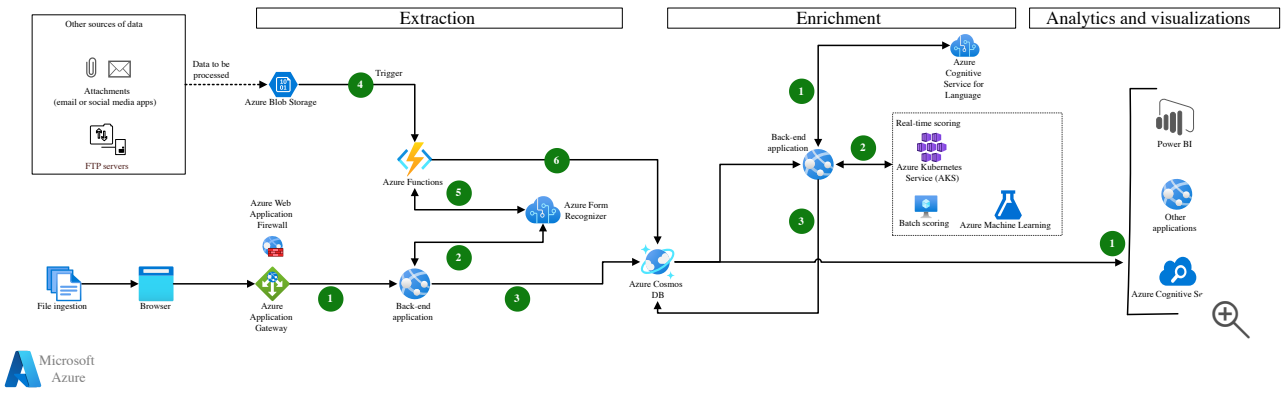
Automate document processing by using AI Document Intelligence

Azure AI Search Azure AI services Azure Cosmos DB Azure AI Document Intelligence Azure Machine Learning

This article outlines a scalable and secure solution for building an automated document processing pipeline. The solution uses AI Document Intelligence for the structured extraction of data. Natural language processing (NLP) models and custom models enrich the data.

Architecture



Download a Visio file of this architecture.

Dataflow

The following sections describe the various stages of the data extraction process.

Data ingestion and extraction

1. Documents are ingested through a browser at the front end of a web application. The documents contain images or are in PDF format. Azure App Service hosts a back-end application. The solution routes the documents to that application through Azure Application Gateway. This load balancer runs with Azure Web Application Firewall, which helps to protect the application from common attacks and vulnerabilities.

2. The back-end application posts a request to an Azure AI Document Intelligence REST API endpoint that uses one of these models:

   - Layout
   - Invoice
   - Receipt
   - ID document
   - General document
   - US tax document models
   - US mortgage document model

   The response from Azure AI Document Intelligence contains raw optical character recognition (OCR) data and structured extractions.

3. The App Service back-end application uses the confidence values to check the extraction quality. If the quality is below a specified threshold, the app flags the data for manual verification. When the extraction quality meets requirements, the data enters Azure Cosmos DB for downstream application consumption. The app can also return the results to the front-end browser.

4. Other sources provide images, PDF files, and other documents. Sources include email attachments and File Transfer Protocol (FTP) servers. Tools like Azure Data Factory and AzCopy transfer these files to Azure Blob Storage. Azure Logic Apps offers pipelines for automatically extracting attachments from emails.

5. When a document enters Blob Storage, an Azure function is triggered. The function:

   - Posts a request to the relevant Azure AI Document Intelligence pre-built endpoint.
   - Receives the response.
   - Evaluates the extraction quality.

6. The extracted data enters Azure Cosmos DB.

Data enrichment

The pipeline that's used for data enrichment depends on the use case.

