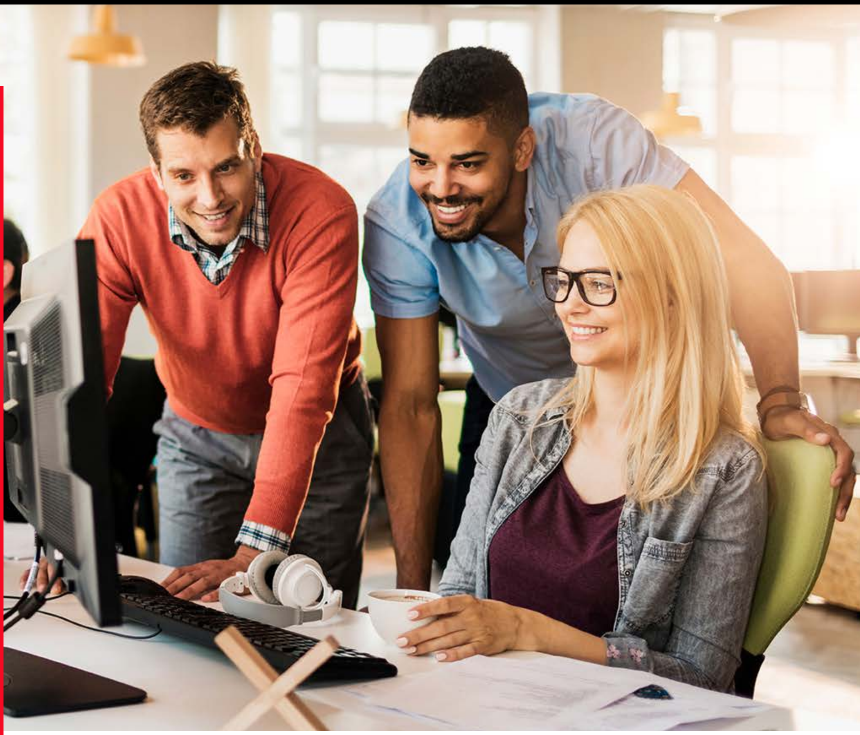
Transforming Enterprise Connectivity and Generative AI with Verizon and Amazon Bedrock



Introduction

Enterprises today confront a significant challenge – the proliferation of disparate data sources and siloed data across hybrid cloud environments. Extracting meaningful insights using generative AI tools has become increasingly arduous as data proliferates across heterogeneous distributed systems and formats. Customers have asked for programmable and secure connectivity to enable generative AI applications in the cloud and across their vast network of data sources. To address these challenges, AWS and Verizon have joined forces to offer a comprehensive solution using Amazon Bedrock and Verizon's private backbone network. This collaboration brings three key benefits:

- **Enhanced connectivity** – Verizon's robust network infrastructure provides high-speed, reliable connectivity to AWS cloud services, improving performance and reducing latency for applications and services.
- **Improved data security** – The solution integrates Verizon's network security features with AWS's cloud security tools, offering enhanced protection for data in transit and at rest.
- **Global reach** – By combining Verizon's network footprint with AWS's global cloud infrastructure, enterprises gain access to a more comprehensive and geographically diverse set of services.

This blog post explores how the combined capabilities of Verizon's Network-as-a-Service (NaaS) offerings and Amazon Bedrock's generative AI solution can address evolving customer requirements around performant, secure networks to unlock robust knowledge bases for Retrieval Augmented Generation (RAG) use cases.

Retrieval Augmented Generation (RAG) Primer

Amazon Bedrock allows enterprises to quickly build and scale generative AI applications with access to a diverse set of foundation models. One of the key features of Amazon Bedrock is the ability to access a choice of leading foundation models (FMs) through a single API. This allows developers to leverage the capabilities of large language models (LLM) without needing to manage the complexities of training and deploying these large-scale AI systems.

Bedrock also integrates Retrieval Augmented Generation (RAG), a technique that combines the strengths of LLMs with the information retrieval capabilities of external knowledge sources. This approach enables Bedrock-powered applications to provide more informed and context-rich responses, drawing upon a wealth of relevant information to enhance the quality and relevance of the generated content.

To ingest the requisite data source required to design robust RAG applications on Amazon Bedrock, Verizon's private backbone network, part of Verizon's network as-a-service (NaaS) portfolio, can help address the bandwidth, scalability, security and resiliency requirements for enterprise customers' hybrid networks. A private backbone is recommended for scenarios where real-time or time sensitive exchanges with AWS Bedrock is required, for exchange of classified or highly secured information or simply when the enterprises cannot afford to transport their data over the public internet due to security or performance requirements.

