

◆◇『スーパーエンジニアの独り言 第5回 “Apache HadoopとBig Data”』◇◆

今回の話題は「Doug Cutting氏の来日講演」の続編です。

Doug Cutting氏は“Apache Hadoop”の創造主です。彼の基調講演を拝聴した(2011年9月1日)その様子をお届けします。講演当日の会場では所狭しとパイプ椅子が敷き詰められ聴講者がひしめき合う満員御礼で、熱気立ち上がる中、プロジェクターとDoug Cutting氏の勇姿を拝見しつつ、彼が発した言葉を一言も漏らさぬ覚悟でミミスが這うような字で手帳にひたすらメモを走り書きした次第です。このメモを元に講演顛末の後半を書き起こします。

“Apache Hadoop: A New Paradigm for Data Processing”
Doug Cutting, Architect, Cloudera, Inc.
(講演されたスライドタイトル一覧は前々回のメルマガをご参照ください。)

Facebook、Twitter、Amazon、eBayといった錚々たるサービス企業でHadoopが採用され、企業向けとしても適用が進んでいる実績がありますが、Hadoopを『新しい基盤』と位置付けているポイントとして挙げたのは以下になります。

- ◇ 「コモディティ化されたハードウェア」
- ◆ 「シーケンシャルファイルアクセス」
- ◇ 「シャーディング」
- ◆ 「オープンソースである」

まず「コモディティ化」は、計算資源が劇的にコストダウンし、性能も大幅にアップしている現象であり、これを活用するためには何千ものコンピュータにスケール可能な(拡張性が柔軟である)仕組みが有効であるという事。

次に「シーケンシャルファイルアクセス」についてです。その対極であるランダムアクセスはハードドライブのシークを発生することになります。ハード的にシークの時間は短縮しておらず「シークタイム=無駄な時間」となります。シーケンシャルなアクセスとすることで無駄を省き実際の操作に集中させる事ができるというのです。一例は、パッチシステムにすることです。

「シャーディング」というのは、データベースを分割することです。データベースを分割することで拡張性が生み出されます。ここで重要なことが、信頼性を保証しながら(マニュアルではなく)自動化を推奨していくことにあります。つまり、分割した上でのフォールトトレラント環境が必要であり、不意のハードウェア故障にも対応できるようになります。

「オープンソースである」ことについての利点は理解されつつあるでしょう。オープンコミュニティで協力して作成していく過程が、対象の汎用性を高め同時にその品質を高めています。閉じられたチームで開発するのではなく、多くの有能な開発者達と作業を通して知り合うことで、開発者自身の向上心を掻き立て、それによりソフトウェアの品質も高まるというものです。これらを満たしている仕組みがHadoopであり、その可能性が期待されています。

講演は、“the future is data” 将来への提示の話で締め括られました。

Web 2.0の教訓で「データが重要である」ことは言うまでもなく、加えて沢山のデータを収集することが重要であることを繰り返して強調しています。収集した生の大量データをシンプルなアルゴリズムで解析すれば良い

ctc201111

という主旨です。複雑なアルゴリズムにおける改良を繰り返すのではなく、データを収集するのに注力した方が得策だとのこと。つまり、必要なデータを全部集めることで、一番効率的な解析が行えるのです。ビジネスにおいて、より多くのデータを集めて解析することが更なる改善に繋げることに有効である、という示唆でした。

この命題は、新しい分散型データOSのカーネルとして「Hadoop」を利用することで実現可能であります、というDoug Cutting氏の締めのお言葉でありました。

講演ではコアのHDFSとMapReduceやHive、Pigなど周辺コンポーネントの簡単な紹介がありましたが、これはまた機会があればということにさせていただきます。因みに、彼は名前の由来であるHadoop（黄色い象のぬいぐるみ）同伴でした。講演後に聴講者と一緒に「Doug + Hadoop」で記念写真を撮られていました。

超高速スーパーコンピューターの登場を待つこともなく、大量のデータを集めて分散並列処理を行うことが可能となる時代が到来しつつあります。コンシューマサービスを生業とする大企業ではHadoopを活用していますが、エンタープライズ適用ではトレーサビリティと仮定しただけでも製造、流通、食品、アパレル、あらゆる業界でビジネスを拡大させます。ビッグデータ (Big Data) を扱う取り組みは既にトレンドになっているのです。

次回は、急速に浸透し普及が進む電子書籍の話題です。お楽しみに。

関連コースの詳細情報はこちら：

「クラウド・仮想化 基礎／入門」関連コース
<http://www.school.ctc-g.co.jp/cldvir/>

「ストレージ関連基礎」コース
<http://www.school.ctc-g.co.jp/vmware/index.html#org>

「Java」関連コース
<http://www.school.ctc-g.co.jp/java/>

■お問合せ・ご意見・ご感想は◆CTC教育サービス◆窓口まで
シーティーシー・テクノロジー株式会社 エデュケーションサービス部
E-Mail : kyouiku@ctc-g.co.jp / TEL : 03-5712-8701

●外部委託について

弊社はメールニュース配信業務をシーティーシー・ビジネスサービス株式会社（CTC100%出資子会社）に委託しております。

●本メールマガジン編集・配信責任者

CTCT エデュケーションサービス部 部長 篠原 義一
所在地：東京都世田谷区駒沢1-16-7 ctc_edu_mail@ctc-g.co.jp

●個人情報保護方針

CTCグループの個人情報保護方針につきましては下記URLをご参照ください。

http://www.ctc-g.co.jp/guide/security_policy.html?top=b_security

●配信中止及びお問合せ対応について

・「CTC教育サービス News&Topics」の配信が不要な場合には、下記URLから配信停止のお手続きを行ってください。

<https://krs.bz/ctc-g/m/ctc-education>

ctc201111

- ・当社では、複数種類のメールマガジンやメールニュースを発行しております。大変お手数ではございますが、CTC教育サービス以外からのメール配信についての受信拒否および個人情報に関するご要求は、各メールに記載の個々の連絡先宛にそれぞれご連絡をお願いします。
 - ・受信者ご本人様からの個人情報の開示・訂正・削除に関するご要求は、随時 ctc_edu_mail@ctc-g.co.jpにてお受けいたします。
-