UFPR - Universidade Federal do Paraná
SEPT - Setor de Educação Profissional e Tecnológica

IAA - Especialização em Inteligência Artificial Aplicada

IAA013 - Big Data

Prof. **João Eugenio** Marynowski – jeugenio@ufpr.br

Roteiro

- Apresentação do Professor
- Introdução ao Módulo
- Apresentação dos Alunos
- Big Data
- Utilização da VM Hadoop
- Ambiente Distribuído
- Programação MapReduce

Quem Sou Eu?

- João Eugenio Marynowski – jeugenio@ufpr.br
- Histórico/Experiência
  - ‘01 Bacharel em Ciência da Computação, UNIOESTE-Foz
  - ‘04 Mestre em Ciência da Computação, UFPR
  - Programador, Professor, Analista de Seg., DBA, Gerente TI, Empresário
  - ‘13 Doutor em Ciência da Computação, UFPR
  - ‘14 e ‘15 Pós-doutorado PUCPR/PPGIa, Segurança
- ‘16 Professor UFPR/SEPT/TADS
- Professor e pesquisador em BD e SD (Big Data)
  - SOLID e CORE
  - Extensão

Especialização em Inteligência Artificial Aplicada (IAA)

- Introdução à Inteligência Artificial
- Estatística Aplicada I
- Linguagem de Programação Aplicada
- Linguagem R
- Estatística Aplicada II
- Arquitetura de Dados
- Aprendizado de Máquina
- Visão Computacional
- **Big Data**
- Ciência de Dados Apl. Saúde
- Frameworks de IA
- Empreendedorismo e Inovação
- Metodologia Científica
- Tópicos em IA
- Seminários
- TCC

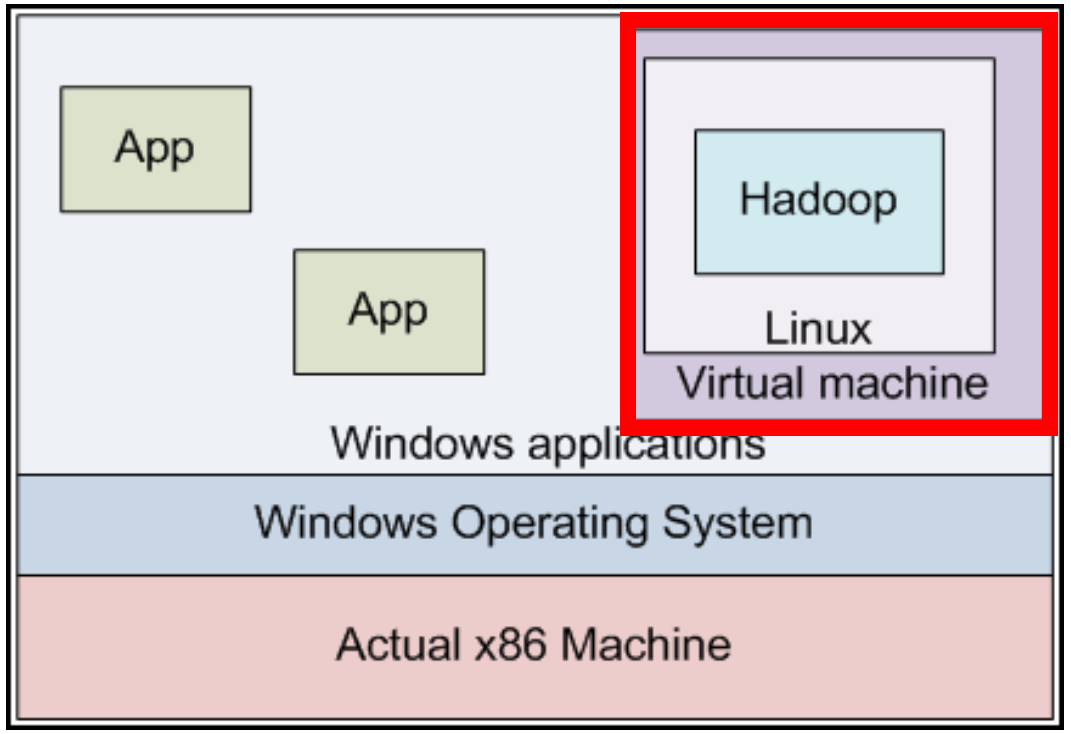
UFPR

Especialização em Inteligência Artificial

SEPT

Sistemas para as Aulas

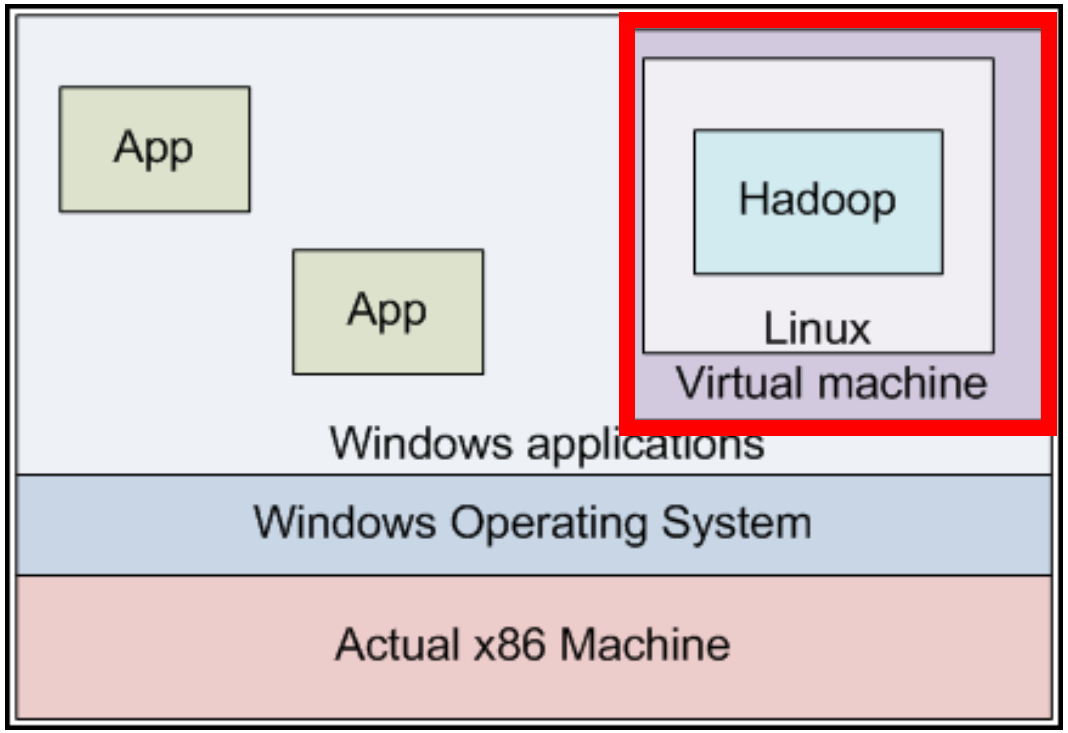
- Sandbox Hortonworks HDP 2.1



  - https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html
  - Bloqueada sem autenticação
  - Permite obter Sandbox mais recentes, mas exige muito recurso (20GB)
  - Então disponibilizei: Hortonworks_Sandbox_2.1.ova

Sistemas para as Aulas

- Sandbox Hortonworks HDP 2.1
- VirtualBox
- Putty e WinScp
- IDE Java (Netbeans)

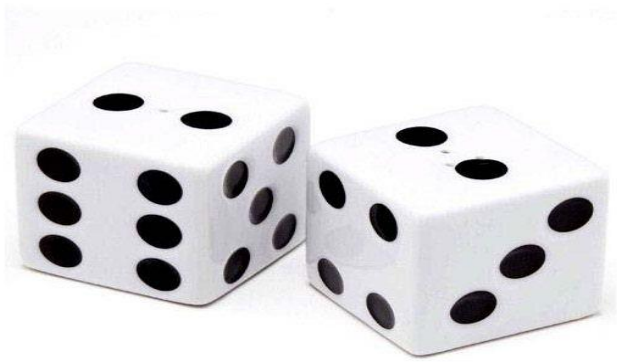


- VirtualBox
  - File, Import Appliance
    - Hortonworks_Sandbox_2.1.ova
    - CPU [2], RAM [4096 ou 2048]
    - Import

Big Data

Big Data

- Muitos dados vem em mente...



- Muitos dados vem em mente...

Big Data na Mídia

- Ah… Big Data é dados!!!
- Mas não estamos na “Era da Informação”?

- http://vimeo.com/31298658

Dado e a Tomada de Decisão

Sabedoria | Decisão

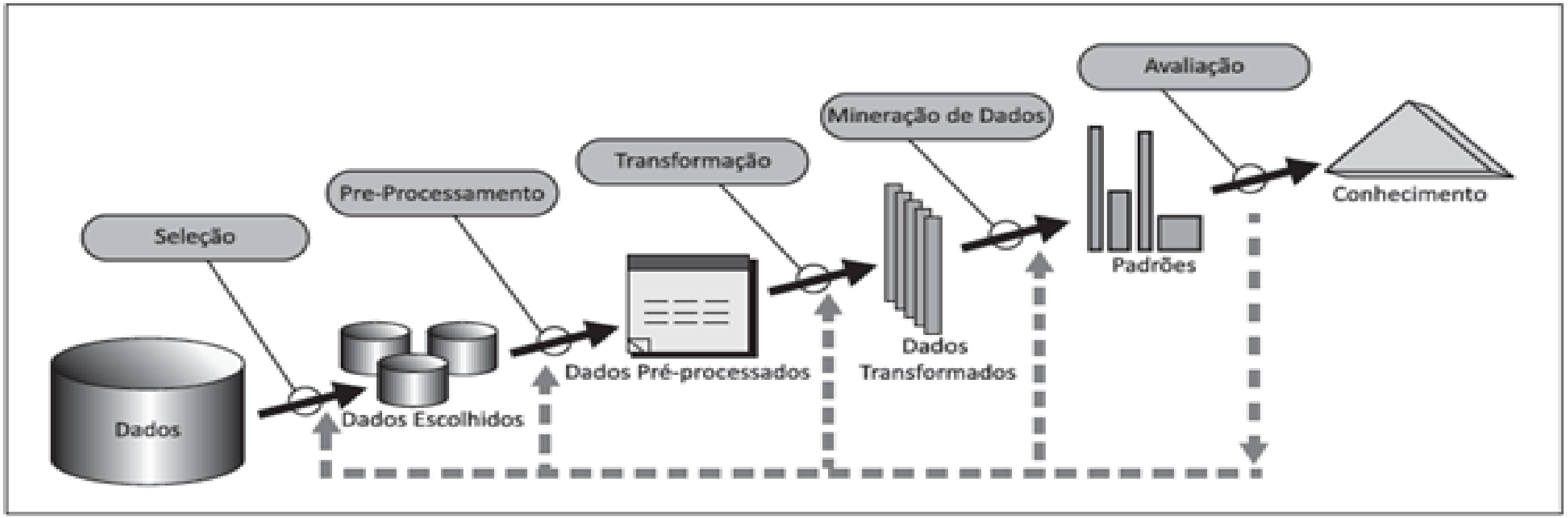
COMPRA!

Tendência de alta

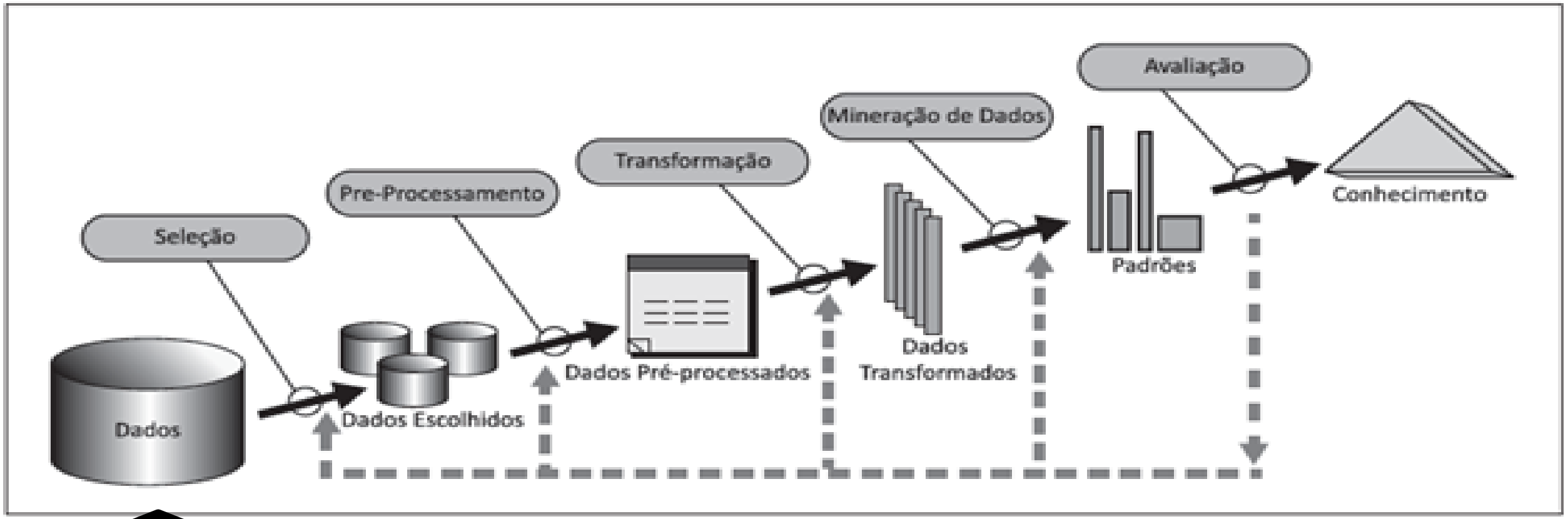
35000 pontos

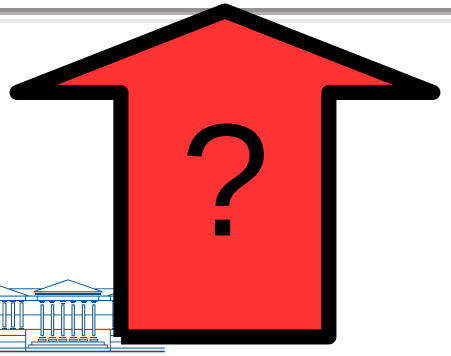
35000

Processo de Descoberta de Conhecimento
KDD - Knowledge Discovery Databases

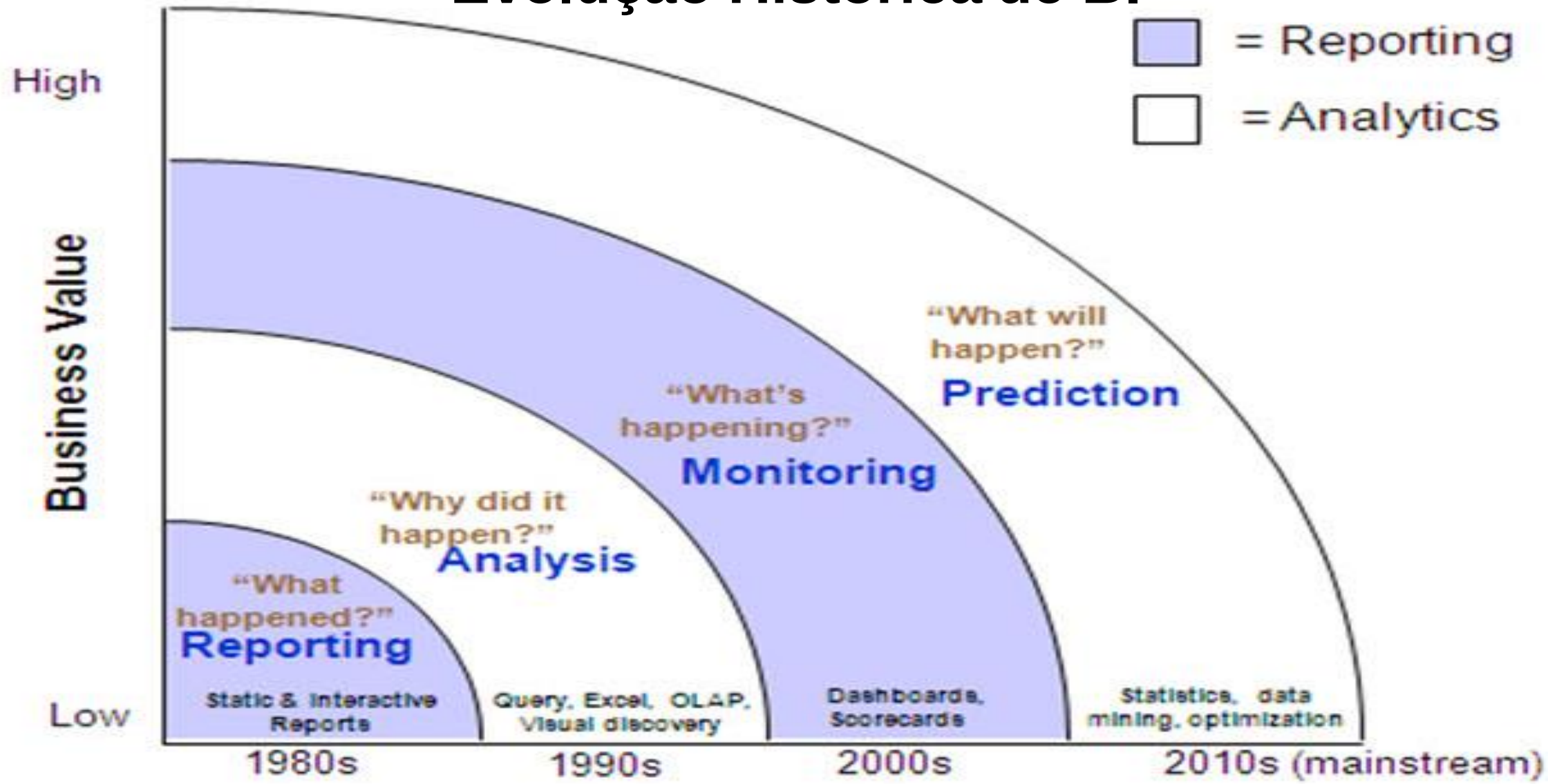


Processo de Descoberta de Conhecimento
KDD - Knowledge Discovery Databases





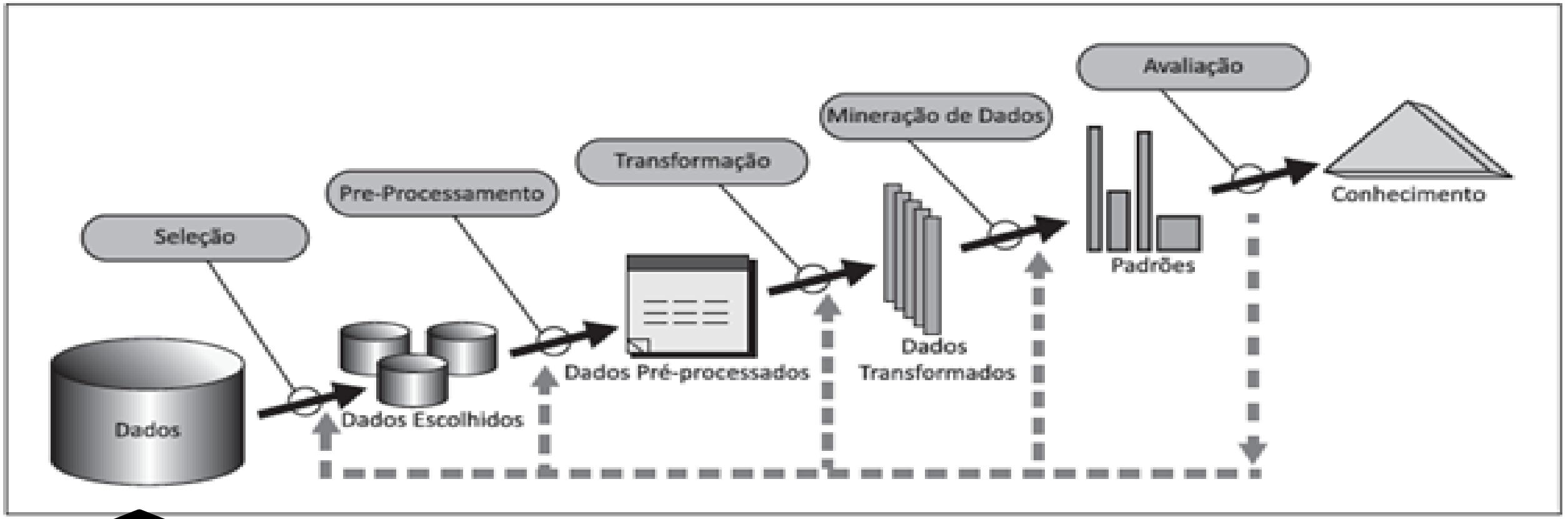
Evolução Histórica de BI



"Essencial"

"Grava Tudo"

Processo de Descoberta de Conhecimento
KDD - Knowledge Discovery Databases



- bi mensagens/semana – 7mil/s!
- mi transações/s
- bi de fotos

TB, PB, EB, ZB,...

Problema

Como processar um grande volume de dados?

1 PB – 1 computador (60MB/s) → 192 dias!

É necessário utilizar **VÁRIOS** computadores...

- FACILMENTE!

