NetApp

Building a Data Lakehouse with PostgreSQL

Dive into Formats, Tools, Techniques, and Strategies

Josef Machytka <josef.machytka@netapp.com>
NetApp Open Source Services / Credativ
2025-01-29 - Prague PostgreSQL Developer Day 2025

© 2025 NetApp, Inc. All rights reserved.

Josef Machytka

- Professional Service Consultant - PostgreSQL specialist at NetApp Open Source Services / Credativ
- 30+ years of experience with different databases.
- PostgreSQL (12y), BigQuery (7y), Oracle (15y), MySQL (12y), Elasticsearch (5y), MS SQL (5y).
- 10+ years of experience with Data Ingestion pipelines, Data Analysis, Data Lake and Data Warehouse
- 2 years of practical experience with different LLMs / AI including their architecture and principles.
- From Czechia, living now 11 years in Berlin.

- LinkedIn: linkedin.com/in/josef-machytka
- ResearchGate.com: researchgate.net/profile/Josef-Machytka
- Academia.edu: netapp.academia.edu/JosefMachytka
- Medium.com: medium.com/@josef.machytka

 © 2025 NetApp, Inc. All rights reserved.

Table of contents

All AI images without credits
were created by the author of this talk
using DeepDreamGenerator

 © 2025 NetApp, Inc. All rights reserved.

What is Data LakeHouse?

- Answer is surprisingly not simple
- Big variety of opinions around this term

- Modern formats like Apache Iceberg, Hudi, Delta Lake
- Object storage with structured and unstructured data
- Data pipelines processing structured and unstructured data
- Mesh of Data Lakes and Data Warehouses in the organization
- Mesh of all existing data sources in the organization
- All of it together with Data Governance
- All of it and AI and ML models

 © 2025 NetApp, Inc. All rights reserved.

Data Formats

Store only what you really need

- Store only data necessary for your operations
- And store them in efficient way

- Avoid storing data just because it ”might be useful in the future”
- Law regulations require Data Retention Policies
- Most companies need in long run only aggregated data
- Some types of raw data can or must be deleted after processing
- You pay for collecting, storing, and processing data
- What about Return Of Investment from this data?

 © 2025 NetApp, Inc. All rights reserved.

Relational Data Warehouses

- 20-30 years ago relational databases dominated
- Mainly proprietary engines: Oracle, DB2, SQL Server
- In new millennia also PostgreSQL
- Engine-specific data storage formats
- Computation and storage were tightly coupled
- Very difficult to scale to more machines
- Almost exclusively row-oriented storage
- All processing done using SQL

 © 2025 NetApp, Inc. All rights reserved.

Early Data Lakes Formats

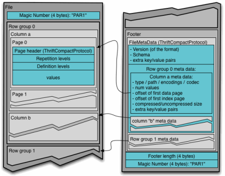
- *JSON* - key-value pairs, supports nested structures
- *Parquet* - compressed columnar storage format, optimized for data analysis
- *Avro* - row oriented, schema-based, binary format
- *ORC* - Optimized Row Columnar - columnar storage format, for read-heavy workloads

- Data in these formats is hard to update - append-only, immutable
- PostgreSQL has FDW for some of these formats or can import them
- New extensions pg_analytics and pg_duckdb aim to allow direct querying of these formats
- FerretDB with DocumentDB extension implements BSON data type and MongoDB wire protocol queries

© 2025 NetApp, Inc. All rights reserved.

Parquet Data Format

- Parquet is a columnar storage format optimized for reading
- Repository: github.com/apache/parquet-format
- Very efficient for numeric data types: INT32, INT64, FLOAT, DOUBLE, BOOLEAN
- Strings are less efficient, stored as BYTE_ARRAY
- Optimized for read-heavy workloads, metadata includes min/max values for columns
- Most popular format, used in all modern Data LakeHouse solutions



(Image from the apache/parquet-format repository)

 © 2025 NetApp, Inc. All rights reserved.

