CALIFORNIA STATE UNIVERSITY, NORTHRIDGE

A Secure and Serverless Approach to Verification of Student Records

A graduate project submitted in partial fulfillment of the requirements

For the degree of Master of Science in Computer Science

By

Karandeep Singh Bhamra

August 2019

The thesis of Karandeep Singh Bhamra is approved:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Li Liu, Ph.D. Date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Robert McIlhenny, Ph.D. Date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Taehyung (George) Wang, Ph.D., Chair Date

California State University, Northridge

Acknowledgements

I am grateful to Dr. George Wang, for being my committee chair and giving me a chance to work on this project. In addition, I would like to express my profound gratitude to Dr. Li Liu, and Dr. Robert McIlhenny for agreeing to be the additional members of my committee.

A special thanks goes out to my wife, Jessica, and my parents. The support my family provided me with went a long way in helping me finish this project.

Also, I would like to thank my friend Jesus Moran Perez, he was the one who initially inspired me with the idea for the project and has been of great assistance in helping me complete this project.

Table of Contents

Signature Page ii

Acknowledgements iii

List of Figures v

Abstract vi

Chapter 1: Introduction 1

Chapter 2: Related Work 5

Chapter 3: Serverless Architecture 6

3.1 Serverless History 6

3.2 Serverless Platform 11

References 13

List of Figures

Figure 1 – Virtual Machine 8

Figure 2 – Containerized Applications 9

Figure 3 – Event-based Functions 10

Figure 4 – Function Integration 11

Figure 5 – Serverless Platform 12

Abstract

A Secure and Serverless Approach to Verification of Student Records

By

Karandeep Singh Bhamra

Master of Science in Computer Science

Most academic institutions do not self-operate the system that is used for the verification of student records. Such a task is contracted out to firms who act as the middleman between the student and the potential employer who might be seeking verification of records. This project proposes a complete overhaul of that system and introduces a fast, and secure alternative system which accomplishes just that while maintaining low operating costs by making use of serverless architecture. The project consists of a scalable frontend website which is used to request a student’s record, the request is then processed by a web framework which routes it to an API endpoint where a serverless function retrieves the student’s record from a scalable database and generates a verified pdf certificate which is sent back to the frontend website client to be displayed. The project also implements tools for administering the blockchain like data structure that holds the student’s records and for verifying the integrity of the student records. The completed program demonstrates the entire workflow from the viewpoint of the client and the school administrator, while showcasing the benefits of implementing the system in a serverless architecture and how the student records are safe guarded against malicious modification.

**Chapter 1: Introduction**

Verifying a degree or any student record takes some amount of time, capital and patience. In addition, the verification process requires potential employers and background check firms to request confirmation of credentials from third party services associated with the academic institution rather than the institution directly. The results are not immediate and the privacy is implied, but not explicit. The data transaction should exist solely between the student, the institution and those that you want to share it with. A third party cannot be completely trusted to handle your private data. On the opposite spectrum, the student cannot be completely, trusted to provide accurate results, as the student providing it can manipulate it before delivering the modified records to those who requested it.

Since it has been established that verification of academic records takes time and money, most academic institutions have contracts with firms that provide the academic record verifications. There is unnecessary overhead related to that process; an employer or background check firm usually has to make a request with those middleman firms, who will then forward the request for records to the institution itself, or fetch the information from their own database which was provided to them [1]. I propose a solution in which the middleman can be cut out, and the verification process is provided directly by the academic institution. Overall, the main purpose of this project will be to create a proof of concept data storage format that borrows from and mirrors the blockchain data structure. In said data structure, the blockchain will hold a growing list of academic records of students, and will allow for the verification of the authenticity of those academic records. Once the format has been implemented, it will then be utilized to show how an academic institution can keep accurate and verifiable records relating to the students, and how a third party might be able to request and receive those records on an infinitely scalable serverless architecture.

Serverless architecture is a relatively modern concept. It is an application design where the business does not have to worry about managing the infrastructure that would support their application. In serverless architecture, there are still servers somewhere in the backend, however, the customer of those servers, is not responsible for managing them in any way. A person, business or any client would not have to worry about having to update the server, software, libraries, or be required to hire employees that would do the managing; instead, cloud providers such as Microsoft, Google, or Amazon, provide and manage the servers on your behalf, thus providing a seamless experience to the end client [2]. Serverless architecture relies heavily on event-driven programming, and executes code without provisioning servers through the use of functions. Functions are the pieces of code that make up the business logic, and they can be run within few milliseconds and the entire functions platform is provided as a service by most cloud vendors [3].

The term blockchain has taken off in popularity due to its connection to cryptocurrency. However, the data structure known as blockchain has many uses besides just cryptocurrency. A blockchain is a construct that acts as a digital database where the transactions are recorded in a secure and public manner. The concept can be simplified to this: anything can be owned, and there are things that we own and would like to share those items, and blockchains act as the middlemen which facilitate said exchange. A block refers to a specific record, or data that can be permanently recorded. A blockchain is a collection of those blocks appended one after another, and the blockchain acts as a historical record of all the total blocks added to date [4].

In a blockchain, the initial block is also known as the genesis block and is always hardcoded into the application. Once a block has been added to the blockchain, it is difficult to change that record [5]. The records are also verified by the network to make sure that they are valid before they are added to the blockchain. Each block also contains a unique code called a hash, and in addition, the current block also contains the hash of the previous block in the blockchain. A hash code is an output of a mathematical function, which is generally a string, that is composed up of series of characters and numbers. So, for a given unique input, the hash function will always yield the same hash; therefore, for someone to modify the blockchain, they will have to compute hash codes for each block and that is an extremely computationally intensive task depending on the underlying hash implementation. The benefit of a data structure like blockchain is that it is a digital log of all transactions that have occurred and due to its nature, it can be quickly searched and verified.

