### 分析平台产品说明

##### 产品介绍

本产品支持构建离线批处理、实时流处理的数据处理产品。平台架构于Apache Spark框架之上。

##### 运行方式

平台提供界面操作，通过界面创建数据源、算子，以拖拉拽方式快速生成业务模型，配置任务运行的资源参数。通过shell将组装的spark-submit命令提交至Yarn集群。平台提供监控任务的运行状态，查看任务运行日志等功能。目前循环任务提供了三种运行方式：a. 正常（每到时间节点正常启动任务运行），b. 保留（上一个任务未完成，则下一个任务不会启动），c. 终止（上一个任务未完成，到下一个时间节点则会杀死上一个任务，启动本次任务）

##### 数据源配置

支持对多种数据源的读写，支持kerberos认证

1. HDFS（支持配置不同目录结构）

catalogPattern：

yyyy-MM-dd/timestamp\*

yyyyMMdd/timestamp\*

yyyy-MM-dd/timestamp/\*

yyyyMMdd/timestamp/\*

yyyy-MM-dd/\*

yyyyMMdd/\*

1. Hive
2. Mysql/Oracle/Postgresql/Gaussdb/Gbase
3. 读取
   1. 全量读取
   2. 增量读取（配置增量字段，选取表中的时间戳字段）
   3. 条件读取（配置条件字段）
   4. 并发读取（配置numPartitions,[partColumnName]）
4. 写入

saveType：

* 1. jdbc（使用sparksession api）
  2. insert（使用数据库连接池，组装SQL）
  3. copy（针对Postgresql/Gaussdb）

saveMode：

1. append（追加写入）
2. overwrite （覆盖写入）
3. errorifexist（表若存在不能写入）
4. ignore（表不存在创建表写入，表存在则跳过）
5. ElasticSearch
6. File（json/csv/txt）
7. Kafka

Kerberos认证配置方式：

security.protocol=SASL\_PLAINTEXT

sasl.mechanism=GSSAPI

sasl.kerberos.service.name=kafka

java.security.auth.login.config=./sparkkafkajaas.conf

java.security.krb5.conf=./krb5.conf

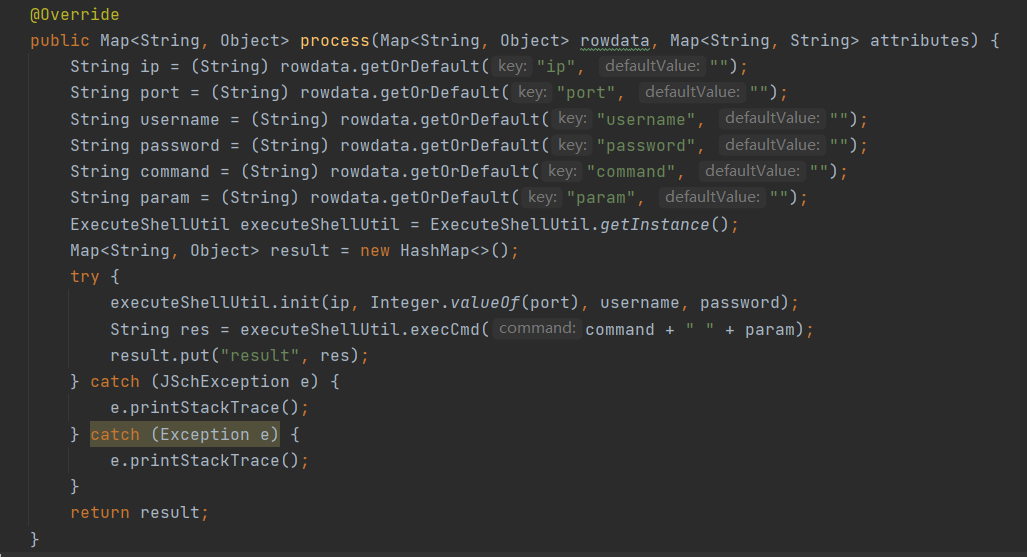
目前平台支持单kafka数据源的消费，不支持kafka+kafka两个集群的消费，不支持kafka+table的关联操作。支持消费的数据格式为：csv、json、avro，支持生产的数据格式为：json。

1. Ftp（json/csv/txt）
2. Http（json）

##### **算子配置**

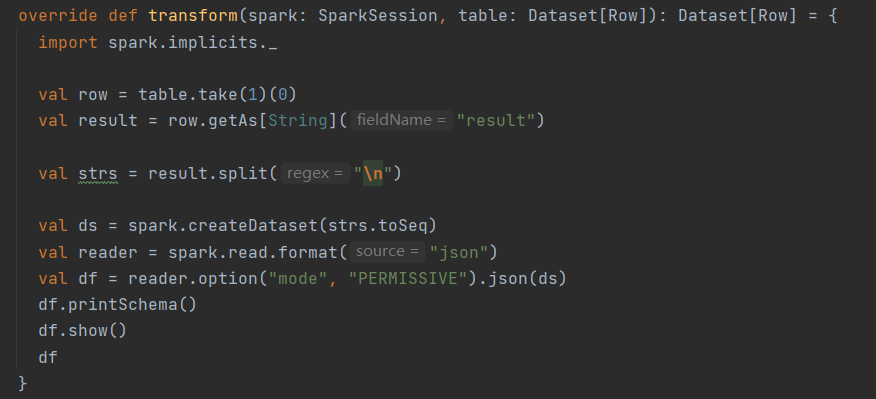
1. Java-jar包

需要实现compute-common包中的ComputeOperator接口，实现Map<String,Object> process(Map<String,Object> rowdata, Map<String,String> attributes); 方法，参数rowdata为从前一个算子流入的一行行数据，简单使用示例：



