



Hadoop ～Yahoo! JAPANの活用について～

2011/01/15

ヤフー株式会社 R&D統括本部

角田直行、吉田一星



自己紹介

角田 直行(かくた なおゆき)

R&D統括本部 プラットフォーム開発本部検索開発部 開発3

-2005年 ヤフー株式会社入社

-ヤフー地図

-ヤフー路線

-ヤフー検索

...

-2010年現在、検索プラットフォームを開発中



自己紹介

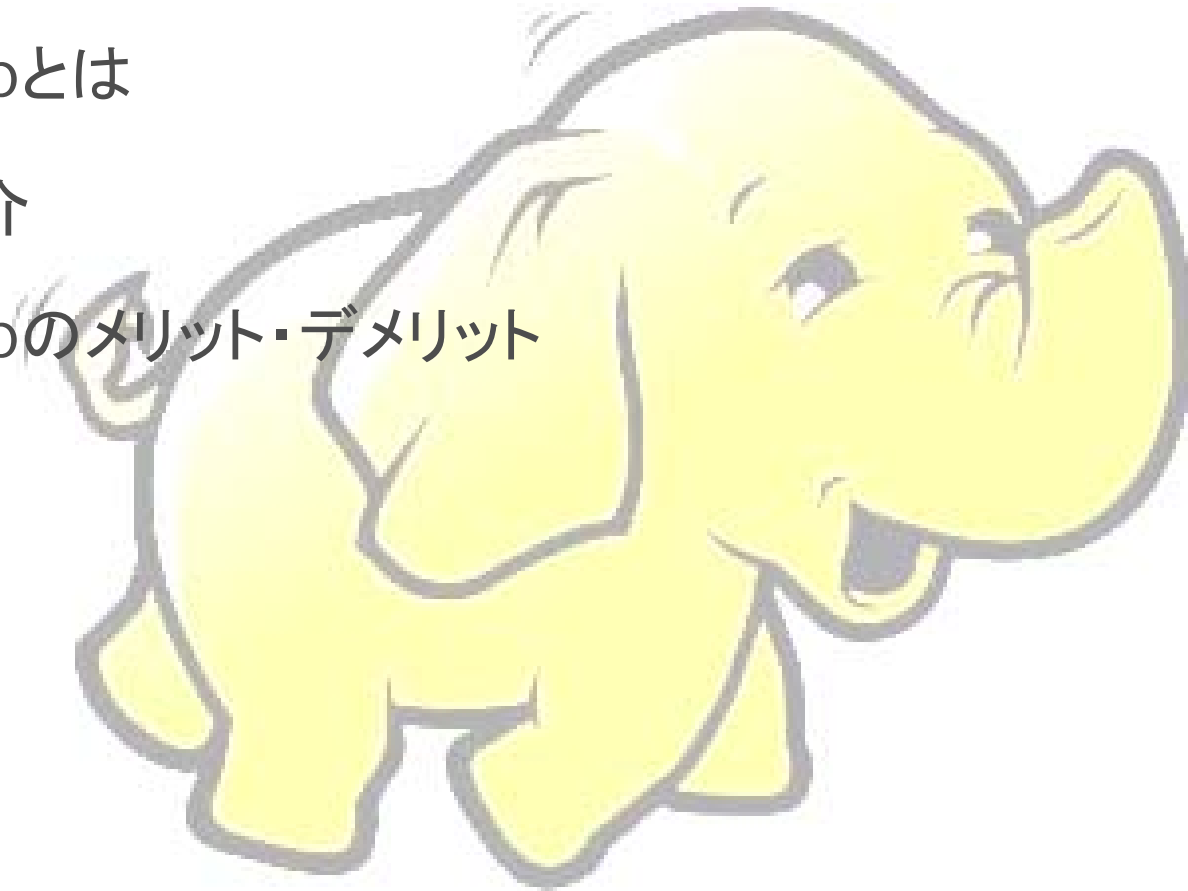
吉田一星（よしだ いっせい）

R&D統括本部プラットフォーム開発本部検索開発部開発3

- 2008年にYahoo! JAPANに入社
- 検索プラットフォームで、Hadoopに関する開発
- 画像処理、iPhone向け技術開発にもかかわる

Y! Agenda

- Introduction
- Hadoopとは
- 事例紹介
- Hadoopのメリット・デメリット
- まとめ





Introduction

Y! 有名なネットサービス

YAHOO!
JAPAN

月間 **496億7100万** PV

twitter™

1日 **5000万** のつぶやき

楽[®]天

商品数 **6800万**

facebook

月間ユーザ数 **5億** 人

各サービスとも日々成長を続けています



莫大なデータ量との闘い

- 成長を続けていくにはアクセスログ解析やデータマイニングなどが必須
- 億単位の行 or テラバイト級のデータを短時間で処理したい



毎日処理しなければならない



Yahoo! JAPANが扱うデータ

- ログは1日分だけでもかなりのサイズになる
- 行数を数えるだけでも数日かかる

Yahoo! JAPAN - Yahoo!検索 | こんにちは、ゲストさん [ログイン] ヘルプ

ウェブ | 画像 | 動画 | ブログ | 辞書 | 知恵袋 | 地図 | 一覧 ▾

hadoop [条件を指定して検索](#)
[検索設定](#)

