**SYLLABUS**

**Fall semester 2022-2023 academic years**

**on the educational program “5В06101- data engineering”**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Discipline’s code** | **Discipline’s title** | **Independent work of students (IWS)** | **Number of credits** | | | | | **Number of credits** | **Independent work of student with teacher (IWST)** |
| **Lectures (L)** | **Practical training (PT)** | | **Laboratory (Lab)** | |
| **VID 1206** | Indroduction to Data Engineering | 4 | 3 | - | | 6 | | 9 |  |
| **Academic course information** | | | | | | | | | |
| **Form of education** | **Type of course** | **Types of lectures** | | | **Types of practical training** | | **Form of final control** | | |
| Full-time | Combined | Problematic, analytic lectures | | | the solution of the problem | | Oral | | |
| Lecturer | Zhienbayev Meiran | | | | | |  | | |
| e-mail |  | | | | | |
| Telephone number | 87012222737 | | | | | |

|  |  |  |
| --- | --- | --- |
| **Aim of course** | **Expected Learning Outcomes (LO)\***  As a result of studying the discipline the undergraduate will be able to: | **Indicators of LO achievement (ID)**  (for each LO at least 2 indicators) |
| The main goal of this course is to prepare students for using technologies and tools which are needed for data engineering. The course also covers some basic topics in data analytics and programming language area. | **LO1 .** They will be able to setup working environments for data engineering work. Understanding of necessary tools to start data engineering tasks. | 1.1. Understanding of using different tools and instruments for starting any particular work.  1.2. Understanding of markdown syntax and general workflow in this course |
| **LO2 (functional).** Using python programming language as a main tool for data manipulations and understanding of different file formats. | 2.1. Understanding of different file formats and how to work with them.  2.2. Scripts to read/write different file formats using python programming language |
| **LO3 (functional).** Main concepts of ETL process. Using pentaho open source ETL instrument and power BI analytical tool for data transformation and visualization. | 3.1. Get familiar with ETL main concepts.  3.2. Get familiar with Pentaho PDI ETL instruments  3.3 Get familiar with Power BI instrument |
| **LO4 (systematic).** Be familiar with the Enterprise Data Warehousing system and main architecture. | 4.1. Understanding of EDW architecture.  4.2 In premise and cloud, and related technologies to develop |
| **LO5 (systematic).** Using python for processing streaming data. Able to build data pipeline using apache kafka, and data processing tool such as ApacheSpark | 5.1 Understanding of streaming data  5.2 Building data pipeline for processing the streaming data |
| **Prerequisites** |  | |
| **Post requisites** | Deep Learning, Data Engineering, Cloud computing, EDW development, Data Analytics, Data Science | |
| **Information resources \*\*** | 1. Mark Lutz. Python Pocket Reference, 5th Edition: Python in Your Pocket, 2014  2. Germán Cocca - Command Line for Beginners – How to Use the Terminal Like a Pro [Full Handbook], 2022 - https://www.freecodecamp.org/news/command-line-for-beginners/  3. Jamie Juviler. What Is GitHub? (And What Is It Used For?), 2021 - https://blog.hubspot.com/website/what-is-github-used-for  4. Rahul Arun. What is Trello and How To Use It?, 2022 - https://www.simplilearn.com/tutorials/project-management-tutorial/what-is-trello  5. Clement Verna. Use HackMD to collaborate on open source projects, 2019 - ​​https://opensource.com/article/19/7/enable-collaboration-hackmd  6. Markdown tutorials - https://www.w3schools.io/file/markdown-introduction/  7. Python tutorials - https://www.w3schools.com/python/python\_intro.asp  8. Julien Kervizic. Files formats for Data Engineers — (Part 1) — Standards Data Formats, 2022 - https://medium.com/analytics-and-data/files-formats-for-data-engineers-part-1-standards-data-formats-98b48dc0bdfc#:~:text=Data%20engineers%20tend%20to%20deal,for%20the%20exchange%20of%20data  9. Postgres official documentation - https://www.postgresql.org/docs/13/index.html  10. Joshua D. Drake, John C. Worsley. Practical PostgreSQL, 2001-2002. http://www.faqs.org/docs/ppbook/book1.htm  11. IBM Cloud Education. ETL (Extract, Transform, Load), 2020 - https://www.ibm.com/cloud/learn/etl  12. Tutorial: Get started creating in the Power BI service, 2022 - https://learn.microsoft.com/en-us/power-bi/fundamentals/service-get-started | |

|  |  |
| --- | --- |
| **Academic policy of the course in the context of university moral and ethical values** | **Academic Behavior Rules:**  All students are required to register for the MOOC. The deadlines for completing the modules of the online course must be strictly observed in accordance with the schedule for studying the discipline. Leave in case of current MOOC or SPOC courses.  **ATTENTION!** Failure to meet deadlines results in loss of points! The deadline for each task is indicated in the calendar (schedule) for the implementation of the content of the training course, as well as in the MOOC. Leave in case of current MOOC or SPOC courses.  **Academic values:**  - Practical trainings/laboratories, IWS should be independent, creative.  - Plagiarism, forgery, cheating at all stages of control are unacceptable.  - Students with disabilities can receive counseling at e-mail \*\*\*\*\*\*\*@gmail.com. |
| **Evaluation and attestation policy** | **Criteria-based evaluation:**  assessment of learning outcomes in relation to descriptors (verification of the formation of competencies in midterm control and exams).  **Summative evaluation:** assessment of work activity in an audience (at a webinar); assessment of the completed task. |

