Presentación Tarea 1

Integrantes:
Fernando Riveros
Víctor San Martín
Alejandro Vasquez
Pablo Vergara

Profesores:
Juan Velasquez
Felipe Vildoso

IN5528
Introduccion al Big Data

Agenda

- Introduccion a Twitter
- Codificación
- Resultados
- Problemas y Conclusiones

Introducción: Twitter

What is Twitter?

Twitter is a service for friends, family, and coworkers to communicate and stay connected through the exchange of quick, frequent messages. People post Tweets, which may contain photos, videos, links and up to 140 characters of text. These messages are posted to your profile, sent to your followers, and are searchable on Twitter search. Learn more about how to use Twitter.

310M
Monthly active users

1B
Unique visits monthly

3,8k
Employees

WHAT HAPPENS IN A TWITTER MINUTE

347,222
Tweets

$2,407
Made in revenue

4,398
Unique site visits

7
Tweets containing racial slurs (in English)

700
Video links shared

1,458,333
Twitter search engine queries

618,725
Tweets when Germany won the World Cup

With nearly 1 billion users, Twitter is one of the top social media networks used by people all over the world. Each time you tweet, there are thousands of others around the world doing the same thing.

1,211,420
Timeline views

$1,744
Made in ad revenue

300
Vines tweeted

300,000
Tweets during Miley Cyrus' VMA performance

108
New accounts

WhoIsHostingThis?

SOURCES:
mediabistro.com,
streamsend.com,
techland.time.com,
hollywood reporter.com,
adexchanger.com, twitter.com,
businessinsider, theverge.com,
statisticbrain.com, usatoday.com

El problema: implementar el Batch Layer en el contexto de la Arquitectura Lambda

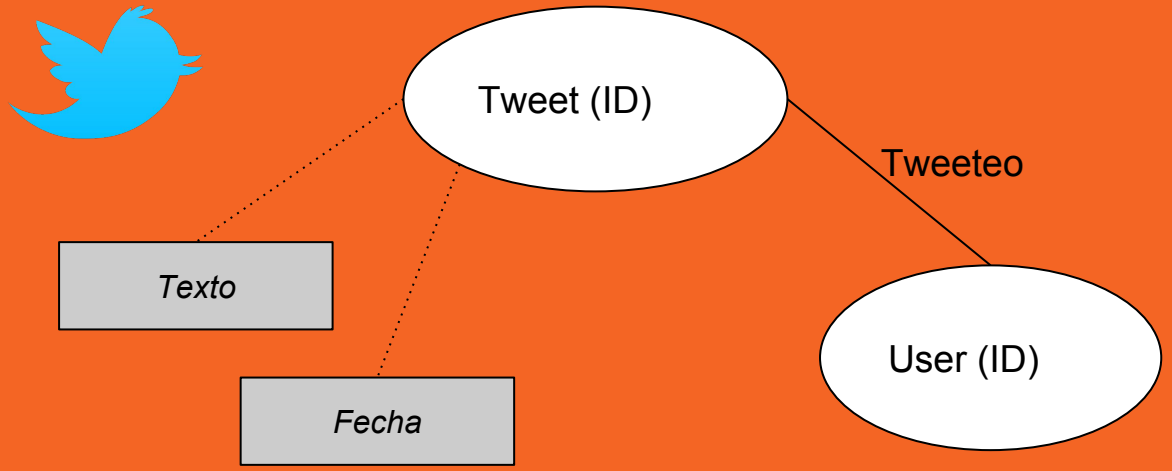
1. Top 10 de usuarios que mencionan @cooperativa
2. Días con más menciones @cooperativa
3. Palabras más nombradas junto a @cooperativa

