

Lenovo AI Platform DigitalBase

AIエージェント構築ガイド

コンテンツ

01	Market Report / 市場動向	03
02	オープンモデルの性能とAPIコストの構造	04
03	Lenovo AI Platform DigitalBaseとは	05
04	社内データ連携とAIエージェントで資料作成	06
05	データ連携基盤とAI高度分析・可視化	08
06	Claude Codeと連携してローカル稼働	10
07	モデル別 推奨Lenovoスペック	12
08	Lenovo プロフェッショナルサービス	13

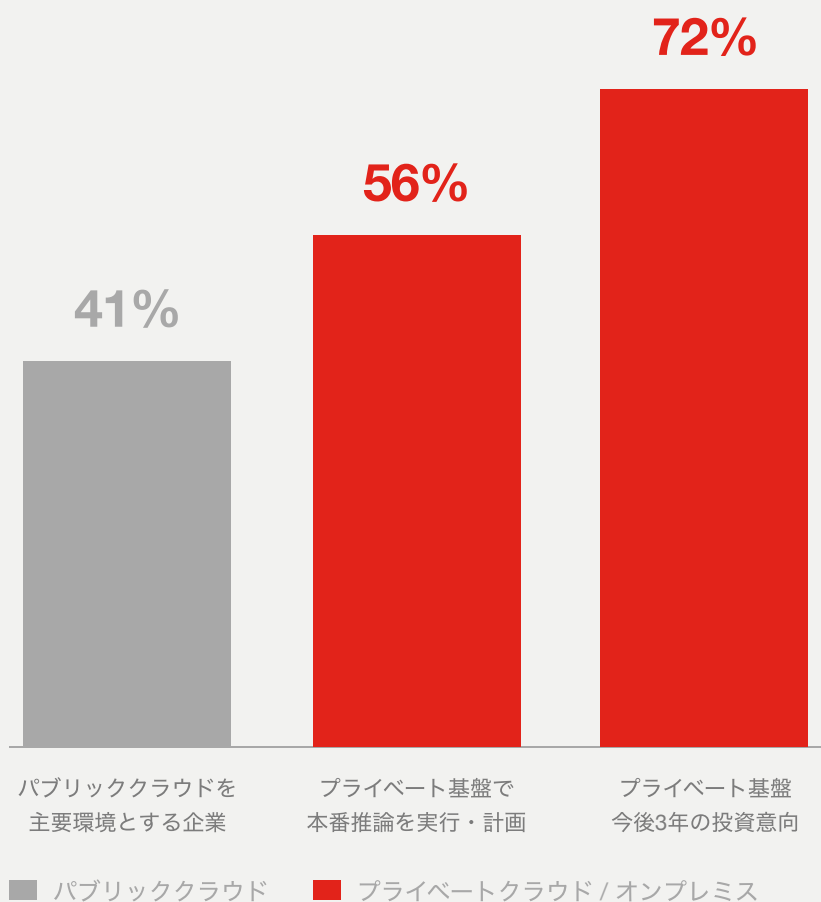


高まるデータ主権意識と、 自社基盤への投資回帰。

2026年、エンタープライズの生成AIは実験フェーズを終え、本番運用の段階に入りました。機密データを自社の管理下に置いたままAIを動かしたいという要請が強まり、推論基盤への投資はパブリッククラウドからプライベート/オンプレミスへと向かっています。

AI実行環境のプライベート/オンプレミス移行

Broadcom 「Private Cloud Outlook 2026」 / 世界8カ国・IT意思決定者1,800名調査



投資意欲

プライベート基盤への投資を増やす企業が3社に2社超へ

今後3年で投資を増やす意向は1年で51%から72%へ上昇。

データ主権

AIワークロードを自社環境へ戻す動きが加速

回帰する企業の43%がAI学習・LLM・推論を移設対象に。

コスト規律

主要環境としてのパブリック利用は縮小へ

主要環境とする割合は56%から41%へ低下しています。

01 コストの予測可能性

IT責任者の62%がAIインフラコストに強い懸念を示し、97%がパブリッククラウド費用に無駄があると回答。うち52%は「無駄が25%超」としています。

02 データ主権・規制対応

機密文書・設計データ・顧客情報をAIに扱わせる場面が増え、データを社外に出さない構成が調達要件に含まれるケースが増加しています。

03 インフラ準備の遅れ

一方でNutanix調査では、82%の企業が「オンプレミスAIを支える現行インフラは十分ではない」と回答。設備の刷新が課題となっています。

出典: Broadcom 「Private Cloud Outlook 2026 — The AI Tipping Point」 (2026年6月、世界8カ国 IT意思決定者1,800名) / Nutanix 「Enterprise Cloud Index 2026」

