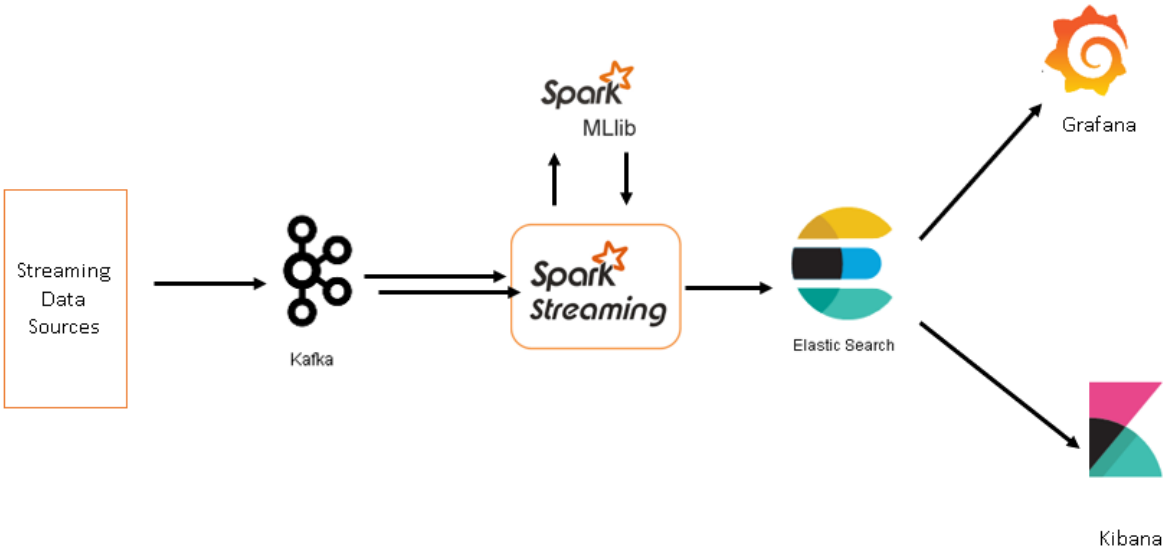
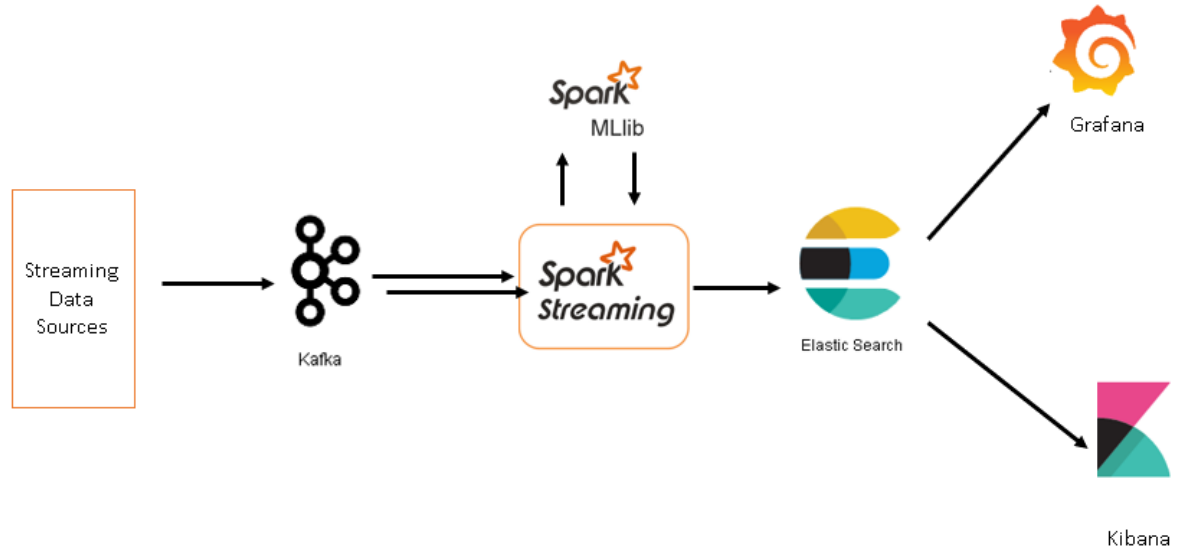
BÁO CÁO LAB5 LƯU TRỮ XỬ LÝ DỮ LIỆU LỚN : PHÂN TÍCH, XỬ LÝ DỮ LIỆU SẢN PHẨM TIKI



Nhóm	MTHH	
Thành viên	Nguyễn Quang Huy	20183554
	Trần Quang Minh	20183594
	Ngô Song Việt Hoàng	20183542
	Nguyễn Văn Thanh	20183632
GVHD	TS. Đào Thành Chung	

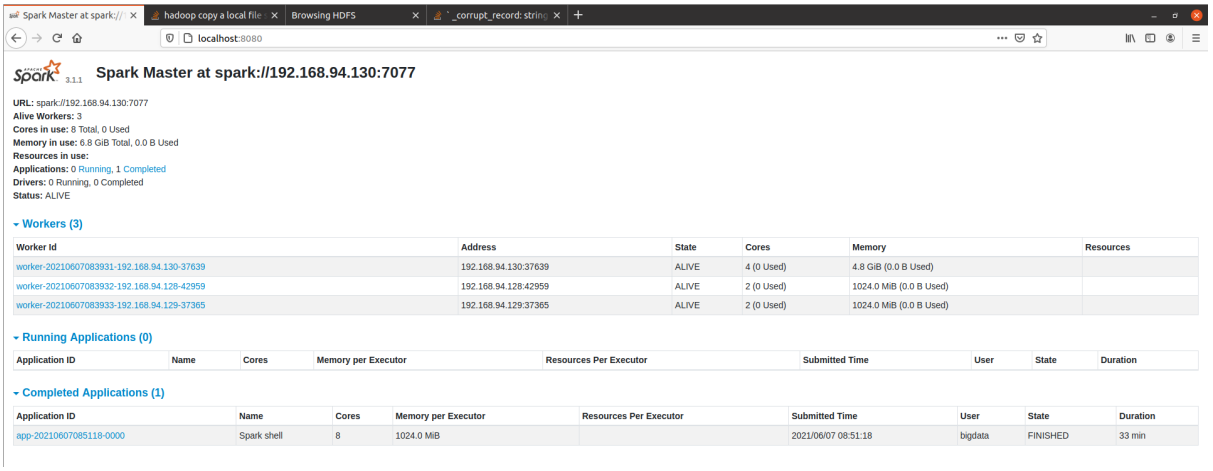
1. Pipeline của bài toán



Bài toán: Crawl dữ liệu từ tiki và thực hiện phân tích, đánh giá các sản phẩm này.

Luồng dữ liệu của bài toán: dữ liệu được crawl từ trang tiki sẽ gửi vào hàng đợi kafka, spark streaming sẽ đọc dữ liệu từ kafka sau đó lưu vào HDFS và elasticseach. Thực hiện phân tích dữ liệu đã được lưu tại elasticsearch qua kibana và kiểm tra lại bằng spark SQL. Thực hiện thuật toán Kmeans sử dụng thư viện Spark Mlib phân chia lại nhãn cho các sản phẩm.

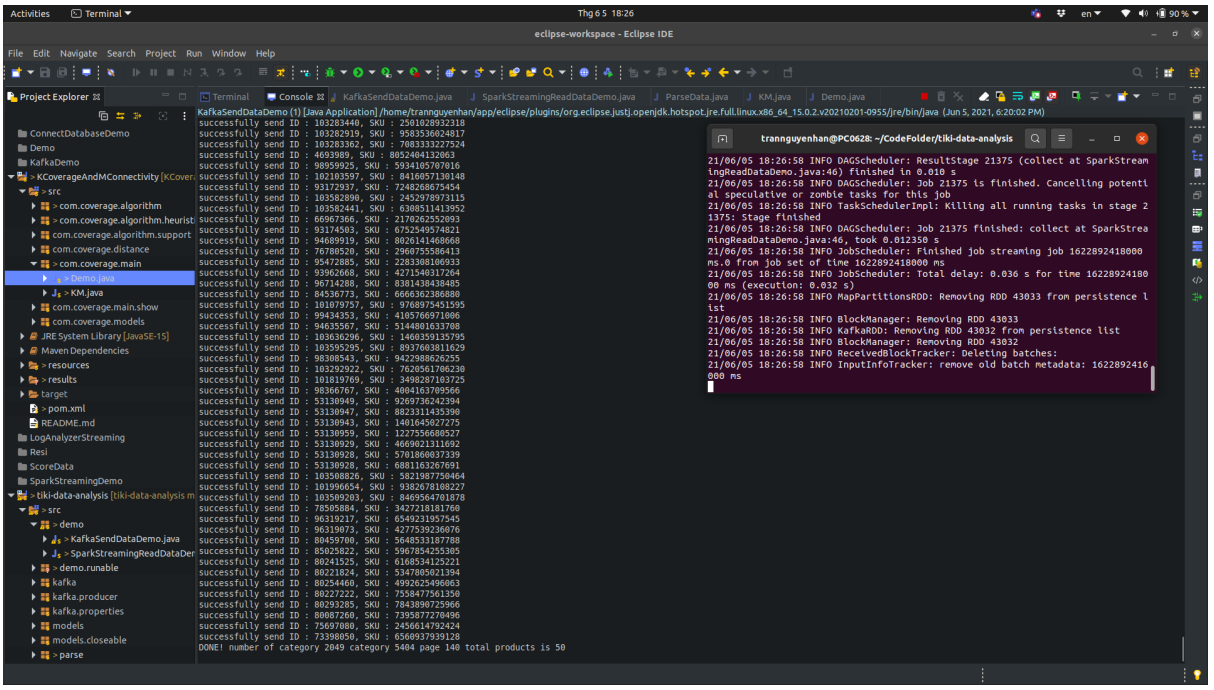
Cấu hình cụm Spark trên 3 máy:

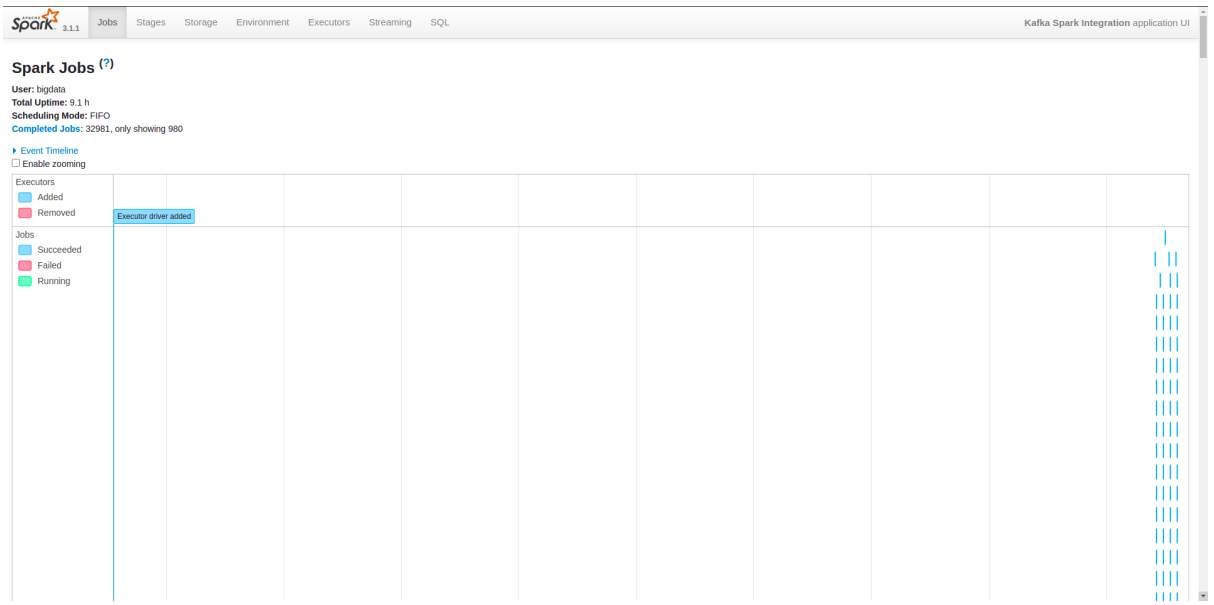


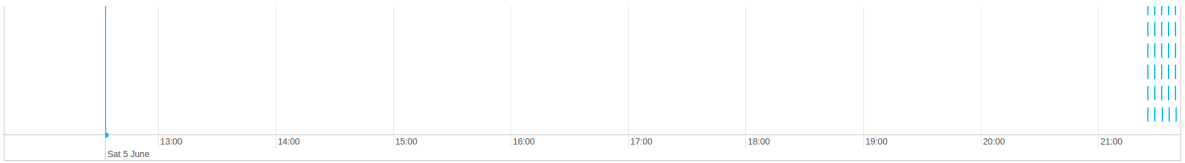
2. Crawl dữ liệu từ tiki

Dữ liệu của bài toán được crawl từ tiki sau đó được đẩy vào kafka, Spark Streaming sẽ lấy dữ liệu từ Kafka sau đó lưu vào HDFS (để lưu trữ), nếu máy có tài nguyên thì sẽ lưu vào Elasticsearch (tuy nhiên thì elasticsearch cần tài nguyên khá lớn, nên khó có thể chạy đồng thời Kafka, Hadoop, Spark, Elasticsearch nên nhóm e sẽ tách rời 2 bước này ra để có đủ tài nguyên phục vụ cho bài toán).

Hình ảnh Job Spark Streaming crawl dữ liệu từ tiki :







Completed Jobs (32981, only showing 980)

Page: 1 2 3 4 5 6 7 8 9 10 >

10 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
32979	Streaming job from [output operation 0, batch time 21:39:07] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:07	8 ms	1/1	1/1
32978	Streaming job from [output operation 0, batch time 21:39:06] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:06	9 ms	1/1	1/1
32977	Streaming job from [output operation 0, batch time 21:39:05] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:05	6 ms	1/1	1/1
32976	Streaming job from [output operation 0, batch time 21:39:04] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:04	18 ms	1/1	1/1
32975	Streaming job from [output operation 0, batch time 21:39:03] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:03	6 ms	1/1	1/1
32974	Streaming job from [output operation 0, batch time 21:39:02] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:02	12 ms	1/1	1/1
32973	Streaming job from [output operation 0, batch time 21:39:01] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:01	10 ms	1/1	1/1
32972	Streaming job from [output operation 0, batch time 21:39:00] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:00	13 ms	1/1	1/1
32971	Streaming job from [output operation 0, batch time 21:38:59] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:38:59	9 ms	1/1	1/1
32970	Streaming job from [output operation 0, batch time 21:38:58] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:38:58	22 ms	1/1	1/1
32969	Streaming job from [output operation 0, batch time 21:38:57] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:38:57	5 ms	1/1	1/1

Completed Jobs (32981, only showing 980)

Page: 1 2 3 4 5 6 7 8 9 10 >

10 Pages. Jump to 1 . Show 100 items in a page. Go

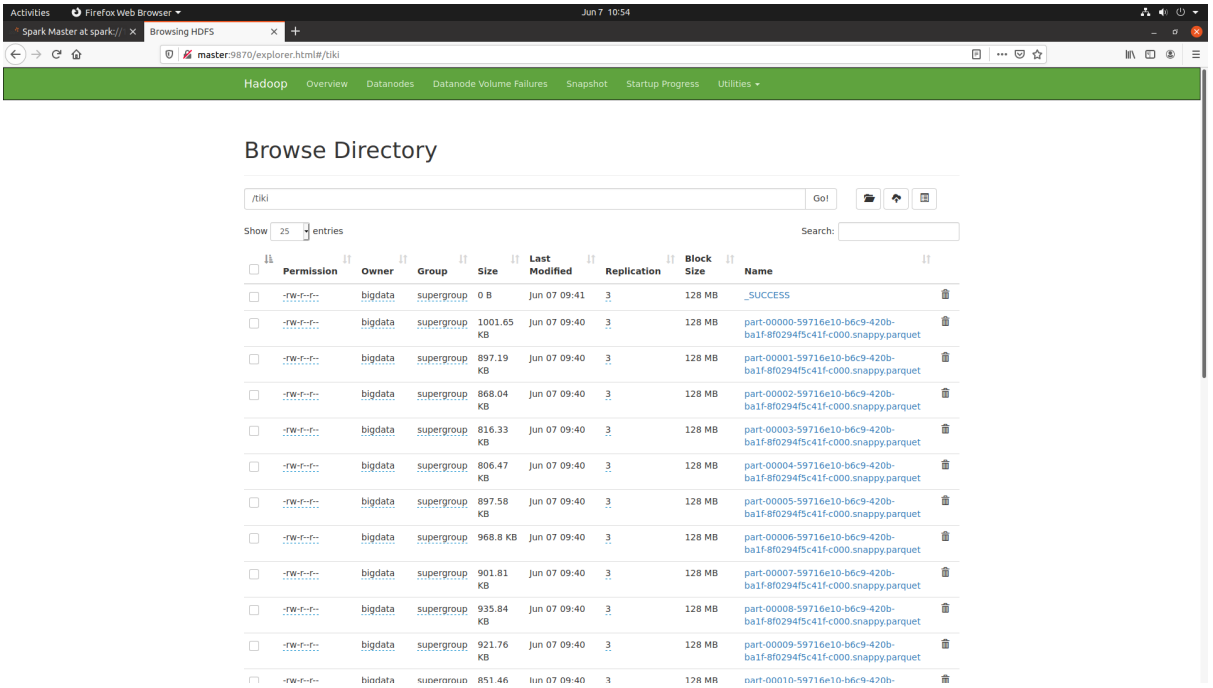
Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
32979	Streaming job from [output operation 0, batch time 21:39:07] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:07	8 ms	1/1	1/1
32978	Streaming job from [output operation 0, batch time 21:39:06] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:06	9 ms	1/1	1/1
32977	Streaming job from [output operation 0, batch time 21:39:05] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:05	6 ms	1/1	1/1
32976	Streaming job from [output operation 0, batch time 21:39:04] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:04	18 ms	1/1	1/1
32975	Streaming job from [output operation 0, batch time 21:39:03] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:03	6 ms	1/1	1/1
32974	Streaming job from [output operation 0, batch time 21:39:02] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:02	12 ms	1/1	1/1
32973	Streaming job from [output operation 0, batch time 21:39:01] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:01	10 ms	1/1	1/1
32972	Streaming job from [output operation 0, batch time 21:39:00] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:39:00	13 ms	1/1	1/1
32971	Streaming job from [output operation 0, batch time 21:38:59] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:38:59	9 ms	1/1	1/1
32970	Streaming job from [output operation 0, batch time 21:38:58] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:38:58	22 ms	1/1	1/1
32969	Streaming job from [output operation 0, batch time 21:38:57] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:38:57	5 ms	1/1	1/1
32968	Streaming job from [output operation 0, batch time 21:38:56] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:38:56	18 ms	1/1	1/1
32967	Streaming job from [output operation 0, batch time 21:38:55] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:38:55	6 ms	1/1	1/1
32966	Streaming job from [output operation 0, batch time 21:38:54] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:38:54	9 ms	1/1	1/1
32965	Streaming job from [output operation 0, batch time 21:38:53] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:38:53	9 ms	1/1	1/1

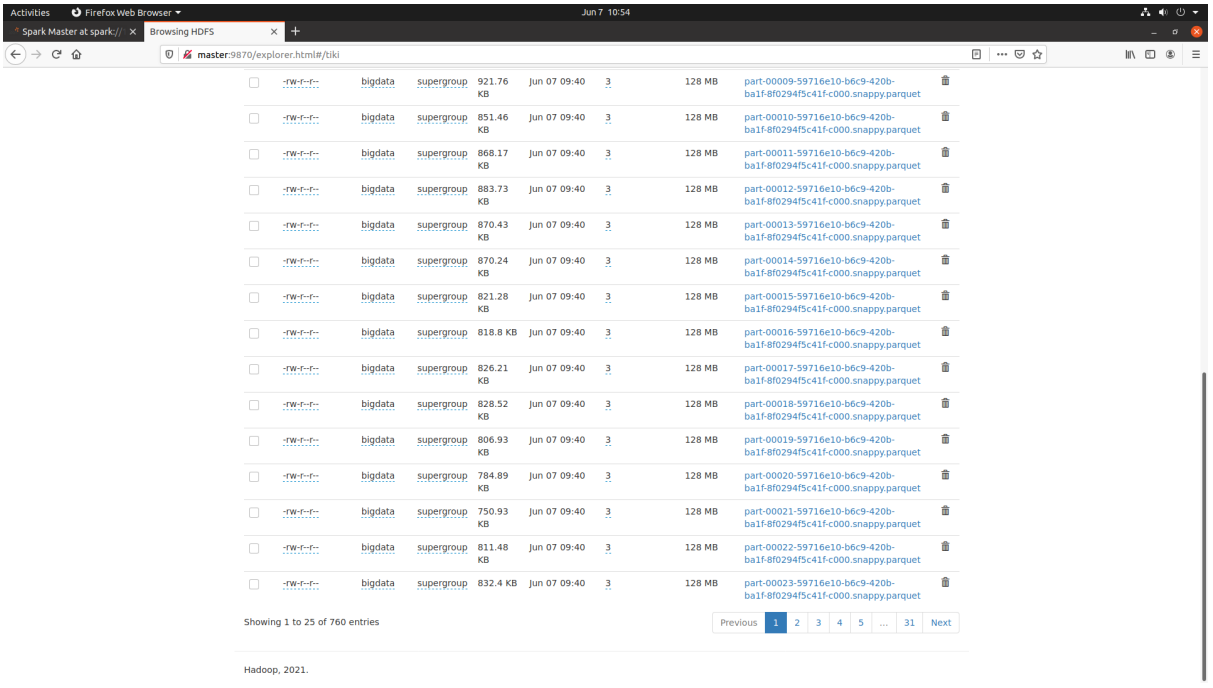
32896	Streaming job from [output operation 0, batch time 21:37:44] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:44	17 ms	1/1	1/1
32895	Streaming job from [output operation 0, batch time 21:37:43] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:43	13 ms	1/1	1/1
32894	Streaming job from [output operation 0, batch time 21:37:42] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:42	22 ms	1/1	1/1
32893	Streaming job from [output operation 0, batch time 21:37:41] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:41	25 ms	1/1	1/1
32892	Streaming job from [output operation 0, batch time 21:37:40] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:40	14 ms	1/1	1/1
32891	Streaming job from [output operation 0, batch time 21:37:39] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:39	32 ms	1/1	1/1
32890	Streaming job from [output operation 0, batch time 21:37:38] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:38	14 ms	1/1	1/1
32889	Streaming job from [output operation 0, batch time 21:37:37] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:37	11 ms	1/1	1/1
32888	Streaming job from [output operation 0, batch time 21:37:36] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:36	8 ms	1/1	1/1
32887	Streaming job from [output operation 0, batch time 21:37:35] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:35	4 ms	1/1	1/1
32886	Streaming job from [output operation 0, batch time 21:37:34] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:34	9 ms	1/1	1/1
32885	Streaming job from [output operation 0, batch time 21:37:33] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:33	4 ms	1/1	1/1
32884	Streaming job from [output operation 0, batch time 21:37:32] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:32	32 ms	1/1	1/1
32883	Streaming job from [output operation 0, batch time 21:37:31] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:31	10 ms	1/1	1/1
32882	Streaming job from [output operation 0, batch time 21:37:30] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:30	8 ms	1/1	1/1
32881	Streaming job from [output operation 0, batch time 21:37:29] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:29	8 ms	1/1	1/1
32880	Streaming job from [output operation 0, batch time 21:37:28] collect at SparkStreamingReadDataDemo.java:46	2021/06/05 21:37:28	6 ms	1/1	1/1

Page: 1 2 3 4 5 6 7 8 9 10 >

10 Pages. Jump to 1 . Show 100 items in a page. Go

Hình ảnh file được lưu về trên HDFS :





Tổng số dữ liệu (số sản phẩm craw từ tiki):

```
scala> data.count()
res23: Long = 1136517
```

Dung lượng lưu dưới định dạng file parquet của tất cả sản phẩm :

```
bigdata@master:~$ hdfs dfs -du -s -h /tiki
437.5 M  1.3 G  /tiki
```

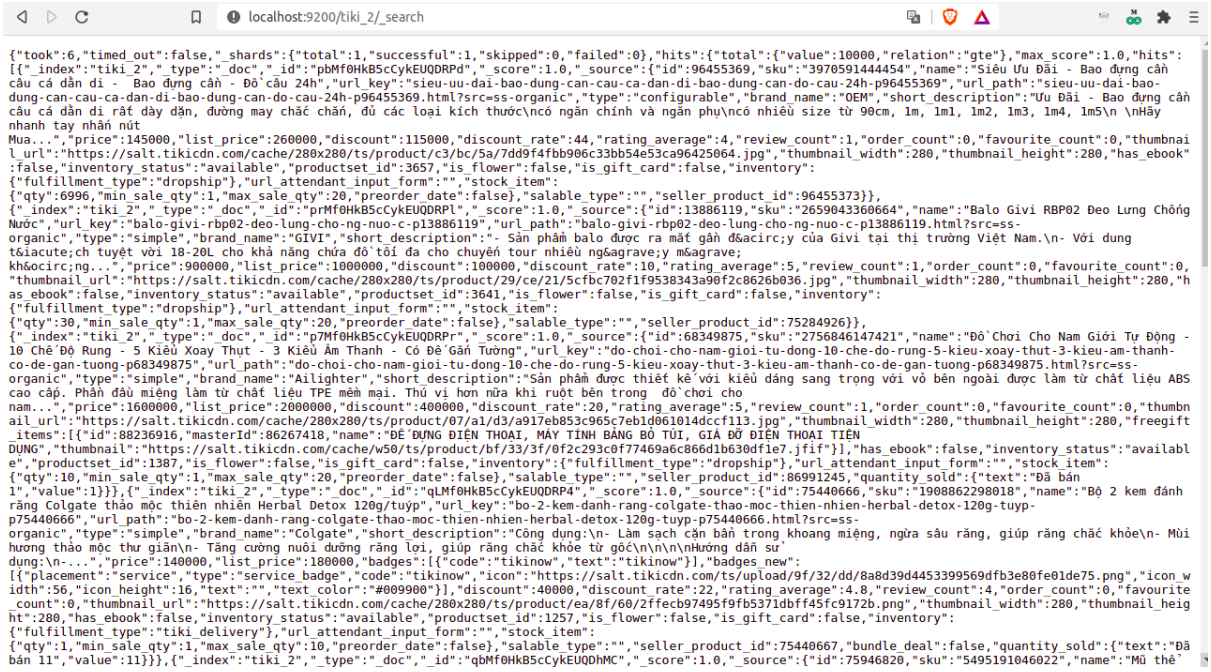
Dung lượng lưu dưới định dạng file json của 1,5 triệu sản phâm :

```
bigdata@master:~$ hdfs dfs -du -s -h /tiki-json
2.2 G  4.5 G  /tiki-json
```

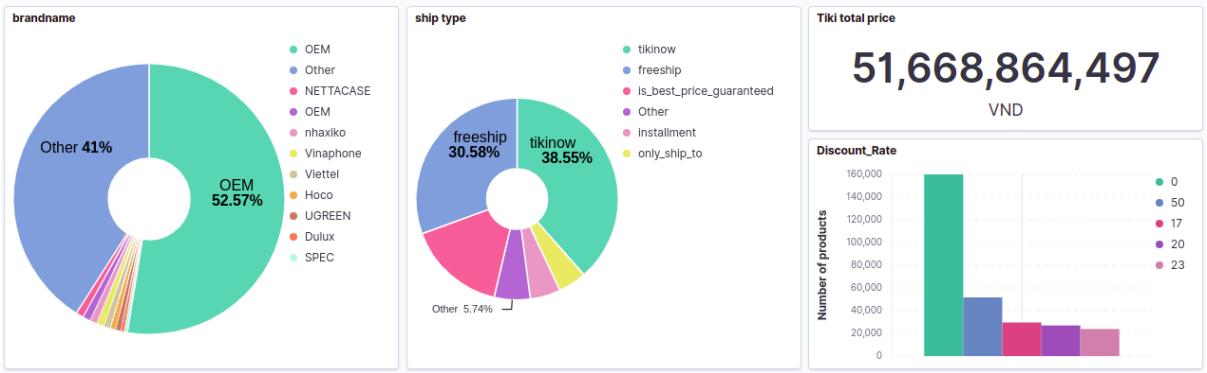
3. Visualization dữ liệu với elasticsearch và kibana

Dữ liệu được đẩy lên Elasticsearch sau đó sẽ visualization bằng kibana.

Tổng số sản phẩm được đẩy lên Elasticsearch :



Dữ liệu được visual theo 1 số trường bằng kibana như sau:

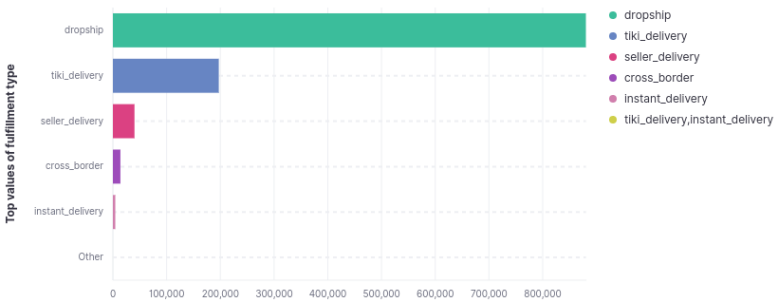


popular_product

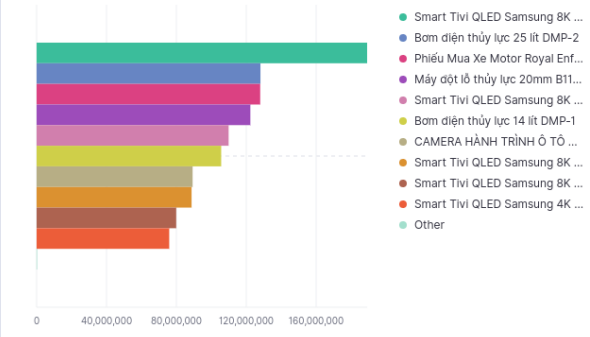
ỐP LƯNG IN DÀNH CHO VSMART BEE
ỐP LƯNG IN DÀNH CHO NOKIA 5.1 PLUS
ỐP LƯNG IN DÀNH CHO NOKIA 2.2
ỐP LƯNG IN DÀNH CHO SONY M4
ỐP LƯNG IN DÀNH CHO NOKIA 8.1
Gối tựa lưng, gối trang trí sofa
ỐP LƯNG IN DÀNH CHO VIVO Y11

Products - Count

fullfillment_type



Top 10 discount price



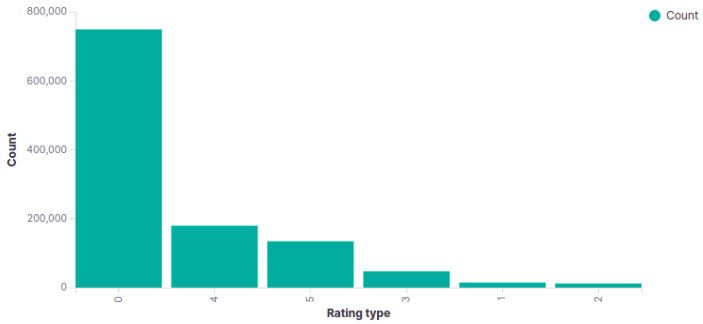
Top 10 expensive

inpopular_product

Đèn ngủ gỗ để bàn MB1508A
Xe trượt Scooter 0072
Lô 2 Cước Cọ Rửa Inox Loving Home
Bột sữa hạt Đòng Đòng Vàng 100gr
✓Béc phun mưa tưới rau
Dụng cụ bào gỗ

Products - Count

rating



4. Kiểm tra lại dữ liệu trên với Spark SQL

Tổng giá trị toàn bộ sản phẩm :

```
scala> data.createOrReplaceTempView("data")

scala> spark.sql("select sum(price) from data").show()
+--------------+
|   sum(price)|
+--------------+
|1353625672528|
+--------------+
```

Top điểm rate được sử dụng nhiều nhất :

```
scala> spark.sql("select discount_rate, count(discount_rate) from data group by discount_rate order by count(discount_rate) desc").show()
+-------------+--------------------+
|discount_rate|count(discount_rate)|
+-------------+--------------------+
|            0|              160103|
|           50|               51727|
|           17|               29575|
|           20|               26988|
|           23|               23862|
|           47|               21287|
|           10|               20553|
|           40|               20332|
|           33|               19540|
|            5|               19249|
|           34|               19160|
|           30|               19034|
|           25|               18116|
|            9|               16983|
|           29|               16449|
|           21|               16306|
|           24|               16270|
|           22|               16225|
|           53|               15535|
|           38|               15449|
+-------------+--------------------+
only showing top 20 rows
```

Top brand name xuất hiện nhiều nhất :

```
scala> spark.sql("select brand_name, count(brand_name) from data group by brand_name order by count(brand_name) desc").show()
+--------------------+-----------------+
|          brand_name|count(brand_name)|
+--------------------+-----------------+
|                 OEM|           599013|
|           NETTACASE|            10999|
|                 OEM|            10781|
|             nhaxiko|            10013|
|           Vinaphone|             9983|
|             Viettel|             9642|
|                Hoco|             7250|
|              UGREEN|             6829|
|               Dulux|             4459|
|                SPEC|             4373|
|               ICASE|             4236|
|           Bát Tràng|             4208|
|              Baseus|             3709|
|               Remax|             3559|
|PHƯƠNG NAM VINANUMIS|             3477|
|                IBIE|             3417|
|               NETTA|             3213|
|                OLSG|             2943|
|            Handtown|             2476|
|           Panasonic|             2412|
+--------------------+-----------------+
only showing top 20 rows
```

Top sản phẩm được giảm giá nhiều nhất :

```
scala> spark.sql("select distinct name, discount from data order by discount desc").show()
+--------------------+---------+
|                name| discount|
+--------------------+---------+
|Smart Tivi QLED S...|189210000|
|Bơm điện thủy lực...|128136000|
|Phiếu Mua Xe Moto...|128000000|
|Máy đột lỗ thủy l...|122360000|
|Smart Tivi QLED S...|109910000|
|Bơm điện thủy lực...|105659000|
|CAMERA HÀNH TRÌNH...| 89400000|
|CAMERA HÀNH TRÌNH...| 89100000|
|Smart Tivi QLED S...| 88710000|
|Smart Tivi QLED S...| 79910000|
|Smart Tivi QLED S...| 75910000|
|Tủ lạnh Panasonic...| 70110000|
|Tủ lạnh Panasonic...| 70010000|
|Phiếu Mua Xe Máy ...| 69400000|
|Smart Tivi QLED S...| 64910000|
|Smart Tivi Neo QL...| 64000000|
|Bàn ăn 4 ghế mode...| 62000000|
|Hệ Thống Lọc Nước...| 61000000|
|Tủ lạnh SBS 3 cán...| 59001000|
|Smart Tivi QLED S...| 58281000|
+--------------------+---------+
only showing top 20 rows
```

Top sản phẩm có giá trị cao nhất:

```
scala> spark.sql("select distinct name, price from data order by price desc").show()
+--------------------+---------+
|                name|    price|
+--------------------+---------+
|Xe Mô Tô Harley D...|809500000|
|Xe Môtô Triumph S...|649000000|
|Xe Môtô Triumph S...|623000000|
|Xe Môtô Triumph B...|619000000|
|Xe Môtô Triumph S...|619000000|
|Xe Môtô Triumph S...|595600000|
|Xe Mô Tô Harley D...|579100000|
|Xe Môtô Triumph B...|573000000|
|Xe Môtô Triumph B...|569000000|
|Xe Môtô Triumph S...|430000000|
|Xe Moto Kawasaki ...|412000000|
|Xe Môtô Triumph B...|410000000|
|Xe Môtô Triumph S...|405600000|
|Xe Môtô Triumph S...|403000000|
|Máy Làm Đá Vảy Hả...|392000000|
|Xe Máy KYMCO AK 5...|375000000|
|Xe Mô Tô Harley D...|365300000|
|Máy Pha Cà Phê Ch...|339999000|
|Lens Canon EF 800...|309308000|
|Phiếu mua xe máy ...|300000000|
+--------------------+---------+
only showing top 20 rows
```

Top fullfilment sử dụng nhiều nhất:

```
scala> spark.sql("select inventory.fulfillment_type, count(inventory.fulfillment_type) from data group by inventory.fulfillment_type order by count(inventory.fulfillment_type) desc").show()
+--------------------+-------------------------------+
|    fulfillment_type|count(inventory.fulfillment_type)|
+--------------------+-------------------------------+
|            dropship|                         880430|
|       tiki_delivery|                         197108|
|     seller_delivery|                          40152|
|        cross_border|                          13953|
|    instant_delivery|                           4482|
|tiki_delivery,ins...|                             19|
|                null|                              0|
+--------------------+-------------------------------+
```

Top sản phẩm xuất hiện nhiều nhất:

```
scala> spark.sql("select name, count(name) from data group by name order by count(name) desc").show()
+--------------------+-----------+
|                name|count(name)|
+--------------------+-----------+
|Ộp Lưng Dành Cho ...|        427|
|Ộp Lưng Dẻo Dành ...|        381|
|Ộp Dẻo Dành Cho N...|        260|
|ỘP LƯNG DẺO DÀNH ...|        214|
|Ộp Lưng Dẻo Dành ...|        184|
|Ộp Lưng Dành Cho ...|        154|
|Ộp Lưng Dẻo Dành ...|        138|
|ỘP LƯNG IN HÌNH D...|        131|
|ỘP LƯNG IN HÌNH D...|        131|
|Ộp Dẻo Dành Cho X...|        128|
|Ộp Lưng Dành Cho ...|        119|
|ỘP LƯNG IN HÌNH D...|        119|
|Ộp lưng Dành Cho ...|        117|
|Ộp Lưng Dành Cho ...|        116|
|ỘP LƯNG IN HÌNH D...|        115|
|Ộp Lưng Dẻo Dành ...|        112|
|Ộp Lưng Dành Cho ...|        111|
|ỘP LƯNG IN HÌNH D...|        110|
|Ộp Lưng Dành Cho ...|        105|
|ỘP LƯNG IN HÌNH D...|        104|
+--------------------+-----------+
only showing top 20 rows
```

Top phương pháp vận chuyển sử dụng nhiều nhất:

```
scala> spark.sql("select badges.code, count(badges.code) from data group by badges.code order by count(badges.code) desc").show()
+--------------------+--------------------------+
|                code|count(badges.code AS `code`)|
+--------------------+--------------------------+
|           [tikinow]|                    209800|
|      [only_ship_to]|                     16098|
|      [cross_border]|                     13459|
|       [installment]|                     10636|
|[only_ship_to, in...|                      8033|
|[tikinow, install...|                      5196|
|[tikipro, fast_de...|                      2396|
|[tikipro, fast_de...|                       880|
|[tikipro, fast_de...|                       444|
|[cross_border, ti...|                       408|
|[tikinow, only_sh...|                       101|
|[tikipro, fast_de...|                        67|
|[cross_border, in...|                        33|
|[fast_delivery, t...|                         9|
|[tikinow, only_sh...|                         9|
|         [tikifresh]|                         8|
|[tikifresh, only_...|                         2|
| [tikinow, tikilive]|                         2|
|                null|                         0|
+--------------------+--------------------------+
```

5. Thực hiện phân cụm sản phẩm với thuật toán KMeans

Thực hiện thuật toán phân cụm với K-means, dữ liệu đầu vào là json:

```
%pyspark
from pyspark.sql.functions import lower, col, split

df = spark.read.json("file:///usr/zeppelin/notebook/dataset/json-tiki-data/*.json")
df.printSchema()
```

```
root
 |-- brandName: string (nullable = true)
 |-- category: long (nullable = true)
 |-- discount: double (nullable = true)
 |-- discountRate: long (nullable = true)
 |-- favouriteCount: long (nullable = true)
 |-- flower: boolean (nullable = true)
 |-- giftCard: boolean (nullable = true)
 |-- hasEbook: boolean (nullable = true)
 |-- id: long (nullable = true)
 |-- inventory: struct (nullable = true)
 |    |-- fulfillmentType: string (nullable = true)
 |-- inventoryStatus: string (nullable = true)
 |-- listPrice: double (nullable = true)
 |-- name: string (nullable = true)
 |-- orderCount: long (nullable = true)
 |-- price: double (nullable = true)
 |-- productsetId: long (nullable = true)
```

Dữ liệu sản phẩm sẽ được phân cụm sử dụng tên của sản phẩm, tiền xử lý dữ liệu thực hiện chuẩn hóa chữ hoa thành chữ thường, chuyển xâu thành list các từ,...:

```
%pyspark
dataset = df.select("id", "name")

dataset = dataset.select("id", lower(col("name")))
dataset = dataset.withColumnRenamed("lower(name)", "name")

dataset = dataset.select("id", split(dataset.name, " "))
dataset = dataset.withColumnRenamed("split(name,  )", "name")
dataset.show()
dataset.count()
```

```
+--------+--------------------+
|      id|                name|
+--------+--------------------+
|23029063|[combo, sách, kin...|
|40981321|[marketing, trong...|
|48585379|[combo, 2, cuốn, ...|
|41372499|[30, ngày, giải, ...|
| 7459867|[combo, nghệ, thu...|
|58232528|[combo, sách, :, ...|
|25147907|[combo, sách, kỹ,...|
|29815359|[10, điều, khác, ...|
|75482723|[bí, quyết, warre...|
|30156463|[tâm, lý, thị, tr...|
|36702049|[mã, vân, giày, v...|
|68130540|[combo, 2, cuốn, ...|
|38517971|[tủ, sách, hay, d...|
|35787156|[tảng, băng, tan,...|
|68134202|[combo, 2, cuốn, ...|
```

Bước tiếp theo sẽ đi vector hóa các đặc trưng này của mỗi sản phẩm sử dụng Spark MLlib:

```
%pyspark
from pyspark.ml.feature import Word2Vec

word2Vec = Word2Vec(vectorSize=4, minCount=0, inputCol="name", outputCol="features")
model = word2Vec.fit(dataset)

result = model.transform(dataset)
result.show()
```

```
+--------+--------------------+--------------------+
|      id|                name|            features|
+--------+--------------------+--------------------+
|23029063|[combo, sách, kin...|[-0.0098653077215...|
|40981321|[marketing, trong...|[0.01220525514621...|
|48585379|[combo, 2, cuốn, ...|[-0.0406815864677...|
|41372499|[30, ngày, giải, ...|[0.04706937345591...|
| 7459867|[combo, nghệ, thu...|[-0.0103190138936...|
|58232528|[combo, sách, :, ...|[-0.1813816154266...|
|25147907|[combo, sách, kỹ,...|[-0.0903325371244...|
|29815359|[10, điều, khác, ...|[-0.0823144884232...|
|75482723|[bí, quyết, warre...|[0.25975030826197...|
|30156463|[tâm, lý, thị, tr...|[0.13589337312926...|
|36702049|[mã, vân, giày, v...|[-0.3250958994030...|
|68130540|[combo, 2, cuốn, ...|[-0.1393234546367...|
|38517971|[tủ, sách, hay, d...|[-0.0509911202603...|
|35787156|[tảng, băng, tan,...|[-0.1932214153930...|
|68134202|[combo, 2, cuốn, ...|[1.87554334600766...|
```

Sử dụng Spark MLlib để phân cụm với K-means:

```
%pyspark
transformed.show()
```

```
+--------+--------------------+--------------------+----------+
|      id|                name|            features|prediction|
+--------+--------------------+--------------------+----------+
|23029063|[combo, sách, kin...|[-0.0098653077215...|        12|
|40981321|[marketing, trong...|[0.01220525514621...|        12|
|48585379|[combo, 2, cuốn, ...|[-0.0406815864677...|        12|
|41372499|[30, ngày, giải, ...|[0.04706937345591...|        12|
| 7459867|[combo, nghệ, thu...|[-0.0103190138936...|        12|
|58232528|[combo, sách, :, ...|[-0.1813816154266...|        12|
|25147907|[combo, sách, kỹ,...|[-0.0903325371244...|        12|
|29815359|[10, điều, khác, ...|[-0.0823144884232...|        12|
|75482723|[bí, quyết, warre...|[0.25975030826197...|        12|
|30156463|[tâm, lý, thị, tr...|[0.13589337312926...|        12|
|36702049|[mã, vân, giày, v...|[-0.3250958994030...|         0|
|68130540|[combo, 2, cuốn, ...|[-0.1393234546367...|        12|
|38517971|[tủ, sách, hay, d...|[-0.0509911202603...|        12|
|35787156|[tảng, băng, tan,...|[-0.1932214153930...|         0|
|68134202|[combo, 2, cuốn, ...|[1.87554334600766...|        12|
```

Took 0 sec. Last updated by anonymous at June 08 2021, 2:03:06 AM.

```
%pyspark
from pyspark.sql.functions import concat_ws

output = transformed.withColumn("name", concat_ws(" ", "name"))

output = output.drop("features")
output.show()
```

```
+--------+--------------------+----------+
|      id|                name|prediction|
+--------+--------------------+----------+
|23029063|combo sách kinh t...|        12|
|40981321|marketing trong c...|        12|
|48585379|combo 2 cuốn sách...|        12|
|41372499|30 ngày giải mã k...|        12|
| 7459867|combo nghệ thuật ...|        12|
|58232528|combo sách : nghệ...|        12|
|25147907|combo sách kỹ năn...|        12|
|29815359|10 điều khác biệt...|        12|
|75482723|bí quyết warren b...|        12|
|30156463|tâm lý thị trường...|        12|
|36702049|mã vân giày vải (...|         0|
|68130540|combo 2 cuốn sách...|        12|
|38517971|tủ sách hay dành ...|        12|
|35787156|tảng băng tan (tặ...|         0|
|68134202|combo 2 cuốn sách...|        12|
```

Took 0 sec. Last updated by anonymous at June 08 2021, 2:25:02 AM.

```
%pyspark
output.write.json("file:///usr/zeppelin/notebook/dataset/kmeans-output")
```

Took 2 sec. Last updated by anonymous at June 08 2021, 2:25:50 AM.

```
%pyspark
from pyspark.ml.clustering import KMeans

kmeans = KMeans().setK(16).setSeed(1)
model = kmeans.fit(result.select("features"))
transformed = model.transform(result)
```

Took 4 sec. Last updated by anonymous at June 08 2021, 2:03:04 AM.