

BIG DATA

CONCEITOS, DESAFIOS E OPORTUNIDADES



WILLER AMORIM SABINO DE ARAUJO

ROTEIRO

- Conceitos
 - O Que é Big Data?
 - Os Vs do Big data
 - Big Data analytics
 - Big Data x BI
 - Alguns números do Big Data
 - Bancos de dados não relacionais
 - Hadoop
- Desafios do Big Data
- Oportunidades
 - Aplicações

O QUE É BIG DATA?

- Termo usado para descrever grandes volumes de dados
- Área de estudo de grandes volumes de dados
- Um conceito sobre a análise de grandes volumes de dados
- Dados tão grandes que não podem ser processados, armazenados e analisados por ferramentas comuns de bancos de dados



OS Vs DO BIG DATA

- É uma das definições mais usadas para o Big Data.
- Inicialmente eram atribuídos 3 Vs para o Big Data, mas hoje existem vários outros Vs “não oficiais”.



VARIETY

structured
unstructured
semi-structured
all of the above

3 Vs of Big Data



VOLUME

terabytes
records
transactions
tables, files



VELOCITY

batch
real-time
streams
near-time

© 2014 Intelligent Software Solutions

OS TRÊS Vs INICIAIS

- **Volume:** Representa a grande quantidade de Dados do Big Data, que é basicamente a sua principal característica.
- **Velocidade:** Os dados do Big Data são gerados/recebidos em grande quantidade e a uma taxa muito alta, necessitando assim de alta velocidade.
- **Variedade:** 70% a 80% dos dados gerados por empresas são não estruturados. Esta é a principal propriedade que diferencia Big Data de aplicações de dados comuns que só armazenam dados estruturados.[1]

ALGUNS Vs IMPORTANTES:

- **Veracidade:** É preciso ter certeza de que os dados a serem analisados são precisos.
- **Valor:** Após todos os processos de análise nos dados, é preciso extrair o valor dos insights que estes dados geraram.
- **Variabilidade:** Um mesmo tipo de dado pode ser apresentado de formas diferentes, e estas diferenças podem ter um grande impacto nas análises de dados.
- **Visualização:** A parte de visualização é crucial para ter um bom resultado em Big Data. Ao término das análises é preciso ter uma boa visualização para entender melhor os resultados.



