

Data Pipeline

Breakout Session

2021-05-27

What is a data pipeline?

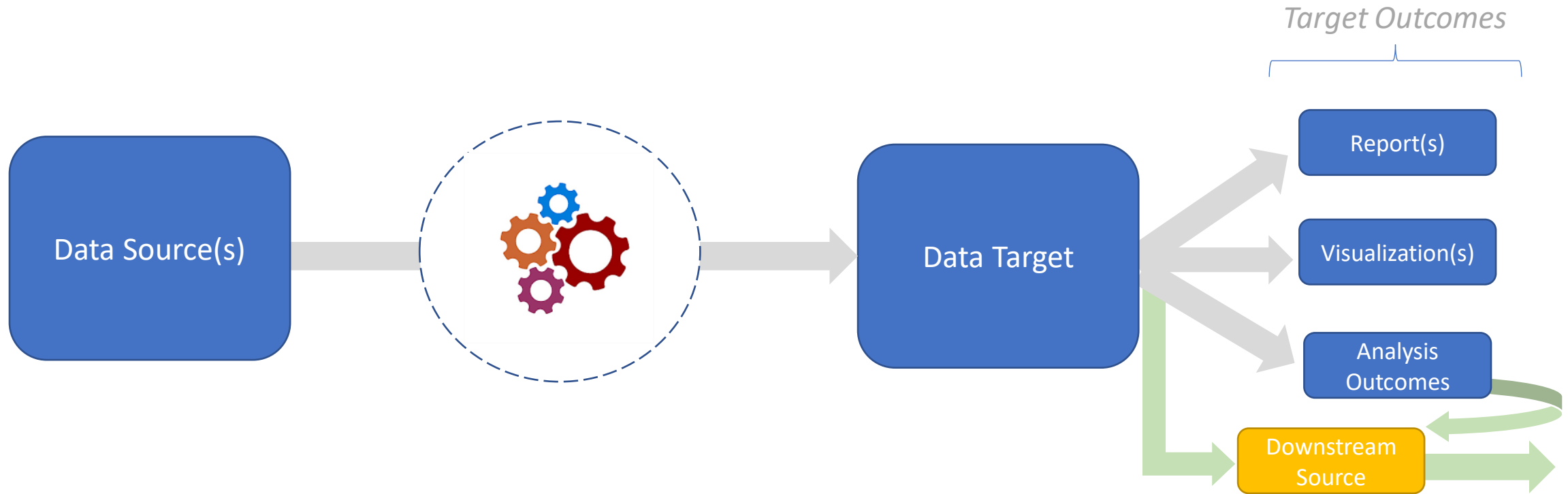
A data pipeline is the series of steps required to make data from one system useful in another. The steps might include ingesting, transforming, processing, publishing, or moving data.

