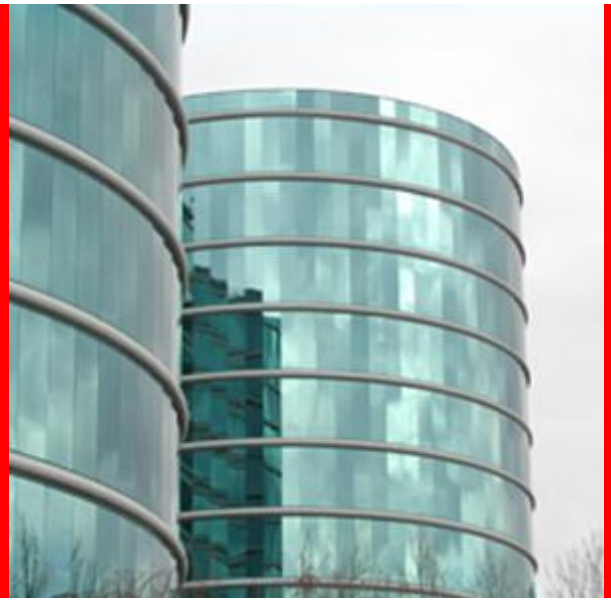


ORACLE

MySQL

MySQL e o Big Data

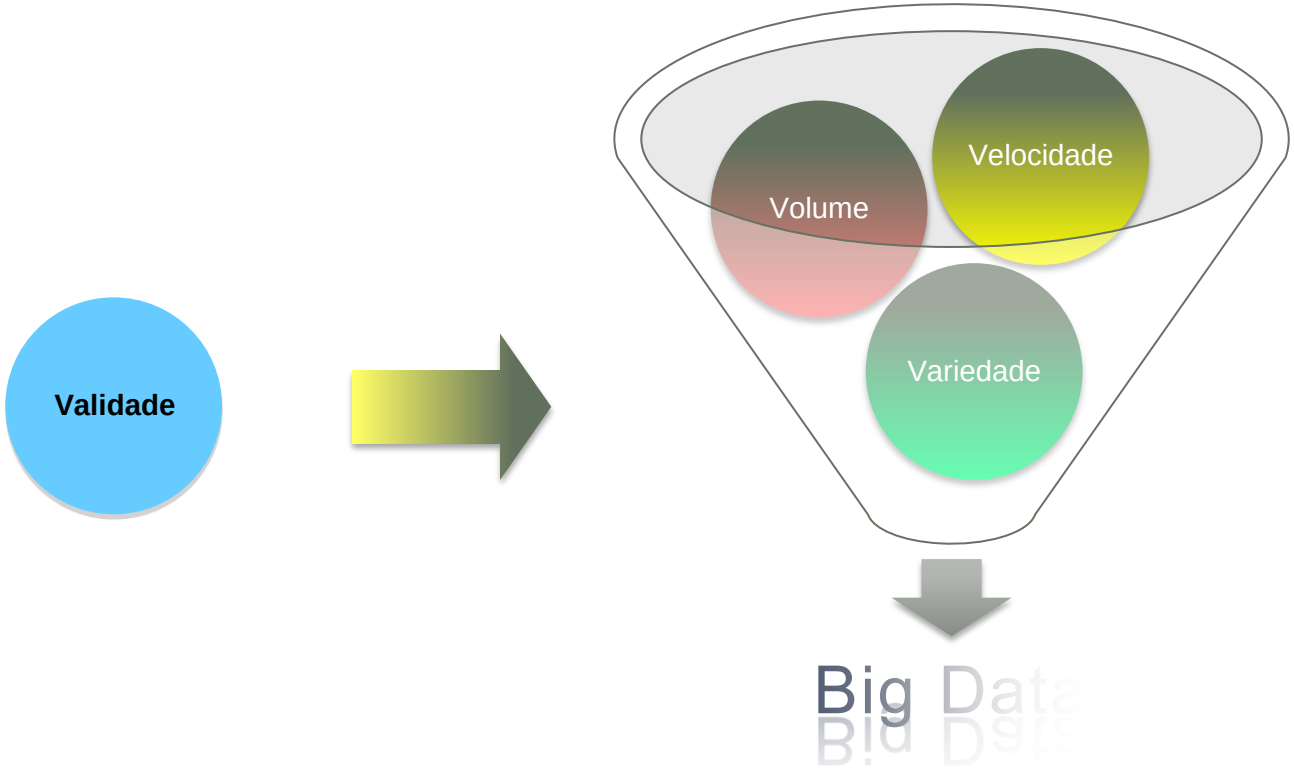
Alexandre M de Almeida
Consultor Senior em Banco de Dados

O que é Big Data?

- Big → Grande
- Data → Dados

O que é Big Data?

- **Quão novo é o conceito?**
- **2001 (3V) de Doug Laney** **(Meta Group/Gartner)**



O que é Big Data?