Velocity

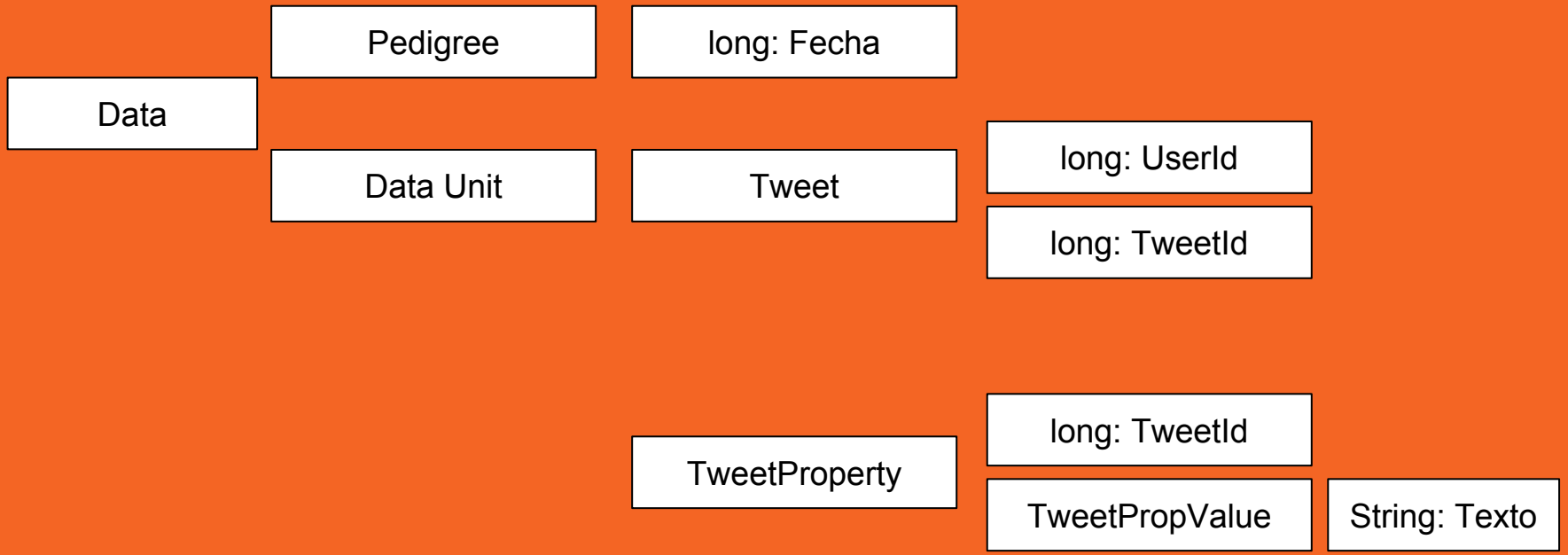
Variety

Volume

Apache Thrift Schema



2 nodos: un Tweet y un Usuario.



