

## Digital security implementation in big data using Hadoop

Gupta, Palak ✉

Shobhit University, India ([palak.gupta2588@gmail.com](mailto:palak.gupta2588@gmail.com))

Tyagi, Nidhi

Meerut Institute of Engineering & Technology, UPTU, India ([mnidhiy@rediffmail.com](mailto:mnidhiy@rediffmail.com))

**Received:** 16 September 2015

**Available Online:** 14 February 2016

**Revised:** 13 November 2015

**DOI:** 10.5861/ijrsc.2016.1334

**Accepted:** 29 November 2015

ISSN: 2243-772X  
Online ISSN: 2243-7797

OPEN ACCESS



### **Abstract**

The security of Big Data is of concern. As the amount of data increases, more and more companies of the large repositories of data for the storage, and extraction of data. Big data offers a tremendous competitive advantage for the companies that help to adapt their products to the needs of data, identifying and minimizing the inefficiency of enterprises, and the dissemination of data with user groups. Fuse the big data supplied with their own challenges apart from a target of great value. This is not fundamentally different from more data security as the security of the traditional data. Big data security challenges arise because of the difference and not incrementally basic elements. In this paper we had discussed all the major and minor challenges of Big Data followed by the literature review in which we have focused on the security issue to be resolved here. Then the work along with the steps is discussed.

**Keywords:** security; big data; challenges; keys; Hadoop; TrustStore; KeyStore

## Digital security implementation in big data using Hadoop

### 1. Introduction

The world is becoming digitalized and interconnected and the amount of data in our world has been exploding. Data is increasing on daily basis then to manage the records it require extremely powerful business intelligence. Big Data refer to the analysis of significantly large collection of data that may contain user data, sensor data or machine data. An analyzed Big Data can deliver new business insights, open new markets and create competitive advantages. An analyzed Big Data can deliver new business insights, open new markets and create competitive advantages. It consists of data sets that are of large magnitude (Volume), large collection of data which diverse representation include structured, semi structured, or unstructured data (Variety), and should arrive fast (velocity). The real value of data is observed after analyzing i.e. after finding patterns, deriving meanings, making decisions, the ultimate data is available whichever required (Gupta & Tyagi, 2015).

Hadoop is an open-source software framework developed in Java for distributed storage and processing of very large data sets on clusters built from corresponding hardware. It is designed to scale from a single server to thousands of clients, with high degree of fault tolerance. Rather than relying on high-end hardware, the flexibility of these clusters comes from the software's capability to detect and handle failures at the application layer.

Apache Hadoop explored a new way of storing and processing data. Instead of relying on expensive, proprietary hardware and different machines to store and process data, It enables distributed parallel processing of large amounts of data across inexpensive, standard servers that both store and process the data, and can scale without limits (Big Data, 2015). With Hadoop, no data is as bigger. And in today's digitally connected world where huge amount of data is being created every day, Hadoop's improved advantages mean that businesses and organizations can now find valuable data that was recently considered useless. The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on belonging hardware. It is very similar to the existing distributed file systems (Zhao & Wu, 2014). However, the differences from other distributed file systems are momentous. HDFS is remarkably fault-tolerant and is designed to be deployed on low-cost hardware. The security of Big Data has also become one of the most challenging factors. As the data sets are increasing dramatically the storage and analysis also increases and because of that security should be increased as an unauthorized person can also download the data of a particular person and can harm him. Different surveillance tools and trackers are used to track the attacker but as the data is increasing the security also should increase. Overall, collection, storage, analysis and usage of personal data are part of our everyday life at all levels of security.

#### *1.1 Security challenges in big data*

Security and privacy concerns are growing as big data becomes more and more accessible. The collection and aggregation of massive quantities of heterogeneous data are now possible. Large-scale data sharing is becoming routine among scientists, clinicians, businesses, governmental agencies, and citizens. However, the tools and technologies that are being developed to manage these massive data sets are often not designed to incorporate adequate security or privacy measures, in part because we lack sufficient training and a fundamental understanding of how to provide large-scale data security and privacy. We also lack adequate policies to ensure compliance with current approaches to security and privacy. Furthermore, existing technological approaches to security and privacy are increasingly being breached, whether accidentally or intentionally, thus necessitating the continual reassessment and updating of current approaches to prevent data leakage (Computing Research Association, 2015).

