Storage and recovery of dairy cattle genotype data from the data science approach

Rennan Silva, Fernanda Almeida, Wagner Arbex

Federal University of Juiz de Fora, Embrapa Dairy Cattle

Abstract

The genotyping process consists of the identification of molecular markers, which may vary from individual to individual. Usually, the records that identify these markers are stored in big text files and contains several informations about each individual, like the animal number and the values associated for each marker. There are different patterns to present the genotyping results, depending on the platform chosen for the job. Usually, the data gathered from this process are provided in text files. Analytical tools are not used to treat this kind of data because they are in a very large volume. There are several limitations on relational databases when data is big and unstructured. The goal of this abstract is to propose a way to store the results of the genotyping process (cattle SNP) and allow this information to be accessed from a middleware of ontologies. Besides, this database should provide a background to cross information about animals, individuals, SNPs, samples, etc. This model foresee the creation of an ontology to map some characteristics of the domain and simplify the queries on a genotype database. This ontology will be used on every possible slice of this process, since the data collect to the queries, enabling the data science on this process. Furthermore, this database will store the description of the processes that the data passed by until the moment they are stored, allowing their provenance. Once the proposed database does not fit on the basic principles of relational databases (atomicity, consistency, isolation and durability), it is necessary to implement this work in a less conventional feature, like NoSQL. This database will not need the delete and update procedures, so it may be off the normalization standards to benefit performance in queries. This work will simplify the access to genotype information and may help future works that depends on research over a molecular marker database in bovine dairy cattle.