streamsets2.6.0导数据常见问题总结

- streamsets2.6.0.1版本使用需要的环境
 - 保证此版本能够正常运行
 - 每个问题都有描述使用场景与解决方法
- streamsets2.6.0.1版本常见问题汇总
- 一、执行任务时有表空间的问题
 - 使用场景
 - 问题描述
 - 问题解答方法
- 二、单表消费oracle数据库表中有Binary_FLOAT字段
 - 使用场景 问题描述

 - 问题解决方法
- 三、读取oracle数据到hive出现null值
 - 使用场景
 - 问题描述
 - 问题解决方法
- 四、读取oracle速度慢的问题
 - 使用场景
 - 问题描述
 - 问题解决方法
- 五、导数据出现字段长度不够的情况
 - 使用场景 问题描述

 - 问题解决方法
- 六、oracle以date类型的字段不能进行实时抓取的问题
 - 使用场景
 - 问题描述
 - 使用联合主键的问题解决方法
- 七、oracle 以Date类型字段按日定时抓取数据
 - 使用场景
 - 问题描述
 - 问题解答方法
- 八、使用streamsets2.6.0.1版本中的Directory组件读取本地文件出现丢数据的情况
 - 使用场景
 - 问题解决方法
- 九、使用jdbc组件向hive导数据时填写Hive元数据处理器注意时的问题
 - 使用场景
 - 问题描述
 - 问题解答方法
- 十、使用JDBC组件导数据时遇到数据库中的字段时中文的情况
 - 使用场景
 - 问题描述
 - 问题解答方法

文档下载:

Streamsets2.6.0导数据常见问题总结.docx

streamsets2.6.0.1版本使用需要的环境

- 保证此版本能够正常运行
- 每个问题都有描述使用场景与解决方法

streamsets2.6.0.1版本常见问题汇总

- 一、执行任务时有表空间的问题
 - 使用场景

以下场景适用于从Oracle或PostgreSQL或其他支持JDBC的数据库,以下错误主要在Oracle数据库中发现,原因在与Oracle中有表空间的概念

· 问题描述

Retries exhausted, giving up after as per stage configuration. First error: java.sql.SQLException: ORA-01652: unable to extend temp segment by 128 in tablespace TEMP



·问题解答方法

这个是由于在oracle数据库中导数据时遇到了大表对数据进行了排序,导致oracle服务器端不能对表进行排序,大的表建议更换为单表消费,直接写SQL的形式。

二、单表消费oracle数据库表中有Binary_FLOAT字段

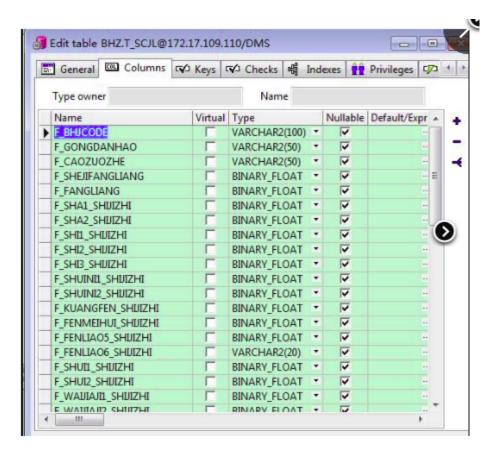
• 使用场景

以下场景适用于从Oracle或PostgreSQL或其他支持JDBC的数据库,以下错误主要在Oracle数据库中发现,原因在与Oracle中有BINARY_FLOAT类型的字段

• 问题描述

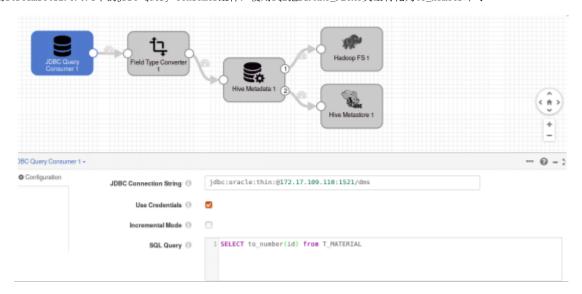
当出现以下错误时,说明JDBC Multitable Consumer读取的数据源中有BINARY_FLOAT字段,JDBC Multitable Consumer组件不支持读取BINARY FLOAT类型的数据





