Earthdata Cloud Analytics Project

Chris Lynnes* and Rahul Ramachandran*
NASA

*U.S. Civil Servant

Earth Observing System Data and Information System (EOSDIS)

data downlink

capture and clean

process

archive

subset

distribute

EOSDIS

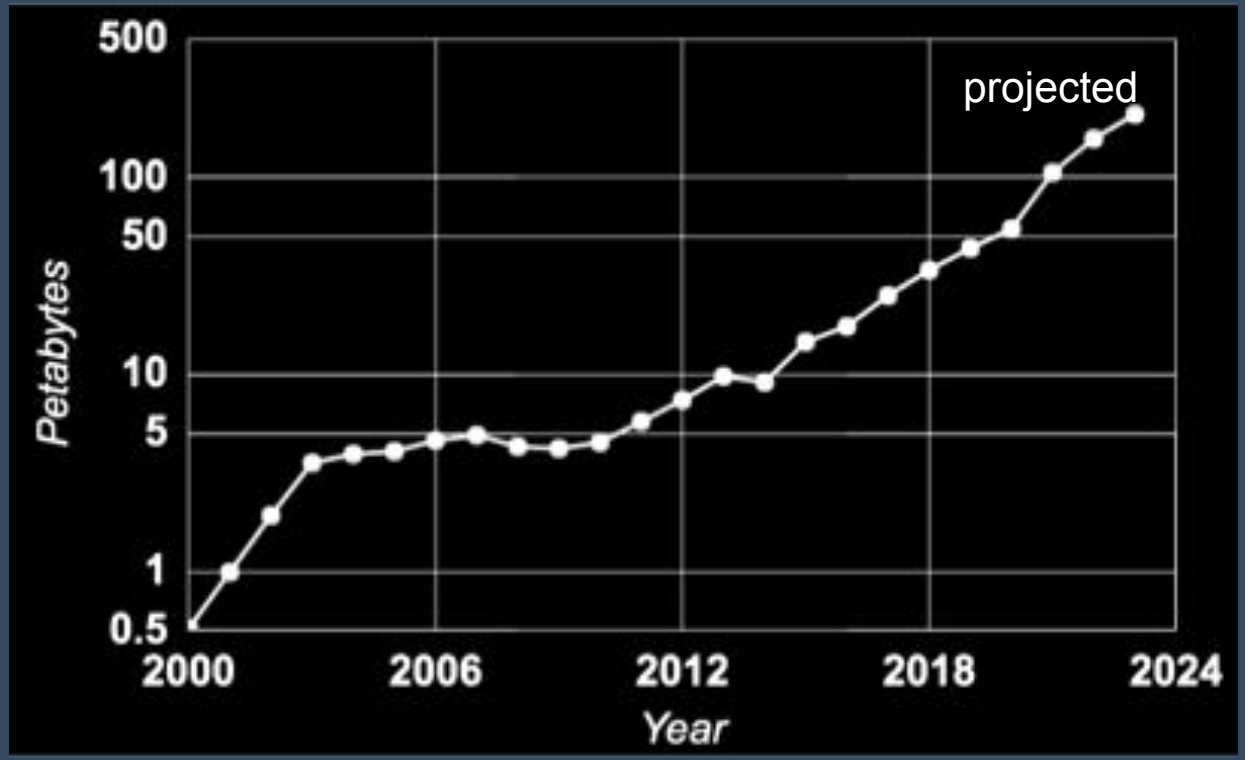
Research

Applications

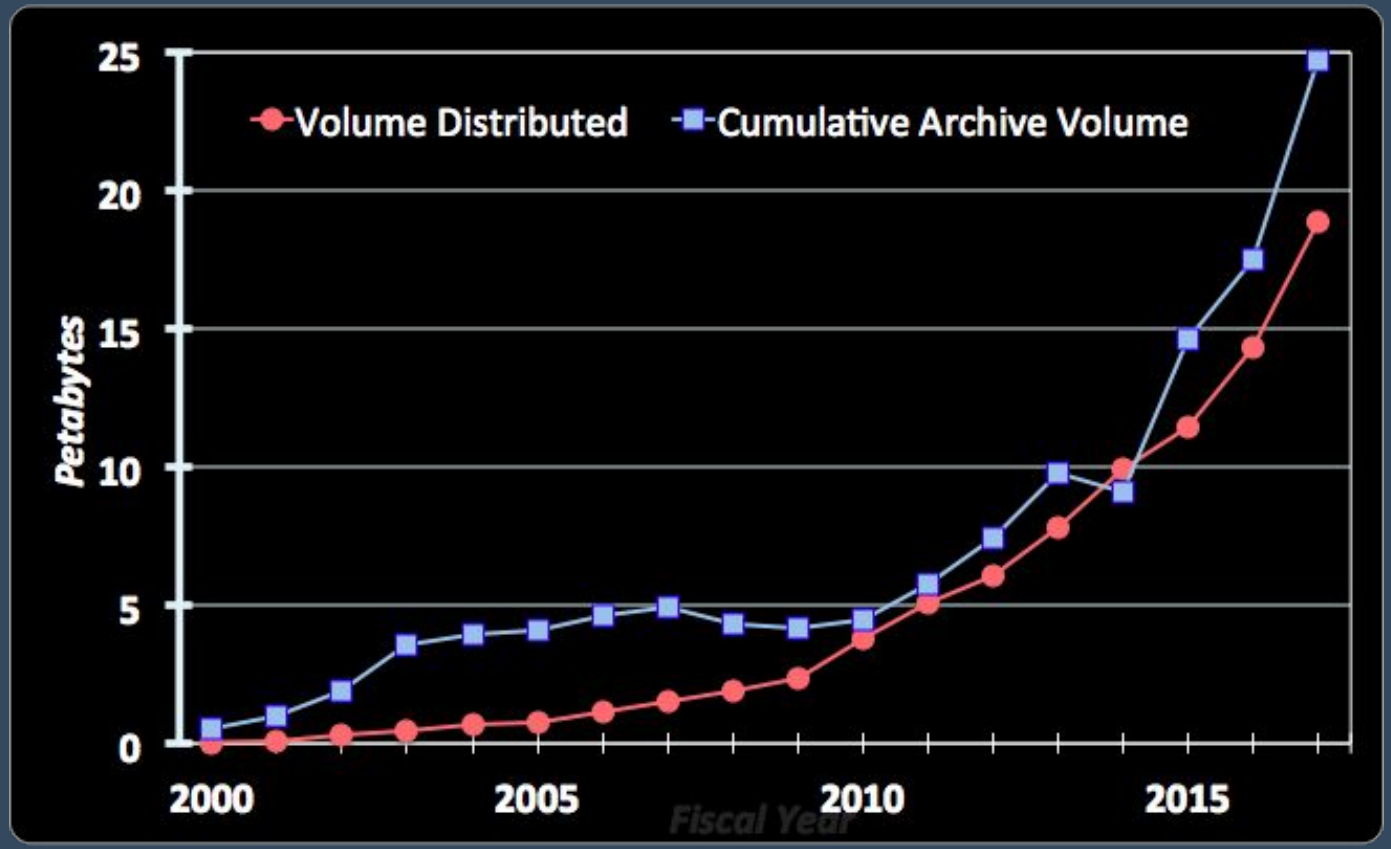
Education

Users

Over time, EOSDIS archive volumes increase exponentially

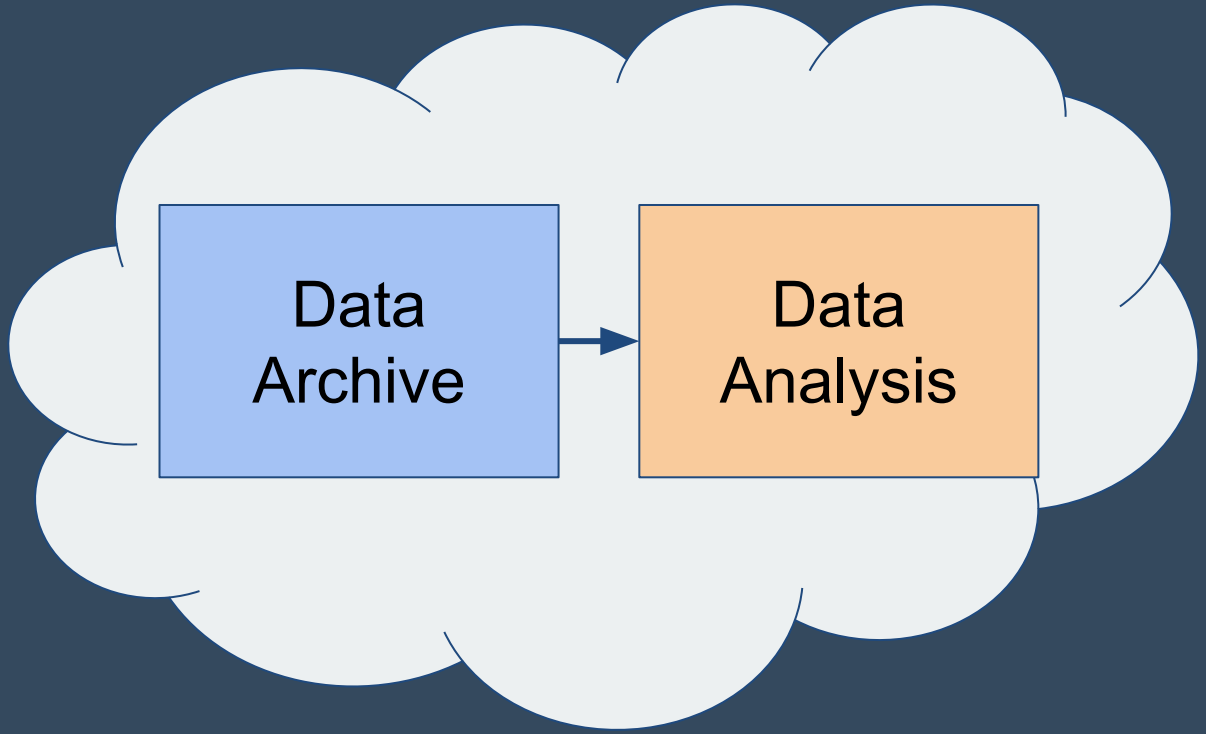


Distribution increases similarly to cumulative volume



How do we support user analysis of very large data volumes?

Solution: Data-proximal Analysis



Goals

1. Enable big compute next to big data
2. Encourage user adoption of cloud for analytics
3. Maximum analytics capability at minimum cost
    a. Use capabilities within NASA more effectively and efficiently
    b. Leverage analytics capabilities of external partners

Key Features

1. Satisfy a diverse user community
2. Support analysis in the cloud without egressing data
3. Facilitate multi-dataset comparison and fusion
4. Support batch, interactive and streaming modes
5. Support distributed filesystems and databases
6. Support cost constraints and cost-sharing

Earthdata Cloud Analytics Guiding Principles

1. Infusion- and innovation-friendly framework and building blocks
2. No monolithic systems
3. Open code and services
4. Interoperability and reuse
5. No unnecessary duplication (“undifferentiated heavy lifting”)