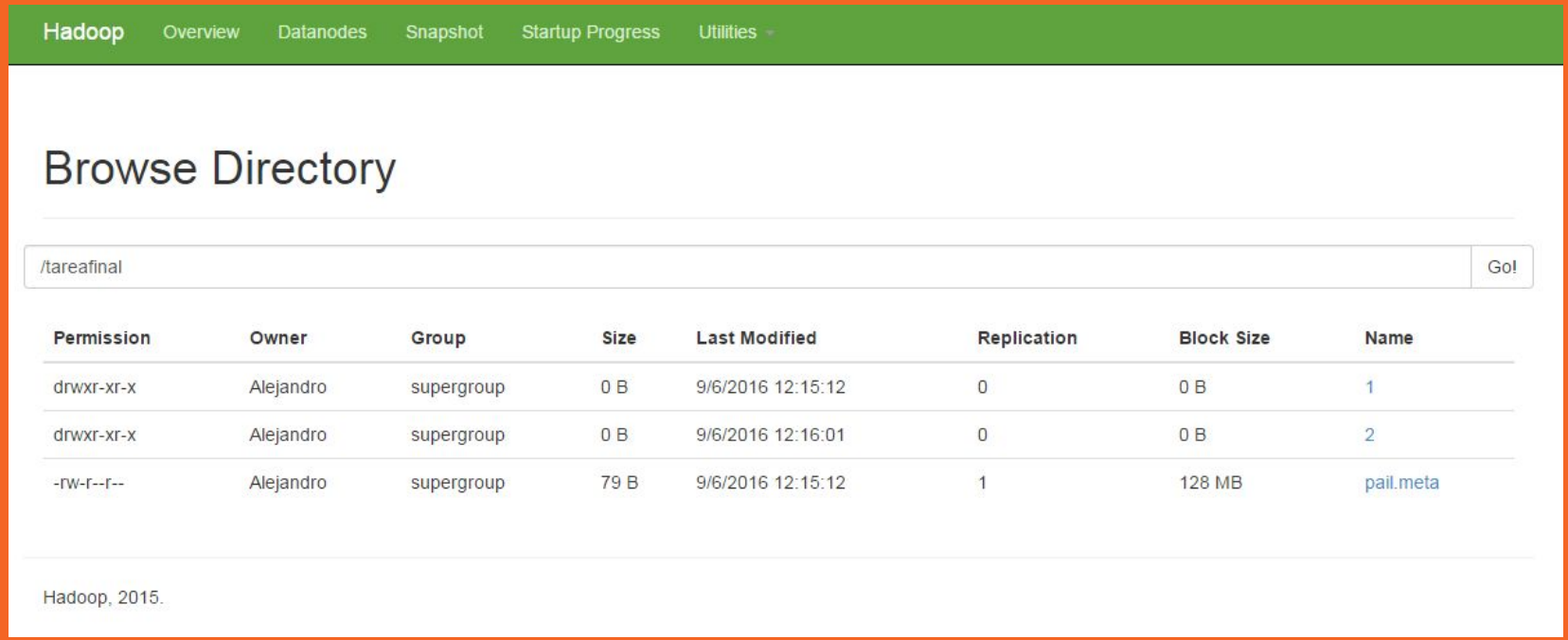
Estructura del Esquema

Pasos del Proyecto

Instalar

Estandarizar Estructura de Proyectos

Creación del Proyecto



Guardado de Datos

Pregunta 1

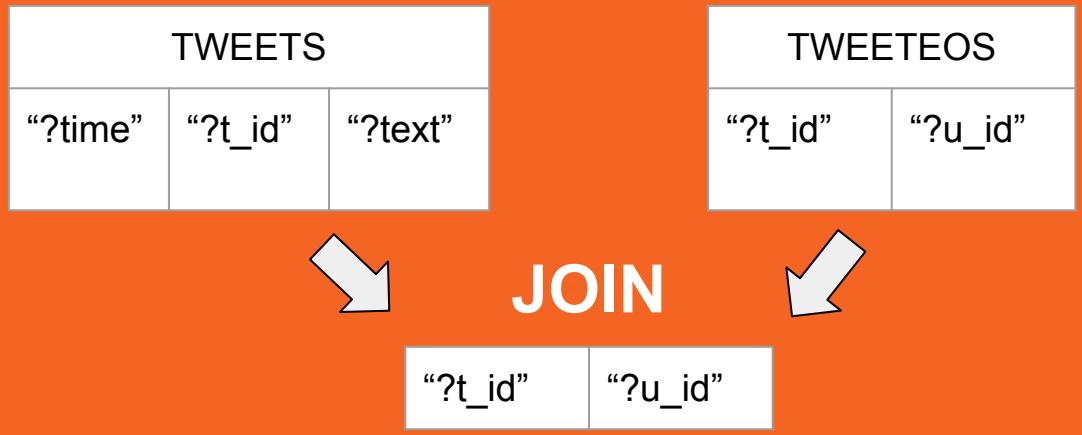
Top 10 de usuarios que mencionan @cooperativa

Pasos

1 Obtener lista con tweets que mencionan @coperativa

2 Obtener lista UserIds y TweetsIds

3



Top 10 de usuarios que mencionan @cooperativa

Query

```
public static void getJoin(){
    PailTap source = getDataTap("/tareafinal/1");
    Subquery tweets = new Subquery("?time", "?text", "?tid")
        .predicate(source, "_", "?data")
        .predicate(new ExtractText(), "?data").out("?time","?text","?tid")
        .predicate(new FText(), "?text");

    PailTap source2 = getDataTap("/tareafinal/2");
    Subquery ids = new Subquery("?time", "?uid", "?tid")
        .predicate(source2, "_", "?data")
        .predicate(new ExtractText2(), "?data").out("?time","?uid","?tid");

    Subquery algo = new Subquery("?uid", "?tid")
        .predicate(ids, "_", "?uid", "?tid")
        .predicate(tweets, "_","_","?tid");


    Subquery algo2 = new Subquery("?uid","?c")
            .predicate(algo, "?uid", "_")
            .predicate(Option.DISTINCT,true)
            .predicate(new Count(),"?c");

            Api.execute(new StdoutTap(),
                    new Subquery("?uid1", "?count1")
                    .predicate(algo2, "?uid", "?c")
                    .predicate(Option.SORT, "?c")
                    .predicate(Option.REVERSE, "?c")
                    .predicate(new Limit(10), "?uid","?c").out("?word1","?count1")
                );
```