1. Spark-jar包

需要实现compute-common包中的SparkProcessProxy接口，实现transform(spark: SparkSession, table: Dataset[Row]): Dataset[Row] 方法，简单示例：



1. Java-code



输入参数对应方法中的参数列表，名称不限制，值是从数据流中选择字段，可以设置为定值

输出参数对应方法的返回值，名称和值都不限制，其中值是给后续算子使用的。

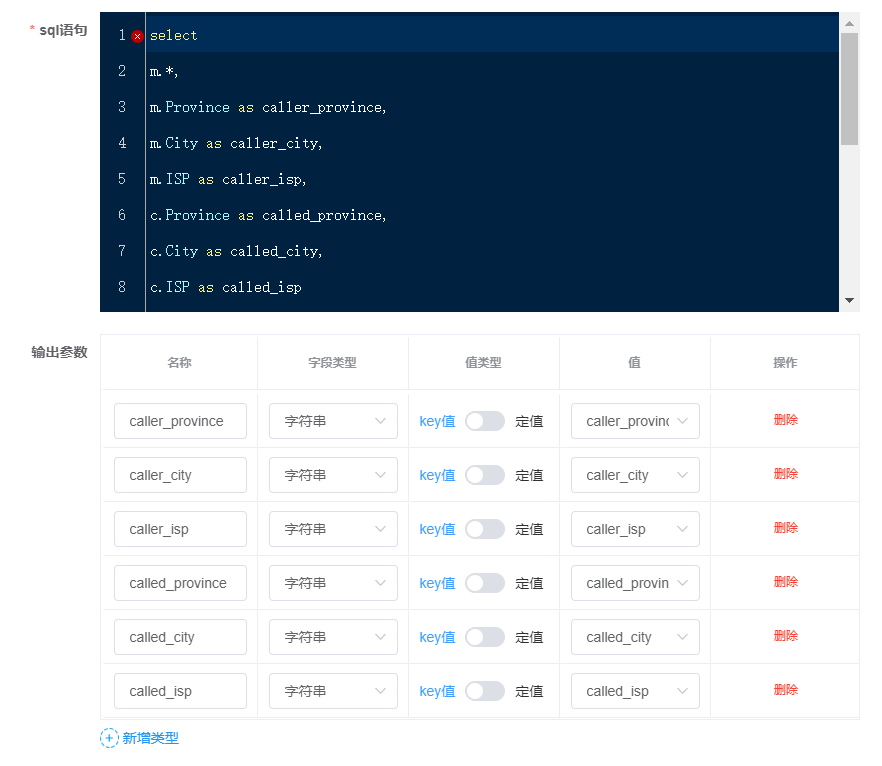
1. Scala-code



Scala-code中的方法参数列表为Seq类型（目前后台支持该种参数列表的解析），输入参数名称不限制，值为数据流中的列或者定值。

输出参数为方法的返回值，名称和值都不限制，其中值是给后续算子使用的

1. SQL



输出参数设置相当于为之前的数据流中增加了一些列，以供后续算子的使用

1. Python算子



1. Shell算子



##### **支持的内置算子**

1. Join (sourceTableName, targetTableName, joinField[List], joinType[default: inner])

关联算子，两个数据源的join操作，提供两个数据源表名、关键字段名、关联类型（默认inner）

1. Repartition (numPartitions)

重分区算子，对原来的数据集重分区，提供分区数量

1. Sample (fraction, limit)

抽样算子，提供抽样因子fraction为0-1之间的double类型数据，系统采用withReplacement=true的方式抽样，limit为-1则拿出所有抽样数据，其他值为只抽取limit条数

1. Checksum (method[CRC32, MD5, SHA1], sourceField, targetField)

校验算子，提供了CRC32、MD5、SHA1三种算法

1. Convert (sourceField, newType)

类型转换算子

1. Constant (constantMap[Map[String, string]])

增加常量列算子

1. Script (JavaCode/ScalaCode/Sql/Python/Shell)

提供编写脚本算子

1. UserDefinedTransform (JavaJar/ScalaJar/JavaCode/ScalaCode/Sql/Python/Shell)

提供上传jar包，编写脚本算子

1. 文件输入

支持json/csv/txt 类型的文件，选择文件上传，如果是csv格式需提供分隔符

1. Siddhi

支持CEP功能，平台集成了Siddhi-CEP模块，支持部分的场景：Filter、Window（lengthBatch/timeBatch）