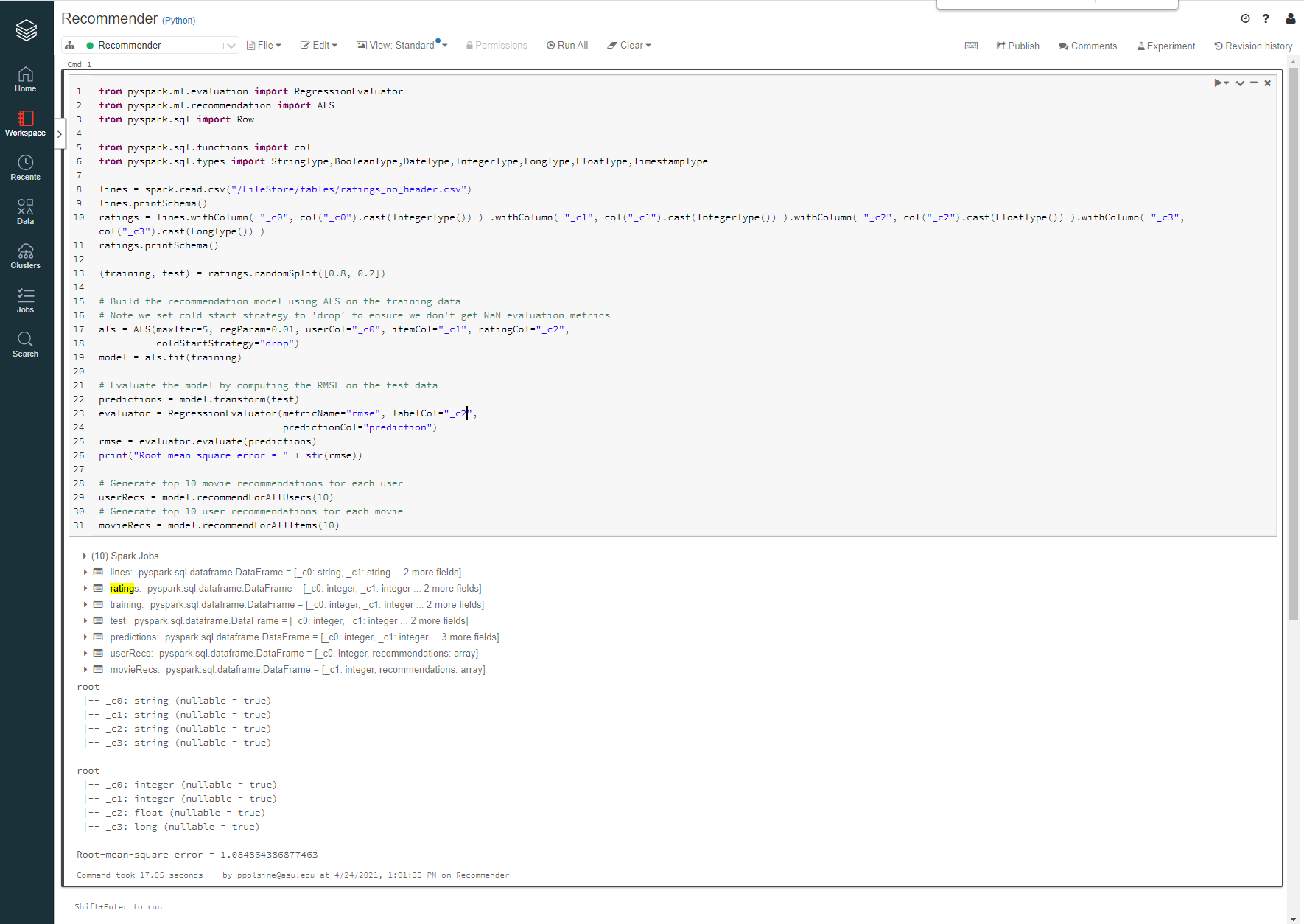
# Lab 12: Automation of a Recommender System

