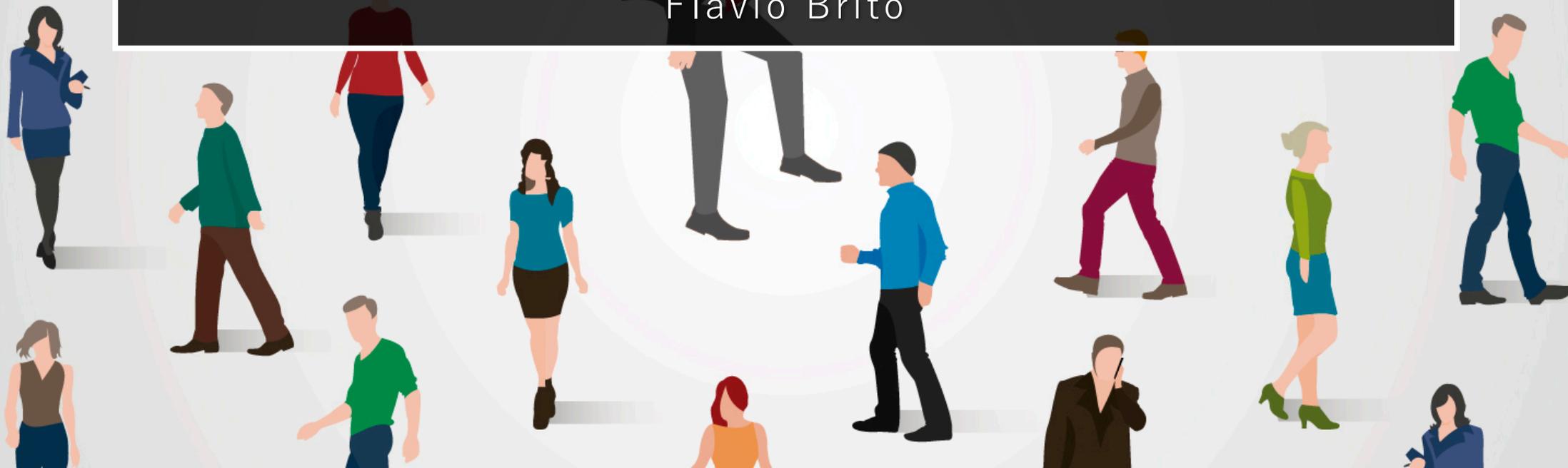


# AUDITORIA DE DADOS

Flávio Brito



# AGENDA

## MÓDULO I

### **Introdução**

Proposta do Curso e objetivos

Material do Curso e dinâmica

Recursos

## MÓDULO 2

### **Fundamentos sobre Auditoria de Dados**

Conceitos gerais e definições

Histórico e utilizações

Aplicações corporativas

O mercado

Preparação profissional

Recursos

# AGENDA

## MÓDULO 3

### Arquivos

#### **Metadados e Armazenagem**

#### **Integração com Excel, Python e R**

#### **Operações e Status**

- Listagem em Diretório
- Permissionamento
- Mudança de Diretório
- Cópia
- Remoção
- Merge
- Append
- Split
- Compactação via script

