Alexander Pierce

December 3rd, 2015

CMP-SCI 5750

Assignment 3

vidsig Architecture Proposal

Overall Architecture

vidsig is a new product, so rather than the complications of moving from legacy systems we get to start with a clean slate. To take full advantage of this opportunity, we attempt to minimize the usage of internal networks and do as much as possible with proven cloud services. The internal network will be used for things not related to the launched *vidsig* application itself, such as incremental development, internal communication, etc. Where possible, SaaS should be embraced to minimize employees working on anything outside of their core competency; there are any number of SaaS options to handle payroll, collaboration tools for coding teams, etc. *vidsig* should never be reinventing a wheel when it can get an expertly designed wheel elsewhere.

We recommend that *vidsig* be launched using Amazon's Web Services (AWS) using a hybrid cloud(s). The *vidsig* companion application on the consumer's device will send a video file to the Amazon Elastic Cloud Computing (EC2) Platform as a Service (PaaS) public cloud portion of the process where it will be analyzed[1]. Once analysis is complete and a signature has been created for the video, our database of known signatures (setup in an AWS S3 private cloud using DynamoDB) is queried for a match. If one is provided, the information provided is returned to the public cloud where it can be transmitted to the consumer.

We recommend Amazon specifically because of its broad, interconnecting services global penetration. Consider the success that Netflix has shown through its adoption of Amazon's cloud services; Netflix is essentially a case study of what to do right, and it would be foolish to ignore the lessons their growth provides. With AWS we can not only process the video into a signature

[1] Video analysis is a compute intensive process; we do not want to attempt to do this on the customer device; further this would open create a greater opportunity for others in image processing to attempt to reverse engineer *vidsig*'s algorithms (if they were locally available), which must be impressive to do what it does. Such analysis would likely include some kind of automatic cropping – so that only the video and not other things caught in the frame like a tv-stand, etc. are passed onto the hashing algorithm – as well as possibly some kind of contextual data; using GPS coordinates and time-stamp it might be possible to cross-reference against broadcast/cable schedules from the uploader's local geographic region in an attempt to increase the probability of correct video identification. Whatever the algorithm, once a video's signature is generated, it must be used to query a database.

(EC2), but store our signature database (in Amazon's S3 using DynamoDB), and further take advantage of the benefits of a provider with a history of audit compliance and security[2]; using AWS will allow *vidsig* to focus on its core competency without trying to manage multiple cloud providers and differing environments. The overall recommendation is to make the fullest use of AWS as is possible for the core *vidsig* architecture. While there might be some concern that *vidsig* should start with IaaS, Netflix is not a fly-by-night operation and, again, their track record with Netflix shows that should *vidsig* take off like a rocket and need excessive resources, that AWS will be able to scale to provide them. Further, IaaS would require additional setup and maintenance; it seems a better trade-off to pay AWS to handle the infrastructure so that *vidsig* can avoid getting caught up in such low-level concerns.

Signature Storage and Creation

vidsig is useless without a database to make comparisons against. We will need to store very large databases[3] and query them constantly (hopefully, as we want *vidsig* to be successful). Further, those databases will need to undergo periodic updates as new videos are processed and given signatures. EC2 is a good choice for the processing involved here, as it is extremely scalable. In addition, AWS DynamoDB is the suggested choice for storing our signature database; not only is DynamoDB built on NoSQL (so the databases are transportable if *vidsig* decides to leave Amazon), but it provides "automatic and synchronous data replication across three facilities in a Region. This helps protect [*vidsig*'s] data against individual machine, or even facility level failures."[4] The cloud processing content to create the database of signatures would be private. The database clouds responding to the signature-check requests would be read-only as read-only databases in SQL have better response time. Additions to the database could be rolled out across duplicate database clouds asynchronously so that there is no overall downtime; at worst some regions might be slightly behind others (and so be unable to identify a video input that a request routed to another, more recently updated region was capable of identifying), but this prevents constant additions to the database slowing (and thus negatively impacting) the overall quality of service.

