BIG DATA

LABDATA

Coordenação:

Prof. Dr. Adolpho Walter Pimazzi Canton

Profa. Dra. Alessandra de Ávila Montini

Disciplina: Aplicações de Big Data com Hadoop

Tema da Aula: Aplicações MapReduce

Profa. Rosangela de Fátima Pereira

Junho de 2016

Currículo

Formação

- Mestrado em Engenharia de Computação pela Escola Politécnica da Universidade de São Paulo (Poli-USP) (em andamento)
- Especialização em Tecnologia Java pela Universidade Tecnológica Federal do Paraná (UTFPR) (2011)
- Tecnologia em Análise e Desenvolvimento de Sistemas pela UTFPR (2011)
- Bacharelado em Administração de Empresas pela Universidade Estadual do Norte do Paraná (UENP) (2007)

Experiência

- Professora de Big Data Analytics em empresas e programas de MBA - FIA (2013 - atual)
- Pesquisadora no Laboratório de Arquitetura e Redes de Computadores (LARC) – USP (2013 - atual)
- Professora de cursos de engenharia na UTFPR (2011 -2012)
- Analista de sistemas na BSI Tecnologia (2009-2010)

LinkedIn: https://br.linkedin.com/pub/rosangela-de-fatima-pereira/68/a10/b56

Apaixonada por **Big Data**!

Objetivo da Aula

Revisar o conteúdo sobre Hadoop

e

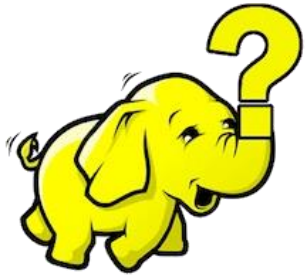
Implementar aplicações MapReduce para compreensão da biblioteca Java do Hadoop

Conteúdo da Aula

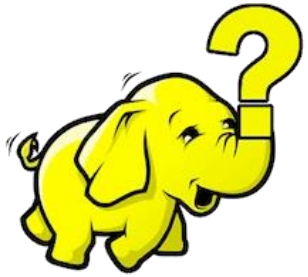
- Execução de aplicação MapReduce
- Revisão sobre Hadoop

Conteúdo da Aula

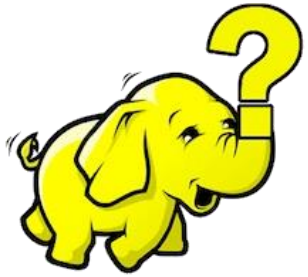
- Execução de aplicação MapReduce
- Revisão sobre Hadoop



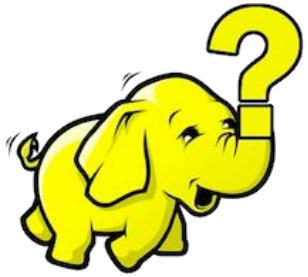
1. Qual a diferença entre escalabilidade vertical e escalabilidade horizontal?



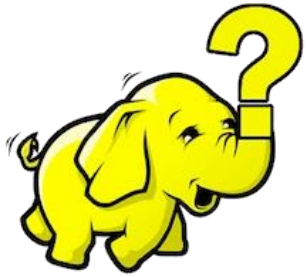
2. Qual a diferença entre dados estruturados, semi-estruturados e dados não estruturados?



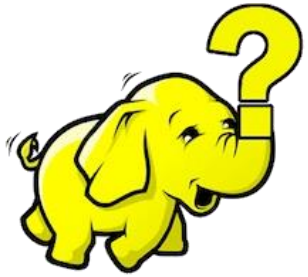
3. Quais são os tipos de analytics existentes?



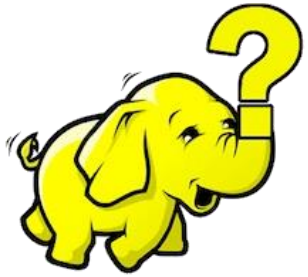
4. Qual a diferença entre a análise descritiva e análise preditiva?



