FIAP

MBA+

MBA EM FULL STACK DEVELOPER

FIAP

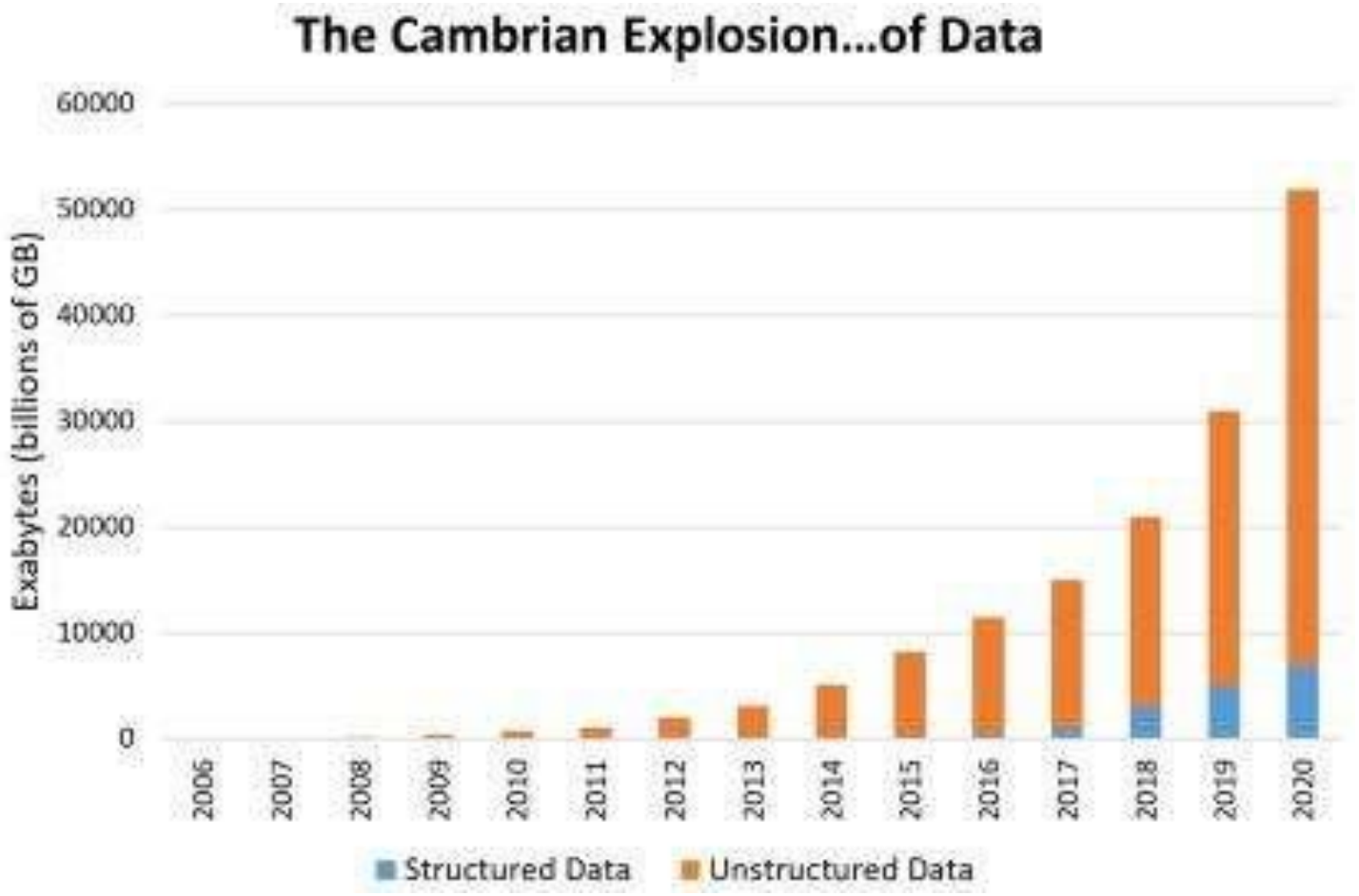
MBA+ | *Big Data Development*

Prof. Thiago Nascimento Nogueira
tnnogueira@gmail.com
Linked-in: https://www.linkedin.com/in/thnogueira/

FIAP

Introdução ao Hadoop

Explosão Big Data



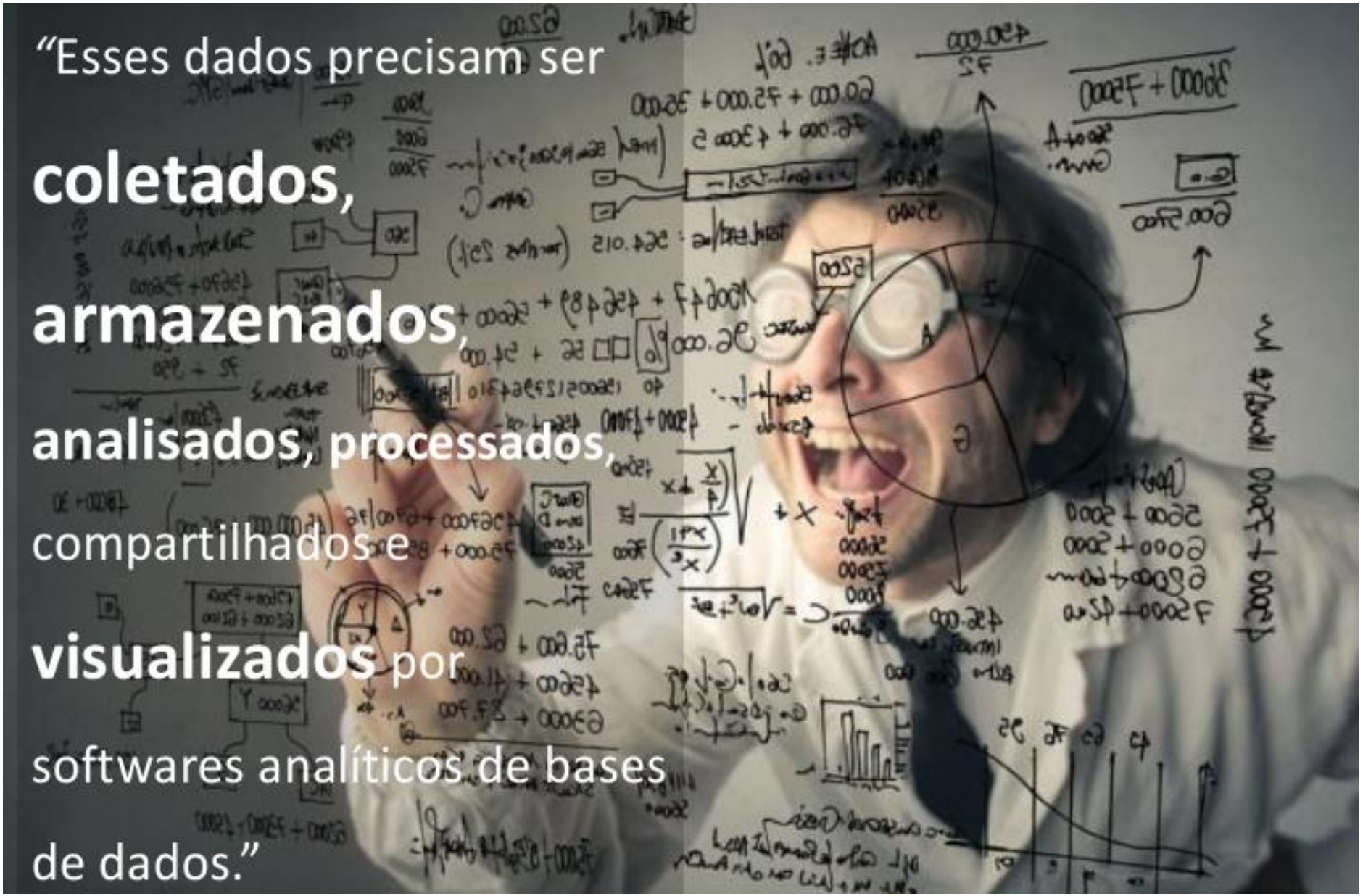
Explosão Big Data

- Produção de dados dobra a cada 18 meses
- 2,5 bilhões de GB são criados diariamente
- Facebook gera 4 PB de dados todo dia
- O mundo gera 2,5 EB de dados por dia
- Este ritmo está acelerando rapidamente com o crescimento do IoT

Byte = 8 bits, Kilobyte = 10^3, Megabyte = 10^6, Giga = 10^9,
Tera = 10^{12}, Peta = 10^{15}, Exa = 10^{18},
Zetta = 10^{21}, Yotta = 10^{24}

https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#3c13224360ba

Contexto


“Esses dados precisam ser
coletados,
armazenados,
analisados, processados,
compartilhados e
visualizados por
softwares analíticos de bases
de dados.”

Contexto

FIAP

V^5

Volume – a escala dos dados:

- As companhias tem em sua maioria 100 terabytes de dados armazenados internamente
- 7.7 bilhões de aparelhos celulares
- 3.1 bilhões de pessoas conectadas a Web
- 25 bilhões de equipamentos conectados a internet perto de 2020

Velocidade: analise do fluxo dos dados:

- A bolsa de Nova York captura 1 terabyte de informação em cada seção de negociação de ações
- Carros modernos possuem 100 sensores que monitoram diferentes sistemas
- Como esses dados podem ser processados em tempo real para produzir informações para decisões?

Variedade – diferentes formas de dados:

- A maior parte das organizações gasta até 80% do seu tempo modelando e preparando suas informações, em vez de estar analisando ela para produzir informações para decisão
- Harmonizar múltiplas fontes de dados pode ser complexo
- Os dados vêm de diferentes fontes e em diferentes formatos

Veracidade – origem dos dados

Valor – negócio e estratégia

Lei de Moore (parafraseado)

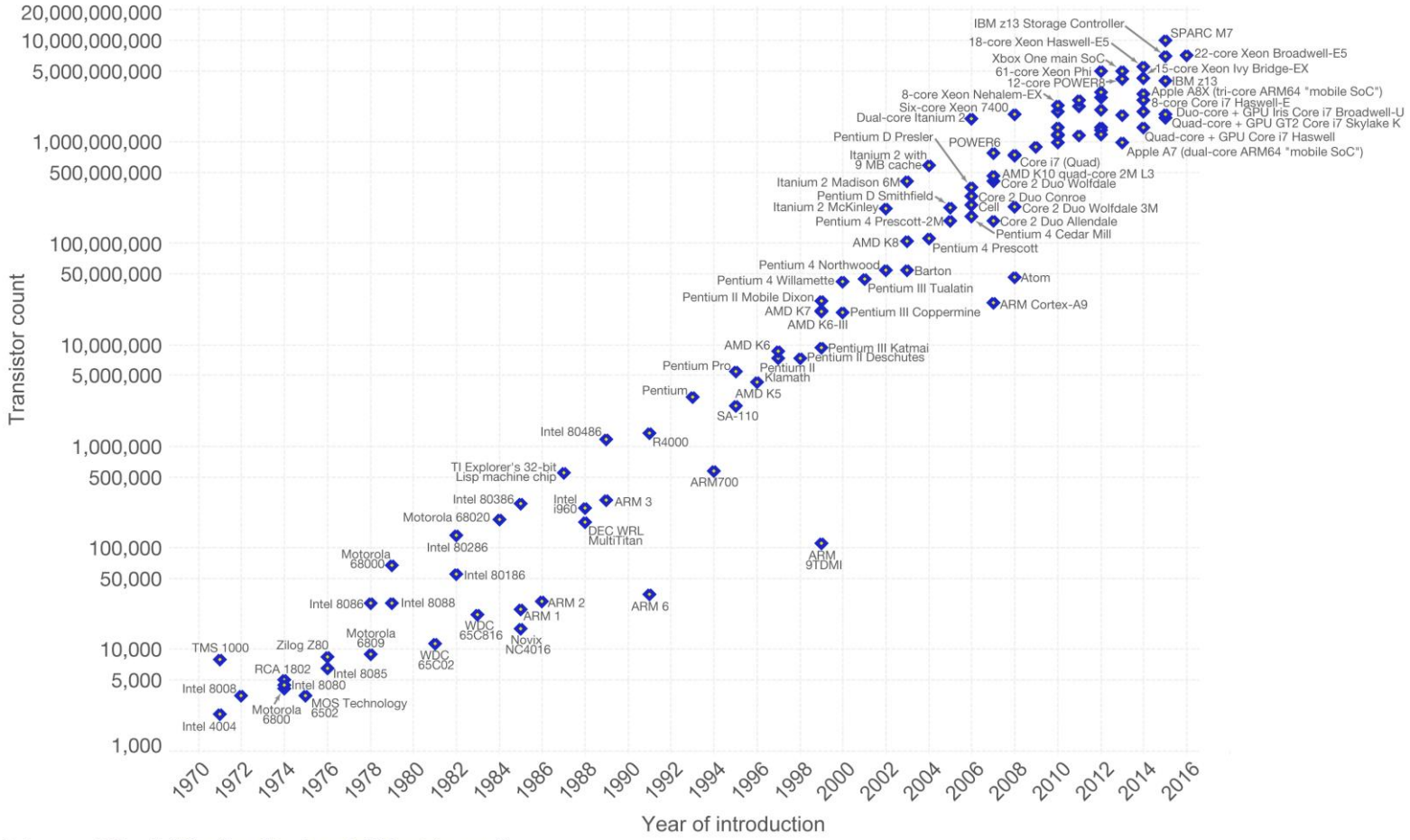
“O número de transistors em um circuito integrado dobra a cada 2 anos.”

