



Bài báo cáo: Lưu trữ, xử lý và phân tích dữ liệu lớn user log activity

Huy Mac Quang¹, Khanh Ho Quoc²

,Ly Tran Thi Cam³, Hang Vu Thi⁴

¹Students of Hanoi University of Science and Technology (Student Code: 20173169)

²Students of Hanoi University of Science and Technology (Student Code: 20173191)

³Students of Hanoi University of Science and Technology (Student Code: 20173250)

⁴Students of Hanoi University of Science and Technology (Student Code: 20173096)

Lưu trữ và xử lý dữ liệu lớn – IT4931 (Class code: 118628)

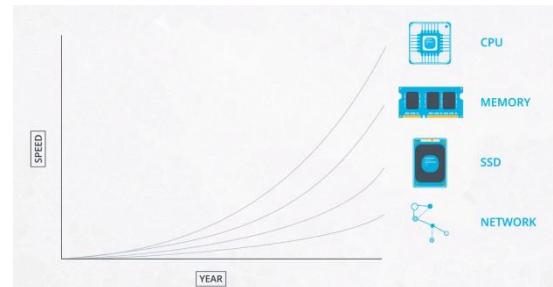
ABSTRACT

<https://github.com/huymq1710/HDFS-MultiNode>

<https://github.com/huymq1710/Spark-cluster-yarn-mode>

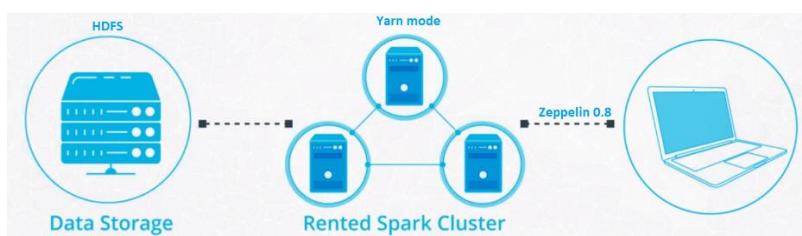
Với 400 users 8GB dữ liệu, có thể lưu và xử lý trong 1 máy nhưng khi lớn hơn. Có thể dùng 12 co-workers laptops, split data cho các máy qua mail, xử lý và mail lại kết quả nhưng “network is suck”. Bài báo cáo môn học sẽ lựa trữ dữ liệu trên cụm HDFS, quản lý tài nguyên bằng Yarn và dùng Spark Yarn mode và Zeppelin cho việc interpreter vào dữ liệu trên cụm để xử lý. Tất cả đều được cài đặt trên cụm kết nối với nhau qua mạng Lan.

Keywords: Hadoop, Yarn, Spark, Zeppelin, Bigdata



I. INTRODUCTION

Project cuối kì sẽ kết hợp các bài thực hành phía trước để phục vụ cho việc phân tích và xử lý dữ liệu lần này. Sơ đồ của toàn bộ project:



File log lưu lại user activity của một ứng dụng nghe nhạc, từ tháng 1/2018 đến 10/2018 mỗi tháng ~2,3GB như sau:



<https://github.com/huymq1710/Spark-cluster-yarn-mode/tree/main/Pyspark%20in%20JupyterNotebook>

II. Các phần mềm đã cài đặt trên toàn cụm

Install Hadoop and Yarn Cluster: <https://github.com/huymq1710/HDFS-MultiNode>

2.1 Hadoop Distributed File System (HDFS)

```
hdfs dfsadmin -report
```

```
hadoopuser@hadoop-master:~
```

```
hadoopuser@hadoop-master:~$ jps
9555 ZeppelinServer
8483 SecondaryNameNode
8202 NameNode
8932 jps
8733 ResourceManager
9078 HistoryServer

hadoopuser@hadoop-master:~$ hdfs dfsadmin -report
Configured Capacity: 2189557296 (203.89 GB)
Present Capacity: 1609768000 (149.60 GB)
DFS Remaining: 12092212736 (112.03 GB)
DFS Used: 39891495333 (37.15 GB)
DFS Used%: 24.98%
Replicated blocks: 0
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0
Non-redundant blocks (with replication factor > 1): 0
Low redundancy blocks with highest priority to recover: 0
Pending deletion blocks: 0
Erasure Coded Block Groups:
  Under replicated block groups: 0
  Blocks with corrupt internal blocks: 0
  Missing block groups: 0
  Non-redundant blocks (with replication factor > 1): 0
  Low redundancy blocks with highest priority to recover: 0
  Pending deletion blocks: 0
-----
Live datanodes (3):
Name: 192.168.100.15:9866 (hadoop-slave2)
Hostname: hadoop-slave2
Decommission Status : Normal
Configured Capacity: 61951623200 (57.70 GB)
DFS Used: 16909610400 (16.00 GB)
DFS Remaining: 45086572864 (15.00 GB)
DFS Used%: 25765974016 (24.00 GB)
DFS Used%: 27.29%
DFS Remaining%: 33.59%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xcursors: 1
Last contact: Sat Dec 26 10:01:41 ICT 2020
Last Block Report: Sat Dec 26 09:55:53 ICT 2020
Num of Blocks: 146

Name: 192.168.100.4:9860 (hadoop-slave1)
Hostname: hadoop-slave1
Decommission Status : Normal
Configured Capacity: 19549730968 (18.21 GB)
DFS Used: 6784432549 (16.32 GB)
DFS Remaining: 11015540519 (10.26 GB)
DFS Used%: 3480537375 (18.00 %)
```

```
hadoopuser@hadoop-master: ~
```

DFS Used: 16904487816 (5.74 GB)
Non DFS Used: 16162972864 (15.66 GB)
DFS Remaining: 25765974016 (24.60 GB)
DFS Used%: 27.29%
DFS Remaining%: 41.59%
Configured Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xclevvers: 1
Last contact: Sat Dec 26 10:01:41 ICT 2020
Last Block Report: Sat Dec 26 09:55:53 ICT 2020
Num of Blocks: 146

Name: 192.168.100.4:9866 (hadoop-slave1)
Hostname: hadoop-slave1
Decommission Status : Normal
Configured Capacity: 19549730968 (18.21 GB)
DFS Used: 6784432549 (6.32 GB)
Non DFS Used: 11915580555 (10.26 GB)
DFS Remaining: 11915580555 (10.26 GB)
DFS Used%: 34.70%
DFS Remaining%: 3.75%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xclevvers: 1
Last contact: Sat Dec 26 10:01:44 ICT 2020
Last Block Report: Sat Dec 26 09:55:59 ICT 2020
Num of Blocks: 52

