Review of Project Telemetry Data Collection and Usage

The following is meant to assist with a review of the project in connection with the project entity’s Telemetry Data Collection and Usage Policy. Participants in the project are requested to provide responses to the following questions, regarding telemetry that is collected by the open source project and for use by the open source project community.

Project: MLflow

Completed by (name and email): Serena Ruan serena.ruan@databricks.com
Date: Jun 4, 2025

1. Specific data proposed to be collected

- Please fill in the following table with details on the specific data elements to be collected.

Data element *e.g., software version; operating system; etc.*	**Could be personal info?** (Yes/No)	**Could be tracking or unique identifier?** (Yes/No)	**Could be end-user / sensitive / business data?** (Yes/No)	**Notes**
Unique session ID	No	Yes	No	A randomly generated, non-customer/non-personally identifiable UUID is created for each session—defined as each time MLflow is imported; a new session (and thus a new UUID) is generated if MLflow is reloaded or the REPL is restarted.
MLflow version	No	No	No	Version of MLflow in use, assuming users are using the public release with no customization (e.g. 2.22.0)
Python version	No	No	No	Version of Python in use (e.g. 3.10.16)
Operating System	No	No	No	The operating system on which MLflow is running (e.g. macOS-15.4.1-arm64-arm-64bit).
API name	No	No	No	Record the API name if

				those APIs are invoked (e.g. log_model, autolog, etc.). See this tab for a full list of API names that'll be recorded
Metadata about GenAI functions usage	No	No	No	See below table for what metadata is logged
Backend store	No	No	No	Record the name of the backend store that's used (FileStore, SqlAlchemyStore, RestStore)

Metadata for APIs that are logged:

API name	Data element	Type	Possible values
log_model	flavor	Enumerated categorical value	catboost, diviner, dspy, h2o, johnsnowlabs, keras, langchain, lightgbm, llama_index, onnx, openai, paddle, pmdarima, promptflow, prophet, pyfunc, pytorch, sentence_transformers, sklearn, spacy, spark, statsmodels, tensorflow, transformers, xgboost
	model	Enumerated categorical value	string, PythonModel, ChatModel, ChatAgent, ResponsesAgent, object
	pip_requirements	Boolean	True, False
	extra_pip_requirements	Boolean	True, False
	code_paths	Boolean	True, False
	params	Boolean	True, False
	metadata	Boolean	True, False
	status	Enumerated categorical value	success, failure
autolog	flavor	Enumerated categorical value	anthropic, autogen, bedrock, crewai, dspy, gemini, groq, keras, langchain, lightgbm, litellm, llama_index, mistral, openai, paddle,

			pydantic_ai, pyspark.ml, pytorch, sklearn, smolagents, spark, statsmodels, tensorflow, transformers, xgboost
	disable	Boolean	True, False
	log_traces	Boolean	True, False
	log_models	Boolean	True, False
genai.evaluate Scorers	scorers	List of enumerated categorical value	Possible values for the list element: answer_correctness, answer_relevance, answer_similarity, faithfulness, relevance, custom_scorer
	predict_fn	Boolean	True, False
	status	Enumerated categorical value	success, failure

- If there is public documentation on the project site describing this data, please also provide a URL to that documentation:
    - N/A at the moment. We will add a page for this prior to the release of the telemetry collection on mlflow.org website.

2. User notification and opt-in

- Please describe how users are *notified* (1) that telemetry will be collected; and (2) which specific data elements will be collected:
    - MLflow will publish clear documentation on its official website (mlflow.org) outlining:
        - How users can enable or disable telemetry collection
        - The specific data elements that are collected
- If there is public documentation on the project site or in the project source code with the particular notices, please also provide a URL:
    - The documentation will be published on MLflow's official website at mlflow.org
    - The source code for the telemetry instrumentation will be publicly available in the MLflow GitHub repository at https://github.com/mlflow/mlflow, excluding any security-sensitive configurations necessary for the secure collection and processing of telemetry data
- Is the telemetry only collected and shared if the user *voluntarily opts into* collection? (As opposed to, collecting data unless the user opts out.)
    - Telemetry collection will be enabled by default; however, users will have the option to disable it at any time—before or after MLflow is imported—by setting an environment variable. This configuration mechanism will be clearly documented in the official MLflow documentation.
- Is the user able to select between only sharing certain data elements, but not others?
    - No, at this time, all data described in the table above will be collected and transmitted as defined.

- How does notification and opt-in function if the software is installed and runs in a *fully-automated* installation (e.g., where there is no user who sees the notice and affirmatively clicks the “I consent” button)? Would telemetry data ever be collected in this type of scenario?
    - MLflow will automatically detect common automated environments (e.g., GitHub Actions, Jenkins) using standard environment variables and will disable telemetry collection in those contexts
    - Release notes and product documentation will inform users about the telemetry collection, and users will be able to opt out by following the guidance provided in the MLflow documentation.

