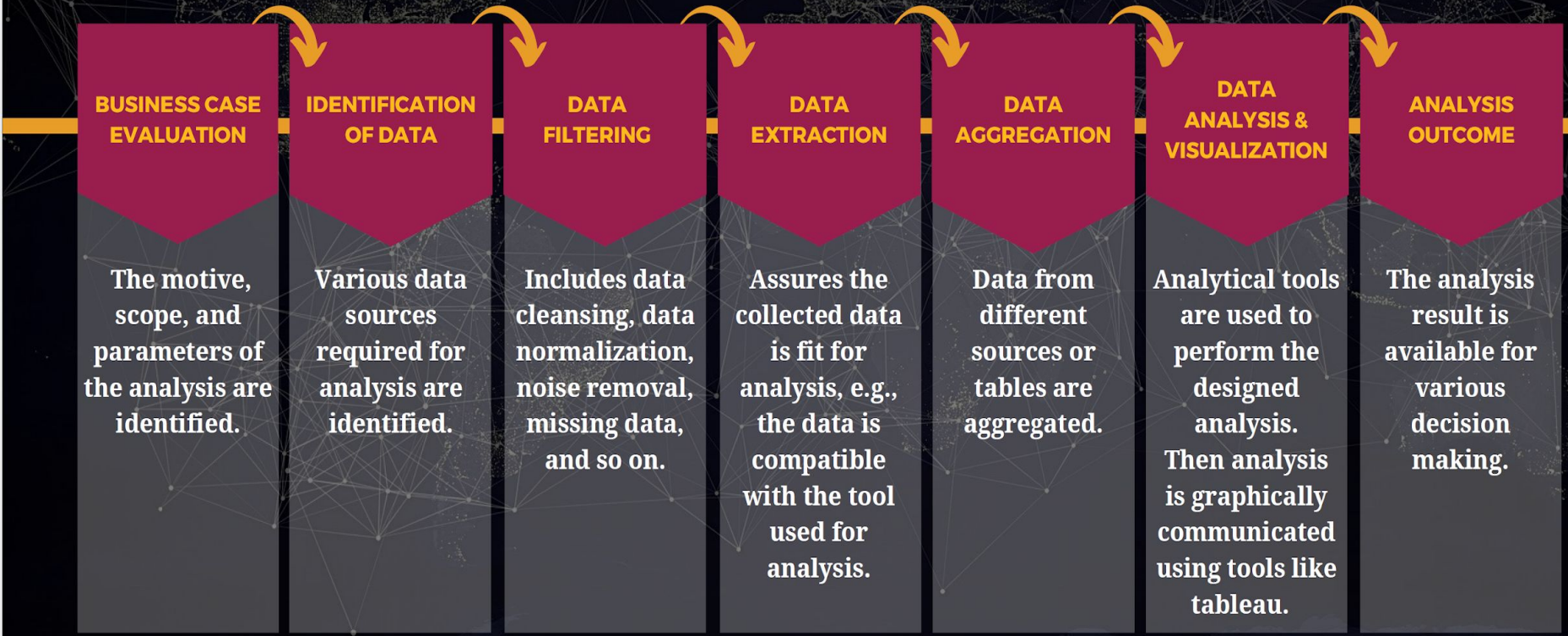
Boston Marathon Data Analytics

By Bibo Gao

BIG DATA ANALYTICS PIPELINE STAGES



BUSINESS CASE EVALUATION

The motive, scope, and parameters of the analysis are identified.

The Boston Marathon is the oldest marathon run in the US as it is the only marathon (other than olympic trails) that most of the participants have to qualify to participate. For the professional runners, it's a big accomplishment to win the marathon. For most of the other participants, it's an honor to be part of it.

Scope:

- **The age distribution for male and female**
- **The average finishing time for different age group**
- **The participants' country distribution**
- **The pace distribution**

IDENTIFICATION OF DATA

Various data sources required for analysis are identified.

- **Boston Marathon Official Website**
  http://registration.baa.org/2015/cf/Public/iframe_ResultsSearch.cfm

  - It includes all the finishers (total 26,596) of the Boston Marathon in 2015, and it contains participants' name, age, gender, country, city, and state (if available), times at different checkpoints, finished time and pace, overall place, gender place and division place.

