Опис стручног курса
Дата инжењеринг
I семестар

1) Циљ курса

Дата инжењеринг курс је састављен од 8 различитих модула.

Основни дата концепти

Data warehouse концепти

Cloud концепти

ETL/ELT (Azure Data Factory, MS SSIS)

Snowflake концепти

Databricks концепти

AirFlow концепти

PowerBI концепти

Његов циљ је приближити студентима дата област и тематику кроз дата пројекте и дата алате. Упознати студенте са различитим алатима за управљање подацима и њихову примену.

2) Очекивана предзнања

Базе података, SQL језик

3) Технологије

SQL, Microsoft SQL Server, Microsoft Server Integration Services, Azure Data Factory, AirFlow, Docker, Pandas, Python, PySpark, Databricks, Snowflake, PowerBI

4) Теме курса

Основни дата концепти

- Шта је Дата? Дефиниција и појашњење.
- Дата Формати

- Чување података (*File Stores, Databases*)
- Датотеке (CSV, JSON, XML, BLOB)
- Базе података (Релационе, не-релационе)
- Трансакционо процесирање података (OLTP)
- Аналитицко процесирање података (OLAP)
- Језера података (Data Lake)
- Обрада података (Batching vs. Streaming)

Data warehouse концепти

- Модели података
- Димензионо моделовање
- Дефиниција и појашњење DWH
- OLTP VS OLAP
- Data warehouse могућности
- OLAP коцке, димензије, мере, типови (Rolap, Molap, Holap))
- DWH шема, типови (Star, Snowflake)
- INMNON/KIMBALL приступ

Cloud концепти

- Cloud koncepti
    - Cloud servisi (SaaS, PaaS, IaaS)
    - Tipovi Cloud-a (Public, Hybrid, Private)
    - AWS, Azure, Google Cloud (шта који нуди, који користи које сервисе)
- Azure Cloud
    - Data Storage
        - Azure Data Lake Storage
        - Azure Blob Storage
        - Azure CosmosDB
- Orchestration
    - Azure Data Factory
    - Azure Logic Apps
- Data Capture
    - Azure Event Hub
    - Azure IoT Hub
- Stream/ Batch processing
    - Azure Data Factory
    - Azure Stream Analytics
    - Azure Databricks
- Analytics Data Store
    - Azure Synapse Analytics
- Analytics&Reporting
    - Azure Analytics Services

- Power BI

***ETL/ELT (Azure Data Factory, Microsoft SQL Server Integration Services*)**

- Увод - шта је и чему служи
- Преглед ETL алата
- ETL vs ELT
- Microsoft SQL Server Integration Services (MS SSIS) упознавање
- Креирање пројекта, конекција и једног пакета
- Променљиве и параметризација
- Оркестрација пакета, тестирање пакета
- Azure Data Factory
- SSIS vs Azure Data Factory
- Azure Data Factory упознавање

Snowflake концепти

- *Snowflake* концепти (оснивачи, историјат, интерфејс, верзије, специфичности и предности у односу на остале сличне алате)
- *Snowflake* архитектура (компоненте, виртуелна складишта и микропартиције, модели наплаћивања, роле)
- Учитавање података у Snowflake (опције учитавања, улитавање полуструктуираних података, *Stage* објекат, тарнсформације, Snwopipe)
- Оптимизација перформанси (скалирање, кеширање, кластеровање)
- Напредне функције (*time travel, zero copy cloning, data sampling, data masking, materialized views*)

DataBricks концепти

- Увод у Databricks
- Складишта података
- Delta Live табеле
- Окружење
- Учитавање фајлова са различитих извора података
- Python функције
- Креирање шеме
- Delta табеле
- Учитавање података у Delta Lake
- Delta Live табеле
- Event Log

AirFlow концепти

- Увод у Airflow
- Архитектура

- Алтернативни/модерни алати за оркестрацију/интеграцију података
- Docker – софтверски контејнери и Windows инсталација
- Airflow инсталација
- Интерфејс окружење
- Креирање структуре фолдера
- Основни пример И пуштање једног DAG-а
- Pandas библиотека за анализу података
- Рад са csv подацима (локално)
- Издвајање
- Трансформација
- Учитивање
- Употреба на комерцијалним пројектима

PowerBI концепти

- Увод
    - Увод у Power BI (Power BI services и Power BI desktop): чему служи алат, основна подешавања.
    - Преглед Power BI desktop алата: даје се општи преглед са импортом excel табеле.
- Прављење извештаја
    - Креирање табела у Power BI
    - Стилови табела и форматирање
    - Матрична визуелизација
    - Промена метода агрегације
    - Карте (cards) и карте са више редова
    - Калкулације процената и рад са датумским типовима података
    - Филтрација података помоћу слицер-а
    - Графови
- DAX Формуле
    - DAX калкулисане колоне
    - Датумске функције
- DAX Мере
    - Увод у DAX мере
- Везе (Relationships)
    - Креирање и управљање везама у Power BI
- Power BI Query Editor
    - Основне трансформације

5) Литература

Azure

- https://learn.microsoft.com/en-us/training/azure/

DWH

- *Building the Data Warehouse , W. H. Inmon (https://ia800202.us.archive.org/9/items/2005BuildingTheDataWarehouse4thEditionWilliamH.Inmon/2005%20-%20Building%20The%20Data%20Warehouse%20%284th%20Edition%29%20%28William%20H.%20Inmon%29.pdf)*
- *The Data Warehouse Toolkit, Ralph Kimball, Margy Ross*

Snowflake

- https://docs.snowflake.com/

Databricks

- Databricks Academy for Partners (https://partner-academy.databricks.com)
- Data Engineer Learning Path (https://github.com/databricks-academy/data-engineer-learning-path/)
- Databricks Community Edition (https://www.databricks.com/product/faq/community-edition)
- Data modelling (https://www.databricks.com/glossary/medallion-architecture
  https://www.databricks.com/blog/2022/06/24/data-warehousing-modeling-techniques-and-their-implementation-on-the-databricks-lakehouse-platform.html
  https://www.databricks.com/blog/2022/10/20/data-modeling-best-practices-implementation-modern-lakehouse.html
  https://www.databricks.com/blog/2022/06/24/prescriptive-guidance-for-implementing-a-data-vault-model-on-the-databricks-lakehouse-platform.html

Airflow

- Airflow:
  - What Is Airflow (https://datascientest.com/en/apache-airflow-what-is-it)
  - Documentation (https://airflow.apache.org/docs/apache-airflow/stable/index.html)

- Алтернативни алати
  - Prefect documentation (https://docs.prefect.io/2.11.5/)
  - Kestra documentation (https://kestra.io/docs)
  - Airflow vs Prefect vs Kestra (https://medium.com/python-in-plain-english/airflow-vs-prefect-vs-kestra-what-is-the-best-data-orchestration-platform-in-2023-899d849743cc)

- Docker
  - Documentation (https://docs.docker.com/)
  - Softverski kontejneri (https://www.helloworld.rs/blog/Sta-je-Docker/69)

- Pandas
  - What Is Pandas In Python (https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/)

PowerBI

- Онлине курс на Udemy: “Complete Introduction to Microsoft Power BI [2023 Edition]“ by Ian Littlejohn: https://www.udemy.com/course/powerful-reports-and-dashboards-with-microsoft-powerbi/
- Списак бесплатних курсева: https://medium.com/javarevisited/10-free-microsoft-power-bi-courses-for-beginners-19ee524008e1