##### 大数据的特征

大数据的4V特性

海量Volume：数据量大，包括采集、存储和计算的量都非常大

多样Variety：种类和来源多样化

价值Value：数据价值高，密度相对较低

高速Velocity：数据增长速度快，处理速度也快，时效性要求高

#### 大数据的处理模式

（批处理 流处理 交互式 ）

复杂的**批量数据**处理，通常的时间跨度在几分钟到数小时之间

基于**实时数据流**的处理，通常的时间跨度在数百毫秒到数秒之间

基于**历史数据**的交互式查询，通常时间跨度在数十秒到数分钟之间

#### 大数据相比传统数据处理的优势

*数据产生方式：主动生成数据。*

*数据采集密度:利用大数据平台，可对需要分析事件的数据进行密度采样,精确获取事件全局数据。*

*数据源：利用大数据技术，通过分布式技术、分布式文件系统、分布式数据库等技术对多个数据源获取的数据进行整合处理。*

*数据处理方式:较大的数据源、响应时间要求低的应用可以采取批处理方式集中计算;响应时间要求高的实时数据处理采用流处理的方式进行实时计算,并通过对历史数据的分析进行预测分析。*

**（答案来自网络）**

#### 大数据和云计算，和互联网有什么关系？为什么大数据会在这个时候出现？



*大数据出现的时间点自有它深刻的原因。2009年至2012年这段时间正是电子商务在包括中国在内的全球全面开花的几年。众所周知，互联网领域有3大类商业模式：广告、游戏和电子商务。而电子商务又是第1个真正将纯互联网经济与传统经济嫁接在一起诞生的混合模式。准确地说，正是互联网与传统经济的碰撞，才真正催生出了今天几乎全民关注的“大数据”。大数据横跨了互联网产业与传统产业，而且大数据真正广阔的应用领域其实也正是比纯互联网经济大得多的传统产业。*

*从数据量的角度来看，在电子商务模式出现以前，传统企业的数量增长缓慢。传统企业的数据仓库中的数据大多数来自于交易型数据，而交易这种行为处于用户消费决策漏斗的最底部，这就决定了交易前的各种浏览、搜索、比较等用户行为数据的都量远远超过交易数据。电子商务模式使得企业可以采集到用户的浏览、搜索、比较等行为，这就导致企业的数据规至少提升了一个数量级。现在日益流行的移动互联网以及将来会流行的物联网又必将使数据量提高两三个数量级。从这个角度来讲，大数据时代是必然会出现的。*

*从IT产业的发展来看，第一代IT巨头大多是2B的，比如IBM、Microsoft、Oracle、SAP这类传统IT企业；第二代IT巨头大多是2C的，比如Yahoo、Google、Amazon、Facebook这类互联网企业。一个有意思的现象是：大数据时代前，这两类公司彼此之间基本是井水不犯河水，我们很少看见这两类公司的老板们在一起坐而论道；但在当前这个大数据时代，这两类公司已经开始直接竞争。比如Amazon已经开始提供云模式的数据仓库服务，直接抢占IBM、Oracle的市场。这个现象出现的本质原因是：在互联网巨头的带动下，传统IT巨头的客户普遍开始从事电子商务业务，正是由于客户进入了互联网，所以传统IT巨头们不情愿地被拖入了互联网领域。如果他们不进入互联网，他们业务必将萎缩。所以第三代 IT 巨头可能会是2B与2C 融合的IT公司。*

**（答案来自网络）**

#### hdfs Hadoop strom yarn hive zokeep等大数据技术栈里的技术有什么用？在什么产品下会用到这些？

1. 大数据分布式文件存储（HDFS）

HDFS（Hadoop Distribute File System）Hadoop 分布式文件系统；

Hadoop 核心组件之一，作为最底层的分布式存储服务而存在；

解决数据存储问题，横跨在多台计算机上的存储系统

分布式文件系统在大数据时代有着广泛的应用前景，它们为存储和处理超大规模数据提供所需的扩展能力。

2、大数据分布式计算系统 （MapReduce）

大规模数据处理的并行计算模型和方法;