- Gordon E. Moore

Lei de Moore

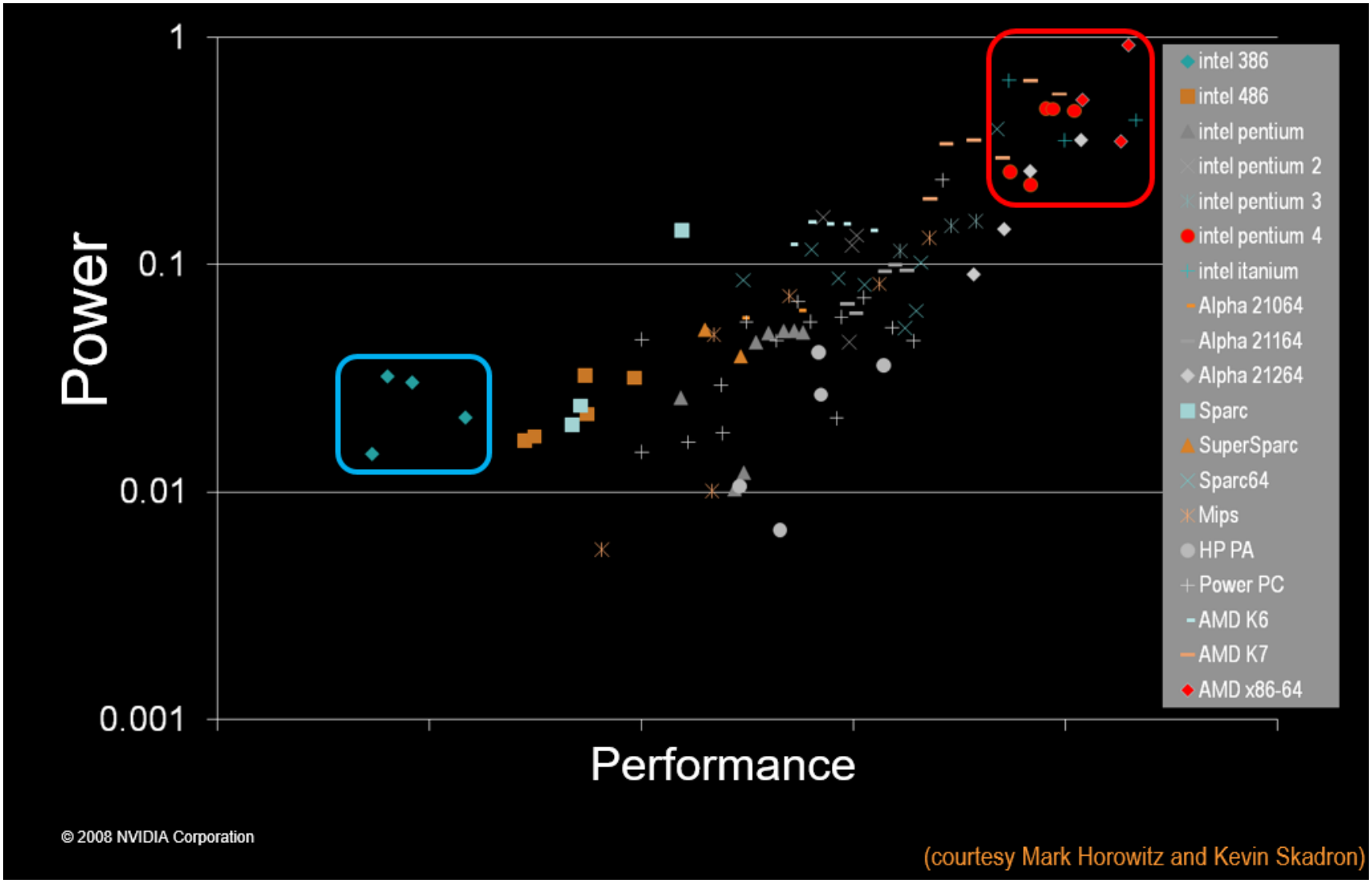
Moore's Law – The number of transistors on integrated circuit chips (1971-2016)

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are strongly linked to Moore's law.



Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

Consumo vs Performance



Limitações da performance serial

FIAP

- Não é possível continuar a escalar a frequência dos processadores (não há chips de 10Ghz)
- Não é possível continuar aumentando a potencia dos processadores (não queremos derreter os chips)

- **Solução**: PARALELIZAÇÃO DA COMPUTAÇÃO!

Nova Lei de Moore

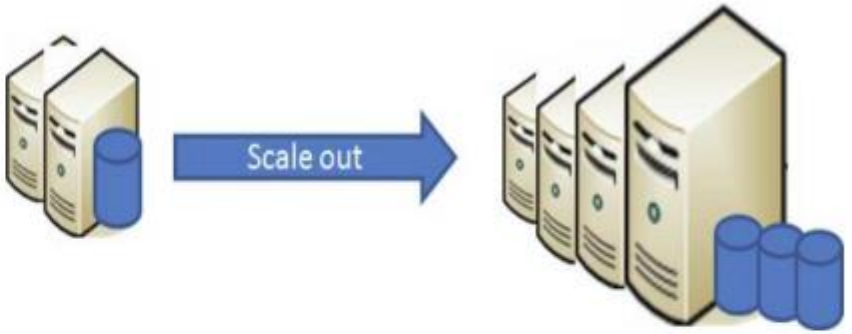
- Computadores não se tornam mais rápidos, e sim mais “largos”
- Precisamos repensar nossos algoritmos para serem paralelos
- Computação paralela é a solução mais escalável

Escalabilidade Vertical VS Horizontal

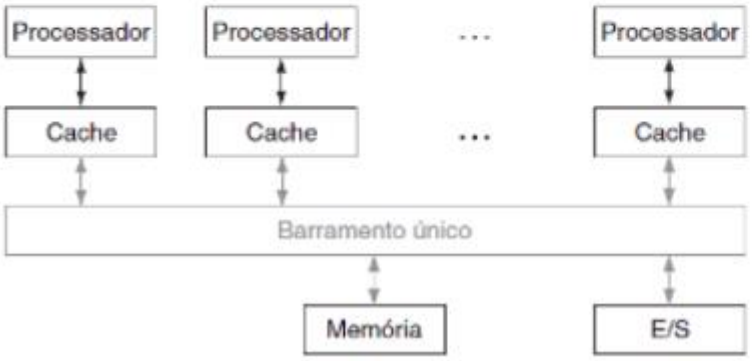
FIAP

Escalabilidade Vertical (Scale-up)

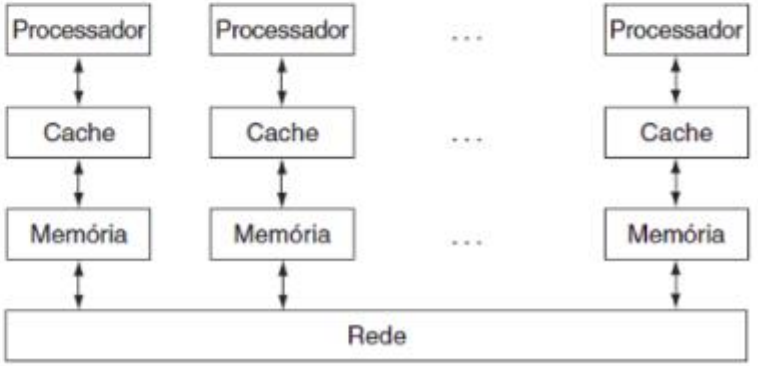
Escalabilidade Horizontal (Scale-out)



SMP (Symmetric Multi-Processing)



Cluster



Hadoop

Mudança de Paradigma

De: Levar o dado para o processamento (servidor)

Para: Levar processamento para o dado (distribuído)

Paper

FIAP

2003

The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung

Google*

2004

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

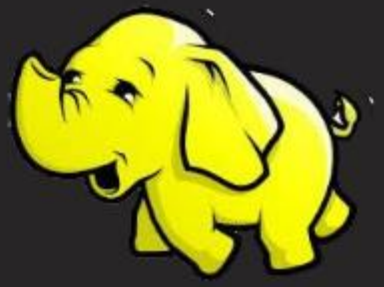
jeff@google.com, sanjay@google.com

Google, Inc.

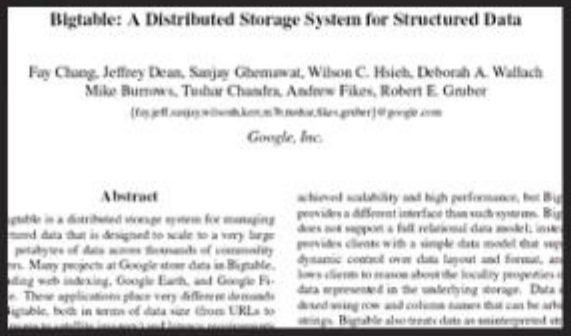
Paper

FIAP

2005: Doug Cutting e Michael J. Cafarella – desenvolveram Hadoop para o projeto do engine de busca Nutch - fundado pelo Yahoo.



2006: Yahoo passa o projeto para Apache Software Foundation.



Bigtable: A Distributed Storage System for Structured Data

Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach
Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber
{fay,jeff,sanjay,wilsonh,kerr,m3b,tushar,fikes,gruber}@google.com

Google, Inc.

Abstract

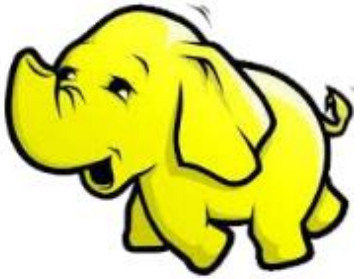
gtable is a distributed storage system for managing
tured data that is designed to scale to a very large
petabytes of data across thousands of commodity
rs. Many projects at Google store data in Bigtable,
ding web indexing, Google Earth, and Google Fi-
e. These applications place very different demands
igtable, both in terms of data size (from URLs to

achieved scalability and high performance, but Big
provides a different interface than such systems. Big
does not support a full relational data model; inste
provides clients with a simple data model that sup
dynamic control over data layout and format, an
lows clients to reason about the locality properties o
data represented in the underlying storage. Data
dexed using row and column names that can be arbi
strings. Bigtable also treats data as uninterpreted st

Paper

FIAP

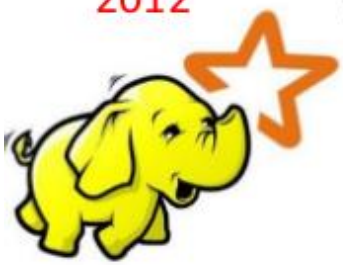
2008 Atinge escala web
Yahoo rodando em 10.000 servidores



2010 Avro, Hive, Pig e Hbase

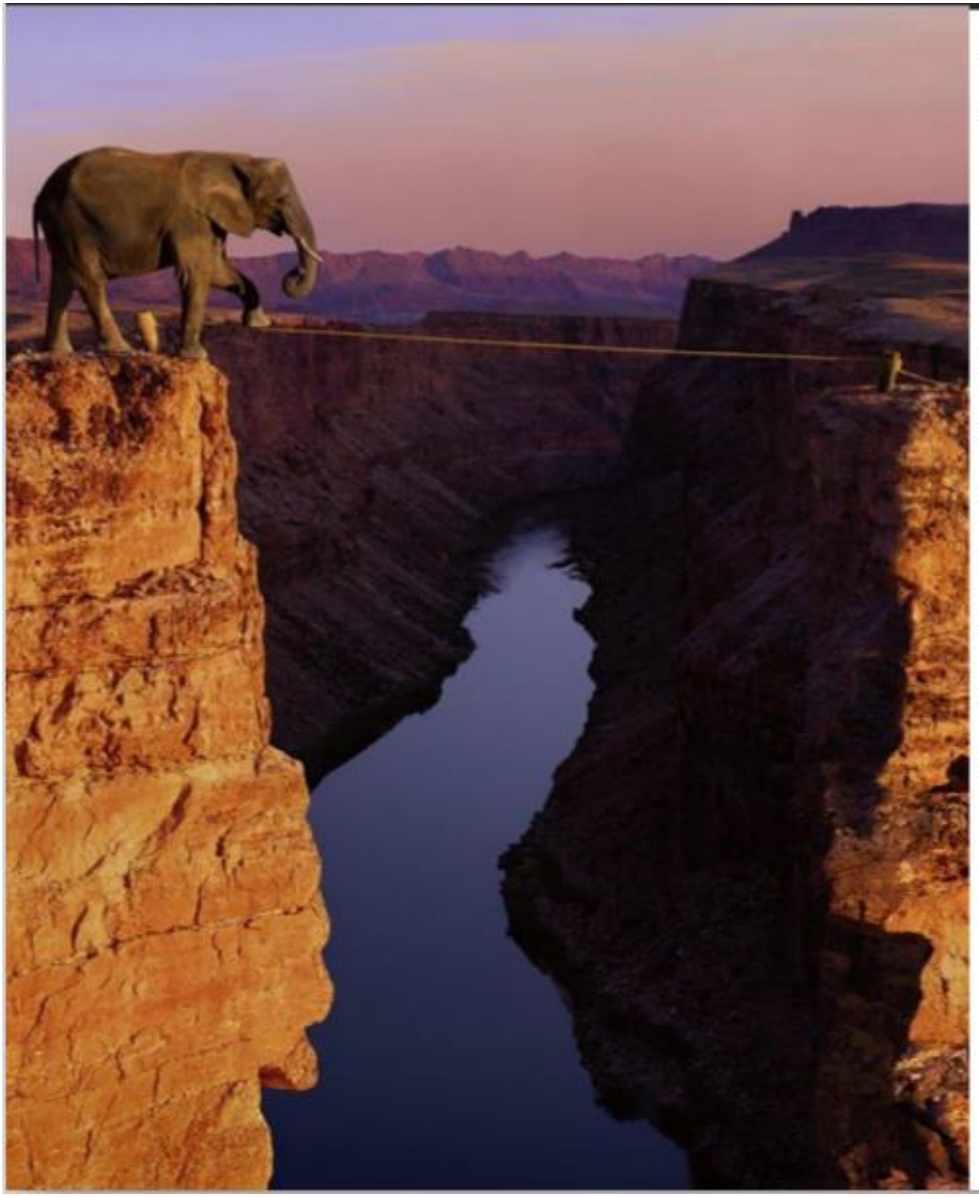
2011 Versão 1.0.0 disponibilizada

2012 Versão 2.0.0 disponibilizada (YARN)



Hadoop

FIAP



- Apache Hadoop [open source]
- Cloudera Enterprise | Cloudera CDH [open source]
- Hortonworks Data Platform HDP [open source]
- MapR
- IBM Big Insights
- Amazon Elastic MapReduce (EMR)
- Microsoft Azure HDInsight
- Google Cloud Dataproc
- Oracle Big Data Cloud Services
- SAP Cloud Platform Big Data Service (formerly SAP Altiscale Data Cloud)

