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

**Huy Mac Quang 1, Khanh Ho Quoc 2**

**,Ly Tran Thi Cam 3, Hang Vu Thi4**

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

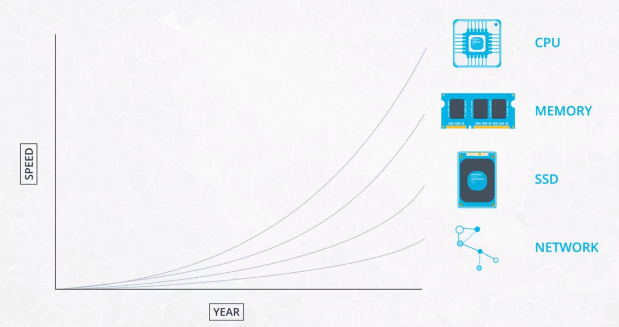
*2Students of Hanoi University of Science and Technology (Student Code: 20173191)*

*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

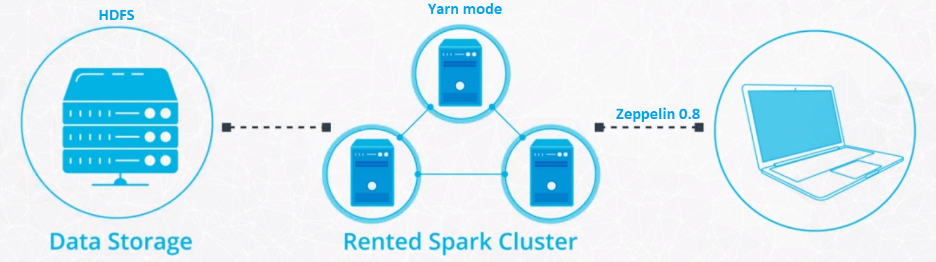


Với 400 users 8GB dữ liệu, có thể lưu và xử lý trong 1 máy nhưng khi lớn hơn. Có thể dùng 12 co-workers laptops, split data cho các máy qua mail, xử lý và mail lại kết quả nhưng “network is suck”. Bài báo cáo môn học sẽ lưa trữ dữ liệu trên cụm HDFS, quản lý tài nguyên bằng Yarn và dùng Spark Yarn mode 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.

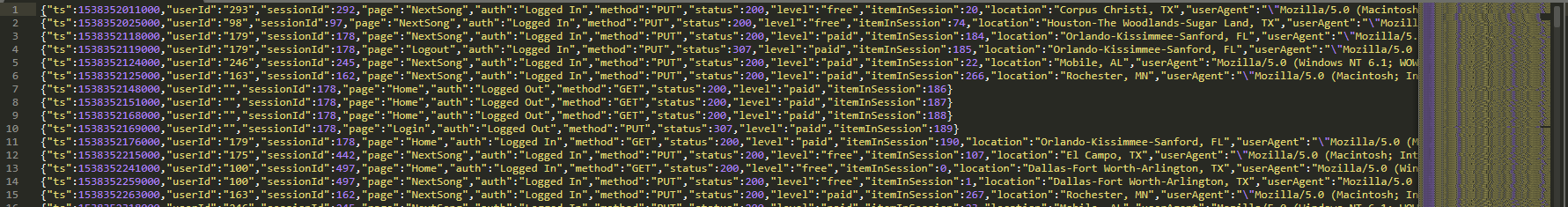
***Keywords: Hadoop, Yarn, Spark, Zeppelin, Bigdata***

# 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:



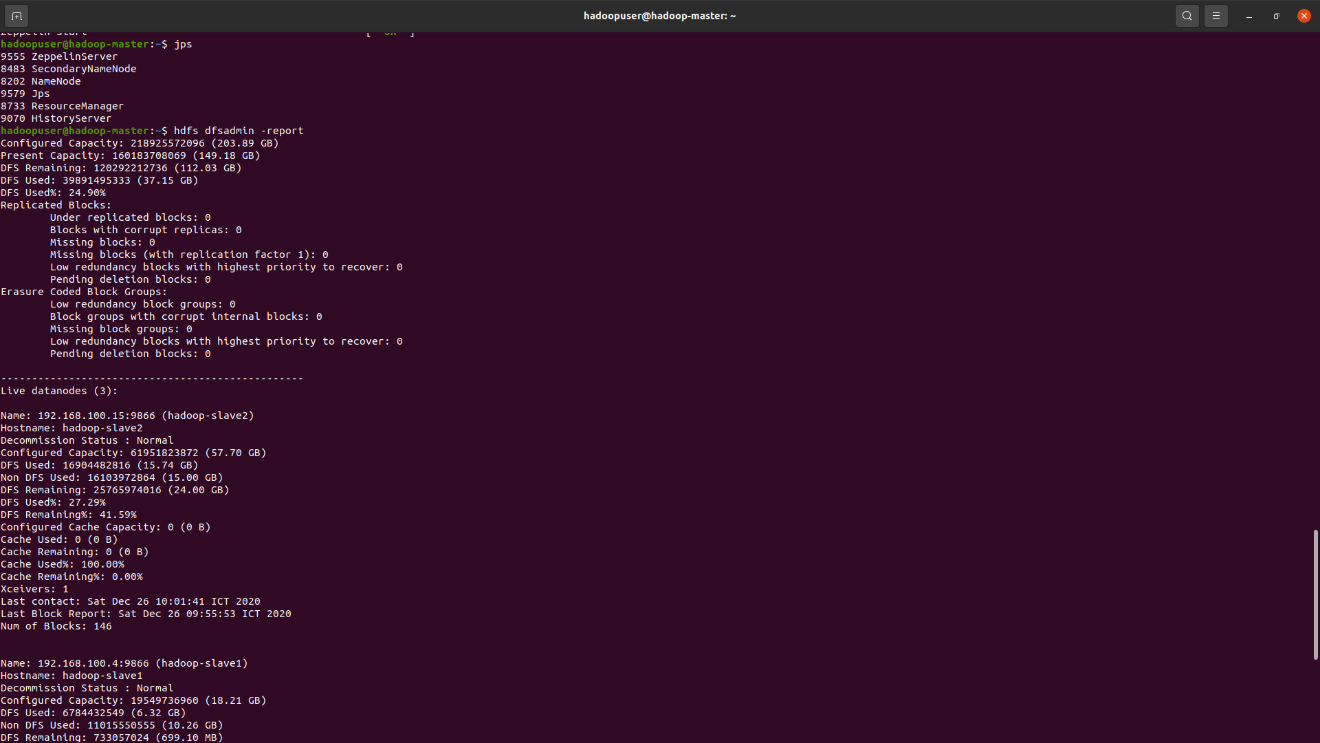
File log lưu lại user activity của một ứng dụng nghe nhạc, từ tháng 1/2018 đến 10/2018 mỗi tháng ~2,3GB như sau:



# II. Các phần mềm đã cài đặt trên toàn cụm

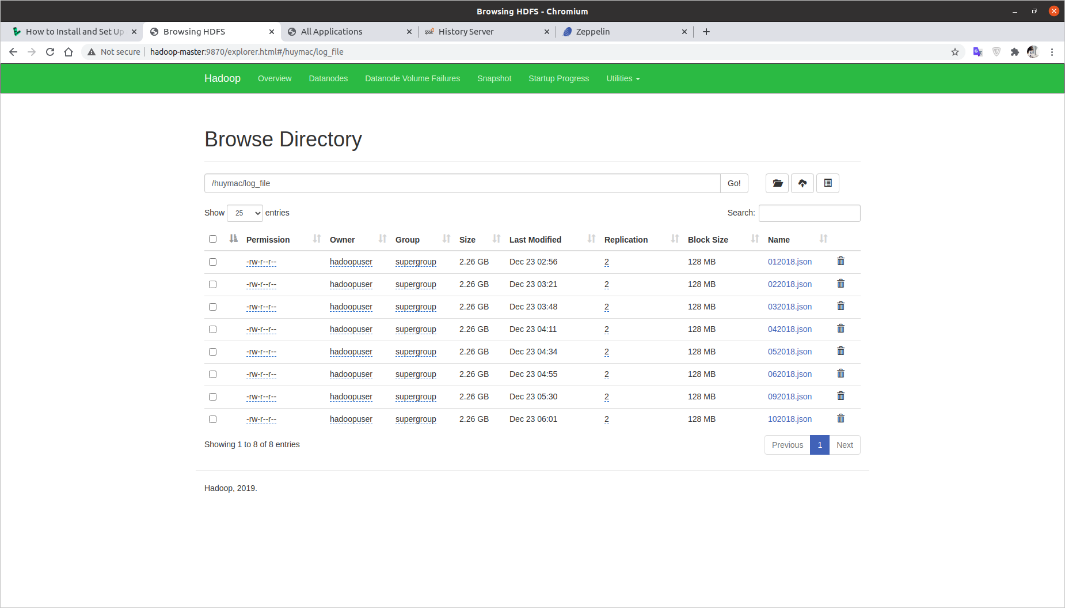
**2.1 Hadoop Distributed File System (**HDFS**)**

