Merck Process Hierarchy

Capstone Spring Semester 2024

Lauren Etsitty, Aiden Astle, Tarif Ahmed, Joseph Sanchez

Business Background

- Merck
  - Pharmaceutical company established in 1891.
  - Headquarters located in Rahway, New Jersey.
  - Specializes in the production of medicines, vaccines, biologic therapies, and animal health products.
  - Known for advancements in cancer immunotherapy, anti-diabetic medications, and vaccines against HPV and chickenpox.

Overall Problem and Solution

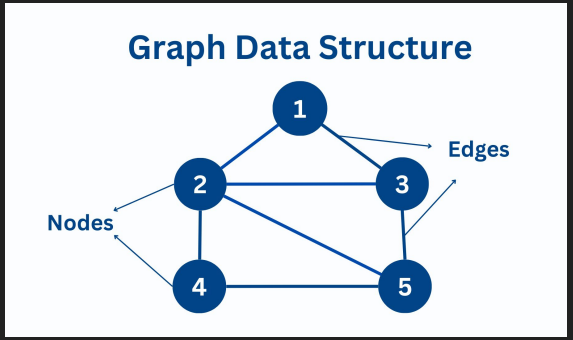
- Extracting relevant data from a data lake is time consuming and inefficient
- Data currently stored in Excel and SQL database model
- We would like to explore different data models
- Improve the performance of our hierarchy modeling and analytics efforts.

Project Solution

- Accessing data quickly & efficiently. Creating an efficient database, likely graph-based architecture.

- The solution includes a Go-based backend and a Next.js/React frontend.

- The goal is to streamline data access for scientists and engineers.

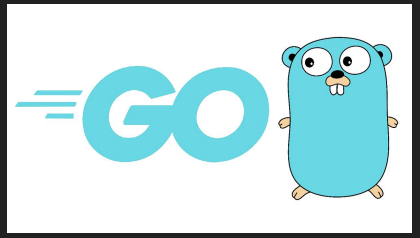


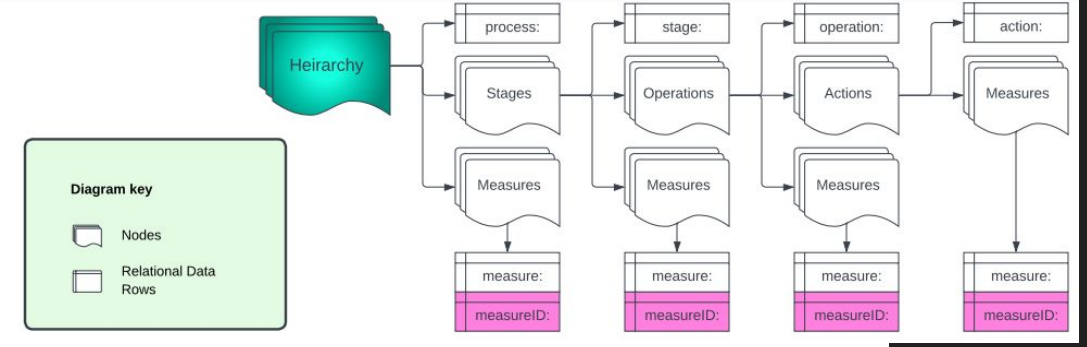
Project Management

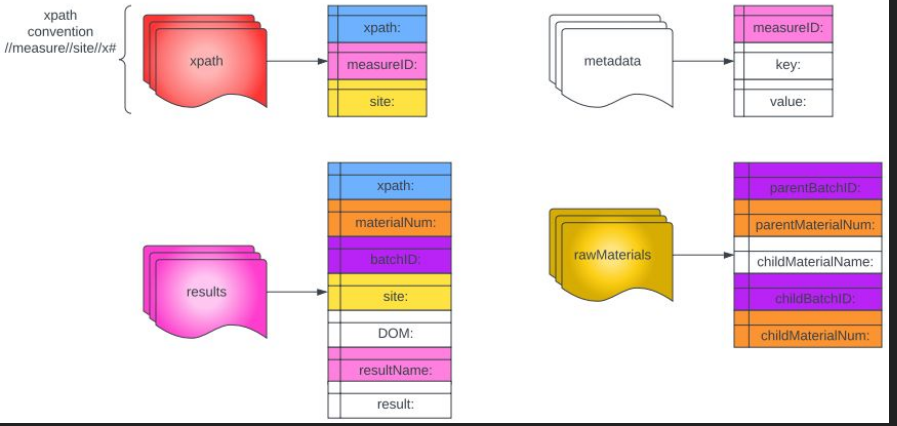
- Scrum methodology
- Weekly Sprints
  - Weekly Meetings
  - Sprint Report
  - Retrospective
  - Sponsor Meeting preparation
  - Meeting Notes
  - Record Sponsor Meetings
- Work progress represented through Jira
- Documentation organized through Confluence
- Communication through Slack/Text