Bib	Name	Age	M/F	City	State	Country	5K	10K	15K	20K	Half	25K	30K	35K	40K	Pace	Official Time	Overall	Gender	Division
3	Desisa, Lelisa	25	M	Ambo	?	ETH	0:14:43	0:29:43	0:44:57	1:00:29	1:04:02	1:16:07	1:32:00	1:47:59	2:02:39	0:04:56	2:09:17	1	1	1
4	Tsegay, Yem...	30	M	Addis Ababa	?	ETH	0:14:43	0:29:43	0:44:58	1:00:28	1:04:01	1:16:07	1:31:59	1:47:59	2:02:42	0:04:58	2:09:48	2	2	2
8	Chebet, Wilson	29	M	Marakwet	?	KEN	0:14:43	0:29:43	0:44:57	1:00:29	1:04:02	1:16:07	1:32:00	1:47:59	2:03:01	0:04:59	2:10:22	3	3	3
11	Kipyego, Ber...	28	M	Eldoret	?	KEN	0:14:43	0:29:44	0:45:01	1:00:29	1:04:02	1:16:07	1:32:00	1:48:03	2:03:47	0:05:00	2:10:47	4	4	4
10	Korir, Wesley	32	M	Kitale	?	KEN	0:14:43	0:29:44	0:44:58	1:00:28	1:04:01	1:16:07	1:32:00	1:47:59	2:03:27	0:05:00	2:10:49	5	5	5
9	Chepkwony, ...	30	M	Koibatek	?	KEN	0:14:44	0:29:45	0:44:59	1:00:29	1:04:02	1:16:07	1:32:00	1:47:59	2:03:18	0:05:00	2:10:52	6	6	6
14	Ritzenhein, D...	32	M	Rockford	MI	USA	0:14:45	0:29:45	0:45:20	1:00:43	1:04:03	1:16:05	1:31:59	1:48:06	2:04:05	0:05:01	2:11:20	7	7	7
1	Keflezighi, Meb	39	M	San Diego	CA	USA	0:14:44	0:29:44	0:44:59	1:00:30	1:04:02	1:16:07	1:31:59	1:47:59	2:04:58	0:05:04	2:12:42	8	8	8
5	Tola, Tadese	27	M	Addis Ababa	?	ETH	0:14:43	0:29:43	0:44:58	1:00:28	1:04:02	1:16:07	1:32:00	1:48:00	2:04:39	0:05:06	2:13:35	9	9	9
16	Shafar, Vitaliy	33	M	Lutsk	?	UKR	0:15:14	0:30:34	0:46:05	1:01:43	1:05:07	1:17:18	1:33:11	1:49:43	2:06:16	0:05:07	2:13:52	10	10	10

DATA FILTERING

Includes data cleansing, data normalization, noise removal, missing data, and so on.

- Remove “Citizen” column from the *.csv files, as there is not any valid input data.
- Remove “Project Time” column from the *.csv files, as the values are duplicated with the values of “Official Time” field.
- Change the value of “Official Time” field of index “2378” to “2:59:59”, as the original value is not correct.

DATA EXTRACTION

Assures the collected data is fit for analysis, e.g., the data is compatible with the tool used for analysis.

- The index / row-key column is available
- The collected data fit into CSV format with comma separator
- The collected data can be divided into three column families:
  - profile
  - time
  - place

```
create view "bm_2015" (
    "index" VARCHAR PRIMARY KEY,
    "profile"."bib" VARCHAR,
    "profile"."lname" VARCHAR,
    "profile"."fname" VARCHAR,
    "profile"."age" VARCHAR,
    "profile"."gender" VARCHAR,
    "profile"."city" VARCHAR,
    "profile"."state" VARCHAR,
    "profile"."country" VARCHAR,
    "time"."5k" VARCHAR,
    "time"."10k" VARCHAR,
    "time"."15k" VARCHAR,
    "time"."20k" VARCHAR,
    "time"."half" VARCHAR,
    "time"."25k" VARCHAR,
    "time"."30k" VARCHAR,
    "time"."35k" VARCHAR,
    "time"."40k" VARCHAR,
    "time"."pace" VARCHAR,
    "time"."total" VARCHAR,
    "place"."overall" VARCHAR,
    "place"."gender" VARCHAR,
    "place"."division" VARCHAR);
```

DATA AGGREGATION

Data from different sources or tables are aggregated.

