Ghislain Fourny

Big Data Exercise

0. Introduction (Jupyter)
Lecture Team

Ghislain Fourny

Ingo Müller (Head TA)

Darko Makreshanski (TA)

Marco Ancona (TA)

Konstantin Taranov (TA)
(Tentative) Exercise Schedule

- Sep 20: Azure + Notebooks
- Sep 27: SQL brush-up
- Oct 4: Azure object storage
- Oct 11: HDFS
- Oct 18: Column Store
- Oct 25: MapReduce
- Nov 1: Spark
- Nov 8: Well-Formedness (XML, JSON)
- Nov 15: Validation (XML, XML Schema)
- Nov 22: Querying/XQuery
- Nov 29: Indexing + MongoDB
- Dez 6: TBD
- Dez 13: Tableau + Excel + ?
- Dez 20: Edgar XBRL filings + XML/Schema/Cube brush-up
(Tentative) Exercise Schedule

- Sep 20: Azure + Notebooks
- Sep 27: SQL brush-up
- Oct 4: Azure object storage
- Oct 11: HDFS
- Oct 18: Column Store
- Oct 25: MapReduce
- Nov 1: Spark
- Nov 8: Well-Formedness (XML, JSON)
- Nov 15: Validation (XML, XML Schema)
- Nov 22: Querying/XQuery
- Nov 29: Indexing + MongoDB
- Dez 6: TBD
- Dez 13: Tableau + Excel + ?
- Dez 20: Edgar XBRL filings + XML/Schema/Cube brush-up

Exercises with
- Azure Cloud services &
- Jupyter
Exercises with
- Oxygen XML Editor
Exercise Format

- Week Nr. X
  - Lecture on X on Tuesday (10-12)
  - Exercises on Wednesday & Friday (13-15)
    - Introduce exercise sheet for X-1
      - Get started with new technology
      - Discuss tips for exercises
    - Discuss solution of exercise sheet for X-2
  - At home
    - Solve the rest of exercises for X-1
    - Read up on topics for X
Plan for Today: Get Started with Jupyter

Access credentials

- Username: <ethzid>@ethzbigdata2016outlook.onmicrosoft.com
  - <ethzid> is your ldap username
- Password: “BigData<LegiNr without dashes>”
  - For a legi nr. of 12-345-678, then password is “BigData12345678”

Access Microsoft’s Notebook Server

notebooks.azure.com
Sharing Your Result Made Easy

With Azure Notebooks, unleash your ideas in the cloud with the Jupyter Notebook

Go to my Notebook Server
Show me some samples
Big Data – Exercises

Fall 2016 – Week 1 – ETH Zurich

1. Jupyter Basics

Welcome to this Jupyter notebook. Jupyter is a web-based open-source tool based on Python that allows you to run python (and other types of) code, visualize results and discuss results, and organize everything into notebooks like this one. We use the notebook server on Microsoft Azure, but you can also install your own. A notebook is organized in cells. Cells of this notebook contain Python code (but other cell types exists). To run a cell, select it, then press ctrl+enter. Try it out!

```
In [ ]: print("Hello World")
```

By default, the last expression is printed. Like this:

```
In [ ]: maxX = 10
   [x * x for x in range(maxX)]
```

You can also edit the text. Just double-click on a cell. It's made with markdown code. After you are done editing, press ctrl+enter.

We will do about half of the exercises with Jupyter. You will learn most things as we go. The notebook you are seeing is your copy stored on your account (including the output of the cells). Write your answers inline and save regularly. (Also save a copy somewhere else from time to time – this service is still beta.)

Read more:
- Jupyter
- Learn Python
- Python documentation

2. Bash Scripts
Jupyter

- Execute cells with ctrl+enter (shift+enter also works)

- Service is in beta phase
  - Might be buggy
  - In cases of problems → kernel and clear output, or close and reopen notebook