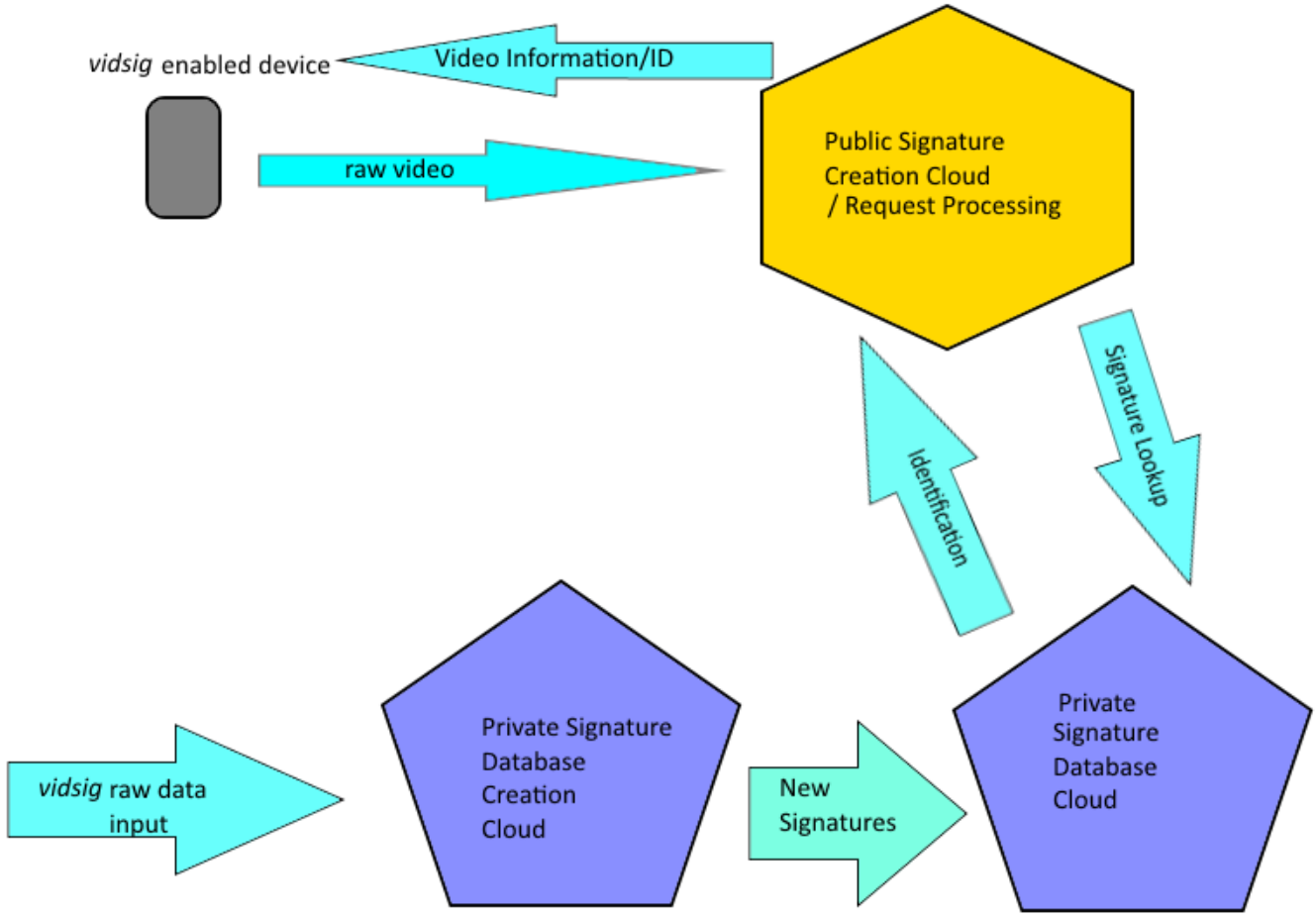
---

[2] https://aws.amazon.com/compliance/ Amazon is compliant with multiple cloud security standards ranging the gamut from PCI DSS to FedRAMP and beyond.

[3] While the individual signatures and characteristics will be relatively small, video is a Big Data topic; even excluding amateur videos such as a baby's first steps (perhaps posted via Vine or YouTube), each day there are 24 hour newscasts, television shows, new released movies, and not just in the United States, but globally. Thus *vidsig* has an immense amount of data to sift, quantify, and store. This data would be raw video with known attributes (title, actors, whatever is considered pertinent) selected by some metric (popularity, regional availability, etc.) and sent to a separate private cloud for signature creation. After the signature is created the video would be deleted from the private cloud. At periodic intervals, the new signatures generated from this data would be added to the queryable private database cloud.

[4] https://aws.amazon.com/dynamodb/details/ In fact, you can further replicate databases across regions, to "build globally distributed applications with lower-latency data access, better traffic management, and easier disaster recovery and data migration." (ibid.)

