DOI 10.29042/2018-3406-3412

Helix Vol. 8(3):3406-3412

A Secure Framework for Authentication of Documents in Cloud

[1] Mr.Srikanth Ganta, [2] Dr Jayaram Pradhan, [3] Dr Narasimham Challa

[1] CSE Deparment, MVGR College of Engineering, VZM, Computer Science Department, Berhampur University, Berhampur, CSE Department, VIIT, Vizag

Email: srikanth@mvgrce.edu.in

Received: 10th Feb 2018, Accepted: 20th March 2018, Published: 30th April 2018

Abstract

Web has become a prominent storage option of late. But there are certain setbacks in the present web which has a shortfall of appropriate mechanism to create and store artifacts -- code, data sets text, image--in digital form which are needed to be unchangeable in any way and be verifiable and permanent. These issues hamper the reliability on the cloud for its functional productivity especially in the domain of science where the re-productivity of outcome process is highly vital. In order to overcome the setbacks, it is proposed a methodology with the stored data processing of encoding and decoding at cloud environment taking the support of cryptographic hash values. In this paper it is presented that how the model work out in verifying the digital artifacts with the help of the format of independent serialization for structured files like nano-publications. It is explained how the documents can be processed using cloud computing environment where the application is integrated with Hadoop installed on Amazon EC2 web service. The approach presented in this work confines to fundamental salient features in the aspect of architecture which is open and decentralized besides completely compatible with prevailing protocols and standards. On its evaluation of approach for its referential implementation exhibits the accomplishment of the design goals indeed stands good practically for the files even larger in size.

Keywords: Hadoop, Map Reduce, Big data, cloud computing

Introduction

With the digitalization, the universe is going at a scorching pace, many domains from small scale organizations to large scale oranizations are experiencing the impact of reassemble change in the amount of data that has been stored, processed and reported .All professionals are confronted with novel challenges and unexpected opportunities even as they come to grips with the changing dynamics of the digital era. Every year , empirical evidence indicates that users are spending more time online, perusing different digital artifacts.

Reasons for document authentication in digitalized word.

Over the past few years, the smart phase revaluation has captured the minds of customers, especially the younger generation , who are now engaged online almost 24x7. Be it bookings, purchasing, reviews or what have you , today's generation is willing to log on initially and use the digital medium to get things done , rather than setup out into the open and expand extra energy on doing the same things. Whether it is Google, Twitter, Facebook, Linkedin, Youtube, blogs and other digital channels, the virtual universe is giving a run for its money to the real world.The digital medium has been gaining rising traction as companies find they can reach they target audience at much lower spends compared to conventional marketing mediums. What's more , digital marketing tools can evaluate the efficacy of specific marketing campaigns. Thereby, digital marketers can decide how to optimise marketing spends more effectively.

So for Digital India large amount of data needs to be maintained and preserved in an orderly manner because if there occurs a case of mismatch or toss of data it may result in complete breakdown of the entire well built system. So to maintain huge amount of data the concepg^f Big Data has come into picture where maintenance and storage becomes easy. Big Data is more about how to organize this data, how to label different kinds of it (structured, unstructured, semi-structuned, internal and external), which technologies one uses to store it and retrieve and many other facets. But the computational speed infrastructure was very difficult because of which cloud computing has come into picture. Before the cloud era, tasks were expensive, technically challenging and possible to only a few Cloud Counting, it is the paradigm of computing on the fly as it can be envisioned. In Cloud Computing everything (or almost) is de-materialized and we don't need to worry a lot about the setup before starting, flje don't either need to care about managing a farm of computers before using powerful tools in today's computing environment, organizations are focusing on reducing cfgj and remain competitive. It departments face grater problems to ensure the needs of business and deliver the desired results in the most efficient and cost effective manner. To meet these challenges IT organizations are increasingly moving away from device-centric views of IT too that is focused on applications, information, and people and more towards the new paradigm of Cloud Computing. Cloud computing is a latest emerging computing techno log}' that uses the internet and central remote servers to maintain data and applications

So for Digital India large amount of data needs to be maintained and preserved in an orderly manner because if there occurs a case of mismatch or loss of

data it may result in complete breakdown of the entire well built system. So to maintain huge amount of data the concept of Big Data has come into picture where maintenance and storage becomes easy.

