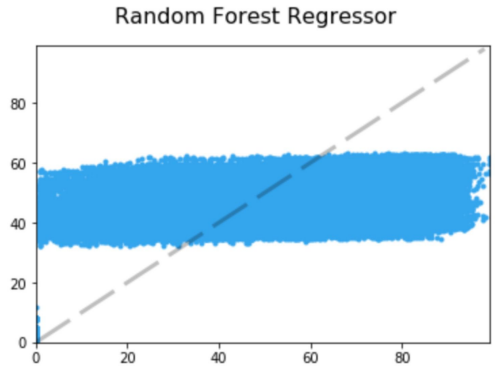
Yifan's P3 retrospective

You've now had the chance to work with both MapReduce and Spark. In your opinion, what are the pros and cons of both?

During working on these two big data frameworks, I find Spark is much faster than MapReduce. However, the RAM size is the limitation of it. In case the resulting dataset is larger than available RAM, Hadoop MapReduce may outperform Spark. MapReduce allows parallel processing of huge amounts of data. If the speed of processing is not critical, MapReduce is also a good choice.

Was there something that you thought would be easy to implement in Spark but it turned out that it wasn't?

When we work on the ML question, it always throws out-of-memory exceptions. Somethings if the features we choose is not good enough, the result will be really bad. Such as



Were there any confusing or surprising aspects of working with Spark? Did you come across some functionality that made your life easier or the computations run faster?

Pysqlite is really easy for us to use. We can just use some SQL code to solve some analysis problems.

Give a rough estimate of how long you spent completing this assignment. What part of the assignment took the most time?

We have a team consists of 3 members. I worked on analysis 1,2,3,5 and option questions. It takes me three days, around 20 minutes. After that, all of the team members work on the ML question. This question takes so much time because the training result is not good enough. We change features many times.

What went well?

Pysqlite is really easy to use, and spark is much faster than MR.

What didn’t go well?

In my opinion, our ML result is not good enough. Although we have changed several features, the Root Mean Squared Error is 7. If we add more test case, it will output out-of-memory exceptions.