A Framework for Leveraging Cloud Computing to Facilitate Biometrics at Large-Scale

John K. Mitchell III and Syed S. Rizvi
Department of Information Sciences and Technology
Pennsylvania State University, Altoona, PA 16601
Email: {jkm5270, srizvi}@psu.edu

Submitted to 2014 International Conference on Security and Management (SAM’14), Track: Biometrics & Forensics

Abstract—The cloud computing paradigm provides an efficient environment for the large-scale use of biometric identification systems. However, there is no established standard for the configuration of the cloud or the type of components necessary to achieve optimal biometric performance. In this paper, we present a conceptual framework which addresses the challenges and the requirements of the biometrics at large-scale while dramatically increasing the performance. In this paper, our contributions are three-fold. First, we describe the related work pertaining to biometric integration with cloud computing. Second, we discuss the issues and challenges preventing biometrics from being adopted at large-scale which influenced the design choices of our proposed framework. Finally, we present the proposed framework and highlight the cloud services that should be utilized with the proper configurations to maximize the biometric efficiency. We also discuss the realization of our proposed framework and the components selected. To show the practicality of the framework, we chose Amazon Web Services even though our framework is applicable to any cloud platform that supports the design criteria.

Keywords—biometrics; cloud computing; large-scale implementation; authentication; MapReduce; Hadoop

I. INTRODUCTION

Biometric technologies are replacing traditional password-based methods as a more effective authentication process in the realm of information security. In addition, biometrics are used for visas and passports, voter registration, border control, and credit card transactions [1]. The concept of biometric identification is especially common in the forensic domain. For instance, the Integrated Automated Fingerprint Identification System (IAFIS) [2] is used by the Federal Bureau of Investigations (FBI) to perform crime scene investigations, criminal background checks, or bank employee checks. Furthermore, biometrics are becoming increasingly popular in other areas of the world as well. For example, India has compiled the largest biometric database in the world [3] with the goal of eventually enlisting 1.2 billion citizens. Therefore, biometrics are expected to continually grow exponentially in the next few years. As the acceptance of biometric identification increases, there will be a higher demand from the public and the private sectors. This demand facilitates the need for large-scale biometric systems. However, there are many challenges and performance issues that prevent large-scale biometric applications from becoming fully realized. As a result, cloud computing has been proposed as a viable solution to the inhibitors associated with the biometrics.

The cloud has many attractive features that could potentially benefit biometric systems and enable large-scale use which include: parallel processing, large storage space, elasticity, real-time support, and automated failover. Moreover, the highly scalable and flexible nature of the cloud reinforces the idea of integration with biometrics to increase performance. When implementing biometrics with cloud computing, there is a steep learning curve to understand the possible cloud services, configurations, and additional components. Our background and experience with the cloud, specifically Amazon Web Services (AWS) [4], allow us to formulate the necessary services and configurations to abstract the challenges and meet the performance requirements. There were several assumptions about the biometric systems which were taken into consideration when designing our framework. We assume that a multimodal system will be utilized since it provides greater reliability than unimodal systems. One example of a multimodal database is the Next Generation Identification (NGI) program [2] that is being developed by the FBI and is scheduled to replace IAFIS in the summer of 2014. A multimodal database significantly increases the quantity of data being stored and computational power required for processing. Additionally, the support of a fused algorithm to compute a decision from multiple modalities may be required. In this research, we design our framework to be user-friendly and simple for non-experts, enable the most efficient integration possible, be applicable to most cloud platforms, and facilitate large-scale use of biometrics.

The rest of the paper is organized as follows: We begin in Section II by describing the related work. In Section III, we identify the challenges and requirements to be addressed in our framework. A comprehensive discussion on our cloud framework will be presented in Section IV with the emphasis on the configurations of the cloud services. In Section V, we provide components that meet the specifications presented in our framework and examine each component individually.
Finally, we conclude the paper and present future directions in Section VI.

II. RELATED WORK

The existing works pertaining to biometric integration with cloud computing are scarce. Moreover, these works do not provide a comprehensive solution to the challenges and requirements of biometrics in the form of a generic framework. However, there are a few relevant publications that are similar in some aspects to our paper.

For instance, the authors [5] primarily discuss the MapReduce function and present an implementation involving similar components described in our paper. On the contrary, they only focus on the advantages of parallelism over sequential processing of data. They do not discuss other benefits of the cloud or offer a comprehensive framework. Peer et al. [6] present a case study that involves fingerprint recognition for the implementation. However, the authors are more concerned with how to move the biometric platform into the cloud and the challenges that could arise rather than the actual implementation.

