**基于大数据的电信话单分析**

课程名称：中间件

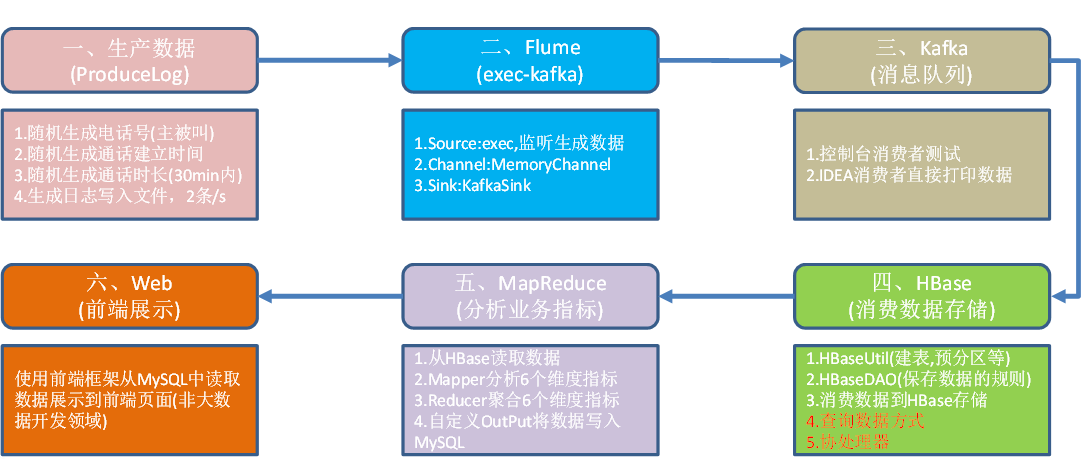
学生：王建欣

学号:2018104155

# 第1节 项目背景

通信运营商每时每刻会产生大量的通信数据，例如通话记录，短信记录，彩信记录，第三方服务资费等等繁多信息。数据量如此巨大，除了要满足用户的实时查询和展示之外，还需要定时定期的对已有数据进行离线的分析处理。例如，当日话单，月度话单，季度话单，年度话单，通话详情，通话记录等等+。通过此背景，学习其中的方法论。我们的需求是：统计每天、每月以及每年的每个人的通话次数及时长。

# 第2节 项目架构



# 第3节 项目实现

**系统环境：**

**表1**

|  |  |
| --- | --- |
| 系统 | 版本 |
| windows | 10 专业版 |
| linux | CentOS-6.5-i386-minimal |

**开发工具**：

**表2**

|  |  |
| --- | --- |
| 工具 | 版本 |
| idea | 2018 |
| maven | 3.3.9 |
| JDK | 1.8 |

**集群环境：**

**表3**

|  |  |
| --- | --- |
| 框架 | 版本 |
| hadoop | 2.7.2 |
| zookeeper | 3.4.10 |
| hbase | 1.3.1 |
| flume | 1.7.0 |
| kafka | 2.11-0.11.0.0 |

**硬件环境：**

**表4**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Spark2upgrade01 | Spark2upgrade02 | Spark2upgrade03 |
| 内存 | 4G | 2G | 2G |
| CPU | 2核 | 1核 | 1核 |
| 硬盘 | 50G | 20G | 20G |

