



PUC Minas

IEC - Instituto de Educação Continuada

Pós-Graduação em Ciência dos Dados e Big Data

**Recuperação da Informação na Web e
em Redes Sociais**

Análise dos Principais Erros Encontrados no SQL Server

Aluno: Raquel Cardoso Lemos

Professor: Zilton Cordeiro Jr.

Março

2019



PUC Minas

IEC - Instituto de Educação Continuada

Pós-Graduação em Ciência dos Dados e Big Data

Projeto Final

Análise dos Principais Erros Encontrados no SQL Server

Trabalho apresentado ao Instituto de Educação Continuada (IEC) da pós-graduação em Ciência dos Dados e Big Data da PUC Minas, como requisito parcial para a obtenção de créditos na disciplina de Recuperação da Informação na Web e em Redes Sociais.

Aluno: Raquel Cardoso Lemos

Professor: Zilton Cordeiro Jr.

Março

2019

Conteúdo

1 Resumo.....	4
2 Introdução.....	4
3 Descrição das Atividades.....	4
4 Análise dos Resultados.....	10
5 Trabalhos Futuros.....	11

1 Resumo

O presente trabalho tem como objetivo demonstrar a utilização da ferramenta KNIME para a análise dos dados relacionados aos principais erros encontrados e corrigidos no Sistema de Gerenciamento de Banco de Dados (SGBD) SQL Server, de forma a identificar quais são os erros mais comuns em todas as versões suportadas atualmente pela Microsoft.

2 Introdução

Os SGBDs são sistemas complexos, com várias camadas de abstração e que suportam operações críticas ao negócio das empresas. Em virtude da complexidade e da importância desses sistemas, não é pouco comum a identificação de bugs (erros) e posteriormente o lançamento de correções para as versões dos produtos suportados pelas empresas fabricantes.

No caso do SGBD SQL Server, quando um bug é identificado e corrigido no produto, a Microsoft disponibiliza uma atualização para ser aplicada nas versões compatíveis. As informações relacionadas aos erros, assim como as correções, são disponibilizadas no site de suporte da Microsoft, em páginas que são conhecidas como KB (*Microsoft Knowledge Base*).

Por meio da análise dos KBs disponibilizados pela Microsoft, é possível identificar quais são os bugs mais comuns identificados nas versões suportadas atualmente: SQL Server 2008, 2012, 2014, 2016 e 2017. Dessa forma, o trabalho tem como objetivo realizar a coleta dos KBs relacionados ao SQL Server, efetuar a análise e identificar quais são os bugs mais comuns encontrados nas versões atualmente suportadas pela Microsoft.

3 Descrição das Atividades

3.1 Coleta dos dados

A coleta dos dados relacionados aos KBs do SQL Server foi feita por meio de RSS Feeds. A Microsoft disponibiliza no site de suporte a opção de subscrição de feeds para

cada produto¹. Dessa forma, foi efetuada a subscrição para os KBs de cada versão do SQL Server.

Os feeds foram coletados utilizando a ferramenta KNIME, e os resultados das coletas foram salvos em um banco de dados SQL Server, havendo uma tabela correspondente para cada versão do SQL Server coletada, conforme ilustrado nas figuras 1 e 2.

