Spark实验：Spark on Yarn安装部署

1、实验描述

- 以Spark on Yarn的运行模式安装Spark集群
- 主要步骤：
  - 解压安装Spark
  - 配置Spark on Yarn
  - 运行Spark on Yarn的测试用例

2、实验环境

- 虚拟机数量：3
- 系统版本：Centos 7.5
- Hadoop版本：Apache Hadoop 2.7.3
- Spark版本：Apache Spark 2.1.1

3、相关技能

- Spark on Yarn安装部署

4、知识点

- 常见linux命令的使用
- 通过修改`.bash_profile`文件配置Spark
- 验证Spark on Yarn安装
- Yarn Resource Manager WebApp的使用

5、实验步骤

1. 解压Spark安装包

   将`master`节点中，路径`/home/zkpk/tgz/spark`下的Spark压缩包，解压至`/home/zkpk/`，并查看解压后的目录中的内容。

2. 配置Hadoop生态组件相关环境变量

   在`master`节点中，编辑路径`/home/zkpk/`下的`.bash_profile`文件，添加Hadoop、HDFS和Yarn的配置文件目录，参考内容如下，然后重新编译使其生效。

   ```
   export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
   export HDFS_CONF_DIR=$HADOOP_HOME/etc/hadoop
   export YARN_CONF_DIR=$HADOOP_HOME/etc/hadoop
   ```

3. 配置`yarn-site.xml`文件

   为了防止在启动Spark on Yarn时出现内存大小错误，导致任务被强制杀死，可以设置取消Yarn运行模式的运行内存检测，在`master`节点中，编辑`/home/zkpk/hadoop-2.7.3/etc/hadoop/`路径下的`yarn-site.xml`文件，添加以下的属性配置：

```
    <property>
        <name>yarn.nodemanager.pmem-check-enabled</name>
        <value>false</value>
    </property>
    <property>
        <name>yarn.nodemanager.vmem-check-enabled</name>
        <value>false</value>
    </property>
```

完成`master`节点的配置后，在`slave01`、`slave02`节点中进行相同的配置，可以通过将`yarn-site.xml`文件远程拷贝来实现。

4. 启动Hadoop集群
    1. 在`master`节点上，运行`/home/zkpk/hadoop-2.7.3/sbin/`路径下的脚本`start-all.sh`。
    2. 在`master`节点上，关闭HDFS的安全模式：

    ```
    hdfs dfsadmin -safemode  leave
    ```

    3. 在`master`节点上利用`jps`命令确认`NameNode`，`SecondaryNameNode`，`ResourceManager`已启动。
    4. 在`slave01`，`slave02`上利用`jps`确认`DataNode`，`NodeManager`已启动。
5. 提交示例应用到Spark
    1. 在`master`节点的`/home/zkpk/spark-2.1.1-bin-hadoop2.7/bin/`路径下，运行`spark-submit`命令，并指定以下选项：
    ***注意：需要以`./spark-submit`的形式键入命令**
        1. 指定应用的主类：`--class org.apache.spark.examples.SparkPi`
        2. 指定Spark运行在Yarn上：`--master yarn`
        3. 指定所需的本地jar包路径：`examples/jars/spark-examples_2.11-2.1.1.jar`
        4. 输入该应用所需参数：`该参数取为同学们的学号后三位 mod 8`
        5. 将应用输出内容重定向到本地文档：`> pi.out`
    2. 键入上述的`spark-submit`命令后，打开`pi.out`文件查看所得结果。
    *** ans ***

    ```
    ./spark-submit --class org.apache.spark.examples.SparkPi --master yarn
    examples/jars/spark-examples_2.11-2.1.1.jar 10
    ```

6. 访问Yarn Resource Manager Webapp

    浏览器中访问http://master:18088，可以查看集群状态，以及提交的应用的信息。