Kohlwey et al. [7] present a prototype system that is based on different requirements identified in their research. However, the system proposed by the authors does not incorporate any cloud services or configurations for the components. Instead, they only use the Hadoop software frameworks as the basis for their prototype system. On the other hand, our main focus is the framework and not the actual system used in our implementation. Another work [8] that is similar in scope utilizes the AWS for their biometric and cloud implementation. They do not utilize some of the important services found in our paper such as the Amazon EMR and Hadoop software. In their paper, the authors mention using the MapReduce framework in order to promote more efficient parallelism. We agree with this argument presented by the authors since the Amazon EMR is essential to the data processing capabilities that our framework can offer.

There are several other works that contain similar topics but proceed in different directions. An example of such work is Gonzales et al. [9] where the authors focus on securing biometric data from open environments using encryption and access control. Similarly, Banrostam et al. [10] propose a unique method for increasing the security of behavioral biometrics in cloud computing. The main difference between the two works is that Banrostam et al. uses a method which gains behavioral data over time while Gonzales et al. incorporates advanced cryptography to minimize the exposure of the private key. Another research work that involves encryption is presented in [11] where the entire biometric database is encrypted and outsourced to the cloud.

After reviewing the related literature, we believe that the cloud is a promising solution to the large-scale issues of biometrics. All of the related works agree that the cloud presents an efficient environment for the growth of biometric applications. However, none of these presents a very detailed framework for optimizing the biometric performance and enabling the large-scale adoption.

III. CHALLENGES AND REQUIREMENTS OF BIOMETRICS

In the current computing world, biometric identification systems are limited by scalability and flexibility constraints as well as several other challenges and performance requirements that cannot be achieved without access to nearly unlimited storage and processing power. To understand the importance of our cloud framework, the biometric issues that motivated our design must be understood as well. Thus, we identify both the known and expected inhibitors and requirements of a large-scale biometric system. The factors that prevent large-scale biometric applications are addressed individually by the components in our framework as shown in Table 1.

A. Scalability

Conceptually, biometric systems should be able to adapt and accommodate to their own growth while providing higher or lower levels of productivity depending on the scale. On the contrary, the increasing population and demand causes scalability issues that need to be resolved in order to promote large-scale biometric applications. For instance, in previous implementations, the biometric databases have to be entirely redesigned whenever demand increases [12]. As a result, the

<table>
<thead>
<tr>
<th>Critical Factors</th>
<th>Framework Components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IaaS</td>
</tr>
<tr>
<td>Scalability</td>
<td>X</td>
</tr>
<tr>
<td>Flexibility</td>
<td>X</td>
</tr>
<tr>
<td>Interoperability</td>
<td>X</td>
</tr>
<tr>
<td>Data Archiving</td>
<td>X</td>
</tr>
<tr>
<td>Data Processing</td>
<td>X</td>
</tr>
<tr>
<td>Near Real-time Speed</td>
<td>X</td>
</tr>
<tr>
<td>Fault-tolerance</td>
<td>X</td>
</tr>
</tbody>
</table>

TABLE 1. SIGNIFICANCE OF EACH FRAMEWORK COMPONENT
modifications are not cost efficient which outweighs the benefits of scaling the application. Furthermore, the procedures and technology for matching and retrieving templates will change over time. The most efficient and effective biometric system must be able to incorporate these design modifications without inducing a sufficient cost upon the consumer.

B. Flexibility

Flexibility is another important factor contributing to the scarcity of a highly efficient biometric system. The ability of a system to change quickly based on the growing requirements is directly related to the scalability. As we mentioned in the previous section, the biometric system must be able to adapt new methods and technology (i.e., next generation scanners) in a cost efficient manner. Additionally, the system should support multiple modalities and multimodal databases. The multimodal databases may require fused algorithms to compile the most accurate results which should be accounted for as well.

C. Interoperability

The existing biometric systems are developed by a variety of vendors adhering to different national standards. As a result, there is an overabundance of data with different formats and structure which causes an interoperability challenge. Specifically, the proprietary algorithms and variety of hardware utilized present challenges when attempting to share information between different vendors [13].

