



第127回スクエアFreeセミナー, Feb 24 2022 18:00-18:30

明日から使える「Apache Hop」
データ統合プラットフォームを使えば
データ統合が簡単で楽しくなります。

Koji Annoura
案浦浩二
koji@annoura.com

Source: Apache Hop, <https://hop.apache.org>

Who am I

❖ グラフ理論LOVE

- ❖ Neo4jグラフデータベースにハマってます

- ❖ Graph Community MVP

- ❖ Neo4j Ambassador

- ❖ Neo4j Speaker

- ❖ Neo4j Ninja

- ❖ Neo4j Developer

❖ User Group

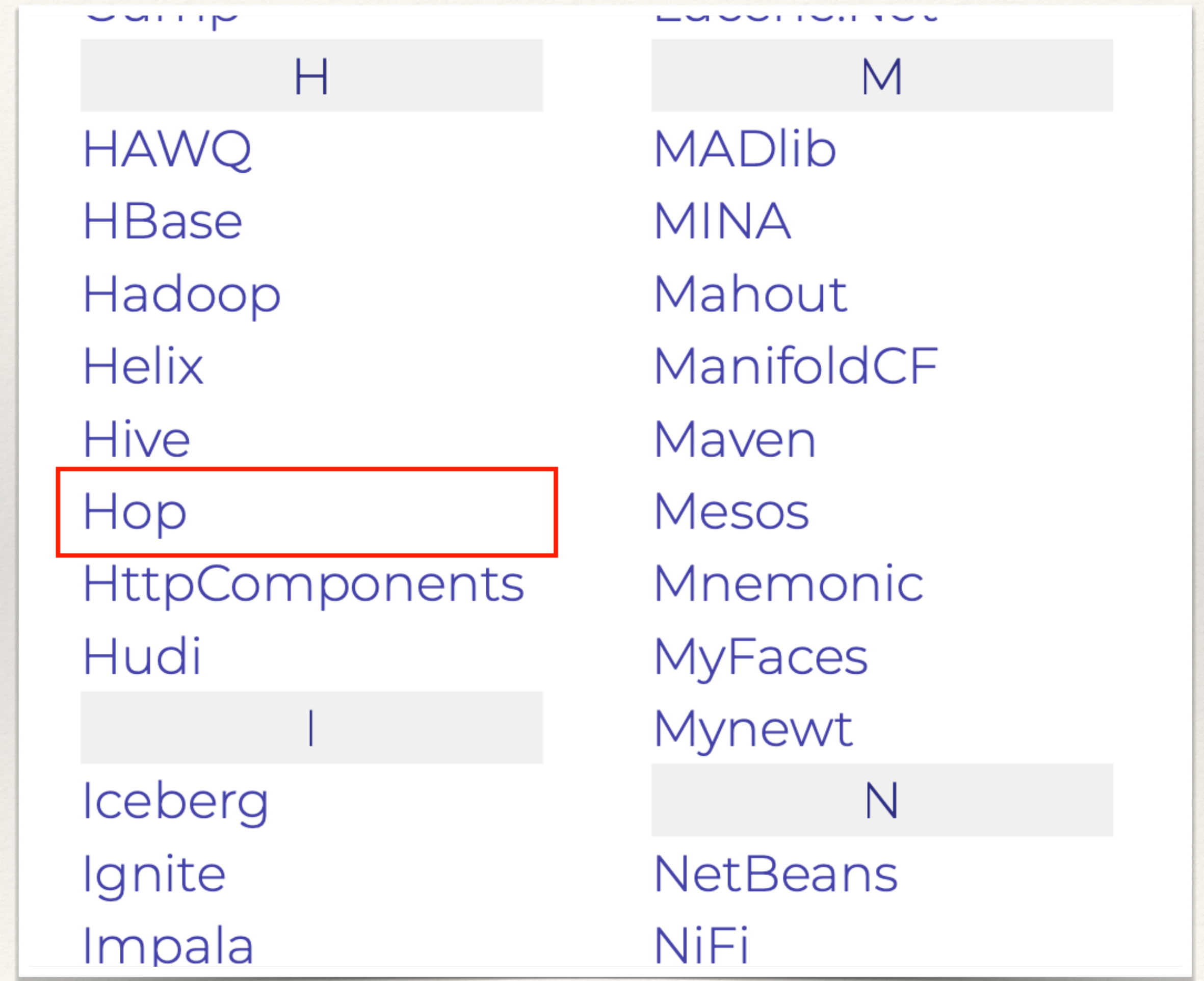
- ❖ Neo4jユーザーグループ

- ❖ 日本Apache Hop ユーザーグループ



Apache® Hop™ as a Top-Level Project (TLP)

- ❖ Apache Hop
- ❖ 2021-12-17 TLP



H	M
HAWQ	MADlib
HBase	MINA
Hadoop	Mahout
Helix	ManifoldCF
Hive	Maven
Hop	Mesos
HttpComponents	Mnemonic
Hudi	MyFaces
I	Mynewt
Iceberg	N
Ignite	NetBeans
Impala	NiFi

Apache Hop 1.1

❖ Apache Hop 1.1.0

❖ Hop Server

- ❖ SSL(https)対応で起動
- ❖ 非同期型Webサービス

❖ New

- ❖ Apache Tika
- ❖ Metadata Input
- ❖ Neo4j Graph Output transform (Add)
 - ❖ multiple relationships
 - ❖ multiple node labels
 - ❖ self-relationships

❖ Hop GUI

- ❖ SCROLL-drag)

