AgenticObjects

Server Sizing and Infrastructure Requirements

What to provide, and how large it must be, to install and operate AgenticObjects inside your own environment.

DOCUMENT VERSION	DATE	APPLIES TO
1.1	9 October 2026	On-premise installations

Every capacity figure in this document is derived from measurements taken on a running installation. Where a number rests on an assumption about your usage rather than on a measurement, it is marked as an assumption, so that you can replace it with your own figure.

1 Purpose and how to use this document

AgenticObjects is installed and operated **inside your own environment**. It is not a hosted service, and your data warehouse is never copied out of your network: the product reads from it in place, through a read-only account, and only aggregated results are returned.

This document tells your infrastructure, security and database teams exactly what to provide before an installation can begin: how many servers, how large each one must be, which software must already be present, which network paths must be open, and how much storage to plan for over time.

How to use it

1. Decide how many people will have a login. Take that row from the sizing table in section 4.1 — it gives the processor, memory and disk for the main server.
2. Check the assumptions in section 4.2 against how your organisation actually works. If they differ, tell us and we will recalculate the row for you.
3. If Microsoft Analysis Services cubes are among your data sources, add the second server described in section 6. If they are not, that section does not apply to you.
4. Work through the preparation checklist in section 12 with your teams.

2 What is installed

The whole product runs as a set of **containers on a single Linux server**, started and kept running by Docker Compose. Nothing else is installed on the server itself: the containers bring their own PHP, Python, Java,

database and web server, so the software your other applications use on that machine is never upgraded or replaced.

Container	Technology	Role
Front door	Apache	The only container reachable from the network: terminates HTTPS and routes requests to the two web containers
Web application	PHP 8.4	Both user interfaces, authorisation, the run and conversation records
Live-activity stream pool	PHP 8.4	Keeps the live progress of open chat and run pages flowing, separately from ordinary requests
Agent engine API	Python 3.13	Accepts run orders from the web application
Chat workers (a pool)	Python 3.13	Answer people's questions
Background workers (a pool)	Python 3.13	Execute scheduled and on-demand agent runs, so that long work never delays a person's question
Scheduler	Python 3.13	Places agents' scheduled runs on the queue; never executes them itself
Query builder	Python 3.13	Turns a structured business question into SQL for your data model
Data gateway	Java 21	Executes that SQL against your warehouse, **read-only by design**
Model Studio worker	Python 3.13	Reads the structure of a data source (its tables, columns and their statistics) while a data model is being built
Mail sender	PHP 8.4	Delivers e-mail only when your administrator turns the channel on
Operations helper	Python 3.13	Lets the administrator restart a service from the product's Services screen; the only container allowed to control the others
Product database	MariaDB 11.8	The system of record: users, agents, runs, evidence, audit trail
Queue	Redis 7	The run queue and a short-lived cache, written to disk so a restart loses no queued work

Every container restarts automatically. Every image of a release is pinned by its digest — a fixed fingerprint of its exact content — so an installation runs exactly the bytes we published. Installation, upgrade, backup, restore and the health check are each one command. An upgrade always runs in the same order: **back up, then update the database schema, then start the new images**.

