PR itself isn't well-formed. See the description, title and what it proposes. Second, it costs Jenkins resource which is limited. When it runs multiple builds, it causes a global test failure which blocks all other PRs. Lastly reviewing cost. I would need to check if there are similar instances to fix at least. Additionally merging it as is might encourage other contributors to form PR like this and seek similar instances to fix.

Why do you think this PR is worth to fix?



HyukjinKwon commented on 21 Jan

More specifically that's an example. I think the example jsons already contain what to show enough. That's not even test case that's previously not covered. What's the gain of this PR? And why are you wasting your time here?

jflittner commented on 22 Jan

Hello @HyukjinKwon and @kmarekspartz. I appreciate all of the conversation and thoughts on this PR. I apologize if I did not follow the PR standards and certainly would not want to set some sort of precedent for less than ideal PR etiquette.

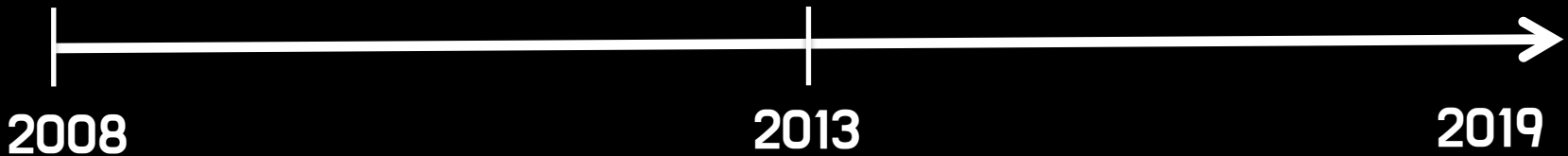
All of that said, @HyukjinKwon I would encourage you and other Spark committers to consider adding some diversity to your people.json and elsewhere in the projects you work on. This project is used as a reference for many in the data community. Anything we can do to be more inclusive is very much worth it.

Thank you very much for your consideration.

HyukjinKwon commented on 22 Jan

Again, I don't see the gain from what this PR proposes considering all potential overhead. The examples already contains possible cases to show. No point of some more people in the JSON examples. Since this project is used as a reference in many places, I would like you to encourage to focus on more important stuff rather than putting efforts on such minor stuff.