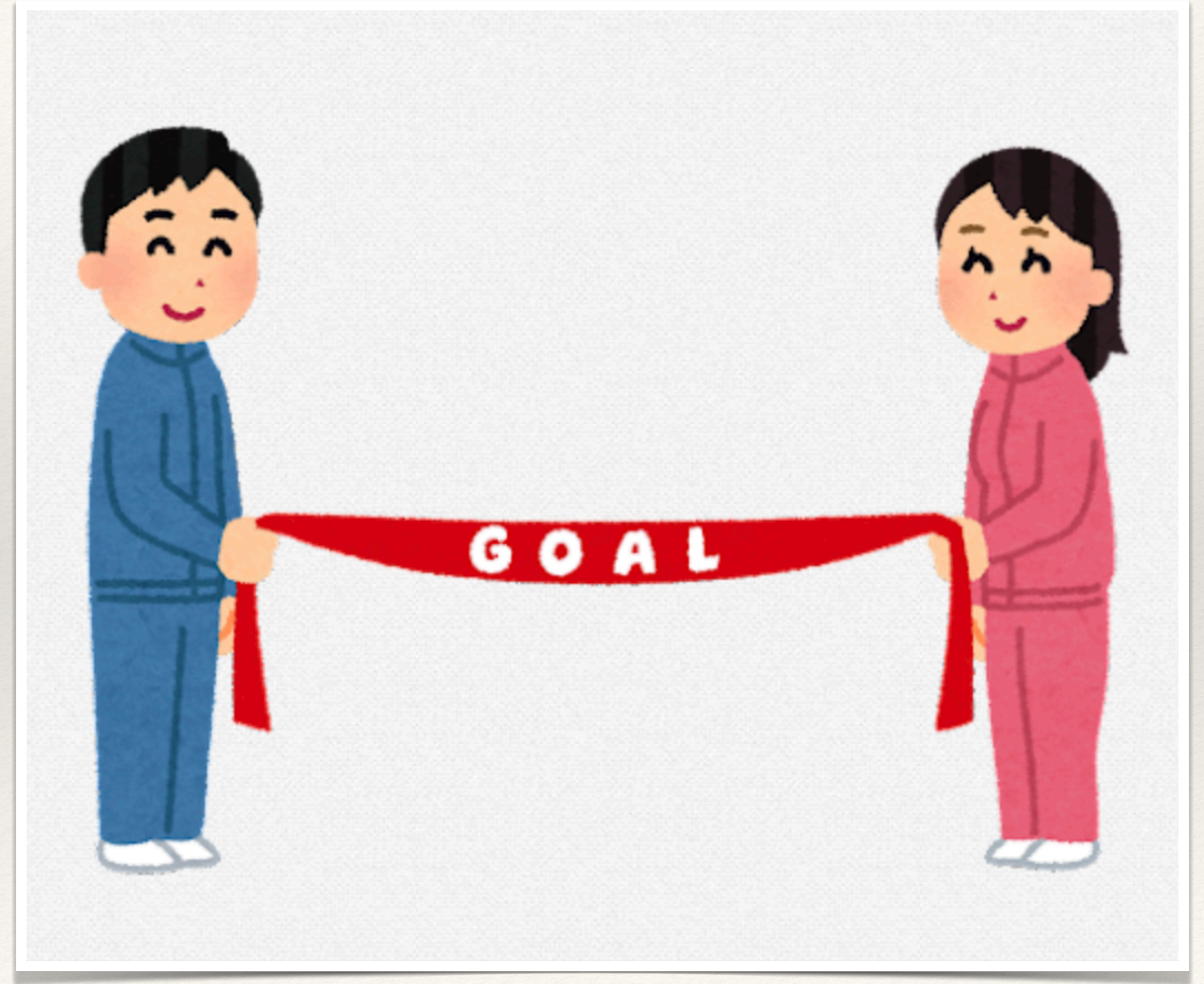
❖ Log4j removal



Source: Apache Hop, <https://twitter.com/ApacheHop>

Goal

- ❖ 今日のゴールは
 - ❖ Apache Hopを知ってもらう
 - ❖ ちょっと使ってみようと思ってもらう
 - ❖ ユーザーグループへのお誘い



どのようにデータを集めていますか？

❖ COVID-19

❖ covid19japan

❖ <https://covid19japan.com>

❖ Problem

❖ 厚労省のサイトにまとまったデータがない

❖ 都道府県・市町村

❖ オープンデータ、Web

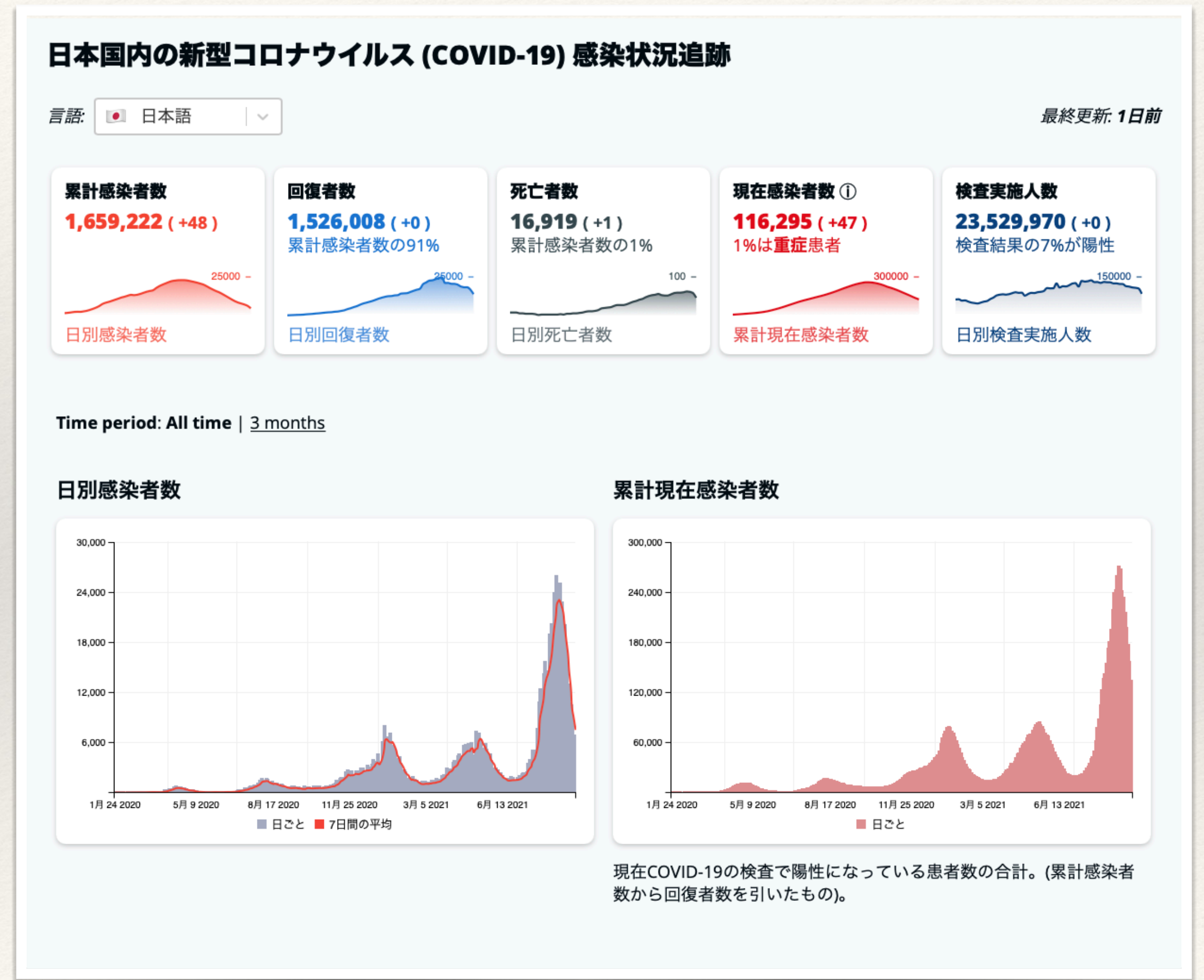
❖ フォーマットが異なる

❖ ヘッダーの名前が微妙に異なる

❖ 発表の間違い

❖ 患者の番号をリナンバーするところがあった

❖ オープンデータやWebの間違いはフィードバック



Source: covid-19 Japan, <https://covid19japan.com>

ETL

- ❖ ETL
 - ❖ Extract
 - ❖ Transform
 - ❖ Load
- ❖ Data Transformer
- ❖ データサイエンティスには必需品



Apache Hop?

❖ Hop?