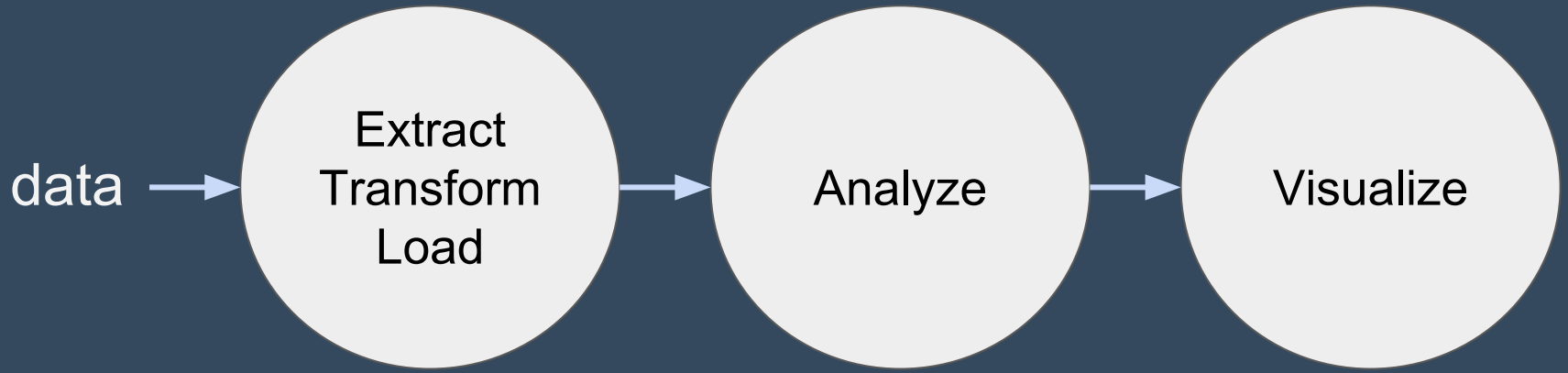
Architectural Concept

Earth Science Data Analytics the Cloud-Native Way:

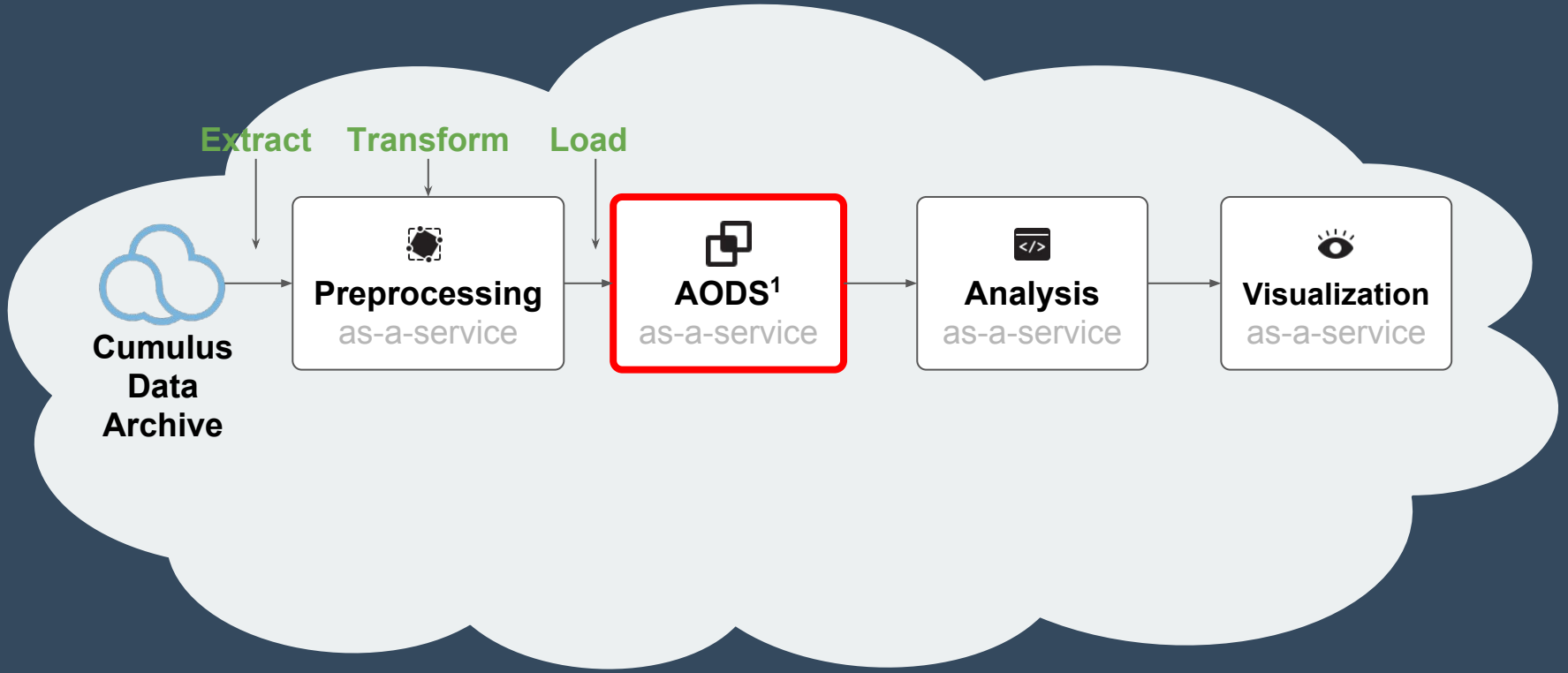
Everything is a Service

This approach produces key important benefits for the user community and EOSDIS

Abstract Analytics Workflow

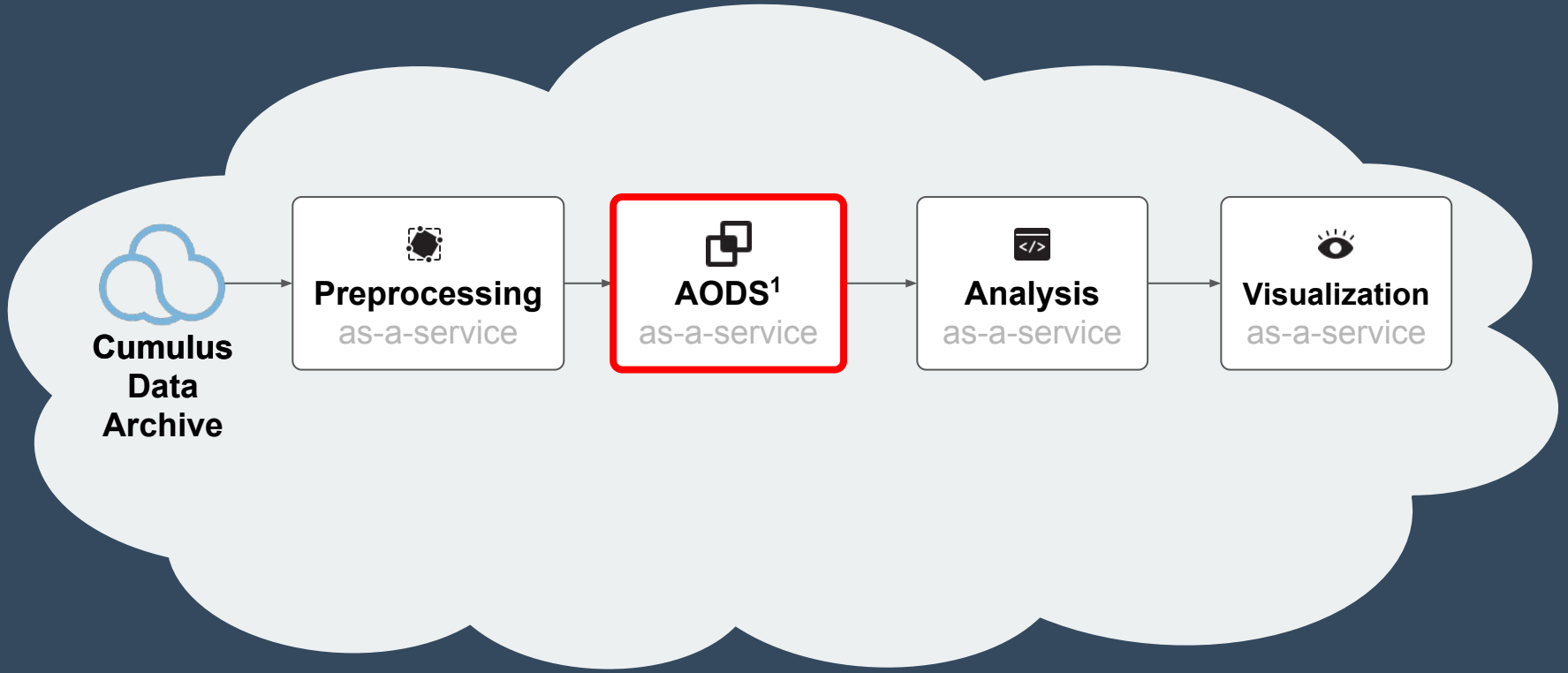


