El papel del Big Data y el Cloud Computing en la Genética

José Ángel Díaz García

Cloud Computing Servicios y Aplicaciones

Contenidos

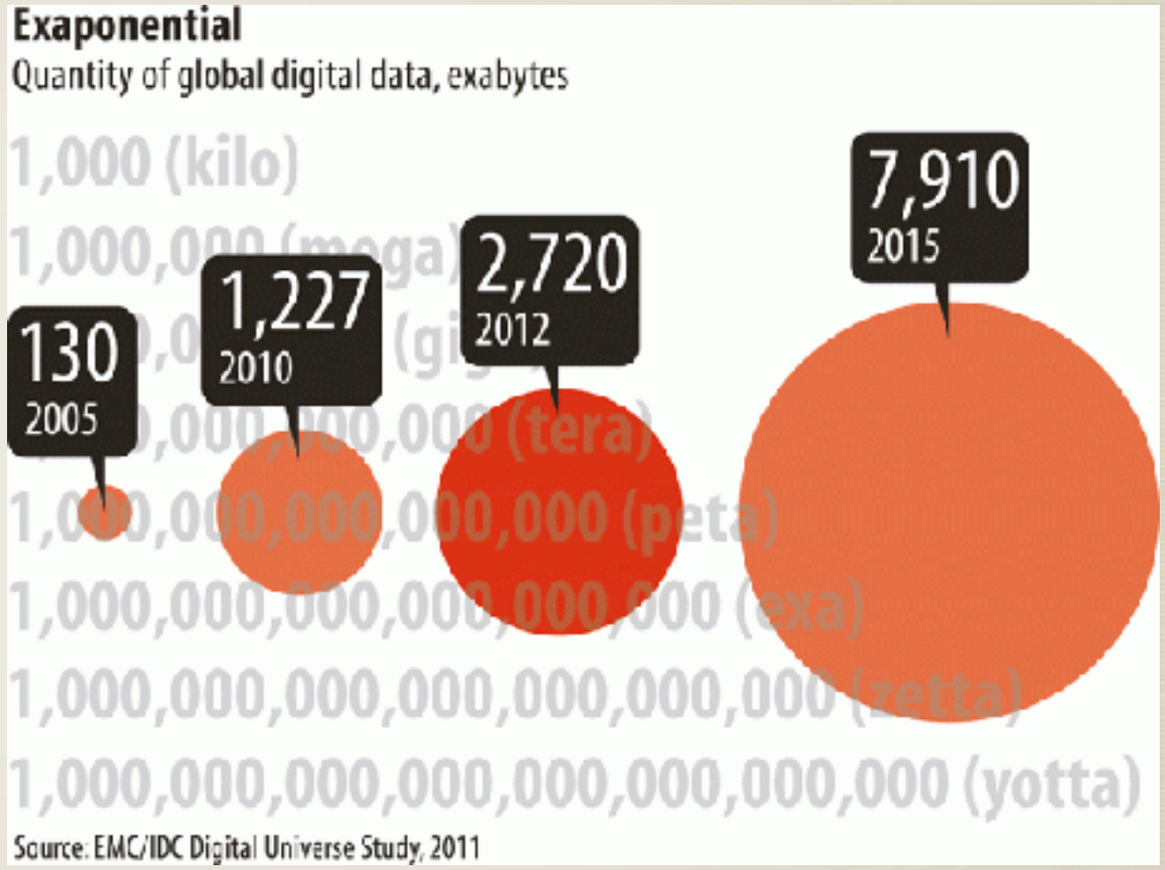
1. Motivación.
2. BigData y Cloud Computing, grandes compañeros.
3. Campos en los que el BigData destaca actualmente.
4. El papel del BigData y el Cloud Computing en la biología.
5. SaaS para bioinformática basados en BigData.
6. Genoma humano y Hadoop.
7. Conclusiones.
8. Referencias.

1- Motivación

En 2011 un estudio realizado por Eric Schmidt cifró en 295 exabytes de información los datos generados por la humanidad entre 1986 y 2007.

En los últimos dos años, se han generado el 90% de los datos de la humanidad y en 2011, hace 6 años, se rozaron los 2 zettabytes.

En la actualidad se generan 2 exabytes de datos al día.