- ❖ データオーケストレーション、データエンジニアリングのプラットフォーム
- ❖ オープンソース
- ❖ Apache v2 license
- ❖ **Kettle 8.2.0.7からFork**
 - ❖ Pentaho Data Integration (ETL) a.k.a Kettle
 - ❖ 敵対的なフォークではない
 - ❖ we've invited Hitachi Vantara to contribute to Apache Hop

Pentaho Data Integration

Pentaho Data Integration (ETL) a.k.a Kettle

Project Structure

- **assemblies:** Project distribution archive is produced under t
- **core:** Core implementation
- **dbdialog:** Database dialog
- **ui:** User interface
- **engine:** PDI engine
- **engine-ext:** PDI engine extensions
- **plugins:** PDI core plugins
- **integration:** Integration tests

Apache Hop

- ❖ Apache Hop
 - ❖ The **Hop** **O**rchestration **P**latform
 - ❖ 2020年12月にApache のTLPになった
 - ❖ 日本**Apache Hop**ユーザーグループ
 - ❖ Apache Hop User Group Japan



Apache Hop Incubating User Group Japan

❖ Meetup

❖ <https://www.meetup.com/apache-hop-user-group-japan/>

❖ Committers (18 people)

❖ Hiromu Hota

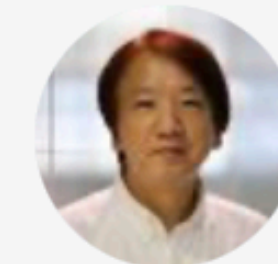
❖ Matt Casters (Neo4j)

❖ Bart Maertens

❖ Facebook

❖ <https://www.facebook.com/groups/apachehopusergroupjapan>

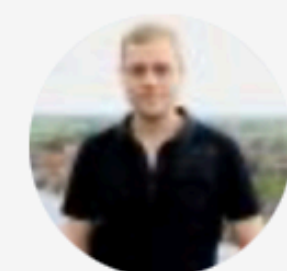
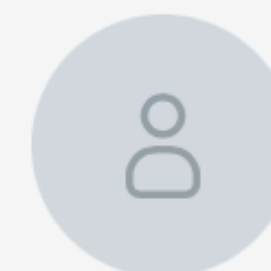
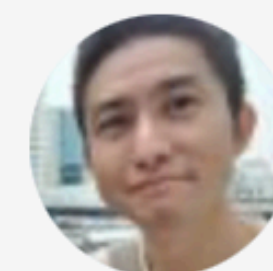
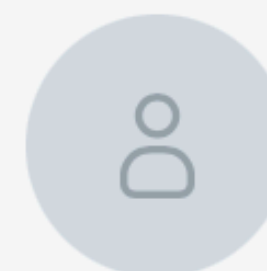
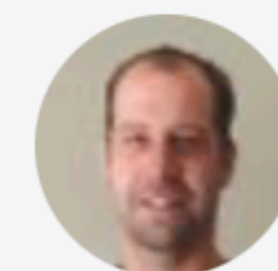
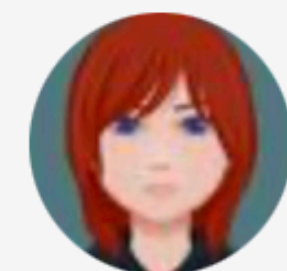
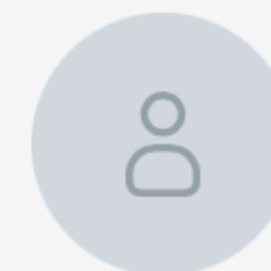
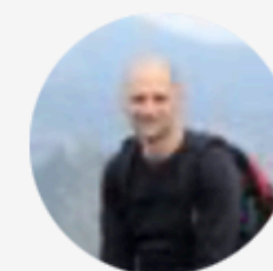
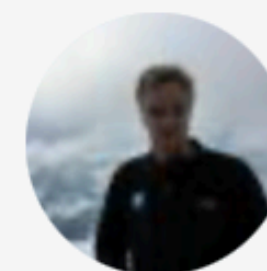
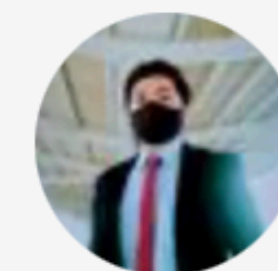
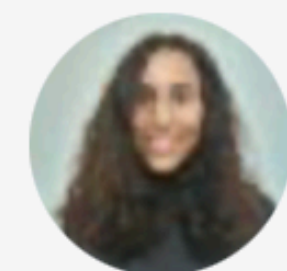
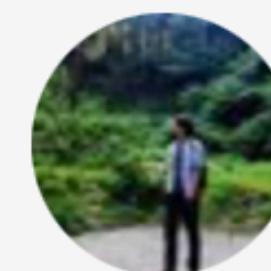
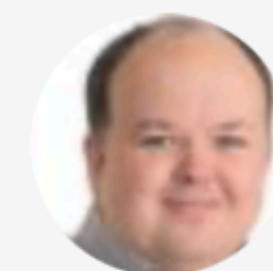
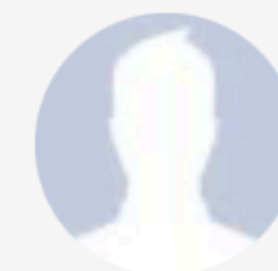
Organizers



Koji Annoura and 1 other
[Message](#)

Members (18)

[See all](#)



