**Statement of Purpose**

Applicant Name: Chen Luo

Program: Ph.D. in Computer Science

People always talk about Big Data, but what does Big Data really mean? For me, a second-year graduate student, the area of Big Data involves numerous research challenges, not just in traditional databases, but also in many emerging topics, such as quality, privacy, and usability. Throughout my undergraduate and graduate studies, I have experienced many Big Data challenges in various aspects, which cultivated my strong interest in the database area and motivated me to undertake further study through a Ph.D. program overseas.

Currently, I am working on the Tsinghua-Intel joint research project entitled *Privacy Checking for Big Data Analytics Systems*, which also constituted my master’s thesis and opened up the opportunity to tackle the privacy challenge of Big Data. During the project, I led a group in building a privacy-integrated data analytics system, which automatically enforces privacy compliances against SQL queries. First, I designed a high-level privacy language for specifying privacy policies for data analytics. The novelty of this language is the explicit support for associated data access and flexible desensitization, which pertain to data analytics but are missing in previous works. To enable formal reasoning about our language, I defined its formal semantics, upon which I further developed several redundancy and consistency analysis algorithms with Satisfiability Modulo Theories (SMT) techniques. To check the compliances of SQL queries, I then implemented the privacy checker into the query processing component of Spark-SQL, upon which our system was built, using program analysis techniques. The system was further integrated with differential privacy to offer stronger privacy guarantees for aggregation. Now the project is at the final stage, and I have summarized the research results into three papers in submission. Moreover, the various techniques involved in building this system also let me realize the importance of inter-field research in handlingreal-world problems.

Dating back to the senior year of my undergraduate study, I experienced the first Big Data challenge during an internship at eBay. Working with my colleague, we redeveloped the internal web traffic analytics system using MapReduce since the original system could not handle the rapidly growing logs in time. I designed a distributed MapReduce algorithm to summarize the logs and then visualize them into interactive reports. To make the system more efficient, I then performed thorough experiments to evaluate the system’s performance, and further developed several optimizations to fix performance issues. The improved system was ultimately able to process daily logs within minutes; moreover, it can be easily scaled out with commodity servers. The successful development of the system significantly improved my problem-solving and engineering skills. It likewise deepened my appreciation of Big Data in terms of turning data into values and motivated my interest to develop scalable solutions to make this happen.

After completing my bachelor’s degree, I continued my graduate study at Tsinghua University to further advance my knowledge and skills. From a course entitled *Advanced Topics in Data Quality*,I witnessed another Big Data challenge in data quality. The course organized a competition called DQCup, which required students to build a data-cleaning tool for a given task. In the process, I applied various techniques, such as data constraints, functional dependencies and sequential dependencies, to design a suitable algorithm. However, the toughest part was to tune the algorithm according to the limited feedback, i.e., some metric scores, from each submission. Thus, I spent a lot of time designing strategies and optimizations, and then validating them based on the received feedback. Although the process was painstaking, the results were exiting - our tool finally detected and repaired 98% of known errors, and was ranked 1st by all metrics among 15 teams. From the experience, I not only learned many useful techniques in data quality, but also greatly improved my capability of algorithm design and optimization. In the meanwhile, I found myself becoming more interested in solving data-related problems, which further determined my future research direction.

In addition to the database area, I also conduct some research in the formal methods area during my graduate study to broaden my eye sight. Entering a new area was a great challenge, but also created an opportunity for me to utilize my knowledge in both areas, which led to my independent research project entitled *Large-Scale Software Model Inference*. Model inference is a technique in formal methods to infer software models from execution logs; however, it may suffer from the scalability issue as the logs could be large in practice. To handle this, I came up with a novel solution resorting to parallel data processing techniques, e.g., MapReduce. I designed two MapReduce algorithms which first slices the log into multiple pieces and then synthesizes the software model from them. To ensure correctness, I formally proved that the model inferred by this approach is identical to the one produced by the traditional approach. Recently, I presented this research at a formal methods symposium, and received a lot of positive feedback. More importantly, the experience in formal methods provided me a unique perspective to reason about computer systems, and greatly strengthened my theoretical foundation in computer science.

My past experiences on Big Data have motivated my strong interest in a wide range of problems in the database area, including data management, processing, integration, quality, privacy, as well as other data-related problems. During my Ph.D. study, I wish to further conduct research in the database area, especially developing scalable and practical algorithms, tools, and systems to facilitate the use of Big Data. After graduation, I plan to join industrial research labs to be an industrial researcher because industries have numerous first-hand experiences on Big Data, which would allow me to place my research in a real-world context to produce more direct impact to the business, even the society.

To strengthen my foundation and skills, I look to the Ph.D. program in computer science at the University of Washington as the ideal platform to achieve my goals. UW’s leading database group and supportive research environment would be extremely helpful for my future research in the database area. In particular, I would like to work with Professor Balazinska. Many of her research projects are very intriguing to me, such as her recent work on data use enforcement and cloud data management. Moreover, UW’s tight collaboration with the industry would also be advantageous for my future career as an industrial researcher. For these reasons, it would be a great opportunity to study in UW under the supervision of its remarkable faculty.