# AGENDA

## MÓDULO 4

### **Carga / Importação**

**Estruturação de Rotinas**

**Boas práticas**

### **Dados**

**Tipos de Dados**

**Metadados e rastreabilidade**

**Limpeza de Dados**

**Filtro de Dados**

**Operações com Variáveis**

**Coerções**

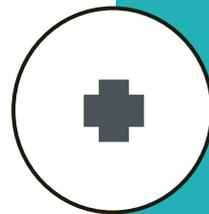
**Pitfalls**

**Outliers**

**Padrões de Informação**

# Introdução

- Proposta do Curso e objetivos
- Material do Curso e dinâmica
- Recursos



# Fundamentos sobre Auditoria de Dados

## Conceitos gerais e definições

- Auditoria de dados é definida como o processo de inspeção, limpeza, transformação, e modelagem de dados com o objetivo de destacar informações úteis, sugerindo conclusões, e apoiar a tomada de decisões. (Diversas fontes)
- Auditoria de dados é um processo analítico pelo qual ideias são extraídas de dados eletrônicos internos e externos a partir de fontes operacionais, financeiras entre outras. Esses insights podem ser históricos, em tempo real, ou preditivos e também pode ser focadas em risco (por exemplo, controle de eficácia, fraude, desperdício, abuso, política de não conformidade regulamentar) ou o desempenho, centrado (por exemplo, no aumento de vendas, redução de custos, melhorou rentabilidade) e frequentemente fornecer o "Como?" e "Por quê?" respostas para a inicial "o quê?" perguntas frequentemente encontradas na informação inicialmente extraídas dos dados. (KPMG)

# Histórico

As aplicações de auditoria são um tipo aplicação voltadas especificamente para processos de auditoria tanto para empresas médias e grandes, especialmente quando algumas das aplicações são desenvolvidas in-house.

Há alguns princípios básicos de aplicações de auditoria que os auditores de TI precisam saber e entender.

- No fim dos anos 1980 – criados por empresas voltadas para o desenvolvimento de softwares genéricos para a área de auditoria
  - ACL, 1987
  - Caseware, 1988
- Charles Carlsaw em Applying Benford's Law to Accounting, 1988
- Continuous Process Auditing System, AT&T Bell Laboratories, 1989

# Aplicações Corporativas

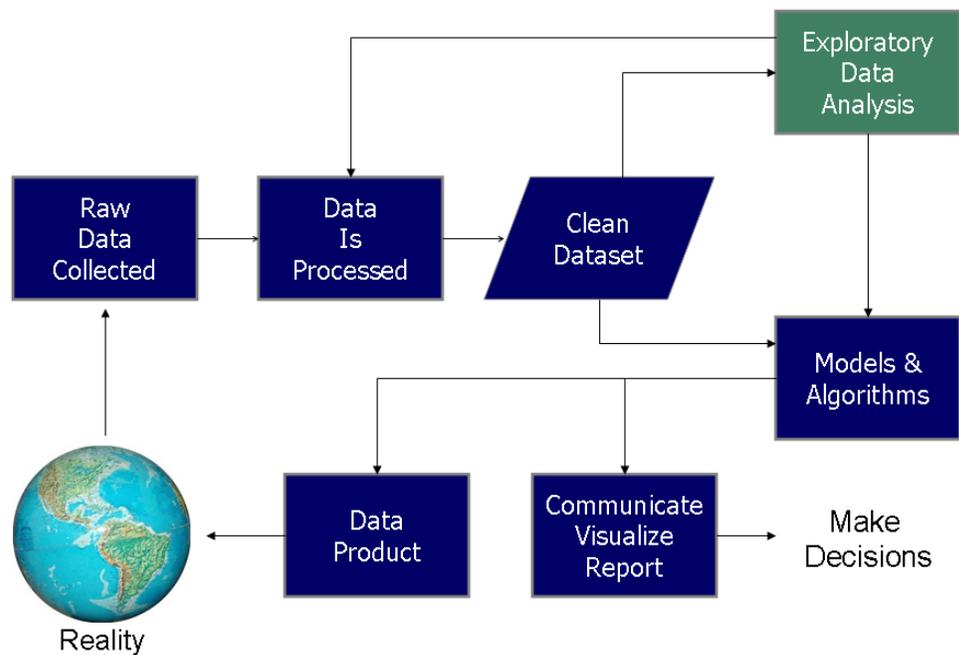
- TIPO
  - Sistemas Transacionais
  - Monitoramento
  - Análise em Tempo Real
  - Predição
- ÁREAS
  - TI
  - Financeira
  - RH
  - Governo
  - Demais Indústrias