hdfs dfsadmin –report

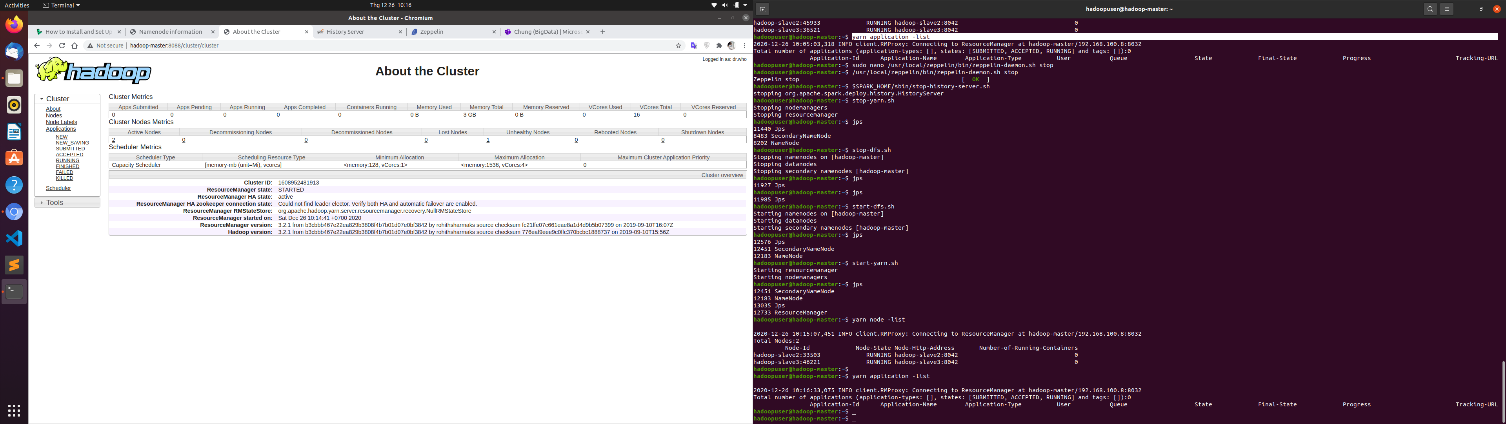




- Data log file trong HDFS

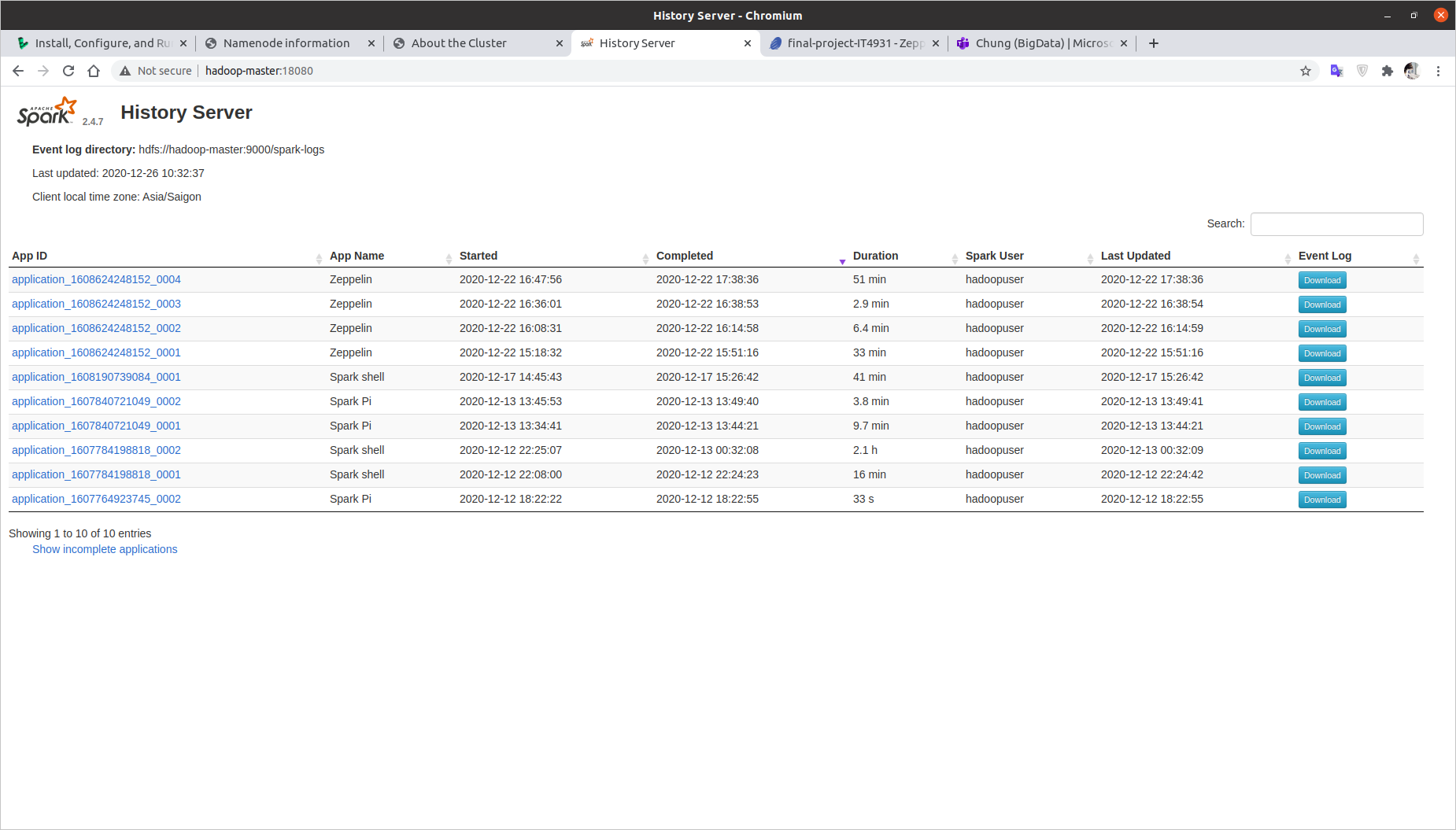


**2.2 Yarn web UI**

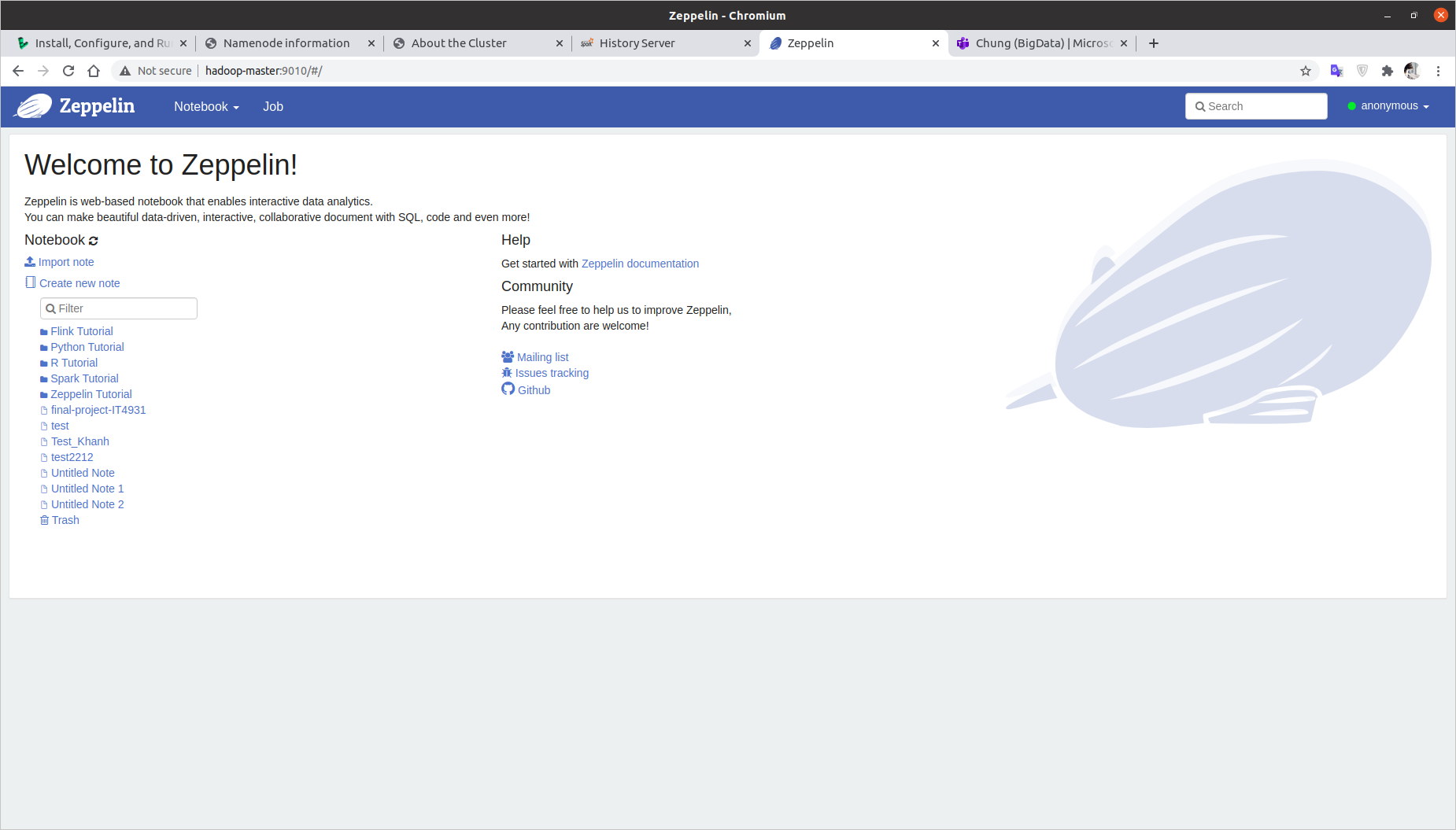


**2.3 Spark cluster Yarn mode**

- Web UI ghi lại các bản log của các job

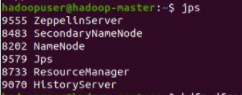


**2.4 Zeppelin 0.8**

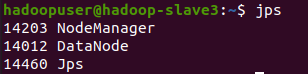


**2.5 Jps trên các node**

Master:

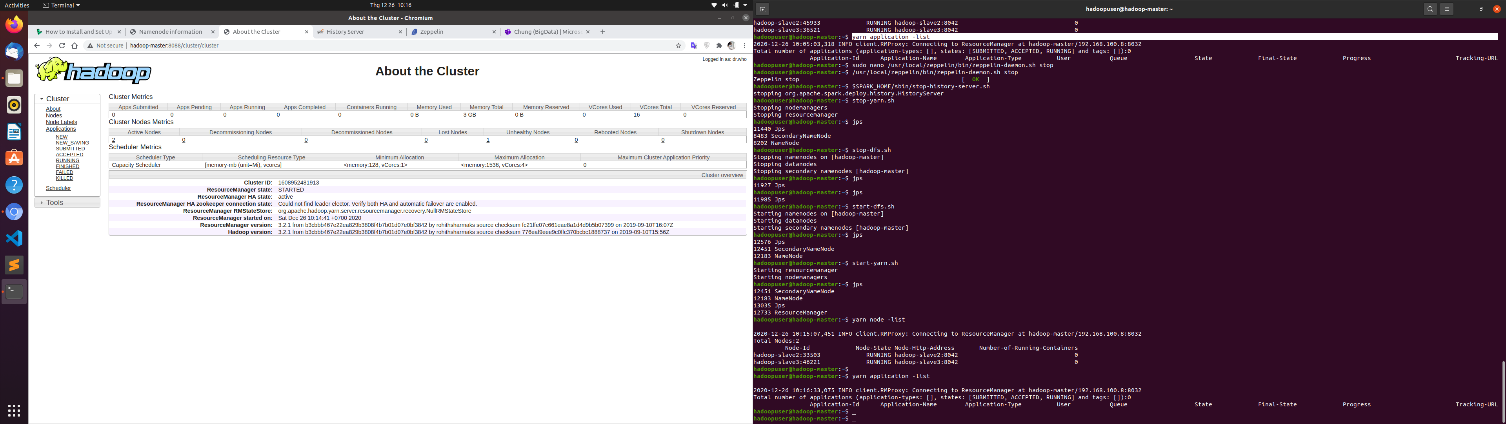


Slave:



2.6 Trường hợp khi 1 node đầy bộ nhớ

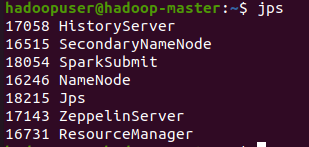
Node đó sẽ bị xếp thành unhealthy nodes



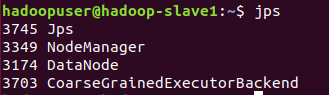
# III. Thực thi job

**3.1 Jps khi job đang được submit**

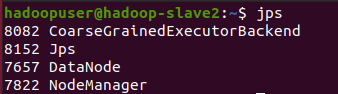
* Master:



* Slave 1:



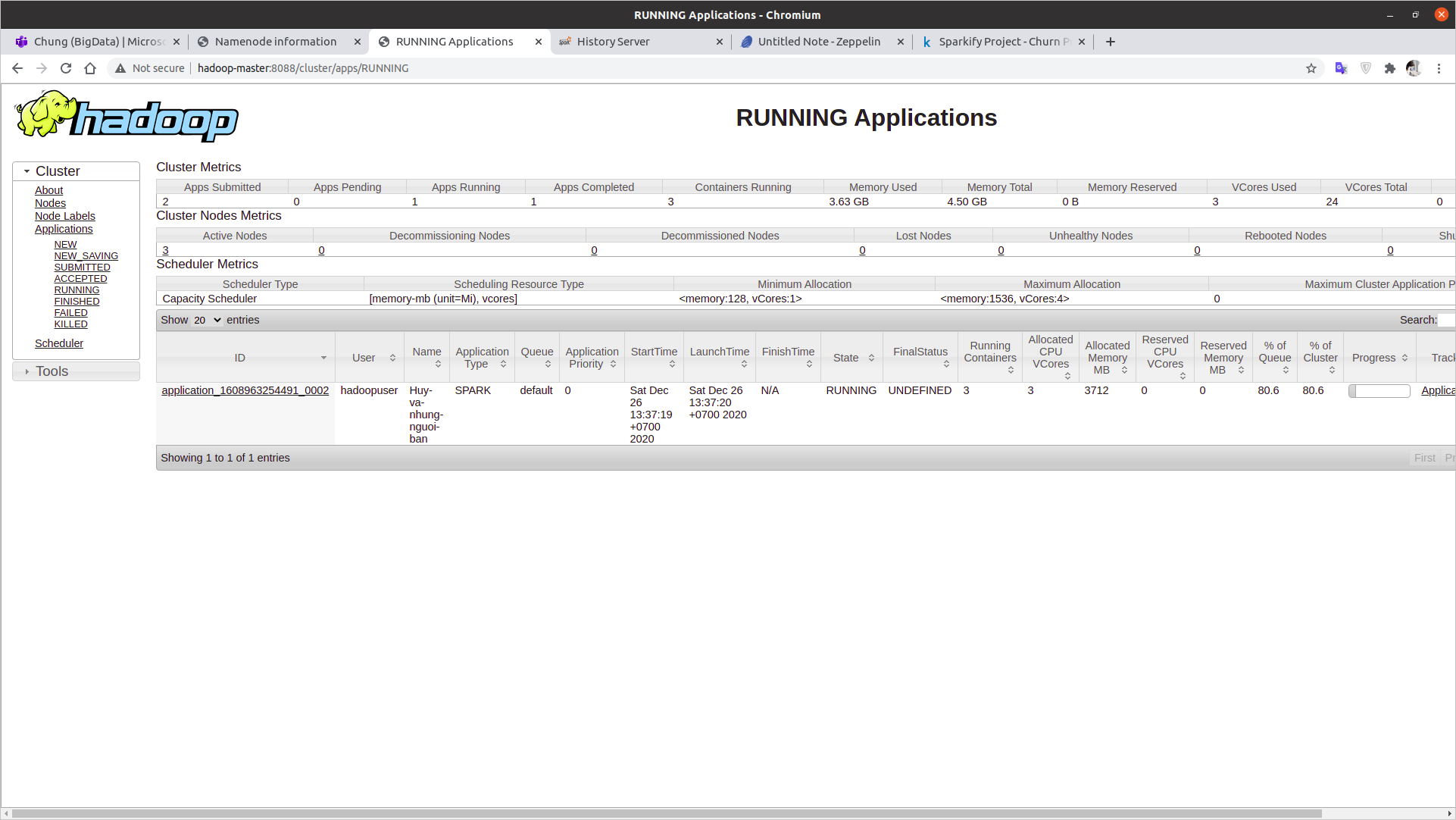
* Slave 2:



* Slave 3:



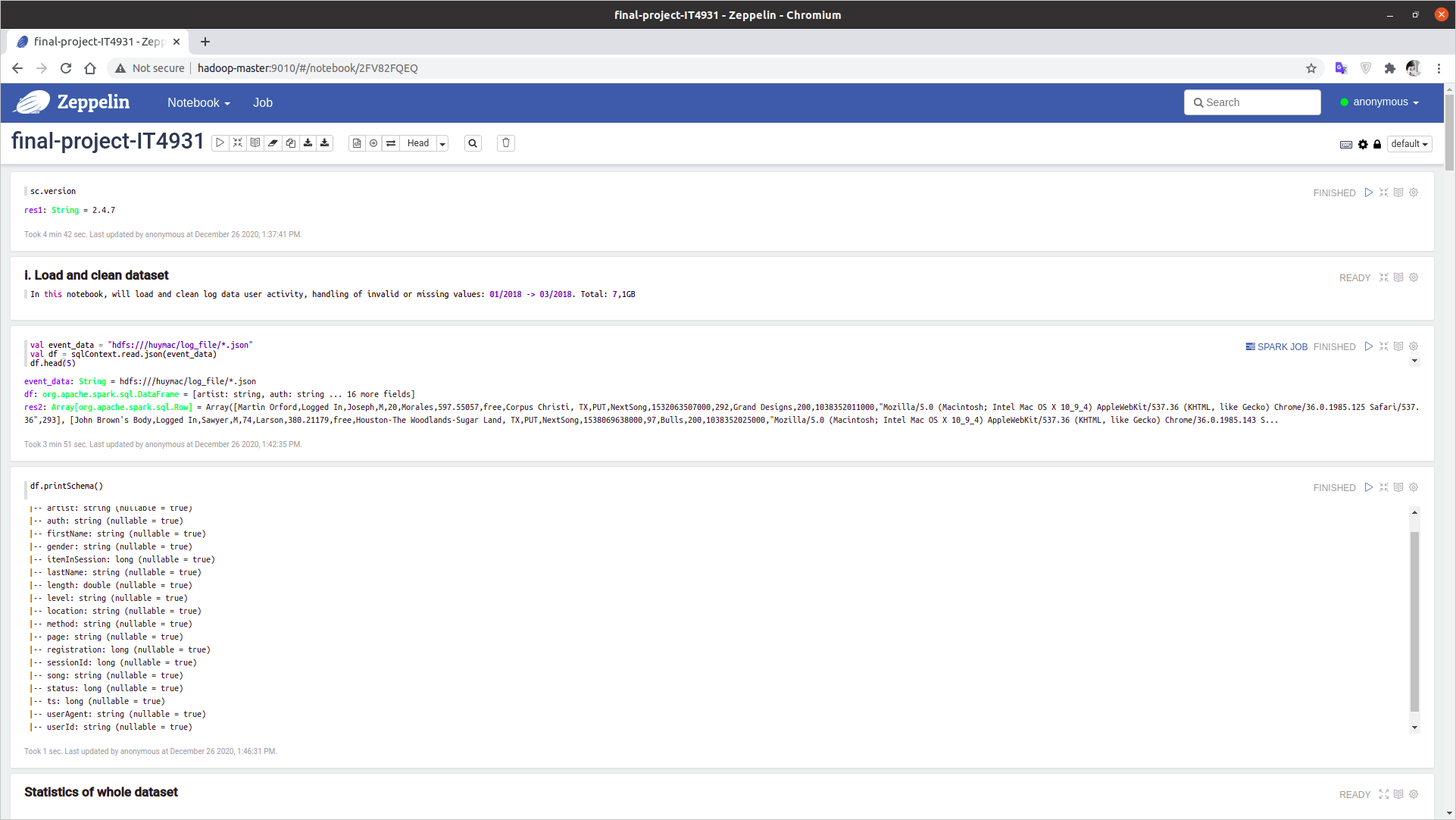
## 3.2 Check Yarn web UI khi job đang chạy