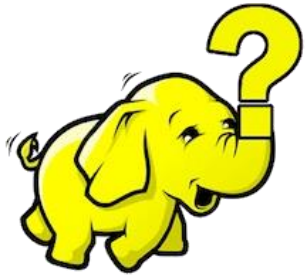
5. O que é NoSQL e quais são as 4 categorias existentes nesse conceito?



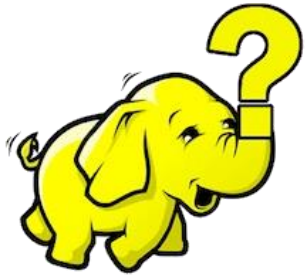
6. Quais são os componentes core do Hadoop?



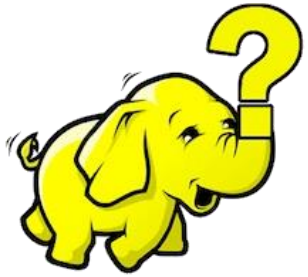
7. O que é Hadoop streaming?



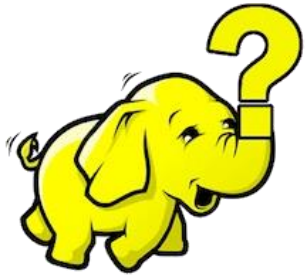
8. Descreva 3 características do framework Hadoop



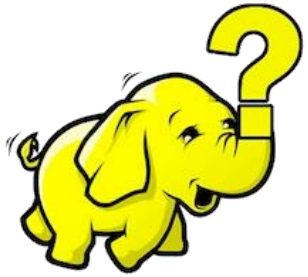
9. Dê exemplo de 3 situações em que o Hadoop pode ser benéfico para uma empresa.



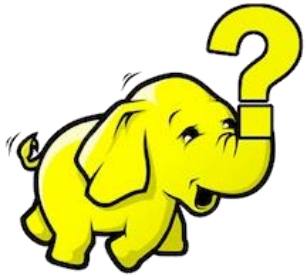
10. Explique brevemente a história do Hadoop.



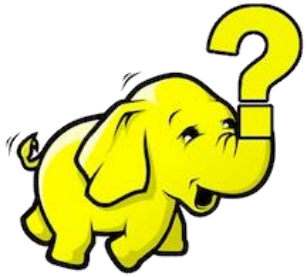
11. Descreva o conceito de data lake.



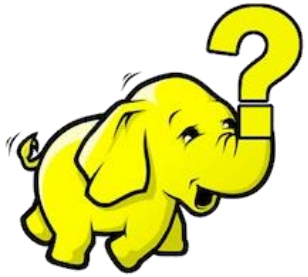
12. Cite 5 frameworks do ecossistema Hadoop.



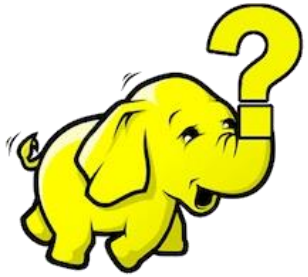
13. O que é hardware commodity? Por que Hadoop permite a utilização desse hardware?



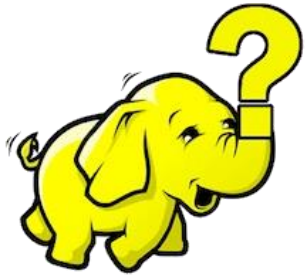
14. O que é uma distribuição Hadoop? Dê exemplo de 2 distribuições



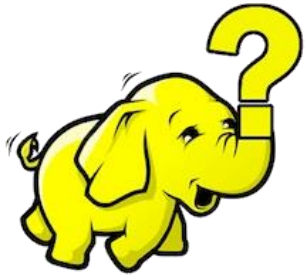
15. Explique as diferenças entre um SGBDR e do Hadoop



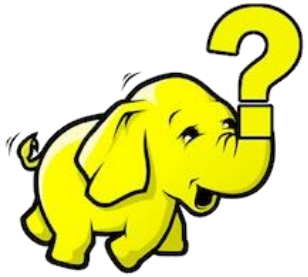
16. O que faz o comando jps?



