The IBM Advantage for Conversation Cloud Architecture

Table of Contents

Executive overview

An artificial intelligence (AI) business is an organization that creates knowledge from data to expand expertise, continually learns, and adapts to predict the needs of the marketplace. AI systems transform how organizations think, act, and operate in the future with technologies that use natural language, hypothesis generation, and evidence-based learning.

The AI architecture discussed in this document is IBM's approach to describe the flows and relationships between business capabilities and architectural components for AI applications that use cloud computing infrastructure, platforms, or services. The elements of this architecture are used to instantiate a conversation system by using the IBM Cloud™ platform, IBM Cloud Pak for Data, or both, on premises or on any other cloud platform.

Before you read this paper, familiarize yourself with the core AI concepts in the AI glossary.

Critical success factors for AI systems

For an AI system to be successful, you must have the following factors in your environment.

The right cloud platform

To create an efficient AI environment, your enterprise-grade cloud platform must be built on a data-first architecture that gives you the choice of using a public, private, or proprietary hybrid architecture. The cloud platform must be user-friendly and created with scalability and resiliency in mind.

The IBM Cloud platform provides several benefits for enterprises:

- Control over where the customer's data resides
- A level of security that enables the secure movement of content from other cloud providers and customer's data centers into the IBM Cloud platform
- Capabilities to encrypt and store data securely
- Capabilities for secure access of information and systems

IBM Cloud is an industrialized cloud that enables integration between data and applications and also between public, private, and proprietary clouds. Additionally, IBM Cloud is an industry-centric cloud, offering capabilities that are designed for industry-specific data or content and regulations. A trained conversation system provides insights for decision-making and can use the IBM Cloud platform services to act based on insights from AI systems. The IBM Cloud platform provides over 120 services and includes IBM Watson® APIs, services, and software that can help you enable your business.

New forms of data to gain insight

Gaining insight from invisible data, which is sometimes referred to as *dark* or *unstructured* data, such as music, literature, pictures, and videos, can improve your decision-making. An enterprise's unstructured data is proprietary, and with the right tools, the enterprise can harness insights from that data. But

sometimes you can't make informed content decisions based only on your unstructured data. You need to use external sources to supplement your data.

IBM Cloud offers access to content like weather data and security insights that can enhance your decision-making. Other content providers can also supplement content.

Robust ecosystem

Your business's ecosystem plays a major role in the successful transition to an AI business. IBM Watson services can be strengthened by content that is captured in the peer cloud of companies that instrumented the physical world to create a robust ecosystem for solving business challenges.

You can use Watson services with ecosystem partners to gain insights into data in these ways:

- Content that is captured by car manufacturers can be used with Watson services to provide car-related information to drivers and for self-driving cars.
- Content that is captured by health-monitoring devices, such as blood sugar monitoring devices, can be used to give recommendations to doctors about changes in medication or to remind patients to take medication at the right time.

Amplification of human cognition

AI systems should amplify human intelligence. Humans train IBM's AI systems to amplify human cognition. The systems are designed not to replace a human's cognitive capabilities, but to enhance them.

Application of AI insights

When you gather AI insights, you need to apply those insights to improve business processes, to make more informed business plans, and to optimize your business operations. IBM's cloud platform, content, and the ecosystem offer comprehensive insights that can help you make better decisions.

IBM Cloud customer AI reference architecture

As shown in Figure 1, the IBM AI reference architecture can be categorized into 3 broad capabilities:

1. **Watson Assistant**: IBM's AI conversation capabilities are trained to assist in decision-making by using natural language conversation. In situations where there is a conversation or a dialog, IBM Watson Assistant service offers an intent-based understanding and a conversation model that is driven by dialog that you can use to determine the best course of action. To understand how those capabilities are realized, see the Conversation Reference Architecture.
2. **Discovery:** IBM's AI discovery capabilities ingest and enrich information, annotate the information that is stored in multiple documents, and prepare corpus for discovering insights to assist in better decision-making. For more information on how these capabilities are realized, see the Discovery Reference Architecture.

3. **Extend (AI APIs):** You can extend IBM's conversation and discovery capabilities with AI services that take broad or unstructured data and create meaningful, actionable, and valuable information for users, and that can be domain specific. By using a variety of capabilities such as Watson Speech to Text, Text to Speech, Tone Analyzer, Visual Recognition, Natural Language Understanding, and Personality Insights, businesses can turn previously "dark data", such as contact center recordings, images, unstructured text, and video, into valuable, actionable insights and assets.

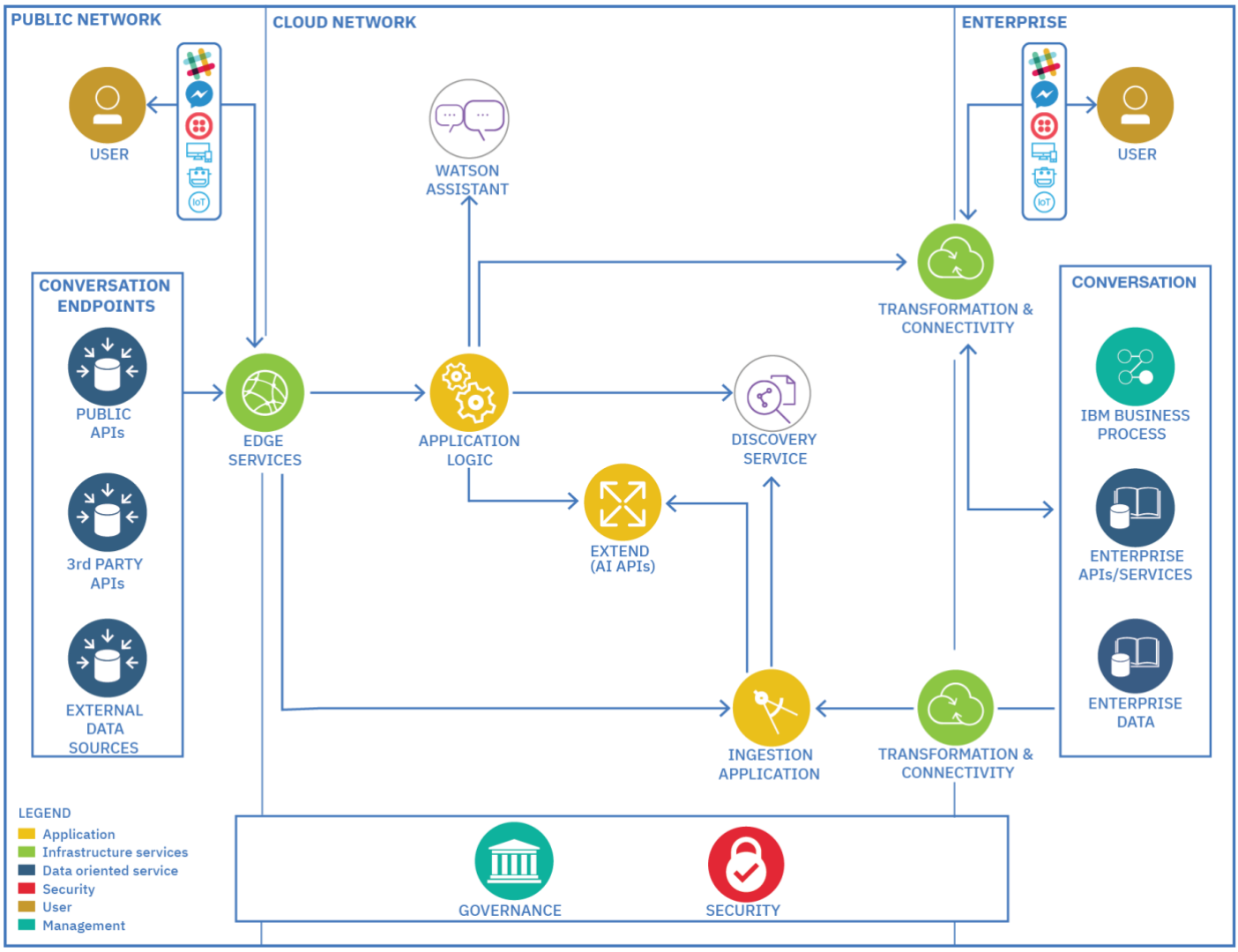


Figure 1. AI reference architecture

Figure 2 shows an example usage of this architecture:

