Hi guys, our project is called “Reputation First”, a public opinion research based on media articles.

As we all know, Facebook is experiencing the largest reputation crisis recently, leading to millions of losses. Today, the public reputation matters for every company. That’s also why we want to dive into social media data and try to predict reputation from them.

We use distributed internet spider to crawl articles from Reuters, NewYork Times and Wall Street Journals, after which we use Spark to clean data and apply textrank algorithm to extract the summary from original text. In the next step, python NLTK is a good solution to handle sentiment analysis and extract media attitude. Finally, these messages are combined to predict the reputation score.

One of the largest challenges is how to process 2GB data and 600 Million tokens efficiently in 3 different tasks.

Faced with the big data challenge, we use a spark cluster with 8 cores to accelerate the workload from 73.9 hours to 9.8 hours.

In this user case, we find a reputation increasing of facebook in Apr.29, the reason of which is facebook begins a corporation with UK government in data protection.

Now, We are looking forward to extending this project into a realtime analysis tool in the future. Thanks for your attention!