“*Big Data é um modelo para administrar enormes quantidades de dados (acima da capacidade dos RDBMS atuais). Dados estes, normalmente, não estruturados, de diversas fontes e, em sua, maioria não tratados (sem confiabilidade), e, que apesar de seu alto volume possa ser manipulado em tempo razoavelmente baixo*. Big Data utiliza-se do alto volume e diversidade como forma de, estatisticamente, sobrepujar a baixa qualidade dos dados”

“Perceber, capturar, armazenar e analisar. Este é o objetivo das tecnologia de Big Data”

Computação Big Data: Criando Avanços Revolucionários para o comércio, ciência e sociedade – CCC Computing Community Consortium, 2008

O que é Big Data?

- Grande ***volume*** de dados (10TB+)
  - Bases menores podem ser melhor endereçadas por RDBMS
- Grande ***variedade*** de dados
  - Maior quantidade para mitigar a má qualidade
  - Não estruturados (e/ou estruturados)
  - Formatos e fontes diferentes
- Grande ***velocidade***
  - Processar enormes quantidades de dados em menor tempo
  - Processamento distribuído, paralelismo, + caixas + velocidade
- Medir e mensurar tudo
  - Geração de conhecimento sobre dados difusos
- Essencialmente Analítico
  - O objetivo primeiro, apesar de distorções, é OLAP!

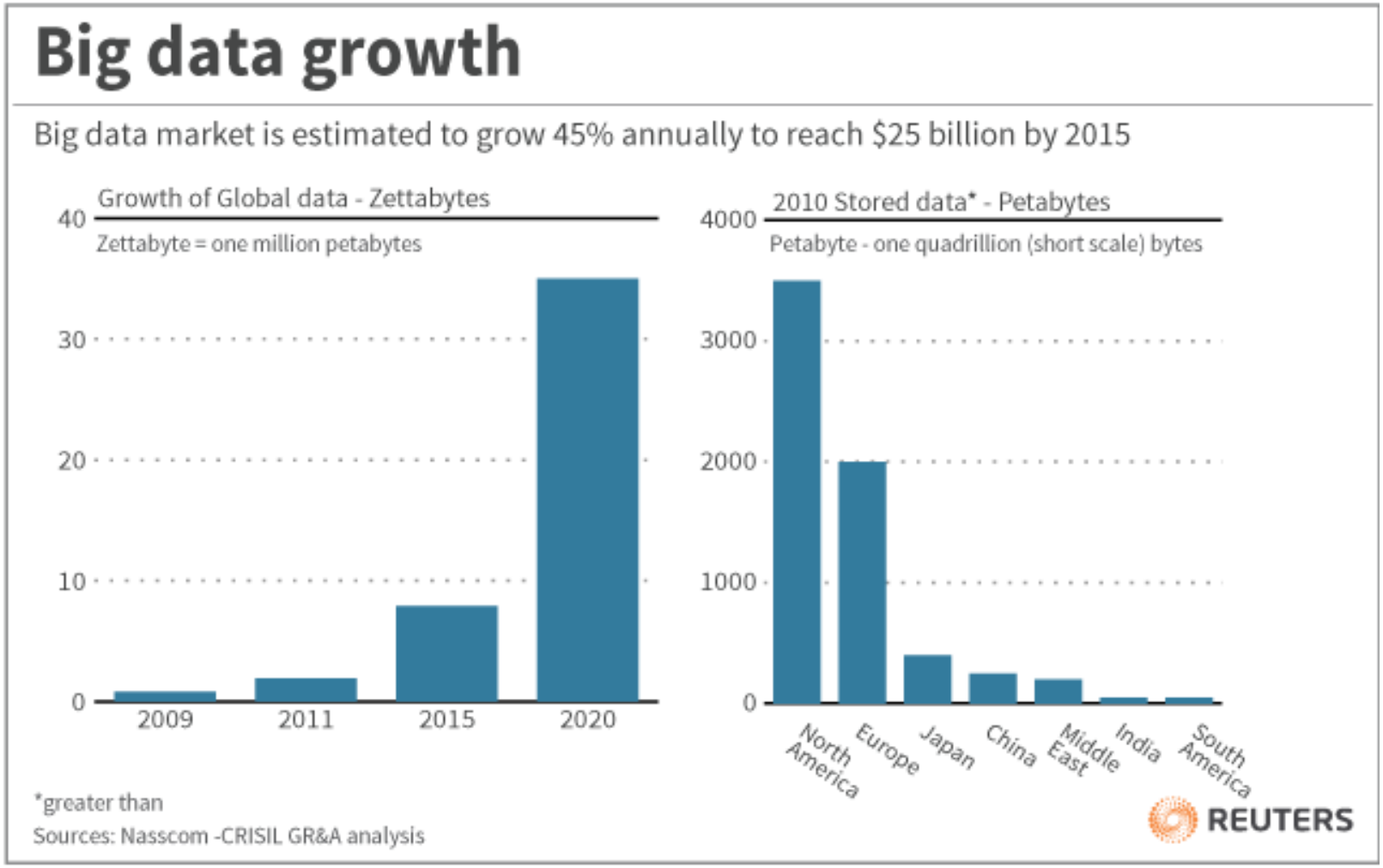
Qual a origem do Big Data?

