链滴

数据仓库计费方式调研

作者：flowaters

原文链接：https://ld246.com/article/1527671836173

来源网站：链滴

<h2 id="背景">背景</h2>
数据仓库是企业用来永久保存数据，并且在上面进行数据分析和建模的服务。
目前的数据仓库除了私有化部署的 Hadoop 系列外，还有一些公有云 SAAS 版本。比如：

Google BigQuery, <a href "https://ld246.com/forward?goto=https%3A%2F%2Fcloud.google.com%2Fbigquery%2Fpricig%3Fhl%3Dzh-cn" target="_blank" rel="nofollow ugc">定价
Amazon Redshift, 定价: Amazon MPP 数据库
阿里云 MaxCompute, 定价
腾讯大数据处理套件 TBDS
UCLOUD 云数据仓库 UDW: 基于 Geenplum 和 PostgreSQL 的 MPP 数据库

本文调研前三个产品
<h2 id="定价调研">定价调研</h2>
<table>
<thead>
<tr>
<th align="center">产品</th>
<th align="center">语法</th>
<th align="center">接口</th>
<th align="center">BI 工具</th>
<th align="center">其它</th>
</tr>
</thead>
<tbody>
<tr>
<td align="center">Google BgQuery</td>
<td align="center">ANSI SQL:2011</td>
<td align="center">ODBC/JDBC</td>
<td align="center">Tableau、MicroStrategy、Looker、Google DataStudio 等</td>
<td align="center">可与 CloudML Engine 和 TensorFlow 进行集成</td>
</tr>
<tr>
<td align="center">Amazon Redshift</td>
<td align="center">标准 SQL</td>
<td align="center">ODBC/PostgreSQL JDBC</td>
<td align="center"></td>
<td align="center"></td>
</tr>
<tr>

<td align="center">阿里云 MaxCompute</td>
<td align="center">非标准 SQL, MapReduce, Graph</td>
<td align="center"></td>
<td align="center"></td>
<td align="center"></td>
</tr>
</tbody>
</table>
<h2 id="Google-BigQuery">Google BigQuery</h2>
每月可免费分析高达 1TB 的数据，并可免费存储 10GB 的数据。

存储: 每月每 G $0.02, 每月前 10 GB 免费
长期存储: 每月每 G $0.01, 每月前 10 GB 免费
流式插入: 每 200M $0.01
查询分析: 每 TB $5, 每月前 1 TB 免费
加载数据：免费
复制数据：免费
导出数据：免费
元数据操作：免费

本文不精确计算，仅定性的调研。
存储费用和流式插入费用，已经非常的明确了。查询分析部分的费用如下：
<h3 id="查询分析部分">查询分析部分</h3>
根据您选择的列中处理的总数据量向您收取费用，而每列的总数据量是基于该列中的数据类型计的。
官方的费用示例，整理如下：
<table>
<thead>
<tr>
<th>示例查询</th>
<th>处理的字节数</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>SELECT corpus,word FROM publicdata:samples.shakespeare LIMIT 1;</code></td>
<td><code>corpus</code> 列 + <code>word</code> 列的总大小</td>
</tr>
<tr>
<td><code>SELECT corpus FROM (SELECT * FROM publicdata:samples.shakespeare);</code></td>
<td><code>corpus</code> 列的总大小</td>
</tr>
<tr>
<td><code>SELECT COUNT(*) FROM publicdata:samples.shakespeare;</code></td>
<td>没有处理任何字节</td>
</tr>

<tr>
<td><code>SELECT COUNT(corpus) FROM publicdata:samples.shakespeare;</code></td>
<td><code>corpus</code> 列的总大小</td>
</tr>
<tr>
<td><code>SELECT COUNT(*) FROM publicdata:samples.shakespeare WHERE corpus = 'hamlet';</code></td>
<td><code>corpus</code> 列的总大小</td>
</tr>
<tr>
<td><code>SELECT shakes.corpus,wiki.language FROM publicdata:samples.shakespeare AS shakes JOIN EACH publicdata:samples.wikipedia AS wiki ON shakes.corpus = wiki.title;</code></td>
<td><code>shakes.corpus</code>、<code>wiki.language</code> 和 <code>wiki.title</code> 列的总大小</td>
</tr>
</tbody>
</table>
即写 SQL 的时候，写法和费用是相关的。
原理估计是 SQL 语法树解析，根据字段列和表，来查询元数据信息，计算费用。
<h2 id="Amazon-Redshift">Amazon Redshift</h2>
定价有三个选项:

按需定价：没有预付成本，您只需基于群集中的节点类型和数量支付小时费用。

Amazon Redshift Spectrum 定价：让您能够对 Amazon S3 中的 EB 级数据运行 SQL 查询，只需为扫描的字节数付费。
预留实例定价：通过承诺使用 Redshift 1 年或 3 年，相比按需费率最多可节省 75% 的费用。

