Whatever-OZS Project Update

Data Preprocessing

Tasks Achieved in Data Preprocessing:

- Hourly PageViews Collected
- MapReduce Scripts Coded and Tested
- AutoDownloading Scripts*
- EMR Clusters Scripts*

Integration between Storage, Computing and Analytics Services

* These Scripts could be used for the automatization of the Data Preprocessing.

Peak Pages Finding

Once we have all the data, the next steps will be as following:

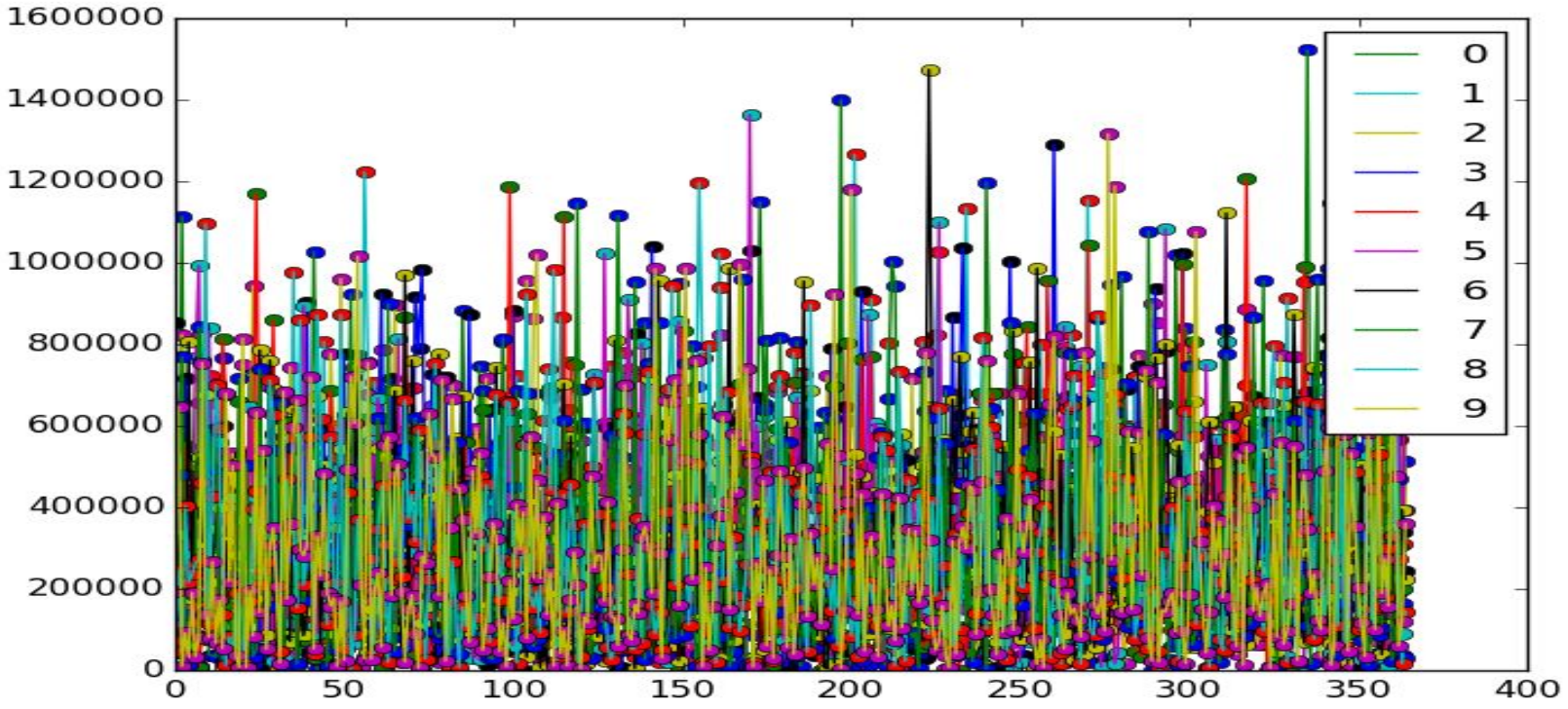
First Step:

For each day, we will find all the pages that reach the peak at this day. The peaks are simply the most page views in a 15 days interval (the day we chose + former 7 days + later 7 days)

Second Step:

After finding all the pages that reach peaks at a day we chose. We will later implement a score function to give each page a score, and then rank these pages by their scores.

