Fusion V9 Architecture Guide
Planning for an on-premise deployment
Table of Contents

Introduction........................................................................................................................................... 1
Overview.................................................................................................................................................. 1
Fusion V9.x Platform Requirements ...................................................................................................... 2
  O/S and Software Support .................................................................................................................. 2
    Web Layer ......................................................................................................................................... 2
    Data Layer ......................................................................................................................................... 2
    Processing Layer ............................................................................................................................. 2
Virtualization ............................................................................................................................................ 2
Concepts and Components ...................................................................................................................... 3
  Web Layer (API Farm) .......................................................................................................................... 3
    Fusion Web Service API ................................................................................................................... 3
    Biometric API .................................................................................................................................... 4
    Management API ............................................................................................................................. 4
    API Security ....................................................................................................................................... 4
Secure encrypted access .......................................................................................................................... 4
Role-based method permissions ............................................................................................................ 4
Fully anonymized metadata .................................................................................................................. 5
Management UI Portal ............................................................................................................................ 5
Data Layer (Biometric Data Store) ......................................................................................................... 5
    Fusion Engine .................................................................................................................................... 6
    Content Addressable Storage .......................................................................................................... 6
Biometric Processing Layer (Engine Farm) ............................................................................................. 6
Scalability and High Availability ........................................................................................................... 8
  Example enterprise deployment .......................................................................................................... 8
    Web Layer ......................................................................................................................................... 8
    Data Layer ......................................................................................................................................... 9
    Processing Layer ............................................................................................................................. 9
Appendix 1: Base-level Production System .......................................................................................... 10
  Example Requirements ......................................................................................................................... 10
  Application Performance ...................................................................................................................... 10
Typical Base-level System Components

Web Layer
Data Layer
Processing Layer
Load Balancing
Introduction

This architecture document is designed to assist in the planning of an on-premise deployment of the VoiceVault Fusion V9 voice biometric system.

The typical intended audience for this guide would be IT Manager/Director, Security Manager, Architect, Project Manager and/or Application Development.

Overview

VoiceVault Fusion consists of a multi-tier asynchronous architecture built on a standard commercial enterprise server platform. Diagram 1 below is a conceptual diagram showing the relationships between the various tiers of a typical Fusion system for a basic IVR-based application.

![Diagram 1: Infrastructure overview](image-url)

The software platform used is based on Microsoft technology, as detailed in the following section. The specific architecture hardware and software is determined from the application requirements, particularly the maximum transaction rate, scalability growth and need for redundancy and disaster recovery.
Fusion V9.x Platform Requirements

Each layer will typically consist of one or more servers depending on the application requirements.

O/S and Software Support

Web Layer
- Windows Server 2012 R2 x64, Standard or higher
- IIS 8.0 or above
- .NET Framework 4.5+

Data Layer
- Windows Server 2012 R2 x64, Standard or higher
- SQL Server 2012 x64, Standard or higher\(^1\)
- .NET Framework 4.5+\(^2\)

Processing Layer
- Windows Server 2012 R2 x64, Standard or higher
- .NET Framework 4.5+

Virtualization
Due to the use of standard commercial platforms throughout the architecture, virtualization should be supported using most of the commonly available systems. VoiceVault Fusion has specifically been tested on the following:
- VMWare ESX 4.1 Update 3
- VMWare ESXi 5.0 Update 2
- Microsoft Hyper-V on Windows Server 2012 and above
- Citrix XenServer 6.x
- Amazon EC2 Xen-virtualized images

\(^1\) It is feasible to use SQL Server Express when performance or scalability is not important, such as for development purposes

\(^2\) Required if Data Migration Service is installed on the Database server(s)
Concepts and Components

The following shows the major components of the Fusion multi-layer architecture (Diagram 2).