由Map和Reduce构成，采用“分而治之”策略。

存储在分布式文件系统中的大规模数据集，会被切分成许多独立的分片，这些分片可以被多个Map任务并行处理。

3、大数据分布式数据仓库 （Hive）

基于 Hadoop 的一个数据仓库工具

可以将结构化的数据文件映射为一张数据库表，并提供类 SQL 查询功能（本质是将 SQL 转换为 MapReduce 程序）

用途：用来做离线数据分析（比直接用 MapReduce 开发效率更高

4、Hadoop数据库（HBase）

一个分布式、可扩展、大数据的存储

是一种NoSQL数据库（非关系型的数据库）

大数据快速读写

Hbase适合用来进行大数据的实时查询；Hive需要很长时间才可以返回结果，不应该用来进行实时的查询

1. 实时大数据处理框架（Storm）

用于实时处理数据，流式计算

特点：低延迟，高可用，分布式，可扩展，数据不丢失，提供简单容易的接口，便于开发

6、大数据分布式内存计算（Spark）

一种快速、通用、可扩展的大数据分析引擎

基于内存计算的大数据并行计算框架

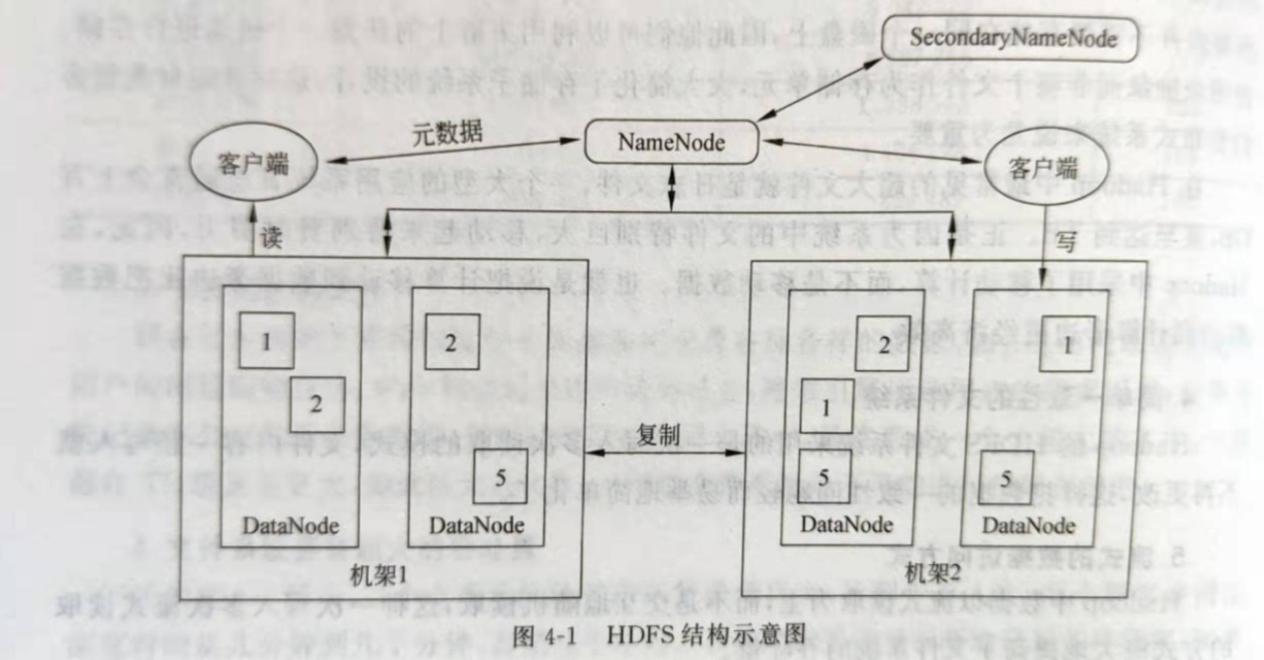
已经发展成为一个包含多个子项目的集合

包含： SparkSQL、Spark Streaming、GraphX、MLlib 等子项目

2014年，代替MapReduce成为Hadoop的默认执行引擎；Spark 实现了高效的 DAG 执行引擎，可以通过基于内存来高效处理数据流。与MapReduce 相比，Spark 基于内存的运算更快。

#### 第二章

#### hdfs 分布式存储是怎么存储的？



1.NameNode

NameNode是HDFS系统中的管理节点，它管理文件系统的命名空间、记录每个文件数据块在DataNode上的位置和副本信息、协调客户端对文件的访问、记录命名空间内的改动和空间本身属性的改动。

