

Apache NiFi 1.0 を使ってみた

2016/11/18 - BigData JAWS 勉強会 #3

FAST RETAILING 業務改革推進
アーキテクチャ・インフラストラクチャ
遠山敏章

目次

1. 自己紹介
2. ビジネス背景
3. Apache NiFi とは
4. ユースケース
5. デモ
6. Hadoop Summit Tokyo 2016 ミニ報告
7. 参考資料

目次

1. 自己紹介
2. ビジネス背景
3. Apache NiFi とは
4. ユースケース
5. デモ
6. Hadoop Summit Tokyo 2016 ミニ報告
7. 参考資料

自己紹介

- 遠山敏章
- ソフトウェアエンジニア@ファーストリテイリング
 - 所属: 業務改革推進 アーキテクチャ インフラストラクチャ
 - 担当: Data Analytics Infrastructure
- 経歴
 - Cyber Agent (6年)
 - ネット広告のMarketing Platform
 - Smartphone Platformのデータ分析基盤
 - Fast Retailing (9ヶ月)
- データ分析のワークフロー管理ツール利用履歴
 - Cron + Hadoop MR in Java (2010)
 - Patriot workflow Scheduler + Hive (2013)
 - Luigi + ETL on AWS (2016)
 - Jenkins + bash
 - Drake

目次

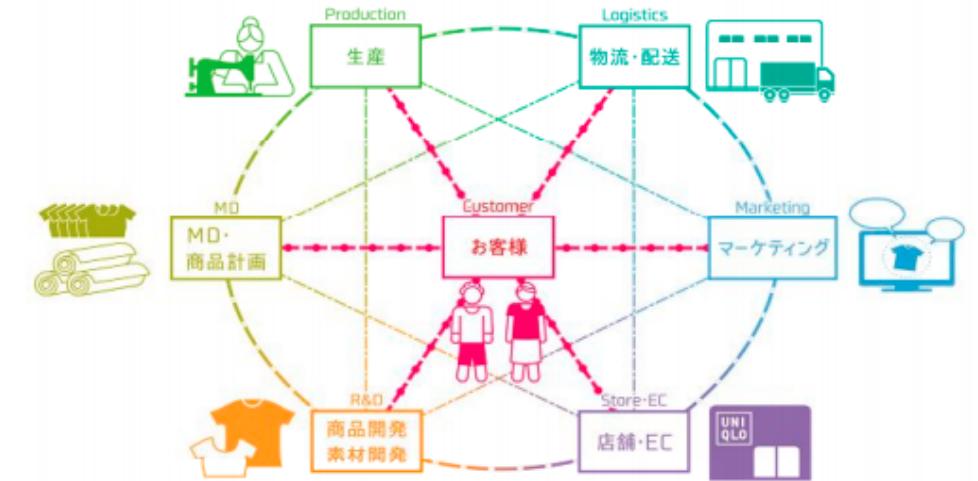
1. 自己紹介
2. ビジネス背景
3. Apache NiFi とは
4. ユースケース
5. デモ
6. Hadoop Summit Tokyo 2016 ミニ報告
7. まとめ
8. 参考資料

