Amazon S3 Export Integration v2

Amazon Simple Storage Service (Amazon S3) is an object storage service that offers scalability, data availability, security, and performance. Amazon S3 provides features for data organization and configuration of access controls for your business, organization, and compliance requirements.

This TD export integration allows you to write job results from Treasure Data directly to Amazon S3.

What can you do with this Integration?

- **Create buckets**: Create and name a bucket that stores data.
- **Storing data**: Store an infinite amount of data in a bucket.

Differences between Amazon S3 Export Integration v2 and Amazon S3 Export Integration v1

Review information in the following table to understand the differences and potential advantages between v2 and v1.

Feature	Amazon S3 v2	Amazon S3 v1
Server-side Encryption with Customer Master Key (CMK) stored in AWS Key Management Service	**X**	
Output data size limit	**5TB**	**100GB**
Support for Quote Policy for output data format	**X**	

This topic includes:

Prerequisites

- Basic knowledge of Treasure Data, including the TD Toolbelt.
- For AWS: the IAM User
    - with s3:PutObject, s3:AbortMultipartUpload permissions
    - with kms:Decrypt, kms:GenerateDataKey* permissions when selecting sse-kms setting

Requirements and Limitations

- The default query result limit for export to S3 is 100GB. you could config part size setting up to 5000 (MB), the file limit will be about 5TB.
- The default export format is CSV RFC 4180.
- Output in TSV, JSONL format is also supported.

About S3 Server-Side Encryption

You can encrypt upload data with AWS S3 Server-Side Encryption. You don't need to prepare an encryption key. Data will be encrypted at the server side with 256-bit Advanced Encryption Standard (AES-256).

Use the Server-Side Encryption bucket policy if you require server-side encryption for all objects that are stored in your bucket. When you have server-side encryption enabled, you don't have to turn on the SSE option. However, job results might fail if you have bucket policies to reject HTTP requests without encryption information.

About KMS Server-Side Encryption

You can encrypt upload data with Amazon S3-managed encryption keys (SSE-S3).

When you enable AWS KMS for server-side encryption in Amazon S3

- if not input KMS key id, it will create/using the default KMS key
- if input KMS Key ID, you must choose asymmetric CMK, not asymmetric CMKs
- The AWS KMS CMK must be in the same Region as the bucket

About File Formats for S3

For both CSV, TSV, JSONL formats, the following table lists options you can use to customize the final format of the files written into the destination:

Name	Description	Restrictions	CSV default	TSV default	JSONL
format	Main setting to specify the file format		csv	csv (Use 'tsv' to select the TSV format)	Use JSONL to select JSONL format
delimiter	Use to specify the delimiter character		, (comma)	\t (tab)	parameter ignored
quote policy	Use to determine field type to quote		MINIMAL	MINIMAL	parameter ignored
quote	Use to specify the quote character	not available for TSV format	" (double quote)	(no character)	parameter ignored
escape	Specifies the character used to escape other special characters	not available for TSV format	" (double quote)	(no character)	parameter ignored
null	Use to specify how a 'null' value is displayed		(empty string)	\N (backslash capital n)	parameter ignored
newline	Use to specify the EOL (End-Of-Line) representation		\r\n (CRLF)	\r\n (CRLF)	\r\n (CRLF)
header	Can be used to suppress the column header		column header printed. Use 'false' to suppress	the column header printed. Use 'false' to suppress	parameter ignored

The following example shows a default sample output in CSV format when no customization is requested:

```
code,cnt
200,4981
302,
404,17
500,2
```

When the format=tsv, delimiter=|, and null=NULL options are specified. The output changes to:

```
code|cnt
200|4981
302|NULL
404|17
500|2
```

When the format=jsonl. The output changes to:

```
{"code": 200, "cnt": 4981}
{"code": 302, "cnt": null}
{"code": 404, "cnt": 17}
{"code": 500, "cnt": 2}
```

Use the TD Console to Create a Connection

In Treasure Data, you must create and configure the data connection prior to running your query. As part of the data connection, you provide authentication to access the integration.