# Mercados de Trabalho

- Data Auditor
- Data Analyst
- Data Scientist

## Data Science Process



# Utilizações Funcionalidades Desejadas nas Ferramentas

- Capazes de manipular grandes bases de dados
- Vasto número de funcionalidades estatísticas e analíticas
- Programável
- Registro de procedimentos – Rastreabilidade
- Facilidade de execução na mudança de dados

```
ngScope, element, attr, ngSwitchController) {  
  scope = attr.ngSwitch || attr.on,  
  transcludes = [],  
  elements = [],  
  selectedElements = [],  
  scopes = [];  
  
  watchExpr, function ngSwitchWatchAction(value) {  
  
    i = previousElements.length; i < ii; ++i) {  
      elements[i].remove();  
  
      scopes.length = 0;  
  
      i = selectedScopes.length; i < ii; ++i) {  
        selectedElements[i];  
        [i].$destroy();  
        scopes[i] = selected;  
        selected, function() {  
          scopes.splice(i, 1);  
  
          length = 0;  
          th = 0;  
  
          ides = ngSwitchController.cases['!' + value] || ngSwitchC  
          range);  
          nscludes, function(selectedTransclude) {  
            = scope.$new();  
            h(selectedScope);  
          }  
        }  
      }  
    }  
  }  
}
```

# Tipos de Ferramenta

- Excel
- Bancos de Dados Corporativos
- Desktop software (Microsoft Access)
- Software genérico para auditoria
  - ACL (1987)
  - IDEA DATA ANALYTICS



```
ngScope, element, attr, ngSwitchController) {  
  scope = attr.ngSwitch || attr.on,  
  transcludes = [],  
  elements = [],  
  elements = [],  
  scopes = [];  
  
  watchExpr, function ngSwitchWatchAction(value) {  
    i = previousElements.length; i < ii; ++i) {  
      elements[i].remove();  
  
      scopes.length = 0;  
  
      selectedScopes.length; i < ii; ++i) {  
        selectedElements[i];  
        destroy();  
        = selected;  
        selected, function() {  
          elements.splice(i, 1);  
  
          length = 0;  
          th = 0;  
  
          des = ngSwitchController.cases['!' + value] || ngSwitchC  
          range);  
          nscludes, function(selectedTransclude) {  
            = scope.$new();  
            h(selectedScope);  
          }  
        }  
      }  
    }  
  }  
}
```

# Alguns Desafios para a Auditoria de Dados



- Escolha do Software
- Fonte de Dados
- Segurança e Privacidade das Informações
- Fonte de Dado e Saída de Informação (Visualização)
- Qualificação e Experiência da Equipe
- Poder de Processamento

# Software



- Capacidade de Processamento

## EXCEL

### Especificações e limites de planilha e de pasta de trabalho

Recurso	Limite máximo
Pastas de trabalho abertas	Limitado pela memória disponível e pelos recursos do sistema
Tamanho da planilha	1.048.576 linhas por 16.384 colunas

### Especificações do banco de dados

Lista de tabelas a seguir se aplica aos bancos de dados do Access 2016:

## ACCESS

#### Geral

Atributo	Máximo
Tamanho total para um banco de dados do Access 2016 (.accdb), incluindo todos os objetos de banco de dados e dados	2 gigabytes menos o espaço necessário para objetos do sistema. <b>OBSERVAÇÃO</b> : Você pode contornar esta limitação de tamanho vinculando a tabelas em outros bancos de dados do Access. Você pode vincular a tabelas em vários arquivos de banco de dados, cada um dos quais pode ser tão grande quanto 2GB. <b>DICA</b> : Para obter mais informações sobre como reduzir o tamanho do seu banco de dados, consulte <a href="#">impedir e corrigir problemas de arquivo de banco de dados usando compactar e reparar</a> .

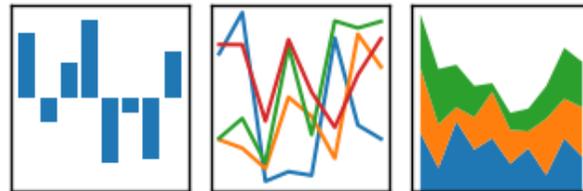
# Tipos de Ferramenta

- Ferramentas Estatísticas e de Análise de Dados
  - R (1993)
  - SAS (1976)
  - Python (1991) – Pandas e Scikit-learn (2011)



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



# Software



DataCamp  
Learn data analysis for free,  
*interactively*

# DATA SCIENCE WARS



VS.



