Q1 客户端Web访问日志分析报告

Q1.1 数据准备与预处理

数据路径

- HTTP 日志路径：`maccdc-2012/{00,01,02,03,04,05}/http.log.gz`
- DNS 日志路径：`maccdc-2012/{00,01,02,03,04,05}/dns.log.gz`

读取与处理

使用 PySpark 读取 JSON 格式的日志文件，并对时间戳字段进行格式转换：

```
df_http = spark.read.json(http_path)
df_http = df_http.withColumn("ts", F.from_unixtime("ts").cast("timestamp"))

df_dns = spark.read.json(dns_path)
df_dns = df_dns.withColumn("ts", F.from_unixtime("ts").cast("timestamp"))
```

注册为临时视图以支持 SQL 查询：

```
df_http.createOrReplaceTempView("http_log")
df_dns.createOrReplaceTempView("dns_log")
```

Q1.2 最频繁访问的 URI 排行

分析目标

统计 HTTP 状态码为 `200` 且请求方法为 `GET` 的访问中，最常访问的 URI。

查询逻辑

- 筛选 `status_code == 200` 且 `method == "GET"` 的记录；
- 以 URI 为维度进行分组并计数；
- 按访问次数降序排列。

Top 20 URI 访问统计

URI	访问次数
/	9475
/admin/config.php?pass=admin	556
/main.php?logout=1	194
/top.php?stuff=15...	191
/top.php	179
/main.php?stuff=1...	172
/get_latest_id.php	159
/admin/config.php...	138
/cacti/index.php	129

URI	访问次数
/en-US/api/message...	118

Q1.3 访问 URI 的 TCP 协议使用比例

分析目标

在任务二的基础上，进一步分析每个 URI 所关联的 DNS 请求中，使用 TCP 协议的比例。

查询逻辑

- 左连接 HTTP 和 DNS 日志（连接键：`uid`）；
- 筛选 GET 请求和状态码为 200 的记录；
- 计算每个 URI 的 HTTP 访问次数；
- 计算这些访问中 DNS 请求使用 `proto = tcp` 的比例。

Top 20 URI 的 TCP 协议比例统计

URI	HTTP访问次数	TCP使用比例（%）
/	9475	0.0
/admin/config.php?pass=admin	556	0.0
/main.php?logout=1	194	0.0
/top.php?stuff=15...	191	0.0
/top.php	179	0.0
/main.php?stuff=1...	172	0.0
/get_latest_id.php	159	0.0
/admin/config.php...	138	0.0
/cacti/index.php	129	0.0

Q2.1 - 2.2客户流失预测：生存分析报告

1. 数据准备与预处理

数据来源：客户流失数据集（共3,351行 × 20列）

关键步骤：

- 筛选出**按月付费且拥有DSL或光纤网络服务的客户。**
- 创建二元标签变量 `churn`（流失=1，留存=0）。
- 将分类变量（如 `dependents`，`techsupport`）转换为哑变量。
- 将所有列名统一为小写格式。

数据预处理快照示例：

customerid	gender	seniorcitizen	...	techsupport_yes	churn
7590-VHVEG	Female	0	...	0	1
3668-QPYBK	Male	0	...	1	0

---

2. 探索性生存分析

分层分析：

- **按性别**：无显著差异（log-rank 检验 p = 0.15）
- **按网络类型（DSL vs Fiber）**：
  - DSL用户在早期具有更高生存概率：

月份	DSL生存概率	Fiber生存概率
1	0.90	0.85
6	0.78	0.65

---

3. Cox比例风险模型

模型摘要：

- **一致性指数（Concordance）**：0.64

显著变量：

特征	风险比 (Hazard Ratio)	p值
onlineBackup_Yes	0.46（风险降低54%）	<0.001
techSupport_Yes	0.53（风险降低47%）	<0.001
internet_service_DSL	0.80（风险降低20%）	<0.001

比例风险假设检验：

以下变量不满足比例风险假设（p < 0.001），说明其影响随时间变化：

- internet_service_DSL
- onlineBackup_Yes
- techSupport_Yes

建议：对上述变量进行**分层建模**（Stratification）。

---

4. 参数生存模型（对数正态AFT模型）

核心结果：

- **中位生存时间预测**：135.51个月

特征的生存时间加速效应：

特征	加速系数 (exp(coef))	解读
onlineBackup_Yes	2.25	生存时间延长125%
techSupport_Yes	1.99	生存时间延长99%
internet_service_DSL	1.47	生存时间延长47%

---

5. 财务影响分析

关键指标（假设：每月利润 = $30，折现率 IRR = 10%）

合同月份	生存概率	累积净现值 (NPV)
12	59%	$251.40
24	43%	$405.44
36	[预测]	$615.20

6.

留存关键因素：

- 使用**在线备份服务**的用户流失风险**降低54%**
- 提供**技术支持**的用户中位生存时间提升近一倍

服务优化建议：

- 对DSL用户重点推广**安全/备份打包服务**
- 强化客户初期（前12个月）支持体验以延长生命周期

客户生命周期价值（CLV）提升：

- **12–24个月**为投资回报率（ROI）峰值区间，应重点投入留存资源

总结

生存分析有效量化了客户留存驱动因素，并预测客户生命周期价值（CLV）。
推荐使用**分层Cox模型**或**对数正态AFT模型**以提升预测准确性，从而助力更精准的客户留存策略制定。

Q2.3

ChatGPT

```
{
  "sql": "SELECT COUNT(*) AS continent_count FROM continents;"
}
```

Deepseek

```
{"sql": "SELECT COUNT(*) AS count FROM continents;"}
```

Q2.4