Laporan Akhir Big Data
Analisa Facebook Data

Kelompok 8

Rionaldi Dwipurna Wongsoputra	**2006577460**
Benedicto Matthew	**2006577435**
Matthew Eucharist	**2006577580**
Adri Ahmad An Nabaa	**1706043153**

Pendahuluan

Social media merupakan media yang digunakan untuk berkomunikasi dan berinteraksi antar personal secara jarak jauh. Media sosial seperti Facebook, Twitter, dan LinkedIn belakangan ini menjadi popular dengan disertai banyaknya user yang menggunakannya. Facebook merupakan platform media sosial yang digunakan untuk berkoneksi dengan orang-orang. Pada tahun 2021, Facebook merupakan media sosial terbesar dengan jumlah pengguna aktif perbulannya mencapai 2.85 miliar pengguna. Pada kesempatan ini, kelompok kami akan menganalisa user dari Facebook yang dapat berguna untuk bisnis untuk mengambil *intelligent decision* untuk berkoneksi kepada penggunanya.

Kami menggunakan dataset yang diambil dari website Kaggle yang berisi informasi mengenai user berupa umur, gender, dan lain lain.

Tools Big Data

Tools yang akan kami gunakan adalah Hadoop, Apache Spark, Apache Drill, dan Apache Zeppelin.

Hadoop

Hadoop merupakan framework open-source yang berguna untuk memproses dan menyimpan dataset berukuran besar yang berkisar dari gigabit ke pentabit. Dengan menggunakan Hadoop kita tidak harus menggunakan komputer berukuran besar untuk menganalisis dataset berukuran besar, tetapi data tersebut diproses secara paralel dengan melakukan clustering terhadap komputer-komputer. Hadoop terdiri dari 4 (empat) modul yaitu

- HDFS (Hadoop Distributed File System)
  Adalah sebuah sistem file terdistribusi yang berjalan pada standard atau low-end hardware. HDFS menyediakan data throughput yang lebih baik dari file system tradisional menyediakan th
- Yet Another Resource Negotiator (YARN), mengatur dan memantau node cluster dan peggunaan resource. YARN menjadwalkan jobs dan tasks
- MapReduce, framework yang membantu program untuk melakukan komputasi secara paralel pada data. Map tasks akan mengambil input data dan mengkonversinya ke dalam sebuah dataset yang bisa dikomputasi dalam bentuk pasangan key value. Output dari map task kemudian di gunakan oleh reduce task untuk mengagregasi output dan menghasilkan output yang diinginkan
- Hadoop Common → menyediakan library umum milik Java yang bisa digunakan semua modul.

Apache Spark

Apache Spark adalah sebuah open source data-processing engine untuk data sets berukuran besar. Apache Spark ini didesain untuk memberikan kecepatan komputasi, skalabilitas , dan programmability yang dibutuhkan oleh Big Data, spesifiknya adalah untuk streaming data, memetakan data ke dalam bentuk grafik, machine learning, dan aplikasi Artificial Intelligence (AI)

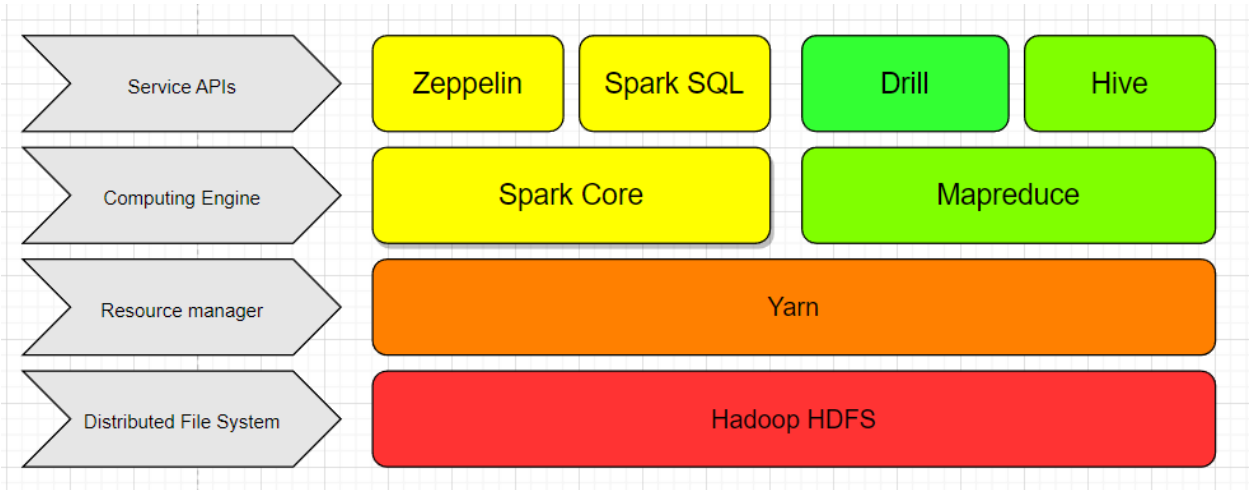
Apache Drill

Adalah mesin query yang schema-free dan memiliki latency rendah untuk big data. Drill menggunakan JSON document model sehingga ia bisa melakukan query data dengan berbagai macam struktur. Drill bekerja dengan bermacam jenis non-relational data stores seperti Hadoop, NoSQL databases, serta cloud storage seperti Amazon S3, Azure Blob Storage dan lain-lain. User bisa meng-query data menggunakan tools dasar SQL dan BI, yang tidak membutuhkan pembuatan skema.

Apache Zeppelin

Adalah open web-based multi-purpose notebook yang mendukung analisis data secara interaktif. Apache Zeppelin adalah tools yang membawa fitur data ingestion, data exploration, visualization, sharing and collaboration kepada Hadoop. Dengan menggunakan Apache Zeppelin maka user bisa lebih produktif melakukan analisis data dengan mengembangkan, mengorganisir, mengeksekusi, dan membagikan kode data dan hasil visualisasi tanpa perlu memberikan cluster details. Zeppelin tidak hanya mendukung bahasa pemrograman Python, tetapi juga seperti Scala, Hive, SparkSQL, shell dan markdown.

Skema



Kami menggunakan Hadoop File System sebagai file system dan Yarn sebagai *resource manager* dari proyek kami. Digunakan Spark Core untuk mengolah data untuk menghitung

jumlah null dari dataset facebook. Map reduce digunakan untuk membuat pasangan key value untuk menghitung jumlah user berdasarkan umur. Digunakan spark SQL untuk mengimplementasikan hal tersebut. Drill dan Hive digunakan untuk melakukan query dari database untuk kami analisa lebih lanjut. Zeppelin digunakan untuk melakukan visualisasi data dari data yang telah kami query

Implementasi Kode

Kode untuk mengecek jumlah null yang ada pada table


```
from pyspark.sql import SparkSession
from pyspark.sql import Row
from pyspark.sql import functions

def parseInput(line):
    fields = line.split(',')
    return Row(value = str(fields[i]))

if __name__ == "__main__":
    # Create a SparkSession (the config bit is only for Windows!)
    spark = SparkSession.builder.appName("SanityCheck").getOrCreate()

    # Get the raw data
    lines = spark.sparkContext.textFile("hdfs://localhost:9000/facebook/pseudo_facebook.csv")


a=["userid","age","dob_day","dob_year","dob_month","gender","tenure","friend_count","friendships_initiated","likes","likes_received","mobile_likes","mobile_likes_received","www_
"www_likes_received"]
    for i in range(15):
        # Convert it to a RDD of Row objects with (value)
        x = lines.map(parseInput)
        # Convert that to a DataFrame
        xDF = spark.createDataFrame(x)

        # Compute count of Null Values
        counts = xDF.filter(xDF["value"]=="NA").count()

        # Print them out
        print ("%s : %d"%(a[i],counts))

    # Stop the session
    spark.stop()
```

Kode map reduce pada file csv, kode ini akan menghitung age dan jumlah orangnya


```
from mrjob.job import MRJob
from mrjob.step import MRStep

class WhatAgeUsesFacebook(MRJob):
    def steps(self):
        return [
            MRStep(mapper=self.mapper_get_ages,
                   reducer=self.reducer_count_ages),
            MRStep(reducer=self.reducer_sorted_output)
        ]

    def mapper_get_ages(self, _, line):
        (userid, age, dob_day, dob_year, dob_month, gender, tenure, friend_count, friendships_initiated, likes, likes_received, mobile_likes, mobile_likes_received, www_likes,
 www_likes_receved) = line.split(',')
        yield age, 1

    def reducer_count_ages(self, age, ones):
        yield str(sum(ones)).zfill(5), age

    def reducer_sorted_output(self, count, ages):
        for age in ages:
            yield age, count

if __name__ == '__main__':
    WhatAgeUsesFacebook.run()
```

Menjalankan kode pengecekan null:

```
ubuntu@ubuntu-VirtualBox:~/Facebook Data Analysis$ python3 SanityCheck.py
22/12/18 15:20:03 WARN Utils: Your hostname, ubuntu-VirtualBox resolves to a loopback address: 127.0.1.1; using 192.168.110.75 instead (on interface enp0s3)
22/12/18 15:20:03 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/12/18 15:20:08 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
userid : 0
age : 0
dob_day : 0
dob_year : 0
dob_month : 0
gender : 175
tenure : 2
friend_count : 0
friendships_initiated : 0
likes : 0
likes_received : 0
mobile_likes : 0
mobile_likes_received : 0
www_likes : 0
www_likes_received : 0
```

Menjalankan kode map reduce:

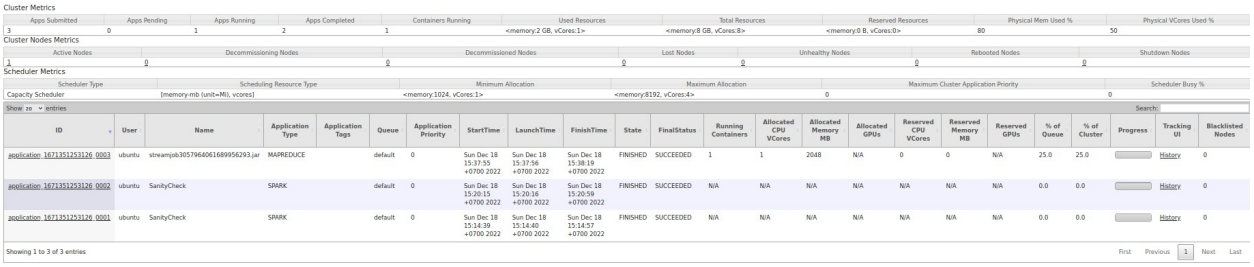
```
ubuntu@ubuntu-VirtualBox:~/Facebook Data Analysis$ python3 map_reduce1.py -r hadoop --hadoop-streaming-jar /home/ub
untu/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar hdfs://localhost:9000/facebook/pseudo_facebook.csv
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /home/ubuntu/hadoop/bin...
Found hadoop binary: /home/ubuntu/hadoop/bin/hadoop
Using Hadoop version 3.3.4
Creating temp directory /tmp/map_reduce1.ubuntu.20221218.083737.608581
uploading working dir files to hdfs:///user/ubuntu/tmp/mrjob/map_reduce1.ubuntu.20221218.083737.608581/files/wd...
Copying other local files to hdfs:///user/ubuntu/tmp/mrjob/map_reduce1.ubuntu.20221218.083737.608581/files/
Running step 1 of 2...
  packageJobJar: [/tmp/hadoop-unjar6796213944065581509/] [] /tmp/streamjob3057964061689956293.jar tmpDir=null
  Connecting to ResourceManager at /0.0.0.0:8032
  Connecting to ResourceManager at /0.0.0.0:8032
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ubuntu/.staging/job_1671351253126_0003
  Total input files to process : 1
  number of splits:2
  Submitting tokens for job: job_1671351253126_0003
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1671351253126_0003
  The url to track the job: http://ubuntu-VirtualBox:8088/proxy/application_1671351253126_0003/
  Running job: job_1671351253126_0003
  Job job_1671351253126_0003 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1671351253126_0003 completed successfully
  Output directory: hdfs:///user/ubuntu/tmp/mrjob/map_reduce1.ubuntu.20221218.083737.608581/step-output/0000
```

```
        File Input Format Counters
                Bytes Read=5221110
        File Output Format Counters
                Bytes Written=1341
        File System Counters
                FILE: Number of bytes read=895055
                FILE: Number of bytes written=2630685
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5221314
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=1341
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=11
                HDFS: Number of write operations=2
        Job Counters
                Data-local map tasks=2
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Total megabyte-milliseconds taken by all map tasks=17529856
                Total megabyte-milliseconds taken by all reduce tasks=5335040
                Total time spent by all map tasks (ms)=17119
                Total time spent by all maps in occupied slots (ms)=17119
                Total time spent by all reduce tasks (ms)=5210
                Total time spent by all reduces in occupied slots (ms)=5210
                Total vcore-milliseconds taken by all map tasks=17119
                Total vcore-milliseconds taken by all reduce tasks=5210
        Map-Reduce Framework
```

```
job output is in hdfs:///user/ubuntu/tmp/mrjob/map_reduce1.ubuntu.20221218.083737.608581/output
Streaming final output from hdfs:///user/ubuntu/tmp/mrjob/map_reduce1.ubuntu.20221218.083737.608581/output...
"age"   "00001"
"109"   "00009"
"110"   "00015"
"112"   "00018"
"111"   "00018"
"87"    "00042"
"92"    "00052"
"97"    "00056"
"89"    "00060"
"88"    "00061"
"96"    "00070"
"90"    "00071"
"104"   "00073"
"91"    "00076"
"86"    "00076"
"95"    "00077"
"82"    "00078"
"105"   "00080"
"99"    "00083"
"85"    "00083"
"84"    "00086"
"98"    "00093"
"107"   "00098"
"81"    "00108"
"79"    "00112"
"106"   "00125"
"80"    "00136"
```

```
"15"    "02618"
"26"    "02815"
"24"    "02827"
"22"    "03032"
"16"    "03086"
"17"    "03283"
"25"    "03641"
"21"    "03671"
"20"    "03769"
"19"    "04391"
"23"    "04404"
"18"    "05196"
Removing HDFS temp directory hdfs:///user/ubuntu/tmp/mrjob/map_reduce1.ubuntu.20221218.083737.608581...
Removing temp directory /tmp/map_reduce1.ubuntu.20221218.083737.608581...
ubuntu@ubuntu-VirtualBox:~/Facebook Data Analysis$
```

Pada directory HDFS,



Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Mem Used %	Physical VCores Used %
3	0	1	2	1	<memory:2 GB, vCores:1>	<memory:8 GB, vCores:8>	<memory:0 B, vCores:0>	80	50

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority	Scheduler Busy %
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0	0

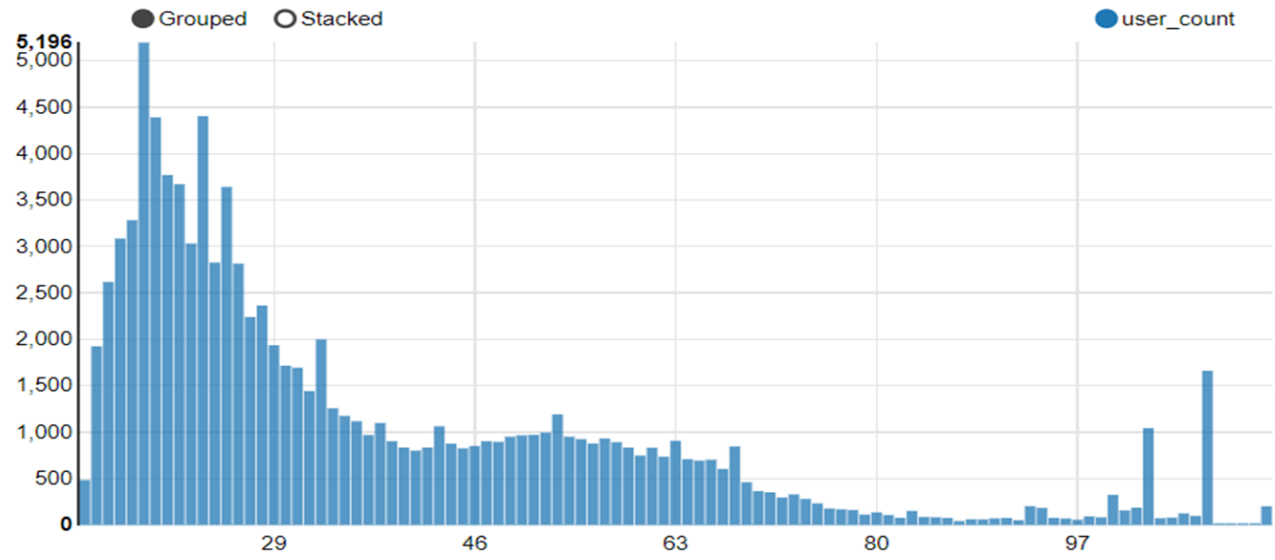
Show 20 entries Search:

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memory MB	Reserved GPUs	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1671351253126_0003	ubuntu	streamjob3057964061689956293.jar	MAPREDUCE		default	0	Sun Dec 18 15:37:55 +0700 2022	Sun Dec 18 15:37:56 +0700 2022	Sun Dec 18 15:38:19 +0700 2022	FINISHED	SUCCEEDED	1	1	2048	N/A	0	0	N/A	25.0	25.0		History	0
application_1671351253126_0002	ubuntu	SanityCheck	SPARK		default	0	Sun Dec 18 15:20:15 +0700 2022	Sun Dec 18 15:20:16 +0700 2022	Sun Dec 18 15:20:59 +0700 2022	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0
application_1671351253126_0001	ubuntu	SanityCheck	SPARK		default	0	Sun Dec 18 15:14:39 +0700 2022	Sun Dec 18 15:14:40 +0700 2022	Sun Dec 18 15:14:57 +0700 2022	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0

Showing 1 to 3 of 3 entries First Previous 1 Next Last

Dapat dilakukan visualisasi data sebagai berikut,

Jumlah umur pada pengguna facebook,



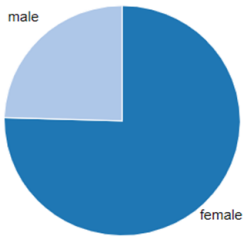
Menggunakan Apache Drill dan Zeppelin sebagai SQL framework, dapat didapatkan data sebagai,

Query:

SELECT gender,avg(likes) AS AVG_Likes_Given FROM hive.facebook_db.facebook GROUP BY gender ORDER BY AVG_Likes_Given DESC

Perbandingan antara gender dan jumlah rata-rata like yang diberikan:

gender	AVG_Likes_Given
female	260.0513240920157
NA	138.50857142857143
male	84.6778946290163



Query:

SELECT userid, gender, likes AS Total_Likes_Given FROM hive.facebook_db.facebook ORDER BY Total_likes_Given DESC LIMIT 10

Data user dengan like terbanyak:

userid	gender	Total_Likes_Given
1684195	male	25111
1656477	male	21652
1489463	female	16732
1429178	female	16583
1267229	female	14799
1783264	male	14355
1002588	female	14050
1412849	female	14039
1878566	female	13692
2104503	female	13622

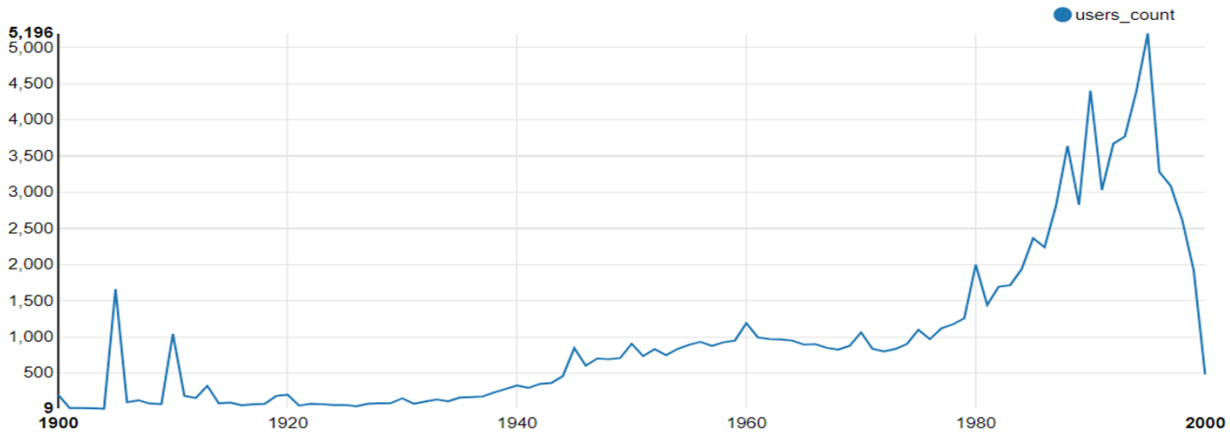
Jumlah mobile like:

SELECT gender,avg(mobile_likes) AS mobile_likes_given, avg(mobile_likes_received) AS mobile_likes_received, avg(www_likes) AS www_likes_given, avg(www_likes_received) AS www_likes_received FROM fb WHERE gender <> "NA" GROUP BY gender

gender	mobile_likes_given	mobile_likes_received	www_likes_given	www_likes_received
female	172.91293	147.10088	87.1383	104.33445
male	60.26133	40.83301	24.41655	27.07853

Data kelahiran dari user:

SELECT dob_year,count(userid) AS users_count FROM fb GROUP BY dob_year



Analisa Hasil

Dari hasil yang terlihat, beberapa kesimpulan dapat diambil:

- Perempuan mendapatkan dan memberi like lebih banyak dibandingkan laki-laki
- Jumlah user terbanyak berada di rentang 15-28
- Like yang diberikan lebih banyak melalui aplikasi mobile dibadingkan aplikasi web

Kesimpulan

Dengan menggunakan Hadoop maka kita bisa melakukan analisis terhadap dataset yang berukuran begitu besar, dalam hal ini contohnya adalah dataset Facebook. Dengan menggunakan hadoop dengan tools-tools yang dikembangkan diatasnya, kita bisa mendapatkan hasil yang diiginkan yaitu *intelligent decision* yang berguna untuk bisnis kedepannya.