Create a New Authentication

1. Open **TD Console**.
2. Navigate to **Integrations Hub** > **Catalog**.
3. Search for S3 and select AmazonS3.

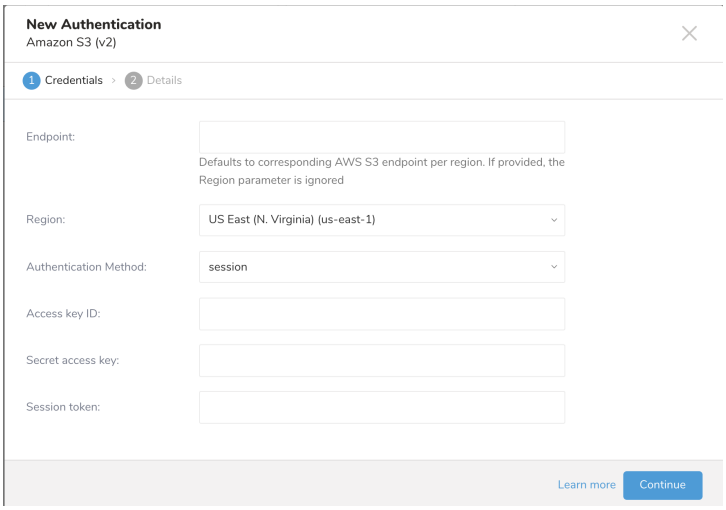
4. Select Create Authentication.
5. Type the credentials to authenticate:

Parameter	Description	
Endpoint	S3 service endpoint override. You can find region and endpoint information from AWS Document. (Ex. *s3.ap-northeast-1.amazonaws.com*) `When specified will override region setting`	
Region	AWS Region	

Authentication Method	**basic**	• Uses access_key_id and secret_access_key to authenticate. See AWS Programmatic access.
		  • Access Key ID
		  • Secret access key

	session (Recommended)	• Uses temporary-generated access_key_id, secret_access_key and session_token.
		  • Access Key ID
		  • Secret access key
		  • Secret token

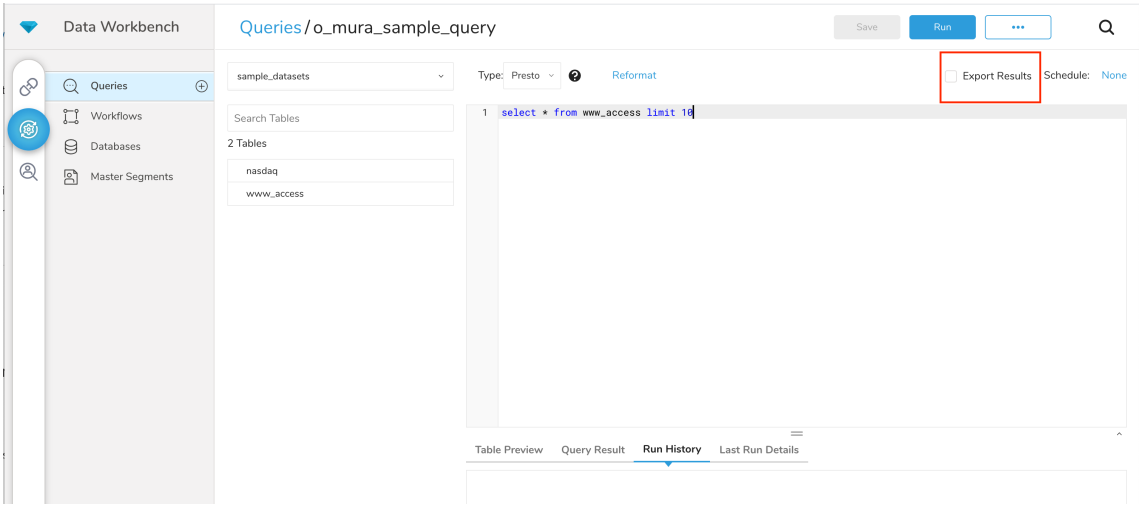
	anonymous	Not Support
Access Key ID	AWS S3 issued	
Secret Access Key	AWS S3 issued	