Cloud computing is a model for enabling ubiquitous, on-demand access to a shared pool of configurable computing resources (e.g., computer networks, servers, storage, applications and services). Cloud computing has a lot of advantages over traditional computing. The benefits of deploying applications using cloud computing include reducing runtime and response time, minimizing the risk of deploying, physical infrastructure, lowering the cost, and increasing the pace of innovation. Cloud computing offers both the software and hardware as a service over the internet.The vendors/service provider of the cloud computing have major responsibility in developing confidence among their clients in terms of authentication, intangibility, privacy e.t.c from unauthorized access. The digital artifacts such as data sets, codes, texts, and images must be verifiable and permanent. For this purpose, a combination of message digest and security hash is proposed to handle the data set and to generate tokens corresponding to various data sets. At the receiving end , these tokes will be used to derive the identification and authentication of the data.

Related Work

Suggested multiple output identification methods from a hash function. It proposed a novel scheme for the purpose of mapping these to HTTP URLs, binary and human speech formats. These formats are such that they doesn't require a strong link to the referenced object which results in the same degree of authentication as that of the reference to it. The (ni) s will serve as standard methods in using standard hash functions outputs in names.

This paper uses SHA – 256 algorithm to generate and accept names based on the same algorithm. The implementation might support some more hash algorithms and may be used for specific names. The hash value is capable of producing name – data integrity bond between names and bytes. The truncated hash values are used but for certain security properties will be affected in the security considerations.

T. Kuhn and M. Dumontier[3] proposed about Trusty URIs which are genuine, inconvertible and everlasting digital artifacts for linked data. Their proposal emphasizes on inclusion of cryptographic hash values in URIs. They called them trusty URIs and detailed the usage for approaches like nanopublications to verify Specific resources as well as the entire tree. The trusty URI's included the cryptographic hash values in the URI's. There are three implementation part in the code part, they are- java, Perl, python in the trusty URI's. Identification of Digital artifacts can be done both on byte level as well as more abstract level such as in the case of RDF graphs. This means the resources retain their hash values even after presenting in different formats.The hash generation and checking on nano publication using the function transform nanopub. .Java and python render in adequate to check these two lines containing metadata in the case of RDF implementation. These trusty URIs might contribute quite significantly in future web publishing.

H. Van de Sompel et.al[4]. proposed Persistent identifiers for scholarly resources and on the web: In this paper they discussed about the need for an unambiguous mapping and explained how these document identifiers play a major role in identifying a wide range of web resources like research papers, datasets, images etc.. Initially the HTTP and the web conceived as a research communication, concerns about the long term of HTTP URI. It identified a decoupled from locating and access is introduced in the result. The communication in research is both understandable and justifiable in the use of Persistent Identifiers(PID's). In PID and HTTP URI's mapping an unambiguous mapping is proposed between the PID oriented paradigm and HTTP oriented web. The research communication environment more machine friendly and can pave the way for new value added applications.

M. Bellare, O. Goldreich, and S. Goldwasser[5] proposed "Incremental cryptography: The case of hashing and signing," The authors says about this document is introduced incremental algorithms for cryptographic functions. By this algorithm the digital signature is used as an example in the functions and the underlying messages are modified made easy update upon using digital signatures. the increment is suitable for undergoing cryptographic transformation which are altered versions of documents. The hash functions we need to extend and implement some usual definitions to allow independent security considerations as a parameter. we have to expand the scope of this research. Insertion and deletion are the complex update operations in the messages. In other way we can consider the fingerprints and message authentication.

McCuskeret.al[6] in their research publication proposed about "Functional requirements for information re- source provenance on the web," and discussed how many different ways the semantics of HTTP transactions can be interpreted. The mechanisms related to these abstractions such as content negotiations are well established but not the semantics behind these abstractions.Informational resources are understanding critical requests in the URI's. The provenance of web information resource access should represent on the web. Functional Requirements for Bibliographic Resources are the mature model which is taken from the science community. The Functional Requirements for Informational Resources are extension for the FRBR. In the FRBR, FRIR helps to describe the relation

between the URI and the HTTP to represent. The focus on the URI's is the identity of the information resources.

R. D. Peng[7] in his paper explained about "Reproducible research in computational science," .He explained that in the case where there is no possibility of full independent replication of a study, this reproducibility serves as minimum standard for the judgment of scientific claims. The reproducibility standards are based on the experiments that are done in the computational science the direction in which the research is on the rise and rapid these days. These computer codes that available to others provide a detail level of regarding the analysis on analogous non-computational experimental descriptions. The field of science will not change overnight, but it simply changes the notion of reproducibility to make a difference.

