Notas

la predicción y demás fácilmente pues se pueden integrar a otras fuentes de datos como más de datos como lagos de datos como Cassandra, BigQuery en el caso de Google Cloud y demás.

En resumen, SparkML es ideal para proyectos que involucran Machine Learning a gran escala, donde los datos, claramente, no hay. El procesamiento de los datos, pues, no cabe en la máquina.

¿Listo?

veamos un poquito entonces lo que tenemos en el script de nuevo tenemos una carga de nombres de películas solamente para tener un diccionario con los respectivos nombres de películas. sus funciones y demás, ya hemos visto pues que esa estructura de datos no tiene como un interior, entonces aquí se le crea una estructura customizada.

asignándole el ID, el MOVID, el RATING y el momento en el que se hizo ese RATING es otra vez este modo de esquema, el cual se usa cuando se carga el ataque mortal listo

cuando se carga la lata como tal

Bueno, aquí tenemos un print del train recommendation model, y aquí es donde vendría el concepto para lan.

aquí es donde tenemos la parte de la configuración de un modelo als que pues Dentro de SparkML, la librería Alternating Lease Squares en Spark es un algoritmo usado para sistemas de recomendación colaborativos. y a la parte de sistemas de recomendación, en la clase pasada andamos un poco de lo que es la semillitud de coseno, ¿no es cierto?, como para identificar cuándo un vector es más o menos parecido a otro vector, cuándo son iguales, cuándo son completamente diferentes

califican películas, pues fácilmente cada una de las películas o cada uno de los usuarios se pueden ver como vectores, sí, eso es lo mejor. para los que no lo vieron en ese momento cuando habló de la matriz me refiero a esto tenemos una matriz con la que estamos trabajando en este momento de películas donde tenemos usuarios, tenemos películas

calificación y una forma de que los sistemas de recomendación de identificar que tenemos algún tipo de similitud de pronto entre películas pues es visualizar

esta columna de acá y esta columna de acá como un vector o ver si algunos usuarios se parecen respecto a lo que han calificado

pues esto de acá se puede ver como un vector y por lo tanto, pues, por esta naturaleza de los datos por esta naturaleza de los datos, es que se puede aplicar una estrategia de similitud de porcenos para identificar qué tan parecidos o qué tan... diferentes pueden ser estas calificaciones, o en otras palabras, las películas o los usuarios. Por eso se llama colaborativo, porque se necesita que exista la calificación de películas, que existan múltiples películas y que existan múltiples...

usuarios. Si ese volumen de información no existe, pues es muy difícil que esto sea colaborativo. Por lo tanto, hay que repartir ese tipo de estrategias algorítmicas. ¿listo? y... bueno les comentaba antes en esta linea pues básicamente se inicializa un modelo de factorización matricial

a hablar un poquito de lo que es la factorización matricial, pero básicamente lo que busca es realizar el... digamos como un número de interacciones, de entrenamiento para ver cuál es, digamos como la más disquena de su poder. si ajusta a la hora de darles algún tipo de recomendación. Me pregunto a ver si tengo todavía ahí un slide.

No, no, no, no, no, no, no, no,

No, no, no, no, no, no, no, no, Subtítulos realizados por la comunidad de Amara.org

Subtítulos realizados por la comunidad de Amara.org

Bueno, básicamente... Ah, espérate, yo lo acepto de aquí... Ya sabes... Básicamente, eh... recomendación y como es un proceso de optimización pues se le establece un parámetro que es denominado regularización para evitar factores en la frente a la hora de establecer cuál es la mejor base que se ajusta a un modelo de computación.

aquella que representa los ítems a recomendar, en este caso sería los movimovieid. ¿Y cuál es la columna donde se tienen las calificaciones que el usuario han hecho sobre los ítems? ¿Listo? Digamos que en resumen este modelo de ALS lo que está haciendo es entrenar durante 5 interacciones

¿Cuál es la mejor matriz? ¿Cuál va a ser ese hiperparámetro de regularización? ¿Cuáles van a ser los identificados? los identificadores de usuarios, cuáles son

los identificadores de las columnas y cuál es la columna correspondiente al rating. Ya después de definir ese modelo, pues se usa algo como este...

als.fit que básicamente es entrenarlo con un dataframe de entrenamiento en este caso, pues, que sería... los reyes, ¿listo?