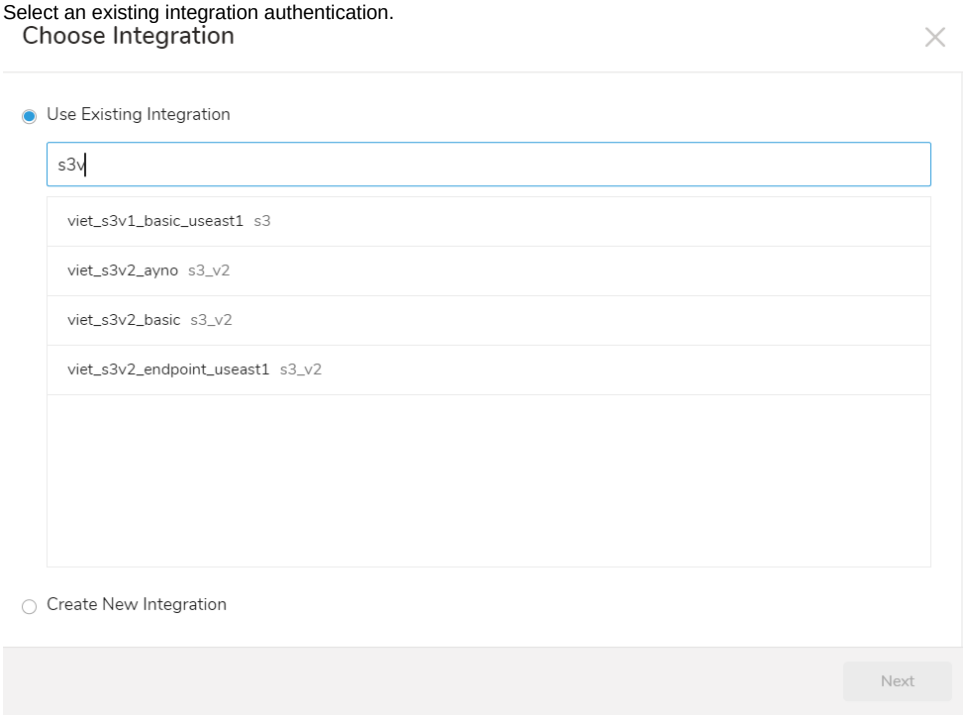
6. Select **Continue**.
7. Type a name for your connection.
8. Select Done.

