## hadoop大数据平台架构与实践

```

掌握大数据存储与处理技术的原理

掌握hadoop的使用和开发能力

```

Google大数据技术

1、MapReduce

2、Bigtable

3、GFS

革命性的变化：

1、降低成本，PC机替代大型机

2、硬件故障视为常态，用软件保证高可靠

3、简化了并行计算模式，无需控制节点同步和数据交换

核心：

1、HDFS：分布式文件系统 存储海量数据

2、MapReduce：并行处理框架，实现任务分解和调度

## Hadoop用来做什么？

A: 搭建大型数据仓库，PB级数据的存储、处理、分析、统计等业务。

如搜索引擎、日志分析、商业智能、数据挖掘

## Hadoop的优势：

1、高扩展

2、低成本

3、成熟的生态圈

## Hadoop的生态：

HDFS+MapReduce+Hive+Hbase+zookeeper

## hadoop的安装和配置

wget http://mirror.bit.edu.cn/apache/hadoop/common/hadoop-1.2.1/hadoop-1.2.1.tar.gz

1、下载解压JDK、安装JDK、配置环境变量

2、下载解压HADOOP+HADOOP环境变量

3、4个配置文件

配置文件1：hadoop-env.sh（外部环境）

配置java目录

配置文件2：mapred-site.xml（任务调度器）

设置map-reduce的job-tracker如何去访问

```sh

<configuration>

<property>

<name>mapred.job.tracker</name>

<value>localhost:9001</value>

</property>

</configuration>

```

配置文件3：hdfs-site.xml（数据存放目录）

//hdfs文件块的目录

```sh

<configuration>

<property>

<name>dfs.data.dir</name>

<value>/hadoop/data</value>

</property>

</configuration>

```

配置文件4：core-site.xml（工作目录）

```sh

<configuration>

//hadoop的临时工作目录

<property>

<name>hadoop.tmp.dir</name>

<value>/hadoop</value>

</property>

//hadoop的元数据目录

<property>

<name>dfs.name.dir</name>

<value>/hadoop/name</value>

</property>

//文件系统访问

<property>

<name>fs.default.name</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>