Private businesses, hospitals, and biomedical researchers are also making tremendous investments in the collection, storage, and analysis of large-scale data and private information. While the aggregation of such data presents a security concern in itself, another concern is that these rich databases are being shared with other entities, both private and public (Govindaraji & Taeib, 2015).

While regulations regarding data access and use apply to health care providers and biomedical researchers, few (if any) regulations exist to protect the individual consumer from unauthorized use of his/her data by industry, and few (if any) incentives exist for private industries or other entities to comply with established privacy and security policies and procedures (Gao et al., 2013).

There are many other challenges that Big Data is facing. It includes (Brust, 2015):

**Big Data Analytics** - Big Data analysis need to store efficient data and to query large data sets, thus the techniques which make complete data sets instead of sampling are focused. These implications are in areas like machine learning, pattern recognition, and many others. Thus some of the methods are required to move beyond standard data mining techniques it includes:

- A new efficient algorithm
- A technology platform with adequate development skills to be implemented.

**Security** - Because the Big Data is stored in data organizations the problem which is faced is of encryption. Data cannot be sent encrypted by the users if the data need to perform operation over data. When producing information for big data, organizations have to ensure that they have the right to balance between utility and privacy.

**Context Awareness** - Demand of Context awareness is increasing with the increase of integrating heterogeneous data. When the data from different data repositories are integrated then the requirement of the data arises which is task relevant. Then that data can only be achieved when one is aware of the context of which the data is integrated. Context awareness has its crucial role to achieve optimized management of resources, systems, and services in many application domains. Context awareness focuses on some portions of data by hitting on the task relevant or application relevant events.

Context awareness is useful for resource consumption by generating Big Data only on the sources that are depending on currently applicable context.

**Visual Analysis** - How data seems to be for a human eye? It is the combined effort of humans and the electronic data processing. This can be achieved through graphics, different techniques are required which includes a variety of multidimensional, spatial, temporal data and solving problems. Visual analysis includes data analysis, problem solving and decision making (Intel IT Center, 2015).

**Data Efficiency** - Efficiency is based upon the performance and scalability to deal with the huge volume of data to be stored and processed by Big Data technologies. Effective solutions are needed to deal with data volumes per second to find out the cost effective scalable, feasible and processing enormous quantities of data. Another issue for Big Data analysis includes time constraints, the data should be available as required by various organizations.

**Correlation** - The correlation between the attributes of data is required between the structural contextual data which signifies that with the manipulation in one type of data another attributes get affected. Techniques are needed for quality preferences or requirements to different tasks and to their interworking relationships.

**Distributed Storage** - Distributed data sets are used for Big Data Analytics because within one organization there are number of data sets that require analysis. To maintain the two characteristics volume and velocity is also a good challenge, and if security and legal issues are considered then the nature of data sets is the most

complex problem to solve (Purdue Education, 2015).

**Content Validation** - Validating the large amount of data is a major challenge, as there are a large number of sources such as social networking platforms, blogs, different types of contents such as comments, articles, tweets, etc. thus to validate that type of content the organizations are taking recommendations from the users. The only need is to develop an algorithm that requires the feedback from the user and then can update the rules accordingly

## 2. Literature review

The biggest challenge for the big data from the point of view of security is the protection of the data of the user. Big data often contains large amounts of personal information and the privacy of the users is therefore a concern. Because of the large number of recorded data, the most important data can relate to criminal offenses more devastating consequences, the lack of data we are normally in the press. Because a major breach of security of the data affect a number a lot more people with consequences not only a position reputation, but with huge legal consequences. In the production of information for the big data, organizations need to ensure that the correct balance between the value of the data and the right to privacy. Before the collection of data must be appropriately anonymized, provides a unique identifier for the user. Alone can be a challenge in the area of security and the abolition of the unique identifier may not be sufficient to enable the information remain anonymous. Therefore, the clouds to carry out operations with encrypted data without the knowledge of the value behind the plain text.