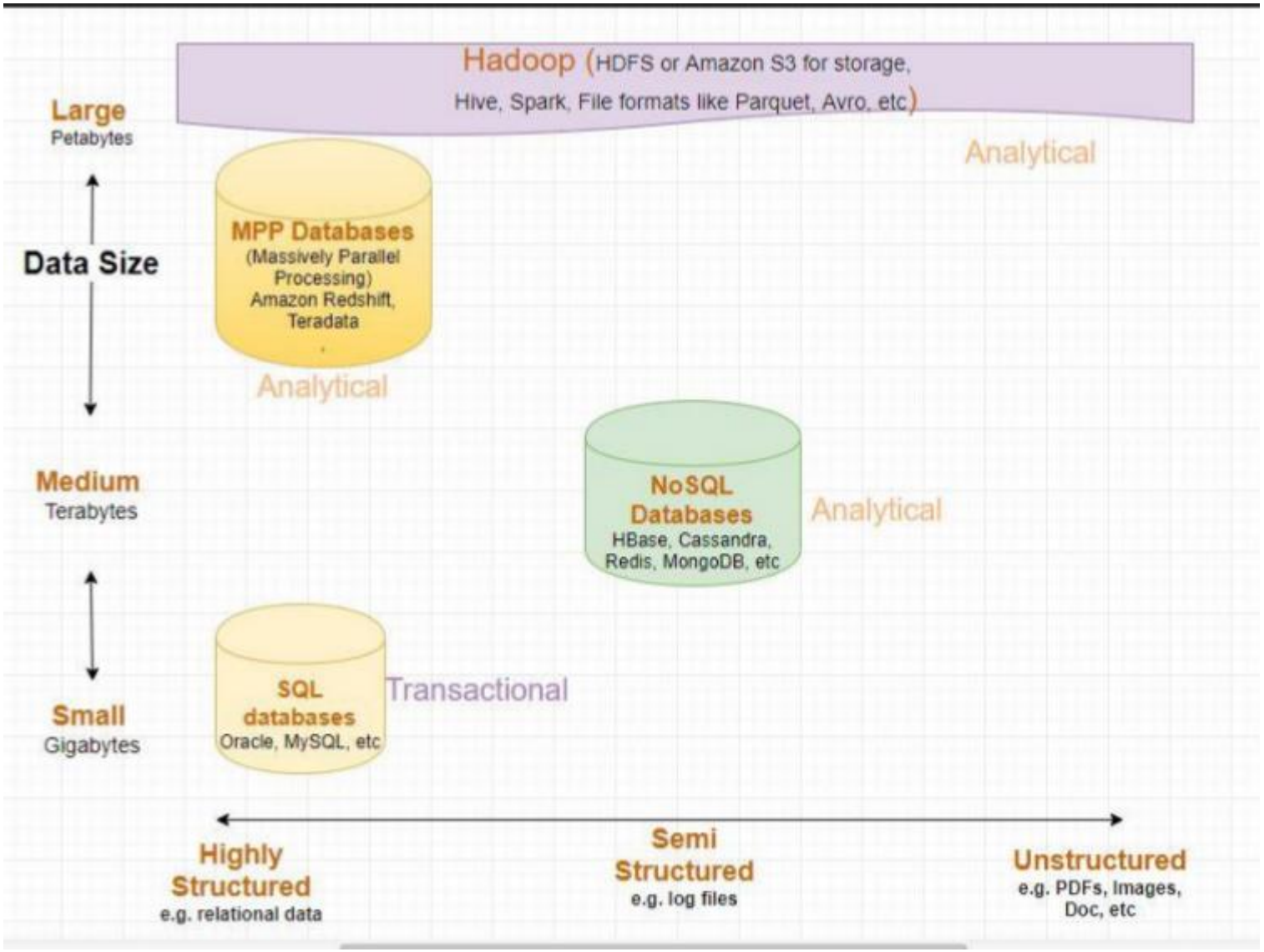
cloudera

MAPR TECHNOLOGIES

Hortonworks

amazon web services

Contexto Big Data

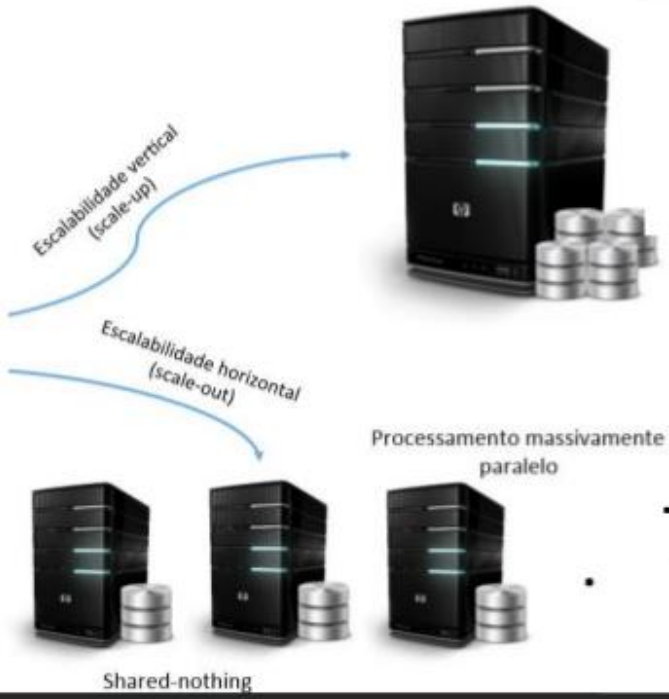


Características

FIAP

Características Infraestrutura

Alto grau de **paralelismo**



Recovery automático

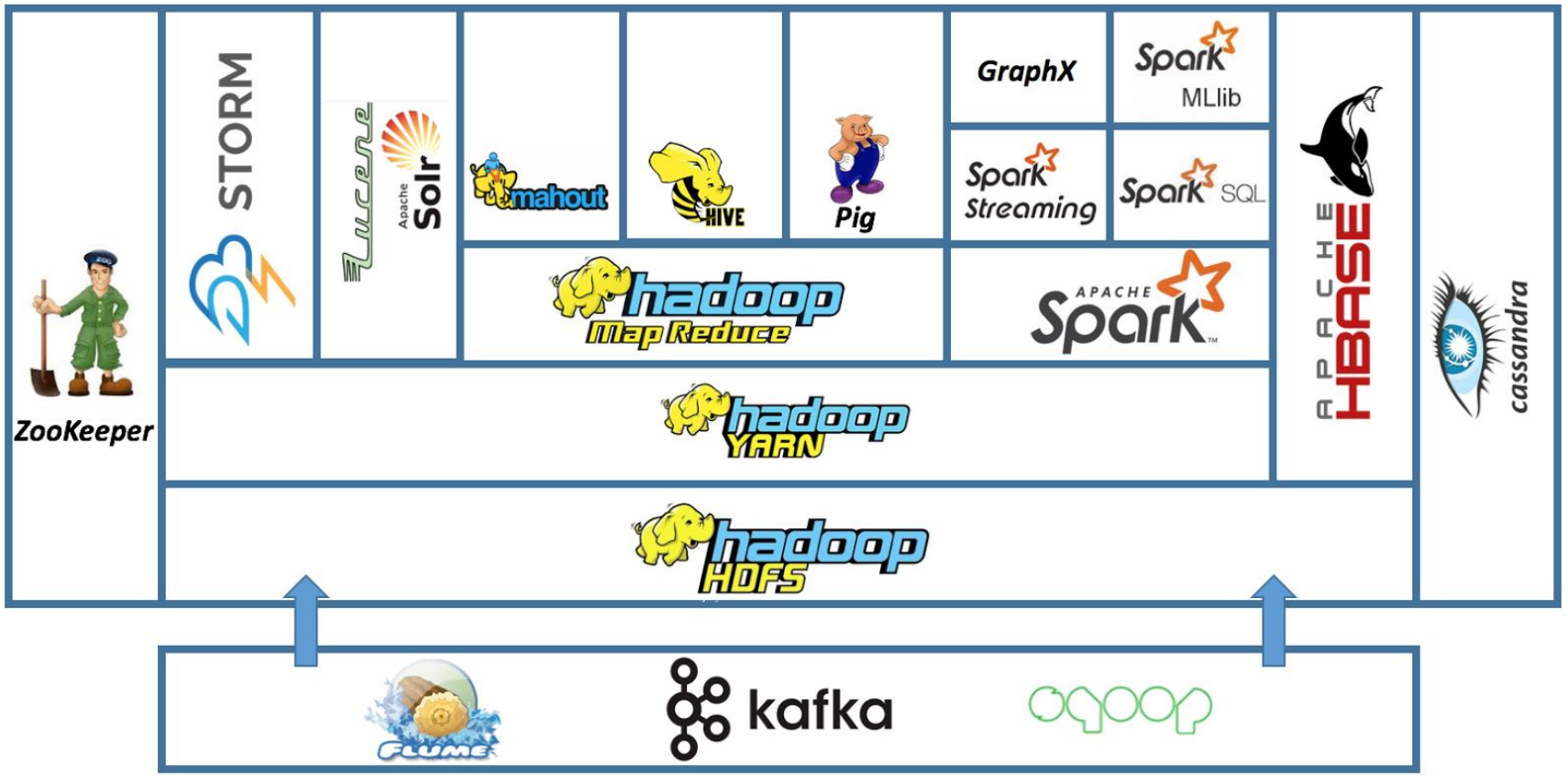
Processamento de dados **distribuído**

Tolerância a falhas

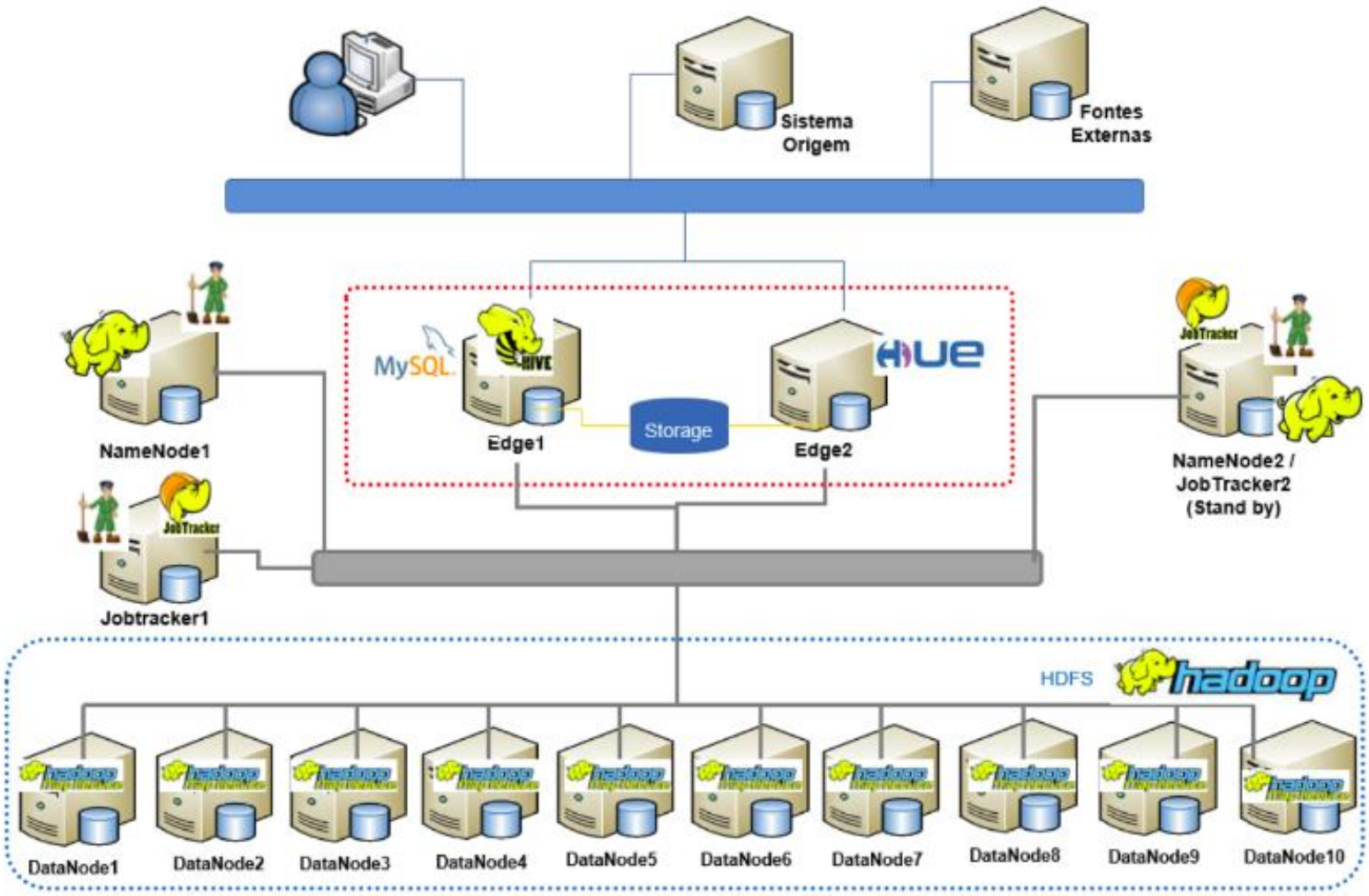
Alto **Throughput**

Escalabilidade linear

Ecossistema Hadoop



Arquitetura da infraestrutura Hadoop



Uso de Hadoop

Vantagens

- Software livre: gratuito, comunidade ativa, rápida evolução, apoio de grandes corporações;
- Roda em hardware commodity tanto máquina quanto rede
- Serviços em nuvem – Amazon Elastic MapReduce, Google App Engine MR;
- Gerencia replicação dos dados e armazena metadados;
- Escalabilidade horizontal – cluster com milhares de nós
- Desenvolvedores podem focar apenas na abstração do problema de negócio – Hadoop faz o trabalho pesado

Desvantagens

- Não dá pra resolver problemas não paralelizáveis
- Não vale a pena para arquivos pequenos (overhead)
- Muito processamento em um pequeno conjunto de dados

Hadoop Distributed File Systems (HDFS)

O que é Hadoop Distributed File Systems (HDFS)?

FIAP

- **Definição:** Um File System distribuído que é executado em grandes clusters utilizando hardware commodity. Foi inspirado no Google File System (GFS).

- **Objetivo:** Ser utilizado como um RAIN (Redundant Array of Nodes) para armazenar dados utilizados por diversos componentes do ecossistema Hadoop.

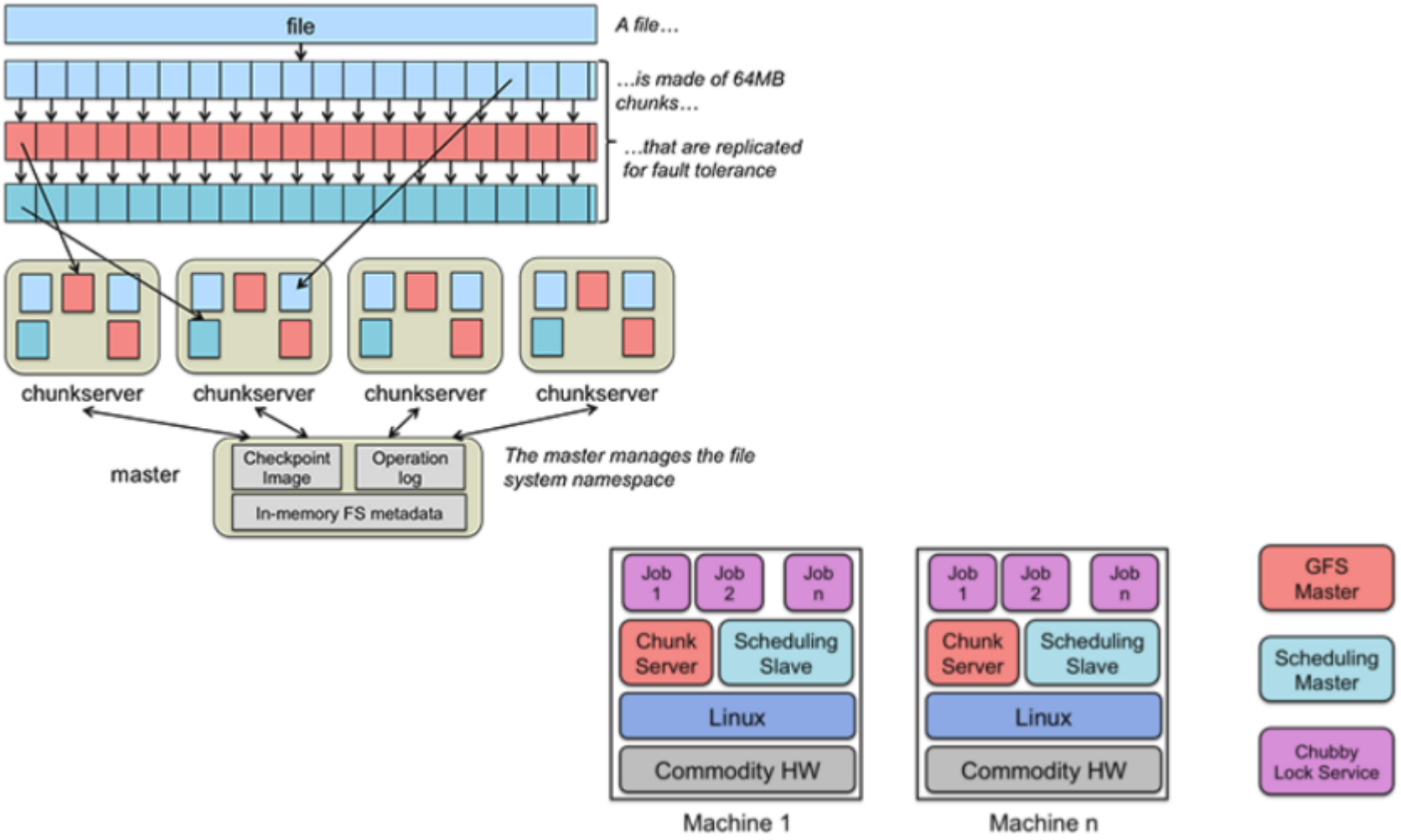
- **Motivações:**
  - A capacidade de armazenamento dos discos cresceu massivamente. No entanto, a velocidade de acesso aos dados não acompanhou esse crescimento.
    - » Um disco comum em 1990 podia armazenar 1,370 MBs de dados e tinha uma velocidade de transferência de 4.4 MB/s. Era possível ler todo o conteúdo do disco em aproximadamente 5 minutos.
    - » Após 20 anos, discos de 1 TB são comuns, mas a transferência de dados aumentou apenas para 100MB/s. Para ler todo o conteúdo do disco leva mais de 2 horas e 30 min.
    - » Solução: para reduzir o tempo de acesso aos dados, é necessário realizar o acesso à vários discos de uma vez. Distribuindo 1 TB em 100 discos, serão armazenados 10 GBs em cada disco, o acesso a todo o dado armazenado leva aproximadamente 2 minutos.