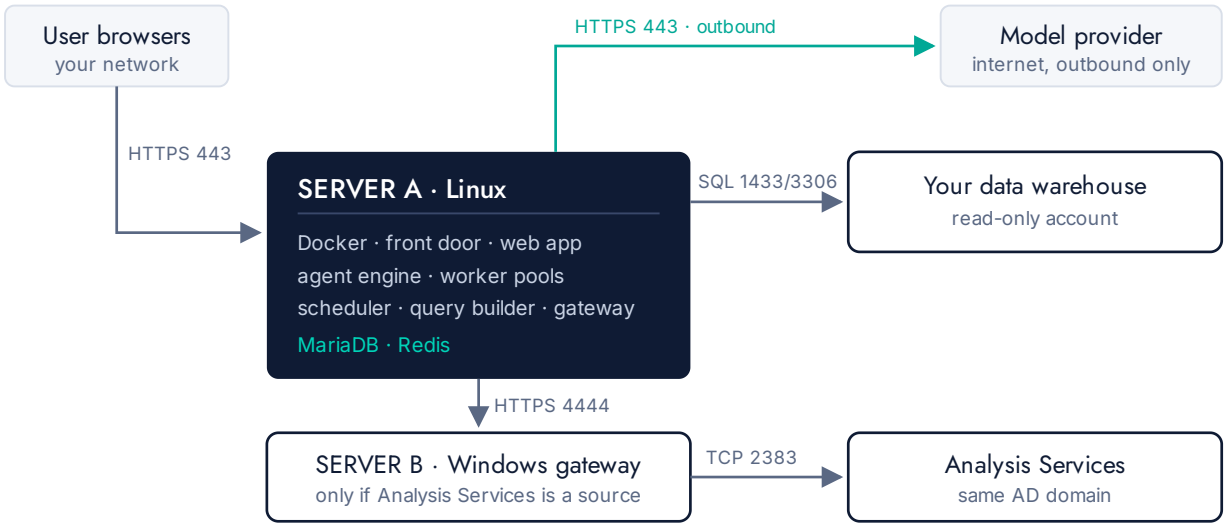
3 Deployment topology

Server A — Linux application server (required). Everything in section 2.

Server B — Windows data gateway (required only if Microsoft Analysis Services cubes are a data source). Two acceptable placements:

- **Option 1 — on the cube server itself.** Add the resources in section 6 to that existing machine. No additional server needed.
- **Option 2 — on its own Windows server.** It must be a member of **the same Active Directory domain as the Analysis Services machines**, because the gateway authenticates to the cube as a domain user.

Your data warehouse stays where it is today. It is not installed on Server A, and Server A does not store a copy of your data. Analytical queries run on your warehouse and only the aggregated result crosses the wire.



Server A is always required. Server B and the Analysis Services path apply only to customers with cubes.

4 Server A — sizing by number of users

4.1 Recommended configurations

Named users	vCPU	RAM	Disk (SSD)	Sustained capacity
25	4	16 GB	100 GB	~125 agent runs/day
50	6	16 GB	150 GB	~250 agent runs/day
75	8	24 GB	200 GB	~375 agent runs/day
100	8	32 GB	250 GB	~500 agent runs/day

Notes that apply to every row:

- **"Named users"** means people with a login, not people using the system at the same time.
- vCPU figures assume modern cores of 2.4 GHz or faster.
- Disk must be SSD or NVMe. The workload is many small writes; 3,000 IOPS is comfortable at every tier.

- These figures assume the data warehouse runs on a **separate** server. If you intend to place the warehouse on Server A as well, its own requirements must be added on top — please tell us and we will re-issue this table.

4.2 The assumptions behind the table

If any of these are wrong for your organisation, the numbers change and we will recalculate:

Assumption	Value used
Peak simultaneous users	20% of named users
Agent runs per user per working day	5
Average duration of one agent run	51 seconds (measured over 30 days, 429 runs)
Longest run observed in the same period	11 minutes
Storage consumed by one run, all records included	44 KB (measured)

4.3 Where the memory goes

This is the working basis for the RAM column, so that it can be checked rather than trusted. The worker figures come from the memory every worker reports about itself once a minute, read over four days of operation:

Consumer	Plan for	Basis
Chat workers (16 started)	2.5 GB	Measured about 65 MB per waiting worker and up to 230 MB per worker while it runs
Background workers (4 started)	1 GB	Same measurement as the chat workers
Live-activity stream pool	1.5 GB	64 processes × about 23 MB, measured
Web application	0.5 GB	16 processes × about 32 MB, measured
Data gateway (Java)	1 GB	Measured 447 MB in use
Agent engine API, query builder, scheduler, Model Studio worker, mail sender, front door, operations helper	1 GB	Measured 66 MB (engine API) and 47 MB (query builder) when waiting; the rest are small single processes
Product database	2 GB up to 50 users, 4 GB above	Database cache recommendation
Queue (Redis)	0.5 GB	Measured 1.2 MB in use
Operating system, Docker, logs	2 GB	Standard allowance

The total is about 12 GB up to 50 users and about 14 GB above, which is why no row of section 4.1 asks for less than 16 GB.