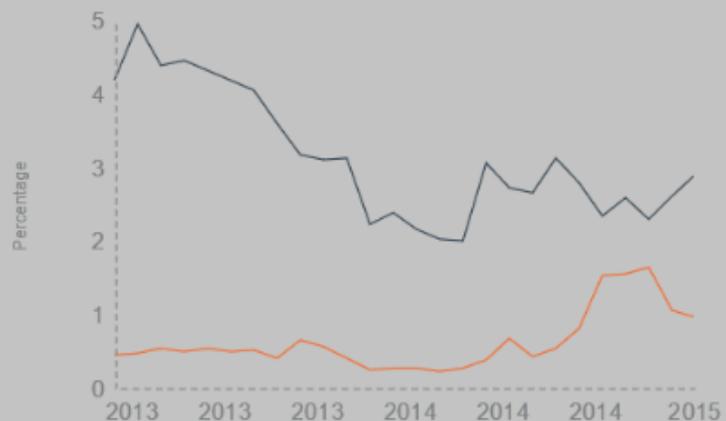
python

# Software

## R and Python: The Numbers

### Popularity Rankings

R and Python's popularity between 2013 and February 2015 (Tiobe Index)



### Jobs And Salary?

2014 Dice Tech Salary Survey:  
Average Salary For High Paying Skills and Experience

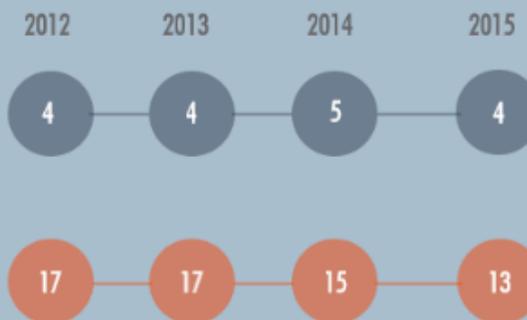


**\$ 115,531**

Redmonk ranking, comparing the relative performance of programming languages on GitHub and Stack Overflow (September 2012 and January 2013, 2014, 2015)

Python

R



**\$94,139**

# Software

#2

## The Data Analysis Battlefield

### Usage

R is mainly used when the data analysis tasks require standalone computing or analysis on individual servers.

Python is generally used when the data analysis tasks need to be integrated with web apps or if statistics code needs to be incorporated into a production database.

### Task

For exploratory work, R is easier for beginners. Statistical models can be written with a few lines of code.

As a full-fledged programming language, Python is a good tool to implement algorithms for production use.

### Data Handling Capabilities

R is handy for data analysis because of the huge number of packages, readily usable tests and the advantage of using formulas.

The infancy of Python packages for data analysis was an issue in the past, but this has improved a lot!

# Software

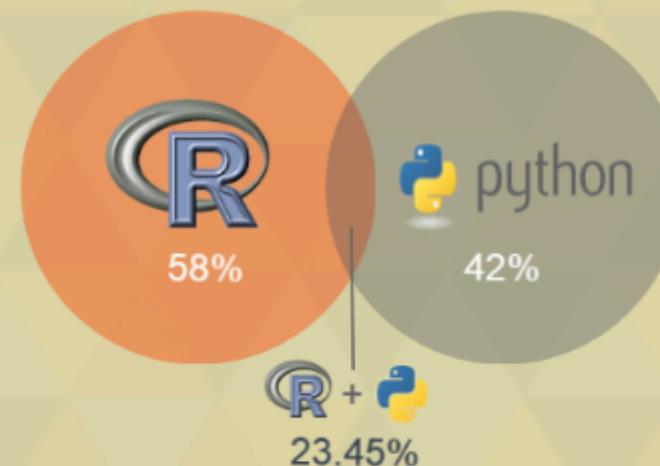
## R And Python: The Quantified Battlefield

### General

Languages for data analysis used in 2014 (KDnuggets polls)



