BUSINESS INTELLIGENCE 4.0: MANIPULANDO ALTO VOLUME DE DADOS DE MANUFATURA DE FORMA DISTRIBUÍDA E *IN-MEMORY* *

Alexandre Keunecke Hardt[1]
Christian Pinto de Souza[2]
Felipe Chagas Rabello[3]
Gustavo Vieira Machado[4]
Roberto Resque de Freitas[5]

Resumo

A dinâmica competitiva do mercado globalizado demanda informações confiáveis no dia-a-dia das empresas. Essas informações são preciosas, pois estabelecem vantagens que permitem que as empresas mantenham suas lideranças no mercado. Enquanto sistemas de *Business Intelligence* (BI) normalmente pertencem ao mundo da gestão e da estratégia, os conceitos da Indústria 4.0 trazem um conjunto de novas oportunidades para obtenção e extração de dados direto do chão-de-fábrica, permitindo com isso que decisões operacionais e táticas suportadas por esses dados sejam tomadas quase em tempo real. Uma solução BI desenvolvida em tecnologia OLAP (*Online Analytical Processing*) possui na etapa de ETL (*Extract Transform Load*) um de seus maiores gargalos. Este trabalho propõe uma ferramenta de processamento desenvolvida em *Spark*, denominada neste trabalho como BI 4.0, que resolve esse problema ao realizar o ETL de forma escalável, distribuída e com alta performance, trazendo ganhos significativos para o negócio.

Palavras-chave: Indústria 4.0; BI 4.0; MES; OLAP; Processamento distribuído; Spark.

BUSINESS INTELLIGENCE 4.0: PROCESSING HIGH VOLUME MANUFACTURING DATA IN A DISTRIBUTED AND IN-MEMORY APPROACH

Abstract

The competitive dynamics of the globalized market demands reliable information on the internal and external reality of corporations. This information is a precious asset and is responsible for establishing key advantages to enable companies to maintain their leadership. Whereas Business Intelligence systems (BI) are usually a managerial and strategic information realm, Industry 4.0 concepts brings a whole set of new opportunities to gather and extract data straight from the shop-floor, therefore allowing operational and tactical decisions supported from these data to be taken almost in real time. A BI solution developed in accordance to OLAP (Online Analytical Processing) technology has in the ETL phase (Extract Transform Load) one of its main points of bottleneck. This work proposes a processing tool developed in Spark, entitled BI 4.0, which addresses this problem by performing the ETL step in a scalable, distributed and high performance approach, dropping the processing time and adding value to the business.

Keywords: Industry 4.0; Business intelligence 4.0; MES; OLAP; Distributed computing; Spark.

[1] *Engenheiro de Controle e Automação, Consultor de projetos MES da Accenture do Brasil, Belo Horizonte, Minas Gerais, Brasil.*
[2] *Engenheiro Eletricista, Consultor de projetos MES da Accenture do Brasil, Belo Horizonte, Minas Gerais, Brasil.*
[3] *Engenheiro Eletricista, Analista de projetos MES da Accenture do Brasil, Belo Horizonte, Minas Gerais, Brasil.*
[4] *Engenheiro de Controle e Automação, Analista de projetos MES da Accenture do Brasil, Belo Horizonte, Minas Gerais, Brasil.*
[5] *Engenheiro de Controle e Automação, Gerente de projetos MES da Accenture do Brasil, Belo Horizonte, Minas Gerais, Brasil.*

1 INTRODUÇÃO

As empresas do setor metalúrgico e mineral estão inseridas em um contexto de alta competitividade num mercado cada vez mais globalizado, configurando-se assim um cenário bastante desafiador. Como alternativa para diferenciação perante aos concorrentes, muitas organizações têm focado boa parte dos seus investimentos na redução de custos operacionais e no aumento de produtividade. Muitos desses investimentos envolvem novas tecnologias e/ou *softwares* para automação dos processos produtivos e até mesmo processos gerenciais.