- 1944 – Arthur Freemont Rider
  - Conteúdo universitário dobra a cada 16 anos
- 1961 – Derek Price
  - Publicações científicas dobram a cada 15 anos
- 1964 – Harry J. Gray e Henry Ruston
  - Publicam artigo sobre a "***explosão dos dados***"
- 1967 – B.A. Marron
  - "Compressão de dados automática"
- 1971 – Arthur Miller
  - "O homem será medido pela quantidade de bits que o descrevem"
- 1975 – Japão conduz censo sobre "Fluxo da Informação"
  - Introduz a contagem de palavras para medir volume de dados
- 1980 – I.A. Tjomsland
  - "Para onde vamos agora?" sobre como armazenar os dados produzidos
- 1981 – Centro de Estatística da Hungria
  - Pesquisa sobre o volume de dados do país e como gerenciá-lo

Qual a motivação para o Big Data?

- Vivemos em novo mundo orientado à “informação”
- Sensores, robôs, rastreadores, capturam toneladas de bytes (e até o Obama)
- Geramos informações tão difusas quanto complexas
- Analisar estas informações de várias formas
- Em tempo que não afete nossa ansiedade
- Produzir conhecimento em qualquer área

Crescimento de dados e impacto

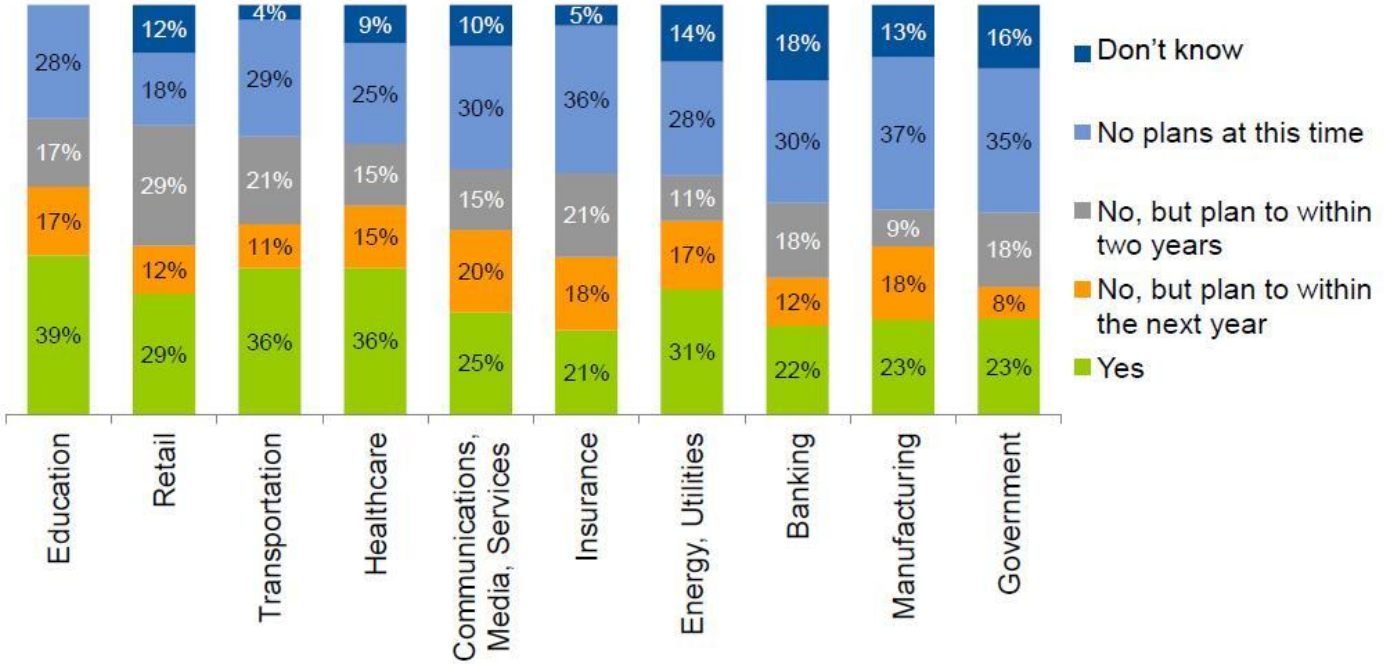


Reuters graphic/Catherine Trevethan 05/10/12

Crescimento de dados e impacto

Big Data Investments by Industry

Has your organization already invested in technology specifically designed to address the big data challenge?



