

Running on-prem AI agents for NOC environments

cisco Live !

Adrian ILIESIU
Developer Advocate, @aidevnet

Session ID: CISCOU-1620

Cisco Webex App

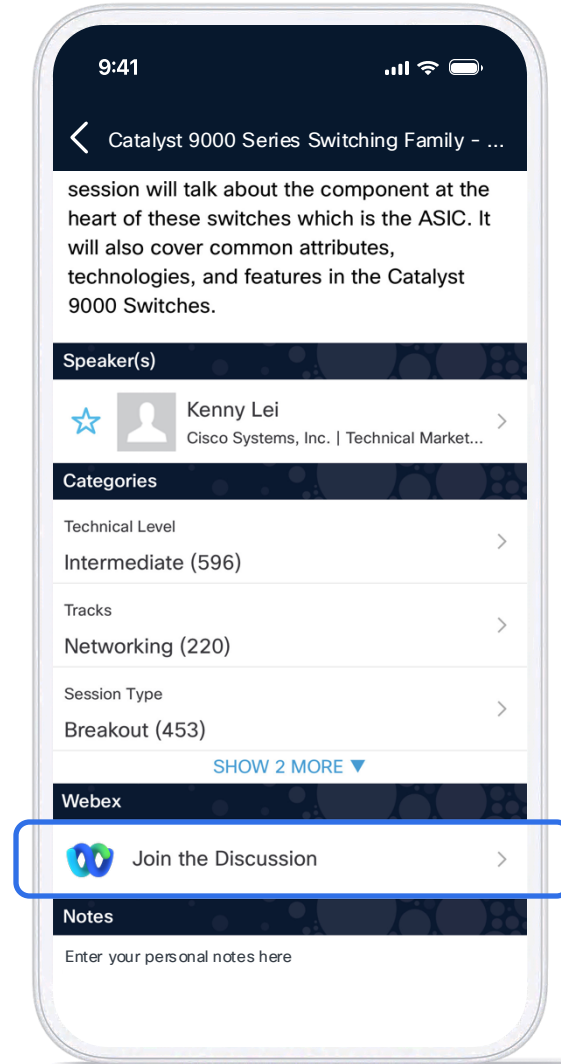
Questions?

Use Cisco Webex App to chat with the speaker after the session

How

- 1 Find this session in the Cisco Live Mobile App
- 2 Click “Join the Discussion”
- 3 Install the Webex App or go directly to the Webex space
- 4 Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until 14 November 2025.



<https://ciscolive.ciscoevents.com/ciscolivebot/#CISCOU-1620>

Agenda

- 01 AI for network automation
- 02 Introduction to n8n
- 03 AI pipeline in n8n
- 04 Demo
- 05 Conclusion
- 06 Resources
- 07 Q&A

AI for network automation

AI for network automation

- Enhanced Monitoring and Anomaly Detection*
- Automated Troubleshooting and Remediation
- Network Optimization and Resource Management
- Security and Compliance
- Service Management and Orchestration
- Digital Twins

