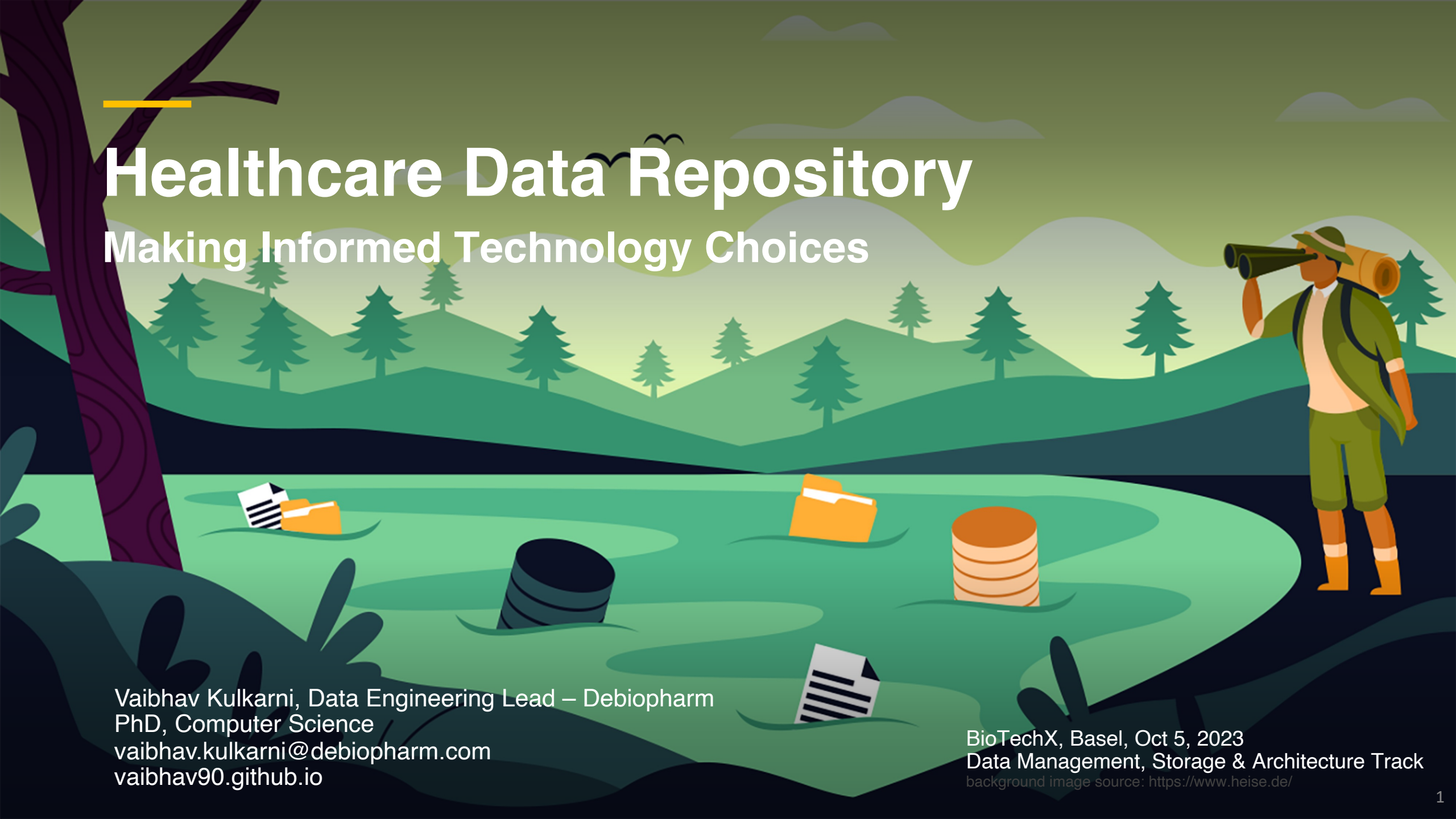


Healthcare Data Repository

Making Informed Technology Choices

Vaibhav Kulkarni, Data Engineering Lead – Debiopharm
PhD, Computer Science
vaibhav.kulkarni@debiopharm.com
vaibhav90.github.io