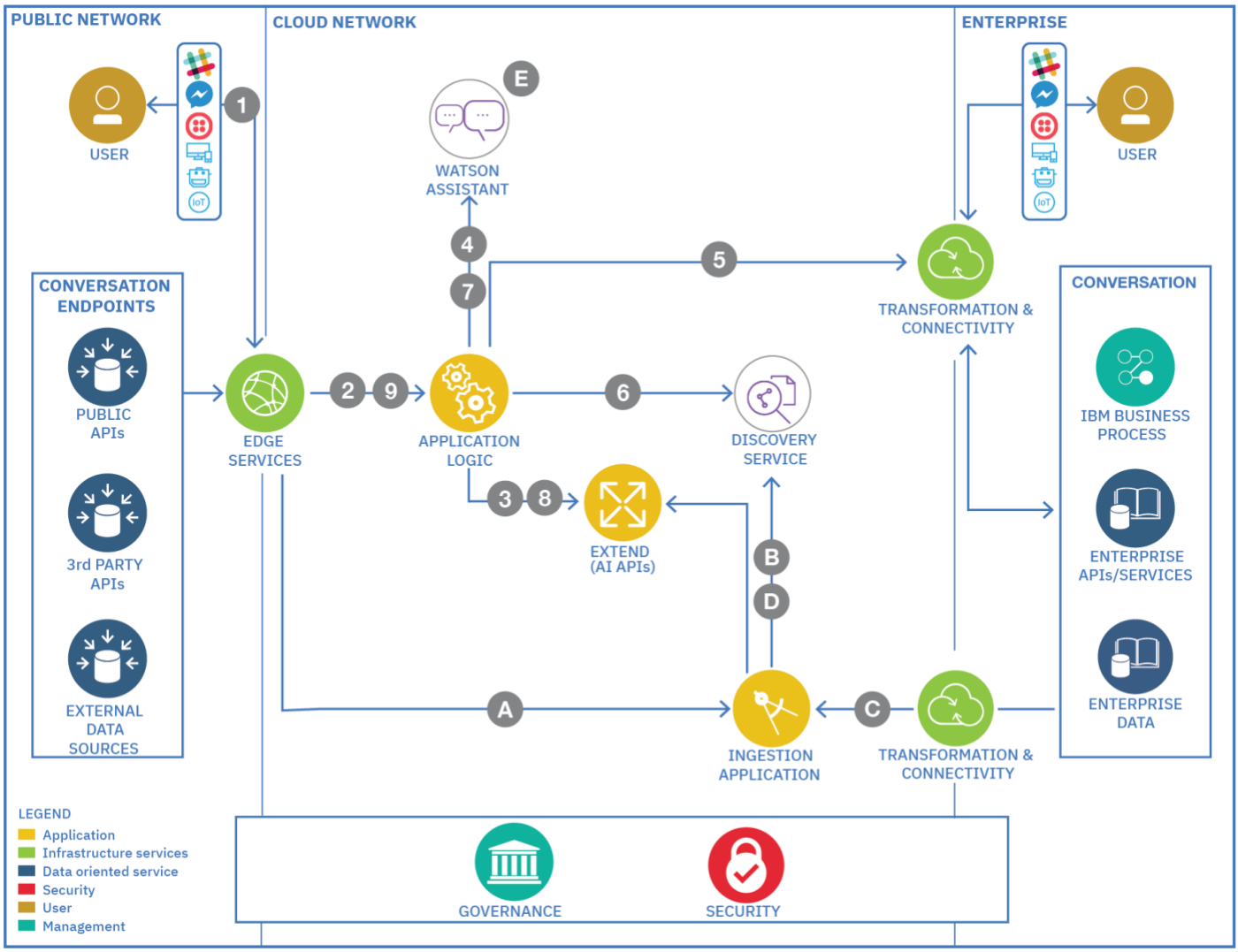


Figure 2. AI reference architecture usage flow

A home improvement store improves its customer service with an AI decision assistant for do-it-yourselfers, professionals, and store associates. The flow is divided into two parts, training and serving.

Training flow

A. The ingestion application crawls customer feedback and comments on social media.

B. The ingestion application uses the discovery APIs to add the content to the collection.

C. The ingestion application crawls product information, product catalogs, descriptions, and product manuals that are stored in the customer data center.

D. The ingestion application uses the discovery APIs to add the data center content to the collection.

E. Subject matter experts (SMEs) train the conversation.

Serving flow

1. A customer accesses the mobile application by using voice to ask what kind of connector is needed for a dishwasher. (The manufacturer does not provide the connector.)
2. Application logic determines that the request is a voice request and invokes the extend speech-to-text component.
3. The extend speech-to-text component converts the voice request to text.
4. Application logic uses the text to check with the Watson Assistant component for a trained response. Watson Assistant is not trained with a specific response on connectors for a dishwasher.
5. Application logic uses the API to determine whether the information is available in the core SAP systems. The SAP system has information about the type of connector but does not have detailed information about how to use the connector to connect the dishwasher to the garbage disposal. The system does have the technical product manual.
6. Application logic checks to learn whether Watson Discovery is trained to get specific information. Discovery is trained with technical product manuals and gives the information to application logic.
7. Application logic uses the Watson Assistant component to get the correct response.
8. Application logic uses extend text-to-speech to translate the response from text to voice.
9. Application logic sends the voice response to the customer's mobile application.

The AI reference architecture positions Watson Assistant, Watson Discovery, and the extend Watson APIs capabilities in relation to each other. This paper covers the planning, preparation, and implementation that is required to enable a conversation system or services, whether through a self-service chatbot or as a call center agent assistant.

You must train IBM AI systems and services for decision assistance and planning as the first step. In the planning phase, business architects identify the knowledge data sources and the relevant content within the data sources to be used. Input from common utterances that are noticed by SMEs are crowdsourced. The collection of this information becomes what is known as the *ground truth*.

After you complete the planning phase, you must prepare the system. In this phase, an engineer prepares the dialog flow, intents, and entities and defines their relationship by using IBM Watson Assistant tools. For definitions of those terms, see the AI glossary.

For decision assistance, you can use trained IBM AI services or systems in any form factor, including mobile, kiosk, car dashboard, web, voice response unit, and others. IBM AI systems learn from continuous interactions and from identified patterns. When the conversation system doesn't have a trained response, it uses the IBM Watson Discovery service. This service examines relevant documents and annotates them with potential response information as a type of knowledge repository.

Because AI systems are part of the IBM Cloud platform and are available on premises or on any other cloud by using IBM Cloud Pak for Data, you can use them to create a conversation system for any industry. The conversation system can use other services of the platform to drive actions that result from the insights it has gained.

The critical success factor for AI depends on having the right cloud platform and actions taken based on insights gained from analytics. IBM Cloud provides capabilities to develop these AI systems of engagement solutions by using services such as analytics, security, Internet of Things (IoT), blockchain for digital transformation foundation and processing, decision management, integration, and collaboration for implementing actions from the insights.

For a thorough discussion and practices, see The IBM Advantage for Implementing the CSCC Cloud Customer Reference Architecture for Internet of Things (IoT).

Creating a conversation-based AI system

The 3 phases for creating a conversation-based AI system are planning, preparing, and implementation (runtime). The following sections show architectures that relate to all 3 phases.

Phase 1: Planning

Figure 3 shows the flow and the sequence of tasks that are needed to plan and design the conversation system for the preparation of ground truth.

