- q background?

I have been working in the field of data scientist a couple of years. I have a lot of industry experience in performing data analytics, data modeling, data engineering, natural language processing in production, marketing and advertising area by taking advantage of statistical tools, machine learning, neural networks and other advanced tools.

- q more details? (most recent project)

Sure, I have done a couple of meaningful industry projects for several big companies in these years to help them to catch up on machine learning and big data for achieving their organization goal. For example, I just finished a predictive analytics project for Shell Oil company in Houston, Texas. That project was supposed to enhance oil field production and cuts cost by finding optimal well settings and forecasting equipment failures and some other potential problems.

- q data sources?

Usually, data we use are from company's database. These data is being collected for months or even years. This project computer models replicating above and below ground well behavior for artificial lift equipment. We used data collected from sensors that spanned last several years tracking oil wells in almost every major North American basin. The data included information on drill and operational data from thousands of shell's wells and hundreds of mills of low-pressure piplines.

- q what was in the standard dataset?

There are many public standard datasets like simple Irish Flower dataset, carlifornia housing price dataset. Data in these datasets are usually properly annotated without any biased. It often covers range which we are trying to predict. In the shell project, we have for example bit record, mud type, pipe size, real time pressure value, gas kick, climate information etc. these data are very useful when we need to measure the performance of the drill and predict the future output.

- q how do you supplement your clients data?

It really depends, for example if we want to know if a people wanna buy some financial asset, we can track those people who show strong willingnese to buy new iphones or new google pixel 3 phone on social media platform.

- q output of the work?

Analysis of the data really revealed critical issues with field deployed equipment. For example, in Alaska, some driller performs much worse than we expected before. It can actually achieve 70% of its standard efficiency. Also our pipline data told us too much energy was wasted in transport these oils in old piplines. Even though I am not there anymore, Some of my former workmates told me that the company may plan to invest over 1 billions dollars to change those inefficient drills and piplines based on our final project.

- q specific responsibilities you have?

As a data scientist, I was mainly in charge of data mining and data analytics part. There are some interesting job to do. Sometimes data are not clean enough, there are many uncommon and missing values, before manipulating data, I need to contact to local equipment supervisor, ask them what happended there? Once I get good data. I also need to transform structured and unstructured data collected into analytical models, sometime we need to design algorithms. implementing neural network in random forest, decision trees and logistic regression.

- q can you describe your deep learning models? CNN? LSTM?

CNN is usually used to recognize image data but LSTM is mostly utilized in continues data like text data, A CNN will learn to recognize patterns across space so it is usually used to learned to recognize components of an image first such as line and curves, and then learn to combine these components to recognize larger structures. Long Short-Term Memory (LSTM) networks are an extension for RNN - recurrent neural networks, which basically extends their memory. LSTM's enable RNN's to remember their inouts over a long period of time. Therefore it is well suited to learn from important experiences that have very long time lags in between.

- q spark / keras?

Spark is usually seen as a more accessible and more powerful replacement for Hadoop or complement to Hadoop and other technologies. Spark is a general-purpose data processing engine that is suitable for use in a wide range of circumstances. We can do streaming, processing, machine learning, integration by using spark. Keras is an open source neural network library written in python . It is designed to enbale very fast eexperimentation with deep learning neural network.

- q are you familiar with deployment side?

Yes I ham pretty familiar with the deployment side. Deployment is usually not easy, I need to help buid and use the right cloud infrastructure on the right cloud provider. Like Amazon and Google their platform fit different companies and projects. I also need to design and implement public and internal APIs for model usage in the company, team. Also I have to integrate with data piplines and cinsistently update our models. I usually assist in continual improvement of AWS data lake environment.

- q how many models have you built?

I have built linear models, random forests, decision trees, linear and logistic regression, support vector machine, clustering, neural networks, principle component analysis and I also have a good knowledge on Recommender Systems.

- q tools set / tech stack you are familiar with?

I'm familiar with tensorflow, keras and scikit-learn those machine learning tools. Also I used a lot NLP tools like NTLK, natural language tookkit, spacy which is a very powerful industrial strength NLP library/

- q at capital one - sun trust - what would you say your level of understand is…

As far as I know, they are all very successful financial service providers. They have been providing customized reliable finanical service to the public for decades. These companies usually have a very high standard of security and commit to protect customers' privavy while doing any research.

- q have you worked when the data is time series?

Yes, I have worked for Fleetcor, a global company leading in coporate payments. When I was working for them, there was tons of transaction data. These transactions happend all over the world, different time zone, and are mainly indexed by the time.

- q location?

I am now in Chicago now but I'm open to any kind of opportunites in the United State.

- q how did you get into data science?

I've been working as a data scientist since my first jo. It is called the sexist job in 21st century, so why not?

- q when did you get your masters?

I have got my master degree for a while.

- q data engineering at all?

Yes, data sicentists have to do all stuff now. Sometimes I need to gather and collect the data, stores it by myself, do batch processing or real-time processing on it, and serves it via an API to a whoever else can easily query it.

- q test you have taken?

Yes, I have done a lot of statistical test, like hypo test, A/B test before.

- q python?

Yes, I almost use python everyday.

- q finance experience?

Yes, I have worked for Fleetcor, a global company leading in coporate payments. When I was working for them, there was tons of transaction data. These transactions happend all over the world, different time zone, and are mainly indexed by the time.

- q have you used spark?

Yes, Spark is usually seen as a more accessible and more powerful replacement for Hadoop or complement to Hadoop and other technologies. Spark is a general-purpose data processing engine that is suitable for use in a wide range of circumstances. We can do streaming, processing, machine learning, integration by using spark.

- q kafka?

Kafka is designed for distributed high throughput systems. Kafka tends to work very well as a replacement for a more traditional message broker. In comparison to other messaging systems, Kafka has better throughput, built-in partitioning, replication and inherent fault-tolerance, which makes it a good fit for large-scale message processing applications.

- q producers?

The producer is thread safe and sharing a single producer instance across threads will generally be faster than having multiple instances.

- q what is the intermediate stage / stage server?

A staging server is a type of server that is used to test a software, website or service in a production-similar environment before being set live. It is part of a staging environment or staging site, where it serves as a temporary hosting and testing server for any new software or websites.

- q how do you achieve standardization of a schema?

I can turn image, text data into pure numerical data. I can try PCA to compress these data.

- q databases?

I need to make query to extract data from data bases, cloud databases everyday. Also some time I need to help the company improve their pipeline for the quality of data a transformation.

- q what do you look at?

I am looking for a chance to contribute to your company’s success if I do not misunderstand your question.

- q scale up?

I usually use scala, spark, or sparkmil specifically to handle increasing streaming data.

- q background?

As I said, I have been working in the field of data scientist a couple of years. I have a lot of industry experience in performing data analytics, data modeling, data engineering, natural language processing in production, marketing and advertising area by taking advantage of statistical tools, machine learning, neural networks and other advanced tools.

- q hobbies?

My hobbies come and go. Currently, I study, photograph, travel, ski, enjoy cooking (and eating), and follow the cryptocurrency space. In the past, I've enjoyed xiangqi, chess/bughouse, soccer, table tennis, rock climbing, shabu-shabu, and probably a few other things.

- q why are you looking?

I just finished my last project, I did it very well, so I’m looking for new opportunites