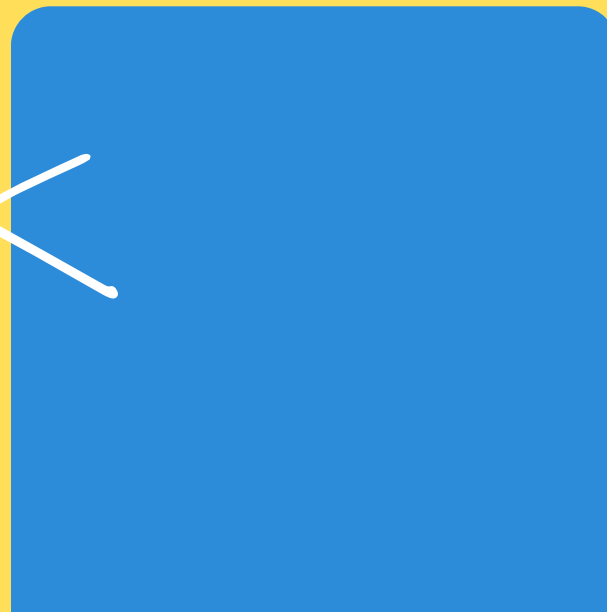
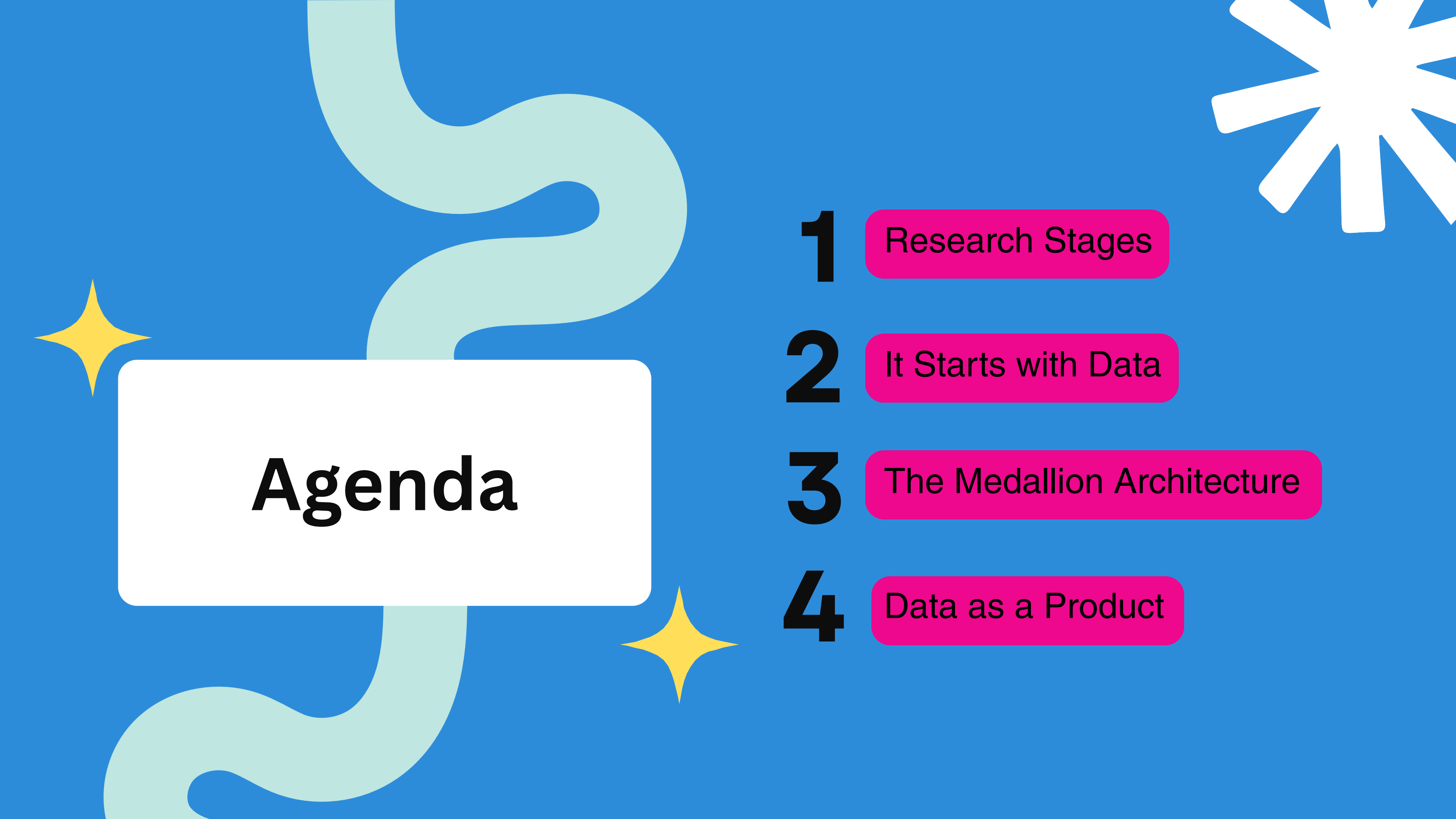