```

4、hadoop namenode -format

5、启动

## Hadoop的核心-HDFS

**HDFS的基本概念**

HDFS的文件以块存储，默认64MB

NameNode：

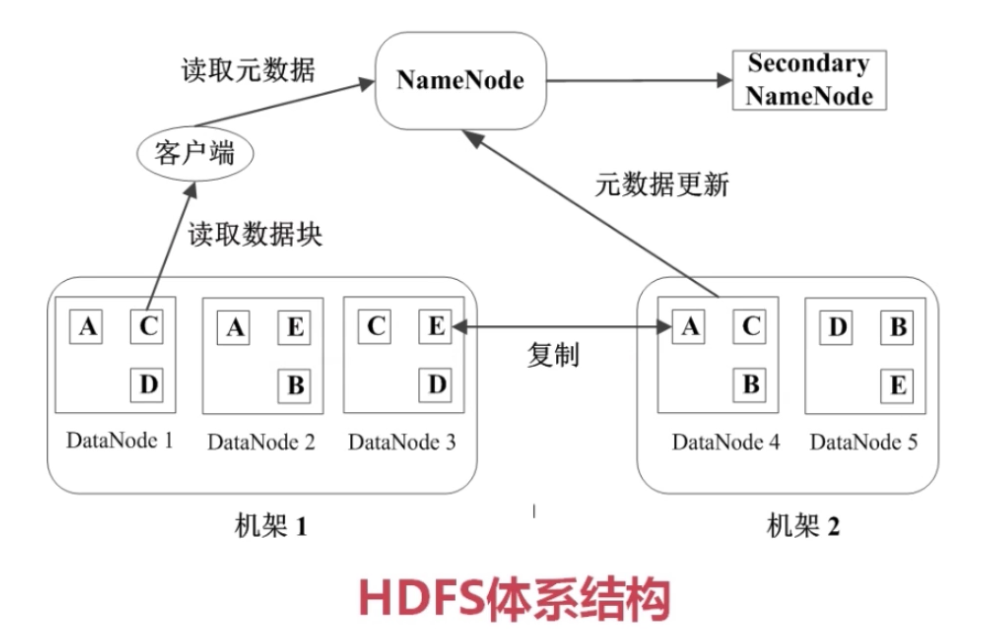
管理节点：存放文件元数据

1、文件与数据块的映射表

2、数据块与数据节点的映射点

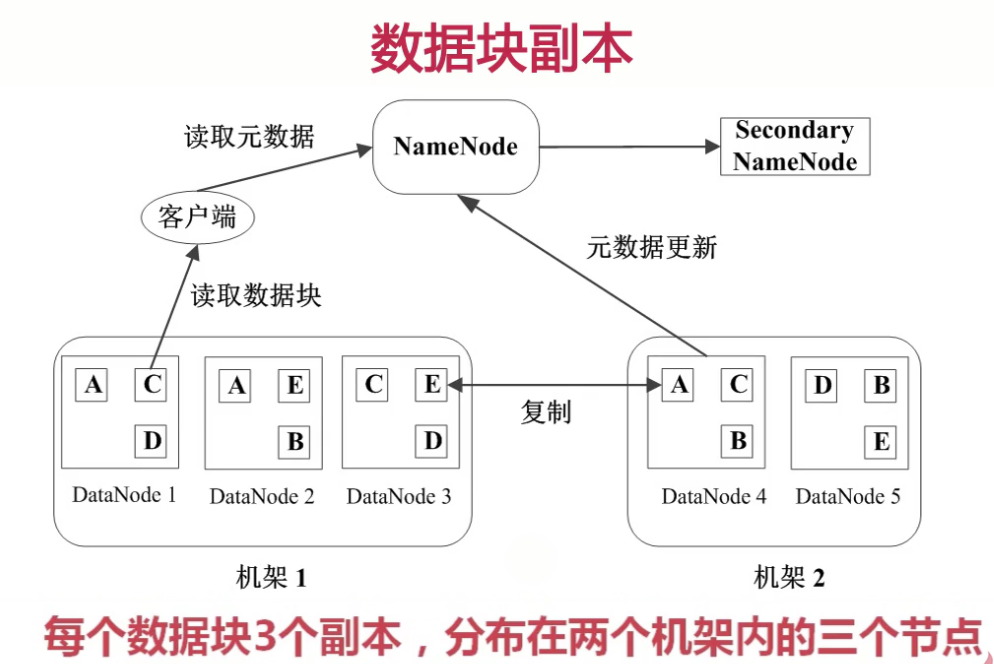
DataNode：

工作节点：存放数据块

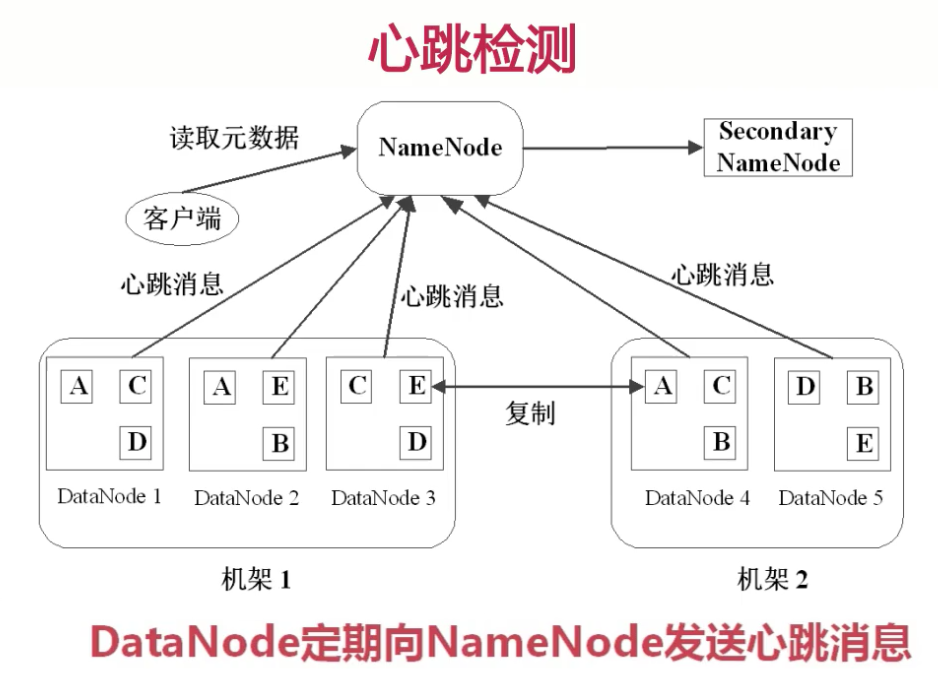


