PERFORMANCE EVALUATION OF A MONGODB AND HADOOP PLATFORM FOR SCIENTIFIC DATA ANALYSIS

M.Govindaraju and L. Ramakrishnan

Alfredo Fernandes / José Fernandes

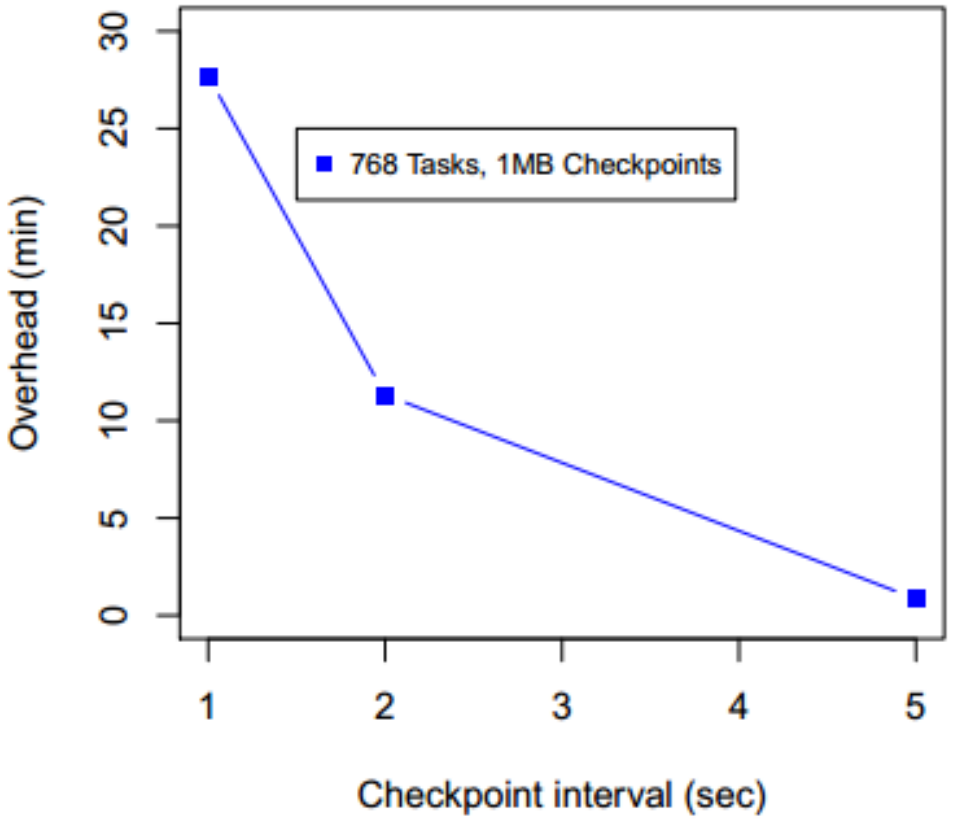
6. Avaliação

- Nesta secção é apresentada as avaliações de performance para quantificar as diferenças mongo-hadoop
- As experiencias foram realizadas 3 vezes
- Os resultados aqui apresentados correspondem à média dos valores obtidos
- Checkpoints de cada segundo até intervalos de 5 segundos