Em paralelo, a evolução tecnológica dos equipamentos e dispositivos utilizados na produção e a maior interconectividade entre as máquinas, impulsionada pelos conceitos da Indústria 4.0 [1] (sobretudo a Internet das Coisas), permitiu um aumento significativo da disponibilidade de dados a respeito do processo produtivo. Tais dados, se receberem o devido tratamento, podem ser muitas vezes a chave para o entendimento dos gargalos operacionais e/ou identificação de oportunidades "escondidas" que permitam a tão desejada diferenciação competitiva.

Essa situação traz uma nova necessidade para as empresas: a necessidade de serem capazes de manipular um volume de dados de manufatura cada vez maior, tendo como desafio integrá-los e transformá-los em informações precisas e relevantes ao negócio. Essas informações, articuladas com os direcionamentos estratégicos da organização e disponibilizadas em tempo real, representam ativos preciosos, pois podem ser usadas tanto nas atividades cotidianas (operacional) quanto como subsídios para processos de decisão (gerencial).

A consolidação das informações de manufatura tipicamente é uma responsabilidade dos sistemas de *Business Intelligence* (BI). O BI é um tipo de sistema de suporte principalmente aos gestores, que apoia os processos de decisão. Ele é formado por um conjunto de arquiteturas, banco de dados, técnicas e ferramentas analíticas voltados para o cruzamento, integração, análise e apresentação de todas as informações existentes na organização e necessárias em seus processos de tomada de decisão. Ele permite que as informações geradas pelas instituições sejam unificadas, estruturadas, modeladas e apresentadas de uma forma simples e objetiva, proporcionando a realização das análises adequadas, convertendo-se em conhecimentos sobre o desempenho corporativo que vão auxiliar a tomada de decisão.

Um alto volume de dados, dependendo da aplicação, pode fazer com que o processo de consolidação das informações do BI demore horas para ser concluído e isso pode significar em atraso na disponibilização de uma informação importante para a tomada de decisão do negócio. Assim sendo, com o aumento da disponibilidade de informações, fomentada pelos conceitos da indústria 4.0, torna-se mandatório repensar as estratégias utilizadas para concepção dos sistemas de BI, de forma que eles consigam processar os dados em tempo suficientemente pequeno a ponto de se aproveitar do benefício da nova gama de informações de manufatura disponíveis. Como consequência dessa otimização, torna-se possível repensar também sobre a abrangência dos sistemas de BI, que podem passar a ser utilizados não somente para informações gerenciais, como também para informações operacionais.

Este trabalho propõe uma ferramenta distribuída, paralelizada e tolerante a falha, como alternativa de processamento de dados em memória para sistemas de BI, de forma a suportar um alto volume de dados de manufatura e a necessidade em tempo real inerentes da Indústria 4.0. Desenvolvido em *Spark* [2], um framework de computação em cluster, o BI 4.0 (como está sendo referenciado neste artigo)

apresenta todas as características desejáveis em um sistema distribuído: escalabilidade, tolerância a falha e alta performance.

2 CONCEPÇÃO DO BI 4.0

2.1 OLAP (*Online Analytical Processing*)

O OLAP [3], ou *Online Analytical Processing*, é uma tecnologia, inserida no contexto de *Business Intelligence* e *Data Warehouse* [3], utilizada para analisar dados multidimensionais de forma interativa e a partir de múltiplas perspectivas. Assim, o OLAP permite que sejam analisados grandes volumes de informações de forma dinâmica, rápida e com diferentes níveis de detalhamento. O modelo de dados OLAP é organizado na forma de um cubo, contendo dois tipos básicos de dados: as medidas, que são dados numéricos, e as dimensões, que são as categorias responsáveis por organizar as medidas. Dessa forma, o cubo agrega as medidas a partir dos vários níveis e hierarquias de cada dimensão que se deseja analisar.
Conforme figura abaixo, o fluxo completo de disponibilização de dados em um cubo OLAP começa na extração de informações do (s) sistema (s) de origem, passa pela sua transformação e, concluída esta transformação, é feita a carga destes dados na tabela ”fato”. A partir das informações contidas na tabela ”fato” e nas tabelas de dimensões, o cubo é processado.