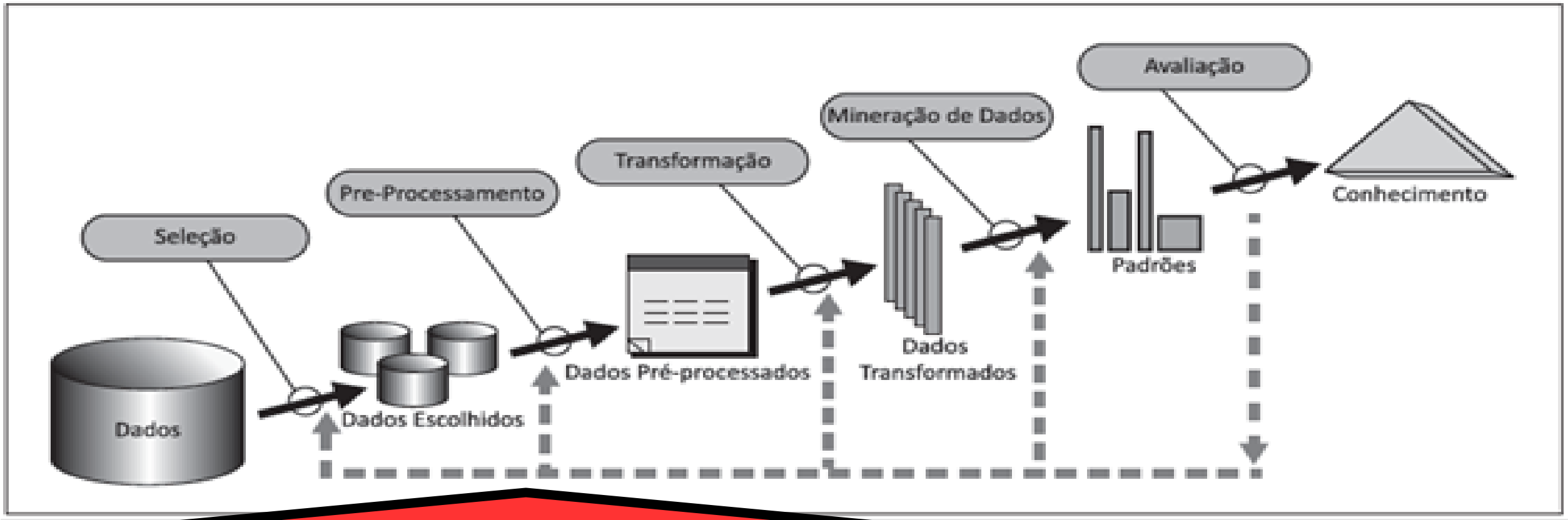
Vários Computadores...

100, 1.000, 10.000, ...

Sistemas Big Data

Sistemas

Processo de Descoberta de Conhecimento
KDD - Knowledge Discovery Databases



ETL?

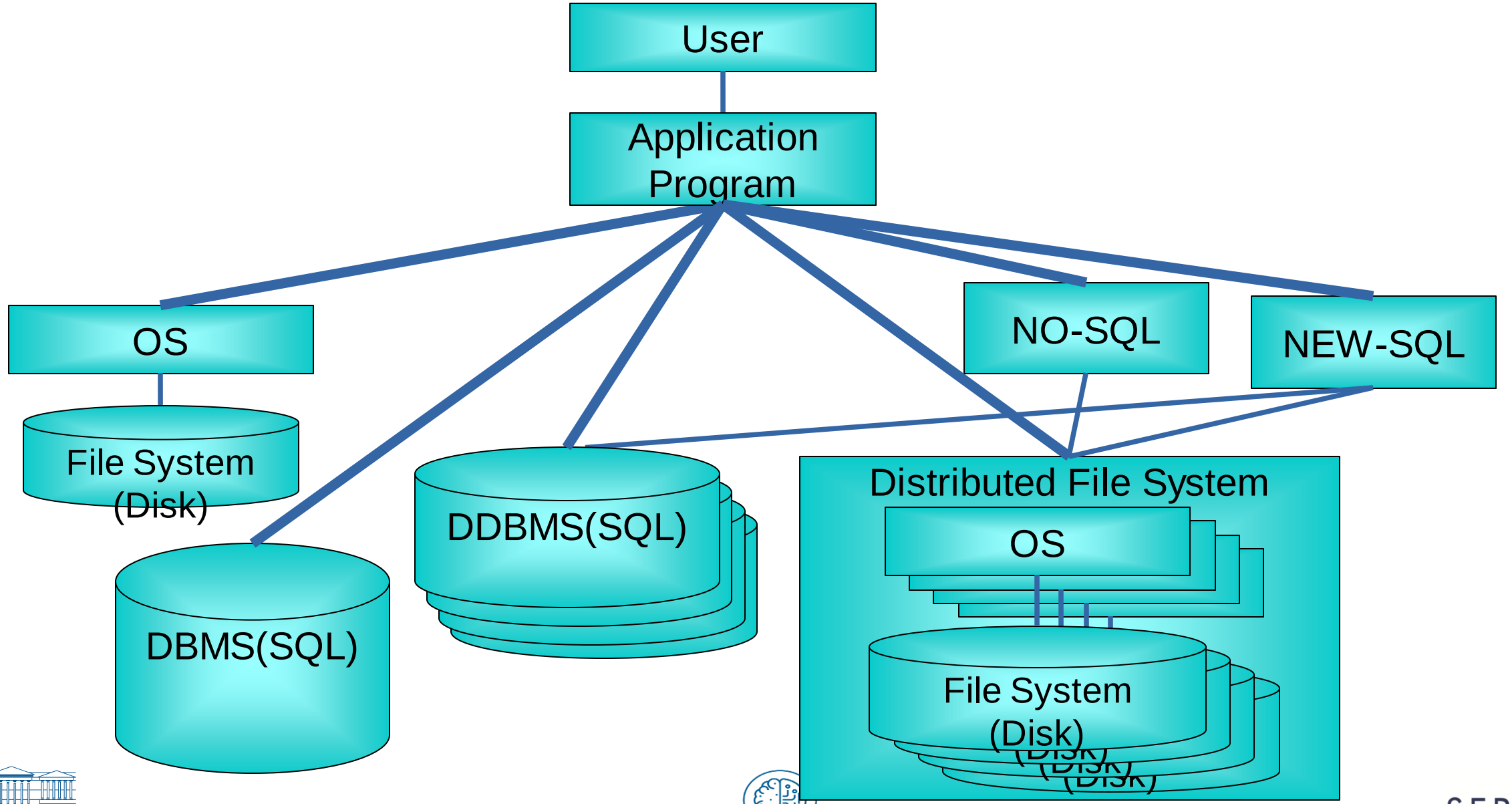
Modelos de Dados para Big Data

- Relacional, Objeto e O-R
- Dimensional/Multidimensional (DW)
- Dados Geográficos (GIS)
- Colunar
- Chave-valor
- Documento
- Grafo (Hierárquico e Rede)
- Não estruturados
  - Excel/CSV, Áudio, Vídeo, Logs, …

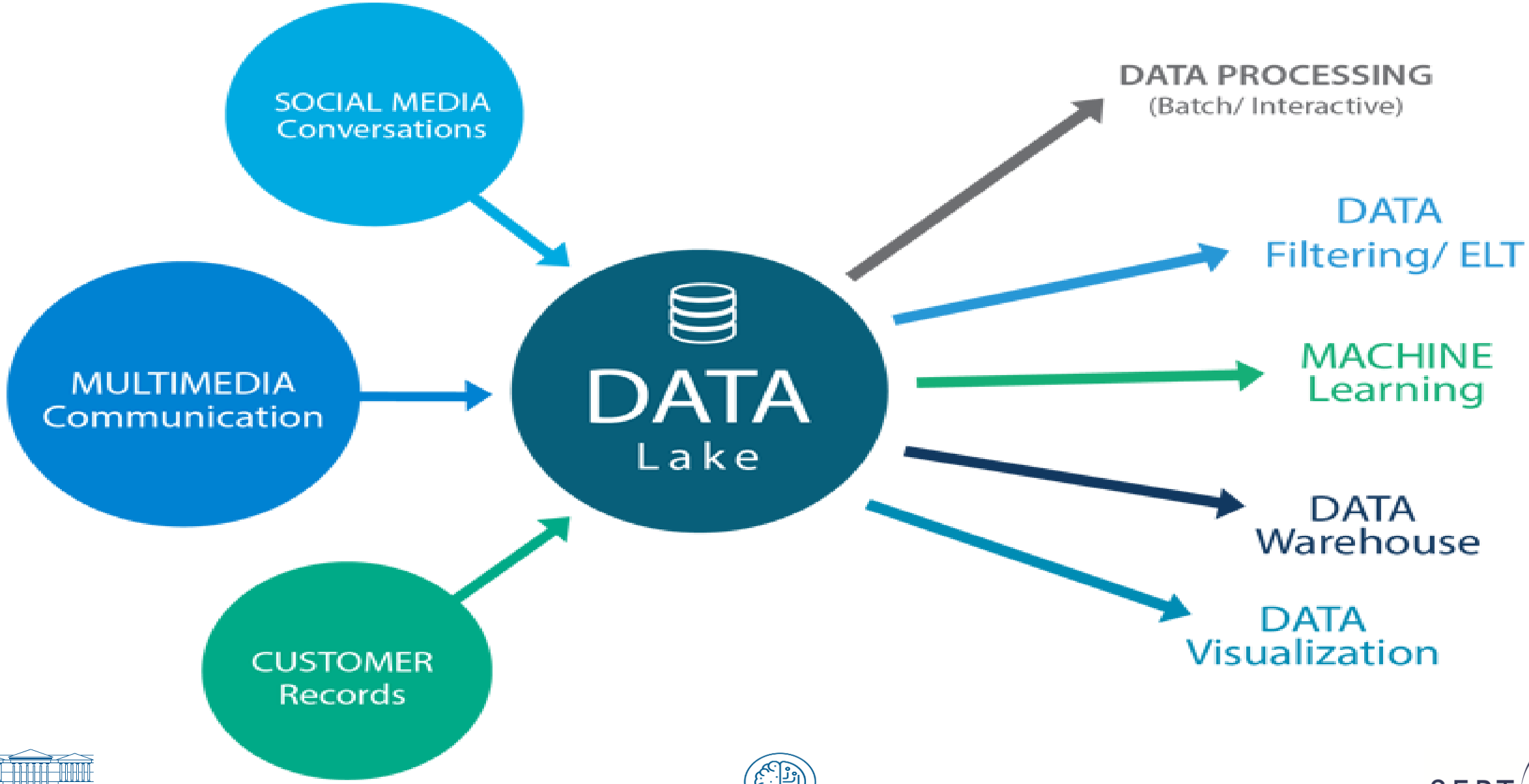
NoSQL

NewSQL

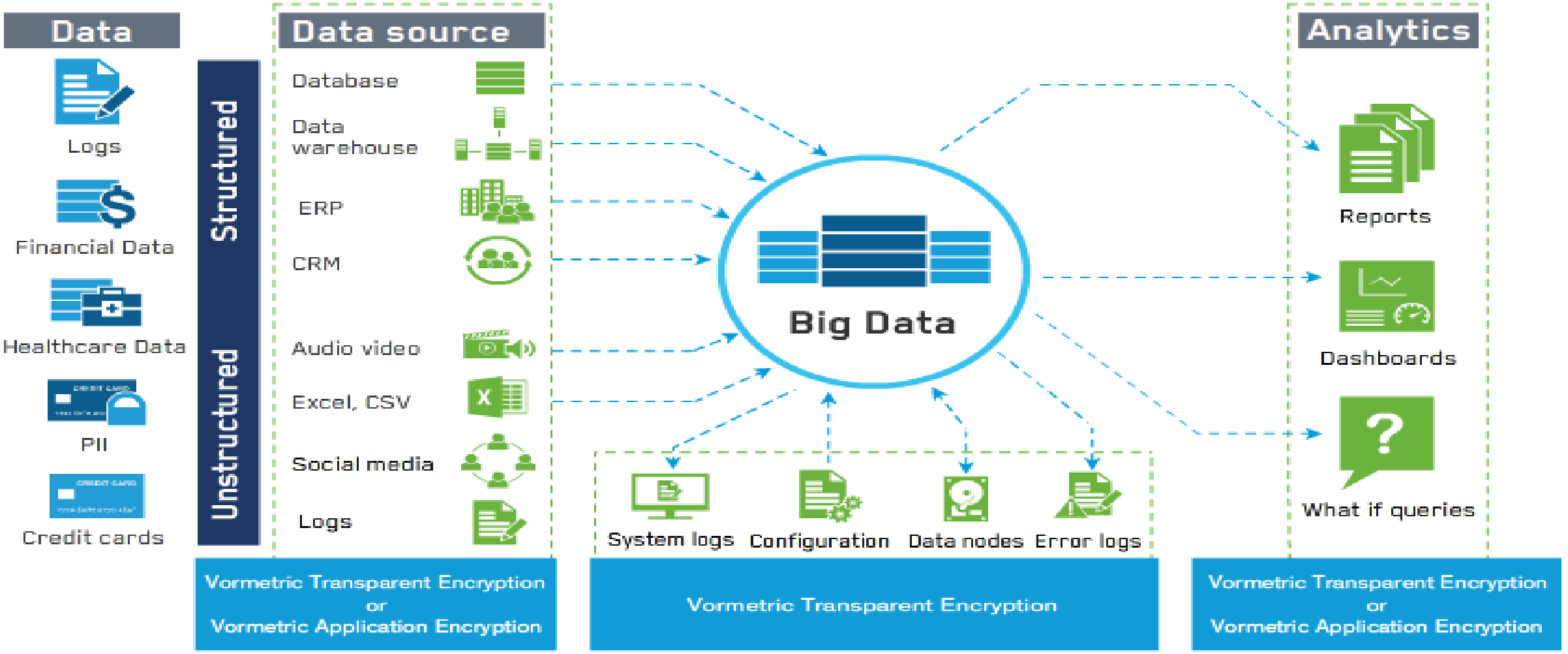
Sistemas Big Data



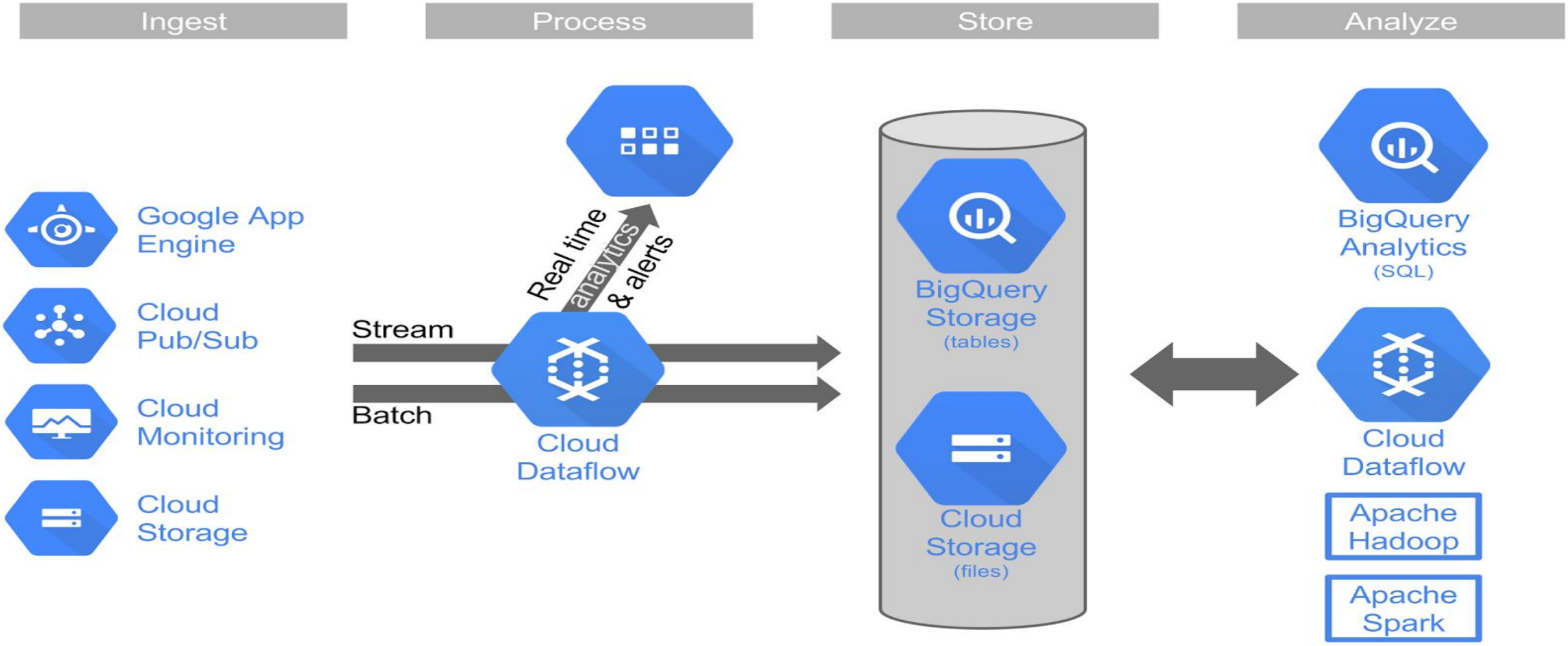
Data Lake


SOCIAL MEDIA
Conversations
MULTIMEDIA
Communication
CUSTOMER
Records
DATA
Lake
DATA PROCESSING
(Batch/ Interactive)
DATA
Filtering/ ELT
MACHINE
Learning
DATA
Warehouse
DATA
Visualization

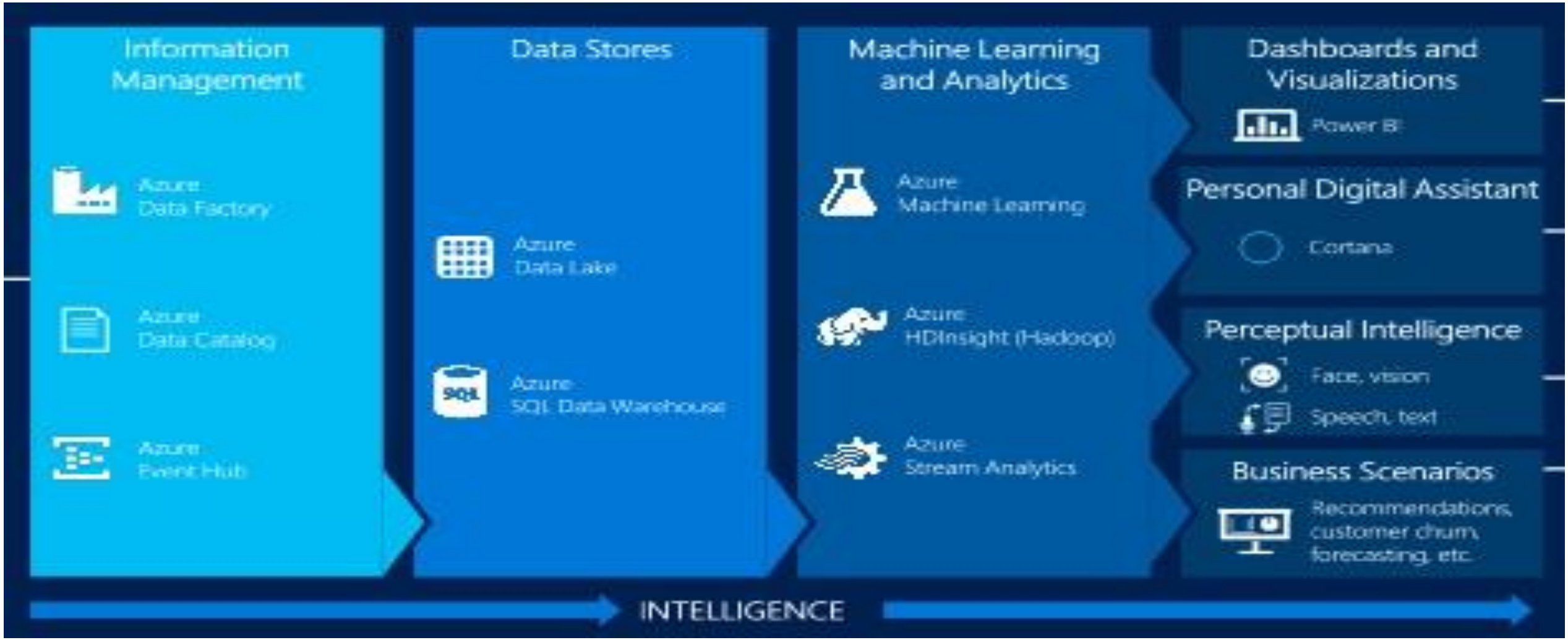
Big Data e Data Lake



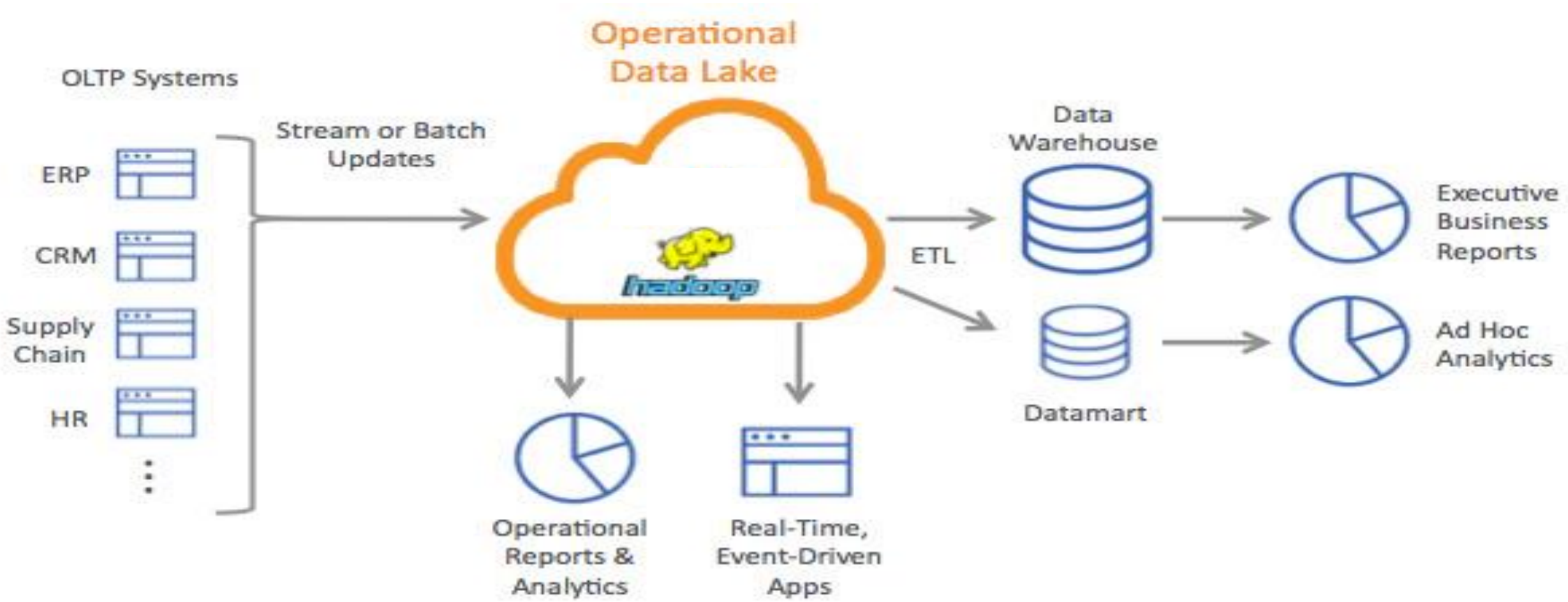
Google Big Data Cloud



Microsoft Data Analytics



Data Lake


OLTP Systems
ERP
CRM
Supply Chain
HR
Stream or Batch Updates
Operational Data Lake
hadoop
ETL
Data Warehouse
Executive Business Reports
Datamart
Ad Hoc Analytics
Operational Reports & Analytics
Real-Time, Event-Driven Apps

Programa

- Introdução à Big Data
- Data Engineering (Gerenciamento/Ferramentas para Big Data)
- NoSQL e NewSQL
- Dados em movimento – Processamento de Streaming

  - Utilização da VM hortonworks 2.1
  - Ambiente distribuído: DFS, MapReduce/Hadoop e Ambari
  - Personalização de aplicações MR
  - Ecosistema Hadoop HDP 2.1

Referências

- Hurwitz, J.; Nugent, A.; Dr. Halper, F.; Kaufman, M. **Big Data Para Leigos**. Alta Books Editora, 2016.
- White, T. (2012). **Hadoop: The Definitive Guide**, 3rd Edition (3rd ed., p. 688). O'Reilly Media. http://it-ebooks.info/book/635/
- Sakr, S., Liu, A., & Fayoumi, A. G. (2013). **The Family of MapReduce and Large-Scale Data Processing Systems**. ACM Computing Surveys, 46(1), 1–44.
- SILBERSCHATZ, A. KORTH, H; SUDARSHAN, S.. **Sistema de Banco de Dados**. 6.ed. Campus, 2012.
- Diversos Artigos
- Dean, J., & Ghemawat, S. (2004). **MapReduce: Simplified Data Processing on Large Clusters**. In Proc. of the OSDI - Symp. on Operating Systems Design and Implementation (pp. 137–149). USENIX.