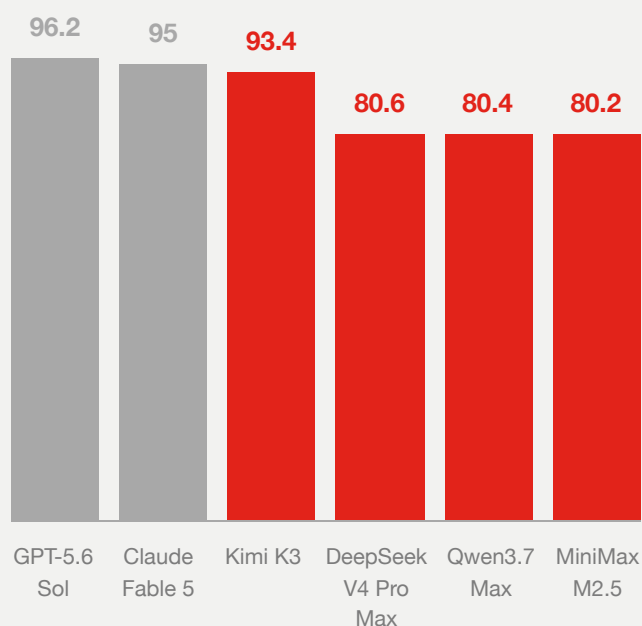
WHY ON-PREMISE NOW

性能差は縮小し、 API単価は上がり続ける。

オープンウェイトモデルは主要ベンチマークでフロンティアモデルに肉薄し、実務用途では十分な精度に達しました。一方でフラグシップAPIの単価は世代を追うごとに上昇しており、実務品質を求めるほど高価なティアを使い続けることになります。

01 オープンモデルの性能

SWE-bench Verified スコア(%) / 2026年7月時点

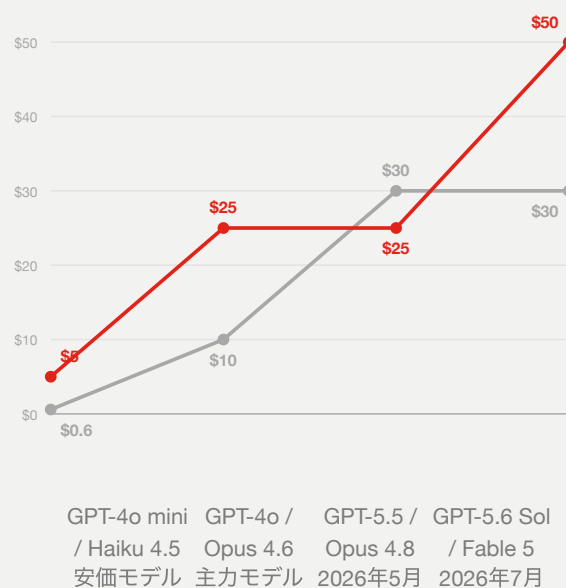


■ クローズド ■ オープンウェイト(自社実行可)

自社実行できるオープンウェイト勢が80%台に到達し、上位のKimi K3は93.4%とクローズド最上位に肉薄。社内文書のRAG、コード生成、データ分析といった実務用途では十分な精度に達しています。

02 ChatGPTとClaudeのAPI単価の推移

出力トークン単価(USD / 100万トークン)



— OpenAI API単価 — Anthropic API単価

上昇が急なのは出力側です。GPT-4o mini \$0.60 から GPT-5.6 Sol \$30 へ50倍、Haiku 4.5 \$5 から Fable 5 \$50 へ10倍。エージェント運用は出力トークンを大量に消費するため、この差が総額に直結します(入力側は \$0.15→\$5、\$1→\$10)。

API依存が高コスト構造になる三つの理由

世代交代のたびに 単価が引き上がる

出力単価はGPT-4o mini \$0.60からGPT-5.6 Sol \$30へ、Haiku 4.5 \$5からFable 5 \$50へ。最新世代を使うほど高くなります。

エージェント運用で 出力量が跳ね上がる

1タスクに数十回の推論が走る自律エージェントは出力トークンを大量に消費し、最も高い単価が最も多く積み上がります。

利用量に比例して 予算が読めない

IT責任者の62%がAIインフラコストを懸念。97%がパブリッククラウド費用に無駄があると回答しています。

出典: SWE-bench Verified — Vals AI 独立ハーネス(GPT-5.6 Sol / Claude Fable 5 / Kimi K3)および IIm-stats ベンダー公表値集計 (DeepSeek-V4-Pro-Max / Qwen3.7 Max / MiniMax M2.5)、いずれも2026年7月時点 / API単価 — OpenAI・Anthropic 公式価格ページ (2026年7月30日時点) / 市場データ — Broadcom 「Private Cloud Outlook 2026」

