

Modelos de Implantação em Ambientes de Computação em Nuvem: Diferentes Abordagens da Tecnologia MapReduce

Vinícius Renato Rocha Geraldo¹

¹Centro de Desenvolvimento Tecnológico - Universidade Federal de Pelotas (UFPel)
Pelotas - RS - Brazil

vrrgeraldo@inf.ufpel.br

Resumo. *Com a crescente demanda por dados e a evolução constante da internet, aumenta a necessidade por meios de comunicação mais eficientes, com uma plataforma mais robusta e segura. Ambientes para computação em nuvem tornaram-se uma solução aceitável para muitas aplicações de plataformas e softwares em empresas que necessitam desses serviços, em forma de flexibilizar e reduzir custos de computação. Para isso, conceitos de serviços para manejar vem sendo implementados para a melhor solução dos altos conjuntos de dados que provém de clientes e servidores de forma a obter grandes aplicações. Em vista disso, tecnologias de tratamento dos dados acabam tornando-se indispensáveis para computações em nuvem, dada a alta demanda dos dados e grande solicitações de serviços requerem frameworks de bom desempenho, que entreguem o resultado com integridade e verificação de dados. Nesse artigo, apresenta-se uma abordagem acerca das implementações em ambientes de computação em nuvem, serviços, aplicações, descrições das tecnologias de MapReduce e como estas são utilizadas para benefícios dos serviços associados.*

1. Introdução

No decorrer dos anos a necessidade de obter serviços dedicados para cada modelo de tarefa tem se visto cada vez mais fundamental para que o cliente utilize o essencial para quando for requisitado, assim tendências desses estilos tecnológicos vem sendo altamente utilizado em *Cloud Computing* ou Computação em Nuvem. Essa tendência recente disseminada em várias plataformas de desenvolvimento e muitas empresas provém desse modelo de serviço por ser de extremo interesse em oferecer uma visão aos usuários do serviço que podem usufruir de possuir uma infraestrutura de armazenamento de dados ou até conter um software que um conjunto de técnicos necessita para o trabalho conjunto e centralizado em um único ambiente para fácil acesso. Essa concepção de implantação de sistema em nuvem pode ser empregado em diferentes tecnologias onde abordam esse tipo de modelo como em computadores, celulares, base de dados, entre outros que estão apresentados na Figura 1. Essa vantagem de obter todos os dados em nuvem é estudado por grandes empresas para oferecer serviços especializados, como por exemplo Google e Amazon, que possam fazer toda a conexão desses dados com a sua base de dados e a oferecer ao cliente uma ampla oportunidade de recursos computacionais em diferentes aplicações para usufrir de seus dados, softwares associados e oferecendo até recursos de hardware em diversos tipos de nuvem.

Ultimamente muitos desenvolvimentos acabaram sendo transportados para esse tipo de projeto pois podem oferecer gamas de processamento de dados e resolução de



Figura 1. Computação em Nuvem de Modo Geral nas Aplicações

programas que um computador convencional acaba por não entregar esses tipos de características e assim ser discutidas em grandes relevâncias a quem e para que podem ser disponibilizados esse modelo de computação.

A computação em nuvem pode oferecer a portabilidade e a praticidade de recursos de tecnologias serem oferecidos como forma de serviço, assim permitindo aos clientes acessarem os serviços sem precisar conhecer as formas que foram desenvolvidos esses modelos. Porém nesse artigo iremos abordar um pouco mais sobre conceitos e aplicações sobre implantações de serviços de computação em nuvem. Esse artigo será dividido da seguinte forma: a seção 2 descrever os principais conceitos de computação em nuvem. A seção 3 descrever as tecnologias de MapReduce e Hadoop. A seção 4 demonstrar como foi realizada a revisão sistemática ao longo da literatura e comentando sobre modelos de MapReduce nas possíveis melhorias presentes, onde assim é discutido sobre os resultados obtidos na literatura. A seção 5 uma breve conclusão sobre o assunto e a seção 6 as referências que foram utilizadas.

