

Jhenny Kelly do C. Dias



Agenda

- Apresentação e introdução
- SEFAZ/ES – do BI ao Big Data Analytics
- Big Data
- Analytics
- Hadoop
- Casos de sucesso
- Conclusão

Apresentação e Introdução

- Analista de tecnologia da informação
 - Prodest/ES, lotada na Sefaz.
- Supervisora da equipe de Business Intelligence e Data Administration da GETEC/SEFAZ
 - Email: jkdias@sefaz.es.gov.br
 - Telefone: 27 3347 5363

SEFAZ/ES – do BI ao Big Data Analytics

- Experiência da Sefaz/ES com BI e Big Data
 - ... Até 2013
 - Data Marts isolados com acesso através de ferramenta limitada;
 - Necessidade de cruzamentos mais avançados;
 - Volume de dados de difícil tratamento.

SEFAZ/ES – do BI ao Big Data Analytics

- Início de estudos de ferramentas de BI;
- Edital para aquisição de ferramentas de BI;
- Solução SAS Enterprise BI como vencedora do certame.

SEFAZ/ES – do BI ao Big Data Analytics

- 2013 até os dias atuais
 - Construção de produtos mais avançados
 - Indicadores de irregularidades fiscais;
 - Inconsistências EFD: Notas fiscais não escrituradas ou escrituradas com irregularidades;
 - Data Mart da Nfe com os produtos;
 - Cruzamentos entre assuntos diferentes
 - Nfe x Arrecadação x EFD;
 - Nfe x Arrecadação x PGDAS;
 - Cartão de Crédito x EFD;
 - Etc.
 - Acesso direto aos dados de documentos fiscais e obrigações acessórias.

SEFAZ/ES – do BI ao Big Data Analytics

- Mesmo com uma nova solução de BI, ainda enfrentamos algumas dificuldades
 - Velocidade de acesso a informação.
 - Janela de carga muito longa
 - Processamento do Data Mart para identificação de empresas laranja somente roda 1 vez por semana!

SEFAZ/ES – do BI ao Big Data Analytics

- Além disso...
 - As origens dos dados do BI é limitada somente aos bancos de dados relacionais: Oracle e SQL Server;
 - Não é possível consumir dados de:
 - Mídias sociais;
 - PDFs;
 - Emails;
 - Xml.
 - Carga de dados no mínimo D+1.

SEFAZ/ES – do BI ao Big Data Analytics

- 2017 até os dias atuais
 - Criação de um grupo de trabalho para estudos de novas tecnologias analíticas: Big Data Analytics;
 - Objetivo de trazer mais modernidade para o processo de apoio a decisão da Sefaz/ES;
 - Realização de diversas provas de conceitos;
 - Escrita de um termo de referência
 - **Projeto Big Data Analytics.**

[illegible]