### **Part I: Collaborative Filtering with Spark’s ML Library**

**What is the number of users, movies and ratings in this dataset?** 610 users, 9,742 movies, and 100,836 ratings

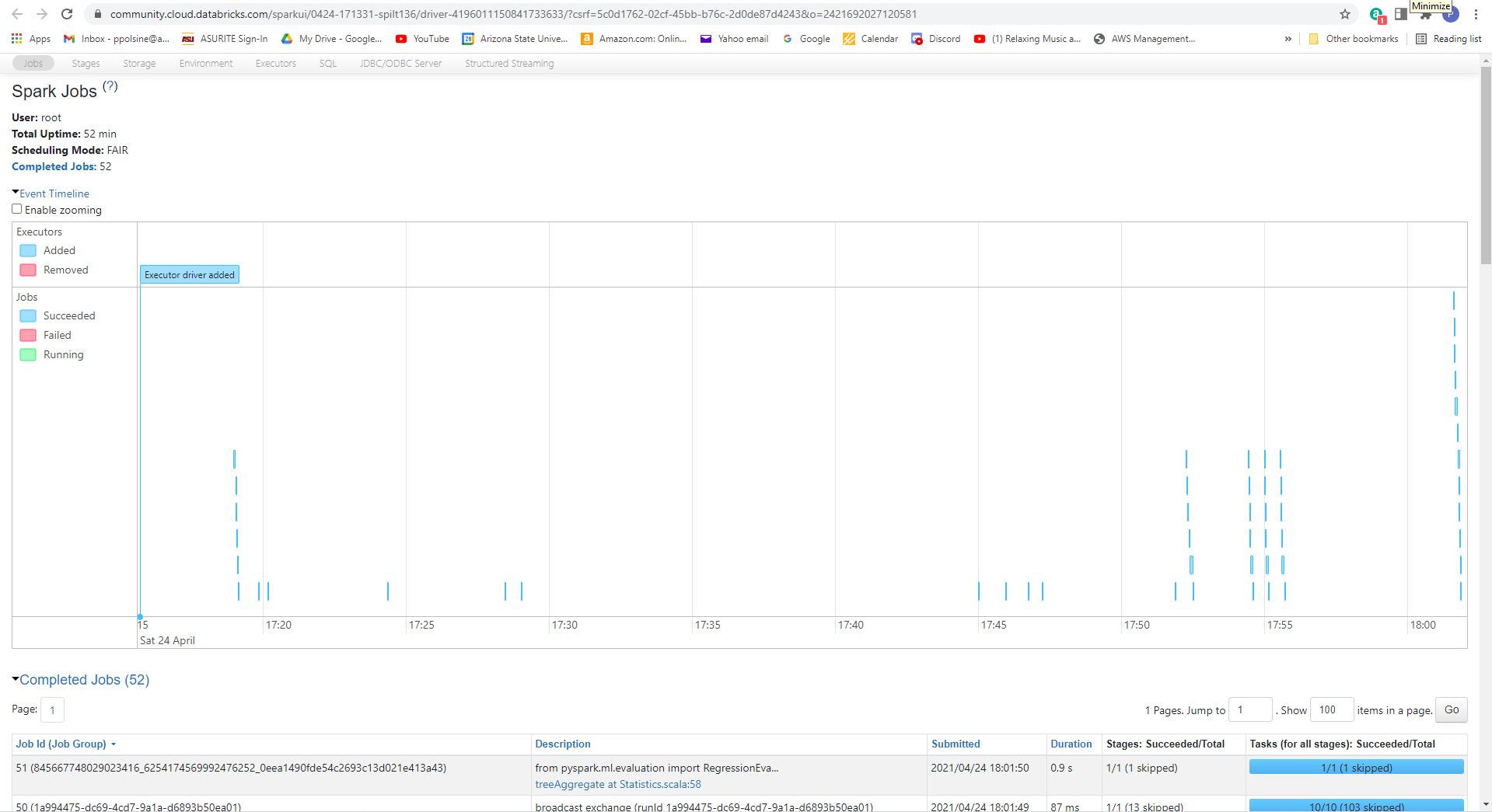
**From these numbers, specify the number of data points, the number of data attributes, and the number of existing values in the dataset.** There are 100,836 data points, 3 attributes, and 302,508 values; (not including timestamp).

