Project 3 : Spatiotemporal Analysis with Spark

- A report by Adarsh and Ayush

1. **Climate Chart** : We first read the data from the sample files on Orion, into a dataframe. We are using predefined python functions to compute geohash based on latitude and longitude as input parameters. We are converting the time into the month of the year. We are then reading features such as temperature and precipitation.
   The below image shows 12 month data, for a given geohash ‘dnd’.

	1_time	max(temperature_surface)	min(temperature_surface)	avg(total_precipitation_surface_3_hour_accumulation)	avg(temperature_surface)
0	1430341200000	296.62330	293.87330	0.000000	294.977467
1	1445396400000	280.52850	278.65350	0.000000	279.307346
2	1447038000000	282.31354	274.06354	0.000000	276.626040
3	1443236400000	290.90260	289.52760	1.442308	290.075677
4	1442026800000	291.46582	286.59082	0.005682	290.227184
5	1442847600000	298.00000	296.00000	0.000000	297.178571
6	1431874800000	302.32275	298.07275	0.116071	300.215607
7	1430924400000	301.92456	295.92456	0.000000	299.538196
8	1444467600000	287.68700	283.06200	0.237500	285.387000
9	1427749200000	292.63990	290.63990	0.000000	291.814900
10	1437987600000	295.24365	293.49365	0.041667	294.843650
11	1433257200000	294.03955	290.16455	0.062500	291.557407
12	1444402800000	299.62524	294.62524	0.125000	298.316416

2. **Travel Startup** : We are calculating the comfort Index based on three features from the dataset. They are temperature_surface, relative_humidity and pressure_surface. We have considered ideal temperature range to be (294 - 304) , ideal humidity range (5 - 25) and ideal pressure range (101000 - 102000). To calculate the comfort index we are considering, how much the values deviate from the means of these ranges and then computing the average of those deviations. A lower value of comfort index means more comfortable in our case.

Initially, we filter out the data which doesn't lie in these ranges. Then, from the remaining, we compute the top 5 locations, according to the best comfort index, which we recommend as a part of our travel startup.
Below image, shows the top 5 geohashes and the time of the year, we got on running our program.

	5_hash	1_time	temperature_surface	relative_humidity_zerodegc_isotherm	pressure_surface	c_idx
0	d7b	01	298.348927	16.909580	101501.593649	1.2840521994996514
1	9vp	10	298.301239	23.413302	101478.392518	10.107340817496015
2	95x	11	298.109089	19.937802	101474.337802	10.23636284003633
3	d6v	03	300.360270	5.042528	101518.578263	10.298668363849922
4	9s0	02	296.744381	18.621229	101526.536313	10.471053765362091

3. **Solar Wind** : We have found the locations of solar farms based on the feature temperature_surface and wind farms based on pressure_maximum_wind. The ideal temperature we have considered lies between 308 and 338. The ideal pressure_maximum_wind we have considered is more than 20000.

   The top 3 locations for solar farms, based on temperature_surface are as below :

	5_hash	1_time	pressure_maximum_wind	temperature_surface
0	9tbq	07	21111.943557	311.928496
1	9tbm	07	21408.059874	311.491012
2	9tbq	08	20832.068599	311.478068

   The top 3 locations for wind farms, based on wind pressure are as below :

	5_hash	1_time	pressure_maximum_wind	temperature_surface
0	f6b6	05	33926.435958	270.893060
1	cdyh	04	33784.938600	256.736114
2	f4fu	04	33710.226613	258.950405

The top 3 locations for solar and wind farms, combining both features are as below :

	5_hash	1_time	pressure_maximum_wind	temperature_surface
0	9se5	06	24738.138428	308.946569
1	9sdu	06	24571.477879	308.757447
2	9se3	06	24040.346235	308.524853

4. **Climate Change** : We calculated the geohashes from latitude and longitude values of the dataset. Based on the geohashes, we calculated the average temperature over the period of 5 years.

```
wide_fmt.limit(5).toPandas()
```

	5_hash	2014	2015	2016	2017	2018	2019
0	f2	273.6225	310.83887	null	null	312.14615	310.55997
1	c0	285.7475	317.34937	null	null	311.72736	309.67
2	f6	254.6225	305.88745	null	null	304.06995	306.63
3	cc	267.8725	312.84448	null	null	317.43997	315.60962
4	bc	281.9975	290.5183	null	null	290.585	291.3078

We then found, in which of these cases, is the temperature increasing.