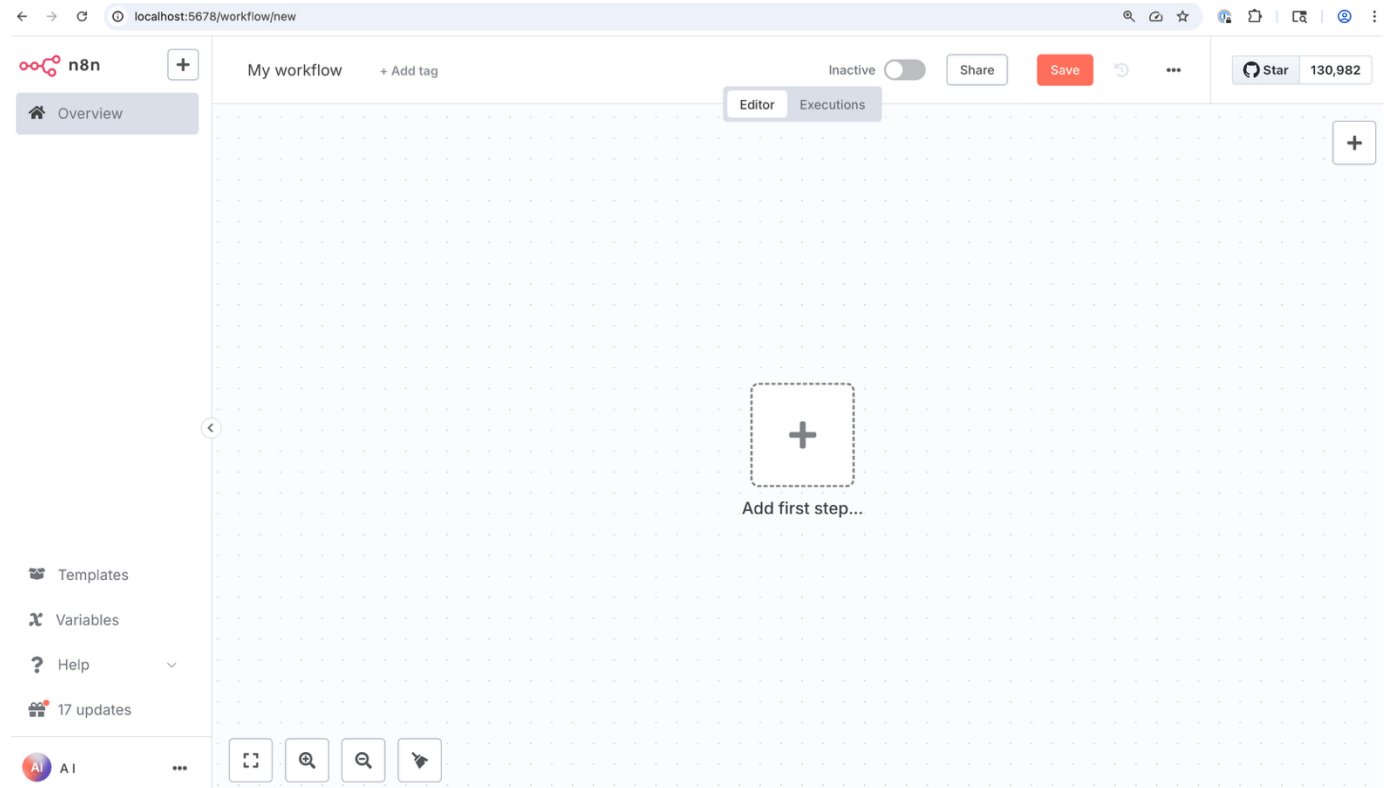
Introduction to n8n

What is n8n?

- **n8n** is a powerful, workflow automation tool that helps you connect applications and automate tasks without writing extensive code.
- Digital assistant that handles repetitive tasks by leveraging hundreds of integrations to connect disparate systems.
- Fair-code licensed with the option to run it in the cloud as a SaaS offering or on-prem.
- Privacy focused with the self-hosted on-prem option.

