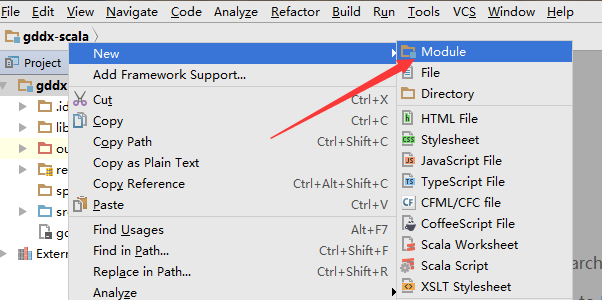
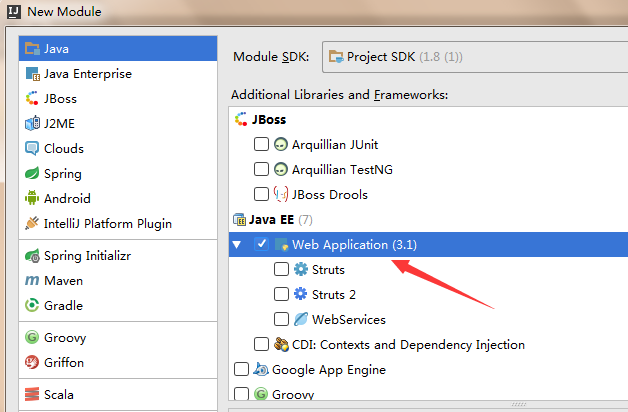
# 课堂笔记