Big data

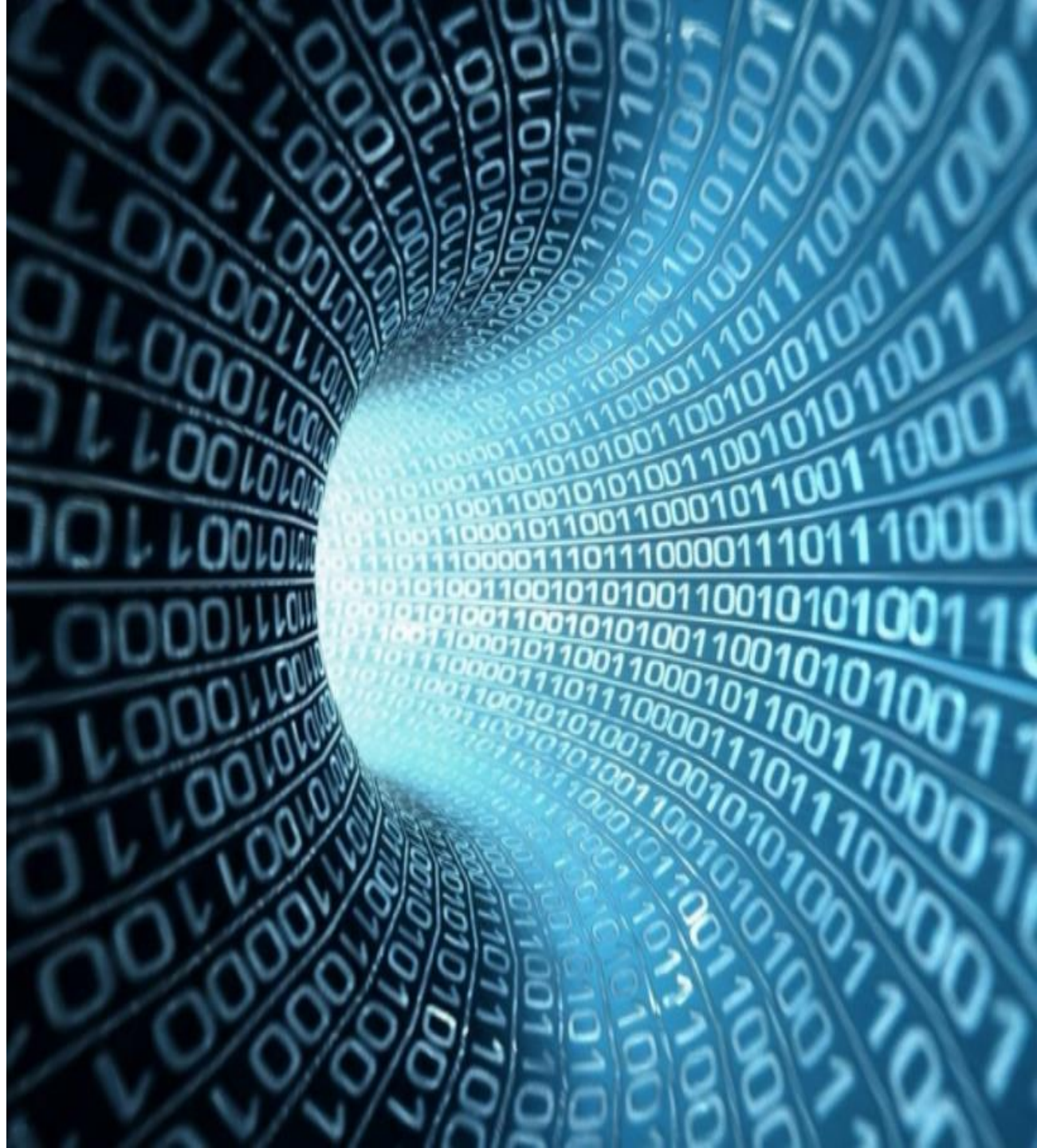
Mas o que é Big Data?

O que é um grande volume de dados?

Gigabytes?

Terabytes?

Petabytes?



Big data

- Para entender melhor, vamos nos perguntar:
 - A nossa organização consegue processar todo volume de dados das aplicações e transações do dia a dia?
 - O tempo de acesso a esses dados está atendendo às necessidades de negócios?
 - Todos os tipo de dados que a organização recebe e trafega são atendidos pelas ferramentas atuais?
 - Existe confiança nos dados apresentados nas soluções de apoio a decisão atuais?
 - Qual a importância dos dados da organização? Possuem valor agregado e podem ser tratados de forma a retornar o investimento das soluções atuais?
- Assim, chegamos aos 5 Vs do Big Data...

Big data



Big data

Volume

Atualmente produzimos mais dados por dia do que se produziu em todos os tempo até alguns poucos anos atrás. Assim, torna-se necessário tratar esse grande volume de forma diferenciada do que a forma atual. Bancos de dados relacionais e modelos ROLAP não suportam mais esses grandes volumes de forma satisfatória.

Big data

Velocidade

Analisar dados históricos não é mais suficiente para alguns tipos de tomadas de decisão. As fraudes ocorrem a todo momento, quanto mais rápido a Sefaz conseguir identificar as fraudes praticadas por empresas laranja, menor será a perda para a administração pública. Analisar dados em tempo real já é uma realidade.

Big data

Variedade

Os tipos de informações a serem analisadas em processos decisórios ou de identificação de fraudes não se limita mais somente aos dados históricos oriundos de bancos de dados relacionais, é preciso considerar os dados não estruturados originários de mídias sociais , emails, pdfs, documentos eletrônicos, planilhas, etc.

Big data

Veracidade

Dados devem ser autênticos e devem fazer sentido no contexto de sua análise.

Big data

Valor

É necessário que a implementação de um projeto dessa natureza retorne o investimento realizado, ou seja, informação tem valor e esse valor deve saltar aos olhos em retorno de um projeto Big Data.

Big data

Ou seja...

Soluções Big Data garantem que o imenso volume de dados gerados a cada dia possam ser tratados de forma mais rápida, flexível e confiável, uma vez que o volume de dados excede a capacidade das tecnologias atuais de os tratarem adequadamente.



[illegible]

Analytics

E o Analytics?

- O que queremos como todos os recursos proporcionados pelo Big Data?
- Quais problemas precisamos resolver?
- Quais fraudes insistem em aterrorizar a nossa organização?