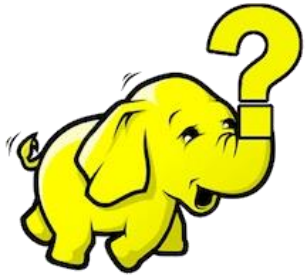
17. Em qual nó é recomendável ter a melhor configuração de hardware em um cluster Hadoop? Por qual motivo?



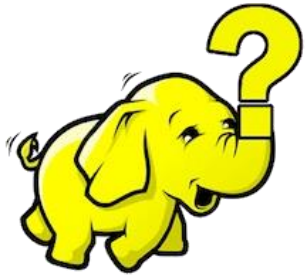
18. Quais são os processos mestres do Hadoop e que operações eles realizam?



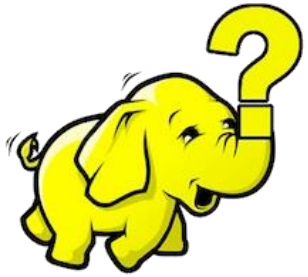
19. Quais são os processos escravos do Hadoop e que operações eles realizam?



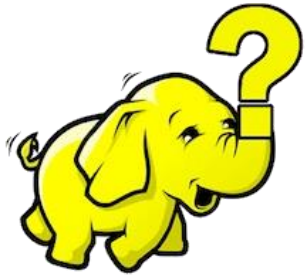
20. O SecondaryNameNode é um substituto do NameNode? Como ele funciona?



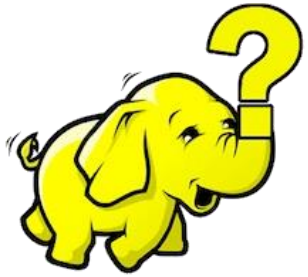
21. Hadoop pode ser executado em 3 diferentes modos, quais são eles?



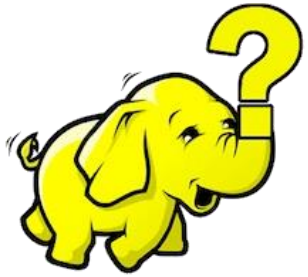
22. Como o HDFS oferece tolerância a falhas?



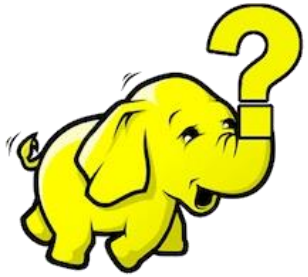
23. Se um arquivo tem 50 MB de dados, mesmo assim o bloco ocupará o tamanho padrão de 64MB ou 1 28MB?



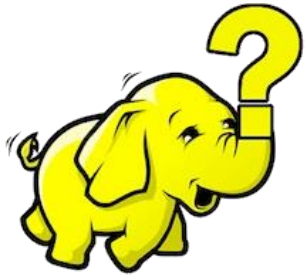
24. O que é o InputFormat no Hadoop?



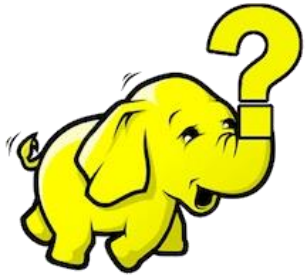
25. Quais são os 3 InputsFormats mais comuns do HDFS?



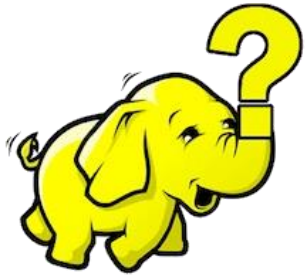
26. Como funciona o formato de entrada TextInputFormat?



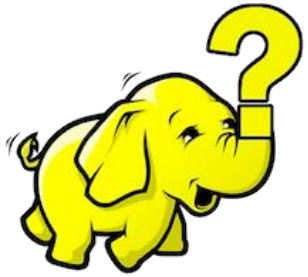
27. Quais são os 4 módulos oferecidos pela biblioteca Java do Hadoop?



