CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

Proyecto №4

Lenguajes y frameworks más utilizados

CC5212 - Procesamiento masivo de datos

Profesor:

- Aidan Hogan

Auxiliares:

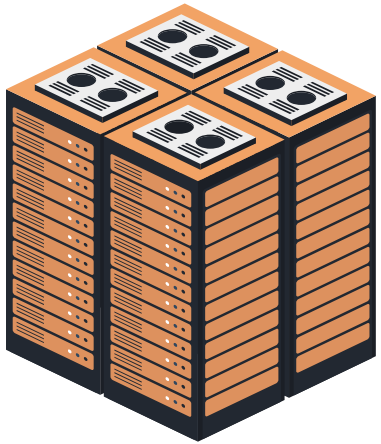
- Alberto Moya
- Felipe Manen

Integrantes:

- Benjamín Mellado
- Samuel Sánchez
- Alonso Vargas

dcc
CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

Objetivos



Motivacion principal

Utilizar herramienta vista durante el curso para tratar datos masivos

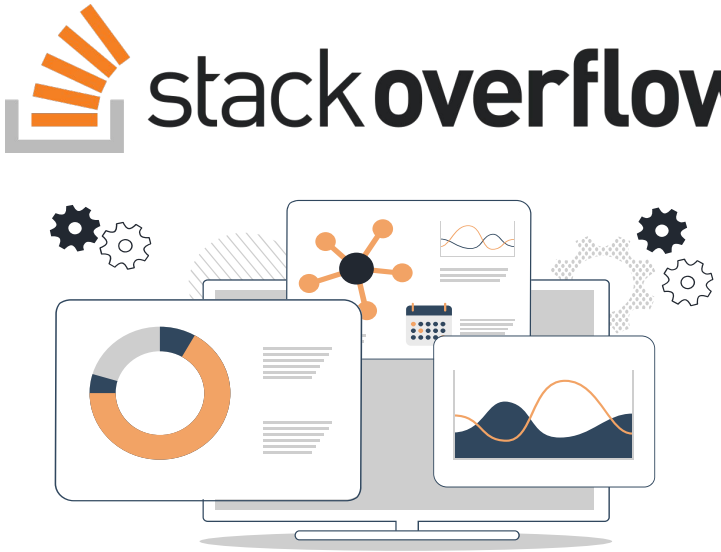
Objetivo del proyecto

Evaluar la evolución en la utilización de lenguajes, frameworks y otros.

Aprendizaje Esperado

Tratar con volúmenes masivos de datos utilizando Apache Spark en Python

Datos - 10% Stack Overflow



¿Qué es?

Sitio de preguntas y respuestas para programadores profesionales y aficionados.

Acerca de los datos

3 tablas en formato CSV con el texto del 10% de preguntas y respuestas de StackOverflow entre agosto del 2008 y octubre del 2016.