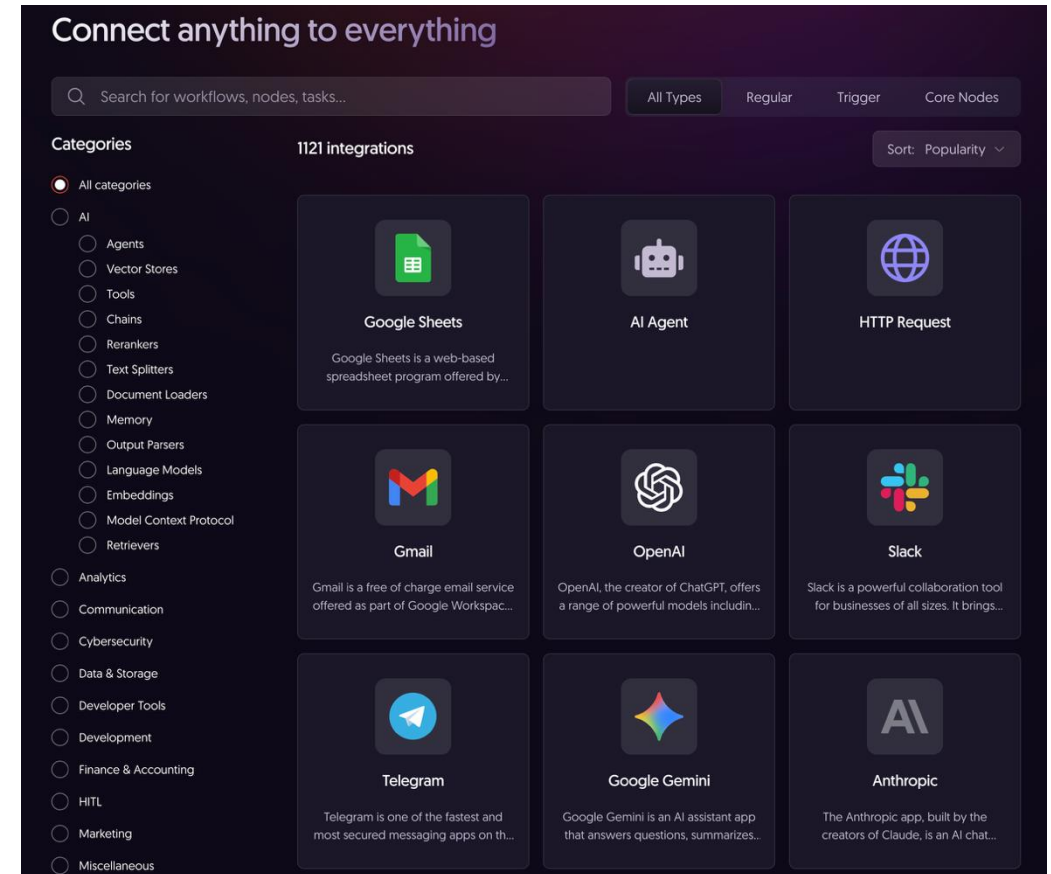
Key Concepts – Visual Workflow Builder

- Drag-and-drop interface to design complex automation workflows.
- Connect different applications and services seamlessly.



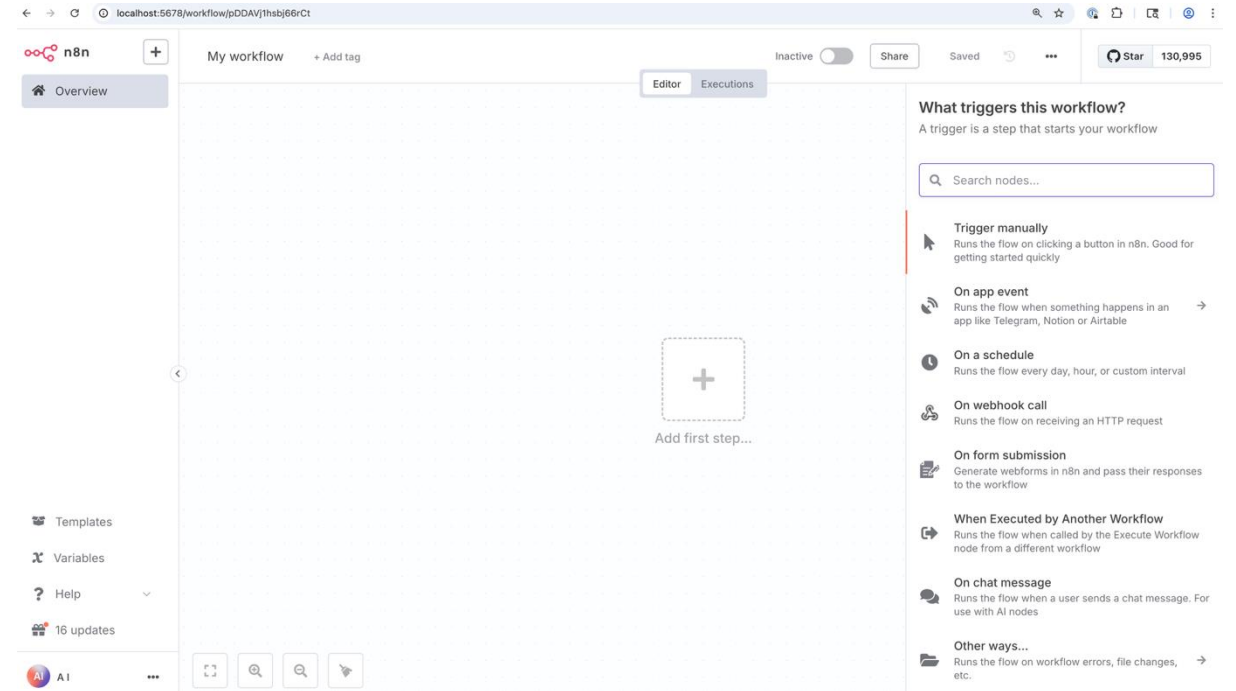
Key Concepts - Nodes

- Each step in a workflow is a "node" (e.g., "Trigger," "Google Sheets," "Slack," "HTTP Request").
- Hundreds of pre-built integrations for popular apps and services.