YAHOO! JAPAN

ウェブ検索結果 hadoop で検索した結果 1~10件目 / 約9,710,000件 - 0.06秒

[hadoop 読み方](#) [hadoop world](#) [hadoop 読み](#) [hadoop Wiki](#) で検索

Hadoop - Wikipedia
HadoopはGoogleのMapReduceおよびGoogle File System(GFS) .. Hadoopバージョン0.21にはチェック ... それ以前のバージョンのHadoopでは、ジョブ ...
ja.wikipedia.org/wiki/Hadoop - ブックマーク: 2人が登録 - キャッシュ

Hadoop, hBaseで構築する大規模分散データ処理システム(2/2) ...
HadoopとhBaseは、Googleの基盤ソフトウェアのオープンソースクローンです。... エンジンを開発している Powerset社のエンジニアが主導して開発しており、Hadoopと同じくJavaで記述されています。hBaseは Hadoop ...
codezine.jp/article/detail/2448?p=2 - ブックマーク: 1人が登録 - キャッシュ

Hadoop - Wikipedia, the free ... - このページを和訳
In earlier versions of Hadoop, all active work was lost when ... Hadoop can in theory be used for any sort of work that is ...
en.wikipedia.org/wiki/Hadoop - ブックマーク: 2人が登録 - キャッシュ

Linux と Hadoop による分散コンピューティング
Hadoop は膨大な量のデータを分散操作することができるソフトウェア・フレームワークです。しかし Hadoop ... ご想像のとおり、Hadoop にとって理想的なのは、Linux を実稼働プラットフォームとして Java™ ...
www.ibm.com/developerworks/jp/linux/library/l-hadoop - キャッシュ

スポンサードサーチ

Hadoopアマゾン公式サイト
全品無料配送実施中(一部を除く)お急ぎ便利用
で当日、翌日にお届け。
Amazon.co.jp

本専門の通販サイト<<bk1>>
最速24時間以内に配送! さらに1500円以上で国内送料無料。
www.bk1.jp
▶ [一覧を見る](#)
▶ [3000円から始めるヤフーの広告掲載](#)



Y! 解決策としてのHadoop

- 大規模な処理、大容量のデータを扱うには
1台のサーバでは不可能
- マルチコアによる並行処理アプローチは複雑すぎる
- 数十～数千台規模で簡単にスケールする環境が不可欠



この発表では、
Yahoo! JAPANがHadoopをどう活用しているか
について事例を交えて解説します



Hadoopとは

Y! Hadoopとは

- 大規模分散処理システム
- Google MapReduce/GFSを論文を元に実装
- 処理時間が数時間以上かかるようなバッチ処理に向いている
 - Webのように、即座に結果が返るようなリアルタイム処理には不向き
- Javaで書かれ、オープンソースとして公開



Y! Hadoopとは

- Doug Cutting氏が生みの親
 - 全文検索ライブラリLuceneなどの
他有名OSSも開発
- Yahoo! Inc. 在籍時はフルタイムで開発
- 現在はClouderaに在籍



(出典元:Wikipedia)

Y! Hadoopとは

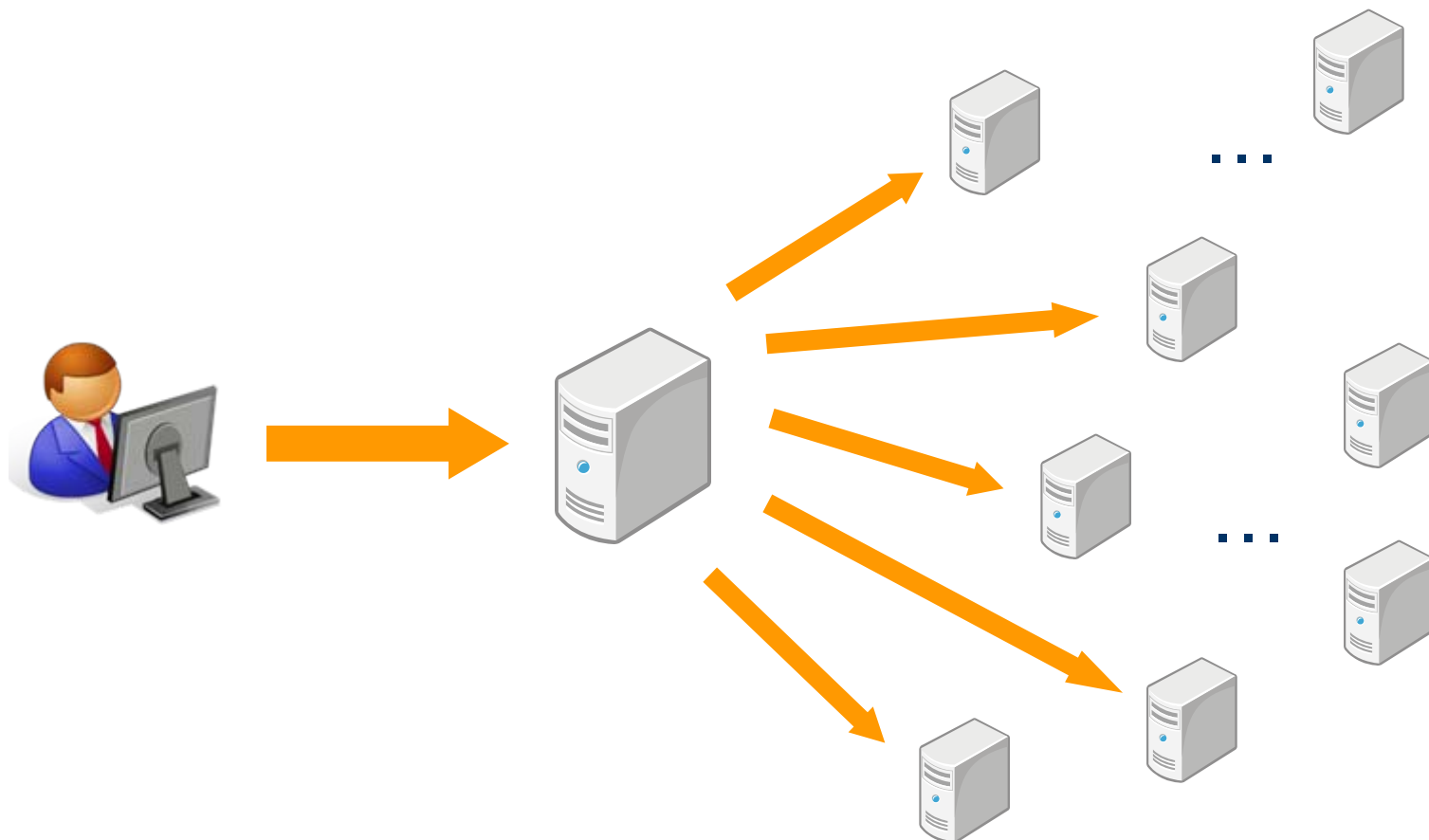
大きく**MapReduce**と**HDFS**
(分散ファイルシステム)に分かれる