Name: 192.168.100.9:9866 (hadoop-slave3)
Hostname: hadoop-slave3
Decommission Status : Normal
Configured Capacity: 203761041264 (127.99 GB)
DFS Used: 15262537668 (15.49 GB)
Non DFS Used: 20376104960 (18.98 GB)
DFS Remaining: 93793181696 (87.35 GB)
DFS Used%: 11.79%
DFS Remaining%: 68.25%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xclevvers: 1
Last contact: Sat Dec 26 10:01:41 ICT 2020
Last Block Report: Sat Dec 26 09:55:53 ICT 2020
Num of Blocks: 146

- Data log file trong HDFS



Browsing HDFS - Chromium

How to Install and Set Up... x Browsing HDFS x All Applications x History Server x Zeppelin x +

Not secure | hadoop-master@10.0.0.10:8080/explorer.html#/hadoop/log/file

Hadoop Overview Databases DataNode Volume Features Snapshot Status Progress Utilities

Browse Directory

#hadoop-log_file

Show 25 entries

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	-rw-r--r--	hadoop.root	supergroup	2.26 GB	Dec 23 02:56	2	128 MB	012018.json
□	-r--r--r--	hadoopuser	supergroup	2.26 GB	Dec 23 03:21	2	128 MB	022018.json
□	-rw-r--r--	hadoopuser	supergroup	2.26 GB	Dec 23 03:48	2	128 MB	032018.json
□	-rw-r--r--	hadoopuser	supergroup	2.26 GB	Dec 23 04:11	2	128 MB	042018.json
□	-rw-r--r--	hadoopuser	supergroup	2.26 GB	Dec 23 04:34	2	128 MB	052018.json
□	-rw-r--r--	hadoopuser	supergroup	2.26 GB	Dec 23 04:55	2	128 MB	062018.json
□	-rw-r--r--	hadoopuser	supergroup	2.26 GB	Dec 23 05:30	2	128 MB	082018.json
□	-rw-r--r--	hadoopuser	supergroup	2.26 GB	Dec 23 06:01	2	128 MB	102018.json

Showing 1 to 8 of 8 entries

Previous Next

2.2 Yarn web UI

The screenshot shows a dual-monitor setup. The left monitor displays the Apache Hadoop Cluster Management interface, which includes sections for Cluster Metrics, Active Nodes, and Resource Manager. The right monitor shows a terminal window titled 'hadoop@ip-10-0-0-102:~\$' running on a Red Hat Enterprise Linux host. The terminal output shows the configuration of a YARN ResourceManager, including the setting of 'yarn.resourcemanager.address' to '0.0.0.0:8032'.

2.3 Spark cluster Yarn mode

<https://www.linode.com/docs/databases/hadoop/install-configure-run-spark-on-top-of-hadoop-yarn-cluster/>

- Web UI ghi lại các bản log của các job

History Server - Chromium							
App ID		App Name	Started	Completed	Duration	Spark User	Last Updated
application_1608624248152_0004		Zeppelin	2020-12-22 16:47:56	2020-12-22 17:38:36	51 min	hadoopuser	2020-12-22 17:38:36
application_1608624248152_0003		Zeppelin	2020-12-22 16:36:01	2020-12-22 16:38:53	2.9 min	hadoopuser	2020-12-22 16:38:54
application_1608624248152_0002		Zeppelin	2020-12-22 16:08:31	2020-12-22 16:14:58	6.4 min	hadoopuser	2020-12-22 16:14:59
application_1608624248152_0001		Zeppelin	2020-12-22 15:18:32	2020-12-22 15:51:16	33 min	hadoopuser	2020-12-22 15:51:16
application_1608190739084_0001		Spark shell	2020-12-17 14:45:43	2020-12-17 15:26:42	41 min	hadoopuser	2020-12-17 15:26:42
application_1607840721049_0002		Spark PI	2020-12-13 13:45:53	2020-12-13 13:49:40	3.8 min	hadoopuser	2020-12-13 13:49:41
application_1607840721049_0001		Spark PI	2020-12-13 13:34:41	2020-12-13 13:44:21	9.7 min	hadoopuser	2020-12-13 13:44:21
application_1607784198618_0002		Spark shell	2020-12-12 22:25:07	2020-12-13 00:32:08	2.1 h	hadoopuser	2020-12-13 00:32:09
application_1607784198618_0001		Spark shell	2020-12-12 22:08:00	2020-12-12 22:24:23	16 min	hadoopuser	2020-12-12 22:44:2
application_1607764923745_0002		Spark PI	2020-12-12 18:22:22	2020-12-12 18:22:55	33 s	hadoopuser	2020-12-12 18:22:55



2.4 Zeppelin 0.8

<https://medium.com/@shehanfernando/zeppelin-on-spark-cluster-6732d8eb1b59>

2.5 Jps trên các node

Master:

```
hadoopuser@hadoop-master:~$ jps
9555 ZeppelinServer
8483 SecondaryNameNode
8262 NameNode
9579 Jps
8733 ResourceManager
9070 HistoryServer
```

Slave:

```
hadoopuser@hadoop-slave3:~$ jps
14203 NodeManager
14012 DataNode
14460 Jps
```

2.6 Trường hợp khi 1 node đầy bộ nhớ

Node đó sẽ bị xếp thành unhealthy nodes



III. Thực thi job

3.1 Jps khi job đang được submit

- Master:

```
hadoopuser@hadoop-master:~$ jps
17058 HistoryServer
16515 SecondaryNameNode
18054 SparkSubmit
16246 NameNode
18215 Jps
17143 ZeppelinServer
16731 ResourceManager
```

- Slave 1:

```
hadoopuser@hadoop-slave1:~$ jps
3745 Jps
3349 NodeManager
3174 DataNode
3703 CoarseGrainedExecutorBackend
```

- Slave 2:

```
hadoopuser@hadoop-slave2:~$ jps
8082 CoarseGrainedExecutorBackend
8152 Jps
7657 DataNode
7822 NodeManager
```

- Slave 3:

```
hadoopuser@hadoop-slave3:~$ jps
2633 ExecutorLauncher
2378 NodeManager
2205 DataNode
3869 Jps
```