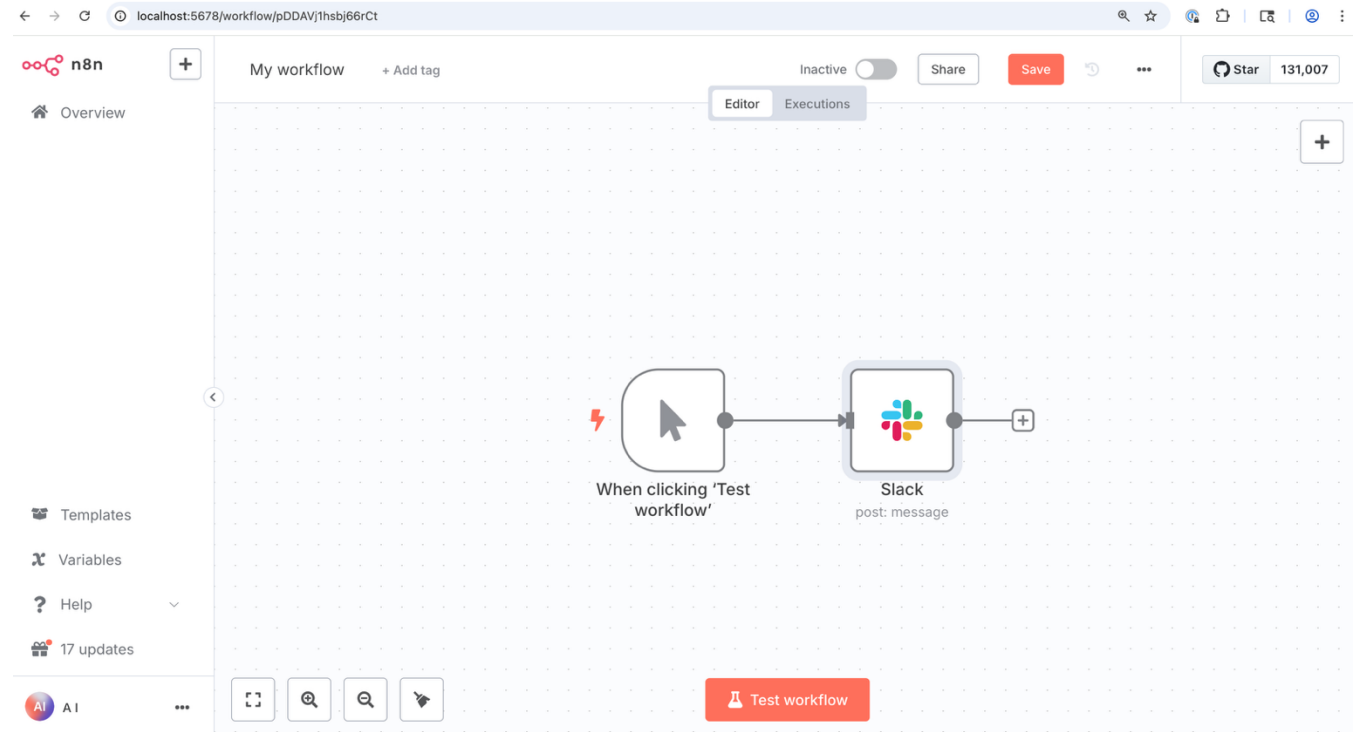
Key Concepts - Triggers

- **Triggers:** Start a workflow (e.g., new email, new entry in a database, scheduled time).

