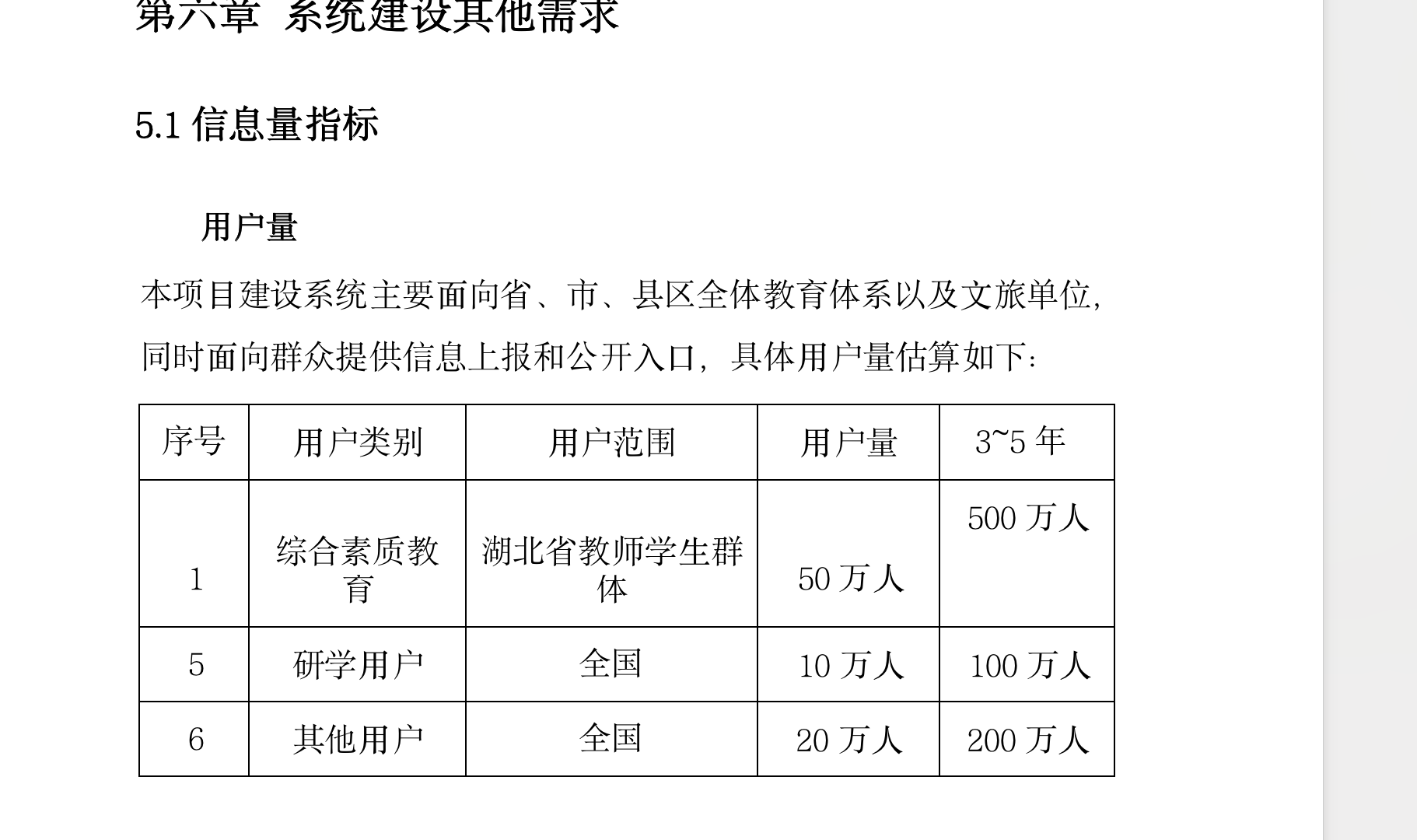
**云上屈原技术方案**

**大数据驾驶舱**

**一. 架构设计**

1. **数据规模&时效**

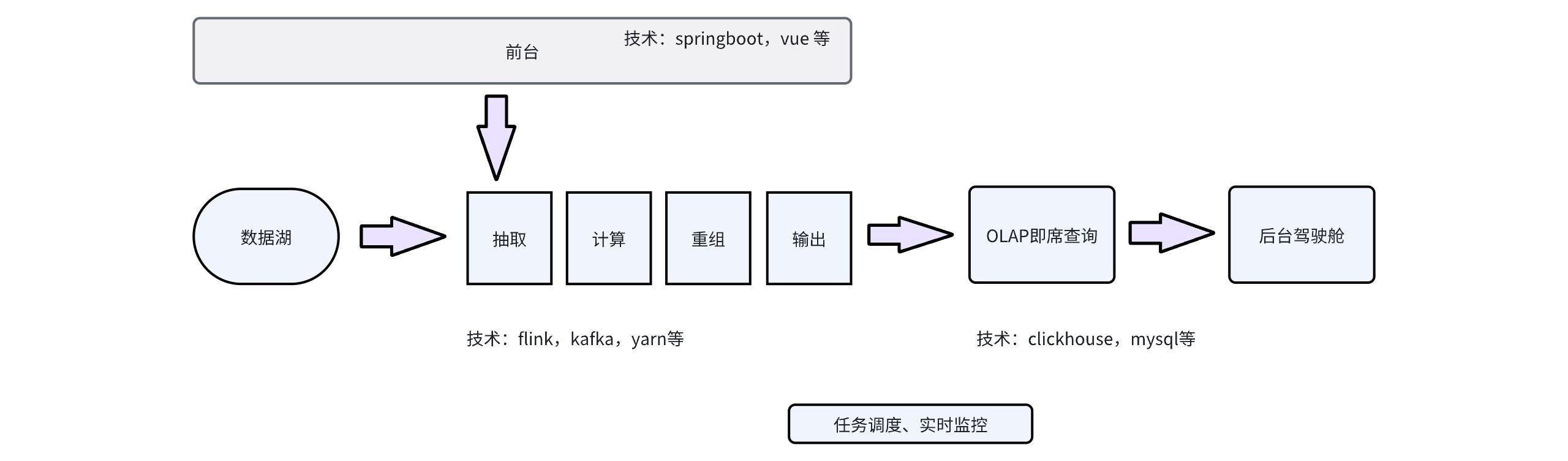


1. 3-5年：800万用户。
2. 数据量预估：
3. 每日1/10的在线用户估算，80万用户在线。日活均值：40万用户在线。
4. 每人每天平均100条行为日志/数据，总量4000万。
5. 每条日志1K大小，数据量大小为40G。
6. 数据量大小：
7. 一年：40\*360 = 14T；三年：40\*365\*3 = 43800G = 42T
8. 三年数据量：2个副本&1/10的压缩 = 8.4T
9. 更新周期：最短1h，所有需求均需要当天内指标统计。

2. **框架版本**

1. Apache/CDH/HDP
2. **Apache：免费，但运维麻烦。需运维自行调研组件，配置组件兼容性。**
3. **CDH：收费。一个节点一万美金/年。**
4. HDP：开源，可二次开发。没有CDH稳定，国内使用少。
5. 云服务：
6. **阿里云EMR**、MaxCompute、DataWorks；亚马逊云EMR；腾讯云EMR；华为云EMR
7. 运维简单，收费，国内用户逐年增长。

3. **技术架构**



4. **技术选型**

