

MO809/MC964
Tópicos em Computação Distribuída

Introdução

Islene Calciolari Garcia

Instituto de Computação - Unicamp

Segundo Semestre de 2015

Sumário

Um pouco sobre mim...

Contexto

Estudo de caso: Twitter

Dropbox

Projeto Hadoop

Objetivos

Experiências anteriores

Critério de Avaliação

Referências

Um pouco sobre mim...

- ▶ Formação e filiação
 - ▶ Instituto de Computação—Unicamp
- ▶ Interesses de pesquisa
 - ▶ Sistemas distribuídos
 - ▶ Sistemas operacionais



SOFTWARE
LIVRE

Objetivos

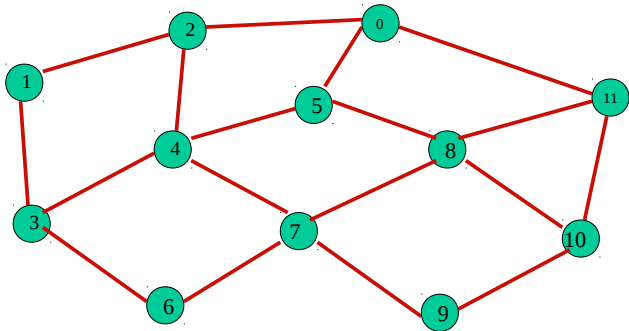
Tópicos em ... ⇒ abordagem livre

Esta disciplina cobrirá tópicos teóricos e práticos de sistemas distribuídos com ênfase em Big Data.

Na parte teórica, serão abordados algoritmos de coordenação e consenso, técnicas de tolerância a falhas e o modelo de programação MapReduce.

Na parte prática, analisaremos o projeto e implementação de sistemas distribuídos reais, que tenham código disponível sob licença livre. Em particular, estudaremos o projeto Apache Hadoop e seu modelo de desenvolvimento.

What is a distributed system?



A channel may be physical (wired, wireless) or logical

Abstract view: It is a **network** of **processes**.

(The **nodes** are processes, and the **edges** are communication channels.)

1

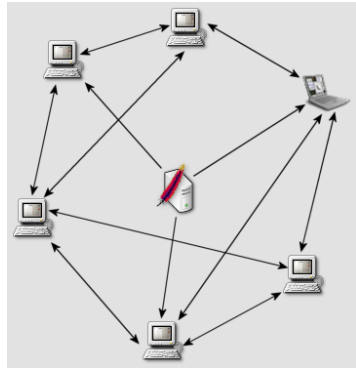
What is a Distributed System?

A collection of independent, autonomous hosts connected through a communication network.

- No shared memory (must use the network)
- No shared clock

Goal of a distributed system

The computers coordinate their activities and to share hardware and software and data, so that users perceive it as a **single, integrated computing service with a well-defined goal.**



Downloading music in Bittorrent

Examples

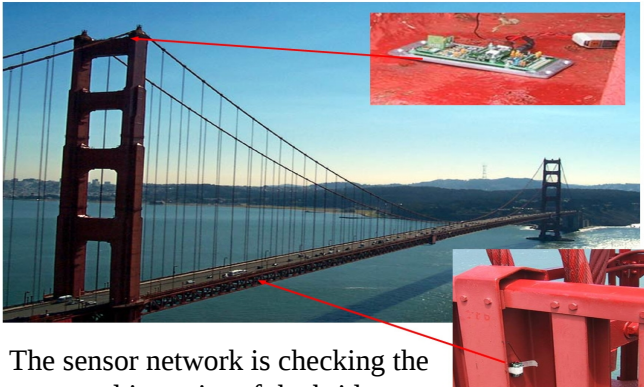
Large networks are very commonplace these days. Think of the **world wide web**. A few examples of distributed systems are:

- eBay for internet-based auction
- Sensor networks
- BitTorrent (P2P network) for downloading video / audio
- Skype for making free audio and video communication
- Facebook (the oxygen of many people)
- Process control networks in engineering factories
- Computational grids (**OSG, Teragrid, SETI@home**)
- Network of mobile robots collectively doing a job
- Distance education, net-meeting etc.
- Netbanking
- Vehicular networking



What are these?