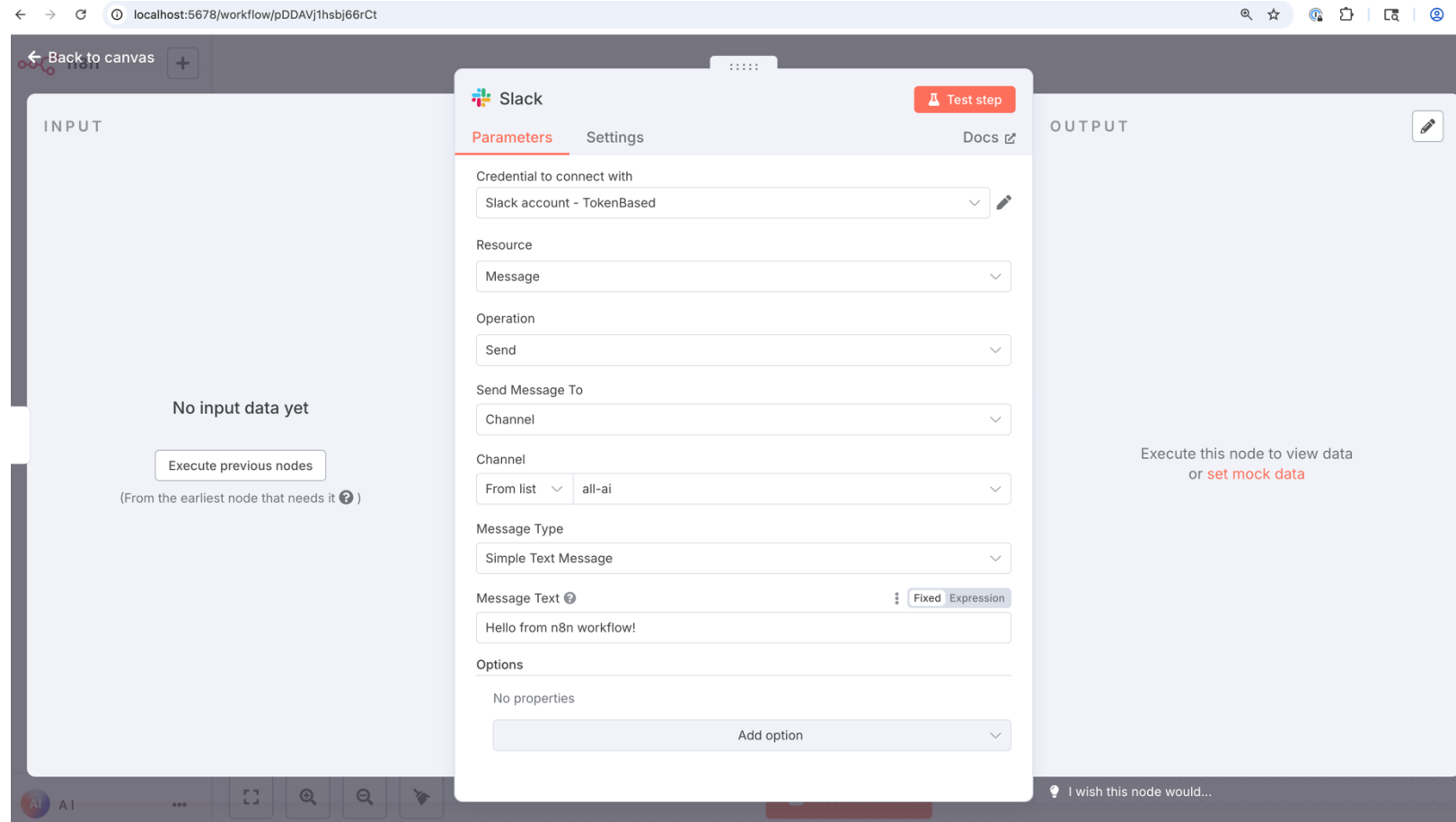


Key Concepts - Actions

- **Actions:** Perform tasks within applications (e.g., send a message, update a record, create a file).



Key Concepts - Actions



Why use n8n?

- **Automate Repetitive Tasks:** Free up time by automating routine data transfers, notifications, and processes.
- **Integrate Disparate Systems:** Connect applications that don't natively "talk" to each other.
- **Improve Efficiency:** Streamline workflows across teams and departments.
- **Reduce Manual Errors:** Automate data handling to minimize human mistakes.
- **Scalable & Flexible:** Adapt to evolving business needs with a highly customizable platform.
- **Data Privacy:** Self-hosting options provide greater control over sensitive data.

Common use cases

- **IT & DevOps:** Automate incident response, system monitoring, and deployment tasks.
- **Data Synchronization:** Keep databases, spreadsheets, and cloud services in sync.
- **Internal Tools:** Build custom internal tools and dashboards by connecting various services.
- **Marketing Automation:** Automate lead nurturing, social media posting, and email campaigns.
- **Sales Operations:** Sync CRM data, automate follow-ups, and generate reports.