ビジネス背景

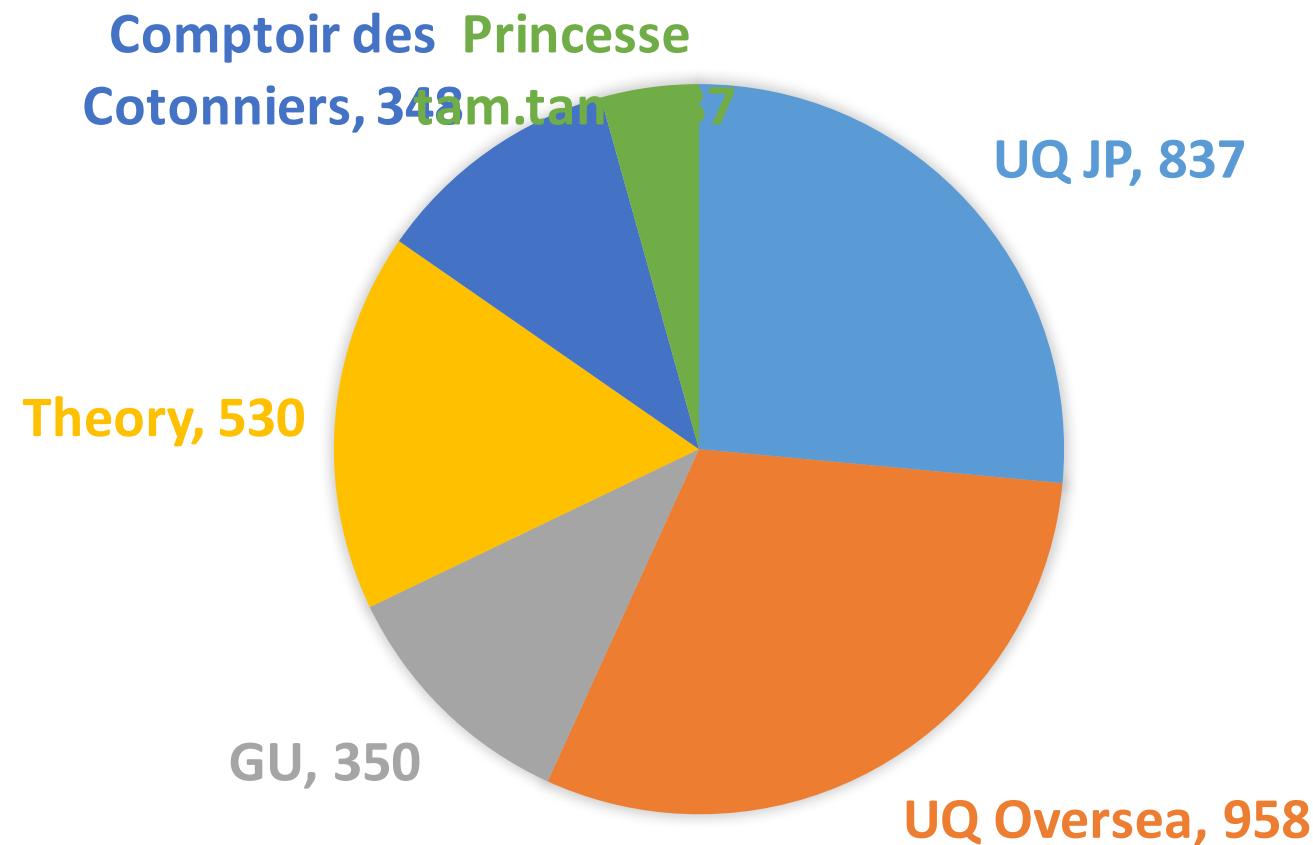
- ・ ファーストリテイリング
 - ユニクロ事業、グローバルブランド事業
 - フルタイムの社員数: 43,639 (2016/8)
 - 店舗数: ユニクロ 1795, グローバルブランド事業 1365 (2016/8)
- ・ 業務改革推進部の役割
 - お客様を中心とした、本当に要望される商品の「素材調達・企画・デザイン・生産・販売までの一貫したサプライチェーン」、「情報製造小売業」への変革 [1]
 - ビッグデータを活用し、お客様の声をすぐに商品化