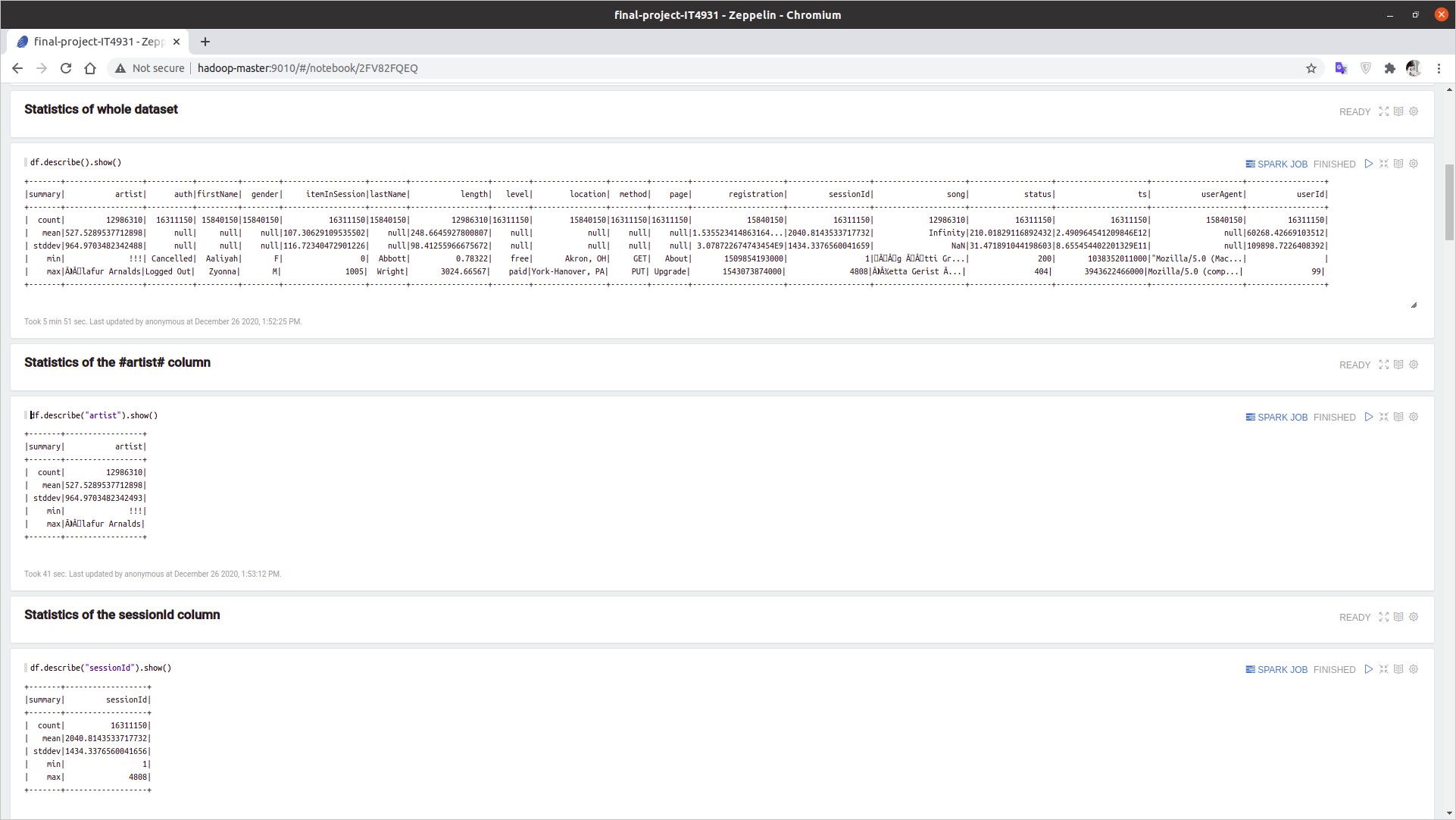


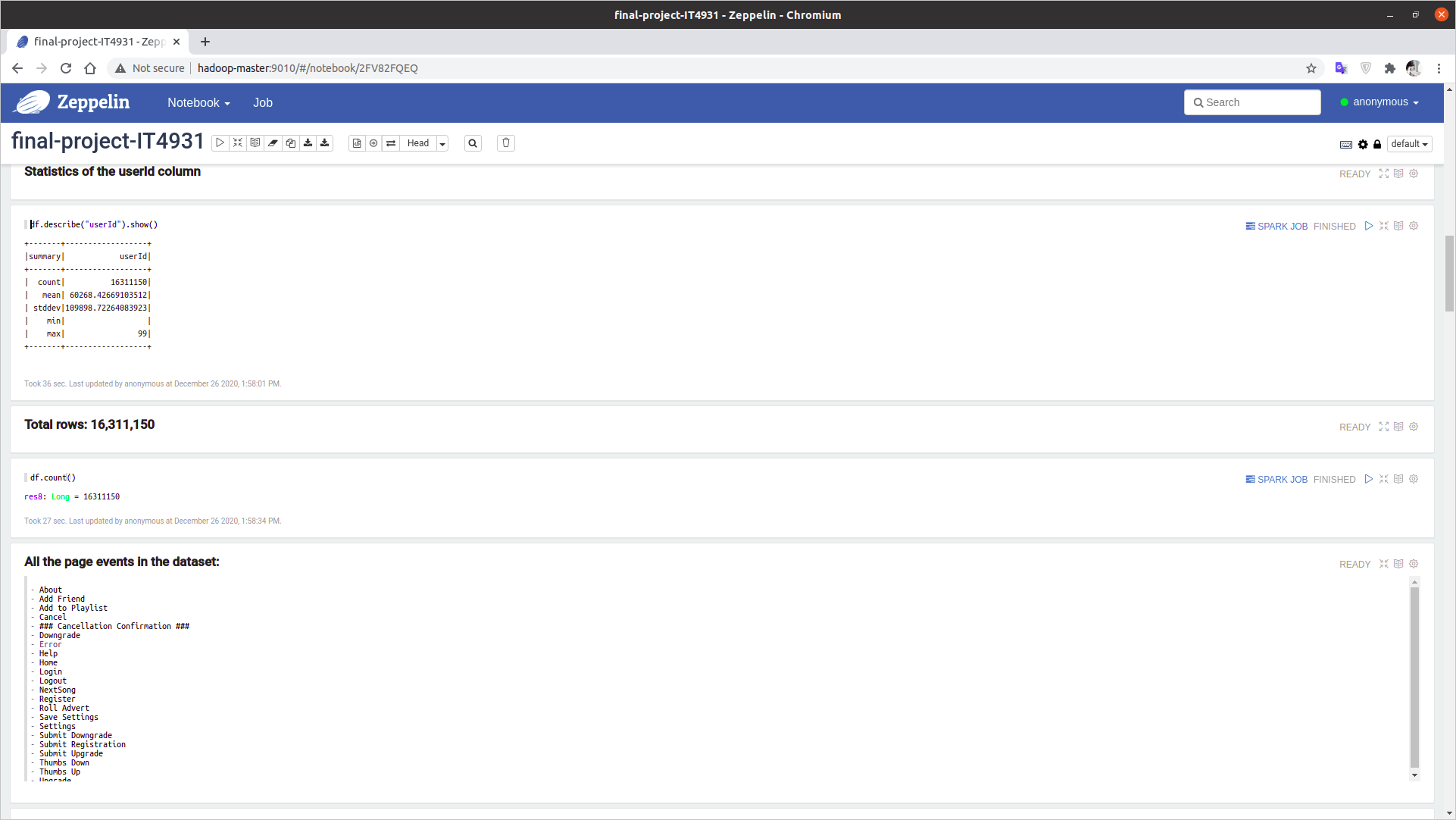
# IV. Code trên Zeppelin

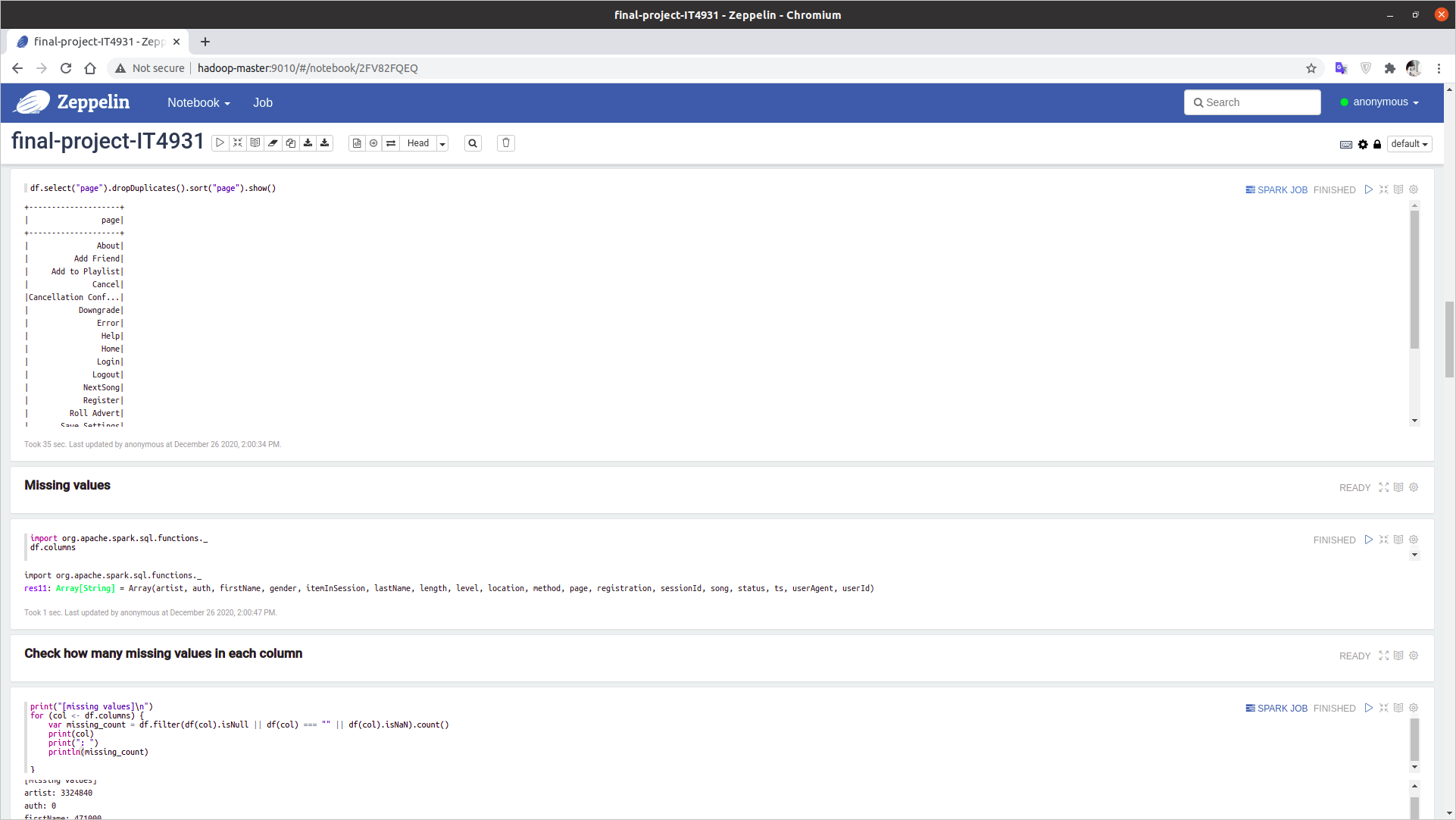
Ngôn ngữ: Scala 2.4.7

