Aula 6

Big Data

Prof. Luis Henrique Alves Lourenço

1

Conversa Inicial

2

Temas complementares

- Componentes Hadoop
- *Data Lake*
- Sistemas de recomendação
- Computação em nuvem
- Design de arquitetura Hadoop

3

Outras tecnologias

4

Impala

- Banco de dados analítico massivamente paralelo
- Fundação Apache
- Consultas SQL + HDFS/HBase
- *Bussiness Inteligence*

5

Accumulo

- Baseado no BigTable (chave-valor)
- Fundação Apache
- Segurança em nível de célula

6

Redis

- *In-memory storage*
- Banco de dados
- Cache
- Agente de mensagens

7

Ignite

- Computação em memória escalável
- Fundação Apache
- Cache + *Storage*
- Propriedades ACID
- Consultas SQL

8

NiFi

- Processamento e distribuição de dados por meio de grafos dirigidos
- Fundação Apache
- Automatizar e gerenciar fluxos
- Altamente configurável

9

Ambari

- Interface Web
- Fundação Apache
- *Core Hadoop*
- Essencial *Hadoop*
- *Hadoop Support*

10

Data Lake

11

- Estratégia de gerenciamento de dados
- Identificação, limpeza e integração dos dados

12

Gerenciamento de dados

- Dados estão espalhados
- Reorganização e reformatação
- Atualização

13

Outras estratégias

- *Data Warehouse*
- *Data Mart*

14

Data Lake

- "Se você pensar em um *data mart* como uma loja de garrafas de água (água limpa, embalada e estruturada para ser consumida facilmente), o *data lake* seria um corpo de água maior em estado natural. O conteúdo do *data lake* flui a partir de uma fonte que preenche o lago e vários usuários do lago podem vir e examinar, mergulhar ou colher amostras" (James Dixon, 2010)

15

Data Swamp

- Degradação de dados
- Gerenciamento
- Acessibilidade
- Governança

16

Níveis de maturidade

- *Data Puddle*
- *Data Pond*
- *Data Lake*
- *Data Ocean*

17

Sistemas de recomendação

18

- Prever a reação dos usuários em relação a um conjunto de opções

19

A cauda longa

20

Tipos de recomendadores

- Editorial ou curadoria
- Agregação simples
- Recomendadores individualizados

21

Recomendadores individualizados

- *Content-based recommendation systems*
- *Collaborative filtering*

22

Modelo formal

- C: conjunto de consumidores
- S: conjunto de itens
- U: CxS → R
- Matriz Utilidade ->

	i_0	i_1	i_2	i_3
c_0	5		2	
c_1		4	3	
c_2	2	4		
c_3	4			5

23

Recomendadores baseados em conteúdo

- Recomendar ao usuário itens parecidos com os que ele avaliou
- Perfil de item
- Perfil de usuário
- Predições

24

Filtragem colaborativa

- Usuários com avaliações semelhantes podem estimar avaliações
- Métrica de similaridade: *Pearson Correlation*
- Predição de avaliações

$$r_{xi} = \frac{\sum_{y \in N} sim(x,y) * r_{yi}}{\sum_{y \in N} sim(x,y)}$$

25

Pearson Correlation

	i_0	i_1	i_2	I_3	i_4	i_5	i_6
u_0	4			5	1		
u_1	5	5	4				
u_2				2	4	5	
u_3			3				3

⇨

	i_0	i_1	i_2	I_3	I_4	i_5	i_6
u_0	2 / 3			5 / 3	-7 / 3		
u_1	1 / 3	1 / 3	-2 / 3				
u_2				-5 / 3	1 / 3	4 / 3	
u_3			0				0

- sim(u0,u1)= cos(ru0,ru1)=0,09
- sim(u0,u2)= cos(ru0,ru2)=-0,56
- sim(u0,u1) > sim(u0,u2)