	5_hash	2014	2015	2016	2017	2018	2019	is_increasing
0	bc	281.9975	290.5183	null	null	290.585	291.3078	true
1	9u	297.9975	323.3811	null	null	325.96997	326.4664	true
2	9x	274.3725	319.81274	null	null	321.16968	321.41	true
3	9s	299.1225	328.09082	null	null	331.566	331.61	true
4	9e	301.9975	326.83325	null	null	328.0671	328.85913	true
5	9p	288.6225	318.38745	null	null	318.41995	318.43	true
6	dp	278.2475	312.4287	null	null	314.58997	316.2473	true
7	bf	281.8725	304.7229	null	null	305.4071	308.69943	true
8	f0	277.3725	310.5537	null	null	312.83	314.73734	true
9	8y	290.9975	296.68628	null	null	296.84616	298.83997	true

Then, from these cases, where temperature is increasing, we found the correlation with humidity.

	5_hash	1_time	relative_humidity_zerodegc_isotherm	temperature_surface
0	8y	2019	44.216567	298.83997
1	9e	2015	50.282616	326.83325
2	bf	2015	77.759759	304.7229
3	9u	2015	41.340677	323.3811
4	9u	2017	35.508914	null
5	8y	2017	47.083304	null
6	9e	2014	21.146939	301.9975
7	9s	2017	39.180013	null
8	9s	2018	43.306710	331.566
9	bc	2014	74.665127	281.9975
10	9p	2019	51.262530	318.43

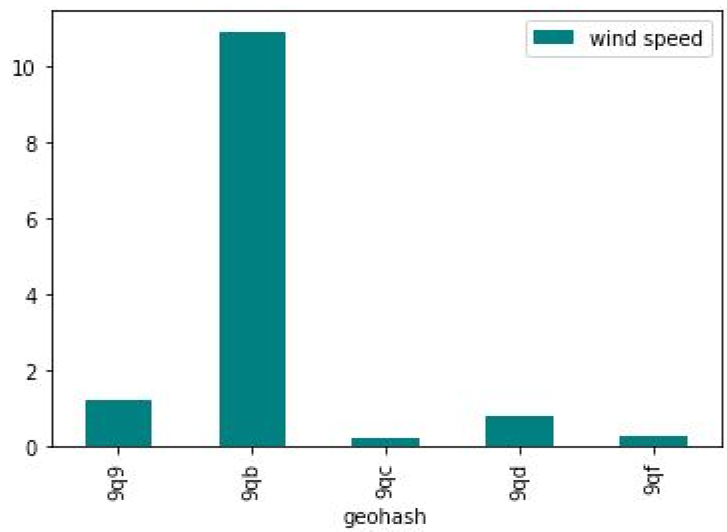
```
+-------+-----------------------+
|geohash|temp_humidity_correlation|
+-------+-----------------------+
|     8y|    -0.46113040974179764|
|     9e|     0.01401291902598702 4|
|     9p|     -0.4509179772252986|
|     9s|      0.4328976712939623|
|     9u|     0.03709686305438359|
|     9x|      0.2750060043533481|
|     bc|     -0.7771805988932008|
|     bf|     0.14249577226255644|
|     dp|     0.31984315570757504|
|     f0|     0.06425264460115487|
+-------+-----------------------+
```