BioTechX, Basel, Oct 5, 2023
Data Management, Storage & Architecture Track

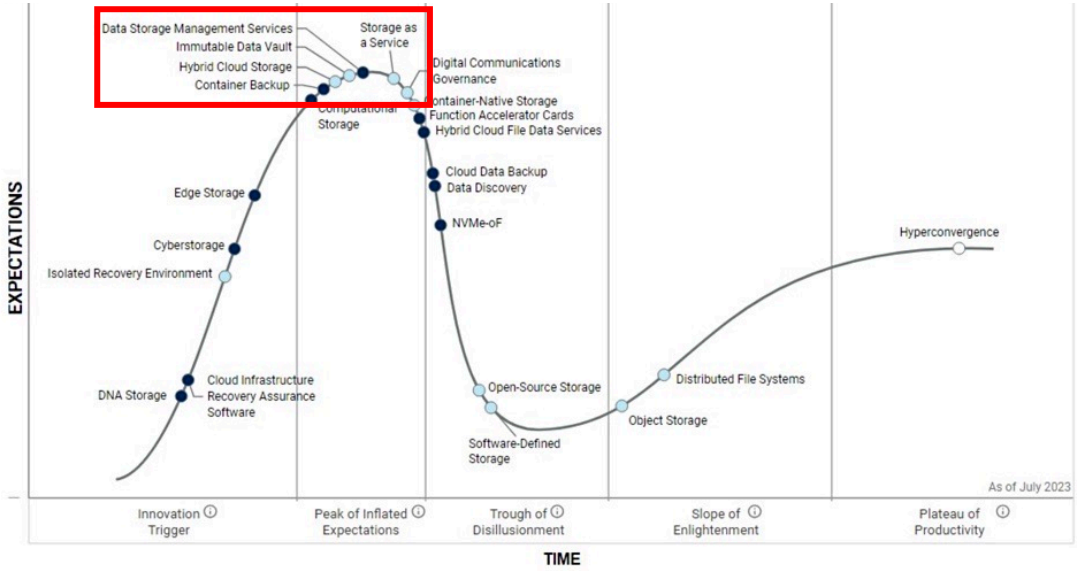
background image source: https://www.heise.de/

Agenda

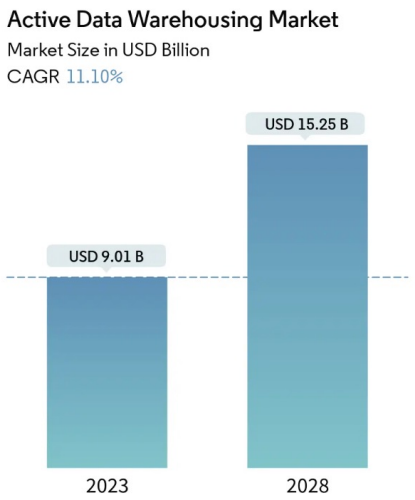
- Healthcare data repository - motivation
- Technical architecture - key components
- Today’s technology landscape - tools/platforms
- Informed selection - data-driven methodology
- Prevalent business cases
- Lessons learnt

Building Future-Ready Healthcare Data Repository

- Data Centralization: Transform raw data into actionable insights
- Innovation Catalyst: Expedite AI-driven drug discovery process



exponential data growth pushes data-store technologies on top of the hype cycle



Study Period	2018-2028
Market Size (2023)	USD 9.01 Billion
Market Size (2028)	USD 15.25 Billion
CAGR (2023 - 2028)	11.10 %
Fastest Growing Market	Asia-Pacific
Largest Market	North America
Major Players	ORACLE, aws, Microsoft, Hewlett Packard Enterprise, SAP

