Tez适配主要步骤:

1. 修改parquet-format-2.1.0源码包:
2. 下载parquet-format-2.1.0源码包:

<https://github.com/apache/parquet-format/archive/refs/tags/parquet-format-2.1.0.tar.gz>

1. 修改src/thrift/parquet.thrift中的enum CompressionCodec{ }, 添加上DATABOX枚举。
2. 本地安装好thrift, 并且需要让parquet-format-2.1.0源码包中的pom.xml中的如下依赖的thrift version改为与本地环境安装的thrift的版本一致。即需要修改:

    <dependency>

      <groupId>org.apache.thrift</groupId>

      <artifactId>libthrift</artifactId>

      <version>0.9.1</version>

     </dependency>

1. 再将parquet-format-2.1.0项目里的pom.xml的maven-thrift-plugin插件的版本由0.1.10改为0.1.11(或更高版本), 因为现在maven-thrift-plugin在各个maven中央仓库中的最低版本就是0.1.11, 如果保持为0.1.10,则用maven会始终编译不成功。
2. 编译并打包发送到本地仓库: mvn install -DskipTests。注意是install, 要将其发布到本地仓库。
3. 在target找到parquet-format-2.1.0.jar。将未修改源码和修改源码后编译的jar包分别解压, 并用diff对比。找到两个jar包的差别, 将差别部分替换掉集群中的相应jar包的部分。
4. 修改源码包parquet-mr-parquet-1.5.0:
5. 下载源码包parquet-mr-parquet-1.5.0:

<https://github.com/apache/parquet-mr/archive/refs/tags/parquet-1.5.0.tar.gz>

1. 修改文件parquet-mr-parquet-1.5.0/parquet-hadoop/src/main/java/parquet/hadoop/metadata/CompressionCodecName.java。
2. 并且把parquet-mr-parquet-1.5.0/pom.xml里面的以下内容注释掉:

     <module>parquet-cascading</module>

     <module>parquet-protobuf</module>

     <module>parquet-scrooge</module>

     <module>parquet-thrift</module>

     <module>parquet-hive</module>

     <module>parquet-hive-bundle</module>

1. 编译: mvn package -DskipTests。
2. 在parquet-mr-parquet-1.5.0/parquet-hadoop-bundle/target目录找到parquet-hadoop-bundle-1.5.0.jar。将未修改源码和修改源码后编译的jar包分别解压, 并用diff对比。找到两个jar包的差别, 将差别部分替换掉集群中的相应jar包的部分。
3. 在parquet-mr-parquet-1.5.0/parquet-hadoop/target目录下找到parquet-hadoop-1.5.0.jar。未修改源码和修改源码后编译的jar包分别解压, 并用diff对比。找到两个jar包的差别, 将差别部分替换掉集群中的相应jar包的部分。
4. 修改集群中hive-exec-1.2.2-bc1.4.0.jar:
5. 将hive-exec-1.2.2-bc1.4.0.jar解压, 找到parquet.thrift, 查看enum CompressionCodec{ }是否正确。
6. 将之前编译的CompressionCodec.class替换掉parquet/format/ CompressionCodec.class。
7. 将之前编译的CompressionCodecName.class替换掉parquet/hadoop/metadata/ CompressionCodecName.class。

适配完成, 运行集群, set hive.execution.engine=tez, 即可进行测试。

**Tez测试使用语句:**

**parquet:**

**SNAPPY:**

create table dns\_parquet\_snappy\_tez stored as parquet tblproperties("[parquet.compression"="](http://parquet.compression/)SNAPPY") as select \* from test\_dns\_500g;

select \* from dns\_parquet\_snappy\_tez where length='716529006' and xdr\_id='593f8aa00000005408175400405ca1d7' and imsi='460028477815285' and imei='8673520217878300' and msisdn='8615849138392' and sgw\_ggsn\_ip\_add='100.232.255.48' and enb\_sgsn\_ip\_add='100.205.133.28' and APN='tmcan.ncb002.hkk460.rpqa' and PROCEDURE\_START\_TIME='1497336478067228' and PROCEDURE\_END\_TIME='1497336481163059' and USER\_IPV4='37.94.228.238' and APP\_SERVER\_IP\_IPV4='22.138.91.253';

