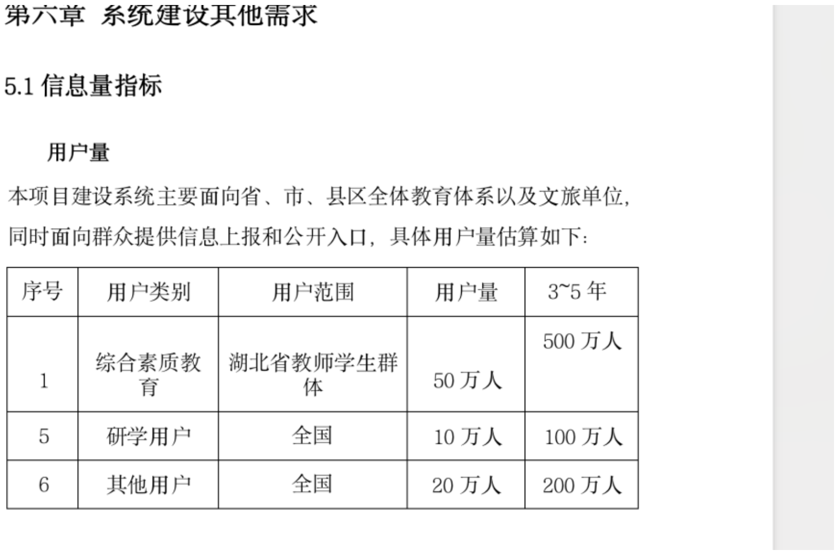
Qu Yuan technical solution on the cloud

Big Data cockpit

One. Architecture design

1. Data scale & timeliness



第六章 系统建设其他需求

5.1 信息量指标

用户量

本项目建设系统主要面向省、市、县区全体教育体系以及文旅单位，同时面向群众提供信息上报和公开入口，具体用户量估算如下：

序号	用户类别	用户范围	用户量	3~5年
1	综合素质教育	湖北省教师学生群体	50万人	500万人
5	研学用户	全国	10万人	100万人
6	其他用户	全国	20万人	200万人

1. 3-5 years: 8 million users.
2. Estimated data volume:
   a. 1 in 10 daily online users estimated, 800,000 users online. Daily active average: 400,000 users online.
   b. Average 100 behavior logs/data per person per day, total 40 million.
   c. Each log is 1K in size, and the data volume is 40GB.
   d. Data size:
      i. One year: 40*360 = 14T; Three years: 40*365*3 = 43800G = 42T
      ii. Three years of data: 2 copies &1/10 compression = 8.4T
3. Update period: 1h minimum, all requirements require intra-day indicator statistics.

2. Framework version

1. Apache/CDH/HDP

    a. **Apache: Free, but O&M hassle. O&m needs to research components and configure component compatibility.**

    b. **CDH: Charge. $10,000 per node per year.**

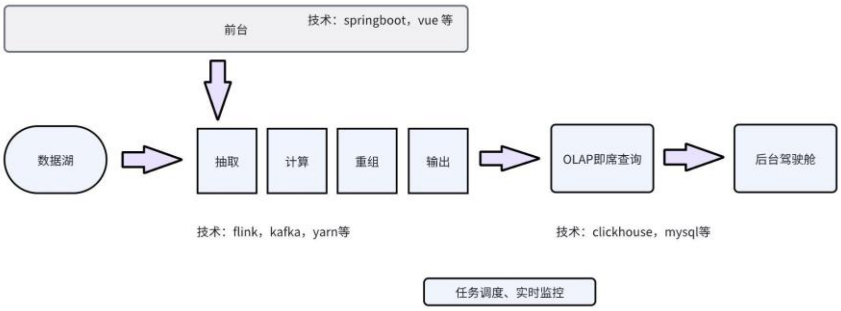
    c. HDP: Open source, secondary development. No CDH stable, less domestic use.