오픈 소스를 안 후



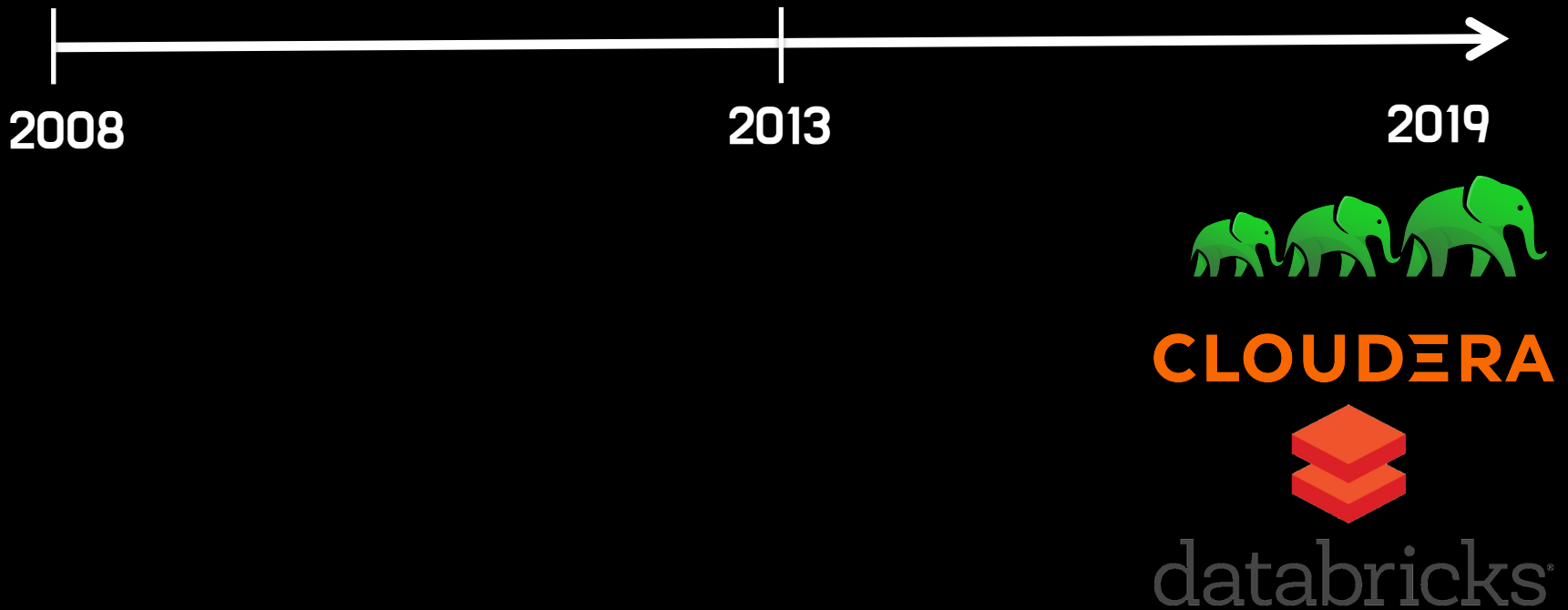
배운점

- 나의 능력치는 평범 혹은 그 이하이다.
- 코드는 최대한 이해하기 쉽게.
- 로직 작성 10% 쉽게 다시 작성 40% 테스트 50%
- 아는척 나대지말자. 결국 탄로난다.
- 알아도 겸손. 중요한게 아니면 그냥 저춰도된다.
- 작은건 그냥 넘어가라.
- 문제 해결에만 집중하자. 그 외에 것, 평판 걱정, 정치, 다 쓰잘대기 없다. 그럼 문제가 빨리 해결된다
- 구글해서 열어본 페이지 끝까지 정독하기
- 프로젝트 별 규칙 및 방법론 준수하기
- 익숙해 질때가지, 코드 한 줄 한 줄을 구글, 일반적
으로 받아들여지는 최선의 코드를 손에 익혀놓는다.