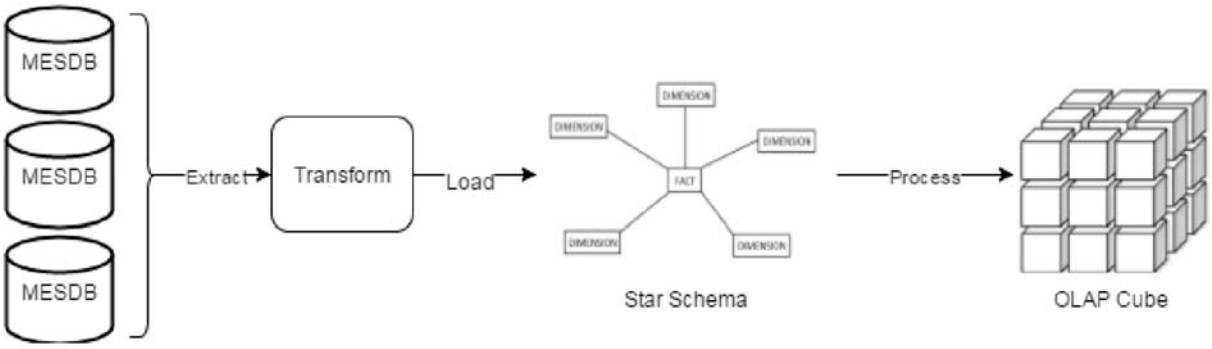

Figura 1. Desenho esquemático das etapas de atualização/cálculo de um cubo OLAP.

As etapas de extração transformação e carga, mais conhecidas como ETL, ou *Extract-Transform-Load*, é um dos principais, senão o principal, gargalo no processamento das informações para cubos OLAP e, com o aumento do volume de informações a serem tratadas e a eventual necessidade de redução da granularidade dos dados, o problema pode aumentar exponencialmente.

2.2 Escolha do *Spark*

Dentre todos os *frameworks* de computação em *cluster*, o *Spark* mostrou-se o mais adequado para esta aplicação. Conforme apresentado pelos autores do *framework,* para aplicações que carregam todos os dados em memória, sem etapas intermediárias de persistência, o *Spark* apresenta desempenho até 100 vezes melhor que o *MapReduce* [4,5]. Outro ponto a favor do *Spark* é a fácil integração entre seus vários módulos, que disponibilizam tudo o que é necessário para o processamento de dados massivos, eliminando a necessidade de ferramentas adicionais.

Outro ponto desfavorável do *Hadoop* é o seu modelo orientado a ”mapa e redução”, restringindo a sua aplicabilidade e obrigando os desenvolvedores a implementar suas ferramentas dentro deste paradigma, mesmo quando esta abordagem não é a mais eficiente para tratar o problema em questão. No caso do *Spark*, seu modelo mais genérico permite o uso de mais operadores, além do *map* e do *reduce*, o que o torna um *framework* extremamente versátil, podendo ser utilizado em uma gama maior de problemas.

2.3 Arquitetura do BI 4.0

O BI 4.0 utiliza a arquitetura proposta pelo *Spark*, ou seja, dentro do *cluster* tem-se o gerenciador do *cluster* (mestre) e os *workers*, responsáveis por gerenciar os recursos computacionais (*executors*) de cada nó. Os *executors* são responsáveis por realizar tarefas computacionais.
Dessa forma, o BI 4.0 descreve, na forma de um grafo acíclico dirigido (GAD), o fluxo de tarefas da aplicação, ou seja, o BI 4.0 é o *driver program* da arquitetura. Ao submeter uma tarefa para o cluster, uma série de *workers* são inicializados, que por sua vez ativam os executors. Os *executors* realizam, assim, as tarefas computacionais requeridas pelo *driver program*.
Como o *driver program* é representado na forma de um GAD, o *Spark* armazena a ”genealogia” das tarefas executadas. Dessa forma, quando um nó falha, o gerenciador do *cluster* pede para outro nó reprocessar as tarefas que estavam sendo executadas pelo nó que falhou, garantindo assim, tolerância à falha.

