# HW_1 Spark architecture

FINISHED

## Урок 1. Архитектура Spark. Принципы исполнения запросов. Сохранение и чтение данных

Took 0 sec. Last updated by 305_koryagin at January 27 2021, 10:10:11 PM. (outdated)

```
spark.sparkContext.applicationId
```

FINISHED

```
res1: String = application_1611766633932_0001
```

Took 32 sec. Last updated by 305_koryagin at January 27 2021, 10:10:45 PM. (outdated)

## Домашнее задание 1. Визуализация

FINISHED

https://s3.amazonaws.com/apache-zeppelin/tutorial/bank/bank.csv (https://s3.amazonaws.com/apache-zeppelin/tutorial/bank/bank.csv)

1. Построить распределения клиентов по возрастам
2. Распределение по возрасту с динамическим численным параметром `max_age`
3. Распределение по возрасту с динамическим параметром `marital`

Took 0 sec. Last updated by 305_koryagin at January 27 2021, 10:10:31 PM. (outdated)

```
%pyspark
bank_df = spark.table("homework.bank")
bank_df.cache()
```

FINISHED

```
DataFrame[age: int, job: string, marital: string, education: string, default: string, balance: int, housing: string, loan: string, contact: string, d
ay: int, month: string, duration: int, campaign: int, pdays: int, previous: int, poutcome: string, y: string]
```

Took 1 sec. Last updated by 305_koryagin at January 27 2021, 10:19:59 PM. (outdated)

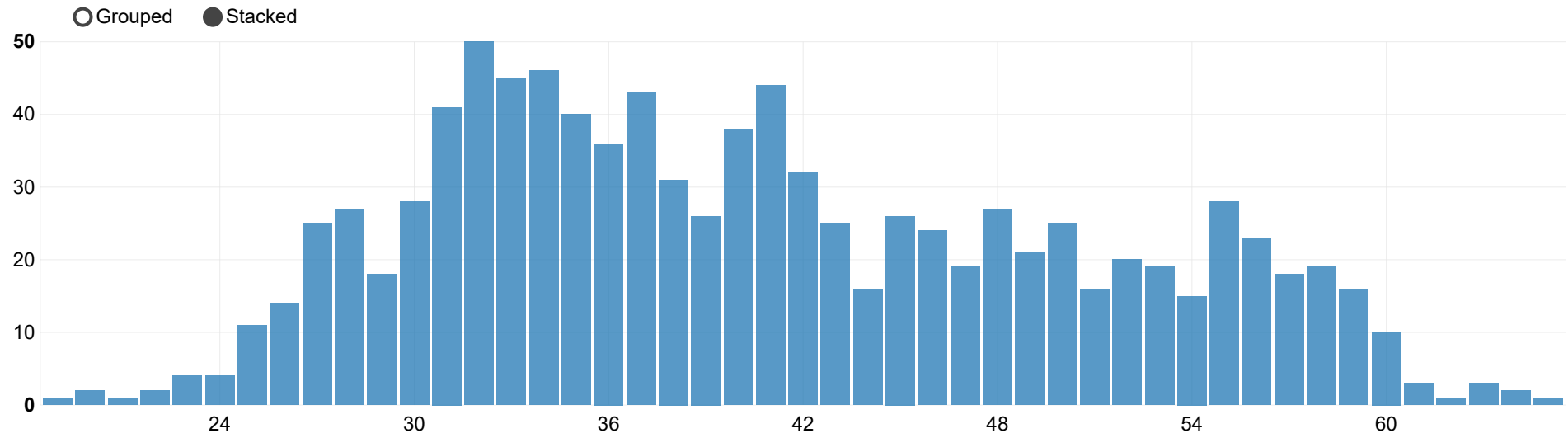
# HW_1 Spark architecture

## 1. Построить распределения клиентов по возрастам

SPARK JOB (http://bigdataanalytics-head-0.novalocal:4040/jobs/job?id=4) FINISHED

```
%pyspark
z.show(bank_df)
```

settings




**Output is truncated** to 1000 rows. Learn more about **zeppelin.spark.maxResult**

Took 1 sec. Last updated by 305_koryagin at January 27 2021, 10:20:11 PM. (outdated)