Esta gran cantidad de datos presenta retos como, **el almacenamiento, el procesado, la seguridad y la capacidad para acoger nuevos datos (5 Vs del BigData)**. Es necesario trabajar con **nuevos paradigmas de computación** que nos permitan operar con estos grandes volúmenes de datos.

2- Big Data y Cloud Computing, grandes compañeros.

Esta gran cantidad de datos ha hecho que surjan nuevos paradigmas que permitan a empresas y organizaciones obtener valor de esta vorágine de datos siendo la computación en la nube la única opción viable para tal menester en la mayoría de los casos ya que:

- Ofrece posibilidad de almacenamiento ilimitada.
- Mayor seguridad en los datos.
- Reducción de costes.
- Gran capacidad de computo.
- Sistemas muy escalables.

Podríamos decir por tanto, que las **limitaciones del Big Data**, son totalmente **cubiertas por el Cloud Computing**.

3-Campos en los que el BigData destaca.

El BigData está presente hoy en día en casi la totalidad de los sectores económicos, destacan ámbitos como la **economía y el marketing**, pero nuevas áreas están cada vez más interesadas en estos paradigmas como la agricultura o la **biología** donde junto con la **física,** es una de las ciencias donde estás tecnologías (BigData y CloudComputing) más destacan actualmente.


EL PAÍS

TECNOLOGÍA

OPINIÓN

El Big Data al servicio de la ciencia en el CERN

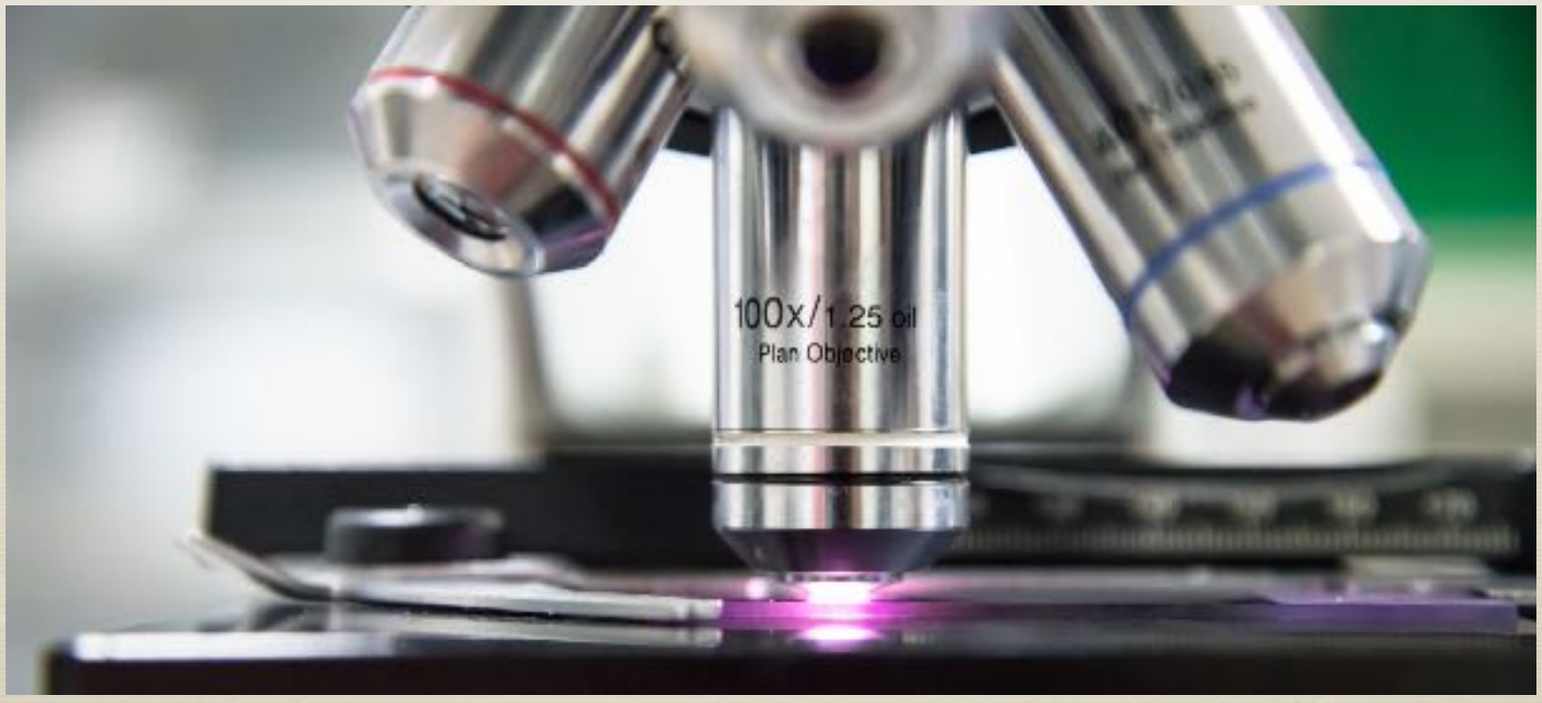
Helix Nebula quiere desarrollar una nube científica que forme un mercado abierto para la ciencia