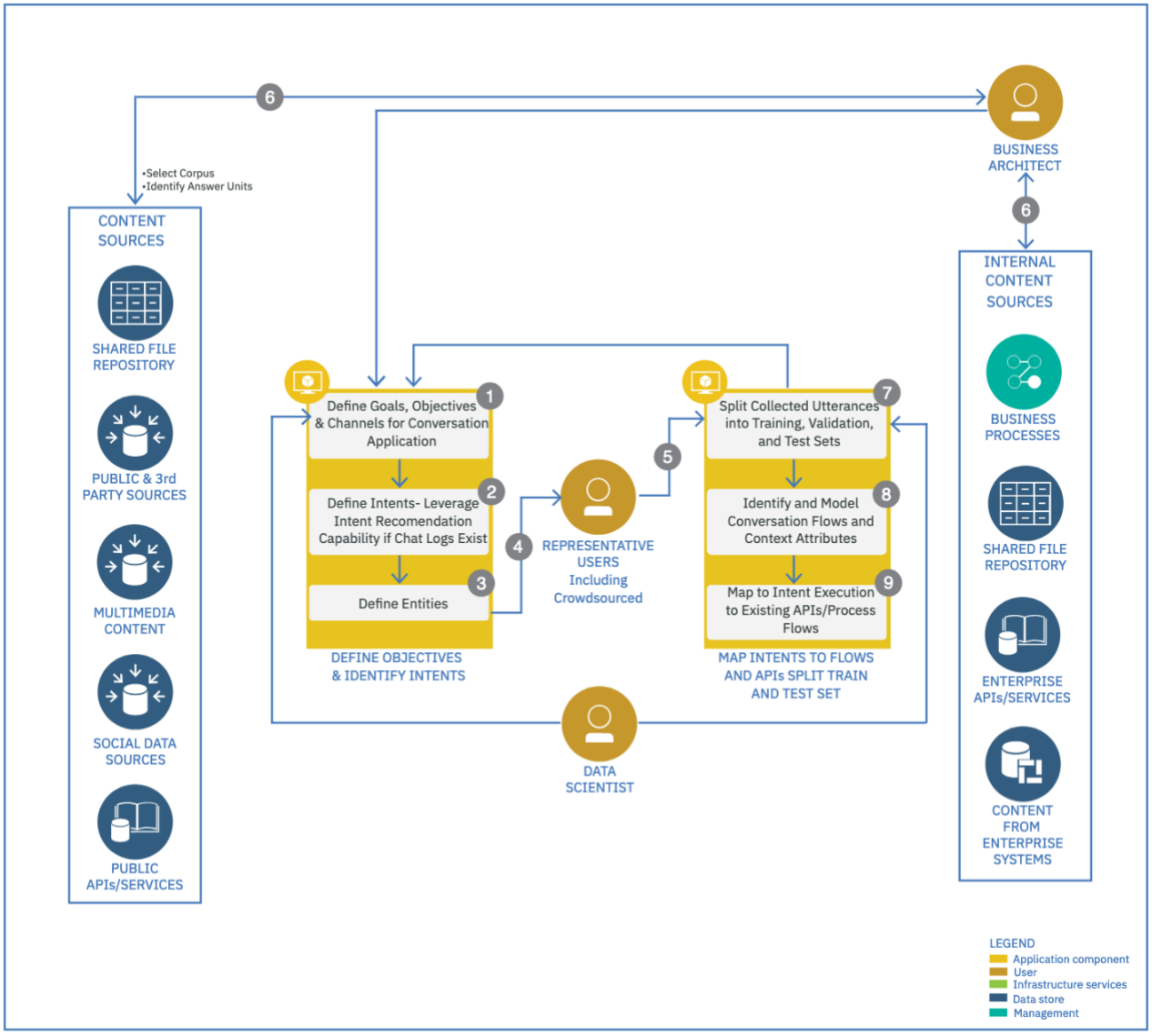


Figure 3. Planning for an AI solution

The diagram shows the 3 personas that are present in the planning phase.

- **Business architect**: This person knows the source of information, whether the source is a training manual, product manual, testing manual, or external, publicly available information. The business architect defines the goals and objectives of the conversational application or bot, including the channels that the application needs to support (web, mobile, social, and others).
- **Data scientist**: The data scientist helps and supports the business architect to understand the right type of information that can be used to train the conversation system. Data scientists have deep knowledge of information that can be used to extract insights.
- **Representative users including crowdsourcing**: These users include resources that have the knowledge and understanding of various terms, utterances, and the specific vernacular that the

conversation system needs to understand. The representative users can be product experts, call center supervisors, scientists, doctors, or engineers. For example, one of the representative users for an appliance manufacturer might be a field technician.

This planning phase involves 2 broad categories of information sources.

- **Internal content sources:** Includes processes and data sources that are within a given enterprise. They typically contain the data that is generated and owned by the enterprise as part of its business operations. The conversation application doesn't automatically ingest the information from these sources. The business architect uses these sources to plan and identify the answer units that form the textual response to the user. Alternatively, these sources can also serve as the process endpoints, which might need to be invoked to fulfill the intent.
    - **Business processes:** These processes are enterprise-level business processes that the chatbot might need to interact with to process and respond to the user's intent.
    - **Enterprise APIs or services:** These sources include APIs or services that might need to be accessed or invoked in response to user queries and responses on the chatbot. For example, the user might ask to place an order, which requires invocation of a certain service endpoint to place the order. Most systems of records involve an API to serve the data that they generate or control.
    - **Shared file repository:** This source includes information that is kept in file systems that are shared between users and locations and are accessible through FTP and other mechanisms.
    - **Content from enterprise systems:** Data from various enterprise systems, including but not limited to catalogs, order or transaction data, and enterprise content management (ECM) repositories.
- **External content sources:**
    - **Shared file repository:** This source includes information that is kept in file systems that are shared between users and locations and are accessible through FTP and other mechanisms.
    - **Public and third-party sources:** Public and third-party sources include information sources that are available for public consumption. This set of information is neither owned by the enterprise nor is generated by the enterprise as part of the business operations. These sources include both public domain data, which is available free of cost, and data that is controlled by other parties. Examples include weather data and domain-specific catalogs that are made available by third-party vendors.
    - **Multimedia content:** This content includes audio, video, or images that are available on the internet.
    - **Social data sources:** This subset of public and third-party sources specifically involves social media such as Twitter, Facebook, and others.
    - **Public API or sources:** These sources include data that is accessible for public

consumption, requiring the invocation of an API.

To plan and design a conversation system for preparation of ground truth, follow these steps. In the first 3 steps, you define objectives and identify intents.

1. The business architect defines the goals and objectives of the conversational application or bot, including the channels that the application needs to support (web, mobile, Twitter, and others).

   The business architect, SME, and data scientist collaborate on the following steps:

2. Identify, define, and model the intents that the conversational application needs to detect from the utterances. You can use the content catalog, which is available as part of Watson Assistant and includes many predefined intents and corresponding user utterance examples across a range of common industries and use cases.

3. Identify, define, and model the entities to detect from the user utterance what will be used to clarify a user's intent.

4. For the identified intents, collect and gather actual user questions, commands, or other utterances from real representative users. Also crowdsource the collection of these utterances to get a better understanding of the input that the conversation application might receive. An excellent source of representative user utterances are actual historical chatlogs. You can use the Watson Assistant intent recommendations feature to speed time to value by ingesting the chat logs and letting Watson Assistant identify the right intents and map the user utterances to those intents.

5. Map the utterances to intents or review and update recommended intents and mapped utterances. Continuously iterate over the utterances collection. Potentially identify new intents and entities that the application might need to handle.

6. Identify the answer units from the corpus (public and private) that must be provided as a textual response to the user. Identify the processes and APIs that might need to be invoked to fulfill the intent. Identify documents to include in a knowledge corpus to expand the bot’s capabilities in addressing user utterances. Use the Watson Discovery Smart Document Understanding capability to train a model for segmenting documents into answer units based on the structure of the documents. Identify the process and APIs that might be needed to be invoked to fulfill the intent.

7. Split the collected and mapped utterance examples into train, test, and evaluation sets. Typically, use a randomized 70-20-10 % split.

8. Identify and model any contextual attributes that are passed to the conversation application or bot from the user application based on which flows and answers might differ.

9. For intents that need to run a series of simple steps, use business process flows and business rules. Identify how to integrate with service desks as demanded by the business operations.

Phase 2: Preparation

The second step in designing a conversations system is to prepare the ground truth for consumption at runtime.