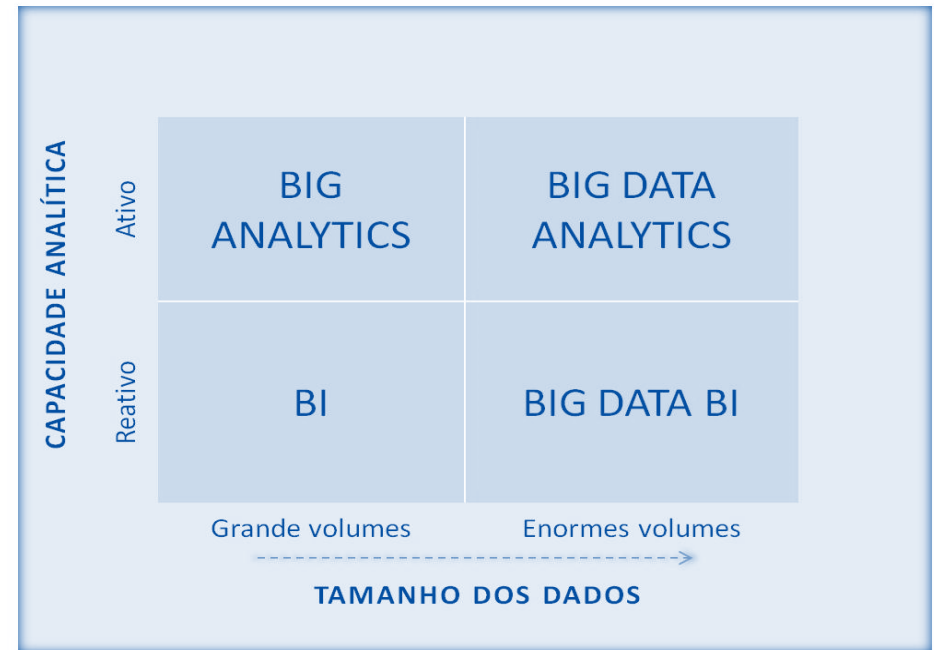
Analytics

O Big Data Analytics permite que sejam desenvolvidas soluções de negócios que resolvam os problemas mais críticos da organização, no caso da Sefaz

- detecção de indícios de emissões de notas fiscais por empresas laranjas;
- identificação de redes sociais de atores envolvidos em processos de fraudes.

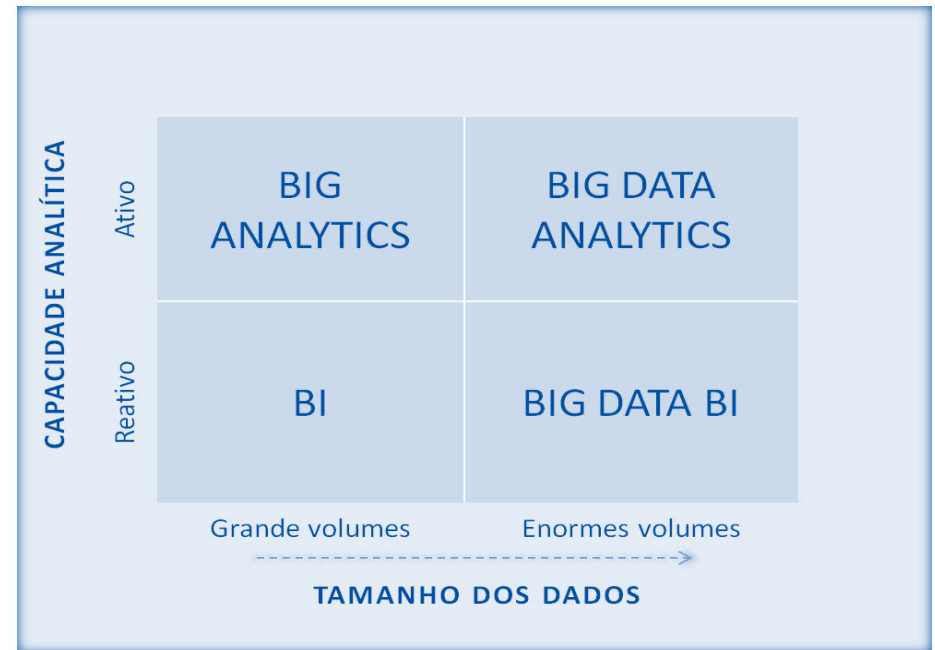
Analytics

Tratamento reativo de grandes volumes de dados pode ser traduzido em BI, ou seja, análise histórica com decisões de negócios baseadas em ocorrências do passado.



Analytics

Tratamento reativo de enormes volumes de dados pode ser traduzido em Big Data BI, ou seja, podemos implantar uma solução Big Data (5 Vs) e ainda continuar atuando em ocorrências do passado.

