

KIP-678: New Kafka Connect SMT for plainText => Struct(or Map) with Regex

- [Status](#)
- [Motivation](#)
- [Public Interfaces](#)
- [Proposed Changes](#)
- [Compatibility, Deprecation, and Migration Plan](#)
- [Rejected Alternatives](#)

Status

Current state: Voting

Discussion thread: [discuss mail](#) [discuss mail2](#)

JIRA:

 Unable to render Jira issues macro, execution error.

PULL REQUEST: <https://github.com/apache/kafka/pull/12219> (before review PR <https://github.com/apache/kafka/pull/7965>)

Please keep the discussion on the mailing list rather than commenting on the wiki (wiki discussions get unwieldy fast).

Motivation

There are no existing SMT for make a plain text to Struct.

this makes easy to control and transform data structure only with the kafka-connect connector configuration.

Use Case :

1. Parsing AccessLog

Input

```
111.61.73.113 - - [08/Aug/2019:18:15:29 +0900] \"OPTIONS /api/v1/service_config HTTP/1.1\" 200 - 101989 \"http://local.test.com/\" \"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_5) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/75.0.3770.142 Safari/537.36\""
```

Output

```
{
  "IP" : "111.61.73.113"
  , "RemoteUser" : "-"
  , "AuthedRemoteUser" : "-"
  , "DateTime" : "08/Aug/2019:18:15:29 +0900"
  , "Method" : "OPTIONS"
  , "Request" : "/api/v1/service_config"
  , "Protocol" : "HTTP/1.1"
  , "Response" : "200"
  , "BytesSent" : "-"
  , "Ms" : "101989"
  , "Referrer" : "http://local.test.com"
  , "UserAgent"
    : "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_5) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/75.0.3770.142 Safari/537.36"}
}
```

2. Parsing Patterned String (such as URL)

Input

```
https://kafka.apache.org/documentation/#connect"
```

Output

```
{
  "protocol" : "https"
  , "domain" : "kafka.apache.org"
  , "path" : "documentation/#connect"
}
```

Public Interfaces

Briefly list any new interfaces that will be introduced as part of this proposal or any existing interfaces that will be removed or changed. The purpose of this section is to concisely call out the public contract that will come along with this feature.

name	description	type	default	valid values	importance
regex	String Regex Group Pattern	String		grouped regular expression string	medium
mapping	Ordered Regex Group Mapping Keys	String		comma seperated names	medium

Example Config

Usecase1. Transform Config

```
"transforms": "RegexTransform",
"transforms.RegexTransform.type": "org.apache.kafka.connect.transforms.ParseStructByRegex$Value",

"transforms.RegexTransform.regex": "^([\\d.]+) (\\S+) (\\S+) \\[[([\\w:/]+\\s[+\\-]\\d{4})\\]\\ \\\"
(GET|POST|OPTIONS|HEAD|PUT|DELETE|PATCH) (.?) (.?)\\\" ([\\d]{3}) ([0-9|-]+) ([0-9|-]+) \\\"([\\^\\"]+)\\\" \\\"([\\^\\"]+)\\\"\\\"

"transforms.RegexTransform.mapping": "IP,RemoteUser,AuthedRemoteUser,DateTime,Method,Request,Protocol,Response,
BytesSent,Ms,Referrer,UserAgent"
```

Usecase2. Transform Config

```
"transforms": "RegexTransform",
"transforms.RegexTransform.type": "org.apache.kafka.connect.transforms.ParseStructByRegex$Value",

"transforms.RegexTransform.regex": "^((https?):\\/\\/\\/([\\^\\/]*))/(.*)"

"transforms.RegexTransform.mapping": "protocol, domain, path"
```

Proposed Changes

Only one main abstract class added : ParseStructByRegex,

this include

1. describe/declare part
2. main functions

1. describe/declare part

```
public abstract class ParseStructByRegex<R> extends ConnectRecord<R> implements Transformation<R> {
    public static final String OVERVIEW_DOC = "Generate key/value Struct objects supported by ordered Regex Group"
        + "<p>Use the concrete transformation type designed for the record key (<code>" + Key.class.getName()
+ "</code>) "
        + "or value (<code>" + Value.class.getName() + "</code>).";

    private interface ConfigName {
        String REGEX = "regex";
        String MAPPING_KEY = "mapping";
    }

    public static final ConfigDef CONFIG_DEF = new ConfigDef()
        .define(ConfigName.REGEX, ConfigDef.Type.STRING, ConfigDef.NO_DEFAULT_VALUE, new GroupRegexValidator(),
ConfigDef.Importance.MEDIUM,
        "String Regex Group Pattern.")
        .define(ConfigName.MAPPING_KEY, ConfigDef.Type.LIST, ConfigDef.NO_DEFAULT_VALUE, ConfigDef.Importance.
MEDIUM,
        "Ordered Regex Group Mapping Keys");

    private static final String PURPOSE = "Parse Struct by regex group mapping";
    ...
}
```

2. main functions interface

[Detail Code](#)

```
@Override
public R apply(R record) {
    if (operatingSchema(record) == null) {
        return applySchemaless(record);
    } else {
        return applyWithSchema(record);
    }
}

private R applySchemaless(R record) {
    ...
}

private R applyWithSchema(R record) {
    ...
}
```

Compatibility, Deprecation, and Migration Plan

this is a new SMT. no older consider. justconfigure connectors with new SMT Config

Rejected Alternatives