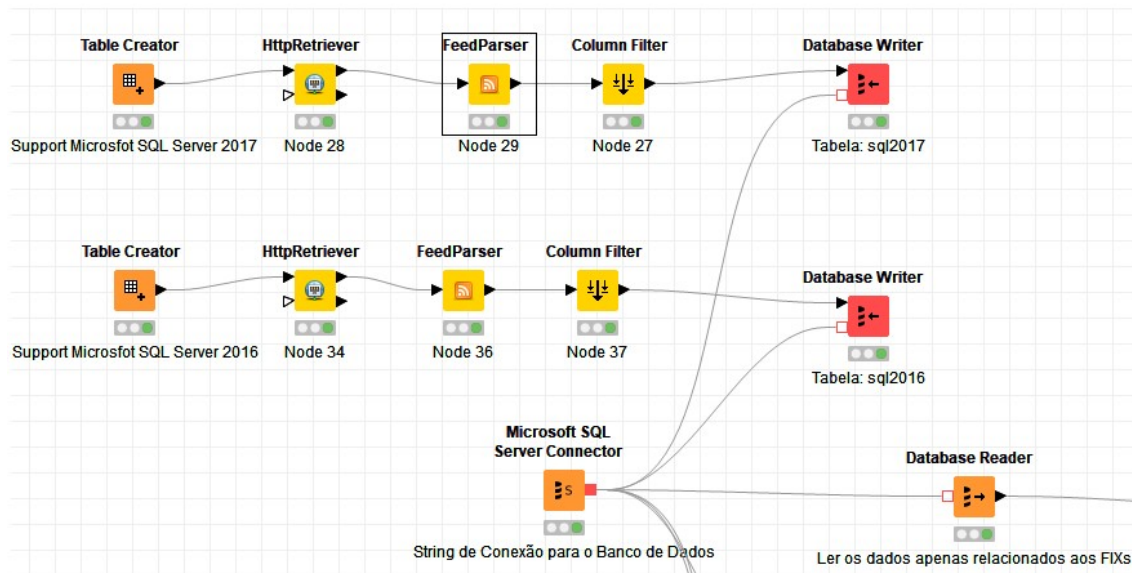


Figura 1 - Coleta dos feeds das versões do SQL Server

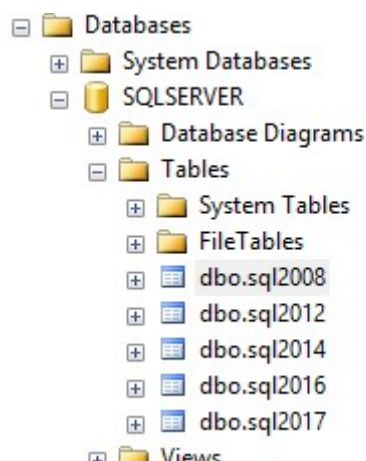


Figura 2 - Tabelas no banco de dados

Foram coletados 3963 feeds, sendo que 2426 eram associados às correções de bugs (Fixs) do produto, os quais foram utilizados para efetuar a análise proposta pelo presente trabalho.

¹ Disponibilizado em <https://support.microsoft.com/en-us/help/4089498>

A recuperação dos dados coletados e armazenados no banco de dados foi efetuada utilizando o node Database Read do KNIME, por meio da consulta SQL nas tabelas envolvidas e utilizando filtro pelos KBs cujo título começava com a palavra FIX. A consulta executada para recuperação dos dados é ilustrada na figura 3.

```
SQL Statement

select 'SQL Server 2017' as sql_version, 'FIX' as type, title, description
from sql2017
where title like 'FIX%'
union
select 'SQL Server 2016' as sql_version, 'FIX' as type, title, description
from sql2016
where title like 'FIX%'
union
select 'SQL Server 2014' as sql_version, 'FIX' as type, title, description
from sql2014
where title like 'FIX%'
union
select 'SQL Server 2012' as sql_version, 'FIX' as type, title, description
from sql2012
where title like 'FIX%'
union
select 'SQL Server 2008' as sql_version, 'FIX' as type, title, description
from sql2008
where title like 'FIX%'
```

Figura 3 - Consulta para recuperar os dados do banco

3.2 Tratamento inicial dos dados

Com o objetivo de efetuar uma normalização inicial dos dados, e evitar que palavras muito frequentes e comuns impactassem a etapa de análise, foi realizada a remoção dos seguintes termos: SQL, Server, fix e fixes. Os termos citados foram removidos do título e da descrição dos documentos coletados na etapa anterior. A figura 4 ilustra a etapa de tratamento inicial dos dados.

