Welcome to the 3Si Technical Assessment!!

Overview:

The purpose of this assessment is to gauge your technical skills in use cases that are very relevant to our day to day jobs. In other words, if you don’t enjoy this test, you won’t like the job. However, if you do enjoy these types of data problems, there are some things to keep in mind:

1. Honor system
   1. We ask that you, and only you, complete this assessment. This is not only for our benefit but yours as well. If you pass this test but cannot perform the day to day duties of the job, it will be noticeable very quickly and your employment will be short lived.
2. Write down all the steps you take and send them in.
   1. Even if you cannot solve the problem, it’s important that we can assess your logical path for how you tried.
   2. **If you cannot complete the test, you are not necessarily excluded from consideration.** With the right aptitude, these skills can be taught on the job but we need to know what we’re working with.
3. We’ll help you complete the test even if not considered
   1. At 3Si, we believe these use cases are extremely relevant for all data analytics in many industries.
   2. **We are committed to helping you complete this test to enhance your skillsets.** We want to create an honest, transparent, and enjoyable recruiting experience for everyone being considered.
   3. Keep at it! Data analytics should be considered a “practice” with continuous self- development. We don’t have a rule that states “if you don’t pass you’re excluded from future consideration for X time”. You will be more than welcome to apply for the next opportunity and will have insights into what we expect.
4. Time considerations
   1. For a skilled person, this should take 2-3 hours.
   2. However, if you start struggling, this could take a considerable amount of time.
   3. If you start exceeding each question’s “estimated time consideration” feel free to reach out to the 3Si team for a hint ([jwren@team3si.com](mailto:jwren@team3si.com))

*Questions on next page….*

**Question #1: What is the answer hidden in the dataset?**

Use Case: Explore an unknown dataset in SQL

Estimated time: 30 mins

Dataset: <https://3si-recruiting-tests.s3-us-west-2.amazonaws.com/prob1.zip>

Description: This dataset comes with no information whatsoever. There is a hidden message in there that gives you a specific answer. Find it!

Instructions/Rules: Using whatever database you wish, upload this data and write the query(s) to find the answer. This is a SQL test so, please, use SQL.

**Question #2: What is the overlap in these tables?**

Use Case: Working with different data sets and address level data

Estimated time: 60 mins

Datasets:

<https://3si-recruiting-tests.s3-us-west-2.amazonaws.com/headstart_wa.csv>

<https://3si-recruiting-tests.s3-us-west-2.amazonaws.com/ncesdata_2D86566_wa.xls>

Description: These datasets are real world datasets we use at 3Si. They are public data downloaded from the following sites: <https://eclkc.ohs.acf.hhs.gov/center-locator> and <https://nces.ed.gov/ccd/schoolsearch/>.

Instructions/Rules: Using whatever database you wish, upload these files and write the queries to determine the following:

1. How many unique addresses are there in each file?
2. What is the overlap? (How many of the same address exists in both files?)
   1. NOTE: For time purposes, feel free to choose the most simplistic way and simply describe how you could make it better given more time.
3. Create a query that includes everything from the Headstart data and an indicator whether there is a match in the NCES data

**Question #3: Do problem #1 or #2 in Python**

Estimated time: 1 hour

Description: 3Si uses primarily SQL and Python for data processing. Feel free to choose one of the above tests and rewrite it in Python.

Instructions/Rules: Avoid using SQL wrapped in Python and utilize Python data objects as best as you can. Jupyter notebooks are encouraged.

**Question #4: Scale Testing**

Use Case: Working with large datasets

Estimated time: 1-2 hours

Datasets:

<https://3si-recruiting-tests.s3-us-west-2.amazonaws.com/clickstream-enwiki-2020-01.tsv.gz>

<https://3si-recruiting-tests.s3-us-west-2.amazonaws.com/clickstream-enwiki-2020-02.tsv.gz>

<https://3si-recruiting-tests.s3-us-west-2.amazonaws.com/clickstream-enwiki-2020-03.tsv.gz>

<https://3si-recruiting-tests.s3-us-west-2.amazonaws.com/clickstream-enwiki-2020-04.tsv.gz>

<https://3si-recruiting-tests.s3-us-west-2.amazonaws.com/clickstream-enwiki-2020-05.tsv.gz>

<https://3si-recruiting-tests.s3-us-west-2.amazonaws.com/clickstream-enwiki-2020-06.tsv.gz>

Description: Data is getting larger and larger every day. Not everyone has worked with large data sets and this will challenge anyone that hasn’t worked within 10 of millions of data records at a time. This is clickstream data from Wikipedia (<https://dumps.wikimedia.org/other/clickstream/readme.html>)

Instructions: Using whatever you want (e.g. cloud databases, Python, punch cards on a Mainframe, slide rules, etc) summarize the top 50 records in the consolidated dataset (merge all together) based on the number of occurrences. For more information about the files, see the readme.html above.

Happy coding!

3Si tech team