3.2 Check Yarn web UI khi job đang chạy

ID	User	Name	Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Reserved CPU Vcores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking URL
application_1608963254491_0002	hadoopuser	Huy-van-nhung-nguoi-ban	SPARK	default	0	Sat Dec 26 13:37:19 +0700 2020	Sat Dec 26 13:37:20 +0700 2020	N/A	RUNNING	UNDEFINED	3	3	3712	0	0	80.6	80.6	Application	



IV. Code trên Zeppelin

Ngôn ngữ: Scala 2.4.7

4.1 Load và clean dataset

```
final-project-IT4931 - Zeppelin - Chromium
Zeppelin Notebook Job final-project-IT4931 [ ] Head [ ] Search anonymous [ ] default [ ]
```

final-project-IT4931 [] Head [] Search anonymous [] default []

```
| sc.version
res1: String = 2.4.7

Took 4 min 42 sec. Last updated by anonymous at December 26 2020, 1:37:41 PM.
```

I. Load and clean dataset:

In this notebook, will load and clean log data user activity, handling of invalid or missing values: 01/2018 -> 03/2018. Total: 7.3GB

```
| val event_data = "hdfs:///hadoop/log_file/*.json"
val df = sqlContext.read.json(event_data)
df.head(5)

event_data: String = hdfs:///hadoop/log_file/*.json
df: org.apache.spark.sql.DataFrame = [event_id: string ... 15 more fields]
res2: Array[org.apache.spark.sql.Row] = Array(Martin.Orford Logged In,Joseph,M,29,Morales,597,55057,Free,Corpus Christi,TX,PUT,NextSong,1532063507000,292,Grand Designs,200,1838352011000,"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/36.0.1985.125 Safari/537.36",293),(John Brown,He,Logged In,Sueper,M,74,Larson,380,21179,Free,Houston-The Woodlands-Sugar Land,TX,PUT,NextSong,1538069638000,97,Bulls,280,1838352025000,"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/36.0.1985.143 S...)
```

Took 4 min 51 sec. Last updated by anonymous at December 26 2020, 1:42:09 PM.

```
| df.printSchema()
root
|- artist: string (nullable = true)
|- auth: string (nullable = true)
|- firstName: string (nullable = true)
|- gender: string (nullable = true)
|- lastGeoipCity: string (nullable = true)
|- lastGeoipCountry: string (nullable = true)
|- length: double (nullable = true)
|- level: string (nullable = true)
|- location: string (nullable = true)
|- method: string (nullable = true)
|- page: string (nullable = true)
|- registration: long (nullable = true)
|- sessionId: long (nullable = true)
|- song: string (nullable = true)
|- status: string (nullable = true)
|- ts: long (nullable = true)
|- userAgent: string (nullable = true)
|- userId: string (nullable = true)

Took 1 sec. Last updated by anonymous at December 26 2020, 1:46:31 PM.
```

Statistics of whole dataset

```
final-project-IT4931 - Zeppelin - Chromium
Zeppelin Notebook Job final-project-IT4931 [ ] Head [ ] Search anonymous [ ] default [ ]
```

final-project-IT4931 [] Head [] Search anonymous [] default []

```
| df.describe().show()
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| summary | artist | auth|firstName| gender| lteInSession|lastName| length| level| location| method| page| registration|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| count | 12986310 | 16311150 | 15840160 | 15401600 | 16311150 | 16311150 | 15840160 | 16311150 | 16311150 | 16311150 | 16311150 |
| mean | 529537212209 | null |
| stddev | 164.9783482342488 | null |
| min | null |
| max | null |
| +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

Took 5 min 51 sec. Last updated by anonymous at December 26 2020, 1:52:25 PM.

Statistics of the #artist# column

```
final-project-IT4931 - Zeppelin - Chromium
Zeppelin Notebook Job final-project-IT4931 [ ] Head [ ] Search anonymous [ ] default [ ]
```

final-project-IT4931 [] Head [] Search anonymous [] default []

```
| df.describe("artist").show()
+-----+-----+
| summary | artist |
+-----+-----+
| count | 12986310 |
| mean | 52.529537212209 |
| stddev | 164.9783482342488 |
| min | null |
| max | null |
| +-----+-----+
```

Took 41 sec. Last updated by anonymous at December 26 2020, 1:53:12 PM.

Statistics of the sessionId column

```
final-project-IT4931 - Zeppelin - Chromium
Zeppelin Notebook Job final-project-IT4931 [ ] Head [ ] Search anonymous [ ] default [ ]
```

final-project-IT4931 [] Head [] Search anonymous [] default []

```
| df.describe("sessionId").show()
+-----+-----+
| summary | sessionId |
+-----+-----+
| count | 16311150 |
| mean | 2048.854553771772 |
| stddev | 1434.3376560041656 |
| min | 1 |
| max | 4000 |
| +-----+-----+
```



final-project-IT4931 - Zeppelin - Chromium

Zeppelin Notebook Job

final-project-IT4931

Statistics of the userId column

```
| df.describe("userId").show()
```

```
+-----+-----+
|summary|userId|
+-----+-----+
| count| 16311150|
| mean| 69268.42669193512|
| stddev|189898.72264083923|
| min|    99|
| max| 16311150|
+-----+
```

Took 36 sec. Last updated by anonymous at December 26 2020, 1:58:01 PM.

Total rows: 16,311,150

```
| df.count()
```

```
res8: Long = 16311150
```

Took 27 sec. Last updated by anonymous at December 26 2020, 1:58:34 PM.

All the page events in the dataset:

```
- About
- Add Friend
- Add to Playlist
- Cancel
- ### Cancellation Confirmation ###
- Downgrade
- Error
- Help
- Home
- Logout
- Logout
- NextSong
- Register
- Roll Advert
- Session Settings
- Settings
- Submit Downgrade
- Submit Registration
- Submit Upgrade
- Thumbs Down
- Thumbs Up
- Userinfo
```

final-project-IT4931 - Zeppelin - Chromium

Zeppelin Notebook Job

final-project-IT4931

```
| df.select("page").dropDuplicates().sort("page").show()
```

```
+-----+
| page|
+-----+
| About|
| Add Friend|
| Add to Playlist|
| Cancel|
| ### Cancellation Confirmation ###
| Downgrade|
| Error|
| Help|
| Home|
| Logout|
| Logout|
| NextSong|
| Register|
| Roll Advert|
| Userinfo|
+-----+
```

Took 35 sec. Last updated by anonymous at December 26 2020, 2:00:34 PM.

Missing values

