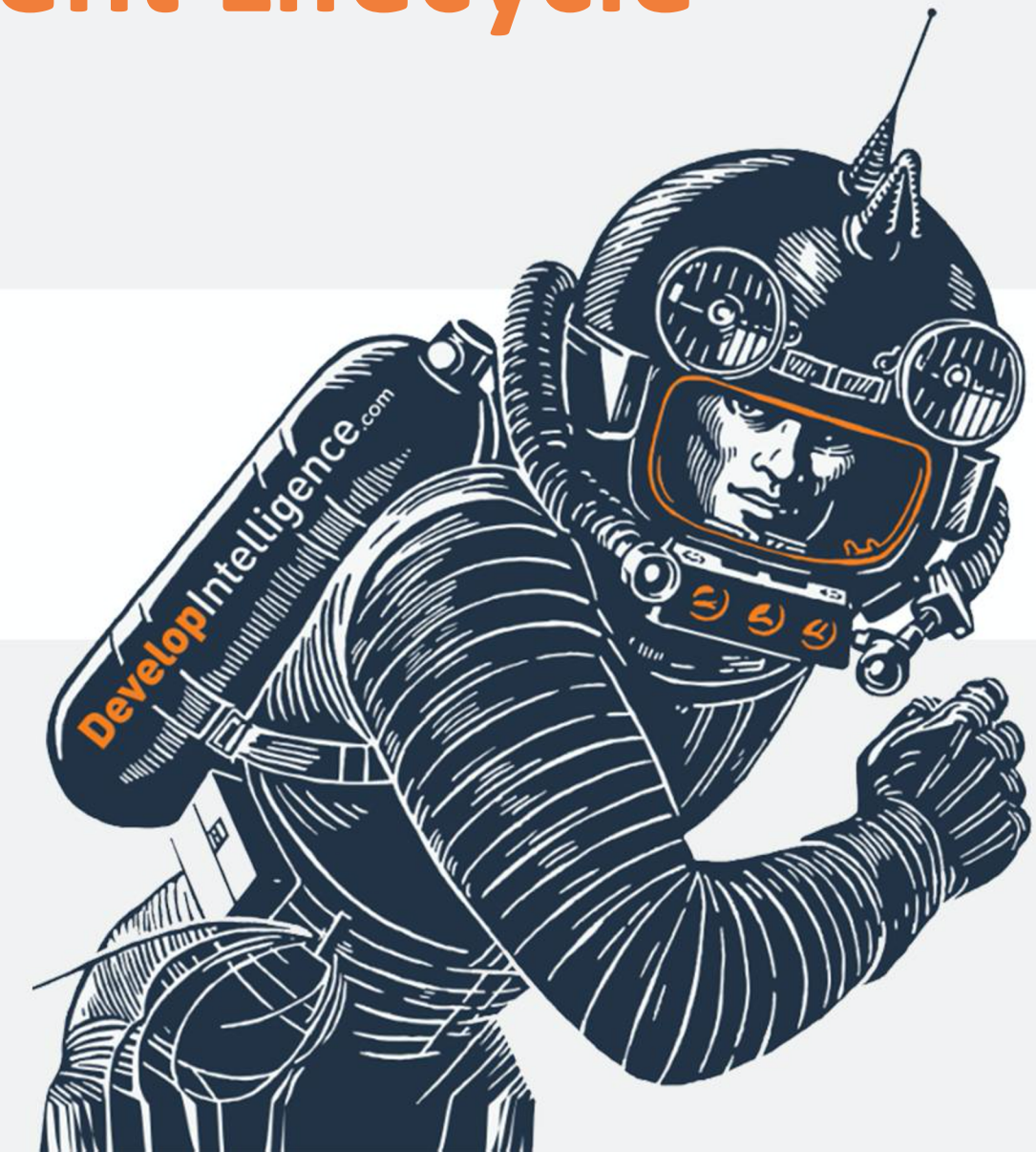
Data Engineering Best Practices

Develop Intelligence

Kerry Nakayama

Software Development Lifecycle



Develop Intelligence

Kerry Nakayama

Objectives

- Should Data engineering be treated like software engineering
  - Waterfall development
- Plan, **Design**, & Implement
  - Planning
  - Designing
  - Implementing
- Root Causes
- Planning
  - ERD
  - Data Dictionary
  - Core KPIs
  - Error handling
- Design and Environments
  - Snowflake
  - Databases
  - Uses
  - Security and permissions