Sensor Network

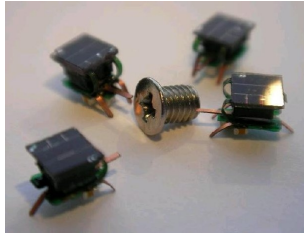


The sensor network is checking the structural integrity of the bridge

Mobile robots

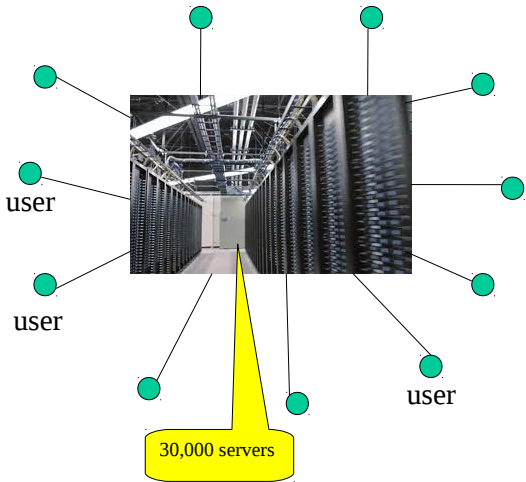


I-Swarm Robot
(See a video of the I-Swarm
Robots on YouTube)



The I-Swarm project, consisting of 10 research institutes, is coordinated by Professor Heinz Wörn and Jörg Seyfried of the University of Karlsruhe in Germany.

The Case of Facebook



The new Facebook data center in Prineville, Oregon. The new servers have been redesigned are networked, for energy efficiency, speed-up and for fault-tolerance.

The set up mimics client-server kind of operation, with the servers having a **high level of parallelism**. However, the network of servers also form a **distributed system**.

Objetivos

Tópicos em ... ⇒ abordagem livre

Esta disciplina cobrirá tópicos teóricos e práticos de sistemas distribuídos com ênfase em Big Data.

Na parte teórica, serão abordados algoritmos de coordenação e consenso, técnicas de tolerância a falhas e o modelo de programação MapReduce.

Na parte prática, analisaremos o projeto e implementação de sistemas distribuídos reais, que tenham código disponível sob licença livre. Em particular, estudaremos o projeto Apache Hadoop e seu modelo de desenvolvimento.

Big Data

- ▶ *Volume, Velocity, Variety, Veracity*
- ▶ Escalabilidade
- ▶ Disponibilidade
- ▶ Tolerância a falhas
- ▶ Segurança
- ▶ Como as grandes empresas enfrentam estes desafios?
- ▶ Da prática para a teoria: muitas lições a serem aprendidas

Estudo de caso: Twitter

Twitter!

Architecture and Scalability

Aditya B
05IT04

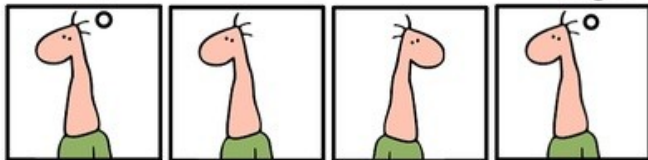
WHAT IS TWITTER ?

- micro-blogging platform
- text-based posts
- 140 characters in length
- followers receive updates

Its addictive

WHAT A BUSY
WORLD!
IN THE NEXT 5
MINUTES I WILL
DO NOTHING

MAYBE I
SHOULD
TWITTER
ABOUT IT



geek and poke

SOUNDS OF SILENCE

**arbitya**

giving a seminar on Twitter tomorrow!
posting for a screenshot :P

2 minutes ago from TweetDeck

10 easy steps to advanced Photography Skills <http://bit.ly/j2oKl>.
check out beautiful photo of the Taj Mahal <http://bit.ly/sEKNa>

about 11 hours ago from TweetDeck

@pachax first success of nitforum! suntex :)

about 11 hours ago from TweetDeck in reply to pachax

<http://www.lyricsplugin.com/> Lyrics Plugin. good one, check it out

about 22 hours ago from TweetDeck

@pachax true dude.. usability first. open-source next :)

about 23 hours ago from TweetDeck in reply to pachax

Hazards of being a bl**dy do-gooder <http://bit.ly/lBGsk> for trying
to do their bit to create a mindset change towards cycling

9:21 PM yesterday from TweetDeck

using safari web-browser on windows. looks good.