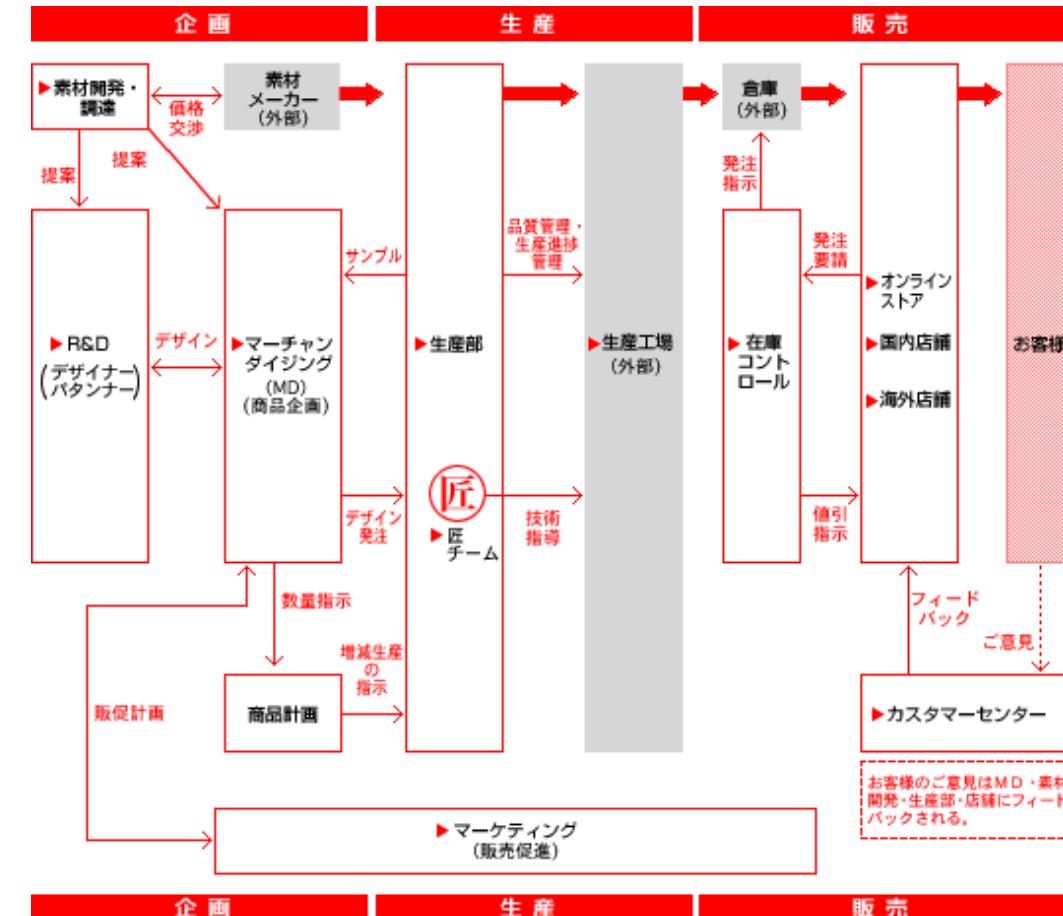
ビジネスの変革



店舗数



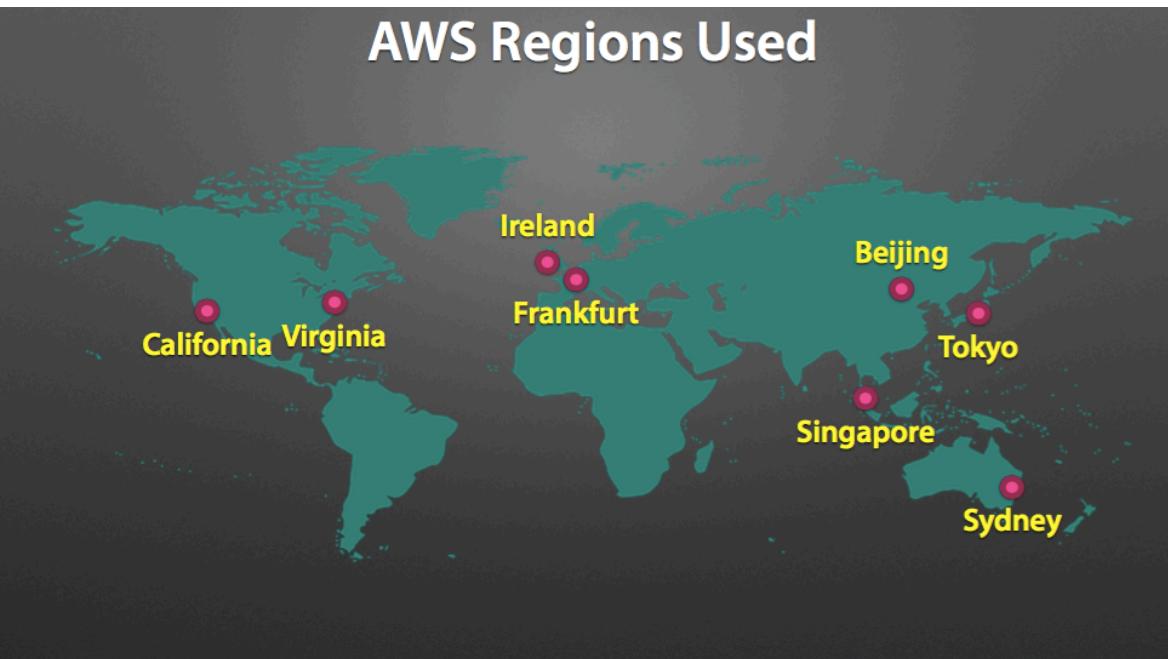
ビジネスモデル



ユニクロのビジネスモデル[2]