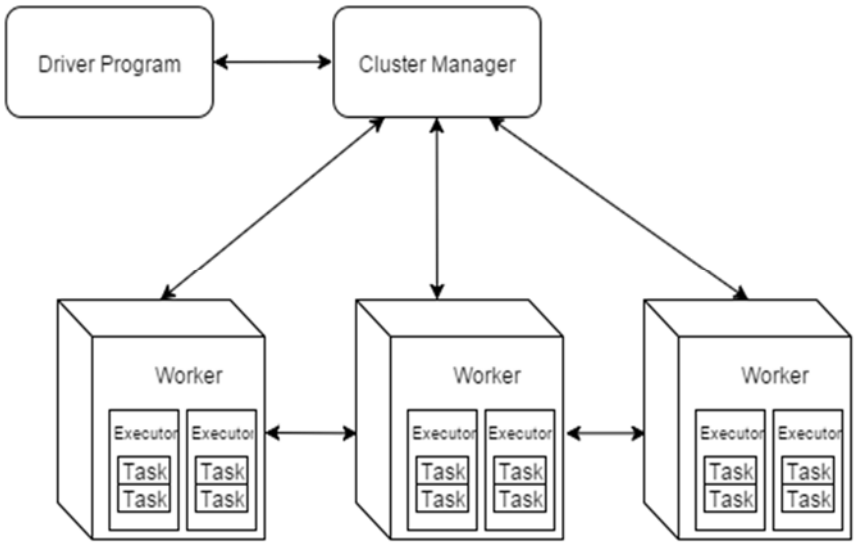

Figura 2. Desenho esquemático da arquitetura do BI 4.0

2.4 Funcionamento

Na etapa de extração é utilizada uma ferramenta, fornecida pelo próprio *Spark*, para leitura de informações em base de dados relacionais. Esta ferramenta permite que os dados sejam carregados no *cluster* de forma distribuída, organizando-os através de uma chave. No caso do BI 4.0, como a aplicação foi para dados de manufatura, a chave utilizada foi o código do equipamento produtivo. Desta forma, cada *executor* do *cluster* abre uma conexão com o sistema de origem e carrega somente os dados de

produção do equipamento da chave em questão, paralelizando, assim, o processamento entre os equipamentos.
Os dados lidos do sistema de origem são transformados em RDDs (*Relational Distributed Datasets*), uma abstração utilizada pelo *Spark* para manipulação de dados de forma distribuída. Essa transformação gera RDDs de cada uma das entidades do modelo de dados do BI 4.0, que são, posteriormente, utilizados na etapa de transformação.
Concluída a extração, o BI 4.0 inicia o processo de transformação por meio de um algoritmo, desenvolvido pelos autores, de unificação e quebra de informações baseado em chave-valor. No caso, todas as informações são analisadas no domínio do tempo, onde são identificadas interseções entre os dados de produção, os status do equipamento e os turnos, com o objetivo de gerar uma entidade de dados, representando todos os indicadores de produção quebrados em granularidades correspondentes às interseções.
Esta etapa, portanto, aproveita a etapa de extração, que particionou os dados de produção por código do equipamento, e agrupa todos estes dados particionados, gerando tuplas em que a chave é o código do equipamento e os valores são listas das entidades, geradas na etapa de extração, relacionadas ao equipamento em questão. Como os cálculos são independentes entre os equipamentos, esta separação permite que o *Spark* paralelize e distribua as tarefas de equipamentos distintos.
Feito o agrupamento, as informações que possuem algum tipo de interseção no domínio do tempo são analisadas para posterior cálculo de seus indicadores de produção, gerando uma lista de resultados, na qual a menor granularidade é representada pelos pontos de interseção entre todas as entidades buscadas para o equipamento em questão
Realizada a identificação de interseções e os cálculos de indicadores, o resultado final deste processo é carregado na tabela ”fato”, finalizando todo o processo ETL. Como as informações estão particionadas por equipamento e o *Spark* permite que cada agrupamento (partição) trabalhe de forma independente, os dados são carregados na tabela ”fato” também de forma paralelizada e distribuída.