Avaliação

- Presenças
  - 70% (aulas e/ou atividades)

- Atividades
  - Entrega via UFPR Virtual
  - Nota >= 70

Quando?

9 às 12h e das 13:30 às 17h

Sábados:
19/11
26/11
03/12

UFPR

Especialização em Inteligência Artificial

SEPT
SETOR DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
UNIVERSIDADE FEDERAL DO PARANÁ

Sandbox no Virtualbox

- File, Import Appliance
  - Hortonworks_Sandbox_2.1.ova
  - CPU [2], RAM [4096 ou 2048]
  - … Import

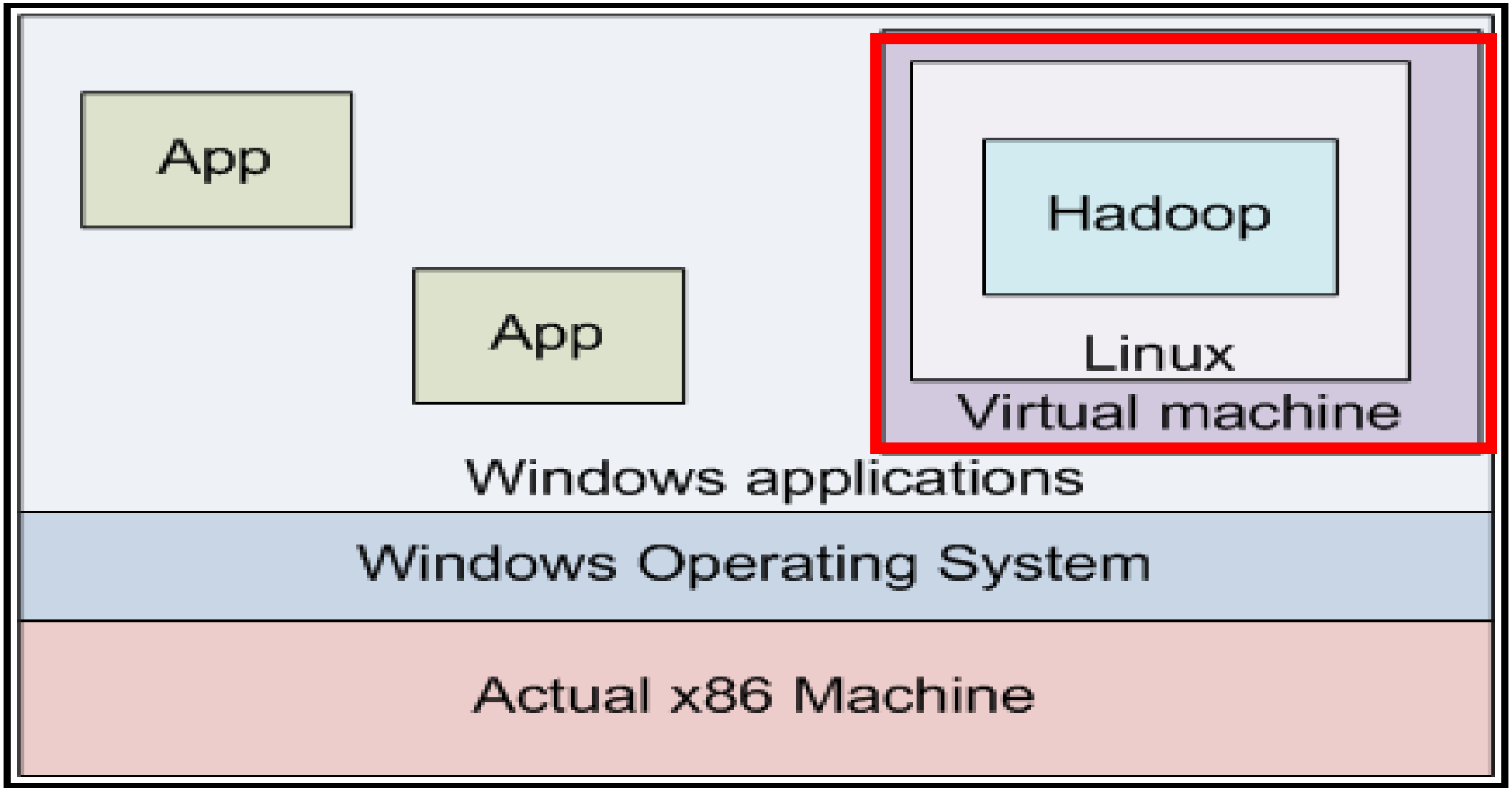
Quem São Vocês?

- Nome
- Experiência relacionada com Big Data
- Objetivos futuros

Sandbox no Virtualbox

- File, Import Appliance
  - Hortonworks_Sandbox_2.1.ova
  - CPU [2], RAM [4096 ou 2048]
  - … Import
- …
- **Start**

Sandbox HDP 2.1



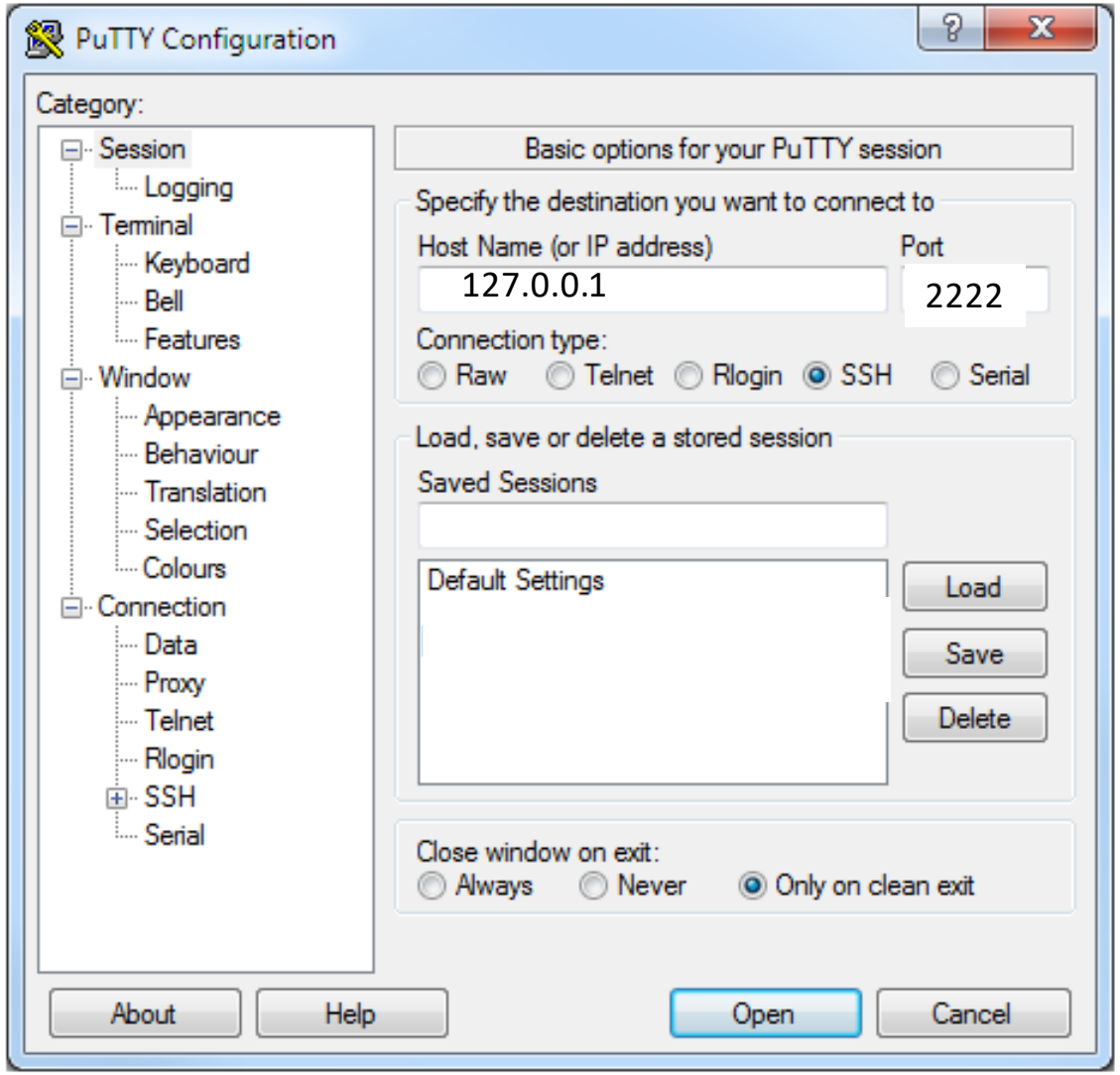
Acesso à Sandbox HDP

- Linux: CentOS, usuário **root** senha **hadoop**

- Acesso Local
  - Terminal → Alt + F5

- Acesso “Remoto” via ip 192.*.*.* ou 127.0.0.1
  - `ssh root@<IP> [-p 2222]`
  - **Putty (Windows)**

Putty


PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Serial
Basic options for your PuTTY session
Specify the destination you want to connect to
Host Name (or IP address)
Port
127.0.0.1
2222
Connection type:
Raw
Telnet
Rlogin
SSH
Serial
Load, save or delete a stored session
Saved Sessions
Default Settings
Load
Save
Delete
Close window on exit:
Always
Never
Only on clean exit
About
Help
Open
Cancel

Atividade 1 (Salvem prints da tela para envio)

1) `$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 1 100`

2) `$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 1 500000`
3) `$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 5 100000`

- Comentem sobre a diferença de execução dos exercícios 2 e 3

(Quais diferenças? Tempo? Por que?)

Atividade 1 (Salvem prints da tela para envio) FEITO!

1) `$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 1 100`

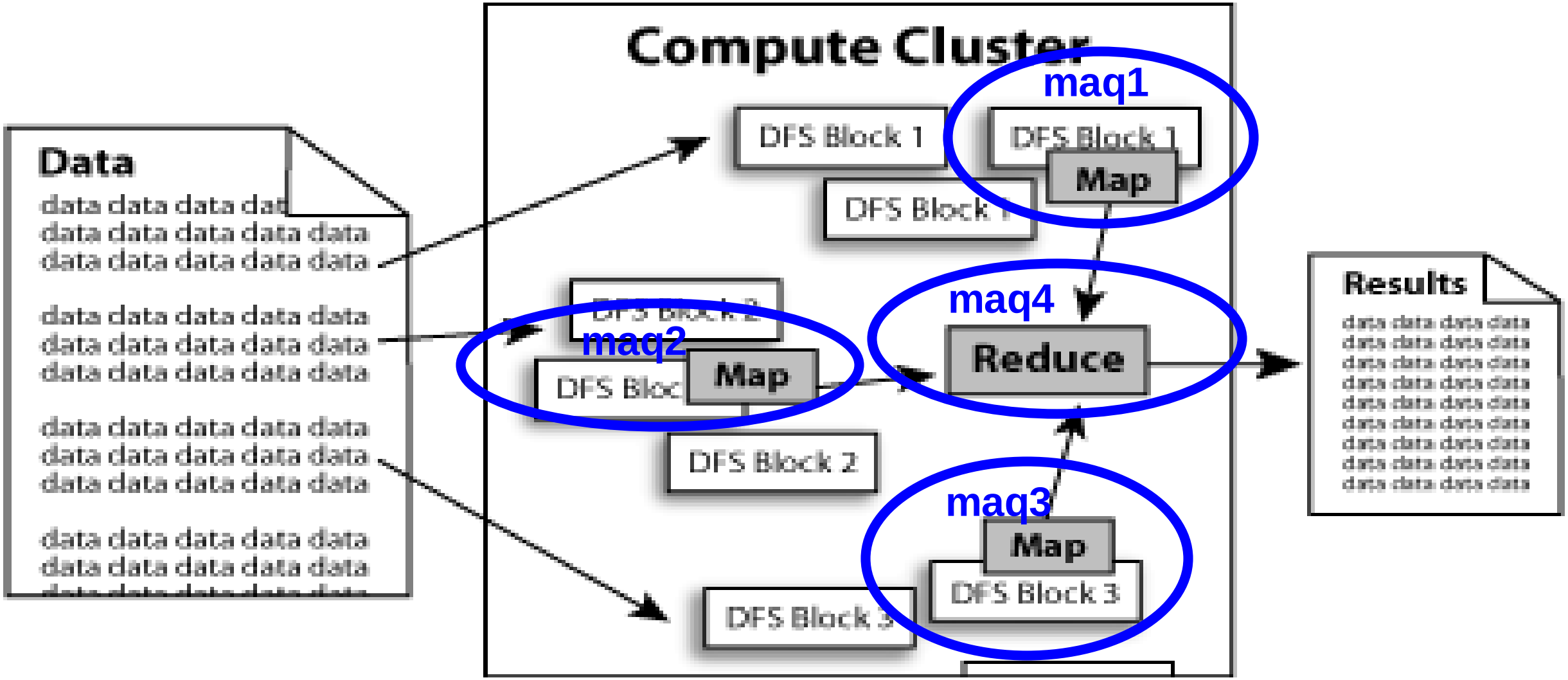
2) `$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 1 500000`
3) `$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 5 100000`

- Comentem sobre a diferença de execução dos exercícios 2 e 3

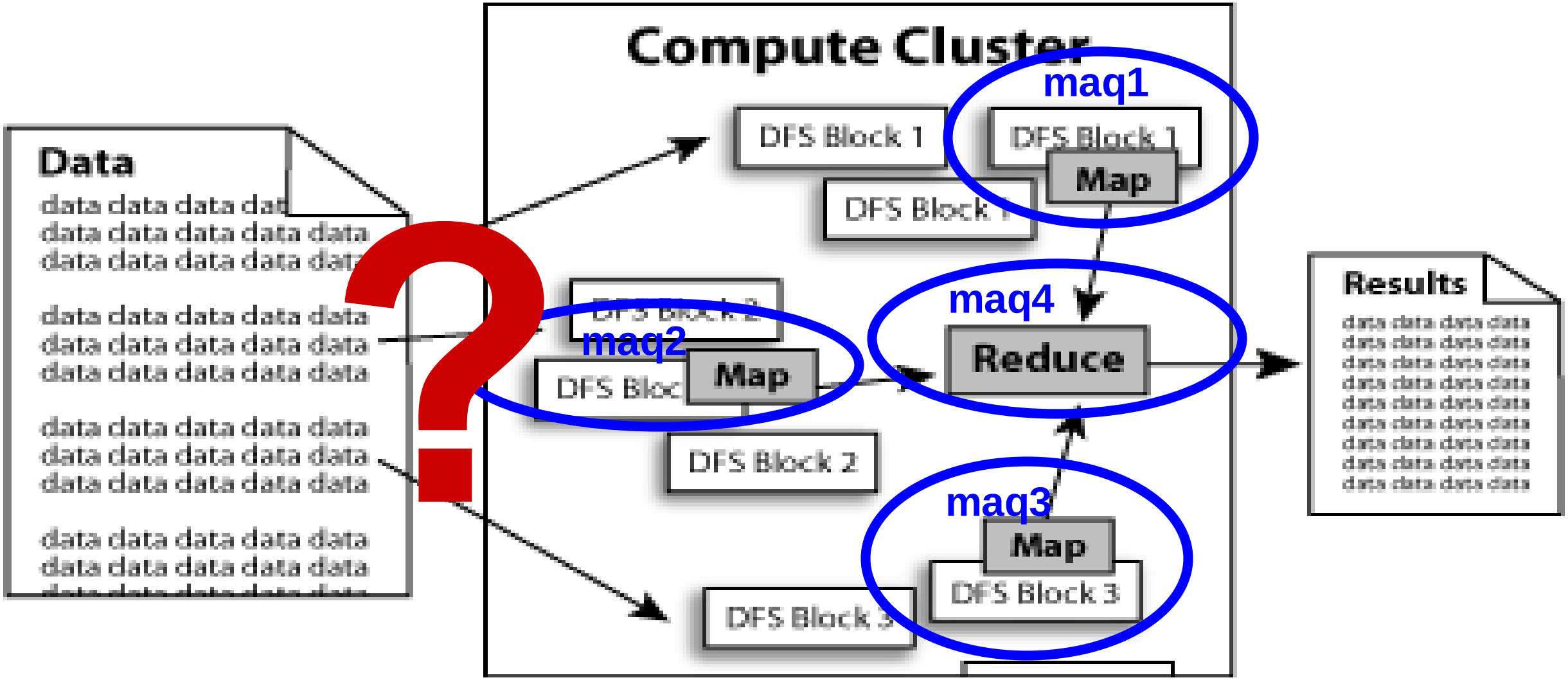
(Quais diferenças? Tempo? Por que?)

Como processo um arquivo?

Distributed File System (DFS) + MapReduce



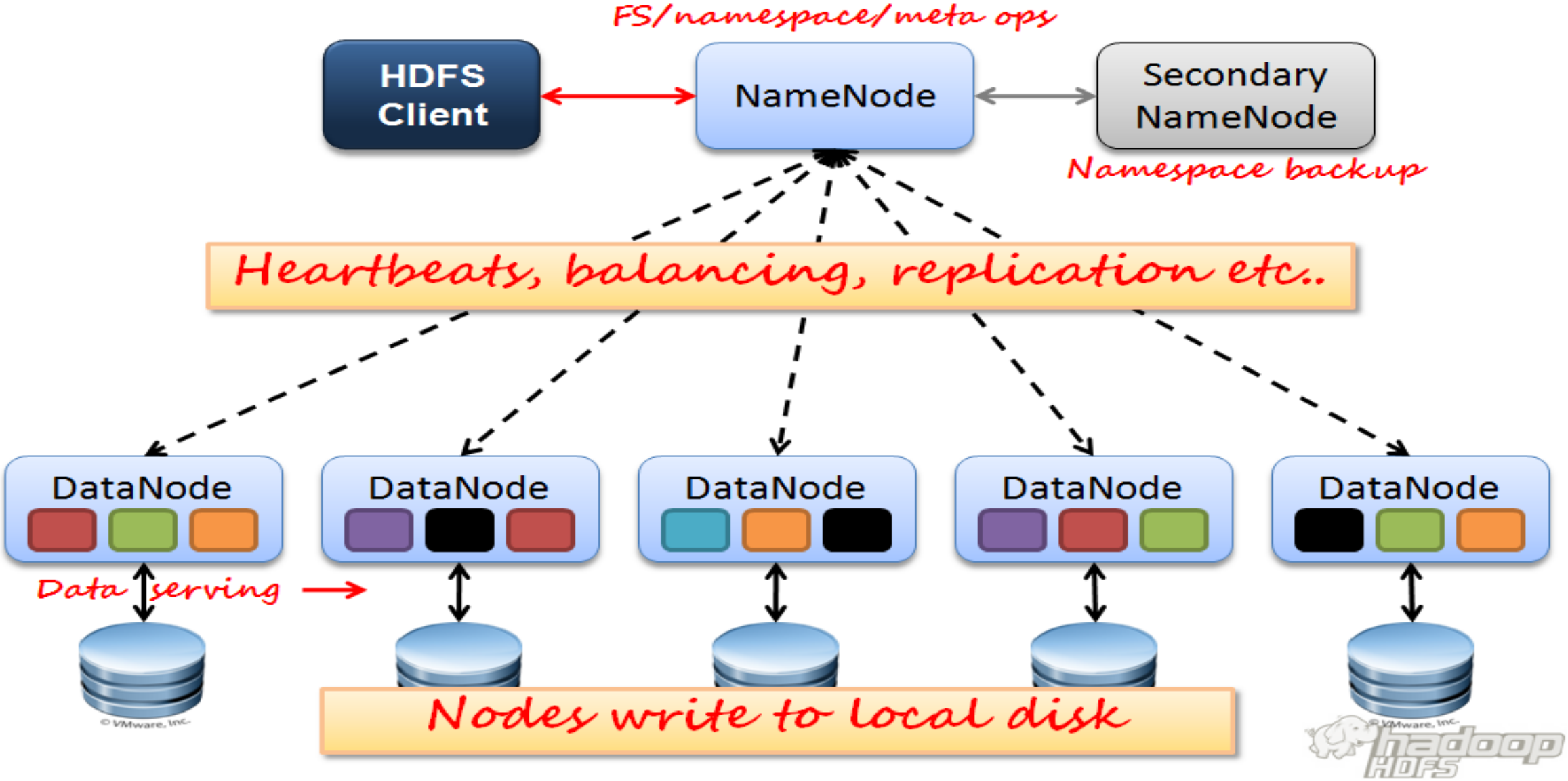
Distributed File System (DFS) + MapReduce