Source: Gartner (July 2012)

14

Gartner

ORACLE

Crescimento de dados e impacto

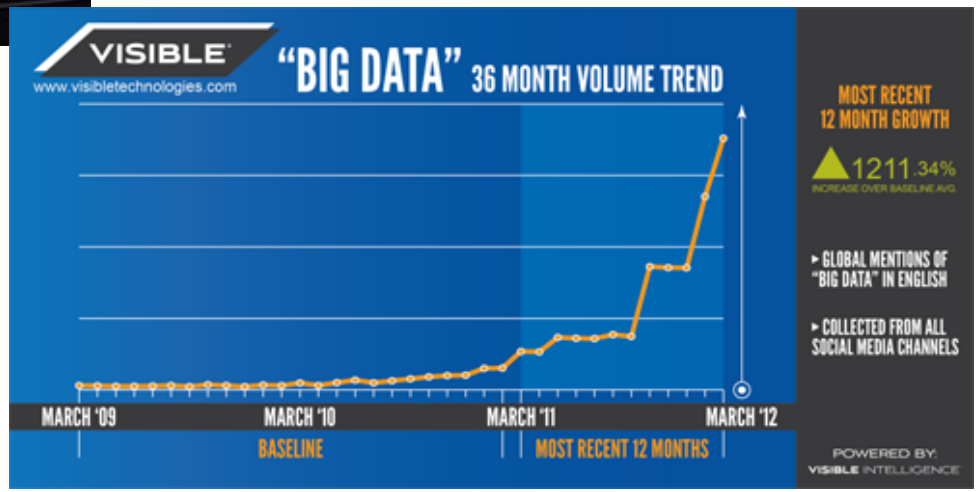
Os dados são produzidos a todo instante, por toda sorte de fonte.

Como vamos coletá-los?

Como vamos armazena-los?

Como vamos recupera-los e interpreta-los em tempo razoável?

Como vamos produzir conhecimento corporativo, científico ou governamental em prol do lucro ou da humanidade?



Como o Big Data afeta pessoas e empresas

- Pessoas
  - Quem não faz parte de algum tipo de rede social?
  - O Big Data produz um dossiê vivo, online
  - Pessoas se conectam à pessoas
  - Interagem melhor com o mundo
- Empresas
  - Entendem melhor seu consumidor, competidores e mercado
  - Investem melhor em áreas conhecidas e desconhecidas
- Governo
  - Conhece melhor seu povo, costumes, necessidades
  - Prevê tragédias e antecipa-se à necessidades futuras
- Ciência
  - Simulações de todo tipo com “zilhões” de combinações (dados)
  - Mapeia o universo externo e o nosso universo interno (genética)
  - Construir máquinas, equipamentos, estruturas virtuais e testá-las

Big Data e o RDBMS

“Nenhum de nós nesta sala irá ver o fim dos RBMS, e sim, seu melhor uso ao estilo OLTP, e como fonte de apoio, indissolúvel, ao Big Data. O Big Data vem para dar uma nova definição de mundo, uma maneira, completamente, nova de explorar o universo dos bytes, produzir conhecimento. O Big Data e sua capacidade de analisar milhares de terabytes será uma contribuição fundamental para a explosão do conhecimento”

Exemplos de uso do Big Data

- Wal-Mart
  - Com sua base de mais de 10 petabyes (10.000 trilhões de bytes) representados por mais de 350 milhões de transações diárias de 6.000 lojas espalhadas pelo mundo, cria estratégias de preços e logística, altamente, eficientes
- LSST (Large Synopit Survey Telescope)
  - Instalado no topo de uma montanha no Chile escaneia o céu e produz 30 trilhões de bytes em imagem por dia! Estes dados serão analisados para conhecer o universo.
- LHC (Large Hadron Collider)
  - O super acelerador de partículas produz 60 terabytes de dados por dia (15 petabytes) por ano. O estudo destes dados possibilitará enormes avanços na ciência.

Exemplos de uso do Big Data

- eCommerce
  - Coletar cada click, cada visualização, cada navegação (tags internas e tags de parceiros):
    - Criar campanhas orientadas para regiões, grupos, e indivíduos
    - Identificar oportunidades (ABC/Custo Oportunidade)
    - Melhorar capacidade logística
- Seguros
  - Coletar dados de diversas fontes: bases internas, segurança pública, redes sociais, climáticas, etc:
    - Análise de risco que afetam prêmio, franquia e lucro
    - Entender as necessidades de proteção (novos produtos)
- Portais
  - Analisar o perfil do internauta para entregar publicidade com alto grau de assertividade
  - Redes sociais podem traçar perfis e sugerir (e encontrar) amigos
- Governo
  - Engenharia de Tráfego: coletar, através de sensores, quantidade de veículos, rotas, saturação e criar semáforos inteligentes
  - Através de sensores metereológicos fazer previsões climáticas que possam ajudar os diversos setores que são impactados por mudanças climáticas, e, prever tragédias