```
| import org.apache.spark.sql.functions...
df.columns
```

```
| import org.apache.spark.sql.functions...
res11: Array[String] = Array(artist, auth, firstName, gender, ltenInSession, lastName, length, level, location, method, page, registration, sessionId, song, status, ts, userAgent, userId)
```

Took 1 sec. Last updated by anonymous at December 26 2020, 2:00:47 PM.

Check how many missing values in each column

```
| print("missing values")|
for col <- df.columns {
  var missing_count = df.filter(df(col).isNull || df(col) === "") || df(col).isNa().count()
  print(f" ${col} : {missing_count}")
}
missing values
artist: 3324840
auth: 0
firstName: 4718000
```



final-project-IT4931 - Zeppelin - Chromium

Zeppelin Notebook + Job Search anonymous default

final-project-IT4931

Check how many missing values in each column

```
print("["missing values"]")
for col in df.columns:
    var missing_count = df.filter(df[col].isNull || df[col] === "") || df[col].isNa).count()
    print(col)
    print(missing_count)
    print("["missing values"]
artist: 3324840
auth: 471000
firstname: 471000
gender: 471000
itemInSession: 0
lastname: 471000
length: 3324840
level: 0
location: 471000
method: 0
page: 0
registration: 471000
sessionId: 0
song: 3324840
status: 0
ts: 0
userAgent: 471000
rows have been removed")
```

Took 13 min 11 sec. Last updated by anonymous at December 26 2020, 2:14:21 PM.

Check user and sessionid

If the below Ids are null or empty, delete those rows:

```
userId
sessionId
```

```
val df_without_null = df.na.drop("any")
val df_without_missing_id = df.filter(df._without_null("userId") != "") // 'userId' should not be empty string
print("["df: " + df.count())
print("["df._without_null: " + df._without_null.count())
print("["df._without_missing_id: " + df._without_missing_id.count())
print("["df._without_missing_id._without_null: " + df._without_missing_id._without_null.count())
print("["df._without_missing_id._without_missing_id: " + df._without_missing_id._without_missing_id.count()) + " rows have been removed")
```

df: 16311159
df._without_null: 12986310
df._without_missing_id: 15840150
471000 rows have been removed

final-project-IT4931 - Zeppelin - Chromium

final-project-IT4931 - Zeppelin - Chromium

Check user and sessionid

If the below Ids are null or empty, delete those rows:

```
userId
sessionId
```

```
val df_without_null = df.na.drop("any")
val df_without_missing_id = df.filter(df._without_null("userId") != "") // 'userId' should not be empty string
print("["df: " + df.count())
print("["df._without_null: " + df._without_null.count())
print("["df._without_missing_id: " + df._without_missing_id.count())
print("["df._without_missing_id._without_null: " + df._without_missing_id._without_null.count())
print("["df._without_missing_id._without_missing_id: " + df._without_missing_id._without_missing_id.count()) + " rows have been removed")
```

df: 16311159
df._without_null: 12986310
df._without_missing_id: 15840150
471000 rows have been removed
df._without_null: org.apache.spark.sql.DataFrame = [artist: string, auth: string ... 16 more fields]
df._without_missing_id: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [artist: string, auth: string ... 16 more fields]

Took 3 min 1 sec. Last updated by anonymous at December 26 2020, 2:17:26 PM.

II. Exploratory Data Analysis

Detect number columns and category columns.

```
num_cols: Number columns (long or double)
cat_cols: Category columns (String)
```

```
var num_cols = new Array[String](6)
var cat_cols = new Array[String](12)
var num_count = 0
var cat_count = 0
for (s <= df.schema()){
    if (s.dataType() == LongType || s.dataType() == DoubleType){
        num_cols(num_count) = s.name
        num_count += 1
    } else if (s.dataType() == StringType){
        cat_cols(cat_count) = s.name
        cat_count += 1
    }
}
num_cols: Array[String] = Array(itemInSession, length, registration, sessionId, status, ts)
cat_cols: Array[String] = Array(artist, auth, firstname, gender, lastname, level, location, method, page, song, userAgent, userId)
cat_count: Int = 12
num_count: Int = 6
```

Took 2 sec. Last updated by anonymous at December 26 2020, 2:17:33 PM.

```
num_cols
result: Array[String] = Array(itemInSession, length, registration, sessionId, status, ts)
```



4.2 Exploratory data analysis

final-project-IT4931-Zep: x +

Not secure | hadoop-master:9010/#/notebook/2FV82FQEQ FINISHED ▶ ✎ ⓘ

```
| num_cols
res5$: Array[String] = Array(itemInSession, length, registration, sessionId, status, ts)

Took 0 sec. Last updated by anonymous at December 26 2020, 2:17:42 PM.
```

```
| cat cols
res6$: Array[String] = Array(artist, auth, firstName, gender, lastName, level, location, method, page, song, userAgent, userId)

Took 0 sec. Last updated by anonymous at December 26 2020, 2:17:51 PM.
```

Number columns READY ▶ ✎ ⓘ

```
| for(s <- num_cols){
|   dfWithoutMissingId.describe(s).show()
| }
| dfWithoutMissingId.show()
| +-----+-----+
| | count| 15840159|
| | mean| 1.53552341e+07|
| | stdDev| 3.078726747445429|
| | min| 1599854193000|
| | max| 1543073874000|
| +-----+-----+
| summary| registration|
| +-----+-----+
| | count| 15840159|
| | mean| 1.53552341e+07|
| | stdDev| 3.078726747445429|
| | min| 1599854193000|
| | max| 1543073874000|
| +-----+-----+
| summary| sessionId|
| +-----+-----+
```

Took 3 min 13 sec. Last updated by anonymous at December 26 2020, 2:21:10 PM.

There are three HTTP status codes: READY ▶ ✎ ⓘ

- 307: Temporary Redirect
- 404: Not Found
- 200: OK

```
| dfWithoutMissingId.select("status").dropDuplicates().show()
+-----+
| status|
+-----+
| 307|
| 404|
| 200|
```

SPARK JOB FINISHED ▶ ✎ ⓘ



```
final-project-IT4931 - Zeppelin - Chromium
```

Final project IT4931 - Zeppelin - Chromium

SPARK JOB FINISHED

```
| df_without_missing_id.select("method").dropDuplicates().show()
```

```
+-----+  
| method  
+-----+  
| PUT|  
| GET|  
+-----+
```

Took 36 sec. Last updated by anonymous at December 26 2020, 2:26:54 PM.