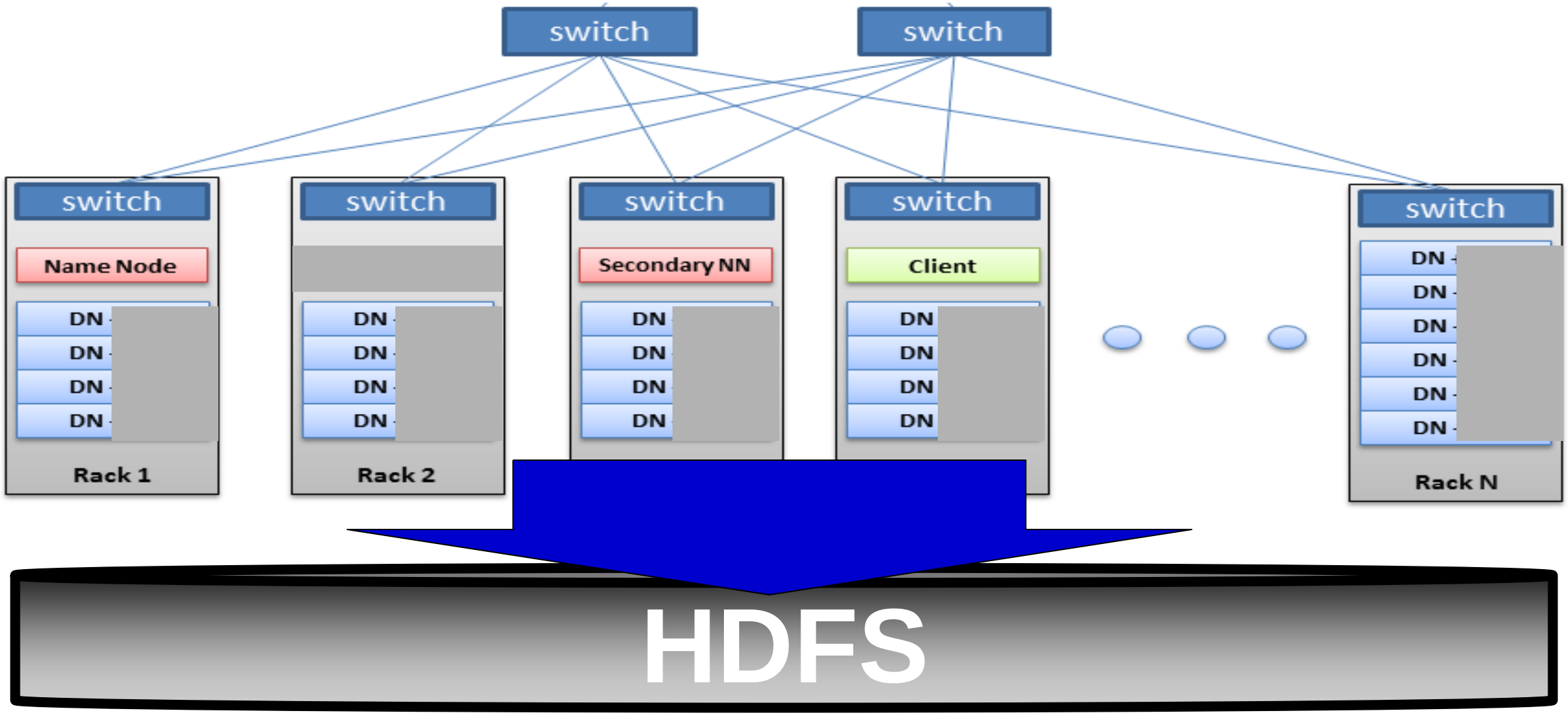


GFS e HDFS

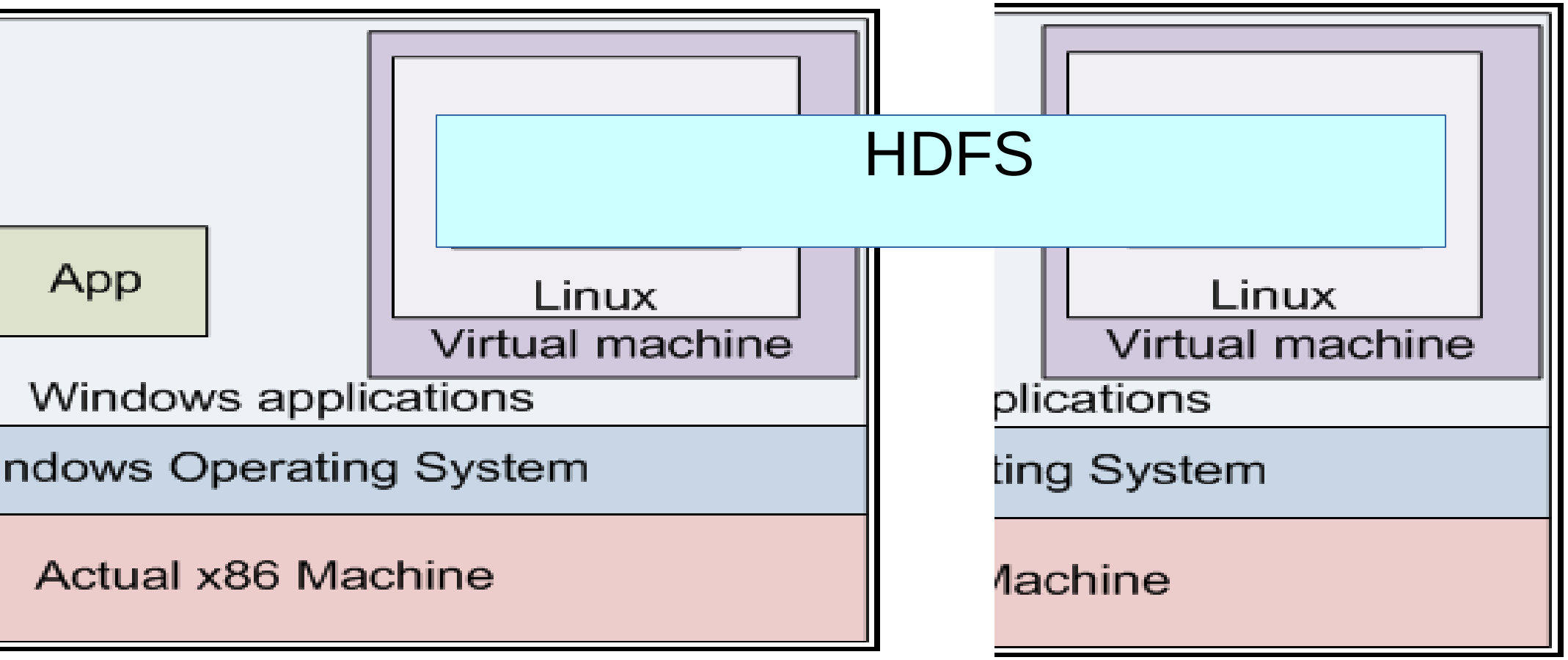
- **The Hadoop Distributed File System (HDFS)**
  - Permite acesso como um disco único
  - Executa sobre sistema de arquivo nativo (EXT3, EXT4, XFS,…)
  - Tolerante a falhas (Máquina, Rede, Disco, ...)
- Usa máquinas “comuns”
  - 2003: 669 dual 1.4 GHz, 2GB, 80GB
  - 2010: +25mil/3500 quad xeon 2.5GHz, 16GB, 4x1TB, 25 PB

Arquitetura do HDFS




switch
switch
switch
Name Node
DN
DN
DN
DN
Rack 1
switch
DN
DN
DN
DN
Rack 2
switch
Secondary NN
DN
DN
DN
DN
switch
Client
DN
DN
DN
DN
switch
DN
DN
DN
DN
DN
DN
Rack N
HDFS

Sandbox Hadoop x2, 3, 4,...



Características do HDFS

- Armazenamento de grandes arquivos
  - GB, TB, PB, EB, ...
  - Milhões ao invés de bilhões de arquivos
  - (arquivo > 128 MB – bloco padrão)
- Escreve um e lê vários (write once and read-many)
- Leitura Sequencial (streaming)
  - Ruim para leitura aleatória
  - Operação de adição ao final (Append)
- Alta vazão (throughput)
  - Alta latência para pequenos pedaços (chunks)
  - (HBase trata bem muitos chunks)

Comandos do HDFS

- **HDP 2.1** → $ hdfs dfs
- Shell commands format

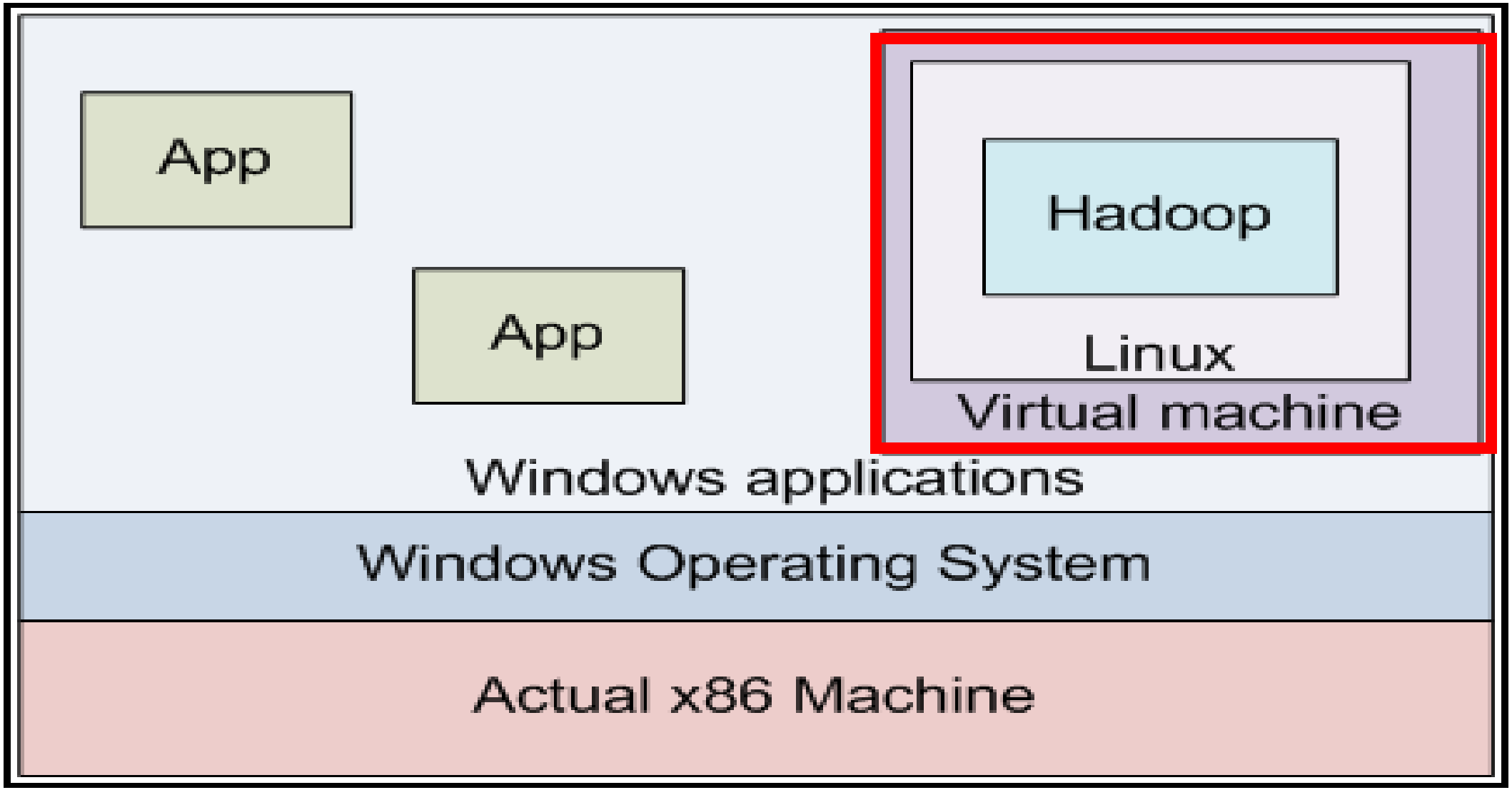
```
$ hdfs dfs -<command> -<option> <path>
$ hdfs dfs -ls /
$ hdfs dfs -ls hdfs://localhost/to/path/dir
$ hdfs dfs -ls file:///to/path/file3
```

- Unix like (cat, du, rm, cp, mkdir, mv, put, get, tail, chmod,...)

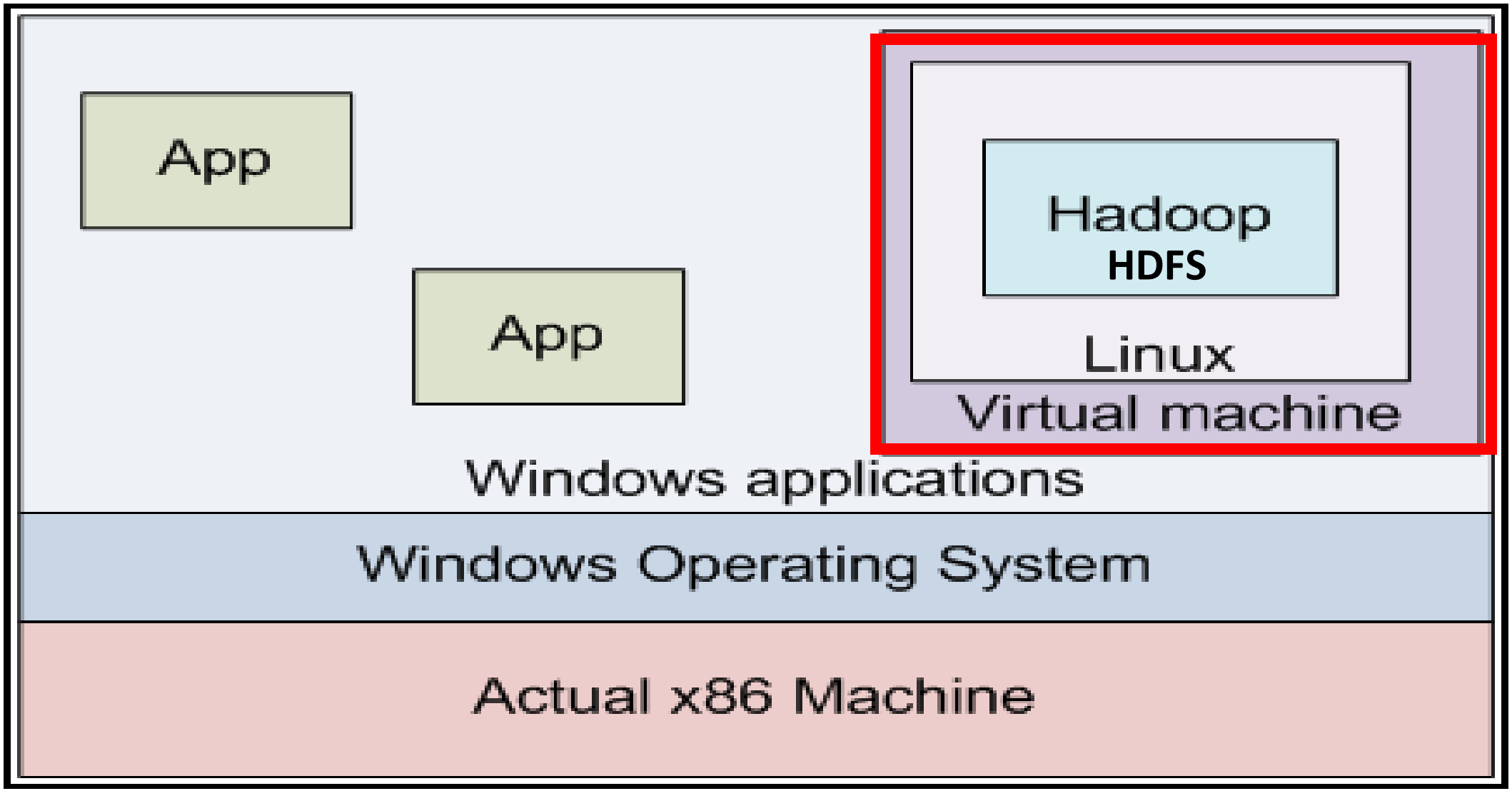
```
$ hdfs dfs -help
$ hdfs dfs -help <command_name>
```

Sandbox Hadoop

HDFS ???



Sandbox Hadoop



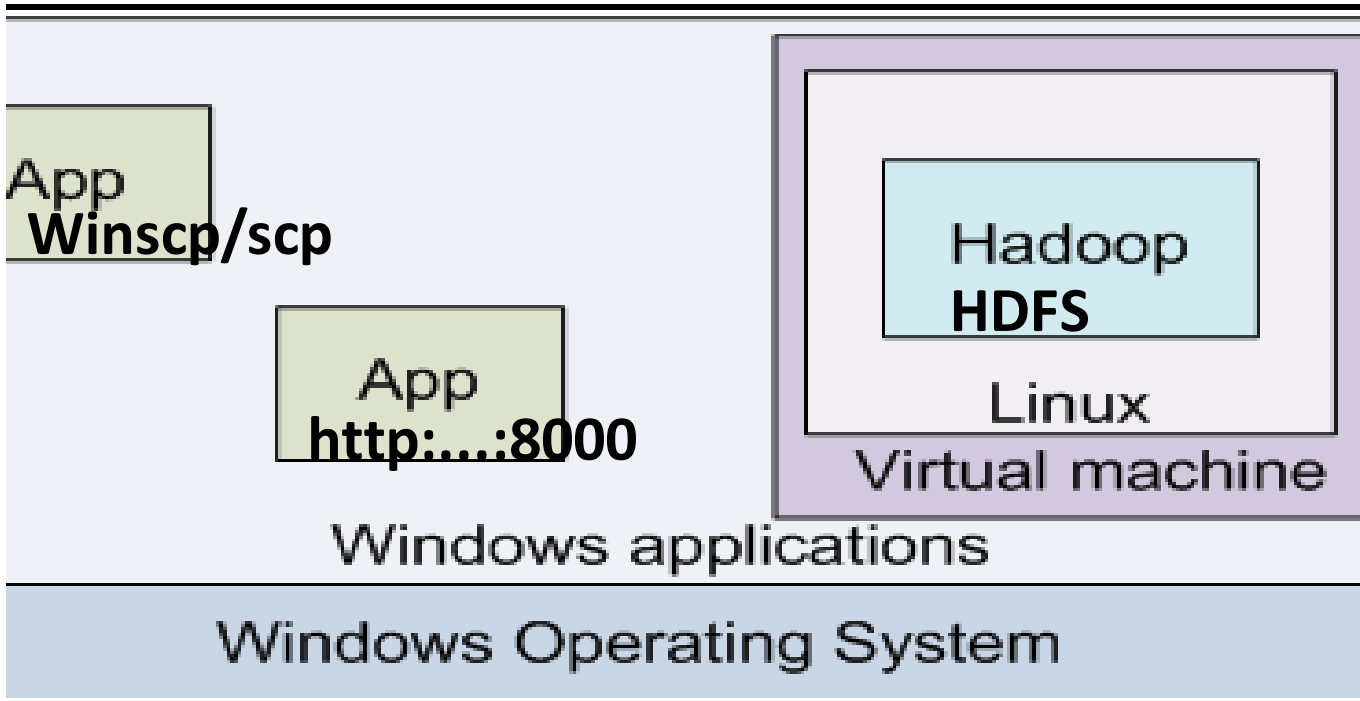
Atividade 2 – Executar o Contador de Palavras (wordcount)

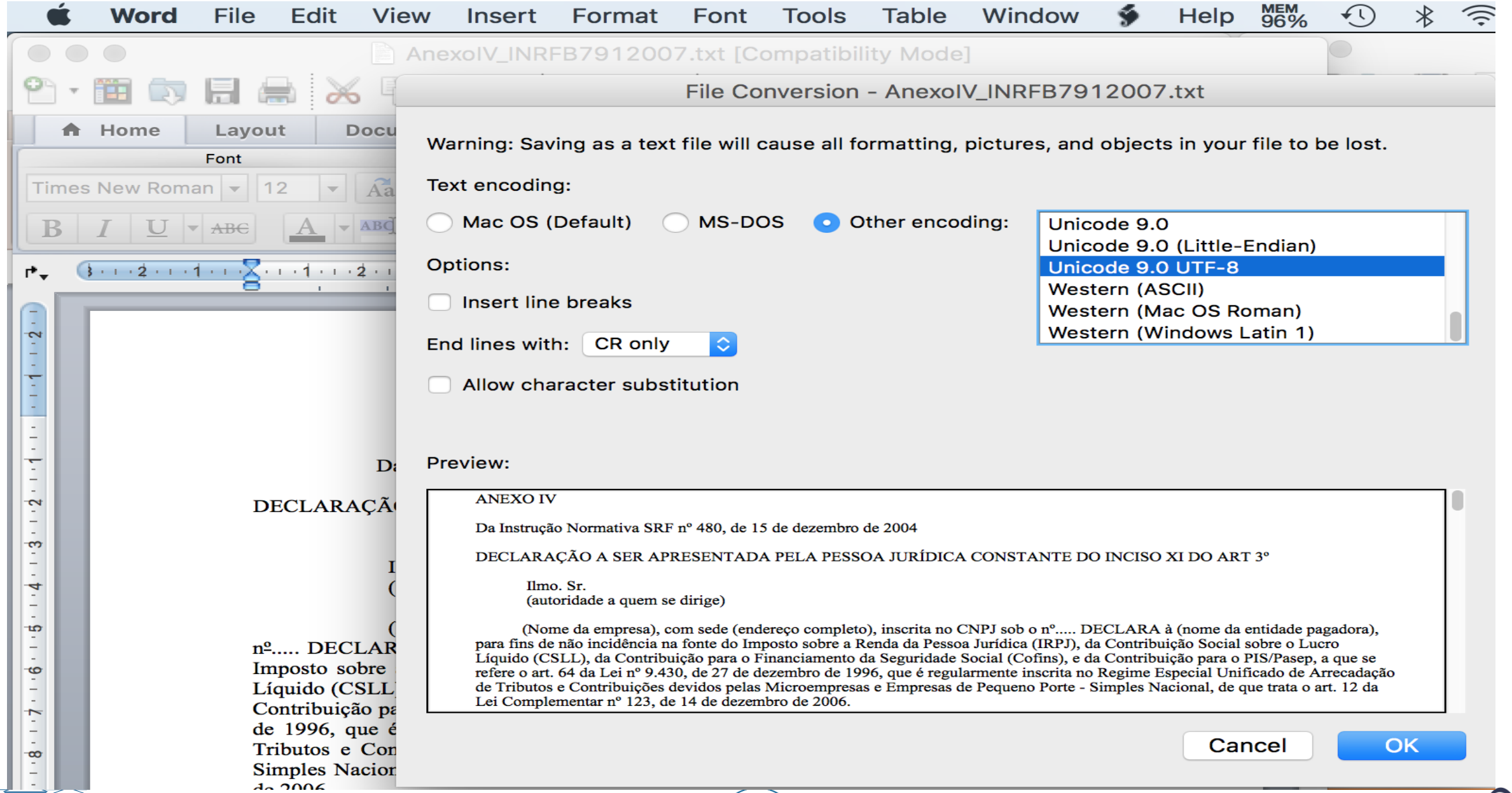
```
$ hdfs dfs -ls /
$ hdfs dfs -put /var/log/mysqld.log /
$ hdfs dfs -ls /
$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-
mapreduce-examples.jar wordcount /mysqld.log /saida
$ hdfs dfs -cat /saida/part*
$ hdfs dfs -get /saida /root/
$ ls -l /root/saida
$ cat /root/saida/part*
```

Como processo um arquivo da minha máquina?