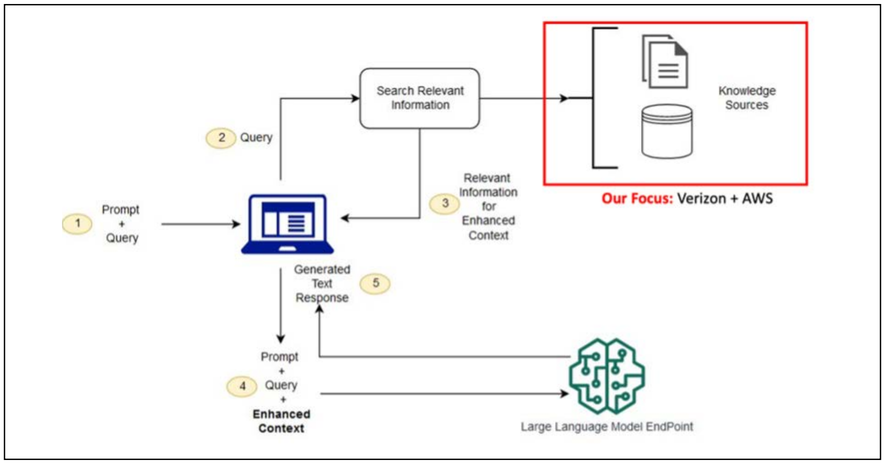

Search Relevant Information
Knowledge Sources
2 Query
Our Focus: Verizon + AWS
3 Relevant Information for Enhanced Context
1 Prompt + Query
Generated Text Response 5
4 Prompt + Query + Enhanced Context
Large Language Model EndPoint

Figure 1: Retrieval Augmented Generation (RAG) workflow of Amazon Bedrock Knowledge Bases

As one such example to connect enterprise customer branches to AWS Direct Connect locations, Verizon SCI is a usage-based model service that charges the customer based on the amount of traffic transmitted in and out of AWS. Alternatively, Verizon SDI and Dedicated Ports are standard port-based services with Committed Information Rate (CIR) for up to 100 Gbit/s. All three are connected to Verizon private backbone and secure layer 3 network. Customers can also add their data centers and other locations to the same backbone network, connecting them to the AWS VPC over a secure, highly performing network, associated with traffic SLAs.

Leveraging Network-as-a-Service for Generative AI

Network connectivity is a critical enabler for generative AI adoption and success for enterprise customers. As businesses increasingly turn to powerful language models and FMs to power their customer-facing applications, the ability to seamlessly connect and integrate disparate data sources across multiple locations becomes paramount. Enterprises have expressed a growing need for programmable, secure, and scalable connectivity solutions that can support the data requirements of advanced generative AI applications.

Our solution consists of the following key components:

Verizon Wireless Network – Across the entire solution, regardless of the underlying device producing data, Verizon Private Wireless Network (4G or 5G) can provide a secure and private wireless connection to both mobile and fixed location users. Alternatively, the wireless option can be used to connect users to AWS Wavelength services.

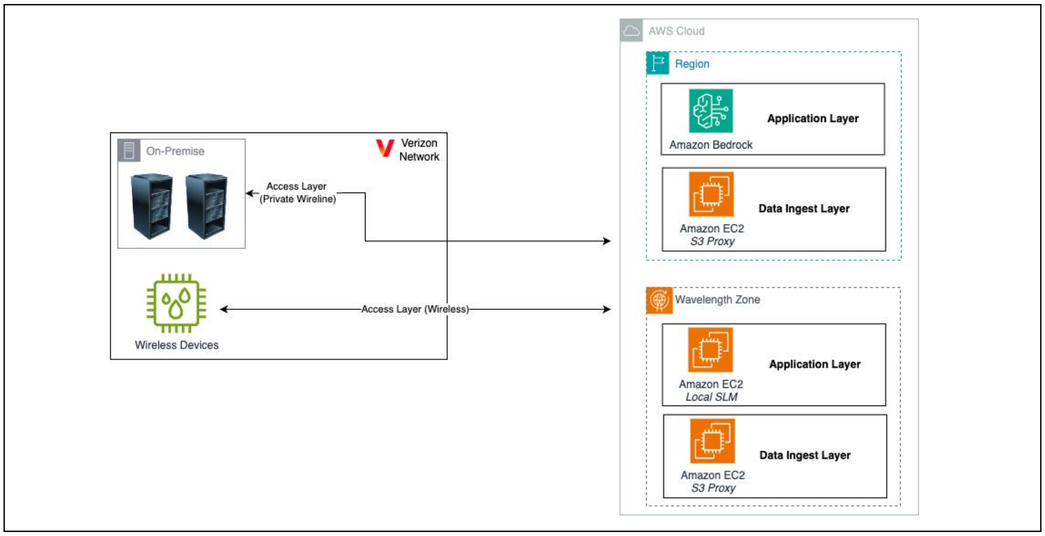


Figure 2: Solution overview of application layer and data ingest layer for both Verizon wireless and private wireline-connected devices.

