

Capítulo 1: Introducción Big Data

Capítulo 2: El Big Data y la ciencia de datos

Capítulo 3: Exploración de Big Data



División de Alta Tecnología

2

El Big Data y la ciencia de datos

Los datos son la nueva ciencia.
El Big Data son las respuestas. (Pat Gelsinger)

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.




Objetivos

Al finalizar el capítulo, el alumno logrará:

- Comprender el proceso del Big Data y el rol del científico de datos.

2 - 2

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Agenda

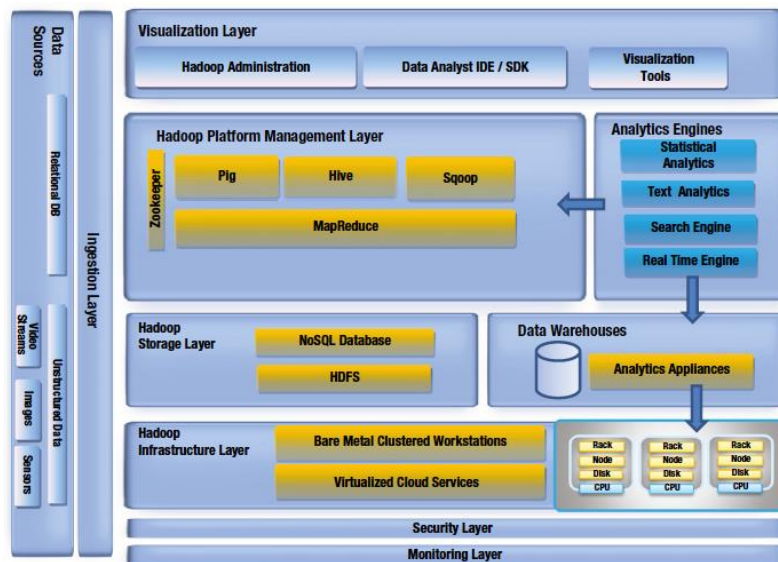
- La plataforma del Big Data
- El científico de datos
- El proceso de la ciencia de datos

2 - 3

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



La plataforma del Big Data

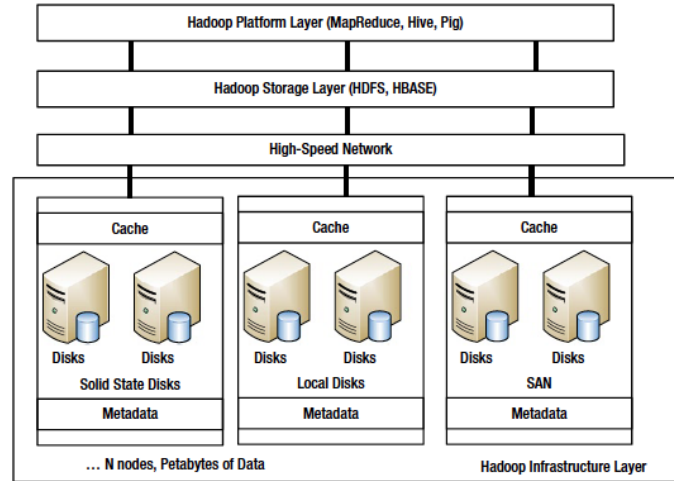


2 - 4

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



La plataforma Hadoop



2 - 5

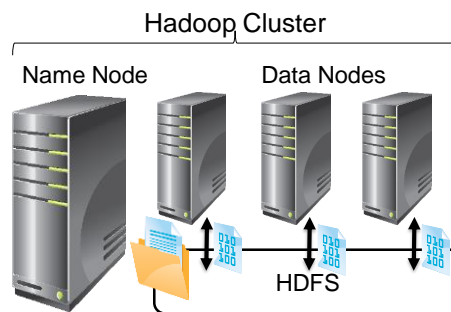
Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



La plataforma Hadoop

Hadoop Cluster

- Múltiples servidores con Sistema de archivos compartidos (HDFS).
- Name Node que atiende las peticiones de los clients.
- Múltiples nodos de datos que utilizan Map Reduce.



