Tell me about yourself:

介绍一个你最喜欢的项目？介绍下简历？介绍下自己？

有没有看职位要求？说说职位要求要找什么人？你是这样的人么？介绍一个你最符合这个职位要求的项目，最后强调你是good fit

1’

My name is Enda Peng, I graduated from Boston University, major in Computer Science. I’ve been working at Two Sigma more than 3 years. I am in the team Modeling data Engineering, my team focus on onboarding external data to research and production environment. More specific, we build ETL libraries and pipeline targeting on large scale data ingestion and processing. I have great passion on solving technical challenges and enjoy the sense of achievement when my work brings benefit to my customers. That’s my motivation of applying the role here, I would expect to embrace bigger challenges and make larger impact at [company name]

你最大的缺点/失败？

翻译：你一个无伤大雅的小缺点/失败是什么？你从以前的哪个项目知道自己有这个缺点/失败？知道以后学到了什么教训？在后面哪个项目中吸取了这个教训，做了什么，取得了什么结果？

I think it is the courage to say NO, or say to be smart on pushing things back. As a junior in two sigma, I only have to work with my mentor, do whatever he assigns to me. But suddenly he was promoted to my team manager and I have to take over the entire project. I think most junior has similar experience, I’d like to show muscle to my customer and manager, I take whatever requests my customer has, but soon it results in a situation that I complain to my manager about overwhelming work, my customer complains about some key results were not delivered in time. That is a bad case.

The takeaway for me is that work is endless, my goal shoudn’t be getting everything done, it should be the maximum overall satisfaction, to optimize my time ROI (return over investment). I start to pay attention to our sync up meeting, try to understand each work’s value and the approximate time investment. Try to reach to an agreement before start and share it with stackholder. The situation is greatly improved, my customer knows the progress and priority so that they can plan their work, my manager knows my plan and difficulties and indeed I get less work and can be more focus.

你最大的优点？

翻译：我知道你很牛，你哪个特质最符合这个职位的要求，并且在最后强调你的某某优点让你是一个good fit for this position

I always keep refreshing my technical knowledge and I am eager to bring them to my team in order to increase our team’s productivity and our service reliability. In the past three years, I spend most of my time on cloud upgrade so that I get the chance to expose to most of the latest technologies. For example, oh feel free to interrupt me if you’d like to further discuss any of the projects, technologies I mentioned, in data-over-cloud project – the project which targets on combining TS private cloud with public cloud, I bring BigQuery, AWS S3 as alternative storage. In addition, kubernetes for computation-over-cloud and CI for general development cycle. Those effort cuts my team budget, increased our service reliability, computation capacity and reduce our support burden.

Compared with other fund or companies at financial industry, I think Two sigma is very friendly on embracing new technologies, but obviously can’t match tech companies. I’d be very happy to join a company which not only using but also have the chance to create those industry leading technologies.

为什么选我们这个公司？

翻译：公司的mission是什么？我的career goal和你们公司的mission完美契合；职位的要求是什么？我的背景和能力和这个职位的要求完美契合。最后强调你是good fit

[Facebook]: First, from the culture perspective, I like facebook’s culture, break things, move fast and make impact. Actually, I interviewed facebook three years ago but didn’t receive an offer, but your company principle inspires me a lot in the past three years. I am always doing the second and third point at two sigma except “break things”, that’s a terrible word at two sigma, our mission is keeping the money printer running forever, don’t break it. Just kidding, ok, so my principle and company culture is perfectly matched, I believe [company name] is the right one.

Second, it is the mission of company, I am looking for my next employer where my work will have impact over millions of people and bring them happiness. [describe some apps, services for this company], on Ins I feel happy when my post receives a lot of like, on Fb, I can easily find those people who share the same interest with me and I make a lot of friend by using fb.

Third, I know that facebook hires the best engineers, give each engineer respect and clear growth path. Fb is on my list if companies where I can be proud of telling my friends that I am working there.

队友/同事不干活/很难相处咋办？

-baidu 1point3acres  
翻译：你有没有经常和队友/同事主动沟通？你愿不愿意为了团队，帮队友/同事分担一些工作？能不能以非常职业的方式解决这个问题？

- maintain good personal relationship with colleagues

- communication, peer to peer, art, resort to manager

- quantitative method to support myself

Communication, communication and communication. The first communication is on the personal relationship level, I am very active on team building activities, like team out, happy hour, I believe keeping a good personal relationship with colleagues will benefit my work.

I often schedule a friendly in person coffee chat, try to figure out why you feel hard to get clone with this guy/his throughput is low. For example, you feel hard to get alone with a guy is shy and not talkative, that’s totally fine, next time when he meets any difficulties, just be pro-active to ask him whether he needs help or not.

The second is be mindful on any verbal communication, making friend is not my ultimate goal, my goal is to drive project success, so I would always make all of us clear about what we are going to do and what is the deadline.

I do meet a trouble case previously, once we had a new hire. I was assigned to coach his ramp up project. He discovered an unreasonable set up and I explained the legacy reason. I would like him to ignore it but he insists on fixing the issue completely. I was very patient at that time, I help him trace the issue back to the the code written by big boss on 2001 and tell him that the effort it too huge compared with the benefit. I think he would give up his idea after it, but he said he doesn’t like it and still wants to fix it.