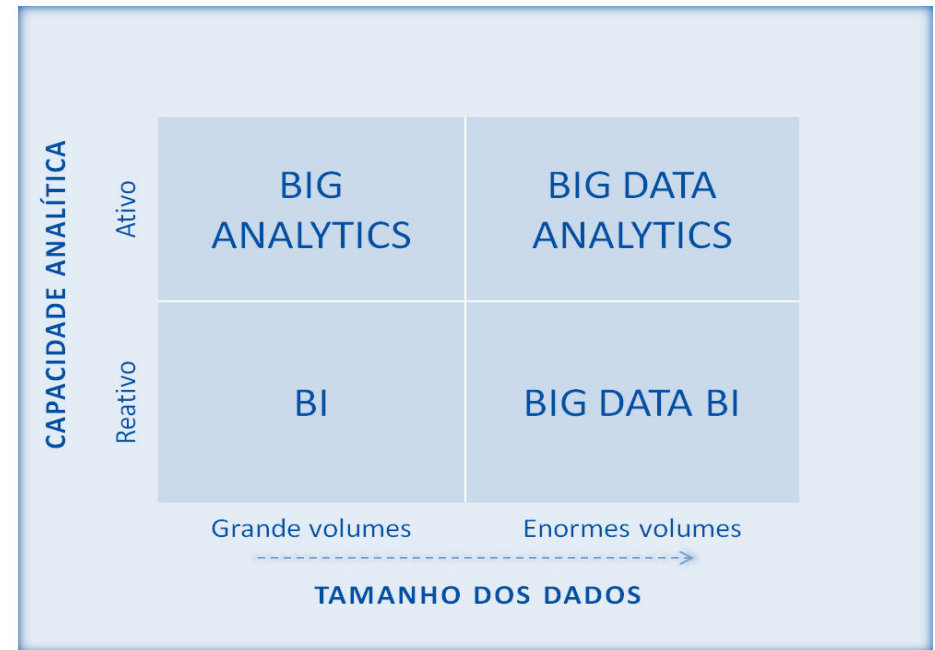


Analytics

Em outra direção, temos a possibilidade de tratar de forma ativa os grandes volumes de dados, ou seja, a organização se antecipa ao cenário que se quer monitorar;

Podemos traduzir essa capacidade analítica em Big Analytics;

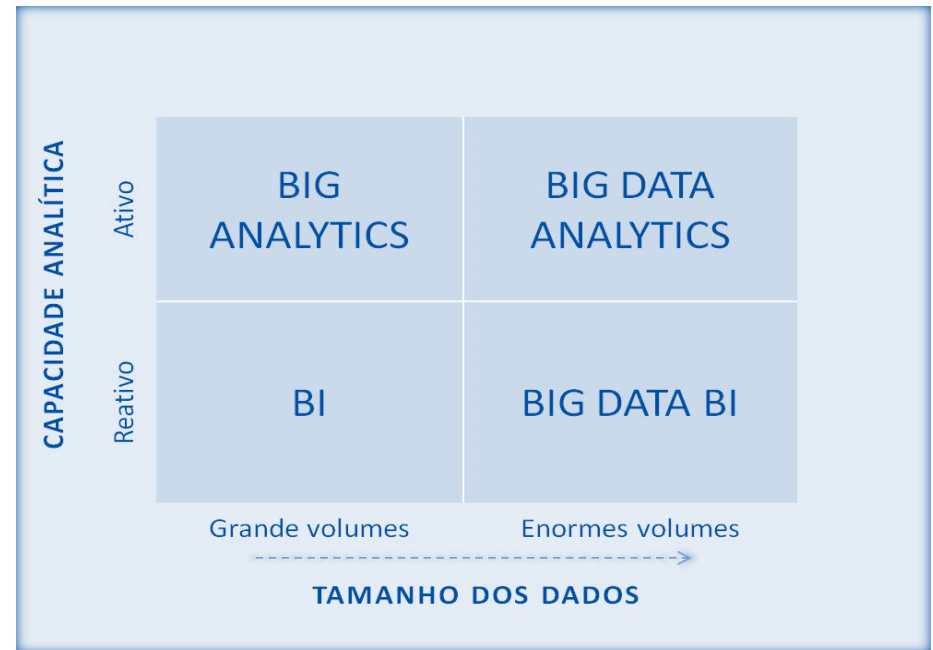
Nesse cenário, temos ambientes tradicionais aliados com ferramentas analíticas, como Tableau, Qlik Sense, SAS Visual Analytics, Data Mining.



Analytics

Por fim, temos a possibilidade de combinar um ambiente Big Data (5 Vs) com ferramentas Analytics, onde a capacidade analítica se dá de forma ativa em um enorme volume de dados;

Nesse cenário, temos ambientes Big Data aliados com ferramentas analíticas, que o cenário mais completo para uma organização que tem necessidade de se antecipar às ocorrências fraudulentas.