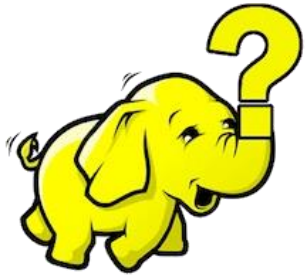
28. Como o Hadoop permite diminuir a quantidade de dados trafegados na rede?



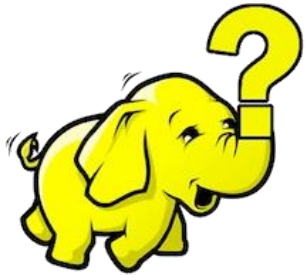
29. Qual é a estratégia de alocação que o HDFS utiliza para armazenar as réplicas dos blocos?



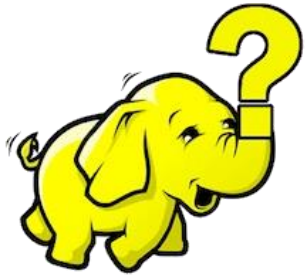
30. O que é o MapReduce?



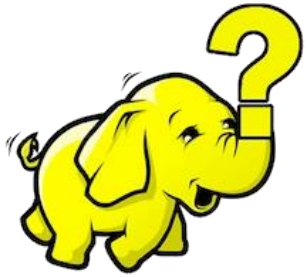
31. Qual o objetivo da fase map e da fase reduce?



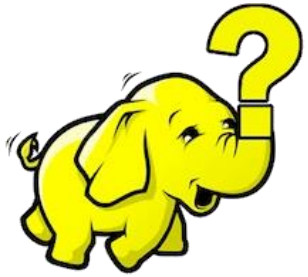
32. Quais são os 4 parâmetros da tarefa Map?



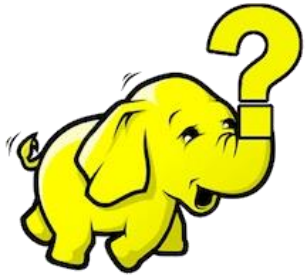
33. O que acontece internamente no Hadoop quando um cliente submete um job ao cluster?



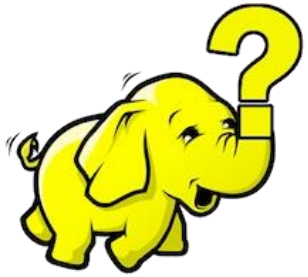
34. Quais são as operações ocorridas na fase Reduce?



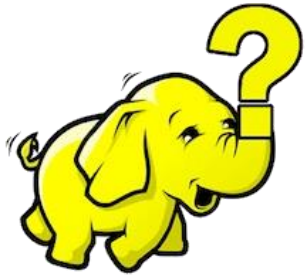
35. Por que não podemos utilizar os tipos de dados primitivos de Java nas tarefas Map e Reduce?



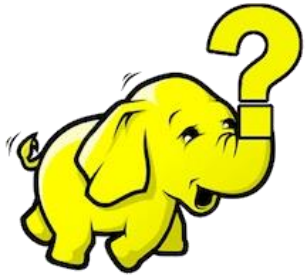
36. Para quais tipos de aplicações o MapReduce não é indicado?



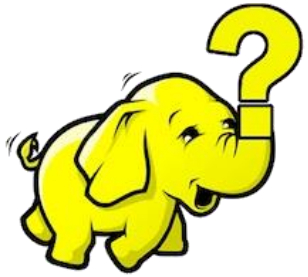
37. O que é o Apache Hadoop YARN?



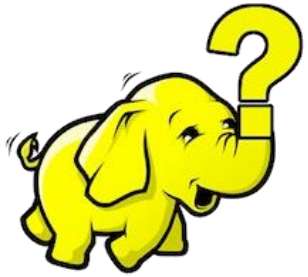
38. Quais foram os benefícios que o YARN trouxe ao Hadoop?



