Big Data

Hadoop

Prof. Jean Paul Barddal

Agenda

Agenda

Sobre

- Jean Paul Barddal
- Professor do PPGIa/PUCPR
- `www.jpbarddal.com.br`
- Áreas de Pesquisa
  - Data stream mining
  - Machine Learning aplicado
- jean.barddal@ppgia.pucpr.br

Datasets

- Durante as aulas, usaremos diferentes datasets
- Datasets e códigos estarão disponíveis em: `www.jpbarddal.com.br`

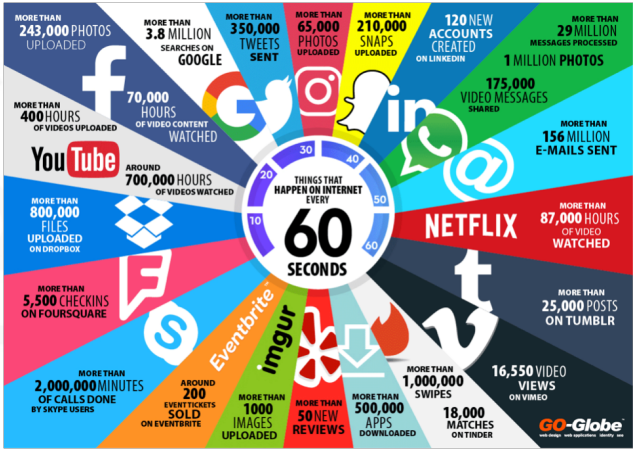
O que nós faremos durante o curso?

- Discutiremos o que é Big Data
- Entenderemos o Hadoop Distributed File System (HDFS)
- Processaremos dados usando Spark (RDDs, Pair RDDs, SQL e Pandas)

Big Data

- Coleção de dados tão grande e complexa que não conseguimos tratar usando bancos de dados e ferramentas de processamento tradicionais (monolíticas)

Motivação


MORE THAN 243,000 PHOTOS UPLOADED
MORE THAN 3.8 MILLION SEARCHES ON GOOGLE
MORE THAN 350,000 TWEETS SENT
MORE THAN 65,000 PHOTOS UPLOADED
MORE THAN 210,000 SNAPS UPLOADED
120 NEW ACCOUNTS CREATED ON LINKEDIN
MORE THAN 29 MILLION MESSAGES PROCESSED
1 MILLION PHOTOS
175,000 VIDEO MESSAGES SHARED
MORE THAN 156 MILLION E-MAILS SENT
70,000 HOURS OF VIDEO CONTENT WATCHED
MORE THAN 400 HOURS OF VIDEOS UPLOADED
YouTube
AROUND 700,000 HOURS OF VIDEOS WATCHED
THINGS THAT HAPPEN ON INTERNET EVERY 60 SECONDS
MORE THAN 800,000 FILES UPLOADED ON DROPBOX
NETFLIX
MORE THAN 87,000 HOURS OF VIDEO WATCHED
MORE THAN 5,500 CHECKINS ON FOURSQUARE
MORE THAN 25,000 POSTS ON TUMBLR
Eventbrite
imgur
MORE THAN 2,000,000 MINUTES OF CALLS DONE BY SKYPE USERS
AROUND 200 EVENT TICKETS SOLD ON EVENTBRITE
MORE THAN 1000 IMAGES UPLOADED
MORE THAN 50 NEW REVIEWS
MORE THAN 500,000 APPS DOWNLOADED
MORE THAN 1,000,000 SWIPES
18,000 MATCHES ON TINDER
16,550 VIDEO VIEWS ON VIMEO
GO-Globe

O que faz de Big Data realmente Big Data?

Definition de Dan Ariely, Duke University

"Big Data is like teenage sex; everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."

7 Vs de Big Data (Características)

- Volume
- Velocidade
- Variedade
- Variabilidade
- Veracidade
- Visualização
- Valor

3, 5, 7, 10, ou 42 Vs?