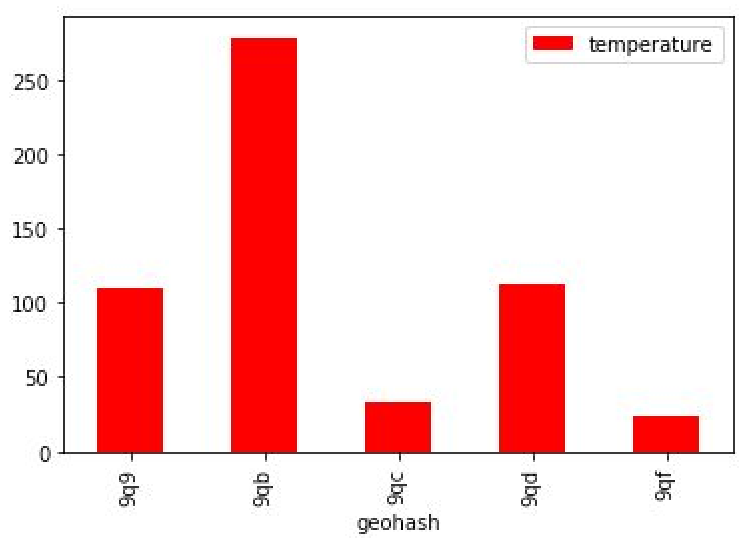
5. **Weather Station** : We made a load_file.py code for streaming the data from the files. The data was received in form of rdd and calculating the mean with the new values coming in. Following shows the similar data for temperature.

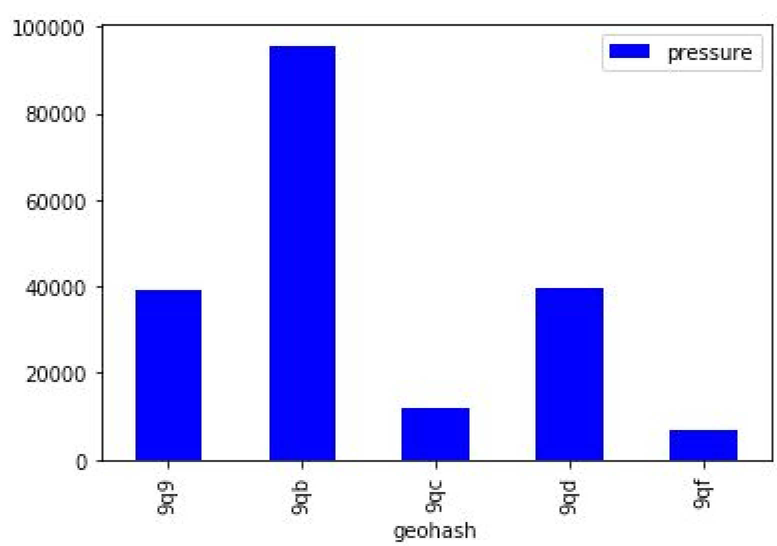
```
-------------------------------------------
{'geohash': '9qd', 'temperature': 107.60127945767196, 'M2': 85847.90365122746}
{'geohash': '9qf', 'temperature': 15.995929740441708, 'M2': 69596.57696703213}
{'geohash': '9q9', 'temperature': 104.92814299903847, 'M2': 84828.38001964623}
{'geohash': '9qb', 'temperature': 278.1117893055556, 'M2': 15.046257003863454}
{'geohash': '9qc', 'temperature': 25.399975841081275, 'M2': 70967.46500002636}

-------------------------------------------
Time: 2020-12-09 23:08:12
-------------------------------------------
{'geohash': '9qd', 'temperature': 112.7402131822264, 'M2': 122978.45124568781}
{'geohash': '9qf', 'temperature': 23.26814274374341, 'M2': 147972.26843974067}
{'geohash': '9q9', 'temperature': 110.24001105182127, 'M2': 124499.99476715154}
{'geohash': '9qb', 'temperature': 278.5114676207086, 'M2': 227.8236074677496}
{'geohash': '9qc', 'temperature': 33.776791623791354, 'M2': 149699.3748616972}
```

We have also uploaded a video of our weather station in action on github. We have shown bar graphs for each feature and each geohash.