```
| df_without_missing_id.select("page").dropDuplicates().show()
```

```
+-----+  
| page  
+-----+  
| Thumbs Down|  
| Home|  
| Downgrade|  
| Roll Advert|  
| Logout|  
| Save Settings|  
| Cancellation Conf|  
| About|  
| Settings|  
| Add to Playlist|  
| Add Friend|  
| NextSong|  
| Thumbs Up|  
| Help|  
| Upgrade|  
| Error|  
| Submit Upgrade|  
+-----+
```

Took 36 sec. Last updated by anonymous at December 26 2020, 2:27:45 PM.

```
| df_without_missing_id.select("userAgent").dropDuplicates().show(10)
```

```
+-----+  
| userAgent  
+-----+  
| "Mozilla/5.0 (Mac...)|  
| "Mozilla/5.0 (Win...)|  
| "Mozilla/5.0 (X11; ...)|  
| "Mozilla/5.0 (Mac...)|  
| "Mozilla/5.0 (Wind...)|  
| "Mozilla/5.0 (Wind...)|  
| "Mozilla/5.0 (comp...)|  
| "Mozilla/5.0 (Win...)|  
+-----+  
only showing top 10 rows
```

Took 37 sec. Last updated by anonymous at December 26 2020, 2:28:36 PM.

```
final-project-IT4931 - Zeppelin - Chromium
```

Final project IT4931 - Zeppelin - Chromium

SPARK JOB FINISHED

```
| df_without_missing_id.filter("page = 'Cancellation Confirmation'").show(10)
```

```
+-----+  
| artist| auth| firstName| gender| itemInSession| lastName| length| level| location| method| page| registration| sessionId| song| status| ts| userAgent| userId|  
+-----+  
| null| [Cancelled]| Olivia| F| 40| Carr| null| free| Fort Wayne, IN| GET| [Cancellation Conf...| 1536738429000| 496|null| 200| [0398400616000| Mozilla/5.0 (Wind...)| 208|  
| null| [Cancelled]| Lillian| F| 234| Cameron| null| paid| Columbus, OH| GET| [Cancellation Conf...| 153947270000| 471|null| 200| [039848279000| Mozilla/5.0 (Wind...)| 231|  
| null| [Cancelled]| Alex| M| 108| Myers| null| paid| Grand Rapids-Hope...| GET| [Cancellation Conf...| 153995579000| 682|null| 200| [0398539817000| Mozilla/5.0 (Wind...)| 236|  
| null| [Cancelled]| Brad| M| 60| Stewart| null| paid| Roanoke, VA| GET| [Cancellation Conf...| 1535618904000| 395|null| 200| [0398770077000| Mozilla/5.0 (Wind...)| 159|  
| null| [Cancelled]| Isaac| M| 193| Miller| null| paid| Seattle-Tacoma-B...| GET| [Cancellation Conf...| 153620182000| 862|null| 200| [039880533000| Mozilla/5.0 (Wind...)| 208|  
| null| [Cancelled]| Ivan| M| 9| Sanchez| null| free| Bridgeport-Stanfo...| GET| [Cancellation Conf...| 153537526000| 1827|null| 200| [0398890595000| Mozilla/5.0 (Wind...)| 138|  
| null| [Cancelled]| Chloe| M| 56| Myers| null| free| Tampa-St. Petersbu...| GET| [Cancellation Conf...| 1534370883000| 480|null| 200| [039890508000| Mozilla/5.0 (Ph...)| 123|  
| null| [Cancelled]| Alexi| F| 7| Warren| null| paid| Spokane-Spokane V...| GET| [Cancellation Conf...| 1532482662000| 1819|null| 200| [039897586000| Mozilla/5.0 (Wind...)| 54|  
| null| [Cancelled]| Peyton| F| 82| Campbell| null| paid| Los Angeles-Long ...| GET| [Cancellation Conf...| 152902754000| 1866|null| 200| [039811456000| Mozilla/5.0 (Wind...)| 39|  
+-----+  
only showing top 10 rows
```

Took 2 sec. Last updated by anonymous at December 26 2020, 2:28:43 PM.

```
| val df_churned = df_without_missing_id.withColumn("churned", when(col("page") === "Cancellation Confirmation", 1).otherwise(0))  
| df_churned.filter("churned = 1").show(10)
```

```
+-----+  
| artist| auth| firstName| gender| itemInSession| lastName| length| level| location| method| page| registration| sessionId| song| status| ts| userAgent| userId|churned|  
+-----+  
| null| [Cancelled]| Olivia| F| 40| Carr| null| free| Fort Wayne, IN| GET| [Cancellation Conf...| 1536738429000| 496|null| 200| [0398400616000| Mozilla/5.0 (Wind...)| 208| 1|  
| null| [Cancelled]| Lillian| F| 234| Cameron| null| paid| Columbus, OH| GET| [Cancellation Conf...| 153947270000| 471|null| 200| [039848279000| Mozilla/5.0 (Wind...)| 231| 1|  
| null| [Cancelled]| Alex| M| 108| Myers| null| paid| Grand Rapids-Hope...| GET| [Cancellation Conf...| 153995579000| 682|null| 200| [0398539817000| Mozilla/5.0 (Wind...)| 236| 1|  
| null| [Cancelled]| Brad| M| 60| Stewart| null| paid| Roanoke, VA| GET| [Cancellation Conf...| 1535618904000| 395|null| 200| [0398770077000| Mozilla/5.0 (Wind...)| 159| 1|  
| null| [Cancelled]| Isaac| M| 193| Miller| null| paid| Seattle-Tacoma-B...| GET| [Cancellation Conf...| 153620182000| 862|null| 200| [039880533000| Mozilla/5.0 (Wind...)| 208| 1|  
| null| [Cancelled]| Ivan| M| 9| Sanchez| null| free| Bridgeport-Stanfo...| GET| [Cancellation Conf...| 153537526000| 1827|null| 200| [0398890595000| Mozilla/5.0 (Wind...)| 138| 1|  
| null| [Cancelled]| Chloe| M| 56| Myers| null| free| Tampa-St. Petersbu...| GET| [Cancellation Conf...| 1534370883000| 480|null| 200| [039890508000| Mozilla/5.0 (Ph...)| 123| 1|  
| null| [Cancelled]| Alexi| F| 7| Warren| null| paid| Spokane-Spokane V...| GET| [Cancellation Conf...| 1532482662000| 1819|null| 200| [039897586000| Mozilla/5.0 (Wind...)| 54| 1|  
| null| [Cancelled]| Peyton| F| 82| Campbell| null| paid| Los Angeles-Long ...| GET| [Cancellation Conf...| 152902754000| 1866|null| 200| [039811456000| Mozilla/5.0 (Wind...)| 39| 1|  
+-----+  
only showing top 10 rows
```

Took 2 sec. Last updated by anonymous at December 26 2020, 2:28:51 PM.

Define Churn

SPARK JOB FINISHED