Modern Data LakeHouse Format Frameworks

- Apache Arrow - platform for in-memory analytics, defines columnar data format
- Apache Iceberg - table format for large-scale data systems
- Delta Lake - storage format for Data Lakehouse architecture
- Apache Hudi - transactional data lake framework

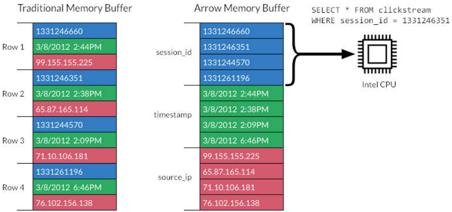
- Designed for managing and processing large data sets
- Optimized for analytical queries and data processing
- Allow limited updates and deletes with ACID transactions

APACHE ARROW ICEBERG DELTA LAKE Apache hudi

 © 2025 NetApp, Inc. All rights reserved.

Apache Arrow - platform for in-memory analytics

- Cross-language platform for in-memory processing of large data sets
- Repository: github.com/apache/arrow
- Standardized, language-independent columnar in-memory format
- Enables zero-copy reads across multiple processes
- Closely integrated with Python for data analysis

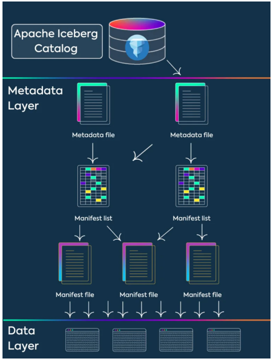


(Image from the article Apache Arrow Overview)

 © 2025 NetApp, Inc. All rights reserved.

Apache Iceberg Table Format

- Originated from Netflix, now an Apache project
- Repository: github.com/apache/iceberg
- Immutable, append-only, and transactional
- Every change creates a new metadata file and snapshot
- Components: catalog, metadata files, manifest files, Parquet data files
- Supports versioning, partitioning, and schema evolution
- Implements time travel to query historical data
- Each snapshot provides full isolation and consistency
- Allows multiple applications to work on the same data

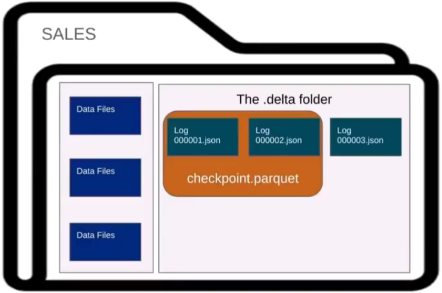


(Image from the article What Is Apache Iceberg?)

 © 2025 NetApp, Inc. All rights reserved.

Delta Lake Storage Format

- Open-source storage format for Data Lakehouse architecture
- Created and maintained by Databricks
- Project page: delta.io
- Transactional storage layer on top of cloud storage
- ACID transactions and scalable metadata handling

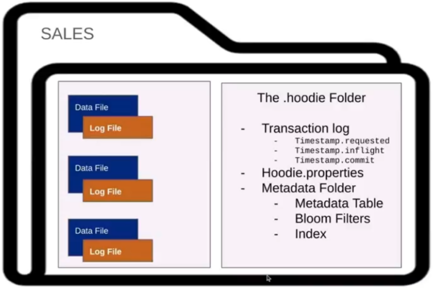


(Image from the LinkedIn course Fundamentals of Apache Iceberg)

 © 2025 NetApp, Inc. All rights reserved.

Apache Hudi

- Originated from Uber, now an Apache project
- Allows multiple updates and deletes
- Changes are stored in a log file
- Brings database and data warehouse features to data lakes
- Project page: hudi.apache.org



(Image from the LinkedIn course Fundamentals of Apache Iceberg)

 © 2025 NetApp, Inc. All rights reserved.

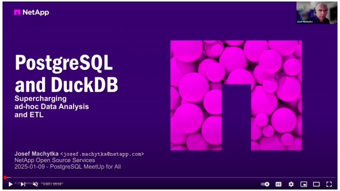
PostgreSQL and Data LakeHouse Formats

- PostgreSQL has columnar storage support for efficient analytics
- Has FDW for CSV, JSON, Parquet, some other formats through JDBC
- Arrow, Iceberg, Delta, and Hudi require more functionality
- Currently closer integration with DuckDB looks very promising
- ParadeDB develops pg_analytics extension
- Goal to provide a unified interface for various data formats & cloud storages
- DuckDB/MotherDuck team work on pg_duckdb extension
- It aims to provide full DuckDB functionality in PostgreSQL

 © 2025 NetApp, Inc. All rights reserved.

