



Big Data para o desenvolvimento Urbano Sustentável

Termo de Referência 8



Conceito e Definições

O Termo “Big Data”

- ❑ O Big Data, ou “Megadados” (em português), em tecnologia da informação é o termo que utilizamos para nos referir à um grande volume de dados armazenados, que são gerados em alta **velocidade** e **variedade**. Essas informações necessitam de ferramentas **não convencionais** para o seu armazenamento e processamento.
- ❑ Quando nos referimos à **grandes volumes** de dados na era atual da informação, estamos nos referindo à valores da ordem de **Tera, Petta, Hexa bytes** ou até mais.
- ❑ Somente quando novos dados são gerados com uma periodicidade muito alta, diariamente ou semanalmente, por exemplo, convém necessariamente a definição de Big Data.

O Termo “Tempo Real”

- ❑ Um sistema de **tempo real** se refere à uma tarefa, ou um conjunto delas, que possuem um tempo de execução rígido independente da carga do sistema. Este tempo pode ser muito curto ou não, isso **dependerá dos requisitos do projeto**.
- ❑ O sistema deverá cumprir a tarefa no prazo ou informar imediatamente que não poderá fazê-la.
- ❑ Sistemas de **trading** tem um requisito de **milésimos de segundo** para computar uma transação.
- ❑ Em um sistema de **notificação de acidentes de trânsito**, o requisito de tempo pode ser **10 minutos**, por exemplo.

Computação na Nuvem e Infraestrutura acerca do Volume de Dados

O Termo “Computação na Nuvem”

- ❑ O termo nuvem faz referência à **internet**. Porque é mais simples imaginá-la como uma **nuvem**, do que como uma **emaranhada rede de conexões, máquinas e dados**.
- ❑ Resumindo, a computação em nuvem inclui servidores, armazenamento, banco de dados, rede, software, análise e inteligência, tudo como um **serviço contratável**.
- ❑ Este tema é muito presente quando tratamos do assunto de Big Data, porque **é muito mais simples escalar um negócio** quando ele reside na “nuvem”.

2,2 Milhões

de Terabytes de novos dados
gerados **todos os dias**

(Instituto Gartner)

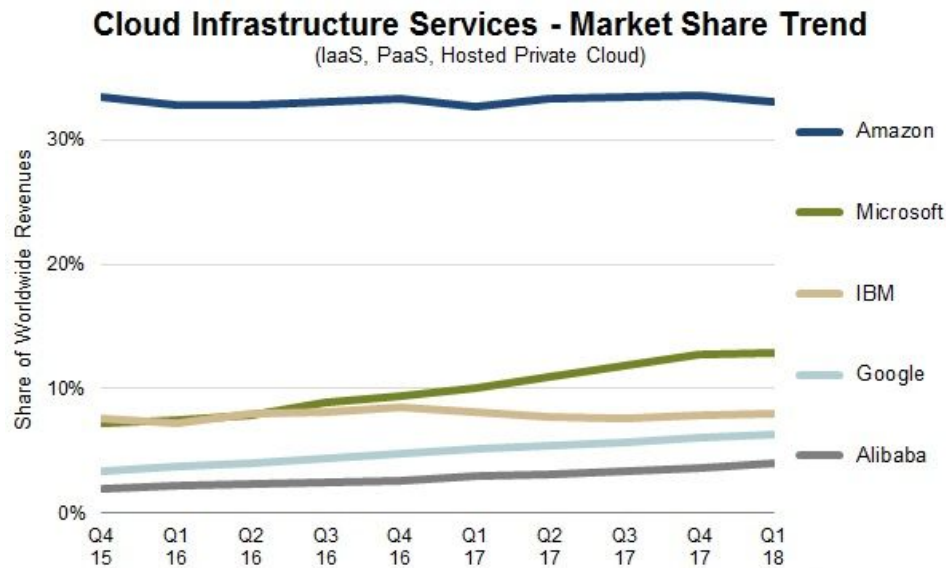
Serviços de Computação na Nuvem

- ❑ Com o crescimento da popularidade dos termos Big Data e Cloud Computing (Computação na Nuvem), os gigantes tecnológicos como Google, Microsoft e Amazon enxergaram excelentes oportunidades de negócio.



