

KYLIN-5390 Dev Design Build tasks support segment coverage

背景

Why

kylin5构建的时候，不支持大时间范围segment覆盖小时间范围segment数据，比如，当前已有segment时间范围（20220801-20220805），需要提交一个构建任务，时间范围为（20220801-20220806），kylin5报错，kylin3正常。这里会影响到用户的使用习惯。

诉求

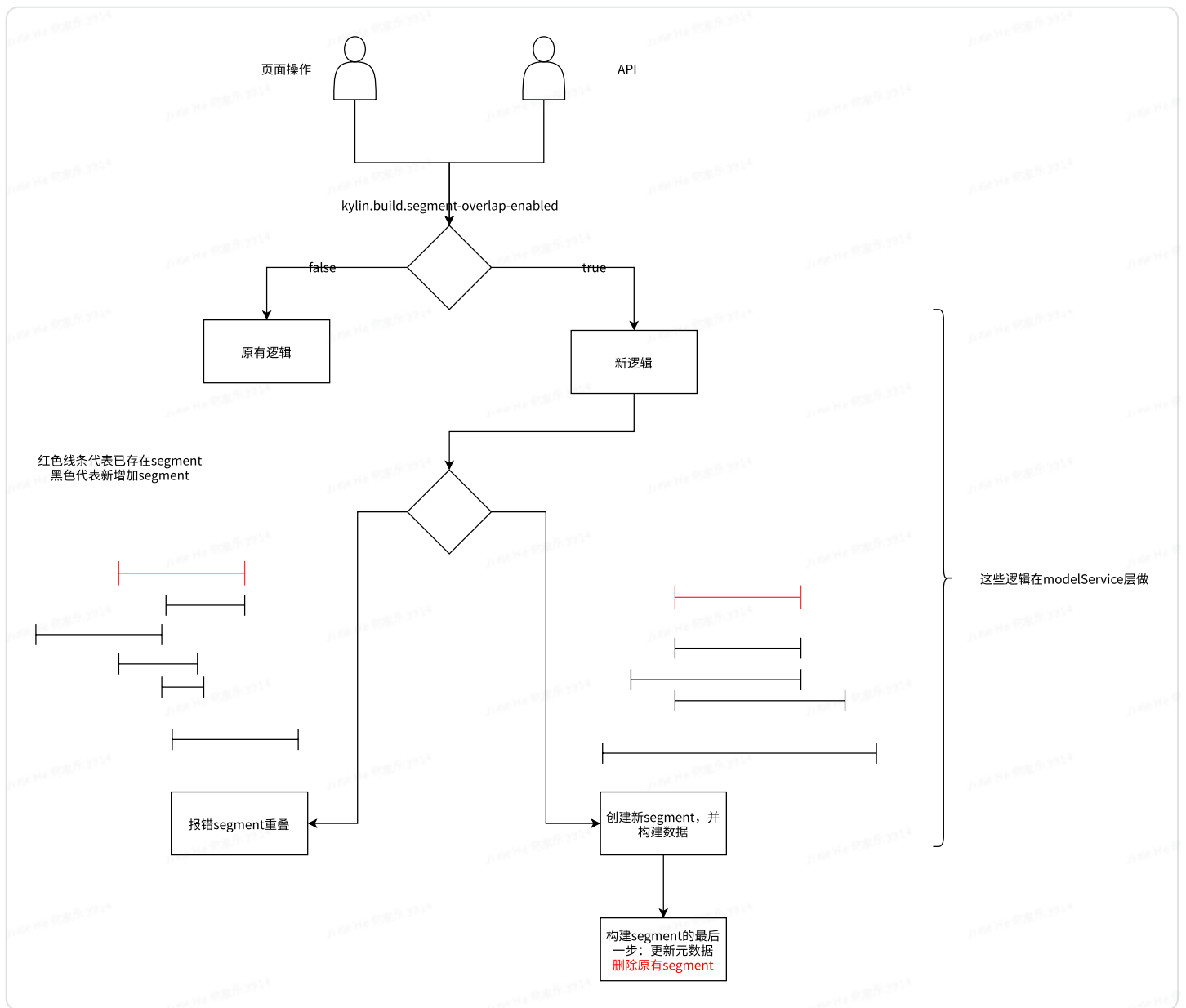
构建时允许大范围segment覆盖小范围segment。KE界面操作和api提交构建/刷新任务都要求支持该功能。

详细设计

Proposol1

新增参数: `kylin.build.segment-overlap-enabled=false(default)`

Segment创建



segment状态更新

当构建新segment时，原有被包含的旧segment需要置为LOCK状态。以及，新segment需要根据覆盖情况设置不同的状态。这些状态什么时候设置？如何设置？

当用户查看segment列表时，会获取当前系统的所有segment(包括新旧segment)，此时设置对用户展示状态

- 旧segment：如果有比其更大的segment处于Loading状态，则将其置为LOCK
- 新segment：如果有和其一样大的segment则将其置为REFRESHE，否则置为LOADING

注：以上行为和kylin4一致

保留segment功能

创建保留segment时，范围包含了旧segment。合规行为？

不支持覆盖，和原有行为一致

< model_test

编辑 构建 更多

基本信息

数据特征

Segment

Segment 列表

+ Segment

刷新

Segment 范围:

开始时间

结束时间

	开始时间	结束时间	索引数	状态	最后更新时间	行数	操作
索引	2022-10-22 00:00:00...	2022-10-26 00:00:00...	3/3	ONLINE	2022-11-03 17:38:52 ...		C 搜索
开发者	2022-10-21 00:00:00...	2022-10-22 00:00:00...	0/3	ONLINE	2022-10-21 18:59:49 ...		C 搜索
	2022-10-20 00:00:00...	2022-10-21 00:00:00...	0/3	ONLINE	2022-10-21 18:59:41 ...		C 搜索