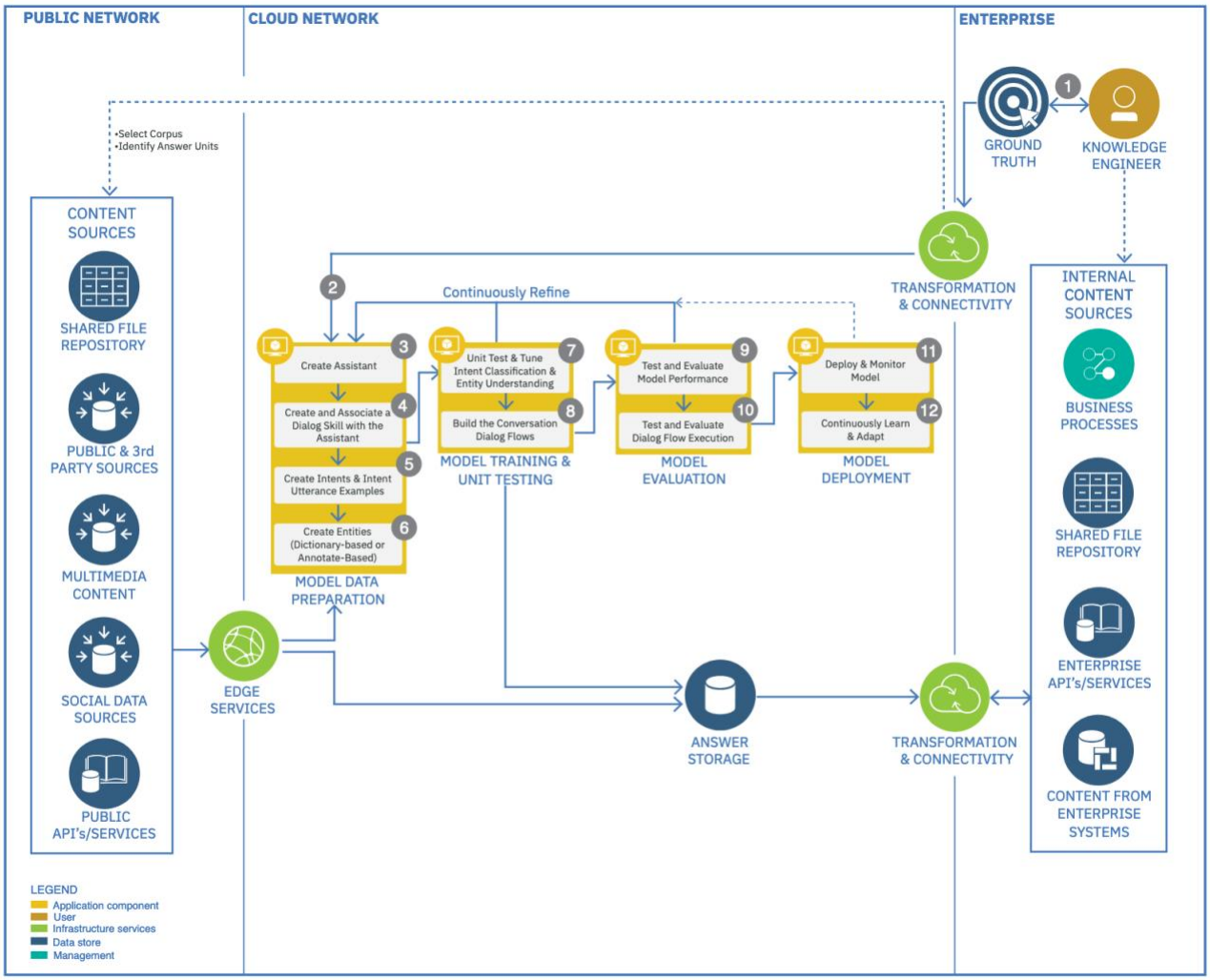


Figure 4. Preparation for an AI solution

This preparation phase involves one persona, the *knowledge engineer*. This person uses the ground truth data and the Watson Assistant service tools to train, test, and configure the Watson Assistant service.

This preparation phase involves 2 broad categories of information sources.

- **Internal content sources:** This includes processes and data sources that are within an enterprise. They typically contain the data that is generated and owned by the enterprise as part of its business

operations. The conversation application doesn't automatically ingest the information from these sources. The business architect uses these sources to plan and identify the answer units that form the textual response to the user. Alternatively, these sources can serve as the process endpoints, which might need to be invoked to fulfill the intent.

  - **Business processes:** These processes are enterprise-level business processes that the chatbot might need to interact with to process and respond to the user's intent.
  - **Enterprise APIs or services:** These APIs or services might need to be accessed or invoked in response to user queries and responses on the chatbot. For example, the user might ask to place an order, which requires invocation of a certain service endpoint to place the order. Most systems of record involve an API to serve the data that they generate or control.
  - **Shared file repository:** This source includes information that is kept in file systems that are shared between users and locations and are accessible through FTP and other mechanisms.
  - **Content from enterprise systems:** This source includes data from various enterprise systems including but not limited to catalogs, order or transaction data, and enterprise content management (ECM) repositories.

- **External content sources:**
  - **Shared file repository:** This source includes information that is kept in file systems that are shared between users and locations and are accessible through FTP and other mechanisms.
  - **Public and third-party sources:** Public and third-party sources include information sources that are available for public consumption. This set of information is neither owned by the enterprise nor is generated by the enterprise as part of the business operations. These sources include both public domain data, which is available free of cost, and data that is controlled by other parties. Examples include weather data and domain-specific catalogs that are made available by third-party vendors.
  - **Multimedia content:** This content includes audio, video, or images that are available on the internet.
  - **Social data sources:** This subset of public and third-party sources specifically involves social media such as Twitter, Facebook, and others.
  - **Public API or sources:** These sources include data that is accessible for public consumption, requiring an invocation of an API.

This architecture lifecycle diagram shows the steps required for the planning, modeling, and training of the ground truth.

1. The cognitive knowledge engineer uses the work that is done by the business architect, including the ground truth that was collected.
2. With the IBM Watson Assistant service that was created in IBM Cloud or in IBM Cloud Pak for Data, the knowledge engineer opens the tools for Watson Assistant.

The business architect, SME, and data scientist collaborate on the following steps.

3. Create an assistant and give it a unique name and description. An assistant is an AI bot or AI virtual assistant. You add skills to the assistant that enable it to interact with your customers in useful ways. Two types of skills can be added to the assistant:
    - Dialog skill: Uses Watson natural language processing and machine learning technologies to understand user questions and requests and respond to them with adequate answers as authored by knowledge engineer.
    - Search skill: For a given user query, uses the IBM Watson Discovery service to search a data source of your knowledge corpus and return an answer.
4. Add a dialog skill to your assistant, giving it a unique name and description, including the language. A dialog skill is a container for the artifacts that define the flow of a conversation that your assistant can have with your customers.
5. Create intents, and for each intent, define representative utterance examples that were collected by using crowdsourcing or chatlogs. The intents and intent utterance examples can be uploaded as a .csv file that is defined by using Watson Assistant.
6. Create the entities either by using the dictionary-based method or by using the annotation-based method.
    - Dictionary-based method: Entities are defined by using synonyms, regular expression patterns, or both. The dictionary method also includes several useful predefined system entities that are part of Watson Assistant, such as date, time, location, and person.
    - Annotation-based method: Entities are defined based on the annotated terms and the context in which they appear. You annotate the actual occurrence of terms in the user utterances that are defined for intents.
7. Unit-test the model by providing sample utterances to check whether the correct intent and entities are being detected. If the detected intents and entities aren’t correct, update the training data, such as intent utterance examples and entity values, synonyms, and patterns, with the correct mapping. Continuously iterate and repeat as necessary and fine-tune the intent examples and entities.
8. Model the dialog that is made up of nodes, which define interactions in the conversation. At each node, define the textual or rich response that the bot must communicate back to the user. Use slots to gather information from users, handle digressions, and define node purposes to support intent disambiguation. As necessary, use the context variables that are identified and make programmatic calls from dialog nodes as needed.
9. Test the model performance by using external scripts with the test and evaluation data sets that were randomly split from the ground truth. Evaluate and analyze the accuracy, F1 score, precision, and recall of the intent classification model.

10. Test and evaluate the dialog flow implementation, especially variations with conditions, context variables, and various entity values. Use external scripts to evaluate coverage and effectiveness of the bot.
    - *Coverage* is the portion of total user messages that your assistant is attempting to respond to.
    - *Effectiveness* refers to how well your assistant is handling the conversations that it is attempting to respond to.
11. Deploy the trained model.
12. Enable continuous monitoring of the deployed AI model and collection of real user utterances. Continuously evaluate the model on collected user utterances to validate model performance and retrain the model when performance starts to degrade.

Phase 3: Implementation (runtime)

This runtime architecture showcases the components that are involved in the use of a trained and deployed conversation system. In the previous architectures, you planned and prepared or trained the conversation system. This architecture flow shows how those parts work together in a conversation system.