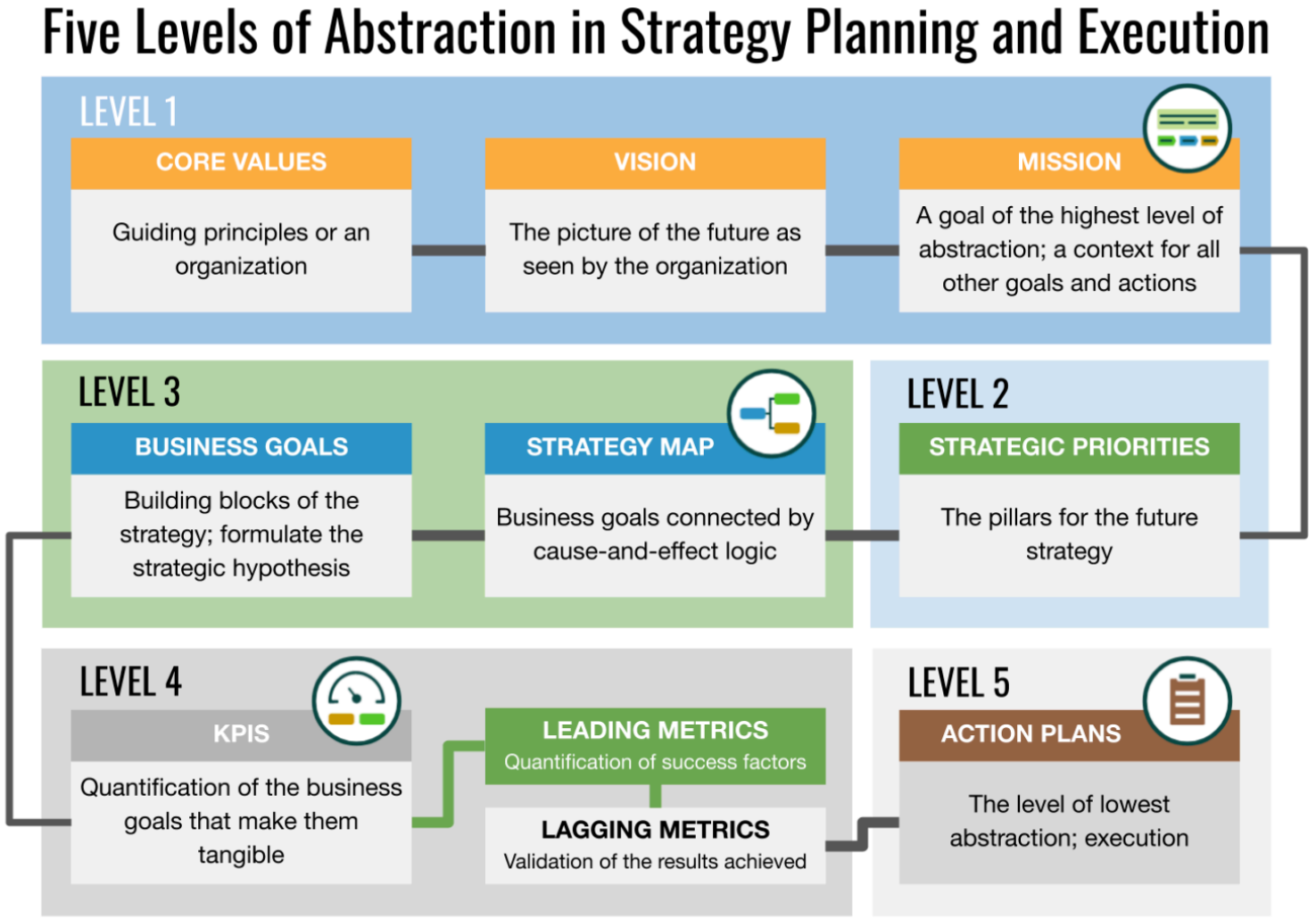
Doing Data Work as it Were Software Engineering

- Develop iteratively.
- Manage requirements.
- Use component architecture.
- Model **software** visually.
- Verify quality.
- Control change.