T. Kuhn et al in their work proposed "Publishing without publishers: a decentralized approach to dissemination, retrieval, and archiving of data," They explained in their proposal that the current methods used for the publication of scientific datasets are not efficient, reliable and acceptable which has become increasingly important now-a- days. The results of verifiability and reproducibility impair seriously on the data related to scientific inventions and they get disconnected from the data. The HTTP GET requests is the one of the possible architectures of the semantic web application. URI's provide the resolvable data based on the application performs the task. The development of core services is to find the advanced services over the sever network. The main limitation here is the disc space.

E. H¨ofig and I. Schieferdecker[9] proposed "Hashing of rdf graphs and a solution to the blank node problem," and explained about the ability to calculate hash values is fundamental for using cryptographic tools, such as digital signatures, with RDF data. It is difficult to implement tamper-resistant attribution or prove- nance tracking, both important for establishing trust with open data. In closed systems the access is strictly regulated and it is possible to record the information in a trustworthy manner. The problems and related work has the even standards seemed directly applicable. In this problem none of the article has not fully explained towards the solution. . It also solves the blank nodes labeling by encoding the complete context of a node. The blank nodes in a graph are smaller when compare to the real execution speed is less.

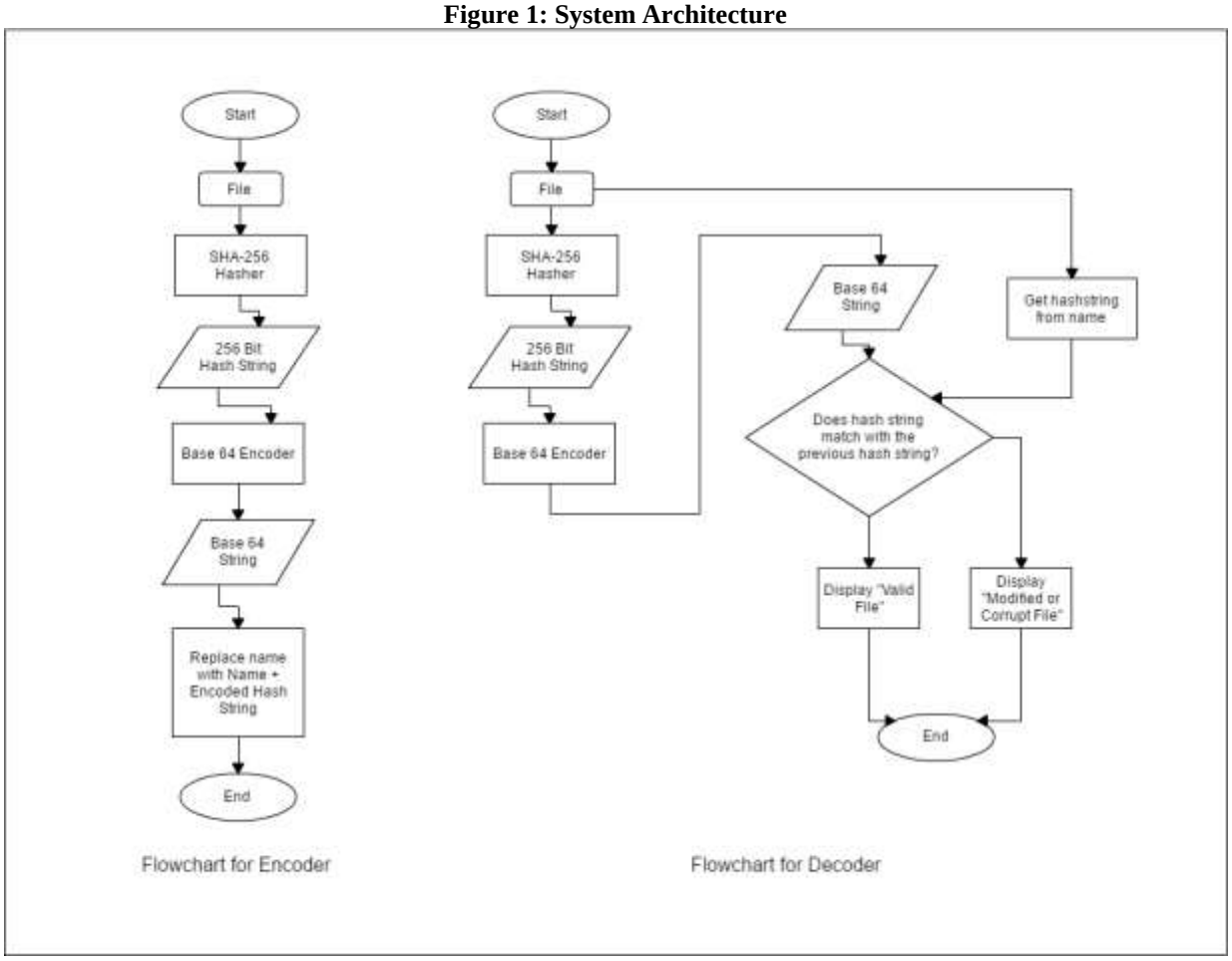
Sean Bechhofer et al in their work "Research Objects: Towards Exchange and Reuse of Digital Knowledge" have explained how the web has become the most preferred platform for publications of scholarly articles and there is a certain need to supporting mechanisms. There are some methods to analyze the data simulate the data that manipulate and produce the data like by scientific workflows, research protocols. The research objects can create and encapsulate the digital knowledge. The research object has principles like reuse; repurpose able, repeatable, reproducible, replay able, traceability, and aggregation. Publication Object is immutable and it is intend of a record of activity. It is considered as distinct objects but it cannot be produced. The electronic publication follows the paper metaphor, which supports the reusable shared research. It leads to greater transparency for the validation results.