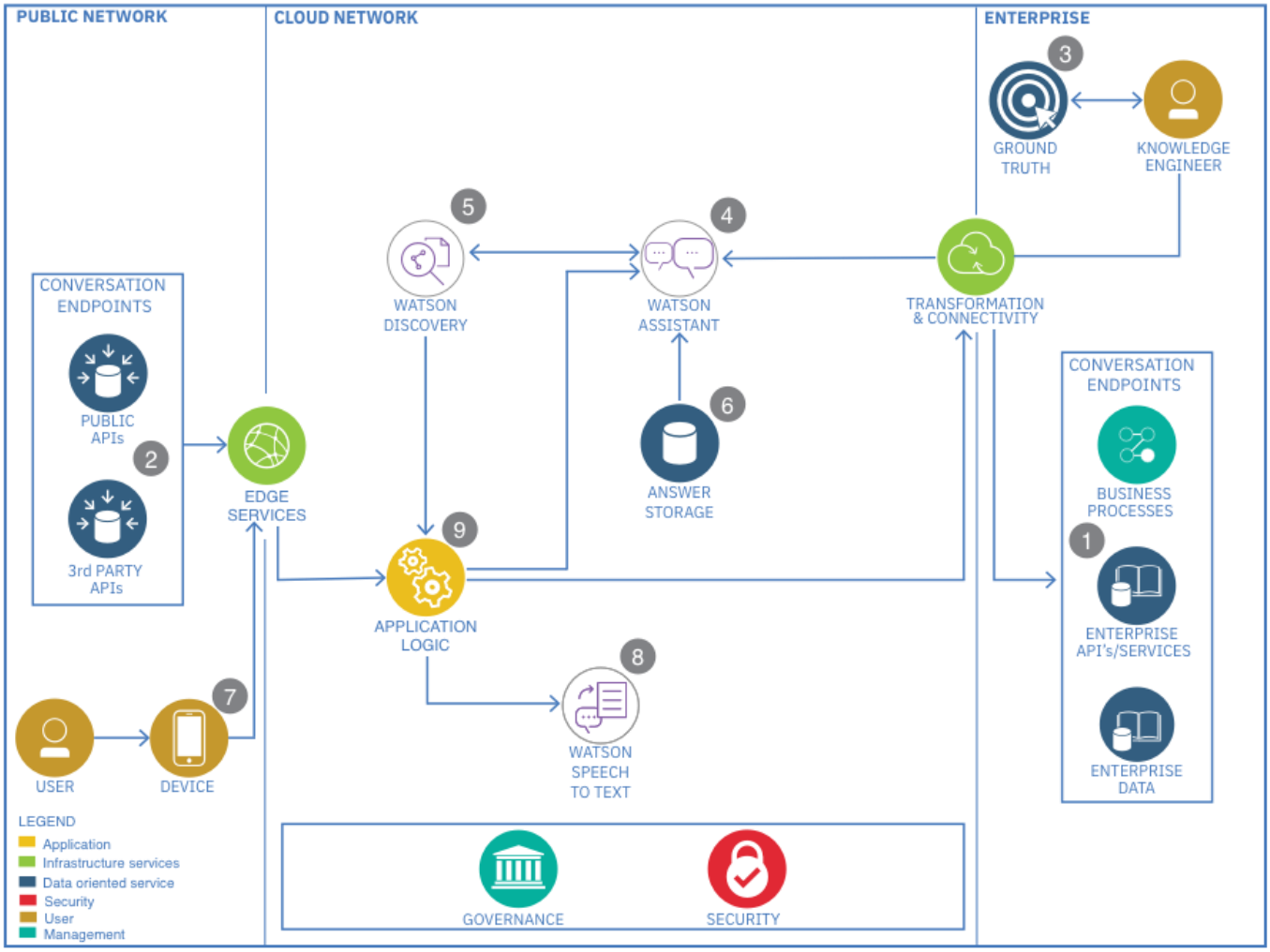


Figure 5. Conversation runtime solution

This example shows how a customer uses a trained AI system and how the system interacts with other components of the cloud platform. It also shows data that is made available from the public source or in the customer's enterprise, either in the form of raw data or in the form of an API that is used to train the conversation. Credentials and permissions of users who access the system are verified. The data that is moved into the enterprise is encrypted whether at rest or in motion.

This diagram has two parts. The first part is the description of the components of the architecture. The second part is the runtime flow that shows the interaction when a customer uses a chatbot.

Component descriptions: Endpoints, ground truth, and Watson Assistant service

1. **Conversation endpoints in the enterprise network**: This component refers to unstructured content or the information that is stored in the enterprise network, including but not limited to FAQ documents, historical customer conversation records, operational manuals, and customer

feedback. This content, along with knowledge from the institutional SMEs, is used to prepare and train the conversation system. The enterprise APIs that can be integrated as part of the overall conversation system from application logic are also included.

2. **Conversation endpoints in the public network**: Enterprise information is strengthened with content or information that includes Wikipedia, news articles, and financial reports. You can use this information to prepare and train the conversation service. Also included are third-party APIs, such as IBM Weather Data that can be integrated as part of the overall conversation system from application logic.
3. **Ground truth**: These sets of artifacts are used to train Watson Assistant. Content from conversation endpoints (both public and enterprise), example utterances, representative SMEs, and crowdsourced inputs are used to model the intents, context attributes entities, and dialog responses and to train the service. Ground truth is typically split into training, testing, and evaluation data.
4. **Watson Assistant (trained and deployed)**: The knowledge engineer uses the ground truth that was collected in step 3 and the conversation API tools to populate the intents, entities, dialog flows, and context for training the conversation API.
5. **Discovery**: Discovery finds the relevant passages in the corpus and answers open-ended questions. It's often used for knowledge expansion or complex scenarios. The knowledge engineer ingests and potentially annotates unstructured documents, including manuals and training handbooks, and trains a ranked model to rank the returned passages for an utterance.
6. **Answer storage**: Answers can be maintained in an external answer storage format. These answers are provided to the user after the intent and entities are understood.

User interaction: Runtime flow

7. **Device**: A customer uses a mobile device or another form factor that has an application with an embedded chatbot to start a conversation with the AI system. Through the application logic component, the AI system returns the requested information to the device on which the conversation occurs.
8. **Speech to text**: For voice-based requests, application logic uses the speech-to-text service to convert the spoken utterances into text before it passes the request to the conversation API.
9. **Application logic**: This logic might be a Node.js or any other runtime application. It first passes the natural language utterance (request) to the Watson Assistant service. When the response from Watson Assistant service is received, application logic checks the level of confidence. If the level of confidence is above a set threshold, it returns the response to the user. The application logic might need to invoke APIs to fetch the answers needed to fulfill the intent detected from the utterance. If the confidence levels are low, application logic checks for possible answers by using a discovery service. It returns the response to the device.
10. **Transformation and connectivity**: Application logic can strengthen the response by supplementing structured data, such as user profile, past orders, and policy information,

from the enterprise network. The connection to the enterprise network is established through the transformation and connectivity component. Results are delivered to users and applications by using transformation and connectivity components that provide secure messaging and translations to and from systems of engagement, enterprise data, and enterprise applications.

Components

These individual components make up the AI architecture.

Public network components

The public network contains elements that exist in the internet: data sources and APIs, users, and the edge services that are needed to access the provider cloud or enterprise network. The public network includes the conversation endpoints.

User

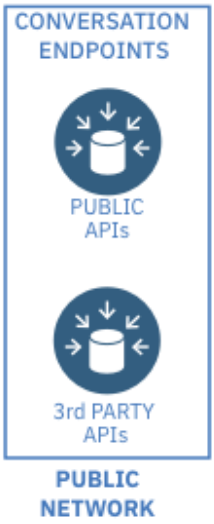
A user is a customer who uses their device to access the conversation system on the cloud provider platform or enterprise network.

Device

A user uses a mobile device or other form factor that has an application with an embedded chatbot to start a conversation with the AI system.

Conversation endpoint in the public network

Enterprise information is strengthened with content or information, including Wikipedia, news articles, and financial reports, that is used to prepare and train the conversation service. Also included are third-party APIs, such as IBM Weather Data, which can be integrated as part of the overall conversation system from application logic.



Cloud network components

Cloud network components include edge services and capabilities.

Edge services

Edge services are distinct network components that are a part of the IBM Cloud platform. These services allow data to flow safely from the internet into the IBM provider cloud and into the enterprise. Edge services also support user applications. The key capabilities in this domain are as follows:

- **Domain name system server**: Resolves the URL for a particular web resource to the IP address of the system or service that can deliver that resource.
- **Content delivery networks (CDN)**: Supports user applications by providing geographically distributed systems of servers that are deployed to minimize the response time for serving resources to geographically distributed users. This distribution ensures that content is highly available and is provided to users with minimum latency. Which servers are engaged depends on server proximity to the user and where the content is stored or cached.

- **Firewall**: Controls communication access to or from a system, permitting only traffic that meets a set of policies to proceed and blocking any traffic that doesn't meet the policies. You can implement firewalls as separate dedicated hardware, as a component in other networking hardware such as a load balancer or router, or as integral software to an operating system.