Sourced 2021-05-25: <https://streamsets.com/learn/data-pipelines/>

Data Pipeline

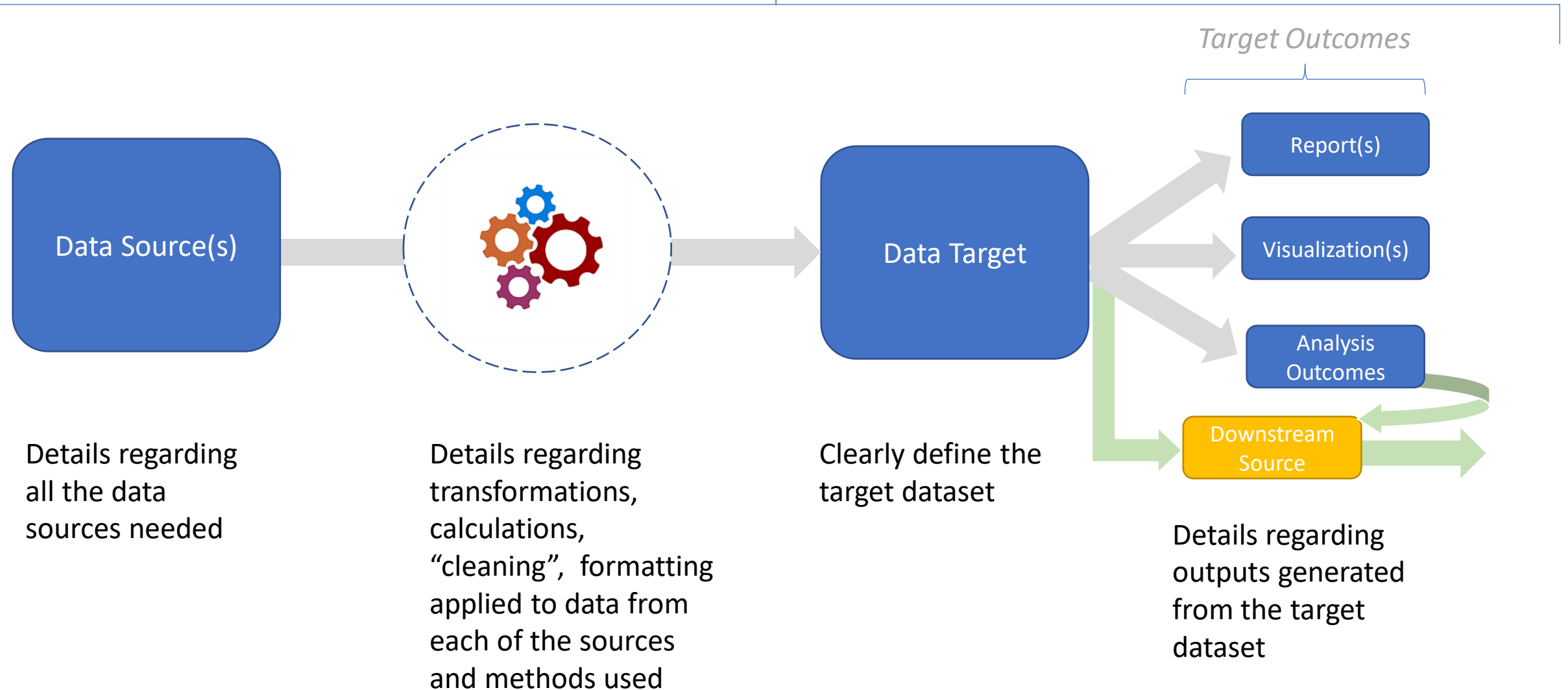


Data Pipeline

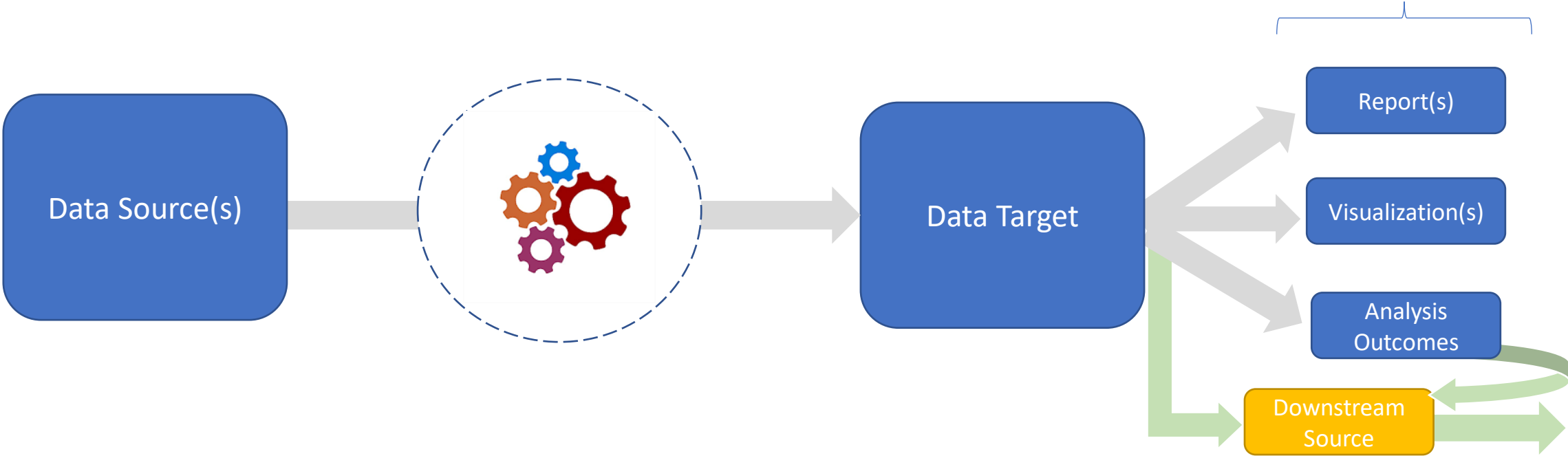


Data Pipeline - Documentation

Document the Current State (as is)



Data Pipeline - Documentation



SourceId
SourceName
SourceSystem
SourcePurpose
SourceSteward
SourceRefreshFrequency
SourceDataNeeded
SourceExtractMethod
SourceExtractSteward
SourceExtractFormat
SourceExtractFrequency
SourceExtractDestination