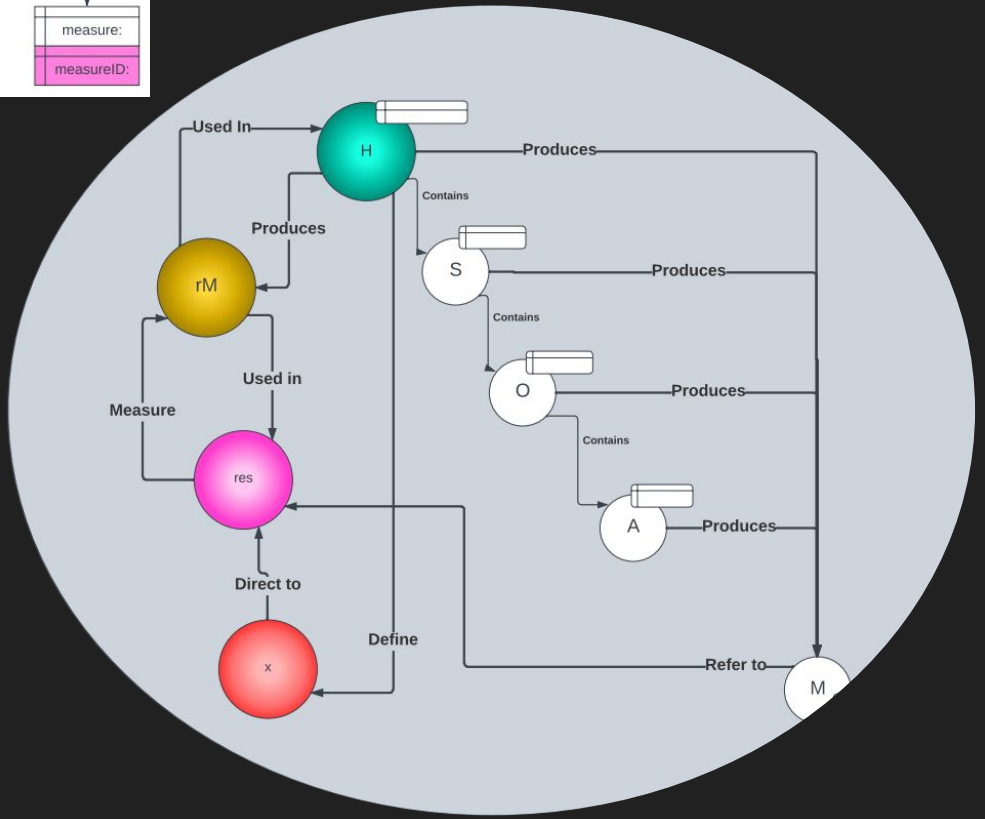
Current Progress

- Learning fundamentals of Go
- Creating backend databases in noSQL, Amazon Neptune, Neo4j, and Dynamodb
  - Gain familiarity
  - Test Programs
- Determining the pros and cons of each database
- Creating a more complex database implementing the following hierarchy provided by Merck
  - https://cse280capstone2024.atlassian.net/wiki/spaces/OS/pages/7929857/Graph+DB+Diagram




Heirarchy
process:
stage:
operation:
action:
Stages
Operations
Actions
Measures
Measures
Measures
Measures
measure:
measureID:
measure:
measureID:
measure:
measureID:
measure:
measureID:
Diagram key
Nodes
Relational Data Rows


xpath convention //measure//site//x#
xpath
xpath:
measureID:
site:
metadata
measureID:
key:
value:
results
xpath:
materialNum:
batchID:
site:
DOM:
resultName:
result:
rawMaterials
parentBatchID:
parentMaterialNum:
childMaterialName:
childBatchID:
childMaterialNum:


H
rM
S
O
A
res
x
M
Used In
Produces
Contains
Produces
Produces
Contains
Used in
Measure
Produces
Contains
Produces
Direct to
Define
Refer to

