PREFACE

In the contemporary digital era, cloud computing and big data eco-system is playing pivotal role in dealing with data storage and computing requirements of enterprises in the real world. Distributed programming frameworks like Hadoop fit into such environment to exploit the power of Graphics Processing Units (GPUs) associated with data centers and the parallel processing phenomena. Hadoop is widely used for storing and processing big data. However, secure computations in distributed programming frameworks for processing big data is the need of the hour and the existing techniques to secure computations are inadequate. According to CSA, it is important to handle four security requirements in such distributed environments. They include data privacy related security, integrity and reactive security, infrastructure security and data management related security.

This thesis deals with four major objectives which are addressed in chapter 3, chapter 4, chapter 5 and chapter 6. The first one is to propose and implement a framework for securing MapReduce infrastructure by detecting rogue worker nodes. This is achieved by analyzing log files and system call traces. It could provide solution to the problem of secure computations in distributed programming frameworks without actually modifying the source code of the framework.

The second objective is to protect big data from privacy attacks. Differential privacy is employed to protect privacy of big data. An algorithm is proposed to take care of adding noise to reduce the chances of adversaries getting succeeded in disclosing sensitive information. Empirical study is made with real cloud environment using Amazon EC2 as part of the computing system. The experiments with EDGAR dataset proved that the adversary is not able to find the presence or absence of an entity through privacy attacks.

The third objective is to propose and implement a framework for securing cloud storage. It does mean that it provides required security mechanisms to ensure that big data is outsourced and retrieved with complete security. The framework takes
care of heterogeneity of big data by considering different approaches in employing cryptographic primitives intelligently. Three algorithms are defined to achieve this. However, the framework does not support searchable encryption. Therefore an improved version of it is presented in Chapter 6.

The fourth objective is to propose a framework for cloud storage security with the provision of supporting arbitrary computations on the outsourced encrypted cloud data. Fully Homomorphic Encryption (FHE) based approach is followed in the framework to ensure that the encrypted big data stored in the cloud can be queried directly. This is the salient feature of the framework that makes the cloud server free from the burden of decrypting the whole data before searching. Instead it directly performs query processing on the encrypted data and returns query results in encrypted format. Then the user will decrypt the query results. With the four objectives fulfilled, this thesis is in tandem with the expected security requirements mentioned by CSA. The results produced with the help of a prototype demonstrate proof of the concept.