Define your Query

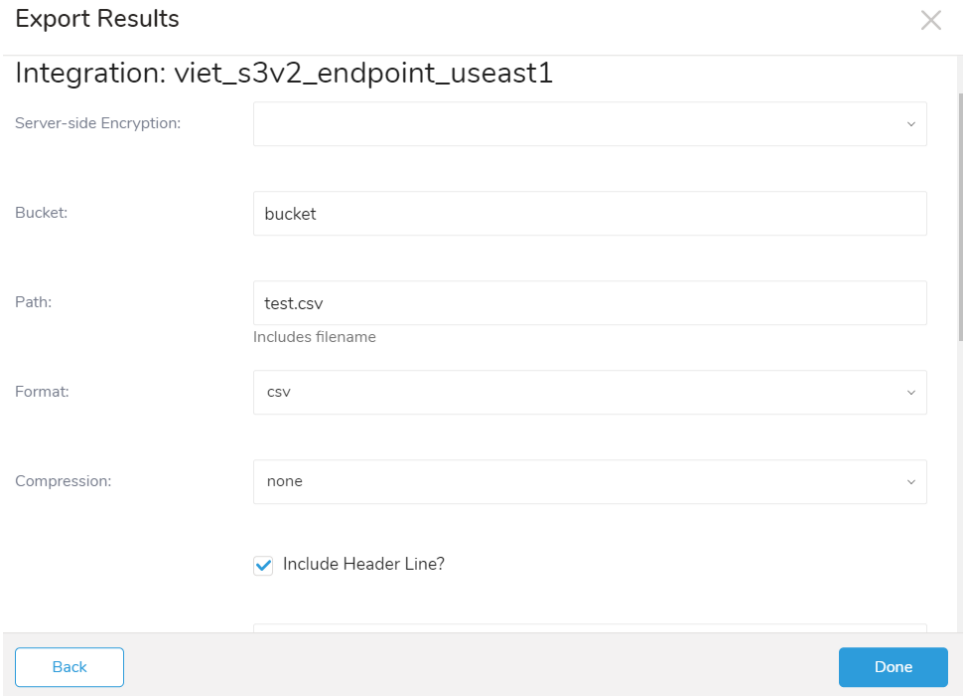
1. Complete the instructions in Creating a Destination Integration.
2. Navigate to **Data Workbench > Queries**.
3. Select a query for which you would like to export data.
4. Run the query to validate the result set.
5. Select **Export Results**.