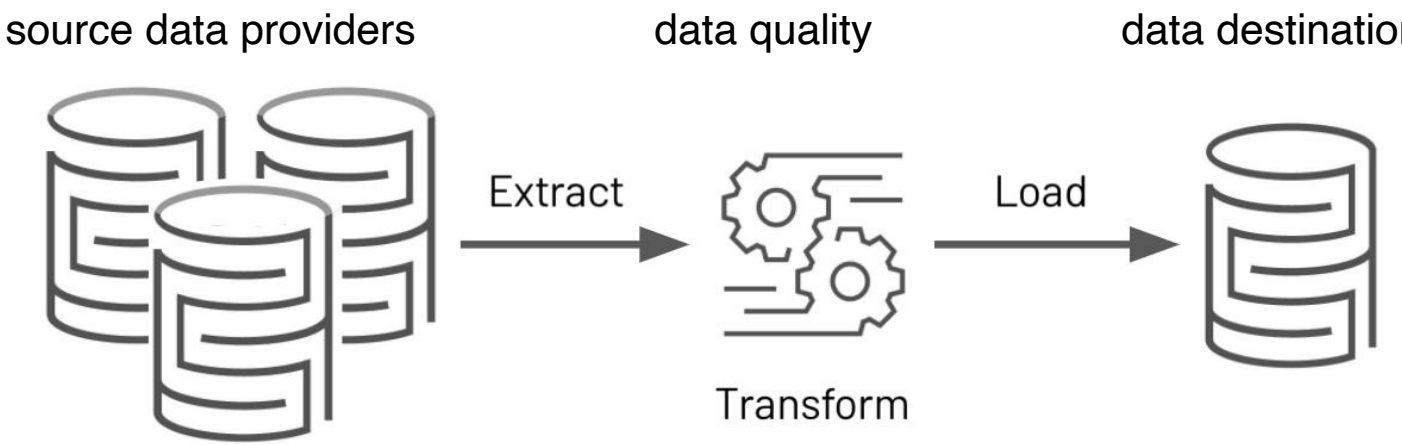
market shift: Increased data, advanced intelligence

sources

https://www.gartner.com/en/documents/4009950 Gartner

https://www.mordorintelligence.com/industry-reports/ Mordor Intelligence

Data Repository Architecture: Key Components



data pipeline

data quality

data storage

Today's Tool/Platform Landscape

Data Pipeline

Apache Airflow
Stitch
Airbyte
Flink
AWS Data Pipeline
MuleSoft
Azure Synapse Analytics
IBM DataStage
Google Dataflow
APACHE nifi

Data Quality

dbt
TRIFACTA
great expectations
APACHE Spark
databricks
DataRobot
beam
Informatica
pandas
SQL

Data Storage

Azure Data lake
amazon REDSHIFT
Azure Cosmos DB
snowflake
PostgreSQL
mongoDB.
Amazon S3
Microsoft SQL Server
Google Cloud Storage
Google Big Query

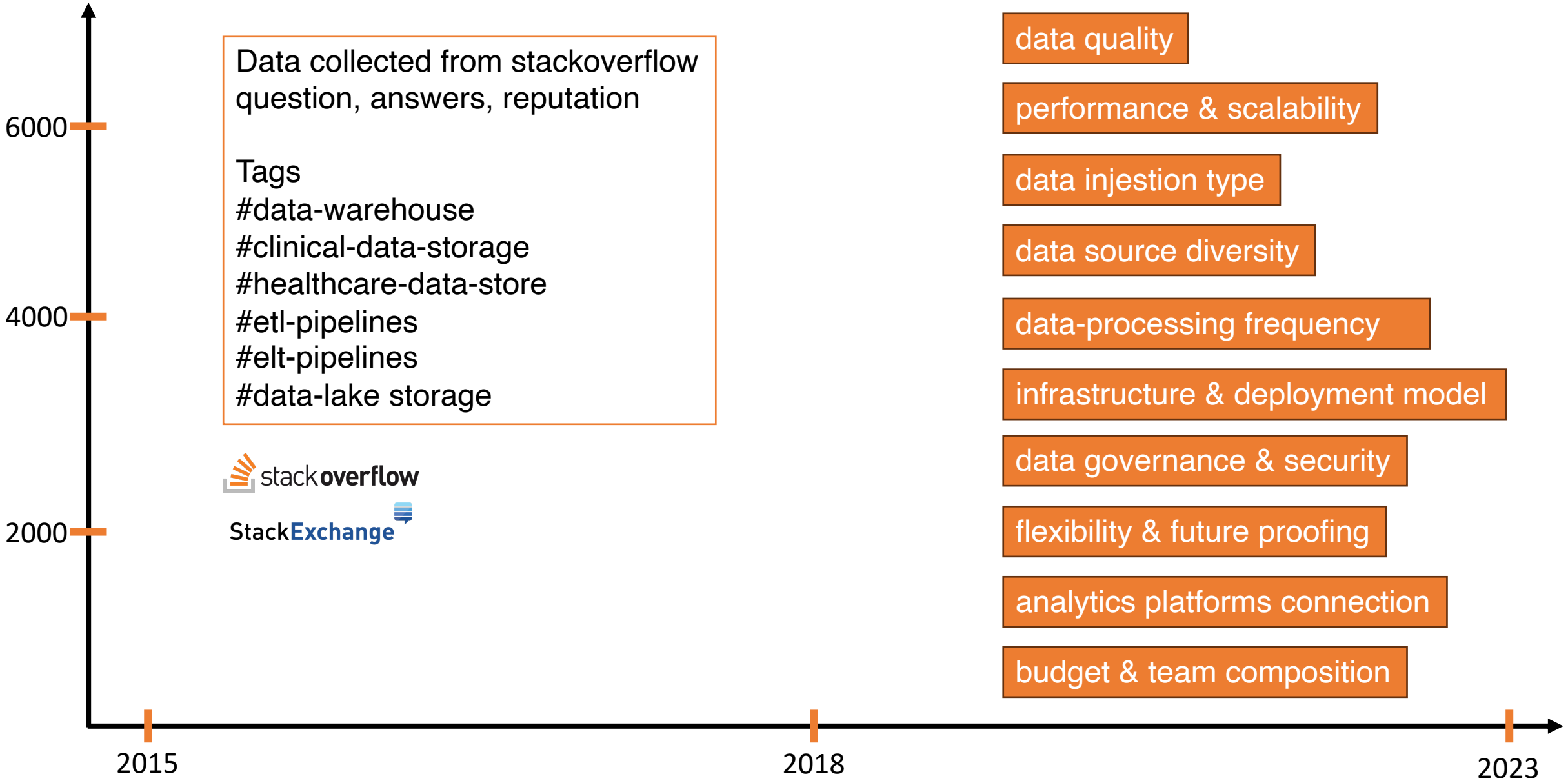
* Top 10 tools/platforms in each category – data from stackoverflow, HackerNews posts and other community forums