Diagram 2: Fusion multi-layer architecture

Web Layer (API Farm)
The Web Layer hosts the VoiceVault Fusion Web Services APIs. All requests to VoiceVault Fusion from customer applications are made over SSL/TLS secured and encrypted connections. Access to the Fusion Engine must go through the Web Layer APIs, and no external access is directly permitted to the Data Layer for reasons of security.

This implementation also facilitates the integration of Fusion in a secure environment where the Web Servers/Farm are isolated from the database in a Data Center, and where they are potentially under the responsibility of completely different teams with no common access rights.

All API methods are designed to be asynchronous, where the client application polls for responses. No server state is required to be maintained by the Web Layer between method calls, enabling almost limitless horizontal scalability within this layer to meet the required API load.

Fusion Web Service API
The Fusion Web Service API is typically the only interface required for most applications, and is implemented as a simple form-based service layer on top of the Biometric and Management
APIs. It exposes the most common VoiceVault Fusion operations required for IVR and mobile app development.

The Fusion Web Service API is called using simple HTTP operations, either POST or GET, and secured using an SSL/TLS transport. Due to the more extensive integration required to use the SOAP APIs, it is recommended for use by all clients that do not need direct access to the detailed reporting methods in the Management API. Refer to the Fusion Management API documentation for further details.

Biometric API
The Biometric API is the main Windows Communication Foundation (WCF) interface for all voice verification functionality. It is exposed as a .NET 4.5 Web Service using HTTP over SOAP bindings and secured using SSL/TLS transport security.

As all the functionality within the Biometric API is available via the Fusion Web Service API, it is primarily designed for systems that specifically require WCF/SOAP integration such as applications developed using Visual Studio.

Management API
The Management API provides a WCF development interface for building system management or reporting directly into client applications. As with the Biometric API, it is exposed as a .NET 4.5 Web Service using HTTP over SOAP bindings secured using SSL/TLS.

Only a few of the methods in the Management API are exposed via the Fusion Web Service API, specifically relating to the registration and management of Fusion claimants³. In almost all applications, integration with the Management API will not be required as all essential functionality is accessible in the Fusion Web Service API, and extensive reporting is available using the Management UI portal. It will usually only be used when direct integration with a client reporting system is necessary.

API Security

Secure encrypted access
Access to all Fusion APIs is secured over SSL/TLS connections, and also requires a valid username and password to be passed for every call made. Within the Fusion database, all passwords are stored only in fully hashed and salted form. All other servers are fully locked-down and totally inaccessible externally, and only the secured SSL/TLS port (e.g. 443) is exposed.

Role-based method permissions
The API methods are additionally accessible only by users with the correct role, allowing fine-grained access control to the Fusion functionality. Refer to the relevant API documentation for details of the extensive role-based permissions available within VoiceVault Fusion.

---

³ A claimant is a registered biometric user required to authenticate within a Fusion-based application

⁴ By default only TLS connectivity is enabled and use of SSL is now deprecated
Fully anonymized metadata
Claimants registered within Fusion for authentication are identified exclusively using a randomly generated UUID for all API calls. No personally identifiable information is required or stored, and therefore cannot be traced back to anybody registered within the system.

Management UI Portal
The Management UI Portal is a fully functional web-based console allowing administrators to view all transactions and audit logs, manage configuration, claimant and processing engine settings, and to obtain system activity reports (diagram 3).

Access to the portal is restricted and secured over SSL/TLS connections using HTTPS certificate(s). As the portal is built on top of the Fusion Web Services API, it also incorporates all the security features and role-based access permissions supported by the Fusion APIs.

![Diagram 3: Fusion Management Portal (screenshot)](image)

Data Layer (Biometric Data Store)
The Data Layer maintains all the data required by the VoiceVault Fusion voice biometric system, as follows:

- **Biometric data**, including enrollment voice models for all claimants
- **Transactional data**, including historical and live enrollment/verification dialogues
- **Configuration data**, including Fusion engine thresholds and processing-related settings
- **Audit data**, including full audit trails for all enrollments and verifications

All the static, dynamic and transient data used by the biometric processors in the Biometric Processing Layer is also stored within the Data Layer.
Fusion Engine

The Fusion Engine consists of the relevant business logic and persistent storage management required for the VoiceVault Fusion voice biometric system. It is built on top of a standard deployment of Microsoft SQL Server, and therefore leverages all the built-in enterprise level functionality to provide proven transactional safety, resilience, redundancy and scalability (such as Database Mirroring, Always-On Availability Groups, etc.).