4.4 Throughput: two lanes, several workers

> People's questions and scheduled work run in **two separate lanes, each with its own pool of workers**, so a long scheduled run never delays a person's question. Several workers share each queue safely: the database decides which worker owns a run, so a run is never executed twice, even when a worker stalls and another takes over.
>
> Measured in a load test: 25 chats and 3 scheduled runs arriving at the same moment, with 8 chat workers and 2 background workers taking work and a stand-in model answering each call in 10 seconds, all completed; the slowest chat finished in 79 seconds from question to answer. The installation starts 16 chat and 4 background workers, and the administrator chooses on the product's Service Configuration screen how many of them take work, next to what the last hour measured.
>
> A worker spends most of a run waiting — for the language model provider and for your warehouse — rather than computing. At a concentrated peak the first limit reached is therefore usually the model provider's request rate, not this server. The product paces itself to that rate: a question then waits visibly in the queue rather than failing.

5 Server A — software prerequisites

Requirement	Version	Note
Operating system	A 64-bit Linux on which Docker Engine is supported — for example Debian 13, Ubuntu 24.04 LTS or RHEL 9	Developed and operated on Debian 13
Processor architecture	x86-64	ARM images are planned, not yet published
Container engine	Docker Engine 20.10 or newer, with the Compose plugin (`docker compose`)	The older stand-alone `docker-compose` program is not enough; the installer checks this before it changes anything
Free disk at installation	8 GB on Docker's data directory	The product images take about 2.1 GB; the rest of section 4.1's disk is for data, logs and backups
Transport security	A TLS certificate for the web address	Your own certificate; or, when the server has a public name, one obtained and renewed automatically from Let's Encrypt; or, until either exists, a certificate the installation signs itself. You may also terminate TLS at your own reverse proxy in front
Mail sending account (optional)	An SMTP account, port 587 with STARTTLS by default	Needed only if briefings and notices are to be e-mailed, and for the operations alarm described in section 11

Administrative access (`sudo`) on Server A is required for installation and upgrades only.

WHAT THE PRODUCT DOES NOT NEED

No GPU — the product runs no language model on your server. No Kubernetes. No PHP, Python, Java, database or web server installed on the server itself — all of them live inside the containers. No message broker beyond the bundled Redis, no separate file storage (data models are stored in the database), and **no inbound connection from the internet**.

6 Server B — Windows gateway for Analysis Services

Only required if Microsoft Analysis Services cubes are among your data sources.

Requirement	Detail
Operating system	Windows Server 2019 or newer
Domain membership	**Same Active Directory domain as the Analysis Services machines**
.NET	**Not required to be installed** — the gateway ships self-contained (.NET 8 included, ~81 MB)
Analysis Services client	The native MSOLAP / Analysis Services OLE DB provider must be present
Service account	A dedicated account with the “Log on as a service” right; a setup script creates and configures it
Listening port	TCP 4444 by default (configurable), HTTPS; the Linux host verifies the certificate by its fingerprint
Administrator rights	Required for installation only; the service itself does not need a URL reservation

Sizing:

Placement	vCPU	RAM	Disk
Option 1 — installed on the existing cube server	add 2 cores	add 2 GB	add 5 GB
Option 2 — its own server	2	8 GB	60 GB

The gateway holds no copy of your data. Cube credentials are stored encrypted on the Windows machine itself and are never sent to the Linux server.