AWS の利用状況

- 3000+ EC2 インスタンス (Tokyo Region)
- TokyoのメンバーがリージョンのAWSの構築・運用を一部サポート

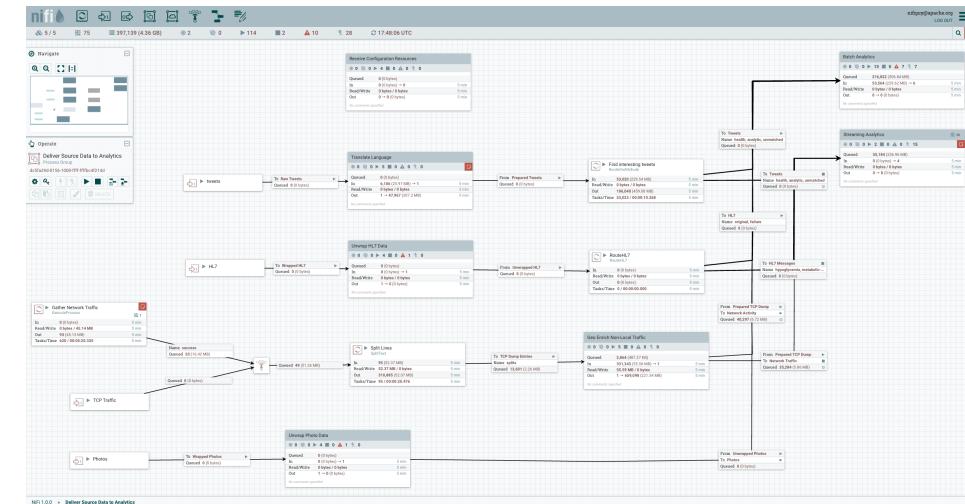


目次

1. 自己紹介
2. ビジネス背景
3. Apache NiFi とは
4. ユースケース
5. デモ
6. Hadoop Summit Tokyo 2016 ミニ報告
7. まとめ
8. 参考資料

Apache NiFi とは

- データフローマネジメントツール
 - 170+ Processorを用いて様々なデータソースとの接続とデータフローの設計・管理がWeb UI上で実現
 - データソース接続: ftp, sftp, udp xml, syslog, http, image, amqp, mqtt
 - データ処理: Hash, encrypt, merge, extract, evaluate, replace
 - フロー制御: Control Rate, Route Content, distributed load



Apache NiFiの歴史

- 2006: アメリカ国家安全保障がNiagaraFiles (NiFi)として開発
- 2014/11: NiFiはApache Software Foundation に寄贈(Apache Incubator)
- 2015/7: ASFのトップレベルプロジェクト

NSA Press Release

For further information contact: NSA Public and Media Affairs, 301-688-6524

NSA Releases First in Series of Software Products to Open Source Community

November 25, 2014



The National Security Agency announced today the public release of its new technology that automates data flows among multiple computer networks, even when data formats and protocols differ. The tool, called "Niagarafiles (Nifi)," could benefit the U.S. private sector in various ways. For example, commercial enterprises could use it to quickly control, manage, and analyze the flow of information from geographically dispersed sites - creating comprehensive situational awareness.

The software is "open source," which means its code is available to the public - in this case, through the Apache Software Foundation. It is the first in a series of releases of in-house software products by NSA's Technology Transfer Program (TTP). Posting the code to open source forums allows the private sector and others to examine the agency's research up close, and potentially benefit from it through additional enhancements and applications. At the same time, the government can gain from related research advances.