1. Data enrichment can include the following NLP capabilities:

   - Named entity recognition (NER)

- The extraction of personal information, key phrases, health information, and other domain-dependent entities

To enrich the data, the web app:

- Retrieves the extracted data from Azure Cosmos DB.
- Posts requests to these features of the AI Language API:
  - NER
  - Personal information
  - Key phrase extraction
  - Text Analytics for health
  - Custom NER, which is in preview
  - Sentiment analysis
  - Opinion mining
- Receives responses from the AI Language API.

2. Custom models perform fraud detection, risk analysis, and other types of analysis on the data:

   - Azure Machine Learning services train and deploy the custom models.
   - The extracted data is retrieved from Azure Cosmos DB.
   - The models derive insights from the data.

   These possibilities exist for inferencing:

   - Real-time processes. The models can be deployed to managed online endpoints or Kubernetes online endpoints, where managed Kubernetes cluster can be anywhere including Azure Kubernetes Service (AKS) .
   - Batch inferencing can be done at batch endpoints or in Azure Virtual Machines.

3. The enriched data enters Azure Cosmos DB.

Analytics and visualizations

1. Applications use the raw OCR, structured data from Azure AI Document Intelligence endpoints, and the enriched data from NLP:

   - Power BI displays the data and presents reports on it.
   - The data functions as a source for Azure Cognitive Search.
   - Other applications consume the data.

Components

- App Service is a platform as a service (PaaS) offering on Azure. You can use App Service to host web applications that you can scale in or scale out manually or automatically. The service supports various languages and frameworks, such as ASP.NET, ASP.NET Core, Java, Ruby, Node.js, PHP, and Python.

- Application Gateway is a layer-7 (application layer) load balancer that manages traffic to web applications. You can run Application Gateway with Azure Web Application Firewall to help protect web applications from common exploits and vulnerabilities.

- Azure Functions is a serverless compute platform that you can use to build applications. With Functions, you can use triggers and bindings to react to changes in Azure services like Blob Storage and Azure Cosmos DB. Functions can run scheduled tasks, process data in real time, and process messaging queues.

- Azure AI Document Intelligence is part of Azure AI services. Azure AI Document Intelligence offers a collection of pre-built endpoints for extracting data from invoices, documents, receipts, ID cards, and business cards. This service maps each piece of extracted data to a field as a key-value pair. Azure AI Document Intelligence also extracts table content and structure. The output format is JSON.

- Azure Storage is a cloud storage solution that includes object, blob, file, disk, queue, and table storage.

- Blob Storage is a service that's part of Azure Storage. Blob Storage offers optimized cloud object storage for large amounts of unstructured data.

- Azure Data Lake Storage is a scalable, secure data lake for high-performance analytics workloads. The data typically comes from multiple heterogeneous sources and can be structured, semi-structured, or unstructured. Azure Data Lake Storage Gen2 combines Azure Data Lake Storage Gen1 capabilities with Blob Storage. As a next-generation solution, Data Lake Storage Gen2 provides file system semantics, file-level security, and scale. But it also offers the tiered storage, high availability, and disaster recovery capabilities of Blob Storage.

- Azure Cosmos DB is a fully managed, highly responsive, scalable NoSQL database. Azure Cosmos DB offers enterprise-grade security and supports APIs for many databases, languages, and platforms. Examples include SQL, MongoDB, Gremlin,

Table, and Apache Cassandra. Serverless, automatic scaling options in Azure Cosmos DB efficiently manage capacity demands of applications.

- AI Language offers many NLP services that you can use to understand and analyze text. Some of these services are customizable, such as custom NER, custom text classification, conversational language understanding, and question answering.

- Machine Learning is an open platform for managing the development and deployment of machine-learning models at scale. Machine Learning caters to skill levels of different users, such as data scientists or business analysts. The platform supports commonly used open frameworks and offers automated featurization and algorithm selection. You can deploy models to various targets. Examples include AKS, Azure Container Instances as a web service for real-time inferencing at scale, and Azure Virtual Machine for batch scoring. Managed endpoints in Machine Learning abstract the required infrastructure for real-time or batch model inferencing.

