Analytics Vidhya

Langfuse for CI/CD

Instructor

Kartik Nighania

MLOps Engineer at Typewise

Ex-Head of Engineering - YC-backed startup

Certified AWS Cloud and Kubernetes Engineer

Open source contributor

Langfuse for CI/CD

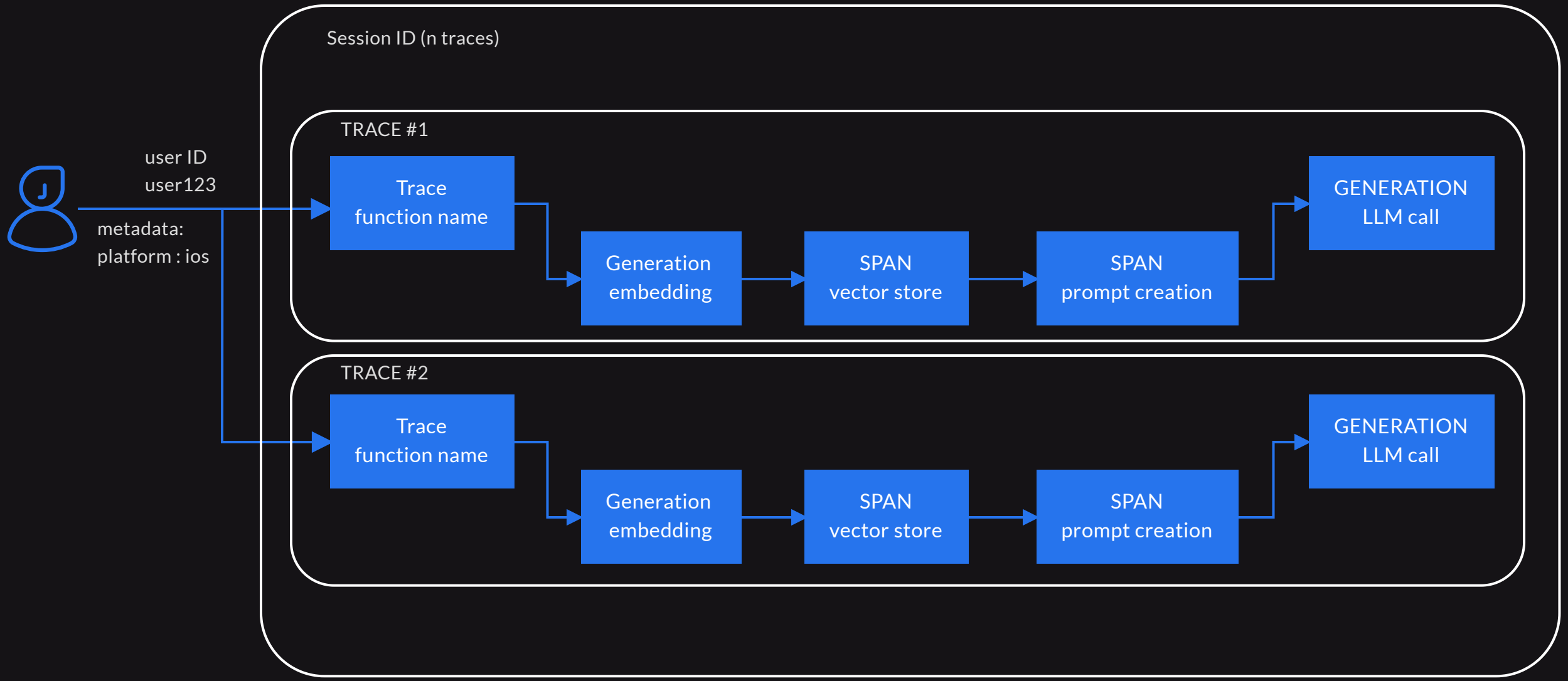
- Langfuse components
- Langfuse integration with Langchain
- Versioning and releases with Langfuse

Analytics Vidhya

Team and data management

- Projects to group experiments
- Users with Single Sign-On
- RBAC support
  - Owner: all ownership
  - Admin: all owner features except project delete and transfer
  - Member
  - Viewer
- Cloud based SaaS solution provided
- On-prem deployment also available
- ISO and SOC2 certifications

