Airavat: Security and Privacy for MapReduce

Summary - This paper presents Airavat, a framework for MapReduce which provides security and rigorous privacy guarantees to data owners and users. It aims at providing efficient distributed computation, support for familiar programming model and to provide precise, rigorous privacy and security guarantees to data owners. It is a wrapper around modified MapReduce, DFS, JVM and SELinux. Cloud provides an ideal environment where distributed computations can provide useful results to research facilities. However, most of the times the data used in these types of computations is private, for example health records of an individual and for this reason sufficient steps should be taken to preserve the privacy of users data. To solve these problems Airavat uses MAC (Mandatory access control) and data privacy to achieve end to end confidentiality and integrity. MAC prevent data leaks through storage channels like network or files whereas differential privacy prevents leaks through the output of the computation. To achieve differential privacy Airavat adds laplacian noise (Lap(Del f / e)) to the output. Intuition is to mask the effect of single input. On the other hand MAC enforces access rules with the help of domains specified by the system administrator at all the time, without user override. Domains discussed in this paper includes trusted and untrusted domains. MapReduce and DFS belongs to trusted domain where as computational provider's mappers belongs to untrusted domain. In this model, the data provider uploads data with certain privacy parameters. The computation provider writes possibly untrusted mapper code. Airavat runs the computation and protects data privacy.

Opinion - This paper addresses the weaknesses of current distributed computations which has been identified as one of the most important research challenges to further increase the viability of the cloud computing. In short, data privacy is getting clouded. Existing approaches involving removal of personally identifiable information like name, address and SSN have been shown inadequate. With all this in mind, in my opinion Airavat is a novel privacy preserving framework for cloud computations. It confines untrusted code with help of MAC and differential privacy. It has an execution overload of 32 % compared to Hadoop jobs without Airavat. Though the privacy parameter controls the tradeoffs between the accuracy of the output and the probability that it leaks the information, it's still better to preserve the privacy. One improvement which could be made is to have a pre built library of permissions which the data providers can use. This relieves the data provider from defining different labels for the data.

Pros and Cons - Lets starts with the pros. This is the first system for privacy preserving in MapReduce computations. It confines untrusted code. Accuracy and privacy can be controlled by the privacy parameters. It uses a programming models similar to MapReduce. This helps developers to better understand the framework. Now for the downsides, Airavat fully trust the cloud provider in the first place. Output keys has to be known in advance which might not be the case. Data provider is burdened with providing security parameter. Most of the time MapReduce jobs uses Pig also. There is no discussion or support for Pig. Also the paper has limited discussion about adding and removing noise from the output.