12:57 AM yesterday from TweetDeck

RT @vinuthomas: "Facebook terms of service compared with
MySpace, Flickr, Picasa, YouTube, LinkedIn, and Twitter"

<http://tinyurl.com/d23mme>

2:13 PM Feb 17th from TweetDeck

<http://tinyurl.com/am5j5f> and <http://tinyurl.com/boew2p>

summarizes final year @ engineering college

6:54 PM Feb 16th from web

Name [adityabheemarao](#)

Location [Bangalore](#)

Web <http://bheemboy.w...>

Bio [haha. lol](#)

42

following

45

followers

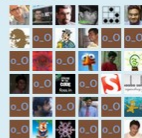
218

updates

Updates

Favorites

Following



RSS feed of arbitya's updates

Who uses twitter

**cnnbrk**

Name CNN Breaking News
Location Everywhere
Web <http://cnn.com/>
Bio CNN.com is among the world's leaders in online news and information delivery.

1 following 236,515 followers 628 updates

Updates

Favorites

Following

RSS feed of cnnbrk's updates

Kyrgyzstan parliament votes to close U.S. military base used as a route for troops and supplies heading to Afghanistan.
about 23 hours ago from CNN Alert

A helicopter carrying 18 people goes down in the North Sea off the coast of Scotland, Royal Air Force officer tells CNN.
12:13 PM Feb 18th from CNN Alert

President Obama signs a \$787 billion stimulus bill into law aimed at stemming and reversing U.S. recession.
12:32 PM Feb 17th from CNN Alert

President Obama signs a \$787 billion stimulus bill he calls "the most sweeping economic recovery package in our history."
12:29 PM Feb 17th from CNN Alert

**BarackObama**

Name Barack Obama
Location Chicago, IL
Web <http://www.barack.com>

288,610 following 299,938 followers 264 updates

Updates

Favorites

Following

View All

is asking you to honor Dr. Martin Luther King, Jr by volunteering in your area. Visit <http://USAservice.org> or text SERVE to 56333 for info.
3:04 PM Jan 19th from web

To participate in the Inauguration visit <http://pic2009.org> or text HISTORY to 56333. Follow the Inauguration on Twitter @obamainaugural
7:52 PM Jan 19th from web

We just made history. All of this happened because you gave your time, talent and passion. All of this happened because of you. Thanks
11:34 AM 2009 gth, 2008 from web

**PiMPY3WASH**

Name PiMPY3WASH

0 following 509 followers 73 updates

Updates

Favorites


A load of laundry finished washing at: Tue Feb 17 19:21:11 2009
2 days ago from web

A load of laundry finished washing at: Sat Feb 14 15:48:12 2009
5 days ago from web

A load of laundry finished washing at: Sat Feb 14 13:44:53 2009
5 days ago from web

A load of laundry finished washing at: Sat Feb 14 13:06:35 2009
5 days ago from web

A load of laundry finished washing at: Sat Feb 14 12:01:00 2009
5 days ago from web

**DellOutlet**

Name Dell Outlet
Location For USA customers
Web <http://DellOutlet.com>
Bio Refurbished Dell™ computers & electronics. Question/comment? Contact @StefanMacDell. Find more Dell Twitter accounts at www.Dell.com/Twitter

21 following 62,214 followers 228 updates

Updates

Favorites

Following

@jeremymeyers @FreshJulius No Studio coupons at the moment, but I may have something I can send to you tomorrow.
about 10 hours ago from HoodStable in reply to jeremymeyers

@ro20 Coupons r sometimes posted on Twitter, but u can get more by signing up for email updates. <http://ow.lym/v>
about 12 hours ago from HoodStable in reply to Ro20

