

Github PR Analysis

Jin Yang and Maria del Carmen Sacristan Benjet

Dataset

- Kaggle: GitHub Pull Requests
 - <https://www.kaggle.com/datasets/stephangarland/ghtorrent-pull-requests>
 - A sampling of GitHub pull requests with metadata, sourced from GHTorrent.
- *Acknowledgements*
 - *Georgios Gousios for creating the GHTorrent dataset.*

Data Explorer

Version 1 (92.53 GB)

ghtorrent-2019-01-07.csv

ghtorrent-2019-02-04.csv

ghtorrent-2019-03-11.csv

ghtorrent-2019-04-15.csv

ghtorrent-2019-05-20.csv

Data Explorer

Version 1 (92.53 GB)

ghtorrent-2019-01-07.csv

ghtorrent-2019-02-04.csv

ghtorrent-2019-03-11.csv

ghtorrent-2019-04-15.csv

ghtorrent-2019-05-20.csv

ghtorrent-2019-01-07.csv (17.61 GB)

Detail Compact Column

actor_login	actor_id	comment_id	comment	repo	language	author_login	author_id	pr_id	c_id	commit_date
<div><div>houndci-bot2%</div><div>codacy-bot1%</div><div>Other (81409107)98%</div></div>	<div><div></div><div>551.7m</div></div>	<div><div></div><div>192m290m</div></div>	<div><div></div><div>4485621unique values</div></div>	<div><div>kibana3%</div><div>elasticsearch3%</div><div>Other (78375085)94%</div></div>	<div><div>[null]23%</div><div>JavaScript13%</div><div>Other (53293849)64%</div></div>	<div><div>danobot1%</div><div>k8s-ci-robot1%</div><div>Other (81629072)98%</div></div>	<div><div></div><div>551.7m</div></div>	<div><div></div><div>464k64.1m</div></div>	<div><div></div><div>8.11m1.42b</div></div>	<div><div></div><div>53440unique values</div></div>
nikolagjorgjievski	34368385	251381442	Do you think its better if we use cards here instead of a list ? \https://getbootstrap.com/docs/4.0/c...	journal-app		TamaraStankovska	13293313	54298419	1282815601	2019-01-07 17:11:19 UTC
nikolagjorgjievski	34368385	246438304	Fix indent of \transition name=\modal\\	journal-app		TamaraStankovska	13293313	53137423	1263914484	2019-01-07 19:32:35 UTC
tn3rb	2327533	248888072	y u no ternary ? \\the previous 12 lines could have been:\\``js\\existingIds = removal ?\\existingIds...	event-espresso-core	PHP	joshfeck	1377750	49682118	1274046759	2019-01-07 22:29:38 UTC
caalador	991111	198711352	`getRel`	flow	Java	pleku	1621377	41569148	1054030538	2019-01-07 05:53:13 UTC
XuHuaiyu	5620059	277629250	Gotcha, never mind.	tidb	Go	XuHuaiyu	5620059	60396802	1371118363	2019-01-07 11:19:59 UTC
lgatto	110974	243489756	A `return` statement is only valid inside a function. I think here you should have\\``r\\stopifnot(va...	pRolocdata	R	ococrook	32435507	51773001	1236218652	2019-01-07 17:23:06 UTC

Hadoop

Python

Spark

Bitnami-centos

Docker

Config

- Windows 11, 21H2
- Docker Desktop for Windows, (Docker version: 20.10.21)
- Spark Docker Image: [bitnami-docker-spark](#), (Spark version: 3.2.0)
- Hadoop 3.2.0

Why use Bitnami Images?

- Bitnami closely tracks upstream source changes and promptly publishes new versions of this image using our automated systems.
- With Bitnami images the latest bug fixes and features are available as soon as possible.
- Bitnami containers, virtual machines and cloud images use the same components and configuration approach - making it easy to switch between formats based on your project needs.




Docker Compose

- 1 master
- 2 workers













Uploading file to HDFS

[Hadoop](#) [Overview](#) [Datanodes](#) [Datanode Volume Failures](#) [Snapshot](#) [Startup Progress](#) [Utilities ▾](#)

Browse Directory

Show entries

<input type="checkbox"/>	 Permission	 Owner	 Group	 Size	 Last Modified	 Replication	 Block Size	 Name	
<input type="checkbox"/>	-rw-r--r--	root	supergroup	18.45 GB	Dec 05 23:47	2	128 MB	ghorrent-2019-03-11.csv	
<input type="checkbox"/>	-rw-r--r--	root	supergroup	16.15 GB	Dec 06 00:25	2	128 MB	ghorrent-2019-05-20.csv	
<input type="checkbox"/>	-rw-r--r--	root	supergroup	440 B	Dec 05 20:48	2	128 MB	words.txt	

Showing 1 to 3 of 3 entries

Hadoop, 2019.