Apache NiFi の特徴

- ウェブベースのUI
 - デザイン・制御・フィードバック・モニタリングのシームレスな操作
- 多様な設定可能項目
 - ロス耐性 vs 転送保証
 - 低レイテンシー vs 高スループット
 - 動的な順序づけ
 - ランタイムのフローの変更
 - バックプレッシャー
- データプロベナンス
 - データフローの初めから終わりまでを追跡
- 拡張を意識した設計
 - 独自プロセッサーの作成
 - Rapid development と効率的なテスト
- セキュア
 - SSL, SSH, HTTPS, encrypted content ...
 - マルチテナント認可と認可/ポリシーの内部管理

Apache NiFi のコアコンセプト

- Flow-Based Programming (FBP)
 - *"Flow-based programming defines applications using the metaphor of a "data factory". It views an application not as a single, sequential process, which starts at a point in time, and then does one thing at a time until it is finished, but as a network of asynchronous processes communicating by means of streams of structured data chunks, called "information packets" (IPs)." -- Wikipedia*

FBP 用語	NiFi 用語	説明
Information packet	FlowFile	各オブジェクトはシステム間を移動。
Black Box	FlowFile Processor	システム間のルーティング、変換、仲介の組み合わせて実行。
Bounded Buffer	Connection	Processor間のリンクは、queueのように振る舞い、様々な処理を異なったrateで相互作用させる。
Scheduler	Flow Controller	どのようにprocessが接続されるかの情報を維持し、すべてのprocessが使用するスレッド、配置を管理。
Subnet	Process Group	プロセスとそのconnectionの集合であり、port経由でデータの受け渡しを行う。Process Groupは新しいコンポーネントの作成を複数のコンポーネントの組み立てることで実現。

Apache MiNiFi

- Apache NiFiのサブプロジェクト
 - NiFiを補完するデータソース上におけるデータ収集の機能を提供
- ゴール
 - 小さく軽量のfootprint
 - エージェントの一元管理
 - データプロベナンスの生成
 - 後続のデータフロー管理のためのNiFiとの統合



Apache NiFi 1.0 の新機能

- モダンなUI
- マルチテナント認可
 - 新しい権限の委譲モデル
 - User ID, action, resource 単位でリクエスト毎に認可
 - Managed Authorizer
 - NiFi上でポリシーを管理し、認可ロジックを設定
 - External Authorizer
 - Apache Rangerと連携し、Ranger上で管理
- Zero Master Clustering
 - Master-Worker モデルによる、NCM(Nifi Cluster Manager)の障害ポイントを除去。
 - ZookeeperがCluster CoordinatorとPrimary Nodeの選定するモデルで、Masterの障害に対応

Apache NiFi に感じた魅力

課題：

活動拠点とタイムゾーンが異なる複数のチームで、共通のソフトウェア資産を共用して大規模なデータフローの管理・運用をしたい。学習コストは最小限に抑えたい。

NiFiが提供するソルーション：

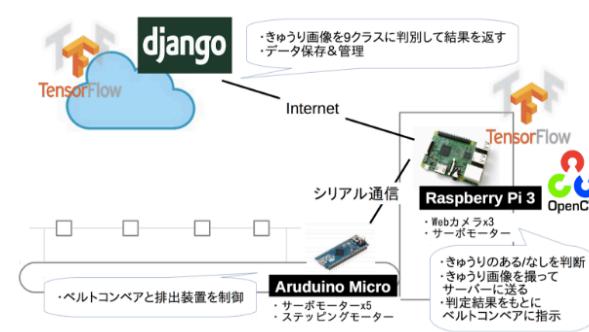
1. 直感的に操作できるリッチなUI
 - データフロー全体を容易に俯瞰できる
 - きめ細やかな警告メッセージ
 - 利用者の学習コストを抑制
2. 豊富なデータソースのConnectorとTemplateが利用可能
3. BackpressureやRate controlといった大規模データに対応した流量制御
4. データフローモデルによるスケールフリーなデータフロー管理メカニズム
 - 複雑なデータフローもProcessor Groupで構造化

目次

1. 自己紹介
2. ビジネス背景
3. Apache NiFi とは
4. ユースケース
5. デモ
6. Hadoop Summit Tokyo 2016 ミニ報告
7. まとめ
8. 参考資料