Data sources:
https://stackoverflow.com
https://news.ycombinator.com
https://dev.to/
https://www.reddit.com/r/dataengineering/

stackoverflow
reddit
HN Hacker News
DEV

data available: https://github.com/vaibhav90/biotechX-CDW-Tech-Choices

Architecting Data Repository - Decision Areas



Data collected from stackoverflow
question, answers, reputation

Tags
#data-warehouse
#clinical-data-storage
#healthcare-data-store
#etl-pipelines
#elt-pipelines
#data-lake storage

stackoverflow

StackExchange

- data quality
- performance & scalability
- data injestion type
- data source diversity
- data-processing frequency
- infrastructure & deployment model
- data governance & security
- flexibility & future proofing
- analytics platforms connection
- budget & team composition

Technology Selection - Metrics

Cost 💰

Setup & licensing
Operational costs

Implementation Ease 🚀

Setup & integration time
User interface & experience

Security & Compliance 🔒

Compliance
Security & user access

Flexibility & Scalability 🛠️

Adaptability & customization
Workload response

Maintenance & Support 🧰

Update frequency
Customer support/engagement

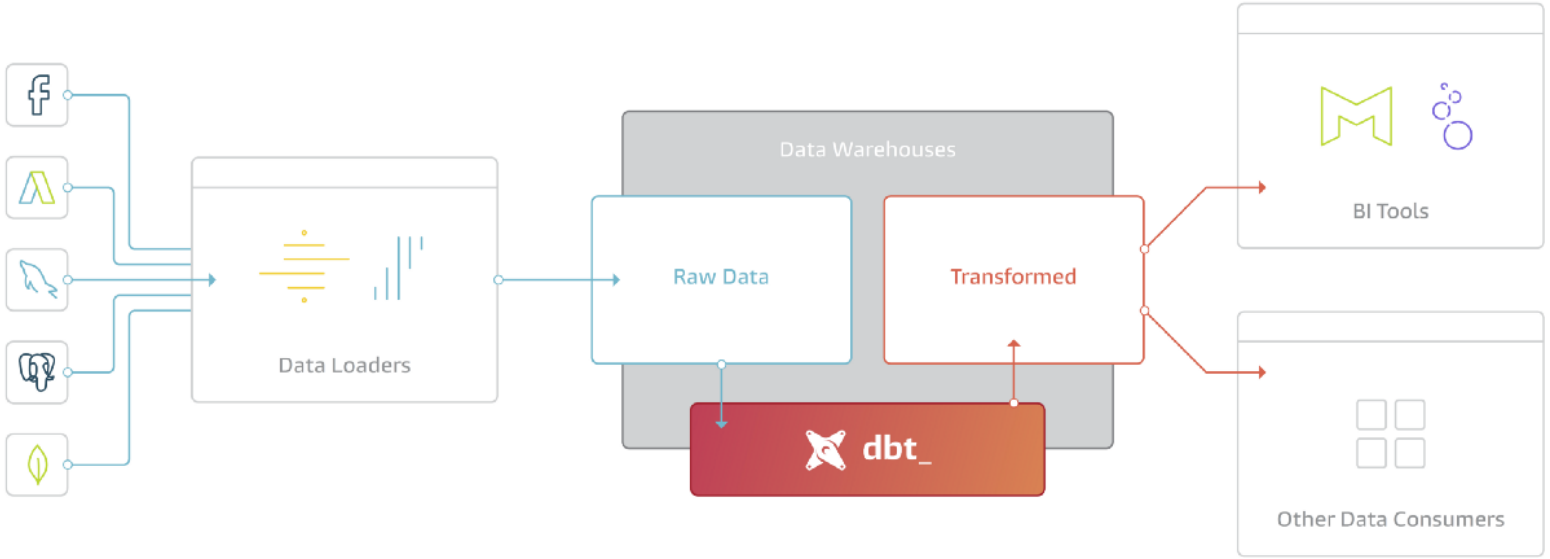
Most Common Technology Stack?