2. Cloud services:

    a. **Alibaba Cloud EMR,** MaxCompute, DataWorks; Amazon Cloud EMR; Tencent Cloud EMR; And Huawei Cloud EMR

    b. Simple operation and maintenance, charges, domestic users are growing year by year.

3. Technical architecture



4. Technology selection

Calculation Engine -Flink

Demand: T+0 within the day, in real time.

Flink: Streaming, state back-end, high accuracy and timeliness.

SparkStreaming: microbatch, timeliness, computation rate is not as good as flink.

Storm: Lack of accuracy.

Data storage -ClickHouse

Requirements: Real-time aggregate queries

Clickhouse: The DWS layer to the ads layer need to be regrouped and aggregated. clickhouse has the advantage of column storage for aggregation operations, and is very good at aggregation analysis of single width tables.

Hbase: k-v storage, which is more suitable for k-v query of single data.

Message Queue -Kafka

Kafka: Decoupling, peaking, suitable for large data volumes.

MQ: Throughput is not as good as Kafka, not suitable for big data scenarios.

Data synchronization -flinkCDC

Reasons for selection:

a. Based on Debezium implementation, real-time synchronization, and low latency. It improves the defects of debezium single machine and can interconnect with distributed system.

b. With flinkCDC, dimension configuration information can be directly synchronized to flink, without going through kafka.

	Query-based CDC	Binlog-based CDC
Open source products	Sqoop, Kafka JDBC Source	Canal, Maxwell, Debezium
Execution mode	Batch	Streaming
Whether all data changes can be captured	no	is
retardation	High latency	Low Latency
Whether to increase database stress	is	no

Data synchronization policy: Incremental synchronization is adopted because the user table has a large amount of data.

mysql: Enable binlog.

Application Data -SpringBoot interface Publishing &MySQL

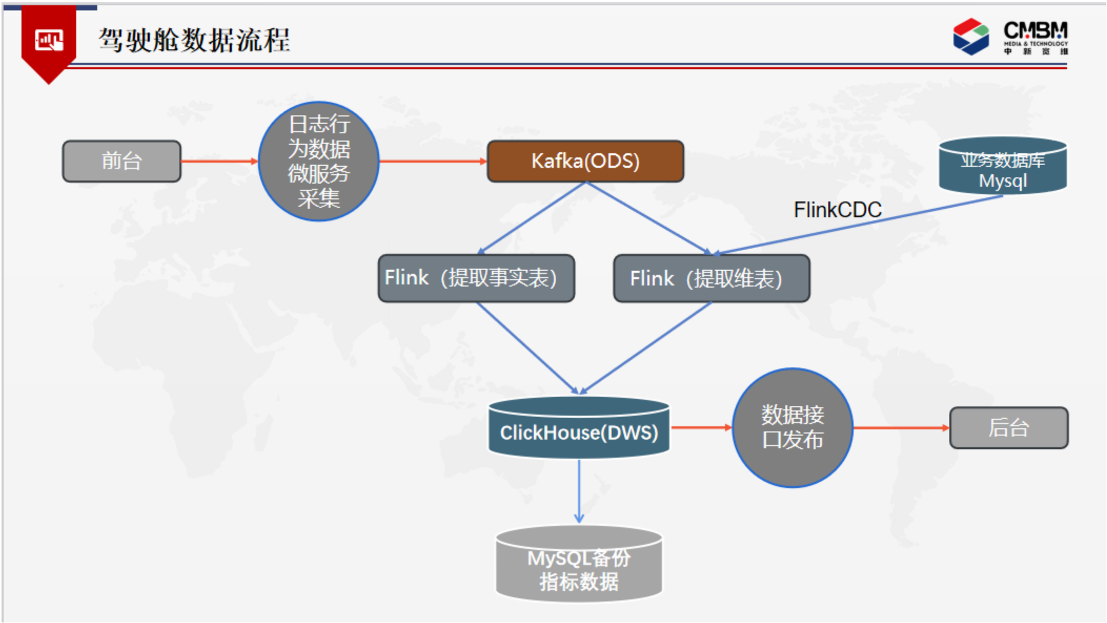
SpringBoot: Use the SpringBoot microservice to query the OLAP library for aggregate analysis, and publish the calculation results into the data interface to the cockpit.