6) Avaliação - MongoDB

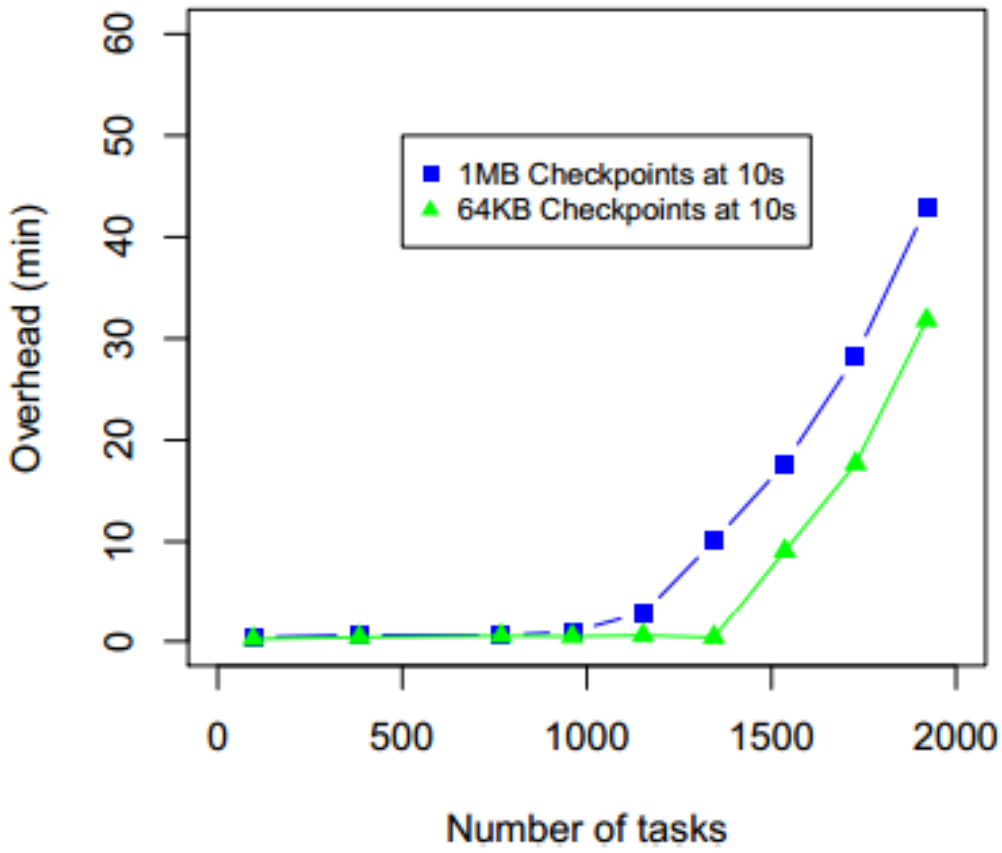
- 1° Teste consiste em analisar capacidade de respostas de MongoDB
- Utilizaram sistema centralizado em que cada node (worker) faz checkpoint periodicamente
- Checkpoint – cada node informa o servidor do seu "status" espaçado por um intervalo de tempo pré-definido (ex: cada 2s)
- Testa a capacidade de MongoDB dar resposta a milhares de coneções ao mesmo tempo

6.1) Avaliação - MongoDB



- 768 tarefas distribuídos por 192 cores (4 tarefas por core)
- Checkpoint de 1s à 5s
- Aumenta-se o Checkpoint o valor do Overhead diminui cecar de 250%
- 1MB provoca tráfego de 768Mb/s para Checkpoints de 1s

6) Avaliação - MongoDB



- Checkpoint em intervalo fixos de 10s e tamanhos diferentes: 1MB e 64KB
- Overhead aumenta com o aumenta do número de tarefas
- O tamanho não influencia o Overhead até 1000 tarefas
- N° de conneções tem mais influencia no Overload do que o tamanho dos dados
- Justificação: MongoDB dispara um thread por cada conecção com um stack próprio.

6.2 Avaliação – HDFS vs MongoDB

- Hadoop Distributed File System (HDFS)
- MongoDB (NoSQL)
- Comparação de operações de READ/WRITE
- Foi desenvolvido um programa em Java + Python faz READ de 37M de records e WRITE 19M records de e para MongoDB e HDFS
- Objectivo: é testar essas duas operações em apenas um node