Load balancers: Provide distribution of network or application traffic across many resources, such as computers, processors, storage, or network links, to maximize throughput, minimize response time, increase capacity, and increase the reliability of applications. Load balancers can balance loads locally and globally. Load balancers must be highly available without a single point of failure. Load balancers are sometimes integrated as part of the provider cloud analytical system components such

as stream processing, data integration, and repositories.

- **MFT (managed file transfer) gateway:** A multi-protocol gateway (AS2, AS4, sftp, ftps, C:D) into and out of the organization that provides security (encryption and decryption), virus checks, data loss prevention, certificate and key management, monitoring, and auditing.

IBM capabilities on edge services

IBM Cloud platform supports various services for DNS, firewalls, load balancing, and CDN. IBM Security Network Protection (IBM XGS) is an intrusion prevention system (IPS) that you can use to monitor network traffic and to provide protection from hidden security vulnerabilities. IBM DataPower® provides load balancing and SSL termination. It can quickly secure, integrate, control, and optimize access to a range of workloads through a single, extensible, DMZ-ready gateway.

Watson Assistant trained and deployed

The knowledge engineer uses the ground truth and the conversation API tools to populate the intents, entities, dialog flows, and context for training the conversation API by using the workspace that is provided by the IBM Watson Assistant service.

You can train and deploy IBM Watson Assistant service to add a chatbot to your website that automatically responds to customers' frequently asked questions. Watson Assistant service can also be used to integrate the assistant you trained to your website with a simple addition of a few HTML lines of code.

Watson Assistant offers several integration options with popular platforms such as Facebook Messenger, Slack, and as web-hosted chat widget. You can also develop custom integration points with other messaging platform chatbots that interact instantly with channel users and allow customers to control your mobile application by using natural language virtual agents. Watson Assistant also supports multiple service desk integration options, such as Intercom and Zendesk. More integrations are being developed.

Watson Discovery service

You can use this discovery service for knowledge expansion or complex scenarios to find the relevant passages in the corpus and answer more open-ended questions. The knowledge engineer ingests and potentially annotates unstructured documents, including manuals and training handbooks, and trains a ranked model to rank the returned passages for an utterance. Given the important of being able to handle a breadth of questions that might be asked, integrating Watson Assistant with Watson Discovery is simplified with a Search Skill capability that can be added to an assistant.

Watson Discovery service helps users find the most relevant information for their query by using a combination of search and machine learning algorithms to detect "signals" in the data.

You load your data into the service, train a machine learning model based on known relevant results, and use this model to provide improved results to your users based on their question or query.

Answer storage

You can maintain answers in an external answer storage format. These answers are returned to the user after the intent and entities are understood.

IBM Watson Assistant service maintains its own answer storage that is tightly integrated with the service.

Speech to text

For voice-based requests, application logic uses the Speech to Text service to convert the spoken utterances into text before it passes the request to the conversation API.

You can use IBM Watson Speech to Text service anywhere that you need to bridge the gap between the spoken word and its written form. This service uses machine intelligence to combine information about grammar and language structure with knowledge of the composition of an audio signal to generate an accurate transcription. It uses IBM's speech recognition capabilities to convert speech in multiple languages into text.

The transcription of incoming audio is continuously sent back to the client with minimal delay and is corrected as more speech is heard. The service also includes the ability to detect one or more keywords in the audio stream. You can access the service by using a WebSocket connection or REST API.

Application logic

Application logic, which might be a Node.js application, first passes the natural language utterance (request) to the conversation service. When it receives the response from the conversation service, application logic checks the level of confidence. If the level of confidence is above a set threshold, it returns the response to the user. The application logic might need to invoke APIs to fetch the answers that are needed to fulfill the intent detected from the utterance. If the confidence levels are low, application logic checks for possible answers by using a discovery service.

IBM Cloud platform provides containers that are portable and allow for consistent delivery of your application without the need to manage the underlying operating systems. IBM Cloud also provides Cloud Foundry services so that you can deploy your application without managing the underlying infrastructure.

The applications that are built for IBM Cloud Foundry-based services such as Node.js or container-based deployments such as Liberty for Java™ are built to orchestrate, choreograph, or enrich decision management or to produce actions that are based on AI or analytical insights. These cloud platform services are essential for the success of AI systems.

Transformation and connectivity

Application logic can strengthen the response by supplementing structured data, such as user profile, past orders, and policy information, from the enterprise network. The connection to the enterprise network is established through the transformation and connectivity component.

In IBM Cloud, you can use the IBM Integration Bus container to integrate applications and infrastructures that are deployed in multiple clouds or in legacy or core applications that are deployed in customers' traditional data centers.

IBM API Connect® is a comprehensive API lifecycle solution that enables the automated creation of APIs, simple discovery of systems of records, self-service access for internal and third-party developers, and built-in security and governance. By using automated, model-driven tools, you can create APIs and microservices that are based on Node.js and Java runtimes—all managed from a single unified console. Ensure secure and controlled access to the APIs by using a rich set of enforced policies. Drive innovation and engage with the developer community through the self-service developer portal. IBM API Connect provides streamlined control across the API lifecycle and helps you to gain deep insights around API consumption from its built-in analytics.

The Secure Gateway service brings hybrid integration capabilities to your IBM Cloud environment. It provides secure connectivity from IBM Cloud to other applications and data sources that run on premises or on other clouds. A remote client is provided to enable secure connectivity.

Enterprise network components

Enterprise network components include conversation endpoints and ground truth.

Conversation endpoints in the enterprise network

The conversation endpoints in an enterprise network encompass unstructured content and information that is stored in the enterprise network, including but not limited to FAQ documents, historical customer conversations records, operational manuals, and customer feedback.

This content, along with knowledge from the institutional SMEs, is used to prepare and train the conversational system. Also included are the enterprise APIs that can be integrated as part of the overall conversation system from application logic.

IBM Watson Assistant combines a number of AI techniques to help you build and train a bot, defining intents and entities and crafting dialog to simulate conversation.

Ground truth

Ground truth is the set of artifacts that is used to train Watson Assistant. Conversation endpoints (both public and enterprise), example utterances, and representative SME crowdsourced inputs are used to model the intents, context attributes entities, and dialog responses to train the conversation service. Ground truth is typically split into training, test, and evaluation data. The ground truth in this instance is documented by using any productivity tools.

Security architecture: Retail scenario

Security is a critical aspect of the AI reference architecture. This diagram shows how the flow of information is secured, including the movement of information, authentication, access control, and auditing of all security requests for conversation.

