# シェルスクリプトを用いた大規模データ処理の提案と授業実践報告

# Proposal of large-scale data processing using shell scripts and class practice report

大野 浩之[*1], 松浦 智之[*2], 當仲 寛哲[*2], 森 祥寛[*1]
Hiroyuki OHNO[*1], Tomoyuki MATSUURA[*2], Nobuaki TOUNAKA[*2], Yoshihiro MORI[*1]
[*1] 金沢大学学術メディア創成センター
[*1]Emerging Media Initiative, Kanazawa University
[*2] ユニバーサル・シェル・プログラミング研究所
[*2]Universal Shell Programming Laboratory Ltd.
Email: mori4416@staff.kanazawa-u.ac.jp

**あらまし**：高等教育においてデータサイエンス教育が求められている中，多くの場合，データ処理に用いられるのは，Python などの高級言語による処理である．それに対して，我々は，シェルスクリプトのコマンドを用いたデータ処理方法を提案しており，そのための新しいコマンド群（PT4A）の開発等も行っている．本発表では，その手法について紹介と，授業実践の報告をする．
**キーワード**：テンプレートシェルスクリプト，PT4A， 大規模データ処理，教育実践

## 1. はじめに

高等教育においてデータサイエンス教育が求められている中，多くの場合，データ処理に用いられるのは，Python などの高級言語による処理である．また大規模データ処理を行うための手法とあげられるのは分散処理技術 Hadoop などであろう．

それに対して，我々は，シェルスクリプトのコマンドを用いたデータ処理方法(1)(2)を提案しており，大野と當仲を中心とする共同研究では，アカデミック向けのコマンド群（PT4A：Personal Tukubai for Academic）(3)の開発なども行っている．

本稿では，シェルスクリプトを用いた大規模データ処理について，その手法の紹介をするとともに，金沢大学で実施した授業実践について報告する．

## 2. シェルスクリプトを用いた大規模データ処理と PT4A

UNIX コマンドによる処理は，ファイルの処理である．これは UNIX の哲学(4)に基づく処理であり，その中のシェルスクリプトのコマンド群は，C 言語などによるプログラミングの結果の活用に他ならない．シェルスクリプトのコマンドを使って，データ処理を行う場合，その多くは，データファイル内に書かれたプレーンテキストを扱うことになり，直観的な作業が可能になる．そして，1 コマンド毎の結果を確認しながら，コマンドとコマンドを「｜（パイプ）」で繋ぐことで，段階を追った作業が可能となり，最終的に，ある特定のデータ処理を行うスクリプトが完成する．これは扱うデータの規模が大きくなっても，ほぼ同じ作業が行える．

一方で，1 コマンドで行える操作が 1 つのため，行いたい作業用コマンドが存在しないことがある．そこで用意されたのが，USP 研究所の手によるエンタープライズ向けコマンド群 uspTukubai であり，前述の PT4A は，これらをパーソナル環境で動作させるためのパッケージである．なお，PT4A はアカデミック分野に対して，無償提供されていて，Windows10/WSL， macOS， GNU/Linux （Ubuntu, CentOS）で利用可能となっている．

## 3. 大規模データの準備

授業で学生が演習に使用する大規模データとして，定型データ（構造化データ）と非定型データ（非構造化データ）の 2 種類を準備した．規模の目安として，Excel などの表計算ソフトウェアで使用できないレコード数（1,048,576 行分以上のデータ）を目指し，ファイルは，シェルスクリプトを使って処理するためテキストベースの CSV ファイルにした．準備したデータは USB メモリに保存し，授業を履修する学生に，授業期間中貸出をした．

### 3.1 定型データ：気象庁アメダスデータ

気象庁では，アメダス（AMeDAS：Automated Meteorological Data Acquisition System：自動気象データ収集システム）と呼ばれる日本の各地 1,300 カ所の気象観測所で構成される無人観測施設がある．ここで取得されたデータは，気象庁の Web ページから公開されており，誰でも取得可能である(5)．またこれらのデータは，「気象業務センター」から一括で購入する事もでき，今回は，大規模データ処理の演習のため，2008 年から 2020 年までの 1 分ごとの 1,320 地点のデータを，共同研究者である USP 研究所が購入した．これは圧縮されたバイナリデータで保存されており，1 年間で約 6 億 8000 万レコード，全部で約 83 億レコード分，1,320GB 分のデータとなる．

演習で使用するために，これらのデータをテキストデータに変換すると共に，10 分毎のデータになるように間引いた．併せて，記録されているデータの項目も整理し，「観測所,時刻,雨量,気温,風速,風向,日照時間」という形で，1 年毎にファイルを分割したデータセットを準備した．この結果，合計でレコード数 828,734,400，およそ 45GB のファイルが用意できた．この内，2018 年と 2019 年のデータについて

は，学生が授業で演習に使用するパソコンの性能を考慮し，さらに月ごとにファイルを分割したものを用意した．

### 3.2 非定型データ：Twitter データ

Twitter でオリンピックについてつぶやかれたつぶやきは，kotoriotoko(6)を用いて収集した．kotoriotoko は，松浦によって作成されたコマンド群で，これを使用してつぶやきの収集が可能となっている(7)．この授業では「オリンピック」をキーワードにつぶやきを収集した．収集期間は 2020 年 1 月から 2021 年 5 月までで，およそ 100GB の非定型データである．ただし，非定型部分はつぶやきの文章部分であり，つぶやかれた日時などについては，定型化された半構造化データの形をとっている．