- AKS is a fully managed Kubernetes service that makes it easy to deploy and manage containerized applications. AKS offers serverless Kubernetes technology, an integrated continuous integration and continuous delivery (CI/CD) experience, and enterprise-grade security and governance.

- Power BI is a collection of software services and apps that display analytics information.

- Azure Cognitive Search is a cloud search service that supplies infrastructure, APIs, and tools for searching. You can use Azure Cognitive Search to build search experiences over private, heterogeneous content in web, mobile, and enterprise applications.

Alternatives

- You can use Azure Virtual Machines instead of App Service to host your application.

- You can use any relational database for persistent storage of the extracted data, including:
  - Azure SQL Database .
  - Azure Database for PostgreSQL .
  - Azure Database for MySQL .

Scenario details

Automating document processing and data extraction is an integral task in organizations across all industry verticals. AI is one of the proven solutions in this process, although achieving 100 percent accuracy is a distant reality. But, using AI for digitization instead of purely manual processes can reduce manual effort by up to 90 percent.

Optical character recognition (OCR) can extract content from images and PDF files, which make up most of the documents that organizations use. This process uses key word search and regular expression matching. These mechanisms extract relevant data from full text and then create structured output. This approach has drawbacks. Revising the post-extraction process to meet changing document formats requires extensive maintenance effort.

Potential use cases

This solution is ideal for the finance industry. It can also apply to the automotive, travel, and hospitality industries. The following tasks can benefit from this solution:

- Approving expense reports
- Processing invoices, receipts, and bills for insurance claims and financial audits
- Processing claims that include invoices, discharge summaries, and other documents
- Automating statement of work (SoW) approvals
- Automating ID extraction for verification purposes, as with passports or driver licenses
- Automating the process of entering business card data into visitor management systems
- Identifying purchase patterns and duplicate financial documents for fraud detection

Considerations

These considerations implement the pillars of the Azure Well-Architected Framework, which is a set of guiding tenets that can be used to improve the quality of a workload. For more information, see Microsoft Azure Well-Architected Framework.

Keep these points in mind when you use this solution.

Availability

The availability of the architecture depends on the Azure services that make up the solution:

- Azure AI Document Intelligence is part of Azure AI services. For this service's availability guarantee, see Service-level agreement (SLA) for Azure AI services .
- AI Language is part of Azure AI services. For the availability guarantee for these services, see SLA for Azure AI services .
- Azure Cosmos DB provides high availability by maintaining four replicas of data within each region and by replicating data across regions. The exact availability guarantee depends on whether you replicate within a single region or across multiple regions. For more information, see Achieve high availability with Azure Cosmos DB.
- Blob Storage offers redundancy options that help ensure high availability. You can use either of these approaches to replicate data three times in a primary region:
  - At a single physical location for locally redundant storage (LRS).
  - Across three availability zones that use differing availability parameters. For more information, see Durability and availability parameters. This option works best for applications that require high availability.
- For the availability guarantees of other Azure services in the solution, see these resources:
  - SLA for App Service
  - SLA for Azure Functions
  - SLA for Application Gateway
  - SLA for Azure Kubernetes Service (AKS)

Scalability

- App Service can automatically scale out and in as the application load varies. For more information, see Create an autoscale setting for Azure resources based on performance data or a schedule.
- Azure Functions can scale automatically or manually. The hosting plan that you choose determines the scaling behavior of your function apps. For more information, see Azure Functions hosting options.
- By default, Azure AI Document Intelligence supports 15 concurrent requests per second. You can increase this value by creating an Azure Support ticket with a quota

increase request.

- For custom models that you host as web services on AKS, azureml-fe automatically scales as needed. This front-end component routes incoming inference requests to deployed services.

- For batch inferencing, Machine Learning creates a compute cluster on demand that scales automatically. For more information, see Tutorial: Build an Azure Machine Learning pipeline for batch scoring. Machine Learning uses the ParellelRunStep class to run the inferencing jobs in parallel.