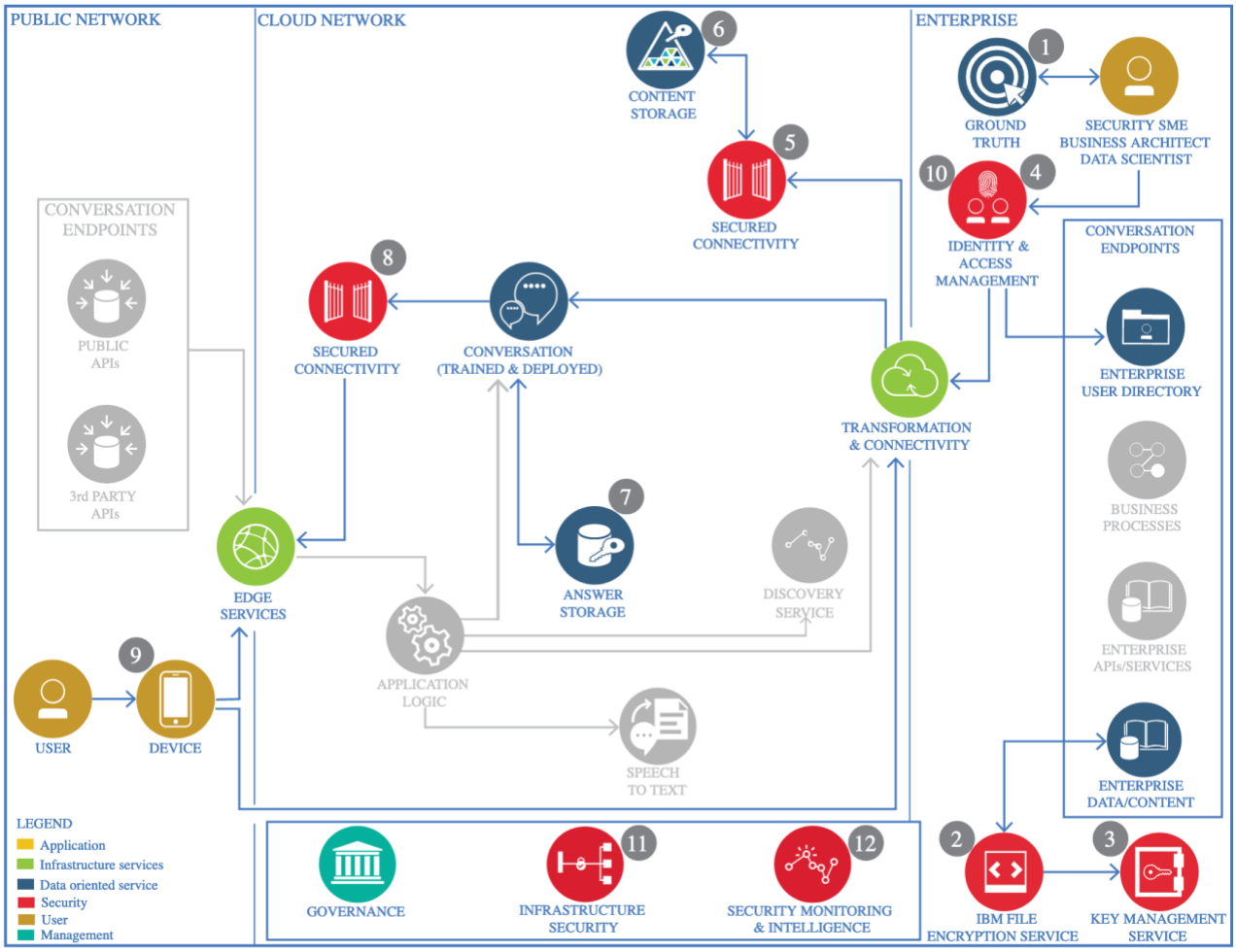


Figure 6. Conversation flow in a retail scenario

In this scenario, a fashion design and retail enterprise stores mood boards, color boards, and images in its enterprise. These fashion designs are trade secrets. However, the data must be moved to IBM Cloud to use IBM Watson services. The retailer also needs to use data that it doesn't have, such as social media data, to understand customer interests. IBM built a fashion designer application with AI for the customer. The customer wants to ensure that the content that is stored in the data center is secured and encrypted when it is in the data center, while it is moving to the IBM Cloud, and when it is accessed by Watson APIs. The fashion designer application must have proper access control.

1. The business architect, security SME, and data scientist identify the data sources and define the security policies for encryption of data in motion and at rest.

2. A system administrator uses the enterprise's own encryption algorithm and encrypts the fashion design mood boards and color board images.

3. The key that is used to encrypt the image is stored in the hardware key vault in the customer's data center. Alternatively, the customer can choose to use the hardware key vault that is provided by the

cloud vendor.

4. Before the content is transferred, the identity of the system administrator is authenticated and access rights are verified. This security is achieved by integration with enterprise user directories and entitlement stores to authenticate the user, validate permissions, and determine access levels to content.

5. VPN and edge services that are provided by the cloud provider secure connectivity from the enterprise network to the cloud. This group of services handles the request and gets it to the right destination securely.

6. After the system administrator is authenticated, the encrypted content is moved to the cloud provider data center to be cleansed. All content is ingested, cleansed, converted, and normalized as necessary. The cleansed, converted, and normalized data is stored in the cloud provider content store. The stored content is encrypted by using the native encryption that is provided by the content store.

7. The answer storage uses its own encryption algorithm to encrypt the content.

8. Data from social media uses transport layer security (TLS) to bring the data for the creation of the corpus. A single sign-on, trusted identity, or both are established between the cloud provider and the social media content provider.

9. A fashion designer uses the application on their laptop.

10. The fashion designer’s user identity is authenticated, and their access rights are verified. The designer has access to the corpus data for decision-making.

11. The infrastructure (compute, network, and storage) that enables the conversation architecture must be secured. Infrastructure security protects against network-level threats and attacks with intrusion prevention and detection, including those that are tunneling through encrypted web transactions and web applications that are deployed within the system. Infrastructure security protects virtual servers and applications against breaches. It tracks and supports regulatory compliance needs for the infrastructure, middleware, and workload.

12. Even after you secure access and provide encryption to data at rest and data in motion for the AI architecture, vulnerabilities can exist. Security intelligence monitoring fills that gap by taking a big data and analytics approach. Security monitoring and intelligence provide security and visibility into cloud infrastructures, data, and applications by collecting and analyzing logs in real time across the various components and services in the cloud. This monitoring provides real-time risk analysis of the workloads that are hosted on the cloud against the myriad of known vulnerabilities and alerts against zero-day attacks.

Although it isn’t shown in the diagram, all security authentication, access, and movement of information is logged for auditing.

IBM capabilities for security in a conversation system

This table maps the IBM capabilities and services to the components in the architecture.

Component	Definition	IBM products
Edge services	Edge services include the services that are needed to allow data to flow safely from the internet.	DNS, CDN, firewall, load balancer
Transformation and connectivity	Transformation and connectivity includes scalable messaging, transformation, and secure connectivity.	IBM Integration Bus container, IBM DataPower, IBM API Connect, IBM Secure Gateway
Conversation service	With the IBM Watson Assistant service, you can create virtual agents and bots that combine machine learning, natural language understanding, and integrated dialog tools to provide automated customer conversations. Watson Assistant provides a graphical environment to create natural conversation flows between your applications and your users.	IBM Watson Assistant service
Key management service	A cloud-based security service that provides key lifecycle management (key creation, usage, and deletion) for encryption keys that are used in IBM Cloud services or customer-built applications, with a "root of trust" that is backed by a hardware security module (HSM).	IBM Key Protect

File encryption service	Safeguards data even when network protection fails. It has built-in and external key management, giving customers control over their encryption keys.	IBM Multi-Cloud Data Encryption
Secured connectivity	Services that offer security connectivity, such as VPN or TLS-based encryption that ensures secure transmission of data from enterprise to cloud or vice versa. Social media providers use TLS-based security to perform a single sign-on to access content.	VPN providers
Identity and access management	Identifies and authenticates the user. Determines access levels by using an enterprise security directory such as LDAP.	IBM Security Access Manager
Security monitoring and intelligence	Provides security and visibility into cloud infrastructures, data, and applications by collecting and analyzing logs in real time across the various components and services in the cloud. Also provides real-time risk analysis of the workloads that are hosted on the cloud against the myriad of known vulnerabilities and alerts against zero-day attacks.	IBM Security QRadar® SIEM
Infrastructure security	Protects against network-level threats and attacks with intrusion prevention and detection, including those that tunnel through encrypted web transactions and web applications that are deployed within the system.	IBM Security Server Protection, IBM Security SiteProtector System

Rental car company bot scenario

In this scenario, a rental car company wants to have a conversation with its customers. When the customer arrives at the airport, the bot provides the customer with information about the weather, restaurants, and sightseeing information, in addition to car selection choices and information about the preassigned car. The following diagram shows the runtime of this bot.