- Poor Quality
- Unacceptable performance
- To hard to maintain or extend
- Inaccurate understanding of end users needs
- Inability to deal with changing requirements
- Late discovery of serious flaws

Plan, Design, Implement

- Plan – What are you trying to solve?
- Design – ERD, architecture, who are the tools for
- Implement – Cultural change not just tool change

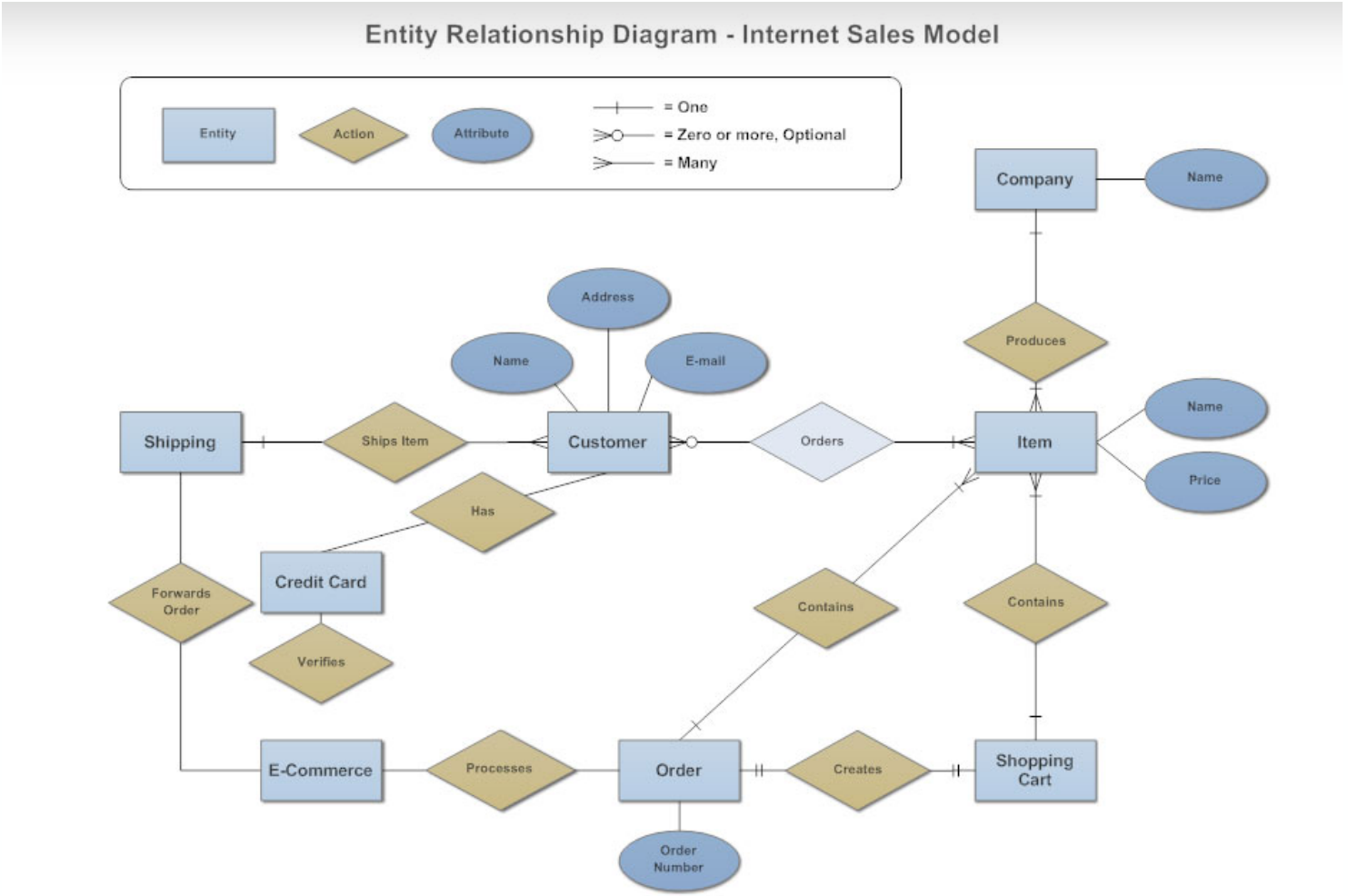


What are the Root Causes of Poor Development?