DuckDB is a Powerful Analytical Database

- Created at the National Research Institute for Mathematics and Computer Science, Amsterdam
- Open-source, column-oriented, in-memory relational database
- Single-node database, intended for embedding in applications, like SQLite
- Designed for heavy parallel analytical workloads
- Columnar-vectorized query processing engine
- Direct selects from multiple formats and cloud storages
- Extremely portable, runs on all architectures, no dependences



(See details in my talk PostgreSQL and DuckDB on YouTube)

 © 2025 NetApp, Inc. All rights reserved.

Examples of Data Pipelines

One Solution does not fit All Cases

- One solution does not fit all cases

- Classical data warehousing theory emphasized centralization
- Also some marketing articles see Data Lakehouse as centralized
- Companies try sell ”one size fits all” solutions
- But special cases require decentralization

(Image from the article
One Size Does Not Fit All)

 © 2025 NetApp, Inc. All rights reserved.

Examples of Data Pipelines

- *Telecommunication software for events from mobile networks*
- *Widget predicting sizes for online stores selling clothes*
- *Software for secure online logins and financial transactions*

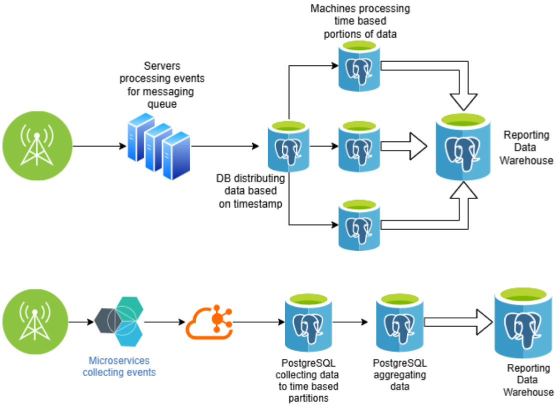
- They all need to calculate output for the clients very quickly
- All companies collect and process a lot data but in different ways
- PostgreSQL heavily used in all these companies
- Multiple PostgreSQL instances in each company

 © 2025 NetApp, Inc. All rights reserved.

Telecommunication software for Events from Mobile Network

- Probes collect events from the mobile network
- Hundreds or tens of hundreds GBs per minute
- Provider needs only aggregated summaries

- Centralized model, storing only aggregated data
- Raw data are discarded after processing
- originally multiple PostgreSQL dbs and PL/proxy
- Later rebuilt with Kafka and quicker hardware

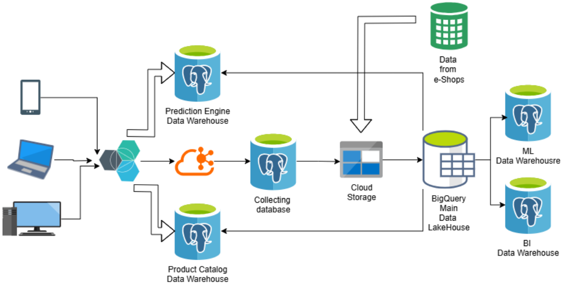


 © 2025 NetApp, Inc. All rights reserved.

Widget predicting sizes for online stores selling clothes

- Calculates the best fit in dozens of milliseconds
- Prediction uses only pre-aggregated data from ML
- Scripts collect events from the website and devices
- Raw data tens or hundreds of GBs per hour

- Mixed model, main Data LakeHouse is BigQuery
- Raw events stored for 2 years for Data Analysis
- Multiple other PostgreSQL instances for other tasks

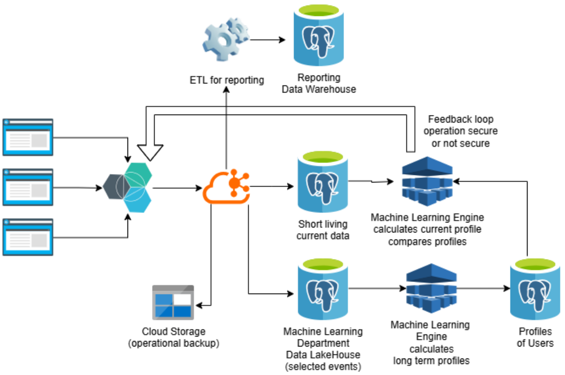


 © 2025 NetApp, Inc. All rights reserved.