AI pipeline in n8n

Challenges in NOC environments

- **High Volume of Messages**
- **Message Noise and Repetition**
- **Message Format and Parsing Issues**
- **Lack of Context and Correlation**
- **Severity Level Misinterpretation**

Possible Solution

- **Centralize syslog messages**
- **Load them in a vector database**
- **Run an LLM on the data**
- **Build a chat interface to interact with the data**
- **Profit 😊**

Vector databases

- **Databases designed to store, index and search efficiently vector embeddings (numeric representations of data).**
- **A vector is an array of numbers that represents complex data (text, images, audio, video, etc.) in a way that preserves semantic meaning.**
- **Example: The word “cat” might be represented as [0.12, 0.98, -0.48, ...] in a 512-dimensional space. The word “feline” would have a similar vector.**
- **Used in:**
 - **Semantic search**
 - **Recommendation systems**
 - **Image similarity search**
 - **Chatbots & RAG***
 - **Fraud detection**

Qdrant

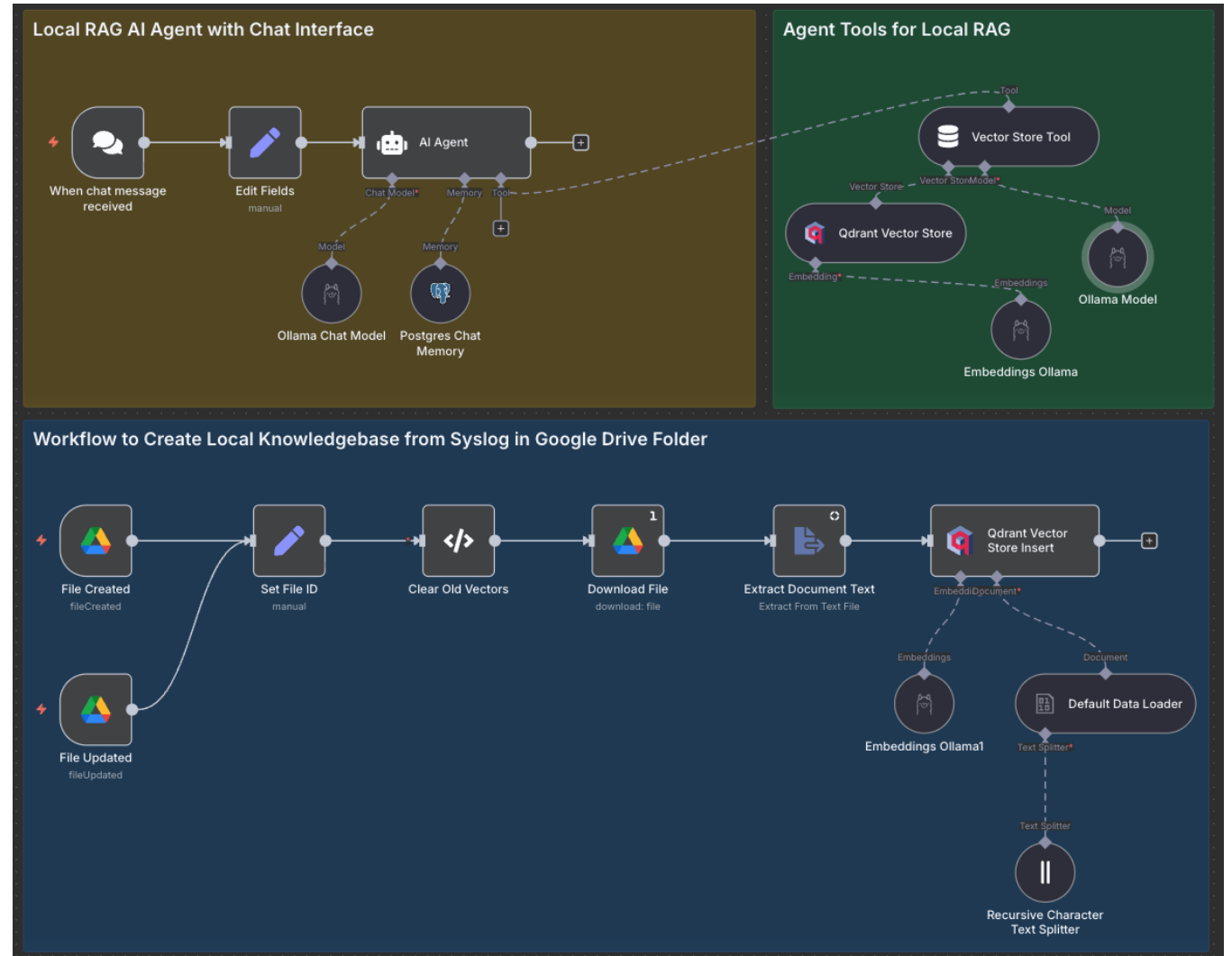
- **Open-source vector database and vector similarity search engine**
- **Written in Rust -> High Performance**
- **Used from small projects to production-level deployments handling billions of vectors -> Scalable**
- **Distributed mode and clustering supported -> High Availability and Throughput**
- **Allows to store metadata (“payload”) together with the vector data -> Rich Data Types**
- **REST and gRPC APIs and client libraries for Python, Go, Rust -> Ease of Use**

