

Índice

Prólogo	15
1. ¿Cómo un hacker acabó de Chief Data Officer en la era del Big Data?	15
Al principio todo fue programar	16
Y en la Universidad las Bases de Datos	17
SQL Injection, LDAP Injection y Connection String Parameter Pollution	19
Y de ahí a Chief Data Officer	21
 Capítulo I	
Antes de Empezar con Big Data	23
1. Qué encontrarás en este libro	23
2. Preguntas que se responderán	24
3. Titulares y datos interesantes	25
 Capítulo II	
¿Qué son los datos?	27
1. Los datos, la información y el conocimiento	27
2. Los datos desde la perspectiva Big Data	30
Ejemplos de tipo de dato estructurado	31
Ejemplos de tipo de dato estructurado	34
Ejemplos de tipo de dato no estructurado	36
 Capítulo III	
Bases de todo proyecto Big Data	39
1. ¿Qué son las ETL?	39
2. Las Bases de Datos: Aliadas en Big Data	44
SQL y las Bases de Datos	44
Ejemplo de una Base de Datos: El banco y los cajeros	46
Operaciones CRUD	47
Las Bases de Datos noSQL	48

Bases de Datos Clave-Valor.....	48
Bases de Datos documentales.....	49
Bases de Datos orientados a grafos	50

Capítulo IV

Business Intelligence, Big Data e Inteligencia Artificial53

1. Definiciones de BI, Big Data e IA.....	53
2. Business Intelligence	57
Beneficios del Business Intelligence y el uso de KPIs.....	58
3. Ejemplos de Paneles o Cuadros de Mando	59
4. Big Data y los datos masivos.....	63
Contar palabras: Lo mejor para entender lo que es Big Data	63
Los Paradigmas Batch, Tiempo Real y Lambda	65
Paradigma Batch Processing.....	65
Paradigma Tiempo Real	66
Paradigma Lambda.....	68
5. Inteligencia Artificial.....	68
Qué es una IA: Lo que todos deberían saber.....	69

Capítulo V

Apache Hadoop y Spark: Herramientas estrella para Big Data71

1. Apache Hadoop y bases técnicas	71
Qué es un clúster	71
Por qué se construye el proyecto Apache Hadoop	72
La escalabilidad horizontal y vertical	73
Los componentes de Hadoop: MapReduce y HDFS.....	75
Map Reduce	75
Ejemplo guiado: conteo de palabras	76
Ventajas de utilizar el framework Hadoop	78
Ecosistema Hadoop: Algunas de las tecnologías	79
Análisis de las limitaciones del funcionamiento de Hadoop	80
2. Apache Spark.....	81
Análisis de ventajas principales de Spark	81
Cómo funciona SPARK: Componentes principales.....	82
Los RDDs y Spark	82
Características de los RDD	83
Transformaciones y Acciones sobre los RDDs en Spark	84

Ejemplos de Transformaciones y Acciones en Spark	84
Spark context, los workers y su importancia	85
Módulos en Apache Spark	87
Hadoop vs Spark: Analizando los dos framework estrella en Big Data	89

Capítulo VI

Trabajo y roles en Big Data.....93

1. Introducción.....	93
2. Salidas Laborales: Un pequeño estudio realizado.....	93
Big Data Architect o Arquitecto de Datos	96
Ingeniero Big Data	96
Data Analyst o Analista de Datos.....	97
Data Scientist o Científico de datos	97

Capítulo VII

3 proyectos Big Data101

1. Proyectos reales Big Data	101
2. Proyecto 1: Sistema de mantenimiento predictivo de un parque eólico.....	101
3. Proyecto 2: Fuga de Clientes aplicado en supermercados.....	103
4. Proyecto 3: Los sistemas Antivirus	104