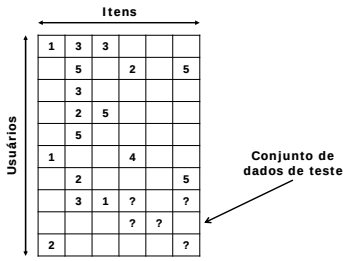
26

Métodos híbridos

- Utilizar métodos de um modelo em outro
- Combinar predições de dois modelos

27

Avaliação de sistemas de recomendação



28

Cloud Computing

29

Recapitulando

- ***Big Data*: captura, armazenamento e análise**
- ***Hadoop clusters***
- **Dados não estruturados**
- **Evolução tecnológica**

30

Cloud Computing

- **Acesso sob demanda**
- **Recursos computacionais**
- **Escaláveis e elásticos**
- **Via internet**
- **Privada ou pública**

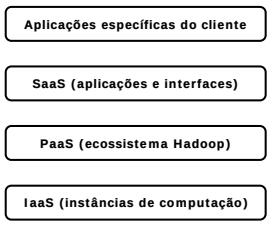
31

Modelos de *Cloud Computing*

- ***Infrastructure as a Service***
- ***Platform as a Service***
- ***Software as a Service***

32

Big Data as a Service



33

Amazon Elastic MapReduce (EMR)

- ***Amazon Web Services***
- ***IaaS + PaaS***
- ***Spark*, *Hive*, *HBase*, *Flink*, *Presto* etc.**
- ***Amazon Elastic Compute Cloud* (EC2)**
- ***Amazon Simple Storage Service* (S3)**

34

Google Cloud Dataproc

- ***Google Cloud PLatform***
- **IaaS + PaaS**
- ***Hive*, *HBase*, *Zeppelin*, *Zookeeper*, *Presto*, *Pig* etc.**
- ***BigQuery, Cloud Storage, Cloud BigTable***
- ***Stackdriver Logging, Stackdriver Monitoring***

35

Microsoft Azure

- **IaaS + PaaS**
- ***Azure Virtual Machines***
- ***Azure Container Services***
- ***Azure Blob Storage***
- ***Azure Batch***
- ***Azure Functions***
- ***Azure HDInsight: Spark, Hadoop***

36

Design de arquitetura *Big Data*

37

Arquitetura *Big Data*

- Combinar componentes do ecossistema *Hadoop*
- Novas possibilidades de negócios
- Novos desafios

38

Obtendo informações

- Quem serão os usuários?
- Quais são os problemas que a aplicação deve resolver?
- Quais os benefícios mais importantes para os usuários?
- Como garantir que sabemos o que os usuários querem e precisam?
- Como é a experiência para os usuários?

39

Entendendo requisitos

- *Amazon Working Backwards*
- Documentação
- Requisitos
- Produto

40

Requisitos de armazenamento

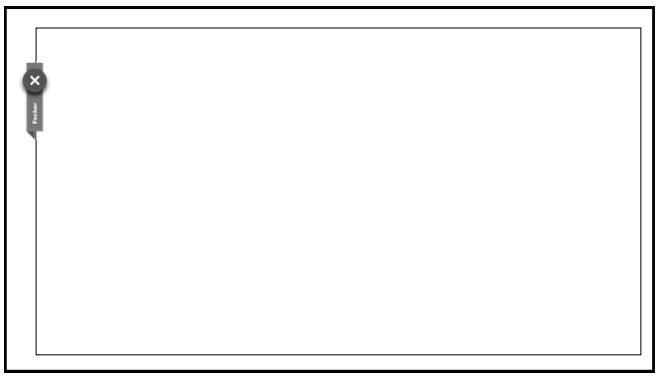
- Tamanho dos dados
- Mover dados pode ser caro
- Ingestão de dados
- Até quando devem ser mantidos?
- Quando devem ser eliminados?

41

Requisitos de dados

- Medidas de segurança
- Teorema CAP
- Padrões de acesso
- Atualização de dados

42



43