2. Computação em Nuvem

Computação em nuvem vem se tornando cada vez mais um atrativo na maioria dos sistemas que desejam obter um acesso ubíquo, adequado e sob demanda via rede a um agrupamento compartilhado e configurável de recursos computacionais, como por exemplo redes, servidores, equipamentos de armazenamento, aplicações e serviços, que pode ser rapidamente fornecido e liberado com esforços mínimos de gerenciamento ou interação com provedor de serviços definição segundo NIST (Instituto Nacional de Padrões e Tecnologia).

O modelo de computação em nuvem foi implementado para desenvolver serviços de métodos de fácil acesso, baixo custo e com certificações na disponibilidade e escalabilidade dos dados. Dessa forma visa como necessário oferecer três eficiências em relação ao modelo desses, onde um deles é reduzir o custo na aquisição dos dados para entregar aos usuários a necessidade de atender a sua requisição ao serviço. A ideia de utilizar baixo custo na aquisição do modelo é oferecer para grandes empresas um produto de alta qualidade e confiabilidade para o processo de requisição dos dados que são feitos sob demanda e com recursos heterogêneos e de menor custo. Seguinte eficiência é na abordagem de flexibilidade que provém desse modelo entregar no sentido de dispor à adição e subtração de recursos computacionais, sendo capaz de escalar tanto em nível de recursos de hard-

ware quanto software para atender os requisitos das empresas e usuários. Como último emprego é prover uma abstração e facilidade de acesso aos usuários destes serviços.

Nesse sentido o ambiente de computação em nuvem provém de um grande número de máquinas físicas ou nós físicos de baixo custo, conectadas através de uma conexão de rede apresentada na Figura 2. A disposição de como é feita a nuvem está diretamente relacionada a qual tipo de serviço é contratado oferecendo ao cliente grandes desenvolvimentos a serem feitos e projetados.

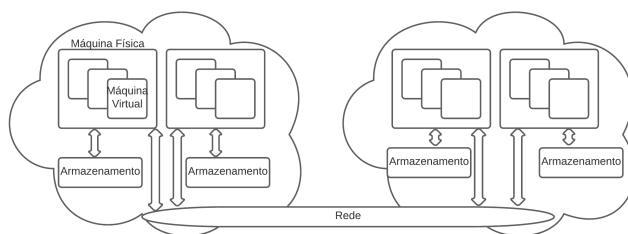


Figura 2. Desenvolvimento de uma Computação em Nuvem

Podendo destacar as principais características que esse tipo de serviço pode oferecer como ser determinado sob demanda em entregar de forma alocável para os recursos, dessa forma colocar o serviço a disposição apenas quando é requisitado e funcionando de maneira que com a necessidade dos recursos o usuário pode extrair sem precisar da interação com os provedores de cada serviço. O amplo acesso via internet possibilita a maior comunicação permitindo a ser utilizada por computadores pessoais, smartphones, acesso a redes corporativas entre outras, isso mostra que esses recursos são apresentados de maneiras padronizadas que podem ser usadas em tais plataformas. Algumas interfaces de computação em nuvem acabam não necessitando que o usuário modifique seu ambiente de trabalho. O compartilhamento de recursos ou como conhecido *polling* oferecem a possibilidade do provedor ser organizado em um *pool* de serviços para atender múltiplos usuários usando o modelo multi-inquilino [Jacobs and Aulbach 2007], disponibilizando de diferentes recursos físicos e virtuais, alocados e atribuídos com a demanda dos usuários. A elasticidade tem a característica de entregar a percepção de recursos ilimitados ao usuário podendo ser adquiridos de forma rápida e elástica caso a necessidade de escalar com o aumento da demanda.

Para esses tais serviços precisamos destacar detalhadamente como funciona exatamente cada um deles e assim mostrar de tal forma que muitas aplicações podem ser separadas ao longo desse tipos de implementações de maneira extrair as características desejadas para o serviço como demonstrado algumas aplicações para cada tipo de serviço na Figura 3 e assim vemos que grandes empresas abordam diferentes modelos de serviço para ser aplicado. Com base nisso, é definido os seguintes modelos para serviço:

- **Software como um Serviço (SaaS):** Nesse tipo de serviço é indicado pelo uso de sistemas de software próprios dos desenvolvedores que são oferecidos para os usuários através da rede de internet. Esses sistemas são acessíveis por quaisquer dispositivos como celulares, computadores pois são entregue em modelos de serviço de web então basta o acesso deles através de um navegador.