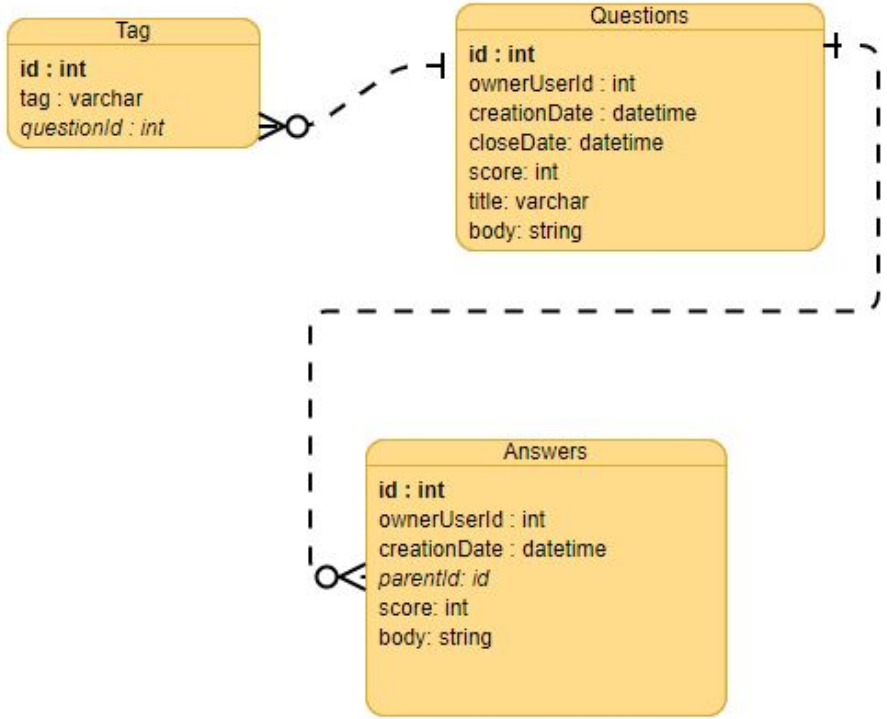
Tamaño y peso

El dataset completo pesa 3.35 GB

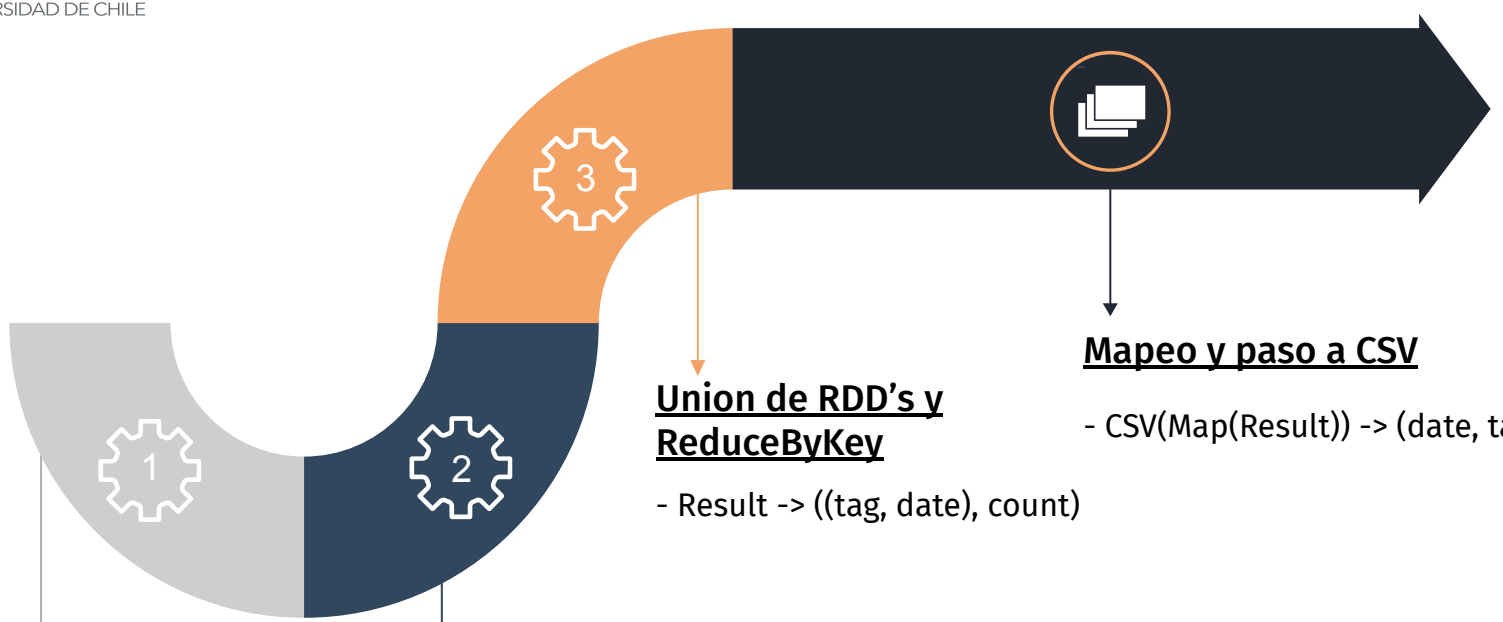
- Tags: 1.25 millones de filas
- Questions: 2.01 millones de filas
- Answers: 3.75 millones de filas

Datos - Modelo Entidad-Relación

- 3 entidades que representan los Tags, Preguntas y Respuestas
- Tag y Answers dependen de Questions y se conectan mediante foreign key a su id.