Desktop, VM, HDFS

- Desktop → VM
  - Windows: Winscp (portable)
  - Linux: scp
- VM → HDFS
  - $ hdfs dfs -[put|get] *
- Desktop → HDFS
  - HUE, filebrowser
  - http://<ip>:8000/filebrowser



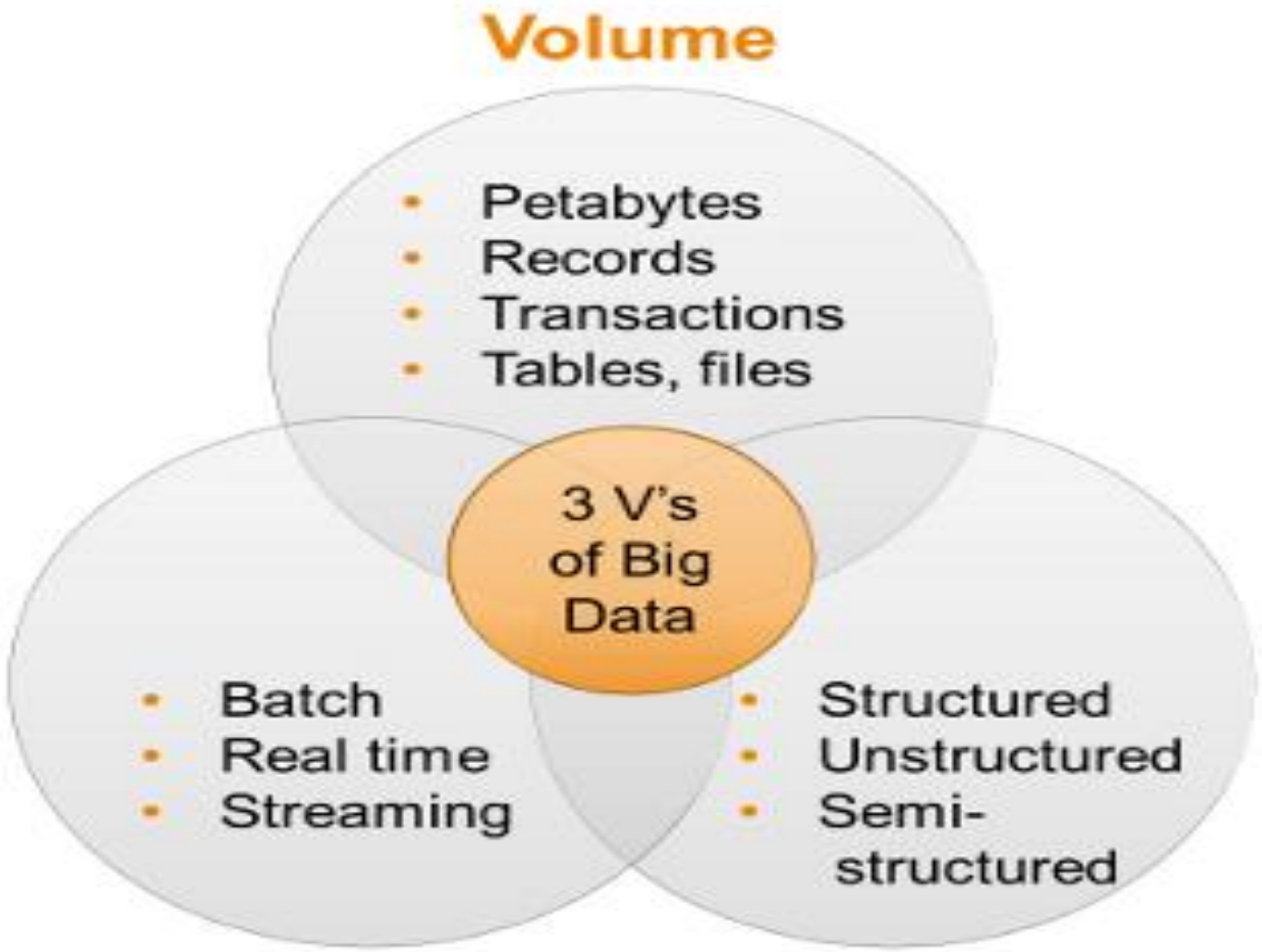

Word File Edit View Insert Format Font Tools Table Window Help MEM 96%
AnexoIV_INRFB7912007.txt [Compatibility Mode]
Home Layout Docu
Font
Times New Roman 12
File Conversion - AnexoIV_INRFB7912007.txt
Warning: Saving as a text file will cause all formatting, pictures, and objects in your file to be lost.
Text encoding:
Mac OS (Default)
MS-DOS
Other encoding:
Unicode 9.0
Unicode 9.0 (Little-Endian)
Unicode 9.0 UTF-8
Western (ASCII)
Western (Mac OS Roman)
Western (Windows Latin 1)
Options:
Insert line breaks
End lines with: CR only
Allow character substitution
Preview:
ANEXO IV
Da Instrução Normativa SRF nº 480, de 15 de dezembro de 2004
DECLARAÇÃO A SER APRESENTADA PELA PESSOA JURÍDICA CONSTANTE DO INCISO XI DO ART 3º
Ilmo. Sr.
(autoridade a quem se dirige)
(Nome da empresa), com sede (endereço completo), inscrita no CNPJ sob o nº..... DECLARA à (nome da entidade pagadora), para fins de não incidência na fonte do Imposto sobre a Renda da Pessoa Jurídica (IRPJ), da Contribuição Social sobre o Lucro Líquido (CSLL), da Contribuição para o Financiamento da Seguridade Social (Cofins), e da Contribuição para o PIS/Pasep, a que se refere o art. 64 da Lei nº 9.430, de 27 de dezembro de 1996, que é regularmente inscrita no Regime Especial Unificado de Arrecadação de Tributos e Contribuições devidos pelas Microempresas e Empresas de Pequeno Porte - Simples Nacional, de que trata o art. 12 da Lei Complementar nº 123, de 14 de dezembro de 2006.
Cancel
OK

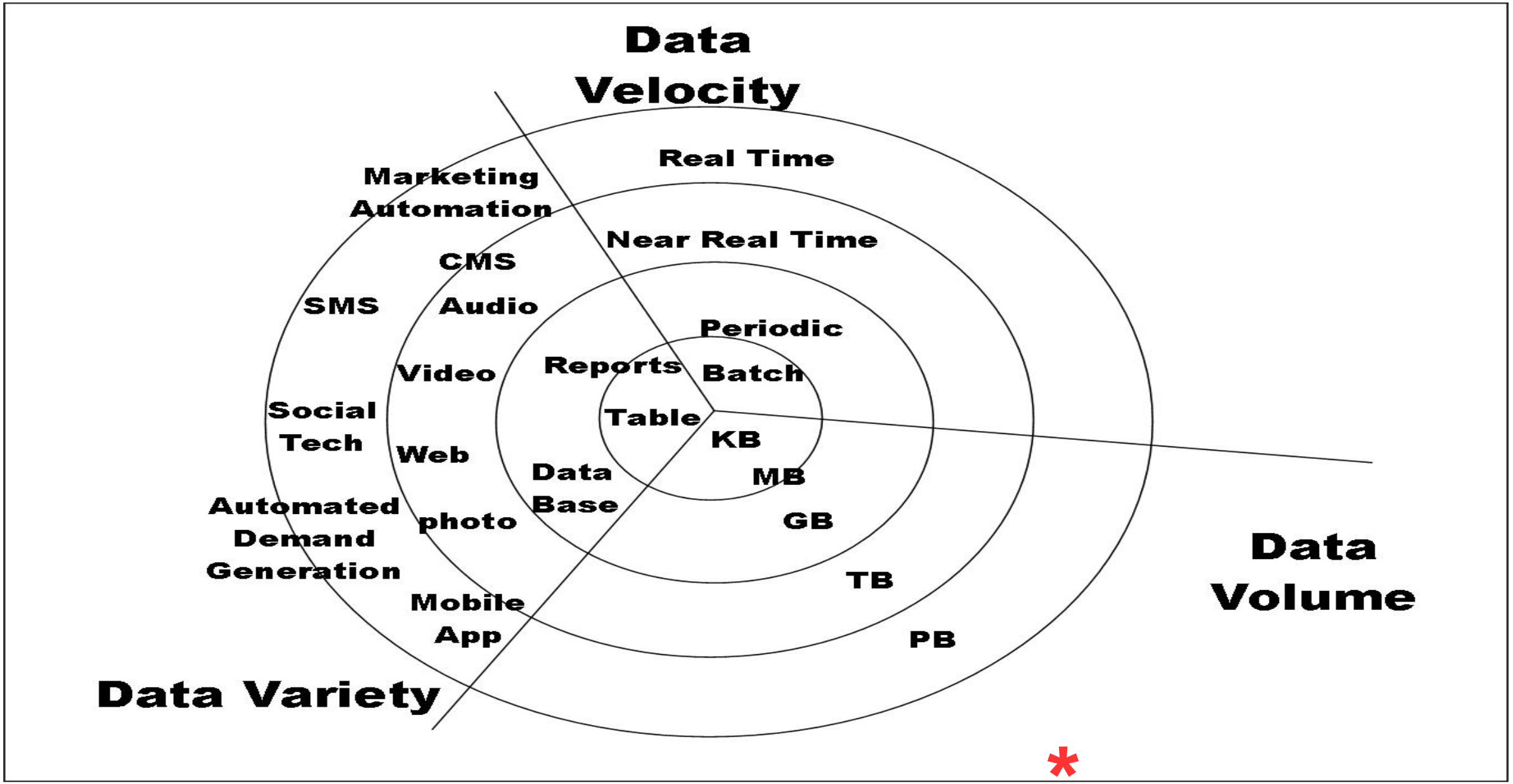
Atividade 3

Executar o *wordcount* de um arquivo do seu computador.

BIG DATA

Big Data ‘00

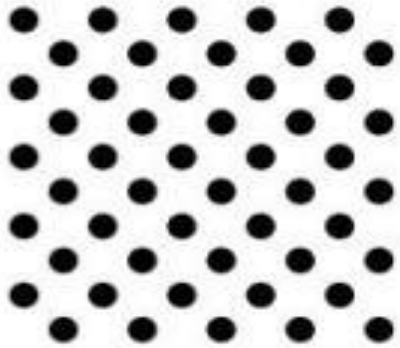

Volume
Petabytes
Records
Transactions
Tables, files
3 V’s of Big Data
Batch
Real time
Streaming
Structured
Unstructured
Semi-structured
Velocity
Variety


Data
Velocity
Real Time
Marketing
Automation
Near Real Time
CMS
SMS
Audio
Periodic
Reports
Batch
Video
Social
Tech
Table
KB
Web
Data
Base
MB
Automated
Demand
Generation
photo
GB
Data
Volume
TB
Mobile
App
PB
Data Variety

*

4 Vs

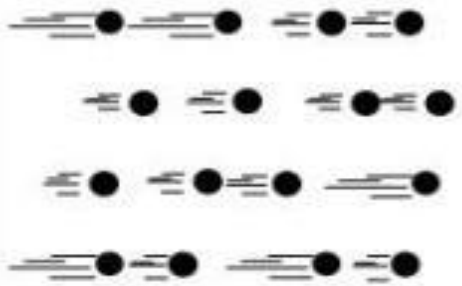
Volume



Data at Rest

Terabytes to exabytes of existing data to process

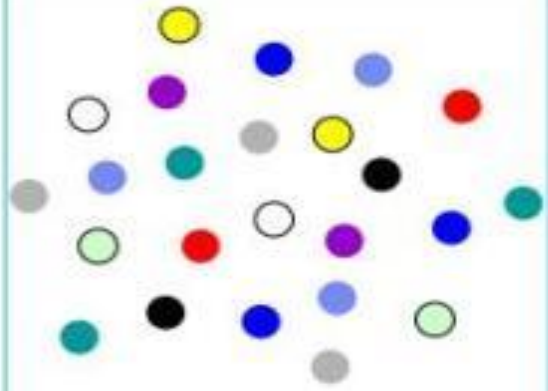
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

Variety



Data in Many Forms

Structured, unstructured, text, multimedia

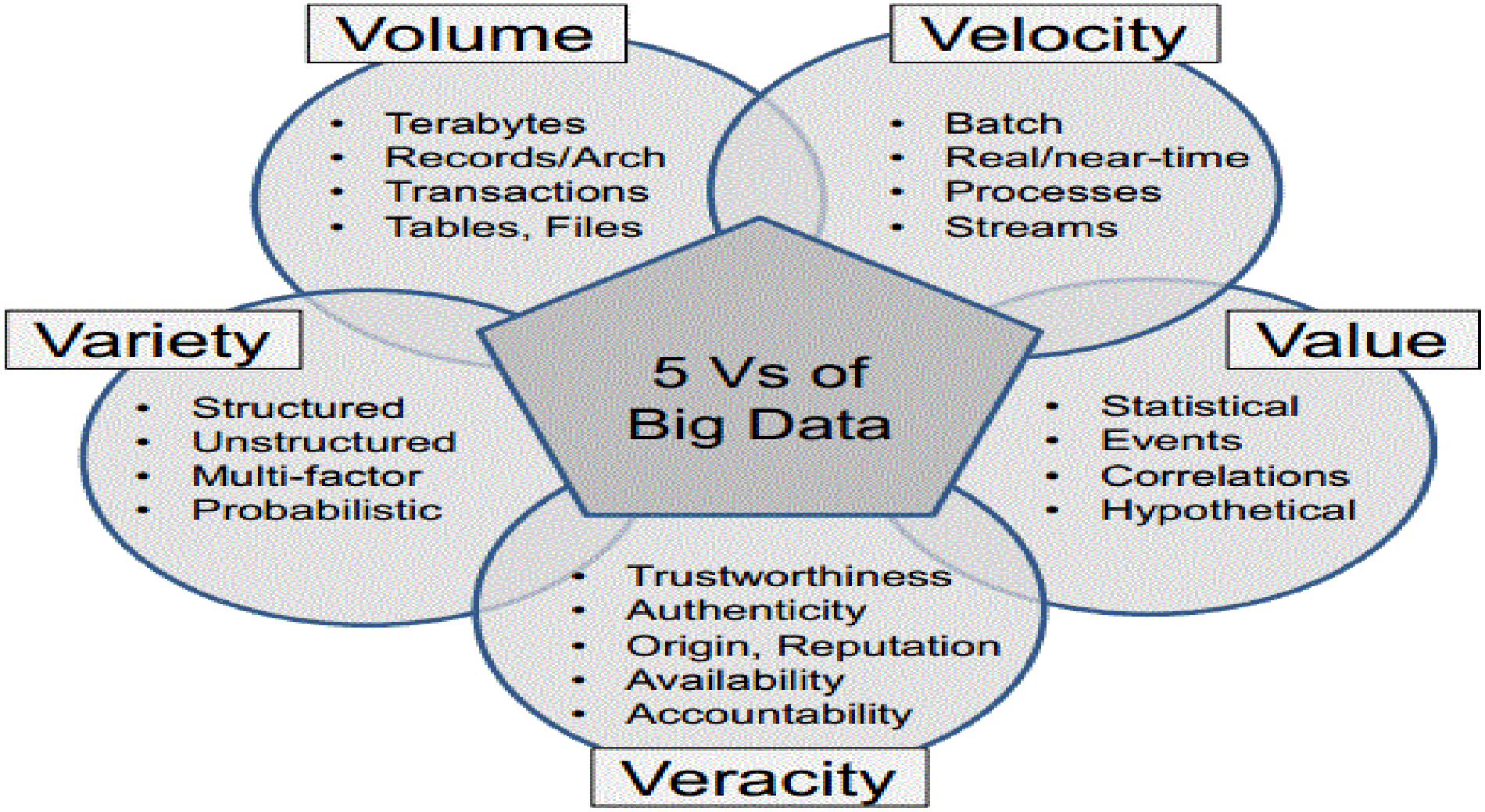
Veracity*

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

[datasciencecentral.com]

5Vs


Volume
Terabytes
Records/Arch
Transactions
Tables, Files
Velocity
Batch
Real/near-time
Processes
Streams
Variety
Structured
Unstructured
Multi-factor
Probabilistic
5 Vs of
Big Data
Value
Statistical
Events
Correlations
Hypothetical
Trustworthiness
Authenticity
Origin, Reputation
Availability
Accountability
Veracity

[ibm]

6 Vs ou Mais…

- Volume
- Variety
- Velocity

- Veracity
- Value
- [Variability | Viability | Visibility | …]

Buzzword

3Vs – Volume, Variedade e Velocidade

Processamento de grande volumes de dados não-estruturados para tomada de decisão em "tempo real"

Caracterização de Big Data

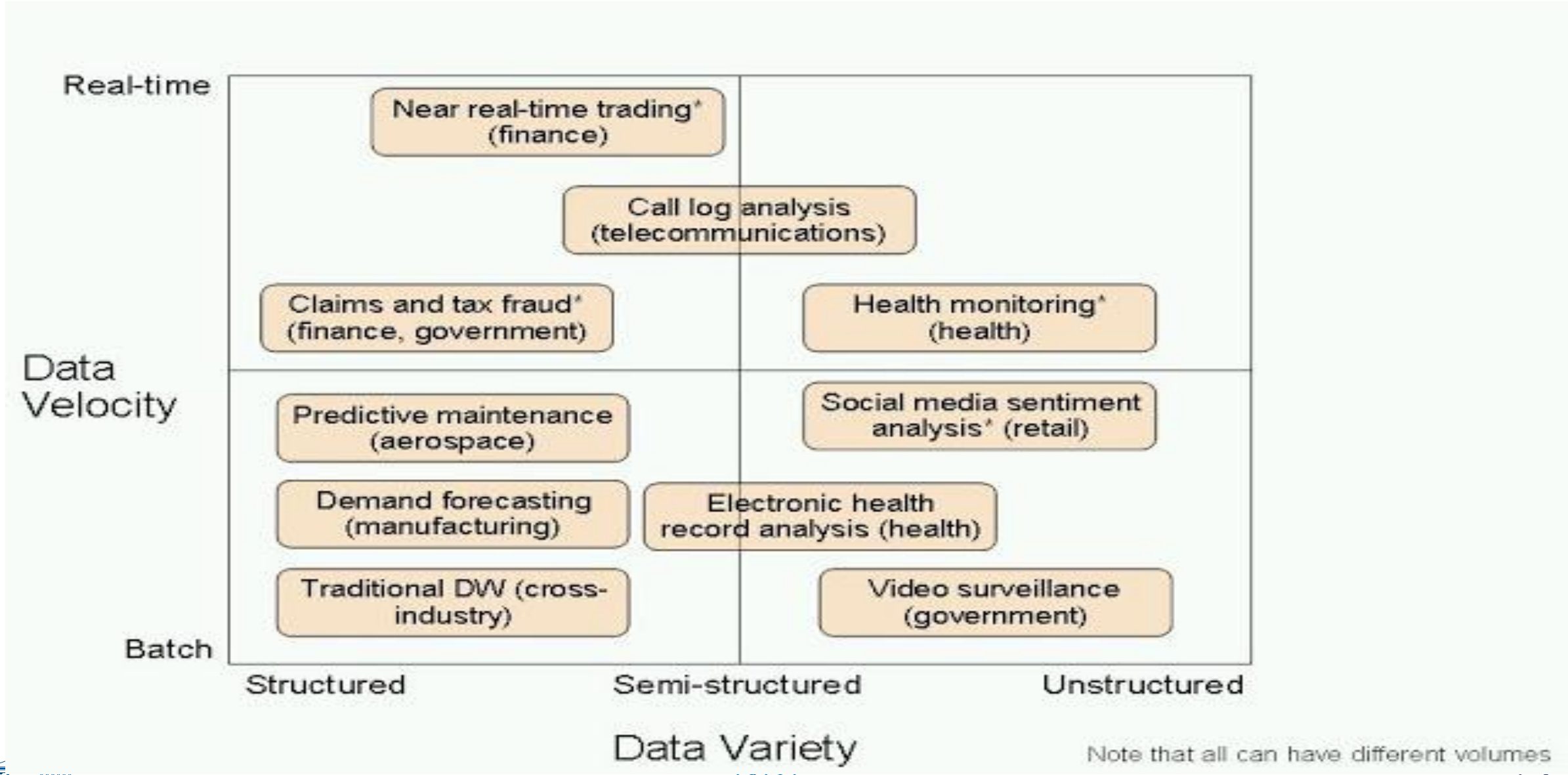
- Big Social Data
- Big Urban Data
- Big Medical Data
- Big ... Data

Big Human Data

Atleta Extremo – Diabético
(Iron Man triatleta e montanhista)
http://bcove.me/w0nks9qg

- Milhares de sensores biométricos
  - Localização (GPS), altitude, temperatura
  - Temperatura corporal, níveis de insulina, caloria, sódio, batimento, …
- Saúde
- Habilidades
- Performance

Velocidade e Variedade



Volume

- Twitter: 1 bi/semana – 6939/s!
- WallMart: 1 mi transações/s
- Facebook: 40 bi de fotos
- IDC: universo digital terá 40 ZB em 2020 (1ZB = 1Bi TB)
- IBM: 2,5 quintilhões dados/dia – 90% criado em 2 anos
- **NYC Taxi – 50GB/ano sem trajeto ou 2TB/ano com trajeto**

Revisando Big Data

- Big Ideas: How big is Big Data (portuguese)
  - https://youtu.be/xWx9THKXd6M

Então...

- Como armazenar e processar Big Data?