## 2. Распределение по возрасту с динамическим численным параметром `max_age`

SPARK JOB FINISHED

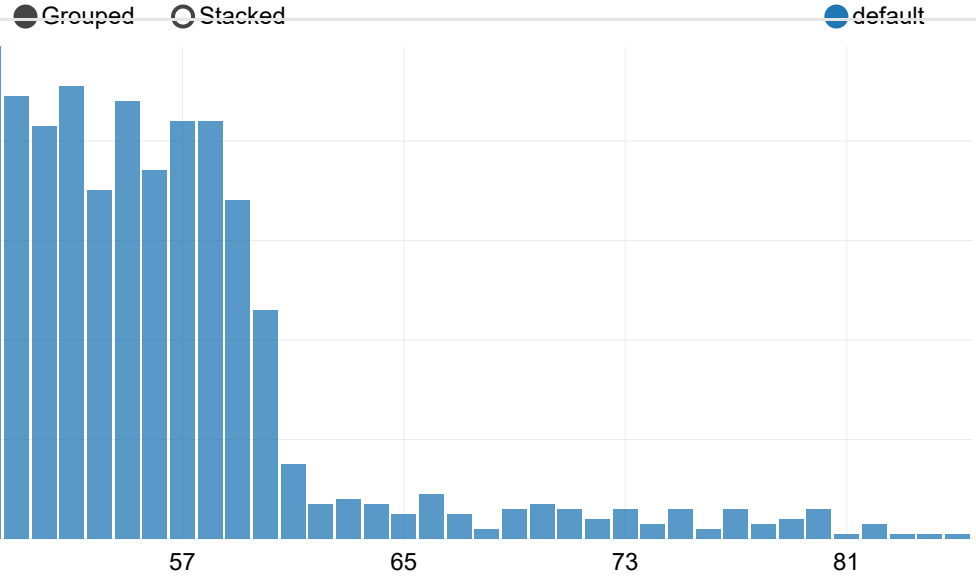
```
%pyspark
z.show(
    bank_df.filter(bank_df["age"] >= z.input("max_age"))
    )
```

**max_age**

50

settings

# HW_1 Spark architecture




**Output is truncated** to 1000 rows. Learn more about **zeppelin.spark.maxResult**

Took 5 sec. Last updated by 305_koryagin at January 27 2021, 10:20:21 PM. (outdated)