Using Spark Shell

```
Command Prompt - docker e X + -
9870->9870/tcp, 0.0.0.0:19888->19888/tcp final-spark-1

D:\final>docker exec -it d4 bash
root@master:/opt# pyspark
Python 3.8.12 (default, Nov 12 2021, 08:41:47)
[GCC 8.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
2022-12-06 05:06:06,116 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
Welcome to

      /---/
     /   /  \  /---/  /---/
    /___/    \ /   /   /
   /___/      \___/   /
  /___/        /___/

version 3.2.0

Using Python version 3.8.12 (default, Nov 12 2021 08:41:47)
Spark context Web UI available at http://master:4040
Spark context available as 'sc' (master = local[*], app id = local-1670303167652).
SparkSession available as 'spark'.
>>> df = spark.read.csv(r"hdfs://master:9000/ghtorrent-2019-03-11.csv").toDF('actor_login','actor_id',
', 'comment_id', 'comment', 'repo', 'language', 'author_login', 'author_id', 'pr_id', 'c_id', 'commit_date')
>>> df.count()
94382017
>>> df.head(20)
```

```
>>> df.show(20)
```

actor_login	actor_id	comment_id	comment	repo	language	author_login	author_id	pr_id	c_id	commit_date
actor_login	actor_id	comment_id	comment	repo	language	author_login	author_id	pr_id	c_id	commit_date
BinaryMuse	36779	239136422	What's the purpos...	electron	C++	QKWWTINH	42065371	50337589	1225661033	2019-03-11 23:15:...
codacy-bot	16780017	208331443	![Codacy](https://...	prisma	Scala	mavilein	1548697	43592438	1094087538	2019-03-11 17:01:...
codacy-bot	16780017	222245297	![Codacy](https://...	prisma	Scala	mavilein	1548697	46158474	1154135408	2019-03-11 11:11:...
codacy-bot	16780017	212660181	![Codacy](https://...	prisma	Scala	mavilein	1548697	44114896	1111436354	2019-03-11 10:14:...
cyrichardson	1898538	227897117	https://login.rac...	rackspace-how-to	HTML	catlook	2844743	48293256	1178221524	2019-03-11 15:33:...
jhugman	113326	219343919	Nit: Rename to ma...	lockbox-android	null	diegolucasb	8237344	46051824	1140462360	2019-03-11 01:10:...
sashei	11525710	218993671	nice	lockbox-android	null	diegolucasb	8237344	46053102	1139308052	2019-03-11 01:10:...
shresthamalik	12456559	220744800	For most ops, we ...	ngraph-tf	C++	jianyinglang	42276008	46380241	1146776553	2019-03-11 21:18:...
fsamin	322226	273806993	We need unit test...	cds	null	richardt	3559940	59244612	1359376382	2019-03-11 07:25:...
JoseEmilio-ARM	39419276	283213493	Hi @alan-baker , ...	SPIRV-Tools	C++	sarahM0	49081552	51419417	1376665791	2019-03-11 18:10:...
redallen	49219175	284262252	Change the functi...	patternfly-react	null	kmcfaul	47663245	62245431	1395232883	2019-03-11 16:34:...
karlnapf	60498	253995599	KMeans is another...	shogun	C++	avramidis	3924844	54811239	1292983693	2019-03-11 14:44:...
dlcjr2015	7777337	243773575	Check this change?	intro-html	null	mariamawuena	49981213	51951797	1077578018	2019-03-11 17:22:...
silvin-lubecki	38037278	232304521	multi-endpoints ?	cli	Go	silvin-lubecki	38037278	49270925	1197341369	2019-03-11 14:59:...
shahzadlone	36995953	260550556	```nullptr```	tensorflow	C++	siju-samuel	39610689	56474495	1310997732	2019-03-11 03:52:...
weekface	150750	229989378	ditto	tidb-operator	Go	UQOHEGBG	49979448	48806401	1187585780	2019-03-11 13:07:...
devinbileck	44195133	255672968	@freimair Since I...	bisq	null	ben-kaufman	42853651	55100928	1294109818	2019-03-11 04:45:...
jstaffor	32387584	201382880	Choose a build l...	mobile-docs	null	finp	33052983	42160835	1066421476	2019-03-11 15:27:...
k8s-ci-robot	30337365	221709060	Golint unexported...	test-infra	null	hongkailiu	1845108	45496759	1131818081	2019-03-11 17:10:...