I realized that if I keep the conversation going, we may miss the deadline of the project, I have to remind my manager of the potential risk at early stage and get more voice involved to help me. I wrote an email to our team senior tech lead and cc our manager, in the email, I fully acknowledge his solid work, then put our conversation in the thread and ask for input from them. I acknowledge his work because indeed his strict attitude deserves encouragement, but in practice I hope he can give up right now and catch project ddl, that’s why I resort to teach lead and manager for help.

介绍项目

Make Python a production support language for modeling computation purpose.

This is the first time that I drive a corss team project without clear goal. I have to what are the problems, what is the best/suboptimal solution and what is the key result to deliver.

First of all, we are a java firm for many years, but recently python shows increasing popularity among our research community, that’s the motivation for us to provide production support for python.

The first challenge is that python itself changes so fast, so does other fundamental packages like pandas, numpy etc. Also our researchers use conda to set up their researcher environment, which is a disaster for version control, you can image that whenever a modeler submit am model, we have to preserve its conda configuration and rebuild it in production in order to fully replicated his result, it also brings security and compliance problem.

To address this problem, I argue with research community that it is engineering team’s responsibility to provide a cross firm consistent python environment and we should stop using conda to prevent the flying packages. The we developed so called mono-repo python, which is fully version controlled, centralized python. With mono-repo python, modelers no longer need to build their own virtual environment and enginnering team doesn’t have to manage all those conda configuration. However, modelers complain about it takes longer for them to onboard a new package, you know, which is as easy as one line command in conda. This is the cost that we are willing to pay, we evaluate each request on bringing new packages to TS and give them feedback. Later we notice a problem that if we pack everything into one python core, it is extreme large. In addition, the nature is that for one team, most of the libraries are not necessary. To solve the problem, we split our core into computation fundamental which includes basic python interpreter and fundamental packages like pandas and numpy, then each team can build their own core on top of it. It solve the size issue.

However, when here comes a lot of cores, our user feel confused, then don’t know how to set up a correct core for their modeling purpose. That’s is the second goal for this project - provide our user a good experience, you know, most of our researchers are from math, physicas and stats, we can’t assume so much computer science background on them.

To provide a easy set up environment, I come out with the idea “cloud hosted”jupyter, in ts, researchers love using juypter to explore their modeling ideas. Previously they have to launch it locally and copy-past the link to browser. With cloud hosted jupyter, the just to to browser, type the app name, we allocate a host to them, our engineering team regiester our cores on this app so that our researcher just need to pick up one from the available list. After we release this feature, both modeling tools team, computation team and my team are very happy because we see more and more active users on our cloud host jupyter.

However, we notice that learning curve is still not so friendly to researchers - they know how to set up python development env, but they write hacky code to achieve their goal. To help them better coding, I designed the first worksheet or say template and integrate with cloud jupyter, the template consist of text hint and unimplemented functions, it is a jupyter notebook, so we can use mark down to emphasize something to our users. E.g write down the event name below. This template is the abstraction of our team’s service, it is the protocol between modeler and our team, my goal here is as long as you implement all the required functions here, I promise that your code can call our team service successfully.

Modeling team gives me very good feedback as their work has been simplified a lot, now their daily work is open browser, choose what service they’d like to call, fill out a form. Submit it. Other engineering team join later, they provide their own core and template to standardize the service call.

Back to engineering side, we provide simple tool to user and leave all complexity to us. First, security, for java code, our company complies java to special secure jar, similarly, we compile those work template to python file, when we run it, we load the encoded package at running time, and only call the protocol functions we defined in template. By doing this, modeler’s can call each team service without exposing their ip sensitive cod.

For consistency problem, at the time that modeler submit their work, we take a snapshot of the env, later when we deploy his code, we link the snapshot interpreter to the particular python script. It guaranteen the consistensy between what researcher see in research and production.

For backward compatibility, we introduce regression test. Each time when we upgrade something, we kick out a regression test which regenerate data for past 1 year, compare with what is in production, send researcher a report of the diff, once they give green light, the snapshot will be updated, if not, on short term, production env remains unchanged and leave enough time for reseachers to figure out why quant diffs are introduced.

Data on cloud

We are trying to get everything running over cloud. When I join the firm my first project is to migrate some pipelines to our next generation host. That was my nighmare, for example, people hard code a host name to find raw data, use specific host to keep status. All the bad smell make migration hard. So 1 year ago when we start our new round of host upgrade, we decide to do a one time investment, get everything running on cloud.

First step is getting data on cloud, once data are downloaded, ingested on our prodcution download host, it is not allowed to keep it on the local disk. We developed storage gate way to upload them to either interval time series storage or external storage like S3, bigquery, based on the data size and ip sensitive or not. This gate way has smart rounter inside, so when people access the data, they don’t have to know so much details behind it, since everything is running on cloud, no more metal host dependency.

Second step is new data processing libraries over cloud , parallel computation works perfectly over most of TS computation as we store everything as time series. We developed a framework with scala, spark to help us process data in parallel, Data is fragmented by time range, processed by the user defined function and merged back to cloud again. I know the key here is that TS store everything as time series, which makes the parallel computation model rather easy.

As data and computation libraries are all cloud orientated, which means there is no baremetal host dedpendency any more. We definitely don’t want to contribute many time on next OS upgrade, so we decide to move pipeline to container, move to k8s. We pack our code to different images and make all pipelines running on k8s. It is fantastic, k8s operation team helps us solve so much capacity, hardware issue on the early stage, all of us sleep better after we finish the migration.