Listo, pues por ahora vamos en el tema del incremento de un modelo para establecer una recomendación. Esto a diferencia de la clase pasada que hicimos el tema de Moisés y el artista, básicamente lo que hicimos fue encuentro.

encuentre para una película cuáles son las 10 películas más similares. ¿Cierto? Hasta ahí no se ha hecho ninguna recomendación, sin embargo, perseguí chocarlos en la recomendación. Lo único que trabajamos fue el tema de similitudes de cosas para encontrar similitudes, ya en este caso es para hacer una recomendación a un usuario.

específico. Entonces, digamos, por consola se recibe, pues se recibe un argumento, que en este caso sería un valor correspondiente a un ID de...

Y bueno, definir... aquí lo que... para un solo usuario, a ver si hay recomendación por eso... Reciben... un dataframe y una estructura, una estructura definida, listo, y esto pues lo hace. Yo, la verdad, no sé por qué, pero sé que lo que recibe es una lista de cantidad.

Básicamente, con este... con este model.recommendation for user subsets... a ese usuario se le va a hacer una recomendación de las potenciales 10 películas que le podrían gustar a ese usuario. Y nada, pues ya después de esto es básicamente recorrer el resultado y entender cuáles serían esas recomendaciones para ese usuario dado. Listo.

al ejecutarlo al ejecutarlo

el resultado, básicamente va a ser de este estilo. Las recomendaciones para el usuario 305. y el estado de las comunicaciones listo la ejecución la ejecución de de eso pues es lo básico El nombre del juego es Sparks of Me, el nombre del juego es un poco extraño, en este caso es muy recomiendable desde el DataFrame.

espacio, el argumento que se le va a pasar, en este caso, representaría el ending del usuario el ending del usuario sobre el que se va a hacer la recomendación, listo Entonces, hasta ahí, la parte de Spartan, listo.

Ahora, lo que vamos a hacer, en dos minutitos, va a ser la ejecución de, no de este mismo algoritmo, pero... hay uno que tenemos por ahí, donde se

establezcan similitudes de películas en un clúster real listo, ahí ya empezamos a grabar la sesión, entonces, porfa, dos minutitos

y hacemos la producción de esto, ¿vale? Gracias

con DataFrame CD y además no tiene nada, está vacío básicamente, las músicas de ahí cuando se ejecute pues que esto va a ser en un clúster, si nosotros le pasáramos aquí la información del local y demás, pues ahí le estaríamos diciendo que... los tasks o el task, pero probablemente le vamos a poner uno para su respectiva ejecución.

Entonces... Si, les decía que el va a tomar por defecto que lo va a estar registrando dentro de... dentro de un cluster y aquí es donde viene la parte de los recursos que van a ser utilizados en AWS en este caso, yo utilizaré una VRS la que, recuerda, voy a poner a grabar con el permiso de todos la sesión, ¿vale? para que lo tengan allí

¡Gracias por ver el video!

Entonces, aquí es donde viene lo bueno, si ustedes notan, recuerden que siempre tenemos pues como la información guardada dentro de nuestro... Nuestro, nuestro local.

Ahorita, la fuente de información nuestra va a estar en un buquete, en un buquete de WS Clip, parece ser, y tampoco un buquete. de S3. Listo. En ese caso aquí vamos a tener los ratings, vamos a tener la información de películas y también vamos a tener

Bueno, luego viene entonces el mapeo, pues, de la data, en este caso de los ratings, para tener un diccionario o los elementos que da valor de, eh... de las calificaciones que ha tenido un usuario sobre una película esto tiene la pareja movie ID, rating, dado por una clave que sería el...

El usuario, hay una especificación de la repartición de esos ratings en las diferentes particiones dentro del clúster aquí se le está diciendo que lo haga sobre 100 particiones lógicas básicamente esto se hace para... para realizar los cálculos de los siguientes pasos, sobre todo cuando se van a aplicar elementos como joins o elementos...

.join pues se está haciendo como un self-join de... de... de... r... de... rdd y el resultado va a ser que cada User ID se va a asociar con todas las combinaciones posibles eliminación de duplicados a través de la función que

se estableció en el método. Sí, porque como son pares de películas, pues la idea es que se tengan películas diferentes.

que no se den la oportunidad de que hayan un par de películas iguales. Luego pues hay una transformación con este... bien

con este map, con el UniqueJoinRatings.map donde básicamente la idea es transformar cada una de las entradas en una clave de valor la clave sería película 1, película 2 y los valores rating 1, rating 2 Listo.

Acá tenemos entonces el mapeo de los valores respecto al computo de la similitud de cosenos Esos pares de películas, ¿cómo sería su desimilitud?