Develop Intelligence

- Ambiguous communication
- Overwhelming complexity
- Insufficient testing
- Insufficient requirements management
- Undetected inconsistencies among requirements, designs and implementations
- Brittle architecture

ERD – Details on Architecture



What is an Entity Relationship Diagram (ERD)?

Data Dictionaries

Data Dictionary

Data Dictionary outlining a Database on Driver Details in NSW

Field Name	Data Type	Data Format	Field Size	Description	Example
License ID	Integer	NNNNNN	6	Unique number ID for all drivers	12345
Surname	Text		20	Surname for Driver	Jones
First Name	Text		20	First Name for Driver	Arnold
Address	Text		50	First Name for Driver	11 Rocky st Como 2233
Phone No.	Text		10	License holders contact number	0400111222
D.O.B	Date / Time	DD/MM/YYYY	10	Drivers Date of Birth	08/05/1956

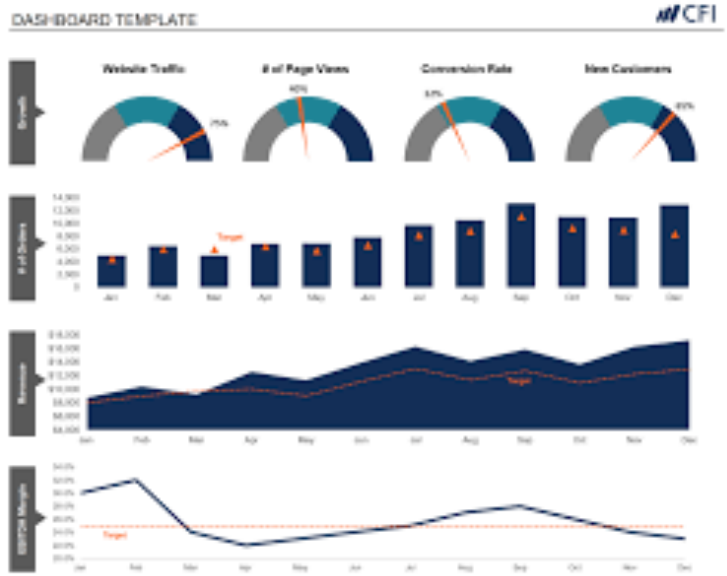
Screencast-O-Matic.com

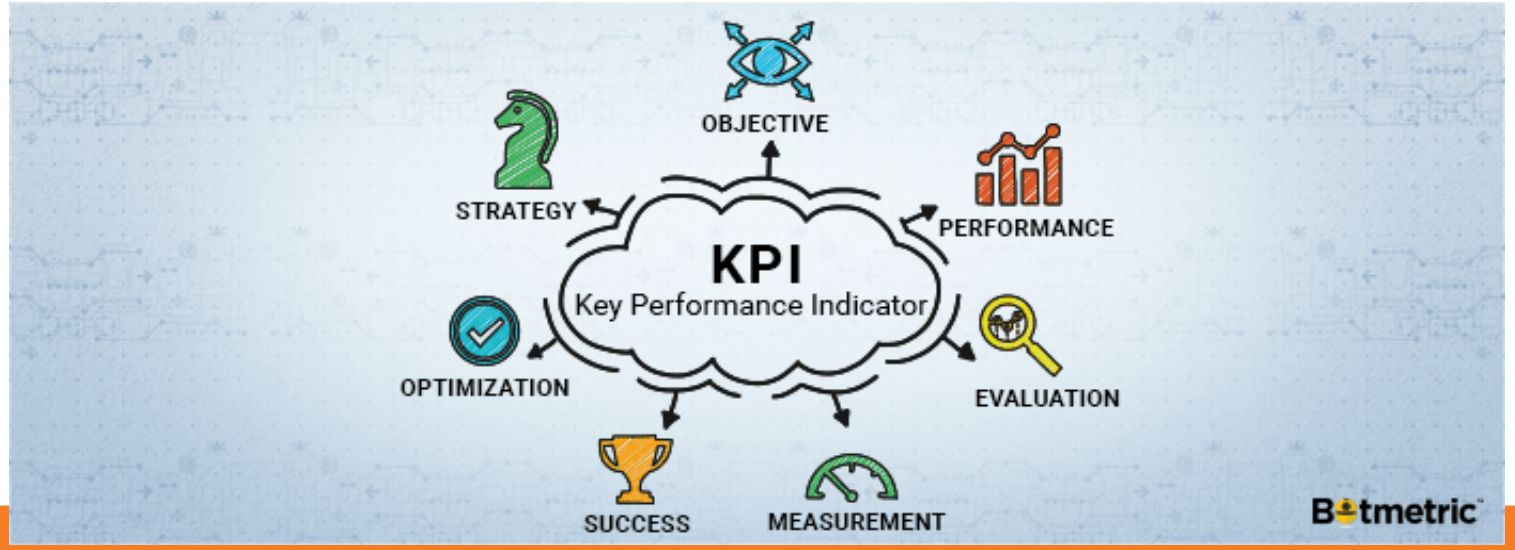
Centralized repository of information about data such as meaning, relationships to other data, origin, usage, and format