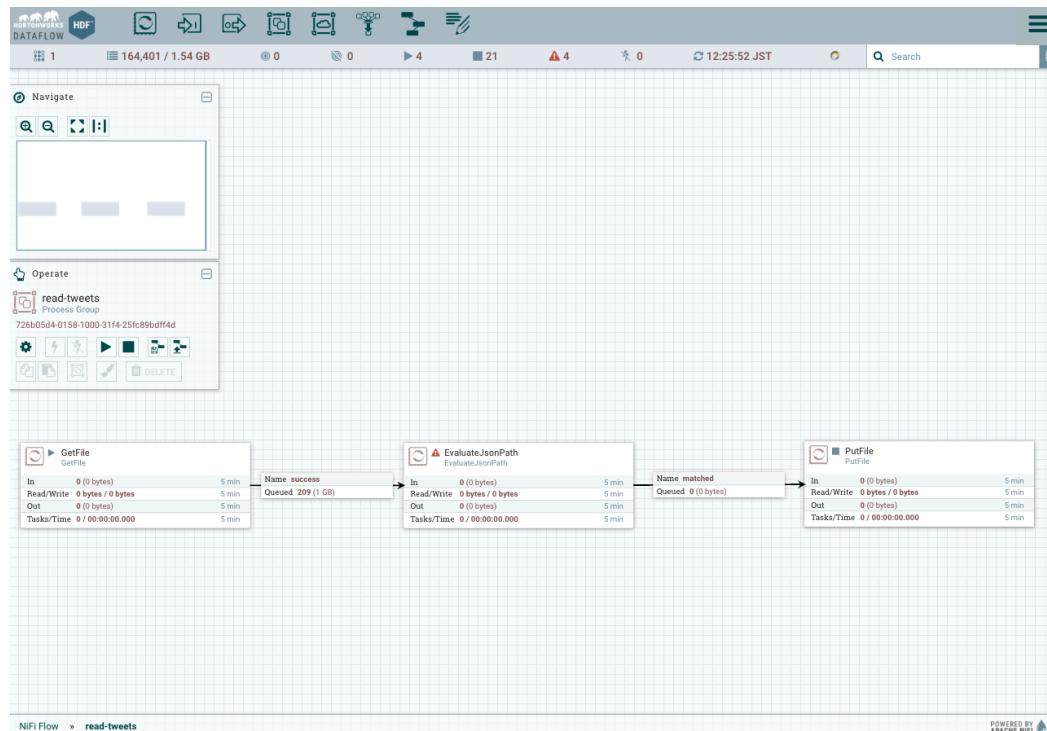
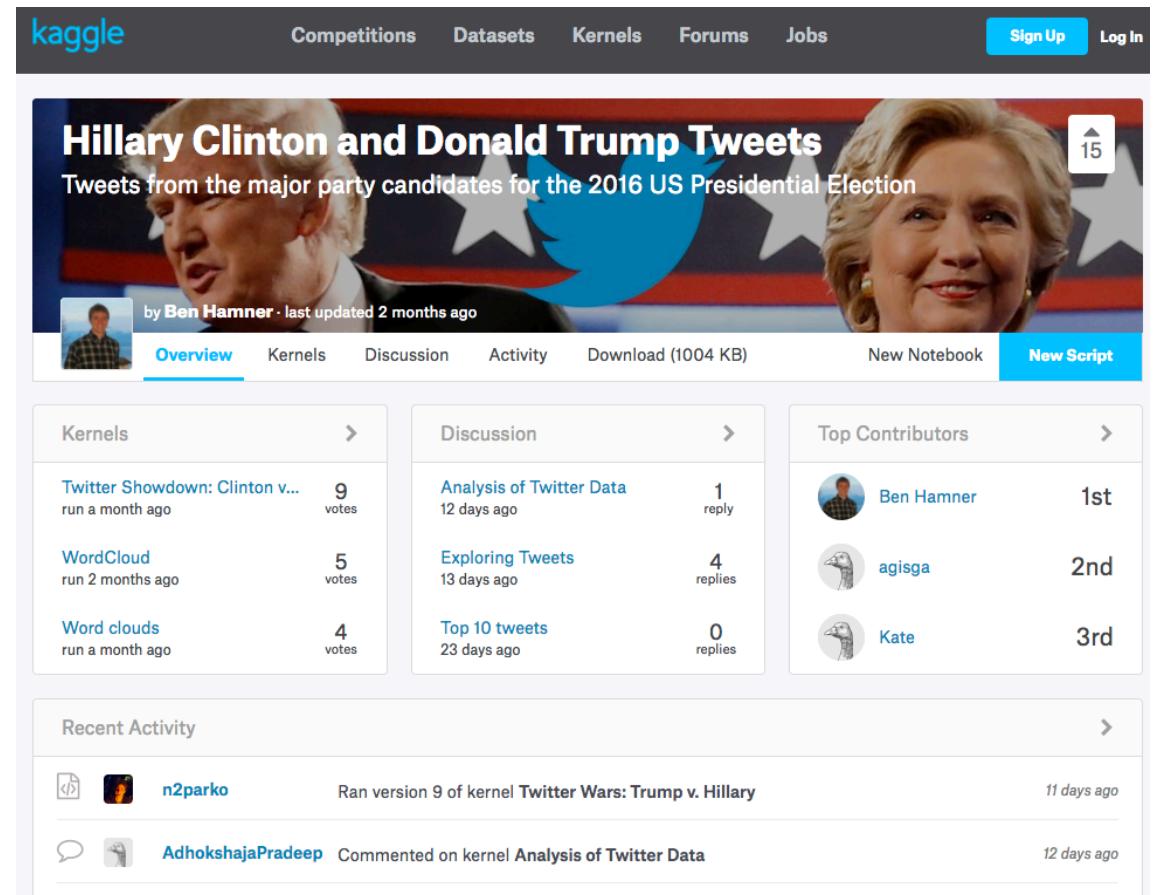
ユースケース

