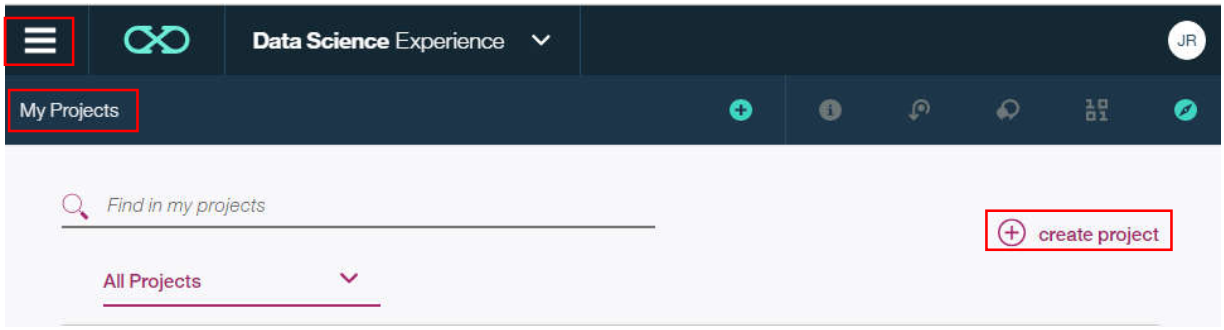
创建 notebook

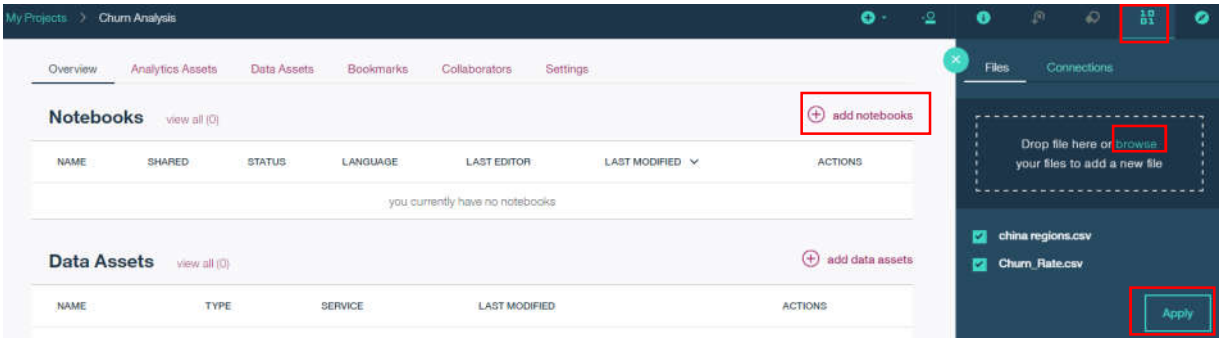
请大家注意，本次活动推荐选用 firefox 浏览器

1. 登录 http://datascience.ibm.com/registration/stepone

2. 创建工程，My Projects -> create project，在 Name 处输入工程名字，点击 Create，稍等片刻



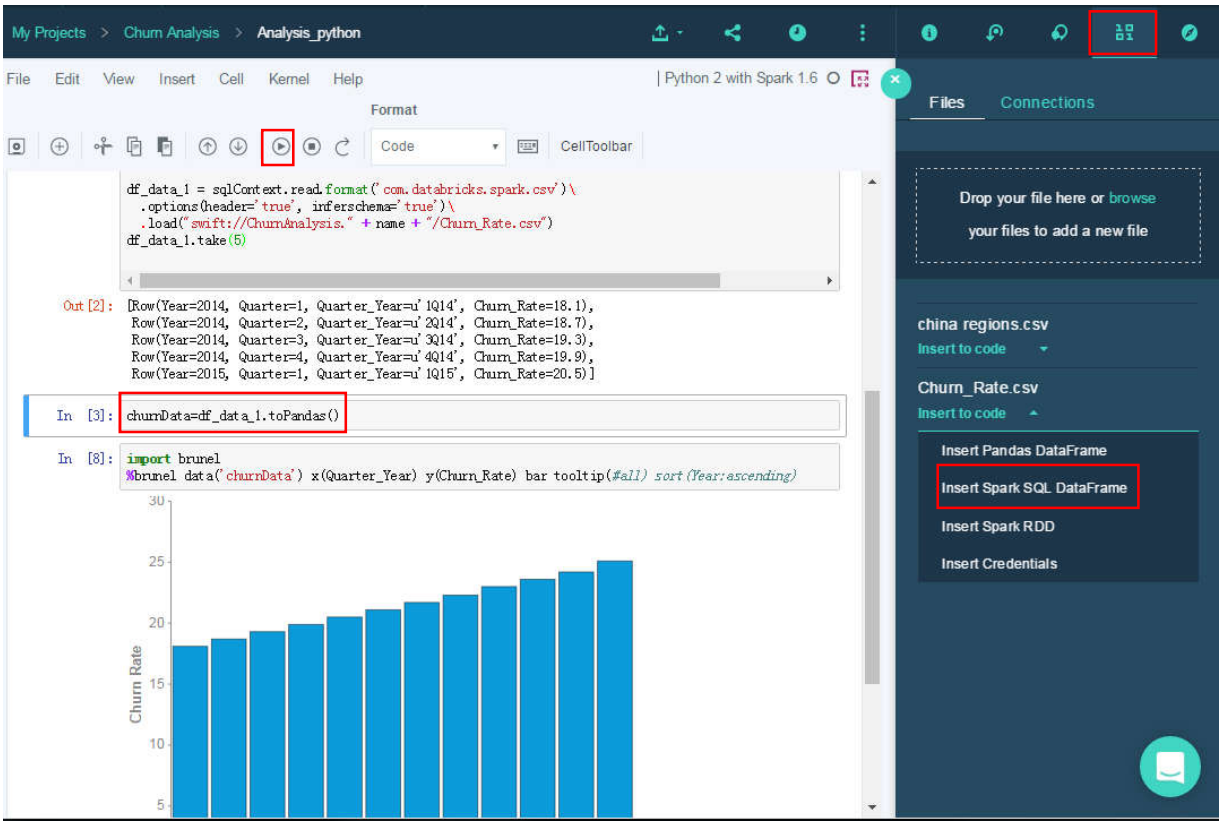
3.添加数据文件，点击右上角 find and add data，把数据文件拖入右方或者点击 browse 添加文件，然后点击 Apply



5.创建 notebook，点击上图 add notebooks, 在 Name 处输入 notebook 名字，选择你需要用的 Spark version 和 Language（如果选择 scala，spark version 选 1.6，如果选 python，spark version 不限），点击 Creat Notebook.