D. Data Archiving

The large volume of data generated by template enrollment presents a challenge when attempting to store the data. The quantity of data depends on the type of modalities used and whether those modalities are used in tandem for a multimodal system. We anticipate that multimodal databases will be commonly employed by the public and private sectors which will result in petabytes of data at their disposal. This requires a database that is capable of supporting a multitude of records ranging from several to hundreds of millions [1].

E. Data Processing

As previously discussed, the large amount of data associated with biometrics requires a vast storage capacity. In addition, an extensive volume of computation is required to process identification requests at the ideal speed for optimal performance. The processing power and memory constraints prevent individual servers or computers from processing the data while satisfying the speed and accuracy demands.

F. Near Real-time Speed

To meet the performance requirements, the biometric identifications need to be performed in near-real-time. One of the performance requirements is that data streams should be continuous and in near real-time for improved data sharing. Furthermore, to increase data consistency, there should be support for near real-time updates among concurrent users [14]. The relative operating characteristic (ROC) curve can be calibrated to accommodate situations that dictate different identification requirements. This can only be achieved by altering the curve in a near real-time speed in order to effectively balance the security and convenience of the system.

G. Fault-tolerance

The possibility of hardware failure for a large-scale application is high and may result in system wide failure. In regards to biometrics, the system must be fault-tolerant to prevent single points of failure and ensure availability for the user. If failure occurs, the system should utilize redundant resources to minimize the interruption of the operation and maintain system reliability [14].

IV. PROPOSED FRAMEWORK

There are multiple cloud services and deployment possibilities that should be considered when implementing biometrics with the cloud. These variables affect the configuration options that are available to the user and ultimately affect the overall performance of the biometric system. In our proposed framework, we describe the cloud services and additional components that should be utilized and the configurations necessary to promote the highest biometric performance achievable. Table 1 shows the significance of each component in regards to meeting the biometric performance requirements.

Fig. 1 shows the relationship among the different components in our framework. The Infrastructure as a Service (IaaS) is used as the hosting environment for the other components as indicated by the arrows. However, the distributed file system extends further than the IaaS level since the distributed file system is used externally for data archiving and backup.

A. IaaS

IaaS was chosen as the underlying cloud service since the designer receives more administrative control than Platform as a Service (PaaS) or Software as a Service (SaaS). This allows for a higher degree of customization, which is important when leveraging the cloud to augment existing biometric systems. In addition, the entire biometrics infrastructure can be migrated into the cloud. Thus, a database modification can be performed easily in minutes. This demonstrates the elasticity of the cloud.

Figure 1: Relationship of framework components
since a biometric application can be scaled quickly based on demand. When accounting for the fixed cost of the cloud service, it is highly cost efficient compared to traditional deployment models.

When examining the virtualization aspect of the cloud, the infrastructure determines the type of hypervisor utilized which affects the performance and capabilities of the system. There are two types of hypervisors known as Type 1 and Type 2, which represent bare-metal and hosted architectures, respectively. Fig. 2 shows the configuration options available to the developer and the design differences between the Type 1 and the Type 2 hypervisors.

In our proposed framework, we chose bare-metal architecture when configuring the virtualization layer since it offers several advantages over hosted architecture. Bare-metal architecture usually provides greater performance due to the direct access of the I/O devices from the virtual machines. Hosted architecture results in less performance since the identification requests would have to be directed through an additional layer which would be the hosted operating system. Furthermore, bare-metal architecture supports real-time running on operating systems located inside the virtual machines, unlike hosted architecture [15]. The only complication, which would make hosted architecture more desirable, is the overhead involved in the installation of Type 1 hypervisor.

After the type of hypervisor is chosen, the underlying virtualization technique must be taken into consideration. Fig. 2 illustrates the three main virtualization techniques and highlights the design choices of our framework. We selected paravirtualization since it is the most advantageous of the virtualization techniques in regards to optimal biometric performance. The other possibilities are binary translation or hardware assisted virtualization. The performance increase granted from each technique varies according to the coupling of the hypervisors. Paravirtualization enables the greatest performance due to the tightly coupled hypervisors while the loosely coupled hypervisors of binary translation cause significantly less performance. In addition, binary translation and hardware-assisted virtualization can actually result in performance degradation since the hypervisor must routinely interrupt the execution of a virtual machine [15]. Another advantage of paravirtualization is the lack of overhead when compared to the other two techniques.

B. Distributed File System

A distributed file system for archiving data is essential to meet the capacity demands of a large-scale biometric database. In our framework, the file system would be hosted on the IaaS, which offers nearly unlimited storage capacity and scalability. In addition, the file system must demonstrate functionality since it would be used externally for the immediate retrieval and backup of data. Furthermore, the system must be fault-tolerant and robust which requires exceptional recovery and failover capabilities.