Apache Hop - YouTube

- ❖ Apache Hop
- ❖ <https://www.youtube.com/c/ApacheHop>
- ❖ チャンネル登録をお願いします



Source: YouTube, <https://www.youtube.com/c/ApacheHop>

Past Events

❖ Recording

❖ <https://www.youtube.com/c/ApacheHop>

❖ Language

❖ English

❖ Spanish, Español

❖ 日本語



Blog

Documentation ▾

Community ▾

Download

Apache ▾

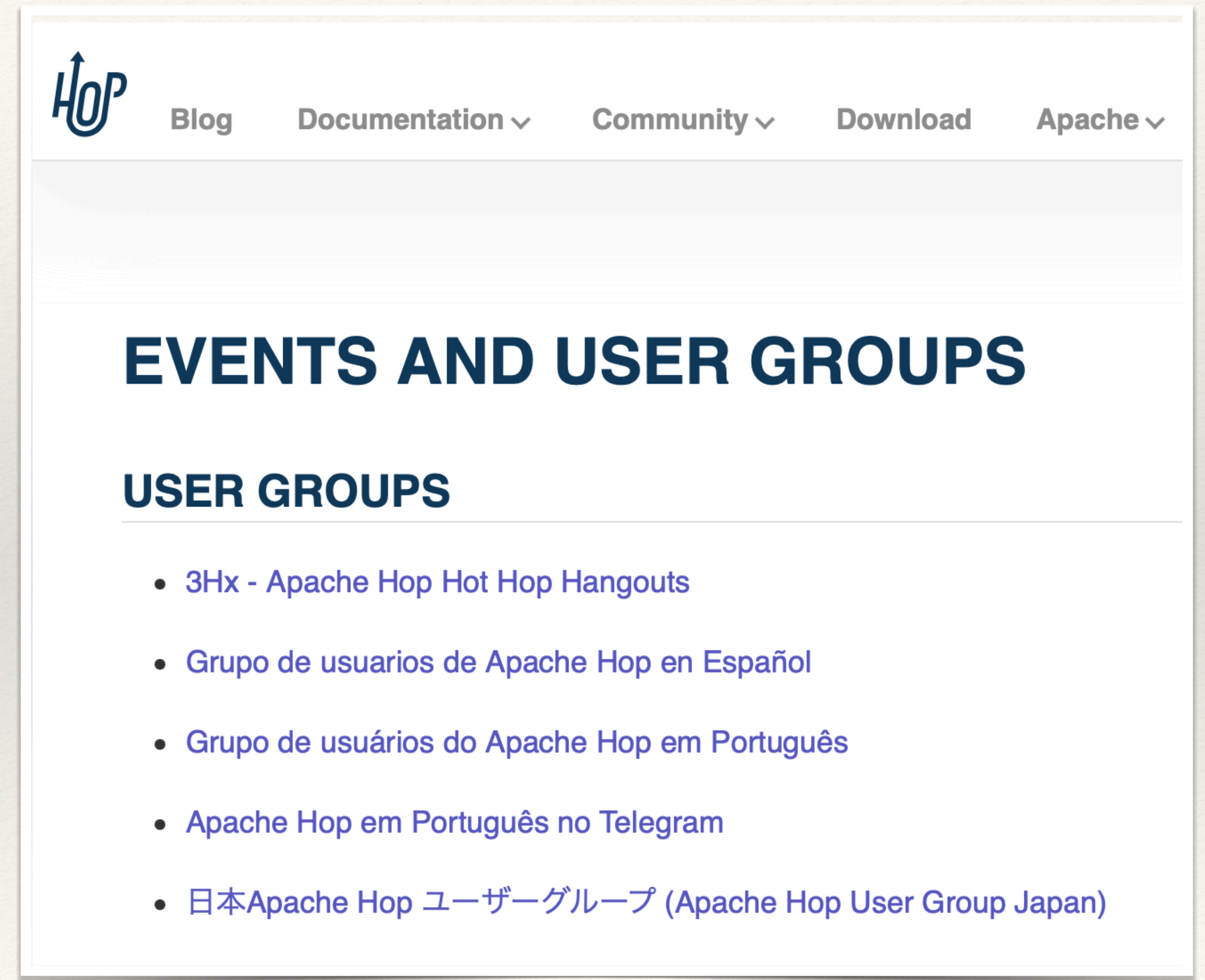
Event name	Description	Date	Language
3H14	Apache Hop in Docker	2021-12-16	EN
3H13	Apache Hop Sucks. Let's un-suck it!	2021-11-18	EN
-	Apache Hop - User Group Japan #1	2021-11-16	JP-EN
3H12	A well-designed Hop project walkthrough	2021-10-28	EN
-	5-fach Speed für Ihr Datenprojekt mit Hop (IT-Novum)	2021-10-26	DE
3H11	Apache Hop 1.0, TLP graduation and beyond	2021-10-07	EN
-	First Brazilian Hop User Group Meetup	2021-7-15	PT
3H10	Apache Hop and Neo4j	2021-07-01	EN
3H9	End-to-end workflows	2021-06-17	EN

Source: Apache Hop, <https://hop.apache.org/community/events/>

User Groups

❖ User Groups

- ❖ 3Hx - Apache Hop Hot Hop Hangouts
- ❖ Grupo de usuarios de Apache Hop en Español
- ❖ Grupo de usuários do Apache Hop em Português
- ❖ Apache Hop em Português no Telegram
- ❖ 日本Apache Hop ユーザーグループ (Apache Hop User Group Japan)



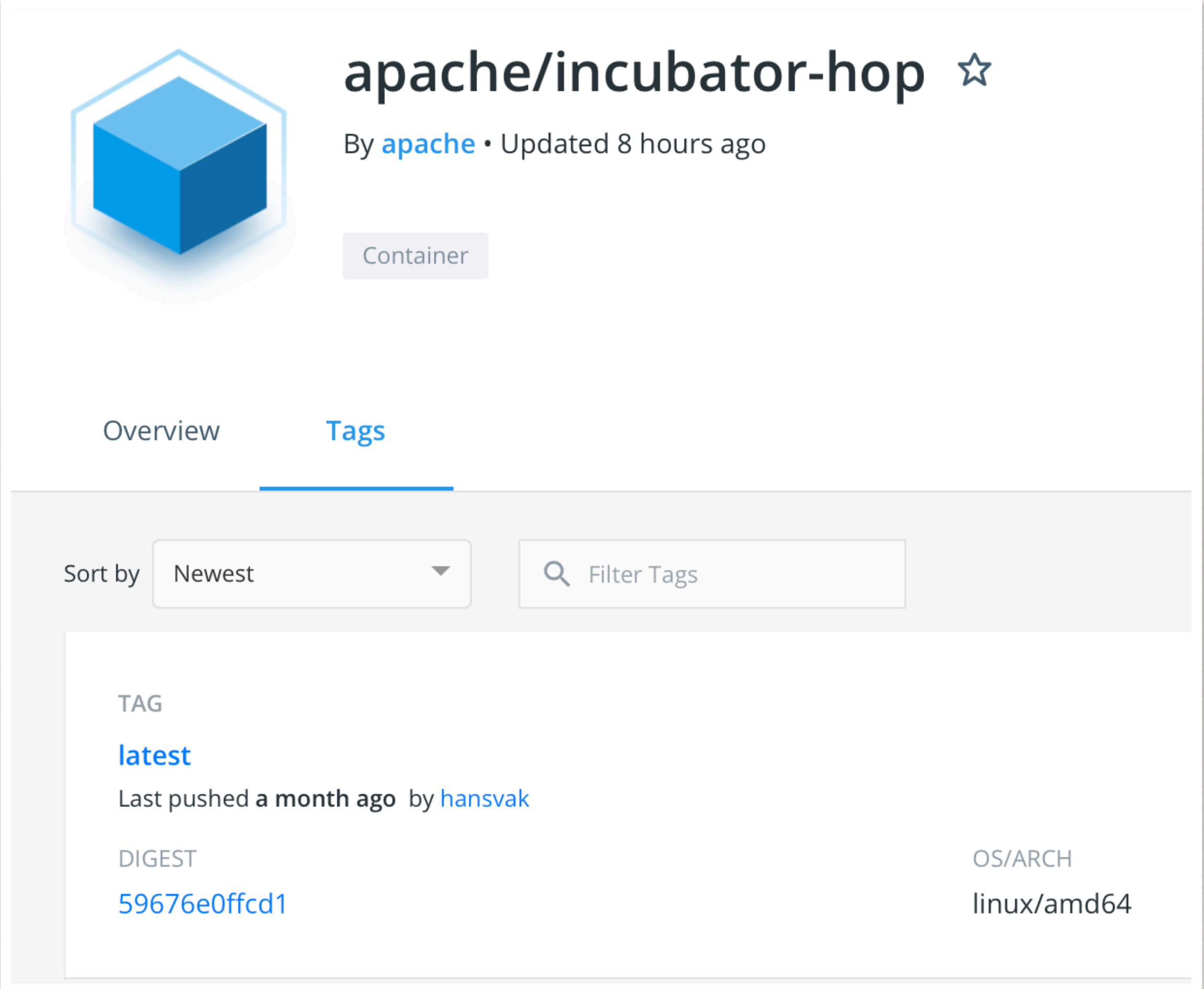
Source: Apache Hop, <https://hop.apache.org/community/events/>

Java

❖ Java

❖ Java 8で開発・テストされています

❖ Java 11正式対応は、2.0.0の予定



The screenshot shows the Docker Hub interface for the `apache/incubator-hop` image. At the top, there is a blue cube icon representing the container. To its right, the text `apache/incubator-hop` is displayed with a star icon, followed by "By [apache](#) • Updated 8 hours ago". Below this, a "Container" label is visible. The interface has two tabs: "Overview" and "Tags", with "Tags" being the active tab. Under the "Tags" tab, there is a "Sort by" dropdown menu set to "Newest" and a "Filter Tags" search box. The main content area shows a table with two columns: "TAG" and "OS/ARCH". The first row shows the tag `latest` with a link to the push event, indicating it was pushed "a month ago" by `hansvak`. The second row shows the digest `59676e0ffcd1` and the architecture `linux/amd64`.