Cada operación es comparable a 40 millones de fotos de segundo con una cámara de 100mpx. Solo una colisión de cada 10.000.000.000.000 es interesante pero se almacenan casi todas. Cada año se añaden más de 25 petabytes y se procesan en 150 CPDs repartidos por el mundo.

4-El papel del BigData y el CloudComputing en la Biología

Los experimentos basados en **BigData** en biología ha sufrido un crecimiento exponencial, versando áreas tan dispares dentro de la materia como **comparación de secuencias genéticas, interacciones entre proteínas, evolución, predicción del florecimiento de algas marinas**...

Tantas fuentes de datos implican, datos muy heterogéneos, desestructurados y de un volumen inmenso (1 secuencia del genoma humano son 140GB) de los que será necesaria su comparación con experimentos y datos previos, aquí es donde entra en juego la unión del **BigData y Cloud Computing**.



4-El papel del BigData y el CloudComputing en la Biología

Los datos y el software de procesamiento están por tanto situados en grandes CPDs en ubicaciones distintas de los laboratorios que trabajan con ellos de modo que no necesitan su propio hardware, y los laboratorios que si disponen de él complementan con esta "**gran base de datos**" que la nube representa en biología (solamente el **EBI** (Instituto Europeo de Bioinformática) almacena 20 petabytes de datos sobre el genoma).

Esto ofrece un gran avance a grandes corporaciones del ámbito de la bioinformática como a pequeños grupos de investigadores que tienen al alcance de la nube todo el potencial de los datos.

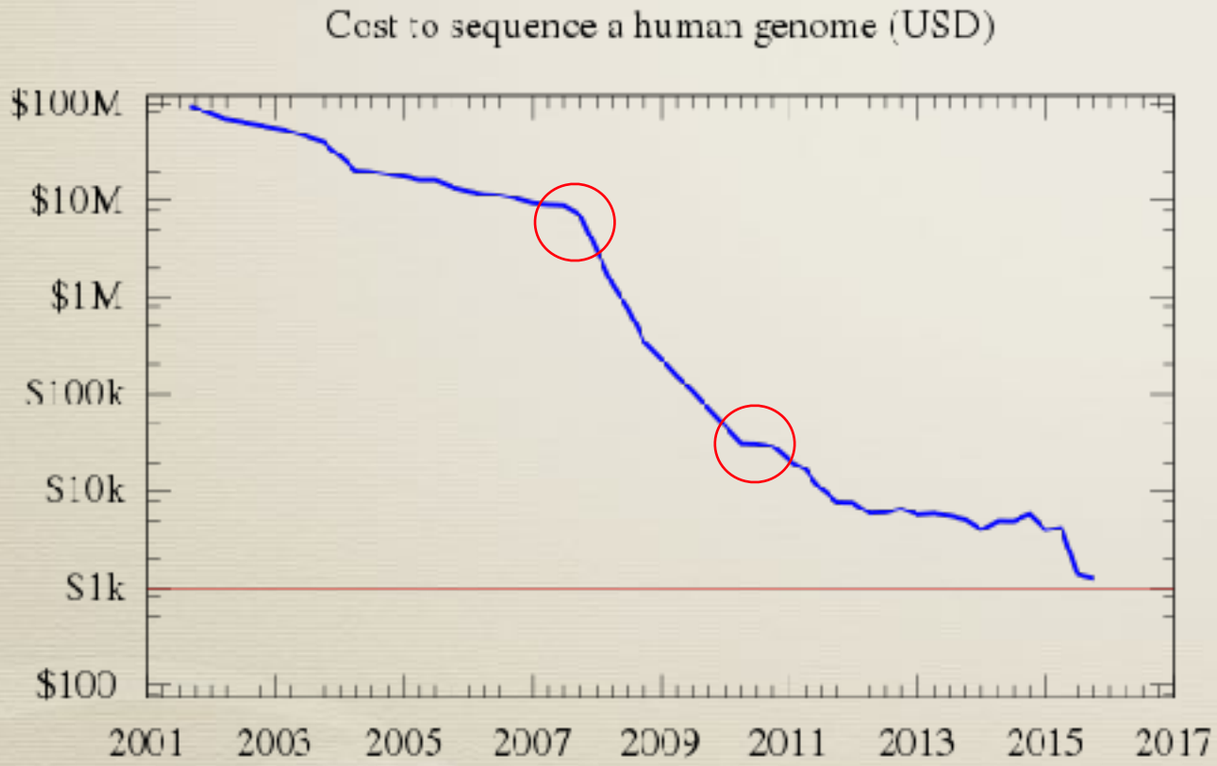
5-SaaS para bioinformática basados en BigData.

