Who judges the agents?

ORCHESTRATING AI EVALS WITH AIRFLOW

RAFAEL V. PIERRE

AIRFLOW® SUMMIT | ONLINE RECONNECT

ABOUT ME

Passionate about building
Novice snowboarder
Full Stack AI Engineer 🦄
Lead AI Engineer @ Citi
Previously MLE @ Hugging Face, SSA @ Databricks



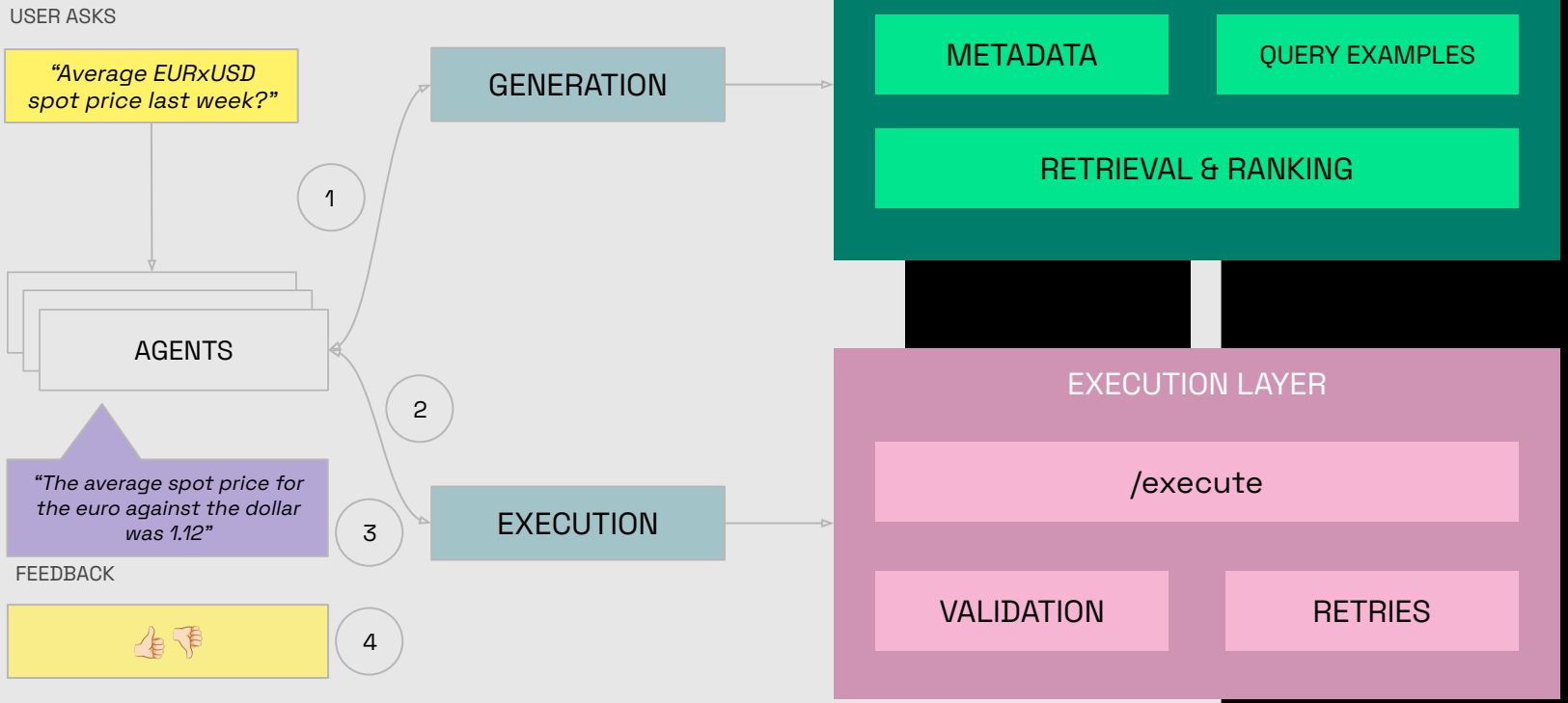
Analytics challenges in F500 land

- ❏ Data silos
- ❏ Sprawling schemas
- ❏ Diverse DBMS engines
- ❏ Niche ones (e.g. kdb+)