**GZIP:**

create table dns\_parquet\_gzip\_tez stored as parquet tblproperties("parquet.compression"="GZIP") as select \* from test\_dns\_500g;

select \* from dns\_parquet\_gzip\_tez where length='716529006' and xdr\_id='593f8aa00000005408175400405ca1d7' and imsi='460028477815285' and imei='8673520217878300' and msisdn='8615849138392' and sgw\_ggsn\_ip\_add='100.232.255.48' and enb\_sgsn\_ip\_add='100.205.133.28' and APN='tmcan.ncb002.hkk460.rpqa' and PROCEDURE\_START\_TIME='1497336478067228' and PROCEDURE\_END\_TIME='1497336481163059' and USER\_IPV4='37.94.228.238' and APP\_SERVER\_IP\_IPV4='22.138.91.253';

**DATABOX:**

create table dns\_parquet\_databox\_tez stored as parquet tblproperties("[parquet.compression"="DATABOX](http://parquet.compression/)") as select \* from test\_dns\_500g;

select \* from dns\_parquet\_databox\_tez where length='716529006' and xdr\_id='593f8aa00000005408175400405ca1d7' and imsi='460028477815285' and imei='8673520217878300' and msisdn='8615849138392' and sgw\_ggsn\_ip\_add='100.232.255.48' and enb\_sgsn\_ip\_add='100.205.133.28' and APN='tmcan.ncb002.hkk460.rpqa' and PROCEDURE\_START\_TIME='1497336478067228' and PROCEDURE\_END\_TIME='1497336481163059' and USER\_IPV4='37.94.228.238' and APP\_SERVER\_IP\_IPV4='22.138.91.253';

**orc:**

**SNAPPY:**

create table dns\_orc\_snappy\_tez stored as orc tblproperties("[orc.compress"="SNAPPY](http://orc.compress/)") as select \* from test\_dns\_500g;

select \* from dns\_orc\_snappy\_tez where length='716529006' and xdr\_id='593f8aa00000005408175400405ca1d7' and imsi='460028477815285' and imei='8673520217878300' and msisdn='8615849138392' and sgw\_ggsn\_ip\_add='100.232.255.48' and enb\_sgsn\_ip\_add='100.205.133.28' and APN='tmcan.ncb002.hkk460.rpqa' and PROCEDURE\_START\_TIME='1497336478067228' and PROCEDURE\_END\_TIME='1497336481163059' and USER\_IPV4='37.94.228.238' and APP\_SERVER\_IP\_IPV4='22.138.91.253';

**ZLIB:**

create table dns\_orc\_zlib\_tez stored as orc tblproperties("orc.compress"="ZLIB") as select \* from test\_dns\_500g;

select \* from dns\_orc\_zlib\_tez where length='716529006' and xdr\_id='593f8aa00000005408175400405ca1d7' and imsi='460028477815285' and imei='8673520217878300' and msisdn='8615849138392' and sgw\_ggsn\_ip\_add='100.232.255.48' and enb\_sgsn\_ip\_add='100.205.133.28' and APN='tmcan.ncb002.hkk460.rpqa' and PROCEDURE\_START\_TIME='1497336478067228' and PROCEDURE\_END\_TIME='1497336481163059' and USER\_IPV4='37.94.228.238' and APP\_SERVER\_IP\_IPV4='22.138.91.253';