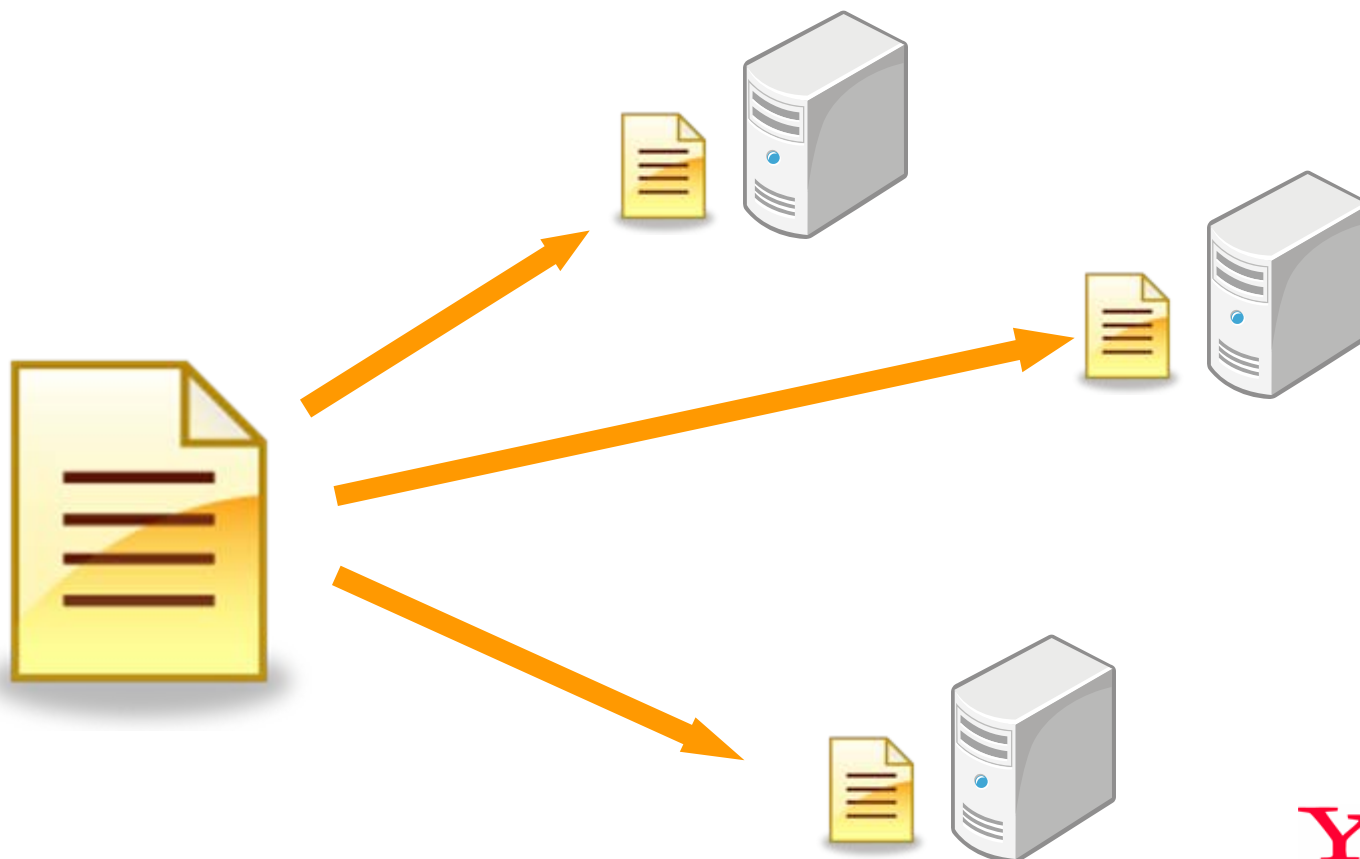
Hadoop MapReduce

長時間かかる巨大な処理を複数台のマシンに分散



Y! Hadoop HDFS

- ・巨大なファイルを複数台に分割
- ・複数サーバの各HDDを1つのHDDのように扱える





Hadoop関連プロダクト



Pig

大規模データ処理用スクリプト言語

```
A = load 'passwd' using PigStorage(' ');  
B = foreach A generate $0 as id;  
dump B;
```



Hive

Facebookが開発 扱いが一般データベースに似ている

```
CREATE TABLE pokes (foo INT, bar STRING);  
SELECT a.foo FROM pokes a;
```




Hadoop関連プロダクト



Oozie

複数のMapReduceジョブなどを
実行制御するワークフロー



HBase

Hadoop上に構築された列指向データベース
Google BigTableのクローン



Mahout

機械学習ライブラリ
Hadoopでスケール可



Hadoopの事例紹介



Hadoopを活用している会社



など・・・
増え続けています！





Yahoo! Inc.での事例紹介



Hadoop at Yahoo! Inc

- Hadoopユーザ、テスター、コミッターの数が最も多い
- Hadoopのクラスタ、台数が最も多い
 - 多数のクラスタがあり、合計25000台以上
 - 1クラスタにつき最大4000台



Yahoo! Inc トップページ

View Yahoo! Sites

MY FAVORITES [Add](#)

- [Yahoo! Mail](#)
- [Autos](#)
- [eBay](#)
- [Finance \(Dow Jones\)](#)
- [Flickr](#)
- [Games](#)
- [Horoscopes](#)
- [Maps](#)
- [Messenger](#)
- [Movies](#)
- [Music](#)
- [MySpace](#)
- [Personals](#)
- [Sports](#)
- [Weather \(65°F\)](#)

RECOMMENDED

- [Deal Of The Day](#)
- [Buzz](#)
- [Shine](#)

[Edit](#) [Add](#)



Man, woman with same name to wed

Kelly Hildebrandt is in love with Kelly Hildebrandt, and soon they'll be married.

- Santa's helpers marry
- Unusual proposals
- Yahoo! Buzz

[How they met](#) NBC Miami



Unique name brings love



Eclipse spooks superstitious



Fighter's cool pool jump



Crowe pulls a 'Robin Hood'

[Prev](#) ● ● ● ● ● [Next](#)

NEWS **WORLD** **LOCAL** **FINANCE**

- Obama pushes back against critics of health care overhaul
- Economic indicators up more than expected in June
- July becomes deadliest month for U.S. in Afghanistan