This project is organized in four distinct parts: serverless backend, the serverless frontend, serverless database, and an administrative student record software. The backend portion relates to the serverless computing provided via functions, which are ultimately responsible for providing the frontend client with the document that was verified and generated on the fly and transmitted over a secure communication channel. The frontend will be a website where the client enters the student information such as name and then the request is routed via a web framework to the corresponding serverless function and the received document is then displayed if the student record exists. The database is necessary for the storage of student records, the focus of the database will be on making it scalable, affordable and integrating the blockchain like data structure with the serverless functions, which when used in cryptocurrency, it is a decentralized system due to peer to peer network; however, for my project it will be a centralized system controlled by the institution due to an easier integration with the statefulness of the project. Lastly, since my project is a complete overhaul of how the records are handled now, it was necessary to showcase how an administrator in charge of this system would utilize it. The tool would allow a person in charge to load student records in a predefined file format, create a new student record, save the student record in the database, and verify the integrity of the records in the blockchain.

For a complete understanding of this project, it will be necessary to discuss other advancements in the area of using blockchains for academic institutions and related works, the underlying technology and algorithms used in serverless applications and blockchains. In addition, we will look at and compare the benefits of serverless versus other traditional architectures, a through comparison of the various cloud services providers, the implementation of the project and its subcomponents, the workflow of the whole system, the testability of the project and any further improvements that could have streamlined the whole operation if not for the for time or resource constraints.

**Chapter 2: Related Work**

As of the writing of this thesis, there are multiple corporations and startups that are starting to look towards blockchain to solve a host of problems. The storage and verification of data is a huge section of the outstanding problems left to be solved. Most recently, the popular hardware and software company IBM has partnered with Sony to make use of blockchain technology. IBM has created their own blockchain called IBM Blockchain, which will be the basis for a new platform geared towards academic institutions [6]. Their goal is to allow students and schools to track each other’s progress. Additionally, such a system will make use of the built-in transparency and accountability due to it being based off of blockchain technology. According to IBM Japan, there are many preparatory schools based around Japan that have already incorporated IBM Blockchain along with student credentials to keep track of coursework and verify transcripts [7]. A fully managed cloud service such as IBM Blockchain allows for institutions to rapidly adopt the technology without having to pay for new research and development for a private blockchain network.

Learning Machine is another small, private company which is hoping to make a big splash in the blockchain scene [8]. Their goal is to provide a complete system to public and private bodies which will allow those parties to issue official, instantly verifiable records anywhere in the world. From the examples shown on their website, they issue verification for any type of company-wide or institution wide record. They have examples of an individual’s doctorate degree being verified and a certificate is issued with their information and the date their degree was conferred upon them. They intend to pioneer the way to allow businesses or schools to easily issue instantly verifiable records.

**Chapter 3: Serverless Architecture**

The tech industry is plagued with buzz words and new ideas that seem to be offshoots of old ideas. Most recently, “serverless computing” (also known as serverless architecture or function as a service, FaaS) is an application design pattern where cloud providers manage the infrastructure and your applications simply build and run on that infrastructure. At a glance, this is a great way to create and deploy applications. It allows for the client of the serverless architecture to focus on their business logic, without having to worry about the hardware, scaling and maintenance aspect of servers. The term serverless is however misleading. There are still servers, because otherwise it would hard for code to run on nothing, but, the person or business that is creating or deploying the application does not have to provision any servers personally [9].

**3.1 Serverless History**

Computers were not as powerful or portable as they are now. With the emergence of personal computers, much of computers moved away from the large, centralized mainframe systems to a smaller client-server model. Early on, when a client wanted to deploy an application, they would have to provision their own physical, bare metal server, install the required software tools, language frameworks, and any other dependencies of the application. In the beginning, all servers were bare-metal servers, which meant that each individual, physical server used by a single client [10]. They would require a dedicated person of expertise whose sole job was to administer the server, and would be known as Server Administrators. The need for such a person was an additional cost for the client of the server, in addition to having to pay for any resources related to the cost of operation such as electricity, and other on-premise infrastructure costs. Eventually, with the development of virtualization, it became easier to rent out the server to multiple tenants. Now, a single machine could host multiple websites. Software and other web applications could be run on these virtual machines with completely different operating systems. It now became possible to run an application that depended on the Linux operating system and at the same time run another application that depended on the Windows operating system. Even with virtualization, there were still costs associated with it. The monetary costs dealt with any software or operating system licensing fees. Overall, virtualization led to a decrease in general costs due to the idea of multiple tenancy [11]. A virtual machine could be created with specified number of virtual processors, dedicated memory and storage space; so, an application that required fewer resources would not be charged the costs associated with the higher resources. It became cheaper to run smaller applications rather than larger, resource hungry monolithic applications. One negative aspect of virtual computing would be that since the resources were divided among the various virtual machines, a larger, resource hungry application could easily starve out other applications. There are also issues with scalability, if your software becomes too large too quickly, the set, finite resources might struggle to keep up with the popularity. Bare metal servers and virtual machines still have their place in computing to this day and age, however, they did lead to newer, and improved ideas.