3. Storage and use of collected data

- Please describe where the telemetry data is collected and stored (e.g., on which servers / repos; where they are physically located, if known):
    - Telemetry data is stored in a Unity Catalog (UC) table within a Databricks workspace owned and maintained by the MLflow open source maintainers. The physical location of the underlying infrastructure depends on the Databricks cloud environment in use. Ultimately, the data is stored in Amazon S3, in accordance with Databricks’ standard security and compliance practices. Access to this workspace is restricted to authorized MLflow maintainers.
- Who administers and has access to the servers where data is stored?
    - Access to the telemetry data (the UC table) is limited to a small group of MLflow maintainers who are authorized and approved to administer the workspace. These individuals are responsible for maintaining the integrity and security of the data.
- Are all participants in the project community permitted to view and use the collected telemetry data? Or only particular participants / community members?
    - MLflow will publish dashboards on its official website mlflow.org to share aggregated insights derived from telemetry data, making the analysis results accessible to all community users. These dashboards will be updated on a regular cadence to ensure timely visibility into usage trends. However, the raw telemetry data itself will not be publicly available for direct access or querying.

4. Security mechanisms

- Is there a documented way that an organization could block the telemetry data from being collected from their systems, even if one of their employees inadvertently approves it?
    - Organizations may choose to block telemetry collection by restricting network access to the known endpoint used by MLflow to transmit telemetry data. However, MLflow does not plan to provide detailed firewall configuration instructions for the wide range of public and private cloud environments in which it may be deployed.
- Is there a reasonable possibility that including telemetry functionality opens up security vulnerabilities?
    - No. Data is transmitted using standard HTTP POST requests to a fixed, well-defined endpoint. The implementation adheres to security best practices to minimize risk, and no executable or untrusted code is received or run as part of telemetry collection.
- If so:
    - What steps are taken to mitigate this?
        - N/A
    - If a user does not opt into telemetry data collection, would this risk be fully mitigated?
        - N/A

5. Future changes

If the project plans to extend the scope of telemetry collection in the future (e.g. to begin collecting new types of data), or if the answers given above would change, please update this form and notify us so that we can quickly review the updated proposal.

2025-06-24 comments from LF review

The proposed telemetry data collection and usage is generally fine, and does not appear to raise the sorts of concerns described in the Telemetry Policy. We had just a couple of follow-up questions; please let us know your thoughts in a separate section below.

- In section 1, regarding the metadata about APIs / GenAI function usage that gets logged: It appears that this is set up to collect just the enumerated values for broadly-available AI tools. If a user were to use MLflow in coordination with their own internal, proprietary AI tooling, would any details (even e.g. its name) be included in the telemetry that gets sent back to the project?
- In section 3, it sounds like the raw telemetry data will be stored in a Databricks hosted environment, with access to raw data limited to a few specific MLflow maintainers.
    - Apologies as I'm not familiar with who the project maintainers are. Just to ask: will any non-Databricks employee maintainers have access? Or if all current maintainers are Databricks employees, then if and when there are additional external project maintainers in the future, can we ensure that they would also have equivalent access?
    - We just want to make sure that access to the telemetry data will be available to a broad set of project participants. Or at least, access to a copy of the data should be provided upon request by any project participants (even if not available for immediate download by anyone). From the Telemetry Policy, it states that any approved telemetry data collection "must make the collected data available to all participants in the project community."

2025-06-26 Responses from MLflow team

- Section 1: No, MLflow does not collect or transmit any identifying information about internal or proprietary AI tooling.
    - The telemetry system only logs usage of MLflow-tracked APIs. For features like genai.evaluate, the telemetry captures enumerated values corresponding to MLflow's built-in classes and tools. If users provide custom or internal tools (e.g., a proprietary scorer), the system replaces any such identifiers with a generalized value like 'custom_scorer', ensuring that no sensitive or specific internal tool names are included in telemetry data.
- Section 3
    - Yes, all maintainers of the MLflow GitHub repository—regardless of whether they are Databricks employees—will have access to the raw telemetry data. Currently, there is one maintainer who is not a Databricks employee, and he will have the same level of access. While the telemetry data is hosted in a Databricks-managed environment, it is stored in a standard Databricks account (not an internal account reserved for employees). Access is granted to the full maintainer group, and future maintainers from outside Databricks will receive the same level of access as internal maintainers.

    - To promote transparency and community participation, we will publish aggregated dashboards summarizing the telemetry data on the MLflow website, ensuring it is accessible to all project participants. If a participant requires a copy of the underlying data, they may submit a request to the MLflow team. Upon approval, we can export the requested data to a destination of the requester’s choice (e.g., S3, Azure Data Lake Storage, etc.), provided they grant us the necessary write access.

2025-06-26 LF Projects approval

- LF Projects has approved MLflow’s telemetry collection as described in this document.