One-Decision Doc: Rate limits for OSS AI Gateway

Intro

Rate limiting support in OSS AI gateway.

Requirements

- [P0] Rate limiting for gateway routes
- [P1] Compatibilities for the existing APIs (mlflow.gateway.get_limits or mlflow.gateway.set_limits)

Differences between OSS and Databricks

	OSS MLflow	Databricks
How to define routes	YAML file (no CRUD APIs)	CRUD APIs
Support for rate limits	No	Yes,
Framework	FastAPI	???
Language	Python	Scala

Example code (for Databrick AI Gateway)

Python

```python
from mlflow.gateway import set_limits, get_limits


set_limits(
    route="my-route",
    limits=[
        # You can define multiple limits on a route
        {
            # 5 calls per user per minute
            "key": "user",
            "calls": 5,
            "renewal_period": "minute"
        },
        {
```

```
            # 50 calls per minute for all users
            "calls": 50,
            "renewal_period": "minute"
        }
    ]
)

get_limits(
    route="my-route"
)
```

Source: https://github.com/databricks/docs/blob/1a11cec796706750db6538013af930895f8071a4/source/machine-learning/ai-gateway/ai-gateway-tutorial.md

Prototype

https://github.com/mlflow/mlflow/pull/9939

[Optional] Non Goals / Out of Scope

List any issues/requirements we are punting on for the future.

Decision 1: How to configure the rate limit settings

Option 1(Preferred): In the config YAML file

Python

```
routes:
  - name: chat
    route_type: llm/v1/chat
    # rate limit parameters
    limit:
      renewal_period: "minute"
      calls: 1
    # or limits? to be consistent with Databricks AI gateway
    # and in case we need to support multiple limits in the future
    # for now, a single limit should suffice.
    # limits:
```

```
    #   - renewal_period: "minute"
    #     calls: 1
    model:
      provider: openai
      name: gpt-3.5-turbo
      config:
        openai_api_key: $OPENAI_API_KEY

  # more routes
  ...
```

Decision 2: Package to use for rate limiting

We use FastAPI in OSS AI gateway. Any useful package we can use for rate limiting?

Stackoverflow: https://stackoverflow.com/questions/65491184/ratelimit-in-fastapi

Package	Stars	Last commit date	Last release on PyPI	Comments
slowapi (preferred)	796	Jun 8, 2023	Apr 7, 2023	Depends on limits. See https://limits.readthedocs.io/en/latest/storage.html for supported storage backends.
fastapi-limiter	324	Nov 16, 2022	Jun 6, 2023	

Decision 3: Should we support `mlflow.gateway.get_limits`?

Option 1(Preferred): Yes

READ operation is easy to support. `mlflow.gateway.get_route` is supported in OSS (while `mlflow.gatway.create_route` and `mlflow.gateway.delete_route` aren't). We should follow the same pattern.

Decision 4: Should we support `mlflow.gateway.set_limits`?

Option 1(Preferred): No

As discussed in the previous decision, no support for `set_limits`.

Decision 5: Should we support the root level config?

Option 1(Preferred): No (but we're open to a request for supporting it)

It's possible to support the root level config as shown below, but we can start with per-route configs:

```python
# Specify limits here.
# This applies to all the routes.
limits:
  - renewal_period: "minute"
  - calls: 1
routes:
  - name: chat
    route_type: llm/v1/chat
    model:
      provider: openai
      name: gpt-3.5-turbo
      config:
        openai_api_key: $OPENAI_API_KEY

  # more routes
  ...
```