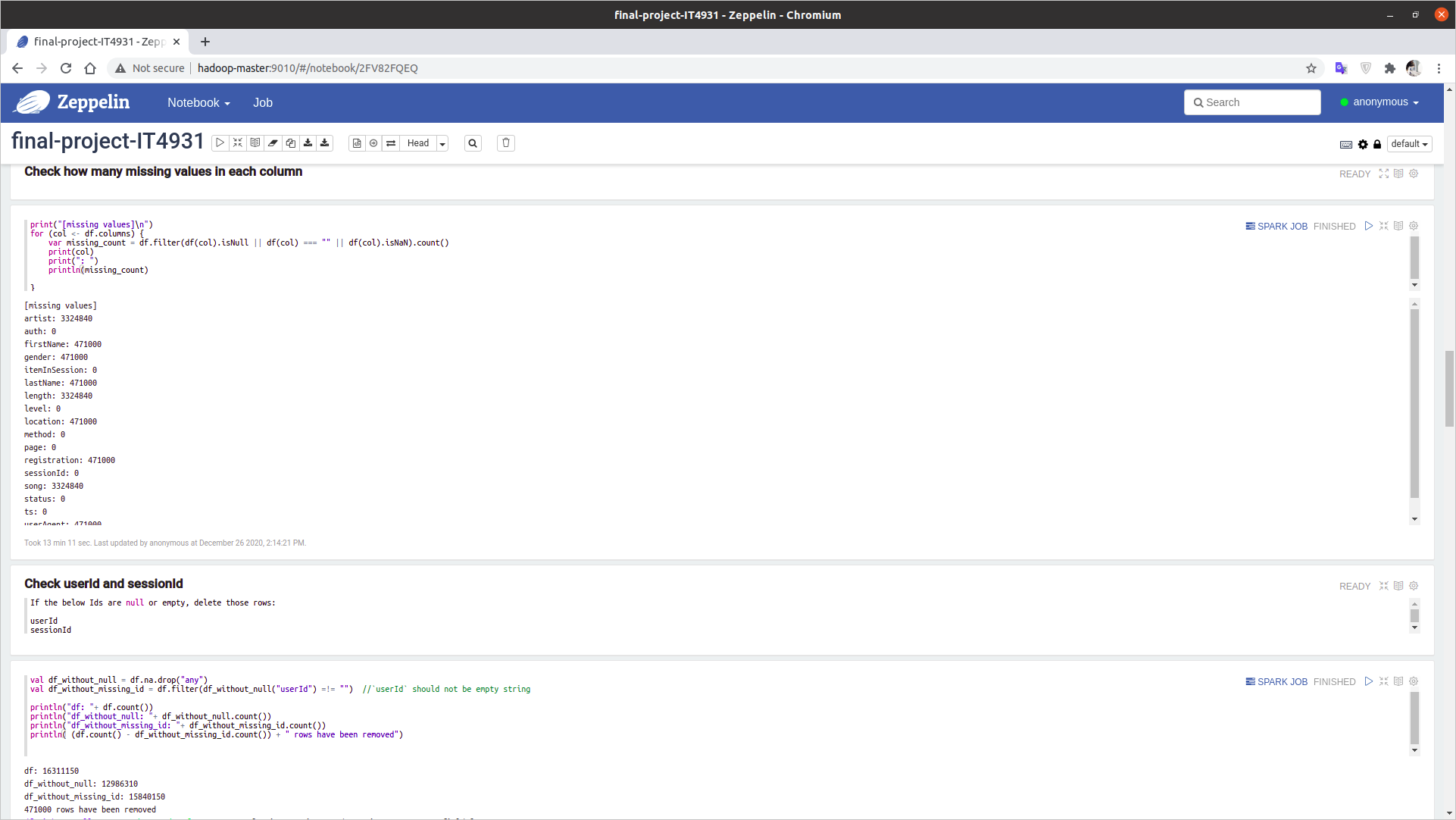
## 4.1 Load và clean dataset

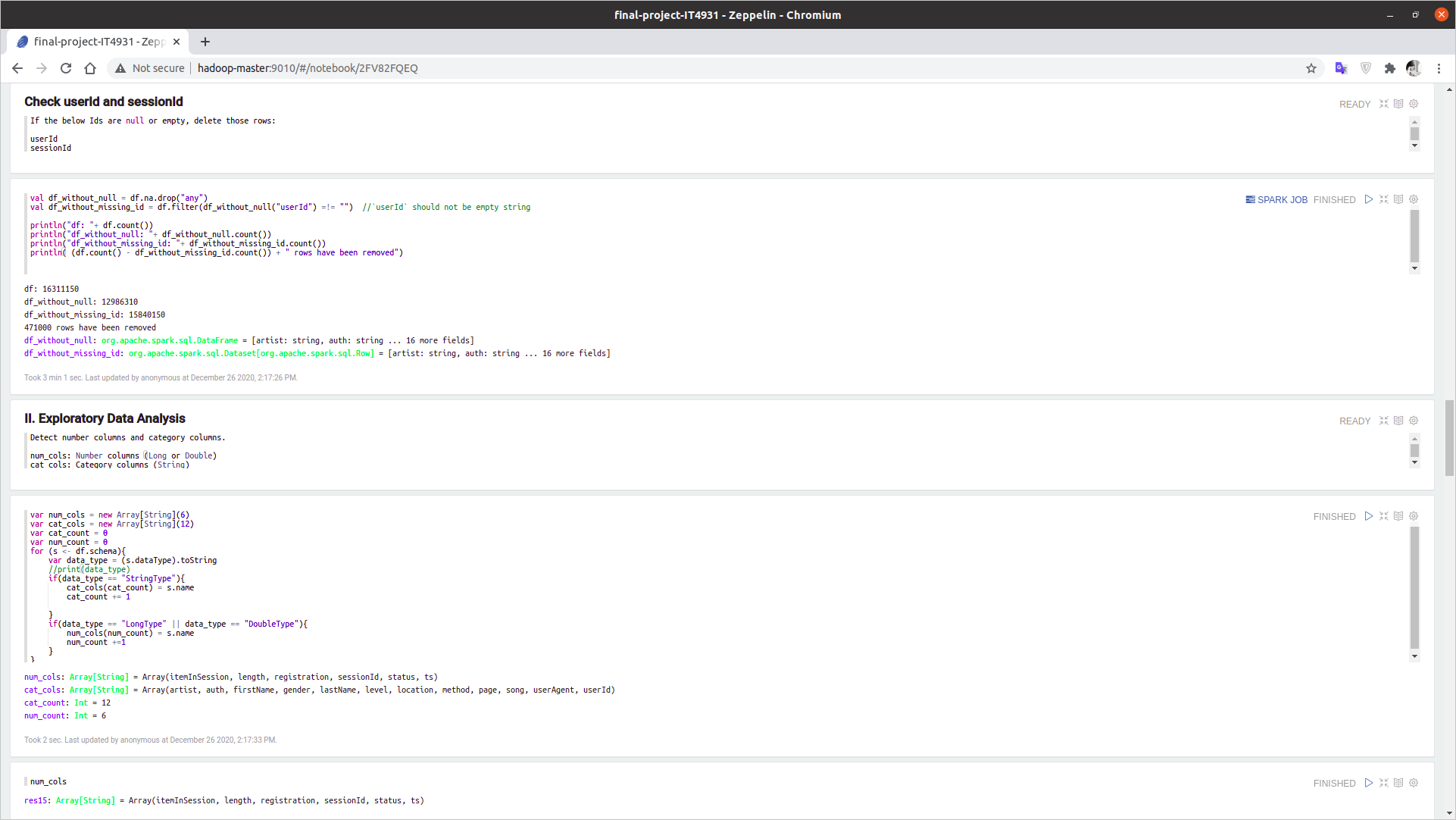




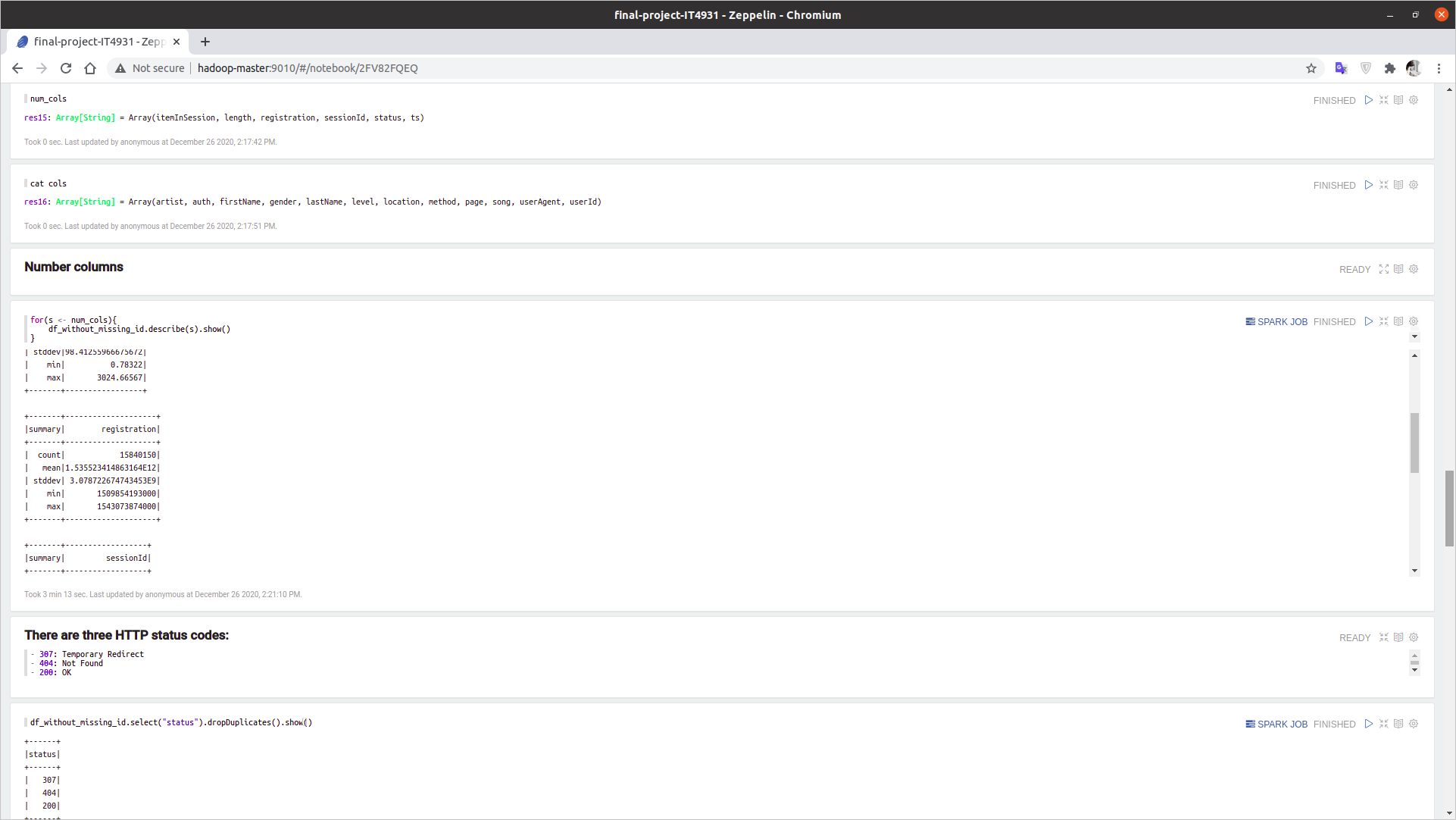


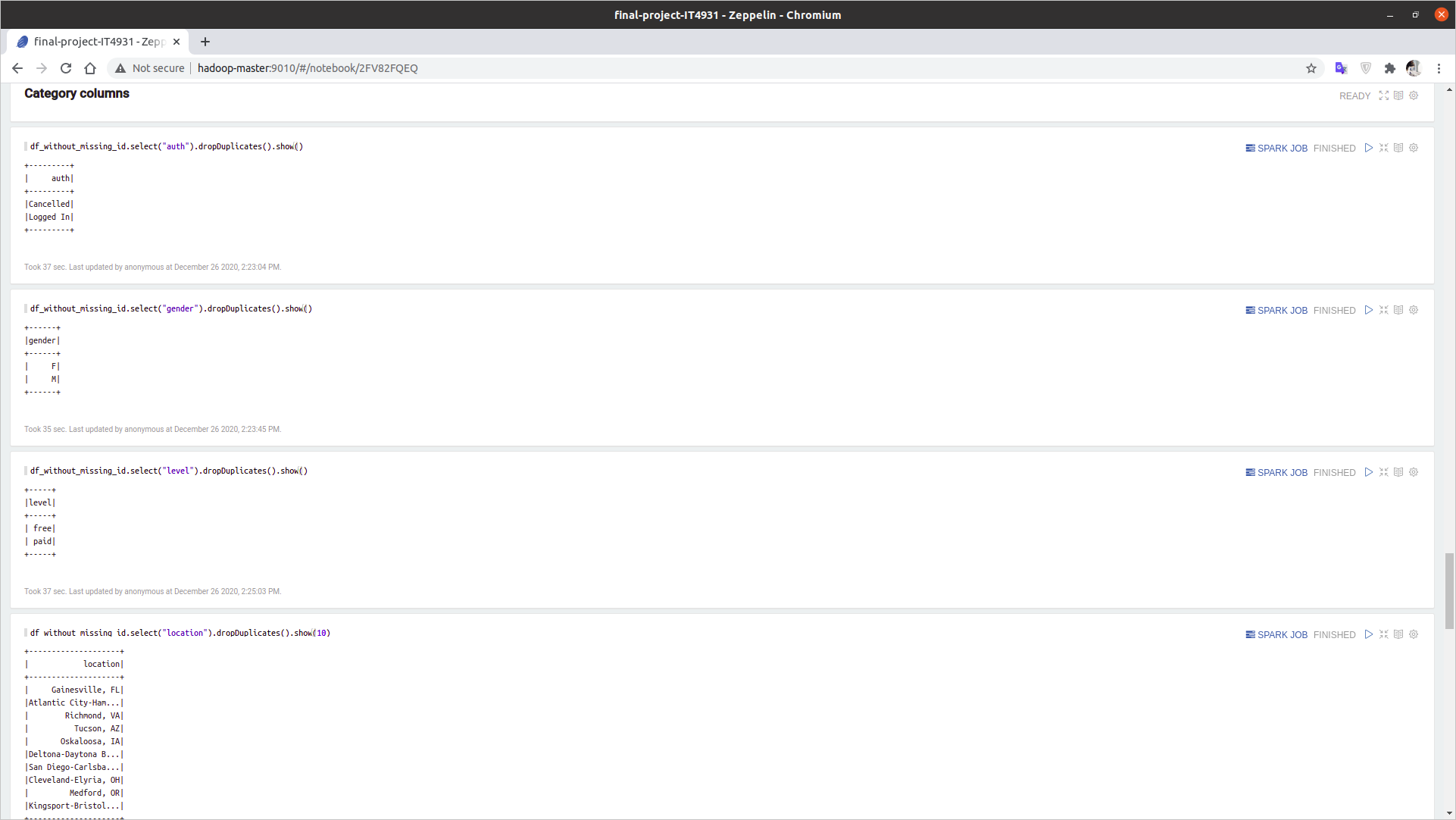