Volume

Variedade

Velocidade

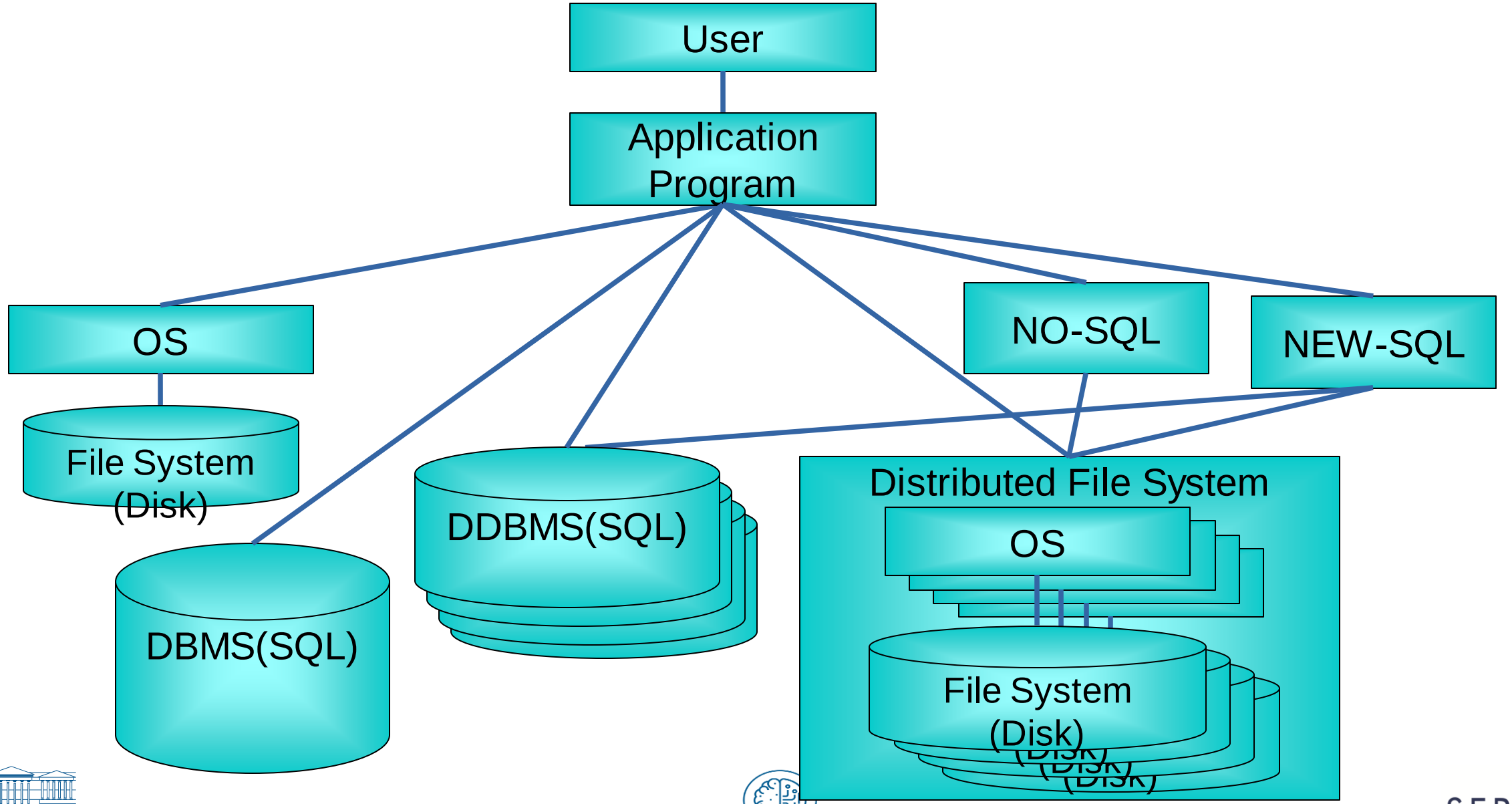
- Normalização → Desnormalização
  - Exclusivo → Redundante
  - Combinação (Join) → Acesso direto
  - “Lento” → “Rápido”
- Essencial → “Tudo”
  - Histórico (log)
  - Tendência

“Grava “tudo” pois poderá ser útil”

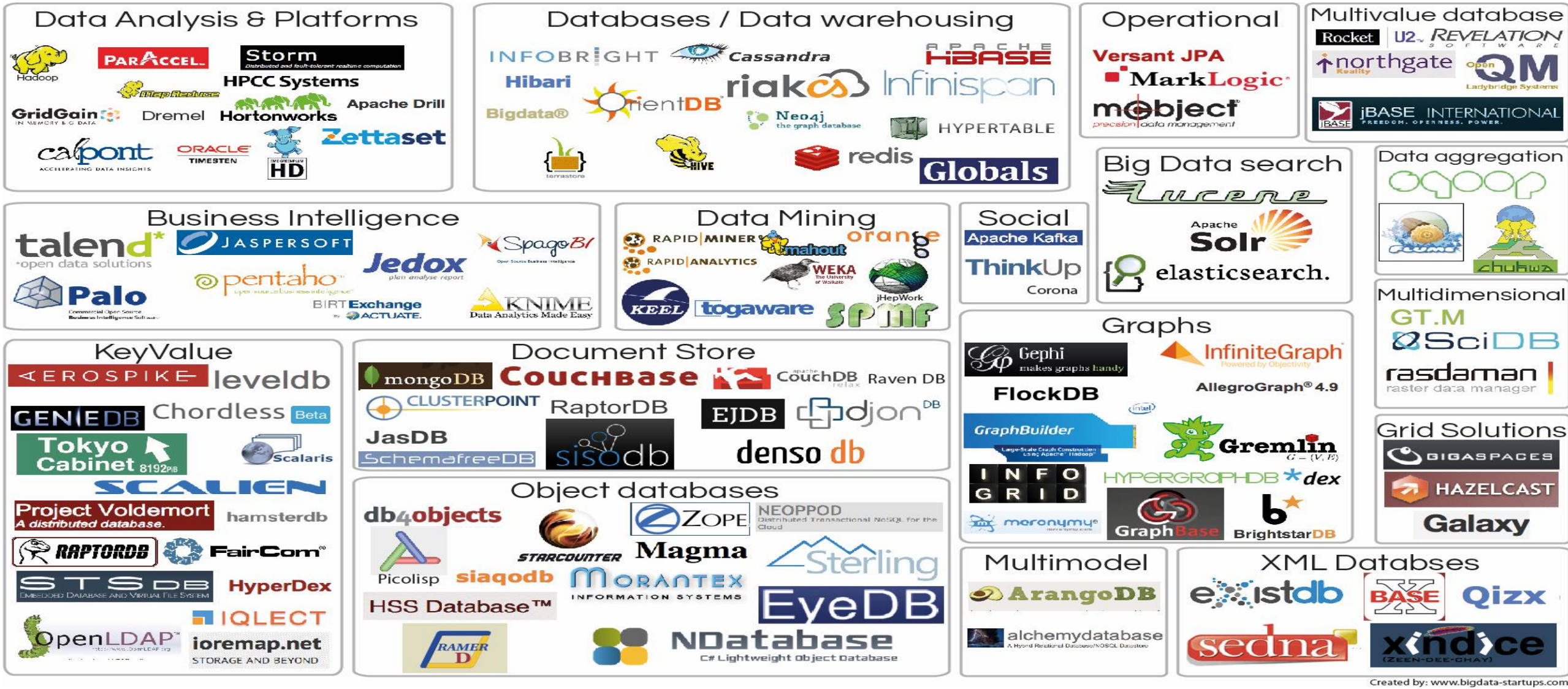
Modelos de Dados para Big Data

- Relacional, Objeto e O-R
- Dimensional/Multidimensional (DW)
- Dados Geográficos (GIS)
- Colunar
- Chave-valor
- Documento
- Grafo (Hierárquico e Rede)
- Não estruturados
  - Excel/CSV, Áudio, Vídeo, Logs, …

Sistemas Big Data



Componentes de Sistemas Big Data



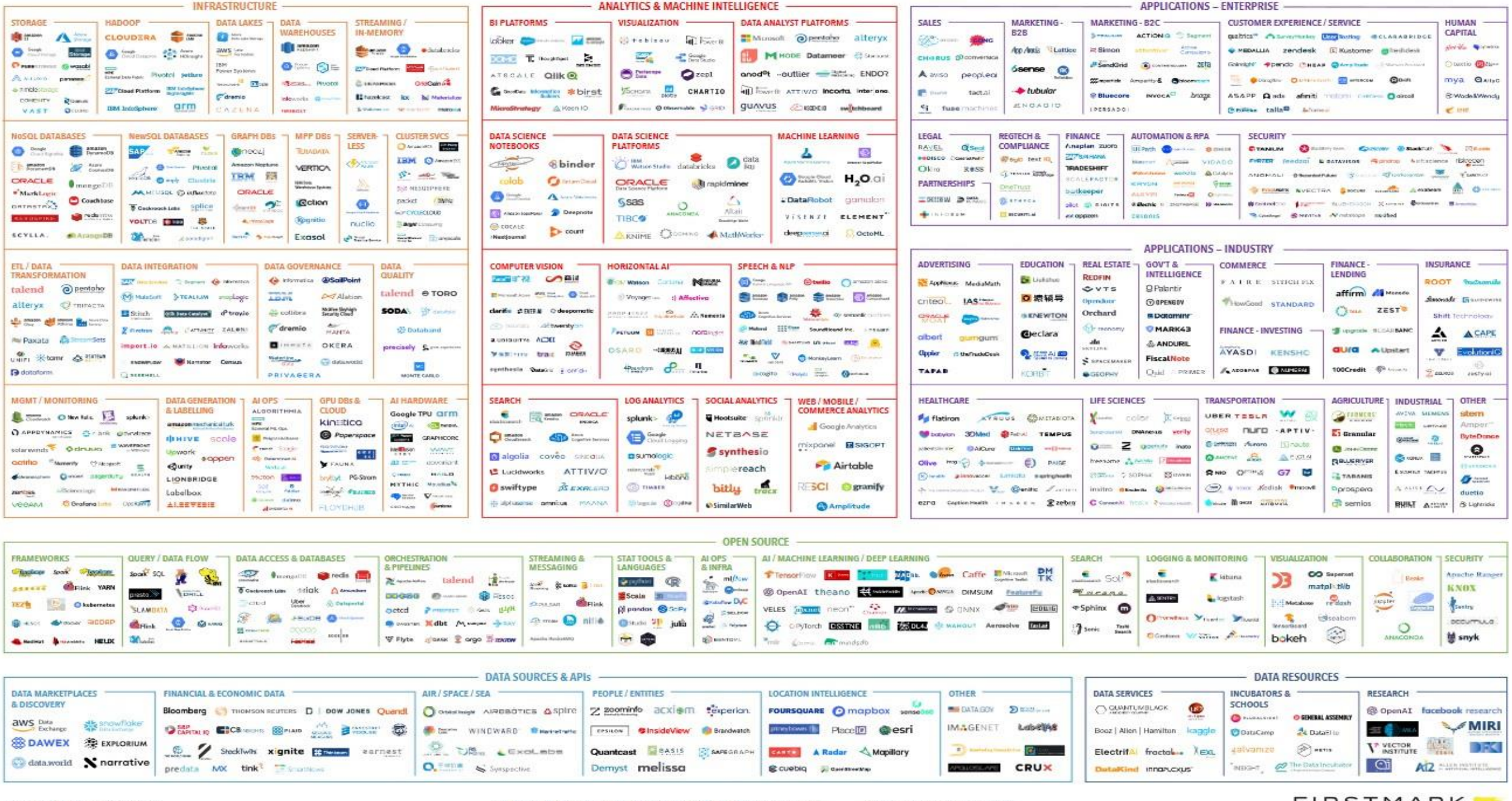
Created by: www.bigdata-startups.com

UFPR

Especialização em Inteligência Artificial

SEPT

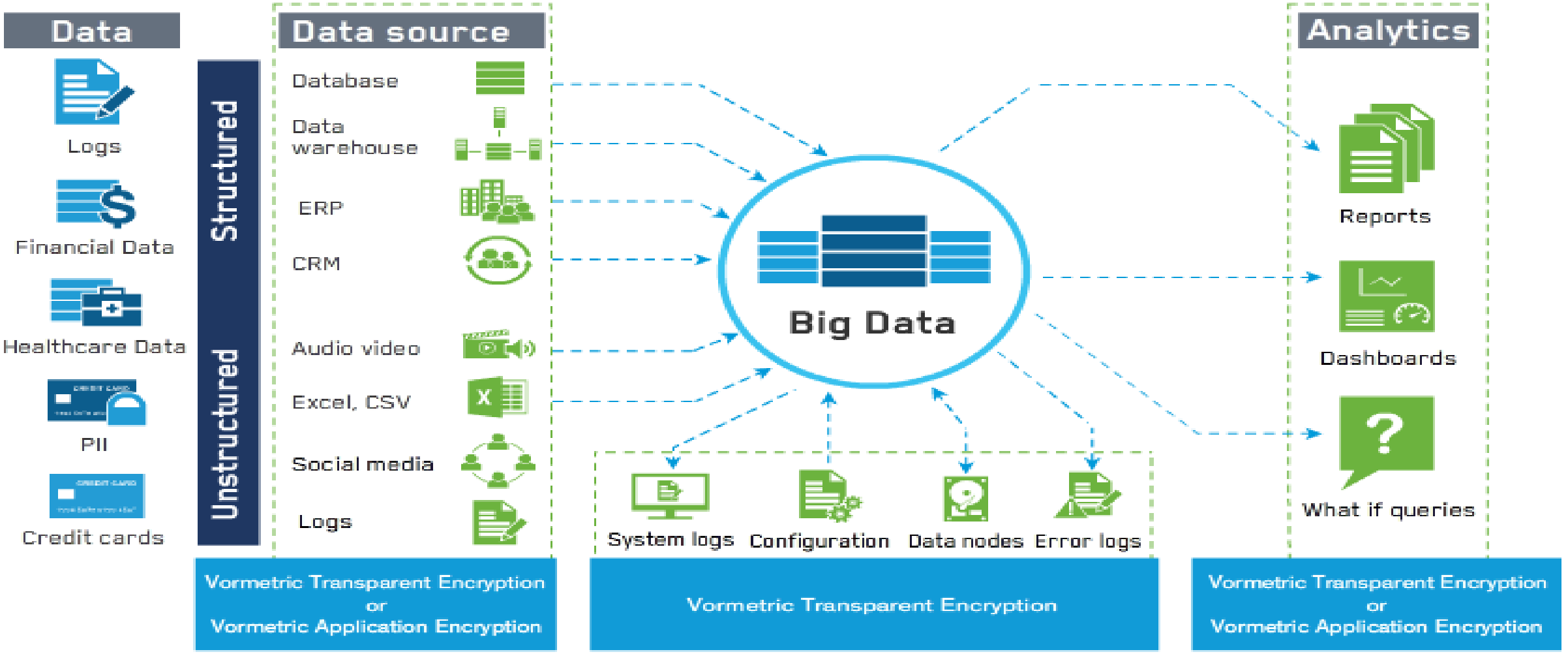
DATA & AI LANDSCAPE 2020

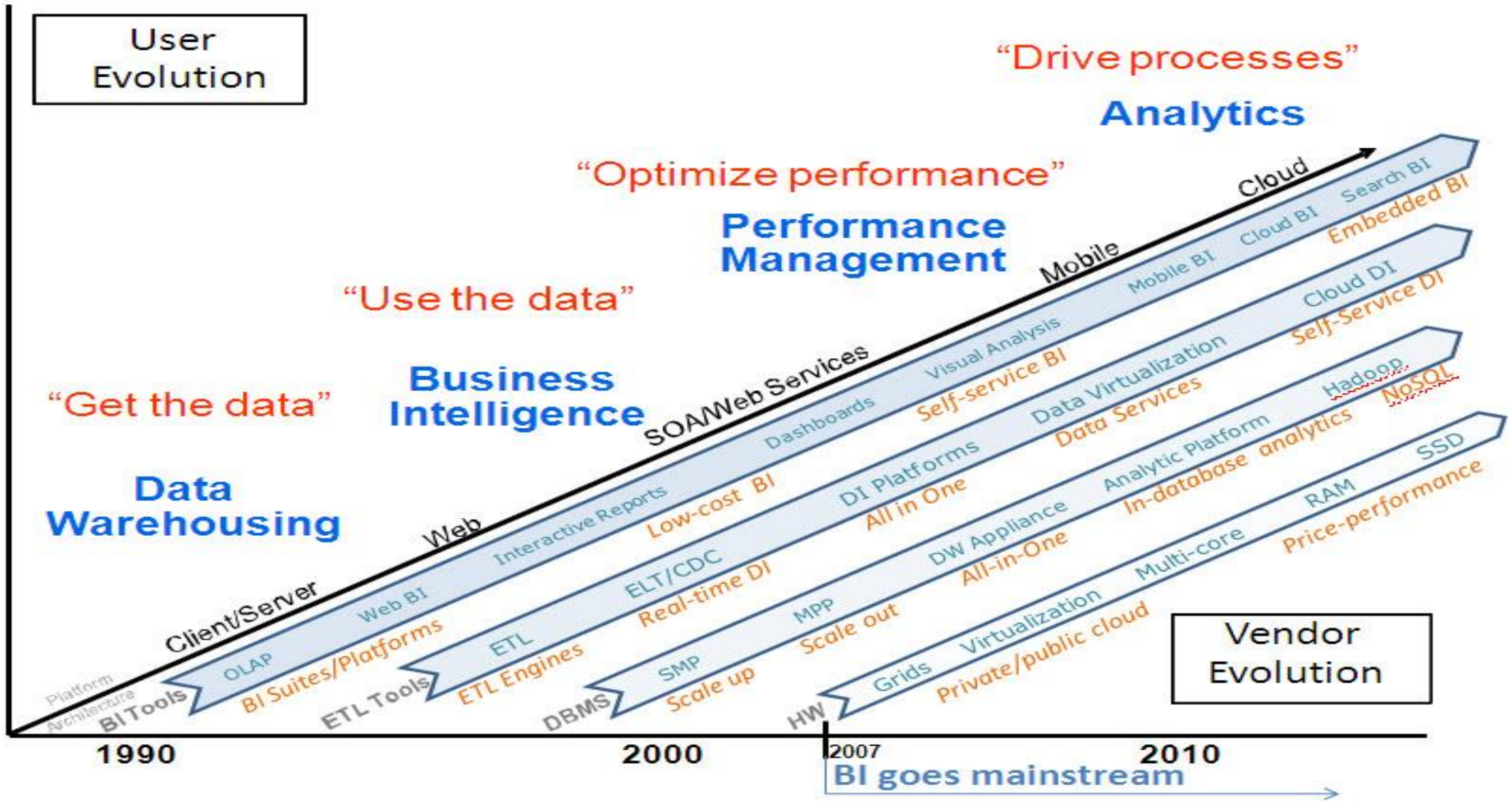


mattturck.com/data2020

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Big Data e Data Lake



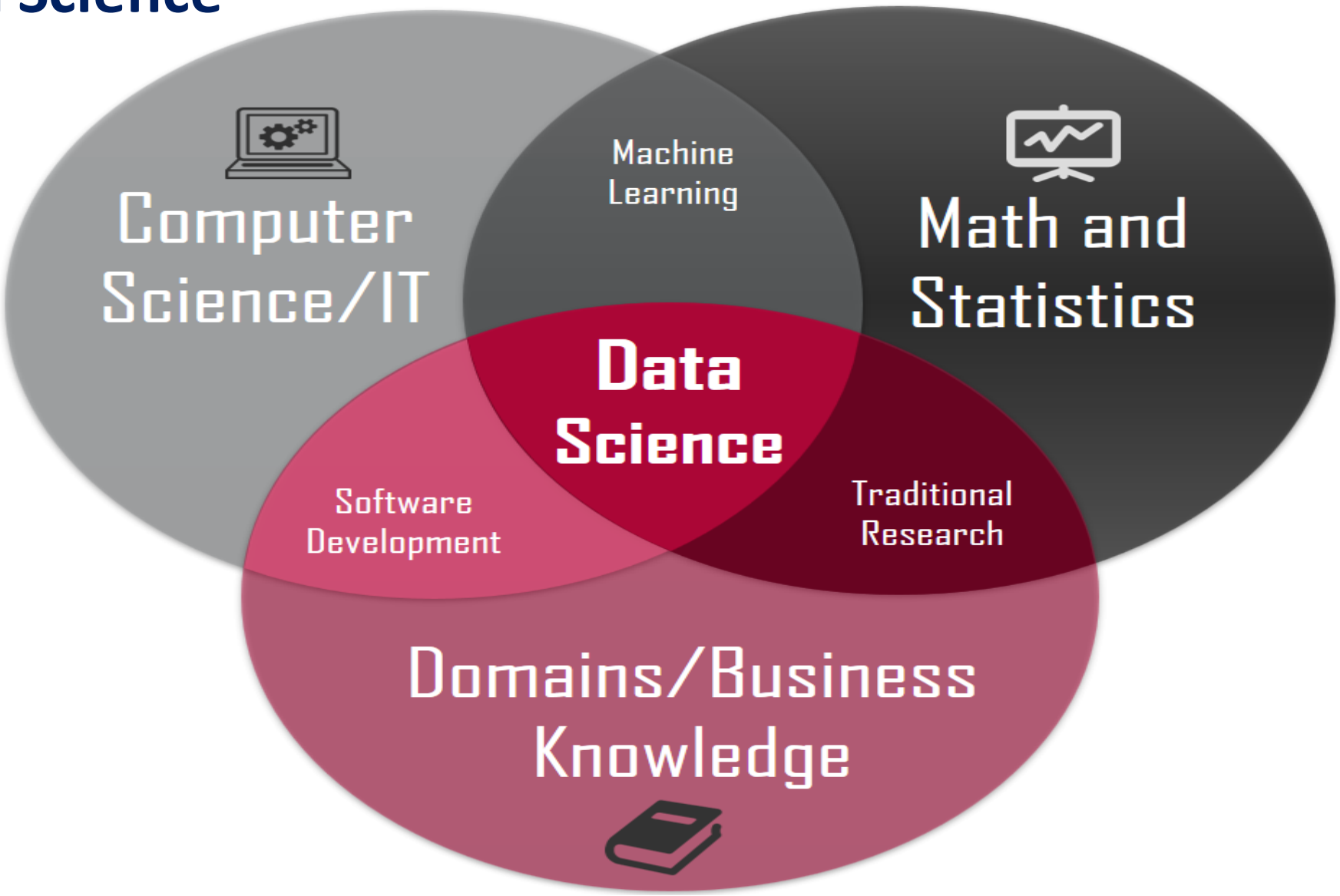

User Evolution
"Drive processes"
Analytics
"Optimize performance"
Performance Management
"Use the data"
Business Intelligence
"Get the data"
Data Warehousing
Cloud
Mobile
SOA/Web Services
Web
Client/Server
Platform Architecture
BI Tools
OLAP
BI Suites/Platforms
Web BI
Interactive Reports
Dashboards
Visual Analysis
Low-cost BI
Self-service BI
Mobile BI
Cloud BI
Search BI
Embedded BI
ETL Tools
ETL
ETL Engines
ELT/CDC
Real-time DI
DI Platforms
All in One
Data Virtualization
Data Services
Cloud DI
Self-Service DI
DBMS
SMP
Scale up
MPP
Scale out
DW Appliance
All-in-One
Analytic Platform
In-database analytics
Hadoop
NoSQL
HW
Grids
Virtualization
Private/public cloud
Multi-core
RAM
SSD
Price-performance
Vendor Evolution
1990
2000
2007
2010
BI goes mainstream