7 Network and firewall requirements

#	From	To	Protocol / port	Purpose
1	User browsers	Server A	TCP 443 (HTTPS); TCP 80 optional	Using the application. Port 80 only redirects to HTTPS and lets an automatic Let's Encrypt certificate renew
2	Server A	Data warehouse	TCP 1433 (SQL Server) or 3306 (MariaDB / MySQL), or your instance's port	Read-only analytical queries
3	Server A	Server B	TCP 4444 (HTTPS)	Analysis Services queries
4	Server B	Analysis Services	TCP 2383, or the named instance port	DAX / MDX queries
5	Server A	Internet	TCP 443	Language model provider — only the providers enabled for your installation: `api.anthropic.com`, `api.openai.com`, `open.bigmodel.cn`; or, for the in-country options, `api.bulutistan.ai` (Bulutistan LLM as a Service) or `evren-llmapi.ssyz.org.tr` (EVREN)
6	Server A	Internet	TCP 443	Our image registry, `registry.agenticobjects.ai`, for installation and upgrades. Not needed if we deliver the images on a disk instead
7	Server A	Internet	TCP 443	Operating system and Docker package repositories, for patching
8	Server A	Your mail server	TCP 587 (optional)	E-mail delivery and the operations alarm, if used (section 5)

Flow 5 is outbound only, and it is required: the reasoning models the agents use are reached over the internet. The content of a model request is the question, the data model's structure and the aggregated result rows the engine has already retrieved — never a bulk copy of your warehouse.

> **IN-COUNTRY MODEL OPTIONS**
>
> If model requests must not leave Türkiye, the product offers two separate in-country services, each with its own models:
>
> - **Bulutistan LLM as a Service** — the language model service of Bulutistan.
> - **EVREN** — the defence-industry artificial intelligence service.
>
> With either, the model runs on in-country infrastructure and flow 5 is opened to that service's address only. No GPU is needed on your side in these options either.

No inbound connection from the internet to any of these servers is required.

8 Data sources supported in this version

Source	Status
MariaDB / MySQL-protocol warehouses	Supported and in production use
Microsoft Analysis Services (tabular and multidimensional)	Supported through the Windows gateway (section 6)
Microsoft SQL Server, direct	Supported. Verified against SQL Server 2022 and SQL Server 2025. The connection is always encrypted; if your server uses a certificate it signed itself, tell us before the installation day
PostgreSQL	Planned for a later phase
Other relational databases with a Java (JDBC) driver	Tell us which one; adding it is a driver and a query dialect, and we will schedule it

Every warehouse connection uses a **read-only account**, and the gateway additionally rejects any statement that is not a read, by inspecting the statement itself rather than trusting the account alone.

9 Storage growth

One agent run consumes **44 KB** of database storage, measured across a full installation and including its execution record, its evidence and its audit entries.

Named users	Runs/day	Database growth/year	3-year database size
25	125	~2 GB	~6 GB
50	250	~4 GB	~12 GB
75	375	~6 GB	~18 GB
100	500	~8 GB	~24 GB

The disk figures in section 4.1 also cover the operating system, the product images (about 2.1 GB), container logs (capped at 100 MB per container) and about a week of daily compressed database backups held locally. The product never deletes a backup on its own: move older ones off the machine according to your retention policy.

IMPORTANT

The product **never deletes records — it archives them.** An agent that is retired, a connection that is removed and a run that was cut short all remain in the database, because the audit trail is part of what the product guarantees. Storage therefore grows in one direction only, and the table above already reflects that.

10 Backup and recovery

The installer schedules a **daily database backup**, and every backup is verified after it is written: its checksum is recorded and the file is restored into a separate check copy of the database, so a backup that could not be restored is known the day it is taken. The health check in section 11 reports how old the newest verified backup is. Two things must also leave the machine:

1. **The backup files themselves** (in `/opt/axoria/backups` by default) — copied to storage that does not depend on this server.
2. **The settings file `/opt/axoria/.env`** — it holds the encryption key for stored warehouse credentials. **This key is deliberately not stored in the database.** If the database is restored without this file, every stored warehouse password becomes unreadable and must be re-entered. Keep a copy in your password vault, separate from the database backup.

A restore is one command. It checks the file, first restores it into a trial copy beside the live database, and replaces the live database only when the trial succeeded and an operator has typed the database name twice.

