Índice

- Introducción
- Big Data
- Map Reduce
- Data Sets disponibles en AWS
- Otros Data Sets

Introducción

- Algunos data sets como los que contienen información sobre el genoma Humano requieren horas o días para su descarga, configuración, análisis etc.
- Con la llegada de la computación en la Nube cualquiera puede acceder a estos datos y analizarlos con instancias (EC2 o EMR).
- Al almacenar estos datos en lugares con capacidad y recursos elásticos, pueden ser procesados de manera más facil, rápida y económica.

Big Data

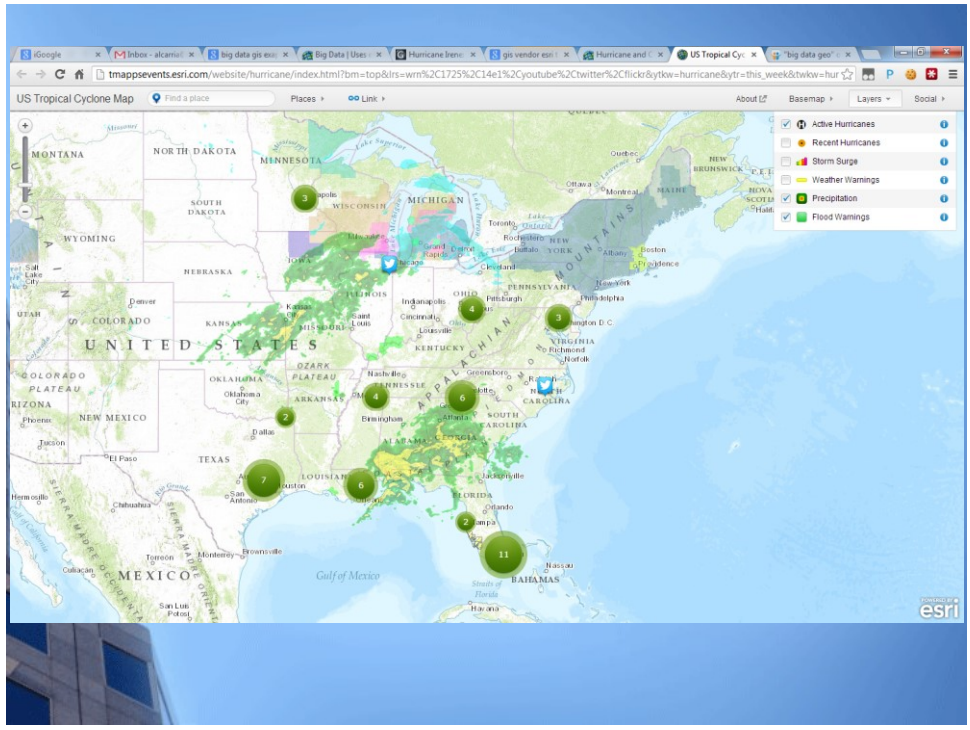
- Tratamiento y análisis de enormes repositorios de datos donde resulta imposible tratarlos con las herramientas de bases de datos y analíticas convencionales
  - Motivado por la proliferación de páginas web, aplicaciones de imagen y vídeo, redes sociales, dispositivos móviles, apps, sensores, internet de las cosas, etc. capaces de generar quintillones de bytes al día
- **El uso de *Cloud Computing* y *Semantic Data* para la gestión de Big Data** en organizaciones es uno de los grandes desafíos en la evolución de la web
  - El paradigma de Linked Data facilita Big Data, al potenciar el principio de hipertexto en documentos RDF.

Big Data

- Big Data promete la generación de conocimiento, crear ventajas sostenibles y competitivas para las organizaciones
- Se necesita Big Data cuando el análisis de información se ve afectado por el Volumen, la Variedad o la Velocidad en el procesamiento de datos:
  - **Volumen**: los datos son demasiado voluminosos para ser gestionados por nuestra infraestructura de datos actual
  - **Variedad**: hay demasiadas fuentes de datos de las que extraer información y en varios formatos (datos estructurados y no estructurados)
  - **Velocidad**: necesitamos de manera ágil obtener conclusiones e información que nos ayude en tiempo real a tomar decisiones