**What is the number of missing values in this dataset?** No missing values are indicated

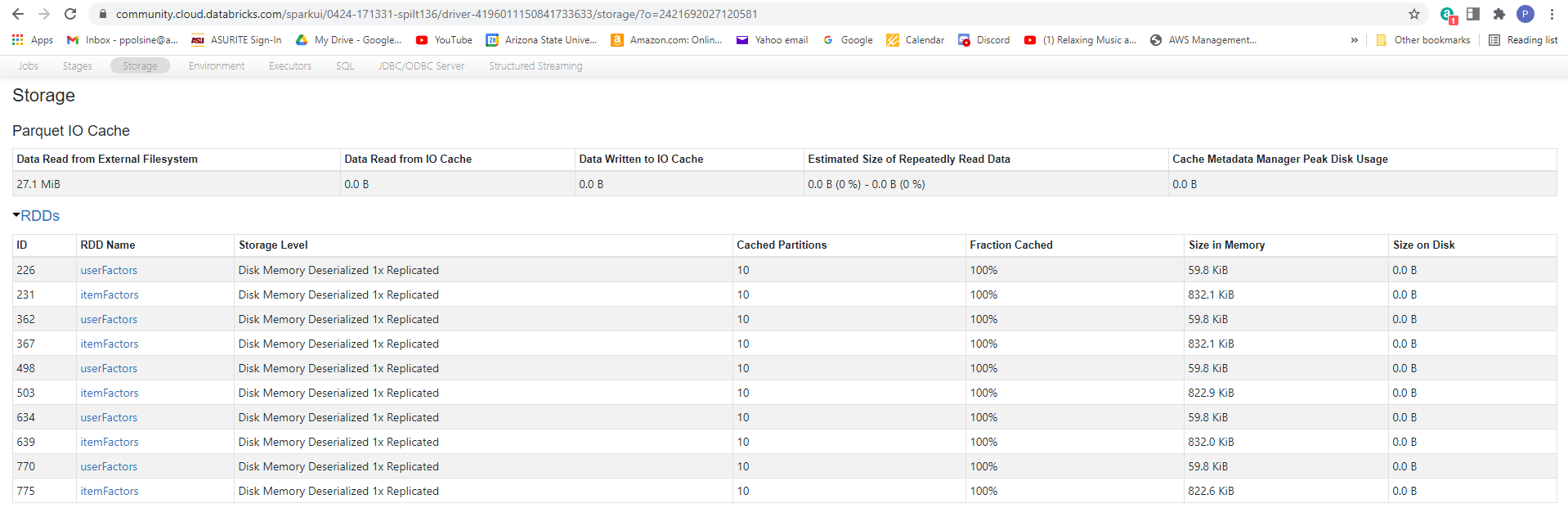


**What is the RMSE on the test set given by the recommender?** Root-mean-square error = 1.084864386877463

**How many seconds did the recommender take