OVERVIEW

Lenovo AI Platform DigitalBase とは。

社内データとAIエージェントを組み合わせ、資料作成からデータ分析、ローカルLLM稼働までを一気通貫で実現するソリューションです。以下の3ステップで、セキュアかつ高速な業務自動化環境を構築します。

- STEP 1

データ収集・連携

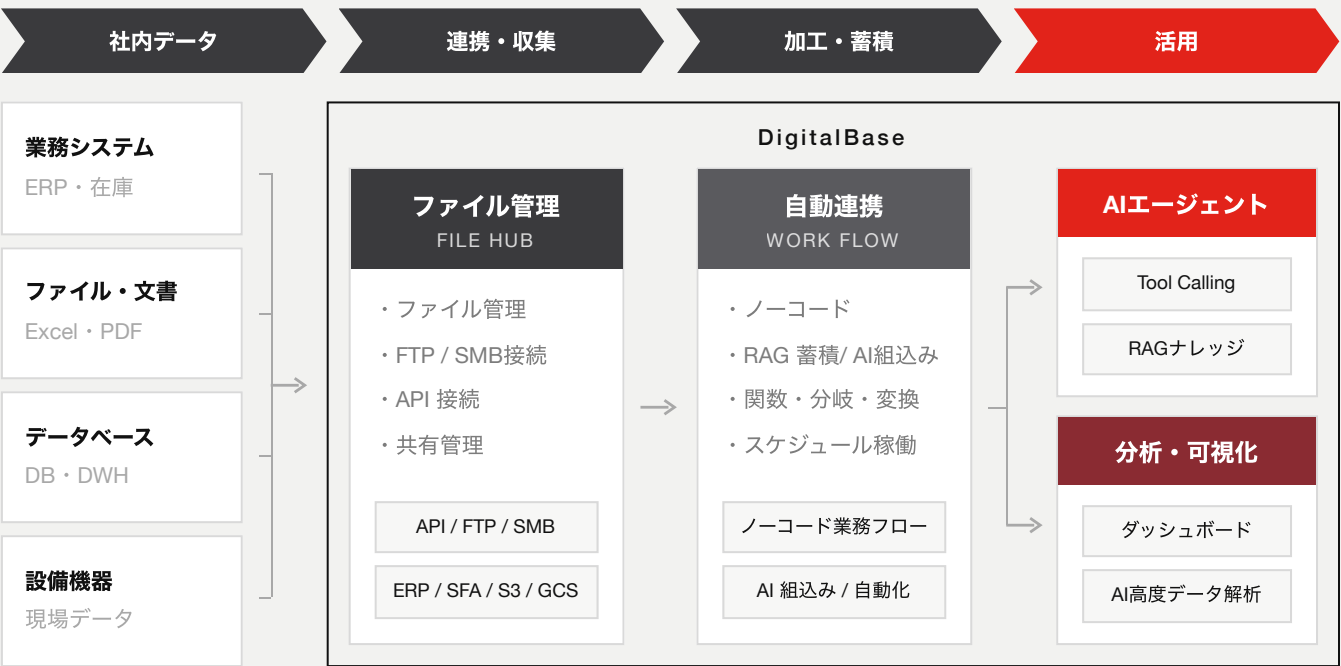
社内の資料・マニュアル・業務データを様々な形式で自動連携し、AIが扱える形に構造化します。
- STEP 2