only showing top 20 rows

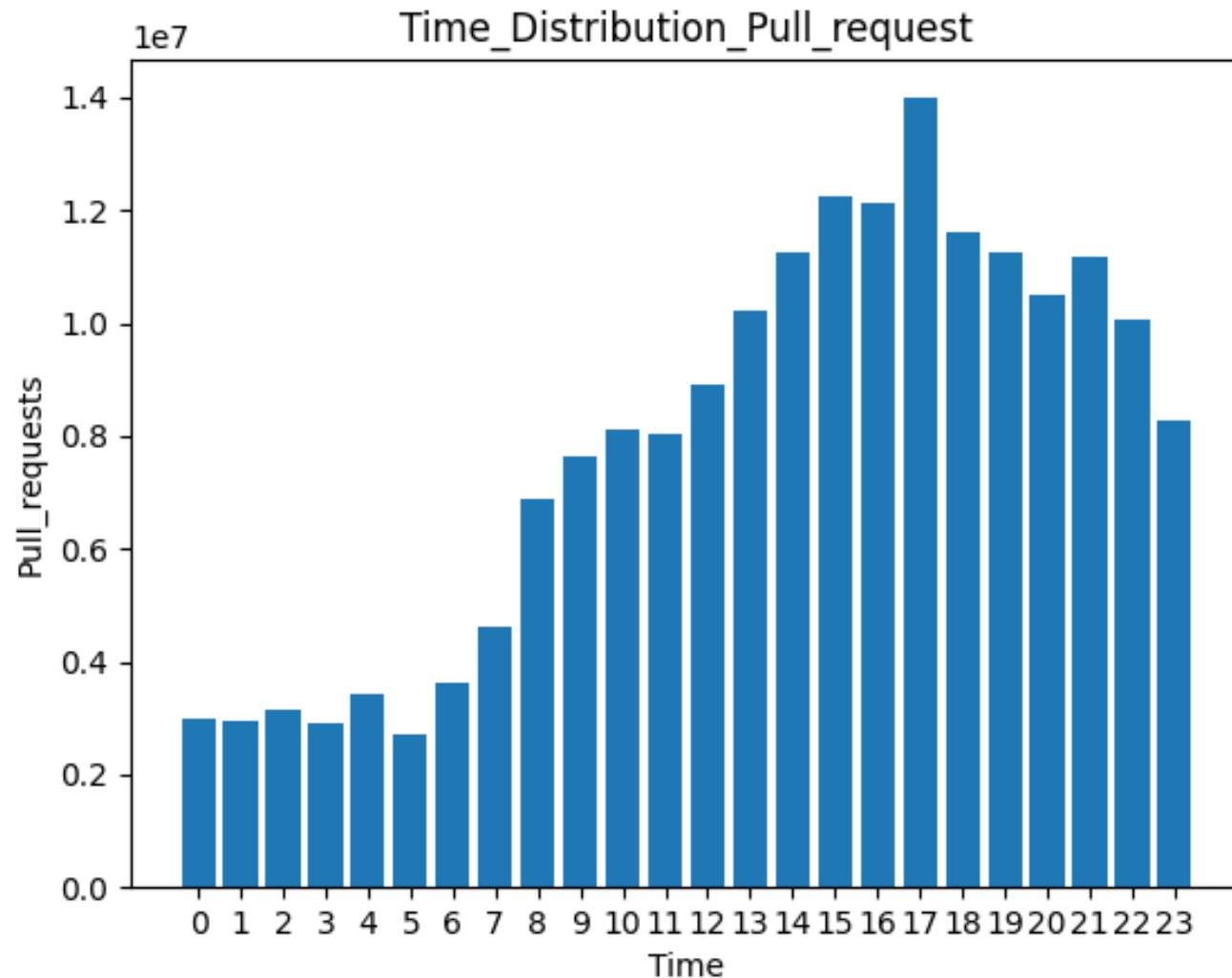
Tasks

- 1. Analyze pull request frequency throughout the day
- 2. Find most active actor and average comment length of individual actors
- 3. Most popular language
- 4. Most popular repo (num of pull requests and contributors)
- 5. Average comment length

1. Analyze pull request frequency throughout the day

```
✓ def commit_time_analysis(df):  
    df2 = df.select("commit_date").withColumn('Hour_of_day_UTC', F.substring("commit_date",12,2))  
    df3 = df2.select("Hour_of_day_UTC").groupBy("Hour_of_day_UTC").count()  
    df3.write.csv("hdfs://master:9000/output/commit_time_analysis.csv")  
    df3.show()
```

1. Analyze pull request frequency throughout the day



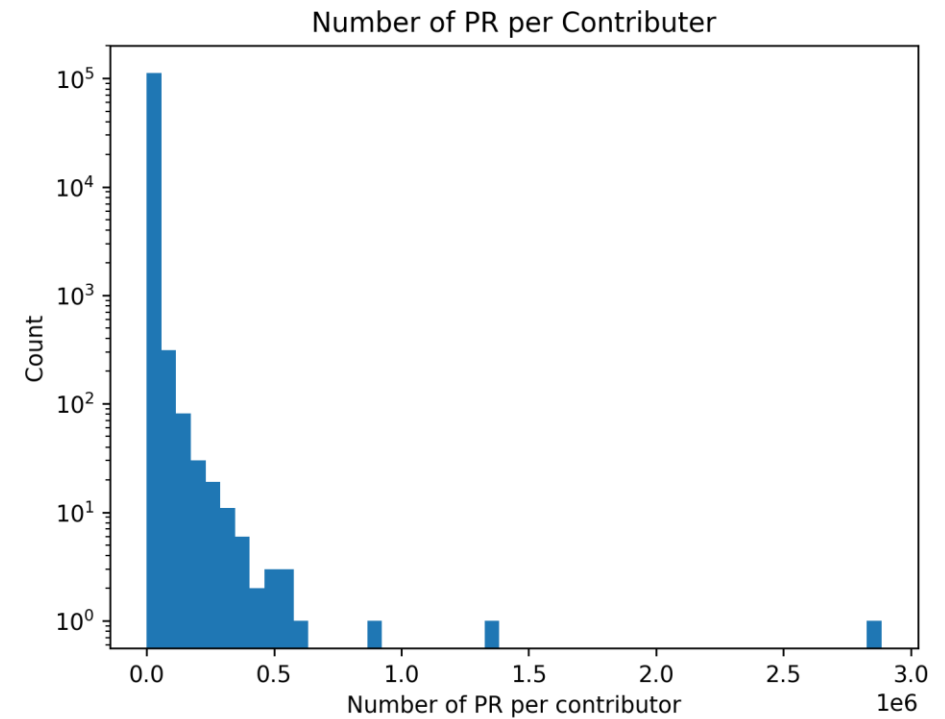
2. Most active actor and average comment length of actors

```
def mostActiveContributors(df):  
    contributorsCommentLength = df.withColumn('length_of_comment', F.length("comment"))  
    c = contributorsCommentLength.groupby(['actor_id', 'actor_login']).agg(F.count('actor_id'), F.mean('length_of_comment'))  
    orderedContributer = c.sort(col("count(actor_id)").desc(), col("avg(length_of_comment)").asc())  
    #print(orderedContributer.show())  
    orderedContributer.write.csv("hdfs://master:9000/output/Most_active_actors.csv")  
    averageCommits(orderedContributer)
```