1. Select an existing integration authentication.

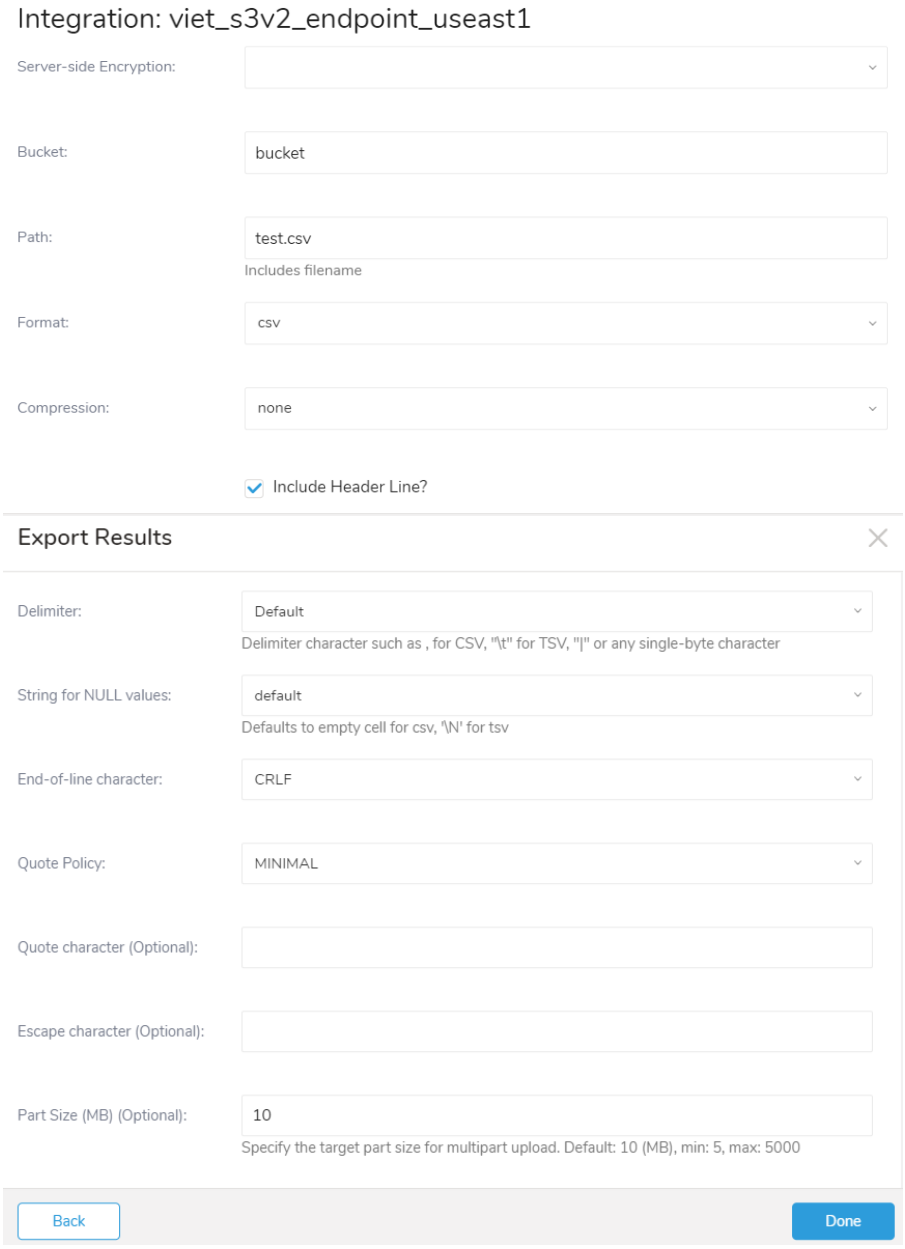


2. Define any additional Export Results details. In your export integration content review the integration parameters. For example, your Export Results screen might be different, or you might not have additional details to fill out:



3. Select **Done.**
4. Run your query
5. Validate that your data moved to the destination you specified.

IntegratioExportn Parameters for S3



Parameter	Data Type	Required?	Supported in V1?	Description
Server-side Encryption	String		yes, only sse-s3	Support values: • sse-s3: Server-side Encryption Mode • sse-kms: new SSE Mode
Server-side Encryption Algorithm	String		yes	Support value: • SEA256
KMS Key ID	String		no	Symmetric AWS KMS Key Id, if not input KMS key id, it will create/using the default KMS key
Bucket	String	yes	yes	Provide the S3 bucket name (Ex. your_bucket_name)
Path	String	yes	yes	Specify s3 filename (object key), include an extension (Ex. test.csv)
Format	String		yes	Format of the exported file: *csv*, *tsv, jsonl*
Compression	String		yes	The compression format of the exported files *(Ex. None or gz)*
Header	Boolean		yes	Include header in the exported file

Delimiter	String		yes	Use to specify the delimiter character *(Ex, (comma))*
String for NULL values	String		yes	Placed holder to insert for null values *(Ex. Empty String)*
End-of-line character	String		yes	Specify the EOL(End-Of-Line) representation *(Ex. CRLF, LF)*
Quote Policy	String		no	Use to determine field type to quote. Support values: • ALL Quote all fields • MINIMAL Only quote those fields which contain delimiter, quote or any of the characters in lineterminator • NONE Never quote fields. When the delimiter occurs in field, escape with escape char Default value: MINIMAL
Quote character (Optional)	Char		yes	The character used for quotes in the exported file(Ex. "). Only quote those fields which contain delimiter, quote, or any of the characters in lineterminator. If the input is more than 1 character, the default value will be used
Escape character (Optional)	Char		yes	The escape character is used in the exported file. If the input is more than 1 character, the default value will be used
Part Size (MB) (Optional)	Integer		no	The part size in multipart upload Default 10, min 5, max 5000

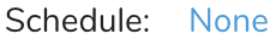
Example Query

```
SELECT * FROM www_access
```