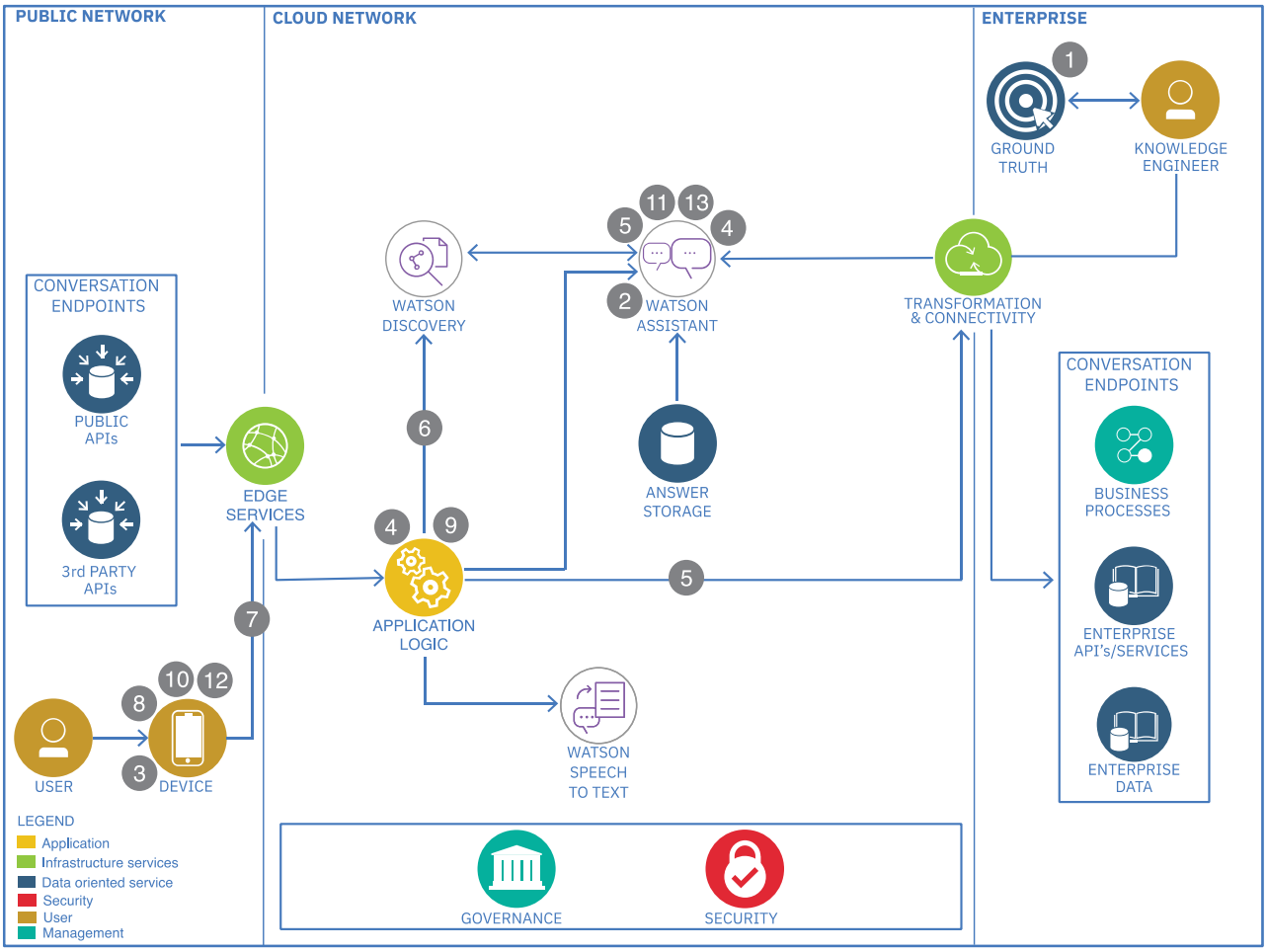


Figure 7. Conversation flow in rental car scenario

1. The business architect identifies Twitter, LinkedIn, enterprise customer service records, regional tourist information, restaurant types and ratings, digital car manuals and add-on information, and weather data as the sources of information. The business architect models intents and entities, then identifies and links entities, intents, and utterances for the preparation and creation of the ground truth.
2. The knowledge engineer trains and deploys the information from ground truth for conversation.
3. The customer's flight lands at the airport. This customer has opted-in to allow the rental car company to interact with them on their mobile device. Based on the airplane landing,

a bot that is powered by IBM Watson Assistant service contacts the customer with a welcome message and information, including the parking lot number, color, make, and model of the assigned rental car.

4. The customer isn't satisfied with the assigned car, starts a conversation with Watson, and requests information about an upgrade.
5. The customer is a business user and is entitled to an upgrade. The application logic checks the customer's profile in the enterprise system and sends the request to Watson Assistant.
6. Watson isn't trained on this question and checks the Watson Discovery service. Watson finds the answer that business users are eligible for an upgrade and sends the upgrade eligibility to the application logic.
7. Application logic checks the enterprise system on available upgrades based on the user profile and sends the upgrade options to the customer on their mobile application.
8. The customer indicates their upgrade choice.
9. The application logic connects with the enterprise system, upgrades the car, and sends new parking lot number and car color, make, and model information to the customer.
10. The customer asks how to reach the rental car facility and indicates that they're hungry and in a hurry to get to a meeting.
11. The conversation bot knows the intent "hungry" and entity "restaurant", and answers that the customer should take a rental car bus at level 1. The bot also identifies the intent that the customer is hungry and is pressed for time. The bot suggests a fast food restaurant that has drive-through service, good quality food, and is on the customer's planned route.
12. The customer is happy and thanks Watson for the food suggestion.
13. Watson bot thanks the customer and provides current weather information. The bot indicates that the customer should reach their meeting place approximately 30 minutes after leaving the rental car facility.

This rental car information bot scenario shows that the IBM Watson Assistant bot is a thinking system that was trained to answer specific questions with a level of accuracy and confidence, whereas a search application returns multiple answers and depends on the user to find a specific answer. When the conversation service isn't trained for a specific answer, it uses the discovery service to get valuable insights from the source of information.

IBM capabilities in a bot scenario

This table maps the IBM capabilities and services to the components in the architecture.

Component	Definition	IBM products
Edge services	Includes services that are needed to allow data to flow safely from the internet.	DNS, CDN, firewall, load balancer
Transformation and connectivity	Includes scalable messaging, transformation, and secure connectivity.	IBM Integration Bus container, IBM DataPower, IBM API Connect, IBM Secure Gateway
Application logic	Orchestrates the service request between the customer, the conversation, and the enterprise.	IBM Cloud workflow, IBM BPM SaaS, Node.js runtime
Conversation service	With the IBM Watson Assistant service, you can create virtual agents and bots that combine machine learning, natural language understanding, and integrated dialog tools to provide automated customer conversations. Watson Assistant provides a graphical environment to create natural conversation flows between your applications and your users.	Watson Assistant API
Public APIs	Information that is in other public clouds that can be accessed by using APIs.	Research data

Third-party APIs	Information that is made available by third parties.	Twitter APIs, IBM Watson Assistant API, IBM Watson Discovery News
Answer storage	Storage of trained content for conversation.	IBM Watson Assistant API, IBM Cloud Object Storage, IBM Cloudant®
Discovery service	Collects content, queries, and relevant answers to improve model.	IBM Watson Discovery service
Ground truth	Identified sources of information, entities, intents, and their relationships.	IBM Watson Assistant API
Enterprise data	Data that is stored in the enterprise data center that runs the customer's business.	IBM Db2, IBM Cloudant, IBM Cloud Object Storage, core systems
Enterprise process	Business processes that are used to run the customer's business.	Processes embedded in legacy applications
Enterprise API	Services that expose enterprise data.	Enterprise APIs such as rental car reservation APIs, hotel, tourist attraction APIs, food reviews, and more

Deployment considerations

The critical success factor for creating conversation systems or services is a secure, user-friendly cloud platform. The cloud platform provides capabilities for actionable insights. IBM Cloud, including the AI service, is available in standard (shared public), premium, or dedicated customer-specific cloud deployment options.

When you deploy AI systems, consider tenancy, privacy, region and language support, and performance and scalability.

Tenancy

This consideration involves deploying to customers that carry client confidential or sensitive private information. In these cases, consider using a premium or a dedicated deployment option to support a single tenant model. However, it might be acceptable to choose a multi-tenant model that is provisioned by using a standard or public deployment option.

Privacy

Don’t store or pass any confidential or protected health information (PHI) when you interact with an AI system. This guideline applies to both standard and dedicated deployment options.

AI systems store users’ conversations or interactions in the form of logs. Those logs might be used for machine learning model improvement. Although the standard for IBM Cloud offerings is to not share any log information, it might be necessary to provide a means to opt out of this capability completely. Services such as Watson Assistant allow the customer or user to opt out of logging.

Region and language support

When you deploy applications that involve multiple geographies and languages, you might need to deploy the services in multiple regions by using the IBM Cloud region settings.

AI systems must be designed and trained against various languages based on the support that is provided by the service. It is the responsibility of the application or the solution to pass the language parameters to the APIs during runtime.

Performance and scalability

To support a large volume of users, create a testing plan that involves load testing. You can use open source frameworks such as JMeter or third-party services such as BlazeMeter in IBM Cloud to create and run load tests.

The load test must include submitting various request sizes and concurrent users. Depending on the performance needs, you might need to scale the service instances in IBM Cloud. IBM Cloud offers capabilities to scale the services both horizontally and vertically. You can employ capabilities such as auto-scaling to configure the scaling based on demand, throughput, and memory usage.

References

- The IBM Advantage for Implementing the CSCC Cloud Customer Reference Architecture for Internet of Things (IoT)
- ISO/IEC 17788:2014 Information technology -- Cloud computing -- Overview and vocabulary
- ISO/IEC 17789:2014 Information technology -- Cloud computing -- Reference architecture
- IBM Watson Assistant API
- IBM Watson Speech to Text
- IBM Watson Discovery
- IBM Watson Developer Cloud
- IBM Cloud Platform
- AI glossary