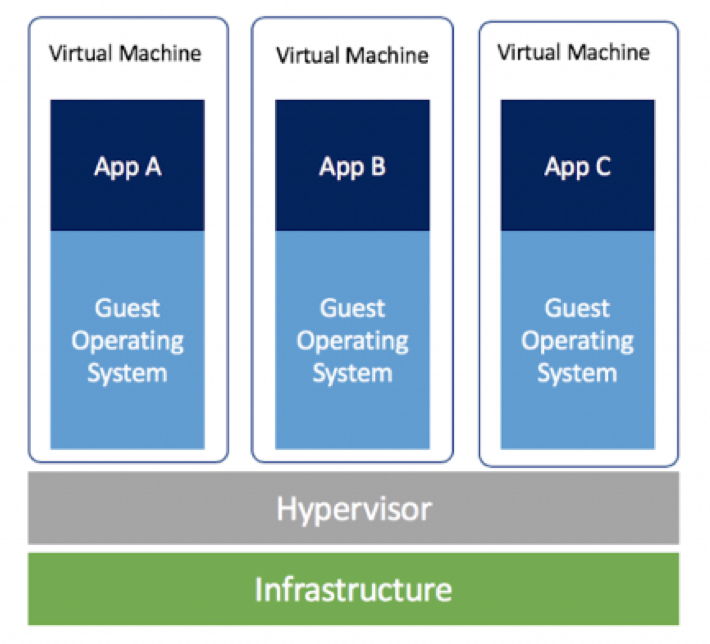


Figure 1 Virtual Machine

After virtualization, the next big idea was containerization. Application containerization is still a type of virtualization, but a container consists of an entire runtime environment. In essence, you are abstracting away the differences of various operating systems and hardware infrastructure, by bundling the application with all of its dependencies and files needed to run it into a single package. Docker is one of the most popular software that performs the operating system level virtualization [12]. A single application can be packaged into a single docker container image which can be downloaded by anyone and installed on any operating system. It is now possible to run applications specific to a single operating system on any operating system platform that has the Docker software. Benefits of containers include consistency when shipping applications, you know the application will behave in a deterministic manner since it was tested and developed in a controlled environment. As popular as Docker is, it is not a one size fits all solution to computing. Some disadvantages of containers are that not all applications benefit from containerization, and there is a speed penalty when running a container. In virtualization or containerization, there is a performance overhead; containers are faster than virtual machines, but still not as fast as bare metal. Lastly, the data inside of a container is not persistent [11]. Due to its design, the data generated when the container is spun up is lost when it is shut down, however, there are ways to save data from a container.

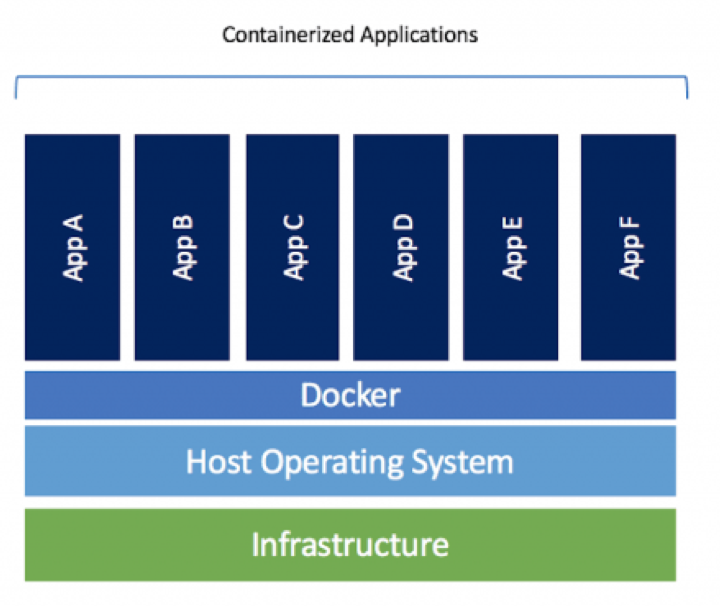


Figure 2 Containerized Applications

The latest fixation brings us to now, which is serverless computing via functions as a service. Functions are useful since they provide an on-demand functionality. Functions are a finite piece of code which are run when they are invoked. It is easy to become familiarized with functions since their underlying technology is based on containers. Functions are event-driven and are therefore invoked by any number of events, upon being invoked, the container where the code resides is spun up and the code is executed and the container is shut down [2]. Therefore, you only pay for the amount of time your code took to execute and not for any of the underlying infrastructure. All function as a service provider also allow for easy and quick scalability of resources. When a function needs to scale up to meet a million requests or scale down to handle a single request, the cloud vendor handles all the underlying allocation and provisioning of resources. Since a function is just a program that will be executed, most vendors allow for programming in plethora of programming languages and have support for many of the popular frameworks. The usefulness of a function is defined by having it connect to various other cloud services such as databases.