Y luego tienen la feedback por parámetro, o por argumento, respecto a...

de la cantidad de ocurrencias, tuvimos la clase pasada, y de ese encuentro de películas que sea lo suficientemente similares al menos un 0,9.

Y que hayan sido ambas, o esos pares de películas, hayan sido rankeadas al menos unas 50 veces. recuerda que el 0.97 viene porque la similitud de cosenos va entre menos uno y uno donde uno va a representar un nivel de similitud alto o el cercano va a tener un nivel de similitud alto y el cercano menos uno va a ser todo lo opuesto

similitud y de ocurrencias igual ya puedes mostrarlo dentro de dentro de la

dentro de la consola las películas más más similares entonces este es el script y se va a hacer sobre un millón de datos, un millón de registros. Necesitamos que tengamos el Movies.Add, ¿cierto?, para los nombres de las películas, y necesitamos también

y el rey listo bueno entonces con base en esto vamos va a pasar a lo que es la consola de AWS ¿Qué es lo que se va a hacer? Se va a crear una instancia de Elastic Map Reduce para tener un clúster real que ejecute el escribir de similitud de películas a través de registros.

Bueno, esto va a necesitar, vamos a necesitar que exista un bucket, si?

Yo, claramente ya tengo un bucket por acá creado, en este caso, este bucket es el ICC Bucket, ¿sí?

¿Y este ojo qué tiene?

Tiene varias cosas. Tiene el registro de información de películas y sus respectivos nombres, tiene el script. Sí.

Tiene los ratings, y tiene los users puntuales. Creo que el de user no lo hemos utilizado. Solamente movies. el script que es el .py y el ratings.dat listo,

entonces ahí ya tenemos el bucket donde va a estar como la fuente de información que se va a utilizar

¿Qué es lo que se va a hacer? Pues que ese clúster se va a utilizar para jalar esa información y guardarlo dentro del servidor. y hacer la ejecución correspondiente ¿listo? eso por un lado es importante por otro lado

Lo que recomendaría, aunque cuando uno empieza a jugar con estas instancias, si de pronto no está creada, cuando uno crea una instancia de MR pues va a requerir que tengas una red privada, una IPC. Y que esa red, esa IPC pues tenga algunas, algunas, algunas redes.

Listo, no vamos a andar mucho en eso, lo que vamos a hacer es crear ya una red, y que esa red tenga algunas subredes, y esto que en este caso... están aquí listas, ¿vale?

Y algo que de pronto puede ocurrir ahorita que estemos creando...

estamos creando la instancia de la Stigma Reduce, es que nos pida unos, no voy a decirlo, como unos usuarios con unos permisos. del uso de recursos. Habrá un usuario que permite manejarla en la instancia de las tic-tac-toe de iOS pero debe tener también los permisos para que desde la instancia de MapReduce se pueda acceder al Pocket

Digamos que no siempre es tan evidente que el usuario lo pueda tener, pero pues si suceden los problemas, pues bueno, ahí lo voy a ver. lo vamos revisando, listo entonces bueno, voy a empezar por aquí a filmar videos

Bueno, realmente he tenido algunos clusters ya funcionando desde hace algunos añitos, otros esta semanita Voy a quitarlos de aquí para no dañar ningún ruido Entonces dentro de MR, damos un MR o damos un MapReduce, claramente los clústeres van a estar basados en instancias de EC2.

¡Listo!

Pero también un segundito...

Bueno, ya quedará el booklet a ver cómo nos va en este caso.

y aquí dice si cluster

Vamos a seleccionar por acá el Spark Interactive, aseguramos de que el Spark 3.5 esté seleccionado. Estos de Hadoop y Jupyter, pues también dejarlos marcados al igual que el Hive, pero sobre todo el Hadoop.

Vamos a dejar que se mantenga un grupo uniforme de instancias S2 de las tareas de ejecución

que de pronto vaya a necesitar o a requerir que el clúster se crezca y se crezca automáticamente y por lo tanto se incremente en... el coro en la parte de ahí. Lo que tenemos es el coro y los taps, ¿sí?

¿Voy a dejar solamente un core y un solo task?

claramente la cantidad de distancias pues ustedes la pueden ajustar si lo necesitan y pues lo que les decía hace un momento es necesario que exista una red privada y PC y una subnet, claramente pues voy a utilizar de las que ya existen, las que ya tengo dentro de esta consola.

de de amazon esta retriba cuando la creé pues la creé con todos los elementos por defecto que me ofrece Amazon ¿listo? todos los elementos por defecto, o sea, no le hice ningún tipo de configuración eeeh Especial. Solamente lo hice con lo que me entrega Amazon por defecto.