Research Data Processing

A Layered Architecture





Agenda

1

Research Stages

2

It Starts with Data

3

The Medallion Architecture

4

Data as a Product

Research Stages

Process for monitoring and strengthening internal systems

STAGE

01



Governance

Create practical policies and procedures to reduce risks and support reliable operational outcomes.

STAGE

02



Continuous Integration

Identify risks that could disrupt objectives and evaluate their possible impact on business operations.

STAGE

03



Define Workflow

Establish clear business goals that align with overall strategy and guide effective internal control measures.

STAGE

04



Set Maintenance

Implement designed measures consistently across departments to strengthen business processes and compliance.

STAGE

05



Deliverables

Track effectiveness of controls, review performance regularly, and adjust measures when improvements are needed.

IT STARTS WITH DATA
BUILD THE FOUNDATION



Images

Documents

Lakehouse

Data gives software meaning. It represents the people, activities, and processes a system is intended to support, making the data model one of the first and most important design decisions. A lakehouse supports structured, semi-structured, and unstructured data within a single platform, allowing diverse research data to be managed without separate storage systems.

To be data-driven is to use trusted information as the basis for both technical and organizational decisions. Designing around data first creates a stronger foundation for analytics, intelligent systems, and meaningful change.

THE MEDALLION ARCHITECTURE

STREAMLINE PROCESSING

The Medallion Architecture

Raw data enters once, and every downstream layer becomes cleaner, more trustworthy, and more useful.

Bronze Layer

First point of ingest: A centralized location that catalogs raw data.

Stores raw, immutable data exactly as it is received from source systems. This layer preserves the original records for auditing, reproducibility, and future reprocessing.

The Medallion Architecture

Raw data enters once, and every downstream layer becomes cleaner, more trustworthy, and more useful.

Silver Layer

The processing pipeline: data is transformed from raw input into trusted information.

Transforms raw data into clean, versioned, validated, and standardized datasets that are ready for integration and analytical use.

The Medallion Architecture

Raw data enters once, and every downstream layer becomes cleaner, more trustworthy, and more useful.