AIによる分析・生成

蓄積したデータをAIエージェントが分析・分類し、報告書や提案資料を自動生成します。
- STEP 3

ローカル環境で稼働

Lenovoのハードウェア上でモデルを完全ローカル稼働。社内専用のセキュアなAI環境を構築します。



USE CASE 01

社内データ連携とAIエージェントで 資料作成、ナレッジ検索

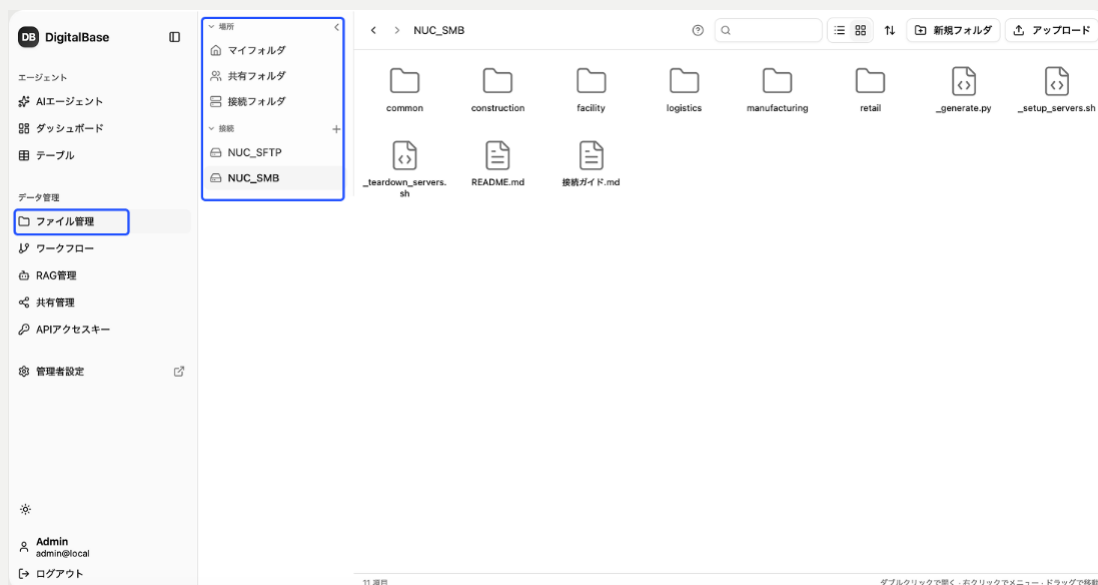
過去資料やマニュアルを様々な形式で自動連携し、報告書・仕様書・提案書・設計書など様々な資料をAIが自動作成します。

01 Tool Calling機能で選択した社内データに自律的な生成AIエージェントを実行



Tool Callingで対象データを指定するだけで、AIエージェントが必要な処理を自律的に判断して実行します。

02 AIエージェントに連携したいデータはファイル管理機能で調整



連携するデータはファイル管理機能で調整。参照範囲を限定し、許可制で安全にデータを管理できます。

03 大きいファイルは事前にRAGナレッジへ登録



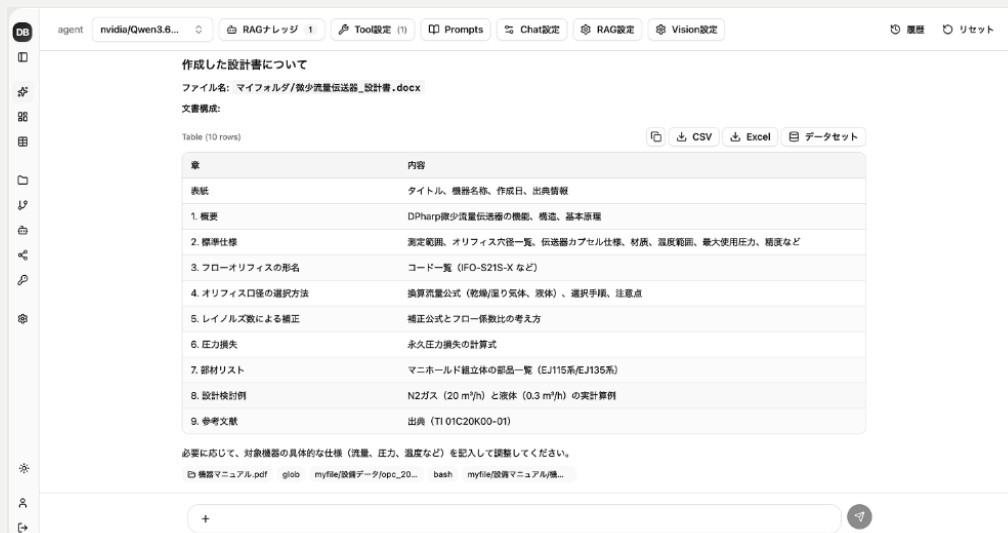
容量の大きいファイルは事前にRAGナレッジへ登録しておくことで、検索精度と応答速度が安定します。

