Data Analysis Research Project – Data Engineer vs. Machine Learning Engineer

In this assignment, you’re asked to create a document that describes the similarities and differences between the Data Engineer job role and the Machine Learning Engineer job role.

|  |  |  |
| --- | --- | --- |
|  | **Data Engineer** | **Machine Learning Engineer** |
| **Goals** | * Set up infrastructure for others to work on; responsible for data storage, data transportation, etc. | * Redefine raw data gathered from data pipelines * Turning theoretical data models to production level models * Build programs that control computers and robots |
| **Responsibilities** | * Create data pipelines and optimize them * Optimize data delivery system and scale it accordingly * Create data tools to help build and optimize products | * Develop machine learning models * Analyze large datasets and derive insights from them * Improve existing machine learning structure * Apply machine learning and data science knowledge to design distributed systems * Collaborate with data engineers to develop model pipelines * Communicate process to business leaders |
| **Requirements** | * Data modeling techniques * Database clustering techniques * High level proficiency in Python, R, pandas, scikit-learn, jupyter notebooks. | * High level proficiency in Python, Java, R, C++, C, JavaScript, Scala, Julia * Communication skills * Skills for model implementation tools like TensorFlow, also tools for model deployment like Microsoft Azure, Amazon SageMaker, Google Cloud ML. |

**First, identify, read (and remember to cite!) at least two great papers, articles, or podcasts that describe these roles. You should create a single chart with a synopsis of what you believe best describes the two roles.**

Data Engineers goals:

* Set up infrastructure for others to work on; responsible for data storage, data transportation, etc.

Data Engineer responsibilities:

* Create data pipelines and optimize them
* Optimize data delivery system and scale it accordingly
* Create data tools to help build and optimize products

Data Engineer requirements:

* Data modeling techniques
* Database clustering techniques
* High level efficiency in Python, R, pandas, scikit-learn, jupyter notebooks.

Machine Learning Engineers goals:

* Redefine raw data gathered from data pipelines
* Turning theoretical data models to production level models
* Build programs that control computers and robots

Machine Learning Engineer responsibilities:

* Develop machine learning models
* Analyze large datasets and derive insights from them
* Improve existing machine learning structure
* Apply machine learning and data science knowledge to design distributed systems
* Collaborate with data engineers to develop model pipelines
* Communicate process to business leaders

Machine Learning Engineer requirements:

* High level efficiency in Python, Java, R, C++, C, JavaScript, Scala, Julia
* Communication skills
* Skills for model implementation tools like TensorFlow, also tools for model deployment like Microsoft Azure, Amazon SageMaker, Google Cloud ML.

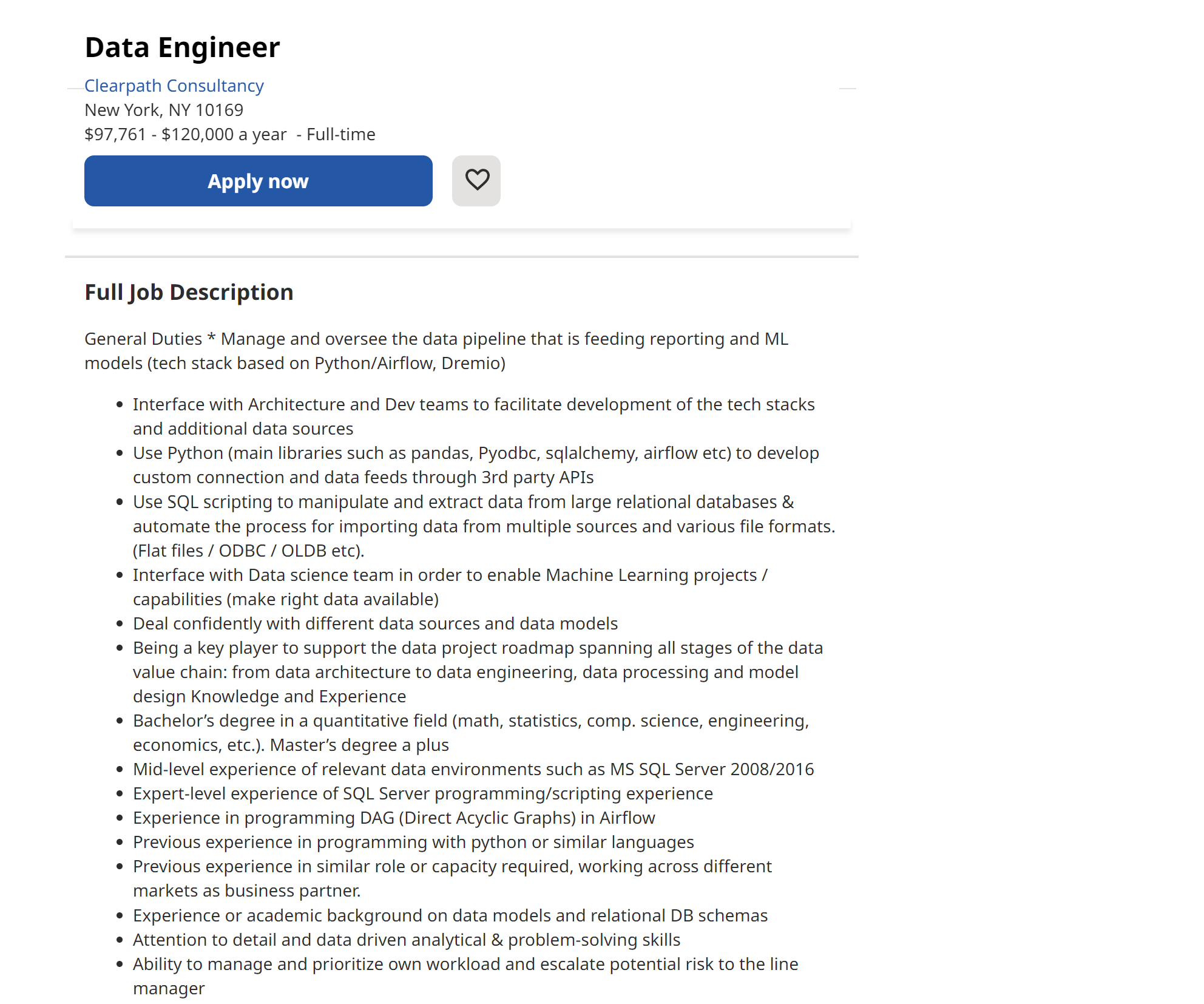
Chart:

Reference #1: ‘Data engineer vs Machine learning engineer vs Data science’, https://sennalabs.com/th/blogs/data-engineer-vs-machine-learning-engineer-vs-data-science

Reference #2: ‘Data Engineering vs. Data Science vs. Machine Learning Engineering’, <https://www.coursereport.com/blog/data-engineering-vs-data-science-vs-machine-learning-engineering-with-springboard>

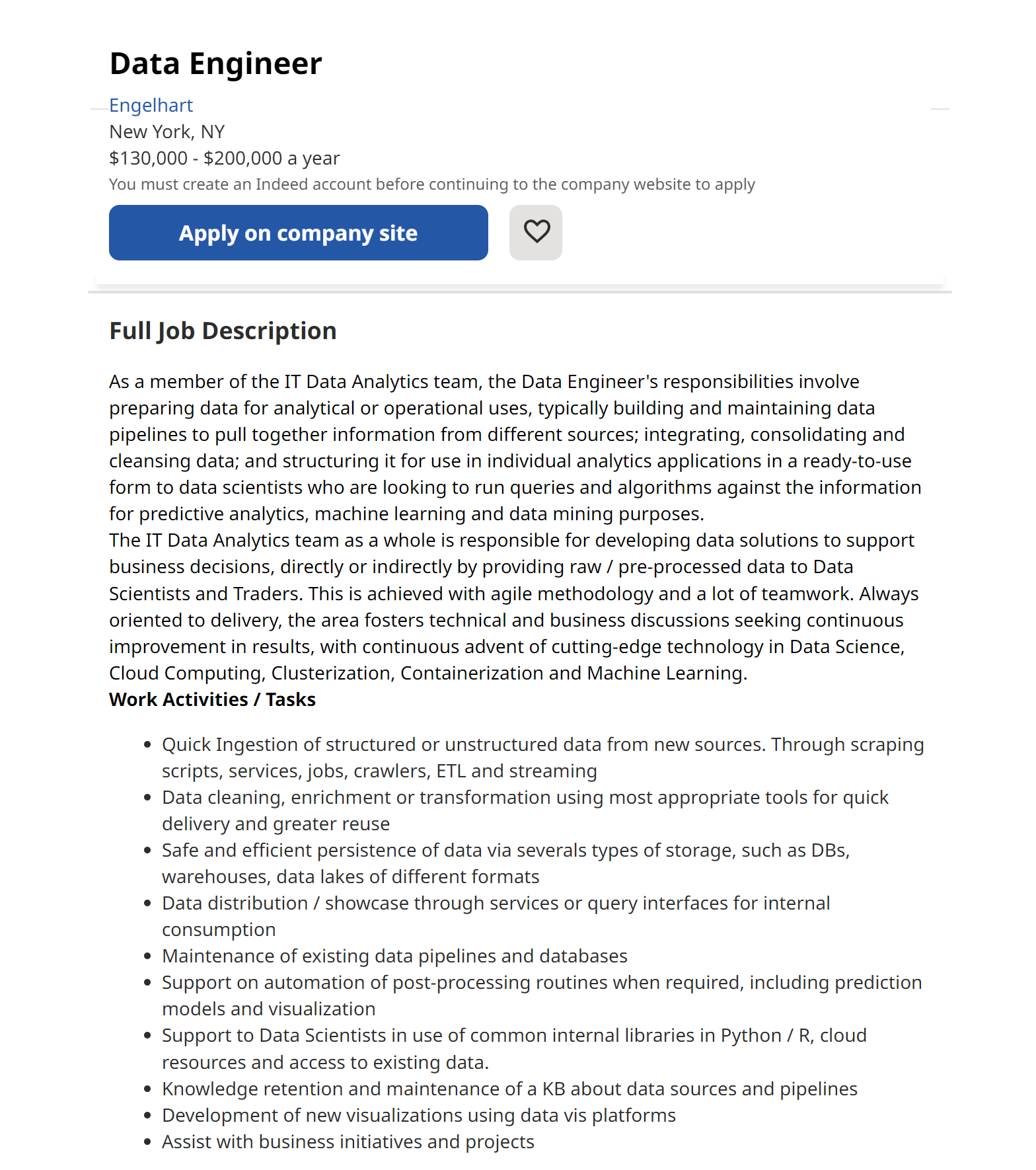
**Next, go to a job board like indeed.com, then pull down and read job requirements for at least three examples of each role. Do the posted job requirements map to what you wrote in the synopsis. Where there is a difference between the job requirement postings and the synopsis of your research, explain which you believe to be correct and why.**