**CALENDAR (SCHEDULE) THE IMPLEMENTATION OF THE COURSE CONTENT:**

|  |  |  |  |
| --- | --- | --- | --- |
| week | Topic name | Number of hours | Max.  score |
| 1 | **Lec 1.** Setup environment | 1 |  |
| **Lab 1.** Markdown, hack.md, trello, terminal basic commands, github | 2 | 6 |
| 2 | **Lec 2.** Setup environment | 1 |  |
| **Lab 2.** Markdown, hack.md, trello, terminal basic commands, github | 2 | 6 |
| **IWST 1.** Writing documentation in markdown and basic commands using terminal | 1 |  |
| 3 | **Lec 3.** Setup environment | 1 |  |
| **Lab 3.** Markdown, hack.md, trello, terminal basic commands, github | 2 | 6 |
| 4 | **Lec 4.** File structure, and python parsing | 1 |  |
| **Lab 4.** Working with json, csv, xml, xlsx using python programming language | 2 | 6 |
| 5 | **Lec 5.** File structure, and python parsingRule Learning Heuristics | 1 |  |
| **Lab 5.** Working with json, csv, xml, xlsx using python programming language | 2 | 6 |
| **IWS 1.** Developing code for file reading and writing and manipulating | 1 | 30 |
| 6 | **Lec 6.** File structure, and python parsing | 1 |  |
| **Lab 6.** Working with json, csv, xml, xlsx using python programming language | 2 | 7 |
| **IWST 2.** Developing concept of functional programming | 1 |  |
| 7 | **Lec 7.** ETL and BI | 1 |  |
| **Lab 7.** Installing Pentaho PDI and creating jobs and transformations. Using Power BI for data visualization. | 2 | 8 |
| **IWST 3. Colloquium (Oral).** | 1 | 25 |
|  | **LEVEL CONTROL 1** |  | **100** |
|  |  |  |  |
| 8 | **Lec 8.** ETL and BI | 1 |  |
| **Lab 8.** Installing Pentaho PDI and creating jobs and transformations. Using Power BI for data visualization. | 2 | 5 |
| **IWST 4.** Developing first ETL process in Pentaho PDI | 1 |  |
| 9 | **Lec 9.** ETL and BI | 1 |  |
| **Lab 9.** Installing Pentaho PDI and creating jobs and transformations. Using Power BI for data visualization. | 2 | 5 |
| 10 | **Lec 10** Introduction to Enterprise Data Warehousing. | 1 |  |
| **Lab 10.** Understanding different architecture and technologies for building EDW | 2 | 5 |
| **IWS 2.** Understanding of technologies and tools | **1** | **30** |
| 11 | **Lec 11** Introduction to Enterprise Data Warehousing. | 1 |  |
| **Lab 11.** Understanding different architecture and technologies for building EDW | 2 | 5 |
| 12 | **Lec 12** Introduction to Enterprise Data Warehousing. | 1 |  |
| **Lab 12.** Understanding different architecture and technologies for building EDW | 2 | 5 |
| **IWST 5.** EDW on premise and cloud | 1 |  |
| 13 | **Lec 13** Data pipeline for steaming data | 1 |  |
| **Lab 13.** Apache Kafka for data pipeline | 2 | 5 |
| 14 | **Lec 14** Data pipeline for steaming data | 1 |  |
| **Lab 14.** Apache Kafka for data pipeline | 2 | 5 |
| 15 | **Lec 15** Data pipeline for steaming data | 1 |  |
| **Lab 15.** ApacheSpark concepts | 2 | 5 |
| **IWST 6. Colloquium (oral)** | 1 | 30 |
|  | **LEVEL CONTROL 2** |  | **100** |

Dean Urmashev B.

Head of Department of

Artificial intelligent and Big Data Mansurova M.

Lecturer Zhienbayev M.

**NOTE:**

The total volume of the syllabus is no more than 5 pages, font 10, Times New Roman

\* LO is based on cognitive (1-2), functional (2-3), systemic (1-2) competencies, total 4-7.

The types and number of competencies (out of 5) are compiled according to the level of education.

\*\* Give no more than 5-7 sources of literature (full bibliographic description), in depth for the last 10 years. (in exceptional cases, 20-30% of irreplaceable classical textbooks), for natural directions - 10 years. Humanitarian direction -5 years

Literature and resources:

1. Basic literature

2. Additional reading

3. Software

4. Internet resources

5. Professional databases

\*\*\*Spreading the assessment of students' knowledge is at the discretion of the compilers of the syllabus.