CS 744: SNOWFLAKE

Shivaram Venkataraman
Fall 2022
ADMINISTRIVIA

- Project Proposals – Due 11am tomorrow!
- Midterm on Thursday! Seating layout?
<table>
<thead>
<tr>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Learning</td>
</tr>
</tbody>
</table>

SparkSQL/Scope: Given a query how do you run it efficiently?

Snowflake: How do you build an elastic data warehouse?
CLOUD COMPUTING STACK

- Machine Learning
- Computational Engines
- Scalable Storage Systems
- SQL
SNOWFLAKE: GOALS

Software-as-a-Service

Elastic

Highly Available

Semi-Structured Data
SNOWFLAKE DESIGN

Cloud Services
- Infrastructure Manager
- Optimizer
- Transaction Manager
- Security

Authentication and Access Control

Metadata Storage

Virtual Warehouse
- Cache
- Cache
- Cache
- Cache

Data Storage
STORAGE VS COMPUTE

Shared Nothing

Multi Cluster, Shared Data
## Storage: Hybrid Columnar

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>32</td>
</tr>
<tr>
<td>Bob</td>
<td>22</td>
</tr>
<tr>
<td>Eve</td>
<td>24</td>
</tr>
<tr>
<td>Victor</td>
<td>27</td>
</tr>
</tbody>
</table>

- Alice, 32, Bob, 22
- Eve, 24, Victor, 27
- Alice, Bob, 32, 22
- Eve, Victor, 24, 27

**Row-oriented**

**Hybrid Columnar**
VIRTUAL WAREHOUSES

Elasticity, Isolation

Local caching, Stragglers
Concurrency Control

Pruning
FAULT TOLERANCE

Snowflake Web UI, BI Tools, ETL Tools, ODBC, JDBC, Python ...

Load Balancer

Cloud Services
Metadata Storage

Always On
On Demand
Infinite

Data Storage
Data Center  Data Center  Data Center
SEMI STRUCTURED DATA

```
{
  first_name: "john",
  last_name: "doe",
  order_id: "1234",
}

{
  first_name: "bucky",
  last_name: "badger",
  order_id: "52342",
  order_date: "3/3/2020",
}
```

Extraction, Flattening operations

Infer types, Pruning
TIME TRAVEL?

SELECT * FROM my_table AT(TIMESTAMP =>
    'Mon, 01 May 2015 16:20:00 -0700':::timestamp);
SELECT * FROM my_table AT(OFFSET => -60*5); -- 5 min ago
SELECT * FROM my_table BEFORE(STATEMENT =>
    '8e5d0ca9-005e-44e6-b858-a8f5b37c5726');

Multiple versions of table (MVCC)

Undo accidental deletes

Cheap to clone / snapshot a table
SUMMARY, TAKEAWAYS

Snowflake
- Cloud computing → Elastic data warehouse
- Key idea: Separation of compute and storage!
- Hybrid columnar storage format
- Elastic compute with virtual warehouses
- Pruning, semi-structured optimizations, fault tolerant
DISCUSSION

https://forms.gle/Trfe62jEp1ZUocJk6
We see how Snowflake leads to the design of an elastic data warehouse. If we were to similarly design an Elastic PyTorch for training how would the design look? What are some design trade-offs compared to existing PyTorch?
NEXT STEPS

Next class: Midterm!