Data pipeline
Excel
Data quality
Excel
Data storage
Excel

Low-Effort Data Repository Tech-Stack

💰💰 | 🚀🚀🚀 | 🔒 🔒 | 🛠️ | 🧰🧰🧰

Enable data integration from common data sources to facilitate seamless querying for data analysts

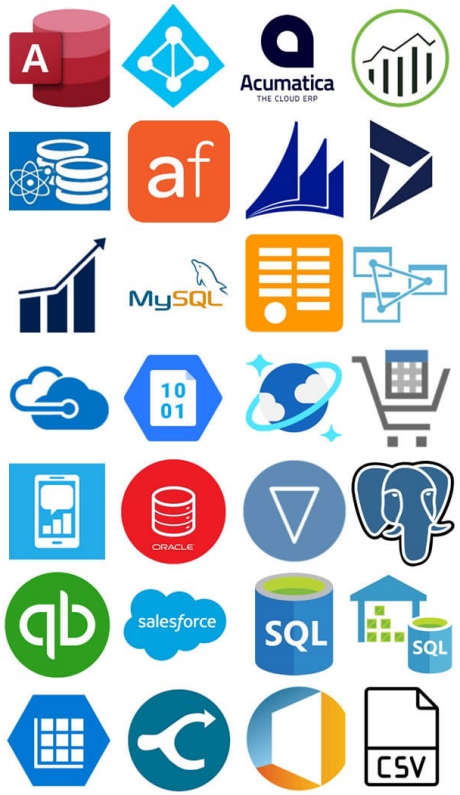


stitch - cloud-native ETL platform offering seamless integrations and rapid deployment

dbt - streamlines data modelling with built-in quality tests and an intuitive design