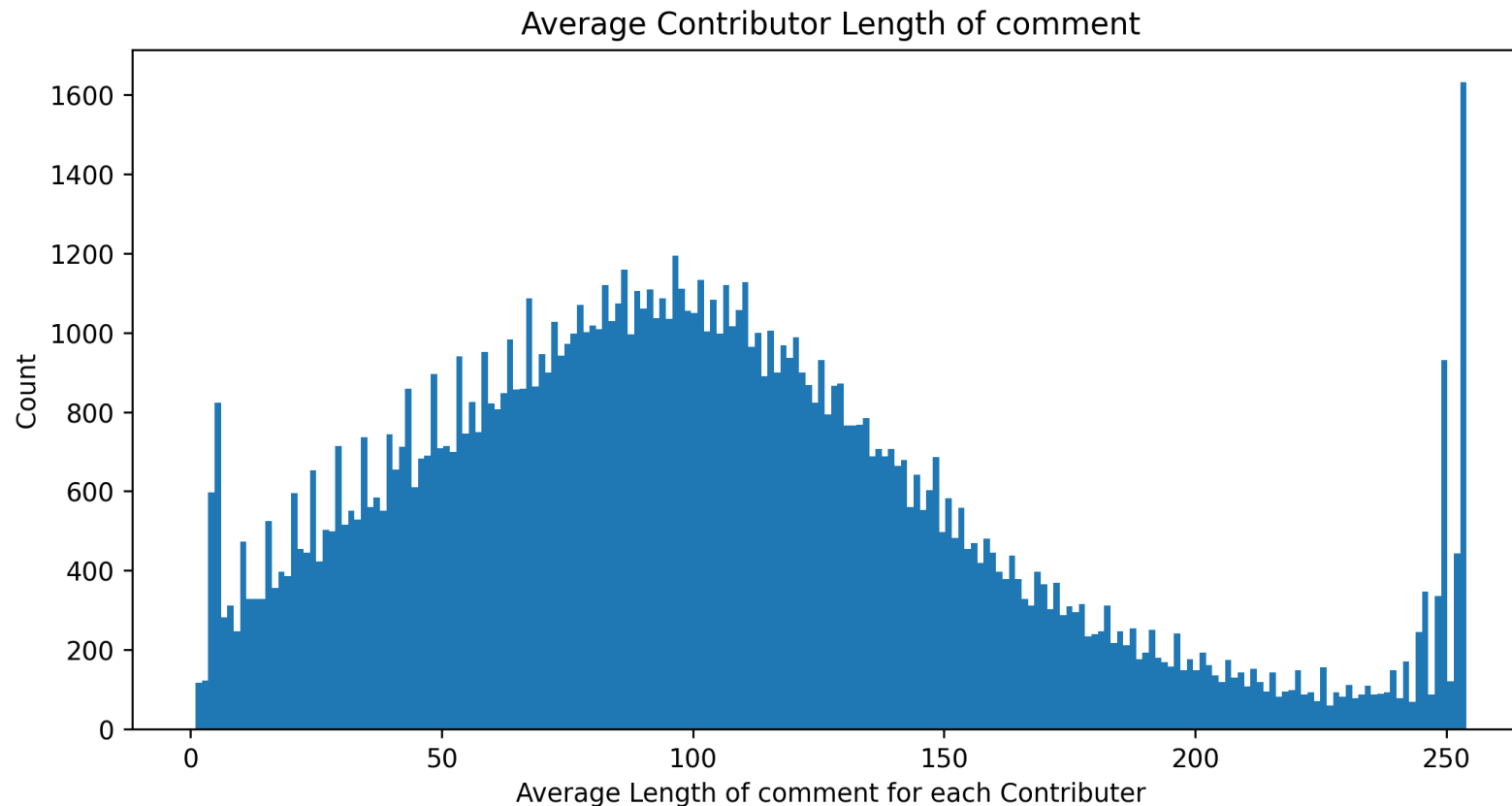
2. Most active actor

	actor_id	actor_login	count	length_of_comment
0	11568779	houndci-bot	2884918	66.316176
1	1435621	MartinHjelmare	1368852	81.186590
2	16780017	codacy-bot	913176	201.223267
3	718011	jreback	585524	68.693642
4	1816524	seanlip	571536	116.972222
...
112888	26609735	sragia	2	254.000000
112889	49831207	ChronicPwnage	2	254.000000
112890	3730222	obicke	2	254.000000
112891	456496	skagedal	2	254.000000
112892	2126894	xipmix	2	254.000000