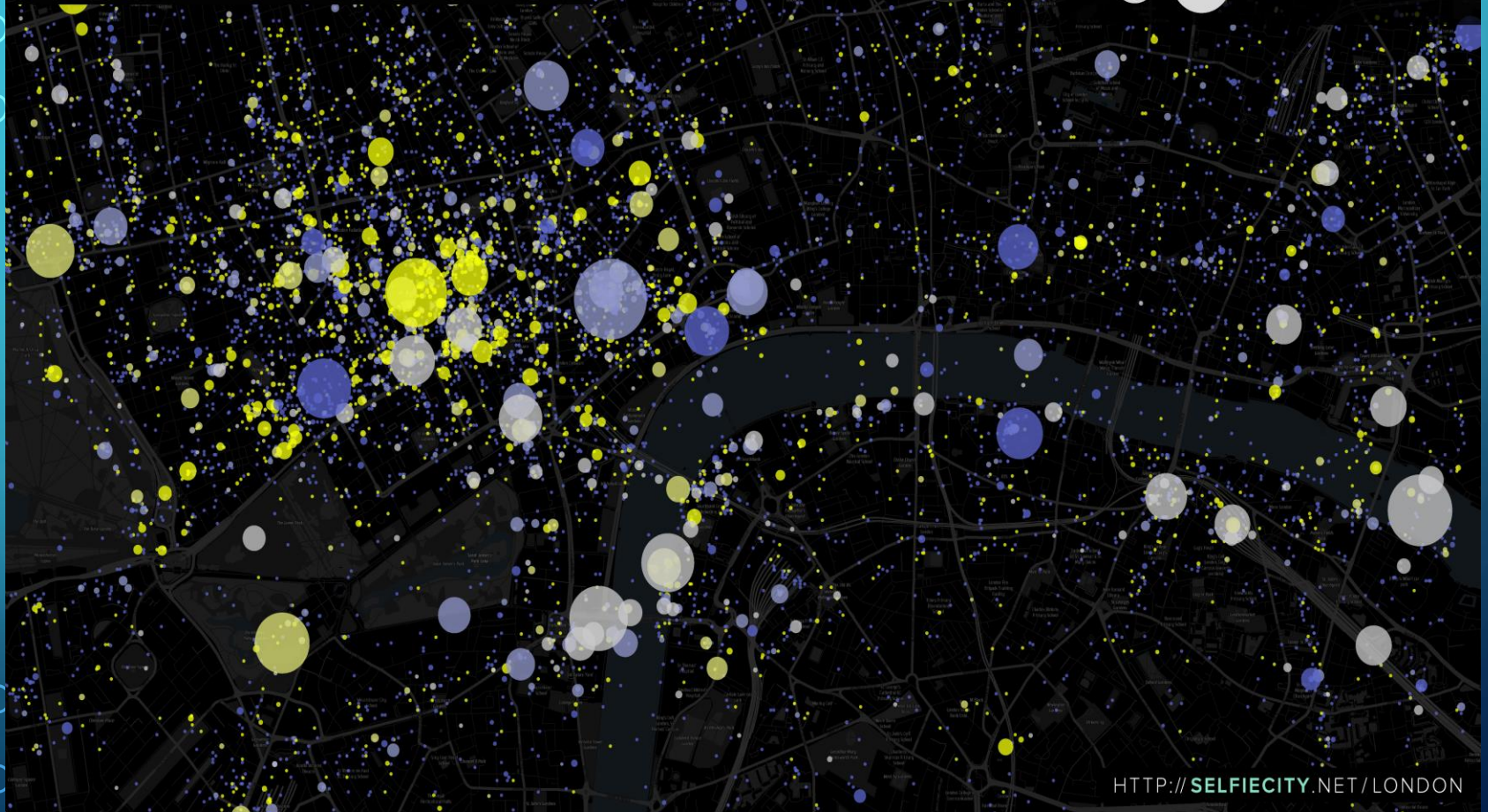
Localização geográfica de 3.6 milhões de tweets com fotos em Londres. Retirados de uma base de dados de 265 milhões de tweets com fotos entre os anos de 2011 e 2014.

Where do Londoners take **the most photos of people**, and where the least?

Instagram
Photos per Day



Percentage of
Photos with Faces



[HTTP://SELFIECITY.NET/LONDON](http://selfiecity.net/london)

Mapa dos lugares onde as pessoas mais tiram fotos no Instagram em Londres.

