# Learning plan – Jakub Cajzl

## First days

**Checklist Item #0** Access to AdaLab (Hue, Jupyter Hub, GitLab)

<https://git.adastragrp.com/>

GIT – kurz na datacamp

IDE – PyCharm, … - propojit s repositářem

**Checklist Item #1** Python libraries: json, csv, xml, subprocess

<https://www.udemy.com/course/complete-python-bootcamp/learn/quiz/178926#overview>

* **general** – timestamp, datetime, try, except, finally
* **json** – mine nested jsons, from string to json, load, export, loops, lambdas, .get(), with/without comments
* **csv** – load, save, change separator, header, encoding
* **xml** – parsing, with/without comments,
* **subprocess** – basic demo
  + send command and return value evaluation – return code – result = subprocess.run(command, shell=True, …) result.check\_returncode()
* **logger** – console, file logging, config
* **gitlab** – create repo and use it for store notebooks/demos
* **requests** – find some Open API – try requests -> json/dict
* **take a look at** PEP8**,** python OOP basics, package/lib installation and custom modules in python

**Checklist Item #2** Pandas, Numpy – Kaggle dataset

* DataFrames – joins
* Basic transformations, lambda

**Checklist Item #3** Hadoop stack – HDFS, Hive, Impala, Yarn – what is it for?

* <https://www.udacity.com/course/intro-to-hadoop-and-mapreduce--ud617>
* Cloudera Data Analyst – kurz nahraný na teams
* Add more sources, links, videos, docs
* **Airflow** – architecture, cron, demo DAG – Airflow on AdaLab? – ask Admins
* **Parquet** – theory, and more formats (AVRO, KUDU)
* **Hive vs**. **Impala** – metastore, differences
* **Hive 3** “New stuff” – materialize views vs classic view and more

**Hive, Impala**

* External, managed tables
* Refresh, Msck repair
* Partitions, Bucketing

**Checklist Item #4** PySpark – online kurz (DataFrame, RDD) – Udemy - BD Teams

* **DBX** kurz <https://partner-academy.databricks.com/learn/signin>

Spark architektura – <https://partner-academy.databricks.com/learn/course/112/introduction-to-apache-spark-architecture>

Spark DF API/SQL – <https://partner-academy.databricks.com/learn/course/63/apache-spark-programming-with-databricks>

* **Jupyter** – Adalab, Local – Virtual, AWS, Azure?
* **DF-Schema** – infer schema, define schema
* **DF** – transformations, lambda, do not use UDFs, export to parquet (with/without schema) – how to rename file, more than one file per DF?
* **RDD –** try some examples
* **BD data formats** – Parquet, Delta,
* **Json** – load json as sring to DF – convert to struct (datatyp struct)
  + clear, enhance, export
  + comments in json files - parsing
* **Struct** –spark DF api – dotazování and structem
* **Optimalizcni cast, Catalyst Optimizer, query plan -> fyzicky plan**
* **Mody spousteni Sparku, local vs cluster ..**

**Checklist Item #5** PySpark – Kaggle dataset, transformations

* Vybrat si nějaké open API nebo kaggle dataset, nejlepe incremental data
* Udělat návrh jak celé řešení bude vypadat (architektura obrázek)
* Data chodi do slozky (jako csv), nacist, zpracovat Sparkem, jak to dostat do databaze (parquet na hdfs, pripojeni do db)
* Skript s configem (cesty do slozek, jine nastaveni) + logovani pres logger
* Retence logů

**Checklist Item #6** Datalake table structure – bronze/silver/gold (DBX naming) … (architecture – landing, stage, int, mart)

## Next steps

**Checklist Item #7** Spark Architecture terms

* Spark config – co se vyplnuje napr.
* Shuffle
* Narrow/wide transformation
* Coalesce vs. Repartition
* RDD
* broadcast variable a proč se používá
* application/job/stage/task
* core/thread
  + virtual cores
* executor/driver/worker
* client/cluster/local execution mode
* how to read/write from spark (API, JDBC, Impala, rel. DB)
  + pripojovani pomoci konektoru k databazi
* Catalist optimizer
* Export a pojmenování souboru/ů jde to?

**Checklist Item #8** NoSQL DB Teorie: Cassandra, HBase, MongoDb, Redis, Neo4j, ElasticSearch + Kibana

* Jaké máme typy
* Architektura
* Na jaké typy dat se hodí

**Checklist Item #9** Výběr certifikace – začít s výběrem

* DBX Spark (dobrý start)
* Cloudy – AWS, Azure, GCP
* Cloudera Developer -X

**Checklist Item #10** Příprava na certifikaci

## Useful Links

1. [Analytika](https://education.cloudera.com/lpaths/4289399/courses/332525/details)
2. [Základy data warehousingu](https://education.cloudera.com/lpaths/4289399/courses/335080/content)
3. [Hive cheat sheet - edited](https://github.com/cherkavi/cheat-sheet/blob/master/hive-cheat-sheet.md)
4. [CDP 3001 Data developer](https://adastrabiz.sharepoint.com/:f:/s/BigDatateam/ErEaBzC-ei1ChNufvaUEQd0BPM55SrSf7-K0N3huNNfa1Q?email=Dominika.Hubova%40adastragrp.com&e=3kJ8hf)
   1. https://www.cloudera.com/about/training/certification/cdp-datadev-exam-cdp-3001.html

## Notes