Problem Statement

The core problem Slack needed to solve was developing an AI system that delivers powerful productivity benefits while maintaining the high levels of security and privacy that enterprise customers require. Specifically:

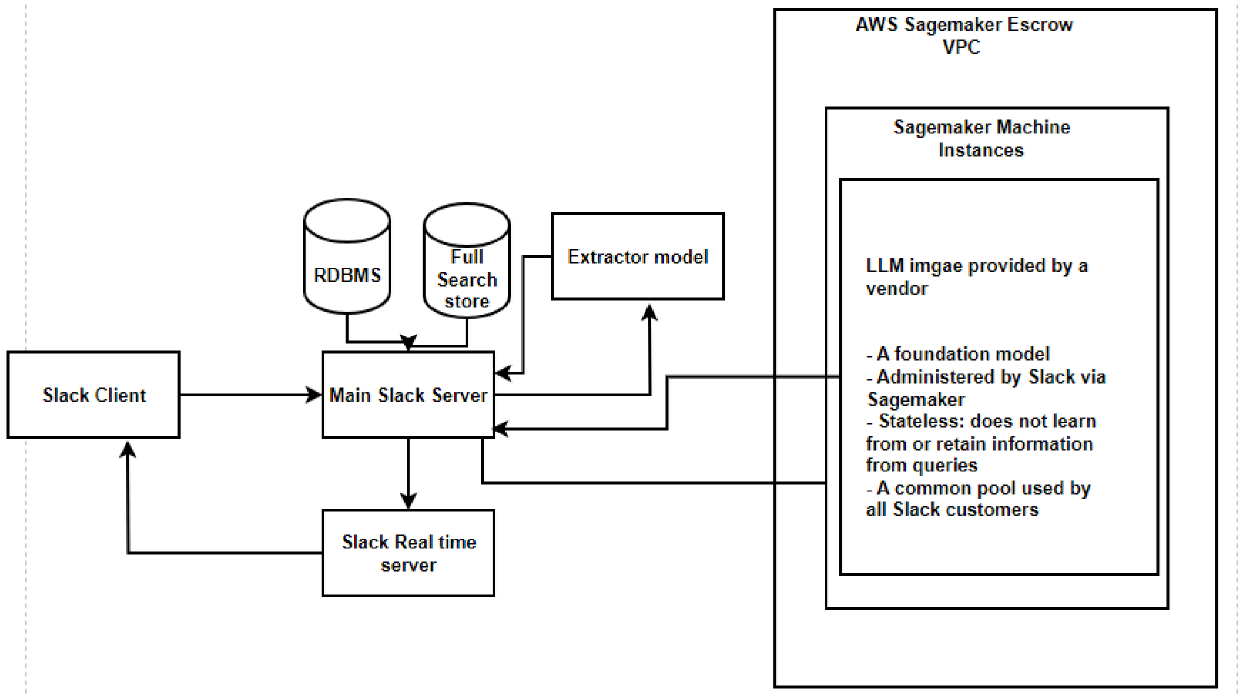
1. Enterprise customers have strict security requirements and confidentiality concerns about their data
2. Traditional AI implementations often create security and privacy risks through data exposure
3. Organizations need confidence that their sensitive information won't be misused or leaked when using AI features
4. Companies in regulated industries have specific compliance requirements that must be satisfied
5. There's an inherent tension between AI's need for data access and organizations' need to protect information

Proposed Solution

Slack's proposed solution is a secure-by-design AI assistant called "Slack AI" that:

1. Processes data within Slack's secure infrastructure rather than sending it to external providers
2. Respects existing workspace permissions and access controls
3. Gives organizations granular control over how AI features can be used
4. Provides transparency about what data is used and how
5. Enables organizations to maintain compliance with regulations while still benefiting from AI

Solution Design


AWS Sagemaker Escrow VPC
Sagemaker Machine Instances
LLM imgae provided by a vendor
- A foundation model
- Administered by Slack via Sagemaker
- Stateless: does not learn from or retain information from queries
- A common pool used by all Slack customers
RDBMS
Full Search store
Extractor model
Slack Client
Main Slack Server
Slack Real time server

Slack's solution design includes several key architectural and operational components:

1. **Secure Data Processing Architecture**:
   - Data remains within Slack's secure infrastructure during processing
   - Custom-built proxy service mediates access to LLM providers
   - System ensures data never persists outside of Slack's security boundaries
   - End-to-end encryption for data in transit
2. **Permission-Based Access Control**:
   - AI features respect existing Slack permissions
   - Users can only generate summaries for channels they have access to
   - Search results are filtered based on user permissions
   - All AI-generated content follows the same visibility rules as normal content
3. **Administrative Controls**:
   - Workspace owners can control which teams can use AI features
   - Options to enable/disable specific AI capabilities
   - Controls for deciding what data can be used for AI features
   - Settings to manage data retention policies
4. **Privacy-Preserving Implementation**:
   - No user data is used to train LLMs
   - Strict data minimization principles
   - Context window management to limit data exposure
   - Clear user interface elements showing when AI is being used
5. **Compliance Framework**:
   - Designed to work within regulated environments
   - SOC 2 Type II certification

    - HIPAA compliance capability
    - Regular security testing and third-party verification

The solution aims to deliver AI-powered productivity benefits while preserving the security, privacy, and compliance characteristics that enterprises require, effectively balancing innovation with protection.

Slack leverages SageMaker JumpStart to securely deploy and manage LLMs on its AWS-owned infrastructure while optimizing latency, scalability, and cost.

Key Takeaways

1. Security & Data Privacy

- Data sent to SageMaker models remains within Slack’s AWS infrastructure.
- End-to-end encryption ensures data security in transit.
- Data is not used for training base models, ensuring privacy compliance.

2. Model Deployment & Optimization

- Slack AI runs on AWS multi-GPU instances to host multiple model copies.
- Efficient resource utilization helps reduce deployment costs.
- SageMaker JumpStart provides access to diverse LLMs, enabling Slack to select the best fit for their use cases.

3. Load Balancing for Performance

- SageMaker's default RANDOM routing distributes traffic uniformly but inefficiently.
- Slack switched to LEAST_OUTSTANDING_REQUESTS (LAR) routing, which prioritizes instances with lower workloads.
- The LAR strategy improved p95 latency by over 39%, enhancing overall performance.