- For AI Language, data and rate limits apply. For more information, see these resources:
  - How to use named entity recognition (NER)
  - How to detect and redact personal information
  - How to use sentiment analysis and opinion mining
  - How to use Text Analytics for health

Security

Security provides assurances against deliberate attacks and the abuse of your valuable data and systems. For more information, see Overview of the security pillar.

- Azure Web Application Firewall helps protect your application from common vulnerabilities. This Application Gateway option uses Open Worldwide Application Security Project (OWASP) rules to prevent attacks like cross-site scripting, session hijacks, and other exploits.

- To improve App Service security, consider these options:
  - App Service can access resources in Azure Virtual Network through virtual network integration.
  - You can use App Service in an App Service Environment, which you deploy to a dedicated virtual network. This approach helps to isolate the connectivity between App Service and other resources in the virtual network.

  For more information, see Security in Azure App Service.

- Blob Storage and Azure Cosmos DB encrypt data at rest. You can secure these services by using service endpoints or private endpoints.

- Azure Functions supports virtual network integration. By using this functionality, function apps can access resources inside a virtual network. For more information, see [Azure Functions networking options][Azure Functions networking options].

- You can configure Azure AI Document Intelligence and AI Language for access from specific virtual networks or from private endpoints. These services encrypt data at rest. You can use subscription keys, tokens, or Microsoft Entra ID to authenticate requests to these services. For more information, see Authenticate requests to Azure AI services.

- Machine Learning offers many levels of security:
  - Workspace authentication provides identity and access management.
  - You can use authorization to manage access to the workspace.
  - By securing workspace resources, you can improve network security.
  - You can use Transport Layer Security (TLS) to secure web services that you deploy through Machine Learning.
  - To protect data, you can change the access keys for Azure Storage accounts that Machine Learning uses.

Resiliency

- The solution's resiliency depends on the failure modes of individual services like App Service, Functions, Azure Cosmos DB, Storage, and Application Gateway. For more information, see Resiliency checklist for specific Azure services.

- You can make Azure AI Document Intelligence resilient. Possibilities include designing it to fail over to another region and splitting the workload into two or more regions. For more information, see Back up and recover your Azure AI Document Intelligence models.

- Machine Learning services depend on many Azure services. To provide resiliency, you need to configure each service to be resilient. For more information, see Failover for business continuity and disaster recovery.

Cost optimization

Cost optimization is about looking at ways to reduce unnecessary expenses and improve operational efficiencies. For more information, see Overview of the cost optimization pillar.

The cost of implementing this solution depends on which components you use and which options you choose for each component.

Many factors can affect the price of each component:

- The number of documents that you process
- The number of concurrent requests that your application receives
- The size of the data that you store after processing
- Your deployment region

These resources provide information on component pricing options:

- AI Document Intelligence pricing
- App Service pricing
- Azure Functions pricing
- Application Gateway pricing
- Azure Blob Storage pricing
- Azure Cosmos DB pricing
- Language Service pricing
- Azure Machine Learning pricing

After deciding on a pricing tier for each component, use the Azure Pricing calculator to estimate the solution cost.

Contributors

This article is maintained by Microsoft. It was originally written by the following contributors.

Principal author:

- Jyotsna Ravi | Senior Customer Engineer

Next steps

- What is AI Document Intelligence?
- Use Azure AI Document Intelligence SDKs or REST API
- What is AI Language?
- What is Azure Machine Learning?
- Introduction to Azure Functions
- How to configure Azure Functions with a virtual network

- What is Azure Application Gateway?
- What is Azure Web Application Firewall on Azure Application Gateway?
- Tutorial: How to access on-premises SQL Server from Data Factory Managed virtual network using Private Endpoint
- Azure Storage documentation

Related resources

- Extract text from objects using Power Automate and AI Builder

Feedback

Was this page helpful? Yes No