-Incluso en el EBI, se accede a servicios y bases de datos en la nube, un recurso muy utilizado es el **Ensembl Genome Browser,** donde investigadores de todo el mundo acceden para procesar y bajar el genoma de diversas especies. El centro principal, estaba en Reino Unido, pero cuando desde Japón y EEUU, comenzaron a usarlo este dio problemas. Solución, replicas espejo alojadas en **AWS**.

-Por otro lado otra solución en la nube es **Cloudera,** que ofrece **SaaS** basado en **Hadoop** y que en colaboración con el **Broad Institute of MIT and Harvard**, el principal centro de investigación biomédica y genómica del mundo está inmerso en una nueva versión de los sistemas **GATK** para la secuenciación del genoma de un paciente de manera mucho más rápida.

6-Genoma humano y Hadoop

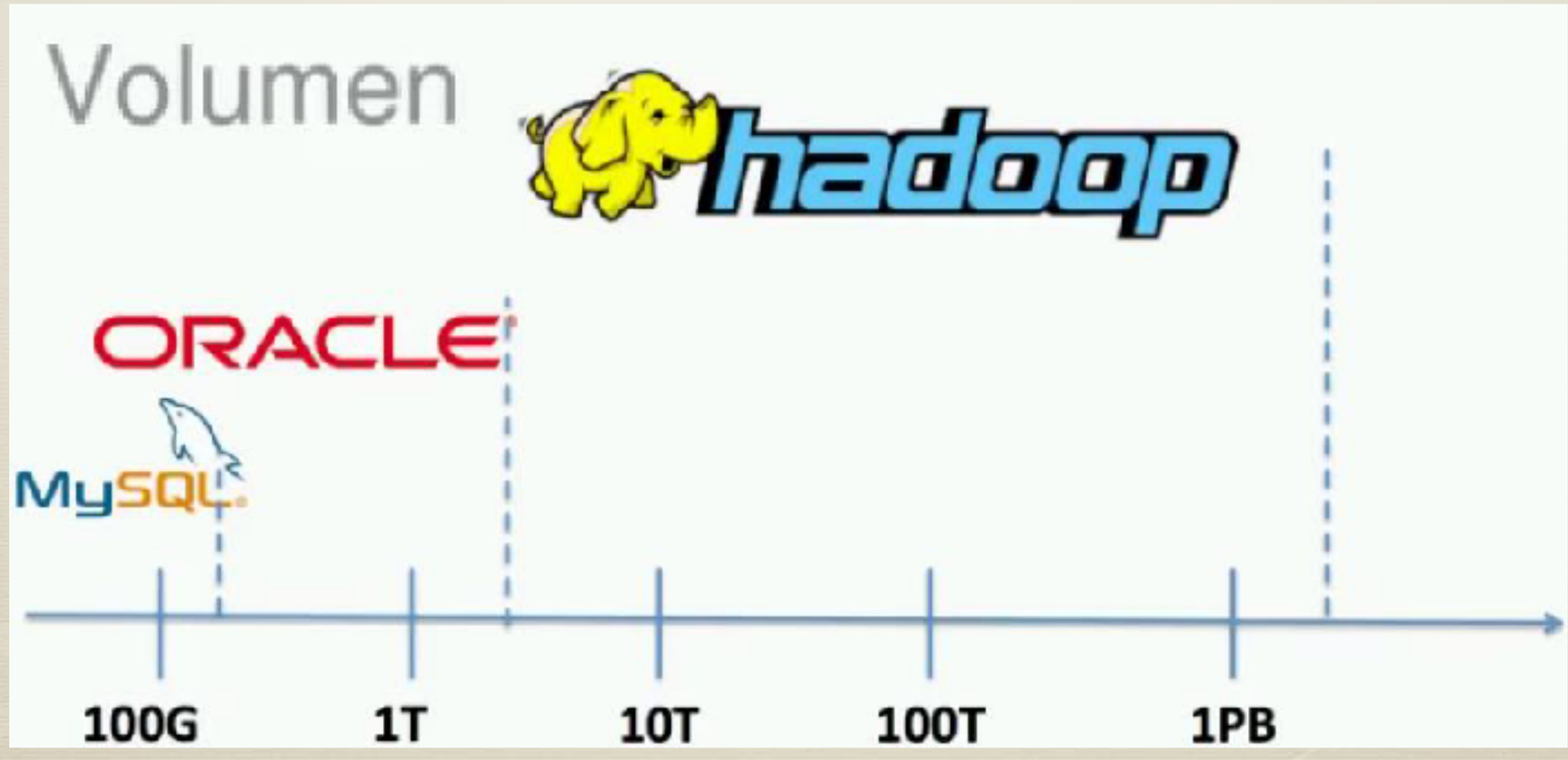
