***Question #1***

*Please describe a project in which you delivered a…*

Answers inline.

**Problem context:**

Q1. How was the problem presented to you?

A1. A set of slides was emailed to me. It presented business view of community detection and KPI generation on call logs using SAS software. This is an example of Social Network Analysis (SNA). The slide deck contained toy examples, results of 90-day analysis and KPI definitions.

Q2. What was the goal?

A2. The goal were: (a) implicit: reproduce SAS results with Open Source Software over Hadoop, and (b) explicit: schedule SNA job over last 90 day call logs every month.

Q3. Who was your customer?

A3. Marketing and business within Freedom Mobile.

**Data gather:**

Q4. How did you identify the data required?

A4. I was asked to perform SNA over call logs.

Q5. How did you get the data?

A5. I used standard Hadoop extracting tool from database systems called Sqoop.

Q6. What was the scale of the data?

A6. Last 90 days. Over 100 million records.

Q7. What challenges did you run into and how did you overcome them?

A7. Data needing extraction was a result of a nested Teradata SQL query. The query joined multiple tables, performed data transformations and used Teradata specific syntax. The challenge was to map the Teradata query to the less rich SQL available in Sqoop, and ensure the final results are identical. I unfolded the Teradata query into parts and mapped them to the SQL in Sqoop, and performed validation on each step along the way.

**Analysis and modelling:**

Q8. Discuss the data analyses you did, whether they are purely exploratory or model-based.

A8. I took the toy examples from the slide deck and research papers, and compared the results of community and KPI detection algorithms that I used. I got similar results. I was trying to build models of underlying communities.

Q9. What aspects of the analysis were particularly interesting or challenging?

A9. The algorithms seem to work well on toy examples. However, finding ground truth on 100 million records was a challenging problem. The algorithms (PSCAN, LDA, k-clique, maximal clique) sometimes returned vastly different communities and were non-terminating on detecting communities or KPIs. The algorithms used by SAS software were unknown, so was unable to compare.

Q10. What types of model did you consider, which did you select, and why?

I use GraphX and GraphFrames in Spark. The algorithms build the model, and they were:

* LDA: did not terminate on 100 million records
* k-clique: community sizes maxed out 2 member
* P-SCAN: fast and terminating but provided non-overlapping communities

**Impact:**

Q11. What measurable impact did your work have? What was the end result?

A11. Tangible results:

* mapping business SNA terms to standard research SNA terms to ensure customer and IT are aligned.
* Scripts and Zeppelin notebooks to identify (non)/overlapping communities and KPIs.

**Reflection:**

Q12. In hindsight, what could you have done better? Is there anything that you will do differently in future?

A12. If I am allowed some autonomy, I’ll work closely with the customer. There were certain part of works that was faster to accomplish and impactful such as discovering central member in communities. There are multiple applications of SNA (e.g. targeted advertisement, prioritizing support) and I’ll let the use case drive the effort.

= ***Question #2*** =

*We have a website, where we monitor when did users log in and logged out. The data is provided in the given format: list of (session id, start date, end date, username) tuples. Please calculate how many users were logged into the system in the different time intervals. The output should be a list of (start date, end date, number of sessions between start & end date) tuples, intervals with 0 session should not appear in the output.*

*Please provide us with a script/code written in the programming language (SQL/Python/...) of your choice that calculates the output according to the specifications above.*

First: assuming the data is in the table ‘mydata’, the required query is:

[SELECT](http://localhost:8888/phpMyAdmin/url.php?url=http://dev.mysql.com/doc/refman/5.5/en/select.html) start\_date, end\_date, [count](http://localhost:8888/phpMyAdmin/url.php?url=http://dev.mysql.com/doc/refman/5.5/en/group-by-functions.html#function_count)(0) AS session\_count FROM (

[SELECT](http://localhost:8888/phpMyAdmin/url.php?url=http://dev.mysql.com/doc/refman/5.5/en/select.html) \* FROM (

[SELECT](http://localhost:8888/phpMyAdmin/url.php?url=http://dev.mysql.com/doc/refman/5.5/en/select.html) start\_date, end\_date, TIMESTAMPDIFF(minute, start\_date, end\_date) AS duration FROM mydata ) AS r )

As q

WHERE duration > 0 GROUP BY start\_date, duration

Assumptions: The required “time interval” or “session” is a bucket determined by start\_date and duration.

