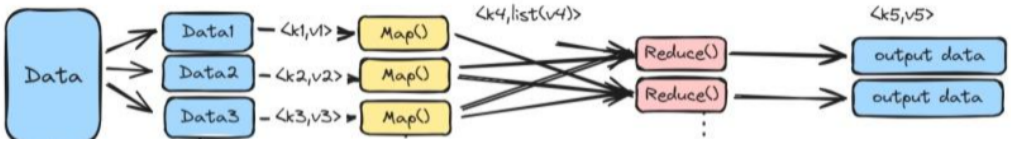
20250627日报

今日工作内容

1、学习spark相关知识，简单了解了Hadoop、HDFS、MapReduce、Hive。

- Hadoop：它是一个开源的分布式计算框架，专为海量数据存储与批处理设计，核心思想是通过集群化实现横向扩展。Hadoop主要由三大模块组成：HDFS（分布式文件系统）、MapReduce（计算模型）和YARN（资源管理器）。
- HDFS：它是一个分布式文件系统，实际上就是对部署在多台独立物理机器上的文件进行管理。HDFS架构主要包含：
  - NameNode：用于存储、生成文件系统的元数据。
  - DataNode：用于存储实际的数据，将自己管理的数据块上报给NameNode。
  - Client：支持业务访问HDFS，从NameNode、DataNode获取数据返回给业务。
- MapReduce：将任务拆分为Map和Reduce两阶段。在Map阶段，并行处理本地数据，生成键值对；在Reduce阶段，汇总相同键的值。对于MapReduce就是实现Map和Reduce两个函数逻辑，而这个目标可以通过Spark的 RDD API实现。总的来说，在分布式环境中MapReduce对数据并行处理的过程为：分片->Map->Shuffle->Reduce。Spark SQL可以实现Map和Reduce逻辑。



- Hive：它是基于Hadoop的数据仓库工具，核心功能是将结构化数据文件映射为数据库表，并提供类SQL查询功能。数据实际存储在HDFS，Hive仅管理元数据（表结构、分区信息等），元数据通常存储在外部数据库（如MySQL）中。HiveQL可以实现Map和Reduce逻辑。
- Spark：这是一个专为大规模数据处理设计的快速、通用计算引擎，它比Hadoop MapReduce快10~100倍，其原因一方面是中间结果优先存于内存而非磁盘，减少磁盘IO；另一方面是通过DAG（有向无环图）调度优化任务流程，减少冗余计算。Spark有五大组件：Spark Core、Spark SQL、Spark Streaming、MLlib、GraphX，并且兼容Hive。Spark SQL可以实现MapReduce逻辑，从而完成hive表的分布式Map()和Reduce()操作。spark定位hive表实际数据的hdfs节点是通过元数据服务实现的，定位元数据服务是通过配置实现的，即可以在.xml配置文件中包含元数据的Thrift服务地址或直接在代码中指定地址。对于首充续充需求，我猜想的处理过程为：
  - 初始化SparkSession
  - 读取Hive表并转换为core.KvMap{}的结构（用Spark SQL编写逻辑）
  - 按时间戳取最新记录，即玩家的最新首充续充记录，并写入redis

明日待办：

1. 首充续充需求的定向。