- Clinton: Officials believe 9/11 ringleaders are hiding in...
- Spacewalk unfolds on 40th moon landing anniversary
- Company designs suit to simulate elderly driving
- Sony bids \$50 million for rights to Jackson rehearsal film
- Defense Secretary Gates announces Army being... - SJ Mercury...
- Frank McCourt - author of 'Angela's Ashes' - S.F. Chronicle
- Annual all-star football game a family affair - Sunnyvale Sun

updated 12:04 pm PDT More: [News](#) [Popular](#) [Buzz](#)

Markets: Dow: 8,821.70 0.88% Nasdaq: 1,900.55 0.73%

[Get Quotes](#)

POPULAR SEARCHES

1. Paula Abdul	6. Katherine Heigl
2. Angela's Ashes	7. Jon Gosselin
3. David Beckham	8. Lupus
4. Moon Landing	9. Tour de France
5. Pearl Jam	10. George Clooney



PROGRESSIVE DIRECT

The Name Your Price® option. New and only from Progressive.

Enter ZIP Code: [Name Your Price](#)

[Compare and Save](#) - [Ad Feedback](#)

SPOTLIGHT [Prev](#) [Next](#)

Hot list: Must-watch videos

-  • Guinea pigs love watermelon
- How Ron Paul got 'punk'd' by 'Bruno'
- Michael Jackson flash mob
- Inside Vogue: 'The September Issue'

[Dog vs. playground slide](#)



GEICO Car Insurance

You could save over \$500 on car insurance. Get a free quote today.

Y! Yahoo! Inc トップページ

View Yahoo! Sites

MY FAVORITES + Add

- Yahoo! Mail** >
- Autos** >
- eBay** >
- Finance** (Dow Jones ↑) >
- Flickr** >
- Games** >
- Horoscopes** >
- Maps** >
- Messenger** >
- Movies** >
- Music** >
- MySpace** >
- Personals** >
- Sports** >
- Weather** (65°F) >

RECOMMENDED

- Deal Of The Day** >
- Buzz** >
- Shine** >

Edit + Add



Man, woman with same name to wed

Kelly Hildebrandt is in love with Kelly Hildebrandt, and soon they'll be married.

