Engenharia de Dados com Hadoop e Spark

Como Tratar a Mudança de Versão de Software

Vou explicar aqui uma questão importante, já que temos alunos que não vieram da área de TI.

Oportunidade de aprendizado. Leiam com atenção.

Como profissionais de dados, precisamos compreender que a tecnologia evolui. E em Ciência de Dados evolui ainda mais rápido. É algo que precisamos incorporar ao nosso trabalho e dia a dia.

Todo software tem versão e release. Uma nova versão costuma trazer diversas mudanças, enquanto um novo release costuma ser correção de bugs e falhas de segurança ou pequenas mudanças.

Por exemplo:

A Linguagem R recebeu recentemente uma nova versão, passando da versão 3.x para a versão 4.x.

Uma mudança de release seria uma atualização de 3.5 para 3.6 (nesse caso chamado de major release) ou 3.6.1 para 3.6.2 (nesse caso chamado de minor release), por exemplo.

Isso vale praticamente para qualquer software, incluindo linguagens de programação. Python, Java, Hadoop, Spark, etc...

Com ferramentas open-source, a evolução é muito rápida e 2 ou 3 releases costumam ser lançados por ano, com uma nova versão a cada 1, 1.5 ou 2 anos, em geral.

Quando muda a versão, como aconteceu recentemente com a Linguagem R, vários pacotes deixam de funcionar. Por quê? Porque o pacote não suporta alguma mudança na nova versão.

Como resultado, ficamos em um período "nebuloso" por 3 a 4 meses, até que os desenvolvedores migrem seus pacotes para funcionar com a nova versão. É assim desde plugins do Wordpress, até apps nos smartphones.

E aí entra onde quero chegar.

Não podemos migrar o curso para a nova versão da Linguagem R agora, pois vários pacotes ainda não foram migrados e não funcionarão e ao manter a versão atual, alguns pacotes já migrados podem não funcionar. E como fazemos então?

Toda solução open-source tem o chamado “archive”, onde mantém TODAS as versões do software desde a versão 1. Todas as versões e releases ficam no archive.

Quando eu acesso o site principal do software e a versão não está disponível, eu sigo esses passos:

1- Abro o Google

2- Digito (por exemplo): spark archive download

3- Baixo a versão do meu interesse

Na maioria dos casos é o primeiro ou segundo link da busca. E então instalo a versão que eu preciso.

No caso da Linguagem R por exemplo, eu ainda estou usando a versão 3.6.3. Só migrarei para a versão 4.x quando o período “nebuloso” a que me referi acima passar, dentro de 3 a 4 meses do lançamento. Estamos agora no segundo mês.

Mas e se um pacote foi migrado e parou de funcionar? O que eu faço? Simples: busco a versão anterior e instalo offline. Sigo esses passos:

1- Pesquiso no Google, por: nome_pacote r package archive

2- Acesso o archive e faço download da versão que preciso (arquivo zip)

3- E então instalo assim:

```
install.packages('caminho_para_o_pacote/package.zip', repos = NULL)
```

Ou seja, instalo a partir do arquivo com a versão anterior. Mas tem que colocar repos = NULL, para que o instalador não tente buscar no repositório ativo da Linguagem R.

Essa dica também é boa para quem usa Linguagem R em empresas que bloqueiam o download de pacotes (caso de muitos alunos). Apenas verifique com a TI se não há problema em fazer a instalação na máquina.

O problema é se o pacote tiver muitas dependências. Ai, tem que repetir o procedimento para cada pacote.

Essa é a regra para qualquer software open-source. Em Python é bem mais fácil, pois basta informar a versão ao usar o pip, por exemplo:

pip install tensorflow==2.2.0 (sinal de igual duas vezes mesmo)

Já pensamos em criar na DSA o nosso repositório com as versões dos softwares que usamos nos cursos, para tentar reduzir os chamados de suporte. Mas eu sou contra isso, pois acho que podemos orientar e ensinar aos alunos um pouco mais do que vieram buscar, já que estão aqui para aprender e são todos profissionais. O Suporte DSA está orientado a explicar ao aluno o problema da mudança de versão e ensinar como resolver o problema, pois isso será útil no futuro quando o aluno estiver usando a solução no dia a dia.

Lidar com mudança de versão de software faz parte do trabalho, pois será realidade com a qual teremos que conviver. Ou será que alguém ainda usa Windows 98?

Estou colocando este texto em um manual em pdf no Capítulo 1 de todos os cursos da Formação Cientista de Dados, que é onde os alunos costumam ter mais dificuldade com as mudanças de versões.

Tão logo a Linguagem R esteja com o release estabilizado, provavelmente em 4.2.1 ou 4.2.2 faremos a migração do material do curso, conforme necessário.

Equipe DSA