- The age data are aggregated over a given period (10-19, 20-29…) to provide statistics such as average finishing time and total count within the age group.

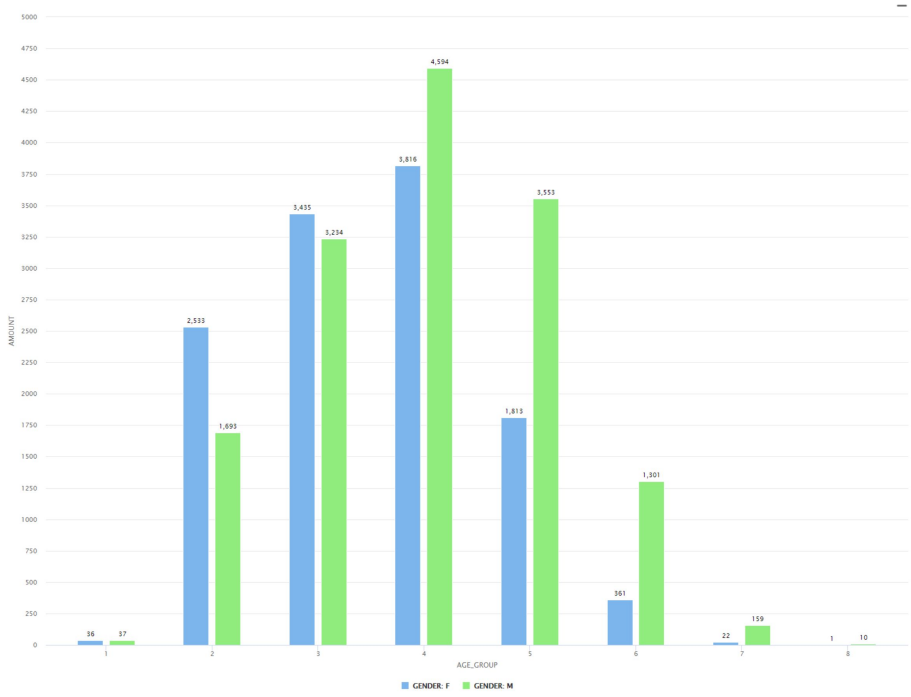
DATA ANALYSIS & VISUALIZATION

Analytical tools are used to perform the designed analysis. Then analysis is graphically communicated using tools like tableau.

- Apache HBase/Phoenix are used to manipulate the data (see project document for details)
- RapidMiner is used to visualize the data (see following slides)

Age&Gender Distribution

```
0: jdbc:phoenix:192.168.1.204> SELECT (CAST(TO_NUMBER("profile"."age") AS INTEGER) / 10) AS AGE_GROUP,
. . . . . .> "profile"."gender" AS gender,
. . . . . .> COUNT(*) AS amount
. . . . . .> FROM "bm_2015"
. . . . . .> GROUP BY age_group, "profile"."gender";
+-----------+--------+---------+
| AGE_GROUP | GENDER | AMOUNT  |
+-----------+--------+---------+
| 1         | F      | 36      |
| 2         | F      | 2533    |
| 3         | F      | 3435    |
| 4         | F      | 3816    |
| 5         | F      | 1813    |
| 6         | F      | 361     |
| 7         | F      | 22      |
| 8         | F      | 1       |
| 1         | M      | 37      |
| 2         | M      | 1693    |
| 3         | M      | 3234    |
| 4         | M      | 4594    |
| 5         | M      | 3553    |
| 6         | M      | 1301    |
| 7         | M      | 159     |
| 8         | M      | 10      |
+-----------+--------+---------+
16 rows selected (0.449 seconds)
```