<h3 id="按需定价">按需定价</h3>
内容整理自 Amazon Redshift 定价页面，如下：
<table>
<thead>
<tr>
<th></th>
<th>vCPU</th>
<th>ECU</th>
<th>内存</th>
<th>存储</th>
<th>I/O</th>
<th>价格</th>
</tr>
</thead>

<tbody>
<tr>
<td>密集计算</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dc2.large</td>
<td>2</td>
<td>7</td>
<td>15 GiB</td>
<td>0.16TB SSD</td>
<td>0.60 GB/s</td>
<td>每小时 0.25 USD</td>
</tr>
<tr>
<td>dc2.8xlarge</td>
<td>32</td>
<td>99</td>
<td>244 GiB</td>
<td>2.56TB SSD</td>
<td>7.50 GB/s</td>
<td>每小时 4.80 USD</td>
</tr>
<tr>
<td>密集存储</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ds2.xlarge</td>
<td>4</td>
<td>14</td>
<td>31 GiB</td>
<td>2TB HDD</td>
<td>0.40 GB/s</td>
<td>每小时 0.85 USD</td>
</tr>
<tr>
<td>ds2.8xlarge</td>
<td>36</td>
<td>116</td>
<td>244 GiB</td>
<td>16TB HDD</td>
<td>3.30 GB/s</td>
<td>每小时 6.80 USD</td>

</tr>
</tbody>
</table>
看起来存储和计算是捆绑起来计费的，不能单独计费。
<h3 id="Redshift-Spectrum定价">Redshift Spectrum 定价</h3>
Spectrum 允许直接对 S3 上的数据进行 SQL 查询。在阿里云中，Amazon S3 对应产品是 OSS，Spectrum 对的功能是访问 OSS 非结构化数据。
Redshift Spectrum 中，查询按扫描的数据量计费，对表的管理操作(即 DDL)不计费。
这里扫描的数据量是按物理存储的数据量。如果使用了列数据，或者压缩存储，那么扫描的数据就会变少，费用会大大降低。
BigTable 中也是对列计费的，但是没有压缩选项。不过 BigTable 的存储单价近似为 Spectrum 的 1/3，可以理解为默认做了压缩。总体算下来存储成本是相当的。
这里 Redshift Spectrum 等列存储格式的选择更灵活，可以选择 Parquet 或者 ORC 格式。</p

具体的计费为：以 10M 为单位，扫描的每 TB 数据 支付 5 USD。
例如，如果扫描 10GB 的数据，则需支付 0.05 USD。如果扫描 1TB 的数据，则需支付 5 USD。/p>
<h3 id="Redshift-Spectrum-定价示例">Redshift Spectrum 定价示例</h3>
假设一个表中有 100 个大小相同的列，以未压缩文本文件的格式存储在 Amazon S3 中，总大为 4TB。如果运行查询以从该表的一个列中获取数据，则 Redshift Spectrum 需要扫描整个文件，为文本格式无法拆分。该查询将扫描 4TB 数据，费用为 20 USD。(5 USD/TB * 4TB = 20 USD)
如果使用 GZIP 将文件压缩，那么压缩比可能为 4:1。这样，您可以获得一个大小为 1TB 的压缩件。Redshift Spectrum 必须扫描整个文件，但由于它的大小是原来的四分之一，所以您只需支付四之一的费用 (即 5 USD)。(5 USD/TB * 1TB = 5 USD)
如果您压缩文件并将其转换为列式格式 (如 Apache Parquet)，那么压缩比可能为 4:1，您可以得一个大小为 1TB 的文件。使用上述查询，Redshift Spectrum 只需扫描 Parquet 文件的其中一列该查询的费用为 0.05 USD。(5 USD/TB * 1TB 文件大小 * 1/100 列，即 10GB 的总扫描量 = 0.05 USD)
<h2 id="阿里云-MaxCompute">阿云 MaxCompute</h2>
费用分为 4 部分: 存储计费, 计算计费, 下载计费和数据导入计算。

存储计费: 按容量阶梯计费
计算计费: 分按量后付费和按 CU 预付费两种计算计费方式
下载计费: 按照下载的数据大小进行计费
数据导入：数据导入 MaxCompute 不计费

其中计算计费分两种方式

按量后付费: SQL 任务，按 I/O 后付费；MapReduce 任务，按量付费。
按 CU 预付费: 仅在阿里云大数据平台提供

<h3 id="存储计费">存储计费</h3>
项目的数据实际存储量大于 0 小于等于 512MB 时
MaxCompute 将收取这个项目 0.01 元的费用。示例如下：

如果您在 MaxCompute 上，某个项目的存储的数据为 100MB，MaxCompute 会直接收取您 0.

