
Ab Initio

Course assignments

A real customer use case
Digital transformation in a large organization

The use case of this assignment

- In this assignment I want to present you a real, recent customer use case
- The customer is a large organization with > 100 million customers
- The customer has (unsuccessfully) tried for a long time to efficiently make operational data available to
  - Data scientists for machine learning
  - Business analysts for analysis and reporting
  - Other internal data consumers
- The vision of the management is to create a data self service environment
  - All operational data should be available in a logical data model
  - People interested in receiving data should be able to “Amazon”-style search for and order data
  - Interested parties subscribe to data from a data catalog and get data delivered at a pre-defined cycle

The goal of this assignment

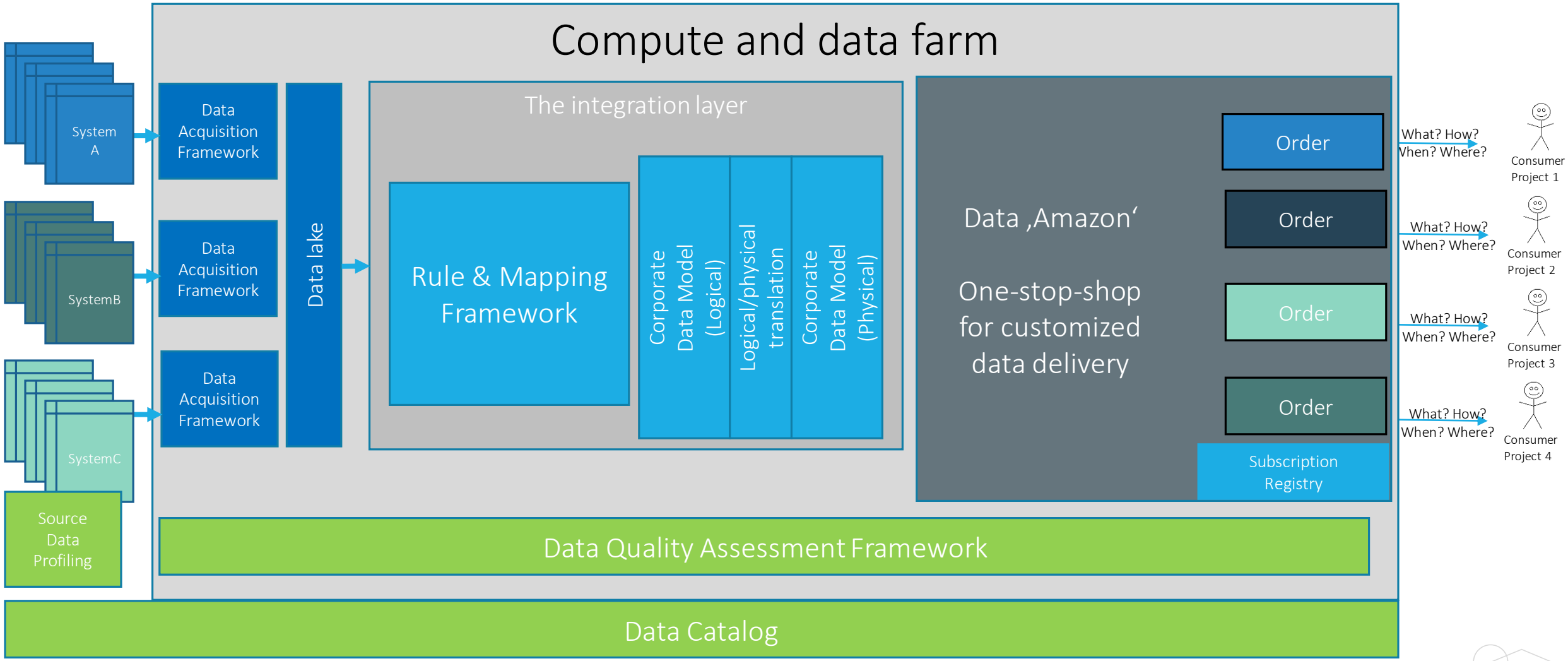
- I want you to analyze the use case from various angles and present a solution that you can envision
  - I help you by giving you buzzwords/hints that you should investigate and consider
  - I would like you to consider important implications of different central roles of each project.
    These roles exist, they are equally important for the success of the project
  - Analyze the situation presented to you, and present your analysis and solution in presentations of 20-30 minutes
- I propose five teams of 2-3 persons each.
  Pick your favorite, all teams are equally important
  - The project management team
  - The data replication architecture team
  - The data integration architecture team
  - The subscription architecture team
  - The data governance team
  - The operations team
- There is no right or wrong solution! Ask me (email at end) if you have any questions about the task.
  But you should work as a team. Within your team and across teams.