Ten pages' year curves

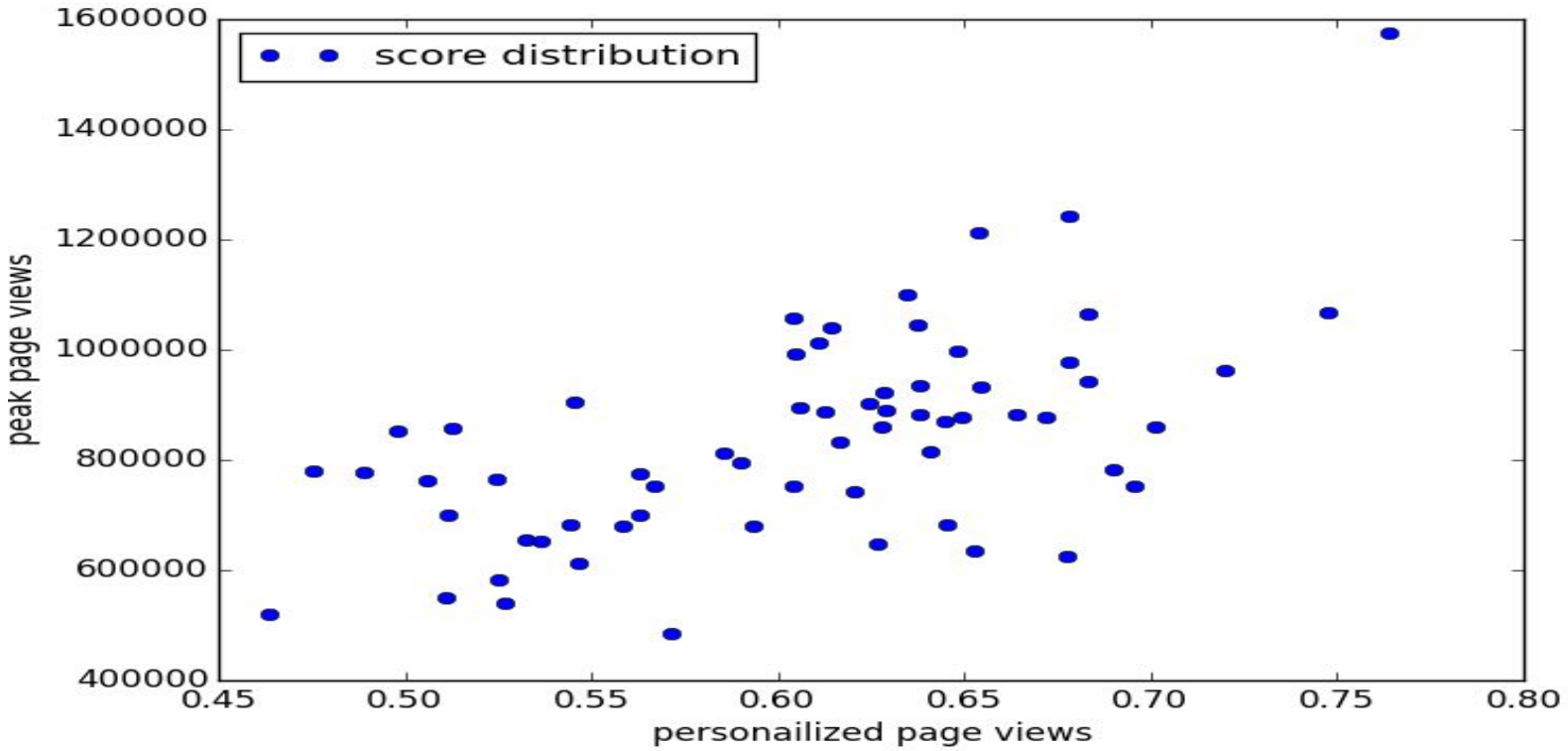


Score function

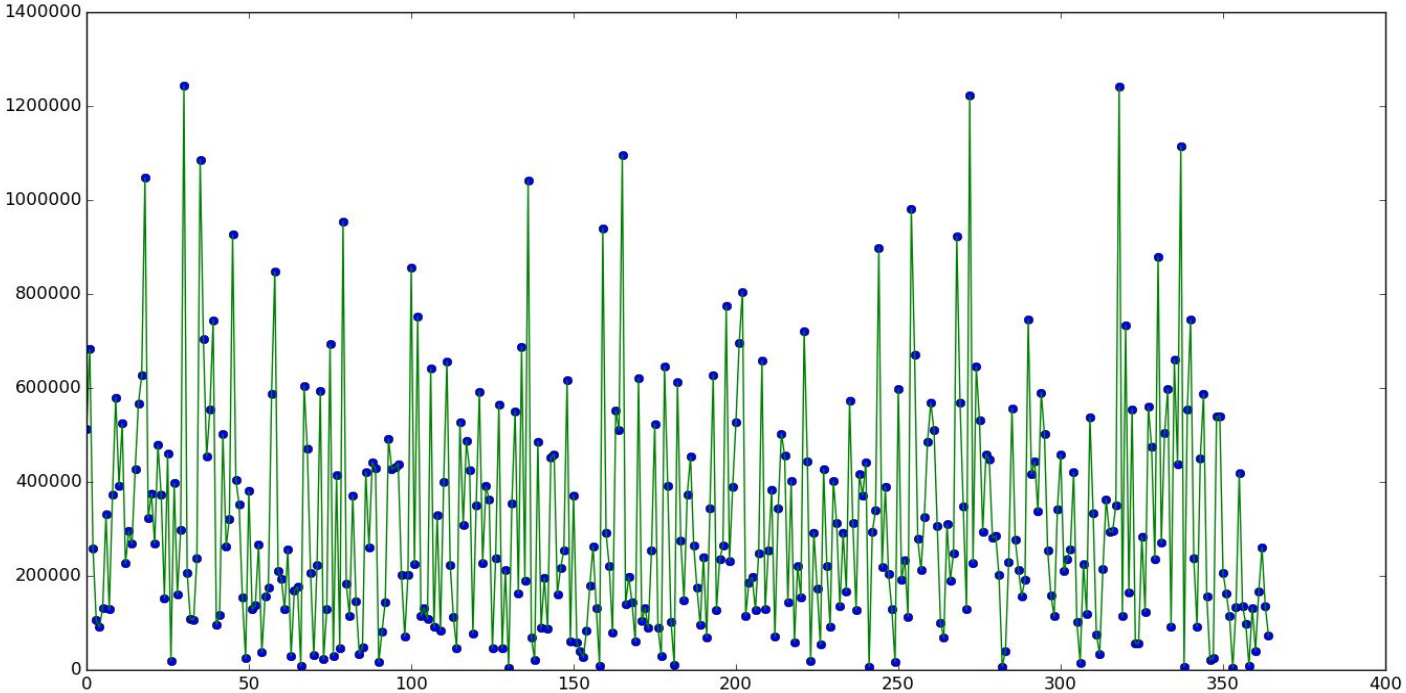
$$personalized\_page\_views = \frac{peak\_page\_views - average\_page\_views}{peak\_page\_views}$$