**DATABOX:**

create table dns\_orc\_databox\_tez stored as orc tblproperties("[orc.compress"="DATABOX](http://orc.compress/)") as select \* from test\_dns\_500g;

select \* from dns\_orc\_databox\_tez where length='716529006' and xdr\_id='593f8aa00000005408175400405ca1d7' and imsi='460028477815285' and imei='8673520217878300' and msisdn='8615849138392' and sgw\_ggsn\_ip\_add='100.232.255.48' and enb\_sgsn\_ip\_add='100.205.133.28' and APN='tmcan.ncb002.hkk460.rpqa' and PROCEDURE\_START\_TIME='1497336478067228' and PROCEDURE\_END\_TIME='1497336481163059' and USER\_IPV4='37.94.228.238' and APP\_SERVER\_IP\_IPV4='22.138.91.253';

**text:**

**SNAPPY:**

set hive.exec.compress.output=true;

set mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.compress.SnappyCodec;

set mapreduce.output.fileoutputformat.compress.type=BLOCK;

create table dns\_text\_snappy\_tez stored as TEXTFILE as select \* from test\_dns\_500g;

select \* from dns\_text\_snappy\_tez where length='716529006' and xdr\_id='593f8aa00000005408175400405ca1d7' and imsi='460028477815285' and imei='8673520217878300' and msisdn='8615849138392' and sgw\_ggsn\_ip\_add='100.232.255.48' and enb\_sgsn\_ip\_add='100.205.133.28' and APN='tmcan.ncb002.hkk460.rpqa' and PROCEDURE\_START\_TIME='1497336478067228' and PROCEDURE\_END\_TIME='1497336481163059' and USER\_IPV4='37.94.228.238' and APP\_SERVER\_IP\_IPV4='22.138.91.253';

**Gzip:**

set hive.exec.compress.output=true;

set mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.compress.GzipCodec;

set mapreduce.output.fileoutputformat.compress.type=BLOCK;

create table dns\_text\_gzip\_tez stored as TEXTFILE as select \* from test\_dns\_500g;

select \* from dns\_text\_gzip\_tez where length='716529006' and xdr\_id='593f8aa00000005408175400405ca1d7' and imsi='460028477815285' and imei='8673520217878300' and msisdn='8615849138392' and sgw\_ggsn\_ip\_add='100.232.255.48' and enb\_sgsn\_ip\_add='100.205.133.28' and APN='tmcan.ncb002.hkk460.rpqa' and PROCEDURE\_START\_TIME='1497336478067228' and PROCEDURE\_END\_TIME='1497336481163059' and USER\_IPV4='37.94.228.238' and APP\_SERVER\_IP\_IPV4='22.138.91.253';

**DATABOX:**

set hive.exec.compress.output=true;

set mapreduce.output.fileoutputformat.compress.codec=com.hadoop.compression.databox.DataboxCodec;

set mapreduce.output.fileoutputformat.compress.type=BLOCK;

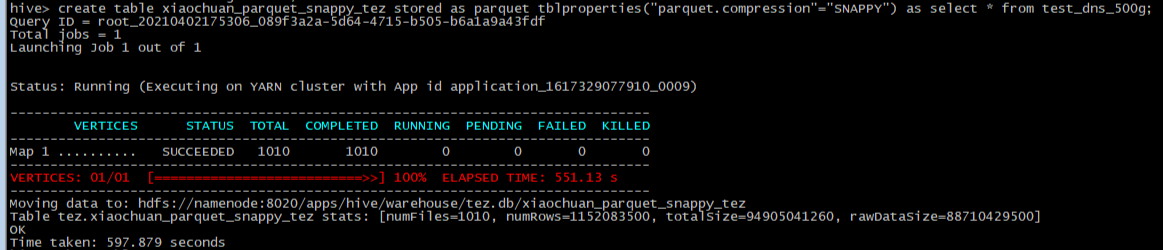
create table dns\_text\_databox\_tez stored as TEXTFILE as select \* from test\_dns\_500g;

select \* from dns\_text\_databox\_tez where length='716529006' and xdr\_id='593f8aa00000005408175400405ca1d7' and imsi='460028477815285' and imei='8673520217878300' and msisdn='8615849138392' and sgw\_ggsn\_ip\_add='100.232.255.48' and enb\_sgsn\_ip\_add='100.205.133.28' and APN='tmcan.ncb002.hkk460.rpqa' and PROCEDURE\_START\_TIME='1497336478067228' and PROCEDURE\_END\_TIME='1497336481163059' and USER\_IPV4='37.94.228.238' and APP\_SERVER\_IP\_IPV4='22.138.91.253';

