Statement of Purpose

of Md Tarikul Islam Papon (PhD applicant for Fall–2019)

My long-term career goal is to be actively involved in research as a faculty member or as a researcher so that I can contribute to the human civilization by making novel and significant innovations. My research interest lies in the field of Data Science and Distributed Systems. More specifically, I am interested in database, big data management, distributed computing, and large-scale analytics. The Department of Computer Science at the University of Maryland, College Park is making major contributions in the aforementioned fields. I believe that this is one of the best places for me to obtain a Ph.D. degree which will be the first step towards my long-term goals. With my academic background (UG rank: 3, GPA: 3.92/4.00, PG GPA: 4.00/4.00), research background (Distributed Computing, Data Analytics, and Machine Learning) and teaching experience (Lecturer, Department of Computer Science & Engg. (CSE) at Bangladesh University of Engg. & Tech. (BUET)), I am confident of meeting the high standards of University of Maryland.

I grew a knack for research while I was working on my undergraduate thesis under the supervision of Dr. A.K.M. Ashikur Rahman, Professor, CSE, BUET. I worked on developing a low-cost device that can calculate the heart rate and approximate blood pressure (BP) of a person in a noninvasive way. In addition, a centralized heart rate monitoring system was developed where the nodes could communicate with the master over RF communication. Our works have been published in the *International Conference on Networking Systems and Security (NSysS)*.

My enthusiasm for research increased during my M.Sc. thesis under Professor Rahman where I worked on a joint project with Ubicomp Lab at Marquette University. The goal was to detect Diabetic Retinopathy (DR) by classifying the images of interior surface of eye (known as *fundus image*) using machine learning techniques. While studying the literature, I noticed the fact that most of the retrospective methods to extract features have been implemented on generic and very small datasets with little or no variation of illumination among the images. In contrast, the real dataset obtained from Ubicomp Lab was extremely challenging because of variable illumination, different orientation, shape, and coloring of the images. I applied techniques like object localization, transfer learning and sequence learning to solve these problems and extracted some features to pass into my classification model. The results were promising but to achieve an even better performance, I needed to extract some highly intricate features which is virtually impossible for this dataset. So I started to devise an alternate approach. Consequently, I applied my skills of image processing to preprocess the raw fundus images and employed deep learning techniques on them using a *convolutional neural network*, which provided excellent results. Now, I am working on submitting my work for publication along with preparing and submitting my M.Sc. dissertation.

My interest in Data Science and Distributed Systems started when I was attending my M.Sc. courses titled "Data Management in Cloud" and "Distributed Systems". As part of the former course, I needed to implement a project on cloud computing. We are living in an era where every sector of our lives are impacted by the amount and variety of data generated. It has been a while since *big data* became a buzzword. I wanted to explore this interesting realm of data by concentrating on the usefulness of data in an efficient manner. So I decided to explore a highly investigated research topic - finding the most influential spreader in a social network. Since online social networks have a huge volume of data which is very difficult to analyze in a single machine, I decided to implement a distributed variant of the *k-core decomposition* algorithm on a Twitter dataset.

While implementing this project, I faced quite a few interesting challenges. I wanted to investigate how such a high volume of data can be retrieved and managed. To achieve this, I gathered 5 million tweets over one week using the public sample API. I used MongoDB to process and store this unstructured data. Then, I sampled this dataset based on some hashtags to extract the tweets related to a specific news (Holy Artisan Attack) which occurred earlier that week. The next conundrum was to build the *follower-followee* graph. I needed the information of who is following whom in the Twitter dataset. However, because

of the API restriction, it took more than one month to gather that information. Meanwhile, to solve the last challenge of implementing a distributed version of the *k-core decomposition* algorithm, I grasped the concept of distributed computing and learned Apache Giraph, which is an iterative graph processing system that uses Hadoop's MapReduce implementation. I ran Giraph on my graph over Hadoop File System (HDFS) on an Amazon AWS cluster. Lastly, I modified the algorithm to incorporate node information and after some experimentation, my algorithm outperformed the original algorithm in both accuracy and runtime. Currently, I am working on to improve the performance more by deploying Natural Language Processing (NLP) on the nodes' historical data. I am extending my work on Stanford Large Network Dataset Collection. We are preparing the manuscript and expecting to publish this work in a journal soon. All these research experiences have augmented up my skills and confidence in pursuing a Ph.D. in my field of interest.

I was fascinated by the enormous potential for research in Distributed Systems. So I studied the important topics related to this field during my Master's course "Distributed Systems". I learned the key concepts of distributed computing like clock synchronization, mutual exclusion, ordered multicast, consensus, etc. in this course. As part of the course, I gave a presentation on *Dynamo*, which is Amazon's highly available, and scalable NoSQL key-value database service. Through this, I introduced myself to many fundamental concepts like consistent hashing, data replication, eventual consistency, fault tolerance, etc. I also explored Google's NoSQL database services- *Spanner* and *Bigtable*. I was particularly impressed by the techniques (Commit Wait, 2-Phase Commit, TrueTime, etc.) adapted to maintain external consistency of transactions in *Spanner*. At the same time, I was conducting my aforementioned project in Cloud Computing and realized the importance of distributed computing for big data analytics.

To analyze and manage big data efficiently, there is no viable way other than distributed computing. In my doctoral research, I would like to delve deeper into these fields. The Department of Computer Science at the University of Maryland, College Park is constantly carrying out groundbreaking research in these fields. I would be thrilled to have the opportunity to get involved in research as a part of this department. I am very keen to involve in a project related to query optimization. I envision to take part in developing a scalable big data platform that can perform large-scale analytics in a secure and efficient way using parallel database. I am also open to working on other projects related to Data Science and Distributed Systems if I get the opportunity. I believe that my academic background, experience in data analytics and distributed computing, and research skills have given me a solid foundation to explore my fields of interests to advance towards my career goal.

I am particularly intrigued by the research activities conducted by the database research group. In particular, I am interested in the project "DataHub" led by Prof. Amol Deshpande. I have always dreamt of developing a smart version controlling system. In addition to dataset's version controlling, this project will enable collaborative data analytics, which seems like a fascinating idea. I would love to work with Prof. Amol Deshpande as his research work meets my interest most. I believe I can be a good fit for his research group because of my past research experience in data analytics, distributed computing and applied machine learning. I am also interested to work with Prof. Daniel Abadi. His recent review on Deterministic Database Systems is fraught with many valuable insights. Specially the advantages of these type systems like higher performance database replication, improved scalability, and increased transaction concurrency are quite promising. I am also interested to work with Prof. Leilani Battle on the interdisciplinary field of database and human-computer interaction. Her recent work on using "Beagle" to collect, label and analyze visualizations created on web seems very interesting. I am also open and would be happy to work with others as well, who have an interest in the fields of Data Science and Distributed Systems if such opportunity occurs.

My desire to remain in academia has turned into my resolution due to my predilection for my current job as a lecturer. I believe that the Department of Computer Science at the University of Maryland, College Park is a place where I can enhance my skills as an academician as well as a researcher. I am looking forward to the diverse research experience that my stay at University of Maryland might bring.