What you should consider in each group

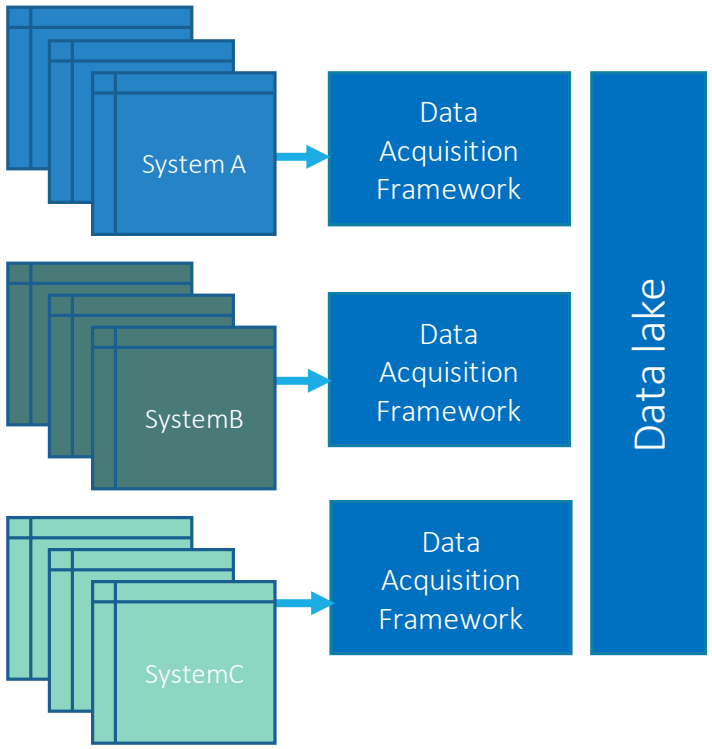
- You will have to do some research on your own. Start by googling the hints I give you.
- There is no right or wrong here! This is not an exam and you will not be judged!
  Just think about a solution of your task, be creative.
- What (open-source) technologies could be applied to solve the task?

  Of course you could do everything with Ab Initio 
- Which technologies would you pick and why?
- What issues or problems can you imagine?
  Think of data volumes, think of changing/unclear requirements, think of unclean or old data
- Make sure that your solution works with the solution designed by the next team.
- Present your solution, explain why you made your choices
- I normally estimate that I talk three minutes per slide
  So you'd need roughly ten slides for the presentation
  Remember: a well-explained picture is better than 1000 words
- Be prepared to answer questions from the other students after the presentation

The self-service data supermarket



The data replication team



- The customer has *many* operational systems which produce *lots* of data
- Sometimes data comes as full files, sometimes as delta files
- All data gets replicated and historized in a “data lake”
- The data lake must be accessible via Hive/SQL for other systems to consume and analyze
- There are lots (dozens) of source systems, so the solution must be a framework