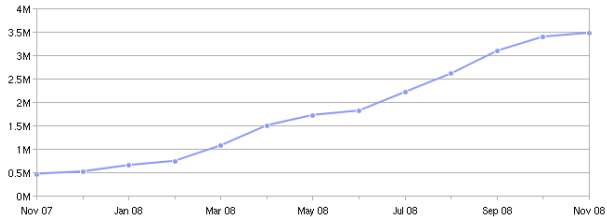
@ro20 Yes! For Dell Outlet home and home office PCs, you can use your Dell Preferred Account (DPA) or sign up here: <http://ow.lym/v>
about 11 hours ago from HoodStable in reply to Ro20

@sunflower For home or business use?
about 11 hours ago from HoodStable in reply to sunflower

Web Traffic

Unique Visitors

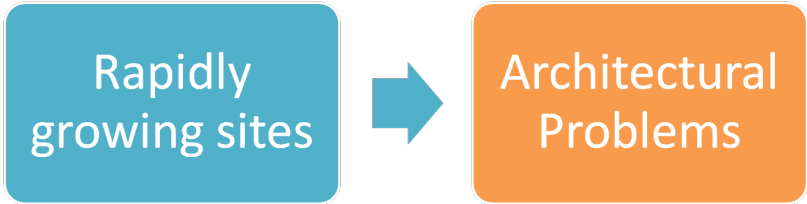
— twitter.com

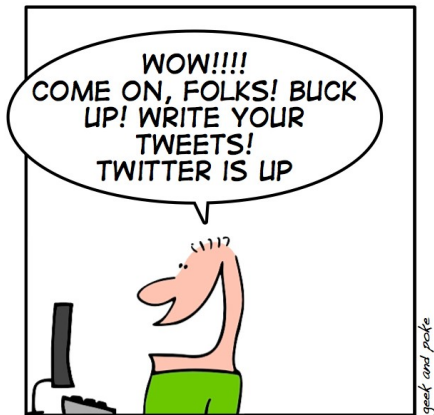


Twitter's web-based traffic

- Plus Twitter's API Traffic which is 10x the Site's

As it often happens..





A TWITTERER'S LUCKY MOMENTS

Abuse Prevention

- Bots crawl the site and add everyone as friends.
- 9000 friends in 24 hours. It would take down the site.

Saraha

9000	14	2
Following	Followers	Updates

- **Be ruthless. Delete them as users.**



Scalability -- Doing It Right

- Asynchronous event-driven design
- Partitioning/Shards
- Parallel execution
- Replication (read-mostly)

Are we all doomed to go through this painful process when we are successful?

- Time-To-Market

Vs

- Architecture

Good, Fast, Cheap - pick two

:P

LESSONS LEARNED

1. Talk to the community.
2. Treat your scaling plan like a business plan
3. Build it yourself
4. Build in user limits
5. Don't make the database the central bottleneck of doom
6. Make your application easily partitionable from the start

7. Optimize the database
8. Cache the hell out of everything
9. Most performance comes not from the language, but from application design
10. Turn your website into an open service by creating an API.

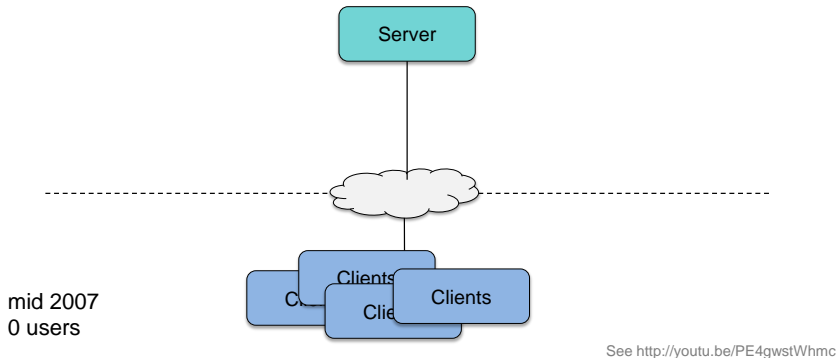
Their API is the single most powerful reason for Twitter's success.