Por que AWS?

- ❑ Falaremos majoritariamente das soluções apresentadas pela Amazon na AWS.
- ❑ Essa escolha deve-se a dominância da empresa no mercado de Cloud Computing. Um estudo do **Synergy Research Group** mostrou que a AWS compartilhou predominantemente de **33%** das receitas destinadas à área entre os anos de 2015 e 2018, mesmo com o mercado quase **triplicando** de tamanho no mesmo período.



(fonte: Synergy Research Group)

Por que AWS?

Outro estudo mais recente, realizado pela **Canalys Cloud Channels Analysis** em Fevereiro de 2020, reportou que a AWS compete com **32.4%** do mercado, Azure com 17.6%, Google Cloud com 6%, Alibaba Cloud com 5.4% e outras soluções de cloud com 38.5%.

Worldwide cloud infrastructure spending and annual growth
Canalys estimates, Q4 2019

Cloud service provider	Q4 2019 (US\$ billion)	Q4 2019 market share	Q4 2018 (US\$ billion)	Q4 2018 market share	Annual growth
AWS	9.8	32.4%	7.3	33.4%	33.2%
Microsoft Azure	5.3	17.6%	3.3	14.9%	62.3%
Google Cloud	1.8	6.0%	1.1	4.9%	67.6%
Alibaba Cloud	1.6	5.4%	1.0	4.4%	71.1%
Others	11.6	38.5%	9.3	42.4%	24.4%
Total	30.2	100.0%	22.0	100.0%	37.2%



Note: percentages may not add up to 100% due to rounding
Source: Canalys Cloud Channels Analysis, January 2019

Só funciona com AWS?

- ❑ Embora exista uma clara preferência pelas soluções da AWS, também é possível que as cidades utilizem as soluções da Google Cloud Platform, Microsoft Azure e afins para implementar os projetos que descreveremos aqui, pois existe uma equivalência entre os serviços disponibilizados por elas e é possível alternar entre as soluções por um baixo custo de transição.
- ❑ As cidades podem avaliar quais as soluções são mais convenientes para si, tendo, portanto, autonomia para utilizarem outras opções.

Infraestrutura acerca do Volume de Dados

Amazon Simple Storage Service (Amazon S3)

“O S3 é um serviço de armazenamento de objetos que oferece escalabilidade, disponibilidade de dados, segurança e performance. Isso significa que clientes de todos os tamanhos e setores podem usá-lo para armazenar qualquer volume de dados em uma grande variedade.”

- ❑ Considerar o S3 é interessante quando para o seu problema, faz sentido armazenar os arquivos em formato de texto, e/ou até JSON e CSV (compactados ou não).
- ❑ Existem soluções, como Amazon Athena e Amazon Glue, que trabalham diretamente com o S3 para analisar e processar os dados direto no seu formato de arquivo. Porque no Big Data é importante que o processamento seja levado até o dado e não o contrário.

Amazon Athena

“O Amazon Athena é um serviço de consultas interativas que facilita a análise de dados no Amazon S3 usando SQL padrão. O Athena não precisa de servidor. Portanto, não há infraestrutura para gerenciar e você paga apenas pelas consultas executadas. Basta apontar para os dados no Amazon S3, definir o schema e iniciar as consultas usando SQL padrão. Com o Athena, não há necessidade de trabalhos complexos de ETL para preparar dados para análise. [...]”

- ❑ O Amazon Athena é uma ferramenta excelente para analisar dados que estão armazenados em um Bucket do Amazon S3.
- ❑ Trabalha com um grande volume de dados e é ideal para visualizar os dados e extrair alguns insights, mas não é viável para integrar em aplicações que requisitam resposta imediata.

Amazon Kinesis

“O Amazon Kinesis facilita a coleta, o processamento e a análise de dados de streaming em tempo real, permitindo que você obtenha insights oportunos e reaja rapidamente às novas informações. Permite processar e analisar dados assim que são recebidos e responder instantaneamente, em vez de aguardar a conclusão da coleta de dados para poder iniciar o processamento.”