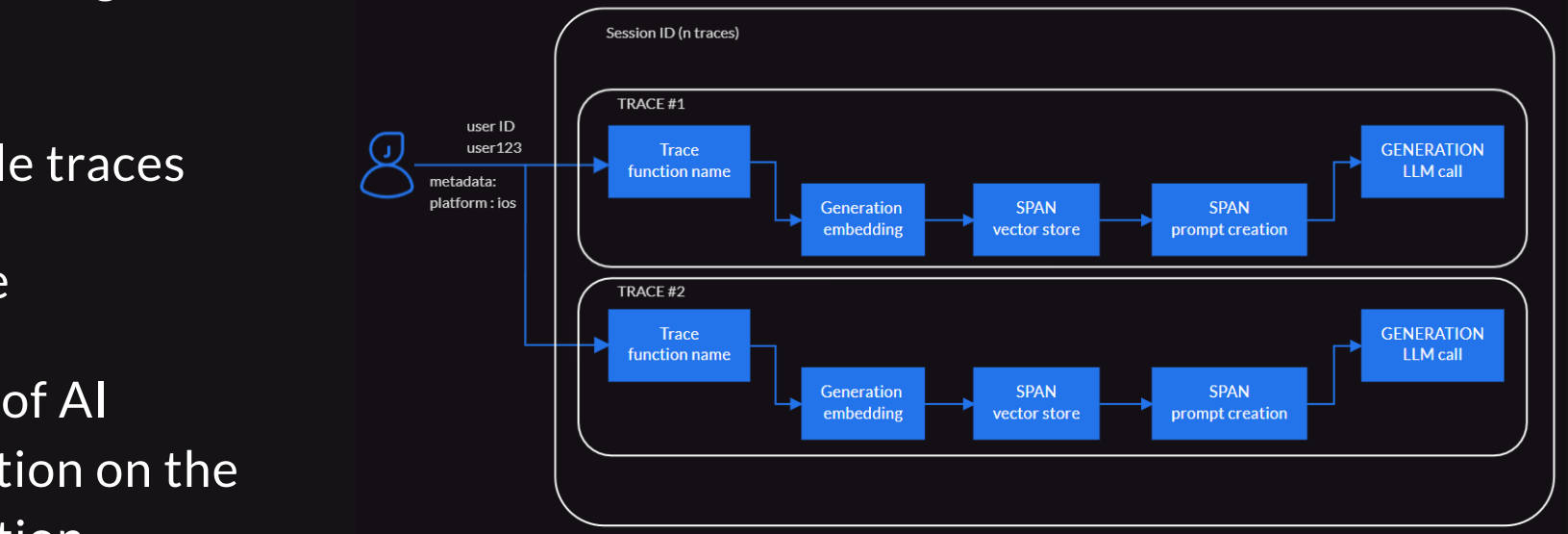
Tracing in Langfuse



Analytics Vidhya

Tracing in Langfuse

- Trace: typically represents a single request or operation
- Session: grouping of multiple traces
- Span: unit of work in a trace
- Generation: log generation of AI models. Additional information on the prompt, model, and completion



Analytics Vidhya

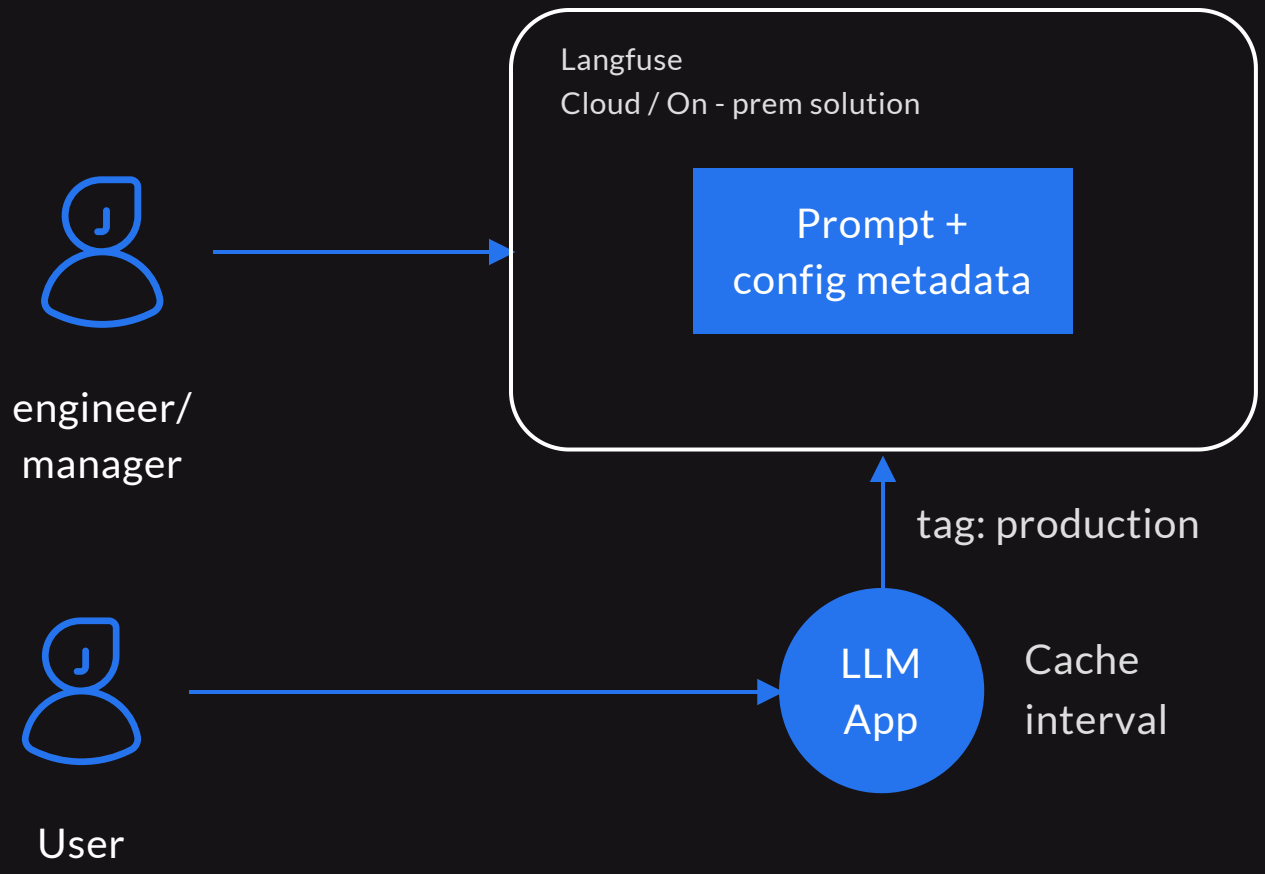
Langfuse Python Decorator

The @observe() decorator can be applied to functions we want to trace.