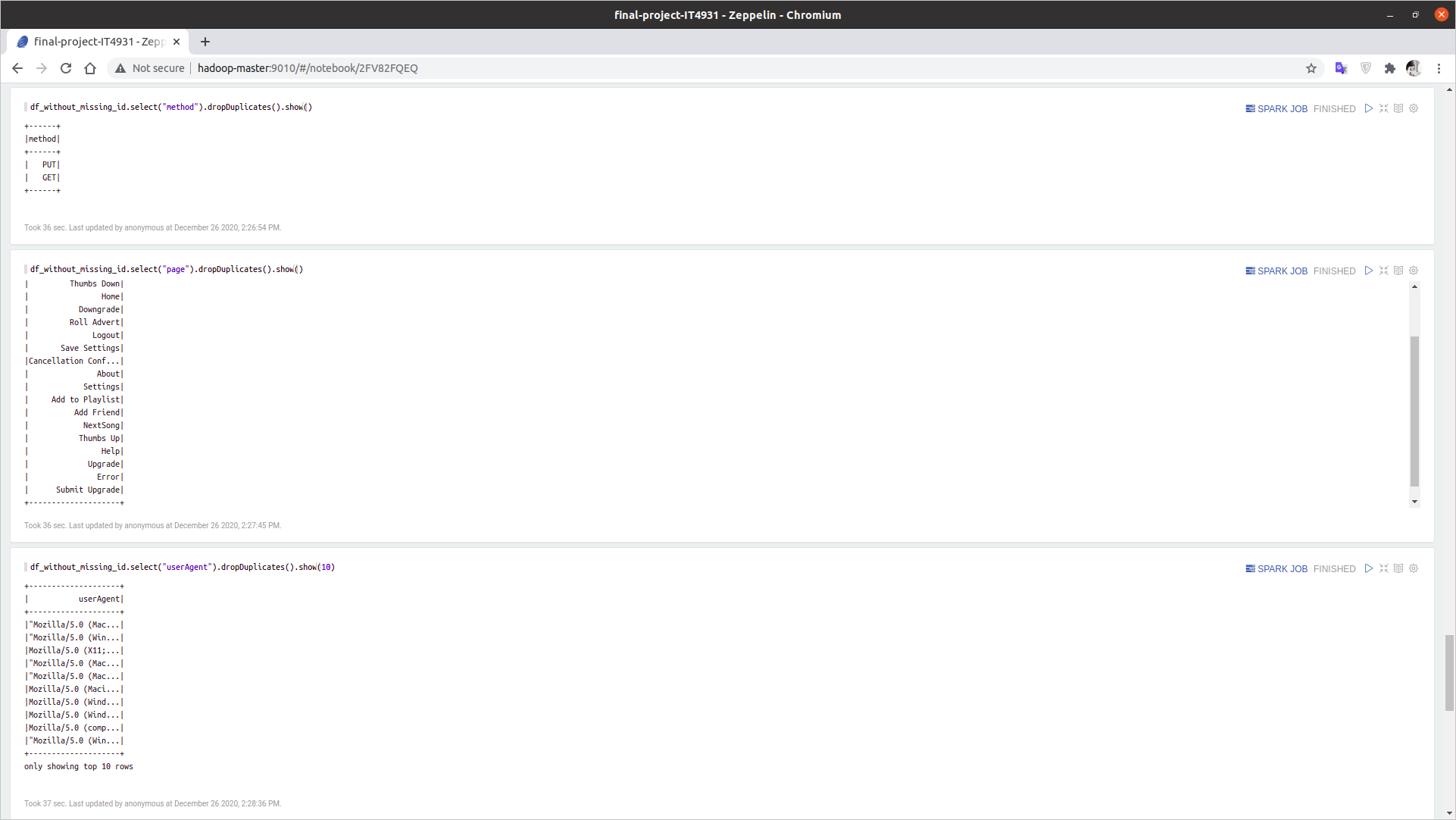


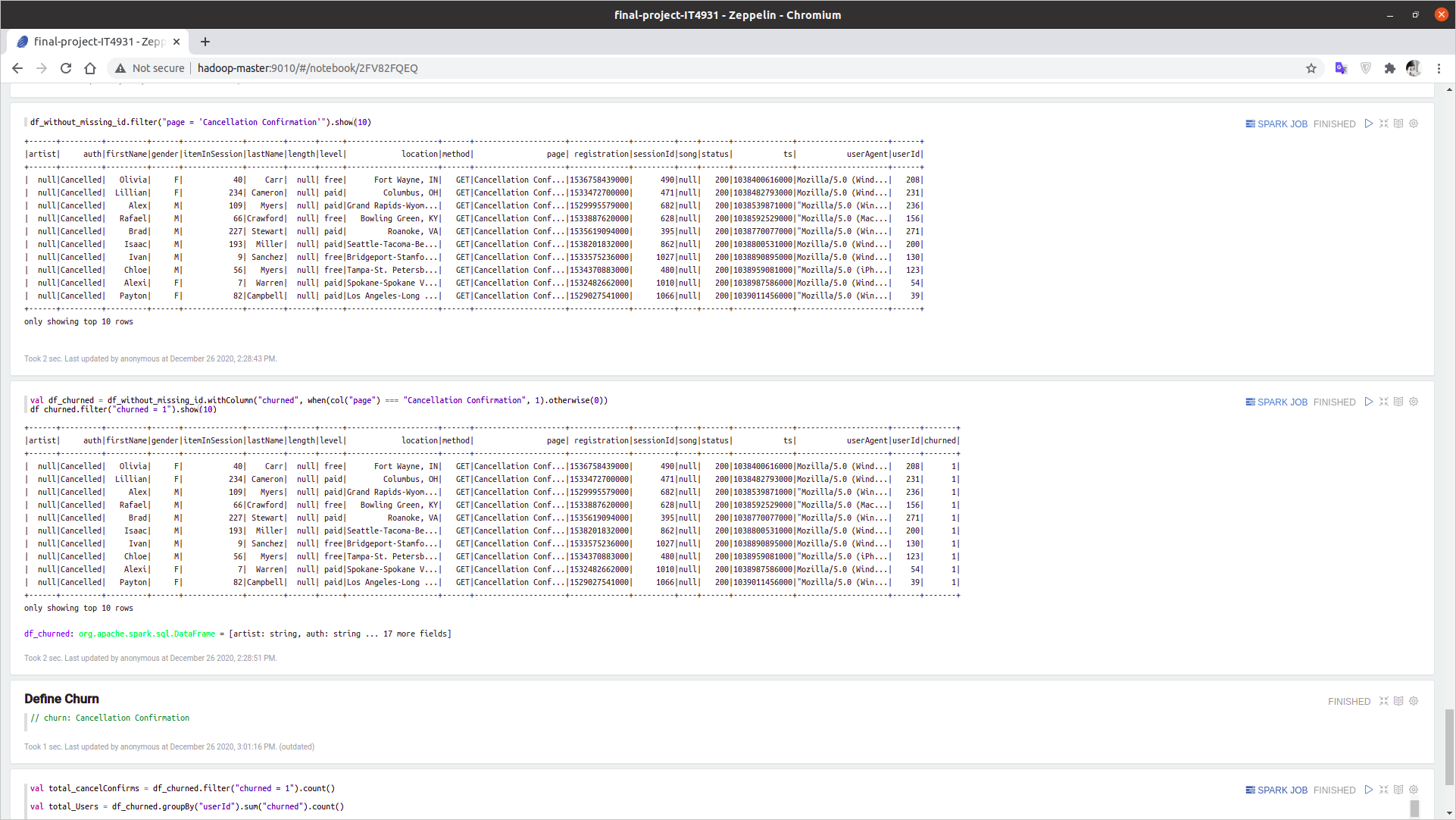


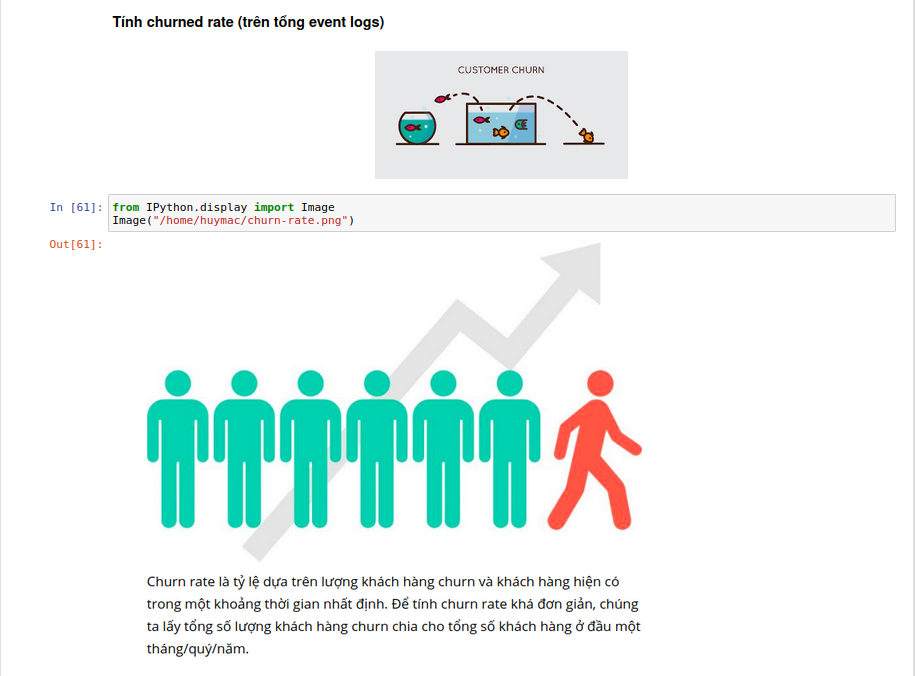
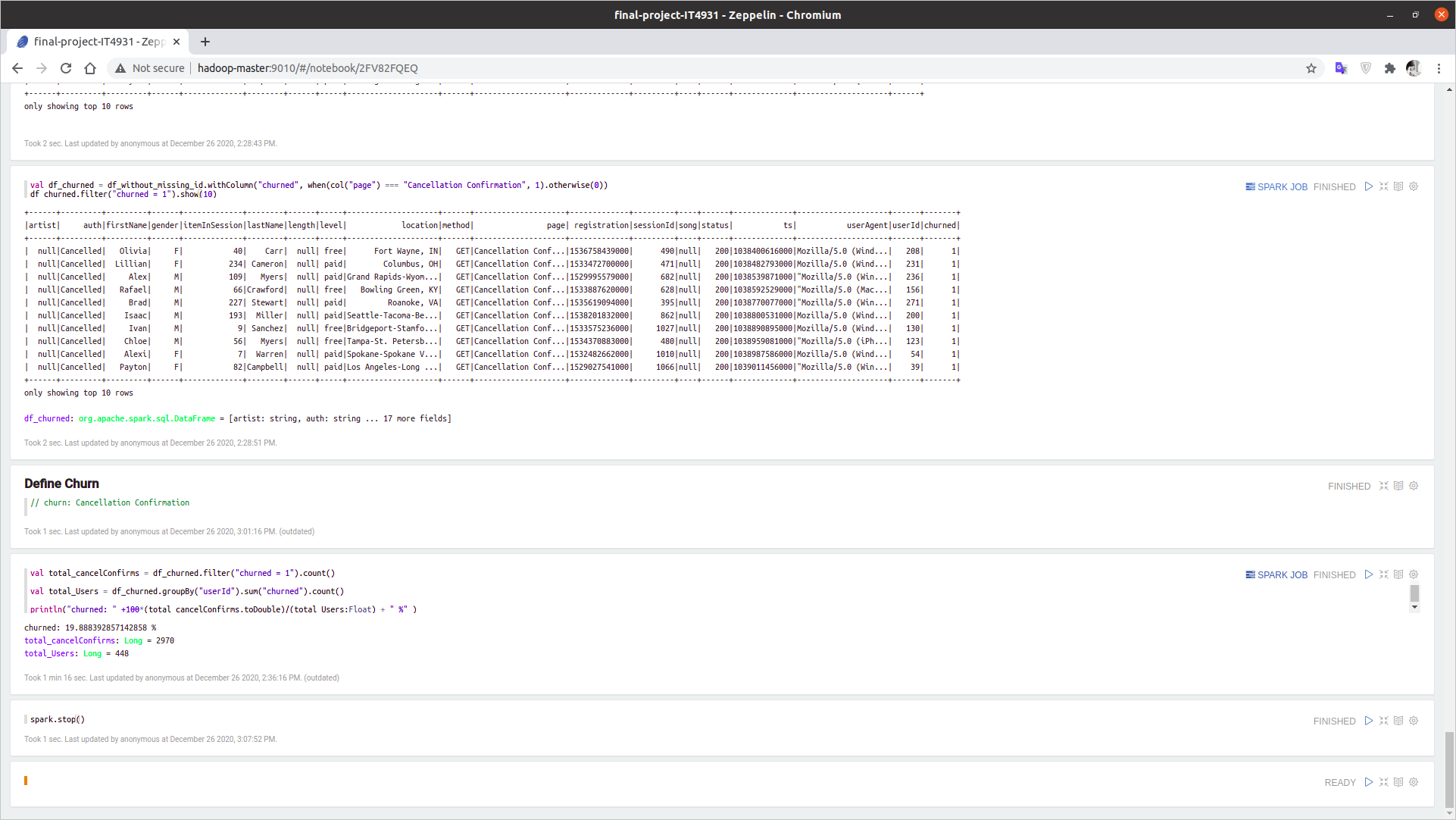
## 4.2 Exploratory data analysis