In the use of large data is a big challenge, how the property of the information. If the data is to be stored in the cloud a limit value of trust should be between the owners of data and the storage of data owners. Mechanisms for monitoring a reasonable access are essential for the protection of privacy. The access control has been traditionally provided by operating systems or applications restriction of access to information, the presentation in the general, all information is the system or application.

## 3. Methodology

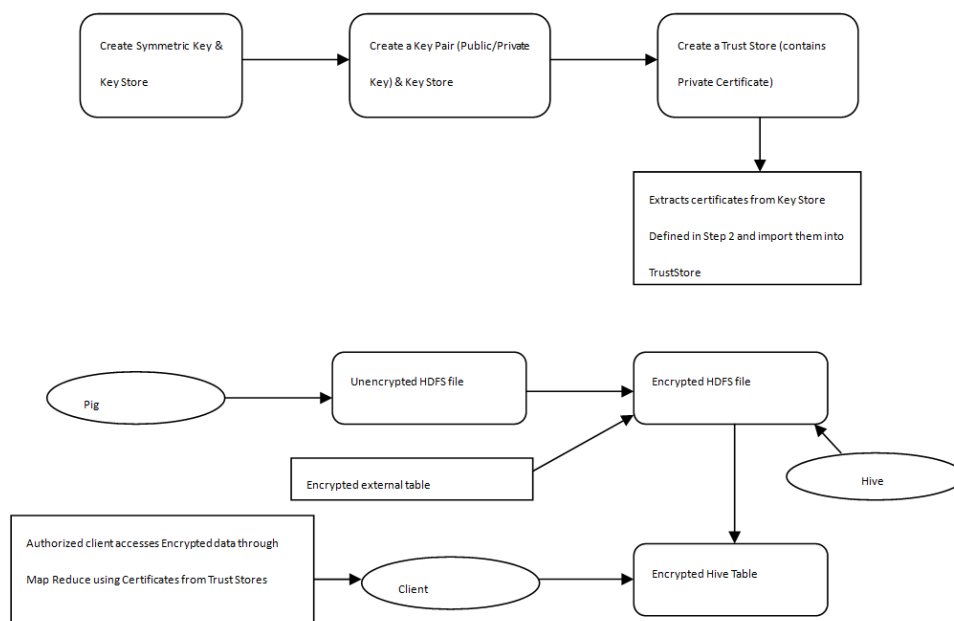


Figure 1: Data Security Methodology

Public key encryption systems are ideally suited to digital security. For example, a publishing company can

first encrypt a contract using their own private key and then encrypt it again using the author's public key. The author can use his private key to decrypt the first level of encryption, and then use publisher's public key to decrypt the inner encryption to get to the contract. After that, the author can "sign" it by creating a hash value of the contract and then encrypting the contract and the hash with his own private key. Finally, one more layer of encryption is added by encrypting again using the publisher's public key and then e-mail the encrypted contract back to the publisher. Because only the author and publisher have access to their private keys, the exchange clearly is enforceable and uniquely authentic.

The hash function and checksum confirm immutability (assuming an initial checksum of the contract was computed and saved for comparison), while the frequency and timestamps of the e-mails ensure one-time recent usage.

