**云上屈原建设文档**

**一、架构图**

**数据输入**

Kafka : 日志行为数据

Mysql : 用户数据

**处理引擎**

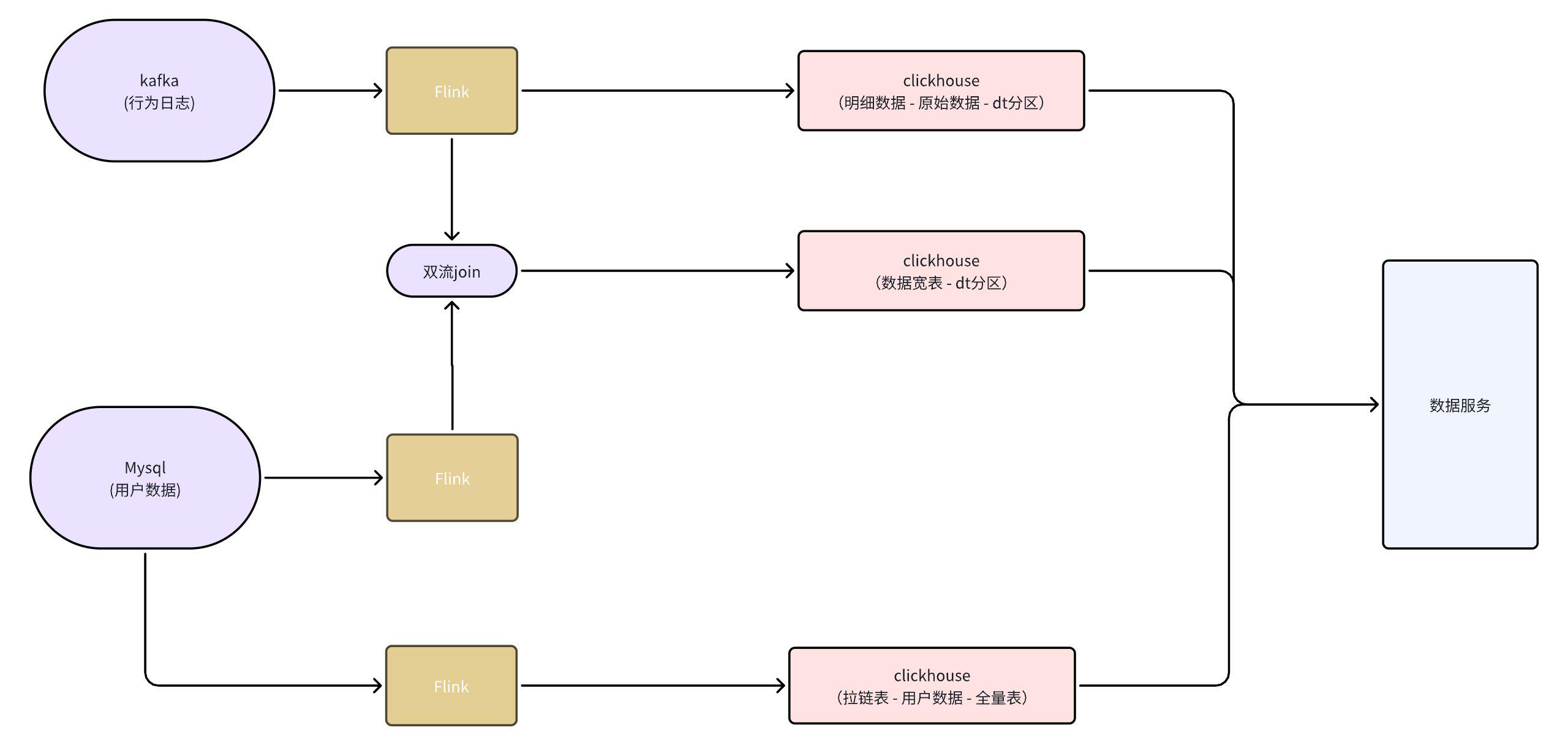
Flink : 支持流批计算

**存储数据库**

Clickhouse ： 离线实时，多种库表引擎，擅长指标统计

**数据输出**

Springboot ： 微服务



**二、模块**

1. **游客数量**

当前在线人数 ： 当前在线的用户+访客数量，实时

今日浏览人数 ： 当天访问次数pv，每小时更新一次

本周浏览人数 ： 本周访问次数pv，每小时更新一次

本月浏览人数 ： 当月访问次数pv，每小时更新一次

自定义周期pv ： dt分区，每天数据统计 （最少以天为单位）

2.  **用户画像**

年龄 ：各年龄段比例

性别 ：男女性别比例

地区 ：全国省市用户数量（同一个注册最近10次访问次数最多的IP的省份）

职业 ：各职业占比，普通用户、教师、学生、学者、其他

3. **核心指标**

访问人数趋势 ：自定义统计周期的访问次数PV，最少以天为单位

注册用户趋势 ：自定义统计周期的注册人数，最少以天为单位

4. **来源趋势**

搜素引擎 ：自定义统计周期的通过搜索引擎跳转到云上屈原的次数，最少以天为单位

直接访问 ：自定义统计周期的直接通过域名访问到云上屈原的次数，最少以天为单位

外部链接 ：自定义统计周期的通过其他外部链接访问到云上屈原的次数，最少以天为单位

自定义来源 ：自定义统计周期的不属于上面3种访问到云上屈原的次数，最少以天为单位

5. **来源占比**

来源占比 ： 自定义统计周期的搜素引擎、外部链接、直接访问、自定义来源4种方式的占比

6. **排行榜**

外景排名 ：自定义统计周期，所有外景访问次数top的区域名称和次数，最少以天为单位

内景排名 ：自定义统计周期，所有内景访问次数top的区域名称和次数，最少以天为单位

数字人排名 ：自定义统计周期，所有数字人访问次数top的区域名称和次数，最少以天为单位

展品排名 ：自定义统计周期，所有展品访问次数top的区域名称和次数，最少以天为单位

**三、表**

1. **用户表**

表名 ： MYSQL : test.cloud\_quyuan\_user\_test

更新周期 ： 实时

处理逻辑 ： 修改、删除、增加

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 字段说明 | 数据类型 | 非空 | 样例数据 |
| user\_id | 用户ID (主键) | INT | Y | 1 |
| user\_name | 用户名 | VARCHAR | Y | 张飞 |
| gender | 性别 | INT | Y | 男 |
| age | 年龄 | INT | N | 48 |
| address | 地区 | TEXT | N | 四川 |
| occupation | 职业 | VARCHAR | N | 学生 |
| create\_time | 创建时间 | DATE | Y | 2023-10-13 09:54:02 |
| upadte\_time | 更新时间 | DATE | Y | 2023-10-13 09:54:02 |

2. **行为日志表 (ods)**

表名 ： CLICKHOUSE : quyuan.quyuan\_ods\_src\_behavior\_log

数据源 ： KAFKA ： *cloud\_quyuan\_test*

分区 ： dt

排序键 ： 模块、名称

表引擎 ： Distributed表引擎 + ReplicatedMergetree表引擎 (2副本)

更新周期 ： 实时

处理逻辑 ： 数据原貌入库

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 字段说明 | 数据类型 | 非空 | 样例数据 |
| location | 位置（前台/后台） | String | Y | 前台 |
| model | 模块 | String | Y | 登录页面 |
| name | 名称 | String | Y | 用户登录成功 |
| describe | 描述 | String | N | 点击登录按钮，通过校验，登录成功 |
| action | 动作类型 （点击/停留） | String | N | 点击 |