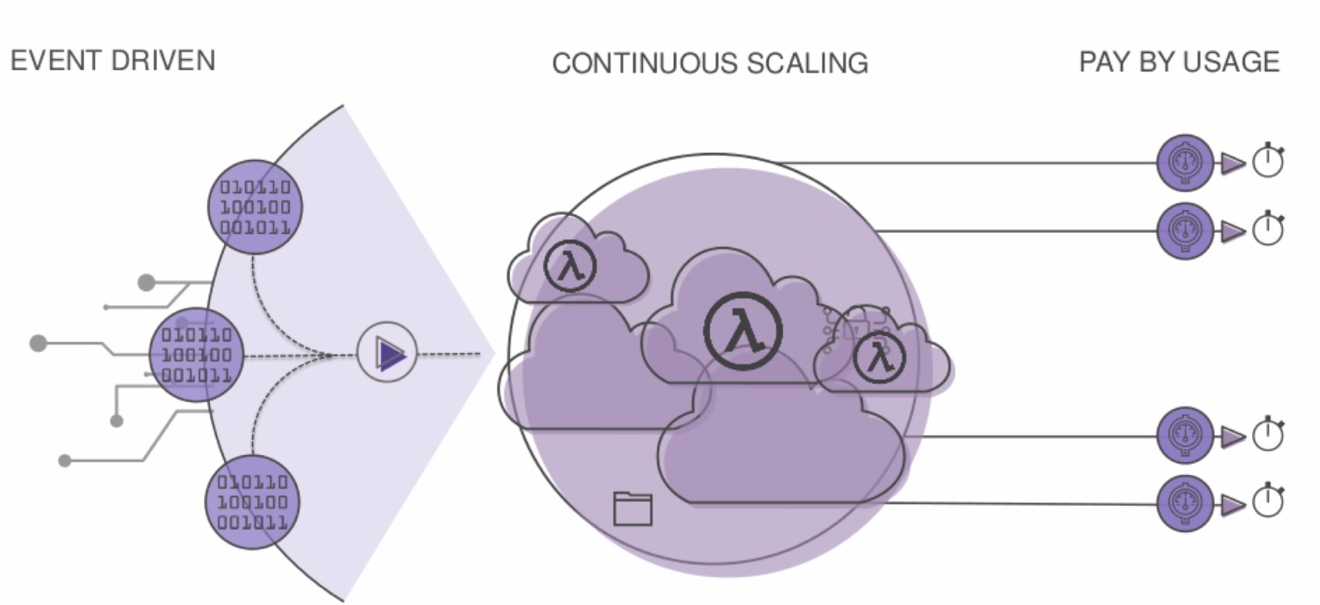


Figure 3 Event-Based Functions

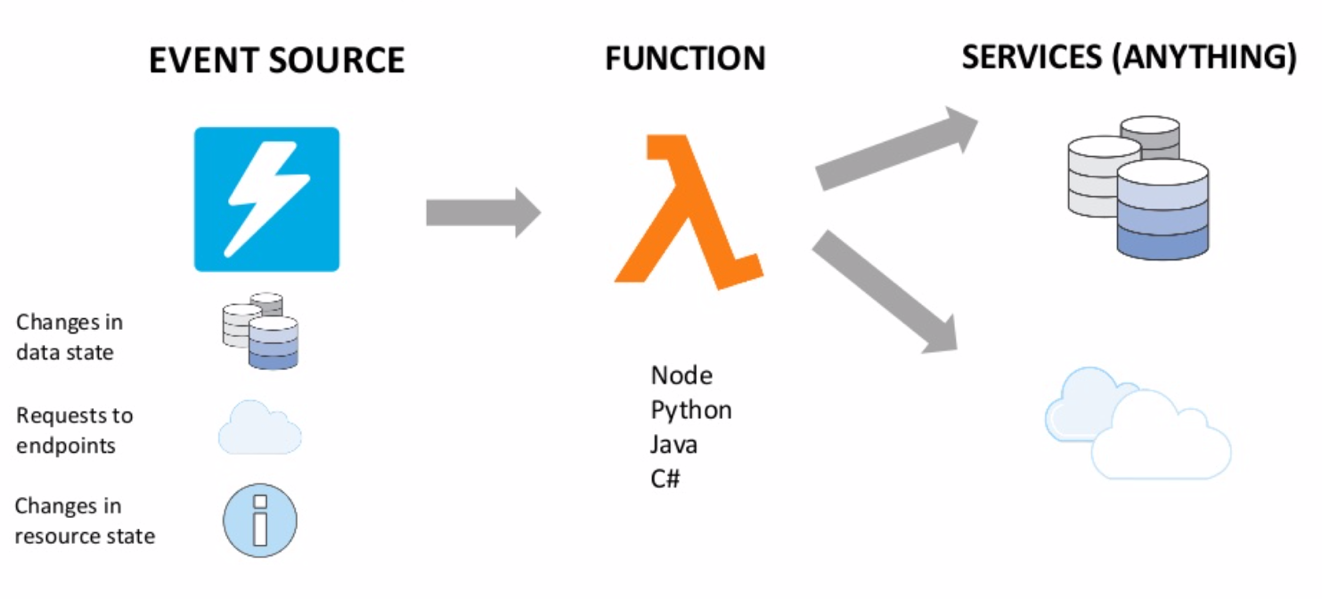
****

Figure 4 Function Integration

**3.2 Serverless Platform**

Functions are not the only cloud computing service that makes up a serverless architecture. As previously discussed, databases are also provided as a scalable service. Manually having to manage the capacity of a database can be time consuming; it is difficult to predict when there might be a spike of incoming or outgoing requests, so you have to be able to handle those requests dynamically. Many cloud services providers now have serverless databases that are simple, scalable, cost effective and highly available. Availability is one of the most important attributes associated with databases. If a database is unavailable, then the users of its data such as clients, and applications are unable to access it. Having an unavailable database is a detrimental to anyone relying on them. Due to the takeover of responsibility by the provider, developers and operators can save time and money by focusing on their business logic part of the application and less dependent on infrastructure support teams. In addition to databases, serverless platform also depends on managed services for API endpoints, message queueing service, continuous integration and continuous delivery pipelines, user management, and storage. Just about every important part of computing can be extracted to a managed service and offered to clients in an easy to use manner.