Metodología



Archivos a RDD

- Tags -> (id, tag)
- Questions -> (id, date)
- Answers -> (id, date)

Cruces de RDD's y mapeo para contar

- Map(Join(Questions, Tags)) -> ((tag, date), 1)
- Map(Join(Answers , Tags)) -> ((tag, date), 1)

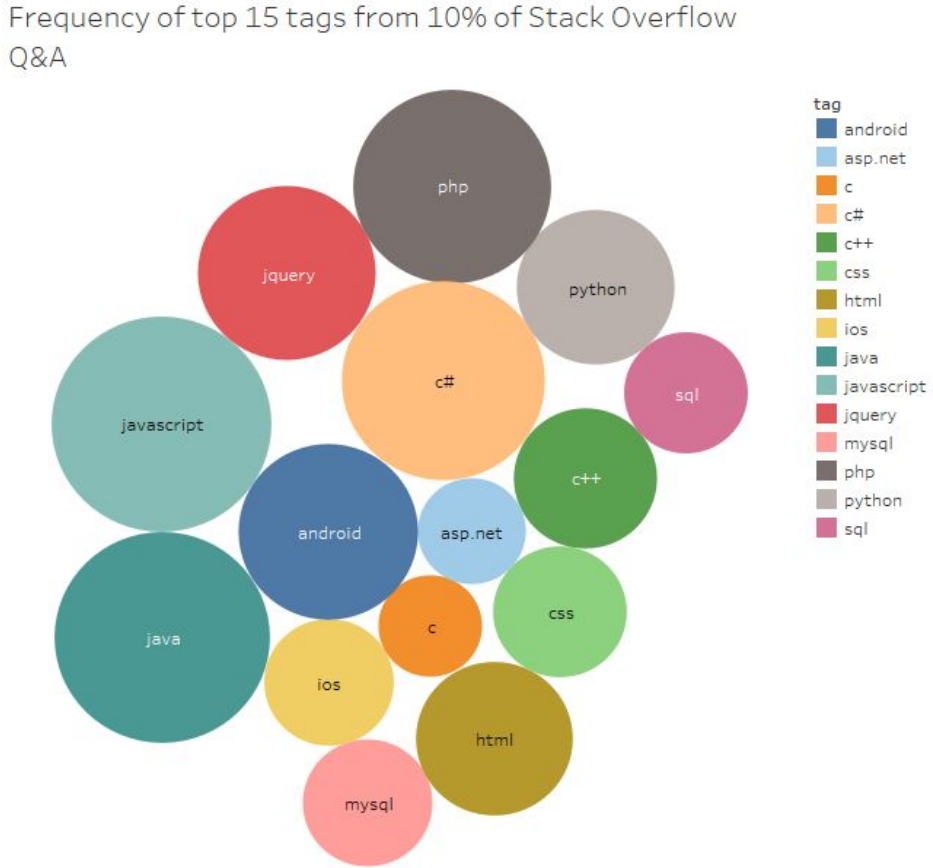
Union de RDD's y ReduceByKey

- Result -> ((tag, date), count)