Fonte: White (2012)

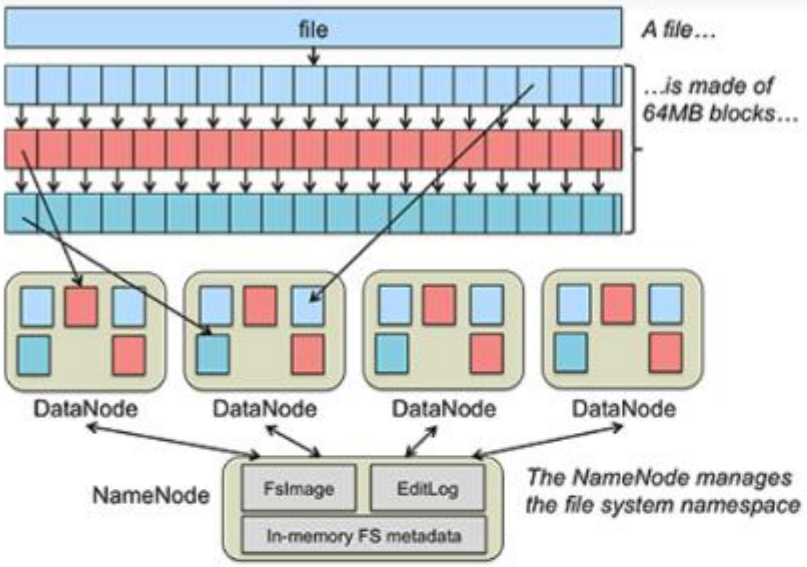
Principais Características

- Alto desempenho - leitura e escrita em paralelo com alto throughput e baixa latência.
- Tolerância a falhas – replicação dos dados e rack awareness (ex: 3 réplicas – primeira réplica em um DataNode, segunda em outro DataNode no mesmo rack e a terceira em outro DataNode outro rack).
- Administração relativamente simples – Arquitetura Master-Slave.
- Escalabilidade horizontal – milhares de servidores e discos.
- Otimizado para processamento do MapReduce – processamento local.
- Write-Once, Read-Many (WORM).
- Otimizado para armazenar arquivos grandes – preferencialmente igual ou maior do que o tamanho do block size utilizado. Tamanho ideal > 1 GB.
- Desenvolvido em Java.
- Possibilidade de fácil integração com APIs.
- Otimiza armazenamento e transferência de dados pela rede por meio da compressão de dados.
- Não suporta todas as funcionalidades do padrão Portable Operating System Interface (POSIX) que geralmente é utilizado em file systems no UNIX e Linux. Possui suporte às funções básicas: criar, deletar, abrir, fechar, ler e escrever.

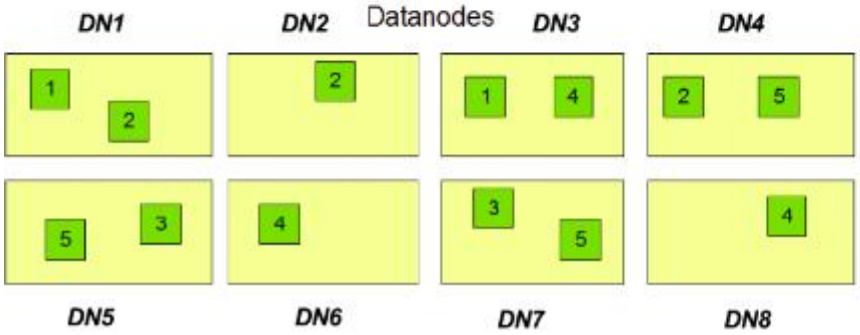
Arquitetura do GFS

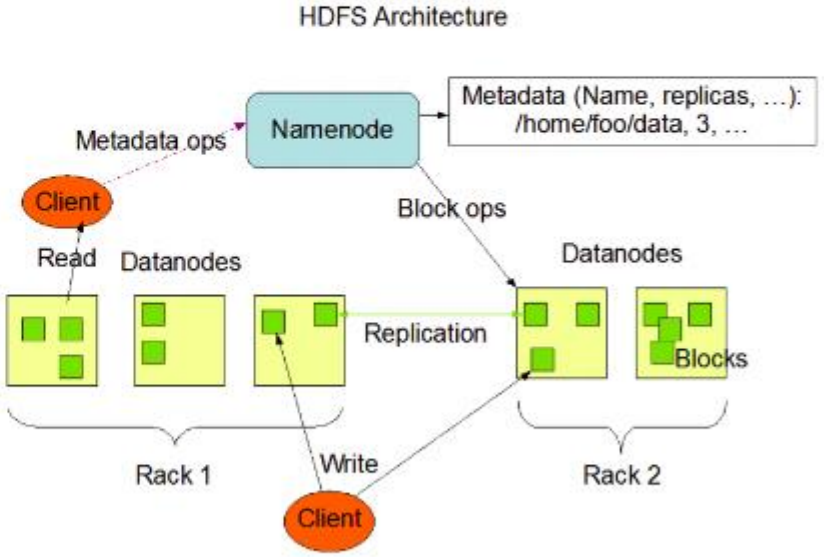


Arquitetura do HDFS



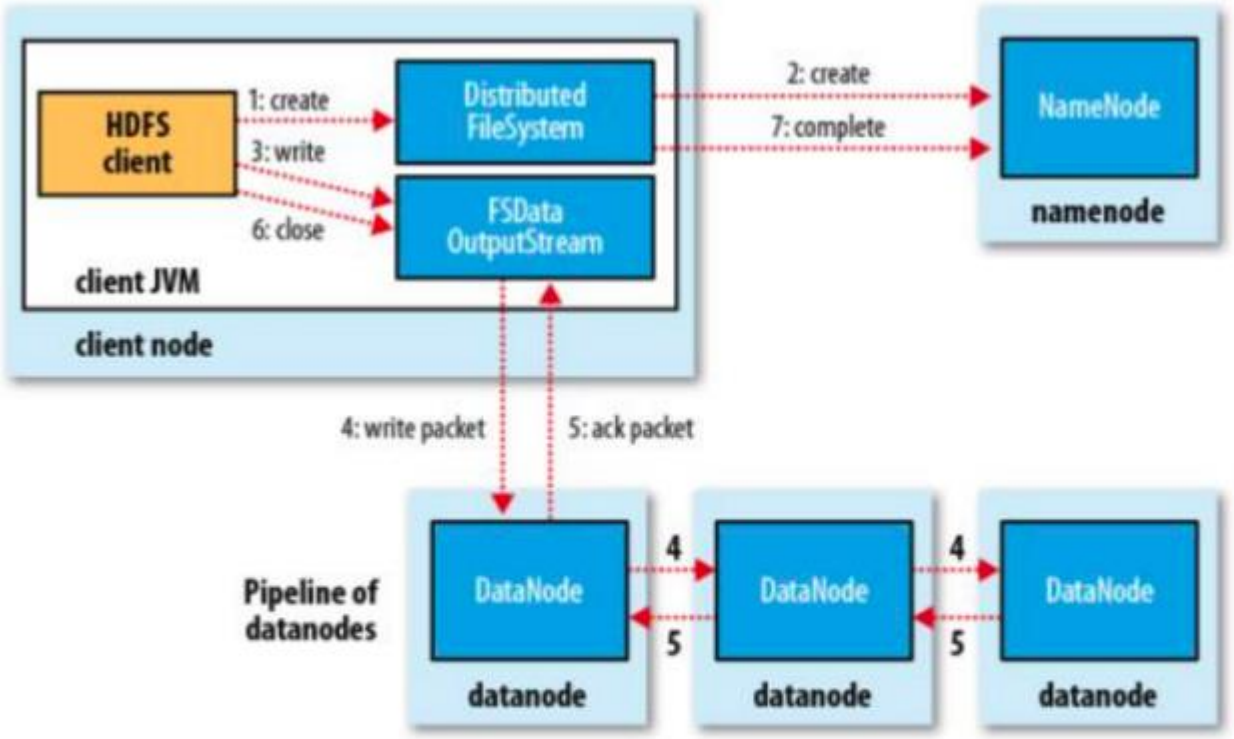
GFS Name	HDFS Name
Master	NameNode
Chunkserver	DataNode
chunk	block
Checkpoint image	FsImage
Operation log	EditLog





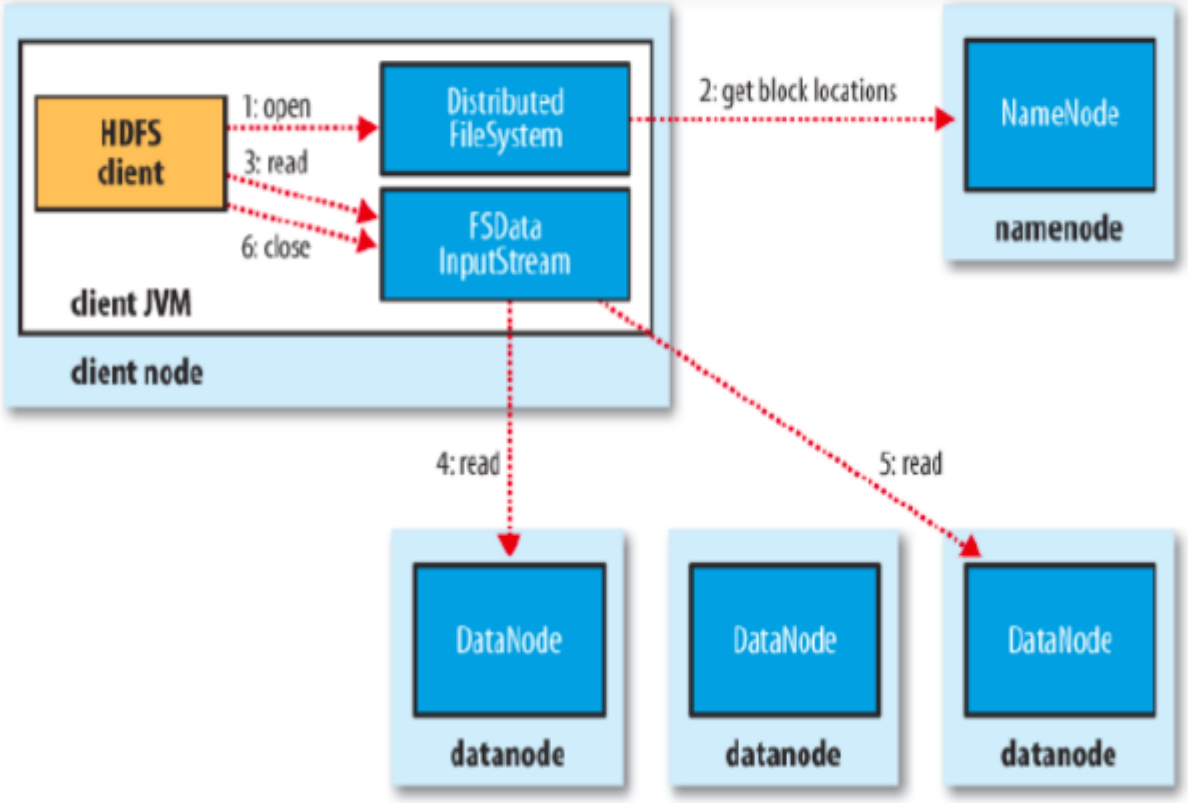
Fonte: APACHE (2013), GHEMAWAT et al. (2003) e Krzyzanowski (2012).

HDFS - Fluxo de Escrita



Fonte: White (2012)

HDFS - Fluxo de Leitura



- NameNode não é um gargalo.
  - O dado nunca passa por ele tanto para escrita, leitura ou replicação.
  - Ele carrega todos os metadados em memória.
  - Cada objeto no metadados consome entre 150 e 200 bytes de RAM. Objetos: nome de arquivo e informações sobre os blocos.

Fonte: White (2012)

Racional para definir Sizing do HDFS

Identificaçao do Consumo	Descrição	Total Disponível (TBs)
1	Espaço Bruto (10 DNs x 12 Discos x 4 TBs)	480
2	Saldo = (ID 1) - (Espaço consumido por formatação do disco e file systems ext4 = 10%)	432
3	Saldo = (ID 2) - (Espaço reservado para Tarefas MapReduce - Arquivos intermediários - área de work. Parâmetro: dfs.datanode.du.reserved = 25% de cada disco = 0,9 TB))	324
4	Saldo = (ID 3) / (Quantidade de replicas = 3)	108
5	Saldo = (ID 4) - (ID 4)*(% de espaço reservado backup)	54
6	Saldo final disponível para uso	54

MapReduce

Características do MapReduce

- Paralelização e distribuição de tarefas de forma automática e transparente.
- Possui tolerância a falhas. Ex: caso um Servidor Slave falhar o processo é submetido a outro servidor que possua as cópias dos blocos que serão processados.
- Cria uma camada de abstração para os desenvolvedores que, em muitos casos, não precisam conhecer nem mesmo código Java.
- Monitora os status dos Jobs.
- Processamento local – Submete Tasks para os servidores que possuem os blocos que serão processados.
- Eficiência no tráfego de rede.
  - Os arquivos intermediários também são escritos localmente na etapa de map para diminuir o tráfego na rede.
  - Entre a etapa de Map e Reduce ocorre o processo de shuffle e sort (transferência pela rede entre Map e Reduce Tasks).
  - Os arquivos intermediários podem ser comprimidos antes de serem transferidos aos DNs onde serão executados os reducers.
- Redundância na execução processos para mitigar riscos impactos de performance – Backup Task.