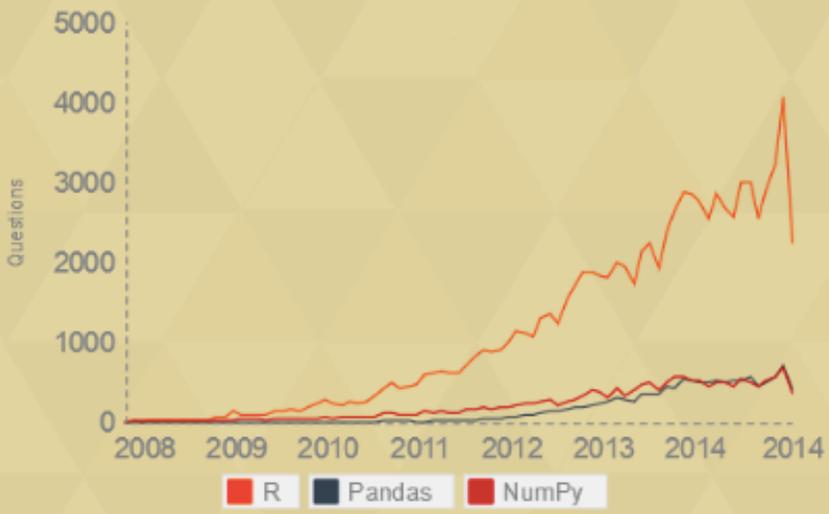
Analysis of R and Python used together in 2014 (KDnuggets polls)



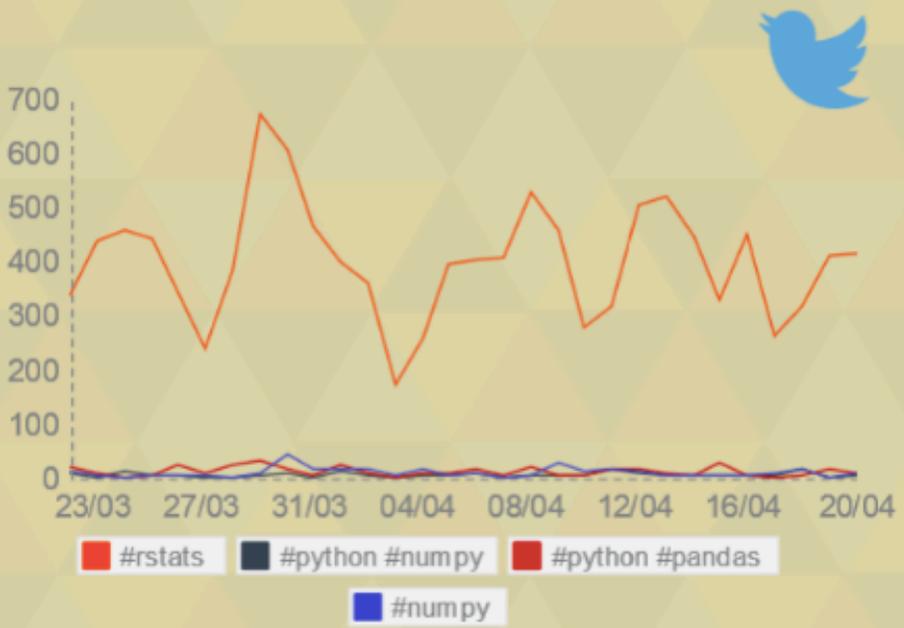
# Software

## Community?

Stack Overflow Questions tagged "R" and/or "Python", "Pandas" between 2008 and April 15, 2015