London Selfie Demographics

PERCENT OF IMAGES

62.8%

AVERAGE AGE

23.7



PERCENT OF IMAGES

37.2%

AVERAGE AGE

28.0

EST. AGE — 20 ————— 30 —————>

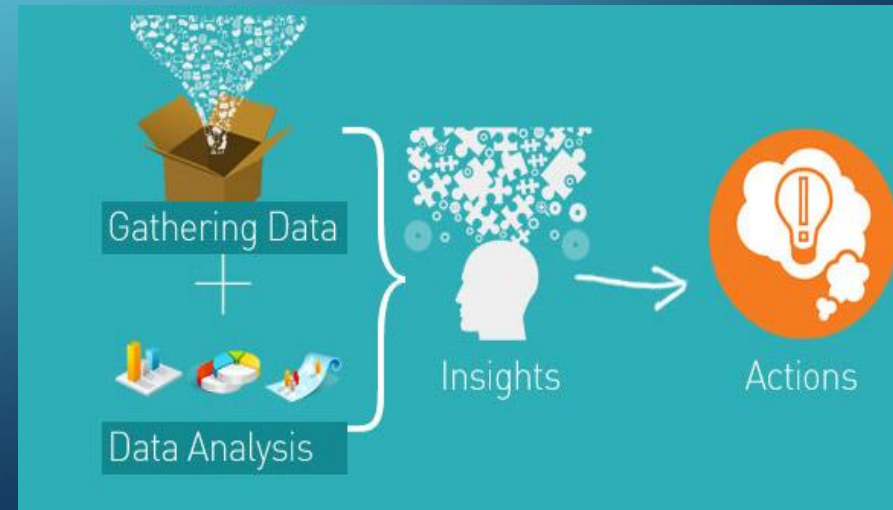


[HTTP://SELFIECITY.NET/LONDON](http://selfiecity.net/london)

Infográfico com várias Selfies de pessoas em Londres. Organizadas por idade, gênero e humor.

BIG DATA ANALYTICS

- É o processo de análise de grandes conjuntos de dados que tem como objetivo desvendar padrões, correlações, tendências de mercado, preferências de consumidor e outras informações relevantes.
- Estas informações são chamadas de insights.
- Empresas que utilizam Big Data tem 5 vezes mais chances de tomar decisões mais rapidamente que os seus competidores.



BI

x

Big Data

- Decisões baseadas em dados passados.
- Foco na análise dos dados.
- Universo de dados limitado.
- Análise intuitiva para ajudar a criação de diretrizes na empresa.
- Dados estruturados.

- Decisões baseadas em dados de tempo real.
- Foco na descoberta de tendências, correlações, aprender com os dados.
- Universo ilimitado de dados.
- Análise preditiva dos dados.
- Vários tipos de dados, estruturados e não estruturados.












ALGUNS NÚMEROS DO BIG DATA

- 4.4 milhões de empregos de TI serão gerados internacionalmente na área de big data até o fim de 2015 e cerca de 1.9 milhões destes apenas nos EUA. [1]
- O mercado de big data, em 2014, valia 28.4 bilhões de dólares e pode atingir a marca de 50.1 bilhões até 2017. [2]
- O Facebook atingiu a marca de 1.13 bilhões de usuários ativos diariamente em junho de 2016. Destes, 1.03 também acessam o site via mobile. Neste mesmo período, o Facebook atingiu a marca de 1.71 bilhões de usuários ativos mensalmente. [3]
- Em 2013, a cada minuto aconteciam cerca de 100 horas de upload de vídeo no Youtube, 20 milhões de visualizações de fotos no Flickr, mais de 2 milhões de buscas no Google, 204 milhões de e-mails enviados e 83 mil vendas na Amazon. Tais números só tendem a crescer a cada ano. [4]