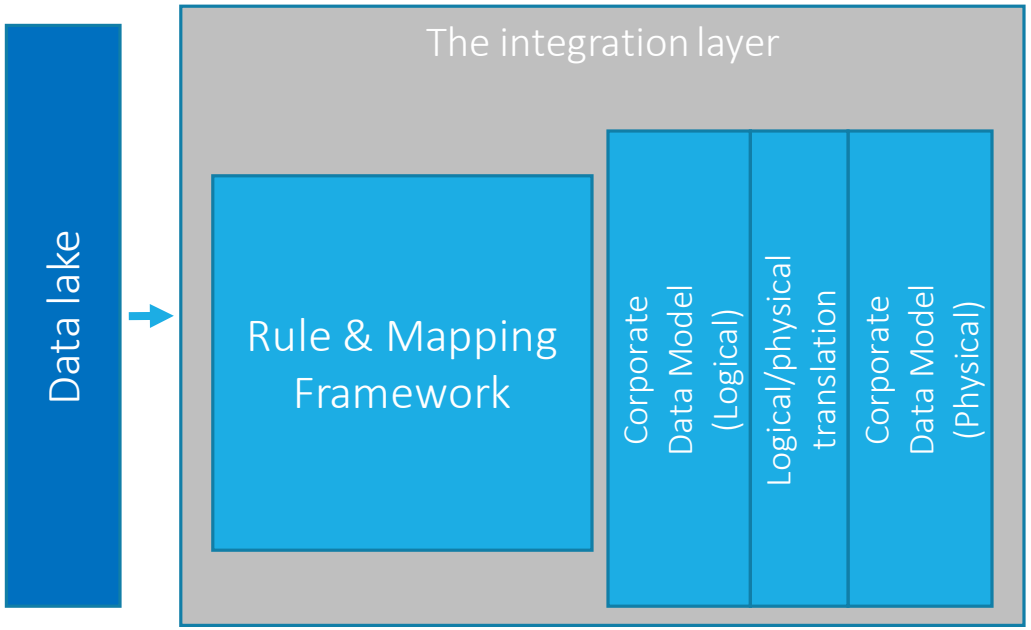
The data replication team
Issues to consider and to answer

- Estimate or guess, how many departments, how many different “tables” and what data volume you have to consider. Roughly…
- Data is replicated into a “Data Lake” including data history.
- Your first task is to replicate the data into the lake.
  Describe in short how you would do this.
  Individual applications? Reusable frameworks? Why?
- What challenges do you face inside the data lake (Volume? Data swamp?)
  Can you estimate the size of the lake? Can you guess?
- What if data comes a small “delta” files just with changes.
  Why is HDFS bad in this case?

Some useful buzzwords to help you

- What are full/snapshot files, what are delta files?
- Change Data Capture from the sources
  What is it? Why is it useful?
- What is Apache Hudi and why could it be useful?
- What is DualTable (on arxiv.org) and why could it be helpful?
- What are Parquet files, what are its benefits and disadvantages here?

The data integration team



- You have raw, historized source data inside your data lake.
  These are just raw copies of your source data!
- You must curate/cleanse it and bring it into a integrated corporate data model.
- First you must map the data into a logical data model.
  The logical data model is clean third normal form (3NF) data model
- Then you must map it into a physical data model which may depend on the database technology you use.

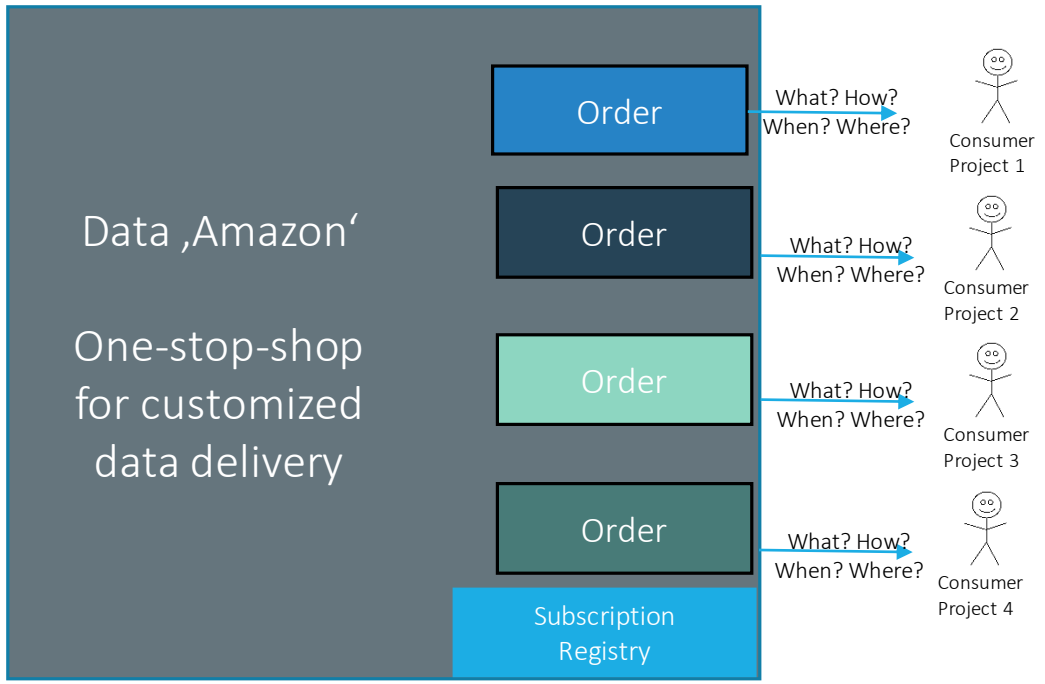
The data integration team
Issues to consider and to answer