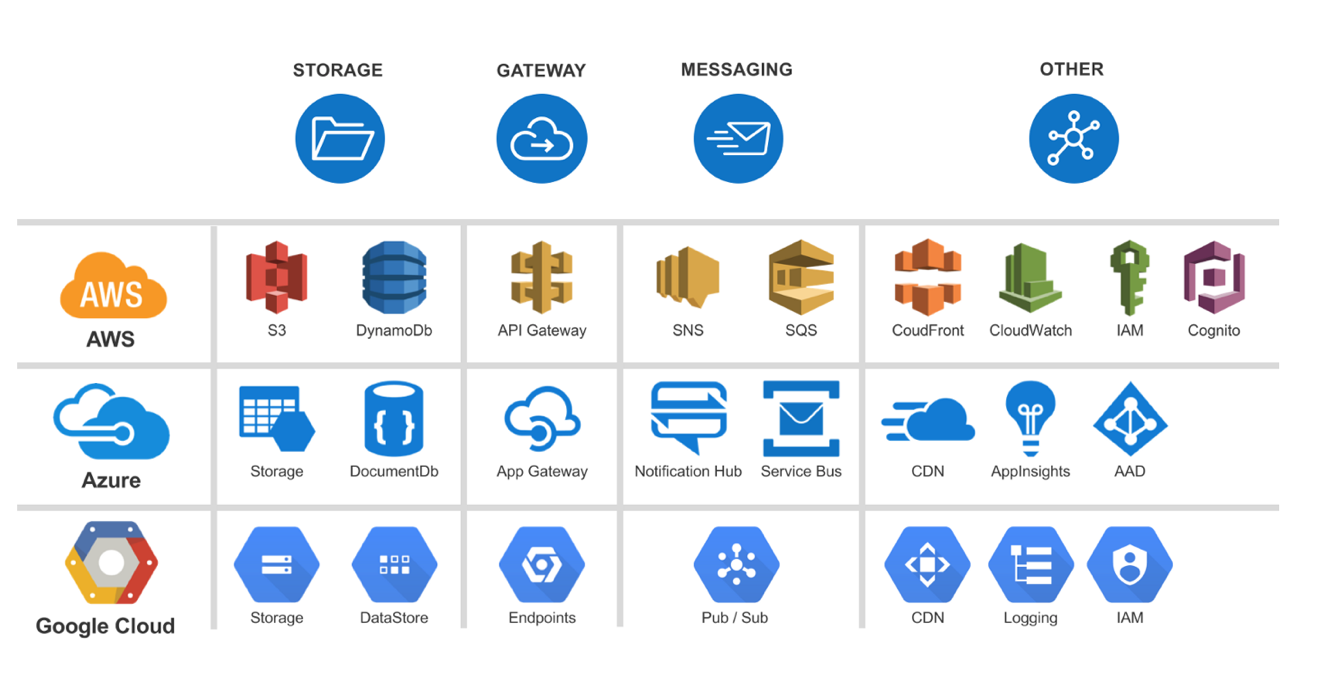


Figure 5 Serverless Platform

**Chapter 4: Underlying Project Technology**

Due to this project’s reliance on the serverless architecture and concepts, there are a few technologies that are used heavily. For starter, this project made use of serverless functions as the compute engine, because there was a need to execute code as required by the workflow of the project. The serverless functions helped facilitate the logic required to add student information in to the database, to query the database, and to generate the documents which will be returned to the end user and much more. In addition to compute functions, another technology that was essential to the project was a database. Almost every single program that needs data to persist after a power cycle has to make use of a storage medium one way or another; it is literally quite impossible for data to exist unless its stored in a medium somewhere such as memory, files or a database. This project needed a database which would act as the basis of the blockchain like structure. The requirements for the structure mandated that the database be scalable. Scalability was the primary factor, because it needs to be able to handle a dozen requests to multimillion requests, even though in reality it might never actually quite reach that limit. However, the application needs to be highly responsive and always online, so the compute functions and the scalable database were essential. In addition to the components that make up the serverless architecture, blockchains also rely heavily on concepts from cryptography such as hashing via hash functions and a related topic of collisions. The next few sections will dive deeper into the listed topics from above and we will see how they played a vital role throughout the project.

**4.1 Serverless Functions**

Serverless functions refer to pieces of relatively small code which will get executed in a completely managed environment. Serverless functions can be a valid tool in most applications. It is hard to advocate for them for every application or role, but they have their uses and the persons deploying them has to think and play to their strengths. The strengths of functions are that they are relatively easy to set them up and get them running code you might have. Most cloud vendors have their own implementation of how they interact with their cloud services and most of the cloud vendors have the same relative restrictions on the functions. As opposed to full on virtual machines or servers with language software development kits (SDK) installed and almost no limitations on how the programming language can interact with the filesystem via its application programming interface; functions are limited on which service they can access with the provided permission roles and how long they can execute for, and how much memory they are able to occupy and make use of. However, at the same time, they also benefit from significantly reduced operational costs, complexity and engineering lead time [2].

Serverless functions can be as small as few lines of code which can be edited directly in the cloud providers web interface and can be deployed directly to live sources from the online interface. It is however not advisable to perform all the programming or testing from the web interface. The difficulty comes in that such interfaces are often quite clunky and different that how a person, a team or company prefer them. For example, Microsoft Azure Functions can be edited and deployed directly from their dedicated Azure Functions webpage. In addition, one is able to write code directly via the available online code editor; if online editing is too cumbersome, you are able to write code offline on your device and then upload the project and all its files via the online upload files options. Even if that is not up to the developer’s liking, they are able to use Microsoft’s development tools such as Visual Studio or their text editor Visual Studio Code which have support via officially released plugins which allow you to create functions, write code and then upload it from the software. In essence, Microsoft provides flexibility and choices to the developers. On a personal experience, I have found their tools and support to very helpful and friendly towards developers.

Most cloud vendors such as Google, Microsoft ad Amazon also support multiple programming languages for their serverless compute functions. For instance, Microsoft currently supports JavaScript, Java, Python and their programming language C# [13]. On the other hand, Amazon supports JavaScript, C#, Java, Go, Python, Ruby, PowerShell and also provides a Runtime API which allows developers to use any additional languages to create the functions [14]. When choosing a cloud vendor, a team has to be aware of all the services and languages that are supported for that given function provider. On a personal note, it is also advisable to see if the cloud provider provides the developer with an offline SDK which can be used to test the function before deploying it to a live resource. Since the cost of the function is computed when it is run and the calculation incorporates variables such as the execution time, the amount of ram used and if it interacted with any other services, it is useful to test it on a development environment to make sure that the team is not being charged for testing while developing.

On an operational level, it is also important to discuss how a serverless function is executed. Generally, when a team has decided on developing their software with functions in mind, they are also going to be making use of other cloud services such as an API gateway or a database. There is generally a single way that a serverless function can be executed, and it is via some received event. The receiving event can be an HTTP method such as GET or POST. The Hypertext Transfer Protocol (HTTP) is used to communicate between a client and a server, it is a request-response protocol [15]. An HTTP request can be made to the web address that is exposed for the serverless function, which will in turn raise an event which is intercepted and is responsible for the function executing either based on the request payment or other determined workflow. The functions underneath is run on container technology, meaning that when the request is received, the container where the code for the function is hosted, is spun up and the code is then run and the output is logged and sent to wherever specified and the container is then shut down. This is the biggest drawing point of the serverless architecture, you only pay for the code for as long as it ran and the resources it used. The container does not exist for much longer than the time taken to execute the code. The concept of just having your code run and only worrying about the business logic, and having the rest be fully managed is a refreshing idea.

**4.2 Serverless Databases**

In a serverless architecture, it is important that as many possible components or services be serverless as well. The main goal is to be highly scalable and highly available. It does not matter if the application is ground breaking or the next greatest thing, if no one is able to access the application due to it being overloaded, it will never be popular. The application should be accessible from a dozen clients to a million clients with no perceivable degradation. A database is generally the most important part of the workflow of any given software. The database is responsible for storing information that is going to be used in the business logic, so it needs to be available at any given time. Generally speaking, a database is usually installed on a server and the database users are created and the database management system is the middle layer which interacts between the data and clients.