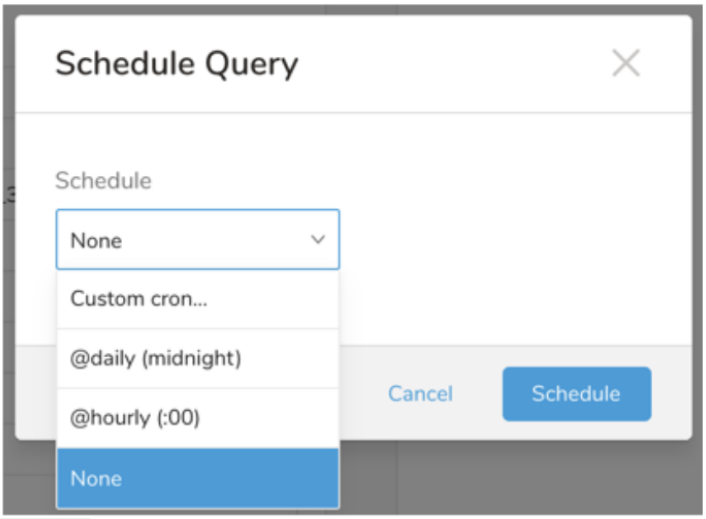
(Optional) Schedule the Query

You can use Scheduled Jobs with Result Export to periodically write the output result to a target destination that you specify.

1. Navigate to **Data Workbench > Queries**.
2. Create a new query or select an existing query.
3. Next to **Schedule**, select None.

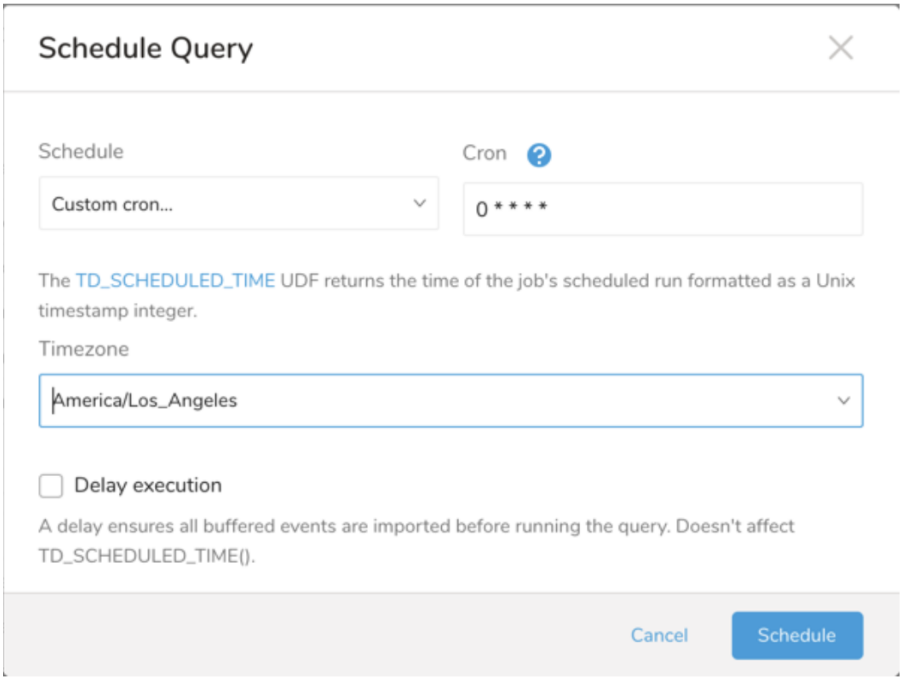


4. In the drop-down, select one of the following schedule options.



Drop-down Value	Description
Custom cron...	Review Custom cron... details.
@daily (midnight)	Run once a day at midnight (00:00 am) in the specified time zone.
@hourly (:00)	Run every hour at 00 minutes.
None	No schedule.

