

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

MAESTRÍA EN CIENCIA DE DATOS

SDD - Social Data Discovery

Manual de Instalación

Data Product Architecture

Mayo 2016

HOJA DE CONTROL

<i>Clave</i>			
<i>Título</i>	Manual de Instalación SDD (Microservicios)		
<i>Autor</i>	Integrantes Social Data Discovery (SDD)		
<i>Versión</i>	01	<i>Fecha Versión</i>	26/05/2016
<i>Revisado/Validado por:</i>		<i>Fecha Revisión/Validación</i>	
<i>Aprobado por:</i>		<i>Fecha Aprobación</i>	
		<i>No. Total Páginas</i>	10

REGISTRO DE CAMBIOS

<i>Versión</i>	<i>Causa del cambio</i>	<i>Responsable del cambio</i>	<i>Fecha del cambio</i>
01	Versión inicial	Adrián Vázquez	26/05/2016

Índice

1. Propósito	3
2. Instructivo de Instalación	3
2.1. Prerrequisitos	3
2.1.1. Consideraciones para iniciar con el procedimiento de instalación	4
2.1.2. Instalación de los Componentes Desarrollados	4
3. Definición de siglas y abreviaturas	10

1. Propósito

El propósito del documento es detallar el proceso de instalación final de la solución para microservicios.

Este documento está dirigido a:

Audiencia	Propósito
Líder técnico ETL	Realizar el documento
Sistemas operativos Base de datos Seguridad	Servir como insumo para habilitar la instalación de los componentes requeridos
Oficina técnica	Revisar el documento y dar las observaciones correspondientes
Control de liberaciones	Instalar los componentes de informática, así como todos los elementos de infraestructura, software y hardware requeridos

2. Instructivo de Instalación

2.1. Prerrequisitos

Descripción	Son los requisitos de software y hardware previamente instalados para el correcto despliegue del Proyecto Social Data Discovery - Ingesta, ETL, Visualización; se hace mención a requerimientos necesarios por parte de Sistemas Operativos y Control de Liberaciones (de aplicar)
-------------	--

Los requerimientos se detallan y organizan en la siguiente tabla:

	Requerimiento	Solicitar a:
1	Se entrega vía GitHub el repositorio con el siguiente contenido: -directorio <code>ambiente/</code> -directorio <code>documentacion/</code> Es el software base para replicar la solución	Integrantes del proyecto SDD -Laura Vargas -Ricardo Lastra -Ana Paula Alonzo -Adrián Vázquez
2	Se debe tener instalado el siguiente software: - <code>docker</code> versión 1.11.1, build 5604cb - <code>docker-compose</code> versión 1.7.0, build 0d7bf73 - <code>git</code> versión 2.7.4	Dr. Adolfo de Unánue

PROCEDIMIENTO

Descripción	Clonación del repositorio <path>/social-data-discovery.git
-------------	--

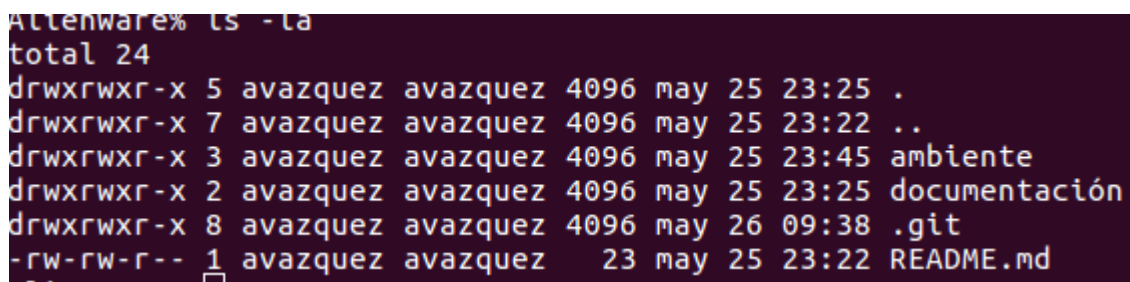
2.1.1. Consideraciones para iniciar con el procedimiento de instalación