Databases themselves come in a variety of flavors and support different model system. The two most popular database models are relational and document based. The relational database model supports collections of items with set, pre-defined relationships between them. Relational databases also support a single unified query language that is used to interact with the database and the data, it is called the Structured Query Language (SQL). SQL was adopted in the 1980s as the industry standard [16]. For relational databases, such a standardization led to its huge growth in popularity, one only had to learn a general form of the SQL standard and suddenly they had the knowledge to work with almost all the SQL databases. Not every single database has the same SQL language implemented, there tend to small differences from one to another. That’s why the standard was created and even then, not every vendor supports the whole standard. SQL works well with databases with relations between the objects. In a database, there is a table which consists of rows, each row can have multiple columns for various data required. Tables are the main components of a database, since that is where the data is stored. A row can have one or more columns of various datatypes. For example, there can be a table called Users where each key can be a unique user and the column for that row can have information such as first name, last name, and age. Every column must have a datatype, and it is up to the developer to find the best datatype that fits the information that will be going into the column [16]. SQL databases are by far the most popular these days and they have remained so due to nature of the data that the application is manipulating.

The other popular type of databases is referred to as NoSQL databases. NoSQL which stands for “no SQL” or “non-relational” is an alternative which provides ways around some of the pitfalls of the relational databases. They store and manage data in such a way that allows for developers to optimize their applications for high speed and extreme flexibility. NoSQL databases were pioneered by some of the giants in the tech field such as Google, Amazon and Facebook. The biggest benefit of NoSQL databases is that it can be scaled horizontally across plethora of servers, which a SQL database cannot accomplish in an easy manner [17]. There are plenty of disadvantages to NoSQL as well; firstly, if the data is highly relational, it becomes harder to represent in a non-relational database schema. Additionally, NoSQL does not provide the same level of data consistency as a traditional SQL database [17]. This is primarily because where the traditional SQL databases focus on the reliability of data transactions, the NoSQL sacrifices that aspect to focus on speed and scalability. One additional benefit of NoSQL is also that any given data can be stored in any other record. There are four popular types of NoSQL databases, and they are document databases, key-value stores, wide column stores and graph databases. We will be focusing on the first one, document database.

This project needed a scalable database and the data for each student was already a bit based on a “page” of the student’s record. So, it made sense to choose a document-based NoSQL database such as Microsoft’s Azure Cosmos DB. Azure Cosmos DB is Microsofts answer to a globally distributed, scalable database as a service. It allows users to scale the throughput required and storage across their highly available Azure regions worldwide. Additionally, when designing the project, I did not have to worry about a database schema. Designing a database schema can be time consuming in itself, so, when you can just store data as a JSON formatted page, it allows one to focus on other aspects of the project. By far the biggest benefit of using Azure Cosmos DB was that it had support for a range of APIs for querying the database [18]. Since I was already familiar with the SQL style of querying databases, it was easy to use the SQL API to query my document base NoSQL database. For someone that might not have the familiarity with a NoSQL type database, Cosmos DB is an excellent entry point due to its familiarity for those with SQL experience.

**4.3 Hash Functions**

Cryptography is used heavily throughout blockchains due to their nature of being cryptographically secure. Even though my project borrowed heavily from blockchain without actually being a blockchain per se, I still needed to incorporate hashing and hash functions into the project. Hashing is the process of making use of a hash function to generate a constant output for a given unique input. So, a hash function takes in an input, generally a string of any length, and it maps the input to an output which is a value of a specific length [19]. In the context of blockchains in cryptocurrencies, the transactions on the blockchains are taken as an input and run through a hashing algorithm which returns an output. So why is hashing required and important? For example, if we look at a popular hashing algorithm such as SHA-512, we know that our output will always have fixed length of 512 bits. This allows the input of any given length to always result in a hash with its length being 512 bits. There are a few properties of hash functions that make it ideal for this project. They are: deterministic in nature, able to be quickly computed, pre-image resistance, small change in input changes the output, being a one way computation and lastly being collision resistant [20]. The first property of being deterministic refers to the concept that regardless of the number of times a certain input is put through a hash function and hashed, it yields the same output, every time. The ability to be deterministic is required because a student’s records will be hashed and compared to check the authenticity of the records; if the hash of the student’s record was to change every time it was hashed, the resulting output would have little to no meaning. Secondly, the hash must be relatively be quick in its computation, if it takes too long, it might not fit the execution speed/time requirements of an application. However, for this project, I relied on two different hashing algorithms, the first was SHA-512 which was used for its ability to be quickly computed and the second being Bcrypt which can be modified to make it slower which can be helpful against brute-force attacks [21].

The concept of pre-image resistance states that given a function and its input, it should be unfeasible and impractical to determine the input. It is difficult to use concrete information such as impossible instead of words like impractical which allow for a bit of leeway because there have been hashing functions which have been broken and are no longer effective. If the input to the hash function is modified even by a bit, it is no longer uniquely equivalent to its old version and thus should yield a new output. It’s a notion that just naturally lines up with the requirements of the hash function. A string is modified so that it is a new string, and the new string gets its own unique hash output. A hashing algorithm is a one-way mathematical function, you take an input and map it to an output; for the algorithm to be useful in cryptographical situations, it must be a concrete rule that it cannot be reversed. It the output was able to be reversed, it would defeat the purpose of it being used as a way to quickly store private information. Lastly, the property that is of utmost importance states that it should be infeasible for two different inputs to map to the same hash value. So, the hash of input A cannot ever be equal to hash of input B, each must have its own unique hash. If there is ever a case where they are equal, the case is called a hash collision. The existence of a case of a collision can lead attackers to exploit the weakness and perform malicious actions. The collision can be exploited by any application which compares two hashes [22]. For example, if a website checks the login of a user by comparing the hash of the provided password, it could be made possible that another carefully constructed password which yields the same hash be substituted in and passed to the server as an authentic login. The server would have no choice but to accept the provided information and server the requested information. Even in the most unsecure hashing algorithms, they were broken because cases were found where collisions occurred, and those odds of those occurrences were extremely low, but still non zero. For a hashing algorithm to be taken as cryptographically secure it must not have any occurrences of collisions. As of right now, there are cryptographic hashing algorithms which are considered secure such as SHA-512 and Bcrypt which have not been broken.

Hashing collisions are the primary way hashing algorithms are deemed broken. Hashing an input is not used to encrypt an input, it is inefficient and ineffective since a hash is a one-way function. There is no way to take the resulted output and put it back through the hashing algorithm to get the original text back. If one intends to do that, there are actual cryptographic functions which provide the desired encryption of plaintext and decryption of ciphertext.