Big Data Analytics

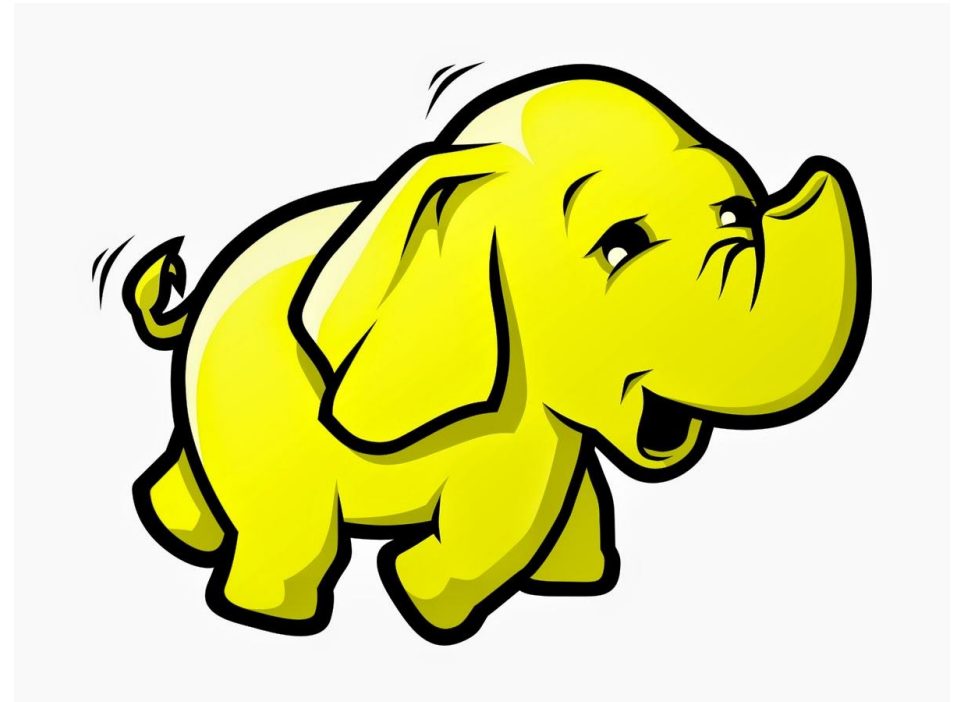
- É importante observar que o Big Data Analytics não significa eliminar os tradicionais sistemas de BI existentes;
- Pelo contrário, eles devem coexistir.



Hadoop

Hadoop

O Hadoop é um projeto de código aberto que implementa uma plataforma de computação distribuída que oferece armazenamento de qualquer tipo de dado, a um custo baixo, em grande escala e permite a possibilidade de análises de dados complexos rapidamente.