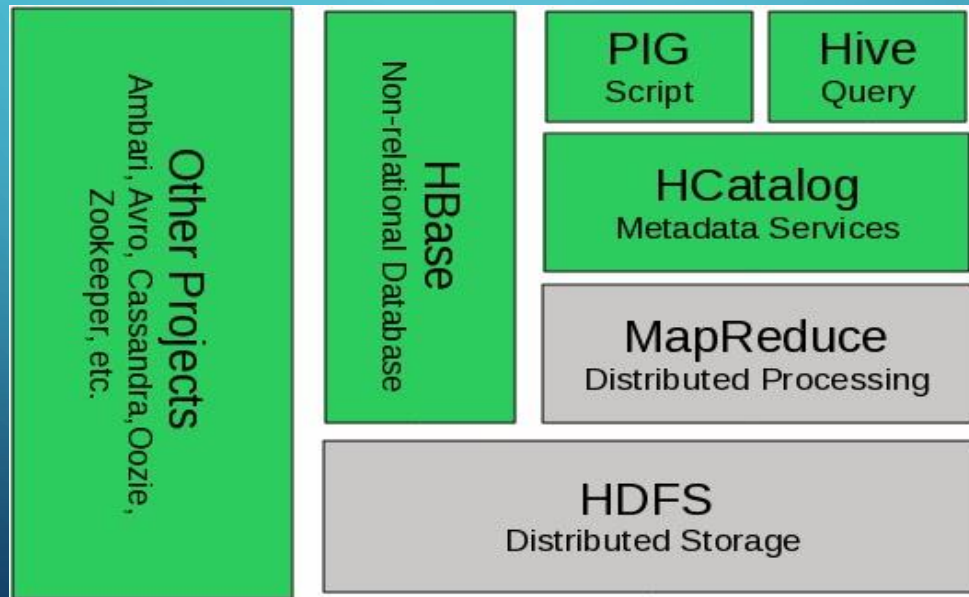
BANCOS DE DADOS NÃO RELACIONAIS

- Bancos de dados NoSQL:
 - Seus esquemas são dinâmicos, novas informações podem ser adicionadas durante a execução.
 - Manipulação através de APIs orientadas a objeto.
 - Maior escalabilidade.
 - Armazenam vários tipos de dados diferentes.

Types of NoSQL DBs	
GRAPH DATABASE	 Neo4j  TITAN
KEY VALUE DATABASE	 amazon DynamoDB  Cassandra  ORACLE BERKELEY DB
COLUMN DATABASE	 APACHE HBASE  Google BigTable
DOCUMENT DATABASE	 CouchDB  mongoDB

HADOOP

- Framework para processamento e armazenamento de grandes quantidades de dados.
- Conjunto de projetos open-source, foi baseado no Google MapReduce.
- 63% das empresas que lidam com Big Data usam Hadoop[1].



DESAFIOS

- Transferência e armazenamento: Tecnologia atual de discos rígidos não é tão eficiente quanto a de CPUs.
- Dificuldade no uso dos softwares: O Hadoop por exemplo não é simples de usar e como é uma tecnologia relativamente nova poucas estão familiarizadas com o seu manuseio.
- Visualização: É inviável plotar gráficos com bilhões de pontos e múltiplas variáveis.
- Falta de mão de obra qualificada: O número de cientistas de dados ainda é muito pequeno.
- Qualidade de dados: Empresas nos EUA perdem 600 bilhões de dólares todo ano devido a “dirty data”[6].

OPORTUNIDADES

- **Segurança:** Detecção de padrões de comportamento físico ou digital que ofereçam algum tipo de risco.
- **Saúde:** Diagnósticos preditivos, análise de dados genéticos, descobertas de novos tratamentos.
- **Marketing:** Otimização de alocação de recursos, identificação de tendências, descobertas de nichos de mercado.
- **Educação:** análises de evasão escolar, classificação de perfis de aprendizado diferenciados.

APLICAÇÕES

- Aumento de segurança em infraestrutura de TI
- Reduzir custos da empresa
- Otimização de processos
- Previsão de tendências e movimentação de mercado
- Transformar os dados em informações exploráveis
- Analisar os fluxos de dados
- Analisar a "e-reputação"
- Analisar e prever os comportamentos dos consumidores:

REFERÊNCIAS

- 1- Big Data: Survey, Technologies, Opportunities, and Challenges: <https://www.hindawi.com/journals/tswj/2014/712826/abs/>
- 2- Statista - Number of big data jobs in the United States and globally in 2015: <https://www.statista.com/statistics/255952/number-of-big-data-jobs-in-us-and-globally/>
- 3- Big Data Vendor Revenue and Market Forecast 2013-2017: http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017
- 4- Facebook company info and statistics: <http://newsroom.fb.com/company-info/>
- 5- What Happens in an Internet Minute?: http://scoop.intel.com/what-happens-in-an-internet-minute/?utm_source=dataflog&utm_medium=ref&utm_campaign=dataflog
- 6- Causes of dirty data: <https://www.qubole.com/blog/big-data/causes-of-dirty-data-and-how-to-combat-them/>

The image features a dark blue gradient background with white circuit-like lines in the corners. These lines consist of straight paths that branch out and terminate in small circles, resembling a printed circuit board (PCB) layout. The lines are positioned in the top-left, top-right, bottom-left, and bottom-right corners, framing the central text.

DÚVIDAS?