- Santa's helpers marry
- Unusual proposals
- Yahoo! Buzz

How they met NBC Miami


Unique name brings love


Eclipse spooks superstitious


Fighter's cool pool jump


Crowe pulls a 'Robin Hood'

<< Prev Next >>

NEWS **WORLD** **LOCAL** **FINANCE**

- Obama pushes back against critics of health care overhaul
- Economic indicators up more than expected in June
- July becomes deadliest month for U.S. in Afghanistan
- Clinton: Officials believe 9/11 ringleaders are hiding in...
- Spacewalk unfolds on 40th moon landing anniversary
- Company designs suit to simulate elderly driving
- Sony bids \$50 million for rights to Jackson rehearsal film
- Defense Secretary Gates announces Army being... - SJ Mercury...
- Frank McCourt - author of 'Angela's Ashes' - S.F. Chronicle
- Annual all-star football game a family affair - Sunnyvale Sun

updated 12:04 pm PDT More: [News](#) [Popular](#) [Buzz](#)

Markets: **Dow:** 8,821.70 **0.88%** **Nasdaq:** 1,900.55 **0.73%**

Get Quotes

POPULAR SEARCHES

1. Paula Abdul	6. Katherine Heigl
2. Angela's Ashes	7. Judy Genselin
3. David Beckham	8. Lupus
4. Moon Landing	9. Tour de France
5. Pearl Jam	10. George Clooney



Compare and Save - Ad Feedback

SPOTLIGHT [<<Prev](#) [Next >>](#)

Hot list: Must-watch videos

-  • Guinea pigs love watermelon
- How Ron Paul got 'punk'd' by 'Bruno'
- Michael Jackson flash mob
- Inside Vogue: 'The September Issue'

Dog vs. playground slide >>



Y! Yahoo! Inc トップページ

MY FAVORITES + Add

- Yahoo! Mail >
- Autos >
- eBay >
- Finance (Dow Jones ↑) >
- Flickr >
- Games >
- Horoscopes >
- Maps >
- Messenger >
- Movies >
- Music >
- MySpace >
- Personals >
- Sports >
- Weather (65°F) >

RECOMMENDED

- Deal Of The Day >
- Buzz >
- Shine >

Edit + Add



Man, woman with same name to wed

Kelly Hildebrandt is in love with Kelly Hildebrandt, and soon they'll be married. Santa's helpers marry
Unusual proposals

>> **How they met** NBC Miami School Buzz



Unique name brings love | Eclipse spooks superstitious | Fighter's cool pool jump | Crowe pulls a Robin Hood'

<< Prev Next >>

NEWS WORLD LOCAL FINANCE

- Obama pushes back against critics of health care overhaul
- Economic indicators up more than expected in June
- July becomes deadliest month for U.S. in Afghanistan
- Clinton: Officials believe 9/11 ringleaders are hiding in...
- Spacewalk unfolds on 40th moon landing anniversary
- Company designs suit to simulate elderly driving
- Sony bids \$50 million for rights to remake 'The Godfather' film
- Defense Secretary Gates announces Army being... - SA Monthly...
- Frank McCourt, author of 'Angela's Ashes' - S.F. Chronicle
- Annual all-star football game a family affair - Suncoast Sun

updated 12:04 pm PDT More: [News](#) [Popular](#) [Buzz](#)

Markets: Dow: 8,821.70 **0.88%** Nasdaq: 1,900.55 **0.73%**

Enter stock symbol [Get Quotes](#)

POPULAR SEARCHES

1. Paula Abdul	6. Katherine Heigl
2. Angela's Ashes	7. Judy Gensel
3. David Beckham	8. Lupus
4. Moon Landing	9. Tour de France
5. Pearl Jam	10. George Clooney



PROGRESSIVE DIRECT

The Name Your Price® option. New and only Progressive.

Enter ZIP Code: [Name Your Price](#)

Compare and Save - Ad Feedback

SPOTLIGHT <<Prev Next >>

Hot list: Must-watch videos



- Guinea pigs love watermelon
- How to read 'The Book of Bruns'
- Michael Jackson flash mob
- Inside VeggieTales September Issue
- Dog vs. playground slide >>



GEICO Car Insurance

You could save over \$500 on car insurance. Get a free quote today.

Y! サーチアシスト



- 入力した検索ワードに関連のありそうな単語を自動で補完
- データベースの構築にHadoopを使用
- 3年分のデータと、20ステップのMapReduce

	Hadoop使用前	Hadoop
時間	26日	20分
言語	C++	Python
開発期間	2~3週間	2~3日



Yahoo! JAPANでの事例



検索ログプラットフォーム

- 社内の検索サービスのログ解析全般
 - Hiveを独自に拡張して使用している
- 様々なYahoo! JAPANのサービスにデータを提供

- 関連検索ワード
- キーワード入力補助
- ショートカットの表示制御
- 検索ログプラットフォームのデータが元になっている


ウェブ 画像 動画 ブログ 辞書 知恵袋 地図 一覧 ▾

江 [条件を指定して検索](#)
[検索設定](#)