- **Verizon Wireline Network** – Verizon private wireline backbone network enables you to connect securely to your growing cloud environment over connections that are completely separated from the public internet. Customers can connect to leading cloud service providers including Amazon Web Services. Verizon's private network connectivity services to the cloud such as Secure Cloud Internet (SCI) and Software Defined Interconnect (SDI) are two Verizon examples that can be used to provide that connectivity.
- **Data Ingest Layer (AWS)** – Through Elastic Load Balancing and VPC endpoints, customers can create a data ingest layer that allows data sent through the private Verizon backbone to securely transit to Amazon S3. Once stored in S3, these data sources can be quickly integrated into Amazon Bedrock Knowledge Bases for RAG use cases.
- **Application Layer (AWS)** –Amazon Bedrock Knowledge Bases allows customers to build RAG applications using data ingested to AWS via Verizon NaaS.
- **Access Layer (AWS)** – For wireless devices, Verizon can manage programmable 5G connectivity through its Quality on Demand (QoD) APIs to an AWS Wavelength Zone (Verizon 5G Edge) to reach the AWS Region-based generative AI application – or even a localized, edge-based small-language model (SLM) hosted a generative AI application at the edge.

This solution can extend, over time, to other Verizon connected devices through the rich portfolio of both wireless and wireline solutions. To illustrate this architecture in action, consider a pharmaceutical manufacturer tasked with ensuring the proper handling and delivery of temperature-sensitive vaccines or biologics.

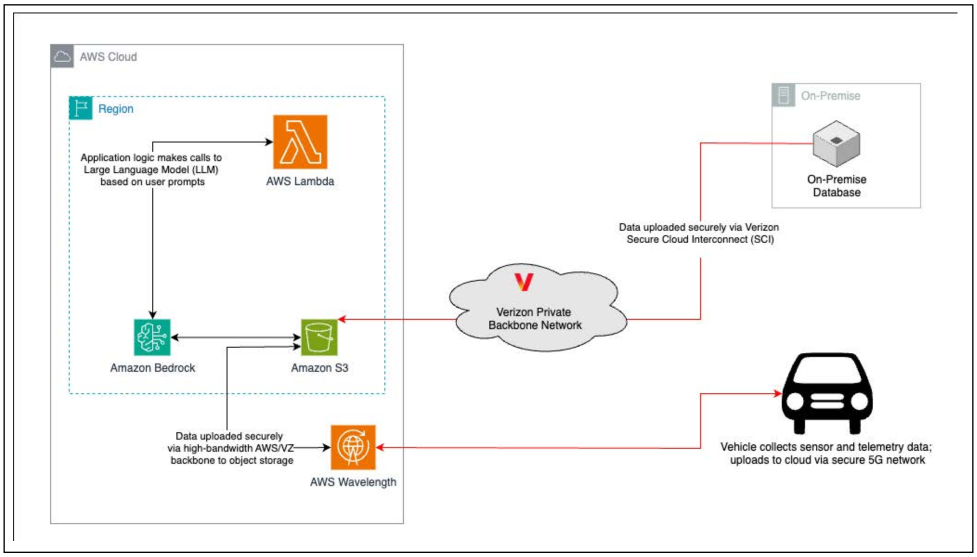

AWS Cloud
Region
Application logic makes calls to Large Language Model (LLM) based on user prompts
AWS Lambda
On-Premise
On-Premise Database
Data uploaded securely via Verizon Secure Cloud Interconnect (SCI)
Verizon Private Backbone Network
Amazon Bedrock
Amazon S3
Vehicle collects sensor and telemetry data; uploads to cloud via secure 5G network
Data uploaded securely via high-bandwidth AWS/VZ backbone to object storage
AWS Wavelength

Figure 3: Sample generative AI use case leveraging both Verizon wireless and private wireline-connected data sources.

Scenario: Logistics and manufacturing companies responsible for the production and distribution of high-value, time-sensitive goods are constantly looking to improve the health, safety, and auditability of their end-to-end operations. The ability to maintain real-time visibility and responsiveness, in the absence of specialized digital twin and computer vision technology, becomes particularly challenging.