- Se debe tener instalado el Software solicitado en el paso 2.1
- Se debe tener acceso al servidor de Github correspondiente.
- Se debió generar el request especificado por los integrantes del equipo Social Data Discovery en el repositorio <https://github.com/ITAM-DS/data-product-architecture/tree/master/proyecto_grupal>

2.1.2. Instalación de los Componentes Desarrollados

Para la instalación de los componentes desarrollados, es necesario ubicarse en la carpeta <path_master>/social-data-discovery y llevar a cabo los siguientes pasos:

1. Una vez en la carpeta social-data-discovery/ ejecutar el comando `ls -la`



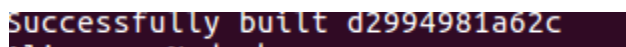
```

Allenware% ls -la
total 24
drwxrwxr-x 5 avazquez avazquez 4096 may 25 23:25 .
drwxrwxr-x 7 avazquez avazquez 4096 may 25 23:22 ..
drwxrwxr-x 3 avazquez avazquez 4096 may 25 23:45 ambiente
drwxrwxr-x 2 avazquez avazquez 4096 may 25 23:25 documentación
drwxrwxr-x 8 avazquez avazquez 4096 may 26 09:38 .git
-rw-rw-r-- 1 avazquez avazquez  23 may 25 23:22 README.md

```

Figura 1: Listado de archivos de instalación

2. Ingresar al directorio ambiente mediante el comando `cd ambiente`, posteriormente ejecutar el comando `docker-compose build`, con esto se genera el ambiente del proyecto Social Data Discovery (SDD). Se debe obtener el siguiente mensaje de salida:



```

successfully built d2994981a62c

```

Figura 2: Notificación de ejecución correcta del build

3. Ejecutar el comando `docker-compose up -d` para validar que el ambiente fue creado correctamente, se obtendrá la siguiente salida:

```

Alienware% docker-compose up -d
Creating ambiente_sdd-data_1
Creating ambiente_sdd-datalake_1
Creating ambiente_sdd-postgres_1
Creating ambiente_sdd-jupyter_1
Alienware% docker-compose ps

```

Name	Command	State	Ports
ambiente_sdd-data_1	sh	Exit 0	
ambiente_sdd-datalake_1	sh	Exit 0	
ambiente_sdd-jupyter_1	tini -- start-notebook.sh	Up	0.0.0.0:8888->8888/tcp
ambiente_sdd-postgres_1	/docker-entrypoint.sh postgres	Up	0.0.0.0:5432->5432/tcp

```

Alienware%

```

Figura 3: Vista del ambiente generado por docker-compose

4. Ejecutar el comando `docker exec -it ambiente_sdd-jupyter_1 /bin/bash`, esto lleva a una terminal del contenedor `ambiente_sdd-jupyter_1`
5. Ejecutar el comando `./sdd-ingesta.sh & exit` el cual ejecuta el Streaming con Twitter para el *hashtag* *EgyptAir*. Se genera la siguiente salida:

```

sdd-ingesta.py sdd-ingesta.sh sdd-visualizacion.py
jovyan@sdd-jupyter:~/work$ ./sdd-ingesta.sh & exit
[1] 19
exit
Mostrando los nuevos tweets de #EgyptAir:
on_data called
Received: 735860525549113344
^C

```

Figura 4: Salida de Streaming Twitter

NOTA: este programa está en Streaming y estará arrojando salidas de mensajes hasta los 1800 tweets, posteriormente se ejecutará nuevamente.

6. En un explorador ir al servidor de Notebook <<http://localhost:8888/tree>> que está preparado para ejecutar el Análisis de Sentimiento, el contenido es el siguiente:

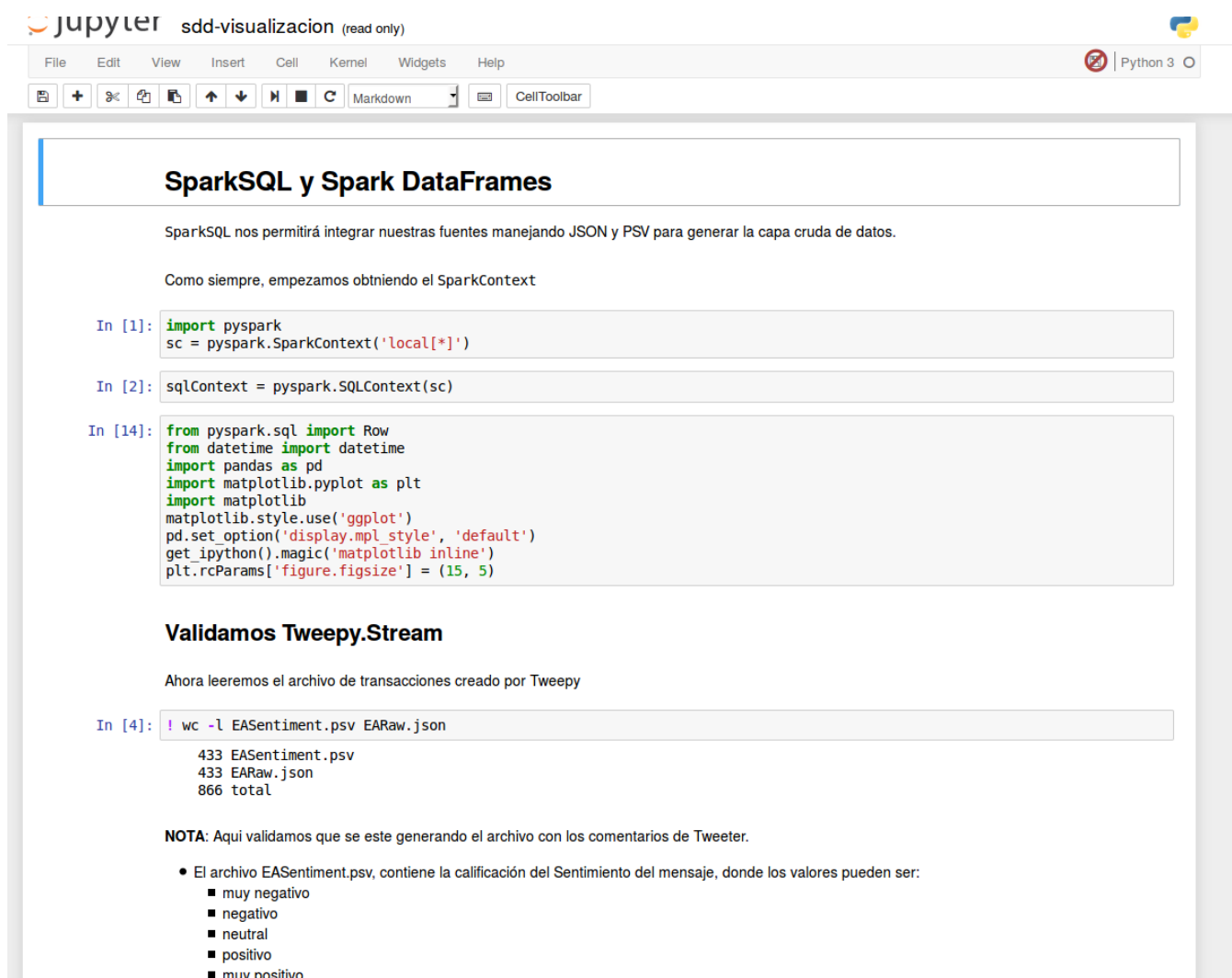


Figura 5: Vista del servidor de Notebook

7. El contenido del Home es el siguiente:

Nombre del archivo	Tipo	Descripción
<code>sdd-visualizacion.ipynb</code>	Código	Notebook Python3 para hacer análisis de sentimientos, basado en PySpark, SQLContext, RDD, Data Frame, Pandas, Matplotlib.
<code>sdd-ingesta.py</code>	Código	Código Python3 para generar el Streaming con Twitter, está basado en tweepy, json, textblob.
<code>sdd-ingesta.sh</code>	Shell	Shell básico para ejecutar el código <code>sdd-ingesta.py</code> desde el start up de los servicios. NOTA: No se pudo ejecutar ya que el contenedor sólo ejecuta un servicio a la vez, se dejó de forma manual.
<code>EARaw.json</code>	JSON	Archivo de trabajo temporal para almacenar de forma incremental los tweets en formato JSON y manipularlos en el Notebook <code>sdd-visualizacion.ipynb</code> para su procesamiento.
<code>EASentiment.psv</code>	psv	Archivo de trabajo temporal para calificar y almacenar el sentimiento por cada tweet en formato psv, y manipularlos en el Notebook <code>sdd-visualizacion.ipynb</code> para su procesamiento.