Team Roles

Aiden Astle

- Neo4j

Tarif Ahmed

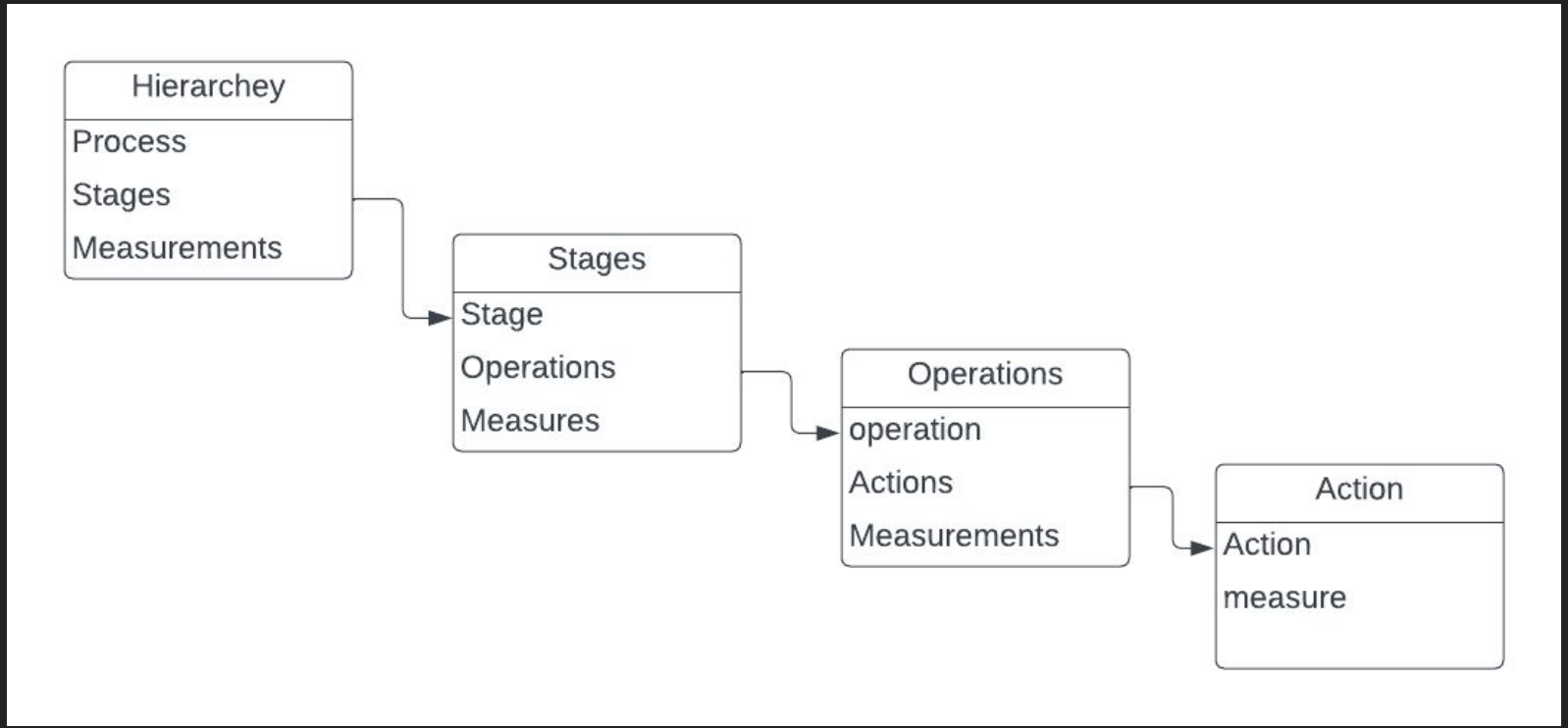
- PostgreSQL

Joseph Sanchez

- NeptuneDB

Lauren Etsitty - Project Manager

- DynamoDB


Hierarchey
Process
Stages
Measurements
Stages
Stage
Operations
Measures
Operations
operation
Actions
Measurements
Action
Action
measure

Future Schedule

Sprint 4-5:

1. Complete Go language Tutorials
2. Solidly graph database platform

Sprint 6-8:

1. Running Database with mock data
2. Testing database efficiency

Sprint 9-10:

1. Refine data organization
2. Compare DBs and determine final choice

Sprint 11-13

1. Expedite data retrieval
2. Organize data with statistics

Risks

- Cost
  - Using cloud services
  - Free tier quotas

- Graph data models are more complex than relational tables (PostgreSQL, noSQL)

Thank you