Earthdata Cloud Analytics Reference Architecture



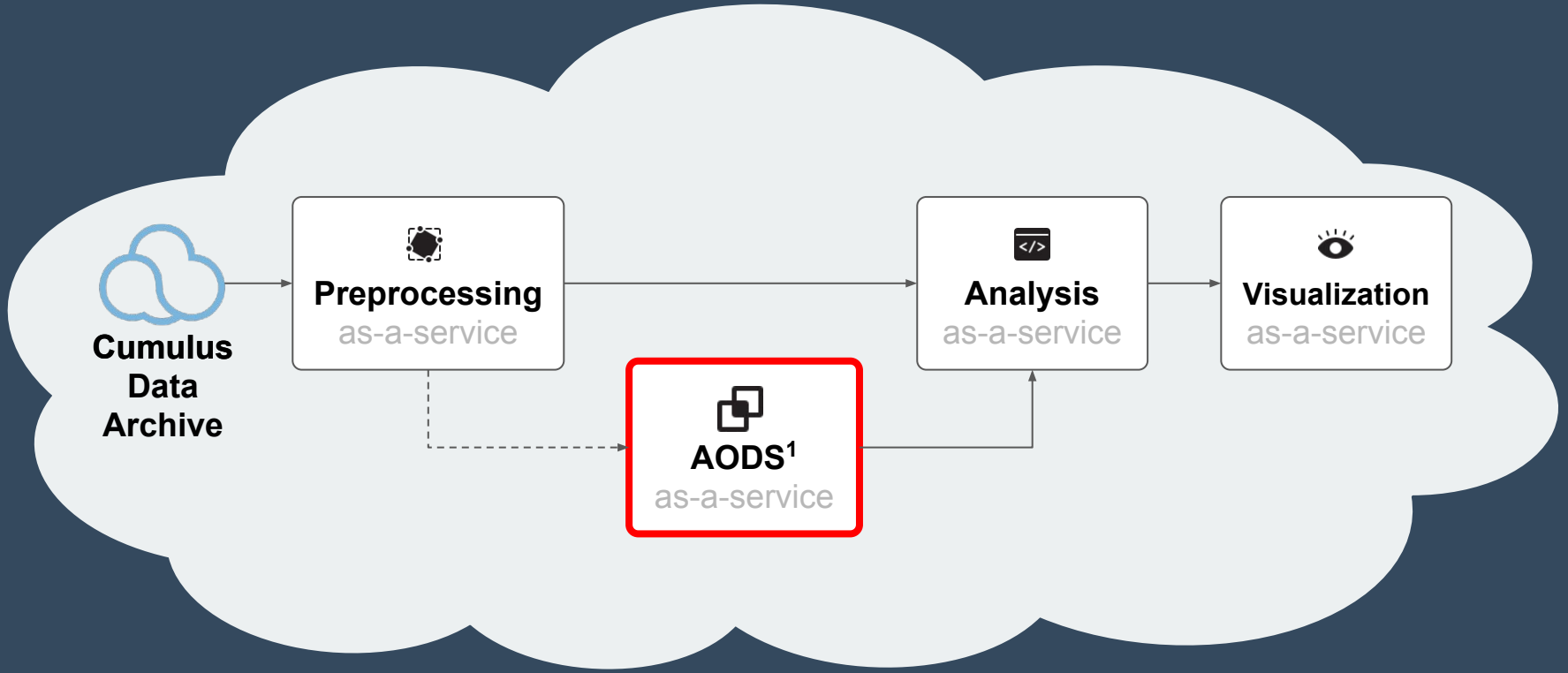
[1] Analytics Optimized Data Store

Interactive Mode: Analytics-Optimized Storage



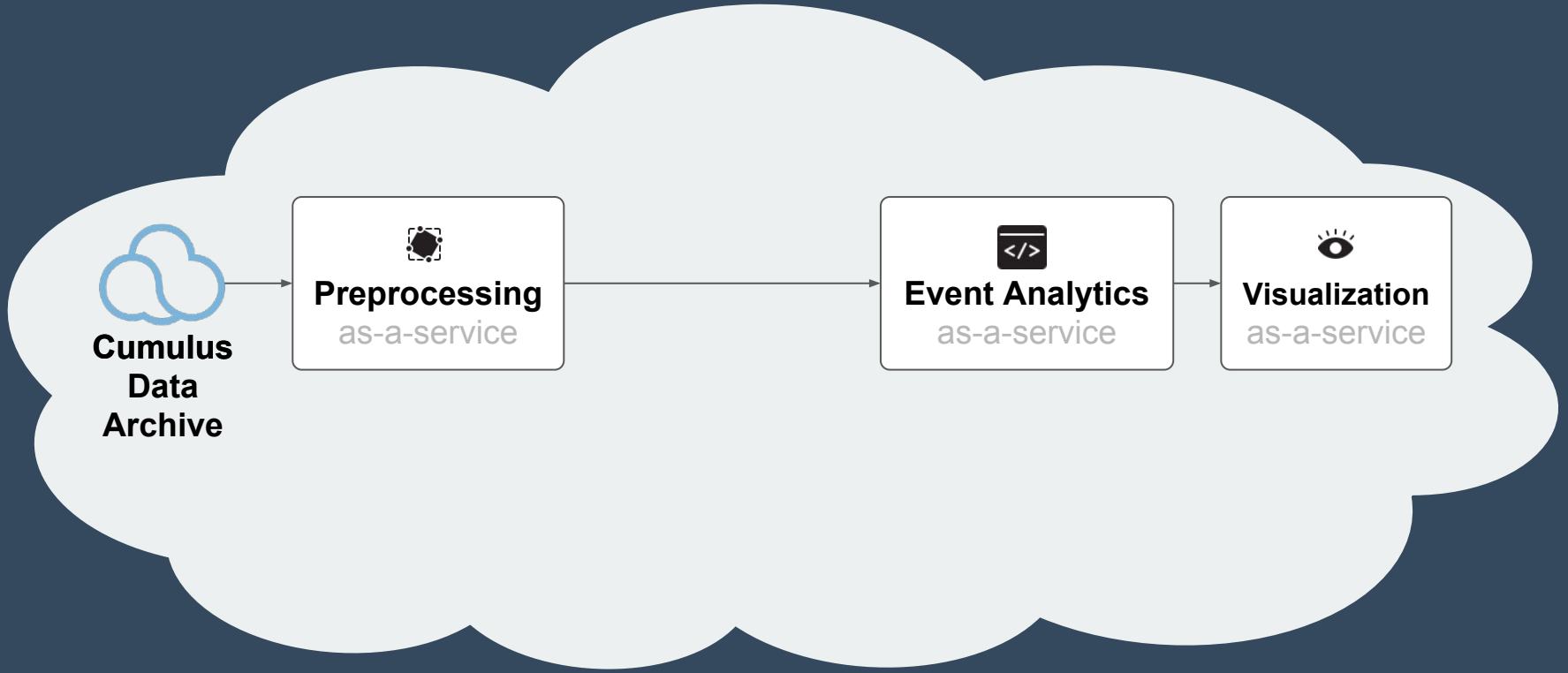
[1] Analytics Optimized Data Store