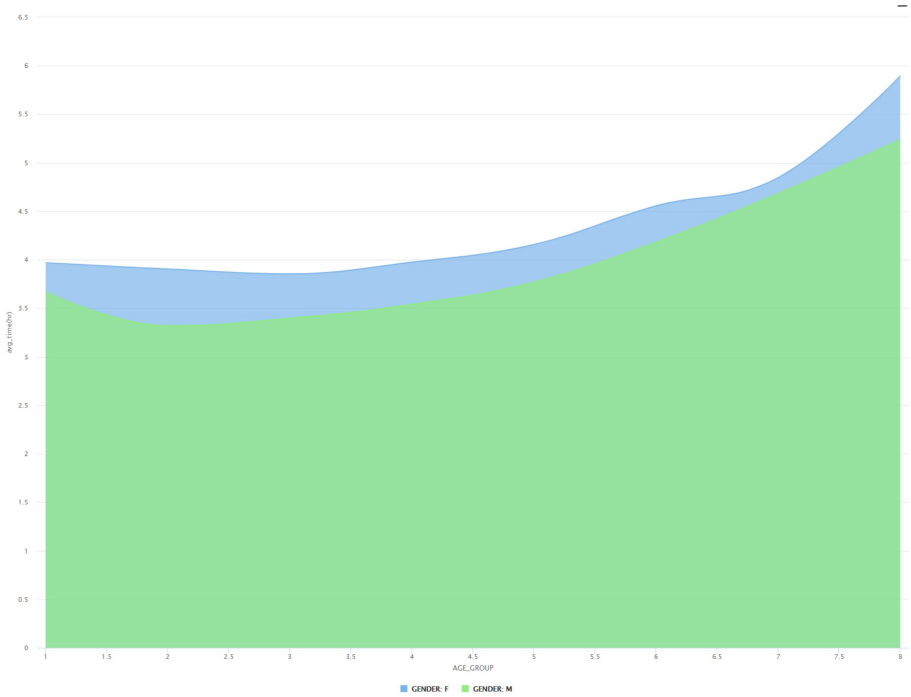
Amount of finishers per age group

Time Distribution

```
0: jdbc:phoenix:192.168.1.204> SELECT (CAST(TO_NUMBER("profile"."age") AS INTEGER) / 10) AS age_group,
. . . . . .>                "profile"."gender" AS gender,
. . . . . .>                AVG(TO_NUMBER(TO_TIME("time"."total",'HH:mm:ss')) / 1000) AS avg_time
. . . . . .> FROM "bm_2015"
. . . . . .> GROUP BY age_group, "profile"."gender";
```

AGE_GROUP	GENDER	AVG_TIME
1	F	14292.5
2	F	14058.0525
3	F	13876.6387
4	F	14310.3629
5	F	14968.1737
6	F	16402.144
7	F	17449.8181
8	F	2.122E+4
1	M	13198.081
2	M	11946.2451
3	M	12228.1376
4	M	12744.1242
5	M	13573.2079
6	M	15055.6587
7	M	16858.8301
8	M	18861.2

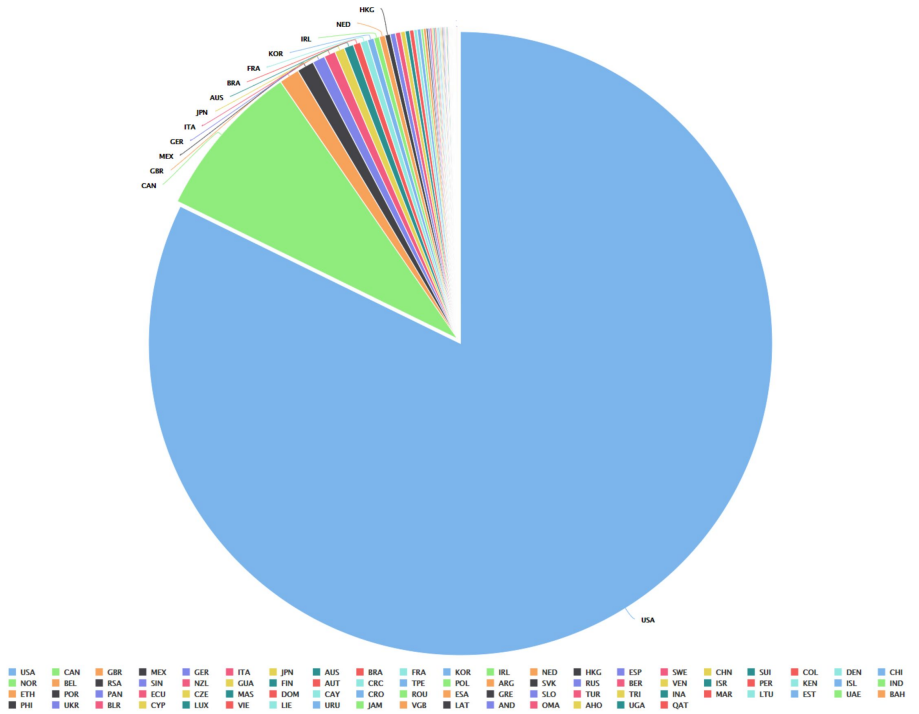
16 rows selected (0.436 seconds)