1 元/天的费用。

如果您有多个项目，且每个项目实际存储量小于 512MB，MaxCompute 会对每个项目收取 0.01 元。

项目的数据实际存储量大于等于 512MB 时
<table>
<thead>
<tr>
<th>基础价格</th>
<th>大于 100GB 部分</th>
<th>大于 1TB 部分</th>
<th>大于 10TB 部分</th>
<th>大于 100TB 部分</th>
<th>1PB 以上部分</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0192 元/GB/天</td>
<td>0.0096 元/GB/天</td>
<td>0.0084 元/GB/天</td>
<td>0.0072 元/GB/天</td>
<td>0.006 元/GB/天</td>
<td>请通过工单联系我们</td>
</tr>
</tbody>
</table>
<h3 id="计算计费">计算计费</h3>
<h4 id="SQL任务按量后付费">SQL 任务按量后付费</h4>
SQL 任务按量后付费即按 I/O 后付费：您每执行一条 SQL 作业，MaxCompute 将根据该作业的输入数据及该 SQL 的复杂度进行计。该费用在 SQL 执行完成后产生，并在下一天做一次性的计费结算。
MaxCompute SQL 任务的按 I/O 后付费会针对每个作业产生一次计量。天的所有计量信息将在第二天做一次性汇总收费。
SQL 计算任务的计费公式为：

<code>一次 SQL 计算费用 = 计算输入数据量 * SQL 复杂度 * SQL 价格</code>

价格如下：
<table>
<thead>
<tr>
<th>计费项</th>
<th>价格</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQL 价格</td>
<td>0.3 元/GB</td>
</tr>

</tbody>
</table>

计算输入数据量：指一条 SQL 语句实际扫描的数据量，大部分的 SQL 语句有分区过滤和列裁剪所以一般情况下这个值会远小于源表数据大小：

列裁剪：例如您提交的 SQL 是 <code>select f1,f2,f3 from t1;</code> 只计算 t1 表中 f1，f，f3 三列的数据量，其他列不会参与计费。

分区过滤：例如 SQL 语句中含有 where ds >“20130101”，ds 是分区列，则计费的数据只会包括实际读取的分区，不会包括其他分区的数据。

SQL 复杂度：先统计 SQL 语句中的关键字，再折算为 SQL 复杂度，具体如下：

SQL 关键字个数 = Join 个数 + Group By 个数 + Order By 个数 + Distinct 个数 + 窗口函数数 + max（insert into 个数-1, 1）。

SQL 复杂度计算：

SQL 关键字个数小于等于 3，复杂度为 1。

SQL 关键字个数小于等于 6，且大于等于 4，复杂度为 1.5。

SQL 关键字个数小于等于 19，且大于等于 7，复杂度为 2。

SQL 关键字个数大于等于 20，复杂度为 4。

复杂度计量命令格式：

<code>cost sql <SQL Sentence>;</code>

示例如下：

<code>odps@ $odps_project >cost sql SELECT DISTINCT total1 FROM</code>

<code>(SELECT id1, COUNT(f1) AS total1 FROM in1 GROUP BY id1) tmp1</code>

<code>ORDER BY total1 DESC LIMIT 100;</code>

<code>Complexity:1.5</code>

示例中 SQL 关键字的个数是 4（该语句中有一个 DISTINCT，一个 COUNT，一个 GROUP BY 一个 ORDER），而 SQL 复杂度是 1.5。如果表 in1 的数据量为 1.7GB（对应账单为 1.7GB×10243=825361100.8Byte），则实际消费为：<code>1.7*1.5*0.3=0.76 元</code>。
<h2 id="定价总结">定价总结</h2>
<table>
<thead>
<tr>
<th align="center">方式</th>
<th align="center">BigQuery</th>
<th align="center">RedShift Spectrum</th>
<th align="center">MaxCompute</th>
</tr>
</thead>
<tbody>
<tr>
<td align="center">存储</td>
<td align="center">每月每 G: 0.01$</td>
<td align="center">每月每 GB: 0.039$</td>
<td align="center">每天每 G: 0.0192 元</td>
</tr>
<tr>
<td align="center">实时流入</td>
<td align="center">每 200M $0.01</td>
<td align="center"></td>
<td align="center"></td>
</tr>
<tr>
<td align="center">SQL 查询</td>
<td align="center">每 TB $5 </td>
<td align="center">每 TB 5$</td>
<td align="center">计算输入数据量 * SQL 复杂度 * SQL 价格(0.3 元/GB)</td>
</tr>
</tbody>
</table>

存储成本：三者的成本是差不多的，具体和数据量和压缩方式相关。
SQL 查询成本：前两者是一刀切；MaxCompute 计算粒度更细。具体也和查询条件相关。

<h2 id="参考">参考</h2>

Google BigQuery
<a href="https://ld246.com/forward?goto=https%3A%2F%2Faws.amazon.com%2Fcn%2Fredshift%2Fpricing%2F%3Fp%3Dps" target="_blank" rel="nofollow ugc">Amazon Redshift</

><
阿里云 MaxCompute
使用 edshift Spectrum 查询 S3 数据
笔记: GCE BigQuery vs AWS Redshift vs AW Athena: 没有测试到 Spectrum + Parquet 的情形
