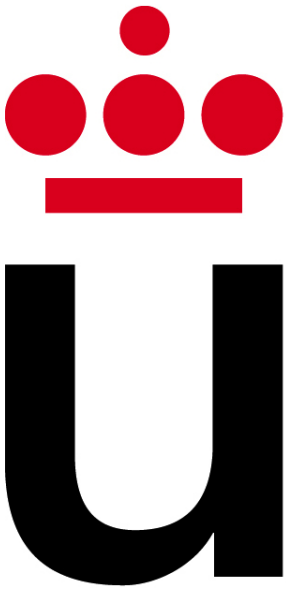
Foro de preguntas

Análisis de Big Data

Preguntas Tema 6

José Ignacio Escribano



Móstoles, 26 de marzo de 2016

Índice

1. Preguntas

1.1. Cloudera Manager es el principal componente de la plataforma Cloudera Enterprise que no está disponible como software libre. Responda a las siguientes preguntas:Resuma brevemente las principales funciones que ofrece. ¿Cuál sería el componente equivalente en la distribución Hadoop de Hortonworks? ¿Considera que es una buena estrategia por parte de Cloudera mantener este componente como propietario? ¿Qué ventajas espera conseguir? ¿Qué riesgos asume ante la aparición de plataformas alternativas como ODPI?

Cloudera Enterprise es la herramienta de Cloudera que permite la gestión de una distribución Hadoop en producción.

Las principales características son:

- Integración y procesado de gran cantidad de datos en menor tiempo (reducción de carga del ETL, creación de pipelines en tiempo real)
- Análisis SQL y Business Intelligence (optimzación del EDW, disponibilidad de dashboard para análisis y descubrimiento de datos)
- Aplicaciones en tiempo real para monitorización y detección.
- Streaming en tiempo real

El equivalente en Hortonworks es Hortonworks Data Platform (HDP). Las principales características son:

- 100 % open-source
- Gran cantidad de tecnologías Apache soportadas (Spark, Storm, Hive, HBase, Pig,Solr, Flume, ...)
- Integración con herramientas de análisis de datos como SAS, SAP, Excel, etc.
- Despliegue tanto en infraestructura propia como en cloud.

Mantener Cloudera Enterprise como componente propietario no parece una buena estrategia, ya que se debe trabajar en mantener dos versiones de Cloudera de forma simultánea (Manager y Enterprise), aunque hasta ahora la estrategia les ha dado buenos resultados, ya que son los líderes del mercado, y además obtienen ingresos licenciando la versión Enterprise.

Si finalmente plataformas como ODPI se afianzan como estándar de Big Data, Cloudera seguramente perderá su posición dominante en el mercado, y deberá adecuarse a este nuevo estándar.
Fuentes:

```
[1]: https://www.cloudera.com/products.html
[2]: ttps://www.odpi.org/
[3]: http://hortonworks.com/hdp/
```

1.2. Apache Zeppelin se está convirtiendo rápidamente en uno de los proyectos que más actividad y atención concentra en el ecosistema Apache Hadoop. Recabe información sobre el proyecto y conteste a las siguientes preguntas: ¿Cuáles son las principales funciones que ofrece Zeppelin? ¿Está ya integrado en alguna de las distribuciones Hadoop de referencia (Cloudera, MapR, Hortonworks, etc.)? ¿Por qué considera que el producto está orientado principalmente a su integración con Apache Spark?

Apache Zeppelin es un creador de notebooks basados en las web para el análisis de datos interactivos.

Algunas de sus características son:

- Documentos colaborativos, transmitiendo los cambios en tiempo real (igual que Google Docs)
- Soporte para Scala y Python (con Apache Spark), SparkSQL, Hive, Markdown y Shell.
- integración con Apache Spark.
- 100 % open-source

Apache Zeppelin es todavía un proyecto en incubación por lo que no está incluido actualmente en ninguna de las plataformas dominantes (MapR, Hortonworks y Cloudera), aunque es posible instalarlo de forma manual en cada una de ellas. Para instalarlo en MapR se puede consultar [2], para hacerlo en Hortonworks, ver [3]. Por último, un tutorial para Cloudera se puede encontrar en [4].

Apache Zeppelin tiene integración incorporada con Apache Spark. Esto hace que no sea necesario construir un módulo o plugin separado para ello.
Esta integración provee de inyección de los contextos de Spark y SQL (SparkContext

y SQLContext, respectivamente), carga de dependencias (en forma de .jar) desde un sistema de archivos local o desde Maven, y es posible cancelar jobs y visualizar su progreso.
Fuentes:

[1]: https://zeppelin.incubator.apache.org/
[2]: https://community.mapr.com/docs/DOC-1493
[3]: http://hortonworks.com/hadoop/zeppelin/#section_3
[4]: http://blog.cloudera.com/blog/2015/07/how-to-install-apache-zeppel