Fonte: White (2012) e Impetus (2009)

O que é MapReduce?

- **Definição:** É um modelo de programação para processamento de dados com chave e valor. Não é uma linguagem ou uma plataforma. Foi inspirado no paper MapReduce do Google - 2004.
  - Consiste basicamente em três etapas:
    - » Map: Extrai algo que você espera de cada registro.
    - » Sort e Shuffle
    - » Reduce: Agrega, sumariza, filtra ou transforma os dados.
- **Objetivo:** Facilitar a distribuição das tarefas e execução em paralelo entre os nodes de um cluster Hadoop.
- **Motivações:**
  - Facilitar o desenvolvimento e execução de aplicações utilizando processamento paralelo.
  - Processar um volume massivo de dados utilizando uma infraestrutura de HW commodity.

Fonte: GHEMAWAT et al. (2004)

Fluxo MapReduce - Exemplo

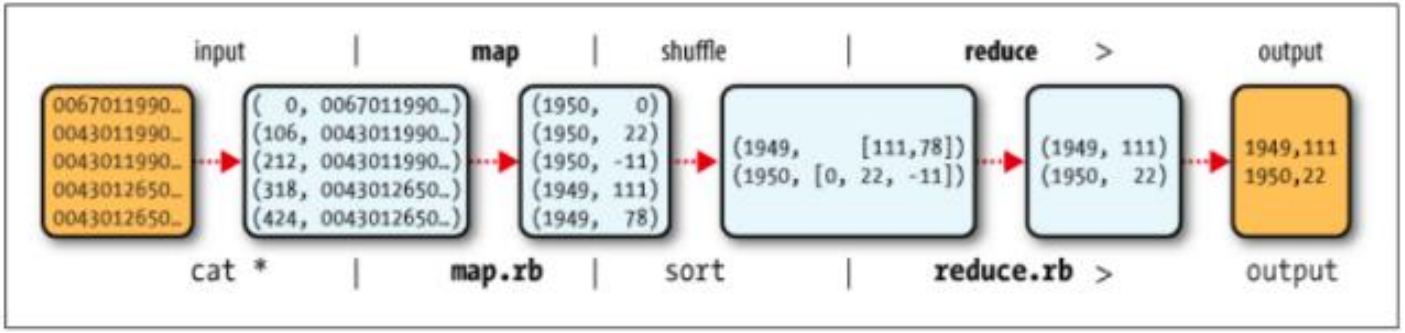
- Obter a maior temperatura (em Graus Celsius) em cada ano.

Arquivo de input

```
(0, 0067011990999991950051507004...9999999N9+00001+99999999999...)
(106, 0043011990999991950051512004...9999999N9+00221+99999999999...)
(212, 0043011990999991950051518004...9999999N9-00111+99999999999...)
(318, 0043012650999991949032412004...0500001N9+01111+99999999999...)
(424, 0043012650999991949032418004...0500001N9+00781+99999999999...)
```

Input:
Chave (Offset – início da linha)
Valor (linha)

Output:
Chave (Ano)
Valor (temperatura)



Map – Shuffle – Reduce

FIAP

Map

Cada nó na fase **map** emite pares chave-valor com base no registro de entrada, um registro por vez.

Shuffle

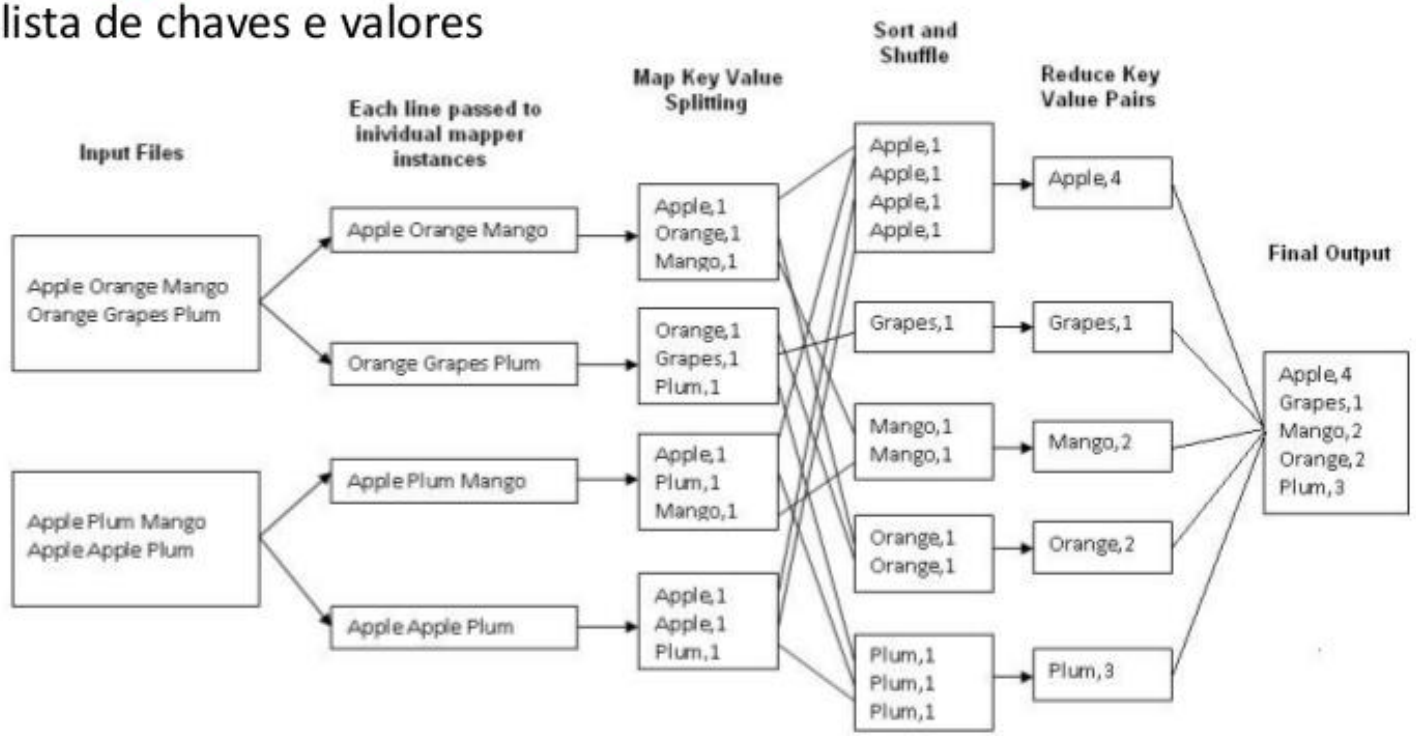
A fase **shuffle** é tratada pela estrutura Hadoop. Ela transfere os resultados dos **mappers** para os **reducers** juntando os resultados por meio **chave**.

Reduce

A saída dos **mappers** é enviado para os **reducers**. Os dados na fase **reduce** são divididos em partições onde cada **reducer** lê uma **chave e uma lista de valores** associados a essa chave. Os reducers emitem **zero ou mais pares chave-valor** com base na lógica utilizada.

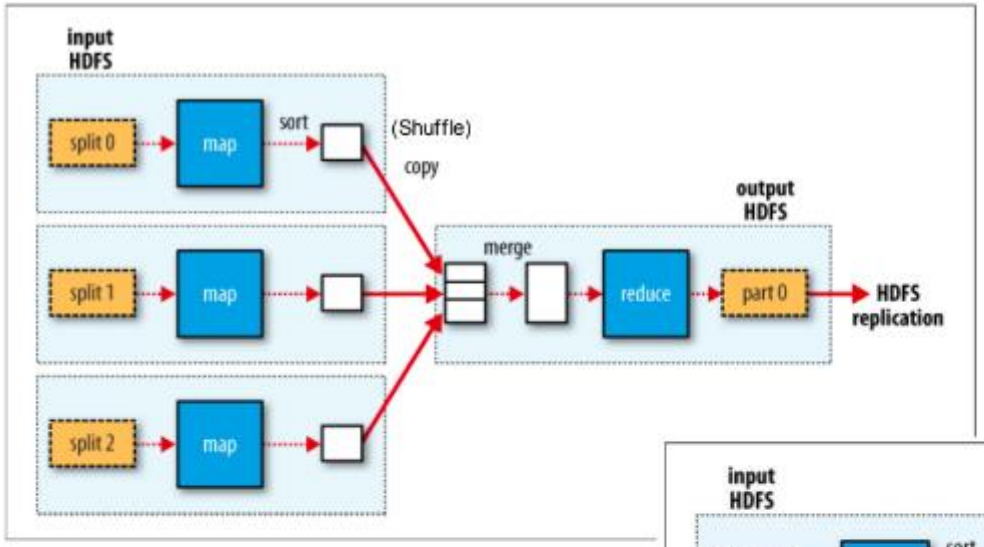
Fluxo de Processamento MapReduce

- A função **Map** atua sobre um conjunto de entrada com chaves e valores, produzindo uma lista de chaves e valores



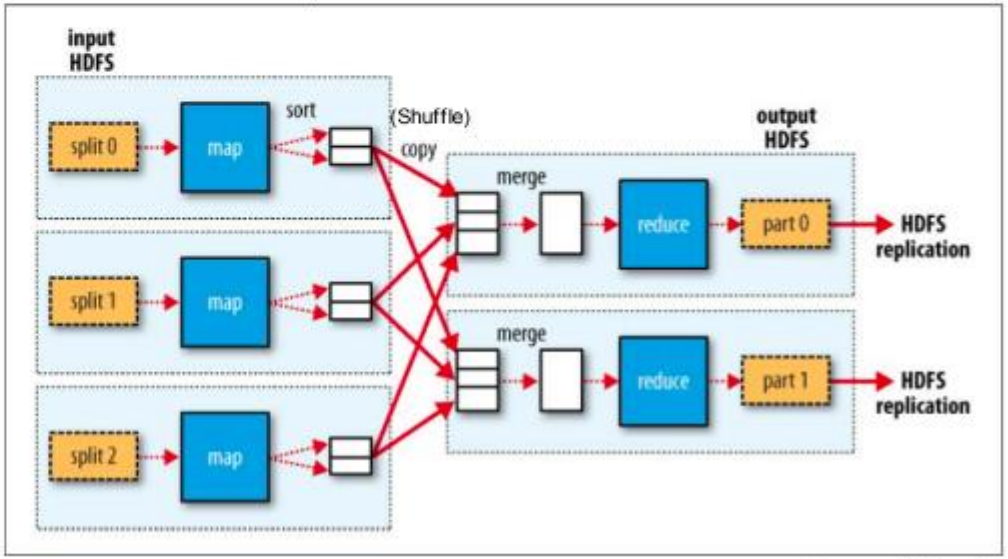
- A função **Reduce** atua sobre os valores intermediários produzidos pelo Map para, normalmente, agrupar os valores e produzir a saída

Fluxo de Processamento MapReduce



Único Reducer

Dois reducers



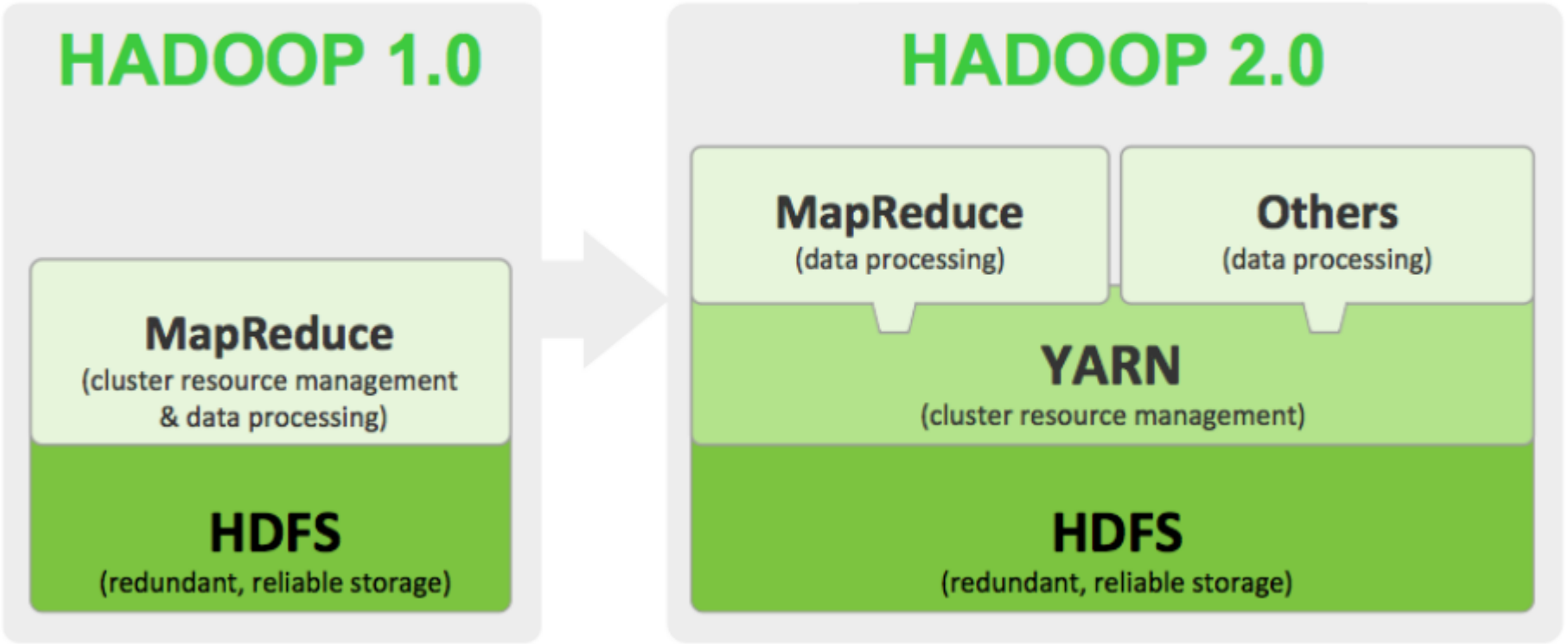
Fonte: White (2012)

YARN

Yet Another Resource Manager

FIAP

MapReduce vs YARN



YARN – Definições

FIAP

- Um host é o termo Hadoop para um computador (também chamado de nó, na terminologia YARN).

- Um cluster é composto por dois ou mais hosts conectados por uma rede local de alta velocidade.

- Do ponto de vista de Hadoop, pode haver vários milhares de hosts em um cluster.

- No Hadoop, existem dois tipos de hosts no cluster
  - O *master* host é o ponto de comunicação para um programa cliente.
  - Um *master* host envia o trabalho para o resto do cluster, que consiste em *worker* hosts.