- ❑ Um dos grandes requisitos para processar quantidades massivas de dados e obter respostas em tempo real é o processamento em paralelo. É fortemente necessário que o sistema seja capaz de trabalhar paralelamente com a geração de dados.

Estatística aplicada ao Big Data

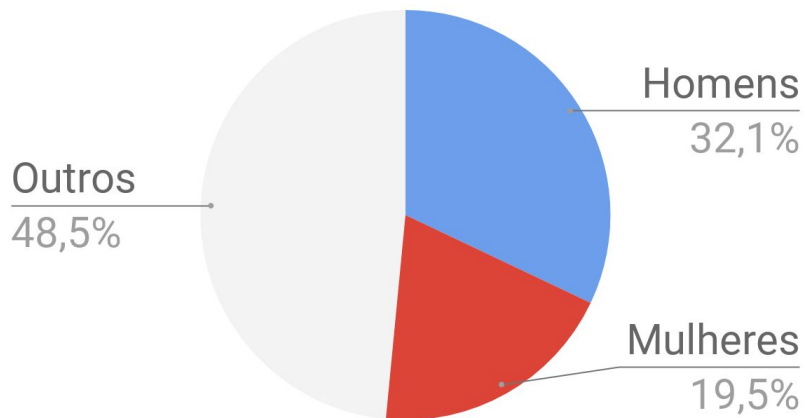
A Incerteza

- ❑ Quando queremos fazer perguntas aos nossos dados, é importante decidirmos o quanto certos queremos estar sobre a resposta.
- ❑ Imagine que você queira calcular a porcentagem de homens e mulheres de uma base de dados que armazenou os dados de consumo de vários usuários.
- ❑ Uma inteligência artificial foi criada especialmente para analisar o padrão de consumo e inferir se aquele consumidor é homem ou mulher.