A popular attack for a hashing algorithm is a rainbow table. A rainbow table is used for cracking passwords, or other sensitive data by precomputing the hashes for a predefined set of inputs [23]. For example, a rainbow table for the popular MD5 hash would look like a two-column table, where the first column has the input string such as the letters of the alphabet and dictionary words and the second column would have the corresponding hash of that given input. Once you have a large enough rainbow table, it is fairly easy to search through the table to see if any of the leaked hashes in your possession are part of the table. In doing you, it is immediately apparent which string input yielded the given hash, thus giving away the original password or sensitive data. Such a tactic only works if the match occurs on the whole string, if a string you are in search of is a composite, then the rainbow table will be quite useless unless the computed hash in the table was generated from the composite string. Composing strings would have to be left up to pure chance because there are seemingly infinite combinations of strings, and the table would never be complete. In an essence, a rainbow table is a massive list of common passwords, or inputs and their hashes. A rainbow table is a step up from trying every possible combination of characters against the hashed password, the time taken to guess the correct choice increases exponentially as password length and keyspace increases.

One way to mitigate a rainbow table attack is by adding a salt to the hash. Hashing by itself is simply not enough to protect passwords or other fields. Salts are short random characters that are appended to the end of the password before they are hashed. So, a combination of the password and salt will thwart any rainbow table attack as long as the salt is also not leaked. In general, the salt is stored as plaintext alongside the hash of the password, so that the authentication system knows what the salt was and can verify the hash of the password plus the salt. It may seem counter intuitive that the salt is stored in plain view since you are essentially giving the hacker part of the password, and dictionary attacks and brute force can still be a problem if the hacker knows where to put the salt in the guesses However, rainbow attacks will be ineffective due to their nature. Additionally, most developers also make use of a pepper. A pepper is a short string or character appended to the end of a password. By their nature, peppers are random and different for each password and it is common for them to be around a character or two. So, the new hash stored is the hash of the password, appended with a random pepper and the salt. Suppose this system is used for a login system, a user will enter their username and their password, and the system will append all the possible combinations of the pepper and the designated salt and generate the hashes of every combination until one matches the stored hash. If the hashes match, then the user is allowed to log into the system. The whole point of this is that the pepper is not stored. No one knows what the pepper is, not even the website, until it goes through all the possible options to find it. All of this leads to extra time taken when compared to a system using no pepper. The use of a pepper is not an issue to the user, since they do not know the underlying strategy for hashing, however, it adds an extra layer of complication for anyone looking to crack the system.

The takeaway idea from hashing is to avoid any novel homemade hashing functions or methods that people invent and to avoid reused values. Many developers insist on designing their own hashing functions and think that the concept of secure cryptographic design means to throw together every kind of cryptographic and non-cryptographic operation that can be thought of [24]. The developer thinks that because their own function is over engineered and complex, that the attackers themselves will be unable to crack their methodology. In reality, it’s only the developer who might have confused himself due to the complexity rather than the attacker. Anyone wanting to make use of a password hashing function should look into the currently standardized hashing functions such as SHA or Bcrypt. At one point in time, MD5 was a popular hash function, it is now considered broken. It was commonly used for password hashing due to it being very fast. The primary goal of the salt is to be as unique as possible; the usefulness of salt goes down if the salt value is reused anywhere.

**Technologies Used**

Throughout this project, there were various technologies that were utilized for the successful execution of the workflow of the project.

The front-end website app makes use of HTML, JavaScript with the CSUN website template to match the look and feel. The application is intended to be hosted on Microsoft Azure App Service, if the entire Microsoft stack is used. The frontend is a Node.js app which routes the form data to a simple Azure Function URL, and receives the pdf binary blob and reconstructs it to be displayed to the user. The simplicity and portable nature of the app allows for the hosting to be on any service which allows hosting of Node.js apps, thus bypassing the typical vendor lock-in.

The main part of the project was the Azure Functions Service. The Azure Functions were implemented in Microsoft’s C# programming language. The code responsible for generating the pdf that would be returned to the user, utilized the DinkToPdf wrapper library [25]. DinkToPdf was immensely useful since it allows the user of the library to design the layout of the pdf using html and model out the layout of the file. Additionally, Microsoft’s Azure Documents library was used to query Cosmos DB and execute other requests against the service. There were two hashing algorithms that were employed in the project as well.

Initially, the project was designed with a single hashing algorithm known as the Secure Hashing Algorithm-2 (SHA-2). More specifically, the SHA-2 family consists of functions with various length hash outputs, the specific length used was SHA-512. It is one of the industry standards used when the primary concerns are security and speed [26]. The second hashing algorithm was chosen with security and slowness as the primary attributes. The algorithm that best fit the criteria is known as Bcrypt [21]. It is the default hashing algorithm in some Linux distributions and on OpenBSD operating system. It is resistant to brute-force search attacks because one can specify the number of iterations if must do in its calculations, thus making the algorithm slower.

The administrator’s utility tool was created in C# and the Windows Forms graphical class library, which is part of the Microsoft’s .NET Framework. Microsoft’s graphical framework allowed for the quick prototyping and testing of the parts needed in this project. The benchmarking was performed with BenchmarkDotNet [27]. BenchmarkDotNet is an easy to use benchmarking library that generates reports in various file formats and allows for multiple iterations of the code in order to measure the performance once the code is warmed up. Lastly, the JSON data storage format was chosen to represent the class models, and the Newtonsoft’s Json.NET framework was used extensively due to its serializing and de-serializing capabilities [28].

The testing of the project was carried out locally on machine using Visual Studio 2017 on Windows 10 operating system. Once the project’s dependencies were chosen and installed, their versions were frozen for maximum future compatibility and stability. The project was tested progressively throughout its many phases making use of Microsoft’s Cosmos DB Emulator, Azure Functions offline deployment and the API testing software Postman [29]. Testing was done locally to avoid the costs of utilizing Azure services for a proof of concept project.