TAG	OS/ARCH
latest Last pushed a month ago by hansvak	
59676e0ffcd1	linux/amd64

Source: docker hub <https://hub.docker.com/r/apache/incubator-hop>

Download

❖ Version

- ❖ 2.0.0 SNAPSHOT
- ❖ 1.2.0 SNAPSHOT
- ❖ **1.1.0 (2022-01-24)**
- ❖ 1.0.0 (2021-10-05)
- ❖ 0.99 (2021-08-05)

❖ docker hub

❖ Hop

- ❖ <https://hub.docker.com/r/apache/hop>

❖ Hop Web

- ❖ <https://hub.docker.com/r/apache/hop-web>

APACHE HOP

		Description	Download Link
Version	Date		
1.1.0	January, 24th, 2022	Binaries	apache-hop-client-1.1.0.zip
		Sources	apache-hop-1.1.0-src.tar.gz
1.0.0	October, 5th, 2021	Binaries	apache-hop-client-1.0.0-incubating.zip
		Sources	apache-hop-1.0.0-incubating-src.tar.gz

Source: Apache Hop, <https://hop.apache.org/download/>

Apache Hop

❖ Metadata Driven



Source: Apache Hop, <https://hop.apache.org>

Apache Hop

- ❖ Design Once, Run Anywhere
 - ❖ **Native Hop runtime**
 - ❖ Apache Spark
 - ❖ Apache Flink
 - ❖ Google Dataflow through the Apache Beam runtimes.



Apache Hop

- ❖ **Visual Development**
 - ❖ intuitive **drag and drop** interface
 - ❖ **workflows**
 - ❖ **pipelines**



Apache Hop

- ❖ **Built-in Life Cycle Management**

- ❖ **Hop Gui**

- ❖ projects
 - ❖ environments
 - ❖ runtime configurations
 - ❖ manage git versions
 - ❖ etc.



Hop Tools

❖ URL

- ❖ <https://hop.apache.org/manual/latest/getting-started/hop-tools.html>

❖ Tools

- ❖ Hop Gui
- ❖ Hop Conf
- ❖ Hop Encrypt
- ❖ Hop Run
- ❖ Hop Search
- ❖ Hop Server
- ❖ Hop Translator (日本語も)

HOP TOOLS

CONTENTS

- Hop GUI
- Hop Conf
- Hop Encrypt
- Hop Run
- Hop Search
- Hop Server
- Hop Translator

Hop contains a number of tools that are optimized for the users and administrators need to perform.

Hop vs. Kettle (Concepts)

Feature	Kettle	Hop
Parallel scalable data pipeline	Transformatio	Pipeline
An operation in a parallel pipeline	Step	Transform
Sequential series of actions	Job	Workflow
An action in a workflow	Job Entry	Action
Shared metadata container	Metastore	Hop Metadata
The graphical user interface	Spoon	Hop Gui
Script to run data pipelines	Pan	Hop Run
Script to run workflows	Kitchen	Hop Run
Server for remote execution	Carte	Hop Server
Script for configuration	-	Hop Config

Hop vs. Kettle (Engines)

Engine	Kettle	Hop
Unit Testing	Plugin	Yes
Apache Spark Support	No (PDI EE)	Yes (Beam)
Apache Flink Support	No	Yes (Beam)
Google Cloud DataFlow Support	No	Yes (Beam)

Hop vs. Kettle (Features)

Feature	Kettle	Hop
Projects and Lifecycle Configuration	No	Yes
Search Information in projects and configurations	No	Yes
Configuration management through UI and	No	Yes
Standardized shared metadata	No	Yes
Pluggable runtime engines	No	Yes
Advanced GUI features: memory, native zoom, ...	No	Yes
Metadata Injection	Yes	Yes
Mapping (sub-transformation/pipeline	Yes	Yes
Web Interface	WebSpoon	Hop Web
APL 2.0 license compliance	LGPL	Yes
Pluggable metadata objects	No	Yes
GUI plugin architecture	XUL based	Java

Hop Server

- ❖ Hop Server
 - ❖ lightweight server
 - ❖ `http://localhost:8081`
 - ❖ run **workflows** and **pipelines**
 - ❖ **REST** services

Hop Server Status



Pipelines

▶ ■ 🔍 👁 ✕				
Name	Hop Server Object ID	Status	Start date	Last log time
remote-pipeline	c1451bfb-b867-4c76-b123-c29d2b05da17	Finished	2021/04/16	21:00:05.754
remote-pipeline	25ea91b0-1cc3-43fc-aa06-12f922082b34	Finished	2021/04/16	20:58:28.818

Workflows

▶ ■ 👁 ✕				
Name	Hop Server Object ID	Status	Start date	Last log time

Configuration details:

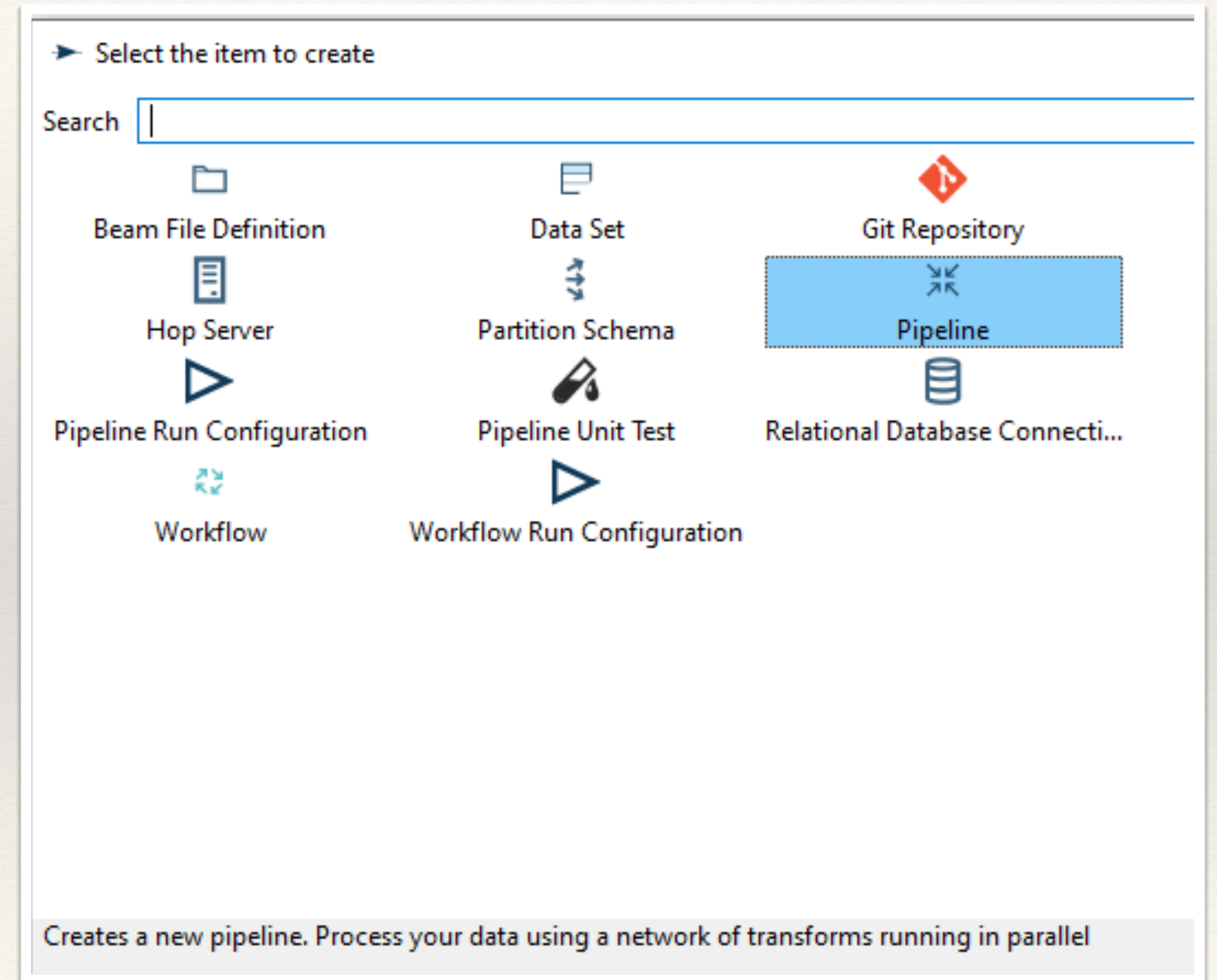
The maximum size of the central log buffer	No limit
The maximum age of a log line	No limit
The maximum age of a stale object	No limit

These parameters can be set in the hop server configuration XML file: (Using defaults)

Pipeline

❖ Pipeline

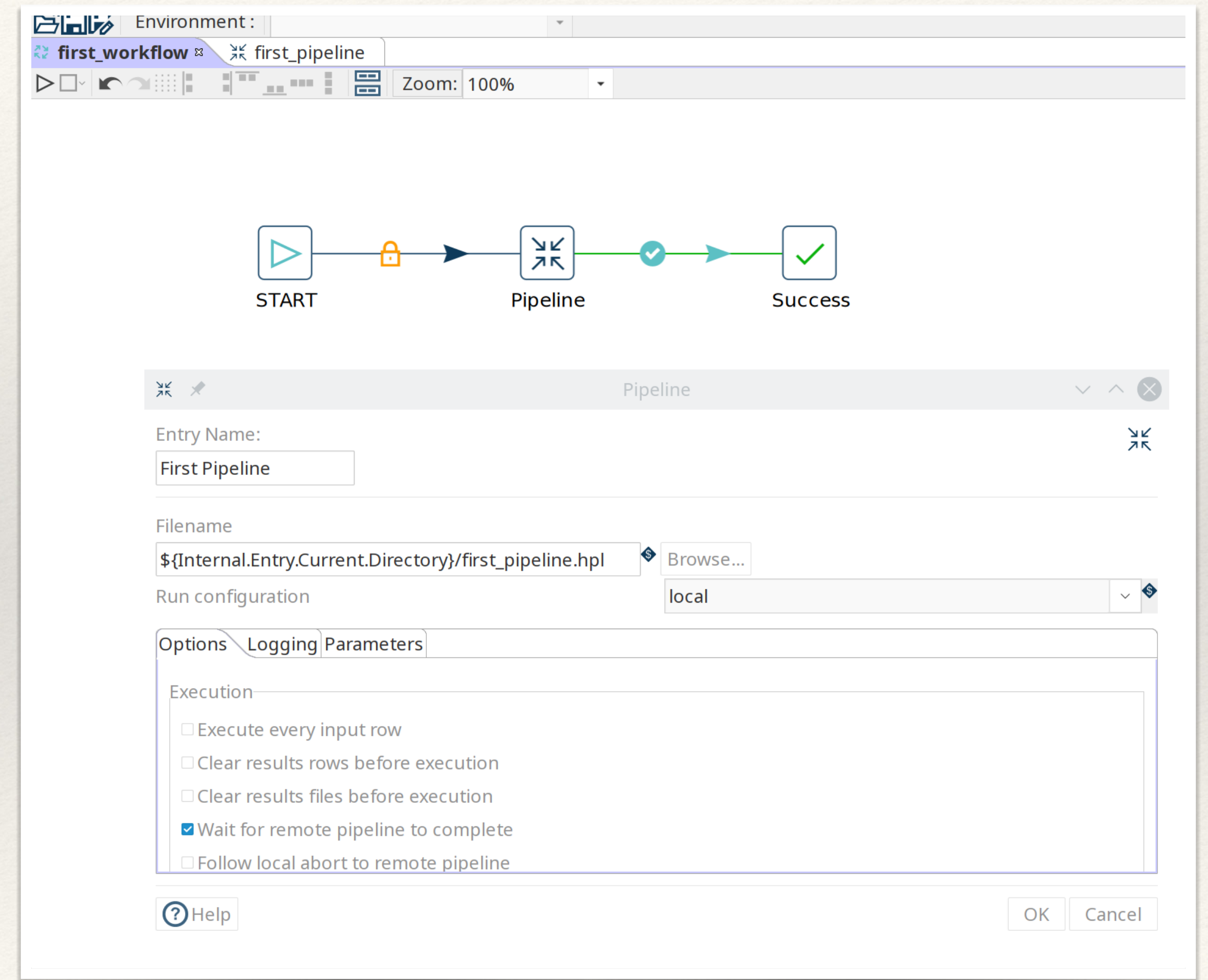
- ❖ 1つまたは複数のソースからデータを読み込む
- ❖ いくつかの操作（結合、ルックアップ、フィルタなど）を実行する
- ❖ 最終的に1つまたは複数のターゲットプラットフォームにデータを書き込む