## 3. Распределение по возрасту с динамическим параметром `marital`

SPARK JOB FINISHED

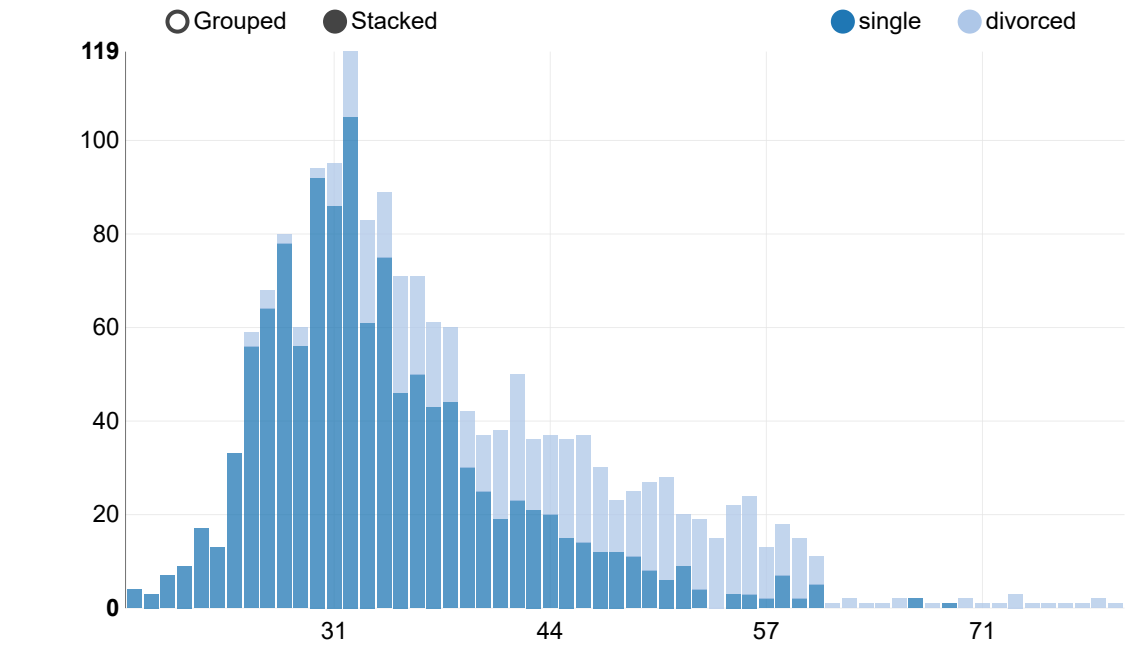
```
%pyspark
marital_status = [("'single'","single"), ("'married'","married"), ("'divorced'"
    ,"divorced")]
marital_status_list = ", ".join(z.checkbox("marital_status", marital_status,
    ["'single'"]))
sql_select = "select bank.marital, age, count(*) as count from homework.bank where
    bank.marital in (" + marital_status_list + ") group by age, bank.marital"
z.show(spark.sql(sql_select))
```

marital_status

single married divorced

# HW_1 Spark architecture

settings




Took 2 sec. Last updated by 305_koryagin at January 27 2021, 10:23:58 PM. (outdated)

## Домашнее задание 2. Fire Station onboarding

FINISHED

/user/admin/sf-fire-calls.csv (/user/admin/sf-fire-calls.csv)

1. What were all the different types of fire calls in 2018?
2. What months within the year 2018 saw the highest number of fire calls?

3. Which neighborhood in San Francisco generated the most fire calls in 2018?
4. Which neighborhoods had the worst response times to fire calls in 2018?
5. Which week in the year in 2018 had the most fire calls?
6. Is there a correlation between neighborhood, zip code, and number of fire calls?
7. How can we use Parquet files or SQL tables to store this data and read it back?

# HW_1 Spark architecture

Took 1 sec. Last updated by anonymous at January 17 2021, 3:49:49 PM. (outdated)

```
%pyspark

path = '/user/admin/sf-fire-calls.csv'

fire_station_df = spark.read.option("header", True).csv(path)
fire_station_df.createOrReplaceTempView("fire_station_table")
```

SPARK JOB (http://bigdataanalytics-head-0.novalocal:4040/jobs/job?id=19) FINISHED

Took 0 sec. Last updated by 305_koryagin at January 27 2021, 10:28:40 PM. (outdated)

```
%pyspark
fire_station_df.printSchema()
```

FINISHED

```
root
 |-- CallNumber: string (nullable = true)
 |-- UnitID: string (nullable = true)
 |-- IncidentNumber: string (nullable = true)
 |-- CallType: string (nullable = true)
 |-- CallDate: string (nullable = true)
 |-- WatchDate: string (nullable = true)
 |-- CallFinalDisposition: string (nullable = true)
 |-- AvailableDtTm: string (nullable = true)
 |-- Address: string (nullable = true)
 |-- City: string (nullable = true)
 |-- Zipcode: string (nullable = true)
 |-- Battalion: string (nullable = true)
 |-- StationArea: string (nullable = true)
 |-- Box: string (nullable = true)
 |-- OriginalPriority: string (nullable = true)
 |-- Priority: string (nullable = true)
 |-- FinalPriority: string (nullable = true)
```