wind speed
0
2
4
6
8
10
9q9
9qb
9qc
9qd
9qf
geohash


temperature
0
50
100
150
200
250
9q9
9qb
9qc
9qd
9qf
geohash


pressure
0
20000
40000
60000
80000
100000
9q9
9qb
9qc
9qd
9qf
geohash

6. **Anomaly Detector** : We are considering a feature to be anomaly, if the new value exceeds 120% of the previous mean value. The previous mean value is considered for the previous 10 values. We are using a function called deviate checker for this. We are using the streaming data, same as the previous task.

```
def deviateChecker(mean, new):
    '''
    true means anomaly, i.e more than 120% of prev value
    '''
    return mean*1.2 < new
```

```
def computeVal(new, old):
    for i in range(6):
        if len(old[i][0]) == 10:
            mean = calculateMean(old[i][0])
            is_anomaly = deviateChecker(mean, new[i][0][0])
            if not is_anomaly:
                old[i][0].pop(0)
                old[i][0].append(new[i][0][0])
            else:
                old[i][1] = True
        else:
            old[i][0].append(new[i][0][0])

    return old
```

We are then showing, which of the data points in the streaming data are Anomaly and Not Anomaly and showing it by printing it.

The idea for taking the previous 10 values is to take into consideration, that the features could change eventually but there shouldn't be a sudden change.

```
Time: 2020-12-10 00:09:14
------------------------------------------
['9hs | ', 'surface_temp :  Not Anomaly ', 'pressure :  Not Anomaly ', 'humidity :  Anomaly ', 'precipitation :  Anomaly ', 'v
isibility :  Not Anomaly ', 'wind_speed :  Anomaly ']
['f0q | ', 'surface_temp :  Not Anomaly ', 'pressure :  Not Anomaly ', 'humidity :  Anomaly ', 'precipitation :  Not Anomaly
', 'visibility :  Not Anomaly ', 'wind_speed :  Anomaly ']
['c10 | ', 'surface_temp :  Not Anomaly ', 'pressure :  Not Anomaly ', 'humidity :  Anomaly ', 'precipitation :  Anomaly ', 'v
isibility :  Not Anomaly ', 'wind_speed :  Anomaly ']
['cbh | ', 'surface_temp :  Not Anomaly ', 'pressure :  Not Anomaly ', 'humidity :  Anomaly ', 'precipitation :  Not Anomaly
', 'visibility :  Anomaly ', 'wind_speed :  Anomaly ']
['dqt | ', 'surface_temp :  Not Anomaly ', 'pressure :  Not Anomaly ', 'humidity :  Anomaly ', 'precipitation :  Anomaly ', 'v
isibility :  Anomaly ', 'wind_speed :  Anomaly ']
['9vh | ', 'surface_temp :  Not Anomaly ', 'pressure :  Not Anomaly ', 'humidity :  Anomaly ', 'precipitation :  Anomaly ', 'v
isibility :  Anomaly ', 'wind_speed :  Anomaly ']
['bc9 | ', 'surface_temp :  Not Anomaly ', 'pressure :  Not Anomaly ', 'humidity :  Anomaly ', 'precipitation :  Anomaly ', 'v
isibility :  Anomaly ', 'wind_speed :  Anomaly ']
['f8j | ', 'surface_temp :  Not Anomaly ', 'pressure :  Not Anomaly ', 'humidity :  Anomaly ', 'precipitation :  Not Anomaly
', 'visibility :  Anomaly ', 'wind_speed :  Anomaly ']
['dp2 | ', 'surface_temp :  Not Anomaly ', 'pressure :  Not Anomaly ', 'humidity :  Anomaly ', 'precipitation :  Anomaly ', 'v
isibility :  Anomaly ', 'wind_speed :  Anomaly ']
['cb8 | ', 'surface_temp :  Not Anomaly ', 'pressure :  Not Anomaly ', 'humidity :  Anomaly ', 'precipitation :  Anomaly ', 'v
isibility :  Not Anomaly ', 'wind_speed :  Anomaly ']
...
```

7. **Prediction (Travel Startup)** : We are considering our Travel Startup task, for ML analysis and predicting the comfort index based on past statistics, using linear regression.

   We first calculate the comfort Index and geohashes, same as the previous task, in part 2.