Software for secure logins and financial transactions

- Software analyzes behavior of users
- Calculates current behavioral profile
- Compares with stored profiles
- Decides if operation is secure
- Response needed in milliseconds

- Strongly decentralized model
- Storing only aggregated data
- PostgreSQL used multiple times
- Raw data tens or hundreds of GBs per minute
- Discarded soon after processing



 © 2025 NetApp, Inc. All rights reserved.

PostgreSQL as Important Part of Data Pipelines

- PostgreSQL can play multiple roles in Data Pipelines
- Multiple features and extensions for different tasks
- Very powerful partitioning, pg_partman
- Row level security for fine-grained access control
- JSONB implementation and multiple types of indexes
- Multiple FDWs for different databases and file formats
- Citus - distributed PostgreSQL, columnar format
- TimescaleDB - sharding and columnar format
- Hydra and OrioleDB (beta) aim to improve analytical performance
- PostGIS for geospatial data
- Powerful build in full text search features, pg_search

 © 2025 NetApp, Inc. All rights reserved.

Data Governance & Legal Aspects

Discipline is essential

- ”Discipline is essential, you foolish lads!”
  Without it, you would be climbing trees like monkeys!”

- Every Data Lakehouse requires clear Data Governance
- Properly defined Data Life Cycles are crucial
- Clear Data Catalog and Lineage are necessary
- Without these, we can ”climb” every new technology
- Jumping ”like monkeys” from one new buzzword to another
- But we will end up in the same mess...

(Image by Mikoláš Aleš
from the book
”The Good Soldier Švejk”)

 © 2025 NetApp, Inc. All rights reserved.

Do not be overwhelmed by Data Governance

- Abundance of articles and books on Data Governance
- Majority of them overly maximalistic and complex
- Marketing articles always sell something
- Some proprietary solution, or consulting services

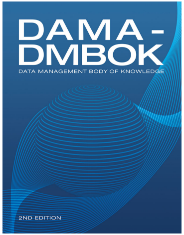
- Big companies really need complex Data Governance
- But smaller companies can start with simple rules

(Picture from article Data Governance)

 © 2025 NetApp, Inc. All rights reserved.

DAMA - Data Management Body of Knowledge

- DAMA International is a non-profit organization
- Website: data.org
- DAMA-DMBOK2 is a comprehensive guide to Data Management
- It provides a very detailed view of Data Management



(Screenshots from dama.org page)

 © 2025 NetApp, Inc. All rights reserved.

Business will suffer with poor quality data

- If data belong to no one, no one will care about quality
- If no one checks data quality, it will be poor

- **Data Quality** - garbage in, garbage out
- To know what is garbage, you need to know what is good
- Basic **Data Catalog** and **Data Definitions** are needed
- **Quality Checks** and **Data Profiling** based on Data Catalog
- **Data Producers/ Owners/ Stewards** responsible for data quality

(Title page of
EU Commission Data Governance
Document)

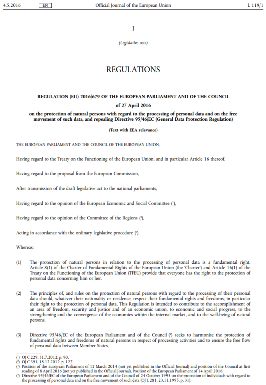
 © 2025 NetApp, Inc. All rights reserved.

Security and Privacy are crucial

- **Data Security** - protect data from unauthorized access
- Security is about safeguarding data
- **Data Privacy** - protect data from unauthorized use
- Privacy is about safeguarding user identity

- **GDPR** - General Data Protection Regulation
- **CCPA** - California Consumer Privacy Act
- **HIPAA** - Health Insurance Portability and Accountability Act

- Principle of Least Privilege & Only for limited time
- Minimize use of customer data, minimize access to them
- Delete or anonymize data at the end of their lifecycle



4.5.2016 EN Official Journal of the European Union L 119/1

I

(Legislative acts)

REGULATIONS

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular Article 16 thereof,

Having regard to the proposal from the European Commission,

After transmission of the draft legislative act to the national parliaments,

Having regard to the opinion of the European Economic and Social Committee [1],

Having regard to the opinion of the Committee of the Regions [2],

Acting in accordance with the ordinary legislative procedure [3],

Whereas:

(1) The protection of natural persons in relation to the processing of personal data is a fundamental right. Article 8(1) of the Charter of Fundamental Rights of the European Union (the 'Charter') and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) provide that everyone has the right to the protection of personal data concerning him or her.