·问题解决方法

使用streamsets2.6.0.1中的JDBC Query Consumer组件,使用SQL把BINARY_FLOAT类型转化为to_number即可



三、读取oracle数据到hive出现null值

• 使用场景

以下场景适用于从Oracle或PostgreSQL或其他支持JDBC的数据库,以下错误主要在Oracle数据库中发现

• 问题描述

3HM1D0BD1



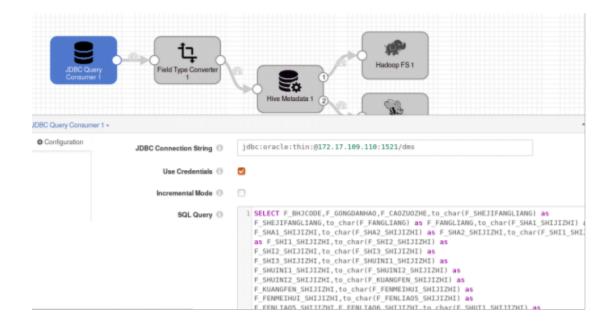
查询历史记录 Q 🖄 保存的查询 Q 🗗 查询生成器

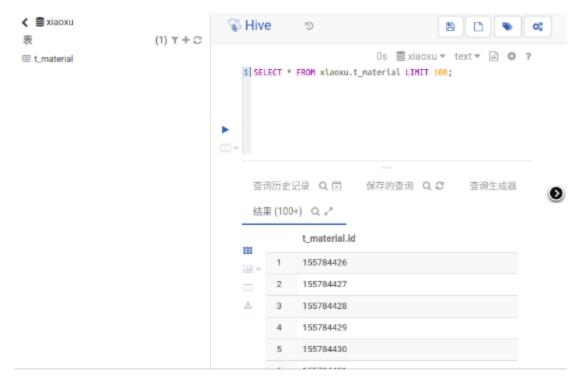
结果(10) Q 🛂

NULL

		t_material.id	t_material.pieceid	t_material.m
.ltl =	1	NULL	8988	3HM1D0BD1
	2	NULL	8988	3HM1D0BD1
<u>±</u>	3	NULL	8988	3HM1D0BD1
	4	NULL	8988	3HM1D0BD1
	5	NULL	8988	3HM1D0BD1
	6	NULL	8988	3HM1D0BD1

8988





• 问题解决方法

解决方式使用streamsets2.6.0.1中的JDBC Query Consumer组件,在SQL Query处写SQL,并使用oracle上的to_char函数来进行转换即可

四、读取oracle速度慢的问题

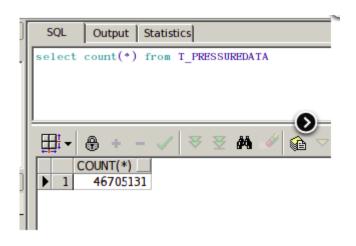
• 使用场景

以下场景适用于从Oracle或PostgreSQL或其他支持JDBC的数据库,以下错误主要在Oracle数据库中发现

• 问题描述

使用streamsets2.6.0.1版本,从oracle数据库中把T_PRESSUREDATA表中的数据导入到Hive中时,出现以下速度慢的问题,目前这个表有46705131行数据,有主键,没有联合主键,使用全量的方式导数据,出现速度在39/S





Name	Virtual	Type		Nullable	Default/Expr.	Storage	Comments	
F_GUID		VARCHAR2(50)	•		211			
F_SBBH		VARCHAR2(500) 🔻	✓	***		777777	
F_RTCODE		VARCHAR2(100		✓	222		????	
F_WTBH		VARCHAR2(100		✓	***		????	
F_SJBH		VARCHAR2(100		✓	***		????	
F_QDD]		VARCHAR2(100		✓			????	
F_SJCC		NUMBER(20)		✓	***		????????	
F_ISW]]		VARCHAR2(200) 🔻	✓			???????	
F_LQ		NUMBER(20)		✓	***		??	
F_KYLZ		NUMBER(10,3)	•	✓			????	
F_SYS]		DATE	•	✓	***		????	
F_OPERATOR		VARCHAR2(200) 🔻	✓			???	
F_YSKYLZ		CLOB		✓	***		m 2	
F_SOFTCOM		VARCHAR2(500) -	✓			?????	
DEVICEID		VARCHAR2(500	•	✓	***		????	
F_SJCC_UPLOAD		VARCHAR2(100		✓			????	
CRRQ		DATE	•	✓	sysdate		????	
STATE		NUMBER(1)	•	✓	0		mmm	
F_VENDER		VARCHAR2(500) 🔻	✓	***		777777	
F_SYLX		VARCHAR2(500		✓	***		??27????	
F_IFSUCCESS		NUMBER		✓	1		????(0,?;1,?)	
F_COLCOM		VARCHAR2(500) 🔻	✓	***		?????	
F_KYQD		NUMBER(10,3)		✓	***		777777	
F_CYMJ		NUMBER(10,3)	•	✓	***		?????	
F_ZCYZ]JL		VARCHAR2(100		V	***		???????	
F_PZ		VARCHAR2(100	•	✓			??	
F_ZZ		VARCHAR2(100		✓	***		??	
F_ISSNKY		VARCHAR2(10)	•	✓	***		`177777 `0777777	
6	V		٠	V	***	***		