2 - 6

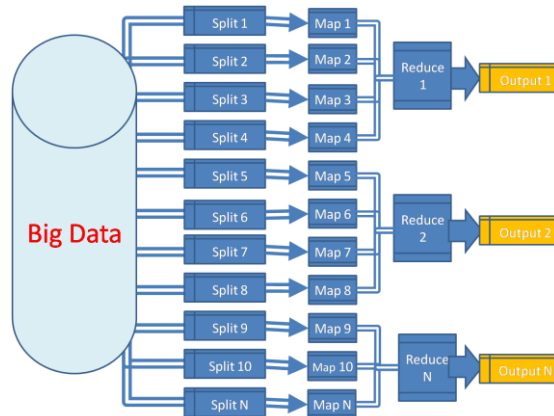
Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



La plataforma Hadoop

Map Reduce

- Map() – dividir el problemas en problemas más pequeños.
- Reduce() – combinar los resultados

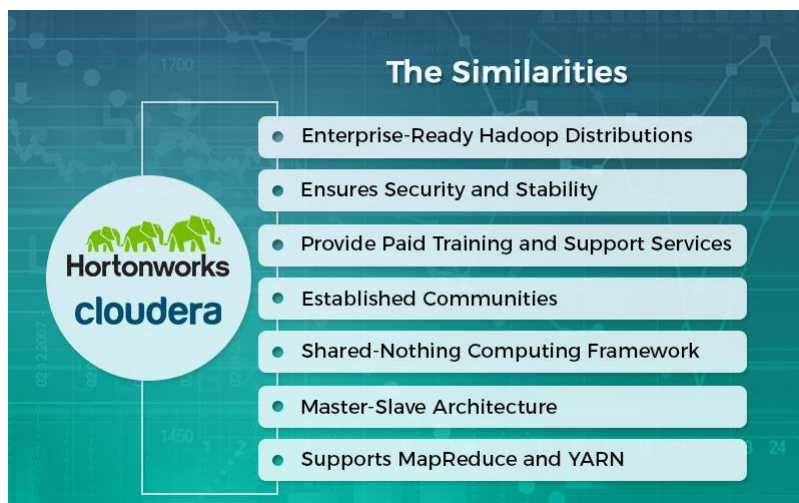


2 - 7

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Distribuidores de Hadoop



2 - 8

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Distribuidores de Hadoop

The Differences	
Hortonworks	Cloudera
Hortonworks offers Apache Foundation certified software	Cloudera sells Commercial software
Hortonworks Embeds Hadoop into existing data platforms	Cloudera Embeds with other commercial software providers
HDP a Native component	Cloudera CDH is Not a native component
No Proprietary Software	Proprietary Software
Open Source License	Commercial License
Free	Paid

2 - 9

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El científico de datos

El científico de datos es una nueva profesión que hoy es considerada clave en el mundo de las tecnologías y es una de las mejores pagadas.



2 - 10

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El científico de datos



49%
\$106 - \$120K
MEDIAN SALARY

BI & VISUALISATION FOCUSED

- More likely to use Tableau
- Most popular technical skill is BI, less likely to use predictive analytics



23%
\$130 - \$130K
MEDIAN SALARY

TRADITIONAL ANALYSTS

- More likely to use SAS Enterprise Miner, SAS Enterprise Guide and Visual Analytics
- Most common technical skills include inferential statistics and predictive analytics



22%
\$140 - \$160K
MEDIAN SALARY

DATA SCIENCE PROFESSIONALS

- A full range of technical skills & broadest tool usage of all segments
- More likely to use big data and cloud technologies



6%
\$125 - \$139K
MEDIAN SALARY

ANALYTICAL INTEGRATORS

- Very limited usage of analytical tools such as SAS or R
- Technical skills include operational analytics, business intelligence, data governance, and systems integration
- Use SQL more than the average respondent

2 - 13

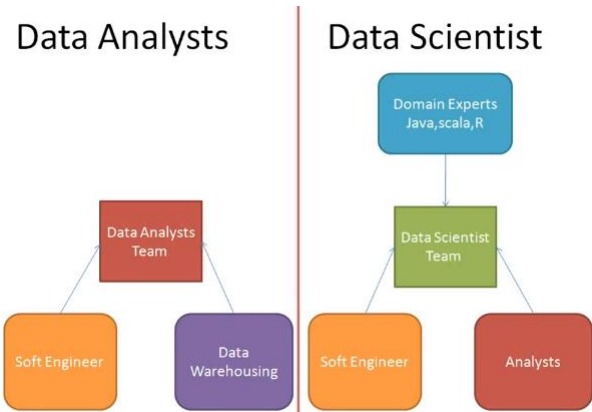
Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El científico de datos