(2) The principles of, and rules on the protection of natural persons with regard to the processing of their personal data should, whatever their nationality or residence, respect their fundamental rights and freedoms, in particular their right to the protection of personal data. This Regulation is intended to contribute to the accomplishment of an area of freedom, security and justice and of an economic union, to economic and social progress, to the strengthening and the convergence of the economies within the internal market, and to the well-being of natural persons.

(3) Directive 95/46/EC of the European Parliament and of the Council [4] seeks to harmonise the protection of fundamental rights and freedoms of natural persons in respect of processing activities and to ensure the free flow of personal data between Member States.

[1] OJ C 229, 31.7.2012, p. 90.
[2] OJ C 391, 18.12.2012, p. 127.
[3] Position of the European Parliament of 12 March 2014 (not yet published in the Official Journal) and position of the Council at first reading of 8 April 2016 (not yet published in the Official Journal). Position of the European Parliament of 14 April 2016.
[4] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (OJ L 281, 23.11.1995, p. 31).

(Title page of
GDPR official text)

 © 2025 NetApp, Inc. All rights reserved.

PostgreSQL and Data Governance

- Check constrains and triggers can help with data quality
- Comments on all objects can help with data catalog
- pgTAP extension - unit testing framework / data quality checks
- pg_analytics with DuckDB SQL features for data profiling

- Data Governance uses mostly external tools and processes
- Great Expectations / dbt for Data Quality checks
- Apache Atlas / OpenMetadata for Data Catalog/Lineage
- OpenLineage for Data Lineage for AI / ML
- Marquez for open source Metadata Service

 © 2025 NetApp, Inc. All rights reserved.

AI and Data LakeHouse

Over Promising AI Marketing Hype

- Do not believe every new AI marketing hype

- Everything these days is ”AI powered”, ”AI driven”
- Many marketing ebooks over promise AI capabilities
- Yes, AI is the future, no doubt about it, but...

- Usefulness of AI in Data LakeHouse depends on use cases
- Commercial AIs can lead to privacy and security issues
- Local Open Source AI solutions give more control
- But are usually not that powerful

 © 2025 NetApp, Inc. All rights reserved.

AI Answers Based on Probability

- We currently have Large Language Models (LLMs)
- LLMs use Transformer architecture
- They use an ”Attention mechanism” to understand context
- LLMs generate text based on training data
- Answers are the ”most probable”, not necessarily ”correct”
- For LLMs, there is no ”correct” or ”incorrect” answer
- Answers depend on activation of semantic associations
- Prompt engineering and system prompts are crucial
- *”Just because it sounds plausible, doesn’t mean it’s true”*

 © 2025 NetApp, Inc. All rights reserved.

Problematic Usability for Niche Topics

- LLMs absolutely depend on the quality of training data
- Amount, quality and topic coverage are crucial
- LLMs work amazingly for general topic data
- Like invoicing, financial reports, warehouse management
- If you have mainly these use cases, AI is perfect for you

- More specialized topics often lead to hallucinations
- But you can definitely use AI for brainstorming
- It can give you new ideas and perspectives
- But you must always double-check the results

 © 2025 NetApp, Inc. All rights reserved.

Most Common Issues with AI Outputs

- **Overgeneralization**: wrong conclusions due to biased data
- **Misinterpretation**: wrong conclusions due to wrong context
- **Underfitting**: model too simple to capture details, too general
- **Overfitting**: AI specialized on training data, cannot generalize

- **Overinterpretation of the Input**: wrong conclusions due to incomplete input, hallucinates missing parts of the input

- **Out-of-distribution Generalization**: wrong conclusions due to topic not covered by training data

 © 2025 NetApp, Inc. All rights reserved.