Custom cron... Details



Cron Value	Description
`0 * * * *`	Run once an hour
`0 0 * * *`	Run once a day at midnight
`0 0 1 * *`	Run once a month at midnight on the morning of the first day of the month
""	Create a job that has no scheduled run time.

```
*    *    *    *    *
-    -    -    -    -
|    |    |    |    |
|    |    |    |    +----- day of week (0 - 6) (Sunday=0)
|    |    |    +---------- month (1 - 12)
|    |    +--------------- day of month (1 - 31)
|    +-------------------- hour (0 - 23)
+------------------------- min (0 - 59)
```

The following named entries can be used:

- Day of Week: sun, mon, tue, wed, thu, fri, sat
- Month: jan, feb, mar, apr, may, jun, jul, aug, sep, oct, nov, dec

A single space is required between each field. The values for each field can be composed of:

Field Value	Example	Example Description
a single value, within the limits displayed above for each field.		
a wildcard `'*'` to indicate no restriction based on the field.	`'0 0 1 * *'`	configures the schedule to run at midnight (00:00) on the first day of each month.
a range `'2-5'`, indicating the range of accepted values for the field.	`'0 0 1-10 * *'`	configures the schedule to run at midnight (00:00) on the first 10 days of each month.
a list of comma-separated values `'2,3,4,5'`, indicating the list of accepted values for the field.	`0 0 1,11,21 * *'`	configures the schedule to run at midnight (00:00) every 1st, 11th, and 21st day of each month.
a periodicity indicator `'*/5'` to express how often based on the field's valid range of values a schedule is allowed to run.	`'30 */2 1 * *'`	configures the schedule to run on the 1st of every month, every 2 hours starting at 00:30. `'0 0 */5 * *'` configures the schedule to run at midnight (00:00) every 5 days starting on the 5th of each month.
a comma-separated list of any of the above except the `'*'` wildcard is also supported `'2,*/5,8-10'`.	`'0 0 5,*/10,25 * *'`	configures the schedule to run at midnight (00:00) every 5th, 10th, 20th, and 25th day of each month.

5. (Optional) If you enabled the Delay execution, you can delay the start time of a query.

Execute the Query

Save the query with a name and run, or just run the query. Upon successful completion of the query, the query result is automatically imported to the specified container destination.

> Scheduled jobs that continuously fail due to configuration errors may be disabled on the system side after several notifications.

(Optional) Configure Export Results in Workflow

Within Treasure Workflow, you can specify the use of this data connector to export data.

Learn more at Using Workflows to Export Data with the TD Toolbelt.

S3 (v2) Configuration Keys

Name	Type	Required	Description
bucket	String	Yes	

path	String	Yes	
sse_type	String		sse-s3, sse-kms
`sse_algorithm`	String		`AES256`
`kms_key_id`	String		
format	String		`csv, tsv, jsonl`
compression	String		none, gz
header	Boolean		Default true
delimiter	String		default , \t |
null_value	String		default, empty, \\N, NULL, null
newline	String		CR, LF, CRLF
quote_policy	String		ALL, MINIMAL, NONE
escape	Char		
quote	Char		
part_size	Integer		

Example Workflow for S3 (v2)

```
_export:
  td:
  database: td.database

+s3v2_test_export_task:
  td>: export_s3v2_test.sql
  database: ${td.database}
  result_connection: s3v2_conn
  result_settings:
          bucket: my-bucket
          path: /path/to/target.csv
          sse_type: sse-s3
          format: csv
          compression: gz
          header: false
    delimiter: default
    null_value:  empty
    newline: LF
          quote_policy: MINIMAL
          escape: '"'
          quote: '"'
          part_size: 20
```

(Optional) Configure Export Results in CLI

To output the result of a single query to an S3 buck add the --result option to the td query command. After the job is finished, the results are written into your s3
You can specify detailed settings to export your S3 via --result parameter.

Example for CLI command for S3 (v2)

```
td query \
--result '{"type":"s3_v2","auth_method":"basic","region":"us-east-2","access_key_id": "************","
secret_access_key":"***************","bucket":"bucket_name","path":"path/to/file.csv","format":"csv","
compression":"none","header":true,"delimiter":"default","null_value":"default","newline":"CRLF","quote_policy":"
NONE","part_size":10}' \
-w -d testdb \
"SELECT 1 as col" -T presto
```