## PROPOSED WORK

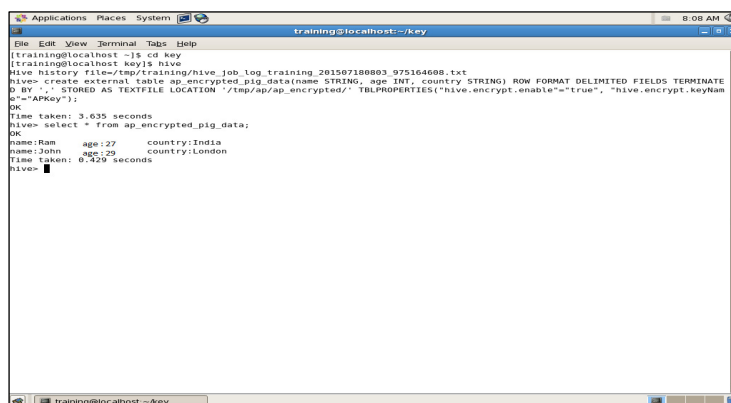
In this paper the technique used for digital security is using **Algorithm**:

- Step 1: Create a Secret Key and KeyStore.
- Step 2: Next is to adjust the permission of the KeyStore with the certificate "600" for giving right to the owner so that he can read or write a file. All other have no right.
- Step 3: Create a Key pair (public/private Key) and KeyStore.
- Step 4: Next is to create a TrustStore.
  - ❖ Extract the certificate from the newly created KeyStore.
  - ❖ Create TrustStore containing the public certificates.
  - ❖ Create clusterpublic.TrustStore ownership to root, group to Hadoop and permissions "644" so that the owner may read and write a file, while all others may only read the file.
  - ❖ Create a file TrustStore.passwords, set its permission to "644", and add password contents.
- Step 5: Copy the /keys directory and all its files to all the other nodes in the cluster. On each node, the KeyStore directory must be in a specific Hadoop folder.
- Step 6: With the TrustStore ready a text file is created to use for testing encryption and copied to HDFS.
- Step 7: Start Pig then after taking it to grunt prompt. Set the environment variables which include KEY\_PROVIDER\_PARAMETERS,AGENT\_SECRETS\_PROTECTOR,AGENT\_PUBLIC\_KEY\_PROVIDER, AGENT\_PUBLIC\_KEY\_PROVIDER\_PARAMETERS, AGENT\_PUBLIC\_KEY\_NAME, pig.encrypt.KeyProvider.Parameters.
- Step 8: Next we have to read the text file from HDFS and then encrypt it, and store it into the same location in a directory meant for encryption.



```
2015-07-18 07:33:48.263 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 30% com
2015-07-18 07:33:51.888 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% co
2015-07-18 07:33:51.818 [main] INFO org.apache.pig.tools.pigstats.PigStats - SCRIPT STATISTICS
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
HadoopVersion  pigVersion  UserId  StartedAt  FinishedAt  Features
1.2.0-cdh3u2  0.9.1-cdh3u2  training  2015-07-18 07:33:35  2015-07-18 07:33:51  UNKNOWN
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
Success!
Job Stats (time in seconds):
JobID  Maps  Reduces  MapMapTime  MinMapTime  AvgMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Alias
-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
job_201507180715_0881  1  0  3  3  3  0  0  0  raw /tmp/app/ap_0n
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
Input(s):
Successfully read 2 records (418 bytes) from: "/tmp/app/ap.txt"
Output(s):
Successfully stored 2 records (168 bytes) in: "/tmp/app/ap_encrypted"
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
Summary:
Total records written : 2
Total bytes written : 68
Spillable Memory Manager spill count : 0
Total heap proactively spilled: 0
Total records proactively spilled: 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
Job Info:
job_201507180715_0881
2015-07-18 07:33:51.016 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encount
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
2015-07-18 07:33:51.017 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
grunt>
```