MySQL: indicator data backup.

Log Data Acquisition -SpringBoot

Common, mature interaction between front and back ends.

5. Data flow



驾驶舱数据流程

前台

日志行为数据微服务采集

Kafka(ODS)

业务数据库 Mysql

FlinkCDC

Flink（提取事实表）

Flink（提取维表）

ClickHouse(DWS)

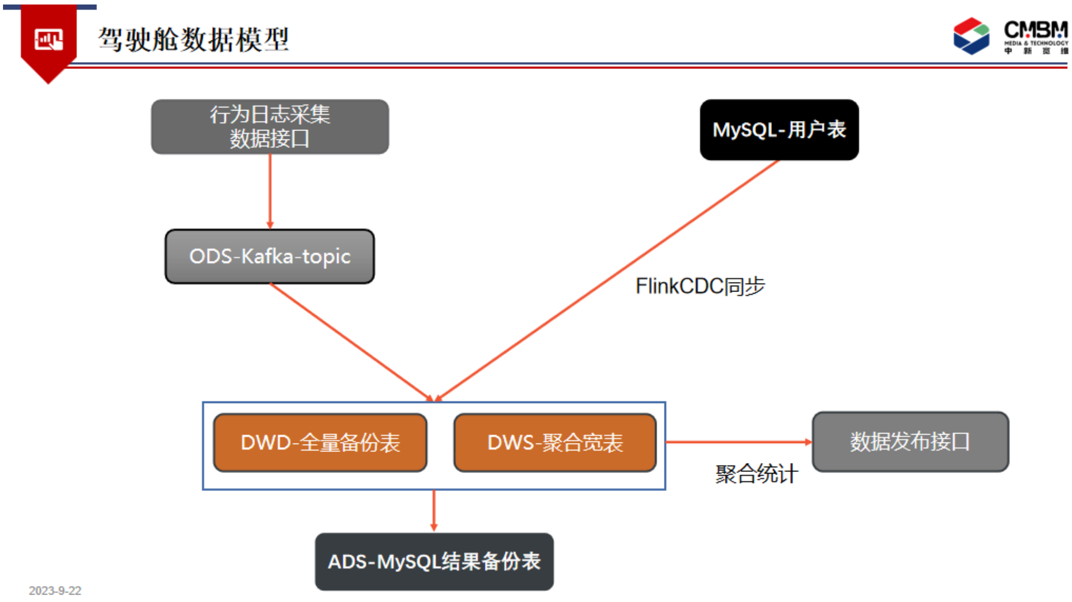
数据接口发布

后台

MySQL备份指标数据

Ii. Data system

Hierarchical model



驾驶舱数据模型

行为日志采集数据接口

MySQL-用户表

ODS-Kafka-topic

FlinkCDC同步

DWD-全量备份表

DWS-聚合宽表

聚合统计

数据发布接口

ADS-MySQL结果备份表

2023-9-22

Dimensional model

1. Facts: Corresponding business processes (individual, indivisible behavioral events)

    a. Behavior log table, subtopic.

2. Dimension: corresponds to the environment in which the business process occurs.
    a. Users
    b. District
    c. Date

DWS

ClickHouse: Log wide table

ODS

Kafka: Behavior Log topic

```
JSON
 Data field: Browsing behavior, behavior time
, Sample data:
 {}
```

MySQL: User business table

Field name	Field description	Data type	Default values	Sample data	Extraction Rules
	User ID				
	Username				
	Gender				
	Age				

	Area				
	Registration time				

Data acquisition API interface

Interface name	Interface path	Request method	Sample data

ADS

MySQL: Metric result sheet

Publishing API interface

Interface name	Interface path	Request method	Sample data