江頭2:50
江森 朋哉
江頭

🔍 [江頭2:50 CM](#) [江頭2:50 名言](#) [江頭2:50 ブログ](#) [江頭2:50 画像](#) で検索

[江頭2:50](#) - Yahoo!人物名鑑



エガシラニジゴジューブン - 芸能人 お笑い芸人
1965年7月1日生まれ / かに座 / 佐賀出身 / B型
(出典: 日本タレント名鑑)

[画像検索](#) - [動画検索](#) - [番組検索](#)

Y! Yahoo! 検索ランキング

- 検索ランキング、急上昇ワードランキングなど
- 都道府県別、性年代別のランキング (Yahoo! ラボ)
- 検索ログプラットフォームが提供したデータをさらに加工している

今年もやります! クリエイティブアワード

Y!

YAHOO! JAPAN 検索ランキング こんにちは、isyoshidさん [ログアウト] Yahoo! JAPAN - ヘルプ

トップ 急上昇ワードランキング 検索総数ランキング トレンドサーフィン RSS一覧 ブログパーツ

斉藤和義 鈴木サチ ポケットドール 高速道路 無料化 荒井聡 6月7日 15時26分更新

急上昇ワードランキング 今がわかる急上昇ワード5 6月7日更新

検索数がうなぎのぼり! 急上昇ワードランキングから、「これはやるかも?」という5つのキーワードをご紹介します。

- 4 鬼龍院花子
テレビ朝日のSPドラマ「鬼龍院花子の生涯」の劇中に登場
- 10 アマダ ブラジル
ブラジル出身のタレント・アマダが「平成教育学院」に出演
- 15 ピークル
ロックバンド・BEAT CRUSADERSの略称。9月で解散を発表
- 25 池田屋事件
新撰組による尊皇攘夷派を襲撃した事件。「龍馬伝」の劇中で
- 27 ワークフォース
イギリスの競走馬。英国がービーで7馬身差をつけ圧勝

急上昇ワードランキング一覧 ▶

検索総数ランキング

総合 人名 テレビ グラフィック スポーツ 6月7日更新

- 1 → YouTube
- 2 → mixi
- 3 → Google
- 4 → Amazon
- 5 → 楽天
- 6 → ニコニコ動画
- 7 → 2ちゃんねる
- 8 → Twitter
- 9 ☆ JRA
- 10 → AKB48



レコメンデーションプラットフォーム

- レコメンデーションサービスの計算処理に利用
- Yahoo!オークションなどに導入

オススメのオークション

 <p>【白ロム】Docomo(ドコモ) SH903iTV(ホ...</p>	 <p>ヴィン☆モノグラムマ ルチカラー☆オーレ...</p>	 <p>アイ・オー・データ USB 3.0/2.0インターフ...</p>
---	--	--

オススメのオークションをもっと見る

Yahoo!オークションのおすすめ

 <p>マニアはあえて銀 塩カメラ</p>	 <p>奥深い銀塩カメラの 世界にふれてみる</p>
 <p>新製品続々。人気 家電をピックアップ</p>	 <p>コンパクトな高性能 デジカメをチェック</p>
 <p>話題の一眼レフ、デ ジカメを手に入れよ う</p>	 <p>ニコンのデジタルー 眼</p>



検索プラットフォーム (ABYSS)

- 社内の検索サービスをホスティングするプラットフォーム
- 様々なサービスに導入されている
- 検索データのストレージとして使用
- 検索インデックス生成、検索データの解析処理

Y! 地図検索

- 地図検索インデックス生成
- クリックログ集計・検索ランキング反映
- 店舗やビルの一意性処理
- 開いているお店検索

- クロール
- 定休日・営業時間抽出
- 検索インデックス生成



Y! その他の事例

- モバイル検索
- 広告プラットフォーム
- 地域APIプラットフォーム (YOLP)
- Yahoo! JAPAN 研究所
- Etc...

- データ解析、データマイニング

- ログ解析、レコメンデーション、テキストマイニングなど

- 検索関係

- 検索インデックス生成、ランキング計算など

→大量のデータを読み込んで解析をする処理、大量の計算が必要な「バッチ処理」がほとんど



Hadoopのメリット・デメリット



Hadoopのメリット・デメリット

- MapReduceを使って、バッチ処理を簡単に分散できる
- × リアルタイム処理には向かない
- HDFSもMapReduceを使ったバッチ処理に最適化されている

Y! HDFSの特性

- ストレージとして使うには特性を理解する必要がある
 - × RDBMSの代用
 - × ユーザから多くのアクセスがあるストレージ
 - △ 小さいデータを多く格納するストレージ
 - ○ アクセスログデータのストレージ
 - ○ 過去の取引履歴データのストレージ

Y! HDFSの特性

- 何GBというような大きなデータを一気に書き込んだり、読み出したりする用途に最適化
 - シーケンシャルアクセス。SSDはあまり意味ない
- データの書き換えは想定されていない
 - ランダム書き込みができない
 - ファイルロック(排他制御)がない
- 秒間何十回といった大量の読み書き処理には向かない
 - ファイルキャッシュがない
 - もちろんRDBMSのようにインデックスがない