[112893 rows x 4 columns]



2. Average comment length of actors



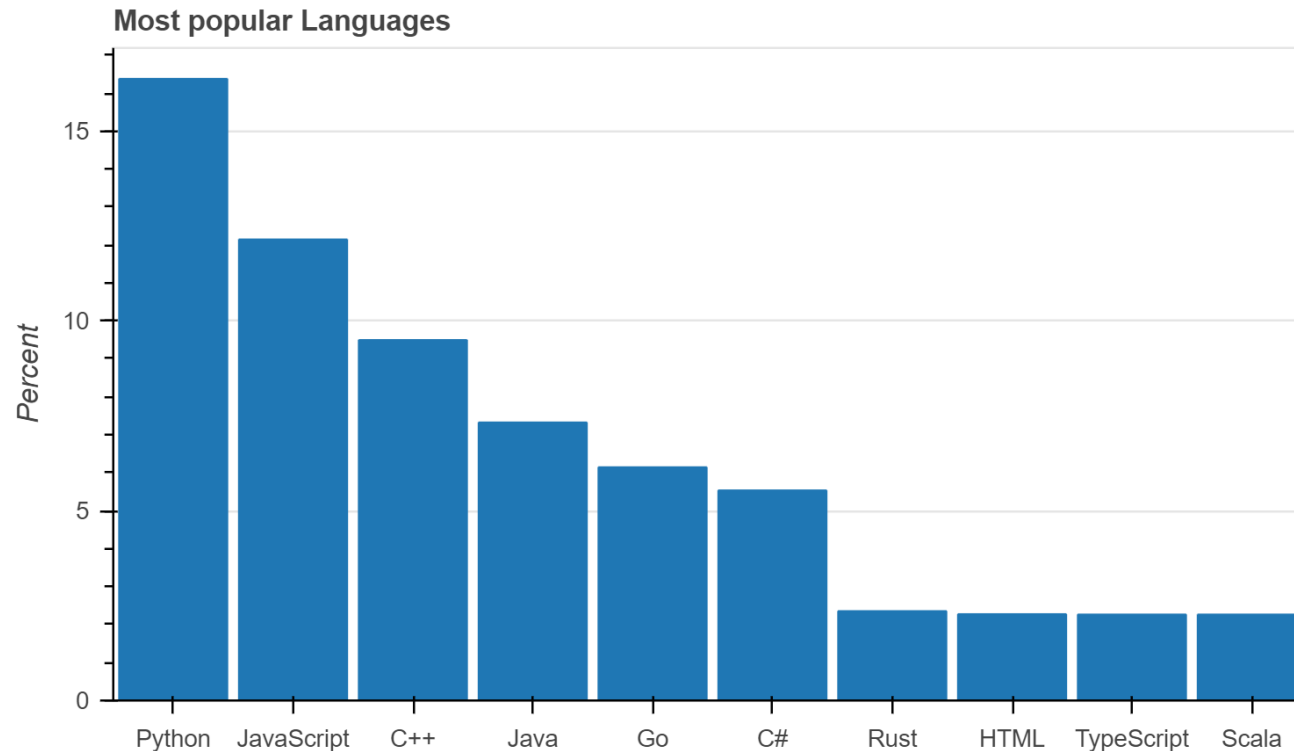
3. Average comment length

- 103.4

```
✓ def averageCommits(df):  
    average = df.agg(F.mean('count(actor_id)'), F.mean('avg(length_of_comment)'))  
    #print(average.show())  
    average.write.csv("hdfs://master:9000/output/average_actor_commits.csv")
```

4. Most popular language

```
def mostCommonLanguages(df):  
    languages = df.groupby(['language']).count()  
    orderedlanguages = languages.sort(col("count").desc())  
    orderedlanguages.write.csv("hdfs://master:9000/output/Most_common_languages.csv")
```