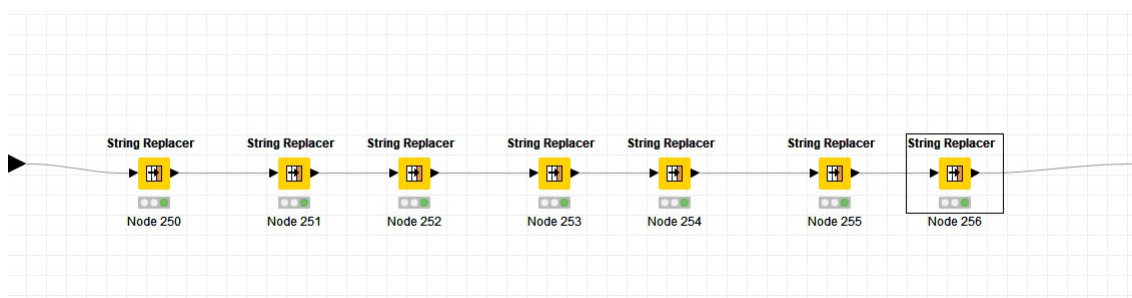


Figura 4 - Tratamento inicial dos dados

3.4 Enriquecimento do texto

Para uma análise mais refinada e assertiva em relação aos documentos coletados e pesquisados, foram utilizados os nodes de enriquecimento do KNIME, que possuem como finalidade principal alterar a granularidade dos documentos por meio de inserção de tags. Para a realização da atividade foram utilizados os nodes POS Tagger, Document Data Extractor, Stanford Tagger e OpenNLP NE Tagger, conforme ilustrado na figura 5.

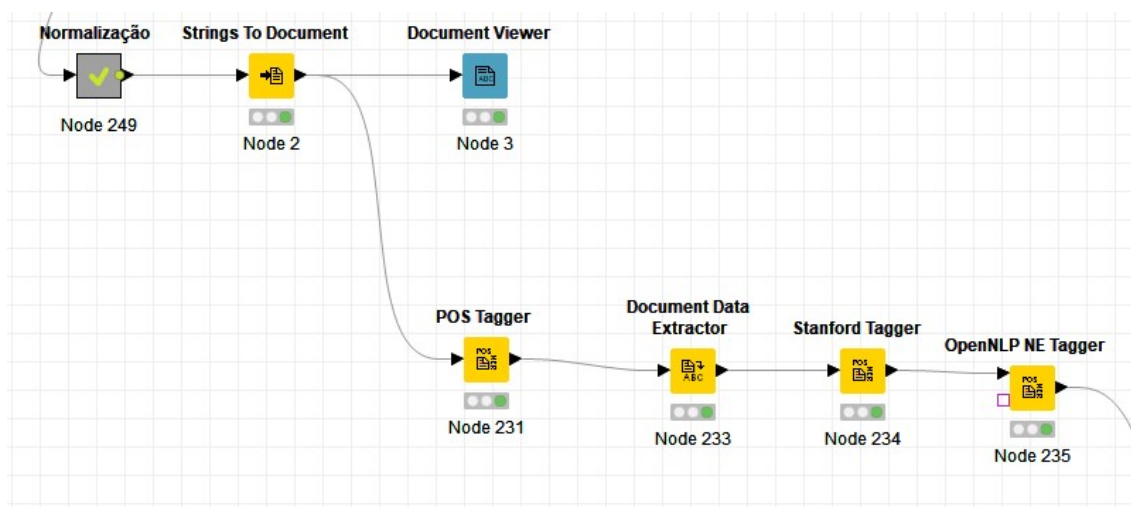


Figura 5 - Nodes de enriquecimento do texto

De forma a possibilitar ainda um refinamento mais adequado para a análise dos documentos, foi utilizado o node Dictionary Tagger, tendo como entrada os documentos refinados com as tags de enriquecimento e uma tabela com os termos relacionados aos erros mais comuns encontrados nos SGBDs, que serão utilizados para pesquisa dos documentos. Os erros comuns foram selecionados com base em uma experiência prévia de administração do SQL Server. A tabela com os termos utilizada nessa etapa do trabalho é apresentada na figura 6.

