Danish Alsayed – 20637067 – HKUST – Master of Science in Big Data Technology

MSBD 5014 Independent Project Proposal

This project aims are implementing an algorithm described in the published research paper "Maintaining Acyclic Foreign-Key Joins under updates" on an SQL query involving at least one join.

The idea is to read data from a data source e.g. file, the query will be streamed via Apache Flink (a real time streaming engine) and the result will be processed via Flink workers in order to return the result. The result could be written in a file and should be the Delta (i.e. the difference in the results with the previous query) for every 100 updates.

The queries will be chosen from the TPC-H benchmark. The TPC Benchmark is a decision support benchmark. It consists of a collection of queries and concurrent data modifications. This benchmark illustrates decision support systems that examine large volumes of data, execute queries with a high degree of complexity, and give answers to critical business questions. The performance metric reported by TPC-H is called the TPC-H Composite Query-per-Hour Performance Metric. Implementing query 3 will suffice for the project's scope as it involves multiple joins. In order to achieve this, the process function should be implemented for each worker, which will involve low-level Flink programming. The aim will be to make the code as modular as possible so code duplication across the process function for each worker can be minimized. A slicing window should be used to stream the data in Flink and the KeyBy operation to distribute the data to the different workers.

The query

```
SELECT
    l_orderkey,
    sum(l_extendedprice * (1 - l_discount)) as revenue,
    o_orderdate,
    o_shippriority
FROM
    customer,
    orders,
    lineitem
WHERE
    c_mktsegment = 'BUILDING'
    AND c_custkey = o_custkey
    AND l_orderkey = o_orderkey
    AND o_orderdate < date '1995-03-15'
    AND l_shipdate > date '1995-03-15'
GROUP BY
    l_orderkey,
    o_orderdate,
    o_shippriority
ORDER BY
    revenue desc,
    o_orderdate
LIMIT 20;
```

Stretch goal

To take the project to the next level the source and sink of the data could be made using Apache Kafka (a real time messaging system) i.e. send the data from a file to a Kafka topic from which it will streamed in to Flink.

Technologies

Apache Flink, Apache Kafka, Java 8 or above, Windows OS