·问题解决方法

查看表中的字段中有一个F_YSKYLZ字段是CLOB类型,其中该字段的长度如下所示:

00--500长度的有: 13812236

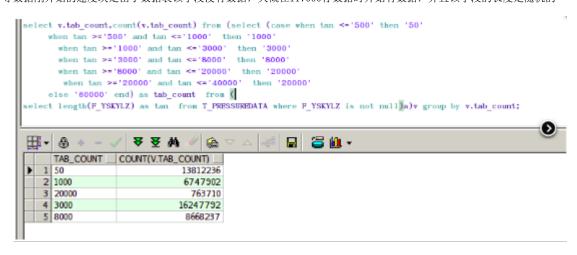
500--1000长度的有: 6747902

8000--20000长度的有: 763710

1000-3000长度的有: 16247792

3000-8000长度的有: 8668237

导数据刚开始的速度块是由于数据表该字段没有数据,大概在117000行数据时开始有数据,并且该字段的长度是随机的



书写的SQL如下:

select v.tab_count, count(v. tab_count, count(v. tab_count) from (select (case when tan <='500' then '50' when tan >='500' and tan <='1000' then '1000' when tan >='3000' and tan <='3000' then '3000' when tan >='3000' and tan <='8000' then '8000' when tan >='8000' and tan <='8000' then '20000' then '20000' when tan >='20000' and tan <='40000' then '20000' else '60000' end) as tab_count from (select length(F_YSKYLZ) as tan from T_PRESSUREDATA where F_YSKYLZ is not null)a)v group by v.tab_count;

经测试如果把F_YSKYLZ表中的字段CLOB类型去掉速度会提升很多,大概在5.8W/S





五、导数据出现字段长度不够的情况

• 使用场景

以下场景适用于从0racle或PostgreSQL或其他支持JDBC的数据库,以下错误主要在0racle数据库中发现

• 问题描述

在streamsets2.6.0.1版本导数据出现HIVE_METADATA_07-Invalld value 0 for precision in filed ID, mininum 1, maxinum 38提示



·问题解决方法

出现这种情况使用streamsets2.6.0.1版本中的JDBC Query Consumer在SQL Query写SQL,使用oracle数据库中的to_number函数来解决或者使用streamsets2.6.0.1中的

Field Tyle Converter 组件对该字段进行转换为STRING即可

六、oracle以date类型的字段不能进行实时抓取的问题

• 使用场景

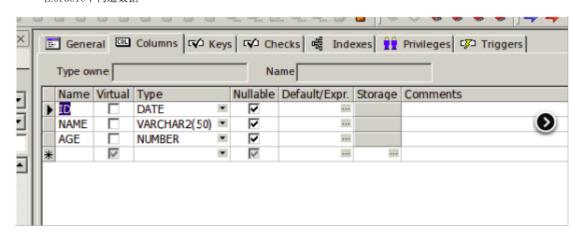
本示例适用于需要实时的读取oracle数据库某个表的数据,但是该表没有ID主键

• 问题描述

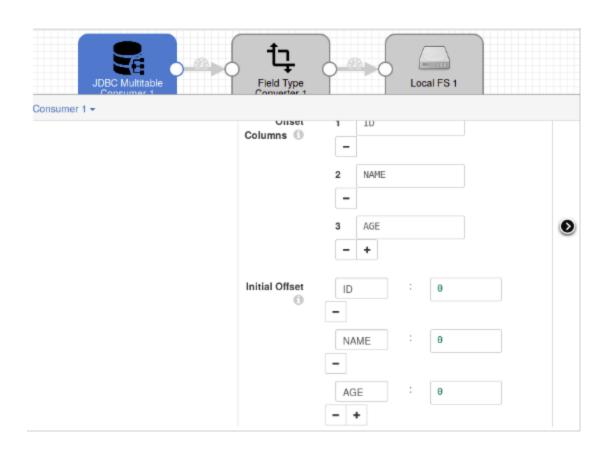
oracle以date类型的字段作为主键不能进行实时抓取,如果主键是number字段且是递增的就可以。原因是date的数据类型精确到日,在这天内可能有多个数据,可能丢数据,不过可以考虑联合主键的方式来使用,不过这种方式也有丢数据的可能性,原因在于联合主键也不能保证这条数据是唯一的