Workflows

❖ Workflows

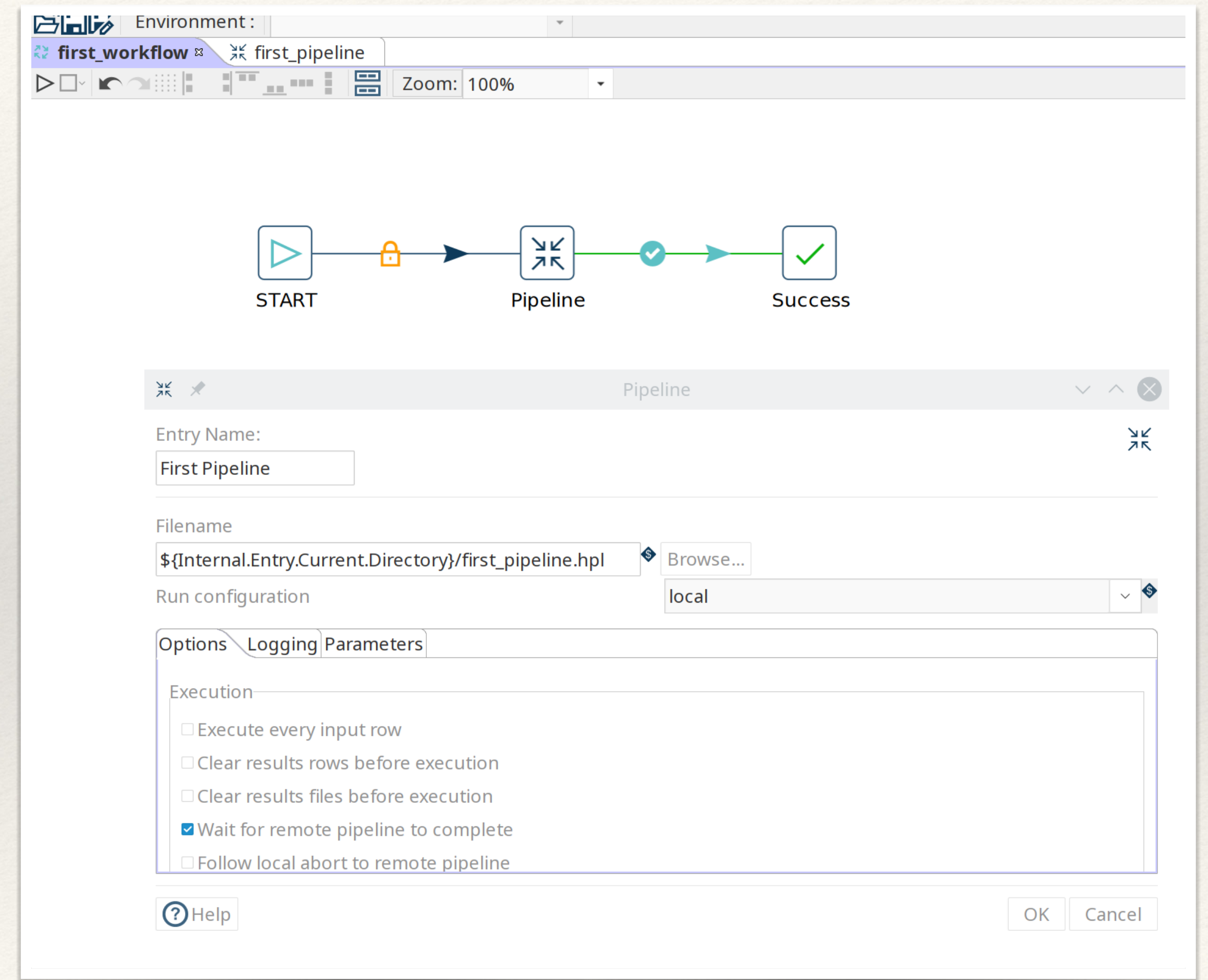
- ❖ Workflowsは、Apache Hopの中核となる構成要素の一つ
- ❖ Pipelineが重いデータ処理を行うのに対し、Workflowsは環境の準備、リモートファイルの取得、エラー処理、child workflowsやPipelineの実行などのオーケストレーション作業を行う。



Workflows (Action Plugins)

❖ Action Pluginsの一部

- ❖ Check Db connections
- ❖ Get a file with SFTP
- ❖ Get Mail from POP
- ❖ Mail Validator
- ❖ MS SQL Server Bulk Loader
- ❖ MySQL Bulk File
- ❖ MySQL Bulk Loader
- ❖ Neo4j Check Connections
- ❖ Neo4j Constraint
- ❖ Neo4j Cypher Script
- ❖ Neo4j Index
- ❖ Ping



Relational Database Connections

❖ Driver

❖ JDBC, Native

❖ RDB

❖ Apache Derby, Apache Doris, AS400, Cache, ClickHouse, DB2, Exasol, firebird, Google BigQuery, Greenplum, Gupta SQLBase, H2, Hypersonic, IBM Universe, InfiniDB, Infobright DB, informix, ingres, interbase, Kingbase ES, **Mariadb**, MaxDB (SAP DB), monetdb, **MS SqlServer**, **MS SqlServer (Native)**, **MySQL**, Netezza, **Oracle**, **Oracle RDB**, **PostgreSQL**, Redshift, Snowflake, SQLite, Sybase, Sybase IQ, Teradata, Vectorwise, Vertica

The screenshot shows a 'Connection information' dialog box with a dark header. Below the header are three tabs: 'General' (selected), 'Advanced', and 'Options'. The 'General' tab contains the following fields:

- Connection name:** A text input field.
- Connection type:** A dropdown menu currently showing 'Generic database'.
- Use ODBC?:** A checkbox, currently unchecked.
- ODBC DSN Source Name:** A text input field.
- Driver class:** A text input field.
- Username:** A text input field.
- Password:** A text input field.
- Manual connection URL:** A text input field.

At the bottom right of the dialog are three buttons: 'OK', 'Test', and 'Cancel'.