Batch Mode

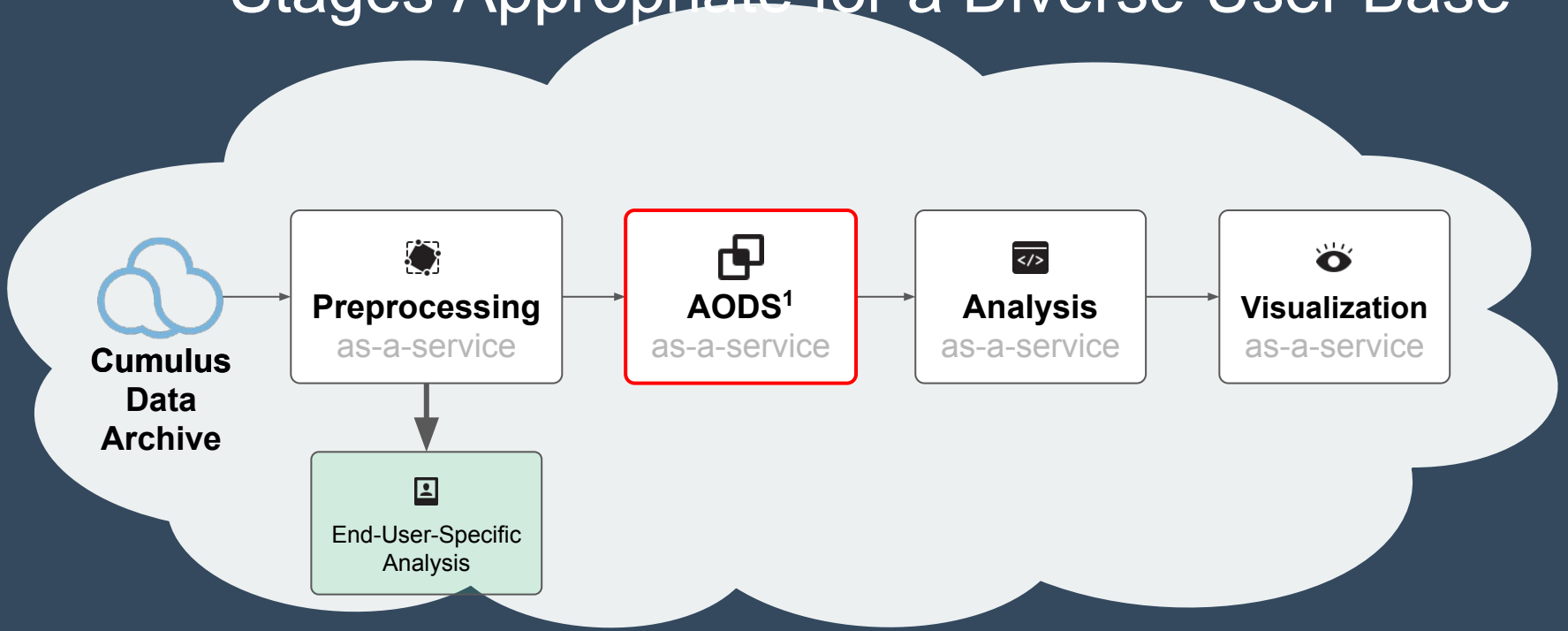


[1] Analytics Optimized Data Store

Streaming Mode

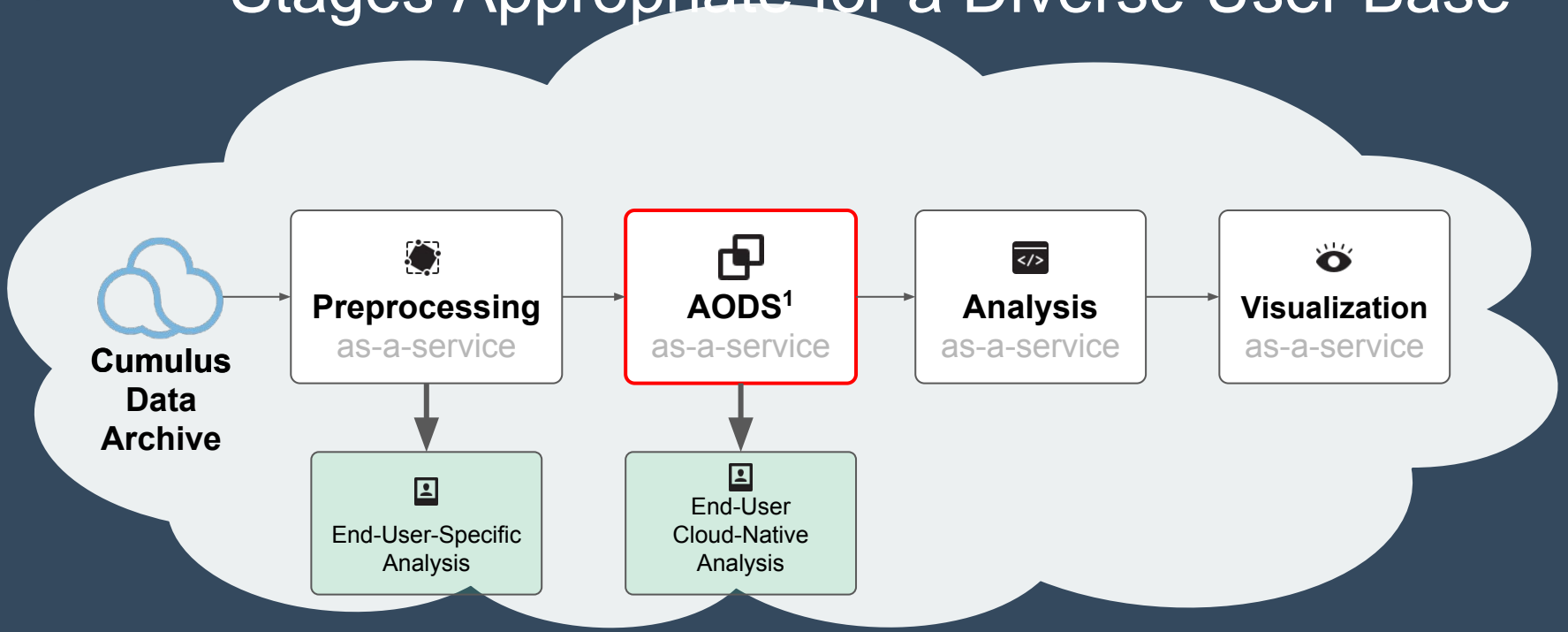


Open Pipeline Provides Outputs at Different Stages Appropriate for a Diverse User Base



