Aquisição de Arquivos e Datasets

Lendo Hadoop

Delermando Branquinho Filho

HDF5

- Usado para armazenar grandes conjuntos de dados
- Suporta armazenar uma gama de tipos de dados
- Formato de dados hierárquico
- *groups* contendo zero ou mais conjuntos de dados e metadados
- Ter um *group header* com nome de grupo e lista de atributos
- Ter um *group symbol table* com uma lista de objetos no grupo
- *datasets* matriz multidimensional de elementos de dados com metadados
- Ter um *header* com nome, tipo de dados, espaço de dados e layout de armazenamento
- Possuir um *data array* com os dados

R HDF5 package

```
source("http://bioconductor.org/biocLite.R")
biocLite("rhdf5")
```

```
library(rhdf5)
created = h5createFile("example.h5")
created
```

```
## [1] TRUE
```

- Isto irá instalar pacotes de Bioconductor http://bioconducto.org/, usado principalmente para genômica, mas também tem bons pacotes de "grandes dados"
- Pode ser usado para interface com hdf5 conjuntos de dados.
- Esta palestra é modelada muito de perto no tutorial rhdf5 que Pode ser encontrado aqui http://www.bioconductor.org/packages/release/bioc/vignettes/rhdf5/inst/doc/rhdf5.pdf —

Cria grupos

```
created = h5createGroup("example.h5","foo")
created = h5createGroup("example.h5","baa")
created = h5createGroup("example.h5","foo/foobaa")
h5ls("example.h5")
```

```
##   group   name     otype dclass dim
## 0     /    baa H5I_GROUP
## 1     /    foo H5I_GROUP
## 2  /foo foobaa H5I_GROUP
```

Escreve em grupos

```
A = matrix(1:10,nr=5,nc=2)
h5write(A, "example.h5","foo/A")
B = array(seq(0.1,2.0,by=0.1),dim=c(5,2,2))
attr(B, "scale") <- "liter"
h5write(B, "example.h5","foo/foobaa/B")
h5ls("example.h5")
```

```
##         group   name       otype  dclass       dim
## 0           /    baa   H5I_GROUP
## 1           /    foo   H5I_GROUP
## 2        /foo      A H5I_DATASET INTEGER     5 x 2
## 3        /foo foobaa   H5I_GROUP
## 4 /foo/foobaa      B H5I_DATASET   FLOAT 5 x 2 x 2
```

Grava um dataset

```
df = data.frame(1L:5L,seq(0,1,length.out=5),
  c("ab","cde","fghi","a","s"), stringsAsFactors=FALSE)
h5write(df, "example.h5","df")
h5ls("example.h5")
```

```
##         group   name       otype   dclass       dim
## 0           /    baa   H5I_GROUP
## 1           /     df H5I_DATASET COMPOUND         5
## 2           /    foo   H5I_GROUP
## 3        /foo      A H5I_DATASET  INTEGER     5 x 2
## 4        /foo foobaa   H5I_GROUP
## 5 /foo/foobaa      B H5I_DATASET    FLOAT 5 x 2 x 2
```

Lendo dados

```
readA = h5read("example.h5","foo/A")
readB = h5read("example.h5","foo/foobaa/B")
readdf= h5read("example.h5","df")
readA
```

```
##      [,1] [,2]
## [1,]    1    6
## [2,]    2    7
## [3,]    3    8
## [4,]    4    9
## [5,]    5   10
```

Escrever e ler pedaços

```
h5write(c(12,13,14),"example.h5","foo/A",index=list(1:3,1))
h5read("example.h5","foo/A")
```

```
##      [,1] [,2]
## [1,]   12    6
## [2,]   13    7
## [3,]   14    8
## [4,]    4    9
## [5,]    5   10
```

Notas e outros recursos

- Hdf5 pode ser usado para otimizar a leitura / gravação de disco em R
- O tutorial rhdf5: * Http://www.bioconductor.org/packages/release/bioc/vignettes/rhdf5/inst/doc/rhdf5.pdf
- O grupo HDF tem informação sobre HDF5 em geral http://www.hdfgroup.org/HDF5/