| content | 数据内容 | String | N | {  "info":"登录成功",  "ID":12,  "time":"202310130954" }  json形式 ， 不同模块名称对应不同内容 [云上屈原管理后台字段表](https://bf6o2s62ws.feishu.cn/sheets/NDiEsWimrhWb9Zt5Sjtc2FmYnUe?sheet=N6WOPk&range=QTpH) |
| timestamp | 时间戳 | DateTime | Y | 2023-10-13 09:54:02 |
| dt | 日期 （分区） | Date | Y | 2023-10-13 |

3. **行为日志表宽表 (dws)**

表名 ： CLICKHOUSE : quyuan.quyuan\_dws\_behavior\_log

数据源 ： KAFKA ： *cloud\_quyuan\_test*

MYSQL ：test.cloud\_quyuan\_user\_test

分区 ： dt

排序键 ： 模块、名称

表引擎 ： Distributed表引擎 + ReplicatedMergetree表引擎 (2副本)

更新周期 ： 实时

处理逻辑 ： 数据清洗、维表关联

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 字段说明 | 数据类型 | 非空 | 样例数据 |
| location | 位置（前台/后台） | String | Y | 前台 |
| model | 模块 | String | Y | 登录页面 |
| name | 名称 | String | Y | 用户登录成功 |
| describe | 描述 | String | N | 点击登录按钮，通过校验，登录成功 |
| action | 动作类型 （点击/停留） | String | N | 点击 |
| content | 数据内容 | String | N | {  "info":"登录成功",  "ID":12,  "time":"202310130954" }  json形式 ， 不同模块名称对应不同内容 [云上屈原管理后台字段表](https://bf6o2s62ws.feishu.cn/sheets/NDiEsWimrhWb9Zt5Sjtc2FmYnUe?sheet=N6WOPk&range=QTpH) |
| user\_name | 用户名 | VARCHAR | Y | 张飞 |
| gender | 性别 | INT | Y | 男 |
| age | 年龄 | INT | N | 48 |
| address | 地区 | TEXT | N | 四川 |
| occupation | 职业 | VARCHAR | N | 学生 |
| timestamp | 时间戳 | DateTime | Y | 2023-10-13 09:54:02 |
| dt | 日期 （分区） | Date | Y | 2023-10-13 |

4. **用户信息表 (ods)**

表名 ：CLICKHOUSE : quyuan.quyuan\_ods\_src\_user\_info

源表 ： MYSQL : test.cloud\_quyuan\_user\_test

分区 ： 不分区

排序键 ： 用户ID

表引擎 ：Distributed表引擎 + ReplicatedVersionedCollapsingMergetree表引擎 (2副本)

更新周期 ： 实时

处理逻辑 ： 修改、增加

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 字段说明 | 数据类型 | 非空 | 样例数据 |
| user\_id | 用户ID (主键) | INT | Y | 1 |
| user\_name | 用户名 | VARCHAR | Y | 张飞 |
| gender | 性别 | INT | Y | 男 |
| age | 年龄 | INT | N | 48 |
| address | 地区 | TEXT | N | 四川 |
| occupation | 职业 | VARCHAR | N | 学生 |
| create\_time | 创建时间 | DATE | Y | 2023-10-13 09:54:02 |
| end\_time | 有效时间 | DATE | Y | 2099-12-31 59:59:59 |

**四、API接口**

http://192.168.99.75:51101/doc.html