Average finished time per age group

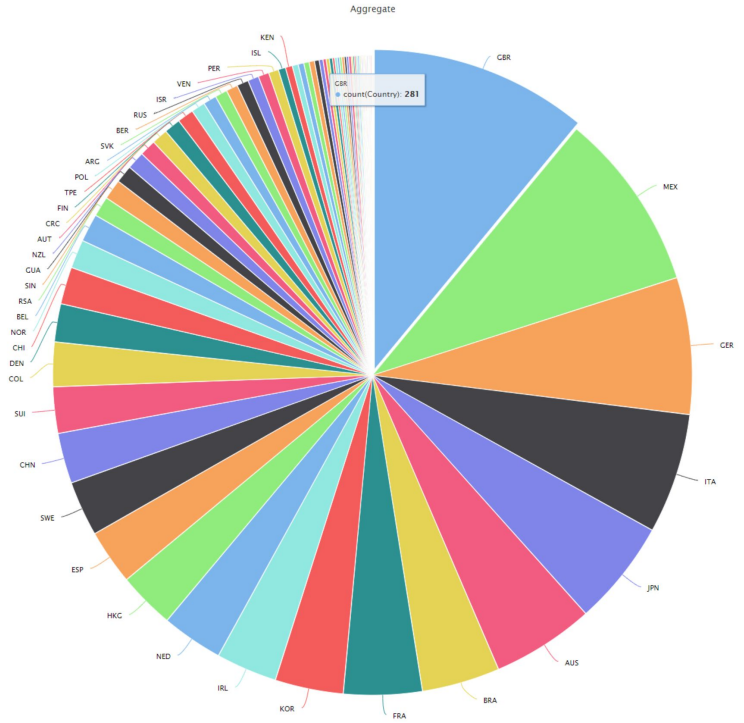
Country Distribution

```
0: jdbc:phoenix:192.168.1.204> select "profile"."country", count(*) as amount
. . . . . .> from "bm_2015"
. . . . . .> group by "profile"."country"
. . . . . .> order by amount desc
. . . . . .> limit 10;
+----------+---------+
| country  | AMOUNT  |
+----------+---------+
| USA      | 21880   |
| CAN      | 2167    |
| GBR      | 281     |
| MEX      | 231     |
| GER      | 176     |
| ITA      | 155     |
| JPN      | 136     |
| AUS      | 132     |
| BRA      | 101     |
| FRA      | 101     |
+----------+---------+
10 rows selected (0.222 seconds)
```



Country Distribution w/o USA & CAN

```
0: jdbc:phoenix:192.168.1.204> select "profile"."country", count(*) as amount
. . . . . .> from "bm_2015"
. . . . . .> group by "profile"."country"
. . . . . .> order by amount desc
. . . . . .> limit 10;
+----------+---------+
| country  | AMOUNT  |
+----------+---------+
| USA      | 21880   |
| CAN      | 2167    |
| GBR      | 281     |
| MEX      | 231     |
| GER      | 176     |
| ITA      | 155     |
| JPN      | 136     |
| AUS      | 132     |
| BRA      | 101     |
| FRA      | 101     |
+----------+---------+
10 rows selected (0.222 seconds)
```



