Questions

1. What architecture do you suggest that will support those requirements?
   1. We could lunch the Spark jobs on top of the AWS EMR
   2. The Spark job aggregate last day events to three aggregated data: activities, modules and unique users
2. How will you schedule this computation?
   1. AWS Data pipline - can schedule EMR jobs
   2. Run a server with aka quartz scheduler with mySql to keep the job state
3. How will you do the actual computation?
   1. The computation will done on
4. Where will you keep the results?
   1. On S3 bucket. Per day and by date
5. How will you make sure the system can handle up to 1 year of data in a timely manner?
   1. Since the data will be save in s3, parquet and partitioned I don’t think it will be a issue
   2. More over, the daily aggregation process can aggregate the data per day and for the last 3, 7, 14, 30, 90, 180, 365 days. And the querying should be faster