**Discussion:**

Sessionization is a large topic and there open source libraries available for the work e.g. Linkedin’s DataFu library.

***= Question # 3 =***

*Write an SQL query which returns a single aggregate metric indicating how well actuals met forecast for a given week. What metric did you choose and why? What assumptions did you make?*

Required output: *single metric* indicating how well actuals met forecast for a *given week.*

Assume, the Forecast and Actual Quantities of a ProductID are placed side-by-side by massaging the input data into a new table called UnitsSoldReport. A sample is shown below.

|  |  |  |  |
| --- | --- | --- | --- |
| **ProductID** | **ForecastQuantity** | **ActualQuantity** | **Date** |
| Widget | 100 | 120 | 6/1/2015 |
| Gizmo | 90 | 80 | 6/1/2015 |
| Doohickey | 200 | 200 | 6/1/2015 |

With this format, it is easy to calculate Mean Absolute Error or MAE

SELECT r.sumerror/r.numOfRecords AS MAE

FROM (

SELECT COUNT(\*) AS numOfRecords,

SUM(ABSOLUTE(ForecastQuantity – ActualQuantity)) AS sumerror

FROM UnitsSoldReport

WHERE DATE BETWEEN ‘$startDate’ AND ‘$endDate’

) AS r;

$startDate and $endDate represent start and end date of a given week, for example, 6/1/2015 and 6/7/2015.

Assumptions,

1. All products are equally important as MAE does not capture this notion.
2. Magnitude of quantity for across products are similar. In the above example, we around hundreds for a given week.

**Discussion**: MAE being an average metric is prone to outliers. We also need to address missing values by removing records or implanting. Both items are outstanding. I choose MAE as a starting point. It is more intuitive to a layman with a simple definition of a “quantity used to measure how close forecasts or predictions are to the eventual outcomes.” [1] compared to root-mean-square-value (rmse).

[1] <https://en.wikipedia.org/wiki/Mean_absolute_error>

***= Question # 4 =***

*Given Flight A arrives at a random time between 2pm and 4pm, and Flight B arrives at a random time between 3pm and 5pm. What are the Odds that Flight A arrives before Flight B? Why?*

The probability that Flight A arrives before Flight B is 0.5625 .

Method:

1. Divide the time for Flight A in two windows, namely 2-to-3 and 3-to-4. The probability of A randomly arriving in the 2-to-3 window is ½ -- similar to a coin toss.
2. Similarly, divide the time for Flight B in two windows, namely 3-to-4 and 4-to-5. The probability of Flight B randomly arriving in the 4-to-5 window is ½.
3. Now, subdivide the time in 3-to-4 in two parts:
   1. 3pm + x minutes, named as early part, where x < 60 min
   2. the probability that Flight A arrives in the early part at a random time = the probability that Flight A arrives in the 3-to-4 window (i.e. ½) AND the probability that Flight A happens to be in 10+x window (i.e. ½) = ½ x ½ = ¼
   3. the probability that Flight B NOT arrives in the early window = the probability that Flight B arrives in the 3-to-4 window (i.e. ½) AND the probability that Flight B is NOT in the early window (i.e. ½) = ½ x ½ = ¼
4. The probability of Flight A arriving in 2-to-3 OR 3pm+x window = I + b = ½ + ¼ = 0.75
5. The probability of Flight B arriving in 4-to-5 OR NOT(3pm+x) window = II + c = ½ + ¼ = 0.75
6. The probability of Flight A arriving before Flight B = IV x V = 0.75 x 0.75 = 0.5625