	5_hash	1_time	temperature_surface	relative_humidity_zerodegc_isotherm	pressure_surface	label
0	9kts	12	293.521624	20.829932	101738.396786	82.901695
1	d5de	12	301.269737	24.305921	101600.328526	37.634727
2	95ys	12	295.523142	20.266667	101838.426671	115.390068
3	9k94	12	293.576559	19.700000	101964.923307	158.015579
4	9krm	12	295.818871	18.223837	101607.346228	37.583733

   We are considering the data from 2014 - 18 as the training data and the 2019 data as test data to check our prediction.

temperature_surface	relative_humidity_zerodegc_isotherm	pressure_surface	label	features	prediction
296.22063469453377	16.177419354838708	101923.21336977495	142.05672	[296.220634694533...	111.92702392093997
296.71955547101464	12.384057971014492	101853.76842753626	119.2216	[296.719555471014...	94.04187596494194
298.46329262135924	24.41747572815534	101505.2668867314	5.049218	[298.463292621359...	58.459319863150085
297.64160194346294	21.5354609929078	101599.75398939928	35.549282	[297.641601943462...	69.82517927896333
297.1748498310811	22.408783783783782	101705.24762500002	71.16052	[297.174849831081...	87.93697620335297
296.8901088850174	21.05944055944056	101500.0001358885	2.3898225	[296.890108885017...	56.98801938623001
299.54940818584066	18.75221238938053	101303.93688053089	67.12158	[299.549408185840...	16.61905971123997
297.9721608945686	15.514376996805112	101483.77256230034	5.5898848	[297.972160894568...	42.040685030642635
294.8797725	14.097014925373134	101850.22715298507	118.08346	[294.8797725,14.0...	102.10367778039472
294.9218261165048	22.915857605177994	101920.59958899676	143.86455	[294.921826116504...	126.8220729473287
298.1120823715414	24.968379446640316	101533.08573913043	14.388734	[298.112082371541...	64.47879812701831
298.2805084166667	20.283333333333335	101386.69748749999	39.62212	[298.280508416666...	35.04321678540873
295.5231421904762	20.266666666666666	101838.42667142859	115.39007	[295.523142190476...	108.69221351056149
298.87458547263685	23.587064676616915	101347.05915422887	54.134167	[298.874585472636...	32.99339826902178
296.353952807571	18.312302839116718	101644.40405047315	49.787468	[296.353952807571...	74.88649550474292
296.19148842592597	24.04320987654321	101269.21573148147	80.54533	[296.191488425925...	30.869796238062918
301.05545801444043	23.72202166064982	101357.75485198559	51.340878	[301.055458014440...	27.98460745482589
295.681796935484	16.496774193548386	101980.63557096773	161.48352	[295.681796935484...	122.41019656068966
294.4526559405941	23.495049504950494	101995.8281617161 7	169.29019	[294.452655940594...	140.08862909830714
296.71685026548687	13.189427312775331	101562.35124336283	21.81499	[296.716850265486...	53.38966538167733

only showing top 20 rows

We are considering the output of prediction as the new comfort indexes for these locations. Based on comfort Index, we are better able to suggest the travel locations, thus improving our travel startup.

The lower the comfort index, the better place it would be for travel.

8. **Final Project Update** : We are starting with our analysis on stackoverflow data for our final project, which we complete later using Google cloud platform.

   Starting with the Users.xml file from our dataset, we try to read the xml files and convert them into rdd. We first filter our data and then extract userid and usernames from the data.

```
+---+----------------+
|id |username        |
+---+----------------+
|1  |Community       |
|2  |Geoff Dalgas    |
|3  |Jarrod Dixon    |
|4  |txwikinger      |
|5  |Nathan Osman    |
|6  |Emmett          |
|7  |Helix           |
|8  |mechanical_meat |
|9  |Andrew          |
|10 |DLH             |
|11 |hannes.koller   |
|12 |Michael Terry   |
|13 |Keith Maurino   |
|14 |Jweede          |
|16 |Jeremy L        |
|17 |tutuca          |
|18 |excid3          |
|20 |ParanoiaPuppy   |
|21 |GeoD            |
|22 |Alan Featherston|
+---+----------------+
only showing top 20 rows
```