- Step 9: Next is the Hive component of Hadoop. In which we can fire our queries and can get the appropriate result.
- Step 10: Set environment variables in Hive i.e. `hive.encrypt.master.keyName`, `hive.encrypt.master.keyProviderParameters.keyStoreUrl`, `hive.encrypt.keyProviderParameters.keyStoreUrl`, Set `mapred.crypto.secrets.protector.class,mapred.agent.encryption.Key.Provider,mapred.agent.encryption.key.provider.parameters`, `mapred.agent.encryption.keyname`, .
- Step 11: Next is to create external table `ap_encrypted_pig_data`.
- Step 12: once the table is created, decrypted data can be viewed by any authorized client(having appropriate key and certificate files within `/usr/lib/hadoop/keys` directory) using the selected query at hive prompt .



```
Applications Places System training@localhost:~/key
[training@localhost ~]$ cd key
[training@localhost key]$ hive
hive history file:/tmp/training/hive job log training_201507180803_075164608.txt
hive> create external table ap_encrypted_pig_data(name STRING, age INT, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATE
D BY ' ' STORED AS TEXTFILE LOCATION '/tmp/ap_encrypted/' TBLPROPERTIES("hive.encrypt.enable"="true", "hive.encrypt.keyNam
e"="APKey");
OK
Time taken: 3.635 seconds
hive> select * from ap_encrypted_pig_data;
OK
name:Ram      age:27      country:India
name:John     age:29      country:London
Time taken: 0.329 seconds
hive>
```

#### 4. Conclusions

Here in this paper the security issue which is tried to resolve is the one related to cryptographically enforced access control and secure communication, which we had tried to resolve with the help of shell programming in Hadoop framework with the help of encrypting the data at source and then decrypting it at the destination. Here we have performed it on a small record but as the Hadoop framework is compatible with large files which are of exabytes or petabytes so the files of this size can also be acceptable without any issue. The greatest advantage to work on Hadoop environment is that it works on Linux platform and which is more secure than any other operating system. Next is that here it includes hive so that if we want to take out the beneficial data out of that huge data after the processing then it can be done easily with the help of any SQL query in hive. In near future this method can be used to solve the issues of securely computations in distributed environment. In which we can transfer the encrypted files to the data in which there are multiple data stores which provides different types and forms of data. So to work in that environment is more complicated but it can be done in Hadoop environment with less complications and more security. The digital security can be used in e-commerce, e-governance fields. The pen and paper can be replaced by bits and bytes and many others. Modern communication tools have created almost limitless opportunities to improve information flow and processes, but they had not eliminated the legal, cultural and practical need for tangible and lasting representation of commitment. The digital security can be used to eliminate it and make it secure.

#### 5. References

- Big Data. (2015). The next frontier for innovation, competition and productivity. Retrieved from [http://www.mckinsey.com/insights/business\\_technology/big\\_data\\_the\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)

- Brust, A. (2015). *Top 10 categories for big data sources and mining technology*. Retrieved from <http://www.zdnet.com/article/top-10-categories-for-big-data-sources-and-mining-technologies/>
- Computing Research Association. (2015). *Challenges and opportunities of big data*. Retrieved from <http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf>
- Gao, W., Zhu, Y., Jia, Z., Luo, C., Wang, L., Li, Z., Zhan, J., Qi, Y., He, Y., Gong, S., Li, X., Zhang, S., & Qiu, B. (2013). *A big data benchmark suite from web search engines*. Paper presented in the Third Workshop on Architectures and Systems for Big Data in conjunction with The 40th International Symposium on Computer Architecture.
- Govindaraji, D., & Taeib, T. E. (2015). Fetching data from an Incognito window for the purpose of big data processing. *International Journal of Advanced Research in Computer Engineering & Technology*, 4(4), 1331-1333/
- Gupta, P., & Tyagi, N. (2015). *An approach towards big data- A review*. International Conference on Computing, Communication & Automation ICCCA, IEEE Conference. <http://dx.doi.org/10.1109/CCAA.2015.7148356>
- Intel IT Center. (2015). *The IT landscape for big data analytics*. Retrieved from <http://www.intel.in/content/dam/www/public/us/en/documents/guides/getting-started-with-hadoop-planning-guide.pdf>
- Purdue Education. (2015). *Challenges and opportunities with big data*. Retrieved from <http://www.purdue.edu/discoverypark/cyber/assets/pdfs/BigDataWhitePaper.pdf>
- Zhao Y., & Wu J. (2014). A data aware caching for big-data applications using the mapreduce framework. In *32<sup>nd</sup> Proceeding of IEEE conference on computer communications* (pp. 35-39).