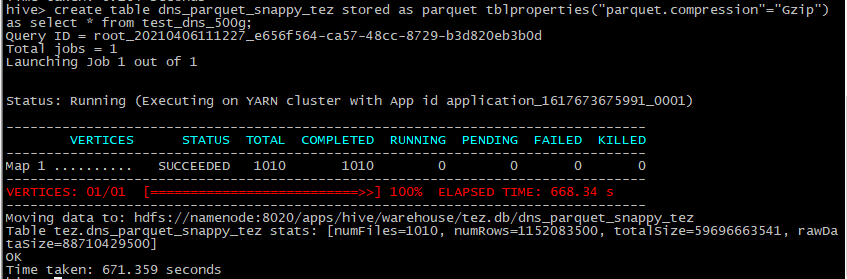
**Tez测试结果截图**

**创建表:**

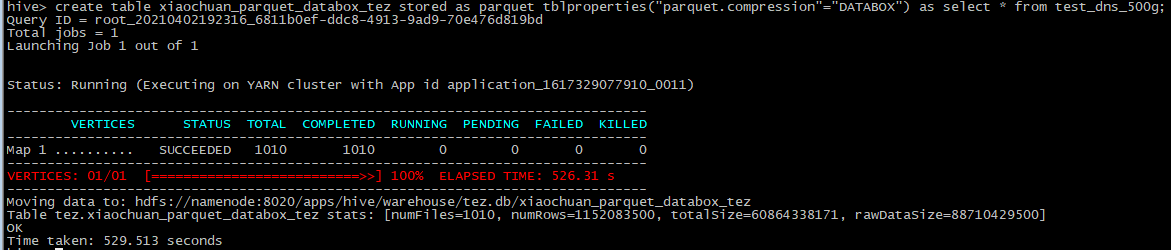
SNAPPY.parquet:



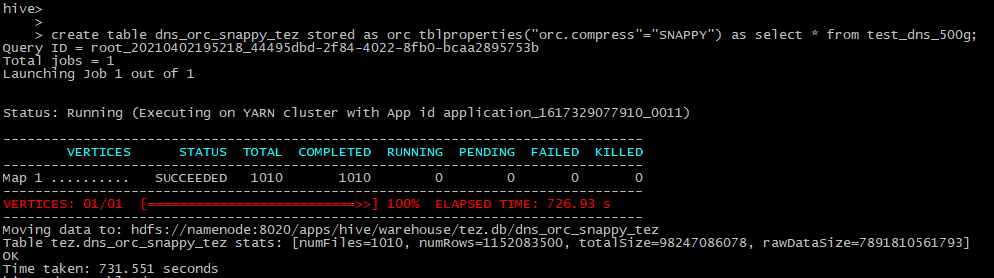
GZIP.parquet:



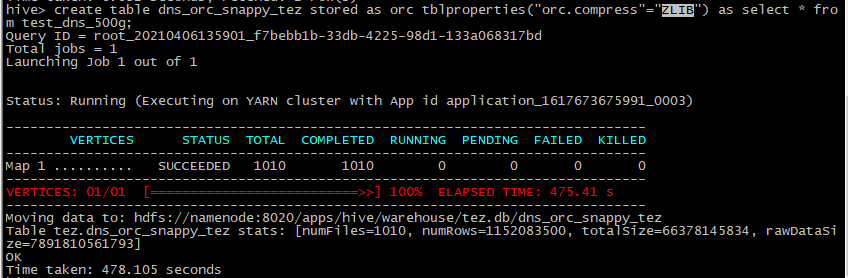
DATABOX.parquet:



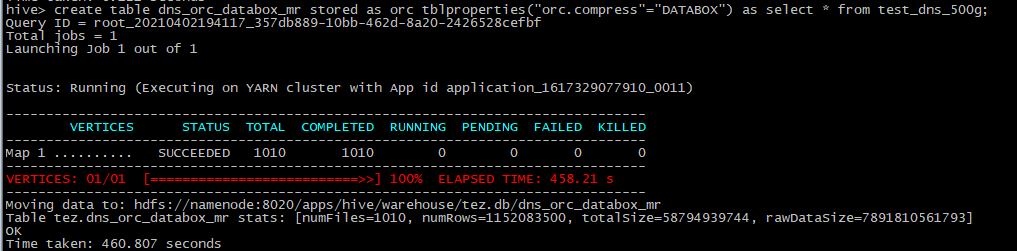
SNAPPY.orc:



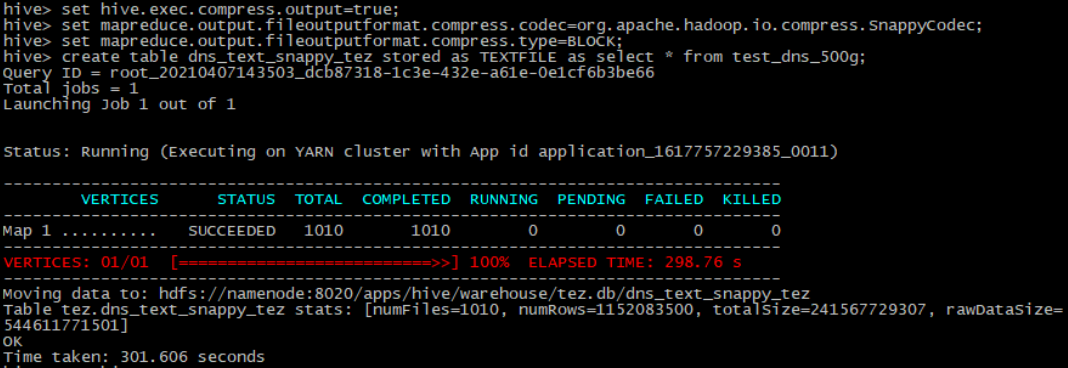
ZLIB.orc



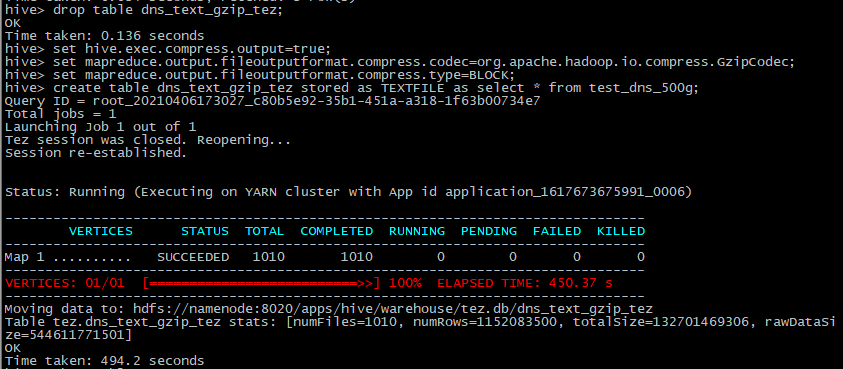