共 3 条

10条/页

1

前往 1 页

加载数据接口

调用加载数据接口，`build_all_indexes`参数为`false`时（效果同保留segment），范围包含了旧segment。产品行为？

当`build_all_indexes`为`false`时，不支持覆盖

Load Segment

- POST `http://host:port/kylin/api/models/{model_name}/segments`
- URL Parameters
 - `model_name` - required string, model name.
- HTTP Body: JSON Object
 - `project` - required string, project name.
 - `start` - optional string, when the model is built in full, it is not required, and it is required when the model is incrementally built. start time of segment (partition column exist), type: timestamp, unit: ms. For example, `694195200000` means `1992-01-01 00:00:00`.
 - `end` - optional string, when the model is built in full, it is not required, and it is required when the model is incrementally built. end time of segment (partition column exist), type: timestamp, unit: ms. For example, `883584000000` means `1998-01-01 00:00:00`.
 - `build_all_indexes` - optional boolean, build all indexes in the new segment, default value is `true`
 - `sub_partition_values` - optional Array, sub-partition value, used for multi-level partition model. the default is empty. For multi-level partition model, when `build_all_indexes` is `true` (all indexes need to be built), this value is required. When `build_all_indexes` is `false` (when creating an empty segment), the value must be empty.

按时间构建segment接口

`buildType`为`REFRESH`时，范围包含了旧segment。产品行为？

kylin5当前行为：如果新segment包含了多个或者不包括旧segment的话则报错。即只能包含一个

上述接口在kylin4中`buildType`没有`REFRESH`，（代码中有）

构建 Cube - 按日期/时间构建

- PUT `http://host:port/kylin/api/cubes/{cubeName}/segments/build`
- URL Parameters
 - `cubeName` - 必选 `string` , Cube 名称
- HTTP Header
 - `Accept: application/vnd.apache.kylin-v2+json`
 - `Accept-Language: en`
 - `Content-Type: application/json;charset=utf-8`
- HTTP Body: JSON Object
 - `startTime` - 必选 `long` , 开始时间, 对应 GMT 格式的时间戳, 如 `1388534400000` 对应 `2014-01-01 00:00:00` , 推荐使用[在线时间戳转换](#)对时间进行处理。
 - `endTime` - 必选 `long` , 结束时间, 对应 GMT 格式的时间戳
 - `buildType` - 必选 `string` , 支持的计算类型, 为: "BUILD"
 - `mpValues` - 可选 `string` , 对应模型的分区字段值
 - `force` - 可选 `boolean` , 强制提交任务选项, 默认值为 `false`
 - `yarnQueue` - 可选 `string` , 指定该任务使用的 YARN 队列, 在系统级别或项目级别设置参数后使用: `kylin.engine-yarn.queue.in.task.enabled` (是否允许为任务指定 YARN 队列, 默认不开启)、`kylin.engine-yarn.queue.in.task.available` (可供设置的 YARN 队列, 多个队列时用英文逗号

多级分区segment

kylin4行为: 新segment包含旧segment的情况下, 如果cube分区值一样则删除原有旧segment, 否则保留新segment。此时系统存在两个重叠segment, 它们的cube分区不一样。

+ Cube 导入

所有模型

Q 输入cube名称筛选

名称	模型	状态	存储空间	源数据条目	最后构建时间	所有者	更新时间	操作
kylin_sales_clone...	kylin_sal...	READY	0.00 KB	0	2022-11-03 20:45:0...	ADMIN	2022-11-03 19:53:2...	

Overview Segments SQL Patterns JSON SQL

刷新 合并 导出 删除

1970-01-01 00:00:00 - 1970-01-01 00:00:00

KYLIN_SALES.OPS_REGION
KYLIN_SALE... shenzhen

Segment ID	Segment 名称	存储空间	源数据条目	起始时间/范围	结束时间/范围	操作
2f0424d0	20220102000000_202201030000...	0.00 KB	0	2022-01-02 00:00:00	2022-01-03 00:00:00	

总大小: 0.00 KB

共 1 条 < 1 > 前往 1 页

kylin5行为: 是否需要和kylin4保持一致。如果保持一致的话, 这里的查询行为一致么?

不支持覆盖

局部构建 (segment异构)

如果segment存在异构, 这里的合并行为?

```
AfterBuildResourceMerger.java × Segments.java × SegmentController.java × BuildSegmentsRequest.java × FusionModelService.java × Mo
BaseController.java × ModelBuildService.java × IncrementBuildSegmentParams.java × ModelService.java × NDataflowManager.java × Segme
ErrorCodeServer.java × CubeRebuildRequest.java × NDataflow.java ×

46     @JsonProperty("sub_partition_values")
47     private List<String[]> subPartitionValues;
48
49     @JsonProperty("build_all_sub_partitions")
50     private boolean buildAllSubPartitions = false;
51
52     private int priority = ExecutablePO.DEFAULT_PRIORITY;
53
54     @JsonProperty("partial_build")
55     private boolean partialBuild = false;
56
57     @JsonProperty("batch_index_ids")
58     private List<Long> batchIndexIds;
59
60     @JsonProperty("yarn_queue")
61     private String yarnQueue;
62
63     @JsonProperty("tag")
64     private Object tag;
```

索引不一致时，即异构，不支持覆盖

当开启开关 `kylin.build.segment-overlap-enabled=true` 让构建支持segment覆盖时，如果开启分层存储会有问题。

这里先将开启分层存储不支持构建重叠segment作为一个已知限制。