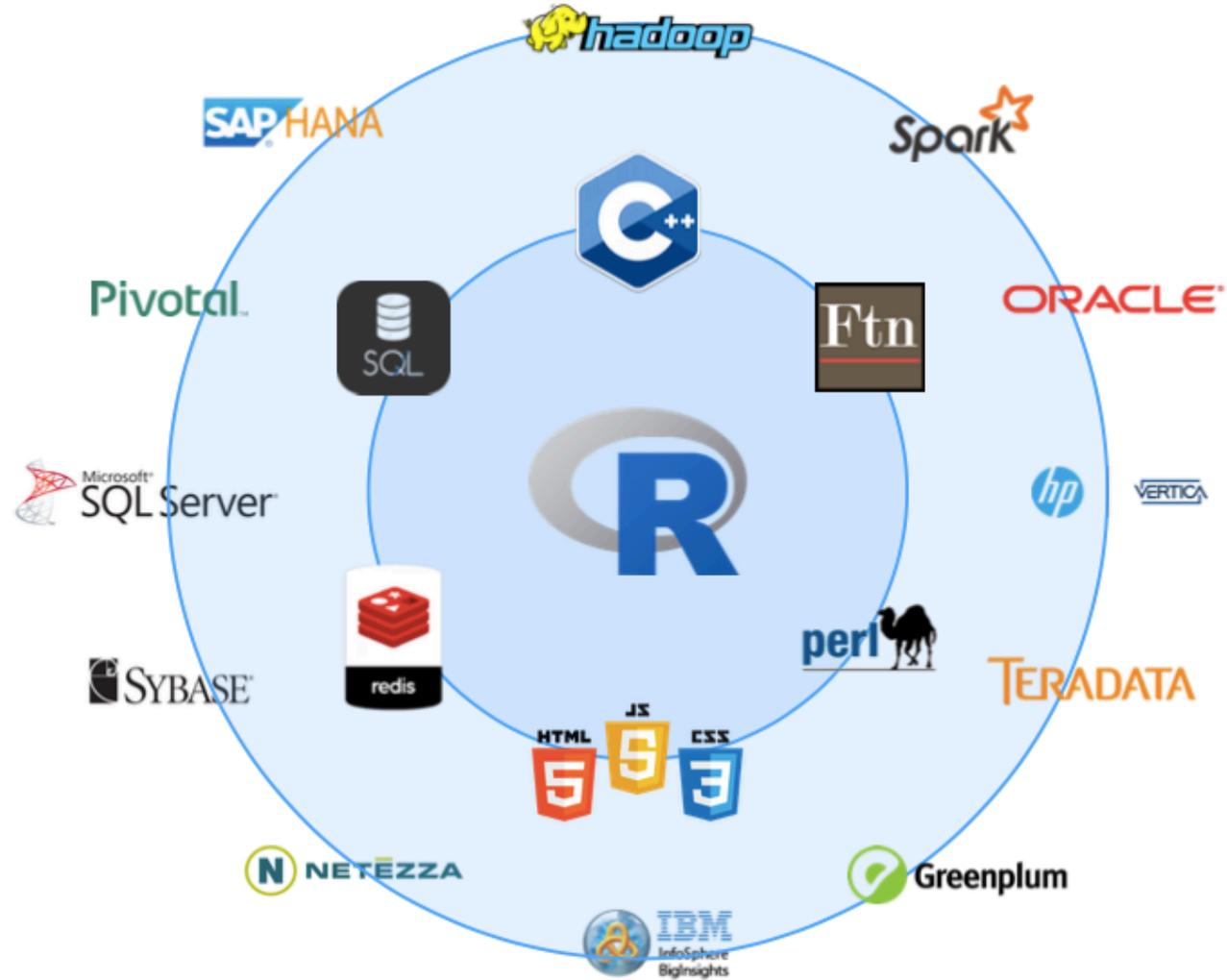


Twitter activities between March 12 and April 10, 2015



# Software

R Studio



# Software



# Planejamento

Audiência	Aquisição	Valor
Demografia Interesse Geo Localização Linguagem Dispositivo Rede Comportamento / Engajamento Tempo para análise	<p data-bbox="698 625 1847 882">Busca Orgânica                      Custo da Busca e Análise Acesso Direto ou por Referência (WS) Palavras-chave / Indexação</p> <p data-bbox="698 882 1847 1025"><b>Comportamento</b></p> <p data-bbox="698 1025 1847 1322">Rollback mais fácil Entender os pontos fracos Enteder os pontos fortes Acelerar os resultados</p>	Conversão (Sucesso) Caminho de Converção Perfomanace