04 ワークフローで自動更新



ワークフローに組み込めば、元データの更新に合わせてナレッジも自動で最新化されます。

05 RAGナレッジやTool Callingで社内資料を元に資料を作成



RAGナレッジとTool Callingを組み合わせ、社内資料を根拠とした文書をAIが作成します。

USE CASE 02

データ連携基盤とAI高度分析

データ分析・可視化

監視データ・センサーデータ・事業部毎データなどを自動で収集・加工・蓄積し、AIが分析可能な状態に整えます。

01 ワークフローでデータを収集



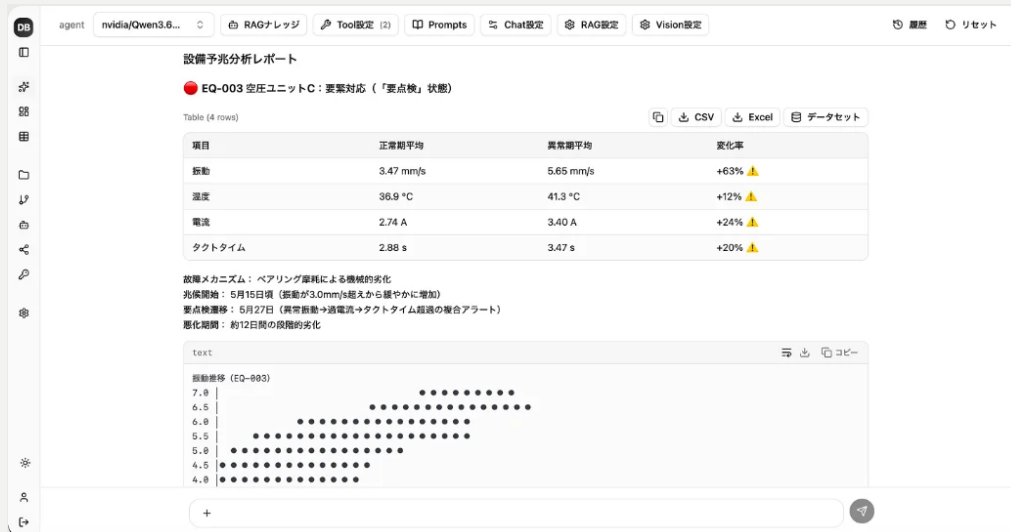
ワークフローを設定するだけで、各種システムからのデータ収集をノーコードで自動化できます。

02 収集したデータを可視化・解析しやすいように型変換



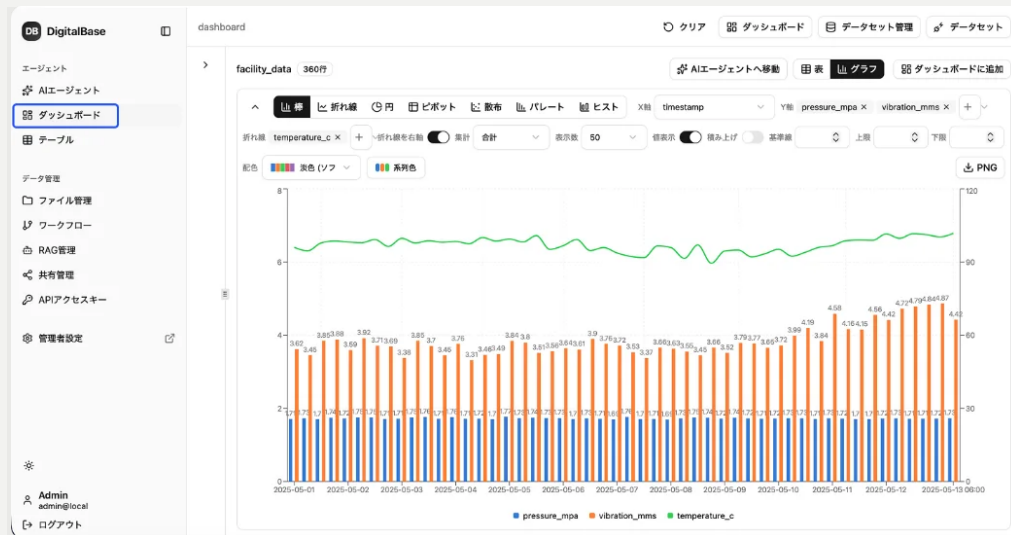
収集したデータは可視化・解析しやすい形式へ自動で型変換し、分析可能な状態に整えます。