**HDFS的数据管理策略**

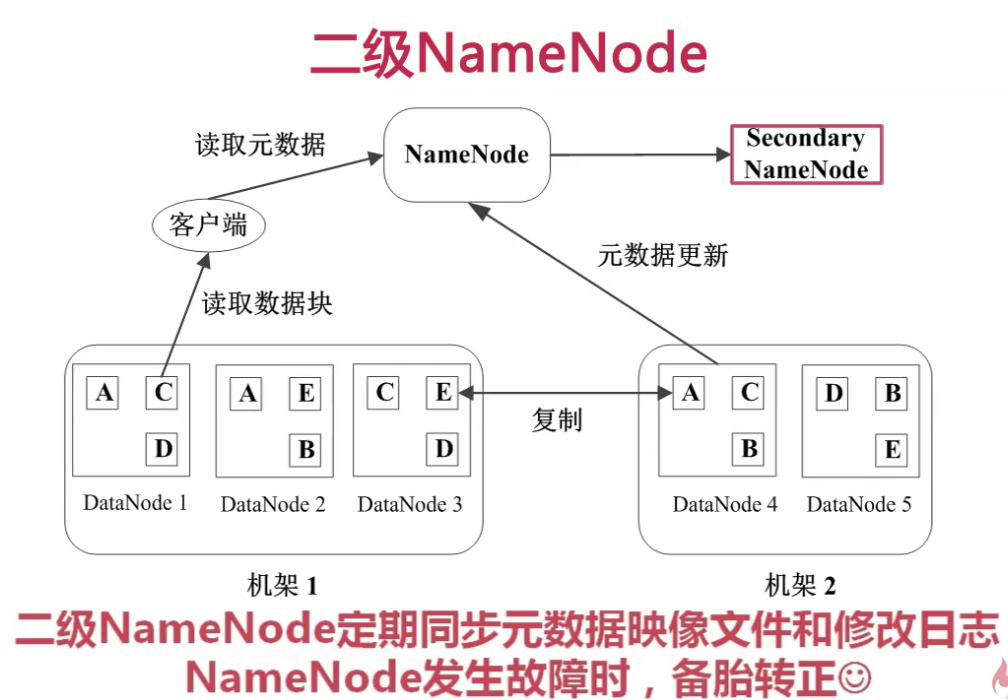
1、数据块副本



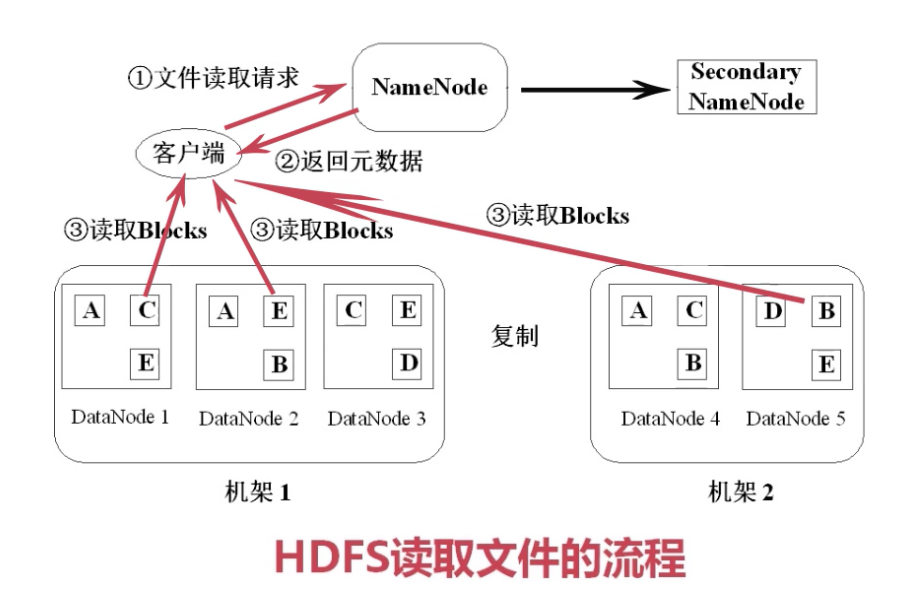
2、心跳检测

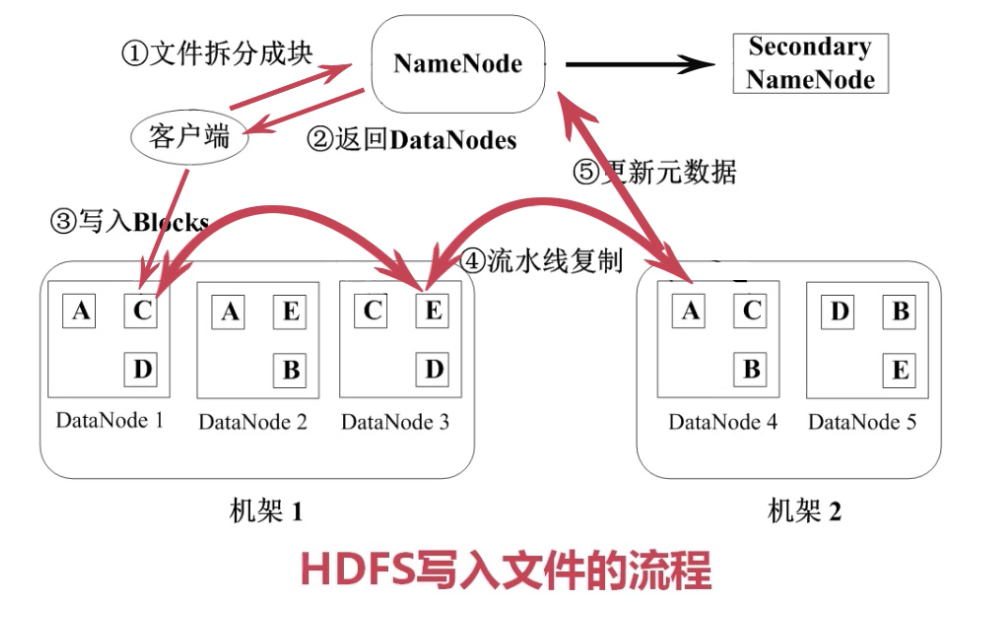


3、二级NameNode



**HDFS的文件读写**





**HDFS的特点**

1、数据冗余，硬件容错

2、流式的数据访问（写一次，读多次）

3、存储大文件

适用性和局限性

* 适合数据批量读写，吞吐量高