$$Score = \lambda_1 \times peak\_page\_views + \lambda_2 \times personalized\_page\_views$$

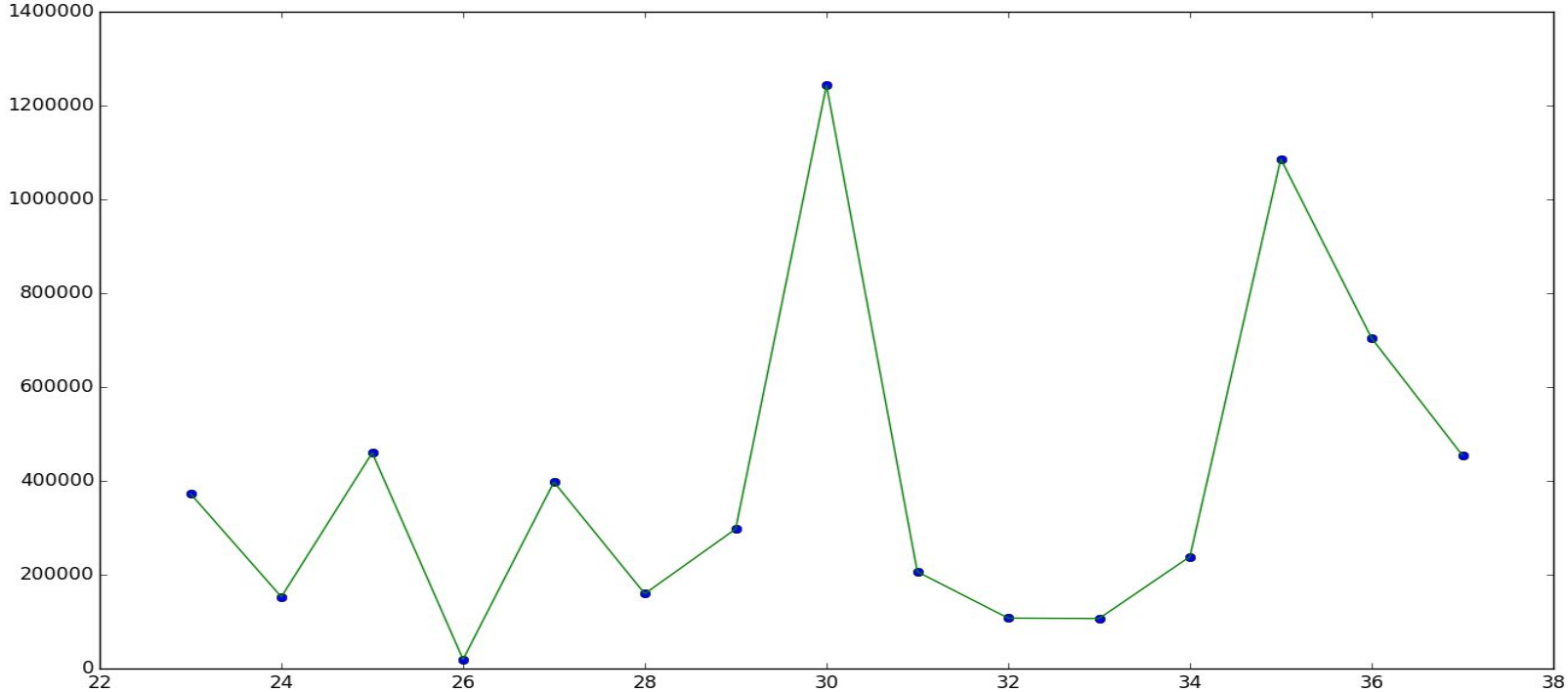
For January 30th, the top 50 peaks we find