To quickly answer critical questions like "Is there any upcoming risk of temperature excursion in the shipment of product batch XYZ?" or "What is the current location and estimated time of arrival for the delivery van carrying the rush order for hospital ABC?", a multi-modal foundation model with both text-generation and computer vision capabilities is essential. But this AI-driven insight is only as valuable as the underlying data it can access, which requires ultra-reliable, high-bandwidth network performance from the factory floor, distribution centers, and even the delivery vehicles en route to the final destinations.

Solution (Wireless) – Using AWS Wavelength, the pharmaceutical company can bring AWS compute and storage resources closer to Verizon's 4G/5G network, enabling their fleet of delivery vehicles to stream real-time telemetry, sensor, and video data to low-latency endpoints. These Wavelength Zone access points can then proxy the most critical information, such as temperature readings, GPS coordinates, and computer vision-processed footage of the cargo areas, directly to Amazon S3 over Verizon's private IP network, in conjunction with SCI or SDI, to serve as the data foundation for the company's generative AI application.

Solution (Wireline) – Complementing the wireless connectivity, the pharmaceutical company leverages Verizon Network-as-a-Service to establish dedicated, high-bandwidth links between their manufacturing sites, distribution hubs, and the AWS cloud. Using Verizon Private IP and private cloud connectivity products such as SCI or SDI, they can reliably transmit large volumes of process control data, quality inspection reports, and video streams from their on-premises computer vision systems. This data is then fed into the knowledge base powering their Amazon Bedrock application, allowing the company to ask detailed, multimodal questions about the end-to-end status of their supply chain operations.

In the customer's application using Amazon Bedrock, the two juxtaposed data domains within a single knowledge base – coupled with tactful prompt engineering – will allow the customer to ask detailed questions about the health and safety of their operations without manually reconciling multiple data sources.

This use case, however, is fundamentally industry-agnostic. Enterprises across a wide range of sectors can benefit from its capabilities, including: healthcare applications analyzing real-time medical imagery or financial services firms executing near-real time trades based on its own operations.

This example clearly demonstrates how the generative AI capabilities of Amazon Bedrock coupled with the secure, high-performance connectivity through Verizon NaaS enable enterprises to extract more value from their data. Enterprises can dramatically improve the efficiency and accuracy of their customer-facing applications by automating the process of generating tailored responses, whether they are chatbots, RFP response systems, or other interfaces that requires the synthesis of large amounts of data.

Solution architecture

To illustrate a vertical-agnostic, end-to-end traffic flow of the solution described above, consider the following step-by-step guide:

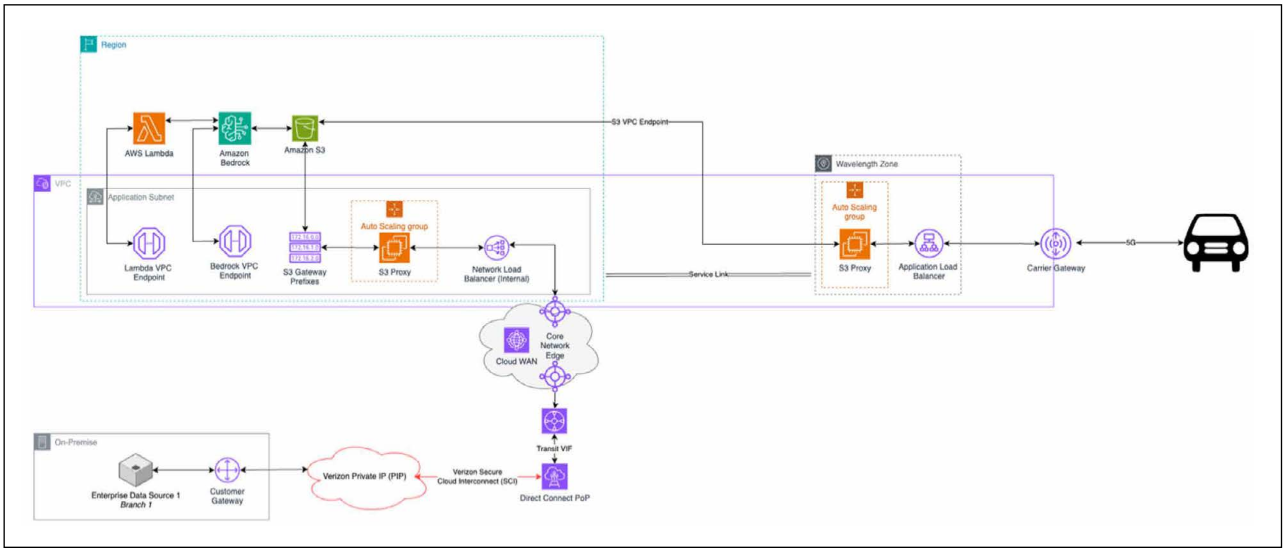


Figure 4: End-to-end architecture of sample generative AI use case leveraging both Verizon wireless and private wireline-connected data sources.