By default, it captures:

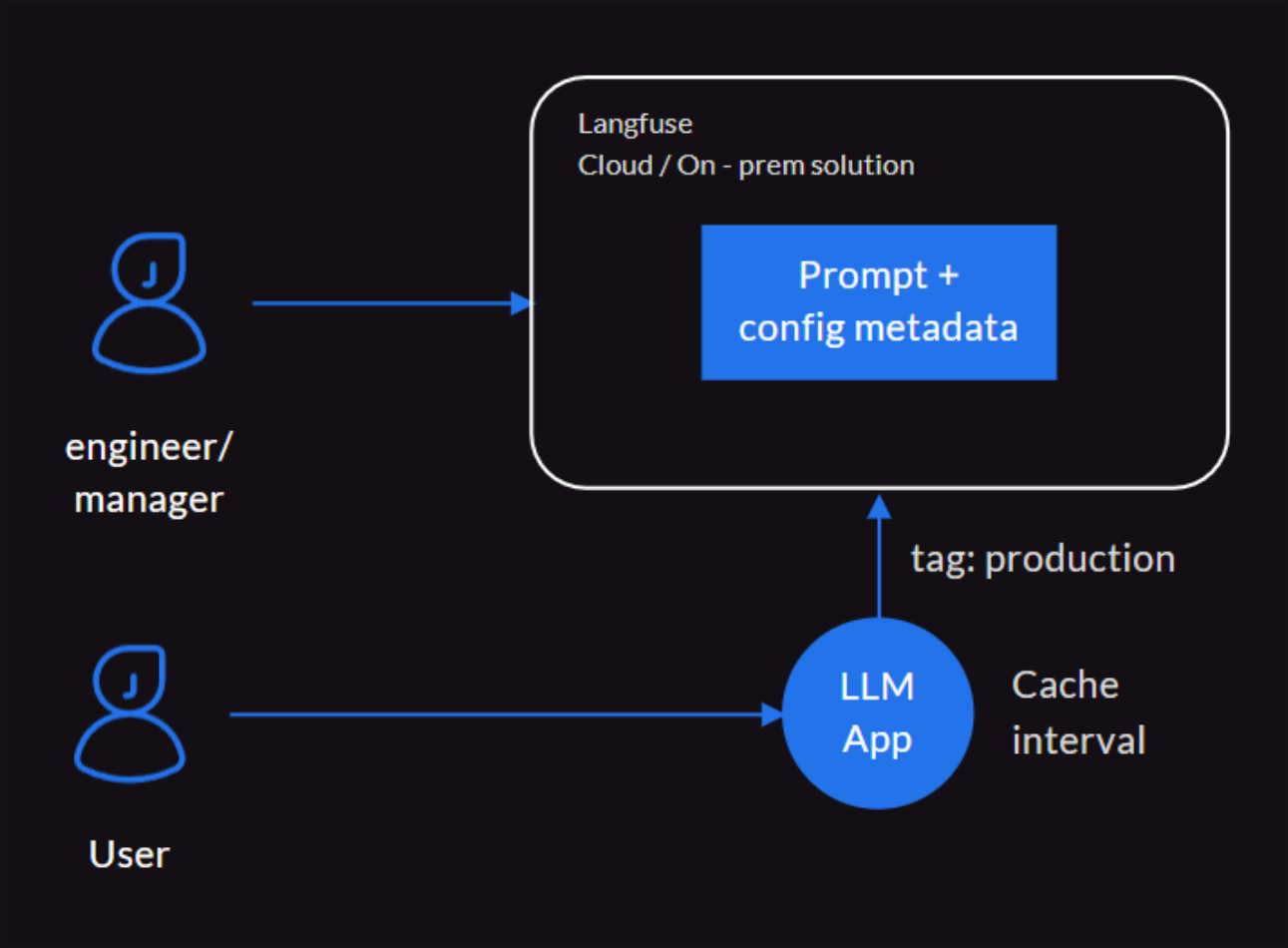
- timings/durations
- function name
- args and kwargs as input dict
- returned values as output
- additional tags, metadata, and IDs can be added
- support for callback in LangChain