Hadoop

- Distribuições Hadoop do mercado:
 - Cloudera
 - Hortonworks
 - MapR

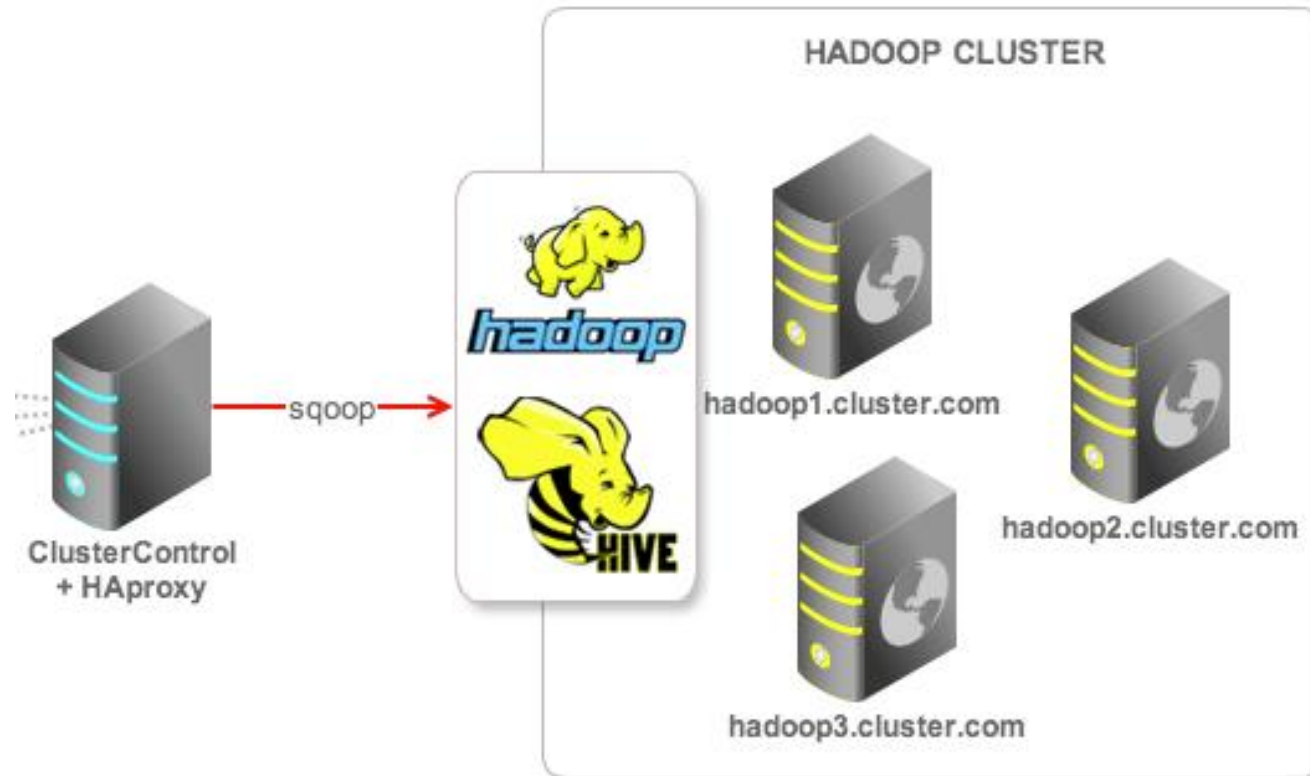


Hadoop

Mas como é possível o tratamento de grandes volumes de dados variados de forma rápida?

Resposta: devido a arquitetura de hardware e software proposta pelas soluções Hadoop, cujo hardware possibilita o uso de equipamentos padrões ligados em clusters ao invés de um super hardware¹.

As máquinas utilizadas em um cluster Hadoop trabalham em conjunto para atender aos 5 Vs do Big Data.



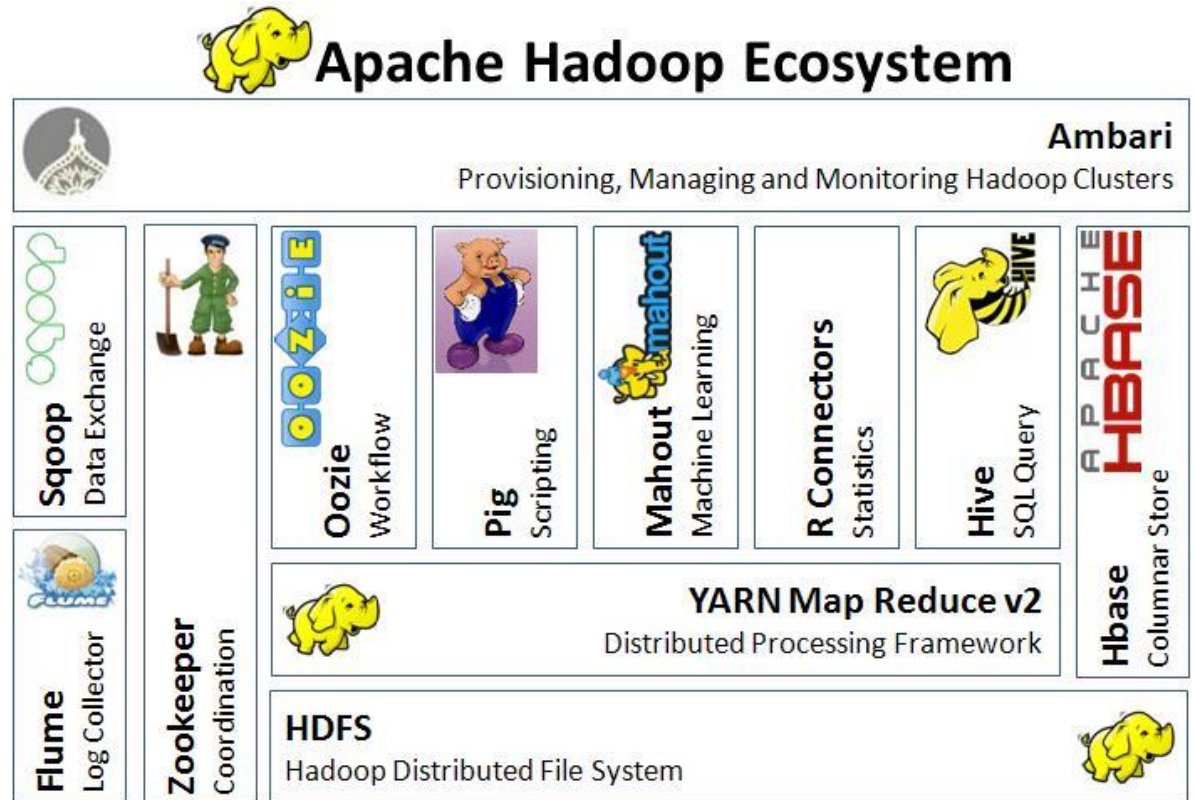
1: Cluster é o nome dado ao conjunto de computadores que trabalham de forma sincronizada para funcionar como um único computador. São diversas máquinas que se comportam como uma única máquina.