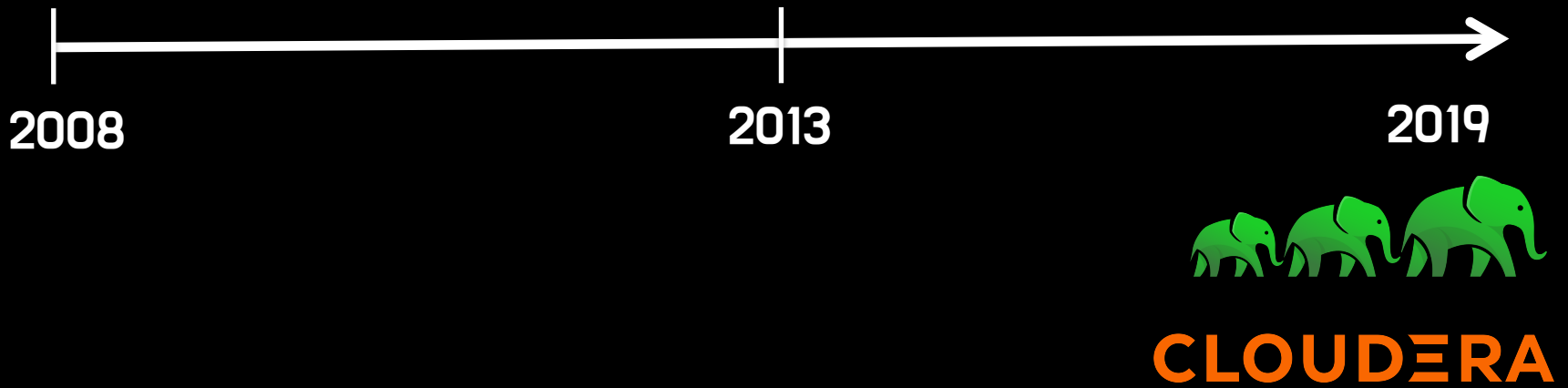
부작용

- 열등감 만빵
- 수면 부족
- 스트레스

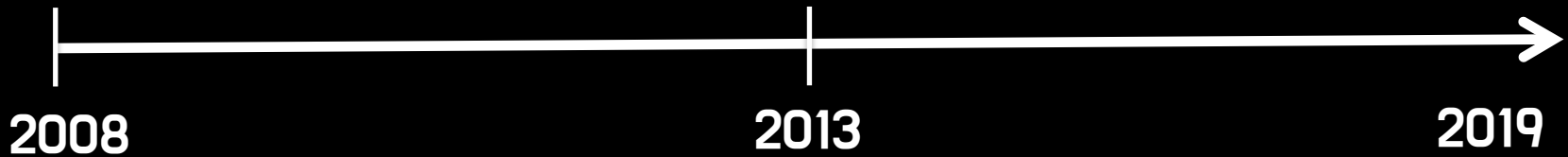
오픈 소스 본격 시작



오픈 소스 본격 시작



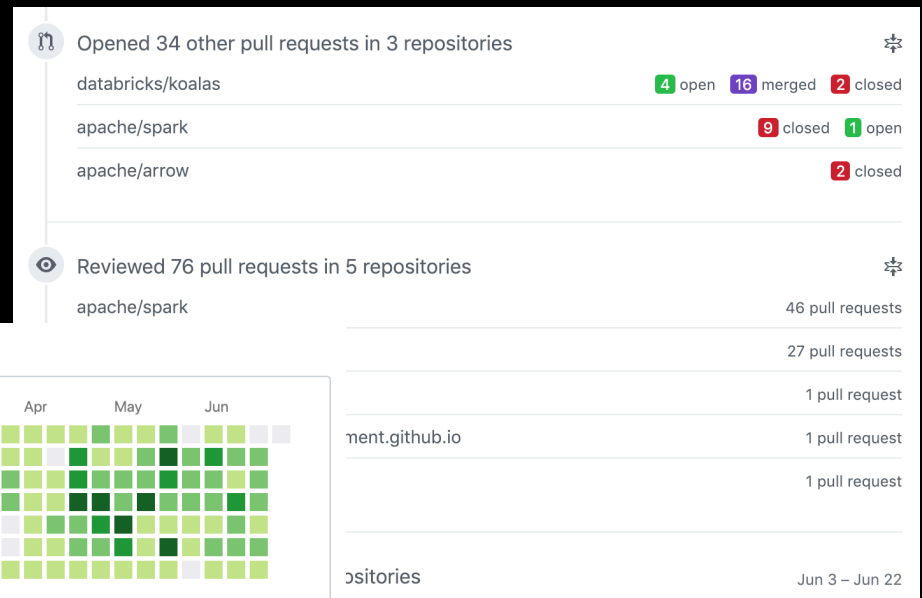
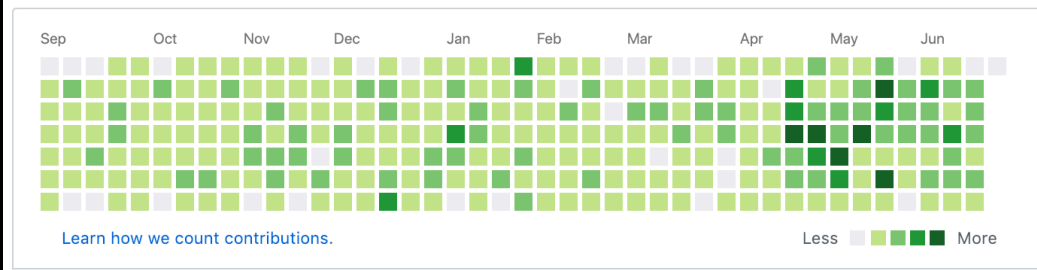
오픈 소스 본격 시작