James P. McCusker et al proposed a model for "parallel identities for managing open government data" and elucidated about how government bodies from local level to national level are publishing their data for usage of the public. This public data can potentially improve the quality of life for the society, for doing business and even for the government for offering better services. In the integrating government data it develops and combines the design of URI and methodology for data transformation, retrieve, collect and enhance the original government data. the library science of Functional Requirements for Bibliographic Records to manage bibliographic resources. By using cryptographic digests the digital information resources can be apply by FRBR abstraction of levels. The ability of transformations has simpler users to provenance of the information.

System Model

The system model for implementing the methodology is shown in the following figure 1 where the data sets will be processed by encoder algorithm with SHA-256 and the same will be applied by using decoder algorithm to know whether the artifacts are modified or not in the cloud environment.

Figure 1: System Architecture



Methodology

The Framework is implemented by using the Enocder algorithm1 in which python libraries are used.

The algorithm uses SHA256 for calculating hash for the artifacts transmitted through cloud. At the receiver side the decoding is achieved by using algorithm2

Encoder: Algorithm1

```
d = get directory name from user
Change directory to d
For all files in directory do
{
content_string = read content from file
encoded_string = SHA256(content_string)
base64_string = base64(encoded_string)
rename file with base64_string
}
```

Decoder: Algorithm2

```
d = get directory name from user
Change directory to d
For all files in directory do
{
content_string = read content from file
encoded_string = SHA256(content_string)
base64_string = base64(encoded_string)
if name of file == base64_string
```



```
    display that the file is valid
else
    display that the file is not valid/corrupted
}
```

Experimentation

The experimentation was done on a cluster, equipped with Hadoop 2.7.1 on top of Ubuntu 14.04LTS installed where there are four data nodes and 1 name node were configured with python as programming language. The framework deployed on Amezon EC2 webserver.

Results

The following figure shows the screenshot for executing checker methodology for given file or set of files or dirctory.

Figure: checker methodology for given data set

The following figure shows the screenshot for executing encoder methodology for given file or set of files or dirctory.