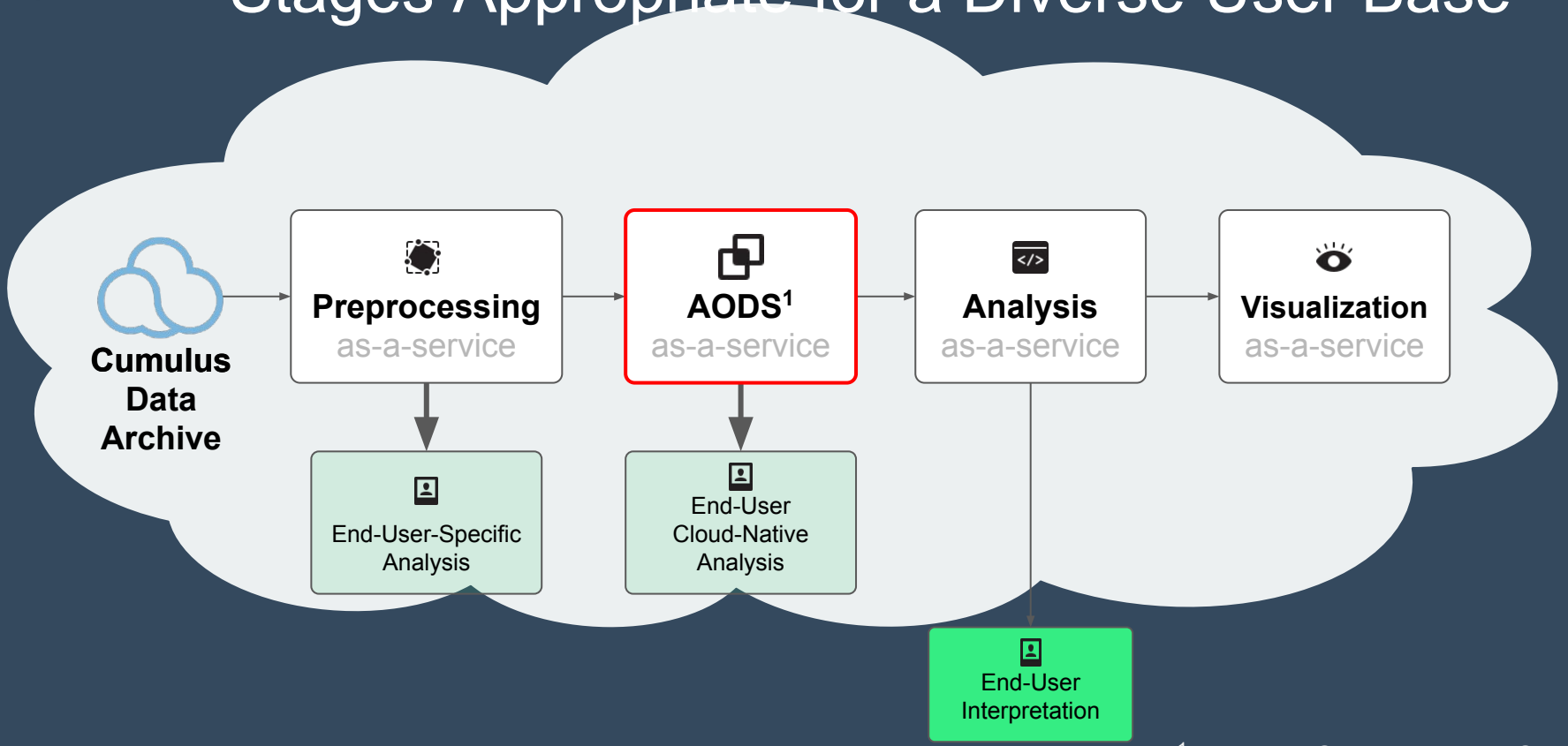
[1] Analytics Optimized Data Store

Open Pipeline Provides Outputs at Different Stages Appropriate for a Diverse User Base



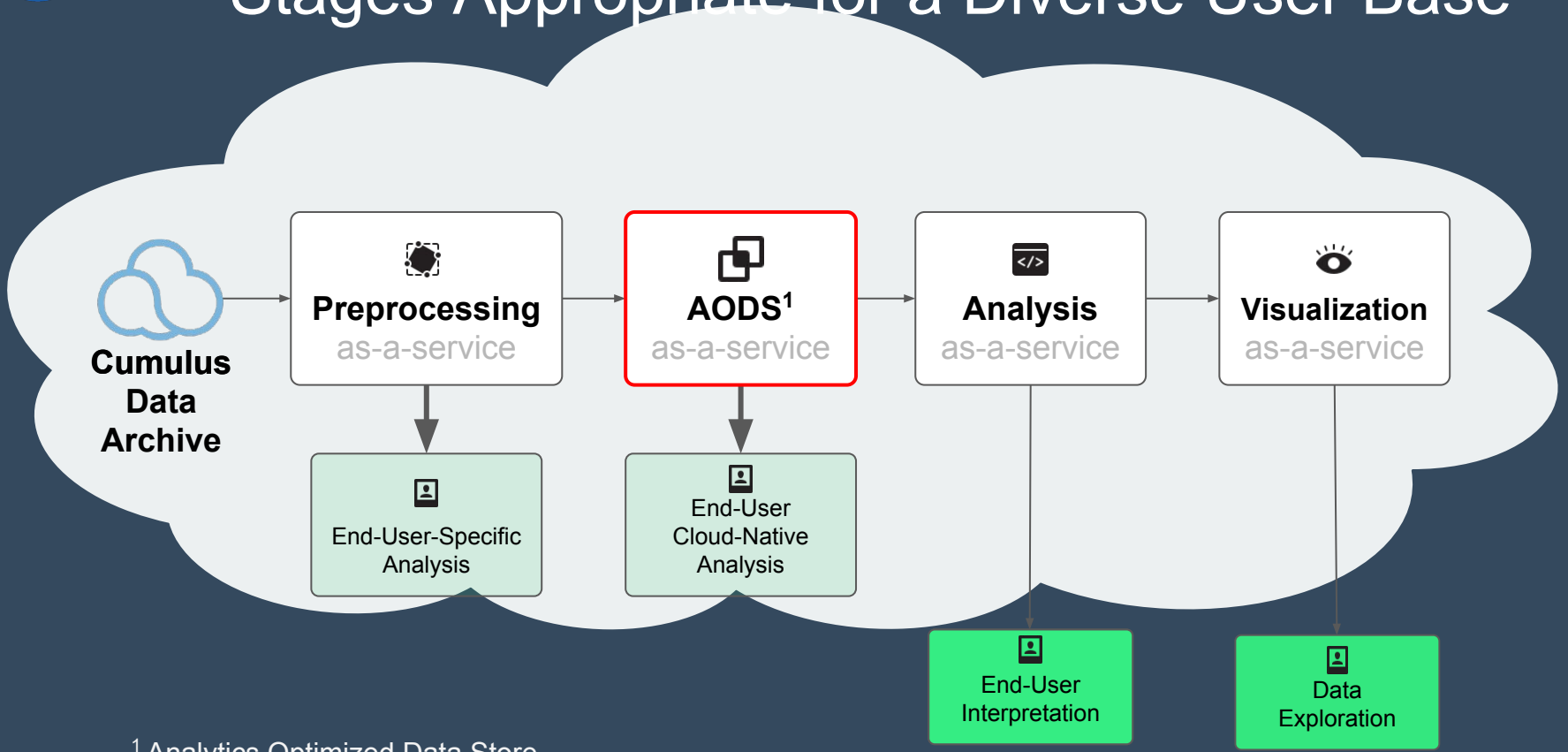
[1] Analytics Optimized Data Store

Open Pipeline Provides Outputs at Different Stages Appropriate for a Diverse User Base