Nesse modelo de serviço o usuário acaba por não obter o acesso da infraestrutura de como a nuvem é disposta então apenas tem acesso as funcionalidades e configurações específicas, assim podemos mostrar como exemplo aplicações da Google, tais Google Docs ou até Google Colab, que são serviços na web para desenvolvimento de documentos ou programas que são oferecidos para os usuários. Dessa forma, os custos são mais reduzidos e vem do cliente querer maiores funcionalidades ou até mais recursos de hardware é pago por parte do usuário.

- **Plataforma como um Serviço (PaaS):** Para serviços de plataforma é entregue um modelo que fica entre o SaaS e o IaaS, oferecendo uma proposta mais robusta e flexível na utilização de recursos de tecnologia, assim trazendo que um recurso possa ser desenvolvido para um programa com a disponibilidade de ferramentas oferecidas em nuvem. Sendo assim o usuário possui uma configuração mais ampla no serviço que está implantando seu sistema. PaaS oferecem um sistema operacional, linguagens de programação e ambientes de desenvolvimento para as aplicações, auxiliando a implementação de sistemas de software, de maneira que existe ferramentas de desenvolvimento.

Descrevemos algumas aplicações como mostrado na Figura3 o Google App Engine que oferece uma máquina para que o usuário possa fazer seus testes e amplamente utilizado para mineração de dados e grande manipulação deles de maneira que são oferecido serviços até uma certa quantidade para disponibilidade do usuário caso venha querer mais um custo adicional é necessário para esse estilo de pagamento é chamado de *pay-per-use*.

- **Infraestrutura como um Serviço (IaaS):** Esse modelo de serviço é responsável por conter toda a infraestrutura necessária para a PaaS e o SaaS. A finalidade principal nesse serviço é tornar mais fácil e acessível o fornecimento de recursos, tais como servidores, rede, armazenamento e outros recursos de computação fundamentais para contruir um ambiente sob demanda, que podem incluir sistemas operacionais e aplicativos. De modo geral, o usuário não administra ou controla a infraestrutura da nuvem, porém contém o controle sobre os sistemas operacionais, armazenamento, e aplicativos implantados de tais modos que aplicações como Amazon Web Service EC2 oferecem esses modelos de serviços.







Software como um Serviço (SaaS)	Plataforma como um Serviço (PaaS)	Infraestrutura como um Serviço (IaaS)
  Google Docs	 Google App Engine  Microsoft Azure	 

Figura 3. Aplicações de serviços para Computação em Nuvem

Nos modelos de computação em nuvem possuem disponibilidade em diferentes estilos de implantação que são necessários descrever em relação a restrição ou liberação dos acessos em determinadas partes de seus serviços, isso mostra que empresas podem utilizar desses modelos, mas não querer que determinada configuração seja exposta para os seus usuários que possuem apenas um acesso para realizar suas alterações específicas em determinada parte da aplicação. Para isso a obrigação de conter serviços mais restritos cabe as implantações que são empregadas as quais são nuvem pública, privada, comunidade e híbrida [Jacobs and Aulbach 2007]. As especificações de cada modelo está detalhada a seguir:

- **Nuvem Privada:** A nuvem é gerenciada por uma organização para uso restrito, no entanto, a evidência casual indica que os recursos oferecidos por esse modelo de implantação é a oportunidade de não disponibilizar dados para outras empresas. Este modelo de nuvem pode ser utilizada localmente ou remotamente e são administradas pela própria empresa a fim de extrair características em nível de gerenciamento de redes, configurações dos provedores de serviços e a utilização de tecnologias de autenticação e autorização.
- **Nuvem Comunidade:** Nesse tipo de implantação de nuvem comunidade define pelo compartilhamento por diversas organizações de uma nuvem, sendo esta sustentada por uma comunidade específica que partilhou seus interesses, tais como a missão, os requisitos de segurança, política e considerações sobre flexibilidade. Pode existir localmente ou remotamente e geralmente é administrado por alguma empresa da comunidade ou por terceiros.
- **Nuvem Pública:** De modo diferente da nuvem privada esse modelo oferece uma infraestrutura para ser utilizada no público em geral, sendo acessado por qualquer usuário que conheça a localização do serviço. Neste modelo de implantação restrições de acesso não são aplicadas quanto ao gerenciamento de redes, e também nas utilizações das técnicas para autenticação e autorização.
- **Nuvem Híbrida:** O modelo de nuvem híbrida pode extrair características de duas ou mais implantações de nuvem para um único desenvolvimento de tecnologia, assim dessa maneira pode explorar a disponibilidade de aplicações e dados em diferentes plataformas.

3. Modelo de MapReduce

Na computação em nuvem existem paradigmas ao longo do aperfeiçoamento dos modelos de tecnologias presentes a fim de obter resultados das aplicações conforme a necessidade de conter a resposta do serviço em relação a demanda dos usuários, assim para que isso aconteça de forma a conseguir um maior desempenho na execução das aplicações. Como podemos descrever essa ideia de tecnologia provém de manusear alto volume de dados nas suas execuções onde para isso são criados modelos de programação para o nível de computação em nuvem.

Dessa forma ao longo dos anos o desenvolvimento em tecnologias baseadas em MapReduce é largamente implementada em computações em nuvem no propósito de oferecer um grande manipulamento de volume dos dados em paralelo [Dean and Ghemawat 2010], dividindo as execuções a ser realizada através de tarefas. MapReduce é desejável para tolerância a falhas, distribuição dos dados e balanceamento

de carga destacando para esse modelo a facilidade de desenvolvimento. Muitos desenvolvimento ao longo da literatura são aprofundados por ser amplamente utilizado esse modelo de desenvolvimento onde realizam implementações a fim de melhorar o desempenho do serviço, a segurança, e também na ideia de entregar modelos de serviços que são abordados em diferentes proposições de novos modelos de computação em nuvem que existe.

Para esse modelo de programação MapReduce são desenvolvidas duas funções para classificar os dados onde é visto na Figura 4

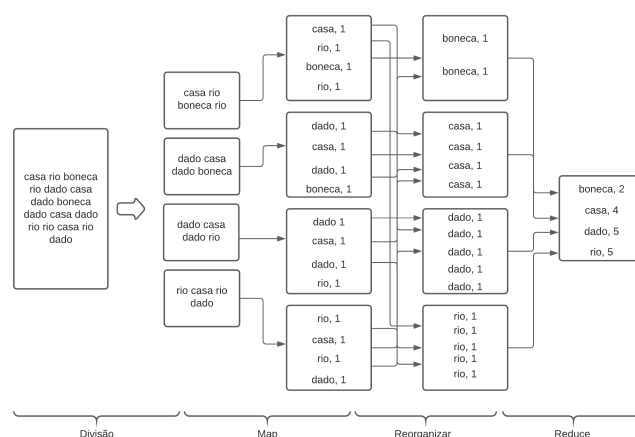


Figura 4. Ferramenta MapReduce Detalhada

Nas implementações de MapReduce podemos destacar a presença das funções Map e Reduce presentes nesse modelo obtendo assim uma programação simples e eficiente para manipulação dos dados. Basicamente a função Map com as entradas dos dados como mostra na Figura 4 é utilizado uma estrutura de dado do tipo par chave-valor (chave, valor) de maneira a realizar um mapeamento dos dados no decorrer da distribuição dos dados recebidos como entrada. Dessa forma uma função com uma entrada de par de chave gera um par de chave intermediário para que na função Reduce seja utilizada como entrada. Na manipulação da função Reduce é realizado a mesma ideia de par chave-valor, porém é feita uma lista com todos os valores para cada valor de chave correspondente gerando assim outro par de chave-valor.

