# The IBM Advantage for Cognitive Conversation

## Cloud Architecture

## Table of Contents

The IBM Advantage for Cognitive Conversation Cloud Architecture ........................................ 1

Executive overview .................................................................................................................. 2
  Critical success factors for cognitive systems ........................................................................ 2
    The right cloud platform ....................................................................................................... 2
    New forms of data to gain insight .......................................................................................... 3
    Robust ecosystem ............................................................................................................... 3
    Amplification of human cognition ....................................................................................... 4
    Application of cognitive insights ........................................................................................... 4

IBM Cloud customer cognitive reference architecture ............................................................. 4

Creating a conversation-based cognitive system ..................................................................... 8
  Phase 1: Planning ..................................................................................................................... 8
  Phase 2. Preparation ............................................................................................................... 12
  Phase 3: Execution ................................................................................................................ 14

Components ............................................................................................................................. 17
  Public network components .................................................................................................. 17
    User ...................................................................................................................................... 17
    Device .................................................................................................................................. 17
    Conversation endpoint in the public network .................................................................... 18
  Cloud network components .................................................................................................. 18
    Edge services ....................................................................................................................... 18
    Conversation trained and deployed ................................................................................... 19
    Watson Discovery service .................................................................................................... 19
    Answer storage .................................................................................................................... 20
    Speech to text ...................................................................................................................... 20
    Application logic ................................................................................................................ 20
    Tranformation and connectivity ......................................................................................... 21
  Enterprise network components ............................................................................................ 22
    Conversation endpoints in the enterprise network ............................................................ 22
    Ground truth ....................................................................................................................... 22

Security architecture: Retail scenario .................................................................................... 22

Copyright © 2017 IBM Corporation
Executive overview

A cognitive business is an organization that creates knowledge from data to expand expertise, continually learning and adapting to outthink the needs of the marketplace.

Cognitive systems will transform how organizations think, act, and operate in the future with technologies that leverage natural language, hypothesis generation, and evidence-based learning.

The cognitive architecture discussed in this document is IBM’s best practices approach to describe the flows and relationships between business capabilities and architectural components for cognitive applications that use cloud computing infrastructure, platforms, or services. The elements of this architecture are used to instantiate a cognitive conversation system using the IBM® Cloud platform.

Before reading this paper, familiarize yourself with some core cognitive concepts described in the cognitive glossary.

Critical success factors for cognitive systems

For a cognitive system to be successful, you must have the following factors present in your environment.

The right cloud platform

To create an efficient cognitive 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 gives enterprises the following:

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

IBM Cloud is an industrialized cloud, which enables integration between data and applications and also between public, private, and proprietary clouds. And IBM Cloud is an industry-centric cloud, offering capabilities designed for industry-specific data or content and regulations. A trained cognitive conversation system provides insights for decision-making and can leverage the IBM Cloud platform services to take actions that are based on insights from the cognitive systems. IBM Cloud platform provides 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, 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 the existing data.

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

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

Here are a few examples of how you can use Watson services in conjunction with ecosystem partners to gain insights into data:

• Content captured by car manufacturers can be used with Watson services to provide car-related information to drivers and for self-driving cars.
• Content 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
Cognitive systems should amplify human intelligence. Humans train IBM’s cognitive systems to amplify human cognition. The systems are designed not to replace a human’s cognitive capabilities, but to enhance them.

Application of cognitive insights
When you’ve gathered cognitive 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 cognitive reference architecture

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

Figure 1: Cognitive reference architecture
1. **Conversation:** IBM’s cognitive conversation capabilities are trained to assist in decision-making 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 driven by dialog that you can use to determine the best course of action. Refer to the [Cognitive Conversation Reference Architecture](#) to understand how these capabilities are realized.

2. **Discovery:** IBM’s cognitive discovery capabilities ingest and enrich information, annotate the information 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 [Cognitive Discovery Reference Architecture](#).

3. **Extend:** You can extend IBM’s cognitive conversation and discovery capabilities with cognitive services that take broad or unstructured data and create meaningful, actionable, and valuable information for users, and can be domain-specific. Using a variety of capabilities such as Watson Speech to Text, Text to Speech, Tone Analyzer, Visual Recognition, Natural Language Classifier, 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 2 shows an example usage of this architecture.
A home improvement store improves its customer service with a cognitive 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.