Data Analysts

Data Scientist



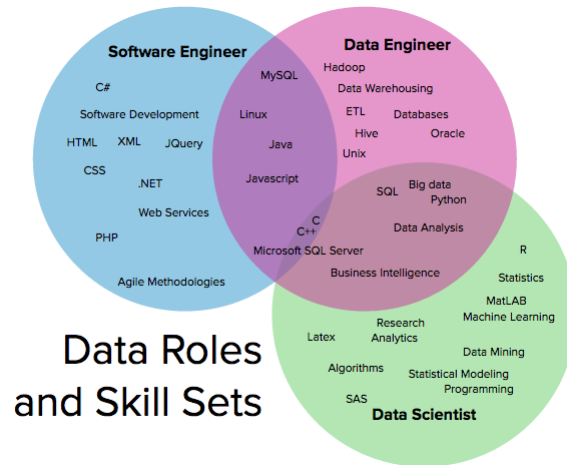
<http://bigdatasimplified.blogspot.in/>

2 - 14

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El científico de datos



Data Roles and Skill Sets

2 - 15

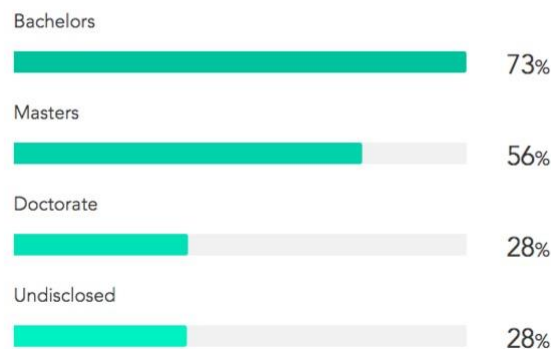
Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El científico de datos

What are the degrees prerequisites for Data Scientists jobs?

There are roughly the same number of applicants with a **Bachelors** as there are with a **Masters**. 72% of applicants have a Bachelors. 56% have a Masters. 28% have a PhD. 28% have no degree.



2 - 16

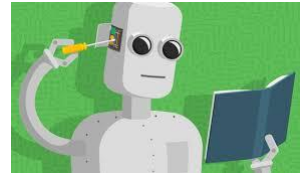
Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El científico de datos

Conocimientos de Machine Learning

- Supervised Learning
- Linear Regression, SVM, Random Forest, Logistic Regression , KNN, etc...
- Unsupervised Learning
- K-Means Clustering, PCA, etc...
- NLP, Model Validation, K-Folds, etc...
- Bias-Variance Trade-Off
- Gradient Descent
- L1 / L2 Regularization
- Bagging / Boosting



2 - 19

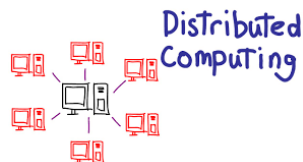
Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El científico de datos

Conocimientos de Ingeniería de Software

- Algorithms and Data Structures
- Databases (SQL)
- Distributed Computed (Spark)
- Data Visualization Products or Services

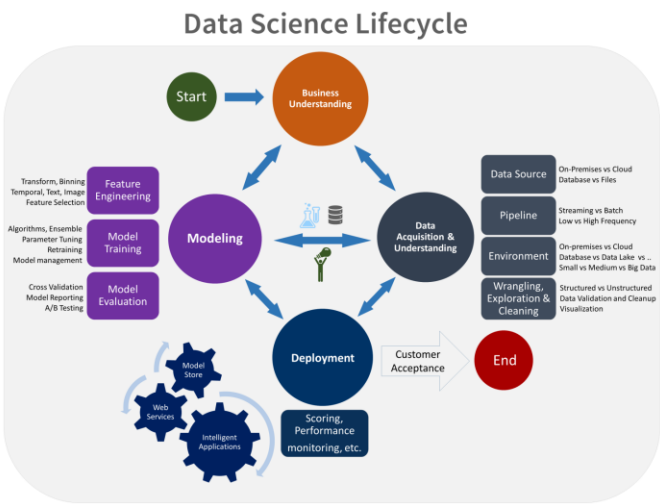


2 - 20

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El proceso de ciencia de datos

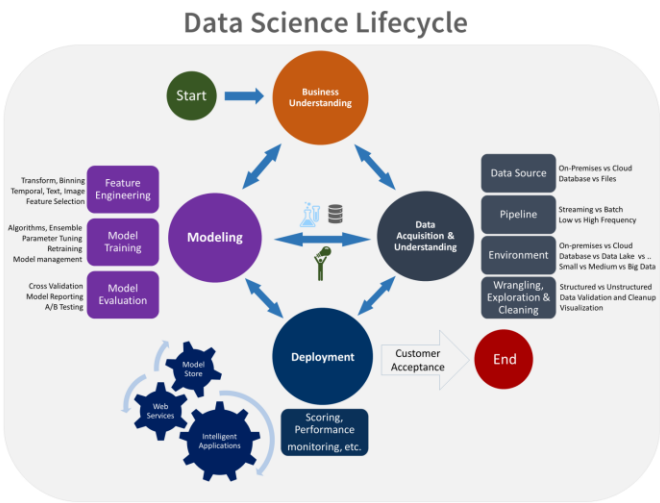


2 - 21

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El proceso de ciencia de datos

