Scale workloads with Snowflake and Dask

Your team uses data warehouses for SQL queries, but you may have noticed that for more advanced operations such as machine learning or exploratory analysis, a general-purpose language like Python is more efficient.

It's common to use a database to filter and join different datasets, and then pass that result off to Python for custom computations. As you work on larger datasets, and results grow beyond the scale of a single machine, passing the results between a database and general-purpose computation system becomes challenging and slow. If you're ready to grow beyond a single machine, then using the distributed powers of Dask with your Snowflake Data Cloud will speed up your work beyond the limitations of a single machine.

Try the Dask-Snowflake Connector

○ 3 Ways to Pass Data from Snowflake to Dask

Traditional Approaches Work but are Cumbersome and Slow the Flow

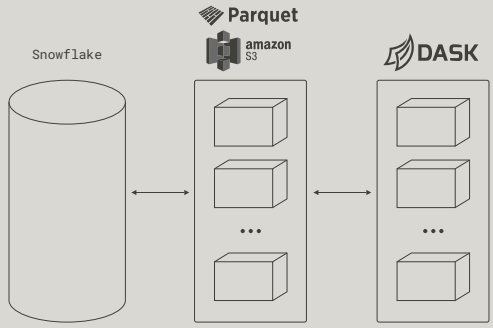
1. Break into many subqueries

For larger datasets, you can break one large table-scan query into many smaller queries and then run those in parallel to fill the partitions of a Dask dataframe.

```
import dask.dataframe as dd
df = dd.read_sql_table(
    'accounts',
    'snowflake://user:pass@...warehouse=...role=...',
    npartitions=10,
    index_col='id'
)
```

2. Bulk Export to Parquet

You can also perform a bulk export. Both Snowflake and Dask can read and write Parquet data on cloud object stores. You can perform a query with Snowflake, write the output to Parquet, and then read in that data with a Dask dataframe.



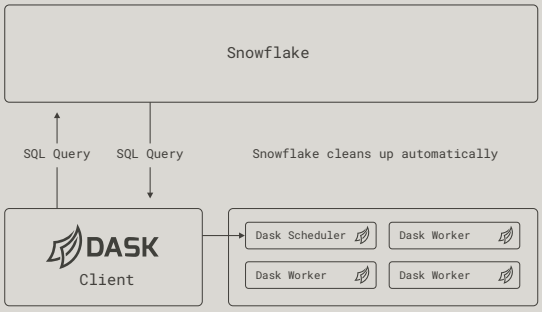

```
import dask.dataframe as dd
import snowflake

query = “””
COPY INTO ‘s3://my_storage_location’
        from <Table name> file_format = (type = parquet)
        credentials = (aws_key_id=’xxxx’ aws_secret_key=’xxxxx’ aws
token=‘xxxxxx’);
“””

con = snowflake.connector.connect(
   user=‘XXXX’,
   passwoard=‘XXXX’,
   account=‘XXXX’,
)
con.curson().execute(query)

df = dd.read_parquet(‘s3://my_storage_location’, ...)
```

3. Direct Parallel Reads and Writes

Stay in the (Python) flow and save time. Snowflake now supports a parallel fetch capability that can be used by external computation systems such as Dask.It combines the raw performance and support for complex queries used in bulk exports, with the central management of directly reading SQL queries from the database. Snowflake can now take the result of any query, and stage that result for external access.

Given this new capability, we developed a connector that can perform parallel read/writes from Dask to Snowflake, resulting in a smooth and simple approach that combines the best of all previous methods. It is easy to use, performs well, and maturely handles any query.

This is the recommended approach by Snowflake and Dask.




```
import dask_snowflake
import snowflake

with snowflake.connector.connect(...) as conn:
    ddf = dask_snowflake.from_snowflake(
       query=”””
       SELECT * FROM TableA JOIN TableB ON ...
       “””,
       conn=conn,
    )
```

See Dask and Snowflake Integration in action

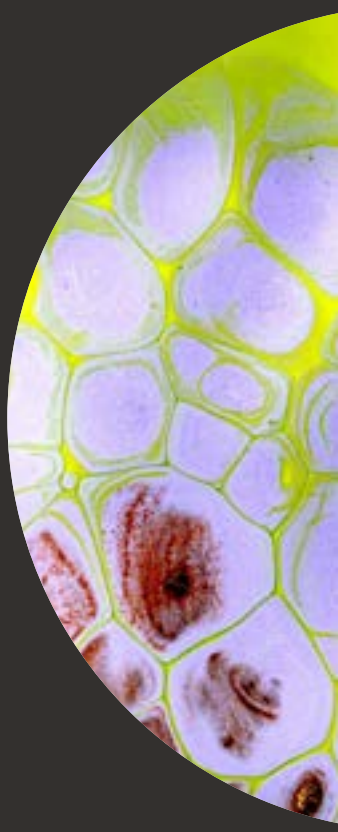
Try the Dask-Snowflake Connecter

About Coiled

Coiled scales Python to the cloud for data professionals. Based on Dask, the leading Python-native solution for distributed computing, Coiled has hosted more than 100M tasks for data professionals, scientists, and researchers around the globe including Capital One, Anthem Health, and the Air Force to solve challenges in business, research, and science. Coiled is a remote-first company with the best and brightest working from around the globe. Founded by the initial author of Dask, Coiled is on a mission to create a platform that gives Data Scientists the power of the cloud and machine learning, freeing them from today's limitations so they can solve important problems.

Try It Now

Follow Coiled



About Snowflake

Snowflake delivers the Data Cloud — a global network where thousands of organizations mobilize data with near-unlimited scale, concurrency, and performance. Inside the Data Cloud, organizations unite their siloed data, easily discover and securely share governed data, and execute diverse analytic workloads. Wherever data or users live, Snowflake delivers a single and seamless experience across multiple public clouds. Snowflake's platform is the engine that powers and provides access to the Data Cloud, creating a solution for data warehousing, data lakes, data engineering, data science, data application development, and data sharing. Join Snowflake customers, partners, and data providers already taking their businesses to new frontiers in the Data

Learn More

Follow Snowflake

Coiled | snowflake® | DASK