Tungsten Replicator Master Class

Advanced: Working with Data Warehouse Targets

Chris Parker, Customer Success Director, EMEA & APAC

continuent

The MySQL Availability Company

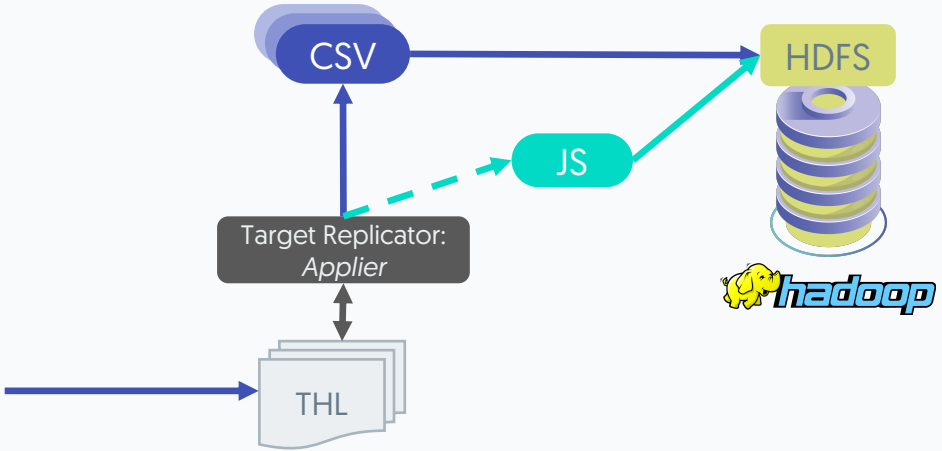
Topics

In this short course, we will

- Review replicator flow
- Explore Hadoop, Redshift and Vertica specific pre-requisites
- Review configurations
- Demo

Replicator Flow

How Hadoop Replication Works



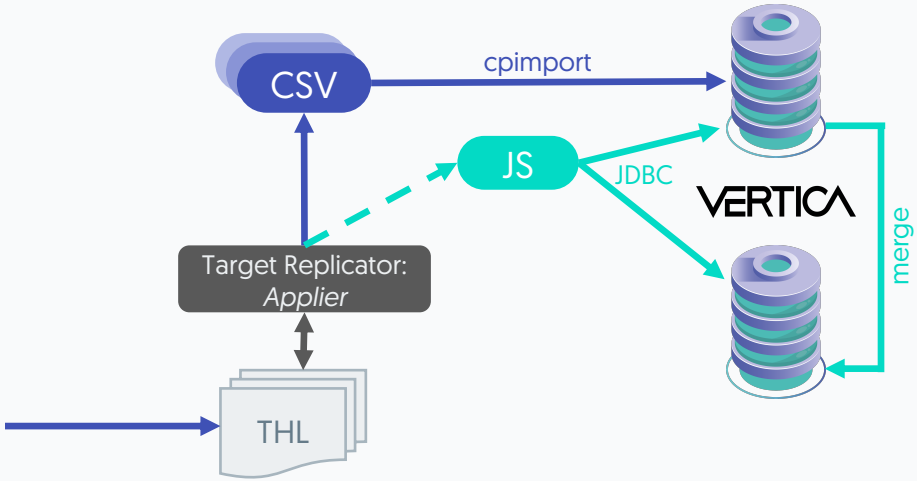
How the Hadoop Materialisation Works

```
insert into t1
   values (1,"Hello World!");

insert into t1
   values (2,"Meet Continuent");

update t1
   set msg="Goodbye World"
   where ID = 1;
```

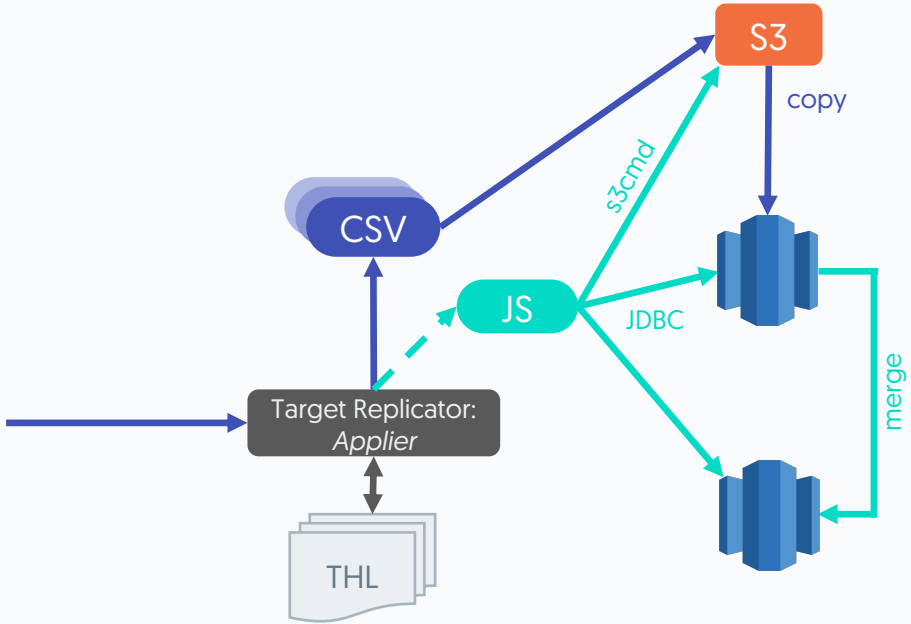
Op	Seqno	ID	Msg
I	1	1	Hello World!
I	2	2	Meet Continuent
D	3	1	
I	3	1	Goodbye World