**Project Implementation and Workflow**

The workflow of the project will be described from the perspective of the end user and the perspective of the administrator. The overall high-level workflow leads the data from the front-end website app to the function responsible for searching the database for the student’s academic information and compiling said information into a pdf and transmitting it back to the user so their requested pdf file can be displayed in the matter of a few seconds. The alternative workflow showcases how the student’s data is organized, compiled, uploaded to the database, and secured against any future tampering. Each section will also touch upon the idea of accomplishing this in a secure manner on a serverless architecture.

**End User** **Workflow**

As stated in the introduction of this section, the workflow for the end-user, be it the student themselves or a business requesting the student’s records, was designed to be simple and fast.

[1] A. © C. S. University, N. 18111 N. Street, Northridge, and C. 91330 P.-1200 / C. Us, “Transcripts and Diplomas,” *California State University, Northridge*, 08-Jul-2013. [Online]. Available: https://www.csun.edu/alumni/transcripts-and-diplomas. [Accessed: 10-Jun-2019].

[2] M. Fowler, “Serverless Architectures,” *martinfowler.com*. [Online]. Available: https://martinfowler.com/articles/serverless.html. [Accessed: 10-Jun-2019].

[3] F. Bashir, “What is Serverless Architecture? What are its Pros and Cons?,” *Hacker Noon*, 28-May-2018. [Online]. Available: https://hackernoon.com/what-is-serverless-architecture-what-are-its-pros-and-cons-cc4b804022e9. [Accessed: 10-Jun-2019].

[4] M. Nofer, P. Gomber, O. Hinz, and D. Schiereck, “Blockchain,” *Bus. Inf. Syst. Eng.*, vol. 59, no. 3, pp. 183–187, Jun. 2017.

[5] “Blockchain explained,” *Reuters*. [Online]. Available: http://graphics.reuters.com/TECHNOLOGY-BLOCKCHAIN/010070P11GN/index.html. [Accessed: 10-Jun-2019].

[6] S. W. 10 O. March, “Blockchain for Student Records,” *Bold Business*, 10-Oct-2017. .

[7] “IBM unveils Blockchain as a Service based on open source Hyperledger Fabric technology,” *TechCrunch*. .

[8] “About,” *Learning Machine*. [Online]. Available: https://www.learningmachine.com/about. [Accessed: 23-Jun-2019].

[9] M. Techlabs, “What is Serverless Architecture? What are its criticisms and drawbacks?,” *Medium*, 04-May-2017. .

[10] “Bare-metal server,” *Wikipedia*. 28-Mar-2019.

[11] “A brief history of serverless (or, how I learned to stop worrying and start loving the cloud),” *freeCodeCamp.org News*, 05-Apr-2018. [Online]. Available: https://www.freecodecamp.org/news/a-brief-history-of-serverless-or-how-i-learned-to-stop-worrying-and-start-loving-the-cloud-7e2fc633310d/. [Accessed: 24-Jun-2019].

[12] May 26 and 2017, “Docker Downsides: Container Cons to Consider before Adopting Docker,” *Channel Futures*, 26-May-2017. [Online]. Available: https://www.channelfutures.com/open-source/docker-downsides-container-cons-to-consider-before-adopting-docker. [Accessed: 24-Jun-2019].

[13] ggailey777, “Supported languages in Azure Functions.” [Online]. Available: https://docs.microsoft.com/en-us/azure/azure-functions/supported-languages. [Accessed: 26-Jun-2019].

[14] “AWS Lambda – FAQs,” *Amazon Web Services, Inc.* [Online]. Available: https://aws.amazon.com/lambda/faqs/. [Accessed: 26-Jun-2019].

[15] “HTTP Methods GET vs POST.” [Online]. Available: https://www.w3schools.com/tags/ref\_httpmethods.asp. [Accessed: 26-Jun-2019].

[16] E. Geschwinde and H.-J. Schoenig, “An Introduction to SQL,” 2002.

[17] S. Yegulalp, “What is NoSQL? NoSQL databases explained,” *InfoWorld*, 07-Dec-2017. [Online]. Available: https://www.infoworld.com/article/3240644/what-is-nosql-nosql-databases-explained.html. [Accessed: 27-Jun-2019].

[18] timsander1, “Getting started with SQL queries in Azure Cosmos DB.” [Online]. Available: https://docs.microsoft.com/en-us/azure/cosmos-db/sql-query-getting-started. [Accessed: 27-Jun-2019].

[19] “Hash function,” *Wikipedia*. 31-May-2019.

[20] “What Is Hashing? [Step-by-Step Guide-Under Hood Of Blockchain],” *Blockgeeks*, 06-Aug-2017. [Online]. Available: https://blockgeeks.com/guides/what-is-hashing/. [Accessed: 27-Jun-2019].

[21] “bcrypt,” *Wikipedia*. 05-Apr-2019.

[22] “Learn Cryptography - Hash Collision Attack.” [Online]. Available: https://learncryptography.com/hash-functions/hash-collision-attack. [Accessed: 27-Jun-2019].

[23] “Rainbow table,” *Wikipedia*. 13-Jul-2019.

[24] “appsec - How to securely hash passwords?,” *Information Security Stack Exchange*. [Online]. Available: https://security.stackexchange.com/questions/211/how-to-securely-hash-passwords. [Accessed: 18-Jul-2019].

[25] R. Dvojmoč, *C# .NET Core wrapper for wkhtmltopdf library that uses Webkit engine to convert HTML pages to PDF.: rdvojmoc/DinkToPdf*. 2019.

[26] “Wayback Machine,” 26-May-2013. [Online]. Available: https://web.archive.org/web/20130526224224/http://csrc.nist.gov/groups/STM/cavp/documents/shs/sha256-384-512.pdf. [Accessed: 21-Jul-2019].

[27] *Powerful .NET library for benchmarking. Contribute to dotnet/BenchmarkDotNet development by creating an account on GitHub*. .NET Foundation, 2019.

[28] J. Newton-King, *Json.NET is a popular high-performance JSON framework for .NET: JamesNK/Newtonsoft.Json*. 2019.

[29] “Postman | API Development Environment,” *Postman*. [Online]. Available: https://www.getpostman.com. [Accessed: 21-Jul-2019].