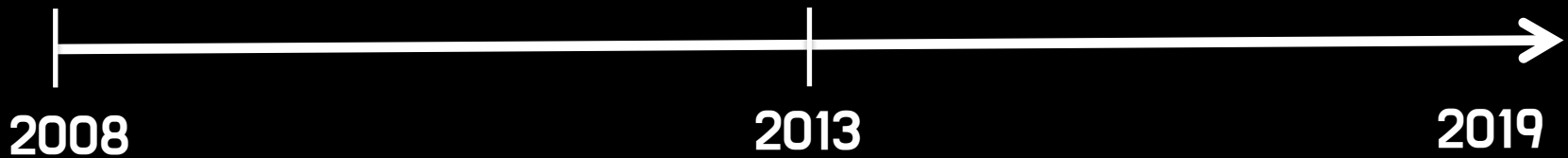
CLOUDERA



1,822 contributions in the last year



오픈 소스 본격 시작



CLOUDERA

SPEAKERS



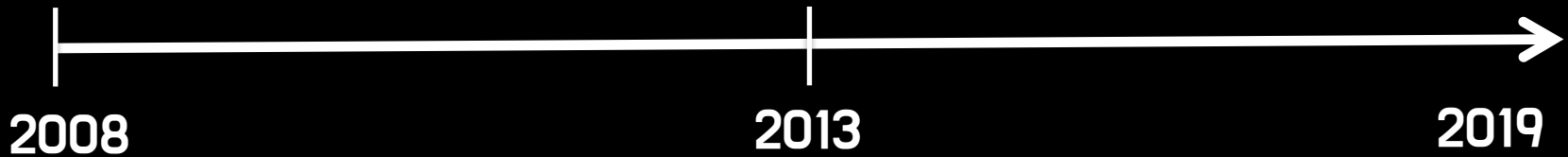
Hyukjin Kwon
SQL, Python, Open source

HYUKJIN KWON
Software Engineer
Hortonworks

DATA WORKS



오픈 소스 본격 시작



CLOUDERA

A screenshot of the Apache Spark Open issues page. The header includes the Apache Software Foundation logo, a search bar, and a 'Log In' link. The main content area shows 'Open issues' with a 'Switch filter' dropdown and a link to 'View all issues and filters'. Below this, there's a section for 'Order by Priority' and a list of issues, including one titled 'Spark / SPARK-27815'.

Apache Spark Developers List [Login](#) [Regist](#)

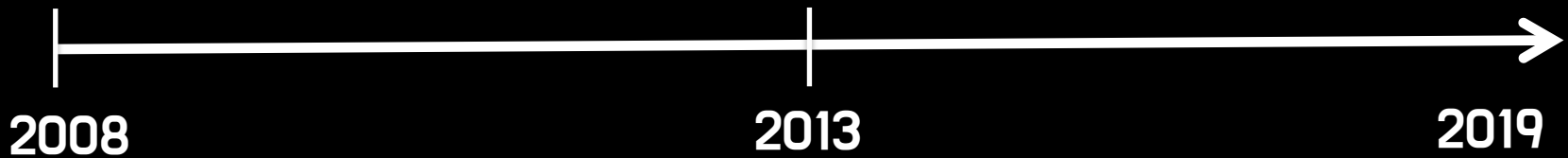
People in Apache Spark Developers List

[Users & Groups](#) [Online Users](#)