Base de Consumidores

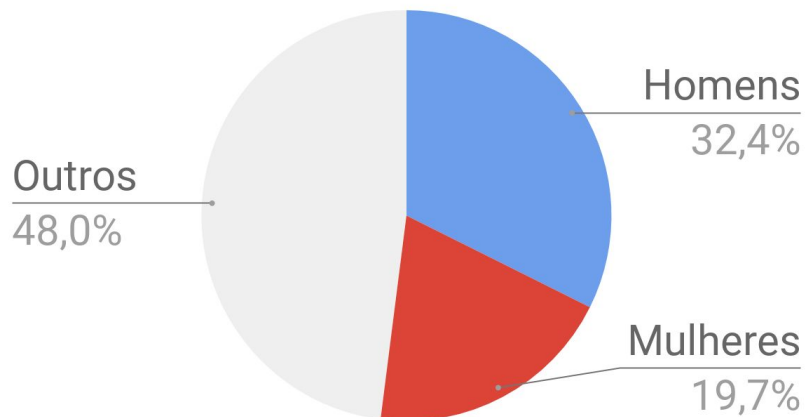
Com 100% de certeza

População de 488 Milhões de Consumidores



Com 99% de certeza

Amostra Aleatória de 16 mil consumidores



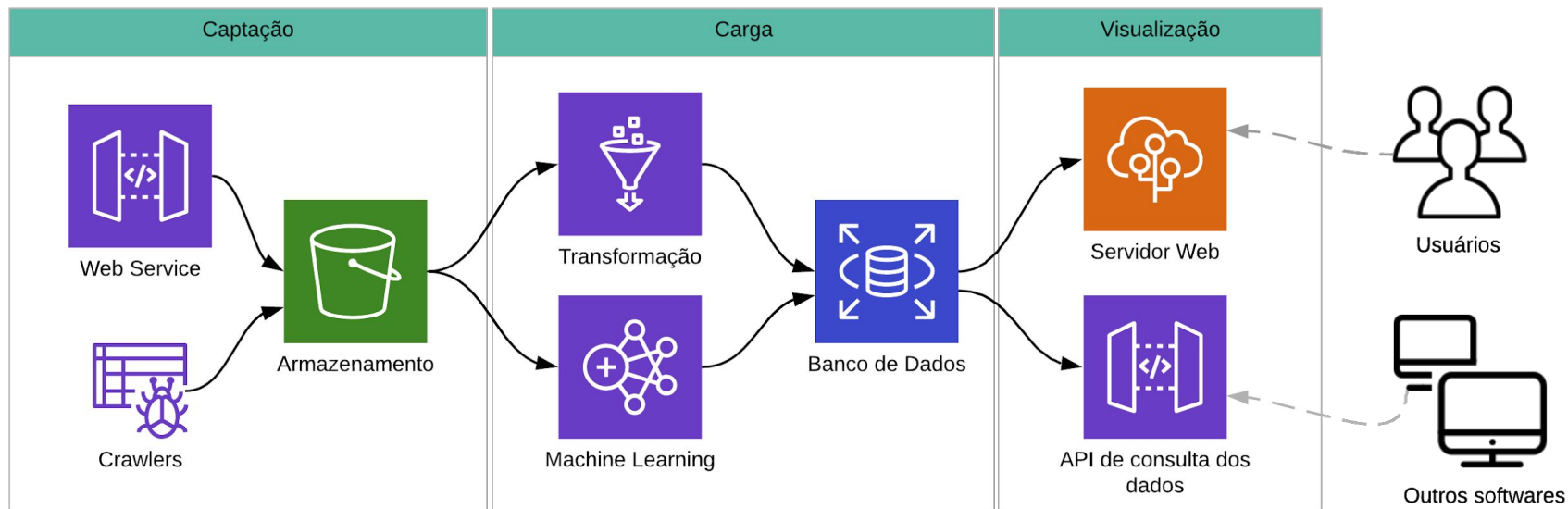
(Dados fictícios criados para fins de exemplificação)

Valor no Big Data

- ❑ Quanto maior a riqueza de dados, mais importante é saber realizar as perguntas certas no início de todo processo de análise. (Eric Brown, 2014).
- ❑ É necessário estar focado para a orientação do negócio, o valor que a coleta e análise dos dados trará para o negócio. Não é viável realizar todo o processo de Big Data se não se tem questionamentos que ajudem o negócio de modo realístico. Da mesma forma é importante estar atento aos custos envolvidos nessa operação, o valor agregado de todo esse trabalho desenvolvido, coleta, armazenamento e análise de todos esses dados têm que compensar os custos financeiros envolvidos. (C. Taurion, 2013).

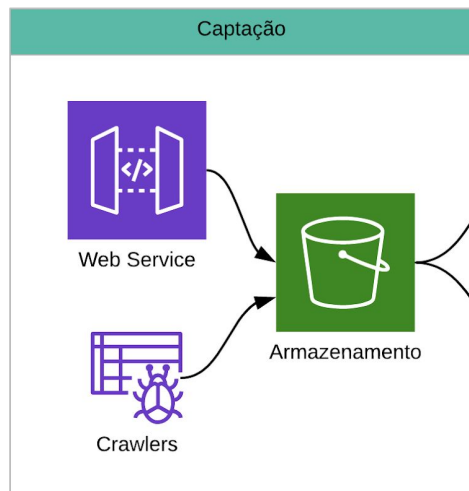
Arquitetura e dimensionamento de custos

Dividimos o projeto em 3 grandes partes



Captação / Coleta

- Na captação definimos como iremos acessar os dados puros e armazená-los para o uso no passo seguinte.



Web Service

Web Service: É uma solução utilizada na integração de sistemas e na comunicação entre aplicações diferentes.

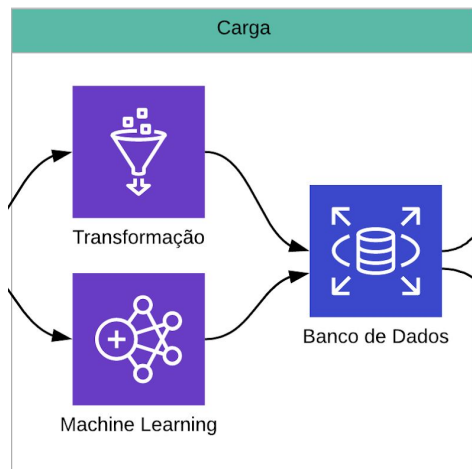


Crawlers

Web Crawlers: softwares-robôs capazes de navegar na internet e extrair qualquer tipo de informação contida no texto e enviá-la para um local de processamento.