User ID	Numero menciones	Ranking
938296908	23	1
171913028	22	2
1705850354	15	3
2161635434	15	4
2196448152	15	5
2196493872	15	6
1681742905	12	7
1861391822	12	8
1912255273	12	9
1912375015	12	10
1968897702	12	11
2161614498	12	12
2221952394	12	13
2332158937	12	14
2895803727	12	15

Resultados


AYUDA
TWEETS
102K
FOLLOWING
2,249
FOLLOWERS
7,085
Ayuda por Favor
@FavorAyuda
La unión hace la fuerza. Ayudemonos !!
¿Le darías RT a nuestros twiits?
Chile
Joined November 2014
Tweets
Tweets & replies
Media
Ayuda por Favor Retweeted
Vínculo.cl @VinculoCL · May 27
¿Necesitas un Servidor?
VPS - Servidor Privado Virtual
>> v.c1.cl/?c=fc328

Pregunta 2

Menciones de @cooperativa por día

Pasos

1 Obtener lista con tweets que mencionan @coperativa

TWEETS		
“?time”	“?t_id”	“?text”

2 Sobre esa lista, contar los elementos únicos

CONTEO	
“?time”	“?count”

Menciones de @cooperativa por día

```
// menciones por dia /P2
public static void getTuits2(){
    PailTap source = getDataTap("/tareafinal/1");
    Subquery q = new Subquery("?time", "?text", "?id")
        .predicate(source, "_", "?data")
        .predicate(new ExtractText(), "?data").out("?time","?text","?id")
        .predicate(new FText(), "?text");

    Api.execute(new StdoutTap(), new Subquery("?tiempo","?c")
        .predicate(q, "?tiempo", "_", "_")
        .predicate(Option.DISTINCT,true)
        .predicate(new Count(),"?c")
    );
}
```

```
public static class FText extends CascalogFilter {
    public boolean isKeep(FlowProcess process, FilterCall call){
        String text = (String) call.getArguments().getString(0);
        return text.contains("@cooperativa");
    }
}

public static class ExtractText extends CascalogFunction {
    public void operate(FlowProcess process, FunctionCall call){
        Data data = ((Data) call.getArguments().getObject(0));
        long time = data.getPedigree().getDate();
        String text = data.getData_unit().getTwit_property().getProperty().getTexto();
        long id = data.getData_unit().getTwit_property().getTwit_id().getTwit_id();
        call.getOutputCollector().add(new Tuple(time,text,id));
    }
}
```

Query

Fecha	Numero menciones	Fecha	Numero menciones
21-04-2016	8	03-05-2016	13
22-04-2016	13	04-05-2016	28
23-04-2016	8	05-05-2016	25
24-04-2016	7	06-05-2016	241
25-04-2016	39	07-05-2016	29
26-04-2016	27	08-05-2016	7
27-04-2016	15	09-05-2016	59
28-04-2016	35	10-05-2016	10
29-04-2016	31	11-05-2016	36
30-04-2016	55	12-05-2016	6
01-05-2016	45	13-05-2016	3
02-05-2016	11	14-05-2016	1

Cantidad de menciones de @cooperativa por día : días con más menciones: 6 de Mayo y 9 de Mayo.


Gabriela Infante
@gabinfanti
Follow
Manifestantes de Chiloé lanzaron salmones a la pileta de La Moneda. Cerca de 10 detenidos @cooperativa
View translation
RETWEETS
215
LIKES
76
11:49 AM - 6 May 2016

Principales Problemas

1. Creación de nodos
2. Identificar las dependencias de Maven
3. Familiarización con Java
4. Comprensión del guardado de datos con Pail
5. Creación de consultas con JCascalog

Conclusiones

1. Hadoop es una potente herramienta para procesar Big Data en una computador común y corriente
2. Se necesita más preparación a la hora de realizar un proyecto como este, el cual necesita conocimientos en múltiples idiomas.

Presentación Tarea 1