1. SNS上のお客様の声の分析
 - Tweetの商品の評判分析
2. サーバーのログのリアルタイム分析
 - Syslog コネクターでサーバーログをNiFiに転送
 - 異常値を検知し、管理者に通知
3. Raspberry pi によるIoTアプリのRapid Prototyping
 - MiNiFi で Raspberry Piからデータを収集



IoTデバイスの活用事例: Raspberry Pi 3 + TensorFlow によるキュウリの仕分けアプリ [4]

デモ

kaggle Competitions Datasets Kernels Forums Jobs Sign Up Log In

Hillary Clinton and Donald Trump Tweets

Tweets from the major party candidates for the 2016 US Presidential Election

by Ben Hamner · last updated 2 months ago

Overview Kernels Discussion Activity Download (1004 KB) New Notebook New Script

Kernels

- Twitter Showdown: Clinton v... 9 votes run a month ago
- WordCloud 5 votes run 2 months ago
- Word clouds 4 votes run a month ago

Discussion

- Analysis of Twitter Data 1 reply 12 days ago
- Exploring Tweets 4 replies 13 days ago
- Top 10 tweets 0 replies 23 days ago

Top Contributors

Rank	User	Contributions
1st	Ben Hamner	15
2nd	agisga	10
3rd	Kate	5

Recent Activity

- n2parko Ran version 9 of kernel Twitter Wars: Trump v. Hillary 11 days ago
- AdhokshajaPradeep Commented on kernel Analysis of Twitter Data 12 days ago

運用面の課題

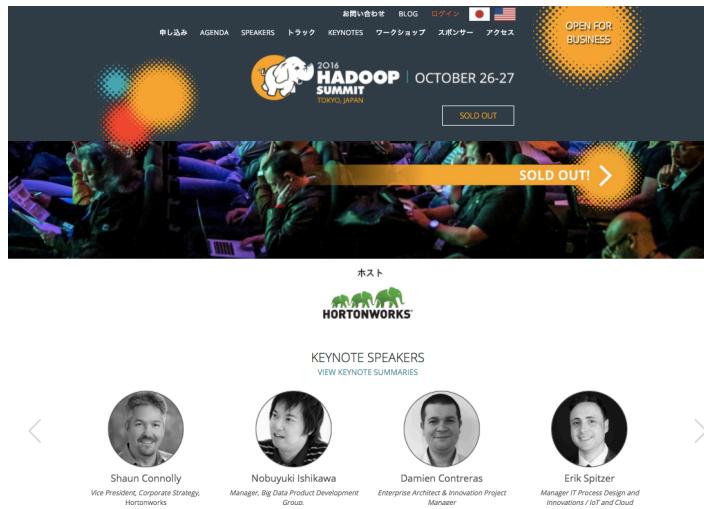
- データフロー作成のサポート機能
 - 間違って消してしまったフローをどうやってリカバリーするか?
 - Undo/redoが欲しい
 - フローの個別カスタマイズ・作り込みのための学習コスト
 - テンプレートにパラメータを渡して、データフローをカスタマイズしたい
- 野良データフローや俺俺スタイルのデータフロー乱立問題
 - データフローの標準化・自動のスタイルチェック・バリデーションをしたい
- データフローの24時間/365日の運用体制
 - 現実的なSLAを設定
 - E.g. 当日の「速報値」は誤差を許容し、翌日のバッチ集計値を「確定値」とする
 - 適切なアプリケーションでの利用
 - アワリー・デイリー集計で済む場合は使わない