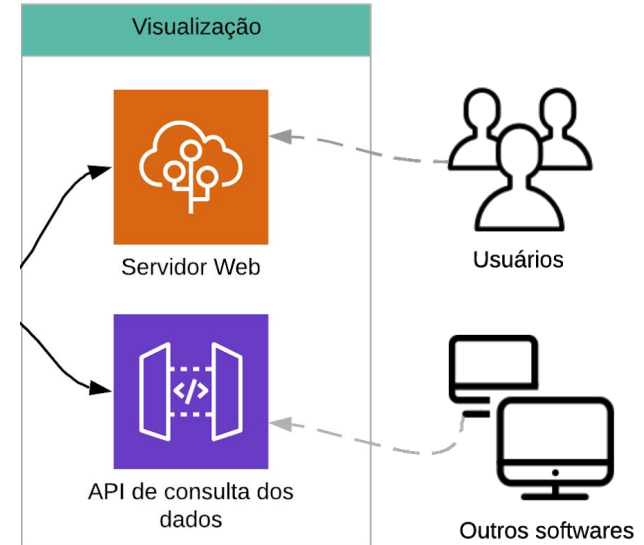
Carga / Armazenamento

- ❑ Nesta fase já podemos nos preocupar em transformar os dados. Nesta etapa também integramos algoritmos de machine learning. Todos com o propósito de estruturar o acesso a informação para a etapa de visualização.



Visualização / Análise

- Quando chega esta etapa, seus dados já estão em uma base de dados que pode ser facilmente integrados com alguma aplicação web para mostrar gráficos, mapas, estudos etc. Também é possível criar APIs de consulta de dados para que outros softwares possam fazer análises nos dados apresentados.



Estimativa de Preço

Estimativa da “Captação”

Realizaremos a estimativa de custo relativas somente a etapa que descrevemos como “Captação”. Os cálculos também serão realizados como um exemplo num número limitado de requisições, todavia os custos da solução em cenários reais podem escalar para 10 vezes mais o preço estimado.

Amazon S3

- ❏ Por fim, o Amazon S3 cobra pelo armazenamento e pela transferência dos dados. O ato de inserir os dados não contabiliza custos de transferência, mas quando direcionarmos esses dados armazenados para um ETL, haverá uma taxa cobrada pela transferência.

Amazon S3 - Standard (Armazenamento padrão)	
Primeiros 50 TB/mês	0,0405 USD por GB
Próximos 450 TB/mês	0,039 USD por GB
Mais de 500 TB/mês	0,037 USD por GB

Amazon API Gateway

- Selecionamos alguns serviços da Amazon Web Services que irão compor uma versão piloto do produto. O primeiro deles é o **Amazon API Gateway**, que é um serviço para criação de **API Rests**.
- Este serviço atuará como a porta de entrada dos dados em uma possível integração via **Web Service**.

Amazon API Gateway	
Número de solicitações (por mês)	Preço (por milhão)
Próximos 667 milhões	4,25 USD
Próximos 19 bilhões	3,61 USD
Mais de 20 bilhões	2,29 USD

Amazon Kinesis

- ❑ O segundo serviço é o Amazon Kinesis Data Stream, no qual discutimos anteriormente.
- ❑ No Amazon Kinesis Data Stream, o preço é definido a partir da Unidade de Carga PUT (que equivale a 25 KB). A unidade de carga PUT é cobrada com base em uma taxa por milhão de unidades de carga PUT.
- ❑ Outra unidade utilizada é o estilhaço, uma unidade básica de throughput de um stream de dados do Amazon Kinesis. Você especifica o número de fragmentos necessários no stream de acordo com os requisitos de taxa de transferência. Para cada fragmento, é cobrada uma taxa horária.

