

puiglogo.png


Análisis de datos con Apache Cassandra y Python

MEMORIA

CFGS DESARROLLO DE APLICACIONES MULTIPLATAFORMA

Enrique Villarreal
Adrián Asensio

Curso **2016-2017**
10 de febrero de 2017



creativecommons.png

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

Resumen

Nullam eu ante vel est convallis dignissim. Fusce suscipit, wisi nec facilis facilisis, est dui fermentum leo, quis tempor ligula erat quis odio. Nunc porta vulputate tellus. Nunc rutrum turpis sed pede. Sed bibendum. Aliquam posuere. Nunc aliquet, augue nec adipiscing interdum, lacus tellus malesuada massa, quis varius mi purus non odio. Pellentesque condimentum, magna ut suscipit hendrerit, ipsum augue ornare nulla, non luctus diam neque sit amet urna. Curabitur vulputate vestibulum lorem. Fusce sagittis, libero non molestie mollis, magna orci ultrices dolor, at vulputate neque nulla lacinia eros. Sed id ligula quis est convallis tempor. Curabitur lacinia pulvinar nibh. Nam a sapien.

Palabras clave: big data, data science, nosql, cassandra, python

Abstract

Nullam eu ante vel est convallis dignissim. Fusce suscipit, wisi nec facilis facilisis, est dui fermentum leo, quis tempor ligula erat quis odio. Nunc porta vulputate tellus. Nunc rutrum turpis sed pede. Sed bibendum. Aliquam posuere. Nunc aliquet, augue nec adipiscing interdum, lacus tellus malesuada massa, quis varius mi purus non odio. Pellentesque condimentum, magna ut suscipit hendrerit, ipsum augue ornare nulla, non luctus diam neque sit amet urna. Curabitur vulputate vestibulum lorem. Fusce sagittis, libero non molestie mollis, magna orci ultrices dolor, at vulputate neque nulla lacinia eros. Sed id ligula quis est convallis tempor. Curabitur lacinia pulvinar nibh. Nam a sapien.

Keywords: big data, data science, nosql, cassandra, python

Índice general

| | |
|---|----------|
| 0.1. Introducción | 1 |
| 0.1.1. Contexto y justificación | 1 |
| 0.1.2. Objetivos | 1 |
| 0.1.3. Metodología | 1 |
| 0.1.4. Planificación del proyecto | 1 |
| 0.1.4.1. Planificación horaria | 1 |
| 0.1.5. Descripción de los capítulos | 2 |
| 1. Anexo | 3 |
| 1.1. Instalando CentOS 7 | 3 |
| 1.1.1. Localización | 4 |
| 1.1.1.1. Fecha y hora | 4 |
| 1.1.1.2. Teclado | 4 |
| 1.1.1.3. Soporte de idiomas | 4 |
| 1.1.2. Discos | 4 |
| 1.1.3. Continuando la instalación | 6 |
| 1.2. Trabajando remotamente | 6 |
| 1.2.1. Consola remota mediante el proxy SOCKS | 7 |
| Glosario | 8 |
| Bibliografía | 9 |

Índice de figuras

| | |
|---|---|
| 1.1. Pantalla principal del instalador Anaconda en CentOS | 3 |
| 1.2. Selección de discos | 4 |
| 1.3. Añadiendo punto de montaje | 5 |
| 1.4. Editando punto de montaje | 5 |
| 1.5. Configuración del proxy en Firefox | 6 |

0.1. Introducción

0.1.1. Contexto y justificación

Las bases de datos relacionales llevan con nosotros desde hace más de 40 años. Como bien dice su nombre, están basadas en el modelo relacional, descrito por Edgar F. Codd en 1970.

Dos de las características deseables de una base de datos relacional bien diseñada, son la *normalización* y las *transacciones*. A través de ellas, la consistencia de los datos mejora, es decir, nos aseguramos de que la base de datos se mantiene en un estado coherente.

0.1.2. Objetivos

Los objetivos principales del proyecto son:

- Estudiar y conocer el SGBD NoSQL Apache Cassandra.
 - Identificar los usos ideales para Apache Cassandra.
- Entender el modelo de datos y consultas de Apache Cassandra.
- Diseñar un sistema para almacenar estadísticas sobre el tráfico de red en Apache Cassandra.
- Entender y familiarizarse con el flujo de trabajo del *data scientist*.
 - Analizar mediante Python los datos obtenidos de Cassandra, y intentar predecir el tráfico de red en el futuro.

0.1.3. Metodología

El proyecto se dividirá en dos partes bien diferenciadas: la primera consistirá de un estudio de las bases de datos NoSQL, en concreto Apache Cassandra, y una comparativa en base al modelo relacional.

La segunda, consistirá en emplear lo aprendido en la primera parte para implementar un sistema que aproveche las aptitudes de Cassandra para el *big data* y el análisis de datos. A éstos efectos, se utilizará Python y un variado conjunto de librerías para explorar los datos recogidos, visualizarlos, y intentar predecir el tráfico de red estimado en un punto concreto del futuro.

0.1.4. Características técnicas

En la siguiente lista se detallan las tecnologías y herramientas que se utilizarán durante el proyecto:

- **Documento de la memoria:** \LaTeX
- **SGBD NoSQL:** Apache Cassandra
- **Lenguaje de programación:** Python **Librerías:**
 - Pandas
 - statsmodels
 - matplotlib
 - Jupyter Notebook
- **Presentación:** Reveal.js

0.1.5. Planificación del proyecto

0.1.5.1. Planificación horaria

0.1.6. Descripción de los capítulos

1. Estado del arte:
2. Conclusiones:

Glosario

.

.

.

.

.

.

Bibliografía

- [1] VMware. *Understanding Full Virtualization, Paravirtualization and Hardware Assist*. http://www.vmware.com/files/pdf/VMware_paravirtualization.pdf. [Online; visitado el 29 de mayo de 2016]. 2007.
- [2] Red Hat. *oVirt Quick Start Guide*. http://www.ovirt.org/Quick_Start_Guide. [Online; visitado el 15 de enero de 2016]. 2015.
- [3] Gionatan Danti. *ZFS, BTRFS, XFS, EXT4 and LVM with KVM - a storage performance comparison*. <http://www.ilsistemista.net/index.php/virtualization/47-zfs-btrfs-xfs-ext4-and-lvm-with-kvm-a-storage-performance-comparison.html?limitstart=0>. [Online; visitado el 20 de mayo de 2016]. 2015.