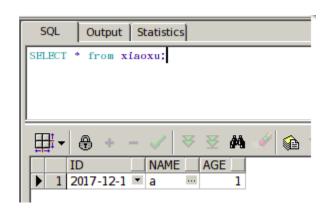
· 使用联合主键的问题解决方法

在oracle中构造数据



在streamsets2.6.0.1中使用JDBC Multitable Consumer组件配置联合主键





由于联合主键也不能确定唯一的记录,这种方法也有可能丢数据,详细请查看官网说明

https://streamsets.com/documentation/datacollector/latest/help/index.html #Origins/MultiTableJDBCConsumer.html

七、oracle 以Date类型字段按日定时抓取数据

• 使用场景

本示例适用于需要实时的读取oracle数据库某个表的数据,但是该表没有ID主键

• 问题描述

一张表中没有主键,但有以date形式的字段作为每天增量的日期,按照日期每天抓取一次,来达到增量的效果

·问题解答方法

官网详细的介绍: https://streamsets.com/documentation/datacollector/latest/help/#Origins/JDBCConsumer.html

中搜索SQL Query for Incremental Mode关键字,可以看出可以在SQL中使用Date来作为增量的字段,其中\${OFFSET}代表date类型的字段,要用单引号扩起来使用,详细如下配置:

使用认证	0	
增量方式	0	
SQL 查询	0	1 select id,name,briday from test where briday > '\${OFFSET}' order by briday
初始化偏移量	0	2017-12-01
偏移列	0	briday
查询间隔	0	\${24*HOURS]}

后台代码会先获取初始化偏移量的值,然后把查询间隔时把获取最新的briday的值重新作为下一个的偏移列,此配置dateTime类型也支持

资料参考:

 $\verb|https://ask.streamsets.com/question/528/cant-jdbc-query-consumer-incremental-mode-work-with-date time-type-off set-column/alternative-type-off set-column/$

八、使用streamsets2.6.0.1版本中的Directory组件读取本地文件出现丢数据的情况

• 使用场景

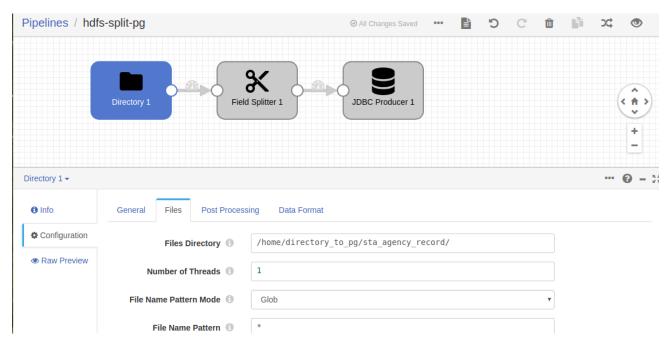
本实例适用于从本地读取带有目录的实时文件的,并把文件进行切割保存到oracle数据库中,也适用于读取Hive的数据

问题描述

使用streamsets2.6.0.1版本中的Directory组件读取刚从hadoop get下的hive数据,数据结构为如下,会出现读取刚下载下来的数据读不到的情况,出现读取数据丢失,本地的数据目录为:

```
# 11
total 0
drwxr-xr-x 5 root root 66 Feb 27 22:01 statist_date=20180202
drwxr-xr-x 3 root root 26 Feb 27 22:02 statist_date=20180203
# 11
total 0
drwxr-xr-x 2 root root 21 Feb 27 22:01 exception_p=1
drwxr-xr-x 2 root root 21 Feb 27 22:01 exception_p=3
drwxr-xr-x 2 root root 21 Feb 27 22:01 exception_p=5
# 11
total 20
-rw-r--r-- 1 root root 18268 Feb 27 22:01 000000_0
# 11
total 4
-rw-r--r-- 1 root root 2031 Feb 27 22:01 0000000 0
```

配置如下,



·问题解决方法

官网说明: https://streamsets.com/documentation/datacollector/latest/help/index.html#datacollector/UserGuide/Origins/Directory.html

解决方法,先把hive数据hadoop get到本地,使用linux mv命令移动到一个新的目录使streamsets2.6.0.1版本中的Directory组件 监控新的目录即可