03 蓄積したデータをTool設定の高度データ分析と組み合わせて解析



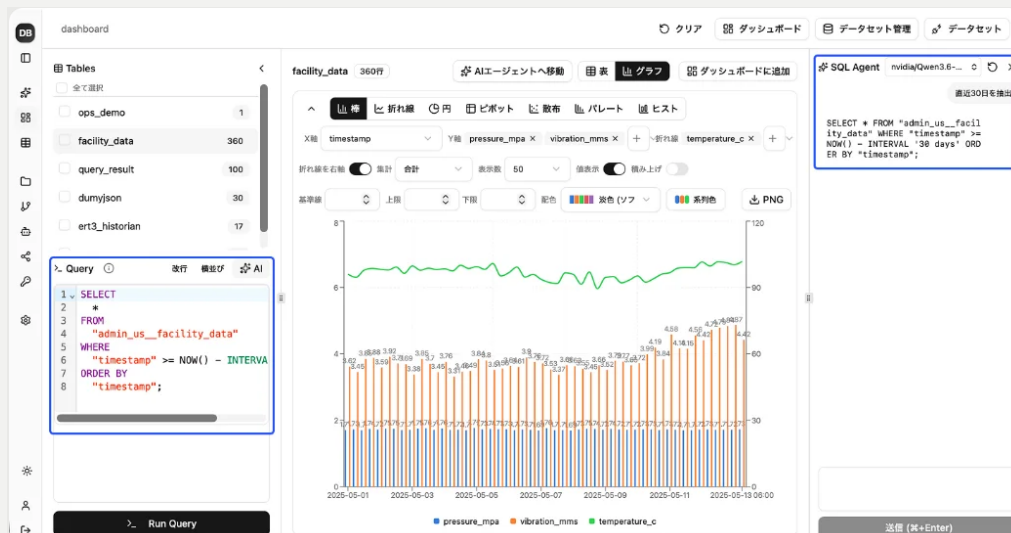
蓄積したデータをTool設定の高度データ分析と組み合わせ、複雑な条件でも解析できます。

04 蓄積したデータをダッシュボードでリアルタイム可視化



ダッシュボードに接続すれば、蓄積したデータをリアルタイムで可視化・監視できます。

05 クエリはAIが自動で作成、保存



クエリはAIが自動で作成し保存。専門知識がなくても繰り返し同じ分析を実行できます。

USE CASE 03

Claude Codeと連携して 専用のローカルClaude Code稼働

設定済みのAIエージェント環境を外部アプリとAPI連携。社内専用のローカルClaude Code環境を構築できます。

01 APIアクセスキーを発行して外部アプリケーションと連携

APIアクセスキーを発行するだけで、外部アプリケーションから同じAI環境を利用できます。

02 APIを発行するとClaude Codeへの連携先が表示される

API発行後は接続情報が画面に表示され、Claude Codeへの連携先としてそのまま設定できます。

03 Claude Codeをインストールして、APIアクセスキーを連携

```
root@localhost:~# curl -fsSL https://claude.ai/install.sh | bash
Setting up Claude Code...

Claude Code successfully installed!

Version: 2.1.220
Location: ~/.local/bin/claude

Next: Run claude --help to get started

Installation complete!

root@localhost:~# export ANTHROPIC_BASE_URL=https://127.0.0.1:8080
export ANTHROPIC_AUTH_TOKEN=sk-2ZCrdz4le0S_Y4fetT3QV_Gf4EWf-w3gAyDt7Za6wVU
export ANTHROPIC_API_KEY=sk-2ZCrdz4le0S_Y4fetT3QV_Gf4EWf-w3gAyDt7Za6wVU
export ANTHROPIC_MODEL=nvidia/Qwen3.6-35B-A3B-NVFP4
root@localhost:~#
```

Claude Code v2.1.220

Welcome back!



nvidia/Qwen3.6-35B-A3B-NVFP4 · API Usage Billing /root

Tips for getting started

Run /init to create a CLAUDE.md file with instructions for Claude

Notes: You have launched claude in your home directory. For the best experience, launch it in a project directory instead.

What's new

Bug fixes and reliability improvements

Added Claude Opus 5 ('claude-opus-5'), now the default Opus model – 1M context, fast mode at \$10/\$50 per Mtok

Added 'sandbox.network.strictAllowList' setting to deny non-allowlisted hosts for sandboxed commands without prompting

/release-notes for more

▲ Both ANTHROPIC_AUTH_TOKEN and ANTHROPIC_API_KEY set · auth may not work as expected

· to use ANTHROPIC_AUTH_TOKEN: Unset the ANTHROPIC_API_KEY environment variable, or claude /logout then say "No" to the API key approval before login.

· to use ANTHROPIC_API_KEY: Unset the ANTHROPIC_AUTH_TOKEN environment variable.

» Try "how do I log an error?"

» manual mode on · ? for shortcuts · ~ for agents

Claude Codeをインストールし、発行したAPIアクセスキーを設定するだけで接続が完了します。

04 APIアクセスキー発行時に設定したRAGナレッジを元に資料作成

```
Last login: Fri Jul 31 02:44:39 2026 from 210.101.53.241
root@localhost:~# curl -fsSL https://claude.ai/install.sh | bash
Setting up Claude Code...
Claude Code v2.1.220
```

Welcome back!



nvidia/Qwen3.6-35B-A3B-NVFP4 · API Usage Billing /root

Tips for getting started

Run /init to create a CLAUDE.md file with instructions for Claude

Notes: You have launched claude in your home directory. For the best experience, launch it in a project directory instead.