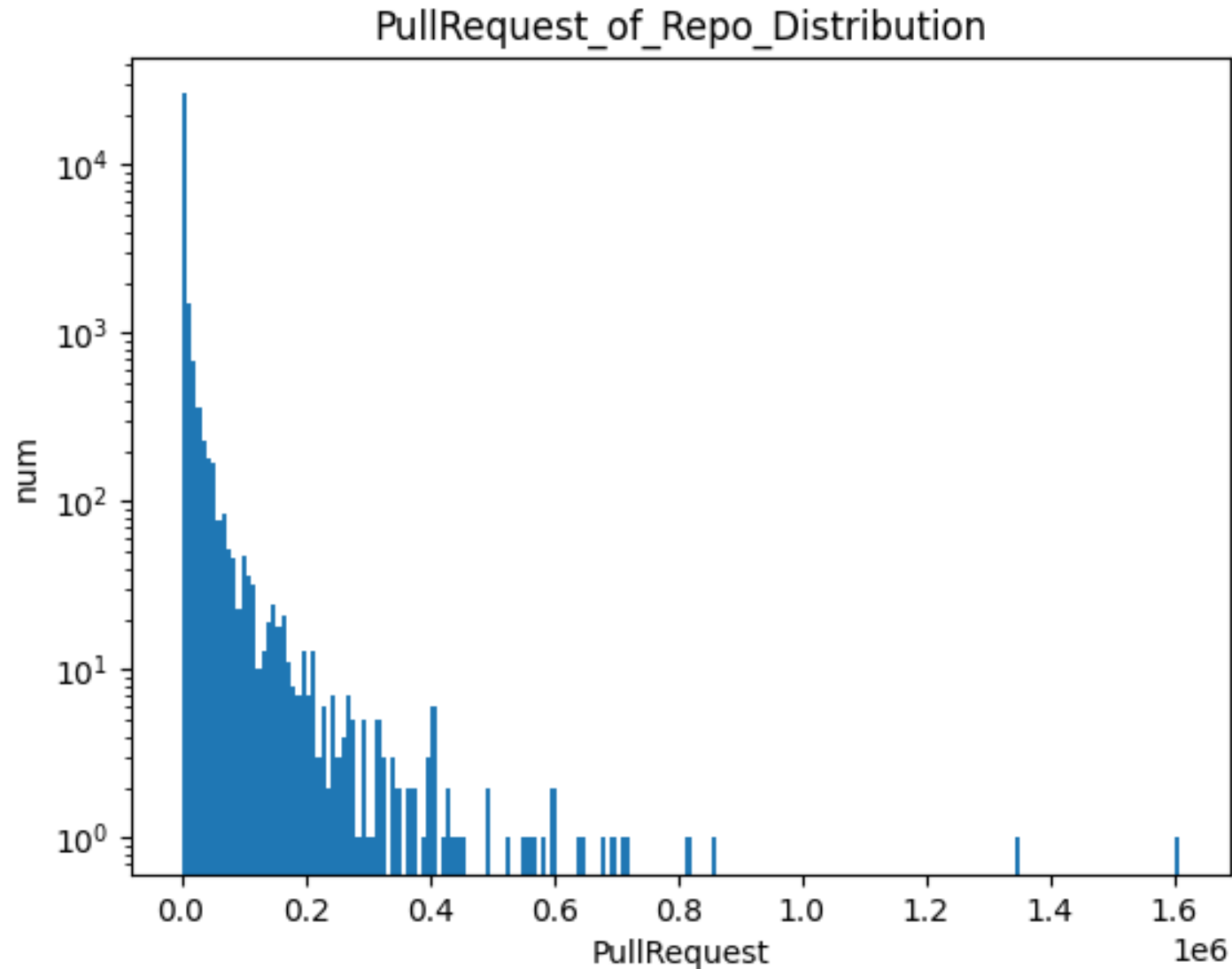


5. Most popular repo (num of pull requests and contributors)

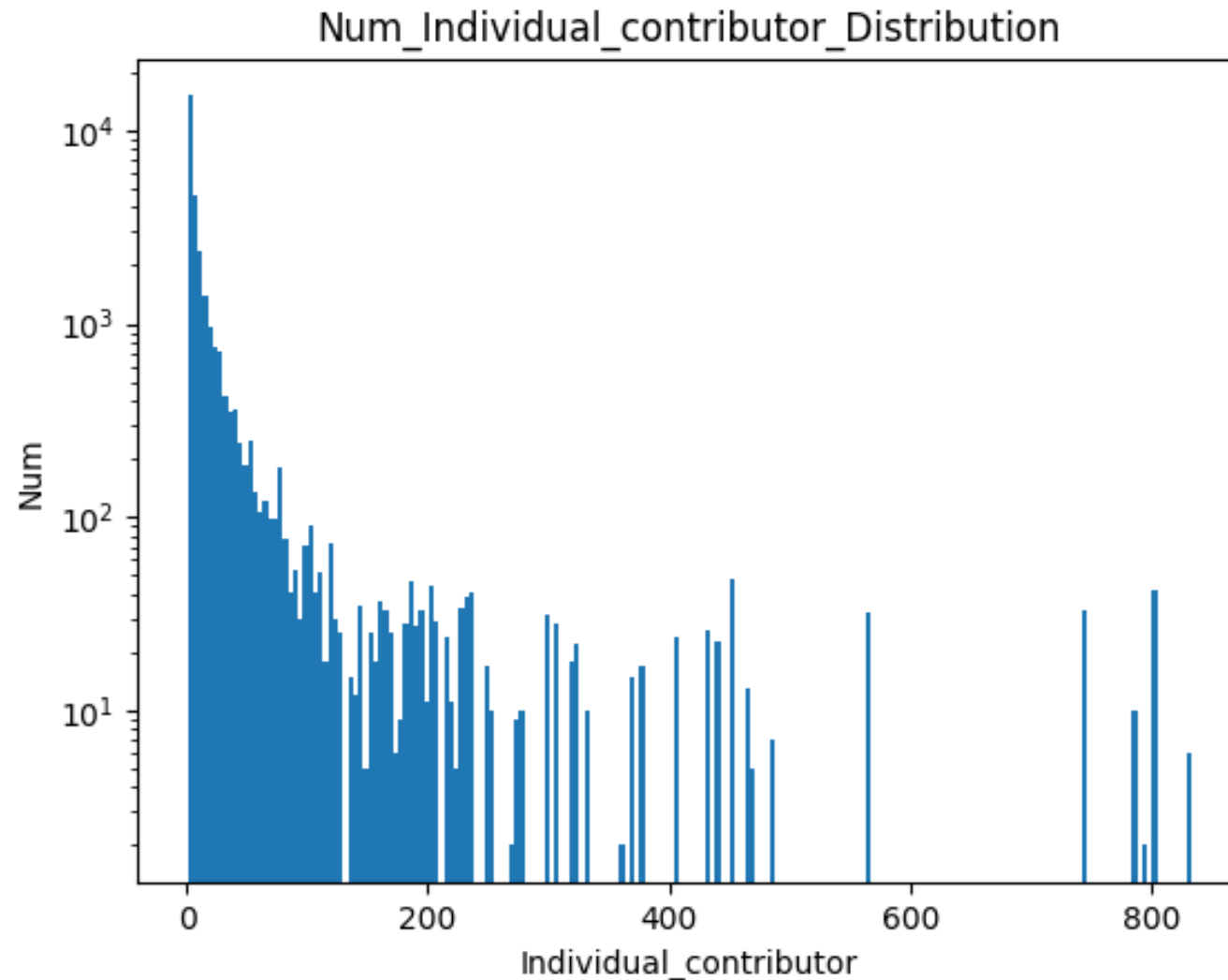
```
def repos(df):
    df1 = df.withColumn('length_of_comment', F.length("comment"))
    df2 = df1.withColumn('Hour_of_day_UTC', F.substring("commit_date", 12, 2))
    r = df2.groupby(['repo', 'author_login', 'author_id'])\
        .agg(F.countDistinct('actor_id').alias('Distinct_Contributors'),
             F.count('pr_id').alias('Pull_Requests'),
             F.mean('length_of_comment'),
             F.first('language'),
             F.percentile_approx("Hour_of_day_UTC", 0.5).alias("median_hour_of_day"))
    rOrdered = r.sort(col("Pull_Requests").desc())
    rOrdered.write.csv("Repo_PR.csv")

if __name__ == '__main__':
```

5. Most popular repo (num of pull requests)



5. Most popular repo (num of contributors)



5257	
5258	
5259	
5260	
5261	egonelbre
5262	
5263	
5264	
5265	
5266	
5267	
5268	
5269	
5270	
5271	
5272	
5273	
5274	
5275	
5276	
5277	
5278	
5279	janetzki
5280	
5281	
5282	
5283	

```
| 131542| Gargol| 1|
```

```
>>> fdf=df.filter(df.author_id=="janetzki")
>>> fdf.show()
```

actor_login	actor_id	comment_id	comment	repo	language	author_login	author_id	pr_id	c_id	commit_date
lanice	1554776	216090108	"I know that "pq... but I still woul...	hyrise	C++	janetzki	39369435	45379269	1126310152	
mrzzzrm	963349	222835159	""The first `_bi... doesn't it?"	hyrise	C++	janetzki	39369435	45626777	1155396180	
mrks	1266579	287940759	"I believe we cha... how about you?"	hyrise	C++	janetzki	39369435	55408549	1390896848	
mrzzzrm	963349	288168465	"Cocky answer: ""... because they are..."	hyrise	C++	janetzki	39369435	55408549	1390896848	
mrzzzrm	963349	249424286	"Right now it is ... right?"	hyrise	C++	janetzki	39369435	53896752	1276436347	