- A document describing a database or collection of databases
- An integral component of a DBMS that is required to determine its structure
- A piece of middleware that extends or supplants the native data dictionary of a DBMS

Core KPIs

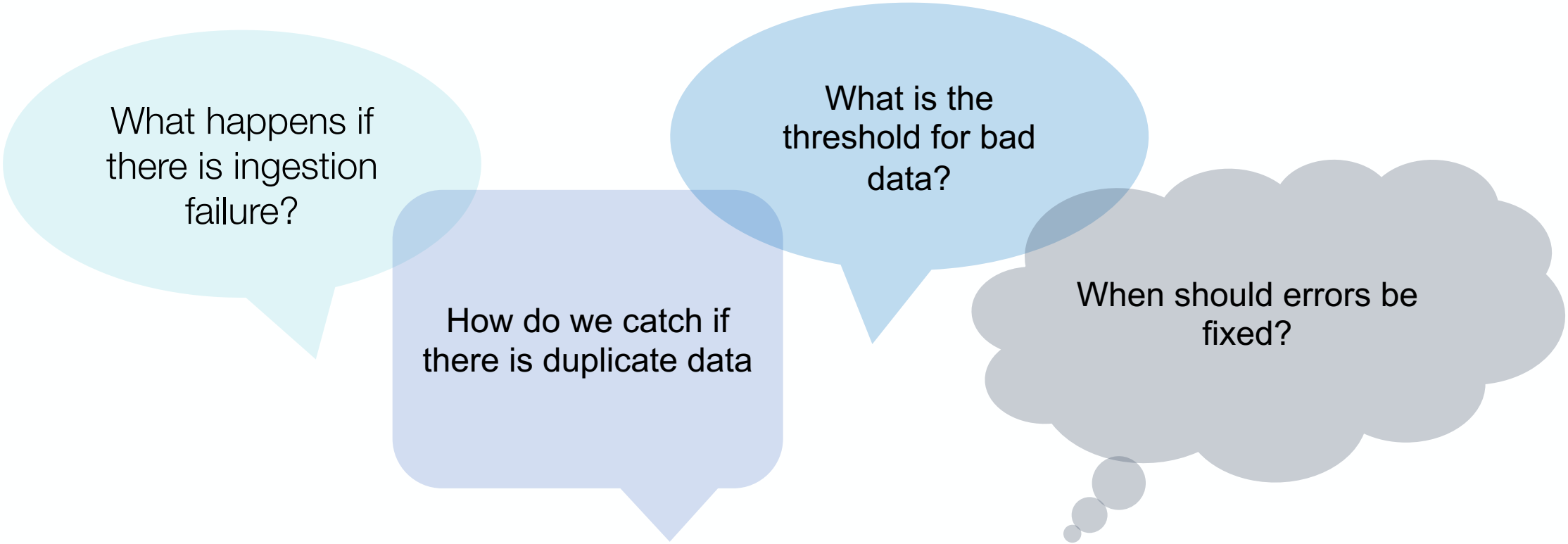
- Key Performance Indicator
  - Why are they important?
  - What happens when they are not clearly defined?
  - Trust begins to break down… hard battle back
  - Anyone here ever have this happen?