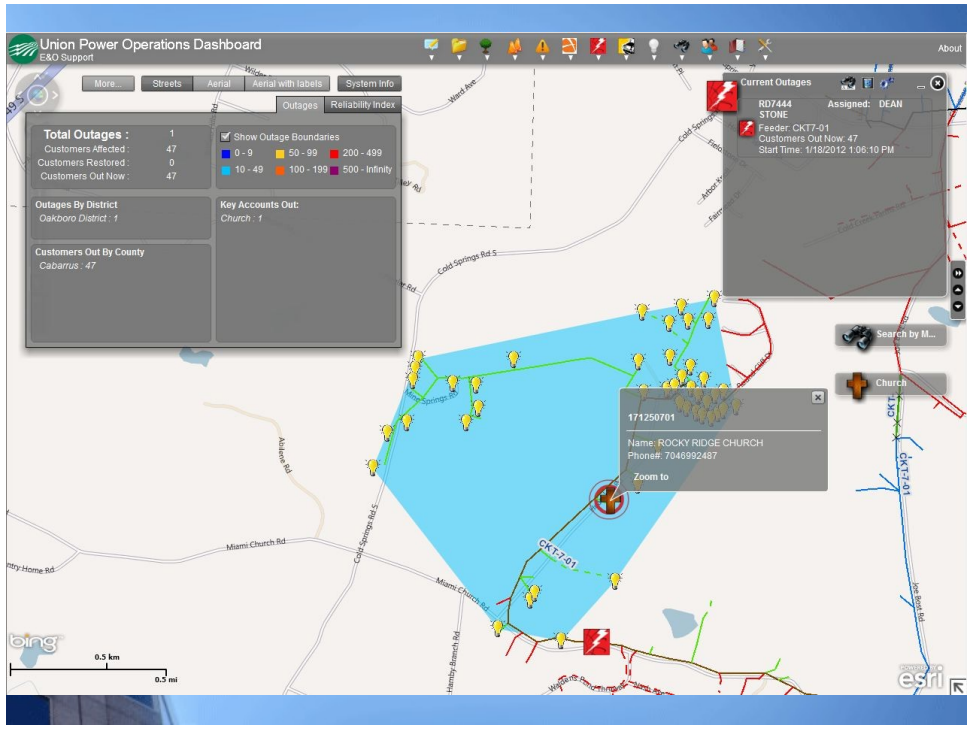
Ejemplos de aplicación de Big Data con visualización Geoespacial

- Planes de contingencia para catástrofes
  - Ver ESRI ArcGIS Mega-aggregator para localizar el Huracán Irene
  - Info de National Hurricane Center, Open Street Map, Telvent e integración con Flickr, Twitter y Youtube.
- Análisis de mercado
  - Compañías de tarjetas de crédito asocian datos de localización con transacciones, información del consumidor y redes sociales.
- Administración pública
  - Detección de fraudes, vigilancia de plagas y epidemias
- Seguros
  - Asociar el lugar de origen de mensajes en redes sociales durante desastres ayuda a las compañías de seguros a monitorizar el impacto económico.
- Recursos naturales
  - La industria petrolera utiliza Tb de datos sísmicos para actividades exploración y extracción



Ejemplos de aplicación de Big Data con visualización Geoespacial

- Marketing
  - Monitorización y filtrado de redes sociales para enviar ofertas dependientes de la localización.
- Telecomunicaciones
  - Los call centers generan inmensas cantidades de datos para detectar patrones y debilidades en las infraestructuras de telecomunicaciones.
- Energía
  - Representación de sensores de medición de consumo para distribución de carga



Map Reduce

- Ejemplo: The New York Times utilizó, en 2007, 100 instancias de Amazon EC2 para procesar 4 TB de imágenes TIFF (guardadas en S3) para generar 11 millones de archivos PDF en 24 horas (240$).
- Amazon ha mejorado los servicios de procesamiento de Big Data con la introducción de EMR (Amazon Elastic MapReduce) en 2009.

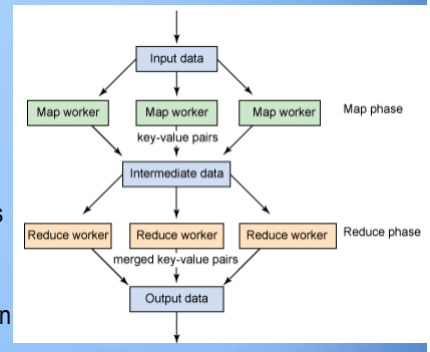
MapReduce

- Introducido por Google en 2004 en el paper “MapReduce: Simplified Data Processing on Large Clusters”.
- Objetivo: permitir la computación paralela sobre grandes colecciones de datos permitiendo abstraerse de los grandes problemas de la computación distribuida.
- Usos: Geodesia, intersección de polígonos, carreteras, elevaciones

Map Reduce

Esta técnica consiste en dos fases: Map y Reduce

- Las funciones **Map** y **Reduce** se aplican sobre pares de datos (clave, valor).
- **Map** toma como entrada un par (clave,valor) y devuelve una lista de pares (clave2,valor2). Se realiza en paralelo.
- El framework MapReduce agrupa todos los pares generados con la misma clave.
- **Reduce** se realiza en paralelo tomando como entrada cada lista de las obtenidas en el Map y produciendo una colección de valores.