Solution: AI



AGENTIC INTELLIGENCE PLATFORM





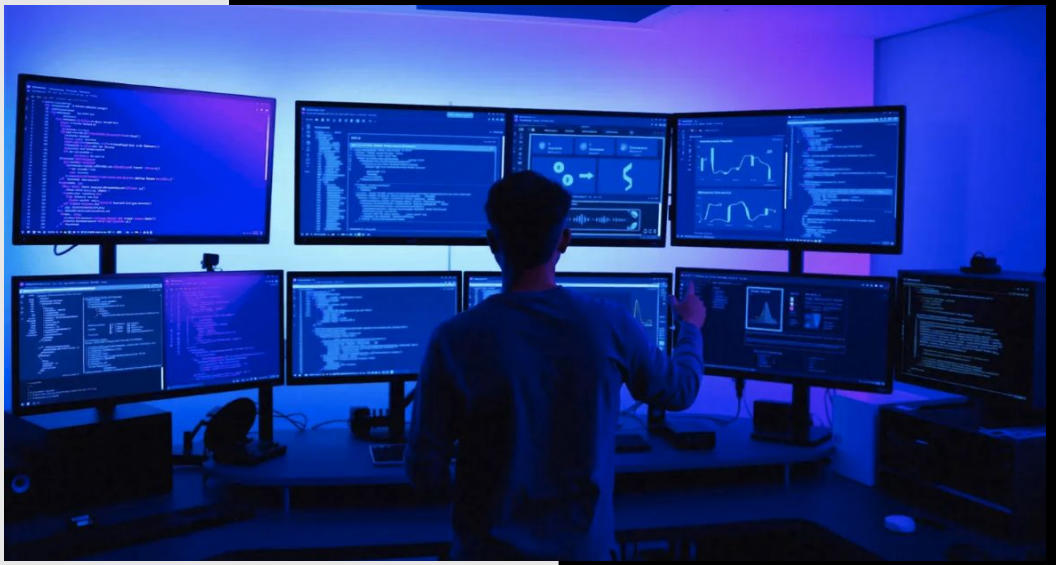
1000+ USERS
AMER
EMEA
APAC

THE PROBLEM(s) WITH AGENTS

But we don’t wanna end up like this!

Agents and LLMs make mistakes

Someone needs to test them…



Solution: AI



LLM AS A JUDGE

Provide a label, a generated example, a rubric and a scale

Compare the source of truth and the generated query

Give it a score based on how close the pair is and a scale (0-5)