to generate top-10 recommendations for the 600 users?** 17.05 seconds

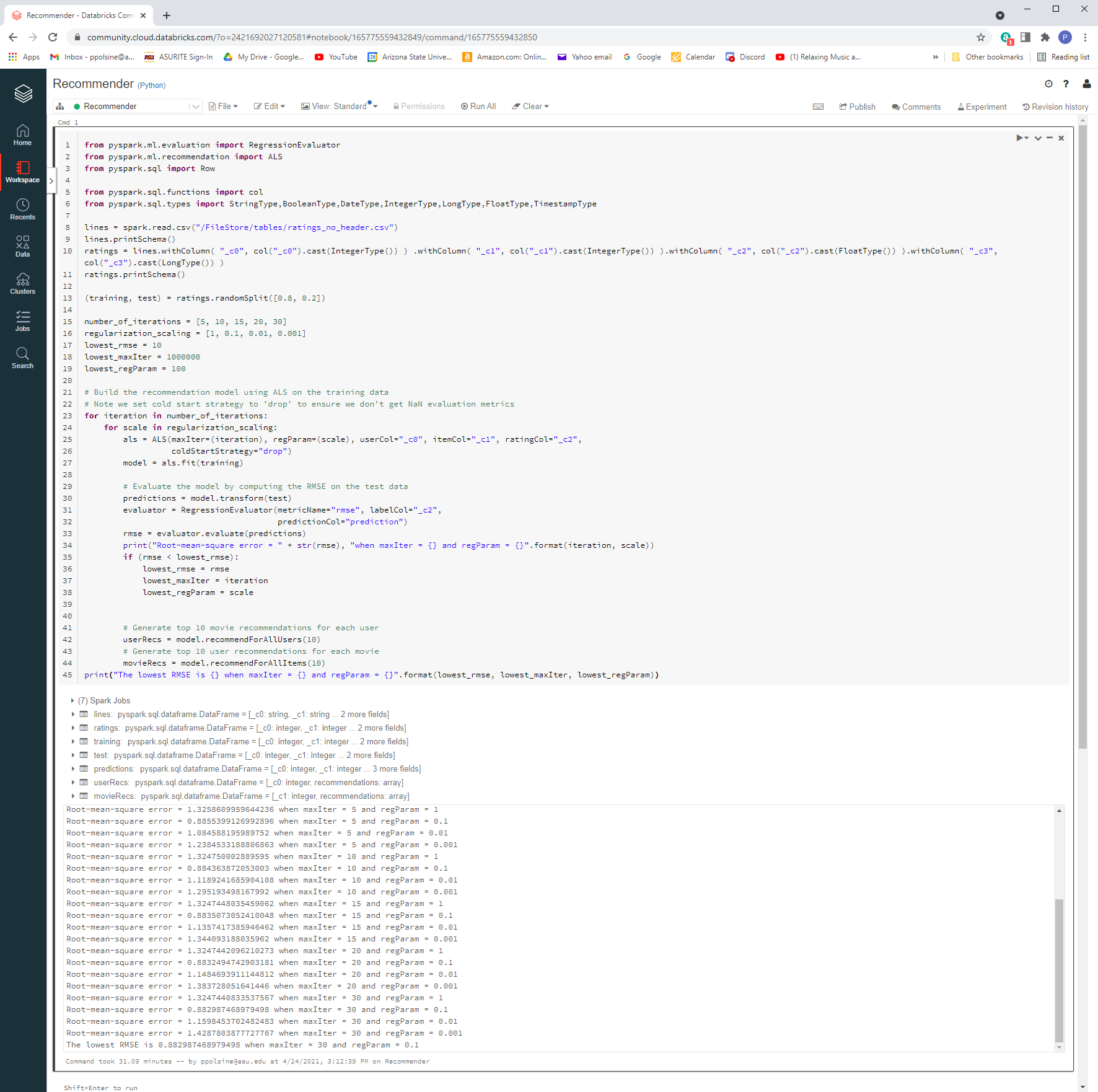


**How many jobs did Spark submit to run your code?** 52



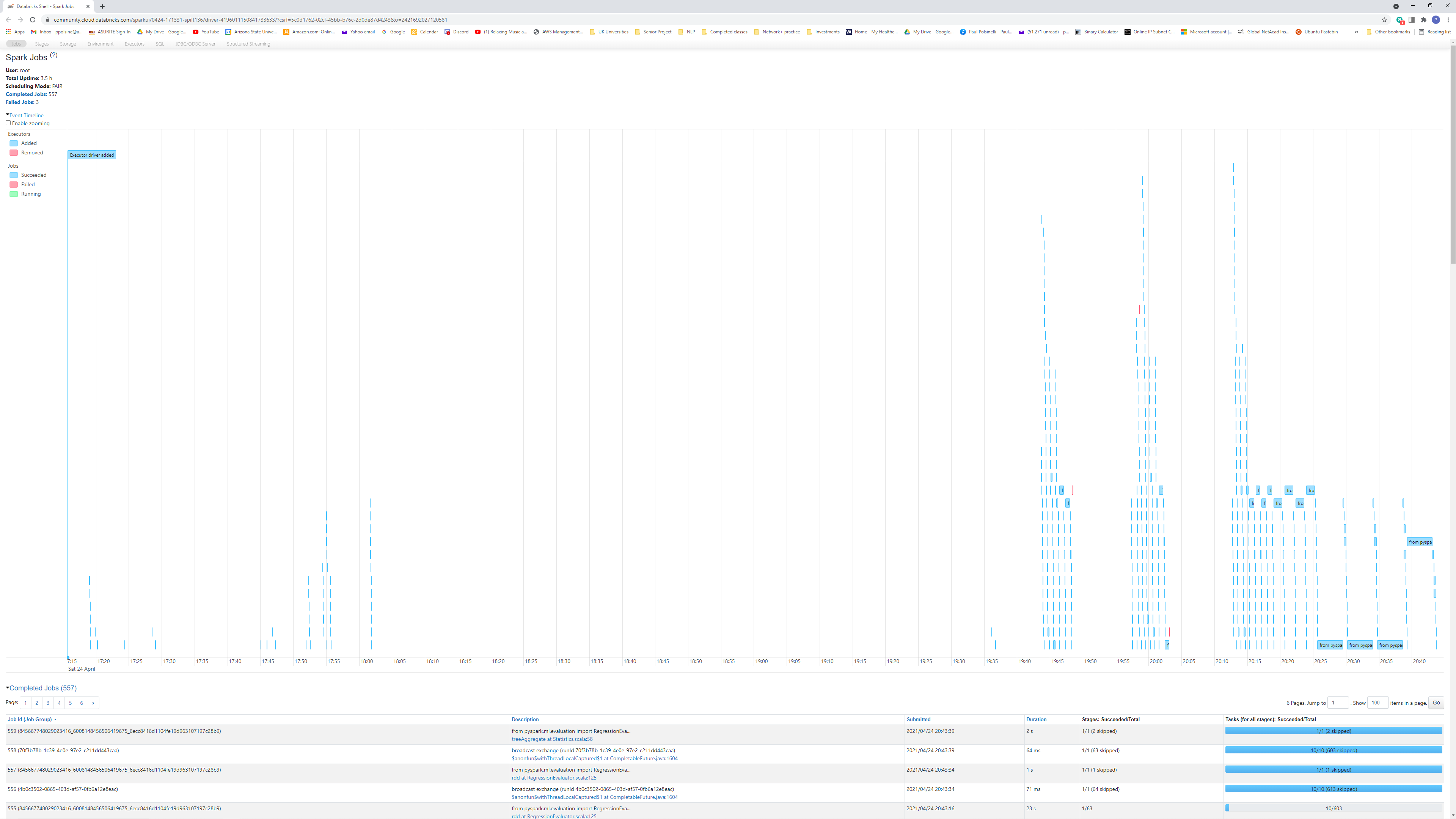
**How many blocks was the RDD holding the dataset divided into?** 10