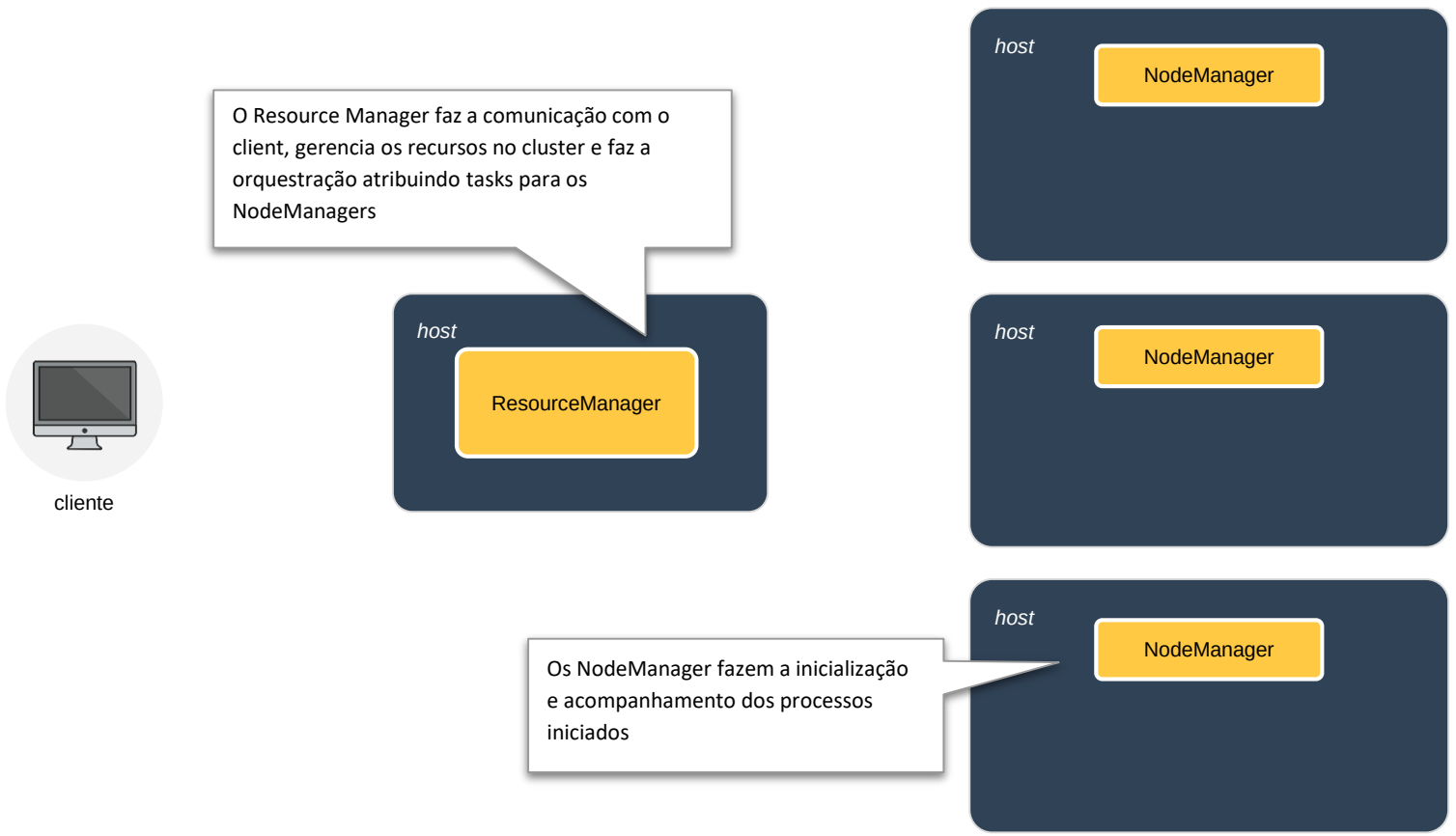
Componentes do MapReduce v2 – YARN (Yet Another Resource Negotiator)

- **ResourceManager:** O ResourceManager é a autoridade final que arbitra recursos entre todos os aplicativos no cluster. Um por cluster.

- **JobHistoryServer:** Armazena as métricas e metadados dos Jobs. Um por cluster.

- **NodeManager:** É essencialmente limitado à gestão de containers, não se preocupa com o gerenciamento de estado por aplicativo, pode escalar muito mais facilmente seu código. Um por slave node.
  - **Containers:** Alocação de recursos como (Memória, CPU, Network, Disco ...).
  - **Application Master:** Tem a responsabilidade de negociar os recursos apropriados (containers) com o ResourceManager, acompanhar o seu status e monitorar o progresso.

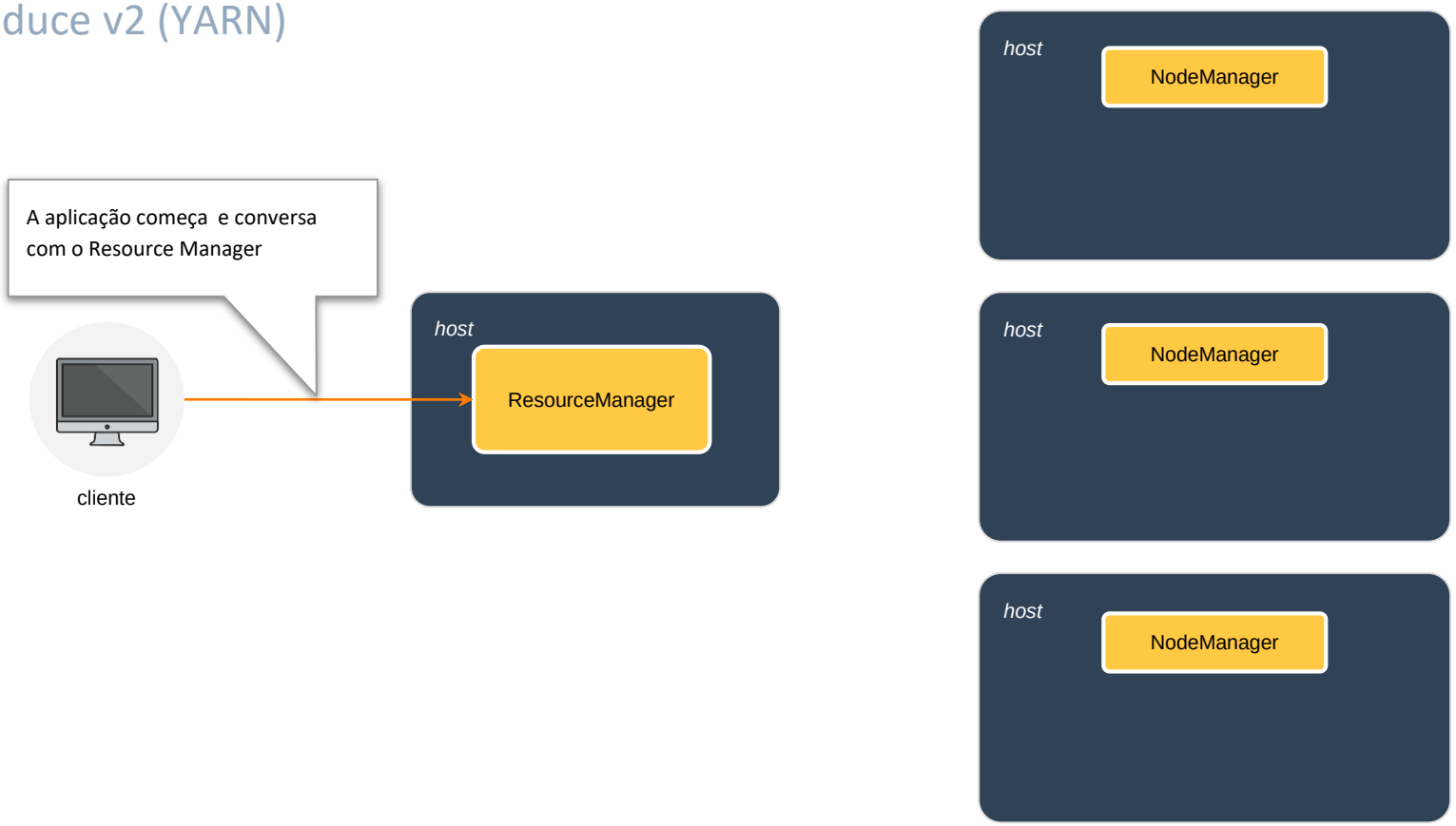
Fonte: White (2012)


O Resource Manager faz a comunicação com o client, gerencia os recursos no cluster e faz a orquestração atribuindo tasks para os NodeManagers
host
ResourceManager
cliente
host
NodeManager
host
NodeManager
host
NodeManager
Os NodeManager fazem a inicialização e acompanhamento dos processos iniciados

YARN – Fluxo de execução do Job

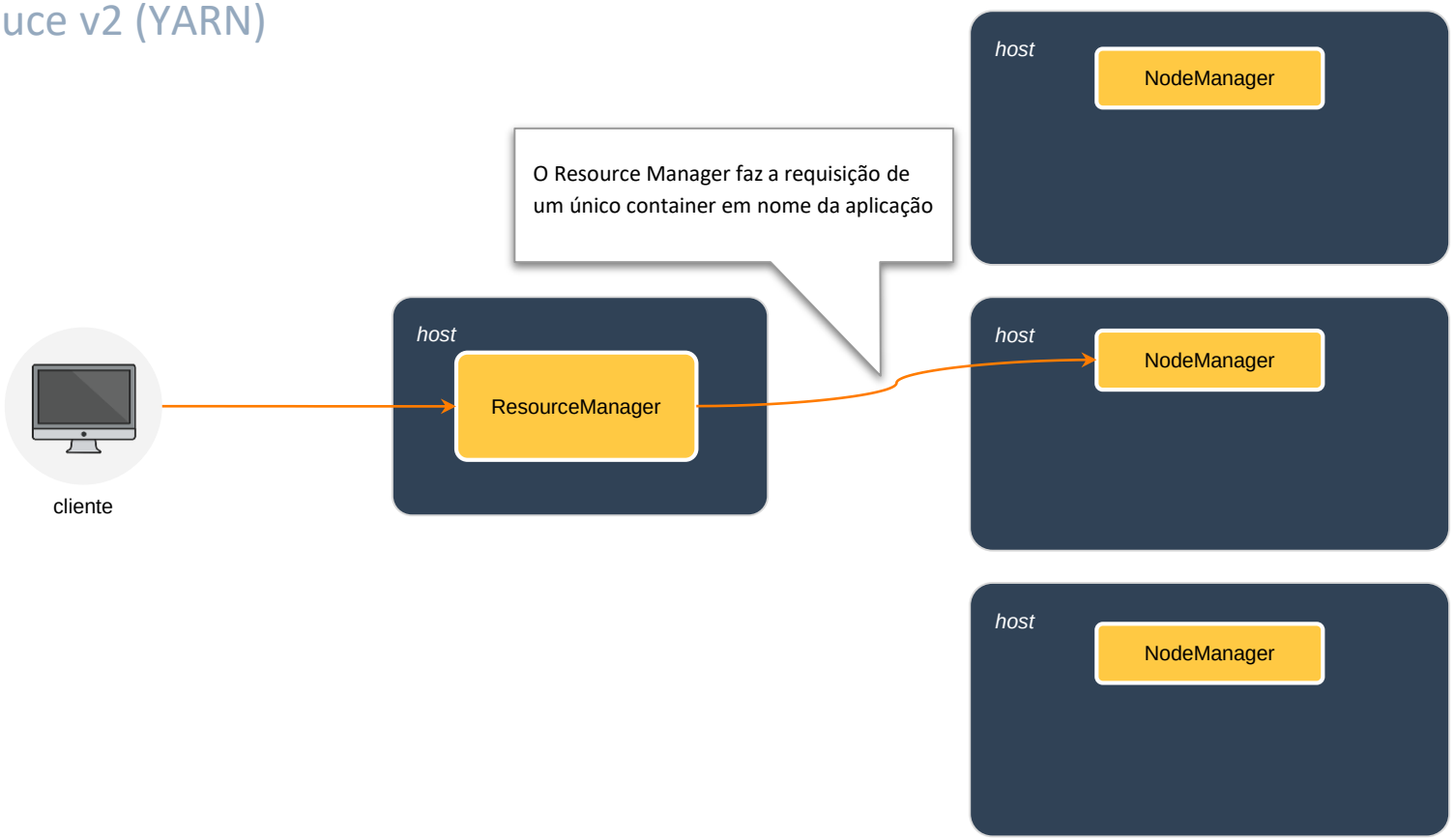
YARN

MapReduce v2 (YARN)



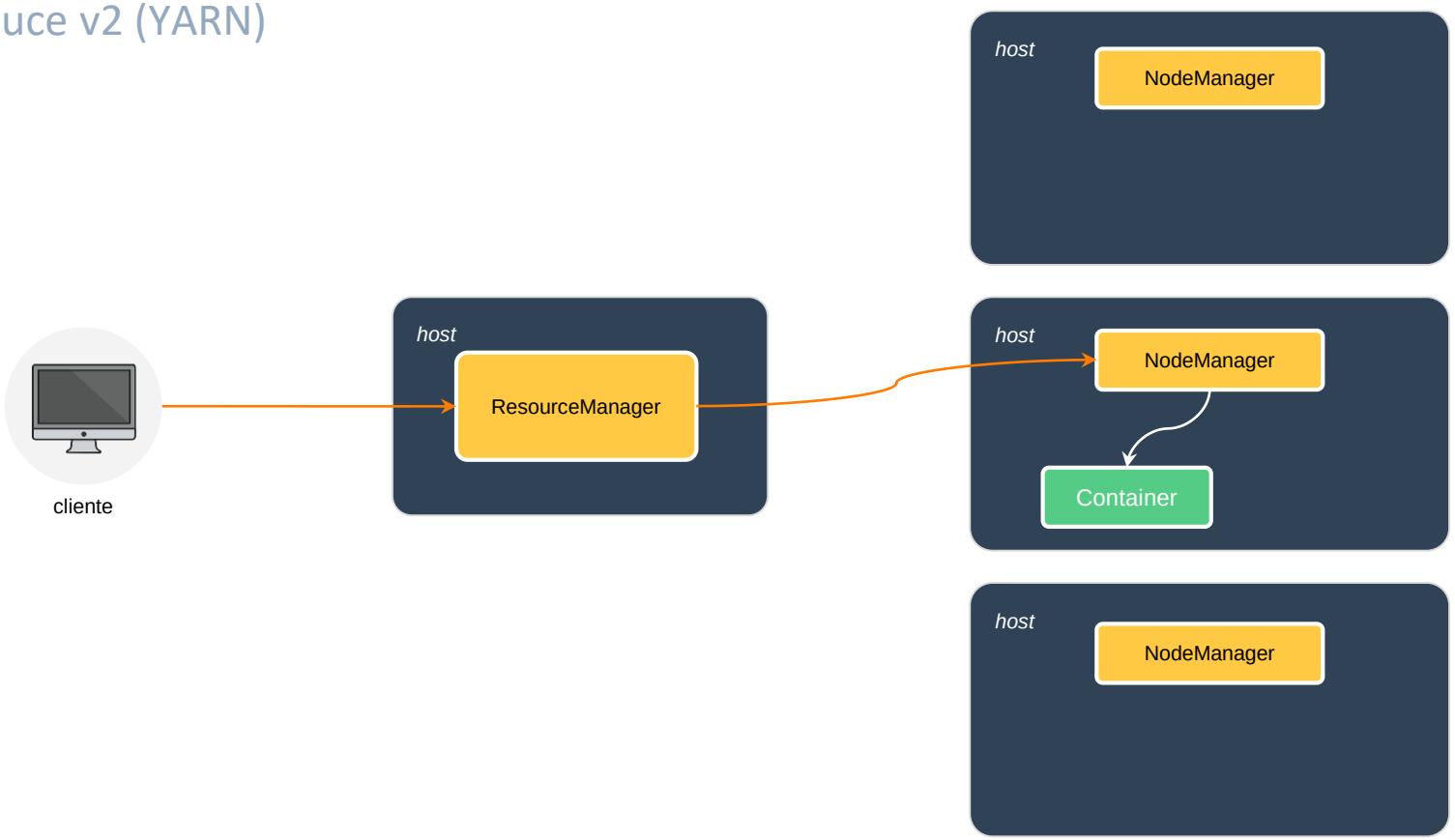
YARN – Fluxo de execução do Job

YARN
MapReduce v2 (YARN)