- 3 Vs: https://www.flydata.com/blog/3-vs-of-big-data/
- 5 Vs: https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/
- 7 Vs: https://impact.com/marketing-intelligence/7-vs-big-data/
- 10 Vs: https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx
- 42 Vs: https://www.elderresearch.com/blog/42-v-of-big-data

Volume

- A característica mais óbvia
- A quantidade astronômica de dados gerada, armazenada e processada
- Examples:
  - 300 horas de video são enviadas ao Youtube a cada minuto
  - 1.1 trilhões de fotos tiradas em 2016 e 9% a mais em 2017
  - Visa trata aprox. 24 mil transações de cartão de crédito por segundo nos EUA

Velocidade

- Frequência com que dados se tornam disponíveis e são processados
- Examplos:
  - Facebook diz armazenar aproximadamente 300PB de dados, mas isso aumenta em 600TB a cada dia
  - Google tem aproximadamente 40k queries por segundo, resultando em 3.5B de queries por dia
- Dados sequencialmente gerados: data streams
- Como processamos data streams?
  - Periodicamente? (Batch)
  - Em tempo real (ou quase real)?

Variedade

- Dado de acordo com os tipos de dados e respectivas origens
- Dados estruturados, semi-estruturados e não estruturados
- Como combinar tipos de dados?

Variabilidade

- Variabilidade não é o mesmo que Variedade
- Se uma cafeteria vende 6 tipos de café: variedade
- Se você pede o mesmo café todos os dias, mas ele sempre tem sabor diferente: variabilidade
- Como discernir entre um outlier e uma mudança de conceito?

Veracidade

- Uma das características mais infelizes de Big Data
- Mais dados, menos confiáveis eles podem ser
- Razão: os dados podem ser obtidos de diferentes fontes e nem todas as fontes são confiáveis

Veracidade

- Exemplo: ao obter estatísticas sobre dados, pense sobre:
  - Quem criou a amostra de dados
  - Qual foi a metodologia de aquisição?
  - Os dados foram sumarizados?
  - Os dados foram editados, verificados, ou corrigidor?
- Importante: dados são obtidos ou processados por pessoas
  - E se as pessoas estão mentindo?
  - E se alguém processou ou modificou os dados de forma errada?

Visualização

- Como visualizamos Big Data?
- Dificuldades:
  - Muitos dados (linhas e colunas), não cabe em RAM
  - Como plotar milhões/bilhões de pontos em um gráfico?
- Técnicas específicas: clustering, tree maps, sunbursts, parallel coordinates, cone trees, etc
- Como visualizamos tipos de dados diferentes que se tornam disponíveis ao longo do tempo?

Valor

- As demais características são irrelevantes caso não tragam valor
- Financeiro ou social
  - Entender clientes, segmentação: marketing direcionado
  - Otimização de processos
  - Identificação de fraudes, etc
  - Foco analítico/preditivo (Data Science/Machine Learning)

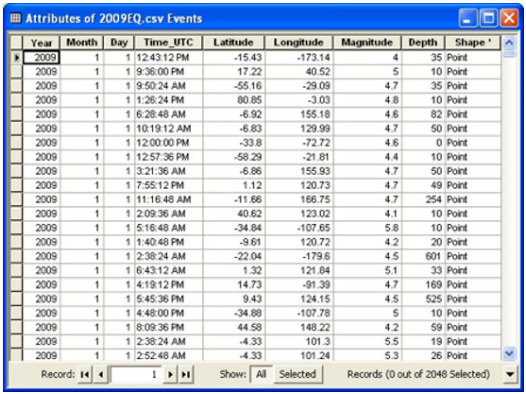
Agenda

Dados e Big Data

- Tudo é armazenado: logs, textos, imagens, vídeos, áudios, etc
- Data são categorizados:
  - Estruturados
  - Semi-estruturados
  - Não-estruturados

Dados estruturados

- Dados tabulares com schema
- Linhas, colunas e tipos de colunas
- Linhas são chamadas de exemplos, instâncias, amostras, registros
- Colunas são chamadas de atributos, características, variáveis