What's new

Bug fixes and reliability improvements

Added Claude Opus 5 ('claude-opus-5'), now the default Opus model – 1M context, fast mode at \$10/\$50 per Mtok

Added 'sandbox.network.strictAllowList' setting to deny non-allowlisted hosts for sandboxed commands without prompting

/release-notes for more

▲ Both ANTHROPIC_AUTH_TOKEN and ANTHROPIC_API_KEY set · auth may not work as expected

· to use ANTHROPIC_AUTH_TOKEN: Unset the ANTHROPIC_API_KEY environment variable, or claude /logout then say "No" to the API key approval before login.

· to use ANTHROPIC_API_KEY: Unset the ANTHROPIC_AUTH_TOKEN environment variable.

» RAGナレッジを元に新しい設計書を作成して

● RAGナレッジは、ガス流量計（熱式質量流量計など）の補償公式に関する技術文書です。温度・圧力・温度補償のための換算式が含まれています。

まず、どこにどのような形式で設計書を作成するか確認してください。

Listed 1 directory (ctrl+o to expand)

● RAGナレッジから設計書を作成しますか。作成にあたり、いくつか確認があります。

📄 設計書タイプ

設計書の形式と内容は？

1. 技術仕様書 (Markdown)

流量補償アルゴリズムの数式定義、数値仕様、計算手順をまとめた設計書

2. 実装設計書 (Reactコンポーネント)

流量計算UIのReact実装設計（入力フォーム、計算ロジック、表示部）

3. API設計書

流量計算APIのエンドポイント定義、リクエスト/レスポンススキーマ

4. 両方（技術仕様書＋実装設計）

技術仕様に加えて、実装デザインの詳細も併記

5. Type something.

6. Chat about this

Enter to select · / to navigate · Esc to cancel

API発行時に指定したRAGナレッジを参照し、社内文書を根拠とした資料作成が行えます。

05 社内データと連携した形でローカル環境でClaude Codeが利用可能

```
●では、RAGナレッジに基づき、流量補償アルゴリズムの技術仕様書を作成します。
● Write(technical-spec-flow-compensation.md)
  1. Wrote 415 lines to technical-spec-flow-compensation.md
  2. 流量補償アルゴリズム 技術仕様書
  3. ## 改訂履歴
  4.
  5. | 版数 | 日付 | 変更内容 | 担当 |
  6. |-----|-----|-----|-----|
  7. | 1.0 | 2026-07-31 | 初版作成 - RAGナレッジ（機器マニュアル PDF）に基づく | - |
  8.
  9. ...
  10. ... +486 Lines (ctrl+o to expand)
```

● 完成しました。作成した設計書 /root/technical-spec-flow-compensation.md の概観：

構成

セクション	内容
第2章	全体フロー図（温度判定 → 乾飽型/湿り気体型の分岐）
第3章	乾燥気体変換（式 A） - 基本公式と併用計算
第4章	湿り気体変換（式 B） - 飽和氷蒸気圧 Antoine 近似付き
第5章	圧力比ガードループ（ $P_2/P_1 > 0.75$ チェック）
第6章	例題（ N_2 ガス, 20 m³/h, 30°C, 300 kPa） - 手計算検証済み
第7章	実装上の注意点（エッジケース、精度、性能）
Appendix	ガス密度一覧表、Python疑似コード実装テンプレート

ナレッジからの主要要素

- 式 A(乾燥気体型) → p_{fg} 算出 → $Q_{m,fg} = 0.8734 \cdot Q_{m,fg} \cdot \sqrt{p_{fg}}$
- 式 B(湿り気体型) → Antoine 近似で飽和氷蒸気圧 P_{fs} を計算し、分圧補償
- 圧力比ガードループ → $P_2/P_1 > 0.75$ で checked flow 判定
- 例題検証 - 原文の N_2 ガス例題 ($p=2.248 \text{ kg/m}^3$, $Q_m=26.3 \text{ Nm}^3/\text{h}$, 438.7 Nl/min) を再現検証済

