Kafka Cluster Topic Key Distribution Hot Partition Analyzer Tool

Efficient **Kafka key distribution** is fundamental to building scalable, high-performance event-driven systems. Kafka uses each record's key to determine which partition it belongs to—governing **data ordering**, **load balancing**, and **parallelism** across the cluster. When key distribution is uneven, some partitions become hot, processing far more traffic than others. These **hot partitions** lead to broker overload, consumer lag, and throttled throughput, undermining the scalability of your Kafka workloads.

This tool helps you **test**, **visualize**, and **validate** how record keys are distributed across topic partitions in your Kafka cluster. It generates records using configurable key patterns, publishes them to a target topic, and then consumes the data to analyze partition utilization and message distribution metrics.

By surfacing patterns of **data skew**, **low-key cardinality**, or **biased hashing**, the analyzer reveals whether your partitioning strategy is truly balanced. The results empower you to:

- Detect and diagnose **hot partitions** before they degrade performance.
- Experiment with **key-salting** or **hashing strategies** to improve balance.
- Optimize **consumer parallelism** and **broker load** for predictable throughput at scale.

Use this tool as a **proactive performance lens** on your Kafka topics—ensuring your cluster's data distribution is as efficient, scalable, and reliable as the workloads it powers.

Table of Contents

1.0 To get started

Download ---> ***Configure*** ---> ***Run*** ---> ***Results***

1.1 Download the Tool

Clone the repo: `shell git clone https://github.com/j3-signalroom/kafka_cluster-topic-key_distribution-hot_partition_analyzer-tool.git`

Since this project was built using **`uv`**, please install it, and then run the following command to install all the project dependencies:

```
uv sync
```

1.2 Configure the Tool

Now, you need to set up the tool by creating a `.env` file in the root directory of your project. This file will store all the essential environment variables required for the tool to connect to your Confluent Cloud Platform and function correctly. Additionally, you can choose to use **AWS Secrets Manager** to manage your secrets.

1.2.1 Create a Dedicated Service Account for the Analyzer Tool

The service account needs to have OrganizationAdmin, EnvironmentAdmin or CloudClusterAdmin role to provision Kafka cluster API keys and the MetricsViewer role to access the Metrics API for all clusters it has access to.

1. Use the Confluent CLI (Command-Line Interface) to create the service account:

    > **Note:** If you haven't already, install the Confluent CLI and log in to your Confluent Cloud account using `confluent login`. Moreover, the account you use to log in must have the OrganizationAdmin role to create the **Cloud API key in Step 5**.

    ```
    confluent iam service-account create <SERVICE_ACCOUNT_NAME> --description "<DESCRIPTION>"
    ```

    For instance, you run `confluent iam service-account create recommender-service-account --description "Service account for Recommender Tool"`, the output should resemble:

    ```
    +-------------+--------------------------------+
    | ID          | sa-abcd123                     |
    ```

```
| Name        | recommender-service-account    |
| Description | Service account for            |
|             | Recommender Tool               |
+-------------+--------------------------------+
```

2. Make note of the service account ID in the output, which is in the form `sa-xxxxxxx`, which you will assign the OrganizationAdmin, EnvironmentAdmin or CloudClusterAdmin role, and MetricsViewer role to in the next steps, and assign it to the `PRINCIPAL_ID` environment variable in the `.env` file.

3. Decide at what level you want to assign the OrganizationAdmin, EnvironmentAdmin or CloudClusterAdmin role to the service account. The recommended approach is to assign the role at the organization level so that the service account can provision API keys for any Kafka cluster in the organization. If you want to restrict the service account to only be able to provision API keys for Kafka clusters in a specific environment, then assign the EnvironmentAdmin role at the environment level. If you want to restrict the service account to only be able to provision API keys for a specific Kafka cluster, then assign the CloudClusterAdmin role at the cluster level.

   For example, to assign the EnvironmentAdmin role at the environment level:

   ```
   confluent iam rbac role-binding create --role EnvironmentAdmin --principal User:<SERVICE_ACCOUNT_ID> --environment <ENVIRONMENT_ID>
   ```

   Or, to assign the CloudClusterAdmin role at the cluster level:

   ```
   confluent iam rbac role-binding create --role CloudClusterAdmin --principal User:<SERVICE_ACCOUNT_ID> --cluster <KAFKA_CLUSTER_ID>
   ```

   For instance, you run `confluent iam rbac role-binding create --role EnvironmentAdmin --principal User:sa-abcd123 --environment env-123abc`, the output should resemble:

   ```
   +-----------+------------------+
   | ID        | rb-j3XQ8Y        |
   | Principal | User:sa-abcd123  |
   | Role      | EnvironmentAdmin |
   +-----------+------------------+
   ```