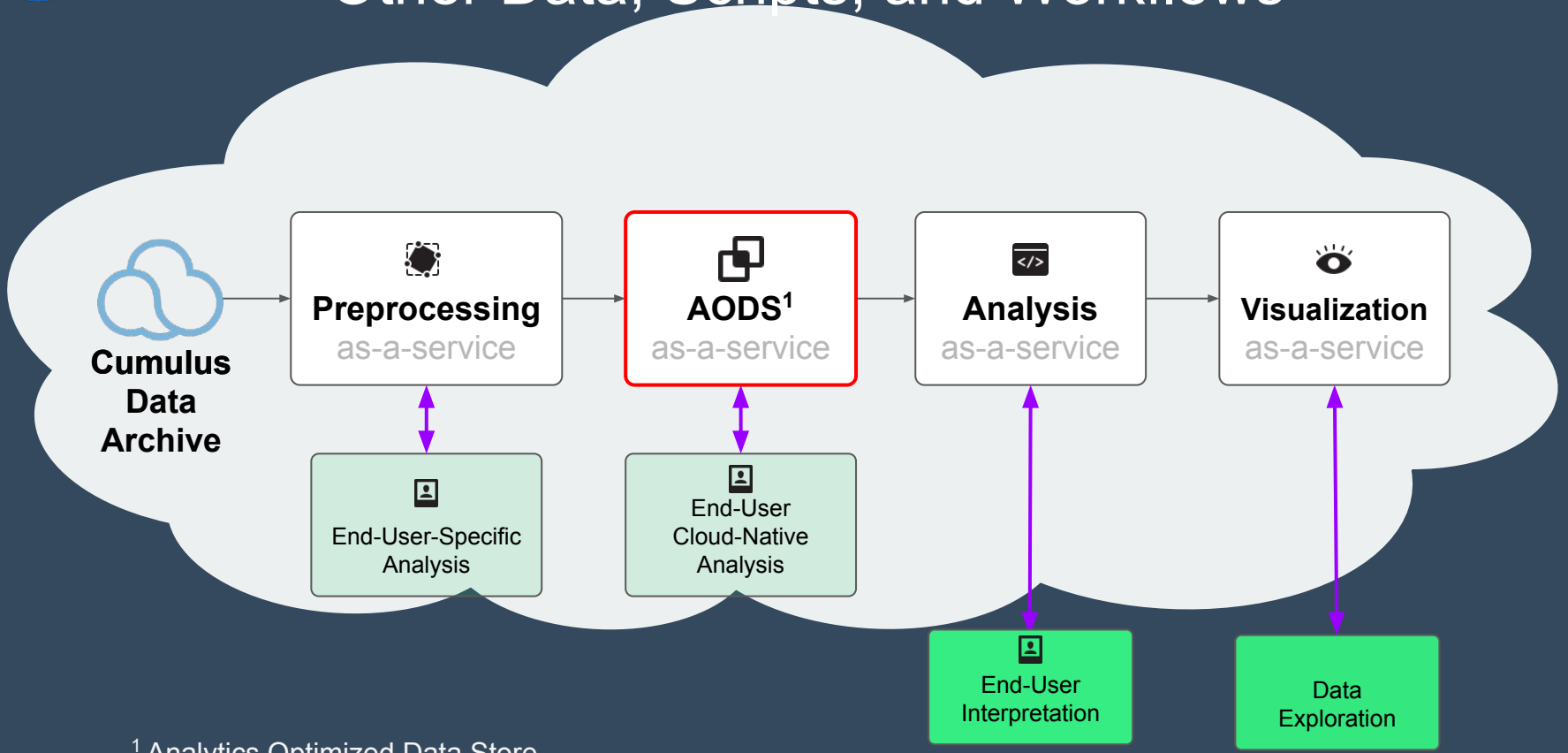
[1] Analytics Optimized Data Store

Open Pipeline Provides Outputs at Different Stages Appropriate for a Diverse User Base



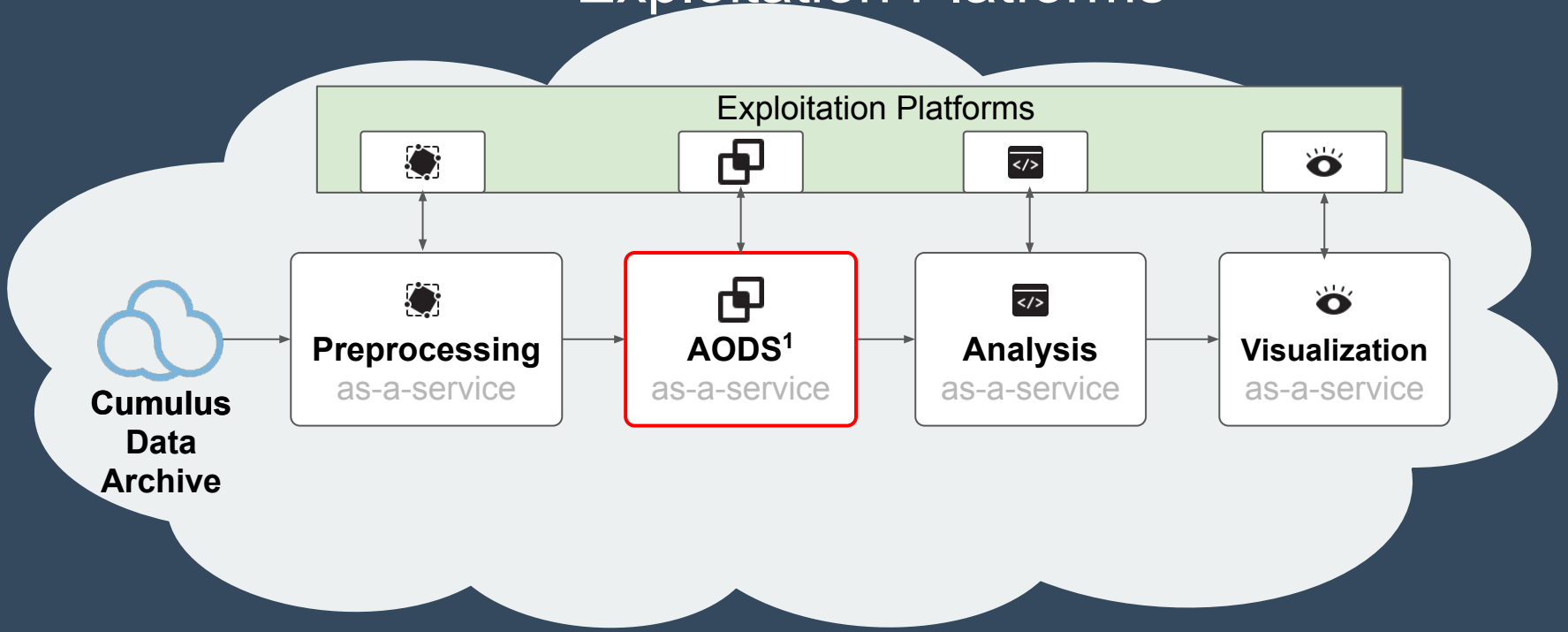
[1] Analytics Optimized Data Store

Open Pipeline Enables Integration with Other Data, Scripts, and Workflows



[1] Analytics Optimized Data Store

Open Pipeline Enables Integration with Exploitation Platforms



[1] Analytics Optimized Data Store