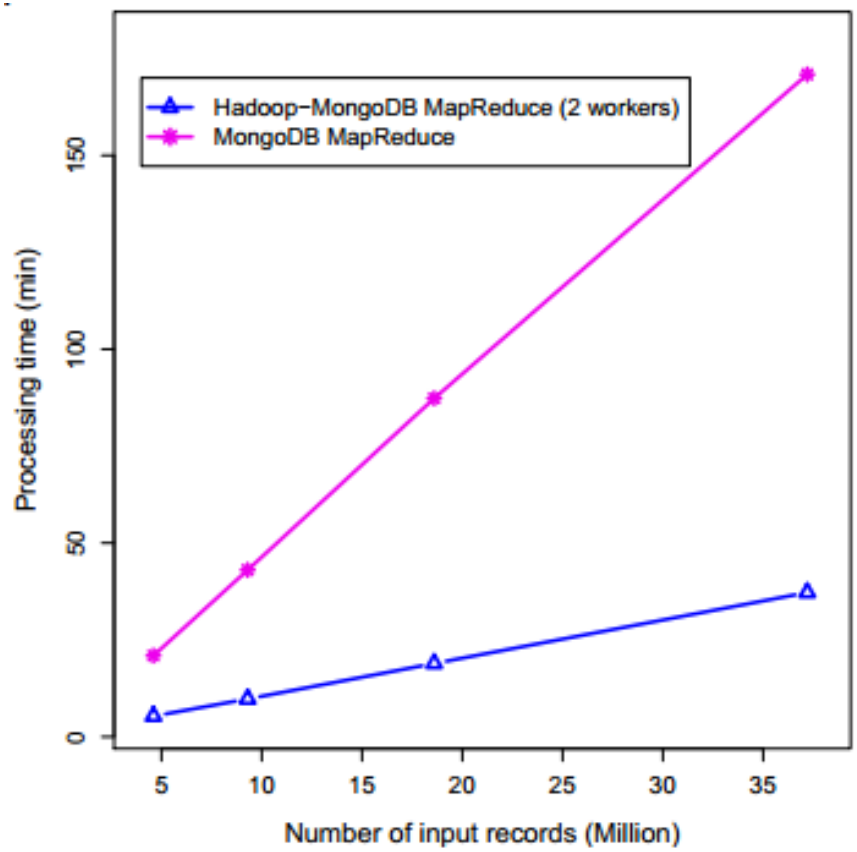
6.2 Avaliação – HDFS vs MongoDB

- Resultados:
  - Foram configurados 2 nodes HDFS e 2 servidores MongoDB
  - Cada node lê 37.2 milhões de records
  - Performance:
    - ratio de 3:1 entre HDFS e MongoDB em volumes grandes
    - Ratio de 24:1 em volumes pequenos

	READ	WRITE
HDFS	9,3 Milhões/ min	19 Milhões / **15 seg**
MongoDb	2,7 Milhões/min	19 Milhões / **6 min**

6.2 Avaliação – MongoDB MapReduce

- MongoDB possui nativamente, a sua implementação de MapReduce



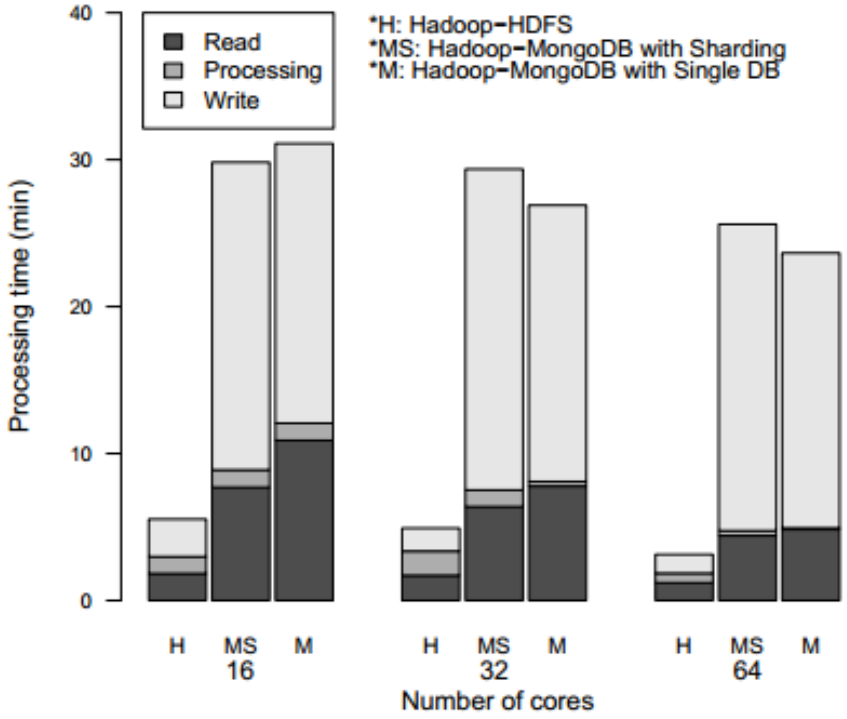
- O gráfico mostra que o ganho da Hadoop-MondoDb em relação a MongoDB Map Reduce nativo é significativamente, tanto a nível , do tempo e do número de records
- Hadoop-MondoDb é mais escalável que MongoDB Map Reduce

6.3 Avaliação – Scalability

- Testaram 3 tipos de combinação entre as duas tecnologias :
  - Hadoop -HDFS
  - Hadoop-MongoDB com Servidores partilhados
  - Hadoop-MongoDb com único Servidor

