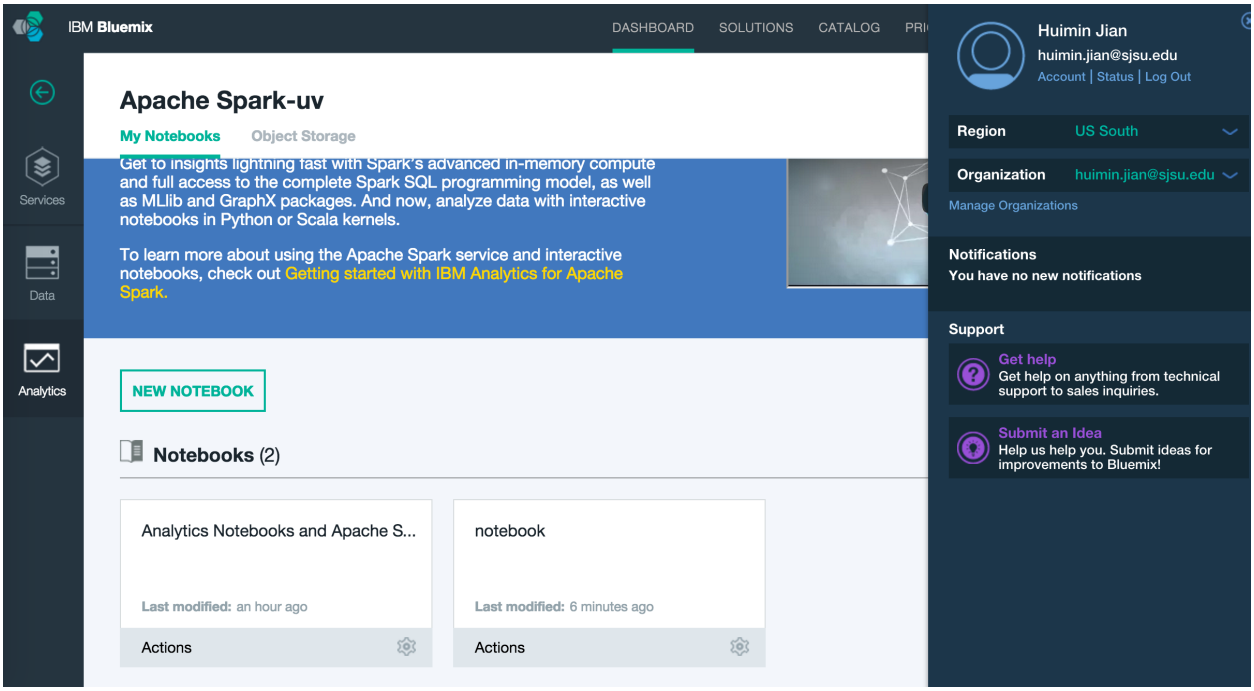
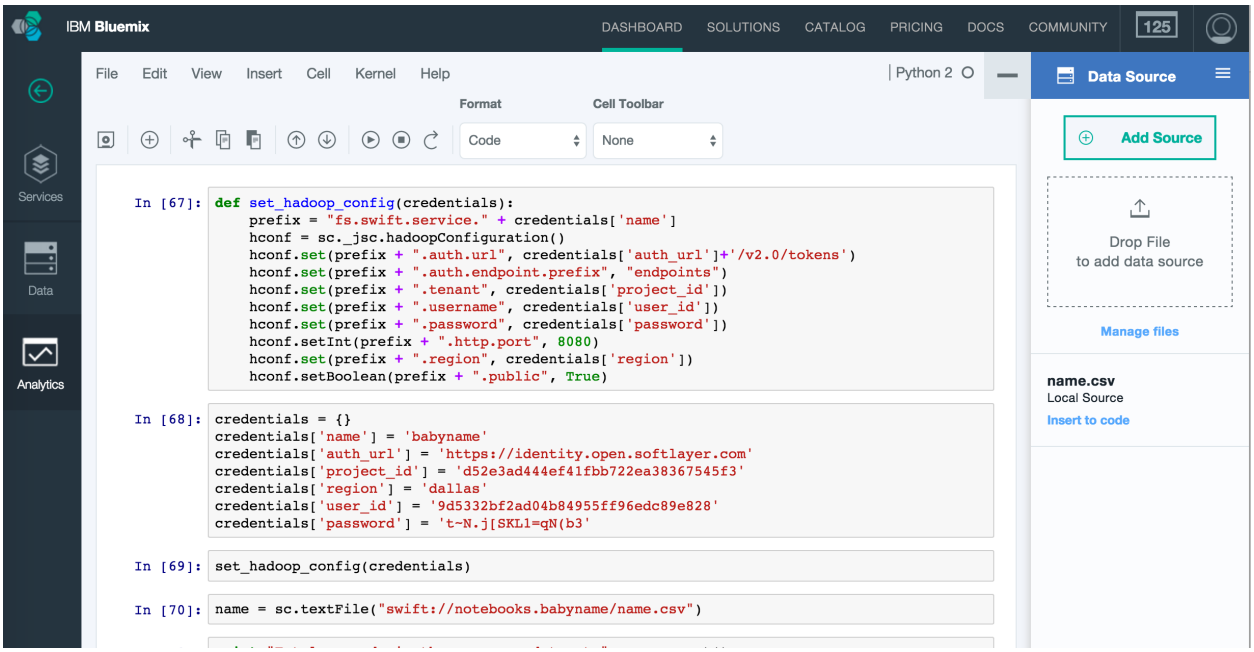
My assignment is use popular baby name datasets from https://www.ssa.gov/oact/babynames/rankchange.html and Spark to predict the most popular baby name in 2016.