```
| // churn: Cancellation Confirmation
```

Took 1 sec. Last updated by anonymous at December 26 2020, 3:01:16 PM. (submitted)

```
| val total_cancelConfirms = df_churned.filter("churned = 1").count()  
val total_Users = df_churned.groupby("userId").sum("churned").count()
```

SPARK JOB FINISHED



```
final-project-IT4931 - Zeppelin - Chromium
final-project-IT4931 - Zeppelin - Chromium
[1]: val df_churned = df.without_missing_id.withColumn("churned", when(col("page") === "Cancellation Confirmation", 1).otherwise(0))
df_churned.filter("churned = 1").show(10)

+-----+-----+-----+-----+-----+-----+-----+-----+
| auth|firstName|gender|id|lastName|lengthLevel|location|method|
+-----+-----+-----+-----+-----+-----+-----+-----+
| null|Cancelled|Olivia|F|234|Cameron|null|paid|Port Wayne, IN|GET[Cancellation Conf...|1530758409000|400|null|200|1038806510000|Mozilla/5.0 (Win...)|1| |
| null|Cancelled|Lillian|F|234|Cameron|null|paid|Columbus, OH|GET[Cancellation Conf...|1533472700000|471|null|200|1038482792000|Mozilla/5.0 (Win...)|223|1|
| null|Cancelled|Alex|M|109|Myers|null|paid|Grand Rapids-Hy...|GET[Cancellation Conf...|1529955790000|682|null|200|1038539871000|Mozilla/5.0 (Win...)|236|1|
| null|Cancelled|Rafael|M|66|Crawford|null|free|Bowling Green, KY|GET[Cancellation Conf...|1533876200000|628|null|200|1038592529000|Mozilla/5.0 (Mac...)|156|1|
| null|Cancelled|Brad|M|227|Stewart|null|paid|Roanoke, VA|GET[Cancellation Conf...|1535619904000|395|null|200|1038770077000|Mozilla/5.0 (Win...)|271|1|
| null|Cancelled|Isaac|M|193|Miller|null|paid|Seattle-Tacoma-Be...|GET[Cancellation Conf...|1538201820000|862|null|200|103880531000|Mozilla/5.0 (Win...)|208|1|
| null|Cancelled|Vivian|M|193|Miller|null|paid|Seattle-Tacoma-Be...|GET[Cancellation Conf...|1538201820000|862|null|200|103880531000|Mozilla/5.0 (Win...)|208|1|
| null|Cancelled|Ollie|M|56|Myers|null|free|Tampa-St. Petersb...|GET[Cancellation Conf...|1534270883000|480|null|200|1038959001000|Mozilla/5.0 (VPh...)|123|1|
| null|Cancelled|Alexi|M|7|Warren|null|paid|Spokane-Spokane V...|GET[Cancellation Conf...|1532482662000|1830|null|200|1038987586000|Mozilla/5.0 (Win...)|54|1|
| null|Cancelled|Payton|F|82|Campbell|null|paid|Los Angeles-Los ...|GET[Cancellation Conf...|1529827541000|1066|null|200|1039011456000|Mozilla/5.0 (Win...)|39|1|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 10 rows

df_churned: org.apache.spark.sql.DataFrame = [auth: string, auth: string ... 17 more fields]

Took 2 sec. Last updated by anonymous at December 26 2020, 2:28:43 PM.
```

Define Churn

```
// churn: Cancellation Confirmation
```

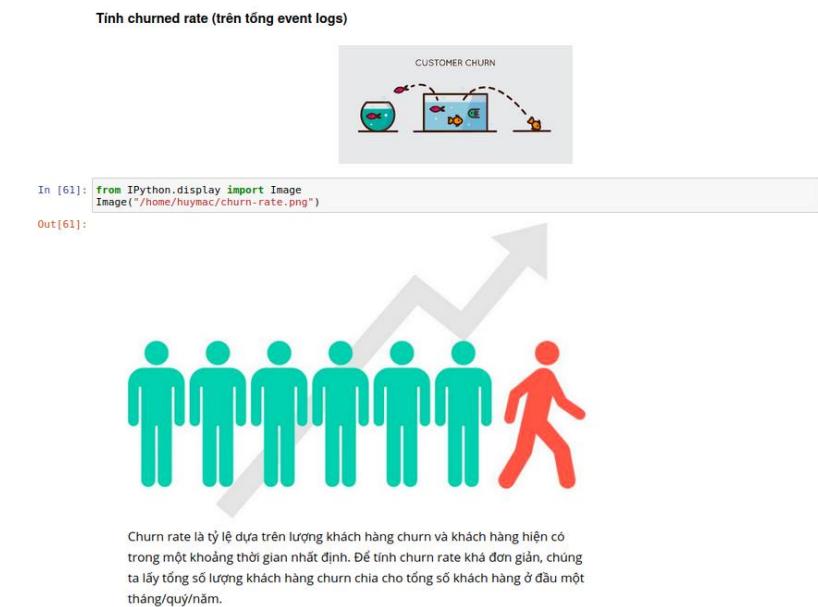
Took 1 sec. Last updated by anonymous at December 26 2020, 3:01:16 PM. (updated)

```
val total_cancelConfirms = df_churned.filter("churned = 1").count()
val total_Users = df_churned.groupBy("userId").sum("churned").count()
println(s"churned: ${100*total_cancelConfirms/(total_Users.toFloat)} %")
churned: 19.888302857142858 %
total_cancelConfirms: Long = 408
total_Users: Long = 448