DATABOX.orc



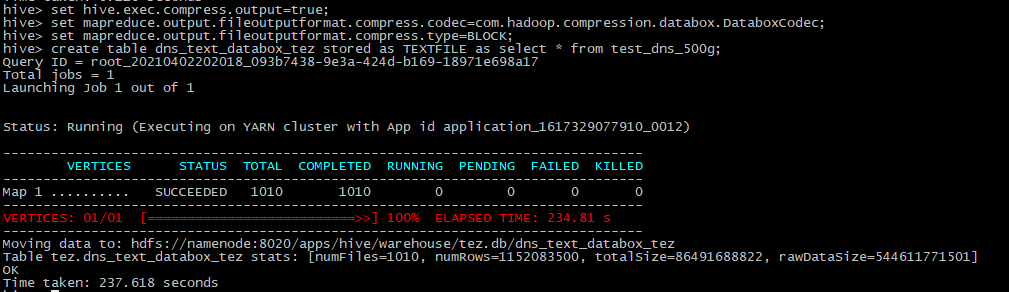
SNAPPY.text



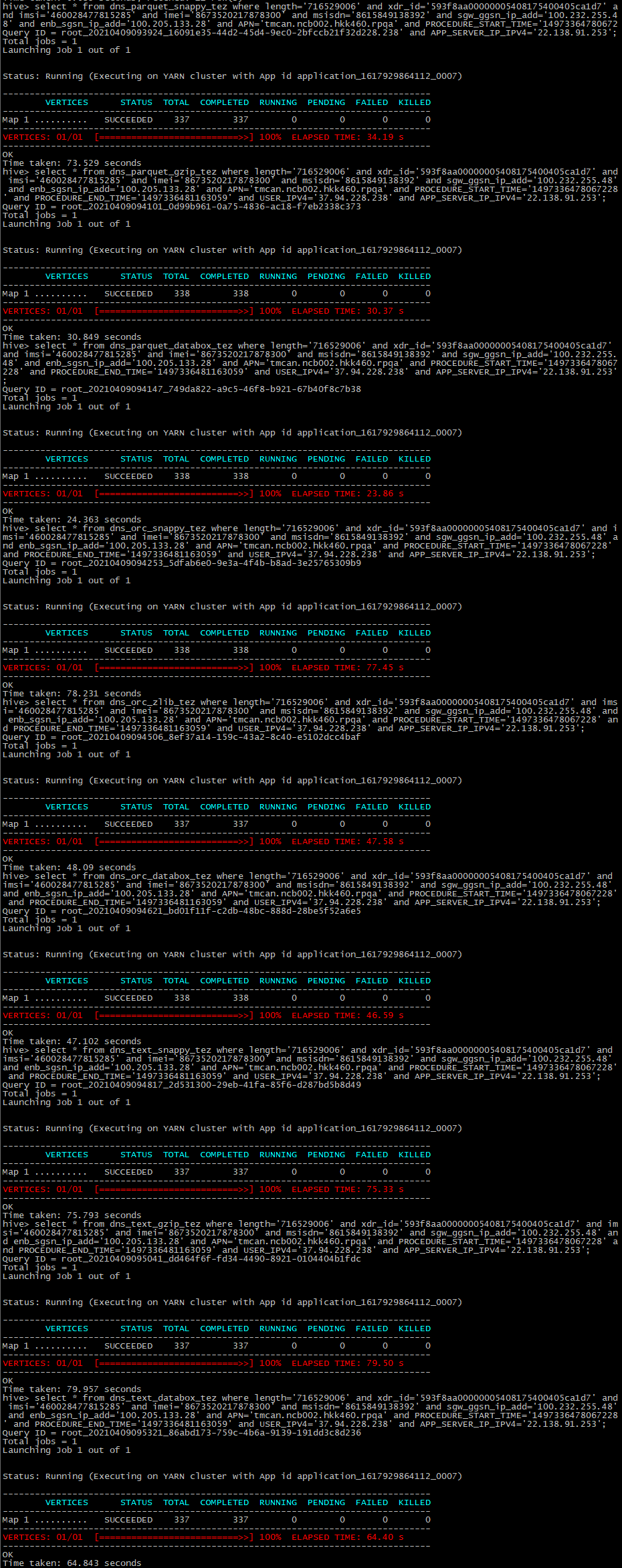
GZIP.text



DATABOX.text



**查询表:**

****