- There are dozens of source systems, with hundreds or thousands of tables in your lake.
  Who should build the mapping rules and logic to transform the data from source format into the corporate data model? IT department? If yes, why? If not, who else?
- What frameworks or products do you know or can you find to do the source-to-target mappings from your data lake to the corporate data model?
- Are they useful for IT? Are they useful for non-technical business analysts?
- Can they support your ultra high volumes?
- The CTO wants that every data access only happens against the logical data model. Can you envision the technical consequences of this?

Some useful buzzwords to help you

- What is ETL, what is source-to-target mapping?
- What is a logical data model, what is a physical data model?
- What is a data vault?
- Can you find open-source products that help with visual ETL/source-to-target mappings?

The subscription architecture team



- This part is really hard
- The CTO wants that non-technical users can subscribe to data.
  They browse data in a “Data Catalog” (like “Credit Card Transactions”), specify how data should be transformed, and when it should be delivered.
  Like:
  - I want all new credit card contracts from the day before, but filtered by by location
  - I want it delivered every day by 6:00
- The rules to retrieve the data could be simple (one-to-one copy of the data) or complex (a dimensional data mart)
- Also, subscriptions are dynamic, go into production, come out of production, at any time.

The subscription team
Issues to consider and answer

- The subscription could be a trivial one-to-one copy, or a complex data mart. Can you find a technology that can supports both, very simply data copies and very complex processing?
- The subscription should be available to non-technical users. Can you imagine a simple solution for non-technical users, a complex solution for the technical experts?
- What technologies or languages could you see to implement this?
- How would you execute this? One big server which knows and runs all subscriptions? Could you run this in containers?
- The CTO wants that all subscriptions only the logical data model, never physical fields? Can you imagine why, and what the consequences are?

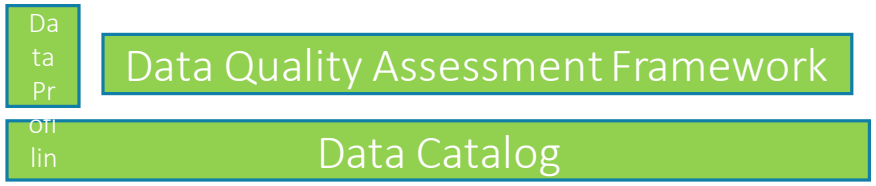
Some useful buzzwords to help you

- Federated queries/federated SQL
  What is it and how could it help?
- Dimensional data model
  What is it and how is it used?
- One-to-one mapping
- Docker and Kubernetes
  What are they and why could it be useful?
  (Imagine running a subscription inside)
- When you want to do machine learning, which data would you want? Raw data? The corporate data model?

The data governance team

- The data flow must be controlled in several aspects
  - All data flowing through the system must be cataloged
  - All data in the corporate data model must be cleansed
  - Source systems must be able to check the quality of their own data, so they do not send bad data