### **Part II: Automation of Model Training (Experimenting with Different Parameter Values)**

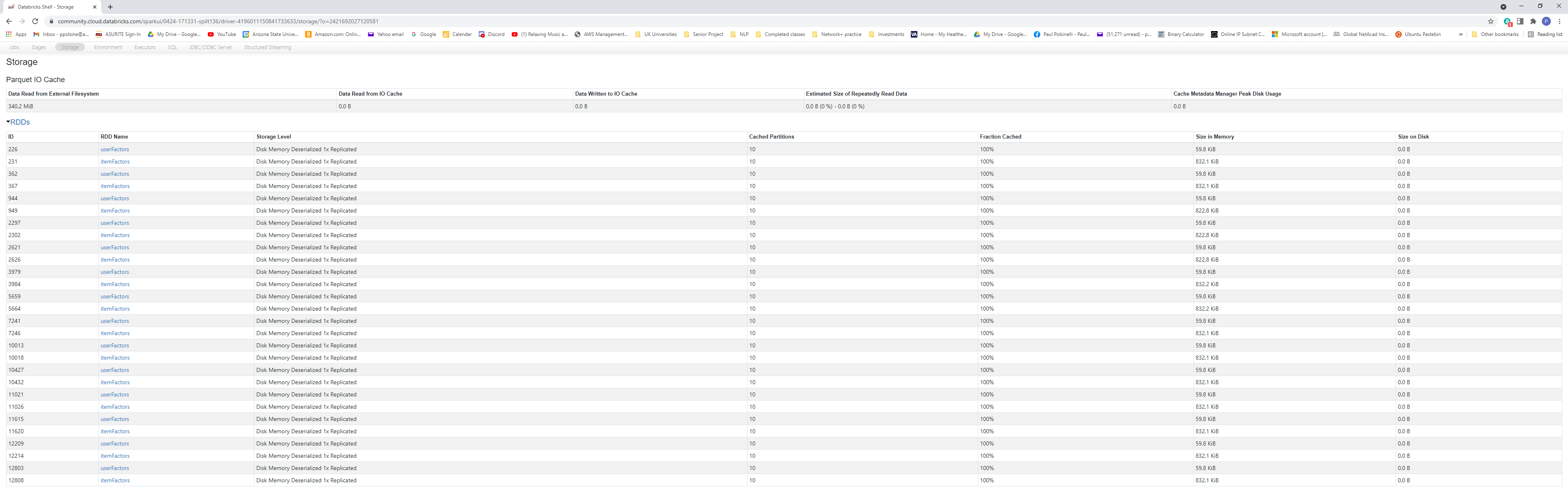


The lowest RMSE is 0.882987468979498 when maxIter = 30 and regParam = 0.1

**How much time did your code take to run?** Command took 31.09 minutes

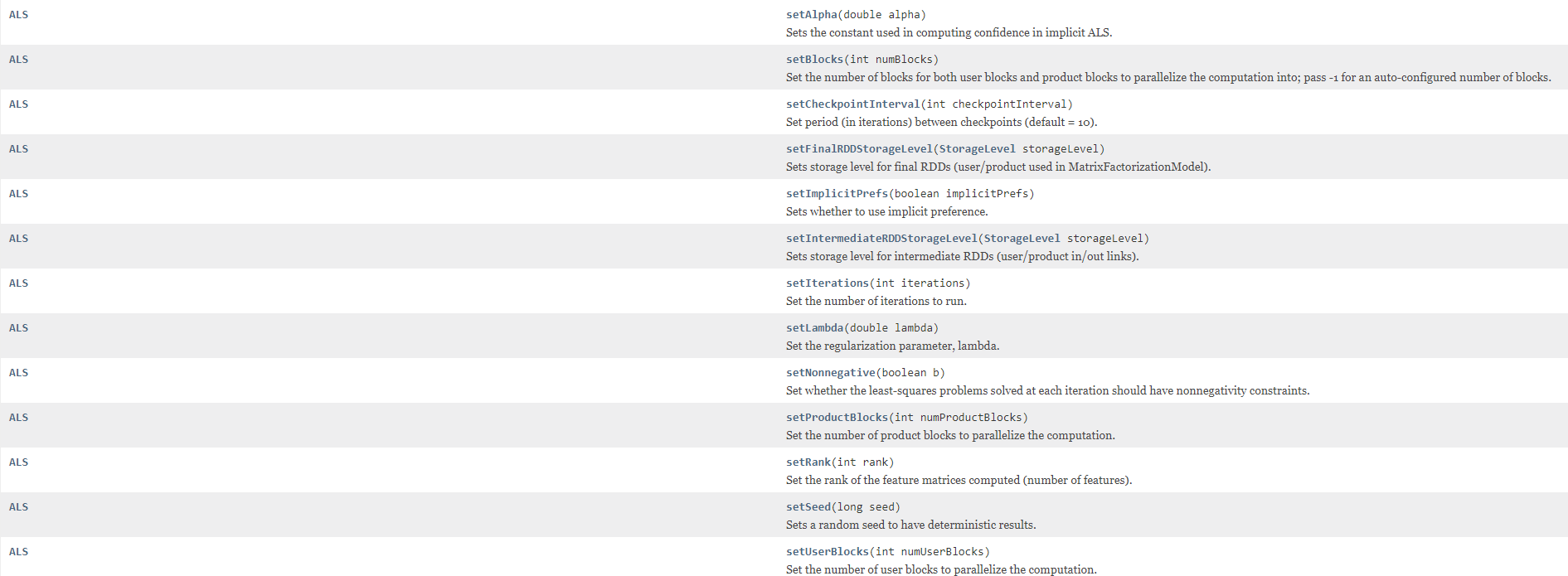