Y! リアルタイム処理には？

-リアルタイム処理の選択肢はたくさんある



PostgreSQL



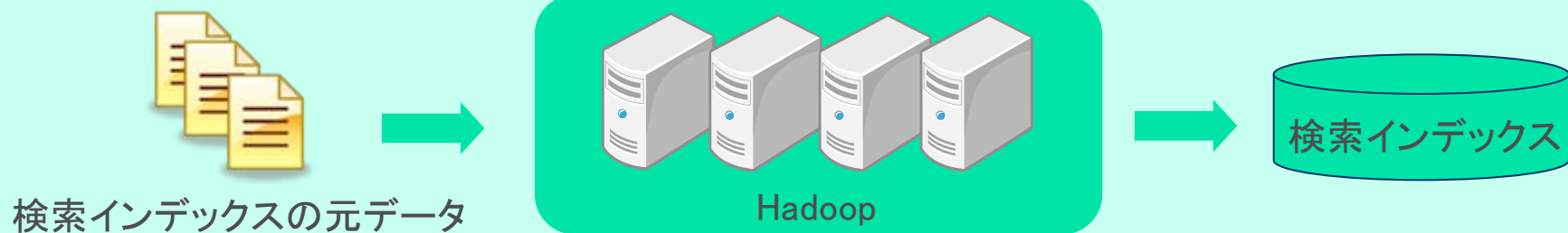
使い分けが重要！

S4 distributed stream
computing platform

Y! 使い分けの例

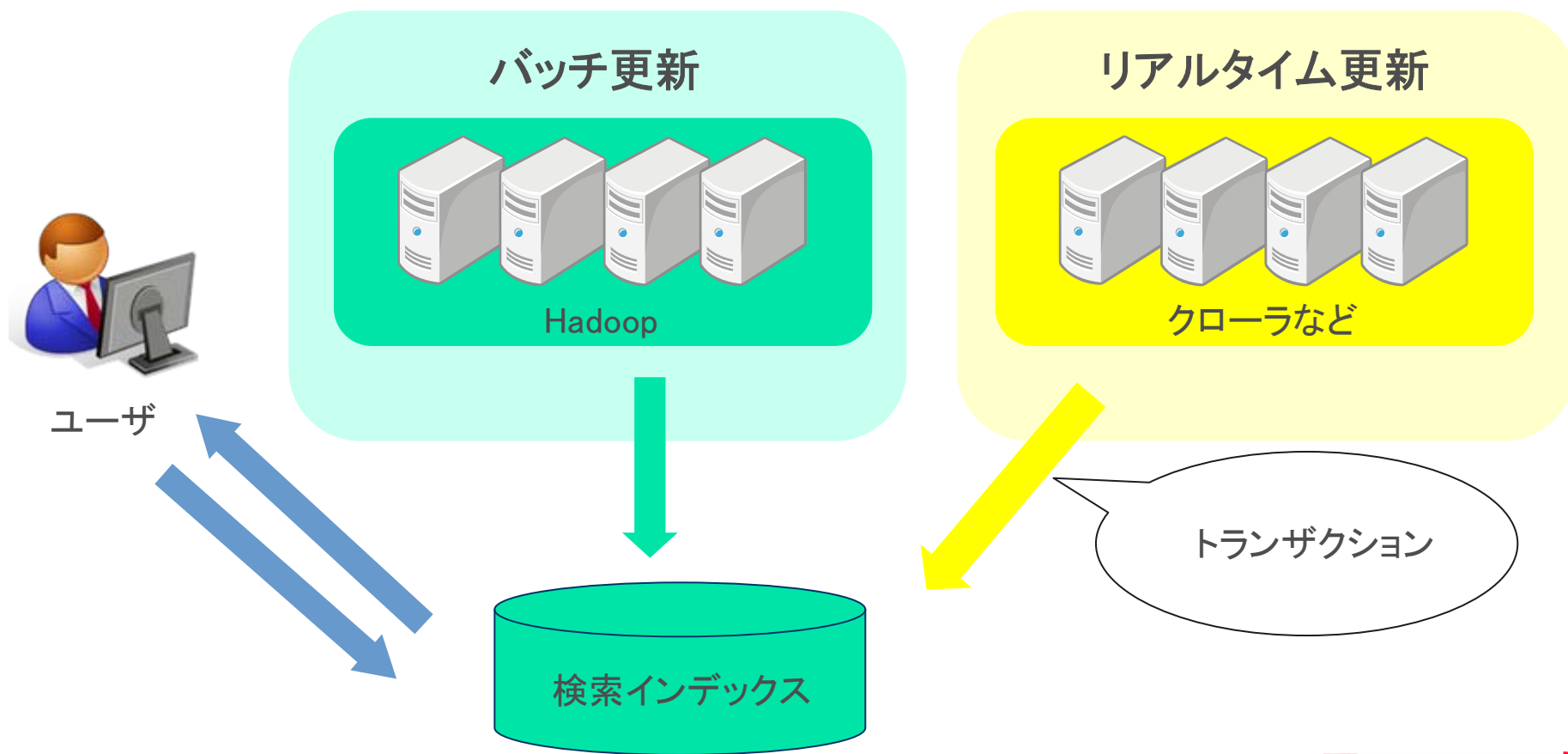
- 検索プラットフォーム、ABYSSの例

検索インデックスのバッチ更新



Y! 使い分けの例

- ユーザからのアクセス、リアルタイム更新はHadoop以外で





- Hadoopは大規模なデータを複数のマシンに分散して処理できるプラットフォーム
- Hadoopを使う企業は増え続けていて、不可欠な技術になりつつある
- Hadoopは、大規模データを扱う処理や、大量の計算が必要なバッチ処理に向いている
- Yahoo!JAPANはこれからもHadoopを活用していきます