Pace Distribution

```
0: jdbc:phoenix:192.168.1.204> select "time"."pace" as pace_group,
. . . . . .> "profile"."gender" as gender,
. . . . . .> count(*) as amount
. . . . . .> from "bm_2015"
. . . . . .> group by pace_group, gender
. . . . . .> having "profile"."gender" = 'F'
. . . . . .> order by count(*) desc
. . . . . .> limit 10;
```

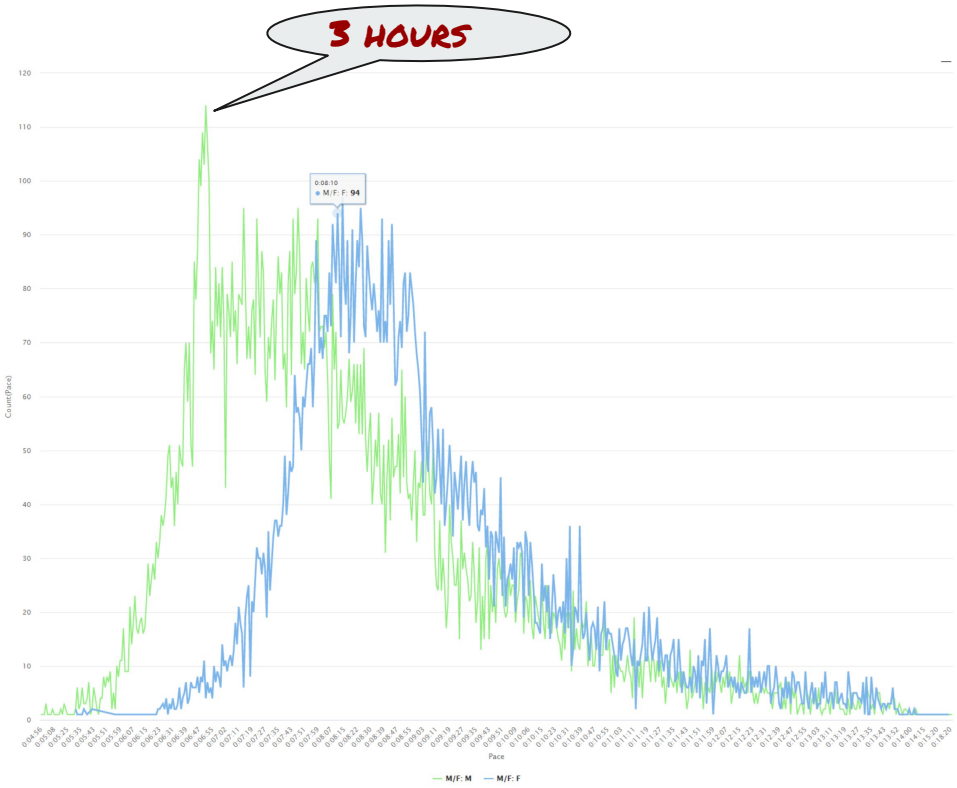
PACE_GROUP	GENDER	AMOUNT
0:08:14	F	97
0:08:24	F	95
0:08:10	F	95
0:08:38	F	93
0:08:08	F	92
0:08:44	F	92
0:08:20	F	91
0:08:17	F	90
0:08:22	F	89
0:08:25	F	89

10 rows selected (0.186 seconds)

```
0: jdbc:phoenix:192.168.1.204> select "time"."pace" as pace_group,
. . . . . .> "profile"."gender" as gender,
. . . . . .> count(*) as amount
. . . . . .> from "bm_2015"
. . . . . .> group by pace_group, gender
. . . . . .> having "profile"."gender" = 'M'
. . . . . .> order by count(*) desc
. . . . . .> limit 10;
```

PACE_GROUP	GENDER	AMOUNT
0:06:51	M	114
0:06:49	M	109
0:06:52	M	107
0:06:47	M	104
0:06:50	M	103
0:06:53	M	100
0:06:48	M	99
0:07:14	M	95
0:07:47	M	95
0:07:59	M	93