Standard backup tools for SQL Server can also be used to safely back-up the entire Fusion Engine and associated data, securing all historical data and audit trails for corporate or legislative purposes, as well as fully supporting Disaster Recovery processes.

Content Addressable Storage

The new Content Addressable Storage (CAS) functionality built in to the Fusion system provides a mechanism for storing crucial biometric data in a way that is both secure and independent of any specific storage technology or location. In conjunction with the biometric Data Migration Service, it supports secure storage and migration across multiple data stores, centers or silos.

The integration of CAS within the Fusion system significantly improves security and resilience of essential biometric data:

- All stored data is identified and secured via a cryptographic hash, to prevent tampering and to support e-signature applications.
- Biometric data can now be stored simultaneously in multiple disparate locations, including network connected and remote storage repositories (e.g. SAN/NAS, Amazon S3 buckets), providing full redundancy and supporting disaster recovery. Hardware and software encryption can be implemented to provide additional security.
- The Data Migration service supports flexible data backup and archival, allowing older data to be automatically transitioned to lower cost or archival storage based on age, while still ensuring data is accessible for use by the system.
- As part of the migration schedule, biometric data can also be automatically deleted and removed as required to meet corporate data protection policies.

Biometric Processing Layer (Engine Farm)

The biometric processors in the Processing Layer are responsible for executing all VoiceVault’s proprietary voice biometric algorithms, including voice model generation, voice verification, Signal Quality Measurement (SQM) analysis and replay attack detection.

To support the asynchronous Fusion architecture, a pull model is implemented, where work requests from the API calls are queued in the Data Layer to be processed by the next available biometric processor. After completing the request, the processor writes the results back to the Data Layer to be retrieved by the application.

The biometric processors are implemented as a stateless Windows service, and multiple processors can be installed per server or across multiple servers. Therefore, almost any number can be deployed in many different configurations within a system to support even the highest biometric transaction rate. Additionally, they can be associated with individual configurations.
(e.g. languages, digit support, multiple passphrases), enabling **flexible and extensive scalability** of the Processing Layer itself.
Scalability and High Availability

VoiceVault Fusion is specifically designed for high availability and scalability within enterprise environments. Using a multi-tier asynchronous architecture which is scalable at every level, and through leveraging the enterprise-level platforms that it is built on, it is possible to build a system to meet even the most demanding requirements. Careful design of the system to ensure that each tier is independently scalable and not tightly-coupled means that it also supports all standard load balancing technologies, therefore minimizing both the requirements and cost.

Using an architecture based on Amazon EC2 instances, VoiceVault has built Fusion deployments that can support over 100 simultaneous authentications per second, and can scale an enterprise system to meet almost any throughput and availability requirements.

Example enterprise deployment

At the most basic level, a Fusion system could be implemented on a single server if only a low throughput was required without redundancy or disaster recovery, e.g. for development purposes. For a typical enterprise-scale deployment, the architecture would instead be designed to support scalability and with partly or fully redundant servers for resilience and high-availability (Diagram 4).

Diagram 4: Typical Fusion deployment for IVR application

Web Layer

The Web Layer can be scaled out through the addition of API server nodes, which are load balanced using any standard stateless HTTP load balancing solution:

- Round robin DNS
- Weighted least-load routing
- Microsoft Network Load Balancing
- Content Switching Appliance (such as Citrix NetScaler)
Load balancing at the API layer is also used to support high-availability, and API servers can even be located across different Data Centers as required. The exact implementation will be dependent upon requirements for availability and to support the expected API load.