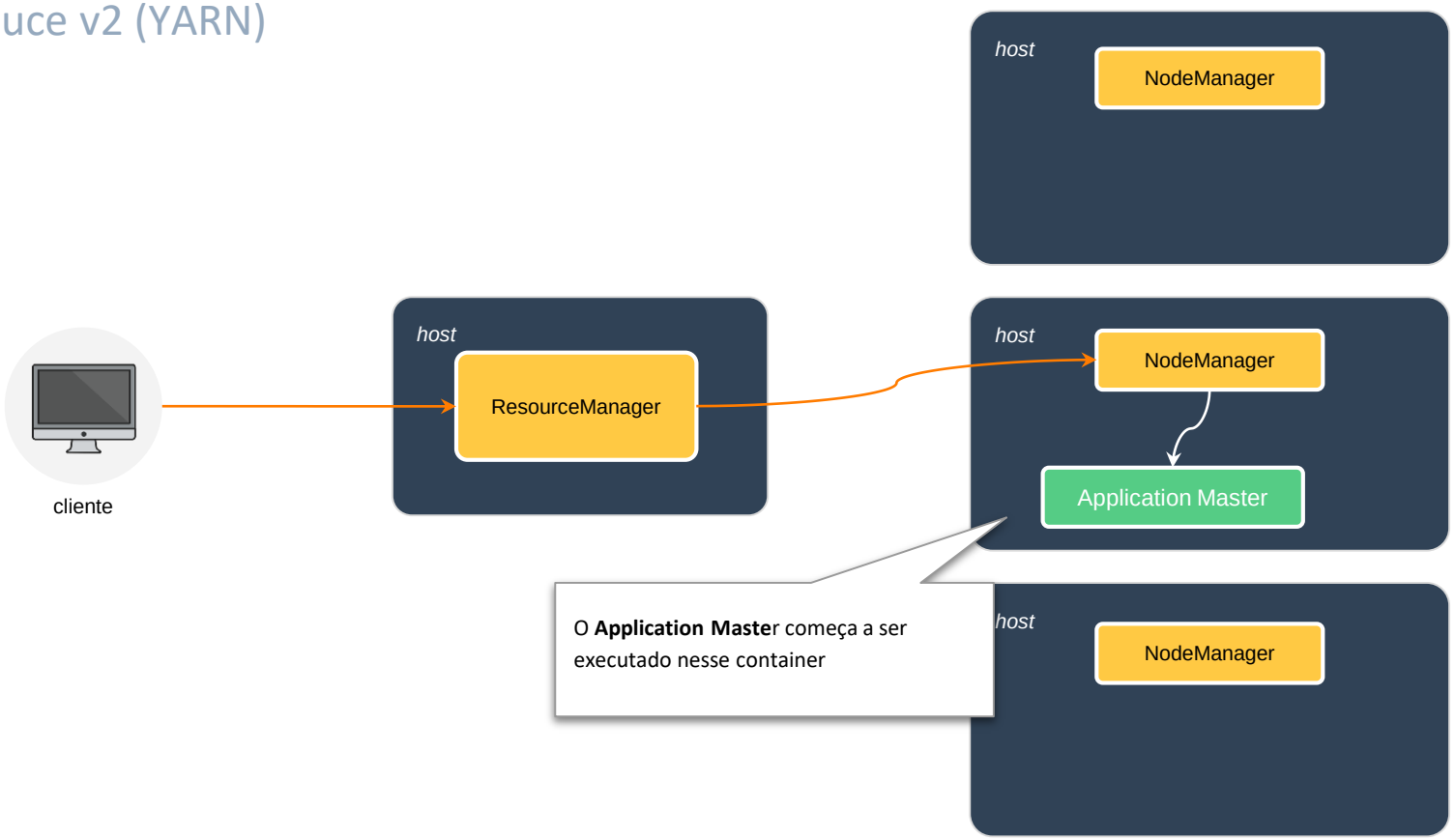
YARN – Fluxo de execução do Job

YARN
MapReduce v2 (YARN)



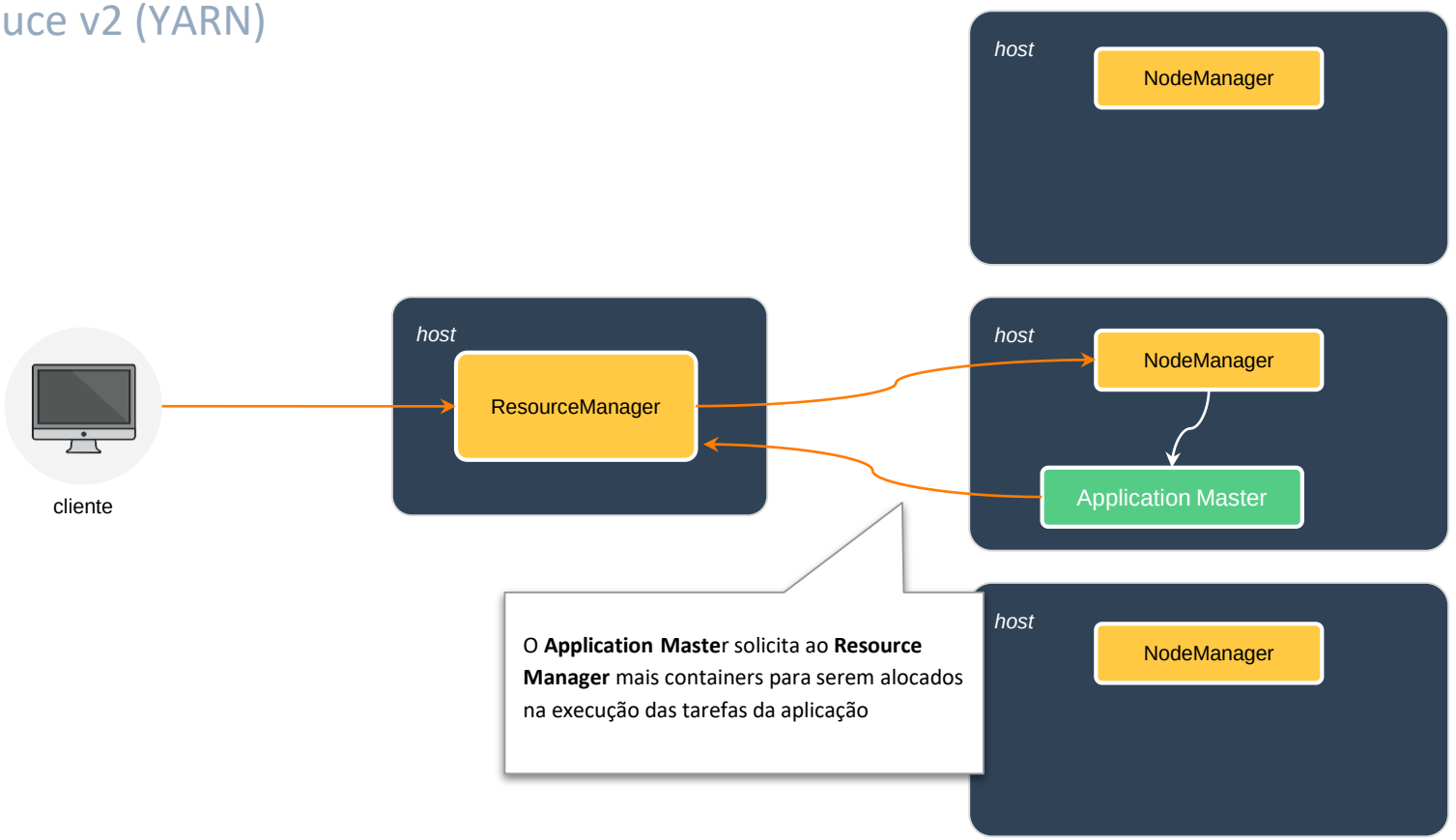
YARN – Fluxo de execução do Job

YARN
MapReduce v2 (YARN)



YARN – Fluxo de execução do Job

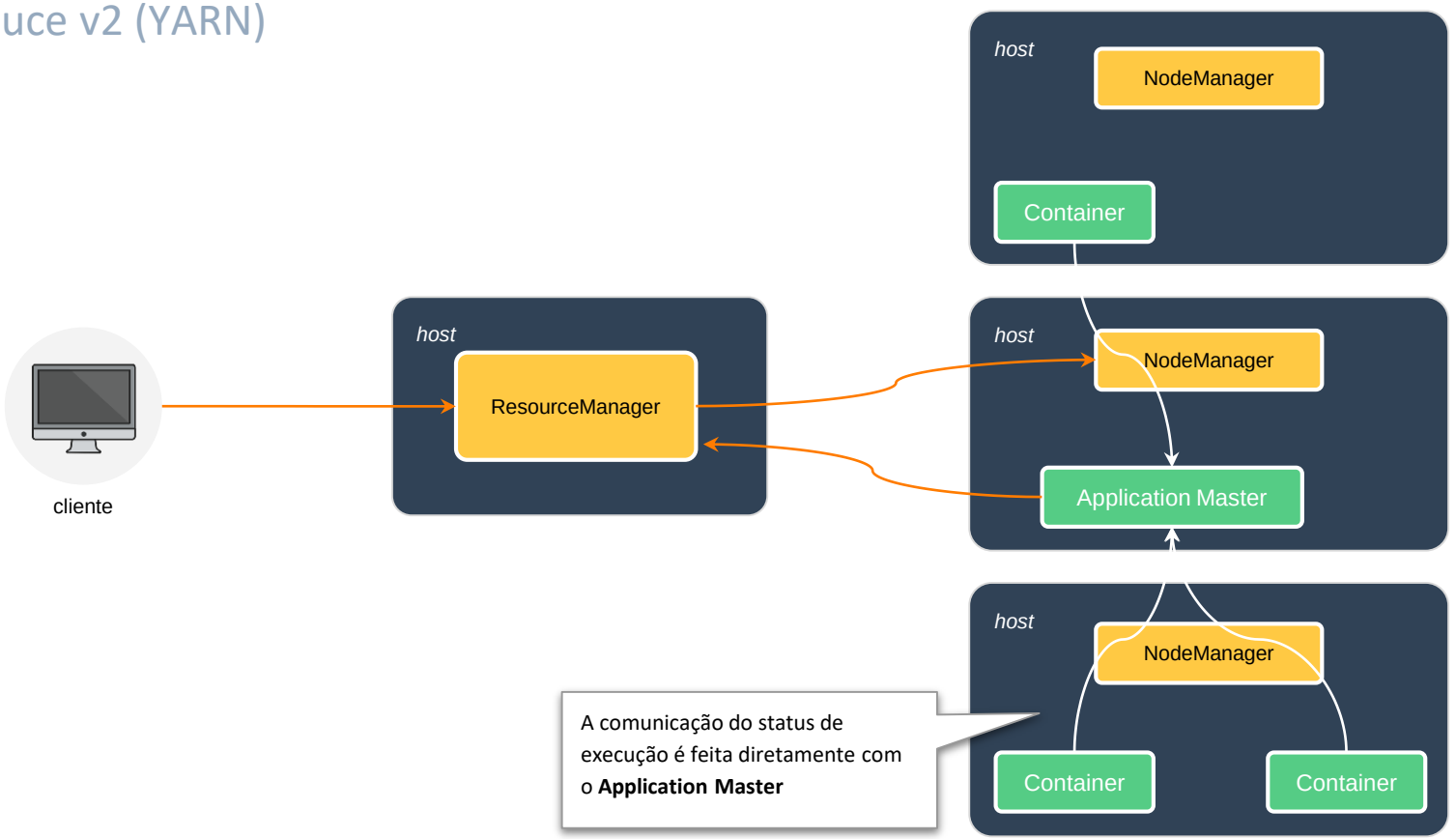
YARN
MapReduce v2 (YARN)