Some Older AI Promises stayed Unfulfilled

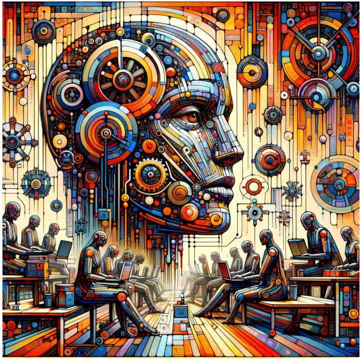
- **Fine-tuning** on domain specific data can shift performance
- Model specializes on new data
- But struggles with more general data
- Can even lead to ”catastrophic forgetting”

- **Retrieval-Augmented Generation (RAG)**
- Depends fully on the quality of additional data
- Very useful for chat-bots, and help-desk systems
- Not that great for analysis of complex data
- Highly specific data require examples and explanations

 © 2025 NetApp, Inc. All rights reserved.

Will AI-agents do better?

- **AI-agents** are new hype
- They can run additional tasks like internet browsing
- Could also run machine learning models
- Could use multiple knowledge sources
- Capable of multi-step reasoning
- But still depend on quality of LLMs
- We can expect best performance on well known topics
- Niche topics can lead to multiple levels of hallucinations



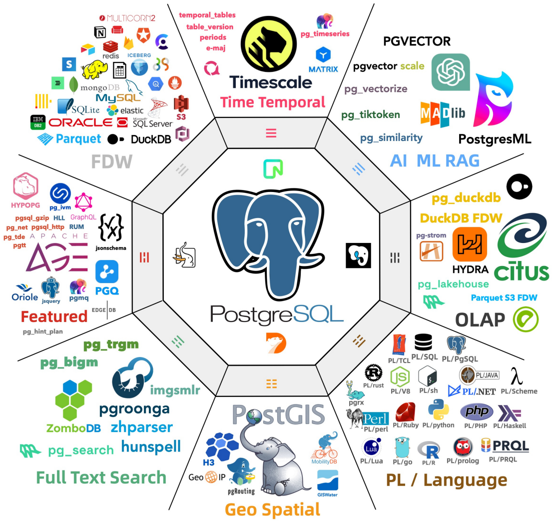
 © 2025 NetApp, Inc. All rights reserved.

PostgreSQL and AI + ML

- AI, ML and PostgreSQL are a great match
- Has multiple extensions for AI and ML
- pgvector for vector similarity search for RAG
- Timescale pgvectorscale improved pgvector extension
- Timescale pgai automates creation of embeddings for RAG
- PostgresML for machine learning in PostgreSQL

 © 2025 NetApp, Inc. All rights reserved.

Summary

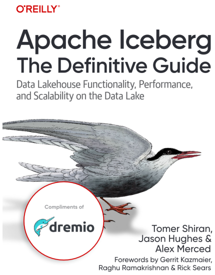


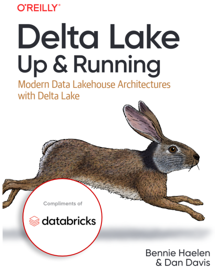
(Image from the article Postgres is eating the database world on PIGSTY (PostgreSQL In Great STYle) blog)

 © 2025 NetApp, Inc. All rights reserved.

Resources

- Other resources used for the talk, not mentioned in slides:
- Articles:
  - What is a Data Lakehouse?
  - History and evolution of data lakes
  - What is Data as a Product
  - Apache Iceberg main web page
  - Dremio.com: What is Apache Iceberg
  - Sqream.com: What is Apache Iceberg
  - Estuary.dev: Apache Iceberg vs Hudi
- E-books:
  - T.Shiran, J.Hughes, A.Merced: Apache Iceberg, The Definitive Guide - O'Reilly
  - Bennie Haelen, Dan Davis: Delta Lake, Up & Running - O'Reilly
  - Dremio white paper: Optimizing the supply chain with a data lakehouse
  - A.Kaplan, A.Kara: Data Lakehouse for Dummies - Databricks
- AI tools:
  - NetAppAI GPT-4o, NetApp GitHub CoPilot AI
  - Paid tier ChatGPT-4o/ o1, Google Gemini Advanced 1.5/ 2.0





 © 2025 NetApp, Inc. All rights reserved.

THANK YOU

- **Questions?**
- Josef Machytka `<josef.machytka@netapp.com>`

NetApp

 © 2025 NetApp, Inc. All rights reserved.