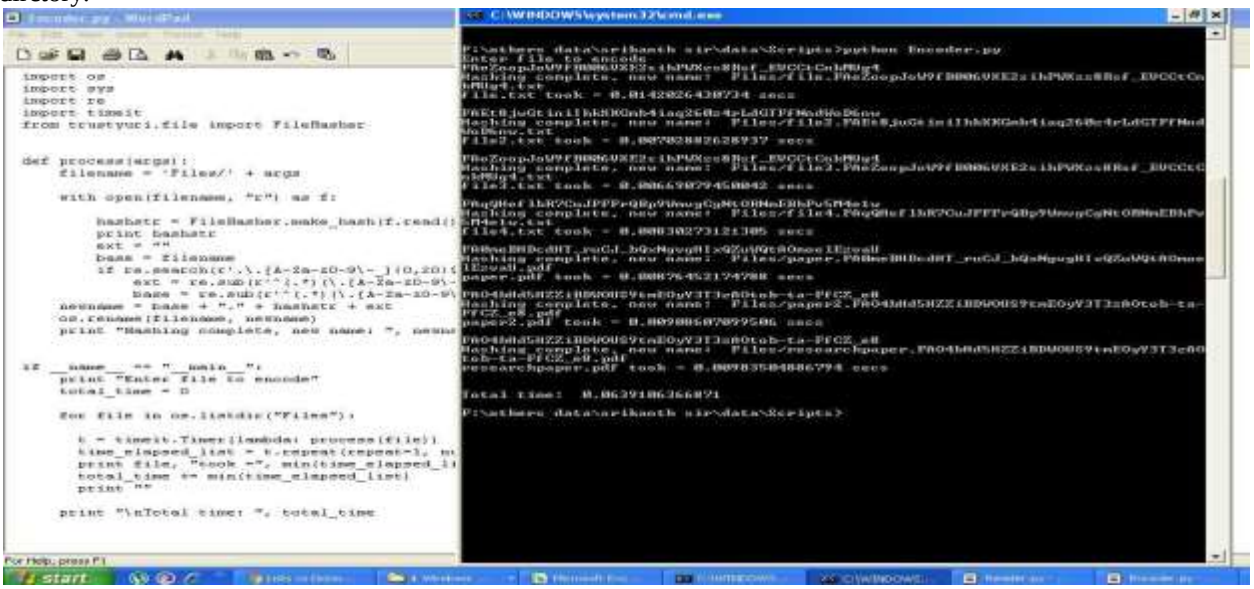


Figure: Encoder methodology for given data set

The following figure shows the screenshot for executing