References

- <http://twitter.com/>
- <http://highscalability.com/scaling-twitter-making-twitter-10000-percent-faster>
- <http://www.slideshare.net/Blaine/scaling-twitter>
- <http://dev.twitter.com/2008/05/twittering-about-architecture.html>
- <http://www.danga.com/memcached/>
- <http://geekandpoke.com/>

Dropbox: architecture evolution: version 1

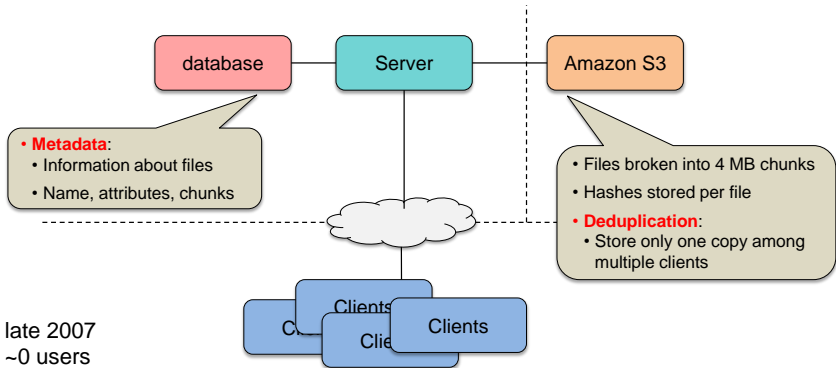
- One server: web server, app server, mySQL database, sync server



See <http://youtu.be/PE4gwstWhmc>

Dropbox: architecture evolution: version 2

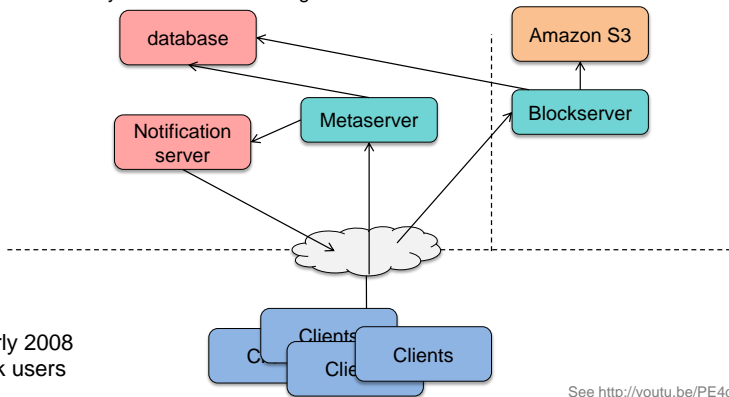
- Server ran out of disk space:
moved data to Amazon S3 service (key-value store)
- Servers became overloaded: moved mySQL DB to another machine
- Clients periodically polled server for changes



See <http://youtu.be/PE4gwstWhmc>

Dropbox: architecture evolution: version 3

- Move from polling to notifications: add **notification server**
- Split web server into two:
 - Amazon-hosted server hosts file content and accepts uploads (stored as blocks)
 - Locally-hosted server manages metadata