Figure 2: Cognitive reference architecture usage flow
Serving flow

1. A customer accesses the mobile application 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 conversation component for a trained response. Conversation 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 if 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 conversation 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 cognitive reference architecture positions conversation, discovery, and extend capabilities in relation to each other. This paper covers the planning, preparation, and execution 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 Cognitive systems and services for decision assistance and planning is 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. Inputs from common utterances that are noticed by subject matter experts 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, a cognitive engineer prepares the dialog flow, tone, intents, and entities and defines their relationship using IBM tooling. (See the cognitive glossary for definitions.)

You can use trained IBM cognitive services or systems in any form factor (mobile, kiosk, car dashboard, web, voice response unit, and others) for decision assistance.
IBM systems learn from continuous interactions and from identified patterns. When the conversation system does not have a trained response, it uses the IBM Discovery service. This service examines relevant documents and annotates them with potential response information as a type of knowledge repository.

Because cognitive systems are part of the IBM Cloud platform, you can leverage 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 cognitive depends on having the right cloud platform and actions taken based on insights gained from analytics. IBM Cloud provides capabilities to develop these cognitive systems of engagement solutions using services like analytics, security, Internet of Things (IoT), blockchain for digital transformation foundation and processing, decision management, integration, and collaboration for implementing actions from the insights.

Refer to this IBM advantage paper on IoT for a thorough discussion and best practices.

- [The IBM Advantage for Implementing the CSCC Cloud Customer Reference Architecture for Internet of Things (IoT)]

**Creating a conversation-based cognitive system**
This paper describes three phases (planning, preparing, and execution) to create a conversation-based cognitive system. This section shows architectures that relate to all three phases.

**Phase 1: Planning**
Figure 3 shows the flow and the sequence of tasks needed to plan and design the cognitive conversation system for the preparation of ground truth.
Figure 3: Planning for a cognitive solution

The diagram shows the three 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 cognitive conversation system. Data scientists have deep knowledge of information that can be used to extract insights.
• **Representative users including crowdsourcing:** This includes 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 could be product experts, call center supervisors, scientists, doctors, or engineers. For example, one of the representative users for an appliance manufacturer could be a field technician.

This planning phase involves two broad categories of information sources.

• **Internal content sources:** This includes processes and data sources that are within a given enterprise. They typically contain the data generated (and owned) by the enterprise as part of its business operations. The conversation application does not 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 may need to be invoked to fulfill the intent.
  - **Business processes:** These are enterprise level business processes with which the chatbot may need to interact in order to process and respond to the user’s intent.
  - **Enterprise APIs or services:** This includes APIs or services that may need to be accessed or invoked in response to user queries and responses on the chatbot. For example, the user may 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 includes information that is kept in file systems shared between users and locations accessible through FTP and other mechanisms.
  - **Content from enterprise systems:** This is data from various enterprise systems including but not limited to catalogs, order or transaction data, and ECM repositories.

• **External content sources:**
  - **Shared file repository:** This includes information that is kept in file systems shared between users and locations accessible through FTP and other mechanisms.
  - **Public and third-party sources:** This includes information sources that are available for public consumption. This set of information is neither owned by the enterprise nor generated by the enterprise as part of the business operations. These include both public domain data (available free of cost) and data controlled by other parties. Examples include weather data and domain-specific catalogs made available by third-party vendors.
Multimedia content: This includes content such as audio, video, or images available on the Internet.

Social data sources: This is a subset of public and third-party sources specifically involving social media such as Twitter, Facebook, and others.

Public API or sources: This includes data accessible for public consumption, requiring an invocation of an API.

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

1. The business architect defines the goals and objectives of the conversational application or chatbot, including the channels that the application must support (web, mobile, social, and others).
2. Identify, define, and model the intents the conversational application must detect—and fulfill—from the utterances.
3. Identify, define, and model the entities to detect from the user utterance that 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 idea of the types of inputs that the conversation application might receive.
5. Map the utterances to intents. Continuously iterate over the utterances collection. Potentially identify new intents and entities that the application may have to handle.
6. Identify and map answer units or sections from the corpus (public and private) that must be provided as a textual response to the user. Identify the processes and the APIs that you may need to invoke to fulfill the intent.

