Securely Searching Big Data

Author: Rohith Perumalla

Date: 9/11/17

Subject: Encrypted Search & Cluster Formation in Big Data

Citations:

Siwach, Gautam, and Amir Esmailpour. "Encrypted Search & Cluster Formation in Big Data." ASEE, 3 Apr. 2014.

Summary:

Big Data is a specialization in analytics where extremely large sets of data are analyzed computationally to understand patterns, trends, and associations. Big Data also looks into how all these patterns relate to humans their behaviors and interactions. The data comes together forming clusters that interconnect with databases. The security of these clusters and databases is very important because if decoded can contain lots of personal information about numerous users. A proposed solution recommends the implementation of a key and integrated search function to allow for fast safe results.

Analysis:

Big Data Analytics revolves around large data sets, extraordinarily large data sets. These data sets are often protected with encryption. The encryption is set in place to ensure the integrity of the data keeping private data private while maximizing the accuracy of the results. Big Data is also an evolving technology, with the expectations in accuracy to increase. The recent growth in devices in the Internet of Things is increasing the amount of data gathered exponentially, in addition to increased awareness about data collection calls to attention the importance of keeping things secure.

In Big Data there are several points at which data is vulnerable in its travels from databases to clusters to individual users. 4 core properties are used to judge the reliability of a database: Atomicity, Consistency, Isolation, and Durability (ACID). Data can travel from different networks to different devices processing it, to even different racks within the same storage facility. As data traverses these network and even when it is inactive in storage it is susceptible

to being misused or stolen. Data protection is extremely important as the data stored ranges "from personal, financial details to the data containing national security." The amount of data stored continues to grow as it grew almost 200% every year from 2010-2014. At every point of transfer of data, there are processes in place to ensure the integrity and confidentiality of the data. While there are numerous methods in place to encrypt data, they can make accessing the data difficult and time-consuming.

A proposed solution for this problem comes from Gautam Siwach, and Dr. Amir Esmailpour. They theorize a solution where a unique key is used to identify parts of encrypted data that the user wants and has access to, after identifying the data the user can pull it from the "Data Lake." one concern of implementing this approach would be that Atomicity, Consistency, Isolation, and Durability (ACID) could be compromised. However, testing revealed that this method could actually maintain Atomicity, Consistency, Isolation, and Durability (ACID) ensuring the integrity and safety of the data. Traditionally a user would use a search function which would return few to no results as most data was encrypted, and decryption would take too long especially with the large amounts of data being handled. The proposed integrated search function would use the key combined with the encryption type to identify parts of data that the user has access to and matches the search request to deliver data at much faster rates. Siwach and Esmailpour claim that the encryption process will "generate a unique code, which shall correspond to the ciphering technique" which will aid in maintaining the confidentiality of the data as each key is unique.

Big Data has been growing in popularity and will continue to grow as we continue to collect data from various sources, the introduction of new technologies like IoT's will facilitate exponential growth in the amount of data gathered. With all this data it is important to keep it secure, but also while making it accessible. Especially in light of the recent Equifax hack (Fall 2017), the confidentiality of data is extremely important. This new proposed solution to accessing data if performing as theoretically expected will allow data to stay secure and while being easily accessible.