**Introdução ao Aprendizado de Máquina através de algoritmos com interface *Myriad* na plataforma *HPCC Systems***

Mauro Donato Marques (LexisNexis Risk Solutions)

*Abstract*

*The open-source HPCC Systems platform uses a distributed data architecture and a parallel processing methodology to work with large data sets. The data pipeline in the platform follows data from its source to its ingestion into the Thor cluster, where it is refined through data cleansing and standardization, and then enriched, before being made available to real-time query applications hosted on the Roxie cluster. Together with these two cluster architectures, and forming the so-called Power Trio, the ECL language is present, being the only language required to express data algorithms across the HPCC Systems platform. On the Thor cluster, the ECL language expresses data workflows consisting of data loading, transformation, linking, indexing, etc. On the Roxie cluster, the ECL language defines data query services. The ECL language is implicitly parallel, so the same ECL code written to run on a single-node cluster can run just as easily on a cluster with hundreds of nodes. The programmer does not need to worry about implementing parallelization, and the ECL language has an optimization function that ensures the best performance for a specific architecture. Integrated into the ECL language, there are several Machine Learning libraries available that can be applied to the various categories of Machine Learning, where each algorithm has its own peculiarities, which must be considered to maximize predictive accuracy, considering that all bundles operate in a very similar way, with some minor variations. For these reasons, it is very important not to use Machine Learning to produce products or claim skills without first consulting the documentation that accompanies each bundle to use it effectively. Finally, with the aim of demonstrating the application of Machine Learning algorithms with the Myriad interface on the HPCC Systems platform, a case study will be presented during the “Minicourse” within the ERAD-SP 2025 program.*

*Resumo*

*A plataforma HPCC Systems de código aberto utiliza arquitetura de dados distribuída e uma metodologia de processamento paralelo para trabalhar com grandes conjuntos de dados. O "pipeline" de dados na plataforma acompanha os dados desde a origem até a sua ingestão no cluster Thor, onde eles são refinados através de limpeza e padronização dos dados e, em seguida, enriquecidos, para depois, serem disponibilizados para aplicações de consulta em tempo real hospedadas no cluster Roxie. Em conjunto com essas duas arquiteturas de clusters e, compondo o chamado Power Trio, se faz presente a linguagem ECL, sendo a única linguagem necessária para expressar algoritmos de dados em toda a plataforma HPCC Systems. No cluster Thor, a linguagem ECL expressa “workflows” de dados que consistem em carregamento dados, transformação, vinculação, indexação etc. No cluster Roxie, a linguagem ECL define serviços de consultas de dados. A linguagem ECL é implicitamente paralela, portanto, o mesmo código ECL desenvolvido para ser executado em um cluster de nó único pode ser executado com a mesma facilidade em um cluster com centenas de nós. O programador não precisa se preocupar em implementar a paralelização, e a linguagem ECL possui uma função otimizadora que garante o melhor desempenho para uma arquitetura específica. Integrado à linguagem ECL encontram-se disponíveis diversas bibliotecas de Machine Learning aplicáveis às várias categorias de aprendizado de máquina, onde cada algoritmo tem suas próprias peculiaridades, as quais devem ser levadas em consideração para maximizar a precisão preditiva, tendo em conta que todos os "bundles" operam de maneira muito semelhante, com algumas variações menores. Por esses motivos, é muito importante que não seja utilizada a exploração de Machine Learning para produzir produtos ou reivindicar habilidades sem antes consultar a documentação que acompanha cada "bundle", a fim de usá-lo de forma efetiva. Finalmente, objetivando demonstrar a aplicação de algoritmos de Aprendizado de Máquina com interface Myriad na plataforma HPCC Systems, será apresentado um estudo de caso durante a realização do “Minicurso” dentro da programação do ERAD-SP 2025.*