# Direto ao Ponto

PROBLEMA	SOLUÇÃO: Dividir para Conquistar	
<b>Dados</b> <ul style="list-style-type: none"><li>- Acesso<ul style="list-style-type: none"><li>- Segurança</li><li>- Periodicidade</li><li>- Tipo de ...</li></ul></li><li>- Metadados</li><li>- Workflow</li><li>- Carga</li><li>- Disponibilidade</li><li>- Crescimento dos Dados</li></ul>	<b>Planejamento</b> <ul style="list-style-type: none"><li>• Data sources</li><li>• Databases metadata</li><li>• Métricas</li><li>• KPI</li></ul>	<ul style="list-style-type: none"><li>• Análise &amp; Modelagem para Predição</li><li>• Armazenamento</li></ul>
	<b>COMPORTAMENTO</b>	
	Rollback mais fácil Entender os pontos fracos Enteder os pontos fortes Acelerar os resultados	

# A Auditoria de Dados pode lhe dizer

Datasources	Databases & Arquivos de metadados	KPI						
Acesso Disponibilidade Qualidade dos Metadados Qualidade dos Dados Qualidade do Armazenamento Tempo de Carga Tempo de Disponibilidade de Conexão Espaço de Armazenagem Crescimento dos Dados Tamanho da Base	O propósito da criação da base Propósito do Dado Hora e Data de Criação Criador e Responsáveis pela manutenção do Dado Localização, Arquivos, Tamanho, Granularidade	Crescimento da Base Engajamento para Acerto de Dados sem Gravidade Custo/Benefício por Auditoria Efetividade da Auditoria						
	<h2>Métricas</h2> <table border="0"> <tr> <td>Total de Metadados / Campos / Bancos / Arquivos</td> <td>Total de Dados Missing / Inconsistentes</td> </tr> <tr> <td>Total de Registros processados</td> <td>Retenção de Qualidade</td> </tr> <tr> <td>% de Falhas / Objeto</td> <td>Custo por Processamento de Extração / Auditoria</td> </tr> <tr> <td>Taxa de Auditoria</td> <td>ROI</td> </tr> </table>		Total de Metadados / Campos / Bancos / Arquivos	Total de Dados Missing / Inconsistentes	Total de Registros processados	Retenção de Qualidade	% de Falhas / Objeto	Custo por Processamento de Extração / Auditoria
Total de Metadados / Campos / Bancos / Arquivos	Total de Dados Missing / Inconsistentes							
Total de Registros processados	Retenção de Qualidade							
% de Falhas / Objeto	Custo por Processamento de Extração / Auditoria							
Taxa de Auditoria	ROI							

# O Auditor de Dados também pode lhe falar ...

- **Predição e Estimativa**
- **Modelagem**



```
function( scope, element, attr, ngSwitchController ) {
  var selected = attr.ngSwitch || attr.on,
      selectedTranscludes = [],
      selectedElements = [],
      previousElements = [],
      selectedScopes = [];

  scope.$watch( watchExpr, function ngSwitchWatchAction( value ) {
    var i, ii;
    for ( i = 0, ii = previousElements.length; i < ii; ++i ) {
      previousElements[i].remove();
    }
    previousElements.length = 0;

    for ( i = 0, ii = selectedScopes.length; i < ii; ++i ) {
      var selected = selectedElements[i];
      selectedScopes[i].$destroy();
      previousElements[i] = selected;
      $animate.leave(selected, function() {
        previousElements.splice(i, 1);
      });
    }

    selectedElements.length = 0;
    selectedScopes.length = 0;

    if ( (selectedTranscludes = ngSwitchController.cases['!' + value] || ngSwitchController.defaultCase) ) {
      scope.$eval( attr.change );
      forEach( selectedTranscludes, function( selectedTransclude ) {
        var selectedScope = scope.$new();
        selectedScopes.push( selectedScope );
      });
    }
  });
}
```