ファイルは Markdown 形式なので、そのままレビューやチーム共有が可能です。

Brewed for 1m 4s

» manual mode on · ? for shortcuts · ~ for agents

社内データと連携した状態のClaude Codeが、完全ローカル環境で利用できます。

HARDWARE

モデル規模別 推奨ハードウェア構成

稼働させるモデル規模と利用人数に応じて、3つの構成をご用意しています。＊構成・世代により
変動の可能性があります。

PGX AI SUPERCOMPUTER



小型AIコンピュータ
NVIDIA GB10 搭載

推奨モデル

30B～70B

Qwen3.6-35B-A3B

用途

通常業務
数人利用

HW詳細

GPU: NVIDIA GB10
(Grace Blackwell)
統合メモリ: 128GB LPDDR5X
メモリ帯域: 273GB/s
サイズ: 150×150×51mm

ThinkStation GPU WORKSTATION



PXシリーズ
NVIDIA RTX PRO 6000 2～4枚

推奨モデル

100B～300B

DeepSeek-V4-Flash-284B-A13B

用途

高度な業務
複数・部門利用

HW詳細

GPU: RTX PRO 6000 Blackwell
VRAM: 192～384GB
(96GB×2～4)
メモリ帯域: 1.8TB/s /枚
GPU間: PCIe Gen5 x16
サイズ: タワー型
(約165×450×550mm)

ThinkSystem SERVER



SRシリーズ
NVIDIA B200 8枚

推奨モデル

300B～

Kimi K3 / DeepSeek-V4-Pro

用途

最先端AIモデル稼働
全社利用

HW詳細

GPU: NVIDIA B200 (Blackwell)
VRAM: 1,536GB (192GB×8)
メモリ帯域: 8TB/s /枚
GPU間: NVLink 5 1.8TB/s /枚
外部接続: 800Gb/s /GPU
筐体: 8U～ラックマウント

導入から運用定着まで、 Lenovo プロフェッショナルサービス

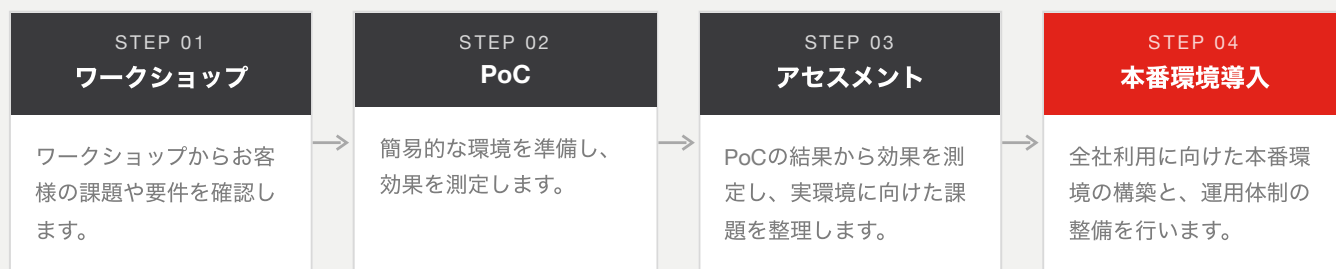
AI基盤の導入では、ハードウェアの調達・構築とAI活用の企画・検証を並行して進める必要があります。レノボ・エンタープライズ・ソリューションズは、その両面をワンストップで支援します。

ハードウェア導入の一括支援

機器調達	設置・キitting	設計・構築	保守・運用
Lenovo製品に加え、他社製スイッチやソフトウェア、備品まで一括で調達	米沢工場で検品・FW更新・OS導入まで実施し、現地作業を最小化	基本設計／運用設計／監視設計、ラッキングから現地構築・管理ツール導入まで	ヘルスチェック、ファームウェア更新、Premier Supportによる24x7対応

AIディスカバリーワークショップ

現行業務の課題を整理し、AIによる業務の効率化をお手伝いします



こんなお困りごとを AIで解決します

- ・知識の集約や活用が進まない
- ・突発的な作業が頻繁に発生する
- ・どのようにPoCを進めればよいかわからない
- ・手間のかかるルーティン作業がある
- ・オンプレミスとクラウドで迷う
- ・AI活用の効果をどう測ればよいかわからない

主な事業領域

AIサービス

AIアドバイザリー／AIシステム導入／ビッグデータ活用

移行

仮想基盤・ストレージ・DBのマイグレーション、本番移行作業まで

Lenovo製品 設計導入

基本・運用・監視設計／機器設置・現地構築／管理ツール導入

プロジェクト管理

定例会・進捗管理・課題管理。PM／PCサービスとして提供

サードパーティ製品

H3C・アライドテレシス等のスイッチ調達から設計構築まで

水冷サーバー

DC準備から直接水冷システム導入、ベンチマークまで

Lenovo AI Platform DigitalBase

問い合わせ先

利用マニュアル

発行

<https://docs.digital-base.co.jp>

2026年8月