Manually created table - 0:119 - Tabl

File Hilite Navigation View

Table "default" - Rows: 20 Spec - Column:

Row ID	S Dictionary
Row0	cpu
Row1	memory
Row2	disk
Row3	security
Row4	buffer
Row5	connect*
Row6	IO
Row7	security
Row8	SSPI
Row9	crash
Row10	violation
Row11	processor
Row12	deadlock*
Row13	performance issue
Row14	cach
Row15	access
Row16	transaction*
Row17	communication
Row18	odbc
Row19	utilization

Figura 6 - Termos utilizados para enriquecimento

3.5 Análise dos dados

Para a etapa de análise dos dados, foram utilizadas três abordagens distintas, sendo elas:

- Utilização de N-gramas: objetivo de extrair dos documentos sequências de palavras encontradas. Para o trabalho optou-se pela utilização de trigramas, por ser mais significativo para a análise proposta.
- Análise por meio de TF: objetivo de efetuar a análise por meio dos termos com maior frequência nos documentos.
- Análise por meio de coocorrência de termos: objetivo de verificar a frequência de termos que aparecem em conjunto nos documentos.

A etapa de análise dos dados é apresentada na figura 7. Para a visualização dos dados foi utilizado o node Tag Cloud, que gera uma nuvem de palavras com o resultado obtido do processamento do texto.

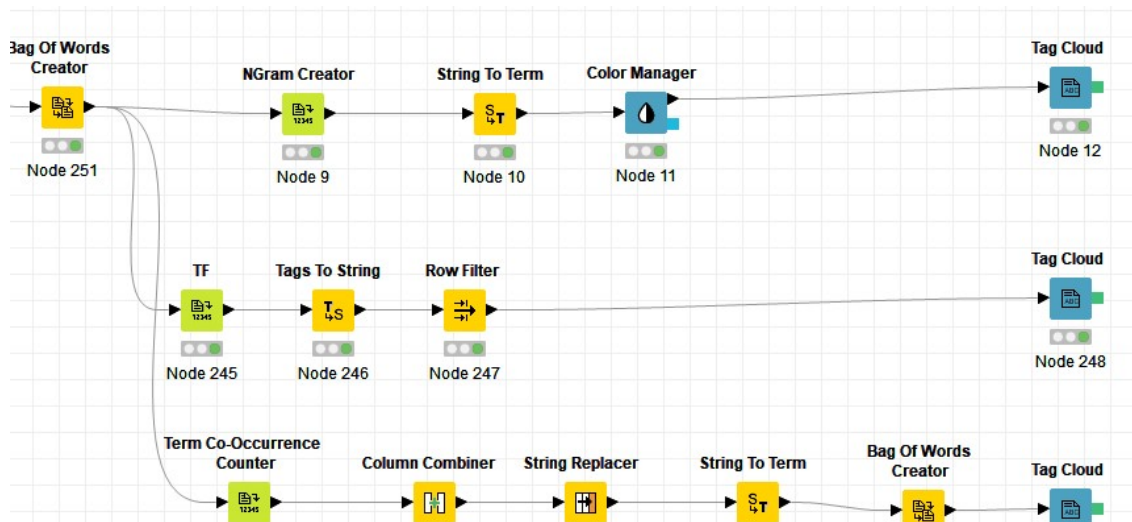


Figura 7 – Análise dos dados

As nuvens de palavras geradas pra cada um dos processamentos efetuados são ilustradas nas figuras 8,9 e 10, respectivamente.