encoder methodology for given file or set of files or dirctory.

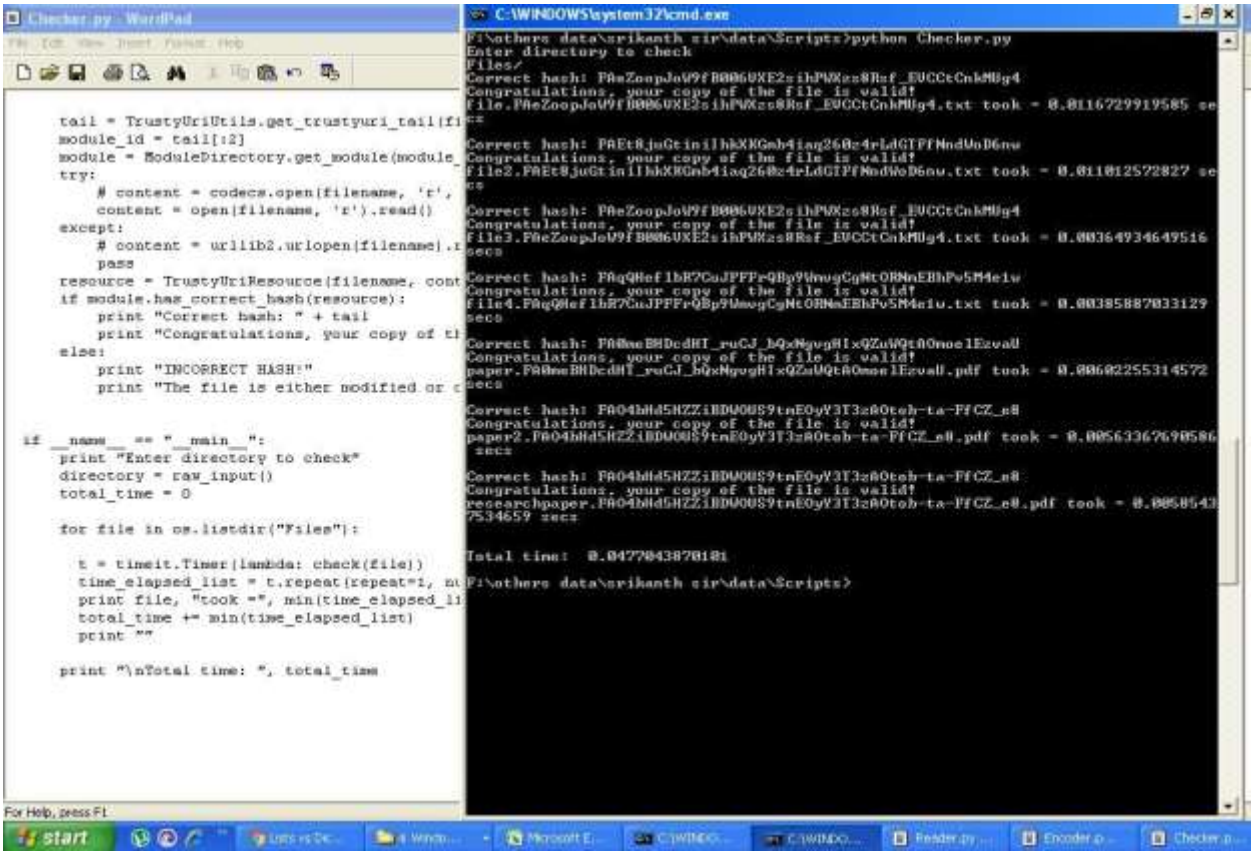
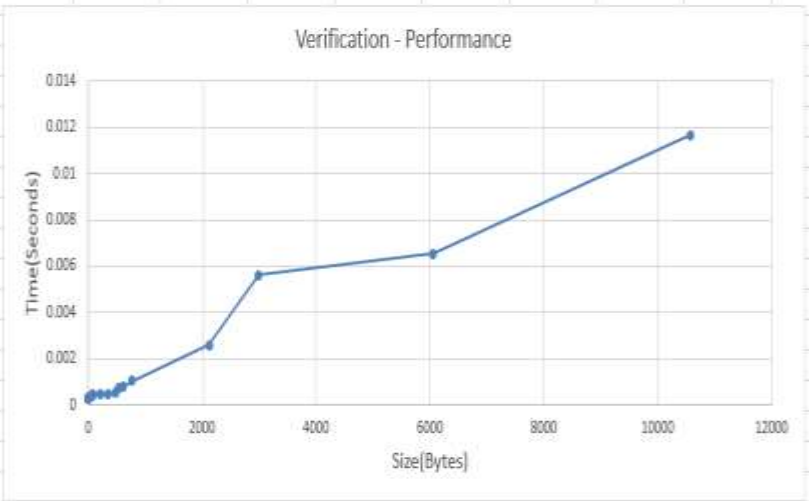