Python 版本

1. 导入数据，在 notebook 右侧点击创建工程时导入的数据文件，选择 Insert Spark SQL DataFrame, 则 notebook 出现下图代码，点击运行 Run cell，输出结果，则生成的 df_data_1 即为 Spark DataFrame 类型数据。



2.如果使用 brunel 分析数据，需要选用 Insert Pandas DataFrame, 或者把 Spark DataFrame 类型数据用 toPandas 方法转成 Pandas DataFrame 类型

3.导入 brunel, 输入 Import brunel.

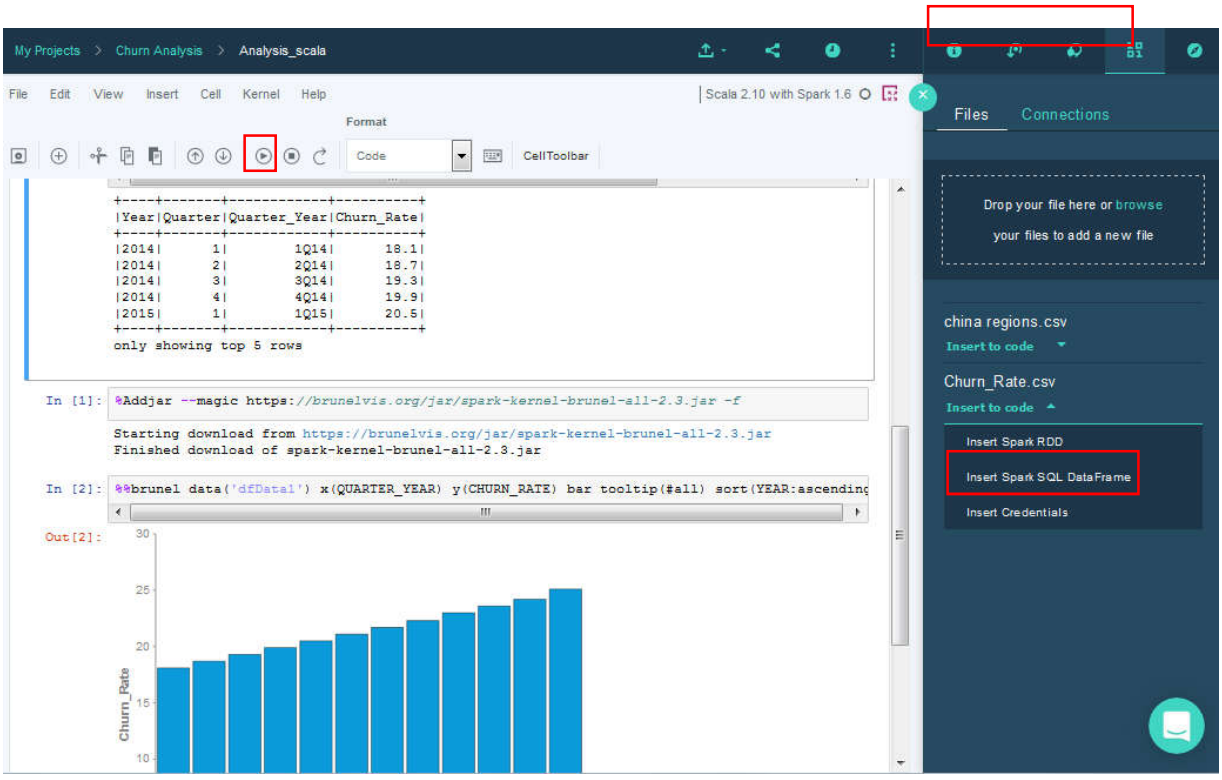
4.用 brunel 画图，例：%brunel data('dfData1') x(QUARTER_YEAR) y(CHURN_RATE) bar tooltip(#all) sort(YEAR:ascending)

