Big Data Application and Privacy Protection, Can't Have Both?

(Feb 25, 2021)

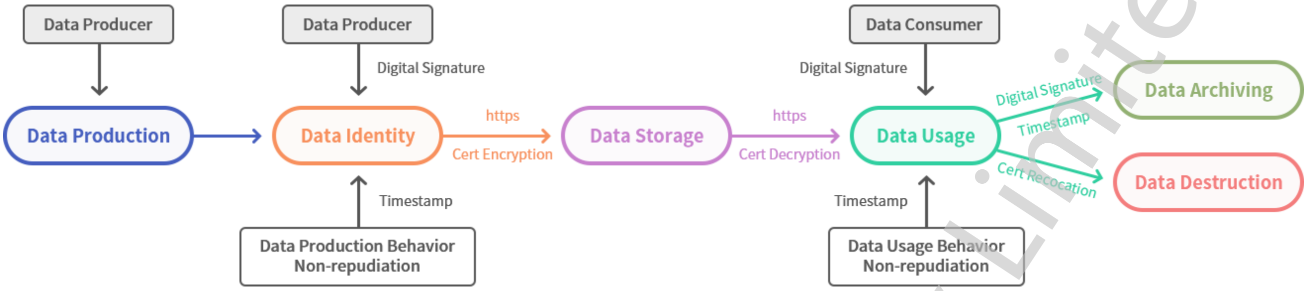
Many countries have begun to implement the national big data strategy, which mainly includes vigorously promoting the innovation and development of the big data technology industry, building a digital economy with data as a key element, using big data to improve the modernization of national governance, using big data to promote protection and improve people's livelihood, and practical protection national data security, etc. Under the guidance of this grand strategy, many countries big data application development has made certain achievements, but the resulting big data privacy protection issues are also many, and it has become an important key indicator of the healthy and stable development of big data. Big data application and privacy protection must have both in order to develop healthily. This article intends to put forward some personal opinions on big data collection, use, and storage.

Let’s talk about the life cycle of data first. All data will go through five periods. From the beginning of data generation to the identity the data, to data storage, and then to data use. Finally, it may be that data archiving is no longer used or data is destroyed. Of course, PKI (Public Key Infrastructure) technology is the only reliable technology to ensure the security of big data, which completely solves the four headache data security issues (**PAIN**): (1) the confidentiality of data (Privacy); (2) the identity authenticity of the data producer and consumer (Authentication); (3) data integrity (Integrity); (4) non-repudiation of data production behavior and usage behavior (Non-repudiation).

The important application of PKI technology is the digital signature and encryption application of digital certificates. As shown in the above figure, let's see how PKI technology protects big data security.

1. **Data Production**

The data producer can be a person or a thing (data collection terminal), and the producer produces the data.



2. **Data Identity**

After the data is produced, the data producer's identity certificate should be used to digitally sign the data to prove the identity of the data producer. Of course, the digital signature and timestamp are used to prove the non-repudiation of the data production behavior and the trusted data production time. At the same time, it can also prove the ownership of the data and the data identity, so that each piece of data has a trusted identity, which is used for data copyright protection, data traceability and legality validation before data processing, to prevent illegal data from entering the database to pollute big data. This step is very important, with multiple responsibilities to protect data from the source and ensure data quality. But this step of the current operation is basically missing.

3. **Data Storage**

After the data is produced, it has an identity and needs to be submitted to the cloud server for storage. The data submission must be encrypted and transmitted using https to prevent the data from being illegally stolen and tampered with during transmission. This security measure is very important, but the reality is that a large amount of data collection is submitted to the cloud server in cleartext, which is in urgent need of improvement. If the data is confidential data, you can also encrypt the data with the public key of the person who has the right to read the data before submitting it to the cloud server, to ensure that only the person who has the right to read the data can decrypt it, and to prevent the confidential data from being illegally used.

4. **Data Usage**

Data consumer must of course use https to browse data or download data, otherwise there is no guarantee that the data will be illegally stolen and tampered with during the data distribution process. If the data is encrypted, the data consumer needs to decrypt it with the private key of the encrypting certificate to read it normally. The most important thing is that not only the use of data must be controlled by user's right, but the user must use a digital signature to prove his legal identity and attach a timestamp signature to prove the time of data use. The digital signature and timestamp can not only confirm the user's identity, but also can effectively guarantee the non-repudiation and trusted time of the data use behavior for follow-up audit and traceability.

5. **Data Archiving** and **Data Destruction**

This step is the end of the data life cycle. You can use the archiving certificate digital signature and timestamp to archive and solidify the data. The purpose is not only to ensure that the data cannot be tampered with anymore, for future auditing and traceability if required, and the timestamp signature can prove the trusted archiving time. If you decide to destroy the data, it can be a physical deletion on the server. However, to destroy the data completely that has been delivered to the user, the encryption certificate can be revoked to destroy the data completely, to ensure that the data can no longer be used. This destruction operation requires that the data must be encrypted with a certificate before being delivered to the user.

As you can see from the above PKI data protection measures, if we use digital signature and encryption technology throughout the data life cycle, we can guarantee the security of the data throughout the life cycle. Big data applications and privacy protection are both available!

For everyone to understand the above concepts, let me illustrate with email data. Email is the second largest data source on the Internet. The number of emails per day is as high as 269 billion. These data contain a large amount of personal privacy information and business secrets, and even many highly confidential information such as financial payment and credit card information, but they are all transmitted in cleartext at present and stored in cleartext in the mail server, very insecure.