Langfuse for Prompt management



Langfuse for Prompt management

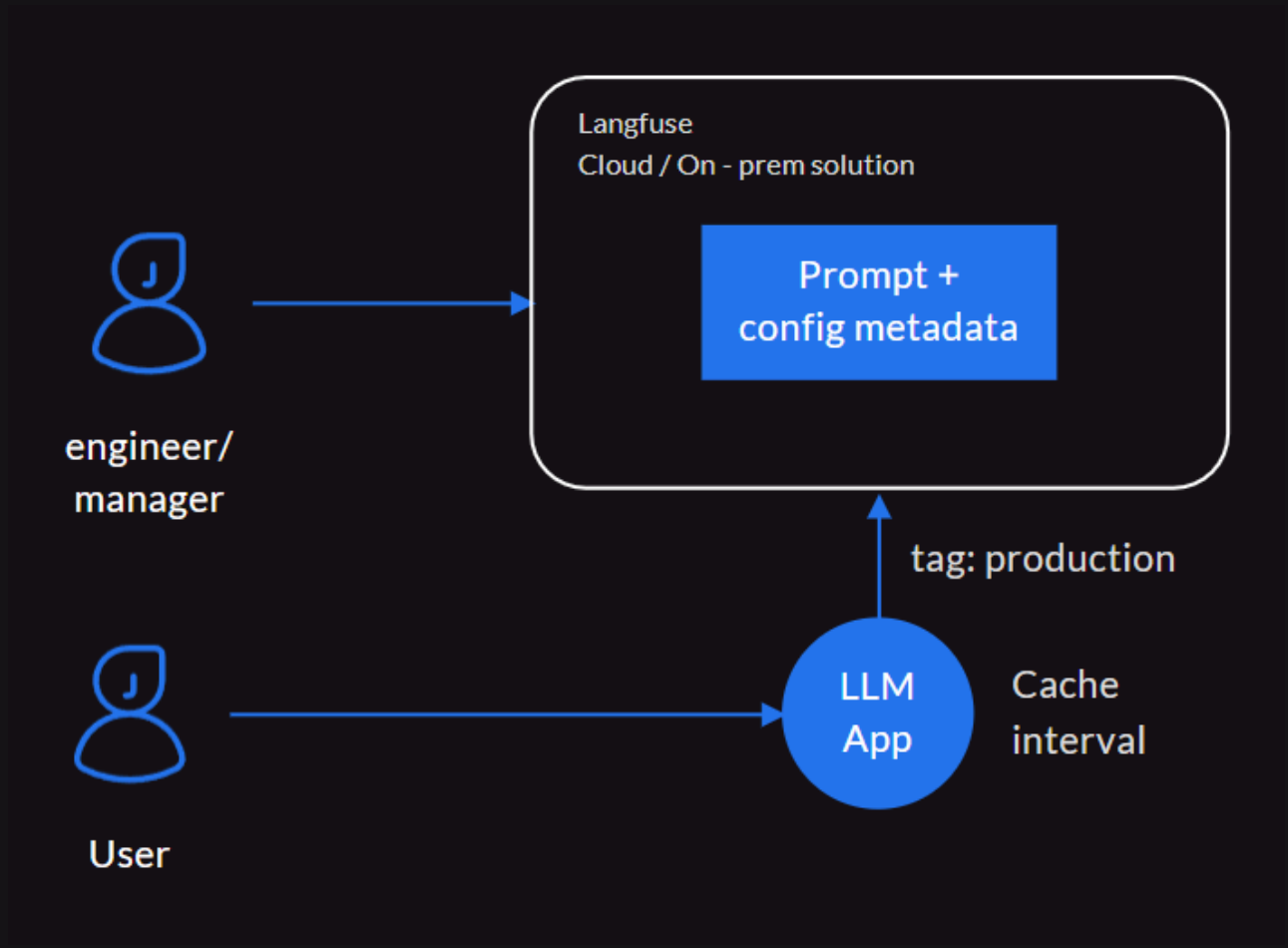
- Decoupling: deploy new prompts without re-deploying your application
- Non-technical users can create and update prompts via Langfuse Console
- Quickly rollback to a previous version of a prompt



Langfuse for Prompt management

Platform benefits:

- Track performance of prompt versions in Langfuse Tracing.



Langfuse Scores and Evaluations

- Scores serve as an object to store evaluation metrics in Langfuse
- Score Components:
  - DataTypes:
    - NUMERIC
    - CATEGORICAL
    - BOOLEAN
  - Attributes:
    - name
    - value / stringValue
  - Additional Metadata:
    - comment
    - source
    - various IDs

Langfuse Scores and Evaluations

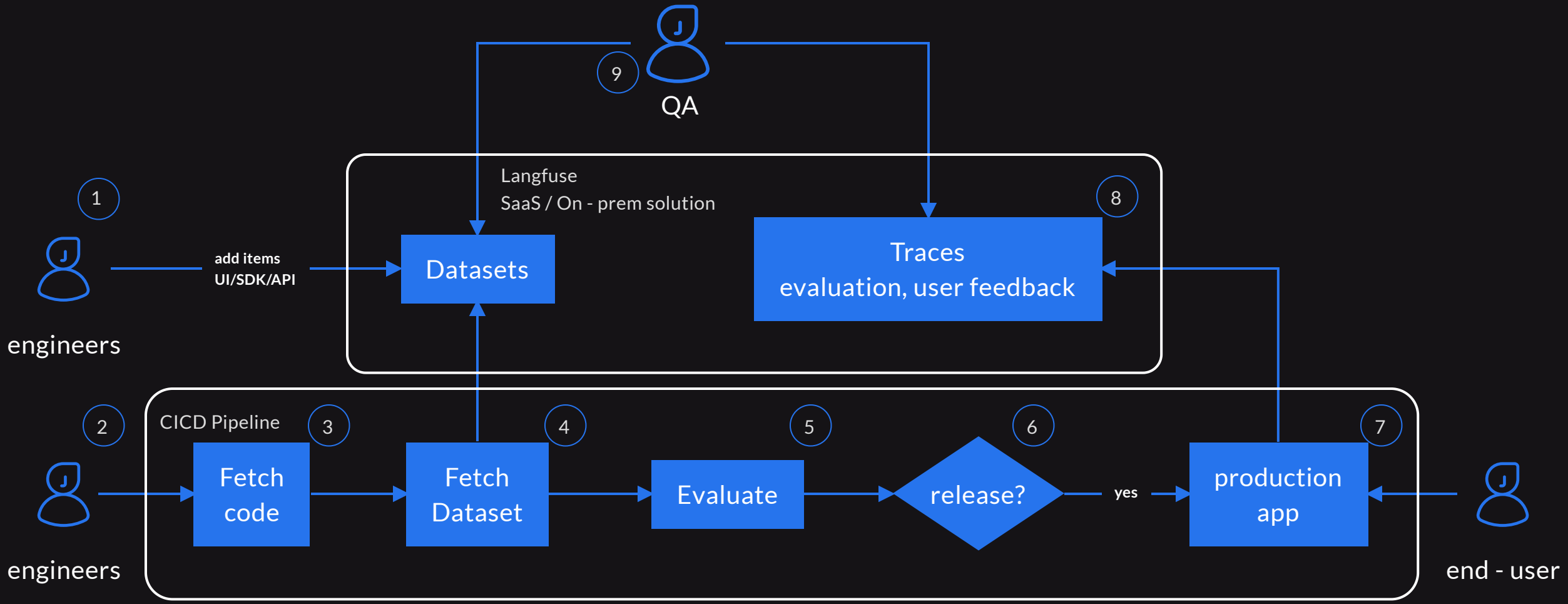
Scores can be of various types:

- Self-annotated in UI:
  - Manual scores given by the team
- User Feedback:
  - Can be integrated in the UI and directly shown to the user
- Model-based Evaluation:
  - We can use LLMs to score trace calls based on varied criteria like toxicity, hallucination, etc.
  - We can create own custom model evaluation
  - These can be run automatically after every trace based on certain filter criteria
  - We can also fetch the runs and upload the computed scores to the UI

Langfuse Datasets

Datasets in Langfuse are a collection of inputs (and expected outputs) of an LLM application.

They are used to benchmark new releases before deployment to production

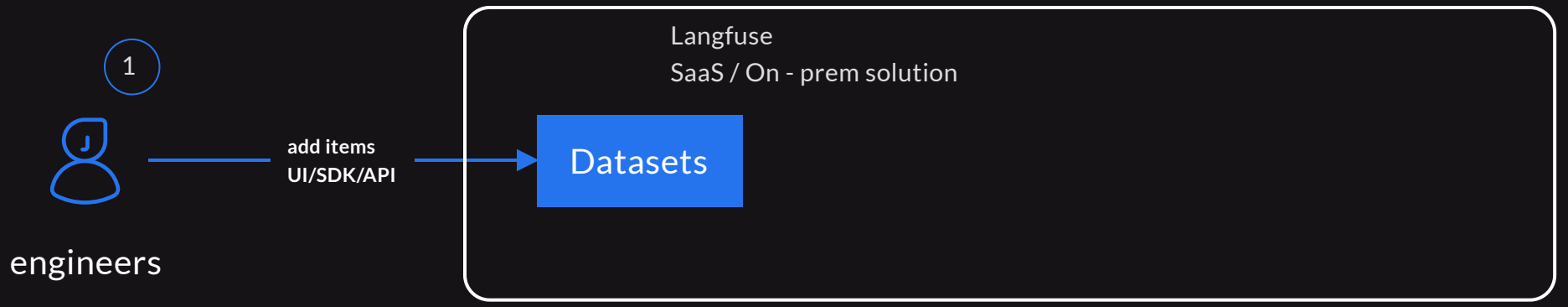


Analytics Vidhya

Langfuse Datasets

Datasets in Langfuse are a collection of inputs (and expected outputs) of an LLM application.

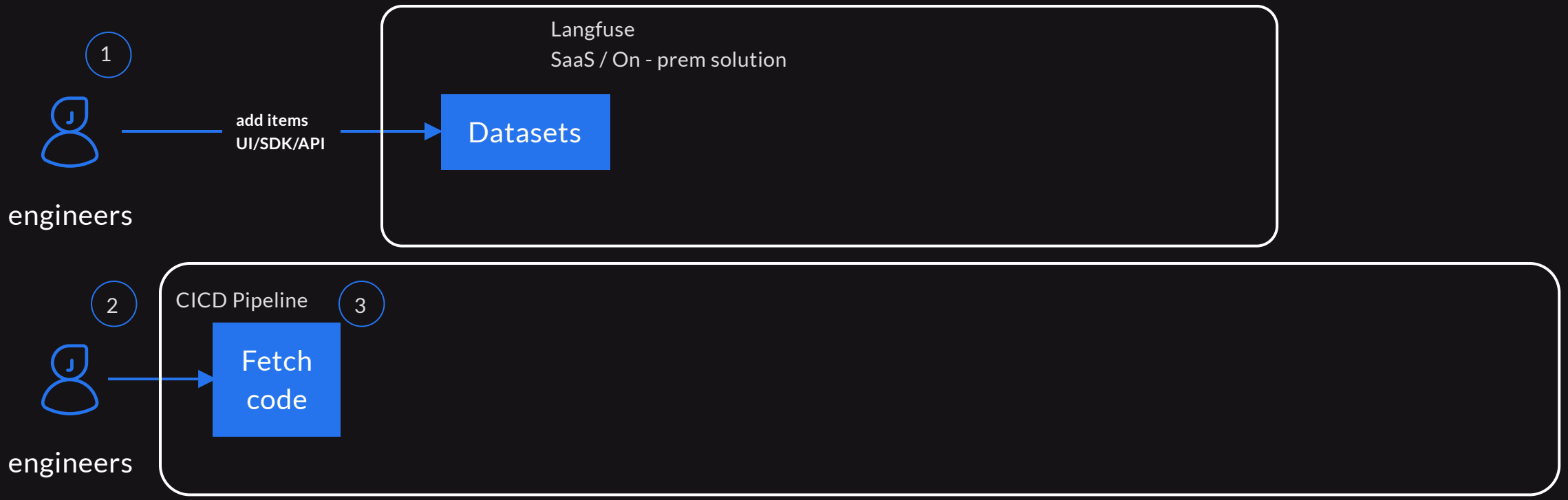
They are used to benchmark new releases before deployment to production



Langfuse Datasets

Datasets in Langfuse are a collection of inputs (and expected outputs) of an LLM application.