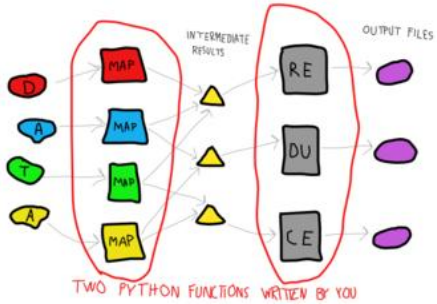
Exemplos de uso do Big Data

- Segurança Pública
  - Capturar emails, twitters, blogs, ligações telefônicas, redes sociais (Volume e Variedade de dados não estruturados) e prever ataques terroristas, manifestações (***), identificar e capturar criminosos
- Supermercados
  - Através de PDV's ou RFID's coletar e analisar informações sobre consumo de cada indivíduo ou comunidade para ofertar produtos mais adequados, gerando maior valor agregado e, evitando, perdas com produtos pouco atrativos
- Infraestrutura e Segurança de Redes
  - Analisar navegação e emails de funcionários para prevenir roubos e fraudes
  - Analisar milhões de linhas de logs de servidores para identificar e prevenir problemas

Quem é quem no cenário de Big Data?

- ***Apache Hadoop***
- Google BigQuery
- HPPC Systems
- Twitter/Backtype Storm
- Amazon Redshift
- Nokia Disco
- Berkeley Spark
- GraphLab
- E mais outros 123 players
  - Variantes do Hadoop
  - Variantes do MapReduce
  - Linguagens
  - Indexadores
  - Frameworks

Google



HPCC Systems

Quem é quem no cenário de Big Data?

- Alternativas interessantes:
  - MongoDB
  - ***Cassandra***
  - RavenDB
  - Microsoft Azure Table Storage
  - DynamoDB

Alternativas tecnológicas

- Existem várias, cada qual com suas virtudes e vicissitudes

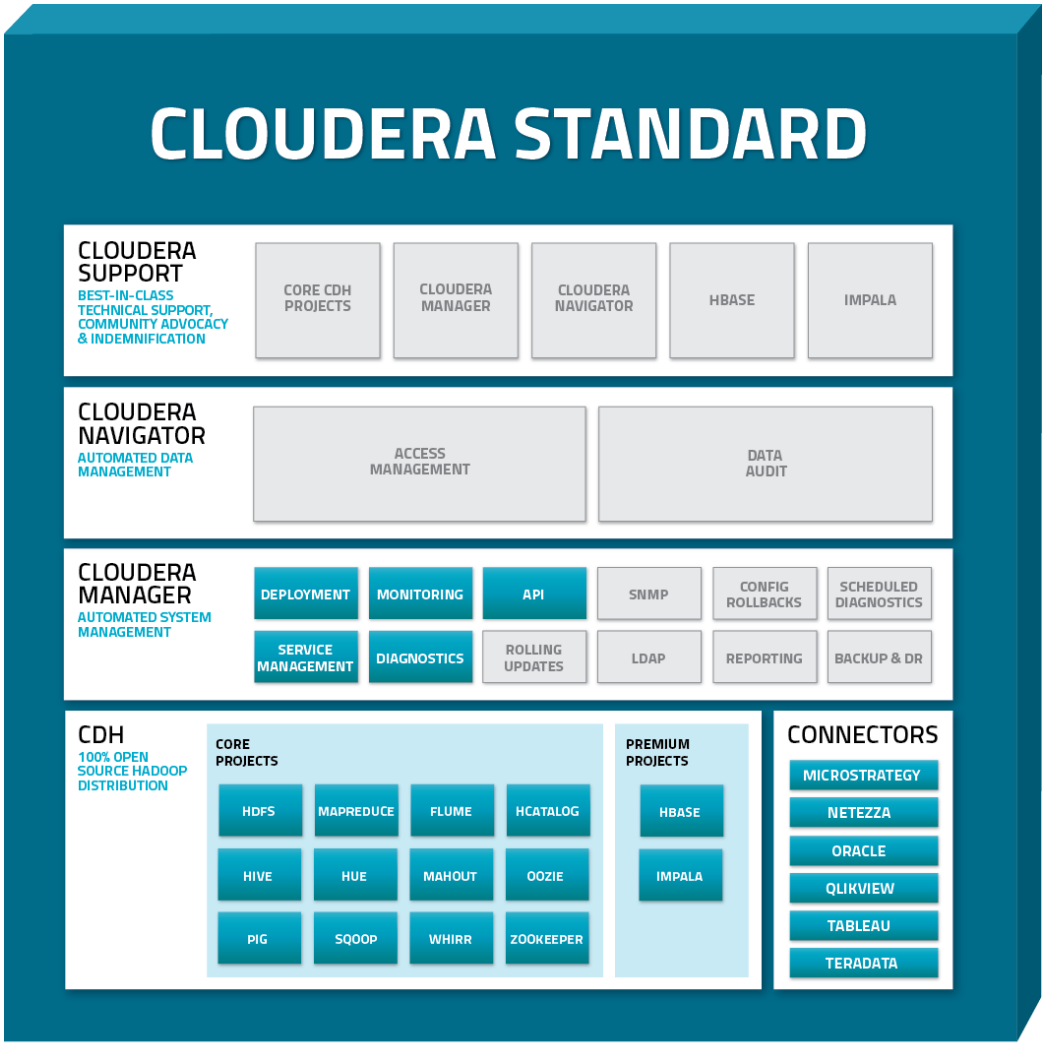
- Para escolher é preciso saber:
  - Tamanho mínimo e máximo do dataset
  - Latência
  - Sustentação (instalação, gerenciamento, disponibilidade, etc)
  - Acessibilidade (como os dados são acessados, linguagens)
  - Consultas em tempo real, consultas “Ad hoc”, etc