So, based on the five different periods of the big data life cycle mentioned above, let me talk about how MeSign Technology applies PKI technology to protect email data security and protect email privacy.

(1) **Data Production**

The user completes the data generation after writing the email, and the email account owner is the data producer.

(2) **Data Identity**

The user can use the digital signature function of MeSign App to add a digital signature to each email to prove the identity of the email, prove the identity of the data, and ensure that the data cannot be illegally tampered with. Once tampered, when the recipient receives the email, MeSign App or Outlook will prompt that the digital signature is invalid, and the data has been tampered with. At the same time, the MeSign email timestamp service attached automatically can also prove that the production time of this data is trusted.

(3) **Data Storage**

MeSign App sent email has digital signature, encryption and timestamp by default, and the mail server also has a TLS implementation like https encryption to ensure encrypted transmission of emails to the mail server. The certificate encryption of the email itself ensures that the email is stored in ciphertext in the mail server, which ensures the storage security of confidential information, especially for users using cloud email services, which can ensure that the confidential information of the email will not be stolen, illegally tampered and will not be leaked.

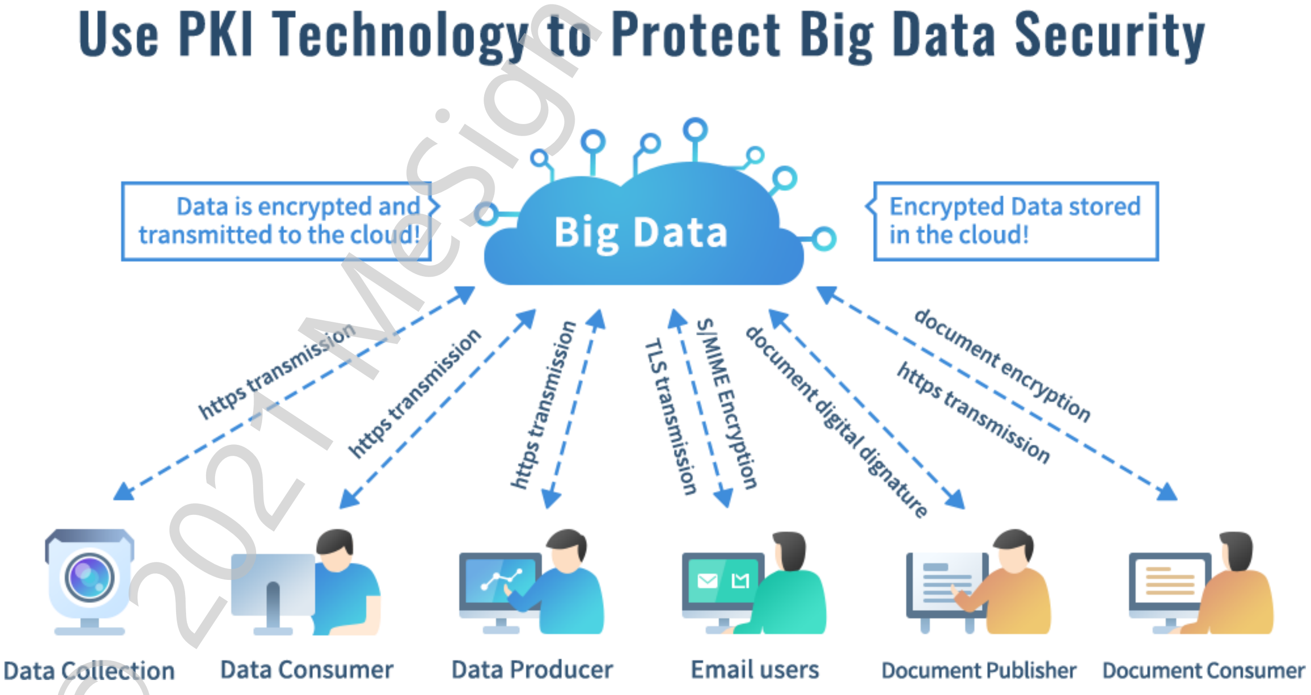
(4) **Data Usage**

After the recipient receives the email, MeSign App will automatically decrypt the encrypted email, validate the digital signature, display the sender's identity, and effectively prevent email fraud. Since the recipient also have a digital certificate, clarifies the real identity of the data consumer, being able to decrypt this email proves that the data consumer's use behavior is non-repudiation.

(5) **Data Archiving and Data Destruction**

If the email itself has been digitally signed and encrypted, it can be automatically archived without signing it again. If you want to destroy the data, simply delete the email completely in the mail server.

As you can be seen from the above example, MeSign App implements automatic email encryption and digital signature and timestamping, which can effectively help users realize the secure use of email data and privacy protection. I believe that no matter what data is, especially the data collected by data collection terminals, all data should be protected by applying PKI technology to ensure the security of its entire life cycle and identify the identity of the data from the source of the data, encrypted transmission to the cloud, and encrypted storage in the cloud, and it should also be encrypted for data exchange and data use. And take effective measures to truly destroy the data that should be destroyed and use digital signature and timestamp to solidify the archived data. Only in this way can we ensure that both big data applications and privacy protection have both.



-- END --

Want to contact me to discuss this topic? Please use MeSign App (👤 - ⛶) to scan the QR code and send me an encrypted email, I will reply to your encrypted email ASAP.