KDD e ETL

September 13, 2022

1 KDD - KNOWLEDGE DISCOVERY FROM DATABASES

Descoberta de Conhecimento em Bases de Dados

Conceito Proposto por Fayyad

Quantidade de dados produzida no mundo duplica a cada 2 anos

"[...] até 2011 a previsão é que 1,8 zetabyte de dados tenha sido criado e replicado."

1.1 O que é Descoberta de Conhecimento em Bases de Dados (KDD)?

A Descoberta do conhecimento em Bases de Dados é um processo **interativo** para identificação de padrões que sejam **novos** e **válidos** nos dados e que sejam úteis na **interpretação dos dados** e **tomada de decisão**.

1.2 Tipos de dados

1.2.1 * Dados não estruturados

1.2.2 * Dados semi-estruturados

1.2.3 * Dados estruturados

2 Data warehouse

Armazéns de dados

2.1 Datasmart

2.2 Áreas relacionadas à área de KDD:

- Mineração de dados
- Sistemas de Apoio a Decisão
- Análise Inteligente de dados
- BI

2.3 Processo

- Seleção -> Dados que serão analisados
- Pré-Processamento de dados -> Dados pré-processados
- Transformação -> Dados Tratados
- Mineração de dados -> Padrões dos dados

- Interpretação -> Conhecimneto

2.4 ETL - Extract Transform Load

Definição -> processo de carga de dados, utilizado em integração de sistemas baseados em software.

2.5 Divisão do Processo ETL

Extração Consiste em se comunicar com outros sistemas ou bancos de dados para captura de dados que serãi inseridos no destino.

Transformação Apresenta diversas etapas sendo elas: * Padronização * Limpeza * Qualidade

Carga (Load) Etapa final, nela os dados são lidos da área de preparação de dados e carregados no Data Warehouse ou DataSmart

2.6 Vantagens das Ferramentas de ETL

- Garantia da Qualidade de dados devido a solução de problemas complexos;
- Funicionalidades prontas para execução;
- Manutenção das cargas, mais simples em manutenbilidade de código;
- Metadados gerados e mantidos de forma autonoma;
- Performance
- Transferência - possibilidade de portabilidade para os mais diversos sistemas e padrões de armazenamento;
- Conectividade
- Reinicialização da carga de dados
- Segurança e estabilidade

[3]:
```
import petl as etl
```

[ ]:
```

```