ADMINISTRIVIA

- Assignment 1 grades out!
- Assignment 2 by mid-week
- Midterm on Thursday! Seating layout?
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
- SQL
- Computational Engines
- Scalable Storage Systems
SNOWFLAKE: GOALS

Software-as-a-Service

Elastic

Highly Available

Semi-Structured Data
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>

**Row-oriented**

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

**Hybrid Columnar**

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

Elasticity, Isolation

Local caching, Stragglers
FAULT TOLERANCE

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

Load Balancer

Cloud Services

Metadata Storage

Data Storage

Data Center

Data Center

Data Center

Always On

On Demand

Infinite
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 operation

Flattening

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
Hierarchical key management

Key rotation, re-keying
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/buUDM9nRs6Gg9tURA
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 class: Midterm!