## 4. 授業での実践

我々は，2016 年度から，ものグラミング(*)と POSIX 中心主義(*)を題材として，大学コンソーシアム石川いしかわシティカレッジ(8)提供科目として，「クラウド時代の「ものグラミング」概論」と「シェルスクリプト言語論」の 2 つの授業の開講から授業での実践を開始し，2019 年度からは集中講義「シェルスクリプトを用いた「ものグラミング」演習－POSIX 中心主義に基づく電子工作－」としても開講している．本研究では，2021 年度 4 月に開講した「シェルスクリプトを用いた「大規模データ処理」演習」の授業について報告する．

表 1 授業内容

| | 授業内容 |
|---|---|
| 第 1 回 | 講義概要，PC の環境確認・設定，PT4A 設定 |
| 第 2 回 | CUI 入力と簡単なシェルスクリプトを用いた演習 |
| 第 3 回 | PT4A の使い方と演習 |
| 第 4 回 | 特別講演と大規模データ処理における課題提示 |
| 第 5 回・第 6 回 | 提示された課題解決に向けた大規模データを用いた演習 |
| 第 7 回・第 8 回 | 成果発表と追加課題のていじとまとめ |

この授業では，学生にノートパソコンを準備してもらい(9)，そのパソコンを使い，シェルスクリプトによる大規模データ処理の演習をする．授業内容は表 1 の通りである．授業では，前節の内容で用意した大規模データを，USB メモリに保存して学生に貸与した．これは使用するデータが大規模であるため，ダウンロードすることが適さないためである．また，そのファイルサイズ故に学生が準備している携帯型パソコンのストレージに収まらない可能性を考慮した．

シェルスクリプトを使用する環境として Windows10 の場合は WLS を，macOS の場合はターミナルを使用する．授業第 1 回から第 3 回までは，パソコンの設定（WSL の導入から，PT4A などのコマンド群のインストールなど）を行うと共に，GUI ではない CUI による操作方法，コマンドライン入力とは何か，コマンド入力の方法などを説明している．そこで大野，森で，これら前段の知識をオンラインで学習可能な教材の作成を行い，授業開始時にできるだけ同じ状態で作業を始められるように整えている．第 4 回授業では，當仲によるシェルスクリプトを用いたシステム構築やデータ処理が実際のビジネスの現場でどのように扱われているかを示し，学生が本授業の教授 j 内容を何故学ぶ必要があるのかを明確にした．第 5 回以降は，前述で用意した大規模データを実際に使用した演習を行った．

## 5. まとめ

さまざまな場面でデータサイエンスを活用する場合，そこにある大規模データをどのように処理するかが大きな問題となる．本研究では，その処理方法としてシェルスクリプトを用いた．この処理方法だけで，全てが行えるわけではないが，データを分かりやすく扱う方法としては非常に有意であろう．また IoT デバイスなどを用いたデータ収集との親和性も高く，インターネットを介してデータを集め，それを逐次処理していく作業にも適している．

我々は，今後，これらの手法を取りまとめて，データ・教材・教育方法を併せて，広く社会に開示することを目指している．

## 謝辞

本研究は，金沢大学学術メディア創成センターと USP 研究所の共同研究として推進された．関係各位のご厚意ご高配に，深く感謝する．

## 参考文献

(1) 中村和敬，當仲寛哲:“Unix シェルスクリプトによる企業システム構築”，情報処理学会第 77 回全国大会, 2A-01,（2015）.
(2) 松浦智之:“すべての UNIX で 20 年動くプログラムはどう書くべきかデプロイ・保守に苦しむエンジニア達へ贈る［シェルスクリプトレシピ集］”，シーアンドアール研究所,（2015）
(3) USP 研究所，PT4A: Personal Tukubai for Academic, https://www.usp-lab.com/pt4atop.html（2021-06-09 アクセス確認）
(4) Mike Gancarz, 芳尾桂監訳, UNIX という考え方, 2001 年オーム社
(5) 気象庁，過去の気象データ・ダウンロード, https://www.data.jma.go.jp/gmd/risk/obsdl/index.php（2021-06-09 アクセス確認）
(6) 秘密結社シェルショッカー日本支部，恐怖！ 小鳥男（オンライン），入手先 ⟨https://github.com/ShellShoccar-jpn/kotoriotoko⟩（参照 2019-02-04）
(7) 松浦智之, 當仲寛哲, 大野浩之，“大量ツイートの収集・分析を個人で手軽に実現可能にする方法の提案”, デジタルプラクティス 11(1), 173-190，2020-01-15
(8) 大学コンソーシアム石川いしかわシティカレッジ, ⟨https://www.ucon-i.jp/newsite/city-college/index.html⟩（2021-06-09 アクセス確認）
(9) 森 祥寛 , 大野浩之，NAKASAN CHAWANAT 他, “金沢大学における携帯型パソコン必携化に関する 12 年間の取組 ”, 学術情報処理研究 23(1), 29-42, 2019