3 APLICAÇÃO DO BI 4.0

O BI 4.0 pode ser aplicado em qualquer tipo de indústria e processo de produção que possua um ou múltiplos sistemas de gestão da manufatura (MES/MOM) e com alto volume de dados para serem manipulados/processados em um curto espaço de tempo. Especialmente em processos com alto nível de integração com o chão-de-fábrica, onde o volume de informações disponibilizadas é elevado, a sua aplicação tende a ser mais relevante.
A solução pode processar dados de múltiplas fontes de dados de uma mesma planta ou até mesmo dados de diversas plantas diferentes, provendo assim dados de manufatura *cross-site.* No caso deste trabalho, como estudo de caso, foi utilizada a base de dados de uma solução MES de três plantas de uma empresa siderúrgica, contendo em cada base cerca de 40GB de dados de produção.
Para avaliar a eficácia da ferramenta, foram realizados três experimentos distintos, conforme apresentados a seguir.

1. BI 4.0 *versus* Abordagem Tradicional: este experimento teve como objetivo avaliar o desempenho do BI 4.0 em comparação a uma das abordagens tradicionais de extração e consolidação de cubos OLAP, que é baseada em

consultas diretas ao banco de dados, que no caso em questão foi utilizado o SQL Server como SGBD (Sistema de Gerenciamento de Banco de Dados);

2. Escalabilidade e Desempenho do BI 4.0: o segundo experimento teve como objetivo avaliar a escalabilidade e o desempenho do algoritmo em um *cluster;*
3. Tolerância a Falhas do BI 4.0: o terceiro e último experimento teve como objetivo avaliar o comportamento do BI 4.0 durante eventuais falhas nos nós do *cluster*, o que é uma situação comum em um ambiente corporativo de sistemas de manufatura.

3.1 Primeiro experimento: BI 4.0 *versus* Abordagem Tradicional

A solução representada pela abordagem tradicional consiste em uma série de procedures, executadas dentro do servidor do banco relacional (neste caso, *Microsoft SQL Server*), que agregam as informações, transformam-nas e salvam-nas na tabela ”fato”. Seu funcionamento é sequencial e sem qualquer tipo de distribuição. A infraestrutura desta solução é constituída por um servidor com 24 núcleos e 32GB de memória RAM.

Já para o BI 4.0, foi utilizado um cluster de três servidores, cada um com 4 núcleos e 8 GB de memória, sendo um mestre e três *workers*. Cada *worker* foi configurado de forma a ter quatro *executors*. Portanto, o cluster conta com doze recursos para processamento.

Como pode ser observado, a arquitetura dedicada à solução tradicional, é superior em números absolutos a estrutura do *cluster* e o objetivo de tal condição é justamente provocar uma reflexão a respeito da melhor utilização dos recursos computacionais disponíveis.

Para avaliar o desempenho das soluções foram realizados testes variando-se o período de tempo de busca de dados de produção (“janela dos dados”). Os períodos utilizados foram doze horas, dois dias e sete dias de histórico de dados.

A Tabela 1 abaixo apresenta os resultados do experimento.

Tabela 1. Resultados do primeiro experimento (BI 4.0 versus Abordagem Tradicional)

Período (número de dias)	0.5	2	7
Tempo Abordagem Tradicional (min)	60	180	600
Tempo BI 4.0 (min)	0.4	0.6	0.9

A Figura 3 apresenta graficamente os resultados.