Step 1: The solution leverages AWS Direct Connect to AWS Transit Gateway and, optionally, AWS Cloud WAN, providing the requisite routing between Verizon network and the customer VPC where the generative AI application is deployed.

For user access, numerous options are offered by Verizon. Standard xDSL, Ethernet access, Fixed Wireless Access (FWA) and Verizon 4G/5G Mobile Private Network in conjunction with a Private Wireless Gateway can be leveraged to access Verizon private backbone network. SCI or SDI can then be used to connect the aforementioned AWS Direct Connect to that same private backbone network, extending the user connections to AWS Cloud. Verizon wireless network can be also leveraged to connect to an AWS Wavelength service when closer proximity to a mobile user is required.

Step 2: There are many ways to transfer data to an Amazon S3 bucket. In this particular example, we've chosen to use an S3 proxy, consisting of an Auto Scaling Group (ASG) of EC2 instances running a Squid proxy. This proxy is exposed via an Internal Network Load Balancer (NLB) and proxies requests from the customer premises to an Amazon S3 bucket. To learn more about additional approaches to transfer data to S3, visit Optimizing Amazon S3 data transfers over Direct Connect.

Step 3: From the S3 bucket(s) where the data is currently uploaded, a knowledge base is created to provide a fully-managed RAG workflow from ingestion to retrieval and prompt augmentation, without having to build custom integrations to data sources and manage data flows.

Step 4: To create the business logic that invokes the knowledge base, a Lambda function is exposed that receives user prompts, and forwards the prompt to the Knowledge Base using the RetrieveAndGenerate API. When you send a prompt like "Is there any upcoming risk of temperature excursion in the shipment of product batch XYZ?" to Amazon Bedrock, it utilizes the RetrieveAndGenerate API to interact with your knowledge base. This API first retrieves relevant information from your indexed documents, such as shipment logs, temperature sensor data, and quality control reports. Then, it passes this retrieved context along with your original query to a LLM. The LLM processes this information to generate a comprehensive and contextually relevant response about potential temperature excursions for the specific product batch. This approach combines the power of information retrieval with natural language generation, enabling more accurate and informed answers based on your organization's specific data and documents.

Step 5: The Lambda function itself is invoked from a static web application served from Amazon S3, an Amazon CloudFront distribution, or a locally-hosted web application using Streamlit or React.

Conclusion and getting started

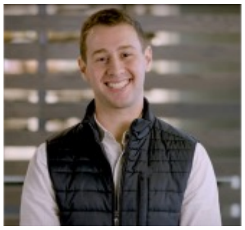
In this blog post, we complement the existing breadth and depth of features that RAG on Amazon Bedrock can provide with Verizon's network as-a-service (NaaS) portfolio to address the bandwidth, scalability, security and resiliency requirements for enterprise customers' networks.

To learn more about the solution, reach out to your AWS or Verizon account manager or get started with Building Contextual Chatbots using Amazon Bedrock Knowledge Bases and Verizon Connectivity Designed for the Cloud.

Authors

Scott Wainner

Scott Wainner is a Principal Solutions Architect focused on identifying and developing modernized methods and patterns that accelerate the customer's journey to the AWS. He has over 35 years of experience in designing, building, and operating complex information systems. He has extensive experience in networking, telecommunications, managed services, and management support systems.



Robert Belson

Robert is a Senior Solutions Architect in the AWS Worldwide Telecom Business Unit, specializing in generative AI at-scale enablement and jGTM activities. As an AWS Hybrid Edge specialist and host of the AWS On Air weekly show, he focuses on working with the developer community and large enterprise customers to solve their business challenges using automation, hybrid networking and the edge cloud.

Pavan Gupta

Pavan Gupta is a Product Manager at Verizon Business Group leading product strategy and vision for cloud connectivity products securely connecting enterprises to cloud providers. Pavan specializes in several areas of networking products including SDN, NFV, 4G/5G Packet Core Networks, wireline/wireless and data center networking.

Ghassan Semaan

Ghassan Semaan is an Associate Director - Solution Architect at Verizon where he leads a team that focuses on developing advanced and innovative networking solutions. His team is currently working on network solution designs that support the development and deployment of AI models.