SQL в анализе данных 101

Решаем прикладную задачу сегментации

План

- Немного теории про Spark и RFM
- Разбираем код для решения на Spark SQL в Databricks
- Строим дашбордик и публикуем в public.tableau.com
- Обсуждаем вопросы
- Happy end

RFM_spark_sql_tableau (Python)

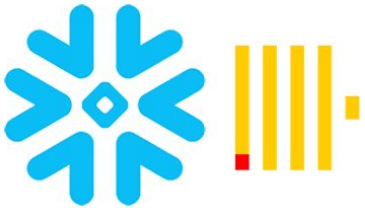
Spark_3_0_1 | File | Edit | View: Standard | Permissions | Run All | Clear

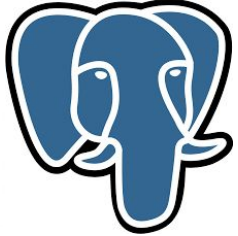
Command took 7.13 seconds -- by kudashev.eugene@gmail.com at 09/12/2020, 22:10:13 on Spark_3_0_1

Cmd 17

```
sql = """
WITH orders AS (
SELECT
  InvoiceNo
  , InvoiceDate
  , CustomerID
  , SUM(Quantity*UnitPrice) AS Price
  , MAX(InvoiceDate) OVER() + INTERVAL '1' DAY  AS Now
FROM
  sales
WHERE
  CustomerID IS NOT NULL
GROUP BY
  InvoiceNo
  , InvoiceDate
  , CustomerID
)
, base AS (
SELECT
  CustomerID
  , MIN(datediff(Now, InvoiceDate)) AS Recency
  , COUNT(DISTINCT InvoiceNo) AS Frequency
  , SUM (Price) AS Monetary
FROM
  orders
WHERE
```

Почему SQL?



- Традиционные БД (MS SQL, Postgres, MySQL, Oracle)
- NoSQL (MongoDB, Cassandra, Neo4j, Hadoop)
- MPP (Vertica, Exasol, BQ, Redshift, Greenplum)
- SQL Engines (Presto, Spark)
- ETL tools(dbt, luigi, airflow)
- BI Tools
- Other (ClickHouse, Kafka)

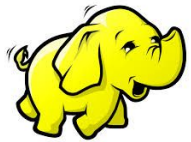




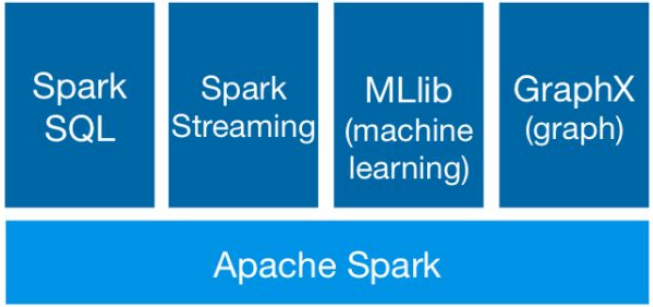








Spark



- Scala, Java, Python, R and SQL (Kotlin, clojure)
- https://spark.apache.org/
- Dataframe Api
- SQL (HIVE SQL or PG SQL)
- https://www.waitingforcode.com/apache-spark-sql/what-new-apache-spark-3-postgresql-feature-parity/read

Runs Everywhere

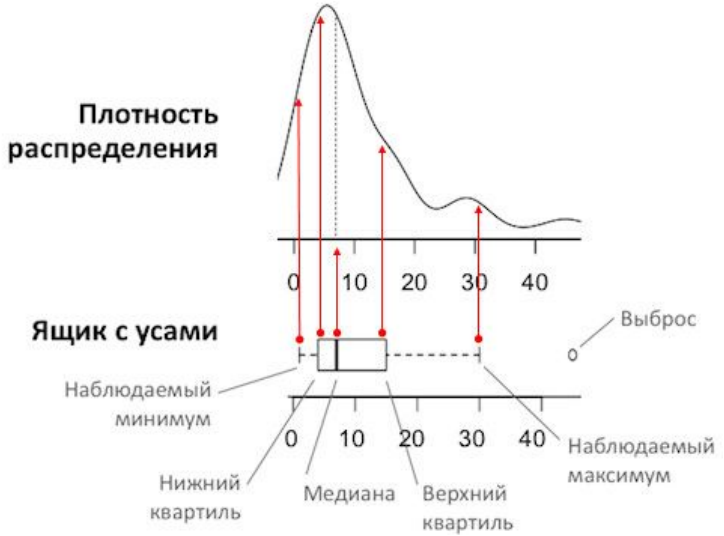
Spark runs on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud. It can access diverse data sources.

You can run Spark using its standalone cluster mode, on EC2, on Hadoop YARN, on Mesos, or on Kubernetes. Access data in HDFS, Alluxio, Apache Cassandra, Apache HBase, Apache Hive, and hundreds of other data sources.

RFM

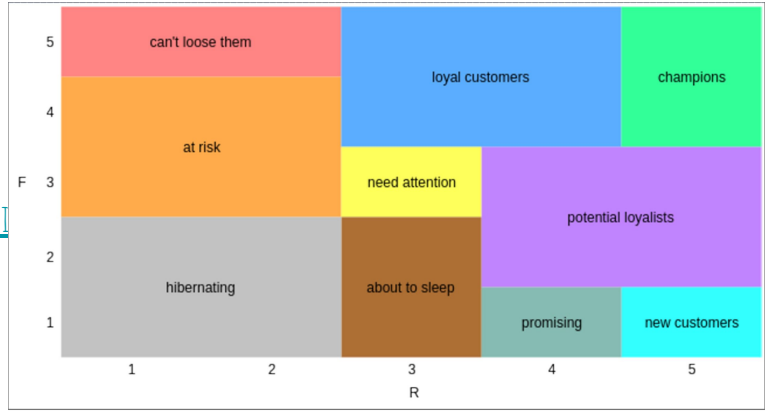
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01.12.2010 8:26:00	2.55	17850	United Kingdom
2	536365	71053	WHITE METAL LANTERN	6	01.12.2010 8:26:00	3.39	17850	United Kingdom
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01.12.2010 8:26:00	2.75	17850	United Kingdom
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01.12.2010 8:26:00	3.39	17850	United Kingdom
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01.12.2010 8:26:00	3.39	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01.12.2010 8:26:00	7.65	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01.12.2010 8:26:00	4.25	17850	United Kingdom

Showing the first 1000 rows.



- Recency (R) – давность последней покупки
- Frequency (F) – суммарная частота покупок
- Monetary (M) – объём покупок

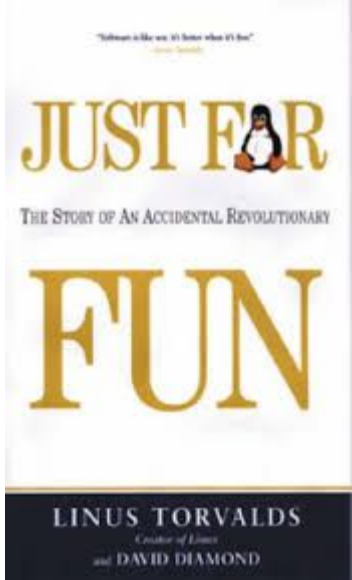
- https://sputnik.com/blog/rfm-segmentaciya
- https://community.tableau.com/s/question/0D54T00000G54TkSAB/rfm-segmentation
- https://medium.com/analytics-vidhya/marketing-analytics-rfm-modeling-855ebec18014
- https://habr.com/ru/company/plarium/blog/431520/
- https://towardsdatascience.com/find-your-best-customers-with-customer-segmentation-i
- https://towardsdatascience.com/dynamic-customer-analytics-i-9135b2a8854b
- https://guillaume-martin.github.io/rfm-segmentation-with-python.html



Что надо сделать чтоб начать?

- https://databricks.com/try-databricks or https://community.cloud.databricks.com/login.html
- https://public.tableau.com/s/
- https://public.tableau.com/en-us/s/download
-

Let's Code


JUST FOR
THE STORY OF AN ACCIDENTAL REVOLUTIONARY
FUN
LINUS TORVALDS
Creator of Linux
and DAVID DIAMOND

Что читать и смотреть?

- https://mode.com/sql-tutorial/
- https://mode.com/blog/finding-user-sessions-sql/
- https://exacademy.exasol.com/courses/course-v1:Exasol+ESSENTIALS+X/course/
- https://exacademy.exasol.com/courses/course-v1:Exasol+PERF+X/course/
- https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

Announcements

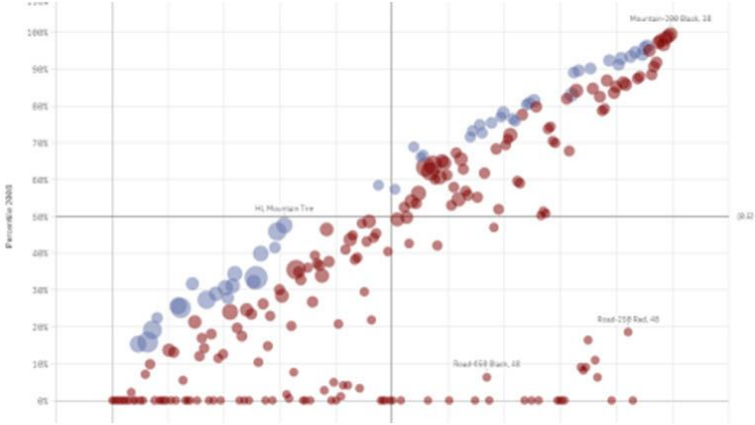
Free Exasol Certifications

Starting November 23rd, we have decided to offer free vouchers for Exasol Certifications permanently! To be eligible, you must score 70% on the Hands-on Exercises.
More Details

Use this opportunity to check out our Course Offering and get up to date on your Exasol certifications !

Что еще поделать?

- Добавить возможность расчета по странам (Rollup + Tableau filters)
- Попробовать Postgres SQL для Spark
- Перенести всю логику в Tableau + Динамические пороги для RFM
- http://blog.atkcg.ru/percentil-alternativnyj-vzglyad-na-dannye/
-



Спасибо что дотерпели до Happy End-а

Вопросы?