Platform Engineer – Persona Report

(v0.2 · 30 May 2025)

0 | Front-Matter

Field	Details
Document	*RHOAI v3 UI – Platform Engineer (PE) Persona Workshop Report*
Session	"Re-organising RHOAI for GenAI" — Day 4 (Persona & Capability deep-dive)
Date & Time	30 May 2025 · 09:00-12:30 ET (210 min)
Venue / Modality	Google Meet (hybrid) + live Miro board
Transcript Source	Tactiq recorder (.txt) — archived in project Drive
Facilitators	Dash Copeland (PM/UX) · Peter Double (UX)
Participants	Adel Zaalouk · Andy Braren · Ann Marie Fred · Burr Sutter · Dash Copeland · Jason Greene · Jenn Giardino · Jon Nemargut
Authors	ChatGPT-o3 (draft) → to be reviewed by D. Copeland & P. Double
Status	Working draft — v0.2 (updated per final MoSCoW)
Purpose	Document Day-4 outcomes (persona, pains, MoSCoW) for Eng & PM leads

1 | Executive Summary

Day 4 converted sticky-note chaos into a laser-focused Platform-Engineer backlog.

- **Persona crystallised:** "Pat" the Platform Engineer — automation-minded guardian of GPUs, catalogs, and security gates.

- **Pain‑to‑Opportunity:** GPU scarcity, manual audits, rogue assets, and missing governance map to Catalog + AuthZ + Approval workflows.
- **MoSCoW decisions: Must = Authorization + Validated Catalog (incl. Inference Mgmt & Model/MCP approval).** Dashboards, usage limits, token metrics, and gateway integration are **Shoulds** for the November 10 preview.
- **Journey Rev B:** four‑stage pipeline (Request → Provision → Operate → Optimise) annotated with feelings & 40 capabilities.

Next‑step actions 1 Freeze sections 3‑7 by 3 Jun → inline review. 2 Create Jira EPICs for Must/Should items. 3 Stand‑up AuthZ PoC (Keycloak + project‑RBAC) by 17 Jun. 4 Ship "Catalog alpha" with one curated model + MCP stack by 28 Jun.

---

2 | Workshop Highlights

Theme	Highlight
Governance mantra	"Catalog first; anything outside it is a risk." — Burr Sutter
Security flashpoint	Fine‑grained vs. project‑level RBAC → agreed to start pragmatic, evolve later
Catalog clarity	Single dropdown of PE‑approved MCP servers/models demonstrated live
Quote of the day	"Three things: Auth, Catalog, Approval — or nothing ships." — Peter Double

---

3 | Persona Refinement — "Platform Engineer"

3.1 Persona Card

Field	Snapshot (validated 30 May 2025)
Archetype	Risk‑averse, automation‑minded infrastructure steward
Environment	Hybrid ROSA/OpenShift; shared 8‑GPU pool; GitOps first
Role / Day‑job	Sr. Platform Engineer / SRE lead for multiple AI squads

Field	**Snapshot (validated 30 May 2025)**
Signature Quote	“Show me a cost and security plan before you get GPUs.”
Success Metric	≤ 5 days from request → approved inference endpoint

3.2 Key Traits

- Budget‑accountable
- Security‑first
- GitOps loyalist
- Metrics‑driven
- Automation‑minded
- Skeptical of hype

3.3 Goals & Motivations

- Govern catalog of approved assets
- Provide endpoints quickly
- Pass audits automatically
- Expose token usage to teams

3.4 Pain → Design Opportunity

Pain (verbatim)	**Design Opportunity**
“Who is allowed to touch which model?”	**Authorization (RBAC + OAuth)**
Rogue models outside platform	**Validated Catalog + Approval Flow**
“Need an endpoint now!”	**Inference Management (hit‑to‑run)**

Representative quotes & confidence notes unchanged from v0.1.

---

4 | Validated Pain‑Points & Opportunities

Abbreviated: focus shifted to governance & auth (Cost dashboards removed).

4.1 Matrix

Theme	**UX Friction**	**Opportunity**
Governance	Unknown model provenance	PE-approved Catalog
Security	RBAC gaps	AuthZ core (project-level to start)
Delivery	Endpoint lag	Inference Mgmt wizard

Deep-dive sections updated to point at Catalog + Auth templates; Cost sections trimmed.

---

5 | PE Journey Map (Rev B)

Stage	**Trigger & Goal**	**Feelings**	**Key Capabilities**
Request	New AI idea → need GPUs	😐 Guarded	**Catalog dropdown + Request form**
Provision	Approval granted	😰 Stressed	**AuthZ setup + Inference Mgmt**
Operate	Endpoint live	🙂 Watchful	**Token Usage Metrics (Should)** + Quota alerts
Optimise	Governance review	😎 Confident	Versioning + audit logs

---

6 | Capability Backlog (MoSCoW)

Priority	**EPIC**	**Capability**
Must	AUTH-CORE	Authorization: project-RBAC + OAuth scopes + API keys
	CAT-CORE	Validated Catalog & Registry (incl. PE-approved MCP list)
	CAT-SUB	Inference Management (endpoint wiring)
	CAT-SUB	Model Approval workflow
Should	OBS-UI	Dashboards & Consoles (unified view)

Priority	**EPIC**	**Capability**
	USAGE-LIM	Usage Tracking & Soft-Quota limits
	TOK-MET	Token Usage Metrics (time-to-first-token, volume)
	GW-INT	AI Gateway integration (3scale/Kong)
Could	MCP-APR	Ability to approve additional MCPs
	HELP-PE	Inline help / quick-starts for PEs
	LOG-AI	AI-specific Logging & Tracing additions
	MET-SLI	Additional Metrics & SLIs
	CI-EXMP	CI/CD pipeline examples
	VER-ART	Versioned assets registry
Won't (Nov)	AGENT-APR	Approve Agents
	AGENT-MGMT	Full Agent Management
	COST-ADV	Advanced Cost Mgmt dashboards

7 | Next 90 Days – Action Plan

Action	**Owner**	**Due**	**Note**
Finalise backlog → Jira EPICs	Peter Double	3 Jun	Reflect MoSCoW priorities
AuthZ PoC (Keycloak + project RBAC)	Adel Zaalouk	17 Jun	Covers API keys & OAuth
Catalog alpha (1 model + MCP)	Dash Copeland	28 Jun	Dropdown UI + approval toggle
Token metric scraper	Observability team	24 Jun	Expose TTF-token & volume
PE inline docs draft	Andy Braren	20 Jun	For HELP-PE EPIC

Action	Owner	Due	Note
KubeCon demo storyboard	Burr Sutter	10 Jul	Needs Auth + Catalog flows