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: Spinnaker

Date: October 29, 2019

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 Spinnaker instance ID	No	Yes	No	SHA256 hashed Unique ID generated with every installation and stored in local config file - not customer/personally identifiable
Spinnaker OSS Version	No	No	No	Version of Spinnaker in use, assuming 7user is using the public release with no customization
Unique Application ID	No	No	No	SHA256 hashed application name salted with the unique local/internal Spinnaker ID
Execution Detail - UniqueID	No	No	No	SHA256 hashed execution ID
Execution Detail - Type	No	No	No	Pipeline, Orchestration, Managed Pipeline

				Template V1, Managed Pipeline Template V2
Execution Detail - Trigger	No	No	No	Type of trigger (Git, Docker, Webhook, etc.)
Execution Detail - Stages	No	No	No	Stage, Cloud Provider, Status
Deployment Method **(added 2019-12-20)**	No	No	No	The installation type (i.e. none, other, halyard, operator, minnaker, <and any other installation types>)
Deployment Method Version number **(added 2019-12-20)**	No	No	No	The deployment method's version number (e.g. halyard v1.28)

- If there is public documentation on the project site describing this data, please also provide a URL to that documentation:
    - **The full proto definition of data being sent can be seen here.**

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:
    - **Initial notice of this stems from this public RFC: https://github.com/spinnaker/spinnaker/issues/4738.**
    - **This functionality has not been released yet but would be configurable from the Halyard command-line tool (Halyard is the installation CLI for Spinnaker). We plan to both update public documentation and the Halyard CLI messaging to users to explain (a) how to toggle telemetry collection and (b) what data is being collected. We also plan to notify users at the time of installation/upgrade, where they will have the option to disable telemetry collection, if they choose.**
- If there is public documentation on the project site or in the project source code with the particular notices, please also provide a URL:
    - **This would go into the Halyard public documentation, when the feature is made available in a release.**
- 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, but users will have the option to disable telemetry collection at installation time or after-the-fact via a configuration edit. Disabling/Enabling collection and customization of the data collection endpoint (should**

users want to configure their own data storage and reporting) will be covered in the Spinnaker documentation.

- Is the user able to select between only sharing certain data elements, but not others?
    - **No, at this time, all data conforming to the .proto definition supplied above will be sent.**
- 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?
    - **Halyard can be used to install Spinnaker programmatically but our experience is that automated upgrades of Spinnaker are rare. Release notes and product documentation will alert users to this scenario.**

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 into a Google BigQuery service that is owned by a CDF-funded spinnaker.io account.**
- Who administers and has access to the servers where data is stored?
    - **The actual servers storing the data are in the cloud fabric so this is really a question of who has access to the underlying BigQuery dataset . Our current plan is to document an access process whereby interested users would need to make a request to the Spinnaker TOC. Once verified, access would be granted for some time interval, in order to maintain access only for interested users. Previously mentioned documentation related to telemetry collection functionality would also provide instructions for accessing this data.**
- Are all participants in the project community permitted to view and use the collected telemetry data? Or only particular participants / community members?
    - **Users with appropriate permissions can access data using a publicly accessible Google Data Studio project that can be used to view/export data from the underlying BigQuery data warehouse.**

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?
    - **This can be done by blocking the known endpoint at the organization level. We do not plan to provide detailed firewall instructions for this operation across all public/private cloud environments where Spinnaker could possibly be deployed.**
- Is there a reasonable possibility that including telemetry functionality opens up security vulnerabilities?
    - **No. The data transmitted is taken from Spinnaker's existing event bus and sent via HTTPS POST operations.**
- If so:
    - What steps are taken to mitigate this?
    - If a user does not opt into telemetry data collection, would this risk be fully mitigated?

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.

2019-11-14 comments from LF review

The proposed telemetry data collection and usage generally looks fine. There are a few follow-up comments and questions – please let us know your thoughts.

- For the unique application ID, I see it is salted with the unique Spinnaker instance ID. Are we receiving in telemetry the actual Spinnaker instance ID that is also used as the salt? Or a hashed / modified version of it?
    - If the instance ID received through the telemetry data is also the salt, does that make it possible for someone to use it and the hashed Application ID to obtain the application name? And/or, to use the salt together with a list of common application names to back into the application name?
    - Alternatively, is there a way to obtain a unique Application ID that can still be used as an identifier, but that doesn't derive from the application name at all?
- In the proto file, for the Stage type and CloudProvider id/variant fields, I see that those are strings. Is there a possibility that those could be sent back with data that a user might not want disclosed?
    - Alternatively, could these be some form of enums / categories instead of strings, to avoid this possibility?
- For the opt-out vs. opt-in at install time, the proposed method should be fine, as long as the installing user is clearly notified about the details of the telemetry and has an easy option to disable it. It would be great if that stage of the install process included a link to the specific public documentation for more details.

2019-11-19 responses from Spinnaker team

- "Are we receiving in telemetry the actual Spinnaker instance ID that is also used as the salt? Or a hashed / modified version of it?"
    - We're getting a hash of the instance ID, but not the raw value. So, to be clear:

End-user has	**Data collected**
Spinnaker Instance ID (random ULID, generated once by lifecycle tool)	Hashed instance ID (no salt)
Application ID (name of application)	Hashed application ID (salted with unhashed Instance ID)
Execution ID (random UUID/ULID)	Hashed execution ID (no salt)

- Maybe? It's possible that a user has forked and developed a custom cloud provider that is exclusive to them (such as Netflix's Titus). I'm not opposed to making the cloudProvider a whitelisted enum. As for the `variant` field, this is not yet implemented but the hope is to somehow be able to distinguish between hosted Kubernetes (like GKE, EKS, and AKS) and vanilla Kubernetes.

- Regarding opt-out vs opt-in at install time, we'll make sure this is covered before release (see https://github.com/spinnaker/spinnaker/issues/5146). I'm going to assume we're good to go on this

point and we'll move forward. I'm happy to report back here (or in some sort of follow-up final step?) as the release is made available.

2019-12-02 LF approval and subsequent notes

- From LF: “Yes, on the "cloudProvider" field I'd ask if that could be some form of enum instead. I don't have a particular preference for what values are used, just something that will help ensure that we aren't inadvertently picking up string data that might indicate a particular 'unique' cloud user (such as the Titus example listed in the responses).”

- In subsequent emails with team, also confirmed that IP addresses will not be collected, and that an exclusion has been added to omit IP addresses from collected data.

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

2019-12-20 Additional data review request

- Received request to add installation type and version number to collected data set.

- Question from LF: For the two new data elements, are they free text strings, or a limited set of predefined values (such as integers or defined enums)? See the question in the 2019-11-14 2nd bullet above and response in 2019-11-19 2nd bullet above -- that’s the situation we want to make sure we don’t inadvertently fall into here.

- Response from project team:
    - enum (We could turn the installation type to an enum i.e. none, other, halyard, operator, minnaker, <and any other installation types>)
    - free text strings (The deployment method’s version number would be difficult to turn into an enum as there’s going to be multiple installation types with various versions that change constantly)

2020-01-09 LF notes and approval

- Thanks for the clarifications -- OK with enum for deployment method and with free text string for deployment method version number.

- LF has approved the addition of these two data points to Spinnaker’s telemetry collection as described in this document.