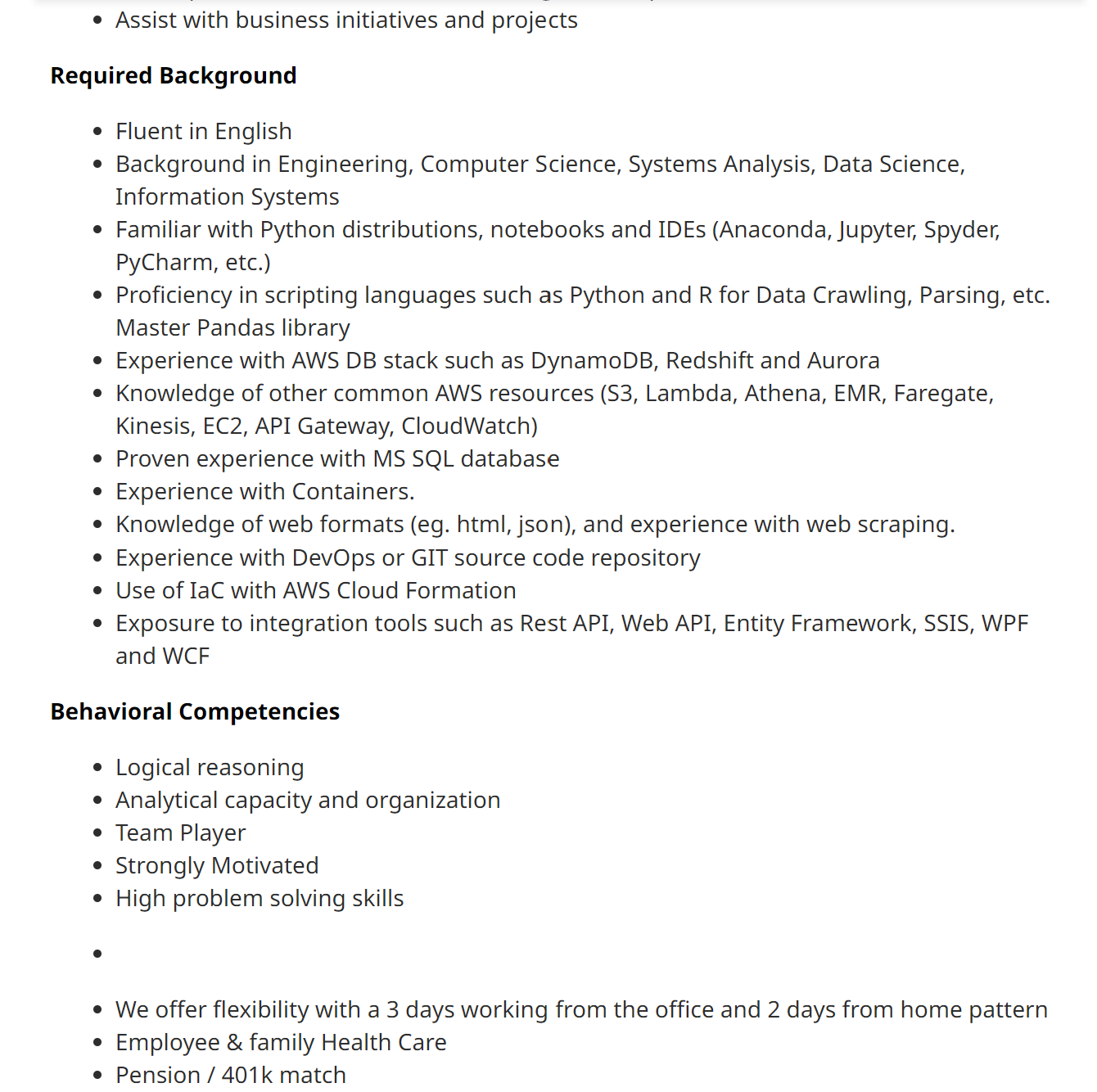
Data Engineer Example #1:



This job requirement posting does map to what I wrote in synopsis.

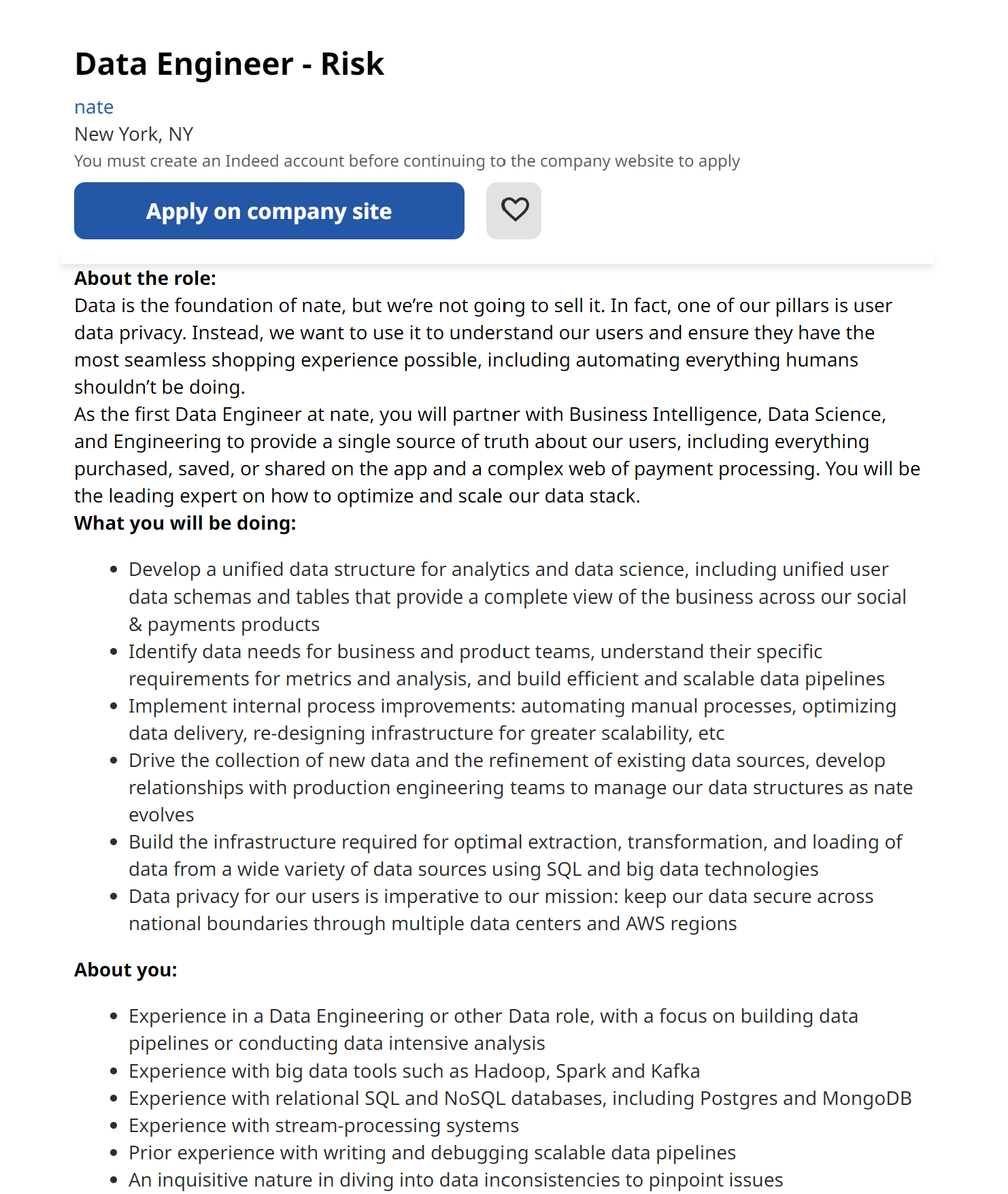
Data Engineer Example #2:

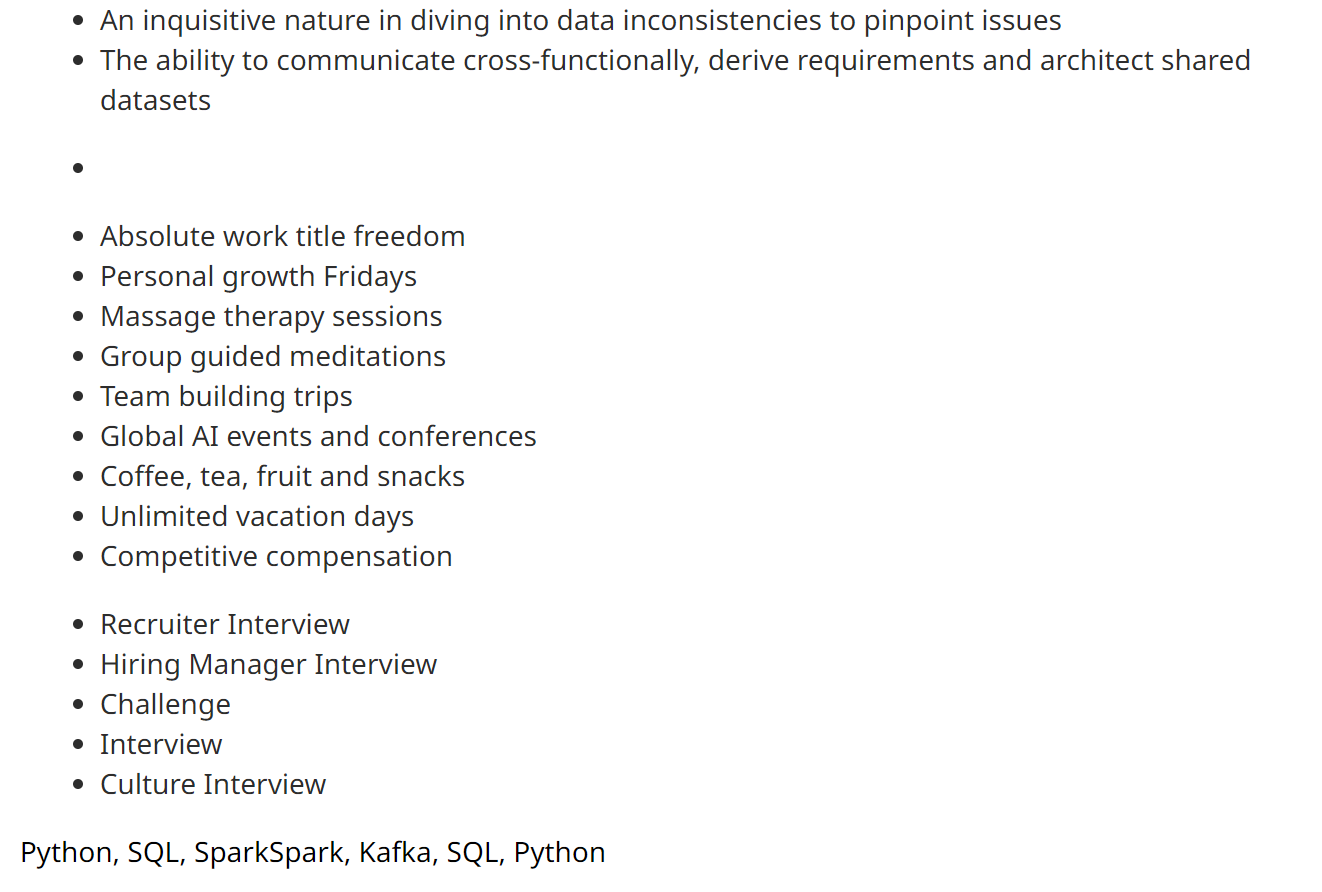




This one mostly map to what I have in the synopsis, while specifying data cleaning responsibilities in the job description. This is not a contradiction to my synopsis as most of what I wrote there are based on the assumption of cleaned datasets, and I should add this in my synopsis as well.