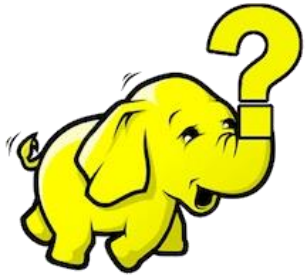
39. Como o MapReduce é utilizado na versão 2 do Hadoop?



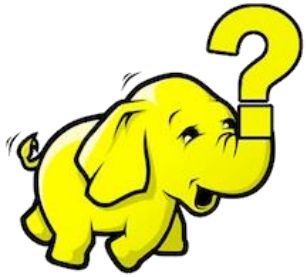
40. Qual o benefício do MapReduce em relação a outras linguagens de programação paralela?



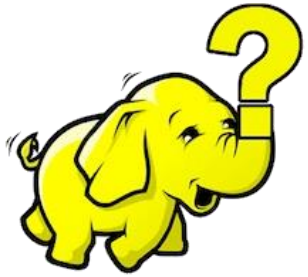
41. Explique porque Java é considerado uma linguagem portável.



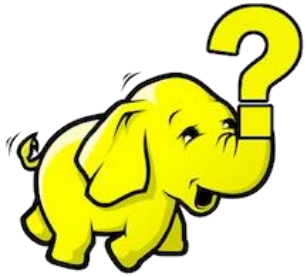
42. Quais são os componentes do YARN?



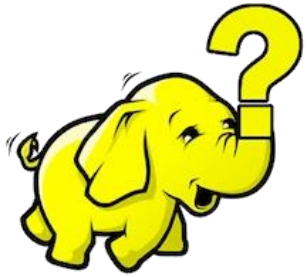
43. Explique o que é e o que faz o application master do YARN.



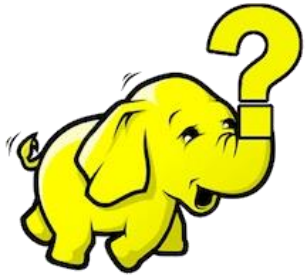
44. O que deve ser implemento na classe Driver do MapReduce? Dê exemplos de 2 comandos.



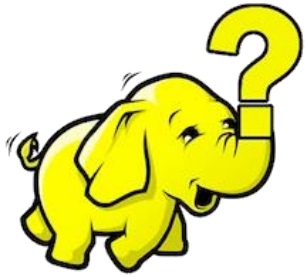
45. Para que serve o Sqoop?



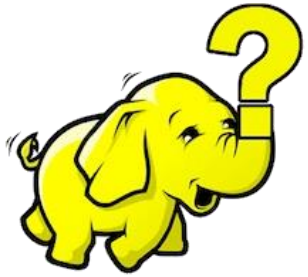
46. O que é o Hive metastore?



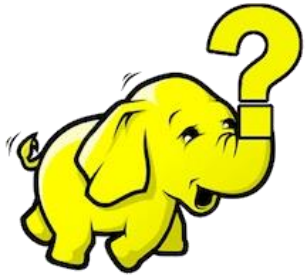
47. Qual framework do ecossistema Hadoop oferece uma biblioteca de aprendizado de máquina?



48. Qual framework do ecossistema Hadoop oferece mecanismos para a coleta e transferência de registros de log?

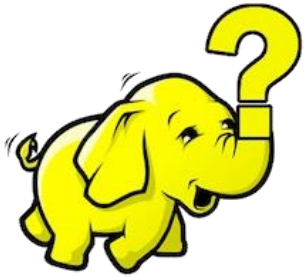


49. Em quais linguagens posso implementar uma aplicação MapReduce?



50. Qual a limitação de implementar uma aplicação MapReduce em uma linguagem diferente de Java?

Perguntas



rpereira@larc.usp.br

Referências Bibliográficas

WHITE, Tom. **Hadoop: The definitive guide**. " O'Reilly Media, Inc.", 2012.

PERERA, Srinath. **Hadoop MapReduce Cookbook**. Packt Publishing Ltd, 2013.