¿Alguna de las redes por defecto ¿Alguna de las redes por defecto ¿Alguna de las redes por defecto es esta misma?

Esto de Steps, aquí no le vamos a apurar nada, pero básicamente lo que va a representar que para mi pues...

corresponden aquellos scripts o acciones que quiero que se ejecuten cada vez que la instancia de MRC inicie Si, por ejemplo, quiero que se ejecute un script que establezca una configuración o que establezca un job, que ejecute pues algo a la tarea. con cierta periodicidad no sé, por ejemplo que tenga un script allí adentro que es dejar la información de una base de datos o de un bucket

en un storage, ¿sí? Básicamente, eh, indicarle, eh, lo que va a ser. Mmm, bueno, eso termina. automáticamente

este es importante y bueno, antes de continuar con este punto ¿Hay algo que les cuente? ¿Alguien quiere comentar algo?

No. Gracias. Todo bien, todo bien. Lo que les iba a comentar es que es importante...

Vamos a tener un keypair para poder acceder a LMR, porque vamos a hacerlo a través de un servicio de SSH. Sí.

Entonces, permítanme un segundito, se me ha perdido el descargue.

Bueno, la idea es que ustedes, dentro de S2 vengan a este punto de keypairs y creen una nueva llave esta llave lo que les va a permitir es por poder acceder a ese clúster con esa llave obviamente no podemos dejar ese clúster abierto igual eso tampoco se nos va a permitir yo en mi caso lo que hice es bueno uno le colocó una llave, este es G-Pair por ejemplo

si, le selecciono el RCA, no recuerdo en este momento la diferencia entre ambos, lo que si se es que yo seleccione poder accederlo a través de OpenSSH que pues siento que es como más fácil simplemente lo hago con la terminal de mi equipo y demás Pero, si de pronto tienen Windows, o si quieran hacerlo a través del Patic, pues lo pueden hacer, ¿sí? Y esa clave la van a utilizar para conectarse a la máquina virtual a través del Patic.

yo en este caso pues lo hago con este punto P el te crea la llave y se la descarga ¿listo? descárgala ya vale yo en mi caso ya la tengo descargada es una .pdm y no se llama ICICI Keypad sino que se llama my keybord, listo.

por aquí la terminal activa, y por acá la consola

ahí todavía en la consola en cierto entonces lo que les decía ya después de que ustedes crean esa esa clave pues vienen aquí y le indican

cual va a ser esa llave, en mi caso mi llave se denomina

la pueden buscar desde acá ¿Listo?

Y...

Aquí básicamente pueden, digamos, como ustedes saben que con este sistema de cloud se tiene un acceso a través de IAM. en Google, que se llama Google IAM también, para la gestión de las identidades y la gestión del acceso a los diferentes recursos. Entonces se puede seleccionar un rol ya existente, o pedirle a MR que cree en él.

Es que ya simplemente lo creen Entonces...

A ver si el yoke lo cree.

Va a ponerlo a crear, a crear un nuevo rol, y creo uno nuevo, cuando el va a crear uno nuevo pues habla sobre que recursos va a tener permiso, entonces sobre la misma... red privada, sobre la misma red privada que ya tenemos, sobre la misma subnet, espero es que se esté seleccionando la misma subnet que me seleccionó.

allá arriba y bueno algunos grupos de de seguridad listo esto para y esto para el acceso o el rol que va a tener dentro de las instancias s2

sobre este bucket, lo que el va a hacer, le esta creando unos buckets donde va a colocar unos locks, los locks que van a representar todo lo que se vea en la ejecución Y eso especialmente, pues, si hay errores, pues, que sus errores queden allí.

Allí guardados, ¿listo? Ahora que recuerdo... Voy a ver si le puedo dar de una vez acceso... A este SSJWalkerz

Y lo voy a dar a ver.

Freedom Right ¿Esto por qué? Porque en algún momento cuando estaba creando este DMR para la clase... los archivos que necesitamos para la tarea

Si no tiene acceso a los archivos que hablamos ahorita para la tarea, pues entonces me va a generar problemas, me va a generar errores. que se pueden solucionar sencillamente, ¿listo? Entonces lo que estoy haciendo aquí es decirle oiga la existencia de S2 también acceda a SysEasyBlock.

