Big Data Mining: In-Database Oracle Data Mining over Hadoop

Zlatinka Kovacheva[1, a)], Ina Naydenova[2, b)], Kalinka Kaloyanova [3, c)] and Krasimir Markov[4, d)]

[1] *Middle East College, Department of Mathematics and Applied Sciences, Muscat, Oman*
[2] *Technologica, Department of Software Development, Sofia, Bulgaria*
[3] *University of Sofia, FMI, Department of Computer Informatics, Sofia, Bulgaria*
[4] *Bulgarian Academy of Sciences - IMI, Sofia, Bulgaria*

[a)] Corresponding author: zkovacheva@hotmail.com
[b)] inaydenova@technologica.com
[c)] kkaloyanova@fmi.uni-sofia.bg
[d)] markov@folbg.com

Abstract

Big data challenges different aspects of storing, processing and managing data, as well as analyzing and using data for business purposes. Applying Data Mining methods over Big Data is another challenge because of huge data volumes, variety of information, and the dynamic of the sources. Different applications are made in this area, but their successful usage depends on understanding many specific parameters.

In this paper we present several opportunities for using Data Mining techniques provided by the analytical engine of RDBMS Oracle over data stored in Hadoop Distributed File System (HDFS). The paper aims to evaluate different approaches for extraction of data over Hadoop for the needs of Oracle Data Mining models building and to give a practical direction for using Oracle Big data decisions.

Some experimental results are given and they are discussed. The results show that it is possible to create scenarios when the data is stored in Hadoop and it is used dynamically in data mining workflows on the base of the Data mining component of Oracle Advanced Analytics. These scenarios can be used successfully in practice.

The dynamical extraction of data from text files stored in Hadoop using SQL Connectors for HDFS is performed in a quite acceptable time, though the productivity in the process of building the data mining model can be slightly improved if the data has been transferred to Oracle tables.

Using such scenarios, the process of the building and usage of a data mining method can remain almost transparent for data scientists acquainted with the data mining component of Oracle Advanced Analytics.

The performed experiments and the comparison analysis can be extended including implementations using Oracle Transparent Gateways and/or R scripts.

Извличане на знания от големи данни: In-Database Oracle Data Mining през Hadoop

Златинка Ковачева[1, а)], Ина Найденова[2, б)], Калинка Калоянова [3, в)] и Красимир Марков[4, г)]

[1] *Близкоизточен колеж, Катедра по математика и приложни науки, Маскат, Оман*

[2] *Технологика, отдел за разработка на софтуер, София, България*

[3] *Софийски университет, ФМИ, катедра "Компютърна информатика", София, България*

[4] *Българска академия на науките – Институт по математика и информатика, София, България*

а) Кореспондент: zkovacheva@hotmail.com

б) inaydenova@technologica.com

в) kkaloyanova@fmi.uni-sofia.bg

г) markov@folbg.com

Резюме

"Големите данни" (big data) поставят предизвикателства по отношение на различни аспекти от съхранението, обработката и управлението на цифрови данни, както и анализирането и използването им за бизнес цели. Едно не по-малко предизвикателство е прилагането на методи за извличане на знания върху масивите от тип "Големи данни" поради огромния обем и разнообразието на информацията, съхранявана в тях, както и динамиката на източниците им. В тази област се разработват различни приложения, но успешното им използване зависи от разбирането на много специфики на данните и правилната им параметризация.

В настоящата статия представяме няколко възможности за прилагане на техники за извличане на знания, предоставени от аналитичните компоненти на СУБД Оракъл върху масиви от тип "Големи данни", съхранявани в разпределената файлова система Hadoop (HDFS). Статията има за цел да оцени различните подходи за извличане на данни, съхранявани в Hadoop за нуждите на изграждането на Oracle Data Mining модели и да даде практическа насока за използване на аналитичните решения на Oracle върху "Големи данни".

В статията се обсъждат получените експериментални резултати. Те показват, че е възможно да се създават сценарии, при които данни съхранявани в Hadoop да се използват динамично в установени работни процеси за извличане на знания, базирани на Data Mining компонента на Oracle Advanced Analytics. Тези сценарии могат да се използват успешно в практиката.

Резултатите показват също, че динамичното извличане на данни от текстови файлове, съхранявани в Hadoop с помощта на SQL Connectors за HDFS се извършва за напълно приемливо време. Производителността на процеса на изграждане на модела за извличане на знания може да бъде леко подобрена, ако данните са прехвърлени в Oracle таблици, но в общия случай това не е необходимо.

Използвайки такива сценарии, процесът на изграждане и използване на метод за извличане на знания може да остане почти прозрачен за изследователите, запознати с компонента за извличане на знания на Oracle Advanced Analytics.

Като насока за бъдещо развитие, извършените експерименти и сравнителният анализ могат да бъдат разширени, включвайки използване на Oracle Transparent Gateways и/или R скриптове.