Dados semi-estruturados

- Dados não estritamente tabulares
- Demandam pré-processamento para conversão em tabela
- Exemplos: XML, JSON, etc
- Estrutura irregular e implícita
- Estrutura definida de acordo com seu conteúdo, e.g., tags XML definidas de acordo com os atributos da aplicação

XML

```
<menu id="file" value="File">
  <popup>
    <menuitem value="New" onclick="CreateNewDoc()" />
    <menuitem value="Open" onclick="OpenDoc()" />
    <menuitem value="Close" onclick="CloseDoc()" />
  </popup>
</menu>
```

JSON

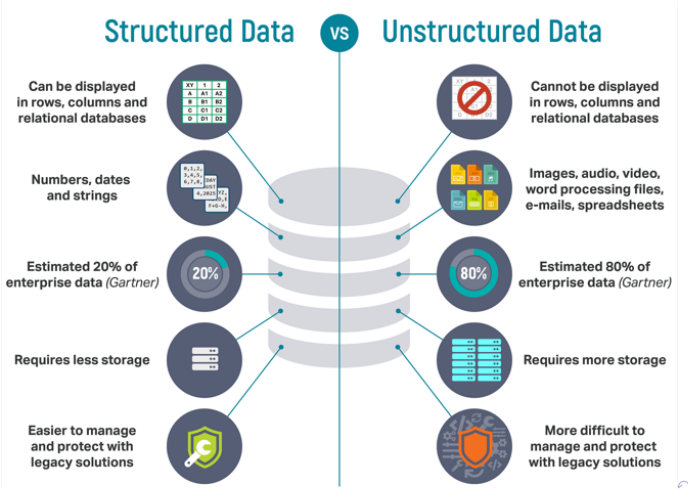
```
{
"menu":{
  "id":"file",
  "value":"File",
  "popup":{
    "menuitem":[{"value":"New","onclick":"CreateNewDoc()"},
        {"value":"Open","onclick":"OpenDoc()"},
        {"value":"Close","onclick":"CloseDoc()"}]
        }
  }
}
```

Log

```
tecmint@TecMint ~ $ tailf /var/log/apache2/access.log
127.0.0.1 - - [31/Oct/2017:11:11:37 +0530] "GET / HTTP/1.1" 200 729 "-" "Mozill
127.0.0.1 - - [31/Oct/2017:11:11:37 +0530] "GET /icons/blank.gif HTTP/1.1" 200
fox/56.0"
127.0.0.1 - - [31/Oct/2017:11:11:37 +0530] "GET /icons/folder.gif HTTP/1.1" 200
efox/56.0"
127.0.0.1 - - [31/Oct/2017:11:11:37 +0530] "GET /icons/text.gif HTTP/1.1" 200 5
ox/56.0"
127.0.0.1 - - [31/Oct/2017:11:11:38 +0530] "GET /favicon.ico HTTP/1.1" 404 500
127.0.0.1 - - [31/Oct/2017:11:12:05 +0530] "GET /tecmint/ HTTP/1.1" 200 787 "ht
0"
127.0.0.1 - - [31/Oct/2017:11:12:05 +0530] "GET /icons/back.gif HTTP/1.1" 200 4
01 Firefox/56.0"
127.0.0.1 - - [31/Oct/2017:11:13:58 +0530] "GET /tecmint/Videos/ HTTP/1.1" 200
101 Firefox/56.0"
127.0.0.1 - - [31/Oct/2017:11:13:58 +0530] "GET /icons/compressed.gif HTTP/1.1"
) Gecko/20100101 Firefox/56.0"
127.0.0.1 - - [31/Oct/2017:11:13:58 +0530] "GET /icons/movie.gif HTTP/1.1" 200
0/20100101 Firefox/56.0"
```