详情参考：Visualizationwith open source package brunel
http://datascience.ibm.com/blog/brunel-interactive-visualization-in-jupyter-notebooks-2/

Scala 版本

1. 创建工程时注意选择 spark1.6 版本，需选用 firefox 等其他浏览器。如果后续 chrome 看不到图像，可在 chrom 右键属性，在 target 处设置允许不安全内容，参考 https://superuser.com/questions/487748/how-to-allow-chrome-browser-to-load-insecure-content。

2. 导入数据，在 notebook 右侧点击创建工程时导入的数据文件，选择 Insert Spark SQL DataFrame，则 notebook 出现下图代码，点击运行 Run cell，输出结果，则生成的 dfData1 即为 Spark DataFrame 类型数据。



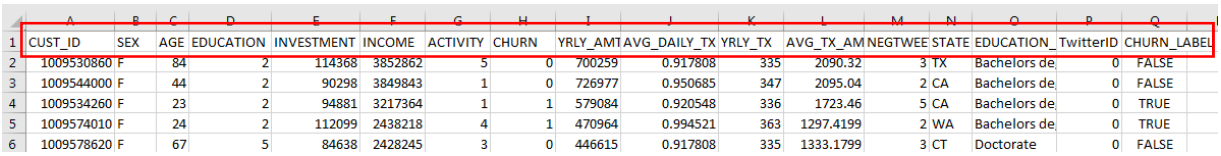
3.导入 brunel, 输入%Addjar --magic https://brunelvis.org/jar/spark-kernel-brunel-all-2.3.jar -f, 运行

4.用 brunel 画图，例：%%brunel data('dfData1') x(QUARTER_YEAR) y(CHURN_RATE) bar tooltip(#all) sort(YEAR:ascending)

详情参考：Visualizationwith open source package brunel

http://datascience.ibm.com/blog/brunel-interactive-visualization-in-jupyter-notebooks-2/

数据文件 CUST_SUM.csv 的数据结构说明



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	CUST_ID	SEX	AGE	EDUCATION	INVESTMENT	INCOME	ACTIVITY	CHURN	YRLY_AMT	AVG_DAILY_TX	YRLY_TX	AVG_TX_AM	NEGTWEE	STATE	EDUCATION_	TwitterID	CHURN_LABEL
2	1009530860	F	84	2	114368	3852862	5	0	700259	0.917808	335	2090.32	3	TX	Bachelors de	0	FALSE
3	1009544000	F	44	2	90298	3849843	1	0	726977	0.950685	347	2095.04	2	CA	Bachelors de	0	FALSE
4	1009534260	F	23	2	94881	3217364	1	1	579084	0.920548	336	1723.46	5	CA	Bachelors de	0	TRUE
5	1009574010	F	24	2	112099	2438218	4	1	470964	0.994521	363	1297.4199	2	WA	Bachelors de	0	TRUE
6	1009578620	F	67	5	84638	2428245	3	0	446615	0.917808	335	1333.1799	3	CT	Doctorate	0	FALSE

从左往右数，每列内容依次为：

信用卡用户 ID、性别（string）、年龄、教育程度（Int）、一年内投资的费用、收入、社区活动、是否流失（Int）、累计交易额、日均交易笔数、一年总交易笔数、平均每笔交易金额、在社交平台上对银行所做的评价量、所在地区(string)、教育程度(string）、Twitter 账户、是否流失（bool）

注：教育程度分 Int 和 string，是同样的意思。是否流失分 bool 和 Int 类型，也是同样的意思。