What is a Cloudy Benchmark?
Measuring software that runs on the cloud
Use Cases for Hadoop@Facebook

- Analytics warehouse using Hive
  - Close to 100 PB in single HDFS cluster
  - 100+ Million files, 50K tables
  - 100K concurrent clients
- Backups and Archival
  - Backup thousands of MySQL db into HDFS
- Semi OLTP workload via HBase
  - 6 Billion messages/day
<table>
<thead>
<tr>
<th></th>
<th>My Asks for a Cloudy Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Elasticity</td>
</tr>
<tr>
<td>2</td>
<td>Fault Tolerance</td>
</tr>
<tr>
<td>3</td>
<td>Data Skew</td>
</tr>
</tbody>
</table>
Elasticity of Resources

- Why we need it?
  - Provision machines online for 24*7 operations

- How do we measure it?
  - Add machines at a defined pace while benchmarking is running
  - Decommission machines at a defined pace when benchmark is running
Fault Tolerance

Why do we need it?
- Use low-end commodity machines
- Faults are the norm rather than the exception
- Anomalous behavior rather than complete failure
  - 10% of machines are always 50% slower than the rest

How we do it?
- Kill 10% of machines over the course of a benchmark
- Abrupt kill of machines via a power reset
Data Skew

- Why do we need it?
  - Most big data systems take in un-curated data
  - Presence of outliers
- How do we do it?
  - Generate artificial data with specified distributions that generates data skew
Comments https://www.facebook.com/hadoopfs