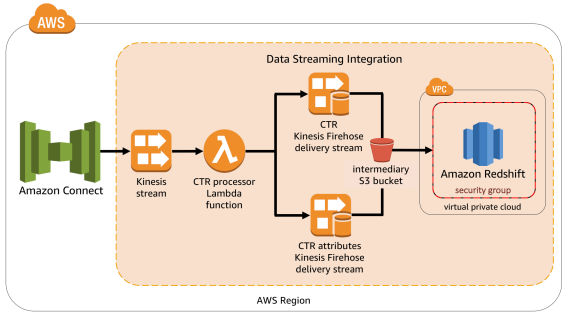


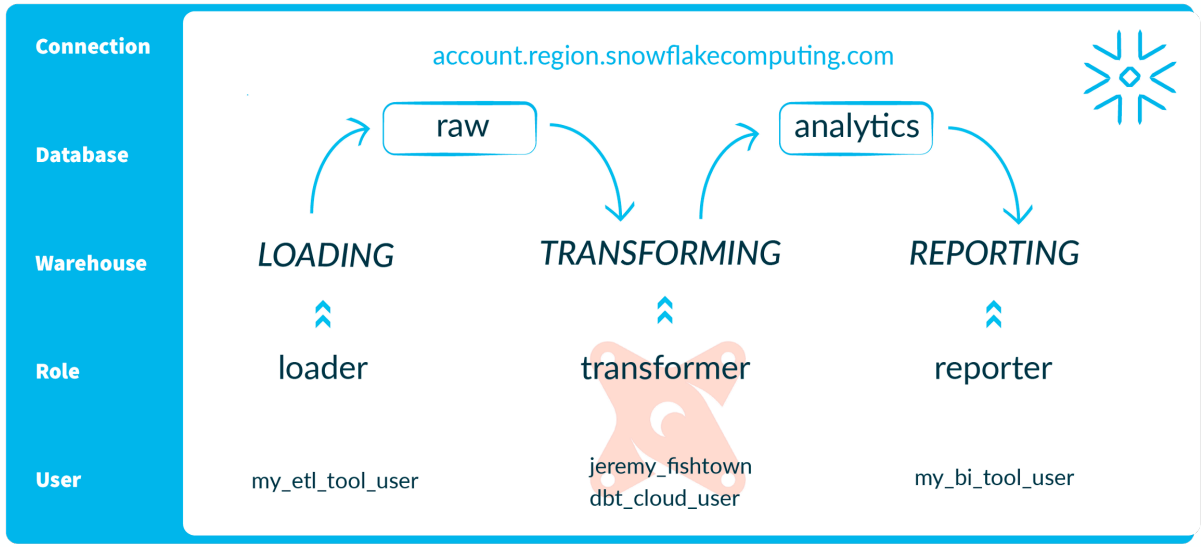
Quick Design Pattern Around How to Ensure Safe Design

- Error handling
  - What is one of the biggest areas where an error can happen for Data Engineers?