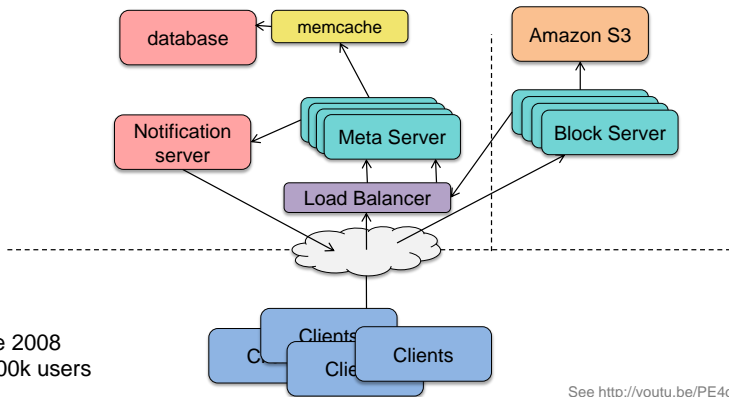


early 2008
50k users

See <http://youtu.be/PE4gwstWhmc>

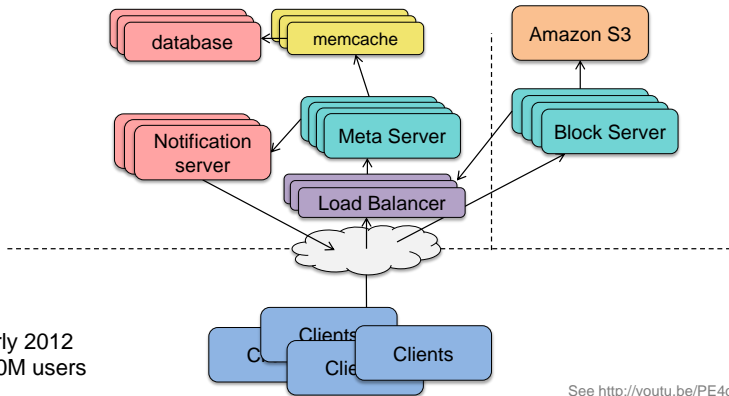
Dropbox: architecture evolution: version 4

- Add more metaservers and blockservers
- Blockservers do not access DB directly; they send RPCs to metaservers
- Add a memory cache (memcache) in front of the database to avoid scaling



Dropbox: architecture evolution: version 5

- 10s of millions of clients – Clients have to connect before getting notifications
- Add 2-level hierarchy to notification servers: ~1 million connections/server



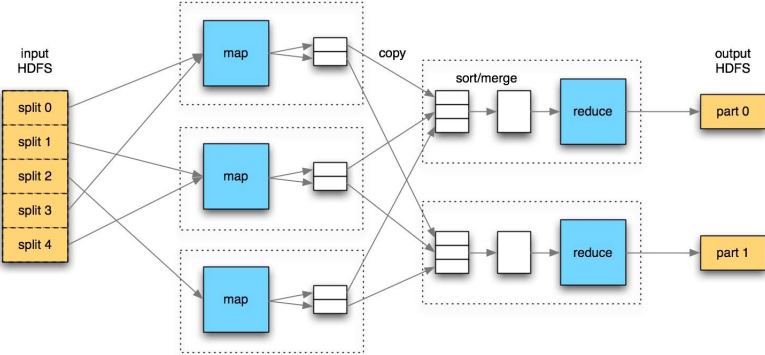
early 2012
>50M users

See <http://youtu.be/PE4gwstWhmc>

História do projeto Hadoop

- ▶ 2002-2004: Doug Cutting e Mike Cafarella trabalham no projeto Nutch.
 - ▶ Nutch deveria indexar a web e permitir buscas
 - ▶ Alternativa livre ao Google
- ▶ 2003-2004: Google publica artigo sobre o Google File System e MapReduce
- ▶ 2004: Doug Cutting adiciona o DFS e MapReduce ao projeto Nutch
- ▶ 2006: Doug Cutting começa a trabalhar no Yahoo!
- ▶ 2008: Hadoop se torna um projeto Apache
- ▶ 2013: Yarn (Hadoop 2)

MapReduce



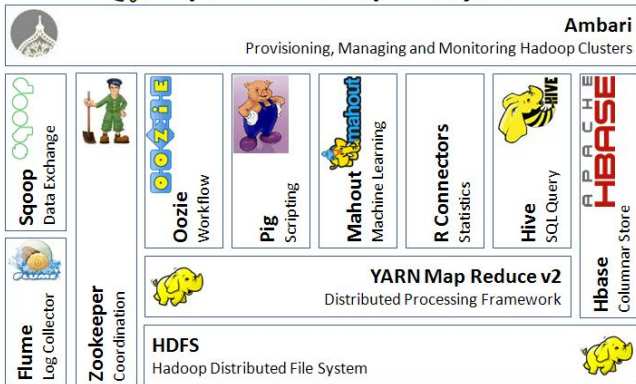
Tom White

Apache Hadoop Ecosystem

Um ecossistema em evolução



Apache Hadoop Ecosystem



Explorando o Apache Hadoop



- ▶ Dimensão de usuário:
 - ▶ MapReduce
 - ▶ Outros sistemas: HBase, ZooKeeper, Hive, Pig
- ▶ Dimensão de desenvolvedor:
 - ▶ Como contribuir
 - ▶ Jira
- ▶ Comparação com outros sistemas/abordagens

Objetivos

Tópicos em ... ⇒ abordagem livre

Esta disciplina cobrirá tópicos teóricos e práticos de sistemas distribuídos com ênfase em Big Data.