**How many jobs did Spark submit to run your code?** 557



**How many blocks was the RDD holding the dataset divided into?** 28

**Spark ALS parameters, all set functions of the form setXYZ()**

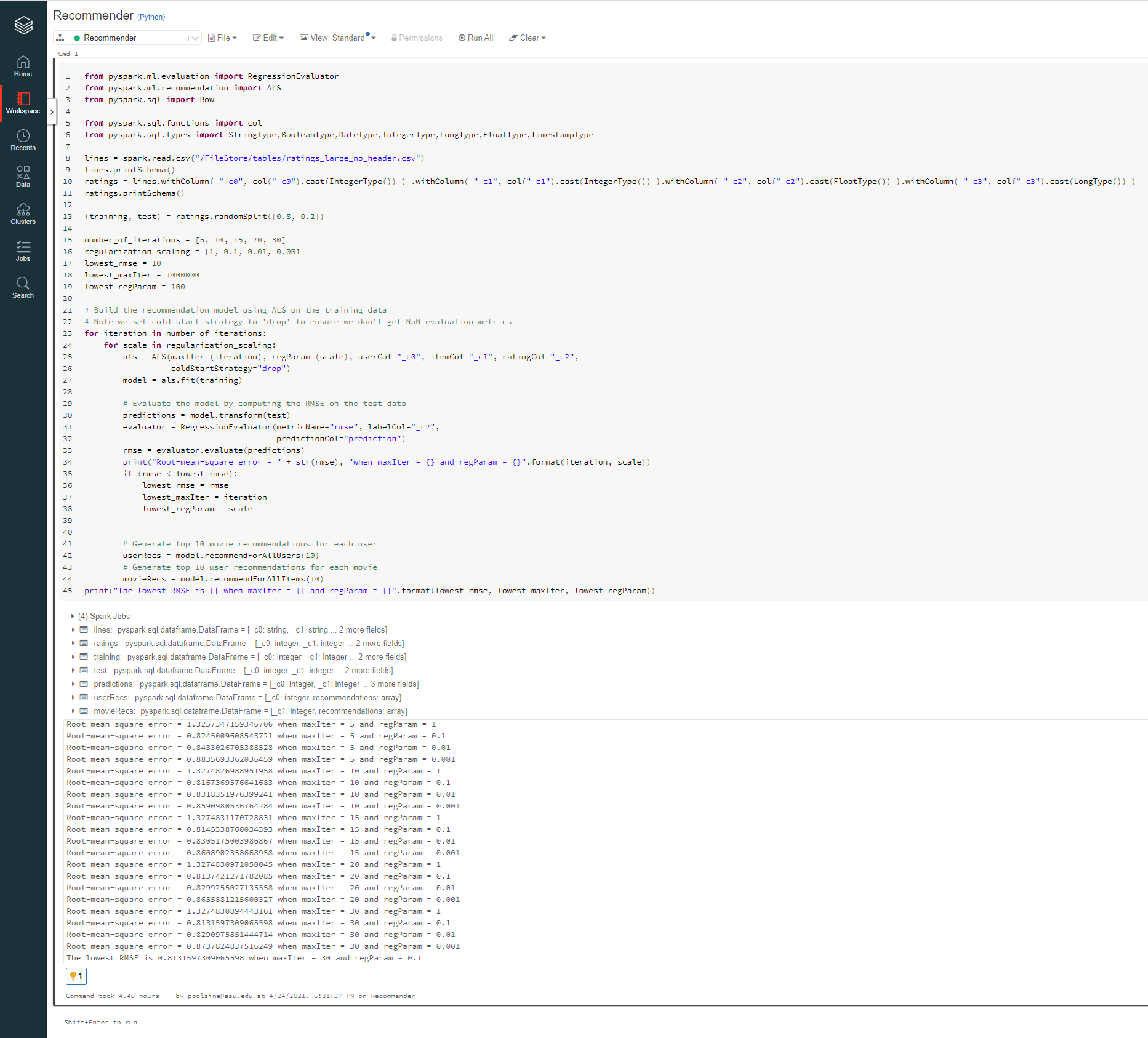


### **Part III: Experimenting with a Big Dataset: The MovieLens 20-Million Rating Dataset**

**What is the number of users, movies and ratings in this dataset?