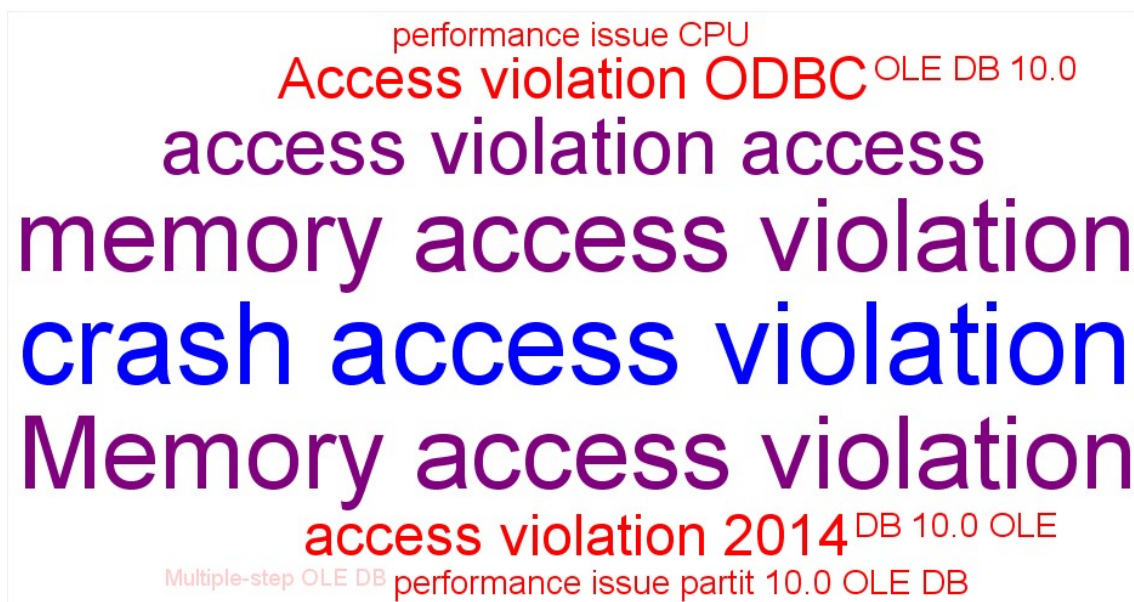


Figura 8 - Visualização NGram



Figura 9 - Visualização por frequência



Figura 10 - Visualização por coocorrências

4 Análise dos Resultados

Por meio da análise dos dados obtidos é possível perceber que a maioria dos bugs (erros) no produto SQL Server estão relacionados com problemas de acesso ao endereçamento de memória (*Memory Access Violation*). Esse tipo de erro ocorre quando o SQL Server tenta acessar um endereçamento de memória que é inválido ou não está mais disponível, e é caracterizado como um bug do produto.

Quando o tipo de erro de endereçamento de memória ocorre, a consulta que estava em execução falha e o SQL Server emite um erro e gera um arquivo com o log, conhecido como arquivo de dump. A solução de contorno para a maioria dos cenários envolvendo

o tipo de erro é a aplicação de correções, que podem ser efetuadas por meio de aplicações de *Services Packs* e *Cumulative Update*.

Conforme apresentado na introdução do trabalho, SGBDs são sistemas críticos que suportam o funcionamento de muitas aplicações importantes para as empresas. Dessa forma, a análise dos principais bugs apresenta a importância de sempre mantermos os SGBDs atualizados com as últimas atualizações disponibilizadas pelo fabricante, de forma a evitarmos indisponibilidades ou erros indesejados.

5 Trabalhos Futuros

Para os trabalhos futuros pode-se verificar a possibilidade de um cruzamento das informações contidas em blogs e fóruns de discussão, como o Stack Overflow e MSDN, juntamente com as informações dos KBs da Microsoft, coletadas e analisadas no presente trabalho. O cruzamento das informações permitiria verificar se os erros reportados pelos usuários da comunidade técnica já foram previamente corrigidos e divulgados nos KBs da Microsoft, ou se são bugs ainda sem correção oficial no produto.