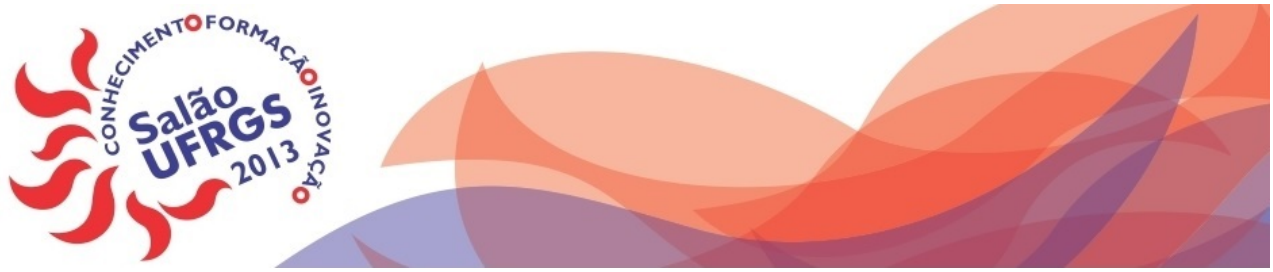

Evento	Salão UFRGS 2013: SIC - XXV SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2013
Local	Porto Alegre - RS
Título	Além do Hadoop, uma análise sobre os novos paradigmas de Computação em Nuvem
Autor	OTÁVIO MORAES DE CARVALHO
Orientador	PHILIPPE OLIVIER ALEXANDRE NAVAUX

O processamento de grandes volumes de dados tem sido dominado pelo modelo de programação MapReduce, originalmente proposto pela Google e largamente adotado através da implementação Apache Hadoop. É um modelo inspirado nas operações de *map()* e *reduce()* da programação funcional, onde o *map* representa operações de filtro de ordenação, e o *reduce* representa a operação de redução, para obtenção de resultados sobre os dados processados.

Porém, desde a sua criação até os dias atuais, foram identificados pontos fracos na abordagem de processar dados da forma como o MapReduce opera, fortemente otimizada para aplicações do tipo *batch*, que operam de maneira sequencial sobre grandes volumes de dados, não apresentando capacidades amplas de análise sobre os dados processados, características largamente presentes nos modelos de banco de dados relacionais existentes.

O objetivo desse trabalho é realizar um levantamento das ferramentas criadas para a evolução do paradigma, buscando identificar quais dos problemas elas se propõem a resolver, bem como as oportunidade de pesquisa oferecidas pelas ferramentas. Dentre as ferramentas estudadas, temos:

O Google BigTable, que é uma implementação de armazenamento de dados distribuída. Inspirou o surgimento de ferramentas como o HBase e o Facebook Cassandra, que tornaram possível não somente processar os grandes volumes de dados, mas também armazená-los e consultá-los de forma mais eficiente.

O Google Pregel, que consiste em uma ferramenta para o processamento de dados agrupados na forma de grafos, através do qual surgiram ferramentas como o Apache Giraph, Apache Hama e, mais recentemente, o GPS.

O Google Dremel, ferramenta distribuída para processamento de consultas ad-hoc interativas, ferramentas como o Apache Drill, Apache Pig e Apache Hive, se desenvolveram. Porém, o Apache Pig e o Apache Hive, apesar de serem evoluções, ainda operam sobre o Hadoop.

O Google Percolator, criado para atualizar incrementalmente os dados armazenados pela Google, inspirou o surgimento de diversas ferramentas para o processamento de fluxos contínuos de dados, implementados como processadores complexos de eventos, tais como o Twitter Storm e o Apache S4.

Além disso, existem outras aplicações sendo desenvolvidas paralelamente, agregando características das anteriores, mas também evoluindo essas idéias em outras direções. O que é o caso da implementação em memória do MapReduce, denominada Spark, desenvolvida pelo AMPLabs da UC Berkeley, que vem ao encontro de outras ferramentas de processamento semelhantes, como o SAP Hana e o Druid.

Portanto, conseguimos identificar que durante o desenvolvimento das novas abordagens, para a resolução dos diferentes problemas apresentados, foram criadas inúmeras aplicações. Essa situação levou a uma dispersão do mercado dentre as diversas aplicações disponíveis.

Esse trabalho representa o esforço inicial na escolha do ferramental a ser utilizado para uma avaliação mais profunda e, posteriormente, para a realização da avaliação da aplicabilidade das mesmas, para a resolução de problemas que necessitem de processamento de alto desempenho.