In steps 7, 8, and 9, you map intents to flows and APIs, and split the train and test sets.

7. Split the collected and mapped utterance examples into train, test, and evaluation sets. You can use a randomized 70-20-10 percent split, where training the model for conversation ground truth is 70 percent, testing the model is 20 percent, and evaluation or feedback is 10 percent.
8. Identify and model any contextual attributes that will be passed to the conversation application or bot from the user application based on which flows and answers may differ. For example, you should have proper contextual variables for an insurance bot coming from Texas or one coming from New York.
9. For intents that need to execute a series of simple steps, leverage existing business process flows and business rules.

**Phase 2. Preparation**

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

![Diagram](image)

*Figure 4: Preparation for a cognitive solution*

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

This preparation phase involves two broad categories of information sources.
• **Internal content sources:** This includes processes and data sources that are within a given enterprise. They typically contain the data generated (and owned) by the enterprise as part of its business operations. The conversation application does not 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 may need to be invoked to fulfill the intent.

  o **Business processes:** These are enterprise level business processes with which the chatbot may need to interact in order to process and respond to the user’s intent.

  o **Enterprise APIs or services:** This includes APIs or services that may need to be accessed or invoked in response to user queries and responses on the chatbot. For example, the user may 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.

  o **Shared file repository:** This includes information that is kept in file systems shared between users and locations accessible through FTP and other mechanisms.

  o **Content from enterprise systems:** This includes data from various enterprise systems including but not limited to catalogs, order or transaction data, and ECM repositories.

• **External content sources:**

  o **Shared file repository:** This includes information that is kept in file systems shared between users and locations accessible through FTP and other mechanisms.

  o **Public and third-party sources:** This includes information sources that are available for public consumption. This set of information is neither owned by the enterprise nor generated by the enterprise as part of the business operations. These include both public domain data (available free of cost) and data controlled by other parties. Examples include weather data and domain-specific catalogs made available by third-party vendors.

  o **Multimedia content:** This includes content such as audio, video, or images available on the Internet.

  o **Social data sources:** This is a subset of public and third-party sources specifically involving social media such as Twitter, Facebook, and others.

  o **Public API or sources:** This includes data 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 leverages the work done by the business architect, including examining the collected ground truth.
2. With the IBM Watson Assistant service created in IBM Cloud, the knowledge engineer launches the tooling.
3. The knowledge engineer creates a workspace, giving it a name and a description, including the language. Each workspace contains the artifacts for an instance of your conversation application.
4. Upload a CSV file or create the intents.
5. Create the entities by giving them a canonical value and its synonyms (other surface forms) so that the same value might be entered. For example, "Beverage Type" can be an entity name, and "Coffee" can be a value with associated synonyms "dark roast" and "latte".
6. Enter each of the intent or utterance examples that were collected from the representative users. Note that the intents and the intent examples can be uploaded as a CSV file into the Watson Assistant tool.
7. Unit test the model by providing sample utterances to check if the correct intent for a given input is detected. If the intent that is displayed is not the correct intent for that input, you can add the input to a different intent. Ensure that the right entities are detected. Continuously iterate and repeat steps 5 and 6 as necessary to fine-tune the intent examples and entities.
8. Model the dialog that is made up of nodes, which define steps in the conversation. If you need to provide textual answers, use the answer units identified by the business architect. If the dialog is a simple process, implement the flow identified by the business architect. Use the context variables identified as necessary.
9. Test the model accuracy using (external) scripts with the test and evaluation data sets that you randomly split from the ground truth. Evaluate and analyze the accuracy, F1 score, and precision of the model.
10. Test and evaluate the dialog flow execution, especially variations with condition, context variables, and various entity values.
11. Deploy the trained model.
12. Enable for continuously monitoring the model and capturing the user utterances.

**Phase 3: Execution**

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

Figure 5: Cognitive conversation runtime solution

This example shows how a customer uses a trained cognitive 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 is the description of the components of the architecture. The second is the runtime flow showing 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 refers to unstructured content or the information stored in the enterprise network, including but not limited to FAQ documents, historical customer conversation records, operational manuals, and customer feedback. These, along with knowledge from the institutional SMEs, are used to prepare and train the cognitive conversation system. The enterprise APIs are also included that can be integrated as part of the overall conversation system from application logic.