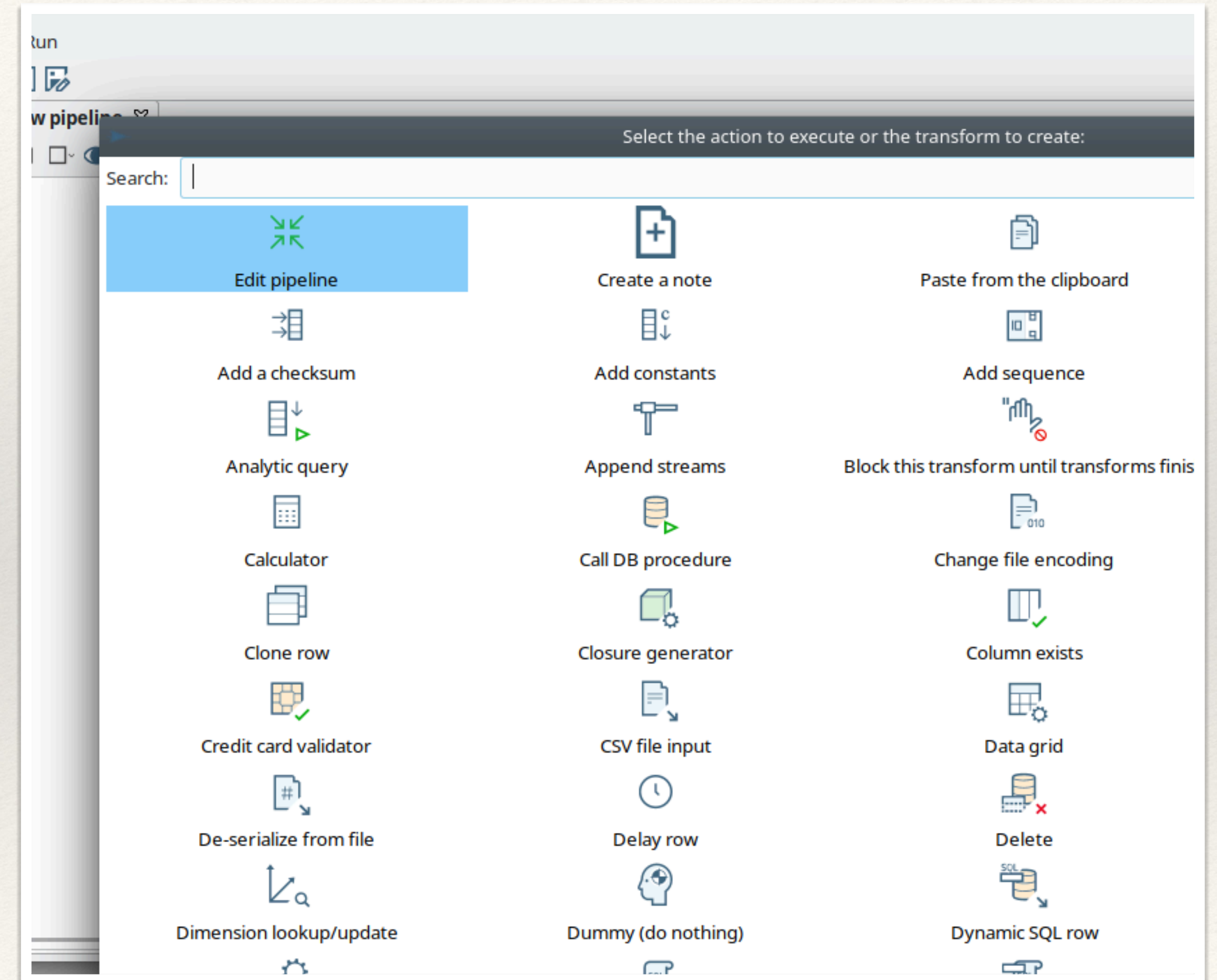
Transform

❖ Transform Plugins

❖ 200近くの変換が可能

❖ List

- ❖ Abort
- ❖ Add a checksum
- ❖ Add Constants
- ❖ Add Sequence
- ❖ (中略)
- ❖ XML Join
- ❖ XML Output
- ❖ XSD Validator
- ❖ XSL Transformation
- ❖ Yaml Input
- ❖ Zip file

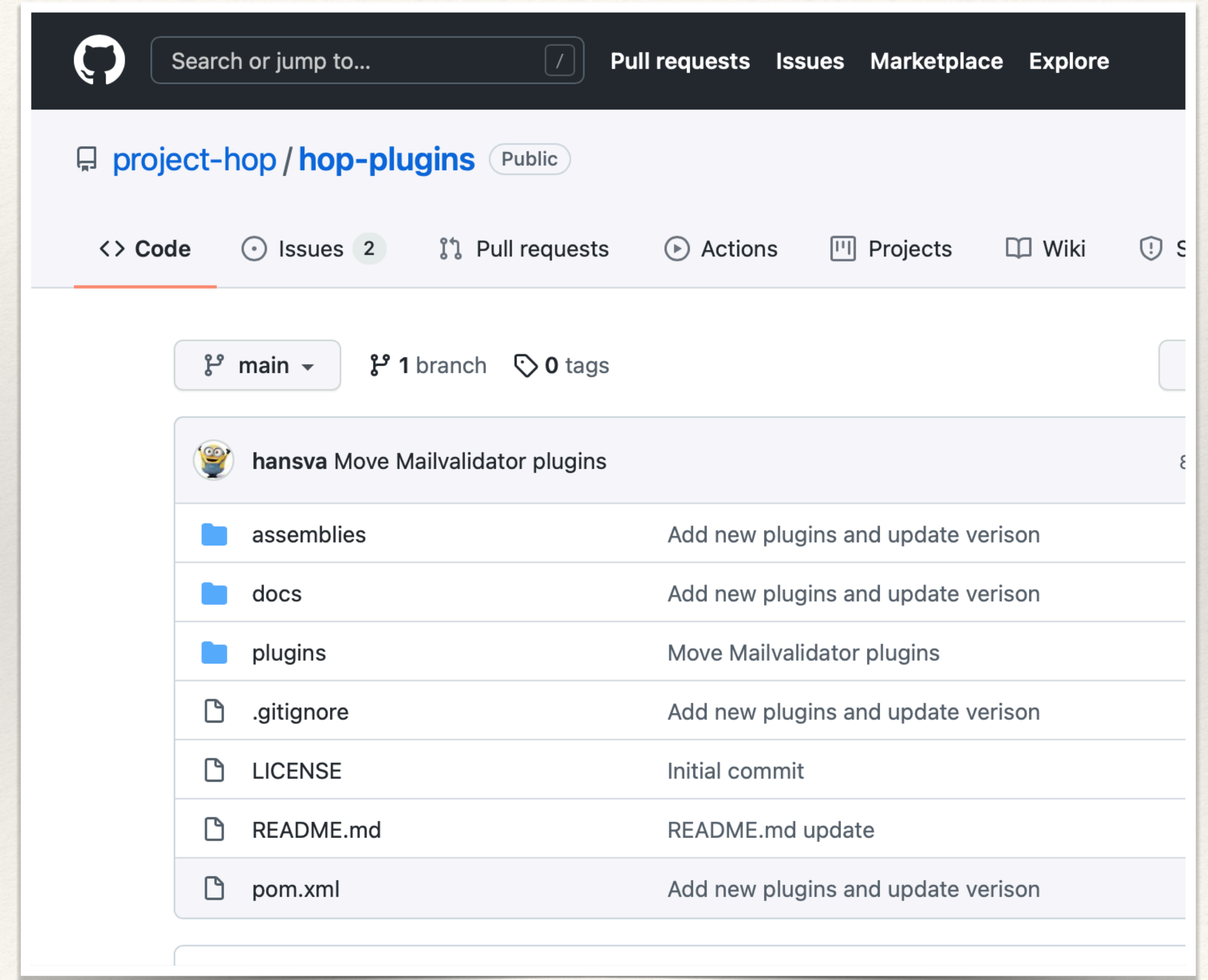


Source: Apache Hop, <https://hop.apache.org/manual/latest/pipeline/create-pipeline.html>

External Plugins

❖ TRANSFORMS

- ❖ **Dropbox** Input
- ❖ **Dropbox** Output
- ❖ **Excel** Output
- ❖ **Google Analytics**
- ❖ **Google Sheets** Input
- ❖ **Google Sheets** Output
- ❖ **LDIF** (Lightweight Directory Interchange Format), LDAP
 - ❖ **LDIF** Input
- ❖ **MQTT** (Message Queue Telemetry Transport)
 - ❖ **MQTT** Input
 - ❖ **MQTT** Output
- ❖ **Send message** to syslog



Source: GitHub, <https://github.com/project-hop/hop-plugins/>

ありがとうございました

❖ ご質問があれば、koji@annoura.com まで