Ollama

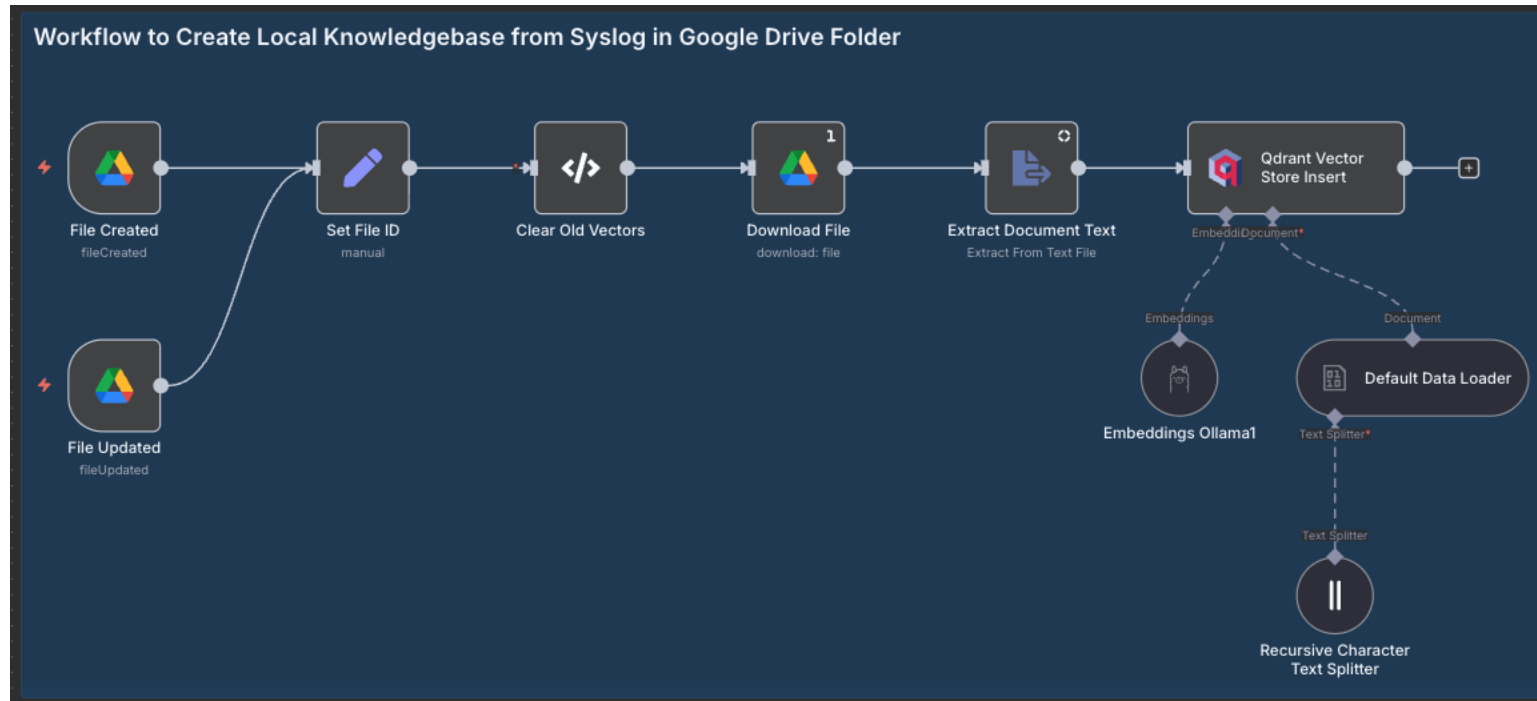
- **Open-source tool designed to make it very easy to run large language models (LLMs) locally on your computer**
- **Described as “Docker for LLMs”**
- **Simple CLI: ollama run llama3**
- **Extensive model library:**
 - **Llama 3 from Meta**
 - **Mistral and Mixtral from Mistral AI**
 - **Gemma from Google**
 - **Phi-3 from Microsoft**
- **Built-in API server compatible with OpenAI API format**
- **Cross-platform support: macOS, Windows and Linux**