Amazon Kinesis

Amazon Kinesis Data Stream	
Unidades de carga PUT, por 1.000.000 de unidades	0,028 USD
Hora de estilhaço (entrada 1 MB/segundo, saída 2 MB/segundo)	0,03 USD

Assumindo um cenário

- ❏ Vamos precisar estimar alguns valores para calcular o custo mensal da etapa de captação. Para isto, vamos assumir a aplicação receberá 1000 (mil) eventos por minuto e que cada evento terá em média 10 Kilobytes de dados.

Calculando...

Amazon API Gateway	$1000 \text{ chamadas/minuto} * 60 \text{ minutos/hora} * 730 \text{ horas/mês} =$ $43.000.000 \text{ chamadas/mês} * 4,25 \text{ USD/milhão} =$ = 186,15 USD / mês
--------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Calculando...

Amazon Kinesis Data Stream	$\begin{aligned} &1000 \text{ registros/minuto} / (60 \text{ segundos/minuto}) = \\ &= 16.7 \text{ registros /segundo} * 2628000 \text{ segundos/mês} = \\ &43.800.000 \text{ PUT/mês} * 0.000000028 \text{ USD} = \mathbf{1,23 \text{ USD}}. \end{aligned}$ $\begin{aligned} &1 \text{ estilhaço} * 730 \text{ horas/mês} = 730.00 \text{ estilhaço-hora/ mês} \\ &* 0.03 \text{ USD} = \mathbf{21,90 \text{ USD}}. \end{aligned}$ $= \mathbf{23,13 \text{ USD / mês}}$
----------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Calculando...

Amazon S3	$10 \text{ KB} * 1000 \text{ chamadas/minuto} * 60 \text{ minutos/hora} * 24 \text{ horas/dia}$ $* 30 \text{ dias/mês} = 432 \text{ GB/mês} =$ $= 438 \text{ GB} * 0,0405 \text{ USD} =$ $= 17,74 \text{ USD / mês}$
-----------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Resultado!

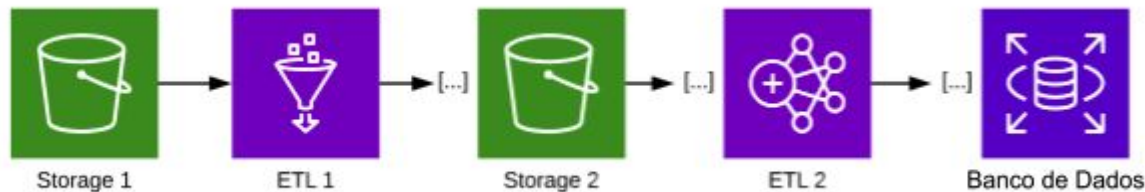
Amazon API Gateway	186,15 USD / mês
Amazon Kinesis Data Stream	23,13 USD / mês
Amazon S3	17,74 USD / mês
TOTAL	227,02 USD / mês

Demonstração

<https://fgv.urbbox.com.br>

O que é ETL?

- ❑ É um acrônimo de Extract Transform Load, são ferramentas de software cuja função é a extração de dados de diversos sistemas, transformação desses dados conforme regras de negócios.
- ❑ ETLs são estruturas bem versáteis que podem ser acopladas uma após a outra criando um pipeline de transformação dos dados.



Sobre o Projeto

O projeto **Big Data para o Desenvolvimento Urbano Sustentável** que tem como objetivo desenvolver um modelo piloto, replicável e escalável de governança de big data para as cidades de:

-  São Paulo (Brasil)
-  Miraflores (Peru)
-  Montevideo (Uruguai)
-  Quito (Equador)
-  Xalapa (México)

O projeto visa...

- Conhecer o ecossistema de dados existentes e os marcos regulatórios em matéria de governança de dados para o acesso e a utilização nas cidades do projeto;
- Desenvolver ferramentas comuns de âmbito legal, técnico e institucional para acessar, integrar e analisar os dados;
- Testar modelos de governança a partir do desenvolvimento de projetos com foco em mudanças do clima e mobilidade em cinco cidades da região;
- Criar unidades de análise de Big data em cidades da região;
- Consolidar o conhecimento gerado de maneira a facilitar sua replicação e disseminação em nível regional.