Na parte teórica, serão abordados algoritmos de coordenação e consenso, técnicas de tolerância a falhas e o modelo de programação MapReduce.

Na parte prática, analisaremos o projeto e implementação de sistemas distribuídos reais, que tenham código disponível sob licença livre. Em particular, estudaremos o projeto Apache Hadoop e seu modelo de desenvolvimento.

Sistema de Rastreamento de Bugs e Melhorias

- ▶ Comunidades de software livre expõem seu desenvolvimento via ferramentas específicas



Bugzilla



- ▶ Bugs
 - ▶ só que não
 - ▶ simples ou complexos
 - ▶ prioritários ou não prioritários
- ▶ Melhorias
 - ▶ Ideias para o avanço dos projetos
 - ▶ Entradas guarda-chuva
- ▶ Oportunidade para aprender e participar!

Como escolher uma issue para trabalhar?

- ▶ A quanto tempo a issue está aberta?
- ▶ Quem reportou o problema ou melhoria?
- ▶ A comunidade deu algum retorno?
- ▶ Houve discussão?
- ▶ Existe possibilidade de contribuição?
- ▶ Se a issue estiver fechada? Valeria pelo estudo do processo, mas as chances de contribuição são mínimas...

Em busca da issue perfeita (2010)

- ▶ ZooKeeper: *Because coordinating distributed systems is a Zoo*
- ▶ Possível participação no GSoC 2011
 - ▶ *Flip bits not burgers*
 - ▶ Estudantes recebem cerca de US\$ 5.000,00!!!
 - ▶ Google seleciona os projetos que selecionam os estudantes
 - ▶ Envolvimento inicial aumenta as chances de sucesso



A primeira contribuição! (2011)

- ▶ Turma pequena (quase todos tinha cursado no verão)
- ▶ Ainda as receitas do ZooKeeper



The screenshot shows the Apache Software Foundation logo and navigation menu at the top. Below that, the issue title "Bug in WriteLock recipe implementation?" is displayed next to a "ZooKeeper / ZOOKEEPER-645" link. The issue is represented by a blue rocket icon.

- ▶ Envio de patch

▼  Andre Esteve added a comment - 19/Apr/11 17:16

compareTo.patch aims to correct ordering of ZNodeName objects used to validate lock ownership.

The code at WriteLock gets a list of znodes and for each znode creates a ZNodeName object which is added to a sorted list.

The sorting was based on the full znode name, i.e. x-sessionID-ephemeral_number. As earlier connected clients appear to have lower sessionID values than those which connected later, who connects first gets the lock disregarding anyone who has already the lock.

This patch simply changes compareTo overload at ZNodeName to just consider the sequence number instead of the full znode name, as this class' objects are used only for this purpose, this seems to have done the trick =)


However, getSessionID not being thread-safe is still an issue.

Could someone try it out and post the results?


[A discussion about this bug and some other issues on lock recipe, as well as this patch contributors, can be found here (in Portuguese) <http://www.lsd.ic.unicamp.br/mc715-1s2011/index.php/Grupo01>]

E o retorno da primeira contribuição?

▶ Um mês depois...

▼  Patrick Hunt added a comment - 23/May/11 20:04

Hi, has anyone tried this?

▼  Matt Abrams added a comment - 30/May/11 23:09

I've tried it. I did see the starvation behavior with the original compareTo method from ZNodeName. When I applied the compareTo patch the starvation issue went away.

▶ Mais de um ano depois

- ▶ Patch seria combinado com outro e incluído
- ▶ Grupo comemorou :-)

▶ O bug continua em aberto...

Kazoo (2013)

- ▶ Comunidade ZooKeeper demora para dar retorno?
- ▶ Netflix apoiou o desenvolvimento do Curator!
- ▶ Kazoo é uma interface em Python para facilitar o uso do ZooKeeper
- ▶ Comunidade mais receptiva!
- ▶ Aluno de mestrado contribuiu e seu patch foi aceito!



E se formos completamente ignorados?