2. **Conversation endpoints in the public network**: Enterprise information is strengthened with content or information including Wikipedia, news articles, and financial reports. You can use this information to prepare and train the cognitive conversation service. Third-party APIs like IBM Weather Data or IBM Insights for Twitter are also included 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 the conversation API. Content from conversation endpoints (both public and enterprise), example utterances, and representative SMEs crowdsourced inputs are used to model the intents, context attributes entities, and dialog responses and to train the conversation service. Ground truth is typically split into training, testing, and evaluation data.

4. **Conversation (trained and deployed)**: The knowledge engineer uses the ground truth that was collected in step 3 and the conversation API tooling 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 long-tail scenarios. The knowledge engineer ingests and potentially annotates unstructured documents, including manuals and training handbooks, and trains a ranker model to rank the returned passages for a given utterance.

6. **Answer storage**: Answers can be maintained in an external answer storage format. These answers are provided to the user once 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 cognitive system. Through the application logic component, the cognitive 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 passing the request to the conversation API.
9. **Application logic**: This might be a Node.js application. It first passes the natural language utterance (request) to the conversation service. When the response from the conversation 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 may 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 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 using transformation and connectivity components that provide secure messaging and translations to and from systems of engagement, enterprise data, and enterprise applications.

### Components

Let’s examine the individual components that make up the cognitive architecture.

#### Public network components

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

**User**

A cognitive user is a customer who uses his device to access the cognitive 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 cognitive 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 cognitive conversation service. Also included are third-party APIs like IBM Weather Data or IBM Insights for Twitter that can be integrated as part of the overall conversation system from application logic.

Cloud network components

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.

Key capabilities in this domain include:

- **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 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 does not 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 reliability of applications. Load balancers can balance loads locally and globally. Load balancers should be highly available without a single point of failure. Load balancers are sometimes integrated as part of the provider cloud analytical system components like stream processing, data integration, and repositories.

• **MFT (managed file transfer) gateway**: This is 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 a next-generation intrusion prevention system (IPS) that you can leverage to monitor network traffic and to provide protection from hidden security vulnerabilities. IBM DataPower® provides load balancing and SSL termination. It helps quickly secure, integrate, control, and optimize access to a range of workloads through a single, extensible, DMZ-ready gateway.

**Conversation trained and deployed**
The knowledge engineer uses the ground truth and the conversation API tooling to populate the intents, entities, dialog flows, and context for training the conversation API using the workspace 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. You can build Twitter, Slack, Facebook Messenger, and other messaging platform chatbots that interact instantly with channel users and allow customers to control your mobile application using natural language virtual agents.

**Watson Discovery service**
You can use this discovery service for knowledge expansion or long-tail 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 ranker model to rank the returned passages for a given utterance.
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. Built on Apache Solr, you load your data into the service, train a machine learning model based on known relevant results, and leverage 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 once 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 passing the request to the conversation API.

You can use IBM Watson Speech to Text service anywhere there is a need to bridge the gap between the spoken word and its written form. This easy-to-use 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. Additionally, the service includes the ability to detect one or more keywords in the audio stream. You can access the service via 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 may 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 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.
Cloud also provides Cloud Foundry services, so that you can deploy your application without managing the underlying infrastructure.

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

**Tranformation 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, the IBM Integration Bus container allows you to integrate applications and infrastructures deployed in multiple clouds or in legacy or core applications deployed in customers’ traditional data center.

IBM API Connect® is a comprehensive end-to-end 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. Using automated, model-driven tools, you can create new APIs and microservices based on Node.js and Java runtimes—all managed from a single unified console. Ensure secure and controlled access to the APIs 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 enable 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 running on-premises or in other clouds. A remote client is provided to enable secure connectivity.
Enterprise network components

Conversation endpoints in the enterprise network
The conversation endpoints in an enterprise network encompass unstructured content and information stored in the enterprise network, including but not limited to FAQ documents, historical customer conversations records, operational manuals, and customer feedback. These, along with knowledge from the institutional SMEs, are used to prepare and train the cognitive 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 cognitive techniques to help you build and train a bot—defining intents and entities, and crafting dialog to simulate conversation.

Ground truth
This is the set of artifacts used to train the conversation API. Content from 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 using any productivity tooling.

