Udacity Nanodegree - Data Engineering

Darjan Salaj
Dominik Schüssele
Patrick Ruoff

Karlsruhe, 08. Juli 2020

Agenda

- Big Picture
- Syllabus / Course - Structure
- Projects
- Conclusion

Big Picture

Summary

1. Data Modeling
2. Cloud Data Warehouses
3. Spark & Data Lakes
4. Data Pipelines with Airflow
5. Capstone Project

- 248 / 359€ per month
- 5 months, 5-10 hours / week
- video lectures, quizzes, jupyter notebooks, exercises, project reviews, references
- prerequisites: “intermediate Python & SQL” - skills

- real-world projects from industry experts
- technical mentor support
- flexible learning plan
- personal career coach & career services
- student community

https://www.udacity.com/course/data-engineer-nanodegree--nd027

Syllabus / Course - Structure

Data Modeling

relational data models (SQL)

- OLAP vs. OLTP
- normalization
- star / snowflake schemas
- upserts

NoSQL data models

- cap theorem
- eventual consistency
- denormalization
- clustering columns / composite key

Cloud Data Warehouses

- business vs. technical perspective
- data warehouse architectures
- dimensional modeling & ETL
- OLAP cubes
  - slice & dice
  - roll up & drill down
  - grouping sets

- basics cloud computing & AWS
- infrastructure as code on AWS
- table design / ETL / AWS Redshift

Spark & Data Lakes

- big data & hadoop basics
- spark-cluster & use-cases
- data wrangling w/ spark
- spark on AWS
- debugging & optimization
- data lake concepts
- data lake vs. data warehouse

Data Pipelines w/ Airflow

- introduction to data pipelines
- airflow basics & concepts
- ensuring data quality w/ airflow
- extending airflow
- monitoring

Projects

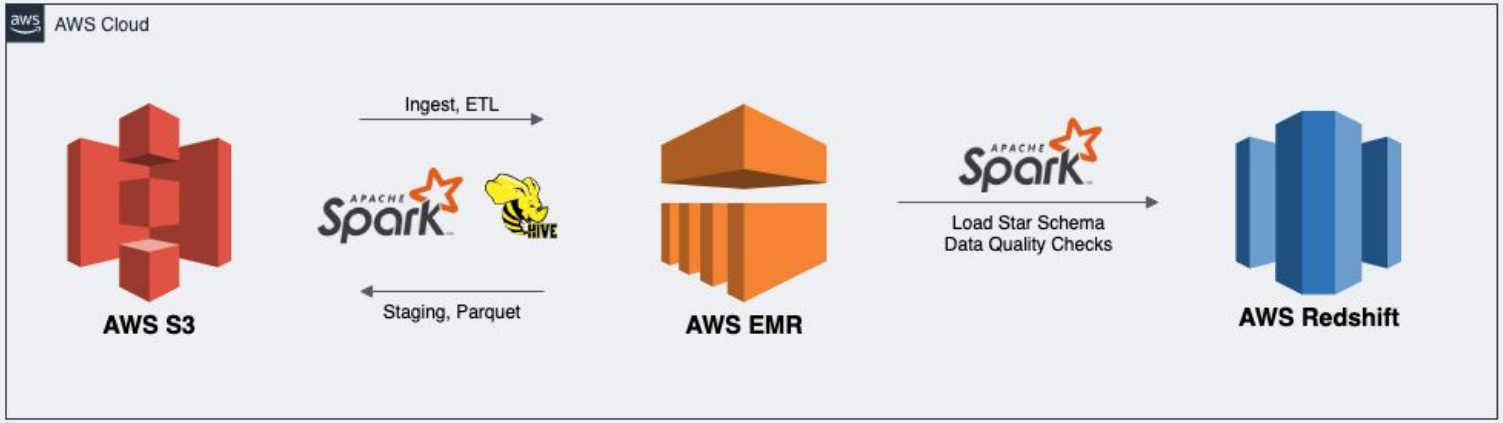
Capstone Project

- combine & demonstrate everything you learned
  - project-rubric
- udacity provided project vs. your own
- additional resources

Guidelines:

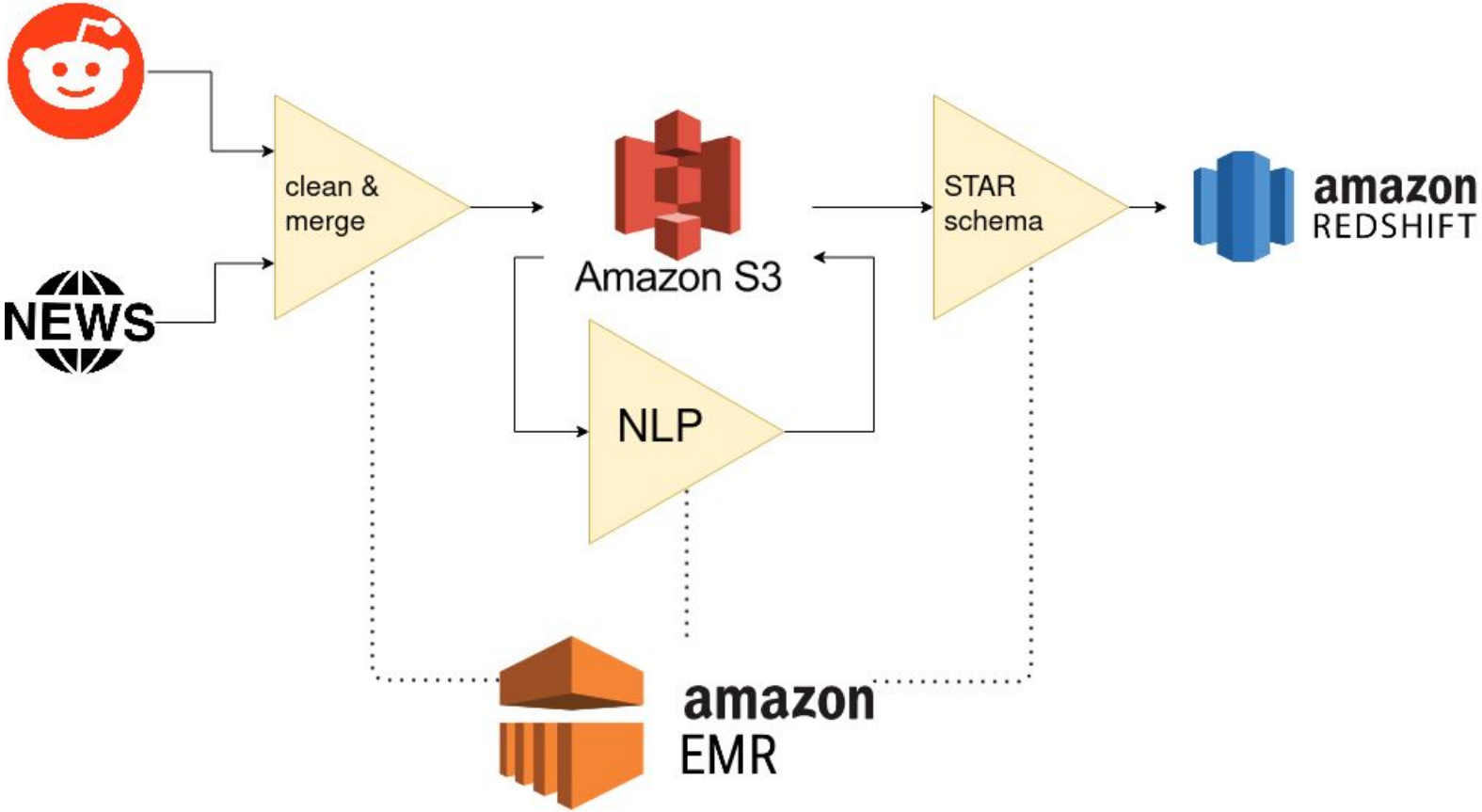
1. define the scope of the project yourself
2. gather & explore data
3. define a data-model
4. build ETL to create a data-warehouse
5. project write-up

What we built!



inovex

Simple trends detection



inovex

Conclusion

Pro's

- hands-on
- short, concise, intuitive lectures (2 - 10 minutes)
- small exercises, quizzes in-between & projects at the end of each section force understanding
- core curriculum & extracurricular
- student / mentor community
- career services

Con's

- some course - sections were not well structured
- technical project workspaces / infrastructure (sometimes) unreliable
- often only core-concepts & not very in-depth
- outdated instructions for AWS
- big data technologies on small data sets

Conclusion

- relevant technologies
- flexible and diverse
- high quality material & workspace
- (mostly) professional feedback
- community / mentoring
- could’ve covered more details - depends on your skill-level!
- read syllabus and decide individually!

Conclusion

For beginners:

- SQL / NoSQL refresher
- intro to AWS DevOps
- challenging Capstone Project
- one superficial review
- later lessons of lower quality

Vielen Dank

Darjan Salaj
Dominik Schüssele
Patrick Ruoff

inovex GmbH
Ludwig-Erhard-Allee 6
76131 Karlsruhe