UFPR

Especialização em Inteligência Artificial

SEPT

Data Science



MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

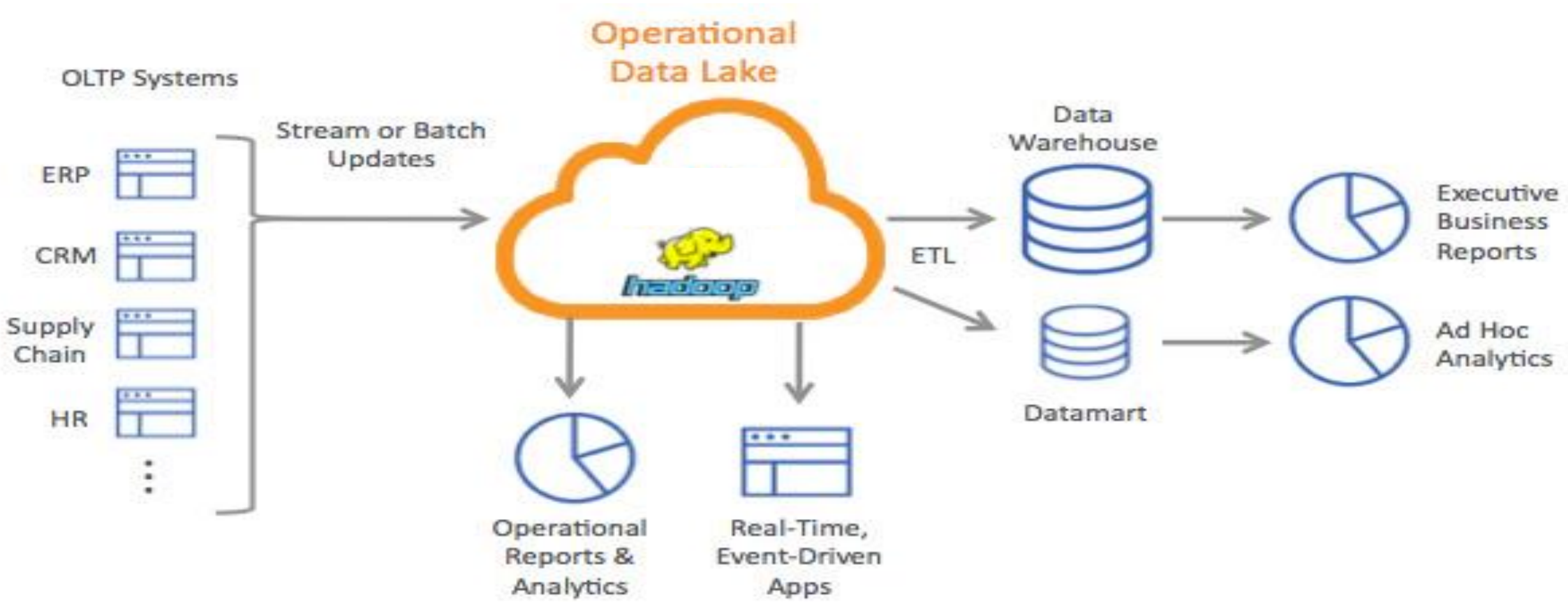
DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

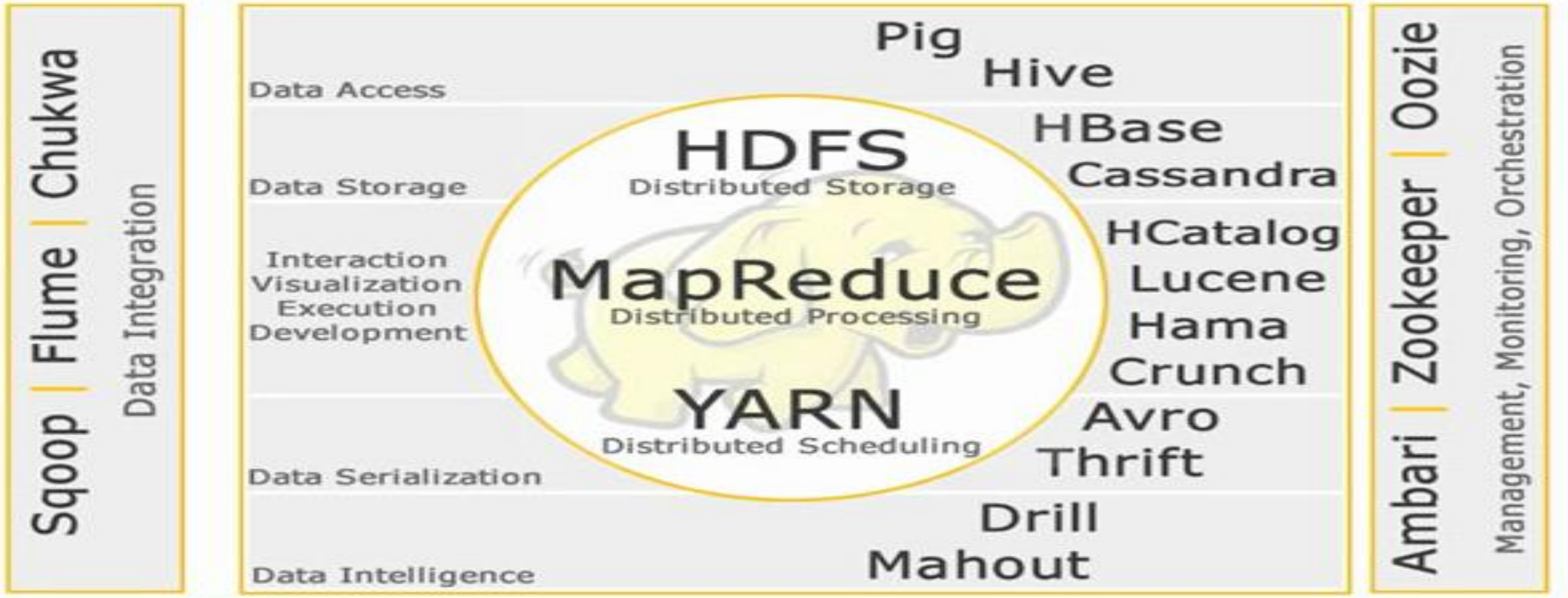
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Data Lake


OLTP Systems
ERP
CRM
Supply Chain
HR
Stream or Batch Updates
Operational Data Lake
hadoop
ETL
Data Warehouse
Executive Business Reports
Datamart
Ad Hoc Analytics
Operational Reports & Analytics
Real-Time, Event-Driven Apps

Hadoop e o Ecosistema Hadoop

http://hadoop.apache.org



Referências

- ...

- Dean, J., & Ghemawat, S. (2004). **MapReduce: Simplified Data Processing on Large Clusters**. In Proc. of the OSDI - Symp. on Operating Systems Design and Implementation (pp. 137–149). USENIX.

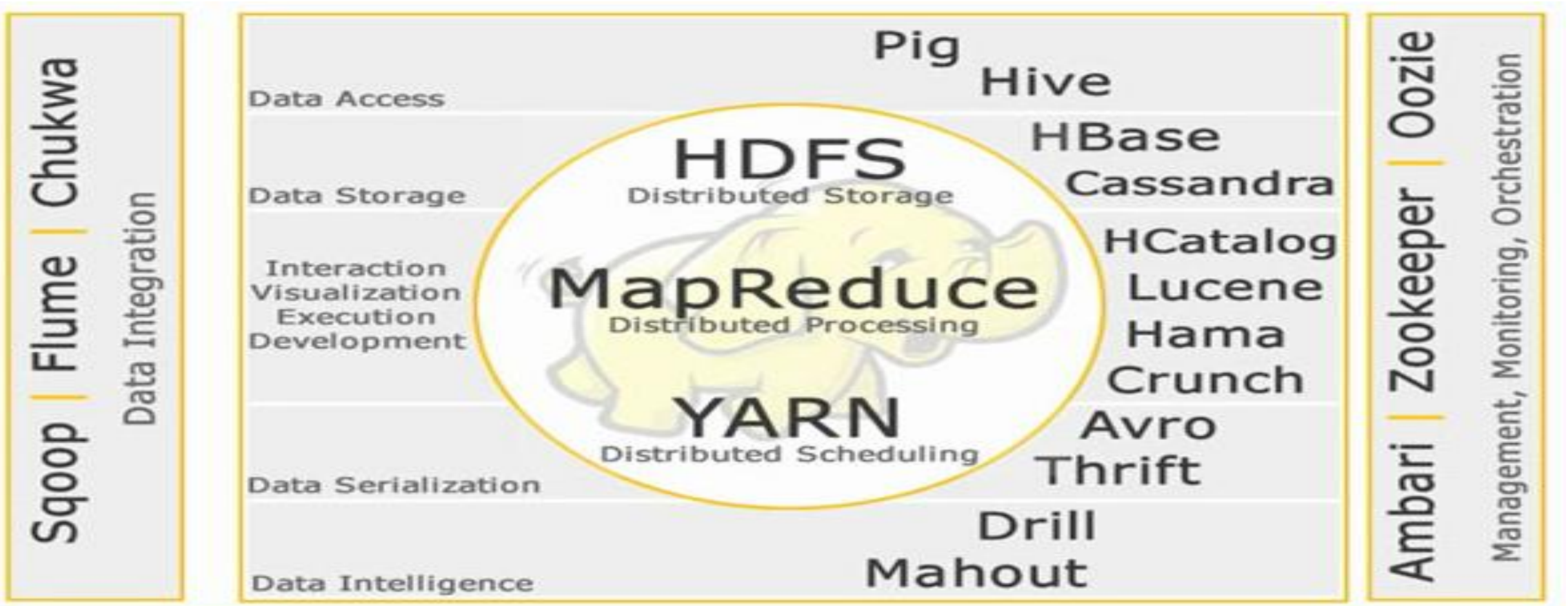
MapReduce (MR) [by Google'04]

- Modelo de programação e **uma implementação/framework**
  - Programação personalizada
  - Distribuição e Paralelização automática
  - Milhares de máquinas "comuns"
  - Balanceamento de carga
  - Otimização de rede e transferência de arquivos
  - Tolerância à falhas
  - Ambiente de desenvolvimento comum
    - Melhorias beneficiam todos usuários

Hadoop e o Ecosistema Hadoop

[Yahoo et al., 2006]

http://hadoop.apache.org

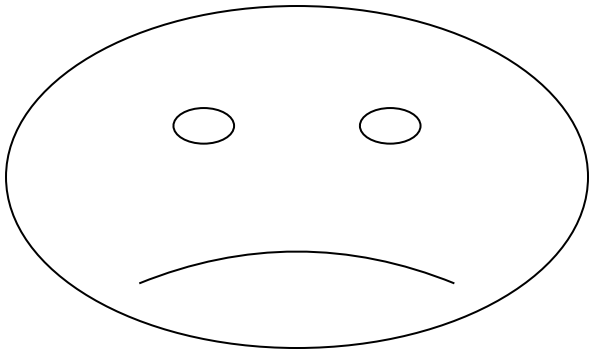


Mão na Massa...

- Instalar Hadoop (na mão)
  - http://hadoop.apache.org/
  - Local, pseudo-distributed, distributed

- E os outros sistemas?
  - Pig, Hive, Cassandra, Hbase,...

- Atualizações?



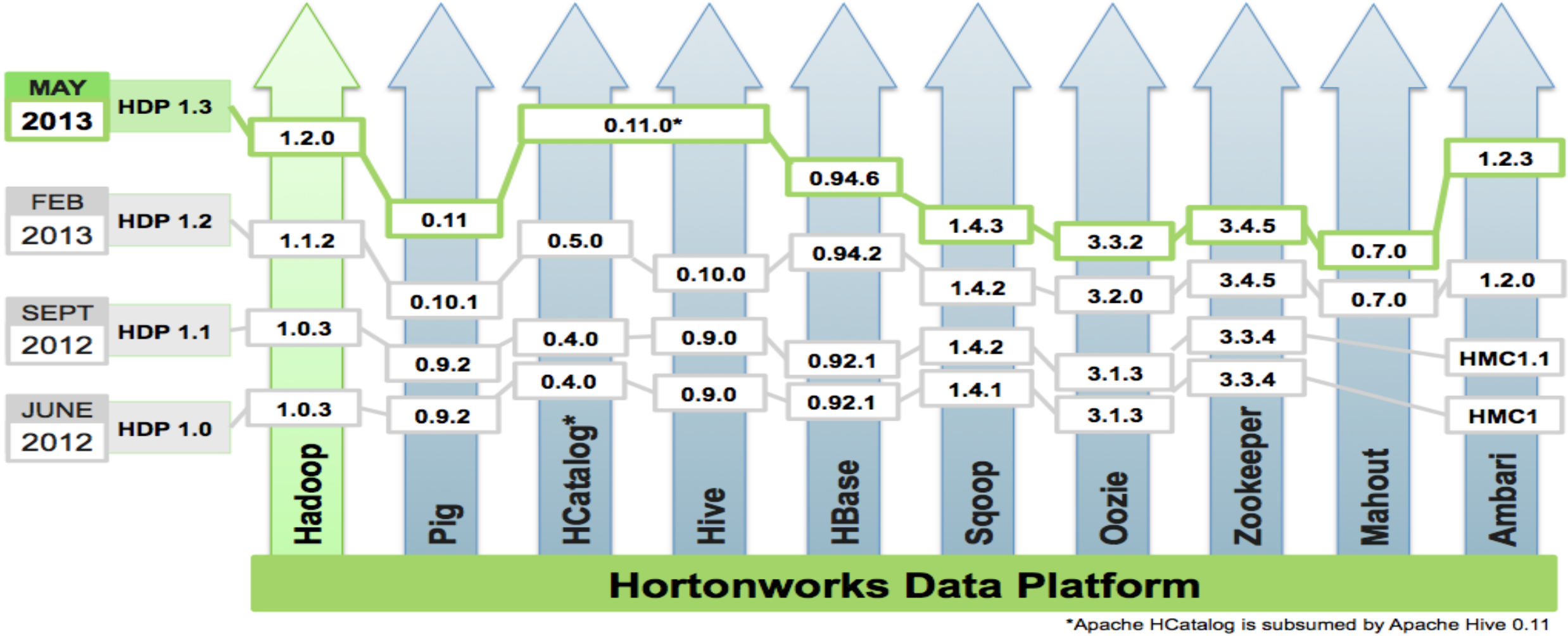
Distribuições Hadoop

- Problemas de incompatibilidade entre versões dos sistemas Hadoop (eco-sistema)
  - Pacotes Linux, VMs, tarballs
  - Scripts adicionais para manipular sistemas Hadoop
  - Suporte/Backport de características e atualizações feitas pela Apache
  - Distribuidores empregarão sistemas Hadoop e ajudarão a atualizar/manter o Apache Hadoop

Distribuições Hadoop

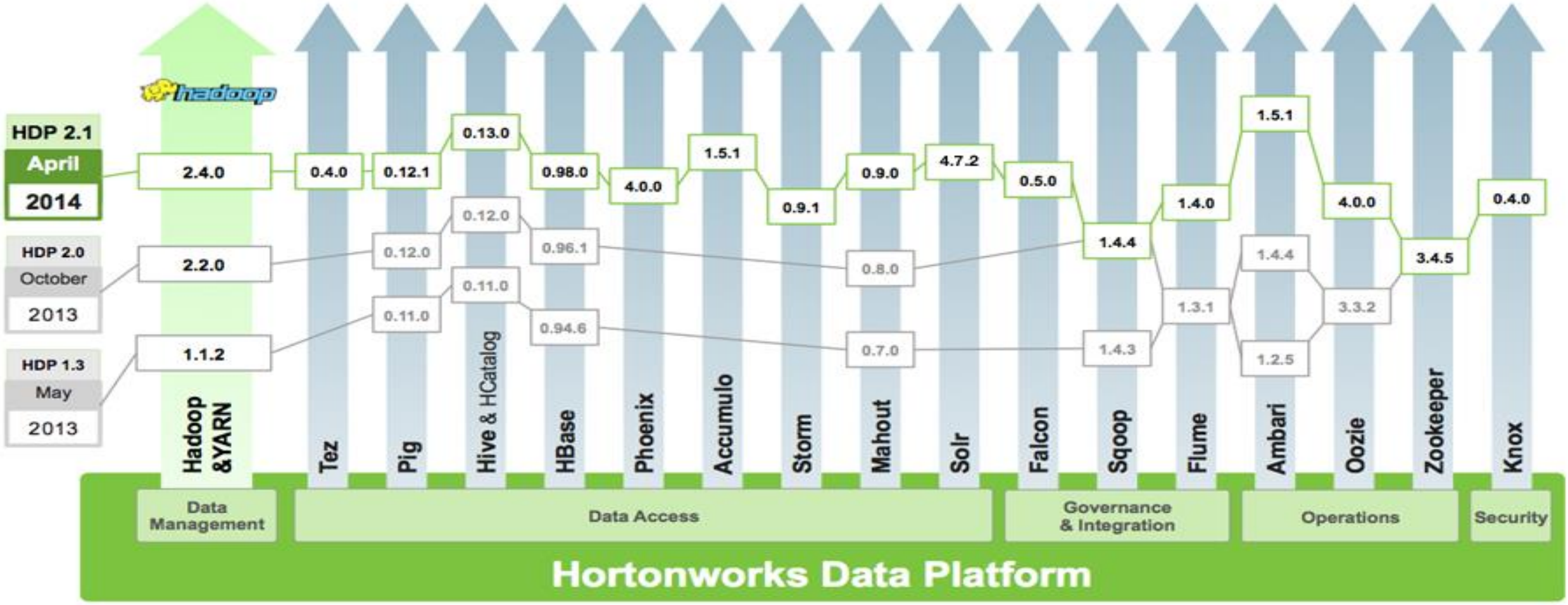
- *Hortonworks Data Platform (HDP)*
  http://hortonworks.com

HDP 1.3

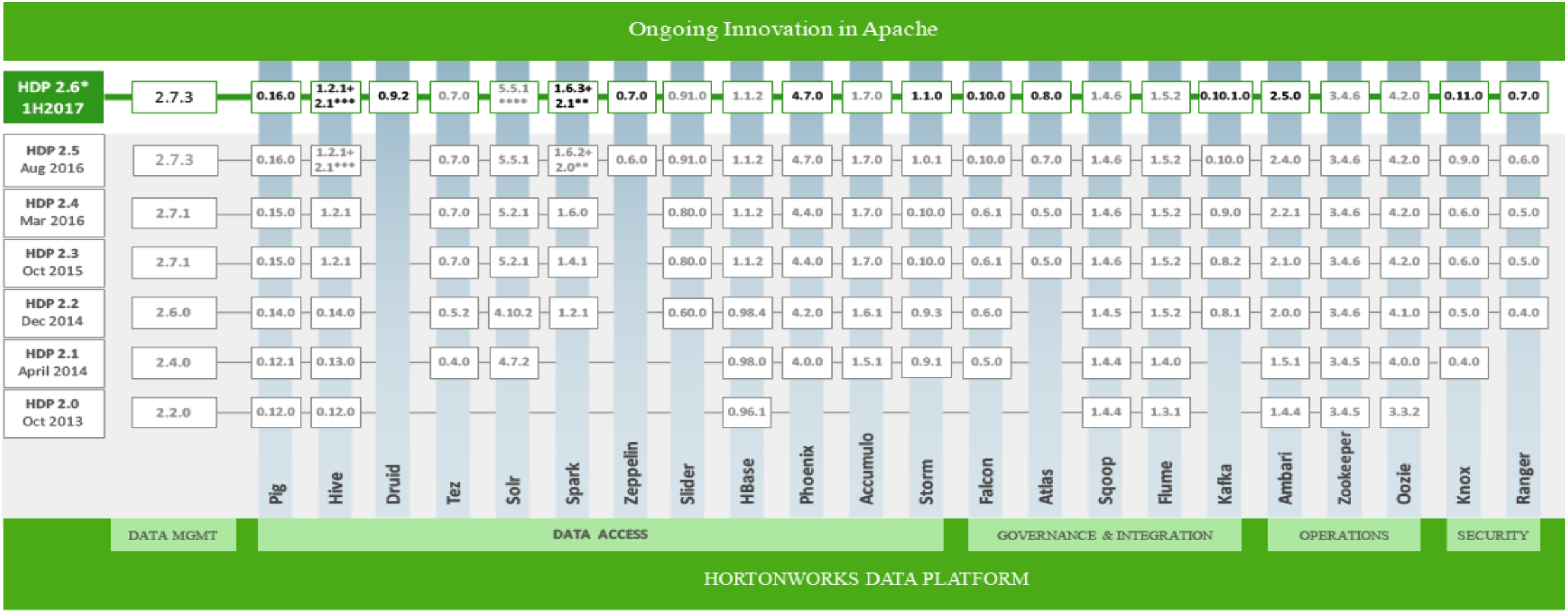


*Apache HCatalog is subsumed by Apache Hive 0.11

HDP 2.1


HDP 2.1
April
2014
HDP 2.0
October
2013
HDP 1.3
May
2013
2.4.0
2.2.0
1.1.2
0.4.0
0.12.1
0.13.0
0.98.0
4.0.0
1.5.1
0.9.1
0.9.0
4.7.2
0.5.0
1.4.4
1.4.0
1.5.1
4.0.0
0.4.0
3.4.5
0.12.0
0.12.0
0.96.1
0.8.0
1.4.4
0.11.0
0.11.0
0.94.6
0.7.0
1.4.3
1.3.1
3.3.2
1.2.5
Hadoop &YARN
Tez
Pig
Hive & HCatalog
HBase
Phoenix
Accumulo
Storm
Mahout
Solr
Falcon
Sqoop
Flume
Ambari
Oozie
Zookeeper
Knox
Data Management
Data Access
Governance & Integration
Operations
Security
Hortonworks Data Platform