**计算引擎-Flink**

需求：T+0当天内，实时。

Flink：流式，状态后端，准确性和时效性都高。

SparkStreaming：微批次，时效性、计算速率不如flink。

Storm：准确性欠缺。

**数据存储-ClickHouse**

需求：实时聚合查询

Clickhouse：DWS层到ads层需重新分组再聚合，clickhouse是列式存储做聚合操作有优势，且非常擅长单宽表的聚合分析。

Hbase：k-v存储，更适合单条数据的k-v查询。

**消息队列-Kafka**

Kafka：解耦、削峰、适用于大数据量。

MQ：吞吐量不如Kafka，不适合大数据场景。

**数据同步-flinkCDC**

选型原因：

1. 基于Debezium实现，实时同步，且低延迟。且改进了debezium单机的缺陷，可对接分布式系统。
2. 使用flinkCDC，可直接把维度配置信息同步至flink，不需要再经过kafka。

|  |  |  |
| --- | --- | --- |
|  | 基于查询的CDC | 基于Binlog的CDC |
| 开源产品 | Sqoop、Kafka JDBC Source | Canal、Maxwell、Debezium |
| 执行模式 | Batch | Streaming |
| 是否可以捕获所有数据变化 | 否 | 是 |
| 延迟性 | 高延迟 | 低延迟 |
| 是否增加数据库压力 | 是 | 否 |