Capítulo VIII

Empezar con Big Data I: Python.....107

1. Programación y Python	107
Bases rápidas para entender la programación	108
Nivel 1: Tipos de datos simples y operaciones básicas en Python.....	108
Guía para empezar a programar en Python rápido.....	110
Algoritmo de programar.....	110
Ejemplo de código aplicando el Algoritmo de programar	110
Consejos para empezar a programar	111
2. Variables y Tipos de Datos Simples en Python	111
Ejercicios para practicar variables y tipos de datos	112
Enunciado ejercicio 1	112
Enunciado ejercicio 2	113
Enunciado ejercicio 3	113
3. Estructuras Condicionales.....	114

Funcionamiento del código	115
Ejercicios para practicar estructuras condicionales en Python	116
Enunciado ejercicio 1	116
Enunciado ejercicio 2	116
4. Estructuras Repetitivas.....	117
For	118
Estilo 1 Range	118
Estilo 2 colecciones.....	118
Ejercicios para practicar el for en Python	118
Enunciado ejercicio 1	118
Enunciado ejercicio 2	118
While	119
Ejercicios para practicar el for en Python	119
Enunciado ejercicio 1	119
Enunciado ejercicio 2	120
Enunciado ejercicio 3	120
5. Estructuras de datos Complejas y Colecciones de Datos.....	121
Listas	122
Funcionalidades de las Listas	123
Ejercicios para practicar con las Listas en Python	124
Enunciado ejercicio 1	124
Enunciado ejercicio 2	124
Textos o String	125
Funcionalidades de los Textos.....	126
Ejercicios para practicar con los Textos en Python.....	127
Enunciado ejercicio 1	127
Enunciado ejercicio 2	127
Diccionarios.....	128
Funcionalidades de los Diccionarios.....	128
Ejercicios para practicar con los Diccionarios en Python.....	129
Enunciado ejercicio 1	129
Enunciado ejercicio 2	129
6. Librerías	130
3 librerías de interés para empezar en Big Data.....	130
Librería nativa CSV	130
Lectura de líneas y transformación de CSV en Python.....	131
7. Jupyter.....	131
Qué es Jupyter y por qué es tan interesante	131
Instalación de Jupyter	132
Cómo usar los cuadernos de Jupyter.....	133

Conector MySQL a Base de Datos	135
pySpark	135

8. Adicional: Instalación de Python	135
--------------------------------------------------	------------

Capítulo IX

Empezar con Big Data II: SQL.....	139
------------------------------------------	------------

1. Bases de Datos y SQL.....	139
-------------------------------------	------------

Guía para empezar SQL rápido	139
------------------------------------	-----

La regla de oro para empezar SQL.....	140
---------------------------------------	-----

2. Bases rápidas en SQL y tipos de operaciones	140
-------------------------------------------------------------	------------

3. Filtrado de datos y la cláusula WHERE	141
-------------------------------------------------------	------------

Ejercicios para practicar el Where	142
------------------------------------------	-----

Enunciado ejercicio 1	142
-----------------------------	-----

Enunciado ejercicio 2	142
-----------------------------	-----

4. Los filtros IN, BETWEEN y LIKE	143
------------------------------------------------	------------

Ejercicios para practicar el IN, BETWEEN y LIKE.....	143
------------------------------------------------------	-----

Enunciado ejercicio 1	143
-----------------------------	-----

Enunciado ejercicio 2	144
-----------------------------	-----

Enunciado ejercicio 3	144
-----------------------------	-----

Enunciado ejercicio 4	144
-----------------------------	-----

Enunciado ejercicio 5	144
-----------------------------	-----

Enunciado ejercicio 6	144
-----------------------------	-----

5. Operadores lógicos: AND OR NOT	145
------------------------------------------------	------------

6. Operaciones de agregación y agrupamientos	145
-----------------------------------------------------------	------------

Operaciones de agregación	145
---------------------------------	-----

Comando Group By	146
------------------------	-----

Ejercicios para practicar Agregaciones	148
----------------------------------------------	-----