## 3.1 数据生产(模拟生成数据)

**a)** 创建Java集合类存放模拟的电话号码和联系人；

**b)** 随机选取两个手机号码当作“主叫”与“被叫”（注意判断两个手机号不能重复），产出call1与call2字段数据；

**c)** 创建随机生成通话建立时间的方法，可指定随机范围，最后生成通话建立时间，产出date\_time字段数据；

**d)** 随机一个通话时长，单位：秒，产出duration字段数据；

**e)、**将产出的一条数据拼接封装到一个字符串中；

**f)、**使用IO操作将产出的一条通话数据写入到本地文件中

### 3.2.1 数据采集

**思路：**

**a)** 配置kafka，启动zookeeper和kafka集群；

**b)** 创建kafka主题；

**c)** 启动kafka控制台消费者（此消费者只用于测试使用）；

**d)** 配置flume，监控日志文件；

**e)** 启动flume监控任务；

**f)** 运行日志生产脚本；

### 3.2.2 数据消费

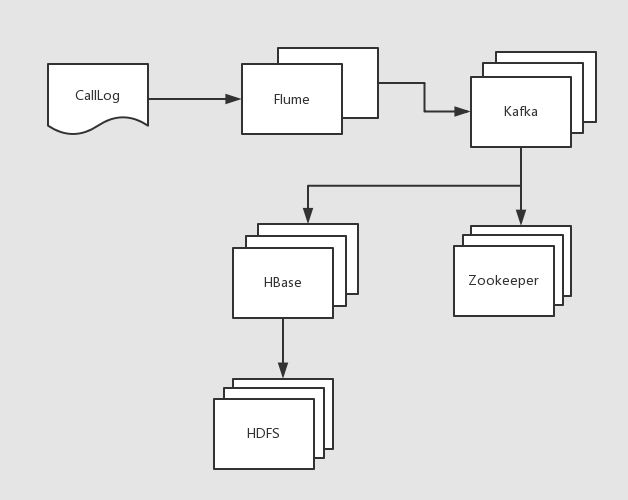
如果以上操作均成功，则开始编写操作HBase的代码，用于消费数据，将产生的数据实时存储在HBase中。

**思路：**

**a)** 编写kafka消费者，读取kafka集群中缓存的消息，并打印到控制台以观察是否成功；

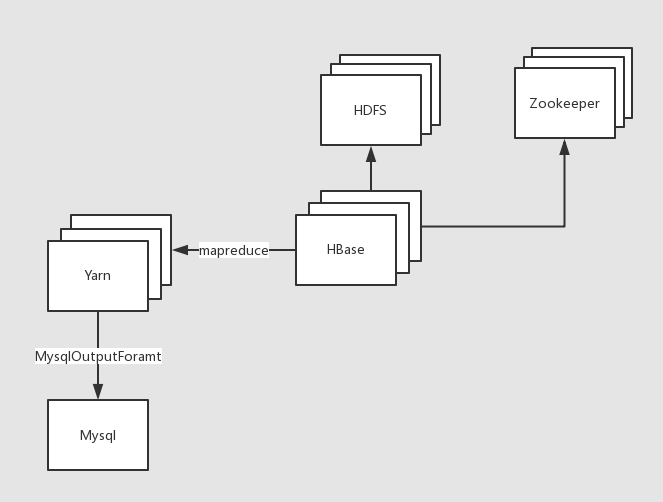
**b)** 既然能够读取到kafka中的数据了，就可以将读取出来的数据写入到HBase中，所以编写调用HBaseAPI相关方法，将从Kafka中读取出来的数据写入到HBase；

**c)** 以上两步已经足够完成消费数据，存储数据的任务。



## 3.3 数据分析

数据已经完整的采集到了HBase集群中，采取一个mapreduce分析多个业务指标的方式来进行任务。



**业务指标：**

a) 用户每天主叫通话个数统计，通话时间统计。

b) 用户每月通话记录统计，通话时间统计。

### 3.3.1 需求分析

根据需求目标，设计出下述（3.2.2）表结构。我们需要按照时间范围（年月日），结合MapReduce统计出所属时间范围内所有手机号码的通话次数总和以及通话时长总和。

### 3.3.2 Mysql表结构设计

我们将分析的结果数据保存到Mysql中，以方便Web端进行查询展示。

**1) 表7：ct\_user**

用于存放用户手机号码与联系人姓名。

**表7 ct\_user**

|  |  |  |
| --- | --- | --- |
| 列 | 备注 | 类型 |
| id | 自增主键 | int(11) NOT NULL |
| telephone | 手机号码 | varchar(255) NOT NULL |
| name | 联系人姓名 | varchar(255) NOT NULL |

**2) 表8：ct\_call**

用于存放某个时间维度下通话次数与通话时长的总和。

**表8 ct\_call**

|  |  |  |
| --- | --- | --- |
| 列 | 备注 | 类型 |
| id\_date\_contact | 复合主键（联系人维度id，时间维度id） | varchar(255) NOT NULL |
| id\_date\_dimension | 时间维度id | int(11) NOT NULL |
| id\_contact | 查询人的电话号码 | int(11) NOT NULL |
| call\_sum | 通话次数总和 | int(11) NOT NULL DEFAULT 0 |
| call\_duration\_sum | 通话时长总和 | int(11) NOT NULL DEFAULT 0 |

**3) 表9：ct\_date**

用于存放时间维度的相关数据

**表9 ct\_date**

|  |  |  |
| --- | --- | --- |
| 列 | 备注 | 类型 |
| id | 自增主键 | int(11) NOT NULL |
| year | 年，当前通话信息所在年 | int(11) NOT NULL |
| month | 月，当前通话信息所在月，如果按照年来统计信息，则month为-1。 | int(11) NOT NULL |
| day | 日，当前通话信息所在日，如果是按照月来统计信息，则day为-1。 | int(11) NOT NULL |

## 3.4 数据展示

将某人按照不同维度查询出来的结果，展示到web页面上。