NameNode使用事务日志(EditsLog)记录HDFS元数据的变化。使用映像文件存储文件系统的命名空间，包括文件映射、文件属性等。

2.DataNode

DataNode节点是HDFS系统中保存数据的节点，负责所在物理节点的存储管理，一次写入，多次读取。文件由数据块组成，默认的块大小是64MB，数据块以元余备份的形式分布在不同机架的不同机器上，DataNode定期向NameNode提供其保存的数据块的列表，以便于客户端在获取文件元数据后直接在DataNode上进行读写。

3. SecondaryNameNode

在Hadoop中只有一个NameNode，所以存在单点问题，即当NameNode节点宕机时，整个系统就瘫痪了。为了解决这个问题，在Hadoop中增加了一个 SecondaryNameNode 节点，它一般运行在单独的计算机上，定期从NameNode节点获取数据，形成NameNode的备份，一旦NameNode出现问题， SecondaryNameNode 便可以顶替NameNode。Hadoop 1.X中没有真正解决单点问题，到2.X后解决了系统的单点问题，提高了系统的可用性，请参照4.6节内容。

4.客户端

客户端是Hadoop集群的使用者，他通过HDFS的Shell和API对系统中的文件进行操作。

5.机架

机架(Rack)是容纳组成集群的普通商用计算机(节点)的架子，各节点被分到不同的机架内，这样做是为了管理和施工的需要，也是为了防灾容错的需要，一个机架内的机器因为灾难、掉电、故障等原因一起宕机的可能性比较大，把机器分配到不同机架后，所有机器一起宕机的可能性则比较小，这也是HDFS文件系统安全的原因。一般来说，一个机架内的节点间数据传输速率要快，而不同机架内的节点间数据传输速率要慢一些。

6.数据块

在操作系统中有数据块的概念，比如磁盘中有扇区，它的大小为512字节，这是进行数据读写的最小单位，操作系统中的数据块是磁盘扇区块的整数倍。在HDFS中也有数据块的概念，不过要大得多，默认是64MB，与磁盘数据块类似，数据块（Chunk）是Hadoop处理数据的最小单位，一个Hadoop文件被分成若干数据块，根据Hadoop数据管理的策略，数据块被放置在不同的数据节点(DataNode)中，当数据块的副本数小于规定的份数时，Hadoop系统的错误检测与自动恢复技术就会自动把数据块的副本数恢复到正常水平。