Judge outputs a score and the reasoning for it

LLM AS A JUDGE

- ❑ How? **3 times per test**
- ❑ When? **weekly**
- ❑ Why? 🤔

LLM AS A JUDGE

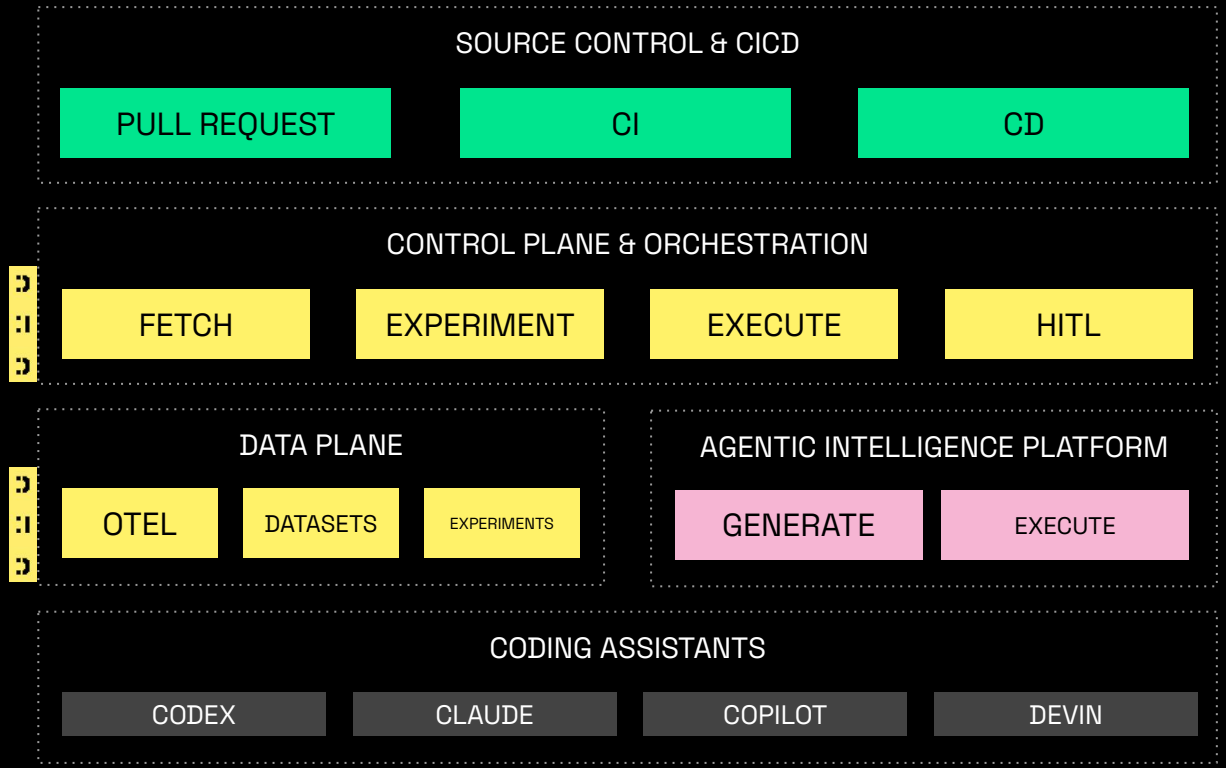
- ❏ How? **3 times per test**
- ❏ When? **weekly**
- ❏ Why? 🤔

Judges are not enough.
We needed a **maestro**.



PEGASUS: OUR EVALS MAESTRO

- ❏ AUTOMATED EXPERIMENTS AND EVALS
- ❏ BUILT-IN OTEL OBSERVABILITY
- ❏ SYNCHRONOUS (ON A SCHEDULE)
- ❏ ASYNCHRONOUS (WEBHOOKS)
- ❏ HUMAN IN THE LOOP


SOURCE CONTROL & CICD
PULL REQUEST
CI
CD
CONTROL PLANE & ORCHESTRATION
FETCH
EXPERIMENT
EXECUTE
HITL
DATA PLANE
OTEL
DATASETS
EXPERIMENTS
AGENTIC INTELLIGENCE PLATFORM
GENERATE
EXECUTE
CODING ASSISTANTS
CODEX
CLAUDE
COPILOT
DEVIN

ACHIEVEMENTS

Some cool outcomes.

TIME SAVINGS

- **80% reduction** in QA hours for backend logic

RELEASE CYCLES

- **Before:** once a month
- **Now:** twice a week

BONUS

- **70% less incidents**
- Target: 90%

NEXT STEPS

There’s always room for more

SCALE

- ❏ Increase dataset sizes
- ❏ Dataset variety
- ❏ Migrate to **LLMBatchOperator**

SCOPE

- ❏ **Now**: Judges analyse query correctness
- ❏ **Next**: Judges analyse query execution results

THE HOLY GRAIL

- ❏ **Frontend Testing**
- ❏ **Self improvement:** Have Pegasus analyse Prompts, OTEL and open PRs
- ❏ **Automate, automate, automate**

THANK YOU!

Stay in touch 📸

rafaelpierre.com