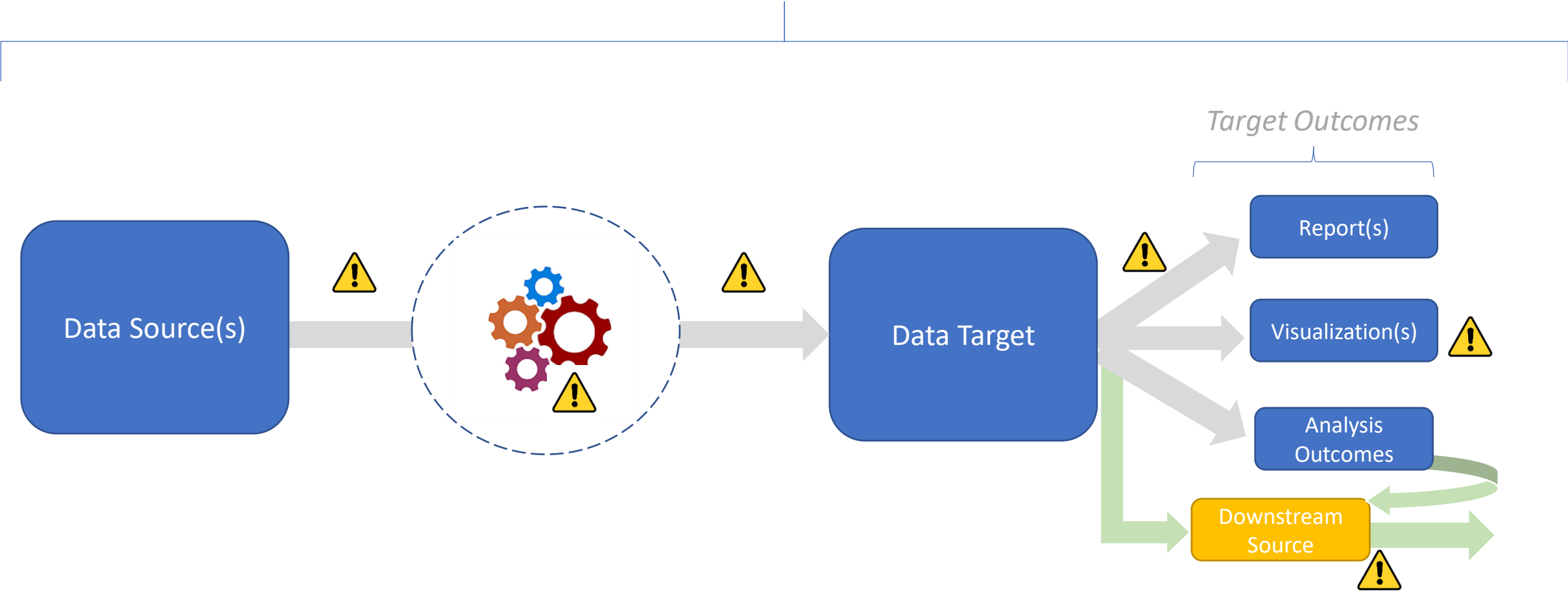
TransformId
TransformName
TransformPurpose
TransformMethod
TransformSystem
TransformSteward
TransformDestination
TransformFormat
TransformParentId

TargetId
TargetName
TargetSystem
TargetPurpose
TargetSteward
TargetFormat
TargetRefreshFrequency

OutcomeId
OutcomeName
OutcomeSystem
OutcomePurpose
OutcomeSteward
OutcomeConsumer
OutcomeExtractMethod
OutcomeFormat
OutcomeRefreshFrequency
OutcomeDestination

