Interview Memo

I. Summary

comScore, Inc., the client I work with, is a global media measurement and analytics company. They mainly provide the services of behavioral analyses across screens, such as PC, mobile, OTT (over-the-top), etc. The data at comScore are mostly behavioral data, which the company use the data for behavioral analyses, such as market share of mobile phone, top video view/impression, accuracy of advertisement delivery, etc.

II. Researcher & Project Info.

Yuvraj Singh is the manager of Ad-operation team at comScore, Inc. He has been working on this project for over three years, and currently he is mainly responsible for the raw data and aggregate data.

The project with K company (due to confidentiality, comScore would rather not to disclose the company's name as well as the exact project name) is the one that I am focusing on during the semester, which is also the one that they have been working on for 3 years. This project is about helping both advertiser (the K company) and publisher such as YouTube, Google, etc. to better delivery advertisements to target audience. The project is periodically report-based, which mainly covers advertisement campaign measurements from specific publishers that were gathered from a high-automated process.

III. Interview Memo

Defining you data

1. Data Description:
   The dataset is about advertisement campaign measurements from specific publishers, with 12 tables and 426 columns in total. There are several campaigns in the dataset, and each campaign stands for an advertising project. Each campaign contains "studies" in monthly basis. There are

basically three types of "studies" – PC, mobile, and TDP (Total Digital Population). The TDP stands for the de-duplication of unique individuals of PC and mobile "studies". For example, there is a campaign/ads project last for 12 months. Under this campaign, there will be 12 studies, and each study contains PC, mobile, and TDP types. All these three types share the same measuring methods such as In-target UVs, In-geo UVs, and In-human UVs. Based on the studies, the dataset can be used for measuring ads display information and qualification, which generate performance report to each specific campaign.

Notes: In-target UVs stands for the unique view data that are collected based on demographic information (i.e., 18-24 years old, female); In-geo UVs data are collected based on geographic information; In-human UVs stands for the data that are "real human views".

2. Where does your data come from?
   The data come from "comScore tag". The principle of this is that the client would put sets of code on their websites and ads, and whenever the ads are displayed, information will be send to the server.

3. How much data do you generate?
   It varies depending on each day. On average, the amount of data generated per day is around 500 MB.

4. What file formats do you use?
   Generally, .txt file format is the most commonly used.

Looking after your data

5. What different versions of each data file do you create?
   Raw data, aggregate data, and file loading data.

6. What metadata are you adding to each data file?
   There are two sets of metadata that comScore adds to each data file. One is used for mapping the campaign to each category, like advertisement classification. For example, there are ads categories like food, sports, beauty, etc. Another one is for mapping placement IDs to the actual placements. For example, when there are several advertisements on a website, different places of

a website have different placement IDs.

7. Where do you store your data?
   The raw data are stored in Greenplum, and the final loading data are stored in SQL Server.

8. How do you structure and name your folders and files?
   Here is a most general example that the client provided:
   …/client/client_id/02.data/date/file_id_file_type.txt
   This stands for client name, client ID, what kind of data or project, date, file ID with file type, respectively.

9. How is your data backed up?
   comScore has their own servers that back up the data on daily basis.

10. How will you test whether you can restore from your backups?
    Since the Ads-operations team has never faced this demand, so they haven't been tried to test whether they can restore from their backups.

11. Who is responsible for the immediate day-to-day management, storage and backup of your data?
    There is another team called data warehouse team that is responsible for immediate day-to-day management, storage, and backup of their data when necessary.

Sharing your data

12. Who owns the data you generate?
    Other than the Ad-Operation team, the CMS (comScore Marketing Solution) team also own the data generated.

    Notes: The Ad-Operation team belongs to the Digital Product team, and the Digital Product team has the same level with the CMS team.

13. Are you working with collaborators (sharing your data with them)?
    The CMS team and the CI (Client Insight) team at comScore. The Client Insights work as client services, that are mainly responsible for communicating among K company, Ad-Operation team,

and CMS team.

14. What should and shouldn't be shared and why?
    Only the final loading data can be shared. Internally, the raw data and the aggregate data cannot be shared even to other teams, such as the CI (Client Insight) team. This is because the raw data and the aggregate data usually contain complicated functions built in, the Client Insights would mess up since they are not technically professional.

Archiving your data

15. What should be archived beyond the end of the project?
    All the flat files older than 90 days will be archived beyond the end of the project.

16. For how long should it be stored?
    All the historical data are stored as long as the contract with K company lasts for.

17. When will files be moved into the archive?
    The files will be moved into the archive in every 90 days range.

18. Where will the archive be stored?
    comScore has its own data storage server, and the server stores the archives as flat files and archive files.

19. Who is responsible for moving data to the archive and maintaining it?
    There is no manual process on this. All the data archiving and maintaining are done in automated process.

20. Who should have access and under what conditions?
    Generally, the CMS team has the full access to the data, and the Ad-operation team has access to raw and aggregate data. I haven't been told about "under what conditions" since it might be so complicated and might be under any conditions whenever necessary. The CI team only has access to the final loading data, when it is necessary for them to communicate among the Ad-operation team, the CMS team, and K company, such as when K company has updated demands.

Data management plan questions

21. Who is responsible for making sure this plan is followed?
    Yuvraj Singh, the manager of Ad-operation team.

22. How often will this plan be reviewed and updated?
    It may depend on the progress of the project. Also, we are pretty flexible in ways of contact, either online or offline. In average, the frequency of reviews and updates is in every two weeks.

23. What actions have you identified from the rest of the plan?
    Since we will be doing the timeline and the budget part for the project, I would like to combine my business background (i.e., minimizing the total cost of the process, optimizing division of labor) to keep tracking of the information.

24. What further information do you need to carry out these actions?
    I would need to know whether comScore already has a thorough record/estimation of data process cost, such as the cost from the raw data to the aggregate data, and from the aggregate data to the final loading data. This would also contain the information of how long it costs for each process, and whether there is any measurable or immeasurable extra cost during the process.

Data services questions

25. What data services would assist you in your research?
    I would need to figure out how Greenplum works in order to help me understand the entire process better. Other than data services, I would need to explore more about the knowledge of online advertising.