HDP 2.6



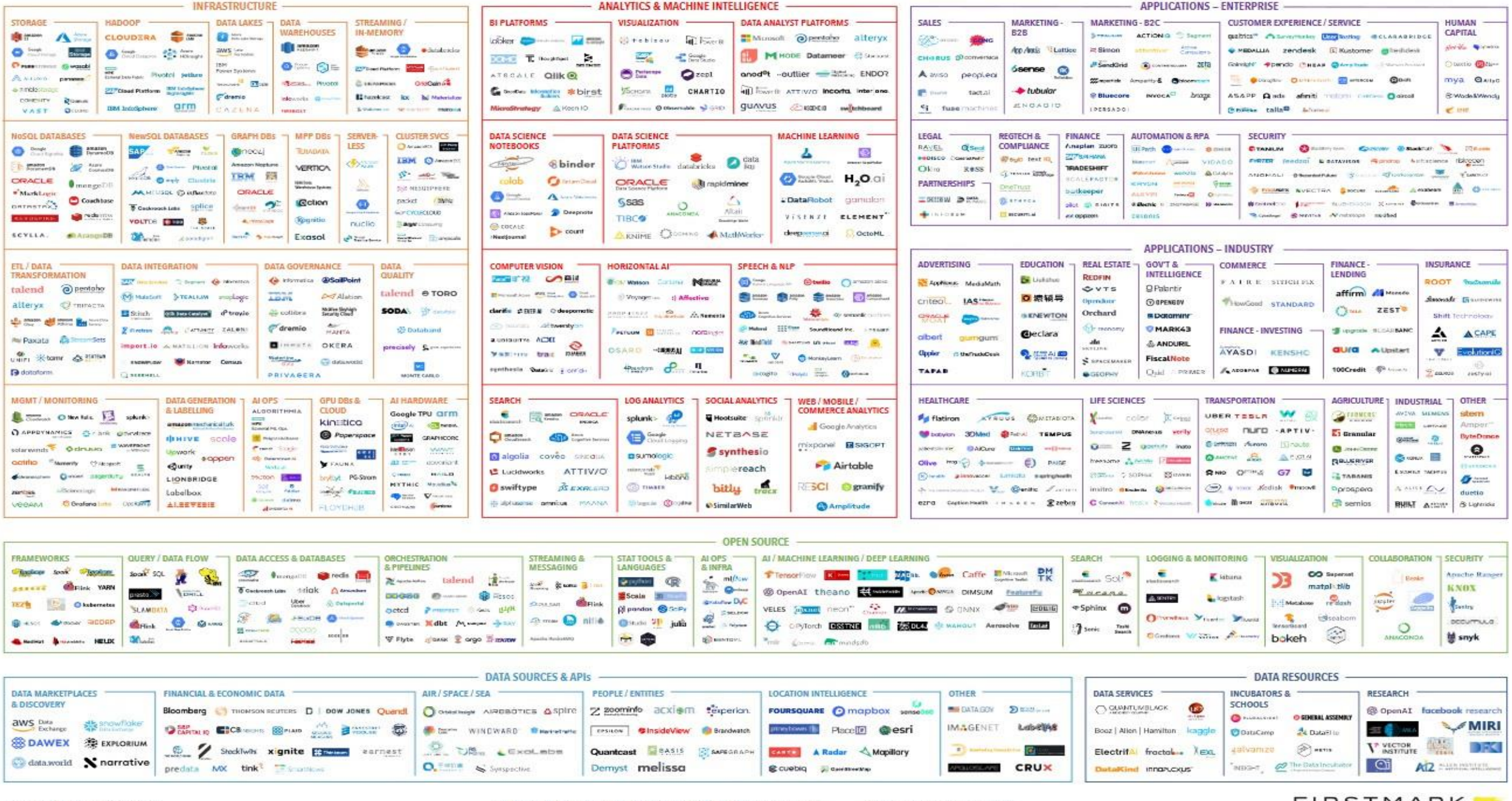
* *HDP 2.6 – Shows current Apache branches being used. Final component version subject to change based on Apache release process.*
** *Spark 1.6.3+ Spark 2.1 – HDP 2.6 supports both Spark 1.6.3 and Spark 2.1 as GA.*
*** *Hive 2.1 is GA within HDP 2.6.*
**** *Apache Solr is available as an add-on product HDP Search.*

UFPR

Especialização em Inteligência Artificial

SEPT

DATA & AI LANDSCAPE 2020

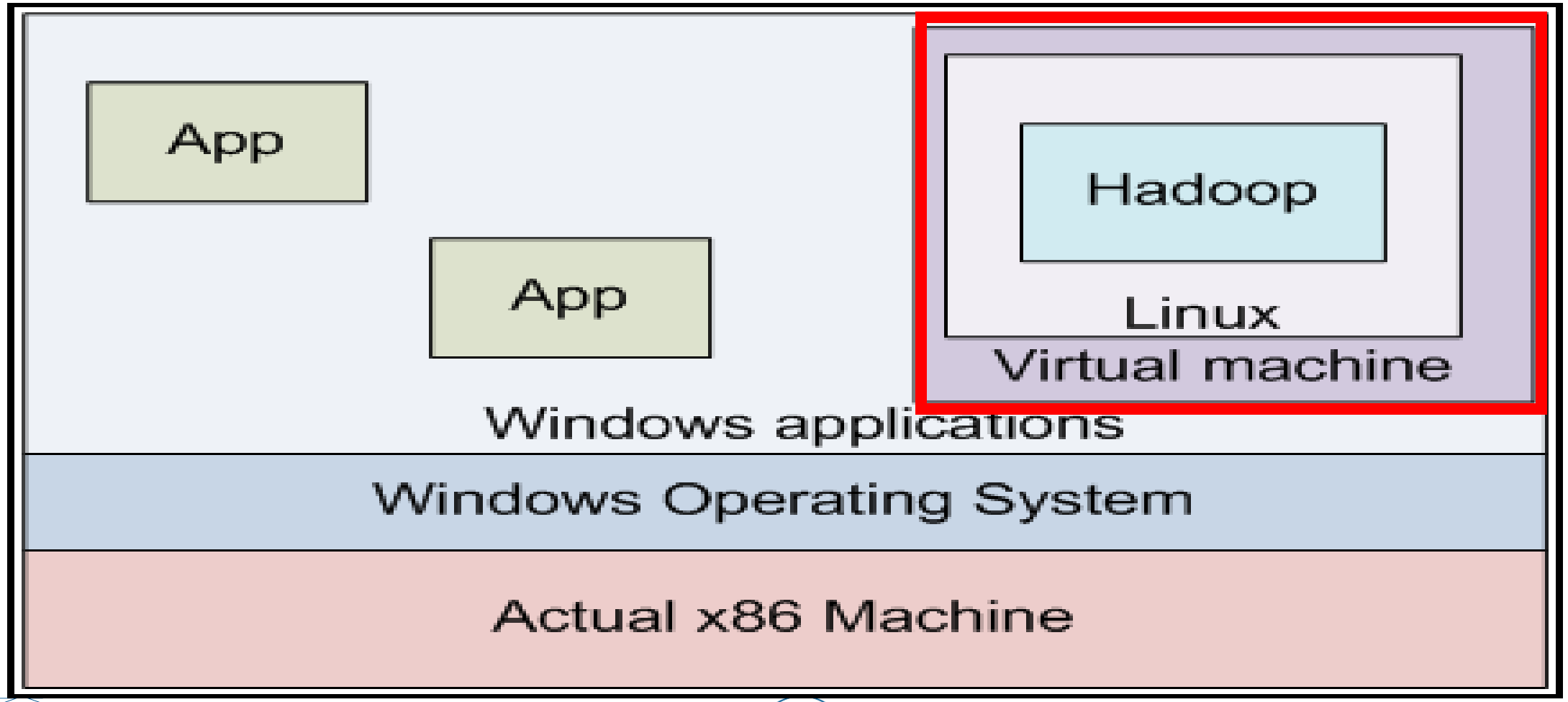


mattturck.com/data2020

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Hortonworks Sandbox 2.1

- Usar uma Sandbox: Virtual Machine (VM) com tudo!



Atividade 4

Dado o conjunto de programas exemplos disponível em:

```
$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar
```

Comente brevemente justificando e apresentando a aplicação de algum outro programa exemplo.

Como faço a minha aplicação MR?

MapReduce (MR) [by Google'04]

- **Modelo de programação (?)** e uma implementação/framework
  - Programação personalizada
  - Distribuição e Paralelização automática
  - Milhares de máquinas "comuns"
  - Balanceamento de carga
  - Otimização de rede e transferência de arquivos
  - Tolerância à falhas
  - Ambiente de desenvolvimento comum
    - Melhorias beneficiam todos usuários

Problema MR Típico

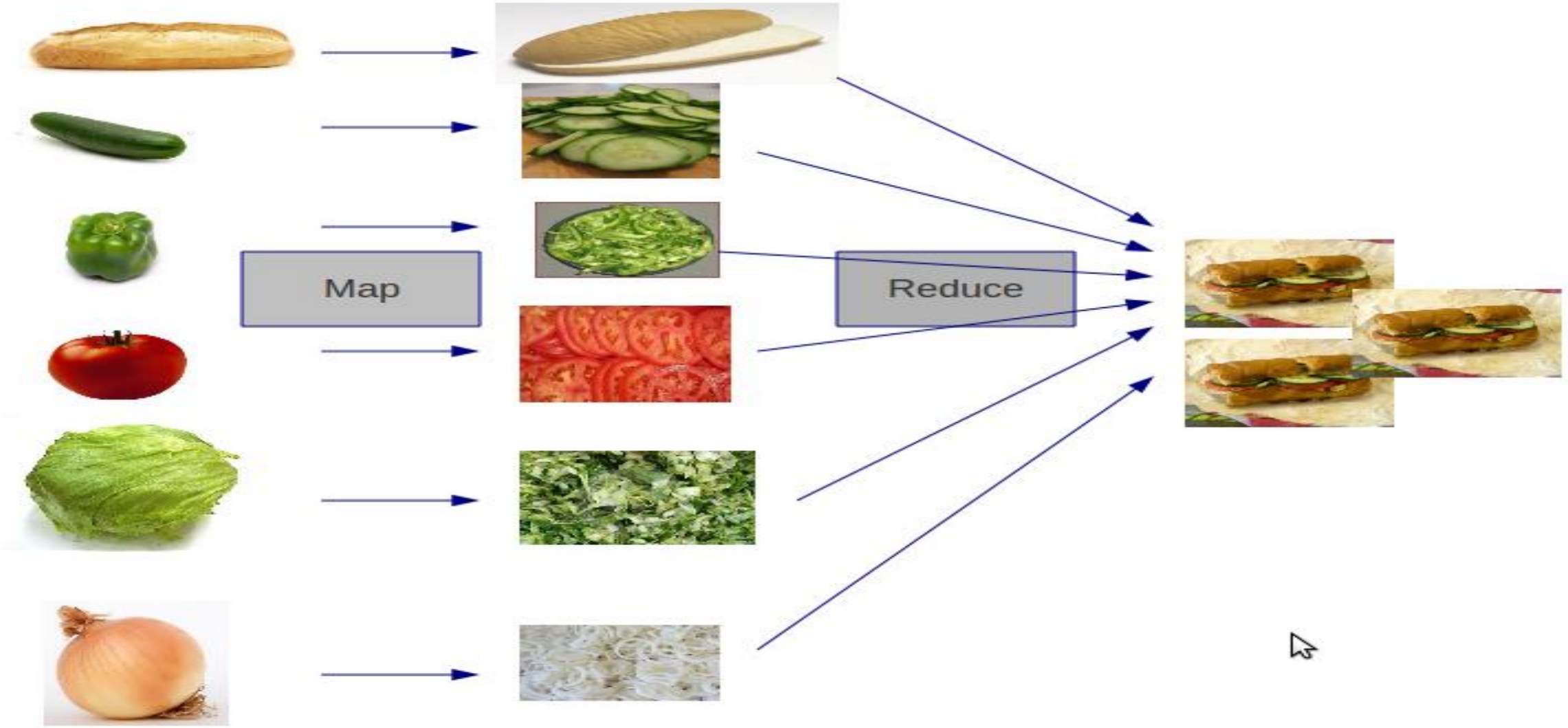
- Leia a entrada
- **Map**
  - identifique os registros (o dado relevante)
- Combine e Agrupe
- **Reduce**
  - agregue, some, resuma, filtre, transforme,...
- Escreva o resultado

- O esquema permanece o mesmo, altera-se o Map e o Reduce para atender cada problema.

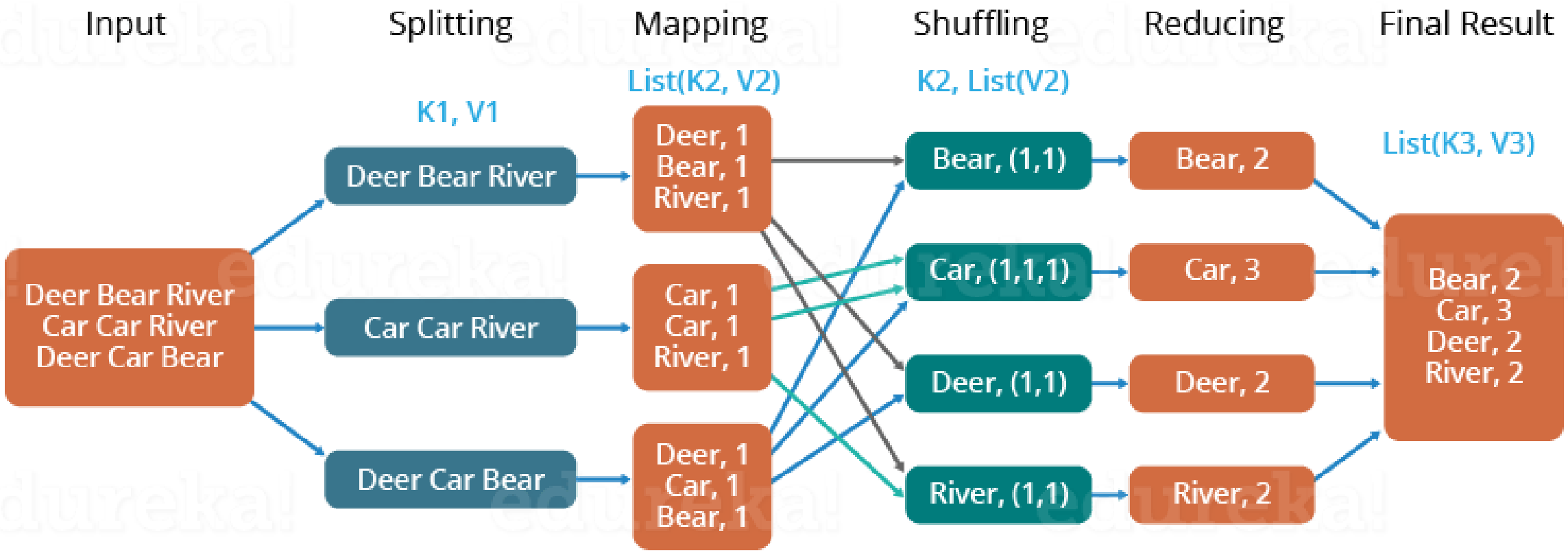
Modelo de Programação

- Computação baseada em <chave, valor>
- **Map** (<chave, valor>)
  - Saída: um conjunto de pares <chave2, valor2>
- **Reduce** (<chave2, lista de valor2>)
  - Saída: um par <chave2, valor3>

Aplicando Map e Reduce...



Aplicando Map e Reduce – Contador de Palavras



Algoritmo Map Reduce para Contar Palavras

```
map(String key, String value)
    // key: nome arquivo
    // value: conteúdo do arquivo
    for each word w in value:
        EmitInternediate(w, "1");

reduce(String key, Iterator values)
    // key: uma palavra
    // values: uma lista de contadores
    result=0;
    for each v in values:
        result+= ParseInt(v);
    Emit(key, AsString(result));
```

Função Map

```
public void map(LongWritable key, Text value,
OutputCollector<Text, IntWritable> output, Reporter
reporter) throws IOException {
  String line = value.toString();
  StringTokenizer tokenizer = new StringTokenizer(line);
  while (tokenizer.hasMoreTokens()) {
    word.set(tokenizer.nextToken());
    output.collect(word, one);
  }
}
```

Função Reduce

```
public void reduce(Text key, Iterator<IntWritable> values,
OutputCollector<Text, IntWritable> output, Reporter
reporter) throws IOException {
    int sum = 0;
    while (values.hasNext()) {
        sum = sum + values.next().get();
    }
    output.collect(key, new IntWritable(sum));
}
```

Temperatura Máxima

Arquivo com cidades e temperaturas

Toronto, 20
New York, 22
New Jersey, 26
Toronto, 22
...

- Como é a função Map e Reduce para obter as temperaturas máximas das cidades?
  - Map (?, ?)
  - Reduce (?, ?)

Temperatura Máxima

- **Map** (arquivo, conteúdo)

  Para cada cidade, emita <cidade, temp>

(Toronto, 20) (Whitby, 25) (New York, 22) (Rome, 33)(Toronto, 18) (Whitby, 27) (New York, 32) (Rome, 37)(Toronto, 32) (Whitby, 20) (New York, 33) (Rome, 38)(Toronto, 22) (Whitby, 19) (New York, 20) (Rome, 31)(Toronto, 3 (Whitby, 22) (New York, 19) (Rome, 30)

- **Reduce** (cidade, valores)

  Emita o maior valor dentre os valores das temp.

(Toronto, 32) (Whitby, 27) (New York, 33) (Rome, 38)

Dados Geográficos

Sistema de mapeamento de estradas

- Objetivo
  - Identificar as conexões das estradas

- **Map**
  - Cria pares de pontos conectados
  - (estrada, intercessão) ou (estrada, estrada)
- **Reduce**
  - Agrupa os pontos de conexão para cada estrada

Grafo de Emails

- Contar quantos e-mails foram trocados entre cada par
  - Joao, Pedro: 30
  - Pedro, Ana: 14
  - Ana, Joao: 7
  - Daniel, Pedro: 43
- Como seria uma aplicação MapReduce?

Grafo de Emails - INPUT


```
"_id" : ObjectId("4f2ad4c4d1e2d3f15a000000"),
"body" : "Here is our forecast\n\n ",
"filename" : "1.",
"headers" : {
  "From" : "phillip.allen@enron.com",
  "Subject" : "Forecast Info",
  "X-bcc" : "",
  "To" : "tim.belden@enron.com",
  "X-Origin" : "Allen-P",
  "X-From" : "Phillip K Allen",
  "Date" : "Mon, 14 May 2001 16:39:00 -0700 (PDT)",
  "X-To" : "Tim Belden ",
  "Message-ID" :  "<18782981.1075855378110.JavaMail.evans@thyme>",
  "Content-Type" : "text/plain; charset=us-ascii",
  "Mime-Version" : "1.0"
```

- Map(?,?)

- Reduce(?,?)

MR para Grafo de Emails

- Map(arquivo, conteúdo)
  - Para cada destinatário
    - Grave(“remetente, destinatário”, 1)
- Reduce(par, valores)
  - Para cada valor em valores
    - Some cada valor
  - Grave o resultado (par, soma)

O que faz essa Aplicação MR?

```
Map (file, social.person records V1)
    for each social.person record in V1 do
        let Y be the person's age
        let N be the number of contacts the person has
        produce one output record (Y,(N,1))

Reduce ( age (in years) K2, list of (N,C) )
    for each input record (Y,(N,C)) do
        Accumulate in S the sum of N*C
        Accumulate in Cnew the sum of C
    let A be S/Cnew
    produce one output record (Y,(A,Cnew))
```

Qual o Problema?

Average of the top 100 salaries for each unique department (key), if that average is larger than 100,000

```
public void reduce(String key, Iterator<Integer>
salaries) {
    int sum = 0; int i = 0;
    while (salaries.hasNext() && i<100) {
     sum += salaries.next();
     i += 1; }
    emit((i>0 && sum/i > 100000)? sum/i : -1);
}
```

[Csallner, C., Fegaras, L., & Li, C. (2011). New Ideas Track: Testing MapReduce-Style Programs. In Proc. of the ESEC/FSE - SIGSOFT Symp. and the European Conf. on Foundations of Software Engineering (pp. 504–507). ACM.]

Como implementar e implantar a sua aplicação MR? (hadoop-mapreduce-examples.jar)

- `$ hadoop jar /usr/lib/hadoop-mapreduce/`**`hadoop-mapreduce-examples.jar`** **`pi`** `10 10`
- `$ hadoop jar /usr/lib/hadoop-mapreduce/`**`hadoop-mapreduce-examples.jar`** **`wordcount`** `/mysqld.log /saida`

Implementação e Implantação de uma Aplicação MR

- Gerar o jar com a aplicação MR no desktop *
- Enviar o jar para a sandbox ou cluster Hadoop
  - Winscp, scp ou [filebrowser]! :)
- Executar na sandbox ou cluster

```
$hadoop jar seu.jar /mysqld.log /saída
```

* Seu ambiente de desenvolvimento (Sandbox) não tem JDK e fonte do Hadoop

Gerar Jar

- Versão hadoop na sandbox? hadoop version
- https://archive.apache.org/dist/hadoop/common/
- hadoop[version].tar.gz (src não tem os jar)
- Netbeans, new project, java application
- Java: https://tinyurl.com/AulaWC-java (baseado em http://tinyurl.com/htjlklc)
  - Adequar o pacote (package)
  - Renomear arquivo para ficar igual a classe
- Adicionar bibliotecas (Project, Properties, Library, Add JAR)
  share/hadoop/common/hadoop-common
  share/hadoop/mapreduce/hadoop-mapreduce-client-core
- Project properties, sources, source/binary format
  - Versão Jdk na sandbox: java –version
- Netbeans, run, build project

Executar a sua Aplicação MR

- Enviar o jar para a sandbox ou cluster Hadoop
  - Winscp, scp ou [filebrowser]! :)
- Executar na sandbox

```
$hadoop jar seu.jar /mysqld.log /saida
```

Problemas na Sandbox

- Processos param/congelam sem resposta
  - Espaço em disco
  - Apagar Logs
    - du
    - rm -rf ….
- Travar processo
  - Reiniciar a VM ou serviço via Ambari
- Ajuste de horário
  - yum install ntp; chkconfig ntpd on; service ntpd restart
  - rm /etc/localtime; ln -s /usr/share/zoneinfo/America/Sao_Paulo /etc/localtime
  - date mmddHHMM; hwclock -w

Próximas Aulas

- Fundamentos de Big Data
  - Big Data, Data Lake e Data Science
- Map Reduce e Hadoop
  - Utilização da Sandbox/VM
  - Personalização de aplicações Map Reduce
- Data Engineering (Gerenciamento/Ferramentas para Big Data)
  - NoSQL e NewSQL
  - Dados em movimento – Processamento de Streaming