Map Reduce

- Explicación del MapReduce

1. Nuestros datos

```
(3414, 'the cat sat on the mat')
(3437, 'the aardvark sat on the sofa')
```

2. Los pasamos al Map

```
map(String input_key, String input_value)
    foreach word w in input_value:
        emit(w, 1)
```

3. El Map genera

```
('the', 1), ('cat', 1), ('sat', 1), ('on', 1),
('the', 1), ('mat', 1), ('the', 1), ('aardvark', 1),
('sat', 1), ('on', 1), ('the', 1), ('sofa', 1)
```

4. El framework (Hadoop) agrupa los datos con la misma clave

```
('aardvark', [1])
('cat', [1])
('mat', [1])
('on', [1, 1])
('sat', [1, 1])
('sofa', [1])
('the', [1, 1, 1, 1])
```

5. Los pasamos al Reduce

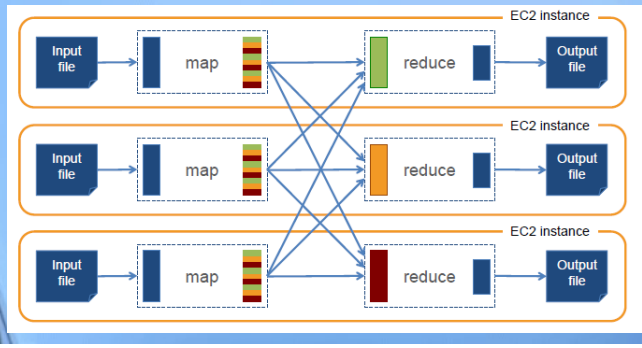
```
reduce(String output_key,
                Iterator<int> intermediate_vals)
    set count = 0
    foreach v in intermediate_vals:
        count += v
    emit(output_key, count)
```

6. Que genera el resultado

```
('aardvark', 1)
('cat', 1)
('mat', 1)
('on', 2)
('sat', 2)
('sofa', 1)
('the', 4)
```

Amazon EMR

- Permite abordar procesamiento de BigData
- Proporciona herramientas parar resolver problemas específicos
- Integrada con todos los servicios AWS (EC2, S3)



Data Sets disponibles en AWS

- Astronomy
- Biology
- Chemistry
- Climate
- Economics
- Encyclopedic
- Geographic
- Mathematics

55 conjuntos de datos, clasificados en distintas categorías:

Los más utilizados:

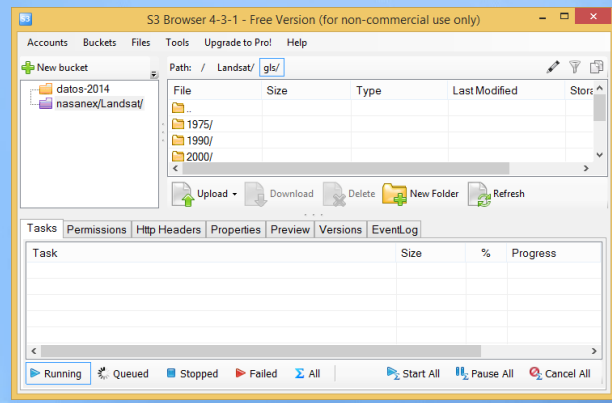
- NASA NEX: A collection of Earth science data sets maintained by NASA, including climate change projections and satellite images of the Earth's surface
- Common Crawl Corpus: A corpus of web crawl data composed of over 5 billion web pages
- 1000 Genomes Project: A detailed map of human genetic variation
- Google Books Ngrams: A data set containing Google Books n-gram corpuses
- US Census Data: US demographic data from 1980, 1990, and 2000 US Censuses
- Freebase Data Dump: A data dump of all the current facts and assertions in the Freebase system, an open database covering millions of topics

Data Sets disponibles en AWS

- Lo más útiles para nosotros
  - **s3://nasanex/Landsat** -> Space-based moderate-resolution land remote sensing data
  - DBPedia -> DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web
  - OpenStreetMap Rendering Database
  - Twilio/Wigle.net database of mapped US street names and address ranges
  - 2008 TIGER/Line Shapefiles for American states, counties, subdivisions, districts, places, and areas

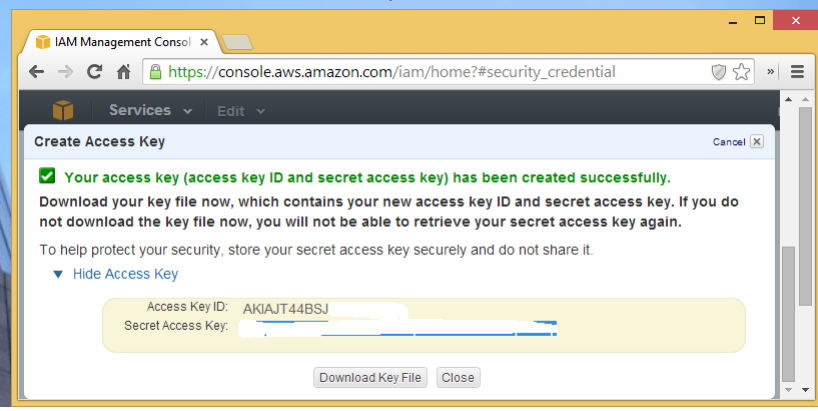
Data Sets disponibles en AWS

- Acceder a los DataSets de AWS
  - **Algunos datasets están en Amazon RDS**
  - **Otros están en S3 -> http://s3browser.com/**



Data Sets disponibles en AWS

- Acceder a los DataSets de AWS
  - **Algunos datasets están en Amazon RDS**
  - **Otros están en S3 -> http://s3browser.com/**



Otros Datasets