数据同步策略：因用户表数据量不小，采用增量同步。

mysql：开启binlog。

**应用数据-SpringBoot接口发布&MySQL**

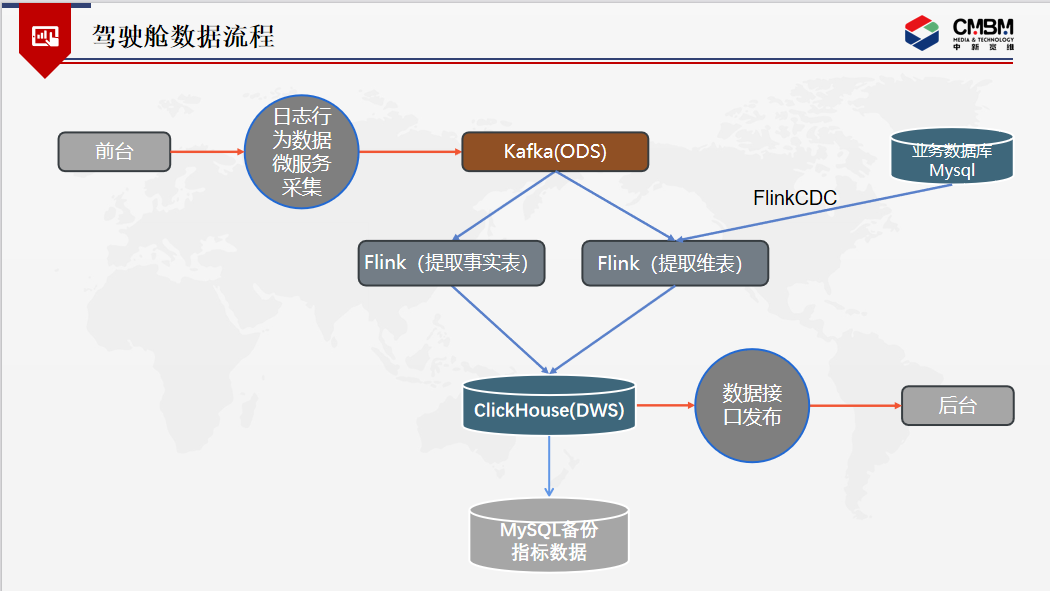
SpringBoot:使用SpringBoot微服务查询OLAP库进行聚合分析，计算结果发布成数据接口至驾驶舱。

MySQL：指标数据备份。

**日志数据采集-SpringBoot**

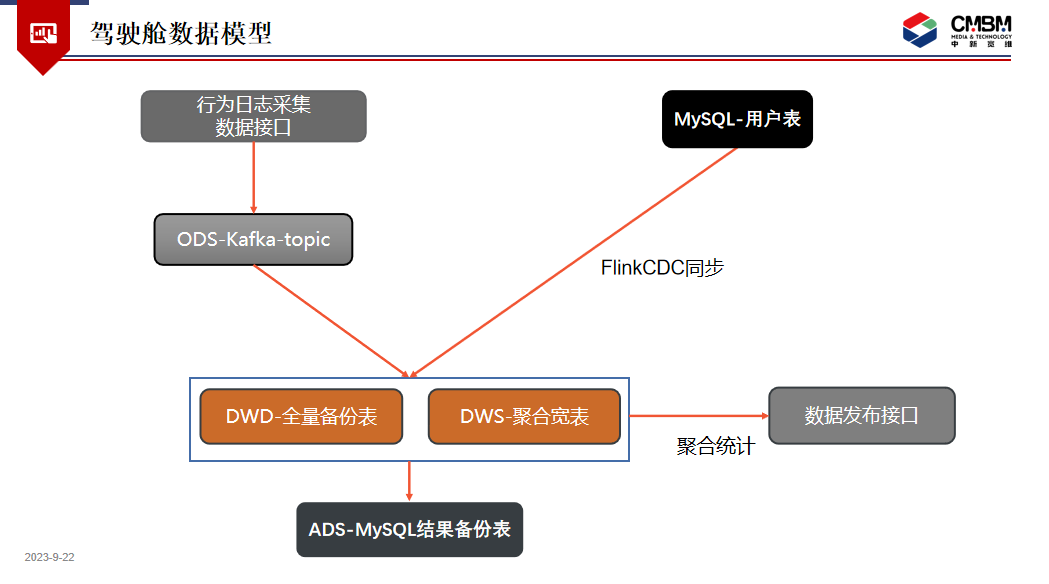
前后端通用、成熟交互方式。

5. **数据流程**



**二. 数据体系**

**分层模型**



**维度模型**

1. 事实：对应业务过程（一个个不可拆分的行为事件）
2. 行为日志表，分主题。
3. 维度：对应业务过程发生时所处的环境。
4. 用户
5. 地区
6. 日期

**DWS**

ClickHouse：日志宽表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**ODS**

**Kafka：行为日志topic**

|  |
| --- |
| JSON 数据字段：浏览行为，行为时间， 样例数据： {} |

**MySQL：用户业务表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 字段说明 | 数据类型 | 缺省值 | 样例数据 | 提取规则 |
|  | 用户ID |  |  |  |  |
|  | 用户名 |  |  |  |  |
|  | 性别 |  |  |  |  |
|  | 年龄 |  |  |  |  |
|  | 地区 |  |  |  |  |
|  | 注册时间 |  |  |  |  |
|  |  |  |  |  |  |

**数据采集API接口**

|  |  |  |  |
| --- | --- | --- | --- |
| 接口名称 | 接口路径 | 请求方式 | 样例数据 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**ADS**

**MySQL：指标结果表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**发布API接口**

|  |  |  |  |
| --- | --- | --- | --- |
| 接口名称 | 接口路径 | 请求方式 | 样例数据 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |