Bogotá R Users Group

BIG DATA CON R: SPARKLYR Y SPARKR

José Fernando Zea

jfzeac@gmail.com

febrero 11 de 2020

AGENDA

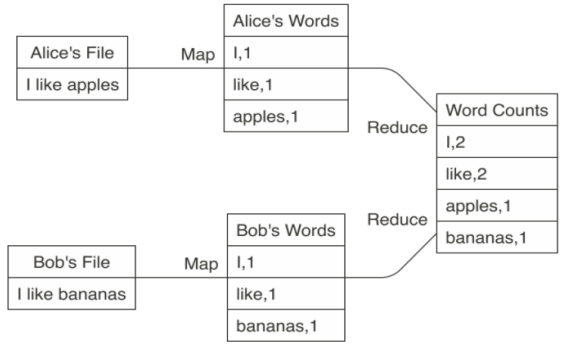
1. Contexto e historia
2. Puesta a punta de herramientas
3. Infraestructura
4. Configuración Databrics Community Edition
5. Ejemplos
6. ¿Cómo profundizo?

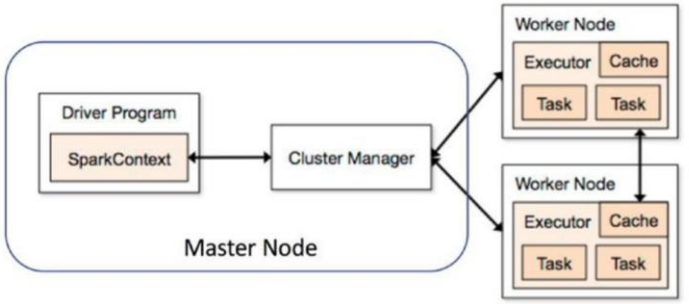
UN POCO DE HISTORIA . . .

Hadoop y MapReduce (2004, 2006): Google, agregar y combinar. HDFS

Hive (2008): Facebook, MapReduce con SQL (Hive)

Spark (2009): Fundación Apache, datos en memoria.
Resilient Distributed Dataset





ANTES DEL WORKSHOP

1. Crear cuenta en databrics community edition:
   https://community.cloud.databricks.com/

2. Crear *cluster* (colocar nombre a cluster. Usar configuración por defecto), esperar aprox 5 mins.

3. En el icono superior dar click y crear nuevo notebook: colocarle nombre y escoger de lenguaje R. El *cluster* creado aparece.

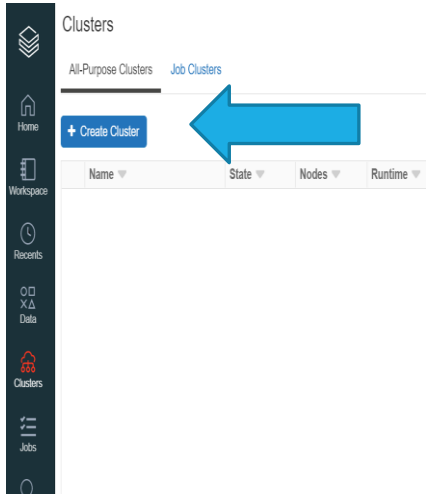
4. Importar Datos: ir al menú Data y a continuación picar en créate table.

Verificar que se selecciona la pestaña Upload File y arrastar el archivo veredas.csv descargado de github (o dar click en browse y buscar el archivo).