No, no lo había hecho antes. Vamos a ver si esto me evita ahorita tener que ir a las instancias de IAM y tener que manualmente darle permiso sobre el SSBot, ¿listo? Y listo, creo que por ahora eso es respecto a la creación de DMD, voy a darle aquí Create Cluster

chicos mmm debe poder crearse en al menos unos 5 a 10 minutos Si pasa de los 10 minutos, hay problemas. Realmente no sé por qué van a ocurrir problemas, me ocurrió a... en un principio que nunca lo hacía y después de una hora es que el estatus decía no se puede arrancar.

vida, ¿por qué? ¿Listo?

Bueno, dame unos minutitos aquí, por...

Si.

que quería comentarles y tiene que ver con el acceso a la instancia como yo ya creé una red y una subred y sobre estas redes que están creadas Estas instancias de S12, por lo tanto el clúster, es importante que accedan a través de SSH, pero para acceder a través de SSH...

Y es importante que se tengan los permisos para hacer, en el sentido, en los permisos, digamos, respecto a las restricciones que se tienen de red. Cuando eso se crea, pues, la red es bastante... es algo restrictivo, entonces es importante que el acceso a la máquina pues permita

Permita su respectivo acceso. Entonces, esta ya es como la visual de las instancias de S2. Esta es nuestra instancia. Va a acceder aquí un momentito.

Vamos a ingresar aquí a Security, a Security Groups.

Y aquí en Security Groups chicos, vamos a hacerlo otra vez, esta es nuestra instancia. ingresamos a la lista de instancias de s2

la distancia que teniamos es esta ingresamos a

¡Gracias por ver el vídeo!

y aquí en Security Groups, ustedes observen que debe existir un Inbound Rule corresponda al accesorio tipo ssh protocolo tcp por el puerto 22 si esto lo que va a indicar es que el recurso por el cual va a acceder a través de SSH, pues va a ser la IP, mi IP en este caso, ¿listo?

Eso lo pueden hacer aquí, editando Inbound Flows, adicionando un arreglo, seleccionando FSH. puerto 22 y aquí le dicen que lo coloque Voy a cancelarlo porque pues ya lo tengo ahí listo. Entonces, desearle eso. Eso requisanciado se debe. Para mí esto es un esquivo. Se dice que es impudable. Y la verdad es que lo ya saben y que lo tienen. Y me voy a conectar al servidor otra vez, en el caso pues lo voy a hacer a través de la web.

Aprendí materia, otra vez me acoliste, y nos vemos. simplemente lo que voy a hacer es.

a la llave que yo descargué, tenerla ahí a mano para saber dónde pues va a estar guardada y sobre este elemento pues digamos reemplazar esto por el path correspondiente a donde esté, a donde esté aguillado, y esto pues nada, pues es básicamente la instancia y la dirección de la instancia hacia donde nos vamos

donde nos vamos a conectar entonces voy a copiar esto aquí voy a abrir la terminal por acá

Pegarlo acá

Y, como les decía, básicamente.

¿Qué es lo que?

lo que vamos a hacer ahorita, ya validamos el acceso a DSSH y ahorita lo que vamos a hacer es lo siguiente, a este servidor vamos a descargar dos cosas el .py

y vamos a descargar el movismillarities1million.py y vamos a descargar el movismillarities1million.py listo

Es carga de luz.

¡Migaritis!

Entonces allí le estamos diciendo vaya el S3, vaya el S4, vaya el S5, vaya el S6, vaya ¿En dónde estamos parados?

dicho listo al parecer lo descargó sin problema y lo hemos descargado Vamos a hacer lo mismo para WSS3CP

Vamos a descargarlo desde el S3 y vamos a descargar el mouse. y le voy a decir que lo descargue aquí mismo donde estamos parados. Listo, ya lo descargué.

Y nada, ahorita ya es echarle candela a la ejecucion de... de esas similitudes de películas y vamos a hacerlo para el AIDIDO 260, es una película cualquiera Y creo que es la de Star Wars o me acuerdo que es.

Se desasimula el llamado Spark, Spark Satmilpara micro LID.

Subtítulos realizados por la comunidad de Amara.org

esto es básicamente lo que quería mostrarles respecto a

respecto a la generación de...

a la generación de NR, de la ejecución de un clóset, como tal. utilizando Spark. ¿Listo?

Hasta ahorita, no sé si de pronto hay algún comentario o alguna pregunta.

Y listo, ahora entonces queda todo bien por la parte de la ejecución del Sparking Cluster de Elastic Node Reduce. vamos para la siguiente parte de la clase chicos

que es ya el taller en parejas listo

Entonces...

Y eso está dicho.