They are used to benchmark new releases before deployment to production

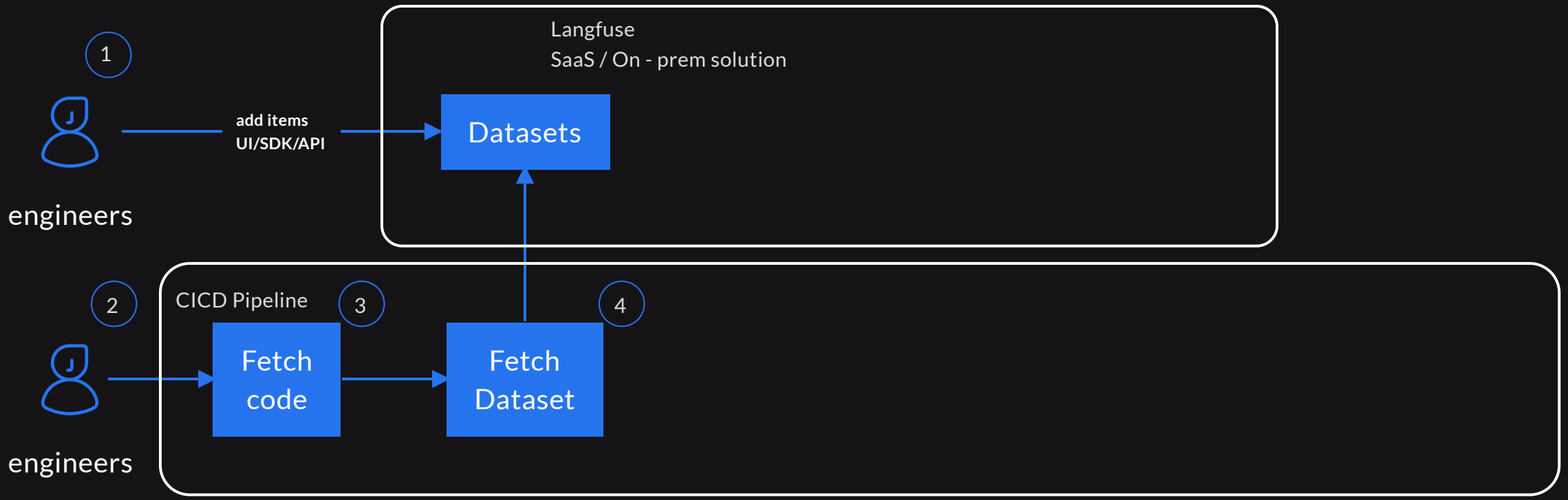


Analytics Vidhya

Langfuse Datasets

Datasets in Langfuse are a collection of inputs (and expected outputs) of an LLM application.

They are used to benchmark new releases before deployment to production

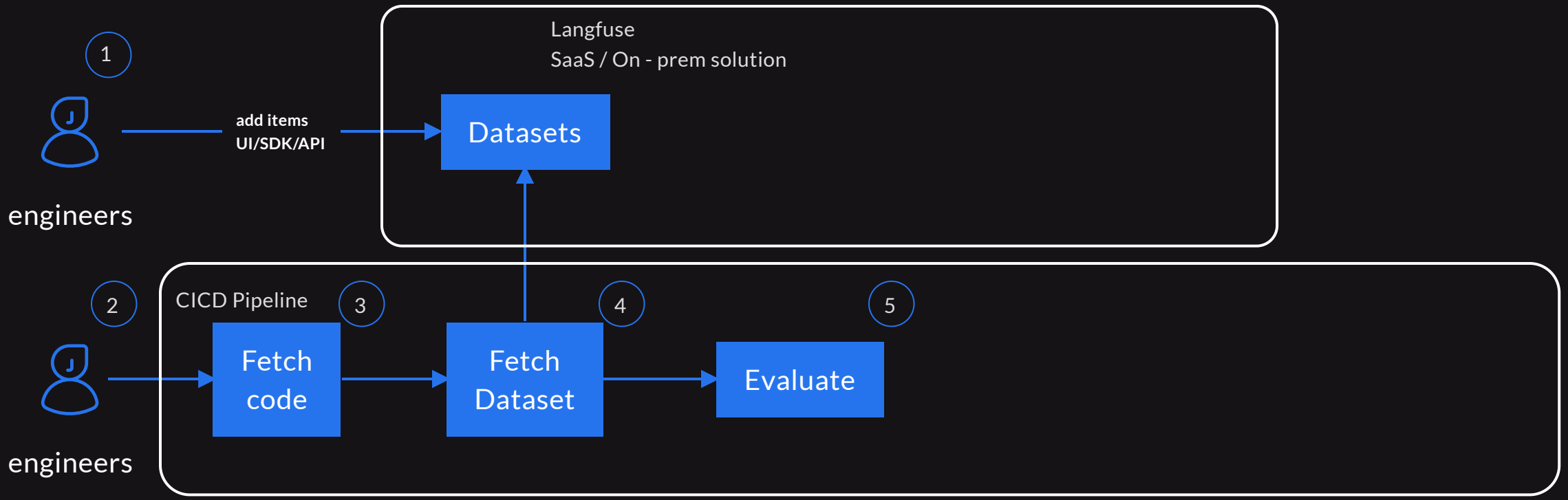


Analytics Vidhya

Langfuse Datasets

Datasets in Langfuse are a collection of inputs (and expected outputs) of an LLM application.