Took 1 min 16 sec. Last updated by anonymous at December 26 2020, 2:36:16 PM. (updated)
```

```
spark.stop()
```

Took 1 sec. Last updated by anonymous at December 26 2020, 3:07:52 PM.



4.3 Khai phá dữ liệu sử dụng pyspark trên Jupyter notebook



III. Khai phá dữ liệu

Explore Data: So sánh giữa 2 nhóm churned users và not churned users => Có sự khác biệt nào giữa 2 nhóm này

Các cột sẽ được kiểm tra:

- artist: Số lượng artist
- gender: Convert về 0 và 1
- length: Tổng length
- Level: Convert về 0 và 1
- page: Số lượng Thumbs Up
- Số lượng Thumbs Down
- song: Số lượng bài hát

Convert churn(0 or 1) thành Not Churn or Churn

- Cả matplotlib + seaborn đều yêu cầu pandas dataframe(chứ không phải là pyspark dataframe) => Cần convert //Chuyển đổi từng phần nhỏ dữ liệu
(Tránh việc chuyển all dataset -> mất thời gian + bộ nhớ)

```
In [50]: func_churn_label = udf(lambda x: 'Churn' if x == 1 else 'Not Churn')
In [51]: df_churn_user = df_churn.groupby("userId").max("churn").withColumnRenamed("max(churn)", "churn").select(["userId", "churn"])
In [52]: pd_gender = df_churn.select(["userId", "gender", "churn"]).withColumn("churn", func_churn_label("churn")).toPandas()
pd_gender.head()
```

Out[52]:

userId	gender	churn
0	F	Churn
1	F	Churn
2	F	Churn
3	F	Churn
4	F	Churn

```
In [53]: sns.countplot(x="gender", hue="churn", data=pd_gender);
```

```
In [53]: sns.countplot(x="gender", hue="churn", data=pd_gender);
```

level

```
In [54]: pd_level = df_churn.select(["userId", "level", "churn"]).withColumn("churn", func_churn_label("churn")).toPandas()
pd_level.head()
```

Out[54]:

userId	level	churn
0	free	Churn
1	free	Churn
2	free	Churn
3	free	Churn
4	free	Churn

```
In [55]: sns.countplot(x="level", hue="churn", data=pd_level);
```



jupyter IT4931-FinalProject-GiuChungTa-huymq1710 Last Checkpoint: 18 hours ago (autosaved)

In [56]:

```
pd_song = df_churn_user.join(df_churn.groupby("userId") \
    .agg({"song": "count"}) \
    .withColumnRenamed("count","song_count"), ["userId"]) \
    .withColumn("churn", func_churn_label("churn")).toPandas()
```

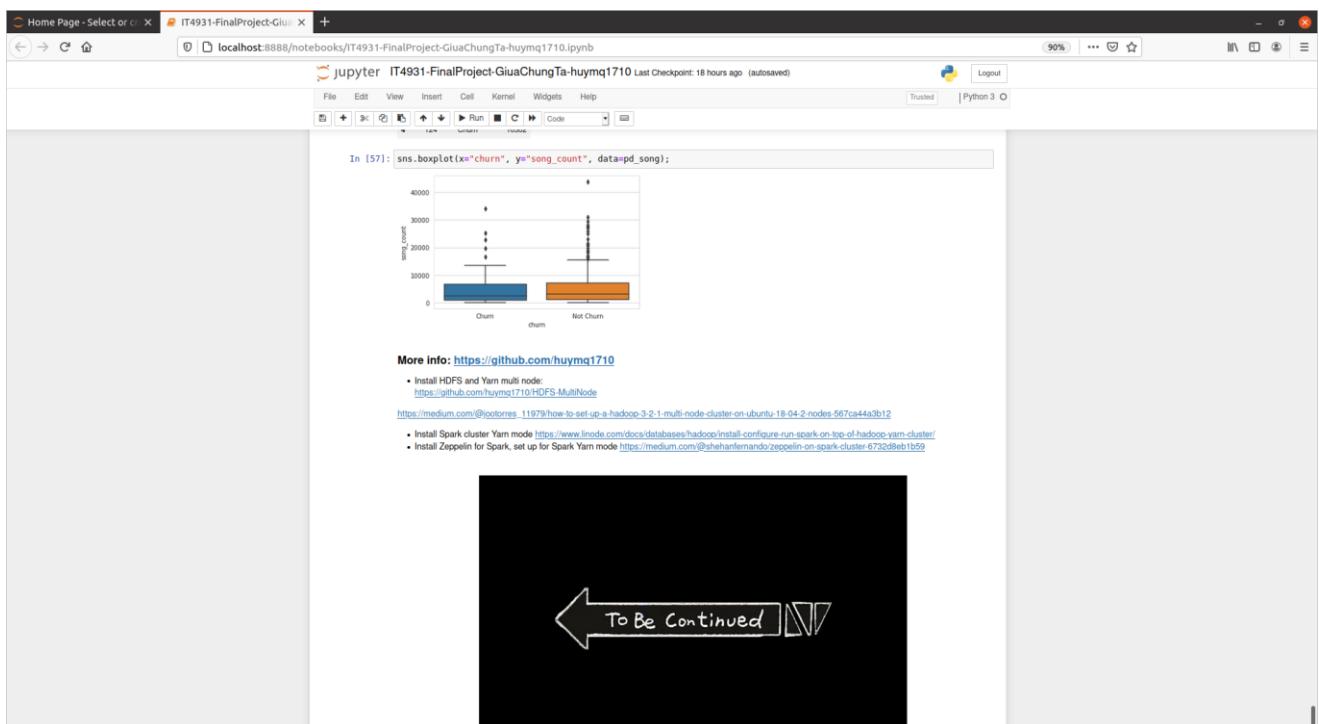
Out[56]:

userId	churn	song_count
0	Churn	480
1	Churn	1550
2	Churn	672
3	Not Churn	368
4	Churn	10302

In [57]: sns.boxplot(x="churn", y="song_count", data=pd_song);

More info: <https://github.com/huymq1710>

- Install HDFS and Yarn multi node:
https://github.com/huymq1710/HDFS_MultiNode
- https://medium.com/@yootrees_11979/how-to-set-up-a-hadoop-3-2-1-multi-node-cluster-on-ubuntu-18-04-2-nodes-567ca44a0b12
- Install Spark cluster Yarn mode <https://www.linode.com/docs/databases/hadoop/install-configure-run-spark-on-top-of-hadoop-yarn-cluster/>
- Install Zeppelin for Spark, set up for Spark Yarn mode <https://medium.com/@shehanfernando/zeppelin-on-spark-cluster-6732d8eb1b59>





4.4 Job done

History Server - Chromium

Chung (BigData) | Micro... | Browsing HDFS | Application application_1608963254491_0002 | History Server | Untitled Note - Zeppelin | Sparkify Project - Churn | Spark Interpreter Group | +

Not secure | hadoop-master:18080

Spark 2.4.7 History Server

Event log directory: <hdfs://hadoop-master:9000/spark-logs>
 Last updated: 2020-12-26 15:08:31
 Client local time zone: Asia/Saigon

Search:

App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
application_1608963254491_0002	Huy-va-nhung-nguo-i-ban	2020-12-26 13:33:21	2020-12-26 15:07:51	1.6 h	hadoopuser	2020-12-26 15:07:52	Download
application_1608963254491_0001	Huy-va-nhung-nguo-i-ban	2020-12-26 13:20:31	2020-12-26 13:31:47	11 min	hadoopuser	2020-12-26 13:31:48	Download
application_1608953423670_0001	Zeppelin	2020-12-26 10:39:31	2020-12-26 10:46:12	6.7 min	hadoopuser	2020-12-26 10:46:13	Download
application_1608624248152_0004	Zeppelin	2020-12-22 16:47:56	2020-12-22 17:38:36	51 min	hadoopuser	2020-12-22 17:38:36	Download
application_1608624248152_0003	Zeppelin	2020-12-22 16:36:01	2020-12-22 16:38:53	2.9 min	hadoopuser	2020-12-22 16:38:54	Download
application_1608624248152_0002	Zeppelin	2020-12-22 16:08:31	2020-12-22 16:14:58	6.4 min	hadoopuser	2020-12-22 16:14:59	Download
application_1608624248152_0001	Zeppelin	2020-12-22 15:18:32	2020-12-22 15:51:16	33 min	hadoopuser	2020-12-22 15:51:16	Download
application_1608190739084_0001	Spark shell	2020-12-17 14:45:43	2020-12-17 15:26:42	41 min	hadoopuser	2020-12-17 15:26:42	Download
application_1607840721049_0002	Spark Pi	2020-12-13 13:45:53	2020-12-13 13:49:40	3.8 min	hadoopuser	2020-12-13 13:49:41	Download
application_1607840721049_0001	Spark Pi	2020-12-13 13:34:41	2020-12-13 13:44:21	9.7 min	hadoopuser	2020-12-13 13:44:21	Download
application_1607784198818_0002	Spark shell	2020-12-12 22:25:07	2020-12-13 00:32:08	2.1 h	hadoopuser	2020-12-13 00:32:09	Download
application_1607784198818_0001	Spark shell	2020-12-12 22:08:00	2020-12-12 22:24:23	16 min	hadoopuser	2020-12-12 22:24:42	Download
application_1607764923745_0002	Spark Pi	2020-12-12 18:22:22	2020-12-12 18:22:55	33 s	hadoopuser	2020-12-12 18:22:55	Download

Showing 1 to 13 of 13 entries
[Show incomplete applications](#)

hadoop-master:18080/api/v1/applications/application_1608963254491_0002/logs

eventLogs-ap...zip [Show all](#)

Application application_1608963254491_0002 - Chromium

Chung (BigData) | Micro... | Browsing HDFS | Application application_1608963254491_0002 | History Server | Untitled Note - Zeppelin | Sparkify Project - Churn | Spark Interpreter Group | +

Not secure | hadoop-master:8088/cluster/app/application_1608963254491_0002

Logged in as: dr.who

hadoop

Application Overview

User: hadoopuser
 Name: Huy-va-nhung-nguo-i-ban
 Application Type: SPARK
 Application Tags:
 Application Priority: 0 (Higher Integer value indicates higher priority)
 YarnApplicationState: FINISHED
 Queue: default
 FinalStatus Reported by AM: SUCCEEDED
 Started: Sat Dec 26 13:37:19 +0700 2020
 Launched: Sat Dec 26 13:37:20 +0700 2020
 Finished: Sat Dec 26 15:07:52 +0700 2020
 Elapsed: 1hrs, 30mins, 32secs
 Tracking URL: [History](#)
 Log Aggregation Status: DISABLED
 Application Timeout (Remaining Time): Unlimited
 Diagnostics:
 Unmanaged Application: false
 Application Node Label expression: <Not set>
 AM container Node Label expression: <DEFAULT_PARTITION>

Application Metrics

Total Resource Preempted: <memory:0, vCores:0>
 Total Number of Non-AM Containers Preempted: 0
 Total Number of AM Containers Preempted: 0
 Resource Preempted from Current Attempt: <memory:0, vCores:0>
 Number of Non-AM Containers Preempted from Current Attempt: 0
 Aggregate Resource Allocation: 20145355 MB-seconds, 16282 vcore-seconds
 Aggregate Preempted Resource Allocation: 0 MB-seconds, 0 vcore-seconds

Show 20 entries

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
attempt_1608963254491_0002_000001	Sat Dec 26 13:37:19 +0700 2020	http://hadoop-slave1:8042	0	0	0

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

V. Kết luận

Bằng việc sử dụng Hadoop, việc lưu trữ, quản lý được quản lý một cách dễ dàng đồng thời có tính chịu lỗi cao khi xảy ra các trường hợp xấu như: node dead, lỗi mạng, out of storage,... Spark được xử lý và ghi trên Ram và được



Yarn quản lý tài nguyên, tăng tốc độ xử lý so với Hadoop . Zeppelin cung cấp một giao diện UI cho việc code và debug job phù hợp với việc báo cáo.

REFERENCES

- [1] Install-Hadoop-and-Yarn:<https://github.com/huymq1710/HDFS-MultiNode>
- [2] Install-Spark:<https://www.linode.com/docs/guides/install-configure-run-spark-on-top-of-hadoop-yarn-cluster/>
- [3] Data and code: <https://github.com/huymq1710/HDFS-MultiNode>
- [4] Hadoop: <https://hadoop.apache.org/>
- [5] Spark: <https://spark.apache.org/examples.html>

[View publication stats](#)