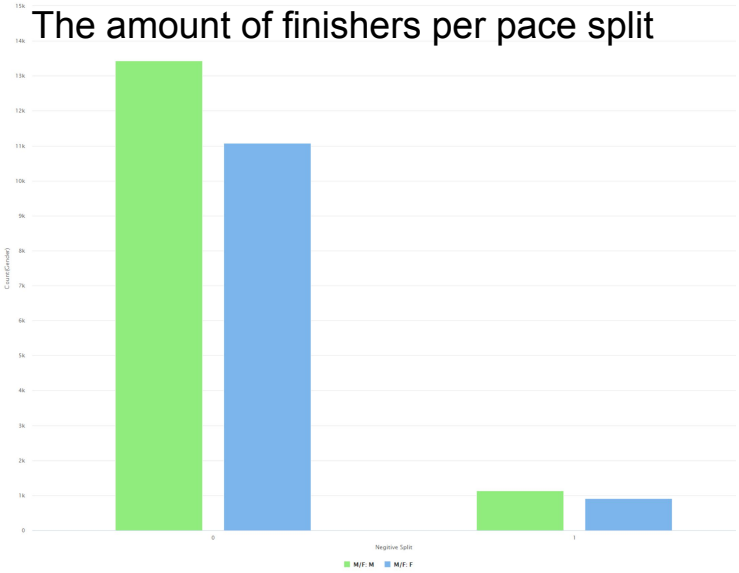
10 rows selected (0.155 seconds)

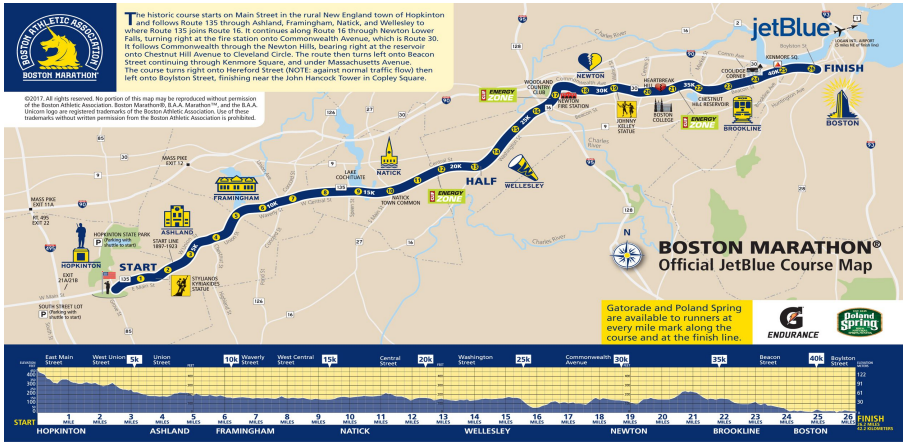


The amount of finishers per different pace

Pace Split Distribution

```
0: jdbc:phoenix:192.168.1.204> select (((cast(to_number(to_time("time"."total",'HH:mm:ss')) as integer) / 1000) -
2 * (cast(to_number(to_time("time"."half",'HH:mm:ss')) as integer) / 1000)) > 0) as positive_split,
. . . . . .>        "profile"."gender" as gender,
. . . . . .>        count(*) as amount
. . . . . .> from "bm_2015"
. . . . . .> where "time"."half" != '-'
. . . . . .> group by positive_split, gender;
+---------------+---------+---------+
| POSITIVE_SPLIT | GENDER | AMOUNT |
+---------------+---------+---------+
| false         | F       | 916     |
| true          | F       | 11081   |
| false         | M       | 1135    |
| true          | M       | 13438   |
+---------------+---------+---------+
4 rows selected (0.313 seconds)
```





ANALYSIS OUTCOME

The analysis result is available for various decision making.

- The age distribution
  - The number of female runner more than the number of male runner before age 40
  - The majority of runners are from age 40 to 49
- The finishing time per age
  - 20 - 29 : the male age of fast runners (3h20m total time)
  - 30 - 39 : the female age of fast runners (3h50m total time)
- The country distribution
  - The majority of runners are from US, CAN, UK
- The pace distribution
  - 06m51s per mile (3h total time)
    - most popular pace for male
    - Benchmark for professional runner
  - 07m59s to 08m44s per mile (3h30m to 3h50m total time)
    - most popular pace for female
  - The most pace split is positive (slower after half distance)