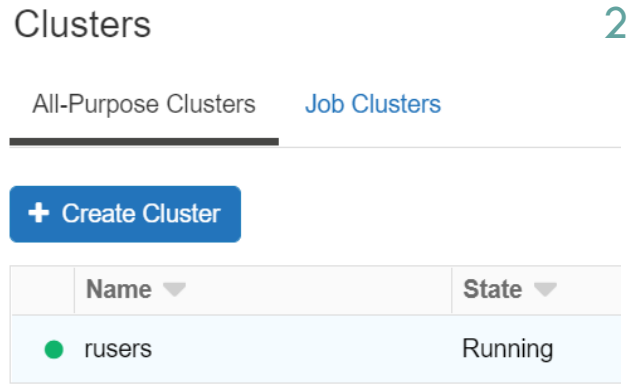
5. Vovler a nuestro notebook creado (En el menú Home)

0. Descargar de veredas.csv de:
https://github.com/bogota-r/big-data-con-r
https://github.com/josezea/BogotaRUsers

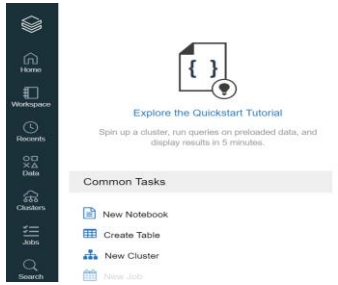
1



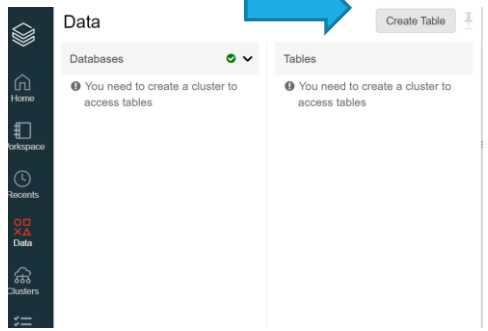
2



3

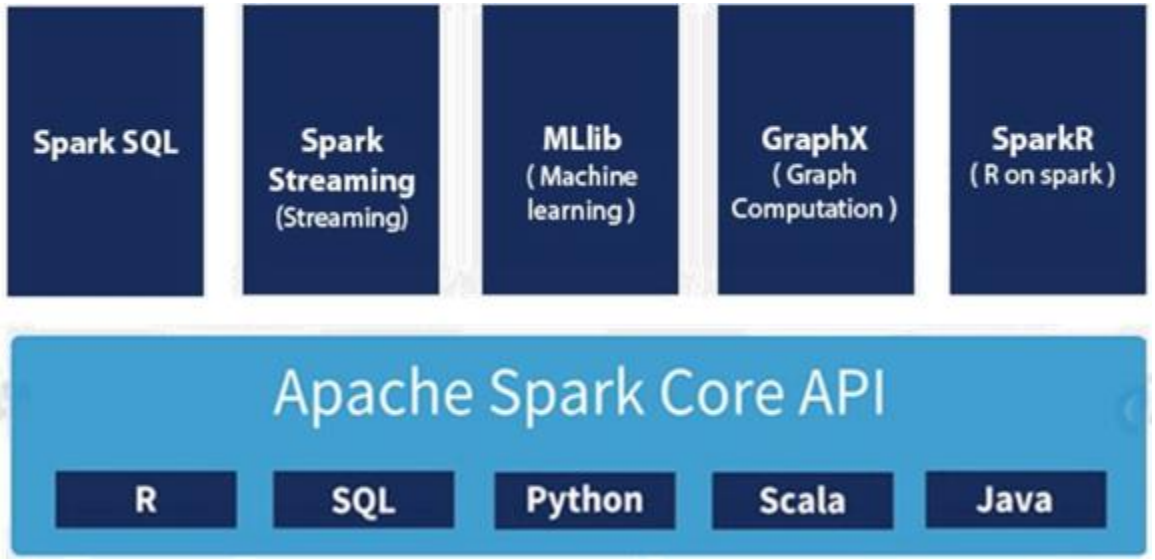


4



Meetup R
Bogotá R Users Group

HERRAMIENTAS DE USO



SparkR

Cloud Computing

amazon EMR

Azure

databricks

Google Cloud

ON-PREMISE

Fuente: Sing (2019), Lurashi et al (2019)

EJEMPLO

- Datos de las veredas del sector rural colombiano:

https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/descarga-nivel-de-referencia-de-veredas/

- https://github.com/bogota-r/big-data-con-r
- https://github.com/josezea/BogotaRUsers