Environments

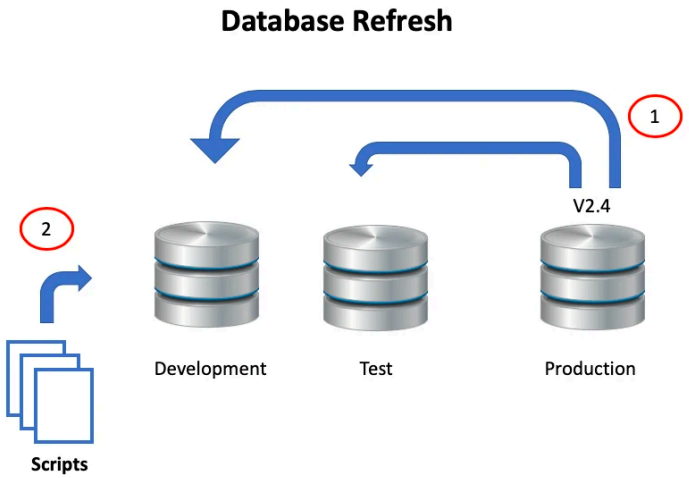


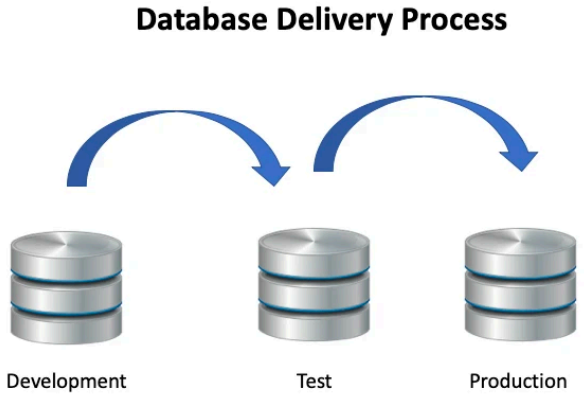


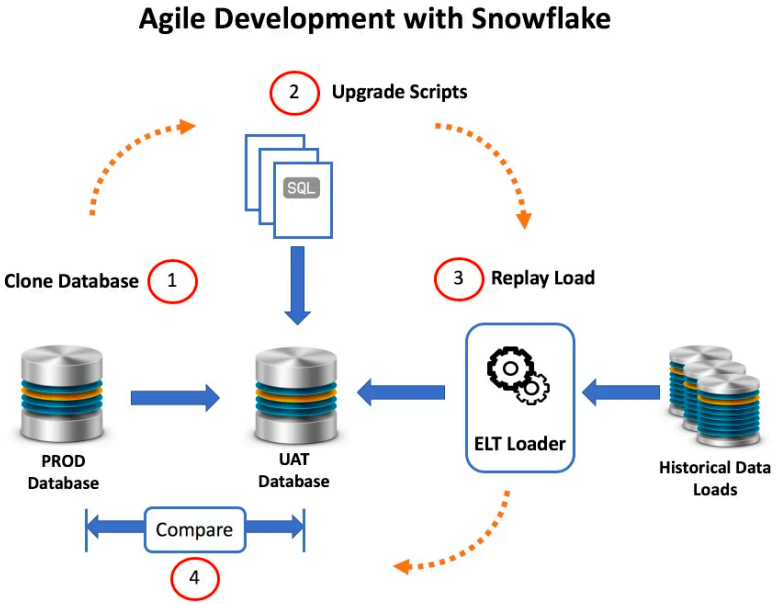
- Snowflake environments
  - Multiple accounts vs Single account
  - How many Databases?
    - 2? Raw and analytics
    - 3-4? Loading, transforming, prod, reporting
  - Who are your users and how many Roles ?
    - 4 Roles?
    - 5 Roles?
  - Security and permissions

How to get the Perfect Copy in Production

- Development, Test, and Production Environments
  - Should there be exact copies of the data?
  - Snowflake has a tool .. Zero copy CLONE and Sampling





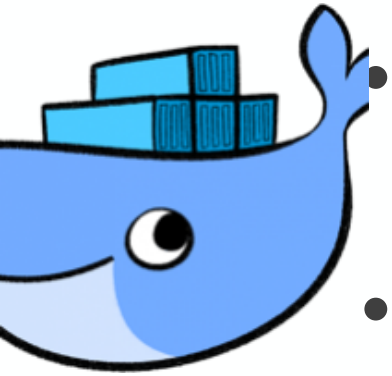


Creating Maintainable & Repeatable Environments

- Snowflake Zero Copy Clones and Time Travel
  - Query by timestamp
  - Query by Statement ID
- Database Backup and Restore
  - Restore a database
  - Recover a database
  - Create an updatable clone copy
  - Maintain multiple independent clones

Nobody needs to perform a backup. You need to be able to recover from data corruption.

Docker/ Kubernetes



- DOCKER (self contained containers intended to wrap around a piece of code and do A thing) to utilizing AWS managed services point to this as the future of coding
- Think of Docker (in the context of this class) as a VM intended to get everyone onto the same OS

- Kubernetes works by managing a cluster of compute instances and scheduling containers to run on the cluster based on the available compute resources and the resource requirements of each container

Provisioners

- Create Virtual machines “as needed”
- ALL MACHINES MUST BE THE SAME THOUGH (or else responses could change)

Confused? ASK QUESTIONS

Break & Lab 9