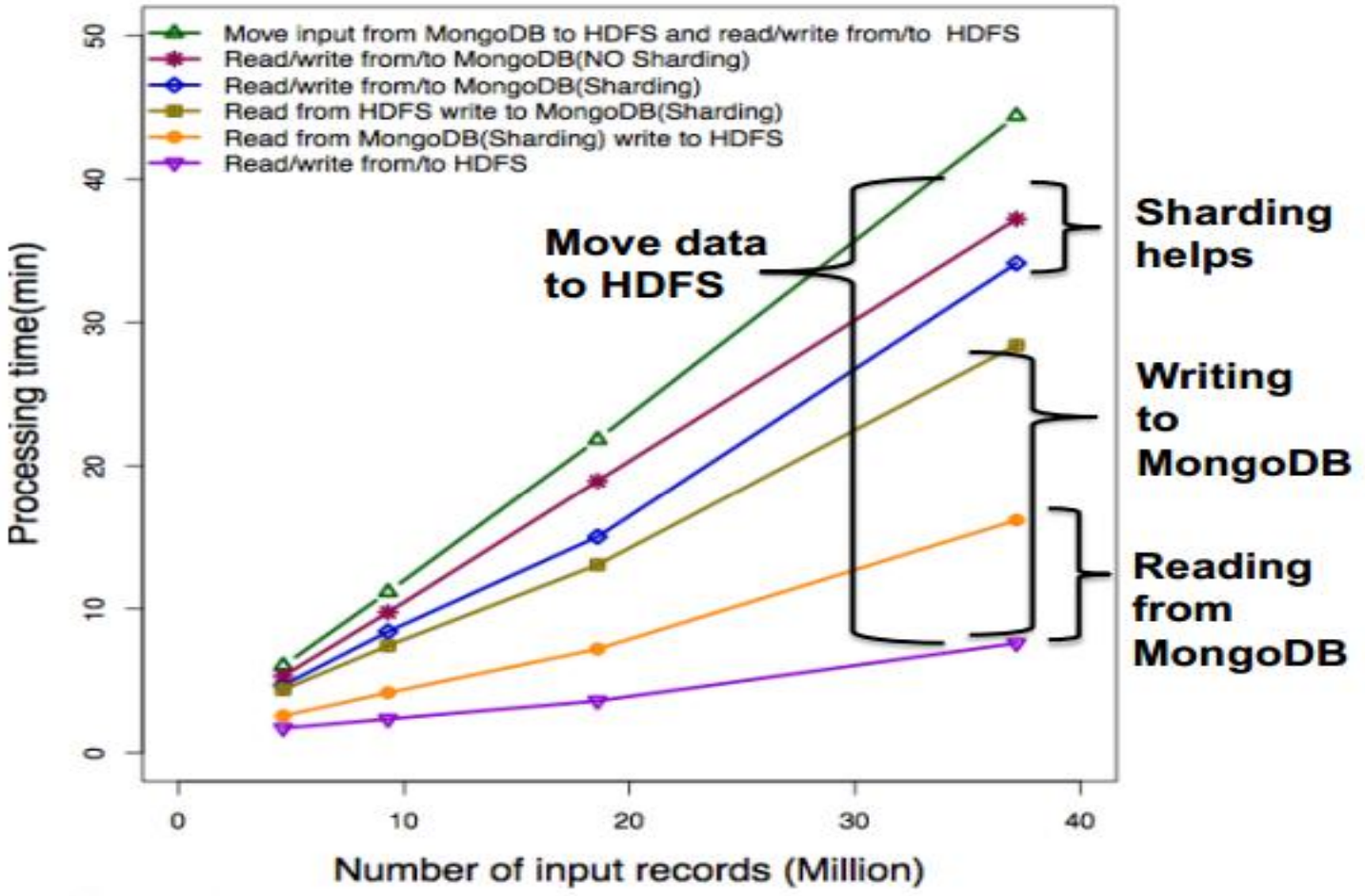
Resultados:

6.3 Avaliação – Scalability



- O Cluster varia entre 16 "cores" a 64 "cores" para de 37 milhões de entradas
- Modelo de servidores partilhados apresenta melhor resultados com o aumento dos cores (logo, aumento dos mappers)
- O gráfico mostra-nos que a operação de escrita é o calcanhar de aquiles do MongoDB