Ideally, we believe the system should incorporate automatic failover. The failure of a node would no longer hinder the operations of the system. Instead, the task of the failed node would be sent to a redundant node with no interruption in the workflow. To the best of our knowledge, there are currently a very limited number of cloud services that include automatic failover in their recovery plans.

C. Distributed Data Processing

Distributed processing involves multiple computers or processors being utilized to execute applications. Parallelism represents the solution to the data processing challenges of biometrics. In parallel computing, data is partitioned across multiple nodes where computations are solved concurrently. When applied to biometrics, the calculations performed by the algorithm responsible for template matching would be broken down and distributed to several nodes. There are many big data processing tools and frameworks that can be utilized to increase the data processing capabilities of a biometric system. Some of these include MapReduce, Microsoft Dryad, Message Passing Interface (MPI), and Swift [15]. In order to choose the most effective framework or tool for biometrics, the size of the data set and the number of tasks must be taken into consideration. Therefore, we have chosen MapReduce since

![Figure 2: Example of virtualization configurations and framework design](image-url)
the biometric systems require large input data sets and small number of tasks.

MapReduce [15], [16], [17] is a data processing paradigm that enables parallel computations for analyzing and generating large data sets. The input data of a user would be sent into the cloud, partitioned, calculated in parallel, compiled, and sent back to the users in the form of a decision. This increases identification speed while promoting the most accurate results. MapReduce provides organization to the database by moving computations to the location of the data that complies with the data locality principle. As a result, the processing and query latency will decrease.

D. Distributed Database

Distributed databases [18] consist of portions of the database being stored in multiple physical locations. These locations are loosely coupled and do not share any physical components. Moreover, a distributed database is dependent on a distributed database management system. Therefore, we included database management as one of the components in our framework. The distributed database is preferable to a relational database for many reasons. The reliability and availability of the biometric system will increase dramatically.

Furthermore, the biometric system expansion and modification is easier and does not affect other related systems. Distribution increases the query processing which improves the biometric performance. In addition, the system becomes more fault-tolerant since offline nodes do not interrupt operations. The distributed database should provide near real-time access to the data located in the biometric database. The record lookups should be in near real-time as well.

E. Distributed Database Management System

A distributed database management system controls the distributed database and routinely integrates data in the database to ensure that any change made by the user is updated. Moreover, the distributed database management system supports a distributed metadata repository for near real-time availability. We believe that metadata should be exploited to allow for easier interoperability among existing systems. Metadata can mitigate the problems caused by different template formats by describing the records and allowing for interchange.

V. REALIZATION OF THE FRAMEWORK

In order to demonstrate the effectiveness of our framework, we designed a system based on the previously discussed specifications. We selected several components to be utilized for biometric integration. The AWS are the primary cloud components with several applications being included that are supported by Amazon. The additional applications originate from Apache Hadoop, which is an open source software framework that is synonymous with the big data.

Fig. 3 illustrates the general workflow of the realized components listed in this Section. The input and output data is sent from Amazon Simple Storage (S3) to Amazon Elastic MapReduce (EMR) where parallel computation is performed. Amazon EMR will have several clusters, based on Hadoop architecture, depending on the modality of the biometric system. The two clusters shown in Fig. 3 indicate that the Apache Hive and the HBase should be run separately to improve performance. The HBase cluster is connected to the Hive cluster by the master nodes. The cluster consists of master and slave nodes. In particular, the slave nodes consist of task and core nodes. The Hadoop Distributed File System (HDFS) is used to store and process data located in the core nodes but does not interact with the task nodes. Finally, the nodes are a product of Amazon Elastic Compute Cloud (EC2).

A. Amazon EC2

Amazon EC2 [19] is an IaaS that presents a virtual computing environment to the user. This service grants the ability to create virtual machines called “instances” on Amazon EC2 and exploit them for improved data processing. The number of virtual machines can be increased or decreased according to the scaling of the system. Moreover, Amazon EC2 utilizes the Xen virtualization in its underlying architecture, which has the benefit of allowing hardware assisted or paravirtualization. As stated in the previous section, we believe paravirtualization is the most effective technique for realizing the full potential of biometrics. It is important to note that the Amazon EC2 costs are included when purchasing the Amazon EMR service. This paper includes the Amazon EC2 in the components since it generates the instances used by Amazon EMR.