** 238,228 users, 58,098 movies, and 27,753,444 ratings

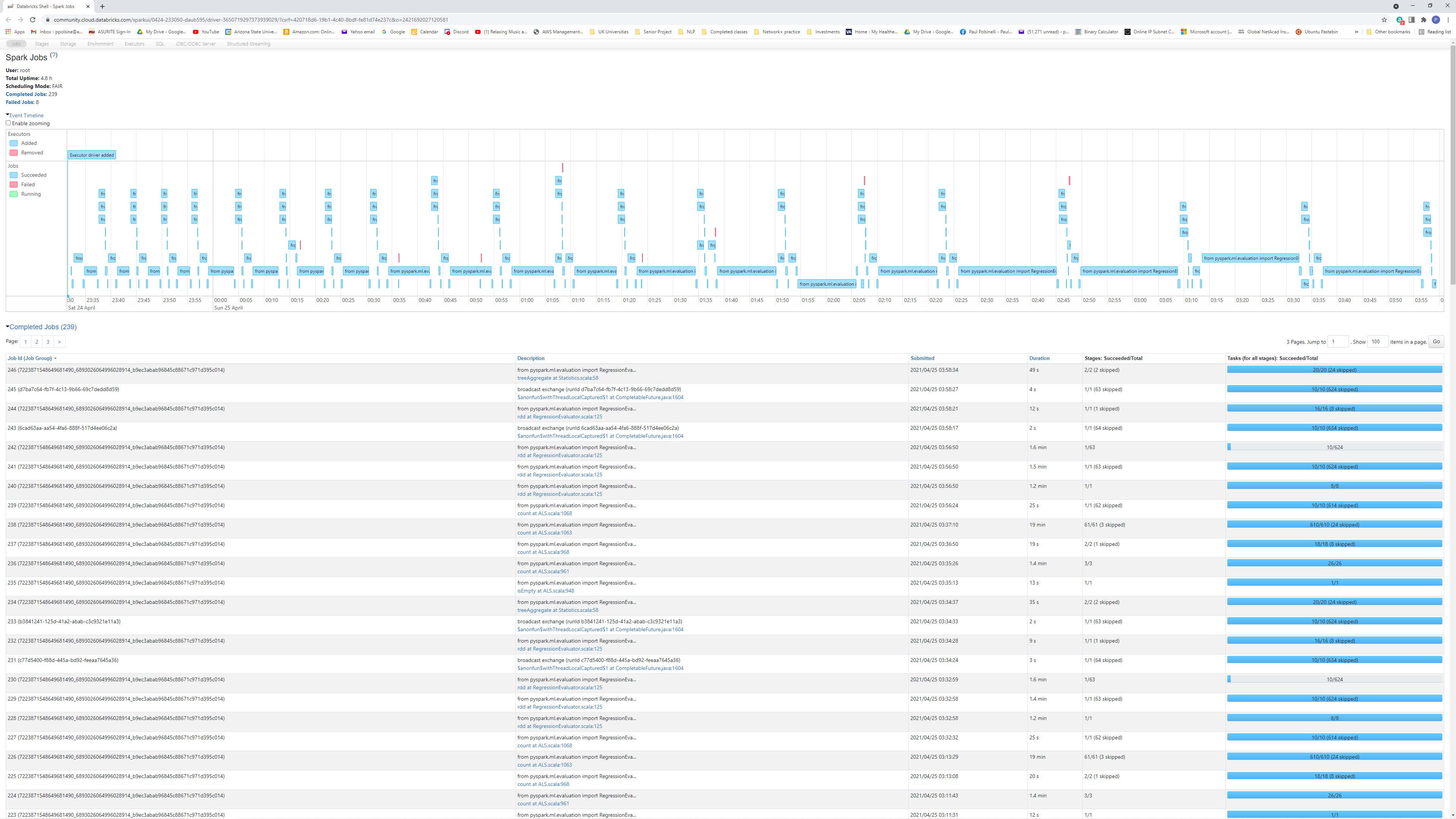
**From these numbers, specify the number of data points, the number of data attributes, and the number of existing values in the dataset.** There are 27,753,444 data points, 3 attributes, and 83,260,332 values; (not including timestamp).