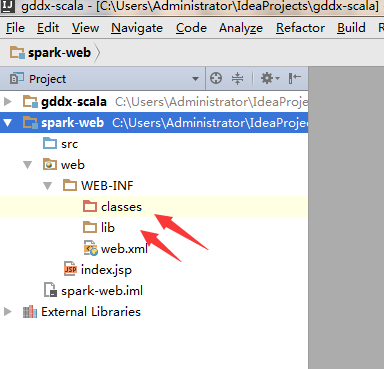
添加Module：



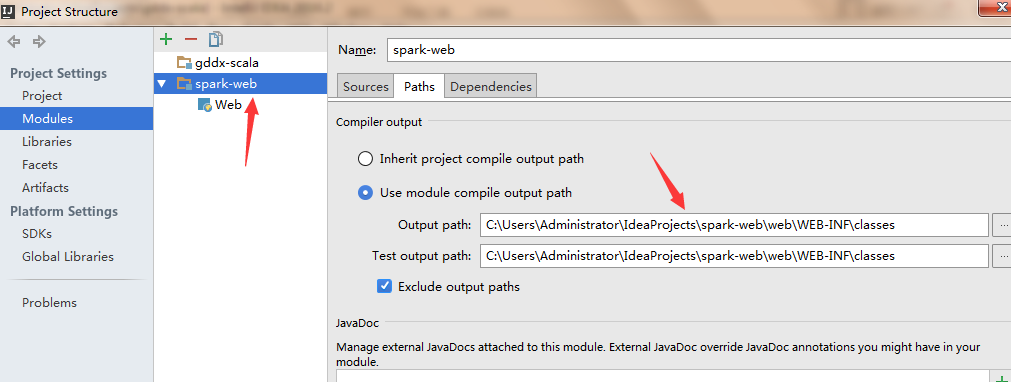
选择Web Application



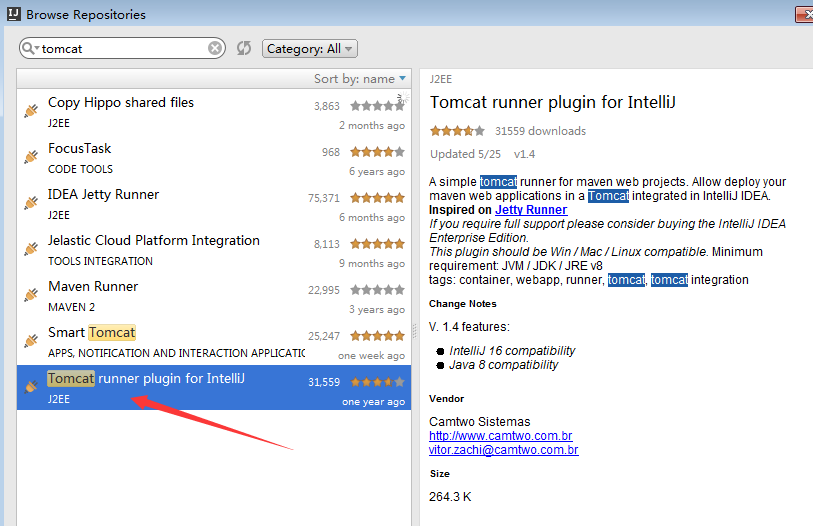
WEB-INF下添加目录：



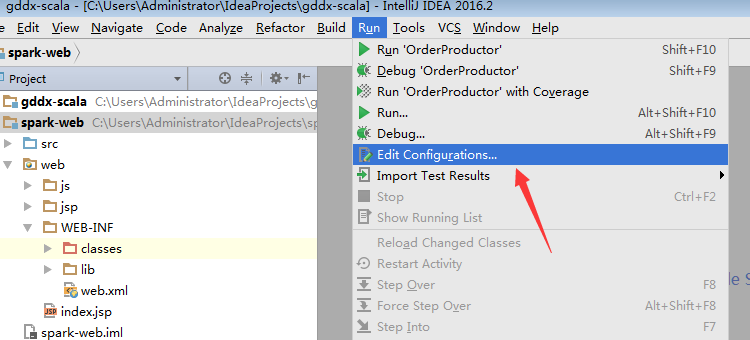
设置out目录：



添加Tomcat Runner：

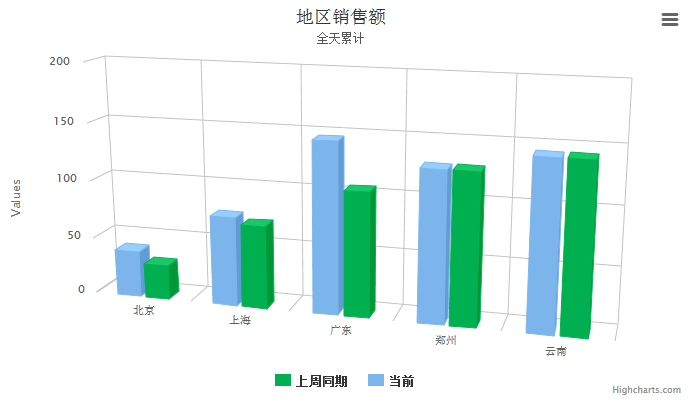


配置tomcat：



## 综合项目需求说明

按天、地区（省份）计算实时销售额，5s延迟，图表展示。

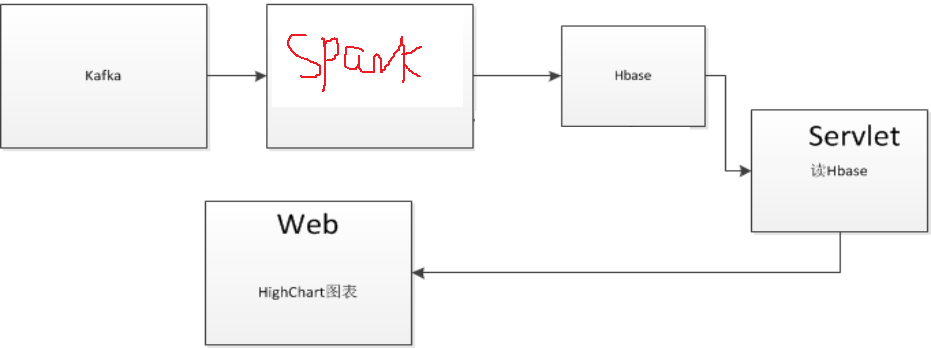


## 架构设计

后台：Spark Steaming进行实时计算

前台：Servlet、Jsp、Highcharts展示

需要做到前后台解耦，互补影响。



优点： 1、前台、后台分离，重启维护互不影响；

2、Spark重启不会影响结果数据，不影响HighCharts图表展示；

3、Tomcat重启不会影响Spark端数据处理；

4、前台后台数据传输通过Hbase，Hbase稳定性强，查询实时性好。

通过HTTP长链接把数据推送到jsp中

学员提问：



答：cdh中Spark最高版本是1.6，所以我们课程用的是apache Spark 2.2

谢香梅 20:47:59

老师，用produce无法把数据写入集群kafka，没有报错，检查看到有建立了topic，用指令在集群上可以生产和消费，这种怎么处理

答：produce无法把数据写入集群kafka，可能是Jar版本对不上，建议把日志级别调整下看更详细日志。

黄秀媛 20:51:27

比如说天猫双十一，滚动显示前几名销售额最大的省份数据

答：这是TopN效果。

黄锡恩 20:58:11

web查清单用thrift去访问hbase，性能行吗？

答：HBase有Rest接口（http）和thrift接口可以查询，也可以用API查询。

性能没问题，因为查询hbase通常都是查小规模数据。

黄锡恩 21:00:57

rest和thrift哪个好使

答：应该差不多。

1、打开area.jsp，触发*init*()调Servlet

2、Servlet里建立http长链接，每隔3s查一次库，并把结果拼成Json格式response.getWriter().write 到jsp的js函数。

3、jsp中执行函数，获得Json结果数据，Set到Serises中。

项目问题：

Spark Steaming作业重启后，发现会从0开始计算并写库。

问：执行Spark Steaming作业即project.AreaAmt时，首先要从库中查出上次最终的结果？不是。

正常情况下会自动从**checkpoint中获取最新计算结果，再累加。**

**解决了！重试后正常了**IMG_256

易爱华 21:44:57

可以讲下spark调优方面的内容吗

**答：1、spark-sql ，主要是sql功底**

1. **spark-submit ，根据作业的数据量合理分配资源**
2. **RDD加载时，利用cache()缓存，利用分区,**

.partitionBy(**new** HashPartitioner(10))

.persist(StorageLevel.*DISK\_ONLY*)

郑博 21:49:21

hbase调优呢？写入慢

答：1、批量写，mutate(List<Put>)

单机测试 8万行/s ,64G，多个硬盘分开挂载

6\*6T 的服务器：

挂载为：data1、data2...

Hdfs的data.dir 设置为data1、data2...

1. 关闭WAL日志