Because nothing is deleted, a restore never has to recover something that was erased by mistake; it only has to recover the machine.

11 Verifying the installation

After installation, one command reports the health of every component and returns exit code 0 when all of them are correct:

```
sudo bash /opt/axoria/health.sh
```

It does not merely check that containers are running. It confirms, among other things, that the database schema has no pending change, that every container passes its own liveness check (for the scheduler and the mail sender: that their loop wrote a heartbeat in the last three minutes), that each worker is actually consuming its queue, that no run is stuck and how old the newest verified backup is. Any failing item is named individually.

The same check runs every five minutes on its own. When an item turns red, or back to green, an **operations alarm** is e-mailed — to our support team by default, so that we often know about a fault before your users do, or to an address you choose. It carries only the health check's own lines — no data from your warehouse and no secret — and it can be switched off at installation.

12 Preparation checklist

Everything below must be in place before the installation day. Items in group F should reach us earlier than the

rest, because they can change the server sizing.

A. Servers

What to prepare	Why it is needed	Usually provided by
The main Linux server, sized according to your user count in section 4.1	This is where the entire product runs, as containers: the web application, the agent engine, the database and the queue. Docker Engine with the Compose plugin must be installed on it (section 5)	Infrastructure / virtualisation team
The Windows gateway server, sized according to section 6 — *or* written confirmation that Microsoft Analysis Services cubes are not one of your data sources	Analysis Services can only be queried from a Windows machine inside the same Active Directory domain as the cubes. If you have no cubes, this server is not needed at all	Infrastructure team

B. Access for the installation

What to prepare	Why it is needed	Usually provided by
Administrative (`sudo`) access on the Linux server	Needed to start the containers, schedule the daily backup and the health check, and later to run upgrades. It is not needed for day-to-day operation	Linux administration team
A domain service account for the Windows gateway, granted the "Log on as a service" right	The gateway authenticates to Analysis Services as this account. A setup script we provide creates and configures it; the account needs read access to the cubes	Windows / Active Directory team

C. Network access

What to prepare	Why it is needed	Usually provided by
The firewall rules listed in section 7	They cover: users reaching the application, the application reaching your warehouse, the two servers reaching each other, the Windows gateway reaching the cubes, and outbound access for the reasoning models, for our image registry, for updates and, if used, for e-mail	Network / security team
Confirmation that outbound access on port 443 to the language model provider is permitted	The agents' reasoning happens at the model provider. Without this path the product cannot answer questions. What is sent is described at the end of section 7	Security team

D. Access to your data

What to prepare	Why it is needed	Usually provided by
A read-only database account on each data warehouse the product should query, with the server address, port and database name	All analytical queries run under this account. Read-only is not only our recommendation — the gateway also rejects any statement that is not a read	Database administration team
Read access to the relevant Analysis Services cubes for the gateway's service account	Same purpose, for cube sources	Analysis Services owner

E. The web address

What to prepare	Why it is needed	Usually provided by
A DNS name pointing at the Linux server, for example `agenticobjects.yourcompany.com`	This is the address your users will open in their browser	Network team
A TLS certificate for that name	Traffic between the browser and the server is encrypted. The certificate is given to the product's front door, or installed on your own reverse proxy if you prefer to terminate encryption there	Security team

F. Information we need from you — please send this first

What to send	Why it matters
The number of people who will have a login	It selects the row in section 4.1, and therefore the size of the server you order
Your own figures for the assumptions in section 4.2, if they differ from ours — in particular how many questions a person is likely to ask per day	These drive both the processor and memory sizing and the storage growth projection in section 9
Which database products your warehouses run on	Section 8 lists what this version can query. If a warehouse runs on something not listed there, we need to know before installation, not during it
Whether the data warehouse will share the main server with the product	The sizing table assumes the warehouse runs elsewhere. If it will not, the server must be larger and we will re-issue the table

AgenticObjects · This document describes the requirements of the version current at the date shown on the cover. Capacity figures are derived from measurements; the assumptions they rest on are stated in section 4.2 so that they can be corrected.