**Big Data - Fall 2019**

**Guidelines for the Project Report and Presentation**

**Project Report**

The project report should contain the following information:

• Title

• Authors

• Abstract: The abstract should summarize the report and give the reader an idea about the work that was undertaken and what you findings/results are.

• Introduction: This is where you need to outline the motivation for your work, state the problem you are addressing, explain why it is important, describe the underlying concepts, and justify if and how big data infrastructure was needed (or useful) for your project.

• Task 1:

1. *Methods:* 
   1. *Describe the methods you designed/implemented and the tools you used to complete this task. Provide a link to the github repo where your code for this task resides.*
   2. *Discuss the challenges you have faced while designing and implementing your solution (e.g., skewed data), time to run the profiling tasks, any optimizations you have implemented to speed up your code, and data quality issues you have identified (e.g., too many missing values, heterogeneous columns -- columns with values of multiple types).*
2. *Evaluation:*
   1. *Summarize your results using plots/visualizations, e.g., a histogram that shows, for each data type, how many columns contain that type. Also report the most common types that co-occur in columns -- this can be done by applying frequent itemsets: you can find the 2-, 3-, 4-frequent itemsets for types in columns. Other information you can summarize: missing values, heterogeneous columns.*

• Task 2:

1. *Methods: Describe the strategies you used to detect the different types, and the benefits and limitations of these strategies. Provide a link to the github repo where your code for this task resides.*
2. *Evaluation: Report the precision and recall for the different strategies. To compute precision and recall, you need the true type of each column: the team members will collaboratively and manually label the columns with their* ***true type****. Note that a given column may have values of different types, therefore, its true type may consist of multiple labels. Include only labels for types that occur frequently, and omit any outliers. Include visualizations that summarize your findings, e.g., a histogram showing for each type, the number of columns in which the type appears; a visualization that shows the prevalence of heterogeneous columns, i.e., columns that have values belonging to multiple types.*

• Task 3:

1. Problem definition: describe the problem you selected.
2. Experimental techniques and methods: You should provide details about the methodology and the tools you used. You should also describe your experimental setup, including the data you used, and the cluster configuration (e.g., node configuration, number of nodes, mappers and reducers)
3. Findings: discuss your findings as well as any issues/challenges you encountered and how you addressed them.

• Individual Contributions: describe the contributions each member of group made to the project

• Summary/conclusions

• References: List any references you have used

The report must follow the ACM Proceedings Format, using the **sample-sigconf.tex** template provided at <https://www.acm.org/publications/proceedings-template> for LaTeX (version 2e), or <https://www.acm.org/binaries/content/assets/publications/word_style/interim-template-style/interim-layout-.docx> for Word. If you choose to use latex, you can use OverLeaf (<https://www.overleaf.com>) to edit your report.

**Project Presentation**

You will have 7 minutes for your presentation and 3 minutes for questions. Present only the most important findings for each task.

You can bring a poster or a set of slides printed on letter size pages. For the latter, you should have between 8 and 10 view graphs.

Make sure your poster/slides list your group name and group members.