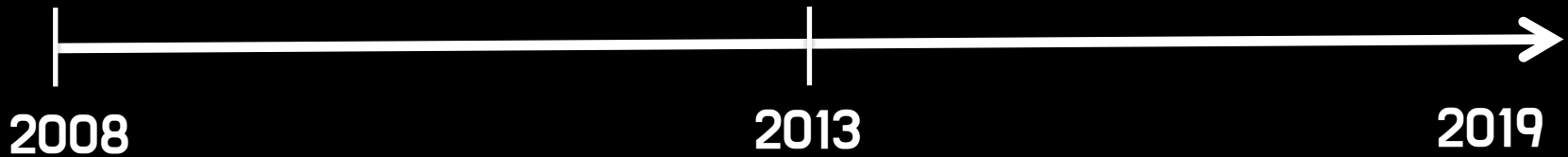
Filter by group [Registered Users](#) 1671 users [1](#) [2](#) [3](#) [4](#) ... [84](#)

Name		Post Count
rxin	Registered	1369
cloud0fan	Registered	251
Hyukjin Kwon	Registered	240

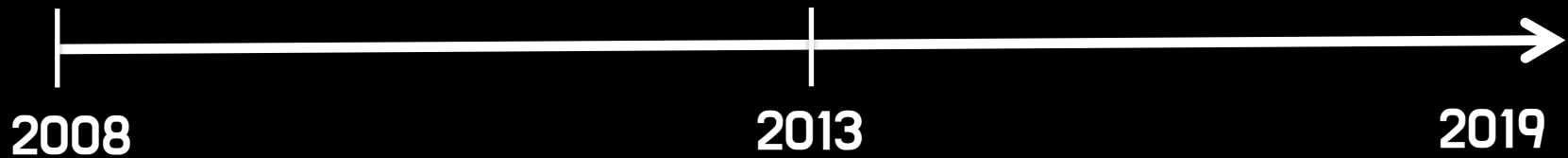
오픈 소스 본격 시작



오픈 소스 본격 시작



오픈 소스 본격 시작

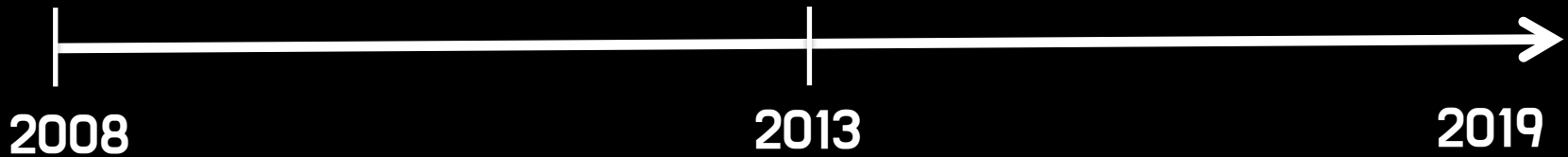


Invitation to join the Apache
Spark PMC Inbox



databricks®

오픈 소스 본격 시작



배운점

- **강약중강약**
- 오픈소스 활동은 영구적 자기소개서 & 실력 증명서
- 하위 호환성
- 눈으로 읽는천재보다 테스트 해보는 바보가 낫다



CLOUDBERA

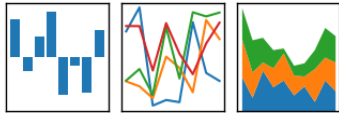


databricks®

Koalas

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



```
import pandas as pd
df = pd.DataFrame({'x': [1, 2], 'y': [3, 4], 'z': [5, 6]})
# Rename columns
df.columns = ['x', 'y', 'z1']
# Do some operations in place
df['x2'] = df.x * df.x
```

```
import databricks.koalas as ks
df = ks.DataFrame({'x': [1, 2], 'y': [3, 4], 'z': [5, 6]})
# Rename columns
df.columns = ['x', 'y', 'z1']
# Do some operations in place
df['x2'] = df.x * df.x
```

<https://github.com/databricks/koalas>

Q & A