Enunciado ejercicio 1	148
-----------------------------	-----

Enunciado ejercicio 2	148
-----------------------------	-----

Enunciado ejercicio 3	148
-----------------------------	-----

Enunciado ejercicio 4	148
-----------------------------	-----

7. Ordenamiento de resultados y Límites.....	149
-----------------------------------------------------	------------

8. Sigüentes pasos: SQL avanzado y noSQL.....	150
------------------------------------------------------	------------

SQL avanzado	150
--------------------	-----

9. Instalación de MySQL	151
--------------------------------------	------------

Configuración del entorno y servidor.....	153
-------------------------------------------	-----

Breve tour de MySQL Workbench.....	154
------------------------------------	-----

Creación de Schema en una Base de Datos mySQL.....	156
----------------------------------------------------	-----

Capítulo X

Empezar con Big Data III: pySpark159

1. Introducción.....	159
Tabla de versiones de librerías para la instalación	159
Instalando Spark: Parte 1 - Instalando Java JDK 1.8.....	160
Instalando Spark: Parte 2 - Descarga de Spark	162
Instalando Spark: Parte 3 - Descarga de Winutils y pySpark	164
2. Inicio y arranque de Spark.....	165
3. Spark UI y creación de RDDs	166
Spark UI	166
Crear RDDs.....	167
4. Transformaciones y Acciones	168
Transformaciones	168
Acciones.....	169
5. Spark SQL y los DataFrames.....	171
Leer un CSV como DataFrame	171
Ejecutar una consulta estilo SQL sobre un CSV en Spark.....	171

Capítulo XI

Ficheros y Python.....173

1. Qué es un fichero	173
2. Punteros o cursores	174
3. Modos de apertura de un fichero	174
4. Sublime: Como manipular datos masivamente.....	175
5. Instalación de Sublime en Windows	176
6. Como abrir ficheros con Sublime por defecto.....	177
7. Conversión de ficheros de texto al vuelo	178
Datos de internet a CSV	178

Capítulo XII

Bases de Datos noSQL181

1. Qué son las Bases de Datos noSQL.....	181
2. MongoDB	182

Qué es MongoDB.....	182
Terminología en MongoDB	182
Utilizando Compass (GUI)	182
Ejemplo de operaciones CRUD en MongoDB con Python	183
Ejemplo de código de lectura de CSV y añadirlo a MongoDB	184
3. Apache Cassandra	185
Operaciones CRUD en Apache Cassandra	185
4. Adicional: Instalando MongoDB y Compass	186
Descargar MongoDB:	186
Ejecutar el instalador:	187
Configurar MongoDB:.....	187
Tour rápido en Compass.....	188
Como importar datos en MongoDB.....	189
Evaluación de servicios en Windows	191
5. Adicional: Instalando Apache Cassandra	193
Descargar Apache Cassandra:.....	193
Descomprimir el archivo ZIP:	193
Configurar las variables de entorno:	194
Arranque del clúster Apache Cassandra.....	195
Ejecutando breves códigos a través de Python	195
Los espacios y tablas en Apache Cassandra.....	196
Añadiendo información en Cassandra	197
Consultando una tabla en Cassandra.....	197

Capítulo XIII

Comunicación Masiva mediante Apache Kafka	199
1. Qué es Apache Kafka y Sus Ventajas	199
2. Funcionamiento principal de Apache Kafka	200
Ejemplo aterrizado a un sistema de Mensajería como WhatsApp	200
3. Como arrancar el clúster de Apache Kafka.....	201
4. Ejemplo de construcción de Topic, Producer y Consumer en Kafka	202
Construcción del Topic.....	202
5. Construcción del Consumidor.....	203
6. Generación del Productor.....	203
7. Adicional: Instalación de Apache Kafka en Windows	204

Índice alfabético	207
Índice de imágenes	211
Otros libros publicados.....	215