Took 0 sec. Last updated by 305_koryagin at January 27 2021, 10:28:58 PM. (outdated)

# HW_1 Spark architecture

## 1. What were all the different types of fire calls in 2018?

SPARK JOB FINISHED

```
%sql
SELECT
    DISTINCT(CallType) as types_of_fire_calls,
    count(UnitID) as number_of_fire_calls
FROM fire_station_table WHERE CallDate LIKE '%/2018'
group by types_of_fire_calls;
```

settings

| types_of_fire_calls | number_of_fire_calls |
|---|---|
| Vehicle Fire | 28 |
| Suspicious Package | 3 |
| Structure Fire | 906 |
| Alarms | 1144 |
| Electrical Hazard | 30 |
| Medical Incident | 7004 |
| Outside Fire | 153 |
| Odor (Strange / Unknown) | 10 |

Took 4 sec. Last updated by anonymous at January 18 2021, 4:43:44 PM. (outdated)

## 2. What months within the year 2018 saw the highest number of fire calls?

SPARK JOB (http://bigdataanalytics-head-0.novalocal:4040/jobs/job?id=57) FINISHED

```
%sql
SELECT
    SUBSTRING(CallDate, 7, 4) as year,
    date_format(concat(SUBSTRING(CallDate, 7, 4), '-', SUBSTRING(CallDate, 1, 2), '
        -', SUBSTRING(CallDate, 4, 2)), 'MMM') as month,
    -- SUBSTRING(CallDate, 1, 2) as month,
```

```
    count(UnitID) as number_of_fire_calls
FROM fire_station_table
WHERE CallDate LIKE '%/2018'
group by month, year
order by number_of_fire_calls desc
```

# HW_1 Spark architecture

settings

| year | month | number_of_fire_calls |
|---|---|---|
| 2018 | Oct | 1068 |
| 2018 | May | 1047 |
| 2018 | Mar | 1029 |

Took 1 sec. Last updated by anonymous at January 18 2021, 2:27:07 PM. (outdated)

## 3. Which neighborhood in San Francisco generated the most fire calls in 2018?

SPARK JOB (http://bigdataanalytics-head-0.novalocal:4040/jobs/job?id=449) FINISHED

```
%sql
select Neighborhood, count(UnitID) as number_of_fire_calls
from fire_station_table
WHERE CallDate LIKE '%/2018' and City = 'San Francisco'
group by Neighborhood
order by number_of_fire_calls desc
limit 1;
```

settings

| Neighborhood | number_of_fire_calls |
|---|---|

| | |
|---|---|
| Tenderloin | 1393 |

Took 2 sec. Last updated by anonymous at January 17 2021, 2:28:48 AM. (outdated)

# HW_1 Spark architecture

## 5. Which week in the year in 2018 had the most fire calls?

SPARK JOB (http://bigdataanalytics-head-0.novalocal:4040/jobs/job?id=124) FINISHED

```
%sql
SELECT
    date_format(concat(SUBSTRING(CallDate, 7, 4), '-', SUBSTRING(CallDate, 1, 2), '
        -', SUBSTRING(CallDate, 4, 2)), 'w') as week,
    count(UnitID) AS number_of_fire_calls
FROM fire_station_table
WHERE CallDate LIKE '%/2018'
GROUP BY week
ORDER BY number_of_fire_calls DESC
LIMIT 1;
```

settings

| week | number_of_fire_calls |
|---|---|
| 22 | 272 |

Took 2 sec. Last updated by anonymous at January 18 2021, 3:27:30 PM. (outdated)