B. Amazon S3

Amazon S3 [20] is a file storage web service that can be used for archiving and analyzing data when implemented with Amazon EC2. We chose Amazon S3 since it is reliable, durable, and supports the other components in our system. Amazon S3 stores and organizes data into “buckets” which are linked to each user account. In addition, Amazon S3 allows users to backup data from the HDFS, Amazon EMR, and the HBase which demonstrates effective recovery of data. The scalability or performance bottlenecks associated with traditional storage databases can be resolved by adding more nodes to increase the system speed, capacity, and throughput.

C. Apache HDFS

Apache HDFS [21] is an open source distributed file system that was chosen to be used in collaboration with Amazon S3. The HDFS is highly portable and is capable of storing large files while replicating data across several hosts to achieve data reliability. One notable feature is the data awareness demonstrated by the interaction of the job tracker and task tracker. The job tracker schedules map or reduce jobs to the task tracker depending on the data location. This reduces the amount of network traffic. The HDFS was primarily selected since it is more fault-tolerant than Amazon S3. In addition, the two file systems overcome many of their own design flaws when implemented together. Thus, it is beneficial but not necessary to include both file systems in our system design to achieve maximum biometric efficiency.

There are several advantages for implementing Amazon S3 and HDFS. The HDFS nodes for storing the data are no longer
required since Amazon S3 provides a massive amount of storage capacity. Additionally, the HDFS data located in an Amazon EMR cluster is lost upon the termination of the cluster. On the other hand, Amazon S3 permanently stores the data ensuring that it is never lost. Another benefit is that HDFS nodes will not become overloaded when using the same data set for multiple jobs [22].

D. Amazon EMR

Amazon EMR [22] introduces the MapReduce function to the virtual machines created with Amazon EC2. It is fully integrated with Amazon S3 and HDFS. This component is vital to the performance of biometric systems since the parallel processing speed is dependent on the MapReduce functions. The Amazon EMR cluster consists of master nodes and slaves. There is only one master node in a cluster and the master node is responsible for coordinating the slave nodes. The slave nodes are responsible for running the computations and storing the data.

E. Apache HBase

Apache HBase [23] is an open source distributed database that uses column compression to store large quantities of data. The HBase needs a file system to store its data which is represented in our prototype system by the HDFS. The purpose of the HBase is to augment the HDFS by adding more functionality and addressing several issues. One of these issues is the individual record lookup speed, which is significantly slow. The HBase allows for fast record lookups and access to a small amount of data without the additional latency. Another issue is the batch inserts and deletions. The HBase is an alternative approach that allows users to efficiently perform batch inserts or deletes as well as updates.

In addition, there are several more reasons why we chose the HBase as the distributed database for our system. The HBase utilizes the Bloom filters for improved query when there is an extensive volume of data. There is a tight integration between the HDFS and the HBase, which allows for easier implementation. Furthermore, the regions that contain the HBase tables are split automatically and distributed as the quantity of data increases. Lastly, it is highly fault-tolerant and supports automatic failover for the RegionServers. The RegionServers are responsible for managing the regions in a distributed cluster.
F. Apache Hive

Apache Hive [24] is an open source data warehouse for managing a distributed database while promoting data query and analysis. The Hive operates using a modified version of the SQL called HiveQL, which enables support for map and reduce functions and different data types. Additionally, unstructured data sources are supported by the Hive and are capable of being processed efficiently. Hive is capable of fabricating a metastore, which is located in MySQL database on the master node of the cluster. The metastore contains the partition names and data types.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented a conceptual framework for facilitating the biometric applications at large-scale. We identified the challenges and requirements of current biometric systems to provide the basis for our framework. We then designed the framework to address these issues while increasing the overall biometric performance. At last, we demonstrated the effectiveness of the framework by selecting components for a system that adheres to the criteria we specified. The abstraction of the learning curve will increase the productivity of developers when implementing the biometric with the cloud computing. We anticipate that researchers will be highly motivated to integrate biometric systems with the cloud, especially with a framework available for the enterprise environment.

Since the framework and realization is conceptual, there are considerations and issues that we will address in the future version. The fused result of a multimodal biometric system could present a challenge for the MapReduce component. If an iteration procedure is proven to be more effective for biometric identification, the framework will probably need to be altered slightly to promote maximum efficiency. To address these issues, we will conduct an experiment using our framework for biometric and cloud integration and analyze the results.

REFERENCES