Esse mecanismo foi desenvolvido em meados de 2004 pela Google com conceitos similares a linguagens funcionais onde é implantado semelhantemente a ferramenta Hadoop [Hadoop 2020] na qual são oferecidos sistemas distribuídos de servidores em clusters para que seja oferecida a ideia de escalabilidade horizontal entre seus serviços para que seja manipulados grandes processamento de dados em paralelo, assim por ser um código *open-source* e se aproximar das resoluções de MapReduce é feita melhorias ao longo desse modelo para entregar as abordagens nas aplicações. No Hadoop é composto de um sistema de arquivos *Hadoop Distributed File System* (HDFS) responsável por controlar toda a distribuição dos arquivos ao longo do servidor de maneira a aumentar a escalabilidade do sistema e ser tolerante a falhas. Por ser uma ferramenta de código livre muitos autores tentam viabilizar esses mecanismos de maneira a extrair maiores funcionalidades ao longo de aplicações que são desenvolvidas como gerenciamento de arquivos em

nuvem, criação de modelo de serviços, aprendizagem de máquina entre outros tipos de processamento de dados.

4. MapReduce na Literatura

Como discutido anteriormente as funções do MapReduce visam otimizar o processamento paralelo dos dados no contexto de computação em nuvem, mas para isso muitas propostas e abordagens sobre esse conceito vem sendo altamente desenvolvidas e aprimoradas ao longo dos anos. Como métrica para esse trabalho foi selecionado na literatura utilizando como base implantações de computação em nuvem que utilizam como funcionalidade da ferramenta MapReduce para seu aperfeiçoamento, assim de maneira satisfatória extrai informações utilizando a ferramenta StArt (*State of the Art through Systematic Review*) onde foi definido a maneira como iria realizar a revisão sistemática de forma a definir um planejamento para realização do protocolo de trabalho. Para isso define primeiramente qual seria feita a *string* de busca nas bases de dados onde utilizei da IEEE e da ACM como referência fazendo a pesquisa com as seguintes palavras-chave: *cloud programming, cloud computing, mapreduce, data storage, applications*, desse modo é obtido vários artigos relacionados sobre essa área de forma que entra a parte da extração da revisão, então define um critério para seleção dos artigos que utilizei para fazer as leituras. Assim foi utilizado apenas uma parcela considerável para a extração das informações do que são implementados e em que área de aplicações são abordadas.

Como aspecto geral é mostrado por [Zhang et al. 2012] mostra um aspecto a cerca de realizar implementações voltadas com grande processamento de dados utilizando a ferramenta MapReduce no desenvolvimento de mais uma camada de arquitetura para preservação da privacidade do usuário. Isso mostra que realizando tarefas provindas do cliente é necessário uma confiabilidade maior dos seus dados, assim os dados originais armazenados na nuvem contém uma confiabilidade maior por conter essa camada extra de acessibilidade. Podemos fazer o comparativo com a Figura 4 que faz a função tradicional do MapReduce de forma que é proposto essa camada entre a realização do Reduce uma segunda utilização em cascata do Map para entregar o resultado ao usuário que requisitou pelo serviço, dessa maneira essa camada é dividida em módulos para anonimização dos dados e garantir uma maior flexibilidade, escalabilidade, dinamicidade e um custo efetivo para que seja cumprido a privacidade.

No modo de obter aplicações mais desenvolvidas dos modelos de serviços na nuvem diversos tipos são incrementados com alguma contribuição gerando resultados extremamente relevantes, como citado em [Iordache et al. 2013] onde comenta exatamente na otimização do MapReduce utilizando uma API compatível a sistemas distribuídos para computação em enfoque na elasticidade oferecida por esses serviços. Como exemplo trabalhado é feito uma arquitetura proposta de *Resilin* onde é abordado utilidades oferecidas e comparadas exatamente em serviços oferecidos em ferramentas Amazon EMR que fornecem essa aplicação, como algumas vantagens está no processamento em vários tipos de nuvens executando serviços em máquinas virtuais pelos clusters.