**What is the number of missing values in this dataset?** No missing values are indicated

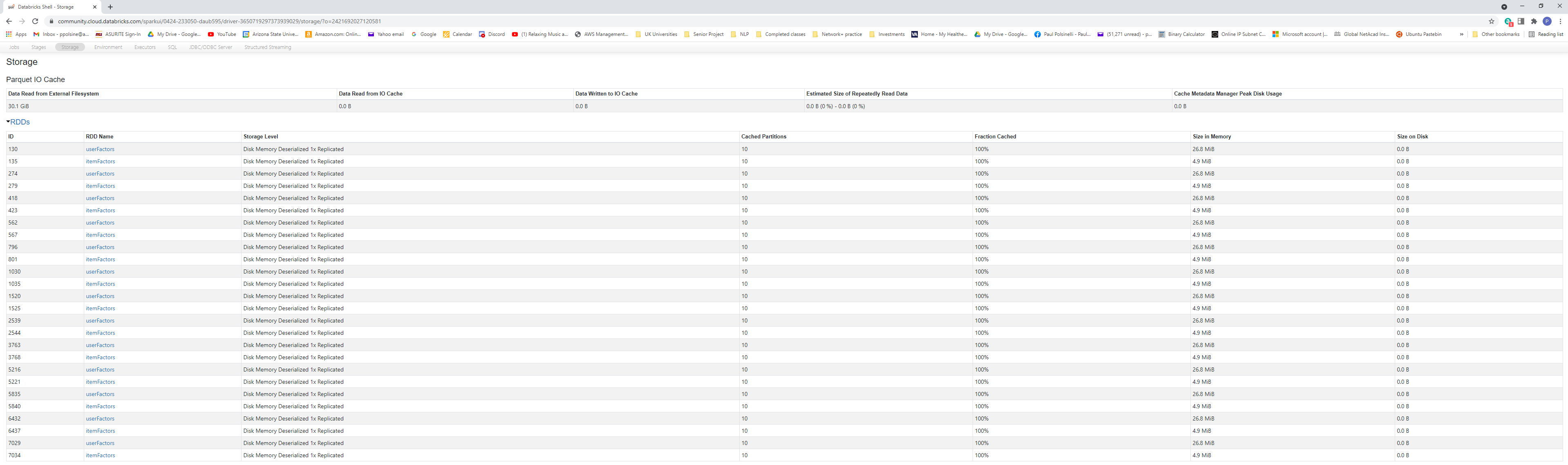


The lowest RMSE is 0.8131597309065598 when maxIter = 30 and regParam = 0.1

**How much time did your code take to run?** Command took 4.46 hours



**How many jobs did Spark submit to run your code?** 239



**How many blocks was the RDD holding the dataset divided into?** 26