#### hdfs什么是元数据？

元数据：描述数据的数据

维护HDFS文件系统中的文件和目录所需要的信息。

#### hdfs备份机制？读机制？写机制？

备份机制

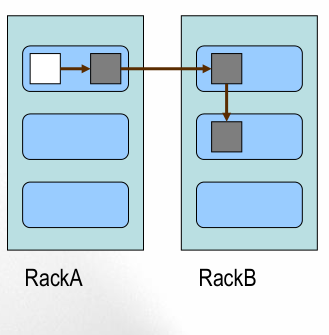
Block的副本放置策略

第一个副本：放置在上传文件的DN；如果是集群外提交，则随机挑选一台磁盘不太满，CPU不太忙的节点。

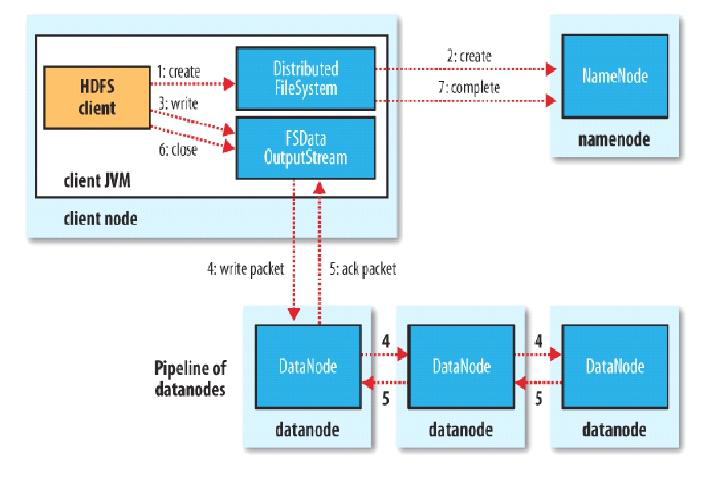
–第二个副本：放置在于第一个副本不同的 机架的节点上。

–第三个副本：与第二个副本相同机架的节点。

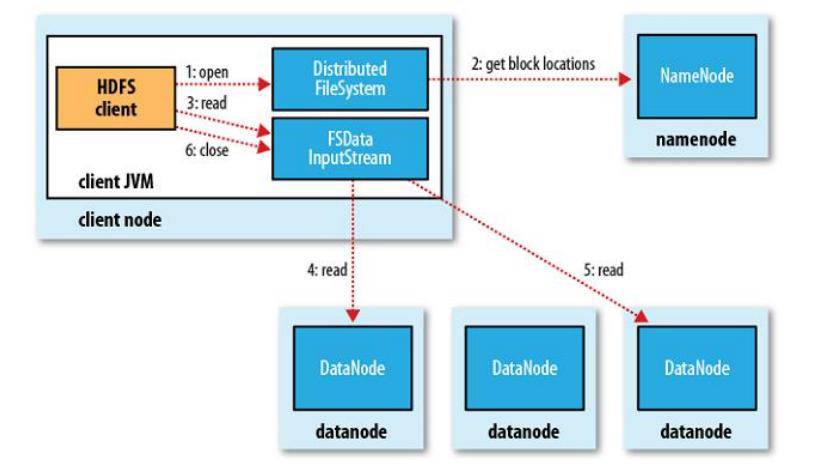
–更多副本：随机节点



HDFS写流程

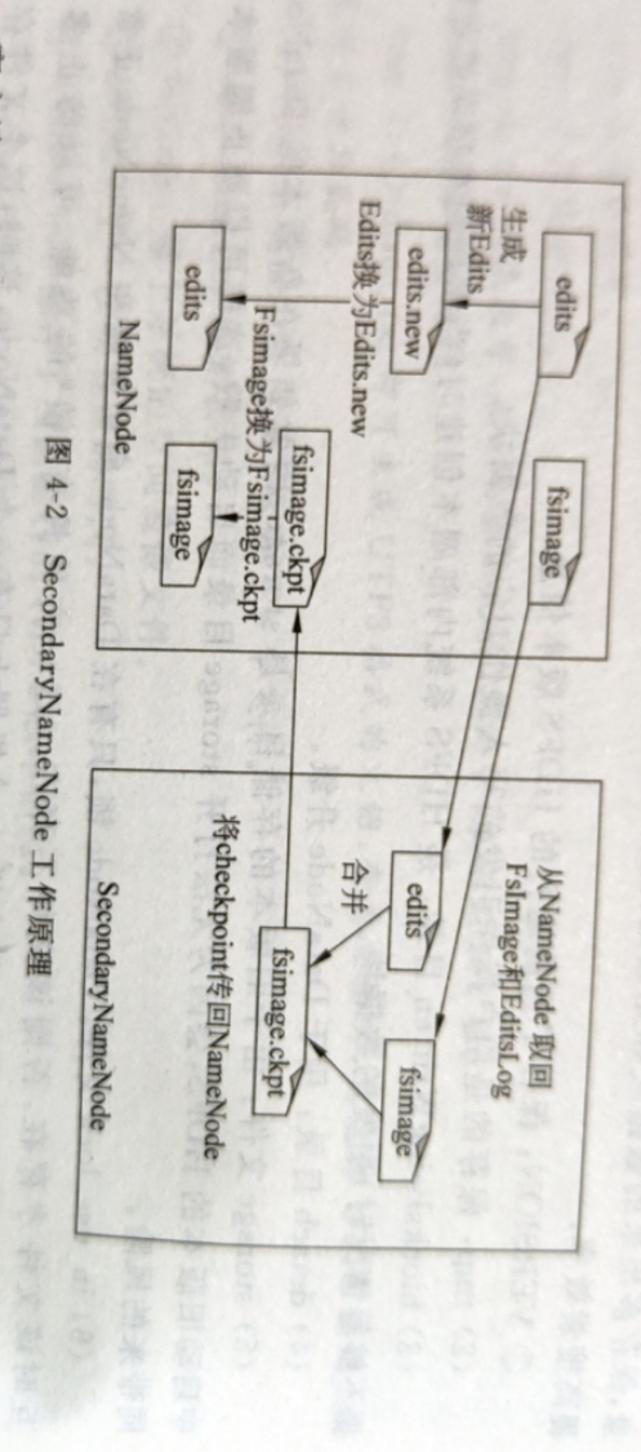


HDFS读文件



鲨鲨比亚:

#### namenode secondarynamenode是怎么更新数据的



鲨鲨比亚:

（p73 和4.3.3等 数据的组成....数据块的备份，读写）

鲨鲨比亚:

MapReduce理解整个流程

#### MapReduce的代码要会写（MapReduce的实验）

#### 为什么会有Hive？

基于类SQL实现大数据数分析（数据仓库）

#### Hive底层是什么？

待填充

#### HBASE是一个什么类型的数据库

Nosql非关系型的数据库

#### 有哪些分类？如键值存储有什么场景 图形存储有什么场景，图形存储有什么场景？

键值(Key-Value)存储数据库Redis

列存储数据库。Cassandra, HBase

文档型数据库MongoDb

图形(Graph)数据库Neo4J

#### HBASE的列族是什么？要会列族创建的代码 课本p171 HBASE...

鲨鲨比亚:

搞清楚列族和列

HBase表中的每个列都归属于某个列族，列族必须作为表模式(schema)定义的一部分预先给出。

列名以列族作为前缀，每个“列族”都可以有多个列成员(column)；如course:math, course:english, 新的列族成员（列）可以随后按需、动态加入；

##### spark 要知道spark和MapReduce有什么关系？有什么不同？

Spark在借鉴Hadoop MapReduce优点的同时，很好地解决了MapReduce所面临的问题

Spark的计算模式也属于MapReduce，但不局限于Map和Reduce操作，还提供了多种数据集操作类型，编程模型比Hadoop MapReduce更灵活

Spark提供了内存计算，可将中间结果放到内存中，对于迭代运算效率更高

Spark基于DAG的任务调度执行机制，要优于Hadoop MapReduce的迭代执行机制

#### DAG有向无环图

https://www.itcast.cn/news/20201229/17080822855.shtml

#### 弄有向无环图的目的是什么？有什么用

**待填充**

#### rdd等概念要掌握

RDD：是Resillient Distributed Dataset（弹性分布式数据集）的简称，是分布式内存的一个抽象概念，提供了一种高度受限的共享内存模型

DAG：是Directed Acyclic Graph（有向无环图）的简称，反映RDD之间的依赖关系

Executor：是运行在工作节点（WorkerNode）的一个进程，负责运行Task

Application：用户编写的Spark应用程序

Task：运行在Executor上的工作单元

Job：一个Job包含多个RDD及作用于相应RDD上的各种操作

Stage：是Job的基本调度单位，一个Job会分为多组Task，每组Task被称为Stage（或者也被称为TaskSet），代表了一组关联的、相互之间没有Shuffle依赖关系的任务组成的任务集

#### 架构...



Spark运行架构包括集群资源管理器（Cluster Manager）、运行作业任务的工作节点（Worker Node）——资源管理

Master/Slave：每个应用的任务控制节点（Driver）和每个工作节点上负责具体任务的执行进程（Executor）——作业管理

资源管理器可以自带或Mesos或YARN

一个Application由一个Driver和若干个Job构成，一个Job由多个Stage构成，一个Stage由多个没有Shuffle关系的Task组成

当执行一个Application时，Driver会向集群管理器申请资源，启动Executor，并向Executor发送应用程序代码和文件，然后在Executor上执行Task，运行结束后，执行结果会返回给Driver，或者写到HDFS或者其他数据库中

#### 运行流程



1）首先为应用构建起基本的运行环境，即由Driver创建一个SparkContext，进行资源的申请、任务的分配和监控

（2）资源管理器为Executor分配资源，并启动Executor进程

（3）SparkContext根据RDD的依赖关系构建DAG图，DAG图提交给DAGScheduler解析成Stage，然后把一个个TaskSet提交给底层调度器TaskScheduler处理；Executor向SparkContext申请Task，Task Scheduler将Task发放给Executor运行，并提供应用程序代码

