# 一 MapReduce流程

[link\MapReduce过程详解(基于hadoop2.x架构) - py小杰 - 博客园.mhtml](link/MapReduce过程详解(基于hadoop2.x架构) - py小杰 - 博客园.mhtml)





# 二 hive

## hive简介

## Hive的sql与传统sql区别

## Hive体系架构

## Hive数据管理

### 内部表和外部表

### Partition（分区）

### Bucket（分桶）

### 数据类型

### Hive的窗口函数

[link\Hive 之 分析窗口函数 - happy19870612's blog - CSDN博客.mhtml](link/Hive 之 分析窗口函数 - happy19870612's blog - CSDN博客.mhtml)

#### 5.1分析函数定义

分析函数又被称为窗口函数。在联机分析(OLAP)的时候，传统是SQL可能需要进行多个自连接，从而导致查询新能很差。又或者有时候需要多维分析需要在不同层级进行聚合，并且聚合后数据行和密钥聚合的数据行都必须返回。传统是SQL都有一定的局限性。分析函数很好的可以解决这些问题，可以不使用自联结就能在一行中取出聚合和非聚合的行

#### 5.2分析函数组成

function(argument1, argument2…)over([partition-by-cluase][order-by-clasue][windowing-clause])

分析函数有三个部分组成：分区子句，排序子句，以及窗口子句。

分区子句： 其实可以理解为根据什么分组

排序子句：对每一个分区的结果集排序

窗口子句：

[ROWS|RANGE] BETWEEN <开始表达式> AND <结束表达式>

<开始表达式>:[UNBOUNDED PRECEDING| CURRENT ROW|

nRECEDING | n FOLLOWING]

<结束表达式>:[UNBOUNDED FOLLOWING|CURRENT ROW|

nRECEDING | n FOLLOWING]

UNBOUNDEDPRECEDING：针对当前分区里面的当前行而言的前面所有行

UNBOUNDEDFOLLOWING：针对当前分区里面的当前行而言的后面所有行

CURRENTROW： 当前行

nRECEDING： 针对当前行而言，前n行

nFOLLOWING：针对当前行而言，后n行

## hive的优化

# 三 Hadoop1和2的区别

## （一）slot的含义

1）slot就是槽的意思，是一个资源单位，只有给task分配了一个slot之后，这个task才可以运行。slot分两种，map slot沪蓉reduce slot。另外，slot是一个逻辑概念，一个数据节点的slots数量既不是CPU的核数，也不是memory chip，一个节点的slot数量用来表示此节点的资源容量或是计算计算能力的大小，也就是说slot是hadoop的资源单位。

2）系统中map slots总数与reduce slots总数的计算方式如下:

map slots 总数=集群节点数\*mapred.tasktracker.map.tasks.maximum（默认是2）;

reduce slots 总数=集群节点数\*mapred.tasktracker.reduce.tasks.maximum（默认是2）;