## 4. Which neighborhoods had the worst response times to fire calls in 2018?

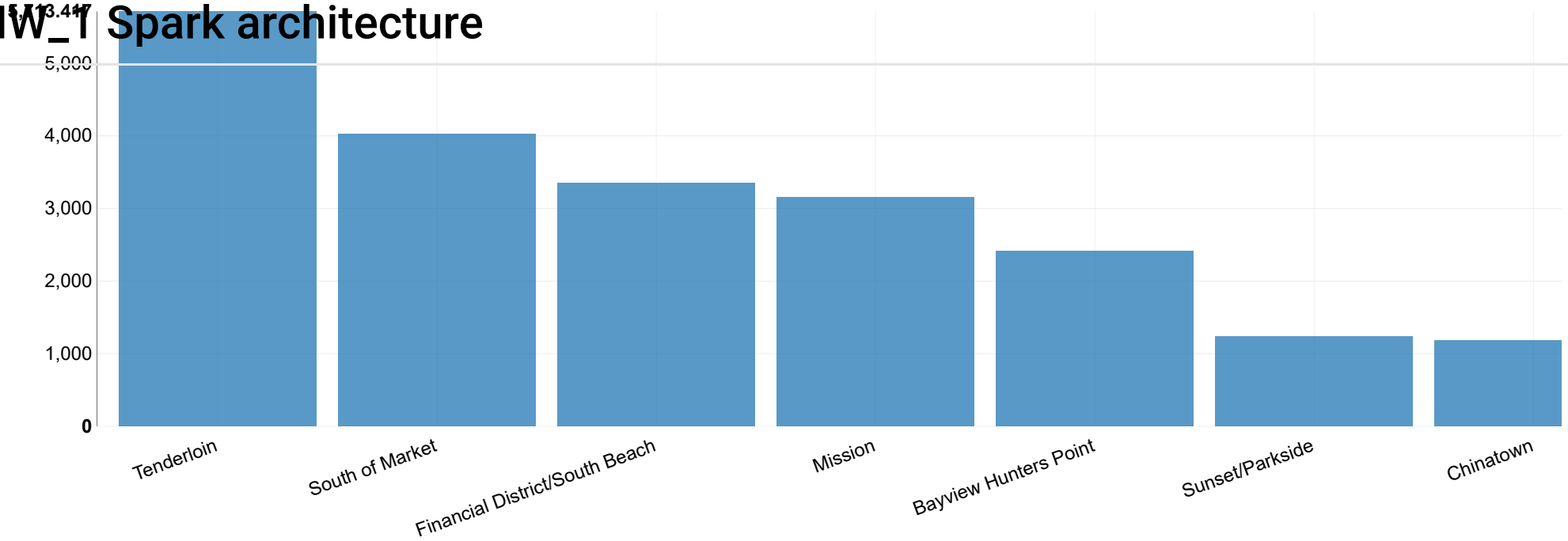
SPARK JOB (http://bigdataanalytics-head-0.novalocal:4040/jobs/job?id=457) FINISHED

```
%sql
SELECT
    Neighborhood,
    sum(Delay) as sum_delay
FROM fire_station_table
WHERE CallDate LIKE '%/2018'
GROUP BY Neighborhood
OREDER BY sum_delay desc
LIMIT 8;
```