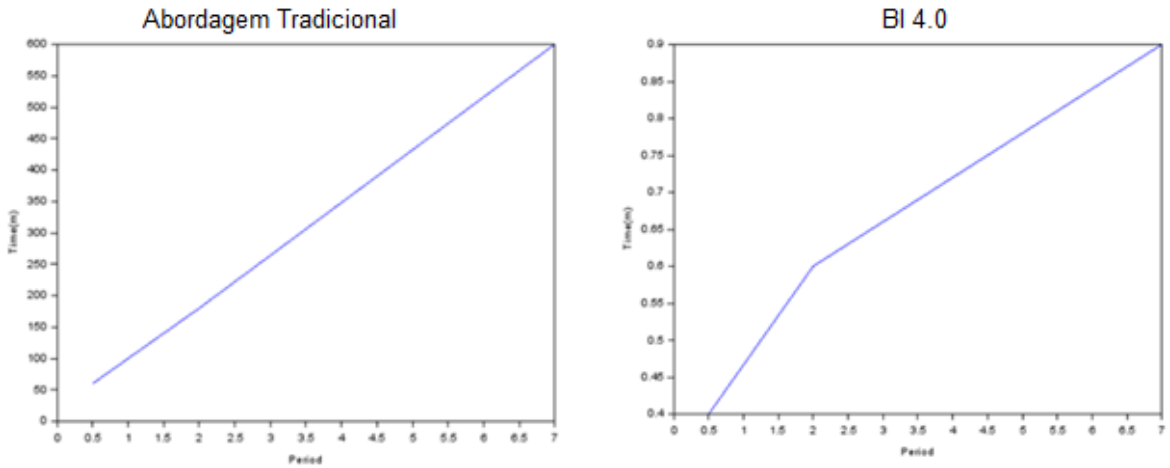

Figura 3. Resultados comparativos da Abordagem Tradicional (esq.) versus BI 4.0 (dir.)

Como é possível observar, o BI 4.0 apresentou um tempo de processamento até 600 vezes mais rápido que a abordagem tradicional. O tempo de processamento de um período de sete dias, que é de dez horas na solução convencional, passou a cinquenta segundos com o BI 4.0.

3.2. Segundo experimento: Escalabilidade e Desempenho do BI 4.0

Para este experimento foram realizados testes variando-se o número de *executors*, disponíveis em um cluster, de um a doze e, para cada configuração do cluster, o BI 4.0 foi executado de forma a processar um período de dados de produção de sete dias, ou seja, mantendo-se a referência constante.
A Tabela 2 abaixo apresenta os resultados do experimento.

Tabela 2. Resultados do segundo experimento (Escalabilidade e Desempenho do BI 4.0)

Número de *Executors*	1	4	8	12
Tempo BI 4.0 (s)	225	138	73	50

A Figura 4 apresenta graficamente os resultados.

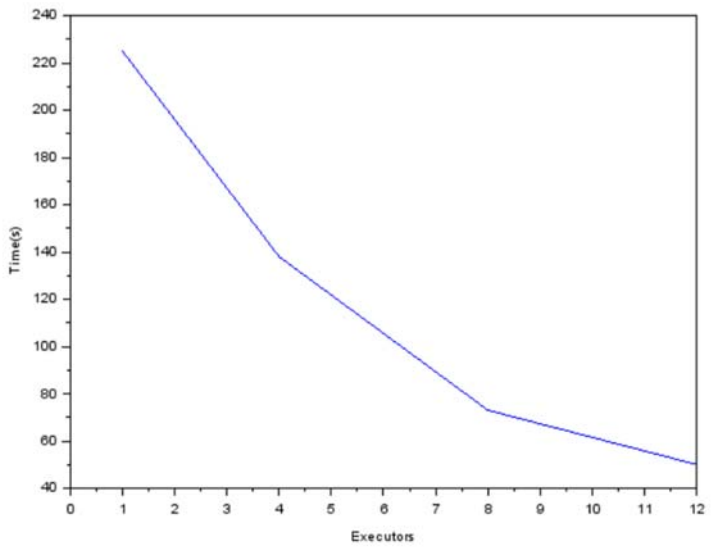

Figura 4. Escalabilidade e Desempenho do BI 4.0.