The following are my assignment progress:

1. Create a Spark service instance “Apache Spark-uv” and a notebook “notebook” under this instance.



2. Download datasets as name.csv file and import it into my notebook.



3. Parse the imported data.

```
In [74]: nameParse = name.map(lambda line : line.split(","))

In [75]: nameParse.first()
Out[75]: [u'NAME', u'2014', u'2013', u'GENDER']

In [76]: nameParse.first()[0]
Out[76]: u'NAME'

In [77]: nameParse.first()[2]
Out[77]: u'2013'
```

4. Separate the dataset to show only girls name and boys name.

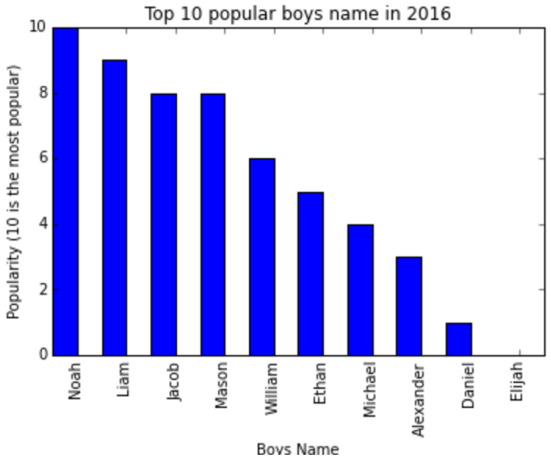
```
In [78]: boyname = nameParse.filter(lambda x: x[3] == "M")

In [79]: boyname.first()
Out[79]: [u'Bode', u'783', u'1428', u'M']

In [80]: girlname = nameParse.filter(lambda x: x[3] == "W")

In [81]: girlname.first()
Out[81]: [u'Aranza', u'607', u'4232', u'W']
```