AI pipeline in n8n

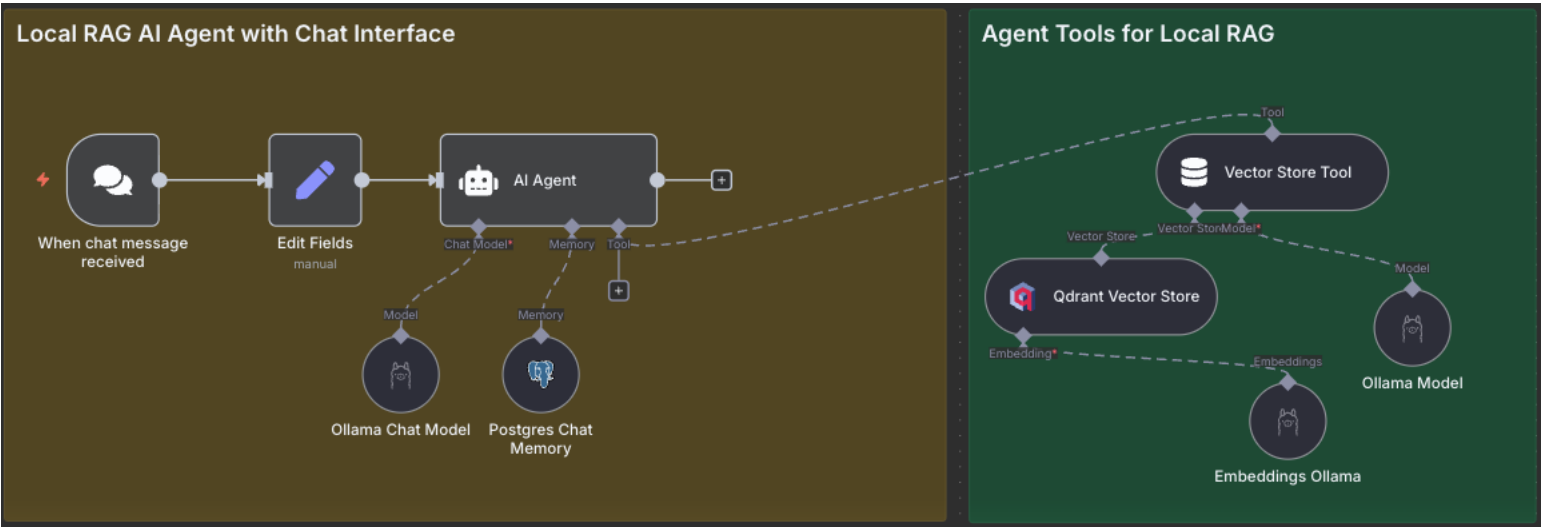
- **Components:**
 - **AI Agent**
 - **Google Drive – Syslog file**
 - **Ollama – LLM**
 - **Qdrant – Vector Database**
 - **Postgres – Relational Database**



Step 1 - syslog data ingestion



Step 2 – Chat interface



Demo

Conclusion

- **Lots of use cases for using AI with network automation including enhanced monitoring**
- **n8n is a powerful, workflow automation tool that helps you connect applications and automate tasks without writing extensive code**
- **AI agent node in n8n makes it very easy to build AI agents and integrate them with local or SaaS LLMs**

Resources

- <https://github.com/aidevnet/local-ai-n8n.git>
- <https://github.com/n8n-io/self-hosted-ai-starter-kit.git>
- <https://developer.cisco.com/site/ai/>
- <https://n8n.io>
- <https://qdrant.tech/>
- <https://ollama.com>
- <https://ngrok.com>

Thank you

CISCO Live !

