# spark checkpoint

## why?

中间结果往往很重要，所以为了保证数据的安全性，要checkpoint

最好把checkpoint到HDFS,这样便于该集群所有节点访问到

在checkpoint之前最好先cache一下，会把数据放到缓存，便于任务调用和checkpoint时可以直接从缓存拿数据

## when?

在发生shuffle之后做checkpoint,因为shuffle太耗时了，结果珍贵

## how?

1. 建立checkpoint存储目录 sc.setCheckpointDir(“hdfs://xxxx”)
2. rdd.cache()
3. rdd.checkpoint()

# spark sql

## DataFrame

dataframe是一种以RDD为基础的分布式数据集，类似于传统数据库中的二维表，与RDD的主要区别在于，前者带有schema信息，即每一列都有名称和类型，这使得Spark SQL 得以洞察更多的结构信息，从而对dataframe背后的数据源以及作用于dataframe之上的变换进行了针对性的优化。

## Spark SQL Core

核心是把已有的RDD，带上schema信息，然后注册程类似sql里的表，对其进行sql操作。主要有两部分，一是生成SchemaRDD,二是执行查询。spark sql查询的优化器可以利用schema信息进行针对性的优化，最后优化过的逻辑执行计划被翻译为物理计划，并最终落实为RDD DAG。

## DataSets vs DataFrames

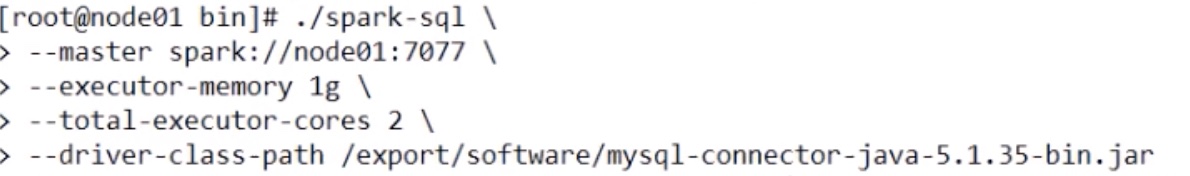
Datasets是数据的分布式集合。Datasets是Spark 1.6中添加的新接口，它具有RDD的优点（强类型输入，使用强大的lambda函数的能力）以及Spark SQL优化的执行引擎的优点。

DataFrame是组织为命名列的Datasets

scala和java中，Dataframe用Row组成的Dataset组成，也就是说是Dataset[Row]类型的别名。

# hive on spark

配置： 把core-site.xml和hive-site.xml放到spark/conf/下。



# spark streaming

## 窗口函数

窗口函数：一段时间内数据发生的变化

两个概念：

窗口长度：窗口的持续时间

窗口间隔： 窗口操作间隔

注意：这两个参数必须是Dstream批次间隔的倍数。

# spark on yarn