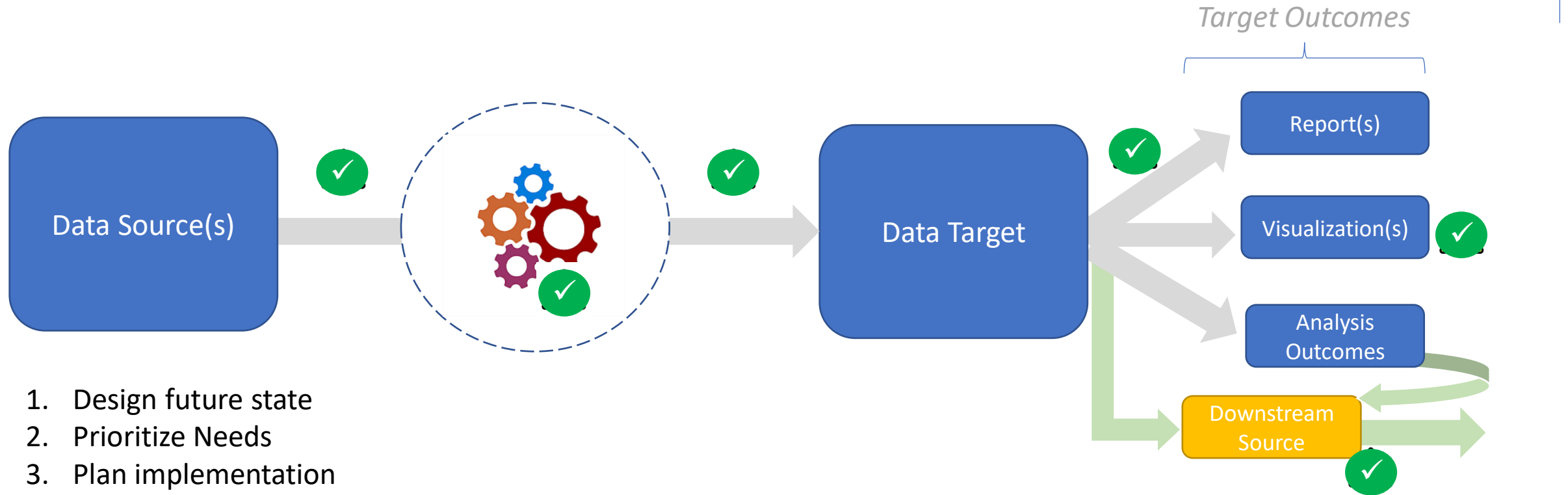
Data Pipeline - Documentation

Evaluation for the Future State
Identify Pain-points/Opportunities for Improvement



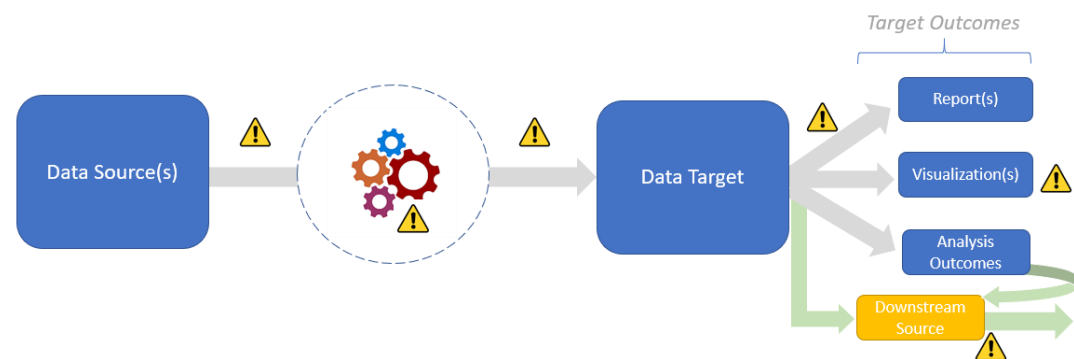
Data Pipeline - Documentation

Plan for the Future State

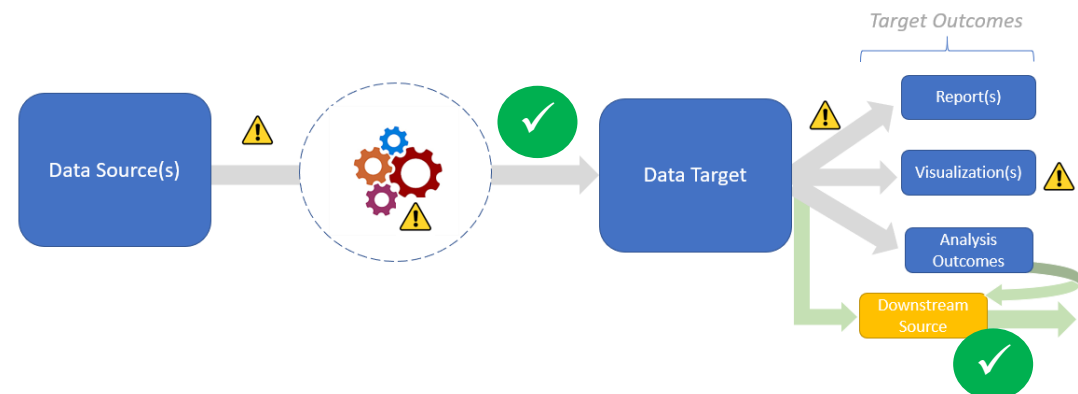


1. Design future state
2. Prioritize Needs
3. Plan implementation
 - Consider constraints (cost, training, tools, other initiatives)
4. Build & Test
 - Run major changes in parallel to validate changes
5. Update documentation
6. Repeat

Data Pipeline - Implementation



Phase 1 -



Phase 2 -



Discussion