Spark安装部署

www.huawei.com

HUAWEI

目标

- 学完本课程后，您将能够:
  - 掌握**Spark**部署规划
  - 熟悉**Spark**常用配置参数说明
  - 掌握**Spark**客户端工具的使用方法

目录

Spark部署原则及约束

服务名称	角色名称	内存最小要求	依赖关系	角色业务部署原则
Spark	SR（SparkResource）	-	依赖于Yarn、HDFS、ZooKeeper、Hive、MapReduce和DBService	无实体进程，不消耗内存。所有数据节点上都要部署，非主备。
	JH（JobHistory）	2GB		分别部署在2个控制节点上，非主备。
	JS（JDBCServer）	2GB		分别部署在2个控制节点上，主备配置。

目录

Spark关键参数配置

运行**Spark**任务的**Executor**参数调整

配置项	说明
SPARK_EXECUTOR_INSTANCES	运行任务启动的**Executor**，默认为**2**
SPARK_EXECUTOR_CORES	每个**Executor**分配的核数，默认为**1**
SPARK_EXECUTOR_MEMORY	每个**Executor**分配的内存，默认为**1000M**

Spark关键参数配置

- **Jobhistory Web UI**允许的最大并发访问数量

配置项	说明
spark.connection.maxRequest	设置客户端访问**Jobhistory**的最大并发数量，默认值为“**5000**”。

- **Jobhistory Web UI**的**Session**超时时间

配置项	说明
spark.session.maxAge	设置会话的超时时间。默认为“**600**秒”。

- 配置**JDBCServer**队列

配置项	说明
SPARK_YARN_QUEUE	用于请求分配**hadoop**队列。

Spark关键参数配置

- 配置事件队列的大小

配置项	说明
spark.evenQueue.size	事件队列的大小，可以根据**driver**的内存做适当的配置，默认为**10000**。

- **Event log**的周期清理

配置项	说明
spark.history.fs.cleaner.enable	是否打开清理功能。**true**为打开，**false**为关闭。
spark.history.fs.cleaner.interval	清理功能的检查周期。默认值为一天。
spark.history.fs.cleaner.maxAge	日志的最长保留时间。默认值为**15**天。

目录

1. Spark部署规划
2. Spark关键参数配置
3. **Spark客户端工具使用示例**

Spark-submit提交spark任务

- 使用示例：

```
./spark-submit --class org.apache.spark.examples.SparkPi --
master yarn-client ${SPARK_CLIENT_HOME}/$}/spark/lib/spark-
examples*.jar
```

- 说明：

--class：**Spark**应用的入口类，应用从此类的**main**函数开始执行。

--jars：业务**jar**包依赖的其它**jar**包通过该参数指定。

--driver-memory：**Driver**进程占用内存大小，在**Driver**（客户业务代码主逻辑）需要大量内存时需要人工设置该参数，否则很容易在**Driver**中遇到**OutOfMemory**异常。

--master：参数为**yarn-client**或**yarn-cluster**模式，若不指定该参数为单机版模式

Spark-shell执行scala语句

- 使用示例：

```
./spark-shell --master yarn-client   //启动spark-shell

val ardd = sc.makeRDD(1 to 10, 20)   //生成一个RDD
ardd.count()                         //统计RDD的数据条数
```

- 说明：

--master：为**yarn-client**或**yarn-cluster**模式，若不指定该参数默认为单机版模式。根据实际业务需要和集群资源情况指定单个执行器所占**CPU**（**--executor-cores**），单个执行器所占内存（**--executor-memory**），以及执行器个数（**--num-executors**）。

Spark-sql执行spark sql语句

- 使用示例：

```
./spark-sql --master yarn-client
show tables;
```

- 说明：

spark-sql是另一种使用**Spark SQL**的工具，该工具不以**JDBC**客户端的形式连接**Spark JDBC Server**，而是申请属于当前应用的资源来进行**Spark SQL**交互式分析。

--master：为**yarn-client**或**yarn-cluster**模式，若不指定该参数默认为单机版模式。

根据实际业务需要和集群资源情况指定单个执行器所占**CPU**（**--executor-cores**），单个执行器所占内存（**--executor-memory**），以及执行器个数（**--num-executors**）。

本章小结

- 本章介绍了**Spark**部署规划涉及的服务进程的部署方法以及资源占用；
- **Spark**关键参数的配置并展示了**Spark**客户端工具的使用方法。

习题

判断题

1. **Spark**修改参数后，需要重新下载安装客户端，使客户端配置生效。
2. **spark.connection.maxRequest**参数可以设置客户端访问**Jobhistory**的最大并发数量，默认值为“**5000**”。

选择题

1. **Spark**依赖于下列哪些服务（ ）

   A. Yarn B. HDFS C.ZooKeeper

   D.Hive E.MapReduce F.DBService

思考

1. 如何提交**spark**任务?
2. 如何检查**JDBCServer**的状态?

Thank you

www.huawei.com