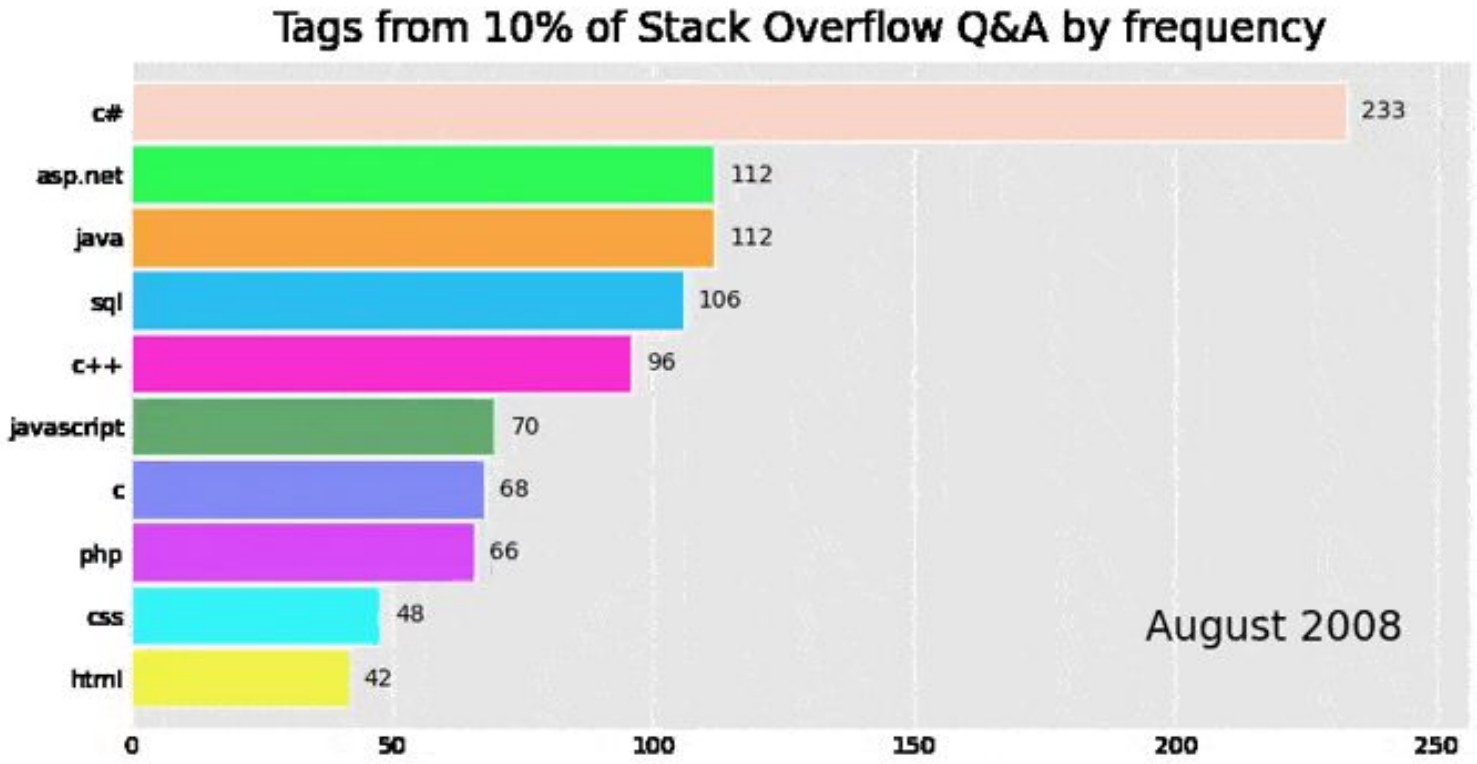
Mapeo y paso a CSV

- CSV(Map(Result)) -> (date, tag, count)

Resultados


Frequency of top 15 tags from 10% of Stack Overflow Q&A
tag
android
asp.net
c
c#
c++
css
html
ios
java
javascript
jquery
mysql
php
python
sql
php
jquery
python
c#
sql
javascript
c++
android
asp.net
java
c
css
ios
html
mysql
Tag. El color muestra detalles acerca de tag. El tamaño muestra suma de count. Las marcas se etiquetan por tag.

Resultados


Tags from 10% of Stack Overflow Q&A by frequency
c#
asp.net
java
sql
c++
javascript
c
php
css
html
233
112
112
106
96
70
68
66
48
42
August 2008
0
50
100
150
200
250

Conclusiones

01

Aprendizaje con respecto a la utilización del framework PySpark

02

Comprensión de las necesidades de frameworks para tratar datos masivos

03

Capacidad de trabajo en equipo en proyectos con datos

04

Proyecto extendible a muestras más grandes de stack overflow