Gold Layer

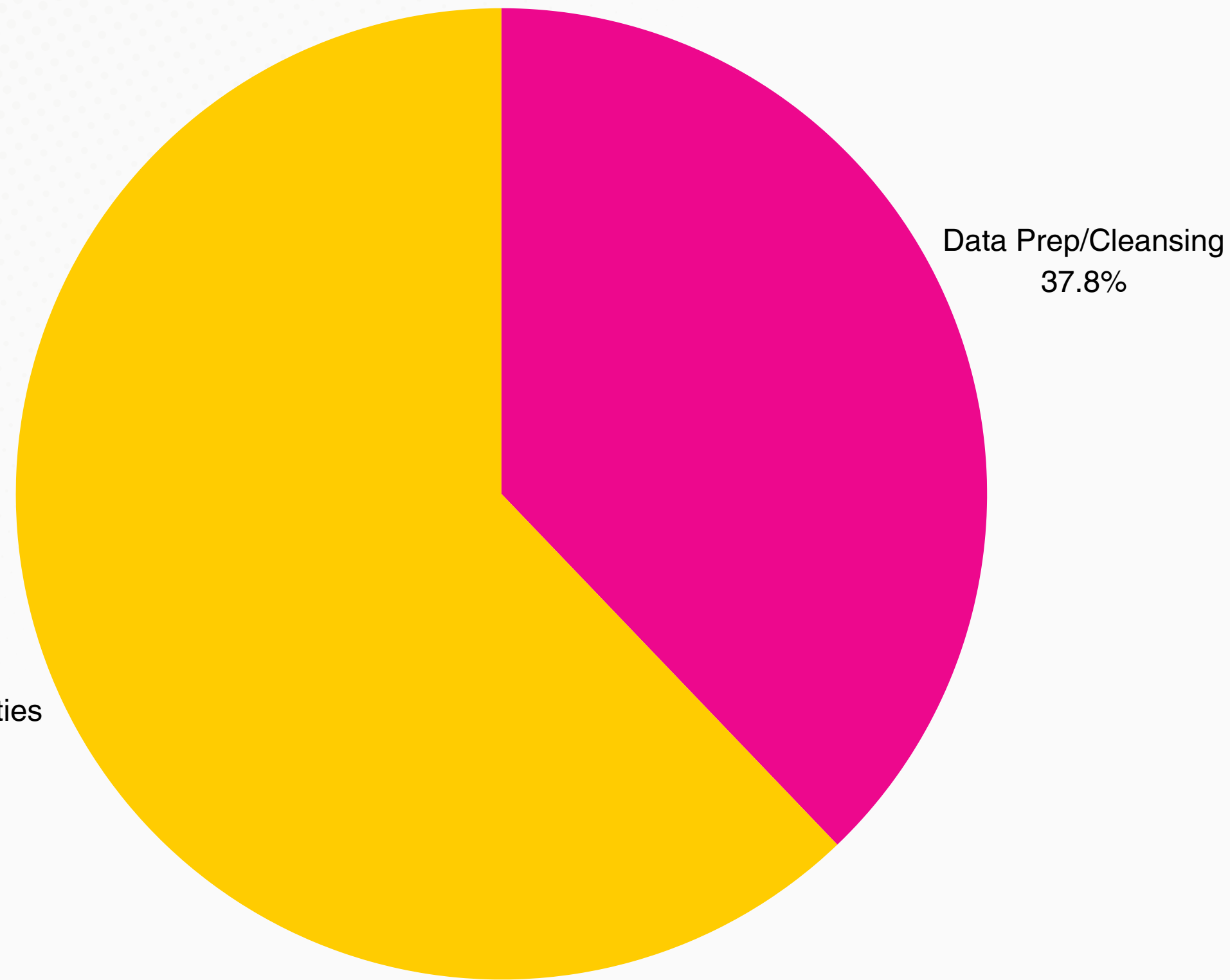
The consumption layer: contains business-ready data products optimized for answering questions.

Organizes trusted information into domain-specific data products that enable analytics, operational reporting, and strategic insights.

OBJECT-ORIENTED **ENTITY DRIVEN SYSTEMS**

THE OBJECT/ENTITY

An object/entity is a representation of a meaningful concept within a problem domain. Defining an object/entity means identifying the responsibilities, relationships, and interactions required for that concept to participate in the larger system.



37.75%

In Anaconda's 2022 State of Data Science survey, practitioners reported spending 37.75% of their time on data preparation and cleansing before analysis, visualization, or modeling.

A LAYERED ARCHITECTURE

LAKEHOUSE

The lakehouse provides a shared environment for structured and unstructured research data.

MEDALLION

The medallion architecture moves data from raw source records to validated information and reusable analytical products.

MESH

Data mesh principles would extend this model by treating curated datasets, transformations, and derived features as governed data products owned by the research teams that understand them best.

THE SYSTEM IS THE WORKFLOW

THINK TOP-DOWN

WORKFLOW DEFINITION

GOVERNANCE

Medallion manages the technical lifecycle of data, not the governance. Are there any rules surrounding access and storage?

What defines the actions or tasks necessary within the workflow? What are the metrics that matter?

DESIGN

TOOLBOX

What's your go to workflow (i.e. preferred tools)? Do you want to be able to run queries against a table? Do you expect the current use case to scale?

What would be your ideal future state? What does the completion of this workflow mean to you?

GROWTH

ANALYSIS

How were you planning to analyze the data? In other words, have you already decided on a data model?

What does output validation need to look like? What are the core parts of the current process?

MAINTENANCE

THANK YOU