- Your task will be to build a data catalog for data governance and data quality



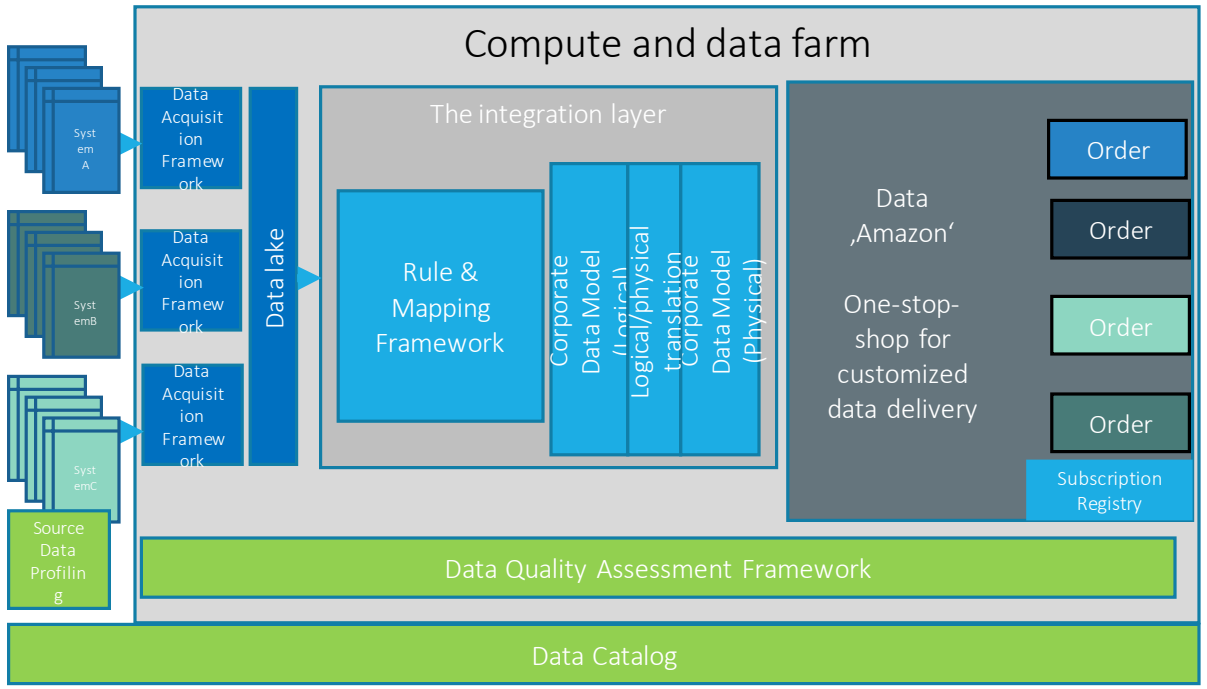
The data governance team
Issues to consider and answer

- What data quality measures can you define?
- What would be technical data quality, what semantic data quality?
- What is “personally identifiying information”, how would you protect it and where would you store information about it?
- What is a business data glossary and why would you need it? Should subscribers search for data in the business glossary or in technical names?
- When doing machine learning, do you want the business glossary, the data from the corporate data model, or the raw data?

Some useful buzzwords to help you

- What methods for statistical data profiling can you find?
- What is the difference between machine learning and data profiling?
- What is semantic discovery and why is it useful?
- When would a user search in the data glossary, when in the raw data? Why?

The project management team



- Your role is to organize the design and development of this system
- You must organize teams, provide time lines for the project and set deliverables.
- Think about roles and responsabilities
- Think about how teams must interact and why.
- How should they communicate?

The project management team
Issues to consider and answer

- What different project management techniques do you know? Which one would you choose?
- Of course, Scrum is one of them.
  How would you organize the teams?
- Explain on a high level Scrum, Kanban and Extreme Programming
- Name and explain the typical stakeholders in a software development project. What are their roles and expectations?
- Explain some methods to estimate development efforts? What if you have to deliver a data warehouse with hundreds almost identical programs?
- (Optional: construct and show a demo of Atlassia Jira and show some user stories, sprints, etc.)

Some useful buzzwords to help you

- Scrum, sprint, product backlog, sprint backlog
- What is scope creep and why is it dangerous?
- What is function point analysis, story points, burn down chart?
- Sponsor, customer, development team, …

The operations team
Issues to consider and answer

You are running the application every day in production.

Your task is to run and protect the production environment.

You must ensure that all processes and jobs running in production will not disrupt the production process. The health of the enterprise depends on you.

Your task is to build a Jenkins pipeline, that can

a) Do some automated system tests upon check-in of your code into a source code repository

b) Show the results of the tests inside Jenkins

c) Set up some example that migrates your code from a fictitious development to test environment

You are completely free in the choice of programming language, source code repository, and kind of application. Just demonstrate how it works

Some useful buzzwords to help you

- Atlassian Jenkins, Projects and Pipelines
- CVS, Git,

Thank you

Dr. Peter Ossadnik

+49 174 248 5629

possadnik@abinitio.com

www.abinitio.com