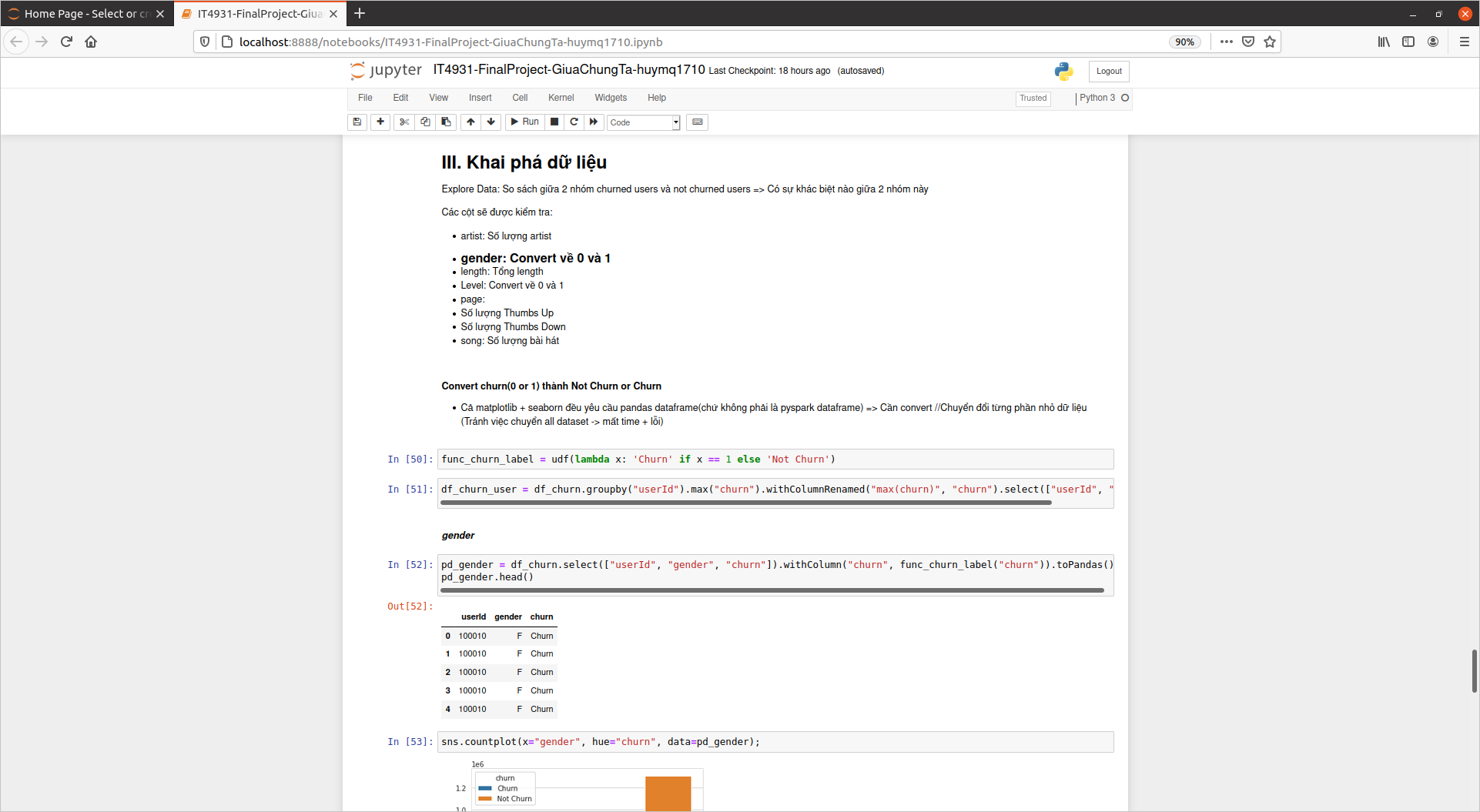


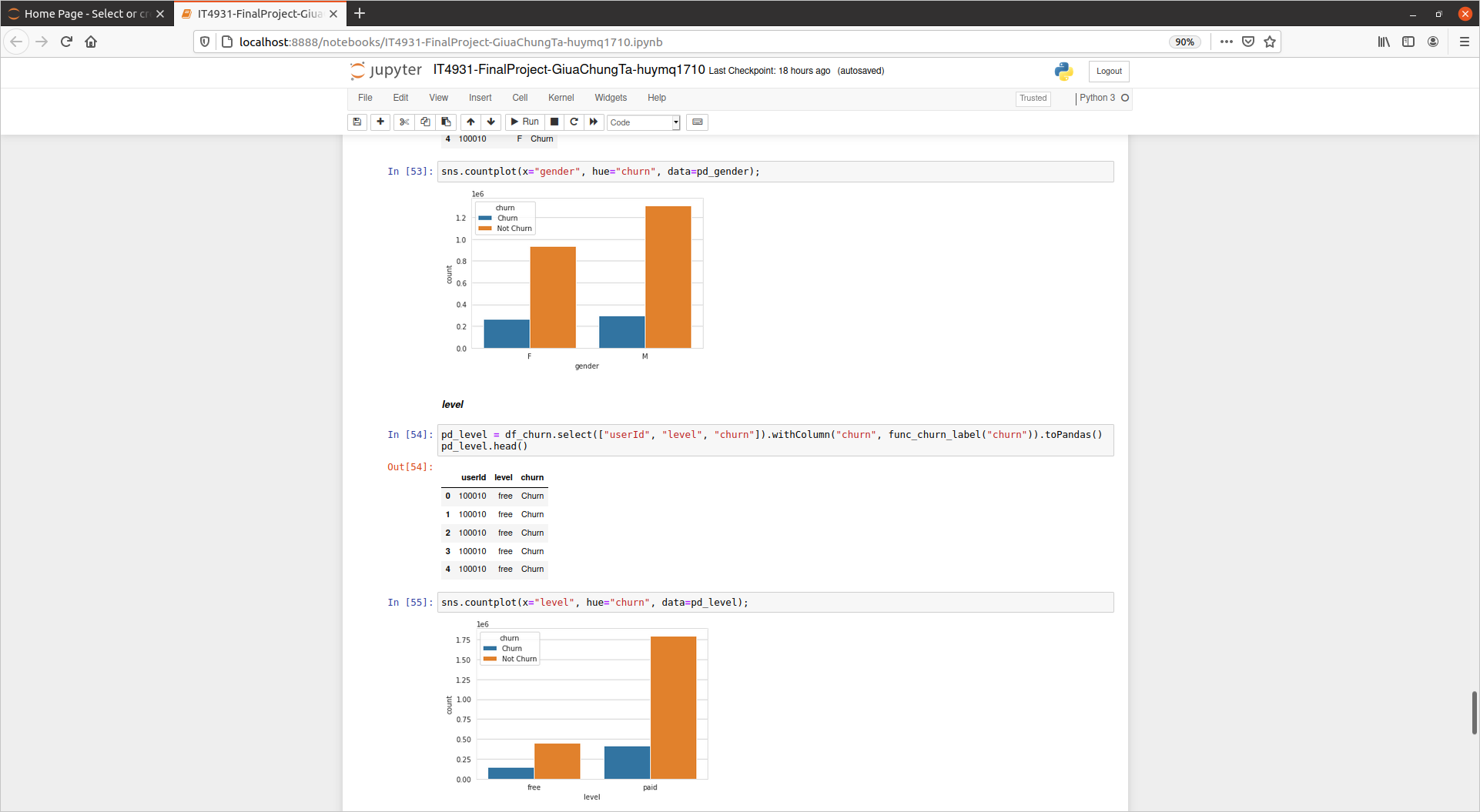


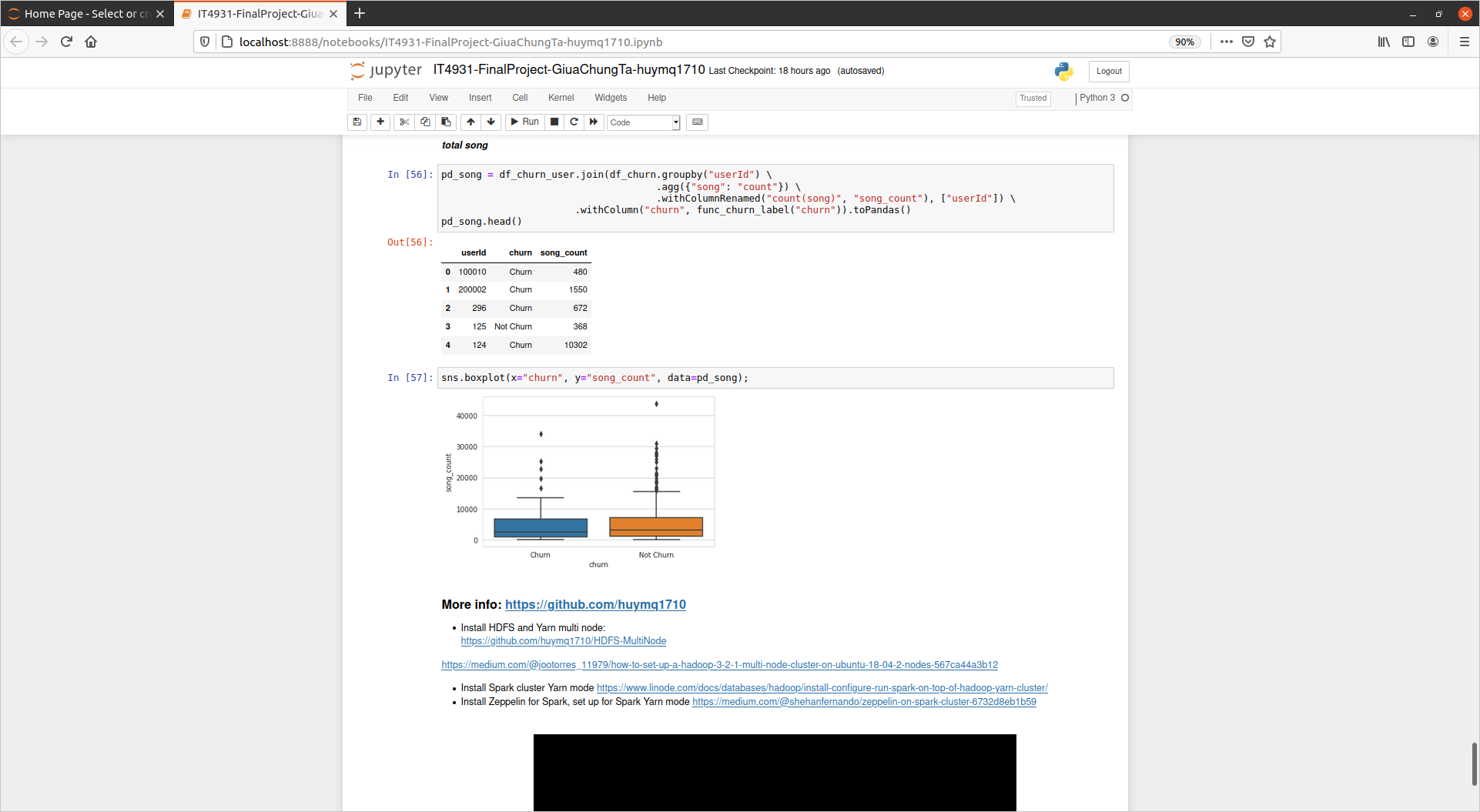


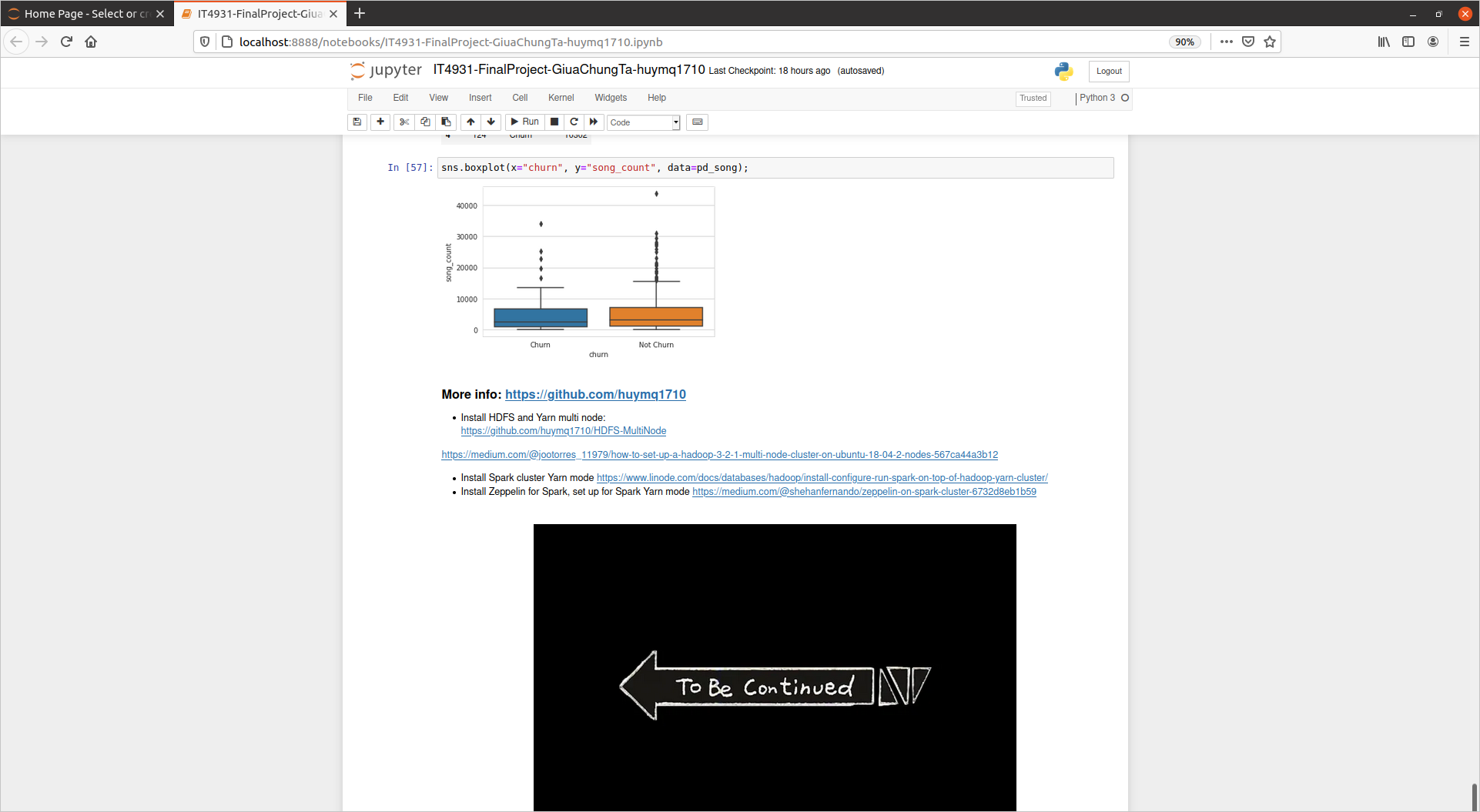


## 4.3 Khai phá dữ liệu sử dụng pyspark trên Jupyter notebook

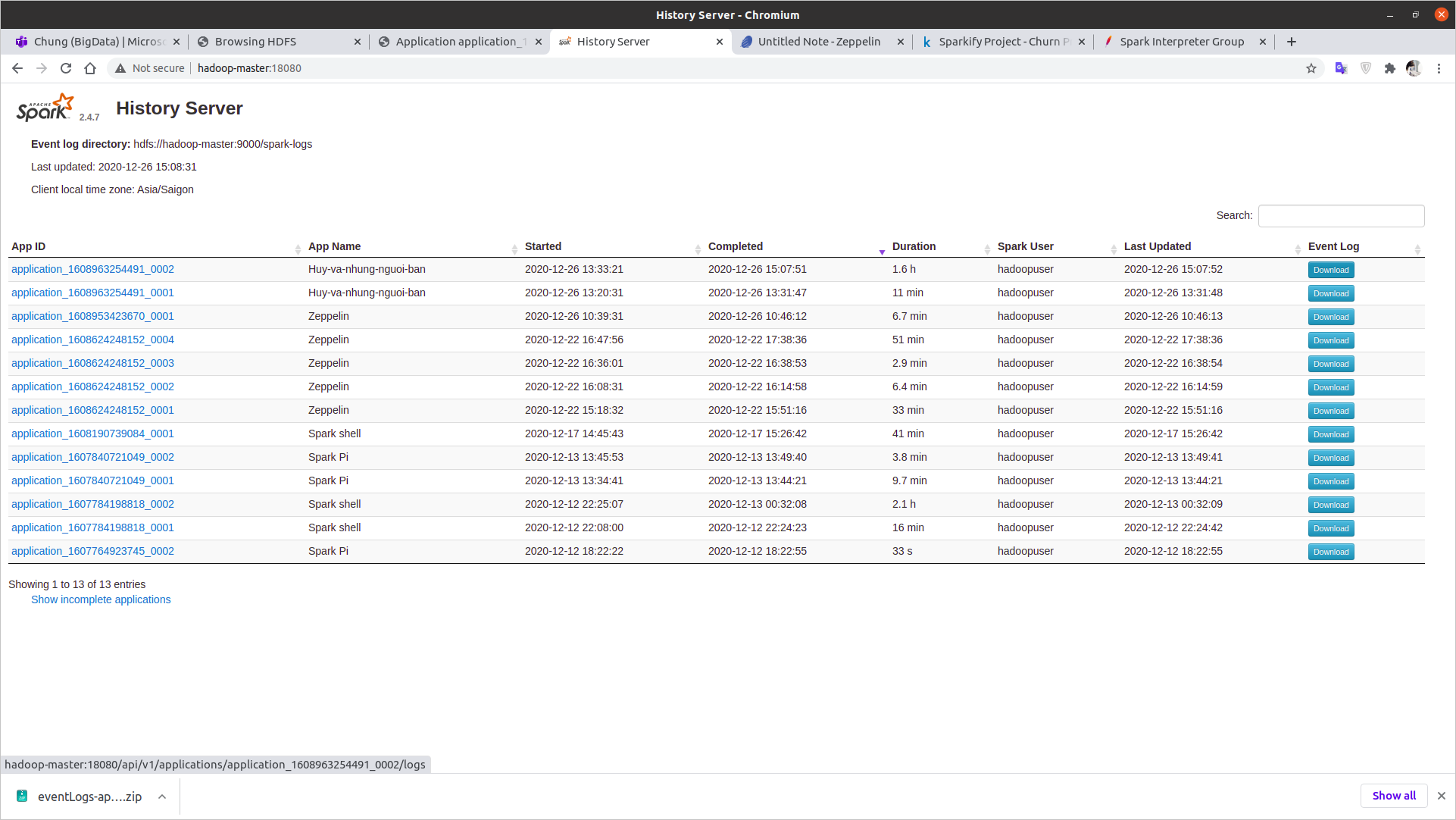


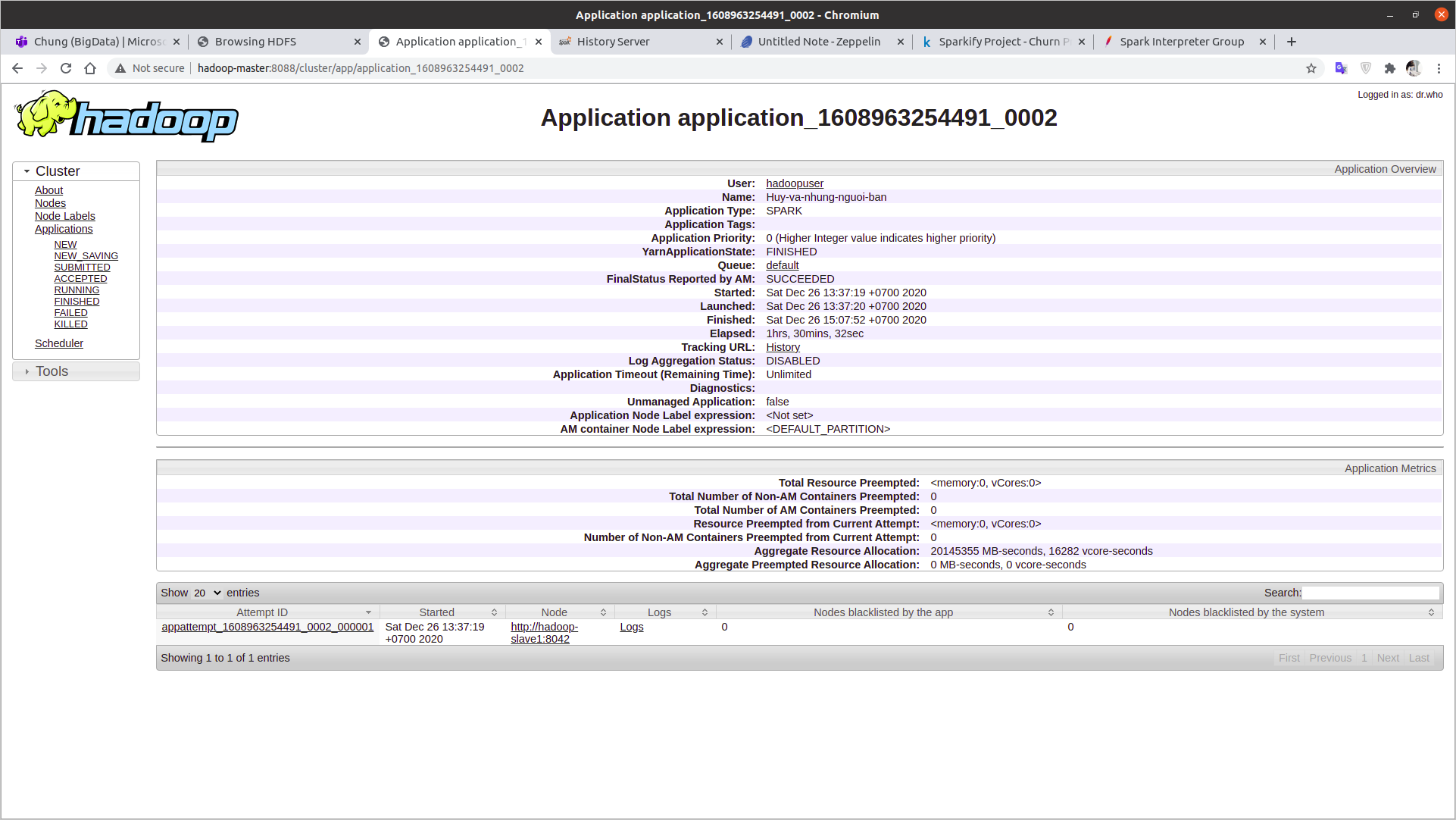






## 4.4 Job done





## 

# 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 xày 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.

# REFRENCES

1. Mobile Marketing Association, 2008, USA 1670 Broadway, Suite 850, Denver, CO 80202.
2. Mobile web, Wikipedia, URL: http://en.wikipedia.org/wiki/Mobile\_Web.
3. Napoli, Marco. (2019). Introducing Flutter and Getting Started. 10.1002/9781119550860.ch1.

[View publication stats](https://www.researchgate.net/publication/325527960)