The Apache Software Foundation

<http://www.apache.org/>

Dashboards ▾

Projects ▾

Issues ▾

Agile ▾



Hadoop YARN / YARN-2299

inconsistency at identifying node

- ▾  Bruno Alexandre Rosa added a comment - 19/Nov/14 17:33

What are the affected versions?

- ▾  Bruno Alexandre Rosa added a comment - 19/Nov/14 17:40

Which*

- ▾  Bruno Alexandre Rosa added a comment - 25/Nov/14 22:24

I tried to reproduce the first case on version 2.5.2 and the bug it is still present. However, instead of host:port1 showing on Lost Nodes, I got host:port2. In the same fashion, I lost track of host:port1. The sum of Lost Nodes remains inconsistent.

- ▾  Jian He added a comment - 02/May/15 00:26

But when we manager inactive nodes(`RMContextImpl.inactiveNodes`), we use only use host. Two nodes with same host but different port are thought to identical.

This has been fixed in 2.8 to use host:port to track inactive nodes. close this.

A melhor experiência

► Bug simples



The Apache Software Foundation
<http://www.apache.org/>

Dashboards ▾

Projects ▾

Issues ▾

Agile ▾



Hadoop HDFS / HDFS-6662

WebHDFS cannot open a file if its path contains "%"

► Envio de patch e retorno rápido da comunidade

▼ Gerson Carlos added a comment - 14/Dec/14 13:54

I was able to reproduce the bug on the latest 3.0.0 code from the git repo.

It looks like the uri wasn't being properly encoded before send the request, i.e., % should be converted to %25. So, I added to explorer.js an encoding command.

But after that, the datanode broke when answering, because it wasn't decoding the uris. To fix that, I added to ParameterParser.java the decoding command from QueryStringDecoder.

See the attached hdfs-6662.patch for the diff code.

▼ Haohui Mai added a comment - 15/Dec/14 06:20

```
+ abs_path = encodeURI(abs_path);  
var url = '/webhdfs/v1' + abs_path + '?op=GET_BLOCK_LOCATIONS';
```

I think it should be `encodeURIComponent()` instead of `encodeURI()`.

Can you add a unit test to ensure that the DN decodes the path correctly?

A melhor experiência

► Uma surpresa

Issue Links

breaks

 [HDFS-7816](#) Unable to open webhdfs paths with "+"



CLOSED

► Mais orientações da comunidade e

► final feliz! :-)





Hadoop HDFS / [HDFS-6662](#)

WebHDFS cannot open a file if its path contains "%"

Agile Board

Details

Type:	 Bug	Status:	CLOSED
Priority:	 Critical	Resolution:	Fixed
Affects Version/s:	2.4.1	Fix Version/s:	2.7.0
Component/s:	namenode		
Labels:	None		

People

Assignee:



Gerson Carlos

Reporter:



Brahma Reddy Battula

Abordagem

- ▶ Estudo de tópicos em sistemas distribuídos (big data)
 - ▶ Sistemas de arquivos distribuídos
 - ▶ Modelo de programação MapReduce
 - ▶ Algoritmos para Coordenação
 - ▶ ...
- ▶ Análise de *issues* do projeto Apache Hadoop
- ▶ Experimentos práticos
- ▶ Artigos científicos

Critério de Avaliação

- ▶ Projetos relacionados aos tópicos (6.0)
 - ▶ experimentos práticos ou
 - ▶ análise de *issues* ou
 - ▶ leitura/apresentação de artigos científicos
- ▶ Seminário (3.0)
- ▶ Participação em seminários/apresentações (1.0)
- ▶ Tabela de conversão de notas:
10...8.5 = A, 8.4...7.0 = B, 6.9...5.0 = C, 5.0...0 = D

Principais Referências

- ▶ [Distributed Systems: Principles and Paradigms](#), Andrew S. Tanenbaum and Maarten Van Steen, Second Edition, Pearson, 2007.
- ▶ [Distributed Systems: Concepts and Design](#), George Coulouris, Jean Dollimore, Tim Kindberg and Gordon Blair, Fifth Edition, Addison Wesley, 2011.
- ▶ [Hadoop: The Definitive Guide](#), Tom White, Fourth Edition, O'Reilly, 2015.
- ▶ [The Hadoop Project](#)