Data Layer
The Data Layer is primarily scalable by leveraging the capabilities available in Microsoft SQL Server, such as Microsoft Cluster Services for SQL Server, Always-On Availability Groups, Database Mirroring (including automated fail-over using a Witness server) and Transactional Replication.

Where an existing SQL Server deployment is available, such as within an enterprise Data Center, then it is feasible to leverage this within the Fusion system as long as the throughput and capacity is suitable.

The Content Addressable Storage module within the Data Layer supports storage within the database (BLOBs), on local or remote filesystems, and even in Amazon S3 cloud services. The flexibility of the CAS and associated Data Migration modules supports both redundancy and scalability for the storage of the biometric data within the system.

Processing Layer
The Processing Layer can be scaled both horizontally and vertically, through the provision of additional processing capability within one or more existing servers and by adding servers to host more processing engines.

The exact number of processing engines that need to be deployed will depend on the number of different configurations supported (e.g. languages, digit modes, passphrases), the peak biometric transaction rates that need to be processed, and the overall system response times required.
Appendix 1: Base-level Production System

For a typical basic Production system, it is assumed that the hardware will be virtualized and potentially cloud-based (public, private or hybrid). Therefore, the hardware requirements are expressed here in terms of CPU cores, RAM and required disk storage space.

Note that these examples are provided as illustrations only and, although they are based on known Fusion deployments, there may be specific implementation details, different usage profiles or additional requirements which mean that similar deployments may vary in experience and performance.

Example Requirements

The examples that follow are for a typical production system, supporting a combined total load of up to 30,000 verification and enrollment transactions per hour using digit-mode only. It is estimated that the system would require a single farm of two biometric processor servers (assuming dual quad core 2.4GHz, 8GB RAM) to support the required throughput and provide 100% redundant capacity.

Application Performance

For a properly sized small/medium-sized system, the end-to-end biometric processing for either an enrollment or verification should be sub-second, and it will typically support in excess of 5 transactions per second.

For mobile applications over a cellular connection, or from a web application over the internet, there will be additional overhead due to the infrastructure and latency of the network, which is influenced by factors such as the device used, network provider, current cell loading, signal strength, backbone network bandwidth, etc.

Trials of mobile connectivity over 4G with a simulated loaded network indicate that download speeds of over 10Mbits/s, upload speeds up to 6Mbits/s, and latencies of around 30 milliseconds are possible. However, real-world results will vary widely, but typically upload speeds would be expected to be well in excess of 2Mbits/s with latencies of between 100-200ms. For a well-designed system, using a typical compressed audio files (A-Law/U-Law), the total end-to-end response should be well under 1 second (for VoiceVault processing and network latency).

In addition, by careful asynchronous design of the application, it is possible to record multiple utterances while submitting and processing previous utterances in an enrollment/verification session, and so this small delay will only be experienced by the user after the last utterance is recorded.
Typical Base-level System Components

Web Layer
- Two Windows Server nodes with IIS deployed:
  - Two dedicated cores running at least 2.4 GHz
  - 4GB RAM
- .NET Framework 4.5

Data Layer
- Two Windows Server nodes with SQL Server deployed:
  - Four dedicated cores running at least 2.4 GHz
  - 16GB RAM
  - 2TB free disk space
- SQL Server 2012 R2 configured for high availability/failover as required

Processing Layer
- Windows Server nodes, variable depending upon virtualization decision:
  - Total of eight dedicated cores - number of Windows OS instances at the discretion of the customer (e.g. 4 dual core Windows Server installs or 2 quad core Windows Server installs)
  - Each node requires minimum 4GB RAM
  - Each node requires 20GB free disk space
- .NET Framework 4.5

Load Balancing
- Load balancing solution at customer discretion
- Recommended with Two dedicated cores running at least 2.4 GHz

---

5 Where biometric data is stored via CAS then these disk requirements apply to the relevant file/network storage, and assumes the biometric data is not replicated across multiple locations