TechBlog

ウェブテクノロジーに特化した技術系ブログ
Yahoo! JAPANの動向や最新情報を発信

< 前の記事 | トップ | 次の記事 >

携帯電話と位置情報：現在の測位 (1)

JavaScript の不思議な面白さ - 第二回

記事の検索

検索

powered by Yahoo! JAPAN

このブログのフィードを取得 [RSS](#)



2009年3月3日

Hadoopで、かんたん分散処理

こんにちは、地域サービス事業部の吉田一星です。

今回は、Hadoopについて、Yahoo! JAPANでの実際の使い

Hadoopとは、大量のデータを手軽に複数のマシンに分散
フォームです。

複数のマシンへの分散処理は、プロセス間通信や、障害
プログラマにとって敷居が高いものですが、
Hadoopはそういった面倒くさい分散処理を一手に引き受

TechBlog

ウェブテクノロジーに特化した技術系ブログ
Yahoo! JAPANの動向や最新情報を発信

< 前の記事 | トップ | 次の記事 >

Yahoo! JAPAN Internet Creative Award 2009
一般の部 グランプリ
『Blogopolis』の浜本階生さんインタビュー

Mac版Yahoo!メッセンジャー3.0のご紹介



2010年1月26日

Hadoopを使いこなす(1)

こんにちは。前回のHadoopの記事では、HadoopやMapReduceについての概要を説明しました
が、

今回は一歩踏み込んで、Hadoopの使いこなし方について書きたいと思います。

今回は、ある程度Hadoopを使ったことのある方、Hadoopのインストールをして、
オフィシャルページのMapReduceチュートリアルなどを試してみた方を対象としています。

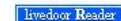
こちらでオフィシャルページの日本語訳もされていますので、試していない方は一度試してみるこ
とをおすすめします。

記事の検索

検索

powered by Yahoo! JAPAN

このブログのフィードを取得 [RSS](#)



Yahoo!サービスのWebサービスを提供
「Yahoo!デベロッパネットワーク」



Tech Blog Twitter (外部サイト)

最近のエントリー

Coolな地図サービスを作ろう！ Mashup Caravan
とCampのご案内

オープンソースカンファレンスにてHadoopセミナー
を行います

サーバーの熱は冷まさなくても良い？

<http://techblog.yahoo.co.jp/>



Hadoop Hack Night

gihyo.jp > イベント > Hadoop Hack Night

Qihyo.jp x YAHOO! JAPAN Presents "Open TechTalk"

HADOOP HACK NIGHT




gihyo.jp > イベント > Hadoop Hack Night Vol. 2

Qihyo.jp x YAHOO! JAPAN Presents "Open TechTalk"

HADOOP HACK NIGHT VOL. 2




Photo by TheLizardQueen
http://www.flickr.com/photos/lizard_queen/ / CC BY 2.0

**米国Yahoo! Hadoopチーム
アーキテクトOwen O'Malley:**

現在、米国Yahoo! Hadoopチームにてアーキテクトを務めるOwen (月)に來日いたします。それに合わせてHadoopの紹介、米国Yahoo! 例などご紹介するセミナーを開催いたします。当日は参加者の方へセッションを設け、Hadoopコミュニティに所属しているOwen氏との交流をいたします。



本イベントのTwitterハッシュタグは『#hadoophn』です。
なお、当日の様子は以下のURLでUstream中継を予定しております。

今回は 「ケーススタディで知るHadoopの可能性」

2010年に入り企業内でのクラウドテクノロジーの導入が進んでいます。とは言うものの、まだまだ黎明期でもあるため、どの技術をどのように、どのタイミングで利用すればよいか悩んでいる方も多いのではないのでしょうか。

今回のHadoop Hack Nightは、「ケーススタディで知るHadoopの可能性」と題して、導入事例から見たHadoopに迫ります。Hadoop技術を牽引するYahoo! JAPAN、さらに十分な導入実績を持つリッテルのセッションに加えて、識者による熱いパネルディスカッションを予定しています。

これからHadoop導入を検討しているSieerの方、これからHadoopについて学びたい方、必見です。
※好評のうちに終了しました、前回のHadoop Hack Nightのレポートは、こちらをご覧くださいませ。

イベント概要	
名称	Hadoop Hack Night Vol. 2
日程	8月4日(水)
時間	19:00~21:00頃 ※変更場合があります。
場所	ヤフー株式会社 [アクセス:東京ミッドタウン(六本木駅直結, 乃木坂駅徒歩10分)]
定員	100名 ※応募多数の場合はお断りさせていただく場合がございます。あらかじめご了承ください。

2010年3月、8月に開催





ご静聴ありがとうございました！