Given the cost (in processing time) of creating the signature database, we further recommend a regular physical backup of the signature database; this would be intended as a check against a malicious act by an employee, or other person with access to/control of our AWS holdings. There will always be multiple copies of the *vidsig* algorithm in house as it is incrementally updated and refined, but the signature database will live in the cloud; to avoid losing all of that labour, secure physical backups make the most sense. Due to the high amount of data constantly being processed, we recommend no less than weekly incremental backups, with semi-annual full duplicates made of the database to mitigate the potential loss due to a damaged incremental backup while also balancing against the costs associated with creating and storing the physical media.



Raw consumer video is sent to public cloud where it is processed into a signature. The public cloud requests matching signature information from the Private Signature Database Cloud. Any information found is returned to the Public Cloud which returns the information to the consumer. To create the private signature database, raw video data (with known attributes) is sent to the Private Signature Database Creation Cloud where signatures are created from the raw video data and associated with the known attributes. These new signatures are periodically sent to the Private Signature Database Cloud to update its database. (Not pictured in this diagram are backups, nor logging.)

Logging

Automated logging should be in place on all clouds, as well as the company's internal network. Splunk, a log analysis SaaS is available through AWS; while there are other options, Splunk is well known and because it is already designed to work with AWS there will be no problems with any of our clouds not working with Splunk. Further, there is an enterprise edition of Splunk that is designed to work hand in hand with the cloud version; we can therefore use Splunk to handle log analysis of the company's internal network as well. This prevents us from relying on two different vendors, having cross-platform difficulties, or even the simple requirement of having to learn two different UIs.

Auditing

vidsig may have to implement specific security/undergo specific audits based on the source of some of the raw video data that is used to create the signature database; due to copyright law, there is a necessity to make certain that no video is being stored for longer than is necessary to process it, nor should any video being processed ever be stored in a public cloud – that just opens the door to potential abuse or lawsuits from rights holders. With the potential exception of the data undergoing processing for signature creation, none of the data stored in the *vidsig* clouds is required to be held to any specific security standard, and so it is not necessary that the clouds implement specific standards to pass such audits. Internal (*vidsig* corporate) auditing should occur to make certain that social engineering or lazy password security does not expose the clouds to security vulnerabilities. To whatever limited extent *vidsig* might be required to safeguard the private data processing cloud, *vidsig* should contract with external audit firms to determine best practices and our state of compliance; access control is a relatively simple problem to handle, but if *vidsig* needs to pass any industry-specific audits, it is best that such work be outsourced as it is outside of the company's core competency and it would be better to do it correctly from the start than to make a best-effort that still might have flaws obvious to industry personnel.

Security

The consumer sending video to the public signature processing cloud neither needs, nor expects privacy (see for example SoundHound, which is a similar application designed to identify audio). As long as the request follows the proper format, and is not abusive (request flooding)[5] there is no reason to worry about who is sending it other than to send the response to the appropriate place. As with most SaaS providers – which *vidsig* will be once all is said and done – *vidsig* should have a thorough Terms of Service / User Agreement eliminating all liability and declaring that all use of the service, application, etc. are "as is" and with no guarantee of security. The physical, and much of the network security of the public and private clouds will be handled by AWS, but obviously it is imperative that *vidsig* implement encryption in all communication with

[5] https://d0.awsstatic.com/whitepapers/DDoS_White_Paper_June2015.pdf

the private clouds. AWS has a marketplace with software designed to work with its cloud offerings to mitigate security risks. We recommend that we deploy a Web Application Firewall available from the AWS Marketplace on the public and private clouds, and only process requests to the private signature database cloud from the public cloud, or the private signature creation cloud. The private signature creation cloud should only process requests from our internal network; this limits the private clouds' exposure to attack as everything must come from a trusted source. AWS has further capabilities which we can take advantage of depending on where we deploy the application; if, for example, *vidsig* is only "live" in North America and Europe (perhaps due to regional licensing contracts that *vidsig* agreed to so that it could get cooperation from content creators to grow its signature database more rapidly), we can deny incoming requests from any non-member country using AWS built in Geo-Blocking.

With those safeguards in place, we also need strict password security, and require that connection to the company network be made either from on-company campus or via VPN, in either case using two factor authorization to further limit potential intrusions. Any malicious action that takes place (as a result of a password being compromised or otherwise) should be limited in scope and the legal repercussions directly assignable due to these requirements.