- Alternativas, realmente, viáveis:
  - ***CDH – Cloudera Distribution of Hadoop***
  - Google BigQuery
  - HPCC (High Performance Computing Cluster)
  - Storm + Cassandra

Olha o passo do elefantinho…

- **Hadoop**
  - Entre 80% e 90% das implementações de Big Data
  - Hadoop tem se provado como melhor alternativa
  - Vários casos de sucesso (*NOAA, NASA, FBI, CIA, Wal-Mart, Expedia, Groupon, eBay, Yahoo!, Google, Facebook, Linkedin, Mercado Livre, Ning, Rakuten, Telefonica, Twitter*)
  - CDH da Cloudera é a melhor distribuição:
    - Fácil de instalar e implementar (NNF)
    - Gerenciador de Cluster
    - Melhor conjunto de complementos:
    - Hadoop, Flume, Hive, Mahout, Oozie, Pig, Sqoop, Whirr, Zoekeeper, Hue, Hbase e Impala
    - Free!!!

CDH – Cloudera Distribution of Hadoop



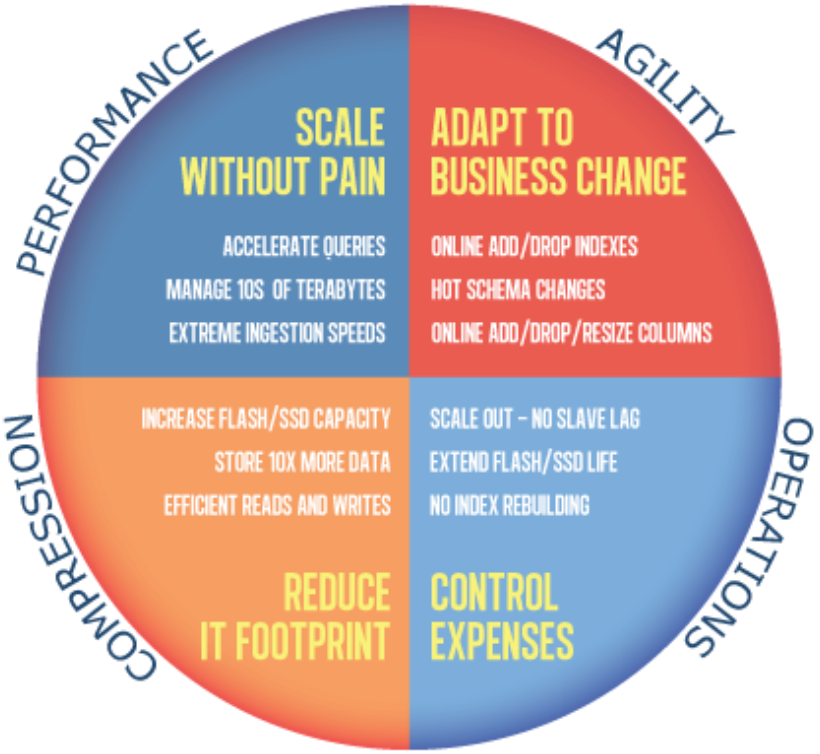
CLOUDERA STANDARD

MySQL e o Big Data

- 5 Terabytes é Big Data?
  - Maioria das implementações c/ MySQL: 150GB e 500GB
  - É possível, somente, com MySQL?
  - Técnicas:
    - Particionamento
    - Sharding
    - Replicação, balanceamento, clusterização
    - Covering Index
    - MyISAM para OLAP
    - InnoDB para OLTP
    - E lógico… Hardware adequado!