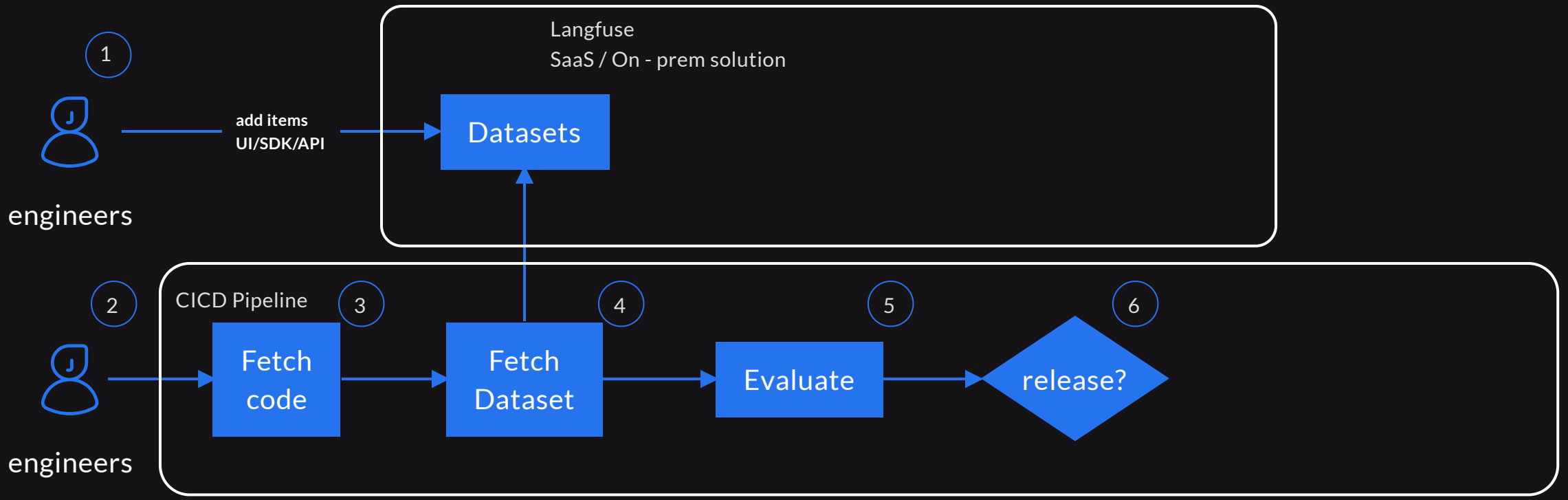
They are used to benchmark new releases before deployment to production



Langfuse Datasets

Datasets in Langfuse are a collection of inputs (and expected outputs) of an LLM application.

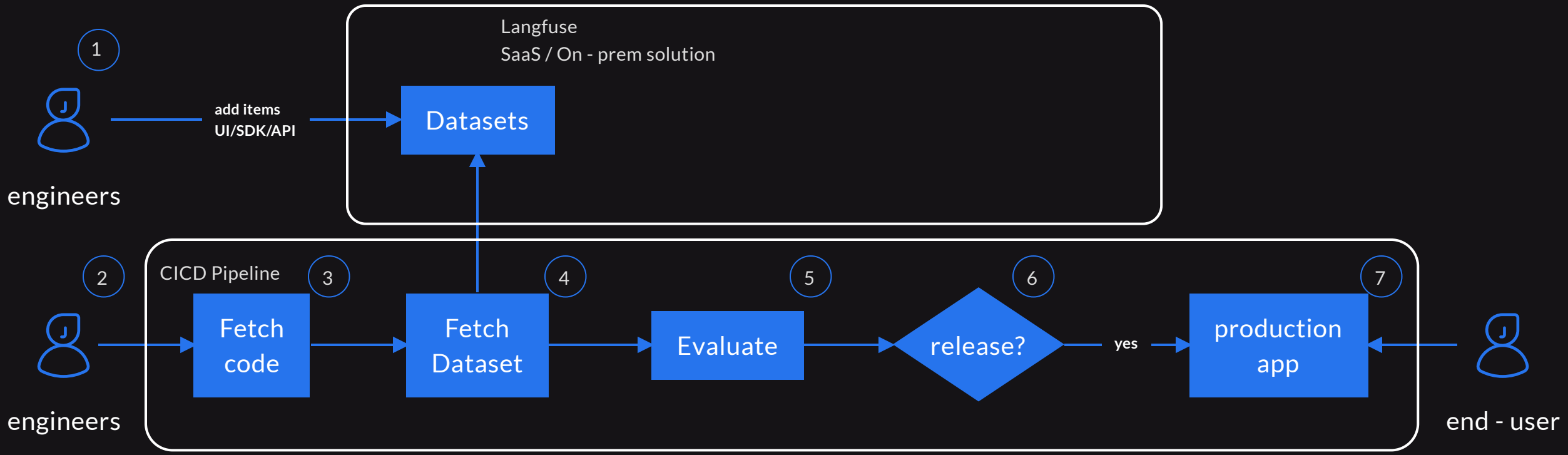
They are used to benchmark new releases before deployment to production



Langfuse Datasets

Datasets in Langfuse are a collection of inputs (and expected outputs) of an LLM application.