Security architecture: Retail scenario
Security is a very important aspect of the cognitive reference architecture. This diagram shows how the flow of information is secured, including securing the movement of information, authentication, access control, and auditing of all security requests for cognitive conversation.
In this scenario, a fashion design and retail enterprise stores mood boards, color boards, and images in their enterprise. These fashion designs are their trade secrets. However, the data must be moved to IBM Cloud to use IBM Watson services. The retailer also needs to leverage data that they don’t have, such as social media data, to understand customer interests. IBM built a cognitive fashion designer application for the customer. The customer wants to ensure that the content that is stored in their data center is secured and encrypted when it is in the data center, while moving the data to the IBM Cloud, and when accessed by Watson APIs. The fashion designer cognitive application needs to 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, using the enterprise’s own encryption algorithm, encrypts the fashion design mood boards and color board images.
3. The key used to encrypt the image is stored in the hardware key vault available in the customer’s data center. Alternatively, the customer can choose to use the hardware key vault 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 via integration with enterprise user directories and entitlement stores to authenticate the user, to validate permissions, and to determine access levels to content.

5. VPN and edge services 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 using the native encryption 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 cognitive fashion designer uses the application on his laptop.

10. The cognitive fashion designer’s user identity is authenticated and his 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 cognitive conversation architecture must be secured. Infrastructure security protects against network level threats and attacks with intrusion prevention and detection, including those tunneling through encrypted web transactions and web applications 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 securing access and providing encryption to data at rest and data in motion for the cognitive architecture, there still might be vulnerability. 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 provides real-time risk analysis of the workloads hosted in cloud against the myriad of known vulnerabilities and alerts against zero day attacks. This includes problem and information security incident management and responding to expected and unexpected events.
13. Although not shown in the diagram, all security authentication, access, and movement of information is logged for auditing.

**IBM capabilities for security in a cognitive conversation system**

This table shows the IBM capabilities and services mapped to the component in the architecture above.

<table>
<thead>
<tr>
<th>COMPONENT</th>
<th>DEFINITION</th>
<th>IBM PRODUCTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge services</td>
<td>Edge services include services needed to allow data to flow safely from the Internet.</td>
<td>DNS, CDN, firewall, load balancer</td>
</tr>
<tr>
<td>Transformation and connectivity</td>
<td>This includes scalable messaging, transformation, and secure connectivity.</td>
<td>IBM Integration Bus container, IBM DataPower, IBM API Connect, IBM Secure Gateway</td>
</tr>
<tr>
<td>Conversation service</td>
<td>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.</td>
<td>IBM Watson Assistant service</td>
</tr>
<tr>
<td></td>
<td>Watson Assistant provides an easy-to-use graphical environment to create natural conversation flows between your applications and your users.</td>
<td>--------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Key management service</td>
<td>A cloud-based security service that provides key lifecycle management (key creation, usage, and deletion) for encryption keys used in IBM Cloud services or customer-built applications, with &quot;root of trust&quot; backed by a hardware security module (HSM).</td>
<td>IBM Key Protect</td>
</tr>
<tr>
<td>Service Type</td>
<td>Description</td>
<td>Product</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>File encryption service</td>
<td>Safeguards data even when network protection fails. It has built-in and external key management, giving customers control over their encryption keys.</td>
<td>IBM Multi-Cloud Data Encryption</td>
</tr>
<tr>
<td>Secured connectivity</td>
<td>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.</td>
<td>VPN providers</td>
</tr>
<tr>
<td>Identity and access management</td>
<td>Identifies and authenticates the user. Determines access levels by using an enterprise security directory such as LDAP.</td>
<td>IBM Security Access Manager</td>
</tr>
<tr>
<td>Security monitoring and intelligence</td>
<td>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 hosted in cloud against the myriad of known vulnerabilities and alerts against zero day attacks.</td>
<td>IBM Security QRadar® SIEM</td>
</tr>
<tr>
<td>Infrastructure security</td>
<td>Protects against network-level threats and attacks with intrusion prevention and detection, including those tunneling through encrypted web transactions and web applications deployed within the system.</td>
<td>IBM Security Server Protection, IBM Security SiteProtector System</td>
</tr>
</tbody>
</table>
Rental car company cognitive 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 cognitive bot provides the customer with information about the weather, restaurants, and sightseeing information, in addition to car selection choices and information about the pre-assigned car. The following diagram shows the runtime of this bot.

