> Written with StackEdit.

- what is a QA environment?
  Ans: the quality assurance environment is intended for you to use to complete quality assurance testing, such as functional or performance tests, and to test integrations. While this environment is fully clustered to support functional and performance testing scenarios, the environment is sized at only a portion of your production environment capacity.

R

- r - best packages for a Hadoop?

  Ans: `plyrmr` apackage is the best package for Hadoop since that provides functions for common data manipulation requirements for large datasets running on Hadoop.

- how do you use r shiny for prediction?

  Ans: Shiny is a good way to demo machine learning model or to submit machine learning challenge so that others can quickly upload test data and get amazed by your nice model. For example, we can build a regularized logistic regression that predicts whether microchips from a fabrication plant passes quality assurance (QA). During QA, each microchip goes through various tests to ensure it is functioning correctly.

- R Package conflict between gam and mgcv; how do you resolve that issues - walking through step by step.

  Ans: The problem is that both gam and mgcv packages install S3 methods for “gam” objects. Usally we can try

  1. installs gam::print.summary.gam
  2. installs mgcv::print.summary.gam
  3. save a pointer before unloading namespaces

SQL

- What is the difference between UNION and UNION ALL?

  Ans: `UNION` removes duplicate records (where all columns in the results are the same), `UNION ALL` does not.

- SQL/mysql - Any way to eliminate the behavior of distinct?

  Ans:

  1. Remove Duplicates Using [Row_Number]
  2. Remove Duplicates using self Join
  3. Remove Duplicates using group By

- What's the difference between TRUNCATE and DELETE in SQL?

  Ans: the TRUNCATE command is like a DELETE command without the WHERE clause with much less of a safety net.

Shell

- How to show only hidden directories, and then find hidden files separately?

  Ans: `ls -ad .*`

- Identify if files or directories are hidden

  Ans: Hidden files and directories have names starting with `.`

Git

- git: How do you create a new project/repository? Why use git? Alternatives? Pluses and Minuses?
  Ans: I usually go to github to create a new project/repo under my account. One of the biggest advantages of Git is its branching capabilities. Unlike centralized version control systems, Git branches are cheap and easy to merge. This facilitates the feature

branch workflow popular with many Git users. Steep learning and binary files are two main disadvantage if git.

#Mixed Topics

- what are the Oracle R Distribution Downloads available?
- what is RJDBC?

  Ans: RJDBC package is an implementation of R's DBI interface using JDBC as a back-end. This allows R to connect to any DBMS that has a JDBC driver.

- Difference between R MarkDown and R NoteBook

  Ans: An R Notebook is an R Markdown document that allows for independent and interactive execution of the code chunks. This allows you to visually assess the output as you develop your R Markdown document without having to `knit` the entire document to see the output.

- what is Apache Drill? how have you used it in your last project.

  Ans: Apache Drill_ is an open-source software framework that supports data-intensive distributed applications for interactive analysis of large-scale datasets.

- what are examples of STRUCTURED, SEMI STRUCTURED AND UNSTRUCTURED DATA? Give me an example of each from your past project. What where the biggest challenges with each one?

  Ans:

  - Structured data is very banal. It concerns all data which can be stored in database SQL in table with rows and columns. They have relationnal key and can be easily mapped into pre-designed fields. Structured data is relatively simple to enter, store, query, and analyze, but it must be strictly defined in terms of field name and type
  - Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze like JSON, XML. With some process you can store them in relation database (it could be very hard for somme kind of semi structured data), but the semi structure exist to ease space, clarity or compute... Semi-structured data is a

form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables.
    - Unstructured data represent around 80% of data. It often include text and multimedia content like social media data, Photographs and video. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents. Unstructured data may have its own internal structure, but does not conform neatly into a spreadsheet or database.

- have you heard of Zaloni? What do you think of there data platform?
  Ans: Zaloni is a leading provider of data lake management software and to the big companies all over the world. I this it is good since it help us access more data sources in more formats delivers governed data lake solutions that provide the agility and flexibility a data-powered business needs to support advanced analytics and business insights.