4. Assign the MetricsViewer role to the service account at the organization, environment, or cluster level, For example to assign the MetricsViewer role at the environment level:

   ```
   confluent iam rbac role-binding create --role MetricsViewer --principal User:<SERVICE_ACCOUNT_ID> --environment <ENVIRONMENT_ID>
   ```

   For instance, you run `confluent iam rbac role-binding create --role MetricsViewer --principal User:sa-abcd123 --environment env-123abc`, the output should resemble:

   ```
   +-----------+------------------+
   | ID        | rb-1GgVMN        |
   | Principal | User:sa-abcd123  |
   | Role      | MetricsViewer    |
   +-----------+------------------+
   ```

5. Create an API key for the service account:

   ```
   confluent api-key create --resource cloud --service-account <SERVICE_ACCOUNT_ID> --description "<DESCRIPTION>"
   ```

   For instance, you run `confluent api-key create --resource cloud --service-account sa-abcd123 --description "API Key for Recommender Tool"`, the output should resemble:

   ```
   +------------+------------------------------------------------------------------+
   | API Key    | 1WORLDABCDEF7OAB                                                 |
   | API Secret | cfltabCdeFg1hI+/2j34KLMnoprSTuvxy/Za+b5/6bcDe/7fGhIjklMnOPQ8rT9U |
   +------------+------------------------------------------------------------------+
   ```

6. Make note of the API key and secret in the output, which you will assign to the `confluent_cloud_api_key` and `confluent_cloud_api_secret` environment variables in the `.env` file. Alternatively, you can securely store and retrieve these credentials using AWS Secrets Manager.

1.2.2 Create the `.env` file

Create the `.env` file and add the following environment variables, filling them with your Confluent Cloud credentials and other required values:

```
# Environment variables credential for Confluent Cloud
CONFLUENT_CLOUD_CREDENTIAL={"confluent_cloud_api_key":"<YOUR_CONFLUENT_CLOUD_API_KEY>", "confluent_cloud_api_secret":
"<YOUR_CONFLUENT_CLOUD_API_SECRET>"}

# Environment and Kafka cluster filters (comma-separated IDs)
# Example: ENVIRONMENT_FILTER="env-123,env-456"
# Example: KAFKA_CLUSTER_FILTER="lkc-123,lkc-456"
ENVIRONMENT_FILTER=<YOUR_ENVIRONMENT_FILTER, IF ANY>
KAFKA_CLUSTER_FILTER=<YOUR_KAFKA_CLUSTER_FILTER, IF ANY>

# Confluent Cloud principal ID (user or service account) for API key creation
# Example: PRINCIPAL_ID=u-abc123 or PRINCIPAL_ID=sa-xyz789
PRINCIPAL_ID=<YOUR_PRINCIPAL_ID>

# AWS Secrets Manager Secrets for Confluent Cloud and Kafka clusters
USE_AWS_SECRETS_MANAGER=<True|False>
CONFLUENT_CLOUD_API_SECRET_PATH={"region_name": "<YOUR_SECRET_AWS_REGION_NAME>", "secret_name": "
<YOUR_CONFLUENT_CLOUD_API_KEY_AWS_SECRETS>"}
```

The environment variables are defined as follows:

Environment Variable Name	Type	Description	Example	Default	Required
`ENVIRONMENT_FILTER`	Comma-separated String	A list of specific Confluent Cloud environment IDs to filter. When provided, only these environments will be used to fetch Kafka cluster credentials. Use commas to separate multiple environment IDs. Leave blank or unset to use all available environments.	`env-123,env-456`	Empty (all environments)	No
`PRINCIPAL_ID`	String	Confluent Cloud principal ID (user or service account) for API key creation.	`u-abc123` or `sa-xyz789`	None	Yes
`KAFKA_CLUSTER_FILTER`	Comma-separated String	A list of specific Kafka cluster IDs to filter. When provided, only these Kafka clusters will be analyzed. Use commas to separate multiple cluster IDs. Leave blank or unset to analyze all available clusters.	`lkc-123,lkc-456`	Empty (all clusters)	No
`CONFLUENT_CLOUD_CREDENTIAL`	JSON Object	Contains authentication credentials for Confluent Cloud API access. Must include `confluent_cloud_api_key` and `confluent_cloud_api_secret` fields for authenticating with Confluent Cloud services.	`{"confluent_cloud_api_key": "CKABCD123456", "confluent_cloud_api_secret": "xyz789secretkey"}`	None	Yes (if not Manager)
`USE_AWS_SECRETS_MANAGER`	Boolean	Controls whether to retrieve credentials from AWS Secrets Manager instead of using direct environment variables. When `True`, credentials are fetched from AWS Secrets Manager using the paths specified in other variables.	`True` or `False`	`False`	No
`CONFLUENT_CLOUD_API_SECRET_PATH`	JSON Object	AWS Secrets Manager configuration for Confluent Cloud credentials. Contains `region_name` (AWS region) and `secret_name` (name of the secret in AWS Secrets Manager). Only used when `USE_AWS_SECRETS_MANAGER` is `True`.	`{"region_name": "us-east-1", "secret_name": "confluent-cloud-api-credentials"}`	None	Yes (if `USE_AWS_` is `True`)

