分布式计算课程项目

设计和实现一个由多个节点组成的分布式系统，完成分布式存储和分布式计算的功能。

一、项目要求

1. 分布式系统的节点数量不少于 4 个，节点可以是物理机，也可以是虚拟机。
2. 系统拥有基本的分布式存储功能，包括：
    1) 文件的上传与下载。从任意一个节点都能上传文件到系统中，也能从任意一个节点访问并下载系统中的每个文件。
    2) 文件的分块与备份。系统对大容量文件以分块的形式存储，并且系统中存储的每个文件都有多个副本，当系统中不超过 20%的节点失效时，也不影响系统中所有文件的访问。
    3) 文件的一致性。从任意一个节点访问并更新某个文件后，其在系统中的副本也相应进行更新。
3. 根据附录中的数据文件，使用分布式系统的计算资源，完成以下计算：
    1) 计算出用户的每日平均通话次数，并将结果以<主叫号码, 每日平均通话次数>的格式保存成 txt 或 excel 文件。
    2) 计算出不同通话类型（市话、长途、国际）下各个运营商（移动，联通，电信）的占比，并画出饼状图。
    3) 计算出用户在各个时间段（时间段的划分如表 1 所示）通话时长所占比例，并将结果以<主叫号码, 时间段 1 占比, ..., 时间段 8 占比>的格式保存成 txt 或 excel 文件。

表 1 时间段划分表

时间段名称	时间段的起止时间
时间段 1	0:00-3:00
时间段 2	3:00-6:00
时间段 3	6:00-9:00
时间段 4	9:00-12:00
时间段 5	12:00-15:00
时间段 6	15:00-18:00
时间段 7	18:00-21:00
时间段 8	21:00-24:00

二、项目提交文档

1. 项目报告，介绍系统的分布式存储和分布式计算的方法，包括系统的架构、文件分块方法、文件备份方法、文件的一致性策略、计算任务分配机制等。

2. 数据文件的计算结果，以 txt 或 excel 文件保存。

三、附录：数据文件

1. 数据文件下载地址：百度网盘 https://pan.baidu.com/s/1bXHpxW

2. 数据文件为 txt 文档，主要包含字段：主叫号码、通话开始时间、通话时长、通话类型、主叫号码运营商等。数据文件的字段说明如表 2 所示。

表 2 数据文件的字段说明表

字段名	字段含义	备注
day_id	日期	
calling_nbr	主叫号码	全部为本运营商加密后的手机号码
called_nbr	被叫号码	g 开头号码表示各运营商各城市固话号码，y 开头号码表示异网手机号码，其它为本运营商手机号码
calling_optr	主叫号码运营商	1：电信； 2：移动； 3：联通； 其它为不详
called_optr	被叫号码运营商	1：电信； 2：移动； 3：联通； 其它为不详
calling_city	主叫号码归属地	主叫号码所归属的城市
called_city	被叫号码归属地	被叫号码所归属的城市
calling_roam_city	主叫号码漫游地	主叫号码所在的漫游城市，没有漫游时则为空
called_roam_city	被叫号码漫游地	被叫号码所在的漫游城市，没有漫游时则为空
start_time	通话开始时间	格式：13:44:25（时:分:秒）
end_time	通话结束时间	格式：13:44:25（时:分:秒）
raw_dur	通话时长	单位：秒
call_type	通话类型	1：市话； 2：长途； 3：漫游
calling_cell	主叫蜂窝号码	所在的基站蜂窝标识或为空