Como pode ser observado, o BI 4.0 com o seu algoritmo baseado em chave e valor, se mostrou altamente escalável, podendo ser utilizado em processamento de dados e *clusters* ainda maiores.
Vale destacar que, com apenas um executor, o BI 4.0 já apresentou um resultado significativamente melhor que a abordagem tradicional. Isso se deve ao processamento paralelo que ocorre no próprio *executor*, ao processamento '*in-memory*' e ao algoritmo desenvolvido pelos autores, com um alto nível de abstração.

3.3. Terceiro experimento: Tolerância a Falhas do BI 4.0

Neste experimento, o BI 4.0 foi iniciado com o *cluster* completo (12 *executors*) e, durante a execução, metade dos servidores foi removido, indo de doze *executors* a seis abruptamente.

Foi observado que a rotina completou sua execução em 80 segundos e sem gerar nenhuma inconsistência nos dados, ou seja, o processamento não foi interrompido e apenas a duração foi estendida, em função da incapacidade de alguns *executors* de executar a tarefa, o que acarretou em menor capacidade de processamento e consequentemente um maior tempo para completude da tarefa.
Dessa forma foi possível comprovar que o BI 4.0 é tolerante a falhas, o que aumenta significativamente a robustez da solução.

4 CONCLUSÃO

As empresas siderúrgicas e de mineração, assim como as outras empresas dos diversos segmentos, possuem um grande volume de dados de manufatura à sua disposição e com uma forte tendência de aumento desses dados, impulsionados pelos conceitos da Indústria 4.0 e pela evolução tecnológica dos equipamentos e dispositivos utilizados no processo produtivo.
Nesse contexto, é imprescindível pensar em formas diferentes de se consolidar os dados de negócio para transformá-los em informações úteis para os processos, na busca por diferenciais competitivos, uma vez que as abordagens tradicionais podem se mostrar incompatíveis com o volume de dados e a velocidade necessária para o processamento dos mesmos.
O BI 4.0 apresenta-se como uma alternativa disruptiva para o processamento de dados de manufatura através de uma plataforma analítica, escalável e de alta performance, com foco em, mas não se limitando a, soluções de manufatura MES/MOM para processos siderúrgicos e de mineração.
Os resultados apresentados neste trabalho comprovam que o BI 4.0, no contexto da indústria, está à frente de outras abordagens comumente utilizadas, permitindo um desempenho de até 600 vezes melhor e com uma arquitetura escalável que permite o tratamento de volumes de dados muito maiores, permitindo abranger por exemplo diversos sites de uma empresa, apenas incorporando mais processadores e memória ao *cluster*.

REFERÊNCIAS

1 Prentice, S., Jacobson, S. F., and Tratz-ryan, B. (2014). *Industrie 4.0 — The Ten Things the CIO Needs to Know*. (October). 2014.
2 Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. (2010). Spark: cluster computing with working sets. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing,* volume 10, page 10.
3 Chaudhuri, S. and Dayal, U. (1997). An overview of data warehousing and olap technology. *ACM Sigmod record*, 26(1):65–74.
4 Dean, J. and Ghemawat, S. (2004). Mapreduce: Simplified data processing on large clusters. *To appear in OSDI*, page 1.
5 Meyer, H., Fuchs, F., and Thiel, K. (2009). *Manufacturing Execution Systems (MES): Optimal Design, Planning, and Deployment*. McGraw-Hill Education, 1 edition.
6 White, T. (2012). *Hadoop: The definitive guide.* ” O’Reilly Media, Inc.”.