1.2.3 Using the AWS Secrets Manager (optional)

If you use **AWS Secrets Manager** to manage your secrets, set the `USE_AWS_SECRETS_MANAGER` variable to `True` and the tool will retrieve the secrets from AWS Secrets Manager using the names provided in `CONFLUENT_CLOUD_API_KEY_AWS_SECRETS`.

The code expects the `CONFLUENT_CLOUD_API_KEY_AWS_SECRETS` to be stored in JSON format with these keys:

- `confluent_cloud_api_key`
- `confluent_cloud_api_secret`

1.3 Run the Tool

Navigate to the Project Root Directory

Open your Terminal and navigate to the root folder of the `kafka_cluster-topic-key_distribution-hot_partition_analyzer-tool/` repository that you have cloned. You can do this by executing:

```
cd path/to/kafka_cluster-topic-key_distribution-hot_partition_analyzer-tool/
```

> Replace `path/to/` with the actual path where your repository is located.

Then enter the following command below to run the tool:

```
uv run streamlit run src/tool.py
```

1.3.1 Did you notice we prefix `uv run` to `streamlit run src/tool.py`?

You maybe asking yourself why. Well, `uv` is an incredibly fast Python package installer and dependency resolver, written in **Rust**, and designed to seamlessly replace `pip`, `pipx`, `poetry`, `pyenv`, `twine`, `virtualenv`, and more in your workflows. By prefixing `uv run` to a command, you're ensuring that the command runs in an optimal Python environment.

Now, let's go a little deeper into the magic behind `uv run`:

- When you use it with a file ending in `.py` or an HTTP(S) URL, `uv` treats it as a script and runs it with a Python interpreter. In other words, `uv run file.py` is equivalent to `uv run python file.py`. If you're working with a URL, `uv` even downloads it temporarily to execute it. Any inline dependency metadata is installed into an isolated, temporary environment—meaning zero leftover mess! When used with `-`, the input will be read from `stdin`, and treated as a Python script.
- If used in a project directory, `uv` will automatically create or update the project environment before running the command.
- Outside of a project, if there's a virtual environment present in your current directory (or any parent directory), `uv` runs the command in that environment. If no environment is found, it uses the interpreter's environment.

So what does this mean when we put `uv run` before `streamlit run src/tool.py`? It means `uv` takes care of all the setup—fast and seamless—right in your local environment. If you think AI/ML is magic, the work the folks at Astral have done with `uv` is pure wizardry!

Curious to learn more about Astral's `uv`? Check these out:

- Documentation: Learn about `uv`.
- Video: `uv` IS THE FUTURE OF PYTHON PACKING!.

If you have Kafka connectivity issues, you can verify connectivity using the following command:

1.3.2 Troubleshoot Connectivity Issues (if any)

To verify connectivity to your Kafka cluster, you can use the `kafka-topics.sh` command-line tool. First, download the Kafka binaries from the Apache Kafka website and extract them. Navigate to the `bin` directory of the extracted Kafka folder. Second, create a `client.properties` file with your Kafka credentials:

```
# For SASL_SSL (most common for cloud services)
security.protocol=SASL_SSL
sasl.mechanism=PLAIN
sasl.jaas.config=org.apache.kafka.common.security.plain.PlainLoginModule required \
  username="<YOUR_KAFKA_API_KEY>" \
  password="<YOUR_KAFKA_API_SECRET>";

# Additional SSL settings if needed
ssl.endpoint.identification.algorithm=https
```

Finally, run the following command to list all topics in your Kafka cluster:

```
./kafka-topics.sh --list --bootstrap-server <YOUR_BOOTSTRAP_SERVER_URI> --command-config ./client.properties
```

If the connection is successful, you should see a list of topics in your Kafka cluster. If you encounter any errors, double-check your credentials and network connectivity.

1.4 The Results

2.0 How the Tool Works

3.0 Resources

- The Importance of Standardized Hashing Across Producers
- librdkafka Configuration