Hadoop

Devemos entender também a orquestração entre o hardware e software no Hadoop, que é o que permite os ganhos de uma solução desse tipo.

No cluster Hadoop, algumas ferramentas que compõem o ecossistema estão representadas na figura.

Não falaremos de todas, mas é importante sabermos um pouco sobre HDFS e Map Reduce.

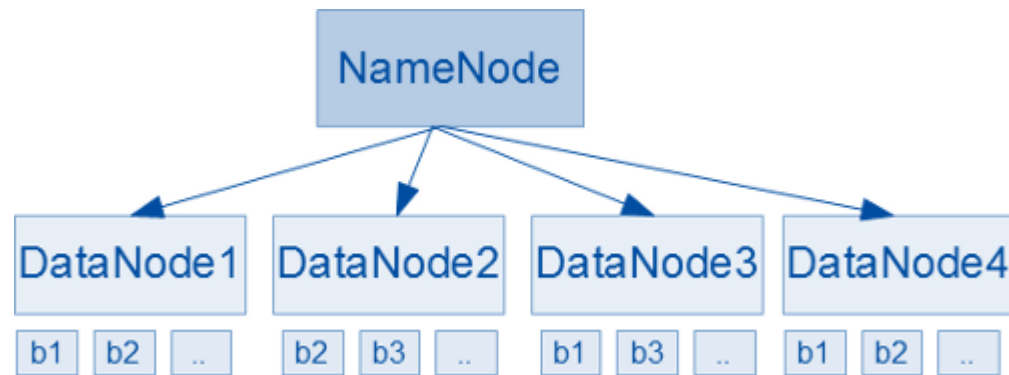


HDFS

Nos últimos anos a capacidade de armazenamento dos discos rígidos aumentou muito, mas a velocidade de leitura e gravação nesses discos não acompanhou o ritmo. Assim, uma solução para resolver esse problema é ler/escrever os dados em paralelo, utilizando vários discos.

Assim, chegamos ao HDFS (Hadoop Distributed File System).

O HDFS é um sistema de arquivos² distribuído, responsável pelo armazenamento das grandes quantidades de dados do Hadoop. O HDFS também permite a conexão entre os nós do cluster Hadoop, onde os blocos de dados são distribuídos.



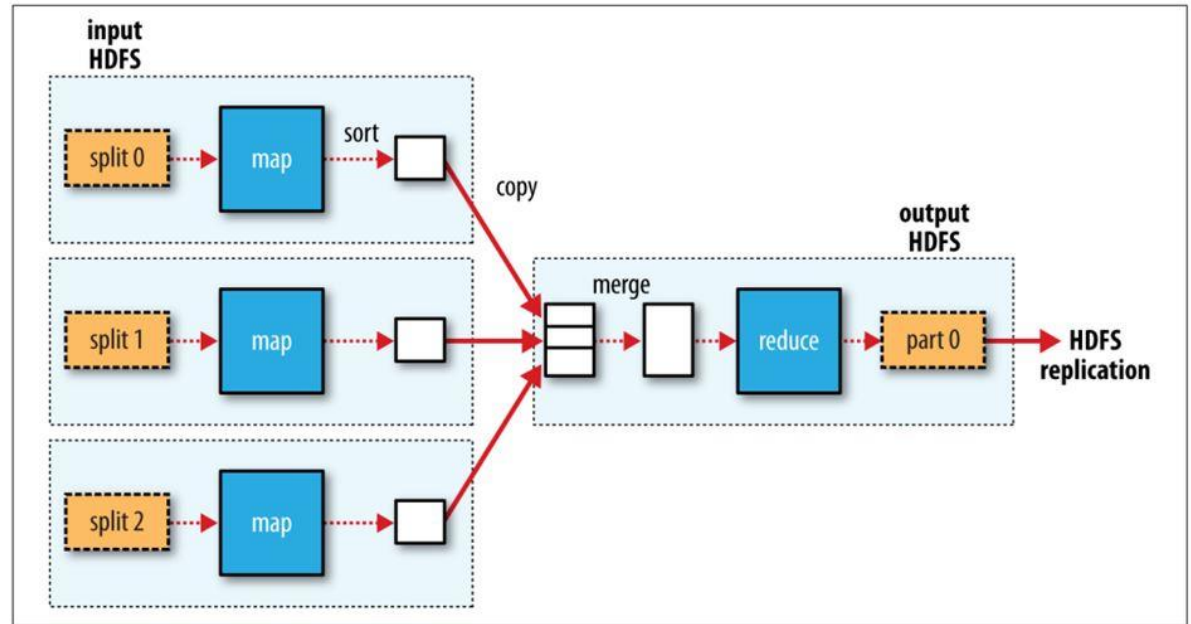
2: Um sistema de arquivos é um conjunto de estruturas lógicas e de rotinas, que permitem ao sistema operacional controlar o acesso ao disco rígido.