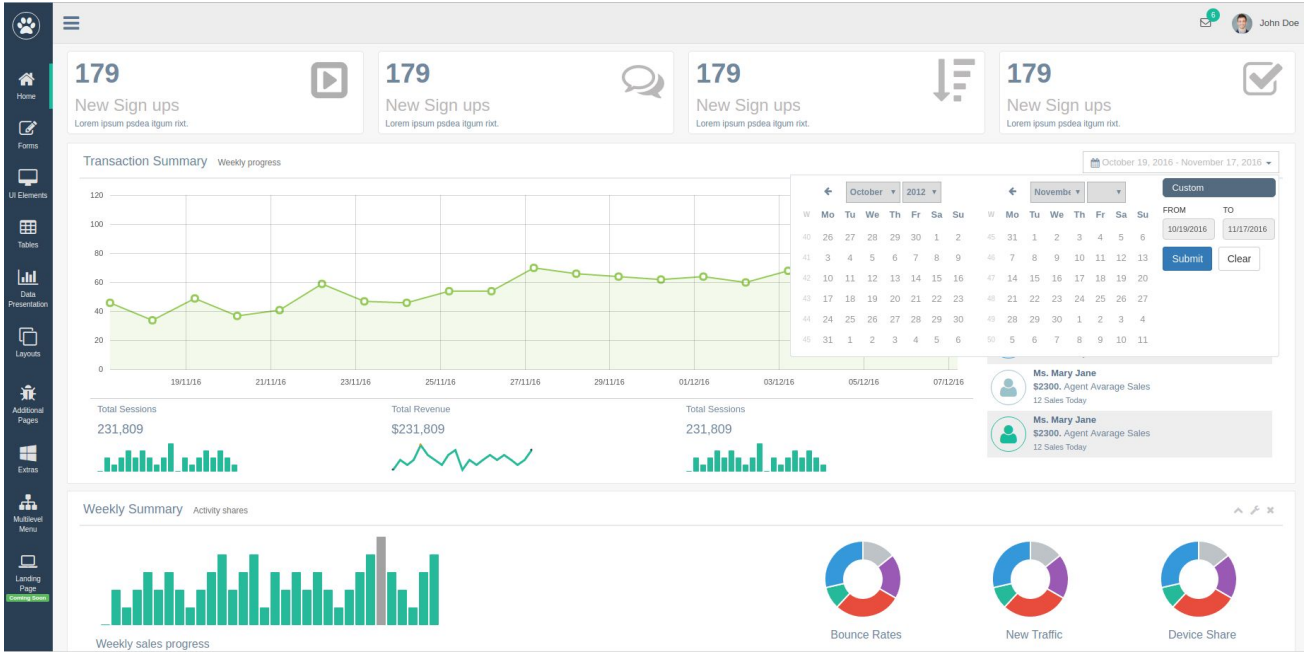
The January 30th highest score page's year curve



The curve within 14 days for January 30th's highest score page



HTML Dashboard



HTML Dashboard

The Dashboard will allow the user to:

- Select a specific date from the year.
- Obtain the Top-N peak articles for the selected date.
- View the daily views for one of the Top-N peak articles in a specific time window.
- Compare the daily views of that Top-N peak article with its other language versions.

The Dashboard will be hosted in the AWS, so it can be consulted by any person.

To-Do's:

- Apply Score Function to the English Wikipedia articles for all days available.
- Storing Scores in DB.
- Implementation of Dashboard Interface.
- Connection of DB to Dashboard (for user queries).
- Deployment of System.