Desde que el proyecto **Genoma Humano**, logró secuenciar el primer genoma humano completo en 2000, los estudios que avalan el conocer el genoma completo de un paciente puede ayudar a **predecir enfermedades** u **ofrecer un tratamiento especializado** para cada paciente han ido en aumento. Tanto es esta expectación que cada vez son más las empresas especializadas en estos aspectos como por ejemplo, **Affymetrix, IBM, Knome, Pacific Biosciences, Complete Genomics**...



Dos cambios importantes de tendencia. **¿Guardan relación con el BigData o el CloudComputing?**

6-Genoma humano y Hadoop

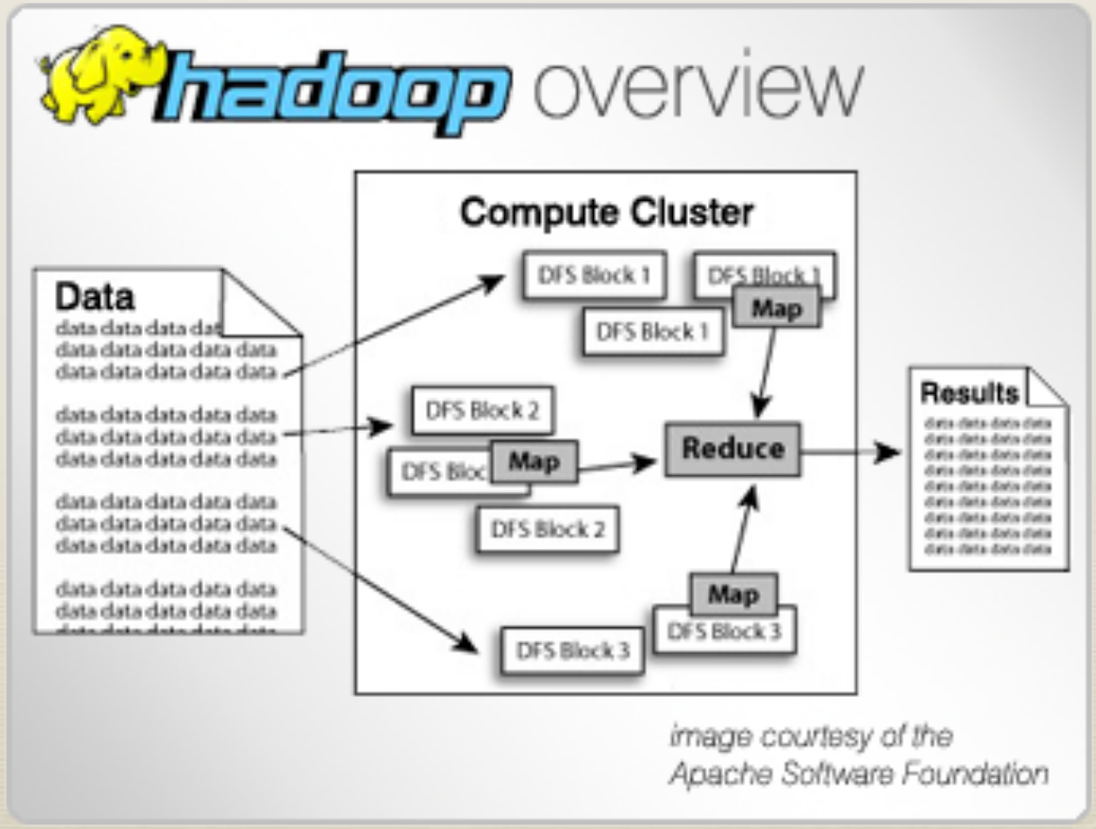
El Release, 0.1.0 de Hadoop es lanzado en 2006, coincide con el nacimiento de las principales plataformas de CloudComputing como Amazon, en Abril de este año, Hadoop ordena 1.8TB usando 188 nodos en 47.9h. Solo dos años después, Hadoop consigue ordenar 1TB en **209 segundos** en un cluster de 910 nodos. En 2008 es lanzado **Cloudera**, que como hemos visto anteriormente es una de las plataformas más usadas en estos menesteres.



