BigQuery Loader

Contents

- Basic Usage
- Specifying Which Columns are Content vs Metadata
- Adding Source to Metadata

Load a BigQuery query with one document per row.

```python
from langchain.document_loaders import BigQueryLoader
```

```python
BASE_QUERY = '''
SELECT
  id,
  dna_sequence,
  organism
FROM (
  SELECT
    ARRAY (
    SELECT
      AS STRUCT 1 AS id, "ATTCGA" AS dna_sequence, "Lokiarchaeum sp. (strain
GC14_75)." AS organism
    UNION ALL
    SELECT
      AS STRUCT 2 AS id, "AGGCGA" AS dna_sequence, "Heimdallarchaeota archaeon
(strain LC_2)." AS organism
    UNION ALL
    SELECT
      AS STRUCT 3 AS id, "TCCGGA" AS dna_sequence, "Acidianus hospitalis (strain
W1)." AS organism) AS new_array),
  UNNEST(new_array)
'''
```

Basic Usage

Skip to main content

```
data = loader.load()
```

```
print(data)
```

```
[Document(page_content='id: 1\ndna_sequence: ATTCGA\norganism: Lokiarchaeum sp.
(strain GC14_75).', lookup_str='', metadata={}, lookup_index=0),
Document(page_content='id: 2\ndna_sequence: AGGCGA\norganism: Heimdallarchaeota
archaeon (strain LC_2).', lookup_str='', metadata={}, lookup_index=0),
Document(page_content='id: 3\ndna_sequence: TCCGGA\norganism: Acidianus hospitalis
(strain W1).', lookup_str='', metadata={}, lookup_index=0)]
```

Specifying Which Columns are Content vs Metadata

```
loader = BigQueryLoader(BASE_QUERY, page_content_columns=["dna_sequence",
"organism"], metadata_columns=["id"])

data = loader.load()
```

```
print(data)
```

```
[Document(page_content='dna_sequence: ATTCGA\norganism: Lokiarchaeum sp. (strain
GC14_75).', lookup_str='', metadata={'id': 1}, lookup_index=0),
Document(page_content='dna_sequence: AGGCGA\norganism: Heimdallarchaeota archaeon
(strain LC_2).', lookup_str='', metadata={'id': 2}, lookup_index=0),
Document(page_content='dna_sequence: TCCGGA\norganism: Acidianus hospitalis
(strain W1).', lookup_str='', metadata={'id': 3}, lookup_index=0)]
```

Adding Source to Metadata

```
# Note that the `id` column is being returned twice, with one instance aliased as
`source`
ALIASED_QUERY = '''
SELECT
  id,
  dna_sequence,
```

Skip to main content

```
FROM (
  SELECT
    ARRAY (
    SELECT
      AS STRUCT 1 AS id, "ATTCGA" AS dna_sequence, "Lokiarchaeum sp. (strain
GC14_75)." AS organism
    UNION ALL
    SELECT
      AS STRUCT 2 AS id, "AGGCGA" AS dna_sequence, "Heimdallarchaeota archaeon
(strain LC_2)." AS organism
    UNION ALL
    SELECT
      AS STRUCT 3 AS id, "TCCGGA" AS dna_sequence, "Acidianus hospitalis (strain
W1)." AS organism) AS new_array),
  UNNEST(new_array)
'''
```

```python
loader = BigQueryLoader(ALIASED_QUERY, metadata_columns=["source"])

data = loader.load()
```

```python
print(data)
```

```
[Document(page_content='id: 1\ndna_sequence: ATTCGA\norganism: Lokiarchaeum sp.
(strain GC14_75).\nsource: 1', lookup_str='', metadata={'source': 1},
lookup_index=0), Document(page_content='id: 2\ndna_sequence: AGGCGA\norganism:
Heimdallarchaeota archaeon (strain LC_2).\nsource: 2', lookup_str='', metadata=
{'source': 2}, lookup_index=0), Document(page_content='id: 3\ndna_sequence:
TCCGGA\norganism: Acidianus hospitalis (strain W1).\nsource: 3', lookup_str='',
metadata={'source': 3}, lookup_index=0)]
```