Complex Data Source Configuration

💰 | 🚀 🚀 | 🔒 🔒 | 🛠️ 🛠️ 🛠️ | 🧰 🧰



Today Airflow supports a total of 176 system connectors out of the box *

Apache Airflow

Airbyte

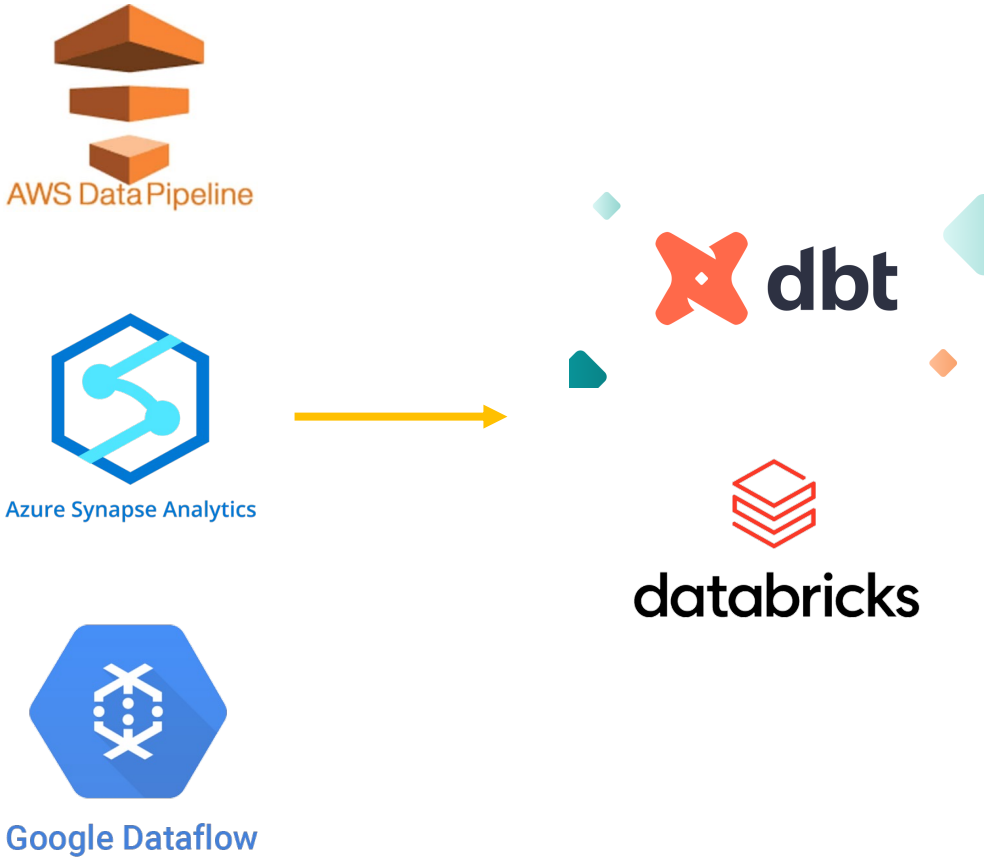
airflow is suited for complex, programmable workflow management where customization is key

stitch is a good choice for simpler, out-of-the-box data replication with less setup & maintenance

*source: https://airflow.apache.org/

Strict Regulatory Compliance & Enterprice Grade Security

💰 💰 💰 | 🚀 🚀 | 🔒 🔒 🔒 | 🛠️ 🛠️ | 🧰 🧰 🧰



- Integrated with your cloud ecosystem
- Serverless environment and support
- Built-in monitoring and management
- Global reach, security, time-to market

Complex Data Quality Checks & Scalable Transformations

💰 💰 | 🚀 | 🔒 🔒 | 🛠️ 🛠️ 🛠️ | 🧰 🧰

Stitch

Apache Airflow

Azure Synapse Analytics

AWS Data Pipeline

Google Dataflow

great expectations

High quality data in your data products

Data documentation & data quality reports

Logging & alerting

APACHE Spark™

Simple.
Fast.
Scalable.
Unified.

Key features

Batch/streaming data

Unify the processing of your data in batches and real-time streaming, using your preferred language: Python, SQL, Scala, Java or R.

SQL analytics

Execute fast, distributed ANSI SQL queries for dashboarding and ad-hoc reporting. Runs faster than most data warehouses.

Data science at scale

Perform Exploratory Data Analysis (EDA) on petabyte-scale data without having to resort to downsampling

Machine learning

Train machine learning algorithms on a laptop and use the same code to scale to fault-tolerant clusters of thousands of machines.

Data Storage Layer

Structured Data

amazon REDSHIFT

PostgreSQL

Microsoft SQL Server

snowflake

Google Big Query

Semi-structured Data

Azure Cosmos DB

💰💰💰 | 🚀🚀 | 🔒🔒🔒 | 🛠️🛠️ | 🧰🧰🧰

mongoDB.

💰 | 🚀 | 🔒🔒🔒 | 🛠️🛠️ | 🧰🧰🧰

Google Cloud Storage

💰💰 | 🚀 | 🔒🔒🔒 | 🛠️ | 🧰🧰🧰

Un-structured Data

Azure Data lake

💰💰 | 🚀🚀 | 🔒🔒🔒 | 🛠️🛠️🛠️ | 🧰🧰🧰

Amazon S3

💰💰 | 🚀🚀 | 🔒🔒🔒 | 🛠️ | 🧰🧰🧰

What Have I Learnt Over the Years?

- There is no such thing as a “correct” architecture
- Focus on the business problem, domain, & team
- Tool selection is easy, business alignment is hard
- Writing clean & maintainable code > tool selection
- Slow down, understand team’s goal, assume nothing