BIG DATA

VARIEDADE DE DADOS

MONGODB
NEO4J
SOLR

Humberto Sandmann
humberto.sandmann@gmail.com

Roteiro do curso

Aula 1: Introdução

Conceitos, mercado, tendências e arquitetura

Aula 2: Ferramentas

Hadoop, MongoDB, Neo4j e Solr

Aula 3: HBase, Hive e tratamento de dados

Aula 4: Machine learning

Perceptron - MLP, DeepLearning, SOM,

Redes probabilística (Redes Bayesianas)

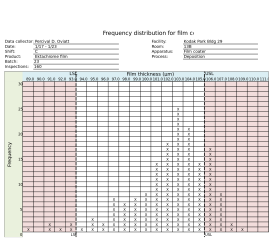
Variedade

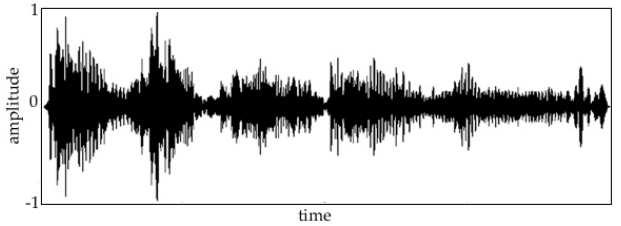
Diferentes fontes de dados

- ERP, CRM, planilhas, Facebook, google, etc

Diferentes formato de dados

- Documentos, planilhas, JSON, vídeos, sons, XML, etc





Teorema CAP

Consistency

Availability

Partition tolerance

Quando se armazena dados, apenas duas características podem ser cumpridas simultaneamente

Modelagem

Dados normalizáveis

- suportam instruções SQL na grande maioria
- indexáveis
- busca ótima
- consistentes

Dados não-normalizáveis

- necessitam de linguagem específica: Not Only SQL
- não indexáveis
- heurísticas de buscas
- nem sempre ou nunca consistentes

Modelagem orientada a documentos

Armazenagem de documentos em geral, sem restrições

- alta performance
- baixíssima ou nenhuma consistência
- schemaless

Ferramentas:

- MongoDB
- Cassandra

MondoDB

Orientado a documentos

- alta disponibilidade
- alta performance
- sem consistência alguma
- utiliza formato baseado em JSON, o BSON, com suporte ao binário

```
{
"_id" : 1,
"name" : { "first" : "John", "last" : "Backus" },
"contribs" : [ "Fortran", "ALGOL", "Backus-Naur Form", "FP" ],
"awards" : [
          {
            "award" : "W.W. McDowell Award",
            "year" : 1967,
            "by" : "IEEE Computer Society"
          },
          { "award" : "Draper Prize",
            "year" : 1993,
            "by" : "National Academy of Engineering"
          }
]
}
```

Exemplo

Modelagem orientada a grafos

Armazenagem de grafos

- consistência
- schemaless
- baixa performance de busca
- estruturado por vértices e arrestas (nós e ligações)

Ferramentas:

- GraphDB
- Neo4j

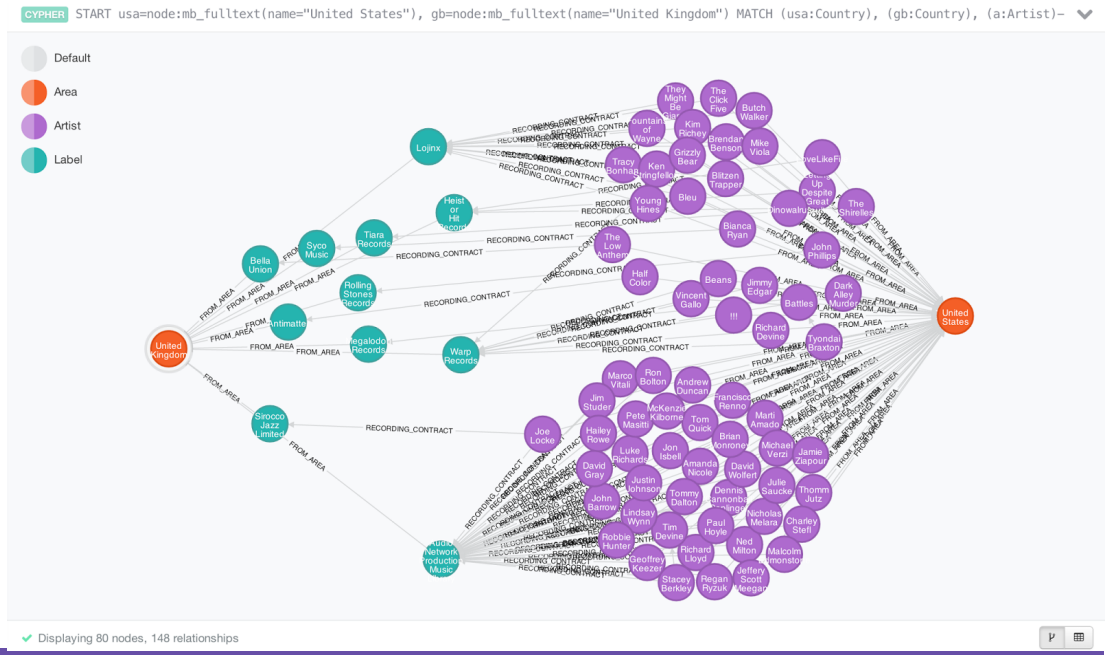
Cenários:

- Internet das Coisas - IoT
- Representação espacial e/ou temporal

Neo4j

Orientado a grafos

- baixa performance
- consistente
- utiliza cypher



Exemplo

Busca em documentos

Realiza buscas em repositório de documentos

- sistema automatizado
- alta performance
- grande variedade de formatos

Ferramentas:

- Solr
- Lucene

Cenários:

- Busca em planilhas, documentos word, JSON, XML, pdf, etc

Solr

Busca em repositórios

- necessita indexação
- índices de busca com complexidade *n log(n)*
- interface RESTful (facilita uso em ajax)
- exporta em diversos formatos

Exemplo