Answers to the questions

1. Similarity Calculation

Pearson correlation coefficient is a measure of the strength of a linear association between two variables. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables and the Pearson correlation coefficient indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

I chose Pearson’s similarity because of the following reasons:

* It quantifies the similarity between two users with the range -1 to 1.
* The Pearson correlation does not take into consideration whether a variable has been classified as a dependent or independent variable. It treats all variables equally.
* Pearson’s correlation coefficient does not indicate the slope of the best fit line. It simply means that there is no variation between the data points and the line of best fit.
* Another reason is that the two variables could be measured in different units as well.
* The cosine similarities of a subset of the original data are the same as that of the original data, which is not true for the Pearson correlation. This can be useful when clustering subsets of the data.
* Also Pearson correlation is invariant to scaling, which was very useful in our case as one user might be a scaled version of another user.
* Under this method, we can also ascertain the direction of the correlation i.e. whether the correlation between the two variables is positive, or negative.

1. Handling Large Data

* I would consider using Big Data platforms such as Hadoop or Spark that allows us to use data, transform it and run machine learning algorithms on top of that.
* I would also consider using Amazon **EC2 (Elastic Compute Cloud)** instances that provide access to scalable computing resources on the cloud and use Services for Data Transfer provided by AWS to migrate the data from local data center to the cloud and use scalable hardware resources.

1. Additional detail

In addition to calculating the similarity scores, I would consider building a recommendation system on top to recommend courses to the current user based on the behavior of similar users and also use the Restful API to give a list of suggested courses.