We are also extracting the data related to postId and the text based on the comments.xml file.

```
+------+-----+--------------------+--------------------+------+
|postId|score|                text|        creationDate|userId|
+------+-----+--------------------+--------------------+------+
|    23|    0|Using /opt helps ...|2010-07-28T19:36:...|    10|
|    18|    0|but popping in a ...|2010-07-28T19:38:...|    10|
|    27|    0|That will revert ...|2010-07-28T19:39:...|    50|
|    31|    0|I think you meant...|2010-07-28T19:41:...|    12|
|    18|    0|@DLH apparently n...|2010-07-28T19:41:...|    63|
|    12|    2|"ssh -X <server> ...|2010-07-28T19:46:...|    96|
|    12|    0|@Suppressingfire:...|2010-07-28T19:48:...|    10|
|    50|    0|Can you please re...|2010-07-28T19:48:...|    56|
|    27|    0|It probably shoul...|2010-07-28T19:49:...|     5|
|    58|    0|Do you mean the c...|2010-07-28T19:50:...|     5|
|    47|    0|Have you checked ...|2010-07-28T19:50:...|     4|
|    47|    1|Might be related ...|2010-07-28T19:51:...|   104|
|    58|    0|Do you use Gnome ...|2010-07-28T19:51:...|     4|
|    60|    0|This causes data ...|2010-07-28T19:52:...|    66|
|    18|    0|no the live CD do...|2010-07-28T19:53:...|     4|
|    52|    0|Does this let the...|2010-07-28T19:55:...|    35|
|    56|    2|LDAP and nfs are ...|2010-07-28T19:56:...|     4|
|    10|    0|Can I use it on a...|2010-07-28T19:56:...|    27|
|    70|    1|That's a good tip...|2010-07-28T19:56:...|    45|
|    70|    0|That is probably ...|2010-07-28T19:58:...|    86|
+------+-----+--------------------+--------------------+------+
only showing top 20 rows
```

We are also doing some analysis combining these two, using joins, which will be helpful in the final project.

id	username	postId	score	text	creationDate	userId
964	Hendrik Brummermann	4602	0	I can confirm thi...	2010-10-13T21:37:...	964
964	Hendrik Brummermann	118087	0	They took it in d...	2012-04-28T06:17:...	964
964	Hendrik Brummermann	638027	0	I have the same i...	2015-08-03T13:26:...	964
1677	eslambasha	84949	0	@fossfreedom i do...	2011-12-03T21:56:...	1677
1697	Frxstrem	16683	0	@Marco, I know, I...	2010-12-08T22:36:...	1697
1697	Frxstrem	16784	0	This seems to be ...	2010-12-09T19:05:...	1697
1697	Frxstrem	16886	1	I only want to di...	2010-12-10T22:26:...	1697
1697	Frxstrem	16892	1	This is not an ac...	2010-12-10T22:28:...	1697
1697	Frxstrem	16988	0	Have you tried bu...	2010-12-11T19:22:...	1697
1697	Frxstrem	17471	0	@Stefano fixed it	2010-12-14T23:14:...	1697
1697	Frxstrem	17892	0	My guess is that ...	2010-12-17T13:50:...	1697
1697	Frxstrem	18014	0	-1 It's too uncle...	2010-12-18T17:53:...	1697
1697	Frxstrem	18273	0	You did replace `...	2010-12-22T17:48:...	1697
1697	Frxstrem	67121	0	Firstly, I have a...	2011-10-15T22:18:...	1697
1697	Frxstrem	108944	0	You should use `t...	2012-03-01T00:30:...	1697
1697	Frxstrem	453415	2	Daily builds can ...	2014-04-23T07:29:...	1697
1697	Frxstrem	223442	0	@user2662639 Simp...	2015-08-26T16:36:...	1697
1697	Frxstrem	223442	0	@user2662639 (I t...	2015-08-26T16:37:...	1697
1697	Frxstrem	17650	2	@Fiksdal I don't ...	2016-03-25T12:21:...	1697
1697	Frxstrem	899129	0	@DavidFoerster Th...	2017-04-01T13:36:...	1697

only showing top 20 rows