mrzzzrm	963349	277178795	"hopefully better... btw."	hyrise	C++	janetzki	39369435	58517864	1369014978	
nilsthamm	12824244	198388028	"My understanding... only invalid_arg..."	hyrise	C++	janetzki	39369435	40799302	1042355086	
mrzzzrm	963349	288166651	"There is no such... not one per bin"	hyrise	C++	janetzki	39369435	55408549	1390896848	
mrzzzrm	963349	232635139	"almost, I'd put ... but that's my pr..."	hyrise	C++	janetzki	39369435	48890888	1196201608	
mrks	1266579	287822713	"For me, ""column... but I find \"\"co..."	hyrise	C++	janetzki	39369435	55408549	1390896848	
mrks	1266579	220870291	""upper bound""... see the STL algo..."	hyrise	C++	janetzki	39369435	45626777	1146478587	
mrzzzrm	963349	266210152	"you could also d... which imo reads ..."	hyrise	C++	janetzki	39369435	55007050	1356648920	
mrks	1266579	220867306	"Does an estimate... can't we combine..."	hyrise	C++	janetzki	39369435	45626777	1146478587	
Benskl	512178	241708092	"I agree. Jenkins... coverage files a..."	hyrise	C++	janetzki	39369435	51385621	1232080057	
mrks	1266579	231210939	""... so in that... simply return th..."	hyrise	C++	janetzki	39369435	49105832	1193125012	
mrks	1266579	231210939	""... so in that... simply return th..."	hyrise	C++	janetzki	39369435	49105832	1193125012	
mrzzzrm	963349	225065097	"> The SQLTransla... so the SQLTransl..."	hyrise	C++	janetzki	39369435	47406661	1167238994	
mrzzzrm	963349	288166651	"There is no such... not one per bin"	hyrise	C++	janetzki	39369435	55408549	1390896848	
mrks	1266579	280357396	"We don't need to... you need to xyz\..."	hyrise	C++	janetzki	39369435	60334194	1372762660	
mrzzzrm	963349	227384191	"Anything but ""S... and a comment th..."	hyrise	C++	janetzki	39369435	47987675	1174538757	

```
only showing top 20 rows
```