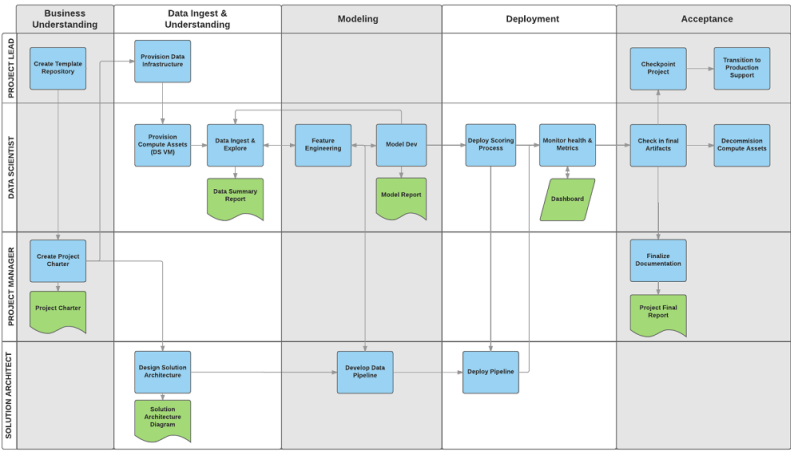


2 - 22

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El proceso de ciencia de datos

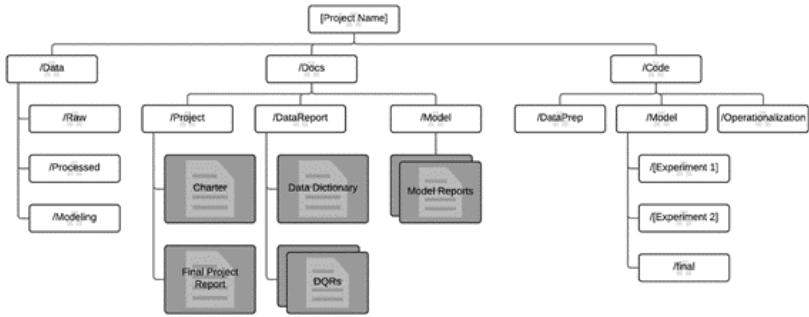


2 - 23

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El proceso de ciencia de datos

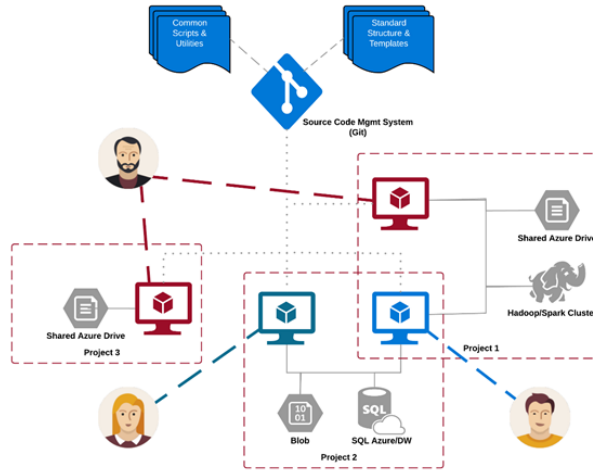


2 - 24

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El proceso de ciencia de datos



2 - 25

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Ejercicio Nº 2.1: Exploración de una solución Big Data

Al finalizar, el alumno logrará:

- Conocer como trabaja una solución de Big Data.

2 - 26

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Tarea Nº 2: Data Science Methodology

Al finalizar la tarea, el alumno logrará:

- Aprender acerca de la metodología que se puede utilizar dentro de la ciencia de datos, para garantizar que los datos utilizados en la resolución de problemas sean relevantes y se manipulen adecuadamente para abordar la cuestión en cuestión.

2 - 27

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Lecturas adicionales

Se sugiere revisar los siguientes enlaces para profundizar en los conceptos tratados en el presente capítulo:

- a) Big Data y sus componentes
- b) Los skills del científico de datos
- c) Importancia de la ciencia de datos

2 - 28

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Resumen

En este capítulo, hemos aprendido de los principales componentes de la plataforma Big Data, las características que tiene el rol del Científico de Datos.

Además, hemos hecho una breve revisión del proceso de la ciencia de datos.