（4）Task在Executor上运行，把执行结果反馈给TaskScheduler，然后反馈给DAGScheduler，运行完毕后写入数据并释放所有资源

#### Storm 是干什么用的，应用场景有哪些？解决了什么问题？

Apachestorm是一个免费和开源的分布式实时计算系统。

ApacheStorm可以轻松可靠地处理无限制的数据流，实现Hadoop对批处理的实时处理。

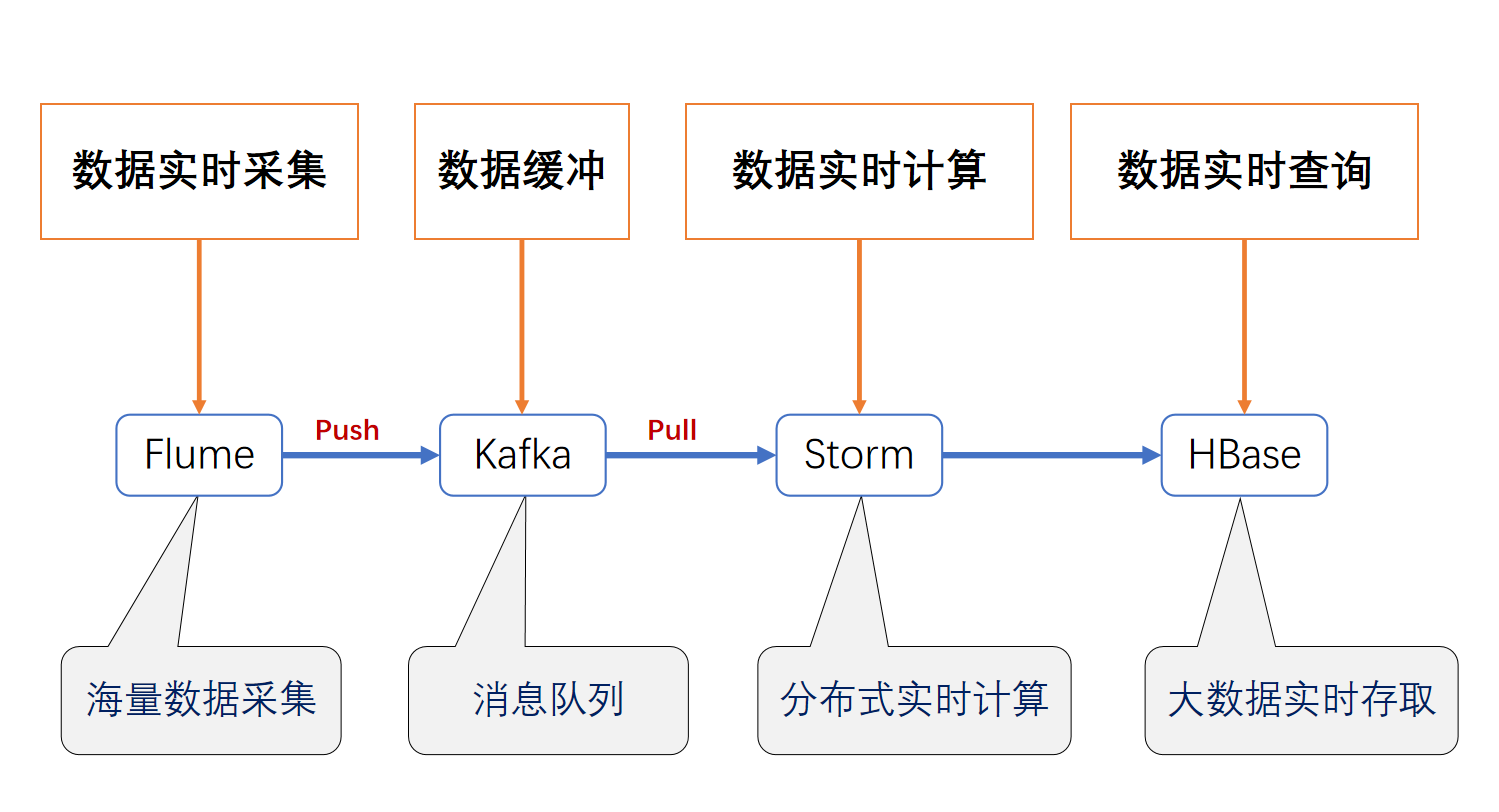
场景：Twitter流式处理大数据分析（话题趋势算法）

网路流量实时分析

基于GPS的实时路况分析

电信数据管理

#### 在进行实时计算的时候基本流程是什么样的？



#### 要明白strom的 拓扑是什么意思 Stream（流） 等什么意思

拓扑是一个有向图，其顶点是计算，边是数据流，包括了整个实时处理程序的逻辑。

Stream（数据流：无限元组序列）

Tuple（元组：Storm提供的一个轻量级的数据结构）

Spout（喷嘴：数据流的来源，负责接受数据源的数据进行分发）

一个喷口是一个拓扑结构中的一个流的来源。

Bolt（处理节点：负责接受数据进行计算，传向下个bolt或进行存储）

拓扑结构中的所有处理都用处理节点完成了。处理节点可以做包括过滤、函数、聚合、连接、与数据库交互等在内的任何事情。

Task（任务）

Component（组件）

Stream groupings（流分组）

#### 最后一次作业要做

#### 六道大题

#### 每道大题下面两个小题