8. Una vez dentro del Notebook, abrir el archivo `sdd-visualización.ipynb`



The screenshot shows a Jupyter Notebook interface with the title "sdd-visualizacion (read only)". The notebook content is as follows:

SparkSQL y Spark DataFrames

SparkSQL nos permitirá integrar nuestras fuentes manejando JSON y PSV para generar la capa cruda de datos.

Como siempre, empezamos obteniendo el SparkContext

```
In [1]: import pyspark
sc = pyspark.SparkContext('local[*]')
```

```
In [2]: sqlContext = pyspark.SQLContext(sc)
```

```
In [14]: from pyspark.sql import Row
from datetime import datetime
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib
matplotlib.style.use('ggplot')
pd.set_option('display.mpl_style', 'default')
get_ipython().magic('matplotlib inline')
plt.rcParams['figure.figsize'] = (15, 5)
```

Validamos Tweepy.Stream

Ahora leeremos el archivo de transacciones creado por Tweepy

```
In [4]: ! wc -l EASentiment.psv EARaw.json
433 EASentiment.psv
433 EARaw.json
866 total
```

NOTA: Aquí validamos que se este generando el archivo con los comentarios de Tweeter.

- El archivo EASentiment.psv, contiene la calificación del Sentimiento del mensaje, donde los valores pueden ser:
 - muy negativo
 - negativo
 - neutral
 - positivo
 - muy positivo

Figura 6: Contenido del Notebook para Análisis de Sentimientos

9. Después de haber seleccionado el archivo, ejecutar la opción **Restart & Run All**, de acuerdo con los siguientes pasos:

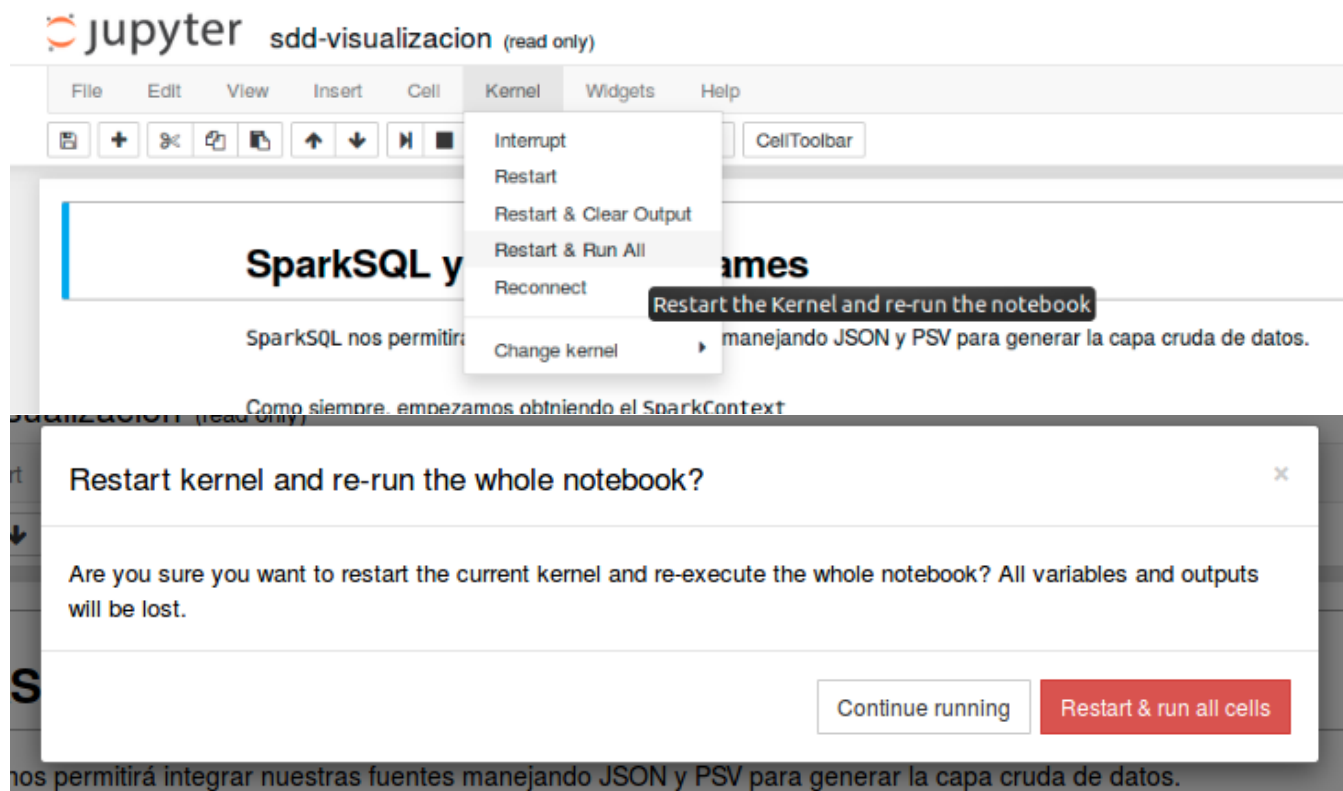


Figura 7: Pasos para ejecutar el Análisis de Sentimientos

10. Validar el código, mismo que puede ser ejecutado por partes o el resultado final que corresponde a gráficas como estas:



- Contar el numero de Tweets por Zona Horaria y poner el Top 10
- Del grafico anterior, se muestran de izquierda a derecha el mayor numero de Tweets acerca del TT EgypAir segun la zona horaria.
- Lo anterior tambien marca un dato curioso que es la **zona de la tendencia**, la cual refleja el interes sobre la nota del avion.

Figura 8: Zona Horaria donde hay tweets con #EgyptAir

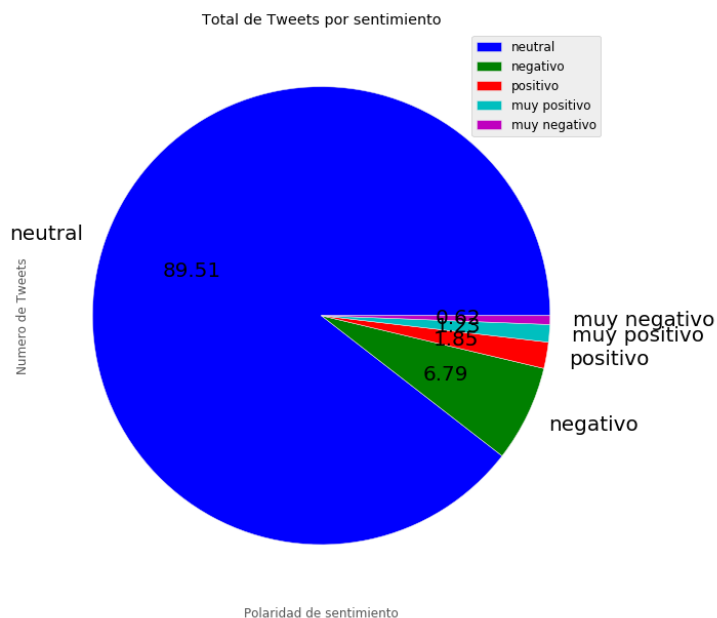


Figura 9: Polaridad del Sentimiento de tweets con #EgyptAir

11. IMPORTANTE, para refrescar las estadísticas se deberá ejecutar el proceso desde el paso 9 en adelante.

3. Definición de siglas y abreviaturas

La tabla muestra las siglas y abreviaturas utilizadas en el documento.

Siglas	Descripción
SDD	Social Data Discovery