5. Write the prediction function for boys name.

```
In [82]: boypredict = boyname.map(lambda p: (p[0], int((int(p[1]) + int(p[2])) / 2)) )

In [83]: boypredict.first()
Out[83]: (u'Bode', 1105)

In [84]: ppTop10=[]
         nameTop10=[]
         for pair in boypredict.map(lambda (x,y) : (y,x)).takeOrdered(10):
             ppTop10.append(pair[0])
             nameTop10.append(pair[1])
             print "Boyname %s has popularities of %f in 2016" % (pair[1],pair[0])

         Boyname Noah has popularities of 1.000000 in 2016
         Boyname Liam has popularities of 2.000000 in 2016
         Boyname Jacob has popularities of 3.000000 in 2016
         Boyname Mason has popularities of 3.000000 in 2016
         Boyname William has popularities of 5.000000 in 2016
         Boyname Ethan has popularities of 6.000000 in 2016
         Boyname Michael has popularities of 7.000000 in 2016
         Boyname Alexander has popularities of 8.000000 in 2016
         Boyname Daniel has popularities of 10.000000 in 2016
         Boyname Elijah has popularities of 11.000000 in 2016
```

6. Draw the prediction result for top 10 popular boys name.

```
index = np.arange(N)
bar_width = 0.5

plt.bar(index, drawppTop10, bar_width,
                 color='b')
plt.xlabel('Boys Name')
plt.ylabel('Popularity (10 is the most popular)')
plt.title('Top 10 popular boys name in 2016')
plt.xticks(index + bar_width, nameTop10, rotation=90)
plt.show()
```

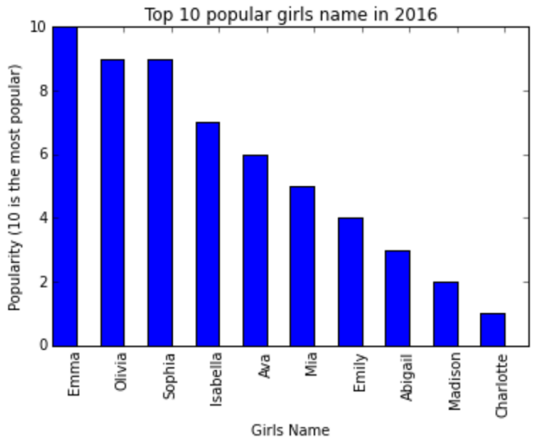


7. Write the prediction function for girls name.

```
In [87]: girlpredict = girlname.map(lambda p: (p[0], int((int(p[1]) + int(p[2])) / 2)) )

In [88]: girlpredict.first()

Out[88]: (u'Aranza', 2419)

In [89]: gppTop10=[]
         gnameTop10=[]
         for pair in girlpredict.map(lambda (x,y) : (y,x)).takeOrdered(10):
             gppTop10.append(pair[0])
             gnameTop10.append(pair[1])
             print "Girlname %s has popularities of %f in 2016" % (pair[1],pair[0])

Girlname Emma has popularities of 1.000000 in 2016
Girlname Olivia has popularities of 2.000000 in 2016
Girlname Sophia has popularities of 2.000000 in 2016
Girlname Isabella has popularities of 4.000000 in 2016
Girlname Ava has popularities of 5.000000 in 2016
Girlname Mia has popularities of 6.000000 in 2016
Girlname Emily has popularities of 7.000000 in 2016
Girlname Abigail has popularities of 8.000000 in 2016
Girlname Madison has popularities of 9.000000 in 2016
Girlname Charlotte has popularities of 10.000000 in 2016
```

8. Draw the prediction result for top 10 popular girls name.

```
plt.bar(index, drawgppTop10, bar_width,
                 color='b')
plt.xlabel('Girls Name')
plt.ylabel('Popularity (10 is the most popular)')
plt.title('Top 10 popular girls name in 2016')
plt.xticks(index + bar_width, gnameTop10, rotation=90)
plt.show()
```



In [ ]:

9. Conclusion: according to my Spark prediction, the most popular baby name in 2016 is Emma for girls and Noah for boys.