* 不适合交互式应用，低延时很难满足
* 适合一次写入，多次读写，顺序读写
* 不支持多用户并发写相同文件

**HDFS的使用**

hadoop fs -put hadoop-env.sh taotaoinput/

hadoop fs -get taotaoinput/hadoop-env.sh hadoop-tao.sh

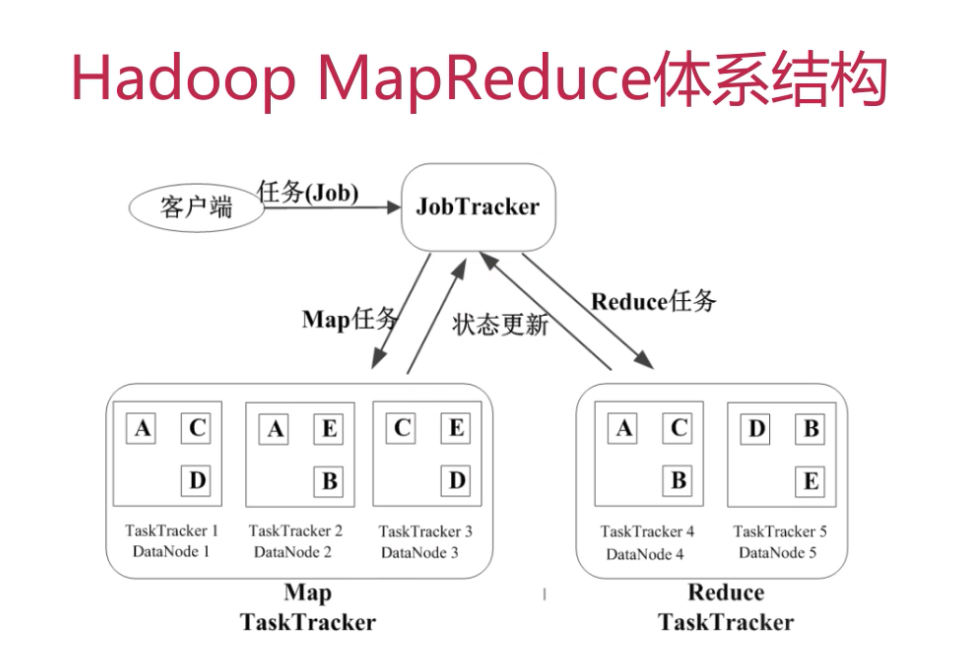
hadoop dfsadmin –report

**MapReduce原理**

原理：分而治之，把一个大任务分成多个小的子任务（map），并行执行后，合并结果（reduce）

**MapReduce的运行流程**

* **Job & Task**
* **JobTracker**
* **TaskTracker**



**JobTracker的角色**

1、作业调度

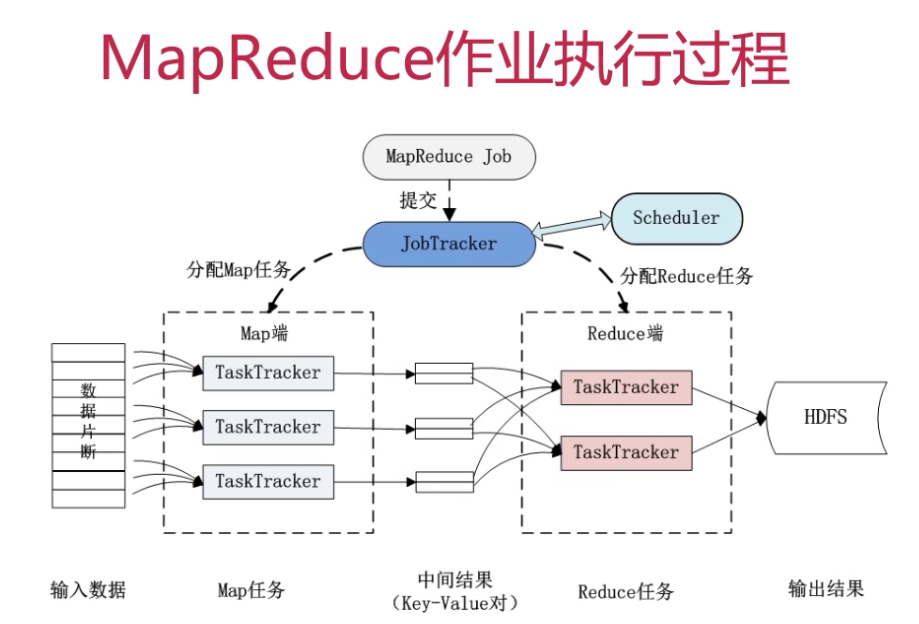
2、分配任务、监控任务执行进度

3、监控TaskTracker的状态

**TaskTracker的角色**

1、执行任务

2、汇报任务状态



**MapReduce的容错机制**

1、重复执行 X4

2、推测执行