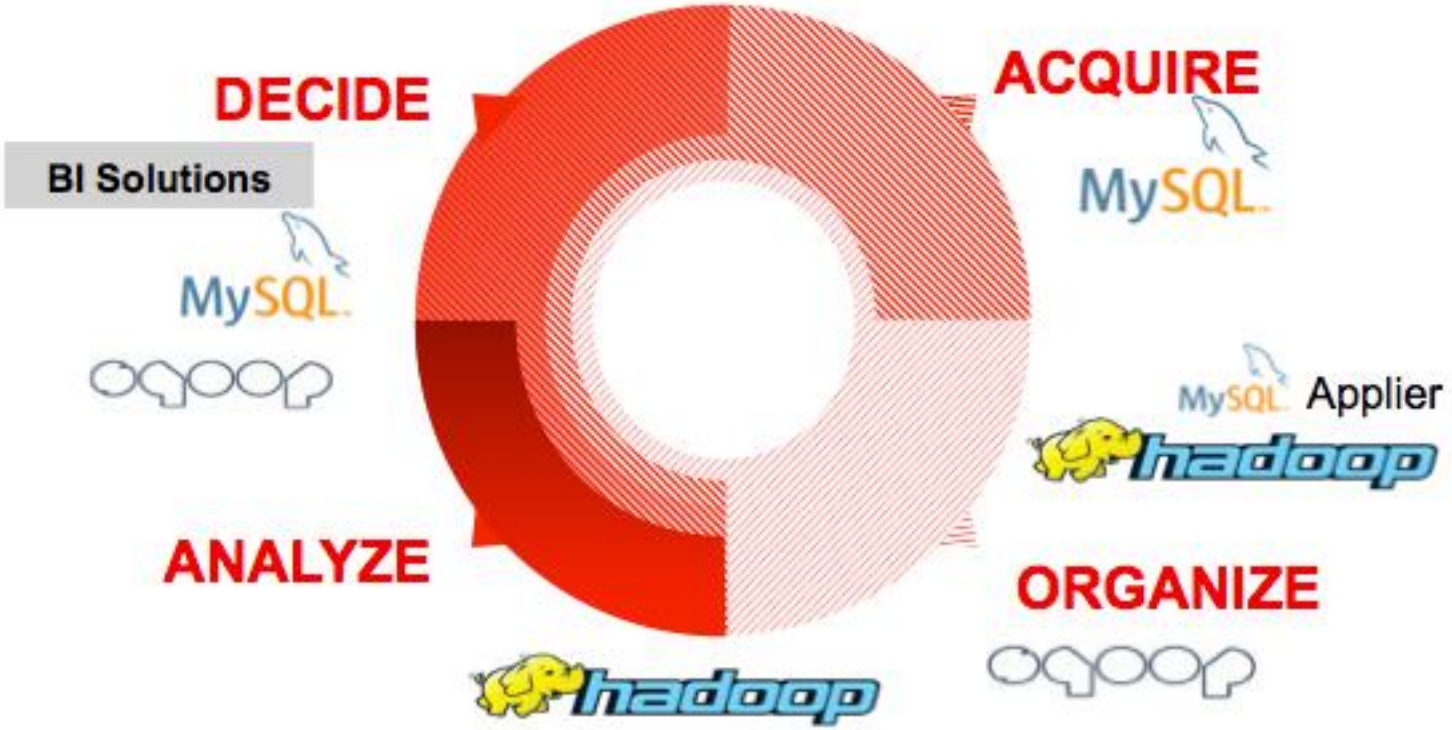
MySQL e o Big Data

- 10 Terabytes é Big Data?
- TokuDB
  - Storage engine by Totutek
  - Transacional (ACID e MVCC)
  - Índice fractal ao invés de B-tree
  - +10,000 inserts/segundo
- InfiniDB
  - Storage engine by CalPont
  - MPP – Massive Parallel Processing
  - Particionamento automático
  - Altamente escalável para múltiplos nós
  - Suporte transacional (ACID e MVCC)
- Resolvem problemas típicos:
  - Performance
  - Load
  - Manutenção



MySQL e o Big Data

- Muitos terabytes é Big Data!
- Integrar o MySQL com o Hadoop é opção PREMIUM



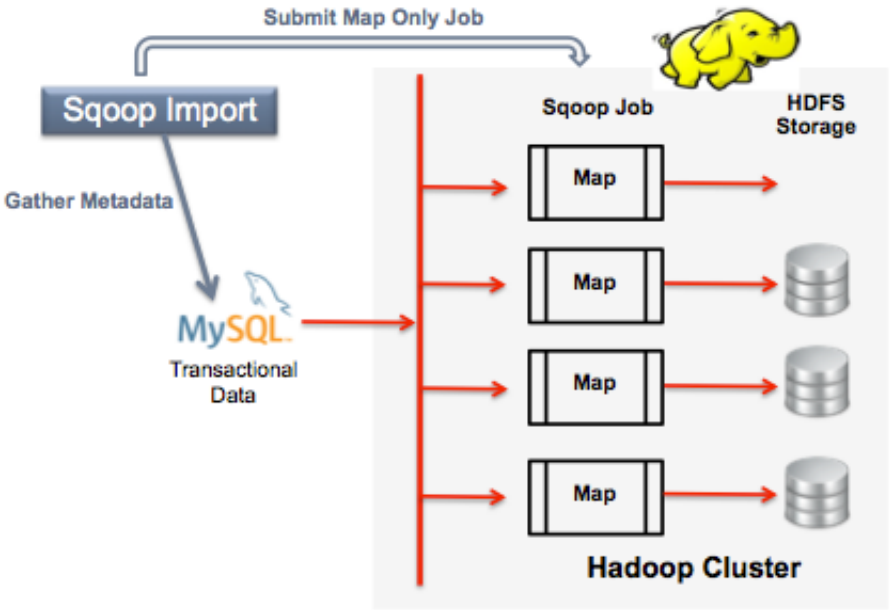
MySQL e o Big Data

- Aquisição de Dados
  - MySQL 5.6 inclui novas características de NoSQL (memcache)
  - Grande capacidade de cargas unitárias (inserts) e/ou batch (loads)
  - ACID e MVCC
  - Sem necessidade de storage engine de terceiros
  - Permite fazer um pré-processamento antes do Hadoop
  - Pré-análise dos dados recém “adquiridos”
  - Aplicar PCI / HIPAA / Sarbanes