![Cognitive conversation flow in rental car scenario](image)

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 him on his mobile device. Based on the airplane landing, a bot powered by IBM Watson Assistant service contacts the customer with a welcome message and information (parking lot number, color, make, and model of the assigned rental car).

4. The customer is not 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 is not 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 his mobile application.

8. The customer indicates his 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 he is hungry and in a rush 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 very 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 his 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 has been 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 is not trained for a specific answer, it uses the discovery service to get valuable insights from the source of information.
IBM capabilities in a cognitive bot scenario
This table shows the IBM capabilities and services mapped to the components in the architecture above.

<table>
<thead>
<tr>
<th>COMPONENT</th>
<th>DEFINITION</th>
<th>IBM PRODUCTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge services</td>
<td>Edge services include services needed to allow data to flow safely from the Internet.</td>
<td>DNS, CDN, firewall, load balancer</td>
</tr>
<tr>
<td>Transformation and connectivity</td>
<td>This includes scalable messaging, transformation, and secure connectivity.</td>
<td>IBM Integration Bus container, IBM DataPower, IBM API Connect, IBM Secure Gateway</td>
</tr>
<tr>
<td>Application logic</td>
<td>Orchestrates the service request between the customer, the conversation, and the enterprise.</td>
<td>IBM Cloud workflow, IBM BPM SaaS, Node.js runtime</td>
</tr>
<tr>
<td>Conversation service</td>
<td>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.</td>
<td>Watson Assistant API</td>
</tr>
<tr>
<td>Public APIs</td>
<td>Information that resides in other public clouds that can be accessed via API.</td>
<td>Research data</td>
</tr>
<tr>
<td>Third-party APIs</td>
<td>Information made available by third parties</td>
<td>Twitter APIs, IBM Watson Assistant API, IBM Watson Discovery News</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Answer storage</td>
<td>Storage of trained content for conversation.</td>
<td>IBM Watson Assistant API, IBM Cleversafe®, IBM Cloudant®</td>
</tr>
<tr>
<td>Discovery service</td>
<td>Collects content, queries, and relevant answers to improve model.</td>
<td>IBM Watson Discovery service</td>
</tr>
<tr>
<td>Ground truth</td>
<td>Identified sources of information, entities, intents and their relationship.</td>
<td>IBM Watson Assistant API</td>
</tr>
<tr>
<td>Enterprise data</td>
<td>Data stored in the enterprise data center that runs the customer's business.</td>
<td>IBM Db2, IBM Cloudant, IBM Cleversafe, core systems</td>
</tr>
<tr>
<td>Enterprise process</td>
<td>Business processes that are used to run the customer's business.</td>
<td>Processes embedded in legacy applications</td>
</tr>
<tr>
<td>Enterprise API</td>
<td>Services that expose enterprise data.</td>
<td>Enterprise APIs such as rental car reservation APIs, hotel, tourist attraction APIs, food reviews, and more</td>
</tr>
</tbody>
</table>

**Deployment considerations**

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

Here are some important things to consider when deploying cognitive systems:
**Tenancy**
This consideration involves deploying to customers carrying 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 may be acceptable to choose a multi-tenant model provisioned using a standard or public deployment option.

**Privacy**
A best practice is not to store or pass any confidential or protected health information (PHI) when interacting with a cognitive system. This applies to both standard and dedicated deployment options.

Cognitive systems store user’s conversations or interactions in the form of logs. These logs may be used for machine learning model improvement. Although the standard for IBM Cloud offerings is to not share any log information, it may 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 deploying applications that involve multiple geographies and languages, the services may have to be deployed in multiple regions using the IBM Cloud region settings. Additionally, cognitive systems must be designed and trained against various languages based on the support 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, you should create a testing plan that involves load testing. You can use open source frameworks like JMeter or third-party services such as BlazeMeter in IBM Cloud to create and execute load tests. The load test should include submitting various request sizes and concurrent users. Depending upon the performance needs, you might have to scale the service instances in IBM Cloud. IBM Cloud offers capability 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 utilization.
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
- Cognitive glossary