Dados não estruturados

- Sem estrutura pré-definida
- Imagem
- Áudio
- Vídeo
- Texto



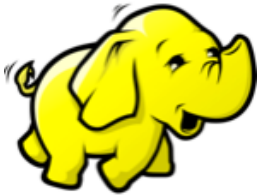
Modos de Processamento

- Batch
  - Dados armazenados no HDFS
  - Periodicamente (ou sob demanda), dados são processados
  - Nosso foco: Spark
- Stream (quase-, tempo real)
  - Dados processados conforme a chegada
  - Opcionalmente, dados são armazenados
  - Ferramentas: Storm e Spark streaming

Agenda

Hadoop

- Big Data se tornou um tópico de interesse por conta do Hadoop
- Objetiva armazenamento e processamento de dados
- Clusters que vão de um único computador a milhares de computadores
- Computadores heterogêneos: hardware e SO
- Detecção de falhas
- Redundância



Composição do Hadoop

- Hadoop Distributed File System (HDFS)
- MapReduce

Características

- Inspirado no modelo de localidade
- Processamento vai aos dados
- Uso reduzido de rede
- Escalabilidade
- Mesmo software roda em 1, 10, 100 ou milhares de computadores
- Performance elevada

Apache Spark

- Framework para processamento de dados em memória
- Altamente escalável
- Permite processamento em lote (batch) e stream
- Otimizado para SQL e Machine Learning
- Programação e otimização baseada em DAGs (directed acyclic graphs)
- Flexível: podemos usar em Java, Scala, Python e R

Teste de Ambiente

- Vamos realizar um teste de acesso ao ambiente e verificar as principais partes disponíveis para uso
- HDFS
- Spark
- Jupyter Notebook

Formulário

- Caso você tenha problemas para acessar o ambiente, favor preencher o formulário abaixo com:
- Nome completo (sem abreviações)
- Email institucional
- Link: `https://forms.office.com/r/UKQv3CPfMh`

Agenda

Como processar massivas bases de dados?

- Desafios:
  - Processamento é lento dadas as quantidades de dados
  - Aquisição de dados a partir de HDs é lenta
  - Quantidade limitada de RAM
  - Idéia-chave: armazenamento e processamento distribuído de dados

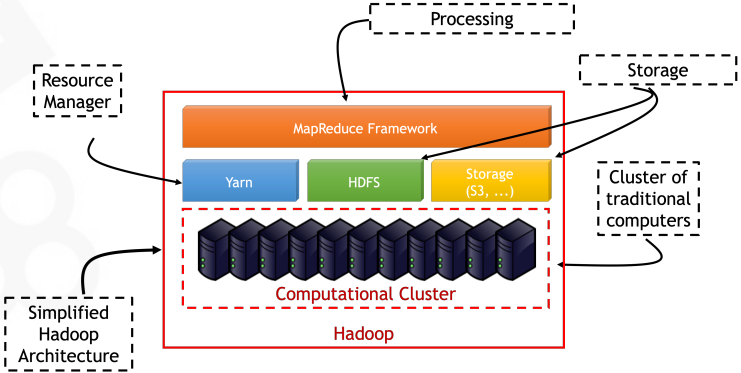
Hadoop

- Hadoop traz uma solução de computação distribuída para:
  - Concorrência
  - Portabilidade
  - Confiabilidade
  - Escalabilidade
  - Heterogeneidade
  - Transparência
  - Tolerância a falhas

Princípios

- O sistema deve ser auto-gerenciável e se recuperar de desastres
- Falhas podem acontecer mas devem ser transparentes para o usuário
- Balanceamento de carga: quando possível, as tarefas devem ser alocadas de acordo com a performance dos computadores
- Performance deve escalar de acordo com a quantidade de computadores disponíveis
- Processamento deve ir até os dados
- Pouco consumo de banda de rede

Visão Geral


Processing
Resource Manager
Storage
MapReduce Framework
Yarn
HDFS
Storage (S3, ...)
Cluster of traditional computers
Computational Cluster
Hadoop
Simplified Hadoop Architecture

Curiosidade

- Criado por Doug Cutting
- “Hadoop” veio do nome do brinquedo do filho