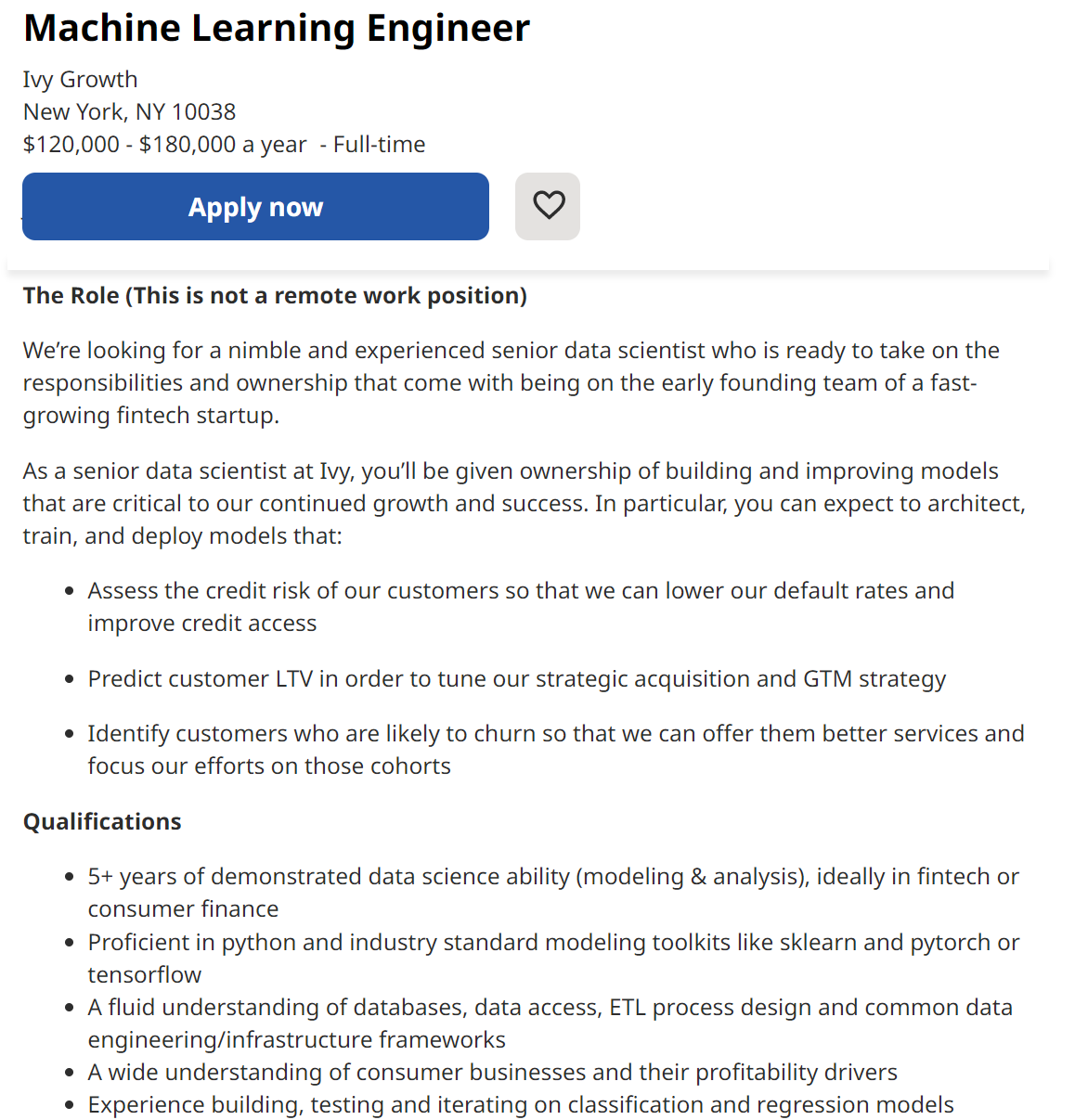
Data Engineer Example #3:

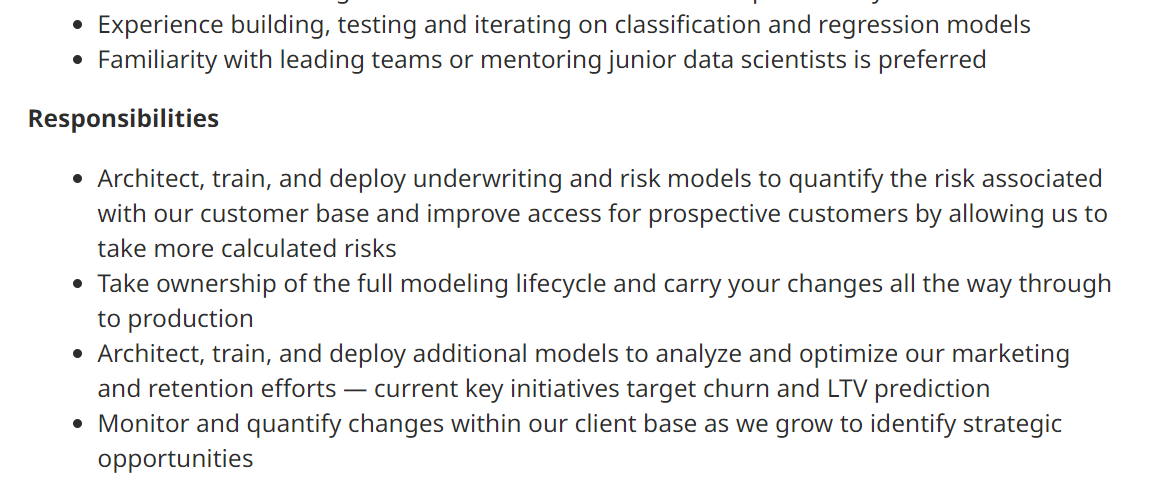




This one is a bit different than what I wrote in the synopsis, it asks its candidates to identify data needs for business and product teams, which to me doesn’t seem to be the right work to be assigned to a data engineer, this task could be assigned to the data department, but would be more fit for someone with a different title than engineer. The requirement for data privacy also seem to be more fit for cyber security department, overall this job posting seems to be looking for candidates who can not only do data engineering tasks but also few other tasks that are not traditionally categorized as data engineer’s duty.

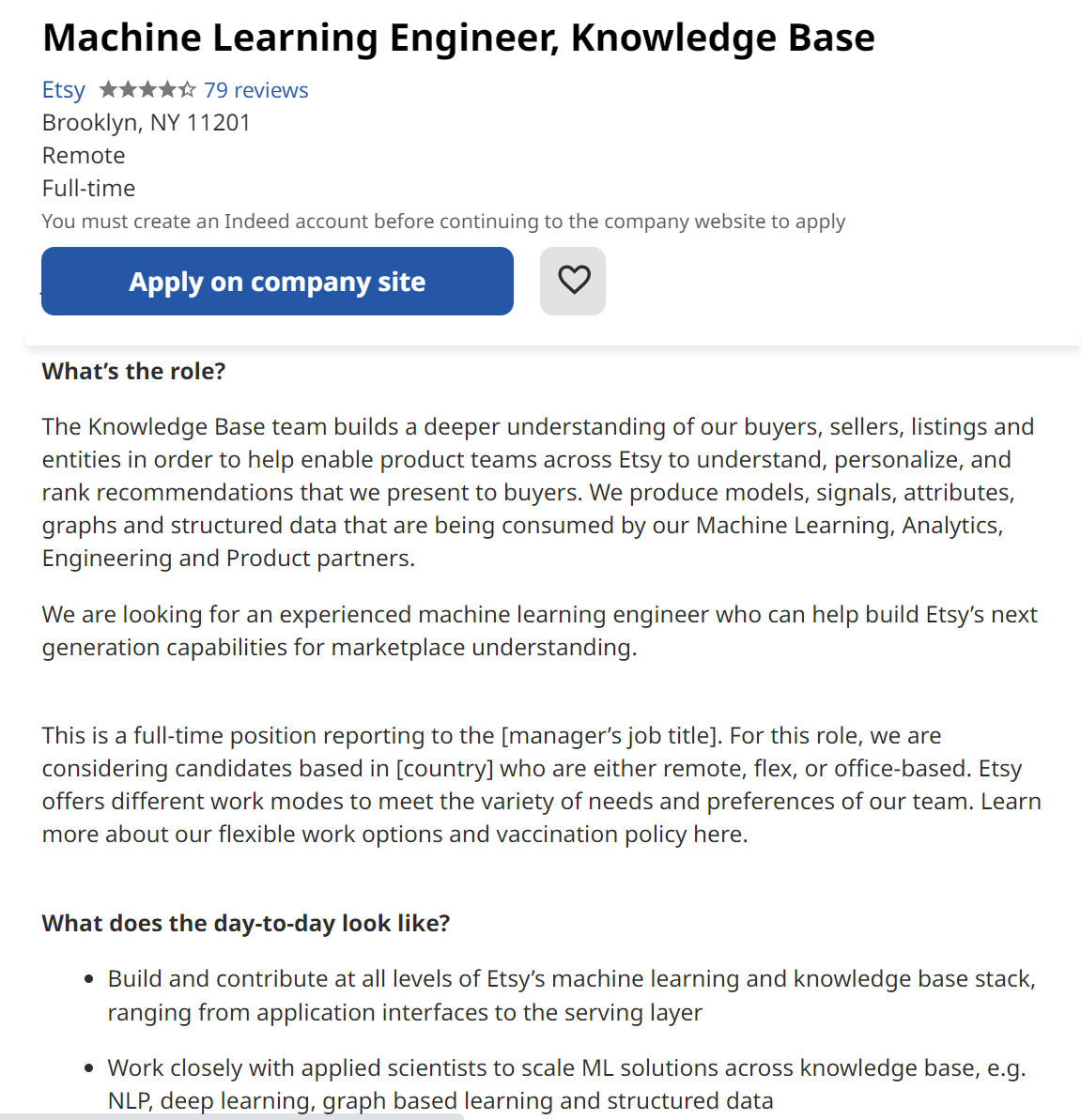
Machine Learning Engineer Example #1:

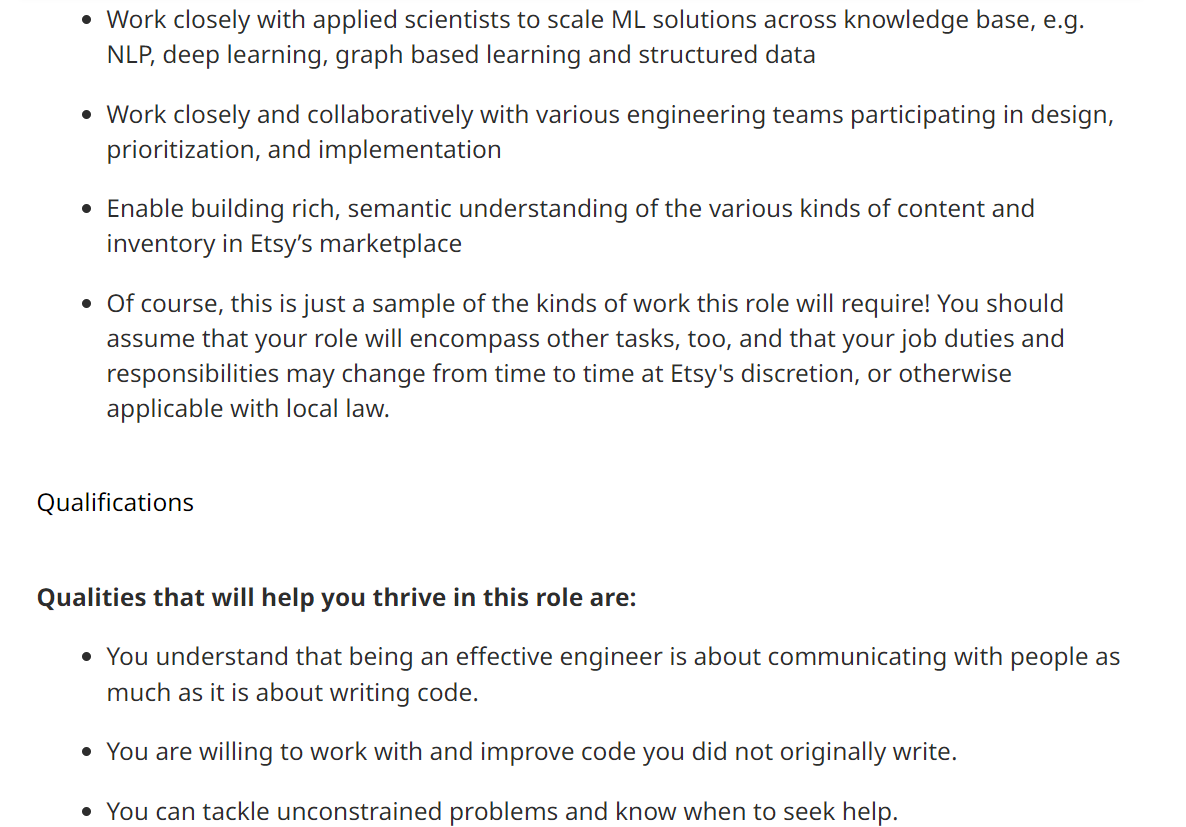


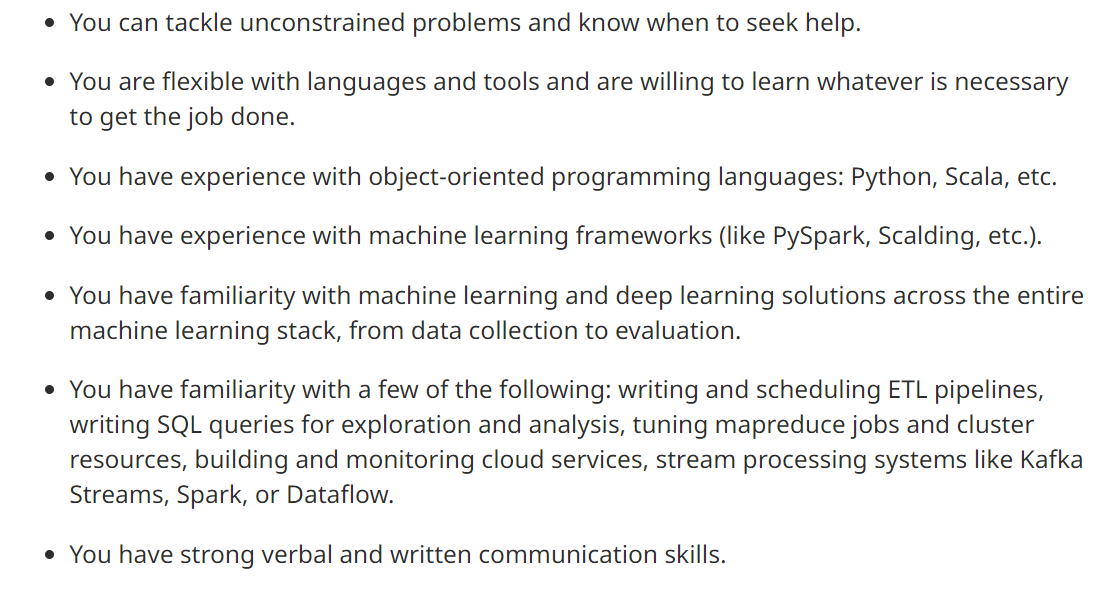


This does map to what I wrote in the synopsis, although this is a pretty specific posting and the responsibilities are comparatively simple.

Machine Learning Engineer Example #2:

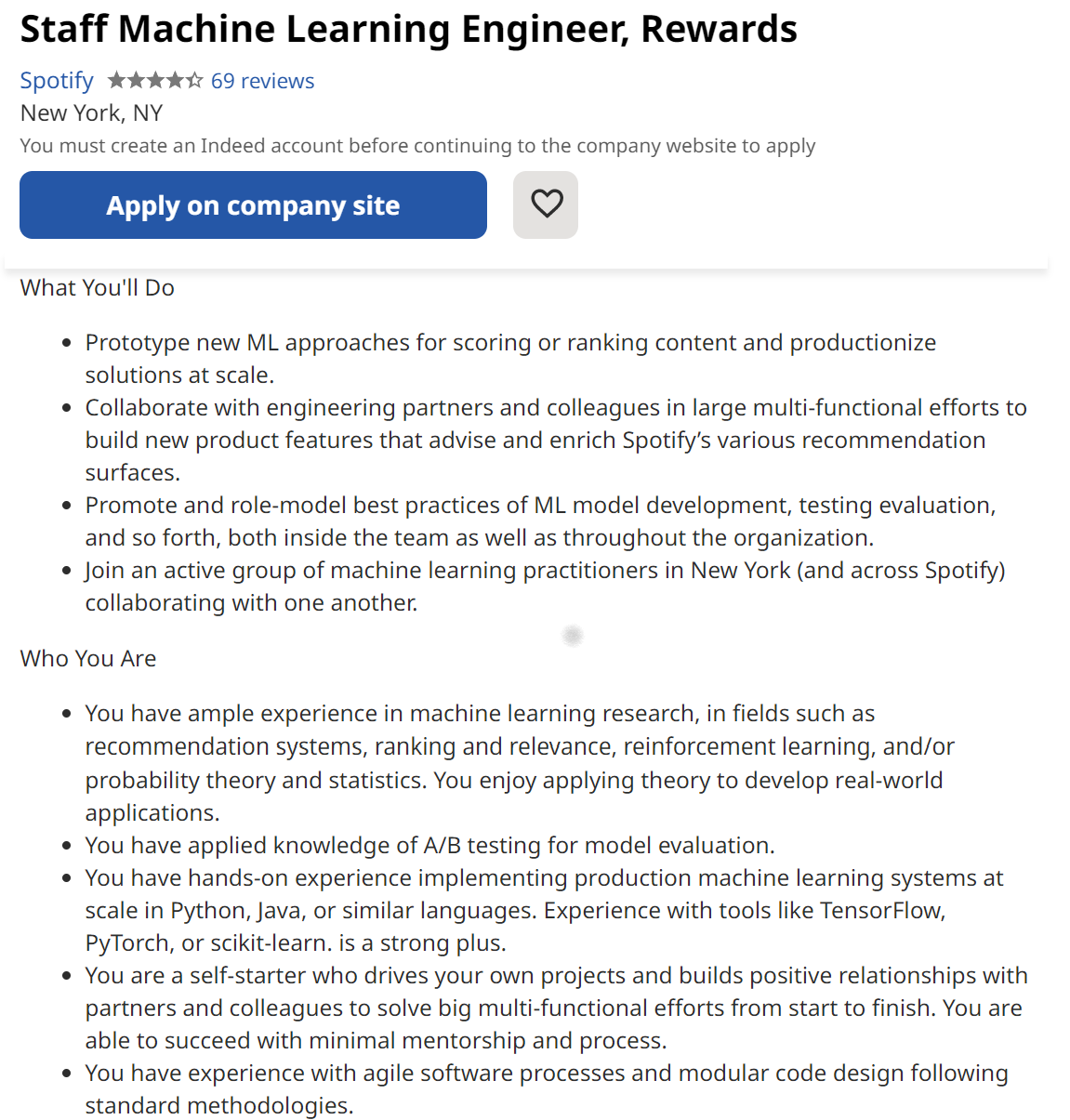


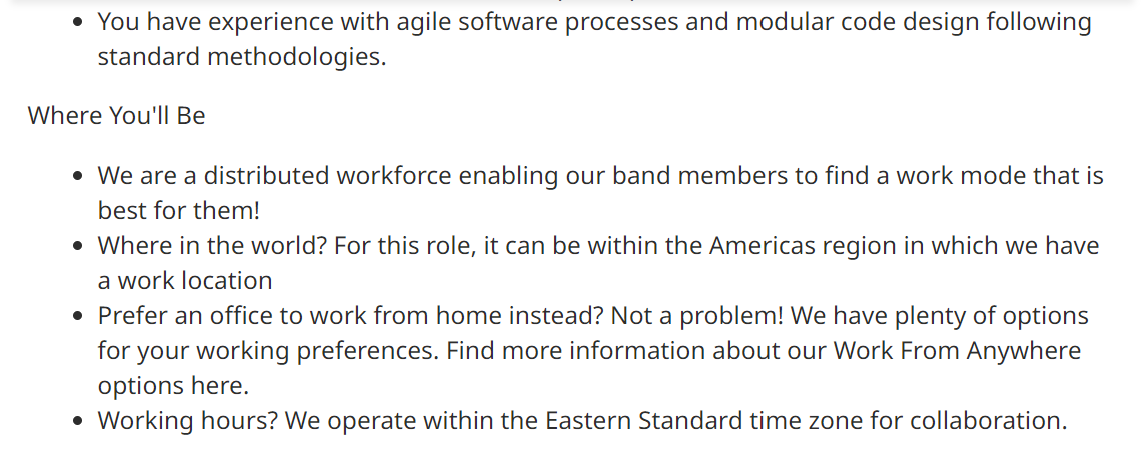




This does map to what I wrote in the synopsis, even though the last claim of ‘this is only a sample and you should expect a various other tasks too’ seems a bit sketchy.

Machine Learning Engineer Example #3:



+

This maps to what I wrote in the synopsis.

**Which of the two job role seems more interesting to you? What would be your “gap analysis” between your current skills and what employers seem to be looking for to make you a strong(er) candidate for the job role that you chose?**

As of right now, data engineer seems more interesting to me.

My gap analysis is that data engineers require skills on building and optimizing data structures, which I lack right now, and also advanced Python, I only possess rather basic Python knowledge, and I also need to expose myself to SQL server programming.

Employers seem to be looking for experience with SQL relational database, creating new visualizations, and motivation to optimize data structures; I think my experience with relational database and visualizations would help me be a stronger candidate. I also have a strong motivation for optimization even though I lack relevant experience.