目次

1. 自己紹介
2. ビジネス背景
3. Apache NiFi とは
4. ユースケース
5. デモ
6. Hadoop Summit Tokyo 2016 ミニ報告
7. まとめ
8. 参考資料

Hadoop Summit Tokyo 2016 ミニ報告

- Hadoop 関連の最大規模の有償カンファレンス
 - 技術とビジネスの両面を扱ったセッション
 - 2016年は世界36カ国から4000人が参加[5]
- 初の東京の開催
 - 2日間で400+人が参加
 - 48 セッション



Apache NiFi 関連セッション

- Apache Nifi の紹介
 - Apache NiFi 1.0 in Nutshell [6]
 - Apache NiFi Crash Course [7]
- 利用事例
 - From a single droplet to a full bottle, our journey to Hadoop at Coca-Cola East Japan [8]
- ユーザーの交流会
 - Bird of Feather: Steaming & Data Flow
 - 1つのグループでApache Nifi について意見交換



Apache NiFi の紹介

- Apache NiFi 1.0 in Nutshell
 - Nifiの全体像の紹介
 - NiFi 1.0の拡張の紹介
 - Edge デバイスagentであるMiNiFiの紹介
- Apache NiFi Crash Course
 - Hortonwork社による3時間のハンズオン
 - データフローとストリーミングの基礎
 - 交通量のデータを処理するデータフローを作成するラボ



Apache NiFi の利用事例

- From a single droplet to a full bottle, our journey to Hadoop at Coca-Cola East Japan
 - 自動販売機の補充量の予測
 - 自販機への補充のためのトラックの訪問回数を削減
 - トラック上の在庫を最適化
 - 売り切れの防止
 - Apache NiFiのデータフローのガイドラインも紹介



目次

1. 自己紹介
2. ビジネス背景
3. Apache NiFi とは
4. ユースケース
5. デモ
6. Hadoop Summit Tokyo 2016 ミニ報告
7. まとめ
8. 参考資料

まとめ

- データフローマネージメントツールApache NiFiを紹介
 - 170+ ProcessorとTemplateを用いて、Web UI上でのデータフローの構築が可能
 - Flow-Based Programmingにより、大規模なデータフローの構築と管理を実現
 - NiFi 1.0では、複数チームによる利用を想定した認可機能や可用性向上させる機能が追加
- Hadoop Summit Tokyo 2016では、Apache NiFi の注目も高く、日本での利用事例も紹介された



<https://www.fastretailing.com/employment/ja/fastretailing/jp/career/it/>

参考資料

[1] ファーストリテイリング 2016年8月期の業績と今後の展望,
http://www.fastretailing.com/jp/ir/library/pdf/20161013_yanai.pdf

[2] ユニクロのビジネスモデル | FAST RETAILING CO., LTD.,
<http://www.fastretailing.com/jp/group/strategy/uniqlobusiness.html>

[3] JAWS Days 2016: Microservice @ Fast Retailing // Speaker Deck,
<https://speakerdeck.com/fastretailing/jaws-days-2016-microservice-at-fast-retailing>

[4] Google Cloud Platform Japan 公式ブログ: キュウリ農家とディープラーニングをつなぐ TensorFlow,
http://googlecloudplatform-japan.blogspot.jp/2016/08/tensorflow_5.html

[5] Hadoopが変えるデータとヒトへのアプローチ —「Hadoop Summit 2016 Tokyo」レポート | gihyo.jp 技術評論社,
<http://gihyo.jp/news/report/2016/10/3101>

[6] Apache NiFi 1.0 in Nutshell,
<http://www.slideshare.net/HadoopSummit/apache-nifi-10-in-nutshell-67930403>

[7] Hadoop Summit Tokyo Apache NiFi Crash Course,
<http://www.slideshare.net/HadoopSummit/hadoop-summit-tokyo-apache-nifi-crash-course>

[8] From a single droplet to a full bottle, our journey to Hadoop at Coca-Cola East Japan,
<http://www.slideshare.net/HadoopSummit/from-a-single-droplet-to-a-full-bottle-our-journey-to-hadoop-at-cocacola-east-japan>