Op	Seqno	ID	Msg
I	2	2	Meet Continuent
I	3	1	Goodbye World

How Vertica Replication Works



How Redshift Replication Works



Prerequisites

- Review online documentation
  - https://docs.continuent.com
  - Download the Prerequisite Checklist
- Extractor/Applier Hosts
  - OS User
  - /etc/hosts
  - sudoers
  - Ruby
  - Java
- Network
  - Review Port Requirements
- MySQL
  - my.cnf settings
  - User accounts
- Hadoop
  - HDFS writeable by replicator user
- Vertica
  - User Accounts
  - JDBC Drivers
- Redshift
  - User Accounts
  - S3 Bucket
  - S3 Tools for uploading
  - AWS JSON Config
- All tables need Primary Keys

AWS JSON Config for Redshift

- /opt/continuent/share/s3-config-<servicename>.json
- `awsS3Path` — the location within your S3 storage where files should be loaded.
- `awsAccessKey` — the S3 access key to access your S3 storage. Not required if awsIAMRole is used.
- `awsSecretKey` — the S3 secret key associated with the Access Key. Not required if awsIAMRole is used.
- `awsIAMRole` — the IAM role configured to allow Redshift to interact with S3. Not required if `awsAccessKey` and `awsSecretKey` are in use.
- `s3Binary` — the binary to use for loading csv file up to S3. (Valid Values: `s3cmd, s4cmd, aws`) (Default: `s3cmd`)
- `cleanUpS3Files` - a boolean value used to identify whether the CSV files loaded into S3 should be deleted after they have been imported and merged (Default: `true`)

Provisioning Options

- Traditional CSV export and import
- Dump and load through Blackhole engine
- If target support standard SQL, extract data as INSERTS
- For Hadoop, use Sqoop

Object Mapping

- Hadoop
  - MySQL Database → HDFS Directory
  - Table → Hive Compatible CSV File
  - Row → Line in the file

- Redshift & Vertica (PostgreSQL Interface and Syntax)
  - MySQL Instance → Database
  - MySQL Database → Schema

DDLScan

- Reverse engineers MySQL objects
- Creates target specific DDL
- Can be used for single objects or entire databases
- Must be configured prior to starting replicators
- Must be run twice, once for base tables, once for staging tables

```
ddlscan -service alpha -template ddl-mysql-redshift.vm -db test >ddl.sql
ddlscan -service alpha -template ddl-mysql-redshift-staging.vm -db test
>ddl-staging.sql
```

Extractor Config

```
[defaults]
user=tungsten
install-directory=/opt/continuent
mysql-allow-intensive-checks=true
profile-script=~/.bash_profile
disable-security-controls=true

[alpha]
master=tr-ext-2
members=tr-ext-2
replication-user=tungsten
replication-password=secret
replication-port=3306
enable-heterogeneous-service=true
```

Applier Configs

Redshift

```
[defaults]
user=tungsten
install-directory=/opt/continuent
profile-script=~/.bash_profile
disable-security-controls=true

[alpha]
master=tr-ext-2
members=tr-app-1
datasource-type=redshift
replication-user=dbadmin
replication-password=Secret123
replication-port=5439
replication-host=redshift-endpoint
redshift-dbname=demo
batch-enabled=true
batch-load-template=redshift
svc-applier-block-commit-interval=30s
svc-applier-block-commit-size=250000
```

Vertica

```
[defaults]
user=tungsten
install-directory=/opt/continuent
profile-script=~/.bash_profile
disable-security-controls=true

[alpha]
master=tr-ext-2
members=verticahost
datasource-type=vertica
replication-user=dbadmin
replication-password=Secret123
replication-port=5433
vertica-dbname=demo
batch-enabled=true
batch-load-template=vertica6
batch-load-language=js
svc-applier-block-commit-interval=30s
svc-applier-block-commit-size=250000
```

Applier Configs

Hadoop

```
[defaults]
user=tungsten
install-directory=/opt/continuent
profile-script=~/.bash_profile
disable-security-controls=true

[alpha]
master=tr-ext-2
members=hadoopapplier
datasource-type=file
property=replicator.datasource.global.csvType=hive
replication-user=tungsten
replication-password=secret
batch-enabled=true
batch-load-template=hadoop
batch-load-language=js
svc-applier-block-commit-interval=30s
svc-applier-block-commit-size=250000
```

Demonstration

Summary

What we have learnt today

- Reviewed replicator flow
- Explored Hadoop, Redshift and Vertica specific pre-requisites
- Reviewed configurations

Next Steps

In the next session we will

- Learn how to use Tungsten Replicator with MongoDB

THANK YOU FOR LISTENING

continuent.com

Chris Parker, Customer Success Director, EMEA & APAC

continuent

The MySQL Availability Company