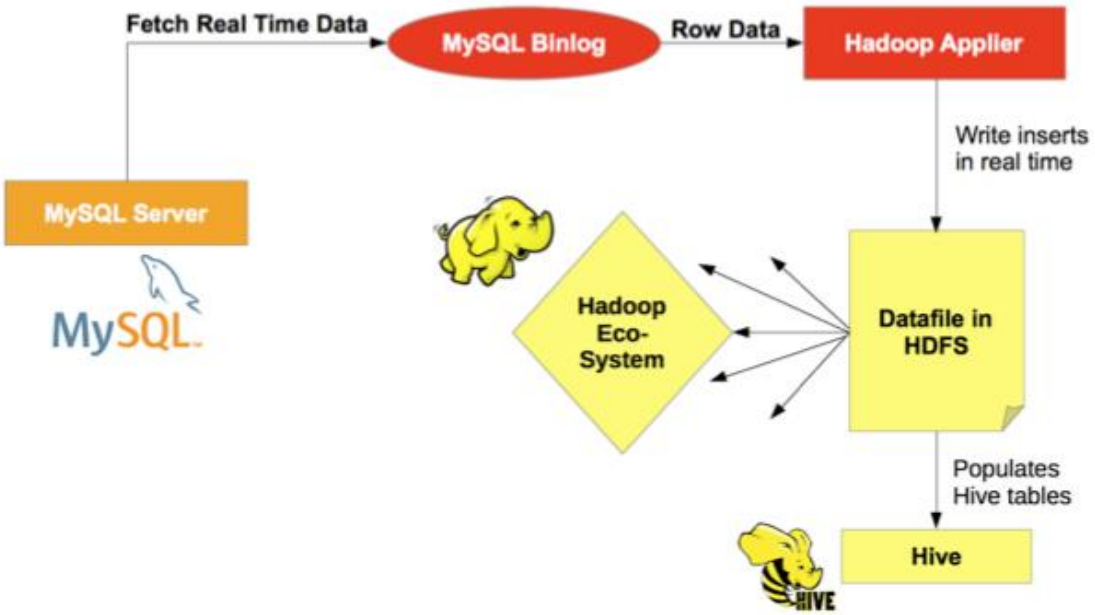
MySQL e o Big Data

- Organização
  - Uma vez que os dados estão no MySQL → Consultas Real Time
  - Transformar, consolidar, higienizar, padronizar, unificar (DQ)
  - Exportar dados
    - Batch (cargas)
    - Real-Time
    - Unidirecional
    - Bi-Direcional
  - Apache Sqooop
  - MySQL Applier

MySQL e o Big Data



MySQL e o Big Data



MySQL e o Big Data

- Exemplos Sqoop:

sqoop-import --connect jdbc:mysql://meu_servidor/meu_banco --driver com.mysql.jdbc.Driver –username meu_usuario –password minha_senha --table user --hbase-table user --hbase-row-key id --column-family data

sqoop-import --connect jdbc:mysql://meu_servidor/meu_banco --driver com.mysql.jdbc.Driver –username meu_usuario –password minha_senha --hbase-table user --hbase-row-key id --column-family data --query "SELECT a., b.* FROM a JOIN b WHERE condição"*

MySQL e o Big Data

- Análise
  - Cadê o MySQL?
  - Nesta fase é onde acontece o processamento de grandes datasets
  - Os dados já foram “transportados” para o HDFS (ou repositório)
  - Geralmente em clusters de servidores
  - Manipulação:
    - Hive
    - Pig
    - Mahout
    - Hue (GUI)
  - Map/Reduce
    - E seus derivativos

MySQL e o Big Data

- Decisão
  - Após o processamento mais pesado na fase de análise, pode-se “transportar” os dados novamente para o MySQL
  - Disponibilização de relatórios e cubos de dados
  - Utilização de ferramentas de BI para produção de conhecimento

Conclusão

- Ou você pega o Big Data ou ele vai pegar você
- É melhor preparar-se do que ser atropelado por ele
- MySQL pode ser utilizado sozinho até 2TB
- Acima de 2TB será necessário técnicas especiais
- Hadoop é o padrão
- Cloudera entrega a melhor distribuição
- Utilizar o MySQL + Hadoop como parte de um portfolio para um Big Data de qualidade, distribuindo carga e tarefas, é a melhor solução

OBRIGADO!

Alexandre M de Almeida

alexandremalmeida.com.br

alex_almeida_db


ORACLE®