They are used to benchmark new releases before deployment to production

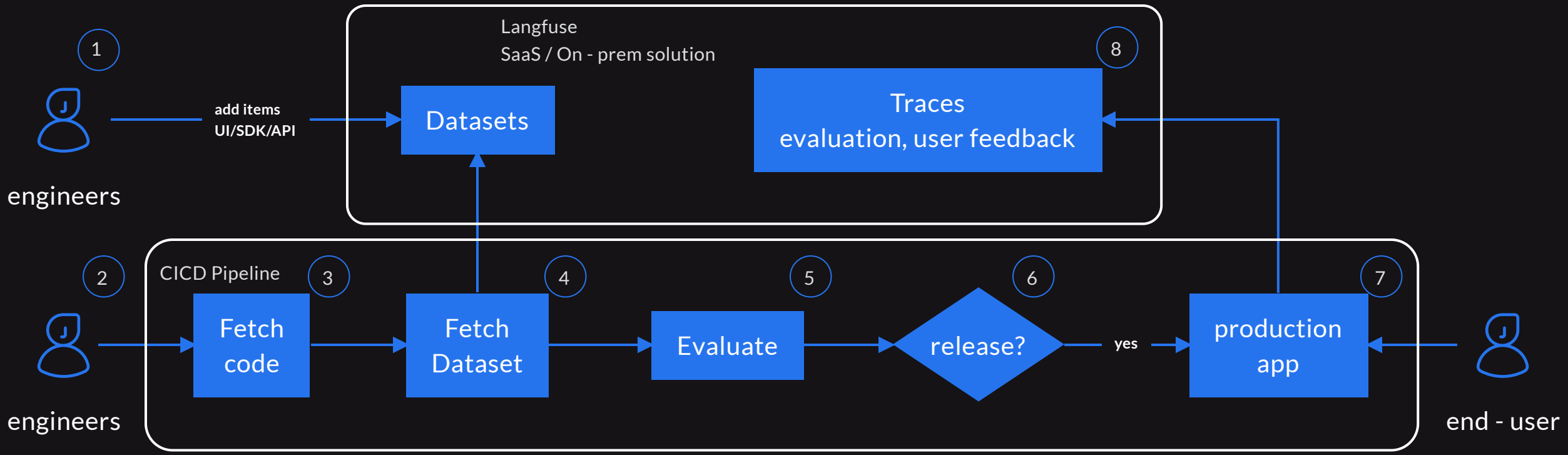


Analytics Vidhya

Langfuse Datasets

Datasets in Langfuse are a collection of inputs (and expected outputs) of an LLM application.

They are used to benchmark new releases before deployment to production

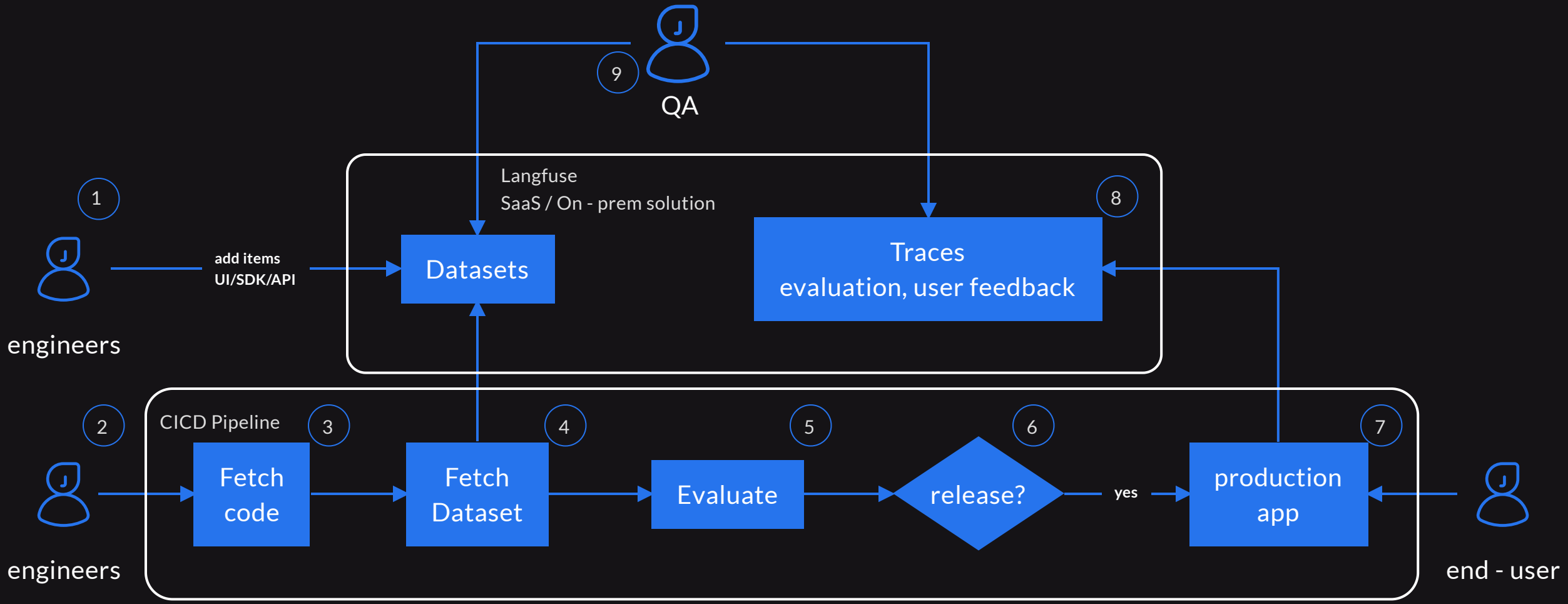


Analytics Vidhya

Langfuse Datasets

Datasets in Langfuse are a collection of inputs (and expected outputs) of an LLM application.

They are used to benchmark new releases before deployment to production



Analytics Vidhya

Thank You