De contra partida a esses autores comentados, muitos serviços necessitam de espera na hora de atender os usuários do seu sistema para que seja oferecidos de maneira a entrega a requisição quando solicitado. Para isso [Tsai et al. 2011] realiza uma abordagem de oferecer serviços que realizem uma replicação dos dados para ser entregue

atendendo entre várias requisições ao longo dos serviços. Para isso foi apresentado a ideia de utilizar MapReduce na sua implementação pelo fato de oferecer um serviço paralelo de distribuição dos dados de maneira a se desenvolver estratégias de replicação dos serviços, assim fazendo uma utilização de serviços de cache para gerenciamento dos dados e assim que requisitados ter um acesso mais rápido provendo para isso um desempenho maior. Como proposta da replicação é desenvolvido duas estratégias de replicação, ativos e passivos com objetivos diferentes de solução no acesso do dado pelo usuário.

Por outro lado muitos artigos comentam sobre aplicações voltados para gerenciamento de *Big Data* que muito autores abordam na literatura em otimizações realizadas para armazenamento de dados de maneira a obter desempenho utilizando funções de MapReduce como mostrados em [Liang and Zhou 2019, Jin et al. 2016, Sheshasaayee and Megala 2017]. Uma outra aplicação que é feita que podemos destacar é no fluxo de execução das funções Map do processamento dos dados através dos diferentes modelos de nuvem [Cao and Wu 2018]. Essa otimização é feita basicamente em mudar os estilo de algoritmos para as funções do MapReduce a fim de extrair uma melhor performance do sistema.

5. Conclusões

Podemos concluir que o sistema de computação em nuvem são utilizados largamente em vários modelos de aplicações e produtos que necessitam de uma maior interação entre os usuários com uma gama de serviços para ser oferecido. Vemos que na maioria das implementações são baseadas utilizando a tecnologia MapReduce e que na literatura há um apelo enorme em otimizar essa função a fim de buscar maiores segurança, escalabilidade, desempenho e flexibilidade em relações a falhas na hora de atender uma requisição ou em momentos de utilização da nuvem. Vale ressaltar que a base de dados que foi utilizada para realizar essa revisão sistemática é focada apenas para aplicações que realizam MapReduce na sua metodologia e que existe maiores aplicações voltadas para computação em nuvem. Muitos conceitos foram retirados de [Sousa et al. 2020].

Referências

- Cao, H. and Wu, C. Q. (2018). Performance optimization of budget-constrained mapreduce workflows in multi-clouds. In *2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pages 243–252.
- Dean, J. and Ghemawat, S. (2010). Mapreduce: A flexible data processing tool. *Commun. ACM*, 53(1):72–77.
- Hadoop (2020). Apache hadoop. In <http://hadoop.apache.org>.
- Iordache, A., Morin, C., Parlavantzas, N., Feller, E., and Riteau, P. (2013). Resilin: Elastic mapreduce over multiple clouds. In *2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing*, pages 261–268.
- Jacobs, D. and Aulbach, S. (2007). Ruminations on multi-tenant databases. In *BTW*, volume 103, pages 514–521.
- Jin, Y., Yan, D., and He, H. (2016). Research on mapreduce-based cloud storage batch auditing. In *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, pages 1317–1322.

- Liang, C. and Zhou, L. (2019). Research on distributed storage of big data based on hbase remote sensing image. In *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 1, pages 2628–2632.
- Sheshasaayee, A. and Megala, R. (2017). A theoretical framework for cloud resource provisioning using mapreduce technique. In *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pages 664–666.
- Sousa, F., Moreira, L., and Machado, J. (2020). Computação em nuvem: Conceitos, tecnologias, aplicações e desafios.
- Tsai, W., Zhong, P., Elston, J., Bai, X., and Chen, Y. (2011). Service replication strategies with mapreduce in clouds. In *2011 Tenth International Symposium on Autonomous Decentralized Systems*, pages 381–388.
- Zhang, X., Liu, C., Nepal, S., Dou, W., and Chen, J. (2012). Privacy-preserving layer over mapreduce on cloud. In *2012 Second International Conference on Cloud and Green Computing*, pages 304–310.