settings

Grouped Stacked

# HW_1 Spark architecture



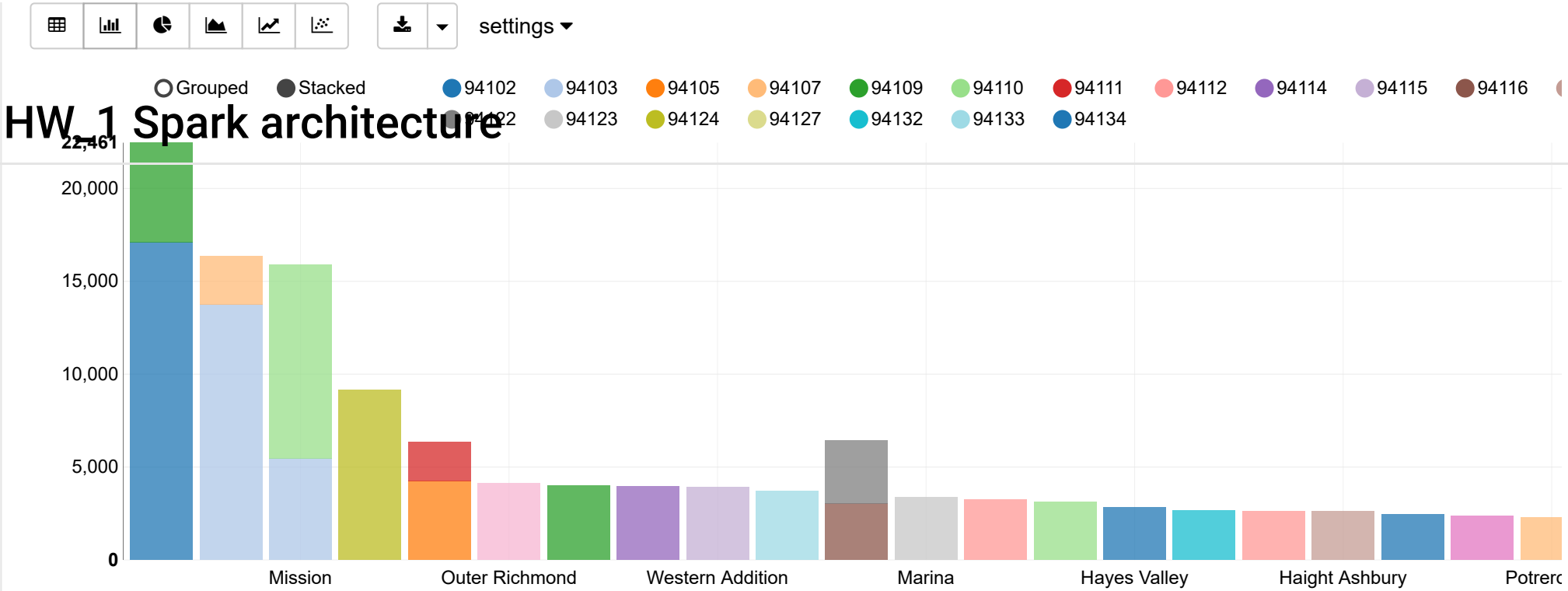

Took 1 sec. Last updated by anonymous at January 17 2021, 2:31:11 AM. (outdated)

## 6. Is there a correlation between neighborhood, zip code, and number of fire calls?

SPARK JOB (http://bigdataanalytics-head-0.novalocal:4040/jobs/job?id=376) FINISHED

```
%sql
SELECT
    Neighborhood, Zipcode, count(UnitID) AS number_of_fire_calls
FROM fire_station_table
GROUP BY Neighborhood, Zipcode
ORDER BY number_of_fire_calls DESC
LIMIT 30;

-- не смог сделать(
```




Took 1 sec. Last updated by anonymous at January 18 2021, 4:31:29 PM. (outdated)

## 7. How can we use Parquet files or SQL tables to store this data and read it back?

FINISHED

```
%md
#### Как мы можем использовать файлы Parquet или таблицы SQL для хранения этих данных и считывания их обратно?

- Мы можем сохранить эти данные в файл parquet или таблицу hive, и работать уже не как с pandas.df, а уже как sql таблицей + исользовать тогда joi
```

### Как мы можем использовать файлы Parquet или таблицы SQL для хранения этих данных и считывания их обратно?

- Мы можем сохранить эти данные в файл parquet или таблицу hive, и работать уже не как с pandas.df, а уже как sql таблицей + исользовать тогда join разных таблиц

Took 0 sec. Last updated by anonymous at January 18 2021, 4:37:56 PM. (outdated)