- `spark.read.csv(r“hdfs://master:9000/ghtorrent-2019-03-11.csv”)`



- `spark.read.csv(r"hdfs://master:9000/ghtorrent-2019-03-11.csv", escape = '“')`

```
Command Prompt - docker e X + v
| 131542| Gargol| 1|
+-----+-----+
>>> fdf=df.filter(df.author_id=="janetzki")
>>> fdf.show()
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|actor_login|actor_id|comment_id|comment|repo|language|author_login|author_id|pr_id|c_id|commit_date|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|lanice|1554776|216090108|"I know that ""pq...|but I still woul...|hyrise|C++|janetzki|39369435|45379269|1126310152|
|mrzzzrm|963349|222835159|""The first `_bi...|doesn't it?"|hyrise|C++|janetzki|39369435|45626777|1155396180|
|mrks|1266579|287940759|"I believe we cha...|how about you?"|hyrise|C++|janetzki|39369435|55408549|1390896848|
|mrzzzrm|963349|288168465|"Cocky answer: ""...|because they are...|hyrise|C++|janetzki|39369435|55408549|1390896848|
|mrzzzrm|963349|249424286|"Right now it is ...|right?"|hyrise|C++|janetzki|39369435|53896752|1276436347|
|mrzzzrm|963349|277178795|"hopefully better...|btw."|hyrise|C++|janetzki|39369435|58517864|1369014978|
|nilsthamm|12824244|198388028|"My understanding...|only invalid_arg...|hyrise|C++|janetzki|39369435|40799302|1042355086|
|mrzzzrm|963349|288166651|"There is no such...|not one per bin"|hyrise|C++|janetzki|39369435|55408549|1390896848|
|mrzzzrm|963349|232635139|"almost, I'd put ...|but that's my pr...|hyrise|C++|janetzki|39369435|48890888|1196201608|
|mrks|1266579|287822713|"For me, ""column...|but I find \"""co...|hyrise|C++|janetzki|39369435|55408549|1390896848|
|mrks|1266579|220870291|""upper bound\"""...|see the STL algo...|hyrise|C++|janetzki|39369435|45626777|1146478587|
|mrzzzrm|963349|266210152|"you could also d...|which imo reads ...|hyrise|C++|janetzki|39369435|55007050|1356648920|
|mrks|1266579|220867306|"Does an estimate...|can't we combine...|hyrise|C++|janetzki|39369435|45626777|1146478587|
|Benski|512178|241708092|"I agree. Jenkins...|coverage files a...|hyrise|C++|janetzki|39369435|51385621|1232080057|
|mrks|1266579|231210939|""... so in that...|simply return th...|hyrise|C++|janetzki|39369435|49105832|1193125012|
|mrks|1266579|231210939|""... so in that...|simply return th...|hyrise|C++|janetzki|39369435|49105832|1193125012|
|mrzzzrm|963349|225065097|>"> The SQLTransla...|so the SQLTransl...|hyrise|C++|janetzki|39369435|47406661|1167238994|
|mrzzzrm|963349|288166651|"There is no such...|not one per bin"|hyrise|C++|janetzki|39369435|55408549|1390896848|
|mrks|1266579|280357396|"We don't need to...|you need to xyz\...|hyrise|C++|janetzki|39369435|60334194|1372762660|
|mrzzzrm|963349|227384191|"Anything but ""S...|and a comment th...|hyrise|C++|janetzki|39369435|47987675|1174538757|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

```
SyntaxError: EOL while scanning string literal
>>> df = spark.read.csv(r"file:///opt/share/ghtorrent-2019-03-11.csv", escape = '').toDF('actor_login','actor_
guage','author_login','author_id','pr_id','c_id','commit_date')
>>> fdf=df.filter(df.author_id=="janetzki")
>>> fdf.show()
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|actor_login|actor_id|comment_id|comment|repo|language|author_login|author_id|pr_id|c_id|commit_date|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```


End

- Thank you for listening.