6.3 Avaliação – Scalability

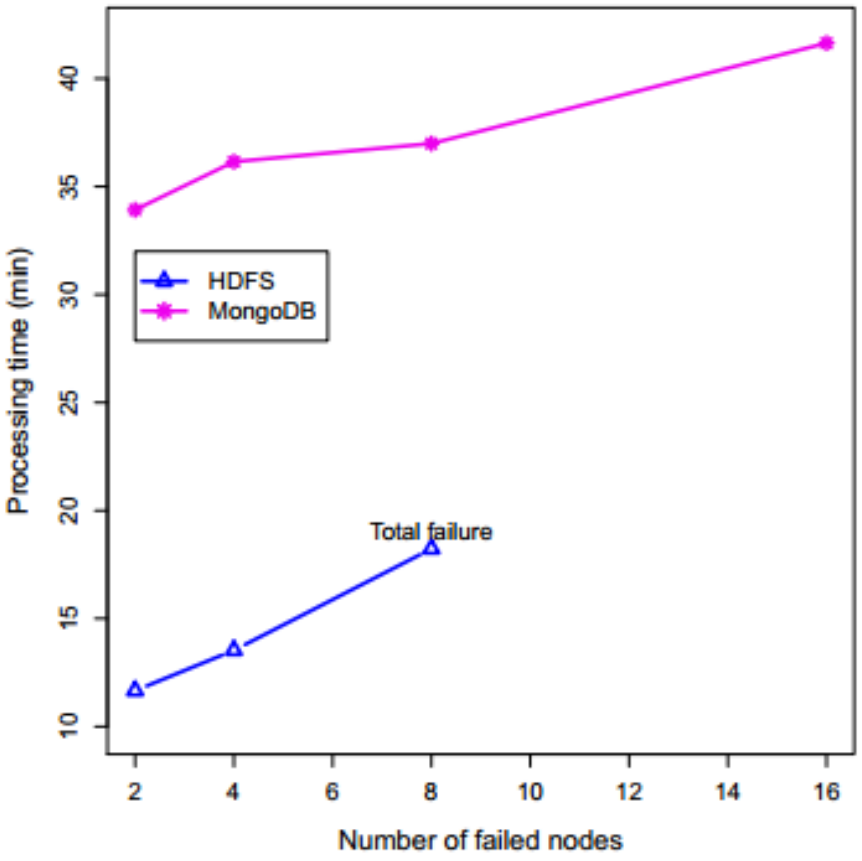


Fonte: http://www.stanford.edu/~paulcon/docs/Hadoop-LBNL-ICME.pdf

6.3 Tolerância a Falhas

- Apresenta-se aqui um gráfico sobre a análise feita sobre a tolerância a falhas
- Teste hadoop-HDFS vd mongoDB-hadoop em 32 nodes num cluster Hadoop
- Após falha em 8 nodes, Hadoop-HDFS apresentou POUCA tolerância a falhas e não termina todas as tarefas MapReduce mesmo que os nodes contenha partes do trabalho já distribuído

6.3 Tolerância a Falhas



- Mongo-hadoop revelou-se MAIS tolerante a falhas, devido ao facto de receber do Servidor MongoDB, logo perda de nodes NÂO, significa perda de dados e a tarefa pode ser concluída, mesmo com metade do cluster disponível.
- Em caso de perda do/os Servidor/os MongoDB deve provocar a mesma falha de tolerância

7. Discussão

- Um único servidor MongoDB revela detioração de performance quando executa entre 1000 e 1500 threads concorrentes (connecções) especialmente na leitura.
- Em casos em que os dados já estão armazenados na MongoDB, a implementação do MapReduce , Hadoop apesenta resultados muito bons a nível de ESCALABILIDADE (5x mais rápido do que MR nativo)

7. Conclusão \ Discussão

- Apesar do volume de dados e número de conecções contribuírem para o overhead do MongoDB, o estudo, mostrou que número de conecções constituem um problema maior que o volume de dados.
- MondoDB apresentou uma maior tolerância a falhas, enquanto que HDFS falhou em completar a tarefa quando houve falha de 8 nodes num cluster de 32

7. Conclusão \ Discussão

- MongoDB trata-se de uma boa alternativa para armazenamento de dados , mas a nível da analise dos mesmos, a Hadoop apresenta melhores resultados
- Introspecção bastante completa sobre o uso de MongoDB e Hadoop
- Apresenta Vantagens e desvantagens em cada operação de READ/WRITE

Referências

- http://docs.mongodb.org/manual/
- http://en.wikipedia.org/wiki/MongoDB
- http://datasys.cs.iit.edu/events/ScienceCloud2013/p02.pdf
- http://www.stanford.edu/~paulcon/docs/Hadoop-LBNL-ICME.pdf