Map Reduce

O Map Reduce é responsável pelo processamento de dados no Hadoop, sendo a solução para o processamento paralelo de dados.

Não é preciso realizar nenhum tipo de programação extra para garantir que os processos serão processados paralelamente, o que garante a performance para os cenários de análise de dados em larga escala.

Existem várias outras ferramentas que compõem o ecossistema Hadoop, mas não falaremos de todas, vamos a alguns casos de sucesso.



Casos de sucesso

- O Ministério da Justiça do Brasil possui um Data Warehouse com mais de 1 bilhão de registros, além de um poderoso supercomputador da IBM, capaz de coletar, agrupar e processar petabytes de dados em frações de segundos.
- O objetivo do Ministério com o uso de Big Data é identificar indícios de ações ilícitas, sobretudo ligadas à lavagem de dinheiro.

Fonte: <http://www.bigdatabusiness.com.br>

Casos de sucesso

- A Nike, maior fabricante de materiais esportivos do planeta, fechou uma parceria com uma empresa especialista em tecnologia, no intuito de desenvolver um software que fosse usado por praticantes de running, informando a eles frequência de batimentos cardíacos, velocidade, quantidade de passos dados, distância percorrida e muitos outros dados. Os dados são integrados as redes sociais, estimulando amigos e fãs de corrida a compartilharem suas informações, propiciando uma certa “competição” entre os atletas.
- O estímulo a esse comportamento multiplicou exponencialmente a quantidade de dados gerados diariamente, os quais são usados pela Nike para compreender seu público, melhorar seus produtos ou desenvolver novos modelos de tênis e roupas esportivas.

Fonte: <http://www.bigdatabusiness.com.br>

Casos de sucesso

A Danone norte-americana tinha a necessidade de melhorar sua cadeia logística na distribuição de um certo tipo de iogurte (o 'Grego', no caso).

Como o produto perecia com rapidez, ele precisava ser produzido e entregue ao varejo em sincronia quase matemática.

Isso se tornou possível apenas com o cruzamento de informações de rotas, tempo de entrega e prazo de validade.

Fonte: <http://www.bigdatabusiness.com.br>

Casos de sucesso

A Rolls-Royce vem implementando soluções de Big Data para otimizar seus processos de manutenção. Os motores e sistemas de propulsão dos veículos da companhia têm centenas de sensores embutidos, que registram todos os detalhes de funcionamento das máquinas e enviam, aos engenheiros da marca dados em tempo real sobre quaisquer mudanças de desempenho delas.

A equipe de engenheiros cruza e analisa estes dados, possibilitando tomadas de decisão rápidas e assertivas para evitar e/ou resolver qualquer problema de funcionamento dos carros.

Segundo a Rolls-Royce, a adoção de tecnologias de Big Data para diagnosticar falhas, corrigi-las e evitá-las de ocorrerem novamente diminuiu significativamente os custos de manutenção de seus veículos.

Fonte: <http://www.bigdatabusiness.com.br>

Casos de sucesso

O Pinterest, rede social de compartilhamento de imagens, chegou recentemente a mais de 150 milhões de usuários ativos mensais.

Um de seus trunfos para o sucesso é a utilização de Big Data e machine learning para personalizar ao máximo a experiência de seus usuários.

A equipe técnica do Pinterest vem refinando o sistema de recomendação da rede ano após ano.

Para alcançar este objetivo, seus cientistas de dados têm implementado inteligência artificial, um sistema de ranking mais rápido e conteúdo localizado.

Fonte: <http://www.bigdatabusiness.com.br>

Conclusão

Big Data já é realidade;

Análise de dados históricos de forma reativa não é mais suficiente para as organizações;

Análise preditiva visa garantir maiores ganhos nos negócios e no caso da Sefaz, a prevenção a fraudes fiscais;

Advento do conceito de fast data, ou seja, análise de dados em tempo REAL;

Nascimento de uma nova carreira: Data Scientist;



Obrigada

Jhenny K. C. Dias

Email: jkdias@sefaz.es.gov.br

Telefone: 27 3347 5363