九、使用jdbc组件向hive导数据时填写Hive元数据处理器注意时的问题

• 使用场景

此场景是用于从支持 jdbc方式的数据库链接,把数据存放到hive中的情况

• 问题描述

在使用Hive元数据处理器组件时,在数据库表达式中使用default数据库报错,错误实例展示



·问题解答方法

改变数据库的名称即可



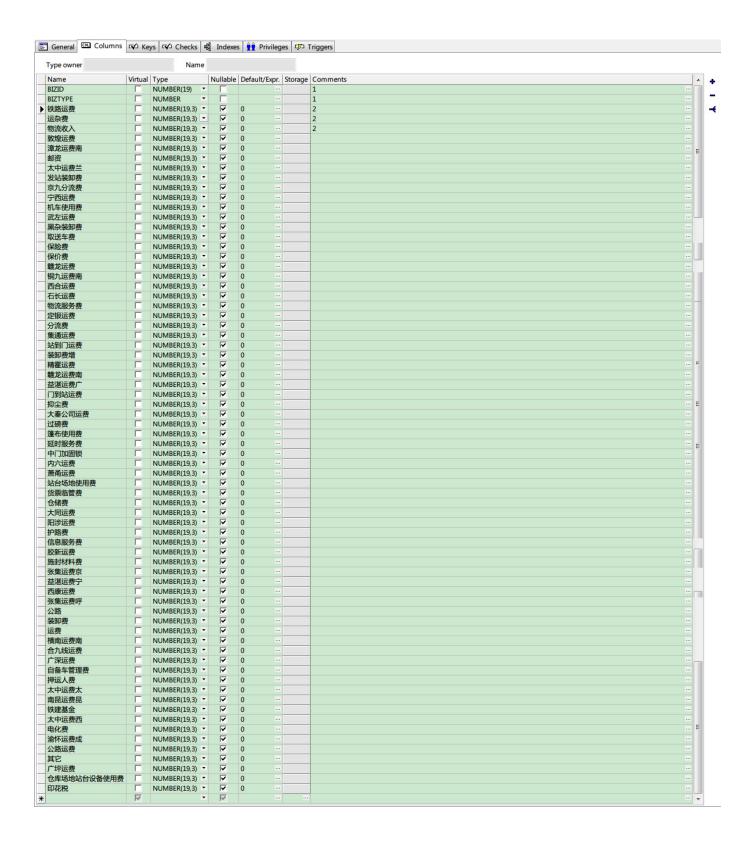
十、使用JDBC组件导数据时遇到数据库中的字段时中文的情况

• 使用场景

此场景是用于从支持 jdbc方式的数据库链接,把数据存放到hive中的情况

• 问题描述

在表中存在中文的情况,在使用jdbc组件导数据时会遇到字段不能解析的情况



	BIZID	BIZTYPE	铁路运费	运杂费	物流收入	敦煌运费	漳龙运费南	邮资	太中运费兰	发站装卸费	京九分流费	宁西运费	机车使用费	武左运费	黑杂装卸费	取送车费
)	1 1013	1	26229.800		2280.000	0.000	0.000	0.000	0.000	0.000		0.000		0.000	0.000	
	2 104	1	28990.800	1842.500	2520.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
	1019	1	32212.100	2002.800	2800.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
	1023	1	23929.000	6495.300	2080.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
	1015	1	26229.800	1705.100	2280.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
	5 1022	1	26229.800	1705.100	2280.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
'	7 105		26229.800			0.000	0.000		0.000	0.000		0.000		0.000		
	1014		26229.800			0.000	0.000		0.000	0.000		0.000		0.000	0.000	
	100		28990.800			0.000	0.000		0.000	0.000		0.000		0.000		
1			28990.800			0.000	0.000		0.000	0.000		0.000		0.000	0.000	
1			32212.100	2002.800		0.000	0.000		0.000	0.000		0.000		0.000	0.000	
1			26229.800			0.000	0.000		0.000	0.000		0.000		0.000	0.000	
1			26229.800			0.000	0.000		0.000	0.000		0.000		0.000		
1			32212.100			0.000	0.000		0.000	0.000		0.000		0.000	0.000	
1			26229.800			0.000	0.000		0.000	0.000		0.000		0.000	0.000	
1			26229.800			0.000	0.000		0.000	0.000		0.000		0.000	0.000	
1			26229.800 28990.800			0.000	0.000		0.000	0.000		0.000		0.000	0.000	
1			32212.100			0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	
2			26229.800			0.000	0.000		0.000	0.000		0.000		0.000	0.000	
2	1016	1	20229.800	1705.100	2280.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00

・问题解答方法

这是由于表的设计不规范所导致的, 修改表字段的名称即可