Figure: Decoder methodology for given data set

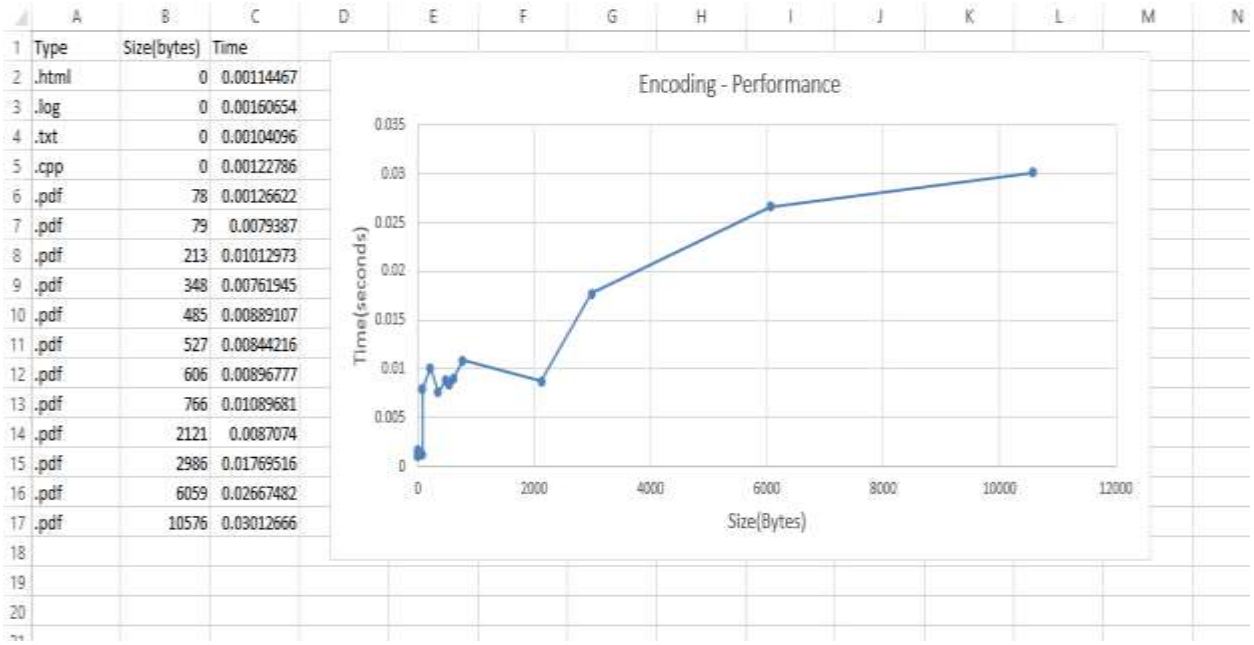
Graphs

To test our approach and to evaluate its implementations, we first took a collection of Ansciallary data set and the resultant graph for encoding is drawn where in time is constantly stable as the data set size is increasing is shown in th graph1.The resultant graph for decoding is drawn where th time is constantly stable as the data set size is increasing is shown in th graph2

	A	B	C
1	Type	Size(bytes)	Time
2	.html	0	0.00026686
3	.log	0	0.0002674
4	.txt	0	0.0003252
5	.cpp	0	0.00025173
6	.pdf	78	0.00036517
7	.pdf	79	0.00044728
8	.pdf	213	0.00048023
9	.pdf	348	0.0004516
10	.pdf	485	0.00054452
11	.pdf	527	0.0007082
12	.pdf	606	0.00079085
13	.pdf	766	0.00103825
14	.pdf	2121	0.00257943
15	.pdf	2986	0.00562667
16	.pdf	6059	0.00653636
17	.pdf	10576	0.01167253
18			
19			



Graph1: Result analysis for encoding



Type	Size(bytes)	Time
.html	0	0.00114467
.log	0	0.00160654
.txt	0	0.00104096
.cpp	0	0.00122786
.pdf	78	0.00126622
.pdf	79	0.0079387
.pdf	213	0.01012973
.pdf	348	0.00761945
.pdf	485	0.00889107
.pdf	527	0.00844216
.pdf	606	0.00896777
.pdf	766	0.01089681
.pdf	2121	0.0087074
.pdf	2986	0.01769516
.pdf	6059	0.02667482
.pdf	10576	0.03012666

Graph2: Result analysis for decoding

References

1. Tobias Kuhn and Michel Dumontier,"Making Digital Artifacts on the Web Verifiable and Reliable", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING Vol NO 99 YEAR 2015
2. S. Farrell, D. Kutscher, C. Dannewitz, B. Ohlman, A. Keranen, and P. Hallam-Baker, "Naming things with hashes," Internet Engineering Task Force (IETF), Standards Track RFC 6920, April 2013.
3. T. Kuhn and M. Dumontier, "Trusty URIs: Verifiable, immutable, and permanent digital artifacts for linked data," in Proceedings of the 11th Extended Semantic Web Conference (ESWC 2014), ser. Lecture Notes in Computer Science. Springer, 2014.
4. H. Van de Sompel, R. Sanderson, H. Shankar, and M. Klein, "Persistent identifiers for scholarly assets and the web: The need for an unambiguous mapping," International Journal of Digital Curation, vol. 9, no. 1, pp. 331–342, 2014.
5. M. Bellare, O. Goldreich, and S. Goldwasser, "Incremental cryp- tography: The case of hashing and signing," in Advances in Cryptology CRYPTO'94. Springer, 1994, pp. 216–233.
6. J. McCusker, T. Lebo, A. Graves, D. Difranzo, P. Pinheiro, and D. McGuinness, "Functional requirements for information re- source provenance on the web," in Provenance and Annotation of Data and Processes. Springer, 2012, pp. 52–66.
7. R. D. Peng, "Reproducible research in computational science," Science, vol. 334, no. 6060, p. 1226, 2011.
8. T. Kuhn, C. Chichester, M. Dumontier, and M. Krauthammer, "Publishing without publishers: a decentralized approach to dissemination, retrieval, and archiving of data," arXiv preprint arXiv:1411.2749, 2014.
9. E. H¨ofig and I. Schieferdecker, "Hashing of rdf graphs and a solution to the blank node problem,"in 10thInternational Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2014), 2014, p. 55
10. Sean Bechhofer, David De Roure, Matthew Gamble, Carole Goble, Iain Buchan "Research Objects: Towards Exchange and Reuse of Digital Knowledge" 2010.
11. James P. McCusker, Timothy Lebo, Cynthia Chang, and Deborah L. McGuinness, Rennsselaer Polytechnic Institute "Parallel Identities for Managing Open Government Data" Published by the IEEE Computer Society, 2012.