YARN – Fluxo de execução do Job

YARN

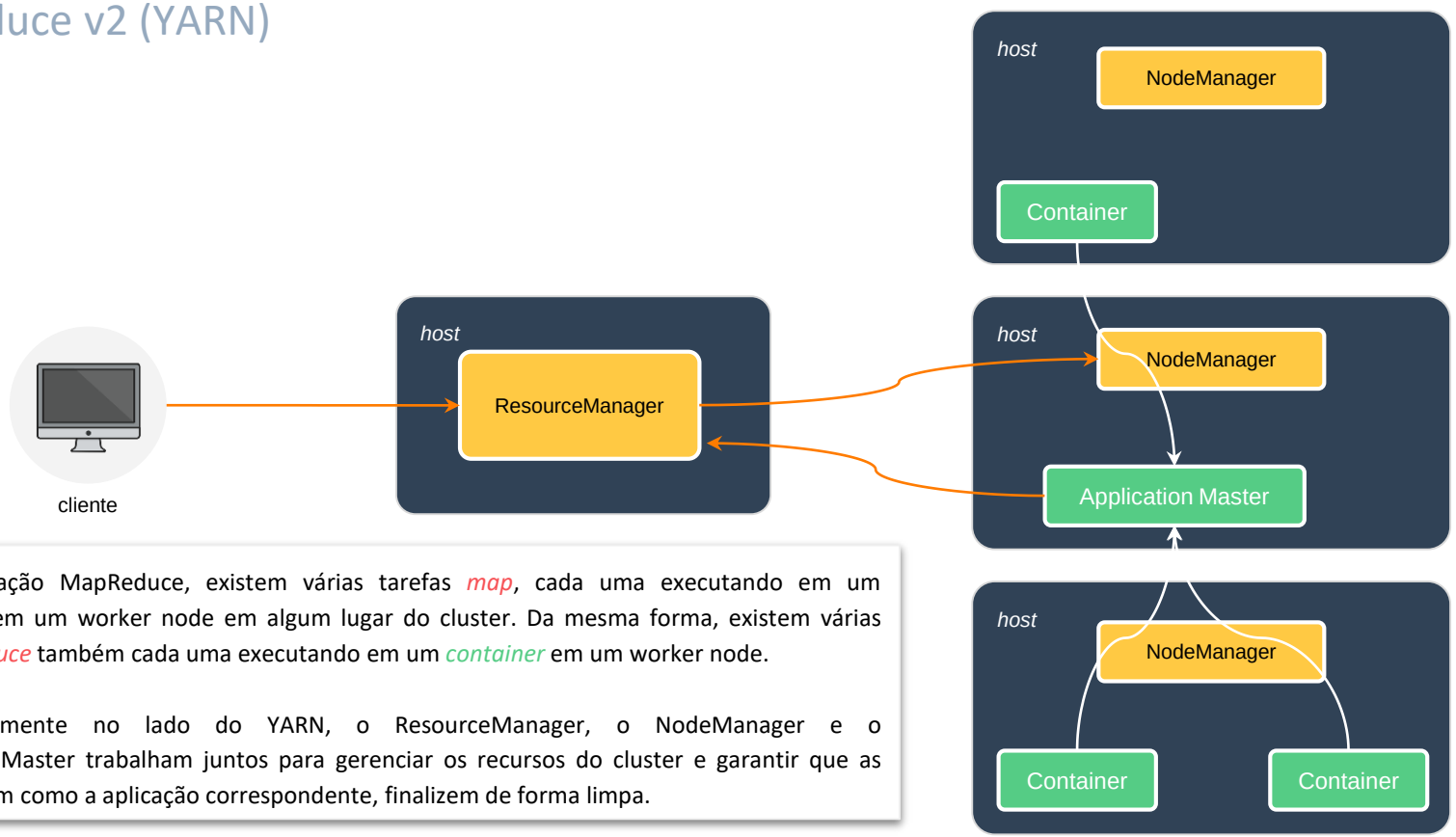
MapReduce v2 (YARN)



YARN – Fluxo de execução do Job

YARN

MapReduce v2 (YARN)



Uma aplicação MapReduce, existem várias tarefas *map*, cada uma executando em um *container* em um worker node em algum lugar do cluster. Da mesma forma, existem várias tarefas *reduce* também cada uma executando em um *container* em um worker node.

Simultaneamente no lado do YARN, o ResourceManager, o NodeManager e o ApplicationMaster trabalham juntos para gerenciar os recursos do cluster e garantir que as tarefas, bem como a aplicação correspondente, finalizem de forma limpa.

Referências

APACHE. HDFS Architecture Guide. 2013. Disponível em: http://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf

BARROSO, Luiz André; DEAN, Jeffrey; HOLZLE, Urs. Web search for a planet: The Google cluster architecture. **Micro, IEEE**, v. 23, n. 2, p. 22-28, 2003. Disponível em: http://www.eecs.harvard.edu/~dbrooks/cs246-fall2004/google.pdf

GHEMAWAT, Sanjay; GOBIOFF, Howard; LEUNG, Shun-Tak. The Google file system. In: ACM SIGOPS Operating Systems Review. ACM, 2003. p. 29-43.

GHEMAWAT, Sanjay; DEAN, Jeffrey. MapReduce: simplified data processing on large clusters. In: **Proc. OSDI**. 2004.

GUO, Zhenhua; FOX, Geoffrey; ZHOU, Mo. Investigation of data locality in MapReduce. In: Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012). IEEE Computer Society, 2012. p. 419-426. Disponível em: http://cgl.soic.indiana.edu/publications/InvestigationDataLocalityInMapReduce_CCGrid12_Submitted.pdf

IMPETUS. Hadoop Performance Tuning. 2009. Disponível em: http://strategisminc.com/wp-content/uploads/2014/07/White-paper-HadoopPerformanceTuning.pdf

Krzyzanowski, P. Distributed File Systems. Rutgers University, 2012. Disponível em: https://www.cs.rutgers.edu/~pxk/417/notes/16-dfs.html

LIN, Jimmy; DYER, Chris. Data-intensive text processing with MapReduce.**Synthesis Lectures on Human Language Technologies**, v. 3, n. 1, p. 1-177, 2010. Disponível em: http://beowulf.csail.mit.edu/18.337-2011/MapReduce-book-final.pdf

MapR. The Future of Hadoop is Right Now, 2014. Disponível em: https://www.mapr.com/wwgd

SHVACHKO, Konstantin V. HDFS Scalability: The limits to growth. v. 35, n. 2, p. 6-16, 2010. Disponível em: https://www.usenix.org/legacy/publications/login/2010-04/openpdfs/shvachko.pdf

TANDON, Prateek; CAFARELLA, Michael J.; WENISCH, Thomas F. Minimizing Remote Accesses in MapReduce Clusters. In: **IPDPS Workshops**. 2013. p. 1928-1936. Disponível em: http://web.eecs.umich.edu/~michjc/papers/tandon_hpdic_minimizeRemoteAccess.pdf

WHITE, Tom. Hadoop: The definitive guide. " O'Reilly Media, Inc.", 2012.

Referências

- AMEX, American Exmpress Our History, 2015. https://secure.cmax.americanexpress.com/Internet/GlobalCareers/Staffing/Shared/Files/our_story_3.pdf
- Apache Hadoop. 2014. http://hadoop.apache.org/
- Baldeschwieler, E. Hadoop at Yahoo!, 2009. Disponível em: https://www.google.com.br/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CB0QFjAA&url=http%3A%2F%2Fwww.hadooper.cn%2Fdct%2Fattach%2FY2xiOmNsYjpwZGY6ODE%3D&ei=ngEmVcTmNoPRsAXYpoSgBg&usg=AFQjCNGw_wSaqgGr32beq8POCxlne13KpA&sig2=IFYc6wfmMKvdFDdJaLP9EQ&bvm=bv.90237346,d.eXY
- Bose, Abhijit, et al. Recommendations @ American Express, 2013. http://pt.slideshare.net/SessionsEvents/ml-conf-axp2013finalversion8am
- Connolly, S. 7 Key Drivers for the Big Data Market. Hortonworks. 2012. br.hortonworks.com/blog/7-key-drivers-for-the-big-data-market/
- Patterson, D.; Hennessy, J. Organização e projeto de computadores: a interface hardware/software. Elsevier, 2005. I. Foster, C. Kesselman, S. Tuecke, “The Anatomy of the Grid: Enabling Scalable Virtual Organizations”, International Journal of Supercomputer Applications, 2001
- D. B. Skillicorn, “Motivating Computational Grids”, 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid, 2002
- ANDREWS, Gregory R. Concurrent programming: principles and practice. Benjamin/Cummings Publishing Company, 1991.
- COLVERO, Taís Appe; DANTAS, Mário; CUNHA, Daniel Pezzi da. Ambientes de Clusters e Grids Computacionais: Características, Facilidades e Desafios. 2005.
- Conway, D. The Data Science Venn Diagram, 2010. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
- Izotov, I. Finding the right people for your Hadoop initiative, 2015. https://www.linkedin.com/pulse/finding-right-people-your-hadoop-initiative-igor-izotov
- Morgan, T. P. Why Hadoop Is The New Backbone Of American Express, 2014. http://www.enterprisetech.com/2014/10/17/hadoop-new-backbone-american-express
- Wikibon, Big Data Vendor Revenue and Market Forecast 2013-2017, 2014.
- http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017
- Wikipedia, History of Yahoo!, 2015. Disponível em: http://en.wikipedia.org/wiki/History_of_Yahoo!
- Zicari, R. Hadoop at Yahoo. Interview with Mithun Radhakrishna, 2014. Disponível em: http://www.odbms.org/blog/2014/09/interview-mithun-radhakrishnan/

Referências

- http://www.top500.org/statistics/overtime
- http://developer.yahoo.com/blogs/ydn/posts/2013/02/hadoop-at-yahoo-more-than-ever-before/
- http://snia.org/sites/default/education/tutorials/2012/spring/storman/AnilVasudeva_NextGen_Storage_Big%20Data.pdf
- http://www.snia.org/
- http://www.research.ibm.com/haifa/projects/systems/dsf.html
- http://mullinsconsultinginc.com/db2arch-sd-sn.html
- http://pt.scribd.com/doc/50751016/Grid-Computing

Principais Players do Mercado de Hadoop:

- http://gigaom.com/2011/03/24/meet-mapr-a-competitor-to-hadoop-leader-cloudera/
- www.computerworld.com/s/article/9217054/EMC_joins_forces_with_Hadoop_distributor_MapR_Technologies
- www.mapr.com/company/press-releases/mapr-closes-110-million-financing-led-google-capital
- http://gigaom.com/2011/02/01/why-yahoo-is-discontinuing-its-hadoop-distribution/
- http://bits.blogs.nytimes.com/2012/02/21/teradata-and-hortonworks-join-forces-for-a-big-data-boost/?_php=true&_type=blogs&_r=0
- http://techcrunch.com/2013/06/25/hortonworks-raises-50m-for-expansion-and-development-in-growing-hadoop-oriented-data-analytics-market/
- http://www.cloudera.com/content/cloudera/en/about/investors.html

Referências

Principais Players do Mercado de Hadoop:

- https://www.datacenterdynamics.com/focus/archive/2014/05/hortonworks-acquires-xa-secure
- http://techcrunch.com/2009/03/16/cloudera-raises-5-million-series-a-round-for-hadoop-commercialization/
- http://venturebeat.com/2009/06/01/cloudera-raises-6m-more-for-serious-data-processing/
- http://venturebeat.com/2010/10/26/cloudera-raises-25m-to-help-deal-with-the-enterprise-data-deluge/
- http://fortune.com/2014/04/01/intels-high-priced-cloudera-investment-looks-a-bit-desperate/
- http://www.infoworld.com/article/2623494/m-a/update--teradata-buys-aster-data-to-boost--big-data--wares.html
- http://www-03.ibm.com/press/us/en/pressrelease/37833.wss
- http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017
- http://databricks.com/categories/partners/
- Hortonworks IPO - Why Now? Or better, who will benefit from the IPO. Link: http://feedproxy.google.com/~r/nosql/~3/PUKLLwqQXWY/102949496417
- http://retailroadshow.com/sys/docView.asp?d=%2Fshows%2FHDP110u_rr%2FDB6KG2B%2Fprosp%2Epdf&di=2442&si=2178
- http://www.infoworld.com/article/2849568/big-data/teradata-brings-mapr-hadoop-into-the-data-warehouse.html
- http://www.infoworld.com/article/2854359/hadoop/why-hortonworks-cant-sustain-a-billion-dollar-unicorn-valuation.html
- http://kellblog.com/2014/11/18/it-aint-easy-making-money-in-open-source-thoughts-on-the-hortonworks-s-1/
- http://blogs.gartner.com/merv-adrian/2014/11/17/hortonworks-ipo-why-now/
- https://www.linkedin.com/pulse/20131003190011-29380071-the-cloudera-model
- http://siliconangle.com/blog/2015/01/16/mapr-ceo-confirms-plans-for-late-2015-ipo/
- http://www.forbes.com/sites/danwoods/2015/01/27/microsofts-revolution-analytics-acquisition-is-the-wrong-way-to-embrace-r/
- http://blogs.microsoft.com/blog/2015/01/23/microsoft-acquire-revolution-analytics-help-customers-find-big-data-value-advanced-statistical-analysis/
- http://projects.revolutionanalytics.com/
- http://blog.revolutionanalytics.com/2015/01/revolution-acquired.html
- http://fortune.com/2015/01/14/mapr-eyes-late-2015-ipo/
- http://blog.cloudera.com/blog/2015/02/got-sql-xplain-io-joins-cloudera/
- https://gigaom.com/2015/02/03/cloudera-acquires-self-service-analytics-startup-xplain-io/
- https://gigaom.com/2015/02/06/exclusive-pivotal-ceo-says-open-source-hadoop-tech-is-coming/
- http://www.businesswire.com/news/home/20150205005118/en/MapR-Earns-Highest-Score-Gigaom-Research-HadoopData#.VPaO8vnF-So
- http://br.hortonworks.com/blog/open-source-communities-developer-kingdom/
- https://blogs.apache.org/foundation/entry/the_apache_software_foundation_welcomes3
- http://appaloud.com/teradata-offers-mapr-within-teradatas-unified-data-architecture/
- http://blogs.gartner.com/nick-heudecker/who-asked-for-odp/

Referências

Principais Players do Mercado de Hadoop:

- http://pivotal.io/big-data/press-release/technology-leaders-unite-around-open-data-platform-to-increase-enterprise-adoption-of-hadoop-and-big-data
- http://opendataplatform.org/
- http://vision.cloudera.com/beware-of-openwashing/
- http://www.dbms2.com/2015/02/18/hadoop-and-then-there-were-three/
- http://www.dbms2.com/2015/02/18/greenplum-is-being-open-sourced/
- http://br.hortonworks.com/blog/pivotal-hortonworks-announce-alliance/
- http://blog.pivotal.io/big-data-pivotal/news-2/pivotal-big-data-suite-open-agile-cloud-ready
- http://community.apache.org/apache-way/apache-project-maturity-model.html
- http://br.hortonworks.com/blog/hortonworks-model/
- http://venturebeat.com/2015/02/13/cloudera-is-getting-ready-for-an-ipo-in-the-next-few-months/
- http://www.infoworld.com/article/2688830/open-source-software/oracle-looks-to-the-future-but-remains-stuck-in-the-past.html
- http://news.investors.com/technology/022315-740424-hortonworks-q4-earnings-preview.htm
- http://www.bloomberg.com/news/articles/2015-02-13/hortonworks-teams-up-with-hitachi-to-develop-hadoop-technology
- https://gigaom.com/2015/03/03/cloudera-ceo-declares-victory-over-big-data-competition/
- http://premium.wikibon.com/boiling-down-the-open-data-platform-debate/
- https://gigaom.com/2013/10/25/cloudera-ceo-were-taking-the-high-profit-road-in-hadoop/
- http://research.gigaom.com/2015/01/hadoop-security-wars/
- http://www.cloudera.com/content/cloudera/en/about/press-center/press-releases/2014/09/30/cloudera-acquires-datapad-technology-assets-and-team-to-strength.html
- http://datapad.io/
- http://br.hortonworks.com/blog/sap-hana-hadoop-a-perfect-match/
- http://blogs.gartner.com/nick-heudecker/cloudera-teradata-partnership-highlights-hadoop-reality/
- http://hortonworks.com/blog/databricks-expanded-partnership-hortonworks/
- http://br.teradata.com/News-Releases/2014/Teradata-Acquires-RainStor/?LangType=1046&LangSelect=true
- http://thinkbig.teradata.com/
- http://br.teradata.com/News-Releases/2014/Teradata-Acquires-Think-Big-Analytics-to-Accelerate-Growth-of-its-Hadoop-and-Big-Data-Consulting-Capability/?LangType=1046&LangSelect=true
- http://rainstor.com/solutions/rainstor-for-teradata/

FIAP

MBA+