6-Genoma humano y Hadoop

Como hemos visto antes, el genoma puede ocupar **140GB**, esto ofrece una gran necesidad de procesamiento que tanto la genética como otras áreas dentro de la biología, necesitan y que es por tanto facilitada por tecnologías como **Hadoop** permitiendo que operaciones complejas cada vez puedan ser realizadas con mayor facilidad y eficiencia.

Sistemas como el mencionado anteriormente, **GATK** (con más de 31000 usuarios activos), están desarrollando nuevas versiones sobre **CloudComputing y Hadoop o Spark** de manera que la **limitación de computacional** versiones anteriores queda suprimida con estas tecnologías.



7- Conclusiones y vías futuras.

1. La aparición del **CloudComputing** y los paradigmas de procesamiento de **BigData**, han supuesto una revolución en la biología permitiendo la compartición, comparación y procesado de grandes volúmenes de datos que de otra manera no podrían ser manejados.
2. La redaccion de costes y versatilidad que el **CloudComputing** ofrece, ha acercado tecnologías punteras a laboratorios e investigadores que de otra manera no podrían haber accedido a ellas.
3. **Hadoop** y los SaaS en la nube basados en esta tecnología, han permitido por tanto que la secuencia del genoma humano de un paciente haya reducido sus costes hasta estar en torno a los **1000$**.
4. Desde la aparición de **Apache Spark**, mucho más rápido que Hadoop, las vías de investigación en el procesamiento y secuencia del genoma humano han girado para adaptar esta nueva tecnología lo que conllevará a redaccion de costes y aumento de la velocidad hasta tal punto que en un futuro próximo obtener el genoma de un paciente pueda ser una tarea rutinaria en centros de salud.

8 -Referencias

1. Artículo BigData en el CERN: http://tecnologia.elpais.com/tecnologia/2013/06/19/actualidad/1371633896_769175.html

2. Revista Nature, Big Data y Cloud Computing en Biología. http://www.nature.com/nature/journal/v498/n7453/full/498255a.html

3. BigData, Hadoop and Cloud Computing in genomics. A. O`Driscoll, J. Daugeaite, R.D Sleator.

4. Artículo Diario TI. Hadoop acelerará la investigación y análisis del genoma humano. https://diarioti.com/apache-hadoop-acelerara-la-investigacion-y-analisis-del-genoma-humano/97432

5. Instituto Europeo de Bioinformática http://www.ebi.ac.uk

6. http://www.ensembl.org/index.html Ensemble Genome Browser

7. https://www.cloudera.com Plataforma Cloudera

Gracias por su atención