

# Codificação para Armazenamento de Dados



Aline Bini  
Ana Livia Franco  
Ana Priss

João Squinelato  
Marcelo Pena  
Thais Siqueira

Trabalho disponível em:  
<https://github.com/mdspena/eEDB-006-2024-2>

# Codificação de Dados

---

Mudanças constantes em aplicações computacionais

Coexistência de dados antigos e novos

Compatibilidade retroativa e futura

Manipulação em memória, em disco ou via Internet


Codificação e decodificação dos dados



# JSON

---

- Codificação textual e inteligível
- Ambiguidade em interpretar tipos dados
- Adoção de esquemas de dados
- *Strings* binárias
- Ocupa maior espaço para armazenamento
- Popularidade e consenso



```
{  
  "userName": "Martin",  
  "favoriteNumber": 1337,  
  "interests": ["daydreaming", "hacking"]  
}
```

tamanho: 81 bytes

# AVRO

- Codificação binária
- Subprojeto do Apache Hadoop
- Adequado para o contexto de Big Data
- Linguagem de definição de interface (IDL)
- Concatenação de valores hexadecimais

## Representação IDL

```
record Person {  
    string          userName;  
    union {null, long} favoriteNumber = null;  
    array<string>    interests;  
}
```

## Dados codificado

```
0c 4d 61 72 74 69 6e 02 f2 14 04 16 64  
61 79 64 72 65 61 6d 69 6e 67 0e 68 61  
63 6b 69 6e 67 00
```

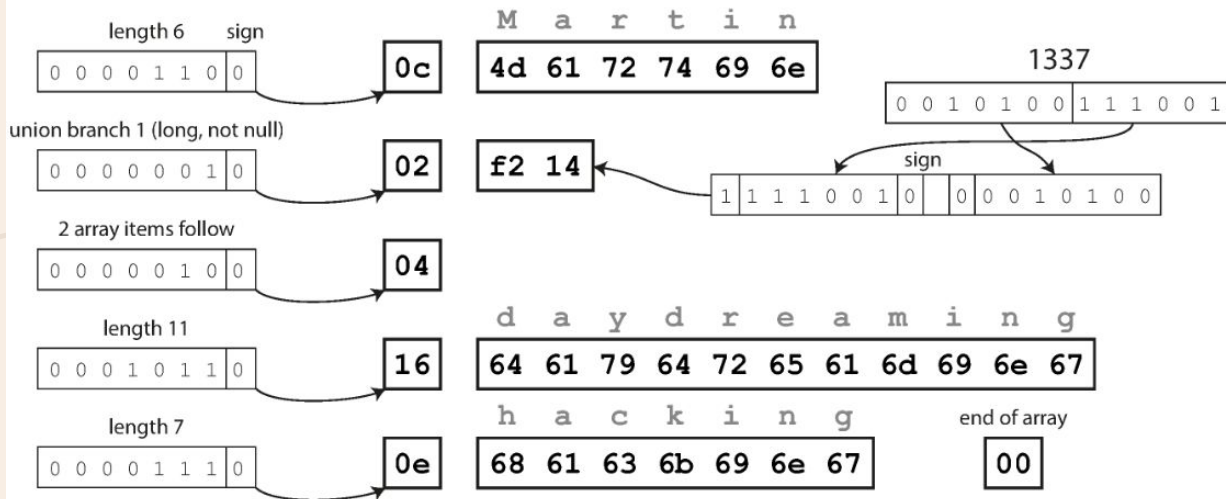
tamanho: 32 bytes

# AVRO

Byte sequence (32 bytes):

0c	4d	61	72	74	69	6e	02	f2	14	04	16	64	61	79	64	72	65	61	6d
69	6e	67	0e	68	61	63	6b	69	6e	67	00								

Breakdown:



# Esquema de Dados em AVRO

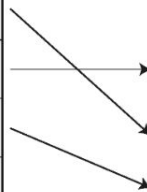
- Esquema de escrita e de leitura
- Compatibilidade entre esquemas de dados
- Documentação dos dados e flexibilidade

Writer's schema for Person record

Datatype	Field name
string	userName
union {null, long}	favoriteNumber
array<string>	interests
string	photoURL

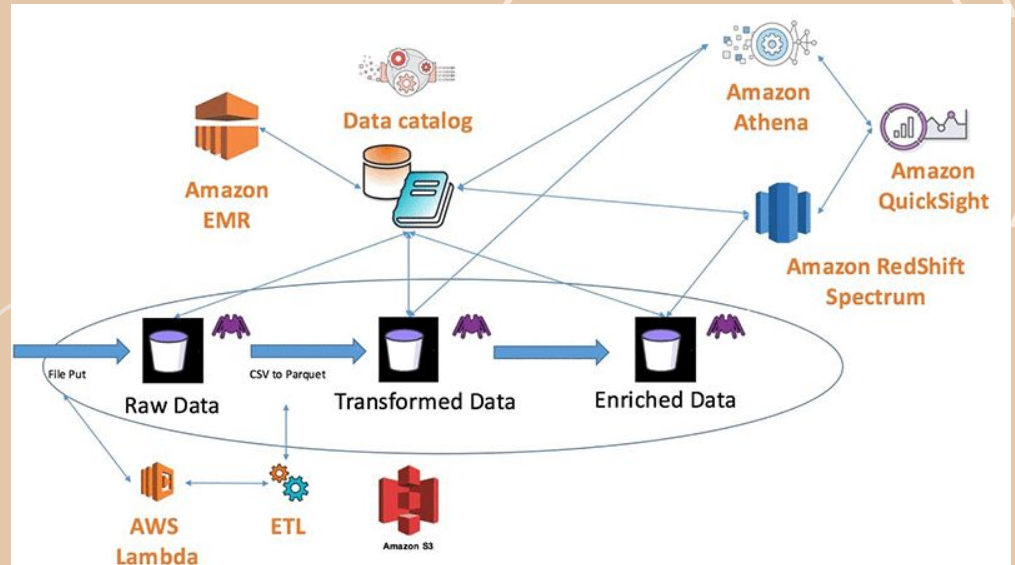
Reader's schema for Person record

Datatype	Field name
long	userID
union {null, int}	favoriteNumber
string	userName
array<string>	interests



# Testes Empíricos

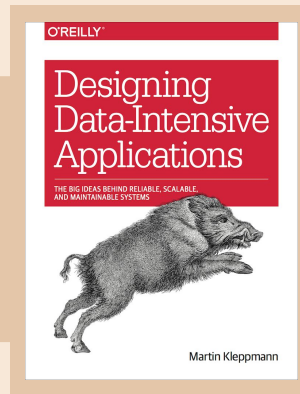
- Censo da Educação Superior
- *Data Lake*
- Diferentes codificações
- Armazenamento
- Performance



# Referências Bibliográficas

---

Designing Data-Intensive Applications, O'Reilly



Trabalho disponível em:  
<https://github.com/mdspena/eEDB-006-2024-2>

